nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0059-2

Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection

Antonio F. Pardiñas[1], Peter Holmans[1], Andrew J. Pocklington[1], Valentina Escott-Price[1], Stephan Ripke[2,3], Noa Carrera[1], Sophie E. Legge[1], Sophie Bishop[1], Darren Cameron[1], Marian L. Hamshere[1], Jun Han[1], Leon Hubbard[1], Amy Lynham[1], Kiran Mantripragada[1], Elliott Rees[1], James H. MacCabe[4], Steven A. McCarroll[5], Bernhard T. Baune[6], Gerome Breen[7,8], Enda M. Byrne[9,10], Udo Dannlowski[11], Thalia C. Eley[7], Caroline Hayward[12], Nicholas G. Martin[13,14], Andrew M. McIntosh[15,16], Robert Plomin[7], David J. Porteous[12], Naomi R. Wray[9,10], Armando Caballero[17], Daniel H. Geschwind[18], Laura M. Huckins[19], Douglas M. Ruderfer[19], Enrique Santiago[20], Pamela Sklar[19], Eli A. Stahl[19], Hyejung Won[18], Esben Agerbo[21,22], Thomas D. Als[21,23,24], Ole A. Andreassen[25,26], Marie Bækvad-Hansen[21,27], Preben Bo Mortensen[21,22,23], Carsten Bøcker Pedersen[21,22], Anders D. Børglum[21,23,24], Jonas Bybjerg-Grauholm[21,27], Srdjan Djurovic[28,29], Naser Durmishi[30], Marianne Giørtz Pedersen[21,22], Vera Golimbet[31], Jakob Grove[21,23,24,32], David M. Hougaard[21,27], Manuel Mattheisen[21,23,24], Espen Molden[33], Ole Mors[21,34], Merete Nordentoft[21,35], Milica Pejovic-Milovancevic[36], Engilbert Sigurdsson[37], Teimuraz Silagadze[38], Christine Søholm Hansen[21,27], Kari Stefansson[39], Hreinn Stefansson[39], Stacy Steinberg[39], Sarah Tosato[40], Thomas Werge[21,41,42], GERAD1 Consortium[43], CRESTAR Consortium[43], David A. Collier[7,44], Dan Rujescu[45,46], George Kirov[1], Michael J. Owen[1*], Michael C. O'Donovan[1*] and James T. R. Walters[1*]

[1]MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [3]Department of Psychiatry and Psychotherapy, Charité, Campus Mitte, Berlin, Germany. [4]Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [6]Discipline of Psychiatry, University of Adelaide, Adelaide, South Australia, Australia. [7]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [8]NIHR Biomedical Research Centre for Mental Health, Maudsley Hospital and Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [9]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [10]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. [11]Department of Psychiatry and Psychotherapy, University of Münster, Münster, Germany. [12]Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [13]School of Psychology, University of Queensland, Brisbane, Queensland, Australia. [14]QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [15]Division of Psychiatry, University of Edinburgh, Edinburgh, UK. [16]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK. [17]Departamento de Bioquímica, Genética e Inmunología. Facultad de Biología, Universidad de Vigo, Vigo, Spain. [18]Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [19]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [20]Departamento de Biología Funcional. Facultad de Biología, Universidad de Oviedo, Oviedo, Spain. [21]iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus, Denmark. [22]National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark. [23]iSEQ, Center for Integrative Sequencing, Aarhus University, Aarhus, Denmark. [24]Department of Biomedicine–Human Genetics, Aarhus University, Aarhus, Denmark. [25]Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [26]NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway. [27]Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark. [28]NORMENT, KG Jebsen Centre for Psychosis Research, Department of Clinical Science, University of Bergen, Bergen, Norway. [29]Department of Medical Genetics, Oslo University Hospital, Oslo, Norway. [30]Department of Child and Adolescent Psychiatry, University Clinic of Psychiatry, Skopje, Macedonia. [31]Department of Clinical Genetics, Mental Health Research Center, Moscow, Russia. [32]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [33]Center for Psychopharmacology, Diakonhjemmet Hospital, Oslo, Norway. [34]Psychosis Research Unit, Aarhus University Hospital, Risskov, Denmark. [35]Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, Copenhagen, Denmark. [36]Department of Psychiatry, School of Medicine, University of Belgrade, Belgrade, Serbia. [37]Department of Psychiatry, National University Hospital, Reykjavik, Iceland. [38]Department of Psychiatry and Drug Addiction, Tbilisi State Medical University (TSMU), Tbilisi, Georgia. [39]deCODE Genetics, Reykjavik, Iceland. [40]Section of Psychiatry, Department of Public Health and Community Medicine, University of Verona, Verona, Italy. [41]Institute of Biological Psychiatry, MHC Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark. [42]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [43]A list of members and affiliations appears at the end of the paper. [44]Discovery Neuroscience Research, Eli Lilly and Company, Lilly Research Laboratories, Windlesham, UK. [45]Department of Psychiatry, University of Halle, Halle, Germany. [46]Department of Psychiatry, University of Munich, Munich, Germany. A list of members and affiliations appears at the end of the paper.*e-mail: owenmj@cardiff.ac.uk; odonovanmc@cardiff.ac.uk; waltersjt@cardiff.ac.uk

SUPPLEMENTARY NOTE

Case sample collection

We collected blood samples from those with treatment-resistant schizophrenia (TRS) in the UK through the mandatory clozapine blood-monitoring system for those taking clozapine, an antipsychotic licensed for TRS. Following national research ethics approval and in line with UK Human Tissue Act regulations we worked in partnership with the commercial companies that manufacture and monitor clozapine in the UK. We ascertained anonymous aliquots of the blood samples collected as part of the regular blood monitoring that takes place whilst taking clozapine due to a rare haematological adverse effect, agranulocytosis. The CLOZUK1 sample was assembled in collaboration with Novartis (Basel, Switzerland). The company, through their proprietary Clozaril® Patient Monitoring Service (CPMS), provided whole-blood samples and anonymised phenotypic information for 6,882 individuals with TRS (5528 cases post-QC), which were included in the in a recent schizophrenia GWAS by the PGC[1]. The CLOZUK2 sample, previously unreported, was assembled in collaboration with the other major company involved in the supply and monitoring of clozapine in the UK, Leyden Delta (Nijmegen, Netherlands). The company, through their proprietary Zaponex® Treatment Access System (ZTAS), provided whole-blood samples and anonymised phenotypic information for 7,417 of those taking clozapine (4973 cases post-QC). Both Clozaril® and Zaponex® are bioequivalent brands of clozapine licensed in the UK[2].

We restricted the CLOZUK1 and CLOZUK2 samples to those with a clinician reported diagnosis of treatment-resistant schizophrenia. The UK National Institute for Health and Care Excellence (NICE) advise prescription of clozapine is reserved for those with schizophrenia in whom two trials of antipsychotics have failed (including one second-generation antipsychotic)[3] which mirrors the criteria for licensed use of clozapine. The sole alternative licensed indication for clozapine in the UK is for the management of resistant psychosis in Parkinson's disease (PD)[4] and, although this is a rare indication, we excluded PD patients (n=8) from the case dataset. We also excluded those with off-license indications, which included those with alternative clinician diagnoses of bipolar affective disorder and personality disorders (n=56). Together with the clinical guidelines outlined, these exclusions ensure that CLOZUK1 and CLOZUK2 samples are from those patients that conform to a clinical description of TRS. We have reported the use of CLOZUK1 as a schizophrenia dataset in previous publications[1,5-7] and have presented evidence to support the use of TRS-defined individuals as valid schizophrenia samples[8], which we have updated and expanded in the next section**,** including validation of a clinician diagnosis of TRS against research diagnostic criteria for schizophrenia.

In addition we also included in our analysis a more conventional cohort of UK-based patients with schizophrenia (CardiffCOGS). Recruitment was via secondary care, mainly outpatient, NHS mental health services in Wales and England. These patients were not exclusively taking clozapine at the time of their recruitment. All cases

underwent a SCAN interview[9] and case note review followed by consensus research diagnostic procedures and were included if they had a DSM-IV schizophrenia or schizoaffective disorder-depressive type diagnosis, as previously reported[5]. The CardiffCOGS samples were recruited and genotyped in two waves: CardiffCOGS1 (512 cases, included in a previous GWAS[1]) and CardiffCOGS2 (247 cases).

Genotyping for these case samples was performed by the Broad Institute (Massachusetts, USA) for the CLOZUK1 sample and CardiffCOGS1 cases, using Illumina HumanOmniExpress-12 and OmniExpressExome-8 chips as described elsewhere[5]. The CardiffCOGS2 cases and the CLOZUK2 sample were genotyped by deCODE Genetics (Reykjavík, Iceland), using Illumina HumanOmniExpress-12 chips.

As all of these samples are intrinsically related and their recruitment and genotyping conforms to research and technical standards, thus we have combined them and used the term "CLOZUK" throughout this manuscript to describe the schizophrenia case dataset.

Case sample validation

Validation of Clinical Diagnosis

In order to validate the clinical diagnosis of treatment-resistant schizophrenia in the CLOZUK sample we used the CardiffCOGS participants for whom we acquired both clinical and consensus research diagnosis. Prior to the research interview we obtained clinicians' diagnoses for all participants. From participants on clozapine we selected those with a clinical diagnosis of schizophrenia and confirmed that this matched the diagnosis provided when the participant was started on clozapine (i.e. treatment-resistant schizophrenia) so as to be equivalent to the samples included in CLOZUK. We then compared this diagnosis with the consensus research DSM-IV diagnosis arrived after following a SCAN interview, note review and diagnostic procedures described above. 214 participants within the CardiffCOGS sample were taking clozapine and had a clinician-assigned diagnosis of treatment resistant schizophrenia. Following consensus research diagnosis, 194 of these participants were identified as having DSM-IV schizophrenia or schizoaffective disorder depressed sub-type, giving a positive predictive value (PPV) of 90.7%. Many international groups and consortia also consider other diagnoses as 'schizophrenia' samples, namely schizoaffective disorder bipolar type, delusional disorder and schizophreniform disorders[1]. If we expand our analysis to include these categories then 210 of 214 (PPV=98.1%) of those on clozapine with a clinical diagnosis of schizophrenia would receive a DSM-IV research diagnosis of one of these schizophrenia spectrum disorders. These results are entirely consistent with equivalent reports of the validity of clinician diagnoses in two Scandinavian studies[10,11].

Genetic Molecular validation of CLOZUK as a schizophrenia dataset

The Schizophrenia Working Group of the Psychiatric Genomics Consortium identified 40 target subgroups within their primary GWAS analysis and performed a leave-one-out analysis[1]. Using risk alleles identified in the remainder of the primary sample, polygenic risk profile scores were calculated for all individuals in the target subgroup; and the ability of these scores to distinguish between cases and controls was then evaluated. The predictive value of the risk profile score when applied to CLOZUK1 was indistinguishable from its performance in other schizophrenia subgroups, indeed the values for Nagelkerke's pseudo-R^2 for CLOZUK are the 5th highest of all subsamples, implying that CLOZUK is one of the samples most highly enriched for schizophrenia risk alleles (see data for 'noclo_clo' in Extended data Figure 6b[1]). In terms of CNVs, the rates of individual confirmed schizophrenia loci in CLOZUK1 are entirely consistent with those of the other schizophrenia studies[12]. As for CLOZUK2, sign test and polygenic score analyses, as described in the Methods section of the manuscript (**Online Methods**, section *"Estimation and assessment of a polygenic signal"*), confirm its similarity to the PGC samples in respect of schizophrenia-related genetic architecture.

Control sample collection

Control samples were collected from publicly available sources (EGA) or through collaboration with the holders of the datasets. Individual datasets were curated using the same procedures as the case-only datasets. In order to maximize the numbers of individuals that could be effectively included in the GWAS without introducing confounders, these datasets were chosen on the basis of having recruited individuals with self-reported UK ancestry (either exclusively or primarily) and having been genotyped on Illumina chips. A summarized view of all the datasets included in the GWAS is provided later in this document**,** which includes further details of the control datasets.

