# 電子カルテ情報を用いた証拠性のある臨床研究手法に関する研究
# （H27-医療-指定-016）
# －測定、画像データ証拠性・安全運用環境検討－
# 遺伝子解析研究への証拠性の導入

研究分担者　作佐部　太也
藤田保健衛生大学医療科学部臨床工学科　　准教授

**研究要旨：　目的：遺伝子解析研究における証拠性を確保する上で、データ改ざんの防止に暗号技術の導入するための方法を検討することを目的とする。方法：塩基配列データについてのハッシュ値を計算する実験を行い、また、改ざん防止のためのシステム構築の可能性についての調査を行った。結果：ハッシュ値の計算の実施には技術的な問題のないことが分かった。一方、改ざん防止システムの構築の必要性が高いことが明らかになるとともに、実現のためには多くの問題があることも明らかとなった。結論：改ざん防止のシステムを構築するためには、研究者や学会などの学術組織だけでなく企業などとも連携した社会的な取り組みが必要であることがわかった。**

## A.研究目的

今日においてゲノムの塩基配列解析研究における次世代シーケンサー（NGS）は、基礎研究を超えて臨床研究のツールとなりつつある。またNGSが生成する塩基配列データ(以下NGSデータ)を活用するための各種の基盤の整備も進んでおり、統計処理やデータベースなどで高度な技術が投入され、更には、それらを自由に利用する文化が根付きつつある。

このことは研究の発展という観点からは良好な状況と考えられるが、研究における不正の防止という観点からはリスクを孕んだ状況でもある。特にデータの改ざんという不正の防止についての検b 討が必要と考えられる。これは画像処理ツールの利用の普及が研究における不正を少なからず助長したという指摘[1]からも、類似の現象がNGSデータにおいても起こることは推測しうることである。

そこで、本分担研究においては、前年度の調査を踏まえ、NGSデータの証拠性確保のため暗号化技術の導入のための方法について、調査研究および提案を行う。

## B.研究方法

1) ハッシュ値の計算実験

**データの改ざんを検出するための基盤となる情報処理はハッシュ値などの要約データの生成であり[2]、どのような高度な改竄防止の仕組みを構築するにしても、ハッシュ値の生成と元データとの関連付けについての実現性は前提となる。そこ**

**で本分担研究ではNGSデータについてのハッシュ値の生成を実際に行いその時間を計測した。また、生成したハッシュ値とNGSデータとの関連付けの方法として最も単純で信頼性の高い埋込みについて、実際にハッシュ値をNGSデータに埋込む実験を行い、NGSデータを解析するシステムに影響を及ぼさない方式を探索的に求めた。**

2) NGS関連環境の調査

**ハッシュ値の生成・埋込みをどのように行うことがデータ改竄の防止に効果的であるかについて検討するために、実際のNGSの運用現場や公共データベース、研究者が利用するソフトウェア等について文献やインターネットなどによる資料に基づいて調査を行った。**

**倫理的配慮**

平成 28 年度の分担研究においては、特に個人情報を取扱うなどの倫理的な課題は発生しなかった。

## C.研究結果

1) ハッシュ値の計算実験

ハッシュ値の計算実験については、添付1、添付2によって報告した。NGS**関連で運用されるサーバと同程度の能力の計算機を用いて全エクソームで30秒程度、全ゲノムでも10分以内であった。元来**NGSデータは巨大であり、どのような処理にしても分単位の時間がかかることから、ハッシュ値の計算にかかる時間はNGSデータの解析処理の中で行っても全体的な処理時間への影響はほとんどないと考えられる。また、ハッシュ値のNGSデータの埋込みについても、解析システムに影響を与えない方式の実現性も確認できた。

2) 生成･埋込みのタイミング

ハッシュ値の生成･埋込みの処理を誰が、また、どのようなタイミングで行うことができうるかについて調査、検討を行った。

なお、NGS自身にそのような機能を付与することは理想的であるが、NGSの製造メーカへの問い合わせなどは本分担研究では行っていない。

NGSを運用する施設において処理を実施するケースについて検討した。

一例として分担研究者の所属する研究機関においては生物学系研究者が独自に解析システムを構築し運用している（中間報告書参照）。解析システムの構成やシステムを構築した研究者の情報処理技術についてのスキルからみて、プログラムを提供すれば、NGSの出力直後にハッシュ値の生成･埋込みを行うプログラムを解析システムに組込むことは可能であると考えられる。

