Custom Targeted Design Workflow for Next Generation Sequencing

B Marosy, DW Mohr, AF Scott, KF Doheny

Johns Hopkins Genomics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD

Introduction

Genome Wide Association and linkage studies that identify a chromosomal region(s) may require additional follow-up sequencing in order to find a causal variant(s) that contributes to the genetic trait. Targeted sequencing is a powerful tool for querying regions of interest in large sample cohorts at high depth for relatively low cost compared to whole exome or whole genome sequencing. However, inefficient design of custom probes across regions of homology, high GC content and repetitive elements can cause a decrease in selection metrics which lead to increased sequencing costs, rendering custom capture as a more expensive option. By utilizing a rigorous design workflow, these challenges can be addressed prior to manufacturing of a custom product, aiding in reducing unforeseen sequencing costs. Here we have developed a workflow for designing custom targeted panels and present cases utilized in optimizing the workflow parameters and potentially improve sequencing quality.

Target Design Workflow

Figure 1 is a visual description of the workflow process. Regions of Interest (ROIs) are obtained from the Investigator (dependent on the research and design of the study) and formatted as genomic coordinates to input for bait design. Types of regions may include:

- Regions based on linkage or association (inclusive of several genes)
- Whole gene
- Gene exons
- Addition of flanking regions – padding to exons, UTRs and promoter regions
- Addition of SNPs

The UCSC genome browser (http://genome.ucsc.edu) provides tools to convert gene names into genomic coordinates, extraction of exon coordinates for genes, and genomic elements of interest are identified for inclusion in the design. Using tools in GALAXY (http://galaxy.psu.edu) allows further manipulation of the bed files for the inclusion of flanks, extraction of exon, UTR regions; merging and collapsing bed files; and determining base coverage.

Design strategies need to account for several factors:

- Regions of masking (using stringency filters) which can cause low percent selection in the hybridization reaction.
- Regions of homology which can cause mappability problems.
- Boosting options (addition of probes across regions of high GC) in conjunction with stringency may help with uniformity (i.e. re-balancing)
- Both masking and homology will affect coverage of the ROI.

The Agilent Sure Design Software (https://earray.chem.agilent.com/suredesign/index.htm) was used to design probes across regions of interest in a tiered fashion (3 passes), reducing stringency parameters for each subsequent pass of uncovered target regions not covered by a probe in a previous pass (**Figure 1**). Boosting options are also varied in conjunction with the masking parameters selected in order to improve uniformity. Using default or low stringency settings in SureDesign may increase coverage across target regions, but with the inclusion of probes regardless of homology.

The use of BLAT (http://www.kentinformatics.com/index.html) against the probes designed will help to identify probes that are targeting regions of homology and repeats. However, BLAT will not account for regions with high GC/AT content. BLAT identifies sequences of ≥95% similarity, in which threshold settings (min score) within BLAT can be scaled to either increase or decrease stringency according to project needs, either increasing coverage or balancing sequencing costs. Combinations from each pass w/wo BLAT filtering can be generated to scale a design from highest coverage to highest stringency.

Summary Statistics of Coverage for each design are then generated using the Coverage tool in GALAXY. Each region of interest is compared to the probes generated to determine the base coverage for each region. Regions with lower probe coverage can be identified quickly for further manual review.

Case 1:

This design was aimed at covering 5 large contiguous regions (3.56Mb). Prior to optimization the SureDesign tool was used to generate the design using least stringent parameters with balanced boosting. Sequencing QC metrics (**Table 1**) of the project samples showed the percent selection was low (21%). Subsequently this project required 50% more sequencing in order to meet OnTarget depth requirements. This identified the need to further optimize custom probe design to identify problem regions prior to manufacturing of a custom panel. Due to the cost of custom capture, typically designs are not screened experimentally prior to ordering. This can be a daunting task, as large custom captures and large sample sizes dictate the need to ensure some amount of upfront expected performance.

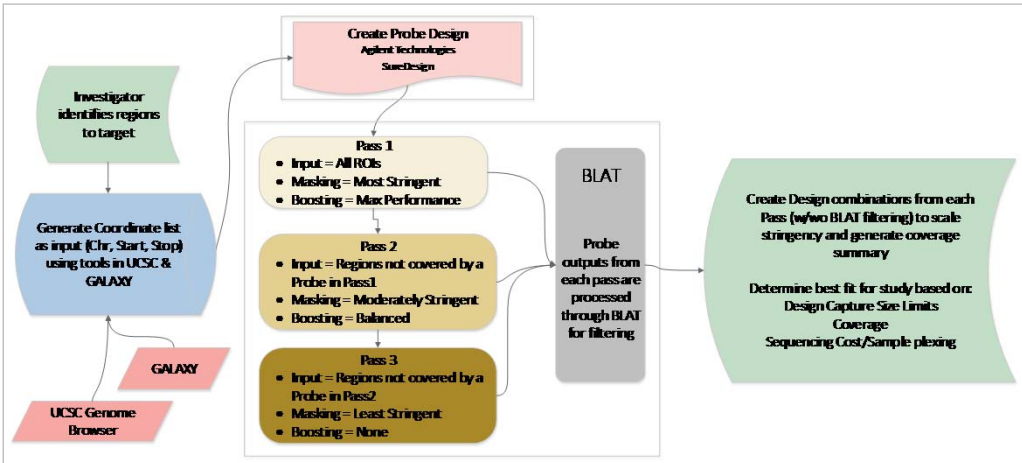


Figure 1 – Optimized Custom Target Design Workflow

Table 1. Average sequencing metrics from Case1 (n=850) that identified need to optimize design Workflow.

