次世代シーケンサー（NGS）を用いた PGS のカバレッジの概念

着床前スクリーニング（PGS: preimplantation genetic screening）の手法は、array-CGH（comparative genomic hybridization）から次世代シーケンサー（NGS: next generation sequencer）を用いた手法に移行しつつあります。

NGS は全塩基配列を調べる手法という考えが根強いように思いますが、実は違います。今回は、PGS におけるカバレッジ（Coverage）の考え方について、詳しい方にご質問してみました。

Question 1 ：

「“カバレッジ”とは何でしょうか？ NGS を用いた PGS でのカバレッジとは、どのようなことか簡単に教えてください。」

Answer 1 ：

カバレッジ（Coverage）とは、簡単に説明すると、同じところを何回重ねて読んだかという指標であり、通常〇〇x カバレッジ（〇には数字が入る）というように表現されます。たとえば同じところを 20 回重ねて読まれていれば、20x カバレッジであると表わされます（図 1）。

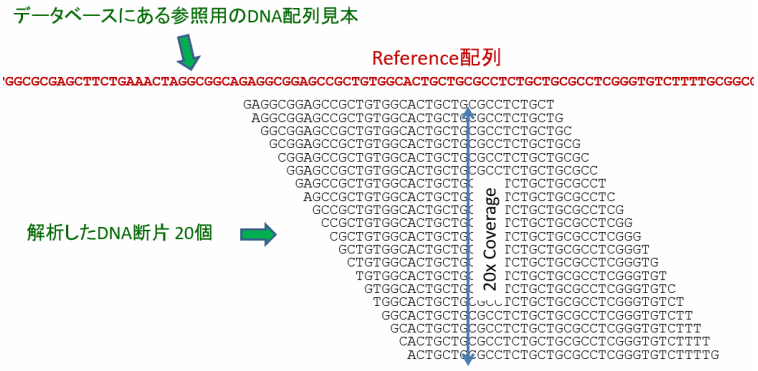

図 1. 20x カバレッジの場合。

このカバレッジの概念を理解するためには、まず NGS から産出されたデータがどのように解析されるのかを知る必要があります。NGS から産出されるデータは、ATGC の 4 種類の文字で構成される DNA 配列（文字列）です。たとえば、I 社の NGS ベースの PGS キットからは 1 断片あたり 36 塩基のデータ（36 文字）が産出されます。一方、参照する元のゲノム DNA 配列も文字列です。ヒトの全ゲノム配列は約 3G（ギガ）塩基（約 30 億文字）です。参照配列（この場合ヒト全ゲノム配列）に対してアライメントという作業（産出された配列を参照配列に対して検索）をしますと、その 36 文字の配列が何番染色体の何番目から何番目の塩基だったかがわかります。この作業はコンピュータ上で行われます。例えるならば、ワープロで特定の文字列を検索するような作業です。このアライメントの結果、この断片の配列が何番染色体のどの位置に由来するのかを知ることができます（図 2）。これを何度も何度も（場合によっては数千万回から数億回）繰り返してゆくと、図 1 のように読んだ配列が隙間なく重なりあうようになります。この重なりの回数がカバレッジと表現されるものになります。

図 2. 産出された 36bp の配列を Reference 配列にアライメント。

さて、それでは I 社の NGS ベースの PGS キットでは、カバレッジがどれくらいになるのでしょうか？このキットでは、1 検体あたり平均 100 万断片の配列情報が得られますので、上記のような検索を 100 万回繰り返します。36 塩基 x 100 万 = 3,600 万塩基ですので、「約 3 千万塩基（得られたデータ）÷ 約 30 億塩基（ヒト全ゲノム配列）」 という計算で、約 0.01x カバレッジになります。意外に少ないと思うでしょうが、染色体のコピー数を推測するアプリケーションですので、カバレッジはこの程度で十分なのです。

Question 2：

「I 社の PGS キットで検査した場合、いわゆる遺伝性疾患の原因変異を検出することは可能でしょうか？」

Answer 2：

いいえ、できません。繰り返しますが、このキットで産出されるデータ量はたったの 0.01x カバレッジにしかなりません。要するに、ゲノム上のたったの 1%の領域を 1 回だけ読む程度です。99%の領域は 1 回も読まれないことになります（図 3）。これでは、スカスカで何も見ることできませんよね。

