ゲノム検査結果の開示によるプライバシ侵害の評価

荒井　ひろみ†　　津田　宏治‡　　佐久間　淳§

†東京大学情報基盤センター
113-8658 東京都文京区弥生 2-11-16
arai@dl.itc.u-tokyo.ac.jp

‡東京大学新領域創成科学研究科
277-8561 千葉県柏市柏の葉 5-1- 5
tsuda@k.u-tokyo.ac.jp

§筑波大学 大学院 システム情報工学研究科 / JST CREST
305-8573 茨城県つくば市 天王台 1-1-1　jun@cs.tsukuba.ac.jp

あらまし　個人ゲノムはプライバシに関わる情報を多く含み，その利用において適切にプライバシを保護する必要がある．個人ゲノム利用は近年盛んになってきており，その応用の一つにゲノム検査がある．我々は情報を適切に運用するための情報開示によるプライベート情報漏えいの監査に着目する．我々はゲノム検査における疾患リスクの開示による個人ゲノムのプライバシ侵害を線形計画問題として定式化し，侵害度合いを定量評価する方法を提案する．さらに，実社会データにおける疾患リスクの開示によるプライバシ侵害を示す．

Evaluation of Privacy Breaches from Genetic Testing Results

Hiromi Arai†　　Koji Tsuda‡　　Jun Sakuma§

†Information Technology Center, The University of Tokyo
2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-8658, JAPAN
‡Graduate School of Frontier Sciences, The University of Tokyo
5-1-5 Kashiwa-no-ha, Kashiwa, Chiba 277-8561, JAPAN
§Graduate School of SIE, University of Tsukuba / JST CREST
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573, JAPAN

Abstract　Personal genome contains much private information so that genome privacy should be preserved appropriately. Personal genome services have been spreading rapidly. One of the major application is the genetic testing services. In this paper, we focus on the auditing of privacy breaches by the information disclosure. We formulate the auditing of privacy breaches from the trait risks in the genetic testings as a linear programming problem. Then, we propose the quantification method for the privacy breaches. We demonstrate on real-world datasets the genetic information breaches from trait risks.

1　はじめに

個人に関する情報の利用が進む現在，プライバシ保護は情報の適切な利用に重要な課題である．なかでもパーソナルゲノムは，個人に関する情報を多く含み取り扱いに慎重を要する．しかし近年パーソナルゲノムの一般利用は急速に進んでいる．病院や DTC のゲノム検査に代表されるゲノム利用サービスが普及しており，個人が自身のゲノムやその二次情報を容易に入手，利用することが可能になってきた．そのため，それらの情報のプライバシレベルに応じた扱いの指針を立てる必要がある．

ゲノム情報はそれ自身に個人特定性および機微性両方を持つ秘匿性の高い情報である．ゲノムは個人固有のパターンを持つため個人照合などに利用できる．また，がんになりやすい体質など疾患や体質に関する機微情報を含む．この疾患リスク情報をゲノムから読み取るルールはゲノムワイド解析 (GWAS) の成果から得られ，未知のルールが日々発見されている．

ゲノム情報はその秘匿性の高さにも関わらず，不用意に公開されてしまう例も後を絶たない．特に，ゲノム検査結果などの明示的にゲノムが記載されていない二次情報を SNS などにおいて公開してしまうような場合がある [7]．

ゲノム情報などの個人情報の保護のためには，データを公開したときの情報漏えいリスクを適切に評価し，それに応じた情報開示を行うことが重要である．本稿ではこれに着目し，ゲノム検査結果の疾患リスクを開示した場合に漏洩するゲノム情報を定量的に評価する．

関連研究として，2 次情報からのゲノム情報推定の研究は，GWAS 報告の統計値からの個人情報の推定 [2,3,10]，ゲノム内の関連構造情報を用いた個人ゲノム上の欠損情報の推定 [6,8] などがある．これらは，各問題についてそれに適した推定モデルを仮定したものであり，複雑な情報処理を行ったデータから元のデータを推定できるもののモデルを用いることによる推定バイアスがある．またゲノム検査による情報漏洩の評価には直接適用することはできない．