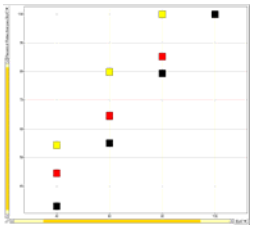
	Case 1
Mean Coverage	85
% OnTarget @20x	97.14
Raw Gb	**1.32**
Expected Gb/sample	**0.37**
Capture Size (Mb)	2.99
% Selection	21.48
Samples plexed /lane	**48**
Expected Sample plexed/lane	**80**

Retrospectively, the design was redone using the optimized BLAT workflow for SureDesign to determine how many regions would have been flagged/filtered (**Table 2**). Utilization of this workflow aided in confirming the decrease seen in percent selection. **Figure 2** shows tracks in UCSC and corresponding IGV plot of regions that were identified from the BLAT analysis as problematic for sequencing.

Table 2 – Case 1 redesign using optimized workflow, 24.53% of the total probes would have been flagged and filtered if the optimized workflow w/BLAT was used with the original design.

Tier	Total Probes	Good Probes	Probes filtered by BLAT	% filtered by BLAT
Pass1	11909	11610	299	2.51%
Pass2	11127	9967	1161	10.43%
Pass3	11717	4652	7065	60.30%
Total	34753	26229	8525	**24.53%**

Figure 3 - results show that as the threshold of BLAT is decreased (100 -> 40), the percent of probes that pass also decreases. Indicating that the lower the BLAT threshold the more stringent the filter. Data colored by 3 different custom designs using the only the 3rd pass and altering the threshold of BLAT. The x-axis is the threshold of BLAT used; y-axis is the percent of probes were retained after BLAT filtering.



Case2:

Using the optimized workflow, two designs were generated from the same ROIs and probes were manufactured and tested experimentally to compare different stringency levels used during the design.

- Design 1 included 3 passes w/o any BLAT filtering (Least Stringent Design). 6.5% of probes are flagged by BLAT in Pass 3 but were included in the manufacturing.
- Design 2 included 3 passes with BLAT/40 filtering on each pass (Most Stringent design). 12% of the probes were filtered out based on BLAT.

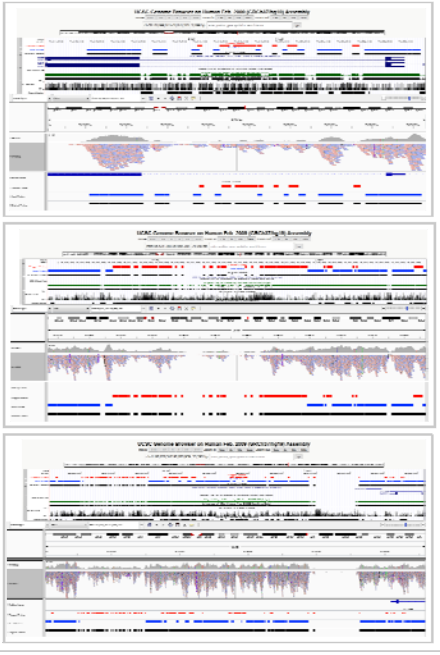
Discussion/Conclusion:

Applying a tiered probe design in conjunction with BLAT allows users to flag and filter probe designs to better suit the needs of a project (taking into account capture size limits, depth requirements and plexing capacity) with more insight as to the expected performance of the design prior to manufacturing. This will enable improved management of lab resources as well as better prediction of sequencing costs for each custom project. Future applications include streamlining of custom clinical panels to aide in the removal of poor performing probes to streamline costs.

Table 3 compares sequencing metrics between the two designs from Case 2. Library prep and capture were performed on 2 HapMap and 14 experimental samples, using standard Agilent processing. Libraries were then clustered and sequenced on the Illumina® HiSeq™ 2500 platform using on-board clustering and Rapid Run SBS chemistry, 2x100. Subsequent data analysis was performed using CIDRSeqSuite v6.0 (http://www.cidr.jhmi.edu/next_gen_seq_serv/sequencing_qc_pipelines_workflow.pdf). The least stringent design (Design 1) had a much higher SNV Count than the more stringent design (Design 2) suggesting a higher rate of false positive variant calls.

	Design 1	Design 2
Stringency	Least Stringent	Most Stringent
% Design Coverage	84	73
Capture Size (Mb)	1.852868	1.627321
% Selection	**60.91**	**66.49**
Raw Gb	0.6756	0.6964
PF_UQ_GIGS_ALIGNED	0.6102	0.6359
Percent of Usable Gb	90.32	91.31
Mean Coverage	180	227
%OnTarget @ 20x	98.08	99.76
% OnTarget @ 30x	96.52	99.58
Uniformity	**92.31**	**95.32**
Sensitivity2Het	99.4078	99.7984
SNV Count	**6043**	**2860**
PERCENT_SNV_ON_BAIT_SNP138	**91.9025**	**97.4518**

Figure 2 – Case 1 probes filtered by BLAT shown in red, Probes flagged as 'good' shown in blue, the original design probes shown in black. IGV plot of sequence reads generated from the original design. Regions of lower coverage correlate to regions that would have been filtered by BLAT.



References:

Kent, W.J. 2002. BLAT -- The BLAST-Like Alignment Tool. Genome Research 4: 656-664

Acknowledgements:

The authors would like to thank Agilent for their technical support.