Annotation of the Bacteriophage 933W Genome: An In-Class Interactive Web-based Exercise

Resource Type: Curriculum: Classroom

Publication Date: 9/29/2008

Authors

David J. Baumler
University of Wisconsin—Madison
Madison, WI 53706
USA
Email: dbaumler@wisc.edu

Kai F. Hung
Department of Bacteriology
University of Wisconsin—Madison
Madison, WI 53706
Email: kfhung@wisc.edu

Eric L. Cabot
Genome Center of Wisconsin
University of Wisconsin—Madison
Madison, WI 53706
USA
Email: ecabot@wisc.edu

Jeremy D. Glasner
Genome Center of Wisconsin
University of Wisconsin—Madison
Madison, WI 53706
Email: jglasner@wisc.edu

John M. Greene
Global Health Sector
SRA International, Inc.
Rockville, MD 20852
Email: John_Greene@sra.com

Nicole T. Perna
Department of Genetics
University of Wisconsin—Madison
Madison, WI 53706
Email: ntperna@wisc.edu

Abstract

Simple genome sequences, like those of bacteriophage (bacterial viruses), are excellent resources for teaching students about genome analysis. This exercise allows students to annotate the genome of the bacteriophage 933W, a virus isolated from pathogenic *Escherichia coli* O157:H7. Students use Internet-based resources in class to add annotations for products and functions to genes from 933W based on queries against genome and protein databases. By the end of class, students will have annotated a simple genome and learned about the various methods and steps involved in adding annotations to genes in a genome database.

Activity

Invitation for User Feedback. If you have used the activity and would like to provide feedback, please send an e-mail to MicrobeLibrary@asmusa.org. Feedback can include ideas which complement the activity and new approaches for implementing the activity. Your comments will be added to the activity under a separate section labeled "Feedback." Comments may be edited.

Learning Objectives.

This activity will help students:

(i) enhance their understanding about how genes are identified in a genome,

(ii) improve their knowledge about hypothetical proteins, and

(iii) be able to use the tools BLAST and InterProScan to derive biological information for adding product and function annotations to a gene in a genome.

Background.

Students should have knowledge of the basics of genetics, the central dogma of molecular biology (DNA to RNA to protein), and an understanding that the entire collection of genes comprise a genome. Extensive computer skills are not required, but the students should be familiar with using a web browser (e.g., Internet Explorer or Mozilla Firefox).

Materials.

A classroom with wireless Internet access (request that students bring wireless-ready laptops) or a classroom or computer lab with Internet access is required for this exercise. Contact the staff at the ASAP database (pmliss@wisc.edu) and request to set up your own unique version of the bacteriophage 933W genome to annotate with your class. Note: you will be required to submit a username and a password. You and your class will all use the same log in information, so choose a username and password that do not include sensitive information. We recommend using all upper or lower case letters, if possible, to make it easier for a large class to all log in. One suggestion is to use the course name for the username (e.g., BACT650) and the school mascot for the password (e.g., badgers). You can request to set up multiple versions of the genome if using it in multiple classes or during multiple semesters. Note: each will require a new username and password.

Student Version.

In response to student feedback, a step-by-step guide showing exactly what to do and where to navigate while using the tools has been created. Students found that they were able to follow along during the instructor-led annotaton of the *int* gene, but when given their own gene they had some difficulties. Therefore an instructional guide for the students was prepared and field tested. Also included is a Shockwave Flash file for an animated student instruction guide for distance learning environments. The .swf file can be dragged into a web-browser to view.

Instructor Version.

A. Introduction to genome annotation and analysis

Have you ever wondered how biologists make sense of the millions of base pairs of DNA sequence that make up a genome? How are annotations, or the biological information, attached to the sequence to create the files available as a GenBank report at the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov)? With the advent of whole genome sequencing, the number of microbial genome sequences available has grown exponentially, yet students are not typically involved in the process of annotating genes and exploring genomic data. When most textbooks address the topic of genomics, they often overlook the steps on how information for annotations is derived. In this exercise, the topic of adding information to newly sequenced genomes is presented. Through the use of bioinformatic tools that query sequence databases such as BLAST and InterProScan, students will examine a predicted open reading frame (ORF) from the bacteriophage 933W genome and add appropriate annotations into the ASAP genome database.

The ASAP database is unique in several ways that facilitate this exercise. An important feature in this system is that every annotation (such as the claim that a particular gene encodes a protein with a specific biological function) has a link to the evidence that supports the assertion made in the annotation. Many databases display the assertion made in the annotation, but do not provide the viewer (or student) with a description of how the information was derived. This feature of ASAP is now highlighted as a model approach to add evidence links for annotations and genes in other genome databases. In other designed modules that use the ASAP database, the link for experimental evidence goes directly to the PubMed abstract, enabling the students to access relevant annotation information quickly and easily. Another reason the ASAP database is useful for this exercise, is that all of the genomes housed in the ASAP database are from the evolutionarily related family the *Enterobacteriaceae*. Members of this family include *Escherichia coli* and *Salmonella* spp., both of which should be familiar terms to students since these organisms are often mentioned in the media in relation to food safety and/or water quality issues. We chose a bacteriophage from *E. coli* O157:H7, since they are important for the microorganism's pathogenesis, evolution, and role in public health. A bacteriophage was selected, as opposed to a whole genome of a bacterium, because we wanted to keep the complexity level of the exercise low. A phage genome is smaller and easier to conceptualize than a bacterial genome, which should help students better understand the genome annotation process.

To introduce the topic of annotations, a slide presentation in PowerPoint format is provided to aid the instructor's presentation of the exercise to the class. The instructor can use these slides to give a brief overview of genomic biology and provide an introduction to the concepts discussed in the exercise described below. We have provided explanations and key topics to address from each slide (see below) in the presentation. All slides are numbered and correspond to the commentary provided.

Slide 1. Introductory slide, with students pictured from an introductory biology course at University of Wisconsin—Madison engaging in this module using wireless ready laptops in a large lecture room equipped with wireless internet. In order to teach topics about genomics, you need to get computers in the students' hands. Based on campus computing survey statistics at UW-Madison in 2007, 77% of undergraduates own their own laptop computer, and laptop checkout programs are prevalent throughout the campus. If you have a small course, it may be feasible to arrange implementation of this module in a computer laboratory setting where students will have access to desktop computers. Alternatively, if your institution is wireless-equipped for the classroom, you can request your students to bring their own laptops. For those students without laptop computers, look into laptop check-out programs at your institution and provide this information to students early in the semester so they may reserve computers for the date that you will conduct this exercise (include in the syllabus). Students can also share a laptop if needed.

Slide 2. Definition of a genome: this slide is meant to serve as an explanation of the basic concept of a genome as the

complete complement of genetic material from an organism.

Slide 3. This slide introduces the first genome ever sequenced. The relative small size of phage genomes made them well-suited to be the first genomes to be sequenced completely. This slide mentions one of the pioneers in genome sequencing, Frederick Sanger; through his efforts, the first genome was sequenced of a small phage that contained only 10 genes. Dr. Sanger was later awarded the Nobel Prize in 1980 for this work that started the emergence of the field of genomics.

Slide 4. History of DNA sequencing and genome (continued). This slide mentions the next genome that was sequenced, another phage, but larger than the first. Dr. Sanger and colleagues developed a new method of sequencing called "shotgun" sequencing, which enabled the sequencing of large strands of DNA at a much faster pace (a brief definition of "shotgun" sequencing is provided by Dr. Francis Collins, Former Director of the National Human Genome Research Institute at http://www.genome.gov/glossary.cfm?key=shotgun%20sequencing). This genome contained 46 genes, and a visual layout of the genome containing the genes is provided in the illustration. By now, it should become obvious to the students that the early history of the field of genomics focused on phage, due to their relatively small genome size. Therefore, to teach the subject of annotations to a class, this exercise also uses a phage since it is relatively small (78 genes), and is a genome that a class may be able to completely annotate and sense a feeling of accomplishment.

