GENI-ACT

Training Manual

The Western New York Genetics in Research and Healthcare Partnership

Revised 7/1/2017

Stephen Koury, Ph.D.

Western New York Genetics in Research and Health Care Partnership

Department of Biotechnical and Clinical Laboratory Sciences

University at Buffalo, 26 Cary Hall

3435 Main Street

Buffalo, NY 14214

Phone: 716-829-5188

Fax: 716-829-3601

Email: stvkoury@buffalo.edu

Project Websites: ITEST: http://ubwp.buffalo.edu/wnygirp

SEPA: http://ubwp.buffalo.edu/wnygirahcp/

Facebook: https://www.facebook.com/groups/635934719884539/

Development of this Manual was Funded by NIH SEPA Award R25OD010536-01A1 and NSF ITEST Award DRL1311902

Table of Contents

The contributions of Dr. Rama Dey-Rao, Dr. Patricia Masso-Welch, Greer Hamilton MSW, Danise Wilson MPH and Anna Gossin MLS to the preparation of this manual are gratefully acknowledged, as is the assistance received from Dr. Brad Goodner of Hiram College and other members of the Microbial Genome Annotation Network (MGAN, http://www.mgan-network.org).

Background Information

Objective

The objective of this chapter is to provide annotators with basic background about DNA structure, transcription and translation that are relevant to gene annotation you will be performing during this project.

DNA Structure

1. DNA is composed of polymers of deoxynucleotides: deoxyadenosine (A), deoxythymidine (T), deoxycytidine (C) and deoxyquanine (G). Each nucleotide consists of a deoxyribose sugar, a phosphate group and a base. The phosphate group is attached at the 5' carbon of the ribose sugar and an –OH (hydroxyl) group is found at the 3' carbon of the sugar (Figure 1).

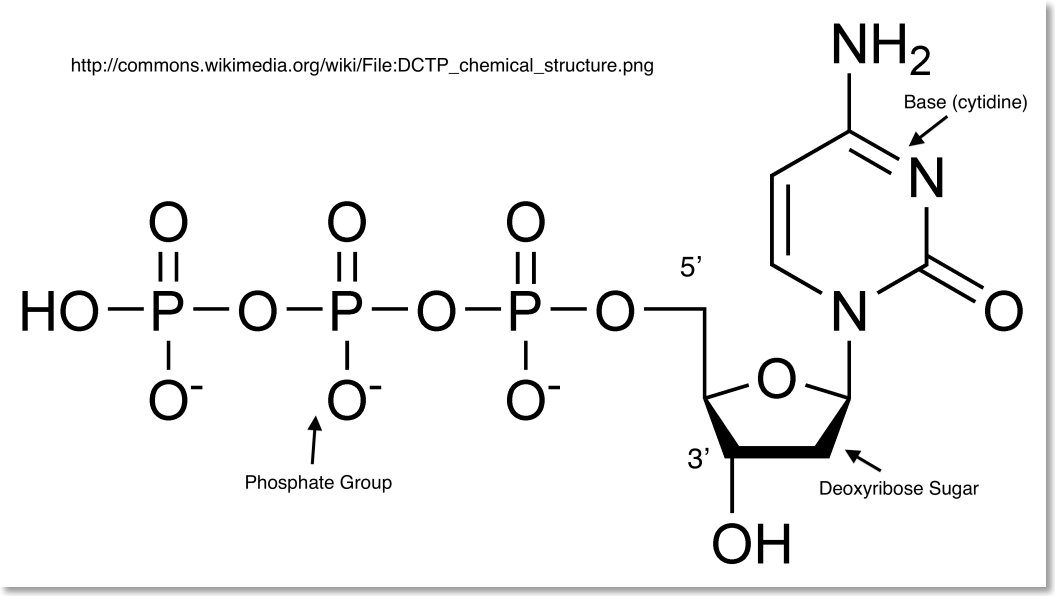

Figure 1. Structure of deoxycytidine.

2. The deoxynucleotides are joined together by way of a phosphodiester bond between the 5' phosphate of one deoxynucleotide and the 3' OH of the other (Figure 2). This gives a strand of DNA polarity, having a free 5' phosphate group at one end of the strand and a free hydroxyl group at the other.
3. Two strands of DNA are held together by hydrogen bonds between A and T bases and between G and C bases. The two strands of DNA in a double stranded DNA are oriented in an antiparallel fashion. The orientation of the strands and the nature of base pairing is illustrated in Figure 3.
4. The two strands of hydrogen bonded DNA form a double helical structure as illustrated in figure 4.

5. When a segment of DNA contains a protein-coding gene, the gene may be located on one strand or the other. The two strands are referred to as the + (also known as the top or forward strand) and – (also known as the bottom or reverse strand) (Figure 4). The top and bottom strand terminology arises from the convention of representing the two strands of DNA as linear, rather than helical, when describing their sequence (Figure 4). The 5'end of the top strand is at the left and the 5' end of the bottom strand is at the right.

