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

Overview

The Custom TaqMan® Gene Expression Assays (formerly named TaqMan® Assays-by-Design® Gene Expression Service) 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 (formerly named TaqMan® Assays-on-Demand™ Gene Expression Assays) offerings. Studies that involve viral detection, species other than *H. sapiens*, *M. musculus*, *R. norvegicus*, *A. thaliana*, *D. melanogaster*, or *C. elegans*, or detection of specific pathogens are some examples of applications that would benefit from this custom design line of products. For human, mouse, rat, *Drosophila*, *C. elegans*, and *Arabidopsis* gene expression assays, the TaqMan® 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 Applied Biosystems web site for regular updates.*

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 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: Protocol: Submission Guidelines
- Online Ordering Procedures Using the File Builder Software: Quick Reference Card
- 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 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

When choosing sequences to submit, first consider the biological significance of the desired assay. The quality assurance on assays carried out during manufacture of the primer 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 one or 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 coordinates (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 to have the TaqMan® MGB probe designed across an exon-exon boundary. The exon boundary location is used as a coordinate 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. Short (fewer than 300 bases) sequences limit the potential number of assays that can be designed.*

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 binding, or no binding, 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 no Ns are 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 coordinate(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 the nucleotide database using Entrez at NCBI or use Vertebrate Genome Annotation (VEGA), which is part of the Ensembl project.

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 DNA. 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.

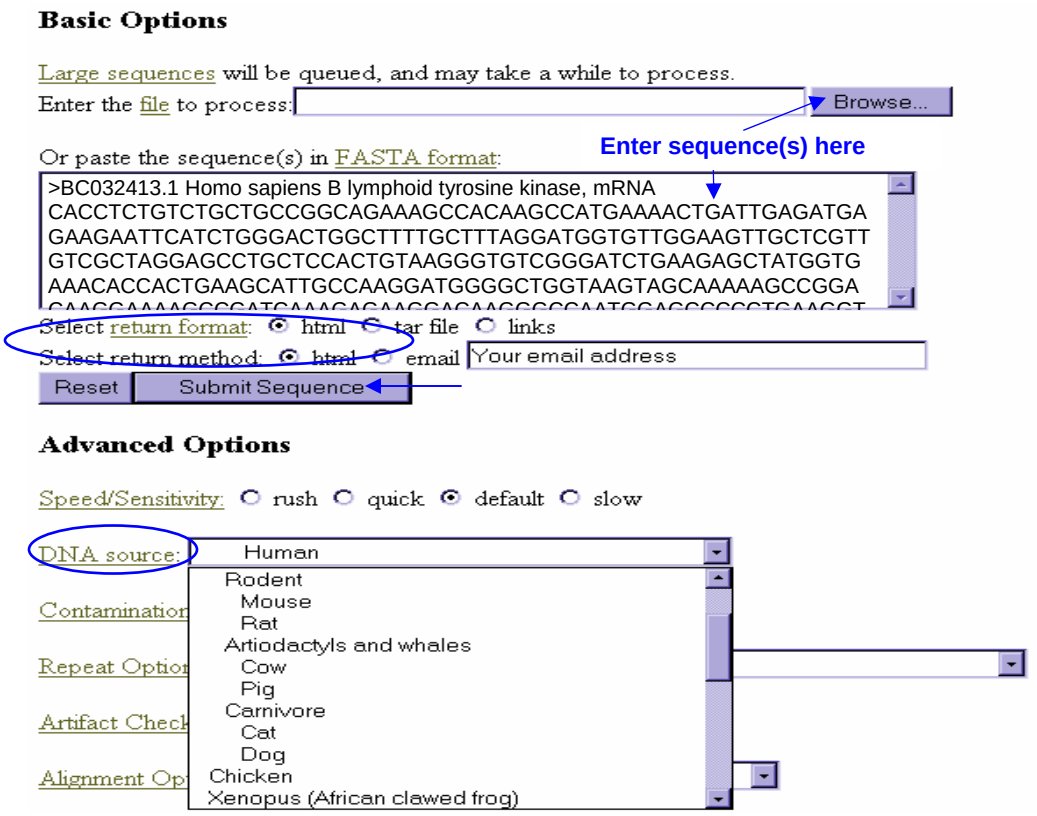
How to use RepeatMasker

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).
- When selecting ‘return format’ and ‘return method’, if you choose “html” for both, your results will 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

Repeat Annotations:

SW score	perc div.	perc del.	perc ins.	query sequence	position in query begin	end	(left)		matching repeat	class/family	position in repeat begin	end	(left)	ID
216	30.8	14.3	0.0	BC032413.1	1882	1972	(279)	+	MIR	SINE/MIR	84	187	(75)	1
477	0.0	0.0	0.0	BC032413.1	2199	2251	(0)	+	(A)n	Simple_repeat	1	53	(0)	2