Slide 5. In 1995, sequencing of the complete genome of *Haemophilus influenza* was finished, making it the first bacterial genome, as well as the first nonviral genome, to be completed. It was also the largest finished genome at the time, with 1,709 genes. The effort to fully sequence the *H. influenza* genome took 13 months. When presenting this slide, I use what we call the "JFK analogy." The day that president John F. Kennedy was shot impacted everyone's lives, and they always remember where they were when the news was announced. Similarly, when the news of the first bacterial genome sequence completion was announced, an instructor declared that this was a hallmark event for science and one that would surely impact everyone's lives, as the field of genomics was preparing to explode and would soon map a new avenue of science in microbiology and eventually human medicine.

Slide 6. The first yeast genome is completed. This was the first organism with multiple chromosomes to have its genome sequenced and represented an increase in the genome size (6,269 genes). As stated on the slide, it took an enormous amount of time (7 years) and labor (74 labs) to complete. This information is presented to provide an appreciation of how DNA sequencing technology will advance in the coming decades to the point where it is believed that human genomes will be sequenced in approximately 15 minutes by the year 2013.

Slide 7. Many other completed eukaryotic genomes were soon to follow with completed genome sequences. This slide mentions the next series of hallmark genome projects, with the first fruit fly (13,000 genes), nematode (19,000 genes), and a plant (26,000 genes). This leads to the obvious question and the mission to sequence the largest and most complex genome yet, the human genome.

Slide 8. The first human genome was completed in 2001, and initially it was a race to be credited as the first team to accomplish this hallmark achievement. The teams under the direction of Drs. Craig Venter and Francis Collins joined forces to complete the sequence and share the credit. With the advent of DNA sequencing technology, it is believed that human genomes will be sequenced in less than 15 minutes at a cost of about $100. This means that every student in the classroom will have their own unique genome sequenced by their health care provider, and soon patients will discuss what does this all mean with their physician. The original sequence is thought to contain between 30,000 to 40,000 genes. Why such a discrepancy in the number of genes? The variability involves different results from the process and steps about how genes are predicted. This leads into the next slides that focus on how genes are predicted.

Slide 9. Regardless of the technology used, the outcomes of all genome sequencing projects are long strings of A's, C's, G's, and T's. So how are the genes predicted? At this point ask the students how they would find a gene in the sequence in the slide. What important patterns would they look for? This is the first process of genome annotation, called "structural annotation" or "gene finding/calling."

Slide 10. Here we define the two main steps involved in genome annotation. For "structural annotation" or "gene finding/gene calling," it would be too time consuming and laborious to find and predict every gene manually. With the advancements in bioinformatics tools, this process has become largely automated. Yet, even with the best computer-based prediction tools, typically in bacterial genomes, 50 to 90% of the "actual" genes can be found. Some of the first bacterial genomes undergo constant updates to find genes that were missed in initial attempts. This emphasizes that a person(s) will still be required to refine and finish these processes on genomes. Once a gene has been identified, the second step of genome annotation is to attach biological information to the protein that is encoded by the predicted gene.

Slide 11. Some of the patterns or criteria that are involved in structural annotation to identify gene boundaries are that the gene must have start and stop codons. This is often referred to as the open reading frame (ORF). Some of the other patterns utilized in predicting genes include sequences upstream of the predicted ORF. These upstream sequences are typically used as recognition sites in either the DNA or mRNA form (transcription and translation control, respectively).

Slide 12. Once a gene has been identified, how do we figure out what the encoded protein is? This is the process of attaching biological information as annotations to a gene. Some of the types of annotations that can be added to a gene are name, function, product, regulation, molecular interactions, cellular location, etc. As a starting point, focus on the annotations for gene names, products, and functions since the latter two are the types that students will annotate in the exercise. Provided are brief explanations of these annotation types. For further information about each annotation type and the process of adding annotations to the ASAP database, refer to the annotation guide found at https://asap.ahabs.wisc.edu /asap/home.php.

Slide 13. Hypothetical proteins are found in any newly sequenced genome. This slide helps to explain when to annotate a product as a hypothetical protein. This is an important subject that students need to understand—not every gene in a newly sequenced genome will have a known product. Even with the most studied microorganism, *E. coli* K-12, there are still a large percentage of genes with encoded proteins that we have no idea what they do or the role they play in the cell. The bacteriophage 933W genome will contain genes that will have to be annotated as hypothetical proteins. This presents an area for discussion. At the end of class, take a survey to find out how many students had genes that were annotated as

hypothetical proteins. This will allow the students to appreciate that genome annotation is not a simple straightforward procedure.

Slide 14. This is a picture of a GenBank file. Once annotations are added to a gene in a genome, all of this information is required to submit a genome file to the largest database of DNA and genomic sequence data at the National Center for Biotechnology Information. For example, every gene has to have a product annotation. Some of the other informative regions of the GenBank page are highlighted to aid in navigating files at NCBI such as organism, name, product, and function. This is important, because when BLAST searches return results, students will choose sequences that are similar to their respective query and this will lead them to other GenBank files to find information for adding annotations in the exercise.

Slide 15. The largest worldwide database of sequenced DNA and genomes is the National Center for Biotechnology Information (NCBI). Some of the resources that are important when annotating genes are tools such as BLAST and PubMed. BLAST is a powerful tool, because if we find a gene is similar to another characterized gene based on BLAST searching, we can hypothesize that what is true for the first gene, is most likely true for the newly sequenced gene product. PubMed allows searching in the archives of published journal articles for information that aids annotations.

Slide 16. This slide is meant to provide some brief general information about phages and also about the 933W phage genome that students will annotate. The animation illustrates the infection of *E. coli* by a phage and demonstrates some of the structural components of a phage, and more importantly, how the phage genome is injected and later integrated into the host genome. Some of the genes that students will annotate will encode structural components, replication proteins, and machinery for integration into the host cell. Bacteriophage 933W was found in the first sequenced genome of *E. coli* O157:H7 strain EDL933 (2).

Slide 17. This is a list of the tools that you and your class will use to annotate the 933W phage genome. For more information about the ERIC genome database, refer to the publication by Glasner et al. (1). The topic of annotations can easily become complex and extensive and there are numerous other tools and resources, but this exercise will focus on two fundamental tools, BLAST and InterProScan. These were selected with the mindset that if you only had 15 minutes to show tools for annotations, select the two that would allow the students to query and learn useful information for the most diverse annotation projects.

Slide 18. Provided are links to additional materials and resources for the ASAP database, NCBI, and Interpro. For simple explanations about using BLAST and InterProScan, refer to the book *Bioinformatics for Dummies,* chapters 6 (InterProScan) and 7 (BLAST).

B. Interactive component: bacteriophage genome annotation

The instructor should provide some basic background information about the genome that they are going to analyze. Viruses that infect bacteria are called bacteriophage, or phage for short. Some types of phage can sometimes insert their genome into the genome of a bacterium they infect, thus providing a mechanism for acquisition of new genes by bacteria. *Escherichia coli* O157:H7, the “bad hamburger” bug, has more than a dozen phage genomes inserted into its genome. One of these, named phage 933W, appears to have carried with it genes for making a nasty molecule called shiga toxin, a protein that causes some of the medical complications in people infected with *E. coli* O157:H7. In this class exercise, students are given the sequence of the 61,670 base pair 933W phage genome, the predicted boundaries of genes (the structural annotation), and instructions for using web-based sequence analysis tools to examine these genes. The students will employ these tools to infer the biological functions of genes and record their observations as annotations in a genome database.

The ASAP website will serve as the central resource for this portion of the exercise. The genome sequence of phage 933W and its predicted genes are presented to students in the ASAP database. The students are provided with a set of instructions that walk them through the process of investigating the potential function of a particular DNA genome segment and adding their findings to the ASAP database. To facilitate the exercise, the instructor should add annotations to the first coding sequence (CDS) in the genome. CDS regions, also called open reading frames (ORFs), are segments of DNA, usually around 1,000 nucleotides long, that begin with a start codon, end with a stop codon, and encode a potential protein. A primary goal of genome sequencing is often to infer the biological function of the hundreds or thousands of genes encoded by the genome. This typically begins by examining CDS's and comparing their DNA or predicted protein sequences to other sequences with known biological functions. During the demonstration, the instructor should familiarize the students with the steps necessary to access the sequence analysis tools students will use, such as following the links in ERIC to access the BLAST program. The instructor should also explain the BLAST result output format, that shows a list of other proteins from a database ("target" proteins) that match a protein used in a search (the “query” protein). The instructor should teach students how to interpret the scores that indicate the reliability of the matches. BLAST produces two scores for each matching target protein, a bit score and an e-value, the higher the bit score and the lower the e-value, the better the match. Show students the alignments of the sequences that were produced by BLAST and used to compute the scores. The instructor should follow the BLAST exercise with a demonstration of another useful sequence analysis tool called InterProScan. This tool also performs searches of proteins against a database, but tends to focus on matches to families of proteins and smaller portions of proteins that might represent domains or functional sites and may shed light on a particular protein’s function. At this point, the instructor may want to demonstrate searches of the PubMed database with keywords such as protein names or descriptions to identify publications in scientific journals that may contain further functional information about the protein.

