**DNA Sequencing & Analysis Using BLAST**

DNA sequencing involves the determination of the sequence of nucleotides (A, T, C and G) in a molecule of DNA. That sequence information can be used to verify that the DNA being sequenced is, in fact, the expected DNA molecule; or, the sequence can be compared to other sequences to analyze similarities and differences across different known molecules (to infer evolutionary relationships, for example). The DNA sample, or ***template***, can be a plasmid or a PCR product, and requires a primer at the 5’ end of the section to be sequenced. In a typical sequencing reaction, usually only about ~500-800bp are reliably sequenced, and because the first stretch of nucleotides sequenced are usually prone to errors, the primer needs to be ~50bp upstream of where accurate sequencing information is needed. Depending on the reaction conditions, the template, and the type of sequencing being performed, errors begin to appear after the first 500bp. Therefore, if more than 500bp need to be sequenced, multiple sequencing reactions with different primers need to be performed on the same template.

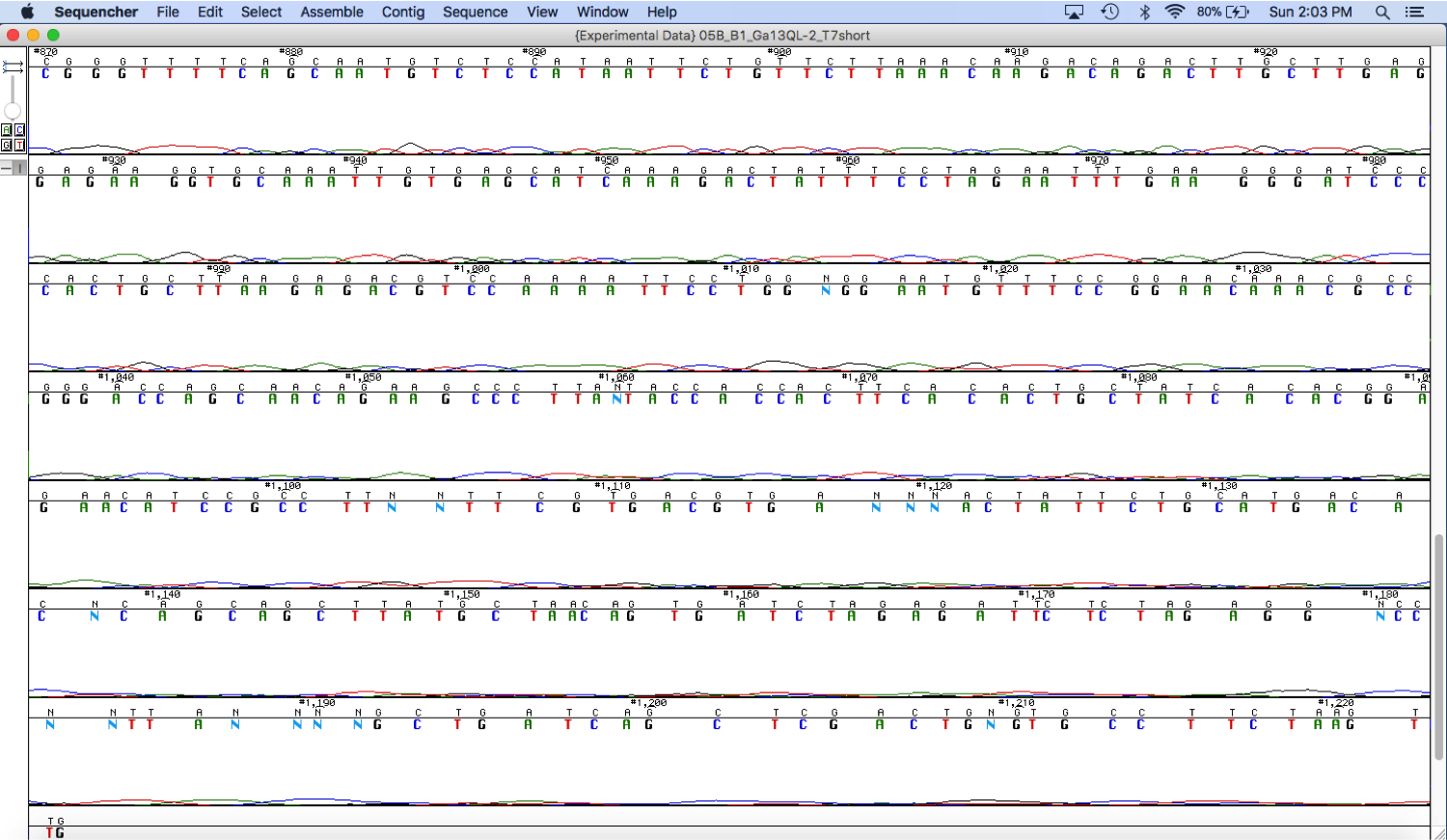
This guide will go over the basics of sending off a plasmid or purified DNA fragment from PCR for sequencing and how to analyze it. When a template is sequenced, the results are displayed in a ***chromatogram*** or an ***electopherogram***. While both chromatograms and electropherograms can look the same and present the same kind of data (i.e., nucleotide sequence), they derive from the type of sequencing that was performed; this guide will refer to chromatograms in the example below.

