**Basic Introduction to Omics Technology**

**Transcriptomics data analysis coursework**

**C135P3 Write-up Instructions**

**What you need to do:**

Analyse a transcriptomics experiment downloaded from the GEO database. Identify differentially expressed genes between appropriate samples and analyse the gene lists and individual genes in more detail.

**Answer the following questions:**

1. **Describe the aim and the experimental detail (e.g. details of the samples) of the experiment that you have analysed. (300 words max, Marks = 7.5)**

Arabidopsis that overexpress ERF96 are more resistant to necrotrophic pathogen like *Botrytis cinerea.* Necrotrophs get energy from killed cells by invading and destroying plant tissue rapidly, and then living saprotrophically on the remains. They enter unspecialised via wounds or natural openings and secrete copious lytic enzymes and toxins. They usually attack weak, young or damaged plants. Necrotrophs are controlled by jasmonate and ethylene dependent host-defence pathways.

ERF96 is one unit of APETALA2/ethylene response factor (AP2/ERF) superfamily of transcription factor (TF). AP2/ERF are integral components of signalling cascades that regulate expression of many downstream target genes related to the plants’ stress response. ERF96 is an activator for transcription by binding to GCC elements present in promoters of jasmonate (JA)-and ethylene (ET)-responsive defence genes like PDF1.2a, PR-3 and PR-4, enhancing the expression of those genes. ERF96 is mainly localised to the nucleus.

The overall aim of the experiment is to determine what gene expressions are up-regulated due to an accumulation of ERF96 and to work out what the putative direct target of the transcription factor is. This experiment used microarrays that contained the RNA for four-week old Arabidopsis overexpressing ERF96 and the WT. The wild types were GSM1404385 (rep1), GSM1404386 (rep2), and GSM1404388 (rep3). The mutant type ERF96OE was used for rep1, rep2 and rep3. RNA of leaves were extracted using Trizol and the biotinylated cRNA were prepared according to the standard Affymetrix protocol. Following the hybridisation protocol, ten micrograms of cRNA were hybridised for sixteen hours on GeneChip ATH1 Genome Array. Microarrays are a collection of DNA probes that are usually bound in defined positions to a solid surface to which sample DNA fragments can be hybridised. The amount of hybridisation detected for a specific probe is proportional to the number of nucleic acid fragments in each sample.

1. **What are the limitations in the experimental design? (100 words max Marks = 7.5)**

One limitation in the experimental design is that only four-week old Arabidopsis were used. The different developmental stages of Arabidopsis will have different expressions of genes, so the experiment limits itself by only focusing on four-week old Arabidopsis. They could have seed buds, 2- weeks and 4-weeks old Arabidopsis plants to show how effective the necrotrophic pathogens are dependent on age. Also, since these plants are not going to be damaged, the experiment will not truly represent how effective nectrophic pathogen will affect damaged plants that are in the wild, which are their natural targets.

1. **How many differentially expressed genes (fold change up and down regulated >1.5, <0.67, p<0.01) are there between appropriate treatments? (Marks = 5)**

There are 105 upregulated genes and 156 downregulated genes.

1. **Present the data in a correctly formatted table on the 10 most up-regulated and 10 most down regulated genes. (Marks = 10)**

**The 10 most upregulated genes in Arabidopsis with overexpressed ERF96**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Upregulated genes** | **Mean expression of wild types (2dp)** | **Mean expression of mutant types**  **(2dp)** | **Fold change**  **(2dp)** | **P-value**  **(4dp)** |
| AT2G43590 | 87.53 | 9889.43 | 112.98 | 0.0047 |
| AT5G44420 | 342.30 | 36994.50 | 108.08 | 0.0064 |
| AT5G43410 | 8.70 | 871.87 | 100.21 | 0.0009 |
| AT4G16260 | 133.80 | 12775.80 | 95.48 | 0.0077 |
| AT5G61160 | 145.23 | 3678.37 | 25.33 | 0.0024 |
| AT3G55970 | 7.07 | 156.63 | 22.17 | 0.0028 |
| AT4G08770 | 24.13 | 495.57 | 20.53 | 0.0089 |
| AT1G61120 | 7.20 | 129.73 | 18.02 | 0.0063 |
| AT5G67080 | 3.80 | 61.60 | 16.21 | 0.0024 |
| AT3G12500 | 105.23 | 1614.83 | 15.35 | 0.0083 |

Ordered by Fold change highest to lowest.