Once logged in to the ASAP database and on the Query Genome Annotations page (Slide 4 in student instructions), obtain a list of all 78 CDS's in the 933W genome—select CDS as the feature type in the Basic Options section, at the bottom of the page in the View section type 78 into the area for Records per page, and click the submit bottom. This will provide the list of the 78 feature IDs. You can copy and paste the information into an Excel spreadsheet to subdivide the list of feature IDs for the class based on birthdays. The first CDS in the list is the *int* gene, which will be used as the example.

When you annotate the first CDS with the class, you should add information for the “product” and explain that the “evidence” that must be provided for each annotation is important, as it describes the method used to arrive at the conclusion.

Annotation table for *int*

Annotation type	Value	Evidence type	Reference
Gene	Int	Published annotation (supply GenPept Accession number)	NP_049461
Product	integrase	Published annotation (supply GenPept Accession number)	NP_049461
Function	cleaves DNA substrates	Protein sequence similarity (supply InterPro domain)	IPR002104
Function	site-specific recombination	Published annotation (supply GenPept Accession number)	NP_049461

After conducting the annotation of this feature with the class, give each student a single CDS to annotate. The students then are assigned a gene to work on, based on birthday month and further divided by weeks of the month if class size is large.

Closing class discussion. After a period of time, go through some of the CDS's that the students annotated and discuss them together as a group. Try to select some that lead to a known product and function and some that had to be annotated as hypothetical proteins. Discuss the nature and quality of evidence available to support each annotation. The discussion can begin as an instructor-led activity, starting with surveying the class with the first set of questions to encourage participation.

Questions for discussion with brief answers.

How many of you annotated the assigned CDS(s) as a hypothetical protein? How many of you were able to add more meaningful product and function annotations? More than 50% of CDS's in the phage genome have unknown products and functions. This illustrates that in a sequenced genome, whether it is from a phage or a microorganism, a large amount of the predicted CDS's still require experiments to determine information about the protein.

How do you know a gene is correctly annotated? There are multitudes of evidence types that could be used for adding annotations. The process of adding annotations ensures that the annotator verifies that the link to the evidence is accurate.

How could annotations be improved? Would you obtain the same results if you annotated the gene 1 year from now or 5 years from now? As more genes are characterized by traditional “bottom up” methods, the amount of information is bound to increase. For example, there will come a day when every gene in *E. coli* is characterized and understood, and consequently annotations would utilize experimental evidence rather than rely on evidence derived from homologous genes.

Common pitfalls encountered.

Some common questions that arose from students and reviewers during field testing are presented, and a response is provided to assist in explanation of the material.

Question 1: What do I do if InterProScan returns no results?

Answer: That means your protein doesn't have any significant matches in the database. Very likely that means the function is unknown. In this case, add the product annotation as a hypothetical protein. The student's instructions contain the step-by-step guide to add this type of annotation in the ASAP database.

Question 2: When the BLAST results are returned, why are there multiple entries that match at 100% identity with my sequence?

Answer: This is due to numerous entries in Genbank for the same gene, i.e., when the genome of the phage was sequenced (3), when the host genome for *E. coli* O157:H7 EDL933 was sequenced (2), and additionally, there has been another strain of *E. coli* O157:H7 (Sakai) sequenced to completion. So there are Genbank entries for the phage genome and the whole bacterial genome. It is likely that the BLAST search will return these four matches with 100% identity. Any of the four files are appropriate for adding as the evidence for annotations, it is up to the students to decide which to use. Students may also be interested in looking at whether there are any substantive differences in the annotation of those redundant entries.

Question 3: Is the first hit on the BLAST or InterProScan list the right one? How do you make the call of one function versus another?

Answer: This question reveals an assumption that many students make, that there is only one correct answer when annotating. This module is designed to show that there are numerous tools and methods for finding information to add biological annotations, and there is a degree of uncertainty in basing annotations on matches in databases. The goal of turning students into annotators is that they learn to use these tools and engage in the decision-making processes required to analyze genomic sequences.

Question 4: Do students ever disagree on an annotation? How is this resolved?

Answer: In this module, each student annotates his or her own gene. In the exception that more than 78 students are in a class or have similar birthdays, there will be multiple students adding annotations to the same gene. It has been our experience that the level of matching similarity from the BLAST search results does not lead to much discrepancy. The database permits addition of multiple entries of annotations, so if multiple students add products and functions, that is fine. The disagreements are what make the process interesting.

Question 5: Will students see the curator approval status—that is, rejected annotations?

Answer: This exercise is set up so that the genome that the class is working on is a "private" genome, which means that only the instructor and the class will be able to see the associated annotations. Further, since this is a previously published genome (3), it does not have to go through the rigorous steps involved in curation of the genome for GenBank upload. This module is really designed to teach the students about how to use the tools and actually go about the process of annotating a gene of a genome. To teach a class on annotation of an newly sequenced genome would require much more time, and it is not within the scope of the learning objectives of this module.

Question 6: How long is one class period?

Answer: During field testing, each class period was 50 minutes. We feel that the exercise can be accomplished in this amount of time. If the class does not finish, students can finish the annotations on their own, and the next class period can be used for discussion and follow-up activities.

Field Testing.
This exercise has been used twice at the University of Wisconsin—Madison, the first time in an introductory biology course with 165 students (Biology 151) and the second in an upper-level microbiology course with 15 students (Bacteriology 650: Advanced Foodborne Pathogens). Since the initial review, this exercise was used a third time with an upper-level microbiology lab course with 12 students at the University of Wisconsin—Milwaukee (Biology 580: Experimental Microbiology). A screen shot of a correctly annotated gene generated by a student is provided in Fig. 1.

FIG. 1. Screen shot of a correctly annotated gene generated by a student in Biology 580 at UW—Milwaukee.

Formal Assessment Methods.

For the students in Biology 151 and Bacteriology 650, the first set of pre- and posttest questions was used to assess student learning. The following questions were posed to the class, and results are provided. Following reviewer's suggestions, many of the questions were rewritten and used to assess learning in an upper-level microbiology lab course at UW—Milwaukee. Overall, the expectations for an introductory biology course are quite different from those for an upper-level course composed of microbiology majors. Therefore, we have provided the results from both scenarios and conducted statistical analysis using the chi-square test when applicable. Additionally, to test if learning objectives were achieved, the percentage of correct product and function annotations were assessed from the microbiology lab course at

UW—Milwaukee. It was determined that 83% and 89% of the annotations were added correctly for product and function annotations, respectively.

Pre- and Posttest questions set #1

1) Which of the following was the first genome sequenced?

