Identifying robust biomarkers of infection through an omics-based meta-analysis

Ashleigh C Myall[1,2], Simon Perkins[1], David Rushton[4], Jonathan David[5], Phillippa Spencer[6], Andrew R Jones[1,&], & Philipp Antczak[1,3,&]*

[1]Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom

[2]Department of Mathematics, Imperial College London, London, United Kingdom

[3]Center for Molecular Medicine, University of Cologne, Cologne, Germany

[4]Defence and Security Analysis Division, Defence Science and Technology laboratory (DSTL), Porton Down Salisbury, United Kingdom

[5]Chemical, Biological and Radiological Division, Defence Science and Technology laboratory (DSTL), Porton Down Salisbury, United Kingdom

[6]Cyber and Information Systems Division, Defence Science and Technology laboratory (DSTL), Porton Down, Salisbury, United Kingdom

*Corresponding author

E-mail: P.Antczak@liverpool.ac.uk (PA)

[&]Joint Senior Authors

Abstract

A fundamental problem for disease treatment is that while antibiotics are a powerful counter to bacteria, they are ineffective against viruses. To ensure a given individual receives optimal treatment given their disease state and to reduce over-prescription of antibiotics leading to antimicrobial resistance, the host response can be measured to distinguish between the two states. To establish a predictive biomarker panel of disease state we conducted a meta-analysis of human blood infection studies using Machine Learning (ML). We focused on publicly available gene expression data from two widely used platforms, Affymetrix and Illumina microarrays, and integrated over 2000 samples for each platform to develop optimal gene panels. On average our models predicted 80% of bacterial and 85% viral samples correctly by class of infection type. For our best performing model, identified with an evolutionary algorithm, 93% of bacterial and 89% of viral samples were classified correctly. To enable comparison between the two differing microarray platforms, we reverse engineered the underlying molecular regulatory network and overlay the identified models. This revealed that although the exact gene-level overlap between models generated from the two technologies was relatively low, both models contained genes in the same areas of the network, indicating that the same functional changes in host biology were being detected, providing further confidence in the robustness of our models. Specifically, this convergence was to pathways including the Type I interferon Signalling Pathway, Chemotaxis, Apoptotic Processes, and Inflammatory / Innate Response. Amongst and related to these pathways we found three genes, *IFI27*, *LY6E*, and *CD177*, particularly prevalent throughout our analysis.

Author summary

Bacterial and viral disease require specific treatments, and whilst there are various treatment options for specific infection types, rapid diagnosis and identification of the optimal treatment remains challenging. Even in wealthier countries with developed healthcare

systems, unnecessary prescription of antibiotics to patients with viral infections is causing phenomena such as multi-drug resistent bacteria. One way to distinguish a viral from bacterial infection is to measure an individual's responses, for example by measuring the expression of particular genes in a blood sample, as different types of infections trigger different types of responses. In our study we analysed thousands of previously collected data sets from human blood, where individuals had either viral, bacterial or no infection (control). We used machine learning to identify "signatures" – small sets of genes that are indicative of the type of infection (if any) carried by an individual. Within data sets we used two different technology platforms had been used to collect data. We demonstrated that their gene-level signatures do not overlap perfectly when derived from the different platforms, the biological networks from which those genes were derived, however, had a high overlap – giving confidence that our models are robust against technology artefacts or bias. We have identified a small set of genes that serve as strong biomarkers of infection status in humans.

Introduction

The varying differences within both classes of bacterial and viral infections cause the body to respond in a distinct way (1). Bacteria can be countered by pathways such as complement-mediated lysis, and the cell-mediated response for those that survive phagocytosis and live within the cell (intracellular bacteria). In this response, cells present bacterial peptides (antigens) on their surface, which are identifiable by Helper T cells that mediate bacterial destruction (2). There are a large variety of viruses and bacteria that affect the host's immune system in various ways. Whilst some response pathways may overlap for bacterial and viral infections, there are however a number key differences (3, 4). In fact, these different response pathways cause varied transcription (expression) of key genes and are the medium for distinguishing disease state based on the host's transcriptional response (5).

Differential expression of certain genes related to immunological responses can be indicative of both (i) disease state and (ii) individual pathogens (6). Such knowledge can be exploited in differentiating between viral, bacterial and control biological states. Previous studies demonstrated this by developing a small set of only seven genes that can accurately discriminate bacterial from viral infections across a range of clinical conditions, whilst simultaneously succeeding to determine with high accuracy which patients do not require antibiotics (7). Simultaneously, there have been numerous other studies looking at diagnosing infection based on the host's transcriptional response (8-12).

Previous work failed to generalise as the data contains a far smaller set of pathogens that would be encountered in 'real world' scenarios, or studies focussed on single technology platforms, specific pathogens, or geographical regions (which contain populations with different HLA alleles, and different local pathogen groups). To address this lack of generalisation this work aims to utilise a larger scale analysis over a more representative sample set to improve biomarker generalisability. To gain statistical power and develop more robust panels, meta analyses of publicly available data have proven to be an effective technique (13). However, analysis integrating several cohorts together face inherent limitations from systematic variations otherwise known as "batch effects". Without proper handling, these batch effects have been shown detrimental in population level gene expression analysis (14). Computational techniques exist to reduce batch to batch variation (15). ComBat (16), used in our study, is a well-known batch correction algorithm, and has been shown successful at removing batch effects between studies whilst retaining relatively high amounts of the biological variation.

Data-driven identification of robust biomarkers is a much-debated subject in the biological field. Several machine learning (ML) approaches have been proposed, with typically good performance on data sets used in a given study, but poorer performance when biomarkers are

taken forward for validation. Important is the distinction between uni- and multi-variate approaches to biomarker discovery. While identifying a single predictive marker might be preferred in theory, multi-variate approaches have enabled the discovery of more complex relationships that can provide performance (accuracy; sensitivity) far exceeding univariate predictive models (17). One particular aspect in multi-variate predictive approaches is the optimisation of the representative model, which rarely can be achieved through brute force testing and relies on feature selection algorithms. In addition, models developed by ML approaches provide a more complete understanding of the underlying biological mechanisms, adding to our understanding of these systems. In this publication we focus on the use of the Random Forest (RF) (18) classifier, which has been demonstrated to perform well in real-world classification problems with high dimensionality and biased data (19). RFs are bagged decision tree models, which classify data points on a subset of features and have been praised for their ability to avoid overfitting (20). Unlike Support Vector Machines or Neural Networks (two frequently used models with high predictive capabilities) RFs forego much of the model selection step using an ensemble approach which builds many weak classifiers into a single strong self-averaging, interpolating model (21). Whilst RFs consist of many weaker models, they have been shown highly effective at capturing non-linear relationships between model predictors and outputs in a number of genomic studies (22, 23).

In recent years bioinformatician seeking predictive models have been faced with increasingly greater dimensionality to their data. With the needs of interpretable models many have responded and used feature selection procedures, which aims to remove redundant and irrelevant model features (24). The results of a smaller feature set not only offers improving model performance, faster computational implementation, and greater interpretation of the underlying generative process (25); but moreover lines up with the original pattern recognition theory, that RFs, like many other ML models were not designed to cope with

large amounts of irrelevant features, often referred to as *the curse of dimensionality* (26). This high dimensionality is especially pronounced in the case of gene expression data with the total human gene set being ~20,000.

