Dissecting the genetics of complex traits using summary association statistics

Bogdan Pasaniuc[1] *and Alkes L. Price*[2,3]

Abstract | During the past decade, genome-wide association studies (GWAS) have been used to successfully identify tens of thousands of genetic variants associated with complex traits and diseases. These studies have produced extensive repositories of genetic variation and trait measurements across large numbers of individuals, providing tremendous opportunities for further analyses. However, privacy concerns and other logistical considerations often limit access to individual-level genetic data, motivating the development of methods that analyse summary association statistics. Here, we review recent progress on statistical methods that leverage summary association data to gain insights into the genetic basis of complex traits and diseases.

Individual-level data
Genome-wide single nucleotide polymorphism genotypes and trait values for each individual included in a genome-wide association study.

Summary association statistics
Estimated effect sizes and their standard errors for each single nucleotide polymorphism analysed in a genome-wide association study.

[1]*Departments of Human Genetics, and Pathology and Laboratory Medicine, University of California, Los Angeles, California 90095, USA.*
[2]*Departments of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA.*
[3]*Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142, USA.*
pasaniuc@ucla.edu; aprice@hsph.harvard.edu

doi:10.1038/nrg.2016.142
Published online 14 Nov 2016

Genome-wide association studies (GWAS) have been broadly successful in identifying genetic variants associated with complex traits and diseases, explaining a significant fraction of narrow-sense heritability and occasionally pinpointing biological mechanisms[1]. These studies have produced extensive databases of genetic variation (typically at the level of common single nucleotide polymorphisms (SNPs) included on genotyping arrays) in large numbers of individuals across hundreds of complex traits. Further analyses of these data can yield important insights into the genetics of complex traits, but privacy concerns and other logistical considerations often restrict access to individual-level data. Nevertheless, summary association statistics are often readily available and can be used to compute *z*-scores (FIG. 1). Here, we define summary association statistics as per-allele SNP effect sizes (log odds ratios for case–control traits) together with their standard errors, although we note that some applications may also require allele frequencies. A list of selected publicly available summary association statistics from large GWAS is provided in TABLE 1. Analyses of summary statistics also offer advantages in computational cost, which does not scale with the number of individuals in the study. These advantages have motivated the recent development of many new methods for analysing summary association data, often in conjunction with linkage disequilibrium (LD) information from a population reference panel such as 1000 Genomes[2].

Here, we review these summary statistic-based methods. First, we review methods for performing single-variant association tests, including meta-analysis, conditional association and imputation using summary statistics. Second, we review methods for performing gene-based association tests by incorporating transcriptome reference data or aggregating signals across multiple rare variants. Third, we review methods for fine-mapping causal variants, including the integration of functional annotation and/or trans-ethnic data. Fourth, we review methods for constructing polygenic predictions of disease risk and inferring polygenic architectures. Finally, we review methods for jointly analysing multiple traits. We conclude with a discussion of research areas for which further work on summary statistic-based methods is needed.

Single-variant association tests

Meta-analysis using fixed-effects or random-effects models. Large consortia often combine multiple GWAS into a single aggregate analysis to boost power for discovering SNP associations with small effects. Studies are combined either by jointly analysing summary association results from each study (meta-analysis) or by re-analysing individual-level data across all studies (mega-analysis)[3]. It has been shown that a meta-analysis attains similar power for association as a mega-analysis, with fewer privacy constraints and logistical challenges (because only summary association data are shared across studies)[4]. A meta-analysis is usually performed using fixed-effects approaches, which assume that true effect sizes are the same across studies. Under the assumption that causal effect sizes may differ across studies, this heterogeneity can be explicitly modelled using random-effects methods. These methods include

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

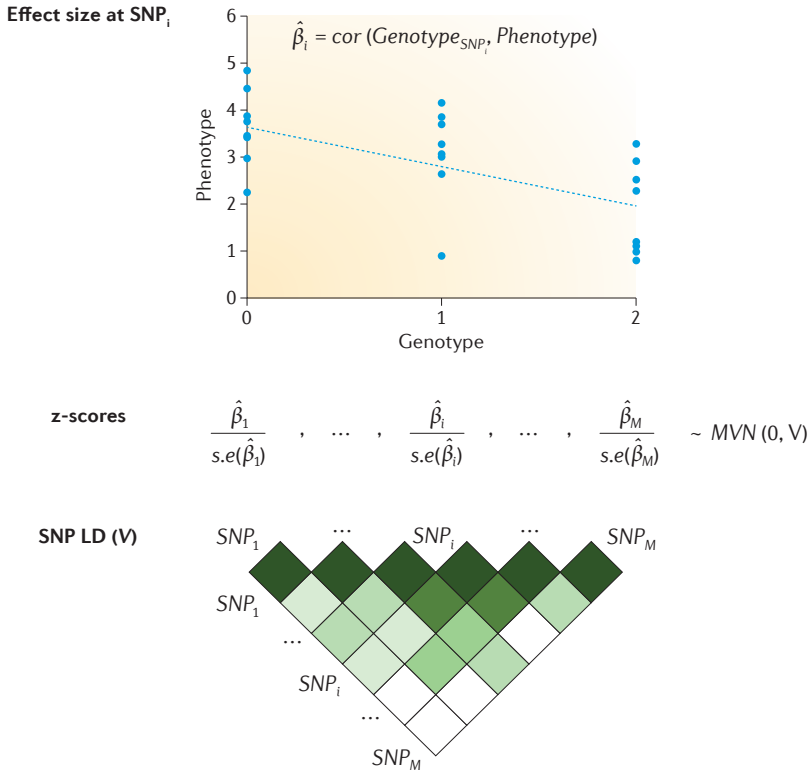


Figure 1 | **Illustration of summary association statistics.** Per-allele single nucleotide polymorphism (SNP) effect sizes (and their standard error (s.e.) values) are typically estimated by regressing the phenotype on the genotype values at the SNP of interest (top). At large sample sizes, the vector of *z*-scores (effect sizes divided by their standard errors) at a locus is approximated by a multivariate normal distribution with mean 0 and variance equal to the linkage disequilibrium (LD) matrix (bottom). MVN, multivariate normal.

an extra variance term in the model to account for heterogeneity. Traditional random-effects methods allow for heterogeneity under the null model, leading to low power even when heterogeneity is present. This limitation of traditional random-effects methods motivated the development of a random-effects method based on a null model of no heterogeneity[5], which has increased power over traditional random-effects methods. Under this framework, a statistical test against a null model of no heterogeneity can be viewed as a summation of a fixed-effect component and a heterogeneity component, thus connecting fixed-effects and random-effects meta-analyses[5]. Subsequent work has introduced the concept of posterior probability for each study to have an effect; this concept can aid interpretation and power under the assumption that a subset of studies have a negligible effect on the trait[6].

Conditional association using LD reference data. Conditional association, in which the association between a SNP and a trait is evaluated after conditioning on the top SNP at a locus, can be used to identify multiple signals of association at a previously identified GWAS locus. Conditional association methods have traditionally required individual-level data to jointly fit multiple SNPs. Recent work has shown that conditional and joint association analyses of multiple SNPs can be approximated using only summary association statistics together with LD information estimated from a population reference panel such as 1000 Genomes[7] (BOX 1). This finding has enabled the discovery of new secondary associations at known loci for height, body mass index and other complex traits and diseases, thus increasing the variance explained by GWAS associations for these traits[8–10]. For example, in a recent GWAS of height, an approximate conditional analysis using summary association statistics data identified 697 genome-wide significant associations, including 34 secondary associations with r^2 >0.1 to a more significant SNP at the same locus (see supplementary table 1 of REF. 8).

