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」

平成 28 年度

総括・分担研究報告書

平成 29 年（2017）4 月

研究代表者　　吉倉　廣

目次

総括研究報告書

分担研究報告書

【報告書の構成】

総括研究報告書は、原案を分担研究者全員が検討し修文の上合意されたものである。ヒトゲノムに関する、人類遺伝学的、法医学的、疫学的、がんその他疾病の予防治療的側面につき、多様な現場での対応を反映させようとしたため、班員全ての意見の細部を総括研究報告書に盛り込む事は難しかった。

なお、最後の補足説明は、個人情報保護委員会事務局から解析方法・データ保存・個人識別の基準各項目につき質問があったので、吉倉・竹内がこれに応じ、作成したものである。

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。**

個人識別性 について：法科学からの視点

分担研究者　大澤 資樹　　東海大学医学部基盤診療学系法医学　教授

ゲノム情報から個人を特定する時には、一人分の配列情報があったとしても、個人を特定することは困難である。他の情報と照らし合わせて始めて、個人が特定できる。対照する方法としては、配列の一致を調べるマッチング（異同識別）が単純で一般的な方法となる。それに加えて、ゲノム情報の特徴として血縁関係を調べることによっても個人が特定できる。このアプローチはファミリアルサーチ（血縁関係推定）と呼ぶ。

マッチングに関しては、非血縁者を仮定する限り、連鎖不平衡や連鎖は必ずしも関係なく、一致確率は集団内の対立遺伝子頻度のみに依存する。どの程度の一致確率が求められるかといえば、数千万分の 1 でも誤認逮捕された英国の事件が有名である。例えば、ある人物の ABO 式血液型が A 型で、別のゲノム情報から A 型遺伝子が得られ、一致したとする。しかし、実際には国内には 5 千万人も A 型の人がおり、A 型が一致したからといって、個人が特定できたかというと、とても言えない。対象とする集団の大きさが重要で、ある程度の数の SNP や繰り返し配列が一致して始めて、特定の個人に結びつく可能性がでてくることになる。

ファミリアルサーチから個人を特定した例としては、厚生労働省が関係する中国残留孤児の身元調査や戦没者慰霊事業が挙げられる。DNA を抽出できるような本人の遺品は残っていないので、もっぱら親族との血縁関係を推定することで個人を特定してきた。逆の言い方をすれば、ゲノム情報から血縁者かどうか分かるリスクがあるということになる。

法科学領域では、この個人識別を STR（short tandem repeat）と呼ばれる繰り返し配列の解析を中心に行ってきた経緯がある。特に、米国連邦捜査局（Federal Bureau of Investigation）は、CODIS（combined DNA index system）と呼ばれる 4 塩基の繰り返し単位からなる 13 STR 座位を犯罪捜査等の個人識別目的で検査する座位と定め、現在では各国で採用され指紋のようにデータベースとして集積されている。市販キットは、CODIS 座位と性別判定用座位を含めて、一度に 16～23 座位を 1 本のチューブ内で増幅し解析するシステムが採用されており、multiplex 系と呼ばれる。

【個人識別性を有する総合一致確率について】

ある集団内で個人識別性を有する可能性のある遺伝型の一致確率は以下の式で表すことができる。

$(1 - p_x)^N \geq 1 - \alpha$

p_x：ある座位 X でのプロファイルが偶然に一致する確率（random match probability）

$1 - \alpha$：信頼水準（confidence level）

N:母集団数（sample size）

(参考文献：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)

集団サイズ（N）を1億人とした時に、この式から要求される一致確率は、次のように計算できる。

	1 － α：信頼水準（confidence level）	
集団サイズ（N）	0.95	0.99
1億人（10^9）	5.13×10^{-10}	1.01×10^{-10}

（参考文献：“Forensic DNA typing” 2nd Ed., John M. Butler, Elsevier, pp 513-515, 2005）

【実際に要求される解析座位数について】

仮に、集団を1億人と仮定した時に、信頼水準0.99で個人を識別するためには、1.0×10^{-10}程度の総合一致確率（P）が求められる。そして、互いに独立した何個の座位における遺伝型の出現頻度（p_x値）を掛け合わせると、このレベルに達するかを考えてゆくことになる。式で表すと以下のようなものとなり、nがどの程度の数値となるかSNPとSTRで検討してみた。

$P = p_1 \times p_2 \times p_3 \times \cdots \times p_n \leq 1.0 \times 10^{-10}$

任意の二検体間における座位Xにおける遺伝型が一致（IBS2）する確率は、

$p_x = \Sigma a_i^4 + \Sigma 4a_i^2a_j^2$ （$i \neq j$, a：対立遺伝子（アリル）頻度)

で計算される。

個人識別に利用される4塩基繰り返し構造のSTR（マイクロサテライト）は、p_x値は0.07～0.10をとる。仮に、一致する確率が0.08ばかりの場合を考えた時に、座位数が増えた時の複合一致率は、以下のようになる。

座位数（n）	複合一致率（P）
1	8×10^{-2}
2	6.4×10^{-3}
…	…
8	1.7×10^{-9}
9	1.3×10^{-10}
10	1.1×10^{-11}
11	8.6×10^{-13}

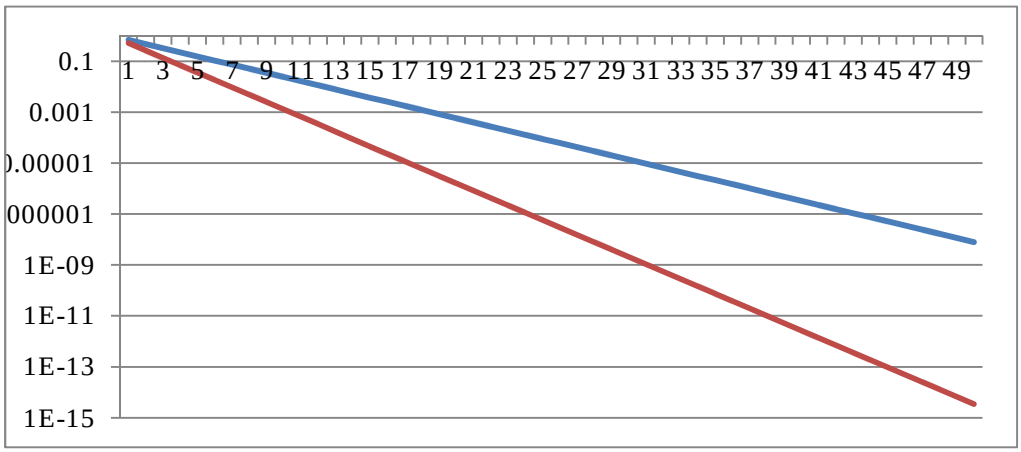
1.0×10^{-10}以下になるのは、9ないし10座位を検出した時と判断した。

次に、SNPにおいては、マイナーアリル頻度を以下のように仮定した場合に、二検体間で座位Xにおいて遺伝型が一致する確率は次の表となる。

minor allele	単独一致率（p_x）
0.05	0.8235375
0.1	0.6886
0.2	0.5136
0.3	0.4246

0.4	0.3856
0.5	0.375

SNP のアリル頻度の分布データを持ち合わせていないので、マイナーアリル頻度が 0.1（青線）と 0.2（茶線）のみの場合の複合一致確率を示したのが、下のグラフになる。



（本来ならば、もっと正確な試算をしなければならないが、今回の期間内ではこの程度までである。）

そこで、個人を 1 億人の中から特定するために求められる解析座位数を概算すると以下の表のようになる。

	p_x 値	必要な個数
SNP	0.38 ～ 0.85	40～50
STR	0.07 ～ 0.10	9～10

日本人集団を想定した時に、個人識別性を有するゲノム情報とは、SNP ではおおむね 40 座位以上のデータを、STR では 9～10 座位以上のデータに含むものに相当すると考えた。ただし、SNP に関しては、座位ごとのアリル頻度により、求められる数はかなり変わってくることはやむをえない。ただ、一般的に、STR の一致率は、SNP の 4 ないし 5 座位分の一致率になるとされており、この観点からも大きな間違いはないと考えている。

