Paul Yenerall, 2011

**How to access gene records in NCBI**

The purpose of this tutorial is to help teach students how to access different types of records in NCBI and explain certain important items displayed by these records. There are multiple types of records NCBI contains, such as nucleotide sequences and protein sequences. This tutorial will focus on accessing nucleotide and protein sequences, but will briefly explain other features of NCBI.

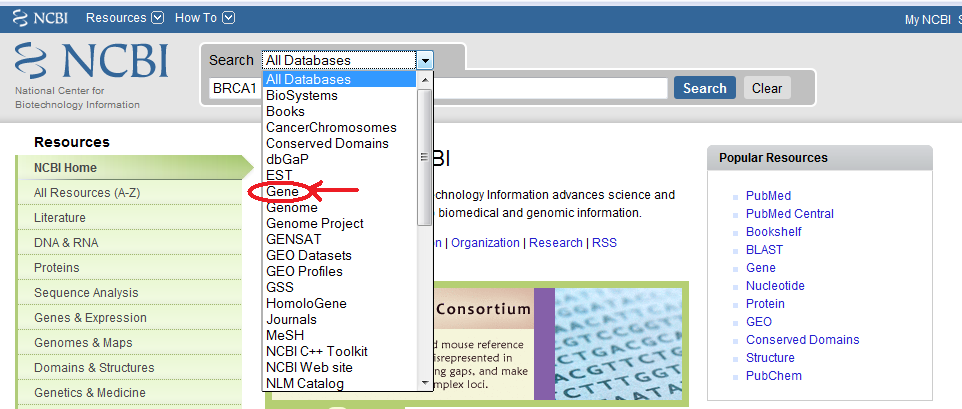
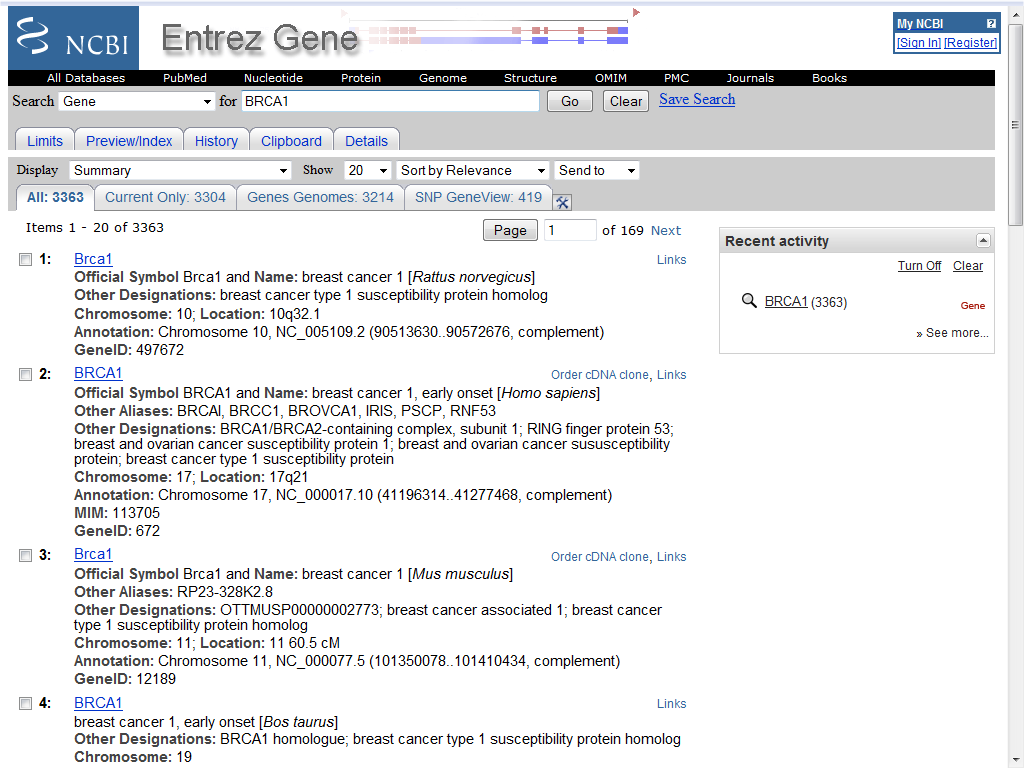
For questions, corrections, or help with this tutorial, please e-mail paulyenerall@gmail.com or lzhou1@pitt.edu

**Accessing a nucleotide and/or protein sequence**

**1)** Open up an internet browser, preferably Mozilla Firefox, and go to the URL: <http://www.ncbi.nlm.nih.gov/>. A page such as below should be displayed:

