平成 28 年度厚生労働科学特別研究事業　ゲノムデータの持つ個人識別性に関する研究

厚生労働行政推進調査事業費補助金 (厚生労働科学特別研究事業)
「ゲノムデータの持つ個人識別性に関する研究」

総括研究報告書

研究代表者
吉倉廣　国立感染症研究所 名誉所員
研究分担者
大澤資樹　東海大学医学部基盤診療学系 法医学教授
荻島創一　東北大学 東北メディカル・メガバンク機構准教授
鎌谷洋一郎　国立研究開発法人理化学研究所 統合生命医科学研究センター 統計解析研究チームリーダー
後澤乃扶子　国立研究開発法人国立がん研究センター 研究支援センター 研究管理部研究管理課長
佐藤智晶　　青山学院大学法学部　准教授
竹内史比古　国立国際医療研究センター・研究所室長
徳永勝士　東京大学大学院医学系研究科・人類遺伝学教授
俣野哲朗　国立感染症研究所 エイズ研究センター長

目的：ゲノムデータの個人識別性に該当する範囲の検討

緒言：

生命科学や情報通信技術など、近年の科学技術の進歩により、世界的に革新的な医療技術が相次いで開発され、我が国でも医療におけるイノベーションが期待されるようになっている。ゲノムに関しては 2003 年 4 月にヒトゲノム配列の解読が終了し、その後様々なゲノム解析技術やそれに伴うゲノム科学が急速かつ著しく進展し、研究、医療、個人認証といった産業、犯罪捜査等に応用が広がりつつある。一方で、ゲノムデータは体細胞変異やリンパ球における遺伝子組換え等を除きおよそ終生不変であり、また一卵性多胎児を除き唯一無二である。また血縁者間で共有されていること等より、個人情報としての適切な保護が必要であるが、どのようなゲノムデータならば個人識別性を帯びるのかについては議論が深まっていない。

なお、H27 年 9 月に改正された個人情報保護法において、「個人情報」は、①氏名、生年月日その他の記述等により特定の個人を識別することができるもの、②個人識別符号が含まれるもの、と定義され（第２条１項)、内閣官房情報通信技術（IT）総合戦略室は、「ゲノムデータは、社会通念上、個人識別符号に該当する」と整理した。また、ゲノム情報を用いた医療等の推進方策を検討するため、H27 年 11 月に設置された「ゲノム情報を用いた医療等の実用化推進タスクフォース」では、個人情報保護委員会に対し、「ゲノムデータの適正かつ効率的な活用」のため、その具体的な範囲について、科学的な観点、海外の動向を踏まえて、総

合的な解釈が示すことが求められるとの見解がとりまとめられたところである。

本研究は、ゲノムデータの持つ個人識別性に関する最新の研究・調査や海外の状況のレビューを行い、医療等の場での情報の取扱いに資することを目的としたものである。本研究班では、実際に具体的な各種技術や状況に応じて生成及び利活用されるゲノムデータについて、一意性の範囲設定の可能性、及び一意性範囲設定に必要な条件を中心に検討した。

注①：本報告書の言葉の定義は以下の通りである。
- **一意性**：他と重複することがないこと。会員サービス ID を例にとると、一つの対象に一つの識別子が付与されている場合、対象が異なっていても同一の識別子が存在する可能性がある場合と比較して、特定の個人を識別することができること。
- **ゲノムデータ**：塩基配列を文字列で表記したもの。
- **ゲノム情報**　：塩基配列に解釈を加え、意味を有するもの。

注②：「特定の個人を識別することができるもの」であるかの判断要素として、国会審議においては 、①個人と情報の結び付きの程度（一意性等）、②可変性の程度（情報が存在する期間や変更の容易さ等）、③本人到達性が示され、これらを総合判断するとされている。本研究ではこのうち特に「一意性」について検討したが、「ゲノムデータの持つ個人識別性」の議論においては、②可変性、③本人到達性についても併せて考慮する必要がある。

【総論】

ゲノム情報は一部の疾病から性格といった特性に関係し得ることから、プライバシーそのものといえる。あるゲノム情報を用いて不特定の個人の中から特定の個人を識別するためには、仮に全塩基配列情報があったとしても、データベースが存在し、突合できる状態になければ個人を識別することは困難である。一方、近年では第三者が特定の個人のプロファイリングを行うことが可能となっており、ある特定の個人に関する複数の情報をプロファイリングする事で、ゲノムデータを用いて個人を識別し得る。本研究班では、どの程度の情報量があれば、ゲノムデータから個人識別性が生じるかを検討した。

ゲノムデータには、人々に共通した部分と、個人により相違のある部分に分けることができる。個人を識別するためには、多型と呼ばれる個人間で相違のある部分を利用する。即ち、一塩基多型(SNP: single nucleotide polymorphism)と呼ばれる一塩基の相違であったり、同じ配列の繰り返し回数の相違だったりする。どこにどのような多型が存在するのかは、大部分特定されており、集団内における出現頻度も併せてデータベースとして登録されている。具体的な検出方法としては、塩基の並びである配列を決定してゆくシークエンス法、相違の部分だけを検出してゆくアレイ法、繰り返し回数を長さから判定するフラグメント解析法が主なものとなる。

