Bioinformatic Evaluation of a Sequence for Custom TaqMan® SNP Genotyping Assays

Overview

The Custom TaqMan® SNP Genotyping Assays (formerly named TaqMan® Assays-by-Design® SNP Genotyping Service) are custom assays that are designed, synthesized, formulated, and delivered as analytically quality-controlled primer and probe sets for single nucleotide polymorphism (SNP) genotyping 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® SNP Genotyping Assay. Specific information is given on how to assess a sequence using a variety of on-line tools.

The Custom TaqMan® SNP Genotyping Assays provide the researcher the opportunity to design an assay that is not currently available through the TaqMan® SNP Genotyping Assays, TaqMan® Validated, Coding and Pre-Designed SNP Genotyping Assays offerings. Studies that involve new SNP discovery or SNPs in non-human species are examples of applications that would benefit from this custom design line of products. For human SNP assays, the TaqMan® SNP Genotyping Assays products should be used. If a particular SNP assay 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® SNP Genotyping 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 steps 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 Procdures Using the File Builder Software: Quick Reference Card
- Ordering TaqMan® SNP Genotyping Assays: Quick Reference Card.

Assessing the Quality of the Sequence

Overview

The most important factor in the success of the Custom TaqMan® 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. Consider the following when selecting your target sequence:

Biological Significance

When choosing sequences to submit, first consider the biological significance of the desired assay. Points to take into consideration include:

- Has the SNP been substantiated by more than one line of experimental evidence, i.e., is the SNP a "double hit" SNP?
- Is there Minor Allele Frequency (MAF) data available for a SNP?
- Has this SNP been identified in the population, e.g., ethnic group, that you are examining?

Such biological qualifiers give confidence that a given SNP is well studied and may be useful as a marker in your particular study.

The quality assurance test of assays carried out during manufacture of the primers and probes ensure that the yield and content of the primers and probes meet specifications. For human SNP assays, an additional functional test is done with 20 genomes representing 3 populations (African American, Caucasian, and Japanese). This test ensures that a passed assay yields an amplified product and that at least one genotypable cluster forms in an allelic discrimination plot. However, Applied Biosystems is unable to guarantee the biological performance of the assays.

Example:

If you are studying a known SNP, try to get information on allele frequencies (in NCBI dbSNP, HapMap, or other project databases). By knowing the minor allele frequency, you can estimate the population size required to detect a given minor allele and to provide statistically significant results.

The Hardy-Weinberg Equilibrium (HWE) equation may be used to assess the likelihood that a SNP with a known MAF in a given population is detectable in the same population of a particular size:

$q^2 + 2qp + p^2 = 1$, where q and p represent the allele frequencies.

The values obtained for q^2, 2qp, and p^2 correspond to the fraction of a given population that would be homozygous q:q, heterozygous q:p, and homozygous p:p. E.g, For a SNP with a MAF of 5% in a given population, the HWE predicted spread of genotypes is: 0.0025 q:q, 0.0095 q:p, and 0.9025 p:p. Thus, in a test of 20 gDNAs from this population, one might expect to see approximately 0 homozygotes for the minor allele, 2 heterozygotes, and 18 samples homozygous for the major allele. It would take a sample size of approximately 400 individuals to detect a homozygote for the minor allele.

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 possibilities, albeit, most SNP assays produce amplicons of <300 bps.
- 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.

Example: You have sequenced some clones, but there are some regions of the sequence that only have a single pass of sequence, or sequence from only one strand of the clone. You send the sequence in for a Custom TaqMan® Assay anyway. You may end up getting no amplification or nonspecific amplification with this assay because it may bind to other targets or to nothing at all, if your sequence contained incorrect sequences to which assay probes and primers were designed.

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 NCBI and dbSNP 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 other SNP sites), albeit the use of Ns may limit assay design choices.

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 SNP assays, make sure that no Ns are within two 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 genomic DNA sequence by verifying that the target sequence is unique within the organism you are studying.

To Ensure Unique Primer and Probe Sequences for gDNA:

1. Follow one of two strategies:
   - Analyze the entire target sequence.
   - Limit analysis to the region around the target site (for example, 100 bp on either side of the SNP).

