

Genetic Basis of Complex Genetic Disease: The Contribution of Disease Heterogeneity to Missing Heritability

Naomi R. Wray · Robert Maier

Published online: 30 September 2014
© Springer International Publishing AG 2014

Abstract The genetic basis of complex genetic disease can be quantified by heritability, which is an estimate of the relative importance of genetic and non-genetic factors in contributing to differences between individuals for any given trait. Heritability is estimated from phenotypic records in data sets of families and represents contributions from genetic variants across the frequency spectrum and of any kind and function. Advances in technology allow direct interrogation of some kinds of DNA variants. Specific DNA variants identified in the era of genome-wide association studies explain only a fraction of the heritability estimated from family studies, as do less common variants identified through whole exome sequencing. If true effect sizes of risk variants are small, studies to date may be underpowered to detect individual risk variants; but the studies may be well-powered to detect the total contribution from common risk variants, and this has explained some of the missing heritability. Here we review explanations for the so-called “still-missing heritability” and focus particularly on the issue of genetic heterogeneity.

Keywords Genome-wide association studies · Genetic heterogeneity · Whole exome sequencing · Genetic architecture · De novo mutations

Introduction

Complex genetic diseases are those that tend to “run” in families yet show no clear pattern of inheritance. Most common diseases are complex genetic diseases including cancers,

heart disease, immune disorders, and psychiatric disorders. Our understanding of causality of these diseases is limited, and this lack of knowledge has contributed to the limited progress made in the development of new treatments. Traditionally, quantification of the genetic basis of disease has been determined by measuring the increased risk of disease in relatives of those affected. Evidence for a genetic risk shared between relatives implies that DNA risk variants are passed from parent to child. This knowledge has underpinned the philosophy that identification of genetic risk variants is a worthy goal that may expose and open new doors towards understanding of the causality of disease, which in turn may lead to new treatments. Strategies to identify DNA risk variants have been dictated by available genotyping technologies. Advances in technology of the last decade have delivered methodologies, notably genome-wide association studies (GWAS) and whole exome sequencing (WES) that have started to deliver DNA risk variants associated with disease. Here we review the portfolio of strategies used to understand the genetic contribution to complex disease. We close by focussing on the issue of genetic heterogeneity of disease.

Heritability

Evidence for a genetic contribution to disease comes from measurement of an increased risk of the disorder in relatives of those affected. However, such increased risks need to be interpreted with care, since close relatives share a common family environment so that recurrence risk in relatives may also reflect non-genetic factors. Estimates of risks of disease in different types of relatives (e.g., monozygotic and dizygotic twins, first and second degree relatives) are needed to disentangle genetic from non-genetic factors. These risks to relatives are used to estimate heritability on the liability scale [1, 2]. Liability to disease is a non-observable or

N. R. Wray (✉) · R. Maier
Queensland Brain Institute, The University of Queensland,
Brisbane, Queensland, Australia
e-mail: Naomi.wray@uq.edu.au

latent, continuous variable with those ranking highest on liability being affected. Heritability on the liability scale, h^2 , quantifies the proportion of variance of liability to disease attributable to inherited genetic factors. Comparison of the relative importance of genetic factors for different disorders is more intuitive on this scale, particularly when comparing diseases of a different lifetime risk. Heritability accounts for genetic factors that are additive on the liability scale; these genetic factors combine non-additively on the disease scale [3], so that the probability of disease is many times higher for individuals carrying a high number of risk alleles compared to those carrying only half the number. Non-genetic factors include identifiable (but perhaps not recorded) environmental factors or measurement error, but also unidentifiable factors which form an intrinsic stochastic noise. Estimates of heritability may vary between populations, across ages and may depend on whether non-genetic factors have been recorded and included in the analysis [4]. They depend on baseline risk of disease in the population, and the degree of sampling variance is often overlooked. Hence, in reality, heritability estimates should be viewed as pragmatic benchmarks representing evidence for low, moderate or high contributions of genetic effects.

Genetic Architecture

While heritability on the liability scale expresses the proportion of the variance in liability that is attributable to genetic factors, it tells nothing about the underlying genetic architecture of the disease in terms of number, frequency, and effect sizes of individual causal variants, nor of the mode of action of causal loci (i.e., additive or non-additive). Lack of evidence that complex disease cases represented single gene disorders generated theories of polygenicity [5]. Empirical results of the last decade provided support for a polygenic model [6]. Under a polygenic model, the liability to disease reflects multiple genetic and non-genetic effects acting additively. Hence, liabilities are assumed to be normally distributed, because such a distribution results from many additively acting effects. All individuals in the population carry some genetic risk variants and likely experience some non-genetic risk factors, but most individuals in the population are not affected. Disease status results when the cumulative load exceeds a burden of risk threshold.

