Genomics Spring 2021

Exam 2 - Due 11:59pm EST, Sunday, 04/11/2021

Please work alone and submit through Blackboard.

Please make one pdf file ("exam2_(your first name)_(your last name).pdf) and include all the appropriate screen shots. For example, I would turn in "Exam2_John_Doe.pdf" which includes all the questions and answers for exam 2 from part 1 to part 4 including all the appropriate screen shots. In addition, you need to attach all the appropriate files you made for the exam2 when you are submitting the exam2 via Blackboard.

Other than clarification on the meaning of the exam questions, we can not give you any hint to be fair to other students.

Good luck!

Part 1 - 9 points

Part 1 uses the two attached files ERR181582a.sam and saccer3_genes_chrI.gtf. The SAM file is from a yeast sequencing run and the GTF file contains annotated genes from Saccharomyces cerevisiae (version sacCer3) on chromosome I. The original SAM data can be found here if you're interested (not needed for the exam).

(3 pts) What is a SAM file and how is a SAM file generated? Be sure to include in your answer what type of data is represented in a SAM file.

- A SAM file refers to a Sequence Alignment/Map file that contains information about how a sequence read aligns to a reference sequence. This file format is generated when a sequence alignment is performed is used to then store that NGS alignment data.

(3 pts) Upload the SAM file to Galaxy. In Galaxy, convert the SAM file to a BAM file. Submit the BAM file.

- File name: ERR181582aBAM.bam

(3 pts) List the Galaxy tool(s) you used and the parameter(s) you set to complete the previous question. Screenshot(s) or text is fine to submit.

- Once the .sam file and .gtf file were uploaded into Galaxy, I used the SAM-to-BAM converter tool offered by Galaxy. The .gtf annotation file was not formatted to FASTA, and therefore was not able to be used as a reference. Therefore, the reference genome: Yeast (Saccharomyces cerevisiae): sacCer3 was used. The SAM file was given the database/build for sacCer3 as well.

Part 2 - 8 points

(2 pts) Find the human TUBB3 gene using the UCSC Genome Browser (hg38). Turn on the Encode Regulation track (HINT: set display mode to “full” for these tracks) and NCBI RefSeq genes. In a few sentences, describe what you see at the TUBB3 locus in terms of the Encode Regulation tracks. Include in your answer what histone modification(s) appear(s) near the transcription start site of the TUBB3 gene. Submit a screenshot of this locus. (HINT: click View > PDF/EPS at the top of the browser page to export a PDF/EPS file.).

- The ENCODE regulation track, when expanded to full, shows multiple regions including a promoter region, 4 proximal enhancer regions, and 6 distal enhancer regions. These regions are near the transcriptional start site of the gene, where the histone modification tracks show a peak region of H3K4me3 methylation where the other two (H3K4me1/H3K27ac) tracks remain quite low.

(2 pts) At the same locus as Part 2.1, configure each Encode track to only show data for the HUVEC cell line. What is the HUVEC cell line? Based on the Encode tracks, do you think this gene is expressed in the HUVEC cell line? Why or why not? Submit a screenshot of this locus.

- The HUVEC cell line stands for Human umbilical vein endothelial cells, and they are cells harvested from umbilical cords which are removed at birth regardless. According to the isolated HUVEC data, one could assume that the TUBB3 gene is expressed in these cells, as the level of methylation seen on the tracks for both H3K4me1 and me3 are not fully filling the peaks that are seen in the other cell lines.

(2 pts) Use Galaxy or any other tool (hg38) to find all UCSC flagged SNPs in (hg38/db147) on chromosome 1. Be sure to import the data as a BED file. Describe the result (number of lines, what's in columns, etc.). Also, is the imported BED file 0- or 1-indexed? How can you tell?

- I used the table “snp147Flagged”. The resulting BED file had 8,787 flagged SNPs. The BED file that was returned was in BED-6 format, including the chromosome, the start locus, the end locus, the name of the variant, the score, and strand. The BED file is 0-indexed, and one can see this from the fact that the SNP’s start and end location are sequential and not the same locus, indicating a 0-based index.

