Supplementary Online Content

Colodro-Conde L, Couvy-Duchesne B, Whitfield JB, et al. Association between population density and genetic risk for schizophrenia. *JAMA Psychiatry*. Published online June 23, 2018. doi:10.1001/jamapsychiatry.2018.1581

This supplementary material has been provided by the authors to give readers additional information about their work.

eAppendix 1. Short Review on the Gap in Schizophrenia Rates Between Urban and Rural Environments

Greater rates of many mental disorders have been reported in urban (versus rural) environments[1,2] and the study of this non uniform distribution of schizophrenia cases has been consistently reported since the ecological studies from Faris and Dunham in 1939[3]. The largest meta-analysis to date[4] included ten studies or national registries from Europe, Australia and the United States and reported that the rate of psychosis in urban areas may be almost twice that of rural areas (OR [95% CI] = 1.72 [1.53, 1.92]). The meta-analysis studied the effect of possible confounders such as age, sex, ethnicity, drug use, social class, family history, season of birth, but none of these could explain the apparent association between urbanicity and psychotic disorders[4]. A more recent meta-analysis, of four studies reported the prevalence of schizophrenia and prodromal score associated with more than two levels of urbanicity exposure in Western countries (OR [95% CI] = 2.37 [2.01, 2.81]). Although most of the research has been conducted in Western societies, the results are consistent in countries like China. Furthermore, subtle expressions of psychosis (with prevalences of 10–20%) were also more prevalent among adolescents raised in urban areas.

Further research differentiated the associations between schizophrenia and place of birth, upbringing and residence. A population-based study in Denmark showed a dose-response relationship between urbanicity during upbringing (first 15 years of life) and schizophrenia risk. Thus, for individuals moving to a more urbanized place during their upbringing, the risk of schizophrenia increased, while for individuals moving to a lower degree of urbanization, the risk decreased. This study also observed that when urbanicity at birth and during upbringing were adjusted mutually, the effect of urbanicity at place of birth vanished while the place of upbringing remained significant. In addition, a Dutch study reported an effect of urban residence at the time of first admission to the clinic, regardless of whether individuals were born or not in urban areas. Finally, a study from the Swedish registry showed that the incidence of psychosis was positively associated with the population density of the area of residence in the Swedish adult population (ages 25-64 years old), thus refining the simplistic contrast of rural versus urban living[5]. Socio-economic status or social deprivation of the area has been also linked to higher rates of schizophrenia[6-8].

There are some excellent reviews in the literature discussing further the relationship between urbanicity and schizophrenia[9,10].

eAppendix 2. Genetic and Phenotypic Data From the QIMR Sample

Since 1980, a series of studies of general health conditions conducted by the Genetic Epidemiology Unit at QIMR have collected longitudinal phenotypic data and genotypes on more than 28,000 Australian twins and their family members[11,12]. For the present study we selected all genotyped participants over 18 years old, that is, a total N of 15,544 individuals from 7,015 families (65.6% females, age mean: 54.4, SD=13.2). Sex was self-reported and verified with the genotypes. Importantly, this is a community based sample representative of the general population according to a number of sociodemographic characteristics [11]. Participants were not screened for schizophrenia and we did not control for disease(s) status in the analyses.

Population density, remoteness and SES were based on the most recent data published by the Australian Bureau of Statistics (ABS), Australia's national statistical agency, in the ABS Census of Population and Housing (version 2016 for population density and 2011 for remoteness and SES). The ABS uses the Australian Statistical Geography Standard (ASGS) for the collection and dissemination of geographically classified statistics, which are comparable and spatially integrated. The boundaries of the units of analysis established by the ASGS take into account continuous changes relating to population and infrastructure [13]. We have linked the postal codes provided by the participants of the present study to the information presented by postal areas or the second level of statistical areas (statistical area 2 or SA2), which represent communities that interact together socially and economically.

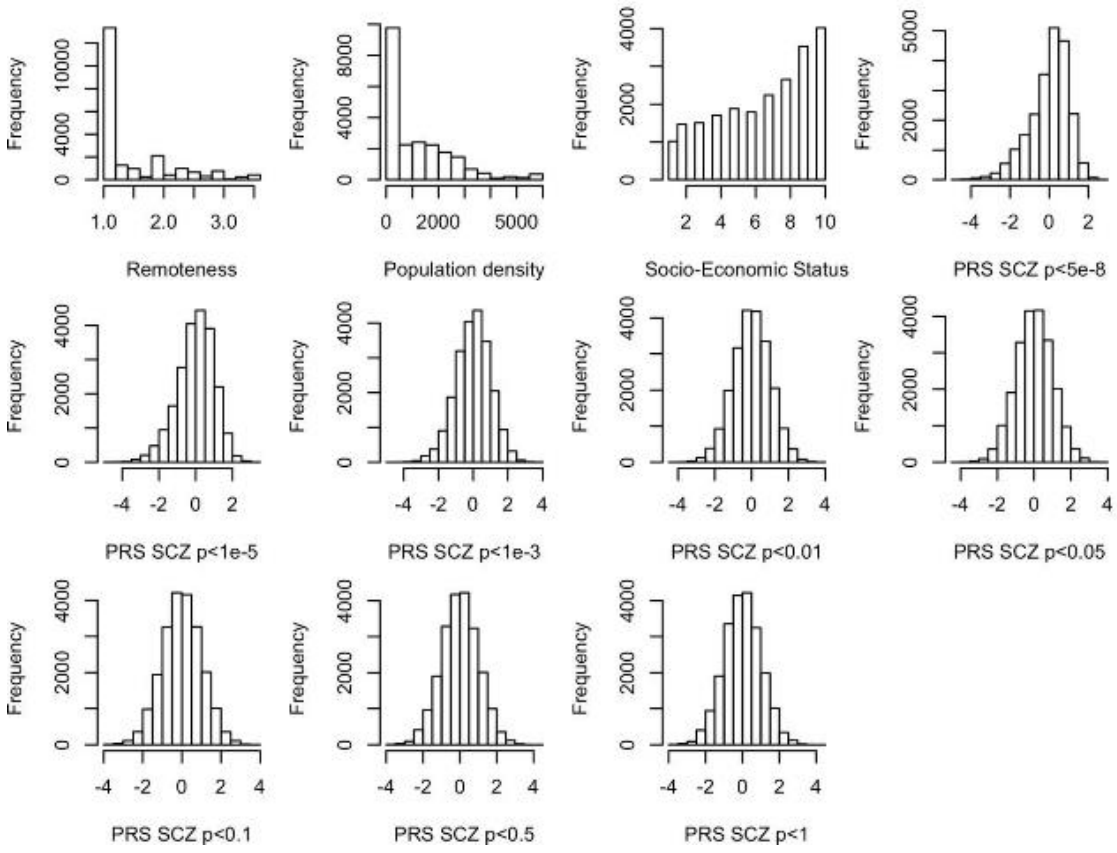
Population density was calculated by dividing the estimated resident population by the km^2 of each statistical area. Remoteness areas are regions in each state and territory, divided on the basis of their relative access to services[14] into five levels: major cities (1), inner regional (2), outer regional (3), remote (4) and very remote (5)[14]. We averaged the remoteness scores of participants living at the border of two different areas and treated the variable as continuous. Outlying values for both population density and remoteness were winsorized to 3 standard deviations. SES was based on the Index of Relative Socio-economic Advantage and Disadvantage (IRSAD)[15], which can be used to measure socioeconomic wellbeing in a continuum, from the most disadvantaged areas (low values) to the most advantaged areas (high values). We used deciles of this index, which allowed comparison within postal areas in the same state/territory. Thus, SES as defined in the present study is a function of the area where a person lives and not of any personal characteristics. See distribution of these variables in **eFigure 1**.

Participants were genotyped using commercial arrays (Illumina 317K, 370K, 610K, '1st generation', or Core Exome plus Omni-family, '2nd generation' [16-18]). Genotype data were cleaned (by batch) for call rate (≥95%); MAF (>=1%); Hardy-Weinberg equilibrium ($p \geq 10^{-3}$; PLINK1. 9 [19]), GenCall score (≥0.15 per genotype; mean ≥0.7) and standard Illumina filters. Data were checked for pedigree, sex and Mendelian errors and for non-European ancestry (6SD from the PC1 and/or PC2 means of European populations). DNA was imputed from a common SNP set to the 1000 Genomes (Phase 3 Release 5) 'mixed population' reference panel (http://www.1000genomes.org) [20,21], a strategy that allows genotype data from different arrays to be combined using a set of SNPs common to the first generation genotyping platforms. Imputation was performed on the Michigan Imputation Server [22] using the SHAPEIT/minimac Pipeline [23,24] and minimac3 [25] or in-house (chromosome X only). '1st generation' and '2nd generation' arrays, with 277,690 and 240,297 core observed markers respectively, were imputed separately due to poor overlap between markers and the two were combined after imputation to maximise sample size. A total of 9,411,304 SNPs were available for analysis, after QC. See **eTable 1** for the number of SNPs included in the each of the thresholds used in the p-value thresholds used in the PRS calculation and **eFigure 1** for distribution of PCs and PRS .

This study was approved by the QIMR Berghofer Medical Research Institute's Human Research Ethics Committee and the storage of the data follows national regulations regarding personal data protection. All the participants provided written informed consent and received a compensation for their participation in some of the studies.

eTable 1. Number of SNPs Included in the Calculation of Each of the QIMR Polygenic Risk Scores

P-value cut-off	N SNPs
$< 5e^{-8}$	96
$< 1e^{-5}$	681
$< 1e^{-3}$	5,078
< 0.01	18,916
< 0.05	50,979
< 0.1	79,569
< 0.5	210,975
< 1	283,990



eFigure 1. Histograms of the Main Variables Used in the Analysis (QIMR)

eAppendix 3. Genetic and Phenotypic Data From in the UK Biobank

The UK Biobank (UKB) collected genetic data on 487,409 unselected volunteers from all over the UK, and we retained the 456,426 participants of European ancestry in the analysis (see below). We used self-reported sex and age (at baseline assessment) of the participants.

From the Easting and Northing coordinates rounded to the kilometre we used https://www.doogal.co.uk/BatchReverseGeocoding.php and the R package *ggmap*[26] to perform reverse geocoding and identify the postcode district that the participants likely lived in (1,182 unique postcode districts in the sample). We crossed this information with the population density by postcode district calculated from the 2011 census. We downloaded the 2011 British and Welsh census data from https://www.nomisweb.co.uk/census/2011/key_statistics, and we obtained the Scottish data from http://www.scotlandscensus.gov.uk/ods-analyser/. See distribution of variables in **eFigure 2**. Ancestry assignment used a two-stage approach. First, the UKB sample was projected onto the first two principle components from the 2,504 participants in the 1000G project. Projections used HM3 SNP with minor allele frequency (MAF) > 0.01 in both datasets and were calculated using GCTA (v1.90.0beta). Allele frequencies for the 1000G GRM were the mean allele frequency from each dataset, weighted by the number of samples. Next, participants from the UKB were assigned to one of the five super-populations from the 1000G project (European, African, East-Asian, South-Asian and Admixed). Assignments for European, African, East-Asian and South-Asian ancestries were based on >0.9 posterior-probability of belonging to the 1000G reference cluster, with the remaining participants classified as Admixed. Posterior-probabilities were calculated under a bivariate Gaussian distribution where this approach generalizes the k-means method to take account of the shape of the reference cluster. We used a uniform prior and calculated the vectors of means and 2x2 variance-covariance matrices for each super-population. A preliminary assignment was made, before filtered of variants for Hardy-Weinberg Equilibrium (pHWE, $p<10^{-6}$). The two steps in the ancestry assignment were repeated. This resulted in the reassignment of only one individual into the EUR set. In the final assignment there were 456,426 Europeans.

Variants from the UKB supplied .bgen files with minor allele count (MAC)>5 and info score>0.3 were extracted for all individuals. Genotype probabilities were converted to hard-call genotypes using the --hard-call 0.1 function in PLINK2 and variants with missingness>0.05 were excluded. Variants were renamed according to chromosome and allele calls to identify the HRC imputed variants. Thus, multi-allelic variants are included in the dataset.