より大規模な研究機関であれば、情報処理の専門技術者や研究者も所属している場合もあり、プログラムの開発を含めて可能であろう。

NGSを運用しているが自前の解析システムを持たない研究機関においてはNGSデータを解析業者に送付する前に、研究者がハッシュ値の計算と埋め込みを行わなければならない。従って計算機操作についての高度なスキルを持たない研究者でも操作できるよう、操作が容易なプログラムを提供する必要ある。また、NGSを持たず試料を委託先に送付してNGSの処理を外部に委託して行う場合には受託業者が行うことになるが、技術的には問題は無いと考えられる。

何れの場合においても、ハッシュ値の生成および埋込みの方法について正確に記述された仕様を策定し配布することは必須である。可能であれば、実際に動作するプログラムを配布することが望ましく、特に解析システムがオープンソースのプログラムによって構築される場合が多いことからも、オープンソースとして配布するのが望ましいと考えられる。

3) 公共データベース

NGSを用いて取得した塩基配列情報の解析に基づく研究では、研究成果の公表の際には際には取得したNGSデータの公共データベースへの登録公開が義務付けられることがある。また、登録されたNGSデータは当該研究の証拠としてだけではなく、以後の別の研究において参照され再利用されることになる。

したがって、公共データベースへの登録前にハッシュ値は埋込まれなければならず、公共データベースから取得するデータにも保持されていなければならない。

公共データベースとして、国際的に協調、集約の動きがあり現在その中心となっているのは、INSDC (The International Nucleotide Sequence Database Collaboration) の活動として米国のNCBI (The National Center for Biotechnology Information)、欧州のEMBL-EBI (The European Bioinformatics Institute)、日本のDDBJ (DNA Data Bank of Japan) が協調して蓄積、管理しているSRA (Sequence Read Archive)である[3]。

SRAへの登録の際にはNGSデータはそのままで、研究や実験に関する情報は別のデータとして登録する。

SRAの内部ではNGSデータは独自のフォーマット (SRA形式) で保存されており、利用するためには解析システムに入力できるFastQ形式などに変換する必要がある。変換のタイミングとしては、利用者がSRA形式のデータをダウンロードして専用のツールにより変換する方法と、専用のダウンロードツールにより自動的に目的のフォーマットとしてダウンロードする方法がある。

従ってSRAの運営組織に対して埋め込んだハッシュ値を維持できるようシステムの改造を要請する必要がある。

4) NGSシミュレータ

NGSの普及とともに、NGSをシミュレーションするシステム(以下NSGシミューレータ)が開発されている[4]。NGSシミュレータは仮想的なNGSデータを生成するソフトウェアである。NGSシミュレータについての調査研究によると、塩基配列の変異について指定してNGSデータを生成できる機能をもつNGSシミュレータがある。これは、NGSデータの解析システムの挙動を詳細に検証するために重要な機能である。一方、別の観点からみると、NGSシミュレータは、実際のNSGによって得られるNGSデータと区別できないようなリアリスティックなデータを生成する可能性があるということである。すなわち、意図的な改ざんを行う上でNGSシミュレータは有効なツールになり得ると考えられる。

## D.考察

1) 改ざんへの障壁の低下

NGSデータのサイズは数ギガバイトから数十ギガバイトのである。今日では一般に市販されているノート型PCでも大容量のものは16GB以上の場合の主記憶装置が搭載されている。従って、データを直接的に編集する

ことすら困難ではなくなっている。

また、NGSデータを処理するプログラムとしてオープンソースのものが多く配布されており、かつ、実際に主要な処理として用いられている。それらの多くはUNIX系統のOS上で動作し、その操作にはGUIではなくコマンドラインを用いるものが多い。プログラムの使用方法についての情報が書籍やインターネット上で掲載されているが[5]、特にインターネットから場合、コマンドラインであればコピー＆ペーストにより簡単に実行させることができる。

加えて、医学生物学系の研究者がApple社製のパーソナルコンピュータを好むことは頻繁に言及されるが、現在、それらのオペレーティングシステム（OS）はUNIX系 統であり、プログラムのインストールから実行についてのスキルの障壁は低いものとなっている。

NGSデータに対して改ざんする意思を持つものが十分な分子生物学的な知識を持つとは容易に想定できる。一方、そのような者が実際に改竄を実施しようとする場合に最初に障壁となると想定されるのが計算機関連の機材の調達やスキルの習得である。そして今日、そのハードルが低くなってきているということである。