*Masked Sequence:

```
>BC032413.1 Homo sapiens B lymphoid tyrosine kinase, mRNA
CACCTCTGTCTGCTGCCGGCAGAAAGCCACAAGCCATGAAAACTGATTGA
GATGAGAAGAATTCATCTGGGACTGGCTTTTGCTTTAGGATGGTGTTGGA
AGTTGCTCGTTGTCGCTAGGAGCCTGCTCCACTGTAAGGGTGTCGGGATC
GTGCTGGCGCAGCCGGCCCGAGGAGCGGCCCACCTTCGAGTTCCTGCAGT
CGGTGCTGGAGGACTTCTACACGGCCACCGAGCGGCAGTACGAGCTGCAG
CCCTAGCCGGCCGCGCCCGCCTGCGCCCCGTGCCCACCTCTGCGCGGACG
ACCCCGACTTCCGTGCCATCCCAGACGGGCCGCGAAGGCGGGGTGTCGCC
TGTGCCCTTTTCTCAGACCCGGAATCCAGTGGGCAGAGGCAGCTTCGCAG
GGGGTCCCCGGACGGACTCCTTCACCGACNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNGCCCCAGTAAGGTGTTCAGGACTGGTAAGC
GACTGTCATCAAGTAAGGCCCCCGTGCTGGGCACCCCCCGTGCTGGCCGC
GTCCCCGCCTCTGCGCCCTGCGTGGACCCCGCCCTGCCCCGCTACAGAAG
CCAGACTGGGTCCCGCGGACGCCAGCAGGGGCAACCCCAGCCTAGGCTGC
GCTCCAGCACTGCGGGGCTTTTCTGCAATAAAGTCACGAGCGTTCGNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

*Sequence was shortened for display purposes.

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

Summary:

```
==================================================
file name: RM2sequpload_2043
sequences:             1
total length:       2249 bp  (2249 bp excl N-runs)
GC level:          0.00 %
bases masked:        144 bp (  6.40 %)
==================================================
```

Number & Percentage of bases masked

Repeat Elements

	number of elements*	length occupied	percentage of sequence
SINEs:	1	91 bp	4.05 %
ALUs	0	0 bp	0.00 %
MIRs	1	91 bp	4.05 %
LINEs:	0	0 bp	0.00 %
LINE1	0	0 bp	0.00 %
LINE2	0	0 bp	0.00 %
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:		91 bp	4.05 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	1	53 bp	2.36 %
Low complexity:	0	0 bp	0.00 %

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

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.

II. BLAST® (Basic Local Alignment Search Tool)

After you have selected a target, there are several things that must be considered before submitting a sequence for a Custom TaqMan® Assay. 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 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 and protein sequences. BLAST®, a sequence comparison algorithm, is available to facilitate nucleotide and protein searching of the NCBI public databases.

A. How to use BLAST® to search for Sequence Similarity

1. **Submitting your sequence / Starting your query**
    - Go to the NCBI BLAST® site
    - Choose **"**Nucleotide-nucleotide BLAST (blastn)" under **Nucleotide**.
    - 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 to find the best exon boundaries with which to work.
    - Enter your sequence into the box provided. You may want to search with your masked sequence generated from RepeatMasker. There are three sequence formats that may be entered into this box. (See pg. 8) For more information on this, click on the word Search next to the box.
    - Choose the appropriate database to search. When searching with a cDNA sequence for a gene expression assay, you would probably want to search at least the 'est' (expressed sequence tags) and the 'nr' databases for the species you are working with.
    - Under 'Options for advanced blasting' you can, among other things, limit your search to a specific organism using the drop down menu, and opt to filter your query for low complexity sequences (not necessary if searching with output form RepeatMasker).
    - Click on 'BLAST!' 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

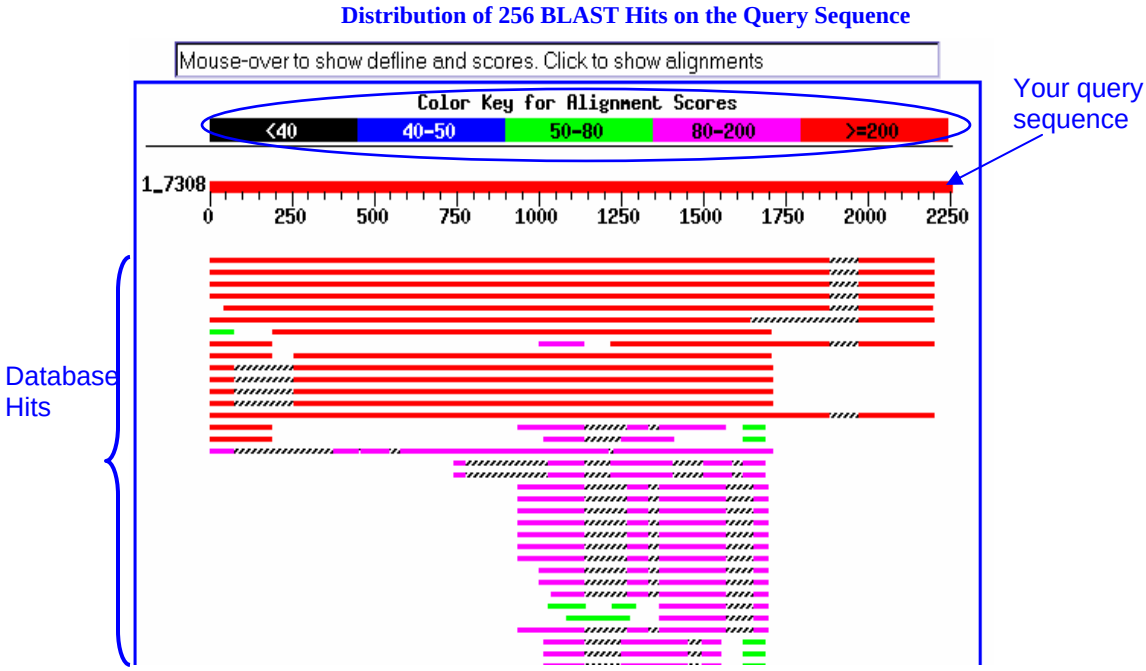
There are three general parts to BLAST® results:

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

These sections are described below (p9–12) to give you a better understanding of what information can be obtained from a BLAST search of the NCBI public nucleotide database.

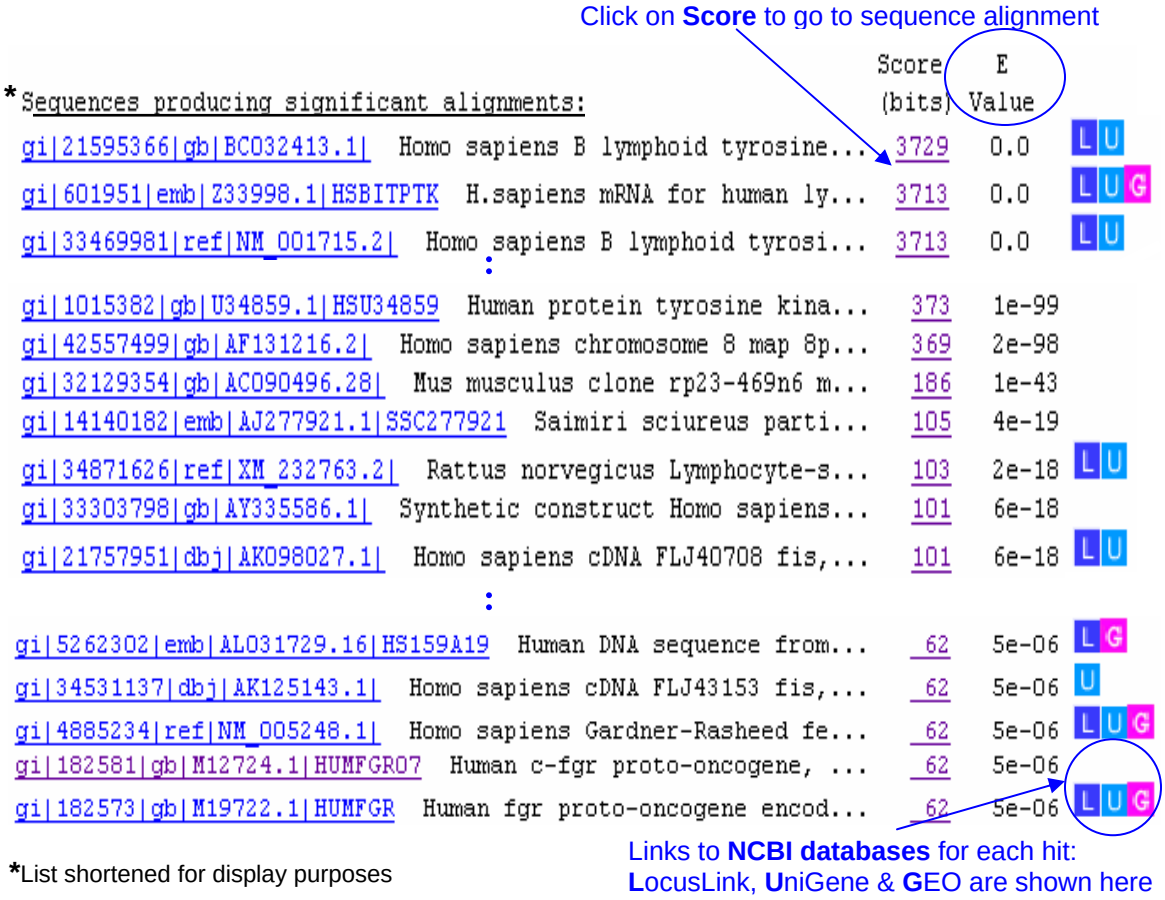
a. Graphical Overview

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 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.



*List shortened for display purposes

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 one that was used for the search in this example, BC032413.1. The score is very high (3729), and the Expect value is 0. Remember that the closer an E-value is to "0" the more "significant" the match. For this particular query, most of the hits are to human tyrosine kinases, which is the same molecular function as the query. 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, so that they will not be considered for assay design.

c. Sequence Alignments

This section is 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, shown here boxed in blue, is the query sequences aligned to itself. This will be the first alignment shown, and will be a 100% match to itself.

```
Score = 4357 bits (2198), Expect = 0.0
Identities = 2198/2198 (100%)
Strand = Plus / Plus
```

The alignment below is an example of another hit from the BLAST search.

