Additive Genetic Variation in Schizophrenia Risk Is Shared by Populations of African and European Descent

Teresa R. de Candia,[1,2,*] S. Hong Lee,[3] Jian Yang,[3,4] Brian L. Browning,[5] Pablo V. Gejman,[6] Douglas F. Levinson,[7] Bryan J. Mowry,[3,8] John K. Hewitt,[1,2] Michael E. Goddard,[9,10] Michael C. O'Donovan,[11] Shaun M. Purcell,[12] Danielle Posthuma,[13,14,15] the International Schizophrenia Consortium,[16] the Molecular Genetics of Schizophrenia Collaboration,[16] Peter M. Visscher,[3,4] Naomi R. Wray,[3,17] and Matthew C. Keller[1,2,17,*]

To investigate the extent to which the proportion of schizophrenia's additive genetic variation tagged by SNPs is shared by populations of European and African descent, we analyzed the largest combined African descent (AD [n = 2,142]) and European descent (ED [n = 4,990]) schizophrenia case-control genome-wide association study (GWAS) data set available, the Molecular Genetics of Schizophrenia (MGS) data set. We show how a method that uses genomic similarities at measured SNPs to estimate the additive genetic correlation (SNP correlation [SNP-r_g]) between traits can be extended to estimate SNP-r_g for the same trait between ethnicities. We estimated SNP-r_g for schizophrenia between the MGS ED and MGS AD samples to be 0.66 (SE = 0.23), which is significantly different from 0 ($p_{(\text{SNP-rg} = 0)} = 0.0003$), but not 1 ($p_{(\text{SNP-rg} = 1)} = 0.26$). We re-estimated SNP-$r_g$ between an independent ED data set (n = 6,665) and the MGS AD sample to be 0.61 (SE = 0.21, $p_{(\text{SNP-rg} = 0)} = 0.0003$, $p_{(\text{SNP-rg} = 1)} = 0.16$). These results suggest that many schizophrenia risk alleles are shared across ethnic groups and predate African-European divergence.

Introduction

Schizophrenia is a severe mental disorder with a lifetime prevalence around 1% and a heritability of 0.7–0.8.[1,2] Using a linear mixed model, we previously found that about a third of the genetic variation in liability to schizophrenia is captured by additive effects of common (minor allele frequency [MAF] > 0.01) SNPs in a large European case-control data set.[3] This approach first derives genetic similarities at measured SNPs among classically unrelated individuals and then uses those similarities to estimate the amount of variability explained by all SNPs together. Because rare causal variants are not well predicted by common SNPs, this finding suggests that a substantial portion of the heritability of schizophrenia is due to additive effects of common causal variants,[3] a conclusion consistent with extrapolations from results by a different methodology.[4] However, it is unclear whether this result also applies to populations of African descent and, if so, whether the proportion of schizophrenia's genetic variation tagged by SNPs is shared between populations of African descent (AD) and European descent (ED). Although rare variants (MAF < 0.01) are largely population specific,[5,6] common variants are typically polymorphic across divergent ethnic groups;[7] therefore, it is possible that a meaningful portion of schizophrenia's genetic variation tagged by common SNPs could be shared between ED and AD individuals. However, different causal variants between ethnicities, or different linkage-disequilibrium (LD) patterns between SNPs and causal variants across ethnicities, could result in lower overlap in additive genetic variation tagged by SNPs. To date, ~96% of genome-wide association study (GWAS) participants have been ED individuals.[8,9] Because SNP associations might differ between ethnicities, genetic studies using European-only samples could yield results with a strong Eurocentric bias.[10]

A traditional approach for understanding whether SNP associations are consistent across ethnicities is to assess the similarity of GWAS results among different ethnic groups[11–14]. However, for most complex traits, only a small minority of truly associated SNPs reach stringent genome-wide significance thresholds,[15] and so overlap in

[1]Department of Psychology and Neuroscience, University of Colorado, Boulder, CO 80302, USA; [2]Institute for Behavioral Genetics, University of Colorado, Boulder, CO 80302, USA; [3]The University of Queensland, Queensland Brain Institute, Brisbane, QLD 4072, Australia; [4]The University of Queensland, Diamantina Institute, Brisbane, QLD 4072, Australia; [5]Department of Medicine, Division of Medical Genetics, University of Washington, Seattle, WA 98195, USA; [6]Department of Psychiatry and Behavioral Sciences, Northshore University Health System and University of Chicago, Evanston, IL 60601, USA; [7]Department of Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, CA 94305, USA; [8]Queensland Centre for Mental Health Research, Brisbane, QLD 4076, Australia; [9]Department of Agriculture and Food Systems, University of Melbourne, Melbourne, VIC 3010, Australia; [10]Biosciences Research Division, Department of Primary Industries, Melbourne, VIC 3001, Australia; [11]MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff CF14 4XN, UK; [12]Broad Institute, Cambridge, MA 02141, USA; [13]Complex Trait Genetics, Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, Amsterdam 1081 HV, the Netherlands; [14]Department of Clinical Genetics, VU University Medical Center, Amsterdam 1081 HV, the Netherlands; [15]Department of Child and Adolescent Psychiatry, Erasmus University Rotterdam, Sophia Child Hospital, Rotterdam 3000 CB, the Netherlands
[16]Full membership lists for the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia Collaboration can be found in the Supplemental Data
[17]These authors contributed equally to this work
*Correspondence: teresa.decandia@colorado.edu (T.R.d.C.), matthew.c.keller@gmail.com (M.C.K.)
http://dx.doi.org/10.1016/j.ajhg.2013.07.007. ©2013 The Authors

