Topic 6: functional genomics

Created	@November 9, 2023 7:12 PM
Reviewed	☐

Topic 6. Functional genomics

2.1 Functional genomics: ArrayExpress and Expression Atlas practical

This practical will introduce you to the data content and query functionality of **ArrayExpress** and **Expression Atlas** through 5 different exercises:

1. **ArrayExpress:** Experiments studying human gliomas with samples of known sex
2. **ArrayExpress:** Experiments looking at chicken with disease or infection
3. **Baseline Expression Atlas:** Look up expression of *TNNI3* gene in normal human tissues
4. **Expression Atlas:** Finding differentially expressed genes in cow mammary epithelial cells challenged by heat-inactivated bacterial antigens
5. **Differential Expression Atlas:** Comparing samples from alopecia areata patients against healthy controls

More information can be found on the dedicated help pages:

ArrayExpress: https://www.ebi.ac.uk/arrayexpress/help/index.htmlExpression Atlas: www.ebi.ac.uk/gxa/help/index.html

Exercise 1. ArrayExpress – Experiments studying human gliomas with samples of known sex

Scenario

Incidence of brain cancers such as gliomas is higher in males than in female, with male:female ratio of about 1.5:1 (Darefsky & Dubrow, 2009). You study gliomas and would like to retrieve experiments where the sex of the glioma samples is known, so you will have the option to treat “sex” as a potential confounding variable during data analysis.

Task

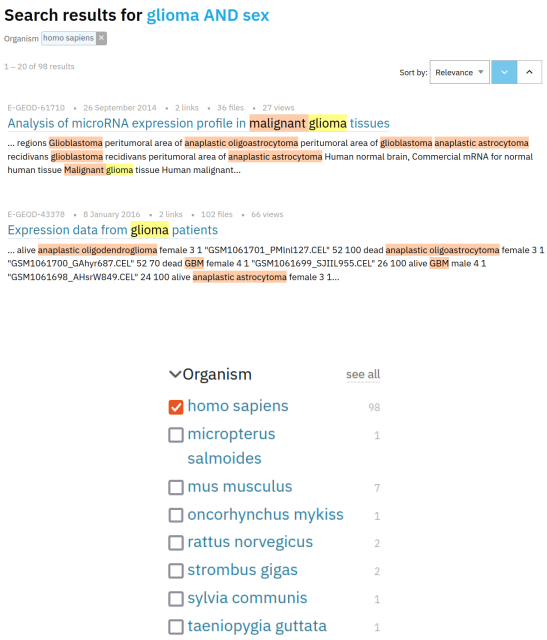
Use ArrayExpress (http://www.ebi.ac.uk/arrayexpress) to find relevant experiments in humans. Try to use advanced search (e.g. *glioma AND human*) to make your search more specific.

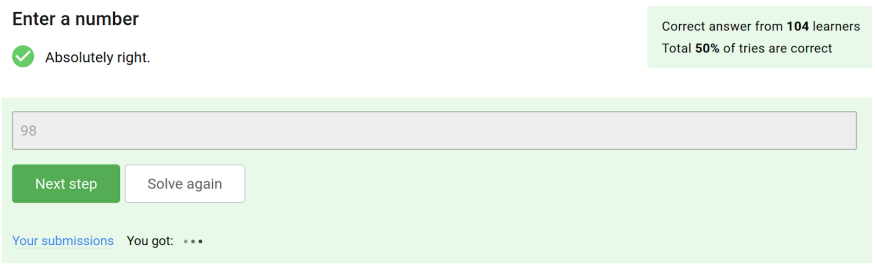
Directions

1. Read the "*Advanced search*" help section of ArrayExpress: https://www.ebi.ac.uk/biostudies/help#advancedsearch
2. Start from the "*Browse*" page: http://www.ebi.ac.uk/arrayexpress/browse.html
3. You are looking for experiments with glioma samples (and not merely mentioning the cancer somewhere in the experiment's record), so the first part of the search term would be glioma. For samples of known sex, use sex. You can use "female OR male" , but since "sex" is an EFO parent term covering both female and male, "sex" will get you the same answer and make the search term less clumsy.
4. Make sure "human" is selected in the "Organism" field at the left.

Question to answer below

How many relevant experiments did you retrieve? 98





Exercise 2. ArrayExpress – Experiments looking at chicken with disease or infection

Scenario

You are mining public data sets from studies which looked at gene expression in chicken affected by diseases or infections, to see if there are any conditions worth studying but have not received much attention in the research community.

Task

Use ArrayExpress (http://www.ebi.ac.uk/arrayexpress/experiments/browse.html) to find relevant experiments. Try to use advanced search (*e.g.* "infect") to narrow down the search space. Once you are satisfied with the list of experiments, take a look at the range of diseases/infections already covered. Choose one experiment and look in more detail, for example:

- Can you tell whether it's a microarray or sequencing experiment?
- What is the intent/aim of the experiment?
- How many samples are used? Any biological or technical replicates?
- Do you know where to download meta-data and data files?

