Evaluating the Strength of Genomic Privacy Metrics

ISABEL WAGNER, De Montfort University

The genome is a unique identifier for human individuals. The genome also contains highly sensitive information, creating a high potential for misuse of genomic data (for example, genetic discrimination). In this paper, we investigate how genomic privacy can be measured in scenarios where an adversary aims to infer a person's genomic markers by constructing probability distributions on the values of genetic variations. We measured the strength of privacy metrics by requiring that metrics are monotonic with increasing adversary strength and uncovered serious problems with several existing metrics currently used to measure genomic privacy. We provide suggestions on metric selection, interpretation, and visualization, and illustrate the work flow using case studies for three real-world diseases.

CCS Concepts:•**General and reference** → **Metrics;**•**Security and privacy** → *Privacy-preserving protocols; Privacy protections;*•**Human-centered computing** → *Visualization; Heat maps;*•**Applied computing** → *Genomics;*

Additional Key Words and Phrases: privacy metrics, genomic privacy, ARCHER

1. INTRODUCTION

In 2001, Celera, Inc was the first to sequence a full human genome at a cost of about 300 million dollars. At the time of this writing, full genome sequences can be obtained at a cost of little more than $1,000 per genome [Wetterstrand 2016]. This has enabled a dramatic increase in the use of genomic data in health care (e.g., personalized medicine and pharmacogenomics [Fredrikson et al. 2014]), research (e.g., genome-wide association studies that correlate the appearance of diseases with specific locations in the genome [Welter et al. 2014]), and forensics (e.g., paternity tests [Naveed et al. 2015]). Unfortunately, the wide availability of genomic data also raises important privacy concerns, because a genome sequence uniquely identifies an individual. Possible violations of genomic privacy range from the re-identification of anonymous participants in genome-wide association studies (revealing a person's disease status [Homer et al. 2008]) to genetic discrimination (for example, denial of insurance because of genetic predisposition [Gottlieb 2001]). Moreover, because related individuals have similar genomes, sensitive information can be inferred not only about an individual but also about her/his kin [Humbert et al. 2013]. Despite these privacy concerns, currently, there is a lack of methods to measure how private a particular genomic technology is (i.e. genomic privacy metrics). As a result, technologies that preserve genomic privacy are still in their infancy.

In this paper, we investigate the strength of genomic privacy metrics. We consider an adversary who targets an individual and aims to infer the target's genome sequence,

The work described in this article is an extension of work described in [Wagner 2015], with the addition of the formalization of the monotonicity criterion, consideration of more scenarios and adversary models (Section 5), and new case studies (Section 6).

Author's addresses: I. Wagner, De Montfort University, The Gateway, Leicester, LE1 9BH, United Kingdom, isabel.wagner@dmu.ac.uk

either in its entirety, or focusing on specific regions of interest. For example, the adversary may be located in a medical unit responsible for conducting genetic tests, or in a biobank that stores genomic sequences (for example a disgruntled employee or an intruder who hacked into the respective systems) [Ayday et al. 013b]. The real-world importance of this scenario will increase with the increasing use of genetic information in routine medical care. In particular, routinely sequencing a patient's genome [Erlich and Narayanan 2014] will increase the availability of (1) genetic information (which may be stored in encrypted form, but may leak during usage [Ayday et al. 2014]) and (2) outcomes of genetic tests (which the adversary can combine with knowledge about the tests to reconstruct the genomic sequence [Goodrich 2009]). We assume that the adversary uses an inference attack to compute a probability distribution for each variation in the target genome. This is a reasonable assumption, because several inference attacks have already been described, for example exploiting linkage disequilibrium [Ayday et al. 013a; Nyholt et al. 2009], exploiting information from kin genomes [Humbert et al. 2013; 2014], exploiting systematic execution of genomic tests [Goodrich 2009], and using statistical information about individuals who participated in genome-wide association studies [Wang et al. 2009]. We note that we are not focusing on any one specific attack. Instead, we consider all attacks that could leak genomic data, including attacks that may be invented in the future. Even though the detailed steps and workings of future attacks are not yet known, it can be assumed that a strong attack will allow the adversary to make highly accurate inferences. We represent this assumption using series of probability distributions that represent attacks of different strength. Apart from this assumption, we do not impose restrictions on who the adversary may be, how they perform the attack, and what information they use.

Contributions. Our three main contributions are: (1) We define a method that allows to evaluate privacy metrics systematically using an ordered sequence of adversaries with different strengths. Adversary strength was measured by how close their inferences of genomic variants were to the true value. (2) We formalize monotonicity as the key indicator of a metric's strength, i.e. we require that metrics show decreasing privacy for increasing adversary strength. To the best of our knowledge, we are the first to formally define a criterion for the strength of privacy metrics. (3) We tested 24 privacy metrics for genomic privacy in four possible attack scenarios: (i) a comparative evaluation with a large number of individuals, (ii) an evaluation of kin privacy considering only related individuals, (iii) an evaluation focusing on risk factors for three real-world conditions (asthma, multiple sclerosis, and vitamin B12 levels), and (iv) an evaluation studying the influence of an individual's population group on the strength of privacy metrics.

Of the metrics we tested, we found that only 7 out of 24 metrics were strong across adversary types and scenarios and could be interpreted easily: the adversary's success rate, the amount of information leaked, health privacy (with information surprisal or relative entropy as base metric), information surprisal, percentage incorrectly classified, relative entropy, and user-specified innocence. Furthermore, we find that none of the metrics we tested are sufficiently reliable when used by themselves. Therefore, we recommend to combine multiple strong metrics to gain insight on as many different aspects of privacy as possible.

Our systematic comparison of genomic privacy metrics enables researchers, clinicians, and policy-makers to make an informed choice about the selection of privacy metrics and privacy-enhancing technologies. In addition, we show how visualization methods from data science, namely heat maps and radar plots, can be applied to privacy metrics to help ensure that new privacy enhancing technologies are evaluated in a consistent and comparable manner.

2. BACKGROUND

2.1. Genomics

Although the human genome consists of about three billion DNA base pairs, genomes from two human individuals differ only in about 0.2–0.4% of base pairs [Tishkoff and Kidd 2004]. Most commonly, this genetic variation comes from differences in single bases, called single nucleotide polymorphisms (SNPs, pronounced *snips*) [Sachidanandam et al. 2001]. In most cases, a SNP has only two variants (alleles) in the human population. Usually, of the two SNP alleles one is more common than the other (called the major allele, A, and the minor allele, a, respectively). Because the genome of a somatic human cell is diploid, that is, it is comprised of two sets of chromosomes – one set inherited from the father, and the other set inherited from the mother – each SNP is present in two copies. Therefore, the genotype of a given SNP can be encoded as 0, 1, or 2 corresponding to the allele combinations AA, Aa, and aa [Humbert et al. 2013]. Population-wide frequencies of alleles A and a can be estimated from a sample of human genomes; this has been done in the 1000 Genomes project[1]. Alleles at different locations in the genome can be correlated, especially when their locations are close to each other. This non-random association between SNPs is called linkage disequilibrium (LD) [Slatkin 2008]. The strength of LD between pairs of SNPs is commonly expressed using the correlation coefficient r^2. Genome-wide association studies can identify SNPs associated with diseases by comparing the incidence of SNP variations between individuals who do and do not have a particular disease [Welter et al. 2014].

2.2. Privacy Metrics

Many privacy metrics have been proposed for different domains [Wagner and Eckhoff 2015]. However, many studies have shown their shortcomings, for example inconsistent metrics [Wagner 2015], metrics that are hard to understand [Diaz et al. 2007], and metrics that work only in narrow scenarios [Kalogridis et al. 2010]. This is problematic, because use of a weak privacy metric can lead to an overestimation of privacy and result in privacy violations, for example the re-identification of individuals in published health data, thus linking individuals to their medical conditions [Sweeney 2002]. Privacy metrics that suffer none of these shortcomings can still be weak if used on their own because some metrics are complementary – they measure different aspects of privacy and thus need to be used in combination to form a more complete measurement of privacy [Murdoch 2014; Liu and Mittal 2016]. These shortcomings show that existing privacy metrics exhibit a lack of consistency, reproducibility, and wider applicability. However, it is unknown which privacy metrics, and in which application domains, produce consistently good measurements of privacy. This can not only impede and slow down privacy research [He et al. 2015; Shokri et al. 2011; Murdoch 2014], but also lead to real-world privacy violations [Sweeney 2002], and has recently led to calls for research on privacy metrics [He et al. 2015; Shokri et al. 2011; Murdoch 2014].

2.3. Genomic privacy metrics

Broadly, privacy metrics measure characteristics of privacy enhancing technologies and quantify how much privacy a technology offers [Clauß and Schiffner 2006], for example, the adversary's probability to break a user's anonymity [Serjantov and Danezis 2002], or the maximum amount of bits of private information an adversary can infer [Diaz et al. 2003]. In the context of genomic privacy, most research applies existing privacy metrics to genomic privacy scenarios [Ayday et al. 2014; Humbert et al. 2013; Ayday et al. 013a; Samani et al. 2015]. Some researchers also propose new metrics

[1] http://www.1000genomes.org/

specific to genomic privacy [Ayday et al. 013a; Ayday et al. 013b; Humbert et al. 2013]. These papers generally propose or describe one or more metrics, and then use these metrics to evaluate a privacy enhancing technology in a given scenario. However, they do not evaluate the strengths of the metrics, or how they differ from other metrics. This paper aims to address this gap. The closest to our work is [Murdoch 2014], which investigates the behavior of anonymity metrics, among them entropy and some of its variations. In previous work, we have published an initial evaluation of metrics for genomic privacy [Wagner 2015].

2.4. Requirements for genomic privacy metrics

Traditionally, a strong privacy metric is one that can (1) indicate, in terms understandable to lay people, how effectively the adversary can succeed [Alexander and Smith 2003]; (2) show both the privacy level and the portion of data not protected [Bertino et al. 2008]; (3) consider accuracy, uncertainty, and correctness as three aspects of the adversary's success [Shokri et al. 2011]; and (4) indicate not only the difficulty for the adversary, but also the amount of resources he needs to succeed [Syverson 2013]. Most of these criteria apply to specific privacy metrics, but cannot be used to compare the strengths of different metrics.

In this work, we introduce a new criterion for strong privacy metrics – monotonicity, which requires privacy metrics to show decreasing privacy for increasing adversary strength (Section 5). Because monotonicity can be quantified, we believe that it can be used to compare the strengths of privacy metrics. Furthermore, we rate understandability based on the results of our case studies (Section 6).

3. PRIVACY METRICS

From our previous survey of privacy metrics [Wagner and Eckhoff 2015], we selected 24 metrics that were applicable to our genomic privacy scenario. The metrics are summarized in Table II, and Table I provides a reference for notation used. Ten metrics have previously been applied in genomic privacy; the remaining metrics have been drawn from the wider privacy literature (see the *Genomics Precedent* column in Table II). The metrics can be grouped into per-SNP metrics that compute values for each SNP separately, and per-individual metrics that compute an aggregate value for all of an individual's SNPs (see the *per SNP* column).

3.1. Excluded Metrics

We excluded a range of privacy metrics that did not fit our assumptions.

Differential privacy [Dwork 2006] offers privacy guarantees for database queries. However, our scenario assumes that the adversary is already one step further in that he has already acquired a probability distribution on the target's genotypes. While differential privacy will not help evaluate privacy in our scenario, it could be used to prevent the adversary from acquiring a probability distribution in the first place.

k-anonymity [Malin 2005] states that an individual cannot be distinguished among at least $k - 1$ other individuals. Since we assume that the adversary already knows the target individual, we know that $k = 1$, and so this metric does not help us analyze privacy further.

3.2. Included Metrics

We group our description of included metrics by the output they measure, according to the taxonomy proposed in [Wagner and Eckhoff 2015]. The notation, summarized in Table I, is the same for each metric.

Table I. Notation

$k \in \{0, 1, 2\}$	Possible genotypes for each SNP
x_i	Estimated genotype of SNP i
y_i	True genotype of SNP i
$p(x_i = y_i)$	Probability to guess true genotype of SNP i correctly
$p(x_i = k)$	Adversary's estimate for the case that SNP i has genotype k
r_i	Minor allele frequency of SNP i
α	Threshold for adversary's probabilities

3.2.1. Metrics Measuring the Adversary's Error. The *expected estimation error* quantifies the adversary's correctness by computing the expected distance between the adversary's estimate and the true genotype for every SNP [Humbert et al. 2013]. In the context of genomics, this distance is computed on the encoded genotypes. Therefore, we have to ensure that the SNP encoding has a meaningful genomics interpretation. For example, the encoding proposed by Humbert et al. [2013] is meaningful, because the encoded value 1 (one each of major and minor allele) lies between 0 (two major alleles) and 2 (two minor alleles). This metric may behave differently with a different encoding.

$$priv_{\text{EEE}} = \sum_{k \in \{0,1,2\}} p(x_i = k)||k - y_i||$$

The *mean squared error* is computed as the squared difference between the true genotype and the adversary's estimate, averaged over all SNPs [Oya et al. 2014].

$$priv_{\text{MSE}} = \frac{1}{|\text{SNPs}|} \sum_{x_i \in \text{SNPs}} \{\|x_i - y_i\|^2\}$$

Other variations of the adversary's error are the *mean error* [Samani et al. 2015] and the *mean error with normalized distance* [Humbert et al. 2015].

