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

Antonio F. Pardiñas[a], Peter Holmans[a], Andrew J. Pocklington[a], Valentina Escott-Price[a], Stephan Ripke[b,c], Noa Carrera[a], Sophie E. Legge[a], Sophie Bishop[a], Darren Cameron[a], Marian L. Hamshere[a], Jun Han[a], Leon Hubbard[a], Amy Lynham[a], Kiran Mantripragada[a], Elliott Rees[a], James H. MacCabe[d], Steven A. McCarroll[e], Bernhard T. Baune[f], Gerome Breen[g,h], Enda M. Byrne[i,j], Udo Dannlowski[k], Thalia C. Eley[g], Caroline Hayward[l], Nicholas G. Martin[m,n], Andrew M. McIntosh[o,p], Robert Plomin[g], David J. Porteous[l], Naomi R. Wray[i,j], Armando Caballero[q], Daniel H. Geschwind[r], Laura M. Huckins[s], Douglas M. Ruderfer[s], Enrique Santiago[t], Pamela Sklar[s], Eli A. Stahl[s], Hyejung Won[r], Esben Agerbo[u,v,w], Thomas D. Als[u,z,aa], Ole A. Andreassen[x,au], Marie Bækvad-Hansen[u,y], Preben Bo Mortensen[u,v,w,z], Carsten Bøcker Pedersen[u,v,w], Anders D. Børglum[u,z,aa], Jonas Bybjerg-Grauholm[u,y], Srdjan Djurovic[ab,at], Naser Durmishi[ac], Marianne Giørtz Pedersen[u,v,w], Vera Golimbet[ad], Jakob Grove[u,z,aa,ae], David M. Hougaard[u,y], Manuel Mattheisen[u,z,aa], Espen Molden[af], Ole Mors[u,ag], Merete Nordentoft[u,ah], Milica Pejovic-Milovancevic[ai], Engilbert Sigurdsson[aj], Teimuraz Silagadze[ak], Christine Søholm Hansen[u,y], Kari Stefansson[al], Hreinn Stefansson[al], Stacy Steinberg[al], Sarah Tosato[am], Thomas Werge[u,an,ao], GERAD1 Consortium[ap], David A. Collier[g,aq], Dan Rujescu[ar,as], George Kirov[a], Michael J. Owen[a*], Michael C. O'Donovan[a*], James T. R. Walters[a*]

[a] MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK

[b] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA

[c] Department of Psychiatry and Psychotherapy, Charité, Campus Mitte, 10117 Berlin, Germany

[d] Department of Psychosis Studies, Institute of Psychiatry Psychology and Neuroscience, King's College London

[e] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[f] Discipline of Psychiatry, University of Adelaide, Australia

[g] Medical Research Council, Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

[h] NIHR Biomedical Research Centre for Mental Health, Maudsley Hospital and Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

[i] Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia

[j] Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia

[k] Department of Psychiatry and Psychotherapy University of Muenster, Muenster, Germany

[l] Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

[m] School of Psychology, University of Queensland

[n] QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia

[o] Division of Psychiatry, University of Edinburgh, Edinburgh, UK

[p] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK

[q] Departamento de Bioquímica, Genética e Inmunología. Facultad de Biología. Universidad de Vigo. Vigo, Spain

[r] Department of Neurology, Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, California, USA.

[s] Pamela Sklar Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[t] Departamento de Biología Funcional. Facultad de Biología. Universidad de Oviedo. Oviedo, Spain

[u] iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Denmark

[v] National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark

[w] Centre for Integrated Register-based Research, Aarhus University, Aarhus, Denmark

[x] Institute of Clinical Medicine, University of Oslo, Oslo, Norway

[y] Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark

[z] iSEQ, Center for Integrative Sequencing, Aarhus University, Aarhus, Denmark

[aa] Department of Biomedicine - Human Genetics, Aarhus University, Aarhus, Denmark

[ab] NORMENT, KG Jebsen Centre for Psychosis Research, Department of Clinical Science, University of Bergen, Bergen, Norway

[ac] Department of Child and Adolescent Psychiatry, University Clinic of Psychiatry, Skopje, Republic of Macedonia

[ad] Department of Clinical Genetics. Mental Health Research Center, Kashirskoe shosse 34. 115522. Moscow, Russia

[ae] Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark

[af] Center for Psychopharmacology, Diakonhjemmet Hospital, Oslo, Norway

[ag] Psychosis Research Unit, Aarhus University Hospital, Risskov, Denmark

[ah] Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, Copenhagen, Denmark

[ai] Department of Psychiatry, School of Medicine, University of Belgrade, Belgrade, Serbia

[aj] Department of Psychiatry, National University Hospital, Reykjavik, Iceland

[ak] Department of Psychiatry and Drug Addiction, Tbilisi State Medical University (TSMU), Tbilisi, Georgia

[al] deCODE genetics, Reykjavik, Iceland

[am] Section of Psychiatry, Department of Public Health and Community Medicine, University of Verona, Verona, Italy

[an] Institute of Biological Psychiatry, MHC Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark

[ao] Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

[ap] Control data used in the preparation of this article were obtained from the Genetic and Environmental Risk for Alzheimer's disease (GERAD1) Consortium. As such, the investigators within the GERAD1 consortia contributed to the design and implementation of GERAD1 and/or provided control data but did not participate in analysis or writing of this report.

[aq] Discovery Neuroscience Research, Eli Lilly and Company Ltd, Lilly Research Laboratories, Erl Wood Manor, Surrey, UK

[ar] Department of Psychiatry, University of Halle, Halle, Germany

[as] Department of Psychiatry, University of Munich, Munich, Germany

[at] Department of Medical Genetics, Oslo University Hospital, Oslo, Norway

[au] NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

* Correspondence and material requests should be addressed to M.J.O. (owenmj@cardiff.ac.uk), M.C.O'D. (odonovanmc@cardiff.ac.uk) or J.T.R.W. (waltersjt@cardiff.ac.uk)

ABSTRACT

Schizophrenia is a debilitating psychiatric condition often associated with poor quality of life and decreased life expectancy. Lack of progress in improving treatment outcomes has been attributed to limited knowledge of the underlying biology, although large-scale genomic studies have begun to provide insights. We report a new genome-wide association study of schizophrenia (11,260 cases and 24,542 controls) and through meta-analysis with existing data we identify 50 novel associated loci and 145 loci in total. Through integrating genomic fine-mapping with brain expression and chromosome conformation data, we identify candidate causal genes within 33 loci. We also show for the first time that the common variant association signal is highly enriched among genes that are under strong selective pressures. These findings provide novel insights into the biology and genetic architecture of schizophrenia, highlight the importance of mutation intolerant genes and suggest a mechanism by which common risk variants persist in the population.

Schizophrenia is characterised by psychosis and negative symptoms such as social and emotional withdrawal. While onset of psychosis typically does not occur until late adolescence or early adult life, there is strong evidence from clinical and epidemiological studies that schizophrenia reflects a disturbance of neurodevelopment[1]. It confers substantial mortality and morbidity, with a mean reduction in life expectancy of 15-30 years[2,3]. Although recovery is possible, most patients have poor social and functional outcomes[4]. No substantial improvements in outcomes have emerged since the advent of antipsychotic medication in the mid-20th century, a fact that has been attributed to a lack of knowledge of pathophysiology[1].

Schizophrenia is both highly heritable and polygenic, with risk ascribed to variants spanning the full spectrum of population frequencies[5-7]. The relative contributions of alleles of various frequencies is not fully resolved, but recent studies estimate that common alleles, captured by genome-wide association study (GWAS) arrays, explain between a third and a half of the genetic variance in liability[8]. There has been a long-standing debate, from an evolutionary standpoint, as to how common risk alleles persist in the population, particularly given the early mortality and decreased fecundity associated with schizophrenia[9]. Various hypotheses have been proposed including compensatory advantage (balancing selection), whereby schizophrenia alleles confer reproductive advantages in particular contexts[10,11]; hitchhiking, whereby risk alleles are maintained by their linkage to positively selected alleles[12]; or contrasting theories that attribute these effects to rare variants and gene-environment interaction[13]. Addressing these competing hypotheses is now tractable given advances from recent studies of common genetic variation in schizophrenia.

The largest published schizophrenia GWAS, that from the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC), identified 108 genome-wide significant (GWS) loci and unequivocally demonstrated the value of increasing sample sizes for discovery in schizophrenia genetics research[5]. Here, we report a large, phenotypically homogeneous GWA

study of schizophrenia which, when combined with previous published data, identifies novel facets of genetic architecture and biology, and demonstrates that the evolutionary process of background selection contributes to the persistence of common risk alleles in the population.

RESULTS

GWAS and Meta-analysis

We obtained genome-wide genotype information for schizophrenia cases from the UK (the CLOZUK sample), which we combined with control datasets obtained from public repositories or through collaboration. The final sample size was of 11,260 cases and 24,542 controls (5,220 cases and 18,823 controls not in previous schizophrenia GWAS**; Methods**; **Supplementary Figure 1** and **Supplementary Figure 2**). At a genome wide level, the association statistics indicated that the common variant architecture in the CLOZUK sample was highly correlated with an independent sample of 29,415 cases and 40,101 controls from the PGC (genetic correlation = 0.954±0.030; $p=6.63x10^{-227}$) and this was further confirmed by polygenic risk score and trend test analyses across the datasets at a range of association p-value thresholds (**Methods** and **Supplementary Table 1** and **Supplementary Table 2).**

Meta-analysis of the CLOZUK and the independent PGC dataset, excluding related and overlapping samples (total 40,675 cases and 64,643 controls; **Supplementary Figure 3**), identified 179 independent GWS SNPs ($p<5x10^{-8}$, **Supplementary Table 3**) mapping to 145 independent loci (**Methods, Figure 1, Supplementary Table 4**). The 145 associated loci include 93 of those that were GWS in the study of the PGC, the majority of which showed a strengthened association **(Supplementary Figure 4, Supplementary Table 5)**. This does not imply the remaining 15 PGC loci were false positives, rather this reflects the expected inflation of effect sizes for GWS SNPs in incompletely powered studies, and as we demonstrate, is

consistent with all 108 PGC loci representing true positives (see **Supplementary Note**). Of the 52 loci not identified by the PGC, two have been reported as genome-wide significant in other studies: the locus at ZEB2[14] and a locus on chromosome 8 (38.0-38.3 MB)[15].

In further independent samples (5,662 cases and 154,224 controls); 43 of the 50 GWS index SNPs showed the same pattern of allelic association, a level that far surpasses chance ($p=1.05x10^{-7}$). Despite the modest number of cases in these samples, 18 of the 50 index alleles reached nominal significance ($p<0.05$), again implausible by chance ($p=1.46x10^{-11}$). None demonstrated evidence for heterogeneity of effect (**Methods**, **Supplementary Table 6**).

Mutation intolerant genes

Recent studies have shown that mutation intolerant genes capture much of the rare variant architecture of neurodevelopmental disorders such as autism, intellectual disability and developmental delay as well as schizophrenia[16-19]. Here, we show that for schizophrenia, this also holds for common variation. Using gene set analysis in MAGMA[20], loss-of-function (LoF) intolerant genes (N=3,230) as defined by the Exome Aggregation Consortium (ExAC)[21] using their gene-level constraint metric pLI $\geq$ 0.9, were enriched for schizophrenia common variant associations in comparison with all other annotated genes ($p=4.1x10^{-16}$).

It has been shown that pLI is correlated with gene expression across tissues, including brain[21], which raises the possibility that the LoF-intolerant gene enrichment in schizophrenia may reflect enrichment for signal in genes expressed in the brain. However, LoF-intolerant gene set enrichment was robust to the inclusion of both "brain expressed" (N=10,360) and "brain specific" (N=2,647) gene sets[19] as covariates in the analysis ($p=1.89x10^{-10}$) or to controlling for FPKM gene expression values in brain[22] ($p=1.03x10^{-14}$).

It has been suggested that clustering of risk alleles in mutation intolerant genes is a hallmark of early-onset traits under natural selection[23,24]. However, LoF-intolerant genes are known to

be enriched for SNPs identified as genome-wide significant in GWAS studies (as listed in the NHGRI-EBI GWAS Catalog[25]) and for broad categories of disorders[21]. To examine whether our finding is a property of polygenic disorders in general, we obtained summary genetic data from a neuropsychiatric and non-psychiatric late-onset disorder (Alzheimer's disease, type-2 diabetes) and a psychological trait (Neuroticism), each of which has been shown to be under minimal selective pressure (see **Methods**). These other phenotypes show at best a weak signal for enrichment of the LoF-intolerant gene set in the MAGMA analysis, not comparable to that seen in schizophrenia (Alzheimer's disease p=0.008, type-2 diabetes p=0.016, Neuroticism p=0.066).

To quantify the contribution of SNPs within LoF-intolerant genes to schizophrenia SNP-based heritability (h^2_{SNP}) we used partitioned LDSR[26] (**Supplementary Table 7)**. Overall, genic SNPs account for 64% of h^2_{SNP}, a 1.23-fold enrichment proportional to their SNP content ($p=5.93x10^{-14}$). Consistent with the analysis using MAGMA, h^2_{SNP} was enriched in LoF-intolerant genes (2.01-fold; $p=2.78x10^{-24}$), which explained 30% of all h^2_{SNP} (equating to 47% of all genic h^2_{SNP}). In contrast, genes classed as non LoF-intolerant (pLI<0.9) were significantly depleted for h^2_{SNP} relative to their SNP content (0.90-fold; $p=5.86x10^{-3}$), although in absolute terms, SNPs in these genes accounted for 34% of h^2_{SNP}. A finer scale analysis of the relationship between LoF intolerance scores and enrichment for association showed that enrichment is restricted to genes with a pLI score above 0.9, precisely those defined as "LoF-intolerant" **(Supplementary Figure 5)**.

Common risk alleles in regions under background selection

Our finding that LoF-intolerant genes are enriched for common risk variants raises the question of how such alleles are found at common frequencies in the population. While the contribution of ultra-rare variation in functionally important genes to disorders associated with low fecundity can be accounted for by *de novo* mutation[16,19,27], this cannot explain the persistence

of common alleles. To address this question, we used partitioned LDSR to test the relationship between schizophrenia associated alleles and SNP-based signatures of natural selection. These included measures of positive selection, background selection, and Neanderthal introgression. We examined the heritability of SNPs after thresholding them at extreme values for these metrics (top 2%, 1% and 0.5%), including in the baseline model annotation sets such as LoF-intolerant genes and genomic regions with extreme linkage disequilibrium patterns **(Methods)**.

We observed strong evidence for schizophrenia h^2_{SNP} enrichment in SNPs under strong background selection (BGS), which was consistent across all the thresholds we examined (**Table 1**). We also found a significant depletion of h^2_{SNP} in SNPs subject to positive selection as indexed by the CLR statistic. These two results are mutually consistent, as the calculation of the CLR statistic explicitly controls for the effect of BGS[28]. This suggests that SNPs under positive selection, but under weak or no BGS, are depleted for association with schizophrenia. No significant relationship between h^2_{SNP} and other positive selection or Neanderthal introgression measures was found after correction for multiple testing **(Table 1)**. An LDSR analysis treating BGS measures as a quantitative trait, rather than as a binary one, confirmed that the relationship between BGS and schizophrenia association was not due to the imposition of arbitrary thresholds to define strong BGS ($p=7.73x10^{-11}$). We also note that the τ_c statistic of the LDSC model is significant for BGS, in both binary (p=0.041) and quantitative (p=0.023) analyses (**Supplementary Table 8)**. The τ_c statistic indicates the enrichment of BGS after controlling for all other annotations in the model (including LoF-intolerant genes)[26], and thus represents a robust and conservative test for the BGS enrichment.