Genotype quality-control (QC)

Given the many data sources used and the variety of genotyping chips available, a stringent quality control (allowing only 2% of missing SNP and individual data) was performed separately in each individual dataset, using PLINK v1.9[13] and following standard procedures[14]. To facilitate merging and to avoid common sources of batch effects[15], all SNPs in each dataset were also aligned to the plus strand of the human genome (build 37p13), removing strand-ambiguous markers in the process. As most control datasets lacked any markers in the Y chromosome or in the mitochondrial DNA, every SNP from these regions was discarded in the combined genotype data. The final merge of all case and control datasets left 203,436 overlapping autosomal

SNPs. For the X-chromosome, we obtained data for all the cases and 13,085 (out of 24,542) controls, which provided 4,612 overlapping SNPs.

All individuals were imputed simultaneously in the Cardiff University high-performance computing cluster RAVEN[16], using the SHAPEIT/IMPUTE2 algorithms[17,18]. As reference panels, a combination of the 1000 Genomes phase 3 (1KGPp3) and UK10K datasets was used, as this has previously been shown to increase the accuracy of imputation for individuals of British ancestry, particularly for rare variants[19].

After imputation, a principal component analysis (PCA) of common variants (MAF higher than 5%) was carried out to obtain a general summary of the population structure of the sample, using the EIGENSOFT v6 toolset[20]. A plot of the first two PCs showed the existence of a large fraction of cases (~20%) with no overlapping controls (**Supplementary Figure 1, A**). A comparison with the 1KGPp3 dataset, performed using PCA and ADMIXTURE[21] estimates, showed that most of these cases were similar in genetic ancestry to non-European individuals, namely from the East Asian or West African superpopulations (**Supplementary Figure 1, B**). In order to use only cases with matching control samples and to ameliorate population stratification in the association analysis[22], all individuals not falling into an area delimited by the mean and 3 standard deviations of the two first principal components of the control samples were excluded from further analyses (**Supplementary Figure 1, C**). By repeating PCA only on the selected individuals, no outliers could be detected in the first two principal components, and ADMIXTURE plots were homogenised as well (**Supplementary Figure 2**).

The CLOZUK sample was further pruned by removing all individuals with inbreeding coefficients (F) higher than 0.2, and leaving only a random member of each pair with a relatedness coefficient ($\hat{\pi}$) higher than 0.2. Furthermore, to ensure the independence of our analyses with previous GWAS conducted by the Schizophrenia Working Group of the PGC, relatedness coefficients of CLOZUK individuals were also calculated with all the individual datasets included in the latest PGC GWAS[1] following approval by the Consortium. Detected genetic relatives (or duplicates) were excluded in CLOZUK in the same way as intra-population relatives. After this process, we excluded 3,103 individuals as PCA-based ancestry outliers, 5 individuals due to heterozygosity and 985 individuals due to relatedness. Finally, 35,802 samples (11,260 cases and 24,542 controls) with 9.66 million imputed markers (INFO>0.3; MAF>0.001; HWE $p > 1x10^{-6}$) remained in the CLOZUK dataset.

Estimating the proportion of true positives in PGC GWAS loci

We used the uniformly minimum variance conditionally unbiased estimator (UMVCUE) of Bowden & Dudbridge [23] to estimate true effect sizes for the genome-

wide significant autosomal index SNPs from the previous GWAS of schizophrenia carried out by the PGC [1]. This method combines replication data (here, the deCODE samples reported in the original study) with the discovery data to minimise upward biases in effect size due to "winner's curse" (i.e. selecting SNPs with $p<5x10^{-8}$). We then estimated the probability that each SNP would be genome-wide significant in the combined CLOZUK+PGC meta-analysis, assuming that the effect size of the SNP in the CLOZUK sample was that estimated by the UMVCUE (i.e. a true positive). We did likewise assuming no effect in the CLOZUK sample (i.e. a false positive), and used these probabilities to estimate the proportion of true positive SNPs, along with a 95% confidence interval.

Of the 108 SNPs reported in the original study, 7 were not available in our meta-analysis, having been excluded in the QC pipeline carried out in the CLOZUK GWAS. Of the 101 remaining SNPs, 18 were not genome-wide significant in the combined CLOZUK+PGC analysis. (Note that this number is slightly higher than that in **Supplementary Table 5** since the latter uses the most significant SNP in the region, which may be different to the original lead SNP). Assuming that all 101 represent true signals, we expect 80 to remain GWS in our meta-analysis following the Bowden and Dudbridge approach (using the formula given in section 9.ii of the procedure described below, setting p=1). We actually observe 83, consistent with all the PGC signals being true positives, with a 95% CI of (0.8,1) –see section 9.iv of the procedure.

Detailed procedure

This was done using a similar pipeline to Hamshere et al. [24]:

1. Use the UMVCUE method to obtain estimates of effect sizes (log odds ratios) and variances for each of the index SNPs from the PGC study using both the discovery (GWAS) and replication (deCODE) samples.
2. Use these to simulate a "true" effect size in the CLOZUK sample: given UMVCUE effect size μ and its variance σ^2, generate a random effect size (β) by sampling from a normal(μ, σ^2). Convert this into an odds ratio $OR=e^{\beta}$.
3. Use this "true" effect size to simulate a log-OR (+variance) in CLOZUK by using its sample size and MAF. Let the minor (reference) allele be *A* and the other allele be *a*. The frequency of the minor allele is *p* and the odds ratio associated with the minor allele is *B.* There are *N* controls and *M* cases in CLOZUK. Consequently the observed frequency *q* of *A* alleles in controls is approximately distributed as a normal(*p*, *p*(1-*p*)/2*N*). Sample *q* from this distribution and calculate (*N*1, *N*2), the corresponding number of *A* and *a* alleles in the controls (=2*Nq*, 2*N*(1-*q*) respectively).
4. The frequency *r* of *A* alleles in cases is given by *r*=*pB*/(1+*pB*-*p*). Observed frequency *s* of *A* in cases is then approximately distributed as a normal(*r*,*r*(1-*r*)/*2M*). Sample *s* and calculate corresponding numbers of *A* and *a* alleles in cases (*M1*, *M2*) = 2*Ms*, 2*M*(1-*s*).

5. Finally, use *N1*, *N2*, *M1*, *M2* to calculate the observed effect size β_S=ln(*M1*N2*/*M2*N1*) and its variance σ_S^2=(1/*N1*)+(1/*N2*)+(1/*M1*)+(1/*M2*)
6. Meta-analyse the simulated CLOZUK log-OR and variance generated in the previous step with the actual log-OR and variance from the PGC GWAS using a fixed effects inverse-variance meta-analysis.
7. Repeat 2)-6) 10,000 times to estimate the probability that the CLOZUK+PGC meta-analysis is genome-wide significant assuming UMVCUE "true" effects (i.e. the PGC GWAS result was a true positive)
8. Repeat 3)-6) 10,000 times to estimate the probability that the CLOZUK+PGC meta-analysis is GWS assuming no effect in CLOZUK (i.e. the PGC GWAS result was a false-positive)
9. Use the probabilities in 7) and 8), combined with the observed number of SNPs that were GWS in the CLOZUK+PGC meta-analysis to estimate the proportion of true positives:
    i. If Pi = probability that SNP i is genome-wide significant (GWS) in PGC+CLOZUK given that it is a true effect and Qi = probability that SNP i is GWS in PGC+CLOZUK given it is a false positive, and p=proportion of true positives, the overall probability that SNP i is GWS = p.Pi + (1-p).Qi.
    ii. So, the expected total number of GWS SNPs is given by

$$E(p) = \Sigma i\ [(p.Pi + (1-p)Qi]$$

    And its variance by

$$V(p) = \Sigma i\ [(p.Pi + (1-p)Qi)\ .\ (p(1-Pi)+(1-p)(1-Qi))$$

    iii. The maximum likelihood estimator of p is the value of p for which E(p) is equal to the observed number of GWS SNPs, O. If O is larger than E(p=1) then this is set equal to 1.
    iv. The 95% confidence interval for p is the set of values of p for which E(p) is not significantly different from O. That is: $(O-E(p))^2/V(p) < 3.841$

Background selection effects on traits under negative selection

The following theoretical analysis aims to characterise how the action of background selection (BGS) can influence the magnitude and frequency of effects that can be detected by GWA studies of negatively selected complex traits. Assuming that rates and distributions of mutational effects are evenly distributed over the genome, we conclude that genome regions under strong BGS can contribute more to heritability than regions under moderate BGS. This conclusion does not hold for neutral traits, for which the expectation is exactly the opposite.

Detecting genotypic effects in a case-control GWAS design

For the sake of simplicity we initially consider a haploid population of size *n*, in which the focus is on a pair of SNPs: One of them ($x = [x_1, x_2,..x_n]$) is neutral and the other one ($y = [y_1, y_2,\ldots,y_n]$) has an effect on a quantitative trait. Let x_j be the dosage (0 vs. 1) of the reference allele of the neutral SNP on individual *j* and y_j the dosage of the risk allele for the causal SNP in the same individual. The risk allele has an effect on a quantitative trait that correlates negatively with fitness. Let p_j be the phenotypic value of individual *j* and, also, let $\alpha \cdot y_j$ be the genotypic value contributed by the causal SNP, where α is the average effect of the allele substitution. Finally, let r^2 be the squared correlation of the allele dosages of both SNPs[25]:

$r^2 = \frac{cov^2(x,y)}{\sigma_x^2 \cdot \sigma_y^2}$, where σ_x^2 and σ_y^2 are variances of x and y, respectively.

The expected χ^2 association test value between the neutral SNP and the phenotypic value in a sample of 2*n* haploid individuals is

$$E[\chi^2] = 2n \cdot E\left[\frac{C^2(x,p)}{V_x \cdot V_p}\right],$$

where $C(x,p)$, V_x and V_p are the covariances and variances of *x* and *p* observed in the sample (dosage of the neutral SNP allele and phenotypic value of the individual, respectively). This expectation can be given in terms of the true variances and covariances of the population, where $\sigma_{Ay}^2 = \sigma_y^2 \cdot \alpha^2 = q(1-q)\alpha^2$ is the genetic variance contributed by the causal SNP with risk-allele frequency *q*[26]. Using the approximation $\sqrt{(1-r^2)/2n}$ for the standard deviation of the correlation coefficient [27], the well-known set of equations for the expected χ^2 in GWA studies are obtained:

$$E[\chi^2] = 2n \cdot \left[\frac{cov^2(x,p)}{\sigma_x^2 \cdot \sigma_p^2} \cdot \left(1 - \frac{1}{2n}\right) + \frac{1}{2n}\right] = \frac{cov^2(x,\alpha y)}{\sigma_x^2 \cdot \sigma_p^2} \cdot (2n-1) + 1 =$$

$$\frac{\alpha^2 \cdot cov^2(x,y)}{\sigma_x^2 \cdot \sigma_p^2} \cdot (2n-1) + 1 = \frac{\alpha^2 \cdot cov^2(x,y)}{\sigma_x^2 \cdot \sigma_p^2} \cdot \frac{\sigma_{Ay}^2}{\sigma_{Ay}^2} \cdot (2n-1) + 1 \; =$$

$$\frac{\alpha^2 \cdot cov^2(x,y)}{\sigma_x^2 \cdot \sigma_{Ay}^2} \cdot \frac{\sigma_{Ay}^2}{\sigma_p^2} \cdot (2n-1) + 1 = \frac{\alpha^2 \cdot cov^2(x,y)}{\sigma_x^2 \cdot \alpha^2 \cdot \sigma_y^2} \cdot h_y^2 \cdot (2n-1) + 1.$$

$$E[\chi^2] = r^2 \cdot h_y^2 \cdot (2n-1) + 1 = r^2 \cdot \frac{q(1-q)\alpha^2}{\sigma_p^2} \cdot (2n-1) + 1.$$