Directions

1. Refer to the "*How to search*" help section of ArrayExpress if you need so: https://www.ebi.ac.uk/arrayexpress/help/how_to_search.html

2. Start from the "*Browse*" page: http://www.ebi.ac.uk/arrayexpress/browse.html

3. For experiments which look at diseases and infections, the common keywords used in sample annotation (and experimental variable declaration) are: "infection", "infect", and "disease". We will use all three keywords in this search. As described in the Advanced search webpage, the search behavior can be modified by using boolean

operators and brackets; e.g., *Leukemia AND (mouse OR human)*. Following this separated with OR, then click "*Search*".

4. Add the terms "*Gallus gallus*", "RNA seq of coding RNA" (for study type), and click "*Filter*".

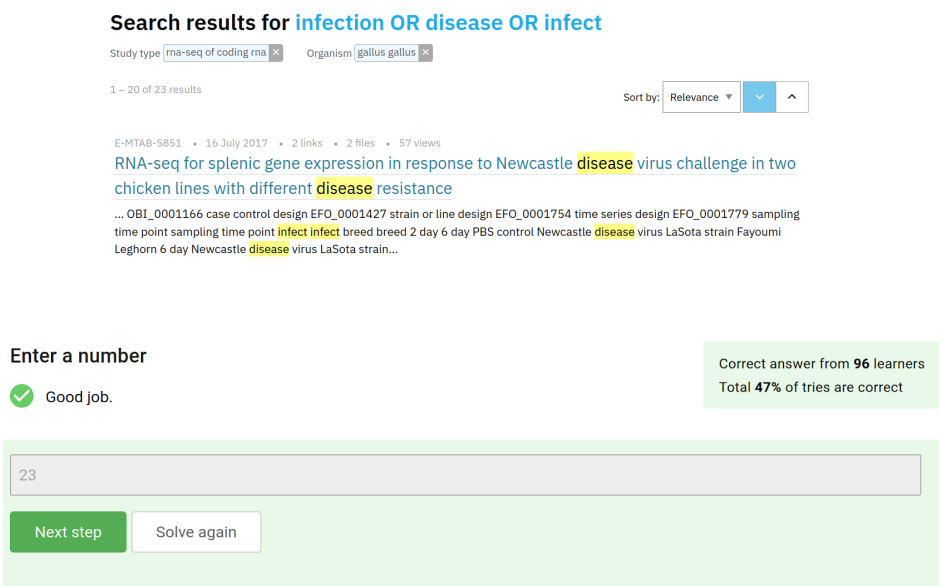
5. Most of the experiment titles, especially those from direct ArrayExpress submissions, should give you an idea about the intent of the experiment. Click at an accession number to see more details about one particular experiment.

6. You can also play around with more filters, e.g. looking at sequencing experiment only, or ArrayExpress direct submissions only, and see how the results list changes.

7. Most of the information you'll need for an experiment can be found its dedicated page. Look at the title and "Description" field for the intent, "Samples" for sample details, "Data files" for all the downloadable files (meta-data and data alike) hosted at ArrayExpress, and "Linked information" for links to European Nucleotide Archive where raw sequencing read files are stored and other relevant information.

Question to answer below

How many relevant experiments did you retrieve? 23



Exercise 3. Baseline Expression Atlas – Look up expression of TNNI3 gene in normal human tissues

Scenario

Imagine you are reading a paper about regulation of liver gene expression in liver cancer patients. The paper's authors presented evidence that one gene, *TNNI3*, is downregulated in cancer patients when compared to healthy controls. *TNNI3* stands for "*troponin I type 3*". You vaguely remember that troponins are responsible for muscle contraction, and since there is no muscle in liver, you're sceptical about the authors' findings in liver. So, you decide to look up *TNNI3* in the Expression Atlas to see in which tissues the gene is expressed in healthy individuals.