2. Run the sequence through a program such as **Repeat Masker*** to detect common repetitive elements found in genomic DNA.

3. If many regions with similar sequences are returned, try using a filter.
For example, limit the search to Human genomic DNA for SNPs.
Note: *The BLAT server at the University of California Santa Cruz carries out searches using the assembled genome, so that when conducting a BLAST search only one species is being queried.*

4. Perform a **BLAST®** search against public databases to detect regions within your sequence that have similarity to other published sequences and repetitive elements.

5. Perform a BLAST® search against SNP databases, such as dbSNP and dbSTS, to determine if there are any other polymorphisms in your target sequence.

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 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

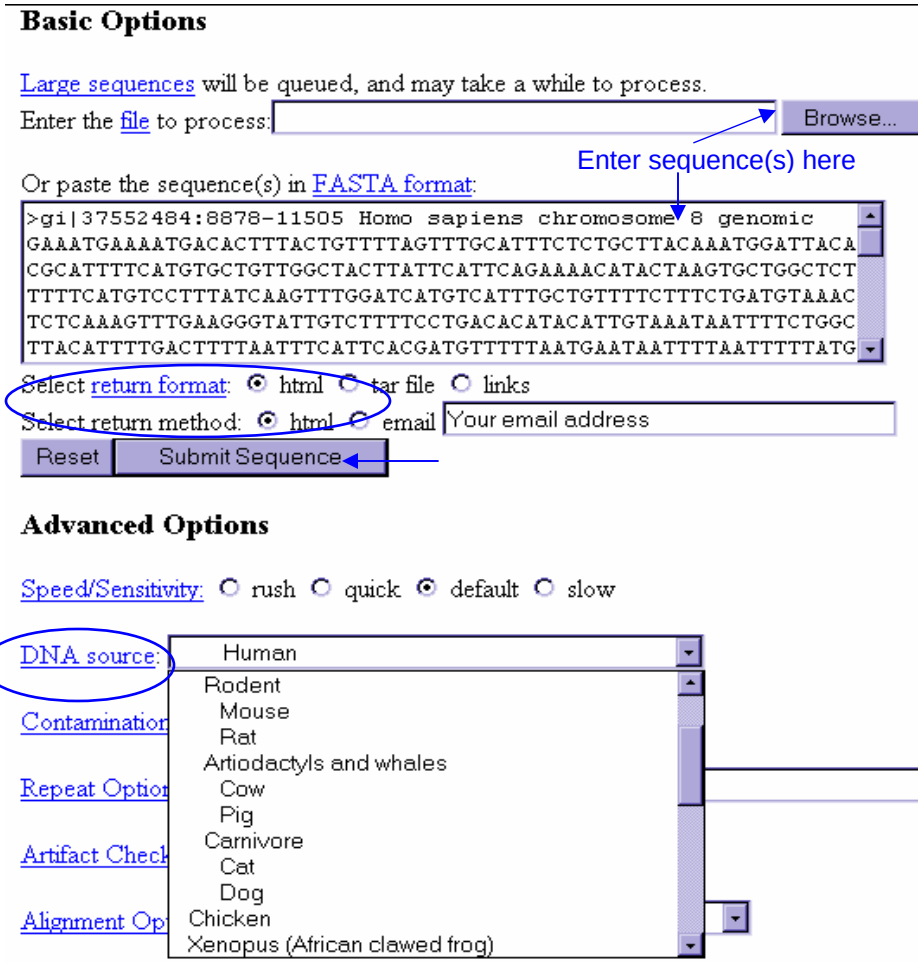
This site is a mirror of the University of Washington site above. The repeatmask help on this site has similar information to that of the University of Washington.

