技术白皮书

平衡全基因组测序的样本覆盖度

Illumina DNA PCR-Free Prep 的标签校正策略

- 提高混合文库的最小覆盖度
- 在文库混合时调整体积，降低标签性能的差异性
- 为高通量分析减少浪费、降低测序成本

illumina®

提高混合样本的最小覆盖度

Illumina DNA PCR-Free Prep, Tagmentation（Illumina DNA PCR-Free）为各种全基因组测序（WGS）应用提供了优化的文库制备解决方案。Illumina DNA PCR-Free 磁珠固化转座酶化学技术，可均匀覆盖整个基因组，并支持按体积进行文库混合的简单操作[1,2]。使用 Illumina DNA PCR-Free 和 NovaSeq™ 6000 测序系统，高通量实验室可以对 WGS 文库进行多重测序，大幅提高效率。平衡多重 WGS 样本的测序产量，使用户能够在每个流动槽中运行更多样本，同时实现可靠变异检出所需的最小覆盖度。

有几个因素会影响每个样本的覆盖深度，包括总测序产量、每个流动槽的样本数量以及样本产量的差异性。要想增加每次运行中达到所需测序覆盖度的样本数量，用户可以增加所有样本的平均覆盖度或减少样本间的差异（图 1）。起始 DNA 质量、移液技术差异性和标签性能都会影响每个样本的测序产量。本技术白皮书重点介绍了减少因标签性能引起的差异的策略。

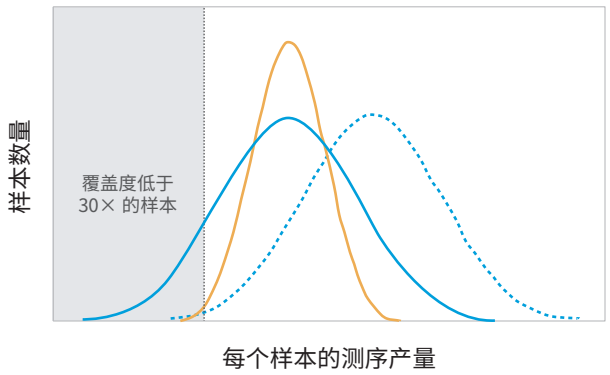

图 1: 提高每个样本的最低测序产量——随着测序产量增加（蓝色虚线）或差异性降低（橙色），更多的样本将实现至少 30× 的覆盖度。

标签性能影响样本覆盖度

IDT for Illumina DNA/RNA Unique Dual (UD) Indexes, Tagmentation 提供了 4 组各 96 个标签对（Set A、B、C 和 D）*，总共 384 个标签对。当与 Illumina DNA PCR-Free† 和 NovaSeq 6000 测序系统‡ 一起使用时，某些标签对始终会生成较高或较低的测序产量。表现不足的标签对会导致这些样本的覆盖度低于预期，而表现过度的标签对会减少其他样本的 read 比例。

为了证实标签序列导致的样本覆盖度差异性，我们采用了 Illumina DNA PCR-Free 文库制备方案的改进版本。使用人细胞系基因组 DNA（Coriell 医学研究所，NA12878）作为起始材料，并且所有样本批量制备，以分别减少与起始 DNA 质量和移液误差相关的噪声干扰。在加标签步骤中，将样本等分至 96 孔板中进行加标签，然后再次混合以进行片段大小选择纯化§。多名操作人员使用手动移液器，对每组 UD Index 进行了共三到五次重复实验。使用 Xp 工作流程在 NovaSeq S4 流动槽的单个泳道上对每个 96 样本组进行测序，获得标签表达。

展示了 UD Index Set A 和 B（表 1）以及 Set C 和 D*（表 2）的归一化标签性能†‡。关于此数据集：

- 平均而言，每组 96 个标签对中有 32 个与性能中位数的偏差超过 15%（蓝色单元格）

* IDT for Illumina DNA/RNA UD Indexes, Tagmentation 的 Set C 和 D 于 2021 年推出。

† Illumina DNA PCR-Free 采用了独特的加标签化学技术；标签性能数据不适用于其他文库制备试剂盒。

‡ 某些标签差异性是由于簇生成造成的；标签性能数据针对 NovaSeq 6000 测序系统。

§ 使用固相可逆化固定（SPRI）磁珠纯化进行片段大小选择时，涉及粘性试剂的多个移液步骤可能会引入额外误差。加标签后立即混合样本，而不是在片段大小选择后再进行混合可以减少内部差异（数据未显示）。