また，集約情報から漏洩する元の情報を推定する問題はクエリ監査問題 [5] として研究されてきた．しかし，ビット値のデータベースにおける集約クエリ等の単純なデータモデルなどの扱いに留まり，このモデルではゲノム検査結果の監査に直接適用することはできず，さらに計算効率が悪いという問題点がある．

本稿では，疾患リスクを開示したときの情報漏えいを評価するための手法を提案，実データを用いて漏洩を検証する．これは，モデルを使わずに直接情報漏洩を評価する．また計算効率の良い方法を取り従来のリスク監査の問題を克服する．第 2 章では疾患リスクの計算方法について概説し，第 3 章で疾患リスクの計算のような複雑な情報処理から元のデータを推定する問題を提案する．第 4 章は疾患リスクから元のゲノムの情報を推測する具体的な計算方法を導入する．第 5 章で，さらに元のゲノム情報から重篤な疾患に関する情報が漏洩する問題を扱う．第 6 章でこれらの漏洩リスクを統計的に評価する基準を導入し，第 7 章で実データを用いて情報漏えいを検証する．

2 疾患リスク計算法

ゲノム検査では個人のゲノムからゲノムに由来する疾患リスクを算定する．ゲノム情報は塩基と呼ばれる 4 種の分子からなる鎖状の高分子であるゲノムにコードされており，その情報は塩基を表す A, T, G, C の 4 文字の文字列として表現することができる．ヒトゲノムは集団間で多くの部分が共通しているが，部分的に多様性が存在する．中でも一文字の違いで表される多様性は一塩基多型（SNP）と呼ばれる．ある SNP における特定の塩基はアレルと呼ばれ，ある体質や疾患のなりやすさと強く関連するアレルはリスクアレルと呼ばれる．例えば，BRCA1 遺伝子における rs222795 という ID で参照される SNP において，”G”がリスクアレルであり，これが存在すると乳がんになる確率が高くなる．個人ゲノムの大部分をしめる常染色体は 23 対存在し，SNP も文字の対で表される．よって上記の例で SNP が ”GG”，”TG ”の場合にリスクアレルはそれぞれ 2 つ，1 つ存在することになる．ある被験者におけるこのような情報をリスクアレル数と呼ぶことにする．ゲノム検査は，このようなリスクアレル数を様々な疾患，SNP において検査し，被験者がある特徴や疾患を持つ確率を評価するものである．

本研究では以下に示すゲノム検査方法を用いる．ある被験者を想定し，被験者が属する人種集団における疾患の相対リスクを被験者の SNP 情報から計算する．検査対象の疾患，体質，特徴などを疾患と総称し，それらの集合を $T_T = \{t_1, \ldots, t_{n_T}\}$ とする．また，疾患のある人種における相対リスクを疾患リスクと呼ぶことにする．SNPi における 1 リスクアレルによって，t_k

の相対リスクが確率 a_{ki} 増えるとする．リスクアレルが複数ある場合は，相対リスクは積算されるとする．これは multiplicative model [4] と呼ばれゲノム検査ではよく用いられるモデルである．ある人種における平均的な相対リスクを $\bar{a}_{k_i}$ とする．SNPi における被験者のリスクアレルの数を $x_i \in \{0,1,2\}$，検査で用いられる全 SNP を L_T とすると，被験者の t_k の疾患リスクは

$$r'_k = \prod_{i \in L_T} s_i = \prod_{i \in L_T} \frac{a_{ki}^{x_i}}{\bar{a}_{ki}}.$$