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 repeat	class/family	position in repeat begin	end	(left)	ID
502	19.7	10.2	1.4	gi|37552484:8878-11505	264	433	(947)	+	MLT1H2	LTR/MaLR	68	214	(270)	1
358	21.1	9.8	3.8	gi|37552484:8878-11505	967	1099	(281)	+	MLT1H2	LTR/MaLR	375	515	(34)	1

*Masked Sequence:

```
>gi|37552484:8878-11505 Homo sapiens chromosome 8 genomic
AATGCAAGTTAAAATAATTCTTTCATTGTGGTTTCTGACATGTCATGCCA
ATAAGGGTCTTCTCCTCCAAGAGCACAGAAATATTTGCCAATACTGTCCT
TAAAATCGGTCACAGTTTCATTTTTTATATATGCATTTTACTTCAATTGG
GGCTTCATTTTACTGAATGCCCTATTTGAAGCAAGTTTCTCAGTTAATTC
TTTTCTCAAAGTGCTAAGTATGGTAGATTGCAAACATAAGTGGCCACATA
ATACTCCCACCTCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCACGAAGGAAACAGAG
CCAACCTGCTGCTTCCTGGGGGGAGACAGTCCCTCAGTCCCTCTGTCTCT
GCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCAGTCCC
TCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCC
CTCTGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGA
AGACAGTCCCTCAGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTT
CCTGGAGGAAGACAGTCCCTCTGTCCCTCTGTCTCTGCCAACCAGTTAAC
CTGCTGCTTCCTGGAGGAAGACAGTCCCTCAGTCCCTCTGTCTCTGCCAA
CCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCTGTCCCTCTGT
CTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCTG
TCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACA
GTCACTCTGTCTCTGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNC
AAAAGCTGAGCATTTATCCCAATCCCAATACTGTTTGTCCTTCTGTTTAT
```

Any repeat regions are automatically converted to 'N's in the submitted sequence.

*Sequence was shortened for display purposes.