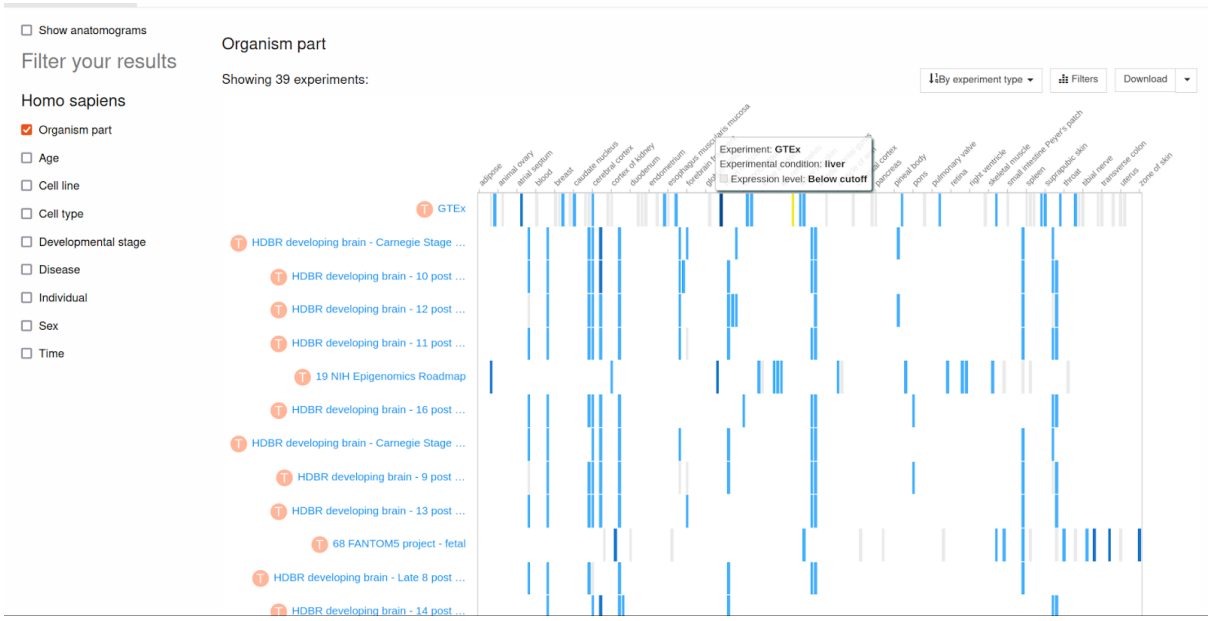
Task

Use the Expression Atlas database (http://www.ebi.ac.uk/gxa/) to search for baseline expression data of human *TNNI3* gene.

1. There are over ten human baseline expression experiments with adult samples in the Atlas. Is *TNNI3* expressed in the liver? Do the experiments agree on this?

Directions

1. Have a look at the Expression Atlas "*Help*" section: https://www.ebi.ac.uk/gxa/help/index.html **2.** Open the Atlas homepage, http://www.ebi.ac.uk/gxa/. **3.** To look at the expression of the gene across a panel of tissues, start typing "*TNNI3*" in "Gene / Gene properties" box (choose from suggestions), select "*Homo sapiens*" for human in the "Organism" list, and click "*Search*". **4.** On the search results page, zoom into the heatmap to reveal "liver" and other tissues (organised in alphabetical order). Uncheck "Show annatomograms" if you would like to have more horizontal space for the heatmap on the screen. **5.** If "liver" is not showing you can find clicking in the "Filters" button and make sure the organ is selected. Sometimes the organ can be found under the "Unmapped" list. Pay attention to the level of expression in this tissue across different experiments (rows). Is *TNNI3* expressed in the liver? Do the experiments agree on this? **Question to answer below** Is *TNNI3* expressed in the liver according to "GTEx"?



Task

2. Take a closer look at the data in experiment “GTEx”. Can you find out more about the liver sample(s) used?

Directions

6. To look at the “GTEx” data set in more details, click at the data set. Once you’re on the GTEx page, decrease the "Expression value" cutoff to 0.0 TPM to display all samples in the heatmap.

