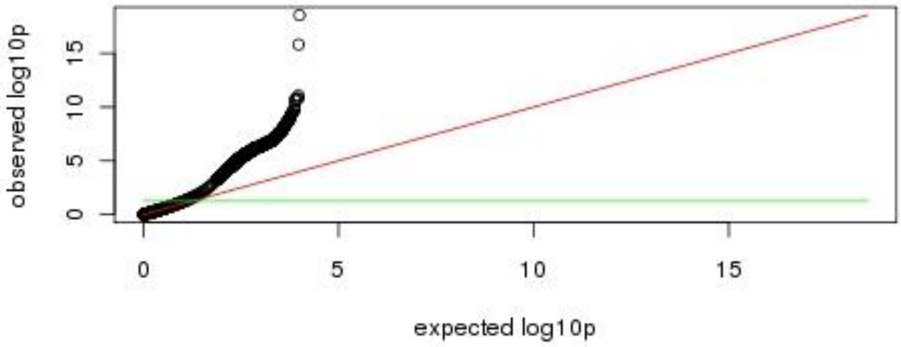
Supplemental Data

American Journal of Human Genetics, Volume *8*

Estimating Missing Heritability for Disease from Genome-wide Association Studies

Sang Hong Lee, Naomi R. Wray, Michael E. Goddard, and Peter M. Visscher



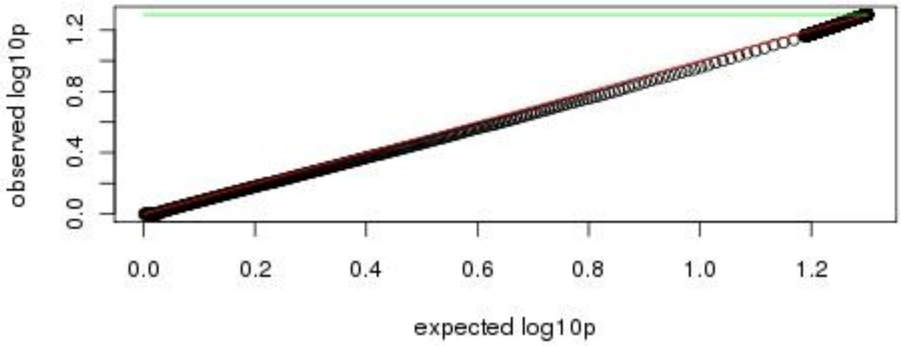
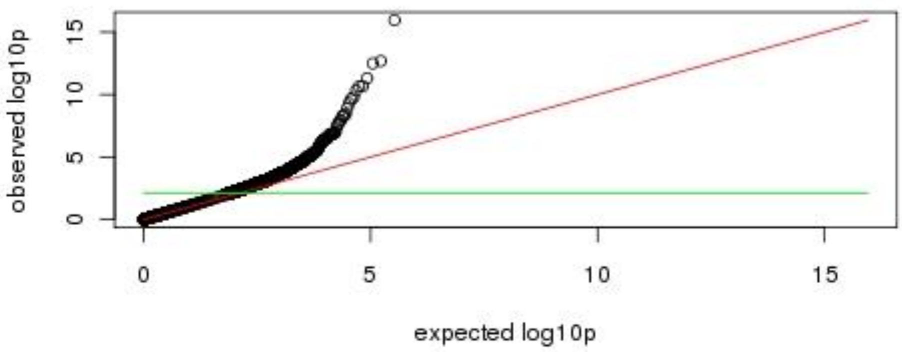


Figure S1. Hardy-Weinberg Test

QQ plot from H-W test for the bipolar case-control data set with filtering out SNPs whose p-value < 0.0001 (upper), or with filtering out SNPs whose p-value < 0.05 (lower). The number of data points that deviated from the expectation (above the green line) was ~35,000 for the upper figure and 0 for the lower figure.