Imputation using summary association statistics. A standard approach to boost association power in GWAS is to leverage LD information from a population reference panel to impute genotypes at variants not typed in the study[11]. Imputation is traditionally performed using individual-level data, which requires substantial computational resources. Moreover, imputation can be logistically cumbersome when new reference panels become available, particularly for large consortia combining data from multiple studies. As an alternative to imputation using individual-level data, approaches have been developed to perform imputation directly at the level of summary statistics[12–18] (providing an alternative to other multivariate tests[19,20]). The key insight of these approaches is that LD induces correlations between *z*-scores, which can be modelled using a multivariate normal (MVN) distribution with the variance equal to the LD correlation matrix[21]. Note that an adjustment in the LD computation is needed for *z*-scores estimated using mixed models[22]. Thus, *z*-scores at untyped SNPs can be imputed from observations at typed SNPs using conditional means and variances of the MVN distribution. Imputation using summary statistics recovers >80% of the information from imputation using individual-level data at common variants[14–16]. This approach is also practical and efficient because the imputed summary statistics are linear combinations of the observed statistics (BOX 1). However, imputation using summary statistics cannot capture nonlinear relationships between SNPs, which are modelled using haplotypes in imputation from individual-level data.

Conditional association and imputation using summary statistics crucially rely on accurate LD information from a population reference panel. Even in the best case, when the reference population closely matches the GWAS population, the relatively small size of reference panels for which LD information is publicly available (typically hundreds or at most thousands of individuals) makes accurate estimation of a large number of LD parameters a challenge. This motivated approaches for the regularization of the estimated LD matrix, both to maximize accuracy and to ensure robustness in the case of imputation using summary statistics, particularly as mis-estimation of the variance of imputed statistics can lead to false-positive associations. A simple approach to regularization is to set all correlations

z-scores
Association statistics that follow a standard normal distribution under the null model; often computed as per-allele effect sizes divided by their standard errors.

Meta-analysis
A method for combining data from different studies in which summary association statistics from each study are jointly analysed.

Mega-analysis
A method for combining data from different studies in which individual-level data from each study are merged and jointly analysed.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Table 1 | **Publicly available summary association statistics***

Trait	*N*	URL	Ref.
Age at menarche	127,884	http://www.reprogen.org/	119
Alzheimer disease	54,162	http://www.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php	120
Bone mineral density	53,236	http://www.gefos.org/?q=content/data-release-2015	121
Body mass index	122,033	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	122
Body mass index‡	322,154	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	9
Coronary artery disease	77,210	http://www.cardiogramplusc4d.org/	123
Crohn's disease	20,883	http://www.ibdgenetics.org/downloads.html	124
Crohn's disease‡	51,874	http://www.ibdgenetics.org/downloads.html	125
Depressive symptoms	161,460	http://www.thessgac.org/data	126
Ever smoked	74,035	http://www.med.unc.edu/pgc/downloads/	127
Fasting glucose	58,074	http://www.magicinvestigators.org/downloads/	128
HbA_{1c} (glycated haemoglobin)	46,368	http://www.magicinvestigators.org/downloads/	129
High-density lipoprotein	97,749	http://www.broadinstitute.org/mpg/pubs/lipids2010/	130
High-density lipoprotein‡	188,577	http://csg.sph.umich.edu//abecasis/public/lipids2013/	131
Height	131,547	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	132
Height‡	253,288	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	8
Hip circumference	213,038	http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	133
Irritable bowel syndrome (Crohn's disease or ulcerative colitis)	34,652	http://www.ibdgenetics.org/downloads.html	124
Irritable bowel syndrome (Crohn's disease or ulcerative colitis)‡	65,643	http://www.ibdgenetics.org/downloads.html	125
Low-density lipoprotein	93,354	http://www.broadinstitute.org/mpg/pubs/lipids2010/	130
Low-density lipoprotein‡	188,577	http://csg.sph.umich.edu//abecasis/public/lipids2013/	131
Neuroticism	170,911	http://www.thessgac.org/data	126
Rheumatoid arthritis (Europeans)	38,242	http://plaza.umin.ac.jp/~yokada/datasource/software.htm	134
Rheumatoid arthritis (Europeans) ‡	58,284	http://plaza.umin.ac.jp/~yokada/datasource/software.htm	134
Rheumatoid arthritis (East Asians)	22,515	http://plaza.umin.ac.jp/~yokada/datasource/software.htm	134
Schizophrenia	70,100	http://www.med.unc.edu/pgc/downloads/	135
Subjective well-being	298,420	http://www.thessgac.org/data	126
Triglycerides	94,461	http://www.broadinstitute.org/mpg/pubs/lipids2010/	130
Triglycerides‡	188,577	http://csg.sph.umich.edu//abecasis/public/lipids2013/	131
Type 2 diabetes	60,786	http://diagram-consortium.org/	136
Ulcerative colitis	27,432	http://www.ibdgenetics.org/downloads.html	124
Ulcerative colitis‡	47,746	http://www.ibdgenetics.org/downloads.html	125
Waist circumference	232,101	http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	133
Waist/hip ratio	212,248	http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	133
Years of education	328,917	http://www.thessgac.org/data	126

*We provide a selected list of publicly available summary statistics from genome-wide association studies with sample sizes of at least 20,000. A more complete list can be found in REF. 137. ‡Includes specialty genotyping array data; not suitable for analysis using linkage disequilibrium score regression and its extensions.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Box 1 | **Conditional association and imputation from summary statistics**

Let X be an $N \times M$ matrix of genotypes, standardized to mean 0 and unit variance, and Y be an $N \times 1$ vector of standardized trait values, where M is the number of single nucleotide polymorphisms at the locus and N is the number of samples. Under a standard linear model, $Y = X\beta + \epsilon$. Let V be an $M \times M$ linkage disequilibrium (LD) matrix of pairwise LD; V is equal to X^TX if individual-level data are available but can otherwise be estimated from a population reference sample (with or without regularization).

Conditional association using LD reference data

We estimate the joint effects of all SNPs using least-squares as $\hat{\beta} = V^{-1}X^TY$ with $var(\hat{\beta}) = \sigma_j^2 V^{-1}$, where σ_j^2 is the residual variance in the joint analysis. However, in a standard genome-wide association study, each SNP is marginally tested one at a time, which can be expressed in matrix form as $\hat{\beta}_M = D^{-1}X^TY$ with $var(\hat{\beta}_M) = \sigma_M^2 D^{-1}$, where D is the (nearly constant) diagonal matrix of V and σ_M^2 is the residual variance in the marginal analysis. It follows that

$$\hat{\beta} = V^{-1} D\hat{\beta}_M$$

$$var(\hat{\beta}) = \sigma_j^2 V^{-1}$$

Summary statistic imputation using LD reference data

Let

$$Z = \frac{\hat{\beta}_M}{s.e.(\hat{\beta}_M)} = \frac{X^TY}{\sqrt{(N)}}$$

be a vector of z-scores (estimated effect sizes divided by their standard errors) obtained by marginally testing each SNP one at a time. Under the null hypothesis of no association, $Z \sim N(0, V)$. Let Z_t and Z_i partition the vector Z into T typed SNPs and $M - T$ untyped SNPs, and let $V_{t,t}$ (covariances among typed SNPs), $V_{i,i}$ (covariances among untyped SNPs), and $V_{t,i}$ (covariances among typed and untyped SNPs) partition the matrix accordingly. It follows that $Z_i|Z_t \sim N(V_{i,t}V_{t,t}^{-1}Z_t, V_{i,i} - V_{i,t}V_{t,t}^{-1}V_{i,t}^T)$. The mean and variance of the conditional distribution can be used to impute summary association statistics at untyped SNPs.

between distal SNPs to zero based on a fixed distance threshold[7] or on approximately independent LD blocks inferred from the data[23]. An alternative is to specify a prior distribution and to compute Bayesian posteriors[12]; data can be combined across multiple ancestry reference panels to further boost accuracy[17,18]. Singular value decomposition-based approaches for LD regularization have also been proposed in other contexts[10]. In general, the accuracy of conditional association and imputation using summary statistics is reduced for low-frequency variants and when the LD structure between typed and imputed SNPs is mis-specified (for example, when the ancestry of the GWAS sample does not exactly match the reference panel). We note that concerns about false-positive associations in imputation using summary statistics can be avoided entirely via the release of in-sample summary LD information.