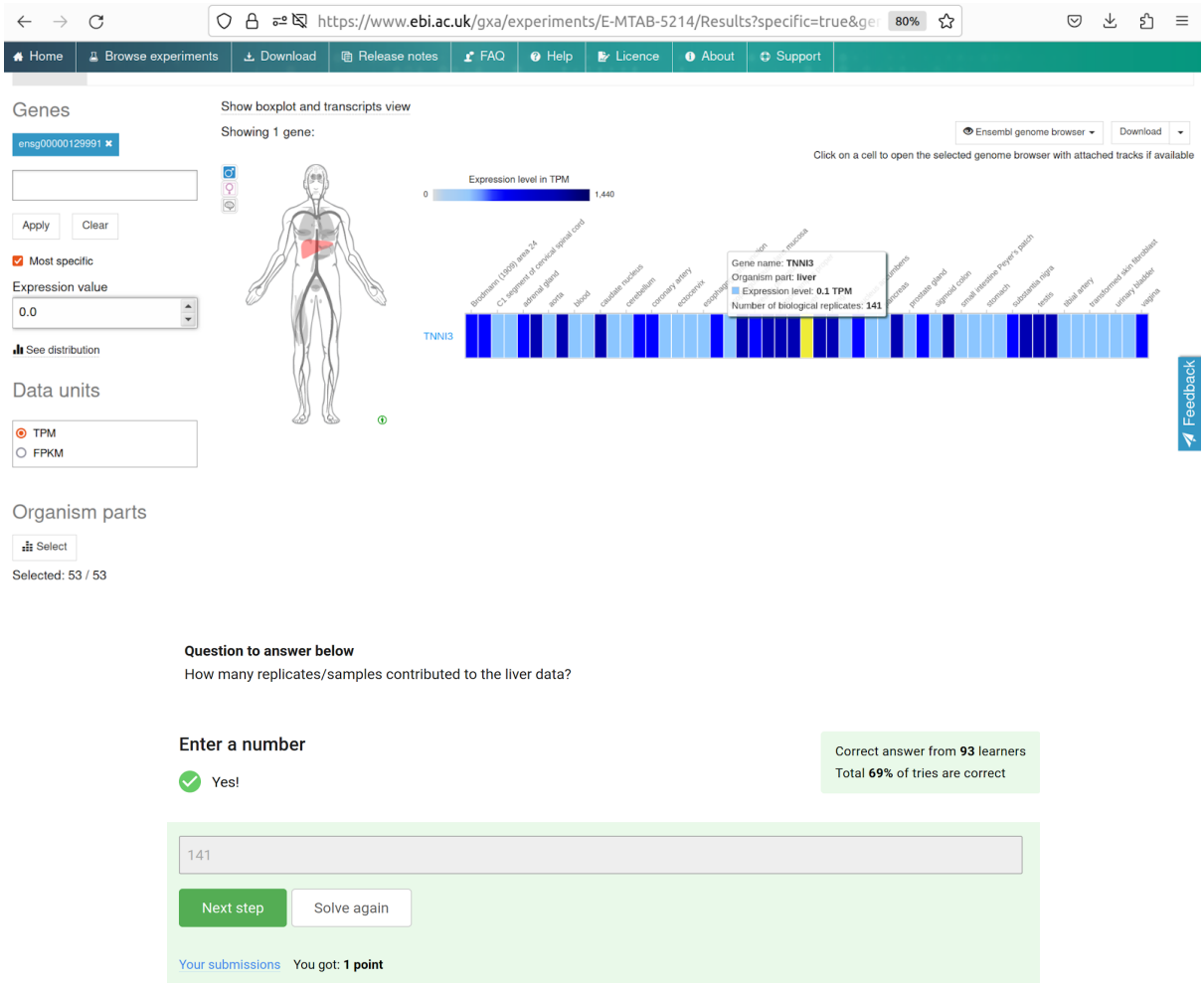
7. Hover your mouse over the “liver” column to find out more about the samples used. What is the expression level in liver? How many replicates/samples contributed to the

liver data?

8. You can also see the full list of samples more comfortably by going to the "Experiment design" page.

Question to answer below

How many replicates/samples contributed to the liver data? 141



Task

3. Find out more about the *TNNI3* gene: How many transcripts are there in *TNNI3* gene? On which chromosome would you find them?

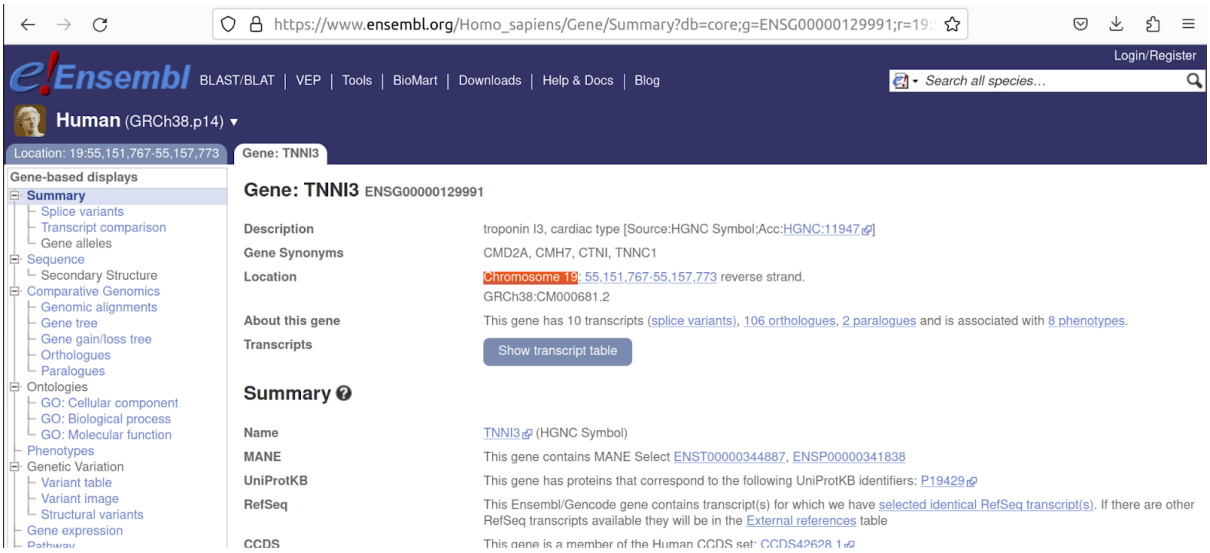
Directions

9. The Atlas links you directly to Ensembl by clicking on a cell in the heatmap.

10. Click the "Gene" tab in Ensembl to find out more about the *TNNI3* gene. How many transcripts are there in *TNNI3* gene? On which chromosome would you find them? Which Gene Ontology terms are associated to the gene? Does the gene have orthologues in other species? Does it have paralogues?

