分子条形码 — 常见问题解答

利用 HaloPlex^{HS} 检测低频等位基因变异

作者

Anniek De Witte[1]、Katie L. Zobeck[1]、Linus Forsmark[1]、Christian LeCocq[1]、Heather Tao[1]、Bahram Arezi[2] 和 Magnus Isaksson[1]

[1]安捷伦科技有限公司，美国加利福尼亚州圣克拉拉市

[2]安捷伦科技有限公司，美国加利福尼亚州拉霍亚

问答 — 前言

什么是 NGS?

新一代测序 (NGS)、大规模平行测序和高通量测序是描述可同时对数百万 DNA 小片段进行测序的创新 DNA 测序技术的相关术语。与标准 Sanger 测序法相比，这一现代化技术得到的结果将大幅增加被测序碱基对的数量。

NGS 已广泛应用于众多领域，其中最常用的是基因组 DNA 变异分析和 RNA 表达分析。这些分析应用可以扩展到整个基因组（WGS 或全基因组测序）和整个外显子组（WES 或全外显子组测序），还可专门测序特定区域和基因组合。

WGS 可捕获所有可能的突变，而 WES（尤其是靶向测序）的优势在于在维持原有经济效益与数据解析复杂程度的同时获得较高的目标区域覆盖率。极高的测序覆盖率对于发现出现比例较低的癌症突变至关重要 (1)。

如何在 NGS 前富集靶标区域?

在 NGS 前，有几种方法可以富集目标区域，最常用的两种方法是：1) 使用靶标特异性探针从测序文库中进行杂交捕获；2) 利用靶标特异性引物直接用样品 DNA 进行 PCR 扩增 (2, 3)。

使用杂交捕获方法时，将 DNA 片段在溶液中与基因组中靶标区域对应的序列特异性捕获探针进行杂交。典型的杂交捕获技术包括 Agilent SureSelect、NimbleGen SeqCap 和 Illumina TruSeq (4)。

在扩增子测序中，在 NGS 前使用一组选定的外显子引物通过 PCR 富集靶标基因。典型的扩增子捕获技术包括 Ion Torrent AmpliSeq、RainDance ThunderBolts 或 Illumina TruSeq 扩增子等仅基于 PCR 的方法，以及 Agilent HaloPlex^{HS} 等杂交和延伸方法 (5)。

Agilent Technologies

NGS 中潜在的错误来源有哪些?

NGS 和靶向序列捕获技术的进步使其可用于测定异质混合物中的低频突变。然而，PCR 和测序方法的系统误差使 NGS 的进一步发展受到了限制。文库制备、靶向序列捕获和测序均采用 DNA 聚合酶以及扩增步骤。这些过程会引入偏差，包括重复、相应的不均匀扩增以及因聚合酶误差导致的假象，这种误差会引入原始样品中本不存在的序列变化 (1)。以最常使用的 Illumina 测序仪来说，误差率在 ~0.05% 到 ~1% 之间，具体取决于读取长度、所用的碱基识别算法和测定的变异类型 (6)。使用分子条形码可使结果更加可靠，对突变率和误差率在同一数量级的临床样品更有帮助。

问答 — 分子条形码

什么是分子条形码?

分子条形码方法是为同一样品的每个原始 DNA 片段连接上一段独一无二的序列编码。分子条形码通常设计为完全随机的核苷酸链（如 NNNNNNN）、部分简并核苷酸链（如 NNNRNYN）或指定核苷酸链（模板分子有限时）。分子条形码可用于计量高覆盖率 NGS 数据中的测序误差和 PCR 误差。

分子条形码和样品条形码有何不同?

样品条形码和分子条形码通常同时存在于同一测序读出序列中。分子条形码可用来减少假阳性结果，而样品条形码也称为标签接头，常用于多数当前的 NGS 流程中，允许在测序前对样品进行混合。样品条形码扮演着标识符或标签的作用，以确定读出序列来源于哪个样品，因此可以在一次运行中同时检测多个样品。样品条形码通常是在 DNA 测序前加入的特异性短序列。这些条形码会与未知样品 DNA 一同被测序。测序结束后，将读出序列按照条形码分类和重组（多重性分解）。

为什么要使用分子条形码?