As the UKB identified a problem with non-HRC imputed variants, we used the unrelated European subset of individuals to filter HRC variants for info score>0.3 missingness<0.05, pHWE<10^{-6}, and MAC>5. This resulted in a set of 28,730,841 variants, which were subsequently extracted for both the European (EUR) and unrelated European (EURu) datasets. Variants names were converted back to rsIDs.

Principal components were calculated with genotyped variants used by the ukb and passing additional QC filters (as applied in the EURu set). Genotyped SNP used by the UKB had already been LD pruned ($r^2<0.1$) and had long-range LD regions removed (Table S12 UKB QC documentation). There were 137,102 SNP included in the analysis. Principal components were calculated for the EURu set using flashPCA [27] then projected onto the complete EUR set (see **eFigure 2** and **eTable 2** and **eFigure 3** for distribution).

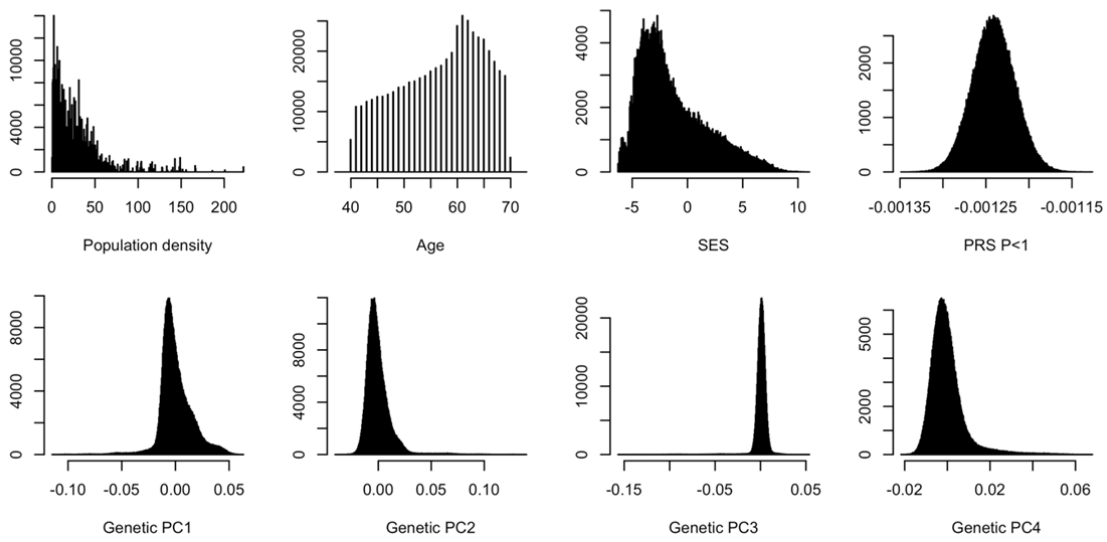
Polygenic risk scores were available on 345,246 unrelated UKB participants (GRM<0.05). First, GWA summary statistics were QCed and aligned to UKBB genotype data. This involved reducing SNPs to MAF > 0.01, INFO > 0.60 (if available) and HWE > $1e^{-6}$. If allele frequencies were not available for summary statistics we imputed them from UKBB data itself and then reduced to MAF > 0.01. SNPs in UKBB genotype data were also reduced to MAF > 0.01 independently of GWA summary statistics. Next, we matched GWA summary statistics with UKBB using rsID (indels were aligned using genomic position and IDs were formatted to match across both datasets). We then checked agreement across both alleles. All non-matching alleles were checked for allele switch (A1=A2 and A2=A1), plus for reciprocal strand (e.g., C=G and T=A). Following this, we checked frequency agreement. If ID's matched, genomic positions and both alleles matched, then we expected minor allele to match too (frequency test was restricted to Frq < 0.4 and Frq > 0.6). Finally, we excluded all ambiguous SNPs (CG, AT, with frequencies between 0.4-0.6) and removed MHC regions except single, most significant hit. Clumping was performed in PLINK (v1.9) [28] using LD calculated from the raw data and standard clumping filters: P1=0, P2=0, r2 = 0.1 and window 500kb.

Informed consent was obtained from all UKB participants. Procedures are controlled by a dedicated Ethics and Guidance Council (http://www.ukbiobank.ac.uk/ethics), with the Ethics and Governance Framework available at http://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf. IRB approval was also

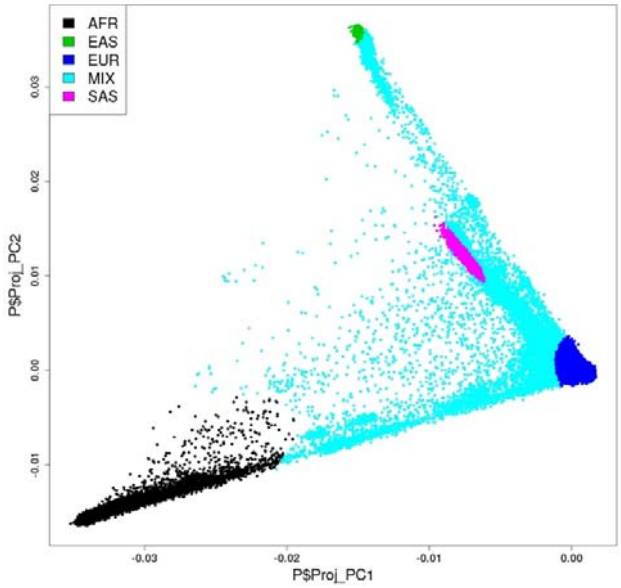
obtained from the North West Multi-centre Research Ethics Committee. This research has been conducted using the UK Biobank Resource under Application Number 12505.

eTable 2. UKB Sample Breakdown by Ancestry

Population	N
African (AFR)	8,231
East Asian (EAS)	1,021
European (EUR)	456,426
Admixed (MIX)	18,435
South Asian (SAS)	3,364



eFigure 2. Histograms of the Main Variables Used in the Analysis (UKB)

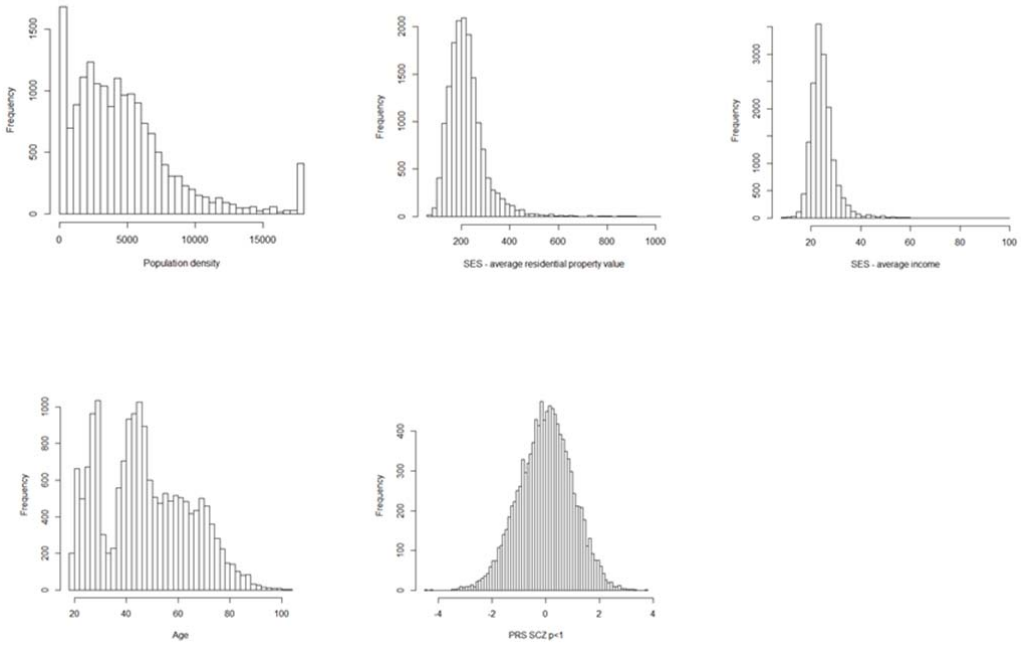


eFigure 3. Scatter Plot of the Genetic Principal Component Projection That Was Used to Determine the Ancestry of Participants (UKB)

eAppendix 4. Genetic and Phenotypic Data From the NTR Sample

The Netherlands Twin Register was established in 1987 by the department of Biological Psychology from the Vrije Universiteit Amsterdam and has collected longitudinal phenotype data in adult twins and their family members by means of a 2- to 3-yearly survey on health, lifestyle and personality[29]. For the replication study, all genotyped participants over 18 year with data on population density were selected, resulting in a total N of 16,434 (59% females, mean age: 47.7 (SD = 17.0) from 6,322 families. Outlying values for population density were winsorized to 3 standard deviations. Neighbourhood level data on average income and/or average property value were available for a subset (N=15,722). Polygenic scores were available for a large subsample of the participants with genotype data (N = 11,212). See **eTable 4** for distributions.

Participants were genotyped using commercial arrays (Illumina Human Quad Beadcip 660, Illumina Omni 1M, Illumina GSA, Affymetrix 6.0, Affymetrix Axiom, Perlegen-Affymetrix). Imputation was performed using the 1000 Genomes (Phase 1 Release 3) reference panel. For the PRSs analysis, SNPs were cleaned for call rate (>90%), MAF (>=0.5%), Hardy-Weinberg equilibrium ($p>10^{-5}$) and Mendelian error rate (N<40). Samples were checked for pedigree, sex, heterozygosity (<-.075 and >.075), genotype call rate (>90%) and Mendelian error rate (SD<5). PRSs were calculated using LDpred, a method that adjusts the effect size for each locus in a linkage disequilibrium (LD) block[30]. For the GWAS analysis, SNPs were cleaned for call rate (>95%), MAF (>=1%), Hardy-Weinberg equilibrium ($p>10^{-5}$) and Mendelian error rate (SD<3). Samples were checked for pedigree, sex, heterozygosity (<-.10 and >.10), genotype call rate (>90%).



eFigure 4. Histograms of the Main Variables Used in the Analysis (NTR)

eAppendix 5. Genetic and Phenotypic Data From the QSKIN Sample

The QSkin Sun and Health Study is a cohort of 43,794 men and women aged 40–69 years randomly sampled from the population of Queensland, Australia in 2011[31]. The cohort was established to study the development of skin cancer and melanoma and collected basic demographic and health information at baseline. For the present study we selected all genotyped participants who did not participate in QIMR studies and who were unrelated (GRM<0.1). The sample for these analyses was composed of 15,726 individuals (54.7% females, age mean: 57, SD = 7.9). Sex was self-reported. Importantly, this is a randomly selected community based sample and as such representative of the general population living in Queensland. Participants were not screened for schizophrenia and we did not control for disease(s) status in the analyses. Note that QIMR was not restricted to Queensland.