表 1: Illumina DNA PCR-Free 和 UD Index Set A 和 B 的 96 孔板布局中的归一化标签性能

Set A	1	2	3	4	5	6	7	8	9	10	11	12
A	1.19	1.18	1.07	1.02	0.84	1.20	1.00	0.90	0.98	1.15	1.03	1.03
B	0.82	1.09	0.74	1.02	1.09	1.03	0.77	0.73	1.11	1.09	1.08	0.96
C	0.95	0.82	0.81	0.83	0.78	1.09	0.89	0.99	0.95	1.07	0.96	0.70
D	1.07	0.76	0.92	1.42	1.08	0.96	1.01	1.18	1.10	0.86	1.05	1.28
E	0.95	0.89	1.01	0.85	1.03	0.87	1.00	0.88	1.42	0.88	0.92	1.01
F	1.21	1.07	1.24	1.04	0.91	0.70	1.32	0.97	1.22	1.36	0.95	1.09
G	1.07	0.98	1.21	0.86	0.84	1.20	1.05	1.27	0.95	0.94	1.08	0.98
H	1.16	0.93	1.14	0.80	0.97	1.09	0.91	0.94	1.15	0.73	0.96	1.10

Set B	1	2	3	4	5	6	7	8	9	10	11	12
A	0.89	1.16	0.94	0.99	1.15	0.98	0.96	1.06	0.81	0.95	0.98	0.85
B	0.78	1.04	0.94	1.12	1.06	1.15	1.04	1.39	1.24	1.13	1.17	1.02
C	0.96	0.85	1.11	1.00	1.22	1.26	0.80	1.43	1.00	0.82	1.13	0.90
D	0.98	0.63	0.95	0.92	1.36	1.05	0.84	0.86	0.75	1.19	0.83	1.00
E	1.00	0.83	0.95	1.00	0.87	1.01	1.28	0.94	0.95	1.19	0.97	1.00
F	0.67	1.27	1.12	0.90	0.76	1.13	1.00	1.03	1.35	1.08	1.01	0.91
G	1.04	1.18	1.19	0.99	0.60	1.22	0.98	0.99	1.18	0.80	1.02	1.00
H	0.94	0.85	0.75	0.94	0.92	1.02	1.09	0.95	0.94	0.94	1.50	0.97

使用 Illumina DNA PCR-Free Prep 手动制备文库，重复数 n = 3（Set A）或 n = 5（Set B），并在 NovaSeq 6000 系统上测序。蓝色突出显示的单元格表示与中位数相差 15% 以上的标签表达。表现不足、与中位数相差至少 30% 以上的标签对（Set A 中 2 个；Set B 中 3 个）以深蓝色突出显示。在 illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-dna-prep-pcr-free-index-correction-tech-note-m-gl-00005/index-correction-illumina-dna-prep-pcr-free.csv 下载包含标签表达数据和计算出的校正因子的逗号分隔值（*.csv）文件。

- 每张板中通常有少数标签对（深蓝色单元格）表现不佳，比性能中位数低至少 30%，使这些标签处于无法覆盖靶标的高风险中
- 一些单独的板列没有标签对，或只有一个标签对，与中位数的偏差大于 15%，表现出比整个板更高的性能

在重复实验中标签性能一致（R^2 = 0.54-0.80）。例如，观察 UD Index Set B（图 2）96 孔板第 2 列的 4 次重复实验，发现了一致的标签过度表达和表达不足现象。

这表明异常值不是单纯的噪声（noise），而是所用标签对的基本属性。一致的标签性能差异指明了与典型的标签性能相关的“标签校正”策略。原则上来说，在测序前混合时调整某些文库的体积，可以弥补标签表达之间的差异。性能出色的标签对应能始终确保较低的差异性，而始终偏离中位数值的标签对则是标签校正的候选对象。

表 2: Illumina DNA PCR-Free 和 UD Index Set C 和 D 的 96 孔板布局中的归一化标签性能

Set C	1	2	3	4	5	6	7	8	9	10	11	12
A	0.98	0.84	1.12	0.92	0.88	1.40	1.04	0.78	0.89	0.89	1.19	0.85
B	0.70	0.97	0.88	1.13	0.80	1.07	1.21	1.03	1.05	0.86	0.97	1.00
C	0.67	0.96	1.17	1.19	1.05	1.09	1.14	1.07	1.20	1.21	0.86	1.05
D	0.80	1.00	1.23	1.02	1.26	0.89	1.21	0.85	0.60	1.01	1.05	0.92
E	1.15	1.07	0.95	0.85	0.84	0.98	1.16	1.10	0.78	0.88	1.04	1.06
F	1.02	1.07	1.26	1.13	0.47	1.09	1.16	0.91	0.95	1.11	0.80	1.22
G	0.77	0.96	1.17	0.84	1.06	1.11	0.96	0.90	1.14	1.13	1.24	1.09
H	0.78	0.91	1.00	1.02	0.91	1.03	0.98	0.98	0.93	0.99	0.89	0.95

