GENOME-WIDE ASSOCIATION STUDIES — OPINION

Assessing and managing risk when sharing aggregate genetic variant data

David W. Craig, Robert M. Goor, Zhenyuan Wang, Justin Paschall, Jim Ostell, Michael Feolo, Stephen T. Sherry and Teri A. Manolio

Abstract | Access to genetic data across studies is an important aspect of identifying new genetic associations through genome-wide association studies (GWASs). Meta-analysis across multiple GWASs with combined cohort sizes of tens of thousands of individuals often uncovers many more genome-wide associated loci than the original individual studies; this emphasizes the importance of tools and mechanisms for data sharing. However, even sharing summary-level data, such as allele frequencies, inherently carries some degree of privacy risk to study participants. Here we discuss mechanisms and resources for sharing data from GWASs, particularly focusing on approaches for assessing and quantifying the privacy risks to participants that result from the sharing of summary-level data.

Population-based genetic studies have the potential to unlock biological mechanisms of disease and reveal their genetic underpinnings. In particular, genome-wide association studies (GWASs) using hundreds of thousands to millions of SNPs have emerged during recent years as a particularly fruitful study design for identifying common variants with subtle genetic effects in complex disorders[1]. A few initial studies made substantial findings by studying only a few hundred individuals, such as in age-related macular degeneration[2]. However, more often the small effect-size of associated SNPs requires the genotyping of thousands of individuals when studying complex diseases across a population.

In the past 2 years there has been a trend towards including tens to hundreds of thousands of individual participants in a GWAS[3]. Examples include: a meta-analysis of 5,539 cases and 17,231 controls for rheumatoid arthritis and 4,533 cases and 10,750 controls for coeliac disease, which collectively identified seven shared loci for these diseases[4]; 6,688 cases and 13,685 controls for Alzheimer's disease, which identified five new genome-wide significant associations[5]; a meta-analysis of 22,233 individuals with coronary artery disease and 64,762 controls taken from 14 GWASs, which identified 13 new susceptibility loci[6]; and a meta-analysis of >100,000 individuals, which identified 59 newly associated loci with cholesterol and blood lipid levels[7]. These studies emphasize how increasing study sizes to tens or hundreds of thousands of individuals is enabling the discovery of multiple genome-wide significant associations, rather than a single or a few loci as was frequently the case in early studies. Often these large-scale GWASs result from a meta-analysis of many previous studies and are inherently enabled by the sharing of genetic data.

A consideration for any genetic study is the need to protect individual participants from the risk of re-identification, and maintaining privacy becomes more complex when data are shared beyond the original study in which the individual agreed to participate. We address these considerations by first discussing frameworks and resources for sharing data from GWASs, and then highlighting some of the risks that are associated with common modes of sharing data. Data can be shared either as information about the individual or as population-level data; we focus particularly on privacy challenges when sharing population-level data (such as allele frequencies) with a large audience, which until recently was regarded as relatively 'safe' in terms of privacy. We describe quantitative approaches and additional considerations in assessing the risk to the privacy of individual participants at varying levels of sharing aggregate data. We consider quantitative approaches in the most depth, as there is currently much deliberation in the research community regarding how risk can be assessed and taken into account when planning studies.

What level of data can be shared?

Sharing individual-level data. Generally, the most comprehensive data-sharing from a GWAS is the distribution of full phenotypic information, accompanied by individual-level genotype data, for each participant. Phenotypic information could be tied to a set of full medical records, such as has been conducted by the electronic Medical Records and Genomics (eMERGE) Network in which genotype data are linked to various conditions including dementia, lipid levels and type 2 diabetes[8], or phenotypic information could be limited to a dichotomous trait, such as a case versus control status. At a simple level the genotype information could be only genotype calls (for example, AA/AG/GG) for >500,000 SNPs, although it could include the raw microarray data that were used for calling genotypes or for identifying copy number variants (CNVs).

Access to the individual-level data has several advantages for analyses across data sets. First, access to individual-level data allows for a joint analysis across all samples, giving greater power to detect associations than a meta-analysis of summary-level statistics[9]. Second, access to individual-level data ensures a uniform analysis across all data sets, in terms of the application of quality control filters (such as SNP missingness), as well as higher-level analyses such as imputation. Imputation is often used to combine data sets that were genotyped on different platforms in order to predict untyped markers[10]. However, the accuracy of imputation can vary depending on the method — such as Beagle, Mach and Impute — or on the training set, such as the 1000 Genomes data sets[11]. Third, the sharing of individual-level data can allow the assessment of multiple variants in combination within a single individual to calculate combined effects (such as SNP–SNP interactions) of multiple associated variants; for example, a cumulative effect of five variants was identified in a meta-analysis of prostate cancer GWASs[12]. Finally, access to the individual-level, raw probe-level microarray data can be used to ascertain evidence for copy number variants that are associated with disease: recently, the enrichment of duplications of the vaso-active intestinal peptide receptor 2 (*VIPR2*) gene was found to be associated with schizophrenia in an analysis that used data from several GWASs for mental health[13].

