Custom panel design for Illumina Complete Long Read Prep with Enrichment, Human

Highly flexible, targeted long-read enrichment for human genomes
Introduction

When performing human whole-genome sequencing (WGS), several regions may be difficult to map with short reads alone. Long-read sequencing can complement standard short-read WGS data to help address these challenging regions. Illumina Complete Long Reads technology uses a standard next-generation sequencing (NGS) workflow to generate contiguous long-read sequences on Illumina sequencing systems with a single analysis pipeline (Figure 1).* Illumina Complete Long Read Prep with Enrichment, Human adds a targeted approach for more cost-effective long-read sequencing.* Illumina Complete Long Read enrichment chemistry provides high flexibility with targets and probe design to help resolve difficult-to-map regions or provide added insight from phased sequencing.

Designing enrichment probe panels for long reads

Illumina Complete Long Read Prep with Enrichment, Human uses a different probe design strategy to capture longer fragments (~7–10 kb) than the approach typically used to capture short fragments (~200–500 bp).

* Requires ≥ 30× standard short-read WGS data from the same sample for analysis. FASTQ files from a previously run sample can be used.

Illumina DesignStudio™ software is a free, user-friendly tool for designing enrichment probe panels. The DesignStudio algorithm considers GC content, target specificity, and probe spacing, ie, how many probes will be in the target region. Standard spacing for 120mer short-read enrichment panels is a 250–350-bp probe window. For long-read enrichment panel design, probe spacing was tested at multiple lengths and a one-kilobase window was found to be optimal for cost-effective, high-efficiency capture.

Hybridization enrichment effectiveness is highly dependent on probe specificity. The percent of on-target enrichment directly affects the amount of sequencing needed to achieve target coverage depth. High specificity is harder to achieve for repetitive regions. However, using a larger probe window allows greater flexibility to exclude poor-performing probes, avoid repeat regions (up to the window size of 1 kb), and maintain enrichment efficiency with fewer probes (Figure 2). The DesignStudio algorithm can use these considerations to recommend probe placement. Third party panels should use similar guidelines for best performance and cost efficiency. Standard enrichment probe spacing is also fully compatible.

Flexibility in probe design and target strategy

Illumina Complete Long Read Prep with Enrichment, Human offers high flexibility to choose and design custom probe panels to match study goals. Individual targeted regions can span single bases up to hundreds of kilobases. The total panel size can range from custom panels as small as...

Figure 1: Part of an integrated workflow—Access cost-effective, targeted long-read WGS data using a scalable, optimized library prep with enrichment protocol, proven Illumina sequencing chemistry, and DRAGEN secondary analysis. Requires ≥ 30× standard short-read WGS data from the same sample for analysis. FASTQ files from a previously run sample can be used.
as 2.5 Mb up to > 95 Mb. Researchers can use targeted long reads to enhance coverage across specific regions known to have low mappability with short-read data. Alternatively, long reads can be targeted to cover entire genes up to long multigene regions to enable phasing of variants and calling haplotypes.

Several predesigned panels are available in the DesignStudio tool (Table 1). These panels target challenging medically relevant genes (CMRG), genes commonly targeted by pharmacogenetic (PGx) testing assays, genes on the American College of Medical Genetics and Genomics (ACMG) secondary findings list (ACMG SF v3.1), or the full major histocompatibility complex (MHC) region. The Illumina Human Comprehensive Panel, which primarily targets discrete low-coverage regions within protein-coding genes, is also available as a predesign or ready-to-ship premanufactured panel (Illumina, Catalog no. 20113836). DesignStudio software supports design of custom panels from BED files or modification of existing predesigns.

**Recommended sequencing depth for custom probe panels**

Illumina Complete Long Read Prep with Enrichment, Human provides highly consistent and robust performance. For the predesigned panels tested, optimal performance was achieved with ~1.5 Gb of sequence data (~5M paired-end reads) per 1 Mb of target panel size (Figure 3). For newly designed panels with unknown performance, 3 Gb of sequence data (~10M paired-end reads) per 1 Mb of target panel size is a recommended starting point, subject to reduction with further optimization.

**High-accuracy coverage and phasing of difficult regions**

Long-read enrichment probe panels focused on enhancing specific low-coverage regions, like the Illumina Human Comprehensive Panel and the CMRG panel, enhance variant calling accuracy in targeted challenging regions (Figure 4). Long-read enrichment with the CMRG panel also helps improve completeness of coverage and detection of variants across protein-coding regions (Figure 5, Figure 6).

---

† BED, browser extensible data format.
Table 1: Predesigned enrichment probe panels for Illumina Complete Long Read Prep with Enrichment, Human

<table>
<thead>
<tr>
<th>Panels&lt;sup&gt;a&lt;/sup&gt;</th>
<th>CMRG panel</th>
<th>PGx panel</th>
<th>ACMG panel</th>
<th>MHC panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes targeted</td>
<td>391 medically relevant genes known to be challenging to resolve with short reads&lt;sup&gt;b&lt;/sup&gt;</td>
<td>98 genes commonly targeted by pharmacogenetic testing assays&lt;sup&gt;c&lt;/sup&gt;–&lt;sup&gt;d&lt;/sup&gt;</td>
<td>78 unique genes from the ACMG secondary findings list&lt;sup&gt;e&lt;/sup&gt; (ACMG SF v3.1)&lt;sup&gt;f&lt;/sup&gt;</td>
<td>&gt; 140 genes across the full MHC region in the GRCh38.p14 assembly&lt;sup&gt;g&lt;/sup&gt;</td>
</tr>
<tr>
<td>Target region size&lt;sup&gt;b&lt;/sup&gt;</td>
<td>22.5 Mb</td>
<td>8.1 Mb</td>
<td>7 Mb</td>
<td>4.9 Mb</td>
</tr>
<tr>
<td>Sequencing output per sample&lt;sup&gt;c&lt;/sup&gt;</td>
<td>~67.5 Gb</td>
<td>~24.3 Gb</td>
<td>~21 Gb</td>
<td>~14.7 Gb</td>
</tr>
<tr>
<td>No. of probes</td>
<td>~22.5K</td>
<td>~8.2K</td>
<td>~6.9K</td>
<td>~5.0K</td>
</tr>
<tr>
<td>N50&lt;sup&gt;d&lt;/sup&gt;</td>
<td>6.1 kb</td>
<td>7.3 kb</td>
<td>7.3 kb</td>
<td>7.3 kb</td>
</tr>
<tr>
<td>Phase block N50&lt;sup&gt;d,e&lt;/sup&gt;</td>
<td>82.8 kb</td>
<td>94.4 kb</td>
<td>94.4 kb</td>
<td>357 kb</td>
</tr>
<tr>
<td>Mean target region size&lt;sup&gt;e&lt;/sup&gt;</td>
<td>58 kb</td>
<td>83 kb</td>
<td>88 kb</td>
<td>5000 kb</td>
</tr>
<tr>
<td>Uniformity&lt;sup&gt;d,f&lt;/sup&gt;</td>
<td>97.9%</td>
<td>99.0%</td>
<td>99.5%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Padded read enrichment (PRE)&lt;sup&gt;d,f&lt;/sup&gt;</td>
<td>80.1%</td>
<td>79.3%</td>
<td>66.3%</td>
<td>67.5%</td>
</tr>
<tr>
<td>% Heterozygous SNV phased&lt;sup&gt;d&lt;/sup&gt;</td>
<td>98.9%</td>
<td>98.9%</td>
<td>99.6%</td>
<td>98.6%</td>
</tr>
</tbody>
</table>

<sup>a</sup> CMRG, challenging medically relevant genes; PGx, pharmacogenomics; ACMG, American College of Medical Genetics and Genomics; MHC, major histocompatibility complex.

<sup>b</sup> Target region size is the sum of padded probe location lengths, merged where they overlap.

<sup>c</sup> Requires 2 × 150 bp sequencing run and 5M–10M paired-end reads (~1.5–3 Gb data) per Mb target region, generating approximately 30× final coverage of Illumina Complete Long Reads. Custom panel data requirements per sample are only a recommended starting point. Users may optimize allotted data based on panel performance.

<sup>d</sup> Data generated using 50 ng HG002 genomic DNA (Coriell, Catalog no. NA24385). Performance may vary with DNA input and sample quality.

<sup>e</sup> Phase block sizes are limited to the sizes of individual contiguous targets regions.

<sup>f</sup> Uniformity of coverage calculated as % > 0.2 * mean. PRE calculated as 100 * (padded target aligned reads / total aligned reads).

Figure 4: Targeted long reads to enhance variant calling accuracy in challenging regions—False negative (FN) plus false positive (FP) variant calls for single nucleotide variants (SNVs) and insertion–deletions (Indels) in HG002 genic regions targeted by the (A) Human Comprehensive Panel or (B) CMRG panel, using Illumina Complete Long Read Prep with Enrichment (orange) compared to standard short-read WGS (blue).
Figure 5: Targeted long reads enhance regions of low coverage—Integrative Genomics Viewer (IGV) plots from long-read sequencing of *HBG1* using Illumina Complete Long Read Prep, Human WGS (top) and Illumina Complete Long Read Prep with Enrichment, Human and the CMRG panel (middle) compared to standard short-read WGS (bottom).

Figure 6: Clear resolution of deletion borders with targeted long reads—IGV plots from long-read sequencing and phasing of *CYP4F3* using Illumina Complete Long Read Prep, Human WGS (top) and Illumina Complete Long Read Prep with Enrichment, Human and the CMRG panel (middle) compared to standard short-read WGS (bottom). Allele 1 in blue, allele 2 in pink.
Long phase blocks to resolve haplotypes

The phase block N50 ‡ of each panel is related to the contiguous length of the target regions (Figure 7, Table 1). The CMRG, PGx, and ACMG panels are designed to target full-length genes of interest and yielded a mean phase block N50 of ~80–95 kb to enable full phasing of heterozygous alleles (Figure 8). The MhC panel targets a single ~4.9 Mb contiguous region and yielded a mean phase block N50 of over 350 kb to enable resolution of the full-length genic region (Figure 9).

‡ Phase block N50 reflects the length of the shortest block of contiguous sequence at 50% of the total assembly length of target regions.

**Figure 7:** Phase block N50 depends on the length of contiguous target regions—The CMRG, PGx, and ACMG panels target full-length genes of interest and yielded mean phase block N50 of ~80–95 kb. The MhC panel targets the full major histocompatibility complex genic region and yielded a mean phase block N50 of over 350 kb. The mean target region size is 58 kb for the CMRG panel, 83 kb for the PGx panel, 88 kb for the ACMG panel, and 5000 kb for the MhC panel.

**Figure 8:** Targeted long reads enable phasing of regions with heterozygous SNVs—IGV plots from long-read sequencing shows full phasing in one phase block for TMEM17, a 21 kb gene, using Illumina Complete Long Read Prep with Enrichment, Human and the ACMG panel. Allele 1 in yellow. Allele 2 in blue.
Summary

Illumina Complete Long Read Prep with Enrichment, Human complements proven Illumina short-read WGS and focuses long-read sequencing where it provides greatest value. Researchers have high flexibility to choose predesigned panels or use the DesignStudio algorithm to design custom panels for long-read targeted enrichment. Enrichment probe panels can be targeted to enhance coverage or add insight by phasing entire genes for cost-effective, highly accurate WGS with a full workflow solution.

Learn more

Illumina Complete Long Read Prep with Enrichment, Human

DesignStudio assay design tool

Long-read sequencing technology

References


