Risk Assessment for Data Sharing in the Type 2 Diabetes Knowledge Portal

Sean Simmons[1], Maria Costanzo[1], Noël Burtt[1], and Jason Flannick[1,2]
[1]Broad Institute, Cambridge, MA, USA
[2]Boston Children's Hospital, Boston, MA, USA

Introduction

As genotype, sequence, and genetic association data become increasingly easy to generate, and the scientific value of data sharing becomes more and more apparent, the number of publicly available resources that present these data is growing. For these resources, the privacy of the study participants who contribute the data is of paramount concern: measures must be taken to protect personal information about their genomes, health, and lifestyles.

The Type 2 Diabetes Knowledge Portal (T2DKP; type2diabetesgenetics.org), one such genomic resource, aggregates, integrates, interprets, and presents genetic data relevant to T2D, with the goal of sparking insights into human health. It is the product of the Accelerating Medicines Partnership in Type 2 Diabetes (AMP T2D), a public-private partnership including government, academic, and non-profit research institutions as well as pharmaceutical companies, that seeks to speed up the validation of new drug targets for T2D by sharing genetic data in a pre-competitive space. AMP T2D partners generate genetic association data and deposit them into the AMP Data Coordinating Center (DCC) at the Broad Institute; from there, these associations are made available for browsing, searching, and interactive analysis in the T2DKP.

Although other genomic resources provide some of the same data types, the T2DKP is unique in offering interactive tools that allow users to create custom subsets of individual-level datasets and run association analysis on those samples; associations for one SNP may be calculated multiple times, using differing subsets. The unprecedented amount of information provided by these unique features may, understandably, lead to concerns from funders and data submitters about whether this level of accessibility is compatible with appropriate privacy protection.

In order to address these concerns, we probabilistically model the genomic data used in the T2DKP. This model allows us to quantify the level of uncertainty that an outside observer (with access to various types of public or private information) has about the private genomic or phenotypic data hidden in the portal. The higher this uncertainty, quantified by a measure known as conditional entropy, the less private information is leaked. In addition, we have benchmarked the accuracy of methods (both known and novel) for inferring private information from genomic data. We have applied these approaches to a few, though far from all, possible uses of the T2DKP.

Privacy risks inherent in access to genetic association data: an overview

The privacy risks associated with presenting genetic data in a publicly accessible resource can be categorized into two major types. The first is the risk of linking disease status or a genotype to a real world identity—a person's name, along with identifying

information such as birth date or address. This risk is exemplified by the research of Latanya Sweeney, who has demonstrated that publicly available information in health records, voter registration databases, newspapers, and other sources can be used to reveal the medical records of individuals.

The second major type of potential risk is that of someone obtaining access to information (in particular, individual level information) that the study participants did not consent to sharing, even if it cannot be linked to a real world identity. For example, this could be a partial genotype, or the knowledge that a person with a particular set of physiological characteristics participated in a study.

Privacy risks inherent in presenting genetic association data via the T2DKP

The first major type of privacy risk—that individuals could be identified by name based on information obtained via the T2DKP—is not likely to be an issue at present. There are currently no known methods that could combine the information accessible in the T2DKP with that in existing public databases to link a person with their genotype or phenotype information.

However, the second type of risk—that some anonymous private information could be obtained—is a more realistic concern for the T2DKP. Below we consider the two types of access to genetic information that the T2DKP web pages and tools provide, access to summary-level information and interactive access to individual-level data, and the privacy risks associated with each.

The frequency with which a disease or trait occurs in the population also impacts the amount of information that could be revealed to an attacker. For a common disease such as T2D, or a common trait such as high blood pressure, learning that an individual with those characteristics participated in a study provides little if any information that could not already be deduced from common knowledge.

Risks in presenting summary statistics

Most of the interfaces and tools in the T2DKP present summary-level information about variants and their genetic associations. Existing methods that can leverage these data to learn private information require knowledge of a target's genotype beforehand, a high bar. As such, these statistics are generally considered to pose a very low privacy risk, and they are often made available on websites for public download by the consortia that generate them.

