Miscellaneous

New data and an old puzzle: the negative association between schizophrenia and rheumatoid arthritis

S Hong Lee,1,2 Enda M Byrne,1 Christina M Hultman,3 Anna Kähler,3 Anna AE Vinkhuyzen,1 Stephan Ripke,4,5,6 Ole A Andreassen,7,8 Thomas Frisell,9 Alexander Gusev,10,11,12 Xinli Hu,13,14,15,16 Robert Karlsson,3 Vasilis X Mantzioris,1 John J McGrath,1,17 Divya Mehta,1 Eli A Stahl,18 Qiongyi Zhao,1 Kenneth S Kendler,19,20,21 Patrick F Sullivan,22 Alkes L Price,10,11,12 Michael O’Donovan,23,24 Yukinori Okada,25,26,27,28 Bryan J Mowry,1,29 Soumya Raychaudhuri25,26,27,30,31 and Naomi R Wray1*; Schizophrenia Working Group of the Psychiatric Genomics Consortium and Rheumatoid Arthritis Consortium International

1The University of Queensland, Queensland Brain Institute, Brisbane, QLD, Australia, 2School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia, 3Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, 4Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA, 5Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA, 6Department of Psychiatry and Psychotherapy, Charité-Universitätsmedizin Berlin, Campus Mitte, Berlin, Germany, 7NORメント KG Jebens Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, 8Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway, 9Clinical Epidemiology Unit, Karolinska Institutet, Stockholm, Sweden, 10Department of Epidemiology, 11Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA, 12Medical and Population Genetics Program, Broad Institute, Cambridge, MA, USA, 13Division of Genetics, 14Division of Rheumatology, Immunology, and Allergy, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA, 15Medical and Population Genetics Group, Broad Institute, Cambridge, MA, USA, 16Health Science and Technology MD Program, Harvard University and Massachusetts Institute of Technology, Boston, MA, USA, 17Queensland Centre for Mental Health Research, Park Centre for Mental Health, Richlands, QLD, Australia, 18Division of Psychiatric Genomics, Mt Sinai School of Medicine, New York, NY, USA, 19Virginia Institute of Psychiatric and Behavioral Genetics, 20Department of Human and Molecular Genetics, 21Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA, 22Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 23Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics, 24Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University School of Medicine, Cardiff, UK, 25Division of Rheumatology, Immunology, and Allergy, 26Division of Genetics, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA, USA, 27Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA, 28Department of Human Genetics and Disease Diversity, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan, 29Queensland Centre for Mental Health Research, Toowoomba, QLD, Australia.
Abstract

Background: A long-standing epidemiological puzzle is the reduced rate of rheumatoid arthritis (RA) in those with schizophrenia (SZ) and vice versa. Traditional epidemiological approaches to determine if this negative association is underpinned by genetic factors would test for reduced rates of one disorder in relatives of the other, but sufficiently powered data sets are difficult to achieve. The genomics era presents an alternative paradigm for investigating the genetic relationship between two uncommon disorders.

Methods: We use genome-wide common single nucleotide polymorphism (SNP) data from independently collected SZ and RA case-control cohorts to estimate the SNP correlation between the disorders. We test a genotype X environment (GxE) hypothesis for SZ with environment defined as winter- vs summer-born.

Results: We estimate a small but significant negative SNP-genetic correlation between SZ and RA (−0.046, s.e. 0.026, \( P = 0.036 \)). The negative correlation was stronger for the SNP set attributed to coding or regulatory regions (−0.174, s.e. 0.071, \( P = 0.0075 \)). Our analyses led us to hypothesize a gene-environment interaction for SZ in the form of immune challenge. We used month of birth as a proxy for environmental immune challenge and estimated the genetic correlation between winter-born and non-winter born SZ to be significantly less than 1 for coding/regulatory region SNPs (0.56, s.e. 0.14, \( P = 0.00090 \)).

Conclusions: Our results are consistent with epidemiological observations of a negative relationship between SZ and RA reflecting, at least in part, genetic factors. Results of the month of birth analysis are consistent with pleiotropic effects of genetic variants dependent on environmental context.

Key words: Schizophrenia, rheumatoid arthritis, genetic relationship, pleiotropy
factor for RA in general population samples (OR > 2). Both disorders have similar lifetime risk (~1%), a waxing and waning pattern of symptoms and increased mortality. On the other hand, there are also dissimilarities including age at onset (16–30 years in SZ vs 25–55 years in RA) and sex:female ratio (nearby 3 females to 1 male for RA and 1 female to 1.4 males for SZ). Sex differences were not considered in early studies of the SZ-RA relationship, but recent population-based studies from Sweden and Denmark accounted for age and sex differences and still reported reduced risks of RA in SZ compared with those without SZ.

Both SZ and RA have a strong genetic component to their aetiology. For both, heritabilities estimated from national hospital records (SZ 64% and RA 40%) are lower than estimates from twin studies (SZ 81% and RA 60%). RA has two distinct subtypes classified on the presence (seropositive) or absence (seronegative) of antibodies to citrullinated protein antigen. Approximately two-thirds of cases are seropositive, and of seronegative cases 4–11% have been estimated to reflect misdiagnosis (for example of ankylosing spondylitis) and 15–37% have been estimated to be undetected seropositive. Using national data from Sweden, the heritability of seropositive RA was estimated as ~50% compared with ~20% for seronegative RA.

