Practical 7 (BS312 – Bioinformatics)

Functional annotation

In the first practical session of this module you already met some gene annotation tools. In this session you will be using these and other tools to extract biologically meaningful information from the bioinformatic analyses we conducted on Down's syndrome expression datasets. Previously, you identified a set of genes in chromosome 21 that were differentially expressed between trisomy-21 and non-trisomy-21 individuals. If you did the previous Tasks correctly, you will have found three genes underexpressed in Down's syndrome (LINC01436, CBR3-AS1 and MIR3687-1) and one overexpressed gene (ITGB2). Now, you're gonna find out what are these genes and what is their known function. Later, you will analyzed a bigger dataset from a whole genome analysis of differentially expressed genes. Our goal is to extract biologically meaningful information from the bioinformatic analyses we conducted in the previous practicals.

1 Annotation of genes and gene products

There are hundreds of databases storing functional information of genes and their products. Some are general databases, some other are more specialized. A good starting point is 'Gene' at the National Center for Biotechnological Information (Table 1). Now, let's find out what the genes you identified are doing.), which compiles information from other databases about human genes. However, if you work with a different species, you may need to use a different database (see Table 1). Now, let's find out what the genes you identified are doing.

Organism	Database
All	`http://www.ncbi.nlm.nih.gov/gene`
Humans	`http://www.genecards.org/`
Mice	`http://www.informatics.jax.org/`
Saccharomyces	`http://www.yeastgenome.org/`
Drosophila	`http://flybase.org/`
Arabidopsis	`http://arabidopsis.org/`

Task 1: What are these genes doing?

1. Open a web-browser and go to `http://www.ncbi.nlm.nih.gov/gene`.
2. Write 'ITGB2' in the search fiels and click on 'Search'. If more than one result appear, click on the first one.
3. The 'summary' and 'genomic context' sections contains valuable information. Read it.
4. Open three more web-browser windows (or tab) with `http://www.ncbi.nlm.nih.gov/gene`, and look for the other three genes of interest.
5. Explore the outputs and answer the questions in the worksheet.

The first thing you will have noticed is that not all genes encode proteins. One is a microRNA, a small RNA regulatory sequence. Two of them are long non-coding RNAs, one of them an antisense transcript of a protein coding gene.

2 Gene Ontology and other annotation resources

Our analysis of gene expression of Down's syndrome is based on a reduced dataset, and only on genes mappin to chromosome 21. If we had enough time we would have analyzed more datasets and the whole human genome. Indeed, that's what the authors of the paper from which we obtained the expression datasets did. To have a more realistic picture of whcih genes are

missexpressed in Down's syndrome, we will download their table of differentially expressed genes, and will annotate their results.

Task 2: Functional annotation of overexpressed genes

1. In a web-browser go to `http://go.nature.com/tnL4FN`. This is the original article we are using in these practicals.

2. Towards the end of the page you will find the Supplementary Information section. There, download the Supplementary Table 1, which is an Excel file.

3. Open the Excel file. This will look like the one you generated with DESeq2 in the previous practical.

4. Genes are sorted by FDR (False Discovery Rate), which in this context is the p-value adjusted to multiple tests. Select the rows of those genes with a FDR of 0.01 or lower.

5. Copy and paste into a new spreadsheet.

6. Sort these genes according to the 7th column (the 'log2FoldChange' you met in the previous practical session).

7. Genes with a negative 'log2FoldChange' are overexpressed in Down's syndrome. Genes with a positive 'log2FoldChange' are underexpressed.

8. Open a new web browser and go to `http://bioinfo.vanderbilt.edu/webgestalt/`. WebGestalt is one of the many available resources for gene set enrichment. From the main page you can learn more about it. You may need to register an account to use this tool.

9. Click on 'START' (top-left).

10. In 'Select the organism of interest' select 'hsapiens' (that is, human)

11. In 'Select gene ID type' select 'hsapiens__gene_symbol' (that is, that you will be using gene names instead of reference numbers).

