

## Exercises in Data Retrieval and Using Blast Searches

[The NCBI web site and its parts are updated periodically, therefore the results given below may change with time.]

More complete instructions are given at the end of the seven exercises.

Step by step instructions with screen shots are available at [http://baygenomics.ucsf.edu/education/traveling/Blast\\_Exercises\\_v4\\_help.html](http://baygenomics.ucsf.edu/education/traveling/Blast_Exercises_v4_help.html)

#1 What is the breadth of information available at NCBI on **cystic fibrosis** in humans?

**Hints:** Do a **All Databases** search at NCBI (<http://www.ncbi.nlm.nih.gov>). Then repeat the search narrowing the return hits to human.

Answer: Lots of data to be explored.

#2 Besides the cystic fibrosis transmembrane conductance regulator gene (CFTR), what other genes are associated with **cystic fibrosis** in humans and what are their roles in the disease?

**Hints:** Perform an **Entrez Gene** search (<http://www.ncbi.nlm.nih.gov>) to find the other genes and their function or relationship to the disease.

Possible Answers:

S100A8 - cystic fibrosis antigen

TGFB1 - mutations modify severity of pulmonary disease in cystic fibrosis patients

TGFB1 - protein and expression correlates with portal tracts showing histological abnormalities associated with cystic fibrosis liver disease

GOPC - CFTR binding

ADRB2 - 2002 polymorphisms contribute to clinical severity and disease progression in cystic fibrosis

2005 - transfected beta3 not beta2-adrenergic receptors regulates CFTR activity via new pathway

ABCB1 - model to study CFTR mutation impacts and possible treatment agent

#3 **Nocturnal asthma** is associated with what gene in humans? What are the RefSeq codes for this gene's mRNA and protein sequences? On the GenBank accession pages, data can be displayed in different formats. What is the difference between default and FASTA formats for these sequence files? How can these RefSeq codes be used to search for similar sequences in other species? What are the results of such a search?

**Hints:** Do a **Gene** search at NCBI (<http://www.ncbi.nlm.nih.gov>), record the codes. Compare the formats of the mRNA and protein sequences. Run a BLAST search, (<http://www.ncbi.nlm.nih.gov/BLAST/>).

Answers:

gene - ADRB2 adrenergic, beta-2-, receptor, surface

RefSeq codes: mRNA Sequence NM\_000024  
Product NP\_000015

FASTA format is very concise, limited to the actual sequence and a identification line that starts with a > symbol. The default format is very verbose, giving all sorts of reference details about the sequence and a version of the sequence that is more easily read by the user.

BLAST searching allows for different types of data entry including the use of accession codes (such as a RefSeq accession code).

ADRB2 contains the **7tm\_1** conserved domain signature which is highly conserved across species.



#7 Are there a knockout mice available to study the AGPAT6 gene? How would you order one of these cell lines?

**Hints:** Find the mRNA FASTA formatted sequence for the AGPAT6 mouse gene by doing an **Entrez Gene** search at NCBI (<http://www.ncbi.nlm.nih.gov>). Then a BLAST search at the International Gene Trap Consortium (IGTC) site (<http://www.genetrap.org>) to see if such knockouts exist.

Answer:

There are three possible knockouts. However, two of them occur at the same place.

Information on how to order a cell line is provided in the Sequence Tag Information section.

In this case:

DTM030 is from BayGenomics and can be ordered from MMRRC

<http://www.mmrc.org/distribution/cellLines.html>

XS0453 and XS0575 are from the Sanger International Gene Trap Resource ordered from SIGTR

<http://www.sanger.ac.uk/PostGenomics/genetrap/clones/>

Step by step instructions for the exercises.

Question #1

What is the breadth of information available at NCBI on **cystic fibrosis** in humans?

1. Go to <http://www.ncbi.nlm.nih.gov>, enter **cystic fibrosis** in the **for** box and click **Go**.
2. The returned **Entrez** page is organized with literature matches in the top box, sequence based information in the middle one and NLM's resources in the bottom box. Items of possible interest are denoted by numbers in a white box next to a topic title.

There are 146 OMIM entries [catalog of human genes and genetic disorders].

OMIM background information with links off to help and frequently ask questions

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

The sequence based information section contains data on cystic fibrosis from all species.

Some of these topics are sensitive to the addition of species information to the search query. OMIM and MeSH are examples of this.