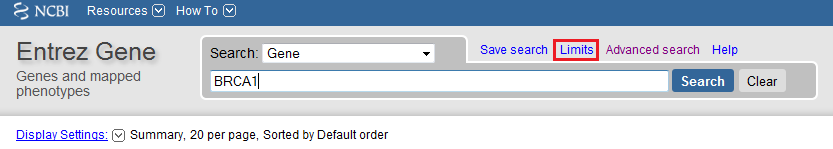


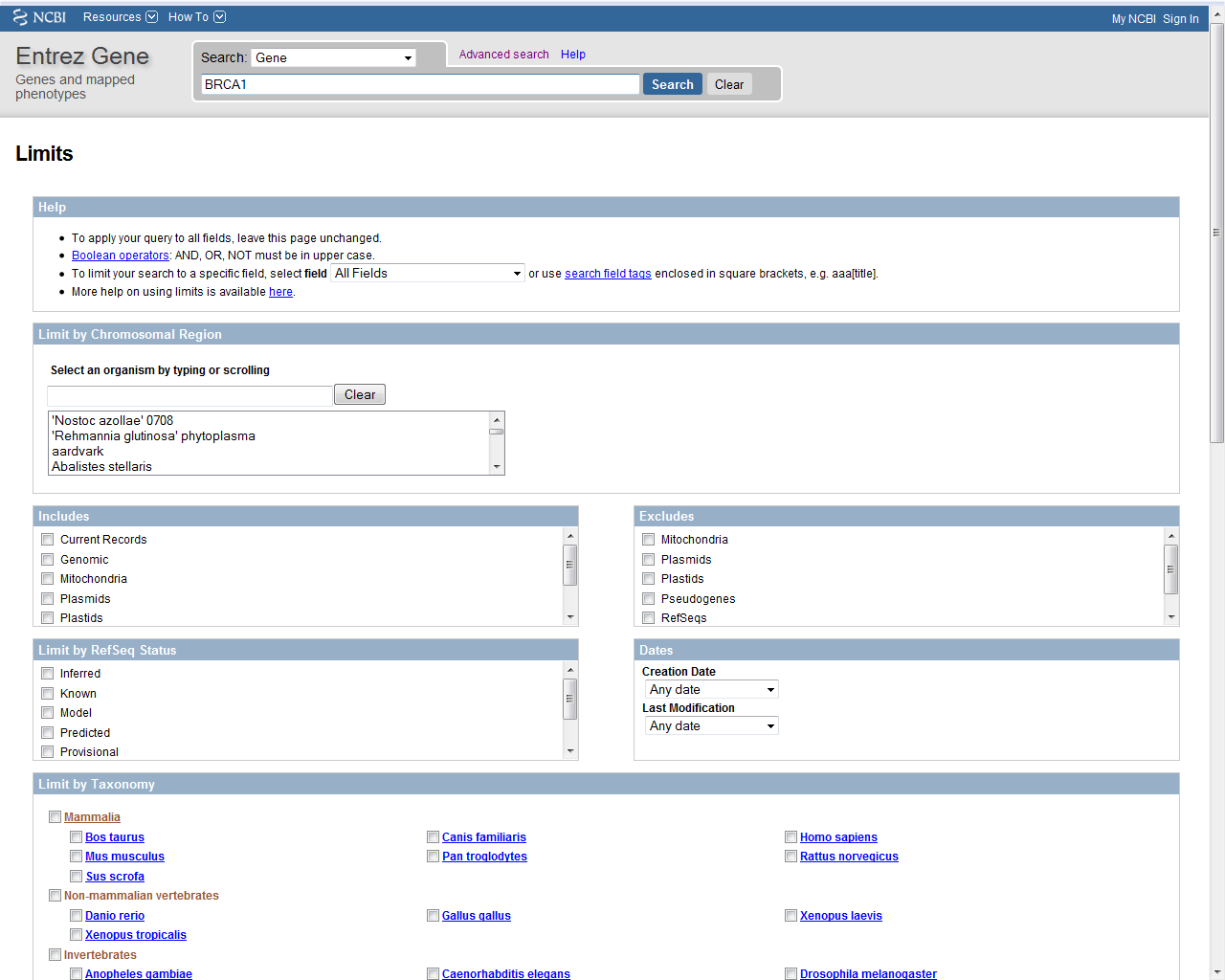
**2)** For illustrative purposes, we will access the nucleotide sequence and protein sequence for the human (*Homo sapiens*) gene *BRCA1*. First, type BRCA1 in the search bar located at top. Next, click the drop down box above the search box you just typed BRCA1 into, as illustrated below. Because *BRCA1* is a gene, select “Gene” from the dropdown box then click search, as illustrated below:

  
The page below should be shown:  


Note: If you do not specify what type of search you want to do (i.e. Gene, Genome, Protein etc.,) from the dropdown box, you will be taken to a general page that asks you what type of search you want to perform.