De Novo Mutations

De novo mutations are genetic variants present in the DNA of a child but not of their parents. Genotyping of parents and their child is used to identify de novo mutations. Whole exome sequencing has identified that de novo mutations play

an important role in Mendelian diseases [7]. Effect sizes of de novo mutations, that are their contribution to the risk of disease, are expected to be both small and large. This is not inconsistent with the expectation that genetic variants of large effect size are more likely to be de novo as they have not been subject to selection. Sequencing studies of the last decade have demonstrated that de novo mutations play an important causal role in some complex diseases and disorders for some individuals [8] (for example, mental retardation [9] and autism [10]). For other diseases and disorders there is evidence of an increased burden of de novo mutations in cases compared to controls [11] without being able to identify which of the de novo mutations are individually causal and increase risk of disease versus those that are benign [12]. In rare instances, somatic de novo mutations have been shown to be causal [13]. De novo mutations are not shared between relatives (except possibly between identical twins, or between siblings as a result of germ line mutations in sperm) and so rarely contribute to explaining heritability [14].

Familial vs Sporadic

It is not uncommon for cases to be referred to as either “familial” or “sporadic”, reflecting whether there is a known family history for the disease. In childhood disorders, cases are similarly referred to as multiplex or simplex depending on the presence or the absence of other affected children. In common parlance, the terms tend to be interpreted as implying a genetic or non-genetic etiology of disease, but this can be misleading. On the one hand, knowledge of family history can be used in optimal experimental design. For example, genetic studies designed to identify de novo mutations would be optimised by genotyping of cases with no family history of disease. In contrast, genetic studies designed to identify common genetic risk variants are optimised by prioritising selection of cases with family history and controls with no family history of disease. On the other hand, it is frequently overlooked that under a polygenic genetic architecture the majority of cases are not expected to report family history. For example, for a disease with lifetime prevalence of 1 % and heritability of 80 %, less than half of cases are expected to report family history when considering all first, second, and third generation relatives [15]. Likewise, for the same disease more than 60 % of monozygotic twins are expected to be discordant for disease status [16].

Missing Heritability

Advances in genotyping technology allow cheap genome-wide interrogation of single nucleotide polymorphism (SNPs). GWAS identify associations between SNPs and disease.

Reported results from association analyses include risk allele frequency (RAF), effect size (expressed for disease as the odds ratio, OR) and p -value of association. The contribution of these associated DNA genetic variants to variance can be calculated on the liability scale [17] to allow direct comparison of the contribution to the risk of each locus on the same scale as the heritability is reported. Assuming independence (and ignoring potential overestimation of effect size due to winner's curse), the contribution of each genome-wide significant (GWS) locus can be summed to determine the proportion of variance in liability explained by these loci together, thus, quantifying the effects of all genome-wide significant SNPs (h_{GWS}^2).

Given the stringent significance threshold applied, the ability to detect risk loci (i.e., the power) depends on whether the sample size is sufficient given the true effect sizes. When the first GWAS were planned, the distribution of expected effect sizes was unknown and sample sizes were powered to detect $OR > \sim 1.3$. The first generation of GWAS yielded few GWS results with h_{GWS}^2 much less than h^2 . This difference has been termed “missing heritability” [18]. As sample sizes have increased, the number of GWS variants have increased for both quantitative traits and diseases (see Figure 2 in Visscher et al. [6]) providing empirical evidence that common variants do play a role in complex genetic disease. Nonetheless, substantial missing heritability remains.