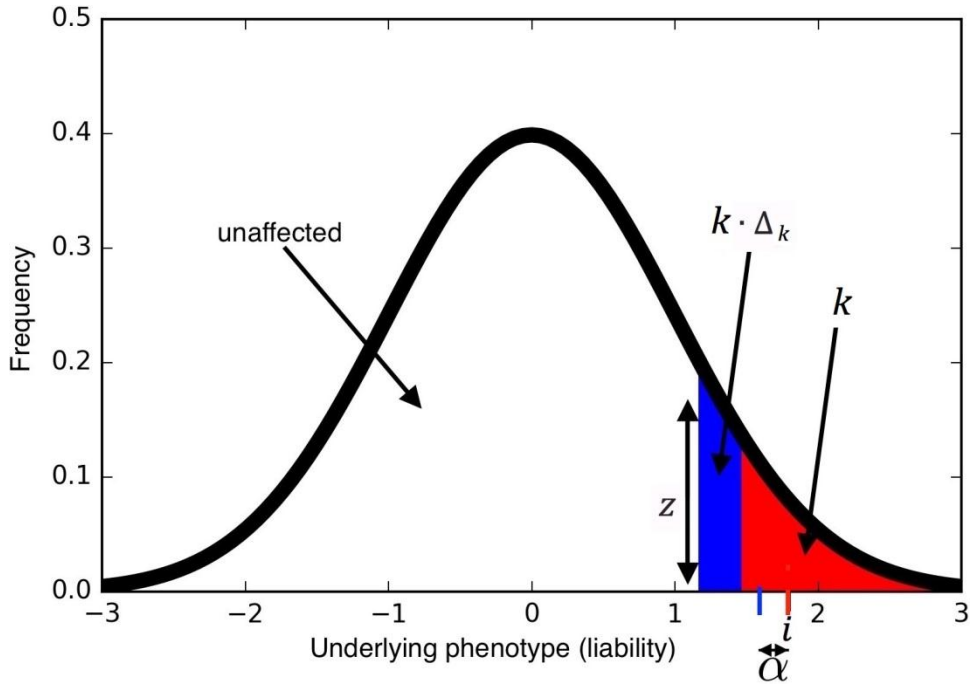
In this set of equations, h_y^2 is the true heritability attributable to the causal SNP and $r^2 \cdot h_y^2 = h_x^2$ is the heritability that is explained by the neutral SNP. These heritabilities are defined here for haploid genomes. However, considering a diploid organism, the corresponding heritability is slightly smaller than twice the haploid heritability (by a factor 1/2*n*) because haploid effects are negatively correlated within diploids due to sampling:

$$E[\chi^2] = r^2 \cdot \frac{q(1-q)\alpha^2}{\sigma_p^2} \cdot (2n-2) + 1 = r^2 \cdot \frac{2q(1-q)\alpha^2}{\sigma_p^2} \cdot (n-1) + 1.$$

Now we consider specifically schizophrenia, which is traditionally analysed as a case-control trait. For such traits, the underlying phenotype is the susceptibility to the disorder (liability), which can be assumed to be normally distributed with a variance $\sigma_p^2 = 1$. Assuming that the population prevalence of schizophrenia is *k* = 0.007 [28], a causal SNP for which the heterozygote increases susceptibility in Δ_k = 1%[1] has an effect on population prevalence of:

$$k + k \cdot \Delta_k \approx k + \alpha \cdot z = k + \alpha \cdot i \cdot k.$$

Here, *p* is the phenotypic value, z is the density of the normal distribution at the liability threshold and i is the mean phenotypic liability of the affected group. These variables are illustrated in a standard liability threshold model below[29]:



From the previous equations, the substitution effect α measured on the liability scale is:

[1] The equation $\mathrm{OR} = (1+\Delta_k)(1-k)/(1-k(1+\Delta_k))$ can be used to transform susceptibility increases into odds-ratios. In this case, Δ_k =1% is equivalent to a marker OR = 1.011 for schizophrenia.

$$\alpha = \frac{\Delta_k}{i} = \frac{0.01}{2.784} = 0.0036 \text{ liability units.}$$

In the present manuscript we describe a meta-analysis of schizophrenia GWAS data of *n* = 105,318 selected individuals: 40,675 cases and 64,643 controls. In this sample, the observable prevalence of schizophrenia is increased 55 times with respect to the population prevalence (from *k* = 0.007 to *k*' = 0.386). This causes that, for a SNP of Δ_k = 1% (as in the previous example), the effect in the sample must be computed from a prevalence of 38.6% (i' = 0.991):

$$\alpha' = \frac{\Delta_k}{i'} = \frac{0.01}{0.991} = 0.0101 \text{ liability units.}$$

According to the sample characteristics, the χ^2 expected value is

$$E[\chi^2] = r'^2 \cdot 2q'(1-q') \cdot \alpha'^2 \cdot (n-1) + 1 = r'^2 \cdot 2q'(1-q') \cdot \alpha'^2 \cdot 105317 \ + 1.$$

The term q' stands for the frequency of the risk allele of the causal SNP in the sample, and it is expected to be close to its frequency in the population unless the effect Δ_k is very large, so $q' \approx q[1 + \Delta_k(k' - k)]$. Consequently, variances, covariances and, particularly, correlations of allele dosages in the sample are not expected to be very different from the corresponding values in the population.

A rough approximation can be made for the statistical power of the present experiment. The critical value for a χ^2 -distribution with 1df for a typical genome-wide significance threshold of 5 x 10^{-8} is 29.72. Using this threshold, the non-centrality parameter (NCP) of a χ^2 distribution which gives a 5% probability of detection is 14.50. So, for the aforementioned SNP to be detected with that probability in this GWA study, the following condition must be met:

$$14.50 < [r'^2 \cdot q'(1-q') \cdot \alpha'^2 \cdot 2 \cdot 105317 \ + 1], \text{ therefore}$$

$$r'^2 \cdot q'(1-q') \cdot \alpha'^2 > 6.409 \cdot 10^{-5}.$$

Given that the maximum values of r'^2 and $q'(1-q')$ are 1 and 0.25 respectively, the minimal α' effect needed for detection is $\alpha' = 0.016$, which corresponds to $\Delta_k = 0.016 \cdot 0.991 \approx 0.016$ and a substitution effect in the population $\alpha = \Delta_k / i = 0.016/2.784 = 0.006$. This is equivalent to a theoretical SNP with OR = 1.017 and MAF = 50%. SNPs with smaller allele frequencies should have larger phenotypic effects to have minimal chances of being detected at this significance threshold. Note that the NCP of a GWAS marker can also be related to other parameters such as the disease risk model and the sample case/control ratio[30], but for simplicity these have been omitted from our calculations.

Detecting associations with causal SNPs in schizophrenia

Schizophrenia has been shown to cause a reduction in fertility rate of $Rf = 0.65$[31]; average in both genders. Assuming a linear effect model on fitness[32], a detectable SNP with an effect α of 0.016 liability units in the meta-analysis ($\Delta_k = 0.01$ in the population) has a selection coefficient of:

$$s = \Delta_k \cdot k \cdot Rf = 0.01 \cdot 0.007 \cdot 0.65 = 0.00004.$$

It is known that mutations with deleterious effects larger than $s = 1/2N_e$, where N_e is the effective population size, are under active negative selection. In this model, it is expected that the larger the effect of the allele on the trait is, the lower the range of the randomly fluctuating frequency will be[33]. This establishes a negative correlation between N_e and the frequency of causal variants. Therefore, it would be unlikely to find large-effect alleles at common frequencies in large populations, as has been consistently shown in human complex traits and psychiatric disorders in particular[34].

Genetic drift also affects the correlation between SNPs, which reflects their linkage disequilibrium. The expected r^2 values of a new mutation with other SNPs are initially of the order of $1/2N_e$. This value is not reduced every generation by recombination as might be intuitively thought, but it is increased by drift up to a maximum value[35]:

$r^2_{max} = \frac{1}{1+4N_e c}$, where c is the recombination rate.

Summing up, genetic drift affects the probability of detection of causal effects on negatively selected traits in two ways. Firstly, it increases the expected frequencies of deleterious mutations. Secondly, it increases the expected correlation between pairs of loci. As indicated above, the product of both terms q and r^2 are included in the equation for the expected χ^2 in GWA studies. Note that, within some parameter ranges, the detection of causal effects for negatively selected traits could be increased in populations with small N_e, as these might harbour alleles of larger effect at higher frequencies and stronger LD. This effect is not expected for neutral traits because the amount of neutral variation, including variation for neutral traits, is proportional to N_e, which largely compensates the contrary but relatively small effect of N_e on r^2.

The same rationale can be extended to the effect of background selection (BGS) on detection of causal effects at different genome regions. The large-scale differences in the amount of neutral variation at genome regions are well explained by the BGS model: Selection on deleterious variants reduces variation at linked neutral sites in a way that is nearly equivalent to a reduction in N_e[36,37], with all of its associated effects. Genome regions with reduced N_e would have increased contributions to variation which can be effectively detected by GWAS of negatively selected traits.

Simulation Study 1

Reductions in N_e allow schizophrenia risk variants to persist at common frequencies and explain more heritability

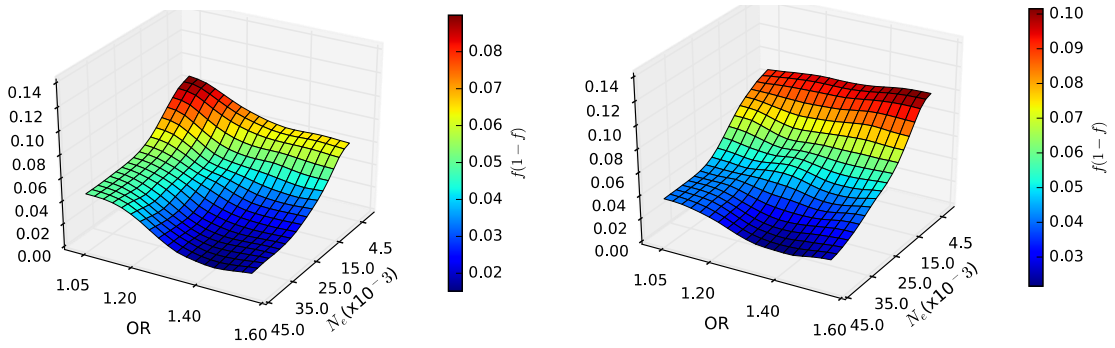
For testing the feasibility of detecting causal alleles in regions under BGS, pairs of causal-neutral biallelic loci were simulated for ten N_e values evenly distributed from 4500 to 45000 diploid individuals. This range of population sizes accounts for the estimations of the effective size of human populations from the out-of-Africa event to current times[38], and also represents a possible range of differences between genome regions. As it will be shown, the general conclusions are not expected to change for combinations of parameters out of this range. For each N_e value, ten effects evenly distributed from α = 0.02 to 0.2 were simulated. Assuming the aforementioned prevalence for schizophrenia of k = 0.007, these α values represent odds ratios from OR = 1.06 to OR = 1.62, which are within the ranges detected by GWAS up to this date. The corresponding selective value s for each α was calculated as indicated above ($s = \Delta_k \cdot k \cdot Rf$) using the reduction in fertility value of Rf = 0.65 and $\Delta_k = \alpha \cdot i$, where $i = 2.784$ as determined by the prevalence. Different recombination rates between the two loci were considered, but since the general trend does not change with differences in recombination, only the average rate c = 0.000225 between neutral SNPs and causal candidates detected in the study was simulated for the whole set of 10 x 10 combinations of N_e and α values. This value of c is equivalent to a linkage between both SNPs of $r^2 \approx 0.1$, which is a commonly used value for LD-based clumping of GWAS results and is assumed to capture the majority of putatively causal SNPs at each locus[39].