Various feature selection procedure exist and have been demonstrated in biological problems (24). For this study we focused on Backwards Elimination (BW) (27) forming a well-established benchmark, and an evolutionary algorithm, a more explorative and parameterizable search approach, to obtain reduced model feature sets (17). BW essentially searches for the optimal feature set by progressively eliminating the least important features from a given dataset and testing whether the new model is significantly more accurate than the previous. Whereas evolutionary algorithms are based on evolving population(s) of models, which are repetitively intermixed, and subject to random point mutations. This evolutionary process is assumed to produce converging model populations in terms of performance and their associated feature sets (28).

The application of different computational pipelines often leads to different outcomes in disease prediction (29). We believe, it is thus important not only to present performance statistics for one given model generated by an ML pipeline, but to explore the underlying biological response of a set of plausible models. By doing so, it is possible to develop a more robust biomarker panel (mitigating overfitting which would generally produce models hard to interpret biologically), and to understand why a given model, or set of similar models, are valid.

In this work, we have performed a meta-analysis over publicly available transcriptomics data (human blood samples where individuals had bacterial, viral or no infection), from two microarray technologies (Affymetrix and Illumina). We applied feature selection and machine learning for biomarker discovery and predictive model generation, and lastly we

explored the biological context of the resulting models by reverse engineering the underlying networks. Representing omics data as a network, has several key benefits. One can often better represent many complex systems as connected components, and the genome is no exception (30). Clustering is one popular method to explore these complex networks and many algorithms exist to reveal insight into these complex structure (31). Visualising a clustered network allows us to explore aspects of this generative process, and how feature selection unfolds over it. However, network construction can often be sensitive to the computational approach and parameterization applied (32, 33). In our approach, we validated our findings and mitigate any potential bias in network generation and clustering by illustrating that the biological driven feature selection is consistent across two separate networks, containing different studies, and derived from different technological platforms.

154 # Materials and Methods

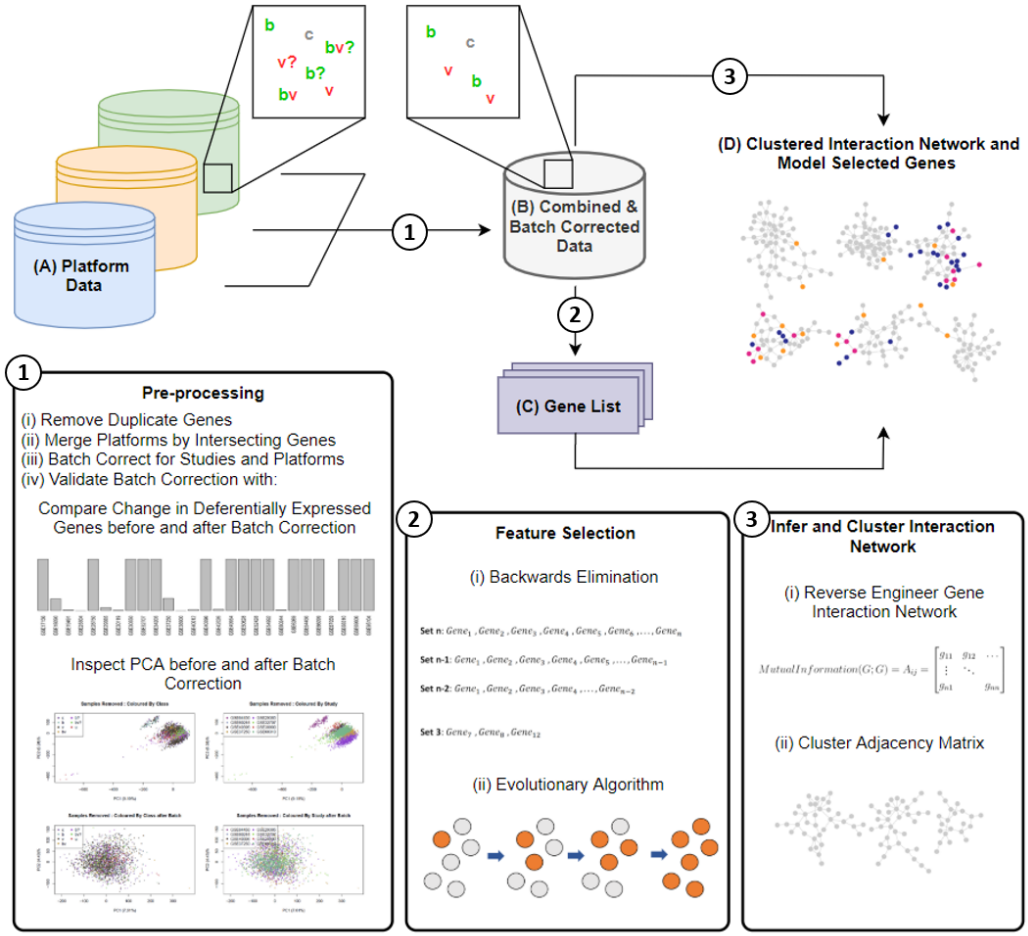

Fig 1. Conceptual overview. Individual data (A), containing bacterial (b), viral (v), control (c), and samples with lower levels of study confidence (?s) is merged by common genes and pre-processed by Step 1. Step 1 outputs a combined and batch corrected dataset (B), where only b/v/c samples are present. Two instances of (B) are formed, one where samples of lower levels of support are integrated into b/v classes, and the other completely omitting uncertain samples. Feature selection is performed on data B in Step 2 using (i) Backwards Elimination, and (ii) an Evolutionary algorithm. Step 2's output is a number of Gene Lists (C) obtained in the feature selection. Data B is also used to infer and cluster a gene interaction network, by (i) reverse engineering the gene interaction network, and (ii) clustering the adjacency matrix. (D) is then formed as the clustered interaction network overlaid with genes found in the best performing mode of each dataset and search procedure.

To identify and validate a panel of biomarkers able to differentiate bacterial and viral infections, we performed a meta-analysis of GEO gene expression data, all from open source microarray human blood infection studies. Our analysis was divided into three major method steps: i) pre-processing, ii) feature selection, and iii) inferring a gene interaction network, to discover and our validate gene lists (Fig 1). Following the major steps, we performed and report the results of a final out of sample test on data not previously used in the training phase for greater validation.

Pre-processing

Data. Datasets from four technological platforms (two from Affymetrix platforms and two from Illumina platforms), consisting of 3868 samples, from 21 different studies, were included in the analysis (Table 1). These datasets were selected from a wider pool identified in an initial scan of online databases, based on a variety of factors including: microarray platform manufacturer (most prevalent platforms – Affymetrix and Illumina) and study set size (larger studies with more predictive power), class pathogen strain distribution (aiming for an equal distribution across the data); and ability to merge with other datasets in our analysis.

Table 1. Summary of platform level Affymetrix and Illumina datasets prior to pre-processing.

Manufacturer	Platform (GPL)	Studies (GSE)	Distinct Genes	Sample Count (%)	Bacterial (b) (%)	Uncertain Bacterial (b?) (%)	Viral (v) (%)	Uncertain Viral (v?) (%)	Control (c) (%)	Other (%)
Affymetrix	GPL570	GSE49954, GSE50628, GSE54992, GSE25504, GSE66099, GSE69606, GSE6269, GSE18090, GSE28750, GSE34205	22,213	615 (100)	27 (4.4)	227 (36.9)	164 (26.7)	0 (0)	156 (25.4)	41 (6.7)
Affymetrix	GPL571	GSE52428, GSE95104, GSE17156	13,383	834 (100)	60 (7.2)	0 (0)	358 (42.9)	348 (41.7)	68 (8.2)	0 (0)

Affymetrix	GPL9188	GSE30550	13,383	268 (100)	0 (0)	0 (0)	132 (49.3)	119 (44.4)	17 (6.3)	0 (0)
Illumina	GPL10558	GSE29385, GSE32707, GSE37250, GSE40396, GSE60244, GSE64456, GSE68310	19,947	2,151 (100)	215 (10.0)	141 (6.6)	1069 (49.7)	0 (0)	467 (21.7)	259 (12.0)

Deduplicating genes by probes and merging datasets. Dataset columns (originally microarray ProbeIDs) were first deduplicated and substituted by their gene mappings. Where duplicate ProbeIDs existed for the same gene we selected a representative ProbeID with the highest average intensity across samples (34). Samples from datasets of the same manufacturer were then merged by common genes, first at the level of studies within the same platform, then by platforms in the same manufacturer.

