トピックス

がんゲノム変異の解析
変異による生体機能への影響の評価、の視点から †

田代晶久*，荒木和浩**，中村建介*

1　はじめに

がんはゲノム変異の蓄積により、個体内の細胞が周囲との協調性を失い、無制限に増殖する能力を獲得することにより発症する疾患である。近年の、分子生物学の進歩、とりわけ、次世代シーケンサーの出現による配列解析技術の進歩にともない、個々のがん細胞についてそのゲノム配列を解読することができるようになり、さまざまなタイプのがんが存在することが明らかになってきた。また、個々の症例についてがん化の要因となっているドライバー遺伝子を特定することで治療の指針を与える「がんゲノム医療」も実際に用いられるようになってきている [1]。上皮細胞がんである肺がん、乳がん、大腸がんなどの癌腫については、がん化のメカニズムが次第に明らかになりつつある一方で [2]、希少がんとも呼ばれる肉腫などでは、未だに解明されていないがん化の要因がまだ多く存在すると考えられている。これら未知のがん化因子を特定していくために必要なツールの一つとして、特定の塩基変異によりどの遺伝子の機能がどの程度影響を受けるのかを評価するプログラムの作成を試みており、本稿ではその概要について簡単に紹介する。

がん化に影響を与え得る変異のタイプとしては、SNP、INDEL などのゲノム上の特定の位置における数塩基の異なる塩基への置き換えによる変化、特定の遺伝子を含む配列領域が重複したり欠失したりするコピー数変化（CNV）、DNA の繋ぎ変えが起きる組み換えや転座などがある。エキソーム解析では遺伝子ごとの CNV の評価は可能であるが、組み換えや転座については解析が難しい場合もある。本稿で主な対象とするのは SNP, INDEL などの微小変異である。これらは、参照ゲノム配列に対して「リードマッピング」を行い、参照ゲノム配列と異なる部位を特定する「変異コール」を行うことにより検出できる。検出された変異のうち既知のがん化因子については ClinVar などの病原性変異データベースとの照合により特定することができる。一般にがんが発症するために必要な変異の数は 70 箇所程度とも言われているが、公開されている SRA データなどで解析する限りでは、こうした既知病原性変異データベースとの照合で検出できる変異は 1 検体あたり 10 箇所程度で、見落とされている変異がまだ多数あるものと考えられる。

変異コールにより得られる数万箇所の変異から、がん化に寄与しているものを特定していくには実験的な手続きが必要であるが、情報解析によりあらかじめ個々の変異による影響の程度を評価できれば、重要な変異の絞り込みに有用であると考えられる。これを厳密に評価することは易しくはないが、産生されるタンパク質のアミノ酸配列を変化させる変異、タンパク質の発現を変化させる変異、であればそうでない変異よりも影響を与える可能性は高いと言える。同様の作業を行うためのプログラムとしては SnpEff が広く用いられている [3]。SnpEff は SnpSift と組み合わせて用いることで、それぞれの変異の重要度について、組み込まれたデータベースを参照しながら評価することができるとされている。我々も当初このプログラムの使用を試みたが、使い勝手にやや癖があること、また、原理的にはシンプルな手続きであり、既存のプログラムをブラックボックス的に用いるよりは、in house でプログラムを作成した方が、実際に行われている工程を正確に理解しながら研究を進めることができると考え、プログラムの開発を開始した。

2　方法

マッピングおよび変異の種類の同定に使うゲノム参照配列としては GRCh38.p13 を用いる。遺伝子アノテーションファイルは gff3 フォーマットで Ensembl より提供されているデータを使用する。次世代シーケンサから fastq 形式で得られたリードデータを bwa により上記参照配列にマッピングし、GATK により変異コールを行い、vcf 形式での変異リストを得る。作成するプログラムは C 言語により記述した。上記、GATK の出力である vcf フォーマットの変異リスト、ヒトゲノム参照配列の fasta ファイル、遺伝子アノテーションの gff3 ファイルの 3 つのファイルを入力として、vcf に列挙された個々の変異について、影響の評価を行い、結果を出力する。

3　結果

検出される個々の変異を評価するにあたって、まず変異箇所がゲノム上のどの領域にあるかを gff3 ファイルのデータを元に評価する。タンパク質をコードする遺伝子領域の外にも遺伝子発現に寄与する領域があるが、現状では考慮せず、一律に遺伝子領域外（OUT_GENE）とする。遺伝子内ではイントロン領域内であれば、スプライシングに影響を与えるドナー/アクセプタ領域であるか否かを判別する。また、エキソン領域内であっても 3’-および 5’-の非翻訳（UTR）領域内であればその旨を記述する。

†　原稿受理 令和４年２月２８日　Received March　28，2022
*　生命情報学科（Department of Life Science and Informatics）
**　群馬県立がんセンター　腫瘍内科