**Sample Submission for Sequencing**

There are a few different commercial vendors that will sequence DNA samples, in addition to some university core facilities; each will have their own specific guidelines for how to submit the samples. Generally, 10ul of sample is required per sequencing reaction, and the concentration required will depend on the size of the template. Some vendors will offer their own reliable universal sequencing primers for certain common plasmids, but customers can submit their own primers as well. There are many primer design tools available online; the University of Michigan has a good overview for primer design for sequencing (<https://seqcore.brcf.med.umich.edu/content/how-do-i-design-my-own-primers>; see also the PCR Guide for general primer design guidelines). When submitting your own primers, 5-10ul is usually required per reaction at about 5uM. Sometimes, a vendor might require that the template DNA be purified in a specific way and resuspended in buffer or water, usually lacking EDTA (the EDTA in TE buffer, for example, will inhibit the sequencing reaction); alternatively, some vendors may offer to purify the template DNA from a PCR reaction for an additional charge. For PCR products, an image of the product on an agarose gel might be required to ensure that the PCR reaction worked and that only one band was amplified. If it’s the first time using a particular vendor, it is probably a good idea to give them a call to talk about what you need from them and what exactly they will need from you.

**Sequencing Results**

Depending on the sequencing vendor, the results will be delivered to you electronically, either by a DropBox-type method or simply by email. Two types of files are usually given: one file contains the sequence (A, T, C, G, and sometimes N if the nucleotide signal is ambiguous) and the other contains the electropherogram/chromatogram, which is the visual representation of the actual data collected during the sequencing run. An example chromatogram is shown below:

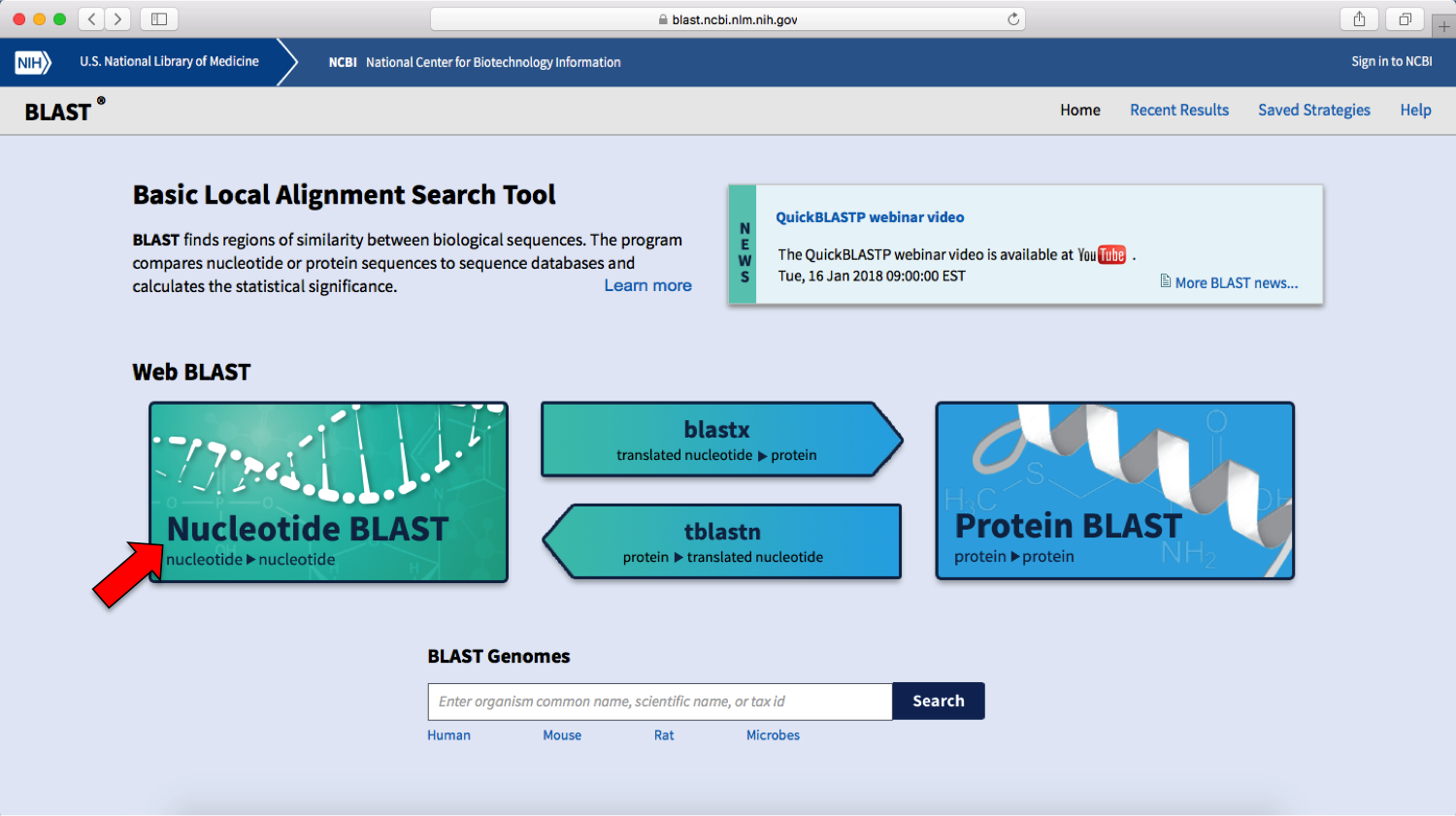


Each nucleotide emits a kind of signal during the sequencing reaction, which is depicted in the graph by a peak, and each nucleotide is assigned a specific color; here, A is green, T is red, C is blue and G is black. The sequence is constructed from the chromatogram by “calling” the nucleotides based on the peaks. The beginning of the run usually starts with ill-defined peaks, then clear sharp peaks appear for a few hundred nucleotides and the signal gradually tapers off to the end of the sequence. At some point, the signal becomes too ambiguous to “call” what the nucleotide is, and more and more “N”s appears in the sequence. There can also occasionally be errors in the sequence, especially when peaks are close together, small or flattened out (which may happen when two or more nucleotides are next to each other), so it is very important to look over the chromatogram and make any corrections to the sequence as needed.

