効率的なターゲット次世代シーケンスのためのオンターゲット率に対するカバレッジ均一性の重要性

Yehudit Hasin-Brumshtein, Ph.D., Maria Celeste M. Ramirez, Ph.D., Leonardo Arbiza, Ph.D., Ramsey Zeitoun, Ph.D.

はじめに

次世代シーケンス(**NGS**)は、研究と臨床の両面において遺伝子の変異を検出できる優れた技術です。シーケンスのコストは着実に下がっていますが、大規模な全ゲノムシーケンスはまだ極めて高価であるため、多くの研究はターゲットシーケンスによって特定の遺伝子や部位にフォーカスされています(Dillon, et al. 2018)。

ターゲットシーケンスでは、シーケンスに先立ち目的のゲノム領域を濃縮する必要があります。エクソームシーケンスでは、例えば、エクソン領域にハイブリダイズするビオチン化合成DNAプローブを設計します。ゲノムDNAサンプルとハイブリダイズさせた後、プローブを精製し、エクソン領域が濃縮されたサンプルを作成します。ターゲットの濃縮はシーケンスコストを下げ、実験をより実現可能でフォーカスされたものにしますが、また、シーケンスエフォートの効率性を損なうバイアスも生じます(Goldfeder et al. 2016, Meynert et al. 2013, 2014)。

ターゲット次世代シーケンスの確率的性質のため、ある程度の非効率性は避けることができず、ターゲットを濃縮するためのプローブパネルの設計および作成に、その労力の多くを必要とします(Warr et al. 2015)。一部のプローブは、非標的部位にクロスハイブリダイズし、これが「オフターゲット」（非特異的）キャプチャになります。プローブパネルはまた、キャプチャ効率の不均衡（均一性の喪失）を起こし、一部のターゲットの過剰な濃縮と、それ以外のターゲットの濃縮不足をもたらすことがあります。研究者は、信頼性の高いデータを確保するためシーケンスの量を増やし、低い読み取り深度で領域のカバレッジを高める必要があります。しかし、この戦略は適切にカバーされた領域の過剰なシーケンシングをもたらし、結果としてシーケンスコストが高まり、効率が低下します。

この「無駄なシーケンス」の程度は、ターゲットシーケンスの全体的な効率を示す均一性とオンターゲット率の2つの指標に反映されます。このホワイトペーパーでは、市販されている典型的なエクソームキットで広く用いられる様々なオンターゲット率と均一性を使用して、この2つの指標の全体的な効率に対する相対的な影響を数学的にモデル化します。ほとんどの市販のプローブパネルはスペックとしてオンターゲット率のみを記載していますが、私たちは均一性こそがターゲットシーケンスの効率にとってより重要であることを示します。

シーケンス要件の評価

シーケンス実験を設計する際の基本作業は、実行可能データ（**読み取りカバレッジ**）のために、サンプルあたりに必要なリード数を決定することです。その答えが、コスト、実現可能性、含めるサンプル数、そして有意な結論に達するための最終的な検出力を決定します。アプリケーションごとに異なる読み取りカバレッジが必要です。例えば、ある指定の位置にアライメントされた10回のリードからの情報（10倍のカバレッジ）は、研究レベルにおける生殖系列変異のコールには十分ですが、臨床レベルにおける体細胞突然変異の信頼性のあるコールとしては不十分かもしれません。当社では希望するカバレッジを**C_D**、実験で実際に観測された平均カバレッジを**C_M**としています。

理想的なシーケンス実験では、ターゲット領域全体にわたり均一かつ排他的な読み取りが行われます（完全な均一性と完全なオンターゲットキャプチャを意味します）。残りのゲノムは読み取りが行われません（**図1A**）。この理想的なシナリオでは、シーケンス効率は100%であり、$C_M = C_D$となります。しかし、不均一でオフターゲットのキャプチャを避けることは難しく、実際にはカバレッジは変動します（**図1B**）。

ほとんどの標的領域をC_Dに到達させるためには、多くの場合は$C_M >> C_D$になるようにシーケンスの量を増加します（**図1B**）。しかし、この戦略では、シーケンス読み取りの大部分が無駄になります。C_M/C_D比は、ターゲットの特定のパーセンテージがC_Dに到達するために必要な過剰シーケンスの量を表します。この値が大きいほど、より多くの過剰シーケンスが十分に利用可能なデータを得るために必要となります。ターゲット次世代シーケンスの効率の最適化では、結果に妥協することなくC_M/C_D比を最小化することが重要です。

