The contribution of genetic variants to disease depends on the ruler

John S. Witte[1,2,3], Peter M. Visscher[4,5] and Naomi R. Wray[4]

Abstract | Our understanding of the genetic basis of disease has evolved from descriptions of overall heritability or familiality to the identification of large numbers of risk loci. One can quantify the impact of such loci on disease using a plethora of measures, which can guide future research decisions. However, different measures can attribute varying degrees of importance to a variant. In this Analysis, we consider and contrast the most commonly used measures — specifically, the heritability of disease liability, approximate heritability, sibling recurrence risk, overall genetic variance using a logarithmic relative risk scale, the area under the receiver–operating curve for risk prediction and the population attributable fraction — and give guidelines for their use that should be explicitly considered when assessing the contribution of genetic variants to disease.

A rapidly growing number of genetic loci have been detected for disease and other traits. These include high-risk Mendelian loci from next-generation sequencing studies and many highly replicated low-penetrance variants from genome-wide association studies (GWASs)[1,2]. Two important questions that follow are: to what degree do such loci and variants affect the overall burden of disease, and how many variants remain to be discovered[3]? They can be assessed using various measures, and many of these have been developed with different goals and within traditionally disparate fields — such as quantitative genetics and epidemiology, the boundaries of which are now blurring in the post-genomics era (FIG. 1). The quantitative genetics approach calculates measures such as heritability of disease liability or sibling recurrence risk that can be explained by genetic variants. A more epidemiological or translational approach might assess their impact on the overall genetic variance (using a logarithmic relative risk (logRR) scale), the area under the receiver–operating curve (AUC) for risk prediction or the population attributable fraction (PAF)[4–6].

Each of these measures can be calculated as a proportion to quantify how much of the underlying genetic basis of disease is explained by known risk loci. The heritability explained is most commonly calculated as the proportion of variance in disease explained by risk loci relative to the overall heritability[5,7]. The proportion of the sibling recurrence risk or the logRR genetic variance explained by the loci provides a similar measure of their impact on disease. The AUC indicates how well known risk loci classify diseased individuals; dividing this measure by the maximum attainable AUC for a genetic risk predictor calculated from the heritability quantifies the proportion of maximum AUC explained[4]. Finally, the PAF approximates the proportion by which disease incidence or death would be reduced in a population in the absence of the identified genetic risk factors.

Although all of these measures are valid and have the same bounds (which range from 0% to 100%), for a given data set they may give different messages about the impact of risk variants on disease. This has resulted in contrasting and confusing use of these measures in the literature. For example, the same association results for the Crohn's disease variants in the *NOD2* (nucleotide-binding oligomerization domain containing 2) gene are reported to explain 1–2% of heritability[8], 5.1% of genetic risk[9] and 18.2% of the PAF[9]. In other words, the apparent proportion of disease 'explained' by risk variants can vary widely across measures, and the particular measure used can therefore result in very different interpretations among geneticists and epidemiologists.

In this Analysis, we compare six measures that are used to assess how much of the genetic basis of disease is explained by risk variants to understand their similarities and differences. We estimate the heritability of liability, approximate heritability, sibling recurrence risk, logRR genetic variance, AUC and PAF that are explained across a range of risk allele frequencies (RAFs) and relative risks (RRs) through empirical calculations and application to data from studies of breast cancer, Crohn's disease,

[1]*Department of Epidemiology and Biostatistics, and Department of Urology, University of California, San Francisco.*
[2]*Institute for Human Genetics, University of California, San Francisco.*
[3]*Helen Diller Comprehensive Cancer Center, University of California, San Francisco, 1450 3rd Street, San Francisco, California 94158, USA.*
[4]*Queensland Brain Institute, The University of Queensland, Building 79, Research Road, Brisbane, 4072, Queensland, Australia.*
[5]*The University of Queensland Diamantina Institute, The University of Queensland, 37 Kent Street, Brisbane, 4102, Queensland, Australia.*
Correspondence to J.S.W. and N.R.W.
e-mails: jwitte@ucsf.edu; naomi.wray@uq.edu.au
doi:10.1038/nrg3786
Published online 16 September 2014

Mendelian loci
Genetic loci that have alleles with discrete effects on the phenotype and that follow Mendel's laws of segregation and independent assortment.

Heritability
The proportion of phenotypic variation in a population that is attributable to genetic variation among individuals.

Disease liability
An underlying or latent continuous variable such that those with a liability above a threshold are considered diseased. The quantitative trait of liability reflects both genetic and environmental factors.

Sibling recurrence risk
The ratio of the probability that a sibling of an individual affected by a disease will also be affected compared to the risk of disease in the general population.

rheumatoid arthritis and schizophrenia. We describe the relationships among these measures and give guidance for their appropriate calculation and interpretation when assessing the overall impact of genetic contributions to disease. Finally, we provide an online tool to calculate these measures from association study summary statistics (that is, RAFs and RRs).

Measures of genetic impact for individual risk loci

Scale matters. A key difference between the measures considered here is the scale on which they are measured (BOX 1; TABLE 1). Assessing the contribution of individual loci to disease risk on the observed (binary) scale is not very informative, as the relationship between increasing burden of risk loci and probability of disease is highly nonlinear[10,11]. Therefore, transformations are made to more informative scales, such as the liability of risk scale or the logarithmic risk scale. Quantitative geneticists commonly use the liability scale to evaluate the genetic basis underlying disease variability in a population[12]. By contrast, epidemiologists more often use logRR models for estimation of genetic effects on disease. These different perspectives, which form the basis of model choices, and calculated measures can ultimately affect inferences and conclusions; that is, the measure of an apparent contribution made by a given locus can depend on the ruler (see below).

Proportion of heritability explained. Using the methods and notation in BOX 1 and TABLE 1, we can estimate the proportion of phenotypic variance on the liability scale explained by risk variant i as $h^2_{L[i]} = V_{AL[i]}/V_{PL} = V_{AL[i]}/(V_{GL[i]}+1)$ (REFS 13,14), where $h^2_{L[i]}$ is the heritability explained, and $V_{AL[i]}$, $V_{PL[i]}$ and $V_{GL[i]}$ are the additive, phenotypic and genetic variance, respectively. On this scale we only consider the additive contribution from the locus ($V_{AL[i]}$), which allows comparison with existing estimates of heritability of liability derived from family data (h^2_L)[13,15,16]. Furthermore, under the assumption of a small RR for a risk variant B (that is, $RR_{Bb} \approx 1$) and a multiplicative model on the observed scale (that is, $RR^2_{Bb} = RR_{BB}$), an approximate heritability is given by $h^2_{L\text{-}approx[i]} = 2p(1-p)(RR_{Bb}-1)^2/v^2$, where p is the frequency of risk allele B[17–19]. In this equation, v is the mean liability of diseased individuals and is approximated as z/K, where z is the height of the standard normal distribution at the threshold T

