Spring 2013 Genomics Exam #1
Genomic Sequences

There is no time limit on this test, though I don't want you to spend too much time on it. I have tried to design an exam that will take less time that exams in the past. You do not need to read any additional papers other than the ones I send to you. There are 4 pages, including this cover sheet, for this test. There are no Discovery Questions on this exam. You are not allowed discuss the test with anyone until all exams are turned in at 10:30 am on Wednesday February 13. **ELECTRONIC COPIES OF YOUR EXAM ANSWERS ARE DUE AT 10:30 am ON WEDNESDAY FEBRUARY 13**. You may use a calculator, a ruler, your notes, the book, and the internet. You may work on this exam in as many blocks of time as you want. Submit your electronic version before 10:30 am (eastern time zone).

The **answers to the questions must be typed in a Word file and emailed to me as an attachment**. Be sure to backup your test answers just in case (I suggest a thumb drive or other removable medium). You will need to capture screen images as a part of your answers which you may do without seeking permission since your test answers will not be in the public domain. Remember to explain your thoughts in *your* own words and use screen shots to support your answers. Screen shots without ***your*** words are worth very few points. Support your answers with data using screen shots liberally.

DO NOT READ or DOWNLOAD ANY NEW PAPERS FOR THIS EXAM. You may search and read abstracts. RELY ON YOUR EXPERIENCE, AND YOUR SKILLS. Spell out your logic for each answer.

-3 pts if you do not follow this direction.
Please do not write or type your name on any page other than this cover page.
Staple all your pages (INCLUDING THE TEST PAGES) together when finished with the exam.

Name (please type):

average grade = 93%
Write out the full pledge and sign (electronic signature is ideal):
"On my honor I have neither given nor received unauthorized information regarding this work, I have followed and will continue to observe all regulations regarding it, and I am unaware of any violation of the Honor Code by others."

How long did this exam take you to complete?

~13 hours average

20 pts

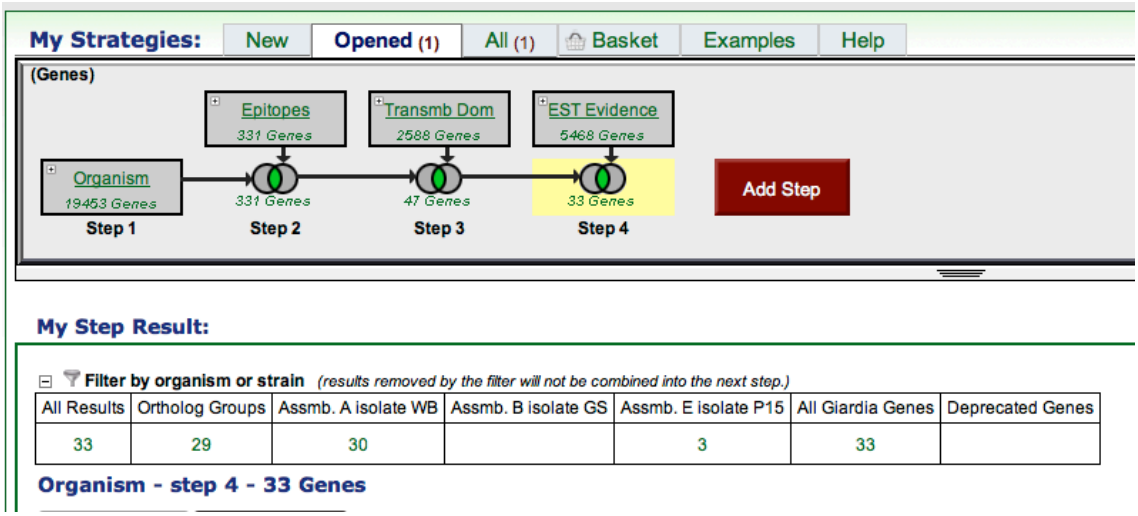
1) I want you to use a database you have never seen before called EuPathDB (http://eupathdb.org/eupathdb/). Use the Giardia portion of this integrated database. We will not use the full power of what this site can do, but you will get a sense of its potential as you work on this problem.

