Significance testing for small annotations in stratified LD-Score regression

Katherine C. Tashman[1], Ran Cui[1], Luke J. O'Connor[1,3], Benjamin M. Neale[1,2], Hilary K. Finucane[1*]

1) Broad Institute, 75 Ames Street, Cambridge, MA 02142; 2) Analytic and Translational Genetics Unit, Massachusetts General Hospital, Richard B. Simches Research Center, 185 Cambridge Street, Boston, MA 02114 3) Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

* Correspondence to finucane@broadinstitute.org

Abstract

S-LDSC is a widely used heritability enrichment method that has helped gain biological insights into numerous complex traits. It has primarily been used to analyze large annotations that contain approximately 0.5% of SNPs or more. Here, we show in simulation that, when applied to small annotations, the block jackknife-based significance testing used in S-LDSC does not always control type 1 error. We show that the inflation of type 1 error for small annotations is due both to the noisiness of the jackknife estimate of the standard error and to the non-normality of the regression coefficient estimates. We use the percent of 0.01 centimorgan blocks in the genome overlapped by the annotation to quantify the size of an annotation and the extent to which the SNPs in the annotation cluster together, and we find thresholds on this value above which type 1 error is controlled. We have implemented a test in the LDSC software that informs users when they compute LD scores for an annotation if the annotation does not pass the threshold for producing controlled type 1 error.

Author Summary

Genetics is a rapidly evolving field that allows us to link our genetic code to the physiological manifestations of disease. A key part of this work is finding regions of the genome that contribute disproportionately to the genetic underpinnings of a disease. A commonly used tool to provide such insight is stratified LD score regression (S-LDSC). S-LDSC allows us to estimate how much a set of genomic regions contributes to the overall heritability of a phenotype, and to test whether this is more than we would expect by chance. Here we show that when we apply S-LDSC to a small set of genomic regions, it does not give an accurate test of whether this set of genomic regions contributes more than we would expect by chance to the phenotype. We characterize what it means to be a "small" set of genomic regions, and we set thresholds to restrict which annotations we test to prevent false positive results.This helps to ensure that as we continue to pursue genetic analyses at scale, we report only truly significant results that will help us further understand the etiology of many of the traits we study.

Introduction

Genome-wide association studies of complex traits have yielded thousands of associated variants; however, such results rarely point to conclusive biological mechanisms. A commonly used tool to gain insight into the biological underpinnings of these traits is enrichment analysis. Stratified LD score regression (S-LDSC)[5] is a widely used enrichment method that estimates the

heritability enrichment of a functional annotation, defined as the proportion of heritability explained by the annotation divided by the proportion of SNPs in the annotation. S-LDSC also estimates the contribution of an annotation to per-SNP heritability in a joint model with other annotations. S-LDSC uses the block jackknife[6], a commonly used statistical tool in population genetics[1,2,3,4], for standard error estimation and significance testing.

S-LDSC is typically applied to large annotations covering approximately 0.5% of SNPs or more. Here, we show in simulation that when S-LDSC is applied to small annotations, block jackknife-based significance testing does not always control type 1 error, especially for less polygenic traits. Specifically, we use the percent of 0.01 centimorgan blocks in the genome overlapped by the annotation as a way to quantify the size of an annotation and the extent to which the SNPs in the annotation cluster together, and we find thresholds on this value above which type 1 error is controlled and below which there can be inflation of type 1 error for the different statistical tests conducted by S-LDSC. We then show that for annotations that do not pass this threshold, the type 1 error inflation can be explained by a combination of non-normality of the test statistic and noisiness of the jackknife estimate of the standard error. We have implemented a test in the LDSC software that informs users when they compute LD scores for an annotation if the annotation does not pass the threshold for producing controlled type 1 error.

Methods

We implemented a simulation framework using 50,000 white British individuals from the UK Biobank and 9.3 million imputed variants to produce three sets of 1250 phenotypes with 200, 1000, and 10000 causal SNPs, respectively. We simulated heritable phenotypes, with $h^2 = 0.6$. The causal SNPs and their effect sizes were chosen independently of any annotation, so the

true value of τ_c was 0 for every annotation *c* except the base annotation containing all SNPs, and every simulated phenotype.

We created 94 annotations with varying size, varying number of jackknife blocks overlapped out of 200 total, and varying average segment length. To create the annotations, we randomly sampled either a set of genes or a percentage of SNPs from the genome from a specified number of jackknife blocks. For example, if we were to create an annotation of 100 genes that overlapped 30 blocks, we would randomly sample 100 genes from 30 of the default 200 jackknife-blocks such that at least 1 gene came from each block. We simulated annotations with 0.25, 0.1, 0.5, 1, 2, 3, or 5 % of SNPs, or 10, 30, 100, 200, 300, 450, or 600 genes, contained in 2, 4, 6, 8, 10, 30 , 60, 80 ,100, 150 or 200 jackknife blocks, excluding impossible combinations of parameters such as 5% of SNPs in 2 jackknife blocks. In addition to these simulated annotations, we included 100 real gene sets sampled from MSigDB as well as the 52 annotations in the baseline_v1.1 model[5], leading to a total of 246 annotations (Table S2).

For each simulated phenotype and each annotation, we ran S-LDSC with the annotation plus the baseline_v1.1 model and calculated both one-sided and two-sided p-values for $\hat{\tau}_C$ as well as two sided p-values for the heritability enrichment estimates. For each of these three tests, we assessed type 1 error control, in aggregate over annotations and simulated phenotypes, in three ways: first, using the proportion of rejections at P=0.05; second, defining a false positive as anything that passed significance after Bonferroni correcting for the 3750*246 hypotheses tested; and third, by visual inspection of a Q-Q plot.

Results

S-LDSC is a method for modeling the contributions of functional annotations to heritability using summary statistics. The method jointly models tens of overlapping annotations and reports two types of output: first, a joint-fit regression coefficient estimate for each annotation c, denoted $\hat{\tau}_c$, that quantifies the contribution of that annotation to per-SNP heritability; and second, the total heritability explained by SNPs in the annotation, including heritability attributable to other overlapping annotations. The former is typically used to identify phenotype-relevant tissues and cell types, while the latter is used to estimate heritability enrichment. S-LDSC uses the block jackknife to estimate standard errors and error covariance for the vector of estimates $\hat{\tau}_c$of τ_c. The standard errors are then used to compute z-scores to test either the null hypothesis that$\tau_c \leq 0$ (one-sided test) or the null hypothesis that $\tau_c = 0$ (two-sided test). To test the null hypothesis of no heritability enrichment -- i.e., to test the null hypothesis that the proportion of heritability in a category equals the proportion of SNPs in that category -- the null hypothesis is transformed into a linear condition on τ and the jackknife covariance of $\hat{\tau}$ is used to test the null hypothesis.

We used S-LDSC to test the null hypothesis of no heritability enrichment, the null hypothesis that$\tau_c \leq 0$ (one-sided test), and the null hypothesis that $\tau_c = 0$ (two-sided test) for each of 3,750 phenotypes simulated with no functional enrichment (true τ_c=0) and each of a set of 246 annotations including many annotations much smaller than is typical for S-LDSC input (see Methods).

Aggregated over all 3,750 simulated phenotypes and over all 246 annotations, our analyses showed controlled type 1 error at a cutoff of P=0.05 but an inflation of very small P-values. Specifically, 3.14% of results were significant at P=0.05, but 78 results passed Bonferroni correction for all annotations and all phenotypes (P < 0.05/(3750*246)) and so were called false

positives. Q-Q plots showed extreme inflation of type 1 error for both the one-sided test of $\tau_c \leq 0$ and the two-sided test of $\tau_c = 0$, and very mild inflation of type 1 error for the two-sided test of heritability enrichment =1 (Figure 1). While results were mostly consistent across polygenicities (Figure S1), we observed more false positives for the two least polygenic settings than for the more polygenic setting for the two-sided test (Figure S2). Restricting to the annotations of the baseline model, we found controlled type 1 error for all three tests (Figure S3).

We anticipated that type 1 error control for an annotation would depend on both the number of variants in an annotation and the extent to which they cluster together. Thus, we characterized each annotation using six metrics: (1) the percent of SNPs in the annotation; (2) the percent of 0.01 centimorgan (cM) blocks of the genome overlapped by the annotation; (3) the percent of 0.1 cM blocks of the genome overlapped by the annotation; (4) the percent of 1 cM blocks of the genome overlapped; (5) the percent of jackknife blocks (average size = 18 cM) overlapped by the annotation; and (6) the effective number of independent SNPs[10] in the annotation, defined as $\frac{|C|^2}{\sum_{j \in C} l(j,C)}$ where $l(j, C)$ is the LD score of SNP *j* to annotation *C*.

For each characterization metric except for number of jackknife blocks overlapped, we found that the most significant P-values tended to occur for annotations with smaller values of the metric: annotations with false positives, on average, had fewer SNPs, overlapped fewer 0.01, 0.1, and 1 cM blocks, and had a smaller effective number of independent SNPs. Specifically, for each annotation, we computed the minimum P-value obtained with the annotation for any of the simulated phenotypes, and for each characterization metric we plotted these minimum P-values across all annotations against the value of the metric (Figure 2a, S4). We then computed the mean value of each metric within annotations with false positives and among all annotations, and found that the mean value of the metric was smaller for annotations with false positives than for all annotations for each of the three statistical tests and five of the six metrics

considered (all except for number of jackknife blocks overlapped). Overall, the two-sided test of τ_c =0 had worse type 1 error inflation than the other two tests, and the inflation was captured less well by the six metrics than the other two tests.

We then evaluated each metric as a diagnostic tool that could be used to exclude annotations with inflated type 1 error. For each characterization metric and each of the three tests, we found the threshold such that excluding all annotations with values of the metric below the threshold would exclude all false positives. We then counted the number of remaining annotations; more remaining annotations indicates that the characterization metric would make a more specific diagnostic tool (Figure S5). We found that results differed for different characterization metrics and for the three tests, but that the percent of 0.01 cM blocks overlapped was a good metric for all three tests (number of remaining annotations = 132, 93, and 161 for the one-sided, two-sided, and enrichment test at a threshold of 1.7, 4.9 and 0.83 percent of 0.01 centimorgan blocks overlapped, respectively.) For the one-sided test and the heritability enrichment test, these thresholds restored type 1 error control. For the two-sided test, a more stringent threshold of 8.3% was needed to restore type 1 error control. We note that the annotations of the baseline model all pass the thresholds for the one-sided and enrichment tests, which are the two most commonly used tests. Because we have shown that type 1 error is controlled for all three tests when restricting to the baseline annotations (Figure S3) we do not recommend excluding annotations from the baseline model even when performing the two-sided test.

