覆盖均一性在高效靶向新一代测序中的重要性大于中靶率

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

前言

在科研及临床中，新一代测序 (**Next-generation sequencing, NGS**) 已经成为变异检测的首选技术。尽管测序成本正稳步降低，但大规模全基因组测序的成本仍然极其昂贵，因此研究往往采用靶向测序技术着重研究特定基因和位点（Dillon 等人，2018 年）。

靶向测序需要在测序前富集目标基因组区域。例如，在外显子组测序中，首先需要针对外显子区域设计并合成与其能互补杂交、生物素标记的 DNA 探针，然后将探针与基因组 DNA 样本杂交、纯化，最终得到外显子区域富集样本。虽然目标区域富集可以降低测序成本，提高实验的可行性和针对性，但也会产生偏好，这些偏好会影响测序工作的效率（Goldfeder 等人，2016 年；Meynert 等人，2013 年，2014 年）。

虽然靶向 NGS 的随机性不可避免地会对效率有所影响，但效率低下的更大原因在于目标区域富集探针组合的设计和生产本身（Warr 等人，2015 年）。一些探针会与非目标区域交叉杂交，导致“脱靶”（非特异性）捕获。另外，由于探针组合中不同探针浓度缺乏有效控制或探针设计时不同区域探针覆盖不一（均一性不足）而出现捕获不平衡，导致一些目标区域过度富集而其他目标区域富集不足。为确保获得高可信度的数据，研究者必须增加测序数据量，以提高低覆盖深度区域的覆盖度。然而，这一策略会导致对已经充分覆盖的区域过度测序，从而提升测序成本，降低测序效率。

这种“浪费性测序”的程度体现在均一性及中靶率上，这两项指标可反映靶向测序的整体效率。在本白皮书中，我们使用市售外显子试剂盒常见的中靶率和均一性进行数学建模，来研究这两个指标对整体效率的相对影响。我们的研究表明，尽管大多数市售探针组合在其说明书中只提及了中靶率，但均一性对靶向测序的效率具有更重要的贡献。

评估测序要求

测序实验设计的基本目的是确定每个样本需要多少测序片段 (reads) 才能获得可用于分析的数据（**覆盖深度**）。每个样本所需的测序片段数决定了测序的成本、可行性、每次测序的样本数，以及研究能够得出有意义的结论的把握度。不同的应用对覆盖深度的要求不同：例如，在胚系变异的研究中，由 10 个比对到给定区域的测序片段（10X 深度）提供的信息可足够确认一个突变，但在临床中，这些信息不足以可靠地确认一个体细胞突变。我们用 $\mathbf{C_D}$ 表示期望的覆盖深度，用 $\mathbf{C_M}$ 表示实验中实际观察到的平均覆盖深度。

理想的测序实验会产生仅涵盖目标区域且均匀分布的测序片段（即实现了完美的均一性和靶向捕获 (on-target capture)），在基因组的其余区域则不会产生测序数据（**图 1A**）。在这种理想情况下，测序效率为 100%，C_M 与 C_D 相等。然而，不均一及脱靶捕获是不可避免的，而且会导致覆盖深度在不同目标区域各不相同（**图 1B**）。

为确保大部分目标区域的覆盖深度达到 C_D，通常会增加测序量，使 $C_M >> C_D$（**图 1B**）。然而，这一策略会浪费大量测序数据。C_M/C_D 比值表示为确保一定比例的目标区域达到 C_D 所需的过度测序量：比值越大，获取足够可用数据所需的过度测序量越大。因此，优化靶向 NGS 的效率涉及在不影响测序结果的前提下将 C_M/C_D 比值最小化。