と計算される．ここで，疾患リスクはサービスにおいて，可読性のために

$$r_k = \mathtt{round}(r'_k, b) = \lfloor r'_k/b + 1/2 \rfloor \cdot b,$$

のように丸めて用いられるとする．通常丸めパラメータは $b \in [0.001, 0.01]$ である．これは，副次的なプライバシ保護の効果もある程度期待される．ここで，リスクアレル数のベクトル表現を $\mathbf{x} = (x_1, \ldots, x_{|L_T|})$，疾患リスク計算関数を

$$f_k(\mathbf{x}) = \mathtt{round}(\prod_{i \in L_T} \frac{{a_{ki}}^{x_i}}{\bar{a}_{ki}}, b). \qquad (1)$$

とする．被験者は T_T に対応する複数の疾患リスクのセット $S_T = \{r_1, \ldots, r_{n_T}\}$ を受け取る．

3 攻撃モデル

被験者が疾患リスクのセット S_T を公開した場合の情報漏洩を評価するために，以下の攻撃モデルを想定する．攻撃シナリオとして，攻撃者が疾患リスクの計算式 1 を知っているとする．多くのゲノム検査サービスで，検査方法は white paper やデモシステムなどで開示されているためこの想定は妥当である．被験者は，個人のゲノム情報自身や，特に重篤な疾患に関連する遺伝情報を公開したくないが不用意に S_T を公開してしまったとする．この場合に攻撃者が推測する情報を評価する．攻撃者はゲノム検査の入力である被験者のリスクアレルの情報を，ゲノム検査関数と出力である S_T から推測する以下の攻撃を行うとする．

Statement 1 (アレル推定攻撃) 被験者がリスクアレル数の集合 $\{x_i | i \in L_T\}$ を秘匿し，それを用いたゲノム検査結果 S_T を公開する．攻撃者は被験者のリスクアレル数 $\{x_i | i \in L_T\}$ を $\{a_{ki} | k \in \{1, \ldots, n_T\}, i \in L_T\}$ と S_T を用いて推論し，その推定値 $\{E(x_i) | i \in L_T\}$ を出力する．

さらに，攻撃者はアレル推定の結果から被験者の持つ重篤な疾患 T_S に関する疾患リスクを推測するとする．これを以下のように定義する．

Statement 2 (センシティブ情報推定攻撃) 攻撃者が被験者のリスクアレル数の推定値 $\{E(x_i) | i \in L_T\}$ を持っているとする．攻撃者は重篤な疾患 T_S に関するリスクアレルを被験者が持っているかどうかを $\{E(x_i) | i \in L_T\}$ から推論する．

これらの問題についての具体的な評価方法を以下の章に述べる．

4 アレル推定評価法

アレル推定を我々は以下に示す方法によって定量的に評価する．我々は $\mathbf{x}$ の推測として，S_T を取りうるデータベースを評価する方法をとる．可能なリスクアレルの集合を $\mathcal{X} = \{0,1,2\}^{|L_T|}$ とする．このうち，S_T を取りうる $\tilde{\mathbf{x}} \in \mathcal{X}$ は

$$\mathcal{X}_r = \{\tilde{\mathbf{x}} \in \mathcal{X} | \{f_k(\tilde{\mathbf{x}}) = r_k | k \in T\}\}.$$

さらに，x_i がある値をとるサブセットは

$$\mathcal{X}_{iv} = \{\tilde{\mathbf{x}} \in \mathcal{X}_r | \tilde{x}_i = v\} \quad (v \in \{0,1,2\})$$

である．ここで，$\tilde{\mathbf{x}}$ は一様確率で $\mathcal{X}$ に分布しているとすると

$$\Pr(\mathbf{x} \in \mathcal{X}, x_i = v | S) = \frac{|\mathcal{X}_{jv}|}{|\mathcal{X}_r|} \qquad (2)$$

