Bioinformatic Evaluation of a Sequence for Custom TaqMan® Gene Expression Assays

Overview

The Custom TaqMan® Gene Expression Assays are custom assays that are designed, synthesized, formulated, and delivered as analytically quality-controlled primer and probe sets for gene expression assays based on sequence information submitted by the customer. The goal of this tutorial is to help the researcher evaluate the quality of their sequence information before submitting an order for a Custom TaqMan® Gene Expression Assay. Specific information is given on how to assess a sequence using a variety of on-line tools.

The Custom TaqMan® Gene Expression Assays provide the researcher the opportunity to design an assay that is not currently available through the TaqMan® Gene Expression Assays offerings. Studies that involve viral detection, species other than human, mouse, rat, *Arabidopsis*, *Drosophila*, *C. elegans*, canine or Rhesus macaque, or detection of specific pathogens are some examples of applications that would benefit from this custom design line of products. For gene expression assays of the species mentioned above, the TaqMan® Gene Expression Assays should be used. If a particular gene target is currently not available then one should consider a custom design.
***Note**: Additional TaqMan® Assays are regularly added to the web site for ordering. Please visit the TaqMan® Gene Expression Assays Search page for regular updates. To learn more about how to order assays, please download the Online Ordering Guide for TaqMan® Gene Expression Assays.*

Process Overview

Ordering Custom TaqMan® Assays involves the following procedures:

1. Selecting a target sequence
2. **Assessing the quality of the sequence**
3. Preparing the submission file using the Custom TaqMan® Genomic Assays File Builder software
4. Formatting the sequence for submission
5. Submitting the order via the File Builder software or e-mail.

Step two, **Assessing the quality of the sequence,** will be covered in this tutorial.

Step 1 and 3-5: Selecting a target sequence, Preparing the submission file, Formatting the sequence for submission, and Submitting the order are covered in:

- Custom TaqMan® Genomic Assays Submission Guidelines Protocol
- Ordering Custom TaqMan® Genomic Assays: Online Ordering Procedures Using the File Builder Software: Quick Reference Card
- File Builder Demo
- TaqMan® Assays-by-Design Service for Gene Expression Assays Quick Reference Card.

Assessing the Quality of the Sequence

Overview

The most important factor in the success of the Custom TaqMan® Gene Expression Assays is the quality of the sequence data that you submit for the design process. Sequence analysis gives one a tool to eliminate poor sequence quality so it does not adversely impact the performance of the assay. Following this section, a variety of on-line tools are presented to help assess your sequence. Consider the following when selecting your target sequence:

- Biological significance
- Sequence length
- Sequence quality
- Masking sequences
- Uniqueness of sequence

Biological Significance

When choosing sequences to submit, one should first consider the biological significance of the desired assay. The quality assurance on assays carried out during manufacture of the primers and probe can ensure only that the yield and content of the primers and probe meet specifications. Applied Biosystems is unable to guarantee the biological performance of the assays.

Examples:

- If you know that your gene of interest has more than one transcript (splice variants) make sure you are submitting a sequence that will detect all of the variants you wish to detect. On the contrary, if you only want to detect one out of five splice variants for a particular transcript, make sure that you have selected your targets (see *Note* below) appropriately, and masked any unwanted regions of that transcript to ensure that the assay you receive is specific only for your transcript of interest.
- If you are studying a gene that has regions of high homology to other members within a gene family, or to closely related genes, you will want to ensure specificity by using areas of sequence unique to the gene of interest and masking homologous regions with Ns.

Note: If you are studying the gene expression of a multi-exon gene, it is important to know the location of the exon junctions within the cDNA sequence that you submit for assay design. The ideal assay design is placement of the TaqMan® MGB probe across an exon-exon junction. The exon boundary information is used as a target in the sequence submission process for a gene expression assay.

Sequence Length

To optimize your assay design, follow these guidelines:

- Submit a sequence length of approximately 600 bases.
  Increasing the sequence length increases the assay design possibilities.
- Select the sequence so that the target site is toward the center of the submitted sequence.

***Note*:** Sequence length can range from 61 to 5000 bases.

Sequence Quality

To assess the quality of the sequence:

1. Obtain confidence in the sequence accuracy. You want to have the most accurate sequence of your desired target before you submit the sequence to have an assay designed. Inaccurate sequences can lead to failed assays due to poor annealing, or no annealing, of primers or probes.
   Note: *If you performed the sequencing yourself, it is strongly recommended that you perform multiple sequencing reactions to remove any ambiguities.*

2. Use other resources, such as public databases with curated sequences such as **RefSeq** (which contains mRNA sequences) or **dbSNP** (which contains documented SNPs) to determine the quality of your sequence.

Masking Sequences

The Custom TaqMan® Assays proprietary software for designing primers and probes will not design probes or primers to a region of sequence containing Ns. You can annotate your sequences with Ns to avoid specific regions of sequence in design (e.g., ambiguous sequences, repetitive sequences, or SNP sites), albeit the use of Ns may limit assay design.

To mask sequences:

1. You may substitute each ambiguous base with an N.

For example:

The **bolded** bases in this sequence are ambiguous:

ACGTGACGTGACGTGACGTGACGTGGAT**Y**GTG**RSRS**TCCT

Where Y= C or T, R=A or G, and S= G or C; they would be substituted as:

ACGTGACGTGACGTGACGTGACGTGGAT**N**GTG**NNNN**TCCT.

2. Minimize the substitution of Ns in the sequence.

Because the Custom TaqMan® Assays proprietary software does not include Ns in the probe or primer, having a sequence with Ns greatly reduces the number of available primers and probes from which to select an optimal assay.

3. Ensure that Ns are not too close to the target site.

Important! No probes can be designed if Ns are too close to the target site. When designing gene expression assays, make sure that there are no Ns within five bases of the target site.

Uniqueness of Sequence

After you have selected a sequence, check whether unique primers and probes can be generated for the cDNA sequence by verifying that the target sequence is unique within the organism you are studying.

1. Substitute Ns to mask small regions of repeats and SNPs. Run the sequence through a program such as **Repeat Masker** to detect common repetitive elements.

2. Perform a **BLAST** search against public databases to detect regions within your sequence that have similarity to other published sequences. If there are large regions of similarity with other sequences in a gene family, use a different area of sequence that is unique to your gene of interest.

