Molecular Validation of the Schizophrenia Spectrum

T. Bernard Bigdeli¹, Silviu-Alin Bacanu¹, Bradley T. Webb¹, Dermot Walsh², F. Anthony O’Neill³, Ayman H. Fanous¹,⁴, Brien P. Riley¹,⁵, and Kenneth S. Kendler*¹,⁵

¹Department of Psychiatry, Virginia Commonwealth University School of Medicine, Richmond, VA; ²Health Research Board, Dublin, Ireland; ³Centre for Public Health, Queen’s University, Belfast, UK; ⁴Mental Health Service Line, Washington VA Medical Center, Washington, DC; ⁵Department of Human and Molecular Genetics, Virginia Commonwealth University School of Medicine, Richmond, VA

*To whom correspondence should be addressed; Department of Psychiatry, Virginia Institute for Psychiatric and Behavioral Genetics of VCU, Box 980126, Richmond, VA 23298-0126, US; tel: (804)-828-8590, fax: (804)-828-1471, e-mail: kendler@vcu.edu

Background: Early descriptive work and controlled family and adoption studies support the hypothesis that a range of personality and nonschizophrenic psychotic disorders aggregate in families of schizophrenic probands. Can we validate, using molecular polygene scores from genome-wide association studies (GWAS), this schizophrenia spectrum? Methods: The predictive value of polygenic findings reported by the Psychiatric GWAS Consortium (PGC) was applied to 4 groups of relatives from the Irish Study of High-Density Schizophrenia Families (ISHDSF; N = 836) differing on their assignment within the schizophrenia spectrum. Genome-wide single nucleotide polymorphism data for affected and unaffected relatives were used to construct per-individual polygene risk scores based on the PGC stage-I results. We compared mean polygene scores in the ISHDSF with mean scores in ethnically matched population controls (N = 929). Results: The schizophrenia polygene score differed significantly across diagnostic categories and was highest in those with narrow schizophrenia spectrum, lowest in those with no psychiatric illness, and in-between in those classified in the intermediate, broad, and very broad schizophrenia spectrum. Relatives of all of these groups of affected subjects, including those with no diagnosis, had schizophrenia polygene scores significantly higher than the control sample. Conclusions: In the relatives of high-density families, the observed pattern of enrichment of molecular indices of schizophrenia risk suggests an underlying, continuous liability distribution and validates, using aggregate common risk alleles, a genetic basis for the schizophrenia spectrum disorders. In addition, as predicted by genetic theory, non-psychotic members of multiply-affected schizophrenia families are significantly enriched for replicated, polygenic risk variants compared with the general population.

Key words: schizophrenia/schizophrenia spectrum/polygene score/GWAS

Both Kraepelin and Bleuler in their classic descriptions of, respectively, dementia praecox and schizophrenia,² observed that close relatives of individuals with schizophrenia have excess rates of unusual personalities, features of which appeared to be mild versions of many of the symptoms seen in their more clearly affected relative.³ Kraepelin noted many “eccentric personalities” in the families of his patients with dementia praecox and considered it likely that they suffered from a latent version of the same “principle malady.”¹² Bleuler articulated a similar view as follows:

If one observes the relatives of our patients [with schizophrenia], one finds in them peculiarities which are qualitatively identical with those of the patients themselves, so that the [patient’s] disease appears to be only a quantitative increase of the anomalies seen in the parents and siblings.²³

In the first series of systematic family studies of schizophrenia conducted in Kraepelin’s Psychiatric Research Institute early in the 20th century, it was already noted that a range of disorders other than classical schizophrenia aggregated in relatives including atypical psychoses and “schizoidia” or schizoid personality.⁴ In his very large classical family study of schizophrenia published in 1938, Kallmann articulated what he alternatively called the “group of schiziform abnormalities” or the “schizophrenic disease-complex.”⁵

The modern concept of the schizophrenia spectrum derives from the Danish Adoption Study of Schizophrenia,⁶ where biological relatives of schizophrenic adoptees were shown to be at increased risk both for classical schizophrenia as well as milder syndromes, alternatively called latent, borderline, or uncertain schizophrenia.⁸ Interviews from this study were used to develop the diagnostic category of schizotypal personality disorder in DSM-III.⁹
Several modern family studies as well as a diagnostic re-review of the Danish Adoption Study findings using operationalized diagnostic criteria and blind diagnosis verified the genetic/familial association between schizophrenia and both personality disorders (especially schizotypal and paranoid) and nonpsychotic psychotic disorders, especially schizoaffective disorder and atypical psychosis.10–15