Percentage incorrectly classified measures how often the highest probability in the adversary's estimate does not correspond to true genotype [Narayanan and Shmatikov 2009].

$$priv_{\text{PIC}} = \frac{|\text{incorrect SNPs}|}{|\text{SNPs}|}$$

3.2.2. Metrics Measuring the Adversary's Uncertainty. *Entropy* quantifies the amount of information contained in a random variable. Used as a privacy metric, it indicates the adversary's uncertainty [Serjantov and Danezis 2002].

$$priv_{\text{ENT}} = H(X_i) = -\sum_{k \in \{0,1,2\}} p(x_i = k) \log_2 p(x_i = k)$$

Entropy can be normalized to a range of $[0, 1]$ by dividing it by Hartley entropy, that is, the logarithm of the number of outcomes [Humbert et al. 2013].

$$priv_{\text{NE}} = \frac{H(X_i)}{H_0(X_i)}$$

Hartley entropy, or max-entropy, has also been used as a privacy metric [Clauß and Schiffner 2006]. It is an optimistic metric because it only accounts for the number of outcomes, but not for additional information the adversary may have. In the context of genomics, however, the number of outcomes per SNP is known to be 3, and therefore max-entropy is not useful and has been excluded from the evaluation.

$$priv_{\text{MXE}} = H_0(X_i) = \log_2 |x_i| = \log_2 3$$

Table II. Privacy Metrics

Metric	per SNP	Genomics Precedent	Inputs	H/L[2]	Priv. Level[4]	Intuitiveness
Adversary's success rate	–	✓	estimate, truth	L	++	++
Amount of information leaked	–	✓	estimate, truth, α	L	++	++
Asymmetric entropy	–	✓	estimate, truth, prior	H	o	–
Asymmetric entropy (per SNP)	–	✓	estimate, truth, prior	H	o	–
Coefficient of determination r^2	–	–	estimate, truth	L	–	o
Conditional entropy	✓	–	estimate, truth	H	+	–
Conditional privacy loss	✓	–	estimate, truth	L	+	–
Cumulative entropy	–	–	estimate	H	+	+
Entropy $H(X_i)$	✓	–	estimate	H	+	+
Expected estimation error	✓	✓	estimate, truth	H	++	o
Genomic privacy	–	✓	estimate, truth, W_i	H	++	–
Health privacy	–	✓	base metric, c_i	H/L	+/++ [3]	+ [3]
Information surprisal	✓	–	estimate, truth	H	++	+
Inherent privacy	✓	–	estimate	H	+	–
Max-entropy $H_0(X_i)$	–	–	estimate	H	–	–
Mean error	–	✓	estimate, truth	H	++	o
Mean squared error	–	–	estimate, truth	H	++	o
Min-entropy $H_\infty(X_i)$	✓	–	estimate	H	+	o
Mutual information	✓	–	estimate, truth	L	+	o
Normalized entropy	✓	✓	estimate	H	+	+
Normalized mutual inf.	✓	✓	estimate, truth	H	+	o
Perc. incorrectly classified	–	–	estimate, truth	H	++	++
Relative entropy	✓	–	estimate, truth	H	++	+
User-specified innocence	–	–	estimate, truth, α	H	++	++
Variation of information	✓	–	estimate, truth	L	–	o

[2] high (H) or low (L) values indicate high privacy

[3] Provided a good/very good (+/++) base metric is used

[4] Metric strength ≤ 30: –; $\in]30, 70[$: o, $\in [70, 90]$: +, > 90: ++

Min-entropy is a pessimistic metric because it is based only on the probability of the most likely outcome, regardless of whether this is also the true outcome [Clauß and Schiffner 2006]. Min-entropy is a conservative measure of how certain the adversary is of his estimate.

$$priv_{\text{MNE}} = H_\infty(X_i) = -\log_2 \max p(x_i)$$

Cumulative entropy is based on the notion that the adversary's uncertainty increases when privacy protection is applied at several independent points. Cumulative entropy is computed as the sum of individual entropies [Freudiger et al. 2007]. In the context of genomics, we sum over the entropies computed for each SNP.

$$priv_{\text{CE}} = \sum_{i=1}^{|\text{SNPs}|} H(X_i)$$

Conditional entropy, or the entropy of Y conditioned on X, measures the amount of information needed to fully describe Y, provided that X is known [Diaz et al. 2007]. For genomic privacy, Y can be chosen as the true genotype and X as the adversary's estimate. This measures how much more information the adversary needs to find the true value.

$$priv_{\text{COE}} = H(Y_i|X_i) = H(Y_i) - I(Y_i; X_i)$$

Inherent privacy [Agrawal and Aggarwal 2001; Andersson and Lundin 2008] and *conditional privacy* [Andersson and Lundin 2008] are derivations of base metrics (entropy and conditional entropy, respectively), each computed as $2^{\text{base metric}}$. While the base metrics are interpreted as bits of information, these metrics can be interpreted as the number of binary questions an adversary has to ask to resolve his uncertainty.

$$priv_{\text{IP}} = 2^{H(X_i)}\ ,\ priv_{\text{CP}} = 2^{H(Y_i|X_i)}$$

Asymmetric entropy is another measure for the adversary's uncertainty. It is tailored to genomics because it assumes that the adversary's estimate is based on population-wide minor allele frequencies, which results in a different maximum value for entropy for each SNP [Ayday et al. 013b].

$$priv_{\text{AE}} = \sum_{i=1}^{|\text{SNPs}|} \frac{p(x_i = y_i)(1 - p(x_i = y_i))}{(-2w_i + 1)p(x_i = y_i) + w_i^2}, \text{ where } w_i = \begin{cases} (1 - r_i)^2 & \text{if } y_i = 0 \\ 2r_i(1 - r_i) & \text{if } y_i = 1 \\ r_i^2 & \text{if } y_i = 2 \end{cases}$$

Asymmetric entropy can also be used as a per-SNP metric to measure privacy for individual SNPs.

3.2.3. Metrics Measuring Information Gain/Loss. The *amount of leaked information* [Wang et al. 2009; Ayday et al. 2014] counts the number of leaked SNPs. A SNP is considered leaked when the adversary's estimate for the true outcome is above the threshold α. A threshold of 1 means that a SNP is considered leaked only if the adversary is absolutely certain. Many scenarios will adopt a more conservative threshold to cover situations when the adversary is reasonably, but not absolutely, certain.

$$priv_{\text{ALI}} = |u| \text{ so that } \forall u_i \in \text{SNPs} : p(u_i = y_i) > \alpha$$

Information surprisal, or self-information, quantifies how much information is contained in a specific outcome of a random variable [Chen et al. 2013]. In the context of genomics, the outcome is the true value of a SNP, and the information content is the probability the adversary assigns to this outcome. Informally, information surprisal quantifies how surprised the adversary would be upon learning the true value of a SNP.

$$priv_{\text{IS}} = -\log_2 p(x_i = y_i)$$

Genomic privacy [Ayday et al. 013a] uses the adversary's estimate for those outcomes when the SNP is present as a variation, i.e. with one or two copies of its minor allele. Each SNP is weighted with a severity W_i which can be derived from scientific studies or tables provided by insurance companies, and all SNPs are then summed up to a per-individual metric. Its value depends strongly on the number of SNPs studied for each individual, and could thus be normalized to make it more comparable.

$$priv_{\text{GP}} = -\sum_{i \in \text{SNPs}} \log_2(p(x_i = 1) + p(x_i = 2)) \cdot W_i, \text{ where SNPs} = \{j | y_j = 1 \wedge y_j = 2\}$$

Mutual information measures how much information is shared between two random variables Y and X [Lin et al. 2002]. As before, Y can be chosen as the true genotype and X as the adversary's estimate.

$$priv_{\text{MI}} = I(Y_i; X_i) = H(Y_i) - H(Y_i|X_i)$$

Normalized mutual information can use either Shannon entropy [Zhu and Bettati 2005] or Hartley entropy [Humbert et al. 2013]. In this paper we use the latter.

$$priv_{\text{NMI}} = 1 - \frac{I(Y_i; X_i)}{H_0(X_i)}$$

Conditional privacy loss [Andersson and Lundin 2008] is derived from mutual information. While mutual information is interpreted as the bits of information shared between the true value and the adversary's estimate, conditional privacy loss can be interpreted as the number of binary questions an adversary has to ask to arrive at the true value.

$$priv_{\text{CPL}} = 1 - 2^{-I(Y_i;X_i)}$$

The *relative entropy*, or Kullback-Leibler divergence, between two random variables Y and X measures the information that is lost when X is used to approximate Y [Deng et al. 2007]. In the context of genomics, good choices for Y and X are the true value and the adversary's estimate, respectively. This measures how many additional bits of information the adversary needs to reconstruct the true value.

$$priv_{\text{RE}} = \sum_{k \in \{0,1,2\}} p(y_i = k) \log_2 \frac{p(y_i = k)}{p(x_i = k)}$$

Variation of information is derived from mutual information so that it fulfills the conditions for a distance metric in the mathematical sense, especially the triangle inequality [Meilă 2007]. It describes the distance between two random variables, chosen as the true value and the adversary's estimate.

$$priv_{\text{VI}} = H(X_i) + H(Y_i) - 2I(Y_i; X_i)$$

3.2.4. Metrics Measuring the Adversary's Success Probability. The *adversary's success rate* captures how likely it is for the adversary to succeed. In the context of genomics, we can define success on a per-SNP basis as the probability of correctly inferring a genotype, and aggregate to a per-individual metric by computing the average probability for all SNPs [Ayday et al. 013a].

$$priv_{\text{ASR}} = \frac{1}{|\text{SNPs}|} \sum_{i \in \text{SNPs}} p(x_i = y_i)$$

User-specified innocence can be seen as a counterpart to the amount of leaked information, because it counts the number of SNPs that remain private [Chen and Pang 2012]. A SNP is considered private if the adversary's estimate for the true outcome is below the threshold α. A threshold of 0 means that a SNP is considered private only if the adversary considers it impossible. Many scenarios will therefore adopt a higher threshold.

$$priv_{\text{USI}} = |u| \text{ so that } \forall u_i \in \text{SNPs} : p(u_i = y_i) \leq \alpha$$

3.2.5. Metrics Measuring Similarity/Diversity. The *coefficient of determination* r^2 describes how well a statistical model approximates data. It is typically used for linear regression where a value of 1 indicates a perfect fit [Kalogridis et al. 2010]. In the context of genomics, the adversary's estimate can be used as statistical model, and the true genotypes represent the data.

$$priv_{\text{R2}} = 1 - \frac{SS_E}{SS_R + SS_E}, \text{ where } SS_E = \sum_i (y_i - x_i)^2, SS_R = \sum_i (x_i - \bar{Y})^2$$

3.2.6. Other Metrics. *Health privacy* focuses on those SNPs known to contribute to a specific disease. Health privacy uses a base metric to compute per-SNP values, and then aggregates to a per-individual metric using a weighted and normalized sum [Humbert et al. 2013]. The weights c_i should be chosen to reflect how much each SNP contributes to the disease. Base metrics discussed in [Humbert et al. 2013] are the expected estimation error, normalized entropy, and normalized mutual information. We extend this list and also investigate relative entropy, conditional entropy, information surprisal, and min-entropy as base metrics.

$$priv_{\text{HP}} = \frac{1}{\sum_{i \in S} c_i} \sum_{i \in S} c_i G_i, \text{ where } G_i \text{ is a per-SNP base metric}$$

4. INITIAL EVALUATION

4.1. Data Sources

We used two publicly available data sources for our initial evaluation. First, we downloaded genomic data from 1857 individuals from openSNP [Greshake et al. 2014]. This dataset consists of genomic data that users acquired from 23andme[5] and FamilyTreeDNA[6] and published on openSNP. On average, each user has data about 730k SNPs. This data serves as ground truth information for all metrics that rely on it (see Table II, column *Inputs*).

Second, we downloaded minor allele frequencies from the Database of Single Nucleotide Polymorphisms (dbSNP) [Sherry et al. 2001]. The minor allele frequencies in this dataset are computed from a sample global population consisting of 1000 genomes. We used minor allele frequencies to construct the *reference* adversary estimate, and for the computation of asymmetric entropy.

4.2. Adversary Models

Our adversary models abstract from the strategies and algorithms a real-world adversary would use, and instead represent the strength of an adversary using probability distributions. For our initial evaluation, we use two different types of adversary estimate. The *reference* model uses the population-wide distribution of minor allele frequencies taken from the dbSNP. Following the Hardy-Weinberg principle and denoting the minor allele frequency as q, the adversary assigns probabilities depending on the number of minor alleles for each SNP: for two minor alleles $p(aa) = q^2$, for two major alleles $p(AA) = (1-q)^2$, and for one each of major and minor allele $p(Aa) = 2q(1-q)$.

The *normal* model uses a series of normal distributions with a small standard deviation ($\sigma = 0.1$, chosen to allow clear distinction between the adversary strength levels), truncated to the $[0,1]$ range, to represent the probability that the adversary assigns to the true genotype. We study six strength levels with mean probabilities $\mu = 0.1, 0.25, 0.4, 0.6, 0.75, 0.9$. Figure 2d shows the average probability the *normal* adversary assigns to the true genotype.

Intuitively, we expect that privacy is higher if the adversary's guesses are far from the truth, and lower if his guesses are close. For the reference estimate, we expect that the adversary's guesses are close to the truth in many cases, because the estimates are chosen to match the majority of the population. An adversary using the reference estimate is very realistic since minor allele frequencies are easy to obtain. It is therefore important to find protection mechanisms that are effective against this kind of adversary.

4.3. Results

To get a high-level overview of how the 24 metrics behave, we computed their values using all genomes in our dataset, but only 10000 SNPs each[7]. We used fixed parameter

[5] https://www.23andme.com/

[6] https://www.familytreedna.com/

[7] We also evaluated the metrics using fewer genomes, but all SNPs for each. The results were very similar, which is why we report our results using the computationally much less demanding scenario with 10000 SNPs per genome. Even so, the results may depend on the dataset, and we make some comparisons with other datasets in Section 5.

values for the three metrics that have parameters. The severities for genomic privacy were sampled from the uniform distribution between 0 and 1. For health privacy, we used 1000 SNPs with equal weights, and the expected estimation error as base metric (we study other parameter settings for health privacy in Section 5.5.1). We set the threshold for amount of information leaked to 0.7 and for user-specified innocence to 0.3, so that the two metrics can distinguish most adversary strength levels (we study variation of the threshold parameter in Section 5.5.2). We computed 15 replications to make sure the results are not due to random variations in the adversary estimate, and computed confidence intervals for the mean. The relative errors for these confidence intervals were below 5% in all cases, indicating that we performed enough replications to achieve highly precise results. We implemented our computations in Python, using SciPy[8] for the entropy-based metrics and scikit-learn[9] for metrics based on mutual information[10].

Figure 1 shows the results. For each privacy metric and adversary strength level, we plot one vertical violin plot [Hintze and Nelson 1998]. The vertical bar shows the range of the data, with horizontal lines indicating the minimum, mean, and maximum. In addition, a kernel density plot on each side of the bar indicates the probability density. The six violins on the left represent the *normal* adversaries with estimates ranging from far to close to the truth. The right-most violin represents the *reference* adversary. Each of the vertical violins aggregates the results for all SNPs, all individuals, and all replications we performed for each metric and each adversary strength level. To illustrate the statistical distribution of the bulk of metric values, we added the median as well as the first and third quartile to the plot. We fitted cubic splines to the medians and quartiles to emphasize how their values change depending on adversary strength. We plot the median spline as a black line, and shade the area between the quartile splines. In addition, we print the value of the mean in boldface at the top of each violin. We do not plot the confidence intervals since they are so narrow that the lower and upper bounds would collapse to a single line on top of the mean.

The most important requirement we look for in a privacy metric is a consistent representation of the privacy level. Privacy should be high for a weak adversary, and decrease with increasing adversary strength, i.e. from left to right in the plots.

4.3.1. Metrics Measuring the Adversary's Error. The expected estimation error (Figure 1a) does not show a big difference between adversary strengths, mainly because the range of values is relatively large compared to where the bulk of the values lie. However, it can be seen that the mean is decreasing with increasing adversary strength. This becomes much more evident when the expected estimation error (a per-SNP metric) is aggregated to a per-individual metric, for example when it is used as a base metric for health privacy (Figure 1b). In this case, the decrease in value is much more pronounced (we investigate different base metrics for health privacy in Section 5.5).

