Schizophrenia Bulletin vol. 36 no. 1 pp. 14–23, 2010
doi:10.1093/schbul/sbp137
Advance Access publication on December 8, 2009

Narrowing the Boundaries of the Genetic Architecture of Schizophrenia

Naomi R. Wray[1,2] and Peter M. Visscher[2]

[2]Genetic Epidemiology and Queensland Statistical Genetics, Queensland Institute of Medical Research, 330 Herston Road, Brisbane 4029, Australia

Genetic architecture of a disease comprises the number, frequency, and effect sizes of genetic risk alleles and the way in which they combine together. Before the genomic revolution, the only clue to underlying genetic architecture of schizophrenia came from the recurrence risks to relatives and the segregation patterns within families. From these clues, very simple genetic architectures could be rejected, but many architectures were consistent with the observed family data. The new era of genome-wide association studies can provide further clues to the genetic architecture of schizophrenia. We explore models of genetic architecture by description rather than the mathematics that underpins them. We conclude that the new genome-wide data allow us to narrow the boundaries on the models of genetic architecture that are consistent with the observed data. A genetic architecture of many common variants of moderate (relative risk > approximately 1.2) can be excluded, yet there is evidence that current generation genome-wide chips do tag an important proportion of the genetic variation for schizophrenia and that the underlying causal variants will include common variants of small effect as well as rarer variants of larger effect. Together, these observations imply that the total number of genetic variants is very large—of the order of thousands. The first generation of studies have generated hypotheses that should be testable in the near future and will further narrow the boundaries on genetic architectures that are consistent with empirical data.

Key words: schizophrenia/complex genetic disease/genetic architecture/polygenic

Introduction

Family history is the most important risk factor for schizophrenia,[1] consistent with a genetic contribution to its etiology. Traditionally, researchers, eg, McGue et al,[2,3] and Risch,[4] employed genetic modeling to see if they could gain insight into the genetic architecture of schizophrenia. They compared patterns of recurrence risks for different types of relatives with those expected under different genetic models. They not only found that multiple genetic models of genetic architecture were consistent with observations but also showed that some simple models could be rejected. The advent of genome-wide association studies (GWAS) has allowed the identification of individual genetic risk loci or at least markers linked to them. This article explores if the new evidence provided from GWAS provides further clues to elucidate an understanding of the genetic architecture of schizophrenia. In doing this, we fully acknowledge the sentiment of the industrial statistician George Box that "all models are wrong, but some are useful." Simple models of genetic architecture allow us to devise hypotheses that can be tested against observable data. A model is useful until observable data allow it to be rejected, thereby narrowing the boundaries on models that remain consistent with observations. In this exploration, we aim to minimize, where possible, the presentation of detailed mathematical foundations presented elsewhere (N.R.W. and M.E. Goddard, PhD, unpublished data, 2009).[5]

What Is Genetic Architecture?

To a complex trait geneticist, genetic architecture comprises 4 factors.

1. The number (n) of risk alleles that contribute to disease in the population, which could include multiple risk alleles within a gene.
2. The frequency (q_i) in the population of each of the risk alleles ($i = 1 \ldots n$), which can be either the major or minor allele.
3. The effect size of a risk allele that encompasses the concept of penetrance.
4. The way in which risk alleles act together, additively or with interaction.

[1]To whom correspondence should be addressed; tel: +61-7-3845-3581, fax: +61-7-3362-0101, e-mail: naomi.wray@qimr.edu.au

Both the way in which effect sizes are described and the way in which we describe the interaction of risk alleles depend on the scale of measurement. We will use the term the "risk scale" to mean the observed scale of disease. On this scale, the phenotypic risk is either affected or not affected, but the genetic risk can be expressed as a probability dependent on the genotype of an individual. On this scale, effect sizes can be expressed as either the relative risk or odds ratio of a risk allele for disease. Risk alleles that combine with interaction (or epistasis) on one scale can combine additively on a transformed scale. Therefore, the same genetic architecture can be described as either multiplicative or additive and leads to confusion because the scale of measurement is often not explicitly stated. Heritability of schizophrenia is usually described on the "liability" scale; on this notional scale, risk alleles are usually assumed to combine additively so that the genetic variance (V_G) is the sum of the variances contributed by each variant,[6]

$$V_G = \sum_{i=1}^{n} 2q_i(1 - q_i)a_i^2, \quad [1]$$

where a_i is the effect size of a risk allele, using "a" to emphasize additive action on this scale. From this equation, we see that for a given effect size of a risk allele (ie, a_i), one with frequency $q_i = 0.5$ will contribute the maximum variance compared with risk variants of higher or lower frequency. Rare variants (ie, q_i close to zero) individually contribute little to the overall variance because $q_i(1 - q_i)$ is also close to zero. Infinite combinations of n, q_i, and a_i can generate the same V_G. If effect sizes are small or if risk variants are rare, then the number of risk loci must be large to account for the genetic variance that we know exists from family studies.

The Evidence for a Genetic Etiology of Schizophrenia

Adoption studies[7,8] and recurrence risk to relatives (table 1) provide direct evidence for a genetic etiology of schizophrenia. From disease prevalence (estimated as 0.72%[9] lifetime morbidity risk) and recurrence risks, we can estimate heritability on the liability scale.[10] Heritability is high: A meta-analysis of twin studies estimated heritability to be 81% (95% confidence interval (CI) = 73%–90%),[11] and the estimate from a Swedish study matching >7 million records from the national multigeneration database with hospital discharge records was 64.3% (95% CI = 61.7%–67.5%).[12] Other risk factors include male gender, advanced paternal age, perinatal events, and recreational drug use (reviewed in Sullivan[1] and Tandon et al[13]) . In addition, shared family environment effects are estimated to be small but significant; they explain 11% (95% CI = 3%–19%)[11] and 4.5% (95% CI = 4.4%–7.4%)[12] of the variance in the twin study meta-analysis and Swedish study, respectively.

