Data supplement for Mahjani et al., The Genetic Architecture of Obsessive-Compulsive Disorder: Contribution of Liability to OCT From Alleles Across the Frequency Spectrum. Am J Psychiatry (doi: 10.1176/appi.ajp.2021.21010101)

Quality control

After quality control described in the main text, we merged the cohorts and performed the following additional quality control steps:

- removed 135 subjects, 6 diagnosed with OCD, deemed to be close relatives (pihat > 0.2).

By contrasting allele frequencies in the different cohorts using measure of allelic variation such as fixation index (FSt), and by analyzing only individuals genetically identified as of European ancestry, we removed variants with:

- FSt > 0.005 (185variants) between controls,
- FSt > 0.005 (6 variants) between all cohorts,
- FSt > 0.005 (2 variants) between EGOS and controls,
- FSt > 0.005 (5 variants) between NORDiC and controls,
- missingness in a cohort > 0.02 (12629 variants),
- (max – min) allele frequency across the control > 0.03 (40540variants).

Next, we sought to remove poorly called SNPs by contrasting allele frequencies from LifeGene (iCON and NORDiC) controls versus LifeGene-ANGI controls using a standard logistic association test, as would be used for a GWAS. We removed 117 variants with p-value < 1e-4.

We removed SNPs with a significant difference in missingness between OCD cases and controls |(missingness – mean missingness)| >0.01 (2894 variants).

The final dataset had 2090 cases and 4567 controls, with 412813 SNPs (56378 variants were removed after merging the cohorts).

TABLE S1. Details of QC for EGOS cases, NORDiC cases, LifeGene iCON, LifeGene NORDiC (batch1)

	Individuals	**SNPs**	**Removed individuals in each step**	**Removed SNPs in each step**
cases/controls	2215/1943	759993	-	-
Phase 1: Pre-QC				
a. Check duplicate marker names	2215/1943	759993	-	0
b. SNPs not containing rs as part of the name	2215/1943	708521	-	51472
c. Remove SNPs without location	2215/1943	701511	-	7010
d. Remove SNPs on PAR and MT	2215/1943	699608	-	PAR:927, MT:976
e. Remove all homozygous SNPs	2215/1943	696155	-	3453
f. INDELs	2215/1943	687102	-	9053
g. Remove SNPs sharing the same location	2215/1943	687102	-	0
h. Remove ambiguous SNPs	2215/1943	677246	-	9856
i. Non call rate on SNPs (0.15)	2215/1943	675308	-	1938
Phase 2: QC on individuals				
a. Check for duplicate samples IDs	2215/1943	675308	0	-
b. Remove samples with platingissues	2215/1943	675308	0	-
c. Non call rate (0.05, autosome)	2143/1912	675308	103	-
d. Sex discrepancy	2142/1905	675308	8	-
e. Heterozygosity (remove <-3SD or >3SD)	2119/1827	675308	101 (23/78)	-
Phase 3: QC, relatedness		675308		
a. Check for Family IDs	2119/1827	675308	0	-
b. Remove close relatives (pihat > 0.2)	2092/1788	675308	66	-
Phase 4: QC on SNPs				
a. Remove ChrY	2092/1788	671902	-	3406
b. Non call rate (0.05)	2092/1788	666322	-	5580
c. $^{+}$Minor allele freq (0.01)	2092/1788	509661	-	156661
d. $^{+}$Hardy-Weinberg equilibrium (0.00125)	2092/1788	505968	-	3693
Phase 5: Check against 1000G (McCarthy tool)				
a. No Match to 1000G	2092/1788	505777	-	191
b. Removed for allele freq diff > 0.2	2092/1788	504959	-	818
c. Palindromic SNPs with freq > 0.4	2092/1788	504959	-	0
d. Non Matching alleles	2092/1788	503570	-	389
e. Duplicates removed	2092/1788	504045	-	525

$^{+}$ Based on European ancestry.

TABLE S2. Details of QC for LifeGene-ANGI-Wave-1 (batch 2)

	Individuals	**SNPs**	**Removed individuals in each step**	**Removed SNPs in each step**
controls	1500	688032	-	-
Phase 1: Pre-QC				
a. Check duplicate marker names	1500	688032	-	0
b. SNPs not containing rs as part of the name	1500	650645	-	37387
c. Remove SNPs without location	1500	650645	-	0
d. Remove SNPs on PAR and MT	1500	650641	-	4
e. Remove all homozygous SNPs	1500	650641	-	0
f. INDELs	1500	650641	-	0
g. Remove SNPs sharing the same location	1500	650641	-	0
h. Remove ambiguous SNPs	1500	642436	-	8205
i. Non call rate on SNPs (0.15)	1500	637487	-	4949
Phase 2: QC on individuals				
a. Check for duplicate samples IDs	1500	637487	0	-
b. Remove samples with plating issues	1500	637487	0	-
c. Non call rate (0.05, autosome)	1500	637487	0	-
d. Sex discrepancy	1496	637487	4	-
e. Heterozygosity (remove <-3SD or >3SD)	1496	637487	12	-
Phase 3: QC, relatedness				
a. Check for Family IDs	1496	637487	0	-
b. Remove close relatives (pihat > 0.2)	1454	637487	30	-
Phase 4: QC on SNPs				
a. Remove ChrY	1454	637487	-	0
b. Non call rate (0.05)	1454	631352	-	6135
c. [+]Minor allele freq (0.01)	1454	491921	-	139431
d. [+]Hardy-Weinberg equilibrium (0.00125)	1454	487997	-	3924
Phase 5: Check against 1000G (McCarthy tool)				
a. No Match to 1000G	1454	487909	-	88
b. Removed for allele freq diff > 0.2	1454	487042	-	867
c. Palindromic SNPs with freq > 0.4	1454	487042	-	0
d. Non Matching alleles	1454	486730	-	312
e. Duplicates removed	1454	486658	-	72
f. [*]Harmonize to batch 1	1454	475953	-	10705