Summary statistics available in the T2DKP for a given variant may include parameters such as p-values and odds ratios or effect sizes for association with a given disease or trait; confidence interval; minor allele frequency (MAF); minor allele count in the whole dataset and in case and control groups; and sample size. They are searchable using the Variant Finder tool, which allows users to choose a phenotype and a dataset, set ranges for various parameters, and then retrieve a list of variants meeting those criteria.

Table 1 summarizes the privacy risks inherent in the presentation of summary statistics in the T2DKP, given certain kinds of background knowledge. If the attacker does not have access to the subject's genotype our model suggests that the conditional entropy is high, meaning that the privacy risk is low. On the other hand, if the attacker has genotype information, conditional entropy is low. However, if an attacker has access to the genotype information, a privacy breach has already occurred, and such a breach seems unlikely in most realistic scenarios. As such, we tend to consider the risk posed by this scenario as low.

Table 1. Privacy risks in presentation of summary-level data in the T2DKP.

With this outside knowledge…	Using this information from the T2DKP…	An attacker could theoretically find out…	Likely risk level
Background MAF **and** an individual's genotype at multiple SNPs	MAF	Whether the individual with that genotype participated in a study	Low
Background MAF **and** an individual's genotype at multiple SNPs	MAF	The disease status of the individual with that genotype	Low
Background MAF	GWAS statistics for multiple SNPs	The partial genotype of an (anonymous) individual	Low

Risks in offering interactive analysis of individual-level data

Individual-level data are indirectly accessible for analysis via the T2DKP. The raw data are currently stored in Data Coordinating Centers (DCCs) at either the Broad Institute or the T2DKP Federated node at EBI, and may be stored at additional federated locations in the future. These data comprise the genotypes or sequences of deidentified individuals, along with the health information collected for those individuals. The data may be used for analysis in two interactive tools in the T2DKP: LocusZoom and the Genetic Association Interactive Tool (GAIT). We focus here on GAIT, since it allows more versatile interactive analysis and thus poses greater potential privacy risks.

For a given dataset, the GAIT interface displays the distribution of many different phenotypes within the samples and allows the user to narrow the sample set by specifying the range of values for each phenotype (Figure 1) before performing association analysis. This allows researchers to detect associations that may only be significant in a particular subset of the experimental subjects, potentially providing important clues about the mechanisms by which variants affect disease.

