**BIO 306 • Genetics**

**Bioinformatics Lab**

Recent advances in high-throughput DNA sequencing (“next-generation” sequencing), gene expression analyses (e.g., microarrays, RNA-Seq), and protein structural analyses have resulted in vast amounts of data stored in publicly accessible databases. With this abundance of data, there is a need for computational tools for searching databases and performing analyses. **Bioinformatics** is the science of biological observation and experimentation *in silico* (contrast with *in vitro* and *in vivo*). Bioinformatics uses mathematics, statistics, and computer science to solve biological problems, and is at the core of modern “-omics” research (genomics, transcriptomics, proteomics, lipidomics, metabolomics, metagenomics, etc.). Although professional bioinformaticians need to have a strong background in computer programming, they have developed a large number of tools that can be used by biologists lacking strong programming skills; indeed, *in silico* analysis is a standard part of the research done in almost any contemporary genetics research laboratory. In today’s lab, you will use some of the standard tools and databases used in genetics research to examine several genetic aspects of a subset of non–small cell lung cancers.

Non-small cell lung cancers (NSCLC) – named to differentiate them from small cell lung cancers (based on tumor cell size determined by microscopic examination) – account for approximately 85-90% of lung cancers [1]. In a study published in 2007 [2], researchers made cDNA copies of mRNAs expressed in an NSCLC tumor surgically removed from a 62-year-old male smoker. Next, a viral delivery system was used to express each cDNA in cultured mouse fibroblast cells. The cell cultures were screened to identify clones that showed signs of transformation to cancer cells. For positive samples, the transforming cDNA was sequenced in order to examine it more closely.

One cDNA capable of inducing cancerous growth in the cell line had the following nucleotide sequence:

GGCGGCGCGGCGCGGCGCTCGCGGCTGCTGCCTGGGAGGGAGGCCGGGCAGGCGGCTGAGCGGCGCGGCTCTCAACGTGACGGGGAAGTGGTTCGGGCGGCCGCGGCTTACTACCCCAGGGCGAACGGACGGACGACGGAGGCGGGAGCCGGTAGCCGAGCCGGGCGACCTAGAGAACGAGCGGGTCAGGCTCAGCGTCGGCCACTCTGTCGGTCCGCTGAATGAAGTGCCCGCCCCTCTGAGCCCGGAGCCCGGCGCTTTCCCCGCAAGATGGACGGTTTCGCCGGCAGTCTCGATGATAGTATTTCTGCTGCAAGTACTTCTGATGTTCAAGATCGCCTGTCAGCTCTTGAGTCACGAGTTCAGCAACAAGAAGATGAAATCACTGTGCTAAAGGCGGCTTTGGCTGATGTTTTGAGGCGTCTTGCAATCTCTGAAGATCATGTGGCCTCAGTGAAAAAATCAGTCTCAAGTAAAGGCCAACCAAGCCCTCGAGCAGTTATTCCCATGTCCTGTATAACCAATGGAAGTGGTGCAAACAGAAAACCAAGTCATACCAGTGCTGTCTCAATTGCAGGAAAAGAAACTCTTTCATCTGCTGCTAAAAGTGGTACAGAAAAAAAGAAAGAAAAACCACAAGGACAGAGAGAAAAAAAAGAGGAATCTCATTCTAATGATCAAAGTCCACAAATTCGAGCATCACCTTCTCCCCAGCCCTCTTCACAACCTCTCCAAATACACAGACAAACTCCAGAAAGCAAGAATGCTACTCCCACCAAAAGCATAAAACGACCATCACCAGCTGAAAAGTCACATAATTCTTGGGAAAATTCAGATGATAGCCGTAATAAATTGTCGAAAATACCTTCAACACCCAAATTAATACCAAAAGTTACCAAAACTGCAGACAAGCATAAAGATGTCATCATCAACCAAGAAGGAGAATATATTAAAATGTTTATGCGCGGTCGGCCAATTACCATGTTCATTCCTTCCGATGTTGACAACTATGATGACATCAGAACGGAACTGCCTCCTGAGAAGCTCAAACTGGAGTGGGCATATGGTTATCGAGGAAAGGACTGTAGAGCTAATGTTTACCTTCTTCCGACCGGGGAAATAGTTTATTTCATTGCATCAGTAGTAGTACTATTTAATTATGAGGAGAGAACTCAGCGACACTACCTGGGCCATACAGACTGTGTGAAATGCCTTGCTATACATCCTGACAAAATTAGGATTGCAACTGGACAGATAGCTGGCGTGGATAAAGATGGAAGGCCTCTACAACCCCACGTCAGAGTGTGGGATTCTGTTACTCTATCCACACTGCAGATTATTGGACTTGGCACTTTTGAGCGTGGAGTAGGATGCCTGGATTTTTCAAAAGCAGATTCAGGTGTTCATTTATGTGTTATTGATGACTCCAATGAGCATATGCTTACTGTATGGGACTGGCAGAAGAAAGCAAAAGGAGCAGAAATAAAGACAACAAATGAAGTTGTTTTGGCTGTGGAGTTTCACCCAACAGATGCAAATACCATAATTACATGCGGTAAATCTCATATTTTCTTCTGGACCTGGAGCGGCAATTCACTAACAAGAAAACAGGGAATTTTTGGGAAATATGAAAAGCCAAAATTTGTGCAGTGTTTAGCATTCTTGGGGAATGGAGATGTTCTTACTGGAGACTCAGGTGGAGTCATGCTTATATGGAGCAAAACTACTGTAGAGCCCACACCTGGGAAAGGACCTAAAGTGTACCGCCGGAAGCACCAGGAGCTGCAAGCCATGCAGATGGAGCTGCAGAGCCCTGAGTACAAGCTGAGCAAGCTCCGCACCTCGACCATCATGACCGACTACAACCCCAACTACTGCTTTGCTGGCAAGACCTCCTCCATCAGTGACCTGAAGGAGGTGCCGCGGAAAAACATCACCCTCATTCGGGGTCTGGGCCATGGAGCCTTTGGGGAGGTGTATGAAGGCCAGGTGTCCGGAATGCCCAACGACCCAAGCCCCCTGCAAGTGGCTGTGAAGACGCTGCCTGAAGTGTGCTCTGAACAGGACGAACTGGATTTCCTCATGGAAGCCCTGATCATCAGCAAATTCAACCACCAGAACATTGTTCGCTGCATTGGGGTGAGCCTGCAATCCCTGCCCCGGTTCATCCTGCTGGAGCTCATGGCGGGGGGAGACCTCAAGTCCTTCCTCCGAGAGACCCGCCCTCGCCCGAGCCAGCCCTCCTCCCTGGCCATGCTGGACCTTCTGCACGTGGCTCGGGACATTGCCTGTGGCTGTCAGTATTTGGAGGAAAACCACTTCATCCACCGAGACATTGCTGCCAGAAACTGCCTCTTGACCTGTCCAGGCCCTGGAAGAGTGGCCAAGATTGGAGACTTCGGGATGGCCCGAGACATCTACAGGGCGAGCTACTATAGAAAGGGAGGCTGTGCCATGCTGCCAGTTAAGTGGATGCCCCCAGAGGCCTTCATGGAAGGAATATTCACTTCTAAAACAGACACATGGTCCTTTGGAGTGCTGCTATGGGAAATCTTTTCTCTTGGATATATGCCATACCCCAGCAAAAGCAACCAGGAAGTTCTGGAGTTTGTCACCAGTGGAGGCCGGATGGACCCACCCAAGAACTGCCCTGGGCCTGTATACCGGATAATGACTCAGTGCTGGCAACATCAGCCTGAAGACAGGCCCAACTTTGCCATCATTTTGGAGAGGATTGAATACTGCACCCAGGACCCGGATGTAATCAACACCGCTTTGCCGATAGAATATGGTCCACTTGTGGAAGAGGAAGAGAAAGTGCCTGTGAGGCCCAAGGACCCTGAGGGGGTTCCTCCTCTCCTGGTCTCTCAACAGGCAAAACGGGAGGAGGAGCGCAGCCCAGCTGCCCCACCACCTCTGCCTACCACCTCCTCTGGCAAGGCTGCAAAGAAACCCACAGCTGCAGAGGTCTCTGTTCGAGTCCCTAGAGGGCCGGCCGTGGAAGGGGGACACGTGAATATGGCATTCTCTCAGTCCAACCCTCCTTCGGAGTTGCACAGGGTCCACGGATCCAGAAACAAGCCCACCAGCTTGTGGAACCCAACGTACGGCTCCTGGTTTACAGAGAAACCCACCAAAAAGAATAATCCTATAGCAAAGAAGGAGCCACACGAGAGGGGTAACCTGGGGCTGGAGGGAAGCTGTACTGTCCCACCTAACGTTGCAACTGGGAGACTTCCGGGGGCCTCACTGCTCCTAGAGCCCTCTTCGCTGACTGCCAATATGAAGGAGGTACCTCTGTTCAGGCTACGTCACTTCCCTTGTGGGAATGTCAATTACGGCTACCAGCAACAGGGCTTGCCCTTAGAAGCCGCTACTGCCCCTGGAGCTGGTCATTACGAGGATACCATTCTGAAAAGCAAGAATAGCATGAACCAGCCTGGGCCCTGAGCTCGGTCACACACTCACTTCTCTTCCTTGGGATCCCTAAGACCGTGGAGGAGAGAGAGGCAATCAATGGCTCCTTCACAAACCAGAGACCAAATGTCACGTTTTGTTTTGTGCCAACCTATTTTGAAGTACCACCAAAAAAGCTGTATTTTGAAAATGCTTTAGAAAGGTTTTGAGCATGGGTTCATCCTATTCTTTCGAAAGAAGAAAATATCATAAAAATGAGTGATAAATACAAGGCCCAGATGTGGTTGCATAAGGTTTTTATGCATGTTTGTTGTATACTTCCTTATGCTTCTTTTAAATTGTGTGTGCTCTGCTTCAATGTAGTCAGAATTAGCTGCTTCTATGTTTCATAGTTGGGGTCATAGATGTTTCCTTGCCTTGTTGATGTGGACATGAGCCATTTGAGGGGAGAGGGAACGGAAATAAAGGAGTTATTTGTAATGAAAAAAAAAAAAAAAAAAAAAAAAAAA

