Introduction to Bioinformatics
Dr. Robert Moss

Bioinformatics is about searching biological databases, comparing sequences, looking at protein structures, and more generally, asking biological questions with a computer. Bioinformatics is now at the center of the most recent developments in biology, such as the deciphering of the human genome, the biotechnologies, new legal and forensic techniques, as well as the medicine of the future. You don't need to install complicated programs on your computer to become familiar with the techniques; many tools for bioinformatics can be run over the Internet via your Internet browser. This lab will introduce to you to the wonderful world of bioinformatics.

1a. MANUAL GENE FINDING: We'll do this together.
1b. MANUAL GENE FINDING: Do on your own. Part 1 of your lab report will consist of answers to the questions in section 1b.

We'll view together: BASICS OF BLAST .PPT

2. Your mitochondrial DNA analysis: Summarize your findings for part 4 of your lab report.
3. Bioinformatics "MUTANT-X": ANALYZE A DISEASE-CAUSING GENE: [Instructions in BIOINFORMATICS_MUTANT file; sequences in BIOINFORMATICS_SEQUENCES file].

You will receive the sequence for a gene or protein that seems to be involved in some human disease. You need to compare this sequence to all known human sequences to identify the gene, and then locate the mutation that seems to be responsible for the disease.
Part 5 of your lab report should be answers to the questions on these sheets, relating to your assigned gene.

4. Flu: We'll do this together. Summarize your findings for part 4 of your lab report.
5. HIV exercise. Do on your own. Part 5 of your lab report will consist of answers to the questions in section 5.

---

## 1a: Manual gene finding:
## Find a Gene Using Protein Evidence

**WHAT DOES A EUKARYOTIC GENE LOOK LIKE?**

Attached is a page with the sequence for a protein (142 amino acids) and a set of 3 pages with DNA sequence (1,200 nucleotides). The DNA sequence contains the gene for the protein on the first page. Feel free to separate the pages.

Underneath the DNA sequence is a translation of this sequence in all three reading frames, RF1 through RF3. The symbol * denotes stop codons in the DNA (check it out, stop codons are either TAA, TAG, or TGA).

Your task is to identify the gene in the DNA sequence by finding within the translated amino acid sequence amino acid stretches that match the sequence of the protein on the first page.

Identify the protein coding region within the translated protein sequence. Highlight the translated amino acid sequences which match the amino acid sequence of the protein. Then highlight the PRECISE DNA portions that encode the highlighted amino acid sequence. You'll need the codon table, and need to identify each intron, to the exact base pair.

NOTE: nearly all introns start with GU, and end with AG.

Answer the questions below. As always, you are encouraged to work together, but you must write out your answers on your own. [You will NOT turn these in]

1. A. What are the sequence stretches that contain coding sequences called?
   B. How many are in this gene?

2. A. What are the sequence stretches in between the coding sequences called?
   B. How many are in this gene?

3. List the exact nucleotide at which each exon begins, and ends.

4. a. Do all exons begin with start codons? Why?
   b. Do all exons end with stop codons? Why?

5. a. Can CODING SEQUENCE “jump” reading frames within a gene? Why?

```
 01   MVLSPADKTNVKAAWGKVGAHAGEYGAEALERFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNA  070

071  VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR 142
```