Question to answer below

In which chromosome is *TNNI3* located? 19



Task

4. Search for human liver-specific genes in the Atlas.

Directions

11. To look for genes preferentially expressed in the liver, stay on the GTEx page as in previous tasks and use the menu on the left to refine your query. Since you're not looking at one particular gene this time, make sure to remove previous genes searched

so the “Genes” box is empty and you click on "Apply". Check the “Most specific” box to promote genes preferentially expressed in liver to the top of the heatmap (if “Most specific” is unchecked, the heatmap will be sorted purely by expression level of genes in liver without considering expression levels in other tissues). Set 0.5 as the "Expression value" cutoff. Use the “Organism parts” box to select for "liver" only. Then click the button “Apply” that is found below the "Genes" box.

Question to answer below

According to the “GTEx” experiment, which is the most strongly expressed liver-specific gene?

Exercise 4. Differential Expression Atlas – Finding differentially expressed genes in cow mammary epithelial cells challenged by heat-inactivated bacterial antigens

Scenario

You study milk production in dairy cows, and came across a published study which looks at how gene expression in immune response pathway changes in cow mammary epithelial cells exposed to heat-inactivated *E. coli* and *S. aureus* antigens:

- *“Comparative kinetics of Escherichia coli- and Staphylococcus aureus-specific activation of key immune pathways in mammary epithelial cells demonstrates that S. aureus elicits a delayed response dominated by interleukin-6 (IL-6) but not by IL-1A or tumor necrosis factor alpha.”* (http://europepmc.org/abstract/MED/21115717)

Microarray data from the study was originally deposited at NCBI Gene Expression Omnibus (GEO). ArrayExpress imported the data set from GEO and analysed it in the Expression Atlas: http://www.ebi.ac.uk/gxa/experiments/E-GEOD-25413 **Task** Navigate around the data set and try to answer the following questions: **1.** How many “comparisons” are there in this experiment? **Directions 1.** Open the Atlas page for this experiment: http://www.ebi.ac.uk/gxa/experiments/E-GEOD-25413. **2.** Note that in the main table, rows are genes and columns are comparisons. **Question to answer below** How many “comparisons” are there in this experiment? (ANSWER: 8)

(Exercise 4 - continued)