**Follow all steps and answer all questions in the exercise below**.

**Turn in via D2L dropbox, one per pair/group of students.**

1. In the cDNA sequence above, what do the last 27 bases represent? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Using the cDNA sequence, conduct a search of GenBank using the BLASTn algorithm. Navigate to the [NCBI BLAST](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) page and paste the cDNA sequence into the “Enter Query Sequence” box. Under “Choose Search Set”, select the “Human genomic + transcript” database, then click on the blue “BLAST” button in the lower left corner.

1. Examine the Graphic Summary and Descriptions portions of the BLAST results. What are the top 2 hits? You can just use the gene abbreviations.
2. Now, look at the Alignment portion of the output for both of your top 2 hits. Which part of your sequence (the query sequence) matches with each of the top 2 hits? List the nucleotide positions of your **query** sequence that match each hit. Also, list the nucleotide positions of the two hits (i.e., the **subject** sequences) that are involved. For each of the two hits, is the entire gene included in your cDNA sequence? If not, which portion of the gene (5’ or 3’) is included in the cDNA? What do you think has happened?

Let’s find out a little more about both of these hits. Click on the Description link for each of your hits; doing so will take you to the Alignment output for that hit. On the right side, under “Related Information,” select the “Gene” link. This link will take you to the NCBI Gene database record for that gene. Scroll down to the “Genomic Context” section. Note the (chromosomal) Location of the gene (top left portion of the Genomic context box). Then navigate to the “Map viewer” link (top right portion of the Genomic context box).

See **Figure 1** on the next page for a navigation image showing the parts of the map viewer relevant to the following questions. **For each hit, note which chromosome it is on, its map location, and the nucleotide positions that it includes** (enter this information below)**.** The chromosomal location (chromosome number + arm + region; e.g., 17p2) can be determined from the ideogram image on the far left of your screen (you may need to zoom in your browser (Ctrl +) in order to see it clearly). The nucleotide positions can be obtained by clicking on the blue accession number to the left of your sequence name, next to the series of light blue vertical bars on the left; clicking on this number will open a pop-up window that contains the information. Note: this accession will have the same orange-pink highlighting that is around your gene name, but you have to look closely; it may not be directly across from the gene name, and it is quite small.

**Hit 1:** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Chromosome / map location: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Nucleotide positions: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Hit 2:** Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Chromosome / map location: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Nucleotide positions: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Are the two hits on the same chromosome or different chromosomes? If on the same chromosome, are they immediately next to one another?
2. Knowing what you now know about the chromosomal location of these two hits, how do you think both of them ended up in the same cDNA sequence?