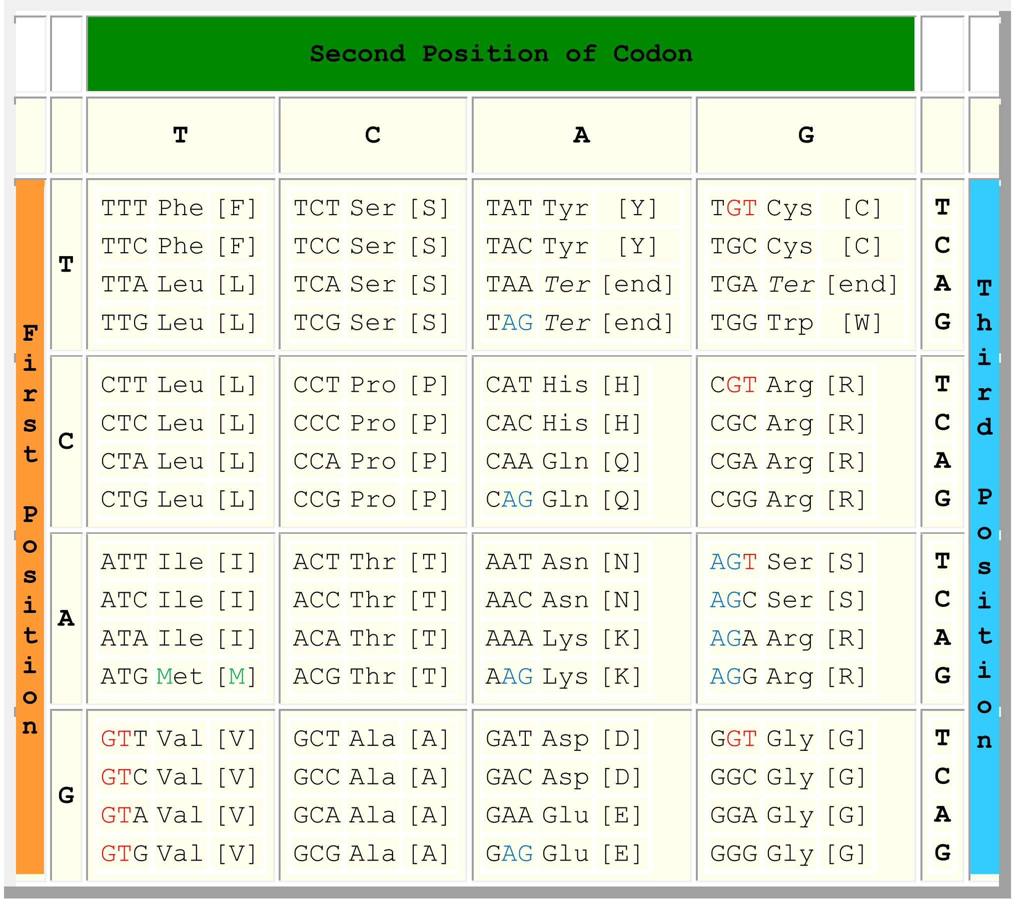
**Genic 2, forward sequence and translation in all three reading frames; M denotes potential start codons; * stop codons**

```
    M  P  P  G  E  R  D  G  R  E  W  S  G  G  W  R  V  E  T  S     F1
     C  P  R  A  S  G  M  G  G  S  G  V  A  G  G  G  W  R  R  P    F2
      A  P  G  R  A  G  W  A  G  V  E  W  R  V  E  G  G  D  V  L   F3
  1 ATGCCCCCGGGCGAGCGGGATGGGCGGGAGTGGAGTGGCGGGTGGAGGGTGGAGACGTCC 60
    ----:----|----:----|----:----|----:----|----:----|----:----|

    W  P  P  P  R  V  H  P  Q  G  R  P  S  P  P  P  G  P  A  Q     F1
     G  P  R  P  A  C  T  P  R  G  G  R  A  R  R  P  A  P  R  R    F2
      A  P  A  P  R  A  P  P  G  E  A  E  P  A  A  R  P  R  A  G   F3
 61 TGGCCCCCGCCCCGCGTGCACCCCCAGGGGAGGCCGAGCCCGCCGCCCGGCCCCGCGCAG 120
    ----:----|----:----|----:----|----:----|----:----|----:----|

    A  P  P  G  T  P  L  R  S  R  P  R  P  G  L  R  A  S  Q  *     F1
```

```
     P  R  P  G  L  P  C  G  P  G  R  A  P  G  S  A  P  A  N  E     F2
      P  A  R  D  S  P  A  V  Q  A  A  P  R  A  P  R  Q  P  M  S    F3
121 GCCCCGCCCGGGACTCCCCTGCGGTCCAGGCCGCGCCCCGGGCTCCGCGCCAGCCAATGA 180
    A  P  P  G  R  A  C  P  R  A  P  S  I  N  P  G  A  L  A  A      F1
     R  R  P  A  G  R  A  P  A  P  Q  A  *  T  L  A  R  S  R  P     F2
      A  A  R  P  G  V  P  P  R  P  K  H  K  P  W  R  A  R  G  P    F3
181 GCGCCGCCCGGCCGGGCGTGCCCCCGCGCCCCAAGCATAAACCCTGGCGCGCTCGCGGCC 240
    ----:----|----:----|----:----|----:----|----:----|----:----|

    R  H  S  S  G  P  H  R  L  R  E  N  P  P  W  C  C  L  L  P      F1
     G  T  L  L  V  P  T  D  S  E  R  T  H  H  G  A  V  S  C  R     F2
      A  L  F  W  S  P  Q  T  Q  R  E  P  T  M  V  L  S  P  A  D    F3
241 CGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCG 300
    ----:----|----:----|----:----|----:----|----:----|----:----|

    T  R  P  T  S  R  P  P  G  V  R  S  A  R  T  L  A  S  M  V      F1
     Q  D  Q  R  Q  G  R  L  G  *  G  R  R  A  R  W  R  V  W  C     F2
      K  T  N  V  K  A  A  W  G  K  V  G  A  H  A  G  E  Y  G  A    F3
301 ACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTG 360
    ----:----|----:----|----:----|----:----|----:----|----:----|

    R  R  P  W  R  G  E  A  P  S  P  A  P  T  R  A  P  R  P  P      F1
     G  G  P  G  E  V  R  L  P  P  L  L  R  P  G  L  L  A  R  P     F2
      E  A  L  E  R  *  G  S  L  P  C  S  D  P  G  S  S  P  A  R    F3
361 CGGAGGCCCTGGAGAGGTGAGGCTCCCTCCCCTGCTCCGACCCGGGCTCCTCGCCCGCCC 420
    ----:----|----:----|----:----|----:----|----:----|----:----|

    G  P  T  G  H  P  Q  P  S  W  P  R  T  Q  T  P  P  L  T  L      F1
     D  P  Q  A  T  L  N  R  P  G  P  G  P  K  P  H  P  S  L  C     F2
      T  H  R  P  P  S  T  V  L  A  P  D  P  N  P  T  P  H  S  A    F3
421 GGACCCACAGGCCACCCTCAACCGTCCTGGCCCCGGACCCAAACCCCACCCCTCACTCTG 480
    ----:----|----:----|----:----|----:----|----:----|----:----|

    L  L  P  A  G  G  S  C  P  S  P  P  P  R  P  T  S  R  T  S      F1
     F  S  P  Q  E  V  P  V  L  P  H  H  Q  D  L  L  P  A  L  R     F2
      S  P  R  R  R  F  L  S  F  P  T  T  K  T  Y  F  P  H  F  D    F3
```

```
481 CTTCTCCCCGCAGGAGGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCG 540
    ----:----|----:----|----:----|----:----|----:----|----:----|

    T  *  A  T  A  L  P  R  L  R  A  T  A  R  R  W  P  T  R  *     F1
     P  E  P  R  L  C  P  G  *  G  P  R  Q  E  G  G  R  R  A  D    F2
      L  S  H  G  S  A  Q  V  K  G  H  G  K  K  V  A  D  A  L  T   F3
541 ACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGA 600
    ----:----|----:----|----:----|----:----|----:----|----:----|

    P  T  P  W  R  T  W  T  T  C  P  T  R  C  P  P  *  A  T  C     F1
     Q  R  R  G  A  R  G  R  H  A  Q  R  A  V  R  P  E  R  P  A    F2
      N  A  V  A  H  V  D  D  M  P  N  A  L  S  A  L  S  D  L  H   F3
601 CCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGC 660
    ----:----|----:----|----:----|----:----|----:----|----:----|

    T  R  T  S  F  G  W  T  R  S  T  S  R  *  A  A  G  R  E  R     F1
     R  A  Q  A  S  G  G  P  G  Q  L  Q  G  E  R  R  A  G  S  D    F2
      A  H  K  L  R  V  D  P  V  N  F  K  V  S  G  G  P  G  A  I   F3
661 ACGCGCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGGTGAGCGGCGGGCCGGGAGCGA 720
    ----:----|----:----|----:----|----:----|----:----|----:----|

    S  G  S  R  G  E  M  A  P  S  S  Q  G  R  G  S  R  G  L  R     F1
     L  G  R  G  A  R  W  R  L  P  R  R  A  E  D  H  A  G  C  G    F2
      W  V  E  G  R  D  G  A  F  L  A  G  Q  R  I  T  R  V  A  G   F3
721 TCTGGGTCGAGGGGCGAGATGGCGCCTTCCTCGCAGGGCAGAGGATCACGCGGGTTGCGG 780
    ----:----|----:----|----:----|----:----|----:----|----:----|

    E  V  *  R  R  R  R  L  R  A  W  A  L  G  P  T  D  P  L  L     F1
     R  C  S  A  G  G  G  C  G  P  G  P  S  A  P  L  T  L  F  S    F2
      G  V  A  Q  A  A  A  A  G  L  G  P  R  P  H  *  p  s  s  l   F3
781 GAGGTGTAGCGCAGGCGGCGGCTGCGGGCCTGGGCCCTCGGCCCCACTGACCCTCTTCTC 840
    ----:----|----:----|----:----|----:----|----:----|----:----|

    C  T  A  P  K  P  L  P  A  G  D  P  G  R  P  P  P  R  R  V     F1
     A  Q  L  L  S  H  C  L  L  V  T  L  A  A  H  L  P  A  E  F    F2
      h  s  s  *  a  t  a  c  w  *  P  W  P  P  T  S  P  P  S  S   F3
```

```
 841 TGCACAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTT 900
     ----:----|----:----|----:----|----:----|----:----|----:----|

     H  P  C  G  A  R  L  P  G  Q  V  P  G  F  C  E  H  R  A  D      F1
      T  P  A  V  H  A  S  L  D  K  F  L  A  S  V  S  T  V  L  T     F2
       P  L  R  C  T  P  P  W  T  S  S  W  L  L  *  a  p  c  *  P    F3
 901 CACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGAC 960
     ----:----|----:----|----:----|----:----|----:----|----:----|

     L  Q  I  P  L  S  W  S  L  G  G  H  A  S  C  P  L  G  L  P      F1
      S  K  Y  R  *  A  G  A  S  V  A  M  L  L  A  P  W  A  S  P     F2
       P  N  T  V  K  L  E  P  R  W  P  C  F  L  P  L  G  P  P  P    F3
 961 CTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCC 1020
     ----:----|----:----|----:----|----:----|----:----|----:----|

     P  A  P  P  P  L  P  A  P  V  P  P  W  S  L  N  K  V  *  V      F1
      Q  P  L  L  P  F  L  H  P  Y  P  R  G  L  *  I  K  S  E  W     F2
       S  P  S  S  P  S  C  T  R  T  P  V  V  F  E  *  S  L  S  G    F3
1021 CCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTG 1080
     ----:----|----:----|----:----|----:----|----:----|----:----|

     G  G  S  L  C  V  P  E  F  F  P  S  A  N  V  P  G  M  G  V      F1
      A  A  A  C  V  C  L  S  F  F  P  Q  Q  T  C  Q  A  W  A  W     F2
       R  Q  P  V  C  A  *  V  F  S  L  S  K  R  A  R  H  G  R  G    F3
1081 GGCGGCAGCCTGTGTGTGCCTGAGTTTTTTCCCTCAGCAAACGTGCCAGGCATGGGCGTG 1140
     ----:----|----:----|----:----|----:----|----:----|----:----|

     D  S  S  W  D  T  H  G  *  n  l  s  a  a  g  *  G  R  K  R      F1
      T  A  A  G  T  H  M  A  R  T  S  L  Q  L  D  R  V  G  K  G     F2
       Q  Q  L  G  H  T  W  L  E  P  L  C  S  W  I  G  *  E  K  A    F3
1141 GACAGCAGCTGGGACACACATGGCTAGAACCTCTCTGCAGCTGGATAGGGTAGGAAAAGG 1200
     ----:----|----:----|----:----|----:----|----:----|----:----|
```



| | | Second Position of Codon | | | | | |
|---|---|---|---|---|---|---|---|
| | | T | C | A | G | | |
| First Position | T | TTT Phe [F]<br>TTC Phe [F]<br>TTA Leu [L]<br>TTG Leu [L] | TCT Ser [S]<br>TCC Ser [S]<br>TCA Ser [S]<br>TCG Ser [S] | TAT Tyr [Y]<br>TAC Tyr [Y]<br>TAA *Ter* [end]<br>TAG *Ter* [end] | TGT Cys [C]<br>TGC Cys [C]<br>TGA *Ter* [end]<br>TGG Trp [W] | T<br>C<br>A<br>G | Third Position |
| | C | CTT Leu [L]<br>CTC Leu [L]<br>CTA Leu [L]<br>CTG Leu [L] | CCT Pro [P]<br>CCC Pro [P]<br>CCA Pro [P]<br>CCG Pro [P] | CAT His [H]<br>CAC His [H]<br>CAA Gln [Q]<br>CAG Gln [Q] | CGT Arg [R]<br>CGC Arg [R]<br>CGA Arg [R]<br>CGG Arg [R] | T<br>C<br>A<br>G | |
| | A | ATT Ile [I]<br>ATC Ile [I]<br>ATA Ile [I]<br>ATG Met [M] | ACT Thr [T]<br>ACC Thr [T]<br>ACA Thr [T]<br>ACG Thr [T] | AAT Asn [N]<br>AAC Asn [N]<br>AAA Lys [K]<br>AAG Lys [K] | AGT Ser [S]<br>AGC Ser [S]<br>AGA Arg [R]<br>AGG Arg [R] | T<br>C<br>A<br>G | |
| | G | GTT Val [V]<br>GTC Val [V]<br>GTA Val [V]<br>GTG Val [V] | GCT Ala [A]<br>GCC Ala [A]<br>GCA Ala [A]<br>GCG Ala [A] | GAT Asp [D]<br>GAC Asp [D]<br>GAA Glu [E]<br>GAG Glu [E] | GGT Gly [G]<br>GGC Gly [G]<br>GGA Gly [G]<br>GGG Gly [G] | T<br>C<br>A<br>G | |

---

## 1B: GENE FINDING, USING PROTEIN EVIDENCE, WITH COMPUTER TOOLS:

You’ll turn in answers to the questions on this one, as part 1 of your lab report.