Clearly, the most important challenge with sharing individual-level genomics data is protecting the privacy of individual participants. As discussed by Heeney and others[14,15], individual genetic data from GWASs are not only uniquely identifying, they can also predict disease risk, and it is possible for consumers outside the scientific community to generate genetic profiles on individuals (such as through the 23andMe personal genomics company). In addition to privacy issues, sharing individual-level data has challenges arising from the size and variability of electronic files that are associated with array-based genotype data. For example, the sharing of probe-level data for CNVs or genotype-level data involves file sizes exceeding 100 megabytes per sample, and these files often contain highly specific formatting or referencing, which are required to avoid strand-assignment errors and inconsistencies across genome builds. Informatically, summary-level data are often preferred over individual-level data when the researcher's goal is the rapid acquisition of allele frequencies and *P* values for exploratory or confirmatory purposes.

Sharing summary-level data. An alternative to sharing individual-level data is sharing aggregate data or summary statistics. Such statistics include genotype counts, allele frequencies, *P* values, odds ratios and other measures of effect size. In dichotomy analyses (case–control), researchers can use genotype counts to calculate summary-level statistics (such as *P* values or odds ratios) under different genetic models (such as dominant, additive, recessive or co-dominant). When study conditions are carefully matched, the counts from different data sets can be used directly in meta-analyses. Also, in the context of a publication, sharing of allele frequencies, *P* values and odds ratios is routine and is essential for future studies to replicate the most-associated SNPs. As highlighted above, many large-scale GWASs have been enabled by meta-analyses that required access to the many associated markers with low effect-sizes that are generally not included in the tables or supplementary data of a primary paper reporting an individual study. Therefore, data sharing is essential for such meta-analyses. Increasingly, large studies are carefully managed efforts involving multiple consortia in which each member contributes summary statistics that are independently prepared in a consistent manner and then combined for meta-analysis[7,16,17].

From the perspective of risk to the privacy of individual participants, aggregating data into summary statistics provides some level of privacy protection. However, as discussed below, some degree of residual identifying information remains in the cumulative analysis of large numbers of SNPs.

> “Quantifying the risk of making summary-level data broadly available is an essential part of the risk-assessment process”

Mechanisms for data sharing

Study consortia. Meta-analyses of GWAS data that include tens to hundreds of thousands of individuals are often carried out by multicentre consortia that have set up mechanisms whereby either summary-level or individual-level data are gathered into a central database. The Coronary Artery Disease Genome-wide Replication and Meta-analysis (CARDIoGRAM) consortium published one such example in which a steering committee provided oversight of several coordinated analysis groups that submitted carefully constructed summary-analysis results to a centralized database where the combined meta-analysis was completed[16]. Another current example is the Gene Environment Association Studies (GENEVA) consortium, which consists of 14 independent GWASs for various phenotypes and includes over 80,000 study participants. Consistent quality control and the use of centralized data deposition to the Database of Genotypes and Phenotypes (dbGaP) for individual- and summary-level data are essential to the GENEVA consortium[18]. International efforts that use centralized computing also include the Psychiatric GWAS Consortium[19], in which individual-level data are uploaded to a common server for structured analyses across a series of psychiatric disorders, including bipolar disorder, schizophrenia, autism and other mood disorders.

Databases for broader distribution. The consortia described above represent major efforts at coordinated and controlled approaches for data analysis. It is expected that the individual-level data and summary-level data will have utility in future studies, and so additional long-term data-sharing mechanisms are essential. For example, individual researchers might be looking to validate the associations of a specific gene or variant that are not listed in the primary publications. Beyond individual researchers, new consortia may wish to rely on historical data; for example, the Population Reference Sample (POPRES)[20] study created a common resource of controls that may be useful for adding power to future case–control GWASs. An important aspect of enabling and realizing the future value of truly large-scale GWASs using thousands of samples has been the emergence of common repositories for distributing both individual-level and summary-level data. These databases provide a centralized resource for sharing data, while providing mechanisms to protect the privacy of individual study participants.