分子条形码方法是将 DNA 片段接上一段独特的序列条形码。具有不同分子条形码的读出序列代表不同的原始 DNA 分子，而具有相同条形码的读出序列则是相同原始分子经 PCR 复制的结果。分子条形码无法阻止 PCR 重复的产生，但可帮助用户追踪这些重复并在下游分析中将其去除。此外，使用分子条形码还能将 PCR 假阳性与原始分子中的真正变异区分开来，从而在变异等位基因频率 (VAF) 极低时进行变异检测。Peng 等人 (1) 开发了一种 NGS 靶向序列捕获流程，可将分子条形码整合到高度多重 PCR 扩增子测序中。他们比较了使用和未使用分子条形码得出的数据，结果表明，在高灵敏度设置下，使用分子条形码可显著减少假阳性结果。

分子条形码能否使杂交捕获方法受益?

尽管杂交捕获前的随机剪切可获得多种可用作每个起始分子特异标识符的随机片段末端，但分子条形码在杂交型靶向序列捕获中仍十分有用。首先，在极高的测序读取深度（10000-50000X 原始读出序列）下，可以发现随机剪切的片段不完全是随机的 (7)。其次，由于成本、通量和产量方面的考量，NGS 将逐渐通过杂交型靶向序列捕获前的酶解片段化取代随机剪切。这种酶切会大大减少产生的随机片段末端，影响追踪不同起始分子以及去除 PCR 重复和相关扩增假阳性的能力。

请介绍一下分子条形码技术的历史

自 2007 年开始，陆续有报道指出可用分子条形码标记单个模板来解决 NGS 中的 PCR 重复和扩增误差问题。过去，分子条形码或分子标记被称为唯一标识符 (UID)、唯一分子标识符 (UMI)、引物 ID 或双条形码等多个名称。

可采用多种不同的技术将分子条形码添加到模板上。其中一种方法是在基因组测序的文库构建步骤中将分子条形码整合到测序接头上。这种“双测序”方法可以有效减少对 DNA 双链中任意一条链单独标记和测序时造成的误差 (8)。另一种方法是将条形码结合到分子倒置探针上以进行靶标体细胞突变检测。通过 smMIP（单分子倒置探针）将单分子标记与多重靶向捕获结合，从而对低频变异等位基因进行有效的高灵敏度检测 (9)。此外，还可将分子条形码结合到靶标特异性 PCR 引物上，这种引物通常为一小段随机碱基。

HaloPlex^{HS} 分子条形码是如何创建的?

HaloPlex^{HS} 是基于扩增子的测序与靶向序列捕获测序的结合。这种方法主要利用限制性内切酶识别、杂交和 DNA 连接的特异性从待测序靶标区域捕获分子。为了能够从通过 HaloPlex^{HS} 制备的文库中识别出重复读出序列，我们在引入的引物载体上添加了一个分子条形码（图 1）。HaloPlex^{HS} 工作

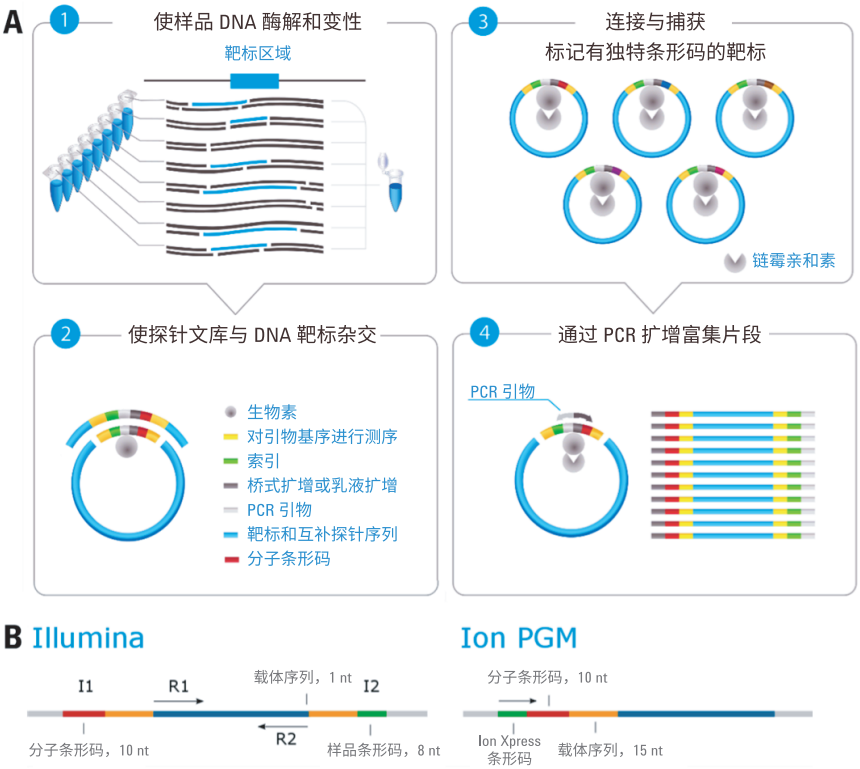

图 1. A) HaloPlex^{HS} 工作流程概览。B) HaloPlex^{HS} 文库示意图。对于在 Illumina 平台上测序的文库，分子条形码将作为 Illumina 双索引系统的一部分添加到文库中。对于在 Ion PGM 平台上测序的文库，将分子条形码结合到读出序列的起始部分。这两种文库均与标准测序方法兼容