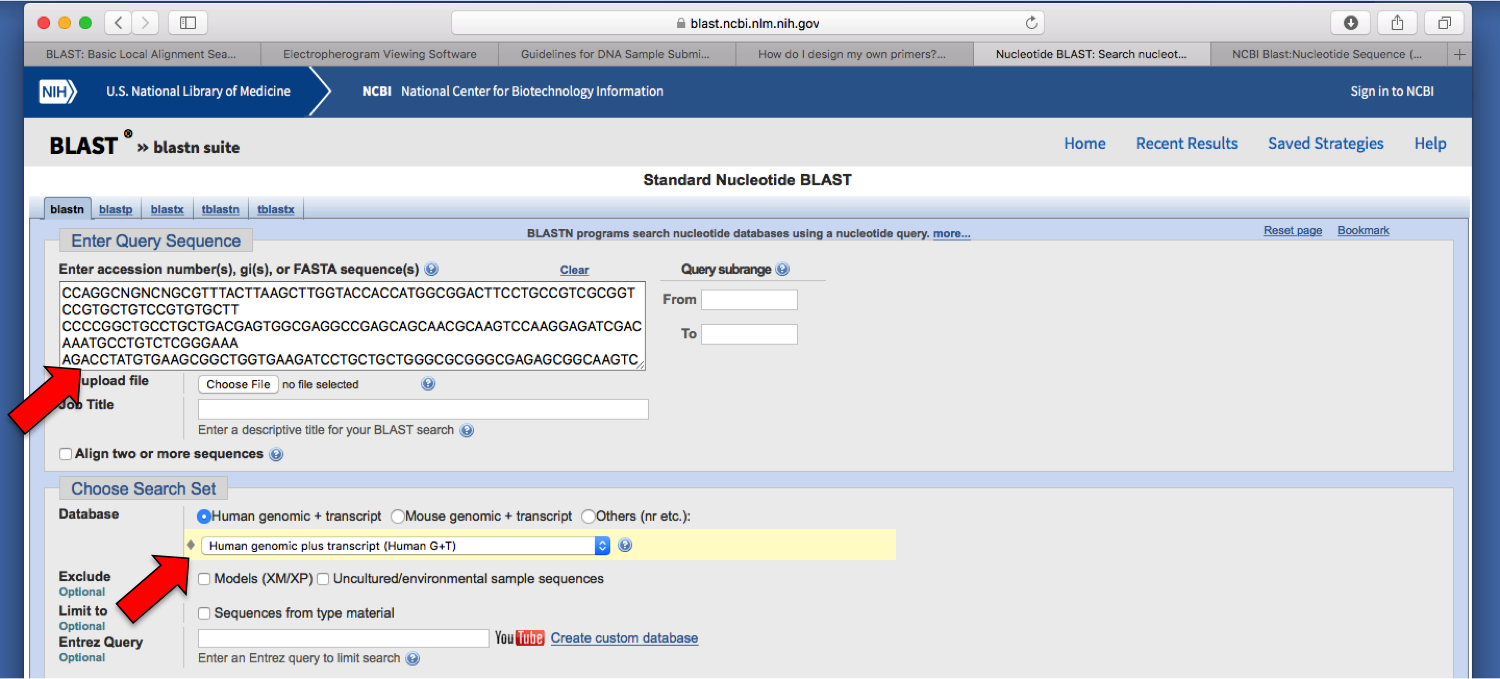
There are a couple different ways these files can be visualized. Sometimes a vendor may have a free online tool for visualizing sequence and chromatogram or electropherogram files; otherwise, there are plenty of free tools available to download. For example, a free viewer called Sequencher can be used to look at a chromatogram; and a free program called Serial Cloner can be used to look at the sequence data (and is also convenient for other molecular cloning applications). Whichever tools you decide to go with, it is helpful to have both the sequence file and the chromatogram file open at the same time, to verify that the sequence accurately represents the data in the chromatogram.

**Using BLAST to Analyze Sequencing Results**

BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is an algorithm for comparing sequences of nucleic acids or proteins and finding regions of similarity between them. A specific sequence can be compared against a database of known sequences of a particular type (e.g., all human sequences), or a few sequences can be compared with each other. After getting sequencing results back and verifying the sequence with the chromatogram/electropherogram, the sequence can be aligned in a suitable BLAST database to verify the identity of the DNA sample.



There are many different types of BLAST analyses these days (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>): nucleotide BLAST, protein BLAST, blastx (translated nucleotide vs. protein), tblastn (protein vs. translated nucleotide), Primer BLAST for designing primers, IgBLAST (specific search for immunglobulins and T cell receptor sequences), etc. In this example we’ll do a straightforward nucleotide BLAST with a sequence from an expression plasmid containing the human G Protein-Coupled Receptor (GPCR) alpha 13 subunit. On the BLAST webpage, copy and paste the sequence data into the Query Sequence box.