(2 pts) Use Biomart (web-based or R) to search among all SNPs (Ensembl Variation 99 or latest version) on human chromosome 1 (hg38). Use the Human Somatic Short Variants database and filter for the eye tumour phenotype. Output the following Attributes: Variant Name, Variant Source, Chromosome name, Chromosome position start, Chromosome position end, Variant start in translation, Variant end in translation, Variant Consequence, SIFT prediction, and PolyPhen prediction. Download unique results and submit a spreadsheet. How many total SNPs are reported? How many SNPs are in noncoding regions? SIFT and PolyPhen scores are optional to include.

- The web-based version of Biomart was used, and the SIFT/Polyphen prediction scores were not included. There are 597 SNPs total in the excel file, with 86 of those variants belonging to noncoding regions.
- File name: Eye Tumor Variants.xls

Part 3 - 8 points

The attached human CNE (human_CNE.fasta) was discovered based on human-chicken identity. Use UCNEbase to answer the following questions. HINT: The CNE ID is in the fasta file.

- CNE ID was retrieved by viewing the FASTA file with notepad
- >chr9_Cameron id=34477 pos=chr9:23842954-23843176

(1 pts) This CNE is intergenic in human. List the most immediate upstream and dowstream protein-coding genes.

- Upstream: **ELAVL2**, downstream: **TUSC1**

(1 pts) Look for conservation of the CNE in birds, reptiles, and fish. Does conservation seem to be absent in any of those three groups? Which one(s)?

- Conservation of this element seems to be absent **in fish.**

(2 pts) Was this specific CNE described by Bejerano et al. in 2004? Why or why not (speculation is encouraged)?

- I don't believe this specific CNE was described in that paper, as there is no provided link to the work, compared to the provided links to ensemble, ECR, and Ancora.

(2 pts) Use the UCSC Genome Browser to display this CNE in the chicken genome (version galGal3). Be sure to include the Conservation track. Submit the image of the browser.

- File name: Chicken GB.pdf

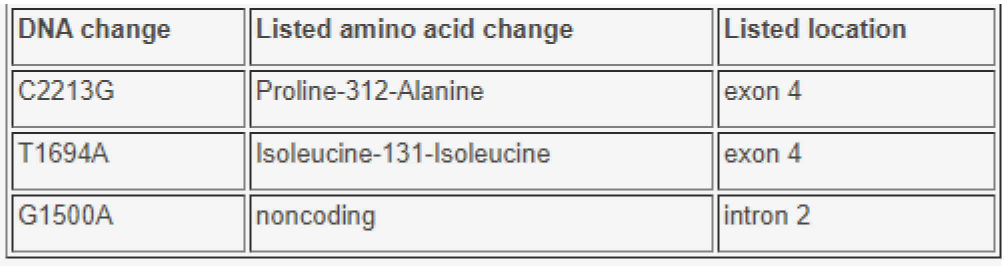
(2 pts) How many Common SNPs (dbSNP 147) are in this CNE in human? Explain why that might be the case, based on what you know about CNEs.

- This region has 5 SNPs listed from dbSNP 147, which makes sense to not have many in a region that is meant to be conserved, as it wouldn't be a conserved element if there was too much variation between individuals and species.

Part 4 - 9 Points

See the attached paper by Sun et al. There are potential flaws in the paper. The authors list three cow SNPs in the CFL2 gene. One is intronic, one is synonymous, and the other is said to be nonsynonymous.

The SNPs are at the mRNA positions in the table (this is a minus strand gene). The DNA positions are based on the unspliced full-length mRNA.



DNA change	Listed amino acid change	Listed location
C2213G	Proline-312-Alanine	exon 4
T1694A	Isoleucine-131-Isoleucine	exon 4
G1500A	noncoding	intron 2

(1 pts) Create a zero-based BED file with the three SNP locations. Label each line with the DNA change. Submit the file. HINT: The SNP at position 1694 is located at position 45,888,611 on chr21 in cow (bosTau7 build). 45,888,611 is the actual position of the SNP nucleotide, not the zero-based position.