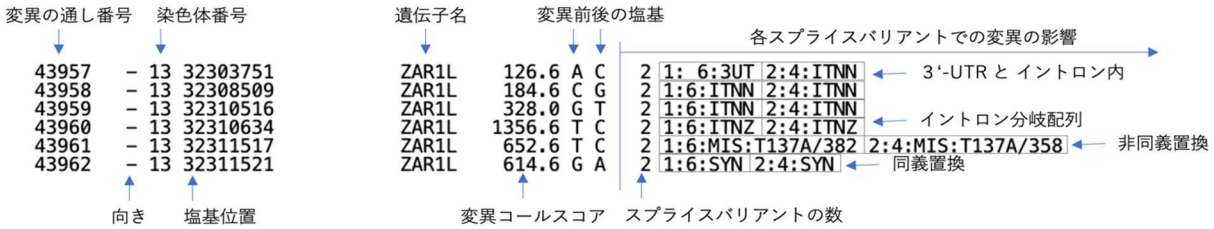

図１　作成したプログラムの出力例
遺伝子ZAR1Lの６つの変異のそれぞれ２つのスプライシングバリアントに対するアミノ酸配列への影響

コード領域の中での変異については、SNPS（一塩基置換）であれば同義置換（SYN）、非同義置換（MIS）、ナンセンス変異（NON）、停止コドンの消失（SLS）などに分類し、同義置換以外は、タンパク質中での変異位置と変異前後のアミノ酸配列の後にスラッシュ記号に続いてタンパク質の全長を記す。例えば、T137A/382、であれば、全長 382 のアミノ酸配列の 137 番目の T が A に置き換わる非同義置換である。また、ナンセンス変異においてはストップコドンを#記号で、T137#/382 のように表記する。INDEL（挿入・欠失）に関しては、挿入・欠失する塩基数が３の倍数であるか否かにより、インフレームインサーション（INS）、インフレームデリーション（DEL）、フレームシフト（FRS）、に区別する。変異前後のアミノ酸配列の表記については、SNPS の場合と同様にインフレームインサーションでは KD821RKD/921 のように、全長 921 のアミノ酸配列の 821 番目の K に続く D が、RKD に変化し１残基が挿入された変異を示す。インフレームデリーションについても同様に、QGQGQG182LG/538 の様に表すが、フレームシフトについては影響の及ぶ範囲が広く、表記が非常に長くなる場合があり、現状では元のアミノ酸配列の長さと、変異の始まる位置の表示にとどめている。

ある遺伝子について変異の影響を評価する場合に問題となることのひとつに選択的スプライシングがある。あるスプライスバリアントではエキソン内にある変異が別のスプライスバリアントではイントロン内である場合等があり、またエキソン内であってもコード領域内の位置が異なる場合もあるため、変異による影響は遺伝子がどのスプライスバリアントを取るかによって変わりうる。今回のプログラムでは一つの変異の一つの遺伝子に対する影響として、すべてのスプライシングバリアント、すなわちアイソフォームに対しての影響を評価し、１行に列挙することで、たとえば非同義置換を導く塩基変異が幾つあるのかを grep コマンド等により容易に判定できるように出力フォーマットを設計した。図１に出力ファイルの一部を示す。

4　現状と今後の課題

今回開発した、ゲノム上の塩基変異が生体機能に与える影響を評価するプログラムをエキソーム解析に適用することで実用性の評価を進めている。

これまでに判っている問題点のひとつとしては、複数の変異があたえる影響の累積がある。たとえば、翻訳時の同一のコドン内の２箇所で SNPS がある場合に、実際に起きるのは２つの塩基変異後のアミノ酸への変異であるが、現状では個々の塩基変異を独立に扱っているため、それぞれの塩基変異が別個に起きた場合の２通りのコドンの解釈によるアミノ酸変異が併記されてしまっている。また、アミノ酸配列のある部位でナンセンス変異あるいはフレームシフトによりストップコドンが生じた場合、その配列部位より下流で起きるアミノ酸配列の変化は意味を持たないため、出力する必要がない。実際にプログラムを開発し動作検証をすることで見えてきたこれらの様々な問題について、今後対応することを考えていく。

結果の最後に述べた様に本研究で開発するプログラムのポリシーとして、一つの塩基変異に対する出力を全てのスプライスバリアントについて１行内に収めるために、個々の表記を可能な限り短くし、有用な情報だけを表示する工夫をした。このため、一般的なアミノ酸配列の変異の表記方法のコンベンションに準拠していない部分もある。たとえばアミノ酸の表記は一般には Ala, Val などの３文字表記が推奨されるが本プログラムの出力は１文字表記による。また、停止コドンの表記は通常＊が使用されるが、本プログラムでは＃とした。これらの点については、オプションによる出力形式の変更を可能としたい。また、出力の文字数を制限したい理由もあって、停止コドンの消失の場合に追加される残基数と追加されるアミノ酸配列、あるいはフレームシフトの場合の変異後のアミノ酸配列の表記、は現時点では行っていないが、適切な表示方法を工夫しながら対応してゆきたい。

参考文献

1) A.A. Friedman, *et al.,* “Precision medicine for cancer with next-generation functional diagnostics”, *Nature Reviews*, **15**, pp. 747-756, (2015).

2) S.-V. Francisco, *et al.*, “Oncogenic signaling pathways in the cancer genome atlas”, *Cell*, **173**, pp. 321-337, (2020).

3) P. Cingolani, *et al.,* “A program for annotating and predicting the effects of single nucleotide polymorphisms. SnpEff: SNPs in the genome of Drosophila melanogaster strain w118; iso2; iso3”, *Fly* (Austin), **6**, pp. 80-92, (2012).