

EXPERT REVIEW

Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses

PM Visscher¹, ME Goddard^{2,3}, EM Derks⁴ and NR Wray¹

¹Queensland Institute of Medical Research, Herston, Queensland, Australia; ²Department of Primary Industries, Biosciences Research Division, Melbourne, Victoria, Australia; ³Department of Agriculture and Food Systems, University of Melbourne, Melbourne, Australia and ⁴Neuroimaging Research Group, University Medical Center, Utrecht, The Netherlands

In this article, we review some of the data that contribute to our understanding of the genetic architecture of psychiatric disorders. These include results from evolutionary modelling (hence no data), the observed recurrence risk to relatives and data from molecular markers. We briefly discuss the common-disease common-variant hypothesis, the success (or otherwise) of genome-wide association studies, the evidence for polygenic variance and the likely success of exome and whole-genome sequencing studies. We conclude that the perceived dichotomy between 'common' and 'rare' variants is not only false, but unhelpful in making progress towards increasing our understanding of the genetic basis of psychiatric disorders. Strong evidence has been accumulated that is consistent with the contribution of many genes to risk of disease, across a wide range of allele frequencies and with a substantial proportion of genetic variation in the population in linkage disequilibrium with single-nucleotide polymorphisms (SNPs) on commercial genotyping arrays. At the same time, most causal variants that segregate in the population are likely to be rare and in total these variants also explain a significant proportion of genetic variation. It is the combination of allele frequency, effect size and functional characteristics that will determine the success of new experimental paradigms such as whole exome/genome sequencing to detect such loci. Empirical results suggest that roughly half the genetic variance is tagged by SNPs on commercial genome-wide chips, but that individual causal variants have a small effect size, on average. We conclude that larger experimental sample sizes are essential to further our understanding of the biology underlying psychiatric disorders.

Molecular Psychiatry advance online publication, 14 June 2011; doi:10.1038/mp.2011.65

Keywords: schizophrenia; bipolar disorder; major depression; GWAS; genetic architecture; heritability

Introduction

A large number of review and commentary articles have been written recently about the likely (molecular) genetic basis of schizophrenia, bipolar disorder and other psychiatric disorders. In this review, we aim not to build more 'theoretical mountains out of empirical molehills',¹ but instead we focus on the empirical evidence that has been accumulated about the genetic basis of psychiatric disease and we will make inferences about what is consistent with observed data. We revisit the often misinterpreted common-disease common-variant hypothesis, the success (or otherwise) of genome-wide association studies (GWAS), the evidence for polygenic variance and the likely success of exome and whole-genome sequencing studies. We limit ourselves to a small

number of psychiatric disorders, namely schizophrenia, bipolar disorder and major depression, and do not address or discuss important issues about definitions of diagnosis and phenotypes. Our main objective is to attempt to understand and explain genetic variation in risk to psychiatric disease in the population as represented by the phenotypically and demographically heterogeneous samples of cases and controls utilised in the current era of GWAS.

Simple and complex inheritance

Some diseases and other traits are controlled by a single gene and show classical Mendelian patterns of inheritance. In some cases, the mutant allele causing the disease is dominant, but more often it is recessive. Mental retardation as a result of phenylketonuria (PKU) is an example. There is usually more than one variant that can cause the disease because there is more than one mutation that interferes with the normal function of the gene product. In some cases, there is more than one gene at which mutations can cause the same disease but, because the disease-causing

Correspondence: Professor PM Visscher, Queensland Institute of Medical Research, 300 Herston Road, Herston, Queensland 4006, Australia.

E-mail: Peter.Visscher@qimr.edu.au

Received 2 February 2011; revised 5 April 2011; accepted 20 April 2011

mutations are rare, within each family the disease is inherited in a simple Mendelian manner.

Common diseases such as psychiatric disorders, diabetes and cardiovascular disease do not segregate in a Mendelian manner within families and are influenced by both genetic and environmental factors. Often such diseases are called ‘complex diseases’. Most continuously variable traits such as height are affected by many genes and environmental factors. The inheritance of disease susceptibility can also be explained by such a model if we assume that a combination of genes and environmental effects control a person’s liability to the disease and that people whose liability is above a threshold show the disease and those below the threshold do not. Many of the features of psychiatric disorders are compatible with this model,^{2–4} whereas these observations are not consistent with Mendelian models. The liability threshold model is consistent with the apparent dichotomy between sporadic cases and cases with family history in the population. For a disease with a population prevalence of 1% and a heritability of liability of 0.6–0.8, more than 60% of probands are predicted to be sporadic (no affected first-, second- and third-degree relatives) in typical families with on average two or three children per couple.⁵ Finally, the liability threshold model predicts discordance between monozygotic (MZ) twins, again consistent with empirical data.⁶ For example, for diseases such as bipolar disorder and schizophrenia (with heritability on the scale of liability of ~0.6–0.8, and a prevalence of about 1%), the threshold model predicts an MZ discordance rate of 80–65% (Figure 1). Hence, this simple model predicts that MZ twins are likely to be discordant for disease, without invoking single causal events such as post-zygotic *de novo* mutations or

epimutations. Sometimes when the simple Mendelian model does not fit the data, the departure is explained as due to incomplete penetrance. That is, there are other genes and environmental factors that determine whether or not an individual exhibits the disease. Incomplete penetrance is consistent with the threshold model when there are many genes and environmental factors that together ‘modify’ the effect of a variant of large effect. In fact, the threshold model can reflect a multitude of genetic architectures of disease in terms of the number of genes that affect the trait, the distribution of their frequencies and effects sizes, and the interactions among alleles within a locus (dominance) and between loci (epistasis). Moreover, other genetic models that may seem quite different to the liability threshold model in fact have similar properties for the parameter combinations, which generate results consistent with observed data (relative risks to different classes of family member). Such models can be considered as ‘exchangeable’.^{7,8} Recognizing this exchangeability makes the threshold model the model of choice because of its simplicity by describing genetic variation through only two parameters, disease prevalence and heritability, consistent with many genetic architectures.

Usually only additive variance is assumed on the liability scale because no empirical data have suggested important contributions of non-additive effects on this scale, but additivity on the liability scale implies considerable epistasis on the disease scale as only those individuals harbouring multiple risk loci will be affected.⁹ We have discussed the effects due to scale in detail in an earlier review.¹⁰

Heritability: what can we learn from family/pedigree studies?

