DAICEL arbor biosciences

Submission Guidelines for myBaits® Custom Designs

To submit your sequences for myBaits Custom DNA-Seq, RNA-Seq, or Methyl-Seq panel design, we accept sequences in either FASTA format (below) or as coordinates from a reference genome (page 2).

Please note that we will design baits from ALL sequences or coordinates that you provide. If you only want specific regions of those sequences in the baitset (e.g. exons only), please first curate your targets to only include those specific regions of interest.

I. SEQUENCES

Acceptable: ***FASTA DNA sequence format, combined into one (1) plain text file***

Names

- All sequence names must be fully unique
- Allowed characters are letters, numbers, and dashes “-” ONLY (no other characters should be used)
- Spaces will be replaced with dashes
- Name length 50 characters or less
- Recommended to incorporate species/locus names

Sequences

- Allowed characters are IUPAC bases
- Alignment gaps ("-") may be present, but will be ignored during bait design
- Minimum target sequence length should correspond to bait length; targets shorter than bait may be omitted.

Note regarding non-ATCG bases: Singleton and/or short stretches of N's will be replaced with T's to facilitate bait design in these regions. Longer stretches (e.g 10+ N’s) will be skipped over during bait placement. Ambiguities (e.g. Y/M/R/S/W/K) are allowed, but will be replaced by ONE random candidate base for manufacturing, since we only synthesize A/T/C/G bases. Sequences that contain on average > 5-7% ambiguous bases are not recommended. If this is a consensus sequence made from a common source (*e.g.* the same gene from multiple genomes), please provide the original individual sequences.

Additionally, we would like a link to up to 5 reference genomes (or close relatives). Part of our pipeline is to look for non-target matches elsewhere in the genome that may be over-represented.

We don’t recommend including baits for both nuclear and organellar genomes in the same design, due to the significantly higher copy number between the mitogenome/plastid genome compared to the nuclear genome.

(next page for **GENOME COORDINATES**)

II. GENOME COORDINATES

Acceptable: ***BED ("Browser Extensible Data") format, in one (1) plain text file***

BED file format details available at: https://genome.ucsc.edu/FAQ/FAQformat.html#format1

- Format:
    - Contiguous target sequences (genes, contigs, etc):
        - Column 1 = chromosome/scaffold name
        - Column 2 = start coordinate
        - Column 3 = end coordinate
        - Column 4 = name to assign sequence (optional)
    - SNP targets:
        - Column 1 = chromosome/scaffold name
        - Column 2 = SNP coordinate
        - Column 3 = name to assign sequence (optional)
- Provide link/copy of exact reference genome, otherwise coordinates will be incorrect
    - If the genome is unpublished, we will keep it private (we can sign a Non-Disclosure Agreement upon request)
- Names of chromosome/contig/scaffolds must match genome entry names exactly
- Plain text file only (tab-delimited)
- Do not submit spreadsheet/excel files
    - If using Excel, please export as ".tsv" file

VCF files may be acceptable for SNP targets, but require discussion with a member of our design team

Contact your myBaits representative for instructions on how to submit your prepared files. Thank you!