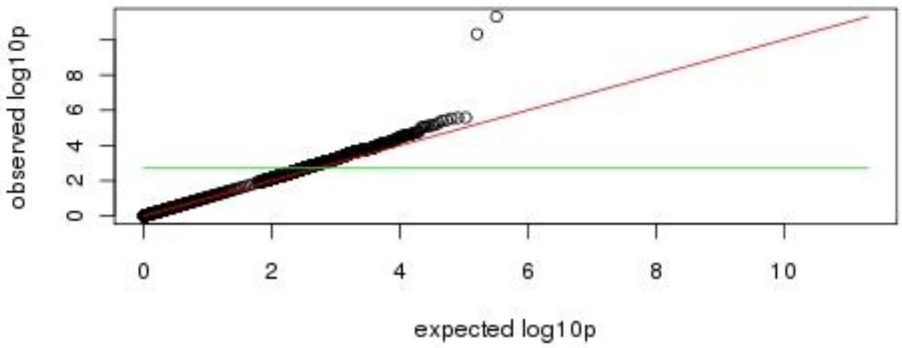
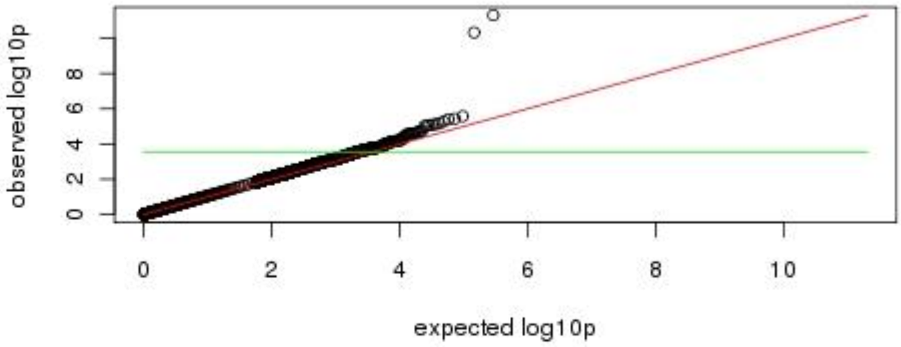


Figure S2. Differential Missingness Test

QQ plot from test statistics for two-locus QC before filtering out problematic SNPs (upper), and after filtering out problematic SNPs (lower) according to their differential missingness. The number of data points deviated from the expectation (above the green line) was 4,117 for upper figure and 963 for lower figure. The bipolar data set was used.



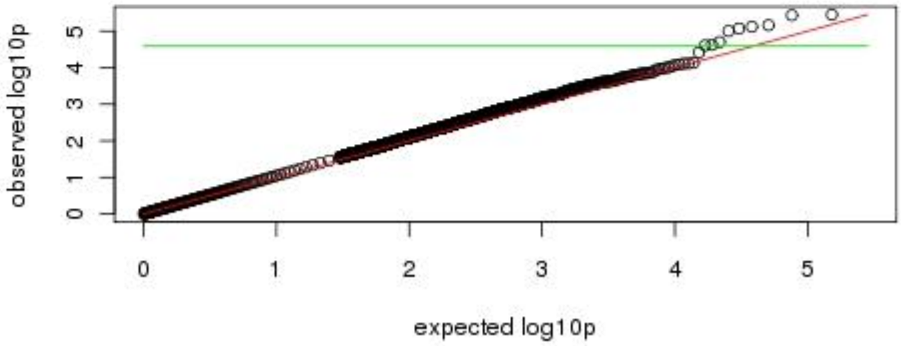


Figure S3. Stringent QC on SNP Missingness

Test statistics from the two-locus QC test for the bipolar disorder data set when excluding SNPs with missing rate > 20/N (upper) and when excluding SNPs with missing rate > 4/N (lower), where N is the total sample size. The number of data points that deviated from the expectation (above the green line) was 145 for upper figure and 7 for lower figure.