Your task is to identify a protein target for a drug to be developed by a company called Mayking Itup, LLC. You will have to figure out how to use EuPathDB to answer most of these questions.

a) What is Giardia and what sort of disease does it cause? Support your answer by providing the URL of your information source(s). **Limit your answer to 3 sentences or less.**

eukaryote, protest, parasite, diarrhea – many possible sites

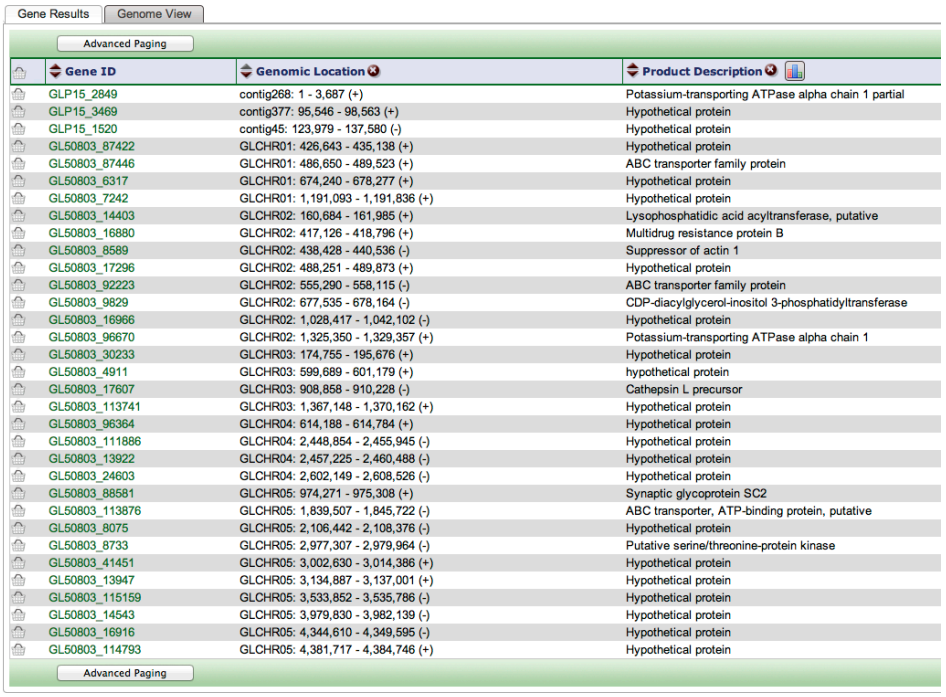
b) Identify a set of proteins whose features include a known epitope and it is an integral membrane protein. You also must be certain that the protein is expressed by using EST data. How many proteins are in this set of genes/proteins? Provide a screen shot to support your answer.



c) Choose one gene/protein from your list above that you would like to inhibit based on its biological process. Name that gene by its common name and its DB accession number.

multidrug resistance protein B

XM_001708156



d) How many transmembrane domains are predicted for your chosen protein? How many amino acids in this protein? Support your answer with data.
14 TM domains
556 amino acids

e) Find another way to independently confirm via computer (prediction) whether the number of transmembrane domains you found for part (d) is correct or not. Support your answer with data.
Kyte-Doolittle or other site is fine. Interpreted screen shot required.