The above analyses accounts for a possible confounding relationship between LoF intolerance and BGS. To illustrate this more clearly, we binned the BGS intensities into four categories of increasing score, and classified SNPs in these bins according to whether they are in LoF-intolerant genes, “all other” genes sets and a non-genic set **(Supplementary Figure 6)**. Note

that the lower boundary of the top bin (BGS intensity > 0.75) corresponds approximately to the top 2% BGS threshold in **Table 1** and is equivalent to a reduction in effective population size estimated at each SNP of 75% or more[29]. We found significant heritability enrichment across all BGS intensity intervals in LoF-intolerant genes that increased progressively with higher intensity scores. Importantly, we also found heritability enrichment for SNPs under BGS pressure in genes that are not LoF intolerant, restricted to the highest BGS intensity bin. Indeed the highest BGS intensity bin in non-LoF genes was enriched for heritability at a level roughly equivalent to all LoF genes. These findings point to BGS and LoF intolerance as making at least partially independent contributions to heritability enrichment in schizophrenia. In contrast, none of the phenotypes we selected on the basis of their minimal impact on fecundity (Alzheimer's disease, type-2 diabetes, neuroticism) showed significant BGS enrichment for heritability using either the BGS τ_c statistic of the LDSR model (minimum $p > 0.24$), or when specifically testing regions of high BGS intensity in genes that are tolerant (pLI<0.9) of functional mutations (minimum $p > 0.40$).

Systems genomics

Using MAGMA, we undertook a primary analysis of 134 central nervous system related gene sets we have previously shown captures the excess CNV burden in schizophrenia[30]. In a GWAS context, we now show that collectively, this group of gene sets captures a disproportionately high fraction of h^2_{SNP} (30% of total heritability; enrichment =1.63; $p= 8.57x10^{-13}$; 46% of genic heritability; **Supplementary Table 7**). Of the 134 sets, 54 were nominally significant of which 12 survived multiple-testing correction (family-wise error rate (FWER) $p < 0.05$, **Supplementary Table 9**), with no notable association for gene sets such as the ARC protein complex and NMDAR protein network, that we have previously implicated in rare variant studies[30,31]. Stepwise conditional analysis, adjusting sequentially for the more strongly associated gene sets, resulted in six gene sets that were independently associated with

schizophrenia (**Table 2** and **Data Supplement**). These extend from low-level molecular and sub-cellular processes to broad behavioural phenotypes. The most strongly associated gene set is constituted by the targets of the Fragile X Mental Retardation Protein (FMRP)[32]. FMRP is a neuronal RNA-binding protein that interacts with polyribosomal mRNAs (the 842 target transcripts of this gene set[32]) and is thought to act by inhibiting translation of target mRNAs, including many transcripts of pre- and post-synaptic proteins. The FMRP target set has been shown to be enriched for rare mutational burden in *de novo* exome sequencing studies of autism[33] and intellectual disability[31]. In schizophrenia studies, it has also been shown to be nominally significantly enriched for association signal in sequencing studies[8,31] and in GWAS[5,8] but only inconsistently in studies of copy number variation[30,34]. Here we provide the strongest evidence to date for the enrichment of this gene set in schizophrenia.

We highlight another five gene sets that are independently associated with schizophrenia. Three of these derive from the Mouse Genome Informatics database[35] and relate to behavioural and neurophysiological correlates of learning; Abnormal Behaviour (MP:0004924), Abnormal Nervous System Electrophysiology (MP:0002272) and Abnormal Long Term Potentiation (MP:0002207). We note that two of these gene sets (MP:0004924 and MP:0002207) were among the five most enriched of 134 gene sets tested in a recent schizophrenia CNV analysis[30]. The remaining two independently associated genes sets were voltage-gated calcium channel complexes[36] and the 5-HT_{2C} receptor complex[37]. The calcium channel finding confirms extensive evidence from common and rare variant studies implicating calcium channel genes in schizophrenia[5,8], including a novel GWAS locus in *CACNA1D* identified in our meta-analysis. Whilst there is less convergent evidence in support of the involvement of the 5-HT_{2C} receptor complex in schizophrenia, the fact that we identify independent association for this gene set implicates these genes in schizophrenia pathophysiology and potentially rejuvenates a previous avenue of 5-HT_{2C} ligand therapeutic endeavour in schizophrenia research[38].

However we interpret this result with caution given the small size of this gene set and the fact that a number of its genes encode synaptic proteins that are structurally related to other receptor complexes[37], not only 5-HT_{2C}.

Systems genomics and mutation intolerant genes

The LoF-intolerant genes and the six conditionally independent ("significant") CNS-related gene sets together account for 39% of schizophrenia SNP-based heritability ($p=5.07x10^{-26}$), equating to 61% of genic heritability **(Figure 2A; Supplementary Table 7)**. This is likely to be an underestimation of the true effect of these gene sets since distal non-genic regulatory elements (not included in this analysis) will add to the heritability explained by these genes. In examining the relationship between the LoF-intolerant and CNS-related gene sets **(Figure 2A)**, genes belonging to both categories were the most highly enriched (2.6-fold; $p=7.90x10^{-15}$), although LoF-intolerant genes that were not annotated to our significant CNS gene sets still displayed enrichment for SNP-based heritability (1.74-fold; $p=9.77x10^{-10}$), while genes that were in the significant CNS gene sets but had pLI<0.9 showed more modest enrichment (1.39-fold; $p= 6.05x10^{-4}$). Notably genes outside these categories were depleted in heritability relative to their SNP content (enrichment=0.79, $p=1.82x10^{-7}$).

This general pattern remained when we focussed on the six significant CNS gene sets individually, in that the enrichment in these gene sets derives primarily from their intersection with LoF-intolerant genes **(Figure 2B)**. Indeed, only the targets of FMRP showed significant enrichment for SNPs in genes that are not LoF intolerant (2.06-fold; $p=4.23x10^{-5}$).

Data-driven gene set analysis

To set the systems genomics results in context, and to ensure we were not missing enrichment in other gene sets by our hypothesis driven approach, we undertook a purely data-driven analysis of a larger comprehensive annotation of gene sets from multiple public databases,

totalling 6,677 gene sets **(Methods, Supplementary Table 10**). Six gene sets survived FWER correction for the full 6,677 gene sets and showed independence through conditional analyses. The LoF-intolerant gene set was the most strongly enriched followed by the two most strongly associated functional gene sets we had specified in our hypothesis-driven CNS gene set analysis (FMRP targets and MGI Abnormal Behaviour genes). The other three sets were calcium ion import (GO:0070509), membrane depolarisation during action potential (GO:0086010) and synaptic transmission (GO:0007268). These are highly overlapping with the independently associated sets from our primary CNS systems genomic analysis. Indeed if we repeat the data-driven comprehensive gene set analysis whilst adjusting for the six independently associated CNS gene sets, then the only surviving enrichment term is the LoF-intolerant genes. These results are consistent with those from CNV analysis[30] in that they do not support annotations other than those related to CNS function, and demonstrate that hypothesis based analysis to maximise power does not substantially impact on the overall pattern of results.

Identifying likely candidates within associated loci

To identify SNPs and genes which might be causally linked to the GWS associations, we used FINEMAP[39] to identify credibly causal alleles (those with a cumulative posterior probability for a locus of at least 95%) and functionally annotated these alleles using ANNOVAR[40]. This identified 6,105 credible SNPs across 144 GWS loci, excluding the MHC region (**Methods**, **Supplementary Table 11**). From these, we defined a highly credible set of SNPs (N=25) as those that are more likely to explain the associations than all other SNPs combined (i.e. with a FINEMAP posterior probability greater than 0.5). Of these, 14 mapped to genes based on putative functionality (exonic SNPs that cause non-synonymous or splice variations or promoter SNPs, n=6) or mapped to regions identified as likely regulatory elements (n=8) through chromosome conformation analysis performed in tissue from the developing brain

using Hi-C[41] physical interactions (**Methods; Supplementary Table 12**). One of the implicated alleles is a nonsynonymous variant in the manganese and zinc transporter gene *SLC39A8*. Nonsynonymous variants in this gene have been associated with severe neurodevelopmental disorders and deficiencies of SLC39A8 with related impaired glycosylation[42], highlighting a mechanism of therapeutic potential.

We also applied Summary-data based Mendelian Randomisation (SMR) analysis[43] to the data in concert with the dorsolateral prefrontal cortex eQTL data from the CommonMind Consortium[44] aiming to identify variants that might be causally linked through expression changes of specific genes. (**Methods**, **Supplementary Table 13**). After applying a conservative threshold ($p_{HEIDI}>0.05$) which prioritises those co-localised signals due to a single causal variant[43], we identified 22 candidates at 19 loci with a false discovery rate $p<0.05$.

In total, the combination of FINEMAP, Hi-C and SMR analyses assigned potentially causal genes at 33 GWS loci and implicated a single gene at 27 of these loci. However, the analyses intersect for a single gene, ZNF823, indicating the need for more comprehensive functional genomic annotations in CNS relevant tissues.

DISCUSSION

In the largest genetic study of schizophrenia to date, we explore the genomic architecture of, and the evolutionary pressures on, common variants associated with the disorder. Our study provides the first evidence linking common variation in LoF-intolerant genes to risk of developing schizophrenia and demonstrates that these genes account for a substantial proportion (30%) of schizophrenia SNP-based heritability. Systems genomic analysis highlights six gene sets that are independently associated with schizophrenia, and point to molecular, physiological and behavioural pathways involved in schizophrenia pathogenesis.

Given mutation intolerance is due to high selection pressure[21,23,24], our finding that schizophrenia risk variants that persist at common allele frequencies are enriched in loss-of-function intolerant genes might appear counter-intuitive. However, novel evidence presented here suggests this can be reconciled by background selection (BGS) which is a consequence of purifying selection in regions of low recombination[45,46]. In such regions, recurrent selection against deleterious variants causes haplotypes to be removed from the gene pool, which reduces genetic diversity in a manner equivalent to a reduction in effective population size[47]. This in turn impairs the efficiency of the selection process, allowing alleles with small deleterious effects to rise in frequency by drift[48]. Such a consequence of purifying selection has been shown to be compatible with the genomic architecture of complex human traits[49] and to influence phenotypes in model organisms[50]. We have explicitly modelled this effect (both theoretically and via simulations; **Supplementary Note**) and provide strong evidence for the feasibility of this effect as explanatory for the effect sizes seen for common alleles in schizophrenia.

We did not find enrichment for any measure of positive selection or Neanderthal introgression. A recent study explained a negative correlation between schizophrenia associations and metrics indicative of a Neanderthal selective sweep as evidence for positive selection or polygenic adaptation in schizophrenia[12]. We do not find any significant correlation in our model, which addresses the contribution of BGS, and hence our results are not consistent with large contributions of positive selection to the genetic architecture of schizophrenia (**Table 1**). Indeed positive selection is not widespread in humans, as reported by other studies that explicitly considered or accounted for BGS[28,51]. Polygenic adaptation, the co-occurrence of many subtle allele frequency shifts at loci influencing complex traits[52], remains an intriguing possibility, but has not been implicated in psychiatric phenotypes, including schizophrenia, in recent analyses[53,54]. In contrast, BGS has been proposed as a mechanism driving Human-Neanderthal incompatibilities, as regions with stronger estimated BGS have lower estimated

Neanderthal introgression[55]. We therefore conclude that the bulk of the BGS signal we obtain is unlikely to be influenced by positive selection[29], challenging theories of selective advantage of schizophrenia risk alleles to explain their high population frequencies.

AUTHOR CONTRIBUTIONS

A.F.P. curated and processed genetic data, performed statistical analyses, contributed to the interpretation of results and participated in the primary drafting of the manuscript.

P.H., A.J.P., V.E-P., A.C. and E.S. performed statistical analyses, contributed to the interpretation of results and participated in the primary drafting of the manuscript.

S.R. curated and processed genetic data and participated in the primary drafting of the manuscript.

N.C. and M.L.H. contributed to the interpretation of results and participated in the primary drafting of the manuscript.

S.E.L., S.B. and A.L. participated in the recruitment of participants for the study and curated and managed their phenotypic information.

D.C., J.H, L.H, E.R. and G.K. contributed and curated data used in the statistical analyses.

K.M. managed the laboratory and genotyping procedures in Cardiff University.

J. H. M, D. A. C. and D.R. supervised the recruitment of the participants for the study.

S. A. M. managed the genotyping of samples for the study.

N. R. W. contributed genotypes of control samples and participated in the primary drafting of the manuscript.

D.H. G, L. M. H., D. M. R., P. S., E. A. S. and H.W. performed statistical analyses and contributed to the interpretation of results.

M. J. O. and M. C. O'D. conceived and supervised the project, contributed to the interpretation of results and participated in the primary drafting of the manuscript.

J. T. R. W. conceived and supervised the project, led the recruitment of the participants and sample acquisition for the study, performed statistical analysis, contributed to the interpretation of results and participated in the primary drafting of the manuscript.

All other authors contributed genotypes of control samples or summary statistics of replication samples.

All authors had the opportunity to review and comment on the manuscript and all approved the final manuscript.

COMPETING FINANCIAL INTERESTS

D. A. C. is a full-time employee and stockholder of Eli Lilly and Company. The remaining authors declare no conflicts of interest.

REFERENCES

1. Owen, M.J., Sawa, A. & Mortensen, P.B. Schizophrenia. *Lancet* (2016).

2. Thornicroft, G. Physical health disparities and mental illness: the scandal of premature mortality. *The British Journal of Psychiatry* **199**, 441-442 (2011).

3. Olfson, M., Gerhard, T., Huang, C., Crystal, S. & Stroup, T. Premature mortality among adults with schizophrenia in the united states. *JAMA Psychiatry* **72**, 1172-1181 (2015).

4. Morgan, C. *et al.* Reappraising the long-term course and outcome of psychotic disorders: the AESOP-10 study. *Psychological Medicine* **44**, 2713-2726 (2014).

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

6. Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nature Neuroscience* **19**, 571-7 (2016).

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

8. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-190 (2014).

9. 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).

10. Huxley, J., Mayr, E., Osmond, H. & Hoffer, A. Schizophrenia as a Genetic Morphism. *Nature* **204**, 220-221 (1964).

11. Shaner, A., Miller, G. & Mintz, J. Schizophrenia as one extreme of a sexually selected fitness indicator. *Schizophrenia Research* **70**, 101-109 (2004).

12. Srinivasan, S. *et al.* Genetic Markers of Human Evolution Are Enriched in Schizophrenia. *Biological Psychiatry* **80**, 284–292 (2016).

13. Uher, R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. *Mol Psychiatry* **14**, 1072-1082 (2009).

14. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics* **45**, 1150-1159 (2013).

15. Shi, Y. *et al.* Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nature Genetics* **43**, 1224-1227 (2011).

16. McRae, J.F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).

17. Kosmicki, J.A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nature Genetics* **49**, 504–510 (2017).

18. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* **46**, 944-950 (2014).

19. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature Neuroscience* **19**, 1433-1441 (2016).

20. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).

21. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

22. Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics* **13**, 397-406 (2014).

23. Smith, N.G.C. & Eyre-Walker, A. Human disease genes: patterns and predictions. *Gene* **318**, 169-175 (2003).

24. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Current biology : CB* **18**, 883-889 (2008).

25. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**, D1001-D1006 (2014).

26. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228-1235 (2015).

27. Takata, A., Ionita-Laza, I., Gogos, Joseph A., Xu, B. & Karayiorgou, M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. *Neuron* **89**, 940-947 (2016).

28. Huber, C.D., DeGiorgio, M., Hellmann, I. & Nielsen, R. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology* **25**, 142-156 (2016).

29. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLoS Genet* **5**, e1000471 (2009).

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

31. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).

32. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-61 (2011).

33. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285-299 (2012).

34. Szatkiewicz, J.P. *et al.* Copy number variation in schizophrenia in Sweden. *Molecular Psychiatry* **19**, 762-73 (2014).

35. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A. & Richardson, J.E. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Research* **42**, D810-7 (2014).

36. Müller, C.S. *et al.* Quantitative proteomics of the Cav2 channel nano-environments in the mammalian brain. *Proceedings of the National Academy of Sciences* **107**, 14950-14957 (2010).

37. Becamel, C. *et al.* Synaptic multiprotein complexes associated with 5-HT(2C) receptors: a proteomic approach. *The EMBO journal* **21**, 2332-42 (2002).

38. Liu, J. *et al.* Prediction of Efficacy of Vabicaserin, a 5-HT(2C) Agonist, for the Treatment of Schizophrenia Using a Quantitative Systems Pharmacology Model. *CPT: Pharmacometrics & Systems Pharmacology* **3**, e111 (2014).

39. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).

40. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164-e164 (2010).

41. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).

42. Park, J.H. *et al.* SLC39A8 Deficiency: A Disorder of Manganese Transport and Glycosylation. *American Journal of Human Genetics* **97**, 894-903 (2015).

43. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**, 481-487 (2016).

44. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).