In the most systematic effort to clarify the nature of the schizophrenia spectrum, Kendler and colleagues applied a multiple threshold model to results from the Roscommon Family Study.16 The model postulated positions on a single dimension of genetic/familial liability for the key diagnostic categories in the following order: schizophrenia, schizoaffective disorder, schizotypal/paranoid personality disorder, other nonaffective psychoses, and psychotic affective illness. This model fit the family data well but the article concluded:

While schizophrenia and psychotic affective illness could be clearly assigned to the two extremes of the schizophrenia spectrum, the proper ordering of schizoaffective disorder, schizotypal/paranoid personality disorder, and other nonaffective psychosis could not be unambiguously determined.16,26

Attempts to verify the schizophrenia spectrum using molecular tools have been more limited. One notable effort utilized linkage in the Irish Study of High-Density Schizophrenia Families (ISHDSF).17 Genome-wide, limit of detection (LOD) scores for schizotypy in non-psychotic members of these families were significantly correlated with LOD scores of schizophrenia in all pedigree members,18 suggesting a sharing of at least some risk variants across the schizophrenia spectrum. Two studies have found that a variant in the risk gene ZNF804A—also associated with schizotypal personality traits21,22—has been observed in several studies to alter liability to schizophrenia19,20 and psychotic affective illness.21

Technological advances have now permitted another look at the genetic relationship between disorders within the schizophrenia spectrum. Using genome-wide association studies (GWAS), polygene scores can be statistically constructed that reflect a broad array of measured molecular variants that in aggregate predispose to illness, typically to high levels of statistical significance, but accounting for small overall proportions of variance.23 In this report, we examine polygene scores for schizophrenia based on GWAS in members of the ISHDSF. Our prediction is that these scores will be highest in members of these pedigrees with classical schizophrenia, lower in members affected with other disorders in the putative schizophrenia spectrum, and lowest in unaffected members of these pedigrees. In addition, we predict that even the unaffected members of these pedigrees—chosen to contain at least two individuals with schizophrenia or poor-outcome schizoaffective disorder (PO-SAD)—will have elevated polygene scores for schizophrenia compared with a control sample.

Methods

Subjects and Single Nucleotide Polymorphism Genotyping

Irish Study of High Density Schizophrenia Families. Fieldwork for the ISHDSF was conducted between April 1987 and November 1992, with probands ascertained from public psychiatric hospitals in Ireland and Northern Ireland with local ethics approval.17 Selection criteria were two or more first-degree relatives meeting DSM-III-R criteria for schizophrenia or PO-SAD. Diagnoses were based on the Structured Interview for DSM-III-R Diagnosis24 and the Structured Interview for Schizotypy.25 Relatives suspected of psychotic illness were always interviewed by trained psychiatrists, and trained social workers would typically interview other relatives. Hospital and/or outpatient records were obtained and abstracted in over 98% of cases with schizophrenia or SAD diagnoses. Family history diagnoses on first-degree relatives by the Family History-Research Diagnostic Criteria (FH-RDC) instrument26 were also routinely obtained.

Independent review of all pertinent diagnostic information (interview, hospital abstract, and family history reports) was made blind to pedigree assignment and marker genotypes by K.S.K. and D.W., with each diagnostician making up to 3 best-estimate DSM-III-R diagnoses. Agreement between the 2 diagnosticians was high (weighted \( \kappa = 0.94 \pm 0.05 \)). Our diagnostic schema included 10 categories ranked by the degree to which they reflected the core vs periphery of the schizophrenia spectrum based on our review of prior family and adoption studies and, in particular, the Roscommon Family Study done by the same research team in the same country using similar diagnostic procedures. For most subsequent analyses, these 10 categories were divided into 4 groups: (1) narrow—schizophrenia, PO-SAD, and simple schizophrenia; (2) intermediate—schizotypal personality disorder, schizophreniform and delusional disorders, atypical psychosis, and good-outcome SAD; (3) broad—psychotic affective illness and paranoid, avoidant, and schizoid personality disorder; and (4) very broad—any other psychiatric illness, particularly nonpsychotic major depression, anxiety disorders, and alcohol dependence. The sample included, of course, individuals who received no lifetime diagnosis of psychiatric illness.

In total, 853 individuals representing 237 pedigrees were selected from the ISHDSF for high-throughput genotyping on the Illumina 610-Quad platform (at Illumina, Inc), and genotypes were called with the BeadStudio software package (Illumina, Inc). Initial exclusion criteria for samples was call rate below 95%. Exclusion criteria for single nucleotide polymorphisms (SNPs) were as follows: third allele observed; pseudautosomal or mitochondrial;
minor allele frequency (MAF) <1%; call rate <98%; and GenCall10 quality score <0.55. Following lift-over to the most recent genome assembly (GRCh37.2), 557 373 autosomal SNPs were available for analysis; genotyping completion was greater than 99.9%.