Estimates of the heritabilities of psychiatric disorders have been reviewed elsewhere, with schizophrenia and bipolar disorder having lifetime probability of disease (0.5–1%) and high heritability ~0.8.^{11,12} In contrast, major depression is much more prevalent (10–15%) and has lower heritability of approximately 0.4,¹³ although higher heritabilities are associated with early onset,¹⁴ recurrence¹⁵ and consideration of severity and reliability (including the use of multiple interviews to account for measurement error) in population-based samples.^{16,17} These heritabilities are expressed on the liability scale and are derived from estimates of risk of disease in relatives compared with risk of disease in the general population. Risk of disease in different types of relatives helps to disentangle genetic from common environmental risk factors and typically estimates of common environmental factors are small.^{12,13,18} The relative consistency of estimates from twin studies for schizophrenia is notable; however, for each study the recruitment of family members is from within a restricted environmental setting, usually a single hospital, and this may lead to biased estimates due to confounding factors.^{12,19} In contrast, heritability

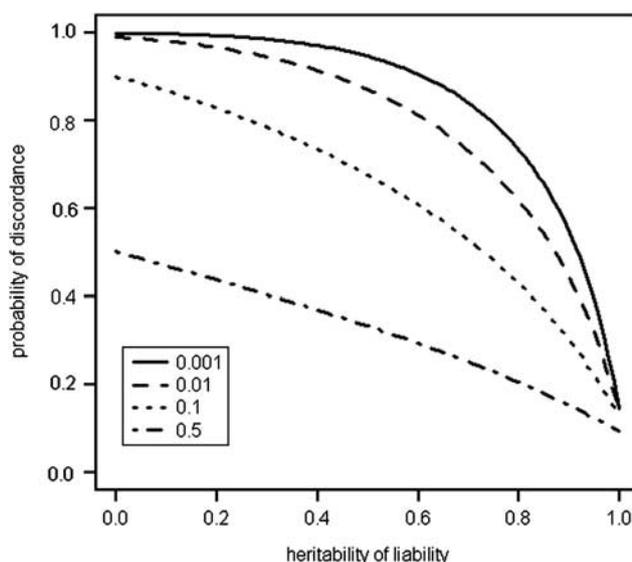


Figure 1 Expected monozygotic twin (MZ) discordance for disease under a homogeneous polygenic (liability threshold) model for difference disease prevalences (0.001, 0.01, 0.1 and 0.5), following refs.^{6,7,3,74}

estimates using national records with family relationships linked to hospital records are lower ~ 0.65 for schizophrenia (Wray and Gottesman, unpublished data),¹⁸ perhaps reflecting environmental factors (such as urban vs rural effects, or clinical practice), which do not contribute to single hospital studies. The estimates of heritability from national records may be more representative of the samples contributing to the large international consortia in the current era of GWAS. Either way, the estimates of heritability are large, reflecting an important contribution of inherited genetic variants to the aetiology of these disorders. Importantly, *de novo* mutations are not shared by relatives and so cannot contribute to the estimates of heritability, except in the case of MZ twins. However, *de novo* mutations in one generation become rare variants if they are passed to the next generation, and indeed heritability estimates reflect sharing of all variants between relatives regardless of their frequencies. Mutation rates affecting complex traits tend to be low, and add about 0.1–1% of heritability per generation when estimated in experimental organisms,²⁰ and heritability estimates of psychiatric disorders from twins are similar to those estimated from other relatives (Wray and Gottesman, unpublished data).^{3,10} Therefore, *de novo* mutations are unlikely to give biased estimates of narrow sense heritability when using MZ twins.

Evolution at genes affecting disease susceptibility

The frequencies and effect sizes of risk variants contributing to psychiatric diseases that are segregating in the population are a consequence of our evolutionary past. The evolutionary forces that are of particular relevance are natural selection, mutation and genetic drift. Natural selection acts upon ‘fitness’, which is easier to define than to measure in practice. New mutations that have an effect on psychiatric disease can also have an effect on other traits (pleiotropy). It is the combined effects of a mutation on all traits that determine their effect on fitness,^{21,22} and it is the effect on fitness that is important when predicting their allele frequency in the population. A mutation can have effects on many phenotypes and the direction of effect is not necessarily the same for all traits. For example, a mutation can increase the risk of one disease (for example, sickle cell anaemia), but be protective for another disease (for example, malaria). Moreover, the effect of a mutation on fitness may change over time or between environments. For example, a mutation increasing the ability to store fat reserves in the body efficiently may be advantageous in times of food shortages, but disadvantageous in times of food abundance. Whereas researchers tend to focus on the effect of a mutation on one particular complex trait (their disease of interest) in the present day, it is the net effect of the mutation on fitness in our evolutionary past that is important in making predictions about its frequency in the population today.

It is well recognised that schizophrenia is associated with increased mortality from natural and unnatural causes and with reduced reproductive rates²³ (reviewed by Uher²⁴). It seems likely that mutations with a large effect on incidence of schizophrenia will have a negative effect on fitness and so selection will eliminate them or keep them to a low frequency in the population. However, it is hard to predict the fitness of mutations with a small effect on susceptibility to schizophrenia, especially considering their possible effects on other traits that may have been important in our evolutionary past. If the effect on fitness is small enough, the mutation will behave like a neutral mutation. Therefore, we will first consider the neutral model and then the effect of natural selection.

From the evolutionary neutral model²⁵ flow some interesting and testable predictions. In particular, under a constant effective population size, the distribution of the frequencies of the segregating causal variants is known.²⁶ We have plotted the cumulative frequency of segregating variants as a function of minor allele frequency (MAF) in Figure 2.

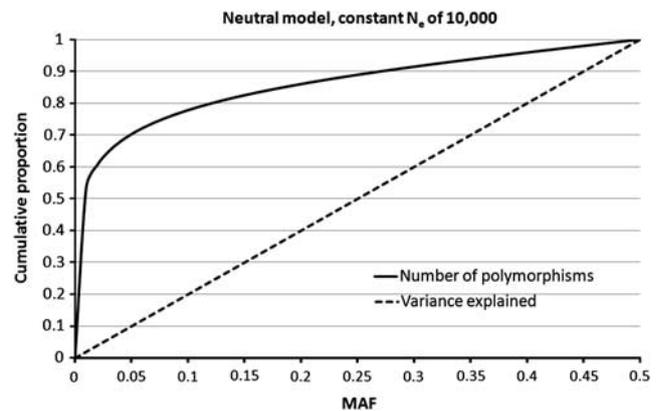


Figure 2 The neutral model: most variants are rare yet most variation is due to common variants. The variance explained by a particular variant depends on its frequency (p) and effect size (a , on some arbitrary scale). If selection against any particular variant is weak and there is random mating, then the genetic variance contributed by the variant is approximately $2p(1-p)a^2$.²¹ For rare variants (small p), the amount of genetic variation contributed by the variant is $\sim 2pa^2$, whereas for a very common variant ($p=1/2$) it is $\frac{1}{2}a^2$. As an example, if the effect sizes (a) are the same, then 250 rare variants with a frequency of 0.1% explain the same amount of genetic variation in the population as one common variant with a frequency of 50%. If p and a are negatively correlated (mutations of large effects are more selected against), which seems plausible, then the question becomes what the value of pa^2 is as a function of allele frequency. Under a neutral model and a constant effective population size, the distribution of segregating causal variants is known (for example, ref.²⁶). Here we have plotted the cumulative frequency of segregating variants as a function of minor allele frequency (MAF). Clearly most variants are rare: approximately 70% of all variants have MAF < 0.05. However, the distribution of the genetic variance explained as a function of MAF is uniform, because the distribution of allele frequency is proportional to $1/[p(1-p)]$ and the variance explained is proportional to $p(1-p)$.⁷⁵

Clearly most variants are rare: approximately 70% of all variants have $MAF < 0.05$. However, the distribution of the genetic variance explained as a function of MAF is uniform, hence the cumulative proportion of genetic variance explained is linear in MAF, so that variants with $MAF < 0.05$ together only explain 10% of genetic variation (Figure 2). Hence, under this basic and simple model, most causal variants are rare, but most variation is due to common variants.