となる．よって SNPi におけるリスクアレル数の期待値は

$$E(x_i) = \sum_{v \in \{0,1,2\}} \Pr(x_i = v | S) \cdot v. \qquad (3)$$

となる．

ここで，$|\mathcal{X}_{iv}|$ および $|\mathcal{X}_r|$ の計算は安易に $\mathcal{X}$ の全要素に対しブルートフォースよって求めると，$|L_T|$ に対して幾何級数的な時間を要する．ゲノム検査サービスでは将来的に $|L_T|$ がかなり大きくなる可能性も考えられ，このアプローチは計算コストが掛かり過ぎると懸念される．よって，我々は $\mathcal{X}_{iv}$，$\mathcal{X}_r$ の条件を不等式制約で表現し，効率的に計算する方法を導入する．

式 1 を満たす $\tilde{\mathbf{x}}$ の条件は，すべての k について

$$f_k(\tilde{\mathbf{x}}) - b/2 \leq \prod_{i \in L_T} a_i^{(k)^{\tilde{x}_i}} / \bar{a}_i^{(k)} < f_k(\tilde{\mathbf{x}}) + b/2,$$

が成り立つことと等価である．両辺の対数をとると

$$\begin{aligned} &\log(f_k(\tilde{\mathbf{x}}) + b/2) \geq \\ &\quad \sum_{i \in L_T} (\log(a_i^{(k)}) \cdot \tilde{x}_i - \log(\bar{a}_i^{(k)})), \\ &\log(f_k(\tilde{\mathbf{x}}) - b/2) < \\ &\quad \sum_{i \in L_T} (\log(a_i^{(k)}) \cdot \tilde{x}_i - \log(\bar{a}_i^{(k)})), \end{aligned} \tag{4}$$

となる．これらは線形不等式制約である．$\mathcal{X}_{iv}$ については線形等式制約 $x_i = v$ が追加される．よって，$|\mathcal{X}_r|$，$|X_{iv}|$ の計算には整数線形計画問題として効率的に解くことができる．我々は上記の制約条件を $\{0,1\}$ ベクトルの形に書き直し，azove 2.0 [1] を用いた．これは，現時点で最も効率的に近似を行わずに計算できる整数線形計画ソルバーの一つである．

5　センシティブ情報漏えい

GWAS 研究から多くの疾患-SNP の関連情報が得られており，今後も未知の関連情報が得られることが予想される．その疾患情報の中には身長や髪の色等のそれほどセンシティブではないものから癌やアルツハイマーなどの，他人に知られることで不利益を被るようなセンシティブな情報がある．

ここで，センシティブでない疾患を T_{NS}，センシティブな疾患を T_S とする．センシティブでない疾患 T_{NS} の疾患リスクから，センシティブな疾患 T_S の疾患リスクが推定される問題を考える．

T_S のうちいずれかの疾患と関連する SNP を考える．疾患 t_k と関連する SNP 集合を L_k とすると，T_S，T_{NS} と関連する SNP はそれぞれ $L_S = \cup_{t_k \in T_S} L_k$，$L_{NS} = \cup_{t_k \in T_{NS}} L_k$ となる．また，L_S と相関する SNP 群も考える．遺伝子には連鎖不均衡 (LD) という，集団内で複数の遺伝子多型の間にランダムでない相関が見られるという現象がある．ヒトのある人種における SNP 間の LD を考える．SNP 対の相関は連鎖不平衡係数 r^2 で評価でき，$r^2 > 0.6$ であると強い相関があると評価される [6]．SNPℓ と j の間の連鎖不平衡係数を $r^2_{\ell j}$ と示す．ここで，被験者は L_S と連鎖不平衡にある SNP

$$L_{LD} = \{\ell | r^2_{\ell j} > 0.6, j \in L_S\},$$

のいずれかを持つ場合，それと LD にある疾患因子を持つと疑われると考えられる．

本論文では，ゲノム検査において，センシティブでない疾患の疾患リスクから推定される上記のセンシティブなリスクアレルの情報を取り扱う．ゲノム検査におけるセンシティブでない疾患は $T_T \cap T_{NS}$，それから推測できるリスクアレルは $L_T \cap L_{NS}$ である．それらから推測できるセンシティブ情報は，センシティブな疾患と関連する SNP