**The 10 most downregulated genes in Arabidopsis with overexpressed ERF96**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Downregulated genes** | **Mean expression of wild types (2dp)** | **Mean expression of mutant types**  **(2dp)** | **Fold change**  **(2dp)** | **P-value**  **(4dp)** |
| AT4G28870 | 23.03 | 1.57 | 0.07 | 0.0003 |
| AT3G44580 | 19.30 | 1.67 | 0.09 | 0.0006 |
| AT5G28340 | 19.80 | 1.80 | 0.09 | 0.0009 |
| AT4G00350 | 28.37 | 3.17 | 0.11 | 0.0088 |
| AT5G46300 | 12.23 | 1.40 | 0.11 | 0.0035 |
| AT1G07330 | 30.07 | 3.83 | 0.13 | 0.0034 |
| AT3G04370 | 39.70 | 5.30 | 0.13 | 0.0001 |
| AT4G27580 | 19.67 | 2.63 | 0.13 | 0.0005 |
| AT4G30040 | 21.67 | 2.93 | 0.14 | 0.0000 |
| AT5G54230 | 17.30 | 2.37 | 0.14 | 0.0099 |

Ordered by Fold change, lowest to highest

1. **Describe the function of 3 up-regulated and 3 down-regulated genes and describe how they relate to the experiment? (400 words max, Marks =15)**

Array element 261713\_at corresponds to locus AT1G32640, and the gene symbol from that locus is MYC2. It encodes a Basic helix-loop-helix (bHLH) DNA-binding family protein that functions in DNA binding, transcription activator activity, sequence-specific DNA binding transcription factor activity and is involved in thirteen processes. One of these processes is that it binds to an extended G-Box promoter motif and interacts with Jasmonate ZIM-domain proteins. It is located in the nucleus and expressed in twenty-six plant structures.

Array element 249052\_at corresponds to locus AT5G44420 and the gene symbol from that locus is PDF1.2. It encodes a protein called plant defensin 1.2 and while molecular function is unknown, it is involved in responses to: ethylene stimulus, jasmonic acid stimulus, jasmonic and ethylene-dependent systemic resistance. It is involved in the defence response and response to insects. It is located in the endomembrane system and cell wall, and is expressed in thirteen plant structures.

Array element 249154\_at corresponds to locus and gene symbol AT5G43410. It encodes an Integrase-type DNA-binding superfamily protein, that function in DNA binding and sequence-specific DNA binding transcription factor activity. It is involved in regulation of transcription and an ethylene-responsive factor. It is located in the nucleus.

All of the three most upregulated genes are involved in jasmonate and ethylene response pathways. The importance of this is that necrotrophic pathogens are controlled by jasmonic acid and ethylene host-defence pathways, so these genes help the plant recognise the presence of these pathogens, allowing for a faster defence response.

Array element 253652\_at corresponds to locus and gene symbol A74G30040. It encodes a Eukaryotic aspartyl protease family protein, that functions in aspartic-type endopeptidase activity and it is involved in proteolysis.

Array element 255717\_at corresponds to locus and gene symbol is AT4G00350. It encodes a MATE efflux family protein that functions in antiporter activity, (drug) transporter activity and it is involved in (drug) transmembrane transport.

Array element 261434\_at corresponds to locus and gene symbol AT1G07G50. It encodes a Leucine-rich repeat transmembrane protein kinase that function in protein serine/threonine kinase activity and ATP binding, and it is involved in transmembrane receptor protein tyrosine kinase signalling pathway and protein amino acid phosphorylation.

The downregulated genes have less information due to the lower expression level overall. Some of the function of the proteins are hypothetical or just best matches. Also, this experiment aim does not focus on downregulated genes.

