REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

Many common human diseases and traits are known to cluster in families and are believed to be influenced by several genetic and environmental factors, but until recently the identification of genetic variants contributing to these 'complex diseases' has been slow and arduous[1]. Genome-wide association studies (GWAS), in which several hundred thousand to more than a million single nucleotide polymorphisms (SNPs) are assayed in thousands of individuals, represent a powerful new tool for investigating the genetic architecture of complex diseases[1,2]. In the past few years, these studies have identified hundreds of genetic variants associated with such conditions and have provided valuable insights into the complexities of their genetic architecture[3,4].

The genome-wide association (GWA) method represents an important advance compared to 'candidate gene' studies, in which sample sizes are generally smaller and the variants assayed are limited to a selected few, often on the basis of imperfect understanding of biological pathways and often yielding associations that are difficult to replicate[5,6]. GWAS are also an important step beyond family-based linkage studies, in which inheritance patterns are related to several hundreds to thousands of genomic markers. Despite many clear successes in single-gene 'Mendelian' disorders[7,8], the limited success of linkage studies in complex diseases has been attributed to their low power and resolution for variants of modest effect[9–11].

The underlying rationale for GWAS is the 'common disease, common variant' hypothesis, positing that common diseases are attributable in part to allelic variants present in more than 1–5% of the population[12–14]. They have been facilitated by the development of commercial 'SNP chips' or arrays that capture most, although not all, common variation in the genome. Although the allelic architecture of some conditions, notably age-related macular degeneration, for the most part reflects the contributions of several variants of large effect (defined loosely here as those increasing disease risk by twofold or more), most common variants individually or in combination confer relatively small increments in risk (1.1–1.5-fold) and explain only a small proportion of heritability—the portion of phenotypic variance in a population attributable to additive genetic factors[3]. For example, at least 40 loci have been associated with human height, a classic complex trait with an estimated heritability of about 80%, yet they explain only about 5% of phenotypic variance despite studies of tens of thousands of people[15]. Although disease-associated variants occur more frequently in protein-coding regions than expected from their representation on genotyping arrays, in which over-representation of common and functional variants may introduce analytical biases, the vast majority (>80%) of associated variants fall outside coding regions, emphasizing the importance of including both coding and non-coding regions in the search for disease-associated variants[3].

[1]National Human Genome Research Institute, Building 31, Room 4B09, 31 Center Drive, MSC 2152, Bethesda, Maryland 20892-2152, USA. [2]National Institutes of Health, Building 1, Room 126, MSC 0148, Bethesda, Maryland 20892-0148, USA. [3]Departments of Medicine and Human Genetics, University of Chicago, Room A612, MC 6091, 5841 South Maryland Avenue, Chicago, Illinois 60637, USA. [4]Duke University, The Institute for Genome Sciences and Policy (IGSP), Box 91009, Durham, North Carolina 27708, USA. [5]National Human Genome Research Institute, Office of Population Genomics, Suite 4076, MSC 9305, 5635 Fishers Lane, Rockville, Maryland 20892-9305, USA. [6]Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA. [7]University of Oxford, Oxford Centre for Diabetes, Endocrinology and Metabolism, Churchill Hospital, Old Road, Oxford OX3 7LJ, UK, and Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [8]GlaxoSmithKline, 709 Swedeland Road, King of Prussia, Pennsylvania 19406, USA. [9]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 733 North Broadway BRB579, Baltimore, Maryland 21205, USA. [10]Yale University, Department of Medicine, Division of Digestive Diseases, 333 Cedar Street, New Haven, Connecticut 06520-8019, USA. [11]deCODE Genetics, Sturlugata 8, Reykjavik IS-101, Iceland. [12]Lewis-Sigler Institute for Integrative Genomics, Howard Hughes Medical Institute, and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA. [13]The Genome Center, Washington University School of Medicine, 4444 Forest Park Avenue, Campus Box 8501, Saint Louis, Missouri 63108, USA. [14]National Human Genome Research Institute, Center for Research on Genomics and Global Health, Building 12A, Room 4047, 12 South Drive, MSC 5635, Bethesda, Maryland 20892-5635, USA. [15]Department of Integrative Biology, University of California, 3060 Valley Life Science Building, Berkeley, California 94720-3140, USA. [16]Stanford University, Health Research and Policy, Redwood Building, Room T204, 259 Campus Drive, Stanford, California 94305, USA. [17]Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, USA. [18]Department of Molecular Biology and Genetics, 107 Biotechnology Building, Cornell University, Ithaca, New York 14853, USA. [19]Howard Hughes Medical Institute and University of Washington, Department of Genome Sciences, 1705 North-East Pacific Street, Foege Building, Box 355065, Seattle, Washington 98195-5065, USA. [20]University of Queensland, School of Biological Sciences, Goddard Building, Saint Lucia Campus, Brisbane, Queensland 4072, Australia. [21]Vanderbilt University, Center for Human Genetics Research, 519 Light Hall, Nashville, Tennessee 37232-0700, USA. [22]Department of Genetics, North Carolina State University, Box 7614, Raleigh, North Carolina 27695, USA. [23]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, NRB 0330, Boston, Massachusetts 02115, USA. [24]Queensland Institute of Medical Research, 300 Herston Road, Brisbane, Queensland 4006, Australia.