Two notable resources include samples genotyped through the International HapMap[21] and 1000 Genomes projects[11], which are broadly distributed through multiple databases including dbSNP[22]. These resources are typically used to observe the variability of population-specific allele frequencies and are generally not used to generate significance-of-association signals. Various other databases allow for controlled or restricted access to individual-level data and/or summary-level data. For US National Institutes of Health (NIH)-funded studies, dbGaP[23] currently holds individual- and/or summary-level data (available through a controlled-access process) for approximately 1,900 data sets covering more than 257,000 individuals. Other sources of individual-level data include: the Genome Medicine Database of Japan (GeMDBJ), which was developed by the study groups of Japan's Millennium

Genome Project (MGP) and maintains over 570,000 SNPs on 2,000 patients in three disease groups; the Wellcome Trust Case–Control Consortium (WTCCC), which maintains over 500,000 SNPs from approximately 14,000 individuals; and the European Bioinformatics Institute (EBI) European Genome–Phenome Archive[24], which distributes data from Phases 1, 2 and 3 of the WTCCC and from 20 other study-specific providers. HuGE Navigator[25], GWAS Central[26], JSNP[27], dbGaP[23] and the Phenotype–Genotype Integrator (PheGenI)[28] (see Further information) provide individual- and summary-level data for several hundred thousand SNPs across diseases and several thousand samples, with multiple levels of controlled access.

An important aspect of databases such as dbGaP is their ability to provide study-specific levels of controlled access to individual- and summary-level data. At one extreme, completely unrestricted access is strongly desired for at least: a subset of variants, such as a list of hundreds to thousands of associated SNPs for tables in a publication; summary measures for individual phenotypic measures; and study protocols and data-collection forms. A more constrained approach is to approve institutional users to allow them to download more-extensive summary-level data (that is, more than hundreds to thousands of SNPs) for studies, without additional approval being required. In the case of dbGaP, open access is available for the broad release of non-sensitive data, and in the case of PheGenI, users can query *P* values across studies for a limited number of SNPs in an open-access manner[28]. Controlled access through dbGaP is available when additional oversight is required for sensitive data sets that involve: individual-level genotype data; individual genome or exome sequences; or comprehensive genome-wide associations for a published study, including the allelic direction of effect. Access is controlled through an application process that has been reviewed by one or more NIH data-access committees (DACs) that oversee the data set (or sets) of interest, with terms of use conveyed through a data-use certification (DUC) agreement.

Privacy for summary-level data

Risks of identification in shared summary-level data sets. As seen from many meta-analyses, summary-level data have great utility when combining multiple studies or even when validating a small number of SNPs. It was originally assumed that summary-level data completely anonymized the participants and hence that this type of data could be openly distributed for all SNPs. For example, if the genotypes of ten individuals for a particular SNP were AA, AT, AT, AA, TT, AT, AA, AT, AA and AT, respectively, the allele frequency summary statistic is 65% A (13 As of 20 total alleles). Intuitively, data users cannot determine much about the individuals when reporting only the allele frequency summary statistic of 65% for a single SNP. Initial views were that this would still apply when considering larger numbers of SNPs, particularly when the average allele frequencies were for hundreds if not thousands of individuals. However, in 2008 Homer *et al.*[29] showed that, in principle, one could estimate whether an individual was a member of a cohort using the marginal information in the allele frequency data across tens of thousands of SNPs. This estimation requires access to genotype data from that individual and access to genetic data from a reference population[30]. Under some circumstances this could be done even in cohort sizes exceeding 1,000 individuals but, as discussed below, the ability to estimate membership is influenced by several factors.

This concept can be illustrated using a simple scenario. First, suppose that there is a data set of ten SNPs in which the allele frequency is 60% for the A allele for all ten SNPs. Next, suppose that we want to determine whether a person is in this data set, with the additional knowledge that this individual has an AA genotype for these ten SNPs. If it was known that the allele frequency of these ten SNPs was actually 50% A in a reference data set, we could construct a statistic that cumulatively accounts for the fact that the observed allele frequency in the data set of interest (as compared with the reference data set) is biased towards the A allele. In this example, we see a shift from the expected 50% A allele frequency to an observed 60% A allele frequency for 10 out of 10 SNPs for which the individual of interest is homozygous AA. One could calculate, by various approaches, a probability that the person is in the data set of interest, based on the observation of 10 out of 10 SNPs being shifted in their average allele frequencies in a direction that is consistent with the allele found in the person of interest.

The publication by Homer *et al.*[29] demonstrated an example cumulative test-statistic which showed that, in principle, one could determine membership in summary-level allele-frequency data sets by comparison with a reference data set. The numbers of SNPs, their minor allele frequency and (to a limited extent) the accuracy of measuring allele frequencies, were all found to influence the ability to improve the estimate of cohort membership. The size of the reference population is also important; this was not considered in Homer *et al.*[29] but is addressed in the discussion below. A series of subsequent studies investigated the implications and statistical aspects of estimating sample membership from aggregate data from GWASs compared

Glossary

Allele frequency
The frequency of the less-common allele of a polymorphism. It has a value between 0 and 0.5 and can vary between populations.

Bayesian
A statistical framework for evaluating a hypothesis. The Bayesian approach assesses the probability of a hypothesis being correct by incorporating the prior probability of the hypothesis.

Discrimination threshold
The significance threshold for rejecting the null hypothesis in a statistical test.

Frequentist
A statistical framework for evaluating a hypothesis. The frequentist approach tests a hypothesis as being correct given the strength of a data set.