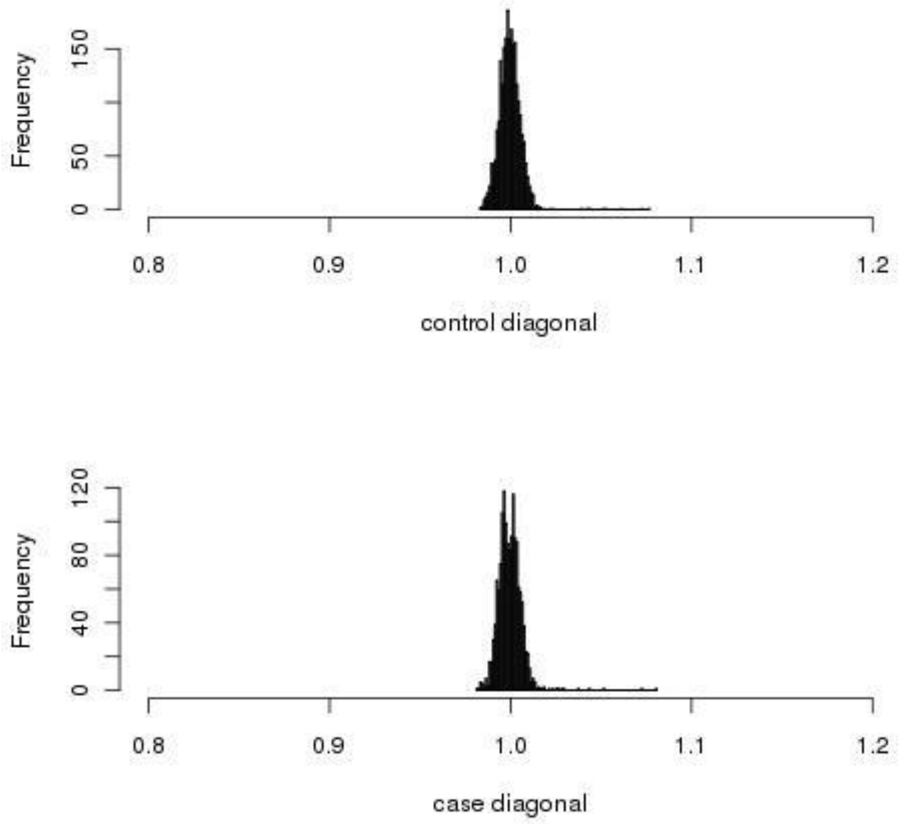


Figure S4. Histogram of Diagonal Elements for the Crohn's Disease Data

The genetic relationships were estimated from 322142 SNPs genotypes on 1504 cases and 2329 controls.

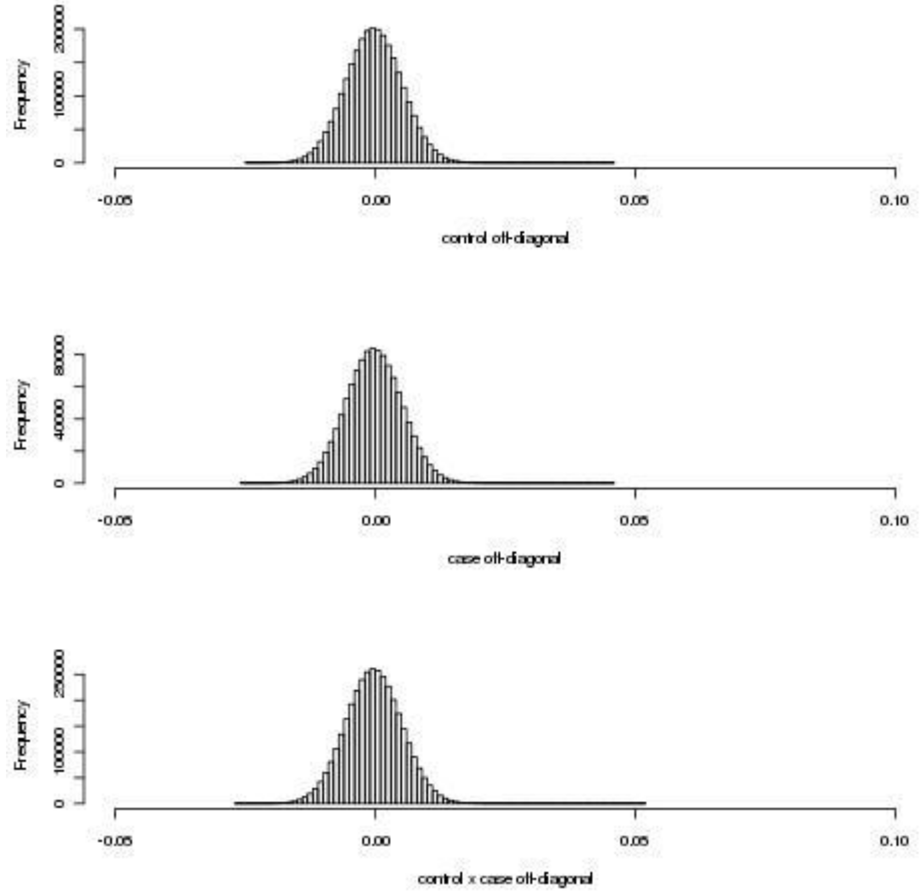


Figure S5. Histogram of Off-Diagonal Elements for the Crohn's Disease

The genetic relationships were estimated from 322142 SNPs genotypes on 1504 cases and 2329 controls.