Hiding Heritability

The observed increase in number of significant association results as sample sizes have been increased [6], This implies that the earlier studies were underpowered to detect the variants given their effect sizes. However, given that collection of larger samples is time consuming and expensive, can we be sure that the same will be true for other diseases? Statistical methods that combine quantitative and population genetic concepts to evaluate the contribution to variance of common SNPs across the whole genome without identifying them individually have been developed [19–24]. These methods use people unrelated in the conventional sense of the word; but given the finite global population size, they share a proportion of their DNA by descent. The proportion of sharing between pairs of individuals can be estimated using genome-wide marker data, and that genomic similarity can be correlated with disease status to estimate genetic variation [20, 21, 25, 26]. By using distantly related individuals, a significant heritability tagged by common SNPs, h_{SNP}^2 , is detected if case-case pairs and control-control pairs have higher genomic similarity than case-control pairs [26]. For most disease traits studied, significant SNP heritabilities have been estimated demonstrating that, although the data sets analysed may have been underpowered to detect the individual small effects as GWS, contributions from common variants exist.

Larger sample sizes are needed for individual detection. Hence, the polygenic analyses have been successful in identifying “hidden heritability”, i.e., the increase from h_{GWS}^2 to h_{SNP}^2 . In theory, with sufficiently large sample size, h_{GWS}^2 can become as large as h_{SNP}^2 .

Explanations for the Still-Missing Heritability

For most diseases the “still-missing” heritability, i.e., the difference between h_{SNP}^2 and h^2 , remains substantial at approximately half of the heritability estimated from family data. It is important to note that it is not necessary to explain all heritability when the goal is to open new biological research doors that may impact treatment, and; indeed, it is likely to be impossible to do so. Nonetheless, seeking further insight for the still-missing heritability may also provide important guidance of future research directions. A number of explanations have been proposed [19–28], which include the following.

a) Over-estimation of heritability from family studies

In human populations, part of the still-missing heritability may simply reflect overestimation of h^2 since typical study designs for estimation of heritability use very close relatives (e.g., full siblings and twins) who share non-additive gene combinations and a common environment, and these confounding factors can be difficult to separate [4, 18]. The difference between estimates of h^2 from family data and the “true” h^2 has been termed “phantom heritability” [29] when the difference is attributable to non-additive genetic variance, but our ability to quantify this based on realistically collectable data is limited. Others have argued that the contribution from non-additive genetic variance to complex traits is likely limited [30, 31], and the presence of important epistasis and small epistatic variance are not inconsistent [32]. The extent to which gene-environment interaction (GxE) or G and E correlation inflate estimates of heritability from twin and family studies is unknown. Nonetheless, it seems intuitive that exposure to environmental risk factors increases risk of disease only in those that are already genetically susceptible and; hence, SNP effect sizes may differ in cases stratified by an environmental exposure. However, GxE studies to date are limited by a dearth of samples that are informative for G and consistently recorded E, and are notoriously underpowered [33]. For this reason, studies of candidate GxE interactions have generally lacked replication, and the field is plagued by publication bias towards studies with positive results [34].

b) Variants not tagged by common SNPs

Part of the still-missing heritability must reflect genomic variants not well tagged by SNPs [21, 27]. Since the SNPs on

SNP chips are generally chosen because both their alleles are common, they cannot be in high r^2 linkage disequilibrium with rare causal variants. For many diseases, copy number variants or other rare variants have been identified, usually through WES studies. In order to have been detected these rare variants requires relatively large effect size; but because they are rare, their contribution to risk in the population is small. A very large number of rare variants are needed to explain the still-missing heritability. For example, a locus with risk allele frequency 0.001 and heterozygous relative risk (RR) of 2.1 explains approximately the same proportion of variance in liability as a locus with allele frequency 0.5 and RR 1.05. It is notable that estimation of h_{SNP}^2 using SNPs imputed to the 1000 Genomes reference panel does not tend to generate higher estimates compared to imputation to the HapMap3 panel [35, 36]. The relative importance of small structural variants to genomic variation is currently not well documented and may not be well represented in sequenced reference panels used for imputation. Since recurrent tandem repeat polymorphisms are known to modulate a range of biological functions [37, 38], these may represent an example of an important, but as yet unprobed, source of disease associated variation. Estimation of h_{SNP}^2 based on haplotypes constructed from SNPs is a field of active research since haplotypes have the opportunity to tag uncommon structural variants not present in imputation reference panels. In practice, such methods may be difficult to apply since they are likely to be very sensitive to genotyping error.

c) Disease heterogeneity

Disease heterogeneity is a possible explanation for still-missing heritability. We have previously noted, for psychiatric disorders at least, that heritabilities estimated from large population samples are lower than those estimated from twin studies. We argued [39] that this may reflect greater diagnostic heterogeneity in large cohorts compared to the carefully collected twin samples, but that the large cohorts may be more representative of the samples currently brought together for analysis in genetic studies.