Gene-based association tests

Gene-based association using transcriptome reference data. GWAS risk variants are significantly enriched for genetic variants that affect gene expression, that is, expression quantitative trait loci (eQTLs)[24]. This motivates the paradigm of transcriptome-wide association studies (TWAS), which evaluate the association between the expression of each gene and a complex trait of interest. Owing to the limited availability of very large samples with measured gene expression and trait values, initial TWAS approaches integrated eQTL and GWAS data to identify susceptibility genes either by matching the association signals[25–27], by mediation analyses[28] or by assessing whether the same causal variant affects both gene expression and trait under a single causal variant model[29–31].

More recent studies have leveraged predicted expression to improve the power of TWAS. Under this paradigm, transcriptome reference data are used to predict gene expression in the GWAS dataset (for example, using *cis* SNPs within 1 Mb of the transcription start site), followed by a test for association between the predicted expression and trait. As an alternative to TWAS using individual-level data[32], TWAS using predicted expression can also be performed using only summary association statistics and summary LD information[33–35]. These studies used expression predictors that do not account for LD[33] or account for LD and allow for sparsity in eQTL effect sizes[34], or used the top eQTL at the locus[35]. The key intuition is that the correlation between a weighted linear combination of SNPs (that is, predicted gene expression) and a trait is equivalent to a weighted linear combination of correlations between SNPs and a trait (that is, summary association statistics from GWAS) (FIG. 2). Because TWAS using predicted expression is conceptually similar to a test for non-zero genetic covariance between gene expression and a trait[34], it can also be performed via a two-sample Mendelian randomization from summary statistics[35]. TWAS using predicted expression can increase power over a standard GWAS when there exist multiple causal variants whose effect on a trait is mediated through expression. TWAS also reduce the multiple hypothesis burden by testing tens of thousands of genes instead of millions of SNPs. TWAS using predicted expression typically use individual-level transcriptome reference data to predict gene expression, but can also be performed using only summary association statistics between SNPs and gene expression, albeit with a reduction in power[34]. The potential power gains of TWAS are underscored by the recent identification of 71 new susceptibility genes across 28 complex traits, of which 17 have no GWAS association within 1 Mb (REF. 36). However, TWAS are underpowered compared to standard GWAS when the true biological mechanism is independent of gene expression or when expression data in the most relevant tissue are not available.

Rare variant association tests. Although most GWAS of complex traits and diseases have focused on common variants that are typed on genotyping arrays or imputed from population reference panels, rare variant associations may also provide a rich source of biological insights, particularly for traits under strong negative selection[37,38]. Because association tests of individual rare variants are likely to be underpowered, rare variant association tests generally aggregate evidence for association across multiple rare variants at a locus. In exome sequencing studies (or exome array studies), rare variants are aggregated at the gene level, making the gene the unit of association. This process can be performed using either burden tests or overdispersion tests, although hybrid omnibus tests are also possible[39]. Recent studies

Summary LD information
(summary linkage disequilibrium information). In-sample correlations between each pair of typed single nucleotide polymorphisms analysed in a genome-wide association study; can be restricted to proximal pairs of typed SNPs to limit the number of pairs of SNPs.

Transcriptome-wide association studies
(TWAS). Studies that evaluate the association between the expression of each gene and a trait of interest; predicted expression may be used instead of measured expression to improve practicality.

Mendelian randomization
A method that uses significantly associated single nucleotide polymorphisms as instrumental variables to quantify causal relationships between two traits.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

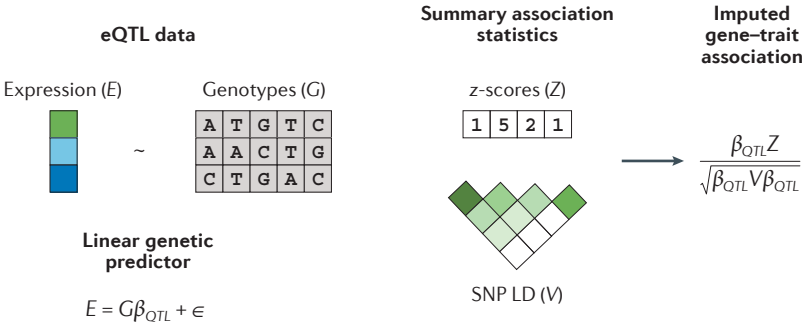


Figure 2 | **TWAS using predicted expression and summary data.** Transcriptome-wide association studies (TWAS) using predicted expression and summary data follow two steps. First, transcriptome reference data are used to build a linear predictor for gene expression, typically using single nucleotide polymorphisms (SNPs) from the 1 Mb local region around the gene with regularized effect sizes (for example, using a Bayesian sparse linear mixed model[81]). Second, this predictor is applied to summary genome-wide association *z*-scores, and gene–trait association *z*-scores are computed, testing the null model of no association between a gene and a trait. eQTL, expression quantitative trait loci; LD, linkage disequilibrium.

have shown that both burden tests and overdispersion tests can be performed using only summary association statistics from each rare variant, together with summary LD information[40–42] (BOX 2). Briefly, burden tests are computed as weighted sums of single-variant *z*-scores whereas overdispersion tests are computed as weighted sums of squared single-variant *z*-scores (analogous to previous work on common variant overdispersion tests using summary statistics[19,20]), with summary LD information used to specify appropriate null distributions in each case. However, a key limitation of these studies is that they require the use of in-sample summary LD information in preference to reference LD information to ensure appropriate null distributions and to avoid false-positive associations. Thus, in contrast to summary statistic-based methods for common variants (see above), both summary association statistics and in-sample summary LD information are required for these methods to be useful (see Conclusions). An additional limitation is that for case–control traits, asymptotic null distributions may not be valid when variant counts or case or control sample sizes are small, necessitating careful scrutiny of quantile–quantile plots.

Fine-mapping

Fine-mapping using posterior probabilities of causality. Statistical fine-mapping aims to identify the causal variant (or variants) that is driving a GWAS association signal, enabling functional experiments to validate biological function. A straightforward approach to fine-mapping is to prioritize variants based on the strength of the marginal association statistics (that is, ranking *P* values)[43]. This strategy is effective in the case of a single causal variant but can be suboptimal when multiple causal variants are present because the SNP with the top *P* value at the locus may be tagging multiple causal variants. An alternative is to compute the posterior probability of causality for every SNP in the region based on the likelihoods of the observed *z*-scores conditional on each possible set of causal variant (or variants)[44]. These posterior probabilities can be used to construct a credible set of SNPs, defined as the smallest set of SNPs that contains the true causal variant with a given probability (typically 90% or 99%). Initial studies approximated the posterior probabilities of causality under a single causal variant assumption. Under this assumption, posterior probabilities of causality can be estimated from *z*-scores without the need for LD information[45]; this approach is both practical and computationally efficient but suboptimal when multiple causal variants are present.