Batch correction and evaluation. Batch corrections targeted two non-biological sources of systematic variation: (i) inter-platform study batch effects (differences between platforms), and (ii) intra-platform batch effects (differences between studies within a batch). Batch correction was implemented with 'ComBat' (16) in a two-step sequential batch correction pipeline (S1 Appendix.docx). We repeated this process for both Affymetrix and Illumina datasets separately to form batch corrected Affymetrix data, and batch corrected Illumina data.

Batch correction was verified to retain biological variation and remove technical variation using two validation steps (Fig 1 Step 1). Firstly, we tested whether pre and post batch correction significant features overlapped significantly. Secondly, we performed Principle Component Analysis (PCA) (35) visualising the data in two dimensions and comparing the PCA plots of before and after batch correction. For a successful batch correction, pre-batch correction sample clustering in the PCA would be visually removed in the PCA plot of post batch corrected data.

Dealing with study sample ambiguity: forming a confirmed and integrated dataset instance. To include more data, including some class ambiguity in the original studies, we formed a modelling dataset which integrated bacterial and viral samples with lower levels of confidence (b?, and v?)(Table 1). This integrated dataset contained only classes labeled b/v/c (Fig 1). For Affymetrix this formed (Affy_I) and similarly, for Illumina this formed Illumina_I. Two additional datasets of confirmed sample classes only, were also generated and included in the study but presented only in the Appendix.

Feature Selection. To search for optimal panels of genes we implemented two search feature selection procedures: (i) the well-known Backward Elimination process (27), and (ii) a genetically inspired search algorithm (GALGO) (17). Both search procedures operated using the RF Classifier, implemented in the R Ranger package (36) a fast and parallelisable implementation of RFs for high dimensional data.

Dataset Preparation. For dealing with un-even class distributions present in our data (Table 1) we employed two strategies. Firstly, we used a study aware data split which insured relatively equal class proportions across both training, test and evaluation data splits. Secondly, we ensured that classification accuracy bias due to larger class proportion of disease states was minimized by weighing smaller classes correspondingly higher (18). This ensures that our model will not be biased to classifying samples with a larger proportion in the dataset.

Backward Elimination. We operated on a 60/20/20 training/test/evaluation data split for each dataset processed in BW (37). On each training set we ran 240 BW search procedures, using Out-of-bag (OOB) error as the minimisation criterion and implementation using the VarSelRF R package (38). Each run generated a single optimal model which minimised OOB.

For each dataset a single representative model was selected from the 240 runs which maximised accuracy on test data.

Genetic-algorithm. The Genetic-Algorithm (GA) optimized approach is an efficient method for creating suitable multivariate models. We used the R library GALGO (17) to identify a small feature model by continuously crossing a number of small feature models (chromosomes of features) with each other, hypothetically identifying better models with successive generations. We used an initialised fitness goal of 0.95, model size (chromosome size) of 15 genes, and k-fold cross-validation to counter overtraining. In the RF, larger classes, namely viral, were also penalized, as to ensure equal predictions across classes. After 250 models, we generated a representative model through a frequency based forward selection strategy which ensures only genes that contributed to predictions are included in the final model (S2 Appendix).

Inferring underlying interaction network

We reverse engineered gene regulatory networks using ARACNe (39) which builds an adjacency matrix of genes with their mutual information from expression data (Fig 1). These networks allow identification of functional relationships between genes and their corresponding products (40, 41). In addition, they can provide insight into the functionally relevant groups of genes for distinguishing disease state, by examining locations of RF selected genes.

To select significant interactions within our dataset we used a p-value threshold < 0.05 in the ARACNe procedure. The approach can then estimate a mutual information threshold that is relevant for the provided dataset and a specified p-value. With our data this resulted in a threshold of $MI > 0.0176$ to be retained. From the gene pairs of mutual information, we formed an edge table which was the basis for our interaction network. Nodes are genes and

edge weights are the mutual information between two genes, where greater mutual information would suggest a stronger relationship. We then loaded our networks in Cytoscape (42) which visualises molecular interaction networks and has support for a number of clustering algorithms.

To identify highly interconnected sub-networks within our reconstructed regulatory network we utilised the Cytoscape clustering plugin GLay (32). GLay uses an implementation of the Girvan-Newman Edge-betweeness algorithm (43) which we used to split our networks it into clusters of connected genes. This resulted in a number of smaller sub-networks and allowed us to inspect their functional roles within the larger network. We then mapped higher level ontologies, such as pathways and gene ontology from gene symbols and used the DAVID (44) tool to provide enrichment analysis. The enrichment analysis looked at several different ontologies, providing an indication of overrepresentation, which we used to infer the likely biological function of a given cluster. Each cluster analysis generated an enrichment table detailing enriched ontology terms along with enrichment ratios and (adjusted) p-values. From the enrichment table we then produced a dotplot which depicted enrichment ratio, p-value and gene count, along with a colour scheme denoting different ontologies, for visual interpretation.

For clusters of genes with enriched and significant terms related to the immune response, we labelled them manually as Functionally Relevant (FR) clusters. These FR clusters allowed us to make inferences about which biological functions hold predictive power, by overlaying model selected genes onto our labelled gene regulatory network.

Out of sample testing

Out of sample testing usually refers to testing a model on data not previously seen in model training and selection (37). Whilst a validation set was held back for both Affymetrix and

Illumina data, the validation data still contained samples from the same manufacturer and group of studies used in training. Hence, within the original 'discovery dataset', gene lists could still be overfit to some non-biological effect persisting in either the manufacturer technology or set of studies present, which was not removed by batch correction.

To properly test generalisability and investigate any discovery data bias, we evaluated the best performing models discovered on both Affymetrix and Illumina data by retraining and testing them on non-discovery data (Affymetrix Gene Lists to Illumina Data, and Illumina Gene lists to Affymetrix Data). These non-discovery datasets contained samples from different studies and technology and therefore represented the ideal validation datasets. With similar error between discovery and non-discovery data one can be confident that models have not overfitted to a given dataset and are suggested to be generalisable.

Results

Pre-processing

Gene de-duplication and data merging was successful for both Affymetrix and Illumina. In the final Illumina datasets 19,947 distinct genes were found intersecting all studies, whereas for Affymetrix Data we found 13,383 (Table 2). This lower Affymetrix count was due to platforms GPL571 and GPL9188 having only 13,383 distinct genes (**Table 1**). This gene loss from intersection resulted in the omittance of 8,830 gene columns, which were present for the 615 samples in GPL570.

Affymetrix platforms were successfully merged and combined via our batch correction pipeline, indicated by non-significant changes in differentially expressed (DE) genes and removal of clustering in our PCA analysis between both study and platform batch corrections (S1 Appendix). Illumina based datasets were represented by a single platform, GPL10558.

Batch correction did not result in significant changes to DE genes and removed the previously observed clustering by study (S Fig 2).