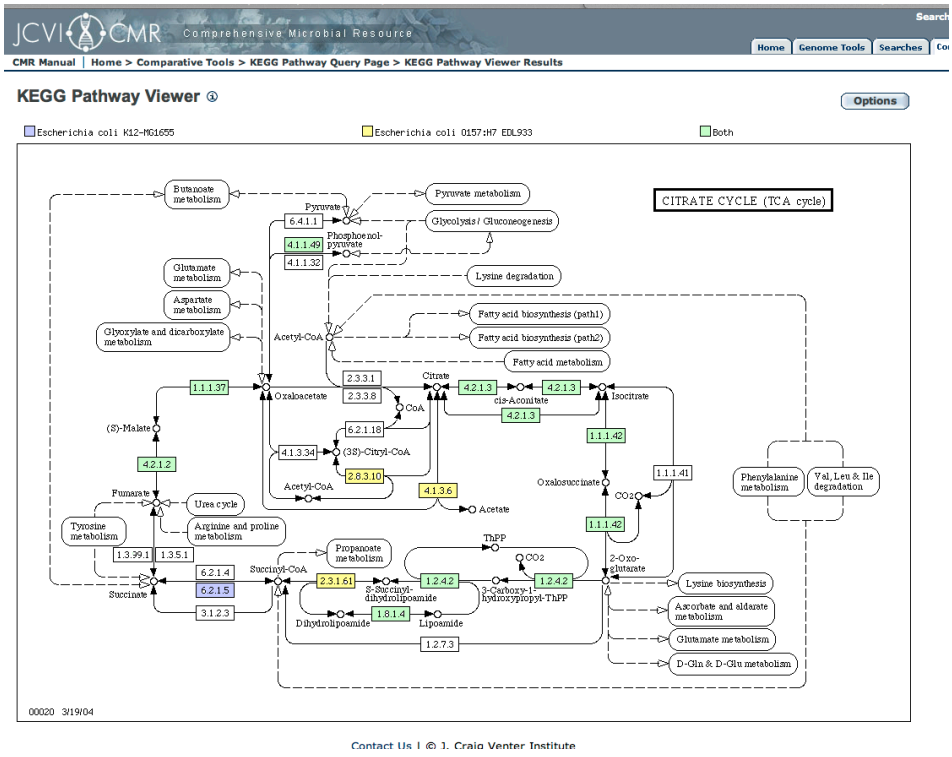
20 pts

2) This time, I want you to use the JCVI CMR (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi). Your task is to compare the predicted metabolic pathways for converting acetate into CO_2, NADH, $FADH_2$ and ATP in two strains of *E. coli*: 1) the first non-pathogenic strain to have its genome sequenced and 2) the pathogenic strain EDL933.

a) What was the original source of strain EDL933? Tell me where you found your answer.
various sources
EDL933 was isolated from ground beef in Michigan linked to an outbreak of hamburgers contaminated with *E. coli* O157:H7

b) Find a fundamental KEGG biochemical pathway in CMR that shows a difference in metabolic capacity between these two strains. You should be looking for a pathway where each genome has at least one enzyme the other lacks. Support your answer with a screenshot showing the differences.

TCA cycle



c) Do you accept that the key metabolic pathway for both strains is accurately annotated in this database? Explain your reasoning. **Limit your answer to 3 sentences or less.**

various

d) Choose one enzyme that is found in only one strain in your screen shot from JCVI CMR and determine if CMR is correct or not about it being absent from the other strain. You will have to tell me where you searched and how you conducted the search. If you can disprove the map, support your answer with data. If you cannot disprove the CMR map, explain why you were unable to find the answer.

various

http://microbes.ucsc.edu/cgi-bin/hgGateway?db=eschColi_O157H7_EDL93

http://microbes.ucsc.edu/cgi-bin/hgGateway?db=eschColi_K12

20 pts

3) Horizontal gene transfer is sometimes called lateral gene transfer (LGT) in order to keep the typical undergraduate confused and to provide yet another TLA. However, you are now on the inside crowd, so I want to ask you some questions about LGT.

a) Below is a figure from a paper that claims to have evidence of bacteria to eukaryote LGT. The method combined fluorescent labeling of a chromosome with FISH. Evaluate these data and tell me if you think the evidence is either 1) inconsistent with LGT, 2) consistent with LGT, 3) compelling evidence of LGT or 4) inconclusive. Support your answer with data. **Limit your answer to 3 sentences maximum.**

consistent with, but controls are missing

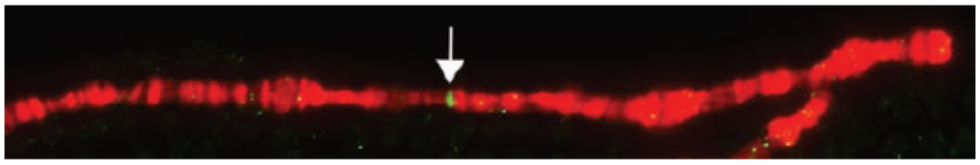


Fig. 1. Fluorescence microscopy evidence supporting *Wolbachia*/host LGT. DNA in the polytene chromosomes of *D. ananassae* were stained with propidium iodide (red), whereas a probe for the *Wolbachia* gene WD_0484 bound to a unique location (green, arrow) on chromosome 2L.

b) The accompanying PDF file called Figure2.pdf contains some data based on genome sequence analysis. The authors present Figure 2 as supporting evidence that LGT has occurred. Your task is to evaluate the data and tell me if you think the data are 1) inconsistent with LGT, 2) consistent with LGT, 3) compelling evidence of LGT or 4) inconclusive. Support your assessment by citing data appropriately. Assume zero sequencing or assembly errors happened in this research. **Limit your answer to 5 sentences maximum.**

c) The accompanying PDF file called Figure3.pdf contains additional data. The table shows 9 examples of LGT. Rank each of the nine examples from most compelling to least compelling and explain your reason for each ranking in one sentence maximum. If you feel a tie is required, then list multiple examples for a single number and reduce the final number accordingly.