Having found that excluding annotations that overlap a small percentage of 0.01 cM blocks of the genome restores type 1 error control, we next sought to understand the source of type 1 error inflation for the annotations that do not pass this threshold. To do this, we focused only on the one- and two-sided test for $\hat{\tau}_c$,and we considered the z-score of the coefficient used to test for significance in these cases. The z-score of the coefficient is equal to the estimate of the

coefficient divided by the estimated standard error. This z-score will have the correct null distribution if the estimate of the coefficient is unbiased and normally distributed, and if the estimated standard error is equal to the true standard deviation of the estimate. We found that the coefficient estimate and standard error estimate were both approximately unbiased (Figure S6), and so we focused on noisiness of the standard error estimate and non-normality of the coefficient estimate as potential explanations for the type 1 error inflation.

To investigate the effect of the noisiness of the standard error estimate on type 1 error inflation, we replaced the jackknife estimate of the standard error in the denominator of the z-score with the true standard deviation of $\hat{\tau}_C$ over simulations. This had the effect of increasing significance for simulations for which the standard error was overestimated and decreasing significance for simulations for which the standard error was underestimated. Overall, this reduced type 1 error inflation by a small amount. However, there was still severe inflation in type 1 error even with this correction (Figure 3a,b). Moreover, while in the original analysis, most false positives for the two-sided test were for negative $\hat{\tau}_C$(2492/2591), after correcting the standard error, most false positives were for positive $\hat{\tau}_C$(723/724; Figure S7).

Having found that correcting the standard error did not suffice to control type 1 error, we then investigated non-normality of the coefficient estimates as a potential explanation for the type 1 error inflation. To do this, for each annotation, we transformed the estimates of $\hat{\tau}_C$to be normally distributed, preserving the standard deviation. Specifically, for each annotation we first chose 3,750 random samples from a normal distribution with mean zero and standard deviation matching the standard deviation of the $\hat{\tau}_C$for that annotation; we then quantile transformed the values of $\hat{\tau}_C$for the annotation to these values. We used the transformed $\hat{\tau}_C$and jackknife standard errors to compute P-values. Overall, this quantile normalization exacerbated the type 1

error (Figure 3a,c, Figure S8a,c). In contrast to the standard error correction, most false positives were for negative $\hat{\tau}_C$both before and after quantile normalization (Figure S9). While neither standard error correction nor quantile normalization sufficed to restore type 1 error control, when applied together they did (Figure 3d, Figure S8d).

We were surprised to find that correcting the standard error left severe inflation of type 1 error and that quantile normalizing $\hat{\tau}_C$in fact exacerbated type 1 error, but that performing both corrections simultaneously restored type 1 error control. Investigating this further, we found two related phenomena. First, for many annotations there was a positive correlation across independent simulated phenotypes between $\hat{\tau}_C$and the standard error estimate, and this correlation tended to be higher for the two least polygenic sets of phenotypes than for the most polygenic set of phenotypes (Figure S10). This correlation resulted in higher standard error estimates and thus more conservative p-values when the coefficient estimate was positive, and lower standard error estimates and thus less conservative p-values when the coefficient estimate was negative. The second phenomenon we observed was that for most annotations that were non-normal ($P < 0.05$ using a Kolmogorov-Smirnov test for normality), the distribution of $\hat{\tau}_C$ had a right skew (mean > median for 183 out of 191 annotations).

Together, these two phenomena explain several aspects of our results. Because the original distribution of $\hat{\tau}_C$tended to be right-skewed, quantile normalizing the $\hat{\tau}_C$distributions mostly reduced right skew and increased left skew, thus mostly increasing significance for the most significant negative $\hat{\tau}_C$. Because of the correlation between $\hat{\tau}_C$and jackknife standard error estimate, standard error correction had the opposite effect: it mostly increased the significance of positive $\hat{\tau}_C$ while decreasing the significance of negative $\hat{\tau}_C$. In the original analysis, the dominant contributor to type 1 error inflation was underestimated standard error for negative $\hat{\tau}_C$.

Quantile normalization exacerbated this problem and standard error correction ameliorated it; hence quantile normalization increased type 1 error inflation while standard error correction decreased it. After standard error correction, right-skewed non-normality became the main source of type 1 error, and most false positives were for positive $\hat{\tau}_C$.

Finally, we ran MAGMA[9], a commonly used gene set enrichment method, on our simulated phenotypes and gene-based annotations to determine whether it also resulted in inflated type 1 error. Aggregated across annotations and phenotypes, MAGMA had mildly inflated type 1 error (Figure S11).

Discussion

We applied a simulation framework to characterize the performance of the block jackknife-based significance testing used in S-LDSC. Using 3750 simulated phenotypes with varying levels of polygenicity, we ran S-LDSC on 94 simulated annotations, the baseline_v1.1 annotations, and 100 real gene sets sampled from MSigDB. For small annotations, we observed significant inflation in the reported one-sided and two-sided p-values for $\hat{\tau}_C$ but only mild inflation in the heritability enrichment p-values from S-LDSC. The inflation is due both to the noisiness of the jackknife estimate of the standard error, and to the non-normality of the regression coefficient estimates. This inflation can be remedied by restricting to annotations that overlap at least 1.7% of 0.01 cM blocks for standard analyses (i.e., for the one-sided test of $\hat{\tau}_C$ and the two-sided test of heritability enrichment) and 8.3% of 0.01 cM blocks for the two-sided test of $\hat{\tau}_C = 0$. We have implemented a test in the S-LDSC software to warn the user when their annotation does not meet the criteria required to produce statistically valid p-values. For small and/or clustered gene sets, MAGMA may provide better type 1 error control; we leave a thorough investigation of type 1 error for MAGMA to future work.

For small annotations, our simulations showed perfect type 1 error control at P=0.05 but a severe inflation of very small P-values. This highlights the need to assess type 1 error control of new methods by looking not only at a single fixed cutoff, but also by examining the tail. This is particularly important for tools such as S-LDSC that are often used to test a very large number of hypotheses, with stringent cutoffs for significance after multiple testing correction.

It is possible that the true null distribution of regression coefficient estimates for a given annotation could be derived or simulated and then used for hypothesis testing for small annotations. However, our results indicate the null distribution will depend on the genetic architecture of the trait being studied, presenting a challenge. Moreover, derivations will involve higher moments of the genotype matrix, leading to potential difficulties in both computation and reference panel mismatch. Since LD score regression controls type 1 error for annotations that are not very small, we propose here a simple restriction on input instead of a new method for significance testing.

We note three limitations of our work. First, the null simulations performed here are of heritable phenotypes with no enrichment in any functional annotation. We caution that, as noted in earlier work[5,8], model misspecification such as enrichment in a category not included in the model can also lead to bias and inflated type 1 error. Model misspecification is best addressed by fitting as flexible a model as possible, and will not be fixed by the threshold on 0.01 cM blocks overlapped introduced here. Second, we simulated a range of annotations including both gene sets and sets of random SNPs, and while we believe these annotations represent typical S-LDSC input, we cannot guarantee that our results extend to arbitrary annotations. Third, all annotations tested in this work are binary annotations. We leave characterization of type 1 error for continuous annotations, including those of the baseline-LD model[11], to future work.

In conclusion, S-LDSC produces well-calibrated p-values when annotations are large and spread throughout the genome, regardless of the level of polygenicity of the trait tested. We recommend performing standard S-LDSC analyses only on annotations that span at least 1.7% of 0.01 cM blocks of the genome, and performing two-sided tests for τ_c= 0 only on annotations that span at least 8.3% of 0.01 cM blocks.

Acknowledgements

We thank L. Abbott, S. Gazal, D. Palmer, A. Price, J. Ulirsch, R. Walters, E. Weeks and O. Weissbrod for helpful discussions. HKF is supported by NIH grant DP5 OD024582 and by Eric and Wendy Schmidt. This research was conducted using the UK Biobank Resource.

References

1. Zou J, Valiant G, Valiant P, Karczewski K, Chan SO, Samocha K, Lek M, Sunyaev S, Daly M, MacArthur DG. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. Nature communications. 2016 Oct 31;7(1):1-5.
2. Keinan A, Mullikin JC, Patterson N, Reich D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nature genetics. 2009 Jan;41(1):66.
3. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D. Genetic evidence for two founding populations of the Americas. Nature. 2015 Sep;525(7567):104-8.
4. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009 Sep;461(7263):489-94.
5. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, Ripke S. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics. 2015 Nov;47(11):1228.
6. Künsch HR. The jackknife and the bootstrap for general stationary observations. Annals of Statistics. 1989;17(3):1217-41.
7. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68-74.
8. Gazal S, Marquez-Luna C, Finucane HK, Price AL. Reconciling S-LDSC and LDAK functional enrichment estimates. Nature genetics. 2019 Aug;51(8):1202-4.
9. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015 Apr 17;11(4):e1004219.
10. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS genet. 2006 Dec 22;2(12):e190.
11. Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, Schoech A, Bulik-Sullivan B, Neale BM, Gusev A, Price AL. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature genetics. 2017 Oct;49(10):1421.