A few pre-QC steps were already performed on these batches, including removing SNPs by a set of standard criteria: without known genomic location, because they fell in the pseudoautosomal region, were part of the mitochondrial genome, and so forth.

+ Based on European ancestry.

*Genotype Harmonizer software was used for strand alignment and format conversion for genotype data integration between different batches. Batch 2 was aligned to batch 1.

TABLE S3. Details of QC for LifeGene-ANGI-Wave-2 (batch3)

	Individuals	**SNPs**	**Removed individuals in each step**	**Removed SNPs in each step**
controls	1500	688032	-	-
Phase 1: Pre-QC				
a. Check duplicate marker names	1500	688032	-	0
b. SNPs not containing rs as part of the name	1500	650645	-	37387
c. Remove SNPs without location	1500	650641	-	0
d. Remove SNPs on PAR and MT	1500	650641	-	4
e. Remove all homozygous SNPs	1500	650641	-	0
f. INDELs	1500	650641	-	0
g. Remove SNPs sharing the same location	1500	650641	-	0
h. Remove ambiguous SNPs	1500	642436	-	8205
i. Non call rate on SNPs (0.15)	1500	638254	-	4182
Phase 2: QC on individuals				
a. Check for duplicate samples IDs	1500	638254	0	-
b. Remove samples with plating issues	1500	638254	0	-
c. Non call rate (0.05, autosome)	1500	638254	0	-
d. Sex discrepancy	1497	638254	3	-
e. Heterozygosity (remove <-3SD or >3SD)	1458	638254	39	-
Phase 3: QC, relatedness				
a. Check for Family IDs	1479	638254	0	-
b. Remove close relatives (pihat > 0.2)	1438	638254	20	-
Phase 4: QC on SNPs				
a. Remove ChrY	1438	638254	-	0
b. Non call rate (0.05)	1438	632687	-	5567
c. $^+$Minor allele freq (0.01)	1438	489693	-	142994
d. $^+$Hardy-Weinberg equilibrium (0.00125)	1438	487930	-	1763
Phase 5: Check against 1000G (McCarthy tool)				
a. No Match to 1000G	1438	487841	-	89
b. Removed for allele freq diff > 0.2	1438	486963	-	878
c. Palindromic SNPs with freq > 0.4	1438	486963	-	0
d. Non Matching alleles	1438	486653	-	310
e. Duplicates removed	1438	486576	-	77
f. *Harmonize to batch 1	1438	476644	-	9932

A few pre-QC steps were already performed on these batches, including removing SNPs by a set of standard criteria: without known genomic location, because they fell in the pseudoautosomal region, were part of the mitochondrial genome, and so forth.

$^+$ Based on European ancestry (the largest clusters in GEMToolss).