ゲノムデータの個人識別符号の範囲と本研究班における検討範囲であるところのゲノムデータの一意性についての報告書

分担研究者　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）

ゲノムデータの個人識別符号の範囲と本報告書における検討範囲について

- ゲノムデータは、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において、社会通念上、個人識別符号に該当するとされ、その範囲は、個人情報保護委員会により、**科学的観点からの一意性**、**可変性**、**及び本人到達性**に基づき、**海外の動向**を踏まえて、総合的な解釈が示されるとされた。
- 本報告書では、このうち**ゲノムデータの一意性を検討範囲**とする。

ゲノムデータの多型・変異の一意性について

- ゲノムデータの一意性のなかでも、本報告書では、**多型・変異の一意性**について論じる。その一意性はその**位置により異なる**。
- 全ゲノムデータ、全エキソームデータは一意性があると考えられる。
- 統計的に独立な30～80座位のSNPのゲノムデータが一意性をもつことがあるという試算の報告があるなど(Lin et al. Genomic Research and Human Subject Privacy, *Science* 2004;**305**(5681):183)、**およそ30～80座位以上の多型・変異から構成されるゲノムデータについては一意性がありうる**と考えられる。
- 30～80座位以下の多型・多様体の一意性の程度は、ゲノムデータが由来する人口集団として、日本人の集団の大規模なゲノム解析により解明されたアリル頻度に基づいて一定程度評価可能と考えられる[参考]。一変異の一意性については評価が必要である。**まれな変異で、希少・難治性疾患の原因である変異は、病歴に関わるゲノム情報として、一定の配慮をもって取り扱われるべき**である。

参考　ファーマコゲノミクスの保険収載、先進医療での遺伝子検査で対象となる多型のゲノムデータを例に、日本人の集団の大規模なゲノム解析により解明されたアリル頻度を示す。

検査区分	検査名称	多型	rs ID	アリル頻度		
				GMAF	GO-ESP	1KJPN
保険収載	UGT1A1遺伝子多型検査	UGT1A1*28	rs34983651	0.3253 (AT)	NA	NA
保険収載	UGT1A1遺伝子多型検査	UGT1A1*6	rs4148323	0.0344 (A)	0.00131 (A)	0.1827 (A)
保険収載	UGT1A1遺伝子多型検査	UGT1A1*27	rs35350960	0.00280 (A)	NA	NA
先進医療	CYP2D6遺伝子多型検査	CYP2D6*4	rs3892097	0.09310 (T)	0.15105 (T)	NA

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」

分担研究報告書

個人識別性について

分担研究者
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　チームリーダー

本稿は名古屋市立大学青木康博先生「DNA 鑑定による法医学的個人識別の確率･統計学的背景」http://www.med.nagoya-cu.ac.jp/legal.dir/aoki/docs/review.pdf を参考にしました。

本報告書は特に、研究班において議論となった「100 万人あたり 1 人〜5 人しかマイナーアレルを持たないような単一のレアバリアントは個人識別符号であるか」という命題に対し、「個人識別符号ではない」と主張する立場から作成しています。

1　用語の整理

表 1: 用語について

用語	意味
ゲノム	遺伝情報の総体を指す。各個人のレベルでは、約 30 億塩基からなる DNA 塩基配列のこと
遺伝的変異（バリアント）	ゲノム上の特定の位置において、人々の間で塩基配列が異なるような場所を指す。
アレル（アリル、アリール）	遺伝的変異において、個人が二本の染色体上にそれぞれ持つ塩基配列を指す。
遺伝型（ジェノタイプ）	遺伝的変異において、個人が二本の染色体上に持つアレルの組み合わせを指す。

（日本人類遺伝学会 2009 年用語改定に従う）

2　一つの遺伝的変異による識別能について

ある遺伝的変異において、アレルが m 個あり、$A_1, ..., A_m$ であるとし、i 番目のアレル A_i の頻度がそれぞれ f_i であるとします。

変異が Hardy-Weinberg's principle に従っているものとすると、任意の i と j について、ホモ接合型 A_i/A_i の遺伝型頻度は p_i^2、ヘテロ接合型 A_i/A_j の遺伝型頻度は $2p_ip_j$ です。

すると、任意に選んできた血縁関係のない二人が同じ遺伝型を持つ確率は、ホモ接合体について p_i^4、ヘテロ接合型について $4(p_ip_j)^2$ となり、合計すると

$$\sum_{i=1}^{m} p_i^4 + \sum_{i,j,i<j}^{m} 4(p_ip_j)^2 \tag{1}$$

が、任意に選んできた二人が、この遺伝的変異において（どれでもいいから）遺伝型が一致する確率となります。

従って「一致しない」確率は

$$1 - \sum_{i=1}^{m} p_i^4 - \sum_{i,j,i<j}^{m} 4(p_ip_j)^2 \tag{2}$$

これを識別能（power of discrimination, PD）とも言うそうです。

2.1　biallelic SNV の場合

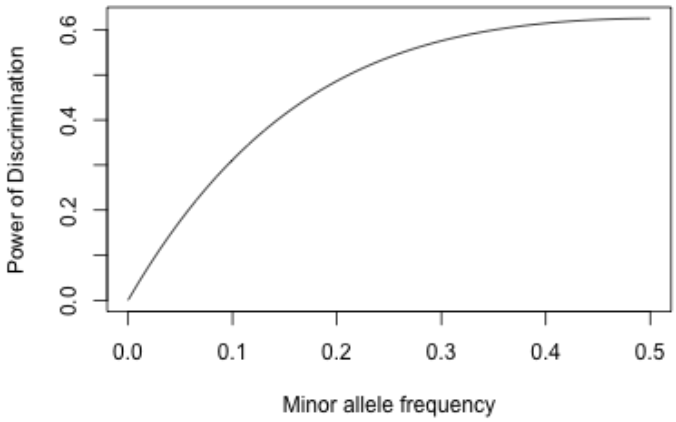
アレルが二種類（biallelic）である一塩基変異（Single Nucleotide Variation; SNV）について検討します[1]。

片方のアレルの頻度が f、二つしかないのでもう片方は $1-f$ なので、識別能は

$$1 - f^4 - (1-f)^4 - 4f^2(1-f)^2 = 2f(1-f)(3f^2 - 3f + 2) \tag{3}$$

これをプロットすると以下のようになります。

[1]任意に選んできた遺伝的配列にほぼ確実に入っている変異は SNV であると言えます。ほとんどの SNV は biallelic です。また、三種類以上（multiallelic）のアレルの場合、SNP array では実験できませんし、次世代シークエンサー（NGS）の場合は原理的にエラー率が高くなることが知られているため、個人識別能はやや低下するものと考えられます。ここではベストエフォートの識別能を検討しているので、マルチアレル変異は検討しないことにします



アレル頻度 50%程度で、識別能が最大になることがわかります。

具体的にどういう数値になるかというと、例えば任意に選んできた二人が、二回に一回は一致しないようになるのはアレル頻度 25%くらいのようです（正確にはこの時 53.9%の確率で一致しない）。逆に見れば、二回に一回は、全く別々の人の遺伝的配列が偶然全く同じになるということであり、これはもちろん「個人識別符号」とは言えません。

STR の場合、式 (2) において（アレル数が多くなるので）それぞれの p が小さくなり、それが乗数効果で大きくなるため、一般に SNV よりも識別能が高いことになります。

ターゲットシークエンスのような状況では、その領域を二つのハプロタイプの組み合わせと見立てることでやはり式 (2) を適用できます。STR と同様にアレル数が多いのと同等の状況になるため、やはり一領域のターゲットシークエンスは、SNV よりも識別能が高いでしょう。