Figures

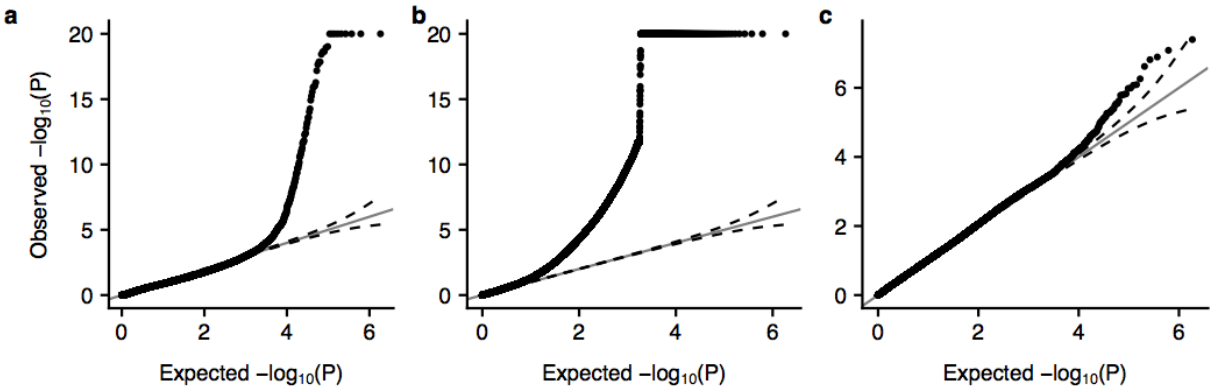


Figure 1: Results of null simulations, aggregated over annotations and simulated phenotypes. Q-Q plots for the **(a)** one-sided test of tauhat <= 0, **(b)** two-sided test of tauhat = 0, and **(c)** two-sided test of heritability enrichment.

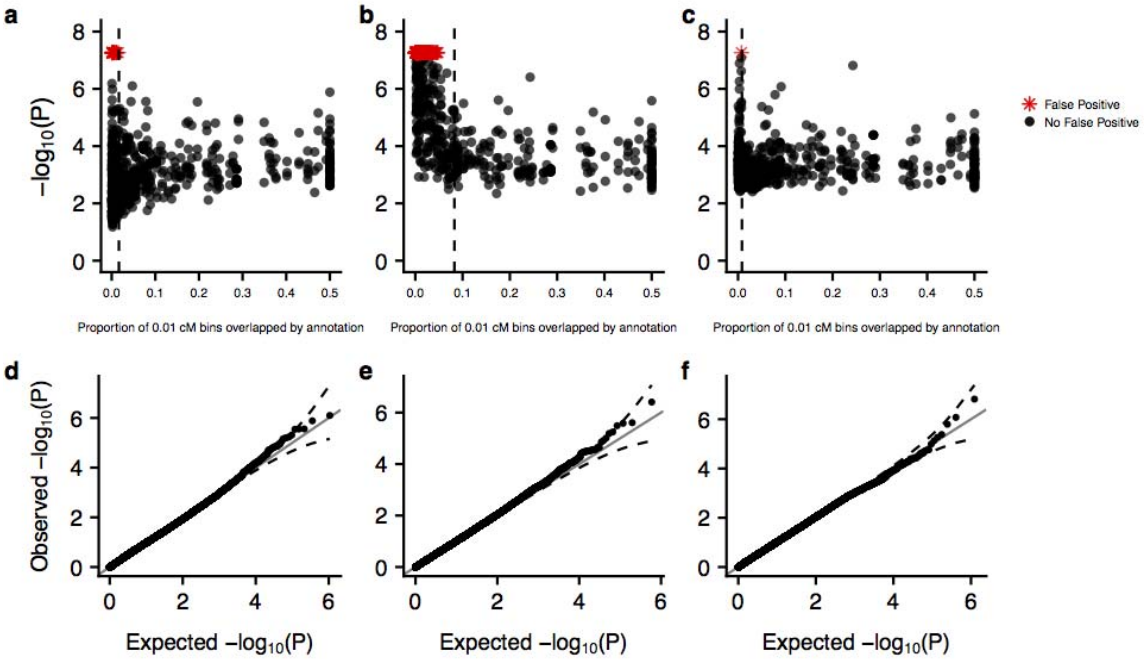


Figure 2: Control of type 1 error is restored by restricting to annotations overlapping sufficiently many 0.01 cM bins. (**a,b,c**) Dependence of type 1 error on the proportion of 0.01 cM bins overlapped by the annotation for the one-sided, two-sided, and enrichment test respectively. Each dot represents one of the 246 annotations tested. Any p-value below the false positive threshold of $P < 0.05/(3750*246)$ was set to that value for visualization and denoted with a red star. The dashed black line indicates the threshold used to recover a well controlled type 1 error. All annotations overlapping more than 50% of 0.01 cM bins were thresholded to 50% for visualization. (**d,e,f**) Q-Q plots for the one-sided, two-sided, and enrichment test respectively, restricting to annotations that pass the threshold depicted in **a**, **b**, and **c**, respectively.

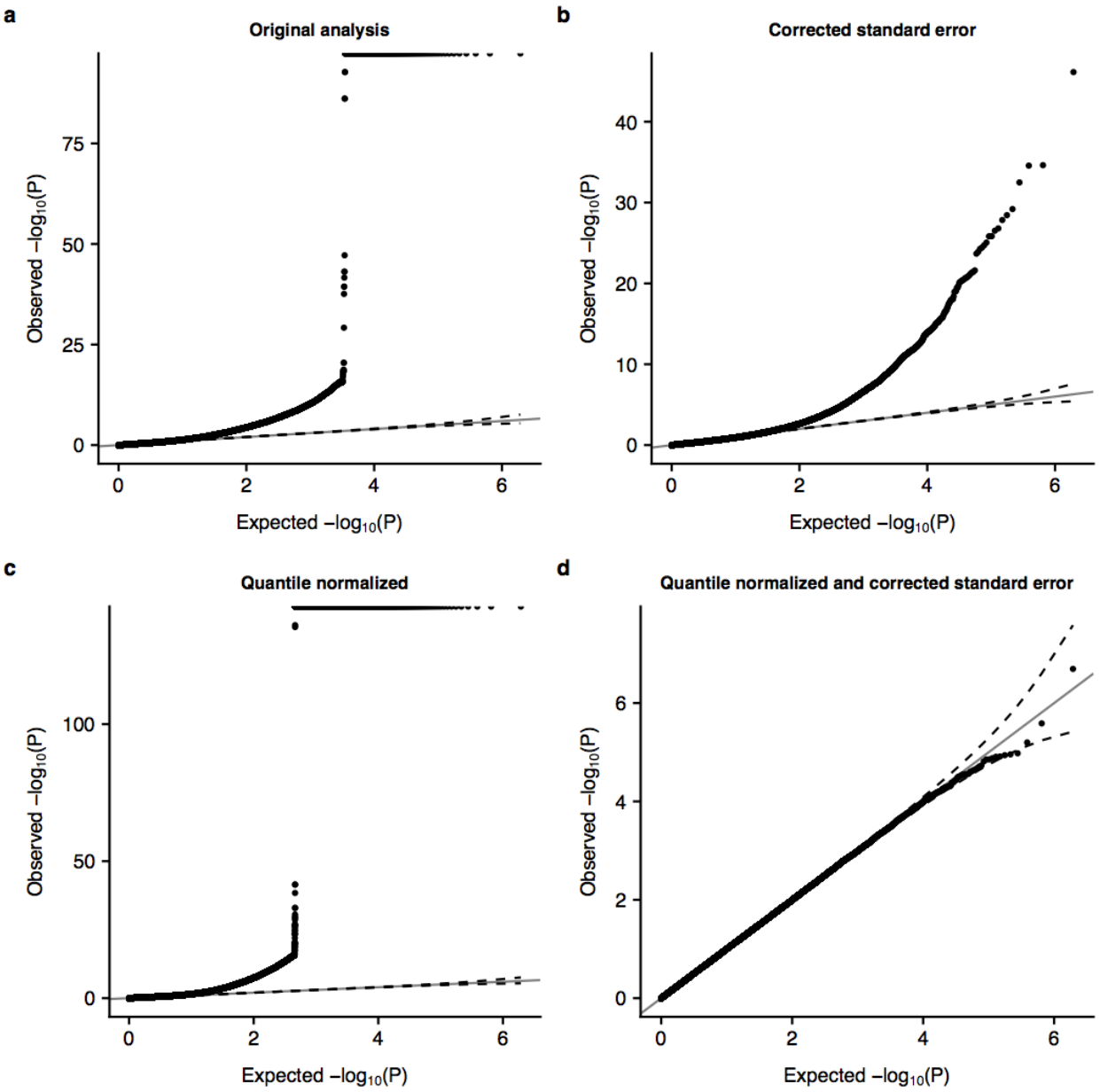


Figure 3: Dependence of type 1 error control on normality of and on the accuracy of the standard error estimate. (**a**) Q-Q plot for the two-sided test of using the jackknife standard error. (**b**) Q-Q plot for the two-sided test of using the corrected standard error. (**c**) Q-Q plot for the two-sided test of the quantile normalized using the jackknife standard error. (**d**) Q-Q plot for the two-sided test of the quantile normalized using the corrected standard error.