45. Charlesworth, B. The Effects of Deleterious Mutations on Evolution at Linked Sites. *Genetics* **190**, 5-22 (2012).

46. Charlesworth, B., Betancourt, A.J., Kaiser, V.B. & Gordo, I. Genetic Recombination and Molecular Evolution. *Cold Spring Harbor Symposia on Quantitative Biology* **74**, 177-186 (2009).

47. 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).

48. Charlesworth, B. Background Selection 20 Years on: The Wilhelmine E. Key 2012 Invitational Lecture. *Journal of Heredity* (2013).

49. North, T.L. & Beaumont, M.A. Complex trait architecture: the pleiotropic model revisited. *Scientific Reports* **5**, 9351 (2015).

50. Rockman, M.V., Skrovanek, S.S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science* **330**, 372-6 (2010).

51. Vitti, J.J., Grossman, S.R. & Sabeti, P.C. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics* **47**, 97-120 (2013).

52. Stephan, W. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Molecular ecology* **25**, 79-88 (2016).

53. Field, Y. *et al.* Detection of human adaptation during the past 2,000 years. *Science* (2016).

54. Key, F.M., Fu, Q., Romagne, F., Lachmann, M. & Andres, A.M. Human adaptation and population differentiation in the light of ancient genomes. *Nat Commun* **7**(2016).

55. Harris, K. & Nielsen, R. The Genetic Cost of Neanderthal Introgression. *Genetics* **203**, 881-891 (2016).

ACKNOWLEDGEMENTS

General

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 279227 (CRESTAR Consortium). The work at Cardiff University was funded by Medical Research

Council (MRC) Centre (MR/L010305/1), Program Grant (G0800509) and Project Grant (MR/L011794/1) and the European Community's Seventh Framework Programme HEALTH-F2-2010-241909 (Project EU-GEI). U.D. received funding by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1; SFB-TRR58, Project C09) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17). E.M.B. and N.R.W. received salary funding from the National Health and Medical Research Council (NHMRC; 1078901, 105363). E.S. and A.C. received funding from Agencia Estatal de Investigación (AEI; CGL2016-75904-C2-1-P), Xunta de Galicia (ED431C 2016-037) and Fondo Europeo de Desarrollo Regional (FEDER). The iPSYCH and GEMS2 teams acknowledge funding from The Lundbeck Foundation (grant no R102-A9118 and R155-2014-1724), the Stanley Medical Research Institute, an Advanced Grant from the European Research Council (project no: 294838), the Danish Strategic Research Council and grants from Aarhus University to the iSEQ and CIRRAU centers.

Case data

We thank the participants and clinicians who took part in the CardiffCOGS study. For the CLOZUK2 sample we thank Leyden Delta for supporting the sample collection, anonymisation and data preparation (particularly Marinka Helthuis, John Jansen, Karel Jollie and Anouschka Colson), Magna Laboratories, UK (Andy Walker) and, for CLOZUK1, Novartis and The Doctor's Laboratory staff for their guidance and cooperation. We acknowledge Lesley Bates, Catherine Bresner and Lucinda Hopkins, at Cardiff University, for laboratory sample management. We acknowledge Wayne Lawrence and Mark Einon, at Cardiff University, for support with the use and setup of computational infrastructures.

Control data

A full list of the investigators who contributed to the generation of the Wellcome Trust Case Control Consortium (WTCCC) data is available from its URL. Funding for the project was provided by the Wellcome Trust (WT) under award 076113. The UK10K project was funded by the Wellcome Trust award WT091310. Venous blood collection for the 1958 Birth Cohort (NCDS) was funded by the UK's Medical Research Council (MRC) grant G0000934, peripheral blood lymphocyte preparation by Juvenile Diabetes Research Foundation (JDRF) and WT and the cell-line production, DNA extraction and processing by WT grant 06854/Z/02/Z. Genotyping was supported by WT (083270) and the European Union (EU; ENGAGE: HEALTH-F4-2007- 201413). The UK Blood Services Common Controls (UKBS-CC collection) was funded by WT (076113/C/04/Z) and by the National Institute for Health Research (NIHR) programme grant to NHS Blood and Transplant authority (NHSBT; RP-PG-0310-1002). NHSBT also made possible the recruitment of the Cardiff Controls, from participants who provided informed consent. Generation Scotland (GS) received core funding from the Chief Scientist Office of the Scottish Government Health Directorates CZD/16/6 and the Scottish Funding Council HR03006. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the WT Clinical Research Facility, Edinburgh, Scotland and was funded by the MRC and Wellcome Trust (grant 10436/Z/14/Z). The Type 1 Diabetes Genetics Consortium (T1DGC; EGA dataset EGAS00000000038) is a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and JDRF. The People of the British Isles project (POBI) is supported by WT (088262/Z/09/Z). TwinsUK is funded by WT, MRC, EU, NIHR-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. Funding for the QIMR

samples was provided by the Australian National Health and Medical Research Council (241944, 339462, 389875, 389891, 389892, 389927, 389938, 442915, 442981, 496675, 496739, 552485, 552498, 613602, 613608, 613674, 619667), the Australian Research Council (FT0991360, FT0991022), the FP-5 GenomEUtwin Project (QLG2-CT- 2002-01254) and the US National Institutes of Health (NIH; AA07535, AA10248, AA13320, AA13321, AA13326, AA14041, MH66206, DA12854, DA019951), and the Center for Inherited Disease Research (Baltimore, MD, USA). TEDS is supported by a program grant from the MRC (G0901245-G0500079), with additional support from the NIH (HD044454; HD059215). In the GERAD1 Consortium, Cardiff University was supported by WT, MRC, Alzheimer's Research UK (ARUK) and the Welsh Government. Kings College London acknowledges support from the MRC. The University of Belfast acknowledges support from ARUK, Alzheimer's Society, Ulster Garden Villages, N.Ireland R&D Office and the Royal College of Physicians/Dunhill Medical Trust. Washington University was funded by NIH grants, Barnes Jewish Foundation and the Charles and Joanne Knight Alzheimer's Research Initiative. The Bonn group was supported by the German Federal Ministry of Education and Research (BMBF), Competence Network Dementia and Competence Network Degenerative Dementia, and by the Alfried Krupp von Bohlen und Halbach-Stiftung.

CRESTAR Consortium Members

Evanthia Achilla[1], Esben Agerbo[2,3] , Cathy L. Barr[4], Theresa Wimberly Böttger[2], Gerome Breen[5,6], Dan Cohen7, David A. Collier[5,8], Sarah Curran[9,10], Emma Dempster[11], Danai Dima[5], Ramon Sabes-Figuera[1], Robert J. Flanagan[12], Sophia Frangou[13], Josef Frank[14], Christiane Gasse[2,3], Fiona Gaughran[15], Ina Giegling[16], Jakob Grove[18,19], Eilis Hannon[11], Annette M. Hartmann[16], Barbara Heißerer[19], Marinka Helthuis[20], Henriette Thisted Horsdal[2], Oddur Ingimarsson[21], Karel Jollie[20], James L. Kennedy[22], Ole Köhler[23], Bettina Konte[16], Maren Lang[14], Sophie Legge[24], Cathryn Lewis[5], James MacCabe[15], Anil K. Malhotra[25], Paul

McCrone[1], Sandra M. Meier[2], Jonathan Mill[5,11], Ole Mors[23], Preben Bo Mortensen[2], Markus M. Nöthen[26], Michael C. O'Donovan[24], Michael J. Owen[24], Antonio F. Pardiñas[24], Carsten B. Pedersen[2,3], Marcella Rietschel[14], Dan Rujescu[16], Ameli Schwalber[19], Engilbert Sigurdsson[21], Holger J. Sørensen[27], Benjamin Spencer[28], Hreinn Stefansson[29], Henrik Støvring[3], Jana Strohmaier[14], Patrick Sullivan[30,31], Evangelos Vassos[5], Moira Verbelen[5], James T. R. Walters[24], Thomas Werge[27].

[1]Centre for Economics of Mental and Physical Health, Health Service and Population Research Department, Institute of Psychiatry, King's College London, London, UK. [2]National Centre for Register-Based Research, Department of Economics and Business, School of Business and Social Sciences, Aarhus University, Aarhus, Denmark. [3]Centre for Integrated Register-based Research, CIRRAU, Aarhus University, Aarhus, Denmark. [4]Toronto Western Research Institute, University Health Network, Toronto, Canada. [5]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [6]NIHR Biomedical Research Centre for Mental Health, Maudsley Hospital and Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [7]Department of Community Mental Health, Mental Health Organization North-Holland North, Heerhugowaard, The Netherlands. [8]Discovery Neuroscience Research, Eli Lilly and Company Ltd, Lilly Research Laboratories, Erl Wood Manor, Surrey, UK. [9]Department of Child & Adolescent Psychiatry, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. [10]Brighton and Sussex Medical School, University of Sussex, Brighton, UK. [11]University of Exeter Medical School, RILD, University of Exeter, Exeter, UK. [12]Toxicology Unit, Department of Clinical Biochemistry, King's College Hospital NHS Foundation Trust, London, UK. [13]Clinical Neurosciences Studies Center, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA. [14]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health,

Medical Faculty Mannheim/Heidelberg University, Mannheim, Germany. [15]Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [16]Department of Psychiatry, University of Halle, Halle, Germany. [17]Department of Biomedicine, Aarhus University, Denmark. [18]Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark. [19]Concentris Research Management GmbH, Fürstenfeldbruck, Germany. [20]Leyden Delta B.V., Nijmegen, The Netherlands. [21]Department of Psychiatry, Landspitali University Hospital, Reykjavik, Iceland. [22]Centre for Addiction and Mental Health, Toronto, Ontario, Canada. [23]Psychosis Research Unit, Aarhus University Hospital, Risskov, Denmark. [24]MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, UK. [25]Division of Psychiatry Research, The Zucker Hillside Hospital, Northwell Health System, Glen Oaks, New York, USA. [26]Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany. [27]Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, Denmark. [28]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [29]deCODE Genetics, Reykjavik, Iceland. [30]Center for Psychiatric Genomics, Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA. [31]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

GERAD1 Consortium Members

Denise Harold[1,9], Rebecca Sims[1], Amy Gerrish[1], Jade Chapman[1], Valentina Escott-Price[1], Richard Abraham[1], Paul Hollingworth[1], Jaspreet Pahwa[1], Nicola Denning[1], Charlene Thomas[1], Sarah Taylor[1], John Powell[2], Petroula Proitsi[2], Michelle Lupton[2], Simon Lovestone[2,10], Peter Passmore[3], David Craig[3], Bernadette McGuinness[3], Janet Johnston[3], Stephen Todd[3], Wolfgang Maier[4], Frank Jessen[4], Reiner Heun[4], Britta Schurmann[4, 5], Alfredo Ramirez[4], Tim Becker[6],

Christine Herold[6], André Lacour[6], Dmitriy Drichel[6], Markus Nothen[7], Alison Goate[8], Carlos Cruchaga[8], Petra Nowotny[8], John C. Morris[8], Kevin Mayo[8], Peter Holmans[1], Michael O'Donovan[1], Michael Owen[1] and Julie Williams[1].

[1]Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics, Neurosciences and Mental Health Research Institute, Department of Psychological Medicine and Neurology, School of Medicine, Cardiff University, Cardiff, UK. [2]King's College London, Institute of Psychiatry, Department of Neuroscience, De Crespigny Park, Denmark Hill, London. [3]Ageing Group, Centre for Public Health, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, UK. [4]Department of Psychiatry, University of Bonn, Sigmund-Freud-Straβe 25, 53105 Bonn, Germany. [5]Institute for Molecular Psychiatry, University of Bonn, Bonn, Germany. [6]Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn. [7]Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany. [8]Departments of Psychiatry, Neurology and Genetics, Washington University School of Medicine, St Louis, MO 63110, US. [9]Present address: Neuropsychiatric Genetics Group, Department of Psychiatry, Trinity Centre for Health Sciences, St James's Hospital, Dublin. [10]Present address: Department of Psychiatry, University of Oxford.

FIGURE LEGENDS

Figure 1. Manhattan plot of schizophrenia GWAS associations from the meta-analysis of CLOZUK and an independent PGC dataset (N=105,318; 40,675 cases and 64,643 controls). The 145 genome-wide significant loci are highlighted in green.

Figure 2. Partitioned heritability analysis of gene sets in schizophrenia. **A**: Heritability of genomic partitions and the six conditionally independent ("significant") gene sets (**Table 2**). Radius of each segment indicates the degree of enrichment, while the arc (angle of each slice) indicates the percentage of total SNP-based heritability explained. No relative enrichment

(enrichment=1) is shown by the dashed red line (and depletion equates to enrichment<1, inside red line). **B**: Heritability of the significant CNS gene sets dissected by their overlap with LoF-intolerant genes. Whiskers represent heritability/enrichment standard errors. Asterisks indicate the significance of each heritability enrichment (* <= 0.05; ** <= 0.01; *** <= 0.001).

		Top 2% of scores (genome-wide)		Top 1% of scores (genome-wide)		Top 0.5% of scores (genome-wide)	
Metric	**Ref.**	**Enrichment**	**2-sided p-value**	**Enrichment**	**2-sided p-value**	**Enrichment**	**2-sided p-value**
Background Selection (B-statistic)	29	<u>1.801</u>	<u>0.001</u>	<u>2.341</u>	<u>9.90×10^{-4}</u>	<u>2.365</u>	<u>0.002</u>
Positive selection (CLR)	28	<u>0.408</u>	<u>6.53×10^{-5}</u>	<u>0.173</u>	<u>5.80×10^{-7}</u>	<u>0.259</u>	<u>0.016</u>
Positive selection (CMS)	106	<u>0.054</u>	<u>0.001</u>	-0.037	0.006	-0.039	0.007
Positive selection (XP-EEH)	105	0.621	0.342	0.383	0.303	0.125	0.268
Positive selection (iHS)	104	0.973	0.946	0.980	0.974	1.633	0.557
Neanderthal posterior probability (LA)	107	0.807	0.347	0.800	0.462	0.858	0.745

Table 1. Heritability analysis of natural selection metrics.

Partitioned LDSR regression results for SNPs thresholded by extreme values (defined as top percentiles vs all other SNPs) of each natural selection metric. All tests have been adjusted for 58 "baseline" annotations, which include categories such as “LoF-intolerant”, “recombination coldspot” and “conserved” (see **Methods**). Enrichment values below 1 indicate a depletion of h^2_{SNP} in an annotation category (less contribution than expected for a given number of SNPs). Negative enrichments should be considered zero (no contribution to h^2_{SNP} by these SNPs). Underlined values indicate results surviving correction after adjusting for all tests (Bonferroni $\alpha = 0.05/18 = 0.0028$). Reference numbers for each metric indicate references in the main text.

Gene Set	Number of genes	Enrichment p value (FWER)	Conditional p value
Targets of FMRP[32]	798	$1x10^{-5}$	$1.9x10^{-8}$
Abnormal Behaviour **(MP:0004924)**	1939	$1.8x10^{-4}$	$1.4x10^{-5}$
5-HT$_{2C}$ receptor complex[37]	16	0.029	0.001
Abnormal Nervous System Electrophysiology **(MP:0002272)**	201	0.003	0.002
Voltage-gated calcium channel complexes[36]	196	0.011	0.016
Abnormal Long Term Potentiation **(MP:0002207)**	142	0.030	0.031

Table 2. Functional gene set analysis highlights six independent gene sets associated with schizophrenia

FMRP: Fragile X Mental Retardation Protein.
MP refers to Mammalian Phenotype Ontology term of the MGI: Mouse Genome Informatics[35], from which gene sets were derived.
FWER: Westfall-Young family-wise error rate, as implemented in MAGMA[20,88].
Conditional p value refers to stepwise conditional analysis that adjusts sequentially for 'stronger' associated gene sets.

METHODS

GWAS and reporting of independently-associated regions

Details of sample collection and genotype quality control are given in the **Supplementary Note**. The CLOZUK schizophrenia GWAS was performed using logistic regression with imputation probabilities ("dosages") adjusted for 11 PCA covariates. These covariates were chosen as those nominally significant ($p < 0.05$) in a logistic regression for association with the phenotype[1]. To avoid overburdening the GWAS power by adding too many covariates to the regression model[2], only the first twenty PCs were considered and tested for inclusion, as higher numbers of PCs only become useful for the analysis of populations that bear strong signatures of complex admixture[3]. The final set of covariates included the first five PCs (as recommended for most GWAS approaches[4]) and PCs 6, 9, 11, 12, 13 and 19. Quantile-quantile (QQ) and Manhattan plots are shown in **Supplementary Figure 7** and **8**.