3. For Gene Expression Assays, choose an exon-exon boundary that is unique for the transcript(s) of interest.

For Custom TaqMan® Gene Expression Assays, the TaqMan® MGB probe, when possible, should be designed across an exon-exon boundary in order to exclude the detection of genomic DNA. The exon boundaries are what will preferably serve as your target(s) in your submission file. If you are working with a gene sequence that is in a public database, there are web resources available to find exon information. One can search nucleotide databases using Vertebrate Genome Annotation (VEGA), which is part of the Ensembl project or Entrez at NCBI.

TOOLS

I. Repeat Masker

While the use of Ns limits assay design (see Masking Sequences), it allows you to eliminate possible assay design in areas of similarity to other unrelated sequences or to regions of low complexity sequence. Neither repeat elements nor low complexity DNA should be used as potential PCR primer or probe sites since they could produce non-specific amplification or probe binding.

On average, close to 50% of the human genomic DNA sequence will be masked by RepeatMasker. It is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). The masked sequence can be used for submission and can also be used in BLAST searches.

Examples of web sites that host RepeatMasker are:

http://www.repeatmasker.org
This website has a lot of useful information on the RepeatMasker program, including FAQs and documentation such as Interpreting Results, Sensitivity, and RepeatMasker uses. "RepeatMasker is most commonly used to avoid spurious matches in database searches. Generally this step is strongly recommended before doing BLASTN or BLASTX equivalent searches with mammalian DNA sequence."

http://woody.embl-heidelberg.de/repeatmask
This site is a mirror of the University of Washington site above. The repeatmask help on this site has similar information to that of the University of Washington.

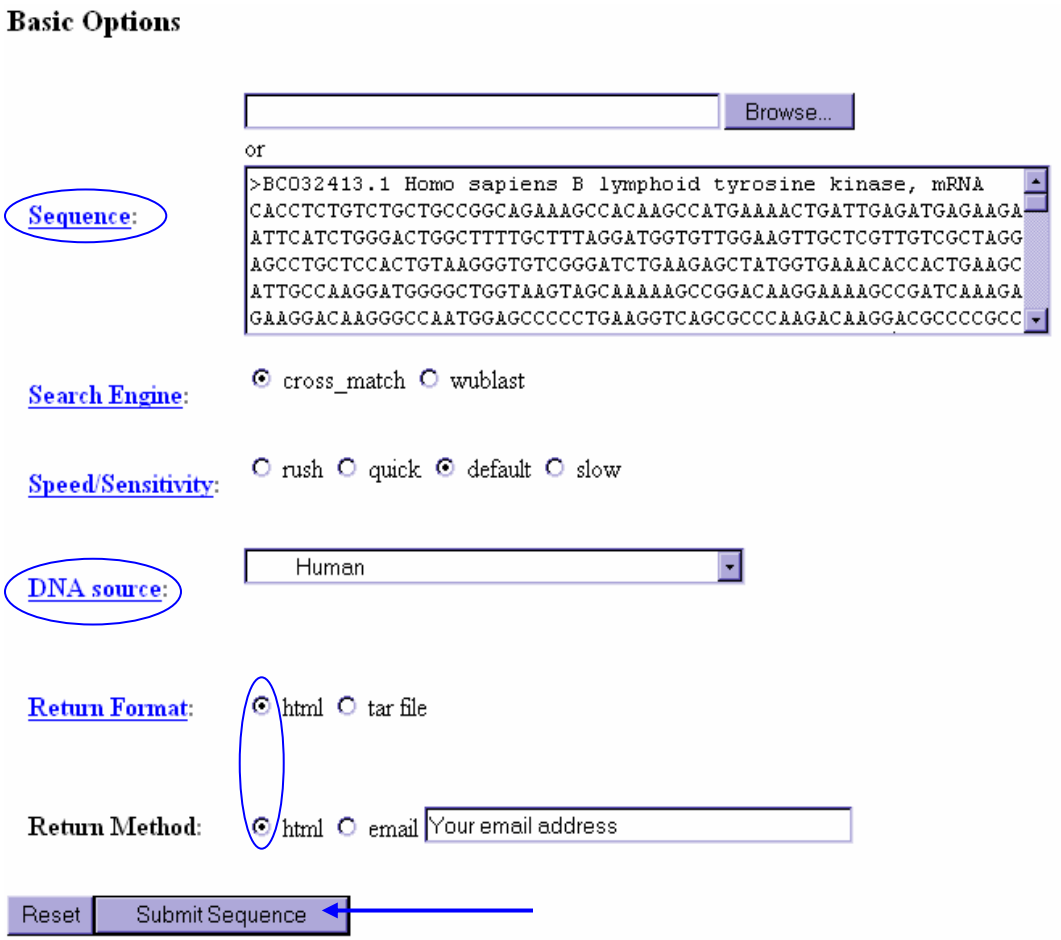
How to use RepeatMasker

A. Submitting your sequence / Starting your query

- You may enter your sequence by either copying and pasting your sequence into the box provided, or uploading it from a file.
- Sequences can be submitted one at a time or in batch form.
- Sequence submissions must be in FASTA format (see input format).
- Accepting the default setting of "html" for both 'Return Format' and 'Return Method' will allow for your results to be displayed in your web browser window.

- Make sure you choose the appropriate source of your DNA. The default genome library is human. Because interspersed repeats are specific to a (group of) species, it is important to select the appropriate repeat library to search.
- Click on 'Submit Sequence'.

RepeatMasker Submission



B. Viewing your Results

- RepeatMasker returns the submitted sequence(s) with all recognized interspersed or simple repeats masked. In the masked areas, each base is replaced with an N, so that the returned sequence is the same length as the original.
- A table annotating the masked sequences as well as a table summarizing the repeat content of the query sequence will be returned to your screen. In the "html" return format all output is returned to your screen in one file.
- The masked sequence can be copied directly from the web browser.
- We strongly recommend that when any sequence is submitted for a Custom TaqMan® Assay, the sequence be masked for repeat elements. This will reduce the possibility of poor sequence quality impacting assays.