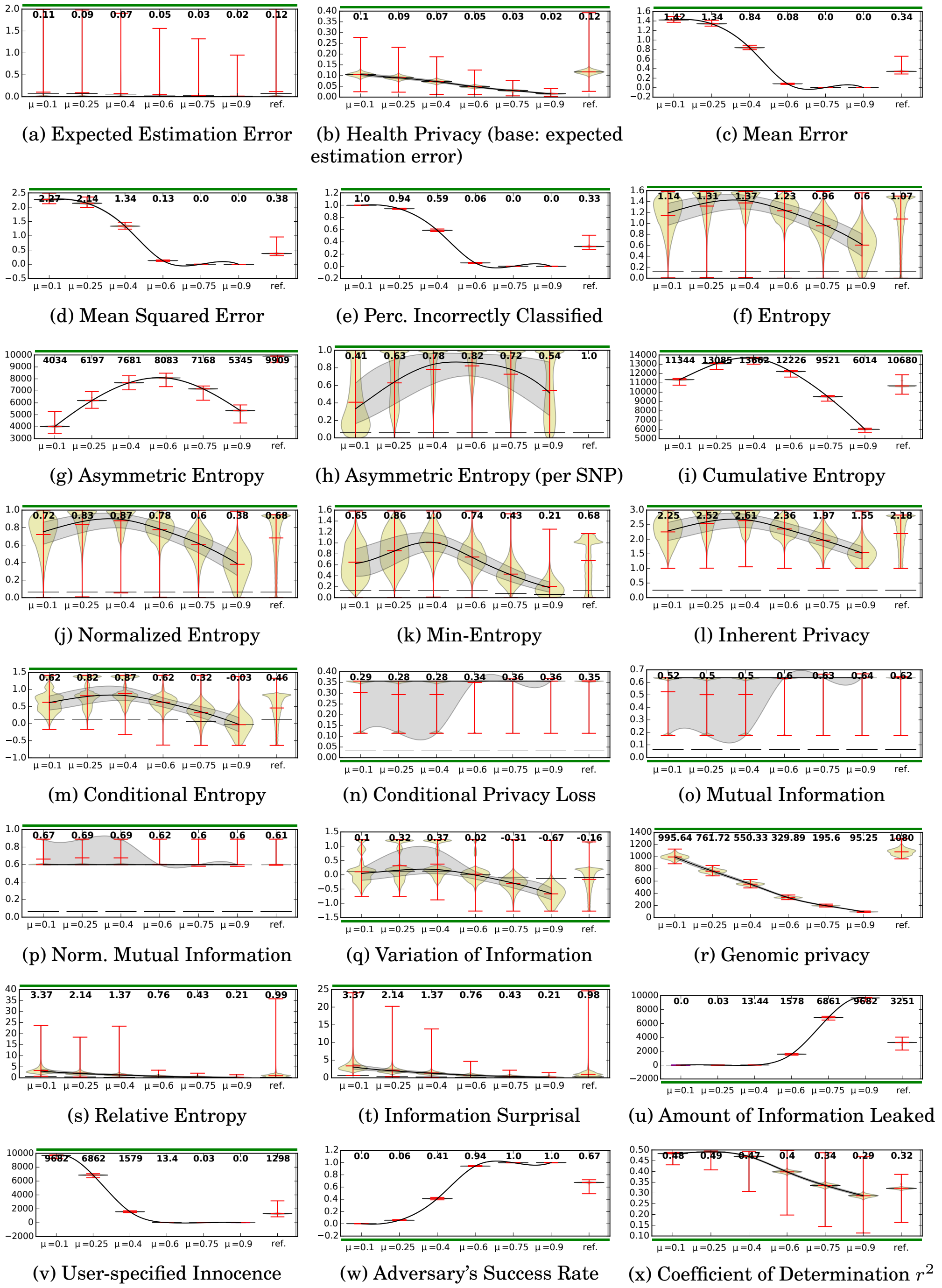
The values for mean error, mean squared error, and percentage incorrectly classified (Figures 1c, 1d, and 1e) are decreasing, but only for the weaker four adversaries. The strongest two adversaries cannot be distinguished.

4.3.2. Metrics Measuring the Adversary's Uncertainty. The entropy-based metrics in Figures 1f–1m peak for medium-strength adversaries and assign similar values to strong and weak adversaries. To explain this, we recall that entropy measures the uncertainty in a random variable. Because of the way we defined the *normal* adversary model, medium-strength adversaries appear more uncertain than adversaries on either end.

[8] http://www.scipy.org/

[9] http://scikit-learn.org

[10] We are working on open sourcing our code. In the meantime, we are happy to share the code on request.



(a) Expected Estimation Error

(b) Health Privacy (base: expected estimation error)

(c) Mean Error

(d) Mean Squared Error

(e) Perc. Incorrectly Classified

(f) Entropy

(g) Asymmetric Entropy

(h) Asymmetric Entropy (per SNP)

(i) Cumulative Entropy

(j) Normalized Entropy

(k) Min-Entropy

(l) Inherent Privacy

(m) Conditional Entropy

(n) Conditional Privacy Loss

(o) Mutual Information

(p) Norm. Mutual Information

(q) Variation of Information

(r) Genomic privacy

(s) Relative Entropy

(t) Information Surprisal

(u) Amount of Information Leaked

(v) User-specified Innocence

(w) Adversary's Success Rate

(x) Coefficient of Determination r^2

Fig. 1. Privacy metrics evaluated according to adversary strength, ordered weakest to strongest from left to right. Green bars show which values indicate high privacy: above the plot (high values = high privacy) or below the plot (low values = high privacy).

While it is certainly good for privacy if the adversary is uncertain, uncertainty alone is not an accurate representation of a user's privacy level.

4.3.3. Metrics Measuring Information Gain/Loss. Mutual information and the metrics derived from it (conditional privacy loss and variation of information) show a similar behavior to the entropy-based metrics, albeit reversed and less pronounced (observe the horizontal bars for the mean in Figures 1n–1q).

Relative entropy and information surprisal, shown in Figures 1s and 1t, are the only two information theoretic metrics that behave as we would expect. Their values decrease with increasing adversary strength.

Genomic privacy (Figure 1r) has decreasing values for increasing adversary strengths and allows to distinguish all strength levels. The values depend on the number of SNPs and the severity for each SNP. Counter to our intuition for the reference adversary, the genomic privacy metric assigns a high value (corresponding to high privacy) to the reference adversary.

The amount of information leaked (Figure 1u) and user-specified innocence (Figure 1v) show the same situation from two different angles: information that is considered leaked versus information that remained private. With the parameter setting we used in this experiment, each metric can only distinguish between five of the six adversaries; the values for the weakest resp. strongest two adversaries are zero. In the other cases, the metrics behave as we expect, with increasing values for information leaked, and decreasing values for user-specified innocence. The value range for these two metrics depends on the number of SNPs in the study; the maximum value of 10000 corresponds to the number of SNPs we investigated. It would thus be easy to normalize these metrics to a range of $[0, 1]$ by dividing by the number of SNPs. The amount of information leaked is the only metric that explicitly counts the number of information items (SNPs) not hidden by a privacy enhancing technology.

4.3.4. Metrics Measuring the Adversary's Success Probability. The adversary's success rate (Figure 1w) increases with the adversary's strength, allowing to distinguish five of the six adversaries. The two strongest adversaries cannot be distinguished because we count a success if the estimate with the highest probability corresponds to the true genotype, regardless of how high this probability is. Because we defined the adversary's success as the exact opposite of incorrect classification, the percentage incorrectly classified (Figure 1e) is a mirror image of the adversary's success rate and conveys exactly the same information.

4.3.5. Metrics Measuring Similarity/Diversity. The values of the coefficient of determination are decreasing for most adversary strengths, as shown in Figure 1x. However, we expect otherwise: the lowest privacy level – a perfect fit between the adversary's estimate and the true outcome – should be indicated by higher values of the coefficient of determination. Figure 1x shows the reverse behavior. The coefficient of determination does therefore not give a correct estimation of a user's privacy level.

Regarding the performance of the reference adversary, we can see that most metrics place it in the middle of our adversary-strength spectrum. Notable exceptions are the expected estimation error, health privacy, and genomic privacy which place the reference estimate among the weakest adversaries.

5. EXTENDED EVALUATION

We then extended our initial evaluation to address a number of open issues: (1) how do the genomic privacy metrics behave for datasets with different characteristics? (2) how do the genomic privacy metrics behave for different adversary models? (3) how can the results be presented in a more compact and user-friendly way? (4) how do the parameter settings influence the metrics' behaviors?

5.1. Data Sources and Definition of Scenarios

We retained the two datasets from the initial evaluation (openSNP and dbSNP) and added two datasets (CEPH/Utah Pedigree 1463 and 1000 Genomes Project) to study the influence of relationships between individuals, population groups, and SNPs associated with specific diseases.

5.1.1. Comparison Scenario. In the *comparison* scenario, we evaluate all 24 privacy metrics using 10.000 SNPs from all openSNP genomes.

5.1.2. Kin Scenarios. To study how relationships between individuals influence the strength of privacy metrics, we identified 13 pairs of blood relatives in the openSNP data[11] and added a dataset with verified blood relatives, the CEPH/Utah Pedigree 1463 [Drmanac et al. 2010], or *Utah* for brevity, which contains the genomes of 17 family members from three generations. In the $kin/opensnp$ scenario, we evaluate all 24 privacy metrics using all SNPs for 13 pairs of related individuals from openSNP data. In the $kin/utah$ scenario, we evaluate all 24 privacy metrics using all SNPs for the 17 related individuals from the Utah data.

5.1.3. Case Study Scenarios. To study whether the strength of privacy metrics varies when applied to a small set of SNPs, we investigate three real-world conditions as case studies: (1) susceptibility to asthma, (2) susceptibility to multiple sclerosis, and (3) vitamin B12 levels (deficiency in vitamin B12 is associated, among others, with cardiovascular disease and cancer). We drew on the NHGRI GWAS Catalog [Welter et al. 2014], a curated resource of SNP-trait associations, to find SNPs associated with these three conditions. We used only SNPs that were present in most genomes in the openSNP dataset and that had a p-value of $p \leq 5.0 \times 10^{-8}$. For asthma and multiple sclerosis, we used the odds ratio for each association as a weight for the health privacy and genomic privacy metrics. For vitamin B12 levels, we used the increase/decrease in B12 level (in pg/ml) as a weight. The SNPs and weights for each of the three case studies are listed in Table V in the appendix. For each of the case study scenarios, we assume that the adversary is only interested in an individual's propensity for the specific condition, and therefore we evaluate all privacy metrics on all genomes using only the SNPs that are associated with each condition.

5.1.4. Population Group Scenarios. Minor allele frequencies can vary widely between population groups. To study how the strength of privacy metrics varies when the adversary knows the minor allele frequencies for the target's specific population group, we used data from the International Genome Sample Resource (IGSR)[12] [The 1000 Genomes Project Consortium 2015]. In contrast to the openSNP data, the IGSR data includes verified metadata for each genome, including the population group. For the *population group* scenario, we evaluate all 24 privacy metrics for 100 genomes from each superpopulation group (i.e., African, American, East Asian, European, and South Asian) using 60 SNPs, i.e. the combined SNPs from the three case studies.

[11] To identify blood relatives in the openSNP dataset, we first identified pairs of genomes that shared more than 80% of genotypes (statistics from http://www.isogg.org/wiki/Autosomal_DNA_statistics). We then attempted to verify a potential relationship using the openSNP user names and user profiles. For 10 of these genomes, the relationship degree can be found either by references in the username (e.g., "father of") or by Google hits on ancestry sites, and another 3 are likely matches judging by the username (same infrequent last name), but with unknown relationship.

[12] http://www.internationalgenome.org/

5.2. Adversary Models

For the extended evaluation, we use two base adversary models: the *normal* adversary model from Section 4 and the new *uniform* model. We extend both models by giving the adversary access to two kinds of prior information: minor allele frequencies and linkage disequilibrium. When studying related individuals in the *kin* scenarios, we also give the adversary access to the kin genome and the degree of the relationship.

5.2.1. Base Adversary Models. In the *uniform* model, the weakest adversary makes an uninformed guess, represented by a truncated normal distribution that comes close to a uniform distribution. With increasing adversary strength, we skew the distribution towards certainty using increasingly narrow truncated normal distributions (i.e. increasingly smaller σ). Specifically, we study seven strength levels, setting the mean to $\mu = 0.99$, and varying the standard deviation $\sigma = 7, 2, 1, 0.5, 0.25, 0.1, 0.05$. Figure 2a shows the average probability the *uniform* adversary assigns to the true genotype.

While the *normal* and *uniform* adversary models do not necessarily represent realistic attacks on genomic privacy, they allow to evaluate monotonicity for uncharacteristically weak adversaries. This important when designing new privacy mechanisms because the strength of a privacy mechanism is essentially dual to the strength of the adversary: a very strong privacy mechanism makes the adversary appear weaker, and a weak privacy mechanism makes the adversary appear stronger. If a metric is used to evaluate a new privacy mechanism, and this mechanism turns out to be very strong (i.e., corresponding to a weak adversary), we need a metric that is monotonous for weak adversaries in order to not misrepresent the strength of the privacy mechanism.

5.2.2. Prior Information: Minor Allele Frequencies. We extend both base adversary models by giving the adversary access to minor allele frequencies (MAFs) as prior information. The adversary uses Bayes' theorem to update the base estimate, which significantly improves the estimate for both the uniform (Figure 2b) and normal (Figure 2e) estimates.

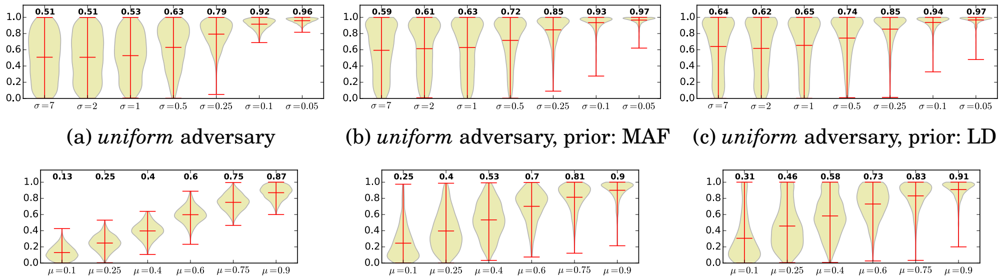
5.2.3. Prior Information: Linkage Disequilibrium. To model stronger adversaries, we give the adversary access to information about linkage disequilibrium in addition to MAFs. Specifically, we assume that for each SNP to be inferred, the adversary knows the true genotype and correlation coefficient for the SNP with the strongest LD relationship. We then use the conditional probability for the unknown genotype to update the adversary's MAF estimate. Figures 2c and 2f show how access to LD improves the uniform and normal estimates, respectively. We used the LD adversary only for a restricted set of SNPs, i.e. the 60 SNPs from the three case studies, and used rAggr[13] to retrieve the LD relationships for each SNP.