The questions arise as to why so much of the heritability is apparently unexplained by initial GWA findings, and why it is important. It is important because a substantial proportion of individual differences in disease susceptibility is known to be due to genetic factors, and understanding this genetic variation may contribute to better prevention, diagnosis and treatment of disease. It is important to recognize, however, that few investigators expected these studies immediately to find all of the variants associated with common diseases, or even most of them; the hope was that they would at least find some[16]. Limitations in the design of early GWAS, such as imprecise phenotyping and the use of control groups of questionable comparability, may have reduced estimates of effect sizes while preserving some ability to identify associated variants[17]. These studies have considerably surpassed early expectations, reproducibly identifying hundreds of variants in many dozens of traits, but for many traits they have explained only a small proportion of estimated heritability[18].

Many explanations for this missing heritability have been suggested, including much larger numbers of variants of smaller effect yet to be found; rarer variants (possibly with larger effects) that are poorly detected by available genotyping arrays that focus on variants present in 5% or more of the population; structural variants poorly captured by existing arrays; low power to detect gene–gene interactions; and inadequate accounting for shared environment among relatives. Consensus is lacking, however, on approaches and priorities for research to examine what has been termed 'dark matter' of genome-wide association—dark matter in the sense that one is sure it exists, can detect its influence, but simply cannot 'see' it (yet). Here we examine potential sources of missing heritability and propose research strategies to illuminate the genetics of complex diseases.

Heritability and allelic architecture of complex traits

It is reasonable to assume that allelic architecture (number, type, effect size and frequency of susceptibility variants) may differ across traits, and that missing heritability may take a different form for different diseases[19], but at present our understanding is too limited to distinguish these possibilities. Age-related macular degeneration may provide the best example of a common disease in which heritability is substantially explained by a small number of common variants of large effect[20], but for other conditions, such as Crohn's disease, the proportion of heritability explained is not nearly so large despite a much larger number of identified variants[21] (Table 1). There are no obvious differences between these two traits in genetic architecture as predicted from clinical and epidemiological data that would explain the differences observed in their allelic architecture. Some apparent differences may simply be due to differences in the stage of investigation across traits. Studies in several conditions have clearly demonstrated that the number of detected variants increases with increasing sample size[22–24].

Population genetic theory suggests an explanation for the paucity of variants explaining a large proportion of disease predisposition, in that decreased reproductive fitness should typically act to reduce the frequencies of high-risk variants. This might explain the relative lack of variants detected so far for some neuropsychiatric conditions, such as autism spectrum disorders, given their low reproductive fitness[25]. Yet for a condition such as type 1 diabetes, which has a similar prevalence, familial risk, early onset and poor reproductive fitness (at least before the discovery of insulin therapy), more than 40 loci have already been reported; this might be because the overall sample sizes studied in type 1 diabetes have been very large[26,27]. Present-day reproductive fitness may correlate poorly with the forces that have shaped variation throughout human evolution; moreover focusing on the reproductive effects of a single disease ignores the pleiotropic effects (effects of the same variant on multiple characteristics or disease risks) of multiple alleles influencing that condition simultaneously with many other conditions[28].

Selection might also be responsible for keeping genetic effect sizes low, as variants of larger effect may be selected against and eventually disappear[19]. Long-term stabilizing selection minimizes the production of individuals at the extremes of a trait[29], in part by reducing the additive genetic effects of alleles already present or those arising *de novo* by mutation[30] to levels potentially beneath the ability of studies of feasible size to detect them. Selection may also contribute to differences in the ability to detect loci in different complex diseases, if genetic susceptibility to some diseases is more strongly affected by selection than other diseases, or if environmental perturbations vary in intensity across diseases. Immune and infectious agents have been recognized as among the strongest selection pressures in human evolution[31], and immune-related genes have been strongly implicated in Crohn's disease and other immune-mediated diseases[3], suggesting either that pleiotropic effects of these variants reduce the efficiency of negative selection, or that strong environmental perturbation in modern societies might expose the disease risk associated with these variants. Selection may thus explain why disease allele frequencies are low and allelic effects are small, but this should manifest as low, rather than missing, heritability.

A probable contributor to the small genetic effect sizes observed so far is that current investigations have incompletely surveyed the potential causal variants within each gene. Relative risks observed for marker SNPs may underestimate the actual risks associated with the true causal variants. Notably, 11 out of 30 genes implicated as carrying common variants associated with lipid levels also carry known rare alleles of large effect identified in Mendelian dyslipidemias, including *ABCA1*, *PCSK9* and *LDLR*[22,32], suggesting that genes containing common variants with modest effects on complex traits may also contain rare variants with larger effects.

An important consideration is that the overwhelming majority of GWAS and other genetic studies have been limited to European ancestry populations, whereas genetic variation is greatest in populations of recent African ancestry[2], and studies in non-Europeans have yielded intriguing new variants[33,34]. Studies of populations of recent African ancestry in particular is likely to increase the yield of rare variants and narrow the large chromosomal regions of association identified in the 'younger' population due to extended linkage disequilibrium, or the tendency for adjacent genetic loci to be inherited together[31]. Isolated populations may also be of value given their potential to be enriched in unique variants[35].