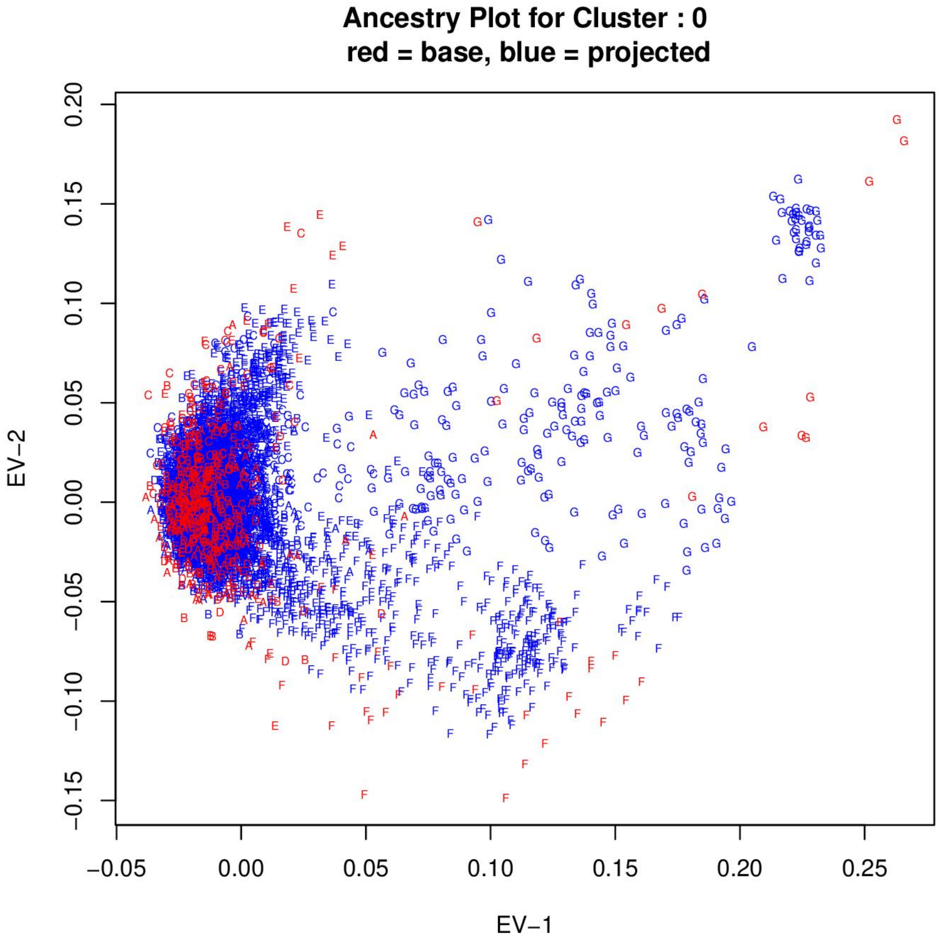
*Genotype Harmonizer software was used for strand alignment and format conversion for genotype data integration between different batches. Batch 3 was aligned to batch 1.

Population stratification, ancestry groups

We used GEMTools to find individuals with recent European ancestry. GEMTools uses spectral graph methods to find a low-dimensional representation of the genetic similarities between individuals, which is referred to as an eigenmap. Assuming an eigenmap is constructed using a representative base sample, additional individuals can be projected onto the map using the Nystrom approximation (1). Non-base individuals are assigned to the cluster of their genetically closest base-neighbor.

Figure S1 illustrates the base and non-base individuals for the first six ancestry vectors. Individuals in clusters A, B, C, and D have the closest ancestry (min.dim=6; GEMTools found two eigenvectors without using min.dim).

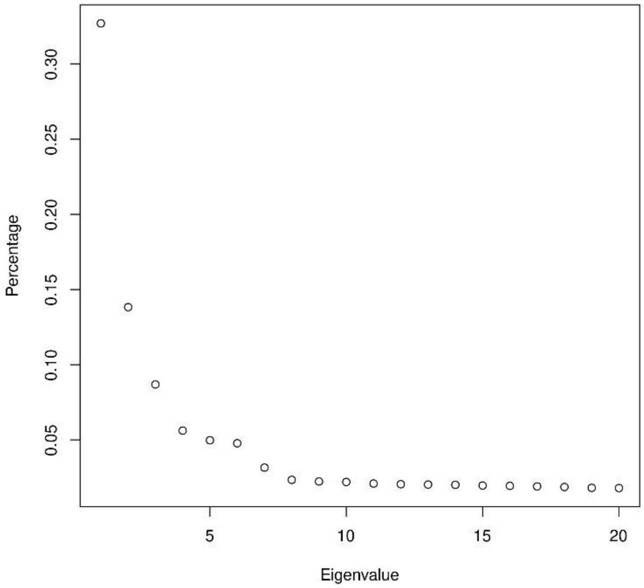
FIGURE S1. Results from GEMTools (colors represent the base and non-base individuals).



Principal component analysis (PCA) for population structure

We used PLINK 2.0 to calculate the first 20 PCAs (after linkage disequilibrium (LD) pruning of the SNPs, --indep 50 5 0.2). The first six PCAs explained around 70% of the variance discovered by the first 20 PCAs (Figure S2). Therefore, we used the first six PCAs to adjust for population structure.

FIGURE S2 The ratio of each eigenvalue to the sum of PCAs.



Heritability for different population prevalences

Table S4 shows the estimate of heritability for different population prevalences (using the first 6 PCAs as covariates). The source population for EGOS is from the Swedish National Patient Register (NPR) and most of the NORDiC cases can be found in NPR. Previously, we estimated 0.0087 as the population prevalence of OCD for individuals born in Sweden between 1982-1990 and have a diagnosis in NPR (2).