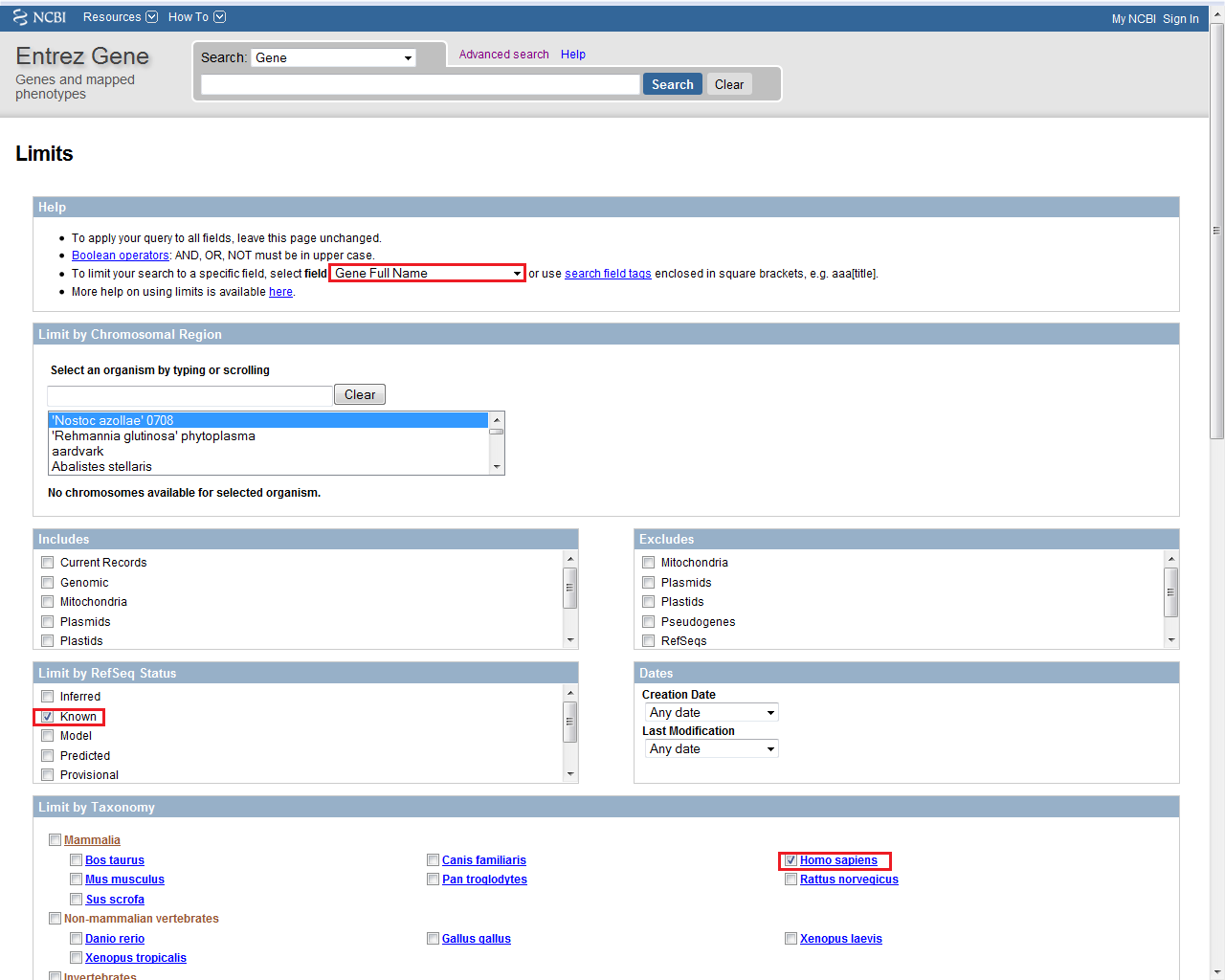
Before we continue, let us narrow our search. While this isn’t always necessary, sometimes you may need to limit the number of returns from a search if you are doing a very broad search, perform specific searches, or a search may involve many homologs. To do so, click on the “Limits” link located above the search box, as illustrated below:



The following page should be displayed after clicking limits:  