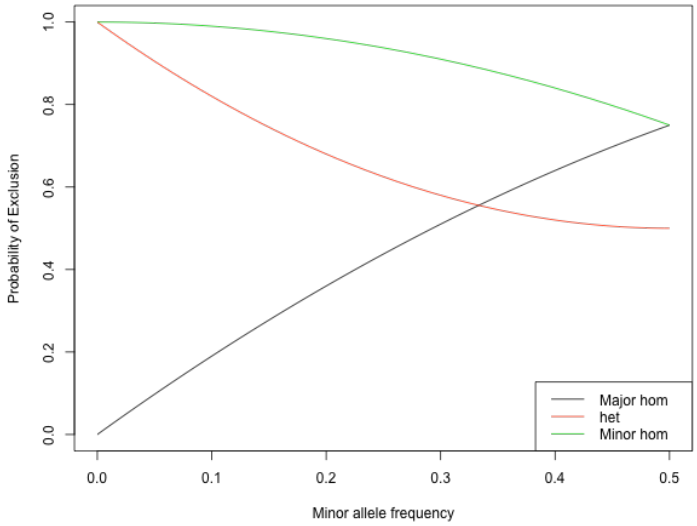
2.2 レアバリアントについての検討

ここまで述べたように、アレル頻度は高いほど「**遺伝的変異の**識別能が高い」ということです。

ところで、上の図をもとにすると、レアバリアントは識別能が低い、ということになるでしょうか。実際にはレアバリアントのアレルを示すデータがあれば、それを保有していると思われる人は少数に限られるはずです（と言っても、現時点で高い精度で頻度の同定が可能な限界と思われる 0.5%のアレルの場合、キャリアの人数は 1 億人中 100 万人程度に絞られるにすぎませんが）。これは、式 (2) は「ある任意の一人」について遺伝的変異の個人識別性

を論じていることによるもので、このレアバリアントの場合、そのレアアレルを持っていない 99%程度の人にとっては全く個人識別性がないからです。

「ある任意の一人」ではなくて、**レアアレルを持つ特定の人**にとっての識別能を検討できます。先ほどのレアバリアントについて、アレル頻度を f とした時、ヘテロ接合体である人について、もう一人を任意に選んだ時に一致する確率は $2f(1-f)$、ホモ接合体である場合は f^2 です。一致しない確率は、それぞれを 1 から除したものです。したがって



この図の左側に行くほどレアバリアントであり、上に行くほど識別能が高いことを意味します。ここに示すように、レアバリアントは、その「遺伝的変異としての識別能」は低いものの、特定の**レアアレル保有者において**、確かに個人識別能が高いことが（当然ですが）わかります。一方、当該レアバリアント部位におけるレアアレル非保有者においては、個人識別能は高くありません。

考え方を整理します。この報告書では、最後に、頻度の高い SNP を基本とした識別能を考慮します。このような**複数の頻度の高い** *SNP* を基本とした考え方の場合、例えば 100 個程度のある程度独立で頻度の高い SNP を用いますと、1 億人いたら 1 億人それぞれを識別できると考えられます。一方、**単一のレアアレル保有者**として考えた時に見られる個人識別能は、そのレアアレルを持つ 1 億人中 100〜200 万人未満[2]に対してのみ有効ということになるでしょう。

レアアレル保持者において、レアバリアントは個人識別性が高いわけですが、それはどの程度の個人識別能を持つかを考えてみます。今述べた「1 億人

[2]マイナーアレル頻度 f である時、そのアレルを持つ人の割合は、Hardy-Weinberg's principle が成立しているとした場合、$2p(1-p)+p^2=2p$ となります

中 100〜200 万人未満に対してのみ有効」というのは逆に見れば 100〜200 万人が同じレアアレルを持つという状況も含んでいます。1 億人の中で 100 万人も同じ情報を持っているようなものは、個人識別符号とは言わないでしょう。ではどれくらいのレアアレルなら個人識別符号と言えるかを次節で考えてみます。ここまで述べた通り、個人識別性はアレル頻度によって規定されます。しかし、次に述べるように、レアバリアントのアレル頻度は正確に定量しづらい（後述の遺伝的浮動を考えれば、「一般的なアレル頻度」を与えることは原理的に不可能）のです。そのため、単一のレアバリアントについて個人識別能を論じることは困難であることを論じます。

2.3 レアバリアントの個人識別性を検討すべきか

レアバリアントは、頻度の高い遺伝的変異とは別の性質で個人識別性を持つことがわかりました。では、このように稀なアレルを保有することの個人識別性を認め、そのため**レアバリアント部位**を含む遺伝的情報は個人識別符号であるとする考え方を取るなら、どのようなゲノムデータが個人識別性を持ちうるかについて考えてみます。

生殖細胞系列突然変異はゲノム上のどの部位にも起きうると考えられています。したがって、30 億塩基のゲノム上のどの部位においても、本日、今の段階で生まれた子が、そこに発生した変異アレルを持っている可能性があります。すなわち、**現時点のデータベースにおいて変異情報のないゲノム上のどの *1* 塩基を取っても、いつかどこかの誰かにとっては個人識別が可能なレアバリアントになりうるため、その全てが個人識別符号である**と言えます。地球上全人類の正確な全ゲノムシークエンスデータを得ていない[3]以上、どの場所をとってもそこがそれではないと主張することはできません。

これは非常に非効率的な考え方であると私は主張します。なぜなら、「ゲノム上の**部位**」といった場合、誰もが持っている情報ということになりますが、そのそれぞれの部位における遺伝的配列情報は、99%以上の人にとっては、個人識別能がないからです。

そうしますと、レアバリアントの個人識別性を議論する場合、重要なのは**場所**（=レアバリアントであるかどうか）ではありません。例えば a 番染色体の部位 b がレアバリアント部位だったとしても、その部位そのものはほとんどの人にとって個人識別符号ではありません。ある特定の個人の遺伝的情報にレアアレルがあった場合、その部位 b のゲノム情報の中でとりわけ**その人のゲノム情報のみ**が個人識別符号であると捉えられます。これは、ありふれた SNP を 100 個という場合のデータ（誰にとっても個人識別符号である）とは違った状況であることは前節で述べています。

[3] しかも日々生まれる全ての子のシークエンスをその時点で得る必要があります

では、どれくらいのレアアレルから、個人識別符号であると言えるようになるでしょうか？前節で述べたように、レアアレル個人識別能もまたアレル頻度によって規定されます。

ところが、レアバリアントのアレル頻度は、次のような理由によりとても不安定であり、確定できません。

- レアバリアントのアレル頻度は、遺伝的浮動の影響により**時間的に不安定**です（次節参照）。おそらく 10 年くらい経てば、データを更新する必要が出てきます。
- レアバリアントのアレル頻度は、**空間的にも不安定**です[4]。日本国内でも地域によって差があると考えられます。
- レアバリアントのアレル頻度は、**シークエンスエラーの影響を受けやすい**です。10000 人あたり 2000 個のマイナーアレルについて 10 個のエラーがあっても頻度は 10%から 10.05%に変わるだけですが、レアバリアントの 20 個のマイナーアレルについて 10 個のエラーがあると頻度が 0.1%から 0.15%に変化し、その変化の割合は 50%にもなります[5]。さらにレアになればより影響は大きくなります。

したがって、個人識別能を規定するために絶対に必要であるアレル頻度情報を、レアバリアントにおいては正確に[6]とることができません。

ですので、レアバリアントにおけるレアアレルの個人識別能については正確な計算が不可能です。これは、裁判の状況を考えれば、その問題点が明らかと思います。裁判において検察側が、レアアレル頻度をもとに個人識別能があるとするレアアレルを事件現場に残された証拠と被疑者が共有しているので犯人であると主張したと考えます。しかしレアアレル頻度自体が不確定ですので、これがその個人を特定するものであるか、他にもある程度の人数で存在しうるのか、科学的に求められないのです。偶然一致した確率を計算できないのです。このようなものは個人識別符号とは言えないことが明らかだと考えます。

ありふれた遺伝的変異のアレル頻度は、上の三つのすべての場合についてレアバリアントよりはるかに頑健であり、現状用意されているアレル頻度を個人識別能の検討のために利用することができると考えられます。

---

[4] Mathieson and McVean. Nat Genet 2012;44:243.

