



Trans-Learn: machine learning for multivariate marker detection

Joshua Colmer¹, Mie Honjo², Anthony Hall¹, Hiroshi Kudoh², Ji Zhou³

1 – Earlham Institute, Norwich, UK, 2 – Center For Ecological Research, Kyoto, Japan, 3 – National Institute of Agricultural Botany, Cambridge

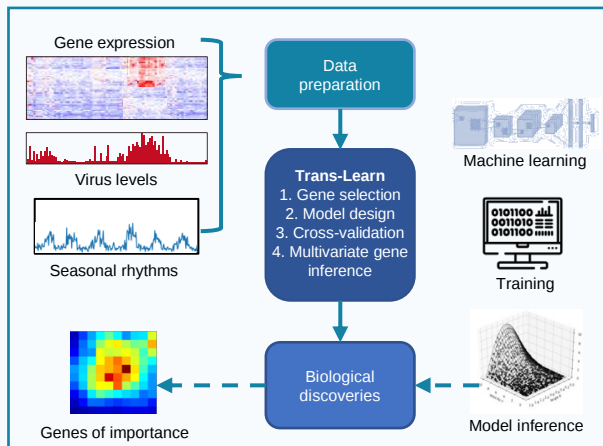


This work was supported by the UKRI-BBSRC Norwich Research Park Biosciences Doctoral Training Partnership  **Doctoral Training Partnership**

1. Introduction

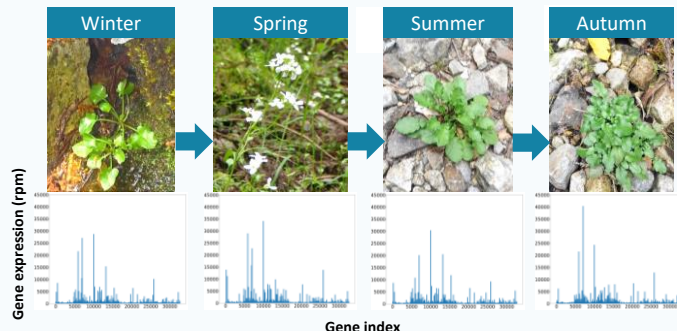
Differential expression analysis is effective at identifying individual genes associated with a condition but unable to identify joint or multivariate relationships.

Trans-Learn is a machine learning based platform with prediction and inference functions designed to identify multivariate marker genes. Through powerful modelling techniques, we identify multivariate marker genes in *Arabidopsis* for pathogen presence, some of which are **not** differentially expressed.



2. Seasonal Data Collection

- Between 2011 and 2017, *Arabidopsis halleri* in-field tissue sampling was conducted in long-term study sites, central Honshu, Japan
- To include a range of environmental variation, seasonal and diurnal data was collected
- Turnip mosaic virus (TuMV) abundance and transcriptome datasets obtained through RNA-Seq
- Gene expression (reads per million, rpm) and virus abundance (rpm) were quantified relative to the total reads of host mRNA
- Aim is to apply Trans-Learn to these datasets to identify multivariate gene markers for pathogen presence and abundance





Trans-Learn: machine learning for multi-variate marker detection

Joshua Colmer¹, Mie Honjo², Anthony Hall¹, Hiroshi Kudoh², Ji Zhou³

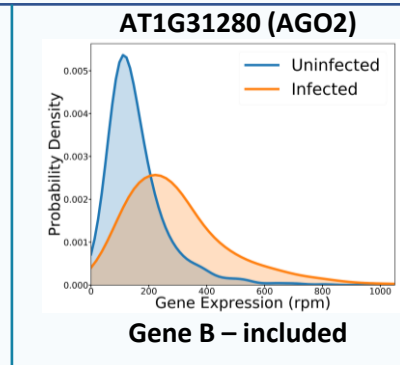
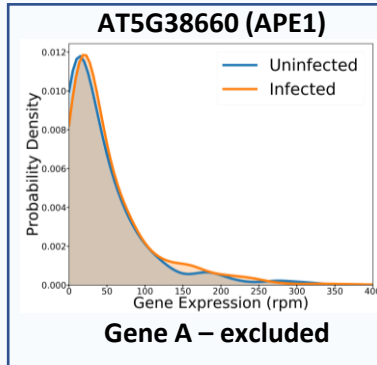
1 – Earlham Institute, Norwich, UK, 2 – Center For Ecological Research, Kyoto, Japan, 3 – National Institute of Agricultural Botany, Cambridge