```
>gi|6984208|gb|AF228313.1|AF228313 Homo sapiens tyrosine kinase LCK mRNA, partial cds
          Length = 1491

 Score =  101 bits (51), Expect = 6e-18
 Identities = 162/199 (81%)
 Strand = Plus / Plus

Query: 1367 gaggggccaagttccccatcaagtggacagccccggaagccatccacttcggggtcttc 1426
            ||||||||||||||| ||||| |||||||||| || ||||||||  ||| ||||   |||
Sbjct: 1153 gaggggccaagtttcccattaagtggacagcgccagaagccattaactacgggacattc 1212


Query: 1427 accatcaaagcagacgtgtggtcgtttggagtcctcctgatggaagttgtcacttatggg 1486
            ||||||||  |||| |||||||| |||||  |||| |||| |||| |||||||  | ||
Sbjct: 1213 accatcaagtcagatgtgtggtcttttgggatcctgctgacggaaattgtcacccacggc 1272


Query: 1487 cgggtgccatacccagggatgagcaaccccgaggtcatccgcaacctggagcgcggctac 1546
            ||  | || |||||||||||||| |||||| ||||| || |  |||||||||| ||||||
Sbjct: 1273 cgcatcccttacccagggatgaccaacccggaggtgattcagaacctggagcgaggctac 1332


Query: 1547 cgcatgccgcgccccgaca 1565
            ||||||  |||||| ||||
Sbjct: 1333 cgcatggtgcgccctgaca 1351
```

```
 Score = 89.7 bits (45), Expect = 2e-14
 Identities = 162/201 (80%)
 Strand = Plus / Plus

Query: 936  ctggacaattcggcgaagtctggatgggttactacaaaaacaacatgaaggtggccatta 995
            ||||||| ||||| || || |||||||| ||||||||     ||| |||||||||  | |
Sbjct: 719  ctggacagttcggggaggtgtggatggggtactacaacgggcacacgaaggtggcggtga 778


Query: 996  agacgctgaaggagggaaccatgtctccagaagccttcctgggtgaggccaacgtgatga 1055
            |||  |||||| |||| | |||||| || || ||||||||||  ||||||||| | ||||
Sbjct: 779  agagcctgaagcagggcagcatgtccccggacgccttcctggccgaggccaacctcatga 838


Query: 1056 aggctctgcagcacgagcggctggtccgactctacgcagtggtcaccaaggagcccatct 1115
            ||   ||||| ||| |||||||||| || |||||||| ||||||||| ||||||||||||
Sbjct: 839  agcagctgcaacaccagcggctggttcggctctacgctgtggtcacccaggagcccatct 898