Task

2. Which samples are consistently used as the “reference” group in comparisons, and why?

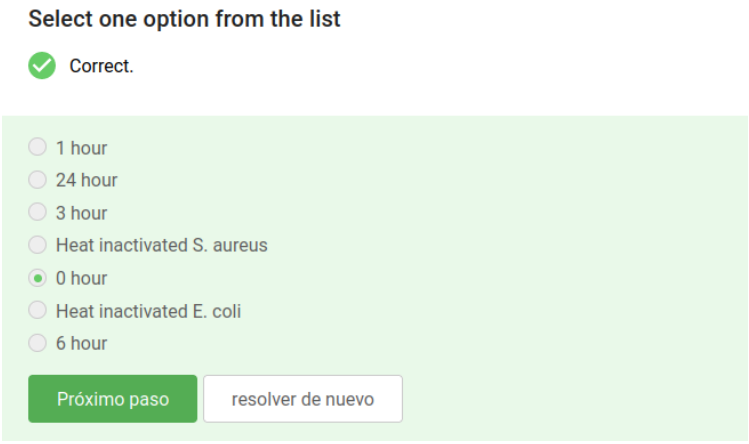
Directions

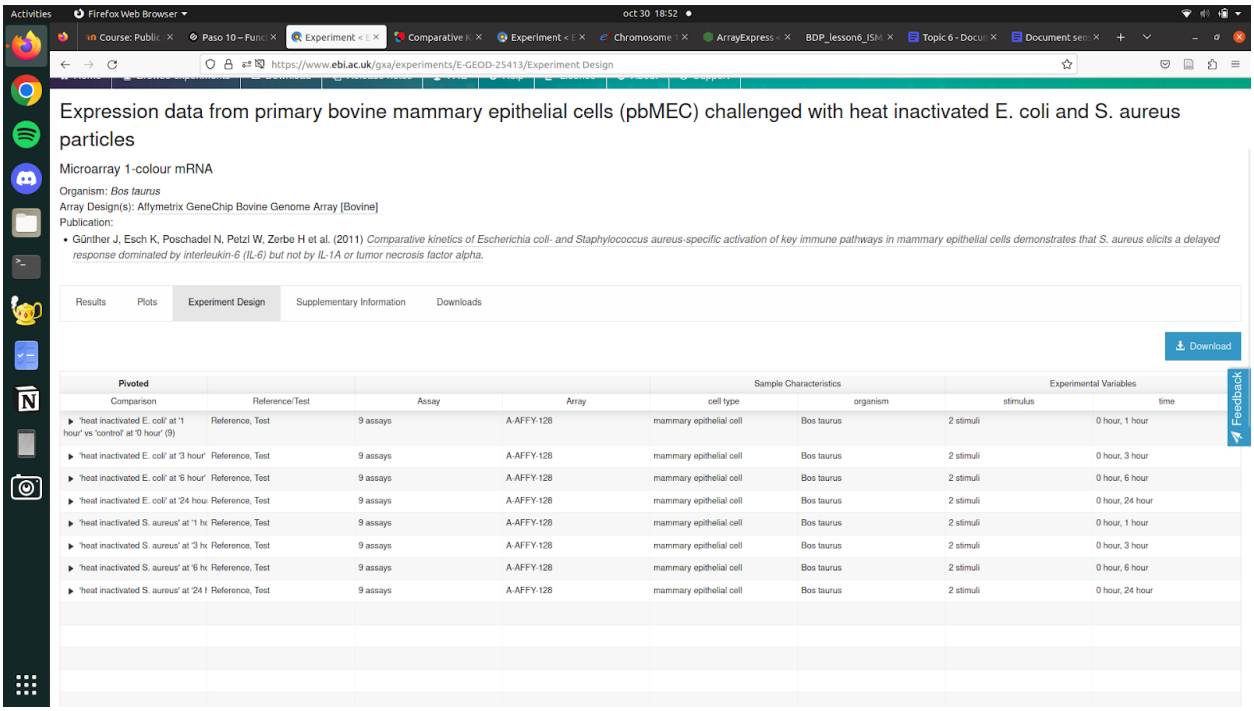
3. To find out about the categorisation of samples into “reference” and “test” groups, check the “Experiment Design” page.

4. Check the "Comparison" column. Text is in the form: "Test conditions" vs "Reference conditions" (number of assays).

Question to answer below

Which samples are consistently used as the “reference” group in the comparisons?





(Exercise 4 - continued)

Task

3. For comparison “heat inactivated S. aureus at 24 hour vs control at 0 hour”, at adjusted p-value cut-off of 0.05 and Log2 fold-change cut-off of 1.0, which gene is most strongly upregulated? What is the Log2 fold-change for that gene?

Directions

5. To explore upregulated genes in one comparison, go back to the “Results” page.

6. Use the form on the left to select “Upregulated only”, 1.0 as the Log2 fold-change cut-off, and 0.05 as the adjusted p-value cut-off. Click the button "Select" in the "Comparisons" section and leave “heat inactivated S. aureus at 24 hour vs control at 0 hour” as the only selected option. Make sure the “Most specific” box is checked, and click “Apply”.

7. In the heatmap, genes are listed in the order of Log2-fold change, with the most strongly

upregulated gene at the top.

8. Hover over your mouse to reveal the foldchange.

Question to answer below

The first gene listed has a log2-fold change of 3. What is the name of this gene?

(Exercise 4 - continued)

Task

4. For the most strongly upregulated gene identified in the previous task, can you find out more information about it, *e.g.* its full gene name, or its expression profile in normal cow tissues?