流程会对每个片段标记一个条形码。当 HaloPlex^{HS} 同时捕获双链（即使用 SureDesign 中的 FFPE 设计选项）时，双链分子会被分别标记。分子条形码含有的十个简并碱基可生成一百万个以上的独特序列用于分子标记。这种条形码可将真正的变异与 PCR 或测序误差加以区别。

需要多高的覆盖率才能达到所需的灵敏度?

结合高保真 DNA 聚合酶和分子条形码的功能区分真正的突变与误差，理论上可以检测变异等位基因频率为 1% 甚至更低的突变。实际上，样本量不足和采样偏差问题会影响分子条形码和深度测序的应用 (10)。为了准确评估模板库中的突变情况，库中的所有模板都应均一地标记，标记过程不得有误，且分子标签在后续扩增中不会受到突变的影响。

Jon Armstrong (11) 在其在线发表的论文中阐述过构建文库时的 DNA 数量（或拷贝数）、产生的测序覆盖率和最小等位基因检测的最低理论阈值之间的关系。如果需要 4 个读出序列以显示替换等位基因并以 0.1% 的灵敏度可靠地测定变异，则应在删除重复后达到 4000X 的深度覆盖（4 个读出序列/0.001）。从理论上而言，我们需要至少 4000 个单倍体基因拷贝、2000 个细胞或 12 ng DNA。换言之，无论产生多少测序读出序列，去除重复的覆盖度都不会超过用于测序的基因组拷贝数。这一数字会随过程噪音、变异和 PCR 扩增而增大，并在分子化流程中减小。

具有分子条形码的 HaloPlex^{HS} 技术是否已应用于临床研究中？

最近一篇论文 (12) 中描述的示例采用分子条形码结合新一代测序技术检测癌症中的体细胞嵌合。这篇文章对患有多种 DICER1 综合征相关原发性肿瘤的儿童进行了研究。在不同部位的多次肿瘤活检中发现，不同研究对象均携带特异性 DICER1 “热点” 核糖核酸酶 IIIb 突变。然而，使用传统技术未在他们的种系 DNA 中检测到突变，因此怀疑是体细胞嵌合。在 Illumina HiSeq 平台上以高覆盖率进行新一代测序前，采用包含分子条形码的 HaloPlex^{HS} panel。为确认/否定嵌合假设，该研究团队采用分子条形码技术测定肿瘤和非肿瘤样品中之前已识别突变的相对丰度。研究团队通过这一方法证实了他们的假设，确认这些儿童体内 DICER1 相关的罕见肿瘤由 DICER1 核糖核酸酶 IIIb 嵌合引起。这篇文章指出，具有分子条形码的 HaloPlex^{HS} 靶向序列捕获系统可提供检测最低 0.24% 的突变等位基因频率所需的灵敏度。

问答 — 分析方法

如何分析分子条形码？

分子条形码分析通常包括五个步骤（图 2）。第一，在文库制备阶段利用分子条形码对单个 DNA 分子进行标记。读出序列对齐时忽略分子条形码。随后，将与相同基因组坐标对齐的读出序列对根据分子条形码进行分组。最后，将分子条形码对应的碱基序列合并为每个分子一个读出序列，并去除 PCR 重复。

能否用 Agilent SureCall 软件分析 HaloPlex^{HS} 分子条形码？

Agilent SureCall 软件针对包含分子条形码的 HaloPlex^{HS} 获得的 NGS 数据分析进行了优化，可识别重复读出序列，因此大大提高了低等位基因频率下的碱基识别准确度。Agilent SureCall 软件可从安捷伦网站（http://www.genomics.agilent.com/article.jsp?pageId=3341）上免费下载