5.2.4. Prior Information: Kin Genome. For the *kin* scenarios, we modify all of the above adversary models by additionally giving the adversary prior information about the true genotypes of a related individual and the degree of the relationship. The adversary uses Bayes' theorem to update the base models, MAF models, and LD models with kin information. Figure 14 in the appendix shows the adversary's estimate for one sample individual when the adversary has prior information about one related individual.

5.3. Formalization of the Monotonicity Requirement

The initial evaluation emphasizes that a strong privacy metric should have decreasing privacy levels for increasing adversary strength. In mathematical terms, this means that privacy metrics should be monotonic with increasing adversary strength. Mono-

[13] http://raggr.usc.edu/



(a) *uniform* adversary

(b) *uniform* adversary, prior: MAF

(c) *uniform* adversary, prior: LD

(d) *normal* adversary

(e) *normal* adversary, prior: MAF

(f) *normal* adversary, prior: LD

Fig. 2. Average probability the different adversary types assign to the true genotype for a sample individual

tonicity is an important requirement, because a nonmonotonic metric can make a privacy-enhancing technology appear stronger than it actually is. If this technology is then used in practice, the use of a weak privacy metric may cause privacy violations.

We formulate an algorithm that evaluates monotonicity based on statistical tests (Algorithm 1). Our algorithm outputs heat maps as a compact and easy-to-understand visualization of the strength of privacy metrics. The algorithm automates the analysis of monotonicity and ensures that the results are consistent and less error-prone than the visual analysis of plots conducted in Section 4. In a monotonic sequence, the differences between successive pairs should all have the same sign. The algorithm therefore awards points for each difference that has the expected sign (positive for metrics where high values indicate high privacy, and negative for metrics where low values indicate high privacy), and penalizes differences that have the wrong sign. Because we have a large number of data points for each adversary strength level, we use statistical tests to evaluate the differences between the means of successive pairs. Welch's t-test tests the null hypothesis that the metric values for the two adversary strengths have identical means (in contrast to the standard t-test, Welch's t-test does not assume equal variance in the two samples), and the Wilcoxon rank-sum statistic tests the null hypothesis that the metric values have been drawn from the same distribution. The results of these tests indicate whether the difference between the mean metric values is positive, negative, or zero, and whether the difference is statistically significant.

We use heat maps to visualize the resulting point values. Figure 3 shows one heat map for each privacy metric. The rows represent scenarios (*comparison*, *kin/opensnp*, *kin/utah*, and the three case studies *b12*, *asthma*, and *multiple sclerosis*) and the columns represent adversary models (normal, uniform, normal with minor allele frequencies, uniform with minor allele frequencies, normal with linkage disequilibrium, uniform with LD). Blue colors indicate a strong metric, green indicates medium strength, and yellow indicates a weak metric. This visualization presents the strengths and weaknesses of a large number of privacy metrics in a compact way and thus helps researchers to select strong metrics. To get a sense of the overall strength of a privacy metric, we aggregate the strengths for all elements in its heat map and show it as a single percentage next to the metric name in Figure 3.

We chose the point values in our algorithm to reflect the monotonicity requirements and to create a high contrast in the visualization, which helps to pinpoint strengths and weaknesses of each metric. We assigned points based on the desired behavior of a metric: a change in the right direction (1 point) is better than no change (0 points), which in turn is better than a change in the wrong direction (-1 points). A peak (-2 points) is undesirable because it means that weak adversaries cannot be distinguished from strong adversaries.

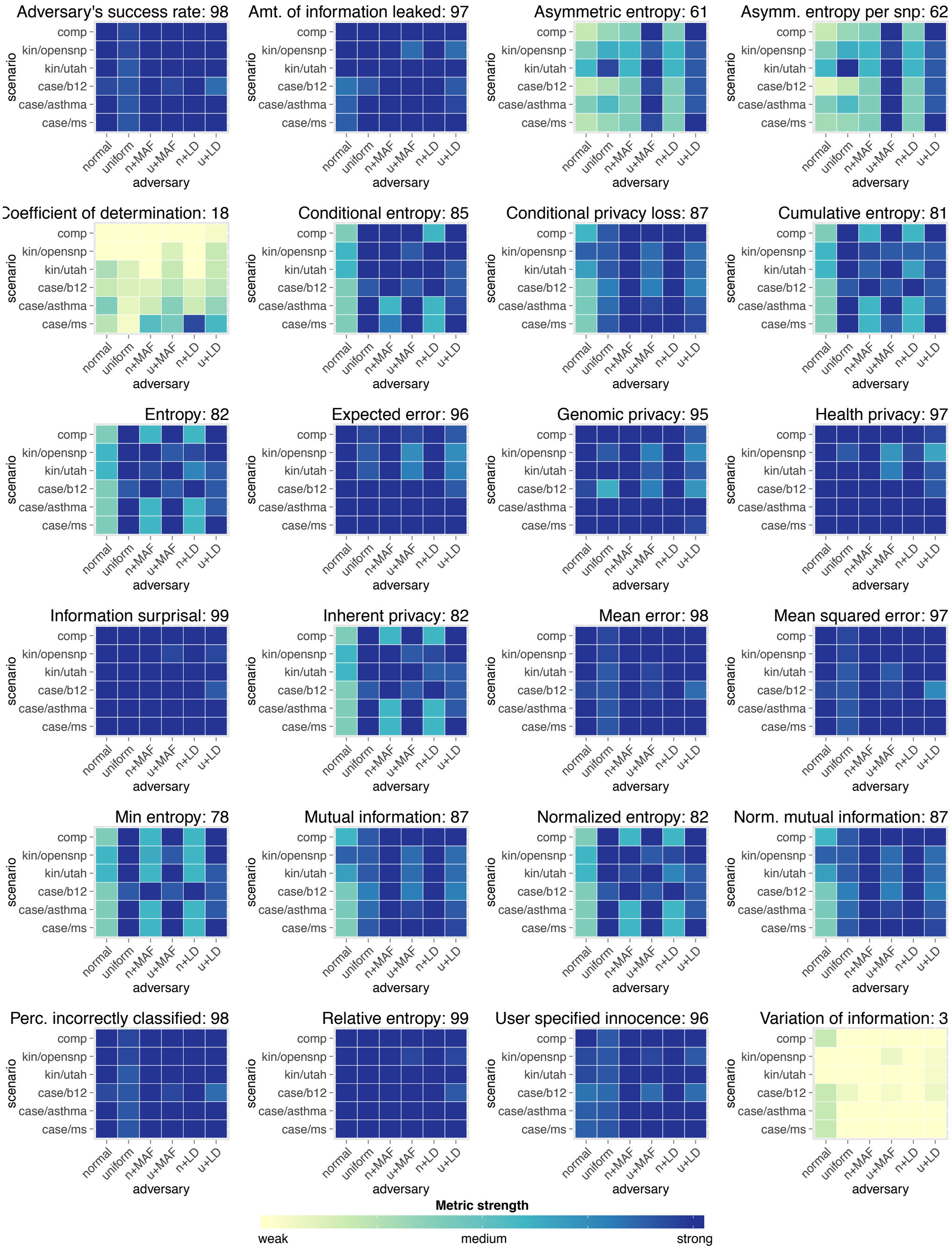
ALGORITHM 1: Monotonicity Computation for one Privacy Metric

```
Input: arrays of metric values for each combination of adversary model and scenario
Output: heat map visualizing the strength of this privacy metric
tests = [Welch's t-test, Wilcoxon rank-sum statistic]
foreach combination of adversary model and scenario do
    $m = 0$ ;                                        // holds the monotonicity points value
    foreach $test \in tests$ do
        $prevResult = 0$;                            // holds result for the previous pair
        foreach pair of successive adversary strengths do
            apply $test$ to pair
            $p$ = statistical significance of test
            $result$ = value of test statistic
            if $p < 0.05$ then                       // test is statistically significant
                if $result > 0$ ($< 0$ for lower-better metrics) then
                    $m = m + 1$ ;                    // difference in the expected direction
                else if $result < 0$ ($> 0$ for lower-better metrics) then
                    $m = m - 1$;                     // difference in the wrong direction
                else
                    ;                                // result is zero, do nothing
                end
            else                                     // test is not statistically significant
                $m = m - 0.2$
            end
            if $result$ and $prevResult$ have different signs then
                $m = m - 2$;                         // penalize peaks in the metric value
            end
            $prevResult = result$;                   // save result to check next pair for peaks
        end
    end
    normalize $m$ to $[-1, 1]$
    save $m$ for plotting
end
plot $m$ in a heat map (rows = scenarios, columns = adversaries)
```

5.4. Results

Figure 3 shows the heat maps for the strengths of all 24 genomic privacy metrics, obtained according to Algorithm 1. Table IV in the appendix lists the SNPs and population groups used for each cell in each heat map. Most entropy-based metrics (asymmetric entropy, conditional entropy, cumulative entropy, min-entropy, normalized entropy, and inherent privacy) behave similarly to entropy, resulting in similar heat maps. These metrics have clear weaknesses for the *normal* adversary type, including the *normal* adversary with prior information, and should therefore only be used in combination with other metrics. A similar behavior, albeit less pronounced, can also be observed for mutual information, normalized mutual information, and conditional privacy loss.

Relative entropy and information surprisal are the only two information theoretic metrics that produce consistently good results, i.e. they produce consistent measurements regardless of the adversary model and scenario. Other strong metrics are the adversary's success and error (expected estimation error, mean error, mean squared error, percentage incorrectly classified), genomic privacy, and metrics measuring the number of SNPs that are leaked or remain private. These metrics can be recommended for use in genomic privacy.



Adversary's success rate: 98

Amt. of information leaked: 97

Asymmetric entropy: 61

Asymm. entropy per snp: 62

Coefficient of determination: 18

Conditional entropy: 85

Conditional privacy loss: 87

Cumulative entropy: 81

Entropy: 82

Expected error: 96

Genomic privacy: 95

Health privacy: 97

Information surprisal: 99

Inherent privacy: 82

Mean error: 98

Mean squared error: 97

Min entropy: 78

Mutual information: 87

Normalized entropy: 82

Norm. mutual information: 87

Perc. incorrectly classified: 98

Relative entropy: 99

User specified innocence: 96

Variation of information: 3

Fig. 3. Strength of 24 genomic privacy metrics shown in heat maps. In each plot, the name of the metric and its overall strength are given in the title, the X axis shows the adversary model, the Y axis shows the scenario, and the colors indicate the strength of the metric (from blue=strong to yellow=weak)

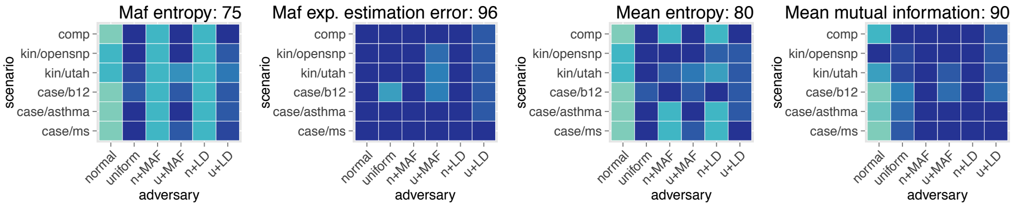


Fig. 4. Strength of four aggregated privacy metrics shown in heat maps. In each plot, the name of the metric and its overall strength are given in the title, the X axis shows the adversary model, the Y axis shows the scenario, and the colors indicate the strength of the metric (from blue=strong to yellow=weak)

The metrics that performed worst in our tests are the coefficient of determination, variation of information, and asymmetric entropy. These metrics do not produce good measurements for most scenarios and adversary models, and can therefore not be recommended for use in genomic privacy.

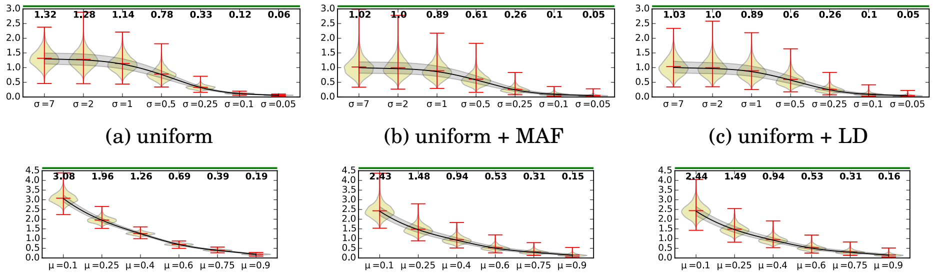
5.4.1. Results for Kin Privacy. The strength of most genomic privacy metrics does not vary when they are applied only to related individuals (*kin*/*opensnp* and *kin*/*utah* scenarios, second and third rows of each heat map in Figure 3) as compared to a large sample of unrelated individuals (*comparison* scenario, first row). Exceptions to this are some of the weaker metrics, especially asymmetric entropy and the coefficient of determination, but also many entropy-based metrics, which appear stronger when applied to related individuals. This is most likely caused by the change in the adversary's estimate when the adversary is given the kin genome and the degree of relationship as prior knowledge.

Other metrics, for example genomic privacy and the amount of information leaked, appear stronger when applied to the *utah* dataset as opposed to the *opensnp* dataset. To explain this, we first note that both the kin and utah scenarios consist only of a small number of individuals (13 and 17, respectively). Second, all individuals in the utah dataset are related to each other, whereas the relationships in the openSNP dataset are between pairs or groups of three. Because the adversary's prior information consists of population-wide allele frequencies, the individuals in the utah dataset (third row) would all tend to have the same deviation from these frequencies, and the deviation in this particular case makes some metrics appear stronger compared to the more random sample of individuals in the openSNP dataset (second row). This is true even though we used minor allele frequencies for the correct population group.

Interestingly, this difference in strength occurs for both weak and strong metrics. This emphasizes the need to use combinations of strong privacy metrics.

5.4.2. Results for Aggregation and Normalization. Normalization aims to bring the metric values for different scenarios into a common value range to allow comparisons. As the heat maps for normalized entropy and normalized mutual information in Figure 3 show, normalization does not change the strength of a privacy metric with regard to monotonicity.

Aggregation aims to combine the metric values for all SNPs belonging to one individual, effectively reducing the amount of data to analyze. Figure 4 shows the strengths of four aggregated metrics, using two different aggregation methods: an arithmetic mean and a mean weighted with population-wide minor allele frequencies (denoted *Maf* in the Figure). As the comparison between the base metrics in Figure 3 and the aggregated metrics in Figure 4 shows, aggregation does not affect the strength of privacy metrics, regardless of the aggregation method used.