RepeatMasker Output

Summary:

```
==================================================
file name: RM2sequpload_1167403373
sequences:                1
total length:          2251 bp  (2251 bp excl N/X-runs)
GC level:             58.29 %
bases masked:           200 bp (   8.88 %)
==================================================
Repeat Elements  number of      length   percentage
                 elements*    occupied  of sequence
--------------------------------------------------
SINEs:                  1          91 bp     4.04 %
      ALUs              0           0 bp     0.00 %
      MIRs              1          91 bp     4.04 %

LINEs:                  1          56 bp     2.49 %
      LINE1             0           0 bp     0.00 %
      LINE2             0           0 bp     0.00 %
      L3/CR1            1          56 bp     2.49 %

LTR elements:           0           0 bp     0.00 %
      MaLRs             0           0 bp     0.00 %
      ERVL              0           0 bp     0.00 %
      ERV_classI        0           0 bp     0.00 %
      ERV_classII       0           0 bp     0.00 %

DNA elements:           0           0 bp     0.00 %
      MER1_type         0           0 bp     0.00 %
      MER2_type         0           0 bp     0.00 %

Unclassified:           0           0 bp     0.00 %

Total interspersed repeats:       147 bp     6.53 %


Small RNA:              0           0 bp     0.00 %

Satellites:             0           0 bp     0.00 %
Simple repeats:         1          53 bp     2.35 %
Low complexity:         0           0 bp     0.00 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

Number & Percentage of bases masked

In this example there is a stretch of sequence that is comprised of 91 bases of MIR sequence, a common repeat element. If a TaqMan® primer or probe were designed across this MIR sequence (because it was not masked before submission) the oligo could bind to any MIR sequence in the genome. This assay would not be very discriminating or specific because of the number of sequences to which it could potentially bind.

Results

Right-click and select "Save As" to save results to your computer or click on the link to view the file in the browser.

Annotation File: RM2sequpload_1167403373.out

Masked File: RM2sequpload_1167403373.masked

Masked sequence

* Masked Sequence:

```
>BC032413.1 Homo sapiens B lymphoid tyrosine kinase, mRNA
CACCTCTGTCTGCTGCCGGCAGAAAGCCACAAGCCATGAAAACTGATTGA
GATGAGAAGAATTCATCTGGGACTGGCTTTTGCTTTAGGATGGTGTTGGA
AGTTGCTCGTTGTCGCTAGGAGCCTGCTCCACTGTAAGGGTGTCGGGATC
TGAAGAGCTATGGTGAAACACCACTGAAGCATTGCCAAGGATGGGGCTGG
CGACCCCGACTTCCGTGCCATCCCAGACGGGCCGCGAAGGCGGGGTGTCG
CCTGTGCCCTTTTCTCAGACCCGGAATCCAGTGGGCAGAGGCAGCTTCGC
AGGGGGTCCCCGGACGGACTCCTTCACCGACNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNGCCCCAGTAAGGTGTTCAGGACTGGTAA
GCGACTGTCATCAAGTAAGGCCCCCGTGCTGGGCACCCCCCGTGCTGGCC
GCGTCCCCGCCTCTGCGCCCTGCGTGGACCCCGCCCTGCCCCGCTACAGA
AGCCAGACTGGGTCCCGCGGACGCCAGCAGGGGCAACCCCAGCCTAGGCT
GCGCTCCAGCACTGCGGGGCTTTTCTGCAATAAAGTCACGAGCGTTCGNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Any repeat regions are automatically converted to Ns in the submitted sequence.

* Sequence was shortened for display purposes.

II. BLAST (Basic Local Alignment Search Tool)

Whether you have sequenced your target or taken the sequence from a sequence database, it is important to determine whether unique primers and probes can be generated for the sequence. Homologs in gene families can present a problem, as can orthologous sequences when working in a transgenic system. It is also important to identify any polymorphisms in your sequence of interest. All of these possibilities should be considered before submitting a sequence for a Custom TaqMan® Assay design.

To do this, you can compare your target sequence to databases of sequences and search for regions of sequence similarities. In order to make your assay as specific as possible, regions of similarity or polymorphism sites can be masked out before submitting your sequence for design, so they are not considered in the assay design. The National Center for Biotechnology Information (NCBI) hosts a database of all published nucleotide sequences, and a database of known sequence polymorphisms. BLAST, a sequence comparison algorithm, is available to facilitate searching of the NCBI public databases.

A. How to use BLAST to search for Sequence Similarity

This section describes the use of BLAST to search the NCBI nucleotide database for sequences similar to your sequence of interest.

