Input data formats for rhAmpSeq assay design

Thank you for choosing the rhAmpSeq Custom Design Service. The type and quality of your target information will have a direct impact on both your Custom rhAmpSeq Panel design and your sequencing results.

Please indicate or provide a reference genome for designs. High quality reference genomes improve *in silico* design quality control.

Please contact CustomrhAmpSeq@idtdna.com with any questions.

Sequence preprocessing: Distance between targets and target size

Two parameters for rhAmpSeq target selection must be met before designing the rhAmpSeq panel. The first is the required minimum distance between targets—600 bp. The second is target size (sometimes termed "insert size" or "amplicon size")—between 50 and 200 bp. Input data files will be preprocessed (i.e., include, exclude, or merge) to ensure workable target distance and size parameters.

Figure 1 illustrates an example of our preprocessing target analysis. In this example, Target_1 will be included in the rhAmpSeq design because it does not have any distance constraints and does not exceed the maximum target size.

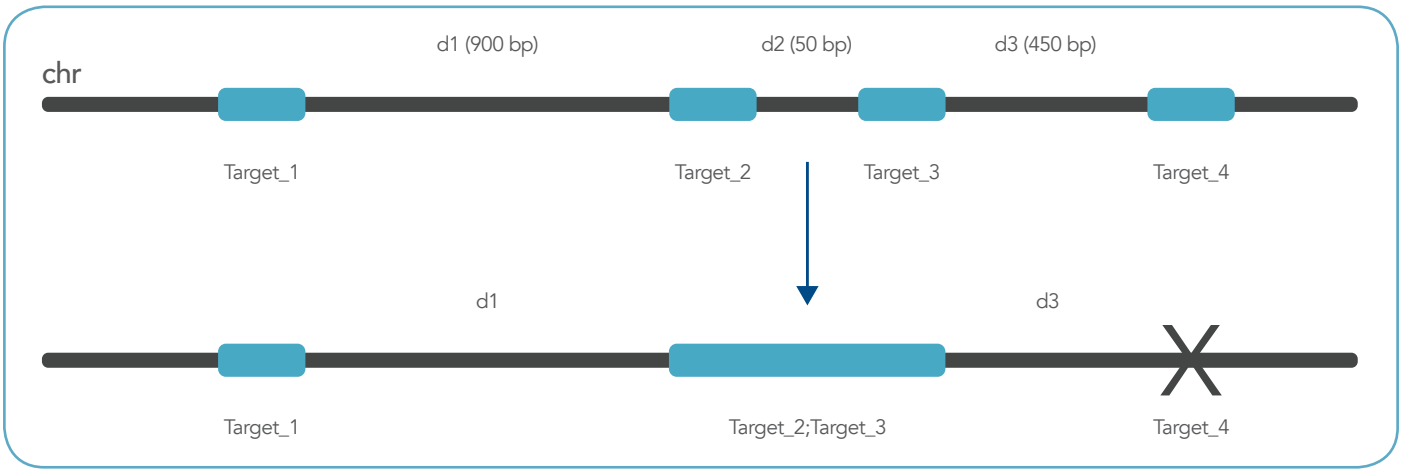

Figure 1. Preprocessing target analysis. The original targets (Targets 1–4) had different distances between them (d1 = 900 bp, d2 = 50 bp, and d3 = 450 bp). After preprocessing, the original 4 targets became 2 targets for rhAmpSeq design. Target_1 did not change because it met the size and distance criteria. Target_2 and Target_3 were merged into a single new target (Target_2;Target_3). Target_4 could not be included for rhAmpSeq design because it was less than 600 bp from Target_2;Target_3.

Adjacent targets separated by less than 60 bp can be merged to generate a new, combined target. For example, Target_2 and Target_3 will be merged into a single, new target (named Target_2;Target_3). However, if the new merged target size exceeds the maximum insert size of 200 bp, the targets will not be merged.

Finally, because Target_4 is 450 bp from Target_2;Target_3, it cannot be merged with "Target2;Target3" (merging distance is >60 bp), and it cannot be included in the rhAmpSeq design (distance between targets is <600 bp). Target_4 is therefore excluded from the rhAmpSeq design.

Data input formats

Option 1: Target in BED file with genome information (preferred)

- **BED file for target of interest** (not including flanking regions)

  The following fields are required:

  chrom: The name of the chromosome (e.g., chr1, chr2)

  chromStart: The starting position of the target in the chromosome

  chromEnd: The ending position of the target in the chromosome

  name: The unique name of the target

- **Genome information**

  It is important to provide the genus, species, and genome build used to generate the BED files (e.g., *Homo sapiens* GRCh38).

- **Example input**

```
chr11   34614227        34614228        rs10488741      0       +
chr11   87518023        87518024        rs7926017       0       +
chr11   88713708        88713709        rs6483391       0       +
chr11   124578234       124578235       rs76543769      0       +
chr12   18867007        18867008        rs10841100      0       +
...
```

Option 2: Target embedded in FASTA format sequences

- **Target SNP**

  Target SNP should be denoted in bracket format with the reference sequence base first and the variant sequence base second (e.g., [A/T], where A is the reference base and T is the variant base).