Set D	1	2	3	4	5	6	7	8	9	10	11	12
A	0.91	0.74	1.01	0.98	0.96	0.73	1.03	1.04	0.97	0.77	0.82	0.84
B	1.22	0.95	1.03	1.19	0.84	0.83	1.02	0.80	0.79	1.05	0.84	0.76
C	1.19	1.14	0.99	0.90	0.93	0.93	1.01	0.80	1.26	1.11	1.02	0.96
D	0.86	0.91	1.16	1.25	1.02	1.12	0.95	1.06	0.99	1.09	1.23	0.91
E	0.69	0.88	1.13	1.03	1.43	0.90	1.23	1.12	1.05	1.30	0.88	1.18
F	0.88	1.13	1.29	0.91	0.82	1.15	0.93	0.93	1.18	1.02	1.24	1.20
G	1.18	0.99	1.07	1.02	1.16	0.93	0.98	0.91	1.03	0.77	1.03	1.03
H	0.96	1.12	1.02	1.23	0.99	0.90	0.95	0.88	1.20	0.80	0.88	0.83

使用 Illumina DNA PCR-Free Prep 手动制备文库，重复数 n = 4（Set C）或 n = 3（Set D），并在 NovaSeq 6000 系统上测序。蓝色突出显示的单元格表示与中位数相差 15% 以上的标签表达。表现不足、与中位数相差至少 30% 以上的标签对（Set C 中 4 个；Set D 中 1 个）以深蓝色突出显示。在 www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-dna-prep-pcr-free-index-correction-tech-note-m-gl-00005/index-correction-illumina-dna-prep-pcr-free.csv 下载包含标签表达数据和计算出的校正因子的逗号分隔值（*.csv）文件。

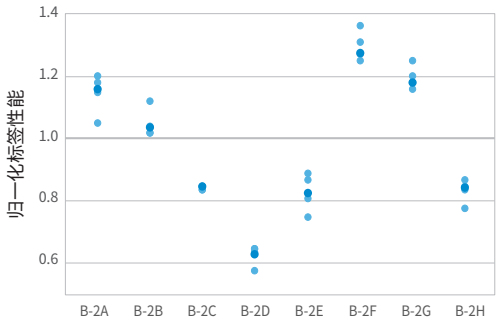

图 2: 始终表现不佳或表现过度的标签对——UD Index Set B 第 2 个板列的 8 个标签对的 4 次重复实验（浅蓝色）和平均（深蓝色）归一化性能。有一个标签对 (B-2D) 的表现始终比中位数低 30%，有一个标签对 (B-2F) 的表现始终比中位数高 20%。

降低标签性能差异

对表 1 和表 2 中所示每个标签对的标签表达值中位数取倒数，获得标签校正因子。在混合时乘以这些因子以调整每个样本的体积——有效地减少加入的过度表达标签的量和增加加入的表达不足标签的量——可以重新平衡每个样本的测序 read 数。

使用中位数的 10% 作为标签校正的阈值

为了测试该策略，在所有 4 组 UD Index 中，对偏离中位数至少 10% 的标签对均应用了标签校正因子（图 3）。按未校正时的平均标签表达对 384 个标签对进行排序后，数据显示出一种线性模式，即 4 次重复实验具有良好的相关性（R^2 = 0.75）和较大的差异性（变异系数 CV = 17%）。标签校正后，标签表达的差异降低（CV = 11%），校正后的值不再与校正前的性能相关（R^2 = 0.006）。

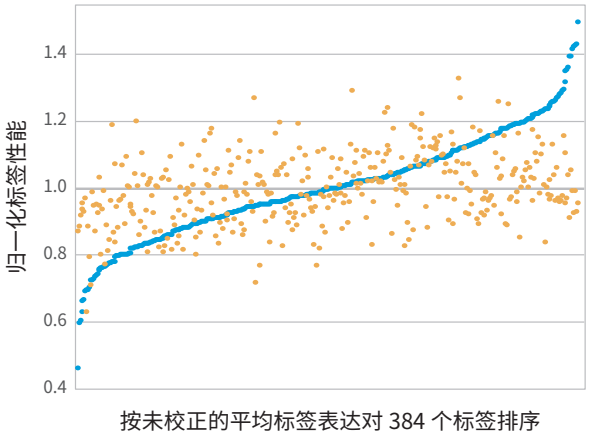

图 3: 基于体积的标签校正消除了标签相关的性能差异——对于未校正的实验，384 个标签对按测序产量从左（最低）到右（最高）排序。未校正标签（蓝色）的变异系数（CV）为 17%。校正后的标签（橙色）显示变异系数（CV）得到改善，为 11%。