The effect of natural selection is to shift both lines in Figure 2 to the left. We used the model from Eyre-Walker²⁷ to quantify the proportion of genetic variation as a function of MAF. The model assumes that mutations that effect the susceptibility to disease inevitably decrease fitness and it uses a parameter (τ) to describe the relationship. For instance, if $\tau = 1$, a mutation that has a twofold larger effect on fitness also has a twofold larger effect on disease.²⁷ However, if $\tau = 0.5$, then the effect on fitness goes up proportional to the square of the effect on disease incidence. In Figure 3, we give the results for three values of τ (0.25, 0.50, 1.00). The results show that under these fitness–disease correlations, a lot of genetic variation is contributed by variants with a frequency of $< 5\%$. For $\tau = 0.25, 0.50$ and 1.00 , a proportion of genetic variance of 45, 88 and 99.97%, respectively, is contributed by variants with $MAF < 0.05$. However, other models of selection would predict a higher contribution to variance from alleles with a frequency

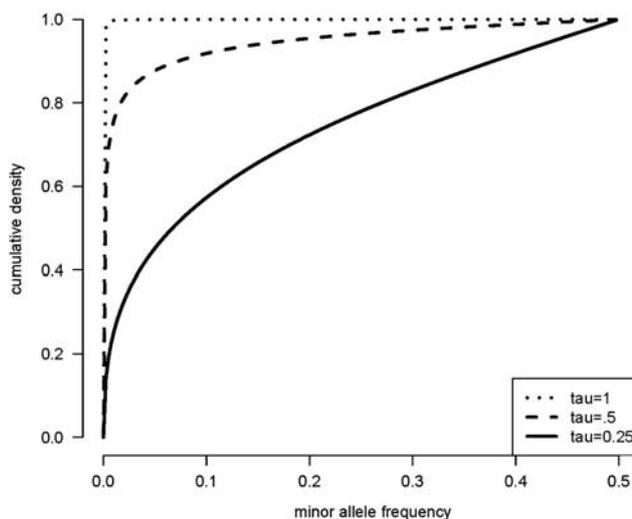


Figure 3 Strong natural selection: cumulative genetic variation explained by minor allele frequency (MAF) of causal variants. The y axis is the cumulative density of genetic variation, which sums to 1.0 for $MAF = 0.5$. We used the model from Eyre-Walker²⁷ to quantify the proportion of genetic variation as a function of MAF. The model assumes that effects of a mutation on fitness and disease are correlated (parameter τ), such that the effect on disease (a in the notation used in Figure 1) is proportional to S^{τ} , with S the strength of selection of the mutation on fitness. Other parameters in the model were chosen as in Eyre-Walker.²⁷ At an extreme, for $\tau = 1$, a mutation that has a twofold larger effect on fitness also has a twofold larger effect on disease.²⁷

of 5% or more. For instance, under the Eyre-Walker model, it is not possible for mutations to have a small negative effect on fitness due to its effect on, say, schizophrenia, being counter-balanced by a small positive effect on some other trait.

Therefore we expect, based on evolutionary theory, that mutations affecting psychiatric disease susceptibility will tend to be at low allele frequency, but we cannot predict with any confidence whether genetic variation will be caused mainly by rare alleles or, more similar to the neutral model, spread more equally across MAF classes. These two extremes can also be described as one in which allelic effect on disease and allele frequency is highly negatively correlated and one in which they are not.

Common disease common variant hypothesis

Currently, the allele frequency at genes causing psychiatric disease is a subject of some debate.^{28,29} The common disease common variant hypothesis is sometimes said to be one side of this debate, with the other side being that disease-causing alleles are typically rare. However, what is the precise ‘hypothesis’ in the common disease common variant hypothesis? Lander³⁰ noted from the then available data that there is a limited diversity in coding regions at genes, in that most variants are very rare so that the effective number of alleles is small. In addition, he provided ‘tantalising examples’ of common alleles with large effects (for example, APOE, MTHFR, ACE). Later, Reich and Lander³¹ presented a theoretical population genetics model that predicted a relatively simple spectrum of the frequency of disease risk alleles at a particular disease locus. They (re)phrased the common disease common variant hypothesis as the prediction that the expected allelic identity is high for those disease loci that are responsible for most of the population risk for disease. These papers did not appear to make any prediction about the number of disease loci and therefore about the effect size. What the authors stated was that if a disease was common, there was likely to be one disease-causing allele that was much more common than all the other disease-causing alleles.^{30,31} Although not the subject of the original common disease common variant hypothesis, the modern debate is about the entire frequency spectrum of disease-causing alleles in the genome. Phrasing the debate as an either/or question is not very helpful as examples of both common and rare alleles are already known, but there is still an open question as to whether most genetic variation in the population is caused by rare variants or common variants or more generally what is the spectrum of allele frequencies of disease-causing alleles.

GWAS results: what are the data telling us?

GWAS for schizophrenia and bipolar disorder have identified a handful of common single-nucleotide polymorphisms (SNPs) and less common copy

number variants (CNVs), which have passed stringent thresholds of significance and/or have replicated across populations. These results have been discussed in reviews elsewhere in terms of their functional relevance (for example, refs.^{32,33}). Here, as our focus is genetic architecture, we present a number

of GWAS identified associations in Tables 1 and 2 in terms of the variance explained in liability. Any calculations on the liability scale require an estimate of the prevalence of disease in the population and we have used both 0.5 and 1% to show that the range of reported prevalences has little impact on the

Table 1 Variance in liability explained by individual robustly associated SNPs and CNVs for schizophrenia

Marker SNP risk allele or CNV	Chromosomal location	Risk allele frequency	Odds ratio	Variance in liability explained ^a assumed disease prevalence		Gene ^b	References
				0.5%	1%		
rs1344706-T	2q32.1	0.59	1.09	0.04	0.05	<i>ZNF804A</i>	76
rs9272219-G	6p21.32–22.1	0.72	1.14	0.08	0.10	<i>MHC</i>	51,77,78
rs12807809-T	11q24.2	0.83	1.15	0.07	0.08	<i>NRGN</i>	78
rs9960767-C	18q21.2	0.06	1.23	0.06	0.07	<i>TCF4</i>	78
rs7341475-G	7q22.1	0.62	1.58	0.57 ^c	0.67 ^c	<i>RELN</i>	79
Total for SNPs				0.82	0.96		
CNV-deletion	1q21.1	0.000214	12.4	0.05	0.06		46,47
CNV-deletion	2p16.3	0.0004	4.8	0.03	0.03	<i>NRXN1</i>	80
CNV-deletion	15q11.2	0.0019	2.9	0.06	0.07		44,47
CNV-deletion	15q13.3	0.00019	11	0.04	0.05		44,47
CNV-duplication	16p11.2	0.00029	26	0.12	0.16		81
CNV-duplication	16p13.1	0.002	3.98	0.11	0.13		44
CNV-deletion	17p12	0.00018	7.81	0.02	0.03		44
CNV-deletion	22q11	0.00033	30	0.16	0.21		46,47
Total for CNVs				0.58	0.74		
Total for schizophrenia				1.40	1.71		

Abbreviations: CNV, copy number variant; GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism. For SNPs, we used identified unique association with $P < 10^{-6}$ as collated in the GWAS Catalog;⁸² for CNVs, we searched the primary literature and used the review.³²

^aThe variance in liability is calculated from the mean liability associated with each genotype class.

^bWe have listed the nearest gene, or the most relevant gene, but this can be debatable, please see primary references for a more complete description.

^cThe listed variance is the variance in liability divided by two as the association is in women only.