This resulting two datasets Affy_I and Illumina_I contained 1676 and 1892 samples respectively (**Error! Reference source not found.**). It is evident there is an uneven class distribution present in both datasets. Both Affy_I and Illumina_I are made up of more than 50% viral samples (66.89% and 56.50%, Table 2). The most underrepresented class is bacterial samples, with both datasets comprising fewer than 20% samples labelled as bacterial (Table 2).

Table 2. Merged and batch corrected modelling dataset description. Merged and batch corrected Affymetrix and Illumina (ambiguous classes integrated) dataset breakdown by distinct genes, platforms, class make up, and sample count.

Dataset	Distinct Genes	Platforms	Bacterial Samples	Viral Samples	Control Samples	Total Samples
Affy_I	13,383	GPL570 GPL571 GPL9188	314 (18.74%)	1121 (66.89%)	241 (14.38%)	1676
Illumina_I	19,947	GPL10558	356 (18.82%)	1069 (56.50%)	467 (24.68%)	1892

Biomarker lists

Running GA and BW on both Affymetrix and Illumina generated an ensemble of models for each method-datasets pair. For BW this was an ensemble of optimal models, one per run of the algorithm. For GA this was the evolved chromosomes obtained by repeats of the search procedure. From this ensemble of models, we computed relative gene selection frequencies (top 16 genes displayed Table 3).

Table 3. Top 16 Gene selection for Affymetrix and Illumina models and their relative selection frequencies. Frequency provided in brackets is based on the model selection frequency in each optimisation run (the number of times a gene was selected across the number of optimised models).

306 Bold genes are included amongst 3 of models top 16 selection, and underlined genes are included in
307 all four.

Affymetrix Genes (relative frequency)		**Illumina (relative frequency)**	
BW	**GA**	**BW**	**GA**
MS4A4A (1.00)	*PCOLCE2 (1.00)*	***IFI44** (1.00)*	***IFI27** (1.00)*
MTHFD2 (1.00)	*CEP55 (0.97)*	*MCEMP1 (1.00)*	*EPSTI1 (0.41)*
RSL24D1 (1.00)	*HBA1.HBA2 (0.88)*	*CD177 (1.00)*	***LY6E** (0.39)*
TSPO (1.00)	*CDC27 (0.66)*	*GPR84 (1.00)*	*SPATS2L (0.34)*
***LY6E** (1.00)*	*TSPO (0.56)*	*EIF1 (1.00)*	*RSAD2 (0.26)*
MMP8 (1.00)	***LY6E** (0.50)*	***IFI27** (1.00)*	*IFIT5 (0.24)*
NSUN7 (1.00)	*MMP8 (0.47)*	*EPSTI1 (1.00)*	***IFI44** (0.24)*
***IFI27** (1.00)*	*PGD (0.47)*	*REPIN1 (1.00)*	*ZDHHC19 (0.22)*
CXCL10 (1.00)	*RSL24D1 (0.47)*	***LY6E** (1.00)*	*FCGR1A;FCGR1CP (0.21)*
ITGAM (1.00)	*SIGLEC1 (0.47)*	*ALKBH5 (1.00)*	*IFI44L (0.19)*
PSMA6;KIAA0391 (1.00)	***IFI44** (0.44)*	*EEF2 (1.00)*	*MCEMP1 (0.19)*
GRB10 (1.00)	*OAS3 (0.44)*	*RBM33 (1.00)*	*PRC1 (0.18)*
GYG1 (1.00)	*WNT10B (0.44)*	*ARRB1 (0.99)*	*HPGD (0.17)*
PGD (1.00)	*ADAMTS3 (0.41)*	*DSCR3 (0.99)*	*OAS2 (0.17)*
CD177 (0.99)	*HPR.HP (0.38)*	*TSPAN18 (0.99)*	*HERC5 (0.17)*
OLAH (0.99)	*OLAH (0.38)*	*FCGR1A;FCGR1CP (0.96)*	*IFITM3 (0.15)*

308

309 BW search procedures in both technologies converged to a small set of genes, indicated by
310 high relative selection rate calculated by the number of times a gene was selected across the
311 multiple runs performed in each optimisation procedure. For Affymetrix 14 were included at
312 a rate of 1.0, whereas for Illumina BW results contain 12 genes at a rate of 1.0 (Table 3).
313 GA's on the other hand contained a much wider gene selection in the evolved chromosome,
314 in both search procedures only a single gene was included at a relative rate of 1.0. which
315 reflects the more varied selection in GA search procedures.

316 Overall search results (aggregated between runs by frequency) from BW and GA in both
317 Affymetrix and Illumina all contained *LY6E* (Lymphocyte antigen 6E, UniProt: Q16553)
318 amongst their 9 most frequently selected genes (Table 3). Amongst the next widely selected

319 genes were *IFI27* (Interferon alpha-inducible protein 27, mitochondrial, UniProt: P40305)
320 and *IFI44* (Interferon-induced protein 44, UniProt: Q8TCB0), both in the top 16 by gene
321 selection frequency for three of the four search procedures (Table 3). These 3 genes (*LY6E*,
322 *IFI27*, and *IFI44*) are all type-I interferon-inducible genes (ISGs), demonstrated to have
323 altered expressions in disease states, and known to be highly effective at countering infection
324 (45-48). Furthermore, an additional number of other ISGs were also found amongst the
325 frequently selected model genes (*MS4A4A, IFI44L, OAS2, and IFIT5*).

326 Additionally, several other most frequently selected genes have been linked to certain disease
327 states in the literature. Particularly increased levels of *MMP8* have been observed in HIV-
328 infected patients, which cross-references well as a high proportion of samples in our
329 modelling data coming from HIV viral studies (49). *SIGLEC1* is a Type I transmembrane
330 protein expressed by a subpopulation of macrophages and was one of fifteen genes found
331 upregulated during *in vivo* respiratory syncytial virus infections (50), whilst also said to
332 initiate the formation of the virus-containing compartment (51).

333

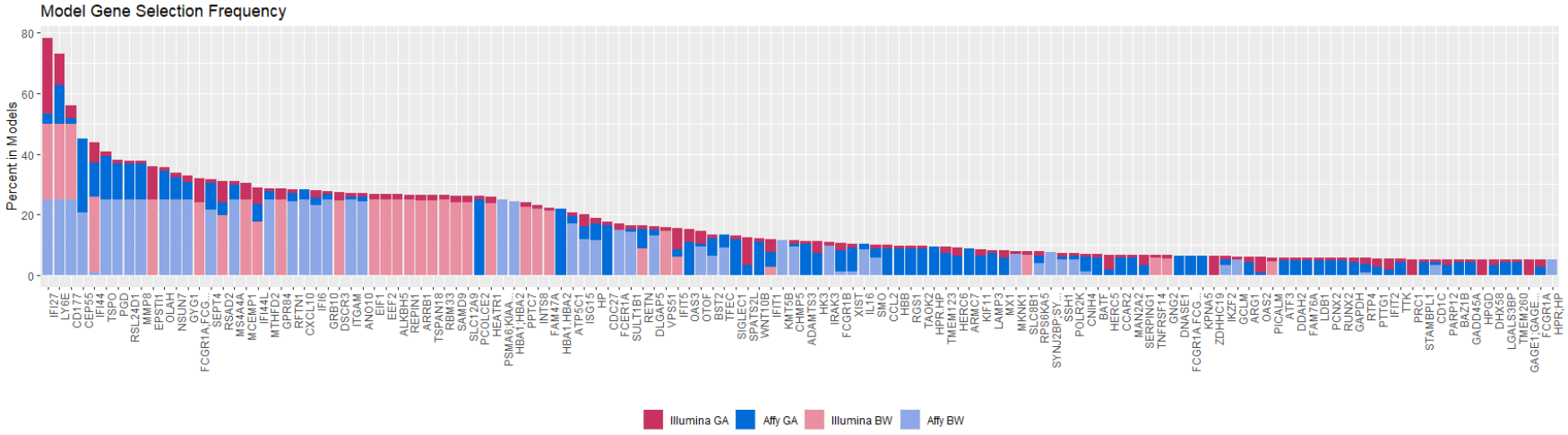

Fig 2. Gene frequency in Affymetrix and Illumina models. Each Model frequency is scaled between 1 and 25. Model overlapping gene frequencies are then stacked and coloured by model-dataset combination. Affymetrix Models by shades of blue and Illumina models by shades of red.