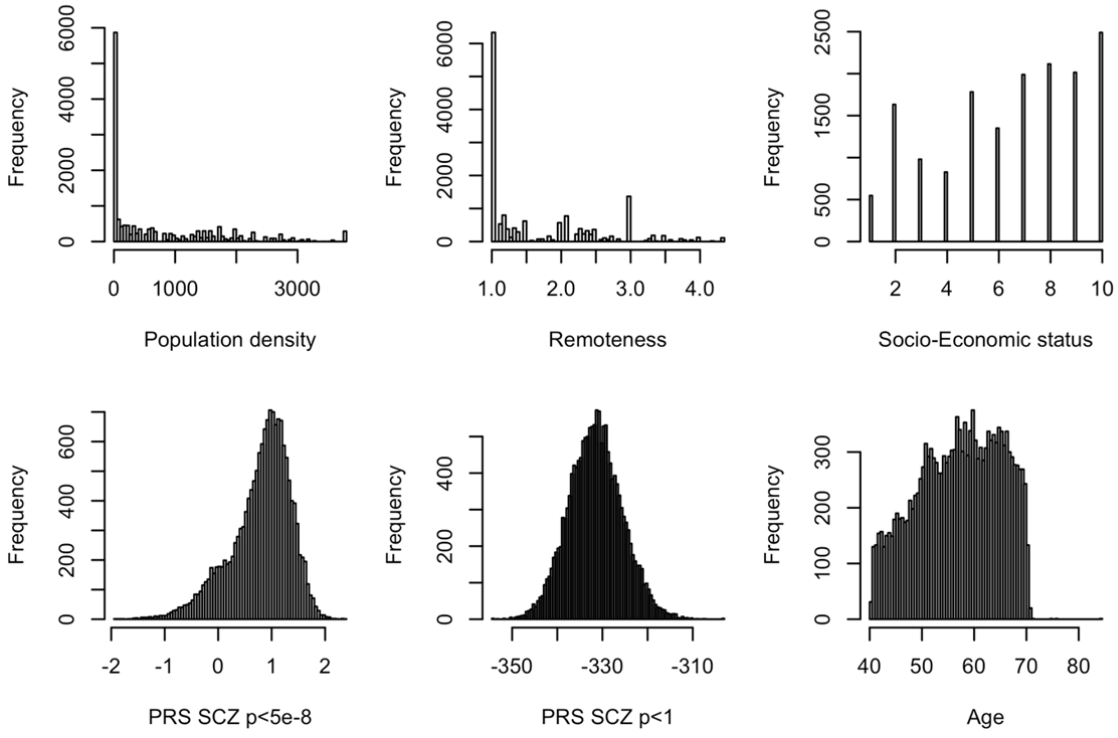
Population density, remoteness and SES were computed following the protocol developed for the QIMR sample. See distribution of these variables in **eFigure 5**.

Participants were genotyped using commercial arrays (Illumina GSA chip). Genotype data were cleaned for call rate (≥95%); MAF (>=1%); Hardy-Weinberg equilibrium ($p \geq 10^{-6}$; PLINK1. 9 [19]), GenCall score (≥0.15 per genotype; mean ≥0.7) and standard Illumina filters. Data were checked for pedigree, sex and Mendelian errors and for non-European ancestry (6SD from the PC1 and/or PC2 means of European populations). DNA was imputed to the 1000 Genomes (Phase 3 Release 5) 'mixed population' reference panel (http://www.1000genomes.org) [20,21]. Imputation was performed on the Michigan Imputation Server [22] using the SHAPEIT/minimac Pipeline [23,24] and minimac3 [25]. A total of 6,735,109 SNPs were available for analysis, after QC. See **eTable 3** for the number of SNPs included in the each of the thresholds used in the p-value thresholds used in the PRS calculation and **eFigure 5** for distribution of PRS.

This study was approved by the QIMR Berghofer Medical Research Institute's Human Research Ethics Committee (P1309; P2034) and the storage of the data follows national regulations regarding personal data protection. All the participants provided written informed consent.

eTable 3. Number of SNPs Included in the Calculation of Each of the QSKIN Polygenic Risk Scores

P-value cut-off	N SNPs
$< 5e^{-8}$	149
$< 1e^{-5}$	709
$< 1e^{-3}$	5,489
< 0.01	21,168
< 0.05	58,620
< 0.1	92,547
< 0.5	250,694
< 1	341,013



eFigure 5. Histograms of the Variables Used in the Analysis (QSKIN)

eAppendix 6. Summary on Twin and Family Studies

Twin and family studies can be genetically informative by contrasting the similarities between pairs of individuals as a function of their different degrees of relatedness. These designs allow the estimation of the contributions of genetic, shared environmental, and residual (also known as unshared or unique environmental) variation[32] to the observed variance of a trait. Heritability is defined as the proportion of variation in a trait that is due to heritable differences between individuals in a population [33].

eAppendix 7. GE Moderator Effect Model

The model is as follows[34]:

Population density = μ + bCovariates +b_1Age + (a + b_aAge)A + (c + b_cAge)C + (e + b_eAge)E

With b and b_1 the fixed effects parameters of age and other covariates effects on population density mean; a, c, e, b_a, b_c, and b_e the parameters of the random effects. A, C and E are the random effect of additive genetics, shared environment and unique environment assumed to follow a normal distribution of mean 0 and of known variance covariance (kinship matrix) for A, family indicator for C, and identity matrix for E.

eAppendix 8. PRS and Prediction Model Used

PRS provide a quantitative measure of the cumulative genetic risk or vulnerability that an individual possesses for a trait. Using GWAS results from an independent sample, PRS are estimated as the sum of risk alleles weighted by their respective independently estimated effect sizes. Importantly, the genetic risk is independent of the occurrence of the disease [35,36].

In order to estimate the variance in population density explained by the PRS, we fit linear mixed models which controlled for relatedness and demographic covariates. The parameters of the model were estimated using GCTA 1.26.0 [37] (Student's t-test, two-tailed) to test the significance of the fixed effects that accounts for twin relatedness using a Genetic Relatedness Matrix (GRM). The linear model used to estimate the variance in population density and remoteness explained by the PRS is as follows:

Population density, remoteness or SES = intercept + Covariates*b + c*PRS + G

With b, c the vectors of fixed effects. The outcome variables and PRS are column vectors of size N, Covariates is a matrix of N row and p columns, b is a vector of p rows, c is a vector of one number.

Covariates = sex, age, age^2, sex*age, sex*age^2, a dummy variable for the GWAS array used, the first 4 genetic principal components. We performed the analyses both with and without SES as a covariate (except for those analyses with SES as outcome).

G is the random effect that models the sample relatedness G~ N(0, GRM*$\boldsymbol{\sigma^2_G}$), with GRM the N*N matrix of relatedness estimated from SNPs (with N the number of individuals, 15,544).

Formally, we performed 40 tests. However, after taking into account the correlation between the variables the number of effective tests is 16 [38-40] (see https://neurogenetics.qimrberghofer.edu.au/ SNPSpDlite/ for scripts and details). Therefore we used a significance threshold of $3.15*10^{-3}$ (Bonferroni correction). This was calculated from an estimated four effective PRS (out of eight used) and four effective phenotypes (out of five considered: population density, remoteness, SES as well as the first two after regressing SES). This approach is a fast and efficient alternative to permutation testing[38], when testing correlated variables.

In the NTR, UKB and QSKIN samples, we used a significance threshold of 0.05 as we aimed to replicate (where available) the results found with the PRS calculated over all independent genomic regions ($p<1$).

We tested for the presence of interactions in the QIMR sample. First, we tested for sex specific effects by adding an interaction term (PRS*sex) to the model and using residualized PRS (we regressed out all covariates but sex and its interactions) to remove the confounding effect of the covariates[41]. Similarly, we tested for a PRS by age interaction.

Due to the large sample size of the UKB, fitting the same mixed model in GCTA would be extremely time and memory consuming. To overcome this problem, we divided the sample into chunks of 50,000 participants and meta-analysed the results obtained from GCTA.

eAppendix 9. Summary on Mendelian Randomization (MR)

Multi instrument MR relates the effect sizes of the exposure SNPs (i.e. genome wide significant SNPs for schizophrenia) to their effect sizes on the outcome (from GWAS of population density and remoteness). The slope is an estimate of the causal effect of the exposure on the outcome. Meta-analysis methods are used to estimate this slope as effect sizes have different precision (from different MAF or different number of observations)[42]. Note that the inference of causation is limited by the instrument SNPs included, that likely only represents a fraction of the biological pathways involved in schizophrenia, due to limited power in GWAS.

MR-base provides estimates of MR using several meta-analysis approaches: fixed effects random effects, maximum likelihood, MR Egger, Weighted Median and Inverse Variance Weighted. In addition, it tests if the results may be confounded by pleiotropy (i.e. different SNPs in LD acting on each phenotype) or by a handful of SNPs (heterogeneity test, like in meta-analyses)[43]. MR-Egger is currently seen as the most robust multi-instrument mendelian randomization approach [42].

First, we merged the GWAS results from the PGC and from our sample, and selected only the genome-wide significant SNPs for schizophrenia. Then, we clumped the SNP list, keeping the most associated SNP (with schizophrenia) per haplotype (default options in MR base: 10MBp window, r^2<0.01 using 1000genomes as reference). This maximises the number of SNPs, hence the power of the analysis. In addition, we used GSMR (Generalised Summary-data-based Mendelian Randomisation)[44], which is on average more powerful than the MR Egger approach (as it models residual LD between SNPs) and more robust (as it allows testing for pleiotropy at a SNP level: HEIDI test[44], vs. testing across all the instruments in MR Egger). We included the same list of pruned SNP in the GSMR analysis and excluded instruments that showed significant pleiotropy (HEIDI test<0.01). We calculated the LD matrix between usable SNPs from our sample using PLINK and GCTA (see http://cnsgenomics.com/software/gsmr/ for code and examples).

To ensure that the results and MR estimates would be comparable across sample we performed the MR analyses on standardised effect sizes (z-scores)[44]. To ensure a fair comparison between the methods, we excluded the SNPs failing the HEIDI test in all the results presented.

We performed reverse MR analysis (testing if population density or SES cause schizophrenia) using the handful of GW significant instruments observed in the UKB (when adjusting SES for population density and conversely). The small number (12) of instrumental variables (SNPs) associated with population density or SES, limits the power of MR analyses and prevent from drawing robust conclusion from the analyses. We presented the results in **eAppendix 15**.

eAppendix 10. Phenotypic, Genetic and Environmental Correlations (95% CI and p-values) Between the Demographic Variables

We present the phenotypic, genetic and environmental correlations of the main variables in the QIMR cohort in **eTable 4**.

In the UKB, the phenotypic correlation between SES and population density was 0.44 (95% CI: 0.43-0.45, $p<1e^{-16}$).

In the NTR, the phenotypic correlation between SES (average residential property value) and population density was -0.37 (95% CI: -0.40 to -0.35, p<.001) and between SES (average income) was -0.12 (95% CI: -0.14 to -0.10, p<.001).

In QSKIN, the phenotypic correlation between population density and remoteness was -0.58 (95%CI: -0.59, -0.57, $p<2.2e^{-16}$), between SES and population density was 0.51 (95%CI: 0.50-0.52, $p<2.2e^{-16}$), and between SES and remoteness was -0.50 (95%CI: 0.52-0.49, $p<2.2e^{-16}$).