More recent studies have computed posterior probabilities of causality under a multiple causal variant assumption[46]. As in the case of imputation using summary statistics, the likelihoods of the observed *z*-scores can be computed based on the MVN distribution with variance equal to the LD correlation matrix, with LD estimated from population reference panels using regularization techniques. In contrast to imputation using summary statistics, which uses the null model of no association (that is, a mean of 0 in the MVN), in fine-mapping the mean is a function of causal effect sizes, which can be heuristically approximated or integrated out using conjugate priors[46,47]. These methods often restrict computations to a maximum number of causal variants (for example, three or six); more recent studies have shown that further efficiencies can be achieved through matrix factorizations[48] or stochastic search[49]. Methods that model multiple causal variants generally improve the accuracy (and calibration) of credible sets at loci with multiple causal variants[46–50], with very limited reductions in accuracy at loci with only a single causal variant[46–52]. A less accurate alternative is to use conditional association analysis to detect multiple signals of association[7,53,54] followed by an estimation of posterior probabilities of causality under a single causal variant assumption for each independent signal. In this case, special care is required in specifying the boundaries of each independent signal and the threshold for the conditional test.

Leveraging functional annotation data. Fine-mapping accuracy can be improved by integrating functional annotation data such as predicted regulatory elements obtained from the US National Human Genome Research Institute's ENCODE (Encyclopedia of DNA Elements) project[55] and the US National Institutes of Health's Roadmap Epigenomics Program[56]. This approach is motivated by early studies showing that disease-associated variants are systematically enriched in chromatin marks that delineate active regulatory regions in disease-relevant cell types[57,58]. Under this paradigm, a statistical model is developed to jointly estimate functional enrichment and update posterior probabilities of causality using functional annotations[47,52,59,60]. Some integrative methods assume that SNPs are unlinked[60] or assume a single causal variant per locus[52,59], with a recent study using a multiple causal variant model[46] to incorporate functional annotation data[47]. In an analysis of

Burden tests
Gene-based rare variant tests in which all rare variants in a gene are assumed to have the same direction of effect.

Overdispersion tests
Gene-based rare variant tests in which rare variants in a gene are assumed to impact trait in either direction.

Posterior probability of causality
The inferred probability that a single nucleotide polymorphism is causal based on association data and optional prior information.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Polygenic risk scores
A method of predicting trait by summing the predicted marginal effects of all markers below a *P* value threshold in a training sample multiplied by marker genotypes in a validation sample.

rheumatoid arthritis summary association data, integrative fine-mapping using this approach reduced the average size of 90% credible sets by 10%[61].

In addition to increasing fine-mapping accuracy, these studies have also provided insights into polygenic architectures (see below) by identifying tissue-specific functional annotations that are enriched for causal disease signals. This result can also be achieved by conducting fine-mapping without integrating functional annotation data (typically under a single causal variant assumption) and then overlapping the resulting credible sets with functional annotation data to assess enrichment[62–64]. Future integrative methods could increase fine-mapping resolution by integrating probabilistic functional annotations (for example, peak intensities of ChIP–seq (chromatin immunoprecipitation followed by sequencing) results) or modelling the strength of association between SNPs and chromatin marks in population-based studies[65,66].

Trans-ethnic fine-mapping. Fine-mapping accuracy can also be improved by leveraging differences in LD patterns across continental populations that have arisen due to differences in demographic events such as population bottlenecks[67–70] (FIG. 3). Intuitively, the set of tag SNPs linked to a causal variant will vary across populations; thus, aggregating evidence of association across populations will dilute signals from tag SNPs and strengthen signals from causal variants. A standard approach to combining information across multiple studies is to compute posterior probabilities of causality from fixed-effects meta-analysis results[67,69,71,72]. Alternatively, posterior probabilities can be computed from the results of random-effects trans-ethnic meta-analysis methods[64,68]. These approaches assume a single causal variant and thus do not require LD information from the underlying populations. More recent studies have introduced hierarchical probabilistic models that allow for multiple causal variants while incorporating LD information from population reference panels[61]. These studies assume that causal variants are shared across populations but allow for heterogeneity in effect sizes across populations and can also incorporate functional annotation data to further increase fine-mapping accuracy[61]. In an analysis of rheumatoid arthritis summary association data in Europeans and Asians (see above), trans-ethnic fine-mapping reduced the average size of 90% credible sets by 25% and by 32% when also integrating functional annotation data[61].

Box 2 | **Rare variant association tests using summary association statistics**

Let X be an $N \times M$ matrix of genotypes, standardized to mean 0 and variance 1, and Y be an $N \times 1$ matrix of standardized trait values, where M is the number of rare variants (for example, in a given gene being tested for association) and N is the number of samples. An $M \times 1$ vector of z-scores (estimated effect sizes divided by their standard errors) can be computed as

$$Z = \frac{X^T Y}{\sqrt{N}}$$

with multivariate normal null distribution $Z \sim N(0, V)$, where V is an in-sample linkage disequilibrium matrix.

Burden tests

Burden tests assume that all rare variants in a candidate gene have the same direction of effect. Burden tests may either assume that standardized effect sizes are the same for each rare variant[112] (that is, per-allele effect sizes are proportional to

$$\frac{1}{\sqrt{p_i(1-p_i)}}$$

where p_i is the allele frequency), or apply weights or thresholds based on allele frequency or functional information[113,114]. If w is an $M \times 1$ vector of weights for each rare variant (including zero weights for rare variants excluded by a threshold), the test statistic for a weighted burden test is $T_{burden} = w^T Z$ with null distribution $T_{burden} \sim N(0, w^T V w)$. This test statistic can naturally be extended to a meta-analysis of burden tests from multiple cohorts (via inverse-variance weighting), and can be extended to variable threshold tests and binary traits[40–42].

Overdispersion tests

Overdispersion tests assume that rare variants in a candidate gene can affect a complex trait in either direction, and can be computed as weighted sums of squared single-variant test statistics[115,116]. If $W = \mathrm{diag}(w_1, \ldots, w_M)$ is an $M \times M$ diagonal matrix of weights for each rare variant, the test statistic for a weighted overdispersion test is $T_{overdispersion} = Z^T W Z$ with null distribution $T_{burden} \sim \Sigma_i \mu_i \chi_i^2$, where weights μ_i for each χ^2 (1 d.f.) distribution χ_i^2 are given by eigenvalues of the matrix $V^{1/2} W V^{1/2}$. This test statistic can be extended to a meta-analysis of overdispersion tests from multiple cohorts (via inverse-variance weighting), and can be extended to binary traits[40–42].