Below you’ll find the sequence of a protein (142 amino acids) and a DNA sequence (1,700 nucleotides). The DNA sequence contains the gene for the protein.

Use the tool called “Six Pack” to get a predicted translation for the DNA sequence, in all three reading frames. http://gander.wustl.edu/cgi-bin/emboss/sixpack The only parameter you should change would be: Set “Display translation of reverse sense?” To “No”. Once you have got your translation in all three frames, print that part out, OR copy it to a file.

The symbol * denotes stop codons in the DNA (check it out, stop codons are either TAA, TAG, or TGA).

Your task is to identify the gene in the DNA sequence by finding within the translated amino acid sequence amino acid stretches that match the sequence of the protein on the first page. You can do this manually, OR use another tool: BLAST2SEQ. Do a Google search for BLAST2SEQ. Bring this tool up. There are many types of blast: Blastn will search a nucleotide sequence with another nucleotide sequence. Blastp will search a protein sequence with another protein sequence. Tblastn will TRANSLATE a nucleotide sequence, in all 6 possible reading frames, and then search that for a protein sequence. That’ s what we want here. So click on “ Tblastn” .

Paste the nucleotide sequence into the “ SUBJECT” box, and the protein sequence into the “ QUERY” box. Properly formatted DNA sequences always start with a comment line, that must begin with a “ >” , that for instance describes the name of the sequence. For instance:

➢ Unknown DNA sequence for translation.

Click “ Blast” . Use the alignments found to help you find the start and stop points to the exons. On your “ sixpack” display. Remember, you need to check all splice junctions, to highlight the *PRECISE* DNA portions that encode the highlighted amino acid sequence.

Answer the questions below. As always, you are encouraged to work together, but you must write out your answers on your own.

I-1 A. What are the sequence stretches that contain coding sequences called?

B. How many are in this gene?

I-2. A. What are the sequence stretches in between the coding sequences called?

B. How many are in this gene?

I-3: Make a list of the exact nucleotide locations of the start of each exon, and the end of each exon. Also, include the location of the stop codon.

I-4. a. Do all exons begin with start codons? Why?

b. Do all exons end with stop codons? Why?

I-5. a. Can CODING SEQUENCE “ jump” reading frames within a gene? Why?

I-6. What do you think the identity of this gene is? [You may have to wait until you learn to use BLASTp before answering this]

## PROTEIN SEQUENCE:

```
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

DNA SEQUENCE:

```
>DNA
Accattggtaaaaatatatataattttctatttttatttatactgactggccaaggctac  60
acatttgcttctgacacaactgtgttcactagcaacctcaaacagacaccatggtgcatc  120
tgactcctgaggagaagtctgccgttactgccctgtggggcaaggtgaacgtggatgaag  180
ttggtggtgaggccctgggcaggttggtatcaaggttacaagacaggtttaaggagacca  240
atagaaactgggcatgtggagacagagaagactcttgggtttctgataggcactgactct  300
ctctgcctattggtctattttcccacccttaggctgctggtggtctacccttggacccag  360
aggttctttgagtcctttggggatctgtccactcctgatgctgttatgggcaaccctaag  420
gtgaaggctcatggcaagaaagtgctcggtgcctttagtgatggcctggctcacctggac  480
aacctcaagggcacctttgccacactgagtgagctgcactgtgacaagctgcacgtggat  540
cctgagaacttcaggtgagtctatgggacgcttgatgttttctttccccttcttttcta  600
tggttaagttcatgtcataggaaggggataagtaacagggtacagtttagaatgggaaac  660
agacgaatgattgcatcagtgtggaagtctcaggatcgtttagtttcttttatttgctg  720
ttcataacaattgtttcttttgttaattcttgctttctttttttttttcttctccgcaat  780
ttttactattatacttaatgccttaacattgtgtataacaaaaggaaatatctctgagat  840
acattaagtaacttaaaaaaaaactttacacagtctgcctagtacattactatttggaat  900
atatgtgtgcttatttgcatattcataatctccctactttattttcttttatttttaatt  960
gatacataatcattatacatatttatgggttaaagtgtaatgttttaatatgtgtacaca  1020
tattgaccaaatcagggtaattttgcatttgtaattttaaaaaatgctttcttcttttaa  1080
tatacttttttgtttatcttatttctaatactttccctaatctctttctttcaggcaat  1140
aatgatacaatgtatcatgcctctttgcaccattctaaagaataacagtgataatttctg  1200
ggttaaggcaatagcaatatctctgcatataaatatttctgcatataaattgtaactgat  1260
gtaagaggtttcatattgctaatagcagctacaatccagctaccattctgcttttatttt  1320
atggttgggataaggctggattattctgagtccaagctaggcccttttgctaatcatgtt  1380
catacctcttatcttcctcccacagctcctgggcaacgtgctggtctgtgtgctggccca  1440
tcactttggcaaagaattcaccccaccagtgcaggctgcctatcagaaagtggtggctgg  1500
tgtggctaatgccctggcccacaagtatcactaagctcgctttcttgctgtccaatttct  1560
attaaaggttcctttgttccctaagtccaactactaaactgggggatattatgaaggcc  1620
ttgagcatctggattctgcctaataaaaaacatttattttcattgc
```

## PART 2: ANALYSIS OF YOUR MITOCHONDRIAL DNA SEQUENCE

1. Align the sequences of everyone in the class with CLUSTALW.

2. Align your sequence with the mitochondrial DNA standard sequence NC_012920, using BLAST2SEQ. Note the positions and sequences of all of your differences.

3. Compare your sequence with those of populations throughout the world: http://www.bioservers.org/bioserver/index1.html

4. Once you have compared your sequence to the "standard", determine your likely "haplogroup": The mtDNAmanager: http://mtmanager.yonsei.ac.kr/index.php [Read about it first at http://www.biomedcentral.com/1471-2105/9/483 ]

# PART 3: MUTATION ANALYSIS:

**GENE MUTATION EXERCISE: - a bioinformatics exercise for undergraduate biology science students**

Robert Moss, Wofford College, Spartanburg, South Carolina
Melissa Rowland-Goldsmith, Chapman University, Orange, CA
Leena Sawant, Houston Community College, Houston, Texas
Michael Fahy, Chapman University, Chapman University, Orange, CA

1. **Project abstract**

Bioinformatics is about searching biological databases, comparing sequences, looking at protein structures, and more generally, asking biological questions with a computer. Bioinformatics is now at the center of the most recent developments in biology, such as the deciphering of the human genome, the biotechnologies, new legal and forensic techniques, as well as the medicine of the future. You don't need to install complicated programs on your computer to become familiar with the techniques; many tools for bioinformatics can be run over the Internet via your Internet browser. This lab will introduce to you to the wonderful world of bioinformatics and will specifically focus on 3 widely used bioinformatics tools.

**Learning objectives:**

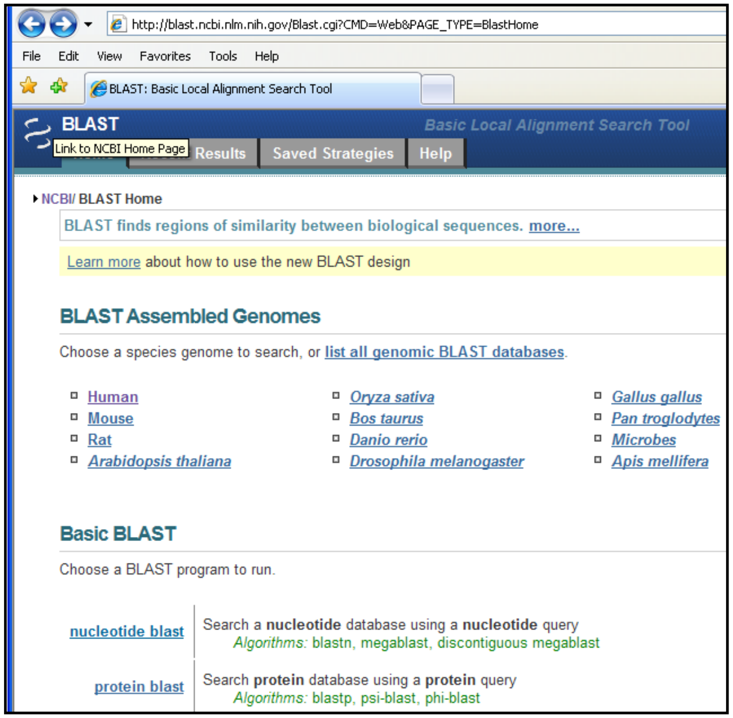
At the end of this interactive exercise, students should feel comfortable navigating in the NCBI website. They should know how to do BLAST searches and find relevant information from such a search. They should know how to navigate ENTREZ and use that site to learn many important features about their gene/ protein sequence. Lastly, they should competent using OMIM to find important information about how a mutated gene can lead to a disease.

1. You will be assigned a gene number. You will find a corresponding gene or protein sequence in a common file your computer can access. Open the file and then copy the corresponding sequence to the clipboard. These sequences are mutated gene sequences, found in patients with particular diseases. You'll first need to find out what the normal gene is, and the nature of the mutation in this patient.

**Demo these procedures with:**