In order to identify independent signals among the regression results, signals were amalgamated into putative associated loci using the same two-step strategy and parameters as PGC (**Supplementary Table 14**). In this procedure, regular LD-clumping is performed ($r^2 = 0.1$, $p < 1x10^{-4}$; window size <3 Mb) in order to obtain independent index-SNPs. Afterwards loci are defined for each index SNP as the genomic region which contains all other imputed SNPs within an $r^2 \geq 0.6$. To avoid inflating the number of signals in gene-dense regions or in those with complex LD, all loci within 250kb of each other were annealed.

Meta-analysis with PGC

A total of 6,040 cases and 5,719 controls from CLOZUK were included in the recent PGC study[5]. We reanalysed the PGC data after excluding all these cases and controls, obtaining a sample termed 'INDEPENDENT PGC' (29,415 cases and 40,101 controls). Adding the summary statistics from this independent sample to the CLOZUK GWAS results allowed for

a combined analysis of 40,675 cases and 64,643 controls (without duplicates or related samples). This meta-analysis was performed using the fixed-effects procedure in METAL[6] with weights derived from standard errors. For consistency with the PGC analysis, additional filters (INFO>0.6 and MAF>0.01) were applied to the CLOZUK and INDEPENDENT PGC summary statistics, leaving 8 million markers in the final meta-analysis results. QQ and Manhattan plots are shown in **Supplementary Figure 3** and **Figure 2**. The same procedure as above was used in order to report independent loci from this analysis (**Supplementary Table 3**, **Supplementary Table 4**). As raw PGC genotypes were not available for the LD-clumping procedure, 1KGPp3 was used as a reference.

Replication of new GWAS loci

In order to validate the association signals from the CLOZUK+PGC meta-analysis, we amalgamated data contributed by other schizophrenia genetics consortia (total of 5,762 cases and 154,224 controls, details in **Supplementary Note**). We sought GWAS summary statistic data for the index SNPs from the 50 novel genome-wide significant loci **(Supplementary Table 4**). These summary statistics were meta-analysed in METAL using the fixed-effects procedure to obtain replication and heterogeneity statistics (**Supplementary Table 6**).

Estimation and assessment of a polygenic signal

Association signals caused by the vast polygenicity underlying complex traits can be hard to distinguish from confounders related to sample relatedness and population stratification. In order to effectively disentangle this issue, we used the software LD-Score v1.0 to analyse the summary statistics of our association analyses, and estimate the contribution of confounding biases to our results by LDSR[7]. An LD-reference was generated from 1KGPp3 after restricting this dataset to strictly unrelated individuals[8] and retaining only markers with MAF>0.01. In order to improve accuracy, the summary statistics used as input were refined by discarding all

indels and restricting SNPs to those with INFO>0.9 and MAF>0.01, a total of 5.16 million SNPs. The resulting LD-score intercept for the CLOZUK GWAS was 1.085±0.010, which compared to a mean χ^2 of 1.417 indicates a polygenic contribution of at least 80%. For the CLOZUK+PGC meta-analysis, the LD-score intercept was 1.075±0.014 (mean χ^2= 1.960), which supports more than 90% of the signal being driven by polygenic architecture. Both of these figures are in line with other well-powered GWAS studies of complex human traits[7], including schizophrenia[5]. This analysis was also used to calculate SNP-based heritability (h^2_{SNP}) for our three datasets (CLOZUK, INDEPENDENT PGC and the CLOZUK+PGC meta-analysis), which we transformed to a liability scale using a population prevalence of 1% (registry-based lifetime prevalence[9]). For reference and compatibility with epidemiological studies of schizophrenia, prevalence estimates of 0.7% (lifetime morbid risk[10]) and 0.4% (point prevalence[10], more akin to treatment resistant schizophrenia prevalence (appropriate for CLOZUK) were used for additional liability-scale h^2_{SNP} calculations (**Supplementary Table 15**).

The LDSR framework allowed us to compare the genetic architecture of CLOZUK and INDEPENDENT PGC, by calculating the correlation of their summary statistics[11]. A genetic correlation coefficient of 0.954±0.030 was obtained, with a p-value of 6.63×10^{-227}. We also examined the independent SNPs that reached a genome-wide significant (GWS) level in the INDEPENDENT PGC dataset, of which there were 76 after excluding the extended major histocompatibility complex region (xMHC). In the CLOZUK sample 76% (n=57) of these GWS SNPs were nominally significant ($p<0.05$). Using binomial sign tests based on clumped subsets of SNPs[12] we found that all but 1 (98.6%) of these 76 GWS SNPs were associated with the same direction of effect in the CLOZUK sample, a result highly unlikely to reflect chance ($p=2.04 \times 10^{-21}$, **Supplementary Table 1**). Moreover, of the 1,160 SNPs with an association p-value less than 10^{-4} in the INDEPENDENT PGC sample, 82% showed enrichment in the

CLOZUK cases ($p=3.44x10^{-113}$), confirming very large numbers of true associations will be discovered amongst these SNPs with increased sample sizes. Additionally, the new sample introduced in this study (CLOZUK2) was compared by the same methods with the PGC dataset and showed results consistent with the full CLOZUK analysis, providing molecular validation of this sample as a schizophrenia sample (**Supplementary Table 1**).

We went on to conduct polygenic risk score analysis. Polygenic scores for CLOZUK were generated from INDEPENDENT PGC as a training set, using the same parameters for risk profile score (RPS) analysis in PGC[5], arriving at a high-confidence set of SNPs for RPS estimation by removing the xMHC region, indels, and applying INFO>0.9 and MAF>0.1 cut-offs. Scores were generated from the autosomal imputation dosage data, using a range of p-value thresholds for SNP inclusion[13] ($5x10^{-8}$, $1x10^{-5}$, 0.001, 0.05 and 0.5). In this way, we can assess the presence of a progressively increasing signal-to-noise ratio in relation to the number of markers included[14]. As in the PGC study, we find the best p-value threshold for discrimination to be 0.05 and report highly significant polygenic overlap between the INDEPENDENT PGC and CLOZUK samples ($p<1x10^{-300}$, Nagelkerke $r^2=0.12$, **Supplementary Table 2**), confirming the validity of combining the datasets. For comparison with other studies we also report polygenic variance on the liability scale[15], which amounted to 5.7% for CLOZUK at the 0.05 p-value threshold (**Supplementary Table 2)**. As in the PGC study the limited r^2 and Area Under the Receiver Operating Characteristic curve (AUROC) obtained by this analysis restricts the current clinical utility of these scores in schizophrenia.

Gene set analysis

In order to assess the enrichment of sets of functionally related genes, we used MAGMA v1.03[16] on the CLOZUK+PGC meta-analysis summary statistics. From these we excluded the xMHC region for its complex LD and the X-chromosome given its smaller sample size. In the resulting data, gene-wide p-values were calculated by combining the p-values of all SNPs

inside genes after accounting for LD and outliers. This was performed allowing for a window of 35 kb upstream and 10 kb downstream of each gene in order to capture the signal of nearby SNPs that could fall in regulatory regions[17,18]. Next we calculated competitive gene set p-values on the gene-wide p-values after accounting for gene size, gene set density and LD between genes. For multiple testing correction in each gene set collection, a FWER[19] was computed using 100,000 re-samplings.

We performed sequential analyses using three approaches:

1. Loss-of-function intolerant genes: We tested the enrichment of the loss-of-function (LoF) intolerant genes described by ExAC[20]. This set comprises all genes defined in the ExAC database[21] as having a probability of LoF intolerance (pLI) statistic higher than 90%. While these genes do not form part of cohesive biological processes or phenotypes, they have been previously found to be highly expressed across tissues and developmental stages[20]. Also, they are enriched for hub proteins[22], which makes them interesting candidates for involvement in the "evolutionary canalisation" processes that have been proposed to lead to pleiotropic, complex disorders[23].

2. CNS-related genes: These gene sets were compiled in our recent study[24], and include 134 gene sets related to different to aspects of central nervous system function and development. These include, among others, gene sets which have been implicated in schizophrenia by at least two independent large-scale sequencing studies[25,26]: targets of the fragile-X mental retardation protein (FMRP[27]), constituents of the N-methyl-D-aspartate receptor (NMDAR[28]) and activity-regulated cytoskeleton-associated protein complexes (ARC[29,30]), as well as CNS and behavioural gene sets from the Mouse Genome Informatics database version 6 [31].

3. Data-driven: The final systems genomic analysis was designed as an "agnostic" approach, with the aim of integrating a large number of gene sets from different public sources, not

necessarily conceptually related to psychiatric disorders, as this has been successful elsewhere[18,32]. We conducted this analysis to test whether additional gene sets were associated in addition to those from the 134 *CNS-related* gene sets. For this, first we merged together the LoF-intolerant gene set and the 134 sets in the *CNS-related* collection. Second, we selected additional gene set sources to encompass a comprehensive collection of biochemical pathways and gene regulatory interaction networks: 2,693 gene sets with direct experimental evidence and a size of 10-200 genes[18] were extracted from the Gene Ontology (GO[33]) database release 01/02/2016; 1,787 gene sets were extracted from the 4th ontology level of the Mouse Genome Informatics database version 6; 1,585 gene sets were extracted from REACTOME[34] version 55; 290 gene sets were extracted from KEGG[35] release 04/2015; and 187 gene sets were extracted from OMIM[36] release 01/02/2016. The total number of gene sets included was 6,677.

Detailed results of the analyses of the CNS-related and data-driven collection are given in **Supplementary Table 9** and **Supplementary Table 10**. Reported numbers of genes in each gene set are those with available data in the meta-analysis. This may differ from the original gene set description as some genic regions had null or poor SNP coverage. Following the data-driven gene set analysis as described we also conducted analysis adjusting for our CNS-related gene sets to determine whether the data-driven analysis was contributing additional findings.

Partitioned heritability analysis of gene sets

It is known that the power of a gene set analysis is closely related to the total heritability of the phenotype and the specific heritability attributable to the tested gene set[37]. In order to assess the heritability explained by the genes carried forward after the main gene set analysis, LD-Score was again used to compute a partitioned heritability estimate of CLOZUK+PGC using the gene sets as SNP annotations. As in the MAGMA analysis, the xMHC region was excluded from the summary statistics. These were also trimmed to contain no indels, and only markers with INFO > 0.9 and MAF > 0.01, for a total of 4.64 million SNPs. As a recognised caveat of

this procedure is that model misspecification can inflate the partitioned heritability estimates[38], all gene sets were annotated twice: Once using their exact genomic coordinates (extracted from the NCBI RefSeq database[39]) and another with putative regulatory regions taken into account using the same upstream/downstream windows as in the MAGMA analyses. Additionally, all SNPs not directly covered by our gene sets of interest were explicitly included into other annotations ("non-genic", "genic but not LoF-intolerant") based on their genomic location. Finally, the "baseline" set of 53 annotations from Finucane et al. 2015[38], which recapitulates important molecular properties such as presence of enhancers or phylogenetic conservation, was also incorporated in the model. All of these annotations were then tested jointly for heritability enrichment. We note that using exact genic coordinates or adding regulatory regions made little difference to the estimated enrichment of our gene sets, and thus throughout the manuscript we report the latter for consistency with the gene set analyses (**Figure 2; Supplementary Table 8**).

Natural selection analyses

We aimed to explore the hypothesis that some form of natural selection is linked to the maintenance of common genetic risk in schizophrenia[40-42]. In order to do this, for all SNPs included in the CLOZUK+PGC meta-analysis summary statistics, we obtained four different genome-wide metrics of positive selection (iHS[43], XP-EEH[44], CMS[45] and CLR[46]), one of background selection (B-statistic[47], post-processed by Huber et al. 2016[46]) and one of Neanderthal introgression (average posterior probability LA[48]). The use of different statistics is motivated by the fact that each of them is tailored to detect a particular selective process that acted on a particular timeframe (see Vitti et al. 2013[49] for a review). For example, iHS and CMS are based on the inference of abnormally long haplotypes, and thus are better powered to detect recent selective sweeps that occurred during the last ~30,000 years[50], such as those linked to lactose tolerance or pathogen response[45]. On the other hand, CLR incorporates

information about the spatial pattern of genomic variability (the site frequency spectrum[51]), and corrects explicitly for evidences of background selection, thus being able to detect signals from 60,000 to 240,000 years ago[46]. The B-statistic uses phylogenetic information from other primates (chimpanzee, gorilla, orang-utan and rhesus macaque) in order to infer the reduction in allelic diversity that exists in humans as a consequence of purifying selection on linked sites over evolutionary timeframes[52]. As the effects of background selection on large genomic regions can mimic those of positive selection[53], it is possible that the B-statistic might amalgamate both, though the rather large diversity reduction that it infers for the human genome as a whole suggests any bias due to positive selection is likely to be minor[54]. Finally, XP-EEH is a haplotype-based statistic which compares two population samples, and thus its power is increased for alleles that have suffered differential selective pressures since those populations diverged[44]. Though methodologically different, LA has a similar rationale by comparing human and Neanderthal genomes[48], in order to infer the probability of each human haplotype to have been the result of an admixture event with Neanderthals.

For this work, CLR, CMS, B-statistic and LA were retrieved directly from their published references, and lifted over to GRC37 genomic coordinates if required using the ENSEMBL LiftOver tool[55,56]. As the available genome-wide measures of iHS and XP-EEH were based on HapMap3 data[57], both statistics were re-calculated with the HAPBIN[58] software directly on the EUR superpopulation of the 1KGPp3 dataset, with the AFR superpopulation used as the second population for XP-EEH. Taking advantage of the fine-scale genomic resolution of these statistics (between 1-10 kb), all SNP positions present in CLOZUK+PGC were assigned a value for each measure, either directly (if the position existed in the lifted-over data) or by linear interpolation. To simplify the interpretation of our results, all measures were transformed before further analyses to a common scale, in which larger values indicate stronger effect of selection or increased probability of introgression. For example, the background selection B-

statistic, in which values of zero indicate the strongest effect (see Charlesworth 2012[59] for its theoretical derivation), was included in all our analyses as 1 - B, which we termed "BGS intensity" in the main text.

Heritability enrichment of these statistics was tested by the LD-Score partitioned heritability procedure. We derived binary annotations from the natural selection metrics by dichotomising at extreme cut-offs defined by the top 2%, 1% and 0.5% of the values of each metric in the full set of SNPs. This approach is widely used in evolutionary genomics, due to the difficulty of setting specific thresholds to define regions under selection[46,49]. Consistent with the previously described LDSR partitioned heritability protocol, enrichment was estimated with all binary annotations included in a model with multiple categories that represent important genomic features. This model included the 3 main categories of our set-based analysis ("non-genic", "genic" and "LoF-intolerant"), 2 categories based on genomic regions with outlying LD patterns (recombination hotspots and coldspots)[60], and the 53 "baseline" categories of Finucane et al. 2015[38].

We then derived the τ_c coefficient[38] (and associated p value) of the significantly enriched natural selection annotations (i.e. the background selection metric), This represents the enrichment of an annotation over and above the enrichment of all other annotations, which is a conservative approach, as most of the categories in our model are partially overlapping. In order to increase our power and for additional validation, we noted that LD-Score allows testing the full range of quantitative metrics, in an extension of the partitioned heritability framework. Results of this analysis are reported in **Supplementary Table 8**.

Analysis of other phenotypes

To explore the specificity of our natural selection results, we retrieved data from other well-powered GWAS of complex traits. We selected three phenotypes for which a) the genome-

wide summary statistic data are publicly available b) the sample size is larger than 50,000 individuals, and c) have minimal impact on fecundity[61-63] (and hence the traits behave as neutral or approximately neutral to selection) d) summary statistics were considered adequate for LD-Score analysis based on baseline Z-scores >4[38,64] (**Supplementary Table 8**). The phenotypes chosen were Alzheimer's disease[65], Neuroticism[66] and Type-2 diabetes[67]. For the LD-Score analyses, as the public release of these statistics did not include imputation INFO scores at the time of this study, we restricted the set of SNPs to those included in the HapMap3 project[68], as recommended[7]. To facilitate comparison with the schizophrenia results, we also restricted our schizophrenia summary statistic data to these SNPs, and repeated the analyses above using BGS as a binary (top 2%) and quantitative trait.

