DIAGNOSTIC META–ANALYSIS FOR DIFFERENTIATION OF VIRAL AND BACTERIAL BIOMARKERS

*Ashleigh C Myall, Simon Perkins, David Rushton, Jonathan David, Phillippa Spencer, Andrew R Jones, Philipp Antczak**
Computational Biology Facility, Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom

INTRODUCTION

A fundamental problem for disease treatment is that while antibiotics are a powerful counter to bacteria, they are ineffective against viruses. Without proper identification equipment, medical practitioners can unnecessarily prescribe antibiotics to patients without bacterial infections, leaving the source disease untreated and fostering anti-microbial resistance. This was highlighted by two defence organisations, the Defence Threat Reduction Agency (DTRA) and Defence Science Technology Laboratories (DSTL) as particularly hazardous for soldiers without access to proper means of diagnosis. Suitably, work was commissioned as a basis to assist in the design of a future, portable diagnosis device that could differentially diagnosis disease state based on profiles of gene expressions.

We conducted a meta-analysis of human blood infection studies using Machine Learning driven biomarker discovery to obtain a panel of genes enabling differential diagnosis of disease state. We presented an optimal panel of genes, as well as some similarly performing alternatives, which are robust and generalisable, enabling technology independent differential diagnosis of disease state. We also provide direct insight into the underlying gene interaction networks which control response to disease, verified by two parallel analyses on Affymetrix and Illumina RNA Microarray cohorts.

METHODS

1. Data Gathering
- Bacterial and viral human blood infection studies were selected and normalised. Included studies covered a variety of both strains and platform versions, but came exclusively from Affymetrix or Illumina RNA-microarrays.

2. Pre-processing
- Technical differences between studies and the need for larger, combined datasets required merged datasets to be batch corrected using a novel multi-step batch correction process.

3. Feature Selection
- Panels of genes enabling differential diagnosis of disease state were obtained using 2 feature selection algorithms: (1) a Genetic search algorithm, and (2) Backward Elimination, both using the Random Forest classifier.

4. Functional Analysis
- We used DAVID [1] which provides enrichment analysis for genes over several different ontologies to check if ontology terms present in our gene lists were over-represented.

5. Infer Interaction Networks
- We created an interaction network of combined datasets to verify which regions of the genome were being selected from in our gene lists and whether there was convergence around key groups of genes.

6. Out-of-sample Testing
- To properly test the generalisability of discovered gene lists, we trained and evaluated them on unseen datasets that they were not discovered in, coming from a different microarray manufacturer.

Figure 1. Methods Flow diagram: 1) Data Gathering, 2) Pre-processing, 3) Feature Selection, 4) Functional Analysis, 5) Infer Interaction networks, 6) Out of Sample Testing.

RESULTS & DISCUSSION

1. Cohort merging and batch correction generated two cleaned data sets. One for each of the largest micrar -ray manufactures (Affymetrix and Illumina). However, some samples classification remained ambiguous, as such we created 2 instances for both Affymetrix and Illumina datasets: (1) where only confirmed classes are included, and (2) where ambiguous classes are integrated (bacterial ? -> bacterial, viral? -> viral).

2. Using the Random Forest classifiers, we discovered gene lists using Backwards Elimination and a Genetically inspired search algorithm for both instances of Affymetrix and Illumina data.

- Backward elimination (confirmed classes only) (**BW-C**).
- Backward elimination (ambiguous classes integrated) (**BW-I**).
- Genetic-algorithm optimized (confirmed classes only) (**GA-C**).
- Genetic-algorithm optimized (ambiguous classes integrated) (**GA-I**).