(a) uniform

(b) uniform + MAF

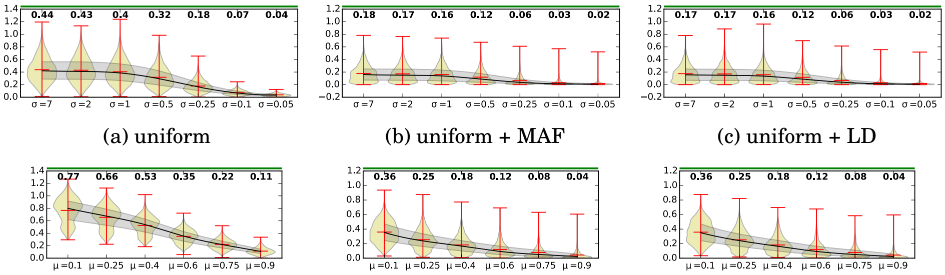
(c) uniform + LD

(d) normal

(e) normal + MAF

(f) normal + LD

Fig. 5. Health privacy based on information surprisal for the asthma case study (IGSR data, prior information based on correct population group)



(a) uniform

(b) uniform + MAF

(c) uniform + LD

(d) normal

(e) normal + MAF

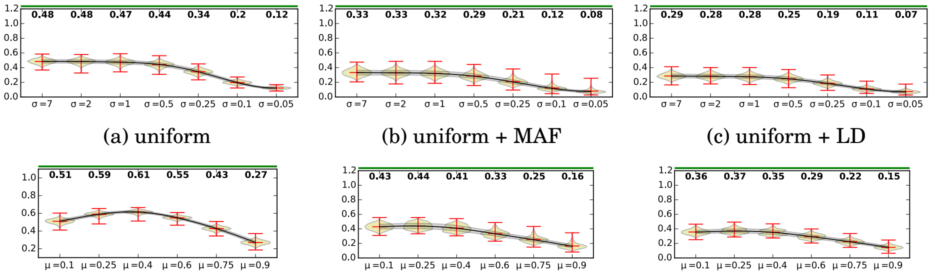
(f) normal + LD

Fig. 6. Health privacy based on the expected estimation error for the vitamin B12 case study (IGSR data, prior information based on correct population group)

5.5. Influence of Parameter Settings

5.5.1. Parameter Settings for Health Privacy. Health privacy presents a way how a per-SNP metric can be aggregated into a per-individual metric. It relies on three parameters: the selection of SNPs, the weights assigned to each, and the base metric that computes per-SNP values. Since the metric is normalized using the sum of SNP weights, the number of SNPs and the composition of the weights do not influence the magnitude of the final value. The value of health privacy therefore depends mostly on the value of the base metric. We evaluate health privacy for seven base metrics: relative entropy, normalized mutual information, normalized entropy, min-entropy, information surprisal, expected estimation error, and conditional entropy. We select SNPs and the corresponding weights according to our three case studies for real-world conditions: asthma, vitamin B12 levels, and multiple sclerosis. Figures 5, 6 and 7 show detailed results for three sample combinations: asthma with information surprisal, B12 levels with the expected estimation error, and multiple sclerosis with normalized entropy. The full set of results is shown in heat maps in Figure 8.

Figure 5 shows that information surprisal is a consistently good (monotonic) base metric for all adversary types and, as Figure 8 shows, also across all three case studies. Figure 6 shows the same for the expected estimation error. Figure 7 shows that use of a weak base metric, in this case normalized entropy, makes health privacy a weak metric. For the *uniform* adversary types, health privacy with normalized entropy fails to distinguish all adversary strength levels, especially for weak adversaries. For the *normal* adversary types, the metric is clearly non-monotonic and ranks privacy for medium-strength adversaries higher than both the weakest and strongest adversaries.



(a) uniform (b) uniform + MAF (c) uniform + LD

(d) normal (e) normal + MAF (f) normal + LD

Fig. 7. Health privacy based on normalized entropy for the multiple sclerosis case study (IGSR data, prior information based on correct population group)

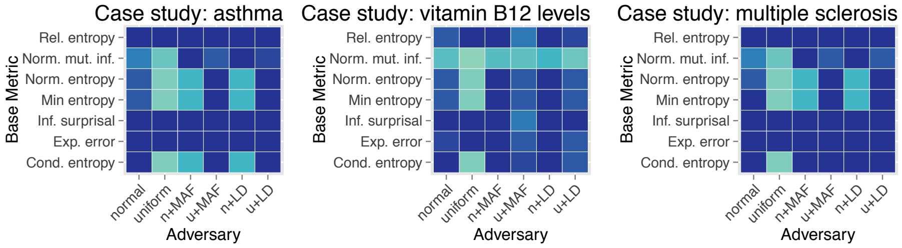


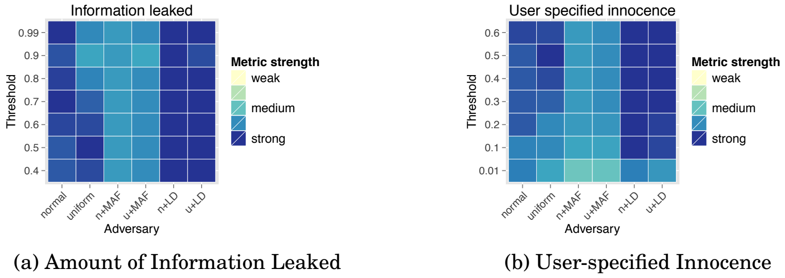
Fig. 8. Strength of health privacy for different base metrics and SNPs/weights for three case studies, shown in heat maps (IGSR data, prior information based on correct population group). The X axis shows the adversary model, the Y axis shows the base metric, and the colors indicate the strength of the metric.

Comparing the strength of different base metrics in Figure 8 with the corresponding heat maps in Figure 3 (first and bottom three rows), we can see that the strength of health privacy corresponds to the strength of the base metric. This means that health privacy is a useful way of aggregating per-SNP metrics into a single per-individual metric, provided that the base metric is appropriate.

5.5.2. Parameter Settings for Amount of Information Leaked and User-Specified Innocence. Both the amount of information leaked and user-specified innocence have a threshold parameter that indicates when a SNP is considered leaked resp. private. We found that setting the threshold for the amount of leaked information close to 1 resulted in zero leaked SNPs for weak adversaries, and 100% leaked SNPs for strong adversaries. The reverse is true for user-specified innocence, when its threshold is set to 0. This setting therefore doesn't allow to distinguish adversaries of different strengths. We have studied the influence of the threshold on each metric's strength, depending on the type of adversary. Figure 9 shows that both metrics are strongest for a threshold of 0.5, with good strengths for thresholds between 0.5 and 0.7 (amount of information leaked), and between 0.3 and 0.5 (user-specified innocence). Combining the two metrics reveals additional information, because in addition to the number of leaked and private SNPs, the combination also shows for how many SNPs the leakage status is uncertain.

5.6. Results for Population Groups

For the initial evaluation in Section 4 and the *comparison* scenario in Section 5, the adversary's prior information about minor allele frequencies was a world-wide estimate computed from a sample of 1000 genomes drawn from different population groups. However, allele frequencies can vary considerably between population groups. There-



(a) Amount of Information Leaked

(b) User-specified Innocence

Fig. 9. Amount of Information Leaked and User-specified Innocence with different thresholds, evaluated using six adversary models, from left to right: *uniform*, *normal*, uniform and normal with minor-allele frequencies as prior information, uniform and normal with LD as prior information.

fore, we now analyze the effect on the strength of privacy metrics when the adversary knows the correct minor allele frequencies for the target's population group.

For each of the five superpopulation groups (African, American, East Asian, European, South Asian), we evaluate all 24 privacy metrics on 100 genomes using the combined SNPs for the three case studies. The results in Figure 10 compare the metric strengths for each population group, once with the correct minor allele frequencies, and once with the world-wide mixture.

Figure 10 indicates that the strength of most metrics depends neither on the target's population group, nor on the kind of minor allele frequencies used by the adversary. Note that this does not preclude the adversary benefiting from knowledge of the target's population group or the correct minor allele frequencies. It only means that the strength of the privacy metric – but not necessarily the value of the metric – is unaffected by this.

Two notable exceptions are the coefficient of determination and conditional entropy. Both metrics are stronger when the correct minor allele frequencies are used, and especially so for some of the population groups. This effect may be due to systematic differences in the genomes of the affected population groups. In practice, the influence of such effects can be minimized by using several privacy metrics in combination.

6. CASE STUDIES

Because genomic privacy is often concerned with protecting information about an individual's susceptibility to genetic diseases, we applied the privacy metrics in three case studies: (1) susceptibility to asthma, (2) susceptibility to multiple sclerosis, and (3) vitamin B12 levels (deficiency in vitamin B12 is associated, among others, with cardiovascular disease and cancer). In this section, we are not primarily concerned with the **strength** of privacy metrics, but rather illustrate the **process** of selecting privacy metrics for a real scenario and interpreting the results. This will allow us to draw further conclusions about the usefulness of metrics and metric combinations. We identified four tasks that are necessary to measure privacy in a real genomic privacy scenario: the choice of SNPs, the selection of privacy metrics, the choice of parameters for the selected metrics, and the interpretation of results.

6.1. Choice of SNPs

The first task is the choice of SNPs. As explained above, we drew on the NHGRI GWAS Catalog [Welter et al. 2014] to find SNPs associated with our three case study conditions. We then restricted our selection in two steps. First, we selected only SNPs with a p-value of less than 5.0×10^{-8}, and second, we selected only SNPs that were present in most genomes in the openSNP dataset.

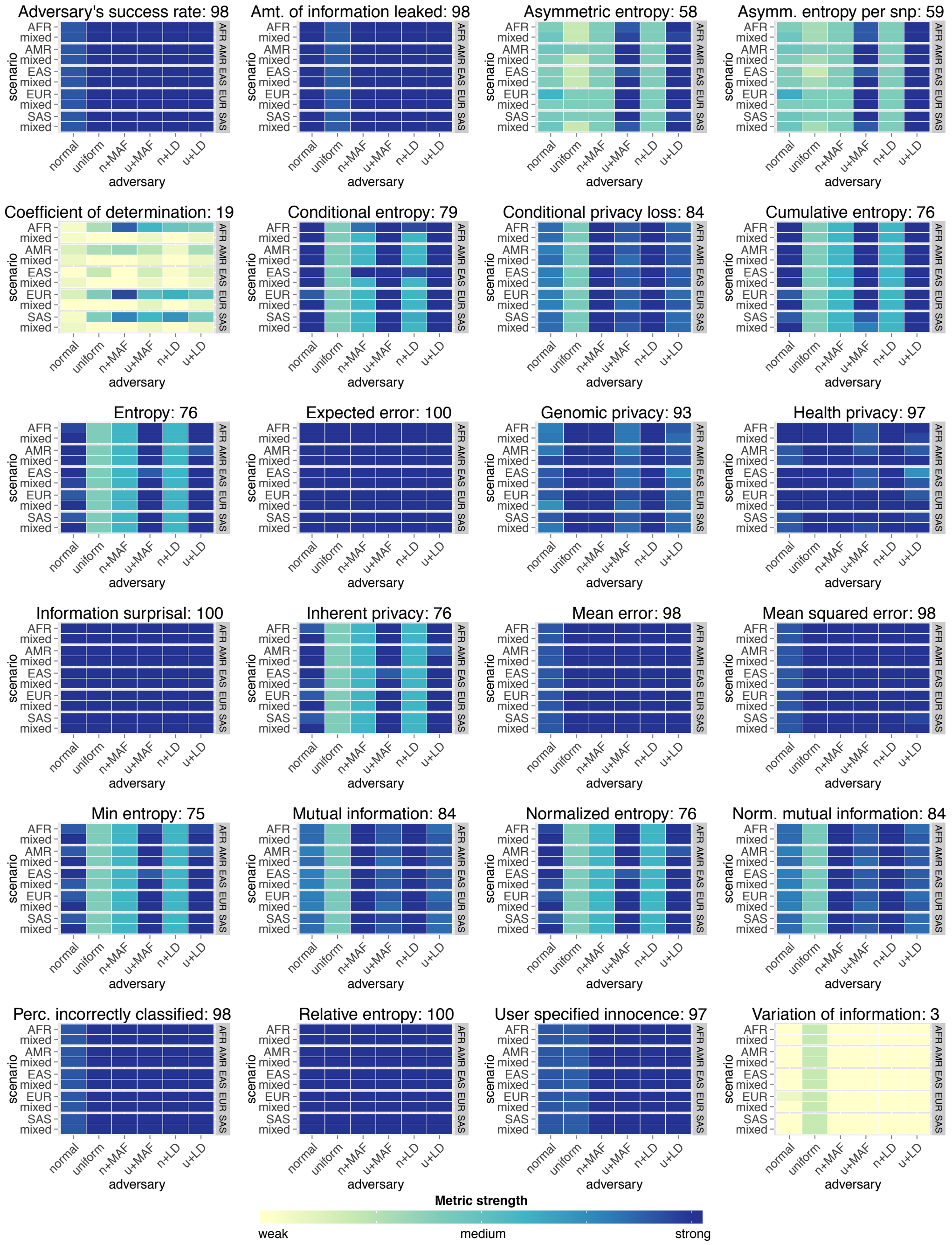


Fig. 10. Strength of privacy metrics depending on the population group. In each heat map, the name of the metric and its overall strength are given in the title, the X axis shows the adversary model, and the Y axis shows the population group (AFR: African, AMR: Americas, EAS: East Asian, EUR: European, SAS: South Asian). The rows marked with *mixed* indicate that the adversary used world-wide minor allele frequencies instead of population-specific ones.

6.2. Selection of Privacy Metrics

The second task is the selection of privacy metrics. Following the process described in [Wagner and Eckhoff 2015], we use eight questions to guide the selection:

(1) **Output measures.** Wagner and Eckhoff [2015] propose eight categories of output measures and suggest that metrics from as many categories as possible should be selected. Our study includes metrics from only five categories (uncertainty, information gain/loss, similarity/diversity, adversary's success probability, and error). However, the only metric belonging to the similarity/diversity category is the coefficient of determination which, as we have shown above, is not a suitable metric for genomic privacy. We will therefore select metrics from each of the other four categories.

(2) **Adversary models**. All metrics in our study are computed using the adversary's estimate. This question therefore does not influence our choice of metrics directly.

(3) **Data source** refers to the data adversaries would use to perform their attack. In our scenario, data could be either observable or published data. Neither data source restricts our choice of metrics.

(4) **Availability of input data.** In this study, we have access to all input data required by different metrics, including knowledge of the adversary estimate, the true outcome, and parameter settings. This question does therefore not influence our choice of metrics.

(5) **Target audience.** Even though this paper is targeted at academics, some target audiences may require metrics that can be interpreted easily. We therefore discuss below how each metric can be interpreted.