The accuracy of current heritability estimates is also important, because experimentally identified variants could never explain all the variance in an erroneously inflated heritability estimate. Heritability of quantitative traits, formally defined as the proportion of phenotypic variance in a population attributable to additive genetic factors (narrow-sense heritability, h^2 (ref. 36)) is typically estimated from

Table 1 | Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual* phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

* Residual is after adjustment for age, gender, diabetes.

family studies, and can be expected to vary across environments. Narrow-sense heritability estimates in humans can be inflated if family resemblance is influenced by non-additive genetic effects (dominance and epistasis, or gene–gene interaction), shared familial environments, and by correlations or interactions among genotypes and environment[36,37]. However, heritabilities estimated from pedigree studies in animals agree well with heritability estimated from response to artificial selection, suggesting that estimates from family studies are not necessarily inflated.

Teasing apart the contributions to heritability of environmental factors shared among relatives will soon be possible because the availability of genome-wide markers now provides empirical estimates of identity-by-descent (IBD) allele sharing between pairs of relatives. For example, full sibs share on average half their genetic complement, but this proportion can vary—in one large study it ranged from 0.37 to 0.62 (ref. 38). By relating phenotypic differences to the observed IBD sharing fraction among sib pairs, marker data were used to generate a heritability estimate of 0.8 for height[38]. This is remarkably consistent with estimates using traditional methods but free of their assumptions, suggesting that for height at least, heritability is not overestimated. Applying such estimation to distantly related or 'unrelated' individuals is now feasible using dense genomic scans[39]; given the number of people with dense genotyping data, heritability estimates could be generated for a wide variety of traits free of potential confounding by unmeasured shared environment.

Improving estimates of all contributors to heritability will facilitate determination of the proportion of genetic variance that has been explained. Despite imprecision in current estimates, it may still be possible to know that 'all the heritability' has been explained by predicting phenotypes in a new set of individuals from trait-associated markers, and correlating the predicted phenotypes with the actual values. If the markers truly explain all the additive genetic variance, the squared correlation between predicted and actual phenotype will be equal to the heritability[40]. Population-based heritability estimates thus provide a valuable metric for completeness of available genetic risk information, but individualized disease prevention and treatment will ultimately require identifying the variants accounting for risk in a given individual rather than on a population basis.

Rare variants and unexplained heritability

Much of the speculation about missing heritability from GWAS has focused on the possible contribution of variants of low minor allele frequency (MAF), defined here as roughly 0.5% < MAF < 5%, or of rare variants (MAF < 0.5%). Such variants are not sufficiently frequent to be captured by current GWA genotyping arrays[14,41], nor do they carry sufficiently large effect sizes to be detected by classical linkage analysis in family studies (Fig. 1). Once MAF falls below 0.5%, detection of associations becomes unlikely unless effect sizes are very large, as in monogenic conditions. For modest effect sizes, association testing may require composite tests of overall 'mutational load', comparing frequencies of mutations of potentially similar functional effect in cases and controls.

Low frequency variants could have substantial effect sizes (increasing disease risk two- to threefold) without demonstrating clear Mendelian segregation, and could contribute substantially to missing heritability[42]. For example, 20 variants with risk allele frequency of 1% and allelic odds ratio (or probability of an event occurring divided by the probability of it not occurring, compared in people with versus without the risk allele) of three would account for most familial aggregation of type 2 diabetes. There are relatively few examples of such variants contributing to complex traits, possibly owing to insufficiently large sample sizes or insufficiently comprehensive arrays.

The primary technology for the detection of rare SNPs is sequencing, which may target regions of interest, or may examine the whole genome. 'Next-generation' sequencing technologies, which process millions of sequence reads in parallel, provide monumental increases in speed and volume of generated data free of the cloning biases and

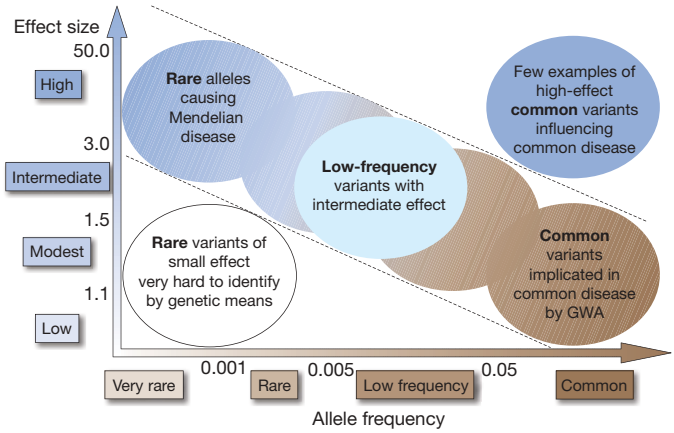

Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

arduous sample preparation characteristic of capillary sequencing[43]. Detection of associations with low frequency and rare variants will be facilitated by the comprehensive catalogue of variants with MAF ≥ 1% being generated by the 1,000 Genomes Project (http://www.1000genomes.org/page.php), which will also identify many variants at lower allele frequencies. The pilot effort of that program has already identified more than 11 million new SNPs in initially low-depth coverage of 172 individuals[44].