We also employed MAGMA on the summary statistics of these additional phenotypes in order to examine whether the LoF-intolerant gene set enrichment displayed specificity to schizophrenia, after excluding the xMHC and APOE regions.

Finemapping, Hi-C and SMR

Accurately locating causal genes ("fine-mapping") for complex disorders is a challenge to GWAS studies, and usually requires multiple approaches[123]. To highlight credibly causal variants, we used FINEMAP v1.1[69] at each of the 145 identified loci (**Supplementary Table 3**), selecting variants with a cumulative posterior probability of 95%. These were then annotated with ANNOVAR[70] release 2016Feb1 (**Supplementary Table 11**). We mapped the SNPs with a FINEMAP posterior probability higher than 0.5 to the developing brain Hi-C data generated by Won et al. 2016[71], following the methodology described therein, which allowed us to implicate genes by chromatin interactions instead of solely chromosomal position (**Supplementary Table 12**). We compiled results from the eQTL analysis of the CommonMind Consortium post-mortem brain tissues[72]. This included 15,782 genes, which were curated to remove any genes with FPKM=0 across >10% of individuals. All the SNPs from the meta-

analysis data were mapped to the eQTL data using RS numbers, position, and allele matching. Both datasets were analysed together using SMR[73], which resulted in 4,276 genes showing eQTLs with overlapping SNPs and genome-wide significant p-values (**Supplementary Table 13**).

URLs

CLOZUK+PGC2 meta-analysis summary statistics: walters.psycm.cf.ac.uk

CRESTAR consortium: www.crestar-project.eu/

Wellcome Trust Case/Control Consortium: www.wtccc.org.uk

People of the British Isles project: www.peopleofthebritishisles.org

Mouse Genome Informatics: www.informatics.jax.org

Psychiatric Genomics Consortium: www.med.unc.edu/pgc/

DATA AVAILABILITY STATEMENT

The gene content of the CNS-related gene sets that survive conditional analysis ("significant") is given in MAGMA format in the **Data Supplement**. Summary statistics from the CLOZUK+PGC2 GWAS are available for download (see URLs).

METHODS REFERENCES

1. Peloso, G.M. & Lunetta, K.L. Choice of population structure informative principal components for adjustment in a case-control study. *BMC Genetics* **12**, 64-64 (2011).
2. Pirinen, M., Donnelly, P. & Spencer, C.C.A. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics* **44**, 848-851 (2012).

3. Bouaziz, M., Ambroise, C. & Guedj, M. Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. *PLoS ONE* **6**, e28845 (2011).

4. Tucker, G., Price, A.L. & Berger, B.A. Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select. *Genetics* **197**, 1045-1049 (2014).

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

6. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191 (2010).

7. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291-295 (2015).

8. Browning, S.R. & Browning, B.L. Identity by Descent (IBD) segment sharing within and between populations. (1000 Genomes FTP site, 2015).

9. Perälä, J., Suvisaari, J., Saarni, S.I. & et al. Lifetime prevalence of psychotic and bipolar I disorders in a general population. *Archives of General Psychiatry* **64**, 19-28 (2007).

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

11. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236-1241 (2015).

12. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature genetics* **43**, 969-976 (2011).

13. Tansey, K.E. *et al.* Common alleles contribute to schizophrenia in CNV carriers. *Molecular Psychiatry* (2015).

14. Dudbridge, F. Polygenic Epidemiology. *Genetic Epidemiology* **40**, 268-272 (2016).

15. Lee, S.H., Goddard, M.E., Wray, N.R. & Visscher, P.M. A better coefficient of determination for genetic profile analysis. *Genetic epidemiology* **36**, 214-224 (2012).

16. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015).

17. Maston, G.A., Evans, S.K. & Green, M.R. Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics* **7**, 29-59 (2006).

18. The Network and Pathway Analysis Subgroup of the Psychiatric Genomics Consortium. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience* **18**, 199-209 (2015).

19. Cox, D.D. & Lee, J.S. Pointwise testing with functional data using the Westfall–Young randomization method. *Biometrika* **95**, 621-634 (2008).

20. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

21. Exome Aggregation Consortium. ExAC browser. (accessed: 04/02/2016).

22. Batada, N.N., Hurst, L.D. & Tyers, M. Evolutionary and Physiological Importance of Hub Proteins. *PLoS Comput Biol* **2**, e88 (2006).

23. Parikshak, N.N., Gandal, M.J. & Geschwind, D.H. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nat Rev Genet* **16**, 441-458 (2015).

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

25. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).

26. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-190 (2014).

27. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-61 (2011).

28. Pocklington, A.J., Cumiskey, M., Armstrong, J.D. & Grant, S.G.N. The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. *Molecular Systems Biology* **2**, 2006.0023 (2006).

29. Fernández, E. *et al.* Targeted tandem affinity purification of PSD‐95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Molecular systems biology* **5**, 269 (2009).

30. Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* **17**, 142-153 (2012).

31. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A. & Richardson, J.E. The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Research* **42**, D810-7 (2014).

32. Pers, T.H. *et al.* Comprehensive analysis of schizophrenia-associated loci highlights ion channel pathways and biologically plausible candidate causal genes. *Human Molecular Genetics* **25**, 1247-1254 (2016).

33. Consortium, T.G.O. Gene Ontology Consortium: going forward. *Nucleic Acids Research* **43**, D1049-D1056 (2015).

34. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Research* **44**, D481-D487 (2016).

35. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457-D462 (2016).

36. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43**, D789-D798 (2015).

37. de Leeuw, C.A., Neale, B.M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat Rev Genet* **17**, 353–364 (2016).

38. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228-1235 (2015).

39. O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733-D745 (2016).

40. van Dongen, J. & Boomsma, D.I. The evolutionary paradox and the missing heritability of schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **162**, 122-136 (2013).

41. Srinivasan, S. *et al.* Genetic Markers of Human Evolution Are Enriched in Schizophrenia. *Biological Psychiatry* **80**, 284–292 (2016).

42. Xu, K., Schadt, E.E., Pollard, K.S., Roussos, P. & Dudley, J.T. Genomic and Network Patterns of Schizophrenia Genetic Variation in Human Evolutionary Accelerated Regions. *Molecular Biology and Evolution* **32**, 1148-1160 (2015).

43. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A Map of Recent Positive Selection in the Human Genome. *PLoS Biol* **4**, e72 (2006).

44. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-918 (2007).

45. Grossman, Sharon R. *et al.* Identifying Recent Adaptations in Large-Scale Genomic Data. *Cell* **152**, 703-713 (2013).

46. Huber, C.D., DeGiorgio, M., Hellmann, I. & Nielsen, R. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular Ecology* **25**, 142-156 (2016).

47. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. *PLoS Genet* **5**, e1000471 (2009).

48. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-357 (2014).

49. Vitti, J.J., Grossman, S.R. & Sabeti, P.C. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics* **47**, 97-120 (2013).

50. Sabeti, P.C. *et al.* Positive Natural Selection in the Human Lineage. *Science* **312**, 1614-1620 (2006).

51. Ronen, R., Udpa, N., Halperin, E. & Bafna, V. Learning Natural Selection from the Site Frequency Spectrum. *Genetics* **195**, 181-193 (2013).

52. Nordborg, M., Charlesworth, B. & Charlesworth, D. The effect of recombination on background selection. *Genetics Research* **67**, 159-174 (1996).

53. Charlesworth, B., Betancourt, A.J., Kaiser, V.B. & Gordo, I. Genetic Recombination and Molecular Evolution. *Cold Spring Harbor Symposia on Quantitative Biology* **74**, 177-186 (2009).

54. Fu, W. & Akey, J.M. Selection and Adaptation in the Human Genome. *Annual Review of Genomics and Human Genetics* **14**, 467-489 (2013).

55. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-1007 (2014).

56. Cunningham, F. *et al.* Ensembl 2015. *Nucleic acids research* **43**, D662-D669 (2015).

57. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).

58. Maclean, C.A., Chue Hong, N.P. & Prendergast, J.G.D. hapbin: An Efficient Program for Performing Haplotype-Based Scans for Positive Selection in Large Genomic Datasets. *Molecular Biology and Evolution* **32**, 3027-3029 (2015).

59. Charlesworth, B. The Effects of Deleterious Mutations on Evolution at Linked Sites. *Genetics* **190**, 5-22 (2012).

60. Hussin, J.G. *et al.* Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nature Genetics* **47**, 400-404 (2015).

61. Ptok, U., Barkow, K. & Heun, R. Fertility and number of children in patients with Alzheimer's disease. *Archives of Women's Mental Health* **5**, 83-86 (2002).

62. Whitworth, K.W., Baird, D.D., Stene, L.C., Skjaerven, R. & Longnecker, M.P. Fecundability among women with type 1 and type 2 diabetes in the Norwegian Mother and Child Cohort Study. *Diabetologia* **54**, 516-522 (2011).

63. Jokela, M. Birth-cohort effects in the association between personality and fertility. *Psychological science* **23**, 835-841 (2012).

64. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).

65. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45**, 1452-1458 (2013).

66. Smith, D.J. *et al.* Genome-wide analysis of over 106,000 individuals identifies 9 neuroticism-associated loci. *Molecular Psychiatry* **21**, 749-757 (2016).

67. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics* **46**, 234-244 (2014).

68. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).

69. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).

70. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164-e164 (2010).

71. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527 (2016).

72. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).

73. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**, 481-487 (2016).

Supplementary Table	Description
Table 1	Sign tests for concordance of datasets used in the present study with the prev
Table 2	Polygenic risk score profile of datasets used in the present study.
Table 3	Independent genome-wide significant association signals from the CLOZUK + I
Table 4	Independent genome-wide significant association signals from the CLOZUK + I
Table 5	Independent genome-wide significant association signals from the PGC study,
Table 6	Replication of 50 newly discovered loci in a combined independent sample (5,
Table 7	Partitioned heritability analysis of broad genomic categories.
Table 8	Partitioned heritability analysis of background selection in schizophrenia and c
Table 9	MAGMA analysis results of the CNS-related gene set collection.
Table 10	MAGMA analysis results of the data-driven gene set collection.
Table 11	List of annotated FINEMAP SNPs with cumulative probability 95%.
Table 12	Hi-C and Functional Characterisation of FINEMAPPED SNPs with Posterior Prol
Table 13	SMR of CLOZUK+PGC2 data in the CommonMind Consortium DLPFC Brain eQT
Table 14	Independent genome-wide significant association signals from the main CLOZ
Table 15	Liability-scale SNP-based h2 of schizophrenia in the datasets used in the prese

vious PGC SZ GWAS.

PGC meta-analysis, clumped.
PGC meta-analysis, clumped and amalgamated into loci.
and their corresponding statistics in the CLOZUK + PGC meta-analysis.
,662 cases and 154,224 controls)

other phenotypes.

bability >50%.
TL data.
UK GWAS, clumped and amalgamated into loci (see Methods).
nt study, for a range of population prevalences of the disorder.

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, $p(1-p)/2N$). Sample *q* from this distribution and calculate (*N*1, *N*2), the corresponding number of *A* and *a* alleles in the controls (=$2Nq$, $2N(1-q)$ 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*) = $2Ms$, $2M(1-s)$.

5. Finally, use *N1*, *N2*, *M1*, *M2* to calculate the observed effect size $\beta_S=\ln(M1*N2/M2*N1)$ and its variance $\sigma_S^2=(1/N1)+(1/N2)+(1/M1)+(1/M2)$

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,…,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

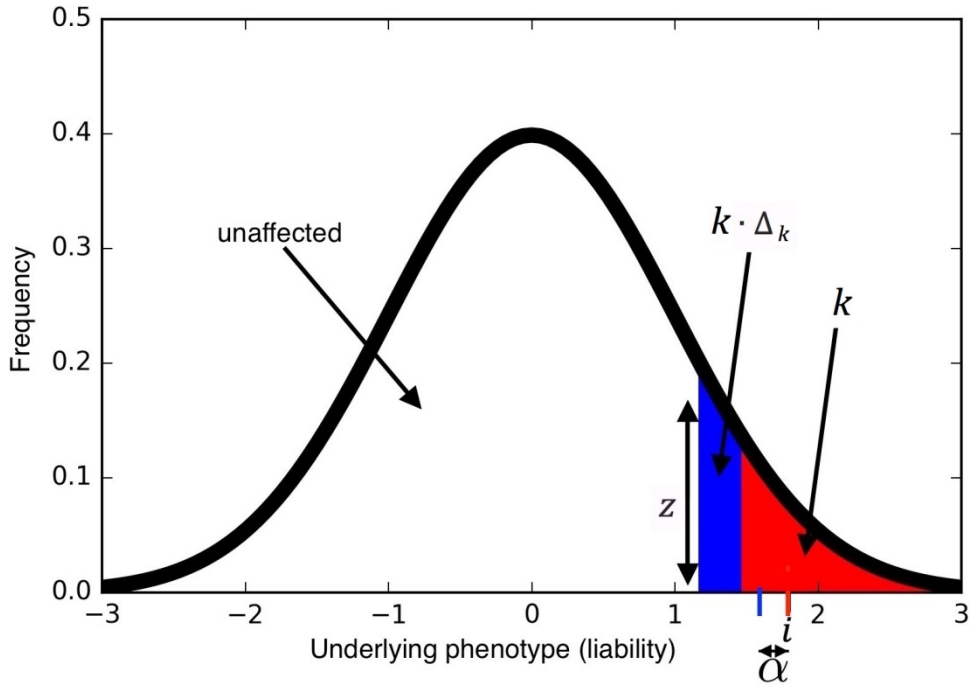
corresponding heritability is slightly smaller than twice the haploid heritability (by a factor 1/2*n*) because haploid effects are negatively correlated within diploids due to sampling:

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

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:

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

[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.

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 5×10^{-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 (Δ_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

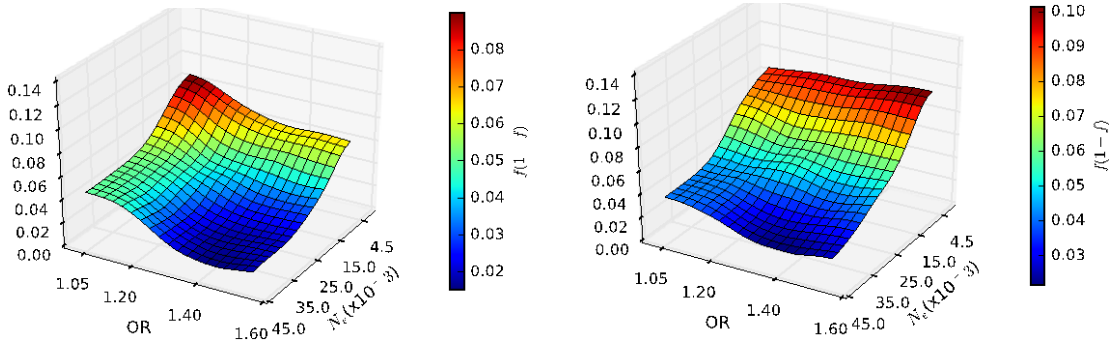
For testing the feasibility of detecting causal alleles in regions under BGS, pairs of causal-neutral biallelic loci were simulated for ten N_e values evenly distributed from 4500 to 45000 diploid individuals. This range of population sizes accounts for the estimations of the effective size of human 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].

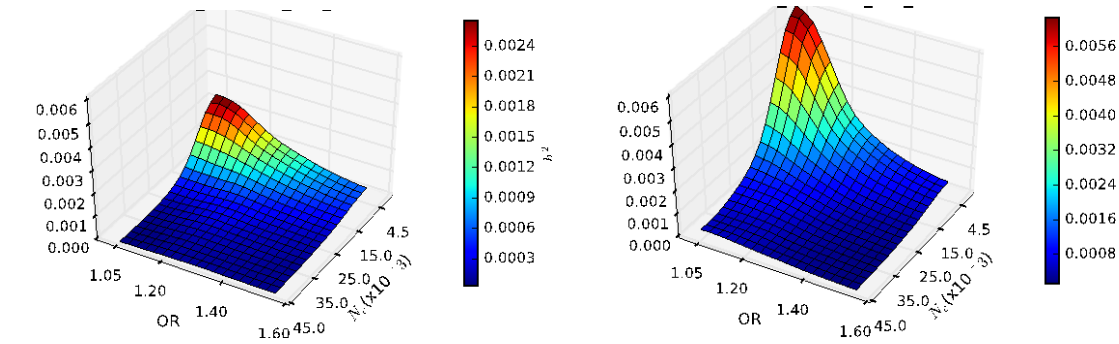
Each combination of parameters was run for 10^8 generations. Every time an allele was lost, the same allele was reintroduced in the population as a single copy, but the results were re-scaled proportionally to N_e. Thus, the number of mutation events was proportional to N_e and the rate of mutation was 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 $\chi^2 > 29.72$ (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.