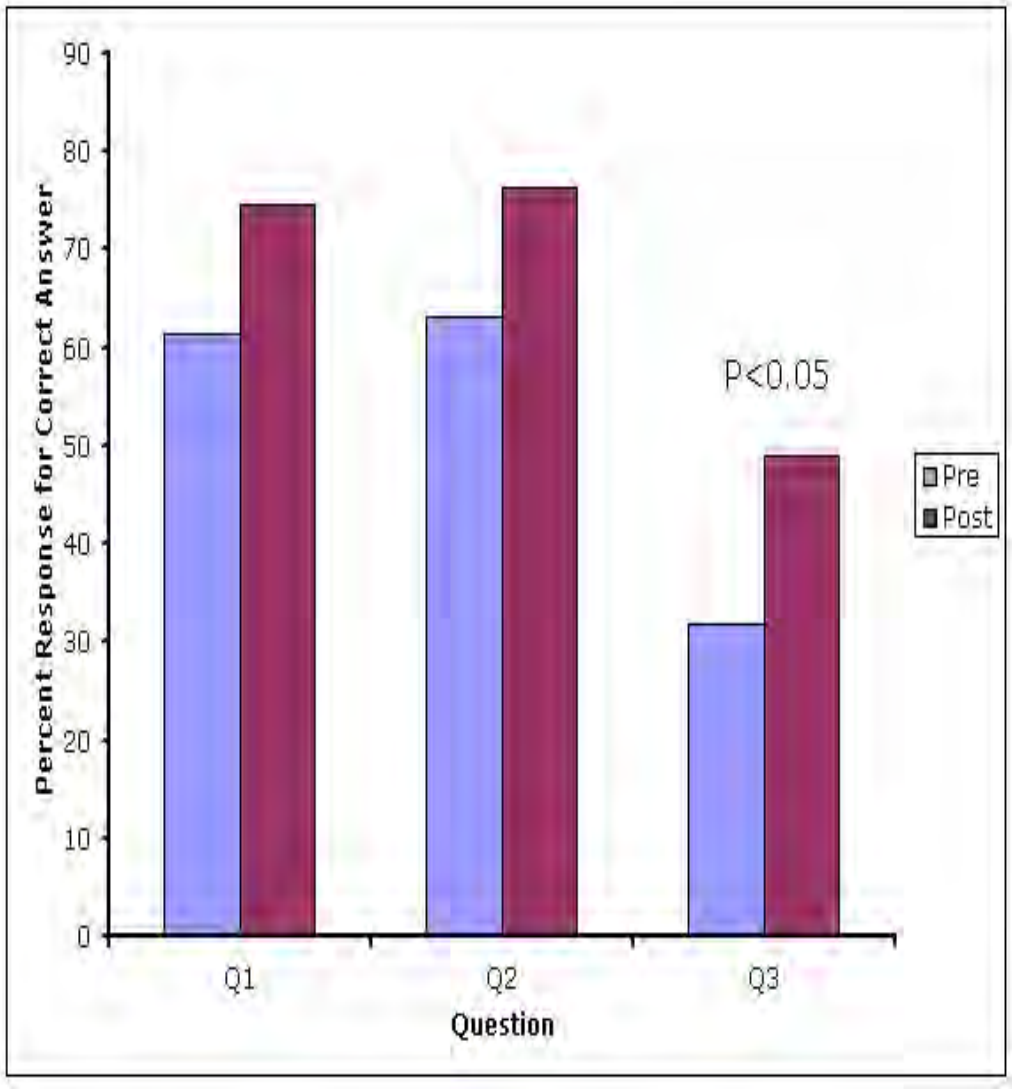
A) bacteriophage
B) bacterial
C) a plant
D) human

2) Which of these web-based resources is useful to find biological information about a gene sequence?

A) BLAST
B) InterProScan
C) Pubmed
D) All of the above
E) None of the above

3) Annotation of a gene in a genome database includes the following steps, except:

A) Identifying an open reading frame
B) Locating regulatory elements
C) Attaching biological information to each identified gene element
D) Conducting experiments to confirm information obtained through database searches



Applying the chi-square test using the pretest answers as expected range and the posttest results as actual range, statistical analysis revealed a significant difference for question #3.

Pre- and Posttest questions set #2

1. Within a sequenced microbial genome, a gene predicted to encode a protein should contain which of the following

characteristics?

A) a region of DNA upstream that corresponds to the promoter binding of RNA polymerase

B) a region of sequence that corresponds to a start codon (ATG)

C) a region of sequence that corresponds to a stop codon (TGA, TAG, or TAA) in the same reading frame as the start codon

D) all of the above

2. What percentage of the protein coding genes do you think automated computer approaches applied to a newly sequenced microbial genome will find?

A) 100% of the actual protein coding genes

B) more than 50% of the actual protein coding genes

C) less than 10% of the actual protein coding genes

3. What type of biological annotation cannot be assigned to a newly sequenced gene based solely on comparisons to known proteins or genes?

A) product

B) function

C) three-dimensional structure

D) name or synonym

4. In a newly sequenced microbial genome, every identified gene produces a protein that is similar to a known protein? True or **false**?