Despite the high heritability of schizophrenia, only a small proportion of cases have a family history of schizophrenia, reported to be less than one-third[13] and estimated to be only 3.81% (95% CI = 3.62–4.00) in the Swedish national study[12] (counting all identified first-, second-, and third-degree relatives). Genotyping studies have provided some direct evidence for a genetic etiology for schizophrenia and direct evidence for genetic architecture either in linkage studies[14] or association studies particularly genome-wide studies using both single-nucleotide polymorphisms (SNPs) and copy number variants (CNVs, submicroscopic structural variants including insertions and deletions).[15–20] Next generation sequencing studies are expected to provide evidence for the relative importance of rare variants.

The Models Rejected Before the Era of GWAS

Although rare variants of large effect size do exist, such as the translocation that disrupts 2 overlapping brain expressed genes on chromosome 1 (*DISC1* and *DISC2*) in a Scottish pedigree,[21] and major chromosomal abnormalities are present in a small proportion of cases (reviewed in Tandon et al[13]), these are very much the exception rather than the rule, and few large pedigrees have been identified. Therefore, the simplest genetic architectures were rejected, recognizing that the observed recurrence risks and segregation patterns within families could not be explained either by a single genetic locus[22] nor by multiple single loci.[4] They also could not be explained by models in which risk loci combine their effects additively on the scale of risk.[4] Knowledge of family history, differentiating between so-called sporadic and familial cases, was found not to be useful in understanding the etiology of schizophrenia,[23,24] and indeed, sporadic cases are expected to be the norm in complex genetic diseases of low prevalence without invoking new mutations of large effect.[25] However, beyond these broad exclusions, a number of genetic models could all explain the observed recurrence risks: either oligogenic models of a few risk loci each of relatively large effect or polygenic models of a large number of loci each of smaller effect. Before imposing the new information generated from GWAS, we will explore some of these genetic models.

Visualizing a Multiplicative Model

We take over from where Risch[4] left off, that the genetic architecture of schizophrenia must be represented at least by a few risk variants that combine in a multiplicative way on the risk scale in order to generate a pattern of recurrence risk to relatives similar to that observed. Can we visualize this multiplicative model? Figure 1 shows simple multiplicative models that are approximately consistent with schizophrenia in that disease prevalence is approximately 0.72% and heritability approximately 0.7. It shows 3 relationships with number of risk loci on the x-axis. Firstly, the bell-shaped curve shows the probability distribution of individuals in the population having x risk alleles; for x-axis a), 50 binomially distributed loci take on a normal distribution about a mean of 2 × 50 ×

Table 1. Observed Recurrence Risks to Relatives and Those Predicted Under the Liability Threshold Model

	Observed			Predicted Using Liability Threshold Model[a]	
	Risch[4] Based on McGue et al[2]	Lichtenstein et al[30]			
		Estimates	95% Confidence Intervals	Using Prevalence and Sibling Risk of Risch[4]	Using Prevalence and Heritability of Lichtenstein et al[30]
Lifetime prevalence (%)	0.85	0.407		0.85[b]	0.407[c]
Recurrence risks					
Parent		9.43	8.26–10.8	8.6	8.6
Offspring	10.0	10.3	8.76–12.2	8.6	8.6
Offspring of 2 affected parents		89	18.8–672	41	44
Full-sibs	8.6	8.55	7.61–9.60	8.6	8.6
Dizygotic twins	14.2			8.6	8.6
Half-sibs	3.5	2.52	1.56–4.05	3.4	3.3
Nephew/nieces	3.1	2.71	2.22–3.21	3.4	3.3
Grand children	3.3	2.95	1.81–4.81	3.4	3.3
Uncles/aunts	3.2	3.04	2.39–3.87	3.4	3.3
Grand parents		3.8	2.75–5.26	3.4	3.3
First cousin	1.8	2.29	1.71–3.07	1.9	1.9
Monozygotic	52.1			37	38
Proportion of individuals with affected family members	—[d]	3.81[e]	3.62–4.00	32[f]	17[f]

[a]From simulations of 10^6 three generation families. Phenotypes (Y) of liability simulated as Y = A + E, where A is additive genetic and E is environmental component. E simulated as $E \sim N(0, 1 - h^2)$, where h^2 is heritability of liability. For founders, $A \sim N(0, h^2)$; for nonfounders, $A = 1/2A_{mum} + 1/2A_{dad} + A_w$, where $A_w \sim N(0, 1/2h^2)$. Individuals are diseased if Y > T, where T truncates the normal distribution at the proportion defined by the disease prevalence.
[b]Uses a heritability of liability of 0.80 which is consistent with disease prevalence 0.85% and sibling recurrence risk of 8.6[10]
[c]Uses a heritability of liability of 0.64 that was estimated from this data.[12]
[d]No estimate provided in these references, but frequency <33% suggested in review.[13]
[e]Counting all identified first-, second-, and third-degree relatives.
[f]Assuming nuclear family size of with Poisson mean 2.2 children and complete knowledge of disease status of all first-, second- and third-degree relatives and assuming no assortative mating and no differences in fertility based on disease status.[25]