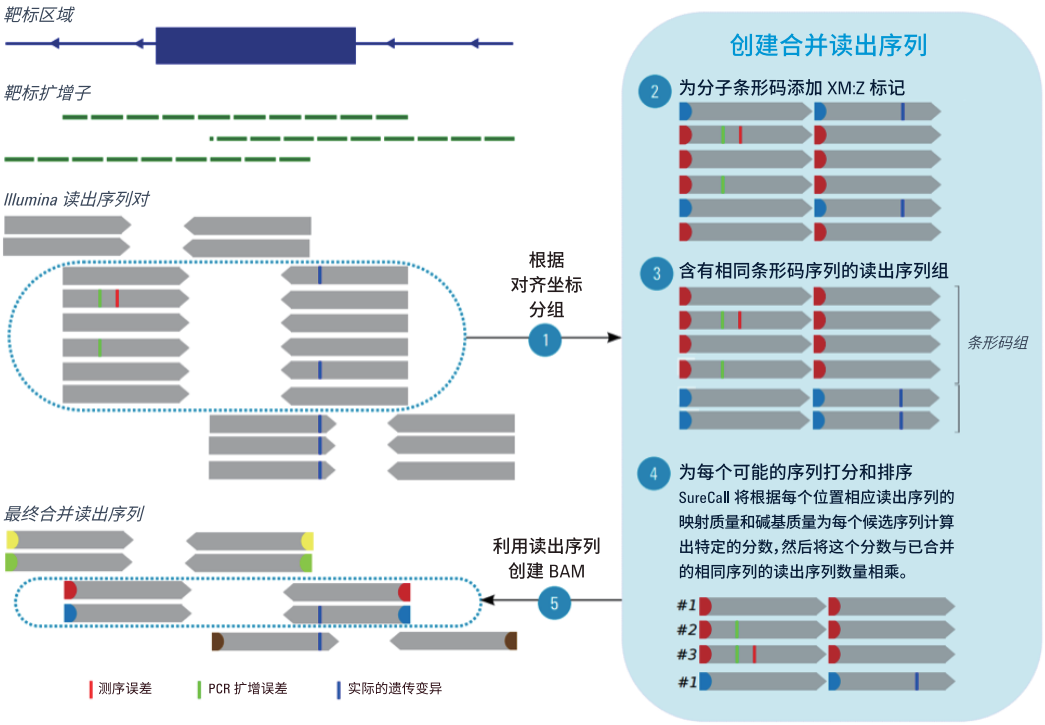

图 2. 采用分子条形码序列信息生成合并读出序列的示意图

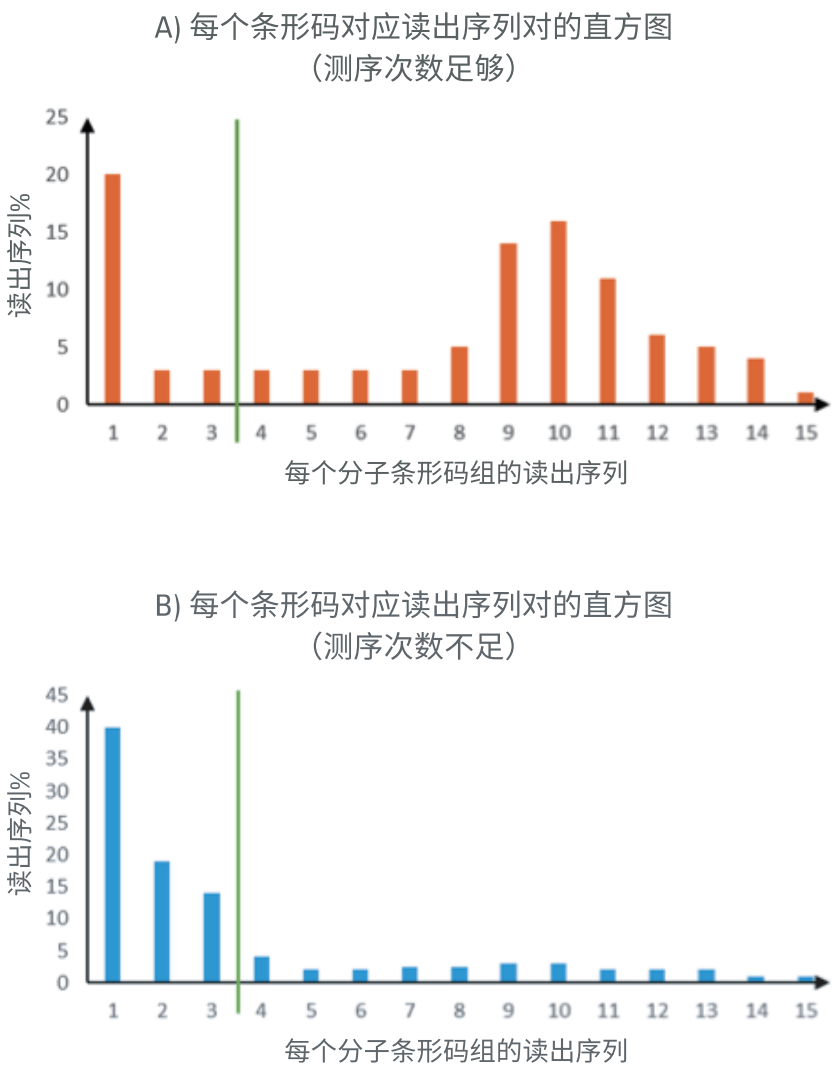

图 3. 每个条形码对应读出序列对数量的理论直方图

Agilent SureCall 软件如何处理每个条形码仅有一个读出序列对的读出序列对？

Agilent SureCall 软件会依据测序次数决定是舍弃还是保留每个条形码仅有一个读出序列对的读出序列对。图 3 是每个条形码对应读出序列对数量的理论直方图。使用默认设置时，为进行误差校正，需要平均 4 个以上的读出序列对（绿色竖线）。图 3A 表明，进行的测序次数足够多时，每个条形码仅有一个读出序列支持的读出序列对会被舍弃，而具有 4 个以上读出序列对的读出序列可用于误差校正。图 3B 显示测序次数不足时获得的结果。虽然舍弃数据中大部分为无效数据，但舍弃仅有一个条形码的读取序列对仍会使可分析的数据明显减少（30%-40% 以上）。

能否用其他软件分析 HaloPlex^{HS} 分子条形码？

如果不想安装 Agilent SureCall 软件，用户可以使用安捷伦基因组新一代测序工具包 (AGeNT)。AGeNT 是基于 Java 的软件模块，专门处理对 HaloPlex^{HS} 文库测序后生成的靶向高通量测序数据的读出序列。AGeNT 可从末端切除低质量碱基、去除接头序列并掩盖酶处理印记。在对齐前以这种方式正确制备读出序列，可有效提高对齐效率并降低假阳性变异识别的比例。AGeNT 还可处理 HaloPlex^{HS} Illumina 数据在对齐后的分子条形码 (MBC) 信息，其中将读出序列对标记在 BAM/SAM 文件（包括由索引 2 FASTQ 文件获得的 MBC 序列）中，还可从该 BAM/SAM 文件中标记或去除 MBC 重复。AGeNT 可从安捷伦网站（http://www.genomics.agilent.com/en/NGS-Data-Analysis-Software/AGeNT/?cid=AG-PT-154&tabId=prod2570007）上免费下载

参考文献

1. Peng *et al*. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC Genomics.* 2015 Aug 7;16:589.
2. Gnirke *et al*. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol.* 2009 Feb;27(2):182-9.
3. Taylor *et al*. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.* 2007 Sep 15;67(18):8511-8.
4. Bodi *et al*. Comparison of commercially available target enrichment methods for next generation sequencing. *J. Biomol. Tech.* 2013 Jul;24(2):73-86.