334 To further investigate gene convergence, we compared the relative model gene inclusion
335 rates for all search procedures together. We scaled each model gene frequency (between 1
336 and 25), then plotted them together as a stacked bar plot. Fig 2 shows the resulting stacked
337 frequency, where genes are visualised for greater than 5% aggregated inclusion across all
338 search procedures. Similarly, to our top 16 gene comparison, *LY6E* is indicated as important,
339 being represented in all search procedures. However, interestingly *IFI27* is also included
340 amongst all search procedures. Furthermore *CD177*, a neutrophil-specific receptor and
341 known to be at increased expression for patients in septic shock (52, 53), was selected
342 relatively frequently and present in all search procedures.

343 One interesting aspect to look at is the intersection of this between Genes frequently selected
344 between Affymetrix and Illumina generated models. We identified 88 genes intersecting
345 between Affymetrix and Illumina (S1 Table) and performed functional enrichment analysis
346 of them using DAVID. We found both highly enriched and significant terms relating to the
347 immune response. Included in the list of significant pathways was, in order of significance

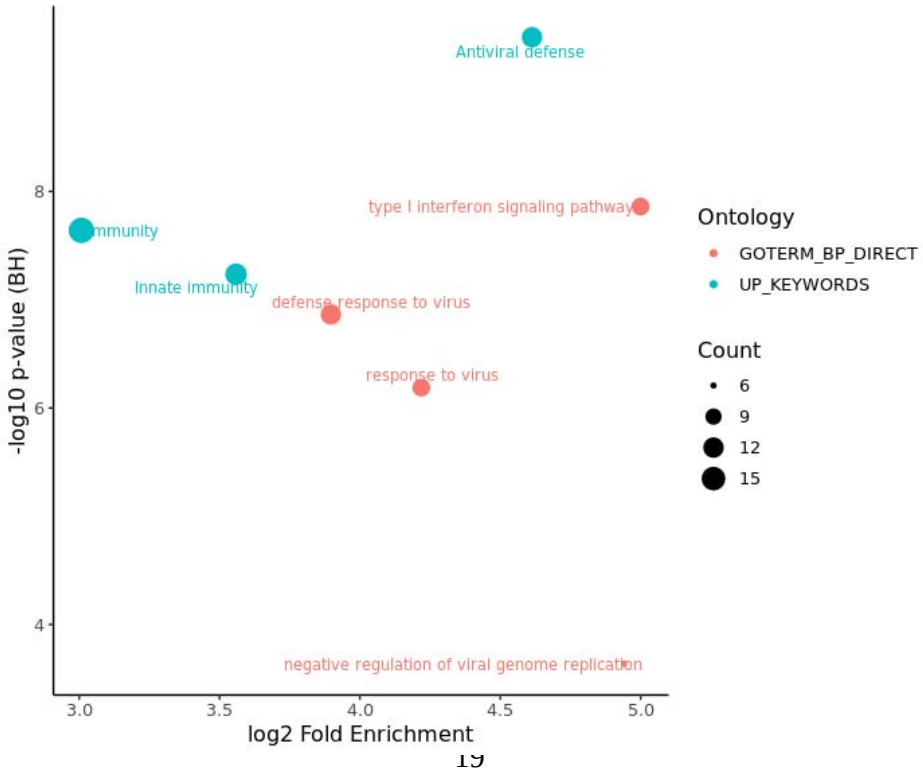

Fig 3. Functional enrichment analysis of the identified 88 genes intersecting between Affymetrix and Illumina search procedures. 'Antiviral defense' is the most significant term, whilst 'type I

348 'Antiviral defense' comprising of 12 genes, the 'type I interferon signalling pathway' which
349 included 10 genes, and 'Immunity' encompassing 17 of the 88 genes intersecting between
350 Affymetrix and Illumina search procedures (**Fig 3**).

351 For each search procedure we obtained a final representative model (Affy_BW, Affy_GA,
352 Illumina_BW, and Illumina_GA) and evaluated its performance on a held-out data split.
353 Model performance was recorded as the size of the gene list and its class-based performance
354 in terms of: Balanced Accuracy, Sensitivity, Specificity, and Mcnemar's Test p-value which
355 tests for consistency in responses and can reveal bias to classifying a certain class (all metrics
356 derived from the evaluation data split) (54).

357 **Table 4. Overall optimal model performance**. Model performance break down by Affymetrix and
358 Illumina data sets on the held out test dataset in terms of final model gene size, Balanced Accuracy,
359 Sensitivity, Specificity, and Mcnemar's Test p-value

	Model-Dataset Combination	Gene-set Size	Balanced Accuracy (B/C/V)	Sensitivity (B/C/V)	Specificity (B/C/V)	Mcnemar's Test p-value
Affymetrix Identified Models	BW	33	0.94 / 0.78 / 0.86	0.90 / 0.57 / 0.97	0.93 / 0.96 / 0.76	3.57e-3
	GA	36	0.93 / 0.82 / 0.89	0.88 / 0.66 / 0.97	0.99 / 0.97 / 0.81	4.90e-10
Illumina Identified Models	BW	30	0.86 / 0.70 / 0.78	0.80 / 0.47 / 0.87	0.93 / 0.92 / 0.87	2.36e-3
	GA	37	0.82 / 0.58/ 0.89	0.83 / 0.58 / 0.89	0.93 / 0.94 / 0.77	4.33e-15
Average	Average	34	0.89 / 0.72 /	0.85 / 0.57 /	0.95 / 0.95 /	5.93e-3

			0.86	0.93	0.80	

Average model size was similar between both Affymetrix and Illumina models (30-37 genes) (Table 4). On average models classified 0.89 of Bacterial, 0.72 of Control and 0.86 of Viral classes correctly across all datasets. In particular, the Affymetrix models, BW and GA, performed particularly well in terms of balanced accuracy on bacterial samples (0.94 and 0.93 respectively). In terms of sensitivity all models performed well for bacterial and viral classes (on average 0.85, and 0.93 respectively), however control sample performance was worse when compared to the viral and bacterial classes (0.57). Evaluating model specificity, bacterial performance was particularly high over all models (averaging 0.95) which would suggest we can determine what a bacterial sample is particularly well regardless of the model used (Table 4).