1. **What are the overrepresented GO terms for the up-regulated and down-regulated genes and how do they relate to the experiment? (400 words max, Marks = 20)**

**Up-regulated**

**Normed to Freq. in arabidopsis set (± *bootstrap StdDev*, p-value)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Biological process* | | | | |
|  | 3.15 | *0.363* | **3.087e-13** | response to stress (Input set freq.: 0.41; 0.13) |
|  | 2.4 | *0.373* | **3.103e-06** | response to abiotic or biotic stimulus (Input set freq.: 0.27; 0.11) |
|  | 2.18 | *0.314* | **1.928e-05** | other biological processes (Input set freq.: 0.27; 0.12) |
|  | 2.15 | *0.506* | **2.359e-03** | signal transduction (Input set freq.: 0.14; 0.06) |
|  | 1.36 | *0.103* | **3.087e-04** | other metabolic processes (Input set freq.: 0.6; 0.43) |
|  | 1.31 | *0.097* | **8.167e-04** | other cellular processes (Input set freq.: 0.6; 0.46) |
|  | 0.23 | *0.086* | **4.090e-07** | unknown biological processes (Input set freq.: 0.05; 0.24) |
| *Molecular function* | | | | |
|  | 1.84 | *0.245* | **3.223e-05** | other enzyme activity (Input set freq.: 0.36; 0.19) |
|  | 1.6 | *0.201* | **3.839e-04** | other binding (Input set freq.: 0.38; 0.23) |
|  | 1.38 | *0.29* | **0.041** | hydrolase activity (Input set freq.: 0.16; 0.11) |
|  | 0.64 | *0.2* | **0.038** | DNA or RNA binding (Input set freq.: 0.09; 0.14) |
|  | 0.57 | *0.135* | **3.632e-03** | unknown molecular functions (Input set freq.: 0.14; 0.24) |
|  | 0.33 | *0.183* | **0.044** | nucleic acid binding (Input set freq.: 0.01; 0.05) |
| *Cellular component* | | | | |
|  | 3.42 | *1.167* | **5.372e-04** | cell wall (Input set freq.: 0.09; 0.02) |
|  | 2.24 | *0.359* | **5.349e-05** | extracellular (Input set freq.: 0.23; 0.1) |
|  | 2 | *0.443* | **2.563e-03** | cytosol (Input set freq.: 0.16; 0.08) |
|  | 1.41 | *0.158* | **2.332e-03** | other cytoplasmic components (Input set freq.: 0.41; 0.29) |
|  | 1.32 | *0.204* | **0.014** | other intracellular components (Input set freq.: 0.33; 0.25) |
|  | 0.79 | *0.122* | **0.025** | nucleus (Input set freq.: 0.28; 0.35) |

**Down-regulated**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Normed to Freq. in arabidopsis set (± *bootstrap StdDev*, p-value)** | | | | |
| *Biological process* | | | | |
|  | 1.42 | *0.369* | **0.048** | signal transduction (Input set freq.: 0.09; 0.06) |
|  | 1.4 | *0.321* | **0.036** | transport (Input set freq.: 0.12; 0.08) |
|  | 0.92 | *0.083* | **0.048** | other cellular processes (Input set freq.: 0.42; 0.46) |
|  | 0.87 | *0.089* | **0.026** | other metabolic processes (Input set freq.: 0.38; 0.43) |
|  | 0.69 | *0.148* | **0.037** | other biological processes (Input set freq.: 0.08; 0.12) |
|  | 0.6 | *0.181* | **0.016** | response to stress (Input set freq.: 0.08; 0.13) |
| *Molecular function* | | | | |
|  | 1.63 | *0.407* | **0.032** | transporter activity (Input set freq.: 0.08; 0.04) |
|  | 1.44 | *0.299* | **0.030** | nucleotide binding (Input set freq.: 0.12; 0.08) |
|  | 1.36 | *0.222* | **0.021** | transferase activity (Input set freq.: 0.18; 0.13) |
|  | 0.59 | *0.158* | **0.010** | DNA or RNA binding (Input set freq.: 0.08; 0.14) |
|  | 0.46 | *0.296* | 0.124 | other molecular functions (Input set freq.: 0.01; 0.02) |
| *Cellular component* | | | | |
|  | 1.49 | *0.257* | **5.803e-03** | plasma membrane (Input set freq.: 0.2; 0.13) |
|  | 1.24 | *0.18* | **0.020** | other membranes (Input set freq.: 0.28; 0.22) |
|  | 0.72 | *0.13* | **0.011** | other intracellular components (Input set freq.: 0.18; 0.25) |
|  | 0.5 | *0.156* | **2.752e-03** | chloroplast (Input set freq.: 0.07; 0.14) |
|  | 0.49 | *0.212* | **0.022** | cytosol (Input set freq.: 0.04; 0.08) |
|  | 0.47 | *0.136* | **3.834e-03** | mitochondria (Input set freq.: 0.06; 0.12) |
|  | 0.37 | *0.15* | **0.010** | unknown cellular components (Input set freq.: 0.02; 0.07) |