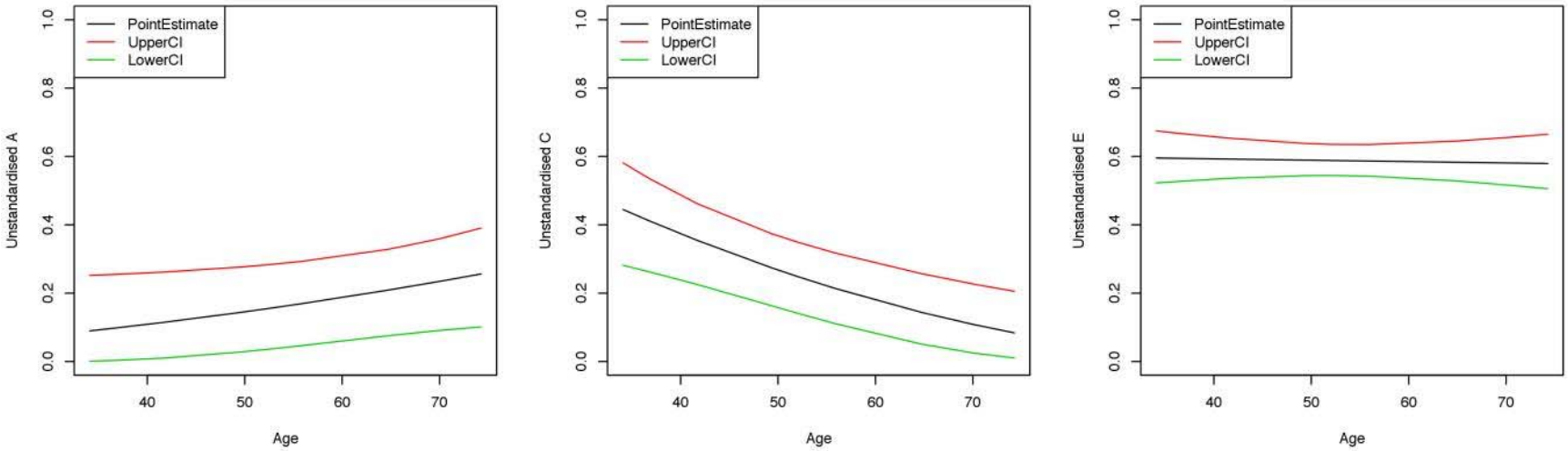
eTable 4. Correlations (and 95% Confidence Intervals) in the QIMR Sample

		Remoteness	SES
Phenotypic correlations	Population density	-0.58 (-0.60 to -0.56) *	0.46 (0.44 to 0.48) *
	Remoteness		-0.48 (-0.50 to -0.46) *
Genetic correlations	Population density	-0.74 (-1.00 to -0.21)	0.35 (-0.77 to 1.0)
	Remoteness		-0.51 (-1.00 to 1.00)
Shared Environment correlations	Population density	-0.66 (-0.84 to -0.44)	0.86 (0.62 to 1.00)*
	Remoteness		-0.65 (-0.86 to -0.43)*
Unique environment correlations	Population density	-0.50 (-0.54 to -0.46) *	0.33 (0.28 to 0.37) *
	Remoteness		-0.40 (-0.45 to -0.35) *

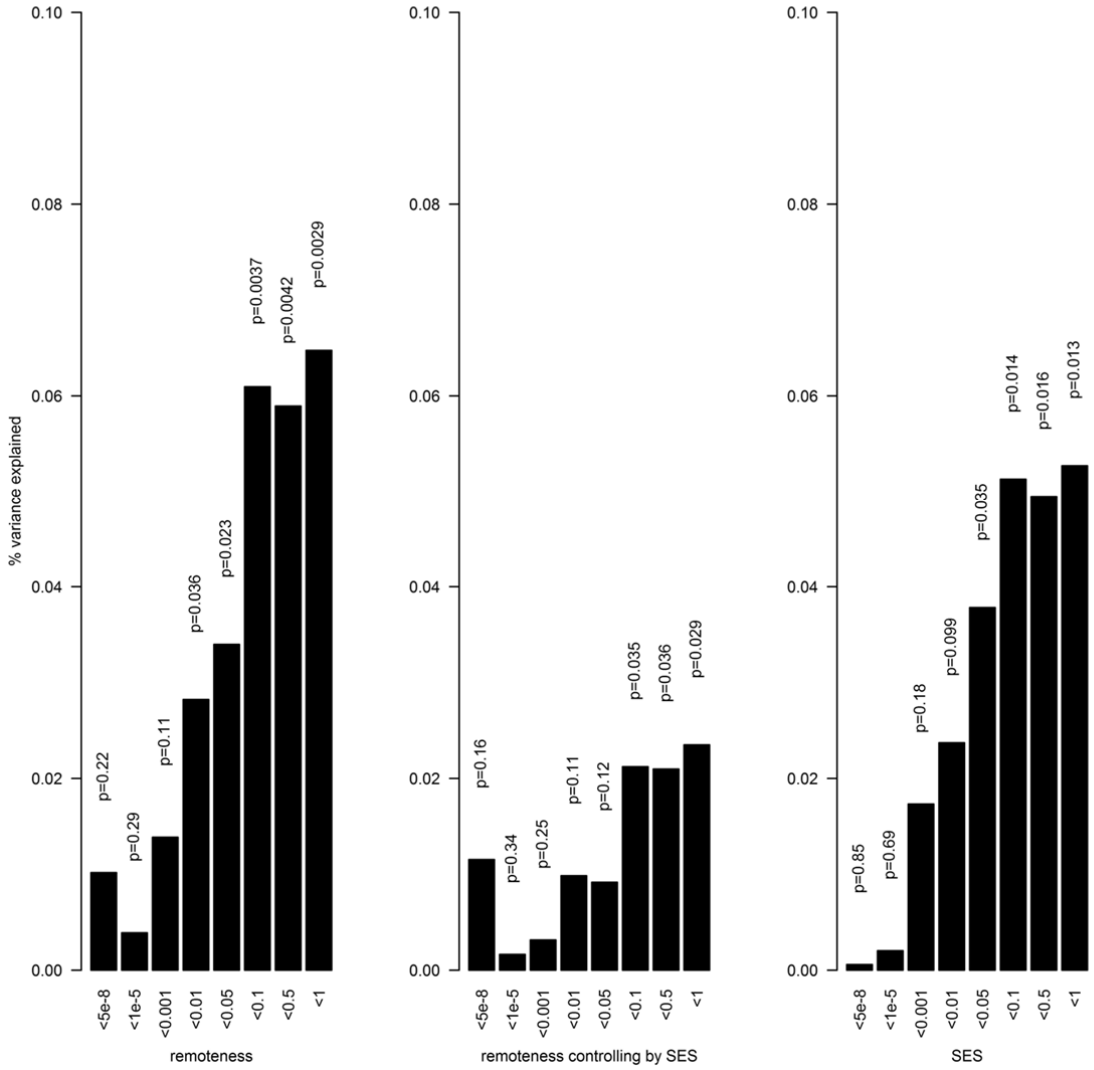
Phenotypic, genetic and environmental correlations were calculated in OpenMx [45], which allows modelling the sample relatedness and shared environment.

* Correlations significant ($p<0.0001$) after correcting for multiple testing.

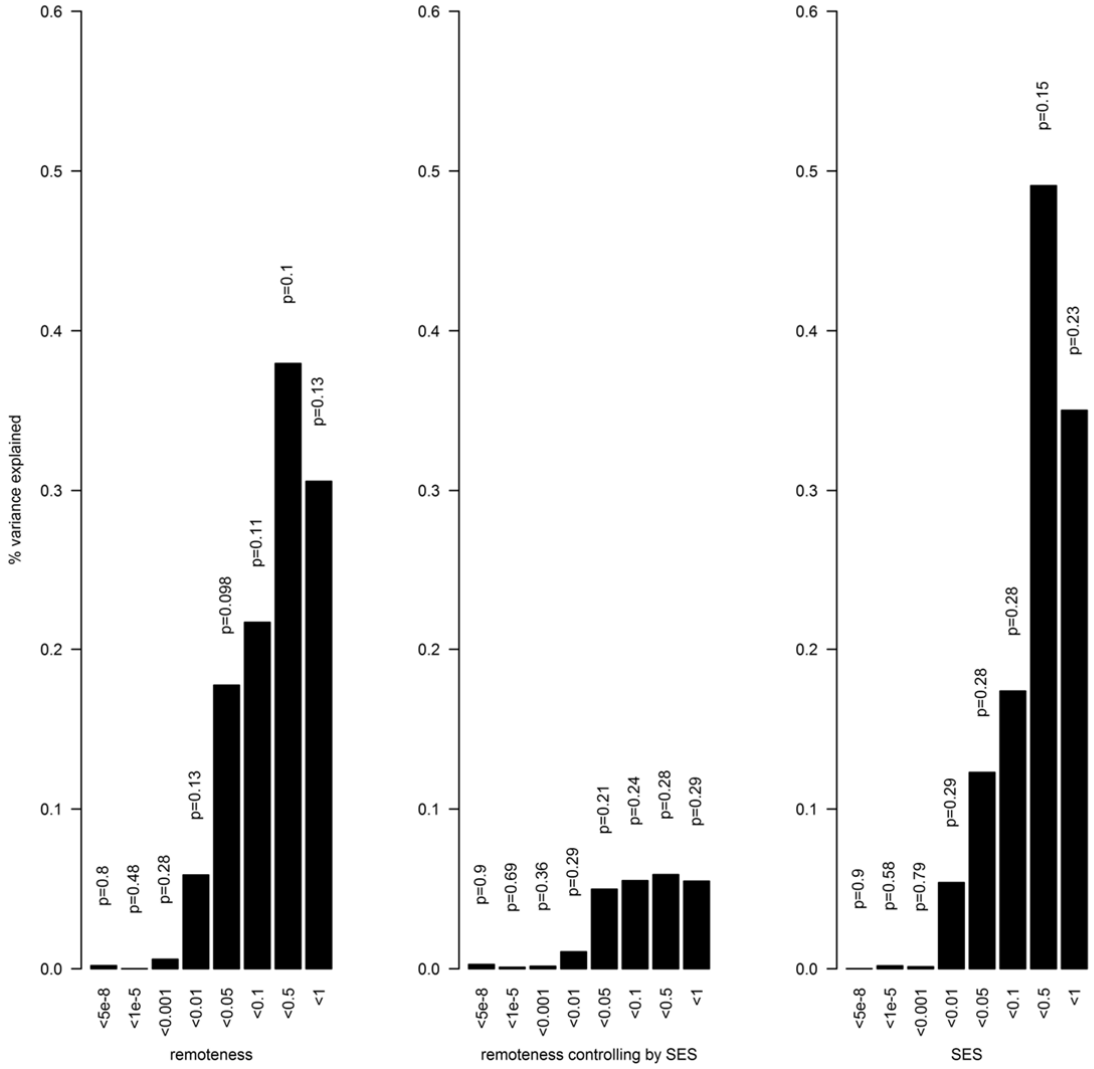
eAppendix 11. Effect of Age on the Genetic (A) and Environmental (Common, C and Unique, E) Sources of Variances for Population Density



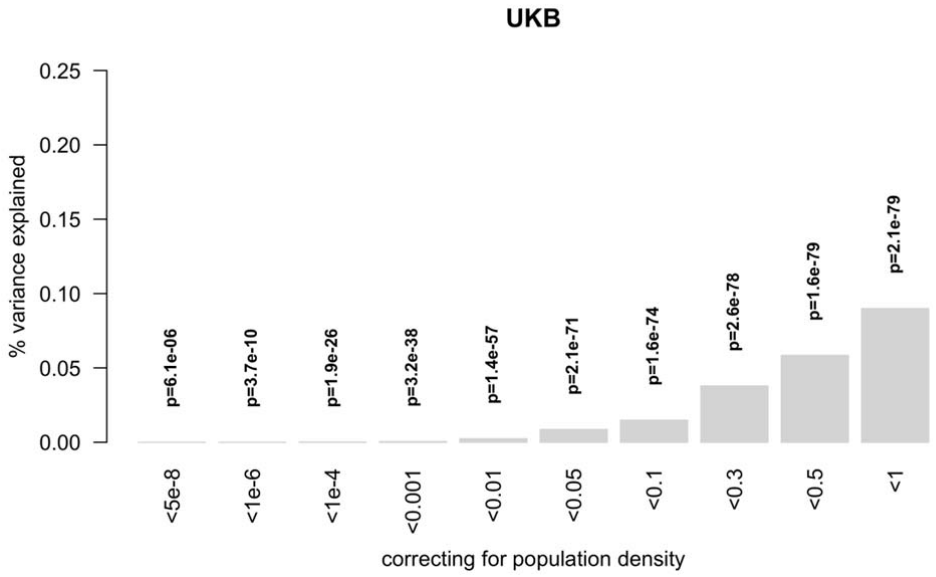
eAppendix 12. Variance of Remoteness and SES Explained by the Genetic Risk for Schizophrenia.