[5] この時、「マイナーアレル」にだけエラーが起こると仮定しているのは、アンフェアではありません。現在のリシークエンス技術は、参照配列にある参照アレル「であるか、否か」を見るので、必然的に非参照アレル＝ほとんどの場合レアバリアントのマイナーアレル、の方がエラー率が高いと考えるのは必然です

[6] ここでいう「正確」とは、個人識別能を検討する際の対象集団についての全体からの頻度を意味します。一般的には、個人識別性を検討する際にはこれは全地球人口と考えられると思います

2.3.1 レアバリアント頻度と遺伝的浮動の関係について

前節にて、遺伝的浮動の影響によりレアバリアント頻度は変化しやすいので、特定の変異について個人識別能を推定できないと記載しましたが、それについて説明を加えます。

Wright-Fisher モデルを考慮すると、遺伝的浮動に基づき、ある人数 N の世代で頻度 p_0 であった遺伝的変異の次世代における頻度 p_1 は、二項分布に基づき次の期待値と分散をとります。

$$\begin{aligned} E[p_1] &= p_0 \\ Var[p_1] &= p_0(1-p_0)/2Ne \end{aligned} \tag{4}$$

ここで Ne は集団の有効な大きさ（effective population size)。

HapMap2 の推定に基づき、日本人集団の Ne を 14269 と仮定します。

すると、それぞれの p_0 の値により、p_1 の分散が求まるため、それを元にして 95%の確率で p_1 の値の取り幅を推定できます。

表 2: アレル頻度ごとの遺伝的浮動により予想されるばらつき

p_0	次世代における頻度の 95%幅	元の頻度に対する増減の割合
50%	49.4 - 50.6 %	1.16%
5%	4.74 - 5.25%	5.06%
1%	0.884 - 1.12%	11.5%
0.1%	0.0633 - 0.137%	36.7%
0.01%	0 - 0.0216%	116%
0.0005%	0 - 0.00309%	519%
0.0001%	0 - 0.00126%	1160%

したがって、個人識別能を規定するために必要である「頻度情報」は、レアバリアントにおいては極めて不安定であることがわかります。これは、有限サイズの集団（つまり現実世界）において本質的な特徴です。

すなわち、あるデータセットを用いてレアバリアントの頻度を求めたとします。それを、個人識別能を算定する際に利用しようというわけですが、年度が違えばアレル頻度が徐々に変化していくため、すぐに利用できなくなると考えられます。

一方、ありふれた遺伝的変異の場合、表 2 に示されるようにアレル頻度は安定しており、個人識別能の推定もまた安定します。

3　複数の遺伝的変異による個人識別

実際に個人識別符号であるかどうかをどのように評価するかについて最後に述べます。

複数の遺伝的変異を利用し、各遺伝的変異が互いに独立であると考えるなら、任意に選んだ二人の遺伝型が一致しない確率は

$$1 - \prod \left[\sum_{i=1}^{m} p_i^4 - \sum_{i,j,i<j}^{m} 4(p_i p_j)^2 \right] \tag{5}$$

となるはずです。しかし、22 個（各染色体に 1 個）を超えた時点で、連鎖不平衡の性質により完全な独立というのは検討できません（もちろん、短腕の端と長腕の端は現実的にはほぼ独立ですが）。biallelic で頻度 50%の SNV について 1 個あたり、式 (3) により一致する確率が 0.375 であることから、22 個による一致しない確率は、$1-(0.375)^{22} = .9999999995747073$ ですので 10^9 レベル、すなわち 10 億人レベルまで識別可能と思われます。日本人なら充分でしょう。30 個持ちいると 10^{12} レベルになりますので、1 兆人レベルなので世界人口に対応できます[7]。

ところが、バイオバンクジャパン 20 万人について 22〜30 個の頻度の高い独立な SNV の組み合わせについて検討したところ、完全な識別はできませんでした[8]。つまり、これらすべての遺伝的変異について、全く同じ遺伝型を持っており、互いに区別できない人がいました。念のため申し上げると、対象としたデータは遺伝的同祖解析により 2 親等程度の近縁関係はすでに除外されています。たかだか 20 万人でも区別できない、これはなぜでしょうか。

1. 調べている個人間において、強くはないが弱い血縁関係が存在し、遺伝型が完全に独立ではないからです。
2. 遺伝的変異間が完全に独立ではないからです。
3. ジェノタイピングエラーが起こりうるからです。

おそらく 1 の問題が大きいです。これが有限サンプルにおける個人識別能推定の問題です。バイオバンクジャパンは全国からの収集でありしかも病院ベースの患者さんバンクなので、1 の問題は地域コホートよりかなり小さいはずですが、それでもこの問題が関係します。

そこで理化学研究所統合生命医科学研究センターでは大きめのマージンを取って、100 SNPs あれば十分個人識別可能としています。ただ、上の方法で

[7]遺伝的浮動、シークエンスエラーや地域差などにより、数 10%のアレル頻度は数%以下の範囲でばらつきますが、それはほとんど影響を与えなさそうです。人種間で大きな差がある SNP、例えばアルコール不耐性変異 rs671 などは存在し、注意を払う必要はあります。ただし、それがどれであるかは十分特定可能です

[8]ランダムに選択した 22〜30 個の頻度の高い独立な SNP の組み合わせが、本当に 20 万人の全てを識別するかを 100 回試したところ、区別できない組み合わせが存在しました

試みた場合、35〜40 SNPs では 100 回の全てで完全に個人識別ができていました。個人識別能は乗数的に増えていくため、40〜50 くらいの頻度が高く互いに独立な SNP があれば、個人識別が可能であろうと思われます。

4 具体的な状況

最後に、ここまでの検討をもとに、私の意見として明らかに個人識別符号であるとするデータを具体的に述べます。

1. 少なくとも 40 個の互いにある程度独立な SNP を含むデータは個人識別符号です。これは、**全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**においては間違いなく成立すると考えられます。

2. 1. に示すよりも小さなデータ（例えば複数以上の領域を含む**ターゲットシークエンスデータ**[9]や**カスタムアレイ *SNP* データ**[10]）については、その中に含む SNP の数と独立性によって個々に決定されると思われます。

3. レアバリアントを含む場合は、一般に全体ではなくその中のレアアレル保有者において、個人識別能は高まります。これは SNP データでなく**シークエンスデータ**において起こりやすい状況であると考えられます。

4. ***30* 個未満の *SNP* データ**であれば、個人識別は不可能であると考えることができます。また、1 領域内数遺伝子以内のターゲットシークエンスも個人識別能ではないと考えられます。

5. *STR* については、別の分担研究者が論じています。

[9] 遺伝子パネルシークエンスを含みます
[10] 具体的には、ImmunoChip、MetaboChip、ExomeChip のようなもの

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」

分担研究報告書

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　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

ゲノム科学の急速な進展により、がん領域においては、遺伝性腫瘍における生殖細胞系列由来の DNA に存在する多型情報･変異情報に加えて、がん細胞に生じた体細胞変異などの後天的に生じるゲノム変化等に着目した研究、さらには医療実装に向けた様々な取り組みが進められている。