【ゲノムデータの持つ個人識別性の計算法】

特定のゲノムデータが持つ個人識別性を計算するには、そのゲノムデータに含まれる配列が、集団内でどの程度の頻度で出現するものなのかを確認する必要がある。例えば、ある常染色体上の部位が、A か C のいずれかのアレル（対立遺伝子）をとる一塩基多型であったとする。日本人集団内における A アレルの出現頻度を f_A、C の頻度を f_C とする時、ある人が A

アレルを二つもつホモ接合の遺伝型(A,A)の時に、その出現頻度は f_A^2 となり、A アレルと C アレルを一つずつもつヘテロ接合の遺伝型(A,C)の時に、その出現頻度は $2f_Af_C$ と計算できる。他の部位にも相違するところがあり、それらが独立して遺伝していると考えられるならば、二つの遺伝型が同時に出現する頻度は、個々の遺伝型の出現頻度を掛け合わせたものに相当する。これを積の法則（product rule）と呼ぶ。そのゲノムデータにｎ個の相違部位があるとするならば、各部位における遺伝型の出現頻度をｎ個すべて掛け合わせた数値が総合頻度（P）になる。

【数値の評価】

ゲノムデータから得られた総合頻度（P）は、集団内でどの程度の確率で出現する可能性があるかを示している。例えば、1.0×10^{-3} という数値が得られたとすると、そのゲノム情報は 1000 人に一人の割合で検出される可能性があると言い換えることもできる。この数値が 1.0×10^{-10} 以下となった時に、0.99（99%）の信頼度で個人が特定できたと解釈されることとなる（Budowle, B., Chakraborty, R., Carmody, G., and Monson, K.L., (2000) Source Attribution of a Forensic DNA Profile. Forensic Science Communications, 2 (3). Available online at:https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2000/index.htm/source.htm)。このレベルに達する多型の個数は、SNP で 40～50 座位程度、4 塩基単位の繰り返し構造（STR：short tandem repeat）で 9～10 座位程度である（参照；個人識別性について：法科学からの視点　分担研究者：大澤)。すなわち、ゲノムデータにこの程度ないしこれ以上の多型部位が含まれた場合には、個人識別性があると想定し得る。

（参考）司法における個人特定では、CODIS（The Combined DNA Index System）と呼ばれる米国 Federal Bureau of Investigation（FBI）が定めた常染色体上の 4 塩基からなる反復配列（STR）13 座位の解析が中心となっている。

【注意点】

１．一卵性双生児の場合には、同じゲノムを共有するので、一般的な手法で遺伝学的に区別することは難しい。

２．親子、同胞といった近縁者を特定してしまう場合が想定される。その意味において、この基準はあくまで目安であり、1.0×10^{-10} 以下の数値が得られなかったとして、個人識別性が無いとは決して言えない。

【個人識別性を考慮したゲノムデータの扱い】

研究や医療の場でゲノムデータを扱う際の指標として、以下のように分類し具体的に示した。

① **「個人識別性がほぼ確かと判断できる」レベル**
全核ゲノムシークエンスデータ、全エクソームシークエンスデータ、全ゲノム SNP データ、※互いに独立な 40 以上の SNP から構成されるシークエンスデータ、STR９~10 座位以上

② **グレイゾーン**
いずれにも該当せず、個別に専門家の判断を要するもの

③ **「個人識別性はほぼ無いと判断できる」レベル**
互いに独立な 30 未満の SNP から構成されるシークエンスデータ、がん細胞等の体細胞変異、単一遺伝子疾患の原因遺伝子の（生殖細胞系列の）ホットスポット変異

※SNP は、30 未満では確実な個人識別性に至らないが、40 を超えるとほぼ確実に個人識別性が生じると、現時点では考えられる（参照；個人識別性について　分担研究者：鎌谷、個人識別性について：法科学からの視点　分担研究者：大澤）。

（注意点①：レアバリアントの取扱について）

まれな変異（レアバリアント）は、そのレアアレルを持っていない大多数の人にとって個人識別性は無いと考えることができる。レアバリアントは遺伝的浮動による影響を強く受け、アレル頻度が変化しやすく、またシークエンスエラーの影響を非常に受けやすい。特定のレアアレル保有者においては、個人識別性が高いと言えるが、このようなまれなアレルを保有することの個人識別性を認めた場合、現時点のデータベースにおいて変異情報の無いゲノム上のどの１塩基をとっても、いつかどこかの誰かにとっては個人識別が可能になり得ると言える（生殖細胞突然変異はゲノム上のどの部位にも起き得る）。以上から、レアバリアントは上記分類に含めず、別途取り扱うべきものと整理した。また、レアバリアントの中で、臨床的意義が明らかな希少性の高い難病等の原因変異については、他の情報との突合により容易に個人識別が可能なものとして、データの取扱には十分注意する必要がある。（参照；個人識別性について　分担研究者：鎌谷）