使用中位数的 15% 作为标签校正的阈值

随后，一名操作人员使用来自 UD Index Set B 的所有 96 个标签对制备文库（图 4A）。为了减少手动移液疲劳造成的噪声，标签校正的阈值从 10% 提高至 15%。加标签后，对文库进行等体积混合，以进行片段大小选择（蓝色），或对偏离中位数至少 15% 的所有标签（橙色）进行校正。标签校正后，整体的变异系数（CV）从 19%（未校正）降低至 10%（校正）。再次注意 UD Index Set B 的第二个板列（图 4B），始终表现不佳（B-2D）或表现过度（B-2F）的标签对变得更为典型。

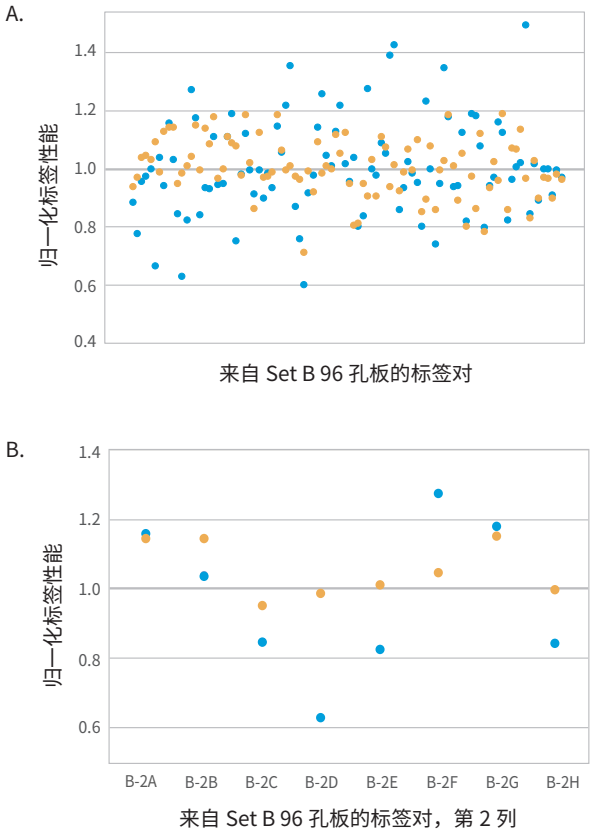

图 4: 标签校正降低了标签性能差异性——（A）来自 UD Index Set B 的 96 个标签对。（B）来自 UD Index Set B 第 2 列的 8 个标签对。未校正标签（蓝色）的变异系数（CV）为 19%。校正后的标签（橙色）显示变异系数（CV）得到改善，为 10%。偏离中位数超过 20% 的标签对（B-2D 和 B-2F）在校正后，恢复到标签性能中位数。

最后一个实验重点关注来自 UD Index Set B 的三个板列（7、8、9），并校正了 11 个通常偏离中位数至少 15% 的标签对。在 NovaSeq S4 流动槽上运行每种条件（校正或未校正）下的所有 24 个样本（图 5）。正如预期，标签表达的变异系数（CV）从 18%（未校正）下降至 7%（校正）。尽管样本之间的平均覆盖度保持一致，但校正后样本的最小覆盖度更高（数据未显示）。

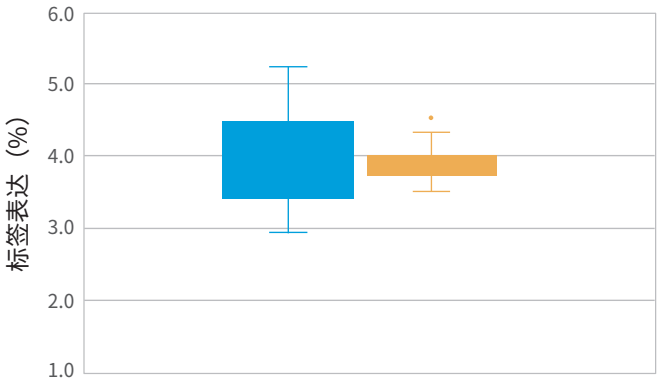

图 5: 标签校正降低了同一流动槽上测序样本之间的差异性——使用来自 UD Index Set B 的 24 个标签对（孔板第 7、8、9 列）的文库的标签性能。未校正标签（蓝色）的变异系数（CV）为 18%。校正后的标签（橙色）显示变异系数（CV）得到改善，为 7%。