がん研究等で用いるゲノムデータの特徴を、①個人と情報との結びつきの程度（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本人到達性の観点から整理する[1-3]。体細胞あるいは生殖細胞系列のホットスポット変異（染色体の特定の箇所で高い頻度で見られる変異）は、複数の患者のがん細胞あるいは非がん細胞が有している変異であり、腫瘍生物学的な説明が付いている変異もあることから、当該がん種の新規症例においても一定の確率で繰り返し出現することが知られている。従って、ホットスポット変異のみ（ゲノム塩基配列上の位置と、その座位での塩基配列の変化の内容）を記述し、その周辺の塩基配列を含まないゲノムデータは、それ単独では①一意性、③本人到達性はないと言える。

一方、ホットスポット変異以外の体細胞変異、生殖細胞系列変異については、30 億塩基対のゲノム塩基配列上のどこにでも、かつ様々なパターンで発生し得るため、当該患者個人のみに認められる、個人に特異的な変異である可能性がある。

個人情報保護法において「特定の個人を識別することができるもの」を個人情報と定義し、氏名、連絡先等の情報が付加されていなくても、顔画像のように、別の画像を本人と対照して具体的な人物を同定できるものは個人と情報と示されたことを踏まえると、ホットスポット以外の遺伝子変異は、特定の個人に限定的なデータである可能性があり、他の研究者が有する電算記号化されたそのゲノムデータのコピーと照合することにより①一意性があると言える（別の言葉で表現すれば traceability がある）。

しかしながら、ゲノムデータのみから特定の個人に到達するには別途その本人の生体試料を用いてゲノム解析を行うか、あるいは当該個人のゲノムデータの全てあるいは一部が公開あるいは別途取得されている必要があり、現実的には③本人到達性は低い。ただし、病歴等臨床情報等の付加により③本人への到達の可能性が大幅に増すこととなる[4,5]。

なお、どの程度の頻度以上で見られる場合「ホット」スポットと呼ぶかという点については、使用される場面、使用する研究者に異なり、数値的な基準は示されていないが、ホットスポットの概念に含まれる重要な点は、それが疾患の発生機構上、生物学的な蓋然性をもつため、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。また、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。

②可変性については、治療やがん自体の進展の影響により、現存する生体の特徴としての体細胞変異情報は、変化する可能性があり、可変性を持つ。

以上より、ゲノムデータが適確かつ正確に取得されていることを前提とし、かつ過去の病歴としてのゲノムデータではなく、現存する生体個体の特徴としてのゲノムデータの個人識別性を考える場合、少なくとも下記のデータは個人識別符号とは考えられない：

1. 体細胞変異のゲノムデータであって、周辺の生殖細胞系列の塩基配列情報を含まないデータ。

   ※治療等の影響により、②可変性があるため。さらにホットスポットであれば①一意性も無い。

2. 生殖細胞系列変異のゲノムデータであって、単一のホットスポット変異のみのデータ。

   ※①一意性が無いため。なお、生殖細胞系列「多型」情報の個人識別性については、がん特有の問題ではないため、本意見書では省略するが、註 6)の議論を参照。

＜註＞

1) 今回の議論では、そもそもこれらの①一意性、②可変性、③本人到達性の指標の定義が必要である。

2) 今回の議論では、ゲノムデータは適確かつ正確に取得されたと仮定して①～③を理論的に議論する。実際にはゲノムデータは 100%正確には取得できず、その不正確さは①～③の全てを低下させる。但し、不正確性を含むデータも、データ量が大量であれば個人識別性が高まる。註 6)参照。

3) 体細胞変異の場合の適確なゲノムデータ取得における留意事項として、多くのがん細胞はゲノム不安定性を特徴としており、新しい変異が発生する確率が高いという性質がある。そのため生体試料を採取した時期や臓器における位置によって異なる変異集合（がんの heterogeneity）を有している。

4) Hayden EC. The Genome Hacker. Nat 497:173, 2013.

5) Bohannon J. Science 339:262, 2013.

6) ゲノム科学的には「一意性／本人到達性」は、①ゲノム配列データの精度、②データ量、③母集団の性質（サイズと集団遺伝学的特徴）に大きく影響される。単純に、categorical に、「このゲノムデータは個人識別符合である・ない」と分類できるものではない。

   ⒜ たとえば精度が悪い（体細胞変異を含むゲノムデータも、エラーのある「精度が悪い」生殖細胞系列ゲノムデータとして扱うことができるだろう）ゲノムデータでも、エラー（体細胞変異）に比して、十分大量の生殖細

胞系列の多型情報を伴っていれば、本人到達性が高くなる。

(b) 逆に精度が高くてもデータ量が少ない（たとえば ABO 血液型のデータのみ）と、本人到達性は低くなるが、それも母集団の「性質」（どの程度の「サイズ」か）による。

(c) さらに、母集団の「性質」として、その集団中の多型の頻度分布にも依存する。たとえばある集団中でヘテロ接合頻度が 1/2 である SNP を 27 箇所選ぶと、その 27 箇所について自分と全く同じ組合せを持つ確率は最大でも 1/134,217,728 となる。すなわち、母集団が 1 億人の集団であれば、27 個の SNP で一意性があると考えられる。アレルの数が多いマイクロサテライト等の多型を使えばもっと少ない座位で一意性に達する。

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」

分担研究報告書

欧米におけるゲノムデータの利用にかかる法制度 ―欧州データ保護指令/規則および米国 HIPAA プライヴァシー規則の匿名化ルールを中心に―

分担研究者　佐藤 智晶　　青山学院大学法学部　准教授

欧米において、いわゆる「ゲノムデータ」は、法制度上、日本でいうところの「個人識別符号」に該当するものと考えられており、その結果として一定の法的保護の対象とされているが、厳密に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[1]。むしろ、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、十分な匿名化を施せば本人同意なしに収集や利用ができる、というのが欧米の見解である[2]。

欧米を比べた場合、やはり欧州の方が匿名化（anonymisation）の要件は厳しく、ゲノムデータを同意なしで利用することは難しい。たとえば、欧州委員会 29 条作業部会の見解（Opinion 05/2014 on Anonymisation Techniques）によれば、ゲノムデータが個人識別性を帯びていることを前提に、提供元が個人特定できない処理を実施して作成したデータセットを第三者に提供し、第三者がデータセットを適切に使う場合（再特定しないで使う場合）、本人の同意なしにゲノムデータを利用する可能性を必ずしも除外していない[3]。しかしながら、十分な匿名化が実施されているか否かについては、提供元に残されている元データや他のデータセットだけでなく、提供先が参照可能な他のデータが考慮されるだけでなく[4]、さらに、個人特定にかかる費用と時間、匿名化処理を施した際

---

[1] 欧州に関する最新の論文としては、たとえば、次のものを参考にされたい。See Jane Kaye,et al., Med Law Rev (2013) doi: 10.1093/medlaw/fwt027 First published online: October 17, 2013, available at http://m.medlaw.oxfordjournals.org/content/early/2013/10/17/medlaw.fwt027.full

[2] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 0829/14/EN, WP216, 10 April, 2014, n. 27, available at http://www.cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf

[3] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 0829/14/EN, WP216, 10 April, 2014, at 10, available at http://www.cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf

[4] Id. at 9 (…"Thus, it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data. Only if the

の技術水準だけでなく今後の技術的な発展のようなすべての客観的な事情が考慮される[5]。今後の技術的な発展まで考慮して個人識別性を帯びるかどうかが判断されるとすれば、提供元としては相当に匿名化を施さなくてはならない。

米国でも、欧州と同じようにゲノムデータは原則として法的な保護対象であるが、十分な匿名化を施すことによって本人の同意なしに利用することができる[6]。HIPAA プライヴァシー規則が制定された当初から、匿名化の方法については賛否両論があったとされる[7]。なぜならば、HIPAA 法の規制対象となるのは、個人識別可能な医 療情報（individually identifiable health information）の取り扱いであったため、 規制対象外となる条件、すなわち、医療情報の適法な匿名化について関心が集まることになった。

先に説明したとおり、同規則における匿名化の方法が実際に変更されたことはないが、 最初の規則（案）からは一度だけ変更されて規則として公表されている。2000 年の 12 月 28 日に公表された規則と、1999 年 11 月 3 日に公表された同規則（案）では、明ら かに匿名化の方法が異なっている。規則（案）では、所定の情報を除去すれば匿名化された情報と推定する、という規定にな

---

data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous. For example: if an organisation collects data on individual travel movements, the individual travel patterns at event level would still qualify as personal data for any party, as long as the data controller (or any other party) still has access to the original raw data, even if direct identifiers have been removed from the set provided to third parties. But if the data controller would delete the raw data, and only provide aggregate statistics to third parties on a high level, such as 'on Mondays on trajectory X there are 160% more passengers than on Tuesdays', that would qualify as anonymous data."