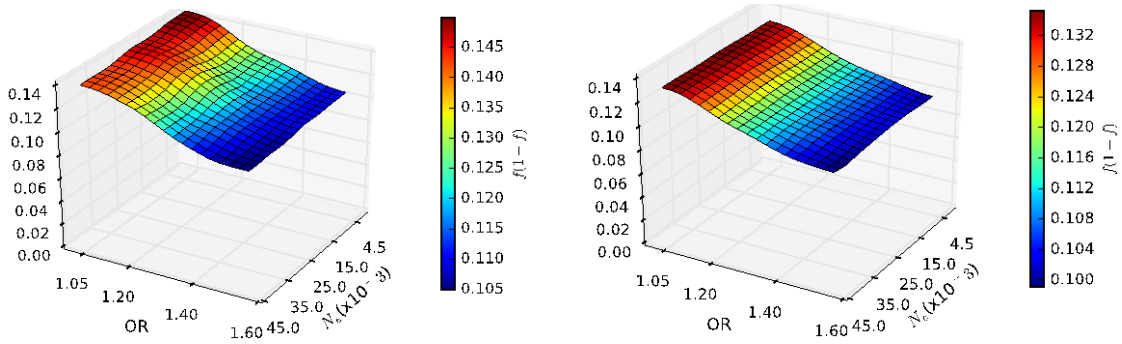
An additional block of simulations of 10 x 10 combinations of N_e and α values was carried out for the same combination of parameters, but this time for a neutral trait (*Rf* = 0). The results show a completelly different pattern: First, the product $f(1 - f)$ does not change with N_e (Supplementary 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.



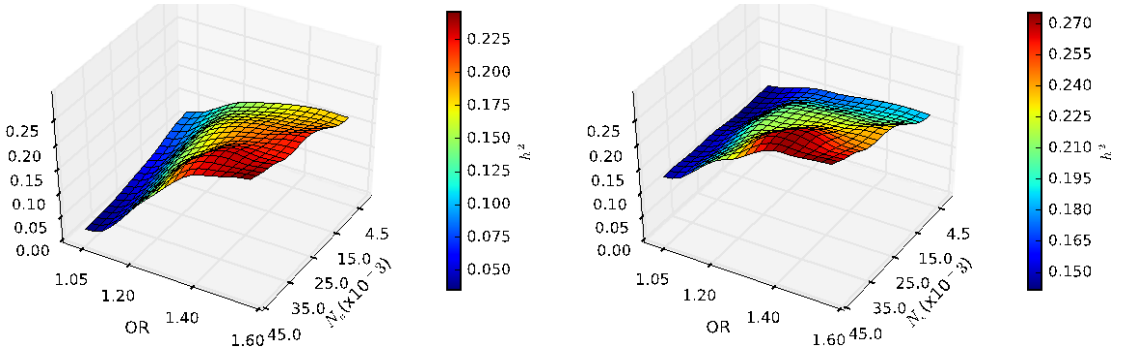
SUPPLEMENTARY NOTE FIGURE 1. Average product $f(1 - f)$ of allele frequencies at neutral SNPs significantly associated with effects on schizophrenia (vertical axis and colour scale). Two sample sizes are shown: *n*=30,000 (left) and *n*=100,000 (right). The surface is given as a function of N_e and the 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.

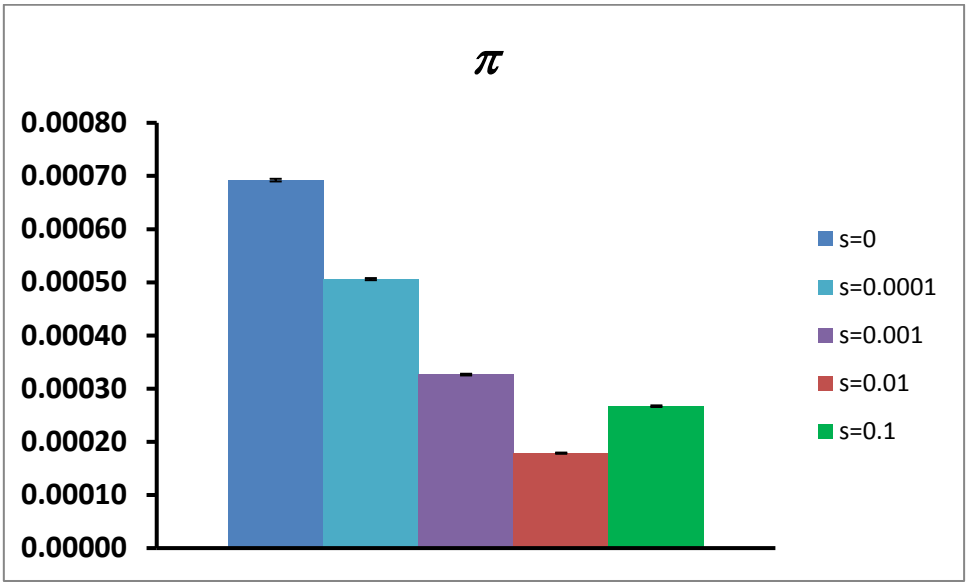
Mutations were assumed to appear at random along the sequence such that 74% of mutations were neutral, 24% were assumed to be deleterious for fitness with a homozygous effect obtained from an exponential distribution with mean s. Five different scenarios were considered using a range of mean values of 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).

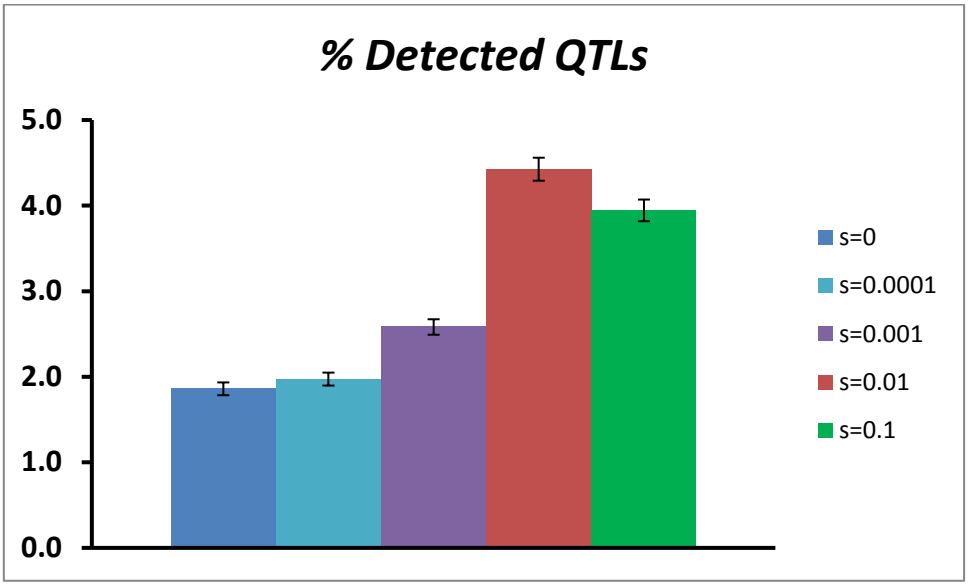
In order to compare the probability of detection of causal SNPs under different levels of BGS, the top twenty SNPs with the lowest probability from the GWAS analysis were considered, and the number of true causal QTLs in these 20 SNPs was recorded. To quantify the magnitude of the genomic reduction in 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).

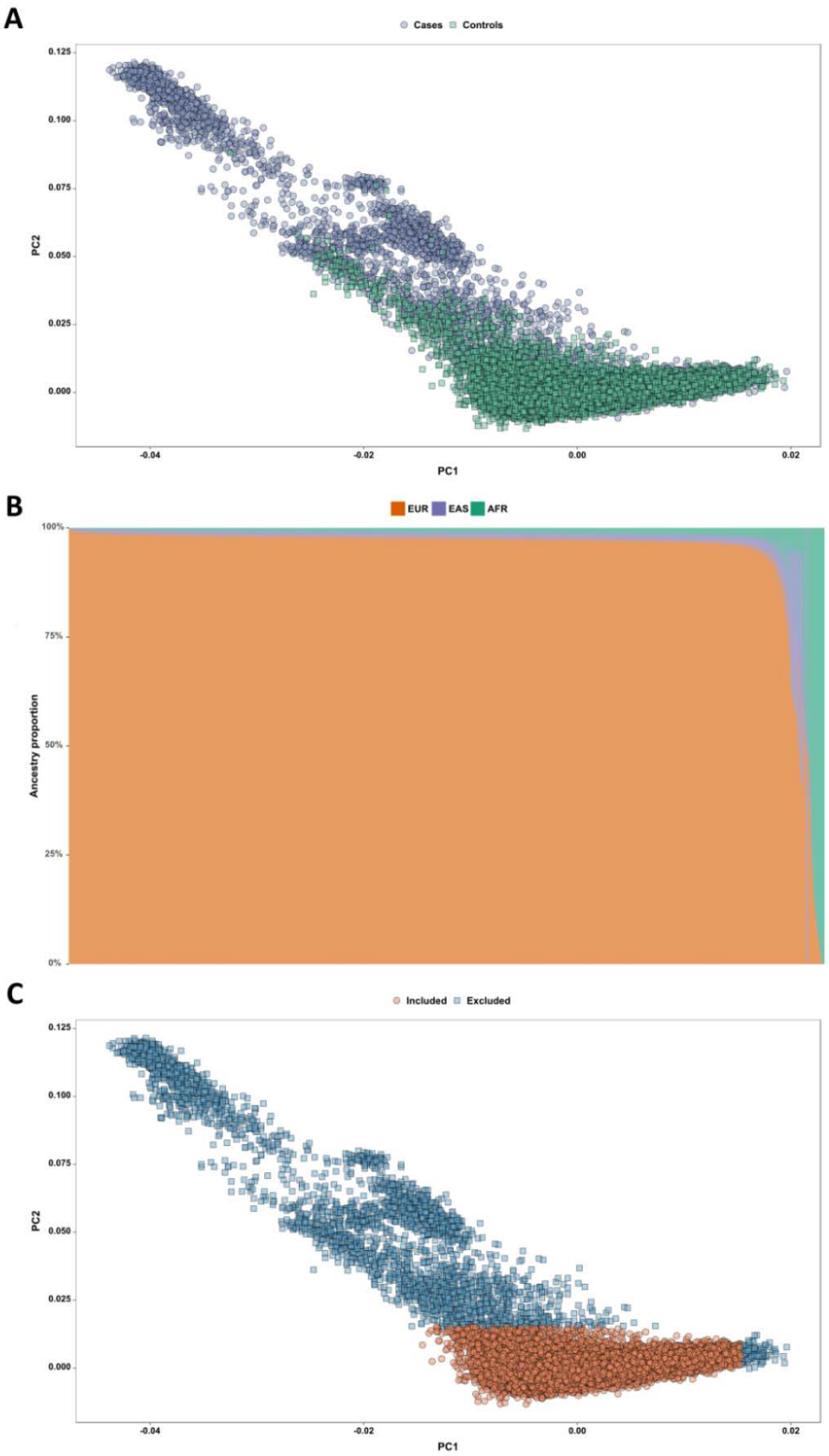
54. Steinberg, S. *et al.* Common variant at 16p11. 2 conferring risk of psychosis. *Molecular psychiatry* **19**, 108-114 (2014).

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).

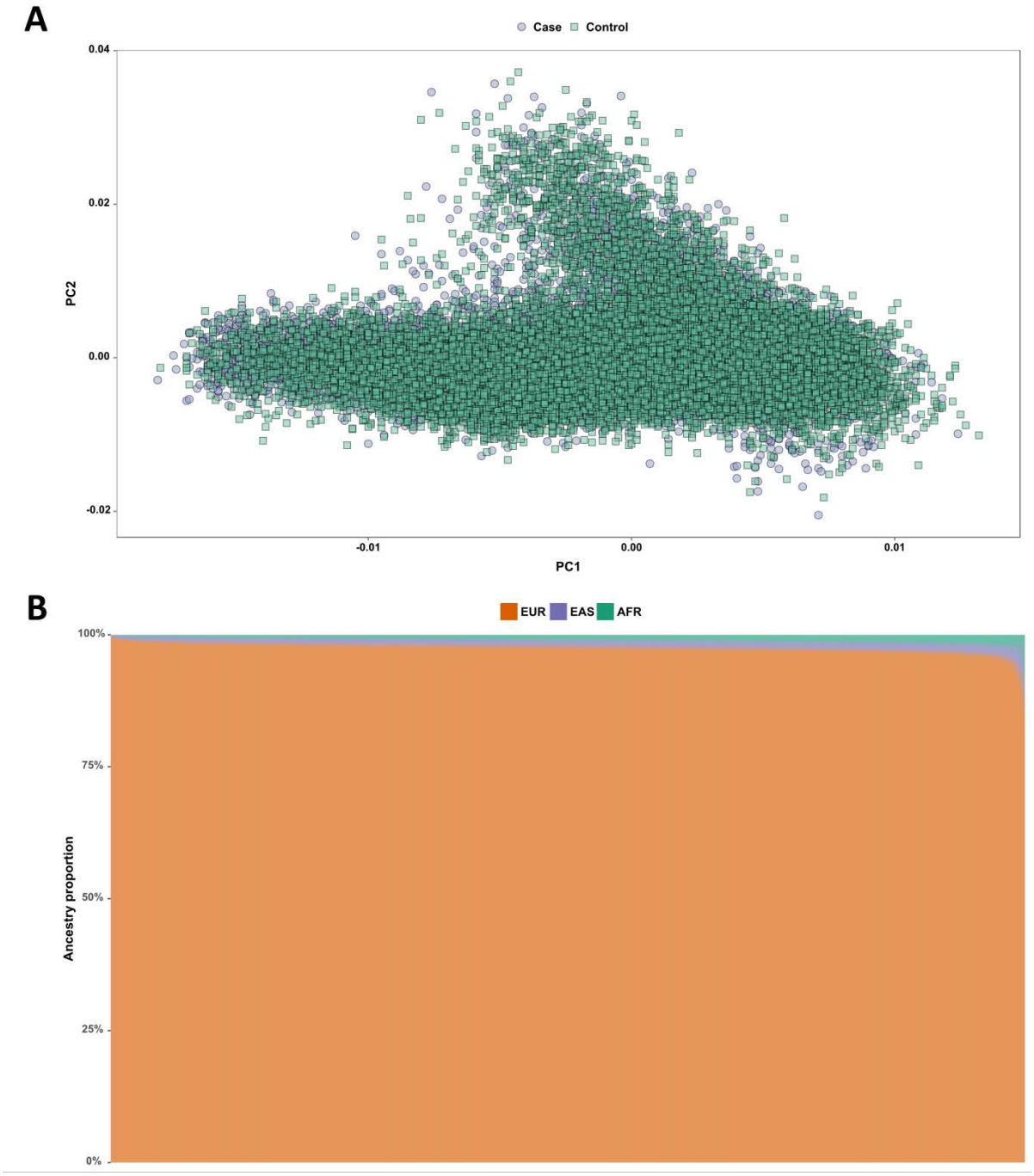
57. Hollegaard, M.V. *et al.* Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source. *BMC Genet* **12**, 58 (2011).