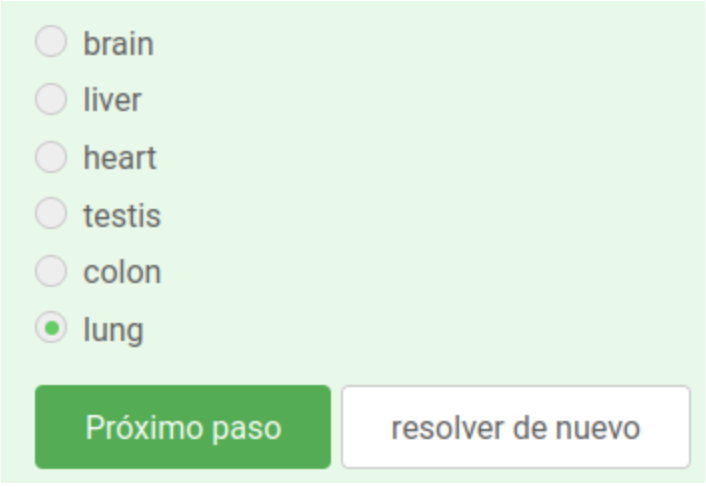
Directions

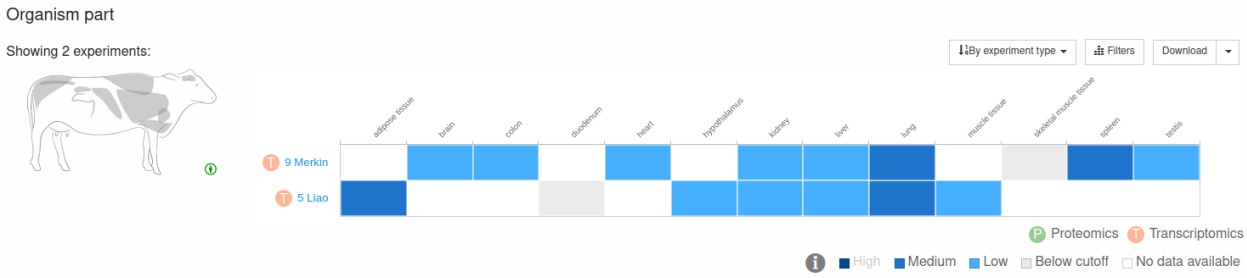
9. Click at the gene link in the heatmap to get to the gene information page.

10. From here you can link to external databases by clicking the corresponding links, or check the expression profile in normal cow tissues by clicking the tab "Baseline expression".

Question to answer below

In which tissue the gene is expressed at the highest level in normal cows?





(Exercise 4 - continued)

Task

5. Related to task 3, for genes which are up/downregulated in the comparison (keeping the adjusted pvalue and Log2 fold-change cut-offs unchanged), could you find the Gene Set overlap results plotted as a graph for Gene Ontology (GO) terms? From the graph, what is the predominant function among the differentially expressed genes?

Directions

11. Go back to the results page of task 3.

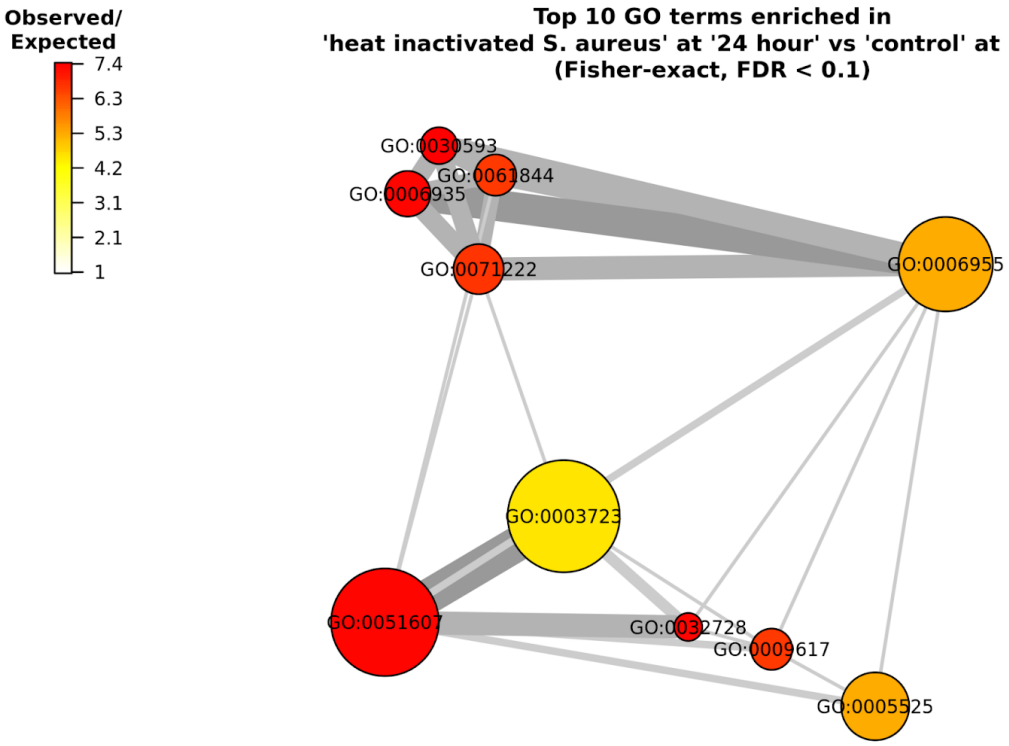
12. Use the form on the left to select “Up- or downregulated”, and leave all other options as in task 3.

13. Click the tab "Plots".

14. Look for the appropriate comparison (“heat inactivated S. aureus at 24 hour vs control at 0 hour”) and click the corresponding link "GO terms enrichment analysis". The size of the circles represent the number of genes matching the GO term, while the color represent significance (the darker the color, the less likely the overlap is caused by pure chance).

Question to answer below

What is the predominant GO function among the differentially expressed genes (biggest circle with dark red)?