$$L_{TS} = (L_T \cap L_{NS}) \cap L_S$$

及び，センシティブな疾患と強い相関を持つ SNP

$$L_{TLD} = (L_T \cap L_{NS}) \cap L_{LD}$$

とし，これらのリスクアレル情報が正確に推測されるならばセンシティブ情報が漏えいしたと考える．

6　プライバシ評価基準

推定攻撃によるプライバシ漏洩を図るための規準を導入する．まず，アレル推定攻撃による情報漏洩は，各被験者の各リスクアレルについてエラー $Err(x_i) = |E(x_i) - x_i|$ が小さいほど大きいと考えられる．

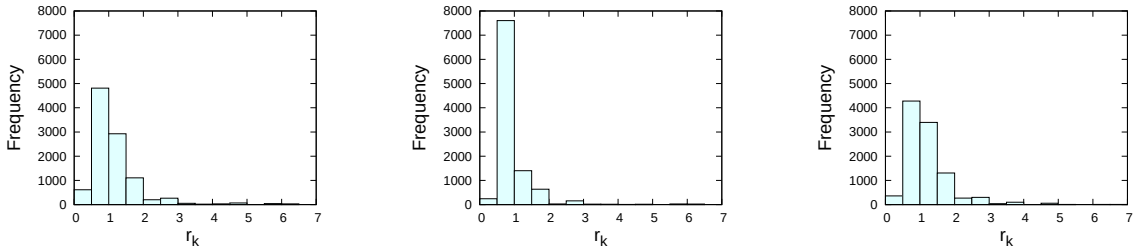

図 1: T_T に関する丸めを行わない場合の疾患リスク値の分布．左から CEU, JPT, YRI データセットについて示す．横軸が疾患リスク値，縦軸が頻度を示す

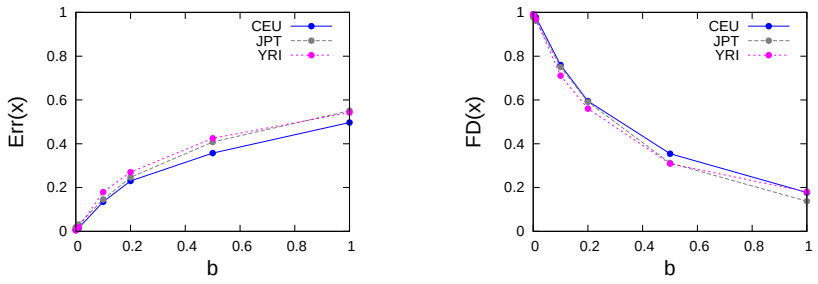

図 2: L_T について CEU, JPT, YRI データセットでのアレル推定攻撃．縦軸は推定のエラー平均 $Err(x)$(左) 及び完全漏洩率 $FD(x)$(右)．横軸は丸めパラメータ b.

ある人口集団において，S_T の開示における平均的なプライバシ漏洩の評価を考える．各人のインデックスの集合を I_X，集団の各人 j のリスクアレルベクトル $\mathbf{x}^{(j)}$ の集合を X とする．ある対象 SNP 集合 L_t について，X における平均的なエラーは

$$Err(x) = \frac{1}{|I_X| \cdot |L_t|} \sum_{j \in I_X} \sum_{i \in L_t} Err(x_i^{(j)}), \quad (5)$$

となる．また，完全に情報が漏洩する割合を，完全漏洩率として，

$$FD(x) = \frac{1}{|I_X| \cdot |L_t|} \sum_{j \in I_X} \sum_{i \in L_t} \delta(Err(x_i^{(j)})),$$