```
CTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAAGCGCGGGAATTACAGATAAATTAAAACTGCGACTGCGCGGCGTGAGCTCGCTGAGACTTCCTGG
ACGGGGGACAGGCTGTGGGGTTTCTCAGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAAGTTCATTGGAACAGAAAGAAATGG
ATTTATCTGCTCTTCGCGTTGAAGAAGTACAAAAGGTCATTAATGCTATGCAGAAAATCTTAGAGTGTCCCATCTG
```

Then go to the NCBI databases http://www.ncbi.nlm.nih.gov/


http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome
File Edit View Favorites Tools Help
BLAST: Basic Local Alignment Search Tool
BLAST
Basic Local Alignment Search Tool
Link to NCBI Home Page
Results
Saved Strategies
Help

▸ NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Learn more about how to use the new BLAST design

**BLAST Assembled Genomes**

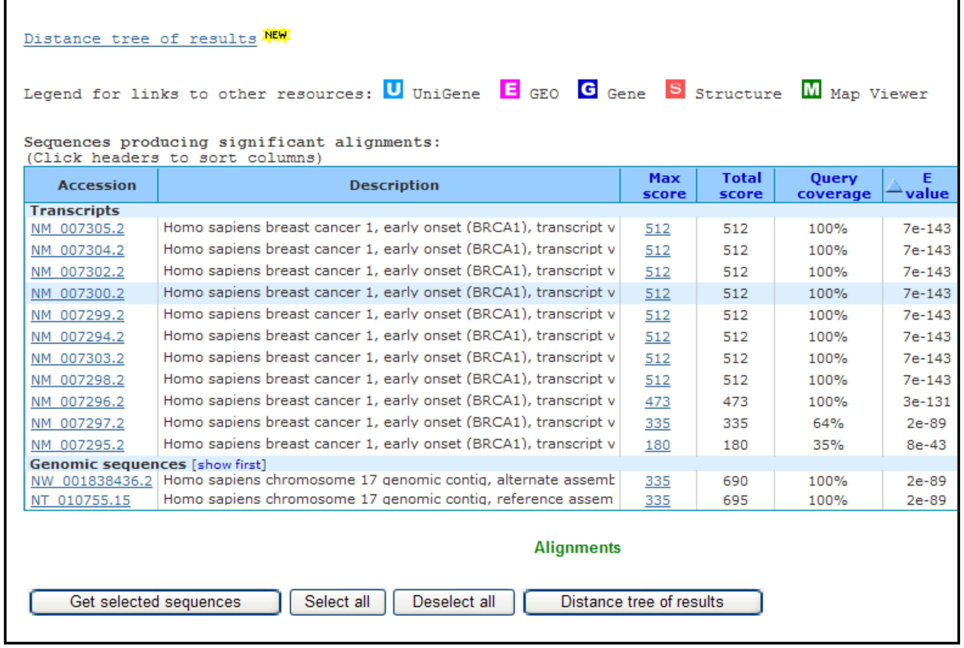
Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

**Basic BLAST**

Choose a BLAST program to run.

| | |
|---|---|
| nucleotide blast | Search a **nucleotide** database using a **nucleotide** query<br>*Algorithms:* blastn, megablast, discontiguous megablast |
| protein blast | Search **protein** database using a **protein** query<br>*Algorithms:* blastp, psi-blast, phi-blast |



Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer

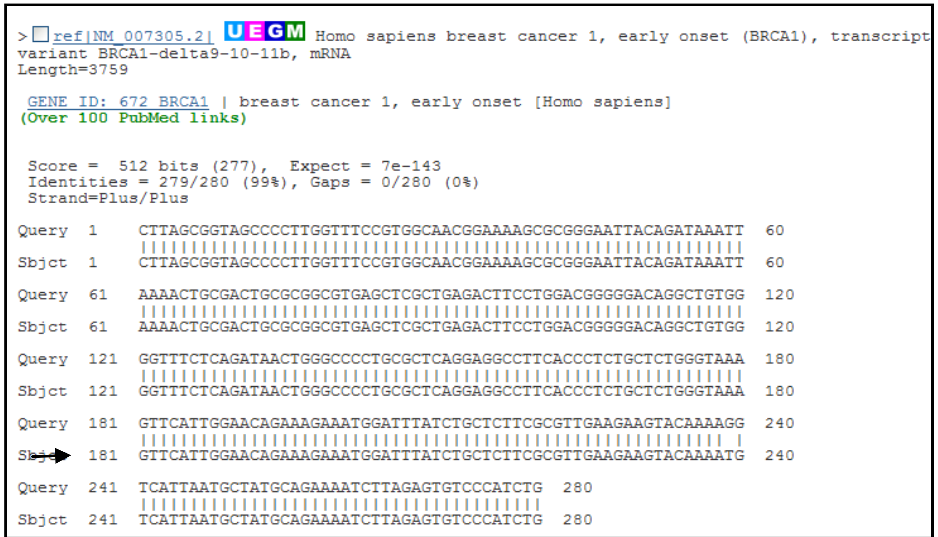
Sequences producing significant alignments:
(Click headers to sort columns)

| Accession | Description | Max score | Total score | Query coverage | E value |
|---|---|---|---|---|---|
| **Transcripts** | | | | | |
| NM_007305.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007304.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007302.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007300.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007299.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007294.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007303.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007298.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 512 | 512 | 100% | 7e-143 |
| NM_007296.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 473 | 473 | 100% | 3e-131 |
| NM_007297.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 335 | 335 | 64% | 2e-89 |
| NM_007295.2 | Homo sapiens breast cancer 1, early onset (BRCA1), transcript v | 180 | 180 | 35% | 8e-43 |
| **Genomic sequences** [show first] | | | | | |
| NW_001838436.2 | Homo sapiens chromosome 17 genomic contig, alternate assemt | 335 | 690 | 100% | 2e-89 |
| NT_010755.15 | Homo sapiens chromosome 17 genomic contig, reference assem | 335 | 695 | 100% | 2e-89 |

Click “BLAST” on the main menu bar; then “Nucleotide blast” or “protein blast”. Do a BLAST search for the gene/protein sequence you have been assigned. Then click on "BLAST". Make sure you have selected a “nucleotide BLAST” if you have a nucleic acid sequence; a “protein BLAST” if you have a sequence of amino acids. Also, for our exercise, select “homo sapiens” for the species.

To compare two specific sequences, click on “BLAST2”. [But that’s not what we’re going to do today!]

Paste your sequence into the search box, and click on "BLAST" at the bottom of the page. You may get a list of sequences. The transcripts at the top of the screen, with very low 'E scores', are most closely related to the search sequence. So start from the top, and look for a “description” that mentions a particular gene sequence. You don’t want a sequence with “putative”, “tentative”, or “predicted” in it; as these are not confirmed as “real” genes. Copy down the “Accession #” for the mRNA you think is most likely the highest one you’d be interested in; here the top one, NM_007305.2



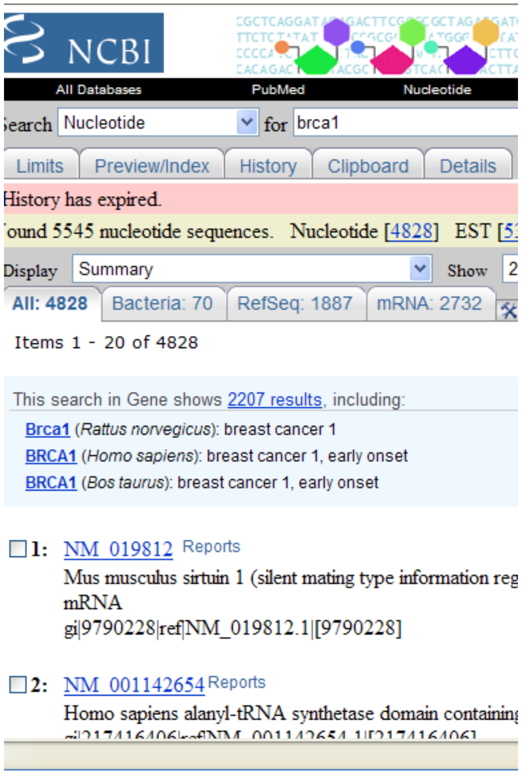
```
>□ref|NM_007305.2| UEGM Homo sapiens breast cancer 1, early onset (BRCA1), transcript
variant BRCA1-delta9-10-11b, mRNA
Length=3759