Most Compelling

1. B. malayi, D. ananasse – tied with >20,000 Wolbachia reads and PCR validated junctions
2. C. pipiens >20,000 reads but no PCR validation
3. D. simulans >7,000 reads but no PCR
4. N. vitirpennis 30 reads and PCR validation
5. N. girault and N. longicornis with 1 or 2 reads and only 1 PCR validation
6. D. sechellia – only one read and no PCR
7. D. immitis – no data, only "trust me"

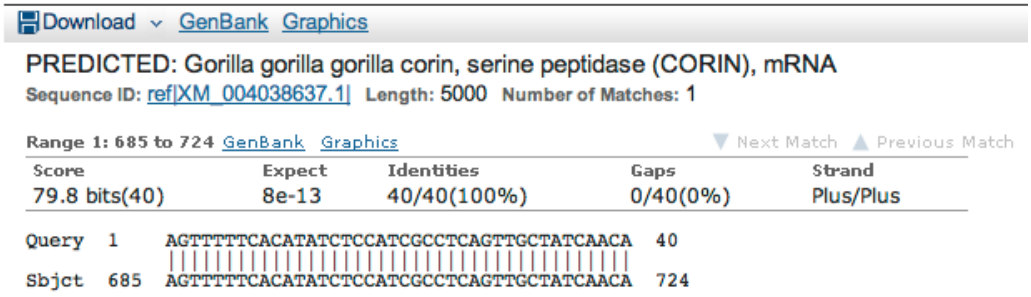
Least Compelling

20 pts

4) Here is a sequence of DNA. You need to answer the following questions using this sequence as your starting place. `agtttttcacatatctccatcgcctcagttgctatcaaca`

a) From which gene and species did this sequence come? Support your answer with evidence and be as accurate as you can be with your answer.

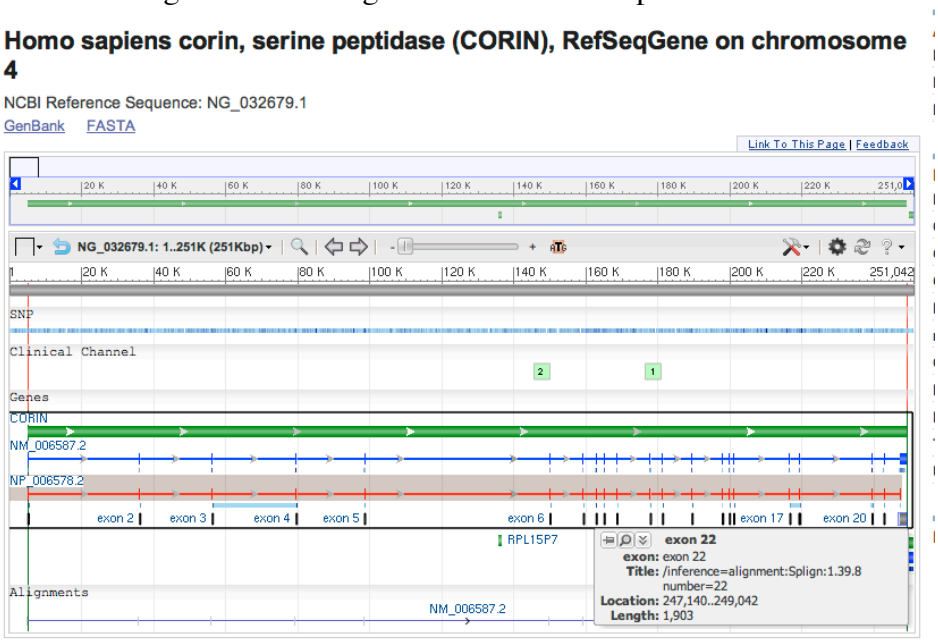
DEFINITION Homo sapiens corin mRNA, complete cds. as well as 16 other primates. Cannot tell which one was the original source of the sequence since they are all identical. Screen shot of Gorilla below.



b) How many exons are in the human ortholog? What is this gene's chromosomal position in humans? Support your answer with data.