0.3 = 30 risk alleles. The S-shaped curve shows the probability of disease on the risk scale; for x-axis a) individuals carrying less than 38 risk loci have probability of disease close to 0, and for those with more than 43 risk loci, disease is guaranteed. The straight line shows the increase in risk or probability of disease on the log scale, illustrating that on this scale the risks of alleles combine additively. Many combinations of number of risk alleles, risk allele frequency, and effect size are consistent with the observed recurrence risks to relatives, illustrated by the alternative x-axes; all generate a steep rise in probability of disease with increasing number of risk alleles, implying that the genetic architecture must include epistasis on the risk scale.[4]

Exchangeable Models

Other genetic models can generate the steep rise in probability of disease for a small proportion of the population. For example, the liability threshold model[6,26–28] assumes that individuals in a population possess an unseen liability to disease, and only those whose liability exceeds a given threshold are affected. The liability has genetic and environmental component on this notional liability scale; risk alleles and environmental risks combine additively. But on the risk scale, risk alleles (and environmental factors) combine with interaction. Figure 2 shows the equivalent relationships to those in figure 1; the bell-shaped curve is the frequency distribution of the genetic liability; the straight line shows (of course) that genetic liability is additive on the liability scale, and the S-shaped curve shows the probability of disease on the risk scale (the Probit transformation of the genetic liabilities). Heritability defines the steepness of the rise in probability of disease, and the disease prevalence determines the position of rise along the x-axis relative to the population distribution of liability. Because the

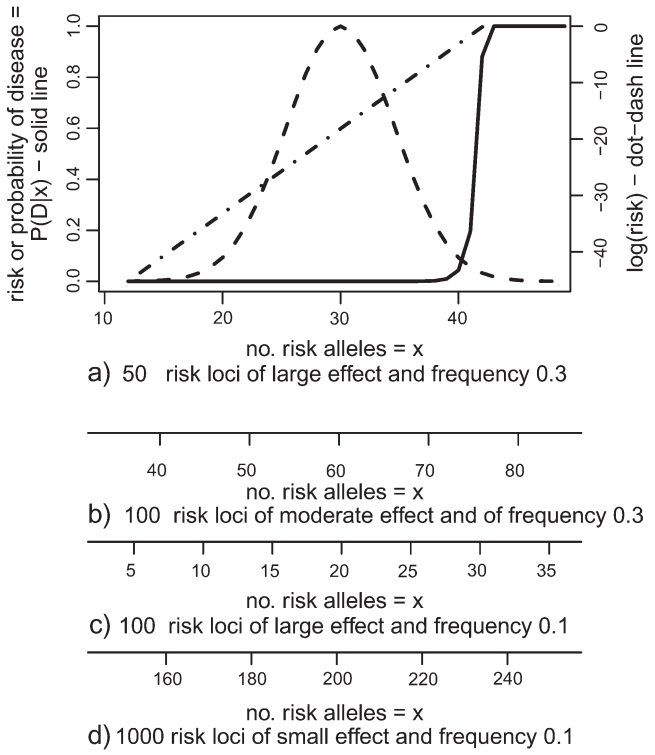

Fig. 1. Visualizing a Genetic Architecture Where Risk Alleles Act Multiplicatively. All examples represent a disease with frequency 0.72% and heritability of ~0.7. Under a simple multiplicative model of n risk loci contributing to disease each with relative risk R, the probability of disease in an individual carrying x risk loci out of the possible $2n$ is $P(D|x) = BR^x$,, assuming multiplicativity of risk alleles both within and between loci. B is the probability of disease in individuals carrying no risk loci, ie, $P(D|x = 0) = BR^0 = B$, with B defined so that $\Sigma P(D|x)P(x)$ = disease prevalence. Because B is very close to 0, x must be high before R^x is big enough to raise BR^x from being close to 0. $P(D|x)$ is constrained to have a maximum of 1. Risch[4] did not recognize the need to impose this constraint that impacts on his predicted results (discussed elsewhere[5]). The dashed bell-shaped line represents the frequency distribution of risk alleles $P(x)$, the straight dot-dashed line represents the additive genetic action on the log(risk) scale, $\log(P(D|x)) = \log(BR^x) = x\log(BR)$, and the solid line represents the multiplicative action of risk alleles on the risk scale, $P(D|x)$. The same shapes of distributions are seen for different genetic architectures as shown by alternative x-axes a)-d).

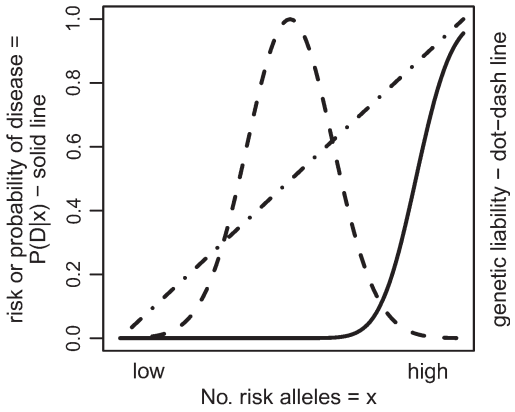

Fig. 2. Visualizing the Genetic Architecture of Complex Genetic Disease Under a Liability Threshold Model for a Disease With Frequency 0.72% and Heritability of 0.7. The model is expressed in terms of the genetic variance and so can represent an infinite combination of number of loci, risk allele frequencies, and effect sizes. The black dashed bell-shaped line represents the frequency distribution of liabilities. The straight dot-dashed line represents the additive genetic action on the liability scale. The solid line shows that on the risk scale the risk alleles combine nonadditively.

threshold model produces the same steep rise (implying epistasis[27]) in risk as the multiplicative model, and because the steep rise is a necessary requirement for a model to generate the observed pattern of recurrence risks, it is a useful model with which to develop theory. It requires no explicit assumptions about 3 key features of genetic architecture, number of loci, risk allele frequencies, and risk allele effect sizes and is simply parameterized in terms of the total variance they explain; there is no requirement that risk variants are common. The mathematical development assumes a normal distribution of liability that is approximately achieved even with a small number of risk loci, and the model is quite robust to deviations from normality as long as the distribution is unimodal.[6] The mathematical tractability of the threshold model makes it a useful way to summarize genetic architecture that applies equally to a range of "exchangeable"[5] genetic models that may appear to be described differently but that cannot be differentiated in practice.