Imputation
A method for inferring untyped variants from neighbouring variants, based on linkage disequilibrium and haplotype structure.

Linear regression
The estimation of a first-order relationship between two variables, which involves fitting a line of best fit to the data.

Missingness
The percentage of samples that do not receive a genotype call for a SNP in a genome-wide association study.

Neyman–Pearson lemma
A theorem that assures the optimality of a likelihood ratio test between simple hypotheses at a given threshold.

Prevalence
The prior probability that a person is in a data set of interest. Alternatively, the term can refer to the fraction of individuals in a data set out of the total number of individuals that could be in the data set.

Reference data set
A data set of samples from individuals who are from the same population that was sampled in the summary-level data set of interest.

with reference populations. Specifically, studies by Sankararaman *et al.*[30] and Jacobs *et al.*[31] formulated statistical frameworks based on likelihood ratios that optimize the power to estimate membership, in part by leveraging the binomial distribution associated with sampling biallelic markers that are pooled in equimolar amounts across a defined number of samples. In fact, by the Neyman–Pearson lemma[32] these methods are an optimal solution. Additionally, Braun *et al.*[33] showed that a high rate of false positives can arise in the original formulation used by Homer *et al.* if the effects of linkage disequilibrium are ignored and a normal distribution is assumed. Conversely, Zhou and colleagues[34] showed that linkage disequilibrium can be leveraged to improve power in a statistical approach that uses multiple correlated markers within long haplotype blocks. Linear regression-based frameworks have also been presented by Visscher *et al.*[35], and Clayton[36] has presented a Bayesian-based alternative to the frequentist approach. Finally, Sampson and Zhao[37] demonstrated methods to address aspects of unknown ancestry by using multiple reference populations.

Box 1 | Risk-assessment definitions applied to sharing GWAS aggregate data sets

In order to consider risk-assessment definitions, it is useful to first recall the standard 'ability of a test to detect a disease' measures of sensitivity, specificity and positive and negative predictive values, as shown in the upper table. Each of these can be converted to an 'ability to classify an individual as being in a genome-wide association study (GWAS) data set', as shown in the lower table.

	Disease positive	Disease negative
Test positive	*a*	*b*
Test negative	*c*	*d*

	Actually in the data set	Truly not in the data set
Classified in cohort	*a*	*b*
Not classified in cohort	*c*	*d*

The risk-assessment definitions in the context of GWAS data sets are listed below.

Type II error. The proportion of times that someone who is actually in the data set is not identified as being in the data set. For example, with 20% type II error, there is a 20% chance of failing to determine that someone is in a data set.

Type I error. The proportion of times that someone is predicted to be in the data set when they are not. For example, with 5% type I error, there is a 5% chance of determining that someone is in the data set when they are not.

Sensitivity. The ability to detect true positives (that is, the correct classification of disease by test result or of people in the data set). In both cases, this would be $(a) / (a + c)$. For example, with a sensitivity of 30%, only 30% of test individuals in the data set will be correctly classified as being in the data set; 70% of those actually in the data set will be missed.

Specificity. The proportion of those people that are not in the data set who are correctly classified as not being in the data set (that is, true negatives). In the lower table, this would be $(d) / (b + d)$. For example, with a specificity of 40%, only 40% of test individuals will be correctly classified as not being in the data set; 60% of those classified as being in the data set actually are not.

Power. The proportion of times that an individual who is actually in the data set will be correctly classified as being in the data set. For example, with 80% power, there is an 80% chance of correctly classifying someone as being in the data set.

Positive predictive value. The positive predictive value (PPV) is defined as the number of true positives divided by the total number of all positives ($(a) / (a + b)$). This measure is frequently used for rare disorders. Similarly, most individuals from a population would not actually be in a GWAS data set. PPV is the proportion of all individuals predicted to be positive from a population that are truly in a data set. With 20% PPV, only 20% of those identified as being in the cohort actually will be; 80% will not (and hence the ratio of false positives to true positives would be 4:1).

What are the implications? The major realization from these papers was that, theoretically, there might be a risk that a data user could determine whether an individual was in a data set, even if only summary-level genotype frequencies were available. This determination is possible, provided that the data user had access to that individual's genotypes for those SNPs and had a sufficiently representative reference set of allele frequencies. Because most GWASs are studies of disease, this implies that there might be a path to determine medically relevant information about participants from summary-level data. Following publication of the Homer *et al.* paper[30], the NIH addressed the sharing of data from GWASs in a paper in *Science*[38], and the NIH and many other groups discontinued openly distributing disease-specific summary-level data sets. The level of risk to participants and the appropriateness of this response have been intensely debated. Krawczak *et al.*[39] have argued that current NIH policy is counterproductive owing to the increased burden on international consortia to comply with the requirements of NIH-based central repositories. Some of this debate was published in a series of articles in *PLoS Genetics,* including one article that provided five views on balancing research with protecting privacy[15]. The authors generally agreed that some privacy risk is inherent to genetic studies and that a balance between research and privacy is needed, although there was less agreement on where the balance should lie. Interim models were suggested whereby the credentials of individuals and institutions could be validated to allow access to full summary-level data, and these models are consistent with a study by Haga and O'Daniel[40], which showed that individuals are more likely to participate in studies that have some restrictions for online access. Further research is ongoing to try to further assess the risks to study participants.