図1

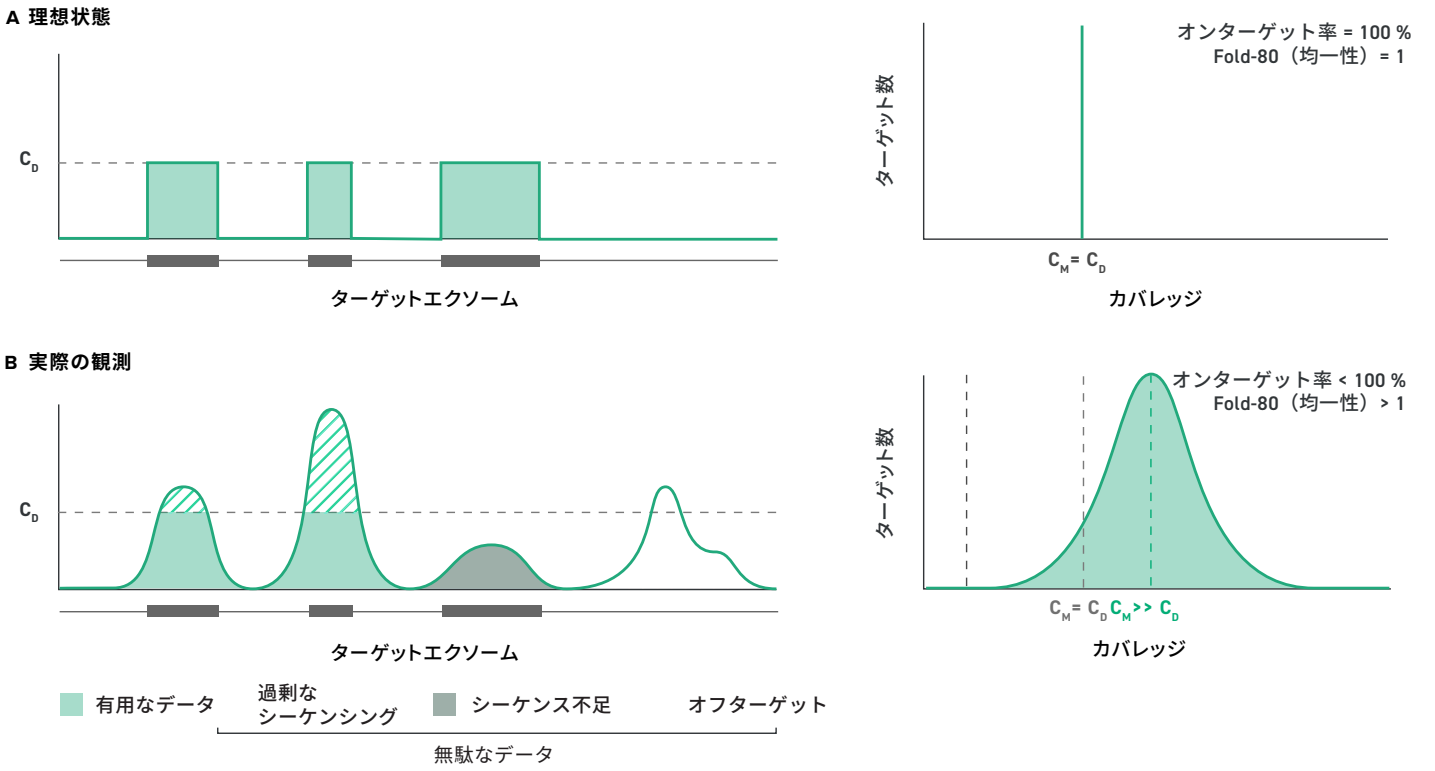


図1. 読み取り分布。A. 理想的な実験における読み取り分布。すべての標的が特異的かつ同等の読み取り深度を持ち、非標的部位の読み取りは存在しません。この場合では$C_M=C_D$となります。**B.** 実際のカバレッジ分布。一部の標的におけるシーケンス不足、あるいは過剰なシーケンス、そしてオフターゲット領域もキャプチャされています。