Table 2 Variance in liability explained by individual robustly associated SNPs for bipolar disorder

SNP risk allele	Chromosomal location	Risk allele frequency	Odds ratio	Variance in liability explained ^a assumed disease prevalence		Gene ^b	Reference
				0.5%	1%		
rs472913-C	1p32.1	0.50	1.18	0.16	0.19	<i>NF1A</i>	83
rs1042779-A	3p21.1	0.63	1.19	0.17	0.20	<i>ITIH1</i>	83
rs17418283-C	5q15	0.28	1.21	0.18	0.21	<i>MCTP1</i>	83
rs10994336-T	10q21.2	0.05	1.45	0.16	0.19	<i>ANK3</i>	84
rs1006737-A	12p13.33	0.32	1.18	0.14	0.17	<i>CACNA1C</i>	84
rs1012053-A	13q14.11	0.84	1.59	0.66	0.77	<i>DGKH</i>	85
rs12899449-A	15q14	0.72	1.2	0.16	0.19	<i>RASGRP1</i>	84
rs420259-A	16p12.1	0.72	2.08 ^c	0.19	0.22	<i>Multiple</i>	86
Total for bipolar disorder				1.82	2.14		

Abbreviations: CNV, copy number variant; GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism. For SNPs, we used identified unique association with $P < 10^{-6}$ as collated in the GWAS Catalog;⁸² robustly associated CNVs specific to bipolar disorder have not been reported.

^aThe variance in liability is calculated from the mean liability associated with each genotype class.

^bWe have listed the nearest gene, or the most relevant gene, but this can be debatable, please see primary references for a more complete description.

^cThe same odds ratio was assumed for heterozygotes and homozygotes reflecting the reported association results.

estimates of variance explained. We note that in many cases the risk increasing SNP allele is very common (frequency >0.3) and often is the major allele. Although the identified SNPs are not expected to be the causal variant, but only in linkage disequilibrium (LD) with them, it is likely that the causal variant is also very common.³⁴ The SNPs in total explain $<1\%$ of the variance in liability for schizophrenia and $\sim 2\%$ of the variance in liability for bipolar disorder, although we note that for both disorders a large proportion of the variance is contributed by a single SNP. At the time of writing, results of the Psychiatric GWAS Consortium for schizophrenia and bipolar disorder have not been published, but the newly identified SNPs by the Psychiatric GWAS Consortium contribute only a very small amount in terms of variance explained. GWAS have also identified both associated CNVs and an increased load of CNVs in schizophrenia cases (reviewed by O'Donovan *et al.*³² and Kirov³³). The associated CNVs are rarer, but have higher association odds ratio (OR). As discussed above, the contribution of an associated variant to the variation in the population reflects both its frequency and its effect size (OR), and so despite the high OR, these rare CNVs individually contribute $<0.2\%$ of the variance. We note that the well-known 22q11 deletion is thought to be a *de novo* deletion in about 90%³⁵ of people who carry it, its prevalence reflecting a chromosomal structure (low-copy repeat sequences) that generates an increased probability of the occurrence of a deletion compared with the rest of the genome.³⁶ The variance contributed by *de novo* mutations, although genetic in the sense that the mutation is in the genome, would be partitioned into the environmental variance in estimation of heritability because the mutation is (mostly) not shared by affected relatives. MZ twins will share *de novo* mutations, but as we argued above, these mutations are unlikely to bias the estimates of narrow sense heritability in a non-trivial way.

Despite only a handful of associated variants identified as being genome-wide significant from the first generation of GWAS, the knowledge we have gained from GWAS should not be understated. We have learnt much about the genetic architecture. A GWAS of 3000 cases and 3000 controls has 99% power to detect common SNPs, which have true association of genotype relative risk 1.4 and should be able to detect 35% of true associations of genotype relative risk 1.25 (equivalent to 0.70 and 0.25% variance in liability, assuming frequency 0.2 and population disease prevalence 0.7% and significance threshold of 5×10^{-8}).³⁷ Similar statements can be made about many other complex genetic diseases.³⁸ So we know that there is substantial genetic variation in the population, but few common SNPs have been detected with stringent genome-wide significance. From these observations, we can conclude that either causal variants that are tagged by the current generation of genotyping arrays must have effect sizes at individual SNPs so small that they do not reach

genome-wide significance in GWAS given current sample sizes, or causal variants are not in sufficient LD with SNPs on the current generation of genotyping arrays to be detected by association.^{39–41} These two explanations are not mutually exclusive, but represent contributions to the genetic architecture of small effects tagged by the genotyped markers vs causal variants that are not in LD with the genotyped markers, mostly these will be rare ($<1\%$) and uncommon ($<5\%$) variants. One way to separate these hypotheses is to estimate the variation that is explained by all common SNPs together.^{39,42} When applied to data of the International Schizophrenia Consortium (ISC), it was estimated that about one-third of variation in liability was in LD with the SNPs of the commercial genotyping array, and that variation was spread over all SNPs, across the entire range of SNP allele frequencies (unpublished results). These results were replicated in the independent Molecular Genetics of Schizophrenia study data and are consistent with many causal variants, also across a range of allele frequencies, segregating in the population.

We know from evolutionary theory (above) that most causal variants are expected to be rare, but also that common variants may contribute most of the variation even if their effect sizes are small. What is the emerging evidence for a contribution from common variants of small effect size? This is explored in the next sections.

What if genetic architecture of psychiatric disease is like that of height?

To date, the largest number of loci identified for any complex trait in human beings is for height. Using discovery and validation samples of $\sim 130\,000$ and $\sim 50\,000$ subjects, Lango Allen *et al.*⁴³ reported 180 loci. The robust associations are with common SNP variants and effect sizes were of the order of 1–4 mm or 0.01–0.06 phenotypic standard deviations (assuming a standard deviation for height of 70 mm). These reported variants each explain approximately 0.02–0.2% of phenotypic variation. Many of the variants were in or near 'meaningful' candidate genes, for example, genes for which major mutants had been reported (in human beings or other species) and genes known to be involved in growth pathways. It appears from these results that height is highly polygenic, with many variants contributing and therefore small effect sizes for the individual variants. It is highly likely that much of the remaining genetic variation for height lies in a spectrum of allele frequencies, with many more common variants with even smaller effect size and many rare variants, some of which may have larger effect size. For example, alleles with a frequency of 1% in the population and an additive effect size of 1 cm would explain 0.04% of variance each, but would not have been in sufficient LD with common SNPs to have been detected by GWAS to date. Statistical power to detect an effect associated with an SNP depends on the variance explained by

the SNP and this is proportional to the LD squared correlation (r^2) between the ungenotyped causal variant and the genotyped SNP. So a genotyped SNP that is in low LD with a causal variant (say, $r^2 = 0.1$), because, for example, the causal variant has a lower frequency than SNPs on a commercial SNP array, will explain only 10% of variation that would be accounted for if the causal variant had been genotyped and tested for association.⁴¹

Many researchers may be uncomfortable drawing an analogy between the quantitative trait of height and the serious psychiatric condition of, for example, schizophrenia, yet an underlying liability to schizophrenia can be conceptualised as an unobserved quantitative trait. Height and liability to schizophrenia have similarly high heritabilities. Even accepting this parallel many researchers would argue that schizophrenia is likely to have a very different genetic architecture from height, on the grounds of past natural selection having operated quite differently on these two complex traits. Apart from the increased prevalence of CNVs in schizophrenia patients^{44–47} compared with healthy subjects (no such increase has been reported for height, but to our knowledge this has not been investigated in detail) and the known Mendelian forms of dwarfism or gigantism for height (no Mendelian forms of schizophrenia have been reported to our knowledge), we do not know of other empirical evidence to say that the underlying architecture of these traits are very different. Results from GWAS (Tables 1 and 2) suggest that for schizophrenia common associated SNPs exist.