Disease heterogeneity can have several interpretations, but at its most tangible there are multiple examples of complex genetic diseases that are now recognised to have biologically determined subtypes reflecting independent, or more likely correlated, diseases that may have different optimal treatment strategies. For example, decades ago, based on clinical symptoms alone, the inflammatory bowel diseases ulcerative colitis and Crohn's Disease would have been indistinguishable and given the same diagnosis. More recently, it has been recognised that diagnosis and treatment of rheumatoid arthritis should consider the presence and the absence of anti-citrullinated protein autoantibodies [40]. The genomics

era has allowed good progress in subtyping of cancers (e.g., ER+ve/ER –ve and overexpression of HER2 as a breast-cancer subtype [41, 42] or K-ras mutations in colorectal cancer and EGFR mutations in lung cancer, as reviewed in [43]). However, other branches of medicine are less able to supply measures of phenotypic heterogeneity in the tissue of relevance for mapping onto the genetic heterogeneity. Given the known examples, it seems likely that other diseases currently treated as a single disease entity may in fact be a diagnostic aggregation of subtypes. How could this impact missing heritability? We consider the impact of disease heterogeneity on estimates of the different parameters of variance explained by genetic factors and demonstrate that it could make an important contribution to still-missing heterogeneity (Fig. 1).

Exploring the Impact of Disease Heterogeneity

To consider the impact of disease heterogeneity on genetic interpretation of disease, we consider an extreme example of two diseases each of lifetime prevalence 0.5 % and heritability 80 % that are phenotypically and genetically independent, but that have such similar clinical presentation that they are indistinguishable and are considered a single disease. Under this composite disease etiology what would be the impact on estimates of h^2 , h_{GWS}^2 and h_{SNP}^2 ?

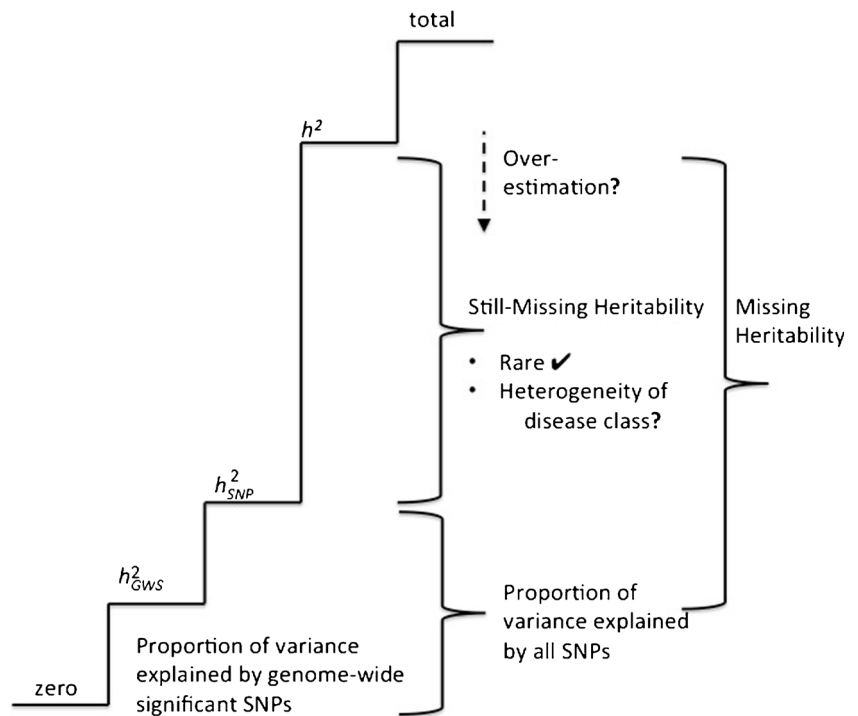
a) Impact on h^2

The composite disease would have a lifetime prevalence of $0.005(2-0.005)=0.998\% \approx 1\%$, and the heritability estimated from the two-disease composite would be estimated as greater than 65 % from a twin design (see Appendix). In fact, for the composite disease the estimates of heritability using the liability threshold model are expected to be slightly inconsistent when estimated from the relative risks of disease from different types of relatives (Fig. 2), but such inconsistencies are expected to be difficult to detect given the sampling error on estimates especially since most studies to estimate heritability use relatively small samples of only twins or first degree relatives [4]. We conclude that high estimates of heritability are possible for a composite disease.