GO:0003723 - RNA binding
GO:0005525 - GTP binding
GO:0006955 - immune response
GO:0009617 - response to bacterium
GO:0061844 - antimicrobial humoral immune response mediated by antimicrobial peptide
GO:0071222 - cellular response to lipopolysaccharide
GO:0032728 - positive regulation of interferon-beta production
GO:0051607 - defense response to virus
GO:0006935 - chemotaxis
GO:0030593 - neutrophil chemotaxis

ANSWER: DEFENSE RESPONSE TO VIRUS

Exercise 5. Differential Expression Atlas – Comparing samples from alopecia areata patients against healthy controls

Scenario

Alopecia areata is a common autoimmune disease which attacks hair follicles in the body. Usually the follicles in the scalp are affected, resulting in patches of baldness. Prevalence is higher among teenagers and young adults.

Task

Use the Expression Atlas database (http://www.ebi.ac.uk/gxa/) to discover experiments which study differential expression between alopecia areata and healthy individuals. Search for "alopecia areata" under “Biological conditions” while filtering by the species "Homo sapiens".

It turned out there is only one experiment in the Atlas studying alopecia: “Human Alopecia Areata Skin Profiling” so all search results come from a single experiment. The results are ranked by the magnitude of change in expression, *i.e.* the most strongly up/down-regulated genes are at the top.

1. Explore the top genes in the list (those with a Log2-fold change > |4|). In which chromosome are all these genes located?

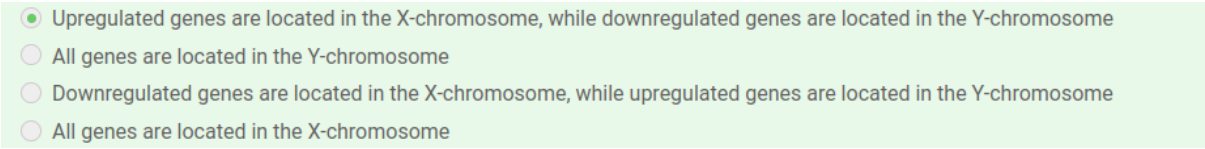
Directions

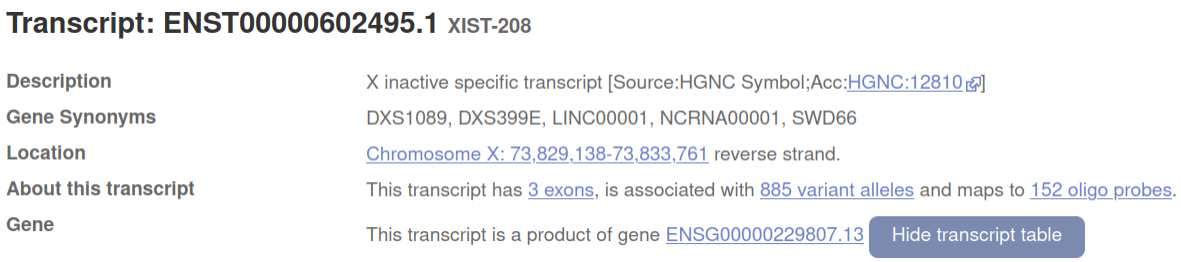
1. Open the Atlas homepage, http://www.ebi.ac.uk/gxa/. Select “*Homo sapiens*” as "Organism", type “alopecia areata” in “Biological conditions” (choose from suggestions), and click “Search”.

2. Click the gene symbol from each of the top genes in the list (those with a Log2-fold change > |4|), go to the gene information tab and link to Ensembl to find their chromosome location.

Question to answer below

In which chromosome are all these genes located?





(THIS IS UPREGULATED)

(Exercise 5 - continued)

The most plausible explanation for this is that the patients are mostly female and the controls are mostly male. In other words, we have inadvertently picked up gender-specific differences in expression, which probably has little to do with alopecia.

Task

2. Can you find out what kind of skin samples were used, and what's the sample size? Has the study been published?

Directions

3. Click at any comparison (they all say *'alopecia areata' vs 'unaffected'*) to get to the experiment's page. You can find a link to the publication on top of the page, at Reference(s).

Question to answer below

What is the Pubmed ID (PMID) of the published study?

ANS:

(Exercise 5 - continued)

The samples are from the scalp and there are 5 samples per group.

Task 3. While on the experiment's page, for the single comparison *'alopecia areata' vs 'unaffected',* can you find the Gene Set overlap results plotted as a graph for Gene Ontology (GO) terms? From the graph, what is the predominant function among the differentially expressed genes? Does it make sense?

Directions

5. To see results of the Gene Set Overlap analysis with gene ontology terms, click the tab "Plots". Then click "GO terms enrichment analysis". The size of the circle reflects the number of
   genes matching the GO term, while color reflects significance (the darker the color, the less likely that the overlap is caused by pure chance).

Question to answer below

In general, what is the predominant function among the differentially expressed genes?

ANS: Functions related to immune response, consistent with alopecia being an autoimmune disease