Finally, we remark that estimating membership in a data set requires access to genetic data from the individual or their relative. Clearly, if these data are available then one can assume that there has already been some loss of privacy; for example, the genetic data can reveal information about disease risks or ancestry. Indeed, in a publication that recommended policies for minimizing identification risks in clinical research data, Malin *et al.*[41] described determining membership from aggregate SNP data as 're-identification' because some aspects of a person's identity are already available. Nevertheless, access to the genetic data of an individual does not render privacy expectations unimportant. First, participation as a phenotyped case in a study more accurately reflects being diagnosed with a disease compared to disease risk predictions from SNP data owing to

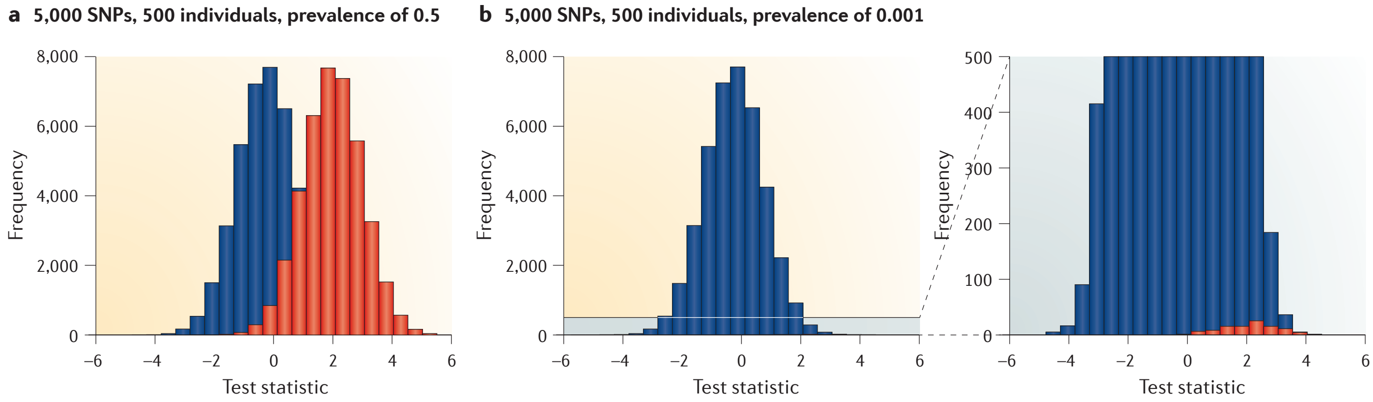

Figure 1 | **Sharing 5,000 SNPs at different prevalence or prior probabilities.** In the plots, we use simulations to show how the prior probability of being in a data set has an impact on the ability to determine whether a person is in that data set, using summary-level allele frequencies from 5,000 SNPs on data sets of 500 individuals. **a** | A histogram of test statistics, based on the approach of Jacobs *et al.*[31], for 100,000 simulations when the person tested is actually in a data set (red bars) and for 100,000 simulations when the person tested is not in a data set (blue bars). Because the simulations are equal between being in a data set or not, the prevalence or prior probability of being in the data set is 0.5. **b** | 100,000 simulations when the person is not in the data set (blue bars) and 100 simulations when they are in the data set (red bars), equivalent to a prevalence or prior probability of being in the data set of 0.001. The graph on the right is a zoomed view of the section within the grey box from the left graph. This shows that a large number of tests of individuals that are not in the data set can obscure the ability to distinguish true positives from false positives. Describing risk in terms of a positive predictive value (PPV) allows the consideration of prevalence for being in a data set as a prior probability, thus increasing the accuracy of assessing the risk of a person in a data set being correctly classified.

the modest effect-size of most associations. Even in strongly associated examples there is only modest predictive value; for example, an odds ratio of >5 for the association of the *APOE*-ε4 allele with Alzheimer's disease gives only a modest predictive value for an individual with mild cognitive impairment developing Alzheimer's disease[42]. Second, the Homer *et al.* paper[29] and subsequent publications (for example, REF. 30) showed that one could potentially also learn something about the immediate relatives of the genotyped individual, owing to shared genetics, even without knowing the exact regions or variants that they share (for example, to learn about a child from information about the parent). Therefore, even in the cases of related individuals with shared genetics there are still important expectations of privacy.