Supplemental Figures and Tables

Baseline_v1.1 Annotations	% of 1 cM blocks overlapped	% of 0.1 cM blocks overlapped	% of 0.01 cM blocks overlapped	Number of independent SNPs	% of jackknife blocks overlapped	Percent of SNPs
CTCF_Hoffman.bed	99	69	18	74145.34	100	2.39
CTCF_Hoffman.extend.500.bed	100	75	24	91283.43	100	7.10
DGF_ENCODE.bed	100	93	51	119180.74	100	13.81
DGF_ENCODE.extend.500.bed	100	97	71	132360.85	100	54.07
DHS_peaks_Trynka.bed	99	92	47	130260.51	100	11.20
DHS_Trynka.bed	99	93	52	134680.27	100	16.80
DHS_Trynka.extend.500.bed	99	96	69	138409.25	100	49.73
Enhancer_Andersson.bed	89	31	5	25973.04	100	0.44
Enhancer_Andersson.extend.500.bed	95	42	9	44440.30	100	1.92
Enhancer_Hoffman.bed	96	59	18	62839.74	100	4.30
Enhancer_Hoffman.extend.500.bed	97	64	23	73014.18	100	9.14
FetalDHS_Trynka.bed	99	89	40	125908.56	100	8.56
FetalDHS_Trynka.extend.500.bed	99	93	56	137199.48	100	28.52
H3K27ac_Hnisz.bed	100	87	53	104598.78	100	39.30
H3K27ac_Hnisz.extend.500.bed	100	87	55	109040.68	100	42.44
H3K27ac_PGC2.bed	100	85	45	101120.92	100	27.29
H3K27ac_PGC2.extend.500.bed	100	86	49	105940.71	100	33.97
H3K4me1_peaks_Trynka.bed	99	93	52	115076.83	100	17.35
H3K4me1_Trynka.bed	100	95	65	124321.84	100	42.89
H3K4me1_Trynka.extend.500.bed	100	96	73	131420.88	100	60.98
H3K4me3_peaks_Trynka.bed	99	79	27	62737.84	100	4.34
H3K4me3_Trynka.bed	100	86	38	67300.12	100	13.72
H3K4me3_Trynka.extend.500.bed	100	89	47	85438.90	100	25.97
H3K9ac_peaks_Trynka.bed	99	73	25	71828.14	100	4.00
H3K9ac_Trynka.bed	99	80	35	82599.93	100	12.91
H3K9ac_Trynka.extend.500.bed	99	84	43	93633.26	100	23.45
Intron_UCSC.bed	91	56	36	71275.16	100	39.43
Intron_UCSC.extend.500.bed	91	57	37	72571.37	100	40.48
PromoterFlanking_Hoffman.bed	94	37	7	34991.10	100	0.86
PromoterFlanking_Hoffman.extend.500.bed	97	48	12	52207.37	100	3.40
Promoter_UCSC.bed	91	38	10	43009.05	100	4.81
Promoter_UCSC.extend.500.bed	91	38	11	44298.41	100	5.89
Repressed_Hoffman.bed	100	90	58	85666.20	100	45.28
Repressed_Hoffman.extend.500.bed	100	92	68	95782.86	100	70.89
SuperEnhancer_Hnisz.bed	79	37	22	55186.19	100	16.98
SuperEnhancer_Hnisz.extend.500.bed	79	38	22	55752.33	100	17.30
TFBS_ENCODE.bed	99	90	45	105635.30	100	13.31
TFBS_ENCODE.extend.500.bed	100	93	58	123902.97	100	34.29
Transcribed_Hoffman.bed	100	94	57	88204.70	100	35.34
Transcribed_Hoffman.extend.500.bed	100	97	75	112670.96	100	76.43
TSS_Hoffman.bed	88	35	8	33298.51	100	1.88
TSS_Hoffman.extend.500.bed	90	39	10	41261.36	100	3.56
UTR_3_UCSC.bed	83	28	6	30205.45	100	1.19
UTR_3_UCSC.extend.500.bed	86	33	8	35388.45	100	2.79
UTR_5_UCSC.bed	82	26	5	20686.02	100	0.59
UTR_5_UCSC.extend.500.bed	89	37	9	35856.61	100	2.82
WeakEnhancer_Hoffman.bed	98	65	17	70218.57	100	2.14
WeakEnhancer_Hoffman.extend.500.bed	99	75	28	91264.02	100	8.95
Coding_UCSC.bed	84	35	9	41088.25	100	1.61
Coding_UCSC.extend.500.bed	86	43	15	51179.17	100	6.74
Conserved_LindbladToh.bed	96	73	24	100752.99	100	2.85
Conserved_LindbladToh.extend.500.bed	97	90	55	124487.84	100	33.66

Table S1: The annotation metrics for the baseline_v1.1 annotations.

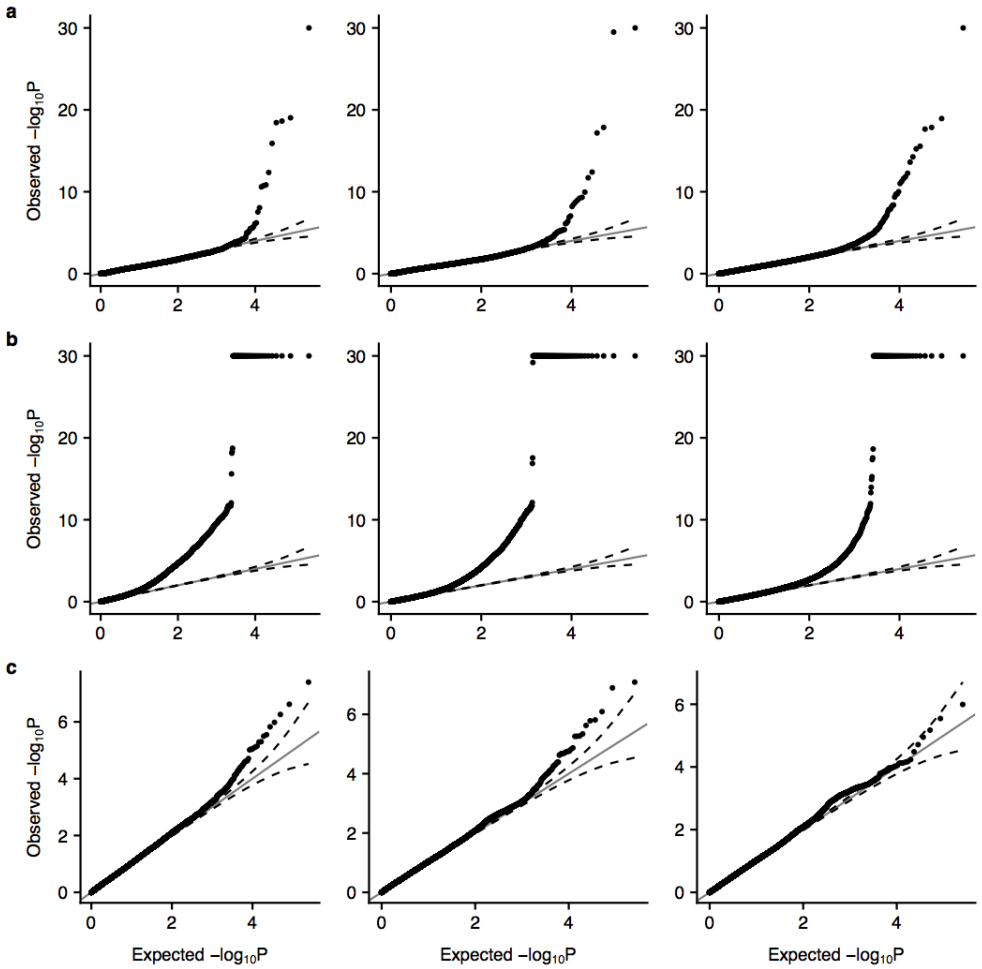


Figure S1: Results are consistent across polygenicities. (**a**) The results from the one-sided test of . (**b**) The results from the two-sided test of . (**c**) The results from the enrichment test. Polygenicity increases from left to right in each case.

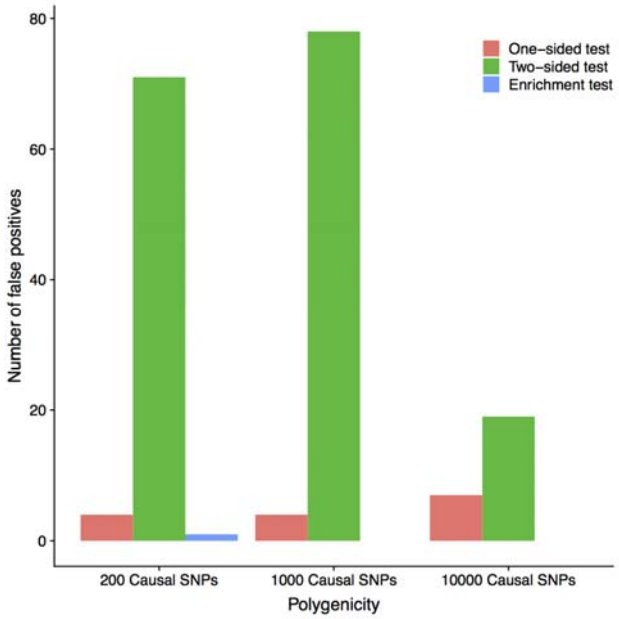


Figure S2: The number of false positives (P<0.05/(3750*246)) for the two-sided test is higher for the two less polygenic sets of phenotypes than for the most polygenic set, while the number of false positives for the one-sided test and enrichment test are stable across polygenicities.