Summary:

```
==========================================================
file name: RM2sequpload_13454
sequences:                1
total length:         1380 bp   (1380 bp excl N-runs)
GC level:            50.07 %
bases masked:          303 bp ( 21.96 %)
==========================================================
                 number of      length   percentage
Repeat Elements* elements*    occupied  of sequence
----------------------------------------------------------
SINEs:                   0             0 bp      0.00 %
      ALUs               0             0 bp      0.00 %
      MIRs               0             0 bp      0.00 %

LINEs:                   0             0 bp      0.00 %
      LINE1              0             0 bp      0.00 %
      LINE2              0             0 bp      0.00 %
      L3/CR1             0             0 bp      0.00 %

LTR elements:            1           303 bp     21.96 %
      MaLRs              1           303 bp     21.96 %
      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:          303 bp     21.96 %


Small RNA:               0             0 bp      0.00 %

Satellites:              0             0 bp      0.00 %
Simple repeats:          0             0 bp      0.00 %
Low complexity:          0             0 bp      0.00 %
==========================================================

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

Number & Percentage of bases masked

In this example there is a stretch of sequence that is comprised of 303 bases of MaLR sequence, a common repeat element. If a TaqMan® primer or probe were designed across this MaLR sequence (because it was not masked before submission) the oligo could bind to any other MaLRs in the genome. This assay would not be very discriminating or specific because of the number of sequences to which the primers could potentially bind.

II. BLAST® (Basic Local Alignment Search Tool)

After you have selected a target, there are other 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. It is also important to identify all polymorphisms in your sequence of interest. 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

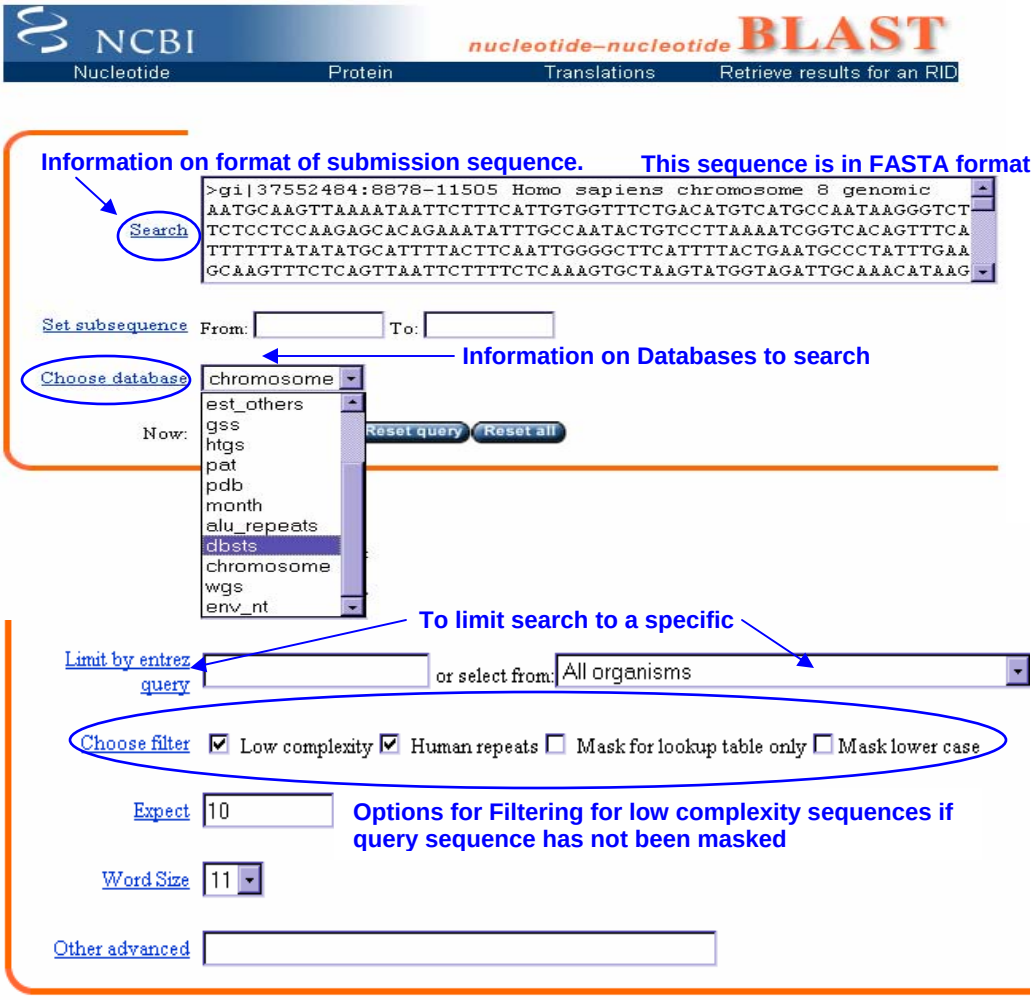
- Go to the NCBI BLAST® site
- Choose **"**Nucleotide-nucleotide BLAST (blastn)" under **Nucleotide**.
- Choose approximately 300 – 600 bases of sequence for your query
- Enter your sequence into the box provided. You may want to search with your masked sequence; the output from RepeatMasker. There are three sequence formats that may be entered into this box. (See pg. 9) For more information on this, click on the word Search next to the box.
- Choose the appropriate database to search.
  Note: For SNP assay design, choose the "chromosome" or "nr" database. Sequences submitted to SNP assay design should not originate from mRNA sequences, as genomic DNA is the template for SNP assays, and assays made to mRNA sequences may be disrupted by intronic sequences.