3. To restrict the sequence data to that from humans, add **homo sapiens** to the current **cystic fibrosis** in the **Search across databases** box and click **Go**.
4. The number of matches for nucleotide, protein, gene topics have decreased, but, there still is a large number of items to sift through.

Question #2

Besides the cystic fibrosis transmembrane conductance regulator gene (CFTR), what other genes are associated with **cystic fibrosis** in humans and what are their roles in the disease?

1. Go to <http://www.ncbi.nlm.nih.gov>, change the **Search** option from **All Databases** to **Gene** using the pull down menu, enter **homo sapiens cystic fibrosis** in the **for** box and click **Go**.
2. Click on at least five diverse hits below the CFTR gene, finding out their relationship to **cystic fibrosis**. Ignore any gene that doesn't have a **NCBI Reference Sequences (RefSeq)** section.

Click on any **PUBMED links** in the middle of the page and scan through the titles for mention of **cystic fibrosis**.

If no papers are listed with **cystic fibrosis** in the description, check out the **MIM** link at the top of the page or the links in the **GeneOntology** section.

Possible answers:

- CFM1 - no RefSeq data (ignored)
- CFM2 - no RefSeq data (ignored)
- S100A8 - cystic fibrosis antigen
- TGFB1 - mutations modify severity of pulmonary disease in cystic fibrosis patients
- TGFB1 - protein expression correlates with portal tracts showing histological abnormalities associated with cystic fibrosis liver disease
- GOPC - CFTR binding
- ADRB2 - 2002 polymorphisms contribute to clinical severity and disease progression in cystic fibrosis  
2005 - transfected beta3 not beta2-adrenergic receptors regulates CFTR activity via new pathway
- ABCB1 - study to see how the common cystic fibrosis mutation might disturb transmembrane segments of the protein using ABCB1 as a model
- ABCB1 - ABCB1 expression increases ATP release in respiratory cystic fibrosis cells potential clinical benefits discussed

Note there is another database that is relevant for getting clinical information, Online Mendelian Inheritance in Man (OMIM or MIM). Although OMIM can not be searched via an actual sequence, it does allow searching by gene symbol, chromosome location, keywords or other features.

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=omim>

Question #3

**Nocturnal asthma** is associated with what gene in humans? What are the RefSeq codes for this gene's mRNA and protein sequences? On the GenBank accession pages, data can be displayed in different formats. What is the difference between default and FASTA formats for these sequence files? How can these RefSeq codes be used to search for similar sequences in other species? What are the results of such a search?

1. Go to <http://www.ncbi.nlm.nih.gov>, change the **All Databases** Search option to **Gene** using the pull down menu, enter **homo sapiens nocturnal asthma** in the **for** box and click **Go**.
2. Check the resulting hits to insure that the summary information on the gene mentions that various types of changes in this gene are associated with the disease.

ADBR2 adrenergic, beta-2-, receptor, surface

3. Scroll down the page to the **NCBI Reference Sequences (RefSeq)** section. Record the mRNA sequence and Product (protein) codes:

mRNA Sequence NM\_000024

Product NP\_000015

4. Click on the mRNA code to see the data on the actual mRNA sequence data. Scroll down the page taking in the format of the information presented.
5. Scroll back to the top of the page and change the **Display** option from **GenBank** to **FASTA**.

The format automatically changes. Note the difference. FASTA format is the sequence format required by many database searching programs.

background information on fasta format

<http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml>

6. Click back to the **Entrez Gene** page and repeat this process with the protein code.
7. After noting the difference, click on the NCBI logo at the top of the page. From the blue navigation bar on the main NCBI page, click on **BLAST**.
8. From the main BLAST page, click on **protein blast** in the **Basic BLAST** section. In the **blastp suite** page, click on the “? **Icon**” above the large box to find out about what sort of inputs this form accepts. Clicking the **more...** link provides additional details.

After reading the presented information, close the popup window and the original “?” block, then enter the protein code into the **Enter Query Sequence** box. Once this is done, information appears in the **Job Title** box. In the **Choose Search Set** section of the page, change the **Organism** button to **Custom** and start to type the term **Vertebrata** into the **Organism** box. As the term is entered, matching **Entrez** terms start to appear. When enough of the word is entered to find the desired term, select this item from the list. Clicking the **BLAST** button at the bottom of the page starts the search. If results are to be displayed in a new window, click on the “Show results in a new window” box prior to clicking the **BLAST** button.