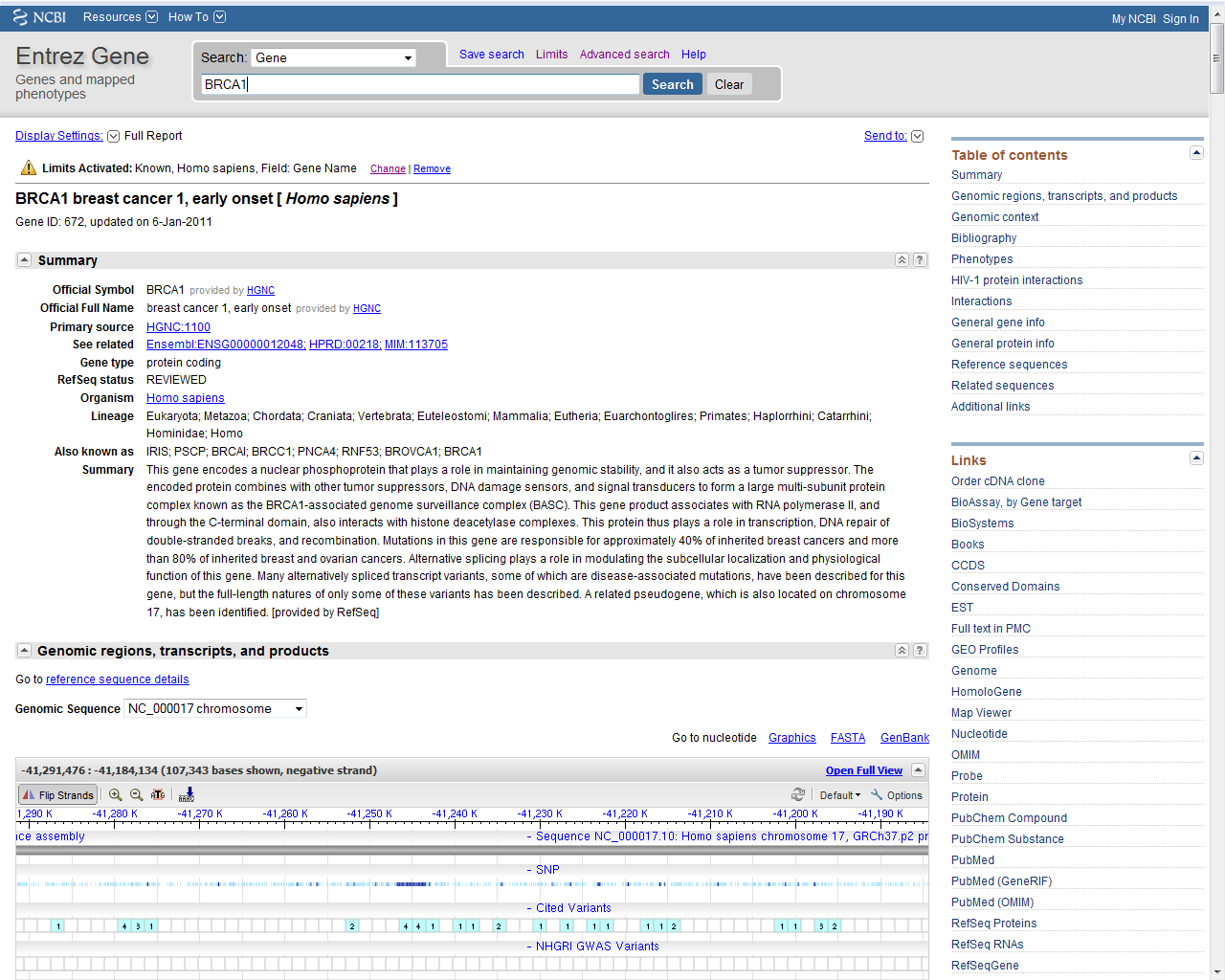
Take a moment to review this page and look over all the possible search limits. While we’ll only use some of the features on this page and you may never see this page again, it is useful to know of these limits. For instance, we can change what the query is searching for (a much more specific search than can be generated from NCBI’s homepage) under the first drop down box under the “Help” section. Also, we can choose to extract only DNA sequences from organelles or DNA sequences from plasmids, or choose to exclude them. Also of use, if we want to locate all the genes on one chromosome or all the genes at a particular locus, we can type the organisms name in the “Limit by Chromosomal Region” box, select our range of nucleotides to search, and clear our search box to locate all genes in the selected region. If we keep anything in the search box, NCBI will search the selected locus for that query.

While these and other features in the limits section are useful, the most common records you’re going to access are records for genes you know the name of in a particular organism. We know we’re looking for *BRCA1* in humans, so let’s limit our search to take us directly to that gene record:



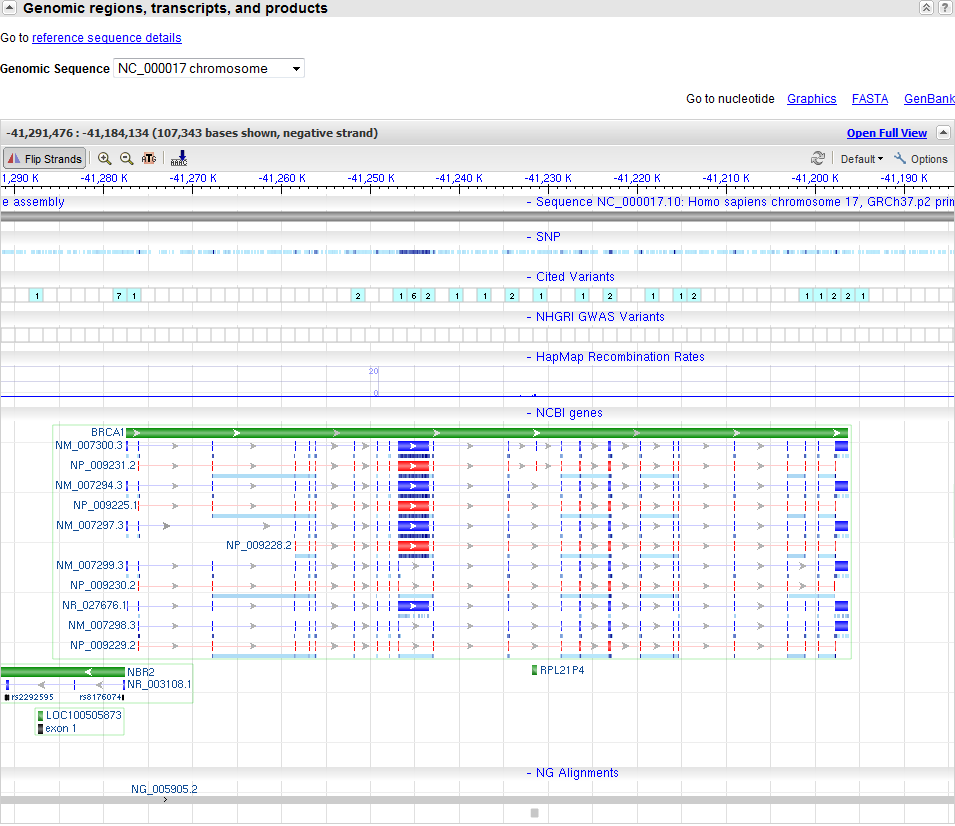
Take a look at the boxes highlighted above. We first change to search only for Gene Name (note: do not put Full Gene Name, the full gene name is the unabbreviated gene name - i.e. for *BRCA1* it is “breast cancer 1, early onset”). This will eliminate a lot of hits for proteins that associate with the entered query. Next, we select “Known” sequences under “Limit by RefSeq Status” because we only want sequences that have been verified to be correct. Records within NCBI may be computationally annotated but may not necessarily be correct. These are “predicted” or may be “inferred” sequences. With a well studied gene like *BRCA1* all hits will most likely be known, however, if you are performing searches on lesser studied genes, this may be useful. Lastly, we select *Homo sapiens* under “Limit by Taxonomy” because we are only looking for the human *BRCA1*.

Once you have selected all these limits, hit the “Go” button to begin your search. The following page should be displayed (note: normally you will have a list of options if you do not put specific limits on your search, but because our limits were specific we went directly to the record for *BRCA1* in *Homo sapiens*):

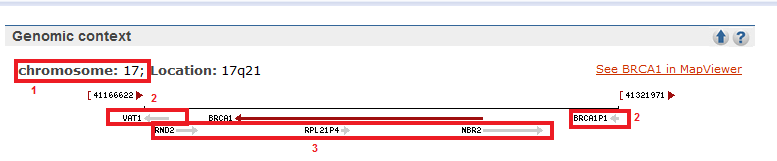


The top of this page shows us some basic information about *BRCA1* (the summary). Take a moment to review this page. The most important information displayed on the top of this page is the “Official symbol” (what the genes symbol is), “Official Full Name” (full name of the gene) “Lineage” (what family this species comes from) as well as a summary of *BRCA1*’s functions (“Summary”) and other names for *BRCA1* (“Also known as”)

**3)** Before we view the nucleotide and protein sequences for *BRCA1*, we will spend a moment reviewing other parts of this page (the NCBI Gene Record) as it contains crucial information about *BRCA1*. Aside from the obvious things (such as the Full Name, Gene Type, Lineage etc.,) listed in the Summary, other information will be reviewed on this page. Scroll down to the portion of the page called “Genomic regions, transcripts, and products” as shown below:

  
This is NCBI’s new Gene Viewer. This will show you the gene’s structure, including the 5’ UTR, exon/intron structure, and the 3’ UTR, as well as all the known transcripts of this gene. You can navigate the gene’s chromosomal locus by clicking down on the mouse and moving the cursor; you can also zoom into the individual nucleotide level. Above the gene’s structure display, Gene Viewer will also show you any relevant information on this gene, such as NHGRI GWAS variants (National Human Genome Research Institute Genome Wide Association Study Variants – SNPs that are included as likely candidates to cause conditions), SNPs (shown as a histogram and under “Cited Variants” which will take you to PubMed records for SNPS) and HapMap Recombination rates (recombination rates generated from the HapMap project). Other genes, pseudogenes, or areas of interest will also be displayed below the gene structure.