- 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 from 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
- 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:

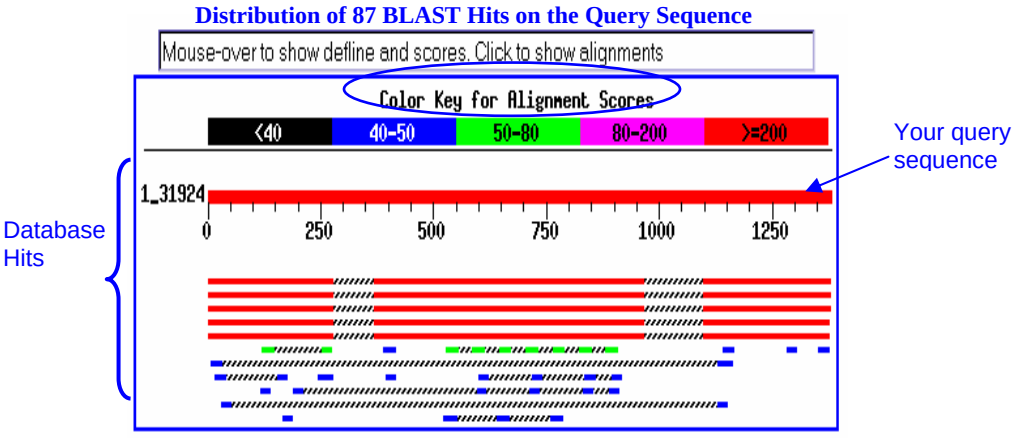
a. Graphical overview
b. List of Sequences producing significant alignments to your query
c. Sequence alignments.

These sections are described below 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 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.

BLAST Output



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.

Click on **Score** to go to sequence alignment

Sequences producing significant alignments:		Score (bits)	E Value
gi|42406292|ref|NC_000008.8|NC_000008	Homo sapiens chromoso...	1181	0.0
gi|42406303|ref|NC_000016.7|NC_000016	Homo sapiens chromoso...	1031	0.0
gi|42406218|ref|NC_000001.7|NC_000001	Homo sapiens chromoso...	898	0.0
gi|42406225|ref|NC_000006.8|NC_000006	Homo sapiens chromoso...	515	e-143
gi|42406224|ref|NC_000005.7|NC_000005	Homo sapiens chromoso...	507	e-141
gi|42406302|ref|NC_000015.7|NC_000015	Homo sapiens chromoso...	52	0.001
gi|42406301|ref|NC_000014.6|NC_000014	Homo sapiens chromoso...	50	0.005
gi|42406297|ref|NC_000010.7|NC_000010	Homo sapiens chromoso...	46	0.079
gi|42406298|ref|NC_000011.7|NC_000011	Homo sapiens chromoso...	44	0.31
gi|42406220|ref|NC_000003.8|NC_000003	Homo sapiens chromoso...	44	0.31
gi|42406221|ref|NC_000004.8|NC_000004	Homo sapiens chromoso...	42	1.2
gi|42406219|ref|NC_000002.8|NC_000002	Homo sapiens chromoso...	42	1.2
gi|50295403|ref|NC_006030.1|	Candida glabrata strain CBS138...	40	4.9
gi|23509224|ref|NC_004317.1|	Plasmodium falciparum 3D7 chro...	40	4.9
gi|31742517|ref|NT_078267.2|	Anopheles gambiae str. PEST ch...	40	4.9
gi|24762252|ref|NT_037436.1|	Drosophila melanogaster chromo...	40	4.9
gi|42406307|ref|NC_000020.8|NC_000020	Homo sapiens chromoso...	40	4.9
gi|42406305|ref|NC_000018.7|NC_000018	Homo sapiens chromoso...	40	4.9
gi|42406304|ref|NC_000017.8|NC_000017	Homo sapiens chromoso...	40	4.9
gi|42406293|ref|NC_000009.8|NC_000009	Homo sapiens chromoso...	40	4.9
gi|42406291|ref|NC_000007.10|NC_000007	Homo sapiens chromos...	40	4.9

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, NC_000008.8. The score is very high (1181), 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 chromosomes, which is the same gene as the query. Keep in mind that what you're looking for is the ability to design an assay that will uniquely detect your sequence of interest. 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.

If you'll notice, the first hit in this list is the query sequences aligned to itself. This will be the first alignment shown, and will be a 100% match to itself.