A number of hypotheses have been proposed to explain the SZ-RA protective relationship, including abnormal tryptophan metabolism, prostaglandin deficiency, an imbalance in corticosteroids, psychosocial factors or consequence of medication. Definitive evidence to support these hypotheses is lacking. There is evidence that factors influencing immune activation, including environmental insults such as infectious agents, are potential pathogenic mechanisms for both disorders. For example, both RA and SZ have been linked, albeit with some controversy, to increased rates of infection by viruses such as Epstein-Barr virus and the parasite Toxoplasma gondii (for a review see). SZ is considered to be a neurodevelopmental disorder, and immune activation in early life may be of particular importance, consistent with perinatal risk factors such as infection and month of birth. RA is an autoimmune disorder, and it has been suggested that there is an autoimmune component to SZ. The autoimmune theory of SZ is supported by epidemiological evidence showing that whereas the relationship between SZ and RA is a negative one, the risk of many other autoimmune disorders is higher in SZ than in controls. An analysis of Danish national records showed a dose-response relationship between risk of SZ and hospitalizations for infection and autoimmune disorders, where three or more infections and an autoimmune disease were associated with an incidence rate ratio of 3.40 [95% confidence interval (CI) 2.91–3.94]. As for all autoimmune disorders, the major histocompatibility complex (MHC) plays an important role in RA but with different alleles being associated with seropositive cases compared with seronegative cases. A role for the MHC in the aetiology of SZ has been proposed for decades, but the empirical evidence has been less consistent than for RA. The first large genome-wide association studies (GWAS) for SZ identified the MHC locus as the most strongly associated locus.

Using the latest published GWAS results, the MHC locus is the only locus that reaches genome-wide significance for both SZ and RA. The most associated single nucleotide polymorphism (SNP) for RA is associated with SZ and vice versa; contrary to expectation, given the negative SZ-RA association, the associated alleles are the same for the two disorders, albeit stronger for RA than SZ (Box 1). This unexpected result may reflect the well-recognized complexity of the MHC region. The primary association for rheumatoid arthritis is within HLA-DRB1 in the class II MHC region, and has a large effect relative to the more modest and less clearly localized effect in schizophrenia. It may be the case that the role of HLA-mediated antigen recognition is simply different in the two diseases, playing a dominant role in rheumatoid arthritis and perhaps a more modest or absent role in schizophrenia. If the negative association between RA and SCZ reflects genetic factors, then it may be driven predominantly by non-HLA genetic factors that are related to immune activation rather than antigen recognition. Genome-wide significant variants explain 3.4% of the variance in liability to schizophrenia and 11.4% of the variance in liability of RA (Box 1). Genome-wide polygenic methods have estimated that for SZ, ~23% of the variance in liability is attributable to common SNPs (or SNP heritability), and 14% to 18% for RA excluding the contribution from the MHC region (~5%). These results imply that more associated loci will be identified for each disorder as sample size increases.

Given the substantial genetic contribution to both disorders, can the negative association between SZ and RA be attributed to genetic factors? This can be investigated from traditional epidemiological studies by measuring risk of RA in relatives of those with SZ compared with relatives of control subjects, and vice versa. As RA and SZ are relatively uncommon, very large cohorts of families with multiple family members measured for both disorders are needed, and this has not been achievable through the family study framework from which estimates of heritability are traditionally derived. National databases provide the only viable strategy to explore a genetic relationship through traditional epidemiological methods, but few countries have suitable national recording frameworks.
Box 1. Comparison of GWAS results of SZ\textsuperscript{36} and RA.\textsuperscript{35}

a. MHC locus. Odds ratios for the two most highly associated SNPs for SZ, rs115329265 (aka rs1233578, hg\textsuperscript{19}:chr 6: 28 712 247 bp), RAF = 0.85 and for RA, rs9268839 (aka rs116633882, hg\textsuperscript{19}:chr 6: 32 428 772 bp) RAF = 0.45. Both are located in the MHC region and the LD\textsuperscript{2} of these SNPs is zero. We note that the RA allele tags the HLA DRB-1 allele, but that the SZ allele is not associated with any classical HLA allele, although it is reported\textsuperscript{36} to be in LD\textsuperscript{2} = 0.32 with an eQTL SNP for HLA-A. The association $P$-values are listed above the error bars.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{MHC Locator}
\caption{MHC Locator}
\end{figure}

b. Variance explained by genome-wide significant (GWS) loci, reported as associated at $P < 5 \times 10^{-8}$.

- For SZ 128, statistically independent GWS loci are reported.\textsuperscript{36} Together these explain 3.4\% of the variance in liability (calculated from reported RAF and OR using INDI-V\textsuperscript{71} assuming lifetime risk of 1\%); all SNPs associated with $P < 0.05$ explain 7\% of variance in out of sample prediction.\textsuperscript{36} Of these 128 SNPs, 102 could be matched to GWAS results for RA. The minimum $P$-value in RA of a SZ GWS locus was 0.004.