Visualizing the Liability Threshold Model

In order to better visualize the liability threshold model, consider an analogy with height. We are all so familiar with the variation in human height that we are intuitively comfortable in recognizing that about 80% of the variation in height that we observe is of genetic origin (ie, the heritability is 80%). Adult children of short parents tend to be shorter than those of tall parents, yet there is variation between the children within families. Indeed, half of the genetic variation in populations occurs between families (the variance of the family means), and half the variation occurs within families,[6] resulting from the unique set of genetic effects received by each child from each parent in the meiotic sampling process. Imagine the "disease" of "loftiness" that affects the top 0.72% of the population. If we lined up the population in height order, then they would be ranked on their phenotypic liability to loftiness. If the population were ordered on their genetic liability, then the order would change but not too much because the heritability is high. Intuitively, we recognize that relatives of those with loftiness would also have an increased risk of loftiness. However, we can also visualize that 0.72% is a small proportion of the population and even families we consider tall might not have many individuals who pass the threshold into loftiness.

Our analogy continues by recognizing that liability to schizophrenia also has a high heritability and that the prevalence of schizophrenia in the population approximately 0.72%, but of course we cannot see liability to schizophrenia, and all we can observe are those individuals who cross the threshold of disease. The epidemiological parameters that we can measure (the combination of disease prevalence and recurrence risks to relatives) fit with what we would expect under the liability threshold model (table 1), further suggesting that it remains a useful model. Under this simple model, apparently sporadic disease is the norm (table 1)[25]; our predicted proportion of affected relatives is higher than that observed in the Swedish national study, but we assume full knowledge of true disease status of all relatives regardless of age. The frequency of sporadic cases was given at least two-thirds in the recent review of Tandon et al,[13] not inconsistent with the predictions. The simple modeling ignores observations of the epidemiology of schizophrenia that could impact on genetic variance over generations: Decreased fertility[29] of diseased individuals would decrease genetic variance while assortative mating,[30,31] and mutations accumulating over generations (including de novo mutations associated with paternal age[32]) would serve to increase genetic variance.

Can GWAS Help Our Understanding of the Genetic Architecture of Schizophrenia?

Technological advances allow us to measure genetic polymorphisms at several 100 000 locations across the genome in large cohorts of individuals. GWAS are designed to identify SNPs or CNVs associated with case-control status.[33] This has the potential to describe the genetic etiology of complex disease in quite different terms to the recurrence risks to relatives. In order to understand how these studies can contribute to an understanding of genetic architecture, we must recognize what the studies are designed to detect. The earlier generation of candidate gene association studies had led us to ensure that sample sizes of GWAS were large, allowing detection of risk alleles of small effect (relative risk > approximately 1.3) despite the unprecedented level of multiple testing. Nonetheless, there was a hope that our poor selection of genes in the era of candidate studies underlay the small number of associated variants that had been detected and that common variants of moderate effect size did exist. As an example, we use the International Schizophrenia Consortium (ISC) study[16] that was one of the larger first-generation GWAS, with 3322 case subjects and 3587 control subjects. The detailed power calculations provided in that study show that it had 100% power to detect a risk allele with frequency 0.2 and relative risk 1.5 at the stringent genome-wide level of significance of 5×10^{-8}. The ISC samples were genotyped on either the Affymetrix 5.0 or 6.0 chips, so that at least 300 000 SNPs survived quality control checks on all samples, but imputation using HapMap samples and the knowledge of correlations between known SNPs allowed association analysis of >1.6 million SNPs. Deep sequencing studies have estimated that the Affymetrix chips tag approximately 70%–80% of the total genomic variance of SNPs[34] and the majority of CNVs.[35,36]

What GWAS Have not Found

Other reviews[37,38] have focused on the handful of interesting rare and common variants that have been identified through GWAS[15–20,39–41] including de novo mutations.[42] In particular, there is mounting evidence that rare CNVs of moderate effect size play an important role in schizophrenia (reviewed in O'Donovan et al[43]). However, an equally important result is perhaps not what GWAS have found but what they have not found. In the ISC study, not one SNP and only one imputed SNP reached genome-wide significance. Because there was 100% power to detect common variants with relative risk of 1.5 and because a high proportion of the genomic variance was tagged by the genotyped SNPs, we can immediately narrow down our expectation of the genetic architecture of schizophrenia by excluding a model based on a relatively small number of common variants of moderate effect. However, from this first observation, we cannot exclude common variants of small effect size or rare variants of small or moderate effect size (important contributions to genetic variance by rare variants of large effect size were already excluded—see above). Either way, we must conclude (from equation 1) that a large number of variants must underlie the genetic etiology of schizophrenia. Can GWAS help us go further in narrowing the genetic architecture of schizophrenia?

