105. 未検出の変異／多型の検出と機能的意義の解明

藤本　明洋

東京大学 大学院医学系研究科 人類遺伝学分野

Key words：遺伝的多様性，体細胞変異，長鎖シークエンス技術，情報解析

緒　言

人類集団には、様々な遺伝的多様性が存在し、疾患のリスクや表現型の個人差に関わっていることが知られている。また、がんにおいても様々な種類の体細胞変異が発癌の原因となることが知られている。近年のゲノム配列解析技術の革新的な発展により、多くの大規模ヒトゲノム、がんゲノム研究が行われており、全世界では100万人のゲノムシークエンスが行われつつある。しかしながら、それらのシークエンスデータは、主に読み取り長が 100 bp 程度の次世代シークエンサー（NGS）を用いて取得されている。そのため、主たる解析対象は一塩基の違いや短い挿入・欠失に限られており、解析が困難な繰り返し領域などの変異／多型についての知見は大きく不足し、革新的な解析技術の開発が課題である。一方近年、長鎖シークエンス技術が発展し、従来のNGSでは検出困難な多型や変異の検出に有効であると期待されているが、高いエラー率などの欠点がある。

我々は、情報解析手法を開発することでこれらの困難を解決し、人類の遺伝的多様性やがんの変異の真の理解と新たな疾患の原因遺伝子の同定を目的とした研究を行った。

具体的には3つの目的を遂行した：1. NGSデータからの検出が困難な変異／多型の検出［1〜3］、2. 長鎖シークエンサーの情報解析手法の開発による変異／多型、新規転写産物の検出［4］、3. 変異／多型の機能的意義の解明［2］である。

方　法

1.　マイクロサタライト変異の検出［1］

マイクロサテライトは、1〜6塩基の繰り返しである（例えば、ATATAT…やGGGGGG…など）。ヒトゲノムの領域をマイクロサテライトとするかは、定義に用いる手法により異なる。我々はRepeatMasker, Tandem Repeat Finder, MISAの3種類のソフトウエアを用いて15,737,726個のマイクロサテライトを定義した。その後、前後の領域の配列がユニークであり、近傍に他のマイクロサテライトがなく、長さが80 bp以下のマイクロサテライトを選出した。この結果、8,817,054個のマイクロサテライトが選出された。

次にマイクロサテライトの多型を検出する方法を開発した。マイクロサテライトにはAの連続、ACの連続などさまざまな塩基のパターンが存在している。それぞれのマイクロサテライトのパターンごとにシークエンスエラー率を推定した。エラー率の推定には、ヘミ接合でありヘテロ接合が存在しない男性のX染色体のデータを用いた。

推定したエラー率を考慮する確率モデルを用いて、エラーと真の変異を区別する方法を開発した。この手法を用いて、国際がんゲノムコンソーシアムの全ゲノム（21がん種、2,717サンプル）を解析した。1サンプルあたり約765万マイクロサテライトを調査した。

2.　中間サイズの挿入欠失多型の検出と機能的意義の解明［2，3］

30〜5,000 塩基の欠失や 30 塩基以上の挿入を正確に検出する方法を考案した。この方法では、重水らが開発したIMSindel法の結果をもとに、多個体のデータを合わせて解析する（Joint call recovery法と命名した）ことで、多型検出の偽陽性と偽陰性を共に低く抑えることができる。174人の日本人ゲノムシークエンスデータに、この方法を適用し

た。

さらに、遺伝子発現データが入手可能な 82 人の日本人のゲノムデータを解析し、中間サイズの挿入欠失多型と遺伝子発現の個人差の関連を解析した。有意な関連が得られた欠失から 2 つを選択し、CRISPR-Cas9 法を用いて細胞株に欠失を導入し、遺伝子発現量との関係を解析した。

3. 長鎖シークエンスデータの解析手法の開発による変異／多型の検出［4］

長鎖シークエンサー（Oxford Nanopore）を用いて 11 人の肝臓癌と正常組織由来の DNA サンプル（合計 22 サンプル）の全ゲノムシークエンスを行った。変異／多型検出手法（CAMPHOR ソフトウエアと命名）を開発した。100 bp 以上の挿入・欠失と逆位・転座を対象とし、遺伝的多様性と体細胞変異を検出した。