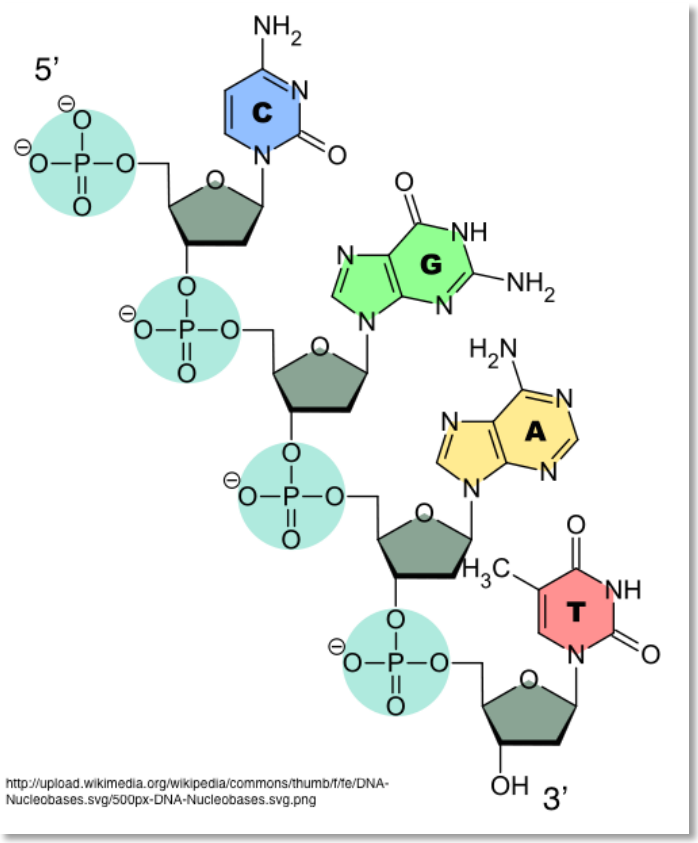

Figure 2. Phosphodiester bonds to create a strand of DNA. The positions of the 3' and 5' carbons are shown.

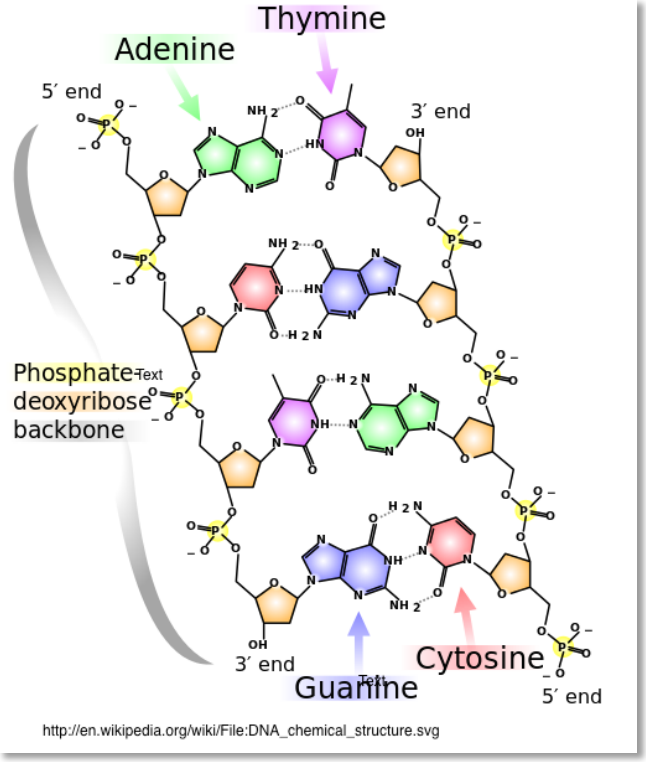

Figure 3. Base pairing of two antiparallel DNA strands. Note the orientation of the left strand with its 5' end oriented toward the top and the right strand with its 5' end oriented

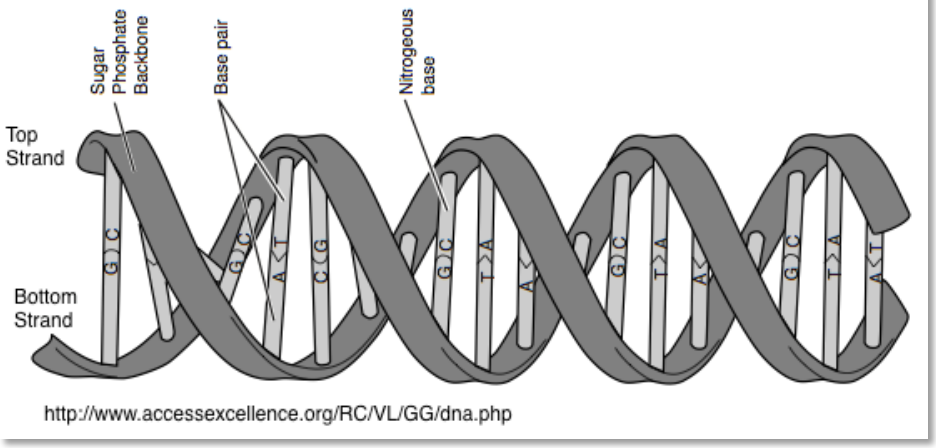

Figure 4. The DNA double helix. The top (+) and bottom (-) strands are indicated.

Prokaryotic Gene Structure, Transcription and Translation

6. The information stored in the DNA of a gene first must be copied into messenger RNA (mRNA) before a protein can be synthesized (note that not ***all*** genes encode a protein), a process referred to as transcription.