 GENE ID: 672 BRCA1 | breast cancer 1, early onset [Homo sapiens]
(Over 100 PubMed links)


 Score =  512 bits (277),  Expect = 7e-143
 Identities = 279/280 (99%), Gaps = 0/280 (0%)
 Strand=Plus/Plus

Query  1    CTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAAGCGCGGGAATTACAGATAAATT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    CTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAAGCGCGGGAATTACAGATAAATT  60

Query  61   AAAACTGCGACTGCGCGGCGTGAGCTCGCTGAGACTTCCTGGACGGGGGACAGGCTGTGG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   AAAACTGCGACTGCGCGGCGTGAGCTCGCTGAGACTTCCTGGACGGGGGACAGGCTGTGG  120

Query  121  GGTTTCTCAGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GGTTTCTCAGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAA  180

Query  181  GTTCATTGGAACAGAAAGAAATGGATTTATCTGCTCTTCGCGTTGAAGAAGTACAAAAGG  240
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |
Sbjct  181  GTTCATTGGAACAGAAAGAAATGGATTTATCTGCTCTTCGCGTTGAAGAAGTACAAAATG  240

Query  241  TCATTAATGCTATGCAGAAAATCTTAGAGTGTCCCATCTG  280
            ||||||||||||||||||||||||||||||||||||||||
Sbjct  241  TCATTAATGCTATGCAGAAAATCTTAGAGTGTCCCATCTG  280
```

If you scroll down on the results page, you'll see an alignment of the sequence you searched. As you can see from this example, the sequence came from **BRCA1**. Copy this gene name down. The mutation is at position 239 where the normal nucleotide ‘T’ (normal BRCA1 Sbjct)is replaced by ‘G’in the mutated query sequence.

Questions [for when you’re looking into your assigned GENE]:

*II-1. Where is the mutation located and what is the nature of the mutation? (example substitution, nonsense mutation, deletion, insertion).*

Now you must use ENTREZ to learn more about the gene. Go back to the NCBI main screen; click on “Entrez Home”, and insert the gene name [or if that doesn’t work, the accession #] this into the ENTREZ search. Then click “go” and click nucleotide.



Search Nucleotide for brca1

History has expired.

ound 5545 nucleotide sequences. Nucleotide [4828] EST [5

Display Summary Show 2

All: 4828 Bacteria: 70 RefSeq: 1887 mRNA: 2732

Items 1 - 20 of 4828

This search in Gene shows 2207 results, including:

Brca1 (*Rattus norvegicus*): breast cancer 1

BRCA1 (*Homo sapiens*): breast cancer 1, early onset

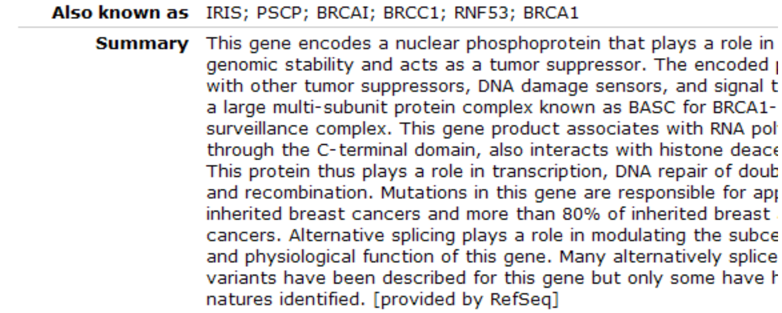
BRCA1 (*Bos taurus*): breast cancer 1, early onset

□ 1: NM_019812 Reports
Mus musculus sirtuin 1 (silent mating type information reg
mRNA
gi|9790228|ref|NM_019812.1|[9790228]

□ 2: NM_001142654 Reports
Homo sapiens alanyl-tRNA synthetase domain containing

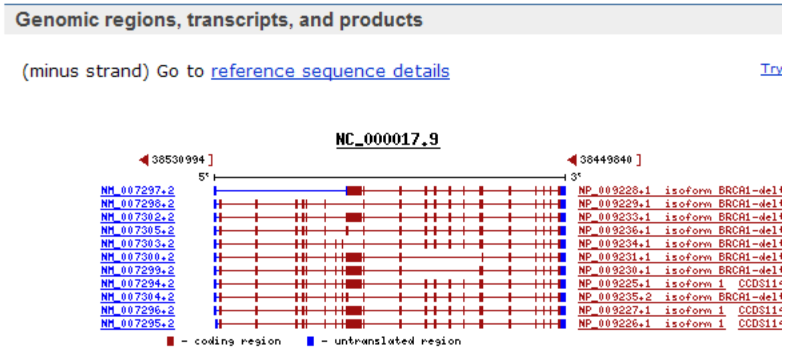
Now click on the gene name link which, in this example, is BRCA1 homo sapiens.

This brings you to the main Entrez screen for the BRCA1 gene. You can get to all information about the gene from here. Bookmark this screen, to make it easy to get back to.



**Also known as** IRIS; PSCP; BRCAI; BRCC1; RNF53; BRCA1

**Summary** This gene encodes a nuclear phosphoprotein that plays a role in genomic stability and acts as a tumor suppressor. The encoded with other tumor suppressors, DNA damage sensors, and signal t a large multi-subunit protein complex known as BASC for BRCA1- surveillance complex. This gene product associates with RNA pol through the C-terminal domain, also interacts with histone deace This protein thus plays a role in transcription, DNA repair of doub and recombination. Mutations in this gene are responsible for ap inherited breast cancers and more than 80% of inherited breast cancers. Alternative splicing plays a role in modulating the subce and physiological function of this gene. Many alternatively splice variants have been described for this gene but only some have natures identified. [provided by RefSeq]

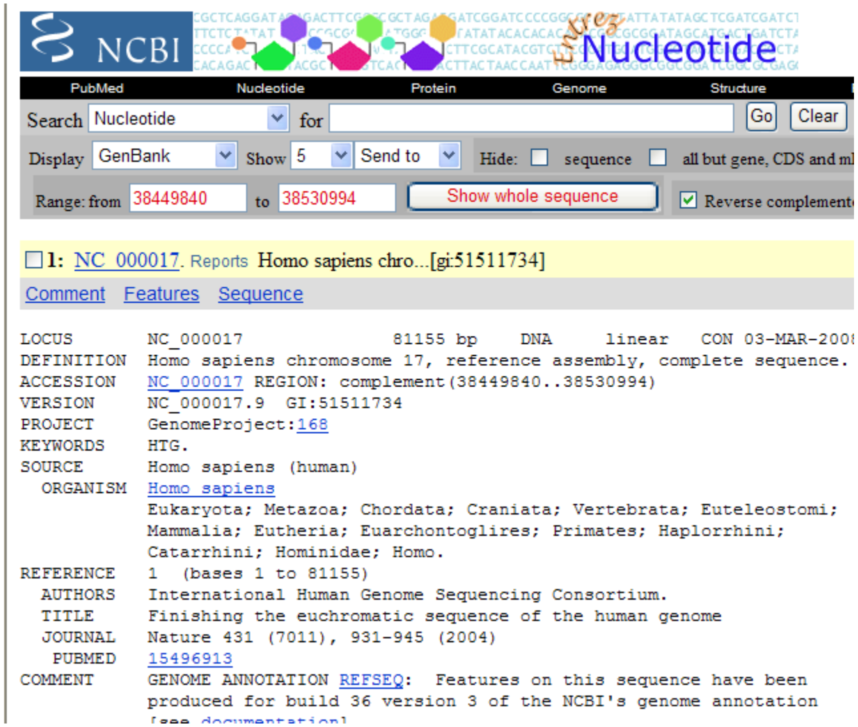


**Genomic regions, transcripts, and products**

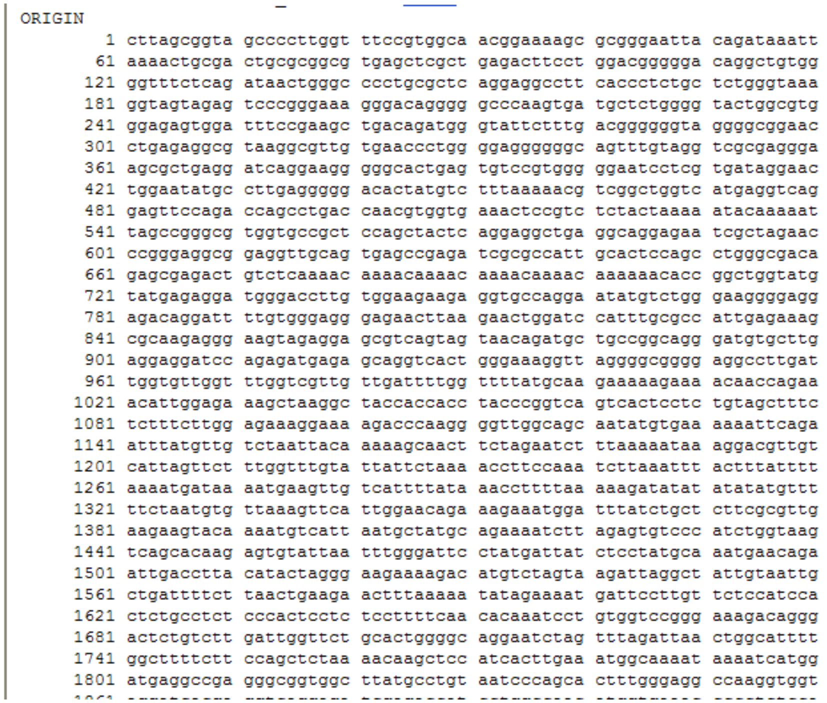
(minus strand) Go to reference sequence details Try

NC_000017.9
38530994
38449840
5'
3'
NM_007297.2
NM_007298.2
NM_007302.2
NM_007305.2
NM_007303.2
NM_007300.2
NM_007299.2
NM_007294.2
NM_007304.2
NM_007296.2
NM_007295.2
NP_009228.1 isoform BRCA1-del
NP_009229.1 isoform BRCA1-del
NP_009233.1 isoform BRCA1-del
NP_009236.1 isoform BRCA1-del
NP_009234.1 isoform BRCA1-del
NP_009231.1 isoform BRCA1-del
NP_009230.1 isoform BRCA1-del
NP_009225.1 isoform 1 CCDS11
NP_009235.2 isoform BRCA1-del
NP_009227.1 isoform 1 CCDS11
NP_009226.1 isoform 1 CCDS11
🟥 - coding region
🟦 - untranslated region

If you click on the NC_ accesssion number, it will go to the DNA sequence of the Chromosomal region.