結果および考察

1. マイクロサテライト領域の体細胞変異の解析［1］

国際がんゲノムコンソーシアムの全ゲノム（21 がん種、2,717 サンプル）を解析した［5］。1 サンプルあたり約 765 万マイクロサテライトを調査したところ、約 20 万マイクロサテライトが 0.1%以上のサンプルで変異していた。これらのマイクロサテライトを用いてサンプルごとの変異率（変異しているマイクロサテライトの割合）を求めたところ、31 サンプル（胃癌、大腸癌、子宮体癌など）の変異率が高く、マイクロサテライト不安定性を呈するがん（MSI）と考えられた（図 1）。また、マイクロサテライトの中から、MSI サンプルの検出に適した 20 マイクロサテライトを選択した。これらのマイクロサテライトは、独立のサンプルにおいても MSI サンプルを従来のマーカーセットとほぼ同等の精度で検出できた。各マイクロサテライトの変異率（各マイクロサテライトにおいて、変異を有するサンプルの割合）を解析したところ、DNA 複製タイミングや DNA の形状がマイクロサテライトの変異率に影響することが示唆された。MSI サンプル（31 サンプル）における DNA 修復に関係する遺伝子の体細胞変異と生殖細胞変異を解析したところ、2 サンプルが DNA 修復系遺伝子に生殖細胞変異を持ち、リンチ症候群であると考えられた。残りの MSI サンプルの DNA 修復系遺伝子においては、構造異常や生殖細胞変異における点突然変異と挿入欠失変異よりも、体細胞における点突然変異と挿入欠失変異が多く、体細胞変異の寄与が大きいことが示唆された。

この研究は、Genome Research 誌に掲載された。また、カバーアートに選出された。

2. 中間サイズ挿入欠失多型の検出と遺伝子発現への影響［2, 3］

日本人集団で多型的な（個人差が大きい）約 8,000 個の挿入欠失を発見した。また、長鎖シークエンサー（Oxford Nanopore 社）を用いて全ゲノムシークエンスを行い、その結果と比較したところ、一致率が 97%であり、今回の方法の精度が高いことが示唆された。挿入多型に注目した解析を行ったところ、461 個のタンデム重複が存在していた。また、タンデム重複は DNA 複製のタイミングが早い領域に多かった。

すでに公開されている日本人集団の血球系の細胞株の遺伝子発現の個人差のデータを用いて、挿入欠失と遺伝子発現量の関連について調べたところ、約 4%の欠失が遺伝子発現の個人差と関連していた。また、遺伝子発現の個人差に関連する欠失をそれ以外の欠失と比較したところ、関連する欠失は遺伝子発現を調整するスーパーエンハンサー領域やプロモーター領域に多く、ヘテロクロマチン領域（ゲノム内の遺伝子発現が抑制されている位置）に少なかった（図 2）。さらに、ゲノム散在し機能的意義が不明のトランスポゾンの欠失の中にも遺伝子発現の個人差に関わるものがあり、トランスポゾンの欠失にも機能的なものが存在することが示唆された。発現に関連する欠失とそれ以外の欠失には長さの差はなく、欠失する領域の特性が重要であることが示唆された。遺伝子発現に関連する欠失のうち 2 つ（イントロンの中にある欠失と約 1 万塩基離れた遺伝子の発現量に関連する欠失）を選び、ゲノム編集技術 CRISPR-Cas9 法を用いて細胞株に導入したところ、遺伝子発現量が変化し、これらの欠失が機能的であることが強く示唆された。

挿入と欠失を比較したところ、挿入の方が遺伝子領域に存在する割合が多かった。また、挿入は欠失と比較し、遺伝子発現と関連する割合はやや低かった。

本研究の結果は、Genome Medicine 誌と Human Genetics 誌に掲載された。

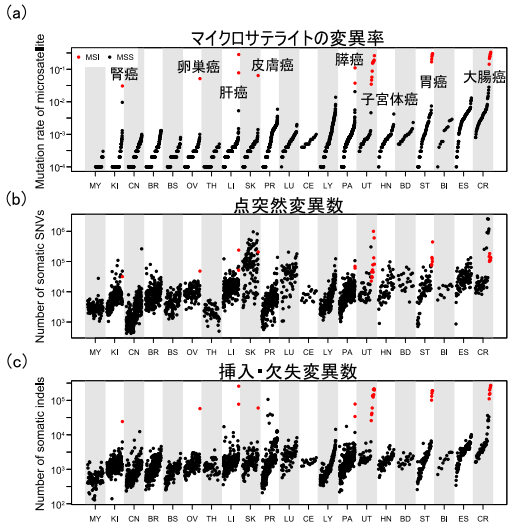

図 1. 21 がん種、2,717 サンプルの体細胞変異の比較

a）マイクロサテライトの変異率、b）点突然変異数、c）挿入欠失数。がん種ごとの比較結果を示している。赤は、マイクロサテライトの変異率が 0.03 以上でありマイクロサテライト不安定性（MSI）と判定されたサンプル。