b) Impact on h_{GWS}^2

We have previously provided theory to estimate the power of association studies in the context of misdiagnosis [20] (see Appendix) which is analogous to the scenario here of a disease composite. In Fig. 3 we show the power of an association study to detect risk alleles of a spectrum of frequencies that have effect size under a multiplicative model of heterozygote relative risk 1.15. For a sample of

Fig. 1 A schematic of heritabilities



10,000 cases of a single genetic disease and 10,000 controls we have >75 % power to detect risk alleles of frequencies 0.2–0.8 at a genome-wide significance of 5×10^{-8} (line A). However, for our composite disease (for which we expect risk alleles to be associated with only one of the underlying diseases) an association study of 10,000 cases, of which only half are from the disease impacted by the risk allele, is totally underpowered to detect risk alleles (line B). To demonstrate that this reflects the

impact of contamination by the phenocopy disease rather than the reduced sample size of the associated disease, we also show the power of an association study of 5,000 cases and 10,000 controls (line C). To consider a range of disease composite scenarios when the proportion of disease 2 cases in the disease composite sample is 0 %, 5 %, 10 %, 20 %, and 50 %, the power to detect a disease 1 risk variant of frequency 0.4 and relative risk 1.15 at the genome-wide significance threshold of

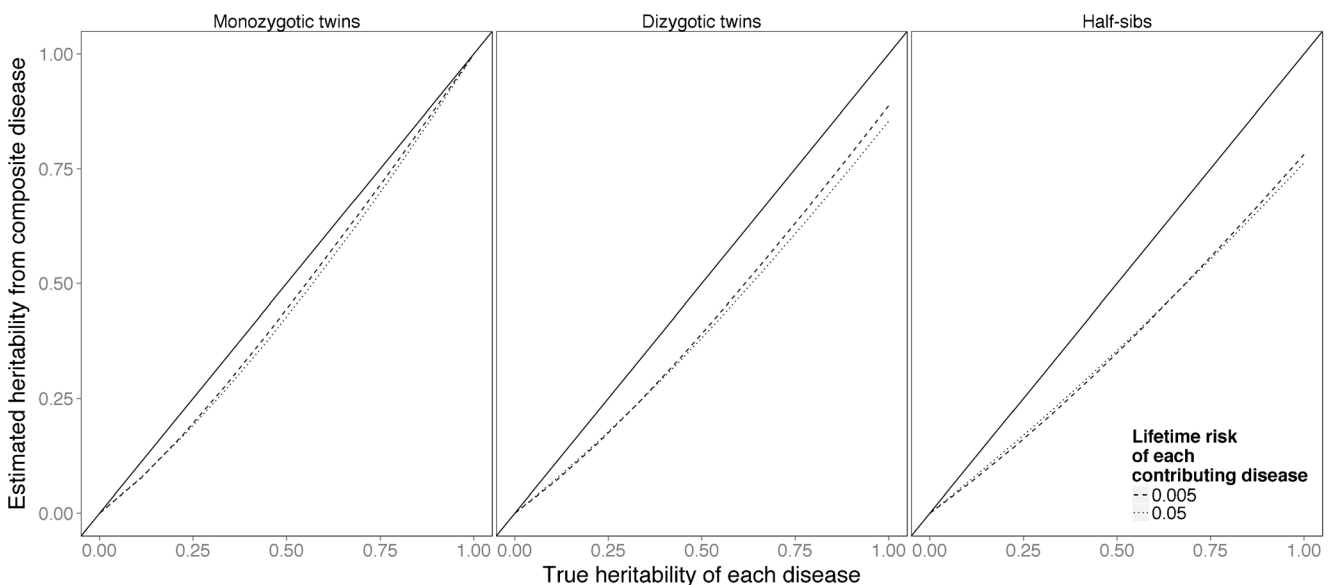


Fig. 2 Estimates of heritability under a liability threshold model calculated from lifetime risk of disease and lifetime risk of disease in relatives of affected individuals for a composite disease that comprises two

independent diseases each of lifetime risk either 5 % or 0.5 % and each of heritability 80 %