significant associations is not necessarily expected, even if the effects of most SNPs are similar across ethnicities. In an attempt to gain insight into the overall consistency of schizophrenia SNP associations between ED and AD populations, a previous study[4] used extremely liberal p value thresholds to define large sets of risk-score alleles in LD in an ED "discovery" sample and used these alleles to generate risk scores in independent "target" samples (the Molecular Genetics of Schizophrenia [MGS] ED and AD samples, as well as the smaller "O'Donovan" ED sample). The most significantly associated set of SNPs in the ED discovery sample did predict case-control status in the AD target sample ($p = 0.008$) but explained substantially less variance ($R^2 = 0.4\%$) than in the ED target samples (MGS ED: $R^2 = 3.2\%$, $p = 2 \times 10^{-28}$; O'Donovan: $R^2 = 2.3\%$, $p = 5 \times 10^{-11}$), which might appear to suggest only modest overlap in schizophrenia SNP associations between EDs and ADs.

A limitation of the genetic-risk-score approach[4,16] is that it requires accurate estimation of effect sizes of individual SNPs in the discovery sample, and error variance of GWAS point estimates accumulates in prediction scores and thus causes estimates of overlap (R^2) to be biased downward as a function of sample size. Furthermore, the genetic-risk-score approach can underestimate overlap between ED and AD populations, in particular when discovery SNPs are pruned to be in LD in an ED discovery sample: because LD tends to be lower in AD populations,[17] more SNPs are required for predicting causal variants in AD populations than in ED populations.

In the present study, we used a bivariate linear mixed-effects model implemented in GCTA[18] to estimate the proportion of schizophrenia-risk variation that is tagged by the additive effects of SNPs (we refer to this estimate as the SNP heritability) in predominately AD individuals (African Americans). We also estimated the additive genetic correlation tagged by SNPs (the SNP correlation) in schizophrenia between ED and AD individuals. With this approach, the effects of SNPs are treated as statistically random and individual SNP effects are not estimated, nor is there a need to prune SNPs for LD or statistical significance. Rather, all SNPs are used simultaneously for estimating genome-wide similarities between pairs of individuals; these similarity estimates are in turn used for estimating variance and covariance parameters in a linear mixed-effects model. This method does not depend on reliable estimation of individual SNP effects and so should provide unbiased estimates of the SNP heritability and SNP correlation regardless of subject sample size (larger samples should decrease the SEs of estimates).

Material and Methods

GWAS Samples and Quality Control

The MGS and International Schizophrenia Consortium (ISC) case-control data sets have previously been described in detail.[3,4,19,20] The study and use of these data sets were approved by institutional review boards at the University of Colorado. Because MGS subjects and just over half of ISC subjects were genotyped for 909,622 autosomal SNPs (on Affy 6 Platforms), whereas just under half of ISC subjects were genotyped for ~500K autosomal SNPs on earlier platforms (Affy 5 and Affy 500K), we used imputed ISC data to maximize the number of subjects included in the analysis. Therefore, we conducted the main analysis by using raw MGS and imputed ISC genotypes. We imputed ISC genotypes with HapMap 3 CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) and TSI (Toscani in Italia) populations as reference panels.[20] We imputed MGS ED genotypes in the same way to allow a comparison of results from raw and imputed data. We found no evidence that use of imputed data affected estimates of SNP heritabilities or SNP correlations; results were virtually identical across models that used imputed versus raw MGS ED data (Table S1, available online). We did not attempt to impute MGS AD genotypes because of the difficulty of imputing admixed AD samples.[21,22] We calculated the MAF and imputation R^2 (ratio of observed to expected variance) for each SNP for quality-control purposes.

Table 1. Cohort Sizes and Univariate SNP Heritabilities and SEs

Sample	Cases	Controls	h^2 (SE)[a]
MGS AD	1,223	919	0.24 (0.09)
MGS ED	2,571	2,419	0.28 (0.03)
ISC ED[b]	3,220	3,445	0.27 (0.02)

[a] h^2 represents SNP heritability.
[b] ISC comprises eight imputation cohorts genotyped at separate sites.