Supplementary Figure 1



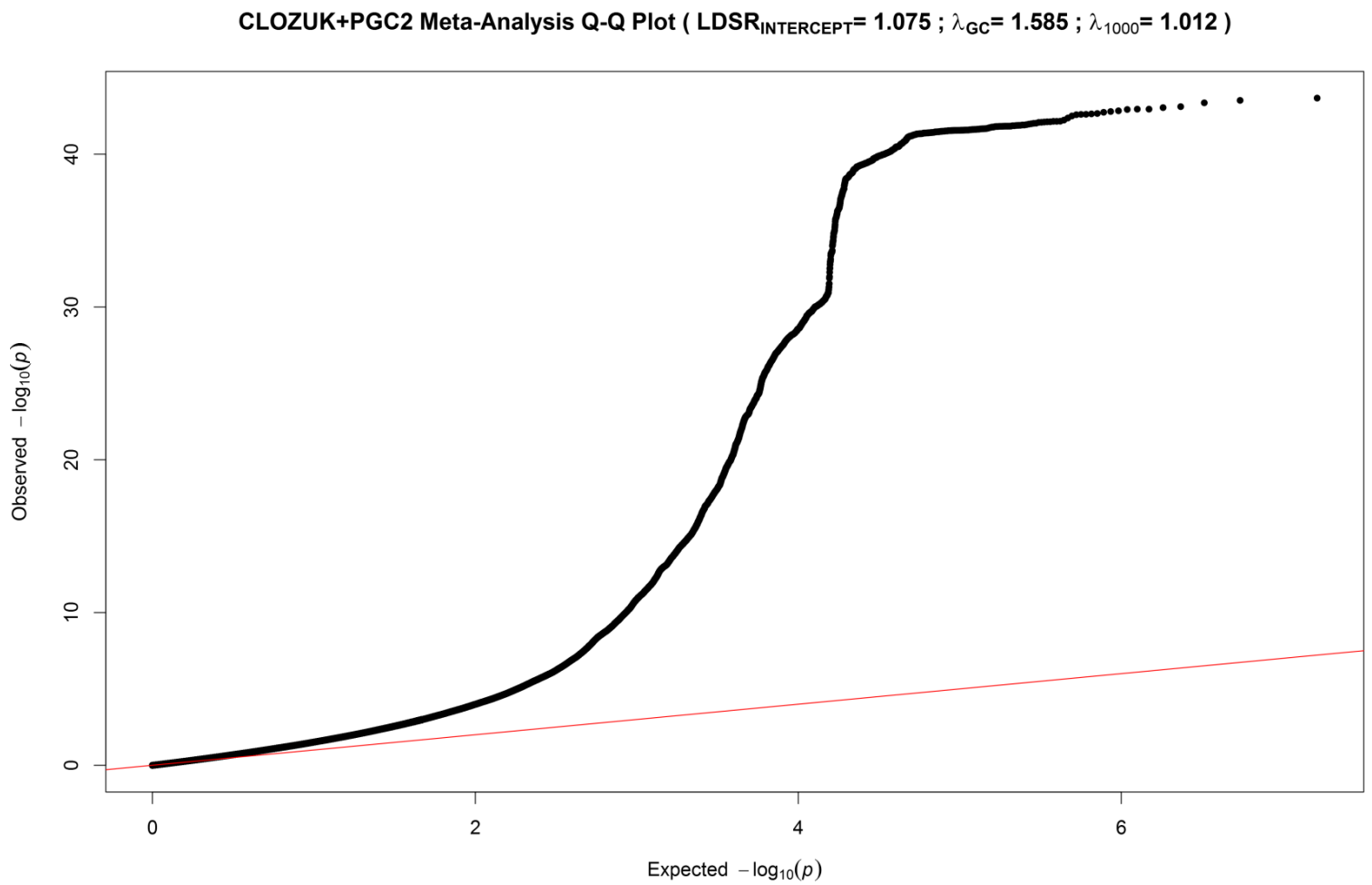
Population structure of the complete CLOZUK dataset. A: PCA showing cases and controls, notice the large spread in the cases. B: ADMIXTURE plot (K=3), names of ancestral components represent the most similar 1KGPp3 superpopulation. C: PCA showing the individuals finally selected for the GWAS.

Supplementary Figure 2



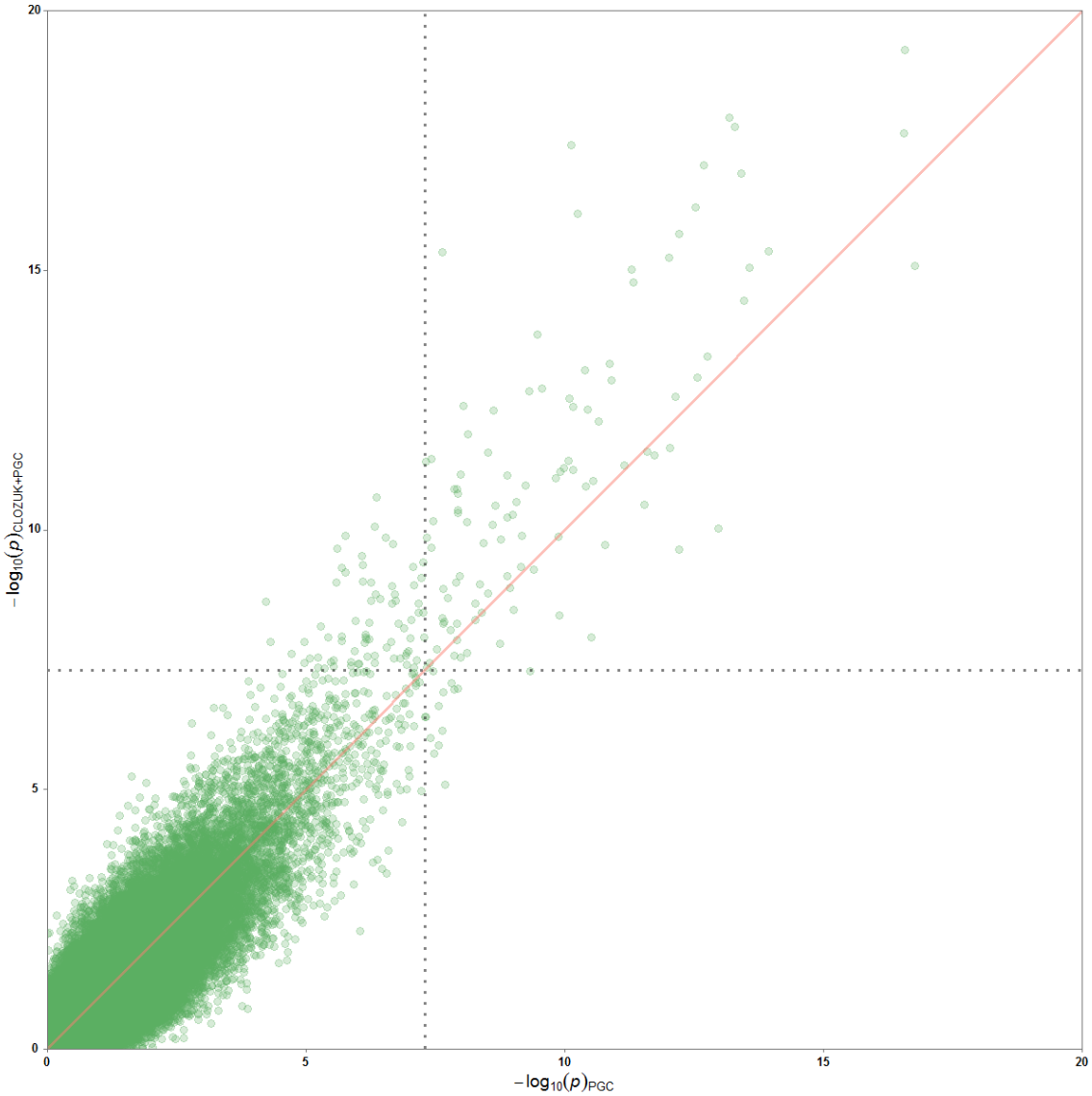
Population structure of the CLOZUK subset selected for the GWAS. A: PCA showing cases and controls, notice the profiles are almost completely overlapping. B: ADMIXTURE plot (K=3), names of ancestral components represent the most similar 1KGPp3 superpopulation.

Supplementary Figure 3



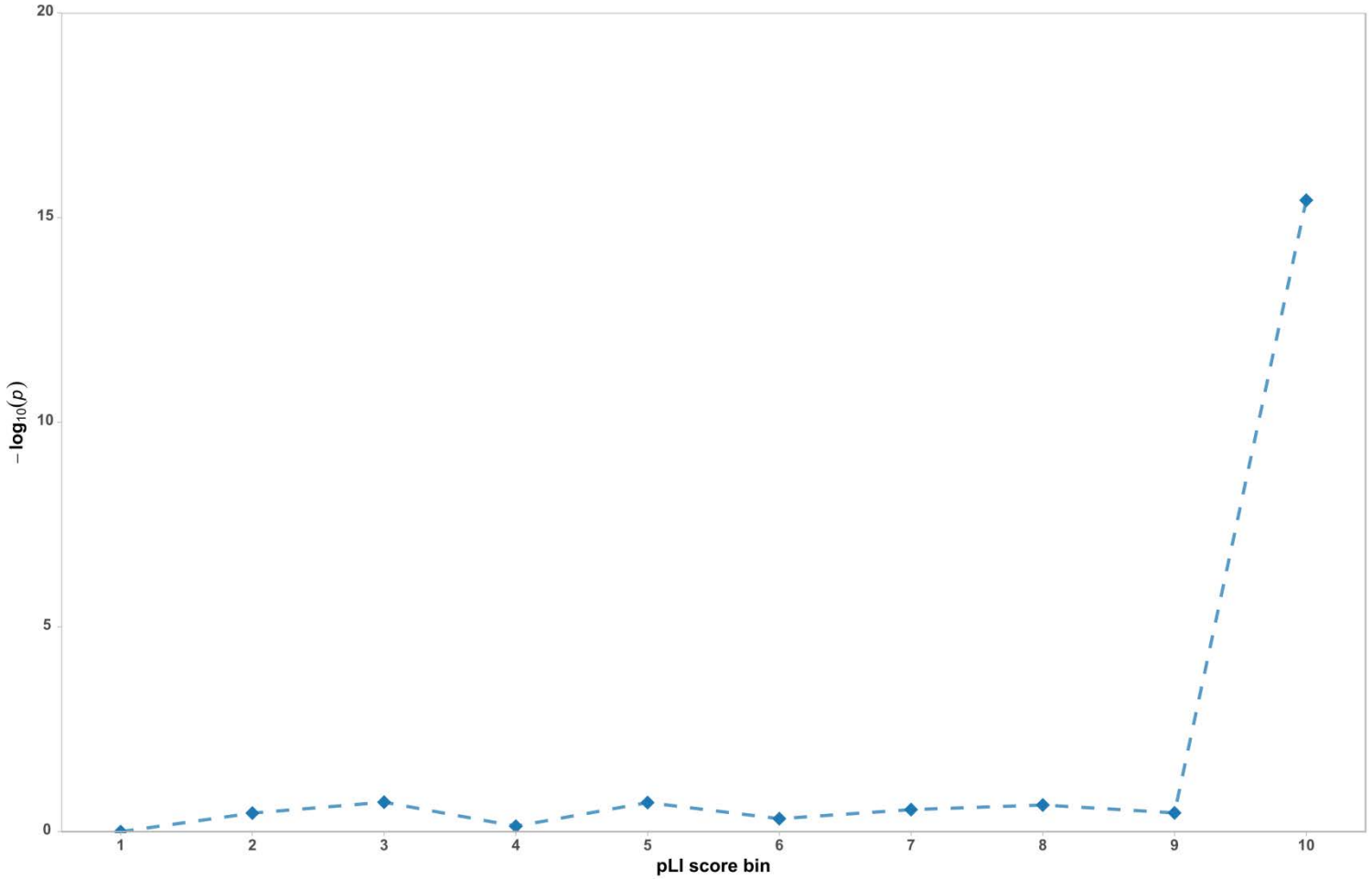
QQ plot of CLOZUK and PGC2 meta-analysis.

Supplementary Figure 4



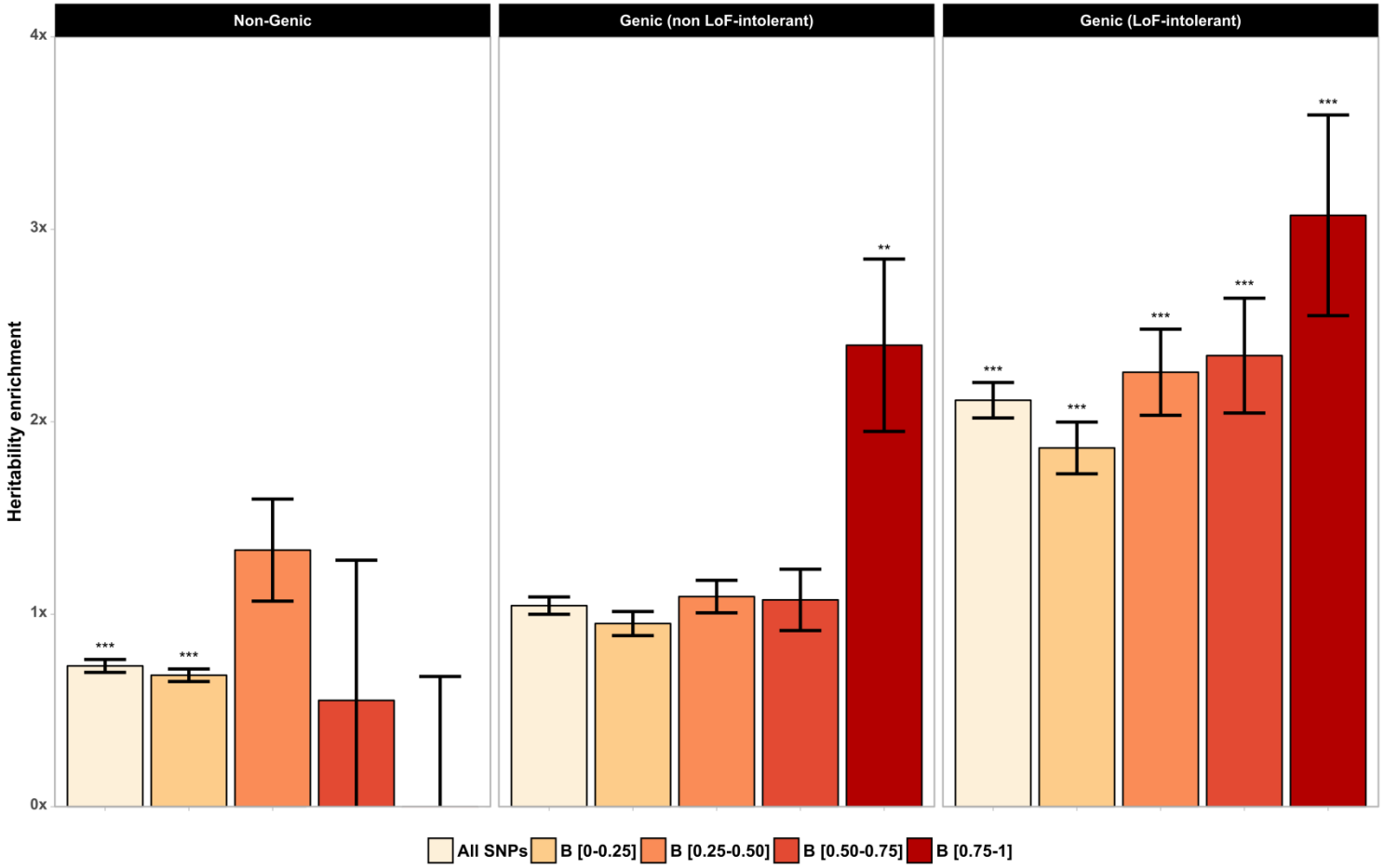
Index SNP p-values for all clumps in the meta-analysis (CLOZUK+PGC) compared with PGC. Dotted lines show the genome-wide significant threshold for the two datasets. The red line indicates a null of equal p-values in both datasets, and thus index SNPs to the left of this line (toward y axis) show increased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