7. One of the two strands of DNA will serve as a template for transcription of RNA. The other strand has the same sequence of nucleotides as in the RNA molecule, with the exception that RNA is composed of ribonucleotides rather than deoxynucleotides and Uracil replaces Thymine in RNA. The sequence of DNA identical to that of the mRNA is the coding strand, while the strand that is used to make mRNA is referred to as the template strand (Figure 5). You will learn during your annotation how it is determined that a gene might be found in a particular stretch of DNA, but for the illustration below a region at the 5' end of the coding strand is indicated where the molecule responsible for transcribing the template strand into an mRNA is indicated.

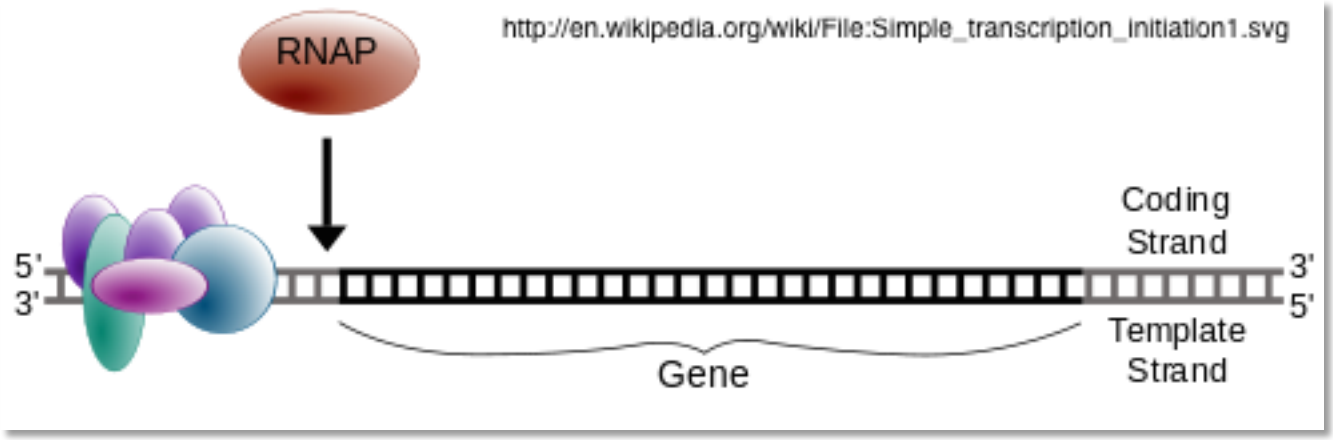

Figure 5. The structure of a prokaryotic gene with the top and bottom strands illustrated. At the 5' end of the top strand is an area that defines where an RNA polymerase molecule (RNAP) can bind.

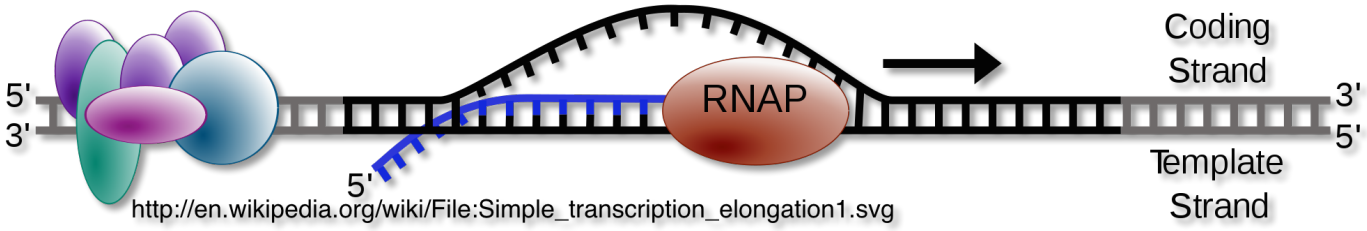

Figure 6. Transcription of an mRNA complementary to the template strand by RNA polymerase. The resulting mRNA has the same sequence as the coding strand of DNA, but is composed of ribonucleotides and uracil is incorporated instead of thymine.

8. The two strands of DNA unwind and the RNA polymerase copies the template strand by incorporating ribonucelotides complementary to the template strand into the mRNA (Figure 6).

9. Once the mRNA for a protein-encoding gene has been transcribed, it associates with ribosomes in the bacterial cytoplasm and is translated into protein.

10. Translation requires that the ribosome "read" the information contained in the mRNA and adds amino acids in the correct order to the growing protein. The language of DNA is based on groups of 3 nucleotides encoding specific amino acids. The code is shown in figure 7 below, which illustrates the combinations in DNA that encode amino acids (called triplets). In the mRNA these combinations are referred to as codons, and U would replace T. As you look at the table you will notice that there are variable numbers of combinations of nucleotides that are translated to a particular amino acid. For example, the amino acid methionine (Met in 3 letter designation and M in single letter designation) is encoded only by ATG in DNA or AUG in mRNA. Methionine and tryptophan (Trp, W) are the only amino acids with a single triplet or codon. In contrast, the amino acid leucine (Leu, L) has 5 different triplets or codons that encode for its addition into a protein. There are 64 possible codons and the fact that all other amino acids other than M and W have more than one codon to encode for their incorporation into proteins illustrates that the code is redundant. You will also notice that there are 3 codons that encode for a STOP. These codons, when encountered, tell the ribosome to stop adding amino acids to the protein and signal the termination of translation of the mRNA.