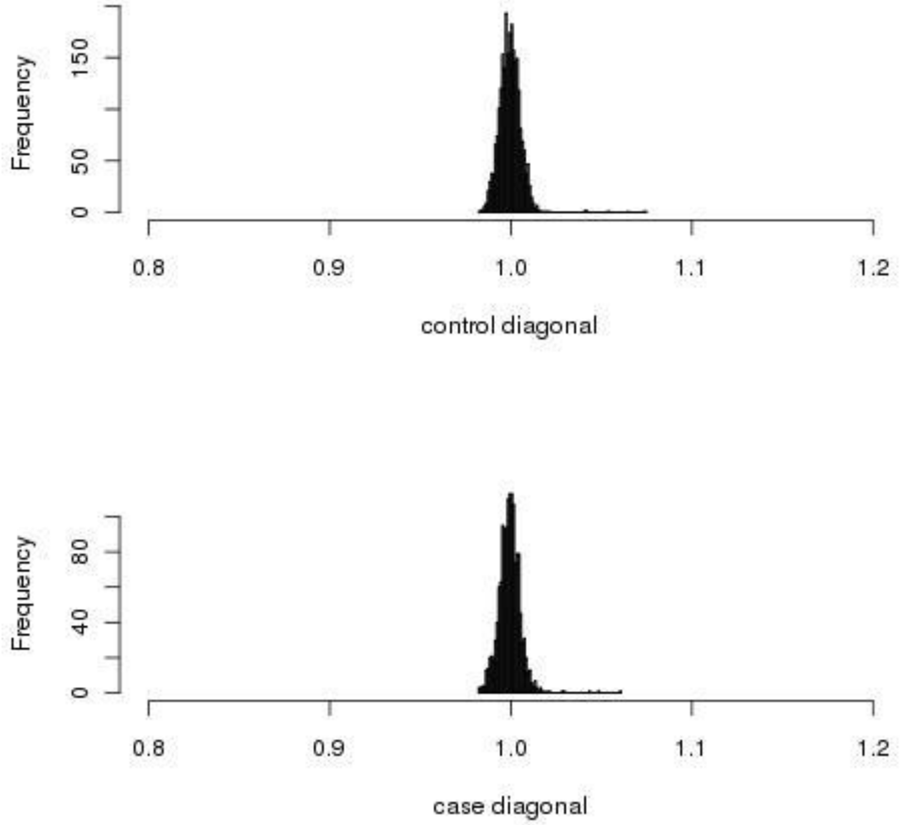


Figure S6. Histogram of Diagonal Elements for the Bipolar Disorder Data

The genetic relationships were estimated from 321605 SNPs genotypes on 1433 cases and 2447 controls.