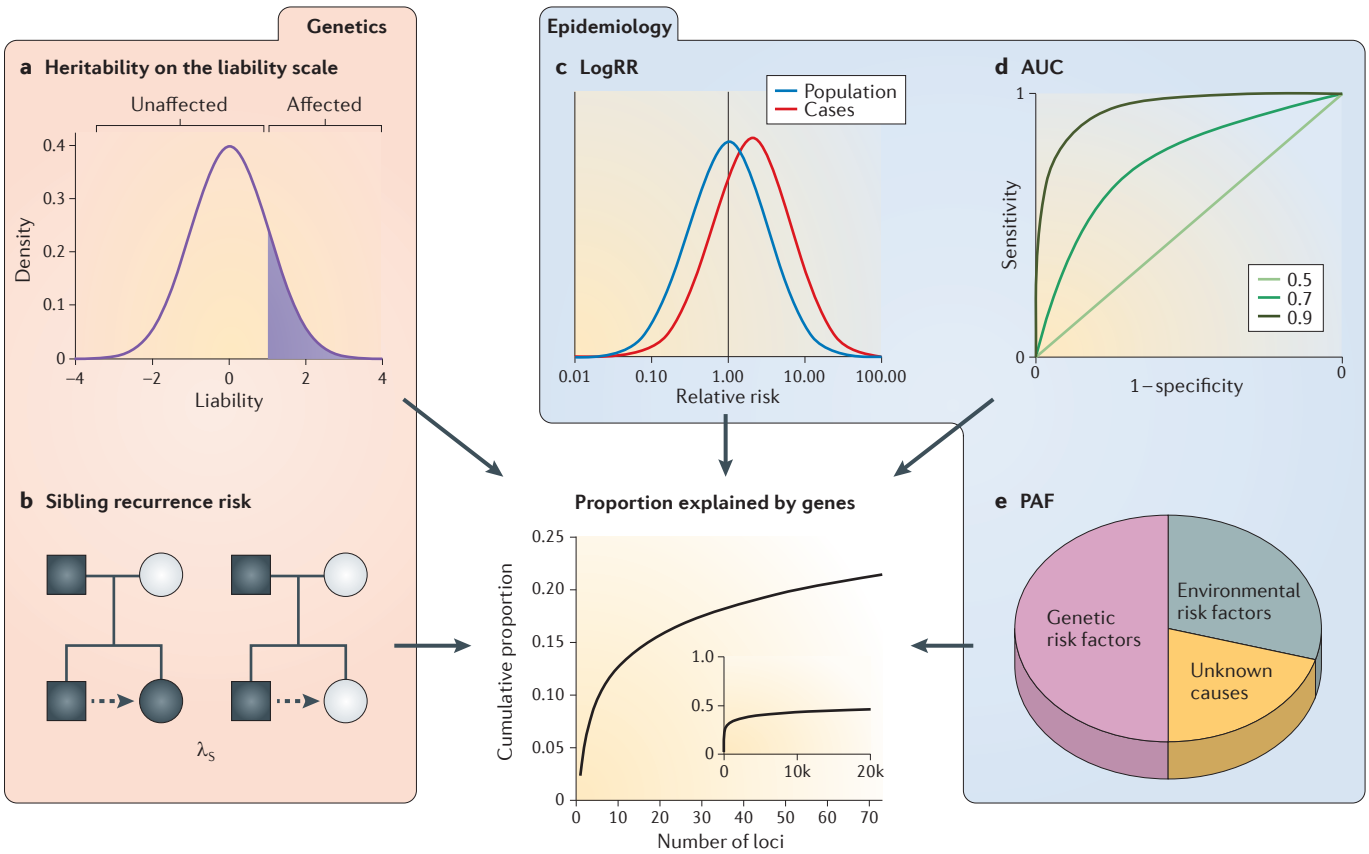


Figure 1 | **Different measures of genetic effects on disease.** Various measures can be used to assess the extent to which known genetic factors contribute to the overall genetic variation in disease. These include heritability (part **a**), sibling recurrence risk (part **b**), logarithmic relative risk (logRR) genetic variance (part **c**), area under the receiver–operating curve (AUC; part **d**) and population attributable fraction (PAF; part **e**). These measures have their bases in traditionally distinct disciplines such as quantitative genetics and epidemiology, which have recently begun to coalesce. Although epidemiological measures were originally developed to address different questions, they are now being repurposed to assess how much genetic variation can be explained. We compare these measures by simulation and applications.

Box 1 | **A matter of scale**

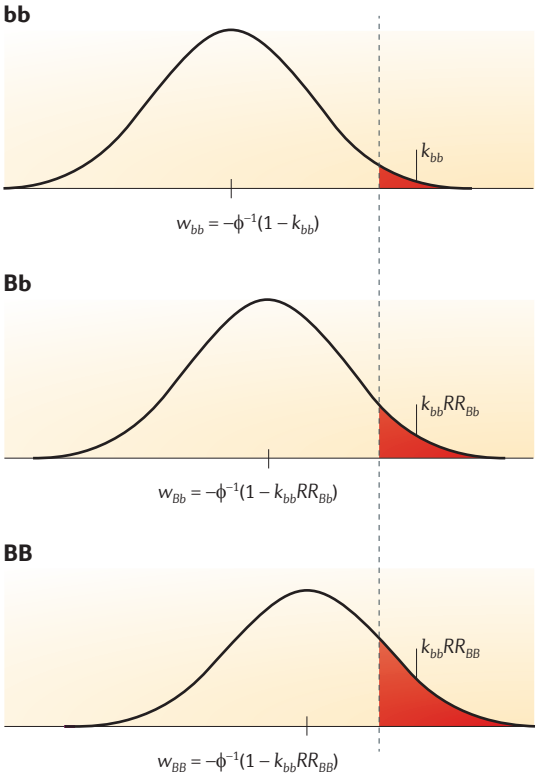
The contribution of genetic loci to disease can hinge on the scale used to assess risk (for example, observed, logarithmic or liability scales). On the observed scale, the risk of disease (D) for individuals carrying zero, one or two copies of risk variant B are $\Pr(D|\mathrm{bb}) = k_{bb}$, $\Pr(D|\mathrm{Bb}) = k_{bb}RR_{Bb}$ and $\Pr(D|\mathrm{BB}) = k_{bb}RR_{BB}$, respectively, where k_{bb} is the overall risk among non-carriers and RR_G is the relative risk for carrying genotype G (that is, Bb or BB) in comparison to the bb genotype. The probability of disease given genotype from a multiplicative model on the observed risk scale can then be represented by the following equation, where x_G is a (0,1) indicator of the genotypes carried by an individual.

$$\Pr(D|G) = k_{bb}RR_{Bb}^{x_{Bb}}\ RR_{BB}^{x_{BB}} \quad (1)$$

The overall risk of disease (K) is represented by the following equation, where p is the frequency of the risk variant B.

$$K = E[\Pr(D)] = \sum_G \Pr(D|G)\Pr(G)$$
$$= k_{bb}((1-p)^2 + 2p(1-p)RR_{Bb} + p^2RR_{BB}) \quad (2)$$

When RR_{Bb}, RR_{BB} and K are known, this can be rearranged to estimate k_{bb}. The overall relative risk due to multiple independent variants can be modelled by extension, in which k_{bb} is replaced by the probability of disease in individuals carrying no risk variants. This model is appealing because it is mathematically tractable; however, it is not constrained, and some combinations of parameters can therefore generate a probability of disease that is greater than one[11,18]. For this reason, it is not the model of choice when considering multiple risk loci. This model is multiplicative on the disease scale but additive on the logarithmic risk scale.



$$\log(\Pr(D|G)) = \log(k_{bb}) + \log(RR_{Bb})\,x_{Bb} + \log(RR_{BB})\,x_{BB} \quad (3)$$

Another possibility is to use the liability risk scale, which assumes that individuals have a latent continuous liability of risk for disease that reflects both genetic and non-genetic risk factors[12]. Disease occurs when the total phenotypic liability exceeds a threshold (that is, when a sufficient number of risk factors are present). For complex diseases, numerous risk factors each of modest effect are expected. The residual variation in liability between individuals of each genotype class at any given risk locus is assumed to have a standard normal distribution about different mean liabilities w_{bb}, w_{Bb} and w_{BB} for the genotype classes bb, Bb and BB, respectively (see the figure). The observed disease risks for each genotype class are converted into thresholds on the liability scale. The difference between the genotype thresholds equals the differences between the mean distributions with a common threshold for disease. The liability risk model is mathematically tractable and easily generalizes to multiple risk loci; it is also constrained so that the probability of disease does not exceed one. Moreover, the contribution of individual risk loci can be parameterized in terms of the variance they explain, which provides a general framework because many different combinations of allele frequency and effect size can generate the same contribution to variance. For these reasons, the liability risk model is usually the model of choice when considering multiple risk loci[18,39–43].

Genetic variance
The variance of trait values that can be ascribed to genetic differences among individuals. The total genetic variance of a trait can be dissected into additive, dominance and other components.

Area under the receiver–operating curve
(AUC). The receiver–operating curve for a predictor (for example, a genetic test) plots the proportion of cases correctly identified by the test against the proportion of controls that are incorrectly classified as cases. The AUC indicates the probability that a factor (for example, a genetic risk score) will predict a higher risk of disease in a randomly selected case than in a control.

Population attributable fraction
(PAF; also known as population attributable risk). For a given disease, risk factor and population, the fraction by which the incidence rate of the disease in the population would be reduced if the risk factor was eliminated.

Overall disease risk
The lifetime probability that an individual will be affected by a disease.

that truncates the proportion K. T is given by $\Phi^{-1}(1-K)$, and K is the overall disease risk (BOX 1). Therefore, $h^2_{L[i]}/h^2_L$ (or $h^2_{L\text{-}approx[i]}/h^2_L$) estimates the proportion of total heritability explained by the i^{th} risk variant.