In order to investigate the possibility of duplicated or erroneously identified DNAs, we compared estimates of genetic relatedness (identity-by-descent) generated using PLINK\textsuperscript{27} against expectation based on known familial relationships. The majority of observed inconsistencies were instances of a single, duplicated sample labeled falsely as representing an affected sib-pair. For sex-disscordant sib-pairs, the true identity of a duplicate sample was resolved by consideration of X-chromosome genotypes. Following exclusion of problematic samples, a total of 843 individuals representing 237 pedigrees remained for analysis.

**Population-Based Controls.** Details of recruitment, screening, and quality control (QC) methods employed in the Wellcome Trust Case Control Consortium (WTCCC) 2 have been published elsewhere.\textsuperscript{26} Controls (\(N = 2048\)) were ascertained with written informed consent from the Irish GeneBank and represented blood donors from the Irish Blood Transfusion Service recruited in the Republic of Ireland. All controls were of Irish origin (born of Irish parents and having all 4 grandparents born in Ireland or the UK) but were not specifically screened for psychiatric illness. Individuals taking regular prescribed medication were excluded from blood donation in Ireland, and donors were not financially remunerated. Meaningful misclassification bias is unlikely because the lifetime prevalence of schizophrenia is relatively low (<1%), and affected individuals typically chronically take medication and probably have a reduced likelihood of volunteering for blood donations. Following QC, the control sample included 1794 participants.

All samples were genotyped using the Affymetrix 6.0 platform totaling 893 634 autosomal SNPs, either at the Affymetrix or Broad Institute laboratories. A subset of samples (\(N = 37\)) was genotyped at both sites and successfully identified as duplicates as part of the QC procedure. For all samples passing laboratory QC, raw intensities (from the CEL files) were renormalized within collections using CelQuantileNorm (see http://outmod-edbonsai.sourceforge.net/). Genotype-calling was performed by GoldenHelix, Inc, utilizing BeagleCall\textsuperscript{29} in an iterative procedure that used normalized genotypes and genotype probabilities generated by BirdSeed (see http://www.broadinstitute.org/mpg/birdsuiCte/birdseed.html) as inputs. Only those individual genotypes with genotype probability of at least 0.97 were retained. Exclusion criteria for SNPs were as follows: third allele observed; pseudoautosomal, mitochondrial, or X/Y; MAF <1%; call rate <97%; and \(P < 1 \times 10^{-6}\) for deviations from Hardy-Weinberg equilibrium. A total of 686 646 autosomal SNPs were available for analysis; genotyping completion was greater than 99.9%.

To assess relatedness among study individuals, we examined estimates of genetic relatedness (identity-by-descent) generated using PLINK\textsuperscript{27} for unrelated sample pairs with allele-sharing greater than 5%; samples involved in more than one such relationship were excluded outright, else one sample in a pair was excluded at random. Given the potential of genotyping site effects to manifest in an aggregate score (data not shown), we elected to include in our final analysis only those controls genotyped at the Affymetrix site (\(N = 929\)).

**Genome-Wide Imputation**

Phasing of entire chromosomes or chromosome arms was performed using SHAPEIT.\textsuperscript{30} Genotype imputation of additional SNPs was carried out with IMPUTE2 v.2.0\textsuperscript{31} using the March 2012 release (v3) of the 1000 Genomes Project data (www.1000genomes.org).\textsuperscript{32} Imputation analysis was performed for genomic windows of 5 Mb with an overlap interval of 500 Kb between adjacent segments. Following the recommendations of the authors of IMPUTE2, we did not limit our imputation procedure to European reference samples.\textsuperscript{33} SNPs were filtered using an information threshold of 0.5 and only those imputed genotypes with a corresponding probability of at least .95 were retained. Phasing, imputation, and postprocessing were performed independently for the ISHDSF and WTCCC samples. Within each study, monomorphic sites were excluded, as well as any SNP that yielded at least one Mendelian inconsistency.