图 1

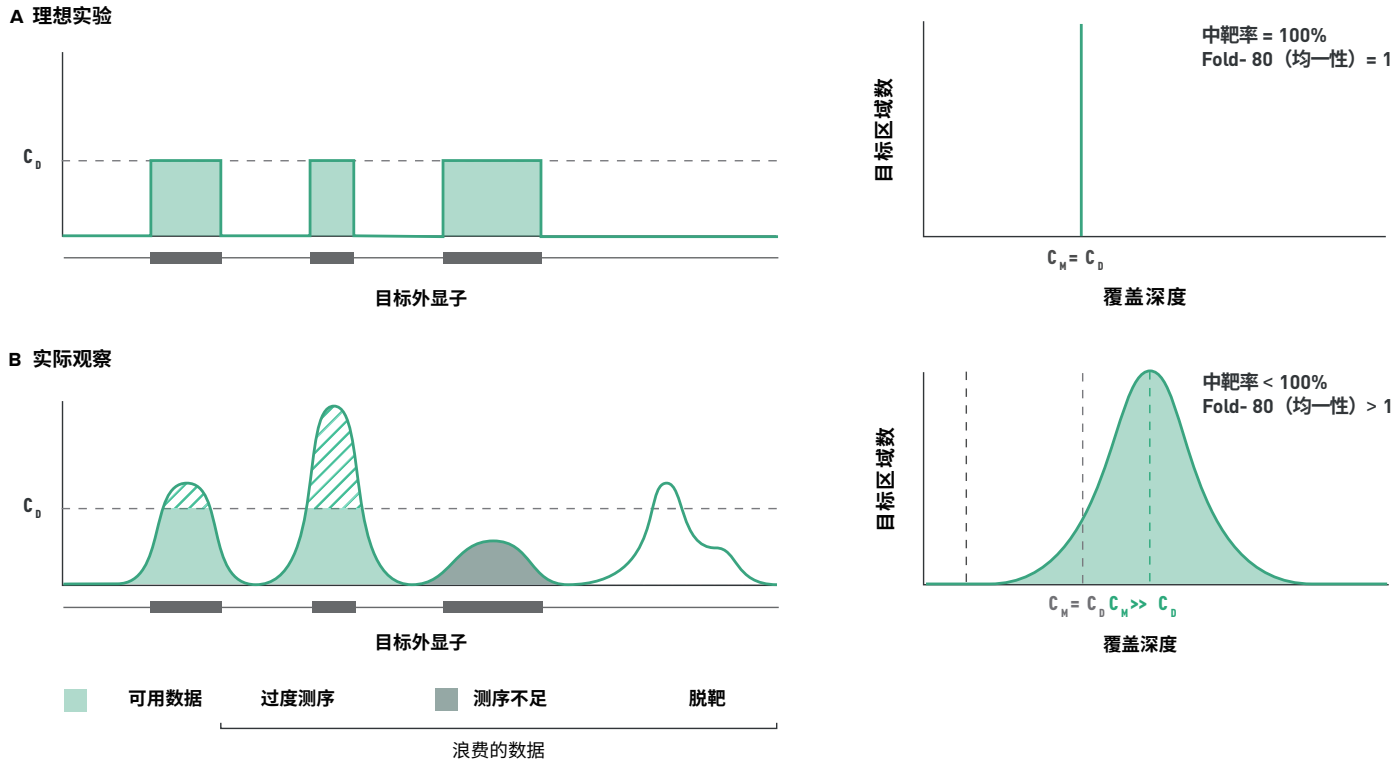

图 1. 测序片段分布。A. 理想实验中的测序片段分布。其中，所有目标区域都具有特定且相等的覆盖深度，而非目标区域无测序片段。在这种情况下，$C_M = C_D$。**B.** 覆盖深度的实际分布。其中，部分目标区域测序不足，其他区域测序过度，同时还捕获了非目标区域。

均一性和 Fold-80 指标

均一性描述了基因组目标区域的数据分布。均一的覆盖可减少所有目标区域达到足够的覆盖深度所需的测序量。均一性是体现 C_M 分布的度量，根据覆盖深度分布的平均值和分位数估计而得到（**图 2**）。

Fold-80 碱基罚分（Fold-80 base penalty，简称 Fold-80）是衡量均一性的一个实用指标。Fold-80 是确保 80% 的目标碱基达到 C_M 所需的额外测序倍数，通过被广泛采用的 Picard[1] 流程计算得到。例如，如果 1M 测序片段的 C_M 为 30X，那么 Fold-80 为 2.0 则表示需要 2M 数据时才能确保 80% 的目标碱基达到 30X 覆盖深度，Fold-80 为 1.4 则表示测序量需要增至 1.4 M 才能确保 80% 的目标碱基达到 30X 覆盖深度。

假设为正态分布，Fold-80 与变异系数（标准偏差与 C_M 的比值）成正比，且大于 1.0（Fold-80 为 1.0 表示完全均一且无方差，图 1A）。较高的 Fold-80 分值具有较宽的覆盖深度分布和较低的均一性，而较低的 Fold-80 分值表示均一性较高（所有目标碱基具有相似的覆盖深度）。