5. Which of these web-based resources are useful to find biological information about a gene sequence?

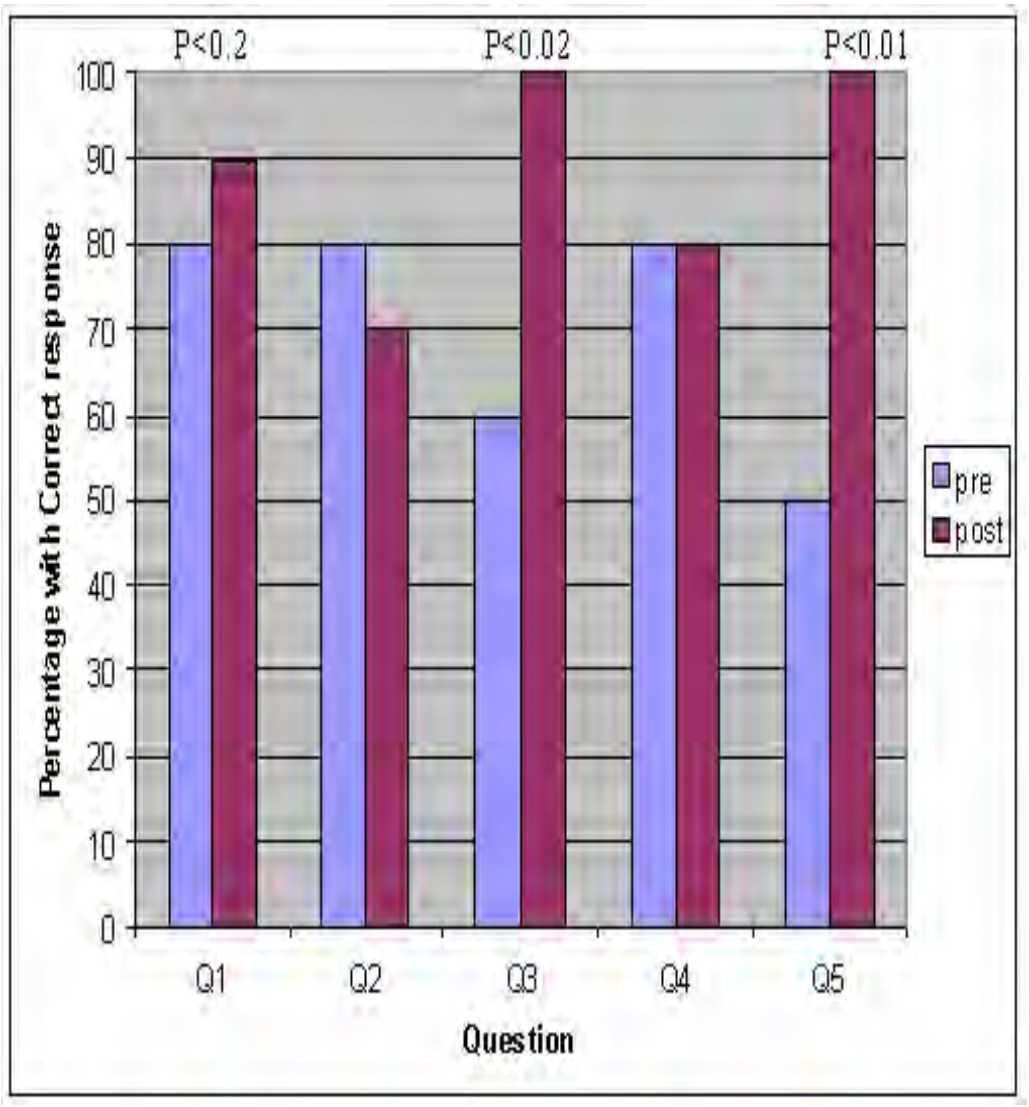
A) BLAST

B) InterProScan

C) Pubmed

D) All of the above

E) None of the above



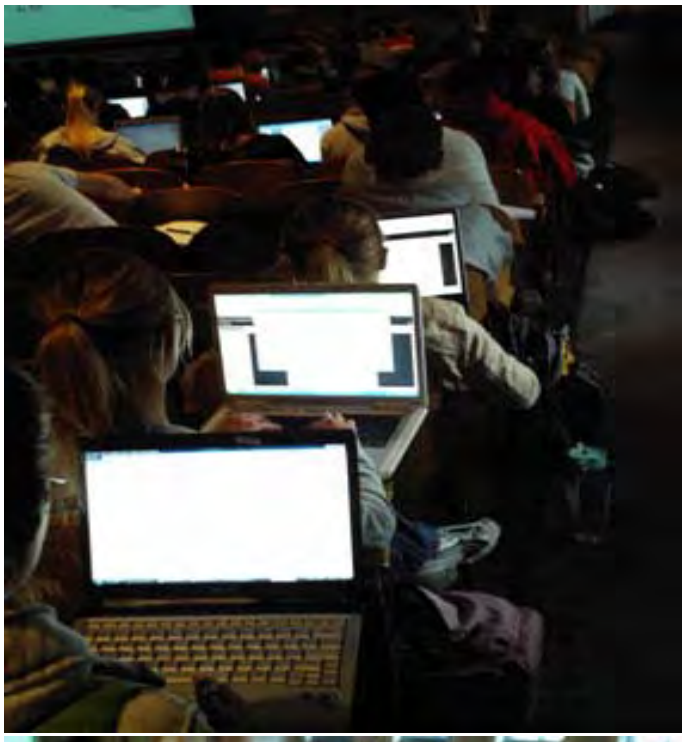
Applying the chi-square test using the pretest answers as expected range and the posttest results as actual range, statistical analysis revealed a significant difference for questions 1, 3, and 5.

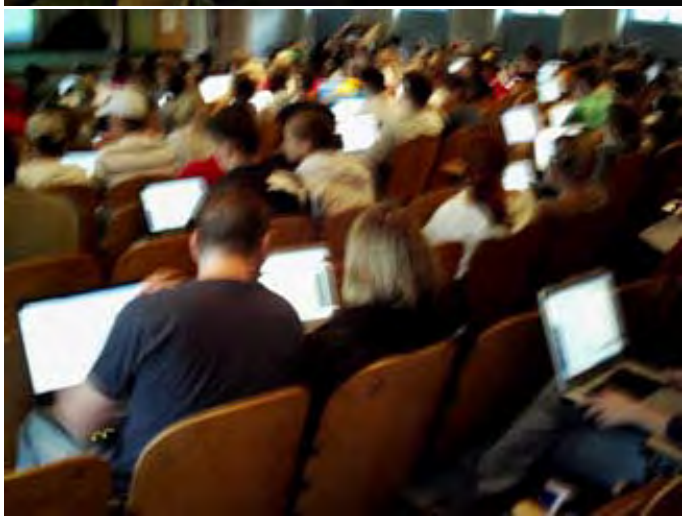
Student testimonials and pictures of students.

"I really enjoyed learning more about bacterial genetics and the tools that are available online for genomic research and gene identification. This is an area of bacteriology that I have little experience in and I think that having experience using these websites will prove valuable as my research continues." –UW—Madison student in Bacteriology 650

"The concepts of using BLAST and InterProScan are pretty neat, and it is great that anyone can access this information, not just the insider scientists that put it together. Thank you for teaching our class how to use these tools! I doubt I would have ever learned this stuff on my own had you not taught us." –UW—Madison student in Bacteriology 650

"It's really helpful to annotate genes, submit references into function and add new interactions for known proteins. It's also great to have a hands-on project instead of learning all of it in theory. I think it's really important to make sure all annotators do follow the same standard, so that product and function annotations are appropriate."
–UW—Milwaukee student in Biology 580





Pictures of Introductory Biology students at UW—Madison annotating the bacteriophage 933W genome using wireless-ready laptop computers in a lecture hall equipped with wireless Internet.

Possible Modifications.

Applying concepts covered by annotation to more advanced topics:

Once students have learned what annotations are and how the information is derived, and are familiar with using the ERIC database, inquiries into any of the other *Enterobacteria* genomes such as *Salmonella* spp., *E. coli*, *Yersinia* spp., *Shigella* spp., or various plant pathogens are possible. Since all members of the *Enterobacteria* are thought to have descended from a common ancestor, concepts can apply to more advanced topics including evolution, biochemical pathways, gene-protein relationships, or comparative genomics. Additionally, more modules have been designed that utilize the ERIC database and the topic of comparative genomics for inquiry-based learning and were presented at the ASM Conference for Undergraduate Educators 2008 meeting. The slides are available for an overview (The Genomics Era: A Vast Resource for Educators by Dave Baumler http://www.asmcue.org/page02d.shtml) and contain information to integrate these more advanced topics into your course(s).

Tailoring the exercise to a smaller class:

An alternative approach for using this exercise in a smaller class is to assign more than one CDS to each student. If the students can't finish the annotations of multiple CDS's in one class, they can finish the work at home, and the class discussion can occur during the next class period. As a result of annotating the whole genome as a class, the discussion will be more interesting since the students will be able to analyze the genome in its entirety.

Tailoring the exercise to an upper-level biochemistry or microbiology class:

This exercise, if used with an upper-level class, can be extended with an additional individual project for students based on the annotation of the genome. The upper-level students could be asked to write a report or prepare a presentation. Some questions to ask include:
1) What did you learn from this activity?
2) Did your annotation agree with others who had the same CDS? If not, explain why this may have happened and which annotation appears to be more accurate?
3) What does the putative protein do?
4) What role does the putative protein play in the bacteriophage? If the CDS was deleted, what affect would this have on the phage and the host? Does the CDS play a role in the virulence of *E. coli* O157:H7?

References.

1. **Glasner, J. D., M. Rusch, P. Liss, G. Plunkett III, E. L. Cabot, A. Darling, B. D. Anderson, P. Infield-Harm, M. C. Gilson, and N. T. Perna. 2006. ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. Nucleic Acids Res. 34: D41-D45.**

2. **Perna, N. T., G. Plunkett III, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Pósfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner**. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature **409:**529–533.

3. **Plunkett III, G., D. J. Rose, T. J. Durfee, and F. R. Blattner.** 1999. Sequence of shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: shiga toxin as a phage late-gene product. J. Bacteriol. **181:**1767–1778.

Answer key:

Pre- and Posttest set #1: 1) A. 2) D. 3) D.

Pre- and Posttest set #2: 1) D. 2) B. 3) C. 4) False. 5) D.

Acknowledgments.

We thank Guy Plunkett III for contributions to this work. The ERIC Bioinformatics Resource Center has been funded with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract number HHSN266200400040C.

Welcome to the ASAP genome database (A systematic annotation package for community analysis of genomes).

Using your web browser,

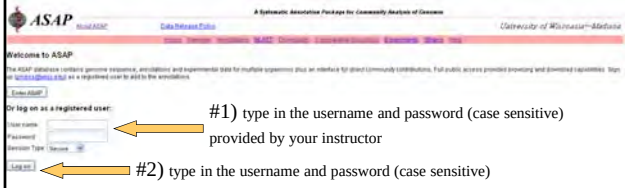
#1) go to https://asap.ahabs.wisc.edu/asap/home.php

#2) in the upper right portion of the screen click on Log on

log on as a registered user.

Then type in the username and password (case sensitive) provided by your instructor and click the log on button.