**Construction of Polygene Score**

We constructed a polygene score based on results from the discovery phase (\(N = 21,856\)) of the Psychiatric GWAS Consortium (PGC) of SCZ.\textsuperscript{34} SNPs present on Illumina 610-Quad or imputed successfully (information >.5; MAF ≥2% in \(N = 286\) independent ISHDSF samples; genotype completion ≥99%) were extracted from the postimputation ISHDSF data set. An identical filtering procedure was performed for Affymetrix 6.0 and imputed SNPs in the WTCCC sample. Of 1.25 M SNPs in the PGC stage-I analysis, 540 576 met our filtering criteria in both cohorts. In constructing the initial polygene score, we identified 289 541 SNPs that yielded a \(P\) value of less than .2 in the PGC stage-I results. Although the choice of \(P\) value threshold is somewhat arbitrary, inclusion of a relatively large number of SNPs is preferable. We decided to use a threshold of \(P < .2\) because it approximates a \(P\) value of .159, which corresponds to the inclusion threshold from Akaike information criterion.\textsuperscript{35} Of these, 137 957 SNPs were available in both cohorts following postimputation filtering. Let \(G_{ij}\) denote the additive genotype for \(i\)th SNP in the \(j\)th subject; let \(p_i\) and \(\theta_i\)
Molecular Validation of the Schizophrenia Spectrum

denote the frequency and OR of the associated allele for the ith SNP. Using an independent subset of k SNPs \((r^2 < .2)\), we constructed a per-individual score, \(S_i\), as 
\[
S_i = \sum_{j=1}^{k} (G_{ij} - 2p_j) \log(\theta_j).
\]
We obtained normalized scores by randomly selecting an individual from each family and calculating the SD; this procedure repeated 1000 times to obtain a final estimate of the SD. For each categorical diagnosis in ISHDSF, we randomly selected an affected individual from each pedigree (where applicable) to obtain an estimate of the mean normalized polygene score; we repeated this procedure 1000 times to obtain a final estimate of the mean.

**Determination of Empirical Significance**

In order to assess the empirical significance of the observed polygene score distribution in the ISHDSF, we generated 1000 null replicates of the PGC stage-I results for SCZ. We simulated \(N = 20000\) diploid individuals from \(N = 379\) European individuals in the April 2012 release of the 1000 Genomes Project data. Autosomal SNP data for simulated individuals was obtained by sampling 44 times (with replacement) from \(379 \times 2 = 758\) reference haplotypes.

Assignment of disease status was random: 10 000 “affected” and 10 000 “unaffected” individuals. These phenotypic definitions were then shuffled 1000 times. Each null phenotype was analyzed using the logistic regression procedure in PLINK\(^27\), as to obtain a direct estimates of the OR. From each set of null association results, we extracted SNPs yielding \(P\) values in the first quintile \((P < .2)\), as described above.

For each null replicate, we used PLINK to obtain an LD-pruned set of independent SNPs based on a panel consisting of available WTCCC samples and \(N = 286\) independent ISHDSF samples. These pruned SNP sets served as the basis for calculating 1000 null polygene scores. In calculating null polygene scores, we used the associated allele frequencies reported in the PGC stage-I results. Empirical significance was defined as the proportion of null replicates yielding a mean difference in polygene score greater than the observed difference between ISHDSF and population controls, \((r + 1)/(n + 1)\).\(^{36}\) If following 1000 permutations, no replicate more significant than the actual finding was observed, we report a conservative estimate of the empirical \(P\) value, ie, \(P \leq .002\).

**Results**

*Polygene Score by Categorical Diagnosis in the Schizophrenia Spectrum*

The numbers of relatives with polygene schizophrenia scores in each of our 5 diagnostic classes were narrow—\(N = 432\) individuals in 231 families; intermediate—\(N = 95\) individuals in 75 families; broad—\(N = 33\) individuals in 31 families; very broad—\(N = 59\) individuals in 49 families; and unaffected—\(N = 217\) individuals in 142 families. The sample sizes in the broad and very broad categories were too small to meaningfully examine on their own and so were combined for subsequent analyses (\(N = 92\) individuals in 71 families).

The mean schizophrenia polygene scores (and \(SEs\)) in each of the 4 diagnostic classes of relatives from the high-density pedigrees are displayed in figure 1. The observed pattern is consistent with that expected under the hypothesis of the schizophrenia spectrum with the narrow category having the highest score and the score decreasing in a nearly monotonic function across the other diagnostic categories. A one-way ANOVA, the single best test for the hypothesized schizophrenia spectrum, but unadjusted for relatedness of subjects within and across diagnostic categories, indicated a significant difference among these diagnostic classes \((P = .03)\). The mean polygene score was highest in the narrow category \((Z = 1.59, 95\%\ CI = [1.47, 1.72])\), similar in the intermediate and broad/very broad categories \((Z = 1.47, [1.24, 1.69] \text{ and } Z = 1.48, [1.23, 1.73],\) respectively) and lowest in the family members without any psychiatric diagnosis \((Z = 1.32, [1.17, 1.47])\).