均一性とFOLD-80指標

均一性とはゲノムの標的領域に沿った読み取り分布を意味します。均一なカバレッジは全ての対象領域に対して十分なカバレッジ深度に達するために必要なシーケンスの量を減らします。均一性はC_Mの周囲への広がりを示す尺度であり、読み取り分布の平均と分位数から推定されます（**図2**）。

均一性を表す便利な指標が**fold-80**ベースペナルティです（fold-80と省略します）。広く使用されているPicard[1]のパイプラインで計算されるfold-80は、C_Mを達成する標的塩基の80%を保証するために必要な追加シーケンスの倍数です。例えば、100万回の読み取りで30倍のC_Mが得られる場合、fold-80が2.0となれば、標的塩基の80%が30倍のカバレッジに到達するために200万回の読み取りが必要であることを意味します。fold-80が1.4であれば、同じ目標を達成するためにシーケンスを140万回の読み取りに増やすことを意味します。

正規分布を想定すると、fold-80は変動係数（C_Mに対する標準偏差の割合）に比例し、1.0より大きくなります（fold-80が1.0であれば、完全な均一性と変動がないことを意味します。図1A）。fold-80スコアが高いほど、より広範なカバレッジ分布と低い均一性を意味し、fold-80スコアが低いほど高い均一性を示します（全ての標的塩基が同様のカバレッジでシーケンスされます）。

図2

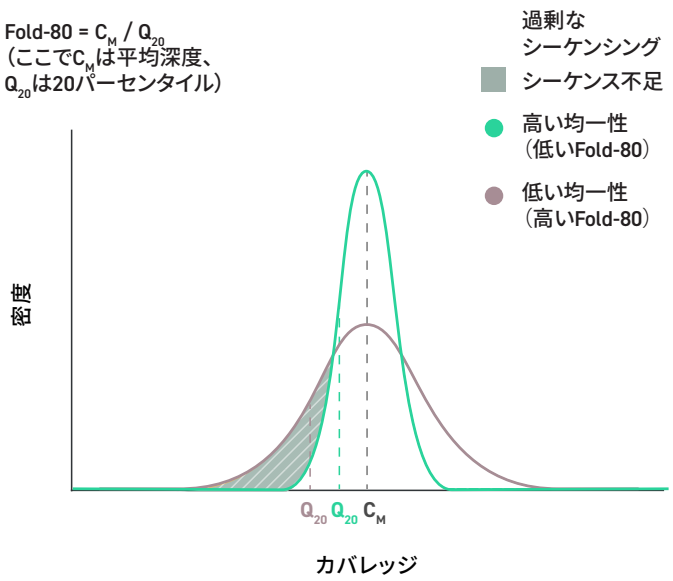


図2. 均一性は分布の形を反映します。低値（緑色）と高値（灰色）のfold-80スコアおよび過剰シーケンスとシーケンス不足を含む読み取りマッピングの相対存在量を示す、2つの異なる仮説読み取り分布曲線。fold-80スコアが下がると（灰色から緑色の曲線へ）より効率的な読み取りが利用され、シーケンス不足の部分が補われて過剰なシーケンス部分が減ります。実際には、均一性が悪い場合は非対称的な分布を示します。

オンターゲット率

オンターゲット率は標的領域にマッピングされるシーケンスデータのパーセンテージを表します。反対にオフターゲット率は、それ以外の領域にマッピングされるシーケンスデータを表します（**図1B**）。これは一般的には、標的領域をカバーするシーケンスの塩基数の、シーケンサーによって出力されたマッピングの塩基数に対する比として表されます（**図3**）。オフターゲットシーケンスを完全に避けることはできません。かなりの割合がプローブパネルに特異的で、無差別的なハイブリダイゼーションによる可能性があります。