5. Chang *et al*. Clinical application of amplicon-based next generation sequencing in cancer. *Cancer Genet.* 2013 Dec;206(12):413-9.
6. Kinde *et al*. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A.* 2011 Jun 7;108(23):9530-5.
7. Poptsova *et al*. Non-random DNA fragmentation in next-generation sequencing. *Sci Rep.* 2014; 4: 4532.
8. Schmitt *et al*. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A.* 2012 Sep 4;109(36):14508-13.
9. Hiatt *et al*. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* 2013 May;23(5):843-54.
10. Kou *et al*. Benefits and challenges with applying unique molecular identifiers in next generation sequencing to detect low frequency mutations. *PLoS One.* 2016 Jan 11;11(1):e0146638.
11. https://cofactorgenomics.com/heterogenous-dna-sequencing-lower-limits-minor-allele-frequency-sensitivity/
12. de Kock *et al*. High-sensitivity sequencing reveals multi-organ somatic mosaicism causing DICER1 syndrome. *J. Med. Genet.* 2016; 53:43-52.

查找当地的安捷伦客户中心：

www.agilent.com/chem/contactus-cn

免费专线：

800-820-3278，400-820-3278（手机用户）

联系我们：

LSCA-China_800@agilent.com

在线询价：

www.agilent.com/chem/erfq-cn

安捷伦科技大学：

http://www.lsca-china.com.cn/agilent

浏览和订阅 Access Agilent 电子期刊：

www.agilent.com/chem/accessagilent-cn

www.agilent.com/genomics/HaloPlexHS

PR7000-0254

2016 年 9 月 19 日，中国出版
5991-7421CHCN

Agilent Technologies