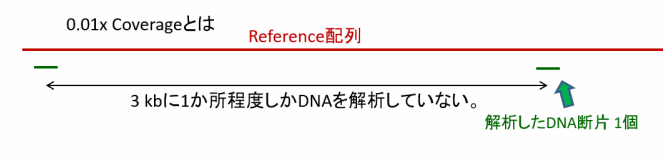

図 3. 0.01x Coverage とは。

通常、変異解析には平均 30x カバレッジが必要であると言われています。例えば、1 塩基置換変異（SNV）の場合、参照配列は「T」なのに対し、得られたデータでは「T」と「A」が半分ずつになっています（図 4 右）。この場合、この箇所は「T/A のヘテロである」と表現されます。また、塩基レベルでの小さな欠失や重複もアライメント作業の後に検出することが可能です。この例では、約半数の断片で 2 塩基（TA）の欠失が起こっていることがわかります（図 4 左）。

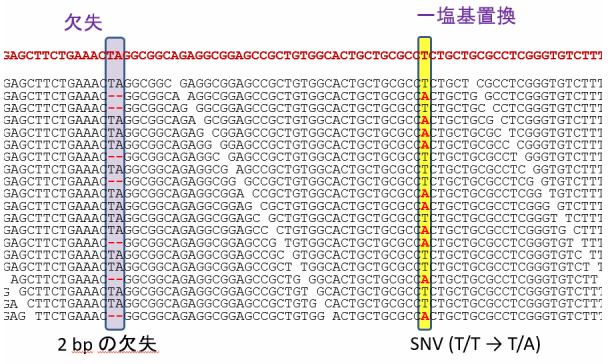

図 4. 一塩基置換と欠失の場合。

ヒトでは、両親からそれぞれ 1 セットずつ受け継いで計 2 セットのゲノムがありますので、メンデル遺伝する 1 塩基変異や塩基レベルでの欠失・重複が、どちらかのゲノムに存在すれば、上記 2 例のように通常 1:1 の比で観察されることになります。この 1:1 で存在する変異を検出するためには、当然、1x カバレッジでは不可能ですし、2x や 3x カバレッジでも確率的にアリルの片方しか検出されないことが多くなるのは容易に想像がつくと思います。10x カバレッジあれば、偶然片方のアリルしか検出できない確率は、$1/2^{10}$ (1/1,024) になるので、平均 30x カバレッジも必要ないと思われるかもしれません。しかし、実際には図 5 のように、ばらつきによりカバレッジが厚いところと薄いところが出てくることになります。また、配列によって読みやすいところと読みにくいところがありますので、全域にわたって十分なカバレッジを得るためには、平均して 30x カバレッジ以上になるように解析することが推奨されています。

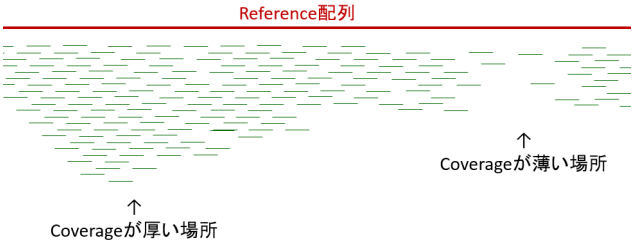

図 5. 実際の DNA 解析には、多少ムラが生じる。

ここで I 社の NGS ベースの PGS キットの話に戻りましょう。このキットでは、1 検体あたりのデータ量では 0.01x カバレッジにしかならないことをご説明しました。もしも、このキットを使って変異解析をするのに十分な 30x カバレッジにするためには、キットの 3,000 検体分をたったの 1 検体に使用しなければいけません。つまり、1 検体あたり 3,000 倍のコスト（数千万円）と 3,000 倍の労力が必要になるということになります。

また、データ量とは別の問題もあります。このキットでは全ゲノム増幅された DNA を解析することになりますが、この全ゲノム増幅に使われている酵素は複製のエラー率が比較的高いため、増幅エラーと真の変異を区別することが非常に困難になってしまうのです。以上のことから、I 社の PGS キットで染色体の異数性以外を検出することは、ほぼ不可能であることがわかっていただけるのではないでしょうか。