Current mechanisms for using sequencing to identify rare variants underlying or co-located with GWA-defined associations include sequencing in genomic regions defined by strong and repeatedly replicated associations with common variants, and sequencing a larger fraction of the genome in people with extreme phenotypes. In the absence of GWA-defined signals, sequencing candidate genes in subjects at the extremes of a quantitative trait (such as lipid levels or the age at onset), can identify other associated variants, both common and rare[45,46]. An important finding from these studies is that much of the information is provided by people at the extremes of trait distributions, who seem to be more likely to carry loss-of-function alleles[47].

Sample sizes used for the initial identification of DNA sequence variants have generally been modest, and sample size requirements increase essentially linearly with 1/MAF. Much larger samples are needed for the identification of associations with variants than those needed for the detection of the variants themselves. They also scale roughly linearly with 1/MAF given a fixed odds ratio and fixed degree of linkage disequilibrium with genotyped markers. Sample size for association detection also scales approximately quadratically with $1/|(OR-1)|$, and thus increases sharply as the odds ratio (OR) declines. Sample size is even more strongly affected by small odds ratios than by small MAF, so low frequency and rare variants will need to have higher odds ratios to be detected.

Complicating matters further, numerous rare variants may be detected in a gene or region but they may have disparate effects on phenotype. Common variants have typically been analysed individually[23,48], but with one or two carriers of each rare variant, pooling them using specific criteria becomes attractive[47,49,50]. Pooling variants of similar class increases the effective MAF of the class and reduces the number of tests performed, but raises several other questions (Box 1).

Determining which of the multitude of variants carried by an individual are responsible for a given phenotype represents a massive task, especially if the causal alleles are relatively anonymous in terms of known functional consequences. Because only a small proportion will have obvious functional consequences for the resultant protein, lesser evidence of association may suffice to implicate variants of this sort. The best approaches for combining functional credibility and statistical support in the evaluation of such variants remain to be

Box 1 | Research strategies using rare and low frequency variants and structural variants

Research strategies using rare and low frequency and structural variants include: (1) using expanding catalogues of human sequence variation[44], by linkage disequilibrium of rare/low frequency/structural variants with GWA-genotyped SNPs and/or improved detection methods, to identify variants underlying association signals identified by SNP arrays. (2) Improving approaches for using common SNPs to predict and control for differences in rare and low frequency SNPs. (3) Using targeted sequencing judiciously, focusing on people with extreme or unusual phenotypes. (4) Including populations of recent African ancestry in sequencing studies to increase yield of rare variants and narrow large linkage disequilibrium blocks; consider isolated or founder populations potentially enriched with unique variants. (5) Focusing discovery efforts on well-phenotyped groups, accessible families with large sibships, and families that allow return to family members for iterative phenotyping. (6) Increasing emphasis on other structural variants such as inversions and translocations. (7) Implementing chromosomal-region-specific matching throughout the genome, to select for each case and for each part of their genome—a control that is more similar to the case within that genomic region rather than matching genome-wide using measures such as geographic ancestry. (8) Pooling rare variants for analysis using logical criteria, by addressing the questions: do the different rare variants increase or decrease disease risk? What classes of variants should be pooled? What is the optimal level of MAF for pooling? (9) Improving CNV detection by developing more extensive population databases in large cohorts to understand allele and mutation frequency, inheritance among unaffected individuals, and CNV calling algorithms.

determined. GWAS have tended to focus almost exclusively on statistical evidence and de-emphasize considerations of biological plausibility, but the challenges of sifting through the millions of rare variants in which two individuals differ may prompt a return to biology if rare variants are to be grouped and analysed properly.

The sheer number of inter-individual differences, mostly rare, to be detected by whole-genome sequencing (roughly 0.4% of 3 billion base pairs[51]) also raises the question of finding appropriate comparison subjects, or allelic matches, because people carrying rare variants at some loci may have important differences in ancestry or other factors from a general population. To reduce the number of variants that must be considered in a case-control comparison it would be useful to implement chromosomal-region-specific matching throughout the genome, to select closely related alleles and regions from the comparison population, thereby greatly reducing the number of incidental allelic differences from cases.

Structural variation and unexplained heritability

Structural variation, including copy number variants (CNVs, such as insertions and deletions) and copy neutral variation (such as inversions and translocations), may account for some of the unexplained heritability if those variants contribute to the genetic basis of human disease and are incompletely assessed by commercial SNP genotyping arrays. Although this type of variation has not been explicitly examined in most GWAS until now, CNVs in particular (regions 1 kilobase (kb) or longer present in variable numbers across individuals) have gained attention as methods to detect them have improved[52,53]. Other forms of structural variation such as inversions, translocations, microsatellite repeat expansions, insertions of new sequence, and complex rearrangements have been implicated in rare Mendelian conditions. For the most part such variation has been largely unexplored in relation to complex traits[54].

Variation due to CNVs arises from a combination of rare and common alleles; as with SNPs most variants are rare but most of the differences between any two individuals arise from a limited set of common (MAF $\geq$ 5%) copy number polymorphisms (CNPs)[55]. Disease-associated CNVs detected so far, like disease-associated SNPs, include rare variants with large associated effect sizes, and common variants with more modest effects but carried by a large proportion of the population (Table 2). An added twist is that rare, highly penetrant CNVs have generally been large (600 kb–3 megabases (Mb), affecting many genes), whereas disease-associated common CNPs have been much smaller (20–45 kb) and have identified specific genomic features for follow-up study. Because both rare and common CNVs are under-ascertained by current methods, the relative affect of these variants will continue to be an important research question for CNVs just as for SNPs. Of note, CNVs arising *de novo* in current cases and shown to be of importance in neuropsychiatric and developmental conditions[56–58] will not contribute to family resemblance and heritability, but could explain some of the variation at present attributed to 'environment'.