- File name: cowSNP.bed

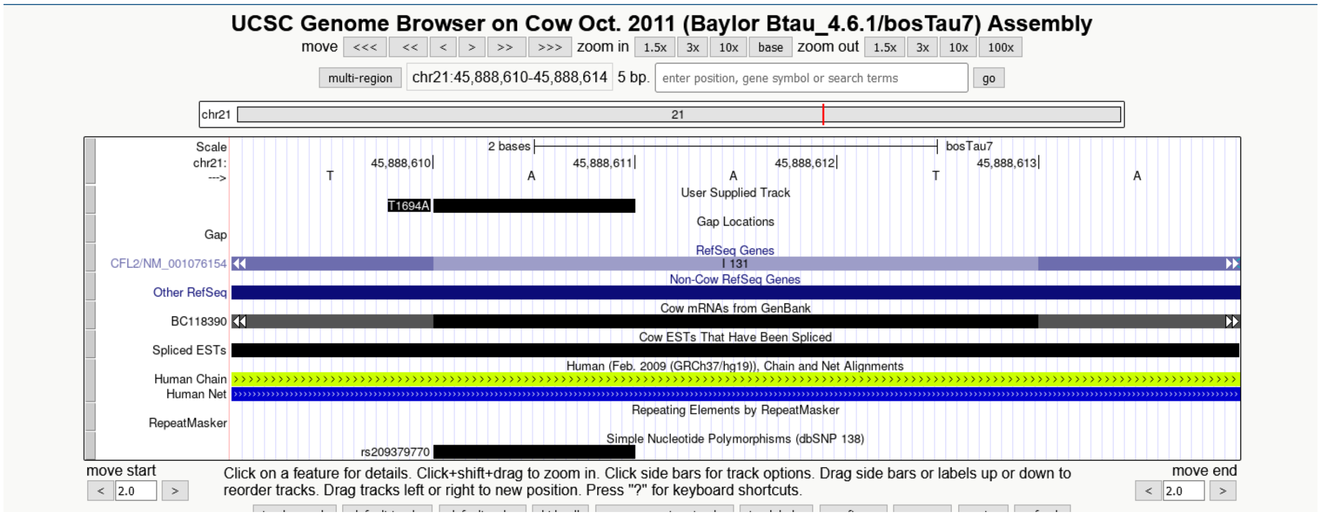
(1 pts) Load the BED file into the UCSC genome browser. Zoom into the CFL2 gene. Be sure the exon pattern and the custom track with the labeled SNPs are viewable. Submit the browser shot.

- See “user supplied track”, File name: cowSNPview.pdf

(2 pts) Based on the two SNPs with amino acid changes, what possible codon changes could cause those amino acids to change? Example: A C-to-A nucleotide change could make a His-to-Asn change by changing a CAU codon to an AAU codon.

- For T1694A, it is the start position of the codon that would normally be AUU in the mRNA(isoleucine), and changing this to would result in an AUA for Isoleucine, meaning this SNP is a silent change.

-

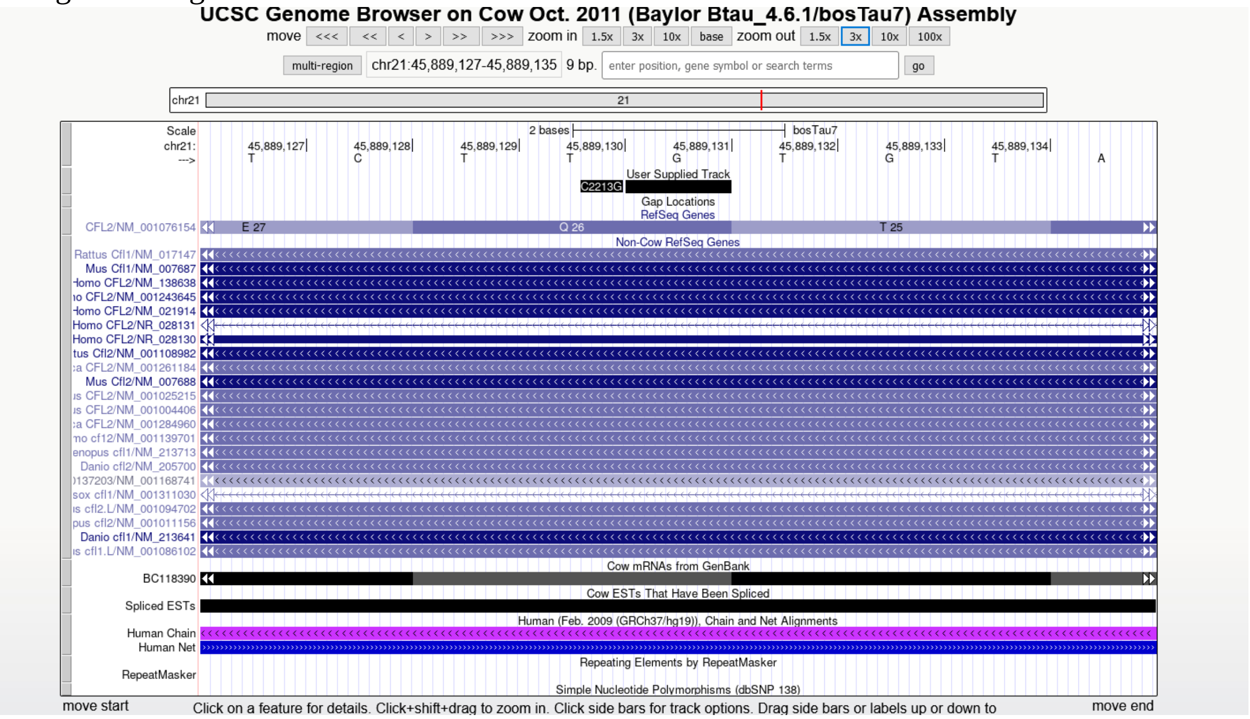


UCSC Genome Browser on Cow Oct. 2011 (Baylor Btau_4.6.1/bosTau7) Assembly

chr21:45,888,610-45,888,614 5 bp.

Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

- For C2213G, it falls at the beginning of a codon that normally codes CAA that codes for Glutamine, and a nucleotide change to GAA for this codon would lead to a change to the AA Glutamic acid, which is a difference in terms of charge as glutamic acid is negatively charged while glutamine is not.



UCSC Genome Browser on Cow Oct. 2011 (Baylor Btau_4.6.1/bosTau7) Assembly

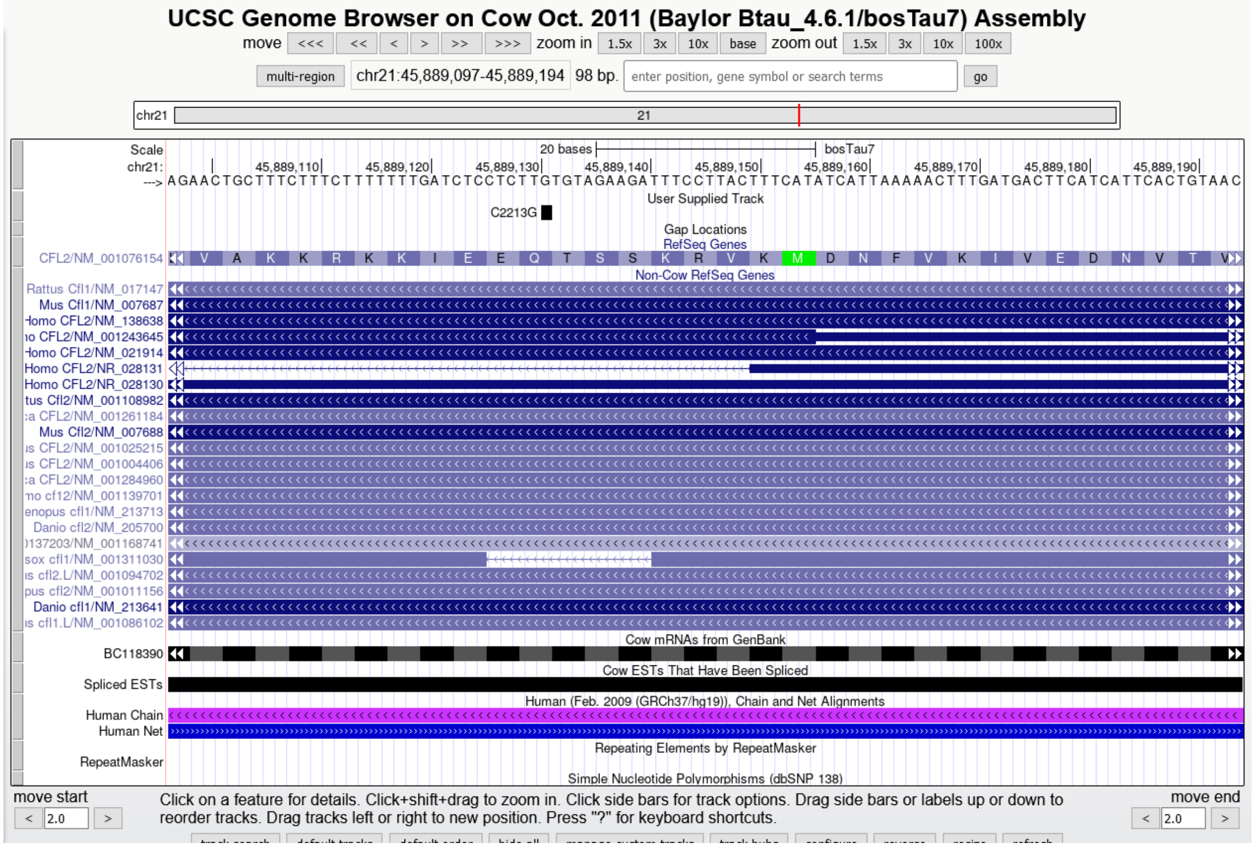
chr21:45,889,127-45,889,135 9 bp.

Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options. Drag side bars or labels up or down to

- Difference seen between the positions here and what was found in the paper can be attributed to the fact that this gene is on the reverse strand in the UCSC genome browser, and listed as exon 2 for C2213G and exon 4 for T1694A in UCSC which also differs from the paper.

(2 pts) The CFL2 protein is 166 amino acids in length. That would make an amino acid change at position 312 a bit difficult. Looking at the exon sequence for CFL2 in UCSC Genome Browser, in which exon is the C2213G SNP? Specifically, what part of the exon: 5'UTR, CDS, or 3'UTR?

- The C2213G SNP can be found in the 2nd of 4 exons in the CFL2 gene, specifically in the CDS region of the gene as it can be found downstream of the starting codon.



UCSC Genome Browser on Cow Oct. 2011 (Baylor Btau_4.6.1/bosTau7) Assembly

chr21:45,889,097-45,889,194 98 bp.

Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

(2 pts) The attached cow_cfl BED files contain all exons for the CFL2 gene and the coding exons for the CFL2 gene. Intersect your SNP BED file to each of those files using Galaxy or command-line BEDtools. How many of the three SNPs intersect with all exons? How many intersect with coding exons?

- The two SNPs at T1694A and C2213G intersct with locations in the coding exons BED file, which matches what was shown in the UCSC genome browser
- All three SNPs are found in exons according to the intersection between their locations and the cow_cfl_all_exons, which does not match the initial assumption from the authors that one of the SNPs was to be found in intron 2.

(1 pt) Submit either a screenshot of your output (Galaxy) or the commands you ran (command line) for the previous answer.

- File name: cow.coding.png
- File name: cow.all.png