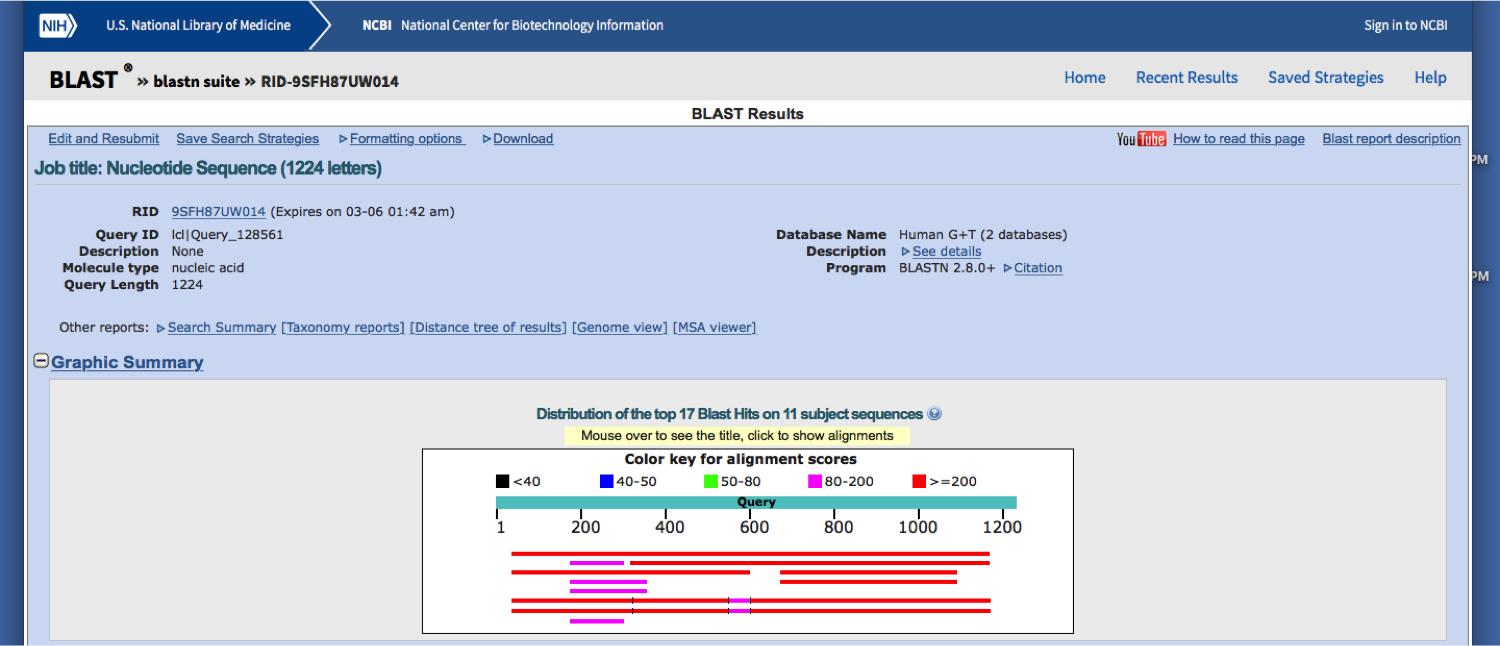


Choose a Search Set based on the database that you think your sequence will show up in. In this example, we want to confirm that the sequence belongs to a human protein, so select “Human genomic + transcript.”