図3

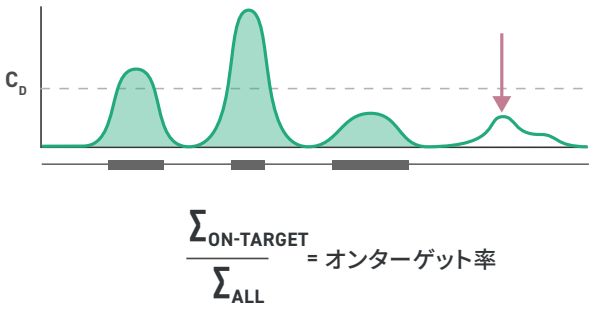


図3. オンターゲット率は、標的領域にマッピングされるシーケンス結果の割合です。オンターゲット率の計算では、シーケンス曲線以下の面積が全シーケンス量(Σ_{ALL})を表し、緑の面積がオンターゲット領域($\Sigma_{On\text{-}Target}$)を表します。オフターゲットシーケンスは矢印で示しています。

[1] *https://broadinstitute.github.io/picard/*

均一性とオンターゲット率の最適化の相対的な影響

均一性(fold-80)とオンターゲット率の両方が、標的シーケンスの効率を定義します。しかし、それぞれの指標はどれくらい影響するのでしょうか？

プローブパネルのライブラリ調製条件が同一である場合、オンターゲット率はほとんど変化せず、シーケンス結果においては、「課税」程度であるとみなすことができます(Chilamakuri et al. 2014)。均一性が完全（つまりfold-80が1.0）である場合、オンターゲット率とC_Mは反比例します。例えば、10倍のカバレッジ(C_D)を想定し、均一性が完全である場合、80%のオンターゲット率は、12.5倍のC_Mを目指すことを意味します。

$$C_M = C_D/\text{オンターゲット率} = 10 / 0.8$$
$$C_M = 12.5x$$

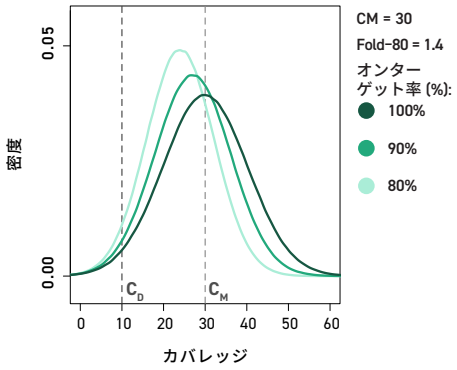
逆に、fold-80の改善が小さい場合でも効率性は大幅に向上できます。均一性を改善すると過剰シーケンスによる標的のカバレッジが減り、シーケンス不足による標的のカバレッジが向上します。

オンターゲット率と均一性の相対的な影響を調べるため、均一性、平均カバレッジ、そしてオンターゲット率の異なる3003個の正規分布をシミュレーションしました[2]。均一性を一定に保ちながら、オンターゲット率を向上させると（**図4A**）、カバレッジ分布はより高い平均(C_M)値に移動し、希望のカバレッジ(C_D)以上にカバーされる塩基の割合が増加します。前述のように、fold-80スコアを改善すると、シーケンス不足領域が補足され、かつ過剰シーケンス領域の割合が減り、結果としてリードの利用率が向上します（**図4B**）。この例では、平均カバレッジ(C_M)値は一定ですが、希望するカバレッジ(C_D)以上にカバーされる塩基の割合が増加します。両方の図とも、有効な塩基の数の違いはC_D以下の曲線間の面積で表されます。

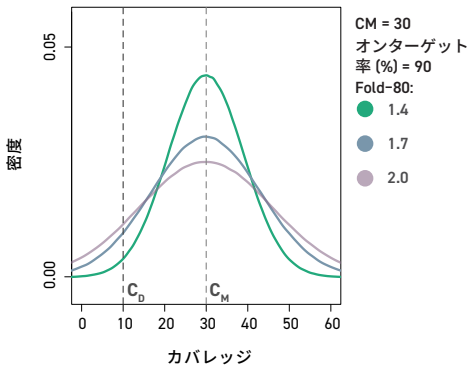
図4Cは、オンターゲット率、fold-80スコア、そして平均カバレッジを変化させたときの複合的な影響を示したものです。各色の曲線は異なるfold-80を示し、曲線の太さはオンターゲット率が80%（曲線の下限）と100%（上限）の間にあるときに取得される有効な塩基の割合を示しています。各曲線では、C_Mが30倍のとき、オンターゲット率を80%から100%に改善することで実質的にすべてのオフターゲットシーケンスを排除し、有効な塩基の割合を1–2%増やすことができます。これに対して、fold-80を1.7から1.4に改善すると、この数値は5–6%と劇的に増加します。