22 exons

Location: 4p13-p12



c) List (numbered list) as many biological processes as you can find for the human ortholog. Are all of the processes very similar, or do you see some pretty diverse categories? **Explain your answer in 1 sentence.** Very diverse given it is just a protease.

List: (no particular order)

1. Protealysis
2. pregnancy (see OMIM for details)
3. blood pressure regulation
4. sodium excretion in urine
5. hormone processing

Process	Evidence Code	Pubs
female pregnancy	IMP	
peptide hormone processing	IDA	PubMed
proteolysis	IEA	
regulation of blood pressure	ISS	
regulation of renal sodium excretion	ISS	
regulation of systemic arterial blood pressure by atrial natriuretic peptide	IMP	
regulation of systemic arterial blood pressure by atrial natriuretic peptide	ISS	

d) Propose a reasonable model to show how mutant alleles of this gene could be passed down from fathers but not mothers. This answer is not intended to be a generic one that could apply to any gene, but you should combine what we have learned in class with the specific role(s) of this protein.

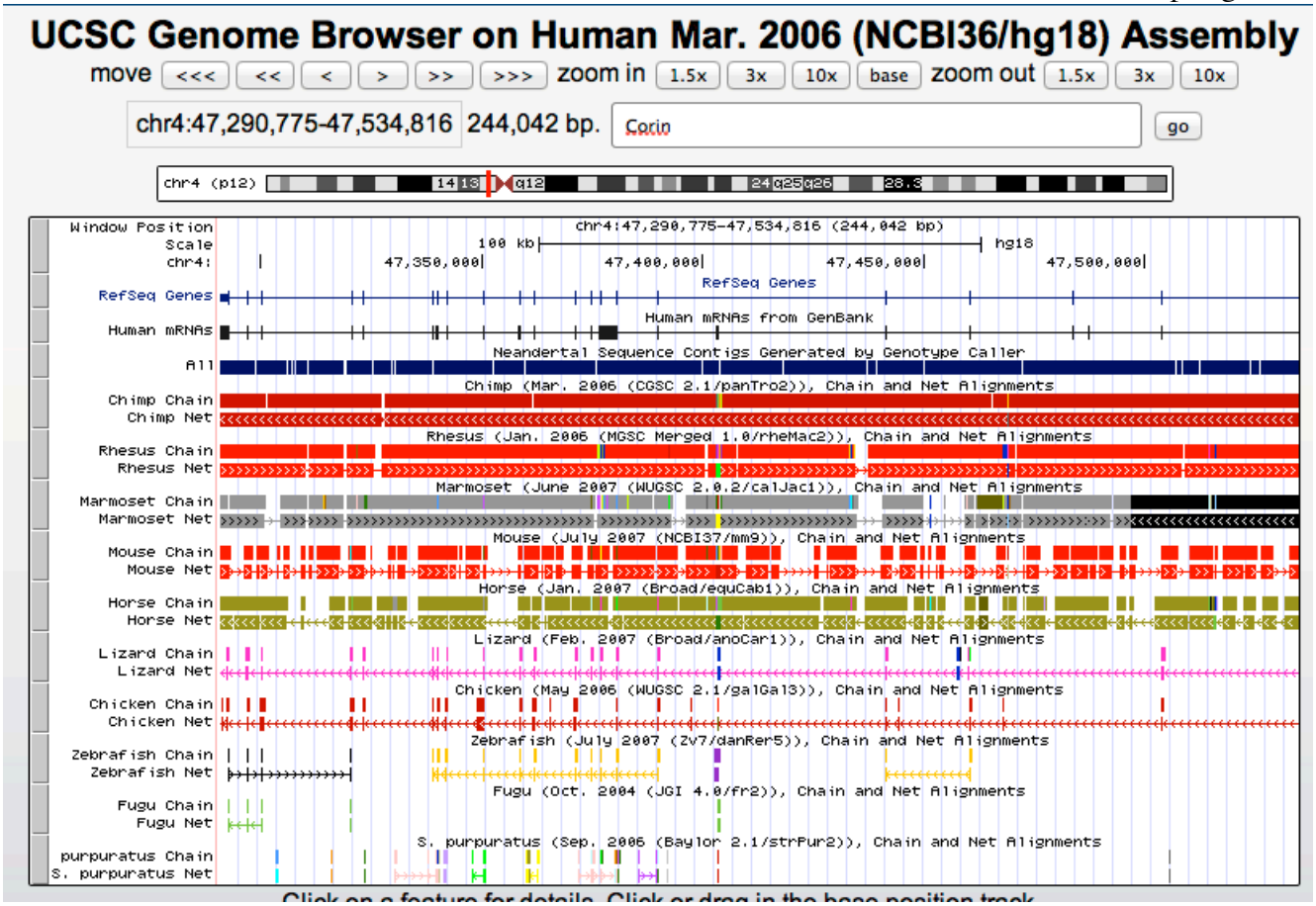
It is possible this gene could be regulated by imprinting and that it is only expressed from the maternal allele. If this is the case, then a woman with a substantial mutation would be sterile.