Assessing risk for summary data

As noted above, in practice some level of open distribution of aggregate data is necessary to communicate results in the literature. Assessing privacy risk is an important aspect when disseminating findings from a GWAS (for example, whether to release information about SNPs that do not reach significance, as well as information about those that do). A dilemma that has been faced by many researchers is what balance should be struck between releasing summary-level data during publication or through searchable databases and minimizing risks to the privacy of study participants. For example, how do researchers determine the number of SNPs that should be placed on the Web or in a supplementary table? Is releasing summary-level data from 1,000 or 5,000 SNPs reasonable? Managing and assessing the risk when sharing summary-level data should balance multiple factors — both quantitative and non-quantitative — and should have a clear deliberation process.

Non-quantitative risk assessment should include consideration of the potential consequences of someone in a particular cohort being identified as a participant. For example, the identification of participants in studies of readily observable common traits, such as obesity or hair colour, would be less concerning than the identification of individuals in studies of alcohol dependence, illegal behaviour or psychiatric conditions. These types of non-quantitative risk considerations are often study-specific, and higher-level restrictions on access may only be warranted for higher-risk studies. In databases such as dbGaP, there is the ability to define access restrictions through the DUC agreement. For example, some data sets require applicants to obtain institutional review board (IRB) approval for access, whereas many other data sets allow for general access after institutions and users agree to adhere to sharing and reporting policies that are standard for GWASs.

Quantifying the risk of making summary-level data broadly available is an essential part of the risk-assessment process, and is a quantification that lends itself to more traditional approaches for risk assessment. BOX 1 introduces several key concepts in risk assessment, such as sensitivity, specificity and positive predictive value (PPV). Each of these metrics gives insight into a specific type of risk. Beyond these metrics, software tools also exist for quantifying the risk that is associated with summary-level data from GWASs. Notably, Sankararaman *et al.*[30] published a method and software tool called SecureGenome, which uses an input genotype data set and a reference data set and determines, from the upper bounds of the optimally solved likelihood ratio test, the number of highly ranked SNPs that can be safely exposed.

Positive predictive value. In this section we discuss a metric, PPV, that can be used in quantitative risk assessments in the context of sharing data. PPV specifically accounts for the size of the sampled population and the fact that most individuals from a population are actually not in the data set. In a concept highlighted by Braun *et al.*[33], false-positive rates are inversely related to the proportion of the population sampled, and PPV as a metric can quantify the risk of correctly identifying an individual as being in a data set, given that most individuals from the population are not actually included in the data set.

Calculating the PPV requires determination of the proportion of the 'at-risk' population that is in the GWAS. As an example, assume that a data user wished to determine whether a person was in a data set of 1,000 European-ancestry individuals as part of the Framingham study (and that the data user had genotype data for this person). Given an estimated 65,000 individuals in Framingham, with approximately 75% of the population being of European ancestry, the 'at-risk' population is approximately 50,000 individuals. The prevalence is thus 1,000/50,000 = 0.02. Without any data, the risk of positively identifying a person who is actually in the data set is therefore 2%. Thus the prevalence allows an estimation of the PPV given this prior knowledge. Prevalence of participants in a study may be low in large-scale studies, or may become relatively high in small 'at-risk' populations such as in a GWAS of the Native Hawaiian populations or the Old Order Amish. The influence of prevalence on risk assessment through PPV is illustrated in FIG. 1, in which a simulation with a high prevalence is compared with a simulation with a low prevalence. With low prevalence, the risk of resolving membership of a cohort is greatly reduced. Therefore, the strength of PPV as a measure is that it inherently accounts for the prior probability that a person selected at random is actually in the data set and inherently accounts for key aspects of the population as part of risk assessment[29,30].

Table 1 | **Risk assessment with different prevalence parameters**

SNPs	Cohort size	Sensitivity	Specificity	PPV
Prevalence = 0.001				
100	100	0.05	0.99	0.010
100	500	0.04	0.99	0.004
100	1,000	0.01	0.99	0.004
500	500	0.19	0.98	0.011
500	1,000	0.12	0.98	0.008
1,000	500	0.36	0.97	0.012
1,000	1,000	0.21	0.97	0.007
5,000	500	0.83	0.99	0.075
5,000	1,000	0.51	0.99	0.038
Prevalence = 0.01				
100	100	0.05	0.99	0.080
100	500	0.06	0.99	0.067
100	1,000	0.04	0.99	0.034
500	500	0.28	0.96	0.061
500	1,000	0.17	0.97	0.049
1,000	500	0.44	0.95	0.076
1,000	1,000	0.27	0.95	0.056
5,000	500	0.89	0.98	0.292
5,000	1,000	0.63	0.98	0.275

The table shows sensitivity, specificity and positive predictive value (PPV) for sharing <5,000 SNPs for <5,000 individuals, assuming a prevalence of 0.001 (upper part of the table) or 0.01 (lower part of the table), based on simulated Framingham SNP Health Association Resource (SHARe) genome-wide association data. In these simulations, summary-level allele frequency data sets were created by randomly selecting a fixed number of individuals from the Framingham SHARe data set into two data sets. From these data sets, SNPs that failed Hardy–Weinberg equilibrium ($<10^{-6}$), minor allele frequency (<0.01), missingness (<0.01) and call rate (<0.97) were removed using the PLINK analysis tool set[43]. Association statistics were calculated for all SNPs, but sharing of allele-frequency data was only assumed for the most associated SNPs by *P* value (5,000 SNPs in the examples shown in the table). Individuals in and not in the data set were evaluated at a defined prevalence with a significance threshold of 0.005, with the entire process repeated until 100,000 simulations were completed.