(6) **Related work** in genomic privacy has used entropy, expected estimation error, adversary's success rate, genomic privacy, and health privacy. We should therefore consider including these five metrics.

(7) **Strength of metrics.** We can refer to the heat maps in Figure 3 for results about the strength of privacy metrics. The bottom row of each heat map indicates the results specific to the case study scenarios we are interested in here. We list the strongest metrics in the *strong metrics* column of Table III.

(8) **Implementation of metrics.** We have relied on generic implementations of entropy and mutual information available in Python packages. To the best of our knowledge, validated implementations of specific privacy metrics are not available, and therefore this question does not influence our choice of metric.

Considering our answers to the eight questions, we see that we need to select strong metrics from four categories and include the five metrics considered in related work. Table III shows how strong metrics and metrics from related work fit into the four categories, with our choice of metrics highlighted in italics. In total, we select eight metrics: four related work metrics[14], three strong metrics to add to the information gain/loss category, and one strong metric to add to the error category.

6.3. Choice of Metric Parameters

The third task is to choose parameter settings for the selected metrics. Health privacy and genomic privacy use weights for individual SNPs, ideally chosen to reflect how much each SNP contributes to the overall disease risk. For the asthma and multiple sclerosis case studies, we chose the weights to correspond to the odds ratios found in the NHGRI GWAS Catalog [Welter et al. 2014]. For the vitamin B12 case study, we

[14] Because of the similarity between health privacy and its base metric, we omit the expected estimation error and instead consider health privacy based on the expected estimation error

Table III. Metric Selection

Category	Strong metrics	Related work metrics
Adversary's success probability	*adversary's success rate* user-specified innocence	*adversary's success rate*
Error	expected estimation error *mean squared error* mean error percentage incorrectly classified	expected estimation error *health privacy (error)*
Information gain/loss	*amount of leaked information* *relative entropy* information surprisal *health privacy (inf. surprisal)*	*genomic privacy*
Uncertainty		*entropy*

chose the weights to correspond to the increase/decrease in B12 levels associated with each SNP. The weights are shown in Table IV in the appendix.

The amount of information leaked uses a parameter for the threshold probability which depends on the privacy preferences of individual users. For our case study, we chose the threshold according to the results in Section 5.5.2, which showed that the metric is strongest for a threshold of 0.5. This threshold means that SNPs are considered leaked if the adversary's estimate of the true genotype is above 50%.

6.4. Interpretation of Results

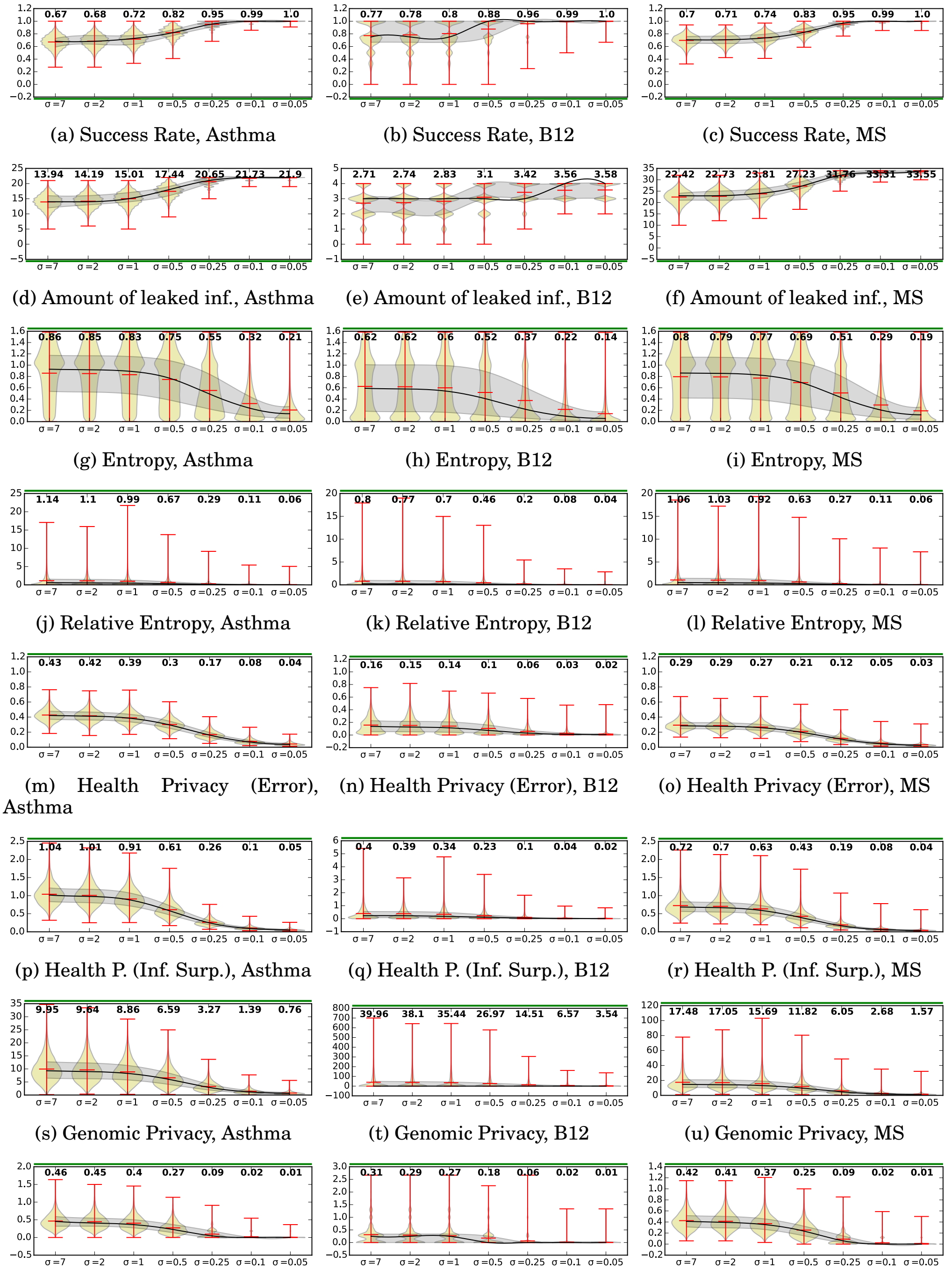
After conducting the privacy measurement, the fourth and final task is to plot and interpret the results.

6.4.1. Interpreting the Values of Privacy Metrics. To interpret what the values of each privacy metric mean, we show violin plots for the eight selected metrics and each of the three case studies in Figure 11. Due to space constraints, we focus on the uniform adversary type with minor allele frequencies as prior information. (For completeness, we show the results for the remaining metrics in the appendix.)

The *adversary's success rate* (Figures 11a–11c) indicates the fraction of SNPs correctly inferred by the adversary. For the B12 case study, which relies on only 4 SNPs, we can see the four individual peaks in the distribution of the adversary's success rate. In the other two case studies these peaks have been smoothed out because they use more SNPs. In all three cases, the values are monotonic (non-decreasing), but the change is not linear: at both ends of the adversary strength spectrum, changes between adversary strength levels are very small.

The *amount of information leaked* (Figures 11d–11f) behaves similarly to the adversary's success rate. The values show the number of SNPs for which the adversary has inferred the correct genotype, which depends strongly on the number of SNPs in the study. Note that the y axes have different ranges for this metric, corresponding to the number of SNPs in each case study: 24 for asthma, 4 for vitamin B12 levels, and 34 for multiple sclerosis.

Entropy (Figures 11g–11i) measures the adversary's uncertainty and therefore does not always reliably indicate the user's privacy level (but it is monotonic for the uniform adversary type). The values of entropy indicate how many bits of information are contained in the adversary's per-SNP estimate, with high values indicating more uncertainty. The entropy values for the B12 case study are consistently lower than for the other two case studies. This indicates that the choice of SNPs influences the entropy value, and comparisons between studies may not be meaningful. Especially for the low and medium-strength adversaries, the distribution of entropy is quite wide, as indicated by the width of the gray shading between the 25% and 75% quantiles. This



(a) Success Rate, Asthma

(b) Success Rate, B12

(c) Success Rate, MS

(d) Amount of leaked inf., Asthma

(e) Amount of leaked inf., B12

(f) Amount of leaked inf., MS

(g) Entropy, Asthma

(h) Entropy, B12

(i) Entropy, MS

(j) Relative Entropy, Asthma

(k) Relative Entropy, B12

(l) Relative Entropy, MS

(m) Health Privacy (Error), Asthma

(n) Health Privacy (Error), B12

(o) Health Privacy (Error), MS

(p) Health P. (Inf. Surp.), Asthma

(q) Health P. (Inf. Surp.), B12

(r) Health P. (Inf. Surp.), MS

(s) Genomic Privacy, Asthma

(t) Genomic Privacy, B12

(u) Genomic Privacy, MS

(v) Mean Squared Error, Asthma

(w) Mean Squared Error, B12

(x) Mean Squared Error, MS

Fig. 11. Privacy metrics evaluated according to adversary strength for the *uniform* adversary with minor allele frequencies as prior information. Green bars show whether high or low values indicate high privacy.

may be problematic in studies with a small number of genomes because they might sample only part of the distribution, which may influence the conclusions.

Relative entropy (Figures 11j–11l) indicates how much additional information, measured in bits, the adversary needs to reconstruct the true genotypes. This amount of information is similar to the amount of surprise the adversary will experience upon learning the true genotype, i.e. the information surprisal metric (see Figures 15s–15u in the appendix).

Health privacy with the expected estimation error as base metric (Figures 11m–11o) indicates the weighted average expected estimation error for each SNP. *Health privacy with information surprisal* as base metric (Figures 11p–11r) indicates the weighted average information surprisal for each SNP, where information surprisal indicates how much additional information (in bits) the adversary would gain upon learning the true genotype. Both variants of health privacy show a significant difference in value between the three case studies, which means that comparisons between studies may not be meaningful.

The *genomic privacy* metric (Figures 11s–11u) does not use weighting, which means that the final value is heavily influenced by both the weights and the number of SNPs. It is unclear how the value of genomic privacy should be interpreted – what does it mean if an individual has a genomic privacy of 9.6? Due to the dependence on the number of SNPs and their weights, the values of genomic privacy are not comparable between studies.

The *mean squared error* (Figures 11v–11x) shows how far, on average, the adversary's guess is from the true genotype. In statistics, the mean squared error is often used to indicate the quality of an estimator (here the adversary's estimate). As such, the mean squared error can be used for comparisons between studies. However, since the error is squared and computed on encoded genotypes, the meaning of the values is not intuitively clear.

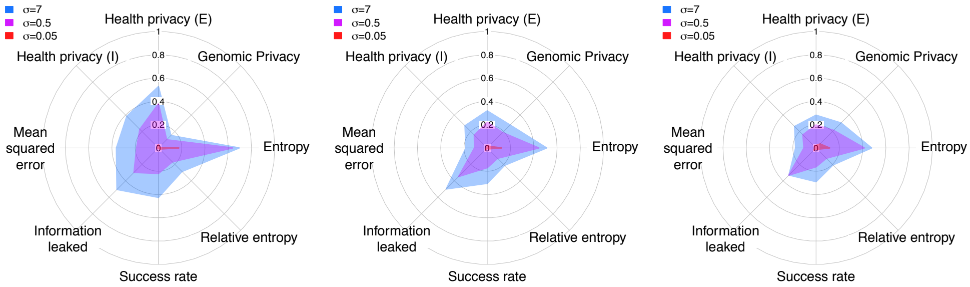
6.4.2. Intuitiveness. The use of privacy metrics with an intuitive interpretation can help communicate findings to the target audience. Based on the findings in this and the previous sections, we rated each metric based on (1) how easy it is to understand what its values mean, and (2) how easily it can be interpreted. We summarize our (subjective) ratings in Table II (column *Intuitiveness*).

6.4.3. Interpreting the Overall Privacy Level. The violin plots presented in Figure 11 are comprehensive, but they make it hard to tell what an individual's overall privacy level is against a specific adversary. In addition, even though we selected mostly strong metrics for the case study, some metrics have small weaknesses when used on their own, as we pointed out in the previous section. Radar plots can visualize the overall privacy level indicated by a combination of metrics.

Figures 12 and 13 show radar plots for different combinations of case study, adversary type, and adversary strength. To keep the plots clean, we plot only three strength levels per adversary type. The values for each metric have been normalized to the $[0, 1]$ value range using the 10th and 90th percentile of values across all adversary strengths. In addition, we inverted the values for lower-better metrics. As a result, a larger area in the plots directly corresponds to a higher privacy level.

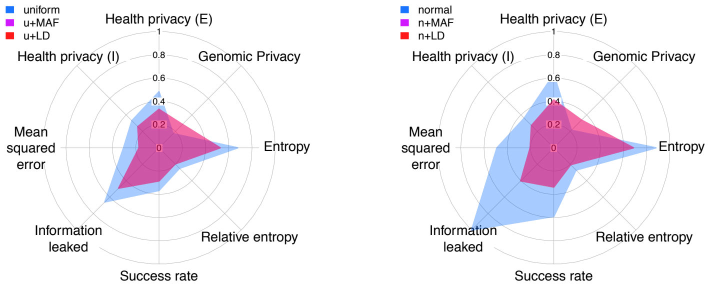
Figure 12 highlights how privacy changes when the adversary is given prior knowledge. Privacy is highest for the base adversary in Figure 12a, and becomes smaller when the adversary knows about minor allele frequencies (Figure 12b) and linkage disequilibrium (Figure 12c).

Radar plots also allow comparisons between different adversary types. For example, the *normal* adversary types (Figure 13b) are much weaker than the *uniform* adversary types (Figure 13a), as indicated by the larger privacy area for the *normal* adversaries.



(a) Uniform adversary (b) Uniform + MAF adversary (c) Uniform + LD adversary

Fig. 12. Radar plots visualizing the privacy level of eight privacy metrics for three strength levels of each adversary type based on the *uniform* adversary for the multiple sclerosis case study



(a) Uniform adversary types (b) Normal adversary types

Fig. 13. Radar plots visualizing the privacy level of eight privacy metrics for each of the three types of prior information (none, minor allele frequencies as prior, linkage disequilibrium as prior) using a medium-strength adversary ($\sigma = 1$ and $p = 0.4$, respectively) for the asthma case study

A real-world evaluation of privacy may not vary the adversary strengths as we have done, but instead vary parameters of a new privacy-enhancing technology. Since the strength of the adversary and the strength of a privacy enhancing technology are essentially two sides of the same coin, we expect that radar plots will be able to highlight differences between privacy enhancing technologies in the same way as differences between adversaries.