Each combination of parameters was run for 10^8 generations. Every time an allele was lost, the same allele was reintroduced in the population as a single copy, but the results were re-scaled proportionally to N_e. Thus, the number of mutation events was proportional to N_e and the rate of mutation was the same for all the different effects. Each generation, frequencies at both loci and correlation between loci were computed. These were used to calculate the expectation $E[\chi^2]$ using a substitution effect α' equivalent to a selected sample with the increased meta-analysis prevalence described in this work (k'=0.386). From that expectation, the probabilities of obtaining values of χ^2 > 29.72 (significant at the level $5 \cdot 10^{-8}$) were in turn computed using the non-central χ^2 distribution for two sample sizes (n = 30,000 and n = 100,000) which are similar to the two GWAS described in the main manuscript. These probabilities were used to calculate two parameters: First, the product $f(1 - f)$, where f is the frequency of a neutral SNP that is significantly associated with the trait (Supplementary Note Figure 1). Second, the product of the sum of probabilities of detection of the effect allele over generations, until the neutral SNP is lost or fixed, multiplied by $f(1 - f)r^2$ (Supplementary Note Figure 2). This product is proportional to h_x^2/α^2, the expected heritability explained by a neutral SNP relative to the squared substitution effect.

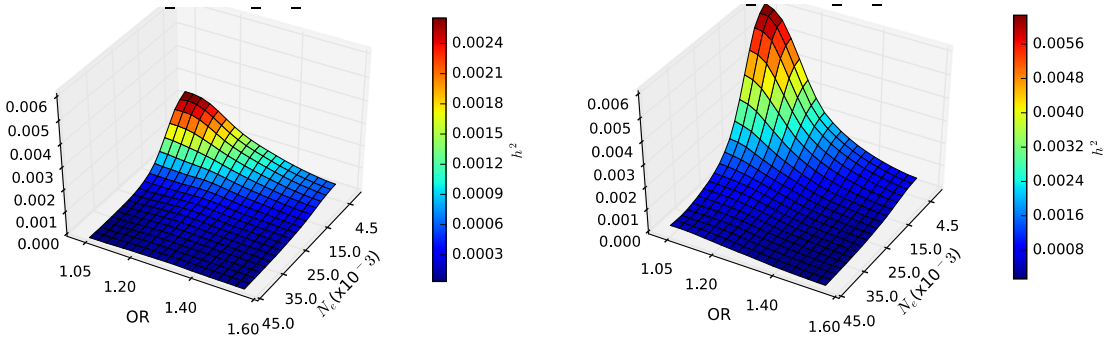
Two main and related conclusions can be obtained from the simulations. First, for any particular effect α (OR in the figures), the frequencies of SNPs significantly associated with causal effects increase as N_e decreases. This is coincident with the observation of the abundance of common SNPs at genome regions under strong

BGS. Secondly, the contribution to heritability increases for decreasing N_e, which supports the mechanism proposed to explain the relationship between genomic regions under BGS and schizophrenia risk loci.

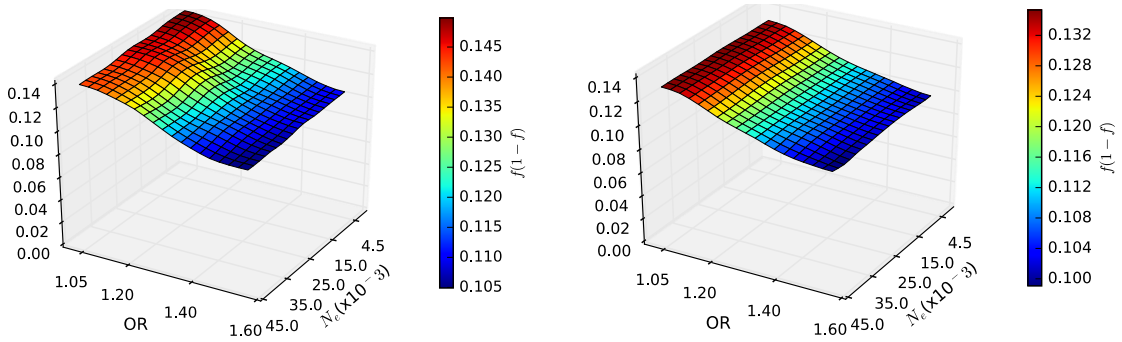
An additional block of simulations of 10 x 10 combinations of N_e and α values was carried out for the same combination of parameters, but this time for a neutral trait (Rf = 0). The results show a completelly different pattern: First, the product $f(1 - f)$ does not change with N_e (Supplementary Note Figure 3). Second, the explained heritability tends to increase as N_e increases (Supplementary Note Figure 4). This allows us to predict that genome regions under strong BGS contribute less to the heritability of neutral traits than weakly selected regions, which is congruent with the expectation of increasing levels of neutral variation as N_e increases.



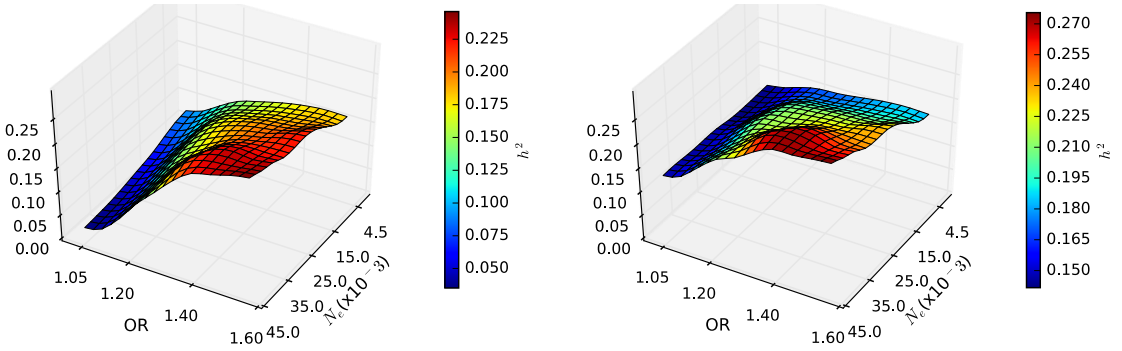
SUPPLEMENTARY NOTE FIGURE 1. Average product $f(1 – f)$ of allele frequencies at neutral SNPs significantly associated with effects on schizophrenia (vertical axis and colour scale). Two sample sizes are shown: n=30,000 (left) and n=100,000 (right). The surface is given as a function of N_e and the odds ratio (OR) of the causal SNP, which is derived from the corresponding effect α in the population (see text).



SUPPLEMENTARY NOTE FIGURE 2. Relative contributions to the heritability of the SNPs significantly associated with effects ($f[1 – f]r^2$) in the schizophrenia simulations. Note that these contributions must be multiplied by α^2 to obtain the absolute contribution to heritability. Two sample sizes are shown: n = 30,000 (left) and n = 100,000 (right). Notice that the scale in which this statistic is reported is arbitrary but both plots have the same scale and can be compared.



SUPPLEMENTARY NOTE FIGURE 3. Average product $f(1 – f)$ of allele frequencies at neutral SNPs significantly associated with effects on a neutral trait (vertical axis and colour scale). Two sample sizes are shown: n = 30,000 (left) and n = 100,000 (right). The surface is given as a function of N_e and the odds ratio (OR) of the causal SNP, which is derived from the corresponding effect α in the population (see text).



SUPPLEMENTARY NOTE FIGURE 4. Relative contributions to the heritability of the SNPs significantly associated with effects ($f[1 – f]r^2$) in the neutral trait simulations. Note that these contributions must be multiplied by α^2 to obtain the absolute contribution to heritability. Two sample sizes are shown: n = 30,000 (left) and n = 100,000 (right). Notice that the scale in which this statistic is reported is arbitrary but both plots have the same scale and can be compared.

Simulation Study 2

Reduction in N_e due to negative selection is caused by BGS and allows for improved detection of risk alleles in GWAS settings

The following study describes computer simulations which illustrate that BGS can also increase the probability of detecting causal SNPs for a quantitative trait in a GWAS setting. It does not intend to be a comprehensive study regarding a range of different scenarios and parameters, but to support the feasibility of the BGS effect in generic negatively-selected traits which might not fit to a liability threshold model.

Forward individual simulations were carried out using the software SLiM[40] for a diploid population of constant size N = 1,000 individuals, run for 100,000 generations. A single genome sequence of 1Mb was assumed where mutations occurred at a rate $\mu = 10^{-7}$ per-nucleotide and generation. The recombination rate between nucleotides was assumed to be $c = 10^{-8}$, constant across the whole sequence, implying an average value of 1cM per 1Mb.

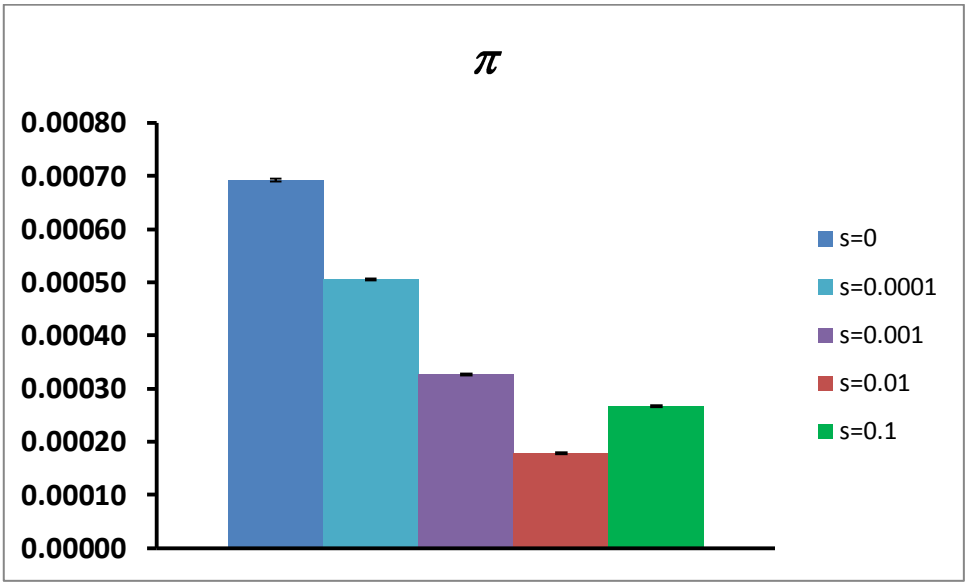
Mutations were assumed to appear at random along the sequence such that 74% of mutations were neutral, 24% were assumed to be deleterious for fitness with a homozygous effect obtained from an exponential distribution with mean s. Five different scenarios were considered using a range of mean values of s (0, 0.0001, 0.001, 0.01 and 0.1) in order to simulate different magnitudes of BGS. The remainder 1% mutations were assumed to be slightly deleterious, with a constant selection coefficient s = 0.001 ($2Ns$ = 2), and to be true quantitative trait loci (QTL) with an effect of one environmental standard deviation. All effects, both for fitness and for the quantitative trait were assumed to be additive. Phenotypes of individuals for the quantitative trait were obtained adding a normal environmental deviation to the genotypic value.

In the last generation a sample of 100 individuals was taken from the population and a GWAS was performed using PLINK, discarding variants with frequency smaller than MAF = 1%. The number of SNPs analysed varied from about 4,000 (s = 0) to about 2,000 (s = 0.1).

