HUTTON OPEN SCIENCE AWARDS

Best Example of a Hutton Project Team Opening Up Data for Scientific or Stakeholder Use

EORNA, a Barley Gene and Transcript Abundance Database

Linda Milne, Craig Simpson, Micha Bayer (ICS/CMS)

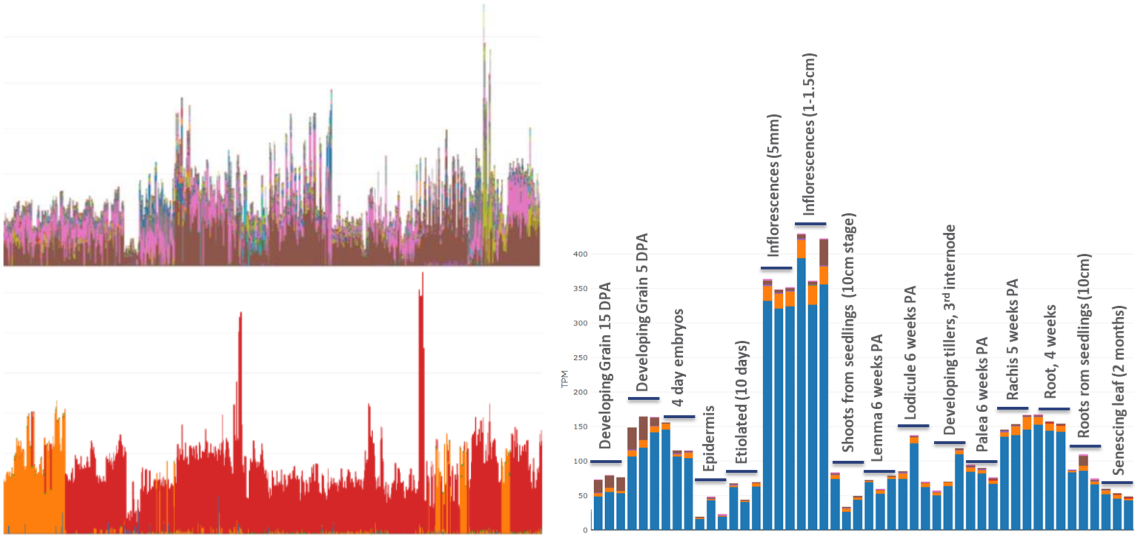
RNA-seq is the primary method to examine and quantify all the expressed gene messages in a biological sample tissue. RNA is extracted from cells and sequenced, and after initial analysis the raw sequence data is deposited in publicly available archives, where it tends to remain unused. Nevertheless, the data is available and valuable for comparative meta-analysis. The RESAS Yr6 project 'Open access to barley genome and gene expression data' successfully developed scripts that created a pipeline to quantify gene expression from public RNA-Seq data, as well as a database and website that display gene and transcript abundance data on demand.

Gene expression analysis is the gateway to understanding phenotype, organ development and environmental response. The analysis of raw RNA-Seq data requires specialist bioinformatics skills. In most cases, access to substantial computational resources is prohibitive as the data volumes for analysis on personal computing equipment are too large. The team recognised that our bioinformatics tools, computational resources and skills may be applied to not only our own in-house datasets but extended to publicly available datasets. Raw sequence data formerly buried in public archives is now accessible to all and we have added value to this by quantifying gene expression from it.

The EORNA project goes beyond standard expectations of open science by reusing vast volumes of underused public data, adding value to it through analysis, and returning the results to the scientific community at no cost and in an easily accessible format. To date, the EoRNA website has recorded approx. 87,300 hits from nearly 4,000 researchers from over 70 countries, demonstrating substantial interest from barley researchers and others outside the barley community. Similarly, free access to the scripts through github for development of a similar database and visualisation we know is being used for other species.

The expansion in data volume, data access and data presentation challenged the team to develop the new tools needed to present this data in a biologically relevant way and link it to relevant genomic data. We learned that experimental metadata is not submitted to public archives in a consistent manner that provides clear information about the underlying experiment. In addition, we need to determine the effect of experimental differences between the datasets. Recent further development of our tools will now allow us to expand the numbers of experimental datasets 4-fold to provide comparison of an even wider range of conditions.

Linda Milne, Craig Simpson, Micha Bayer
https://ics.hutton.ac.uk/eorna/index.html



Explanation of Image: The two images on the left represent the expression of gene and transcript abundances from 22 RNA-seq experiments, covering 843 separate samples. There are over 30,000 genes in the database and the images on the left represent the expression patterns for two of these genes. Each peak represents an RNA-seq sample, and the height of each peak represents the gene abundance. Uniquely, EORNA shows the abundances of the different transcripts for each gene represented by different colours. The gene and transcript expression values that generate these graphic images may be downloaded for further analyses. The image on the right represents a zoomed in part of a graphic image to identify 1 of the 22 experiments showing 16 different barley tissues and the presented metadata that identifies each sample.