We have performed a thought-experiment assuming that the underlying genetic architecture of schizophrenia is similar to that of height. If we take individuals about 2.5 s.d.'s above the mean for height (approximately 1.97 m for men), then ~0.7% of the population is 'affected' with being tall. The GIANT discovery sample of 130 000 people would contain only 910 such cases. A case-control study of 910 cases and 910 controls for being tall would have no power to detect any of the reported findings in Lango Allen *et al.*⁴³—the power to detect the largest effect (an SNP explaining 0.3% of variance) is about 0.02 (using results from ref.⁴⁸ and assuming a type-I error rate of 10^{-8}). If we were to have ascertained 9400 tall people and 12 500 controls (similar to the number of cases and controls for schizophrenia in the Psychiatric GWAS Consortium schizophrenia study), then the power to detect an SNP explaining 0.03% of variance is 0.038, so approximately seven such loci would be detected out of a total of 180. Also, assuming that genome-wide genetic values for height are normally distributed, the polygenic model would predict a discordance of about 70% between MZ twins of being tall (Figure 1), consistent with the observed discordance rate of MZ twins for schizophrenia that is ~40–60%.¹⁹ To our knowledge, nobody has a problem with accepting that for those ascertained 'probands' who are taller than 197 cm and who have an MZ twin, that the majority of their twins are

smaller than 197 cm. The standard deviation of the difference between MZ pairs for height is about 6 cm, and is usually ascribed to random environmental effects. Therefore, even if epigenetic changes^{49,50} or other complex biological processes mediate environmental factors that result in MZ discordance, such processes are not needed to explain the stochasticity, which underlies affection status.

The comparison between height and schizophrenia does not prove anything as such. However, it shows that the results for height and schizophrenia in terms of the number of variants detected by GWAS are consistent with the power of the experiments to date if their genetic architecture were to be similar. To argue that GWAS has not worked for psychiatric disease is premature because the same could have been said for GWAS on height when the first variants were detected that explained a tiny fraction of the total variation, or if height had been dichotomised in tall vs non-tall individuals. Approximately 50 000 cases and 50 000 controls for an association study of schizophrenia are needed to achieve the same power as an association study of 180 000 individuals measured for height.⁴⁸

Missing, hiding and explained heritability

The term 'missing heritability' was coined to emphasise the discrepancy between total heritability, as estimated from family data, and the proportion of phenotypic variation explained by all detected SNPs from GWAS that are associated with the trait with stringent (genome-wide) significance.³⁸ There are two possible not mutually exclusive explanations for this heritability gap: causal variants are not tagged by SNPs on the commercial arrays or they are tagged, but their effect size is so small that genome-wide significance is not achieved. Using height as an example, we recently tried to distinguish between these two explanations.³⁹ About 10% of phenotypic variation in height is explained by 180 genome-wide significant loci,⁴³ using a very large discovery sample of >130 000. We found that all the genotyped SNPs together explain about 50% of the phenotypic variance, implying that there are many variants with effects so small (<1 mm) that they do not reach statistical significance, even with very large sample sizes. We also provided evidence that the allele frequencies of causal variants are, on average, lower than the SNPs on the genotyping arrays.³⁹ So the conclusion for height is that most genetic variation is 'hidden' and not 'missing', in that larger sample sizes will uncover more statistically significant variants.

The results for height are remarkably similar to those for schizophrenia: Purcell *et al.*⁵¹ calibrated their GWAS results with accuracy of prediction in independent case-control samples, and concluded that about 1/3 of variation in liability of schizophrenia was captured by all SNPs together.⁵¹ Subsequent analyses on the same data using the methodology used in the height paper has confirmed that

~30–40% of liability to schizophrenia (unpublished results) and bipolar disorder⁵² is explained by all SNPs together.

In a risk prediction framework, one can estimate genetic effects in a discovery sample, create a predictor from the estimated effects for each individual in an independent data set and correlate the predictor with observed outcome (for example, disease status).^{51,53} One measure to quantify how well the predictor performs is how much variation in outcome it explains,⁵¹ which is typically carried out by performing a (logistic) regression analysis of case–control status on the predictor (‘genomic profile’). When applied to schizophrenia or bipolar disorder, variation of case–control status explained from GWAS SNPs in a subsequent study has resulted in a small proportion of variation in outcome explained (about 1–3% in ref.⁵¹). Yet, the proportion in case–control status that is captured by all SNPs is much larger (about 34% in ref.⁵¹). This is not a paradox. The reason for this observation is that individual SNP effects are estimated with much error because effect sizes at individual SNPs are very small. Estimating the total amount of variation due to all genotyped SNPs does not suffer from these small effect sizes, but the precision of prediction does.⁴¹

What about major depressive disorder (MDD)?

For a review claiming to cover MDD, we have made little mention of it. There have been seven GWAS for MDD,^{54–60} but no associations have reached genome-wide significance or have been solidly replicated. Why might this be so? Firstly, case–control studies of the same size for different disorders do not have the same power; the difference in mean liability between cases and controls is much less for MDD (1.6 liability standard deviation units assuming prevalence of 15% and screened controls) compared with, for example, schizophrenia (2.8 standard deviation units for prevalence 0.7%). Therefore, larger sample sizes are needed for MDD to detect variants that explain the same proportion of variance in liability. Moreover, if we can assume that the *number* of genetic variants underlying MDD is the same as for schizophrenia, and if those variants have the same or more common frequency distribution (which seem like reasonable assumptions) than those affecting schizophrenia, then the effect sizes for MDD must be smaller if the heritability is lower. Sample sizes need to be ~4–5 times greater for MDD compared with schizophrenia to explain the same proportion of *genetic* variance (see ref.⁵⁹ for detailed calculations). Therefore, if sample sizes of 50 000 schizophrenia cases and 50 000 controls are needed to explain the same proportion of variance as 180 000 individuals measured for height, then 200 000 cases and 200 000 controls are needed for MDD. Are such sample sizes unreasonable? Not if there is a willingness to achieve them. Large samples are being generated for other less prevalent diseases; already in 2007 a breast cancer

study had brought together over 20 000 cases and 20 000 (prevalence ~5% in women only).⁶¹ The heterogeneous phenotype is, in part, responsible for the high prevalence and lower heritability for MDD, although many of the GWAS for MDD have already tried to focus on the less prevalent, but more heritable early onset¹⁴ and/or recurrent¹⁵ MDD. Any subtyping strategy has to contend with the unknown balance between power gained from increased genetic homogeneity vs the power lost from smaller sample sizes. Quantitative scores of reliability and severity may offer the best strategy to optimise this balance, but requires consistent recording across collaborating groups. Gains may also be made by selecting more homogeneous environmental exposures, and with this in mind a new international consortium for post-partum depression is being formed. Investments into the collection of cohorts that are both environmentally and genetically informative, and which are longitudinal over the critical period of childhood through adolescence to early adulthood, are essential if we are serious about dissecting the aetiology of MDD. Whatever strategy is taken, large sample sizes will be needed.