Protein searches gets around the problem of multiple codons coding for the same amino acid that impacts nucleotide searches. However, depending in the information being sought, this is not always possible.

9. It may take a few seconds for the search to be completed. While waiting, click on the **7tm\_1** in the image to find out about the conserved domain that was found in the sequence.

**7tm\_1** indicates that the protein being search with contains the transmembrane receptor signature of the rhodopsin family of transmembrane proteins.

This signature is located from residue 50 to 325 in the sequence.

Close the pop up window.

Note the length of the query sequence, this may be given on the Job Title line. **413 letters**

Wait for the results page to appear.

10. Scroll down the results page past the image with it colored horizontal bars to the **Sequences producing significant alignments** section.

Scores are based on the length of the query sequence and the size of the database. Short sequences will never produce great scores. To get a E value of 0.0 requires a match of at least 330 characters. A very long sequence could easily have a match this long and still not have a match that covers a significant portion of the query sequence. Always look at the resulting alignment. The mathematics of the process can sometimes result in the strange ordering of hits.

A hit line gives the GI number of the match sequence, the database it is from, the accession code used for the sequence, the description of the sequence from the database, its Bits score and finally the E value. Hits in the list are ordered by their E value, then their Bits score, which reflects the length of their actual match. Enough of the description may be given to see what species the sequence is from.

Clicking on the link given on the left side of a hit line goes off to the actual sequence information. Clicking on the right side link moves down to the alignment data for that hit.

Notice that there are over 30 hits with an E value of 0.0 at the top of this list and that the protein code entered is not at the top of the list. There are about 130 hits in the list which mention **ADBR2, beta-2 adrenergic receptor** or variations thereof before sequence description changes to something else. The first 12 hits are all from man with from 0 to 2 mismatches in the alignment.

NCBI used to make an effort to remove redundant sequences, but the size of the database increased to such an extent that it was no longer possible to do this quickly enough so that it wouldn't impact the processing of new data.

When an accession code begins with **XP\_**, it means that the data is the results of an automated analysis process. This situation usually occurs when a genome sequencing project is first being analyzed. These sequences have not been checked for accuracy and can be much longer or shorter than their homologs from more mature genome studies. These sequences usually have their description start with **PREDICTED:**.

Check out some of the hits beyond the 0.0 E values and determine where the match is actually taking place within the query sequence.

PREDICTED: similar to beta-2 adrenergic receptor [Gallus gallus] Length=397  
 23 - 399  
 beta-2 adrenergic receptor [Homo sapiens] Length=275  
 52 - 326  
 beta-2 adrenergic receptor [Macaca mulatta] Length=275  
 52 - 326  
 beta-2 adrenergic receptor [Hylobates concolor] Length=275  
 52 - 326  
 beta-2 adrenergic receptor [Ateles fusciceps] Length=275  
 52 - 326  
 ....  
 beta-2 adrenergic receptor [Hippopotamus amphibius] Length=275  
 52 - 326

The match is happening in the **7tm\_1** region of the sequence which appears to be highly conserved.

#### Question #4

Are there any solved protein crystal structure(s) for the **nocturnal asthma** gene? Does the structure include the transmembrane segments? Are the found structure protein and the **nocturnal asthma** protein closely enough related to believe the results?

1. Go to <http://www.ncbi.nlm.nih.gov>, and from the blue navigation bar click on **BLAST**.
2. From the main BLAST page, click on **protein blast** in the **Basic BLAST** section. In the **blastp suite** page, click on the “**? Icon**” at the end of the **Database** menu line in the **Choose Search Set** section to find out information about the databases that can be used in a protein BLAST search. Clicking on the **more...** link provides additional information. Once a suitable structure database name has been located, close the **more...** page and re-click on the “**? Icon**” to close the information block.

From the list given, the structural database to use is **pdb**. The **swissprotein** database was also listed.

Of the protein databases, **swissprotein** is considered to have the best annotation. One of the features they report is transmembrane segment locations when available or predicted.

3. Change the **Choose Search Set Database** option from **nr** to **swissprotein** using the pull down menu, enter the previously found RefSeq protein accession code into the **Enter Query Sequence** box.

To speed things up and reduce the size of the output file, restrict the organisms searched to humans by changing the **Any** organisms option to **Human** by clicking the **Human** button in the **Organism** line. Start the run by clicking the **BLAST** button.