11. Amino acids are brought to the ribosome by molecules called transfer RNAs (tRNA) that have an anticodon on one end (complimentary to the codon on the mRNA molecule) and the attached amino acid specific for that codon. The ribosomal RNA catalyzes the formation of a peptide bond between the last amino acid added to the protein and the one newly arriving on the tRNA (Figure 9). A segment of DNA that encodes a protein will thus have a triplet that signals the first amino acid of the protein (a start

Table of Standard Genetic Code

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

Figure 7. The genetic code. The letters on the right are the first nucleotide of a triplet, the letters across the top of the table are the second nucleotide of a triplet. The 3 letter and single letter designations are shown for each amino acid. Source: http://www.apsnet.org/edcenter/K-12/TeachersGuide/PlantBiotechnology/Pages/Modifications.aspx

codon), a variable number of triplets that encode all the amino acids of the protein and then a stop triplet to end the incorporation of amino acids. In bacteria most proteins have a methionine (ATG) as the first amino acid, but some proteins can begin with either leucine (TTG) or valine (CTG).

12. A protein-coding gene will thus have what is called a long open reading frame that begins with a start triplet and ends with a stop triplet. You may have deduced that since we take 3 nucleotides at time to define an amino acid, there are 3 different potential reading frames for each strand of DNA, depending if we start with the first three nucleotides, or if we start reading triplets from the second nucleotide, or if we start reading triplets from the third nucleotide . This is better illustrated in Figure 8 below, where reading fame 1(blue) begins with AGG, reading frame 2 (red) begins with GGT and reading frame 3 (green) begins with GTG. When a new DNA sequence is analyzed for the presence of genes, all three reading frames are checked for potential start codons. If one exists in a reading frame the triplets that follow are read until a stop codon is encountered. If a long enough reading frame exists, then the sequence has the potential to be a protein-encoding gene. Frames 4, 5 and 6 would be found on the opposite strand of DNA.

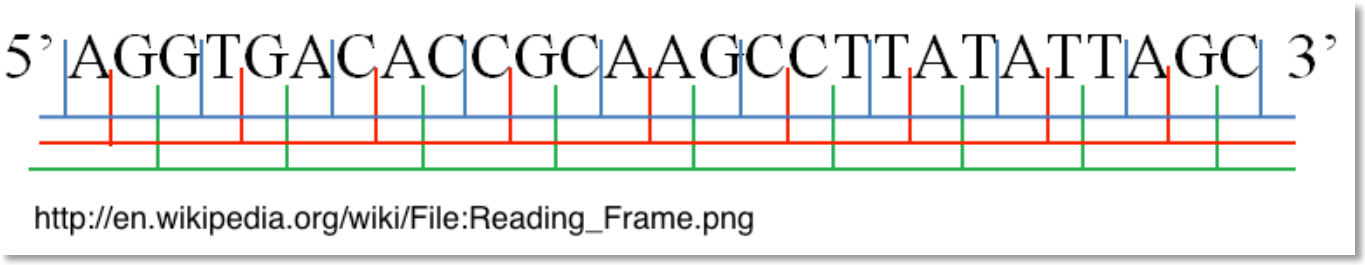

Figure 8. Illustrations of the three possible reading frames in a DNA sequence.

13. In addition to the start codon, long open reading frame and a stop codon, some bacterial genes have a sequence of nucleotides 5' to the start codon called the Shine-Dalgarno sequence, that facilitates the binding of the mRNA to the ribosome to being translation. We will discuss this more in one of the annotation modules in which you will analyze your gene for alternative start codons or reading frames.