An unhelpful legacy of Mendelian disease genetics

In our view, it is rather unfortunate that many of the terms from epidemiology that pertain to Mendelian diseases have been carried over to complex common disease, including psychiatric disorders. For ultra-rare Mendelian diseases, it makes sense to focus on the few families that are affected and to quantify the penetrance of a disease mutation (the (high) probability of becoming affected given one’s genotype). For the same rare Mendelian disease, probands from different families can have different mutations, often in the same gene, and these mutations can cause differences in the phenotype. Therefore, it is logical to use terms such as ‘allelic heterogeneity’ (different mutations, same gene), ‘locus heterogeneity’ (same disease, different gene), ‘phenotypic heterogeneity’ (same gene or same mutation, different phenotype) and ‘phenocopy’ (same phenotype, different causation).

For most complex diseases, including psychiatric disease, these terms are unhelpful because their use implicitly assume that there is a single ‘cause’ (single mutation) of the disease, that is, they assume pseudo-Mendelian disease models. From the empirical evidence on segregation patterns within families, the recurrence risk to relatives and the evidence from GWAS and CNV analyses, it appears beyond doubt that most psychiatric disease is not Mendelian in the sense that they are not caused solely by a single mutation (others disagree with this conclusion^{28,62}). That is not to say that Mendelian forms cannot exist, but if they do, they account for very little of the population variance. The focus on (nuclear) families, for example, by distinguishing between ‘sporadic’ and ‘familial’ cases, may be inefficient. When there

are multiple loci involved in disease that are segregating in the population, the nuclear family is not a natural unit anymore. Different families are related to each other (we all are, and we do not have to go back far in time to find a common ancestor) and *de novo* variants of small effect get passed on to next generations and some become part of the standing variation. Yes, families with multiple early-onset cases can be enriched with variants of large effects, but these families are also likely to be enriched for variants of small effects.⁶³ Yes, psychiatric patients from families with no family history of psychiatric disorders may be, on average, more likely to have had their genetic risk enhanced by a *de novo* mutation of large effect compared to patients with family history, but ‘sporadic’ cases are expected to comprise 70% of cases even under a highly polygenic genetic architecture.⁵

Instead of using the vague term ‘heterogeneity’, which is neither helpful nor biologically relevant, it is better to describe the genetic architecture as the number of segregating variants, their effect sizes (ideally on multiple diseases) and their interaction within and between loci. Similarly, ‘effect size’ (for example, an odds ratio or relative risk) can encompass and replace the term ‘penetrance’. We note that a highly polygenic architecture of many variants of small effect implies that each individual may carry a unique portfolio of risk alleles, so there is ‘genetic heterogeneity’ in the sense that any two affected individuals are highly unlikely to carry the same combination of risk alleles. Figure 4 illustrates

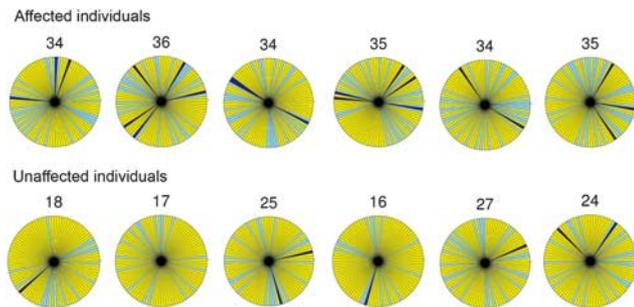


Figure 4 Visualising the unique genetic profiles consistent with a genetic architecture of many variants of small effect. Each pie is divided into 100 segments, representing 100 risk loci. Each risk allele has frequency 0.1, so that homozygotes for the non-risk alleles (yellow segments) have frequency $0.9 \times 0.9 = 0.81$, heterozygotes (blue segments) have frequency $2 \times 0.1 \times 0.9 = 0.18$ and homozygotes for the risk alleles (black segments) have frequency $0.1 \times 0.1 = 0.01$. On average, individuals in the population carry $2 \times 100 \times 0.1 = 20$ risk alleles. The standard deviation (s.d.) for the number of risk alleles is $\sqrt{(2 \times 100 \times 0.1 \times 0.9)} = 4.2$, so that the range of the mean ± 3 s.d. is 7–33 risk alleles. Here we show the diversity in risk allele profiles in individuals who each carry >33 risk alleles (‘affected’) compared with unaffected individuals. The pie charts illustrate the heterogeneity between individuals in their portfolios of risk alleles and illustrate that even unaffected individuals carry a mutational load.

that both cases and controls carry alleles that increase susceptibility to disease, but that cases carry more such alleles. However, this model is consistent with a genetic architecture that is classically described as ‘homogeneous’, in that the distribution of genetic liability is normally distributed. Portfolios of risk alleles between related individuals will be correlated consistent with correlated symptom portfolios of relatives.⁶⁴

Shared genetic aetiology

There is growing evidence for a shared genetic aetiology for psychiatric disorders both from family studies,^{65,66} national record linkage studies^{18,67} as well as emerging evidence from GWAS.^{51,68,69} These observations validate the persistent questioning of the representativeness of the dichotomous classification system between schizophrenia and bipolar disorder in favour of continuous dimensional scales that may have more biological validity.^{70,71} A genetic architecture of many causal variants underlying psychiatric disorders is consistent with a spectrum of symptom profiles, but where individuals with correlated symptom profiles may have correlated genetic profiles, making the boundaries between diagnostic categories rather blurred. Craddock *et al.*⁷² provided a powerful paradigm on using symptom profiles. They found that that an associated SNP in the GABA_A receptor gene GABRB1 identified in a GWAS of bipolar disorder was most highly associated with the subset of cases classified under the Research Diagnostic Criteria as schizoaffective disorder, bipolar type. They went on to test and find support for the hypothesis that this bipolar subtype would show associations in other GABA_A receptor genes, but at a level of association that would never have survived multiple testing in a GWAS using the small number of cases that comprise the subtype class. This approach shows that relationships between symptom profiles both within and between diagnostic classes should be achievable, but only with very large sample sizes and with detailed recording of symptom profiles available for all cases.

The availability of high-density genetic data (GWAS or sequencing data) for many different diseases and traits facilitates the estimation of genetic co-morbidity between diseases, at the level of individual loci but also genome wide,⁵¹ when such studies are difficult or impossible to do for low prevalence diseases in family designs. In the context of the liability threshold model, genetic correlations between the liabilities for different diseases can be estimated, just as can be carried out for quantitative traits with pedigree data, for example, by estimating the genetic correlation between height and weight. This is an exciting prospect and is likely to lead to a better understanding of both the evolution and standing genetic variation of psychiatric disease. Importantly, by utilising cases ascertained for different diseases who are ‘unrelated’ in the conventional sense, it can be assumed that a correlation in liability is not

confounded with (shared) environmental factors, which simplifies the estimation and interpretation of any observed co-morbidity.

Future research and concluding remarks

This review was restricted to inference about the nature of genetic variation underlying psychiatric disease, in particular with respect to schizophrenia, bipolar disorder and major depression. There are, of course, a lot more research topics that are relevant to the genetics of psychiatric disease that we have not touched upon. These include, but are not restricted to, the importance of phenotype classification and measurement, genetic dissection of endophenotype for which the genetic architecture may be different than for disease itself, the interplay between environmental and genetic risk factors, the importance of understanding the biology of risk variants, irrespective of allele frequency in the population, the importance of biological pathways and gene networks, and the genetic basis of response to treatment.