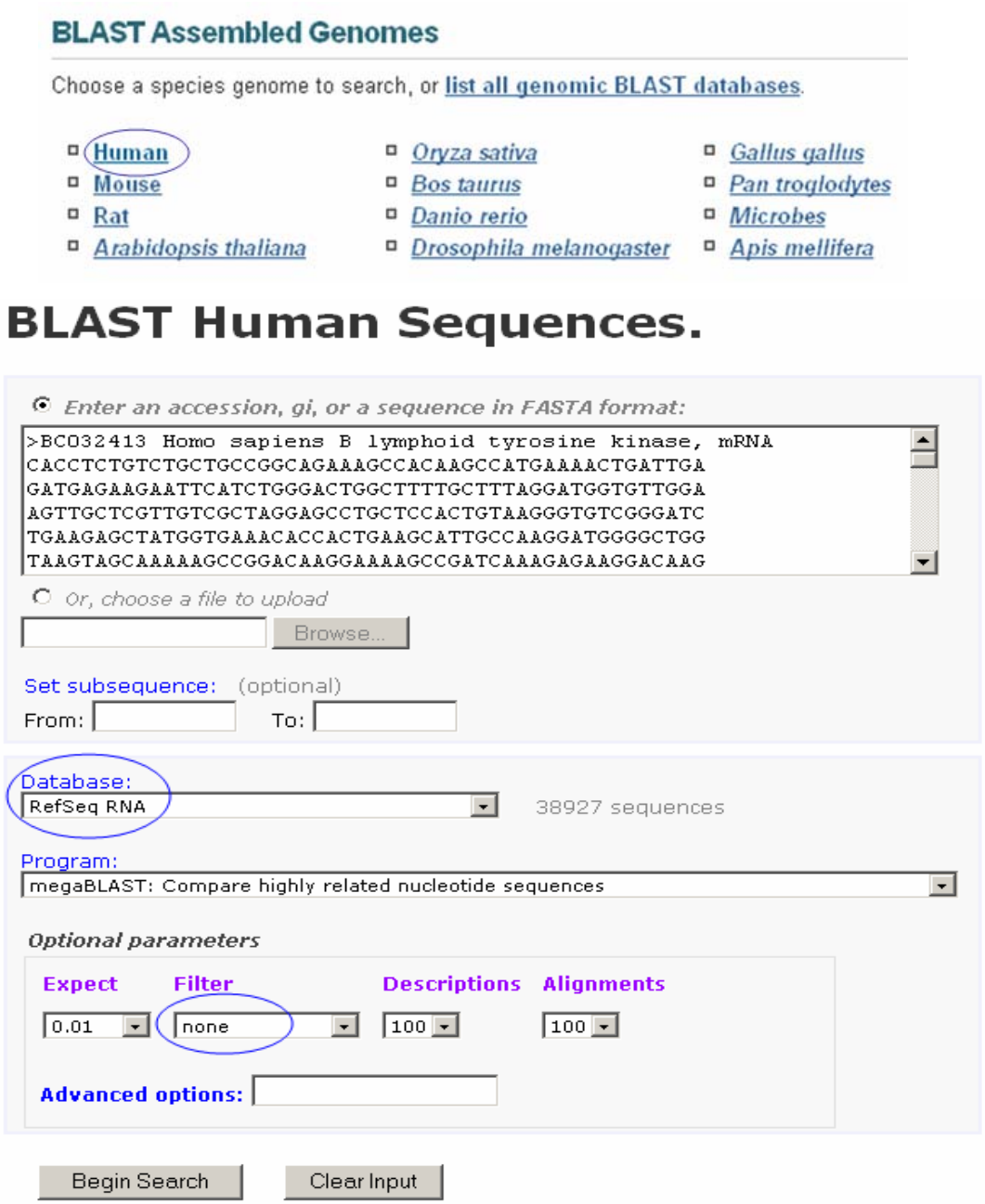
1. Submitting your sequence / starting your query

- Go to the NCBI BLAST site
- Choose your species under "BLAST Assembled Genomes. In the following example, we have selected Human.
- You may choose to BLAST some or all of your cDNA sequence. If you are only interested in a particular region of a transcript, then choose about 300 – 600 bases in that area to BLAST. If you are not sure about where you want the assay located, or you want options, then you may want to BLAST the whole cDNA sequence (masked output sequence from RepeatMasker) to find the best exon boundaries with which to work.
- Enter your masked sequence into the box provided. There are three sequence formats that may be entered into this box. (See pg. 8) For more information on this, click on above the box.
- Choose the appropriate database to search. Here we have chosen the "RefSeq RNA" database.
- Change Filter to "none" since the sequence used has already been masked in RepeatMasker.
- Click on "Begin Search" to submit your search.

2. For more information on how to use BLAST

NCBI has extensive help documentation on the NCBI BLAST website. This includes FAQs and Tutorials. Included on the Tutorials page are also an Introduction to Similarity Searches and a Glossary of Terms.

BLAST Submission



3. **BLAST Results**

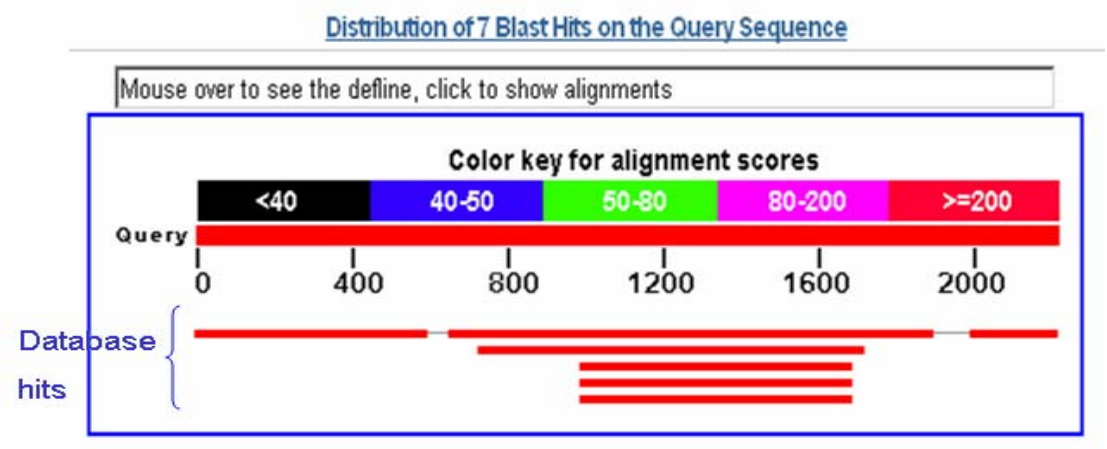
   There are three general parts to BLAST results:

   a. Graphical overview
   b. List of Sequences producing significant alignments to your query
   c. Sequence alignments.

   These sections are described below (p9–11) to give you a better understanding of what information can be obtained from a BLAST search of the NCBI public nucleotide database.

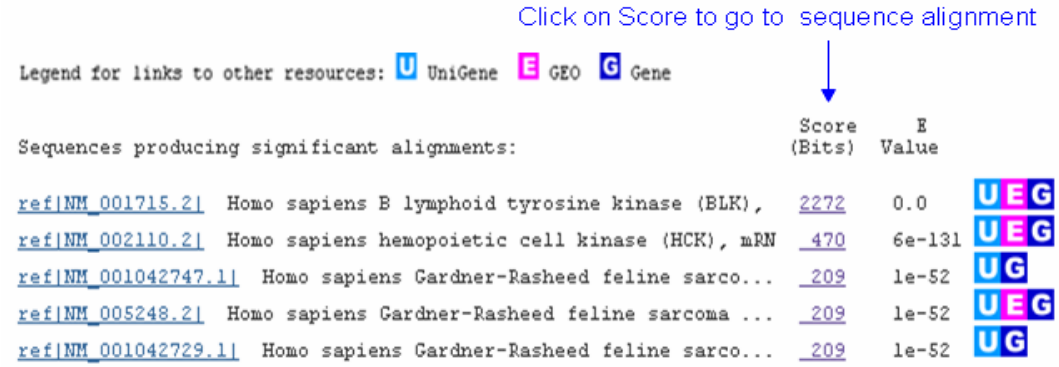
a. Graphical Overview

The graphical overview, as seen below, is a representation of the database sequences (hits) that align to your query sequence, with the query sequence represented by the thick red numbered line at the top of the graph. The color of the line represents the score of the alignment, and a striped line connects multiple alignments to the same database sequence.