14. An overall summary of the process of transcription and translation of mRNA in bacteria is shown in figure 9.

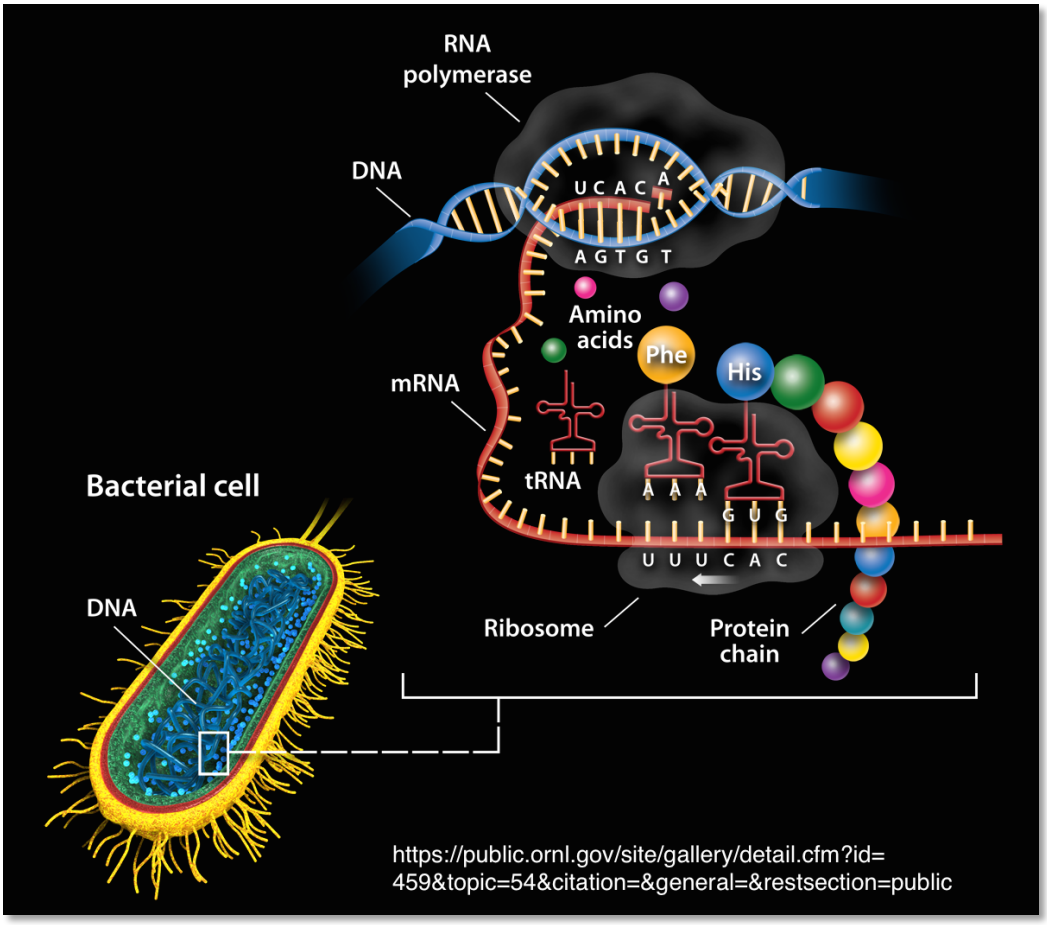

Figure 9. A summary of the transcription and translation of mRNA in bacteria.

General Considerations of Gene Annotation

1. You will be taking a modular approach to annotation of the gene or genes assigned to you as part of this project. Annotation is the process of assigning function or biological significance to a gene.
2. Each participant group will be working on genes from a different clinically significant microorganism. Basic information about the genome on which you will be working can be found by doing a "genome search" at the following link: https://img.jgi.doe.gov/cgi-bin/m/main.cgi .

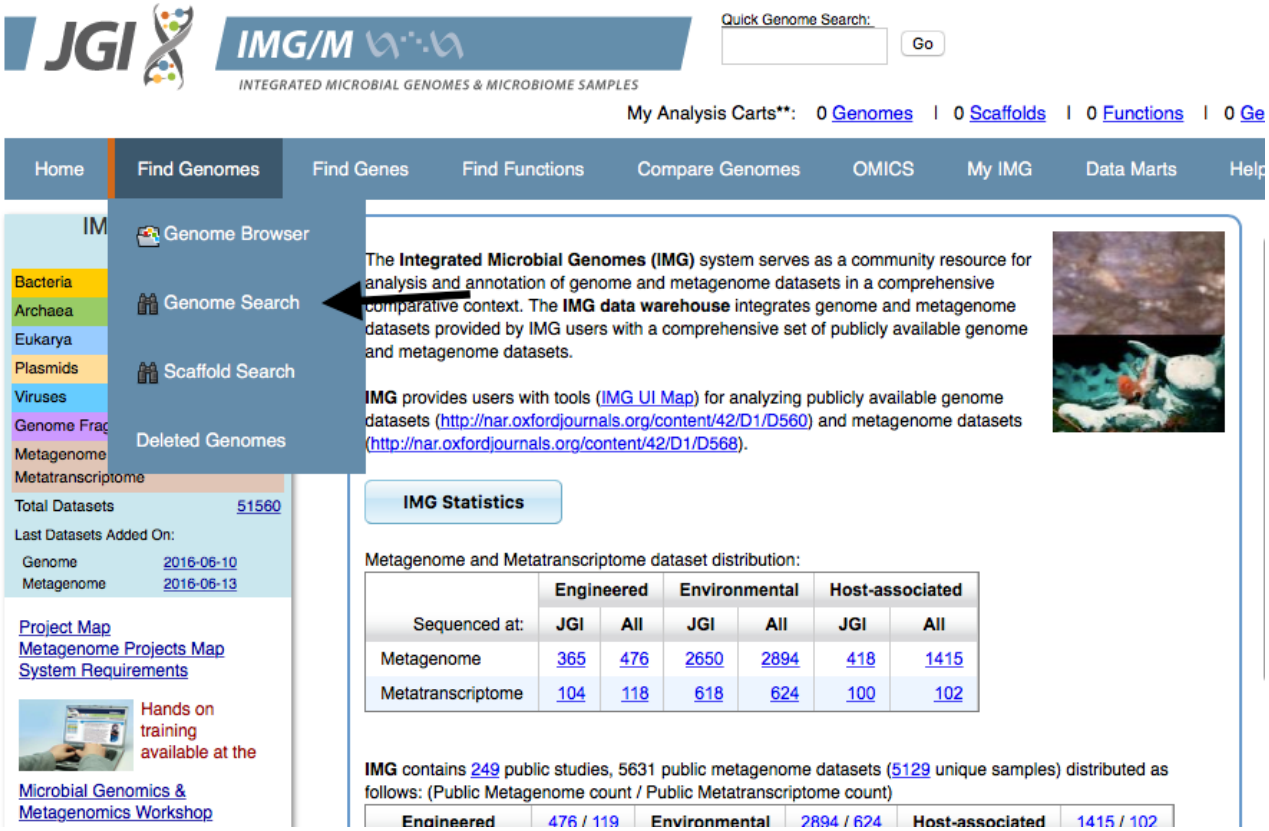

Figure 10. The IMG/EDU entry page. The arrow points to the Genome Search option of the Find Genomes pull down menu.

3. Figure 10 shows the IMG/M entry page with the Find Genomes pull down menu selected. You should select Genome Search from this menu to be taken to a page where you can find basic information about the sequencing project for the genome on which you are working.