3. 中間サイズ挿入欠失多型の検出と遺伝子発現への影響［2, 3］

日本人集団で多型的な（個人差が大きい）約 8,000 個の挿入欠失を発見した。また、長鎖シークエンサー（Oxford Nanopore 社）を用いて全ゲノムシークエンスを行い、その結果と比較したところ、一致率が 97%であり、今回の方法の精度が高いことが示唆された。挿入多型に注目した解析を行ったところ、461 個のタンデム重複が存在していた。また、タンデム重複は DNA 複製のタイミングが早い領域に多かった。

すでに公開されている日本人集団の血球系の細胞株の遺伝子発現の個人差のデータを用いて、挿入欠失と遺伝子発現量の関連について調べたところ、約 4%の欠失が遺伝子発現の個人差と関連していた。また、遺伝子発現の個人差に関連する欠失をそれ以外の欠失と比較したところ、関連する欠失は遺伝子発現を調整するスーパーエンハンサー領域やプロモーター領域に多く、ヘテロクロマチン領域（ゲノム内の遺伝子発現が抑制されている位置）に少なかった（図 2）。さらに、ゲノム散在し機能的意義が不明のトランスポゾンの欠失の中にも遺伝子発現の個人差に関わるものがあり、トランスポゾンの欠失にも機能的なものが存在することが示唆された。発現に関連する欠失とそれ以外の欠失には長さの差はなく、欠失する領域の特性が重要であることが示唆された。遺伝子発現に関連する欠失のうち 2 つ（イントロンの中にある欠失と約 1 万塩基離れた遺伝子の発現量に関連する欠失）を選び、ゲノム編集技術 CRISPR-Cas9 法を用いて細胞株に導入したところ、遺伝子発現量が変化し、これらの欠失が機能的であることが強く示唆された。

挿入と欠失を比較したところ、挿入の方が遺伝子領域に存在する割合が多かった。また、挿入は欠失と比較し、遺伝子発現と関連する割合はやや低かった。

本研究の結果は、Genome Medicine 誌と Human Genetics 誌に掲載された。

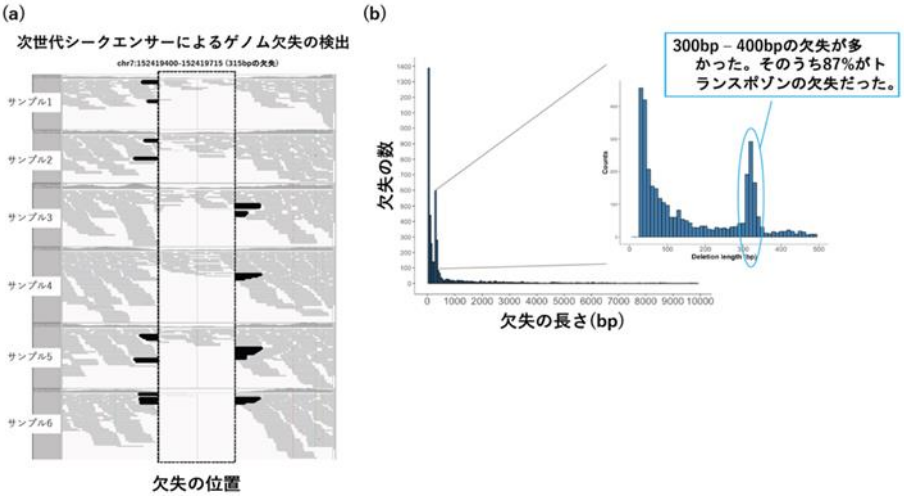

図 2. 欠失の検出

a) 次世代シークエンサーのデータからの欠失の検出 6 サンプルの例を示す。欠失を示唆するリード（シークエンサーが出力する塩基配列）は黒、それ以外は灰色。複数のサンプルでリードの本数が減っている箇所があり、欠失であると考えられる。

b) 検出された欠失のサイズの分布。短い欠失が多い傾向がある。300～400 bp の欠失も多く、それらの 87%がトランスポゾンを単位とする欠失であった。

4. 長鎖シークエンサーを用いた全ゲノムシークエンス［4］

Oxford nanopore シークエンサーを用いた全ゲノムシークエンス結果を開発した CAMPHOR ソフトウエアを用いて解析したところ、多型的な 8,004 挿入、6,389 欠失、27 逆位が検出された。挿入と欠失は短いものが多かった。しかし、挿入には、300 bp 付近の長さのもの、6,000 bp 付近の長さのものが多い特徴があった（図 3）。挿入・欠失の配列の特徴を調べた結果、挿入の 9 割はトランスポゾン由来であることが明らかになった。また、processed pseudogene の多型が 15 個検出された。processed pseudogene の起源は、多くの組織で高く発現している遺伝子であることが明らかとなった。