4. At the top of the actual results page, click on the “Reformat these Results” link. This leads off to a form which allows the changing of the produced results. The number of descriptions, lines in the image and alignments can be restricted using the **Descriptions:**, **Graphical overview:** and **Alignments:** pull down menus. Restrict these three options to 10 each and then click on the **View report** button near the top of the page.
5. Scroll down the results to the significant alignments section and click on the sequence link containing the term **ADRB2\_HUMAN**. It should be the first one on the list.
6. Scroll down the **swissprotein** data file to the FEATURES section. Then read through the listed features to find those regions called “**Transmembrane region**”.

transmembrane segments: 1. 35 - 58    3. 107 - 129    5. 197 - 220    7. 306 - 329  
 2. 72 - 95    4. 151 - 174    6. 275 - 298

7. Return to the protein blast page, re-enter the RefSeq accession code if necessary, and change the database to be used to **pdb**, return **Organism** to its default **Any** value, and then click the **BLAST** button.
8. Wait until results page appears.
9. The best hit comes from Bovine Rhodopsin, but it is not very strong. The alignment does cover the

entire area containing the transmembrane segments.

10. Return to the main NCBI page, change **All databases** to **Gene** and enter **rhodopsin homo sapiens** in the **for** box and click **Go**.

11. The gene symbol for rhodopsin is RHO from the result of this search.

Click on the RHO link to get to the **Entrez Gene** page. Scroll down the page to find the RefSeq protein accession.

Product NP\_000530

12. Return to the main NCBI page, click on **BLAST** from the blue navigation bar. This time on the main BLAST page, click on the "**Align two sequences using BLAST (bl2seq)**" link in the **Specialized BLAST** section.

13. In the BLAST 2 SEQUENCES page, change the Program from **blastn** to **blastp**, enter the protein accession code for ADRB2 in the sequence 1 box and the RHO code in the sequence 2 box and click **Align**.

The results page shows that the two human proteins are not highly related to one another.

Solving transmembrane protein structures is very difficult. Perhaps a few more structures for this protein family should be obtained before believing the alignment results.

#### Question #5

Find proteins that are known to contribute to **pulmonary artery hypertension** and determine if animal models exist in which the disease can be studied. Can a full length dog protein sequence be found?

1. Go to <http://www.ncbi.nlm.nih.gov>, change the **All Databases** option to **Protein** and enter the term **pulmonary artery hypertension homo sapiens** into the **for** box and click **Go**.

2. Choose the top hit, in this case NP\_001195. Clicking on the link will take you to the page for the protein.

3. Check out the length of the protein. The length is the second item on the "LOCUS" line.

(The sequence to be used in the search is 1038 residues long.)

4. To obtain the sequence of NP\_001195 for the BLAST search, change the **Display** option of the page from **GenPept** to **FASTA** using the pull down menu. This automatically changes the format to FASTA. Copy this data, starting with the ">" and continuing to the end of the sequence.

5. Click on the **NCBI logo** in the upper left-hand side of the screen.

6. From the main NCBI page click on **BLAST** in the blue navigation bar. On the main BLAST page, click on **protein blastp** in the **Basic Blast** portion of the page.

7. Paste your sequence into the **Enter Query Sequence** box, be sure that the **Choose Search Set** parameters are at their default values (database **nr** and organism **Any**), and then click the **BLAST** button.

8. Wait until the result page appears.

9. Check out the best non-human hits with a description that appears to be correct. Checking out the best hits to determine the quality of the matches and the species results in the following information.

XP_001101663	[Macaca mulatta]	1038	residues	Identities = 1029/1038 (99%)
NP_031587	[Mus musculus]	1038	residues	Identities = 1003/1038 (96%)
XP_001065181	[Rattus norvegicus]	1038	residues	Identities = 998/1038 (96%)
NP_001001465	[Gallus gallus]	1031	residues	Identities = 918/1021 (89%)
XP_617592	[Bos taurus]	914	residues	Identities = 868/898 (96%)
AAB39883	[Xenopus laevis]	1048	residues	Identities = 784/1028 (76%)

These results would indicate that rehesus monkey [Macaca mulatta], mouse and rat would all be good animal models in which to study this gene and its function.

10. Return to the **blastp suite** page and change the **Organism** option in the **Choose Search Set** section from **Any** to **Canis familiaris** by clicking the **Custom** button and starting to enter this term into the field. Highlight the term when it appears on the list. Click the **BLAST** button to start the run.