更にはNGSシミュレータを悪用することにより高度なねつ造が可能になると考えられる。

NGSデータについてハッシュ値を埋込むことによって、埋込み以後の改竄を検知する技術は実現できるが、運用としてそれを研究者に行わせることは無意味である。なぜならハッシュと取除き、データを改竄し、再度ハッシュ値を埋込むことにより、改ざんの検出はNGSデータ単体ではできなくなってしまう。

従って、NGSによってハッシュ値が埋込まれていなければならない。

2) 電子署名の必要性

NGSによりハッシュ値がNGSデータに埋め込まれたとして、実際に改竄を検出するにはNGSは生成したハッシュ値を研究者が操作できない所に半永久的に格納しておかなければならなくなる。これは全く現実的ではない。

この問題を解決するには公開鍵暗号技術を用いた電子署名を作成し、それをNGSデータに埋込む必要がある。この場合、改ざんを検知するためは、電子署名を生成する際に用いた秘密鍵と対になる公開鍵を用いて電子署名を復号することになる。その為には公開鍵基盤(PKI)との連携が必要になる。日本においては医療関連機関向けにHPKIが運用されているが、NGSデータは国際的に流通するデータであるため一国の組織としてではなく国際的な組織としてPKIを運用する必要がある。またPKIシステムに対する情報処理および管理上の負荷は非常に大きくなるものと予想される。

## E. 結論

NGSデータの証拠性を向上させるため、特に改竄を検知できるようにするためには、学術組織だけでなくNGSメーカや解析のためのソフトウェアやサービスを提供する営利組織、国際的な学術協調組織にまでまたがった取り組みが必要であることが明らかになった。

## 参考文献

1. 榎木 英介，“生命科学の研究倫理

なぜ不正が絶えないのか？”, KEIO SFC JOURNAL Vol.15 No.1 2015, pp.
3. Cryptography Second Edition: protocols, algorithms, and source code in C”, John Wiley & Sons Inc. 1996, p38.
4. Rasko Leinonen, Hideaki Sugawara, Martin Shumway, “The Sequence Read Archive”, Nucleic Acids Research, 2011, Vol. 39, Database issue D19-D21.
5. Merly Escalona, Sara Rocha, David Posada, “A comparison of tools for the simulation of genomic next-generation sequencing data”, NATURE REVIEWS GENETICS Vol. 17, pp.459-469.
6. 清水厚志,坊農秀雅, “細胞工学別冊 次世代シーケンサーDRY 解析教本”, 学研メディカル秀潤社, 2015.

## F. 健康危険情報

平成28年度の本研究においては、生命、健康に重大な影響を及ぼすと考えられる新たな問題、情報は取り扱わなかった。

## G. 研究発表

### 1. 論文発表

作佐部太也, 大内雄矢, 澤智博, 渡辺浩, 中島直樹, 木村通男: 証拠性のある医学研究 – 次世代シーケンサーからのデータの証拠性確 保における暗号技術の利用についての評価と提案第36回医療情報学連合大会, 医療情報学 第36回医療情報学連合大会論文集 36(Suppl.2), 340-362.

2. Bruce Schneier, “Applied 720-721, 2016

澤智博, 渡辺浩, 作佐部太也, 中島直樹, 木村通男: 証拠性のある医学研究 —次世代シーケンサー等のデータソースおよび 解析ソフトウエアの検討第 36 回医療情報学連合大会, 医療情報学 第36回医療情報学連合大会論文集 36(Suppl.2), 718-719, 2016

中島直樹, 渡辺浩, 澤智博, 作佐部太也, 宇山佳明, 山口光峰, 木村通男：証拠性のある医学研究 —病院情報システムからの EDC データ源に関する検討 —第 36 回医療情報学連合大会, 医療情報学 第36回医療情報学連合大会論文集 36(Suppl.2), 714-717, 2016

木村通男, 渡辺浩, 澤智博, 作佐部太也, 中島直樹: 証拠性のある医学研究 — Web 型小病院向け電子カルテシステムを用い た研究ノートの電子化 第36回医療情報学連合大会抄録集 722-723, 2016

### 2. 学会発表

作佐部太也, 大内雄矢, 澤智博, 渡辺浩, 中島直樹, 木村通男: 証拠性のある医学研究 – 次世代シーケンサーからのデータの証拠性確 保における暗号技術の利用についての評価と提案第36回医療情報学連合大会,2016年11月24日,横浜市

## H. 知的財産権の出願･登録状況
（予定も含む）

1. **特許取得**　　**なし**
2. **実用新案登録**　　**なし**
3. **その他**　　**なし**