**Figure 1**. Navigation for questions #3 and #6.

For each hit, go to the NCBI Sequence Viewer by selecting the sequence viewer (“sv”) link in the pinkish-orange highlighted box containing the hit sequence name in the map viewer (**Figure 1**).

The sequence viewer window is depicted in **Figure 2** on the next page. Your gene will be depicted as a thin green line with arrowheads along its length. These arrowheads show the direction of transcription of the gene.

1. What is its direction of transcription for each of your hits (left to right, or right to left)?
2. Using what you now know about the normal position and direction of transcription of the two genes, describe a type of chromosomal rearrangement that could result in these genes being co-transcribed into one mRNA.

****

**Figure 2**. Example of a sequence viewer window, showing a gene with its direction of transcription depicted by arrowheads. Use for question #6.

The type of genetic occurrence represented by this cDNA sequence occurs in approximately 3-5% of NSCLC patients; these patients tend to be approximately 10 years younger than other lung cancer patients, and are often only light smokers or non-smokers. The drug crizotinib has shown success in treating these tumors, but drug resistance often occurs within a year of treatment due to secondary mutations in the target gene, a process of natural selection similar to the occurrence of antibiotic resistance seen in bacterial populations. Therefore, researchers have looked for additional mutations in these tumor cells; the idea is that, if two or more aberrant (mutant) proteins are targeted at once, there is a lower chance that cells can acquire resistance to the drug regimen. A recent study [3] studied a set of tumors with the genetic rearrangement that you previously examined, looking for mutations in a number of other genes. One of these genes is known as MET.

1. Navigate to the NCBI Gene database (<http://www.ncbi.nlm.nih.gov/gene/>) and search for “MET homo sapiens”. Navigate to the appropriate search result. What are some alternative names for this gene (see the “Also known as” line)? What is its chromosomal location (chromosome and map region)?
2. Consult the NCBI Genetics Home Reference (<http://ghr.nlm.nih.gov/gene>) and search for “MET” using the search box at the top of the screen. Navigate to the appropriate search result. What are two diseases associated with mutations in MET?

Download the file “MET\_sequences.txt” from the course website: (http://websites.uwlax.edu/biology/BIO306Genetics.htm). The file contains a reference (wild-type) MET protein sequence and the MET protein sequences from 5 patients. We will look for differences between the sequences using a multiple sequence alignment program. Navigate to the Clustal Omega alignment tool at the European Bioinformatics Institute site (<https://www.ebi.ac.uk/Tools/msa/clustalo/>). Paste the entire contents of the MET\_sequences file into the Input box labeled “STEP 1 – Enter your input sequences.” Under “STEP 2 – Set your parameters,” select the output format “**Clustal w/ numbers**” instead of the default (w/o numbers). Scroll down a little, and select the “Submit” button (do not request to be notified by email; this alignment will only take a minute or two at most). Examine the alignment. Each row corresponds to one of the sequences (i.e., reference sequence or patient sample). Each column corresponds to one amino acid position in the primary structure of the protein. Columns with an asterisk beneath are **conserved** (i.e., all of the sequences contain the same amino acid at that position). Columns with a dot or space below are **variable** (the dots signify that the amino acids are different, yet biochemically similar to one another).

1. How many amino acid positions are **variable** between samples?
2. Examine the variable positions, and identify which patients have mutations. Where are the mutations (i.e., which position number in the amino acid sequence) and what kind of mutation are they (i.e., what is the amino acid in the wild-type sequence, and what is it changed to in the patient sequence?) [Hint: not every patient has a mutation, and no patient has more than one.] Write the mutations in the following format: Original amino acid, position, mutant amino acid. For example, if the 8th amino acid in the sequence is G in the wild-type and K in one of the patient sequences, write it as G8K.
3. Find a codon table (one is available at <http://www.biogem.org/codon.jpg>). For each of the mutations that you identified, write the most likely mutation to have occurred at the DNA level. [Hint: look for codons that require the least number of nucleotide changes to get from the wild-type amino acid to the mutant one. If there is more than one equally-parsimonious option, assume that transitions (purine 🡪 purine or pyrimidine 🡪pyrimidine) are more likely to occur than transversions (purine 🡨/🡪 pyrimidine). If there are still two or more equally-parsimonious options, write all of them. [Note: *parsimony* is the principle that says we should favor the simplest explanation that can explain the data.]

# REFERENCES

# *Want to know more? Literature references are available from your instructor.*