- For RA, 101 independent GWS loci are reported.\textsuperscript{35} Together these explain 11.4\% of the variance in liability (using INDI-V\textsuperscript{71} assuming lifetime risk of 0.7\%). All SNPs could be matched to GWAS results for SZ. The minimum $P$-value in SZ of an RA GWS locus was 2.06e-05 (see MHC locus above).

c. QQ plots: (i) $P$-values from SZ GWAS for 101 RA GWS loci; (ii) $P$-values from RA GWAS for 102 SZ GWS loci.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{QQ Plots}
\caption{QQ Plots}
\end{figure}
A Danish national study\(^{25}\) compared 7704 persons in Denmark diagnosed with SZ and their parents with a sample of matched subjects and their parents. Contrary to a hypothesis of a negative genetic association, this study found significantly increased rates of RA in parents of those with SZ compared with parents of control subjects. A Swedish national study reported that first-degree relatives of schizophrenia patients were not at reduced risk of RA, but the risk for seronegative RA was significantly decreased in children and siblings of SZ probands (hazard ratio (HR) = 0.13; 95% CI 0.02–0.95, and HR = 0.67; 95% CI 0.50–0.91, respectively).\(^{2}\) These studies assumed that risk in first-degree relatives is only attributable to genetic factors, but sharing of environmental factors could also contribute.

The genomics era provides a new opportunity to investigate whether the SZ-RA relationship may be attributable to genetic factors, by determining whether common alleles conferring increased risk to SZ are protective against RA and vice versa. The hypothesis has been considered for candidate genes outside the MHC region, but no shared associations were found.\(^{42,43}\) Comparison of the latest GWAS results for SZ\(^{36}\) and RA\(^{35}\) shows evidence for more association of the genome-wide significant (GWS) loci from each disease in the other disease than expected by chance (Box 1). Across GWS SNPs, there is a positive correlation between the OR of the two disorders (Box 1), although this relationship is dominated by the positive correlation between SNPs in the MHC region described above.

Here we use linear mixed model methods, applied to genome-wide SNPs from case-control cohorts collected for GWAS, to explore the relationship between SZ and RA.\(^{44}\) By comparing additive genetic similarities between SZ and RA cases with their genetic similarities with controls, we quantify the relationship between the disorders by the SNP-genetic correlation.\(^{44,45}\) Since this approach uses unrelated cases and controls to estimate the SNP correlation, estimates are less likely to be confounded with shared environmental factors that can bias estimates from family studies or population studies where individuals are measured for both phenotypes. We use this framework to explore the genetic relationship between SZ and RA.

**Methods**

**Data**

Three RA and two SZ GWAS data sets (see Supplementary Table 1, available as Supplementary data at IJE online) were made available to us. Briefly, the Stahl et al. (‘Stahl’) RA sample comprises 5441 seropositive cases and 22 532 controls of European ancestry from six independent case-control cohorts.\(^{39,46}\) The Okada et al. (‘Okada’) RA sample has 3427 cases (1840 seropositive) and 6837 controls of European descent from five independent case-control cohorts,\(^{35}\) including the Corrona RA cohort. The Epidemiological Investigation of Rheumatoid Arthritis (EIRA) sample comprises 770 seronegative RA cases (EIRA seropositive cases from this cohort were already included in the Stahl sample). The Psychiatric Genomics Consortium (PGC) for Schizophrenia Wave 1 sample comprises data from 17 GWAS cohorts\(^{47}\) and a total of 9431 cases and 12 848 controls; and the Swedish (SWE) sample
comprises 5193 case and 6391 controls in four cohorts defined by genotyping platform which were independent of Swedish samples in PGC. All sample sizes are those after sample quality control (QC).

All data sets were processed through similar QC and imputation pipelines using the CEU + TSI Hapmap Phase 3 data as the reference panel. We augmented these QC so that estimates of genetic variance would not be influenced by artefacts of genotyping. SNPs with an imputation $r^2 > 0.6$ and an MAF $> 0.01$ in all cohorts were retained, resulting in 797,875 SNPs for analysis. Sex chromosome data were not available for all data sets and so were excluded. If any pair of individuals had an estimated similarity relationship coefficient $> 0.05$, one person was excluded at random so that all SZ cases, RA cases and controls were unrelated. The final analysis data set consisted of 8064 seropositive RA cases (including 1131 RA cases of unknown status of which at least two-thirds are expected to be seropositive), 1197 seronegative RA cases and 26,737 controls plus and 12,793 SZ cases and 15,912 controls (Figure 1).

Figure 1. Sample sizes.

Given accumulating evidence that seropositive and seronegative RA should be regarded as different clinical entities, we did not combine seropositive and seronegative RA cases.