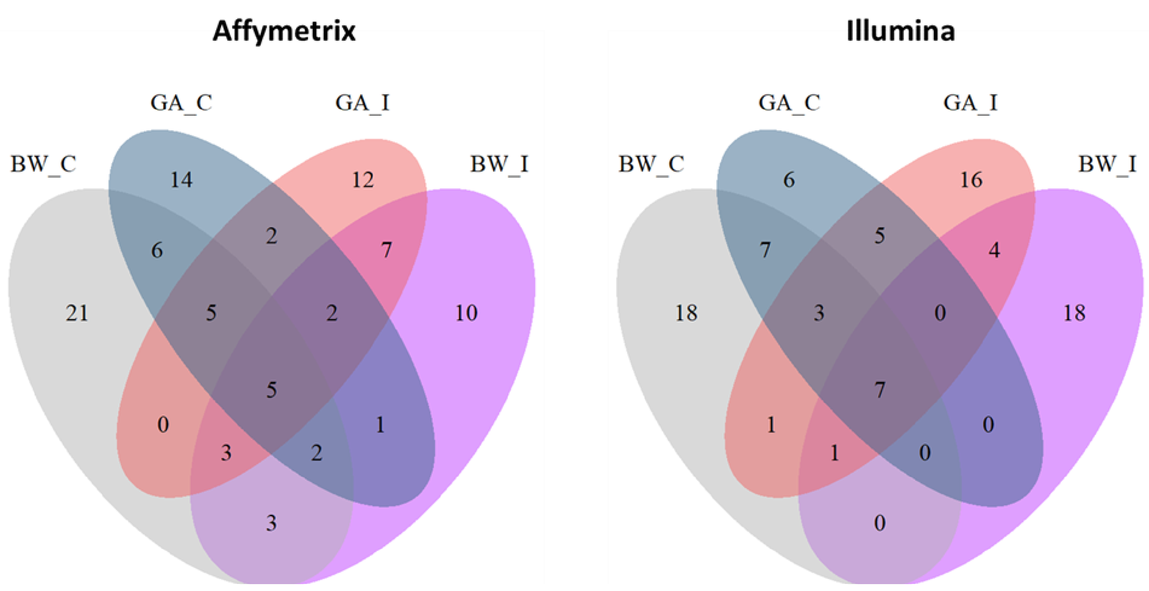

Figure 2. Overlap analysis of BW-C, BW-I, GA-I, and GA-C gene lists discovered on Affymetrix and Illumina Datasets.

3. We found a high degree of convergence amongst optimal models. In the Affymetrix dataset we discovered 5 genes important to discriminate between infection types selected by all models, out of these 2 were also found in all Illumina models (Figure 2).

RESULTS & DISCUSSION (CONT)

5. For better understanding of the underlying biology discovered by models, we performed functional enrichment analysis of the gene lists. For all models we found a number of relevant terms related to the human immune system. Particularly, we found all Affymetrix and Illumina models contained a number of Type I Interferon inducible genes (ISGs) - demonstrated to have altered expressions in disease states [2]

6. We investigated whether gene selection converged around key clusters of functionally related genes by constructing a gene interaction network (Figure 3) and found that all Affymetrix and Illumina models tended to select genes from the same areas, in fact all models contained genes from 4 sub-clusters of cluster 3, each with distinct associated roles in immune responses: (1) Type I interferon-inducible genes (ISGs), (2) Chemotaxis genes, (3) Apoptotic Processes genes, and (4) Inflammatory / Innate Response.

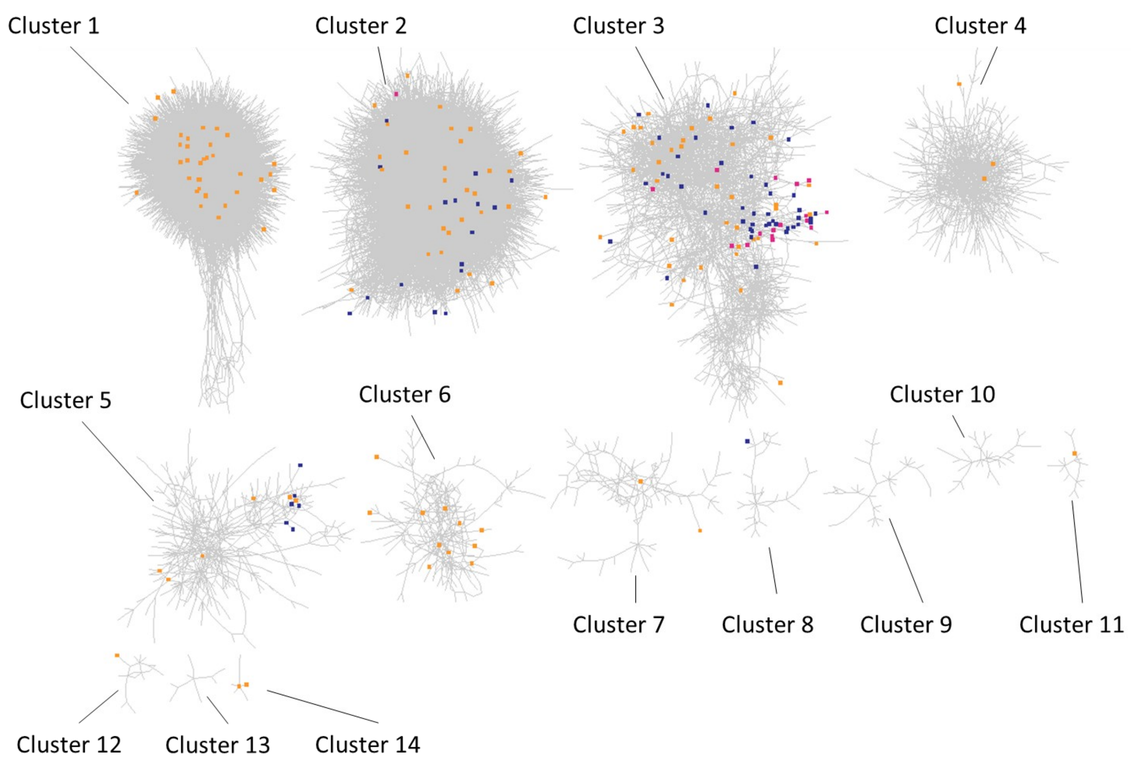

Figure 3. Illumina Interaction network after initial clustering (visualising clusters > 10 Genes), where Illumina models selected genes are blue, Affymetrix selected genes are orange, and those intersecting both manufacturers are pink.