Each data set underwent stringent quality-control procedures separately. In particular, we dropped (1) individuals with average missing rates in raw SNP calls > 0.02; (2) individuals who were more than 5 SDs away from the cohort's (within data set, within ethnicity) mean scores on the first two ethnic principal components (PCs; see "PC Analysis" below), the first of which differentiated ED and AD individuals; and (3) one individual from each pair of individuals with genome-wide similarities > 0.05. We dropped SNPs when (1) MAF < 0.01; (2) missing rates > 0.02 (raw data only); (3) missing rate differences between cases and controls > 0.02 (raw data only); (4) SNP frequency differences to CEU HapMap > 0.15 (because AD individuals tend to be admixed, the MGS AD data set did not undergo this step); (5) Hardy-Weinberg equilibrium test (controls only) $p < 10^{-6}$; and (6) imputation $R^2 < 0.6$ in all cohorts (imputed data only). Only SNPs in common across cohorts were kept for analysis, resulting in 477,664 SNPs. Final cohort sample sizes are summarized in Table 1.

PC Analysis

PC analysis is used for reducing the dimensionality of data with correlated variables. When applied to genomic-similarity matrices computed from SNP data, the first several PCs typically capture ancestry differences between populations and later PCs can also capture technical batch effects. We ran PC analysis on subject pairwise genomic (identity by state [IBS]) similarities across SNPs from the combined MGS AD, MGS ED, and ISC ED samples, along with JPT, YRI, and CEU HapMap reference panels, by using a random subset of 30K SNPs that were pruned to be in linkage equilibrium and that were shared in common across samples. The use of HapMap reference panels "anchored" our PC results, such that the first PC (PC1) distinguished African and European ancestries and the second distinguished Asian ancestry (Figure S1). We used the PC results to remove ethnic outliers (see "GWAS Samples and Quality Control" above), for inclusion as covariates to control for population stratification (see "Linear Mixed Models for Estimating Case-Control Status" below), and to divide the MGS AD sample into subsamples on the basis of degree of European admixture. After quality control, ADs were divided into overlapping subsets of decreasing degrees of European admixture, removing individuals > 2, > 1, and > 0.5 SDs above the mean AD PC1 score (Figure S1).

Linear Mixed Models for Estimating Case-Control Status

For univariate heritability models, we estimated genomic similarities between pairs of individuals separately in the MGS AD sample, in the MGS ED sample, and in the ISC ED sample by using the method described by Yang et al.[23] In particular, we calculated the genomic similarity for each pair of individuals by taking the sum of the products of SNP coefficients between those individuals and then scaling that sum by the expected SNP heterozygosity,

$$\widehat{A}_{ij} = \frac{1}{L}\sum_{l=1}^{L}\frac{(x_{il}-2p_l)\bullet(x_{jl}-2p_l)}{2p_lq_l}\ (i\neq j)$$

$$\widehat{A}_{ii} = 1+\frac{1}{L}\sum_{l=1}^{L}\frac{x_{il}^2-(1+2p_l)x_{ij}+2p_l^2}{2p_lq_l},$$

where $x_{il} = 0$, 1, or 2 according to whether individual i has genotype bb, Bb, or BB, respectively, at locus l (alleles are arbitrarily called b or B), p_l and q_l are the allele frequencies of B and b, respectively, and $2p_l$ is the mean of x_l. These values are scaled to be both positive and negative; therefore, for clarity we use the term "similarity" rather than "relationship."

For bivariate models, we re-estimated the genomic similarities in the combined MGS AD and MGS ED samples, in the combined MGS AD and ISC ED samples, and in the combined MGS ED and ISC ED samples. In the bivariate model, the observed disease status for traits 1 and 2 can be written as

$$y_1 = X_1b_1 + Z_1g_1 + e_1 \text{ for trait 1}$$

$$y_2 = X_2b_2 + Z_2g_2 + e_2 \text{ for trait 2},$$

where y is a vector of observations, b is a vector of fixed covariate effects, g is a vector of genetic effects tagged by SNPs, and e is a vector of residuals. X and Z are incidence matrices for b and g, respectively. The variance covariance matrix (V) is defined as

$$V = \begin{bmatrix} Z_1AZ_1'\sigma_{g_1}^2 + I\sigma_{e_1}^2 & Z_1AZ_2'\sigma_{g_1g_2}^2 \\ Z_2AZ_1'\sigma_{g_1g_2}^2 & Z_2AZ_2'\sigma_{g_2}^2 + I\sigma_{e_2}^2 \end{bmatrix},$$

where A is the genomic similarity relationship matrix based on SNP information across all groups, I is an identity matrix, and σ_g^2, σ_e^2, and $\sigma_{g_1g_2}^2$ are genetic variance tagged by SNPs, residual variance, and covariance between g_1 and g_2, respectively.[18] We treated schizophrenia in each sample as a separate "trait" and estimated SNP heritabilities (h_1^2 and h_2^2) and the SNP liability covariance (cov_{12}) or SNP correlation (SNP-$r_g = cov_{12}/h_1h_2$) between ethnicities.[18,23,24] For all models, we transformed genetic-variance estimates to a liability scale by assuming a disease prevalence of 1%, as described previously,[24] and controlled for gender, 20 PCs, and site (Table 1) as covariates. Although there were large differences in average similarities of ED-AD pairs compared to within-ethnicity pairs, we controlled for these differences by including ancestry PCs as fixed effects, the first of which captured the African-European gradient (see Figure S1). We also included $PC1^2$ in bivariate models in order to control for any possible ascertainment or diagnostic differences that might exist between the most and least admixed individuals. For each model, PCs included as covariates were from an IBS matrix of subjects included in that model. Because MAFs can vary by ethnicity, for bivariate models, elements in the A matrix were standardized with expected allele frequencies that varied by data set.