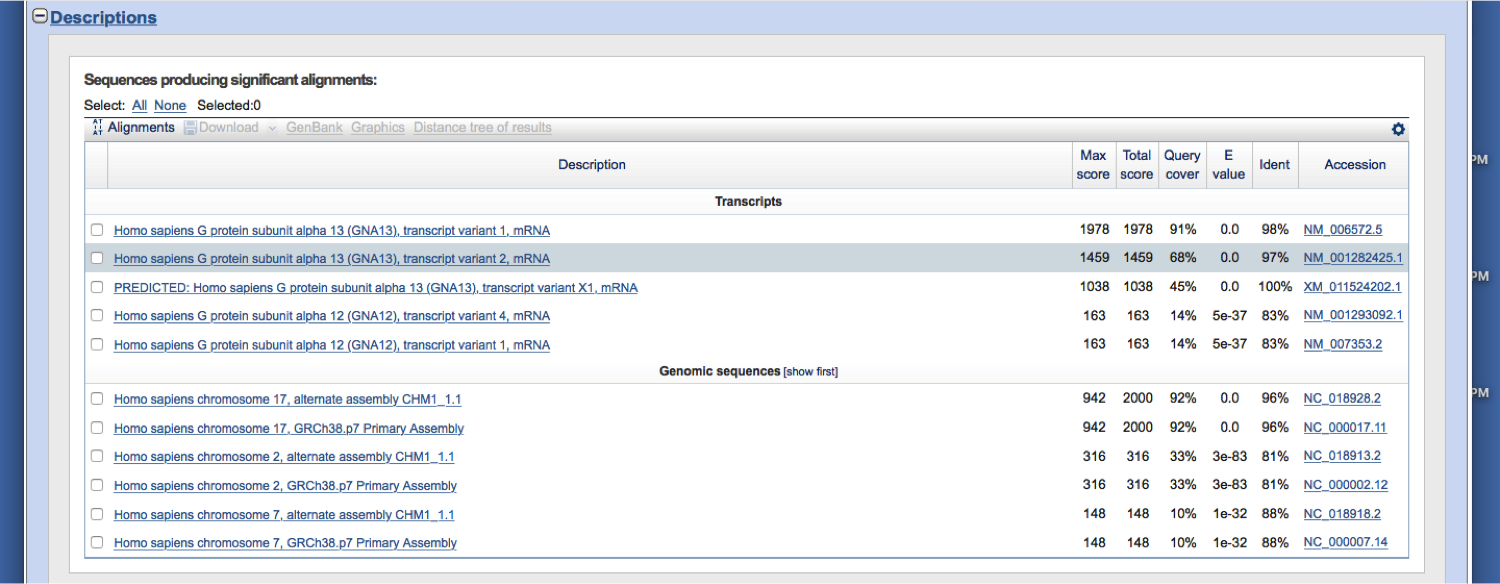
Keep the rest of the default settings, check “Show results in a new window” for convenience, and click the “BLAST” button.