```
Score = 1181 bits (596), Expect = 0.0
Identities = 596/596 (100%)
Strand = Plus / Plus
```

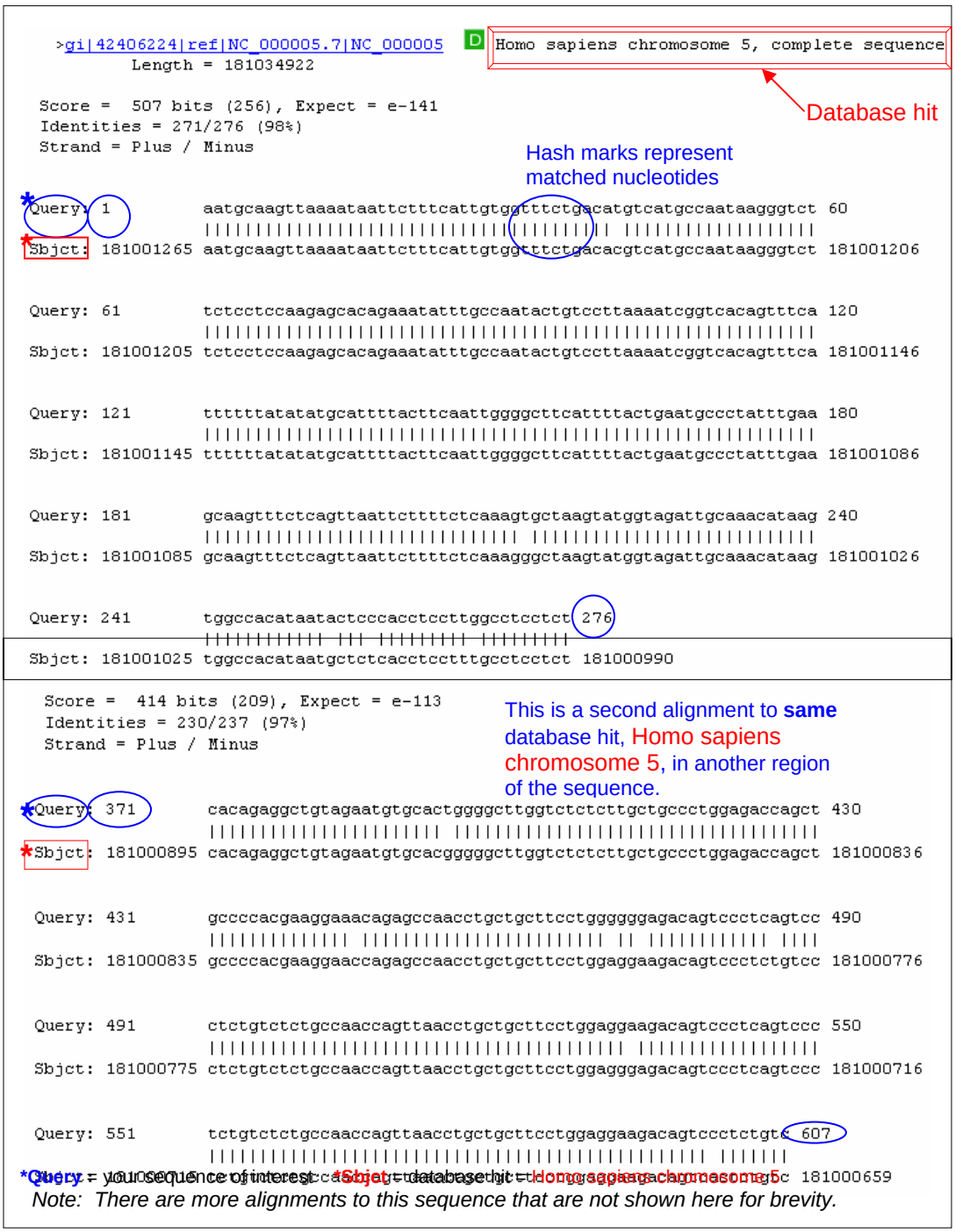
The alignments shown below are from the following genomic sequence from the database.

>gi|42406224|ref|NC_000005.7|NC_000005 Homo sapiens chromosome 5, complete sequence Length = 181034922

Expect = e-141 Identities = 271/276 (98%)
Expect = e-135 Identities = 271/279 (97%)
Expect = e-113 Identities = 230/237 (97%)
Expect = 8e-79 Identities = 158/160 (98%)
Expect = 8e-76 Identities = 153/155 (98%)

Expect = 8e-76 Identities = 153/155 (98%)
Expect = 5e-71 Identities = 151/155 (97%)
Expect = 1e-68 Identities = 144/147 (97%)
Expect = 1e-68 Identities = 144/147 (97%)

Shown above is a list of the Expect values and percent identities for each of the 9 alignments (high-scoring segment pairs; *see HSP in Glossary*) for this database hit, NC_000005. The alignments shown on page 12 are the first and the third alignments listed above from NC_000005, with E values of e-141 and e-113.



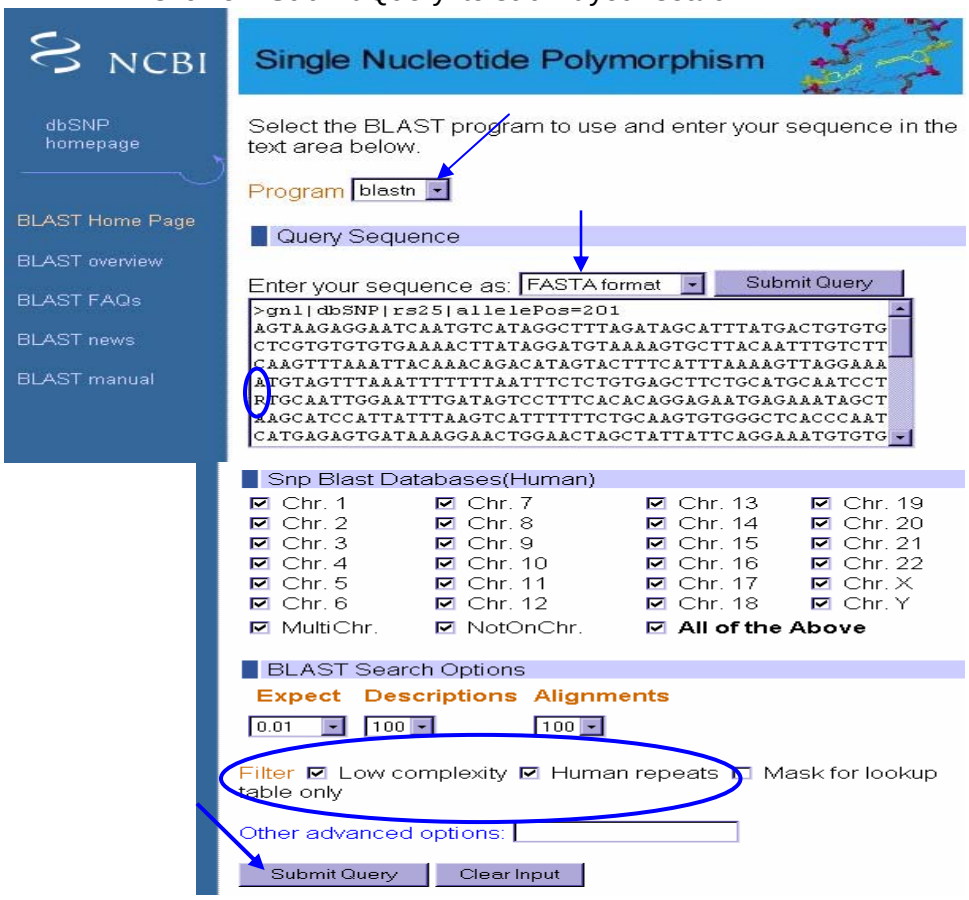
In the example above, notice that the query sequence did not align to this database hit contiguously. The query sequence aligned to this sequence from the database starting at base 1 of the query and ending at base 276, from 371 - 607 of the query sequence, etc. These alignments are shown in order of significance, and have different E-values and scores. Sometimes BLAST alignments of the query to target genomic DNAs are broken into multiple HSP segments if areas of repetitive DNA are contained in these sequences. From the areas masked in RepeatMasker, we know that the sequence between bases 264 – 433 in our query was masked due to MaLR sequences (human repeat elements). This is

the likely reason for the first two HSPs (shown in the above figure). If a segment of your query sequences came up with a significant match to part of a sequence from another gene, you should mask out that region of the sequence in your sequence for submission or simply not include that region in your submission.