4. Figure 11 shows the Genome Search window. Enter the genome name of the organism on which you are working. Be sure to enter the entire name, as for a number of organisms there are multiple numbers of variant that have had their genomes sequenced. In the example shown in Figure 11, the genome name is *Listeria monocytogenes* 08-5578.

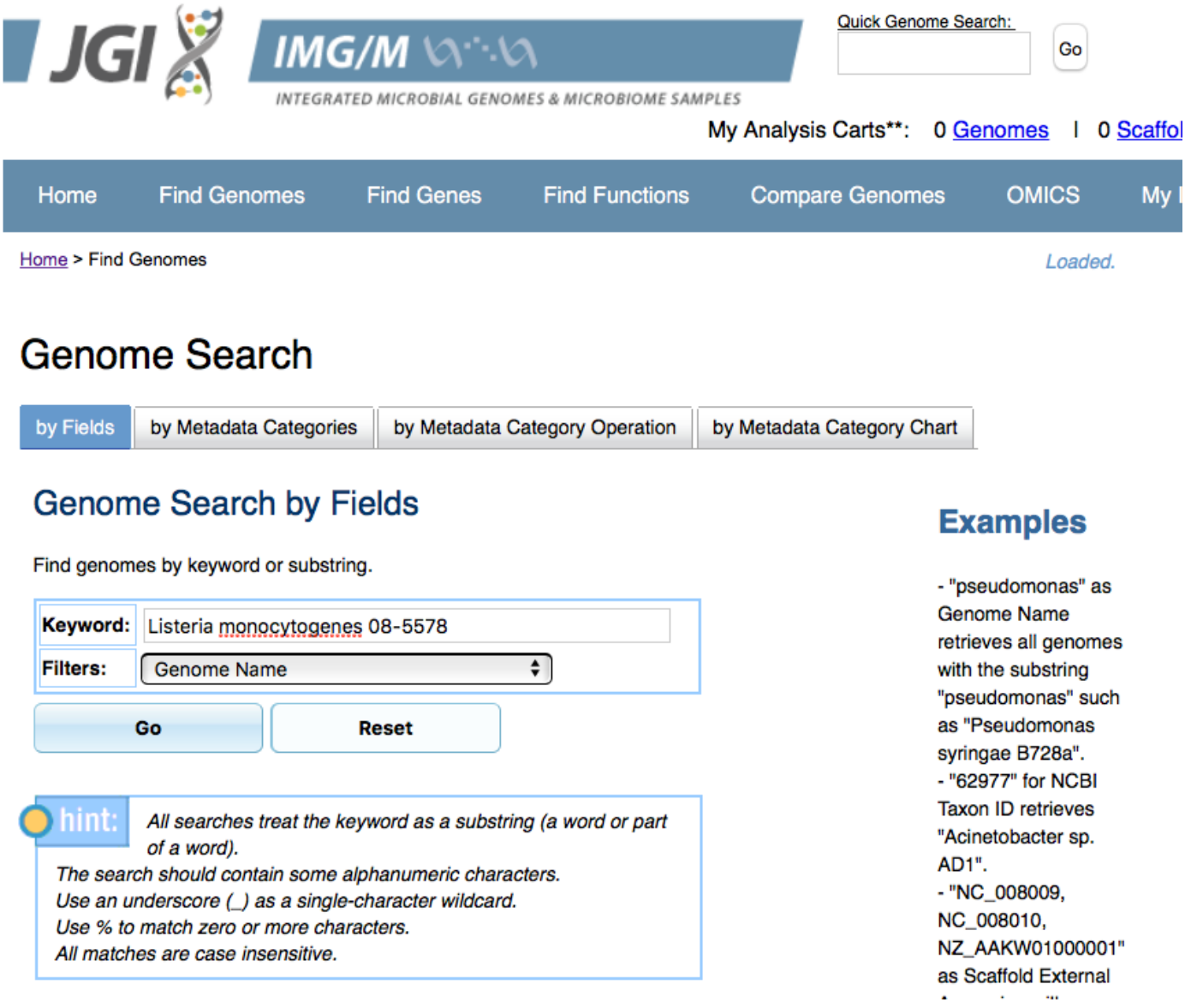

Figure 11. The IMG/M genome search page. The genome name *Listeria monocytogenes* 08-5578 has been entered into the keyword search window.

5. Figure 12 illustrates the results of a Genome Search for *Listeria monocytogenes* 08-5578. The more general your search, the more likely it is that you may find more than one genome listed. Click on the hyperlinked name in the Genome Name column to get to the information page about the genome you are investigating.