Sibling recurrence risk explained. The impact of a risk variant can also be quantified relative to the overall sibling recurrence risk (λ_S)[9]. Siblings share $V_{AO}/2 + V_{DO}/4$ of risk[20], where V_{AO} and V_{DO} are the additive and dominance genetic variance on the observed risk scale, respectively. Thus, the increased risk attributable to the i^{th} risk variant can be represented by the following equation.

$$\lambda_{S[i]} = 1 + \frac{\frac{V_{AO[i]}}{2} + \frac{V_{DO[i]}}{4}}{K^2} \quad (4)$$

From TABLE 1 we can estimate $V_{AO[i]} = k^2_{bb}2p(1-p)(p(RR_{BB} - RR_{Bb}) + (1-p)(RR_{Bb} - 1))^2$, and $V_{DO[i]} = k^2_{bb}p^2(1-p)^2(RR_{BB} + 1 - 2RR_{Bb})^2$. The $\lambda_{S[i]}/\lambda_S$ ratio indicates the impact of the i^{th} variant on the sibling recurrence risk, and λ_S is generally obtained from published estimates. However, $\lambda_{S[i]}/\lambda_S$ can give nonsensical values under the null hypothesis. When $\lambda_{S[i]} = 1$ the ratio incorrectly suggests that the i^{th} variant contributes to the genetic risk,

and when λ_S also equals 1 the ratio equals 1. Instead, the ratio of logarithms ($\log(\lambda_{S[i]})/\log(\lambda_S)$) has been proposed[9]. In this case, when $\lambda_{S[i]} = 1$ the ratio of logarithms appropriately indicates no contribution of the i^{th} genetic variant to risk, and the ratio of logarithms gives values that are more uniformly distributed across the range of zero to one. Of course, shifting scales results in a quantitatively different measure.

Genetic variance on a logarithmic relative risk scale. From a more epidemiological perspective, one can calculate the contribution of a risk variant to the overall genetic variation on the logRR scale. From TABLE 1, the genetic variance attributable to the i^{th} risk variant on the logRR scale is $V_{Glog[i]} = (1-p)^2M^2 + 2p(1-p)(\log(RR_{Bb}) - M)^2 + p^2(\log(RR_{BB}) - M)^2$, where M is the mean value of logRR, and $M = 2p(1-p)\log(RR_{Bb}) + p^2\log(RR_{BB})$. Assuming a multiplicative model this simplifies to $V_{Glog[i]} = 2p(1-p)(\log(RR_{BB}))^2$. For a polygenic disease with numerous risk alleles, the distribution of logRR in the population tends towards normal with variance V_{Glog}. Thus, the fraction of the genetic risk explained by the i^{th} risk variant is given by $V_{Glog[i]}/V_{Glog}$. In practice, V_{Glog} is assumed to approximately equal $2\log(\lambda_S)$ (REFS 17–22). Note that V_{Glog} should not be estimated as $\log(\lambda_{MZ})$, which is the recurrence risk to monozygotic twins, because $\lambda_{MZ} \approx \lambda_S^2$ is an asymptotic result that only holds for diseases of high prevalence (for example, $K > 0.1$) and low heritability[18], and can otherwise give nonsensical results.

Proportion of area under the curve. We can also determine how much of the maximum possible AUC that is attainable with a risk prediction model based on all genetic information is explained by the i^{th} risk variant. We can first estimate the AUC for the i^{th} variant using the heritability on the liability scale explained by this variant ($h^2_{L[i]}$) (REF. 4).

$$AUC_{L[i]} = \Phi\left(\frac{(x-v)h^2_{L[i]}}{\sqrt{h^2_{L[i]}(1-h^2_{L[i]}x(x-T) + 1 - h^2_{L[i]}v(v-T))}}\right) \quad (5)$$

In this equation, $x = -z/K$, T is the population threshold, and $v = -xK(1-K)$ (REF. 13) (see above and BOX 1). Next, we determine the maximum attainable AUC (AUC_{Max}) by substituting into the above equation the overall heritability h^2_L (for example, estimated from twin studies)[4]. Although the AUC upper bound is 1.0, the AUC attainable with genetic factors will generally be lower. We can then estimate the proportion of the maximum AUC explained by the risk variants as the proportion of AUC (pAUC), where

Table 1 | **Measures of a genetic variant's impact on disease are grounded in different scales of risk**

Measures	Genotype*		
	bb	Bb	BB
General notation			
Population frequency‡	$(1-p)^2$	$2p(1-p)$	p^2
Genotype risk§	w_{bb}	w_{Bb}	w_{BB}
Mean genotype risk (M)‖	$(1-p)^2w_{bb}$	$2p(1-p)w_{Bb}$	p^2w_{BB}
Variance of genotype risk (V)‖	$(1-p)^2(w_{bb}-M)^2$	$2p(1-p)(w_{Bb}-M)^2$	$p^2(w_{BB}-M)^2$
Scale-specific genotype risks			
Observed risk¶	k_{bb}	$k_{bb}RR_{Bb}$	$k_{bb}RR_{BB}$
Relative risk	1	RR_{Bb}	RR_{BB}
Logarithmic relative risk	0	$\log(RR_{Bb})$	$\log(RR_{BB})$
Liability threshold#	$-\Phi^{-1}(1-k_{bb})$	$-\Phi^{-1}(1-k_{bb}RR_{Bb})$	$-\Phi^{-1}(1-k_{bb}RR_{BB})$
Quantitative genetics notation			
Genotype risk	$-a$	$d = w_{Bb} - (w_{bb}+w_{BB})/2$	$a = w_{BB} - (w_{bb}+w_{BB})/2$
Deviations from the mean**			
Total	$-a-M = -2p(a+(1-p)d)$	$d-M = a((1-p)-p)+d(1-2p(1-p))$	$a-M = 2(1-p)(a-pd)$
Additive‡‡	$-2p\alpha$	$((1-p)-p)\alpha$	$2(1-p)\alpha$
Dominance	$-2p^2d$	$2p(1-p)d$	$2(1-p)^2d$

For each scale, the genotype risk values can be used to calculate the corresponding mean and variance values. *B denotes the known risk variant. ‡Under Hardy–Weinberg equilibrium. §In general notation, to estimate the scale-specific mean and variance, the genotype risks are substituted for w (for example, logarithmic relative risk or liability). ‖The mean (M) and variance (V) of genotype risk is the sum of the three genotype-specific components. ¶k_{bb} is the overall disease risk for individuals carrying the homozygous non-risk genotype (bb). RR_G is the relative risk of disease for carriers of the risk genotype G (that is, Bb or BB) compared with non-carriers (bb). #Φ is the standard normal cumulative distribution function. **The notation of Falconer and Mackay[14] is used, and the quantitative genetics notation values are assigned such that, in the absence of dominance ($d = 0$), the value of the heterozygote is zero and midway between the values of the two homozygotes. ‡‡$-\alpha = a + d((1-p)-p)$ is the average effect of substituting b with B. The total genetic deviation is the sum of the additive deviations and the dominance deviations with M expressed in the quantitative genetics notation $M = (1-p)^2(-a) + 2p(1-p)d + p^2a$.

pAUC = $[(AUC_{L[i]} - 0.5)/(AUC_{Max} - 0.5)]^2$. We square this measure because it is related to the square root of heritability, thus allowing comparisons with other measures that are visually more intuitive to interpret. This measure will generally range from 0 (when AUC = 0.5) to 1 (when AUC = 1).

Population attributable fraction. The PAF is commonly used to approximate the public health implications of modifying or removing an exposure. Although we cannot currently intervene to remove or nullify risk variants, genetic PAFs are often used to estimate the degree to which a disease can be attributed to the risk variants. We can calculate this from the ratio of the disease due to a risk variant (that is, subtracting off the baseline risk among non-carriers (k_{bb})) divided by the overall risk, as seen in the following equation.

$$\mathrm{PAF} = \frac{K - k_{bb}}{K} = 1 - \frac{k_{bb}}{K} \quad (6)$$

From BOX 1,

$$k_{bb} = \frac{K}{((1-p)^2 + 2p(1-p)RR_{Bb} + p^2RR_{BB})} \quad (7)$$

so

$$\mathrm{PAF} = 1 - \frac{1}{(1-p)^2 + 2p(1-p)RR_{Bb} + p^2RR_{BB}} = \frac{2p(1-p)(RR_{Bb} - 1) + p^2(RR_{BB} - 1)}{1 + 2p(1-p)(RR_{Bb} - 1) + p^2(RR_{BB} - 1)} \quad (8)$$

These equations highlight that the PAF is the effect of 'removing' the genetic risk variant on the overall risk of disease. Note that previous work had a typographical error in the equation for the PAF[21].

