技术说明

# Illumina Complete Long Read Prep with Enrichment, Human 的定制 panel 设计

## 针对人类基因组的高度灵活的靶向长读长富集方案

illumina®

仅供研究使用，不得用于诊断。

## 简介

在进行人类全基因组测序（WGS）时，部分基因区域仅使用短读长可能难以绘制。长读长测序可以作为标准短读长 WGS 数据的补充，帮助解析这些具有挑战性的区域。Illumina Complete Long Reads 技术采用标准的新一代测序（NGS）工作流程，通过单一分析流程在因美纳基因测序仪上生成连续的长读长序列（图 1）[1–3]。Illumina Complete Long Read Prep with Enrichment, Human 提供了一种靶向方法，可实现更经济高效的长读长测序。* Illumina Complete Long Read 富集化学技术在目标和探针设计方面具有高度灵活性，有助于解析难以绘制的区域或通过定相测序提供更多信息。

## 专为长读长设计的富集探针 Panel

与常用于捕获短片段（约 200-500 bp）的方法相比，Illumina Complete Long Read Prep with Enrichment, Human 采用了不同的探针设计策略来捕获更长的片段（约 7-10 kb）。

* 需要来自同一样本的≥ 30× 标准短读长 WGS 数据用于分析。可以使用之前运行的样本中的 FASTQ 文件。

Illumina DesignStudio™ 软件是一款用户友好的免费工具，可用于设计富集探针 panel。DesignStudio 算法考虑了 GC 含量、目标特异性和探针间隔，比如目标区域中的探针数量。120mer 短读长探针富集 panel 的标准间隔为 250-350 bp 探针窗口。对于长读长富集 panel 设计，在多个长度下测试了探针间隔，发现 1 kb 窗口是实现经济高效捕获的理想选择。

杂交富集的有效性高度依赖于探针的特异性。靶向富集的百分比直接影响实现目标覆盖深度所需的测序量。对于重复区域更难实现高特异性。然而，使用更大的探针窗口可以更灵活地排除性能较差的探针，避免重复区域（窗口大小可达 1 kb），并以更少的探针保持富集效率（图 2）。DesignStudio 算法可以根据这些考虑因素来推荐探针位置。第三方 panel 应使用类似的指南以获得出色性能和成本效益。此外，标准富集探针间隔也完全兼容。

### 灵活的探针设计和靶向策略

Illumina Complete Long Read Prep with Enrichment, Human 可以根据研究目标非常灵活地选择和设计定制探针 panel。单个目标区域的覆盖范围可从单个碱基到数百 kb。整个定制 panel 可从小至 2.5 Mb 到> 95 Mb。对于已知短读长