Inferred interaction networks

372 We inferred the underlying gene regulatory networks for both Affymetrix and Illumina

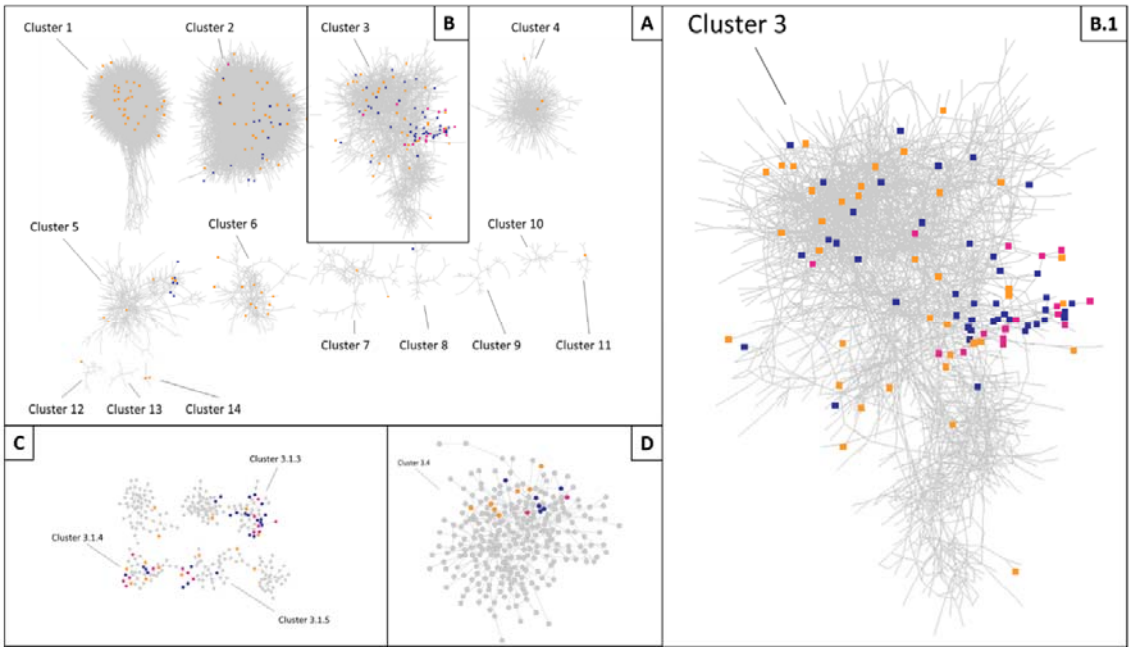

Fig 4. Clustered Illumina interaction network. Illumina models' selected genes are blue, Affymetrix selected genes are orange, and those intersecting both technologies are pink. (A) Illumina Interaction network after initial clustering (visualising clusters > 10 Genes). (B) Cluster 3, containing the most selected genes which intersected between Affymetrix and Illumina models. (B.1) Cluster 3 Enlarged. (C) Highly selected sub clusters of Cluster 3. (D) Cluster 3.4, a sub cluster of Cluster 3 containing two genes which were selected by both Affymetrix and Illumina models.

373 datasets, but present here the analysis on the larger Illumina network (Affymetrix analysis in
374 S3 Appendix). GLay clustering of the gene interaction network initially revealed 14 clusters
375 containing more than 10 genes (Fig 4). To enable a more granular analysis of specific
376 network sections (those indicated to be functionally relevant in the immune response (FR) as
377 indicated by enrichment analysis, or containing genes selected by our models) we further
378 partitioned several of the initial clusters, forming a network hierarchy (limited to a depth of 3).
379 This resulted in 110 distinct groups of genes which we analysed (Table 5).

In Illumina 24 of the 110 clusters had enriched and significant terms related to functions of the immune system in our DAVID analysis (Table 5). Of these 24 FR clusters, 10 had at been selected by at least one Illumina model. These 10 clusters contained 55 genes in the union of Illumina models (68% of all 81 Illumina selected genes in the network). Additionally, a small number of clusters (four) were selected by every model.

Table 5. Illumina interpreted inferred interaction network properties. Clusters have been labelled either functionally related to the immune response (FR). For a cluster to be labelled as FR, functional enrichment analysis of their gene list will have revealed terms both enriched and significant implicated in the host response to disease.

Manufacturer	Nodes (Genes)	Sub-clusters of more than 4 Genes	FR Clusters (% of all)	FR Clusters selected by > 1 Model (% of all)	FR Clusters selected by all four Models (% of all)
Illumina	19839	110 (1.00)	24 (21)	10 (9)	4 (4)

Affymetrix – Illumina cluster comparison. We found a similar number of clusters converged to between both Affymetrix and Illumina gene lists in their respective networks (S3 Appendix). It is clear the RF models are selecting genes from multiple of these uncorrelated clusters, to build a stronger, less correlated model feature sets able to define disease state. For greater biological understanding we compared the most selected clusters from both the Affymetrix and Illumina Interaction Network. In Illumina this was Cluster 3.1.3 (S3 Appendix). Whilst the size between both clusters was not comparable (Affymetrix – Cluster 5 being 435 Genes and Illumina Cluster 3.1.3 being only 47) we found an intersection of 16 Genes (*DDX60, IFI35, IFI44, IFI44L, IFIH1, IFIT1, IFIT2, IRF7, ISG15, MX1, OAS2, SCO2, TIMM10, TRAFD1, TRIM22 and ZBP1*) which was statistically significant (p-value < 3.18e-12), 10 of which known to be ISGs (*IFI35, IFI44, IFI44L, IFIH1, IFIT1, IFIT2, IRF7, ISG15, MX1, OAS2)* (47). Performing DAVID enrichment analysis on both clusters, we find in Illumina Cluster 3.1.3 one highly enriched term ‘type I interferon signalling pathway’

albeit with a non-significant p-value (S3 Appendix). We do not see the same term in the Affymetrix cluster; however, it does contain numerous ISGs, which we saw commonly amongst gene lists. This convergence between independent feature selection across separate manufacturers and different studies reinforces the high predictive power of ISGs for discriminating disease state across infection studies.

Independent cluster convergence between Affymetrix and Illumina models. To examine whether convergence between Affymetrix and Illumina was also to the same clusters containing the same genes we looked at where in the Illumina interaction network Affymetrix gene lists selected from (Fig 4, full break down in S3 Appendix). Although selected genes varied between Affymetrix and Illumina, we indeed found that both converged around the same clusters of genes. Moreover, we found that 19 clusters (including lower level sub clusters) were selected by both Affymetrix and Illumina models in the Illumina interaction network. Interestingly amongst this set, the four sub clusters intersecting across all Illumina gene lists (all from within the larger Illumina-Cluster 3: Fig 4) were also selected by Affymetrix gene lists: Illumina-Cluster 3.1.3, Illumina-Cluster 3.1.4, Illumina-Cluster 3.1.5, and Illumina-Cluster 3.4. All of these clusters contained genes revealed by selection frequency analysis in previous section 4.2.

We investigated all four clusters selected by all Illumina models (Clusters 3.1.3, 3.1.4, 3.1.5 and 3.4) and found they could be separated functionally to different aspects of an immune response. As mentioned, enrichment analysis on Illumina Cluster 3.1.3 revealed the ISGs to be present. However, enrichment analysis also revealed a number of both highly enriched and significant terms related to viral infections (‘response to Viruses’, ‘defense response to virus’), and most prominently ‘Antiviral Defense’ which is no surprise given the high number of interferon related genes in the cluster (S3 Appendix). Comparing the 47 genes in Clusters 3.1.3 to our model frequency analysis revealed 18 overlapping genes (*DHX58*, *EPSTI1*,

427 *HERC5, IFI44, IFI44L, IFI6, IFIT1, IFIT2, IFIT5, ISG15, MX1, OAS2, OAS3, RSAD2, RTP4,*
428 *SAMD9, SPATS2L, and TMEM123*).