From OMIM: "These and other results indicated that corin and ANP are essential for physiologic changes at the maternal-fetal interface, suggesting that defects in corin and ANP function may contribute to preeclampsia." Preeclampsia is high BP due to pregnancy. Mayo Web site says, "Left untreated, preeclampsia can lead to serious — even fatal — complications for both you and your baby."

e) Demonstrate the degree of sequence conservation between the human ortholog and other species. You must provide a screenshot and then write a summary of what you conclude from your screen shot.

Highly conserved in (placental) mammals but more conservation than I expected in other animals, even sea urchin, fish and chicken (with increasing degrees of conservation). Probably due to the protease component since sea urchin does not have closed circulatory system. (see screen shot next page)

Limit your summary to a maximum of 2 sentences.



20 pts

5) This last question has to do with the ENCODE project. *You are NOT allowed to look up any scientific ENCODE papers or ENCODE abstracts*. Therefore, do not perform a PubMed search. If you do a Google search, be sure to screen any hits before clicking to be sure you are NOT reading a scientific paper or abstract.

a) What was the purpose of the ENCODE project? **Limit your answer to a maximum of 2 sentences.**

Determine function of all DNA and not just the genes. Epigomic information, histone modification and DNase hyprsensitivity sites are examples of the ENCODE information they want to correlate with biological function.

b) How many ENCODE papers were published in a coordinated way in late 2012? **Limit your answer to a maximum of 1 sentence.**

64 in 2012.

30 in late 2012

c) What does DHS stand for in the ENCODE project? Describe the physical characteristic of DNA DHS is measuring. **Limit your answer to a maximum of 1 sentence.**

DNase Hypersensitivity Sequence

d) In one sentence, define what TSS means. Support your definition using data from Figure_4.pdf.

Transcriptional Start Site

e) Summarize the two main lessons in panel B of Figure_4.pdf. **Limit your answer to a maximum of 3 sentences.**

1) Intergenic and gene body DHS are about the same sizes – TSS is the only outlier.
2) TSS DHS are about 1200 bp long, about 3X bigger than all other DHS.

f) This is the first time I have ever seen "violin plots". Summarize panel C in Figure_4.pdf. **Limit your answer to a maximum of 2 sentences.**

Although there is overlap, the vast majority of TSS DHS sites contain about 80% of the GC bases compared to the rest of the gene and the intergenic DNA.

g) In panel A of Figure_5.pdf, they tested 19 different cell lines for DHS. Summarize the findings of panels A – C **in six sentences or less (two per panel).**

Panel A: Two different genes have high degrees of DHS near the two TSS in nearly 18/19 cell types tested. Different gene body (and 1 intergenic region) DHSs varied by cell type.

Panel B: Intergenic DNA is the most cell-type specific DHS, followed by gene body. TSS are the least cell-type specific.

Panel C: The higher the percentage of GC for a TSS DHS, the more cell lines had that DNA "open" and thus available to DNase.

h) View bases 201,574,325 to 201,591,603 on chromosome 1. Show me a screen shot of this region with DHS data included in your display as well as the degree of conserved bases in 5 diverse vertebrates. Summarize what you see in your screen shot based on what you learned in Figure_4.pdf and Figure_5.pdf. **Limit your summary to a maximum of 2 sentences.**

One gene in the area (R → L) and DHS sites present, TF binding, and Histone methylation and exons conserved in diverse vertebrates. DHS sites not fully conserved and more at 3' end than 5' end (surprising). I chose one non-vertebrate because I was curious to compare.