Query: 1116 acattgtcaccgagtacatgg 1136
            ||||  |||| || ||||||||
Sbjct: 899  acatcatcactgaatacatgg 919
```

Database hit

Hash marks represent matched nucleotides

This is a second alignment to **same** database hit, Homo sapiens tyrosine kinase LCK mRNA, in another region of the sequence.

***Query** = your sequence of interest

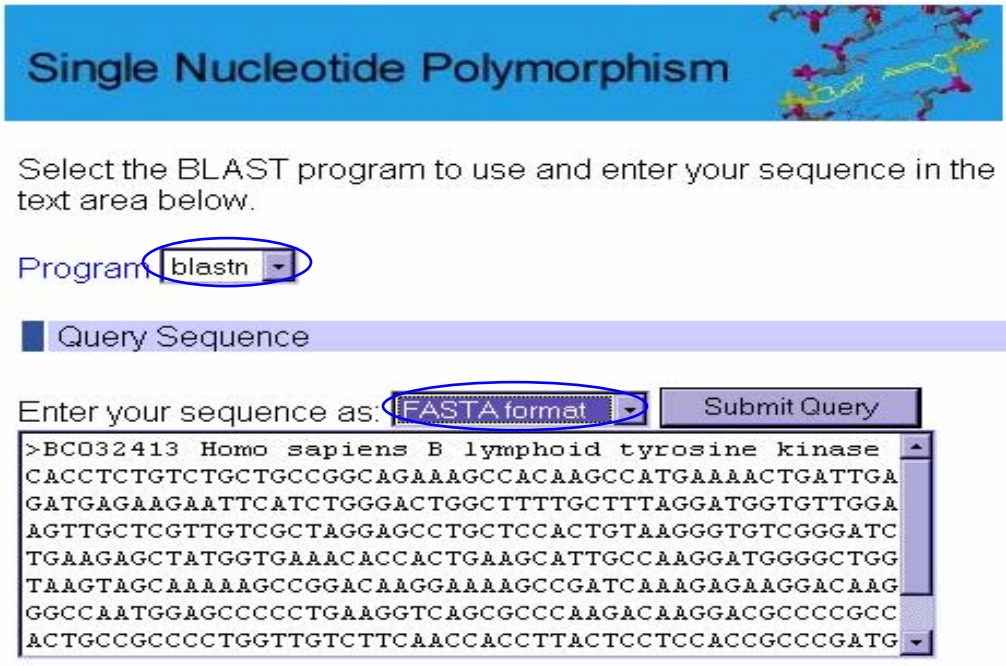
***Sbjct** = database hit = Homo sapiens tyrosine kinase LCK mRNA

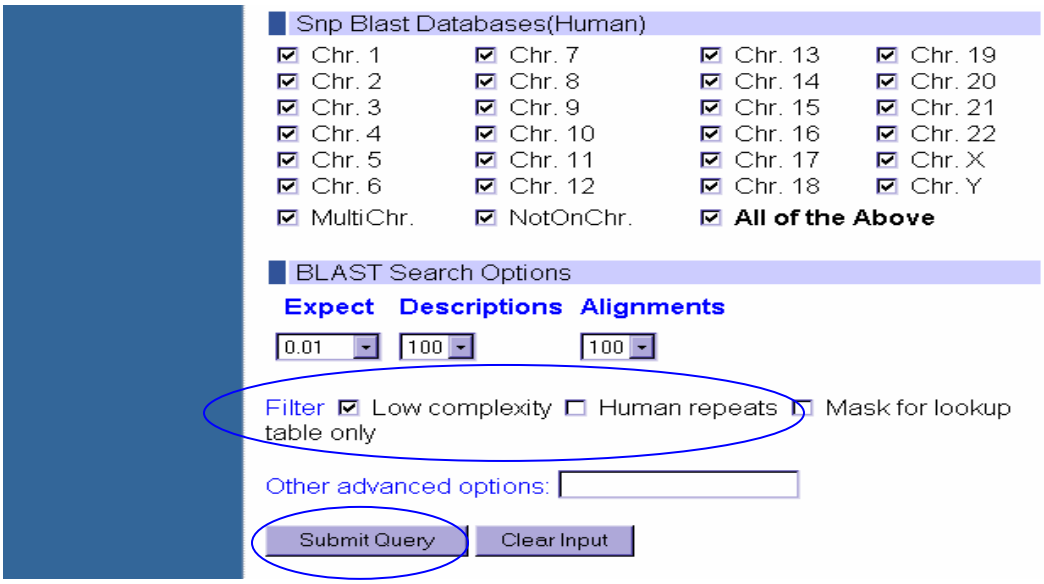
In the example on page 11, the query sequence did not align to this database hit contiguously. Only a part of the query sequence aligned to this sequence from the database: starting at base 1367 of the query sequence, and ending at base 1565, and from 936 -1136 of the query sequence. Also, there is more than one alignment associated with this hit. These alignments are shown in order of significance, and have different E-values and scores. Sometimes BLAST alignments of an mRNA query can bring up hits to genomic DNA. These alignments can be broken into multiple HSP segments, indicating the presence of introns in the gDNA. If a segment of your query sequence came up with a significant match to part of a sequence from another gene, you could either mask out that region of the sequence in your sequence for submission or simply not include that region in your submission and find another region of interest in your gene to submit.

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

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 sequence that you would like to submit based on your BLAST for sequence similarity.
- Enter your sequence 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, as shown on page 13.
- Click on 'Submit Query' to submit your search.





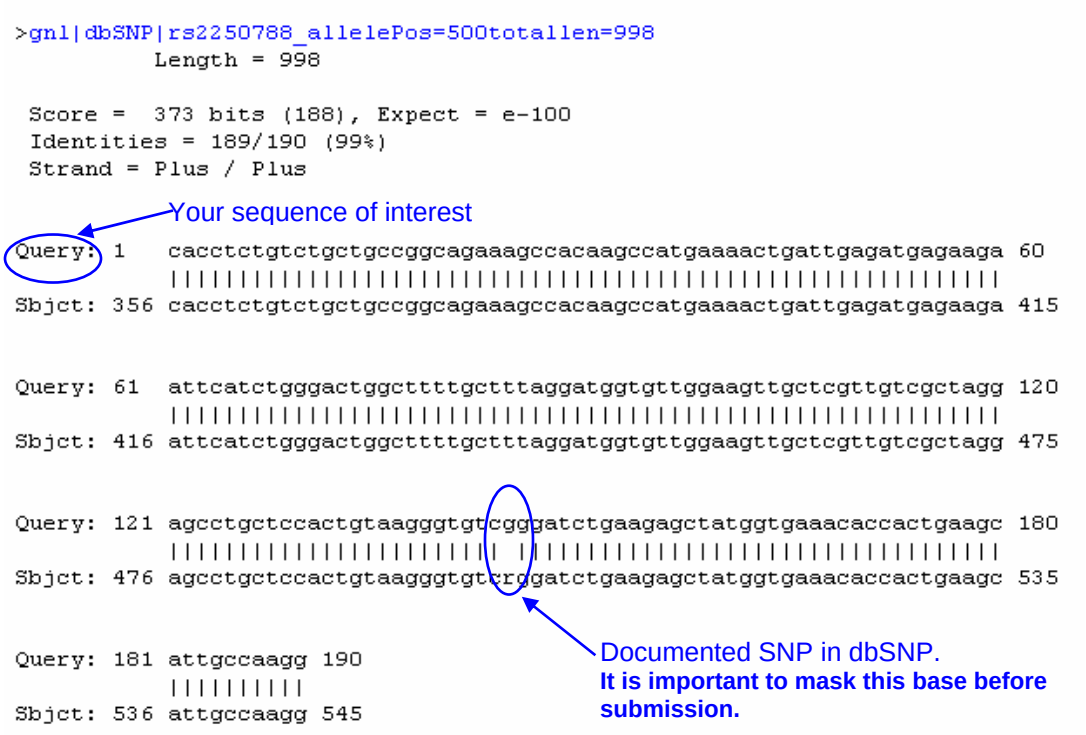
2. dbSNP BLAST® Results

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|rs922483_allelePos=161totallen=577	375	e-101
gnl|dbSNP|rs2250788_allelePos=500totallen=998	373	e-100
gnl|dbSNP|rs2245250_allelePos=201totallen=401	254	1e-64
gnl|dbSNP|rs2245232_allelePos=201totallen=401	228	7e-57