Now, scroll down to the next section entitled “Genomic Context” as shown below. This area tells us where our gene BRCA1 is located, as well as genes in close proximity to BRCA1 (much like the new Gene View does; the old Gene Viewer did not show this).

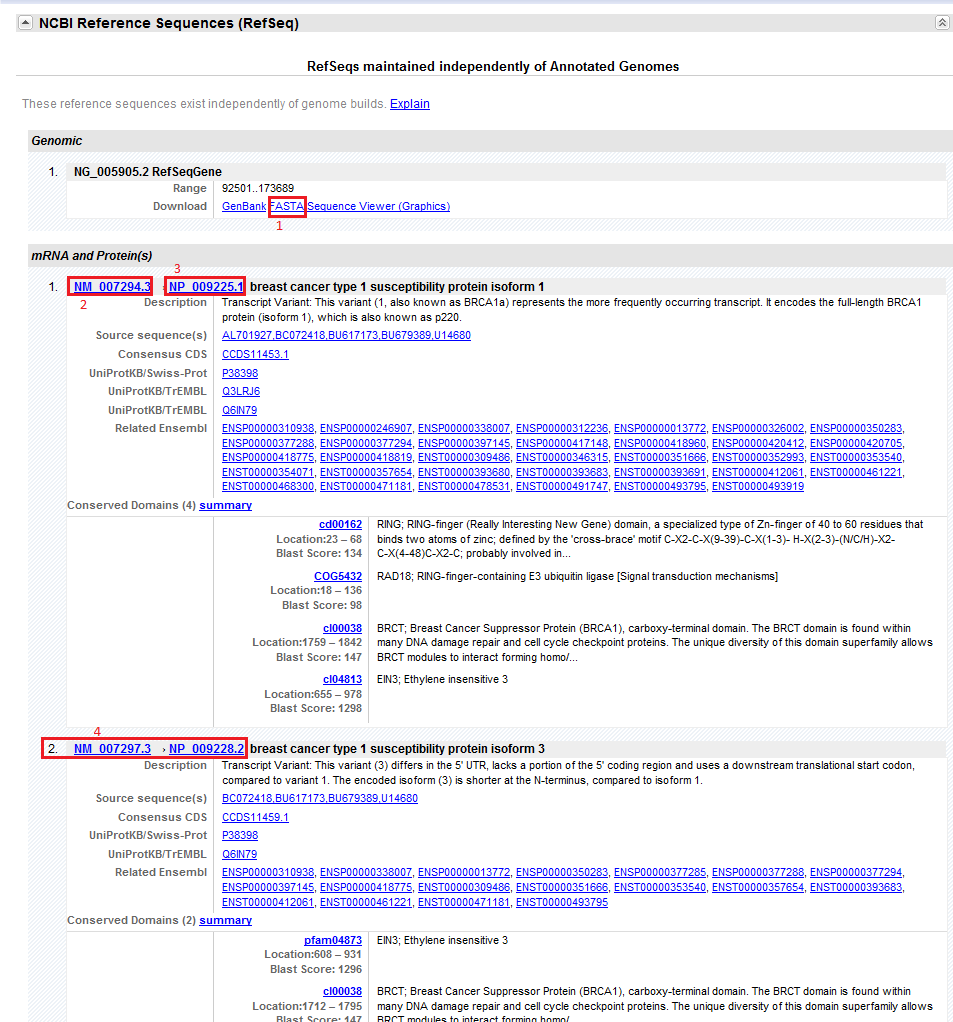


Each section highlighted is briefly explained below:

1 – This part tells us which chromosome our gene is located on (remember, eukaryotes generally have multiple linear chromosomes, whereas prokaryotes generally have one looped chromosomes)  
2 – These two parts indicated show genes close to the nucleotide position of our gene on the same strand as our gene (the minus or reverse strand). You can tell these genes are located on the reverse (minus) strand because the arrow is pointing to the left.  
3 – This part tells us genes close to the nucleotide position of our gene but on the positive (forward strand). You can tell these genes are located on the positive strand because the arrow is pointing to the right.