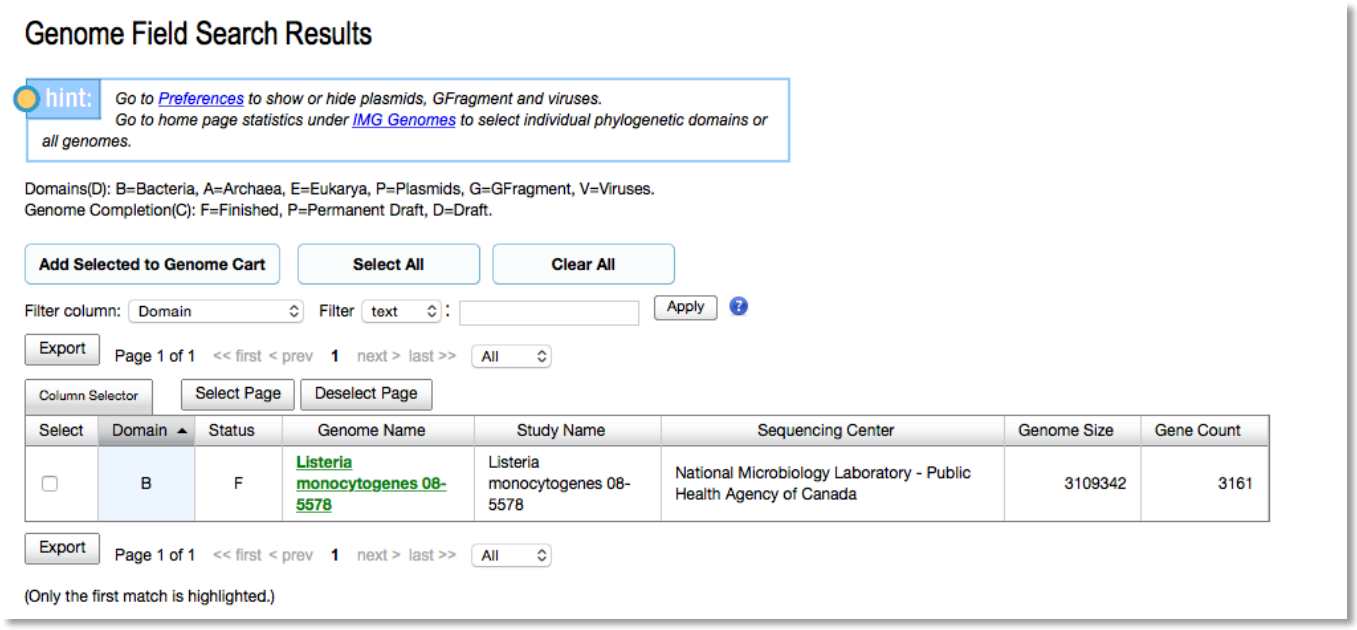

Figure 12. Genome Field Search Results in IMG/M. Clicking on the hyperlink to *Listeria monocytogenes* 08-5578 will open a summary page about the genome.

6. Figure 13 illustrates the upper most section of genome information page for *Listeria monocytogenes* 08-5578. We can see in Figure 13 that the genome sequencing has been completed, where the sequencing took place and links to other sites containing information about the genome. The overview section will also tell you where and why the bacterium was isolated and provide links to the assembled sequence file (NCBI Project ID, in the Project Information Subsection, Figure 14).