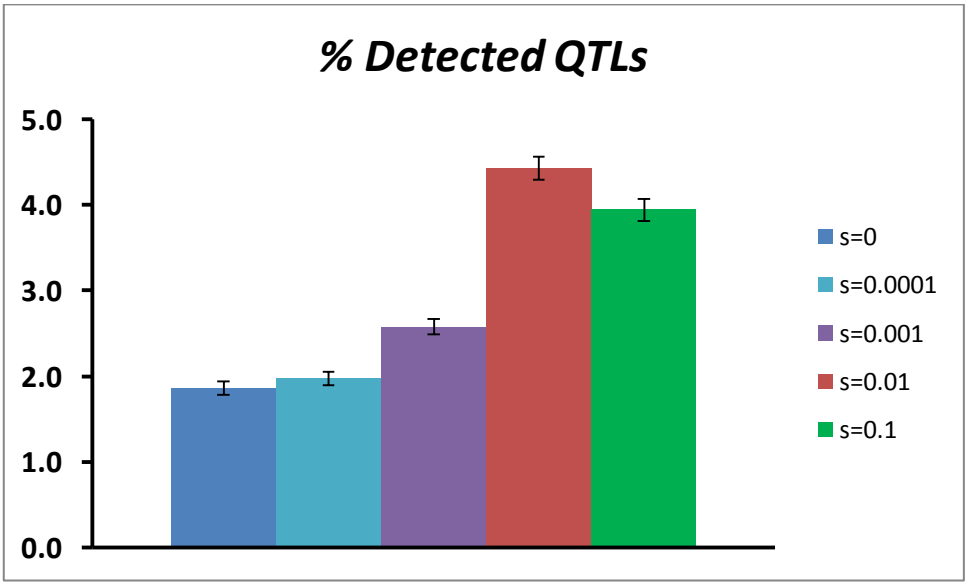
In order to compare the probability of detection of causal SNPs under different levels of BGS, the top twenty SNPs with the lowest probability from the GWAS analysis were considered, and the number of true causal QTLs in these 20 SNPs was recorded. To quantify the magnitude of the genomic reduction in effective population size due to BGS the nucleotide diversity (π) was scored for all neutral SNPs. Each scenario was replicated 1,000 times.

As expected, negative selection was found to result in a reduction in neutral diversity, which is a clear signature of the BGS process and is related to genomic N_e (Supplementary Note Figure 5). Such a reduction was paired with an increased number of detected QTLs (Supplementary Note Figure 6). For strong values of the

selection coefficient (s = 0.1), however, both effects were weaker than for intermediate values (s = 0.01), as mutations of large effect do not result in strong BGS because they persist less time in the population.



SUPPLEMENTARY NOTE FIGURE 5. Average neutral nucleotide diversity for scenarios with different mean selection coefficients of deleterious mutations, where s = 0 implies no BGS. Bars indicate one standard error for the mean across replicates.



SUPPLEMENTARY NOTE FIGURE 6. Percentage of causal variants (QTLs, mutations affecting the quantitative trait) found by GWAS within the 20 top SNPs according to their probability for scenarios with different mean selection coefficients of deleterious mutations, where s = 0 implies no BGS. Bars indicate one standard error for the mean across replicates.

Detail of samples included in the present study

Samples in the CLOZUK study			
DATASET	Samples in GWAS	Genotyping chip	Reference
CLOZUK1	5,528	OmniExpress	24*
CardiffCOGS1	512	OmniExpress	12*
CLOZUK2	4,973	OmniExpress	This study
CardiffCOGS2	247	OmniExpress	This study
WTCCC2	4,641	Illumina 1.2M	41,42*
Cardiff Controls	1,078	OmniExpress	43*
Generation Scotland	6,480	OmniExpress	44
T1DGC	2,532	HumanHap 550	45
POBI	2,516	Illumina 1.2M	46,47
TWINSUK	2,426	Illumina 317/610/660/1M	48
QIMR	2,339	Illumina 317/610/660	49,50
TEDS	1,752	OmniExpress	51
GERAD	778	Illumina 660	52

Samples marked with an asterisk * were included in the PGC schizophrenia study [1]; all other samples have not previously been reported in schizophrenia GWAS

Summarized description of control samples

WTCCC2: Wellcome Trust Case-Control Consortium unscreened controls from the UK Blood Bank and 1958 Birth Cohort (NCDS).

Cardiff Controls: Unscreened blood donor controls recruited in Wales by Cardiff University in collaboration with the NHS Blood and Transplant Authority.

Generation Scotland: Samples from individuals recruited by the Generation Scotland Scottish Family Health Study (GS:SFHS). While in the original design there was no selection on the basis of medical status or history, these controls have been screened for psychiatric disorders using SCID criteria.

T1DGC: Unscreened controls used by the Type 1 Diabetes Genetics Consortium, recruited in the UK through the 1958 Birth Cohort. This recruitment was intended to be independent from WTCCC2, though any sample overlaps were addressed by the GWAS QC pipeline (see **Online Methods**).

POBI: Individuals genotyped for the "People of the British Isles" project, which collected samples from geographically diverse rural communities throughout the UK. The sample is unscreened for psychiatric illness and was recruited from predominantly older age brackets (mode 60-69 years at time of collection).

TWINSUK: This sample consists of individuals recruited through the Twins Health Registry of the Department of Twin Research of King's College London. The samples included in this study were unrelated and screened for self-reported psychiatric disorders. We selected one individual randomly from each twin pair.

QIMR: This sample is a mixture of controls screened for Major Depressive Disorder (MDD) and unscreened controls from an Australian community sample. Unrelated individuals included in this study were ascertained through studies of twin families.

TEDS: Individuals recruited through the Twins Early Development Study. The sample is formed by selected unrelated individuals from the original twin-based design. Though unscreened for psychiatric disorders, these individuals had no severe medical problems nor suffered severe problems peri- or postnatally.

GERAD: This sample was obtained from the Genetic and Environmental Risk for Alzheimer's disease (GERAD) Consortium. All of these controls were elderly and screened for dementia using the MMSE or ADAS-cog assessments.

Replication samples

DATASET	Cases	Controls	Genotyping chip	Reference
deCODE1	681	137,678	HumanHap/OmniExpress	53
deCODE2	885	924	HumanHap	54
iPSYCH	3,226	10,583	HumanCoreExome/PsychChip	56,57
TOP	970	5,039	OmniExpress	-

Summarized description of replication samples

deCODE1: The Icelandic sample consisted of cases and controls who were recruited and diagnosed in Iceland as previously described[53]. Diagnoses were assigned according to Research Diagnostic Criteria (RDC) using the Schedule for Affective Disorders and Schizophrenia Lifetime Version (SADS-L). Controls were recruited as a part of various genetic programs at deCODE and were not screened for psychiatric disorders. The study was approved by the National Bioethics Committee and the Icelandic Data Protection Authority, and all participants provided written, informed consent.

deCODE2: This sample included cases and controls from Italy, Georgia, Macedonia, Russia and Serbia; these individuals were recruited and diagnosed as detailed elsewhere[54]. All studies were approved by local ethics committees, and all participants provided written, informed consent.

iPSYCH: The Danish data consists of two samples (GEMS2 and iPSYCH-SCZ). In both samples cases were identified from the Danish Psychiatric Central Research Register[55], and diagnosed with SCZ by a psychiatrist according to ICD10. Eligible were singletons born to a known mother and resident in Denmark on their one-year birthday. Samples were linked using the unique personal identification number to the Danish Newborn Screening Biobank at Statens Serum Institute, where DNA was extracted from Guthrie cards and whole genome amplified in triplicates as described previously [56,57]. The study was approved by the Danish regional scientific ethics committee and the Danish data protection agency.

TOP: Thematically Organized Psychosis (TOP) Study cases participating in the current study were mainly included from the Therapeutic Drug Monitoring laboratory at Diakonhjemmet Hospital, Oslo. This laboratory is used for monitoring of nearly all schizophrenia patients treated with clozapine and other antipsychotics in the region. We obtained anonymous aliquots of the blood samples collected as part of the regular blood monitoring and DNA was extracted and used in the current study based on approval from the Regional Committee for Medical and Health Research Ethics. The healthy controls were randomly selected from statistical records of persons from the same catchment area as the cases. Participants were between 18-60 years old and healthy based on clinical examination and disease history, and none had any history of severe mental disorders, head injury, neurological disorders, illicit drug use, close relatives with severe mental disorder or medical problems that somehow could interfere with brain function. All participants provided written informed consent and the human subjects protocol was approved by the Regional Committee for Medical and Health Research Ethics and the Norwegian Data Protection Agency. In addition, healthy blood donors from the same region were included in the control sample. They were all thoroughly screened for diseases, and provided blood for DNA analysis, in line with approval from the Regional Committee for Medical and Health Research Ethics.

References

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).

2. Couchman, L., Morgan, P.E., Spencer, E.P., Johnston, A. & Flanagan, R.J. Plasma Clozapine and Norclozapine in Patients Prescribed Different Brands

of Clozapine (Clozaril, Denzapine, and Zaponex). *Therapeutic Drug Monitoring* **32**, 624-627 (2010).

3. National Collaborating Centre for Mental Health. Psychosis and schizophrenia in adults: The NICE guideline on treatment and management. Vol. National Clinical Guideline Number 178 (NICE, London, 2014).

4. Davie, C.A. A review of Parkinson's disease. *British medical bulletin* **86**, 109-127 (2008).

5. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry* **204**, 108-14 (2014).

6. Hamshere, M.L. *et al.* Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Molecular Psychiatry* **18**, 738 (2013).

7. Richards, A.L. *et al.* Exome arrays capture polygenic rare variant contributions to schizophrenia. *Human Molecular Genetics* **25**, 1001-7 (2016).

8. Pocklington, A.J. *et al.* Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia. *Neuron* **86**, 1203-14 (2015).

9. Wing, J.K. *et al.* Scan - Schedules for Clinical-Assessment in Neuropsychiatry. *Archives of General Psychiatry* **47**, 589-593 (1990).

10. Ekholm, B. *et al.* Evaluation of diagnostic procedures in Swedish patients with schizophrenia and related psychoses. *Nordic Journal of Psychiatry* **59**, 457-464 (2005).

11. Jakobsen, K.D. *et al.* Reliability of clinical ICD-10 schizophrenia diagnoses. *Nordic Journal of Psychiatry* **59**, 209-212 (2005).

12. Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol Psychiatry* **19**, 37-40 (2014).

13. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**(2015).

14. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protocols* **5**, 1564-1573 (2010).

15. Zuvich, R.L. *et al.* Pitfalls of merging GWAS data: lessons learned in the eMERGE network and quality control procedures to maintain high data quality. *Genetic Epidemiology* **35**, 887-898 (2011).

16. Advanced Research Computing @ Cardiff (ARCCA). Introduction to RAVEN. (accessed: 29/03/2016).

17. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* **44**, 955-959 (2012).

18. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5-6 (2013).

19. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature communications* **6**(2015).

20. Patterson, N.J., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genetics* **2**, e190 (2006).

21. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655-1664 (2009).

22. Tian, C., Gregersen, P.K. & Seldin, M.F. Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics* **17**, R143-R150 (2008).

23. Bowden, J. & Dudbridge, F. Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. *Genetic epidemiology* **33**, 406-418 (2009).

24. Hamshere, M.L. *et al.* Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Mol Psychiatry* **18**, 708-712 (2013).

25. Hill, W. & Robertson, A. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226-231 (1968).

26. Nielsen, D.M., Ehm, M.G. & Weir, B.S. Detecting Marker-Disease Association by Testing for Hardy-Weinberg Disequilibrium at a Marker Locus. *The American Journal of Human Genetics* **63**, 1531-1540 (1998).

27. Visscher, P.M. & Hill, W.G. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet* **5**, e1000628 (2009).

28. McGrath, J., Saha, S., Chant, D. & Welham, J. Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. *Epidemiologic Reviews* **30**, 67-76 (2008).

29. Dempster, E.R. & Lerner, I.M. Heritability of Threshold Characters. *Genetics* **35**, 212-236 (1950).

30. Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**, 335-346 (2014).