Confirm that the hits are from dog and determine how close the length is to that of the starting human sequence. The first two look the most likely.

XP\_536035            759 aa            linear    MAM 31-AUG-2005  
DEFINITION PREDICTED: similar to Bone morphogenetic protein receptor type II precursor (BMP type II receptor) (BMPR-II) (BRK-3) [Canis familiaris].

XP\_851509            248 aa            linear    MAM 31-AUG-2005  
DEFINITION PREDICTED: similar to Bone morphogenetic protein receptor type II precursor (BMP type II receptor) (BMPR-II) (BRK-3) [Canis familiaris].

11. Return to the main BLAST page, and click on the “**Align** two sequences using BLAST (bl2seq)” link in the **Specialized BLAST** section.
12. On the BLAST 2 SEQUENCES page, change the **Program** from **blastn** to **blastp**, enter the protein accession code for the human protein sequence in the sequence 1 box and the first dog code in the sequence 2 box, click **Align**.

For XP\_536035 the alignment goes from 1 - 758 of the human sequence and the identity is 97%.

Return to the BLAST 2 SEQUENCES page, change the code in sequence 2 box to that of the second hit and click **Align**.

For XP\_851509 the alignment goes from 791 - 1038 and the identity is 92%.

There appears to be about a 30 residue gap in the two dog sequences that needs to be filled before the sequence is complete.

#### Question #6

How conserved are the ATP2A2 proteins across vertebrate species? Should all the available protein sequences be used to make this assessment?

1. Go to <http://www.ncbi.nlm.nih.gov>, change for **Search** option from **All Databases** to **Gene**, enter **ATP2A2** in the **for** box and click **Go**.
2. Work through the list of exact gene name matches. First, go the **Entrez Gene** page for each entry. Scroll down this page to the **NCBI Reference Sequences (RefSeq)** section. If this section is there, record the RefSeq name for the protein product and check to see how long the sequence is by clicking on the link. Record the length as well. If a **Related Sequences** is only available, record the protein’s name and check on it length by clicking on the link.

There are 8 exact gene name matches:

frog [Xenopus laevis], human, dog [Canis familiaris], mouse, chimp, rat, cat [Felis catus], and chicken [Gallus gallus]

From the information contained in the human **Entrez Gene** page, there are two known isoforms for this gene. One whose length is 997 and another whose length is 1042 (NP\_733765). Use the longer one in your analysis runs.

4. Use this RefSeq name to do a **BLAST** search on all vertebrate sequences, by returning to a page with the **NCBI logo** on it and clicking on that logo. From the NCBI main page, click on **BLAST** in the blue navigation. Click on the **protein blast** link in the **Basic BLAST** section to get to the protein search form. Enter your RefSeq name into the **Enter Query Sequence** box, change **Any** to **Vertebrata** in the **Organism** line and change the searched database to **swissprotein** and then click the **BLAST** button.
5. Wait until a results page appears.
6. Scroll down the page to the significant alignments section. Check out the top hits to see what the **swissprotein** family code is for the gene symbol **ATP2A2**.

The swissprotein family name appears to be **AT2A2**.

7. Repeat this process using the **nr** database to find other possible sequences. Return to the Blast submission page and change the database from **swissprotein** to **nr** using the pull down menu. Check out the results to find full sequences (around ~ 1038 residues) from more species.

Possible additional sequences are:

XP_001141715	chimp [Pan troglodytes]	1042 residues	Identities = 1037/1042 (99%)
CAJ42045	horse [Equus caballus]	1042 residues	Identities = 1032/1042 (99%)
AAH98958	frog [Xenopus laevis]	1042 residues	Identities = 961/1042 (92%)

From this list of possible sequences, those starting with XP\_ probably won't work at the ClustalW site.

8. Go to the CLUSTALW site (<http://www.ebi.ac.uk/clustalw/>). Data can be entered in a number of ways on this page. But, notice the new feature at the top of the page. This feature allows the **EBI** local databases to be searched for sequences of interest.

Replace the **Enter Text Here** in the box with a swissprotein accession code and click **Go**.

Protein sequences are needed for the alignment and there is one listed on the page. Click on the **Protein Sequences** link.

There is one match in the UniProt database, to see the actual data file, click on the **in UniProt format** link in the View line.