このデータは、オフターゲット率をゼロまで減らすことができたとしても、オンターゲット率の向上より、fold-80スコア（均一性）の改善の方が、ターゲット次世代シーケンスの効率に大きな影響を及ぼすことを示しています。

A オンターゲットの変更、Fold-80一定



B Fold-80の変更、オンターゲット一定



C 目標カバレッジにおけるFold-80とオンターゲットの塩基の割合に及ぼす影響

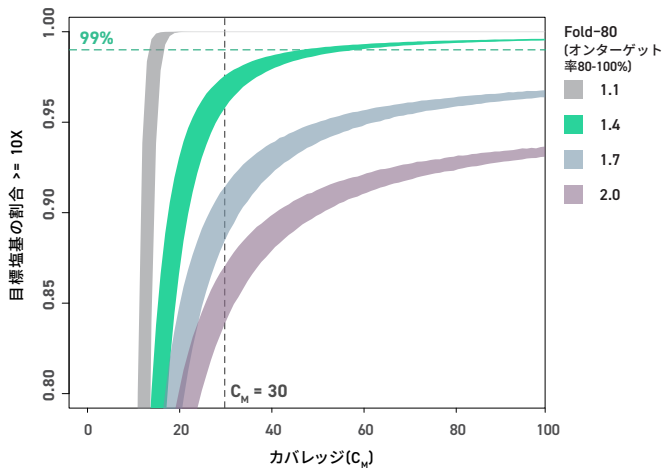


図4. シーケンスの必要な深度における均一性とオンターゲット率の影響。シミュレーション結果は、望ましいカバレッジ(C_D) = 10x、カバレッジ深度の正規分布、平均カバレッジ(C_M)の変化、オンターゲット率(0.8–1.0)、fold-80 (1.1–2.0)を想定しています。**A.** オンターゲット率は変化するが、fold-80とC_M（それぞれ1.4と30）が一定である場合のカバレッジ分布のシミュレーション深度。オンターゲット率の改善は、平均カバレッジを高め、分布を右側に移動させます。**B.** fold-80が変化し、オンターゲット率とC_M（それぞれ0.9と30）が一定である場合のカバレッジ分布のシミュレーション深度。fold-80スコアを改善（低下）すると、過剰シーケンス標的のカバレッジが減り、シーケンス不足標的のカバレッジが向上します。**C.** オンターゲット率、fold-80スコア、平均カバレッジを変化させた際に、10倍以上でカバーされる標的塩基の割合。

[2] オンターゲット率と均一性の概念を直感的に説明するために、正規分布を使用しました。実際のカバレッジ分布は通常は正規分布に従いませんが、分析の一般的な結論は、次世代シーケンスに典型的な分布にも当てはまります（正確な数値は異なることがあります）。

結論と展望

ターゲット次世代シーケンスでは、均一性(fold-80)とオンターゲット率が、シーケンスの効率を評価するための重要な指標です。これら2つの指標は、プローブパネル自体の本質的な性質であり、これらを最適化することで信頼性の高いデータを得るために必要なシーケンスの量を減らすことができます。

最も効率的なターゲット濃縮システムを選択するためには、提供される実際の均一性とオンターゲット率を注意深く評価する必要があります。オンターゲット率も重要ですが、ここではfold-80スコア（均一性）の改善がターゲット次世代シーケンスの効率に極めて大きな影響を及ぼすことを示しました。

参考文献

Chilamakuri CSR, Lorenz S, Madoui M-A, Vodák D, Sun J, Hovig E, Myklebost O, Meza-Zepeda LA (2014) Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15(1): 449.

Dillon OJ, Lunke S, Stark Z, Yeung A, Thorne N, Gaff C, White SM, Tan TY (2018) Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders. Eur J Hum Genet 26(5): 644–651.

Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, Salit M, Ashley EA (2016) Medical implications of technical accuracy in genome sequencing. Genome Med. 8(1): 24.

Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics. 14: 195.

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. 15: 247.

Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. (2015) Exome Sequencing: current and future perspectives. G3: *Genes|Genomes|Genetics*. 5(8):1543–1550.