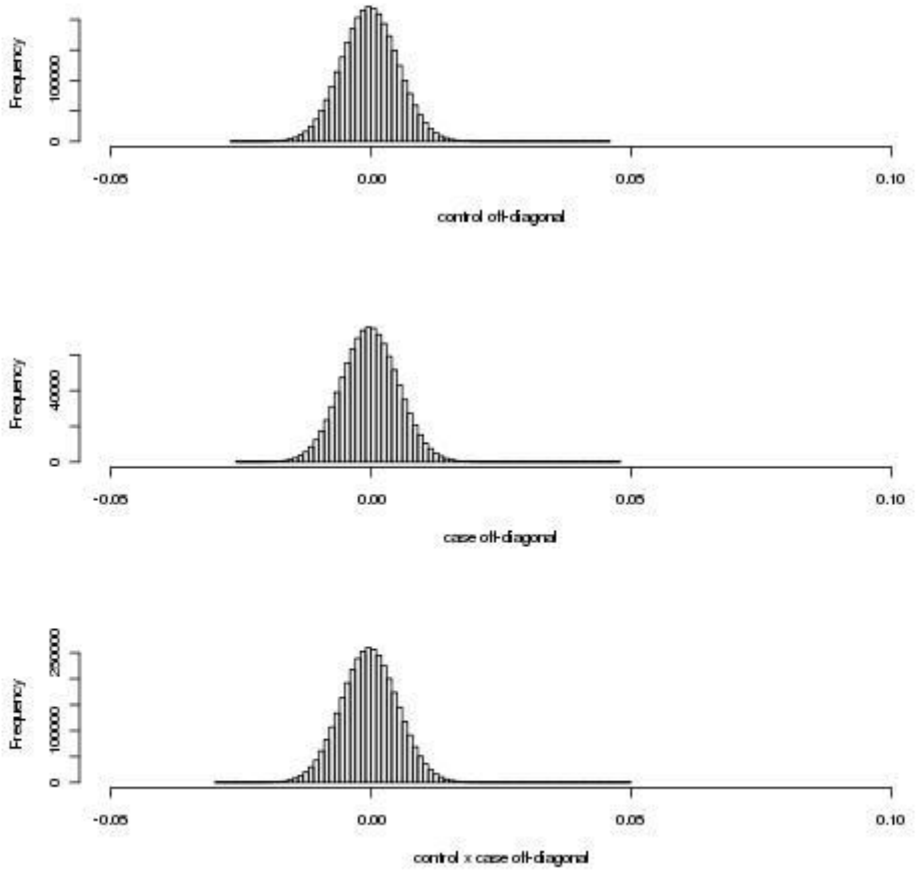


Figure S7. Histogram of Off-Diagonal Elements for the Bipolar Disorder Data

The genetic relationships were estimated from 321605 SNPs genotypes on 1433 cases and 2447 controls.

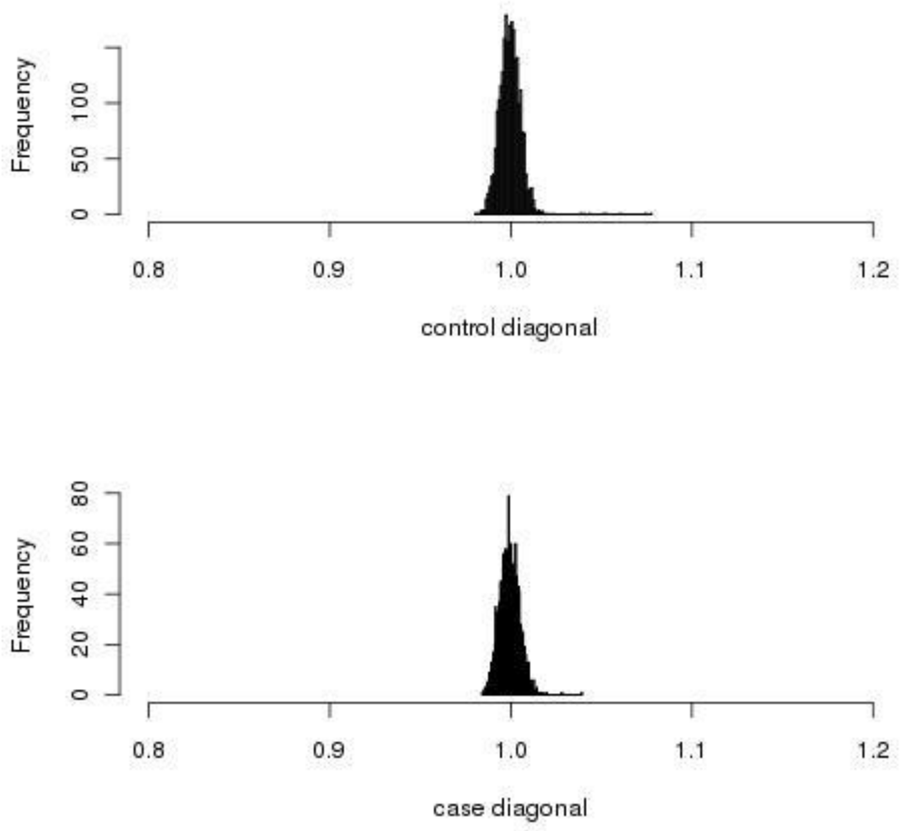


Figure S8. Histogram of Diagonal Elements for the Type I Diabetes Data

The genetic relationships were estimated from 318044 SNPs genotypes on 1640 cases and 2423 controls.