[5] Id. at n. 6. See also The Final version of the EU General Data Protection Regulation, Recital 23, Dec. 15, 2015, available at https://iapp.org/media/pdf/resource_center/2015_12_15-GDPR_final_outcome_trilogue_consolidated_text.pdf (..." To determine whether a person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by any other person to identify the individual directly or indirectly. To ascertain whether means are reasonable likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development. The principles of data protection should therefore not apply to anonymous information, that is information which does not relate to an identified or identifiable natural person or to data rendered anonymous in such a way that the data subject is not or no longer identifiable. This Regulation does therefore not concern the processing of such anonymous information, including for statistical and research purposes."

[6] National Human Genome Research Institute in NIH, Privacy in Genomics, April 21, 2015, available at https://www.genome.gov/27561246#al-5 (..."In 2013, as required by the passage of the Genetic Information Nondiscrimination Act, the Privacy Rule was modified to establish that genetic information is health information protected by the Privacy Rule to the extent that such information is individually identifiable, and that HIPAA covered entities may not use or disclose protected health information that is genetic information for underwriting purposes. There are no such restrictions on the use or disclosure of PHI that has been de-identified.")

[7] HIPAA プライヴァシー規則における匿名化の方法については、次の論文などを参照されたい。たとえば、佐藤智晶「米国と欧州における医療情報法制をめぐる議論」東京大学政策ビジョン研究センターワーキング・ペーパーPARI-WP, No. 9, Jan. 15, 2013, available at http://pari.u-tokyo.ac.jp/policy/working_paper/WP130115_satoc.pdf

っていたのに対し、2000 年の最終規則では次のように変更された。 第 1 に、匿名化された医療情報について、個人識別に用いられるという合理的な理由がない場合、と規定された（Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information）。すなわち、2000 年の規則では、結果責任でも厳格責任でもなく、合理性の基準で匿名化されているかどうかが判断されることになった。 第 2 に、個人識別可能な医療情報の匿名化として、次の 2 つの方法が明記された。当然ながら、上記 2 つの匿名化の方法は、ゲノムデータにも適用される[8]。

1 つめは、情報の受領者が個人を識別してしまうリスクについて、最小化されていることを専門家が確認する方法である。専門家は、統計的または科学的手法によってリス クが最小化されていることを確認するものと規定されている。このような柔軟かつより合理的な別の匿名化を許容する規定は、規則（案）にはまったく含まれていなかった。

2 つめは、18 種類からなる所定の情報を予め除去し、残りの情報では個人識別できないことを確認する方法である（いわゆる、セーフハーバー・ルール）。18 のデータとは、名前、州以下の住所、誕生日等の年月日、電話番号、FAX 番号、E メールアドレス、社会保障番号（SSN）、診療録番号、医療保険の受益者番号、銀行口座の番号、資格等の番号、自動車登録等の番号、医療機器番号、ウェブの URL、IP アドレス、指紋や声紋等の生体認証記録、顔面写真等のイメージ、その他の個人識別コードである。

なお、米国では一塩基多型が 30 から 80 あると個人識別性を帯びうるという指摘があり[9]、逆にいえば、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ことになる。個人識別性のないゲノムデータであれば、当然ながら HIPAA プライヴァシー規則の保護対象にはならない。

---

[8] See, e.g., Simson L. Garfinkel, De-Identification of Personal Information, National Institute of Standards and Technology Internal Report 8053, October 2015, available at http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

[9] たとえば、El Emam K. Methods for the de-identification of electronic health records for genomic research. Genome Medicine. 2011;3(4):25. doi:10.1186/gm239, available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129641/ (“There is evidence that a sequence of 30 to 80 independent single nucleotide polymorphisms (SNPs) could uniquely identify a single person”). See also Lin Z, Owen A, Altman R. Genomic research and human subject privacy. Science. 2004;305:183. doi: 10.1126/science.1095019.

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」
分担研究報告書

ゲノムデータの持つ個人識別性

分担研究者　　竹内 史比古　　国立国際医療研究センター・研究所　室長

定量性

ゲノムデータの持つ個人識別性は、定量的であり、そのように扱う必要がある。ゲノムデータの規模が大きいほど、個人識別性は高くなる。全核ゲノムデータがあれば、特定の個人ないしはその一卵性双生児を識別できる。一方で、ABO 式血液型には 4 種類しかなく、5 人以上を識別できない。大まかには、前者は世界人口約 70 億人を約 70 億通りに識別でき、後者は 4 通りに識別できる。

ゲノムデータが対象とするゲノム領域

ゲノムデータの規模は、どの染色体のどの部分を対象にするかで規定される。具体的な領域の指定には、標準ゲノム配列を用いる。

標準ゲノム配列　http://genome.ucsc.edu/cgi-bin/hgGateway

個人識別性の定量

個人識別性は、シャノンの情報量（エントロピー）で定量するのが数理的に便利である。情報量は、集団を平均的に何通りに識別できるかを 2 のベキ乗として表しており、情報量が 2 なら 4 通り、30 なら約 10 億通りとなる。

情報量　https://ja.wikipedia.org/wiki/%E6%83%85%E5%A0%B1%E9%87%8F

ゲノムデータの持つ個人識別性の計算法

特定領域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より集団が何種類にどのような割合で分かれるかを評価して、情報量の計算式に当てはめればよい。

　集団の分かれ方は、その集団に依ることに注意が必要である。即ち、特定のゲノム領域であって

も、日本人集団なのか、アフリカ人集団なのか、あるいは全世界の集団なのかで異なってくる。個々の集団におけるゲノムの多様性と頻度の評価には、ゲノム多様性の調査研究が利用できる。そのような調査は、世界規模では千人ゲノムプロジェクト、日本国内では東北メディカルメガバンクなどで行われている。

世界のゲノム多様性　http://www.1000genomes.org/

宮城・岩手のゲノム多様性　http://www.megabank.tohoku.ac.jp/tommo/researchers

　ゲノムの複数の領域を合わせたようなゲノムデータの情報量については、領域間が離れていて遺伝的に独立している場合には、各領域の情報量の和で全領域の情報量を近似できる。離れていない領域については、一体のものとしてゲノムの多様性と頻度を評価する必要がある。

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」
分担研究報告書

個人特定性とゲノムデータ・遺伝的識別性の関係について

研究分担者　　徳永 勝士　東京大学大学院医学系研究科・人類遺伝学　教授

ゲノムデータの特徴

身体の部分についての情報とはいえ、顔や指紋等は視覚的に異同が認識しやすいのに対して、ゲノムデータは 4 種類の塩基の長大な配列であるため、目視のみで異同を判断することはできない点で大きな違いがある。したがってゲノム情報は、意図的に情報を抽出・加工することに加えて、個人名などの情報と共に登録されていて利用できるという条件がないと個人の特定につながらないことに特徴がある。