(注意点②：ホットスポット変異について)

ゲノム DNA 中の多様性（変異）の分布は一様ではなく、多様体の頻度が特に高く（100 倍等）なる部分を、遺伝学ではホットスポットと言うが、ここでは医療上の意義に注目し、単一遺伝子病や薬物応答異常等の原因変異であって、独立した複数の発端者（家系を代表する罹患者）に繰り返し同定される変異を指す。ホットスポットの概念に含まれる重要な点は、それが疾患等の発生機構上、生物学的な蓋然性を持つため、新規症例でも出現し得るという点である。そのため、過去に観測されている頻度にかかわらず、任意の新規症例に出現し得るため、一意性が失われている。ホットスポット変異は当該遺伝性疾患や薬物応答異常等の診断や治療の標的として重要であり、実際にいくつかのホットスポット変異は既に診断あるいは研究的診断に広く用いられている市販の多遺伝子パネルにも搭載されて活用されている。ただし、ホットスポットのリストも研究の進捗により随時変化していく。

がん細胞等の体細胞変異については、研究や診断・治療の主たる対象は、生物学的意義を持つ体細胞変異のホットスポット（ドライバー）変異となっているため、その情報には一意性がない。さらに、がんの体細胞変異はゲノム不安定性を背景に、がんの発生・進展・治療に伴い変化し、あるいは消失するため、個人を識別する符号としては不変性が保たれていない。(参照；がん研究におけるゲノムデータの個人識別性について　分担研究者：後澤)。

【海外の見解】

欧米において、遺伝情報がプライバシーに相当するとして、法制化が進んできている。法律の規定する内容としては、検査を受けることや遺伝情報を保存・開示することへの本人の同意の必要性、情報へのアクセス、検体および遺伝情報の個人としての所有権、違反した時の罰則等多岐にわたる。保険や雇用に利用された時には、社会的に大きな影響が及びうるとして、ゲノムデータのもつ個人識別性について、活発な議論が展開されている。ゲノムデータには個人識別性があるという前提のもとで、生体材料(ゲノムデータを含む)取得時に本人の同意を得た上で、十分な匿名化を施せば原則自由に利用ができる、というのが欧米の見解である(ただし、本人と同意した内容については、関連規制法令の問題とは別に、契約上の義務が別途生じる余地があるため、同意時の内容や合意内容には注意が必要である)（Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 0829/14/EN, WP216, 10 April, 2014, at 9; National Human Genome Research Institute in NIH, Privacy in Genomics, April 21, 2015）。

個人識別可能な医療情報（ゲノムデータを含む）の匿名化の方法として、HIPAA プライバシー規則（米国）では二つの方法が明記されている（45CFR§164.514(b)(1) and (2); U.S. Department of Health & Human Services, Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, available at http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html）。

一つ目は、情報の受領者が個人識別するリスクについて最小化されていることを**専門家が確認する**方法である。二つ目は、**18 種類または 16 種類からなる所定の情報**（18 種類の場合は 1. 名前、2. 州以下の住所、3. 誕生日等の年月日、4. 電話番号、5. FAX 番号、6.E メールアドレス、7. 社会保障番号（SSN）、8. 診療録番号、9. 医療保険の受益者番号、10. 銀行口座の番号、11. 資格等の番号、12. 自動車登録等の番号、13. 医療機器番号、14. ウェブの URL、15. IP アドレス、16. 指紋や声紋等の生体認証記録、17. 顔面写真等のイメージ、18. その他の個人識別コード; 16 種類の場合は 2, 3 は除去しなくてもよい）を予め除去し、残りの情報が個人識別に使用されないことを確認する方法である。

なお、**一塩基多型が 30 未満であれば個人識別性がない**ゲノムデータに該当しうるとの見解がある。

（Lin et al. Genomic Research and Human Subject Privacy, *Science* 2004;305(5681):183.）

【留意事項】

今回ゲノムデータを便宜上「個人識別性がほぼ確かと判断できるレベル」「グレイゾーン」「個人識別性がほぼ無いと判断できるレベル」の３つに分類したが、今後の技術の発展等に伴い個々のゲノムデータのもつ個人識別性は常に変化していく事に留意する必要がある。

ゲノムデータは、それ自体は「個人情報」のカテゴリーであったとしても、直接、或いは HIPAA **プライバシー規則にリストされている** 18 **種類の所定の情報などを通して、直接間接生身の人間に繋がらない限り、「誰のゲノムデータ」かは分からない。従って「個人の特定」には至らない。「一意性」は、データの所有者と思われる生身の人間が出現した処で始めて具体的な意味が出てくる。**