- **Surrounding sequences in FASTA format**

  Surrounding sequences around target SNPs are used for primer designs.

  Optimal: at least 200 bp upstream and downstream of target SNP.

  Minimum: 100 bp upstream and downstream of the target SNP.

- **Genome information, if available**

  It is beneficial to provide the genus, species, and genome build used to generate the FASTA file. The genome will be used for *in silico* design quality control.

- **SNP and/or repeat information in surrounding regions** (optional)

  SNP and repeat information are used in quality checks for optimal designs.

  BED file (at least 4 columns) can be used for the surrounding SNP and repeat information. The BED coordinates should be relative to the FASTA (not genome) sequences.

- **Example input**

```
>SNP_1
TCGGGGTTCACGAGCTTGCTCTCCTCTGCCATTCGCAATGGGCTGGCATAGGGCAGCCTCACTGCTTGGCTCCAGCCAGCGACTT
CAGGATGTGGCGATGAAGATGGGGTCTGGATGGGCGTTGGTATTCAGGATGTCGAGCCCCACTGGTCTGGGGGACCAGTTGGCGC
TTCCTGCAAAGGCATGTGCTCTGGGAAGGG
[A/T]
CTGGCCTGCTGCAGCAGCTCTGCAGAGGGGCCGGCCCCAGACTTGCCTATGTCATTAGTATAGCAGGTCCCGTTATTACCCGAGG
AGGCTTACAGGTTATCAGCGAGCTCCAGGAGCCactggagaaaggaagaagataaagaaggatttaaaaagaaaataacaaaaag
aaaaactgtattttctaATCCAAACCTTGC
>SNP_2
ATTGCCTTTCTGTGGAGCAAGGGGTGTTGTACACACAAGCCTCACTGTAGacactgcctcagtttccccataggCATAATGGGTC
CCTTCTAGTTCAGGCAATCTGGATTTGATCTTGAGTTCCAGTGCCAGCCTCTGGAGTCACTCCATTTTCATACCTTTTCATGATC
TCAGGGGCTCTGGGCAGTGGGAGGTGATGG
[C/G]
TTGGACAGATTCTTGGTCATGCTCCCCAACTCTTGGTGGCTCACCACTGAACACTCCAAACCCTGCTTAAAGAAGTTGATCTATC
TGAAAGCCAGGGTAAAGATTGCTAAGGCTTGTCTCCTCTCCCAGTGGGAAGAGAGAGGTTCTGTTGGTGTCCTGGTTGAATTGCT
TTGCAGAGAAGTCAATGCCCATCACCCTTG
...
```

Option 3: Target in BED file + reference sequences (no genome information)

- **Target SNP in BED file**

  Required: Target SNP position in the FASTA sequence.

  The coordinates should be relative to the FASTA sequence name.

- **Surrounding sequences in FASTA format** (reference sequences)

  Surrounding sequences around target SNPs are used for primer designs.

  Optimal: at least 200 bp upstream and downstream of target SNP.

  Minimum:100 bp upstream and downstream of the target SNP.

- **Example input**

```
Target bed file:

Seq1   127        128        target_snp1
Seq2   230        231        target_snp2
...
```

```
Sequences in FASTA format:

>Seq1
TCGGGGTTCACGAGCTTGCTCTCCTCTGCCATTCGCAATGGGCTGGCATAGGGCAGCCTCACTGCTTGGCTCCAGCCAGCGACTT
CAGGATGTGGCGATGAAGATGGGGTCTGGATGGGCGTTGGTATTCAGGATGTCGAGCCCCACTGGTCTGGGGGACCAGTTGGCGC
TTCCTGCAAAGGCATGTGCTCTGGGAAGGGaaaaactgtattttctaATCCAAACCTTGCCTGGCCTGCTGCAGCAGCTCTGCAG
AGGGGCCGGCCCCAGACTTGCCTATGTCATTAGTATAGCAGGTCCCGTTATTACCCGAGGAGGCTTACAGGTTATCAGCGAGCTC
CAGGAGCCactggagaaaggaagaagataaagaaggatttaaaaagaaaataacaaaaag
>Seq2
TTGGTCATGCTCCCCAACTCTTGGTGGCTCACCACTGAACACTCCAAACCCTGCTTAAAGAAGTTGATCTATCTGAAAGCCAGGG
TAAAGATTGCTAAGGTCTGGATGGGCGTTGGTATTCAGGATGTCGAGCCCCACTGGTCTGGGGGACCAGTTGGCGCTTCCTGCAA
AGGCATGTGCTCTGGGAAGGGaaaaactgtattttctaATCCAAACCTTGCCTGGTCTGGATGGGCGTTGGTATTCAGGATGTCG
AGCCCCACTGGTCTGGGGGACCAGTTGGCGCTTCCTGCAAAGGCATGTGCTCTGGGAAGGGaaaaactgtattttctaATCCAAA
CCTTGCCT
```


NGS-10129-BR 10/2018