As explained above, researchers are often faced with the question of how many SNPs should be included in the summary data that they release. The PPV is one way to obtain a quantitative risk assessment for different numbers of SNPs and different study sizes; TABLE 1 provides several examples of PPV as a risk assessment in simulations of releasing between a few hundred and a few thousand of the SNPs with the highest associations (by *P* value) from a study with different prevalence settings. In these simulations we used a prevalence of 0.01, which could be similar to a study of cardiovascular traits in a Framingham population, and 0.001, which could be similar to a study of 1,000 individuals with major depression sampled from a population that is defined to include all people of European ancestry in the United States. The results of these simulations show the importance of considering prevalence: for 5,000 SNPs and a cohort size of 500, the PPV is 29.2% for a prevalence of 0.01 and 7.5% for a prevalence of 0.001, both with a discrimination threshold of 0.001. Further results from TABLE 1 suggest that sharing 1,000 SNPs for data sets with >500 individuals generally leads to a low PPV, regardless of the population size. Taken together, the process of assessing risk with PPV and/or other statistical metrics can be used to inform discussions of non-quantitative risks.

Summary and implications

The path for future GWASs will benefit from, and depend on, data sharing. Recent large-scale efforts showed that careful, coordinated efforts of sharing summary-level data led to the discovery of many new genome-wide significant associations. With hindsight, these associations were often apparent in the original studies, although not at levels that merited follow-up sequencing. Clearly, the sharing of data and the ability to access summary-level data will be an important part of identifying new associations in future studies, and protecting the privacy of participants is an important part of this process. Quantitatively assessing privacy risks using PPV incorporates population size and can inform the discussion of non-quantitative factors, such as the impact of an individual being identified in studies. Therefore, it is our opinion that quantitative tools should play a useful part in assessing the risk of determining that an individual is in a data set when releasing aggregate genome-wide SNP genotyping data sets and subsets of these data sets.

David W. Craig is at the Translational Genomics Research Institute (TGen), Phoenix, Arizona 85004, USA.

Robert M. Goor, Zhenyuan Wang, Justin Paschall, Jim Ostell, Michael Feolo and Stephen T. Sherry are at the National Center for Biotechnology Information (NCBI), Bethesda, Maryland 20892, USA.

Teri A. Manolio is at the National Human Genome Research Institute (NHGRI), Bethesda, Maryland 20892, USA.