Supplementary Figure 5



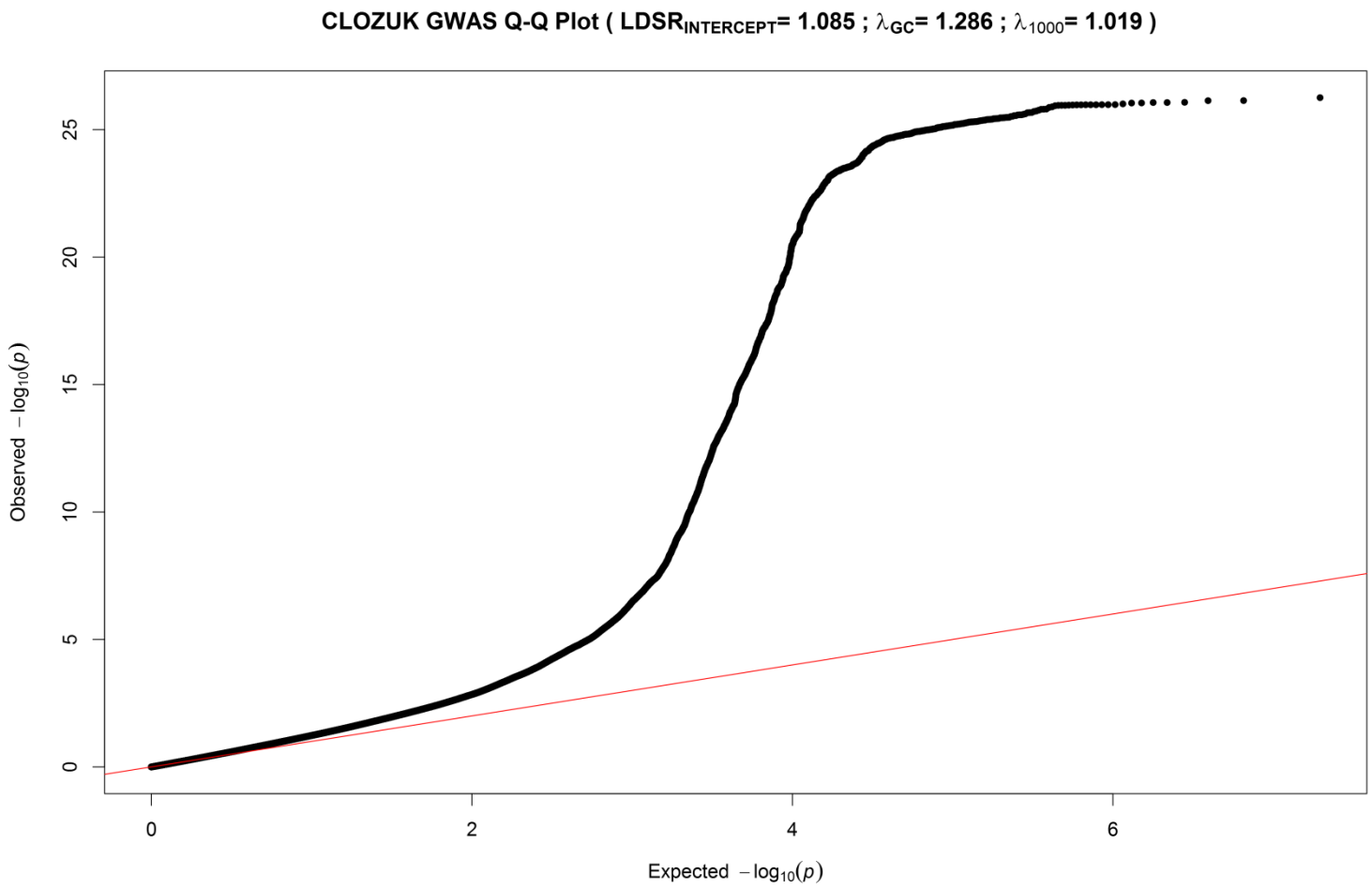
Schizophrenia association for genes within bins of pLI, an ExAC-based measure of intolerance to functional sequence variation. Bins are based on increasing 0.1 intervals of the statistic, and thus all LoF-intolerant genes (defined as pLI > 0.9) are in bin 10. P-values correspond to the statistical significance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

Supplementary Figure 6



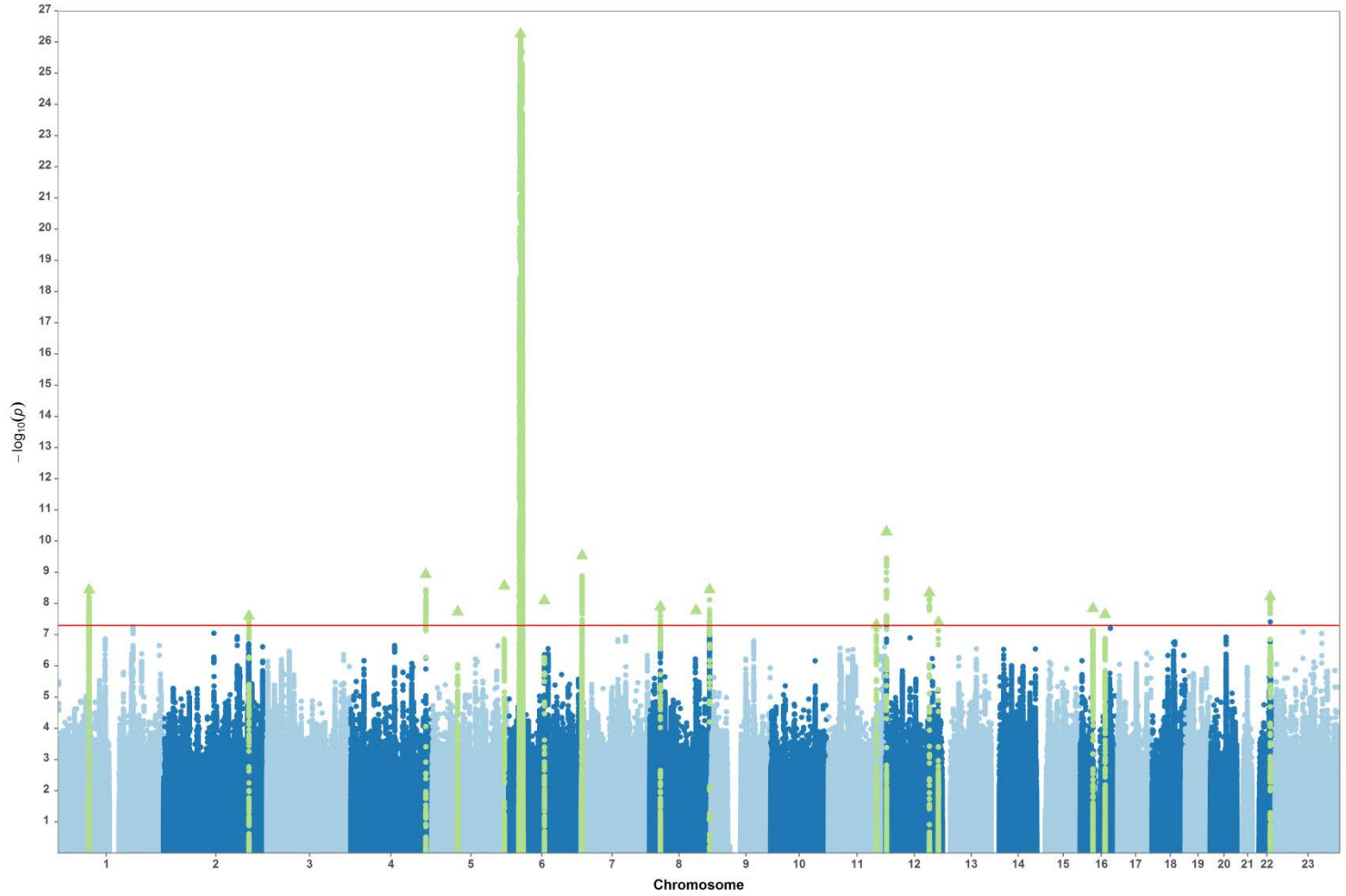
Schizophrenia SNP-based heritability enrichment, as estimated by LDSR, is influenced by the intensity of background selection ("B") and the genomic location. Error bars indicate enrichment standard errors. Asterisks indicate the significance of enrichment for each group of SNPs (* <= 0.05; ** <= 0.01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

Supplementary Figure 7

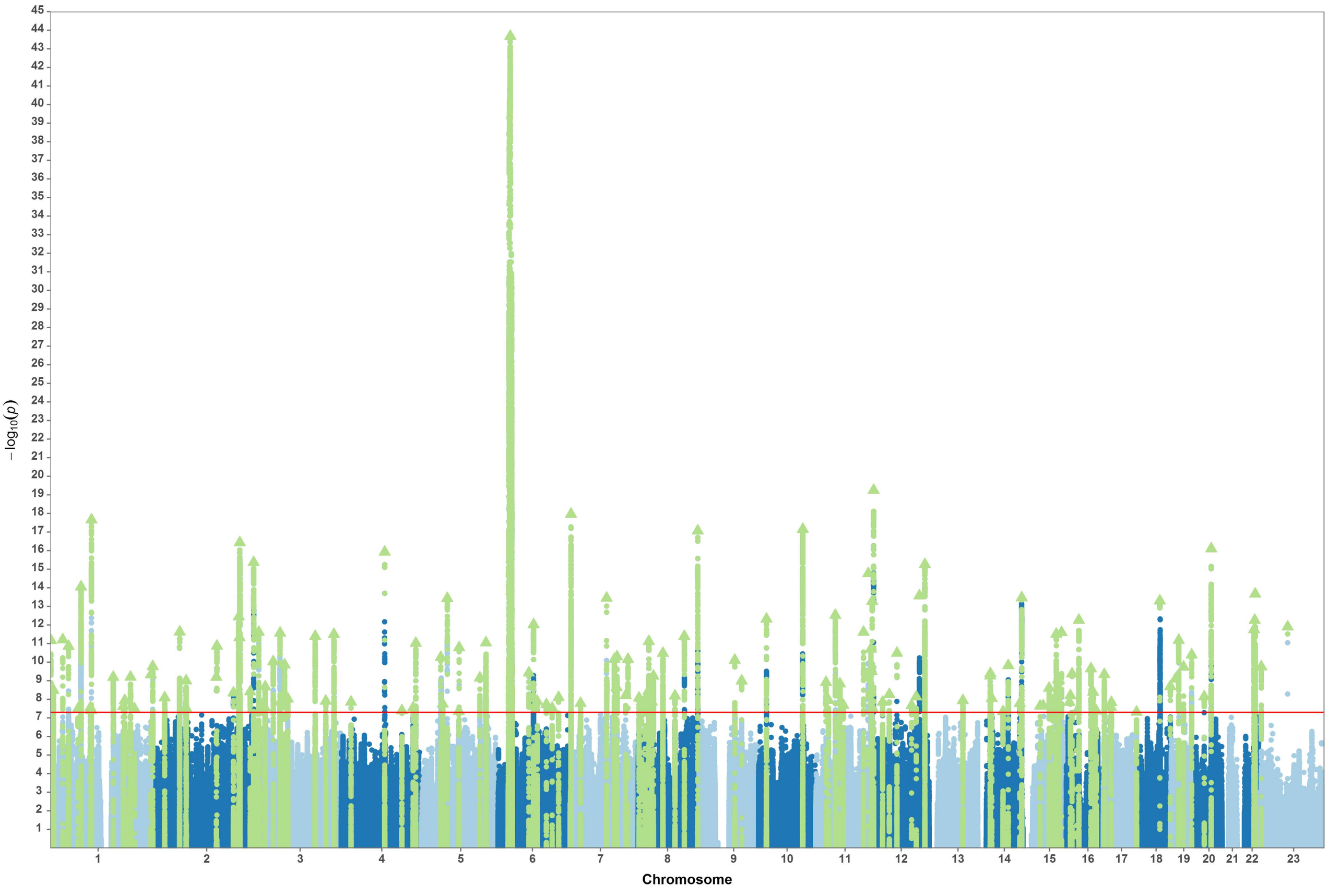


QQ Plot of the CLOZUK GWAS.

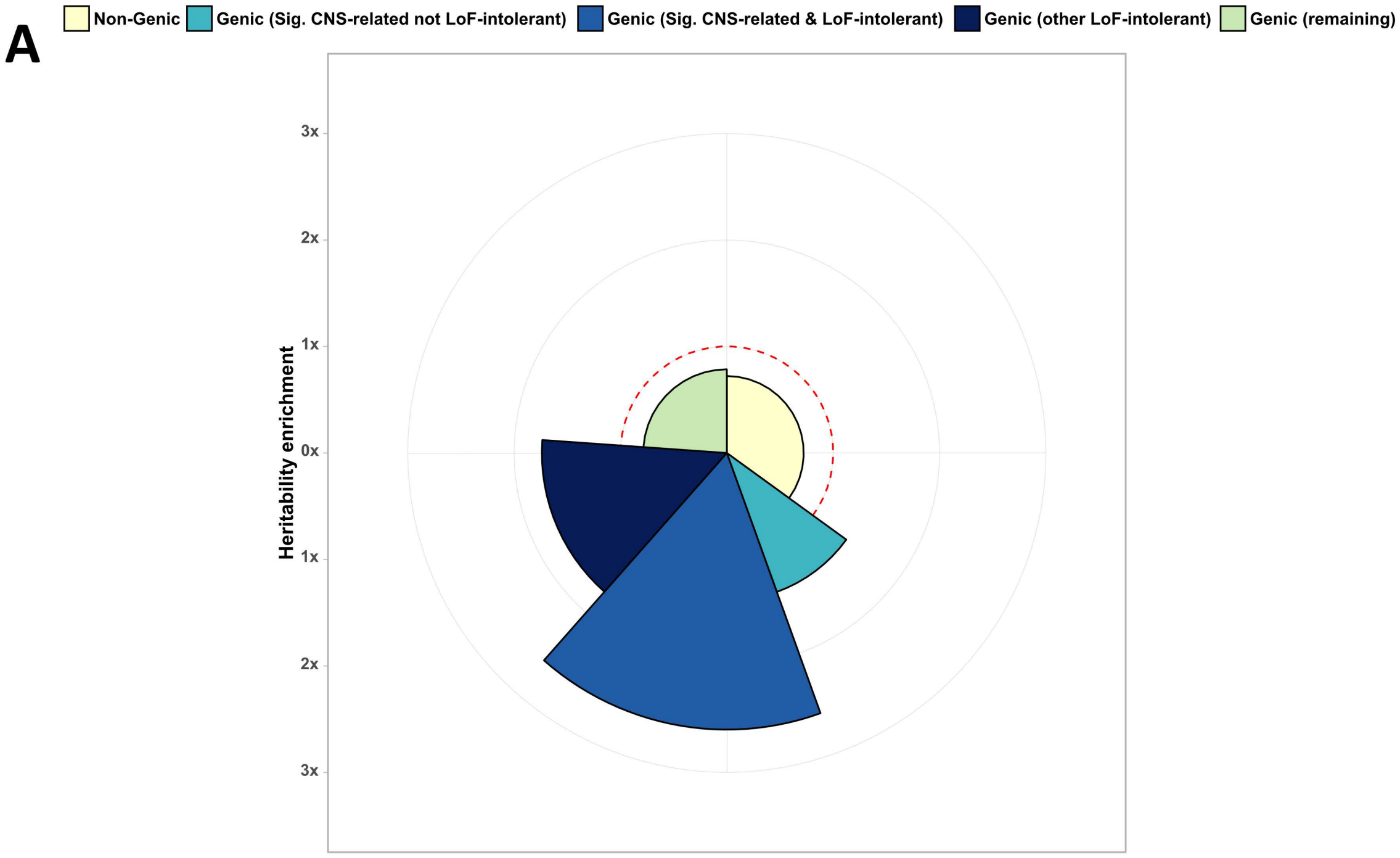
Supplementary Figure 8



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


$-\log_{10}(p)$
45
44
43
42
41
40
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Chromosome

A


Non-Genic
Genic (Sig. CNS-related not LoF-intolerant)
Genic (Sig. CNS-related & LoF-intolerant)
Genic (other LoF-intolerant)
Genic (remaining)
Heritability enrichment
3x
2x
1x
0x
1x
2x
3x

B

ALL GENES
Not LoF-intolerant
LoF-intolerant
Proportion of genic heritability explained
25%
20%
15%
10%
5%
0%

ALL GENES
Not LoF-intolerant
LoF-intolerant
Heritability enrichment
15x
10x
5x
0x
***
***
**
***
**
*
***
***
***
*
***

Abnormal behaviour (n=1939)
FMRP-targets (n=798)
Voltage-gated Ca+ channels (n=196)
Abnormal nervous system electrophysiology (n=201)
Abnormal long-term potentiation (n=142)
5HT-2C (n=16)