429 For cluster 3.1.4, in which *LY6E* resides, it bears relation to cell signalling with by far the
430 most significant and enriched term 'chemotaxis' (S3 Appendix). Chemotaxis is well known
431 to play critical role in host response to infections, and is specifically involved in recruitment
432 of leukocytes, and movement of lymphocytes around the body (55). The intersect of cluster
433 3.1.4 with our model frequency analysis was also large, being 12 of its 40 genes *(ATF3,*
434 *CCL2, CXCL10, HERC6, LAMP3, LGALS3BP, LY6E, OTOF, PARP12, SEPT4, SERPING1,*
435 *and SIGLEC1*).

436 Cluster 3.1.5 contains genes involved in programmed cell death, containing several
437 significant and enriched terms like 'Apoptosis', 'Regulation of apoptotic process' and
438 'apoptotic process' (S3 Appendix). A total of 3 of its 37 genes intersected our model
439 frequency analysis (*CHMP5, FCGR1A, and FCGR1B*).

440 Illumina cluster 3.4 contained genes more related to general innate responses with enriched
441 terms containing 'Inflammatory response' and 'innate immune response' with non-significant
442 p-values (S3 Appendix). Amongst the genes are a number related to the Toll-like receptor
443 family (also an enriched and significant term), which respond to microbial products and
444 viruses, and are key-receptors of the innate immune system (56). Although not visible in the
445 functional enrichment analysis, Illumina Cluster 3.4 also contained a number of Interleukin
446 genes (*IL1B, IL1R1, IL4R, IL18R1, IRAK3*), known to be involved in inflammation and
447 fundamental to innate immunity (57). Out of the 253 genes in cluster 3.4, 15, including
448 CD177, intersected with previous frequency analysis (*BATF, CD177, DDAH2, GADD45A,*
449 *GPR84, GRB10, GYG1, HK3, IRAK3, MAN2A2, MKNK1, NSUN7, SULT1B1, TSPO, and*
450 *ZDHHC19*).

Cross manufacturer gene list performance

We evaluated each of the BW & GA representative models from Affymetrix on the Illumina Data and Illumina Models on the Affymetrix data. Contrasting each model's performance between these two discovery and non-discovery datasets we get the performance results depicted in Fig 5. This figure shows the difference between overall accuracy, and class-based accuracy, speciality and sensitivity when generalising our models to data pertaining from a different technology and set of studies.

In terms of overall accuracy (Fig 5 A) Affymetrix models, both GA ad BW, performed worse when applying to the Illumina data. However, the drop was less than 0.1 for both Affymetrix

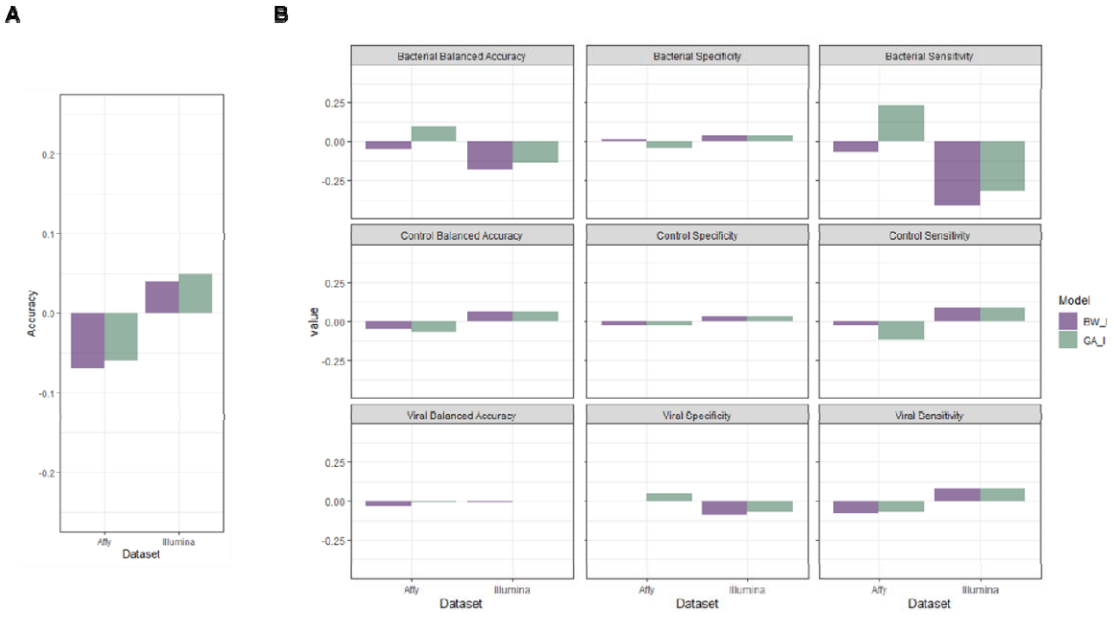

Fig 5. Cross manufacturer model change in performance. Difference in performance when taking Affymetrix derived models and testing on the Illumina data, and the Illumina derived models when testing on the Affymetrix data. (A) Difference in performance in terms of overall Accuracy. (B) Class based performance in terms of Balanced Accuracy, Sensitivity, and Specificity. For each performance measure, bars are grouped by model, and each bar refers to the difference between performance on the original dataset (which each model was discovered on) and the performance on the data it had not been exposed too. For Affymetrix models this would contrast the performance on the Affymetrix data, with the same model's performance on the Illumina data.

GA and BW. Whereas for Illumina, both GA and BW models slightly gained accuracy when applied to the Affymetrix data (0.04 and 0.05 respectively).

Looking specifically at bacterial performance (Fig 5 B), both Illumina models performed worse on the Affymetrix data in terms of bacterial balanced Accuracy (BW_I 0.71 and GA_I 0.73 2dp). Whereas the Affymetrix models performed well on the Illumina data (BW_I 0.89 and GA_I 0.89 2dp). In terms of bacterial specificity there was little change for all models, staying within +/- 0.05 2dp of change in performance. However, in terms of bacterial sensitivity, the Illumina models performed particularly worse on the Affymetrix data (BW_I 0.44 and GA_I 0.47 2dp).

Across viral class specific metrics (Fig 5 B), no model had any large change in Balanced Accuracy (change < 0.05 2dp). The largest metric change was seen in sensitivity, with Affymetrix models slightly decreasing, but with an original score of 0.97 and 0.95 for BW_I and GA_I they are still performing well when ran on the Illumina data.

Overall, both Affymetrix and Illumina models performed well given that data was pertaining from different manufacturers and different groups of studies. Particularly stability around viral performance suggests a robustness to the gene lists for classifying viral samples correctly. However, given that bacterial performance change was very comparable to viral, it too suggests a strong ability to classify bacterial samples, even when moving out of the original dataset.

Discussion

Due to the amount of relevant data, we focused our analysis on studies from two of the largest microarray platforms, Affymetrix and Illumina. Whilst these both determine the expression levels of genes and are common in large-scale population studies, differences in quantification and normalisation of gene expression values create technical difference (58). Studies within manufacturers were successfully batch corrected, indicated by non-significant changes in differentially expressed genes and removal of sample clustering by studies and platforms in PCA analysis. However, the combination of studies between manufacturers was unsuccessful, leading to two parallel analyses on the combined and batch corrected versions of (i) Affymetrix and (ii) Illumina datasets which minimized biological variation loss.

