BI 188 paper 2011

The objective of the paper is for you to explore more deeply one human gene, or locus, and to learn who it has been studied, what is known, and identify the problems it presents in genetic and genomic terms for the future. In the process we want you to do some initial analysis of your own. This analysis will be very basic (see Part III). There is lots of guidance below that is intended to keep you from going too deep or wandering too far. Also consult with us in email. Be certain to address such questions with a clear heading: **BI 188 Paper Question.**

I. Review of your gene and the associated genetic disorder

A. What genetic disease/phenotype(s and ranges) are associated with mutation of your gene

B What are the genetics of the disorder in humans (pattern of inheritance, nature of mutations de novo? Heritability?.)

C. Give a very succinct account of the evidence that this gene is causal (or partly causal) for the genetic disorder. Emphasize the best evidence, even if it was not historically the first to be shown, and say why your find it most convincing.

D. Summarize what is known about the distribution of mutations in the human population (frequencies in different populations, specific alleles in partcular populations, as applicable)

II. Animal models. Are there animal models for study of your gene/disorder ? Breifly say what those models are.

A. How well (if at all) does the best animal model reflect the human disorder?

B. What is the most important thing that has been shown thus far with one of the animal models? Explain why you think it is important (for example, is it medically significant for understanding the disease or testing a treatment, or scientifically important because it illuminates mechanism?).

C. If there is no animal model, discuss why, and review any major alternative (cell models for example).

III. Genomic data at your locus

We will provide more specific help on this in a separate posting; also feel free to email your TA's with questions.

A. Go to the UCSC genome browser to find your gene and relevant surrounding locus. How much intergenic space separates your gene from its two nearest neighbors. Capture a map of the locus and the UCSC gene models by opening that track and taking the figure shot. Use this for reference in the rest of your answer.

B. call up ENCODE data tracks for

1. RNA-Seq (Wold) data; Is your gene expressed detectably in Tier 1 or Tier 2 cell types.
2. Go to RNA-Seq data for the cell type that shows highest expression of your gene. Is there evidence for the expression of more than one splice isoform based on known UCSC gene models and the expression track? Annotate a screen shot to show the major isoforms you find evidence for. [If your gene is not significantly expressed in any of these cells, go to K562 cells and identify the nearest gene to yours that is significantly expressed and evaluate it for one or more isoforms.] If there is evidence more than 3 isoforms, describe the most prevalent 3 and just say that there are others.
3. Does the RNA-Seq data in the highest expressing cell type for your gene support only the previously known 3'UTRs from existing REF-Seq and UCSC gene models? Or something different? Use nearest the neighbor gene significantly expressed in K562, if you have a gene that is not significantly expressed in the ENCODE cell lines. Breifly discuss the biological meaning of different 3' ends.

C Prime. Look at any one of three kinds of ENCODE data at your locus and interpret it realtive to expression of your gene: DNAse1 hypersensitivity track; a transcription factor of your choice; Chromatin histone mark data.

C Alternate. This is an alternative to C, and it focuses on comparative genomics and evolution of your locus. Assess conservation of the coding, intronic and upstream 20kb of your gene. If your gene is larger than 30kb, you should do the analysis on the first 30kb plus the upstream 20kb. You can also use the conservation track on the browser to guide you to a region for more detailed MUSSA analysis or you can use functional track as in 4A to target a region for MUSSA analysis. Thus, if your locus has some especially interesting candidate noncoding sequence, such as the LCR of globin, you can pick that region for focusing a conservation analysis.

Show and explain a key screenshot or two, and interpret what the pattern of conservation may mean (ie, does it say something about different regions of the protein? The functional importantce of the 3'UTR ? The existence of possible alternative exons or regulatory elements within introns? The more or less important parts of the promoter region etc….).

The MUSSA software for Alternate C and its use will be at the website and a detailed videa tutorial on its use is also downloadable. It use will be reviewed for anybody with questions through email to Katherine or in office hours on thuesday May 26. Do MUSSA analysis using human, mouse, dog, cow. Do it again, adding chicken or zebrafish. Explain how the results differ.

IV. **Thinking forward:** Conclude with a brief discussion of what you think is most important to find out next about this disorder and the gene(s) that cause it,

and discuss why. This might be an experiment in an animal model or it could be a study in cells tunors or human genetics and functional genomics. You do not need to apply all the approaches- just pick one or at most two questions to ask and explain the experiment to do and how you would interpret the result..

Bibliography. Primary literature references from the journals are desirable. But you need to reflect this because they are really YOUR source. If it is a core point or study or data, we do want you to look at the primary source. But when you are taking information from a review or from a text or a website that –you should cite the review (unless you really go see what the primary research paper says). Using a good review as source material is OK for information that is less central to your argument and especially to introduce ideas and interpretation that might be unique to a review. Over-documenting where the credit comes from for an idea or fact is far better than under-documenting, but we do not want inflated bibliogrphies.

Pre-approved project topics:

Huntingtin
Dystrophin
BRCA1/BRCA2
the genes of rhabdomyosarcoma (PAX3 and Pax7 and FOXO plus the question of the nature of others that are not these translocations)
c-met
p16/INK4a
beta globin (thalassemias)
Myc (c-myc focus, with attention to its relationship with L-myc and N-Myc)
CNTNAP2 (context is neurexin family)

You are not restricted to the topics on this list, but you must notify one of the TA's of any other topic by Friday, May 20th so we can make sure that the scope of your project is reasonable.

If you are interested in doing any programming as a part of your analysis (not a requirement), you are welcome to contact Katherine and/or Georgi for advice. Additionally, we are willing to teach you some introductory Python if you have little or no programming experience but would like to learn. Email us if you are interested.

Final details:

The paper should be approximately 10-15 pages. The format is Times New Roman or Arial 12-point font; 1-inch margins; double-spaced.