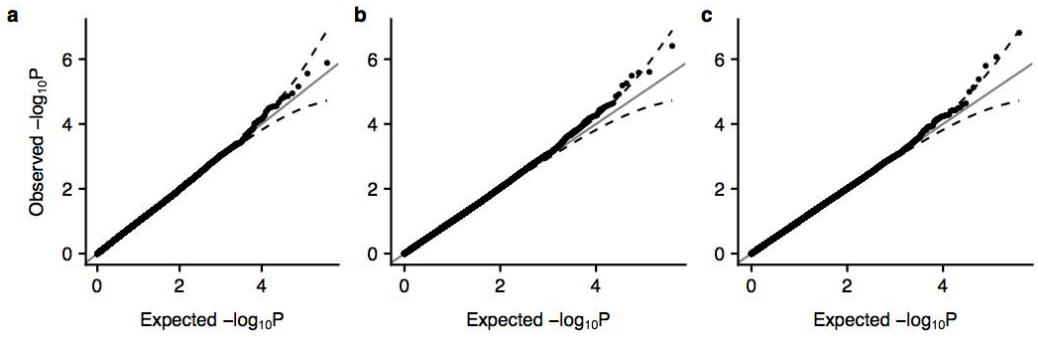
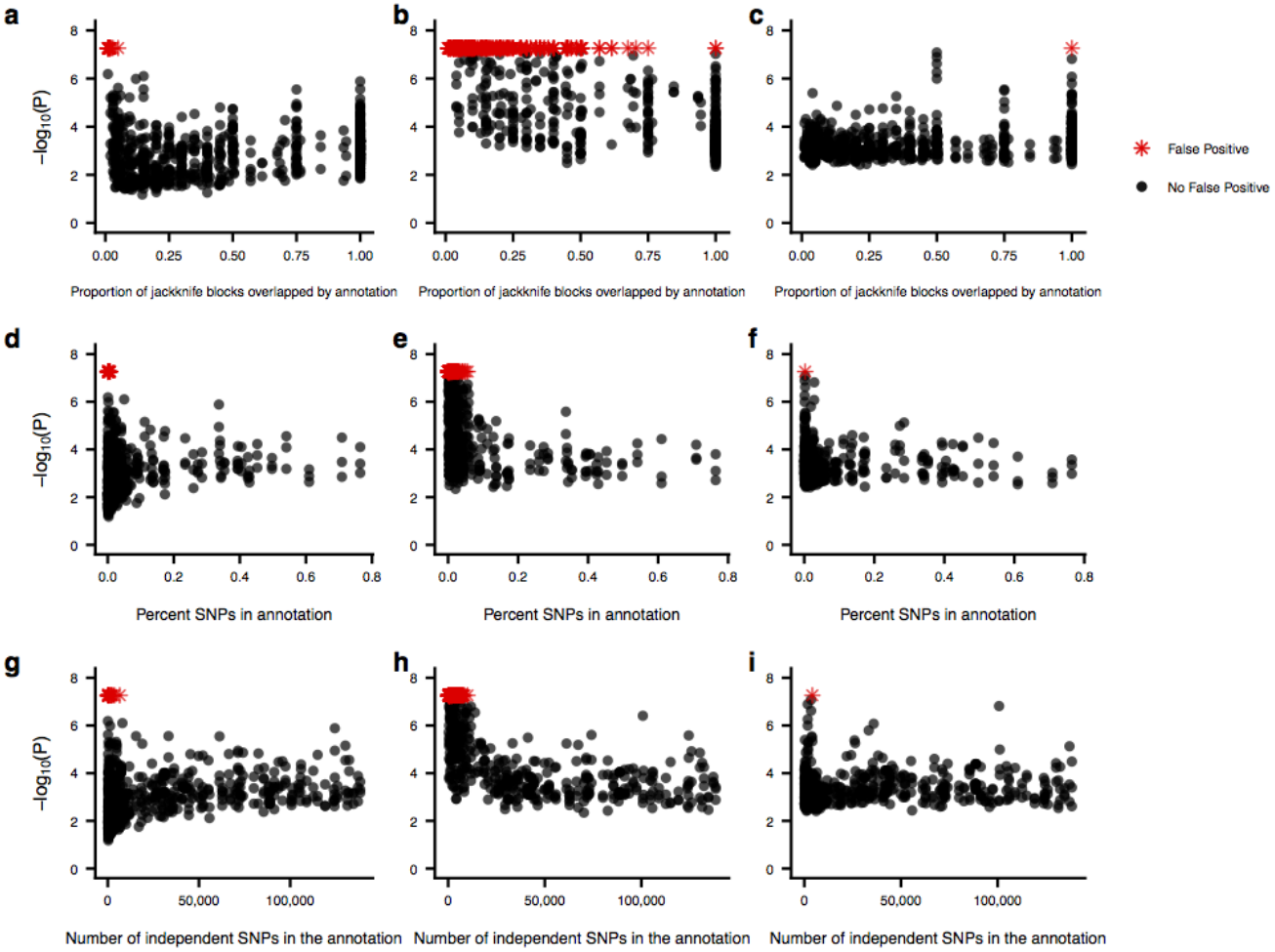


Figure S3: The baseline_v1.1 annotation results have well controlled type 1 error for the (**a**) one-sided, (**b**) two-sided and (**c**) enrichment results.



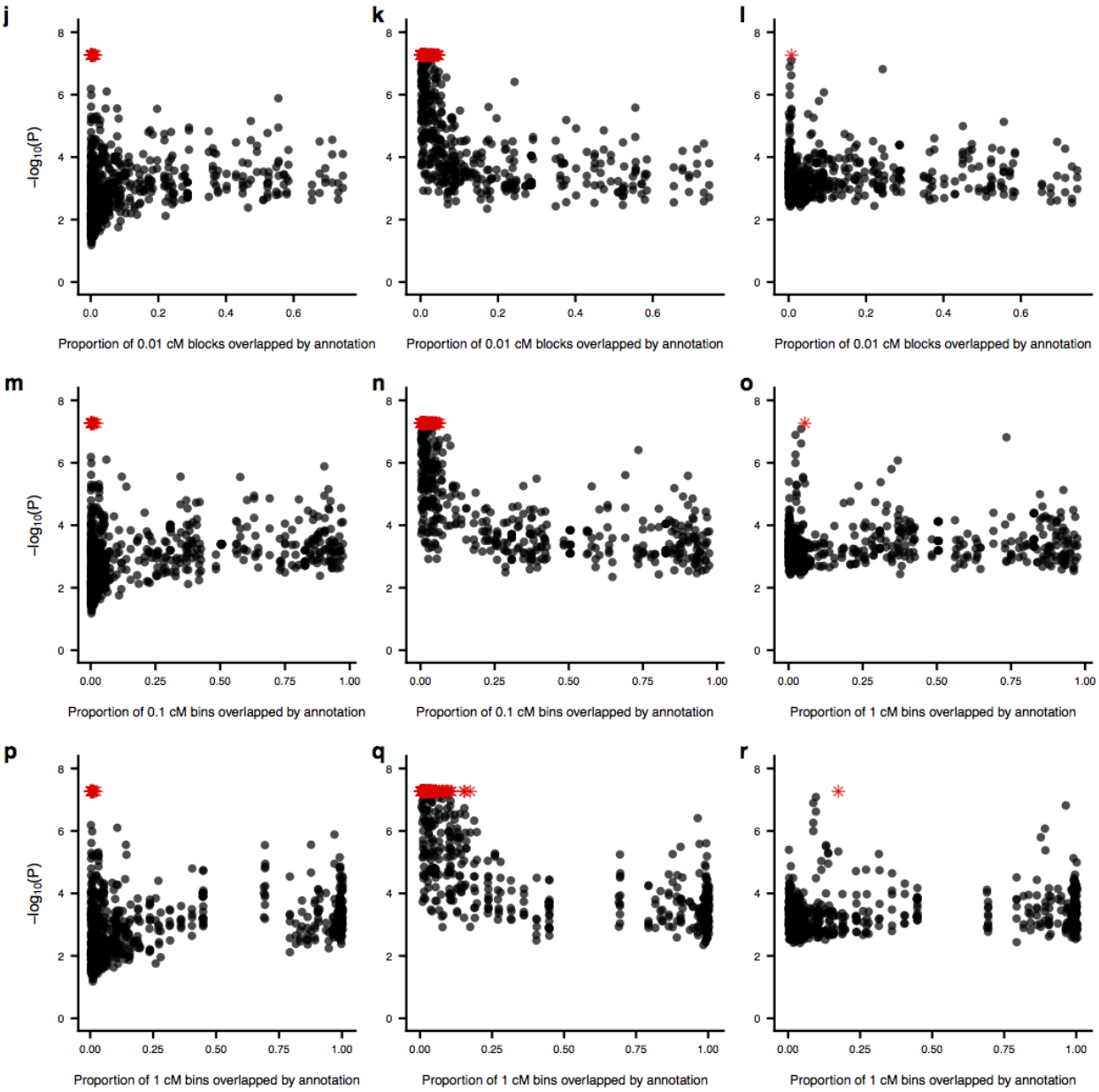


Figure S4: Relationship between significance and the annotation characteristics for the three tests. The annotation characteristics are: (**a-c**) proportion of jackknife blocks overlapped by the annotation, (**d-f**) percent of SNPs in the annotation, (**g-i**) number of independent SNPs in the annotation, (**j-l**) proportion of 0.01 cM bins overlapped by the annotation, (**m-o**) proportion of 0.1 cM bins overlapped by the annotation, (**p-r**) proportion of 1 cM bins overlapped by the annotation. The tests are **(a,d,g,j,m,p)** one-sided test, **(b,e,h,k,n,q)** two-sided test, and **(c,f,i,j,o,r)** test for heritability enrichment. False positives are denoted with red stars.