In the current context, SNP heritability refers to the proportion of phenotypic variance in liability to schizophrenia that is explained by the additive effect of all SNPs together, and SNP correlation refers to the additive genetic correlation (explained by all SNPs together) between the liability of schizophrenia in ADs and the liability of schizophrenia EDs. Evidence of a SNP correlation between ADs and EDs occurs to the degree that AD cases are more genetically similar to ED cases than to ED controls, and vice versa, after main effects of ethnicity are controlled for. Because the two traits (schizophrenia in ADs and schizophrenia in EDs) are measured on different individuals, the only link between traits is in the genomic-similarity scores for each pair of individuals. Therefore, unlike traditional (e.g., twin) models that decompose phenotypic correlations into additive genetic, nonadditive genetic, and environmental sources that can be mutually confounded, in the present model there is no such confounding, and SNP correlations are unlikely to be due to residual (e.g., environmental or nonadditive genetic) sources.

We used a permutation approach in which affection status was permuted within ethnicity in order to exclude the possibility of upward biases of SNP-heritability and SNP-correlation estimates (see Figure S2). In addition, because prevalence rates have been shown to vary across populations,[25] we varied this parameter across populations and found that it did not substantially alter SNP heritabilities or SNP correlations (see Figure S3). That SNP correlations are unaffected by prevalence-rate differences is also evident from the transformation equation in Lee et al.[18] Also, to understand whether some of the genetic overlap in schizophrenia between ED and AD samples might be due to the relatively high level of European admixture in ADs, estimated to be an average of ~15%,[26] we ran additional bivariate models, each of which estimated SNP correlations between EDs and AD subsamples of higher-proportion African ancestry as judged from the first ancestry PC (see "PC Analysis" above).

We used the likelihood-ratio test statistic to compare the fit of full models in which SNP correlations were freely estimated to the fit of reduced models in which SNP correlations were fixed to either 0 or 1. The distribution of differences in log likelihoods between full and reduced models is approximately chi-square distributed with 1 degree of freedom.

Binning SNPs by Rates of Recombination

We examined whether SNPs in regions of high recombination showed lower SNP correlations than SNPs in regions of low recombination. To do this, we used recombination rates estimated from CEU, YRI, and JPT+CHB HapMap populations on 3,303,900 SNPs[27] and interpolated recombination rates for 476,704 SNPs in our data sets. For each SNP, an interpolated recombination rate was derived from a linear regression of recombination rates on base pair positions for its two closest HapMap neighbors. A

Table 2. SNP-Heritability and SNP-Correlation Estimates from Bivariate Models

Sample 1	Sample 2	Sample 1 n	Sample 2 n	Sample 1 h^2 (SE)	Sample 2 h^2 (SE)	cov_{12}[a]	SNP-r_g (SE)	$p_{(SNP\text{-}rg\ =\ 0)}$	$p_{(SNP\text{-}rg\ =\ 1)}$
Across Ethnicity									
MGS AD	MGS ED	2,142	4,990	0.20 (0.09)	0.33 (0.03)	0.17	0.66 (0.23)	0.0003	0.26
MGS AD	ISC ED	2,142	6,665	0.22 (0.09)	0.27 (0.02)	0.15	0.61 (0.21)	0.0003	0.16
Within Ethnicity									
MGS ED	ISC ED	4,990	6,665	0.26 (0.03)	0.27 (0.02)	0.22	0.83 (0.09)	<0.0001	0.09

[a] Calculated by cov_{12} = SNP-$r_g h_1 h_2$.

SNP-correlation estimate was then derived for a model that only included SNPs with above-median recombination rates and separately for a second model that only included SNPs with below-median recombination rates.