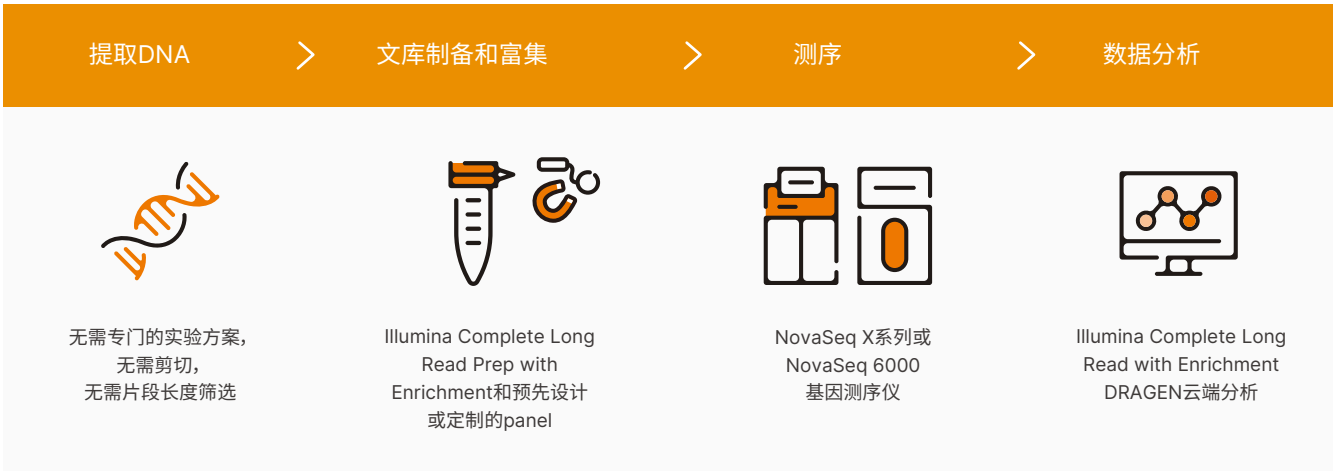


图 1： **集成工作流程的一部分**——使用可扩展、经过优化的文库制备与富集方案、成熟的因美纳测序化学技术和 DRAGEN 二级分析，获取经济有效的靶向长读长 WGS 数据。需要来自同一样本的≥ 30× 标准短读长 WGS 数据用于分析。可以使用之前运行的样本中的 FASTQ 文件。

测序时绘制率较低的特定区域，研究人员可以使用靶向长读长来提高覆盖度。也可以借助靶向长读长覆盖整个基因，乃至长片段多基因区域，以实现变体定相和单倍型检出。

DesignStudio 工具中提供了几种预设计 panel（表 1）。这些 panel 靶向许多不同的区域，包括复杂区域医学相关基因（CMRG）[4]、药物遗传学（PGx）试验检测通常靶向的基因[5–7]、美国医学遗传学和基因组学学会（ACMG）二级研究结果列表（ACMG SF v3.1）基因[8]，或整个主要组织相容性复合体（MHC）区域[9]。Illumina Human Comprehensive Panel 主要靶向蛋白质编码基因内离散的低覆盖度区域，可作为预设计或即用型 panel 提供（因美纳，货号 20113836）[10,11]。DesignStudio 软件支持根据 BED 文件 † 设计定制 panel 或修改现有的预设计。

### 定制探针 panel 的推荐测序深度

Illumina Complete Long Read Prep with Enrichment, Human 可提供高度稳定且可靠的性能。对于所测试的预设计 panel，每 1 Mb 的目标 panel 获得约 1.5 Gb 的测序数据（约 5M 双端 read）即可实现理想性能（图 3）。对于性能未知的新设计 panel，建议以每 1 Mb 目标 panel 获取 3 Gb 的序列数据（约 10M 双端 read）为起点，然后再逐步减少进一步优化。

## 高度准确地覆盖和定相复杂区域

长读长富集探针 panel 专注于增强特定的低覆盖度区域的覆盖深度，例如 Illumina Human Comprehensive Panel 和 CMRG panel，可提高复杂的目标区域的变异检出准确度（图 4）。使用 CMRG panel 进行长读长富集还有助于提高对蛋白质编码区域变异的覆盖度和检测的全面性（图 5、图 6）。

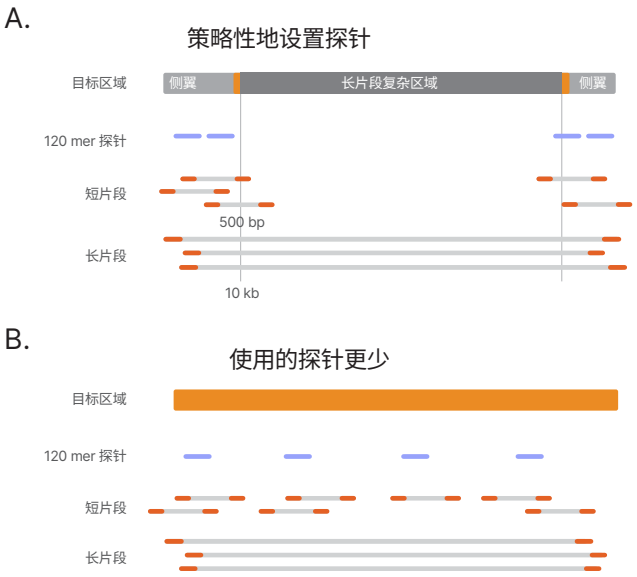


图 2：长片段杂交可提高富集效率——与短片段捕获相比，长片段杂交更具优势，包括（A）将探针策略性地放置在难以设计探针的区域之外，例如 GC 含量过高、低复杂性或重复序列的区域，以及（B）捕获每个目标区域所需的探针更少。DesignStudio 算法以 1 kb 为单位搜索目标区域，寻找最佳 GC 含量和特异性最高的区域来放置探针。