として評価し，この値が大きいほどプライバシ漏洩が大きいとする．

また，アレル推定攻撃における複数疾患リスクを同時に評価する効果を考察するため，式 4 を，各 k ごとに評価した場合のエラーを

$$E_s(x_i) = \frac{1}{|T_C|} \sum_{r_k \in T_C} \Big(\sum_{v \in \{0,1,2\}} \Pr(x_i = v | r_k) \cdot v \Big). \quad (6)$$

とし，すべての k についての条件を同時に評価した場合の式 3 と比較する．

7 評価実験

本章ではゲノム検査結果の疾患リスクの開示におけるプライバシ漏洩を実験的に評価する．ゲノム検査被験者としての人口集団のデータセットを作成し，また疾患リスク計算のための SNP-疾患関係を収集した．それらを用いてアレル推定攻撃によるリスクアレル情報の漏洩及びセンシティブ情報の漏洩を評価した．

7.1 データセット

7.1.1 人口集団データセット

被験者の集団として人種別の人口集団のリスクアレルデータを作成する．人種別集団における各個人 SNP 情報のデータベースである HapMap データベース [9] を用い，CEU，JPT，YRI と呼ばれる人口集団における各 174，96，176 人についてのリスクアレル数のデータセットを作成した．リスクアレルは次節にのべる疾患リスク計算で用いる $|L_T| = 119$SNP について収集した．これらを各々CEU，JPT，YRI データセットとする．

表 1: **複数疾患リスクによる漏洩情報の増加例**

リスクアレル	アレル数	推定アレル数			相対リスク値	
		c_1 and c_2	c_1	c_2	c_1	c_2
rs7335046-G	0	0.25	-	0.879	-	1.26
rs1805007-T	2	2	2	0.697	4.37	1.55
rs12210050-T	0	0.25	-	0.909	-	1.24
rs7538876-A	0	0.25	-	0.818	-	1.28
rs1540771-A	0	0	0	-	1.40	-
rs1042602-C	1	1	0.5	-	1.32	-

CEU データセットにおけるある被験者のリスクアレル数と，そばかす (c_1)，バーゼル細胞癌 (c_2) 及び両方の疾患リスクから推定されたリスクアレル数．また，c_1，c_2 に対する各アレルの相対リスク値を示した．

7.1.2 疾患リスク計算パラメータ

疾患リスク計算法のパラメータである SNP と疾患の関連情報及び相対リスク値を GWAS Catalog [11] の 2013/12/09 時点のデータを用い収集した．これは，GWAS 研究の成果をキュレートしたデータベースであり，ここから，信頼できる情報を以下のように収集した．まず，インパクトの高いジャーナルに掲載された研究を選び，さらに P 値が $5 \cdot 10^{-5}$ 以下である疾患-SNP の関連を選択した．これは GWAS 研究の評価におけるスタンダードなしきい値である．

ここから，関連する SNP が 2 から 5 個である疾患を選択し，最終的に T_T として 56 疾患，$|L_T| = 119$ 種のリスクアレル及び各リスクアレルについての相対リスク値を得た．これらのうち 4 疾患は共通のリスクアレルを持ち，疾患と SNP をそれぞれ T_C，$L_C = \cup_{t_k \in T_C} L_k$ とする．

なお，各人口集団における疾患リスクの値の分布を図 1 に示す．人種間で分布に差が存在するがおおよそ $[0, 3]$ に分布している．よって，疾患リスク情報が有用であるのは $b \leq 0.5$ 程度であると考察される．

7.1.3 センシティブアレルデータセット

GWAS Catalog から T_S 及び T_{NS} をそれぞれ 67 及び 37 疾患収集し，それらと関連するリスクアレル L_S 及び L_{NS} を収集した．また，SNP ペアと r^2 の情報を HapMap データベース [9] における LD データベースから収集し，L_{LD} を作成した．$|L_{TS}| = 4$，$|L_{TLD}|$ は CEU，JPT，YRI それぞれについて $2, 4, 1$ であった．