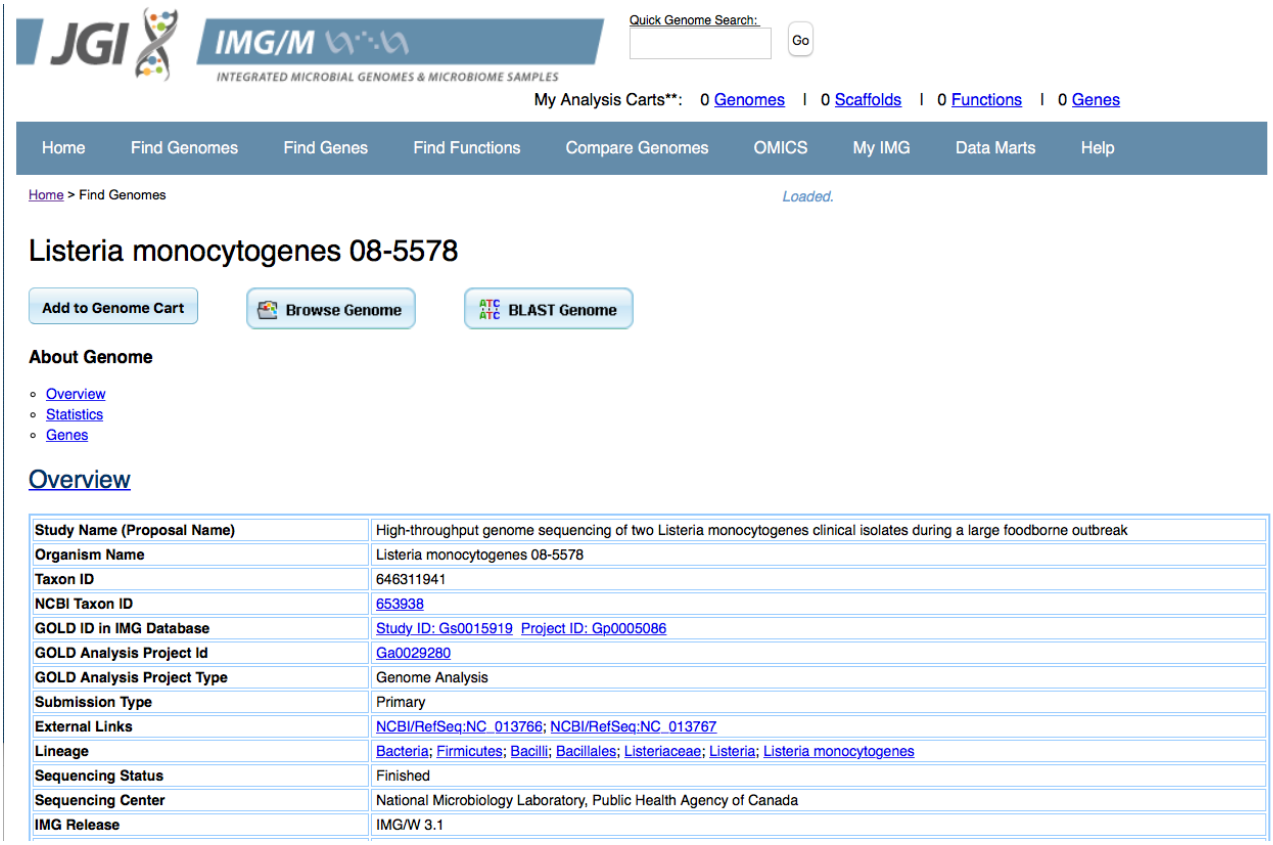

Figure 13. The uppermost portion of the genome search results page for *Listeria monocytogenes* 08-5578

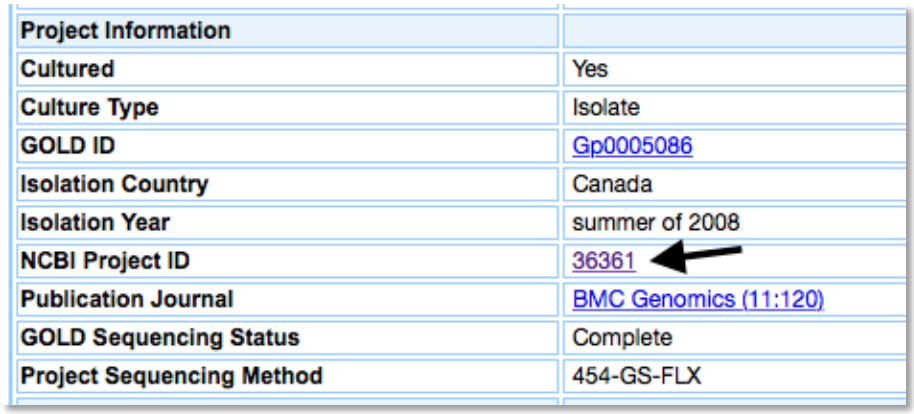

Figure 14. The Project Information portion of the genome information page. The arrow points to a link describing the project at the National Center for Biotechnology Information (NCBI)

7. Scrolling further down the page will lead to a section called Genome Statistics (Figure 15).

Genome Statistics

hint: *To view rows that are zero, go to MyIMG preferences and set **"Hide Zeroes in Genome Statistics"** to **"No"**.*

	Number	% of Total
DNA, total number of bases	3109342	100.00%
DNA coding number of bases	2799663	90.04%
DNA G+C number of bases	1179103	37.92% [1]
DNA scaffolds	2	100.00%
CRISPR Count	1	
Plasmid Count	1	
Genes total number	3161	100.00%
Protein coding genes	3088	97.69%
RNA genes	73	2.31%
rRNA genes	15	0.47%
5S rRNA	5	0.16%
16S rRNA	5	0.16%
23S rRNA	5	0.16%
tRNA genes	58	1.83%
Protein coding genes with function prediction	733	23.19%
without function prediction	2355	74.50%
Protein coding genes connected to SwissProt Protein Product	1	0.03%
not connected to SwissProt Protein Product	3087	97.66%
Protein coding genes with enzymes	841	26.61%
Protein coding genes connected to Transporter Classification	456	14.43%
Protein coding genes connected to KEGG pathways[3]	965	30.53%

Figure 15. A portion of the Genome Statistics output for *Listeria monocytogenes* 08-5578. Information about the size of the genome and characteristics of genes predicted to exist by computer annotation are shown.

8. You can quickly see information about what is known about the genome of your organism from the genome statistics page. For example, as is shown in Figure 15, the genome of *Listeria monocytogenes* 08-5578 has approximately 3.1 x 10^6 nucleotides (see "DNA, total number of bases") and the percentage of those nucleotides that are either G or C is 37.9%. The G+C content will be used later in your gene annotation exercises.

9. Computer analyses have been applied to the raw sequence data and done two different jobs.

    a. **Gene calling** - The first thing the computer has done is to identify sequence of DNA it "thinks" represent genes. This is one place were computer annotation can have errors. Sometimes it calls the wrong start and / or stop positions of a gene and other times it is completely wrong in its identification of a gene, or it fails to call a gene that really is there.

    b. **Function prediction –** the computer looks at the genes it predicts to exist and then compares them to other genes from other organisms that have been sequenced. If the gene under consideration seems to be a good match with other genes that have had their function predicted or experimentally determined, the computer may call the new gene by the same name. If it finds sequence similarity to functional domains in other known genes it may say that the protein has a putative function, for example an ATPase, but not call the gene by name. Two other potential calls are "hypothetical" or pseudogene. Hypothetical genes look like genes to the computer, but the computer cannot determine what function if might have. Pseudogenes are genes that were once functional but have lost their function due to some sort of mutation.

10. Referring back to the Genome Statistics page for *Listeria monocytogenes* 08-5578 in Figure 15, it can be seen that 3161 genes have been predicted by the computer analysis. Of the total of 3161 genes predicted to exist, a function has been assigned to only 733!

11. The human brain has unique properties that allow it to make connections that might not be obvious even to the best supercomputer. Manual annotation of genes, such as you will soon begin to do, allows errors in the computer analysis to be caught and may help to identify function in genes called hypothetical by the computer.