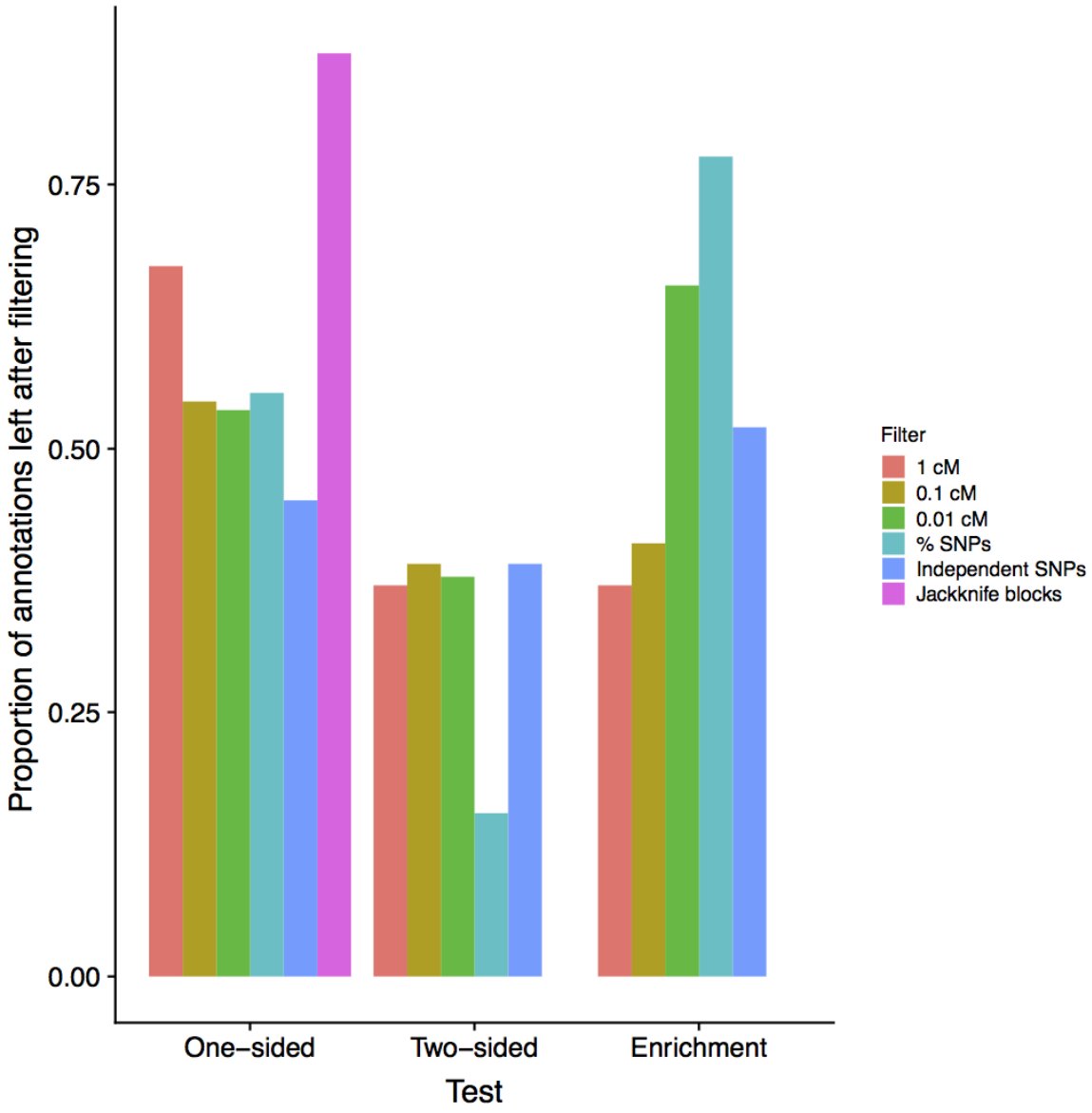


Figure S5: Proportion of annotations remaining after imposing a filter that excludes all false positives. Color denotes the annotation metric used to define the filter.

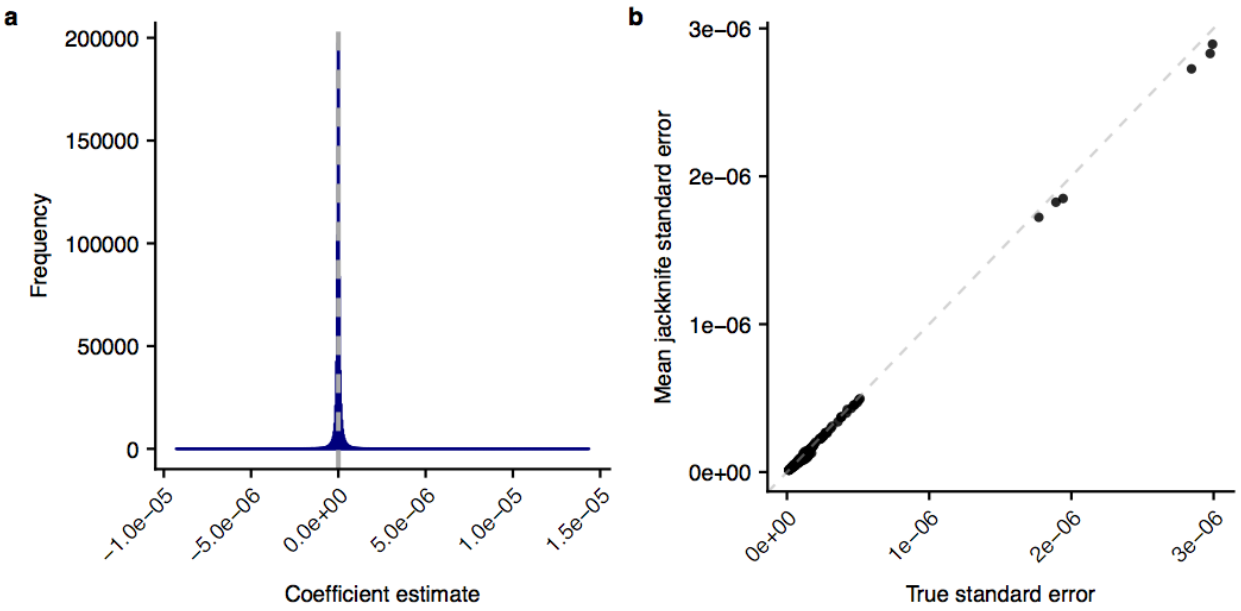


Figure S6: The coefficient estimates and standard error estimates are approximately unbiased. (**a**) Histogram of coefficient estimates over all annotations and phenotypes. The true value for all annotations and phenotypes is =0. (**b**) The true standard error vs. the jackknife standard error for each of the annotations.

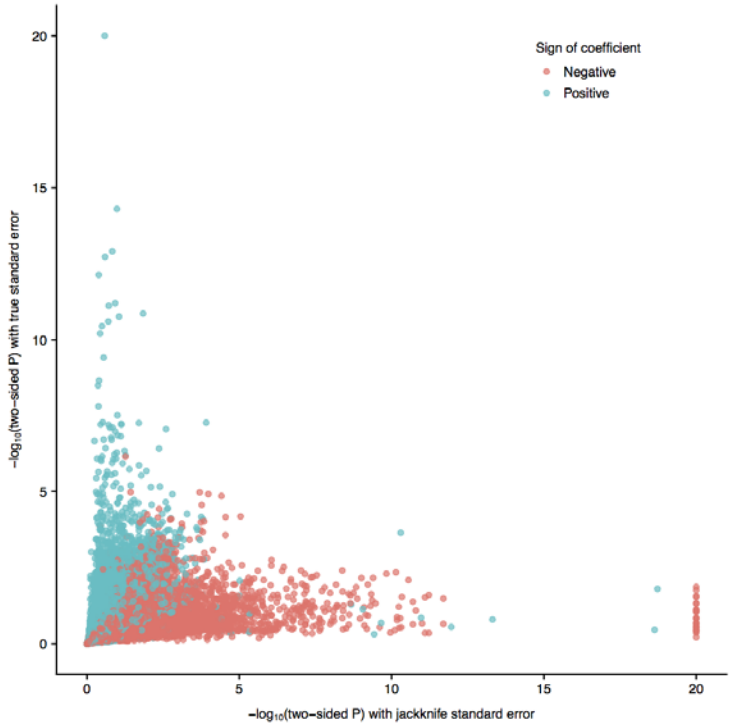


Figure S10: Correcting the jackknife standard error to the true standard deviation tended to increase significance for positive while decreasing significance for negative .

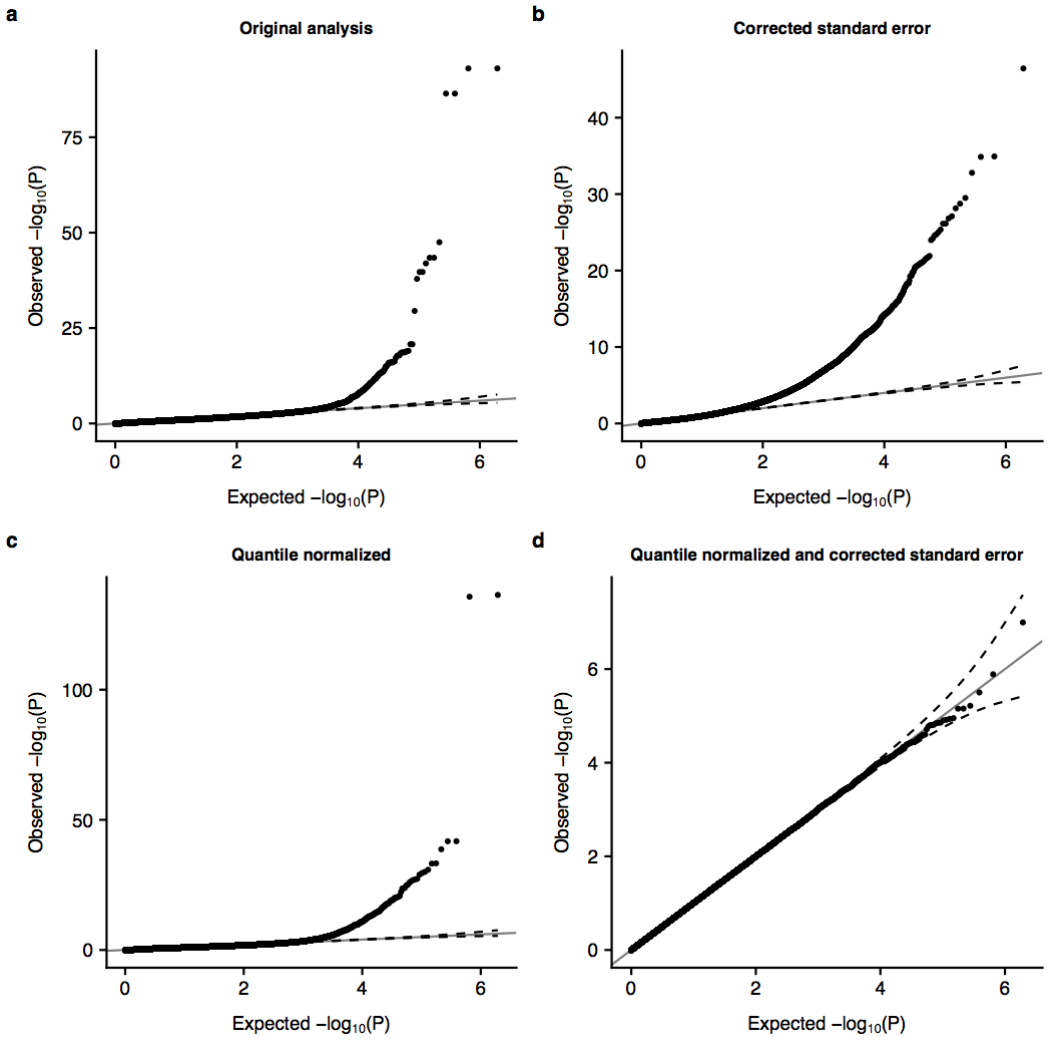


Figure S8: Dependence of type 1 error control on normality of and on the accuracy of the standard error estimate. (**a**) Q-Q plot for the one-sided test of using the jackknife standard error. (**b**) Q-Q plot for the one-sided test of using the corrected standard error. (**c**) Q-Q plot for the one-sided test of the quantile normalized using the jackknife standard error. (**d**) Q-Q plot for the one-sided test of the quantile normalized using the corrected standard error.