SNP heritability and SNP genetic correlation

The bivariate linear mixed model genomic relationship matrix (GRM) restricted maximum likelihood (GREML) approach implemented in GCTA was used to estimate SNP heritabilities, the SNP coheritability and the SNP correlation between the disorders. The standard error (s.e.) of each estimate was calculated by the delta method which has been shown to agree well with s.e. expected from normal distribution theory. We used the estimate and its s.e. to generate a Wald statistic to test hypotheses that SNP heritabilities were different from zero and that SNP correlations were less than zero; the directional hypothesis for the correlation is justified by the epidemiological data reported in the Methods section. The model of analysis estimates SNP heritabilities as the proportion of variance in case-control status attributable to genome-wide SNPs, but estimates of SNP heritabilities and coheritabilities are presented on the liability scale, assuming population lifetime risk of 1% for SZ, 0.7% for seropositive RA and 0.3% for seronegative RA, so that they can be compared with estimates from epidemiological data. We note that when genetic relationships between individuals are small, the relationship between disease and liability scale is approximately linear and so the estimated genetic correlation is independent of scale. A SNP correlation of zero is estimated if the genome-wide relationship between cases of one disorder is the same with the cases as with the controls of another disorder. A SNP correlation reflects the magnitude of the covariance term between the traits relative to the product of the standard deviations, and so can be high even when the variance is low. A genome-wide SNP correlation could represent a uniform correlation across the genome or a weighted average of higher and lower correlations. Hence, we undertook genomic partitioning analyses which included multiple additive genetic random effects terms in the linear mixed model with multiple GRM constructed from non-overlapping SNP sets. Cohort and the first 20 principal components were included as covariates in all analyses. Sex was included as a covariate for SZ and in some analyses for RA. SNP heritabilities are presented on the liability scale. Follow-up analyses were conducted by sex and considering functional annotation.

Sensitivity analysis

We explored the sensitivity of our results and sought to exclude the possibility that genetic outliers could explain our results. We tested this by restricting the coefficient of similarity between any pair of individuals to be $< 0.025$. As one individual from a pair was excluded at random, we constructed 20 randomly drawn samples with restricted ancestry, and drew 20 random samples of the same size from the sample with coefficient of similarity $< 0.05$.

Benchmarking with epidemiological observations

From epidemiological studies we can obtain estimates of the population risk of SZ and RA, $K_{SZ}$ and $K_{RA}$.
respectively, and also for the probability of RA in those with SZ, \( K_{RA|SZ} \). We assume that the phenotypic liabilities of SZ (\( l_{SZ} \)) and RA (\( l_{RA} \)) are distributed as bivariate normal with mean 0, standard deviation 1 and correlation \( R_{SZ-RA} \):

\[
V \begin{pmatrix} l_{SZ} \\ l_{RA} \end{pmatrix} = \begin{pmatrix} V(l_{SZ}) & Cov(l_{SZ}, l_{RA}) \\ Cov(l_{SZ}, l_{RA}) & V(l_{RA}) \end{pmatrix} = \begin{pmatrix} 1 & R_{SZ-RA} \\ R_{SZ-RA} & 1 \end{pmatrix}
\]

The variances and covariance of liabilities among those affected with SZ

\[
V \begin{pmatrix} l_{SZ|DSZ} \\ l_{RA|DSZ} \end{pmatrix} = \begin{pmatrix} 1 - k_{SZ} & R_{SZ-RA}(1 - k_{SZ}) \\ R_{SZ-RA}(1 - k_{SZ}) & 1 - R_{SZ-RA}^2 k_{SZ} \end{pmatrix}
\]

where \( k_{SZ} = i_{SZ}(t_{SZ} - t_{SZ}) \) and reflects the proportional variance reduction as a consequence of ascertainment on SZ status, \( i_{SZ} = \frac{d_{SZ}}{l_{SZ}} \), with \( d_{SZ} \) the height of the normal curve at the threshold \( t_{SZ} \) defined from \( P(t > l_{SZ}) = K_{SZ} \). From \( K_{RA|SZ} \) we define the normal distribution threshold\(^{33,34} \) for RA in those with SZ as

\[
t_{RA|SZ} = \frac{t_{RA} - i_{SZ} K_{RA} R_{RA|SZ}}{\sqrt{1 - K_{RA|SZ}}}.
\]

Solving the quadratic for \( R_{SZ-RA} \) gives

\[
R_{SZ-RA} = \frac{i_{SZ} t_{RA} - \sqrt{t_{SZ}^2 t_{RA}^2 - (t_{RA}^2 + i_{SZ}^2)(t_{RA}^2 - t_{RA}^2)}}{t_{RA}^2 + i_{SZ}^2}
\]

(equation 1)

### Genotype x environment analysis

Our analyses led us to a postulate a hypothesis of genotype x environment interaction for SZ. Specifically we hypothesized that the SNP correlation would be less than 1 for coding and regulatory SNPs in a bivariate analysis in which the two traits are winter-born and non-winter born SZ cases and controls. We undertook a bivariate GREML analysis to test this hypothesis. Month of birth was only available for the SWE SZ sample, which comprised winter-born cases (born January to April\(^{22} \), \( n = 1511 \)) and winter-born controls (\( n = 2036 \)), as well as non-winter cases (\( n = 2962 \)) vs non-winter controls (\( n = 3772 \)); 199 individuals did not have month of birth recorded. To visualize the interaction, we identified 47318 SNPs associated with schizophrenia\(^{36} \) at \( P < 0.05 \) (the threshold that maximized out of sample prediction across multiple cohorts;\(^{36} \) the SNPs were quasi-independent with minor allele frequency > 0.05, pairwise linkage disequilibrium \( r^2 < 0.25 \) in a 250-kb window). We identified the risk alleles of the SNPs that defined the odds ratio to be greater than 1. We undertook association analysis