b. List of Sequences producing significant alignments to your query

The list of sequences is shown from best to worst alignment; the top hit being the best hit (and possibly the sequence with which you queried the database). Public ID information is available as hypertext to the GenBank records that align to your query sequence, as well as a sequence definition. Clicking on the Max score hypertext will take you to the actual sequence alignment. The score reflects the degree of similarity between your sequence and the sequence to which it is being aligned. The higher the score is, the more similar the sequences. You should also be able to understand the E value in order to evaluate the significance of a particular result. The E value represents the number of hits one can "expect" to find by chance when searching a database of a particular size. In this case, the database is the NCBI database that you searched. The lower the E value is, the more significant the match. Hits with E values higher than around 0.1 are unlikely to be very significant.



By just browsing a list of hits one can get a good idea of the types of sequences that have been found to have some identity to your query. Notice that the first sequence in the list is the transcript that was used for the search (in this example, BLK). The Max score is very high (2272), and the Expect value is 0. The closer an E-value is to "0" the more "significant" the match. Remember that what you're looking for is the ability to design an assay that will uniquely detect your sequence of interest, whether it is a unique gene sequence or a unique splice variant. If you find some regions of similarity between your sequence and another, those bases can be masked out in your submission so that they will not be considered for assay design.

c. <u>Sequence Alignments</u>

The Sequence Alignment section displays your query sequence aligned to every sequence on your list of hits. These alignments are to help assess the degree of similarity. The Score and Expect values are displayed underneath the sequence identifiers. The number of bases aligned and percent identity are shown, as well as the strand that was aligned of your query sequence and the database hit.
You'll notice that the first hit in this list above is NM_001715.2; a RefSeq that is, in part, based on the sequence (BC032413) that was used for the search. This is the first alignment shown, and is a 99% match to the query sequence. It is not 100% because there are few single base mismatches between the query sequence, BC032413, and the RefSeq NM_001715.2.

Shown on the next page is a portion of the three alignments from the first hit, NM_001715.2, as compared to the query sequence (sequence used to BLAST the NCBI RefSeq RNA database).

In this example, notice that the query sequence did not align to this database hit contiguously. There are 3 alignments for the first hit. The first alignment is from base 1 of the query sequence to base 592; the second alignment is from base 649 to base 1881, and the third is from base 1973 to base 2198. This is because the sequence used for the search (the query sequence) had masked bases in it, and the gaps represent where the masked regions of the query sequence exist.

If a segment of your query sequence came up with a significant match to part of a sequence from another gene, you should either mask out that region (with Ns) in your sequence for submission or submit only a partial sequence, that only includes unique regions of that gene.

```
> ref|NM_001715.2| UEG Homo sapiens B lymphoid tyrosine kinase (BLK), mRNA
Length=2642

 Score = 1088 bits (589),  Expect = 0.0
 Identities = 591/592 (99%), Gaps = 0/592 (0%)
 Strand=Plus/Plus

Query  1     CACCTCTGTCTGCTGCCGGCAGAAAGCCACAAGCCATGAAAACTGATTGAGATGAGAAGA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  392   CACCTCTGTCTGCTGCCGGCAGAAAGCCACAAGCCATGAAAACTGATTGAGATGAGAAGA  451

Query  61    ATTCATCTGGGACTGGCTTTTGCTTTAGGATGGTGTTGGAAGTTGCTCGTTGTCGCTAGG  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  452   ATTCATCTGGGACTGGCTTTTGCTTTAGGATGGTGTTGGAAGTTGCTCGTTGTCGCTAGG  511

                ⋮ *

Query  541   GAGCCTGGAAATGGAAAGGTGGTTCTTTAGATCACAGGGTCGGAAGGAGGCT  592
             ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  932   GAGCCTGGAAATGGAAAGGTGGTTCTTTAGATCACAGGGTCGGAAGGAGGCT  983
```

```
 Score = 2272 bits (1230),  Expect = 0.0
 Identities = 1232/1233 (99%), Gaps = 0/1233 (0%)
 Strand=Plus/Plus

Query  649   TGAAACCAACAAAGGTGCCTTCTCCCTGTCTGTGAAGGATGTCACCACCCAGGGGGAGCT  708
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1040  TGAAACCAACAAAGGTGCCTTCTCCCTGTCTGTGAAGGATGTCACCACCCAGGGGGAGCT  1099
                      ⋮*

Query  1009  GGGAACCATGTCTCCAGAAGCCTTCCTGGGTGAGGCCAACGTGATGAAGGCTCTGCAGCA  1068
             ||||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  1400  GGGAACCATGTCTCCAGAAGCCTTTCTGGGTGAGGCCAACGTGATGAAGGCTCTGCAGCA  1459
                      ⋮*
```

*Alignments shown have been shortened for display purposes

```
 Score =  412 bits (223),  Expect = 1e-113
 Identities = 225/226 (99%), Gaps = 0/226 (0%)
 Strand=Plus/Plus

Query  1973  GCCCCAGTAAGGTGTTCAGGACTGGTAAGCGACTGTCATCAAGTAAGGCCCCCGTGCTGG  2032
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2364  GCCCCAGTAAGGTGTTCAGGACTGGTAAGCGACTGTCATCAAGTAAGGCCCCCGTGCTGG  2423

Query  2033  GCACCCCCCGTGCTGGCCGCGTCCCCGCCTCTGCGCCCTGCGTGGACCCCGCCCTGCCCC  2092
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2424  GCACCCCCCGTGCTGGCCGCGTCCCCGCCTCTGCGCCCTGCGTGGACCCCGCCCTGCCCC  2483

Query  2093  GCTACAGAAGCCAGACTGGGTCCCGCGGACGCCAGCAGGGGCAACCCCAGCCTAGGCTGC  2152
             ||||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  2484  GCTACAGAAGCCAGACTGGGTCCCGCGGACGCCAGCAGGGGCAGCCCCAGCCTAGGCTGC  2543

Query  2153  GCTCCAGCACTGCGGGGCTTTTCTGCAATAAAGTCACGAGCGTTCG  2198
             ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2544  GCTCCAGCACTGCGGGGCTTTTCTGCAATAAAGTCACGAGCGTTCG  2589
```