体細胞変異解析では、919 個の体細胞変異が検出された。またそのうち、499 個は我々の先行研究で NGS を用いて検出されていたものであり、420 個は新規候補であった。切断点の解析から、欠失や体細胞の構造異常の原因を推定したところ、NAHR、alt-EJ、FoSTeS／MMBIR、NHEJ によると考えられることが明らかになった。特に、NAHR（非相同組換え）による欠失は次世代シークエンサーを用いた先行研究では検出されていなかった。このことは、長鎖シークエンス技術の優位性を強く示唆する。本研究は、長鎖シークエンス技術の解析手法を構築するとともに、ヒトゲノムの多型や変異の全体像の解明に貢献すると考えられる。

本研究の結果は、Genome Medicine 誌に掲載された。また、開発したソフトウエアを利用して、難病のゲノム解析を行っている。

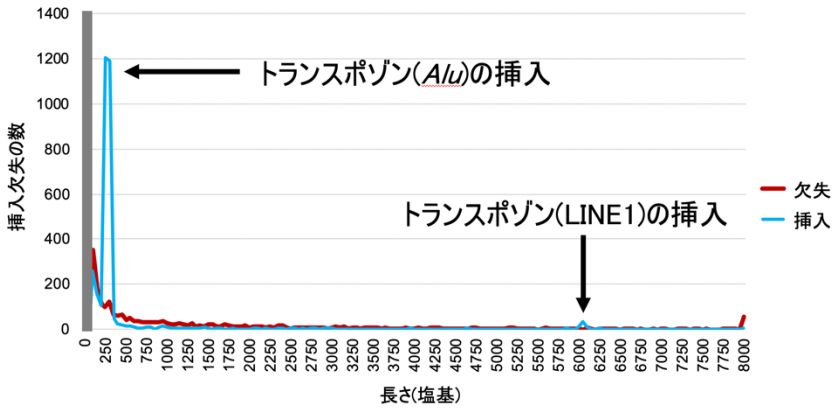

図 3. 日本人 11 人の全ゲノムシークエンスから同定された挿入欠失の長さの分布
100 塩基未満の挿入欠失は解析対象としていないため、0～100 塩基の領域は灰色で塗りつぶしてある。挿入と欠失で長さのパターンが異なる。挿入には、300 bp 付近の長さのもの、6 kbp 付近の長さのものが多く、これらはトランスポゾンの挿入であった。

謝　辞

上原記念生命科学財団の研究推進特別奨励金に大変感謝いたします。貴財団の助成金のおかげで、異動（京都大学から東京大学）に際してもほぼ問題なく研究を継続することができました。また、この研究費で所得させていただいたデータの解析は、学生の研究テーマにもなり、情報解析人材の育成にも貢献しています。

文　献

1) Fujimoto A, et al. Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. Genome Res. 2020 Mar 24;30(3):334–46. doi: 10.1101/gr.255026.119. Epub ahead of print. PMID: 32209592; PMCID: PMC7111525.

2) Wong JH, et al. Identification of intermediate-sized deletions and inference of their impact on gene expression in a human population. Genome Med. 2019 Jul 24;11(1):44. doi: 10.1186/s13073-019-0656-4. PMID: 31340865; PMCID: PMC6657090.

3) Ashouri S, et al. Characterization of intermediate-sized insertions using whole-genome sequencing data and analysis of their functional impact on gene expression. Hum Genet. 2021 May 12. doi: 10.1007/s00439-021-02291-2. Epub ahead of print. PMID: 33978893.

4) Fujimoto A, et al. Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. Genome Med. 2021 Apr 29;13(1):65. doi: 10.1186/s13073-021-00883-1. PMID: 33910608; PMCID: PMC8082928.

5) ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82-93. doi: 10.1038/s41586-020-1969-6. Epub 2020 Feb 5. PMID: 32025007; PMCID: PMC7025898.