Pair-wise comparisons of diagnostic categories, adjusted for pedigree structure using a random-effects model, indicated significant differences (by a likelihood ratio test) between the narrow category and the intermediate, broad/very broad, and unaffected categories (one-sided \(P\) values of .022, .042, 7.68 \times 10^{-4}, respectively); nonsignificant
differences between the intermediate category and broad/very broad \((P = .431)\) or unaffected categories \((P = .216)\); and a nonsignificant difference between the broad/very broad and unaffected categories \((P = .141)\).

**Relative Enrichment Compared With Population Controls**

The mean schizophrenia polygene scores for the 929 Irish controls was estimated at \(Z = 0.96, (0.90, 1.02)\). As calculated from 1000 permutations, this score was significantly lower than those observed in any of the 4 of the diagnostic classes of relatives from the ISHDSF \((P < .002)\). We estimated the variance explained by the polygene score as 6.33\% for the narrow category, 1.87\% and 1.85\% for the intermediate and broad/very broad categories, and 1.61\% for unaffected relatives. A nearly identical pattern of findings was observed for polygene scores based on \(P\) value thresholds of \(0.1\) and \(0.05\) although the variance explained by these scores was lower at progressively more stringent thresholds.

**Discussion**

We sought in this article to validate, using new molecular methods, the concept of a genetically influenced schizophrenia spectrum, the origins of which can be traced back to beginnings of modern psychiatry in the late 19th and early 20th century. Our results replicated earlier family and adoption studies showing a continuum of genetic risk, as assessed by a polygene risk score based on variants found to be associated with schizophrenia in GWAS data from the PGC schizophrenia consortium,\(^1\) in blindly diagnosed members of Irish pedigrees selected for a high risk for schizophrenia. Members of those pedigrees with disorders in the narrow schizophrenia spectrum had the highest polygene risk scores, followed by relatives with diagnoses considered to be in the intermediate, broad, and very broad part of the spectrum. Both of these groups of relatives had schizophrenia polygene scores higher than relatives in these pedigrees judged to be free of any psychiatric illness.

This report can be viewed as a direct attempt to verify our prior modeling of the schizophrenia spectrum in the Roscommon Family Study.\(^6\) Of interest, in that report, we could with some confidence, determine the low- and high-risk ends of the schizophrenia spectrum but were less certain about the proper ordering of disorders in the middle. Our findings, based on molecular variants rather than statistical patterns of occurrence and cooccurrence in relatives, reached a similar conclusion.

In the relatives of these high-density families, the observed pattern of enrichment of molecular indices of the risk to schizophrenia is congruent with an underlying, continuous liability distribution and, we submit, represents preliminary, molecular validation of a common genetic basis for the schizophrenia spectrum disorders.

We also compared the polygene scores calculated in the diagnostic classes of the high-density schizophrenia families with those obtained from a large group of ethnically matched controls. All diagnostic classes in the high-density families were highly significantly enriched for polygene schizophrenia risk. As would be expected for any reasonable model of genetic transmission, individuals who are unaffected but close relatives of multiple individuals with schizophrenia or schizophrenia spectrum disorders are at an elevated genetic risk.

**Limitations**

These results should be interpreted in the context of 3 potentially important methodological limitations. First, the members of familial samples were selected for high-throughput genotyping based on the informativeness of their genetic relationships for association analysis. That is, schizophrenic probands and their unaffected first-degree relatives were preferentially selected for inclusion in a primary GWAS of schizophrenia. We therefore do not have samples on all members of these pedigrees available for molecular analysis. Second, ascertainment for the ISHDSF required at least 2 first-degree relatives with a primary diagnosis of schizophrenia or simple schizophrenia; eligibility did not require the incidence of schizophrenia spectrum disorders in a given pedigree. Therefore, the various diagnostic categories considered herein are not equivalently represented across all pedigrees, with a substantial proportion of spectrum cases originating from a subset of larger families and multiplex sibships. Third, the schizophrenia polygene scores utilized in these analyses account for only a modest proportion of total disease variance and surely do not index all aspects of the genetic risk for schizophrenia. For example, genetic risks that arise from rare variants or large structural genomic changes (ie, “copy number variants”) are not well represented in the polygene score. We cannot be certain if their inclusion would result in a different pattern of findings.

**Funding**

The United States National Institutes of Health (MH-41953, MH-083094, and MH-068881).

**Acknowledgment**

The authors have declared that there are no conflicts of interest in relation to the subject of this study.

**References**