Polygenicity of complex traits

Polygenic risk prediction. Although the main focus of complex disease genetics is to gain insights into disease biology, genetics can also be leveraged to build predictions of disease risk, which may become clinically useful as sample sizes increase[73,74]. A landmark study of schizophrenia showed that polygenic risk scores produced predictions of schizophrenia risk in validation samples that were significantly better than random, and far more accurate than those based on the single genome-wide significant locus identified in the study[75]. This study provided an early demonstration of the advantages of incorporating markers that do not attain genome-wide significance into polygenic risk scores to improve the prediction accuracy for polygenic traits. An important issue in computing polygenic risk scores is that of LD between markers, which has historically been addressed by LD pruning — either without regard for *P* values[75] or via informed LD pruning[76] (also known as clumping), which preferentially retains markers with more significant *P* values. More recent work has shown that explicitly modelling LD using an LD reference panel and estimating posterior mean causal effect sizes can improve prediction accuracy from summary statistics[77].

An alternative to summary statistic-based methods is to fit effect sizes of all markers simultaneously using best linear unbiased prediction (BLUP) methods and their extensions[78–80], which require individual-level training data. Fitting all markers simultaneously is theoretically more appropriate and can produce more accurate predictions, although the relative advantage is small when overall prediction accuracies are modest (BOX 3).

In their simplest form, polygenic risk scores and BLUP methods assume infinitesimal (Gaussian) architectures in which all markers are causal. However, these methods have been extended to increase prediction accuracy in the case of non-infinitesimal architectures.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

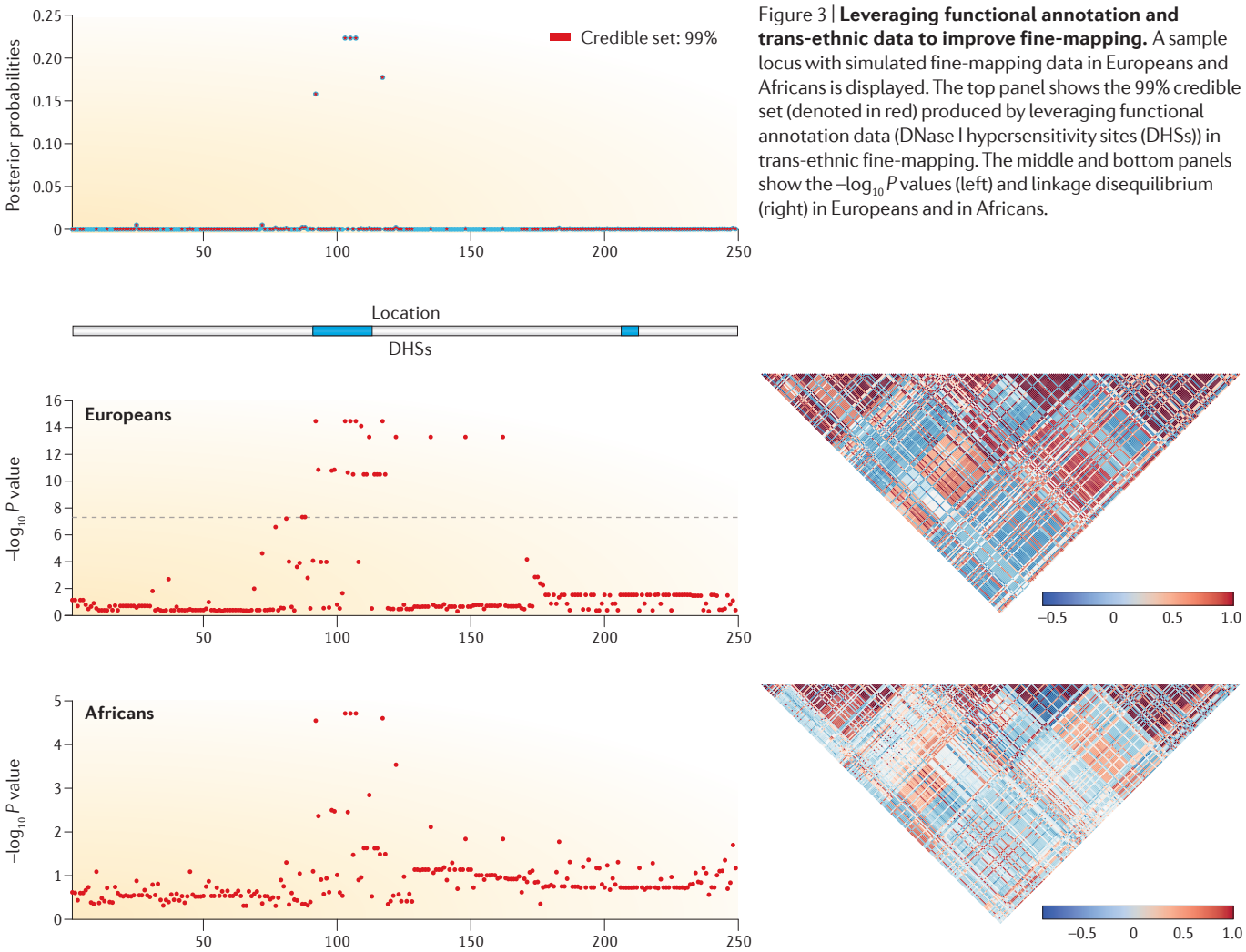


Figure 3 | **Leveraging functional annotation and trans-ethnic data to improve fine-mapping.** A sample locus with simulated fine-mapping data in Europeans and Africans is displayed. The top panel shows the 99% credible set (denoted in red) produced by leveraging functional annotation data (DNase I hypersensitivity sites (DHSs)) in trans-ethnic fine-mapping. The middle and bottom panels show the $-\log_{10} P$ values (left) and linkage disequilibrium (right) in Europeans and in Africans.

Improvement of polygenic risk scores has been accomplished by restricting markers to those below a *P* value threshold[75] or estimating posterior mean causal effect sizes under a point-normal prior[77]. Increased prediction accuracy for BLUP methods has been achieved by estimating (joint-fit) posterior mean causal effect sizes under a normal mixture prior[81,82]. Although polygenic risk scores must await even larger training sample sizes to attain clinical utility, appreciable prediction accuracies have been achieved for some traits, including a Nagelkerke coefficient of determination (R^2) of 0.25 (area under the curve: 75%) for schizophrenia[77]. A crucial caveat is to avoid non-independence of training and validation samples (for example, due to cryptic relatedness or shared population stratification) when constructing and evaluating polygenic risk scores. Non-independence of training and validation samples could cause prediction accuracy to be overstated relative to what could be achieved in an independent validation sample[77,83].

Inferring polygenic architectures. It is becoming increasingly clear that most complex traits and diseases have highly polygenic architectures, with a large number of causal variants with small effects. To understand these polygenic architectures, it is of interest to infer parameters such as the heritability explained by SNPs and the number of variants with non-negligible effects on the trait. Both of these quantities have been estimated using accuracies of polygenic risk scores (see above), as a function of the *P* value threshold that is used to constrain the set of employed markers[75,76]. Computing polygenic risk scores requires individual-level data in the validation cohort, implying that these methods are not strictly summary statistic based. Recent work has shown that the information in polygenic risk scores can be derived from summary-level data in the training and validation cohorts to estimate the heritability explained by SNPs and the number of causal variants[84]. A limitation of this approach is that SNPs are assumed to be uncorrelated, which can be approximately achieved by LD pruning but precludes analyses of dense marker panels. The heritability explained by SNPs can alternatively be estimated from the slope of LD score regression[85], leveraging the fact that SNPs with higher LD scores are expected to contain more polygenic signals[86]. This approach explicitly allows for LD

LD score regression
A method of assessing trait polygenicity by regressing χ^2 association statistics against linkage disequilibrium (LD) scores for each single nucleotide polymorphism (SNP), computed as sums of squared correlations of each SNP with all SNPs including itself.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