What appears next is the local copy of the **AT2A2\_HUMAN** data file. Copy the entire data file and then return to the starting **EBI ClustalW page** and paste it into the **Enter or Paste** box.

Repeat this process with all 8 of the found swissprotein family members. You will need to scroll down to the bottom of the **Enter or Paste** box each time to add your new data. Be sure that there is a cursor below the // of the last entry, below this you paste in your new data.

After all your sequences have been pasted in, click the **Run** button.

9. The waiting page, that lets you know that the site is working on your alignment, is replaced by your results.

In the first part of the page is a box listing the **ClustalW Results** of the search with a **JaView** button. Click it and see if a colored image comes up. [This may not work with all browsers.]

- 10a. If the image appears, the columns are colored according to the residue type. A column that is all the same color and which contains the same character in each row is a conserved position. Care needs to be taken with the A,I,L,M,V locations, as the site considers these residues interchangeable, but they are not. Move across the alignment and see how it changes with position.
  - 10b. If the image doesn't appear, move down the page to the **Alignment** portion. Click the **Show Colors** button. This produces a new page where the alignment position characters are colored according to residue type. Again, care needs to be taken with the A,I,L,M,V locations. All the positions with an "\*" in the line below the members of the alignment have the same character in that location.
11. Depending on how careful you were in selecting sequences, there could be possible changes.

There could be two types of proteins here based on length. Looking at the human **Entrez Gene** page showed that there were two isoforms of this gene. They only seem to differ at the C terminus end, where one form is longer than the other.

If you have more than one sequence from the same species, you might want to drop the shorter one. You might want to try to add additional sequences to this alignment using the codes that were collected. There is a frog code, **AAH98958** you could search for and add to your alignment. EBI doesn't have predicted RefSeq data, but it might have most other protein sequence codes. You just have to try a search and see.

Adding the frog data changed the aligned data. The least conserved area is at the very end, between 993 and 1044. Only the cat and dog sequences seem to be short. Perhaps their sequence data is not for the same isoform as the others.

The members of the alignment appear to be very highly conserved when the same isoform is used across species.

## Question #7

Are there a knockout mice available to study the AGPAT6 gene? How would you order one of these cell lines?

1. Go to <http://www.ncbi.nlm.nih.gov>, change the **Search** option to **Gene**, enter **AGPAT6** in the **for** box and click **Go**.
2. Click on the mouse link on the search results page to go mouse **Entrez Gene** page. Scroll down this page to the **NCBI Reference Sequences (RefSeq)** section. Click on the mRNA sequence link.

mRNA Sequence NM\_018743

3. Convert the sequence in the default format to a FASTA formatted file by changing the **Display** option to **FASTA**. Copy this sequence.
4. Go to the IGTC web site (<http://www.genetrap.org>). Click on **DATA ACCESS** in the blue navigation bar to see the options available. Select the **Blast Search** option.
5. Paste your mRNA sequence into the **Enter sequence below** box and click **Quick Search**.

Using an mRNA sequence for a Blast search at this site allows the detection of standard loss-of-function allele cell lines. To find intronic ones would require the use of the sequence for the genomic region occupied by the mRNA sequence.

6. Scroll through the results and look at the actual alignments.

The IGTC web site uses strict guidelines in associating a cell line with a gene. A match needs to be at least 50% of the cell line length and have an identity of at least 90%.

Using this criteria the following three cell lines are associated with the AGPAT6 gene:

DTM030, XS0453 and XS0575.

The other cell line in the top four CMHD-GT\_184C11-3 does not meet this criteria.

7. Click on one of these cell line links to go off to a cell line annotation page. Here data is presented on the cell line, the gene it is associated with, and an image is given displaying the location of the cell line with respect to the gene's mRNA sequence.

To see all the data that is available on this page, click the **Show All** arrow in the **Additional Information** section. To hide this information again, click on **Hide All**.

In the **Sequence Tag Information** section of the page, information is provided on the source of the cell line and how to order it via a provided link.

In this case:

DTM030 is from BayGenomics  
order from MMRRC

<http://www.mmrrc.org/distribution/cellLines.html>

XS0453 and XS0575 are from the Sanger International Gene Trap Resource  
order from the SIGTR

<http://www.sanger.ac.uk/PostGenomics/genetrap/clones/>

Clicking on the Gene Description link goes off to a Gene Annotation page.

The image shows that cell lines XS0453 and XS0575 occur in approximately the same place, while DTM030 is further down stream.