Results

We found that the univariate SNP-heritability estimate for schizophrenia in ADs ($h^2 = 0.24$, SE = 0.09) was slightly lower than the SNP-heritability estimates in EDs ($h^2 = 0.28$, SE = 0.03 in MGS EDs; $h^2 = 0.27$, SE = 0.02 in ISCs; Table 1), as might be expected given the lower average LD between SNPs in African populations.[17] The schizophrenia-liability SNP correlation captured by autosomal SNPs between the sexes was high (SNP-r_g = 0.70) among ADs, although a strong conclusion is not possible given the large SE (0.51) around the SNP-r_g estimate. Nevertheless, this estimate is consistent with the between-sex genetic correlation we previously reported for schizophrenia in a much larger sample of EDs (SNP-r_g = 0.89, SE = 0.06). To assess whether SNP heritability differed by functional annotation of SNPs, we partitioned the variance explained by SNPs into three components by creating genomic-similarity matrices of SNPs in genes expressed in the CNS, those found in other genes, and those not localized to genes. The proportion of SNP heritability estimated from SNPs in genes expressed in the CNS was 30% for ADs, and although large SEs do not allow us to distinguish whether this is significantly greater than the 20% of the genome these SNPs represent (SE = 0.18, $p_{(30\%\ =\ 20\%)} = 0.30$), these results are also consistent with those we reported previously for EDs (31%, SE = 0.02, $p_{(31\%\ =\ 20\%)} = 7.6 \times 10^{-8}$; see Table S2).

The main goal of our analysis was to understand the degree to which genetic variation tagged by SNPs is shared between ED and AD populations. To do this, we estimated the SNP correlation (SNP-r_g) for schizophrenia between ethnicities. SNP correlations were estimated as SNP-r_g = 0.66 (SE = 0.23, $p_{(SNP\text{-}rg\ =\ 0)} = 0.0003$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.26$) between MGS ADs and MGS EDs and similarly as SNP-r_g = 0.61 (SE = 0.21, $p_{(SNP\text{-}rg\ =\ 0)} = 0.0003$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.16$) between MGS ADs and ISC EDs. By way of comparison, the SNP correlation between MGS EDs and ISC EDs was estimated as SNP-r_g = 0.83 (SE = 0.09, $p_{(SNP\text{-}rg\ =\ 0)} < 0.0001$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.09$) (Table 2). In follow-up analyses, controlling for 6 rather than 20 PCs made little difference to SNP-heritability or SNP-correlation estimates (Table S3). By excluding one chromosome and rerunning these models 22 times, we found no evidence that any individual chromosome disproportionately influenced SNP-heritability or SNP-correlation estimates (Table S4).

Compared to SNPs with low MAF, SNPs with high MAF typically better tag high-MAF causal variants, which are more likely to predate the African-European divergence. We therefore predicted that the genetic correlation between EDs and ADs would be higher for high-MAF SNPs than for low-MAF SNPs. Consistent with this expectation, SNP-correlation estimates derived from genomic-relationship matrices restricted to SNPs with above-median MAFs (MAF > 0.26) were higher (SNP-r_g = 0.79, SE = 0.28, $p_{(SNP\text{-}rg\ =\ 0)} < 0.0001$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.52$ between MGS ADs and MGS EDs and SNP-r_g = 0.67, SE = 0.27, $p_{(SNP\text{-}rg\ =\ 0)} = 0.0004$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.37$ between MGS ADs and ISC EDs) than SNP-correlation estimates derived from genomic-relationship matrices restricted to SNPs with below-median MAFs (MAF < 0.26, SNP-r_g = 0.34, SE = 0.20, $p_{(SNP\text{-}rg\ =\ 0)} = 0.0706$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.0399$ between MGS ADs and MGS EDs and SNP-r_g = 0.42, SE = 0.18, $p_{(SNP\text{-}rg\ =\ 0)} = 0.0126$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.0321$ between MGS ADs and ISC EDs; Table 3). We observed that SNP heritabilities were lower for above-median-MAF ($h^2 \sim 0.13$) than for below-median-MAF ($h^2 \sim 0.23$) SNPs among MGS ADs, although these estimates were not significantly different from each other and are consistent with random fluctuations in estimated effects across models using different genomic-relationship matrices.

On the other hand, we found no strong evidence supporting the prediction that SNP-correlation estimates would be higher in low-recombination regions (SNP-r_g = 0.60, SE = 0.27, $p_{(SNP\text{-}rg\ =\ 0)} = 0.008$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.28$ between MGS ADs and MGS EDs and SNP-r_g = 0.70, SE = 0.25, $p_{(SNP\text{-}rg\ =\ 0)} = 0.0001$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.34$ between MGS ADs and ISC EDs) than in high-recombination regions (SNP-r_g = 0.64, SE = 0.23, $p_{(SNP\text{-}rg\ =\ 0)} = 0.0005$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.23$ between MGS ADs and MGS EDs and SNP-r_g = 0.41, SE = 0.20, $p_{(SNP\text{-}rg\ =\ 0)} = 0.019$, $p_{(SNP\text{-}rg\ =\ 1)} = 0.054$ between MGS ADs and ISC EDs; Table 3).