(logistic regression with 20 principal components, cohort and sex as covariates) in the Swedish sample that had season of birth recorded. We estimated the OR of the risk alleles and compared mean OR for SNPs annotated as C&R (coding/regulatory, the genomic region showing strong negative SNP correlation between SZ and RA, 2820 SNPs, 6%) and not coding/regulatory, testing the hypothesis \( H_0 \) Mean OR for winter-born sample = Mean OR for other sample.

### Results

#### SNP heritability and SNP genetic correlation

The estimated SNP heritability for RA seropositive cases was indistinguishable from zero (\( -0.006, \) s.e. 0.025, \( P = 0.98 \)). Despite the smaller sample size for seronegative cases, the s.e. shows that it was powered to detect SNP heritability > 5%. Given there was no evidence of contribution to risk of common variants for this sample, detection of a genetic relationship between these cases and SZ was not possible. Hence, all reported analyses are for seropositive RA cases only. Given the major contribution of the MHC region in RA (which may violate underlying assumptions of GREML\(^{55,56} \)), we undertook analyses using as the phenotype residuals after adjusting for the 550 SNPs (the number after pruning for SNPs with linkage disequilibrium \( r^2 > 0.99 \)) located within the MHC region (29–34Mb in chromosome 6) and for the other covariates. For SZ, the estimated SNP heritabilities were 0.223 (s.e. 0.006) including the MHC region and 0.212 (s.e. 0.006) after correcting for the MHC region. For RA, the estimated SNP heritabilities were 0.194 (s.e. 0.007) including the MHC region and 0.137 (s.e. 0.007), after correcting for the MHC region. The estimated SNP genetic correlations were –0.046 (s.e. 0.026) and –0.065 (s.e. 0.030) for including the MHC and after correcting for the MHC, respectively, which were significantly less than zero (\( P \)-values = 0.036 and 0.015, respectively) (Figure 2; Supplementary Table 2, available as Supplementary data at IJE online). We confirmed that the method to estimate \( P \)-values was robust, by checking the \( P \)-value from a likelihood ratio test comparing models with and without genome-wide SNP effects.

To aid interpretation and comparisons, we also present the SNP genetic covariance or coheritability, the latter represents the relationship between the disorders on the same scale as the heritability (Supplementary Table 2). Subset analyses (PGC-SZ/Stahl-RA, PGC-SZ/Okada-RA, SWE-SZ/Stahl-RA and SWE-SZ/Okada-RA) showed negative genetic correlations estimated for all combinations except PGC-SZ/Stahl-RA (Figure 2, Supplementary Table 2). We explored the sensitivity of our results and sought to exclude the possibility that genetic outliers could explain our results.
We found that the SNP correlation between SZ and RA was significantly more negative in 20 samples drawn from our data when ancestry was more restricted (similarity relatedness coefficient $< 0.025$) [PC0.054, standard deviation over replicates (s.d.) 0.002 vs $-$0.047 s.d. 0.001] (Supplementary Figure 1). This sensitivity analysis implies that the negative correlation is not driven by ancestry artefacts and provides confidence that SNP correlation between SZ and RA is negative.

Sex analyses

Given that risk of RA is higher in females and risk of SZ is higher in males, we undertook SZ/RA analyses stratified by sex (i.e. four-trait multivariate GREML, in which the four traits were SZ-male, SZ-female, RA-male, RA-female each matched by their sex-specific control set) to determine if SNP correlations (based on autosomal SNPs) were sex dependent. Sex information was missing for $\sim 11\%$ of the RA sample who were excluded in the analysis. Based on reported male:female population ratios, we assumed the male and female baseline risks were 0.42% and 0.98% for seropositive RA, and 1.15% and 0.85% for SZ respectively. SNP heritabilities were significantly greater when estimated from males compared with females for SZ (male 0.258, s.e. 0.010, female 0.214, s.e. 0.012, $P = 0.0053$) but not for RA (male 0.174, s.e. 0.016, female 0.158, s.e. 0.013, $P = 0.43$) (Figure 2, Supplementary Table 2); these estimates must be interpreted with caution, recognizing that they are dependent on the lifetime risk of disease chosen for each sex. SNP correlations between sexes were high but were significantly different from 1 for both RA ($P = 6.1e-06$) and SZ ($P = 2.4e-07$, Supplementary Table 2). All SNP correlation point estimates are negative between male/female SZ/RA analyses.