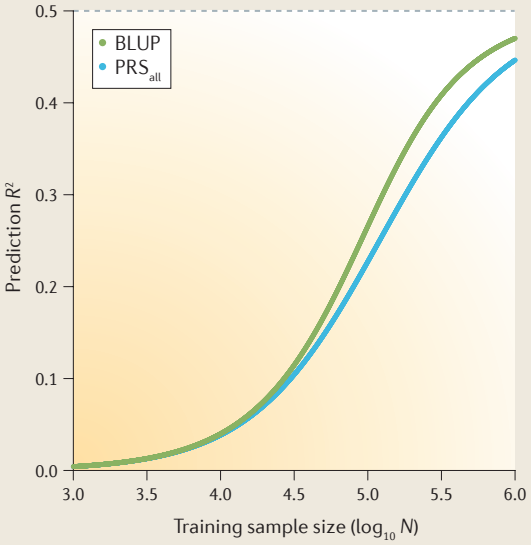
Box 3 | **Polygenic risk prediction using summary versus individual-level data**

Suppose that polygenic risk prediction for a quantitative trait is conducted using a training cohort with N unrelated samples, using M unlinked markers with single nucleotide polymorphism (SNP) heritability[7] equal to h_g^2. We initially consider two polygenic risk prediction methods that assume infinitesimal (Gaussian) architectures: polygenic risk scores computed using marginal effects at all markers with no P value thresholding (PRS_{all}), and fitting effect sizes of all markers simultaneously via best linear unbiased prediction (BLUP). We note that PRS_{all} requires only summary statistics from the training cohort, whereas BLUP requires individual-level data. Prediction accuracy (coefficient of determination; R^2) for each method is given by[83,117]

$$R^2_{PRS_{all}} = \frac{h_g^2}{1 + \frac{M}{Nh_g^2}}$$

$$R^2_{BLUP} = \frac{h_g^2}{1 + \frac{M}{Nh_g^2}(1 - R^2_{BLUP})}$$

These equations can naturally be extended to linked markers (using the effective number of unlinked markers[108]) and case–control traits (using observed-scale SNP heritability[118]). The relative advantage of BLUP over PRS_{all} is small when prediction R^2 is small in absolute terms, but grows larger when prediction R^2 is larger. This effect is illustrated in the figure, which shows prediction R^2 at various training sample sizes based on $M = 60{,}000$ unlinked markers and a SNP heritability of $h_g^2 = 0.5$. These results generalize to non-infinitesimal extensions of polygenic risk scores[75,77] and BLUP[81,82]; in the latter case, the noise reduction from fitting all markers simultaneously remains equal to $1 - R^2$, corresponding to an increase in training sample size of $1/(1 - R^2)$.



between SNPs and can distinguish between polygenicity and confounding. However, the approach assumes a linear model that may not hold in practice and makes strong assumptions about effect sizes of rare variants that only enable robust estimates for common variants. Another recent method models LD while treating SNP effects as fixed rather than random (similar to the method reported in REF. 84), enabling estimation of heritability explained by common SNPs in local regions as well as genome-wide[10]. Overall, summary statistic-based methods provide a useful alternative to methods for estimating heritability explained by SNPs from individual-level data using restricted maximum likelihood (REML) and its extensions[87,88].

The increasing availability of functional annotation data (see above) can also be used to identify functional annotations that are enriched for polygenic signals of disease heritability. A recent study accomplished this goal using a Bayesian hierarchical model that splits the genome into blocks and incorporates functional annotations both coarse-scale at the level of blocks and fine-scale at the level of SNPs[59]. This study was the first to quantify polygenic enrichments for cell type-specific chromatin marks and DNase I hypersensitivity sites across a broad set of complex traits and diseases. For example, polygenic signals for platelet volume and platelet count were enriched at DNase I hypersensitivity sites in $CD34^+$ cells (which are on the cell lineage that lead to platelets), and polygenic signals for Crohn's disease were depleted at repressed chromatin in lymphoblastoid cell lines. Functional enrichments can alternatively be estimated by stratified LD score regression[89], which generalizes LD score regression[85] to regress χ^2 statistics for each SNP against LD scores with each functional category. Fine-mapping methods can also estimate functional enrichments, although these analyses are often restricted to disease-associated loci[47,52,61]. Notably, all of these summary statistic-based methods have been applied to a large number of overlapping functional annotations, whereas methods that analyse individual-level genotypes have only been applied to a small number of non-overlapping functional annotations[88,90]. In addition, stratified LD score regression is not limited by the single causal variant per block assumption of the Bayesian hierarchical model, increasing power in settings of highly polygenic traits[89]. Application of stratified LD score regression identified significant cell type-specific enrichments for many highly polygenic traits, including enrichments for histone marks in the brain for smoking behaviour and educational attainment[89], even though the summary statistics analysed contained only one and three genome-wide significant loci for smoking behaviour and educational attainment, respectively. One limitation of stratified LD score regression is its limited power for functional categories spanning a small percentage of the genome; thus, additional work in this area is required. As both summary statistic and functional annotation datasets grow larger and richer, identifying enriched functional annotations using summary statistic data will likely continue to be a fruitful endeavour.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Cross-trait analyses

Many complex traits and diseases have a shared genetic aetiology, which can be either via a shared genetic variant (or variants) with non-zero effect sizes (pleiotropy) or via a correlation between causal effect sizes (genetic correlation). Indeed, many instances of genetic variants with pleiotropic effects on multiple traits have been identified[91–96]. A recent study applied a Bayesian framework to summary association statistics from pairs of traits to estimate, at each locus in the genome, the probability that an associated variant has pleiotropic effects on both traits[97]. Pleiotropic SNPs can also be used as instrumental variables in Mendelian randomization analyses from summary statistics[98–100], with one such analysis showing that increased body mass index causally increases triglyceride levels[97].

An alternative approach to assessing the genetic overlap between two traits is to estimate the correlation between causal effect sizes across the two traits. Genome-wide genetic correlations can be estimated from individual-level data using bivariate REML[101]. A recent study estimated genome-wide genetic correlations from summary data using the information in polygenic risk scores[84]. However, this approach required LD pruning of the data, which may lead to upwards bias[84]. Another recent study estimated genome-wide genetic correlations from summary data using cross-trait LD score regression[102], which generalizes LD score regression to regress products of *z*-scores against LD scores for each SNP. This method produced estimates that were highly concordant with those from individual-level data[101]. Fitting the underlying MVN model using maximum likelihood instead of linear regression has produced promising results in applications to estimate cross-trait and cross-population genetic correlations, and may also prove useful in other settings[103]. Although genetic correlation analyses restricted to associated variants have also produced important findings[97], the power of methods that leverage polygenic signals in genome-wide data is underscored by the discovery of significant genetic correlations involving traits with zero or few genome-wide significant loci, including a significant negative genetic correlation between smoking behaviour and educational attainment[102]. Notably, recent work has shown that association statistics for unmeasured traits can be computed using summary statistics from genetically correlated traits[104,105].