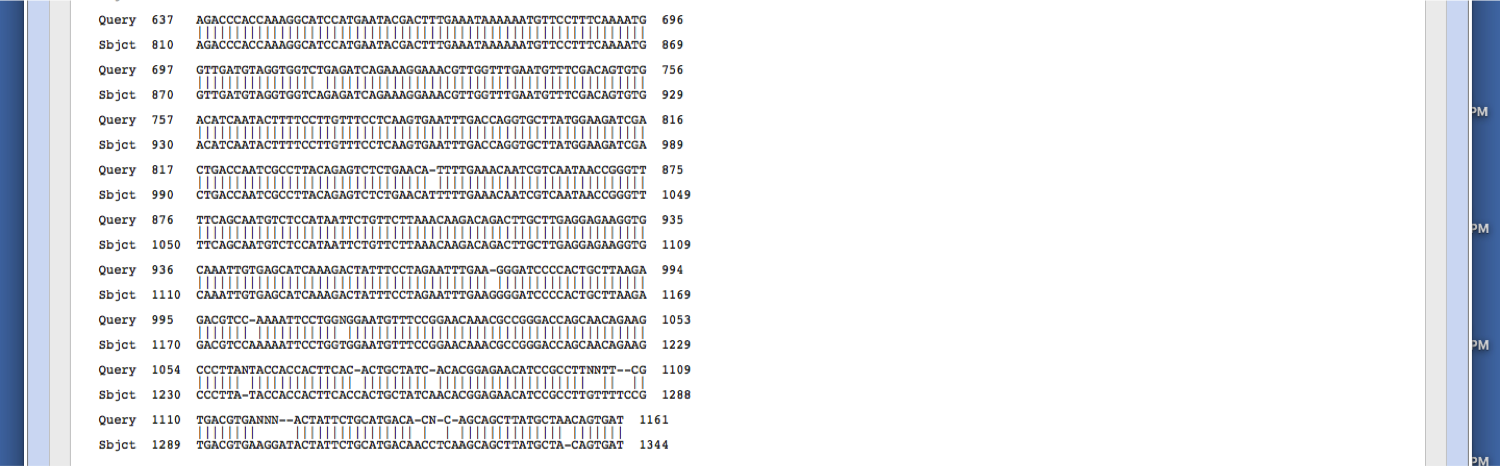
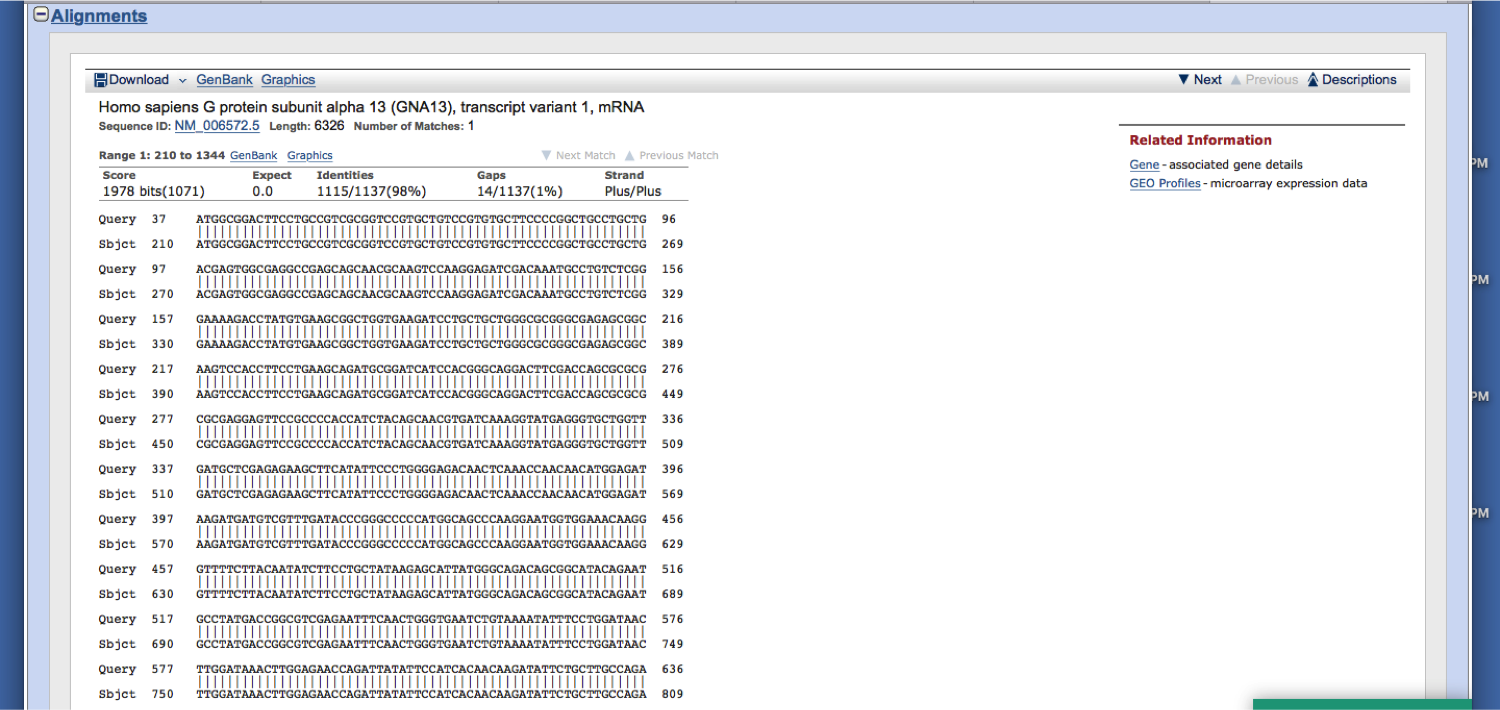
It might take awhile for the results to come up, depending on the type of analysis and the length of the query sequence. The results page will show a graphic summary of the DNA sequences in the database that aligned with the query sequence, with colors to show relative sequence similarity.



Below this is a description of the sequences that produced significant, very similar alignment.



In this case, the query sequence appears to have matched well with the human GPCR alpha 13 mRNA; success! This is what should have been amplified in the PCR reaction. There are a couple of transcript variants that also matched, which makes sense; the variants shouldn’t be too different in the region that was amplified. Below this are the genomic sequences that matched. And below this are the actual alignments, showing the occasional gaps where the sequence didn’t match up perfectly.



Notice that the alignment tends to break up a little towards the end of the sequenced DNA; this is completely normal and reflects the ambiguity in the signal as the sequencing reaction ends, but doesn’t necessarily mean that the actual sequence is misaligning at those nucleotides, it’s just an artifact of the sequencing reaction.