```
LOCUS       NC_000017               81155 bp    DNA     linear   CON 03-MAR-200
DEFINITION  Homo sapiens chromosome 17, reference assembly, complete sequence.
ACCESSION   NC_000017 REGION: complement(38449840..38530994)
VERSION     NC_000017.9  GI:51511734
PROJECT     GenomeProject:168
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 81155)
  AUTHORS   International Human Genome Sequencing Consortium.
  TITLE     Finishing the euchromatic sequence of the human genome
  JOURNAL   Nature 431 (7011), 931-945 (2004)
   PUBMED   15496913
COMMENT     GENOME ANNOTATION REFSEQ:  Features on this sequence have been
            produced for build 36 version 3 of the NCBI's genome annotation
            [see documentation]
```

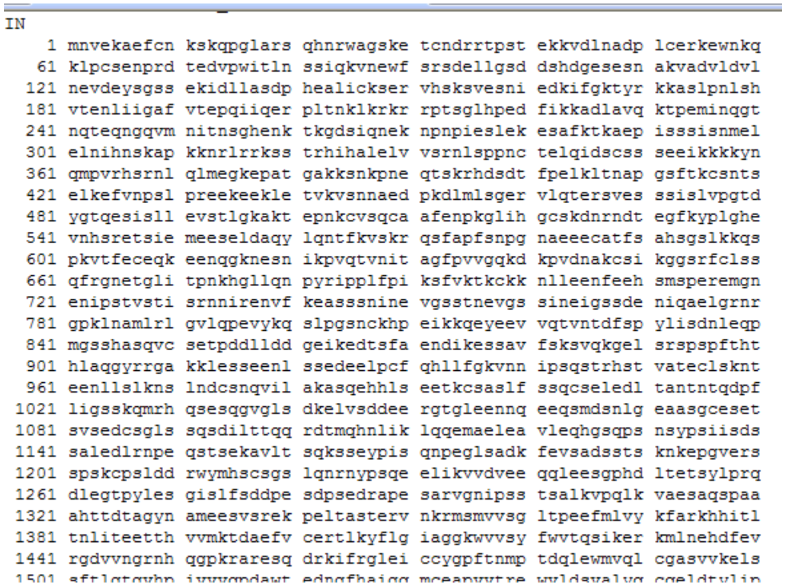


```
ORIGIN
        1 cttagcggta gccccttggt ttccgtggca acggaaaagc gcgggaatta cagataaatt
       61 aaaactgcga ctgcgcggcg tgagctcgct gagacttcct ggacgggggа caggctgtgg
      121 ggtttctcag ataactgggc ccctgcgctc aggaggcctt caccctctgc tctgggtaaa
      181 ggtagtagag tcccgggaaa gggacagggg gcccaagtga tgctctgggg tactggcgtg
      241 ggagagtgga tttccgaagc tgacagatgg gtattctttg acggggggta ggggcggaac
      301 ctgagaggcg taaggcgttg tgaaccctgg ggaggggggc agtttgtagg tcgcgaggga
      361 agcgctgagg atcaggaagg gggcactgag tgtccgtggg ggaatcctcg tgataggaac
      421 tggaatatgc cttgaggggg acactatgtc tttaaaaacg tcggctggtc atgaggtcag
      481 gagttccaga ccagcctgac caacgtggtg aaactccgtc tctactaaaa atacaaaaat
      541 tagccgggcg tggtgccgct ccagctactc aggaggctga ggcaggagaa tcgctagaac
      601 ccgggaggcg gaggttgcag tgagccgaga tcgcgccatt gcactccagc ctgggcgaca
      661 gagcgagact gtctcaaaac aaaacaaaac aaaacaaaac aaaaaacacc ggctggtatg
      721 tatgagagga tgggaccttg tggaagaaga ggtgccagga atatgtctgg gaaggggagg
      781 agacaggatt ttgtgggagg gagaacttaa gaactggatc catttgcgcc attgagaaag
      841 cgcaagaggg aagtagagga gcgtcagtag taacagatgc tgccggcagg gatgtgcttg
      901 aggaggatcc agagatgaga gcaggtcact gggaaaggtt aggggcgggg aggccttgat
      961 tggtgttggt ttggtcgttg ttgattttgg ttttatgcaa gaaaaagaaa acaaccagaa
     1021 acattggaga aagctaaggc taccaccacc tacccggtca gtcactcctc tgtagctttc
     1081 tctttcttgg agaaaggaaa agacccaagg ggttggcagc aatatgtgaa aaaattcaga
     1141 atttatgttg tctaattaca aaaagcaact tctagaatct ttaaaaataa aggacgttgt
     1201 cattagttct ttggtttgta ttattctaaa accttccaaa tcttaaattt actttatttt
     1261 aaaatgataa aatgaagttg tcattttata aaccttttaa aaagatatat atatatgttt
     1321 ttctaatgtg ttaaagttca ttggaacaga aagaaatgga tttatctgct cttcgcgttg
     1381 aagaagtaca aaatgtcatt aatgctatgc agaaaatctt agagtgtccc atctggtaag
     1441 tcagcacaag agtgtattaa tttgggattc ctatgattat ctcctatgca aatgaacaga
     1501 attgacctta catactaggg aagaaaagac atgtctagta agattaggct attgtaattg
     1561 ctgattttct taactgaaga actttaaaaa tatagaaaat gattccttgt tctccatcca
     1621 ctctgcctct cccactcctc tccttttcaa cacaaatcct gtggtccggg aaagacaggg
     1681 actctgtctt gattggttct gcactggggc aggaatctag tttagattaa ctggcatttt
     1741 ggcttttctt ccagctctaa aacaagctcc atcacttgaa atggcaaaat aaaatcatgg
     1801 atgaggccga gggcggtggc ttatgcctgt aatcccagca ctttgggagg ccaaggtggt
```

Scroll down on this screen, and you'll see the actual DNA sequence:

If you click on NM_ it will give the mRNA sequence.   Here, you can determine the transcript size.

If you click NP_ it will give protein sequence information. Here you can find the amino acid sequence and molecular weight.