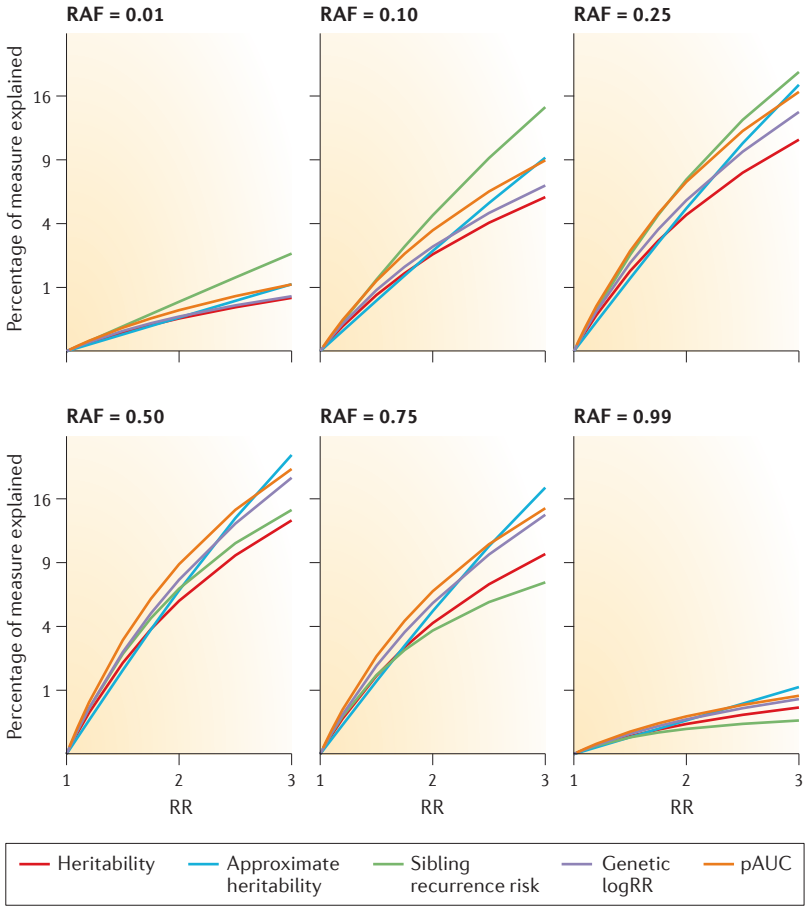


Figure 2 | **Empirical evaluation of measures of genetic effects.** Comparison of heritability, approximate heritability, sibling recurrence risk, logarithmic relative risk (logRR) genetic variance and proportion of area under the receiver–operating curve (pAUC) explained across a range of complex disease architectures is shown. The measures are calculated for single causal variants with risk allele frequencies (RAFs) of 0.01, 0.10, 0.25, 0.50, 0.75 and 0.99, and genetic relative risk (RR) of 1.0–3.0 (assuming a multiplicative model). The overall disease risk is assumed to be 0.01, and the total sibling recurrence risk is 5, which gives an overall genetic heritability on the liability scale of 0.55 and a maximum AUC of 0.95. The percentages of heritability, sibling recurrence risk and logRR genetic variance explained are fairly modest for low RRs and small RAFs, but as these increase the measures start to materially differ. Heritability is always one of the smallest measures and is overestimated by the approximate heritability as the RR increases. The sibling recurrence risk and pAUC are generally the largest measures for lower RAFs.

Comparison of measures for single variants

We first evaluated how the above measures assess the impact of a single genetic variant on disease. Specifically, we calculated the measures across a range of RAFs and genetic RRs for carrying one additional risk allele. We assume an overall disease risk in the population of 0.01 and a sibling recurrence risk of 5 — which are consistent with an overall genetic heritability on the liability scale of 55% — and a multiplicative model of genotype RRs. Note that we present calculations for PAF separately because it generally gives estimates that are an order of magnitude larger than the other measures. The proportion of genetic risk explained by all of these measures is similar and fairly limited for variants that are less common and/or have modest effects on disease (FIG. 2). However, these measures diverge as the RAF increases up to a certain point and as the RRs increase. The conventional heritability estimate always suggests one of the smallest impacts of the genetic variants on disease, irrespective of RAF and RR (FIG. 2, red line). Similar values are given by the approximate heritability when RR < 1.5, but this increasingly overestimates the heritability as the RR increases, as expected from its derivation (FIG. 2, blue line). The sibling recurrence risk explained suggests the largest contribution of the genetic variants to disease when RAF ≤ 0.25 but a smaller amount for larger RAF values (FIG. 2, green line). An opposite trend is seen for the logRR genetic variance explained (FIG. 2, purple line), which is lower than the sibling recurrence risk when RAF < 0.25 and then larger for more common risk variants. Finally, the pAUC consistently indicates one of the highest estimates of genetic basis of disease explained (FIG. 2, orange line). Although these differences may seem slight, they are only for individual variants. When aggregated across numerous risk variants, substantially larger differences in the measures become apparent, as shown in the following applications.

Contribution of multiple risk loci to disease

To determine the contribution of multiple risk loci to disease from summary statistics, the measures for individual loci can be aggregated if they are independent. Specifically, for heritability on the liability scale, approximate heritability, sibling recurrence risk and logRR genetic variance, an aggregate score is calculated from the sum of the contributions calculated for each locus. Similarly, the aggregate heritability of liability is used to calculate the AUC. To calculate the PAF due to multiple risk variants, one cannot simply add together the PAFs of single variants because this ignores the fact that most individuals will carry multiple risk alleles. In fact, summing PAFs of single variants can quickly give an overall PAF that is greater than 100%. Instead, we can calculate a joint PAF across multiple variants, which restricts the total PAF due to all risk variants to be ≤100%. Specifically, if we assume that the risk variants are independent of each other and that their combined effects on disease are multiplicative, then a joint estimate of PAF is given by $PAF_{Total} = 1 - \Pi_i(1 - PAF_i)$.

Application to complex diseases

To further explore how these measures can imply different impacts of genetic variants on disease, we calculate them across studies of breast cancer, Crohn's disease, rheumatoid arthritis and schizophrenia. We selected these diseases because they have so far been well studied and have a range of underlying genetic architectures. For each disease, we selected the loci that have previously been reported at the time of this analysis as independently associated with disease and identified the reported risk allele, as well as its frequency and RR (generally estimated by odds ratios). More specifically, for breast cancer the loci were obtained from the catalog of published GWASs, and for the other three diseases we used the single-nucleotide polymorphisms (SNPs) that were reported and selected as independent by the corresponding publications (cited below). Although the criteria for SNP selection vary depending on the publications and ongoing work continues to discover novel loci for these traits, the SNPs considered here provide a sufficient view of the differences in the measures, and the inclusion of additional SNPs should not materially affect our findings.

Breast cancer. GWASs have detected a large number of common, low-risk variants for breast cancer (see the catalog of published GWASs). Here, we evaluate 65 SNPs from the catalogue that seem to be independently associated with breast cancer using a linkage disequilibrium filter of $r^2 < 0.2$ among Europeans within 100 kb of the most associated SNP. On the basis of the literature, we assume that the overall disease risk is 12% and the sibling recurrence risk is 2.0 (REF. 22); these are consistent with the heritability of liability being equal to 60%. Benchmarked against these values, almost all of the risk variants individually explain <0.5% of the total variation in heritability, sibling recurrence risk, logRR genetic variance and pAUC (FIG. 3a; TABLE 2; see Supplementary information S1 (table)). As expected, the variants with larger effects on breast cancer ($1.3 < RR \leq 2$) explain a larger proportion of these measures (FIG. 3a, blue lines). For breast cancer, the approximate heritability and heritability explained are lower than the other measures, and the sibling recurrence risk is the largest, which is in agreement with our empirical calculations. All breast cancer variants combined are estimated to explain 13% of the approximate heritability, 18% of the heritability, 19% as measured by the pAUC, 21% of the logRR genetic variance and 22% of the sibling recurrence risk (FIG. 3a; TABLE 2). The similarity among the heritability, pAUC, logRR and sibling recurrence risk reflects the uniformly low penetrance and high frequencies across the risk variants. Moreover, the relatively high proportion of these measures explained reflects the high overall risk but modest sibling recurrence risk for breast cancer in the population.