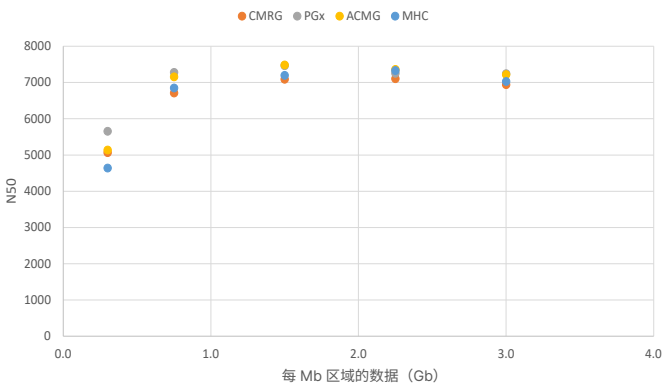


图 3：定制探针 panel 的测序要求——对最大 N50 所需的测序数据进行逐步优化表明，每 Mb 目标区域获得 1.5 Gb（约 5M 双端 read）数据可有效解析目标区域，以生成 Illumina Complete Long Read 数据。

† BED，浏览器可扩展数据格式。

表 1：专为 Illumina Complete Long Read Prep with Enrichment, Human 预设计的富集探针 panel

| Panel[a] | CMRG panel | PGx panel | ACMG panel | MHC panel |
|---|---|---|---|---|
| 靶向基因 | 391 个医学相关基因，已知这些基因难以用短读长解析 [5] | 药物遗传学试验检测中常见的 98 个靶向基因 [6-8] | 来自 ACMG 二级研究结果列表（ACMG SF v3.1）的 78 个独特基因 [9] | GRCh38.p14 组装中的整个 MHC 区域（超过 140 个基因）[10] |
| 目标区域大小 [b] | 22.5 Mb | 8.1 Mb | 7 Mb | 4.9 Mb |
| 每个样本的测序产出 [c] | 约 67.5 Gb | 约 24.3 Gb | 约 21 Gb | 约 14.7 Gb |
| 探针数量 | 约 22.5K | 约 8.2K | 约 6.9K | 约 5.0K |
| N50d | 6.1 kb | 7.3 kb | 7.3 kb | 7.3 kb |
| 相区块 N50[d,e] | 82.8 kb | 94.4 kb | 94.4 kb | 357 kb |
| 平均目标区域大小 [e] | 58 kb | 83 kb | 88 kb | 5000 kb |
| 均一性 [d,f] | 97.9% | 99.0% | 99.5% | 97.8% |
| 侧翼 read 富集（PRE）[d,f] | 80.1% | 79.3% | 66.3% | 67.5% |
| 定相的 SNV 百分比 [d] | 98.9% | 98.9% | 99.6% | 98.6% |

a. CMRG，复杂区域医学相关基因；PGx，药物遗传学；ACMG，美国医学遗传学和基因组学学会；MHC，主要组织相容性复合体。
b. 目标区域大小为填补探针位置长度的总和，在重叠处合并。
c. 需要 2×150 bp 测序运行和每 Mb 目标区域的 5M-10M 双端 read（约 1.5-3 Gb 数据），从而产生约 30× 的 Illumina Complete Long Reads 最终覆盖度。每个样本的定制 panel 数据要求只是一个建议的起点。用户可根据 panel 性能优化分配的数据。
d. 使用 50 ng HG002 基因组 DNA（Coriell，货号 NA24385）生成的数据。性能可能因 DNA 起始量和样本质量的不同而有所不同。
e. 相区块大小受限于单个连续目标区域的大小。
f. 覆盖均一性按> 0.2* 平均值的百分比来计算。PRE 按 100* 计算（填补目标匹配 reads/ 总匹配 reads）。