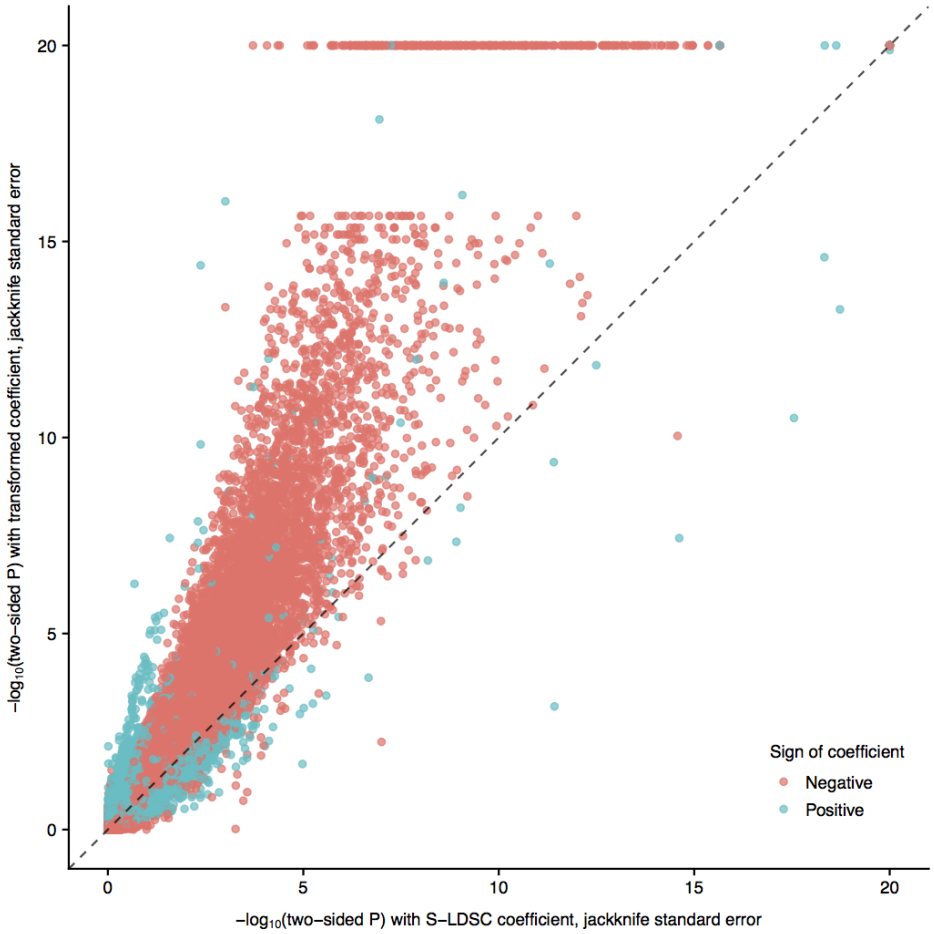


Figure S9: Quantile normalizing the distributions had the effect of mostly reducing right skew and increasing left skew, thus mostly increasing significance for the most significant negative .

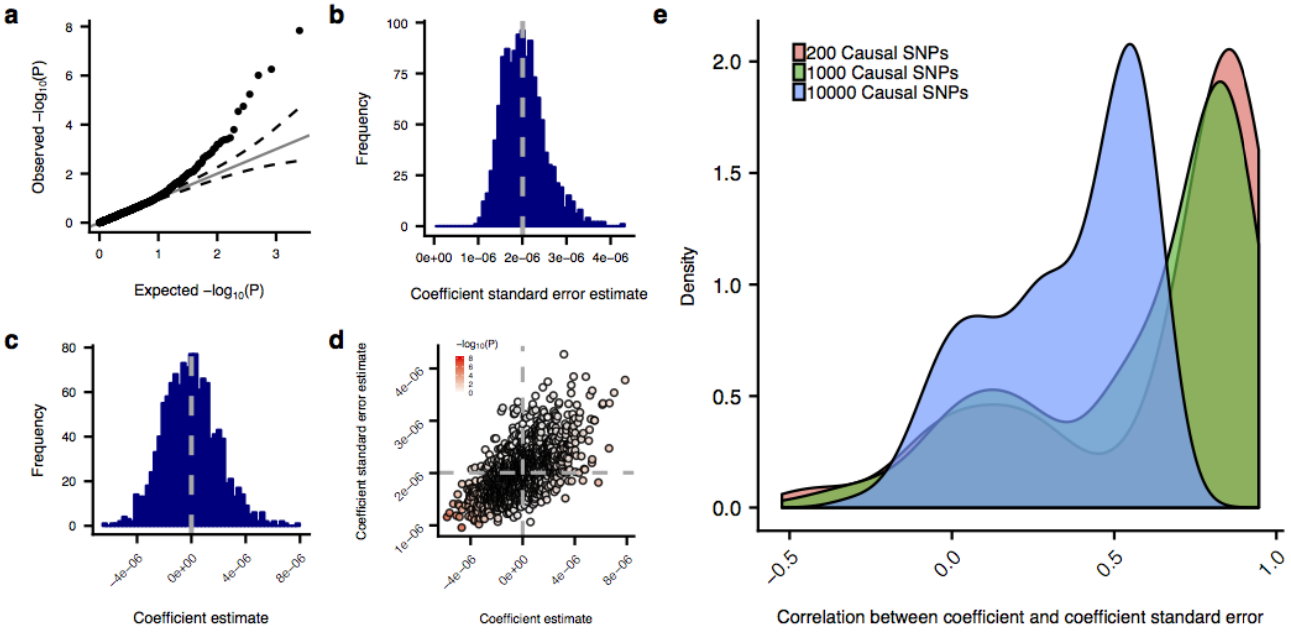


Figure S10: Correlation between and standard error estimate. **(a-d)** Example of inflated type I error in the two-sided test of for an annotation with 0.25% of SNPs overlapping 100 blocks driven by a strong positive correlation between the coefficient estimate and standard error estimate over the mid-polygenic set of phenotypes. (**a**) QQ-plot of p-values from two-sided results. (**b**) Histogram of standard error estimates for over 1250 identical simulations. (**c**) Histogram of over 1250 identical simulations. (**d**) Scatter plot of against the standard error estimate, colored by the -$\log_{10}$P of from the two-sided test of . Each point is one of 1250 identical simulations. (**e**) Density of correlations for each annotation between the coefficient and coefficient standard error estimate, colored by the level of polygenicity.

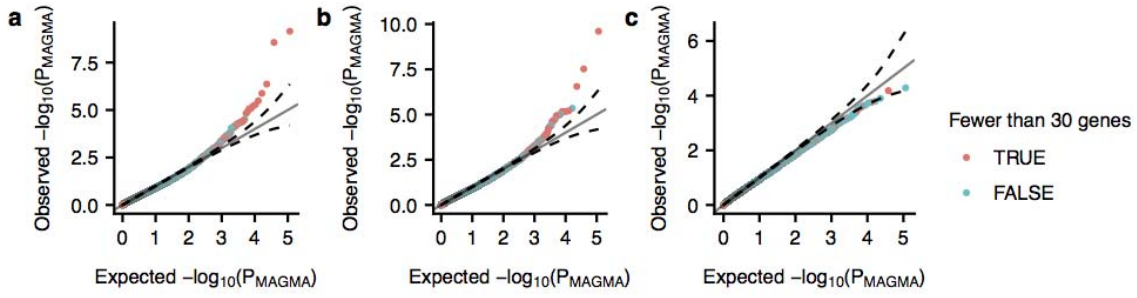


Figure S11: For gene set-based annotations, MAGMA exhibits only mild enrichment. Q-Q plots of MAGMA results on all gene set-based annotations at different levels of polygenicity, ranging from least polygenic to most polygenic (**a-c**). The inflation in panels (a) and (b) appear to be driven by annotations with fewer than 30 genes.

0.25% SNPs, 100 blocks	3% SNPs, 150 blocks
0.25% SNPs, 150 blocks	3% SNPs, 200 blocks
0.25% SNPs, 200 blocks	3% SNPs, 200 blocks
0.1% SNPs, 100 blocks	3% SNPs, 200 blocks
0.1% SNPs, 150 blocks	3% SNPs, 200 blocks
0.1% SNPs, 200 blocks	3% SNPs, 20 blocks
0.5% SNPs, 100 blocks	3% SNPs, 30 blocks
0.5% SNPs, 100 blocks	3% SNPs, 40 blocks
0.5% SNPs, 100 blocks	3% SNPs, 50 blocks
0.5% SNPs, 100 blocks	3% SNPs, 60 blocks
0.5% SNPs, 150 blocks	3% SNPs, 70 blocks
0.5% SNPs, 200 blocks	3% SNPs, 80 blocks
0.5% SNPs, 200 blocks	3% SNPs, 90 blocks
0.5% SNPs, 200 blocks	5% SNPs, 100 blocks
0.5% SNPs, 200 blocks	5% SNPs, 10 blocks
0.5% SNPs, 20 blocks	5% SNPs, 150 blocks
0.5% SNPs, 30 blocks	5% SNPs, 200 blocks
0.5% SNPs, 40 blocks	5% SNPs, 50 blocks
0.5% SNPs, 50 blocks	0.5% SNPs, 10 blocks
0.5% SNPs, 60 blocks	0.5% SNPs, 2 blocks
0.5% SNPs, 70 blocks	0.5% SNPs, 4 blocks
0.5% SNPs, 80 blocks	0.5% SNPs, 6 blocks
0.5% SNPs, 90 blocks	0.5% SNPs, 8 blocks
2% SNPs, 100 blocks	3% SNPs, 10 blocks
2% SNPs, 10 blocks	3% SNPs, 6 blocks
2% SNPs, 150 blocks	3% SNPs, 8 blocks
2% SNPs, 200 blocks	100 genes, 20 blocks
2% SNPs, 50 blocks	100 genes, 30 blocks
2% SNPs, 6 blocks	100 genes, 40 blocks
2% SNPs, 8 blocks	100 genes, 50 blocks
3% SNPs, 100 blocks	100 genes, 60 blocks