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 approximately 300 – 600 bases of sequence for your query
- 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 below.
- If you are searching with a gDNA sequence containing a SNP of interest, use an IUPAC code or 'N' for your SNP of interest so that it is readily identifiable. For example, if your SNP is [A/G], you may want to annotate it is as 'R'. Do not leave it as [A/G] because this will be interpreted as two bases.
- 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|rs25_allelePos=201totallen=691	1206	0.0
gnl|dbSNP|rs2883564_allelePos=500totallen=1000	1202	0.0
gnl|dbSNP|rs2354963_allelePos=499totallen=999	1202	0.0
gnl|dbSNP|rs2040818_allelePos=381totallen=730	909	0.0
gnl|dbSNP|rs57_allelePos=444totallen=730	909	0.0
gnl|dbSNP|rs2040819_allelePos=491totallen=731	887	0.0
gnl|dbSNP|rs58_allelePos=491totallen=731	843	0.0
gnl|dbSNP|rs12531657_allelePos=201totallen=401	675	0.0
gnl|dbSNP|rs26_allelePos=531totallen=731	575	e-161
gnl|dbSNP|rs9937771_allelePos=320totallen=820	371	e-100
gnl|dbSNP|rs9939845_allelePos=388totallen=888	367	1e-98
gnl|dbSNP|rs9925030_allelePos=306totallen=806	367	1e-98
gnl|dbSNP|rs9924427_allelePos=276totallen=776	367	1e-98
gnl|dbSNP|rs4591651 allelePos=401totallen=801	349	4e-93

Sequence Alignments

In the alignment on page 15 there are a few things of which to take note:

1. There is a stretch of 'N's in your query sequence. This is where BLAST® has filtered out a region of low complexity sequence. This region should be masked in your submission sequence.
2. You may or may not be able to see your SNP in the alignment, depending on what part of your sequence is aligned to the database hit.