Table 3. SNP-Heritability and SNP-Correlation Estimates from Bivariate Models for SNP Bins Based on MAF and Recombination

Sample 1	Sample 2	Sample 1 n	Sample 2 n	Sample 1 h^2 (SE)	Sample 2 h^2 (SE)	cov_{12}[a]	SNP-r_g (SE)	$p_{(\text{SNP-rg} = 0)}$	$p_{(\text{SNP-rg} = 1)}$
High-MAF SNP Bin									
MGS AD	MGS ED	2,142	4,990	0.13 (0.07)	0.27 (0.03)	0.15	0.79 (0.28)	<0.0001	0.52
MGS AD	ISC ED	2,142	6,665	0.12 (0.07)	0.22 (0.02)	0.11	0.67 (0.27)	0.0004	0.37
MGS ED	ISC ED	4,990	6,665	0.22 (0.03)	0.21 (0.02)	0.18	0.82 (0.09)	<0.0001	0.0717
Low-MAF SNP Bin									
MGS AD	MGS ED	2,142	4,990	0.21 (0.09)	0.25 (0.02)	0.08	0.34 (0.20)	0.0706	0.0399
MGS AD	ISC ED	2,142	6,665	0.24 (0.09)	0.22 (0.02)	0.10	0.42 (0.18)	0.0126	0.0321
MGS ED	ISC ED	4,990	6,665	0.19 (0.03)	0.23 (0.02)	0.17	0.83 (0.12)	<0.0001	0.19
High-Recombination SNP Bin									
MGS AD	MGS ED	2,142	4,990	0.20 (0.08)	0.28 (0.03)	0.15	0.64 (0.23)	0.0005	0.23
MGS AD	ISC ED	2,142	6,665	0.20 (0.08)	0.22 (0.02)	0.09	0.41 (0.20)	0.0193	0.0541
MGS ED	ISC ED	4,990	6,665	0.21 (0.03)	0.22 (0.02)	0.17	0.80 (0.11)	<0.0001	0.0836
Low-Recombination SNP Bin									
MGS AD	MGS ED	2,142	4,990	0.13 (0.07)	0.23 (0.03)	0.10	0.60 (0.27)	0.0076	0.28
MGS AD	ISC ED	2,142	6,665	0.15 (0.07)	0.21 (0.02)	0.12	0.70 (0.25)	0.0001	0.34
MGS ED	ISC ED	4,990	6,665	0.19 (0.03)	0.20 (0.02)	0.17	0.86 (0.10)	<0.0001	0.19

Estimates above are from independent analyses in which the following SNP bins were included separately for each model: 238,810 SNPs with high average MAF across ADs and EDs, 238,854 SNPs with low average MAF across ADs and EDs, 238,352 SNPs in high-recombination regions, and 238,352 SNPs in low-recombination regions. Recombination rates were estimated from CEU, YRI, and JPT+CHB HapMap populations, and high- and low-MAF and high- and low-recombination bins were formed with median splits of SNP MAFs and recombination rates, respectively. The Pearson correlation between SNP MAF and recombination rate was 0.006 ($p = 0.0001$).

[a] Calculated by $\text{cov}_{12} = \text{SNP-}r_g h_1 h_2$.

To examine whether some of the genetic overlap in schizophrenia between ED and AD samples might be due to the relatively high level of European admixture in ADs, estimated to be an average of ~15%,[26] we rederived SNP-correlation estimates by using subsamples of ADs with increasingly lower proportions of European ancestry as judged from the first ancestry PC. We found modest but nonsignificant decreases in SNP correlation between ADs and EDs (Table 4). In particular, SNP-r_g was estimated to be 0.46 (se = 0.26, $p_{(\text{SNP-rg} = 0)} = 0.064$, $p_{(\text{SNP-rg} = 1)} = 0.23$) between the least admixed MGS ADs and MGS EDs and to be 0.54 (se = 0.24, $p_{(\text{SNP-rg} = 0)} = 0.006$, $p_{(\text{SNP-rg} = 1)} = 0.19$) between the least admixed MGS ADs and ISC EDs.

Discussion

Using the largest combined AD and ED schizophrenia case-control GWAS available, we have demonstrated an application that uses direct estimates of SNP correlations to quantify the amount of shared genetic variance tagged by SNPs between any two ethnically distinct populations for any trait. We found that common genetic liability to schizophrenia is largely shared across ED and primarily AD populations. Because SNPs included in GWASs are unlikely to tag the effects of rare causal variants and because rarer causal variants are increasingly likely to be population specific, these estimates of overlap are unlikely to apply to the portion of schizophrenia heritability caused by rare causal variants. Although we make no assertion that all estimated gene effects are exactly additive, these results are unlikely to be very biased by nonadditive genetic effects because, if they exist, they contribute very little to the similarity of distantly related individuals.

