InterMine use-cases based on FlyMine and HumanMine

Below are a set of use-cases based on real research examples. Work through in order or pick and choose which you want to have a go at.

Use case 1: Analysis of transcripts from a fly transcriptomics study

You have carried out a transcriptomics analysis of genes from two groups of flies under different stress conditions, oxidative stress induced by paraquat and H2O2 and Endoplasmic reticulum (ER) stress induced by tunicamycin. The two sets of transcripts are stored in FlyMine as public lists (UseCase1_transcripts_oxidativeStress and UseCase1_transcripts_ERstress). Answer the following questions:

1. Are there any **genes** in common to both sets, suggesting a core stress response?
2. Do these genes show enrichment of any particular type of protein?
3. Can you identify pathways that appear to operate in oxidative stress but not in ER stress?

Use case 2: Creation of a blast database

You want to create a blast database of all *Drosophila pseudoobscura* genes. How would you output the sequence of each gene in fasta format? Try using the query builder to create this search.

[to use this data for blast you would then need to run the makeblastdb program, but we won't do that for this exercise].

Use case 3: Intron size and expression in Drosophila

You are studying the effects of intron size on gene expression.

1. Identify all genes in *Drosophila melanogaster* that have at least one intron with a size between 100 and 300 Kb.
2. Do any of these introns contain known regulatory regions?
3. Which intron contains the most regulatory regions?
4. Modify the query/results to show the gene symbol for each intron.

Use case 4: Analysis of transcription factor binding sites

Dichaete is important for a number of key developmental processes, including embryonic segmentation, hindgut morphogenesis and nervous system development. You are interested in learning more about the role of Dichaete during early *Drosophila* embryogenesis. You have used a combination of DamID and chIP to generate a set of high-confidence diachete binding intervals. The full set of intervals will take up to 10 minutes to upload. We will therefore use the set of coordinates from chromosome 3 only. You can find the coordinates here:

https://docs.google.com/document/d/1JMbkPvJYqe4jaE-ryiQyH5RwaBAaNywGvsv5Yq2QR4Q/edit?usp=sharing

1. Find genes that overlap the set of Dichaete coordinates.
2. For this set of genes answer the following questions:

    a. Look at the expression of these genes - where are they expressed in particular? Can you identify genes that are expressed in early embryonic development?
    b. Are any of the genes involved in developmental processes?
    c. Dichaete is thought to be a key transcriptional regulator which may activate/repress a number of other transcription factors. Does the set of Dichaete target genes contain other known transcription factors?

Use case 5: Analysis of genes from a Drosophila RNAi screen

You have carried out an *in vivo* RNAi screen in *Drosophila* neurons and assayed for obese or lean phenotypes by measuring changes in levels of stored fats (in the form of triacylglycerides or TAG). 1746 genes were tested in the screen. In the first round of screening 510 genes produced statistically significant changes in TAG levels. You want to analyse these 510 genes further to identify genes/pathways that function in the nervous system to regulate energy balance. Use FlyMine and HumanMine to find the following:

1. Find human orthologues for the 510 Fly genes.
2. Analyse the genes against GWAS data to identify any hits from obesity-related genome-wide association studies. (note: phenotypes such as: metabolic syndrome, fasting glucose levels, triglyceride levels, waist-hip ratios, body-mass index, obesity, type-1 and type-2 diabetes.). (you should get around 120-140 genes)
3. Send this set of genes back to FlyMine.
4. Using the 1746 genes tested in the screen as the background population, examine enrichment statistics for the genes identified in the previous step. (NOTE: for genes sent back from humanMine you will first need to intersect with the original list to remove any orthologues that were not from this list). You might need to adjust the multiple test correction for the enrichment.

5. Examine the expression data for this set of genes. Where are these genes expressed in particular? (note that nearly all genes implicated in monogenic obesity are expressed in the brain).
6. Are any of the genes involved in energy homeostasis - e.g glucose homeostasis or metabolic processes.

The following lists are already saved in FlyMine:
UseCase5_ScreenHits: the 510 genes showing changes in TAG levels
UseCase5_backgroundGeneList: all 1746 genes tested in the screen

Use Case 6: Analysis of genes involved in Parkinson's disease

You are joining a lab investigating Parkinson's disease. You use HumanMine to find out more about the genes involved:

1. Identify the set of high-confidence genes involved in Parkinson's disease according to Genomics England Gene Panels.
2. Parkinson's disease is a complex multigenic disorder. Identify other diseases that these genes are involved in.
3. Are there any clinically significant SNPs (Pathogenic) associated with these genes and this disease?
4. Parkinson's can be early-onset or late-onset. Identify the sets of genes associated with each. Is there any overlap in these sets of genes?
5. Are there any known interactions between genes (or protein product) within the Genomics England panel list.
6. In which tissues have these genes and the proteins for these genes been shown to be located?