100 genes, 70 blocks	100 genes, 10 blocks	H3K4me1_peaks_Trynka.bed
100 genes, 80 blocks	100 genes, 2 blocks	H3K4me1_Trynka.bed
100 genes, 90 blocks	100 genes, 4 blocks	H3K4me1_Trynka.extend.500.bed
200 genes, 150 blocks	100 genes, 6 blocks	H3K4me3_peaks_Trynka.bed
200 genes, 200 blocks	100 genes, 8 blocks	H3K4me3_Trynka.bed
200 genes, 30 blocks	10 genes, 10 blocks	H3K4me3_Trynka.extend.500.bed
200 genes, 60 blocks	10 genes, 2 blocks	H3K9ac_peaks_Trynka.bed
200 genes, 80 blocks	10 genes, 4 blocks	H3K9ac_Trynka.bed
300 genes, 150 blocks	10 genes, 6 blocks	H3K9ac_Trynka.extend.500.bed
300 genes, 200 blocks	10 genes, 8 blocks	Intron_UCSC.bed
300 genes, 30 blocks	Coding_UCSC.bed	Intron_UCSC.extend.500.bed
300 genes, 60 blocks	Coding_UCSC.extend.500.bed	PromoterFlanking_Hoffman.bed
300 genes, 80 blocks	Conserved_LindbladToh.bed	PromoterFlanking_Hoffman.extend.500.bed
30 genes, 20 blocks	Conserved_LindbladToh.extend.500.bed	Promoter_UCSC.bed
30 genes, 30 blocks	CTCF_Hoffman.bed	Promoter_UCSC.extend.500.bed
30 genes, 40 blocks	CTCF_Hoffman.extend.500.bed	Repressed_Hoffman.bed
30 genes, 50 blocks	DGF_ENCODE.bed	Repressed_Hoffman.extend.500.bed
30 genes, 60 blocks	DGF_ENCODE.extend.500.bed	SuperEnhancer_Hnisz.bed
30 genes, 70 blocks	DHS_peaks_Trynka.bed	SuperEnhancer_Hnisz.extend.500.bed
30 genes, 80 blocks	DHS_Trynka.bed	TFBS_ENCODE.bed
30 genes, 90 blocks	DHS_Trynka.extend.500.bed	TFBS_ENCODE.extend.500.bed
450 genes, 150 blocks	Enhancer_Andersson.bed	Transcribed_Hoffman.bed
450 genes, 200 blocks	Enhancer_Andersson.extend.500.bed	Transcribed_Hoffman.extend.500.bed
450 genes, 30 blocks	Enhancer_Hoffman.bed	TSS_Hoffman.bed
450 genes, 60 blocks	Enhancer_Hoffman.extend.500.bed	TSS_Hoffman.extend.500.bed
450 genes, 80 blocks	FetalDHS_Trynka.bed	UTR_3_UCSC.bed
600 genes, 150 blocks	FetalDHS_Trynka.extend.500.bed	UTR_3_UCSC.extend.500.bed
600 genes, 200 blocks	H3K27ac_Hnisz.bed	UTR_5_UCSC.bed
600 genes, 30 blocks	H3K27ac_Hnisz.extend.500.bed	UTR_5_UCSC.extend.500.bed
600 genes, 60 blocks	H3K27ac_PGC2.bed	WeakEnhancer_Hoffman.bed
600 genes, 80 blocks	H3K27ac_PGC2.extend.500.bed	WeakEnhancer_Hoffman.extend.500.bed

MSigDB: ACTIN_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	MSigDB: KEGG_REGULATION_OF_ACTIN_CYTOSKELETON
MSigDB: AMINOPEPTIDASE_ACTIVITY	MSigDB: KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION
MSigDB: AMYLOID_PRECURSOR_PROTEIN_METABOLIC_PROCESS	MSigDB: KEGG_VIBRIO_CHOLERAE_INFECTION
MSigDB: BASSO_B_LYMPHOCYTE_NETWORK	MSigDB: KIM_WT1_TARGETS_6HR_UP
MSigDB: BIOCARTA_AKAPCENTROSOME_PATHWAY	MSigDB: KORKOLA_TERATOMA
MSigDB: BIOCARTA_CARDIACEGF_PATHWAY	MSigDB: KORKOLA_TERATOMA_DN
MSigDB: BIOCARTA_CLASSIC_PATHWAY	MSigDB: KORKOLA_TERATOMA_UP
MSigDB: BIOCARTA_HIVNEF_PATHWAY	MSigDB: LEE_LIVER_CANCER_DENA_DN
MSigDB: BIOCARTA_IL4_PATHWAY	MSigDB: LINDGREN_BLADDER_CANCER_CLUSTER_2A_UP
MSigDB: BIOCARTA_PDGF_PATHWAY	MSigDB: LIPOPROTEIN_METABOLIC_PROCESS
MSigDB: BIOCARTA_TOB1_PATHWAY	MSigDB: MICROBODY
MSigDB: CELLULAR_PROTEIN_METABOLIC_PROCESS	MSigDB: MICROBODY_MEMBRANE
MSigDB: CELLULAR_RESPONSE_TO_STRESS	MSigDB: MICROBODY_PART
MSigDB: CELL_RECOGNITION	MSigDB: MICROTUBULE_CYTOSKELETON
MSigDB: CHEN_HOXA5_TARGETS_9HR_DN	MSigDB: MICROTUBULE_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS
MSigDB: CHEOK_RESPONSE_TO_HD_MTX_DN	MSigDB: MORF_PTPRB
MSigDB: CORTICAL_ACTIN_CYTOSKELETON	MSigDB: MULLIGHAN_MLL_SIGNATURE_1_UP
MSigDB: CORTICAL_CYTOSKELETON	MSigDB: NEGATIVE_REGULATION_OF_APOPTOSIS
MSigDB: CTGCAGY_UNKNOWN	MSigDB: NEGATIVE_REGULATION_OF_CELLULAR_PROTEIN_METABOLIC_PROCESS
MSigDB: CTGRYYYNATT_UNKNOWN	MSigDB: NEGATIVE_REGULATION_OF_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS
MSigDB: CYTOSKELETON	MSigDB: NEGATIVE_REGULATION_OF_PROGRAMMED_CELL_DEATH
MSigDB: CYTOSKELETON_DEPENDENT_INTRACELLULAR_TRANSPORT	MSigDB: NEGATIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS
MSigDB: CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	MSigDB: NEGATIVE_REGULATION_OF_RNA_METABOLIC_PROCESS
MSigDB: DAVICIONI_MOLECULAR_ARMS_VS_ERMS_DN	MSigDB: NUCLEAR_REPLICATION_FORK
MSigDB: FALVELLA_SMOKERS_WITH_LUNG_CANCER	MSigDB: ONE_CARBON_COMPOUND_METABOLIC_PROCESS
MSigDB: FLOTHO_PEDIATRIC_ALL_THERAPY_RESPONSE_DN	MSigDB: ORGANELLE_INNER_MEMBRANE
MSigDB: GALE_APL_WITH_FLT3_MUTATED_DN	MSigDB: POSITIVE_REGULATION_OF_CELLULAR_PROTEIN_METABOLIC_PROCESS
MSigDB: GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS_BLACK_DN	MSigDB: POSITIVE_REGULATION_OF_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS
MSigDB: GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS_YELLOW_DN	MSigDB: POSITIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS
MSigDB: GCM_SMO	MSigDB: POSITIVE_REGULATION_OF_RNA_METABOLIC_PROCESS
MSigDB: GLYCOLIPID_METABOLIC_PROCESS	MSigDB: POSITIVE_REGULATION_OF_TRANSFERASE_ACTIVITY
MSigDB: GLYCOPROTEIN_METABOLIC_PROCESS	MSigDB: POSITIVE_REGULATION_OF_T_CELL_PROLIFERATION
MSigDB: GNF2_SMC1L1	MSigDB: PROGRAMMED_CELL_DEATH
MSigDB: GRAHAM_CML_QUIESCENT_VS_NORMAL_DIVIDING_UP	MSigDB: PROTEIN_METABOLIC_PROCESS
MSigDB: GRESHOCK_CANCER_COPY_NUMBER_DN	MSigDB: PROTEIN_SERINE_THREONINE_PHOSPHATASE_ACTIVITY
MSigDB: HANN_RESISTANCE_TO_BCL2_INHIBITOR_DN	MSigDB: REACTOME_COSTIMULATION_BY_THE_CD28_FAMILY
MSigDB: HUANG_DASATINIB_RESISTANCE_UP	MSigDB: REACTOME_MITOTIC_M_M_G1_PHASES
MSigDB: HUMORAL_IMMUNE_RESPONSE	MSigDB: REACTOME_NUCLEOTIDE_LIKE_PURINERGIC_RECEPTORS
MSigDB: HYDROLASE_ACTIVITY_ACTING_ON_ACID_ANHYDRIDESCATALYZING_TRANSMEMBRANE_MOVEMENT_OF_SUBSTANCES	MSigDB: REACTOME_P75_NTR_RECEPTOR_MEDIATED_SIGNALLING
MSigDB: HYDROLASE_ACTIVITY_ACTING_ON_GLYCOSYL_BONDS	MSigDB: REACTOME_REGULATION_OF_GENE_EXPRESSION_IN_BETA_CELLS
MSigDB: INOSITOL_OR_PHOSPHATIDYLINOSITOL_PHOSPHATASE_ACTIVITY	MSigDB: REACTOME_TRAF6_MEDIATED_INDUCTION_OF_THE_ANTIVIRAL_CYTOKINE_IFN_ALPHA_BETA_CASCADE
MSigDB: INTERLEUKIN_BINDING	MSigDB: REGULATION_OF_APOPTOSIS
MSigDB: INTERMEDIATE_FILAMENT_CYTOSKELETON	MSigDB: REGULATION_OF_CELLULAR_PROTEIN_METABOLIC_PROCESS
MSigDB: KALMA_E2F1_TARGETS	MSigDB: REGULATION_OF_CELL_MORPHOGENESIS
MSigDB: KEGG_JAK_STAT_SIGNALING_PATHWAY	MSigDB: REGULATION_OF_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS
MSigDB: KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM	MSigDB: KEGG_REGULATION_OF_ACTIN_CYTOSKELETON

Table S2: List of 246 annotations used in simulations.