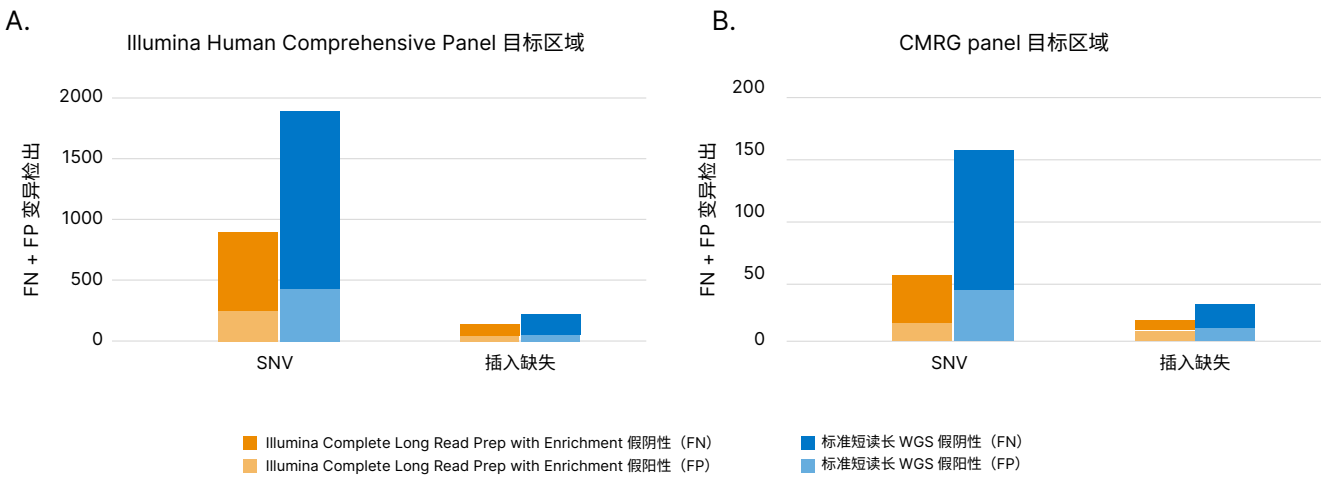


图 4：通过靶向长读长提高复杂区域的变异检出准确性——与标准短读长 WGS（蓝色）相比，使用 Illumina Complete Long Read Prep with Enrichment（橙色）在 (A) Human Comprehensive Panel 或 (B) CMRG panel 靶向的 HG002 基因区域的单核苷酸位点变异 (SNV) 和插入缺失 (indel) 的假阴性（FN）和假阳性（FP）变异检出情况。


图 5： 靶向长读长可增强低覆盖度区域的覆盖深度——与标准短读长 WGS（下）相比，使用 Illumina Complete Long Read Prep, Human WGS（上）和 Illumina Complete Long Read Prep with Enrichment, Human 和 CMRG panel（中）对 *HBG1* 进行长读长测序的 Integrative Genomics Viewer（IGV）图谱。