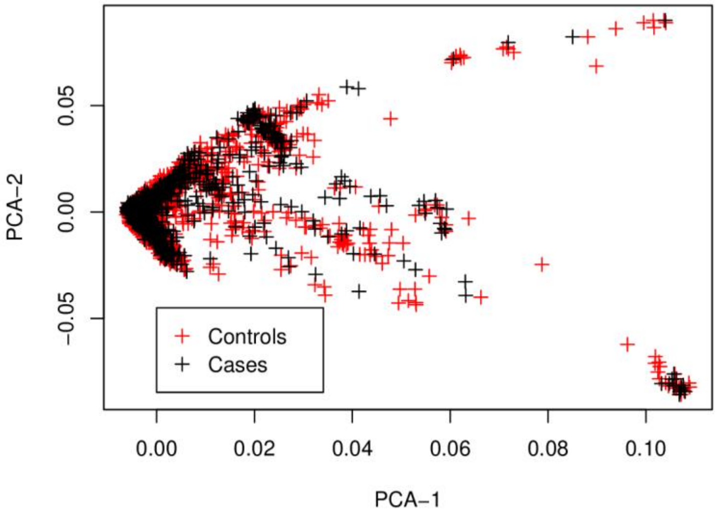
TABLE S4. Estimates of heritability for different population prevalence

Prevalence	heritability (SE)
0.005	25% (4%)
0.01	28% (4%)
0.015	32% (5%)
0.02	34% (5%)
0.025	36% (5%)
0.03	38% (6%)

Comparison of EGOS and NORDiC cases

Principal component analysis of the first two ancestry vectors for cases and controls are illustrated in Figure S3. For illustration purposes, we focused on individuals with PCA-1 < 0 (Figures S3).

FIGURE S3. First two ancestry vectors.



Figures S4 and S5 show the PCAs for the controls and cases, respectively (for PCA-1 < 0). Figures S6 and S7 show the PCAs for EGOS and NORDiC cases.

FIGURE S4. Controls, the first two ancestry vectors.

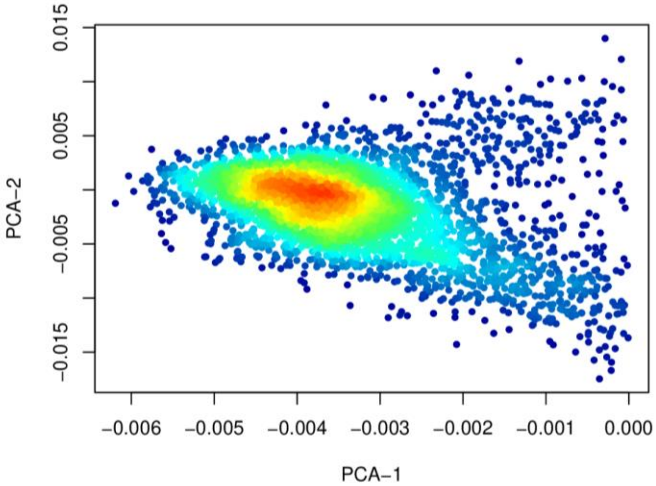


FIGURE S5. All cases, the first two ancestry vectors.

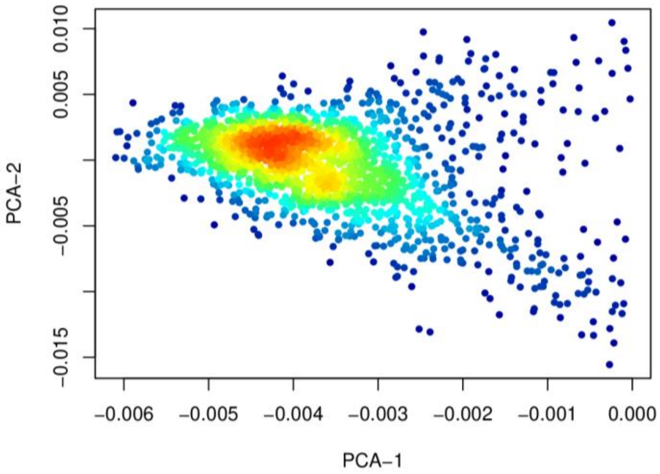


FIGURE S6. EGOS, the first two ancestry vectors.

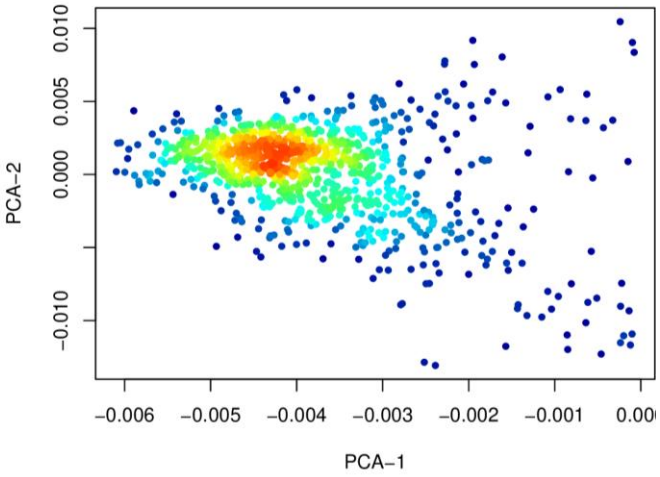
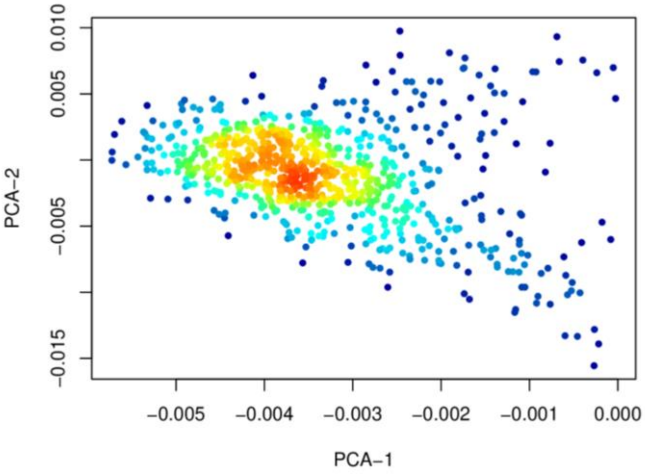


FIGURE S7. NORDiC, the first two ancestry vectors.