gnl|dbSNP|rs12386974_allelePos=201totallen=401	194	1e-46
gnl|dbSNP|rs6994605_allelePos=201totallen=401	172	4e-40
gnl|dbSNP|rs2244938_allelePos=201totallen=401	105	7e-20
gnl|dbSNP|rs2244931_allelePos=201totallen=401	88	2e-14
gnl|dbSNP|rs13272061_allelePos=201totallen=401	82	1e-12
gnl|dbSNP|rs11780851_allelePos=284totallen=484	52	0.001

Sequence Alignments

By looking for mismatches in the alignment (no hash marks) you will be able to identify documented SNPs. These SNPs should also be masked out (changed to N) in your submission sequence so that no primer or probe is designed over this base.



III. Identifying Exon Junctions

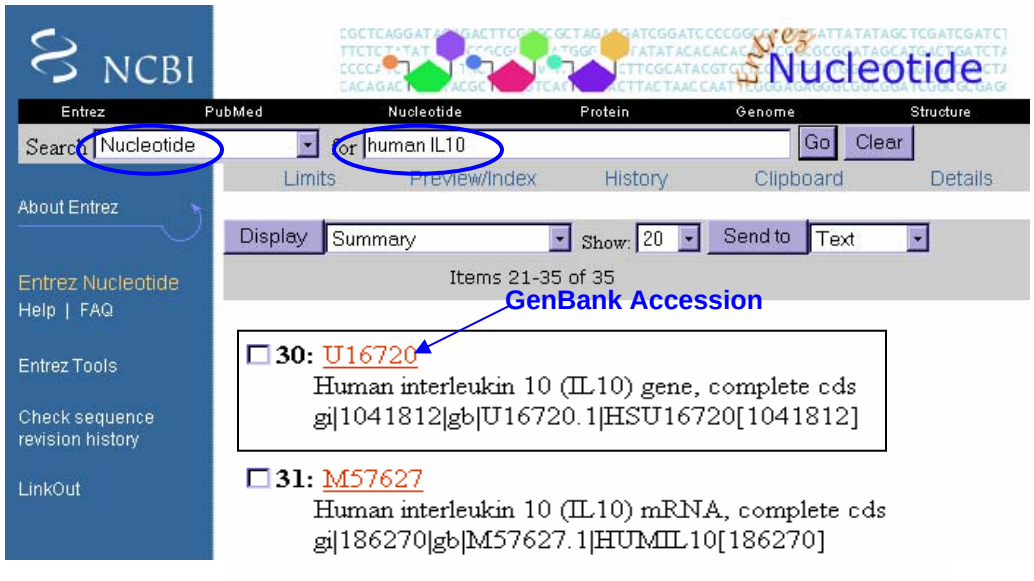
If you are going to order a gene expression assay, it is important to know where the exon junctions are in the cDNA sequence you are submitting for a Custom TaqMan® Assay. The TaqMan® MGB probe, when possible, should be designed across an exon-exon boundary in order to exclude the detection of genomic DNA. The exon boundaries should serve as your coordinate(s) in your submission file. The more coordinates you provide, the better your chances of having an assay designed. While you may provide as few as one coordinate, or as many as you'd 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. Entrez at NCBI (National Center for Biotechnology Information)

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 at NCBI.
- Find your gene of interest with a complete coding sequence.
- Click on the GenBank Accession number.



- Scroll down to find the **mRNA** annotated under the **Features** section. This will list the exon start and stop bases for this gene sequence. For example, in this record the first exon starts at base 4057 and ends at base 4221, the second exon is from bases 5088 – 5147, and so on. If you join these exons together, they make up the mRNA sequence. The bases listed correspond to the bases from the DNA sequence shown in that particular GenBank record.

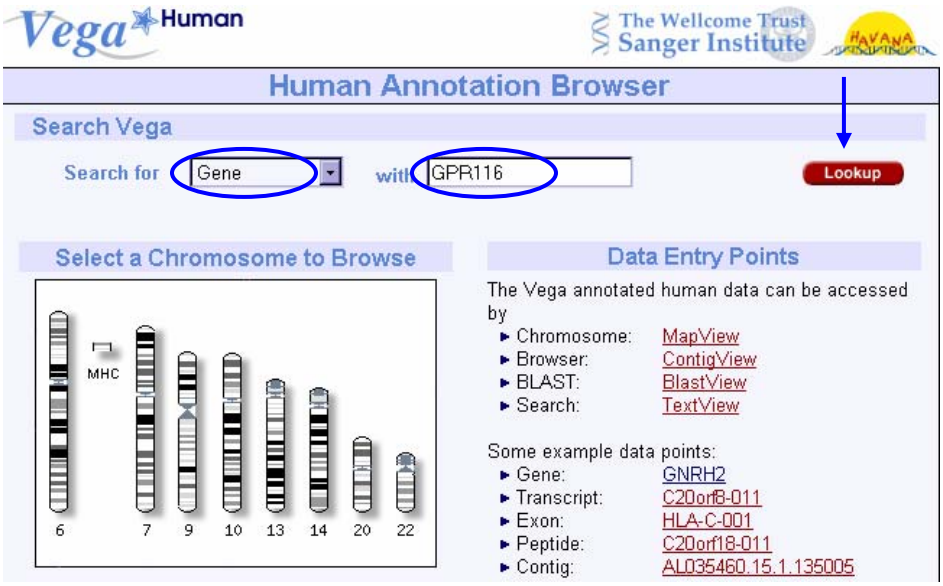
```
1: U16720. Human interleukin...[gi:1041812]