Note your class has been given access to a unique version of the genome, in which you and your fellow classmates will be the only people annotating the phage genome



#1) type in the username and password (case sensitive) provided by your instructor

#2) type in the username and password (case sensitive) provided by your instructor

2

Click on Annotations

#1

Then use the pull down bar to select bacteriophage 933W

then click the OK button

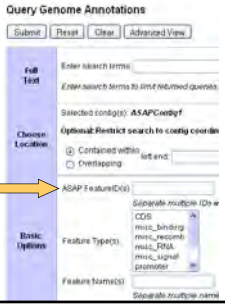
#2

Every gene in a genome in the ASAP database has what we call a feature ID, which consists of three capitol letters a dash and seven numbers For example ABC-1234567

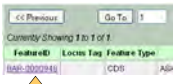
Your genome will have a unique 3 letter code and each gene or coding sequence (CDS) will have a unique seven digit number. Your course instructor will provide you with the feature IDs for your class

To start as a group you will annotate the gene called "int", so type the feature ID for this gene into the box in the Basic Options section that says ASAP Feature ID(s), then click the Submit button

#1



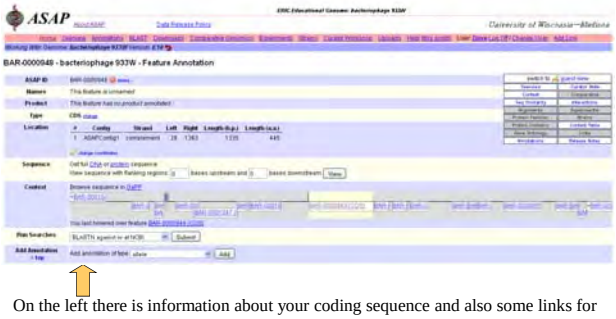
On the next page, click on the link for the feature ID



#2

4

Your webpage should look like this



On the left there is information about your coding sequence and also some links for tools you will be using

Further down is the Add annotations, this is where you will be adding annotations (cont. on next slide)

5

Example of an annotation page for the gene "int"

Here you will see the three things that are important for the annotation, I) the annotation type (name, product, or function), II) the data, and III) the evidence. In this example the function and product has been added by another annotator, in your page you should only see the name annotation type.

Definitions for Product: the primary name of the protein/enzyme

Function: The biological role that the protein performs, usually beginning with a verb (Example: cleave DNA substrates)

6

To find information about your gene (CDS) that you are annotating, there are two approaches you will use:

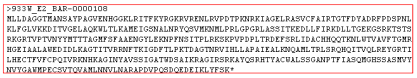
#1) Use Interproscan to search a database of protein families and domains which contain identifiable features from known proteins that can be applied to new protein sequences

#2) Use BLASTP to search all genes and genomes at NCBI

7

You will need the protein sequence to copy and paste into the Interproscan website

Click Protein sequence, and a window will open in your web browser containing the protein sequence, highlight the sequence (starts with a M and ends with a *) and select copy



8

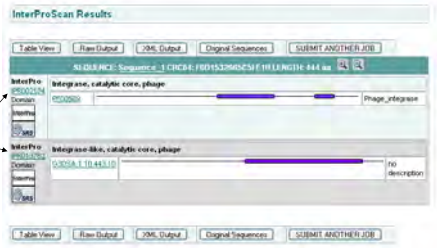
Start by adding a new tab to your web browser and go to http://www.ebi.ac.uk/InterProScan/

Paste your protein sequence into the box, and click submit, and wait for the results

9

Your Interproscan will look like this, it has found similar protein matches from two different protein databases, do they agree?

Click on the InterPro link to view that page with information about the protein match



If your Interproscan returns no results proceed to slide #18

10

There is information that we can add for product & function

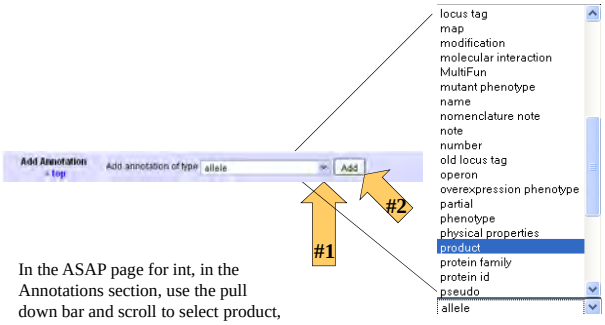
For product annotation, you will need to copy integrase, and also the Interpro ID in this case IPR002104

You will need to search for a function in the InterPro annotation section, (i.e. cleave DNA substrates) and also the Interpro ID in this case IPR002104

Lets add go back to ASAP and add a product annotation

11

In the ASAP page for int, in the Annotations section, use the pull down bar and scroll to select product, then click on the Add button



12

#1) Paste the data, in this case integrase

Then select the type of evidence from the pull down bar,

in this case select: protein sequence similarity (supply Interpro domain)

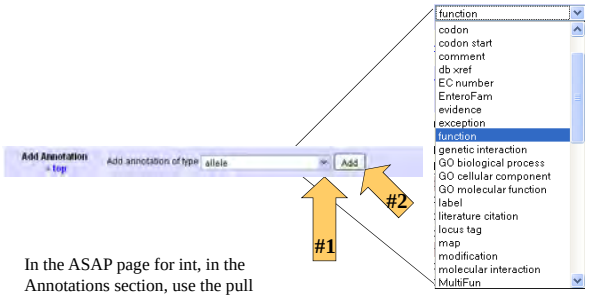
In the box below paste or type in the Interpro ID, in this case IPR002104

Then click Add Annotation

You will be prompted to this screen,

Click here to verify that it leads to the correct webpage in Interpro, if it does, click Yes

Congratulations, you have added a product annotation, and should see it appear in the annotation page for int in the bacteriophage 933W genome in the ASAP database, now try to add a function annotation from the Interpro results.



In the ASAP page for int, in the Annotations section, use the pull down bar and scroll to select function, then click on the Add button

#1) Paste the data, in this case cleave DNA substrates

Then select the type of evidence from the pull down bar,

in this case select: protein sequence similarity (supply Interpro domain)

In the box below paste or type in the Interpro ID, in this case IPR002104

Then click on Add Annotation

You will be prompted to this screen,

Click here to verify that it leads to the correct webpage in Interpro, if it does, click Yes

Congratulations, you have added a function annotation, and should see it appear in the annotation page for int in the bacteriophage 933W genome in the ASAP database.

If Interproscan returns no results

-If there is no good match, it is called a hypothetical protein

-add an annotation for product as hypothetical protein

-use Unpublished Sequence analysis as Evidence

-type in author name, email

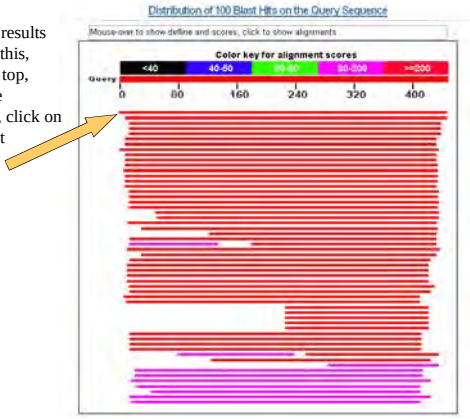
-click Add Annotation

Now lets try using BLASTP to search for similar sequences, on the annotation page for int in the bacteriophage 933W genome in the ASAP database,

all you have to do is select BLASTP against nr at NCBI from the pull down

You will then be directed to a page at NCBI running your BLAST search click on the button that says View Report, (it may take a few moments)

Your BLAST results will look like this, starting at the top, scroll over the matching hits, click on one of the best matches

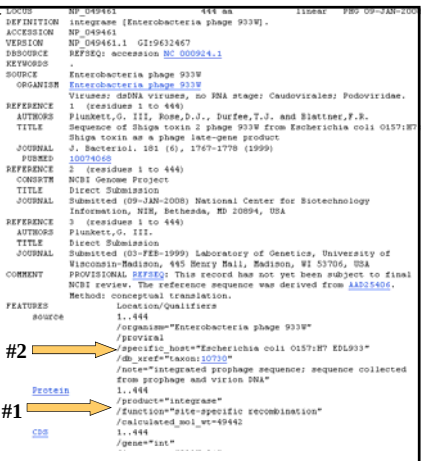


As you see your sequence matched these with 100% identity, click on the first entry to view the GenBank file

Note: Your BLAST match will not always be 100%, so it is a good idea to look at a couple of the matches and decide which to use for deriving annotations