Correspondence to D.W.C
e-mail: dcraig@tgen.org

doi:10.1038/nrg3067

1. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
2. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
3. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
4. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
5. Hollingworth, P. *et al.* Common variants at *ABCA7*, *MS4A6A/MS4A4E*, *EPHA1*, *CD33* and *CD2AP* are associated with Alzheimer's disease. *Nature Genet.* **43**, 429–435 (2011).
6. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genet.* **43**, 333–338 (2011).
7. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
8. Kho, A. N. *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Sci. Transl. Med.* **3**, 79re1 (2011).
9. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genet.* **38**, 209–213 (2006).
10. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Rev. Genet.* **11**, 499–511 (2010).
11. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
12. Zheng, S. L. *et al.* Cumulative association of five genetic variants with prostate cancer. *N. Engl. J. Med.* **358**, 910–919 (2008).
13. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene *VIPR2* confer significant risk for schizophrenia. *Nature* **471**, 499–503 (2011).
14. Heeney, C., Hawkins, N., de Vries, J., Boddington, P. & Kaye, J. Assessing the privacy risks of data sharing in genomics. *Public Health Genomics* **14**, 17–25 (2011).
15. Church, G. *et al.* Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet.* **5**, e1000665 (2009).
16. Preuss, M. *et al.* Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: a genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circ. Cardiovasc. Genet.* **3**, 475–483 (2010).
17. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genet.* **42**, 937–948 (2010).
18. Cornelis, M. C. *et al.* The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet. Epidemiol.* **34**, 364–372 (2010).
19. The Psychiatric GWAS Consortium Steering Committee. A framework for interpreting genome-wide association studies of psychiatric disorders. *Mol. Psychiatry* **14**, 10–17 (2009).
20. Nelson, M. R. *et al.* The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.* **83**, 347–358 (2008).
21. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
22. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
23. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–1186 (2007).
24. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39**, D28–D31 (2011).
25. Yu, W., Gwinn, M., Clyne, M., Yesupriya, A. & Khoury, M. J. A navigator for human genome epidemiology. *Nature Genet.* **40**, 124–125 (2008).
26. Thorisson, G. A. *et al.* HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.* **37**, D797–D802 (2009).
27. Hirakawa, M. *et al.* JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res.* **30**, 158–162 (2002).
28. Hindorff, L. A. *et al.* PheGenI: an integrated resource for browsing genetic association data. *Proc. of the 2011 AMIA Summit on Translational Bioinformatics* [online], http://proceedings.amia.org/16pcs7/16pcs7/1 (2011).
29. 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).
30. Sankararaman, S., Obozinski, G., Jordan, M. I. & Halperin, E. Genomic privacy and limits of individual detection in a pool. *Nature Genet.* **41**, 965–967 (2009).
31. Jacobs, K. B. *et al.* A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genet.* **41**, 1253–1257 (2009).
32. Neyman, J. & Pearson, E. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond. A* **231**, 289–337 (1933).
33. Braun, R., Rowe, W., Schaefer, C., Zhang, J. & Buetow, K. Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS Genet.* **5**, e1000668 (2009).
34. Wang, R., Li, Y. F., Wang, X., Tang, H. & Zhou, X. Learning your identity and disease from research papers: information leaks in genome wide association study. *Proc. of the 16th ACM Conf. on Computer and Communications Security*, 534–544 (2009).
35. Visscher, P. M. & Hill, W. G. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet.* **5**, e1000628 (2009).
36. Clayton, D. On inferring presence of an individual in a mixture: a Bayesian approach. *Biostatistics* **11**, 661–673 (2010).
37. Sampson, J. & Zhao, H. Identifying individuals in a complex mixture of DNA with unknown ancestry. *Stat. Appl. Genet. Mol. Biol.* **8**, 37 (2009).
38. Zerhouni, E. A. & Nabel, E. G. Protecting aggregate genomic data. *Science* **322**, 44 (2008).
39. Krawczak, M., Goebel, J. W. & Cooper, D. N. Is the NIH policy for sharing GWAS data running the risk of being counterproductive? *Investig. Genet.* **1**, 3 (2010).
40. Haga, S. B. & O'Daniel, J. Public perspectives regarding data-sharing practices in genomics research. *Public Health Genomics* 24 Mar 2011 (doi:10.1159/000324705).
41. Malin, B., Karp, D. & Scheuermann, R. H. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J. Investig. Med.* **58**, 11–18 (2010).
42. Elias-Sonnenschein, L. S., Viechtbauer, W., Ramakers, I. H., Verhey, F. R. & Visser, P. J. Predictive value of *APOE*-ε4 allele for progression from MCI to AD-type dementia: a meta-analysis. *J. Neurol. Neurosurg. Psychiatry* 14 Apr 2011 (doi:10.1136/jnnp.2010.231555).
43. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

Acknowledgements

This manuscript represents the views and opinions of its authors and does not necessarily represent the views or policies of the NIH or the US Department of Health and Human Services. This research was supported in part by the Intramural Research Program of the NIH National Library of Medicine. D.W.C. would like to acknowledge support from the US National Heart, Lung and Blood Institute (NHLBI), award U01 HL086528. The authors thank I. Marpuri, S. Buchholtz and L. Gyi for their support in coordinating the development of this work.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

23andMe personal genomics company: https://www.23andme.com

Electronic Medical Records and Genomics (eMERGE) Network: https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page

Database of Genotypes and Phenotypes (dbGaP): http://www.ncbi.nlm.nih.gov/dbgap

dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP

European Genome–Phenome Archive: http://www.ebi.ac.uk/ega

Genome Medicine Database of Japan (GeMDBJ): https://gemdbj.nibio.go.jp/dgdb

GWAS Central (includes policy): http://gwas.nih.gov

HuGE Navigator: http://hugenavigator.net/HuGENavigator/home.do

JSNP: http://snp.ims.u-tokyo.ac.jp

***Nature Reviews Genetics* series on Genome-Wide Association Studies:** http://www.nature.com/nrg/series/gwas/index.html

Phenotype–Genotype Integrator (PheGenI): http://www.ncbi.nlm.nih.gov/gap/PheGenI

Public Population Project in Genomics (P3G): http://p3g.org

SecureGenome: http://securegenome.icsi.berkeley.edu/securegenome

SNP Health Association Resource (SHARe): http://public.nhlbi.nih.gov/GeneticsGenomics/home/share.aspx

US National Genome Research Institute catalogue of published GWASs: http://www.genome.gov/gwastudies

Wellcome Trust Case–Control Consortium (WTCCC): http://www.wtccc.org.uk

ALL LINKS ARE ACTIVE IN THE ONLINE PDF