Standard association analyses are geared to identify associated loci that are unlikely to be false positives. This is important if we are going to follow up identified loci in time-consuming and costly functional studies. In the supplementary information of the ISC study, power was explored from a different angle. The example of an associated variant with relative risk 1.05 and frequency 0.2 was considered. The power was 0 to detect association of this variant at the genome-wide significance type I error rate because it would be expected to have allele frequency of 0.2079 in case subjects and 0.1999 in control subjects. Even with 10 000 case subjects and 10 000 control subjects, power at the relatively low threshold of 1×10^{-6} was only 0.2%. Yet, these power calculations can be turned on their head to recognize that 46% of the time we would expect to see variants of this size with P values less than 0.2, and 72% of the time they will feature in the top half of the list. So if our genotyped SNPs were associated with low effect size, we might expect to see an enrichment of small P values. If many weakly associated variants were detected in an association study, then we would expect the quantile-quantile (Q-Q) plot of

genome-wide associations to rise above the line of expectation. The Q-Q plot represents the relationship between the ranked *P* values obtained from the GWAS and the ranked *P* values expected under the null hypothesis of no association. The Q-Q plot from the ISC study (figure S2 of Purcell et al[16]) shows more small p-values than expected by chance. This is typical of Q-Q plots from GWAS of complex traits (eg, Easton et al[44]), but the excess of lowly associated variants may reflect unknown biases such as population stratification or technical artifacts.

Is the Excess of Lowly Associated Variants in GWAS of Complex Diseases an Artifact or Real?

In the ISC study, the analysis team set out to investigate if the excess of lowly associated variants reflected true positives, arguing that although we could not distinguish individual true from false positives, sets of lowly associated SNPs would be (mildly) predictive of disease status in other datasets if those sets did indeed contain an excess of true positives. The basic idea was that by combining the estimated effect sizes of many SNPs simultaneously we could detect a genome-wide signature of association. We used sets of SNPs including all those with *P* values less than thresholds of 0.1, 0.2, 0.3, 0.4, and 0.5 identified from the ISC study, recognizing that by chance alone 0.1, 0.2, etc, of the SNPs would fall into these categories. For clarity, we used sets of SNPs in linkage equilibrium, although this restriction had little impact on our results. For each SNP in a SNP set, we recorded the associated allele and its odds ratio. For each individual from other ("target") GWAS, we generated a score regardless of their case-control status. The score was a weighted sum of the log(odds) for each associated allele harbored by an individual. Logistic regression of case-control status on profile score in the target studies provided evidence that the SNP sets were indeed predictive of case-control status. The scores only explained 3% of the variance in schizophrenia case-control status, but the large sample sizes ensured that this was highly significant (up to $P = 2 \times 10^{-28}$). Does this result unequivocally represent the detection of true common causal genetic variants? In the ISC study, the analysis team considered carefully whether systematic differences between cases and controls across study samples could explain the results. Population stratification seems an unlikely confounder because the same population strata would be needed in cases and controls between geographically diverse samples. However, unknown technical artifacts such as consistent differences across studies in storage conditions of clinical cases vs population sampled controls may still prevail. Stronger confirmation that the result could not be explained away came from using the Wellcome Trust Case Control Consortium studies of 7 complex genetic diseases as target study samples; critically, the same control sample was compared with each disease sample. For the target samples of coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type I diabetes, and type II diabetes, the ISC-identified SNP sets were not predictive of case-control status ($P > .05$), but for bipolar disorder, the scores explained approximately 2% of the variance in case-control status at $P = 1 \times 10^{-12}$, adding to the growing literature that supports a shared genetic etiology of schizophrenia and bipolar disorder.[12,38,40] The unequivocal conclusion is that current generation genome-wide SNP chips do tag some of the causal genetic variation for schizophrenia. Could these results provide any further insight into the genetic architecture of schizophrenia?

Using GWAS to Generate Hypotheses About the Genetic Architecture of Schizophrenia

As part of the ISC study, the ISC analysis team used simulation to explore what genetic architectures could explain the pattern of results we saw. The simulations used the same sample size and SNPs as in the real data. The parameters varied were (i) the proportion of genotyped SNPs that tagged associated causal variants, (ii) the mean effect size of the causal variants, (iii) the distribution of the effect size of the causal variants, and (iv) the linkage disequilibrium between the genotyped and causal variants. The simulated data were analyzed in the same way as the real data and combinations of parameters that generated the same pattern of variance explained using different thresholds for the SNP sets used in the target samples were identified. Many combinations of the parameters could be rejected, but equally many combinations were consistent with the observed results. As found in early studies relating genetic models to recurrence risks, and as expected from equation 1, the driving force was the total variance explained by the associated loci; many combinations of number of loci, frequency of associated loci, and effect size of associated loci generate the same variance explained. But the simulations did allow some models to be excluded and to generate hypotheses that can be tested as sample sizes, and genotyping density increase.

The simulations allowed the exclusion of models where the number of associated variants is less than approximately 100. In these situations, the effect sizes of the associated variants needed to be large in order to generate a profile that explained 3% of the variance in case-control status. But if the effect sizes were large, then they were easily detected with low association *P* values. This meant that a SNP set defined by a stringent *P* value threshold explained a high proportion of the variance in case-control status, and adding additional SNPs at less stringent thresholds simply added noise; this did not fit with observations.