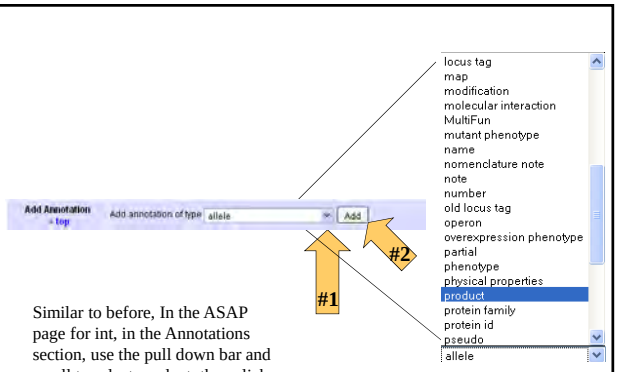
#3

On the GenBank page, the information that you are interested in is the product & function (#1), specific host (#2), and also the locus ID number (#3)



#2

#1



Similar to before, In the ASAP page for int, in the Annotations section, use the pull down bar and scroll to select product, then click on the Add button

23

#1) Paste the data, in this case integrase

Then select the type of evidence from the pull down bar,

in this case select: Published annotation (supply GenPept Accession number)

In the box below it paste or type in the GenPept ID, in this case NP_049461

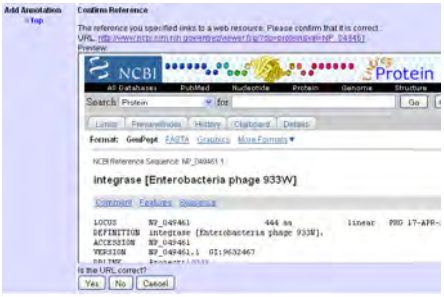
In the lower box that says enter an organism to associate with this annotation, paste or type in Escherichia coli O157:H7 EDL933

Then click Add Annotation

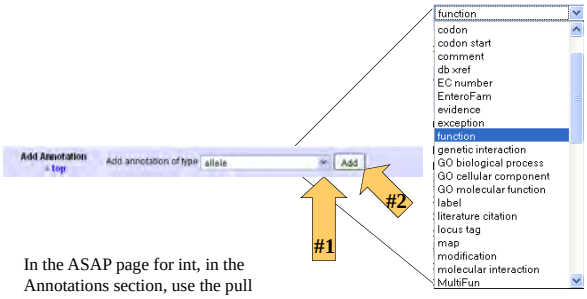
24

You will be prompted to this screen,

Verify that it leads to the correct webpage in GenBank, if it does, click Yes

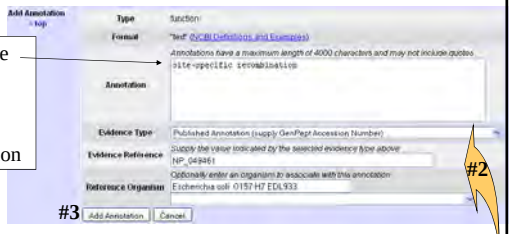


Congratulations, you have added a product annotation, and should see it appear in the annotation page for int in the bacteriophage 933W genome in the ASAP database, now try to add a function annotation from the BLAST results.



In the ASAP page for int, in the Annotations section, use the pull down bar and scroll to select function, then click on the Add button

#1) Paste the data, in this case site-specific recombination



Then select the type of evidence from the pull down bar,

in this case select: Published annotation (supply GenPept Accession number)

In the box below it paste or type in the GenPept ID, in this case NP_049461

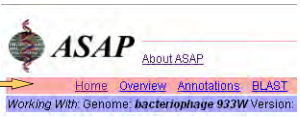
In the lower box that says enter an organism to associate with this annotation, paste or type in Escherichia coli O157:H7 EDL933

Then click Add Annotation

Now its time for your own gene (CDS) to annotate

Now you will be provided with a feature ID for another gene in the bacteriophage 933W genome

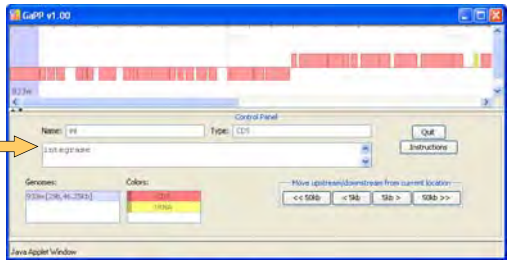
Click on the home link, and proceed to your annotation page for your gene as you did in slide #3



follow the same steps as before to run a Interproscan and also a BLASTP search and add product and function annotations for your gene

Once you have completed your annotations for you gene(s), you can view the genome of the phage and see how your fellow classmates are doing by clicking on “Browse sequence in Gapp”

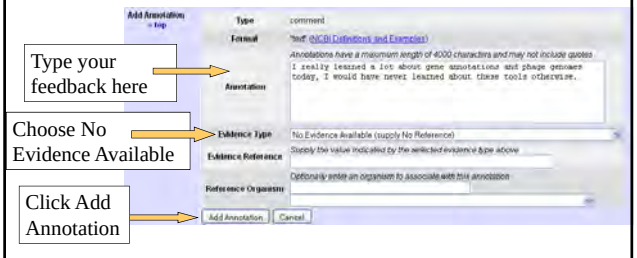
A new window will appear in a few seconds,



The gene you are working on is highlighted in blue, and you are visualizing the entire Bacteriophage 933W genome, scroll over each gene (in pink) and you should see the name and the product information provided in the boxes below the genome, also double click any of the genes, and your web-browser will open the annotation page in ASAP and you can view the function annotation, etc.

Congratulations, you are officially an annotator!!

The ASAP database thanks you and welcomes your feedback, you can add an annotation type called a comment in the page for the gene in ASAP, here you can provide use with useful feedback and constructive criticism.



Annotation of the bacteriophage 933W genome: an in-class interactive web-based exercise

Genome: The entire collection of genetic information of an organism

- Everybody has a genome
  - Viruses, bacteria, archaea, eukaryotes (fungi, plants & animals)
  - They can range from thousands to billions of base pairs (bp) of DNA (some viruses have RNA genomes)
  - Genomes can consist of one or many chromosomes
  - Chromosomes can be linear or circular

History of DNA sequencing and genome research

-Methods for determining the sequence of DNA were developed in the early 1970's.

-Frederick Sanger and colleagues determine the first complete genome sequence of all 5,375 nucleotides of bacteriophage ϕX174 (sequence completed in 1977, Nobel prize awarded in 1980).

10 genes

[Sanger F. *et al*., Nature 265, 687-695 (1977)]

History of DNA sequencing and genome research (cont.)

Sanger developed a method called "shotgun" sequencing and completed the 48,502 bp genome of bacteriophage lambda in 1982

This method allows sequencing projects to proceed much faster and is still commonly used.

A map of the lambda genome

[Sanger F. *et al.* J. Mol. Biol. 162, 729-773 (1982)]

1995 *Haemophilus influenza* sequenced

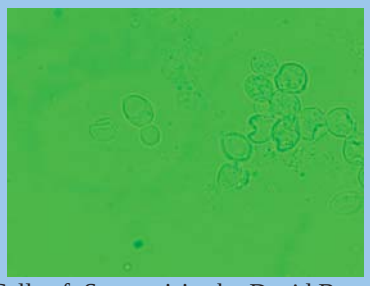
- Craig Venter and colleagues at The Institute for Genomic Research (TIGR) reported the first complete genome sequence of a (nonviral) organism, *Haemophilus influenza*.
- used shotgun sequencing
- assembled ~24,000 DNA fragments into the whole genome using the "TIGR assembler" software
- 1,830,138 bp genome
- 13 months to sequence

1,709 genes

1996 Yeast Genome Sequenced

- ***Saccharomyces cerevisiae* (ale yeast)**
- The yeast genome sequence was completed by an international consortium (74 labs) started in 1989.
- 16 chromosomes, 12,070,900 bp.

~6,269 genes



Cells of *S. cerevisiae* by David Baumler

Other eukaryotic genomes

Increase in genome size

Drosophila melanogaster,

Fruit fly ~13,000 genes, completed 2000

Caenorhabditis elegans

Nematode ~19,000 genes, completed 1998

Arabidopsis thaliana
(plant) ~26,000 genes, completed 2000

Humans???