31. Power, R.A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA psychiatry* **70**, 22-30 (2013).

32. Eyre-Walker, A. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences* **107**, 1752-1756 (2010).

33. Ohta, T. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23**, 263-286 (1992).

34. Sullivan, P.F., Daly, M.J. & O'Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet* **13**, 537-551 (2012).

35. Hill, W.G. Linkage disequilibrium among neutral mutant alleles in finite population. *Advances in Applied Probability* **8**, 10-12 (1976).

36. Charlesworth, B., Morgan, M. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303 (1993).

37. Comeron, J.M., Williford, A. & Kliman, R.M. The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity* **100**, 19-31 (2007).

38. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108**, 11983-11988 (2011).

39. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature genetics* **22**, 139-144 (1999).

40. Messer, P.W. SLiM: Simulating Evolution with Selection and Linkage. *Genetics* **194**, 1037-1039 (2013).

41. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678 (2007).

42. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology* **35**, 34-41 (2006).

43. Green, E.K. *et al.* The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Molecular Psychiatry* **15**, 1016-22 (2010).

44. Amador, C. *et al.* Recent genomic heritage in Scotland. *BMC Genomics* **16**, 437 (2015).

45. Hilner, J.E. *et al.* Designing and implementing sample and data collection for an international genetics study: the Type 1 Diabetes Genetics Consortium (T1DGC). *Clinical Trials* **7**, S5-S32 (2010).

46. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309-314 (2015).

47. Winney, B. *et al.* People of the British Isles: preliminary analysis of genotypes and surnames in a UK-control population. *European Journal of Human Genetics* **20**, 203-210 (2012).

48. Moayyeri, A., Hammond, C.J., Hart, D.J. & Spector, T.D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Research and Human Genetics* **16**, 144-149 (2013).

49. Wright, M.J. & Martin, N.G. Brisbane Adolescent Twin Study: Outline of study methods and research projects. *Australian Journal of Psychology* **56**, 65-78 (2004).

50. Wray, N.R. *et al.* Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol Psychiatry* **17**, 36-48 (2012).

51. Haworth, C.M.A., Davis, O.S.P. & Plomin, R. Twins Early Development Study (TEDS): A Genetically Sensitive Investigation of Cognitive and Behavioral Development From Childhood to Young Adulthood. *Twin Research and Human Genetics* **16**, 117-125 (2013).

52. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* **41**, 1088-1093 (2009).

53. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744-747 (2009).

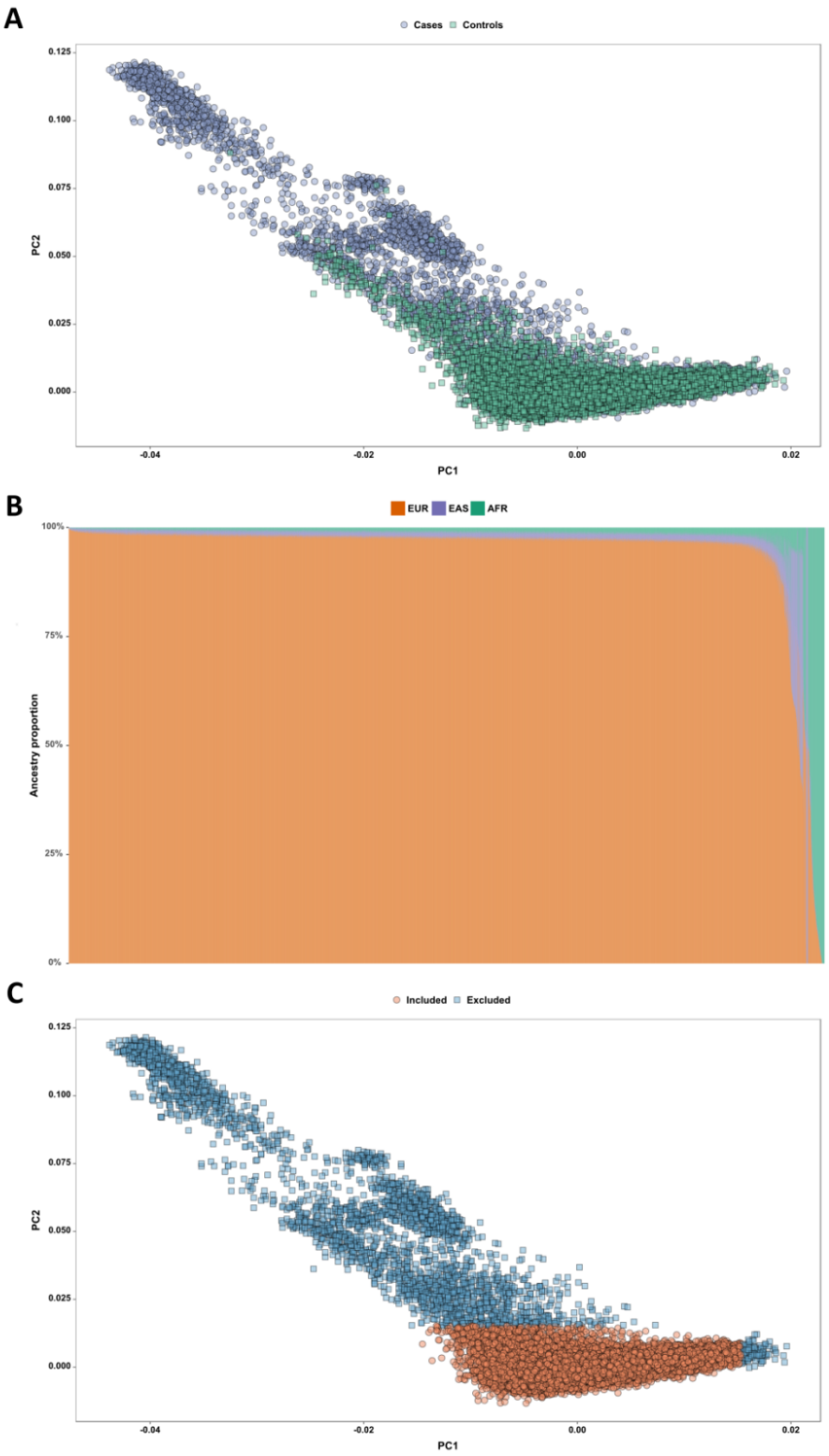
54. Steinberg, S. *et al.* Common variant at 16p11. 2 conferring risk of psychosis. *Molecular psychiatry* **19**, 108-114 (2014).

55. Mors, O., Perto, G.P. & Mortensen, P.B. The Danish Psychiatric Central Research Register. *Scand J Public Health* **39**, 54-7 (2011).

56. Borglum, A.D. *et al.* Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Mol Psychiatry* **19**, 325-33 (2014).

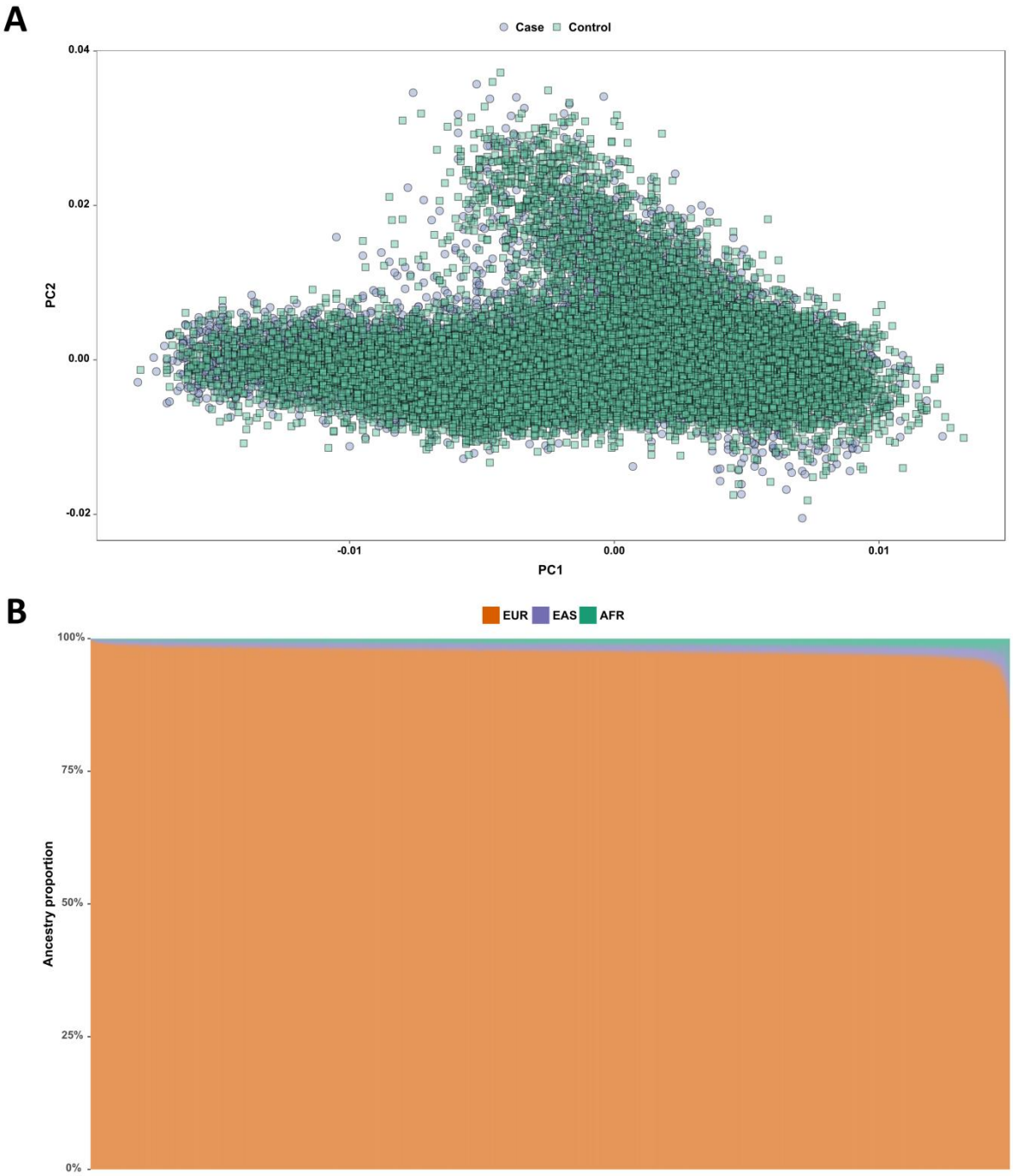
57. Hollegaard, M.V. *et al.* Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source. *BMC Genet* **12**, 58 (2011).