Conclusions

Recently developed methods have made it possible to leverage summary association statistics to perform a wide range of analyses, many of which previously required individual-level data. As the availability of summary association statistics continues to grow (TABLE 1), summary statistics will continue to be broadly used in analyses involving single-variant association tests, gene-based association tests, fine-mapping, polygenic prediction and inferring polygenic architectures, and cross-trait analyses. The use of summary data will entail a loss of accuracy in some applications such as imputation and polygenic risk prediction. For imputation, methods that analyse individual-level data can use haplotypes to model nonlinear structure, whereas for polygenic risk prediction, methods that analyse individual-level data can reduce noise by fitting all markers simultaneously. However, when summary statistics are available in larger sample size than individual-level data, the advantage of larger sample size will outweigh those limitations. In addition, there are some settings when summary statistic-based methods are the method of choice even when individual-level data are available, such as identifying functional annotations that are enriched for heritability, for which methods that analyse individual-level data cannot currently handle a large number of overlapping annotations.

Despite considerable recent progress, there are some areas in which further research on summary statistic-based methods is needed. As population reference panels grow, more accurate modelling of rare and low-frequency variants will become possible, and it will be important to assess the limits of such efforts. It is also of interest to develop methods for inferring polygenic architectures from summary statistics that permit different relationships between allele frequency and effect size. Identifying functional annotations that are enriched for heritability is an application that is particularly likely to produce important biological insights; however, for such applications there is a need for new methods that are adequately powered for functional categories spanning a small percentage of the genome. As the number of functional annotations continues to increase, the integration of such data poses computational and statistical challenges in disentangling the correct functional annotations among many correlated ones.

We conclude by emphasizing the importance of making summary association statistics publicly available. A 2012 editorial in *Nature Genetics* asked its authors to publish or deposit in databases summary association statistics for all SNPs analysed[106]. This editorial elicited a broad impact on the set of publicly available summary statistics in the years that followed (TABLE 1). The public release of summary statistics is a useful compromise in situations when sample consent restrictions or privacy concerns preclude the release of individual-level data in a public repository. Even though the release of summary statistics can in principle lead to privacy concerns[107], more recent work has shown that such privacy attacks have low power when the summary sample size exceeds the effective number of independent markers (currently estimated at 60,000 in typical GWAS datasets[108]), implying that privacy concerns should not preclude the public release of summary statistics from large studies[109–111]. Indeed, some recent studies have created web portals where summary data can be publicly accessed and visualized[63]. Finally, we note the potential benefits of publicly releasing summary statistics that include summary LD information (that is, correlations) between each pair of proximal SNPs. However, the optimal approach to aggregating summary LD information across multiple cohorts in large-scale meta-analyses remains unclear, motivating the need for future work in this area.

Pleiotropy
The existence of a genetic variant (or variants) that affects more than one trait.

Genetic correlation
The signed correlation across single nucleotide polymorphisms between causal effect sizes for two traits.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

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. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
3. Evangelou, E. & Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
4. Lin, D. Y. & Zeng, D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epidemiol.* **34**, 60–66 (2010).
5. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
This study introduces a powerful new random-effects meta-analysis method that uses a null model of no heterogeneity.
6. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet.* **8**, e1002555 (2012).
7. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
This study demonstrates that conditional association analysis can be performed using summary statistics.
8. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
9. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
10. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
11. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
12. Wen, X. & Stephens, M. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann. Appl. Stat.* **4**, 1158–1182 (2010).
This study is the first to show that Gaussian imputation methods can be applied to summary-level genetic data.
13. Kostem, E., Lozano, J. A. & Eskin, E. Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. *Genetics* **188**, 449–460 (2011).
14. Lee, D., Bigdeli, T. B., Riley, B. P., Fanous, A. H. & Bacanu, S. A. DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* **29**, 2925–2927 (2013).
15. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
16. Xu, Z. *et al.* DISSCO: direct imputation of summary statistics allowing covariates. *Bioinformatics* **31**, 2434–2442 (2015).
17. Lee, D. *et al.* DISTMIX: direct imputation of summary statistics for unmeasured SNPs from mixed ethnicity cohorts. *Bioinformatics* **31**, 3099–3104 (2015).
18. Park, D. S. *et al.* Adapt-Mix: learning local genetic correlation structure improves summary statistics-based analyses. *Bioinformatics* **31**, i181–189 (2015).
19. Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
20. Li, M.-X., Gui, H.-S., Kwan, J. S. H. & Sham, P. C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.* **88**, 283–293 (2011).
21. Conneely, K. N. & Boehnke, M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am. J. Hum. Genet.* **81**, 1158–1168 (2007).
22. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B. & Eskin, E. Identification of causal genes for complex traits. *Bioinformatics* **31**, i206–i213 (2015).
23. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
24. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
25. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
26. Xiong, Q., Ancona, N., Hauser, E. R., Mukherjee, S. & Furey, T. S. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res.* **22**, 386–397 (2012).
27. He, X. *et al.* Sherlock: detecting gene–disease associations by matching patterns of expression QTL and GWAS. *Am. J. Hum. Genet.* **92**, 667–680 (2013).
28. Huang, Y. T., Liang, L., Moffatt, M. F., Cookson, W. O. C. M. & Lin, X. iGWAS: integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. *Genet. Epidemiol.* **39**, 347–356 (2015).
29. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
This study introduces a method for performing TWAS using summary statistics by assessing whether a single causal variant affects both gene expression and trait.
30. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
31. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839–846 (2015).
32. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
33. Lee, D. *et al.* JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics* **31**, 1176–1182 (2015).
34. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
This study identifies 69 new genes associated with obesity-related traits using a powerful new method for performing TWAS using summary statistics by assessing the association between predicted gene expression (using all *cis* SNPs) and trait.
35. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
36. Pavlides, J. M. W. *et al.* Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med.* **8**, 84 (2016).
37. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2011).
38. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
39. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
40. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* **93**, 42–53 (2013).
This study is the first of three studies to demonstrate that rare variant burden and overdispersion tests can be performed using summary statistics.
41. Hu, Y.-J. *et al.* Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *Am. J. Hum. Genet.* **93**, 236–248 (2013).
42. Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
43. Faye, L. L., Machiela, M. J., Kraft, P., Bull, S. B. & Sun, L. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS Genet.* **9**, e1003609 (2013).
44. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
45. Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
This study uses posterior probabilities of causality to construct credible sets of causal disease-associated SNPs across multiple loci and diseases under a single causal variant per locus assumption.
46. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
47. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
48. Chen, W. *et al.* Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015).
49. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
50. Newcombe, P. J., Conti, D. V. & Richardson, S. JAM: a scalable bayesian framework for joint analysis of marginal SNP effects. *Genet. Epidemiol.* **40**, 188–201 (2016).
51. Van de Bunt, M. *et al.* Evaluating the performance of fine-mapping strategies at common variant GWAS loci. *PLoS Genet.* **11**, e1005535 (2015).
52. Li, Y. & Kellis, M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res.* **44**, e144 (2016).
53. Udler, M. S. *et al. FGFR2* variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum. Mol. Genet.* **18**, 1692–1703 (2009).
54. Udler, M. S., Tyrer, J. & Easton, D. F. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet. Epidemiol.* **34**, 463–468 (2010).
55. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
56. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
57. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
58. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
59. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
This study uses a Bayesian hierarchical model to estimate posterior probabilities of causality and to identify functional annotations enriched for disease heritability under a single causal variant per locus assumption.
60. Chung, D., Yang, C., Li, C., Gelernter, J. & Zhao, H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet.* **10**, e1004787 (2014).
61. Kichaev, G. & Pasaniuc, B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
This study shows that fine-mapping accuracy can be improved by leveraging functional annotation data and trans-ethnic samples and modelling multiple causal variants per locus.
62. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
63. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
64. Liu, C.-T. *et al.* Trans-ethnic meta-analysis and functional annotation illuminates the genetic architecture of fasting glucose and insulin. *Am. J. Hum. Genet.* **99**, 56–75 (2016).
65. Grubert, F. *et al.* Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* **162**, 1051–1065 (2015).
66. Waszak, S. M. *et al.* Population variation and genetic control of modular chromatin architecture in humans. *Cell* **162**, 1039–1050 (2015).
67. Zaitlen, N., Pasaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).
68. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).
69. Ong, R. T.-H., Wang, X., Liu, X. & Teo, Y. Y. Efficiency of trans-ethnic genome-wide meta-analysis and fine-mapping. *Eur. J. Hum. Genet.* **20**, 1300–1307 (2012).