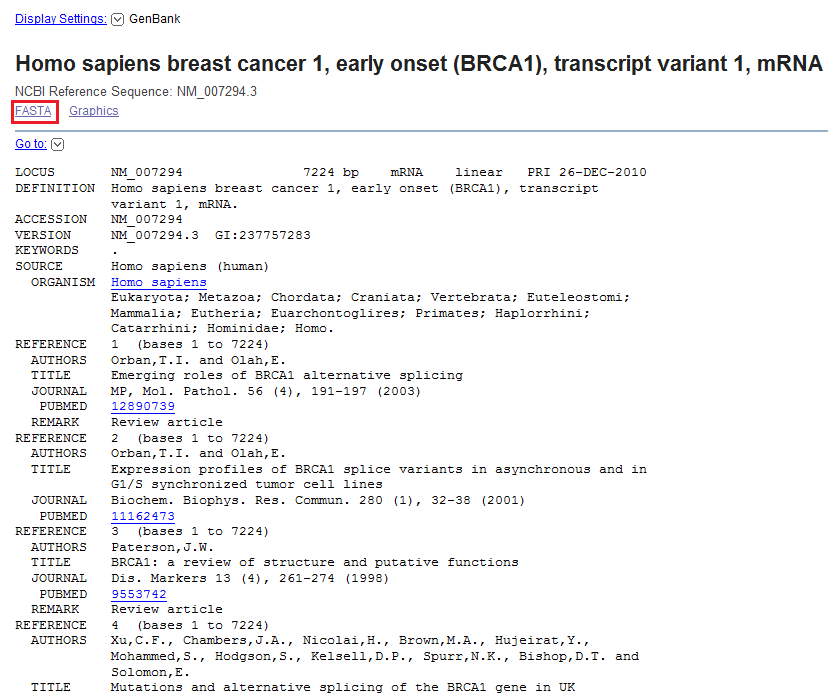
There is a lot of other relevant information listed on this page that can be located by scrolling down the page, but most is self explanatory and will not be highlighted in this tutorial. Note, however, that on this page you can find related PubMed articles to this gene, GeneRIFs (the noted functions of the gene), genes that interact with this gene (“Interactions”), phenotypes (observable characteristic trait of an organism), genotypes, pseudogenes (regions that appear homologous to a gene but are not transcribed), orthologs (the same gene in different species with a common ancestor gene), as well as a description of all the isoforms (“NCBI reference sequences – RefSeq”).

**4)** Now, let’s obtain the mRNA nucleotide and protein sequence for *BRCA1*. Scroll down to the section labeled “NCBI Reference Sequences (RefSeqs)”. Here we can obtain the gene’s nucleotide sequence, the mRNA sequence and the protein sequence for all verified isoforms of *BRCA1* as illustrated below (each step is numbered):



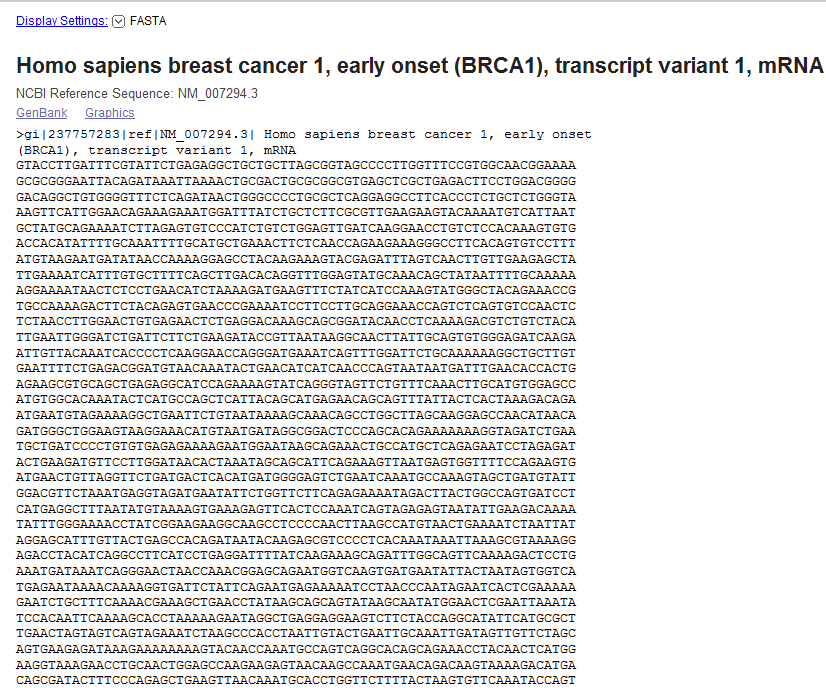
If we want to access the gene’s full nucleotide sequence (including UTRs, exons and introns) click the link labeled “1” in the diagram under “Genomic”. This will take you to the GenBank page with the whole gene’s sequence in FASTA format. If we want to access the mRNA sequence or protein sequence of a specific isoform, use the links under “mRNA and Protein(s)”. Click the link labeled “2” in the diagram (NM\_007294.3) to access Isoform 1’s mRNA entry in GenBank. Click the link labeled “3” in the diagram (NP\_009225.1) to access Isoform 2’s protein entry in GenBank. The links labeled “4” will take you to Isoform 3’s mRNA and protein entry in GenBank, respectively.