eFigure 6. Variance of Remoteness and SES Explained by the Genetic Risk for Schizophrenia (QIMR)



eFigure 7. Variance of Remoteness and SES Explained by the Genetic Risk for Schizophrenia (QSKIN)



eFigure 8. Variance SES Explained by the Genetic Risk for Schizophrenia (UKB)

eAppendix 13. GWAS Results

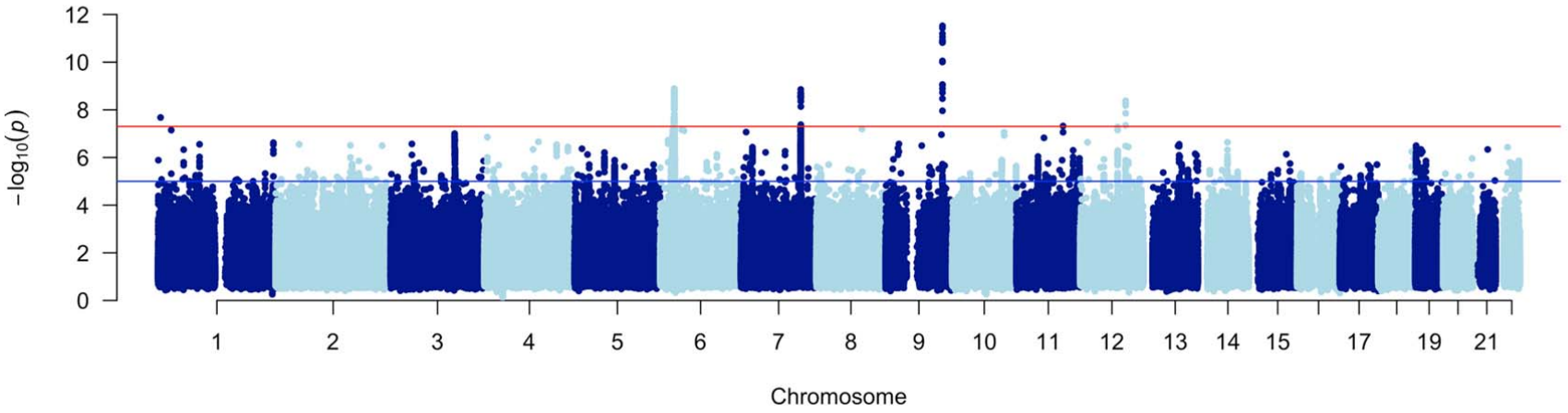
We present here the Manhattan plots for the GWAS of population density, SES, as well as population density corrected by SES and SES corrected by population density performed on the UKB (**eFigures 9 to 12)**, the GWAS of population density for QIMR, NTR and QSKIN **(eFigures 13 to 15**) and the GWAS meta-analysis of population density, which attenuated the results found in UKB (**eFigure 16**). We performed an inverse variance weighted meta-analysis using METAL[46].

eTable 5 SNP heritability and genetic correlation of population density and SCZ. These estimates are obtained using LD score regression from GWAS of population density performed in QIMR, UKB, NTR and QSKIN in this project and from the GWAS meta-analysis of SCZ from the 2014 Psychiatric genomics consortium [47]. The low SNP heritability (SNP_h^2) captured in the QIMR GWAS did not allow the calculation of the genetic correlation (r_g) of these results with the rest. The largest correlation was found for UKB with NTR (r_g=0.61, SE= 0.29, p-value=0.04), what could explain our results in the meta-analysis for population density. The genetic correlation between population density and schizophrenia was only significant when estimated from the population density GWAS in the UKB and cohort.

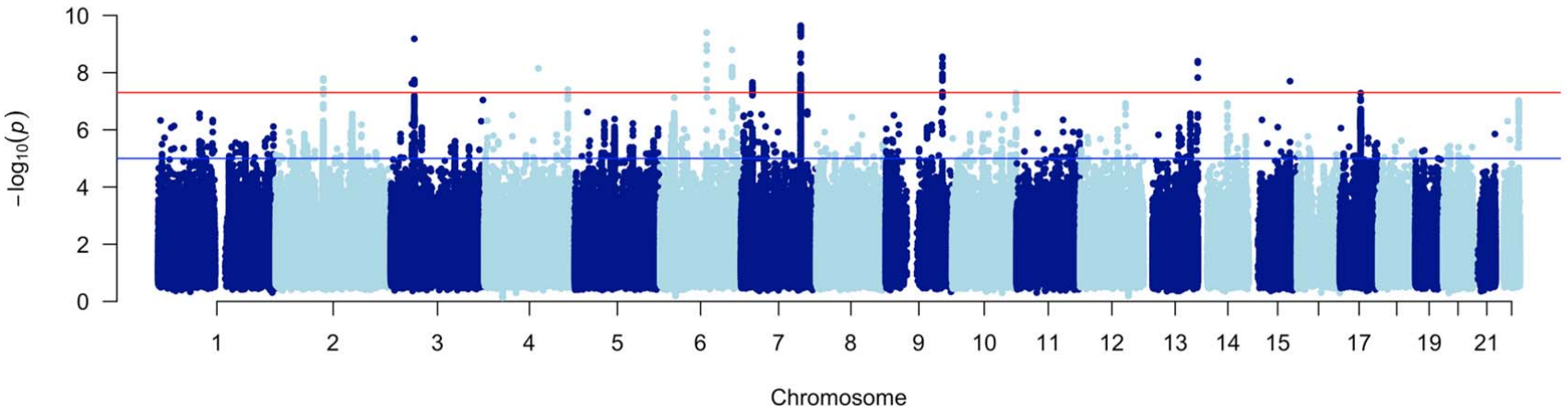
eTable 5. SNP heritability (SNP_h^2) and standard error (SE) from the univariate analyses (left column) and genetic correlations (r_g) and standard error (SE) from the bivariate analyses (right columns) performed in LD score for the GWAS results used in the present study.

	SNP_h^2 (SE)	r_g (p-value)				
		QIMR	UKB	NTR	QSKIN	SCZ
QIMR	0.0056 (0.0328)	1				
UKB	0.0222 (0.0017)	-	1			
NTR	0.0318 (0.0271)	-	0.6121(0.2929)	1		
QSKIN	0.0926 (0.0401)	-	0.5038(0.1417)	0.3017(0.4412)	1	
SCZ	0.2365 (0.0094)	-	0.2285(0.0384)	0.226 (0.1491)	0.0137(0.0774)	1

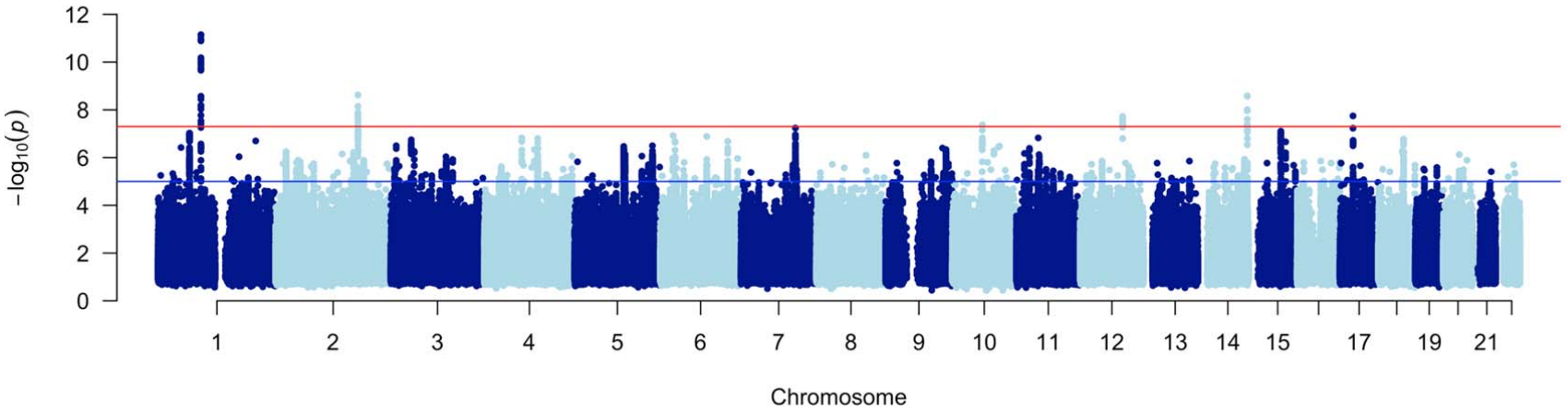
Note: due to the low SNP_h^2 for QIMR, the r_g could not be computed. Differences in genetic correlations may be reflecting differences in the phenotypic correlations between population density and SES in the European and Australian samples.



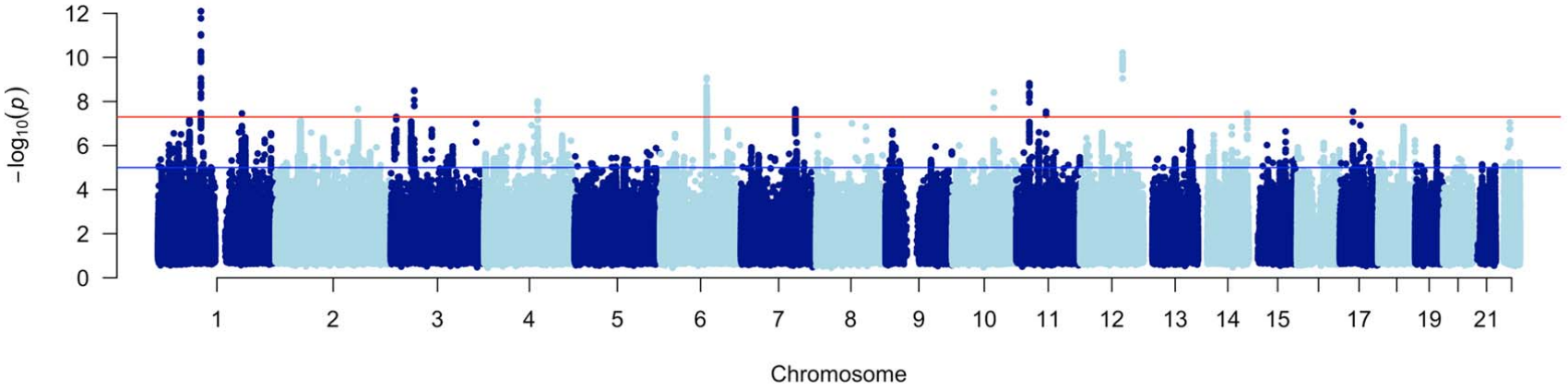
eFigure 9. Manhattan Plot of Population Density of Place of Residence (UKB)



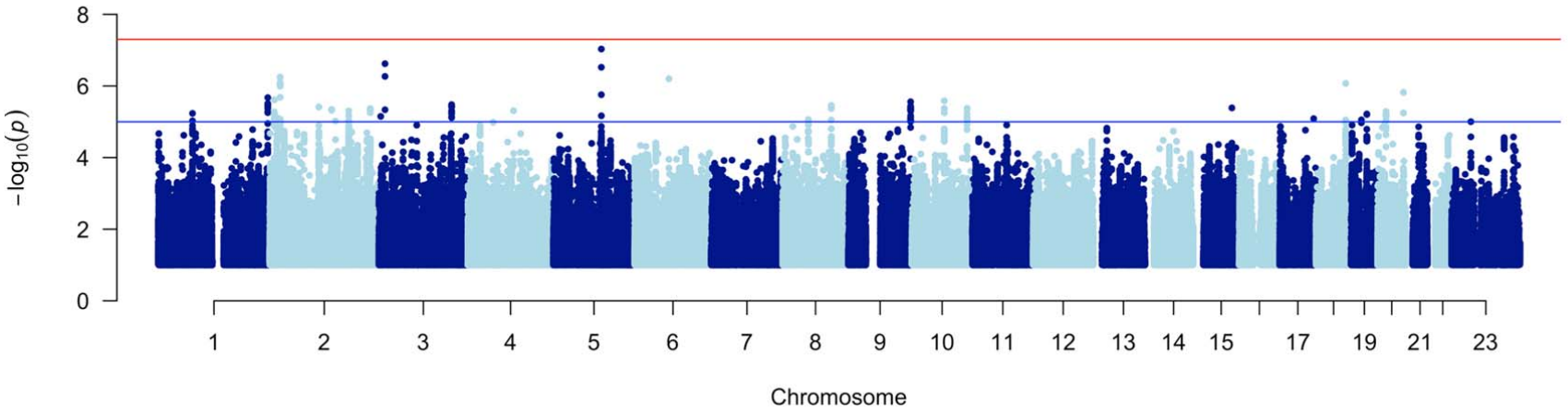
eFigure 10. Manhattan Plot of Population Density of Residence, Correcting for SES (UKB)