Even for comparisons of the same trait within an ethnicity, SNP correlations between different samples are typically less than 1 as a result of loss of real signal (e.g., greater ethnic, phenotypic, and environmental homogeneity within samples) and potential artifacts that can inflate the unshared portions of heritability (e.g., SNP calling or plate confounds that differ systematically between cases and controls). Artifacts are expected to deflate, not inflate, SNP correlations because random directional effects of artifacts are not expected to be consistent across data sets. Because MGS ADs, MGS EDs, and ISC EDs were all collected in different labs and genotyped on separate plates, the SNP correlation between MGS EDs and ISC EDs provides an upper limit on the potential SNP correlation between MGS ADs and MGS or ISC EDs. Given that SNP correlations between samples drawn from the same population are expected to be 1 in the absence of artifacts, we estimate that the SNP correlation might be as high as approximately 0.75 (0.61/0.83 or

Table 4. SNP-Heritability and SNP-Correlation Estimates from Bivariate Models after Exclusion of ADs of Increasing European Admixture

Sample 1	Sample 2	Sample 1 n	Sample 2 n	Sample 1 h^2 (SE)	Sample 2 h^2 (SE)	cov_{12}[a]	SNP-r_g (SE)	$p_{(SNP\text{-}rg\,=\,0)}$	$p_{(SNP\text{-}rg\,=\,1)}$
Between ADs and MGS EDs									
MGS AD (all)	MGS ED	2,142	4,990	0.20 (0.09)	0.33 (0.03)	0.17	0.66 (0.23)	0.0003	0.26
MGS AD($z < 2$)	MGS ED	2,044	4,990	0.18 (0.09)	0.33 (0.03)	0.18	0.73 (0.27)	0.0003	0.41
MGS AD ($z < 1$)	MGS ED	1,792	4,990	0.19 (0.10)	0.31 (0.03)	0.16	0.66 (0.28)	0.0025	0.36
MGS AD ($z < 0.5$)	MGS ED	1,557	4,990	0.22 (0.12)	0.30 (0.03)	0.12	0.46 (0.26)	0.0644	0.23
Between ADs and ISC EDs									
MGS AD (all)	ISC ED	2,142	6,665	0.22 (0.09)	0.27 (0.02)	0.15	0.61 (0.21)	0.0003	0.16
MGS AD($z < 2$)	ISC ED	2,044	6,665	0.19 (0.09)	0.28 (0.02)	0.16	0.70 (0.25)	0.0001	0.35
MGS AD ($z < 1$)	ISC ED	1,792	6,665	0.22 (0.10)	0.28 (0.02)	0.16	0.64 (0.24)	0.0006	0.24
MGS AD ($z < 0.5$)	ISC ED	1,557	6,665	0.23 (0.12)	0.28 (0.02)	0.14	0.54 (0.24)	0.0061	0.19

[a] Calculated by $cov_{12} = \text{SNP-}r_g h_1 h_2$.

0.66/0.83) between EDs and ADs when corrected for artifacts that lower genetic correlations. This SNP correlation might be closer to 0.5 (0.46/0.83 or 0.54/0.83) between nonadmixed AD and ED populations.

There are two principal reasons why genetic correlations across ethnicities are expected to be lower than those within ethnicities. First, some causal variants might have different effects across ethnicities (e.g., because of differences in environmental or genetic backgrounds), and some might be population specific if they arose or were lost after the European-African divergence thought to have occurred 50–100K years ago.[28–30] Our finding that high-MAF SNPs show greater SNP correlations than low-MAF SNPs lends some support to the hypothesis that lower-MAF schizophrenia causal variants are increasingly population specific and/or are differentially tagged by SNPs across ethnicities. Second, lower SNP correlations should occur to the degree that LD between causal variants and the SNPs that predict them differs across ethnicity. However, we did not find strong evidence that SNP correlations were lower for high-recombination regions than for low-recombination regions, which might be expected if differing LD patterns reduce the SNP correlation between EDs and ADs. Nonetheless, our results do not allow us to reliably quantify the relative importance of these two alternatives.

Two limitations to our analysis are worth noting. First, the AD sample size in particular was small, leading to large SEs around estimates of both SNP heritabilities and SNP correlations between ethnicities. This means that we cannot place great confidence in the specific estimates of SNP correlation, and we reiterate that these parameters reflect the allelic spectrum captured by common SNPs and so might not generalize to the covariance attributable to variants in low LD with common SNPs. Nevertheless, our results suggest that there is substantial overlap of variants tagged by common SNPs between ethnic groups (with 95% confidence that SNP-r_g has a lower bound of 0.20) and allow us to rule out the possibility of no overlap with even greater confidence. Second, no AD replication GWAS data set of sufficient size exists for this or any other psychiatric disorder, highlighting the importance of further large-scale genome-wide data collection in non-ED samples.

In summary, we have demonstrated how a random-effects modeling approach fitting all SNPs simultaneously can elucidate the degree to which additive genetic variation tagged by SNPs is shared between ethnically divergent populations. We estimate that about half to three-quarters of such additive genetic variation underlying the risk of schizophrenia is shared between AD and ED populations, suggesting that schizophrenia GWAS discoveries from European samples are likely to be relevant to AD populations and that meta-analyses of schizophrenia can improve power by including results across both ED and AD samples. These conclusions should apply even more strongly to other ethnic groups given their more recent divergence times from the European lineage. Because it is vanishingly unlikely that different causal variants between ethnicities would systematically be in LD with the same SNPs, these findings suggest that many schizophrenia causal variants are ancient and predate European-African divergence.