Comparison of Figures S6 and S7 suggests that EGOS and NORDiC cases have slightly different ancestry distribution. EGOS cases are more concentrated above zero for PCA-2. We observed a similar pattern in the histograms of PCA-2 in Figures S8 and S9. The ancestry distribution of EGOS cases was not a perfect match to that of controls. However, when EGOS and NORDiC were merged, their ancestry distribution matched the controls quite well (Figure S4 and S5).

FIGURE S8. EGOS cases, first two ancestry vectors.

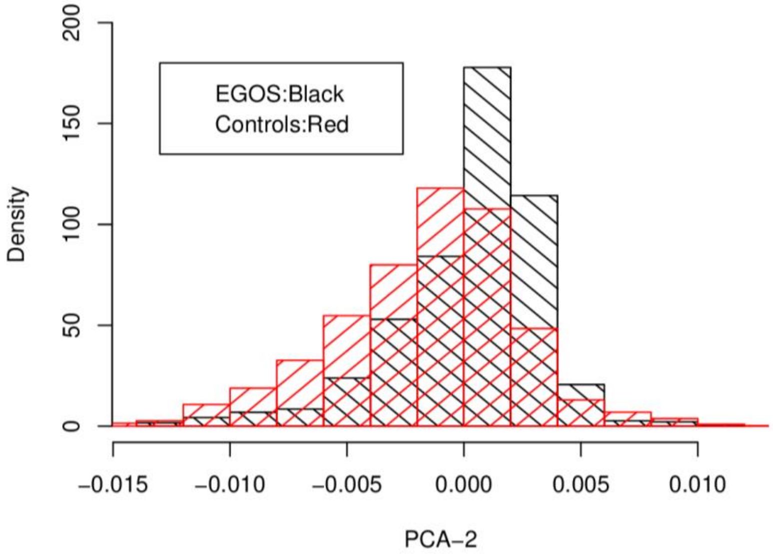
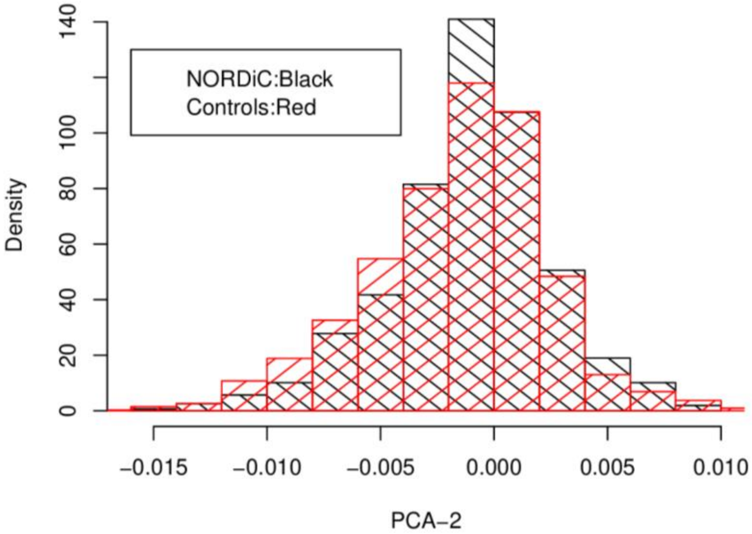


FIGURE S9. NORDiC cases, first two ancestry vectors.



We used 1:1 pair matching using PCA-1 and PCA-2 as the distance function (*pairmatch* function in R). EGOS and NORDiC cases had similar heritability after matching controls (Table S5).

TABLE S5. Estimates of heritability for EGOS and NORDiC cases.

Cohorts	**Heritability (SE)**
EGOS and matched controls	28% (11%)
NORDiC and matched controls	27% (12%)

Heritability analysis partitioned by MAF bins

TABLE S6. Heritability estimates for ten samples of size 180K SNPs. Sampling from each bin was proportional to the percentage of SNPs in that bin in the real data.