12. In the box below the 'Upload gene list' paste the list of genes that are overexpressed in Down's syndrome.

13. Click on 'ENTER'.

14. In the 'Enrichment Analysis' tab select 'GO analysis' This will allow us to find Gene Ontology terms associated to the list of overexpressed genes (see your lecture notes for more details about it).

15. In 'Select Reference Set for Enrichment Analysis' select 'hsapiens' genome.

16. Leave the other options as they are and click on 'Run Enrichement Analysis'

17. After a few seconds a new tab in the browser will be opened. Click on the 'View Results' button.

18. You will find in red the enriched GO term for the three main ontologies (Biological Process, Cellular Component and Molecular Function). You will also find how many genes are annotated in each GO term and the p-value (adjP) associated to the enrichment.

19. Click on any red term to see which genes are annotated to it.

20. Go to the first tab in the web browser (were the main WebGestalt is).

21. In 'Enrichement Analysis' change to 'KEGG analysis'. KEGG will explore enrichment in metabolic pathways. Click on 'Run Enrichment Analysis'.

22. In the new tab, click on 'View Results'.

23. Explore the results. If you click on any term you will be directed to a graph of metabolic interactions with the overexpressed genes in red.

After this analysis we start to understand that in Down's syndrome, a series of genes involved in inflamatory response are overexpressed. But what about those genes that are lowly expressed? Let's have a look.

Task 3: Functional annotation of underexpressed genes

1. Start over a new WebGestalt window at `http://bioinfo.vanderbilt.edu/webgestalt/`.

2. Again, select 'hsapiens' and 'hsapiens__gene_symbol' in the relevant sections.

3. Paste the list of underexpressed genes from the Excell file (remember, those with a positive 'Log2FoldChange').

4. Select 'GO Analysis' in the 'Enrichment Analysis' and 'hsapiens_genome' as a reference.

5. Click on 'Run Enrichment Analysis'.

6. 'View results'.

3 Gene networks

Often, it is useful to see if whithin a list of genes any of the products interact to each other. Interaction means anything from protein-protein physical contact to genetic interactions. We will see how to use this information to extract a bit more of information from our query list. To do so, we will explore the STRING database.

Task 4: Exploring gene interactions

1. Open a web browser and go to `http://string-db.org/`.

2. There are four tab in the query box. Open the 'multiple names' tab.

3. In the box below 'multiple names' paste the list of overexpressed genes from the previous section.

4. Select 'Homo sapiens' in the organism option.

5. Click on the 'GO!' button.

6. Ignore the next window and click on the 'Continue' button.

7. Here you are the interaction network of the overexpressed genes in Down's syndrome.

8. Under 'options' select 'Hide disconnected nodes'.

9. A central cluster (subnetwork of highly connected nodes) becomes clear. Do you recognize these genes?

4 Putting all this together

We've compared the expression pattern of Down's Syndrome individuals with non-Down's people. We identified a set of differentially expressed genes. From the functional analysis in this practical we now know that Down's syndrome's patients show an overexpression of inflammatory response genes. Indeed, neuroinflammatory response is a known symptom in this disease. We discovered that this inflammatory response is triggered by at least 4 cytokines (CXCL1, CXCL2, CXCL3 and CXCL6) which form an interacting cluster with other overexpressed genes. On the other hand, there is a deficiency in the expression of genes involved in neural development, as expected. But more specifically, you have identified three specific genes whose expression is impaired: LHX2, TLX2 and LPPR4.

If you were to continue this project as a researcher, the next step will be to validate the expression of these candidate genes in more samples and different conditions. From there, you would build models that would be tested with further experiments. Eventually, this information will be used by other researchers to develop better treatments. Now you know why bioinformatics is central to biomedical research. If you want to have a career in biomedicine, you may consider become yourself a bioinformatician, and be proud of it!