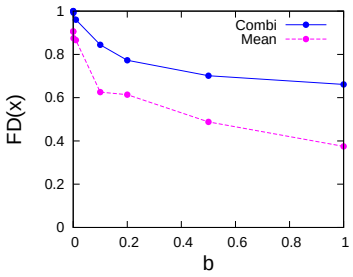

図 3: リスク推定攻撃における疾患リスクの組み合わせ効果，CEU データセットにおける T_C によるプライバシ漏洩を式 3(combi) 及び式 6(mean) で評価した場合の完全漏洩率の変化．

7.2 実験結果

アレル推定攻撃によるリスクアレル数の推定結果は以下のようであった．b は 10^{-3}，10^{-2}，0.1，0.2，0.5，1 と変化させた．丸めパラメータ b に対する変化を図 2 に示す．丸めパラメータが大きいほど，エラーは大きくなり完全漏洩率は減少し，プライバシ漏洩は少なくなる．一方で，$b \leq 0.5$ と十分小さい場合には，ほとんどのリスクアレルの情報が正確に推定されてしまうため疾患リスクの開示における有用性とプライバシ保護の両立は困難であると考察される．結果は CEU, YRI, JPT データセットについて同様の結果を得た．

疾患 T_C の疾患リスクによるリスクアレルの推定結果を図 3 に示す．複数の疾患リスクの効果を同時に評価した場合の方が推定精度があがり，情報の組み合わせ効果によってよりリスクアレル情報が漏洩することが示された．さらに，組み合わせ効果の例を表 1 に示す．結果は CEU

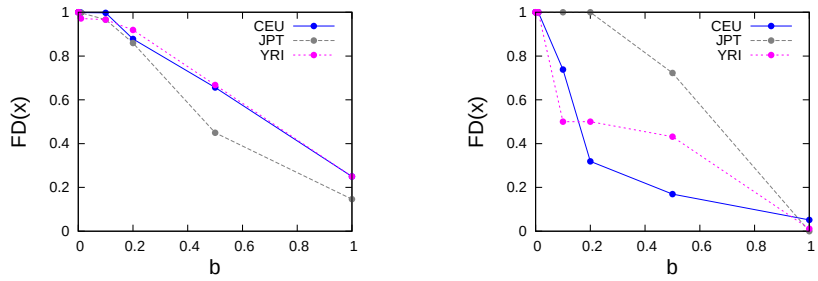

図 4: CEU，JPT，YRI 各データセットについてのセンシティブ情報推定攻撃．L_{TS}(左) 及び L_{TLD}(右) における完全漏洩率の丸めパラメータに対する変化．

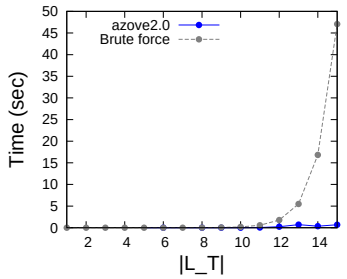

図 5: 対象リスクアレル数の変化に対するアレル推定攻撃の 1 リスクアレルに対する計算時間の変化．azove 2.0 を用いた場合と，brute force の場合それぞれの 100 回平均値を示す．

データセットのみについて記すが，他のデータセットにおいても同様の結果を得た．

センシティブ情報の各人口集団データセットにおける完全漏洩率を図 4 に示す．完全漏洩率は図 2 と同様，b とトレードオフの関係にあり，b が十分に小さい場合は完全に漏洩してしまうことがわかる．また，特に LD による漏洩は人種によって結果に差があり，これは人種によって連鎖不平衡が異なっており，漏洩しやすさがそれに依存して変化するためと考察される．

センシティブ情報漏えいは具体的には，L_{TS} では表 1 のように，そばかすの疾患リスクから癌の因子 rs1805007-T を持つと推定されてしまうケースが存在した．また，L_{TLD} の例では網膜血管の口径と相関する SNP である rs10774625 が慢性腎臓病の因子である rs10774625 と強い連鎖不平衡にあり，s10774625 が完全漏洩するケースがあり，rs10774625 のリスクアレルを持つと疑われた．