This work was supported by the UKRI-BBSRC Norwich Research Park Biosciences Doctoral Training Partnership  **Doctoral Training Partnership**

3. Gene Selection

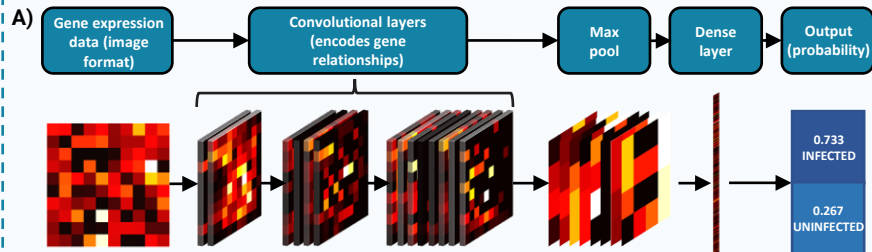
- High dimensional (30,000+ genes) gene expression datasets require feature selection to remove genes not dependent on the pathogen
- We calculate the mutual information between gene expression and the presence of the pathogen
- This ranks genes based on their ability to distinguish between infected and uninfected samples
- Genes unable to distinguish between the two classes of samples are excluded based on a specified threshold



4. Novel CNN Approach

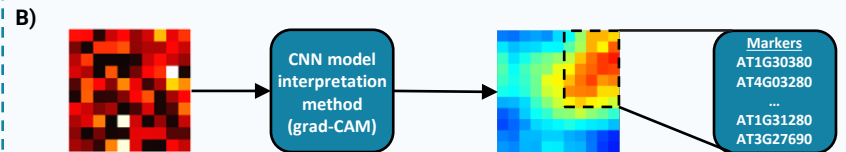
A) Model

- After selecting genes, we encode the expression data in 2D image format
- We use a convolutional neural network (CNN, typically applied to images) to predict probability of pathogen presence



B) Interpretation

- We utilise a published method known as grad-CAM to interpret the encoded gene relationships in our CNN model
- Heatmaps produced displaying clusters of jointly important genes





Trans-Learn: machine learning for multivariate marker detection

Joshua Colmer¹, Mie Honjo², Anthony Hall¹, Hiroshi Kudoh², Ji Zhou³
 1 – Earlham Institute, Norwich, UK, 2 – Center For Ecological Research, Kyoto, Japan, 3 –
 National Institute of Agricultural Botany, Cambridge

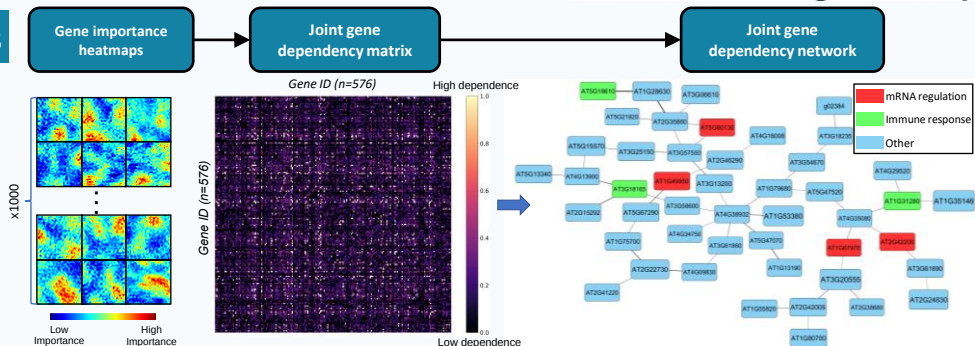


norwich research park
 Doctoral Training Partnership

This work was supported by the UKRI-BBSRC Norwich Research Park Biosciences Doctoral Training Partnership

5. Predictive interactions

- Aggregate 1000+ gene importance heatmaps in a dependency matrix
- Perform clustering on the joint gene dependency matrix to identify clusters of multivariate markers
- Visualise clusters of jointly dependent marker genes, colour-coding each gene by function
- Each cluster contains a set of genes that are highly predictive when analysed together but not necessarily individually



6. Multivariate markers

- The joint mutual information contained in a subset of multivariate markers detected by Trans-Learn is on average higher than a subset of differentially expressed (DE) genes
- This suggests non-differentially expressed genes may have multivariate relationships with a condition that are invisible to univariate analysis
- Trans-Learn is currently being used to identify genes associated with subtypes of breast cancer as well as patient survival time
- A graphical user interface (GUI) has been designed so no programming is required for use