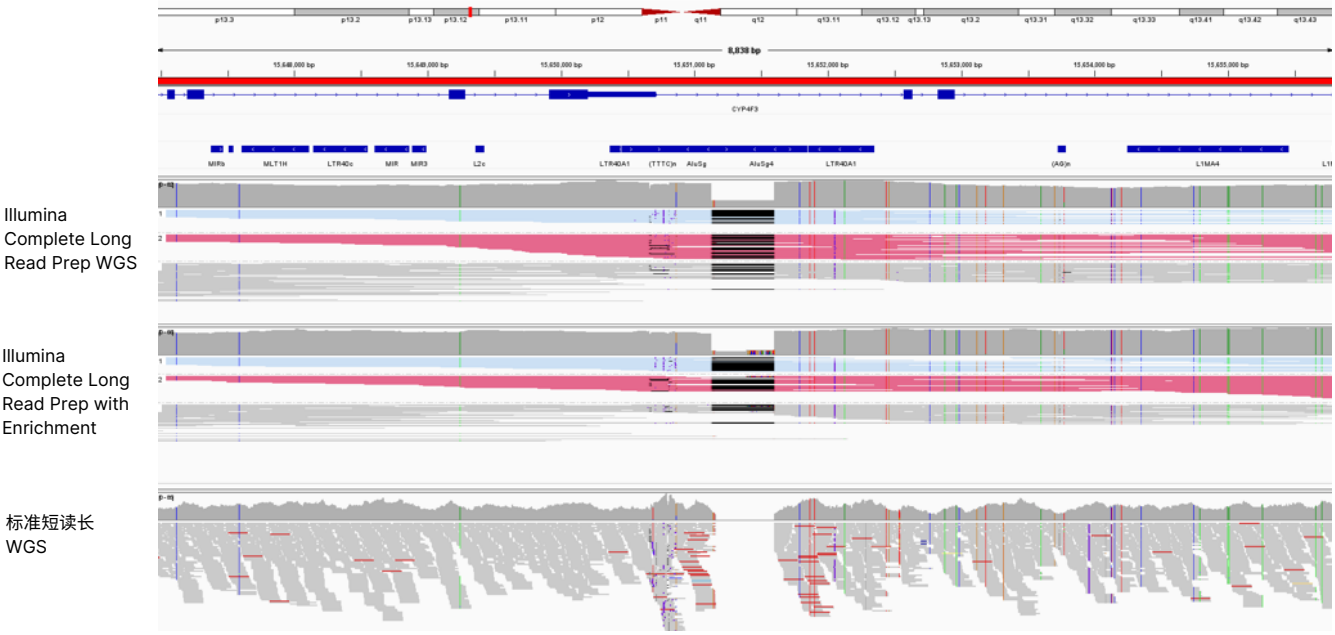


图 6： 通过靶向长读长清晰分辨缺失片段的边界——与标准短读长 WGS（下）相比，使用 Illumina Complete Long Read Prep, Human WGS（上）和 Illumina Complete Long Read Prep with Enrichment, Human 和 CMRG panel（中）对 CYP4F3 进行长读长测序和定相的 IGV 图谱。蓝色表示等位基因 1，粉红色表示等位基因 2。

## 用于解析单倍型的长相区块

每个 panel 的相区块 N50‡ 与目标区域的连续长度有关（图 7、表 1）。CMRG、PGx 和 ACMG panel 旨在靶向全长目标基因，并产生约 80-95 kb 的平均相区块 N50，以实现杂合等位基因的完全定相（图 8）。MHC panel 靶向单个约 4.9 Mb 的连续区域，并产生超过 350 kb 的平均相区块 N50，从而解析全长基因区域（图 9）。

‡ 相区块 N50 反映了占目标区域总组装长度 50% 的连续序列的最短区块长度。

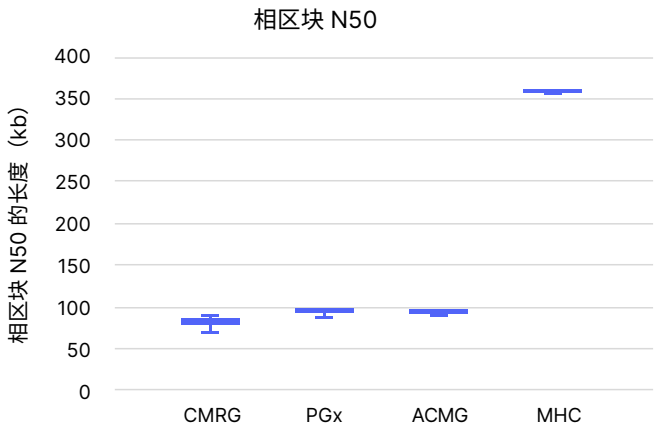


图 7：相区块 N50 取决于连续目标区域的长度——CMRG、PGX 和 ACMG panel 靶向全长目标基因，并产生约 80-95 kb 的平均相区块 N50。MHC panel 靶向整个主要组织相容性复合体基因区域，并产生超过 350 kb 的平均相区块 N50。CMRG panel 的平均目标区域大小为 58 kb，PGx panel 为 83 kb，ACMG panel 为 88 kb，MHC panel 为 5000 kb。