Crohn's disease. At least 140 modest-risk variants and 3 additional high-risk variants have been reported as independently associated with Crohn's disease[23]. We assume that the overall risk of this disease is 0.5%, the sibling recurrence risk is 10.3, and the heritability of liability is 72%[24]. For the common, low-risk variants, the patterns observed are similar to those of breast cancer: heritability is smaller than the logRR genetic variance, which is smaller than the sibling recurrence risk (FIG. 3b; TABLE 2; see Supplementary information S2 (table)). However, for the high-risk variants ($2 < RR \leq 15$), there is more variation in these measures, which reflects different combinations of RRs and RAFs (FIG. 3b, red lines). Specifically, the common allele of rs11209026 — the wild-type allele corresponding to the uncommon interleukin 23 receptor (*IL23R*) coding variant that is protective for Crohn's disease — has a fairly large effect (RR = 2.4) but is extremely common (RAF = 0.93), and this combination explains the most individual heritability (1.4%) but lower sibling recurrence risk (0.96%) (TABLE 2). By contrast, rs5743293 has an even larger effect (RR = 3.1) but is less common (RAF = 0.02); therefore, it explains slightly less heritability (1.1%) but substantially higher sibling recurrence risk (4.0%) (TABLE 2). Taken together, the 143 risk variants of Crohn's disease account for ~16.4% of the heritability but explain a larger proportion of the sibling recurrence risk (25%) and an even larger proportion as measured by the pAUC (34%) (FIG. 3b; TABLE 2). The higher pAUC estimates across all of the risk variants partly reflect the low overall risk of the disease (0.5%).

Rheumatoid arthritis. Here, we evaluate 36 risk variants previously reported as being independently associated with rheumatoid arthritis and assume an overall disease risk of 1% and a sibling recurrence risk of 6.0, which are collectively consistent with a heritability of liability of 63%[16]. Even with so few risk variants we observe a similar proportion of disease explained as for Crohn's disease (FIG. 3c; see Supplementary information S3 (table)). This is due to the substantial impact of a single variant — rs6910071 at the *HLA-DRB1E* locus — on all of the measures (FIG. 3c, red line). This variant

Genetic architectures
The number of risk alleles underlying disease, their allele frequency spectrum, effect sizes and mode of interaction.

Linkage disequilibrium
A measure of whether alleles at two loci coexist in a population in a nonrandom manner. Alleles that are in linkage disequilibrium are found together on the same haplotype more often than expected by chance.

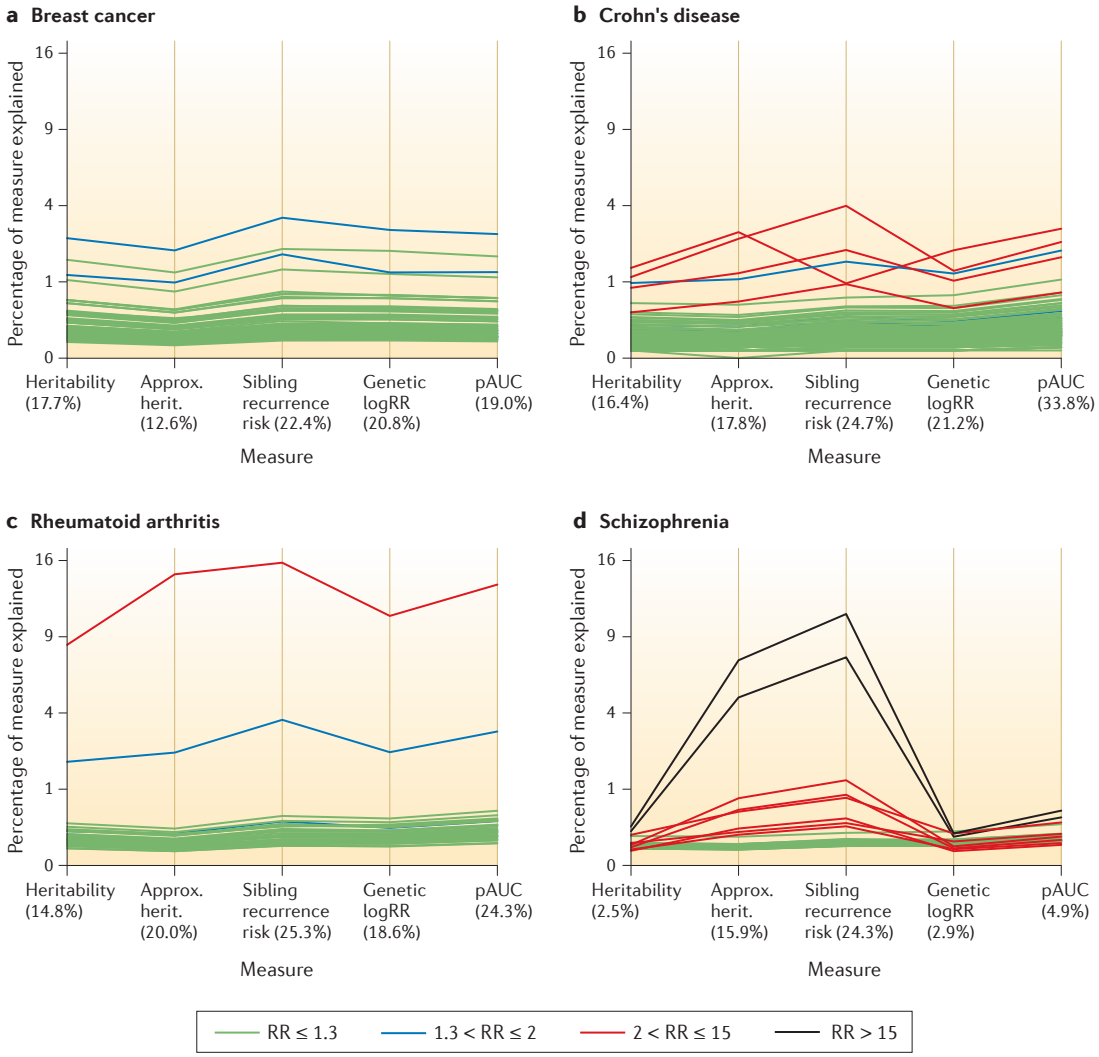


Figure 3 | **Application of measures to four diseases.** Commonly used measures for assessing the impact of known risk variants on disease are compared for four diseases: breast cancer (65 variants; part **a**), Crohn's disease (143 variants; part **b**), rheumatoid arthritis (36 variants; part **c**) and schizophrenia (32 variants; part **d**). The measures considered are heritability explained, approximate heritability explained (Approx. herit.), sibling recurrence risk explained, logarithmic relative risk (logRR) genetic variance explained, and the proportion of area under the receiver–operating curve (pAUC). Each line corresponds to an individual risk variant and indicates the percentage of each measure (for example, total heritability) explained by the variant. Lines are different colours depending on the relative risk (which is estimated by the odds ratio) for the variants. The *y* axes are on a squared scale. The percentages given in parentheses after each measure on the *x* axes indicate the total across all risk variants.

has a large effect on rheumatoid arthritis (RR = 2.88) and is common (RAF = 0.22); therefore, it accounts for an estimated 8% of the heritability, 16% of the sibling recurrence risk, 11% of the logRR genetic variance and 14% as measured by the pAUC (TABLE 2). The twofold range between heritability and sibling recurrence risk of this variant leads to a substantial difference in the overall measures of genetic variation explained: 15% of heritability but 25% of sibling recurrence risk. Thus, single common variants of large effect can result in different estimates across these measures. We note that the latest GWAS for rheumatoid arthritis reports 101 associated loci[25].