Functional annotation analyses

Previous studies have demonstrated that contributions to SNP heritabilities are not distributed equally over the genome. We therefore set out to test if the SNP-correlation between SZ and RA was dependent on SNP annotation. We undertook genomic partitioning analyses in which multiple additive genetic random effects terms were considered in the linear mixed model, with multiple GRM each constructed from SNPs grouped by a functional annotation. Following Gusev et al., SNPs were classified as being in coding/regulatory (in exons, 3' UTR, 5'UTR, 1-kb region up- and downstream of transcription start and end site and noncoding RNA), DNase I hypersensitivity sites (DHS) and intronic or intergenic regions. SNPs with multiple annotations were allocated with hierarchical preference of coding/regulatory, over DHS and over intronic. For SZ, all annotations had estimates of SNP heritability that were significantly greater than zero (Figure 3; Supplementary Table 3, available as Supplementary data at IJE online), although the proportion of total variance allocated to intergenic SNPs was significantly less than expected given the proportion of all SNPs annotated to that group (Figure 3; Supplementary Table 3). For RA, the coding/regulatory and DHS annotations had SNP heritability estimates that were both significantly different from zero and were higher than expected based on the proportion of SNPs in those functional partitions (Figure 3; Supplementary Table 3). As a consequence in RA, and in contrast to SZ, neither intronic nor intergenic variants...
made a significant contribution to SNP heritability, a finding that is underscored by the fact that the proportion of all SNP heritability contributed by these types of variant is significantly less than expected based upon the proportion of SNPs of these classes. Given that in RA, SNP heritability is essentially restricted to SNPs in coding/regulatory and DHS regions, these are the only classes of SNP for which meaningful estimates of SNP correlation between SZ and RA can be made. We note that currently available DHS annotation may be biased towards cell types of relevance to RA compared with SZ. We estimated a significant and stronger genetic correlation for coding/regulatory region

\( \begin{align*} \frac{0.037}{0.14}, & \quad 0.016, \quad 2.7 \times 10^{-06}, \quad 1.02 \times 10^{-07} \end{align*} \)

regions \((-0.322, \text{ s.e. } 0.115, \text{ } P = 0.003\) \) (Figure 3; Supplementary Table 3) than for the whole genome.

**Immune related pathway analyses**

Since RA is an autoimmune disease and since the epidemiological negative relationship between SZ and RA has contributed to the autoimmune hypothesis of SZ, we set out to test if the SZ-RA SNP correlation is more negative in SNPs in immune-related pathways. To avoid multiple testing, we selected a single immune gene set based on previous work using the Stahl GWAS results of Hu et al.\(^5\) Using gene expression from 223 murine immune cell types, they reported over-representation of RA-associated SNPs in genes expressed specifically in CD4+ effector memory T cells, with strongest over-representation in genes expressed in the subcutaneous lymph node subset named T.4Mem44h621.LN.\(^5\) We selected the top 4000 genes expressed in the T.4Mem44h621.LN cells to be in the ‘T4Mem’ set. The arbitrary threshold of 4000 genes generated an SNP set of about 10% of the total SNPs analysed. We tested if the T4Mem genes make an enriched contribution to SNP heritabilities and the SNP correlation, partitioning the coding/regulatory, DHS and intronic partitions into T4Mem and non-T4Mem classes (Supplementary Table 4, available as Supplementary data at IJE online). As expected, there was a significant enrichment for variance attributable to the T4Mem class for RA (28% of the SNP heritability compared with only 11.5% of SNPs, 2.5 fold enrichment, \(P = 2.0e-11\)). Interestingly, there was also significant enrichment of variance attributable to T4Mem regulatory SNPs in SZ (16% of SNP heritability, \(P = 2.4e-04\)). We did not find evidence that SNPs in this group of genes were more negatively correlated than those in the rest of the genome, although the patterns of correlations are difficult to interpret, and the size of the standard errors means that the sample is underpowered (Supplementary Table 4).

**Benchmarking with epidemiological observations**

We can benchmark our estimated SNP correlation between RA and SZ of \(-0.046\) relative to expectation from epidemiological data. Using the meta-analysis result that the risk of RA in those with SZ is 29% of the risk in the general population,\(^1\) we estimate that this implies a phenotypic correlation between the disorders of \(-0.15\) (equation 1; \(K_{SZ} = 0.01\) and \(K_{RA} = 0.01\), respectively, \(K_{RA/SZ} = 0.29 \times K_{RA}\)). More recent epidemiological studies\(^2,7\) imply a substantially smaller negative phenotypic correlation (\(-0.05\)). Phenotypic correlation is considered a reasonable benchmark for genetic correlation,\(^59\) and
therefore our small negative estimate of the genetic correlation is consistent with the epidemiological data. However, the genetic relationship between SZ and RA is complex. As an autoimmune disease, the MHC region contains risk factors for RA which alone explain ~5% of the variance in liability to RA, whereas the most significant individual SNP association for SZ, also in the MHC region, explains only ~0.1% of variance in liability (ref60 and consistent with Table 1 with and without MHC region). These MHC risk alleles are positively correlated in SZ and RA, although clearly the effect sizes are very different. In analyses in which we first removed the contribution to variance of the MHC region, the magnitude of negative correlation between SZ and RA increased (~0.065, s.e. 0.015, \( P = 0.015 \)), and increased further still when considering only SNPs in coding and regulatory regions of the genome (~0.322, s.e. 0.115, \( P = 0.003 \)), indicating that regions of the genome other than the MHC region contribute to the epidemiological observations.