Several approaches have been developed for integrating analysis of CNVs into GWAS, including innovation in the design of GWA arrays (with associated discoveries in neuropsychiatric disorders[59,60]) and the use of the linkage disequilibrium relationships between SNPs and common CNPs (with associated discoveries in Crohn's disease and body weight[52,61]). These approaches are early in their development and have important limitations, although rapid progress is expected as CNV detection algorithms evolve and large-scale sequencing studies produce comprehensive, high-resolution maps of segregating CNPs that can be measured in large reference panels.

Many GWA data sets already have sufficient genotype and intensity information to permit calling of large, rare CNVs even if specific CNV probes were not included. As with non-structural single nucleotide sequence variants, more detailed ('iterative') phenotyping in relatives may reveal subtle phenotypic effects that were not initially appreciated.

Harnessing family studies

Family studies provide several opportunities for the investigation and interpretation of as-yet-unidentified genetic variation of many types

Table 2 | Selected disease associations with rare CNVs and common CNPs

Disease	Locus	Type of CNV	Size (kb)	Population frequency	Case frequency	Effect size (OR)
Rare CNVs						
Autism/IMR[59]	16p11.2	*De novo* deletion	600	1×10^{-4}	1%	100
Autism[59]	16p11.2	Rare duplication	600	3×10^{-4}	0.50%	16
Schizophrenia[60,78]	1q21.1	Rare deletion	1,400	2×10^{-4}	0.30%	15
IMR[79]	1q21.1	Rare deletion	1,400	2×10^{-4}	0.47%	Not observed in 4,737 controls
Schizophrenia[60,78]	15q13.3	Rare deletion	1,600	2×10^{-4}	0.20%	12
Epilepsy[80]	15q13.3	Rare deletion	1,600	2×10^{-4}	1.0%	Not observed in 3,699 controls
IMR[79,81]	15q13.3	Rare deletion	1,600	2×10^{-4}	0.30%	Not observed in 960 controls
Schizophrenia[82]	22q11.2	Rare deletion	3,000	2.5×10^{-4}	1%	40
Common CNPs						
Crohn's disease[83]	*IRGM*	Deletion polymorphism	20	7%	11%	1.5
Body mass index[61]	*NEGR1*	Deletion polymorphism	45	65%	Quantitative trait	<1 kg
Psoriasis[84]	*LCE3C*	Deletion polymorphism	30	55%	65%	1.3

IMR, idiopathic mental retardation.

underlying complex diseases (Box 2). Family studies may facilitate the detection of rare and low frequency variants, and the identification of their associations with common diseases, because predisposing variants will be present at much higher frequency in affected relatives of an index case.

Family studies also permit the investigation of parent-of-origin-specific effects, as have been reported for structural variants[62,63]. If not properly accounted for, such effects could mask associations and diminish the proportion of heritability explained. High-density SNP data in extended pedigrees can be used to localize predisposition genes, as unexpectedly long runs of identity-by-state sharing among affected relatives suggest true IBD that is probably due to an underlying genetic cause[64]. Linkage data can also enhance the power of high-density GWA scans by essentially relaxing *P*-value thresholds in the few instances in which suggestive findings overlap but are not definitive[65]. Family studies may also be useful in identifying gene–gene interactions, because affected relatives are more likely to share two nearby epistatic loci in linkage disequilibrium that would be unlinked in unrelated individuals[66,67].

Strategies for existing and future GWAS

The nearly 400 GWAS published so far represent a wealth of data on the genetics of complex diseases[4]. These studies have provided valuable insights into the genetics of common diseases, particularly about the underlying genetic architecture of complex traits and the predominance of non-coding variants that may have a role in their aetiology. Just as linkage studies demonstrated that complex diseases cannot be explained by a small number of rare variants with large effects, GWAS have shown that they cannot be explained by a limited number of common variants of moderate effect (Fig. 1). The distinction between low frequency and truly rare alleles is largely an operational one, relating to the potential, given realistic effect sizes, for detecting associations with low frequency variants by GWAS at attainable sample sizes. Low frequency variants of intermediate effect might also contribute to explaining missing heritability that should be tractable through large meta-analyses and/or imputation of genome-wide association data.

GWAS will probably remain an efficient way of investigating the remaining heritability, because their association signals may well define the genomic regions where rare variants, structural variants, and other forms of underlying variation are likely to cluster. The value of future studies can be enhanced by expanding to non-European samples and less common diseases and including more precise phenotypes and measures of environmental exposures[48,68] (Box 3). Information on lower frequency alleles emerging from projects such as the 1,000 Genomes will be used to produce even more comprehensive GWA arrays, and will facilitate the investigation of the lower frequency spectrum without the need for *de novo* sequencing.