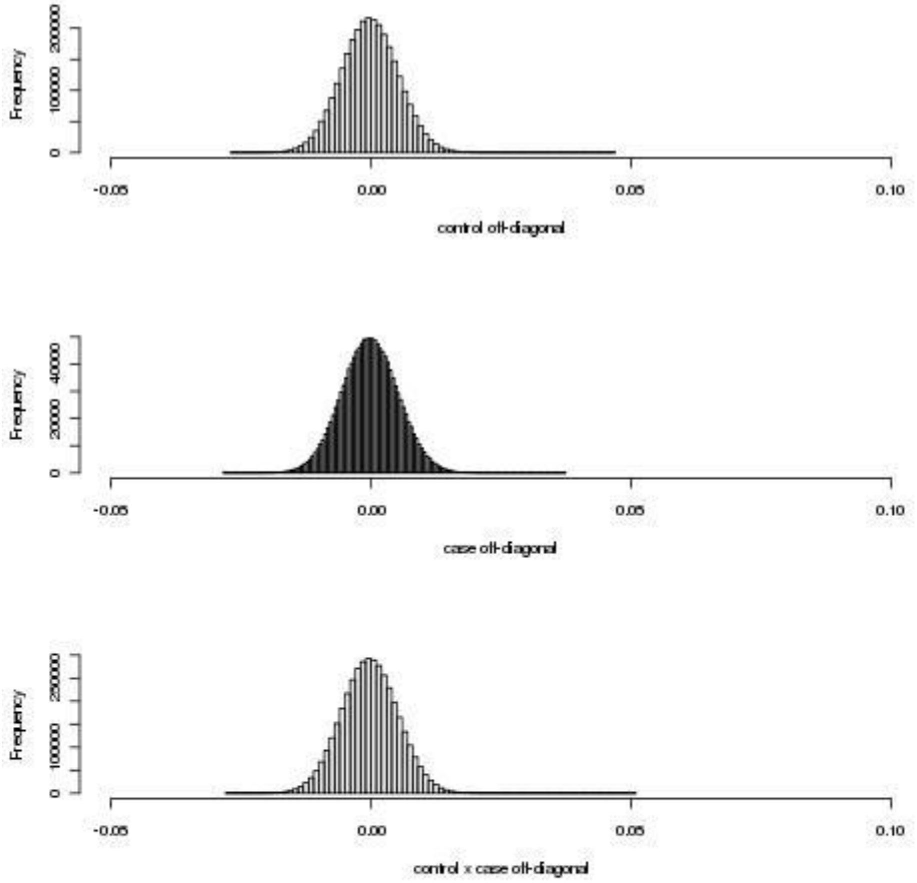


Figure S9. Histogram of Off-Diagonal Elements for the Type I Diabetes Data

The genetic relationships were estimated from 318044 SNPs genotypes on 1640 cases and 2423 controls.

Table S1. Mean and Standard Deviation for Diagonal Values

	controls		cases	
	mean	SD	mean	SD
CD	0.999646	0.006418	0.999569	0.006732
BD	0.99963	0.006411	0.999321	0.006114
T1DB	0.999658	0.006446	0.999254	0.005549

Table S2. Mean and Standard Deviation for Off-Diagonal Values

	controls		cases		controls x cases	
	mean	SD	mean	SD	mean	SD
CD	-0.00025	0.005368	-0.00023	0.005384	-0.00028	0.005373
BD	-0.00024	0.005401	-0.00021	0.00542	-0.00028	0.005405
T1DB	-0.00023	0.005411	-0.00021	0.005409	-0.00027	0.005408

The mean value for the conrol-control and case-case is slightly higher than that for the case-control for all traits.

Table S3. Estimated Genetic Variance on the Observed and Liability Scale Explained by all SNPs for Two Independents Sets of Case-Controls Studies for Three Complex Traits

data	estimate[a] (SE)	LR	transformed[b] (SE)
	Crohn's disease		
original	0.50 (0.07)	54.94	0.18 (0.03)
set 1	0.51 (0.13)	15.02	0.19 (0.05)
set 2	0.59 (0.15)	16.05	0.21 (0.05)
	Bipolar disorder		
original	0.62 (0.07)	92.21	0.32 (0.03)
set 1	0.50 (0.13)	15.44	0.26 (0.07)
set 2	0.72 (0.15)	25.17	0.36 (0.07)
	Type 1 diabetes		
original	0.51 (0.07)	64.74	0.25 (0.03)
set 1	0.46 (0.12)	14.41	0.23 (0.06)
set 2	0.55 (0.14)	16.15	0.27 (0.07)

[a]Estimate of genetic variance proportional to the total phenotypic variance on the observed scale. [b]Transformed genetic variance proportional to the total phenotypic variance on the liability scale. To create two independent case-control studies for the same disease, we randomly divided cases into two sets. One case set was combined with the 1958 birth cohort controls, and the other was combined with the NBS controls. Both of the two case-control sets were analysed separately. There was no significant difference between estimates from the two independent data sets for all traits. We note that the estimates from the joint analysis is roughly midway between the estimates from two independent data sets. These results provided no evidence for bias due to technical artifacts or population stratification.