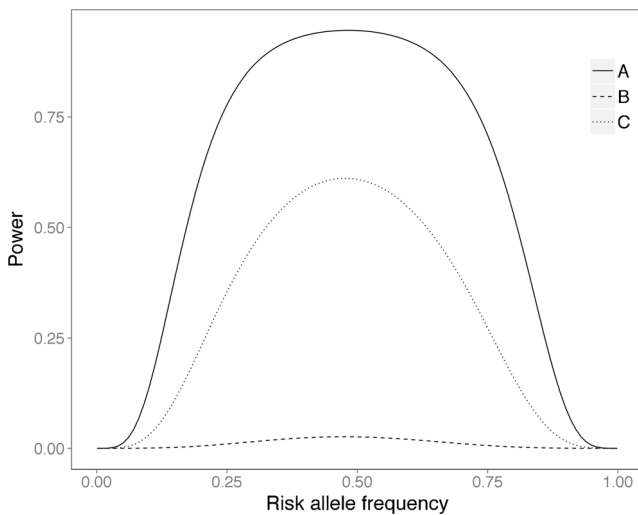


Fig. 3 Power of a genome-wide association study to detect risk variants with heterozygous relative risk of 1.15. A) 10,000 cases of a homogeneous genetic disease of prevalence 1 % and 10,000 screened controls B) 10,000 cases of a composite disease and 10,000 screened controls, the composite disease has prevalence 1 % but comprises two equally represented genetically independent diseases each of prevalence 0.5 % C) 5,000 cases of a homogeneous genetic disease of prevalence 0.5 % and 10,000 screened controls

$p < 5 \times 10^{-8}$ is 93 %, 87 %, 79 %, 55 %, and 3 % (assuming 10,000 composite disease cases and 10,000 controls and 0.5 % lifetime risk of disease 1).

We conclude that disease heterogeneity can severely compromise the power of association studies and, hence, estimation of h_{GWS}^2 .

c) Impact on h_{SNP}^2

The impact of analysing a disease composite to estimate h_{SNP}^2 can also be considered in terms of disease misclassification [20]. The estimated h_{SNP}^2 is a weighted average of the true h_{SNP}^2 parameters of each underlying disease and the SNP-covariance (counted twice). So if the two contributing diseases have equal true h_{SNP}^2 and are independent, the estimated value from the composite disease will be $0.5 h_{SNP}^2$. We conclude that disease heterogeneity can generate underestimates of h_{SNP}^2 compared to when disease classes are genetically homogeneous.

Summary

The genetic basis of complex genetic disease can be quantified by heritability, which is an estimate of the relative importance of genetic and non-genetic factors in contributing to differences between individuals for any given trait. Heritability is estimated from phenotypic records in data sets of families and represents contributions from genetic variants across the frequency spectrum and genetic variants of any kind and

function. Advances in technology allow direct interrogation of some kinds of DNA variants. Specific DNA variants identified in the era of genome-wide association studies explain only a fraction of the heritability estimated from family studies (h_{GWS}^2) as do less common variants identified through whole exome sequencing. If true effect sizes of risk variants are small then studies to date may be underpowered to detect individual risk variants, but they may be well-powered to detect the total contribution from common risk variants (h_{SNP}^2) and such analysis has helped to explain some of the missing heritability. Here we reviewed explanations for the so-called “still-missing heritability” and focus particularly on the issue of disease heterogeneity.

To explore the impact of disease heterogeneity on estimates of h^2 , h_{GWS}^2 and h_{SNP}^2 we considered an extreme example of two independent indistinguishable but equally genetic diseases being lumped together as a disease composite. We have shown that under this scenario the estimates of h^2 from family data are nearly as high as the heritabilities of the contributing individual diseases, yet the estimates of h_{GWS}^2 and h_{SNP}^2 are severely compromised. In reality, this toy example may be too extreme as real presentations of composite diseases may reflect diseases that are genetically correlated rather than totally independent. For example, Crohn’s Disease and ulcerative colitis are estimated to have a genetic correlation based on SNP data of 0.6 [44], the vast majority of SNPs identified in GWAS affect both diseases, but a handful of them have effects in the opposite direction [45]. Clearly, as the genetic correlation between the two contributing diseases approaches 1, the two diseases merge as a single genetic disease entity. For genetically correlated diseases, the power to detect associated loci may be increased by considering the disease composite for loci contributing to both diseases and decreased for other loci. Consideration of these factors can quickly lead to philosophical musings of the definition of disease, since even for a single genetic disease under a polygenic model of disease, each individual could carry a unique portfolio of risk loci. In the genomics era, a disease definition may be at the pathway level, whereby a single genetic disease considers different portfolios of risk loci impacting the same pathway, or, more practically, the class of individuals who respond to the same treatment.