Box 2 | Using family studies to investigate missing heritability

To investigate missing heritability using family studies, the following measures are required: (1) examine phenotypic effects of rare variants, particularly for subtle phenotypic abnormalities. (2) Investigate mutation rates and inheritance patterns of recurrent mutations. (3) Assess inheritance patterns of rare and structural variants. (4) Investigate parent-of-origin-specific effects. (5) Enhance power for identifying associated loci by studying affected sibs, particularly for conditions with substantial genetic heterogeneity. (6) Identify associated loci by unexpectedly long runs of identity-by-state sharing among distantly related affected relatives. (7) Enhance power of GWA scans by up-weighting *P* values in preselected regions based on linkage signals. (8) Identify gene–gene interactions by positive correlations between family-specific logs odds ratio (lod) scores or evidence of linkage disequilibrium among unlinked loci.

Potential of research to explain missing heritability

GWAS were initially designed to focus on the higher end of the frequency-effect size spectrum, so much work remains to be done, both in finding other variants in the lower frequency and larger effect domains shown in Fig. 1, and in understanding their functional and pathophysiological properties. To the extent that there are several causal variants on a common haplotype or that causal variants are in imperfect linkage disequilibrium with genotyped markers, marker SNPs will underestimate the associated disease risk.

The modest size of genetic effects detected so far confirms the multifactorial aetiology of these conditions and suggests that complex diseases will require substantially greater research effort to detect additional genetic influences. Near-term approaches for finding missing heritability on which there seems to be wide agreement include: targeted or whole-genome sequencing in people with extreme phenotypes, especially those with available family members and consent for recontact and iterative phenotyping; use of expanded reference panels of genomic variation such as 1,000 Genomes to enhance coverage of existing and future GWAS; mining of existing GWAS for associations with structural variants and evidence of gene–gene interactions; improved methods for detection of CNVs and other structural variants, applied to large, well-phenotyped groups and families; and expansion of sample sizes for numerous complex diseases through larger individual studies and meta-analyses, including people of non-European ancestry.

Given all that has been learned of the genetic architecture of common diseases in the past few years, it may also be worthwhile to attempt exhaustive characterization of some well-studied traits by cataloguing all the contributing variation, be it in DNA sequence, DNA structure, chromatin structure, environmental modifiers, and defining all its functional implications. Potential criteria for deciding which traits to pursue aggressively in this way might include the strength and robustness of detected associations, evidence that associations are disrupted by varying linkage disequilibrium patterns, documented associations of identified loci with multiple traits, and public health importance of the traits to be studied.

Box 3 | Making the most of existing and future GWAS

The following steps can be used to make the most of existing and future GWAS: (1) ensure the wide availability of data with appropriate protections for consent and privacy. (2) Increase sample sizes and ensure thorough meta- and mega-analyses of comparable data, with increased focus on conditions with relatively small sample sizes studied so far. (3) Expand studies to non-European samples and more diverse diseases. (4) Improve phenotyping by expanding to subtler or more quantitative or precise phenotypes as needed to reduce heterogeneity or explore pleiotropic effects. (5) Capture larger proportion of variation in implicated genes. (6) Enhance the investigation of the X chromosome, particularly as the methods for imputation of X and Y markers improve. (7) Investigate gene–gene interactions, including dominance and epistasis. (8) Investigate gene–environment interactions: measure environment rigorously and analyse it against GWA data; examine rare exposures in common diseases for unusual responders; consider including GWA in monozygotic twins or migrant studies to identify gene–environment interaction interactions; conduct suitably large (several hundred thousand people) prospective cohort studies with GWA genotyping, and reproducible reliable exposure measures at baseline; include routine biobanking of material suitable for epigenetic analysis, such as non-immortalized lymphocytes for DNA methylation or cryopreserved cell or nuclear preparations for chromatin studies; relate quantitative phenotypes to epigenetic variation, which unlike SNPs is inherently quantitative; measure epigenetic variants in appropriate tissues when technically feasible. (9) Measure CNVs: use linkage disequilibrium patterns of SNP data and improved maps and imputation methods to identify common CNPs; use SNP intensity data to identify large CNVs where feasible regions; use best possible CNP typing array until using next generation sequencing for this purpose becomes feasible.

Explaining missing heritability, however intellectually satisfying, will probably have fewer practical applications as an end in itself than as a means to an end. The ultimate goal of this line of research, as with nearly all research in the genetics of complex disease, is to improve understanding of human physiology and disease aetiology so that more effective means of diagnosis, treatment and prevention can be developed. If a genetic variant(s) was found that opened the door to effective new treatments at low cost and with minimal side effects (LDL-receptor mutations and the statin class of drugs comes to mind), one would probably be content to leave some heritability unexplained. It is the expectation that associations identified by GWAS or other genomic methods will eventually enable effective disease prevention or treatment, either through delineation of the functional properties of variants recognized at present, or identification of new variants in which true functionality lies, that primarily motivates the hunt for missing heritability.