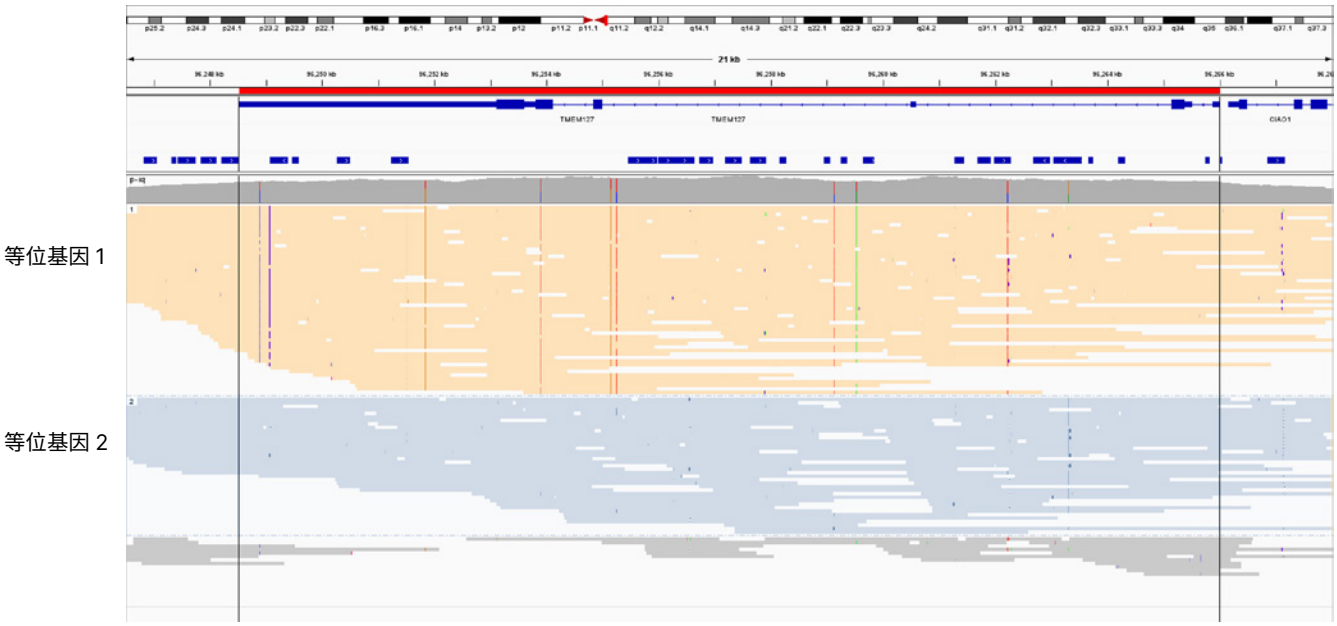


图 8：靶向长读长能够对具有杂合 SNV 的区域进行定相——长读长测序的 IGV 图显示了使用 Illumina Complete Long Read Prep with Enrichment, Human 和 ACMG panel 在单相区块中对 *TMEM17*（一种长度为 21 kb 的基因）完全定相。黄色表示等位基因 1。蓝色表示等位基因 2。

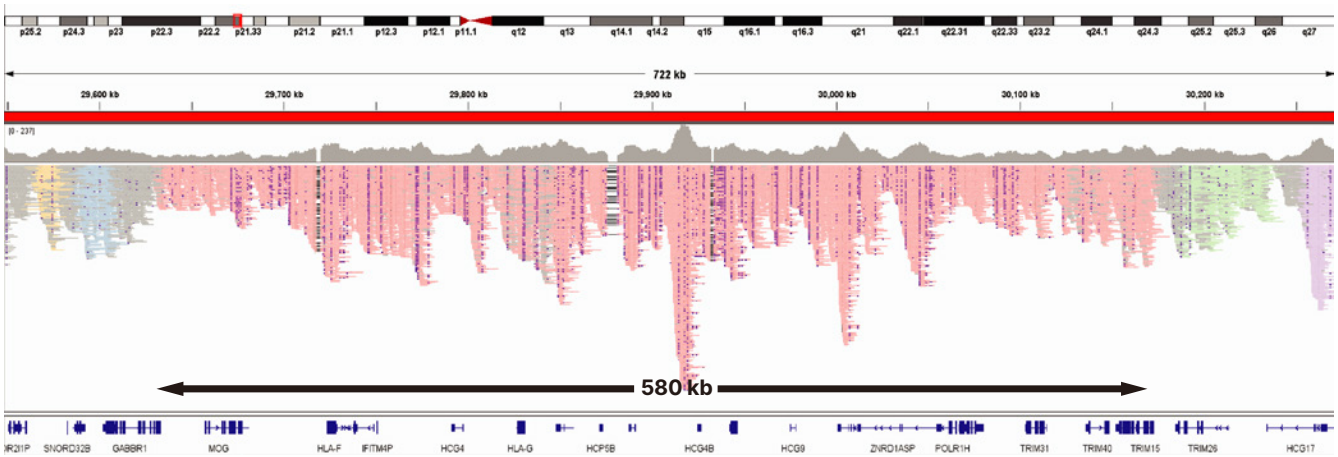


图 9：靶向长读长有助于解析多态性基因中的单倍型——使用 Illumina Complete Long Read Prep with Enrichment, Human 进行长读长测序的 IGV 图谱。MHC 基因位点 722 kb 区域的定相。一个相区块中包含 580 kb 的区域（粉红色）。

## 总结

Illumina Complete Long Read Prep with Enrichment, Human 是经过验证的因美纳短读长 WGS 的补充，专注于长读长测序，可通过长读长提供更高的价值。研究人员可以灵活地选择预设计 panel，或使用 DesignStudio 算法设计定制 panel，以实现长读长靶向富集。靶向富集探针 panel 可以增强覆盖或通过完整的工作流程解决方案对整个基因进行定相来获得更多信息，从而实现经济高效、高度准确的 WGS。

## 了解更多

Illumina Complete Long Read Prep with Enrichment, Human

DesignStudio 实验分析设计工具

长读长测序技术

## 参考文献


1. Illumina. Illumina Complete Long Read Prep, human data sheet. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-datasheet-m-gl-01420/illumina-long-read-prep-data-sheet-mgl-01420.pdf. Published 2022. Accessed September 22, 2023.
2. Illumina. Comprehensive whole-genome sequencing with illumina Complete Long Read Prep, Human technical note. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-prep-human-technote-m-gl-01421/ilmn-long-read-hu-tech-note-m-gl-01421.pdf. Published 2022. Accessed September 22, 2023.
3. Roessler K. Illumina Complete Long Reads software analysis workflow for human WGS. https://www.illumina.com/science/ genomics-research/articles/complete-long-read-softwareanalysis.html. Published 2023. Accessed September 22, 2023.
4. Wagner J, Olson ND, Harris L, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol*. 2022;40(5):672-680. doi:10.1038/ s41587-021-01158-1
5. PharmGKB. ViPs: Very Important Pharmacogenes. pharmgkb.org/vips. Accessed September 22, 2023.
6. National Library of Medicine. GTR: Genetic Testing Registry. Precision HealthPGx Panel (25 Genes). ncbi.nlm.nih.gov/gtr/tests/593428/. Updated November 29, 2022. Accessed September 22, 2023.

7. Pratt VM, Everts RE, Aggarwal P, et al. Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. *J Mol Diagn*. 2016;18(1):109-123. doi:10.1016/j.jmoldx.2015.08.005
8. Miller DT, Lee K, Abul-Husn NS, et al. ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2022;24(7):1407-1414. doi:10.1016/j.gim.2022.04.006
9. Kulski JK, suzuki s, shiina t. human leukocyte antigen superlocus: nexus of genomic supergenes, SNPs, indels, transcripts, and haplotypes. *Hum Genome Var*. 2022;9(1):49. doi:10.1038/s41439-022-00226-5
10. Bekritsky Ma, Colombo C, Eberle MA. Identifying genomic regions with high quality single nucleotide variant calling. Published 2021. Accessed August 30, 2023.
11. Illumina. human Comprehensive Panel data sheet. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/illumina-long-read-enrich-hu-comp-panel-data-sheet-m-gl-02191/long-read-hu-comp-panel-data-sheet-m-gl-02191.pdf.



**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）8441-6900 • 传真（010）8455-4855
技术支持热线 400-066-5835 • chinasupport@illumina.com
市场销售热线 400-066-5875 • china_info@illumina.com • www.illumina.com.cn



M-GL-02189 v1.0

因美纳 因美纳讲堂