Acknowledgments NR Wray is funded by the Australian National Health and Medical Research Council grants 61602 and 1050218.

Compliance with Ethics Guidelines

Conflict of Interest NR Wray and R Maier both declare no conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Appendix

Estimation of Heritability from a Disease Composite

We define a disease composite as a clinically indistinguishable disease comprising two independent diseases. For illustration and simplicity we assume that the two independent diseases (D_1, D_2) have the same lifetime risk of disease of K and the

$$K_{R,C} = \left(P(D_{R1}|D_1) + P(D_{R2}|D_1) - P(D_{R1}\&D_{R2}|D_1) \right) P(D_1\&!D_2|D_1 \text{ or } D_2) \\ + \left(P(D_{R1}|D_2) + P(D_{R2}|D_2) - P(D_{R1}\&D_{R2}|D_2) \right) P(D_2\&!D_1|D_1 \text{ or } D_2) \\ + \left(P(D_{R1}|D_1\&D_2) + P(D_{R2}|D_1\&D_2) - P(D_{R1}\&D_{R2}|D_1\&D_2) \right) P(D_1\&D_2|D_1 \text{ or } D_2).$$

Assuming equal prevalences and heritabilities for D and D_2 , this reduces to

$$K_{R,C} = KK_R \left(2 \left((1-K)/K + (1-K_R)^2/K_R \right) + (2-K_R) \right) / (2-K).$$

From K_C and $K_{R,C}$, which are the risks that would be estimable from family data, we can calculate the heritability of liability. These calculations have been checked by simulation.

Impact of Power of an Association Study in the Context of a Disease Composite

As before, define a disease composite as a clinically indistinguishable disease comprising two independent diseases. A locus is expected to be associated with only one of the two underlying diseases. The underlying disease considered has a lifetime risk K . We consider a causal variant for this disease that has the frequency of the risk allele and protective alleles of p and $(1-p)$, respectively, in the population. Let $(1-p)^2$, $2p(1-p)$ and p^2 be the frequencies of the genotypes (in Hardy-Weinberg equilibrium), and the risks of disease in the genotypes are f_0, f_1 , and f_2 . If we assume a multiplicative model on the disease scale, then $f_1=f_0 \gamma$ and $f_2=f_0 \gamma^2$ where γ is the relative risk of the risk allele compared to the protective allele. We can calculate the frequency of the risk alleles in cases (true cases) and screened controls as

$$p_{case} = \frac{p\gamma}{1 + p(\gamma - 1)} \text{ and} \\ p_{control} = \frac{p}{1-K} \left(1 - \frac{K\gamma}{1 + p(\gamma-1)} \right).$$

same heritability of h^2 . From standard liability threshold theory, we can calculate the risk in family members whose relatives of a given degree of kinship are affected, K_R . The lifetime risk of the composite disease is $K_C=K(2-K)$. The risk of either of the underlying disease in relatives of those affected by either of the underlying diseases $K_{R,C}$ can be written in terms of the probabilities of each of the underlying diseases in the relatives (D_{R1}, D_{R2})

If s is the proportion of cases in the association sample that are from the other underlying disease (we assume that the locus is not associated with this disease and that their allele frequency the same as in controls), then the allele frequency in the composite disease sample is

$$p_{caseC} = (1-s)p_{case} + sp_{control}.$$

The non-centrality parameter (NCP) of the χ^2 test of association is

$$NCP = \frac{N^2(p_{caseC} - p_{control})^2}{Var(\hat{p}_{caseC} - \hat{p}_{control})} = \frac{Nv(1-v)(p_{caseC} - p_{control})^2}{\bar{p}(1-\bar{p})},$$

where $\bar{p} = vp_{caseC} + (1-v)p_{control}$ and where $v = N_{case} / (N_{case} + N_{control}) = N_{case} / N$. We calculate power as the normal probability $p(Z > T)$, where $Z = \sqrt{NCP}$ and T is the normal deviate corresponding to the type I probability level, i.e., 5×10^{-8} for genome-wide association. When $s=0$, the power calculation agrees with the genetic power calculator [46].