Supplementary Figure 1



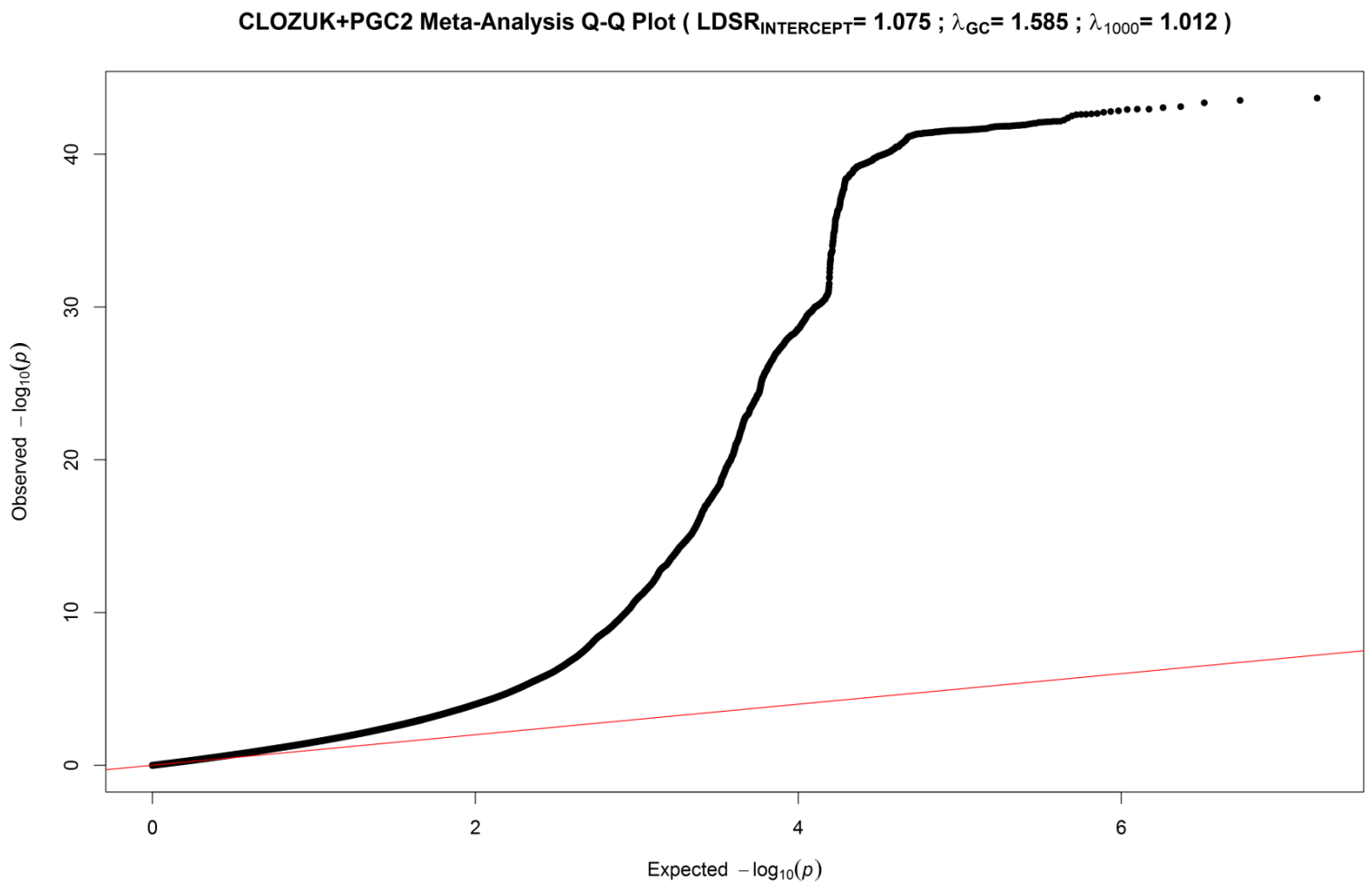
Population structure of the complete CLOZUK dataset. A: PCA showing cases and controls, notice the large spread in the cases. B: ADMIXTURE plot (K=3), names of ancestral components represent the most similar 1KGPp3 superpopulation. C: PCA showing the individuals finally selected for the GWAS.

Supplementary Figure 2



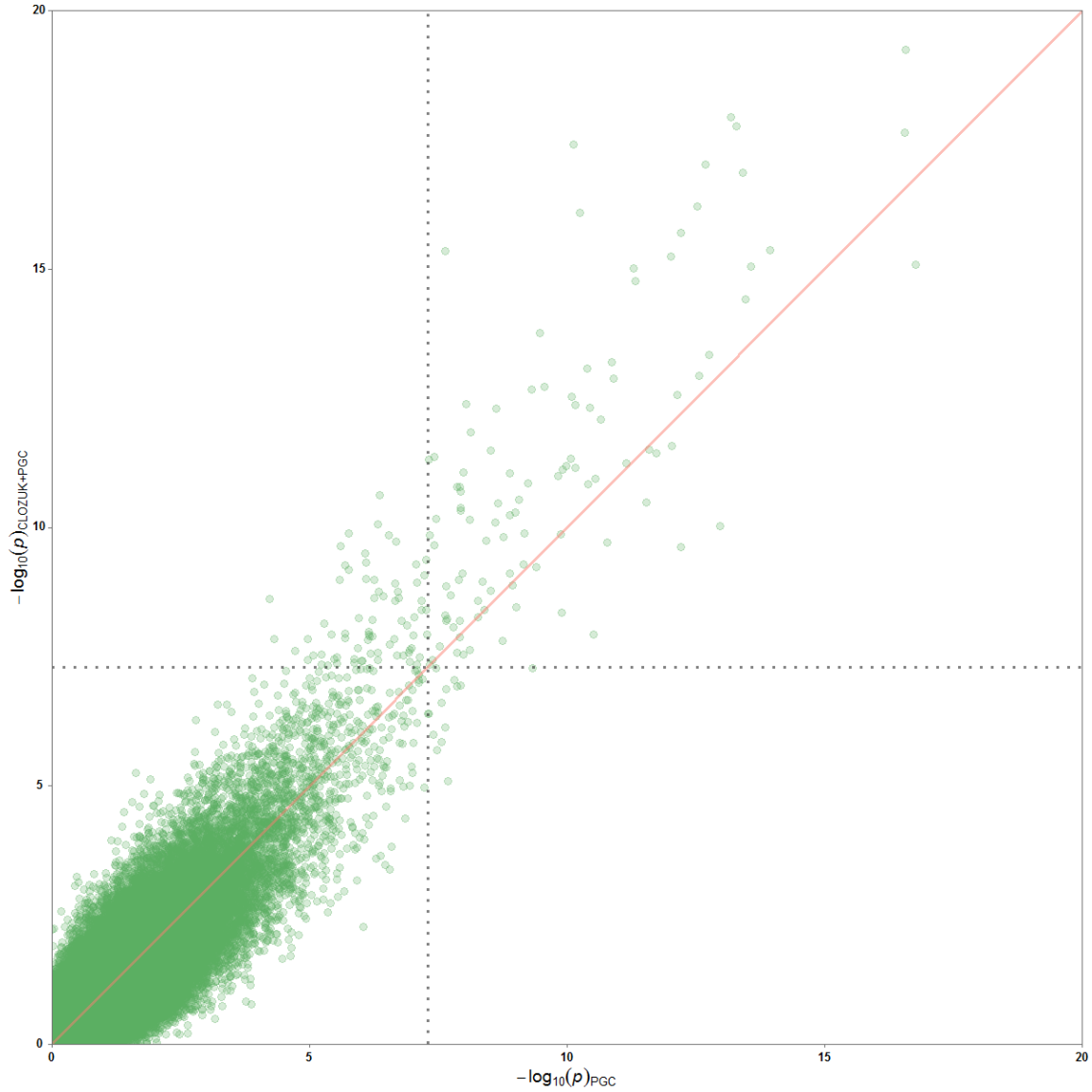
Population structure of the CLOZUK subset selected for the GWAS. A: PCA showing cases and controls, notice the profiles are almost completely overlapping. B: ADMIXTURE plot (K=3), names of ancestral components represent the most similar 1KGPp3 superpopulation.

Supplementary Figure 3



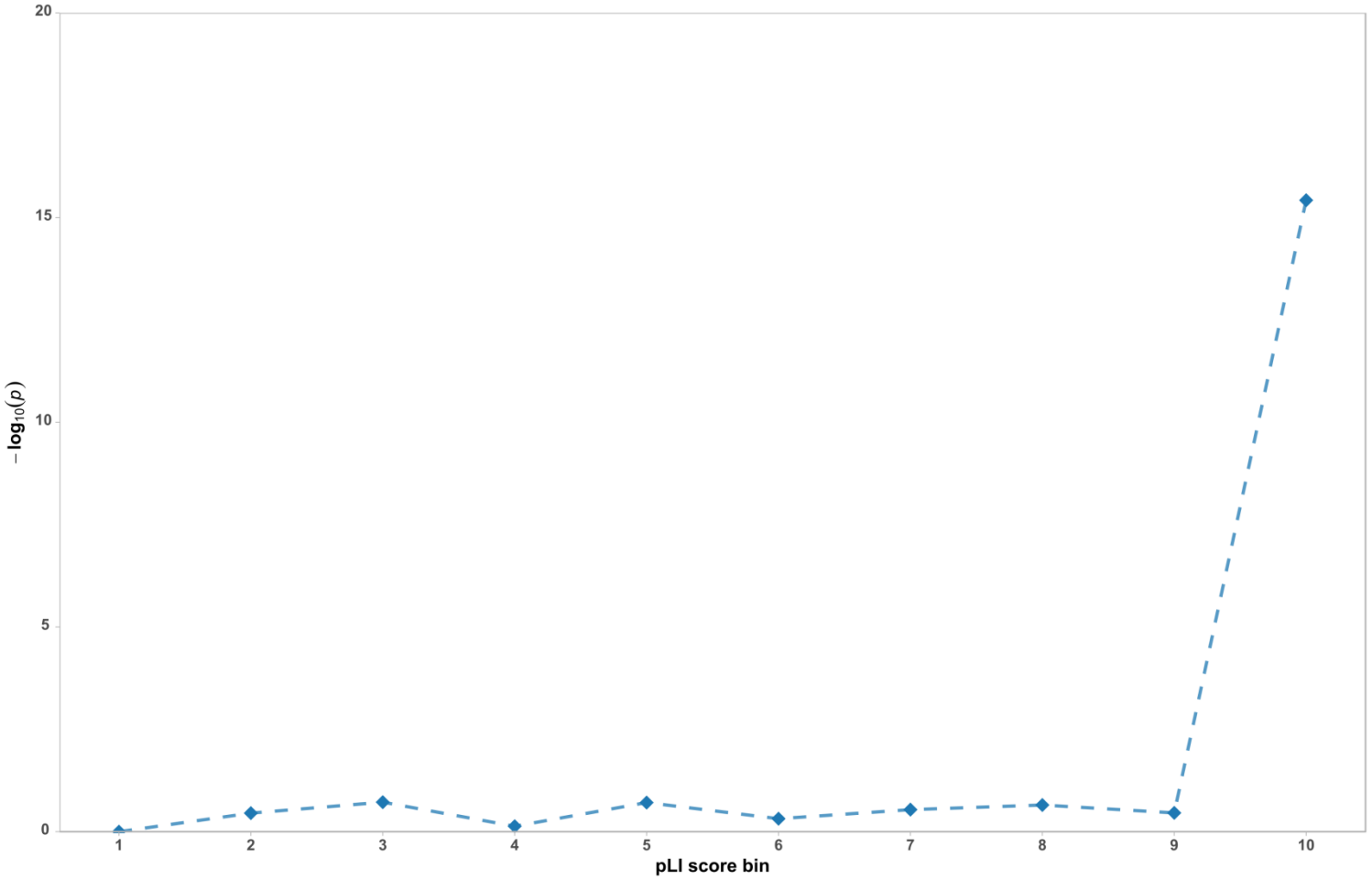
QQ plot of CLOZUK and PGC2 meta-analysis.