Schizophrenia. Here, we consider 24 GWAS risk variants previously reported for schizophrenia[26,27], as well as 8 rare copy-number variants (CNVs) that substantially increase risk of this disease[28–30] (FIG. 3d; see Supplementary information S4 (table)). We benchmark using an overall disease risk of 1% and a sibling recurrence risk of 8.8, which are collectively consistent with the heritability of liability of 81%[31]. As above, the common, low-risk variants explain a small percentage of the measures evaluated here (FIG. 3d, green lines). By contrast, the CNVs give extremely different results across these measures (FIG. 3d, red and black lines). This is especially apparent for the CNVs at 16p11.2 and 22q11,

Table 2 | **Measures of overall impact of risk variants on different diseases with a range of underlying genetic architectures***

			Heritability			Sibling recurrence	LogRR	AUC		
Risk variant [*i*]	**RAF‡**	**RR‡**	$h^2_{L[i]}$ §	$h^2_{L[i]}/h^2_L$ ‖	$h^2_{L[i]approx}/h^2_L$ ‖	$\lambda_{S[i]}$	$\log(\lambda_{S[i]})/\log(\lambda_S)$ ¶	$V_{Glog(RR[i])}/2\log(\lambda_S)$ ¶	$AUC_{[i]}$	$pAUC_{[i]}$ #	**PAF**
Breast cancer											
rs2943559	0.07	1.13	0.07%	0.12%	0.08%	1.001	0.16%	0.14%	0.51	0.13%	1.80%
rs10771399	0.90	1.20	0.22%	0.36%	0.27%	1.003	0.39%	0.45%	0.52	0.39%	28.1%
rs2180341	0.21	1.41	1.49%	2.47%	2.00%	1.02	3.39%	2.83%	0.57	2.65%	15.2%
All variants (n = 65)	–	–	10.7%	17.7%	12.6%	1.17	22.4%	20.8%	0.65	19.0%	95.2%
Crohn's disease											
rs12103	0.18	1.09	0.03%	0.04%	0.03%	1.001	0.05%	0.05%	0.51	0.07%	3.1%
rs11209026	0.93	2.37	1.02%	1.40%	2.73%	1.023	0.96%	2.00%	0.58	2.88%	80.8%
rs5743293	0.02	3.10	0.82%	1.13%	2.45%	1.10	3.99%	1.31%	0.57	2.32%	9.5%
All variants (n = 143)	–	–	11.9%	16.4%	17.8%	1.78	24.7%	21.2%	0.77	33.8%	100%
Rheumatoid arthritis											
rs5029937	0.04	1.40	0.13%	0.20%	0.17%	1.01	0.33%	0.24%	0.53	0.34%	3.1%
rs2476601	0.10	1.94	1.17%	1.85%	2.19%	1.07	3.65%	2.21%	0.58	3.09%	16.4%
rs6910071**	0.22	2.88	5.30%	8.38%	14.6%	1.33	15.8%	10.7%	0.67	13.6%	50.0%
All variants (n = 36)	–	–	9.34%	14.8%	20.1%	1.57	25.3%	18.6%	0.72	24.3%	99.3%
Schizophrenia											
rs171748	0.47	1.08	0.04%	0.05%	0.04%	1.001	0.06%	0.06%	0.52	0.10%	7.0%
rs17504622	0.05	1.24	0.06%	0.08%	0.08%	1.003	0.12%	0.10%	0.52	0.15%	2.3%
CNV (duplication) at 16p11.2	0.0003	26.0	0.16%	0.20%	4.85%	1.18	7.45%	0.14%	0.53	0.40%	1.4%
All variants (n = 32)	–	–	2.02%	2.50%	15.9%	1.69	24.3%	2.87%	0.61	4.93%	90.9%

AUC, area under the receiver–operating curve; logRR, logarithmic relative risk; PAF, population attributable fraction; pAUC, proportion of AUC; RAF, risk allele frequency; RR, genetic relative risk for disease due to carrying a copy of risk variant versus none. *Two sets of results are presented for each disease: selected individual variants and all significant variants combined. Results for all individual variants are given in Supplementary information S1–S4 (tables). Overall population risks of disease assumed from the literature are 12% for breast cancer, 0.5% for Crohn's disease, 1% for rheumatoid arthritis and 1% for schizophrenia. λ_S values are assumed to be 2.0 for breast cancer, 10.3 for Crohn's disease, 6.0 for rheumatoid arthritis and 8.8 for schizophrenia. On the basis of the risk of disease and λ_S values, h^2_L values are 60% for breast cancer, 72% for Crohn's disease, 63% for rheumatoid arthritis and 81% for schizophrenia. ‡RAF and RR are estimated by odds ratios under the assumption of the multiplicative (that is, logarithmic additive) model, so the RR for carrying two risk variants is RR^2. §$h^2_{L[i]}$ is the proportion of variance in disease explained by risk variant *i* on the liability scale. ‖$h^2_{L[i]}/h^2_L$ and $h^2_{L[i]approx}/h^2_L$ represent the proportion of heritability explained by the risk variants and the proportion explained by the approximate estimate, respectively. ¶$\log(\lambda_{S[i]})/\log(\lambda_S)$ and $V_{Glog(RR[i])}/2\log(\lambda_S)$ are the proportion of sibling recurrence risk explained by risk variants and the proportion of logRR genetic variance explained, respectively. #$pAUC_{[i]}$ is the proportion of AUC explained by risk variants compared to the maximum AUC expected from a genetic predictor. The maximum AUC values are estimated from the overall heritability to be 0.9 for breast cancer, 0.98 for Crohn's disease, 0.97 for rheumatoid arthritis and 0.99 for schizophrenia. **HLA-DRB1E* locus.

both of which are rare (RAF = 0.0003) and have very large effects on schizophrenia (RR > 25). Owing to their rarity these CNVs explain a modest proportion of heritability, genetic variance and pAUC (<0.5%); but their large impact on disease results in much higher proportions of approximate heritability (>5%) and sibling recurrence risk (>7.5%) (FIG. 3d; TABLE 2). Thus, when looking at all 32 schizophrenia variants (24 GWAS SNPs and 8 CNVs), estimates of the heritability, sibling recurrence risk, logRR genetic variance and pAUC explained give very different messages about the impact of the variants on this disease. Although the variants explain only 2.5–3% of heritability or logRR genetic variance and 5% of pAUC, they are estimated to account for up to 5 times as much of the approximate heritability and 10 times as much of the sibling recurrence risk (FIG. 2d; TABLE 2). The large increase for the approximate heritability was expected, as this measure departs from heritability for large RRs. However, it was surprising to see such a large departure between the sibling recurrence risk and logRR genetic variance explained. Although the sibling recurrence risk is generally always larger than the logRR genetic variance, the rarity and extremely large effects of the CNVs result in the drastically different results given by these two seemingly similar measures.

PAF: a problematic measure

The PAF can also be used to assess the impact of genetic factors on disease, but this measure has various limitations[32]. The PAF estimates the extent to which a disease

might be reduced if a risk factor was removed from a population. In our empirical comparisons, the PAF generally gives estimates that are an order of magnitude larger than the other measures even when the RAF is 0.01 and the RR is low. As the RAF increases beyond 0.50, the PAF is the one measure that continues to increase because it directly depends on the RAF. Even for a single variant, as the RAF and RR increase, the PAF can approach the upper bound of 100%. For example, in our breast cancer analysis, a variant (rs10771399) with a large RAF (0.90) but a modest impact on disease (RR = 1.20) has a very large PAF (28%) (TABLE 2). Similarly, if a rare genetic variant is protective for disease, then the other (extremely common) allele can give a very large PAF. For example, the protective *IL23R* coding variant (rs11209026) for Crohn's disease (which has a minor allele frequency of 0.07% and a RR of 0.42)[23] yields a PAF of an astonishing 81% (that is, for the risk allele, RAF = 0.93 and RR = 1/0.42 = 2.37) (TABLE 2). By contrast, our schizophrenia analysis shows how a rare variant (CNV at 16p11.2, which has a RAF of 0.0003) with an enormous effect size (RR = 26.0) can have a relatively small PAF (1.4%) (TABLE 2). Looking at all of the risk variants combined, the PAF for the four diseases are all >90%, and only half of the risk variants of Crohn's disease are able to give a PAF of 100% (TABLE 2; see Supplementary information S1–S4 (tables)).

The combined PAF also shows a computational anomaly: the apparent impact of each additional risk variant depends on the variants that have already been incorporated into this measure. For example, assume that there are 2 genetic variants for a disease, each of which has an individual PAF of 0.50 and a corresponding combined PAF of 0.75 ($= 1 - (1 - 0.5)^2$). An intervention that eliminates the effect of a risk variant at any one of these risk loci would decrease the incidence of disease in the population by half. An intervention at the second locus would further reduce the disease incidence by half in the remaining population or by a quarter in the original population. The order in which the exposure is removed will affect the magnitude of its apparent effect on the combined PAF. In other words, the apparent impact of a given risk variant on the combined PAF depends on what has already been discovered. Novel variants from less well-studied traits will seem to have larger effects than more well-studied traits, even if the risk variants have the same magnitude of association and RAF. Moreover, the combined PAF for multiple low-penetrance risk SNPs is not analogous to that obtained by removing a single high-risk environmental exposure from a population, such as reducing smoking to decrease rates of lung cancer. In this case, the difference depends not only on the number of risk factors but also on their penetrance and prevalence, as well as on their potential for modification or therapeutic intervention. As the number of known risk loci continues to increase — many of which are quite common — essentially everyone in the population will carry various risk alleles. At that point, any preventive treatment directed at countering the risk loci would have to be applied to almost the entire population.