The simulations also allowed the exclusion of models of only rare variants. Acknowledging that the current generation of genome-wide chips overrepresent common SNPs, a genetic architecture of ungenotyped rare causal variants whose effects could only be detected through

linkage disequilibrium with genotyped SNPs was investigated. Models with only rare variants of very large effect could not generate the pattern of variance explained in case-control status observed with decreasingly stringent SNP sets. A model with only rare variants of moderate effect could generate this pattern, but in this case, the contribution that alleles of different frequencies made to the variance explained did not match the observed results. Simulation of rare variants only showed that SNPs with low allele frequency would be expected to contribute more to the variance explained in case controls status than was observed. This is because the linkage disequilibrium (r^2) will be higher for SNPs whose minor alleles are coupled to the rare variant,[45] and so they are more likely to generate smaller *P* values in an association test. A model of only rare variants, with multiple rare variants present on common haplotype backgrounds, also could not explain the results.

Consistent with equation 1, the ISC simulations showed that we could not distinguish between a small proportion of the genotyped variants having small effect size and all genotyped variants having a very small effect size and therefore the multitude of combinations in between. However, the consistent combinations all pointed to a genetic variance of liability of 32%–36% tagged by the genotyped SNPs. The simulations have generated hypotheses that will become testable in a relatively short time frame.

1. As sample size increases, using the same genotyped SNPs, the proportion of variance in case-control status explained will increase but will still reflect the same 32%–36% of variance in liability.
2. As sample size increases, the pattern of variance explained in case-control status by SNP sets defined by the stringency of *P* values will change (because there is more power to detect variants of small effect and they will filter toward the top of the *P* value list). The change in pattern may allow us to exclude additional genetic architectures (see figure S8 in Purcell et al[16]) and may shed more light on the relative contributions of rare and common variants.
3. The next generation of SNP chips represents more of the common genomic variance and would be expected to explain a higher proportion of the liability variance. But perhaps the proportion of variance explained will not be much higher because it is likely that a proportion of the liability variance detectable through recurrence risk may never be detectable through association or sequencing if there are many rare causal variants of very small effect.

Visualizing a Polygenic Model

The results from GWAS point strongly to a genetic architecture of many (poly) genetic variants. They also imply that both common[15,16,18] and rare variants[17,41] of small effect contribute to the genetic architecture of schizophrenia. Only time will tell the genetic architecture of the unaccounted variance. Evolutionary genetics leads us to expect an L-shaped[46](or U shaped[47] with pleiotropy) distribution of risk allele frequencies and an inverse correlation between risk allele frequency and effect size.[46] From equation 1, we expect that, individually, rare variants will contribute only a very small part of the overall genetic variance. Their overall contribution to the variance depends on how many there are.

A polygenic model is not inconsistent with the phenotypic heterogeneity that characterizes schizophrenia because each individual will carry their own unique portfolio of risk alleles that may generate a spectrum of phenotypes and phenotypic heterogeneity. For example, for simplicity, let us assume that there are 1000 genetic variants contributing to risk of disease each with frequency 0.1 (figure 1, x-axis d). From binomial theory, all individuals in the population carry at least 150 risk alleles, an average individual carries 200 risk alleles, and when disease prevalence is approximately 0.72% and heritability approximately 0.7, most of those with disease carry 230–250 risk alleles. Each will carry a different set of risk alleles out of the maximum of 2000 (2 per locus).

One consequence of there being many risk variants is that the impact of a given risk variant depends on the genetic (and environmental) background of an individual. In our simplified example, there is no noticeable difference in probability of disease for individuals with 200 or 201 risk loci, yet there is a difference between individuals with 240 or 241 risk loci (figure 1 x-axis d). This means that risk variants detected with a small genotype relative risk can still be biologically very important. Therefore, identified truly associated common variants are worthy of functional investigation even if their estimated effect sizes are very small.

We have shown that the constraints of low disease prevalence and high heritability imply a steep increase in probability with disease for individuals with a high burden of risk alleles. Therefore, a normal phenotype is maintained when individuals harbor a manageable number of risk variants. This implies that biological systems can compensate for minor deviations from the normal equilibrium, eg, through alternate pathways, so that the disease phenotype is only revealed when the system cannot compensate for a large number of perturbations. This further implies a considerable degree of redundancy of genetic material that is completely consistent with a range of studies, eg, knocking out an entire gene of known function often has no or unexpected impact on the phenotype,[48–51] and protein interaction studies have shown that the vast majority of known disease genes are individually nonessential.[52] This robustness allows accumulation of mutations.[53] Moreover, while theories of balancing selection argue against common variants

of large effect but not against many common variants of small effect.[54] Indeed, if each complex genetic disease and trait is underpinned by thousands of genetic variants, models of multiple pleiotropy of each risk variant[55] are necessarily inferred (ie, a variant that has a negative effect on some characteristics may have a positive effective on others). Multiple pleiotropy is a mechanism to maintain genetic variance[47,56] because the contribution to fitness of a risk allele reflects its average contribution across its pleiotropic functions.

Prediction of Genetic Risk

Genetic variants that confer only a small increase risk to disease are individually not useful in predicting a person's genetic risk to disease. However, a risk equation combining presence/absence of each risk variant and its effect size can generate a personalized prediction of genetic risk. We investigated this problem using simulation of GWAS.[57] Our simulations showed that only when GWAS comprise about 10 000 cases and controls, would it be possible for a useful proportion of variance in disease status to be explained, even though the risks conveyed by individual variants are small. If associated variants explain half of the known genetic variance in liability of schizophrenia, then a multilocus genetic risk profile would generate an area under the receiver operator characteristic curve (a well-established tool for determining the efficacy of clinical diagnostic and prognostic tests in correctly classifying diseased and nondiseased individuals) of approximately 0.9 (N.R.W., J. Yang, PhD, M.E. Goddard, PhD, P.M.V., 2009 unpublished data,).