为高通量分析减少浪费、降低测序成本

选择性能最佳的标签对

这些 Illumina DNA PCR-Free 标签性能数据表明，哪些标签有望提供更好的覆盖度，以及哪些列的标签性能差异较小（表 3，以橙色突出显示）。在 96 孔样本板中，标签表达的典型变异系数（CV）为 15%-20%。只需使用部分板的客户可以选择具有最低变异系数（CV）的列，以提高性能而无需校正。

采用标签校正因子减少差异性

这些结果还证明，在文库混合时调整文库体积可以校正标签性能不佳的问题。在来自该数据集的标签校正因子的指导下进行基于体积的样本混合，足以可靠地降低差异。对于手动操作人员而言，一个容易出现错误的地方是需要校正很多标签对。手动校正标签是一个繁琐的过程，可能会因移液误差而导致噪声增加。我们观察到，使用更高的标签校正阈值可以获得更好的结果（即，对性能超过中位数 15% 以上以及超过中位数 10% 以上的标签进行校正）。即使只校正少数表现非常差的标签对，也有望“挽救”一些样本，使其不会因为覆盖率低于预期而导

表 3: 按列和板表示的 IDT for Illumina DNA/RNA UD Indexes 标签性能差异性（CV）

列 [a]	1	2	3	4	5	6	7	8	9	10	11	12	板 [b]
Set A	12.9%	14.7%	18.1%	20.7%	12.5%	16.7%	16.1%	17.5%	14.4%	19.8%	6.4%	16.3%	**15.5%**
Set B	13.8%	22.6%	14.1%	7.0%	25.2%	9.4%	14.9%	19.7%	20.6%	15.4%	18.6%	6.7%	**16.6%**
Set C	19.9%	7.9%	12.6%	13.3%	25.7%	13.4%	9.1%	11.8%	20.5%	12.8%	15.5%	11.2%	**15.7%**
Set D	19.3%	14.4%	9.3%	13.1%	19.3%	14.8%	9.5%	12.7%	14.2%	19.4%	17.1%	16.8%	**15.1%**

a. 根据表 1 和表 2 中的平均标签性能数据计算得出的 CV。使用 Illumina DNA PCR-Free Prep 手动制备文库，重复 3-5 次，并在 NovaSeq 6000 系统上测序。突出显示了标签表达 CV 低于 10%（深橙色）或低于 15%（浅橙色）的列。
b. 列的 CV 范围为 6.4%-25.7%，而 96 孔板的 CV 范围仅为 15.1%-16.6%。

致数据无法被使用。将标签校正视为一个动态过程，随着您收集的数据不断增多而进行迭代调整。这些数据是下一步研究的基础。如果要减少背景噪声并获得所有优势，可能需要使用自动化操作流程。

总结

对于使用 Illumina DNA PCR-Free 和 NovaSeq 6000 测序系统的实验，标签性能是样本覆盖度差异的关键影响因素。一些标签序列会始终获得较高或较低的产量。通过在混合时调整体积，您可以减少同一流动槽中测序样本之间测序产量的差异。虽然这不会增加总体测序产量，但会增加最小覆盖度，减少低于最低覆盖度阈值的样本数量。进行群体基因组学研究和其他高通量人类 WGS 应用的实验室可以把所提供的数据作为一个起点，以优化样本数据量并减少浪费。

了解更多

有关 Illumina DNA PCR-Free 的信息，请访问 illumina.com/products/by-type/sequencing-kits/library-prep-kits/dna-pcr-free-prep.html

参考文献

1. Illumina (2020). Illumina DNA PCR-Free Prep, Tagmentation. Accessed February 17, 2021.
2. Bruinsma S, Burgess J, Schlingman D, et al. Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics*. 2018;19(1):722. doi:10.1186/s12864-018-5096-9.

illumina®

Illumina中国
上海办公室 · 电话 （021）6032-1066 · 传真 （021）6090-6279
北京办公室 · 电话 （010）8455-4866 · 传真 （010）8455-4855
技术支持热线 400-066-5835 · chinasupport@illumina.com · www.illumina.com.cn

M-GL-00005

因美纳 因美纳讲堂

illumina®