B. How to use BLAST dbSNP to search for Sequence Polymorphisms

This section describes the use of BLAST to search the NCBI SNP database, dbSNP, to identify any polymorphisms in the sequence you will submit for assay design. dbSNP is database of known single nucleotide polymorphisms, small-scale insertions/deletions, polymorphic repetitive elements, and microsatellite variation. Here you will use your sequence of interest complete with any bases that you have masked from searches thus far.

1. **Submitting your sequence / Starting your query**
   - Go to the NCBI BLAST SNP site. The default Program is blastn. This is the program you should use.
   - Choose the SNP blast database that you would like to query based on the species of your sequence.
   - Enter your masked sequence, or Accession number, into the box provided. The sequence format should be FASTA. You may either search with your masked sequence (output from RepeatMasker) or have the sequence filtered for you by the program. To have the sequence filtered for you, simply check the appropriate boxes next to the word FILTER.
   - Click on 'Submit Query' to submit your search.

2. dbSNP BLAST Results

The output is typical of BLAST results, a list of sequences producing significant alignments to your query and the sequence alignments. Notice the Scores and Expect values, as well as the public identifiers. These are all discussed in the section entitled "List of Sequences producing significant alignments to your query".

Sequences producing significant alignments:		Score (Bits)	E Value
gnl|dbSNP|rs14053	rs=14053|pos=256|len=511|taxid=9606|mol="cD...	935	0.0
gnl|dbSNP|rs1042689	rs=1042689|pos=203|len=491|taxid=9606|mol...	900	0.0
gnl|dbSNP|rs1042701	rs=1042701|pos=301|len=601|taxid=9606|mol...	671	0.0
gnl|dbSNP|rs7843987	rs=7843987|pos=301|len=601|taxid=9606|mol...	516	9e-143
gnl|dbSNP|rs10097015	rs=10097015|pos=301|len=601|taxid=9606|m...	510	4e-141
gnl|dbSNP|rs10097005	rs=10097005|pos=301|len=601|taxid=9606|m...	462	1e-126
gnl|dbSNP|rs7840433	rs=7840433|pos=301|len=601|taxid=9606|mol...	442	2e-120
gnl|dbSNP|rs13248757	rs=13248757|pos=301|len=601|taxid=9606|m...	394	5e-106
gnl|dbSNP|rs34744472	rs=34744472|pos=301|len=601|taxid=9606|m...	337	8e-89

*List shortened for display purposes

Sequence Alignments

You will be able to readily identify any documented SNPs or sequence mismatches in the alignment as they are represented by red bases. Sequence identity is represented as a dot. Any SNPs identified in your sequence should also be masked out (changed to N) in your submission sequence so that no primer or probe is designed over that particular base.

```
>gnl|dbSNP|rs2306234 rs=2306234|pos=301|len=601|taxid=9606|mol="genomic"|class=1|alleles="C/T"|build=126
Length=601

 Score =  335 bits (181),  Expect = 3e-88
 Identities = 181/182 (99%), Gaps = 0/182 (0%)
 Strand=Plus/Plus
```

Your sequence of interest

```
Query  682  GGTTACTACAAAAACAACATGAAGGTGGCCATTAAGACGCTGAAGGAGGGAACCATGTCT  741
Sbjct  230  ............................................................  289

Query  742  CCAGAAGCCTTCCTGGGTGAGGCCAACGTGATGAAGGCTCTGCAGCACGAGCGGCTGGTC  801
Sbjct  290  ..........Y.................................................  349

Query  802  CGACTCTACGCAGTGGTCACCAAGGAGCCCATCTACATTGTCACCGAGTACATGGCCAGA  861
Sbjct  350  ..................                                            409

Query  862  GG  863
Sbjct  410  ..  411
```

Documented SNP in dbSNP.
It is important to mask this base before submission.

III. Identifying Exon Junctions

If you are going to order a gene expression assay, it is important to know where the exon junctions are in the cDNA sequence you are submitting for a Custom TaqMan® Assay. The TaqMan® MGB probe, when possible, should be designed across an exon-exon boundary in order to exclude the detection of genomic DNA. The exon boundaries should serve as your target(s) in your submission file. The more targets you provide, the better your chances of having an assay designed. While you may provide as few as one target, or as many as you would like, only one assay per sequence will be designed. If you are working with a gene sequence that is in a public database there are many places you may go to find exon information on the web. A few are listed and described below.

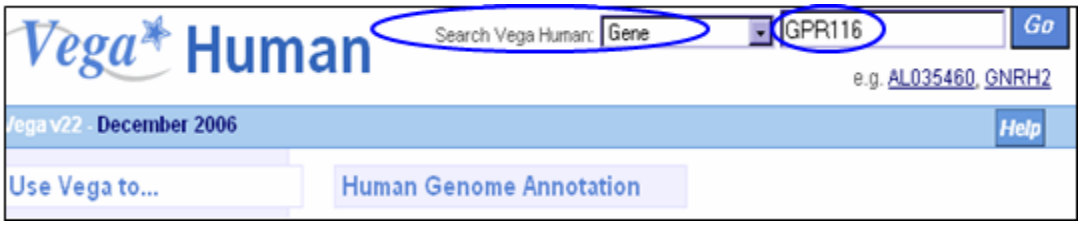
A. Ensembl / Vega

The Ensembl project developed a software system which produces and maintains automatic annotation on selected eukaryotic genomes. The Vertebrate Genome Annotation (VEGA) database is a collection of high quality, frequently updated, manually curated vertebrate finished genome sequences. The VEGA website is built upon code from the Ensembl project. Searching either site will give similar results.