Let’s take a look at the nucleotide sequence for isoform 1, labeled “2” in the diagram (NM\_007294.3) by clicking NM\_007294.3. This will access the GenBank record for the mRNA of Isoform 1 of *BRCA1*, as shown below:



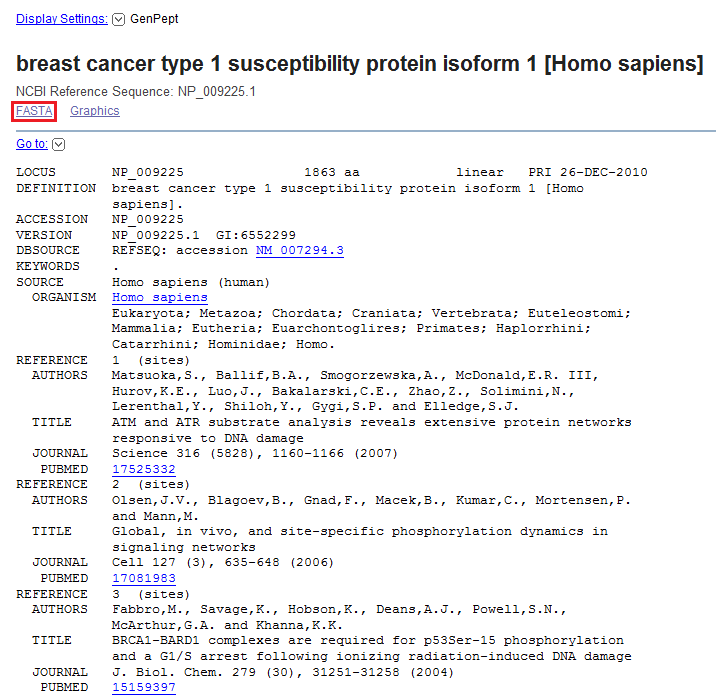
There is a lot of good information on this page, however, to learn more about GenBank records, take a look at the GenBank tutorial. The easiest way to access the nucleotide sequence of this mRNA in FASTA format is to click on the link labeled “FASTA” at the top left of the GenBank record (illustrated in red above).

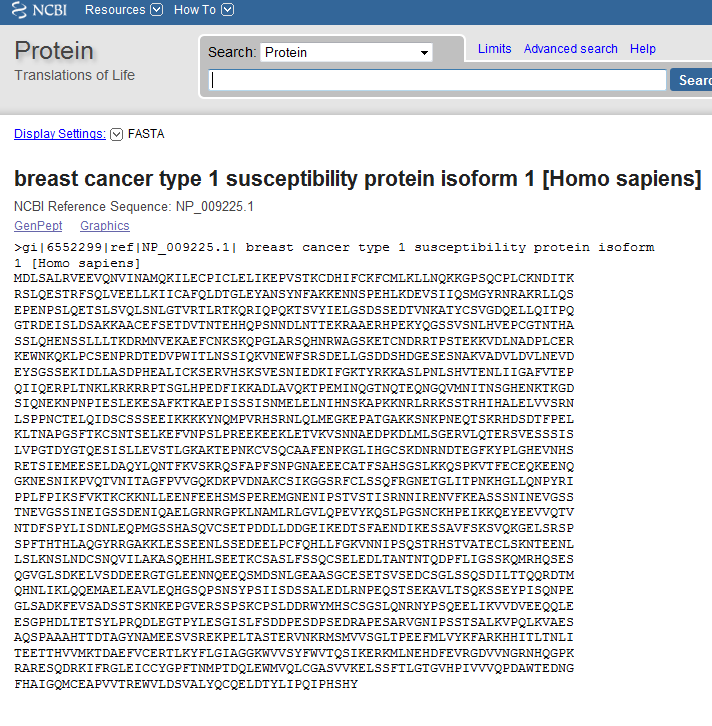
After clicking the link labeled “FASTA”, the following page should be displayed:



Note that in FASTA format, the header is indicated by “>” and ends after the first break. After the break, the rest of the letters are the actual sequence (GTACCTT….)

Now let’s access the protein sequence. Hit the back button once. If you scroll down, you’ll notice that the mRNA sequence for Isoform 1 is actually already translated for you on this page (under Features, then under CDS, and then after “/translation=”). However, there is no header for this sequence, so let’s grab the sequence with the header (and therefore in FASTA format, the format most programs analyzing sequences will prefer). Hit the back button again to return to the GeneRecord for *BRCA1*, and then hit the link labeled “3” in the previous diagram (NP\_009225.1). This should take you to a page that looks like the following (next page):



Again, there is a lot of good information on this page; however, we will not be discussing it in this tutorial. To access the FASTA protein sequence for Isoform 1, hit “FASTA” at the top left of the GenBank record (as illustrated above in red), the following page should be displayed (next page):

We are now done accessing the mRNA sequence and protein sequence for *BRCA1.* If you wish to save any pages on NCBI’s gene record or GenBank, you must make an account. From there, you can save pages and have personal preferences. The link to make this account is always in the top right of the page, as illustrated above:

