UCSC Genome Browser

SEQUENCE ANALYSIS EXERCISES

Part I. Browsing for genomic information

1. Find the human gene PIK4CA in the UCSC Genome Browser:
   - Select the human genome, May 2004 assembly and enter "PIK4CA" as the position.
   - Once you do the search, select any of the choices under "Known Genes" or "RefSeq genes".
   - Under "Genes and Gene Prediction Tracks" at the bottom half of the page, select "full" for Known Genes, RefSeq Genes, MGC Genes, and Ensembl Genes.
   - Click on "refresh" or "jump".
   - Zoom out until you can see all the transcripts named PIK4CA.
2. Relative to the reference chromosome, what strand is the gene on?
3. Try clicking on the gene structures to see what information is linked.
4. How many transcripts are there, according to (a) RefSeq, or (b) Ensembl?
5. How many exons does the longest transcript have? To clearly see the answer for the Ensembl transcript, click on the gene; follow the link to "ENST____" (Ensembl transcript), and then look for a link called "Exon structure".
6. Look at the longest intron. What do you see there?
7. Under "mRNA and EST Tracks", turn on the "Spliced ESTs" tracks to squish, pack, or full. Can you find expression evidence of these transcripts?
8. At the top of the page, what do the "Ensembl", "NCBI", and "PDF/PS" links do?

Part II. Extracting annotated genomic sequence

1. Enter NM_058004 (the RefSeq ID of longer of the PIK4CA transcripts) into the position box and click on "jump" to get the browser to show the width of the gene.
2. What are the genomic coordinates of this transcript?
3. How long is the gene (in genomic context, rather than in cDNA context)?
4. To extract the genomic sequence of the PIK4CA gene, including 5kb upstream and 1 kb downstream of NM_058004, adjust the coordinates in the position window.
   - Since the gene is on the negative strand, adding 5000 to the second coordinate (y, where the position is chr22:x-y) will expand the window to include 5 kb upstream.
   - Subtracting 1000 from the first coordinate will extend the view to the 3' end.
5. What are the expanded coordinates?
6. At the top of the page, click on the "DNA" link and note that you could adjust the coordinates at this time too.
7. Note, however, that "upstream" and "downstream" refer to the reference chromosome (so directions are opposite for a gene on the negative strand, like PIK4CA).

- Check "Reverse complement" since the gene is on the negative strand, and click on "Extended case/color options."
- To capture some of the gene and EST mapping data with your genomic sequence,
  - enter 255 under the Red box for RefSeq genes,
  - enter 255 under the Blue box for Ensembl Genes,
  - check "underline" for Spliced ESTs
  - click on Submit.
- What's the significance of the formatting of the output file?

Part III. Gene-finding with comparative mammalian genomics

1. Find the **human** gene NM_016175 in the human UCSC Genome Browser using the latest assembly (May 2004).
2. Once you're on the browser page, click on the gene (under "RefSeq Genes"). What information does this lead to?
3. Go back to the browser. How many exons does NM_016175 transcript have? Do you think it's the whole gene?
4. To help answer the question, try the next few steps:
   - Change Known Genes, RefSeq Genes, and MGC Genes, and Ensembl genes to "full" and click on "refresh" or "jump".
   - Zoom out 10X and turn on the "Spliced ESTs" tracks to full.
   - Now can you convince yourself that NM_016175 is a full-length transcript or not? If not, how could you identify any longer transcripts?
5. Keep the human browser open, use the sequence of the longest transcript of this gene encoding truncated calcium binding protein (TCBP; BC069051, which you can also get to by clicking on the gene in the browser and following the links), search the latest **mouse** genome with BLAT.
   - When you get back the BLAT Search Results, follow the "browser" link to get the genome browser view.
6. Does BLAT bring you to the longest transcript of mouse TCBP? Why or why not?
7. Are you sure that this is the mouse ortholog? What would it take to convince you?
   - Look at the name of the gene downstream of **mouse** TCBP. What is it called?
   - Look at the name of the gene downstream of **human** TCBP. What is it called?
   - Should this finding be unexpected? What phenomenon are you observing?

Part IV (supplementary). Gene and genome analysis through annotation

1. Find the human gene TGFB3 (Transforming growth factor beta 3) in the Ensembl project:
   - Click on human.
   - Enter TGFB3 and Search for "Anything".
   - After clicking on Lookup, note that the first hit is to Vega, the Vertebrate Genome Annotation (VEGA) database of manually genome annotations. According to Vega, how many transcripts does TGFB3 have?
   - For now, select a match to "Ensembl Gene" that refers to the gene.
2. Follow the link under Genomic Location to view the gene in its genomic location. This presentation of data should look somewhat familiar.
3. Go back to the GeneView page.

4. Peruse the GeneView page, noting the information provided under Orthologue Prediction, Similarity Matches, and SNP information.
5. How many other genes are classified as having growth factor activity? To answer this question,
    - Follow the GO (Gene Ontology) ID link to the GO term that refers to proteins that act as growth factors ("growth factor activity").
    - Note that this term appears in multiple locations in the molecular function hierarchy.
    - While you're viewing the GO tree, look at the column on the right. How many genes are in this category? Get a list of these genes by clicking on the "_ gene(s)" link.
6. Get sequences for all human proteins with growth factor activity:
    - Go back to the Ensembl home page.
    - Go to EnsMart and click on START
    - Select Ensembl Genes as the Focus and Homo sapiens as the species, and click "next".
    - Under REGION, uncheck "Limit to" to search the entire genome.
    - Under GENE ONTOLOGY, check Evidence code for mapping and enter "growth factor activity" next to Molecular Function.
    - (Entering a GO ID should work but doesn't at this time.)
    - Click on "next" and then note the tabs for Features and Sequences.
    - Select the Sequences tab.
    - Select Transcripts/proteins and Peptide
    - Select "Text/Fasta" as output format, gzip as File compression, enter a file name and click on Export.
    - (Optional) Save the fasta-format file and look at it in a text editor.
7. Get orthologs for all human proteins with growth factor activity
    - On the final MartView page, select the Features tab.
    - Under GENE: Ensembl Attributes, check Ensembl Gene ID and Description.
    - Under MULTI SPECIES COMPARISONS: Mouse Homolog Attributes and Rat Homolog Attributes, select Ensembl Gene ID.
    - Select "Text, tab separated" as output format, gzip as File compression, enter a file name and click on Export.
    - (Optional) Save the file and look at it in Excel.