Table S4. Haseman-Elston Regression Testing the 1958 versus the NBS Cohort

threshold[a]	no. SNP	Intercept (SE)	regression slope (SE)	p value
before stringent QC[b]	389528	0.50 (0.0003)	-0.23 (0.05)	1.9e-05***
200/N	309040	0.50 (0.0003)	-0.08 (0.05)	0.13
20/N	297198	0.50 (0.0003)	-0.06 (0.05)	0.25
7/N	266534	0.50 (0.0003)	-0.04 (0.05)	0.44
4/N	226165	0.50 (0.0003)	-0.03 (0.05)	0.58

[a] Excluding SNP whose missing rate > specified threshold (N is total number of samples).

[b] Applying only standard QC without additional stringent QC. We checked the control-control contrast study using Haseman-Elston regression. The model was: z-scores = mu + x + e where x variables were pairwise relationships, and z-scores were 1 for pairs in the same group or 0 for pairs across different groups. The contrast between 1958 cohort and NBS groups was highly significant before stringent QC was applied, but the regression coefficient dramatically decreased and became non-significant after applying more stringent QC. For more stringent QC on the SNP missing rate, the regression coefficients decreased, and their p-values increased. This clear pattern indicated that artificial differences in allele frequencies across the two control populations could be removed by applying stringent QC.

Table S5. Haseman-Elston Regression Testing Each Age Group in Controls

age group	Intercept (SE)	regression slope (SE)	p value
1	0.03 (8e-05)	-0.005 (0.02)	0.77
2	0.12 (0.0002)	0.001 (0.03)	0.97
3	0.17 (0.0002)	-0.01 (0.04)	0.89
4	0.45 (0.0003)	-0.04 (0.05)	0.44
5	0.23 (0.0002)	0.03 (0.04)	0.47
6	0.07 (0.0001)	0.01 (0.03)	0.66

This analysis was performed because age (at which the participants entered a study) might be associated with systematic artifact bias due to batch or plate effects. However, we observed that the relationships within an age group were not significantly higher than those across the rest of age groups for controls (1958 cohort and NBS).

Table S6. Haseman-Elston Regression Testing Each Age Group in Crohn's Disease Cases

age group	Intercept (SE)	regression slope (SE)	p value
1	0.035 (0.0002)	-0.003 (0.032)	0.92
2	0.254 (0.0004)	0.090 (0.076)	0.23
3	0.365 (0.0005)	0.093 (0.084)	0.27
4	0.331 (0.0004)	0.024 (0.082)	0.77
5	0.295 (0.0004)	0.101 (0.080)	0.20
6	0.183 (0.0004)	-0.030 (0.068)	0.66
7	0.100 (0.0003)	-0.033 (0.052)	0.53
8	0.021 (0.0001)	0.013 (0.025)	0.59

The genetic relationships based on genome-wide SNPs for each age group were not significantly different from those across the rest of age groups when using the Crohn's disease case data.

Table S7. Haseman-Elston Regression Testing Each Age Group within Bipolar Disorder Cases

age group	Intercept (SE)	regression slope (SE)	p value
1	0.012 (0.0001)	0.018 (0.019)	0.37
2	0.173 (0.0004)	0.010 (0.068)	0.89
3	0.317 (0.0005)	-0.006 (0.083)	0.94
4	0.417 (0.0005)	-0.078 (0.088)	0.38
5	0.360 (0.0005)	-0.294 (0.086)	0.0006***
6	0.244 (0.0004)	0.110 (0.077)	0.15
7	0.047 (0.0002)	0.017 (0.038)	0.65
8	0.010 (9.5e-05)	0.015 (0.017)	0.39

The genetic relationships based on genome-wide SNPs for each age group were not significantly different from those across the rest of age groups except those for the age group 5 when using the bipolar disorder case data. This was subsequently explored (Table S7-2).