7. CONCLUSIONS AND FUTURE WORK

We measured the strengths of 24 published genomic privacy metrics. We introduced monotonicity as the key indicator of a metric's strength, i.e. metrics should show decreasing privacy for increasing adversary strength. We tested each of the 24 metrics in four different scenarios, for adversaries of different strengths, and found that only 7 out of 24 metrics were strong across scenarios and adversary types and could be interpreted easily. The 7 strong metrics were the adversary's success rate, the amount of information leaked, health privacy (with information surprisal or relative entropy as base metric), information surprisal, percentage incorrectly classified, relative entropy, and user-specified innocence. Furthermore, we found that none of the metrics we tested were sufficiently reliable when used in isolation. Therefore, we recommend that several strong metrics that measure different outputs should be used together. Finally, we showed how heat maps can be used to visualize the strength of privacy metrics in a compact and intuitive way, and how the level of privacy indicated by a combination of metrics can be visualized in radar plots. Our systematic comparison of genomic privacy metrics will enable researchers to make informed and consistent decisions about the selection of privacy metrics and privacy enhancing technologies.

In future work, we will measure the strength of privacy metrics in other application domains, e.g., vehicular networking and smart metering. Future work also needs to study whether there are additional requirements for privacy metrics aside from monotonicity, and whether privacy metrics should satisfy the conditions for metrics in a mathematical sense.

ACKNOWLEDGMENTS

This work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk). The author would like to thank the anonymous reviewers for their insightful comments and suggestions.

REFERENCES

Dakshi Agrawal and Charu C. Aggarwal. 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems (PODS 2001)*. ACM, Santa Barbara, CA, USA, 247–255.

James Alexander and Jonathan Smith. 2003. Engineering Privacy in Public: Confounding Face Recognition. In *Proc. 3rd Int. Workshop on Privacy Enhancing Technologies (PET 2003) (LNCS 2760)*. Springer, Dresden, Germany, 88–106.

Christer Andersson and Reine Lundin. 2008. On the Fundamentals of Anonymity Metrics. In *Proc. 3rd IFIP Int. Summer School on The Future of Identity in the Information Society*. Springer, Karlstad, Sweden, 325–341.

Erman Ayday, Jean Louis Raisaro, Urs Hengartner, Adam Molyneaux, and Jean-Pierre Hubaux. 2014. Privacy-preserving processing of raw genomic data. In *Data Privacy Management and Autonomous Spontaneous Security*. Springer, 133–147.

Erman Ayday, Jean Louis Raisaro, and Jean-Pierre Hubaux. 2013a. Personal Use of the Genomic Data: Privacy vs. Storage Cost. In *Proc. IEEE Global Communications Conf. (GLOBECOM 2013)*. IEEE, Atlanta, GA, USA, 2723–2729.

Erman Ayday, Jean Louis Raisaro, Jean-Pierre Hubaux, and Jacques Rougemont. 2013b. Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine. In *Proc. 12th ACM Workshop on Workshop on Privacy in the Electronic Society (WPES'13)*. ACM, Berlin, Germany, 95–106. DOI:http://dx.doi.org/10.1145/2517840.2517843

Elisa Bertino, Dan Lin, and Wei Jiang. 2008. A Survey of Quantification of Privacy Preserving Data Mining Algorithms. In *Privacy-Preserving Data Mining: Models and Algorithms*. Number 34 in Advances in Database Systems. Springer, Chapter 8, 183–205.

Terence Chen, Abdelberi Chaabane, Pierre Ugo Tournoux, Mohamed-Ali Kaafar, and Roksana Boreli. 2013. How Much Is Too Much? Leveraging Ads Audience Estimation to Evaluate Public Profile Uniqueness. In *Proc. 13th Int. Symp. on Privacy Enhancing Technologies (PETS 2013) (LNCS 7981)*. Springer, Bloomington, IN, USA, 225–244.

Xihui Chen and Jun Pang. 2012. Measuring Query Privacy in Location-based Services. In *Proc. 2nd ACM Conf. on Data and Application Security and Privacy (CODASPY'12)*. ACM, San Antonio, TX, USA, 49–60. DOI:http://dx.doi.org/10.1145/2133601.2133608

Sebastian Clauß and Stefan Schiffner. 2006. Structuring Anonymity Metrics. In *Proc. 13th ACM Conf. on Computer and Communications Security 2006 (CCS'06): 2nd ACM Workshop on Digital Identity Management (DIM'06)*. ACM, Alexandria, VA, USA, 55–62. DOI:http://dx.doi.org/10.1145/1179529.1179539

Yuxin Deng, Jun Pang, and Peng Wu. 2007. Measuring Anonymity with Relative Entropy. In *Proc. 8th Int. Workshop on Formal Aspects in Security and Trust (FAST 2011)*. Springer, Leuven, Belgium, 65–79.

Claudia Diaz, Stefaan Seys, Joris Claessens, and Bart Preneel. 2003. Towards Measuring Anonymity. In *Privacy Enhancing Technologies*. 54–68.

Claudia Diaz, Carmela Troncoso, and George Danezis. 2007. Does Additional Information Always Reduce Anonymity?. In *Proc. 6th ACM Workshop on Privacy in Electronic Society (WPES '07)*. ACM, Alexandria, VA, USA, 72–75. DOI:http://dx.doi.org/10.1145/1314333.1314347

Radoje Drmanac, Andrew B. Sparks, Matthew J. Callow, Aaron L. Halpern, Norman L. Burns, Bahram G. Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B. Nilsen, George Yeung, Fredrik Dahl, Andres Fernandez, Bryan Staker, Krishna P. Pant, Jonathan Baccash, Adam P. Borcherding, Anushka Brownley, Ryan Cedeno, Linsu Chen, Dan Chernikoff, Alex Cheung, Razvan Chirita, Benjamin Curson, Jessica C. Ebert, Coleen R. Hacker, Robert Hartlage, Brian Hauser, Steve Huang, Yuan Jiang, Vitali

Karpinchyk, Mark Koenig, Calvin Kong, Tom Landers, Catherine Le, Jia Liu, Celeste E. McBride, Matt Morenzoni, Robert E. Morey, Karl Mutch, Helena Perazich, Kimberly Perry, Brock A. Peters, Joe Peterson, Charit L. Pethiyagoda, Kaliprasad Pothuraju, Claudia Richter, Abraham M. Rosenbaum, Shaunak Roy, Jay Shafto, Uladzislau Sharanhovich, Karen W. Shannon, Conrad G. Sheppy, Michel Sun, Joseph V. Thakuria, Anne Tran, Dylan Vu, Alexander Wait Zaranek, Xiaodi Wu, Snezana Drmanac, Arnold R. Oliphant, William C. Banyai, Bruce Martin, Dennis G. Ballinger, George M. Church, and Clifford A. Reid. 2010. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* 327, 5961 (Jan. 2010), 78–81. DOI:http://dx.doi.org/10.1126/science.1181498

Cynthia Dwork. 2006. Differential Privacy. In *Proc. 33rd Int. Colloq. on Automata, Languages and Programming (ICALP 2006) (LNCS 4052)*. Springer, Venice, Italy, 1–12.

Yaniv Erlich and Arvind Narayanan. 2014. Routes for Breaching and Protecting Genetic Privacy. *Nature Reviews Genetics* 15, 6 (June 2014), 409–421. DOI:http://dx.doi.org/10.1038/nrg3723

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *USENIX Security*. USENIX.

Julien Freudiger, Maxim Raya, Márk Félegyházi, Panos Papadimitratos, and Jean-Pierre Hubaux. 2007. Mix-Zones for Location Privacy in Vehicular Networks. In *Proc. 1st Int. Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS 2007)*. ICST, Vancouver, Canada.

Michael T. Goodrich. 2009. The Mastermind Attack on Genomic Data. In *30th IEEE Symposium on Security and Privacy*. 204–218.

Scott Gottlieb. 2001. US Employer Agrees to Stop Genetic Testing. *BMJ : British Medical Journal* 322, 7284 (Feb. 2001), 449.

Bastian Greshake, Philipp E. Bayer, Helge Rausch, and Julia Reda. 2014. openSNP–A Crowdsourced Web Resource for Personal Genomics. *PLoS ONE* 9, 3 (March 2014). DOI:http://dx.doi.org/10.1371/journal.pone.0089204

Daojing He, S. Chan, and M. Guizani. 2015. Privacy and incentive mechanisms in people-centric sensing networks. *IEEE Communications Magazine* 53, 10 (2015), 200–206. DOI:http://dx.doi.org/10.1109/MCOM.2015.7295484

Jerry L. Hintze and Ray D. Nelson. 1998. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician* 52, 2 (May 1998), 181–184.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics* 4, 8 (August 2008), e1000167.

Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2013. Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy. In *Proc. 20th ACM Conf. on Computer and Communications Security (CCS'13)*. ACM, Berlin, Germany, 1141–1152. DOI:http://dx.doi.org/10.1145/2508859.2516707

Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti. 2014. Reconciling Utility with Privacy in Genomics. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES '14)*. ACM, Scottsdale, AZ, USA, 11–20.

Mathias Humbert, Kévin Huguenin, Joachim Hugonot, Erman Ayday, and Jean-Pierre Hubaux. 2015. De-anonymizing Genomic Databases Using Phenotypic Traits. DOI:http://dx.doi.org/10.1515/popets-2015-0020

Georgios Kalogridis, Costas Efthymiou, Stojan Z. Denic, Tim A. Lewis, and Rafael Cepeda. 2010. Privacy for Smart Meters: Towards Undetectable Appliance Load Signatures. In *Proc. 1st Int. Conf. on Smart Grid Communications (SmartGridComm 2010)*. IEEE, Gaithersburg, MD, USA, 232–237.

Zhen Lin, Michael Hewett, and Russ B. Altman. 2002. Using Binning to Maintain Confidentiality of Medical Data. In *Proc. AMIA Symp. (AMIA 2002)*. San Antonio, TX, USA, 454–458.

Changchang Liu and Prateek Mittal. 2016. LinkMirage: Enabling Privacy-preserving Analytics on Social Relationships. In *NDSS*.

Bradley A. Malin. 2005. Protecting DNA sequence anonymity with generalization lattices. *Methods of Information in Medicine* 44, 5 (2005), 687–692.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98, 5 (May 2007), 873–895.

Steven J. Murdoch. 2014. Quantifying and Measuring Anonymity. In *Data Privacy Management and Autonomous Spontaneous Security*. Springer Berlin Heidelberg, 3–13.

Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing Social Networks. In *30th IEEE Symposium on Security and Privacy*. 173–187. DOI:http://dx.doi.org/10.1109/SP.2009.22

Muhammad Naveed, Erman Ayday, Ellen W. Clayton, Jacques Fellay, Carl A. Gunter, Jean-Pierre Hubaux, Bradley A. Malin, and Xiaofeng Wang. 2015. Privacy in the Genomic Era. *ACM Comput. Surv.* 48, 1 (Aug. 2015), 6:1–6:44. DOI:http://dx.doi.org/10.1145/2767007

Dale R Nyholt, Chang-En Yu, and Peter M Visscher. 2009. On Jim Watson's APOE Status: Genetic Information Is Hard to Hide. *European Journal of Human Genetics* 17, 2 (Feb. 2009), 147–149. DOI:http://dx.doi.org/10.1038/ejhg.2008.198

Simon Oya, Carmela Troncoso, and Fernando Pérez-González. 2014. Do Dummies Pay Off? Limits of Dummy Traffic Protection in Anonymous Communications. In *Proc. 14th Int. Symp. on Privacy Enhancing Technologies (PETS 2014) (LNCS 8555)*. Springer, Amsterdam, Netherlands, 204–223.

Ravi Sachidanandam, David Weissman, Steven C. Schmidt, Jerzy M. Kakol, Lincoln D. Stein, Gabor Marth, Steve Sherry, James C. Mullikin, Beverley J. Mortimore, David L. Willey, Sarah E. Hunt, Charlotte G. Cole, Penny C. Coggill, Catherine M. Rice, Zemin Ning, Jane Rogers, David R. Bentley, Pui-Yan Kwok, Elaine R. Mardis, Raymond T. Yeh, Brian Schultz, Lisa Cook, Ruth Davenport, Michael Dante, Lucinda Fulton, LaDeana Hillier, Robert H. Waterston, John D. McPherson, Brian Gilman, Stephen Schaffner, William J. Van Etten, David Reich, John Higgins, Mark J. Daly, Brendan Blumenstiel, Jennifer Baldwin, Nicole Stange-Thomann, Michael C. Zody, Lauren Linton, Eric S. Lander, and David Altshuler. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 6822 (Feb. 2001), 928–933. DOI:http://dx.doi.org/10.1038/35057149

Sahel Samani, Zhicong Huang, Erman Ayday, Mark Elliot, Jacques Fellay, Jean-Pierre Hubaux, and Zoltán Kutalik. 2015. Quantifying Genomic Privacy via Inference Attack with High-Order SNV Correlations. In *2015 IEEE Security and Privacy Workshops (SPW)*. 32–40.

Andrei Serjantov and George Danezis. 2002. Towards an Information Theoretic Metric for Anonymity. In *Proc. 2nd Int. Symp. on Privacy Enhancing Technologies (PETS 2002) (LNCS 2482)*. Springer, San Francisco, CA, USA, 41–53.

S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29, 1 (January 2001), 308–311. DOI:http://dx.doi.org/10.1093/nar/29.1.308

Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying Location Privacy. In *Proc. 2011 32nd IEEE Symp. on Security and Privacy (S&P 2011)*. IEEE, Oakland, CA, USA, 247–262. DOI:http://dx.doi.org/10.1109/SP.2011.18

Montgomery Slatkin. 2008. Linkage Disequilibrium —Understanding the Evolutionary Past and Mapping the Medical Future. *Nature Reviews Genetics* 9, 6 (June 2008), 477–485. DOI:http://dx.doi.org/10.1038/nrg2361

Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570. DOI:http://dx.doi.org/10.1142/S0218488502001648

Paul Syverson. 2013. Why I'm Not an Entropist. In *Proc. 17th Int. Workshop on Security Protocols (LNCS 7028)*. Springer, Cambridge, UK, 213–230.

The 1000 Genomes Project Consortium. 2015. A Global Reference for Human Genetic Variation. *Nature* 526, 7571 (Oct. 2015), 68–74. DOI:http://dx.doi.org/10.1038/nature15393

Sarah A. Tishkoff and Kenneth K. Kidd. 2004. Implications of biogeography of human populations for 'race' and medicine. *Nature Genetics* 36 (Oct. 2004), S21–S27. DOI:http://dx.doi.org/10.1038/ng1438

Isabel Wagner. 2015. Genomic Privacy Metrics: A Systematic Comparison. In *2015 IEEE Security and Privacy Workshops (SPW)*. 50–59. DOI:http://dx.doi.org/10.1109/SPW.2015.15

Isabel Wagner and David Eckhoff. 2015. Technical Privacy Metrics: a Systematic Survey. *arXiv:1512.00327 [cs, math]* (Dec. 2015). http://arxiv.org/abs/1512.00327

Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 2009. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study. In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS '09)*. ACM, Chicago, IL, USA, 534–544.