```
IN
        1 mnvekaefcn kskqpglars qhnrwagske tcndrrtpst ekkvdlnadp lcerkewnkq
       61 klpcsenprd tedvpwitln ssiqkvnewf srsdellgsd dshdgesesn akvadvldvl
      121 nevdeysgss ekidllasdp healickser vhsksvesni edkifgktyr kkaslpnlsh
      181 vtenliigaf vtepqiiqer pltnklkrkr rptsglhped fikkadlavq ktpeminqgt
      241 nqteqngqvm nitnsghenk tkgdsiqnek npnpieslek esafktkaep isssisnmel
      301 elnihnskap kknrlrrkss trhihalelv vsrnlsppnc telqidscss seeikkkkyn
      361 qmpvrhsrnl qlmegkepat gakksnkpne qtskrhdsdt fpelkltnap gsftkcsnts
      421 elkefvnpsl preekeekle tvkvsnnaed pkdlmlsger vlqtersves ssislvpgtd
      481 ygtqesisll evstlgkakt epnkcvsqca afenpkglih gcskdnrndt egfkyplghe
      541 vnhsretsie meeseldaqy lqntfkvskr qsfapfsnpg naeeecatfs ahsgslkkqs
      601 pkvtfeceqk eenqgknesn ikpvqtvnit agfpvvgqkd kpvdnakcsi kggsrfclss
      661 qfrgnetgli tpnkhgllqn pyripplfpi ksfvktkckk nlleenfeeh smsperemgn
      721 enipstvsti srnnirenvf keasssnine vgsstnevgs sineigssde niqaelgrnr
      781 gpklnamlrl gvlqpevykq slpgsnckhp eikkqeyeev vqtvntdfsp ylisdnleqp
      841 mgsshasqvc setpddlldd geikedtsfa endikessav fsksvqkgel srspspftht
      901 hlaqgyrrga kklesseenl ssedeelpcf qhllfgkvnn ipsqstrhst vateclsknt
      961 eenllslkns lndcsnqvil akasqehhls eetkcsaslf ssqcseledl tantntqdpf
     1021 ligsskqmrh qsesqgvgls dkelvsddee rgtgleennq eeqsmdsnlg eaasgceset
     1081 svsedcsgls sqsdilttqq rdtmqhnlik lqqemaelea vleqhgsqps nsypsiisds
     1141 saledlrnpe qstsekavlt sqksseypis qnpeglsadk fevsadssts knkepgvers
     1201 spskcpsldd rwymhscsgs lqnrnypsqe elikvvdvee qqleesgphd ltetsylprq
     1261 dlegtpyles gislfsddpe sdpsedrape sarvgnipss tsalkvpqlk vaesaqspaa
     1321 ahttdtagyn ameesvsrek peltasterv nkrmsmvvsg ltpeefmlvy kfarkhhitl
     1381 tnliteetth vvmktdaefv certlkyflg iaggkwvvsy fwvtqsiker kmlnehdfev
     1441 rgdvvngrnh qgpkraresq drkifrglei ccygpftnmp tdqlewmvql cgasvvkels
     1501 sftlgtgvhp ivvvqpdawt ednqfhaiqq mceapvvtre wvldsvalyq cqeldtylip
```

Questions during this phase of the assignment [related to YOUR OWN GENE].

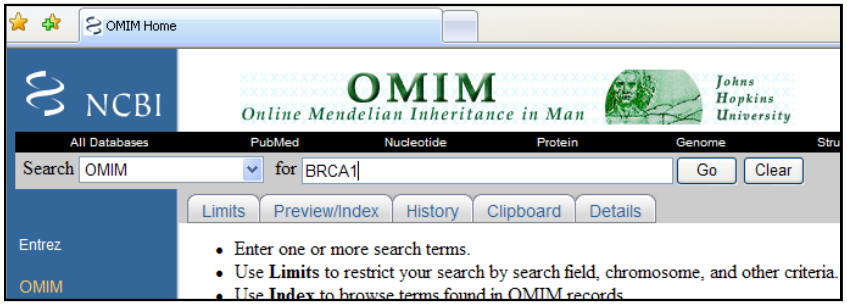
*II-2. How many different transcripts are shown?   How do they differ?*

*II-3: Focusing on the very first transcript:   How many introns and exons are there?*
*What is the length of this mRNA transcript?*

*II-4. What is the number of amino acids of the protein?*

Now you are ready to finally use OMIM to study the biological mechanism of how this mutated gene causes disease (in this example it is breast cancer).   Please go to

NCBI and click OMIM and enter your normal gene (in this example BRCA1) as shown below.

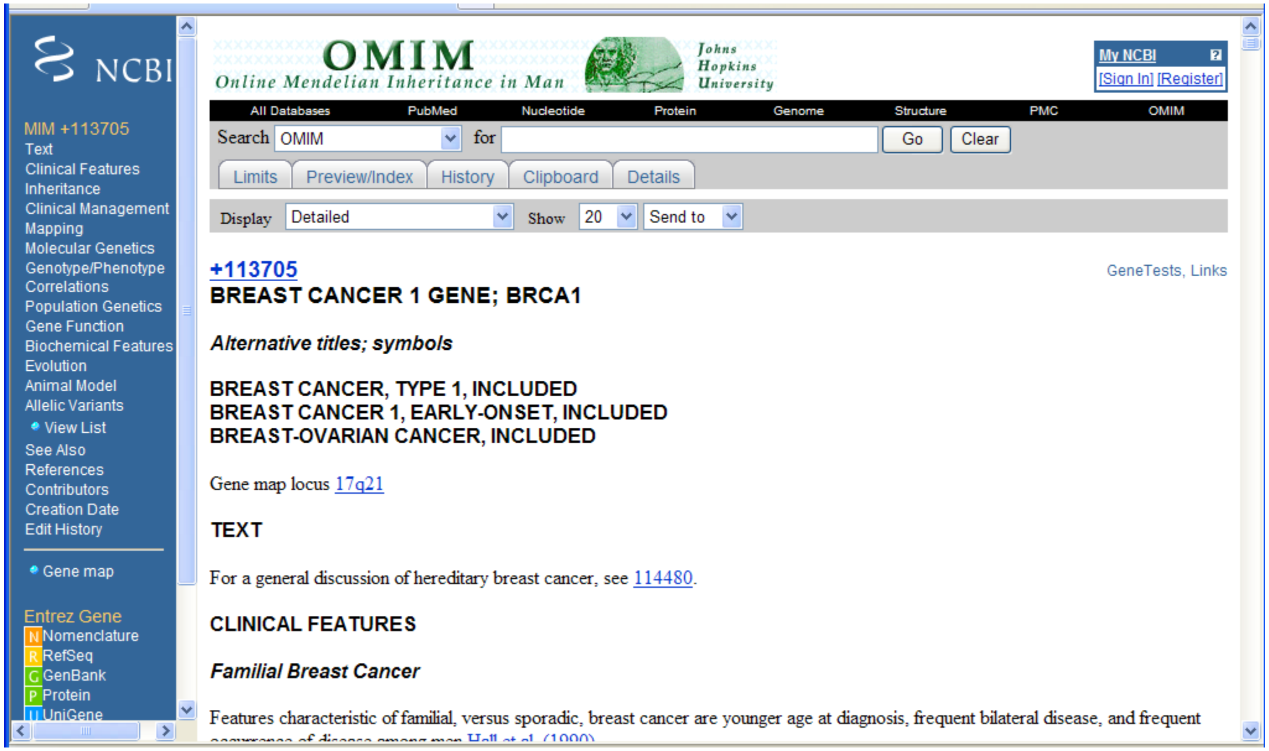


Search OMIM for BRCA1

- Enter one or more search terms.
- Use **Limits** to restrict your search by search field, chromosome, and other criteria.
- Use **Index** to browse terms found in OMIM records.

Click on "Go" and you will see the following screen.

Check to make sure that the first gene is the one of interest, and then click on its number.



Search OMIM for

Display Detailed Show 20

+113705

**BREAST CANCER 1 GENE; BRCA1**

***Alternative titles; symbols***

**BREAST CANCER, TYPE 1, INCLUDED**
**BREAST CANCER 1, EARLY-ONSET, INCLUDED**
**BREAST-OVARIAN CANCER, INCLUDED**

Gene map locus 17q21

**TEXT**

For a general discussion of hereditary breast cancer, see 114480.

**CLINICAL FEATURES**

***Familial Breast Cancer***

Features characteristic of familial, versus sporadic, breast cancer are younger age at diagnosis, frequent bilateral disease, and frequent occurrence of disease among men Hall et al. (1990)

This is the site where you get to play detective and learn about the exciting biology of this gene that causes this disease.

Questions to answer:

*II-5. State which diseases this mutated gene causes.*

*II-6. What chromosome is this gene located on?*

*II-7. What is the function of the normal gene?*

Human genome mutation database HGMD gene search

http://www.ncbi.nlm.nih.gov/

## PART 4: PANDEMIC FLU

The most amazing thing about the 2009 Pandemic flu is, in my opinion, the fact that DNA sequences of the pathogen were posted online in nearly real-time, allowing physicians and scientists around the world to investigate the infection with new computer tools. We will examine those tools here.

Imagine yourself a physician with a patient having a suspected case of H1N1 pandemic flu. You take a swab, and send it to the state lab for testing, which involves using PCR to amplify any flu viruses in the sample, and sequencing the amplified DNA.

Go to http://www.cdc.gov/h1n1flu/
Near the bottom under “Additional Links”, go to the Genbank resources. You might want to bookmark this page.
Restricting your work to this site will limit all searches to influenza viruses, so will make your work easier.

Here is a portion of the sequence found from a virus from your patient:

```
   1 atgaaggcaa tactagtagt tctgctatat acatttgcaa ccgcaaatgc agacacatta
  61 tgtataggtt atcatgcgaa caattcaaca gacactgtag acacagtact agaaaagaat
 121 gtaacagtaa cacactctgt taaccttcta gaagacaagc ataacgggaa actatgcaaa
 181 ctaagagggg tagccccatt gcatttgggt aaatgtaaca ttgctggctg gatcctggga
 241 aatccagagt gtgaatcact ctccacagca agctcatggt cctacattgt ggaaacatct
 301 agttcagaca atggaacgtg ttacccagga gatttcatcg attatgagga gctaagagag
 361 caattgagct cagtgtcatc atttgaaagg tttgagatat tccccaagac aagttcatgg
 421 cccaatcatg actcgaacaa aggtgtaacg gcagcatgtc ctcatgctgg agcaaaaagc
 481 ttctacaaaa atttaatatg gctagttaaa aaaggaaatt catacccaaa gctcagcaaa
 541 tcctacatta atgataaagg gaaagaagtc ctcgtgctat ggggcattca ccatccatct
 601 actagtgctg accaacaaag tctctatcag aatgcagatg catatgtttt tgtggggtca
 661 tcaagataca gcaagaagtt caagccggaa atagcaataa gacccaaagt gagggatcaa
 721 gaagggagaa tgaactatta ctggacacta gtagagccgg gagacaaaat aacattcgaa
 781 gcaactggaa atctagtggt accgagatat gcattcgcaa tggaaagaaa tgctggatct
 841 ggtattatca tttcagatac accagtccac gattgcaata caacttgtca gacacccaag
 901 ggtgctataa acaccagcct cccatttcag aatatacatc cgatcacaat tggaaaatgt
 961 ccaaaatatg taaaaagcac aaaattgaga ctggccacag gattgaggaa tgtcccgtct
1021 attcaatcta gaggcctatt tggggccatt gccggtttca ttgaaggggg gtggacaggg
1081 atggtagatg gatggtacgg ttatcaccat caaaatgagc aggggtcagg atatgcagcc
1141 gacctgaaga gcacacagaa tgccattgac gaaattacta acaaagtaaa ttctgttatt
1201 gaaaagatga atacacagtt cacagcagta ggtaaagagt tcaaccacct ggaaaaaaga
1261 atagagaatt taaataaaaa agttgatgat ggtttcctgg acatttggac ttacaatgcc
1321 gaactgttgg ttctattgga aaatgaaaga actttggact accacgattc aaatgtgaag
```

```
1381 aacttatatg aaaaggtaag aagccagcta aaaaacaatg ccaaggaaat tggaaacggc
1441 tgctttgaat tttaccacaa atgcgataac acgtgcatgg aaagtgtcaa aaatgggact
1501 tatgactacc caaaatactc agaggaagca aaattaaaca gagaagaaat agatggggta
1561 aaactggaat caacaaggat ttaccagatt ttggcgatct attcaactgt cgccagttca
1621 ttggtactgg tagtctccct gggggcaatc agtttctgga tgtgctctaa tgggtctcta
1681 cagtgtagaa tatgtattta a
```

1. Based upon this fragment:
   A. What viral gene is this from?
   B. What strain of the virus does this seem to be from?
   C. Based upon the sequence information available, where would you guess this person was exposed to this virus?
   D. What antivirals do you believe the virus will or will not respond to, based upon the record associated with this gene?

   Download and save the protein sequence.

2. Were this early on in a new infection cycle, you would want to find out what species this virus seems to have come from. Go to the NCBI Influenza Virus Sequence Database. You'll make a tree. Select about 4 sequences each from each:

- USA H1N1 human flu from 2009
- USA H1N1 swine flu, and then avian flu, and a flu from another species, all from 2000-2009

Make sure you're comparing the same gene segments for each virus. Click on "full length", and "Remove identical".

Which viruses are these genes from the human H1N1 flu viruses most closely related to? Print screen and paste a copy of the screen from which you're making your conclusions into this document here.

3. Now that you know you're dealing with a new swine flu, you might want to see if this year's flu vaccine will offer any protection to the public. What part of the protein is most important for vaccines? Open another window, and we'll examine the protein structure. First, go to the protein database "pdb.org". There are many entries for hemagglutinin, but we'd like to see how it interacts with the immune system. So search for "**HEMAGGLUTININ and ANTIBODY".** Take note of the 4 character code next to the check box.

We could examine the protein structure by clicking on the name, but that requires a plug-in that isn't installed on campus computers. So instead, google "firstglance", and search for the code you copied down. "1QFU". Chain 'A' is the viral protein we're looking at; H and L are chains of an ANTIBODY molecule that's bound to the virus.

4. Based upon your aligment, and the "firstglance" model for this protein, if you were to make an 'artificial' vaccine against this new flu, which part of the protein would you want to include [give a rough range of about 50 amino acids], and why?

5. A. Go back to the Genbank page. Find "vaccines" on the menu bar on the left. Look up the contents of the 2008-2009 flu vaccine. What is the name of the H1N1 virus in the vaccine?

B. Using any tool you wish, get the sequence of the hemagglutinin protein of that H1N1 virus that the 2008-2009 vaccine was based upon. Save that sequence somewhere.

C. Use blast2seq to compare the sequence of the protein used in the 2008-2009 vaccine to that of your patient's blood sample. What % identities do you find? Are they very similar in the area you found to be important in question #4?

D. Based upon this result, without further information, would you guess that last year's vaccine would provide much protection against the current pandemic flu? Explain your reasoning in a few sentences.

# Part 5: Investigating a Mutation in HIV-1

**Lab Report: Answer the questions below as your lab report. You may need to do some background research on HIV; Use cdc.gov as a starting resource to find information on HIV.**

**Questions:**

1. Patients A and B are both HIV positive. Patient A has a CD4 count of 650 cells/μL and patient B has a CD4 count of 160 cells/μL. Do both patients have AIDS? Explain why CD4 counts are used as a diagnosis of AIDS.
2. What is meant by the term "lentivirus"?
3. What is proviral DNA?
4. Directions: Draw a haplotype tree [basically a family tree, showing the relationships between the different "clones" or sequences] for the following sequences. These are from a patient, from two different blood draws at different times. The subject was infected with a single clone of HIV which had already evolved into 4 different clones by the time of the second visit. Keep in mind that the haplotype tree should show clones from the second visit evolving from clones from the first visit. (Hint: all clones evolved from V1-1)

V1-1 GAGATAGTAA TTAGATCTGC CAATTTCTCG GACAATACTA AAA 43<br>
V2-1 GAGGTAGTAA TTAGATCTGC CAATCTCACG GACAATGCTA AGA 43<br>
V2-3 GAGATAGTAA TTAGATCTGC GAATTTCACG GACAATACTA AAA 43<br>
V2-2 GAGGTAGTAA TTAGATCTGC CAATCTCACG GACAATGCTA AAA 43<br>
V2-4 GAGGTAGTAA TTAGATCTGC CAATTTCACG GACAATACTA AAA 43

More detailed Procedure:

1. Since all the sequences listed evolved from the V1-1 sequence, use that sequence as your root.

Circle changes from the S16V1-1 sequence in all the other sequences.

2. Start drawing the haplotype tree with the V1-1 sequence as the root. The next sequence(s) should be the one(s) that require the fewest # of changes from V1-1.

On each line connecting two 'clones', write the nucleotide change(s) required to go from one sequence to the next.

3. Continue drawing the tree until all of the clones are included.