MAF	**SNPs**	**% of the total SNPs**	**Heritability (10 Samples)**											**% of heritability**
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**	
0.01-0.05	81360	45.2%	0.018	0.020	0.013	0.026	0.031	0.052	0.018	0.083	0.027	0.074	0.036	16.8%
0.05-0.1	21420	11.9%	0.000	0.009	0.010	0.008	0.000	0.015	0.002	0.009	0.001	0.000	0.005	2.5%
0.1-0.2	25920	14.3%	0.056	0.013	0.025	0.033	0.044	0.014	0.042	0.035	0.047	0.039	0.035	16.2%
0.2-0.3	19980	11.0%	0.089	0.059	0.075	0.064	0.066	0.054	0.045	0.056	0.082	0.052	0.064	29.9%
0.3-0.4	16200	8.9%	0.021	0.045	0.031	0.017	0.045	0.048	0.023	0.031	0.018	0.022	0.030	14.0%
0.4-0.5	15120	8.3%	0.041	0.056	0.060	0.038	0.031	0.048	0.047	0.046	0.039	0.035	0.044	20.6%
Total	180000	100%	0.226	0.201	0.214	0.186	0.217	0.231	0.176	0.261	0.213	0.221		

TABLE S7. Heritability estimates for ten samples of size 180K SNPs. Sampling from each bin was proportional to the percentage of SNPs in that bin from 1000G data.

MAF	**SNPs**	**% of the total SNPs**	**Heritability (10 Samples)**											**% of heritability**
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**	
0.01-0.05	53100	29.5%	0.008	0.015	0.023	0.033	0.019	0.029	0.046	0.031	0.021	0.013	0.024	10.9%
0.05-0.1	25200	14.0%	0.000	0.002	0.000	0.020	0.000	0.000	0.000	0.017	0.000	0.009	0.005	3.7%
0.1-0.2	32940	18.3%	0.053	0.051	0.062	0.025	0.059	0.030	0.044	0.032	0.039	0.045	0.044	17.2%
0.2-0.3	25200	14.0%	0.066	0.085	0.080	0.078	0.063	0.066	0.066	0.080	0.087	0.067	0.074	30.9%
0.3-0.4	22320	12.4%	0.033	0.025	0.020	0.028	0.010	0.059	0.034	0.014	0.028	0.024	0.027	12.1%
0.4-0.5	21240	11.8%	0.066	0.064	0.056	0.068	0.062	0.037	0.047	0.059	0.047	0.067	0.057	25.1%
Total	180000	100%	0.225	0.241	0.241	0.252	0.213	0.220	0.236	0.232	0.222	0.225		

TABLE S8. Heritability estimates for ten samples of size 180K SNPs. 30K samples from each bin.

MAF	SNPs	% of the total SNPs	Heritability (10 Samples)											% of heritability
			1	2	3	4	5	6	7	8	9	10	Mean	
0.01-0.05	30000	16.7%	0.008	0.000	0.015	0.002	0.014	0.000	0.005	0.028	0.015	0.051	0.014	5.9%
0.05-0.1	30000	16.7%	0.010	0.019	0.001	0.015	0.004	0.001	0.010	0.002	0.000	0.006	0.007	3.0%
0.1-0.2	30000	16.7%	0.048	0.040	0.037	0.031	0.051	0.018	0.043	0.033	0.032	0.053	0.039	16.5%
0.2-0.3	30000	16.7%	0.077	0.069	0.084	0.093	0.075	0.082	0.062	0.084	0.075	0.059	0.076	32.6%
0.3-0.4	30000	16.7%	0.026	0.032	0.036	0.019	0.024	0.031	0.040	0.042	0.038	0.031	0.032	13.7%
0.4-0.5	30000	16.7%	0.069	0.067	0.058	0.073	0.061	0.071	0.072	0.060	0.067	0.064	0.066	28.4%
Total	180000	100%	0.238	0.226	0.231	0.232	0.230	0.203	0.231	0.249	0.227	0.265		

TABLE S9. Estimates of heritability partitioned by MAF bins in this study, in the study of the IOCDF-GC sample (3), and proportional to 1000G data. For 1000G proportional to data, the estimate of heritability for each bin is the mean of heritability for that bin for ten samples of size 180K SNP; Sampling from each bin was proportional to the percentage of SNPs in that bin from 1000G data.

	This study			The IOCDF-GC sample			Expected (Proportional to 1000G)		
MAF	Heritability (SE)	SNPs (% of total)	% Heritability	Heritability (SE)	SNPs (% of total)	% Heritability	Heritability (SE)[2]	SNPs (% of total)	% Heritability
0.01-0.05	2.6% (3.7%)	183388 (45.2%)	10.0%	0.0001% (3%)[1]	19605 (5.2)	0%	2.4% (2.4%)	53100 (29.5%)	10.4%
0.05-0.1	0.0% (2.0%)	48313 (11.9%)	0.0%	4% (5%)	47976 (12.8)	11%	0.5% (1.2%)	25200 (14.0%)	2.2%
0.1-0.2	4.6% (2.5%)	58476 (14.4%)	17.7%	8% (6%)	91661 (24.5)	23%	4.4% (2.6%)	32940 (18.3%)	19.0%
0.2-0.3	9.8% (2.3%)	45347 (11.1%)	37.7%	1% (6%)	77193 (20.7)	3%	7.4% (1.6%)	25200 (14.0%)	32.0%
0.3-0.4	2.6% (2.1%)	36359 (9.0%)	10.0%	11% (5%)	70193 (18.7)	31%	2.7% (3.2%)	22320 (12.4%)	11.7%
0.4-0.5	6.4% (2.0%)	34181 (8.4%)	24.6%	11% (5%)	66770 (17.8)	31%	5.7% (2.1%)	21240 (11.8%)	24.7%
Sum	26.0%	406064	100%	35%	373398	100%	23.1%	180000	100%

[1] The reported boundary for this study was 0.001-0.05. [2] The estimated standard error based on the ten samples.