目的・用途による違い

鑑別（親子鑑定、個人識別）のための DNA 検査は、遺伝的な個人識別性を有することを意図して行われる検査であるのに対し、疾患遺伝子探索などの医学研究において得られるゲノム解析データは、そのままでは遺伝的個人識別性は低い。すなわち、このようなゲノムデータから意図的かつ専門的に個人識別性を有する情報を抽出して精度管理しない限り、遺伝的な個人識別性は生じない。

ゲノムデータの違い

理論的には、マイナー型の頻度が 0.5 に近く、互いに独立な（連載不平衡にない）ゲノム多型を 50 個程度検査すれば、一卵性多胎児を除いて遺伝的一意性が高くなる。ただし、個々の多型・変異の頻度分布には集団差があるため、各集団内での識別性と世界の諸集団も対象に含めた場合の識別性の程度には違いがある。

現実のゲノム医学研究においては、germ line（胚細胞系列）のゲノムデータのみならず、somatic mutation（体細胞に生じた突然変異）を解析したゲノムデータやエピゲノム研究から得られるゲノムデータもあり、それぞれ異なる特性を持っている。したがって、それらから個人識別性を有する情報を抽出して精度管理する作業も異なる。

さらに、解析対象とするゲノム領域によって多型・変異の密度が異なり、それらの独立性（連鎖不平衡の程度）が異なることから個人識別性の程度も異なる。

目的外使用の禁止

上記のように、研究を目的として得られるゲノムデータそのものは個人識別性が低い。ゲノムデータから意図的に個人識別性を有する情報を抽出して、研究目的以外に利用する行為を禁止する措置がとられるべきである。

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」
分担研究報告書

ゲノムデータの持つ個人識別性に関する研究

研究分担者　　俣野 哲朗　　国立感染症研究所エイズ研究センター　センター長

ゲノムデータの有する個人識別性について、一般論的に論ずることは困難である。したがって、データの解析目的・使用目的等もふまえ、個別の視点・論点に基づいて考慮する必要がある。

一つの遺伝子の多型解析のデータでは、一般に個人の特定にはいたらないが、非常に稀な多型を有する場合、個人の特定に結びつくことになる。研究の進展により変化する可能性を考慮すると、このような個人特定に結びつく可能性のある遺伝子を全て排除することは困難である。したがって、確率的に極めて低いリスクを有することを考慮したうえで、個人識別性は考えにくいという判断を下すことが妥当なケースが多々あると考える。個人識別性の有無という二者択一の概念ではなく、大小等レベルでの判断が必要で、それをどのように反映させるかが検討課題である。

まず、ゲノムの定義について考える必要がある。ヒト染色体ゲノムだけではなく、エピゲノム、各種 RNA、さらには蛋白質等の情報まで、近い将来、対象となりうることを考慮しておく必要がある。今回、ヒト染色体ゲノムに絞ったとしても、その由来組織・細胞の違いをどの程度考慮すべきかという問題がある。一方、腸内細菌叢や口腔内細菌等々のゲノム情報の問題もある。特に腸内細菌叢の microbiome 解析等、情報収集が急速に進展している現況において、今回の対象とすべきかどうかについての議論は必要である。

今回、国内法との関連で議論がなされているが、対象となるゲノムデータは、国外で得られるものも含まれ、また国外で使用されるものも含まれる。国際共同研究、海外機関で管理されているデータベースおよび国内機関で管理されているデータベースの海外機関の活用等々において、国内での個人識別性の判断およびその対応について、国外と整合性がとれる必要がある。このような状況を十分ふまえた判断が必要である。

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」
分担研究報告書

タイトル：ゲノムデータの持つ個人識別性

研究代表者　　吉倉 廣　　国立感染症研究所名誉所員

１．緒言

「ゲノムデータの持つ個人識別性に関する研究班」開始に当たって、意見交換を行い、ゲノムデータの一意性の検討において考慮すべき点として、次の項目を挙げた。

- ゲノムデータが由来する人口集団(国籍、人種、集団サイズ等)により一意性の程度・到達度が異なり得る
- 対象は、単一遺伝子か、（同一個人に由来する）複数遺伝子か
- ゲノムデータと同時に得られる一意性を上げるゲノム以外のデータの考慮
- ゲノムデータの使用目的により、対象遺伝子の種類が異なる
- 染色体部位によって異なる連鎖不平衡の考慮、等。

2. 一委員としての意見：

１．　ある遺伝子型の人が、対象人口の中にいる確率が、たった一人ならば、その遺伝子には個人識別性がある、とするのが最も単純な考え方である。

２．　この考え方の中で出てくる係数は、遺伝子型頻度(f)、対象人口数(P)、でその遺伝子型の一意性は式

$$f^{*}P \leq 1 \quad (*\text{は掛け算}) \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

で表される。

３．　このように定式化すると、幾つかの問題が浮かび上がる。

- 遺伝子型に由来する問題として：ゲノムデータの全ての遺伝子型についてその頻度 f が分かっている訳ではない。同じ遺伝子型でも頻度 f は母集団により必ずしも一定ではない。
- ゲノムデータ適用対象母集団の問題として：(1)の式から、同じ遺伝子型頻度でも集団サイズが大きければ、一意性は下がり、集団サイズが小さければ一意性は上がる。人種、地域、でその集団本来の遺伝子型頻度は異なりえるので、頻度の高い集団では一意性は下がり、頻度の低い集団では一意性が上がる。
- 遺伝子サイズが大きくなれば、情報量は大きくなり一意性が増す。即ち、遺伝子のサイズは各遺伝子の一意性に影響する。

以上の議論は単一遺伝子座を例にしているが、一個体由来複数遺伝子座をデータベースとする場合についても、問題は同様である。

4．　ゲノムデータベースには、複数の遺伝子情報が存在し、それらの遺伝子の頻度・地理的分布は、遺伝子型により異なり、結果として一意性も異なる。

5．　ゲノムデータには、その、生物学的意味が分かっているものと、分かっていないものがあり、これらを同等に扱うかどうかは議論の余地がある。疾病に関する情報は、当人の保険等に於いて不利となるものもあり、特別な考慮を要する。

6．　以上の様に考えると、一意性の検定により、ゲノム情報の個人識別符号への適合性を機械的に検討する事には可成りの無理がある。むしろ、データベースの構築、使用に視点を移し、よりプラグマテイックな手法が適切かも知れない。

7．　その具体例は、佐藤委員の紹介している米国 HIPAA（Health Insurance Portability and Accountability Act, 2000 の匿名化法である。２つの手法があり、

- 情報受領者が個人を識別するリスクが最小化されている事を専門家が統計的或いは科学的手法により確認する。
- 18 種類の所定の情報（名前、州以下の住所、誕生日等の年月日、電話番号、FAX 番号、E メールアドレス、社会保障番号（SSN)、診療録番号、医療保険の受益者番号、銀行口座の番号、資格等の番号、自動車登録等の番号、医療機器番号、ウェブの URL、IP アドレス、指紋や声紋等の生体認証記録、顔面写真等のイメージ、その他の個人識別コード）を一切除去する。

関連コメント

１．　ゲノムデータベースにつき、2008 年頃の OECD の遺伝子検査に関する検討の中で、国境を越えた生体材料の移動と医療機関を介さない遺伝子検査の問題が指摘されていた。これは産業活動の自由と個人情報保護の双方に関わる問題であった。

２．　Google 検索をすると、多数の遺伝子検査の広告が見られ、中には試料が国外の安価な検査機関に送付され検査されている可能性もある。このような検査機関は、試料を受け取るので、その意図を持てば、営利事業を行いながら大量のゲノム情報（例えば日本人に多い癌関連遺伝子情報）を蓄積し、特定集団の遺伝子データベースを作り、遺伝子標的創薬の為の情報として、商業的に或いは戦略的に利用或いは販売することが出来る。

３．　合理的な予防治療或いは保険加入の為と云う理由で、個人を対象にし、保険会社と提携した遺伝子検査を勧める商業活動がある。このような動きは、検査を受ける個人の利害に関する問題（遺伝子情報に基づく保険掛け金額差別）の他に、長期的には、このような商行為を通して構築したゲノムデータベースを、保険企業が営利目標に使うことを可能にする。