References

1. Falconer D. The inheritance of liability to certain diseases, estimates from the incidence among relatives. *Ann Hum Genet.* 1965;29: 51–76.
2. Reich T, Morris CA, James JW. Use of multiple thresholds in determining mode of transmission of semi-continuous traits. *Ann Hum Genet.* 1972;36:163–84.
3. Dempster ER, Lerner IM. Heritability of threshold characters. *Genetics.* 1950;35:212–36.
4. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet.* 2013;14:139–49.

5. Risch NJ. Searching for genetic determinants in the new millennium. *Nature*. 2000;405:847–56.
6. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet*. 2012;90:7–24.
7. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*. 2011;12:745–55.
8. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*. 2010;11:415–25.
9. Vissers LELM, de Ligt J, Gilissen C, Janssen I, Stehouwer M, de Vries P, et al. A de novo paradigm for mental retardation. *Nat Genet*. 2010;42:1109–12.
10. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*. 2012;485:237–41.
11. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature*. 2014;506:179–84.
12. Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. *Nat Genet*. 2012;44:623–30.
13. Vádlamudi L, Dibbens LM, Lawrence KM, Iona X, McMahon JM, Murrell W, et al. Timing of de novo mutagenesis - a twin study of sodium-channel mutations. *N Engl J Med*. 2010;363:1335–40.
14. Gratten J, Visscher PM, Mowry BJ, Wray NR. Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat Genet*. 2013;45:234–8.
15. Yang J, Visscher PM, Wray NR. Sporadic cases are the norm for complex disease. *Eur J Hum Genet*. 2010;18:1039–43.
16. Smith C. Heritability of liability and concordance in monozygous twins. *Ann Hum Genet*. 1970;34:85–91.
17. Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. *Nat Genet*. 2014. doi:10.1038/nrg3786.
18. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet*. 2008;9:255–66.
19. Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, Liewald D, et al. Genetic contributions to stability and change in intelligence from childhood to old age. *Nature*. 2012.
20. Wray NR, Lee SH, Kendler KS. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *Eur J Hum Genet*. 2012;20:668–74.
21. 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;42:565–9.
22. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88:76–82.
23. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*. 2011;43:519–25.
24. So HC, Li M, Sham PC. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet Epidemiol*. 2011;35:447–56.
25. Powell JE, Visscher PM, Goddard ME. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet*. 2010;11:800–5.
26. 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.
27. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–53.
28. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. VIEWPOINT Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010;11:446–50.
29. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109:1193–8.
30. Stringer S, Derks EM, Kahn RS, Hill WG, Wray NR. Assumptions and properties of limiting pathway models for analysis of epistasis in complex traits. *PLoS One*. 2013;8:e68913.
31. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*. 2008;4:e1000008.
32. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat Rev Genet*. 2014;15:22–33.
33. Dunn EC, Uddin M, Subramanian SV, Smoller JW, Galea S, Koenen KC. Research review: gene-environment interaction research in youth depression - a systematic review with recommendations for future research. *J Child Psychol Psychiatry Allied Discip*. 2011;52:1223–38.
34. Duncan LE, Keller MC. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am J Psychiatry*. 2011;168:1041–9.
35. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet*. 2013;45:1150–9.
36. Gusev A, Bhatia G, Zaitlen N, Vilhjalmsdottir BJ, Diogo D, Stahl EA, et al. Quantifying missing heritability at known GWAS loci. *PLoS Genet*. 2013;9:e1003993.
37. Hannan AJ. TRPing up the genome: tandem repeat polymorphisms as dynamic sources of genetic variability in health and disease. *Discov Med*. 2010;10:314–21.
38. Hannan AJ. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet*. 2010;26:59–65.
39. Wray NR, Gottesman II. Using summary data from the danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. *Front Genet*. 2012;3:118.
40. Han B, Diogo D, Eyre S, Kallberg H, Zernakova A, Bowes J, et al. Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. *Am J Hum Genet*. 2014;94:522–32.
41. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671–9.
42. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*. 1987;235:177–82.
43. Ferraldeschi R, Newman WG. Pharmacogenetics and pharmacogenomics: a clinical reality. *Ann Clin Biochem*. 2011;48:410–7.
44. Chen G-B, Lee SH, Montgomery GW, Wray NR, Radford-Smith GL, Visscher PM. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immuno-chip data. Submitted.
45. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491:119–24.
46. Purcell S, Cherny SS, Sham PC. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. 2003;19:149–50.