Measures depend on the overall disease risk

Of the measures evaluated here, heritability depends on the overall disease risk (*K*). In practice, pAUC may be directly estimated, but here it is calculated from the heritability of liability, which is calculated from the reported RAF and RR, and hence also depends on *K*. For a given RR, both of these measures increase with increasing *K*, as the RR is expressed relative to the risk in the wild-type homozygote, which depends on *K*. The proportion of heritability and AUC explained is actually lower with increasing *K*, and these measures therefore depend on the value assumed for *K*. By contrast, the sibling recurrence risk, logRR genetic variance and PAF do not depend on *K*, which is an advantage of these measures because it is not always straightforward to define *K*. Nevertheless, the possible range in *K* — which can be determined from the literature — will generally be small for most diseases. For example, the ranges of *K* are 10–15% for breast cancer, 0.3–0.5% for Crohn's disease, 1.0–3.6% for rheumatoid arthritis and 0.5–1% for schizophrenia. Such ranges may have a limited impact on the proportion of heritability and AUC explained, which would thus be fairly robust to misspecification of *K*. We note that, although sibling recurrence risk, logRR genetic variance and PAF do not seem to depend on *K*, there is a built-in assumption that the value of *K* is the same in the family data used to calculate sibling recurrence risk as that in the population used to calculate the contribution to risk from an individual variant, as any RR is expressed relative to a baseline. Violation of this assumption may generate misleading results.

To complicate matters further, there is some confusion in the literature over the definition of disease risk, which partly reflects the merging of disciplines. Falconer defines *K* as the incidence of a binary trait[12] "or, in the context of human disease, the prevalence" (REF. 13). Both incidence (that is, the rate at which new cases occur in a time period) and prevalence (that is, the proportion of the population that is affected by a disease at any one time) have precise meaning in epidemiology. In fact, the relevant benchmark for calculation of heritability of liability is the lifetime morbid risk (LMR), which is the lifetime probability of being affected or the lifetime incidence. The most likely reason for this confusion is because, in the context of idealized populations pertaining to logical thinking in quantitative genetics theory, the parameters prevalence and LMR would be the same. In practice, they can be very different. For example, schizophrenia is a disorder with a fairly early age of onset and a long average mean life expectancy after diagnosis (albeit reduced compared with the general population); therefore, annual incidence, prevalence and LMR differ considerably at 2.5, 46 and 72 per 10,000, respectively[33]. As another example, consider motor neuron disease, for which the median age at onset is ~60 years and the life expectancy is only 2–5 years. In this case, estimates of incidence, prevalence and LMR are 0.3, 0.6 and 25 per 10,000, respectively[34]. For less common disorders, the assessment of LMR (or prevalence or incidence) and risk to relatives are associated with considerable sampling variance, and estimates of heritability of liability and

Genomic profile risk
A predicted measure of genetic risk for individuals constructed from a set of loci, the risk alleles and corresponding effect sizes of which have been estimated in an independent sample.

sibling recurrence risk can vary substantially between studies. Finally, in addition to the overall disease risk, study design and time-dependent effects could also affect the measures considered here.

Focus on the mean or the variance?

Another important point to consider when contrasting the different measures is whether emphasis should be placed on assessing the effect of variants on the mean risk in a population or on the genetic variation. Under a simple additive model and assuming that there is no dominance effect ($d=0$), the effect on the mean and the variance are $2pa$ and $2p(1-p)a^2$, respectively (TABLE 1). Therefore, a variant at or near fixation (that is, $p \approx 1$) can have a relatively large effect on the mean and no effect on the variance. Thus, for a given effect size, 'intervening' on more common variants may help to reduce disease risk regardless of the amount of variance explained. Nevertheless, if there are many risk variants for disease, then it will be effectively impossible to remove or affect all of them to decrease risk. In this case, it does not make sense to use measures that focus on the mean (for example, the PAF). Instead, we recommend using measures that help to understand and explain variation around the mean, which is a key component of genetic risk prediction.

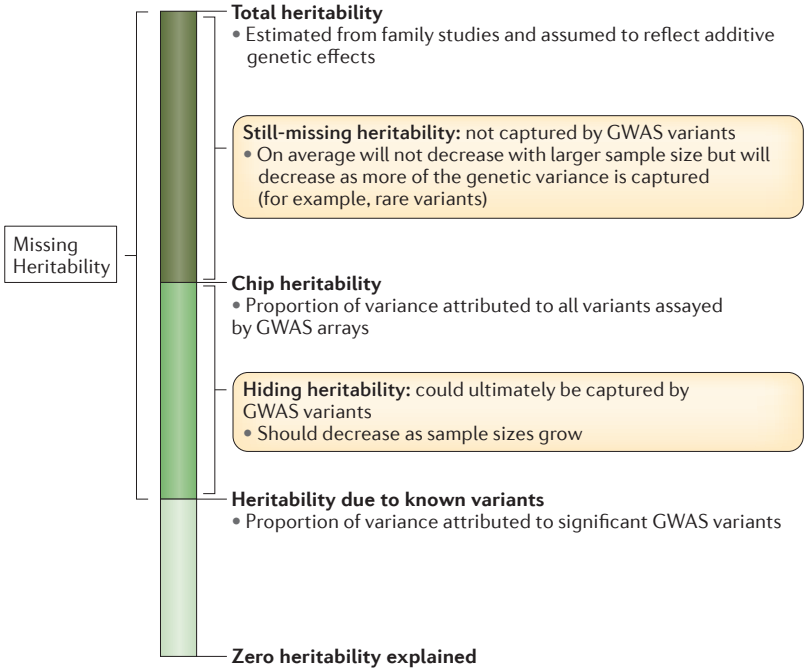


Figure 4 | **Aspects of disease heritability: known, hiding and missing.** A growing proportion of the total heritability estimated from family studies can be explained by known variants detected in existing genome-wide association studies (GWASs). This is one of the key measures considered here. The remaining heritability can be categorized as 'hiding' heritability and 'still-missing' heritability. The hiding heritability can be estimated from genome-wide arrays using the Genetic Relatedness Estimation through Maximum Likelihood (GREML) model[34]. The still-missing heritability may remain even after GWASs and could reflect different genetic architectures (for example, rare variants). Note that the total heritability may be biased upwards owing to confounding by non-additive genetic or non-genetic factors.

Extensions and additional measures

Our focus is on measures for a limited number of variants, in which we extend the one-locus methods to multiple loci under the assumption of independence among risk variants. Hence, the most associated locus from a region is usually used. Necessarily, this requires some arbitrary threshold on linkage disequilibrium, which becomes increasingly unsatisfactory as more associated loci are identified. To overcome this, associated loci can be fitted together in a regression analysis, and the variance explained that accounts for the interdependence between loci can be estimated. If the sample for discovery of the associated loci is used, then there may be some inflation of variance explained compared to the value obtained if the contribution was estimated from an independent sample drawn from the same population. Genomic profile risk scoring[15,35] is a strategy used to test the efficacy of associated SNPs identified in one sample for the contribution to variance in another sample. Briefly, risk alleles and their effect sizes identified by a GWAS carried out in a discovery sample are used to generate genomic profile risk scores (GPRSs) in an independent target sample, using SNPs with P values in the discovery sample that are below some user-defined threshold of significance. A GPRS is calculated for each individual in the target sample as the sum of the count of risk alleles weighted by the effect size in the discovery sample. The profile score is evaluated through regression of the target phenotype on the GPRS after accounting for other known covariates. The efficacy statistic is frequently Nagelkerke's R^2 or AUC, although expression on the liability scale may be more interpretable[4].

To account for the correlational structure among loci and to estimate the overall proportion of variance that is attributable to variants genome wide, one can use more complex mixed models that jointly fit all variants[5,36]. Such methods estimate the variance that is attributable to all variants together, which is known as chip heritability or SNP heritability. One can also partition this variance on the basis of variant annotation, for example, those in loci identified as associated with disease versus all remaining variants. In this case, one fits the genetic contribution from known disease-associated loci as one random effect and the genetic contribution from all other loci as another. Then, the ratio of these will provide an estimate of the extent to which known risk variants explain the overall chip heritability. These different components of heritability explained by genetic variants are illustrated in FIG. 4.