Supplementary Figure 4



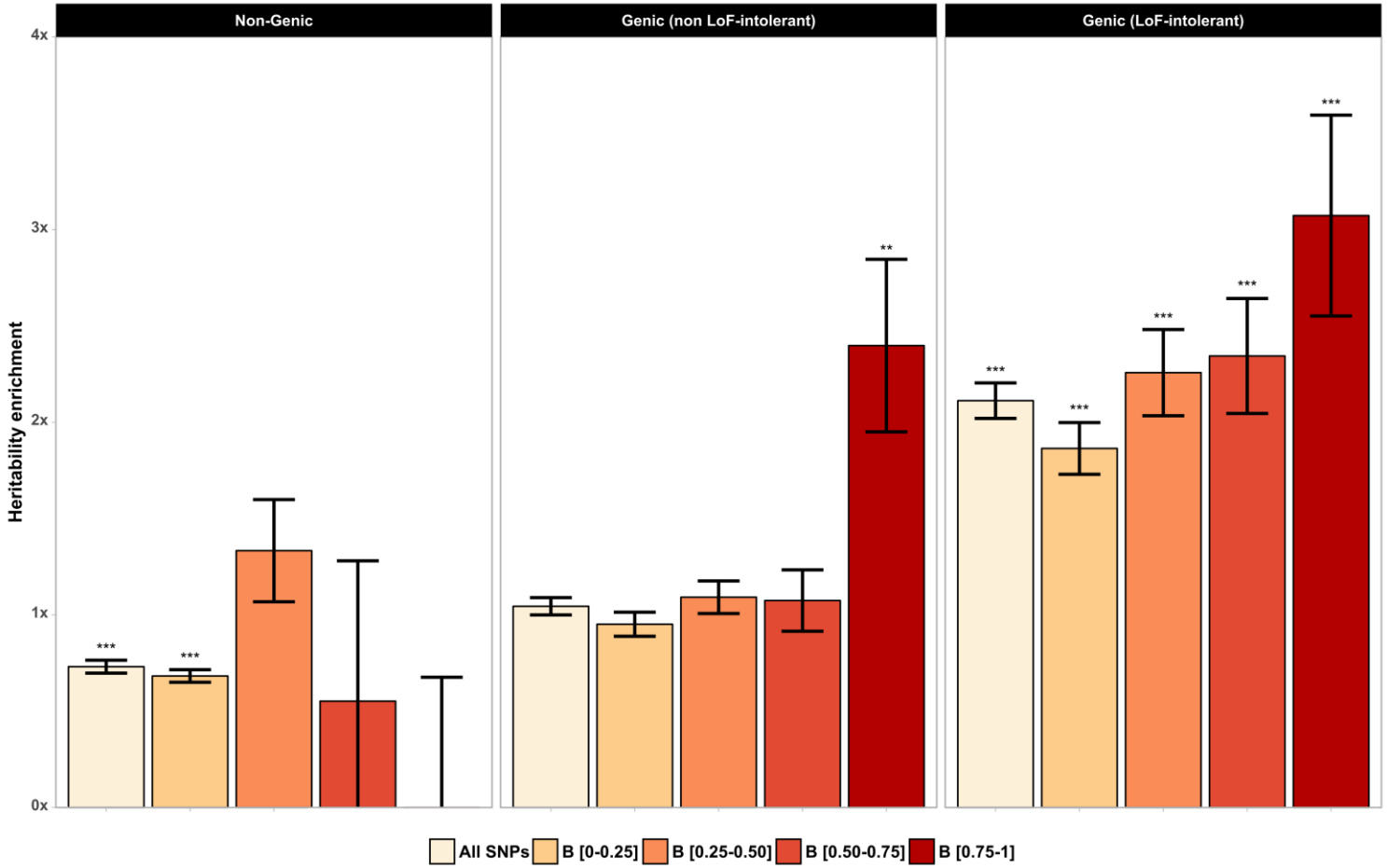
Index SNP p-values for all clumps in the meta-analysis (CLOZUK+PGC) compared with PGC. Dotted lines show the genome-wide significant threshold for the two datasets. The red line indicates a null of equal p-values in both datasets, and thus index SNPs to the left of this line (toward y axis) show increased significance in our meta-analysis. All clumps in the xMHC have been excluded from this plot.

Supplementary Figure 5



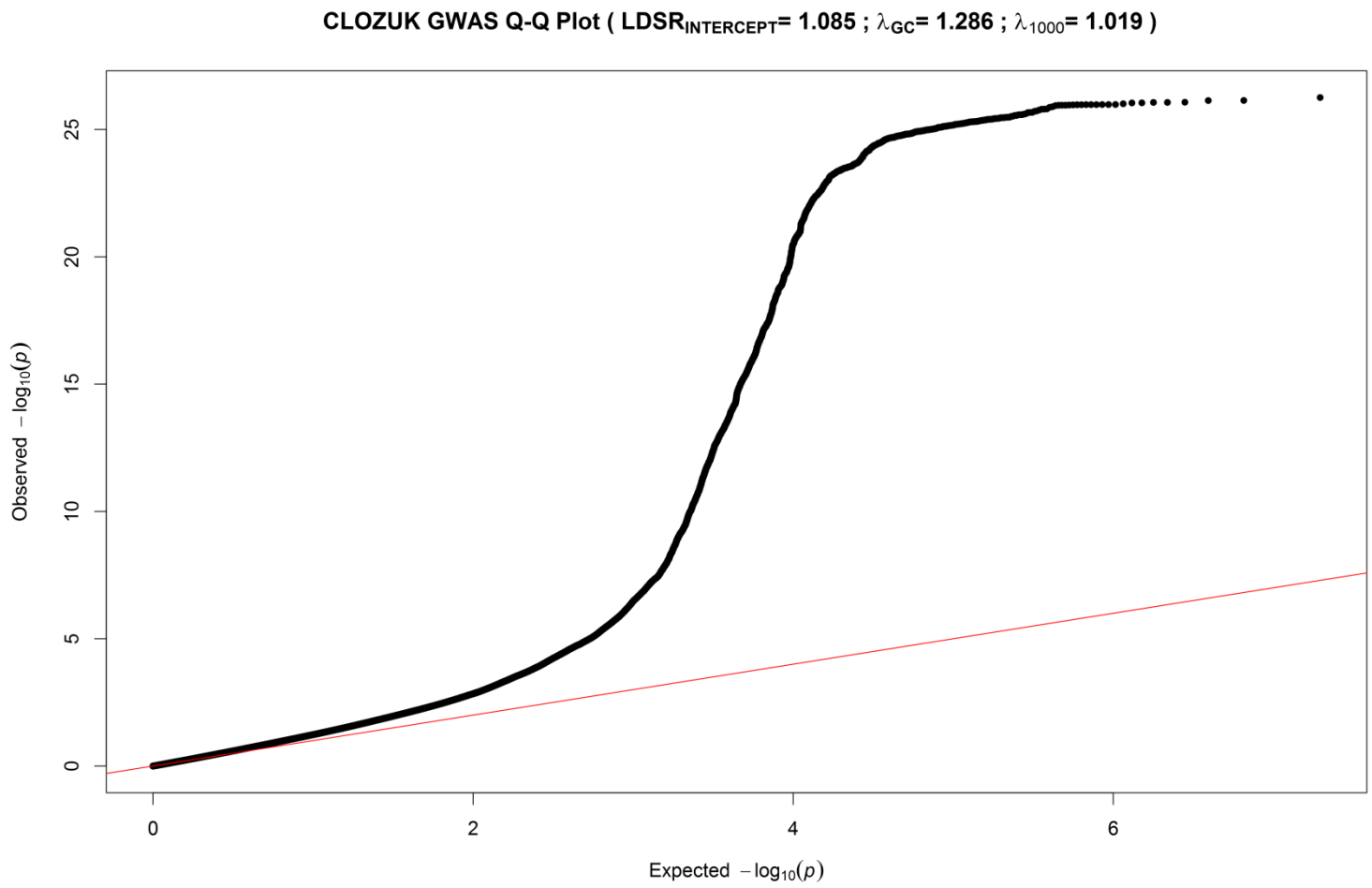
Schizophrenia association for genes within bins of pLI, an ExAC-based measure of intolerance to functional sequence variation. Bins are based on increasing 0.1 intervals of the statistic, and thus all LoF-intolerant genes (defined as pLI > 0.9) are in bin 10. P-values correspond to the statistical significance of a MAGMA competitive gene-set analysis.

Supplementary Figure 6



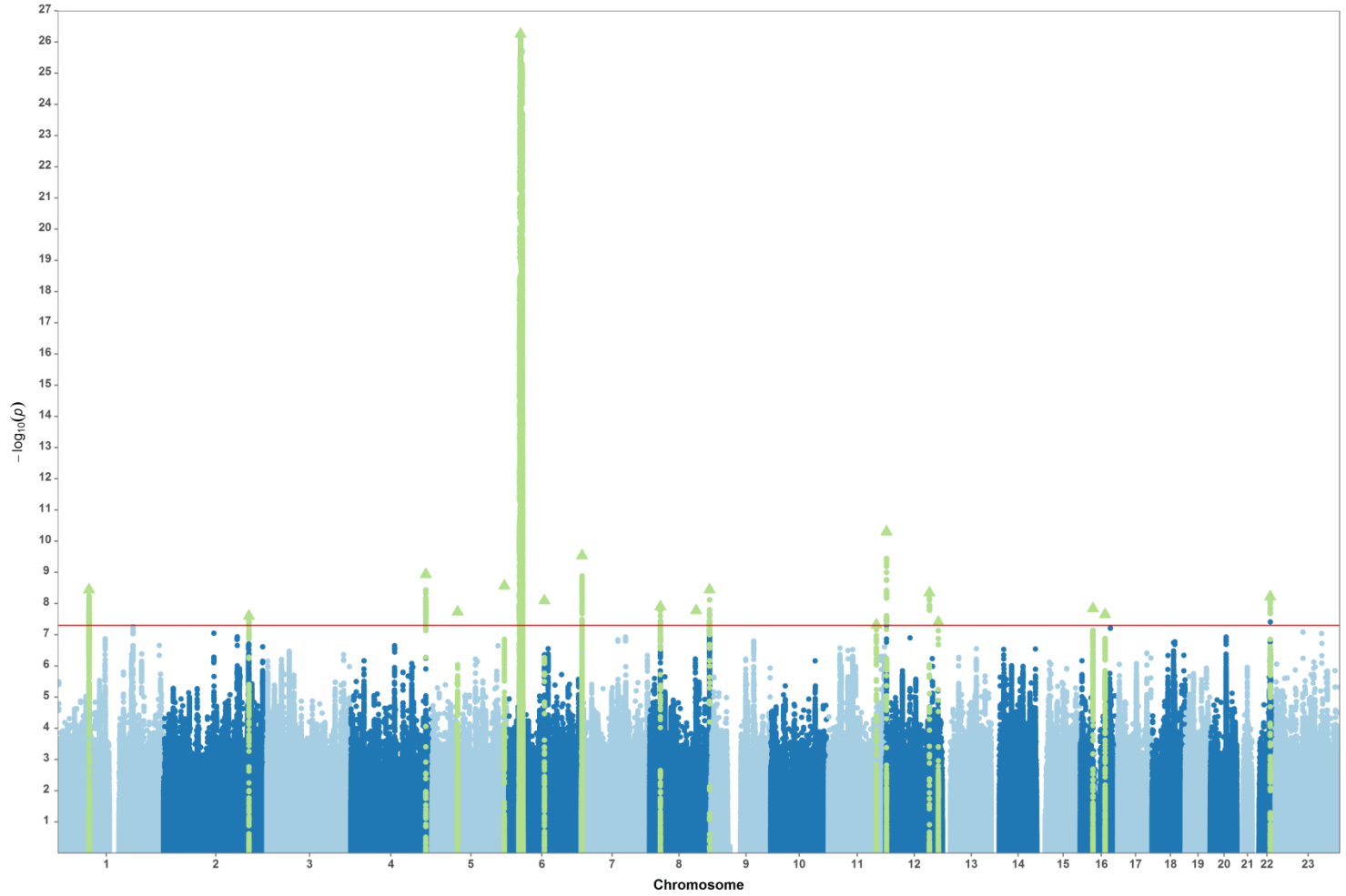
Schizophrenia SNP-based heritability enrichment, as estimated by LDSR, is influenced by the intensity of background selection ("B") and the genomic location. Error bars indicate enrichment standard errors. Asterisks indicate the significance of enrichment for each group of SNPs (* <= 0.05; ** <= 0.01<; *** <= 0.001).

Supplementary Figure 7



QQ Plot of the CLOZUK GWAS.

Supplementary Figure 8



Manhattan plot of the CLOZUK GWAS (N=35,802; 11 260 cases, 24,542 controls).