If you are working with human, mouse, zebrafish, pig or dog use the VEGA website, as these are finished genome sequences. For other species, go to Ensembl. You may also access the VEGA genomes for the above species via the Ensembl site.

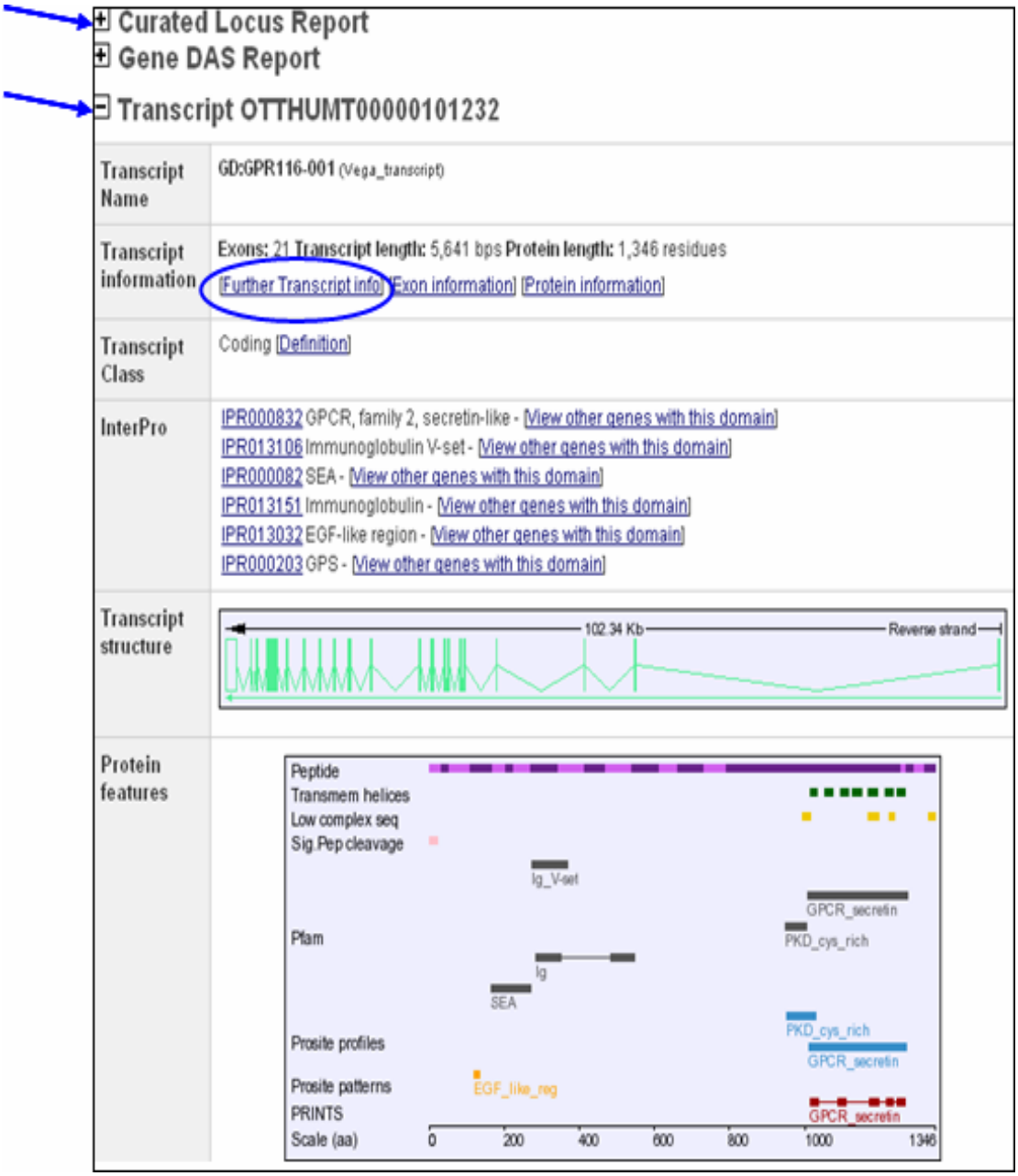
To search for exon information:

- Choose the species of interest.
- Enter in a gene identifier, such as gene name, gene symbol or RefSeq ID and click “Go”



- Once you get the results, click on the link for the transcript of interest. Make sure to read the Description of each result so that you are choosing the transcript in which you are interested and not a psuedogene or a transcript from a different gene. In the example shown below, please note that the first hit shown is to a psuedogene. The second hit is the desired transcript.

- Click on “Further Transcript info” to view the cDNA sequence. Notice that there are several tiers of information in this record, such as “Curated Locus Report”, “Gene DAS Report” and “Transcript”. If there is a plus sign in front of the name of the tier, the information is compressed. If you would like to see the information in a compressed tier, simply click on the “+” to reveal the information.



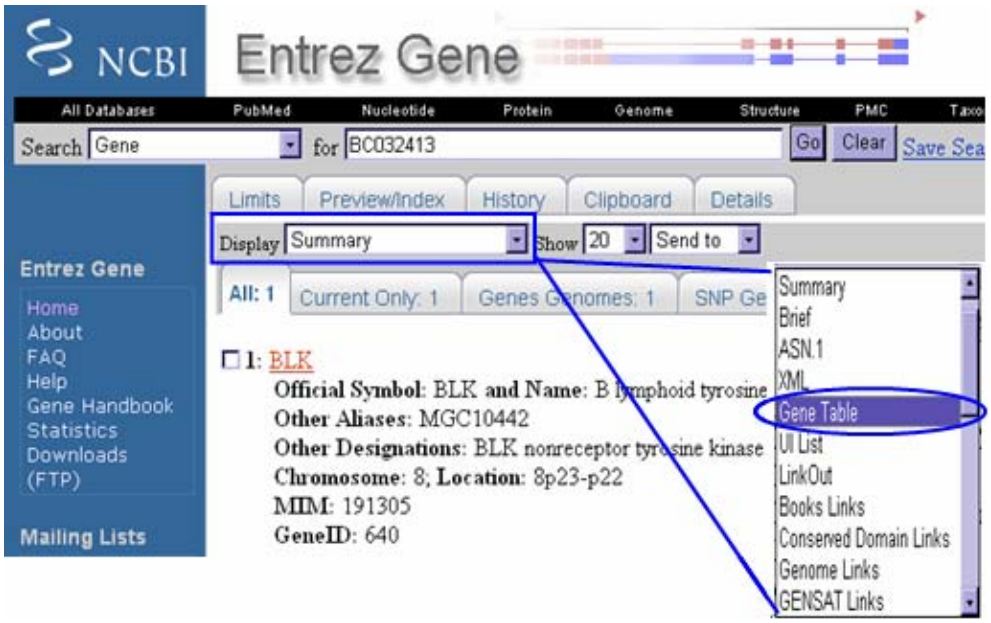
- The transcript sequence will be shown with the exons represented contiguously in alternating blue then black text.