7. In the key functions underlying our gene lists ISGs were particularly prevalent amongst all models. This family of genes had been said to link the innate and adaptive immune systems together [3]. Their interconnected nature likely makes them a good approximator for a much larger group of genes, capturing significant predictive power of the immune system.

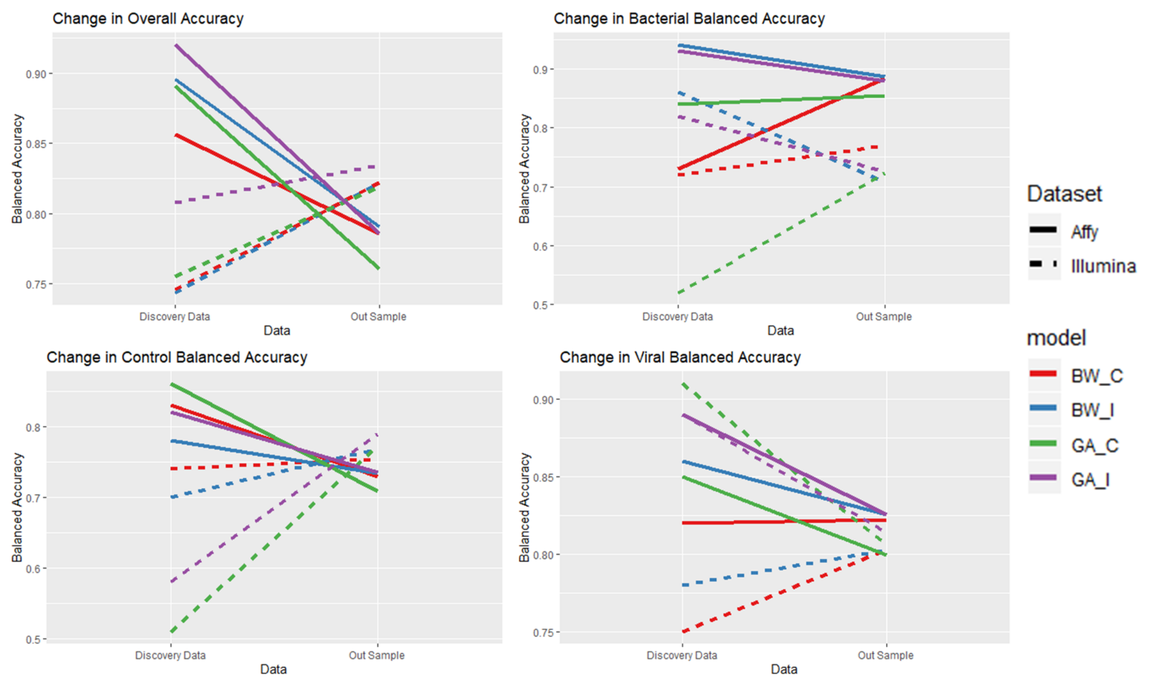

Figure 4. Affymetrix vs Illumina Gene Lists performance on Discovery Set evaluation vs Out of sample on non-Discovery Data. Models are indicated by colour and the Gene Lists discovery dataset is by the line type: Solid for Affymetrix and dotted for Illumina.

8. Out of sample performance remained high between discovery and non-discovery data, demonstrating that our gene lists generalise well to unseen data using different studies, class distribution, and platform (Figure 4).

CONCLUSION

We presented several panels of genes which are indicated to perform well on both discovery data and out of sample testing (on different microarray manufacturers). While there was some inconsistency between manufacturer gene lists, we still observed high convergence around 4 key groups of genes, most predominantly the ISGs – suggesting we have uncovered the underlying biology significant for differential diagnosis of disease state and verified through our parallel analysis. However, no gene list contained genes distinctly from these key groups, which indicates that genes from a core set of functions are required, but always in combination with a wider, more varied set of other functions to best capture disease state.

ACKNOWLEDGEMENTS

I gratefully acknowledge funding from the Computational Biology Facility (http://cbf.liverpool.ac.uk/) for my MRes degree. I also gratefully acknowledge DSTL (www.gov.uk/dstl) for funding the overall project and providing data sets.

This work was supported by the Chem-Bio Diagnostics program contract HDTRA1-12-D-0003-0023 from the Department of Defense Chemical and Biological Defense program through the Defense Threat Reduction Agency (DTRA).

REFERENCES

1. Huang da, W., B.T. Sherman, and R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc, 2009. 4(1): p. 44-57.

2. McNab, F., et al., Type I interferons in infectious disease. Nat Rev Immunol, 2015. 15(2): p. 87- 103.

3. Trunt, D. (2004). Type I Interferon as a Link Between Innate and Adaptive Immunity through Dendritic Cell Stimulation. Leukemia & Lymphoma, 45 (2), pp.257-264.

Contact: P.Antczak@liverpool.ac.uk