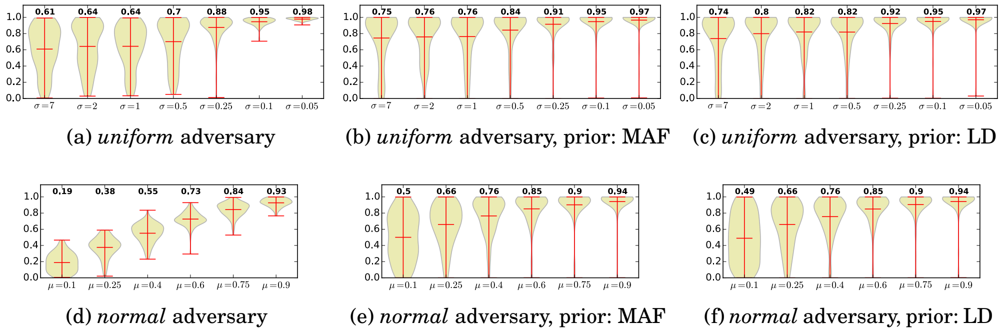
Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. 2014. The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations. *Nucleic Acids Research* 42, D1 (Jan. 2014), D1001–D1006. DOI:http://dx.doi.org/10.1093/nar/gkt1229

Kris Wetterstrand. 2016. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). (2016). https://www.genome.gov/sequencingcostsdata

Ye Zhu and Riccardo Bettati. 2005. Anonymity vs. information leakage in anonymity systems. In *Proc. 25th IEEE Int. Conf. on Distributed Computing Systems (ICDCS 2005)*. IEEE, Columbus, Ohio, USA, 514–524.

APPENDIX

A. KIN ADVERSARY



(a) *uniform* adversary

(b) *uniform* adversary, prior: MAF

(c) *uniform* adversary, prior: LD

(d) *normal* adversary

(e) *normal* adversary, prior: MAF

(f) *normal* adversary, prior: LD

Fig. 14. Average probability the different *kin* adversary types assign to the true genotype for a sample individual

B. SNPS AND POPULATION GROUPS FOR HEAT MAP IN FIGURE 3

Table IV. SNPs and population groups used to produce the heat map in Figure 3. The first line indicates the number of SNPs corresponding to each cell in the heat map and the second line indicates whether *mixed* minor allele frequencies or minor allele frequencies for the *correct* population group were used (where relevant for the adversary model). Note that when we used 60 SNPs, these are the SNPs for the three case studies combined.

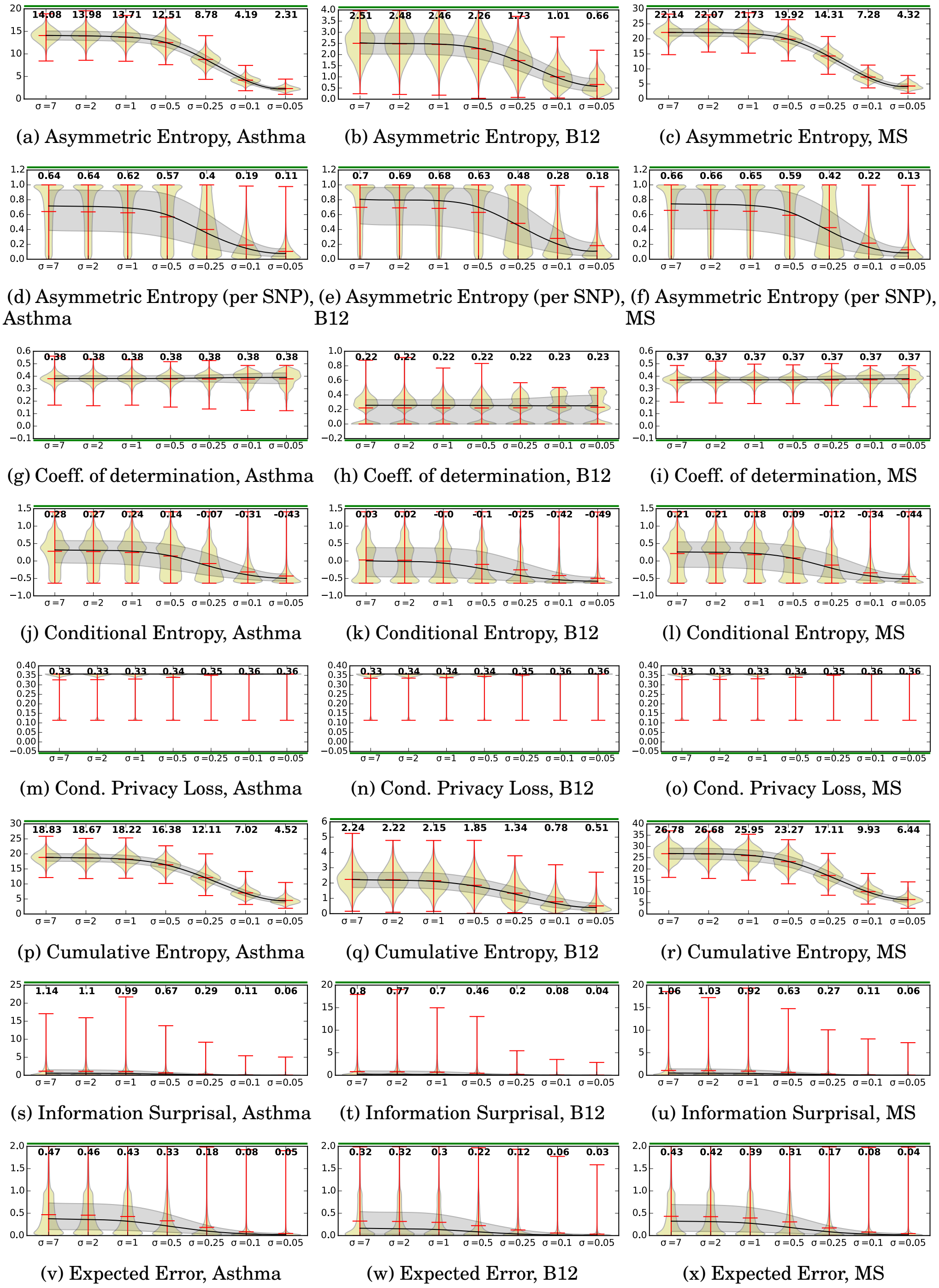
		Adversary Model					
		normal	uniform	normal+MAF	uniform+MAF	normal+LD	uniform+LD
Scenario	comparison	all –	all –	all mixed	all mixed	60 correct	60 correct
	kin/opensnp	all –	all –	60 correct	60 correct	60 correct	60 correct
	kin/utah	all –	all –	60 correct	60 correct	60 correct	60 correct
	case/b12	4 –	4 –	4 correct	4 correct	4 correct	4 correct
	case/asthma	22 –	22 –	22 correct	22 correct	22 correct	22 correct
	case/ms	34 –	34 –	34 correct	34 correct	34 correct	34 correct

C. SNPS AND WEIGHTS FOR CASE STUDIES

Table V. SNPs and weights for case studies (data from the NHGRI GWAS Catalog [Welter et al. 2014])

Multiple Sclerosis		Asthma		Vitamin B12 levels	
rs4613763	1.2	rs2284033	1.12	rs492602	-0.09
rs874628	1.11	rs7686660	1.16	rs10515552	43.93
rs3118470	1.12	rs7216389	1.45	rs2298585	71.8
rs2300603	1.11	rs3771166	1.15	rs3760776	49.78
rs13192841	1.1	rs2069408	1.15		
rs6897932	1.11	rs3129890	1.15		
rs7238078	1.12	rs9268516	1.15		
rs2303759	1.11	rs3117098	1.16		
rs9271366	2.78	rs2786098	1.43		
rs2546890	1.11	rs9275698	1.18		
rs2119704	1.22	rs204993	1.17		
rs3135388	2.75	rs1837253	1.17		
rs7595037	1.11	rs1701704	1.19		
rs3129889	2.97	rs744910	1.12		
rs1077667	1.16	rs10508372	1.16		
rs4902647	1.11	rs3129943	1.17		
rs11581062	1.12	rs7130588	1.09		
rs669607	1.13	rs987870	1.4		
rs3129934	3.3	rs7775228	1.17		
rs7090512	1.19	rs4129267	1.09		
rs771767	1.1	rs404860	1.21		
rs12722489	1.23	rs9500927	1.13		
rs9282641	1.21				
rs17174870	1.11				
rs1738074	1.13				
rs12466022	1.11				
rs9292777	1.19				
rs140522	1.1				
rs2248359	1.12				
rs7923837	1.1				
rs11154801	1.13				
rs1800693	1.12				
rs2019960	1.12				
rs3135338	3.43				

D. SUPPLEMENTARY FIGURES



(a) Asymmetric Entropy, Asthma (b) Asymmetric Entropy, B12 (c) Asymmetric Entropy, MS

(d) Asymmetric Entropy (per SNP), Asthma (e) Asymmetric Entropy (per SNP), B12 (f) Asymmetric Entropy (per SNP), MS

(g) Coeff. of determination, Asthma (h) Coeff. of determination, B12 (i) Coeff. of determination, MS

(j) Conditional Entropy, Asthma (k) Conditional Entropy, B12 (l) Conditional Entropy, MS

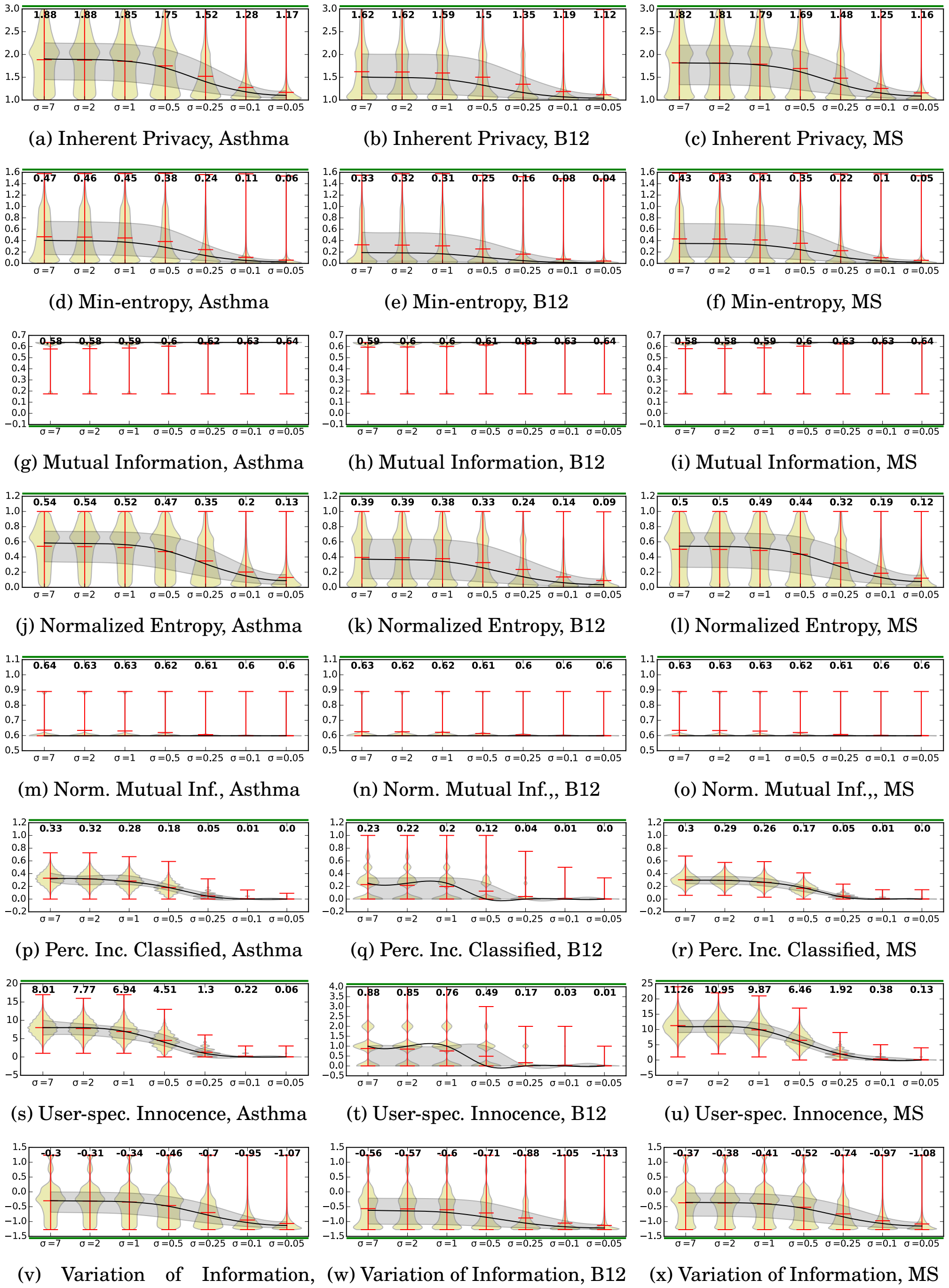
(m) Cond. Privacy Loss, Asthma (n) Cond. Privacy Loss, B12 (o) Cond. Privacy Loss, MS

(p) Cumulative Entropy, Asthma (q) Cumulative Entropy, B12 (r) Cumulative Entropy, MS

(s) Information Surprisal, Asthma (t) Information Surprisal, B12 (u) Information Surprisal, MS

(v) Expected Error, Asthma (w) Expected Error, B12 (x) Expected Error, MS

Fig. 15. Privacy metrics evaluated according to adversary strength for the *uniform* adversary with minor allele frequencies as prior information. Green bars show whether high or low values indicate high privacy.



(a) Inherent Privacy, Asthma

(b) Inherent Privacy, B12

(c) Inherent Privacy, MS

(d) Min-entropy, Asthma

(e) Min-entropy, B12

(f) Min-entropy, MS

(g) Mutual Information, Asthma

(h) Mutual Information, B12

(i) Mutual Information, MS

(j) Normalized Entropy, Asthma

(k) Normalized Entropy, B12

(l) Normalized Entropy, MS

(m) Norm. Mutual Inf., Asthma

(n) Norm. Mutual Inf.,, B12

(o) Norm. Mutual Inf.,, MS

(p) Perc. Inc. Classified, Asthma

(q) Perc. Inc. Classified, B12

(r) Perc. Inc. Classified, MS

(s) User-spec. Innocence, Asthma

(t) User-spec. Innocence, B12

(u) User-spec. Innocence, MS

(v) Variation of Information, Asthma

(w) Variation of Information, B12

(x) Variation of Information, MS

Fig. 16. Privacy metrics evaluated according to adversary strength for the *uniform* adversary with minor allele frequencies as prior information. Green bars show whether high or low values indicate high privacy.