B. <u>Entrez Gene at NCBI (National Center for Biotechnology Information)</u>

Entrez is a tool used to query different databases at NCBI. GenBank is a public database of nucleotide sequences (as well as other sequences), that is updated daily. A good number of sequences are annotated with mRNA sequences, so you may be able to find some exon information on your sequence of interest here.

To do this:

- Search the nucleotide database using Entrez Gene at NCBI. There are several options for search terms. You can search with a particular Accession number (BC032413), the gene name (lymphoid tyrosine kinase) or the gene symbol (BLK).
- If you get more than one hit, select the gene and species of interest and then click on the gene symbol.
- Go to the Display drop-down menu and choose 'Gene Table'.



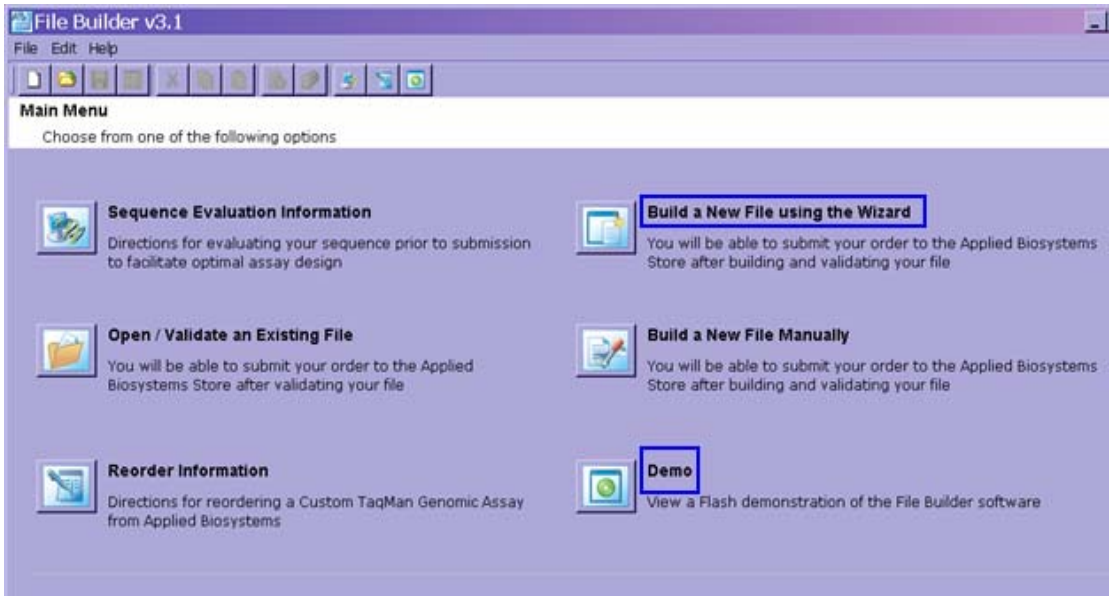
- This will bring up a table that includes the length and location of each exon, coding exon and intron, as shown below.

Exon information:
NM_001715.2 length: 2590 bp, number of exons: 13
NP_001706.2 length: 506 aa, number of exons: 12

EXON		Coding EXON		INTRON	
coords	length	coords	length	coords	length
1 - 580	580 bp			581 - 49212	48632 bp
49213 - 49336	124 bp	49214 - 49336	123 bp	49337 - 52040	2704 bp
52041 - 52092	52 bp	52041 - 52092	52 bp	52093 - 54020	1928 bp
54021 - 54114	94 bp	54021 - 54114	94 bp	54115 - 55012	898 bp
55013 - 55111	99 bp	55013 - 55111	99 bp	55112 - 56147	1036 bp
56148 - 56251	104 bp	56148 - 56251	104 bp	56252 - 60731	4480 bp
60732 - 60878	147 bp	60732 - 60878	147 bp	60879 - 61320	442 bp
61321 - 61473	153 bp	61321 - 61473	153 bp	61474 - 62646	1173 bp
62647 - 62826	180 bp	62647 - 62826	180 bp	62827 - 63950	1124 bp
63951 - 64027	77 bp	63951 - 64027	77 bp	64028 - 67290	3263 bp
67291 - 67441	151 bp	67291 - 67441	151 bp	67442 - 68967	1526 bp
68968 - 69099	132 bp	68968 - 69099	132 bp	69100 - 69891	792 bp
69892 - 70588	697 bp	69892 - 70097	206 bp		

Note: If there is no exon information available for your sequence of interest, you may still submit that sequence for assay design. For your targets, select multiple sites across the sequence to ensure optimal design.

Having evaluated the quality of your sequence information, you are now ready to move on to preparing your submission file using the File Builder software. Start the process by using the New File Wizard. A demo is also available within the software for further assistance, and you may wish to consult Ordering Custom TaqMan® Genomic Assays: Online Ordering Procedures Using the File Builder Software: Quick Reference Card.



For Research Use Only. Not for use in diagnostic procedures.

Notice to Purchaser for Custom TaqMan Gene Expression Assays and Custom TaqMan Probes:

Practice of the patented 5' Nuclease Process requires a license from Applied Biosystems. The purchase of Custom TaqMan Gene Expression Assays and Custom TaqMan Probes includes an immunity from suit under patents specified in the product insert to use only the amount purchased for the purchaser's own internal research when used with the separate purchase of an Authorized 5' Nuclease Core Kit. No other patent rights are conveyed expressly, by implication, or by estoppel. For further information contact the Director of Licensing, Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94404, USA.

127GU07-03

Part Number 4371002 Rev C