Psychiatric diseases are usually not caused solely by a mutation in a single gene but, like other complex traits, are controlled by many genes and by environmental factors. GWAS have found SNPs and CNVs associated with some of the causal variants, but each one explains only a very small amount of the variance and together they only explain a fraction of the known genetic variance. Some of the associated SNPs are common and most likely track common causal variants, whereas the associated CNVs are rare. The 'missing heritability' is most likely due to a combination of common causal variants with effects too small to be detected in current GWAS and causal variants that are rare and hence not in high LD with SNPs on the commercial chips. However, it is unlikely that any of these causal variants explain a large part of the genetic variance. Therefore, large sample sizes will be required to detect them even if we can assay for the causal variants, for instance, by sequencing.

We appear to be in uncertain times with respect to experimental strategies to further our understanding in the genetic basis of psychiatric disease, because on the one hand the enthusiasm from funding bodies to support more GWAS is waning, yet on the other hand the exciting prospects of exome or whole genome sequencing studies have not yet delivered new genes or pathways (although it is early days). As we have tried to emphasise in this review, we do not have sufficient empirical data at present to quantify how much genetic variation is contained in rare variants or, in other words, how strong the relationship is between effect size and the frequency of risk alleles. What we do know is that a substantial fraction of genetic variation is in LD with SNPs on commercial genotyping arrays. Giving up on an experimental design (GWAS) that has resulted in new biological knowledge and has provided evidence for the polygenic nature of a number of psychiatric diseases seems premature, in particular because predictions

about gene discovery and genetic variation explained as a function of experimental sample size have proven correct. The same (il)logic would have stopped researchers in human height and other quantitative traits from pursuing ever larger sample sizes, at the detriment of scientific discovery and progress. Although whole genome sequencing for all samples is our aspiration, with limited research funds, we would argue that SNP genotyping of large numbers of samples (and using imputation) will provide more insight into the biology of mental disorders than genome sequencing of a smaller number of individuals. In genome sequencing, it will be hard to separate the wheat from the chaff unless sample sizes are large.

The important conclusion is that, whether SNP genotyping or sequencing is used, large sample sizes (10 000s of cases and controls) will be required because individual variants explain a very small proportion of the variance. The small effect sizes may mean that the path from discovery of variants to impact on patients may be less straightforward than pre-GWAS (optimistic) expectations, but we must be careful not to apologise for the true state of nature. There is no doubt that identification of more causal variants, whatever their effect size and allele frequency, will impact on the much-needed understanding of the underlying biology of mental disorders. Larger sample sizes that are phenotypically well documented will not just benefit genetic research, but will allow thorough investigation in environmental factors (which also have a frequency and effect size) and in particular into robust research into gene–environment interactions.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

We acknowledge funding from the Australian National Health and Medical Research Council (Grants 389892, 442915, 496688, 613672 and 613601) and the Australian Research Council (Grants DP0770096 and DP1093900 and Future Fellowship to NRW). We thank Adam Eyre-Walker for discussions and correspondence regarding computing Zeta functions, and Bill Hill for helpful comments. We dedicate this paper to the memory of Charlie Smith and Douglas Falconer, who had it all worked out 40 years ago.

References

- 1 McGrath JJ. The surprisingly rich contours of schizophrenia epidemiology. *Arch Gen Psychiatry* 2007; **64**: 14–16.
- 2 Gottesman II, Shields J. A polygenic theory of schizophrenia. *Proc Natl Acad Sci USA* 1967; **58**: 199.
- 3 McGue M, Gottesman II, Rao DC. Resolving genetic models for the transmission of schizophrenia. *Genet Epidemiol* 1985; **2**: 99–110.

- 4 Wray NR, Visscher PM. Narrowing the boundaries of the genetic architecture of schizophrenia. *Schizophr Bull* 2010; **36**: 14–23.
- 5 Yang J, Visscher PM, Wray NR. Sporadic cases are the norm for complex disease. *Eur J Hum Genet* 2010; **18**: 1039–1043.
- 6 Smith C. Heritability of liability and concordance in monozygous twins. *Ann Hum Genet* 1970; **34**: 85–91.
- 7 Slatkin M. Exchangeable models of complex inherited diseases. *Genetics* 2008; **179**: 2253–2261.
- 8 Wray NR, Goddard ME. Multi-locus models of genetic risk of disease. *Genome Med* 2010; **2**: 10.
- 9 Dempster ER, Lerner IM. Heritability of threshold characters. *Genetics* 1950; **35**: 212–236.
- 10 Wray NR, Visscher PM. Narrowing the boundaries of the genetic architecture of schizophrenia. *Schizophr Bull* 2010; **36**: 14–23.
- 11 Cannon TD, Kaprio J, Lonqvist J, Huttunen M, Koskenvuo M. The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study. *Arch Gen Psychiatry* 1998; **55**: 67–74.
- 12 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.
- 13 Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry* 2000; **157**: 1552–1562.
- 14 Weissman MM, Wickramaratne P, Merikangas KR, Leckman JF, Prusoff BA, Caruso KA et al. Onset of major depression in early adulthood—increased familial loading and specificity. *Arch Gen Psychiatry* 1984; **41**: 1136–1143.
- 15 McGuffin P, Katz R, Watkins S, Rutherford J. A hospital-based twin register of the heritability of DSM-IV unipolar depression. *Arch Gen Psychiatry* 1996; **53**: 129–136.
- 16 Kendler KS, Neale MC, Kessler RC, Heath AC, Eaves LJ. The lifetime history of major depression in women—reliability of diagnosis and heritability. *Arch Gen Psychiatry* 1993; **50**: 863–870.
- 17 Foley DL, Neale MC, Kendler KS. Reliability of a lifetime history of major depression: implications for heritability and co-morbidity. *Psychol Med* 1998; **28**: 857–870.
- 18 Lichtenstein P, Yip BH, Bjork C, Pawitan Y, Cannon TD, Sullivan PF et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet* 2009; **373**: 234–239.
- 19 Cardno AG, Gottesman II. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *Am J Med Genet* 2000; **97**: 12–17.
- 20 Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*, 1st edn. Sinauer Associates: Sunderland, MA, 1998.
- 21 Falconer DS, Mackay TFC. *Introduction to Quantitative Genetics*, 4th edn. Longman: Harlow, Essex, UK, 1996.
- 22 Zhang XS, Hill WG. Multivariate stabilizing selection and pleiotropy in the maintenance of quantitative genetic variation. *Evolution* 2003; **57**: 1761–1775.
- 23 Laursen TM, Munk-Olsen T. Reproductive patterns in psychotic patients. *Schizophr Res* 2010; **121**: 234–240.
- 24 Uher R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. *Mol Psychiatry* 2009; **14**: 1072–1082.
- 25 Kimura M. Evolutionary rate at the molecular level. *Nature* 1968; **217**: 624–626.
- 26 Fu YX. Statistical properties of segregating sites. *Theor Popul Biol* 1995; **48**: 172–197.
- 27 Eyre-Walker A. Genetic architecture of complex traits and its implications for fitness and genome-wide association studies. *Proc Natl Acad Sci USA* 2010; **107**: 1752–1756.
- 28 McClellan JM, Susser E, King MC. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry* 2007; **190**: 194–199.
- 29 Craddock N, O'Donovan MC, Owen MJ. Phenotypic and genetic complexity of psychosis—invited commentary on schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry* 2007; **190**: 200–203.
- 30 Lander ES. The new genomics: global views of biology. *Science* 1996; **274**: 536–539.
- 31 Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001; **17**: 502–510.
- 32 O'Donovan MC, Craddock NJ, Owen MJ. Genetics of psychosis; insights from views across the genome. *Hum Genet* 2009; **126**: 3–12.
- 33 Kirov G. The role of copy number variation in schizophrenia. *Expert Rev Neurother* 2010; **10**: 25–32.
- 34 Wray NR, Purcell S, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol* 2011; **9**: e1000579.
- 35 Bassett AS, Marshall CR, Lionel AC, Chow EWC, Scherer SW. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. *Hum Mol Genet* 2008; **17**: 4045–4053.
- 36 Shaikh TH, Kurahashi H, Saitta SC, O'Hare AM, Hu P, Roe BA et al. Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet* 2000; **9**: 489–501.
- 37 Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; **19**: 149–150.
- 38 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ et al. Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
- 39 Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
- 40 Goddard ME, Wray NR, Verbyla K, Visscher PM. Estimating effects and making predictions from genome-wide marker data. *Statist Sci* 2009; **24**: 517–529.
- 41 Visscher PM, Yang J, Goddard ME. A commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010). *Twin Res Hum Genet* 2010; **13**: 517–524.
- 42 Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010; **42**: 565–569.
- 43 Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; **467**: 832–838.
- 44 Kirov G, Grozeva D, Norton N, Ivanov D, Mantripragada KK, Holmans P et al. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* 2009; **18**: 1497–1503.
- 45 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.
- 46 Stone JL, O'Donovan MC, Gurling H, Kirov GK, Blackwood DHR, Corvin A et al. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008; **455**: 237–241.
- 47 Stefansson H, Rujescu D, Cichon S, Pietilainen OPH, Ingason A, Steinberg S et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008; **455**: 232–U261.
- 48 Yang J, Wray NR, Visscher PM. Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet Epidemiol* 2010; **34**: 254–257.
- 49 Haque FN, Gottesman II, Wong AH. Not really identical: epigenetic differences in monozygotic twins and implications for twin studies in psychiatry. *Am J Med Genet C* 2009; **151C**: 136–141.
- 50 Kato T, Iwamoto K, Kakiuchi C, Kuratomi G, Okazaki Y. Genetic or epigenetic difference causing discordance between monozygotic twins as a clue to molecular basis of mental disorders. *Mol Psychiatry* 2005; **10**: 622–630.
- 51 Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–752.
- 52 Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011; **88**: 294–305.
- 53 Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* 2007; **17**: 1520–1528.
- 54 Muglia P, Tozzi F, Galwey NW, Francks C, Upmanyu R, Kong XQ et al. Genome-wide association study of recurrent major