**Genotype x environment analysis**

Although RA is well recognized as an immune disorder, the contribution of immune activation to SZ is open to debate (see Introduction). To add further insight to the complex relationship between SZ and RA, we postulated that if there is any interplay in risk for SZ between environmental risk factors associated with immune challenge and genes that are relevant to the immune response, it is likely to occur at the coding and regulatory regions. This is based on the rationale that these are the sets of variants that both capture the SNP heritability of RA (i.e. are likely to be most enriched for genes with influences on immune activation) and are negatively correlated with SZ (i.e. are both immune activation and SZ relevant). If our hypothesis is correct, we predict that the apparent effect sizes in schizophrenia at these loci will be greater in those exposed to a relevant immune challenge than in those who are not. Consequently, we predict that the negative correlation between SZ and RA will be larger at these loci in cases exposed to an immune challenge. At present, our ability to test this G x E hypothesis is limited by availability of samples that are both genetically informative and recorded for environmental risks. We therefore sought a proxy for immune challenge in SZ.

A robust epidemiological finding in SZ research is that people with the disorder are more likely to be born in winter or spring than summer or autumn61 (odds ratio 1.07, 95% CI 1.05–1.08, estimated from a meta-analysis of 27 studies62). Winter/spring birth is also associated with recognized immune-mediated disorders including RA.63 Candidate exposures underlying this finding include seasonally varying factors such as prenatal vitamin D, or maternal/fetal exposure to infections. Both factors can impact on the immune system.64–66 In the absence then of direct measures of immune activation, we used season of birth as our proxy measure; in doing so, we are aware this proxy measure is only likely to be weakly correlated with exposure and hence its use will adversely affect power.

Month of birth was only available for the SWE SZ sample. We selected quasi-independent SNPs associated at \( P < 0.05 \) and minor allele frequency > 0.05 (47 318 SNPs) from the largest published schizophrenia meta-analysis.36 We next divided them into coding/regulatory (2820 SNPs, 6%) and non coding/regulatory sets. We undertook a bivariate GREML analysis of the SWE datasets in which the two traits were winter-born (January to April) vs non-winter born SZ cases and controls; this division was justified by studies using Swedish data.22 The correlation between winter/non-winter born was significant for coding/regulatory SNPs (0.56, s.e. 0.14, \( P = 0.0009 \)) but not so for other SNPs (0.95, s.e. 0.05, \( P = 0.15 \)) (Supplementary Table 5, available as Supplementary data at IJE online). To visualize this interaction, and to demonstrate that effects sizes of the coding/regulatory SNP set are increased in the winter-born cohort, we present the mean association OR for the different season by annotation classes estimated in the SWE SZ sample (Figure 4). We confirmed that results were robust to the \( P \)-value threshold used for selection of associated SNPs (Supplementary Table 6, available as Supplementary data at IJE online) and we checked the seasonal trend by using sliding window definitions of 4-month season definitions (Supplementary Table 7, available as Supplementary data at IJE online). The results are consistent with our G x E hypothesis of immune-related disruption proxied by winter birth, in concert with risk variants in coding/regulatory regions increasing risk of schizophrenia.

**Strengths and limitations**

The strength of our methodological approach to explore the genetic relationship between SZ and RA is that it is based on genome-wide genotype data and uses independently collected data for the two diseases studied. While under review, two studies have investigated the relationship between SZ and RA based partially67 or fully68 on GWAS summary statistics; one reported a non-significant small positive relationship67 and the other a non-significant small negative relationship.68 Use of genotype data is computationally more demanding but is considered more definitive than methods based on summary statistics,67 and such an approach is needed when the relationship between the disorders is benchmarked by a weak phenotypic correlation of ~0.05 to ~0.15. The results here indicate that there is a subset of
estimates of heritability based on family data, and also consistent with smaller OR for seronegative genetic associations than for seropositive associations. A SNP correlation of 0.98, s.e. 0.165, was estimated in a Han Chinese sample. We note that the proportion of seronegative cases in this study was high (519/952 = 55%) and any misclassification could serve to inflate the correlation. Third, the non-availability of sex chromosome data meant that we could only explore sex differences based on autosomal SNPs. Fourth, we did not have the power to break down the signal attributed to coding/regulatory regions into more finely defined functional categories. Last, given the limitations on the data available to test the G x E hypothesis, alternative explanations for our results cannot be excluded.

**Discussion**

In summary, we have applied a mixed linear model method to estimate the genetic correlation between RA and SZ. Epidemiological evidence has demonstrated decreased prevalence rates of RA in SZ cases, consistent with a phenotypic correlation of liabilities of up to −0.15. We show that there is a small but significant negative correlation across the genome and the signal is stronger for SNPs annotated as coding and regulatory. Given that RA is an immune-related disease and that a role for immune activation has long been hypothesized for SZ, a negative genetic correlation could imply that variants in immune response pathways have different roles in different tissues and/or in response to different challenges. The immune activation hypothesis of SZ is partly founded on an increased risk for SZ associated with month of birth. Our hypothesis that increased effect sizes for SZ-associated SNPs in the coding/regulatory SNP set for a winter-born case-control set was supported by our analyses, although other explanations for these results may be possible. Most importantly, if the complexity of SZ is to be unraveled, then data sets that are informative for both genetic and environmental risk factors are essential. Since SZ is an adult-onset disorder, and yet perinatal and childhood experience, especially infections, are known environmental risk factors, then prospective gathering of data in nationally accessible repositories is needed.