Prediction of genetic risk is perhaps likely to generate the most immediate impact of the results of GWAS in the clinical setting. This is because in prediction of genetic risk, the associated SNPs (or other markers) do not have to be the causative mutations: They just need to be correlated with the causative mutations ensuring that there is a consistent association between the variants used in prediction and disease risk. Ethical considerations[58] govern the use of genetic risk prediction, but to some extent these issues have been bypassed through the availability of direct to consumer testing (necessarily with very limited efficacy at this point). Despite ethical concerns, prediction of genetic risk may be an important tool for identifying schizophrenia in its prodromal phase that is the key to early intervention.[59] Around the world, protocols have been developed for identification of patients at ultrahigh risk of developing psychosis[60] that includes genetic risk through family history; but, as we have shown, a very high proportion of schizophrenia case subjects will have no close relatives with the disorder. Moreover, as half the genetic variance occurs within families,[6] each child of a parent with schizophrenia will have a different genetic risk for disease even though their risk based on family history is the same. Individuals whose genetic risk coincides with the steep rise in probability of disease (figures 1 and 2) are those most vulnerable to environmental risk factors such as recreational drug use.

Conclusion

Early studies showed that many different genetic architectures could explain the observed recurrence risks in relatives of schizophrenics. However, those studies were able to exclude the most simple genetic models and concluded that each individual with schizophrenia harbored at least "a few" genetic risk variants that act multiplicatively.[4] Recent GWAS have allowed us to exclude additional genetic architectures, showing that a genetic architecture of less than 100 risk variants and a genetic architecture of only rare variants are both not consistent with observed data. The GWAS provide evidence that perhaps half of the known genetic variance is tagged by common variants some of which is directly attributable to common causal variants of small effect. As GWAS sample sizes increase and as genotyping chips account for more of the genomic variance, the genetic architecture will become clearer. Genetic theory expects us to anticipate that some of the genetic variance so far not accounted for will be explained by rare variants, even though individually the contribution of each variant will be very small. We would expect the spectrum to include rare variants of small effect, which may never be identified. Several genetic models that represent the way in which risk variants combine to contribute to risk of disease can all explain observed results, and it is unlikely that we will be able to differentiate between them. All models require epistasis on the risk scale, which is essential to produce the steep rise in probability of disease necessary to generate the pattern of observed recurrence risks to relatives. The most tangible and immediate outcome of the GWAS may be prediction of genetic risk to disease, which may be an important tool in ensuring early intervention treatments.

Funding

Australian National Health and Medical Research Council (389892, 442915, 339450, 443011, and 496688); Australian Research Council (DP0770096).

Acknowledgments

We thank the ISC polygene analysis subteam (Shaun Purcell, Stuart Macgregor, Pamela Sklar, Patrick Sullivan, and Michael O'Donovan) and Mike Goddard for many discussions on this subject. We thank Michael O'Donovan, Michael Owen, John McGrath, Michael Berk, and Danielle Posthuma for commenting on the manuscript.