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

70. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).
71. Liu, C.-T. *et al.* Multi-ethnic fine-mapping of 14 central adiposity loci. *Hum. Mol. Genet.* **23**, 4738–4744 (2014).
72. Kuo, J. Z. *et al.* Trans-ethnic fine mapping identifies a novel independent locus at the 3′ end of *CDKAL1* and novel variants of several susceptibility loci for type 2 diabetes in a Han Chinese population. *Diabetologia* **56**, 2619–2628 (2013).
73. Chatterjee, N., Shi, J. & Garcia-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
74. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013).
75. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
This study uses polygenic risk scores to predict schizophrenia risk with appreciable accuracy, implicating a highly polygenic disease architecture.
76. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
77. Vilhjalmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
78. Henderson, C. R. Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447 (1975).
79. de los Campos, G., Gianola, D. & Allison, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* **11**, 880–886 (2010).
80. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–1557 (2014).
81. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
82. Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet.* **11**, e1004969 (2015).
83. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
84. Palla, L. & Dudbridge, F. A. Fast method that uses polygenic scores to estimate the variance explained by genome-wide marker panels and the proportion of variants affecting a trait. *Am. J. Hum. Genet.* **97**, 250–259 (2015).
85. Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
86. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
87. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
88. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
89. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
90. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
91. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7**, e1002254 (2011).
92. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
93. Styrkársdottir, U. *et al.* Nonsense mutation in the *LGR4* gene is associated with several human diseases and other traits. *Nature* **497**, 517–520 (2013).
94. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
95. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).
96. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
97. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
This study applies a Bayesian framework to identify pleiotropic effects across a broad set of complex traits and diseases.
98. Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
99. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
100. Burgess, S., Dudbridge, F. & Thompson, S. G. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat. Med.* **35**, 1880–1906 (2016).
101. Lee, S. H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
102. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
This study introduces a new method for estimating genome-wide genetic correlations from summary statistics.
103. Brown, B. C. *et al.* Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
104. Nieuwboer, H. A., Pool, R., Dolan, C. V., Boomsma, D. I. & Nivard, M. G. GWIS: genome-wide inferred statistics for functions of multiple phenotypes. *Am. J. Hum. Genet.* **99**, 917–927 (2016).
105. Hormozdiari, F. *et al.* Imputing phenotypes for genome-wide association studies. *Am. J. Hum. Genet.* **99**, 89–103 (2016).
106. [No authors listed.] Asking for more. *Nat. Genet.* **44**, 733 (2012).
107. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).
108. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
109. Sankararaman, S., Obozinski, G., Jordan, M. I. & Halperin, E. Genomic privacy and limits of individual detection in a pool. *Nat. Genet.* **41**, 965–967 (2009).
110. Visscher, P. M. & Hill, W. G. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet.* **5**, e1000628 (2009).
111. Erlich, Y. & Narayanan, A. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**, 409–421 (2014).
112. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
113. 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).
114. Price, A. *et al.* Pooled association tests for rare variants in exon resequencing studies. **86**, 832–838 (2010).
115. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322 (2011).
116. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
117. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PloS One* **3**, e3395 (2008).
118. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
119. Perry, J. R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
120. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
121. Zheng, H. F. *et al.* Whole-genome sequencing identifies *EN1* as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).
122. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
123. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
124. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
125. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
126. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–633 (2016).
127. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
128. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
129. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A_{1c} levels via glycemic and nonglycemic pathways. *Diabetes* **59**, 3229–3239 (2010).
130. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
131. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
132. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
133. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
134. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
135. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
136. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
137. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* http://dx.doi.org/10.1093/bioinformatics/btw613 (2016).

Acknowledgements
The authors are grateful to H. Finucane, S. Gazal, N. Mancuso and H. Shi for helpful discussions, and to G. Kichaev and R. Johnson for help with figure 3. The work of the authors is funded by US National Institutes of Health grants R01 HG006399, R01 MH101244, R01 GM105857 and R01 MH107649.

Competing interests
The authors declare no competing interests.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Author biographies

Bogdan Pasaniuc is an assistant professor in the Department of Pathology and Laboratory Medicine and the Department of Human Genetics in the David Geffen School of Medicine at the University of California, Los Angeles, USA. He obtained his Ph.D. in computer science and trained as postdoctoral fellow at the International Computer Science Institute, Berkeley, California, USA, followed by a postdoctoral fellowship at the Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA. His research group develops computational and statistical methods to understand the genetic basis of complex traits, focusing on integrative genomics, fine-mapping and heritability analyses.

Alkes L. Price is an associate professor in the Program in Genetic Epidemiology and Statistical Genetics in the Department of Epidemiology at the Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA, with a secondary appointment in the Department of Biostatistics. He is an associate member of the Program in Medical and Population Genetics at the Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. His research focuses on the development of statistical methods for uncovering the genetic basis of human disease, and on the population genetics underlying these methods. Areas of interest include functional components of heritability, common versus rare variant architectures, and disease mapping in structured populations.

Key points

- Summary association statistics from genome-wide association studies (GWAS) are widely available in large sample sizes across hundreds of complex traits. Analyses of such data can yield important insights, motivating the development of new statistical methods in this area.
- Single variant association analysis (including meta-analyses, conditional association and imputation) can be performed effectively using summary association data. These methods often rely on linkage disequilibrium (LD) information from population reference panels.
- Summary association data can be used to perform gene-based association tests to identify genes influencing complex traits. In particular, expression quantitative trait loci (eQTLs) can be integrated to identify genes whose expression levels influence complex traits, and rare variant association tests can aggregate evidence of association across multiple rare variants in a gene.
- Statistical fine-mapping of causal variant (or variants) at GWAS loci can be performed using summary association data, leveraging information on the strength of association, functional genomic annotations and differences in LD patterns across different populations.
- It is becoming increasingly clear that most complex traits and common diseases have a large number of causal variants with small effects. Summary association statistics can be used to understand these polygenic architectures and leverage them for polygenic risk prediction.
- Summary association statistics have broad utility in cross-trait analyses, including detecting pleiotropic effects and inferring genetic correlations between traits. Pleiotropic effects can be used in Mendelian randomization analyses to draw inferences about causal relationships among traits.

Subject categories

Biological sciences / Genetics / Genome / Genetic variation [URI /631/208/726/649]
Biological sciences / Genetics / Heritable quantitative trait / Quantitative trait loci [URI /631/208/729/743]
Biological sciences / Genetics / Genetic association study / Genome-wide association studies [URI /631/208/205/2138]

ToC blurb

000 **Dissecting the genetics of complex traits using summary association statistics**

Bogdan Pasaniuc and Alkes L. Price

Investigating the genetic basis of complex traits and diseases using individual-level genetic data from genome-wide association studies is often hampered by privacy concerns and logistical considerations. Here, the authors review recent statistical methods that leverage summary association data, which are widely available and can circumvent these issues.

© 2016 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.