图 2

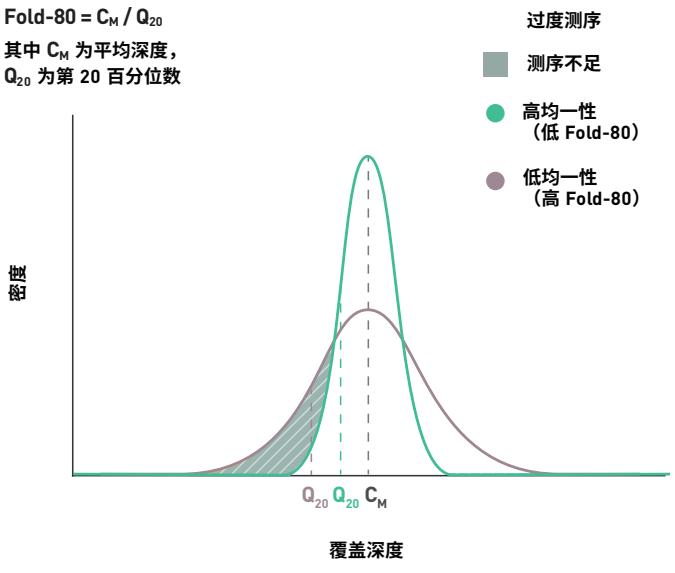

图 2. 均一性影响分布形状。假设的两种不同覆盖深度分布曲线，其中用灰色和绿色曲线分别表示高、低两种 Fold-80 分值，并示出了比对到过度测序区域（绿色阴影）和测序不足区域（灰色阴影）的测序片段相对丰度。降低 Fold-80 分值（灰色曲线至绿色曲线）既可增加测序不足区域的覆盖率，又可减少过度测序区域的比例，从而提高测序数据的利用效率。在实际情况中，均一性差时，分布也会不对称。

中靶率

中靶率表示比对至目标区域的测序数据比率；反之，脱靶率表示比对至其他区域的测序数据比率（**图 1B**）。中靶率通常表示为覆盖目标区域的测序碱基数与测序仪得到的比对上的碱基总数的比值（**图 3**）。一定程度的脱靶是不可避免的；很大一部分脱靶是探针组合特异性的，可能因非特异性杂交所致。

图 3

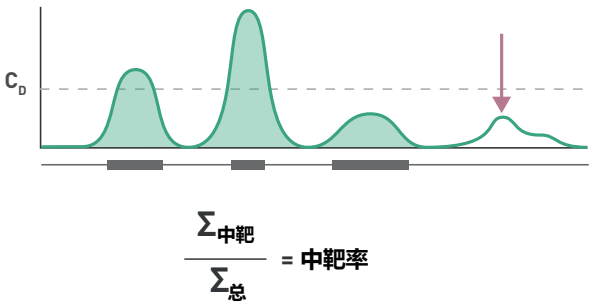

图 3. 中靶率是指比对到目标区域的测序数据比例。计算中靶率时，总的测序数据量（$\Sigma_{总}$）由测序曲线下面积表示，中靶区域（$\Sigma_{中靶}$）由绿色区域面积之和表示。图中箭头指示非靶向测序部分。

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

优化均一性与中靶率的相对影响

均一性 (Fold-80) 和中靶率共同决定了靶向测序的效率，但它们各自的影响有多大呢?

只要探针组合的文库制备条件一致，中靶率的变化通常就很小，可以看作是测序过程中的“税收”（Chilamakuri 等人，2014 年）。当具有完美的均一性（Fold-80 为 1.0）时，中靶率与 C_M 成反比。例如，假设期望的覆盖深度 (C_D) 为 10X 且具有完美的均一性，那么中靶率为 80% 意味着 C_M 应为 12.5 倍：

$$\mathbf{C_M = C_D / 中靶率} = 10 / 0.8$$

$$C_M = 12.5x$$

相反，较小的 Fold-80 改善即可显著提高效率。改善均一性会降低过度测序目标区域的覆盖深度，提高测序不足目标区域的覆盖深度。

为了检验中靶率和均一性的相对影响，我们模拟了 3,003 个具有不同均一性、平均覆盖深度和中靶率的正态分布[2]。在保持均一性恒定的前提下，提高中靶率（图 4A）后平均覆盖深度 (C_M) 值升高，从而增加了超过期望覆盖深度 (C_D) 的碱基比例。如前所述，在中靶率恒定的前提下，通过增加测序不足区域的覆盖率并减少过度测序区域的比例来改善 Fold-80 分值，可以提高测序数据的利用率（图 4B）。在这种情况下，虽然平均覆盖深度 (C_M) 值保持不变，但是超过期望覆盖深度 (C_D) 的碱基比例增加了。在这两幅图中，低于 C_D 的曲线间区域表示可用于分析的碱基数的差异。

图 4C 示出了中靶率、Fold-80 分值和平均覆盖度变化的综合影响。不同颜色的曲线分别代表不同的 Fold-80，曲线宽度代表中靶率在 80%（各曲线的下限）至 100%（上限）范围时覆盖的可用于分析的碱基百分比。在各曲线中，当 C_M 为 30 X 时，将中靶率从 80% 提高至 100%（基本消除了所有非靶向测序）能使可用于分析的碱基比例增加 1–2%。相比之下，将 Fold-80 从 1.7 降低至 1.4 能更显著地增加可用于分析的碱基比例，增加值为 5–6%。