LOCUS       HSU16720                8868 bp    DNA     linear   PRI 28-OCT-1995
DEFINITION  Human interleukin 10 (IL10) gene, complete cds.
ACCESSION   U16720
                    :
FEATURES             Location/Qualifiers
     source          1..8868
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="1"
     repeat_region   1144..1447
                     /rpt_family="Alu"
                     /rpt_type=dispersed
     mRNA            join(<4057..4221,5088..5147,5438..5590,6601..6666,
                     7742..8868)
     gene            join(4057..4221,5088..5147,5438..5590,6601..6666,
                     7742..7834)
                     /gene="IL10"
     CDS             join(4057..4221,5088..5147,5438..5590,6601..6666,
                     7742..7834)
                     /gene="IL10"
                     /codon_start=1
                    .
                    :
```

B. Ensembl / Vega

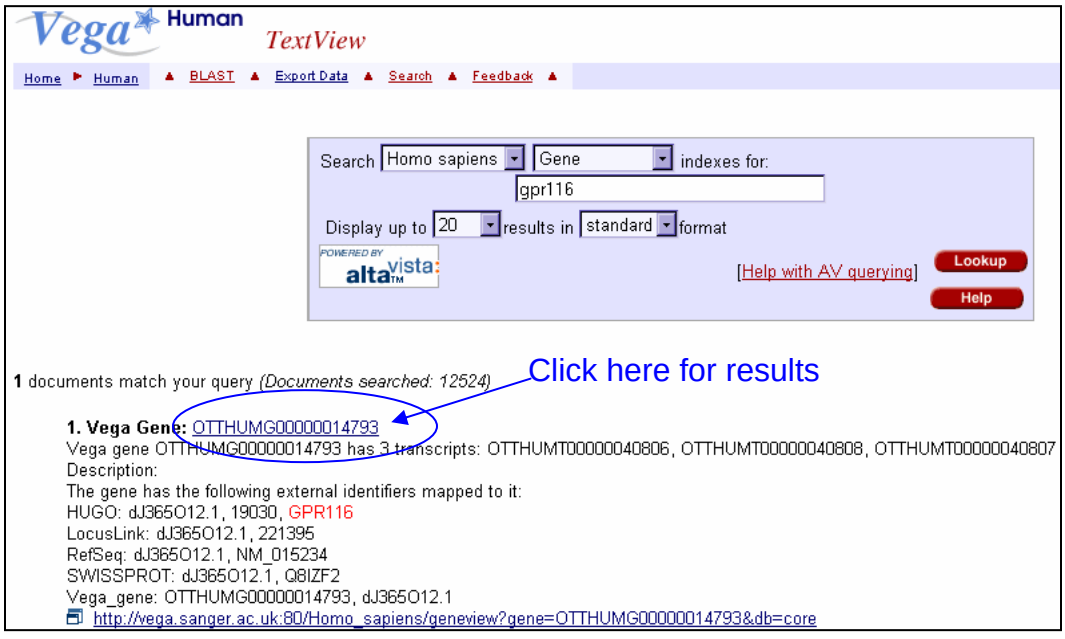
The Vertebrate Genome Annotation (VEGA) is part of the Ensembl project. The VEGA database is a collection of manually curated genome sequences. The current genomes available for searching are human, mouse and zebrafish.

To search this database for exon information:

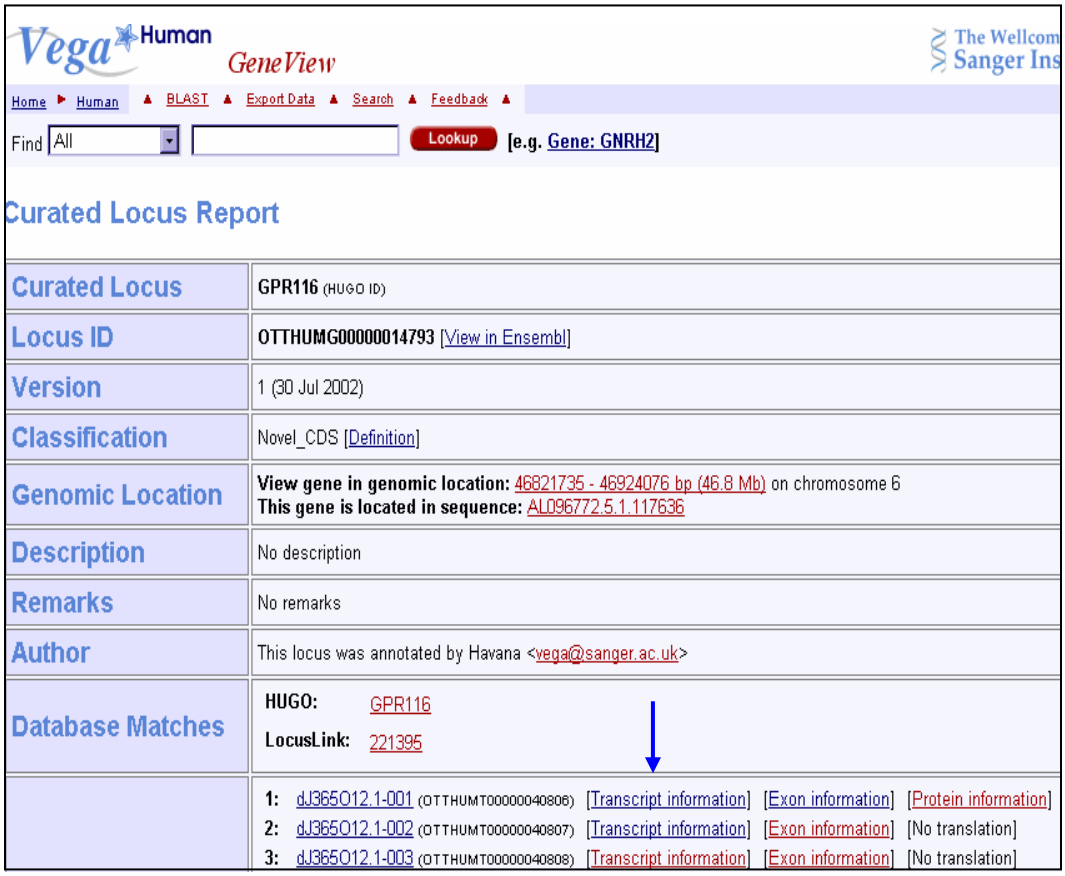
- Choose the species of interest.
- Enter in a gene identifier, such as gene name, gene ID, accession number, etc, and click on “Lookup”

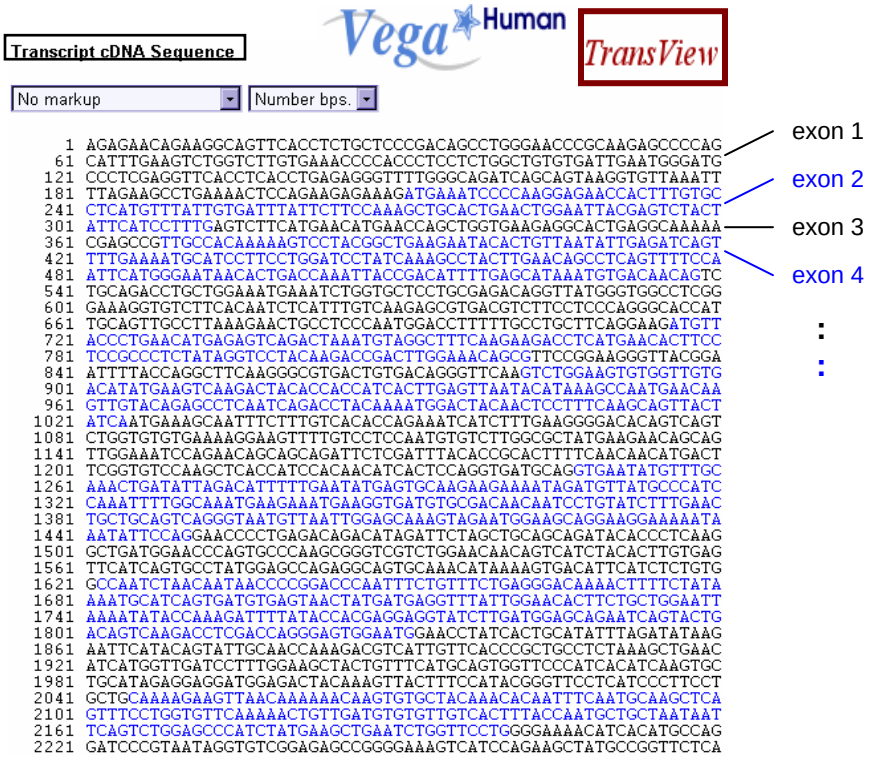


- Once you get the results, click on the link to the gene of interest.



- Click on “Transcript information” to view the cDNA sequence, with the exons shown contiguously in alternating blue then black text.





Note: If there is no exon information available for your sequence of interest, you may still submit that sequence for assay design. For your coordinates, 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.

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

Custom TaqMan Gene Expression products –

Notice to Purchaser: Disclaimer of License for Custom Sequence Detection Primers