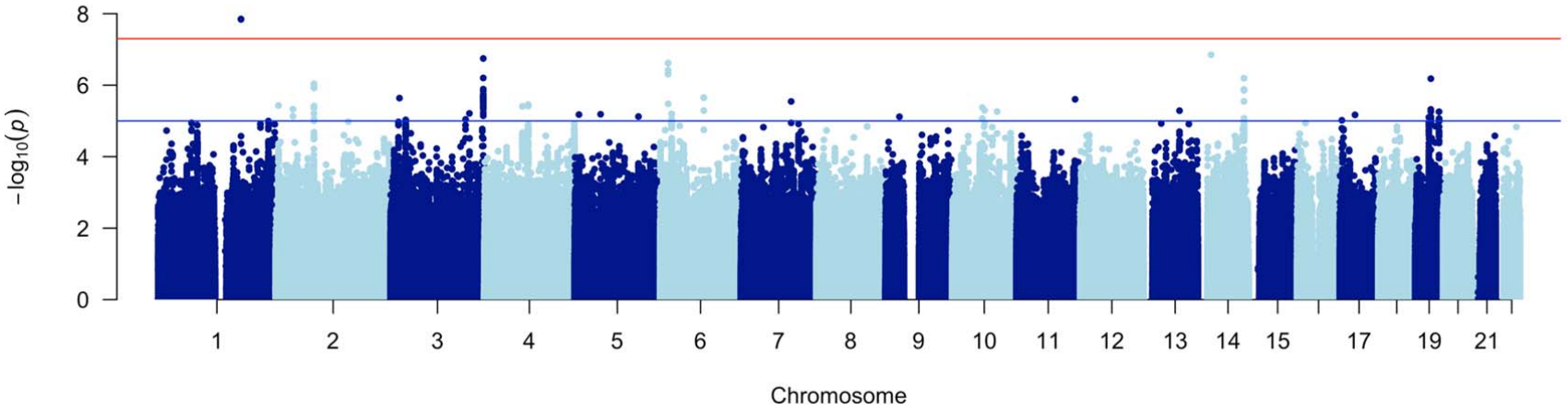
eFigure 11. Manhattan Plot of SES of Residence (UKB)



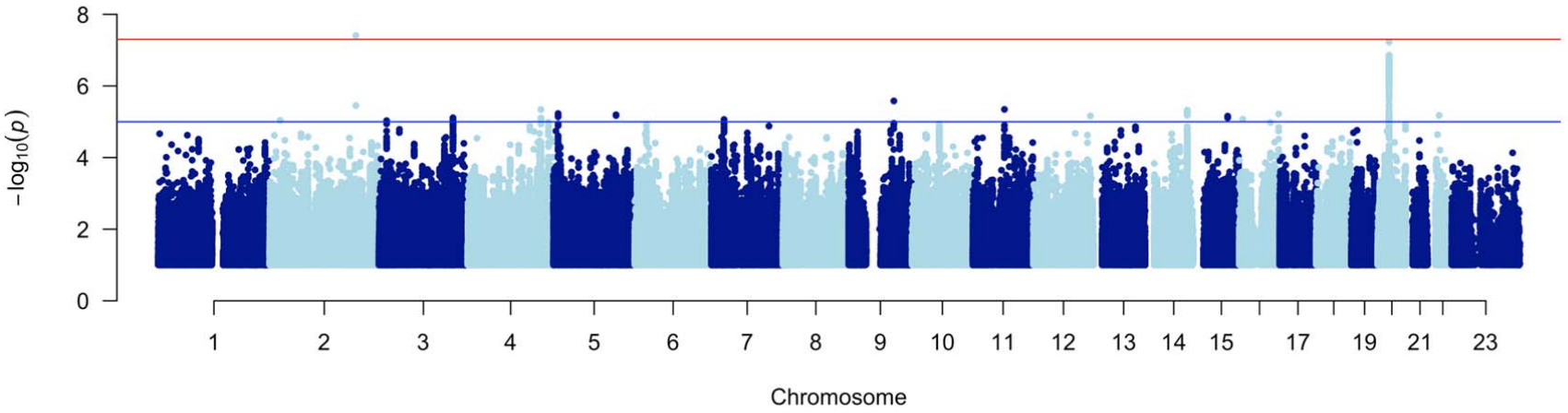
eFigure 12. Manhattan Plot of SES, Correcting for Population Density (UKB)



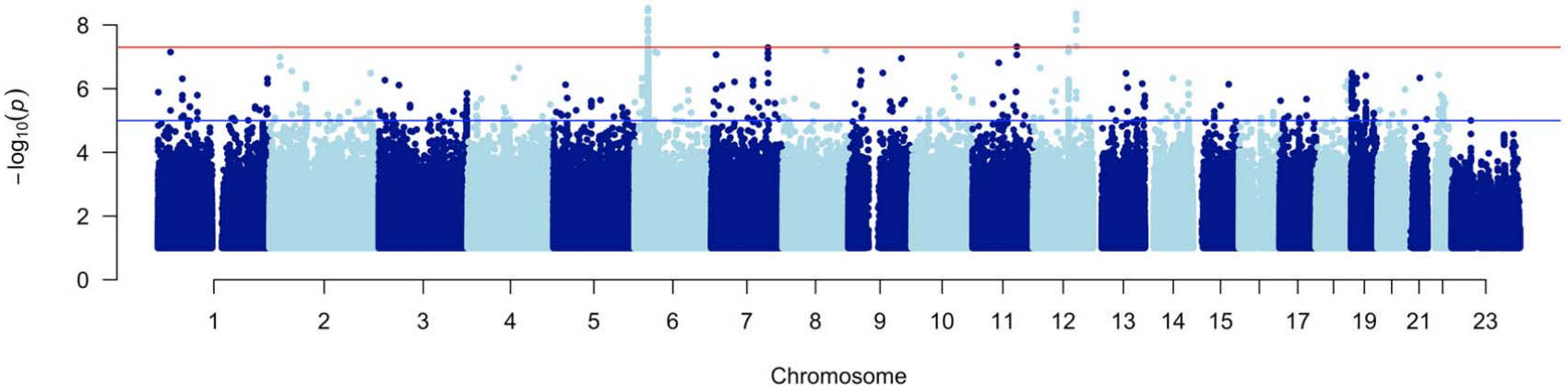
eFigure 13. Manhattan Plot of Population Density of Residence in the (QIMR)



eFigure 14. Manhattan Plot of Population Density of Residence (NTR)



eFigure 15. Manhattan Plot of Population Density of Residence (QSKIN)



eFigure 16. Manhattan Plot of Population Density of Residence in All Cohorts (Meta-analysis of UKB+QIMR+NTR+QSKIN)

eAppendix 14: Detailed MR Results

eTable 6. MR Results Testing the Hypothesis That Schizophrenia Is Causal of Population Density of Residence (Correcting for SES in the GWAS of Population Density) in the QIMR (left panel) and UKB (right panel) Cohorts

	QIMR				UKB			
Method	N SNP	b	SE	p-value	N SNP	b	SE	p-value
Fixed effects meta-analysis (simple SE)	90	0.079	0.051	0.11	91	0.035	0.0095	0.00022
Fixed effects meta-analysis (delta method)	90	0.082	0.051	0.11	91	0.031	0.0097	0.0011
Random effects meta-analysis (delta method)	90	0.081	0.052	0.11	91	0.028	0.013	0.038
Maximum likelihood	90	0.081	0.051	0.11	91	0.037	0.0098	0.00016
MR Egger	90	0.37	0.25	0.14	91	0.12	0.065	0.051
Weighted median	90	0.13	0.076	0.073	91	0.035	0.0159	0.025
Inverse variance weighted	90	0.079	0.053	0.13	91	0.035	0.014	0.012
GSMR	90	0.082	0.051	0.11	91	0.032	0.0098	0.00094

Even when accounting for SES in the GWAS, we observed a significant positive causal relationship between Schizophrenia and population density.

eTable 7. MR Hypothesis That Schizophrenia Causes to Live in Deprived Neighbourhood (measured by SES of the area) in UKB.

Method	N SNP	b	SE	p-value
Fixed effects meta-analysis (simple SE)	94	0.056	0.0093	$1.48e^{-9}$
Fixed effects meta-analysis (delta method)	94	0.050	0.0096	$1.31e^{-7}$
Random effects meta-analysis (delta method)	94	0.051	0.013	0.00021
Maximum likelihood	94	0.060	0.0097	$5.22e^{-10}$
MR Egger	94	0.077	0.065	0.23
Weighted median	94	0.049	0.015	0.0012
Inverse variance weighted	94	0.056	0.014	$6.94e^{-5}$
GSMR	94	0.049	0.0097	$2.60e^{-7}$

Schizophrenia also causes to live in areas with low SES (places with greater deprivation scores).

eTable 8. MR Hypothesis That Schizophrenia Causes to Live in Deprived Neighbourhood (measured by SES of the area) after adjusting for population density in UKB.

Method	N SNP	b	SE	p-value
Fixed effects meta-analysis (simple SE)	94	0.034	0.0093	0.00026
Fixed effects meta-analysis (delta method)	94	0.030	0.0095	0.0013
Random effects meta-analysis (delta method)	94	0.032	0.013	0.018
Maximum likelihood	94	0.036	0.0096	0.00018
MR Egger	94	0.020	0.062	0.74
Weighted median	94	0.025	0.016	0.12
Inverse variance weighted	94	0.034	0.014	0.017
GSMR	94	0.029	0.00964	0.0020

We still observe the causal relationship between SCZ and deprivation when correcting the GWAS of SES score for population density.

eTable 9. Reverse MR Hypothesis That Population Density Corrected for SES Is Causal of Schizophrenia in UKB.

Method	N SNP	b	SE	p-value
Fixed effects meta-analysis (simple SE)	12	0.22	0.085	0.0073
Fixed effects meta-analysis (delta method)	12	0.22	0.086	0.010
Random effects meta-analysis (delta method)	12	0.22	0.086	0.010
Maximum likelihood	12	0.23	0.086	0.0069
MR Egger	12	0.54	0.26	0.061
Weighted median	12	0.30	0.11	0.0078
Inverse variance weighted	12	0.22	0.085	0.0073
GSMR	12	0.22	0.086	0.010

Here, we use as MR instruments the 12 genomic regions reaching genome wide significance in the population density GWAS (correcting for SES). The p-values are suggestive of a reverse positive causal relationship between population density and schizophrenia. The magnitude of the causal effect of population density on SCZ may be 5 time the one of SCZ on population density (0.20 vs. 0.04). Our results should be interpreted with caution as only 12 SNPs are included in the multi-instrument MR.

eTable 10. Reverse MR Hypothesis That Deprived Neighbourhood (measured as SES) Adjusted for Population Density Is Causal for Schizophrenia in UKB.

Method	N SNP	b	SE	p-value
Fixed effects meta-analysis (simple SE)	12	0.23	0.083	0.0055
Fixed effects meta-analysis (delta method)	12	0.20	0.087	0.016
Random effects meta-analysis (delta method)	12	0.22	0.13	0.090
Maximum likelihood	12	0.24	0.087	0.0045
MR Egger	12	-0.19	0.52	0.71
Weighted median	12	0.17	0.13	0.17
Inverse variance weighted	12	0.23	0.13	0.082
GSMR	12	0.20	0.087	0.0165

Here, we use as MR instruments the 13 genomic regions reaching genome wide significance in the population density SES (correcting for population density). One SNP was excluded by the HEIDI test. The p-values are suggestive of a reverse positive causal relationship between population density and schizophrenia. Our results should be interpreted with caution as only 12 SNPs are included in the multi-instrument MR.

eAppendix 15. Sensitivity Analysis in the UKB

Recent publications have highlighted the fact that PRS for schizophrenia could be strongly associated with genetic principal components, with large differences in PRS distribution across ancestry groups[48]. This implies that population admixture (between continents) can have a confounding effect on PRS analyses. Another recent paper focussed on the UKB sample and showed local disparities of PRSs across the UK, raising the question of the influence of within-European admixture on the results[49]. This is an even greater concern in our analyses since we use measures of population density and SES that are also spatially distributed and relate to the household income and education level for which the concerns were raised[49].