FIGURE S10. The proportion of expected and observed heritability explained by different minor allele frequencies (MAF) bins based on A) the real data; B) the average of ten samples of size 180K SNPs, sampling from each bin was proportional to the percentage of SNPs in that bin in the real data; C) the average of ten samples of size 180K SNPs, sampling from each bin was proportional to the percentage of SNPs in 1000 Genomes data; D) the average of ten samples of size 180K SNPs, 30K samples from each bin. MAFs were binned, and we used the average MAF in a bin to plot the results.

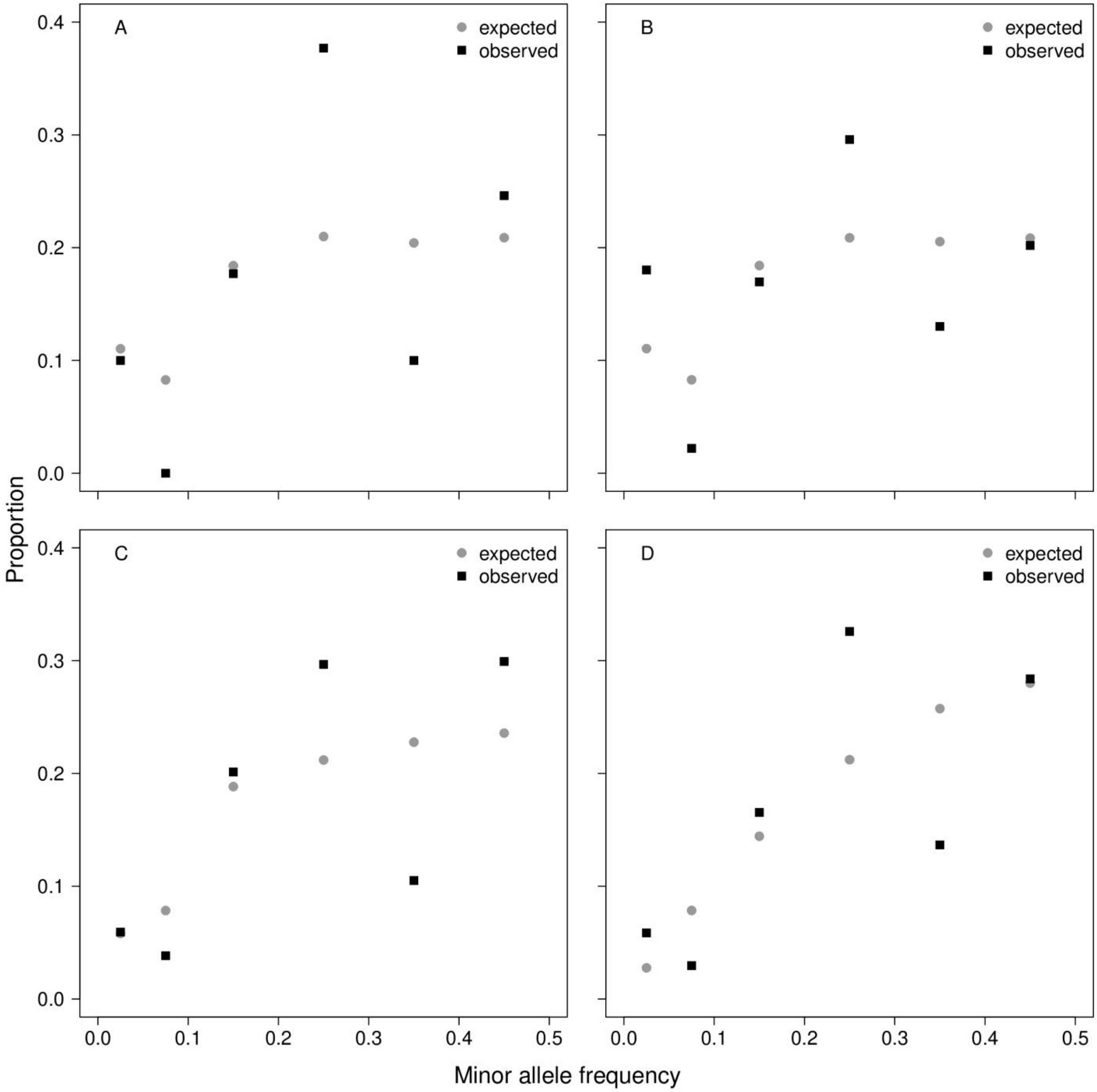
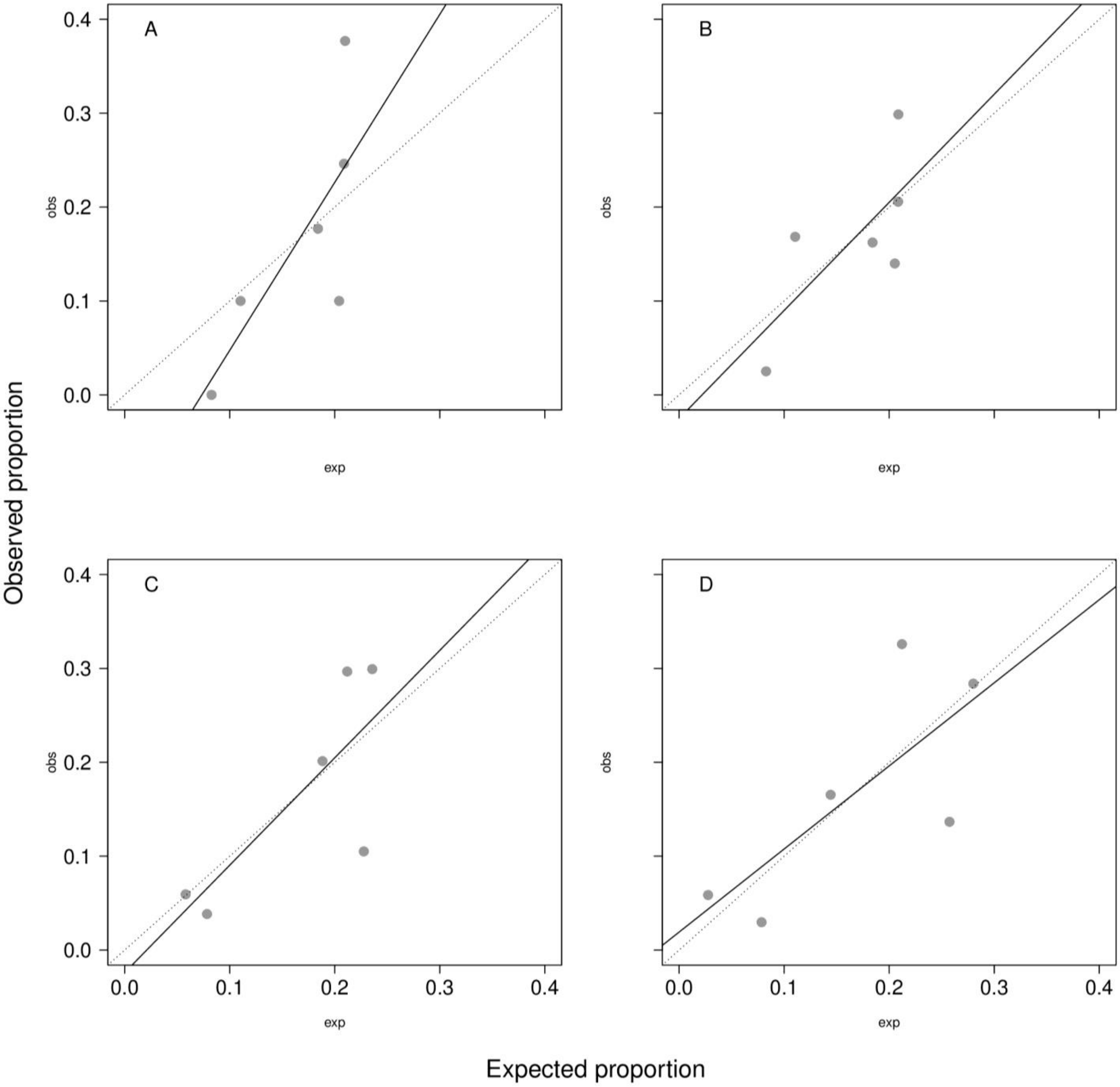


FIGURE S11. The observed proportion of heritability versus its expected proportion based on A) the real data (Adjusted R^2=0.46, p-value=0.082); B) the average of ten samples of size 180K SNPs, sampling from each bin was proportional to the percentage of SNPs in that bin in the real data (Adjusted R^2=0.40, p-value=0.107); C) the average of ten samples of size 180K SNPs, sampling from each bin was proportional to the percentage of SNPs in 1000G data (Adjusted R^2=0.49, p-value=0.073); D) the average of ten samples of size 180K SNPs, 30K samples from each bin. In each plot, the solid line is the regressed line, and the dashed line has slope one and intercept zero (observed=expected) (Adjusted R^2=0.45, p-value=0.085).



References

1. Crossett A, Lee AB, Klei L, et al.: Refining genetically inferred relationships using Treelet Covariance Smoothing. Ann Appl Stat 2013; 7:669–690

2. Mahjani B, Klei L, Hultman CM, et al.: Maternal Effects as Causes of Risk for Obsessive-Compulsive Disorder, in Biological Psychiatry. 2020, pp 1045–1051.

3. Davis LK, Yu D, Keenan CL, et al.: Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. PLoS Genet 2013; 9:e1003864