3. By looking for mismatches in the alignment (no hash marks) you will be able to identify other known, documented SNPs. These SNPs should also be masked out in your submission sequence so that no primer or probe is designed over this area.

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.

Appendix

>gi|37552484:8878-11505 Homo sapiens chromosome 8 genomic
AATGCAAGTTAAAATAATTCTTTCATTGTGGTTTCTGACATGTCATGCCAATAAGGGTCTTCTCCTCCAAGAGCACAGA
AATATTTGCCAATACTGTCCTTAAAATCGGTCACAGTTTCATTTTTTATATATGCATTTTACTTCAATTGGGGCTTCATTT
TACTGAATGCCCTATTTGAAGCAAGTTTCTCAGTTAATTCTTTTCTCAAAGTGCTAAGTATGGTAGATTGCAAACATAA
GTGGCCACATAATACTCCCACCTCCTTGGCCTCCTCTCCCAGGAGGAGATAGCCTCCATCTTTCCACTCCTTAATCTG
GGCTTGGCCATGTGACTTACACTGGCCAATGGGATATTAACAAGTCTGATGTGCACAGAGGCTGTAGAATGTGCACT
GGGGCTTGGTCTCTCTTGCTGCCCTGGAGACCAGCTGCCCCACGAAGGAAACAGAGCCAACCTGCTGCTTCCTGGG
GGGAGACAGTCCCTCAGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCAGT
CCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCTGTCCCTCTGTCTCTGCCAAC
CAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCAGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTC
CTGGAGGAAGACAGTCCCTCTGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCC
TCAGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCTGTCCCTCTGTCTCTGC
CAACCAGTTAACCTGCTGCTTCCTGGAGGAAGACAGTCCCTCTGTCCCTCTGTCTCTGCCAACCAGTTAACCTGCTG
CTTCCTGGAGGAAGACAGTCACTCTGTCTCTGCCAACCCAGTTGACCGCAGACATGCAGGTCTGCTCAGGTAAGACC
AGCACAGTCCCTGCCCTGTGAGCCAAACCAAATGGTCCAGCCACAGAATCGTGAGCAAATAAGTGATGCTTAAGTCA
CTAAGATTTGGGCAAAAGCTGAGCATTTATCCCAATCCCAATACTGTTTGTCCTTCTGTTTATCTGTCTGTCCTTCTCT
GCTCATTTAAAATGCCCCCACTGCATCTAGTACATTTTTATAGGATCAGGGATCTGCTCTTGGATTTATGTCATGTTCC
CACCTCGAGGCAGCTTTGTAAGCTTCTGAGCACTTCCCAATTCCGGGTGACTTCAGGCGCTGGGAGCCCTGTGCATC
AGCTGCTGCTGTCTGTAGCTGAGTTCCTTCACCCCTCTGCTGTCCTCAGCTCCTTCGC

This is the sequence used for the RepeatMasker and nucleotide BLAST sections

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

Custom TaqMan SNP Genotyping 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 135GU01-02

Part Number 4371003 Rev B