In the following, we investigated these issues in the UKB sample and performed sensitivity analyses to attempt to confirm the results of our PRS analysis. In the analyses presented in the main text, we aimed to model the full European ancestry of the sample by fitting a random genetic effect, which is equivalent to fitting jointly all the genetic PCs.

PRS of schizophrenia and population density are associated with PCs in the UKB

Note that we used PCs calculated from the European set only, thus likely represent within-Europe genetic components. As such the first PCs strongly correlate with the Easting and Northing coordinates that align with Welsh-British and Scottish-British gradients in ancestry. We observed strong associations between our PRS for schizophrenia and genetic PCs. Beyond mean differences, the variance of the PRS was also dependent on the PCs (**eFigure 17**), which confirms the heteroscedasticity issue reported before[48,49]. In addition, we also observed associations (in mean and variance) between population density and genetic PCs (**eFigure 18**), which confirms their possible confounding effect in the analyses.

Sensitivity analyses

In order to confirm the results presented in the main text, we performed several sensitivity analyses:

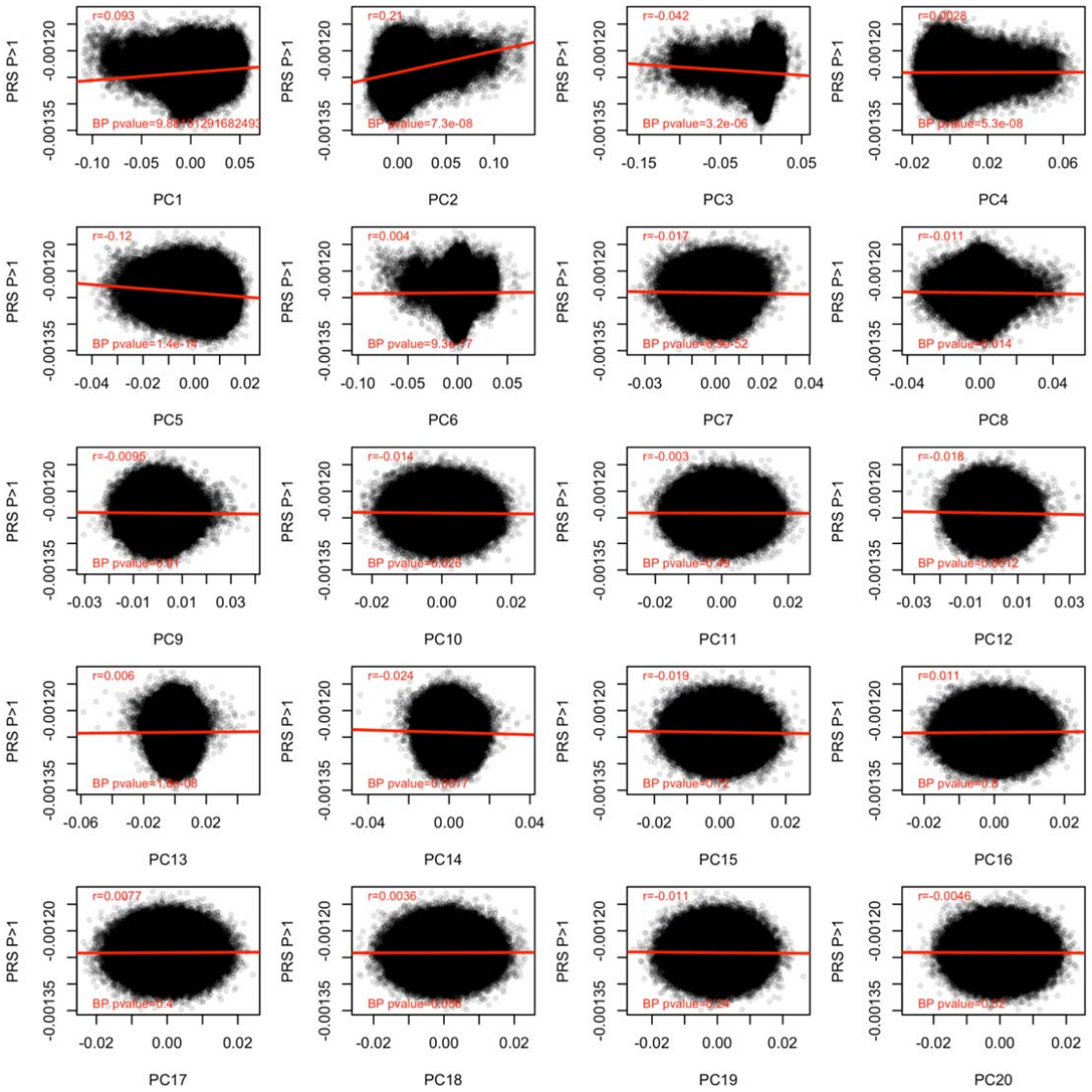
1) We fit linear models of prs ~ age + sex + ses + population density + PCs
2) We fit linear models of prs ~ age + sex + ses + population density + PCs after restricting the sample to genetically homogeneous participants (first 20 PCs in -0.02 0.02)

In each analysis, we used the PRS corresponding to "p<1" and "p<0.05". The first one was the most significant in our analyses, the second one was suggested to be the most predictive of schizophrenia[50] and it should be less influenced by ancestry as it contains fewer SNPs. We also varied the number of PCs used (20 or 40).

In the first sensitivity analysis, using 20PCs, we observed a significant correlation between PRS and population density (PRS "p<1": r=0.0066, r^2=0.0046%, p-value=3.8e^{-4}; PRS "p<0.05": r=0.0075, r^2=0.0057%, p-value=6.5e^{-5}). Though, the Breusch-Pagan (BP) test suggested presence of heteroscedasticity of the residuals (BP pvalue=0.0014 and pvalue=0.0054). When including 40PCs, the association persisted (PRS "p<1": r=0.0068, r^2=0.0047%, p-value=2.5e^{-4}; PRS "p<0.05": r=0.0079, r^2=0.0062%, pvalue=3.2e^{-5}), with less evidence of heteroscedasticity (pvalue=0.055 and pvalue=0.026).

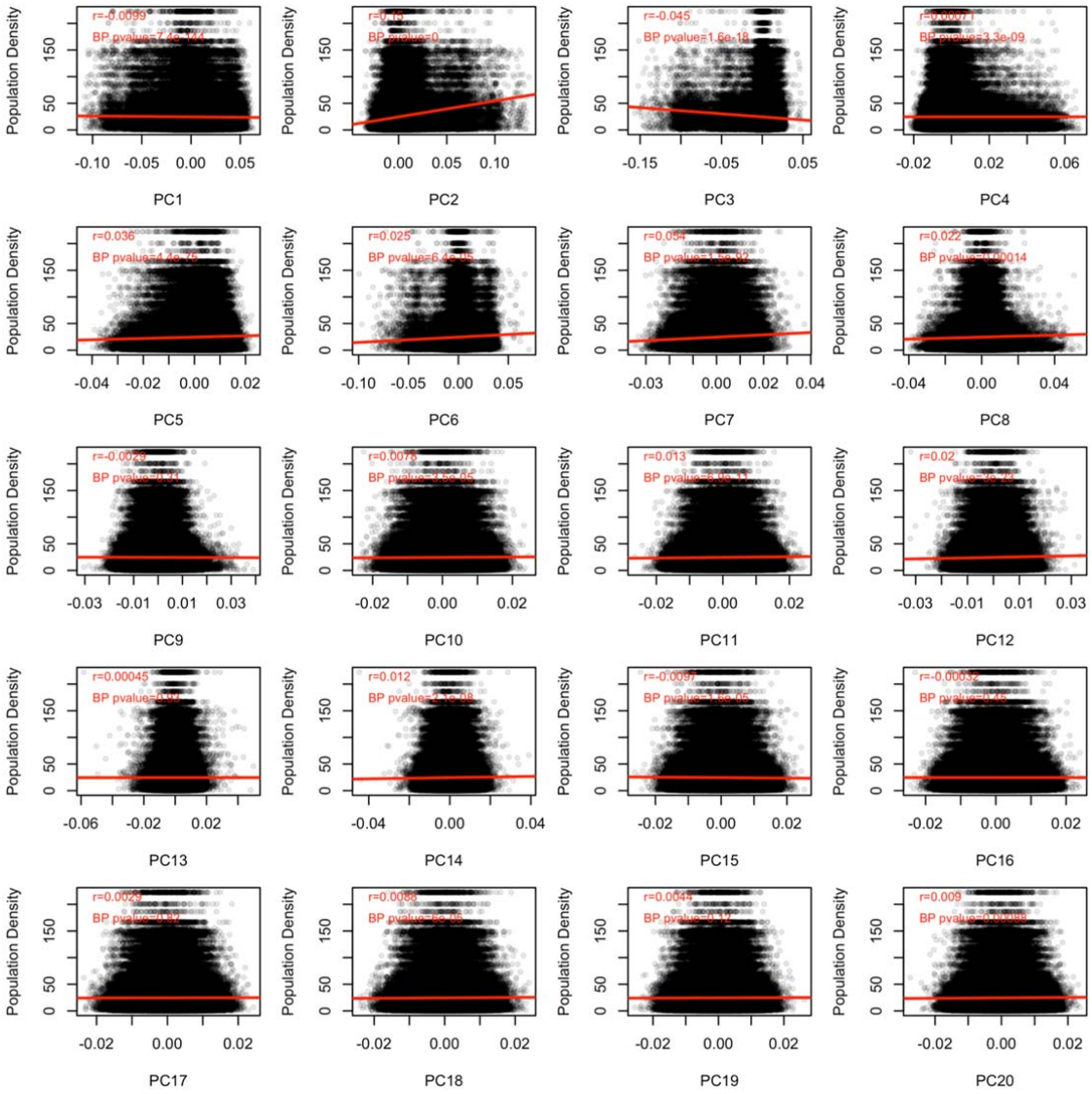
In the second sensitivity analysis, we excluded 68,028 participants with PC values greater than 0.02 or smaller than -0.02. The association between PRS and population density was still observed when fitting 20 genetic PCs (PRS "p<1": r=0.0081, r^2=0.0065%, p-value=1.7e^{-4}, BP pvalue=0.0032; PRS "p<0.05": r=0.0090, r^2=0.0082, p-value=3.4e^{-5}, BP pvalue=0.012). Likewise, when fitting 40 PCs (PRS "p<1": r=0.0083, r^2=0.0068, pvalue=1.3e^{-4}, BP pvalue=0.11; PRS "p<0.05": r=0.0093, r^2=0.0087%, pvalue=1.9e^{-5}, BPpvalue=0.096).

Our sensitivity analyses suggest that the association between schizophrenia PRS and population density is observed when accounting for population ancestry using (a) genetic PCs, (b) using a genetically homogeneous subset of the population or (c) using a GRM to model the (familial and cryptic) relatedness in the sample. However, unaccounted ancestry may explain the increased prediction of the PRS "p<1" compared to the PRS "p<0.05" that was suggested to be the best at predicting schizophrenia[50]. Indeed, the choice of best pvalue cut-off depends on the "training" GWAS sample size as well as the proportion of causal SNPs for the disorder[51]. Adding random signal from non-associated genomic regions may increase the score association with genetic ancestry (though it was not observed in the UKB **eFigures 17 and 19**).