また，リスクアレル推定法の計算効率を評価した．3.20GHz CPU，4GB RAM の Linux マシンにおけるアレル推定攻撃の 1 リスクアレルに対する計算時間の変化を図 5 に示す．実験は CEU データセットを用いて行った．結果より，式 4 をブルートフォースに評価した場合は計算時間が対象 SNP 数に対して幾何級数的に増加するが，azove2.0 を用いた場合には対象 SNP 数が増加しても計算時間は大きく変化せず効率的に計算が実行されていることがわかる．よって，実際に多量の SNP についてゲノム情報漏えいの監査を行う場合にも計算コストを多く要しないことが期待される．

8 終わりに

本稿ではゲノム検査結果の開示における被験者のリスクアレル情報の漏洩および，センシティブな疾患に関する因子の推測を監査する問題を提案した．さらに漏洩の定量評価方法を提案し実データを用いてある人種における平均的なプライバシ漏洩を評価した．

我々の提案した方法はゲノムプライバシ漏洩を整数線形計画問題として定式化し，従来のクエリ監査問題と比べ計算効率性の高く定量評価が可能な手法を導入したという新規性がある．また，我々の監査手法は推定モデルを用いないことでモデルによるバイアスのない評価ができているものと期待される．

また，本稿によってゲノム検査結果によるプライバシ漏洩は避けられないと定量的に評価されたことによって，情報開示におけるプライバシ漏洩度合いが明確になり，適切な情報利用に貢献するものと期待する．

今後の発展として，プライバシ漏洩評価法を

発展させ様々な情報開示問題に適用し情報漏洩を評価することが期待される．

謝辞

本研究は，科学研究費基盤研究 (B)「情報検索システムにおけるプライバシー保護に関する研究」および JST CREST 「ビッグデータ統合利用のための次世代基盤技術の創出・体系化」の助成を受けました．

参考文献

[1] Markus Behle and Friedrich Eisenbrand. 0/1 vertex and facet enumeration with bdds. In *Proceedings of the 9th Workshop on Algorithm Engineering and Experiments*, pages 158–165, 2007.

[2] 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. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[3] Hae Kyung Im, Eric R Gamazon, Dan L Nicolae, and Nancy J Cox. On sharing quantitative trait gwas results in an era of multiple-omics data and the limits of genomic privacy. *The American Journal of Human Genetics*, 90(4):591–598, 2012.

[4] Cathryn M Lewis. Genetic association studies: design, analysis and interpretation. *Briefings in bioinformatics*, 3(2):146–153, 2002.

[5] Shubha U Nabar, Krishnaram Kenthapadi, Nina Mishra, and Rajeev Motwani. A survey of query auditing techniques for data privacy. In *Privacy-Preserving Data Mining*, pages 415–431. Springer, 2008.

[6] Dale R Nyholt, Chang-En Yu, and Peter M Visscher. On jim watson's apoe status: genetic information is hard to hide. *European Journal of Human Genetics*, 17(2):147, 2009.

[7] Lukasz Olejnik, Kutrowska Agnieszka, and Claude Castelluccia. I'm 2.8% neanderthal-the beginning of genetic exhibitionism? In *Workshop on Genome Privacy*, 2014.

[8] Sahel Samani, Zhicong Huang, Erman Ayday, Mark Elliot, Jacques Fellay, Jean-Pierre Hubaux, and Zoltán Kutalik. Quantifying genomic privacy via inference attack with high-order snv correlations. In *2nd International Workshop on Genome Privacy and Security (in conjunction with IEEE S&P; 2015)*.

[9] Gudmundur A Thorisson, Albert V Smith, Lalitha Krishnan, and Lincoln D Stein. The international hapmap project web site. *Genome research*, 15(11):1592–1593, 2005.

[10] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 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*, pages 534–544. ACM, 2009.

[11] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.