The experiment is focusing on upregulated genes and not downregulated genes. Apart from the unknown biological processes, most of the upregulated genes are involved in some form of response to stress or abiotic/biotic stimulus, which matches with our experiment as necrotrophic pathogens will act as biotic stimulus and will provide stress to the plant. The sooner the plant can respond to the stimulus, the sooner it can activate its defence pathway, so this is good for the plants’ overall resistance. In terms of molecular function, binding, whether that’s to nucleic acids like DNA or RNA or others, is very common and so is hydrolase activity; there is also a lot of unknown. This matches with what is known about ERF96 binding to GCC elements in the promoters of genes. Hydrolase activity can be explained by removing the binding to act as a signal to stop the defence process from harming the plant by overexpression. Transporter activity can be explained by the protein being expressed in multiple locations and this may correlate with binding protein/hormone-dependent transport systems in Arabidopsis. Finally, the cellular component tells us that most of the genes are present in the cell wall, extracellular, cytosol and nucleus. This information is critical to the experimental aim of figuring out the direct target of the transcription factor and we can assume most of the proteins end up being located in the cell wall/extracellular as that is the first point of contact between the pathogen and the plant, allowing for the quickest signalling response.

1. **What conclusions can be drawn from your analysis about the experiment? (200 words max, Marks = 20)**

So, one aim of this experiment was to know which genes were upregulated due to ERF96 and 105 upregulated genes have been identified. From the overrepresented GO terms, we can see that most of these genes are involved in response to stress or abiotic/biotic stimulus and are involved in binding which fits with what we know about ERF96 binding to GCC elements in response to necrotrophic pathogen. In order to find out the putative direct target of the transcription factor, the GO terms for the cellular component highlights the nucleus, cytosol, extracellular and cell wall. From the analysis of the three up-regulated genes, we can see that some of the genes may be located in the nucleus but are expressed in multiple structures. As the genes are involved in signalling and responding to stimuli, having the proteins be expressed on the outer part of the plant such as the cell wall means that the plant will recognise the necrotrophic pathogen as soon as possible, allowing for the quickest signalling response. In general, you could say that the direct target of the transcription factor is the cell wall, as this fits with the general knowledge around ERF96 and our data.

1. **How would you confirm the role of a target gene from this analysis in the experimental system? (200 words max, Marks = 15)**

To confirm the role of a target gene, you can see how the organism acts when that gene is not present by studying mutants that have the corresponding deletion in the organism’s nucleotide sequences. Firstly, a genetic screen for isolating mutants of interest would occur in order to identify the target gene responsible for the phenotype expressed that differs from the original plant. By using reverse genetics, you would try to find the target gene by searching for homologous sequences and working out when and where the target gene is expressed.

Gene tagging is one method you could use to create mutants by insertion of a DNA tag. One reason would be that it is highly successful in Arabidopsis, as there is a 95% probability insert for every 6kb of Arabidopsis DNA. To get homozygous mutants you would have to screen 2nd generation lines. By using knockouts, you can identify the role of a target gene. The insertion junctions have been sequenced so you would screen the database for insertions which map within the target gene. You would then order the corresponding seeds online and do a PCR to screen the gene pool for the correct presence of insert within the gene. This process is called reverse genetics.

**Deadline: Friday 30rd November via Moodle**