Table S7-2. Testing Whether Age Group 5 in Cases Had Significant Effects in Estimating Genetic Variance

data	no. samples	estimate (SE)	LR	transformed (SE)
including age group 5 in cases	1433 cases 2447 controls	0.62 (0.07)	92.21	0.64 (0.14)
excluding age group 5 in cases	1095 cases 2447 controls	0.60 (0.07)	70.10	0.66 (0.15)

This further test was conducted to investigate if the more related individuals in age group 5 for the bipolar disorder cases influenced the estimate of genetic variance. The estimate without the age group 5 was not much different from the original estimate.

Table S8. Haseman-Elston Regression Testing Each Age Group in Type I Diabetes Cases

age group	Intercept (SE)	regression slope (SE)	p value
1	0.480 (0.0004)	0.015 (0.079)	0.85
2	0.061 (0.0002)	-0.016 (0.038)	0.68
3	0.041 (0.0002)	-0.002 (0.031)	0.96
4	0.017 (0.0001)	0.010 (0.020)	0.63
5	0.007 (0.0001)	-0.007 (0.013)	0.59
6	0.001 (2.9e-5)	-0.003 (0.005)	0.55
7	0.001 (2.9e-5)	-0.005 (0.005)	0.35

The genetic relationships based on genome-side SNPs for each age group were not significantly different from those across the rest of age groups when using the type I diabetes case data.

Table S9. Estimated Genetic Variance Proportional to Total Phenotypic Variance for a Case-Case Contrast Study

study	no. SNP	estimate (SE)	LR
CD	195977	0.50 (0.07)	54.94
BD	187597	0.62 (0.07)	92.21
T1D	178892	0.51 (0.07)	64.74
CD : BD	212634	0.96 (0.08)	148.66
BD : T1D	198649	1.00 (0.0003)	189.66
CD : T1D	205101	0.92 (0.07)	149.73

These analyses were to investigate if there were systematic correlations between case samples that were not expected among these three complex diseases. If there were, estimated values from case-case contrasts should be less than those from the original case-control study. After the stringent QC applied, variances were estimated for Crohn's disease cases vs. bipolar disorder cases (CD:BD), bipolar disorder cases vs. type I diabetes cases (BD:T1D), or Crohn's disease cases vs. type I diabetes cases (CD:T1D), and compared with those for original case-control studies (CD, BD and T1D). It was shown that case-case contrast study gave much higher estimated values than the original case-control studies. This is expected under the hypothesis that genetic differences between case-case groups for diseases that are genetically uncorrelated are larger than between case-control groups.

Table S10. Bivariate Analysis for Cases and Controls for Each Pair of Diseases

trait 1	trait 2	estimate for trait 1 (SE)	estimate for trait 2 (SE)	genetic corr.
CD	BD	0.60 (0.09)	0.68 (0.09)	0.07 (0.10)
CD	T1D	0.51 (0.08)	0.59 (0.09)	0.02 (0.12)
BD	T1D	0.70 (0.08)	0.56 (0.09)	-0.12 (0.10)

This test was to investigate if there were systematic correlations between case samples that were not expected among these three complex diseases. The analysis was done for cases for each disease with half of the controls, i.e. 1958 cohort controls for one trait and NBS controls for the other trait, i.e. analysis 1. (CD + 1958 cohort controls) vs (BP + NBS controls), analysis 2. (CD + NBS controls) vs (T1D + 1958 cohort), and analysis 3. (BP + NBS controls) vs (T1D + 1958 cohort controls). All estimated pairwise genetic correlations were not significantly different from zero.

Table S11. Quantification of the Proportional Bias in the estImate of Variance Explained by All SNPs when the Prevalence of Disease in the Population Is Misspecified

True prevalence	Proportional bias due to misspecification of prevalence			
	Assumed prevalence ($\hat{K}$)			
	0.5*K*	0.75*K*	1.5*K*	2*K*
K=0.2	0.81	0.92	1.12	1.18
K=0.1	0.81	0.91	1.14	1.24
K=0.01	0.86	0.94	1.10	1.19
K=0.005	0.87	0.94	1.09	1.17
K=0.001	0.90	0.96	1.07	1.13

The ratio was derived as $(\hat{K}(1-\hat{K})/\hat{z})^2 / (K(1-K)/z)^2$ where $\hat{K}$ =0.5*K*, 0.75*K*, 1.5*K* or 2*K*.