References

1. Sullivan PF. The genetics of schizophrenia. *Plos Med*. 2005;2:614–618.
2. McGue M, Gottesman II, Rao DC. The transmission of schizophrenia under a multifactorial threshold model. *Am J Hum Genet*. 1983;35:1161–1178.
3. McGue M, Gottesman II, Rao DC. Resolving genetic models for the transmission of schizophrenia. *Genet Epidemiol*. 1985;2:99–110.
4. Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet*. 1990;46:222–228.
5. Slatkin M. Exchangeable models of complex inherited diseases. *Genetics*. 2008;179:2253–2261.
6. Falconer D, Mackay T. Introduction to Quantitative Genetics. 4th ed. Harlow, UK: Pearson Education Ltd; 1996.
7. Heston LL. Psychiatric disorders in foster home reared children of schizophrenic mothers. *Br J Psychiatry*. 1966;112: 819–825.
8. Kety SS, Rosentha D, Wender PH, Schulsin F. Mental illness in biological and adoptive families of adopted schizophrenics. *Am J Psychiatry*. 1971;128:302–306.
9. Saha S, Chant D, Welham J, McGrath J. A systematic review of the prevalence of schizophrenia. *PLoS Med*. 2005;2:413–433.
10. Reich T, James JW, Morris CA. The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. *Ann Hum Genet*. 1972;36:163–184.
11. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait—evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry*. 2003;60:1187–1192.
12. Lichtenstein P, Yip BH, Bjork C, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*. 2009;373:234–239.
13. Tandon R, Keshavan MS, Nasrallah HA. Schizophrenia, "Just the Facts" what we know in 2008. 2. Epidemiology and etiology. *Schizophr Res*. 2008;102:1–18.
14. Ng MYM, Levinson DF, Faraone SV, et al. Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry*. 2009;14:774–785.
15. Stefansson H, Ophoff RA, Steinberg S, et al. Common variants conferring risk of schizophrenia. *Nature*. 2009;460:U744–U799.
16. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460:748–752.
17. Stone JL, O'Donovan MC, Gurling H, et al. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*. 2008;455:237–241.
18. Shi JX, Levinson DF, Duan J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*. 2009;460:753–757.
19. Shifman S, Johannesson M, Bronstein M, et al. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. *PLoS Genet*. 2008;4(2):e28.
20. Sullivan PF, Lin D, Tzeng JY, et al. Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Mol Psychiatry*. 2008;13:570–584.
21. Blackwood DHR, Fordyce A, Walker MT, St Clair DM, Porteous DJ, Muir WJ. Schizophrenia and affective disorders—cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *Am J Hum Genet*. 2001;69:428–433.
22. O'Rourke DH, Gottesman II, Suarez BK, Rice J, Reich T. Refutation of the general single-locus model for the etiology of schizophrenia. *Am J Hum Genet*. 1982;34:630–649.
23. Eaves LJ, Kendler KS, Schulz SC. The familial sporadic classification—its power for the resolution of genetic and environmental etiologic factors. *J Psychiatr Res*. 1986;20:115–130.
24. Kendler KS. Sporadic vs familial classification given etiologic heterogeneity. 1. Sensitivity, specificity and positive and negative predictive value. *Genet Epidemiol*. 1987;4:313–330.
25. Yang J, Visscher PM, Wray NR. Sporadic cases are the norm for complex disease [published online ahead of print Oct 14, 2009]. *Eur J Hum Genet*. doi:10.1038/ejhg.2009.177.
26. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet*. 1965;29:51–71.
27. Dempster ER, Lerner IM. Heritability of threshold characters. *Genetics*. 1950;35:212–236.
28. Gottesman II, Shields J. A polygenic theory of schizophrenia. *Proc Natl Acad Sci U S A*. 1967;58:199–205.
29. Risch N. Estimating morbidity risks in relatives—the effect of reduced fertility. *Behav Genet*. 1983;13:441–451.
30. Lichtenstein P, Bjork C, Hultman CM, Scolnick E, Sklar P, Sullivan PF. Recurrence risks for schizophrenia in a Swedish National Cohort. *Psychol Med*. 2006;36:1417–1425.
31. Maes HHM, Neale MC, Kendler KS, et al. Assortative mating for major psychiatric diagnoses in two population-based samples. *Psychol Med*. 1998;28:1389–1401.
32. Malaspina D, Harlap S, Fennig S, et al. Advancing paternal age and the risk of schizophrenia. *Arch Gen Psychiatry*. 2001;58:361–367.
33. Cichon S, Craddock N, Daly M, et al. Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *Am J Psychiatry*. 2009;166:540–556.
34. Bhangale TR, Rieder MJ, Nickerson DA. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet*. 2008;40:841–843.
35. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome [published online ahead of print Oct 07, 2009]. *Nature*. doi:10.1038/nature08516.
36. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*. 2008;40:1199–1203.
37. O'Donovan MC, Craddock N, Norton N, et al. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet*. 2008;40:1053–1055.
38. Craddock N, O'Donovan MC, Owen MJ. Psychosis genetics: modeling the relationship between schizophrenia, bipolar disorder, and mixed (or "schizoaffective") psychoses. *Schizophr Bull*. 2009;35:482–490.
39. Kirov G, Zaharieva I, Georgieva L, et al. A genome-wide association study in 574 schizophrenia trios using DNA pooling. *Mol Psychiatry*. 2009;14:796–803.
40. Moskvina V, Craddock N, Holmans P, et al. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol Psychiatry*. 2009;14:252–260.
41. Need AC, Ge DL, Weale ME, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet*. 2009;5(2):e1000373.

42. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet*. 2008;40: 880–885.
43. O'Donovan MC, Craddock NJ, Owen MJ. Genetics of psychosis; insights from views across the genome. *Hum Genet*. 2009;126:3–12.
44. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007;447:1087–1093.
45. Wray NR. Allele frequencies and the r2 measure of linkage disequilibrium: Impact on design and interpretation of association studies. *Twin Res Hum Genet*. 2005;8:87–94.
46. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001;69:124–137.
47. Zhang XS, Hill WG. Genetic variability under mutation selection balance. *Trends Ecol Evol*. 2005;20:468–470.
48. Carlisle HJ, Fink AE, Grant SGN, O'Dell TJ. Opposing effects of PSD-93 and PSD-95 on long-term potentiation and spike timing-dependent plasticity. *J Physiol*. 2008;586:5885–5900.
49. Grant SGN, Odell TJ, Karl KA, Stein PL, Soriano P, Kandel ER. Impaired long-term potentiation, spatial-learning, and hippocampal development in fyn mutant mice. *Science*. 1992;258: 1903–1910.
50. Migaud M, Charlesworth P, Dempster M, et al. Enhanced long-term potentiation and impaired learning in mice with mutant postsynaptic density-95 protein. *Nature*. 1998;396: 433–439.
51. Ahituv N, Zhu Y, Visel A, et al. Deletion of ultraconserved elements yields viable mice. *PLoS Biol*. 2007;5:1906–1911.
52. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007;104:8685–8690.
53. Masel J, Siegal ML. Robustness: mechanisms and consequences. *Trends Genet*. 2009;25:395–403.
54. Keller MC, Miller G. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behavioral and Brain Sciences*. 2006;29:385–404.
55. Taylor CF, Higgs PG. A population genetics model for multiple quantitative traits exhibiting pleiotropy and epistasis. *J Theor Biol*. 2000;203:419–437.
56. Zhang XS, Hill WG. Multivariate stabilizing selection and pleiotropy in the maintenance of quantitative genetic variation. *Evolution*. 2003;57:1761–1775.
57. Wray NR, Goddard ME, Visscher PM. Prediction of individual risk to disease from genome-wide association studies. *Genome Res*. 2007;17:1520–1528.
58. Laegsgaard MM, Mors O. Psychiatric genetic testing: attitudes and intentions among future users and providers. *Am J Med Genet B*. 2008;147B:375–384.
59. McGorry PD, Yung AR, Bechdolf A, Amminger P. Back to the future—predicting and reshaping the course of psychotic disorder. *Arch Gen Psychiatry*. 2008;65:25–27.
60. Yung AR, Killaciey E, Hetrick SE, et al. The prevention of schizophrenia. *Int Rev Psychiatry*. 2007;19:633–646.