It is more difficult to imagine predictive variants accounting for a sizeable proportion of disease risk without also explaining a sizeable proportion of heritability, and the limited incremental value in disease prediction of variants identified so far suggests that genetic prediction of complex diseases on a population basis will be challenging[69–71]. Still, the identification of even many hundreds of risk variants of small effect should permit identification of the small proportion of a population at the highest genetically defined risk, in which targeted prevention strategies should be explored. If testing of such variants was to be conducted across several diseases, as is now feasible with dense genome-wide association genotyping and will be greatly facilitated by whole-genome sequencing, a sizeable number of people could be identified to be at greatly increased risk for at least one disease. Identification of genetic variants that influence disease risk, prognosis, or the response to treatment should enable the development of diagnostic and interventional strategies that are safe, effective and as necessary, individualized[71], although the value of genetic variants in disease prediction and the steps needed to realize this are widely debated[69,70]. Given how little has actually been explained of the demonstrable genetic influences on most common diseases, despite identification of hundreds of associated genetic variants, the search for missing heritability provides a potentially valuable path towards further discoveries.

1. Hardy, J. & Singleton, A. Genomewide association studies and human disease. *N. Engl. J. Med.* **360,** 1759–1768 (2009).
2. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
3. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci.* **106,** 9362–9367 (2009).
Comprehensive analysis of genomic annotations for disease-associated SNPs defined by GWAS, showing great majority of associated loci in intronic or intergenic regions of unknown function.
4. Hindorff, L. A., Junkins, H. A., Mehta, J. P. & Manolio, T. A. A catalog of published genome-wide association studies. Available at ⟨http://www.genome.gov/26525384⟩ (accessed, 18 September 2009).
5. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4,** 45–61 (2002).
6. Todd, J. A. Statistical false positive or true disease pathway? *Nature Genet.* **38,** 731–733 (2006).
7. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261,** 921–923 (1993).
8. Lifton, R. P. Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lect.* **100,** 71–101 (2004).
9. Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genome-wide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* **69,** 936–950 (2001).
10. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).
11. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405,** 847–856 (2000).
12. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17,** 502–510 (2001).
13. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278,** 1580–1581 (1997).
14. Pritchard, J. K. Are rare variants responsible for susceptibility to common diseases? *Am. J. Hum. Genet.* **69,** 124–137 (2001).
15. Visscher, P. M. Sizing up human height variation. *Nature Genet.* **40,** 489–490 (2008).
16. Collins, F. S. 2005 William Allan Award address. No longer just looking under the lamppost. *Am. J. Hum. Genet.* **79,** 421–426 (2006).
17. Pearson, T. A. & Manolio, T. A. How to interpret a genome-wide association study. *J. Am. Med. Assoc.* **299,** 1335–1344 (2008).
18. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).
19. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant.or not? *Hum. Mol. Genet.* **11,** 2417–2423 (2002).
20. Jakobsdottir, J., Gorin, M. B., Conley, Y. P., Ferrell, R. E. & Weeks, D. E. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet.* **5,** e1000337 (2009).
21. 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).
22. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genet.* **41,** 56–65 (2009).
23. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40,** 638–645 (2008).
24. Ahmed S. . *et al.* Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nature Genet.* **41,** 585–590 (2009).
25. Lord, C., Cook, E. H., Leventhal, B. L. & Amaral, D. G. Autism spectrum disorders. *Neuron* **28,** 355–363 (2000).
26. Cooper, J. D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genet.* **40,** 1399–1401 (2008).
27. Barrett J. C. . *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genet.* **41,** 703–707 (2009).
28. Keller, M. C. & Miller, G. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behav. Brain Sci.* **29,** 385–404 (2006).
29. Gibson, G. & Wagner, G. Canalization in evolutionary genetics: a stabilizing theory? *Bioessays* **22,** 372–380 (2000).
30. Gibson, G. Decanalization and the origin of complex disease. *Nature Rev. Genet.* **10,** 134–140 (2009).
31. Campbell, M. C. & Tishkoff, S. A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* **9,** 403–433 (2008).
32. Lusis, A. J. & Pajukanta, P. A treasure trove for lipoprotein biology. *Nature Genet.* **40,** 129–130 (2008).
33. Zheng, W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nature Genet.* **41,** 324–328 (2009).
34. Yasuda, K. *et al.* Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nature Genet.* **40,** 1092–1097 (2008).
35. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genet.* **41,** 35–46 (2009).
36. Falconer, D. S. & Mackay, T. F. C. *Introduction to Quantitative Genetics Addison* 123 (Wesley Longman Ltd, 1996).
37. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era–concepts and misconceptions. *Nature Rev. Genet.* **9,** 255–266 (2008).
Detailed review of strengths, weaknesses and controversies in estimations of heritability from human, agricultural and experimental studies.
38. Visscher, P. M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2,** e41 (2006).
39. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157,** 1819–1829 (2001).
40. Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E. & Visscher, P. M. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet.* **4,** e1000231 (2008).
41. McCarthy, M. I. & Hirschhorn, J. N. Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* **17** (R2), R156–R165 (2008).
Insightful review of initial findings from GWAS, the heritability that they do and do not explain, and potential for progress from other GWAS, identification of rare variants, and studies of epigenetics and gene expression and function.
42. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9,** 356–369 (2008).
43. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).
44. Abecasis, G. R. The 1000 Genomes Project: analysis of pilot datasets. *Biology of Genomes* page 246 (Cold Spring Harbor Laboratory, 5–9 May 2009).
45. Kotowski, I. K. *et al.* A spectrum of *PCSK9* alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* **78,** 410–422 (2006).
46. Cohen, J. C. *et al.* Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl Acad. Sci. USA* **103,** 1810–1815 (2006).
47. Romeo, S. *et al.* Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. *Nature Genet.* **39,** 513–516 (2007).
48. Haiman, C. A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genet.* **39,** 638–644 (2007).

49. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324,** 387–389 (2009).
Four rare variants in *IFIH1* independently lowering risk of type 1 diabetes were identified by sequencing exons and splice sites of 10 genes under GWA-defined peaks, demonstrating the power of intensive sequencing to identify potentially causative variants in follow-up of GWAS.
50. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321 (2008).
51. Crawford, M. H. *Anthropological Genetics: Theory, Methods and Applications* 341 (Cambridge Univ. Press, 2006).
52. McCarroll, S. A. Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* **17** (R2), R135–R142 (2008).
53. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature Genet.* **39** (suppl.), S7–S15 (2007).
54. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).
55. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40,** 1166–1174 (2008).
Initial map of CNVs demonstrating high proportion (>80%) of inter-individual differences in copy number differences due to common CNVs of MAF 5% or greater; >99% of CNVs probably derived from inheritance rather than *de novo* mutation; and most common diallelic CNVs in strong linkage disequilibrium with common SNPs.
56. de Vries, B. B. *et al.* Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77,** 606–616 (2005).
57. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
58. Xu, B. *et al.* Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nature Genet.* **40,** 880–885 (2008).
59. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
60. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232–236 (2008).
61. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41,** 25–34 (2009).
62. Abrahams, B. S. & Geschwind, D. H. Advances in autism genetics: on the threshold of a new neurobiology. *Nature Rev. Genet.* **9,** 341–355 (2008).
63. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40,** 1068–1075 (2008).
64. Thomas, A., Camp, N. J., Farnham, J. M., Allen-Brady, K. & Cannon-Albright, L. A. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Ann. Hum. Genet.* **72,** 279–287 (2008).
65. Roeder, K., Bacanu, S. A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* **78,** 243–252 (2006).
66. MacLean, C. J., Sham, P. C. & Kendler, K. S. Joint linkage of multiple loci for a complex disorder. *Am. J. Hum. Genet.* **53,** 353–366 (1993).
67. Zhao, J., Jin, L. & Xiong, M. Test for interaction between two unlinked loci. *Am. J. Hum. Genet.* **79,** 831–845 (2006).
68. Waters, K. M. *et al.* Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. *Cancer Epidemiol. Biomarkers Prev.* **18,** 1285–1289 (2009).
69. Clayton, D. G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet.* **5,** e1000540 (2009).
70. Khoury, M. J. *et al.* The scientific foundation for personal genomics: recommendations from a National Institutes of Health-Centers for Disease Control and Prevention multidisciplinary workshop. *Genet. Med.* **11,** 559–567 (2009).
71. 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).
72. Maller, J. *et al.* Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nature Genet.* **38,** 1055–1059 (2006).
73. International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN). Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. *Nature Genet.* **40,** 204–210 (2008).
74. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40,** 638–645 (2008).
75. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genet.* **40,** 189–197 (2008).
76. Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genet.* **41,** 334–341 (2009).
77. Prokopenko, I. *et al.* Variants in *MTNR1B* influence fasting glucose levels. *Nature Genet.* **41,** 77–81 (2009).
78. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237–241 (2008).
79. Mefford, H. C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* **359,** 1685–1699 (2008).
80. Helbig, I. *et al.* 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature Genet.* **41,** 160–162 (2009).
81. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genet.* **40,** 322–328 (2008).
82. Bassett, A. S., Marshall, C. R., Lionel, A. C., Chow, E. W. & Scherer, S. W. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. *Hum. Mol. Genet.* **17,** 4045–4053 (2008).
83. McCarroll, S. A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nature Genet.* **40,** 1107–1112 (2008).
84. de Cid, R. *et al.* Deletion of the late cornified envelope *LCE3B* and *LCE3C* genes as a susceptibility factor for psoriasis. *Nature Genet.* **41,** 211–215 (2009).

Acknowledgements This paper is inspired by the deliberations of an expert working group convened by the National Human Genome Research Institute (NHGRI) on 2–3 February 2009, to address the heritability unexplained in GWAS. The authors acknowledge the participation of J. C. Cohen, M. Daly and A. P. Feinberg in the workshop.

Author Contributions T.A.M., F.S.C., N.J.C., D.B.G., L.A.H., D.J.H., M.I.M. and E.M.R. planned and participated in the workshop; L.R.C., A.C., J.H.C., A.E.G., A.K., L.K., E.M., C.N.R., M.S., D.V., A.S.W., M.B., A.G.C., E.E.E., G.G., J.L.H., T.F.C.M., S.A.M. and P.M.V. participated in the workshop; T.A.M., P.M.V., G.G., M.I.M., E.E.E., T.F.C.M. and S.A.M. drafted the manuscript; F.S.C., N.J.C., D.B.G., L.A.H., D.J.H., E.M.R., L.R.C., A.C., J.H.C., A.P.R., A.E.G., A.K., L.K., E.M., C.N.R., M.S., D.V., A.S.W., M.B., A.G.C. and J.L.H. critically reviewed and revised the manuscript for content.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Correspondence should be addressed to T.A.M. (manoliot@mail.nih.gov).