- depressive disorder in two European case-control cohorts. *Mol Psychiatr* 2008; **15**: 589–601.
- 55 Sullivan PF, de Geus EJC, Willemsen G, James MR, Smit JH, Zandbelt T *et al*. Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol Psychiatr* 2009; **14**: 359–375.
- 56 Shi J, Potash JB, Knowles JA, Weissman MM, Coryell W, Scheftner WA *et al*. Genome-wide association study of recurrent early-onset major depressive disorder. *Mol Psychiatr* 2010; **16**: 193–201.
- 57 Shyn SI, Shi J, Kraft JB, Potash JB, Knowles JA, Weissman MM *et al*. Novel loci for major depression identified by genome-wide association study of sequenced treatment alternatives to relieve depression and meta-analysis of three studies. *Mol Psychiatr* 2009; **16**: 202–215.
- 58 Lewis CM, Ng MY, Bulter AW, Cohen-Woods S, Uher R, Pirlo K *et al*. Genome-wide association study of major depression in the UK population. *Am J Psychiatry* 2010; **167**: 949–957.
- 59 Wray NR, Pergadia ML, Blackwood DH, Penninx BW, Gordon SD, Nyholt DR *et al*. Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol Psychiatry* 2010; *Molecular Psychiatry*; advance online publication, 2 November 2010; doi:10.1038/mp.2010.109.
- 60 Rietschel M, Mattheisen M, Frank J, Treutlein J, Degenhardt F, Breuer R *et al*. Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. *Biol Psychiatry* 2010; **68**: 578–585.
- 61 Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG *et al*. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007; **447**: 1087–1093.
- 62 Mitchell KJ, Porteous DJ. Rethinking the genetic architecture of schizophrenia. *Psychol Med* 2011; **41**: 19–32.
- 63 Byrnes GB, Southey MC, Hopper JL. Are the so-called low penetrance breast cancer genes, ATM, BRIP1, PALB2 and CHEK2, high risk for women with strong family histories? *Breast Cancer Res* 2008; **10**: 208.
- 64 Korszun A, Moskva V, Brewster S, Craddock N, Ferrero F, Gill M *et al*. Familiality of symptom dimensions in depression. *Arch Gen Psychiatry* 2004; **61**: 468–474.
- 65 Pope HG, Yurgeluntodd D. Schizophrenic individuals with bipolar 1st-degree relatives—analysis of 2 pedigrees. *J Clin Psychiatry* 1990; **51**: 97–101.
- 66 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.
- 67 Gottesman II, Laursen TM, Bertelsen A, Mortensen PB. Severe mental disorders in offspring with 2 psychiatrically ill parents. *Arch Gen Psychiatry* 2010; **67**: 252–257.
- 68 Moskva V, Craddock N, Holmans P, Nikolov I, Pahwa JS, Green E *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.
- 69 Green EK, Grozeva D, Jones I, Jones L, Kirov G, Caesar S *et al*. The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Mol Psychiatr* 2010; **15**: 1016–1022.
- 70 Craddock N, Owen MJ. The beginning of the end for the Kraepelinian dichotomy. *Br J Psychiatry* 2005; **186**: 364–366.
- 71 Craddock N, Owen MJ. Rethinking psychosis: the disadvantages of a dichotomous classification now outweigh the advantages. *World Psychiatry* 2007; **6**: 20–27.
- 72 Craddock N, Jones L, Jones IR, Kirov G, Green EK, Grozeva D *et al*. Strong genetic evidence for a selective influence of GABA(A) receptors on a component of the bipolar disorder phenotype. *Mol Psychiatr* 2010; **15**: 146–153.
- 73 Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet* 1965; **29**: 51–71.
- 74 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.
- 75 Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* 2008; **4**: e1000008.
- 76 O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskva V *et al*. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 2008; **40**: 1053–1055.
- 77 Shi JX, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I *et al*. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* 2009; **460**: 753–757.
- 78 Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D *et al*. Common variants conferring risk of schizophrenia. *Nature* 2009; **460**: 744–U799.
- 79 Shifman S, Johannesson M, Bronstein M, Chen SX, Collier DA, Craddock NJ *et al*. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. *Plos Genetics* 2008; **4**: e28.
- 80 Kirov G, Rujescu D, Ingason A, Collier DA, O'Donovan MC, Owen MJ. Neurexin 1 (NRXN1) deletions in schizophrenia. *Schizophr Bull* 2009; **35**: 851–854.
- 81 Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM *et al*. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008; **320**: 539–543.
- 82 A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies/ (accessed on 13 December 2010).
- 83 Scott LJ, Muglia P, Kong XQ, Guan W, Flickinger M, Upmanyu R *et al*. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc Natl Acad Sci USA* 2009; **106**: 7501–7506.
- 84 Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L *et al*. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* 2008; **40**: 1056–1058.
- 85 Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B *et al*. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* 2008; **13**: 197–207.
- 86 WTCCC. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.