This product is optimized for use in the Polymerase Chain Reaction (PCR) and 5' nuclease detection methods covered by patents owned by Roche Molecular Systems, Inc. and F. Hoffmann-La Roche Ltd. No license under these patents to use the PCR process or 5' nuclease detection methods is conveyed expressly or by implication to the purchaser by the purchase of this product. A license to use the PCR process for certain research and development activities accompanies the purchase of certain Applied Biosystems reagents when used in conjunction with an authorized thermal cycler, or is available from Applied Biosystems. Further information on purchasing licenses to practice the PCR process may be obtained by contacting the Director of Licensing, Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94404 or at Roche Molecular Systems, Inc., 1145 Atlantic Avenue, Alameda, California 94501, USA.

Notice to Purchaser: Disclaimer of License for Custom TaqMan Probes

This product is optimized for use in the Polymerase Chain Reaction (PCR) and 5' nuclease detection methods covered by patents owned by Roche Molecular Systems, Inc. and F. Hoffmann-La Roche Ltd. No license under these patents to use the PCR process is conveyed expressly or by implication to the purchaser by the purchase of this product. A license to use the PCR process for certain research and development activities accompanies the purchase of certain Applied Biosystems reagents when used in conjunction with an authorized thermal cycler, or is available from Applied Biosystems. Further information on purchasing licenses to practice the PCR process may be obtained by contacting the Director of Licensing, Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94404 or at Roche Molecular Systems, Inc., 1145 Atlantic Avenue, Alameda, California 94501, USA.

Notice to Purchaser

TaqMan® probes are covered by U.S. Patent 5,723,591 and foreign counterparts and patents pending owned by Applera Corporation, and may be covered by U.S. Patents 5,801,155 and 6,084,102 and foreign counterparts licensed to Applied Biosystems.

Publication 127GU07-02

Part Number 4371002 Rev B