这些数据表明，即使脱靶率可降至为零，对靶向 NGS 效率的提升来说，改善 Fold-80 分值（均一性）的效果仍比提高中靶率要大得多。

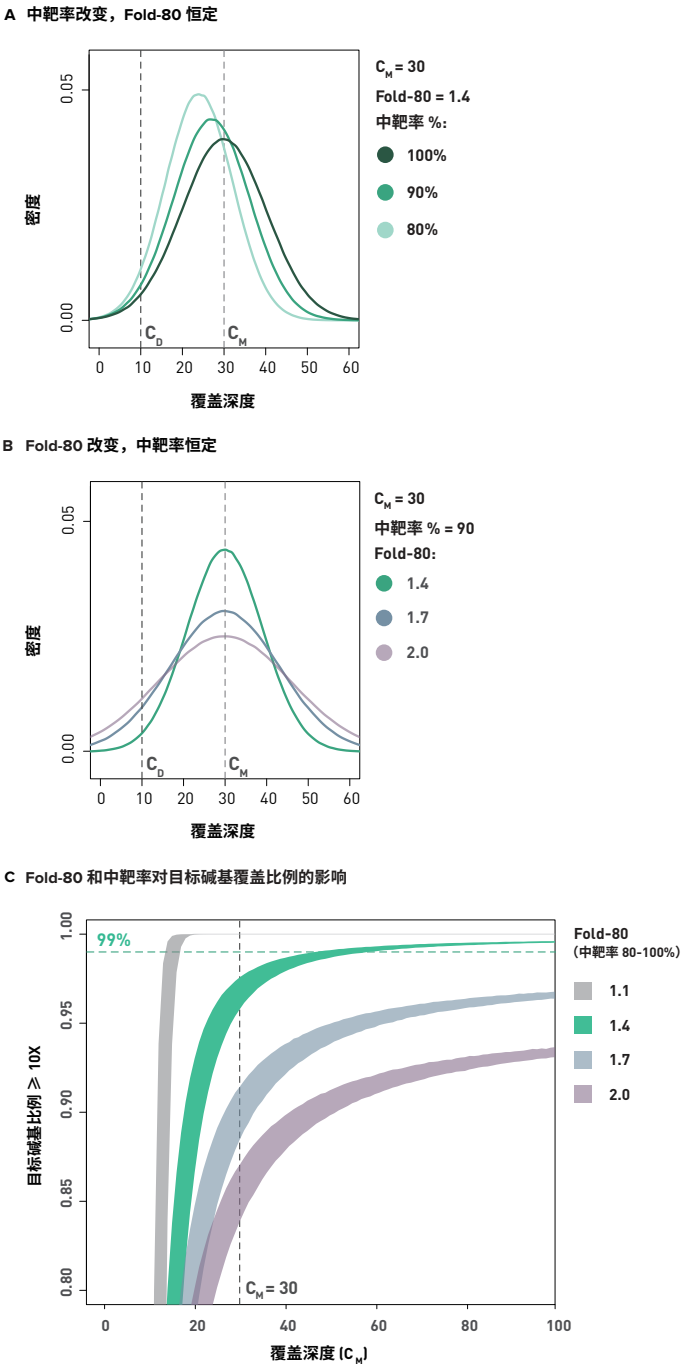

图 4. 均一性与中靶率对所需覆盖深度的影响。模拟结果假设期望覆盖深度 (C_D) 为 10X，覆盖深度为正态分布，平均覆盖深度 (C_M) 各不同，中靶率为 .8-1.0，Fold-80 为 1.1–2.0。**A.** 模拟 Fold-80 和 C_M（分别为 1.4 和 30）恒定时，覆盖深度分布随中靶率的变化。中靶率的提高增加了平均覆盖度，使分布向右移动。**B.** 模拟中靶率和 C_M（分别为 0.9 和 30）恒定时，覆盖深度的分布随 Fold-80 的变化。改善（降低）Fold-80 分值降低了过度测序目标区域的覆盖度，并提高了测序不足目标区域的覆盖度。**C.** 中靶率、Fold-80 分值和平均覆盖度变化时，10X 或更高的目标碱基覆盖比例。

[2] 采用正态分布是为了直观地说明中靶率和均一性的概念。虽然实际覆盖深度分布通常不遵循正态分布，但我们分析的一般结论可扩展到 NGS 中常见的分布（确切的数值可能不同）。

结论

靶向 NGS 中，均一性 (Fold-80) 和中靶率均为评估测序效率的重要指标。这两项指标大多为探针组合本身的固有特性，优化探针组合可以减少获得高可信度数据所需的测序量。

要选出最有效的目标富集体系，需要仔细权衡均一性的实际范围和提供的中靶率。虽然中靶率很重要，但本文研究表明：改善 Fold-80 分值（均一性）对靶向 NGS 的效率具有更显著的影响。

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.