**Funding**

This work was supported by the Australian Research Council [grant number DE130100614 to S.H.L.], the National Health and Medical Research Council [grant numbers 613602, 1078901 to N.R.W.; 1047956 to N.R.W., S.H.L. and B.J.M., 1053639 to E.M.B]; the Arthritis Foundation to S.R., the Doris Duke Foundation to S.R.; the National Institutes of Health [grant numbers 1R01AR063759-01A1, 1U01HG0070033, 5U01GM092691-04 to S.R.; R01 MH077139 for the Sweden SZ Study to P.F.S.]. The Swedish SZ
study was also funded by the Karolinska Institutet, Karolinska University Hospital, and the Swedish Research Council. The RA dataset from Vanderbilt University Medical Center’s BioVU is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH. Other funding acknowledgements can be found in the primary publications from each study, as references. Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) hosted by SURFsara, and financially supported by The Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam.

Acknowledgements

We thank Peter Smart, Jake Carroll, Irek Porebski and the Queensland Brain Institute IT team for technical support.

Author contributions: N.R.W. was responsible for overall direction of the study, with analyses conducted by S.H.L. E.M.B. conducted a literature review and wrote the first draft of the introduction generating hypotheses tested. N.R.W., S.H.L., S Raychaudhuri, E.M.B., B.J.M. and A.P. contributed to decisions about analyses conducted. N.R.W. wrote the first draft of the manuscript with substantive contributions including suggestions for follow-up analyses from S.H.L., S Raychaudhuri, P.F.S, A.P., M.O.D. and Y.O. S Ripke was responsible for initial QC of the GWAS data sets. A.A.E., O.A.A., T.F., A.G., V.M., J.J.McG., D.M., E.A.S., P.S. and Q.Z. contributed to secondary analyses.

Author list continued:

Affiliations: 1Department of Psychiatry, University of California, San Francisco, CA, USA, 2NCIRE (Northern California Institute of Q Research and Education), San Francisco, CA, USA 3Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center, Utrecht, The Netherlands, 4David Geffen School of Medicine, University of California, Los Angeles, CA, USA, 5Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany, 6Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, Switzerland, 7Institute of Neuroscience and Medicine (INM-1), Research Center Juelich, Juelich, Germany, 8Department of Psychiatry, Trinity College Dublin, Dublin, Ireland, 9UCL Genetics Institute, University College London, London, UK. 10NORMENT, KG Jebsen Centre for Psychosis Research, Department of Clinical Science, University of Bergen, Bergen and Department of Medical Genetics, Oslo University Hospital, Oslo, Norway, 11Medical Research Council (MRC) Centre for Neuropsychiatric Genomics and Genomics, Cardiff University School of Medicine, Cardiff, UK, 12Biostatistics and Bioinformatics Unit, Cardiff University, Cardiff, UK, 13Department of Psychiatry and Behavioral Sciences, North Shore University Health System and University of Chicago, Evanston, IL, USA, 14Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University School of Medicine, Cardiff, UK, 15Department of Psychiatry, University of Halle, Halle, Germany, 16Institute of Biological Psychiatry, Copenhagen University Hospital, Roskilde, Denmark, 17Lundbeck Initiative for Integrative Psychiatric Research, iPSYCH, Roskilde, Denmark, 18Department of Genetics, University of North Carolina at Chapel Hill, NC, USA, 19Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA, 20Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK, 21Molecular Medicine Centre, University of Edinburgh, Edinburgh, UK, 22Mental Health Sciences Unit, University College London, London, UK, 23Cognitive Genetics and Therapy Group, Discipline of Biochemistry and School of Psychology, National University of Ireland, Galway, Ireland, 24Institute of Human Genetics, University of Bonn, Bonn, Germany, 25Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA, 26Department of Psychiatry, University of Colorado Denver, Aurora, CO, USA, 27Department of Psychiatry, Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, 28Department of Complex Trait Genetics, VU University, Amsterdam, The Netherlands, 29Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany, 30Department of Psychiatry & Behavioral Sciences, Johns Hopkins University, Baltimore, MD, USA, 31Department of Psychiatry & Psychotherapy, University of Gottingen, Gottingen, Germany, 32Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA, 33Oxford Health NHS Foundation Trust, Marlborough House Secure Unit, Milton Keynes, UK and 34Faculty of Health and Medical Science, University of Copenhagen, Copenhagen, Denmark.


Affiliations: 1KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, 2Department
Conflict of interest: Thomas Werge has acted as lecturer and consultant to the pharmaceutical company Lundbeck A/S (Denmark). Jeffrey D Greenberg is an employee of, and has stock in, Corrona, LLC, and is consultant for AstraZeneca, Celgene, Novartis and Pfizer. Paul Tak is now an employee of GSK and holds stock in GSK after completion of this study; GSK was not involved in this study.

References