The Human genome

~30-40,000 genes

- 1999 First human chromosome sequenced
- 2001 Human genome "completed"
- 23 chromosomes (haploid genome), 3,038,000,000 bp
- Francis Collins and Craig Venter

- September 2007, Venter publishes the sequence of his own diploid genome
- Venter announces plan to sequence 10,000 human genomes in 10 years
- in the future $100 human genomes

Part of a genome sequence

TCAGCGAAGATGAGATAGTTTTTAAAGGTGGGATTTCCCCACCTTTAAAAAGCGAGAAGTCCCGGTTTTAA
AGAGGAGTAAAATCCTCTTTTTCTAGCCCACTCAGGTGGTTTTTTTGGTTTTCGCTCCTTGCCGCATCTTC
TGTGCCTTTGATGGCGGCTGGTTGGGGTGAAAGGCTGCATATTCCAGAATTTCAGACAGTAGATTGTTTTT
GAAATCTTCCGTTTTATCGTTGACGAACTTAACCATCCTGTTGAAATCATCTTCCTTTGATACACCTTCAG
GAAATGCCTTAGGAACTGATGTTTGGCTATCCAAGGCATCTTGCAATATCTGCACGATCTCCGAATTCATT
GATCGCCCATTGGCCTTTGCTCTGGCGGCAACTGCGTCACGCATACCGTCAGGCATCCTAACTGTAAATCT
CTCAATGAAAGCTGGATCTTCTTTTTCAGTCATCATCTTAAACCATAAAAATTTATACAAAACACACTAGC
ATCATATTGACATTACCCACAATGACATCATAATGGTGTCAGGCATCAAAATGATGTCATCATGACAAGGG
GAAAGTAAATGCAAGATGTTCTCTATACAGGTCGTAAGAACGACAGCTTTCAGCTTCGTCTGCCTGAGCGA
ATGAAAGAAGAGATCCGTCGCATGGCAGAGATGGACGGCATTTCGATTAATTCTGCAATCGTGCAGCGCCT
TGCTAAAAGCTTGCGTGAGGAAAGAGTTAATGGGCAGTAAAAACAGCGAAGCCCGGAAGTGTGGGGACACT
AACCGGGCTTCTAATGTCAGTTACCTAGCGGGAAACCAACAATGACCAGTATAGCAATCTTTGAAGCAGTA
AACACTATCTCTCTTCCATTCCACGGACAGAAGATCATAACTGCGATGGTGGCGGGTGTGGCGTATGTGGC
AATGAAGCCCATCGTGGAAAACATCGGTTTAGACTGGAAGAGCCAGTATGCCAAGCTCGTTAGTCAGCGTG
AAAAGTTCGGGTGTGGTGATATCACCATACCTACCAAAGGTGGTGTTCAGCAGATGCTTTGCATCCCTTTG
AAGAAACTGAATGGATGGCTCTTCAGCATTAACCCAGCAAAAGTACGTGATGCAGTTCGTGAAGGTTTAAT
TCGCTATCAAGAAGAGTGTTTTACAGCTTTGCACGATTACTGGAGCAAAGGTGTTGCAACGAATCCCCGGA
CACCGAAGAAACAGGAAGACAAAAAGTCACGCTATCACGTTCGCGTTATTGTCTATGACAACCTGTTTGGT
GGATGCGTTGAATTTCAGGGGCGTGCGGATACGTTTCGGGGGATTGCATCGGGTGTAGCAACCGATATGGG
ATTTAAGCCAACAGGATTTATCGAGCAGCCTTACGCTGTTGAAAAAATGAGGAAGGTCTACTGATTGGCGT
ATTGGAAGGCGCAAAAAGAAAAGCCAGCAGATGGGCTGCTGGCATTCATTGGGTATATGAACTTTCGGAGA
ACATATGAAGTCAATTATCAAGCATTTTGAGTTTAAGTCAAGTGAAGGGCATGTAGTGAGCCTTGAGGCTG
CAAGCTTTAAAGGCAAGCCAGTTTTTTTAGCAATTGATTTGGCTAAGGCTCTCGGGTACTCAAATCCGTCA

What exactly are annotations?

Genome annotation is the process of attaching biological information to sequences. It consists of two main steps:

1.-identifying elements on the genome, a process called "structural annotation" or "gene finding". Today much of this is automated with computers, yet ~50-90% of the actual genes can be predicted, still requires a person(s) to finish predicting them all.

2.-attaching information to these elements such as their molecular and biological functions.

Annotation step #1: Structural Annotation

The genetic code (Courtesy of the National Institutes of Health)

Example of a gene - the start codon is green and the stop codon is red

Structural annotation consists of the identification of genomic elements (e.g. genes).

•Open Reading Frames (ORFs) also called coding sequences (CDSs) must have a start codon and a stop codon

•location of regulatory motifs (such as promoters and ribosome binding sites)

•This step is typically automated using gene prediction software

Annotation step #2

Functional annotation: consists in attaching biological information to genomic elements.

- •biochemical function
- •involved regulation and interactions
- •expression
- •cellular location

Three examples of annotations for one gene:

•**Name/synonym**: a short "word" used to refer to the gene (Ex. *ureC*)

•**Product**: a descriptive protein name (Ex. Urease gamma subunit)

•**Function** : Describes what the protein does (Ex. Catalyzes the hydrolysis of urea to form ammonia and carbon dioxide)

When is the gene product a “Hypothetical Protein”

- When a gene is identified, but the predicted protein sequence doesn’t have an analog in protein database(s) (for example the search in Interproscan returns no result)
- A protein whose existence is predicted, but there is no evidence that it is expressed in vivo
- These are called hypothetical protein, and are added as product annotations
- Sometimes the function of a hypothetical protein can be predicted by searching for domains in a protein database, often though they are annotated as function “unknown”
- Even in the genome of “the most studied microorganism”, non-pathogenic *E. coli* K-12, ~30% of the genes are annotated as hypothetical proteins.

A genbank file

Organism from which the sequence was characterized

List of annotated features

Structural annotation

Name of the gene (*ureC*)

Product

Function

How is all of the sequence data stored and what do we do with it?

National Center for Biotechnology Information (NCBI)

What tools are there to use?

#1 BLAST – search for similar sequences

#2 PubMed – search for related literature

Where are all of the genome sequences?

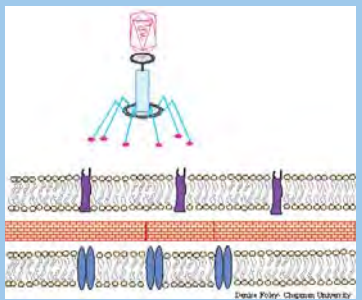
(http://www.ncbi.nlm.nih.gov/)

We are going to annotate a phage genome today

What type of genes should we anticipate finding in the phage genome?

- Structural components of a phage
- Phage replication proteins
- Machinery for integration into the host genome

You are going to annotate the bacteriophage 933W genome. This phage was found in the genome of *E. coli* O157:H7. The phage genome contains the genes *stx2A* and *stx2B* that encode the shiga toxin 2 protein, that contributes to disease in humans.



Animation Courtesy of Microbelibrary.org

Tools you will use to annotate 933W

- #1 ASAP database: this is where you will get the sequences and record your functional annotations.
- #2 BLAST: this is a tool you will use to find similar sequences in the NCBI database of all publicly available known and predicted proteins
- #3 InterproScan: this is a tool you will use to find similar sequences in a database of protein families (groups of related proteins) and domains (functionally significant subregions of proteins)

Links to additional resources

ASAP – A Systematic Annotation Package for Community Analysis of Genomes
- Home page: https://asap.ahabs.wisc.edu/asap/home.php
- Annotation guide: https://asap.ahabs.wisc.edu/asap/home.php

NCBI – National Center for Biotechnology Information
- Home page: www.ncbi.nlm.nih.gov
- BLAST home: www.ncbi.nlm.nih.gov/blast/Blast.cgi
- BLAST guide: www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch16

Interpro – a database of protein families and domains
- Home page: www.ebi.ac.uk/interpro
- Manual: www.ebi.ac.uk/interpro/user_manual.html
- InterproScan: www.ebi.ac.uk/InterProScan

For additional information on using Blast and Interproscan, we recommend the book "Bioinformatics for Dummies"