Supplemental Data

Supplemental Data include Supplemental Acknowledgments, three figures, four tables, and full membership lists for the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia Collaboration and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

The authors acknowledge funding from the National Institute of Mental Health (grants K01MH085812 and R01MH100141 to M.C.K., 5T32MH016880-30 to J.K.H., and R01HG004960 and R01HG005701 to B.L.B.), the Australian National Health and Medical Research Council (grants 389892, 442915, 496688, 613672, 613601, 1047956, and 1052684), and the Australian Research Council (grants DP0770096, DP1093502, FT0991360, and DE130100614). The Molecular Genetics of Schizophrenia data set described in this manuscript was obtained from dbGaP, found at http://www.ncbi.nlm.nih.gov/gap, through study accession numbers phs000021.v3.p2 and phs000167.v1.p1. This manuscript does not necessarily reflect the opinions or views of the funding agencies or institutes who helped in collection of this data. Acknowledgments for the International Schizophrenia Consortium and the Molecular Genetics of Schizophrenia Collaboration can be found in the Supplemental Data.

Received: March 21, 2013
Revised: June 3, 2013
Accepted: July 2, 2013
Published: August 15, 2013

References

1. Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch. Gen. Psychiatry *60*, 1187–1192.
2. Cardno, A.G., and Gottesman, I.I. (2000). Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. Am. J. Med. Genet. *97*, 12–17.
3. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R.; Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ); International Schizophrenia Consortium (ISC); Molecular Genetics of Schizophrenia Collaboration (MGS). (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.
4. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
5. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D., Altshuler, D.L., et al.; 1000 Genomes Project. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA *108*, 11983–11988.
6. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.
7. Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072–1079.
8. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.
9. Haga, S.B. (2010). Impact of limited population diversity of genome-wide association studies. Genet. Med. *12*, 81–84.
10. Weiss, K.M., and Clark, A.G. (2002). Linkage disequilibrium and the mapping of complex human traits. Trends Genet. *18*, 19–24.
11. Zuo, L., Zhang, C.K., Wang, F., Li, C.-S.R., Zhao, H., Lu, L., Zhang, X.-Y., Lu, L., Zhang, H., Zhang, F., et al. (2011). A novel, functional and replicable risk gene region for alcohol dependence identified by genome-wide association study. PLoS ONE *6*, e26726.
12. Fesinmeyer, M.D., North, K.E., Ritchie, M.D., Lim, U., Franceschini, N., Wilkens, L.R., Gross, M.D., Bůžková, P., Glenn, K., Quibrera, P.M., et al. (2013). Genetic risk factors for BMI and obesity in an ethnically diverse population: results from the population architecture using genomics and epidemiology (PAGE) study. Obesity (Silver Spring) *21*, 835–846.
13. Chang, M.H., Ned, R.M., Hong, Y., Yesupriya, A., Yang, Q., Liu, T., Janssens, A.C.J.W., and Dowling, N.F. (2011). Racial/ethnic variation in the association of lipid-related genetic variants with blood lipids in the US adult population. Circ Cardiovasc Genet *4*, 523–533.
14. Waters, K.M., Stram, D.O., Hassanein, M.T., Le Marchand, L., Wilkens, L.R., Maskarinec, G., Monroe, K.R., Kolonel, L.N., Altshuler, D., Henderson, B.E., and Haiman, C.A. (2010). Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet. *6*, e1001078.
15. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7–24.
16. Ikeda, M., Aleksic, B., Kinoshita, Y., Okochi, T., Kawashima, K., Kushima, I., Ito, Y., Nakamura, Y., Kishi, T., Okumura, T., et al. (2011). Genome-wide association study of schizophrenia in a Japanese population. Biol. Psychiatry *69*, 472–478.
17. Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502–510.
18. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540–2542.
19. Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P.A., Whittemore, A.S., Mowry, B.J., et al. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature *460*, 753–757.
20. Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.-Y., Duan, J., Ophoff, R.A., Andreassen, O.A., et al.; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.
21. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
22. Shriner, D., Adeyemo, A., Chen, G., and Rotimi, C.N. (2010). Practical considerations for imputation of untyped markers in admixed populations. Genet. Epidemiol. *34*, 258–265.
23. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
24. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
25. McGrath, J., Saha, S., Chant, D., and Welham, J. (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiol. Rev. *30*, 67–76.
26. Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan,

B., et al. (2009). Characterizing the admixed African ancestry of African Americans. Genome Biol. *10*, R141.

27. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al.; International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
28. Templeton, A. (2002). Out of Africa again and again. Nature *416*, 45–51.
29. Sved, J.A. (2009). Linkage disequilibrium and its expectation in human populations. Twin Res. Hum. Genet. *12*, 35–43.
30. Sved, J.A., McRae, A.F., and Visscher, P.M. (2008). Divergence between human populations estimated from linkage disequilibrium. Am. J. Hum. Genet. *83*, 737–743.