Note that genetic variation as evaluated here is not the only measuring stick for the utility of identified risk variants. A set of variants may have good clinical utility in a particular context (that is, for some patients) while not explaining much variation in the population and vice versa. Moreover, various measures besides the AUC have been proposed to assess the risk prediction properties of known variants[37]. However, as many of these measures do not yield a single bounded summary value and are context dependent, they are not useful for assessing genetic variation per se.

Conclusions and future perspectives

In genetic studies it is a common and useful practice to quantify the contribution to disease risk of each associated variant, the total contribution for all associated variants and the additional contribution compared to previous studies. Quantifying such successes across research projects can be hampered if different studies use different measures. Here, we present the different measures side-by-side, and compare the similarity and differences of these commonly used measures. We also provide an online tool to calculate these measures from association study summary statistics (see INDI-V online tool).

Although geneticists and epidemiologists often interpret different measures of the impact of risk variants on disease as providing similar information, as shown here they are not interchangeable and can give different messages. For common, low-risk variants the measures are fairly uniform. However, for risk variants with a range of RAFs and RRs, heritability explained is often substantially lower than sibling recurrence risk and logRR genetic variance. For rare, high-penetrance variants, the approximate heritability[16] and sibling recurrence risk can be an order of magnitude larger than other measures. The pAUC may be larger or smaller than the other measures depending on the nature of the risk alleles; the PAF gives much larger estimates than all other measures and has philosophical and computational limitations. As we move into the era of discovering both common and rare variants with varying penetrance for disease, we recommend investigators to focus primarily on the heritability of liability or the logRR genetic variance explained, as these seem to give estimates that are less sensitive to rare, high-risk variants than other measures.

Although the measures of the contribution to risk considered here may have similar underlying intentions, they can be on different scales and include different types and amounts of information. Depending on the measure, the apparent impact of genetic variants can hinge on the assumed overall risks of disease which, despite their apparent simplicity, are often difficult to pin down. All of the measures considered here, except the PAF, can be expressed relative to a maximum specified by parameters measured in twin or family studies: the 'denominator' (for example, total heritability, sibling recurrence risk and maximum AUC). The denominator measures are themselves difficult to estimate, may be contaminated by non-genetic factors and, for less common diseases, are subject to considerable sampling variance[38]. Moreover, these denominator estimates can be dependent on the study context owing to real differences that reflect environmental factors such as country, age, year and many other complexities of real-life data. Valid comparison of the numerators and denominators requires samples to be drawn from the same population. Thus, we recommend that investigators undertake sensitivity analyses that explore how their results vary when using a range of assumed underlying risks. The important message is that given such uncertainty, the concept of individual loci 'explaining' disease is less straightforward than it may seem at first sight, and all quantifications should therefore be considered in terms of benchmarking rather than as precise measures. In addition, calculating multiple different measures may provide valuable information about the sensitivity of results to the underlying assumptions.

Genetic and epidemiological study designs and analytic methods have nicely coalesced to help investigators to detect large numbers of risk variants for complex diseases. However, the different views of these disciplines can shade the interpretation and apparent implications of such findings. By juxtaposing the different models and measures used to assess the impact of genetic variants on disease, we highlight their strengths and weaknesses, and make various recommendations for their use. With this information and software provided as an online tool to calculate the measures considered here (see INDI-V online tool), one can judge what is truly meant when a study concludes that genetic variants explain or account for a particular proportion of disease.

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
2. Witte, J. S. Genome-wide association studies and beyond. *Annu. Rev. Publ. Health* **31**, 9–20 (2010).
3. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
4. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6**, e1000864 (2010).
5. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
6. Cole, P. & MacMahon, B. Attributable risk percent in case–control studies. *Br. J. Prev. Soc. Med.* **25**, 242–244 (1971).
7. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genet.* **44**, 247–250 (2012).
8. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40**, 955–962 (2008).
9. Wang, K. *et al.* Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.* **86**, 730–742 (2010).
10. Dempster, E. R. & Lerner, I. M. Heritability of threshold characters. *Genetics* **35**, 212–236 (1950).
This study explores the relationship between heritability on disease and liability scales.
11. Slatkin, M. Exchangeable models of complex inherited diseases. *Genetics* **179**, 2253–2261 (2008).
12. Falconer, D. The inheritance of liability to certain diseases, estimates from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
This paper presents a formal derivation of the relationship between disease risk in relatives and heritability, and also provides a thoughtful exploration of scenarios and caveats.
13. Falconer, D. & Mackay, T. F. *Introduction to Quantitative Genetics*, (Pearson Education, 1996).
14. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
This paper describes variance explained by a single locus on the disease and liability scale.
15. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
16. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genet.* **44**, 483–489 (2012).
17. Pharoah, P. D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genet.* **31**, 33–36 (2002).
This is a clear presentation of the logRR model.
18. Wray, N. R. & Goddard, M. E. Multi-locus models of genetic risk of disease. *Genome Med.* **2**, 10 (2010).
19. Pharoah, P. D., Day, N. E., Duffy, S., Easton, D. F. & Ponder, B. A. Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int. J. Cancer* **71**, 800–809 (1997).
20. James, J. W. Frequency in relatives for an all-or-none trait. *Ann. Hum. Genet.* **35**, 47–49 (1971).
21. Pharoah, P. D., Antoniou, A. C., Easton, D. F. & Ponder, B. A. Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**, 2796–2803 (2008).
22. Park, J. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet.* **42**, 570–575 (2010).
23. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
24. Chen, G.-B. *et al.* Estimation and partitioning of (co) heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum. Mol. Genet.* **23**, 4710–4720 (2014).

25. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
26. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genet.* **45**, 1150–1159 (2013).
27. Kirov, G. *et al.* Neurexin 1 (*NRXN1*) deletions in schizophrenia. *Schizophr Bull.* **35**, 851–854 (2009).
28. Kirov, G. *et al.* Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum. Mol. Genet.* **18**, 1497–1503 (2009).
29. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
30. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
31. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait — evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
32. Rockhill, B., Weinberg, C. R. & Newman, B. Population attributable fraction estimation for established breast cancer risk factors: considering the issues of high prevalence and unmodifiability. *Am. J. Epidemiol.* **147**, 826–833 (1998).
This study considers the limitations of the PAF.
33. Saha, S., Chant, D., Welham, J. & McGrath, J. A systematic review of the prevalence of schizophrenia. *PLoS Med.* **2**, e141 (2005).
34. Alonso, A., Logroscino, G., Jick, S. S. & Hernan, M. A. Incidence and lifetime risk of motor neuron disease in the United Kingdom: a population-based study. *Eur. J. Neurol.* **16**, 745–751 (2009).
35. Wray, N. R. *et al.* Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* http://dx.doi.org/10.1111/jcpp.12295 (2014).
36. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
37. Gail, M. H. & Pfeiffer, R. M. On criteria for evaluating models of absolute risk. *Biostatistics* **6**, 227–239 (2005).
38. Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nature Rev. Genet.* **14**, 139–149 (2013).
39. So, H. C., Gui, A. H., Cherny, S. S. & Sham, P. C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet. Epidemiol.* **35**, 310–317 (2011).
40. So, H. C., Li, M. & Sham, P. C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet. Epidemiol.* **35**, 447–456 (2011).
41. So, H. C., Kwan, J. S., Cherny, S. S. & Sham, P. C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.* **88**, 548–565 (2011).
This study uses variance explained by loci and considers complications of age-related risk.
42. Do, C. B., Hinds, D. A., Francke, U. & Eriksson, N. Comparison of family history and SNPs for predicting risk of complex disease. *PLoS Genet.* **8**, e1002973 (2012).
43. Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case–control association studies. *PLoS Genet.* **8**, e1003032 (2012).

Acknowledgements
The authors thank C. Nolan and B. Beyamin for developing the companion website, M. Robinson for help with the figure in Box 1, T. Hoffmann for help in plotting Figure 3, and J. Liu for linkage disequilibrium filtering of the breast cancer SNPs. This work is supported by the US National Institutes of Health grants R01 CA088164, U01 CA127298, U01 GM061390 and P30 CA82103, and by the Australian National Health and Medical Research Council grants 613602, 613601, 1011506, 1050218 and 1048853.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
Catalog of published GWASs: http://www.genome.gov/gwastudies/
INDI-V online tool: www.cnsgenomics.com/software

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table) | S3 (table) | S4 (table)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF