Guide to Reviewing and Approving Custom KAPA Target Enrichment Designs

Overview

This document describes how to review and approve proposed custom KAPA Target Enrichment designs based on the genomic regions you provided.

The purpose of reviewing a design is to verify that all desired regions are adequately covered and that the design meets your needs. This is a required part of the process, as a custom design needs to be reviewed and approved before it can be released to manufacturing.

If the regions that are vital to your experiment are not adequately covered, you might need to modify the parameters of your design, or contact your local Roche representative for additional options such as working with an expert designer.

Getting Started

To review a design (created using the automated design tool or by working with expert designers), go to the HyperDesign home page and choose the **Your Designs** option.

Roche provides design files in three formats:

- **BED (.bed) files:** Viewable in web-based genome browsers, e.g. UCSC Genome Browser, Integrative Genomics Viewer (IGV, Broad Institute), or Ensembl.
- **Coverage Summary (.txt) files**: Viewable with a text editor or spreadsheet software.
- **Design Report (.pdf) file**: Viewable with any PDF viewer, contains relevant information about your design.

When reviewing your design, keep in mind the following:

- For the purpose of coverage visualization, use the three BED files provided with the design:
    - primary_targets.bed
    - capture_targets.bed
    - predicted_no_coverage_regions.bed

  See Appendix: Design Files for more information.

- Summary files distinguish between coverage and estimated coverage as well as capture and estimated capture. The distinctions are as follows:
    - **Capture** is the sequence obtained when the probe hybridized with the DNA library fragments.
    - **Coverage** is the sequence of the target regions, or regions of interest.
    - **Estimated** is an estimated capture or coverage of sequence adjacent to the probe. The laboratory protocol results in probes reliably capturing up to 100 bp of sequence on either side of the probe target. The value given by the estimated capture (or coverage) metric

refers to the sequence targeted by the probe and the adjacent sequence predicted to be captured following the KAPA HyperCap Workflow.

NOTE: *if the value is specified as just capture or coverage, it refers to only the sequences targeted by the probes.*

NOTE: *for degraded samples (e.g., FFPET, ancient DNA) estimated values* are a less reliable prediction for sequence coverage.

- Focus on gaps in the *capture_targets* track that do not provide coverage for the *primary_targets* track. These gaps represent portions of a target region not directly covered by the probes. Review the region-by-region coverage file for detailed direct probe coverage or estimated probe coverage for each region. Review the *predicted_no_coverage_regions* file for regions estimated to have no coverage.
- If two or more target regions overlap, Roche automatically merges them into a single region. Therefore, “Initial regions count” and “Final regions count after consolidation” may differ.
- Regions not covered by the design are typically repetitive regions, which, if included, cause capture of other homologous regions in the genome and decrease capture efficiency. Therefore, most KAPA Target Enrichment experiments benefit from excluding these regions in the design. Check the region-by-region coverage file (ending in coverage.txt) for information on regions not covered due to repetitive regions.
- The stringency filter Roche uses, by default, does not include low complexity regions in the design. For customers working with an expert designer, and these regions are necessary to answer a specific research question, please note this when submitting the design request (use the “Additional details” field when completing the electronic design specification form [eDSF]). Be aware that using less stringent criteria during design generation may provide more genomic coverage at the cost of a decrease in capture efficiency. There may be more off-target reads when the captured DNA is sequenced.

Review a Custom Design

Step 1. Review the Coverage Summary File and Other Design Files

The design and coverage files describe the properties of the KAPA Target Enrichment custom design.

1. Using a text editor or spreadsheet software (such as WordPad, Notepad, or Microsoft Excel), open the *coverage_summary.txt* file.

	A	B	C
1	Genome build	hg38	
2	Number of regions	9859	
3	Length of regions (bp)	1996105	
4			
5	Statistics	Probe_Coverage	Estimated_Coverage
6	Target Bases Covered	980684	1297072
7	% Target Bases Covered	49.1	65
8	Targets with no coverage	3119	3110
9			
10	Target Bases Not Covered	1015421	699033
11	Due to N's	0	0
12	Due to repeats	905352	648552
13	% Target Bases Not Covered	50.9	35
14	Due to N's	0	0
15	Due to repeats	45.4	32.5
16			
17	Total capture targets	7463	
18	Total capture space (bp)	1395394	

Figure 1: Review the coverage summary

2. Refer to Appendix: Design Files for the definition of each field.
3. Review each field to ensure that the design meets the specifications for your KAPA Target Enrichment project.
4. Open the *coverage.txt* file. It is recommended that you view this file with a spreadsheet program, such as Microsoft Excel.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	REGION_NAME	CHROMOSOME	START	STOP	LENGTH	BASES_PROBE_COVERAGE	FRAC_PROBE_COVERAGE	BASES_ESTIMATED_COVERAGE	FRAC_ESTIMATED_COVERAGE	BASES_W_NO_PROBE_COV	BASES_W_NO_PROBE_COV_DUE_TO_N	BASES_W_NO_PROBE_COV_DUE_TO_REPEATS	BASES_W_NO_EST_COV	BASES_W_NO_EST_COV_DUE_TO_N	BASES_W_NO_EST_COV_DUE_TO_REPEATS
2	chr1:987432-987532	chr1	987382	987582	200	27	0.135	127	0.635	173	0	173	73	0	73
3	chr1:987821-987922	chr1	987772	987972	200	43	0.215	152	0.76	157	0	157	48	0	48
4	chr1:1088358-1088458;	chr1	1088308	1088508	200	200	1	200	1	0	0	0	0	0	0
5	chr1:1248013-1248113;	chr1	1248013	1249113	1100	112	0.102	212	0.193	988	0	0	888	0	0
6	chr1:1342249-1342349;	chr1	1342199	1342399	200	200	1	200	1	0	0	0	0	0	0
7	chr1:1379044-1379144;	chr1	1378994	1379194	200	112	0.56	200	1	88	0	88	0	0	0
8	chr1:1589263-1589363	chr1	1589213	1589413	200	0	0	0	0	200	0	200	200	0	200
9	chr1:1589528-1589628	chr1	1589478	1589678	200	0	0	0	0	200	0	176	200	0	176
10	chr1:1590379-1590479	chr1	1590329	1590529	200	0	0	0	0	200	0	122	200	0	122
11	chr1:1590610-1590710	chr1	1590560	1590760	200	0	0	0	0	200	0	200	200	0	200
12	chr1:1995427-1995528	chr1	1995378	1995578	200	89	0.445	189	0.945	111	0	111	11	0	11
13	chr1:2204826-2204926;	chr1	2204776	2204976	200	42	0.21	200	1	158	0	0	0	0	0
14	chr1:2299099-2299199;	chr1	2299049	2299249	200	200	1	200	1	0	0	0	0	0	0
15	chr1:2417370-2417470;	chr1	2417320	2417520	200	200	1	200	1	0	0	0	0	0	0
16	chr1:2837998-2838098;	chr1	2837948	2838148	200	200	1	200	1	0	0	0	0	0	0

Figure 2: Review the coverage details

This file displays region-by-region coverage information, with relevant details on why regions might not have full coverage. Review this file thoroughly to ensure that the design meets the specifications for your Target Enrichment project. To quickly identify regions with little or no coverage, sort the file by percent coverage. For a detailed list of column header descriptions, see Appendix: Design Files.

Step 2. Review the Design BED files

To review the design BED files, refer to the appropriate instructions for the software in use. The following uses UCSC Genome Browser as an example.

1. Save to your computer the *primary_targets.bed*, *capture_targets.bed*, and *predicted_uncovered_targets.bed* files provided in your design file deliverables.
2. Go to the UCSC Genome Browser home page at http://genome.ucsc.edu.
3. On the menu, click **Genomes**. The Genome Browser Gateway page opens.
4. Enter the species or common name, choose the genomic assembly or build, and click **GO**.
5. Click the **add custom tracks** button located in the middle of the buttons below the genome browser display.
6. To select and upload the BED file, on the Add Custom Tracks page, click **Choose File** and click **Submit**. The Manage Custom Tracks page manages all added custom tracks. If loading multiple tracks, edit the User Track name and description with a unique identifier before submitting another custom track/file to help visualize the data.
7. Click **User Track** to edit the custom track:
    a. Change the configuration of existing tracks. Figure 3 shows an example of changing a track configuration using **Edit configuration**.

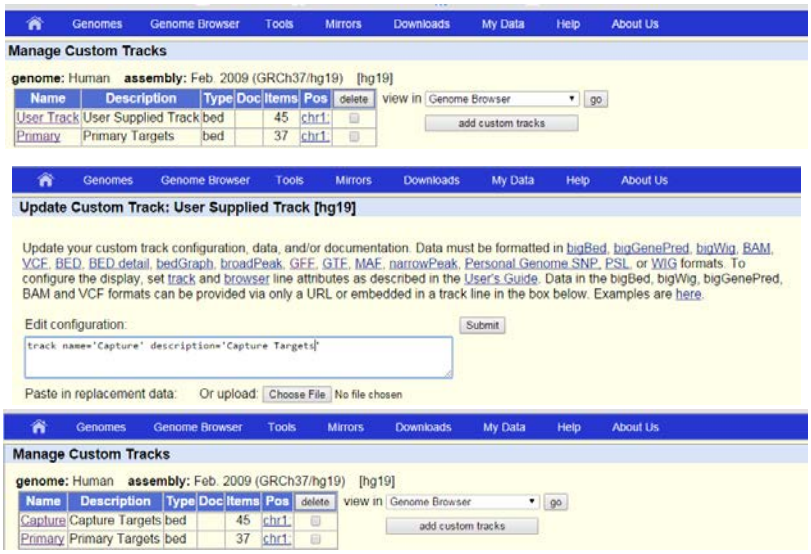

Figure 3: Managing Custom Tracks in the UCSC Genome Browser

    b. To add additional BED files, repeat steps 5 and 6.
    c. Click the **go** button to view the custom track(s). Figure 4 illustrates an example design displayed in the UCSC Genome Browser. Note that the default tracks will be different from the view in Figure 4. Visibility of UCSC provided tracks may be changed by right clicking the bars on the left side of the browser or from the track selections below the image and drop-down controls at the bottom of the page.

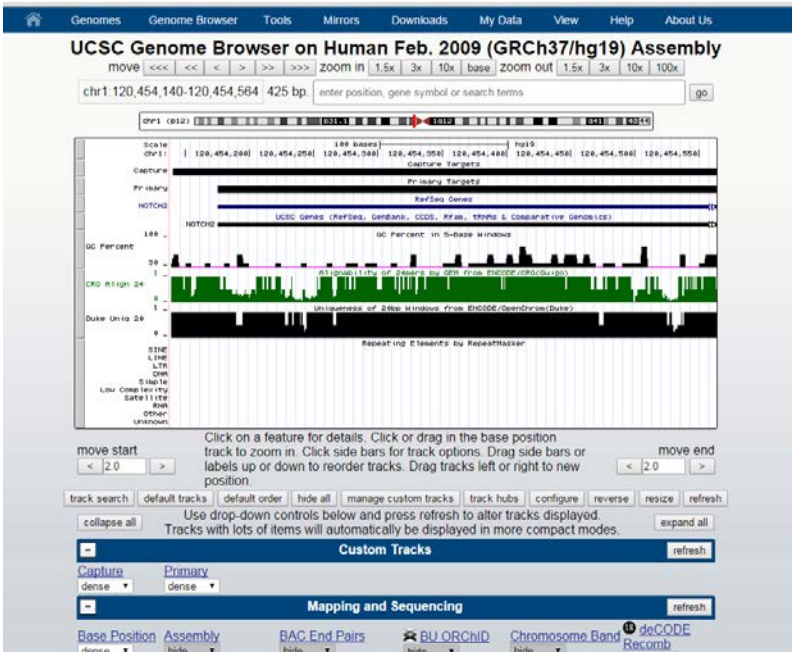

Figure 4: UCSC Genome Browser

8. Review the design using the following UCSC Genome Browser functions:

   - **Zoom**: Click the **zoom in** and **zoom out** buttons to zoom in or out on the center of the annotation tracks window by 1.5-, 3-, or 10-fold.
   - **Scroll**: Click the **move** buttons to scroll to the left or right.
   - **Display**: To display a different position in the genome, in the **position/search** text box, enter the coordinates and click the **jump** button.
   - **View base composition**: Click the **base** button to view the base composition of the sequence underlying the current annotation track display.

   UCSC Genome Browser provides useful tracks to load including Mappability and RepeatMasker tracks. These can be used to diagnose regions uncovered due to repeats.

9. For additional details about the UCSC Genome Browser's capabilities, click the **Help** link.

Step 3. Approve the Design

- For custom designs made with the automated design tool:
  If satisfied with the design, click the **Approve design** button. For a design made using the automated design tool, an Internal Reference Number (IRN) will be assigned to the design upon approval. Provide the IRN to a Roche representative when placing an order for the design.

  To make further modifications, choose the **Clone design** option to keep the input regions and modify the design parameters to create a new design.

- For custom designs made working with expert designers:
  If satisfied with the design, *you must provide written approval via email to the designer*. The final design deliverables will appear under **Your Designs** after the purchase order of the probes is processed.

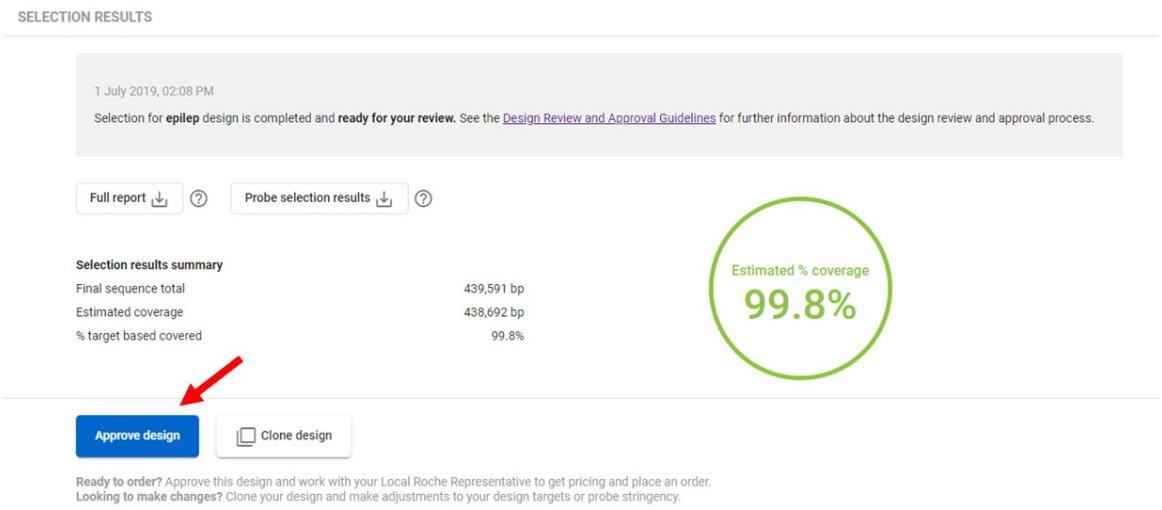

Figure 5: Approving Design

For any questions, please contact Roche Technical Support (sequencing.roche.com/support.html).

Appendix: Design Files

File Formats for Regions of Interest

HyperDesign Software allows you to type or paste design coordinates or gene names/identifiers, or import a design coordinate list or gene name/identifier file. The software allows three formats for specifying regions of interest:

- 1-column text file

  A text file with the information in 1 field: chromosome:start-stop

- 3-column text file

  A tab-delimited text file with 3 fields: chromosome<tab>start<tab>stop.

- 4-column text file

  A tab-delimited text file with 4 fields: chromosome<tab>start<tab>stop<tab>name. The fourth field can contain comments or a region name that will be carried through to the final enrichment design BED files. The contents of the fourth field will not affect the design.

Consolidated Regions

The BED file used as input for the probe selection algorithm. The consolidated regions are obtained from either merging overlapping coordinates provided by the customer or from the coordinates obtained for the identifiers in the gene list.

Capture

The sequence matched by the probe and captured during the KAPA HyperCap Workflow. These often overhang the consolidated regions.

Coverage

The sequence of the consolidated regions both matched by the probe and captured during the KAPA HyperCap Workflow without including any overhang.

Estimated Capture/Coverage

The KAPA HyperCap Workflow results in probes reliably capturing up to 100 bp of sequences on either side of the probe. In other words, the targeted sequence of the probe and the 100 bp adjacent sequence are both to be captured. Estimated capture sequence is the captured sequence plus the 100 bp adjacent to the capture sequence. Estimated coverage is the coverage plus the lesser of either the end of the consolidated region or 100 bp adjacent to the coverage. The 100 bp capture padding was validated with Illumina paired-end sequencing, using a typical library size of ~200 bp. This number may not be accurate for libraries with larger or smaller insert sizes, or single end reads.

Overhang

The number of bases that a probe might overhang the end of the specified target. For smaller targets, probes may overhang the ends up to a maximum of 125 base pairs. For all other targets, an overhang of zero will restrict probe placement to be within the targeted regions.

capture_targets.bed

Probe coverage regions where each base is covered by at least one probe. This is a tab-delimited coordinate file with no header, in BED format (*http://genome.ucsc.edu/FAQ/FAQformat.html#format1*), and suitable for viewing in various genome browsers.

primary_targets.bed

Customer requested regions of interest, with overlapping ranges consolidated - that overlap at least 1 bp with a probe. This is a tab-delimited coordinate file with no header, in BED format (*http://genome.ucsc.edu/FAQ/FAQformat.html#format1*), and suitable for loading into various genome browsers. Note that any requested regions with no probes selected against them will not appear in this file.

predicted_no_coverage_regions.bed

All positions from the *regions.bed* (regions of interest) that are not within 100 base pairs of any probe. This is a tab-delimited coordinate file with no header, in BED format (*http://genome.ucsc.edu/FAQ/FAQformat.html#format1*), and suitable for loading into various genome browsers. If the estimated coverage of a design is 100%, this file will not be generated.

coverage_summary.txt

The global coverage properties of a KAPA Target Enrichment custom design. Use a text editor or spreadsheet software to open the tab-delimited text file. The following is a description of the fields included in the file.

Field	Description
Genome build	Genome and build targeted by the design (*e.g.* GRCh38/hg38).
Number of regions	Number of regions after consolidation.
Length of regions (bp)	Sum total of all region sizes (in base pairs) after consolidation. The base pair length of the consolidated regions. Used as inputs in the probe selection.
Probe_Coverage	Direct probe coverage of consolidated regions.
Estimated_Coverage	Probe coverage plus 100 bp of adjacent sequence to the probe coverage or up to the end of the consolidated region. This number is not accurate for libraries with much larger or smaller insert sizes.
Target bases covered	Sum of all bases from the consolidated regions that are covered (in base pairs) by at least one probe or by predicted captured sequence. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Percent target bases covered	Percentage of all bases from the consolidated regions that are covered by one or more probes. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Targets with no coverage	Number of consolidated regions with no captured sequence. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Target Bases Not Covered	Number of target bases in the consolidated regions that are not covered by any probe. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Target Bases Not Covered (due to N's)	Number of target bases in the consolidated regions that are not covered by any probe due to the source genome having N's or ambiguous bases within the target range. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Target Bases Not Covered (due to repeats)	Number of target bases from consolidated regions that are not covered by any capture due to the source genome having low complexity or highly repetitive DNA within the target range. Roche avoids selecting probes in regions of low complexity or high repeat content to reduce the chance of capturing off-target sequences. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Percent Target Bases Not Covered	Percentage of target bases from the consolidated regions that are not covered by any probe. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Percent Target Bases Not Covered (due to N's)	Percentage of target bases from consolidated regions that are not covered by any probe due to the source genome having N's or ambiguous bases within the target range. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Percent Target Bases Not Covered (due to repeats)	Percentage of target bases from consolidated regions that are not covered by any probe due to the source genome having low complexity or highly repetitive DNA within the target range. Roche avoids selecting probes in regions of low complexity or high repeat content to reduce the chance of capturing off-target sequences. Calculations for Probe_Coverage and Estimated_Coverage are provided.
Total capture targets	Total number of regions in the capture target files. This may be different from the number of regions above. If the coverage of a target has a gap in the probes, it will be considered two regions rather than one. If two regions are close enough that probes are selected across the gap of the two regions, it will be considered a single region.
Total capture space (bp)	Total number of bases covered by the capture targets. This can be very different from the primary target space, and provides an idea of the total amount of sequencing that will be needed for each sample. Use this size for categorization of panel capture target size in Chapter 5 of *KAPA HyperCap Workflow v3.0 Instructions for Use*.

coverage.txt

A tab-delimited text file displaying region-by-region coverage information of the consolidated input regions of interest. The headers are as follows.

Header	Description
REGION_NAME	Customer-provided name from the 4th column of their input BED file(s). If no 4th column was provided, then a default name of the selection region will be used instead. This name takes the format of CHROMOSOME:START-STOP.
CHROMOSOME	Target chromosome, or sequence identifier, for the region.
START	Region start coordinate.
STOP	Region stop coordinate.
LENGTH	Length of the region.
BASES_PROBE_COVERAGE	Number of bases in the region that are directly covered by a capture probe.
FRAC_PROBE_COVERAGE	Fraction of the region that is covered using direct coverage. A value 1.000 means that every base of the target is covered by one or more capture probes. A value of 0.460 means that 46% of the region is covered by one or more capture probes.
BASES_ESTIMATE_COVERAGE	Number of bases in the region directly covered by a probe or by indirect/adjacent coverage. This is an estimate of the actual amount of sequence that may be captured by a capture probe, determined in empirical tests, reflecting that capture probes may hybridize to the end of library insert and extend coverage away from the probe. The 100 bp capture padding was validated with Illumina paired-end sequencing, using a typical library size of ~200 bp. This number may not be accurate for libraries with much larger or smaller insert sizes, or single end reads.
FRAC_ESTIMATED_COVERAGE	Fraction of the region that is covered including indirect/adjacent coverage. A value 1.000 means that every base of the target is covered by one or more capture probes. For example, a value 0.982 means that 98.2% of the target is covered directly or indirectly by one or more capture probes.
PREDICTED_NO_COVERAGE_BASES	Number of bases in the region that are not likely to be captured.
BASES_W_NO_PROBE_COV	Number of bases in the region that are not directly covered by a capture probe.
BASES_W_NO_PROBE_COV_DUE_TO_N	Number of bases in the region that are not covered directly by probes due to the region containing ambiguous bases in the source. Roche cannot design probes against sequences containing non-ACGT characters.
BASES_W_NO_PROBE_COV_DUE_TO_REPEATS	Number of bases in the region that are not covered directly by probes due to the region containing low complexity or highly repetitive sequence. Roche avoids selecting probes in regions of low complexity or high repeat content for the purposes of reducing off-target sequencing results.
BASES_W_NO_EST_COV	Number of bases in the region not directly or indirectly covered by a probe.
BASES_W_NO_EST_COV_DUE_TO_N	Number of bases in the region that are not covered directly or indirectly due to the region containing ambiguous bases in the source.

Header	Description
BASES_W_NO_EST_COV_DUE_TO_REPEATS	Number of bases in the region that are not covered directly or indirectly due to the region containing repetitive sequence(s).

Published by:
Roche Sequencing Solutions, Inc.
4300 Hacienda Drive
Pleasanton, CA 94588

HYPERCAP, HYPERDESIGN and KAPA are trademarks of Roche.
For Research Use Only. Not for use in diagnostic procedures.

©2020 Roche Sequencing Solutions, Inc. All rights reserved. MC--03807