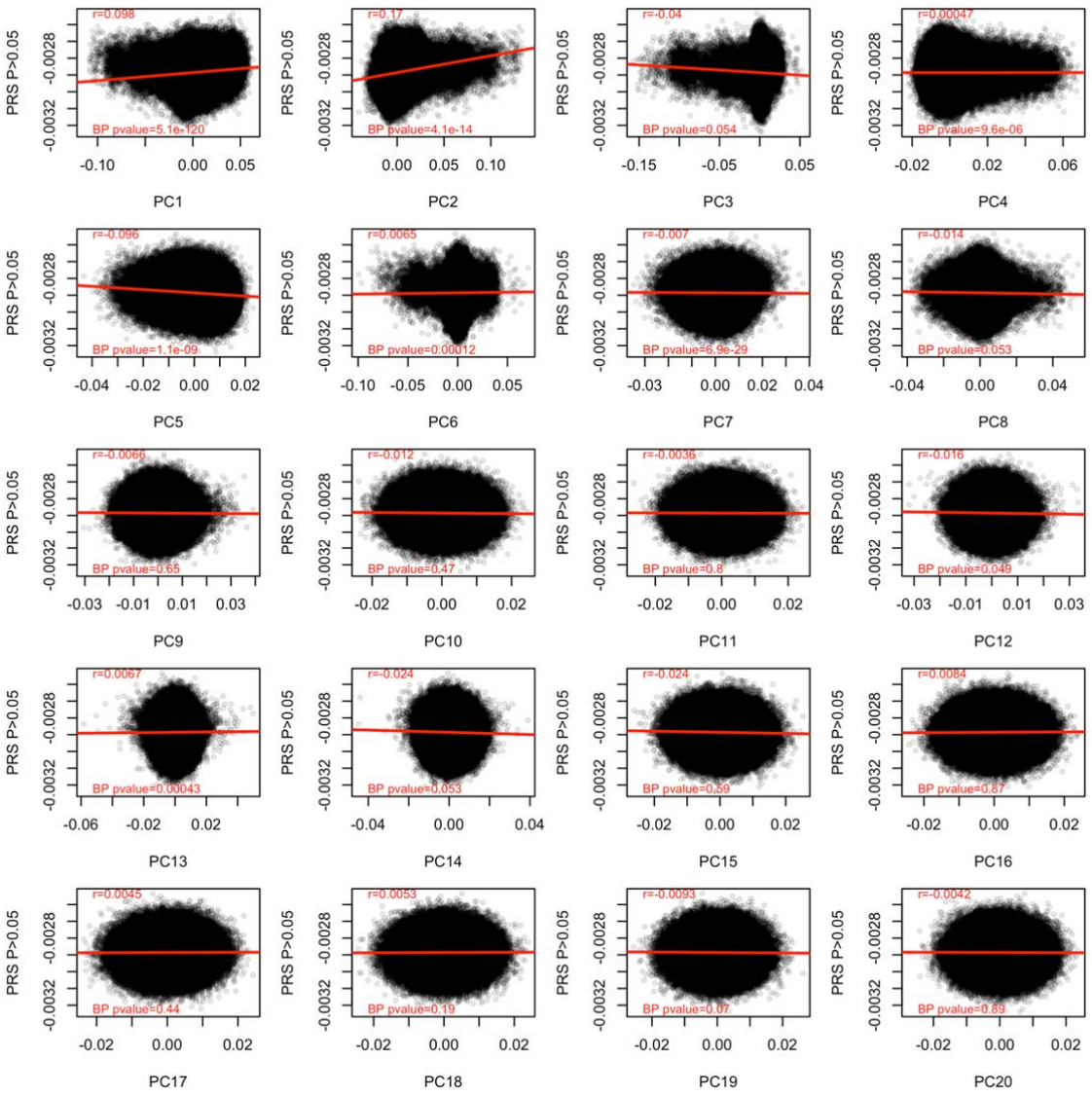
eFigure 17. Relationship Between PRS "p<1" for Schizophrenia and the First 20 Genetic PCs

Correlations are reported at the top of each plot (r=) as well as the p-value of the test of heteroscedasticity (bottom of each plot). We used the studentised Breusch-Pagan test whose null hypothesis is that the residuals of the regression of PRS on a PC have a constant variance (homoscedasticity).



eFigure 18. Relationship Between Population Density and the First 20 Genetic PCs

Correlations are reported at the top of each plot (r=) as well as the p-value of the test of heteroscedasticity (below, NP pvalue=). We used the studentised Breusch-Pagan test whose null hypothesis is that the residuals of the regression of PRS on a PC have a constant variance (homoscedasticity).



eFigure 19. Relationship Between PRS "p<0.05" for Schizophrenia and the First 20 Genetic PCs

Correlations are reported at the top of each plot (r=) as well as the p-value of the test of heteroscedasticity (bottom of each plot). We used the studentised Breusch-Pagan test whose null hypothesis is that the residuals of the regression of PRS on a PC have a constant variance (homoscedasticity).

eAppendix 16: Sample Overlap Between Schizophrenia GWAS and UKB

Sample overlap, even limited to a few percentage of the sample may inflate the estimates from PRS and MR analyses. QIMR, NTR and QSKIN we can be confident that the overlap is very small to none, as these samples did not contribute to (or were excluded from) the GWAS meta-analysis conducted by the PGC in 2014[50]. For the UKB on the other hand we do not know how many participants have participated in both studies. Note that this concern also applies to close relatives of participants as they are genetically similar.

We used the GWAS summary statistics from the PGC and from the UKB analysis on population density to estimate the number of participants present in both samples. More precisely, we used the intercept from the bivariate LD score regression analysis[52] presented in **eAppendix 13**.

As per the LDscore regression paper the intercept of such LD score analysis is a function of, the GWAS sample sizes (N1 and N2), the phenotypic correlation between the traits (r) and the sample overlap(Ns):

InterceptBivarLDscore=r*Ns/(sqrt(N1*N2))

Thus, Ns= InterceptBivarLDscore*sqrt(N1*N2)/r

We do not know the phenotypic correlation between schizophrenia (liability) and population density but we used a conservative, realistic estimate of half the genetic correlation estimated via LDscore regression hence: r=0.228/2=0.114. The LDscore intercept was $2.8354e^{-5}$, yielding an estimated sample overlap of 64 individuals. This represents 0.01% of the UKB full sample of 448,679, which should have a minimal impact on the PRS and MR results.

eReferences

1. Vassos E, Agerbo E, Mors O, Pedersen CB. Urban-rural differences in incidence rates of psychiatric disorders in Denmark. *The British journal of psychiatry : the journal of mental science.* 2016;208(5):435-440.

2. Peen J, Schoevers RA, Beekman AT, Dekker J. The current status of urban-rural differences in psychiatric disorders. *Acta Psychiatr Scand.* 2010;121(2):84-93.

3. Faris REL, Dunham HW. *Mental Disorders in Urban Areas.* Chicago, Ill: University of Chicago Press; 1939.

4. Krabbendam L, van Os J. Schizophrenia and urbanicity: a major environmental influence--conditional on genetic risk. *Schizophr Bull.* 2005;31(4):795-799.

5. Sariaslan A, Fazel S, D'Onofrio BM, et al. Schizophrenia and subsequent neighborhood deprivation: revisiting the social drift hypothesis using population, twin and molecular genetic data. *Transl Psychiat.* 2016;6.

6. Bhavsar V, Boydell J, Murray R, Power P. Identifying aspects of neighbourhood deprivation associated with increased incidence of schizophrenia. *Schizophr Res.* 2014;156(1):115-121.

7. Kirkbride JB, Hameed Y, Ankireddypalli G, et al. The Epidemiology of First-Episode Psychosis in Early Intervention in Psychosis Services: Findings From the Social Epidemiology of Psychoses in East Anglia [SEPEA] Study. *The American journal of psychiatry.* 2017;174(2):143-153.

8. Sariaslan A, Fazel S, D'Onofrio BM, et al. Schizophrenia and subsequent neighborhood deprivation: revisiting the social drift hypothesis using population, twin and molecular genetic data. *Transl Psychiatry.* 2016;6:e796.

9. March D, Hatch SL, Morgan C, et al. Psychosis and place. *Epidemiologic reviews.* 2008;30:84-100.

10. McGrath J, Saha S, Welham J, El Saadi O, MacCauley C, Chant D. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC medicine.* 2004;2:13.

11. Heath AC, Bucholz KK, Madden PA, et al. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. *Psychol Med.* 1997;27(6):1381-1396.

12. Knopik VS, Heath AC, Madden PA, et al. Genetic effects on alcohol dependence risk: re-evaluating the importance of psychiatric and other heritable risk factors. *Psychol Med.* 2004;34(8):1519-1530.

13. Statistics ABo. *Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas [ABS Catalogue No. 1270.0.55.001].* 2011.

14. Statistics ABo. *Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness Structure.* 2011.

15. (ABS) ABoS. *Measures of Socioeconomic Status [ABS Catalogue no. 1244.0.55.001].* 2011.

16. Cuellar-Partida G, Springelkamp H, Lucas SE, et al. WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Human molecular genetics.* 2015;24(17):5060-5068.

17. Medland SE, Nyholt DR, Painter JN, et al. Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am J Hum Genet.* 2009;85(5):750-755.

18. McEvoy BP, Montgomery GW, McRae AF, et al. Geographical structure and differential natural selection among North European populations. *Genome Res.* 2009;19(5):804-814.

19. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81(3):559-575.

20. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.

21. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* 2015;526(7571):75-81.

22. Whitcher **B**, Schmid V, Thornton A. Working with the DICOM and NIfTI Data Standards in R. *Journal of Statistical Software.* 2011;44(6):1-28. .

23. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2012;9(2):179-181.

24. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 2013;10(1):5-6.

25. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics.* 2012;44(8):955-+.

26. Kahle D, Wickham H. ggmap: Spatial Visualization with ggplot2. *R Journal.* 2013;5(1):144-161.

27. Abraham G, Inouye M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *Plos One.* 2014;9(4).

28. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.

29. Willemsen G, Vink JM, Abdellaoui A, et al. The Adult Netherlands Twin Register: Twenty-Five Years of Survey and Biological Data Collection. *Twin Research and Human Genetics.* 2013;16(1):271-281.

30. Vilhjalmsson BJ, Yang J, Finucane HK, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet.* 2015;97(4):576-592.

31. Olsen CM, Green AC, Neale RE, et al. Cohort profile: the QSkin Sun and Health Study. *Int J Epidemiol.* 2012;41(4):929-929i.

32. Grasby KL, Verweij KJH, Mosing MA, Zietsch BP, Medland SE. Estimating Heritability from Twin Studies. *Methods Mol Biol.* 2017;1666:171-194.

33. van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. *Nat Rev Genet.* 2012;13(9):640-653.

34. Martin NG, Eaves LJ, Heath AC. Prospects for detecting genotype X environment interactions in twins with breast cancer. *Acta Genet Med Gemellol (Roma).* 1987;36(1):5-20.

35. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. *Curr Opin Genet Dev.* 2008;18(3):257-263.

36. Wray NR, Lee SH, Mehta D, Vinkhuyzen AAE, Dudbridge F, Middeldorp CM. Research Review: Polygenic methods and their application to psychiatric traits. *J Child Psychol Psyc.* 2014;55(10):1068-1087.

37. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.

38. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet.* 2012;131(5):747-756.

39. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity.* 2005;95(3):221-227.

40. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet.* 2004;74(4):765-769.

41. Keller MC. Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry.* 2014;75(1):18-24.

42. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol.* 2015;44(2):512-525.

43. Hemani G, Zheng J, Wade KH, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv.* 2016.

44. Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *bioRxiv.* 2017.

45. Boker S, Neale M, Maes H, et al. OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika.* 2011;76(23258944):306-317.

46. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* 2010;26(17):2190-2191.

47. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421-427.

48. Curtis D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *bioRxiv.* 2018.

49. Haworth S, Mitchell R, Corbin L, et al. Common genetic variants and health outcomes appear geographically structured in the UK Biobank sample: Old concerns returning and their implications. *bioRxiv.* 2018.

50. Ripke S, Neale BM, Corvin A, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421-+.

51. International Schizophrenia C, Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009;460(7256):748-752.

52. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47(3):291-295.