４．　しかし、この行為には、営利目的の遺伝子検査だとして、排除出来ない側面がある。検査会社が消費者のゲノム検査を行い、生命保険会社はこの検査会社と契約を結ぶ。生

命保険会社は保険金を払う段階で、その消費者の疾病又は死亡原因を知り、検査会社の取得したゲノムデータに対照すれば、多数の遺伝子の種々の疾病への関与を、大量のデータを持って解析出来る。この仕組みは、効率の良い prospective study の一種であり、商業活動をしながら医療の向上に貢献出来る。大きな産業に繋がる。

5. 23andMe のアン・ウオジツキーの云う様に、もしも「消費者は自らのデータを使った医学の発展に貢献したがっている」(毎日新聞 2016.6.2) のであれば、DTC (Direct to Consumer)の遺伝子検査では、消費者が自分の意志で自己の生体材料を送っている、と云う事になる。当然自分の住所氏名、払い込み等の銀行番号を検査会社に送っているのであるから、消費者は「個人識別性の極めて強い情報をセットにして」、会社に提供しているという理屈になる。つまり、DTC に於いては、そもそも消費者は自己の個人情報の保護を考慮していない。もし、そうであれば、この件は消費者と検査会社との商取引の問題に帰し、「個人情報保護」の範疇から外れる、との解釈も成り立ち得る。

6. 多くの人々が、Facebook、Twitter、LINE、Google+、YouTube 等のソーシャルメデイアでプライバシーを公開していることを考えると、個人の意志による遺伝子情報の公開を止める根拠は、少なくとも見つからない。現在の個人情報保護の議論は、このような社会的変化に追いついていないかも知れない。

補足説明

個人情報保護委員会事務局から解析方法・データ保存・個人識別の基準各項目につき質問があったので、これに応えて作成したものが以下の文書である。

解析方法

1　DNA 塩基配列決定（https://en.wikipedia.org/wiki/DNA_sequencing）

i.　**古典的手法**

初期の DNA 塩基配列：**プライマー延長法を原理とするもの**

全ゲノム塩基配列決定：Direct-Blotting-Electrophoresis-System GATC 1500（**EU のゲノム解読計画で使用**）;AB370I **半自動** DNA **シーケンスシステム**（Applied Biosystem）、Genesis 2000 (Dupont)　(1980 **年代終わり頃**)；**ショットガン法**。

ii.　**次世代**

Single-molecule real-time sequencing (Pacific Bioscience)

Ion semiconductor (Ion Torrent sequencing)

Pyrosequencing (454LifeSciences)

Sequencing by synthesis (Illumina)

Sequencing by ligation (SoLiD sequencing)

Chain termination (Sanger sequencing)

iii.　**開発中**

Nanopore DNA sequencing

Tunnelling currents DNA sequencing

Sequencing by hybridization

Sequencing with mass spectrometry

Microfluidic Sanger sequencing

Microscopy-based techniques

RNAP sequencing

In vitro virus high-throughput sequencing

2　短い縦列反復配列タイピング:　PCR して電気泳動により反復回数を測定

3　一塩基多型タイピング：BeadChip (Illumina)、Axiom (Affymetrix)、TaqMan (Life Technologies)

注意：以上のリストは、上市或いは上市前のものを含め、全ての解析法を網羅している訳では無い。

保存

1．一般に電子媒体で保存され、そこから情報を引き出し使用されている（例 http://www.ncbi.nlm.nih.gov/gene/）。

現在一般的なゲノム解析法のいずれにおいても、ヒトゲノム約 30 億塩基対のうちの一部分がゲノムデータとして取得されている。即ち、

- 短い縦列反復配列（Short Tandem Repeat, STR）タイピング解析では、特定遺伝子座における STR の反復の回数を測定し、その回数をデータとして取得する。回数は、例えば 3 回だったり 4 回だったりする。
- 一塩基多型（Single Nucleotide Polymorphism, SNP）タイピング解析では、特定遺伝子座に位置する SNP のアレルを測定し、その塩基をデータとして取得する。アレルは、例えばアデニン（A）だったりシトシン（C）だったりする。
- DNA 配列解読解析では、特定遺伝子座における数百塩基に渡る領域について、塩基配列を測定してデータとして取得する。塩基配列は、例えば AAGCCTACAC のように塩基が連なっている。

ヒトゲノム　（https://en.wikipedia.org/wiki/Human_genome）

ゲノム情報の一意性の議論においては、ヒトゲノムがどのような構築になっているかの理解が必要である。

1. ヒトの遺伝情報は、２２対の常染色体に加えて、女性では 1 対の X 染色体、男性では一本の X 染色体と一本の Y 染色体に存在する。従って、常染色体上の遺伝子は原則 1 対（２個）あることになる。加えて、必ず母親に由来するミトコンドリアの遺伝子がある。人間とチンパンジーの間のゲノムの違いは 1.2%、人間の間での違いは 0.1%と云われている。
2. ヒトの全ゲノムの内、タンパクをコードしている領域は全体の 1.5%で残りはタンパクをコードしない。遺伝子をコードしない領域は（非コード領域）、

   Pseudogene

   Noncoding RNA

   Introns and untranslated regions of mRNA

   Regulatory DNA sequence

   Repetitive DNA sequences（反復配列）：minisatellite, microsatellite, satellite; short interspersed nuclear element (SINE）; DNA transposon; LTR retrotransposon; long interspersed nuclear element (LINE); rDNA 等

   （http://www.nature.com/nrg/journal/v13/n1/box/nrg3117_BX1.html）

   これら反復配列は、全ゲノムのほぼ半分になる。
3. この内、反復配列は、個人特定の為に法医学で汎用されている技術である（http://aboutforensics.co.uk/dna-analysis/）。
4. ヒトのゲノム解読は緒についたばかりで、特に、非コード領域の機能については不明な点が多い（例：脳細胞での発現が高い retrotransposon の生理的意味など）。解読進展に伴いゲノム情報の個人情報としての各情報の評価も変化する可能性がある。

一意性の判断

1. 多くの研究班員の指摘するように、
   - 同一対象者から情報が得られる遺伝子座の数、
   - 当該遺伝子座の対立遺伝子、
   - 遺伝子頻度（調査対象集団で替わり得る）、
   - 調査母集団人口、
   - 調査対象集団の地域特異性等、

   が一意性に影響する。
2. 個人識別の目安として次の様な提案がなされている。
   - ゲノムデータによる個人識別としては、便宜的ではあるものの、100 万人から 1 人を特定できるなら個人識別できると便宜的に定義できる（しかし、1 億人から 1 人を特定できると書き換えても、便宜的である事には変りない）。
   - 対象集団で多様性が高く、しかも互い連鎖不平衡にない遺伝子座であれば、30 遺伝子座を測定すれば、約 100 万人から 1 人を特定でき、個人識別できることになる。
   - DNA 塩基配列解読においては、極めて低頻度な配列が発見されることがあり、その場合はその配列を持つ人が特定できることになる。
   - ゲノム解析法においては、多くの遺伝子座を測定するほど、より高い確率で個人を識別できるようになる。
3. ゲノムデータの為の生体試料は「ヒトゲノム・遺伝子解析研究に関する倫理指針」、或いは、OECD 指針に基づいて作成された「遺伝子関連検査に関するベストプラクテイスガイドライン」に従って入手されている筈なので、インフォームドコンセントを得る手続がある筈である。この段階で、ゲノムデータから個人を特定する可能性(一意性)についての説明が求められた場合には前項のような説明が可能である。

別紙４

研究成果の刊行に関する一覧表

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書　籍　名	出版社名	出版地	出版年	ページ
該当なし							

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
該当なし					