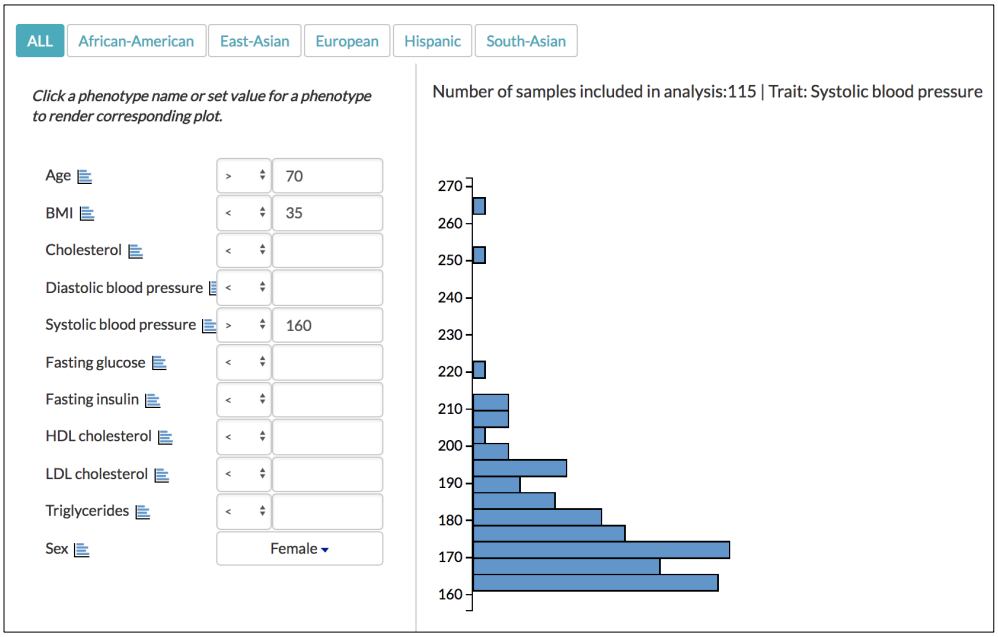

Figure 1. GAIT displays phenotype distributions across a sample set and allows users to filter samples by multiple phenotypic criteria before performing association analysis. In this example, the sample set has been narrowed down to 115 samples from individuals who are female, over 70 years old, with BMI under 35 and systolic blood pressure over 160. In the next step, a T2DKP user could perform association analysis for a particular variant and trait using this sample set.

Table 2 summarizes the privacy risks inherent in the interactive analysis offered by GAIT. Two types of information could theoretically be deduced from GAIT queries. First, since the interactive histograms offer an overview of the phenotypic characteristics of the sample set, repeated filtering could in theory enable an attacker to learn the phenotype of one individual in the study; however, the fact that GAIT does not allow the user to view a set of fewer than 100 individuals would make this difficult. It is extremely unlikely that these characteristics, most of which are recorded in private medical records, could be matched with a person's identity using data that are currently publicly available. If medical data become more widely available in the future, it is possible that this risk assessment will change.

Table 2. Privacy risks in interactive analysis of individual-level data using GAIT.

With this outside knowledge…	Using this information from the T2DKP…	An attacker could theoretically find out…	Likely risk level
None	GAIT's interactive histogram	The phenotypic profile of an (anonymous) individual in the study	Moderate
Background MAF	Results from numerous association analyses using GAIT	An (anonymous) individual's partial genotype	High

The second privacy risk inherent in GAIT is that the results of repeated interactive queries on slightly differing sample sets could be combined to learn about the genotype of an individual with a particular phenotype. Here, the conditional entropy is low and the risk is relatively high. But, as above, the likelihood of connecting this information with a person is very small using current approaches. In particular, approaches that attempt to infer last names using genomic data fail due to the lack of Y-chromosome information, while current methods for linking publicly available phenotypic traits to genotype data are not yet powerful enough to work at a population scale.

Currently, several measures are in place in the T2DKP to mitigate privacy risks. As mentioned above, GAIT does not allow users to create a set smaller than 100 individuals. All users must register and log in to access the site, agreeing to rules of conduct as they do. And all queries to the data are recorded in an audit log; while not routinely monitored, this could be checked in case of suspicious activity.

Conclusions

We feel that there is no need for concern that the privacy of study subjects could be breached by allowing summary statistics to be displayed in the T2DKP. Providing tools for interactive analysis of individual-level data does present some risk of revealing private information such as the traits possessed by an individual in a study or their partial genotype. However, the risk of an attacker being able to connect such information to an identifiable individual via the T2DKP is extremely low. This risk can be weighed against the potential rewards of enabling interactive analysis in the T2DKP: speeding up the identification of new T2D treatments.

We are aware that the types of information that can be either openly accessed or obtained by malicious attacks are continually changing. The tools and interfaces offered by the T2DKP are constantly evolving as well. We will continue to actively investigate potential privacy risks inherent in the T2DKP and Knowledge Portal platform along with methods to mitigate those risks.

Acknowledgments

The Knowledge Portal platform was built by our team at the Broad Institute in collaboration with colleagues in the Accelerating Medicines Partnership in Type 2 Diabetes (AMP T2D). We thank AMP T2D for supporting the creation of the T2DKP and the KP platform.