Simpler solutions are more specifically justifiable and allow for greater interpretation, which is the motivation for feature selection amongst models in biological data. We employed two feature selection algorithms using the Random Forest Classifier over our Data: Backwards Elimination and GALGO – both essentially cutting the noise and finding the most significant biological variation responsible for predicting disease state. It is unknown without a brute force search whether a *truly* optimal combination of genes has been found, however both BW and GA approaches converged around a small group of genes located in uncorrelated and functionally separable clusters. Models were found to be strongly enriched for the ISGs. In fact, *IFI27* and *LY6E* (both ISGs) were included in all Affymetrix and Illumina models. *IFI27* is involved in various signalling pathways affecting apoptosis (59-61). Whereas, *LY6E* belongs to a class of interferon-inducible factors that broadly enhance viral infectivity (62). *LY6E* has also been attributed a diverse set of effects, including attenuating T-cell receptor signalling (63) and suppressing responsiveness to *Lipopolysaccharide* which stimulate immune responses (64). Moreover, *IFI27* was shown by Tang et al. to be a *single–gene* biomarker that discriminates between influenza, and other viral and bacterial infections in

505 patients with suspected respiratory infection (65). However, this single-gene biomarker
506 approach lacks generalisability and robustness when predicting a more varied pathogen set.
507 As we have observed, performance in our meta-analysis was greatly improved by including
508 more genes in our models.

509 Our larger set of RF selected genes contained numerous examples confirmed by previous
510 studies to be implicated in disease states. For instance, our results coincide with recent meta-
511 analysis, by Andres-Terre et al., looking at transcriptional signatures of infections,
512 specifically in distinguishing influenza from other viral and bacterial infections, which found
513 127 multi-gene signatures, 27 of which were also present in our representative models (*ATF3,*
514 *BST2, CXCL10, EIF2AK2, HERC5, HERC6, IFI27, IFI44, IFI44L, IFI6, IFIT1, IFIT2, IFIT5,*
515 *ISG15, JUP, LGALS3BP, LY6E, MRPL44, MTHFD2, MX1, OAS1, OAS2, OAS3, OASL,*
516 *RSAD2, RTP4, SERPING1, SPATS2L*) serving to validate our successful data integration and
517 biological findings (66). Notably amongst these coinciding genes are *IFI27* and *LY6E*, again
518 confirming the validity of our converging feature selection.

519 By inferring the underlying interaction network, we discovered that convergence was not
520 only happening to a set of genes, but also, and more prominently, convergence was focusing
521 around particular groups of functionally similar genes. This gene-group convergence only
522 emerged as part of an in-depth investigation into the driving forces of feature selection from a
523 biological network perspective. When representative members of these uncorrelated gene
524 clusters are taken together, they can form highly predictive gene lists. With the ability to
525 define the host response to viral and bacterial infections, genes of our identified clusters are
526 likely good at approximating key functions important in disease state prediction. Notably, the
527 four functional groups of genes were indicated to be: Type I interferon-inducible genes
528 (ISGs), Chemotaxis genes, Apoptotic Processes genes, and Inflammatory / Innate Response
529 genes, which were prevalent in every model (both Affymetrix and Illumina). Within this

cluster convergence we found a highly selected group of genes to be ISGs (the most frequent between both Affymetrix and Illumina models). This is no surprise, given Type I Interferons serve as a link between the innate and adaptive immune systems (67) and have a broad range of effects on both innate and adaptive immune cells during infection with viruses, bacteria, and parasites (47). Their varying sensitivity to particular forms of pathogens is likely why a number can be used in conjunction for classification with RFs. While ISGs exact function are not fully understood, it appears our RF models have identified their strong connection to disease state (68, 69). Whilst convergence was prominent around four functional groups of genes, we also note that both in Affymetrix and Illumina, a greater more variable set of functional gene groups were used in addition within our gene lists. Hence, there is a degree of variability in gene solutions, and it seems there is an interchangeable portion of our gene lists in which a number of genes from uncorrelated functional groups of genes can be used to achieve high performance in defining disease state.

Finally, we verified our gene lists for generalisability by retraining and evaluating on data from a different manufacturer to which they were discovered in (Affymetrix Gene lists to Illumina and Illumina Gene lists to Affymetrix). It is apparent that all gene lists tend to do better on Affymetrix data, regardless of which set they were discovered on, which suggests that the dataset, not the gene lists, is influencing performance. Hence, we have uncovered the differentiating biological signatures underlying able to define bacterial and viral infections.

Conclusions

Our meta-analysis of Affymetrix and Illumina human blood infection data has revealed several panels of genes which are able to distinguish well between bacterial and viral infections. The difference in technology and gene coverage between Affymetrix and Illumina did not allow for a direct integration in our analysis. However, we were able to confirm that convergence was occurring independent of the technology, to both the same genes and the same functional groups of genes. This technology independent differentiable signal is learnable, and we demonstrated its presence by reconstructing the underlying regulatory gene network and overlaying models from the two datasets.

Acknowledgments

We thank all the contributing studies for generating and making publicly available their respective datasets. We also gratefully acknowledge DSTL (www.gov.uk/dstl) for providing support.

This work was also 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).

570 ReferencesBibliography
571
572 1. Shi Z, Gewirtz AT. Together Forever: Bacterial-Viral Interactions in Infection and Immunity.
573 Viruses. 2018;10(3):122.
574 2. Chaplin DD. Overview of the immune response. The Journal of allergy and clinical
575 immunology. 2010;125(2 Suppl 2):S3-S23.
576 3. Rock KL, Reits E, Neefjes J. Present Yourself! By MHC Class I and MHC Class II Molecules.
577 Trends Immunol. 2016;37(11):724-37.
578 4. Yewdell JW, JR B. Mechanisms of Viral Interference with MHC Class I Antigen Processing and
579 Presentation. In: Annual Reviews Collection Bethesda (MD): National Center for Biotechnology
580 Information (US);. 2002.
581 5. Manger ID, Relman DA. How the host 'sees' pathogens: global gene expression responses to
582 infection. Current Opinion in Immunology. 2000;12(2):215-8.
583 6. Suarez NM, Bunsow E, Falsey AR, Walsh EE, Mejias A, Ramilo O. Superiority of transcriptional
584 profiling over procalcitonin for distinguishing bacterial from viral lower respiratory tract infections in
585 hospitalized adults. J Infect Dis. 2015;212(2):213-22.
586 7. Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via
587 integrated host gene expression diagnostics. Sci Transl Med. 2016;8(346):346ra91.
588 8. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, et al. Gene expression patterns
589 in blood leukocytes discriminate patients with acute infections. Blood. 2006;109(5):2066-77.
590 9. Hu X, Yu J, Crosby SD, Storch GA. Gene expression profiles in febrile children with defined
591 viral and bacterial infection. Proceedings of the National Academy of Sciences. 2013;110(31):12792-
592 7.
593 10. Nascimento EJM, Braga-Neto U, Calzavara-Silva CE, Gomes ALV, Abath FGC, Brito CAA, et al.
594 Gene expression profiling during early acute febrile stage of dengue infection can predict the disease
595 outcome. PloS one. 2009;4(11):e7892-e.
596 11. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, Lucas J, et al. Gene Expression Signatures
597 Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans. Cell Host &
598 Microbe. 2009;6(3):207-17.
599 12. Dawany N, Showe LC, Kossenkov AV, Chang C, Ive P, Conradie F, et al. Identification of a 251
600 gene expression signature that can accurately detect M. tuberculosis in patients with and without
601 HIV co-infection. PloS one. 2014;9(2):e89925-e.
602 13. Lagani V, Karozou AD, Gomez-Cabrero D, Silberberg G, Tsamardinos I. A comparative
603 evaluation of data-merging and meta-analysis methods for reconstructing gene-gene interactions.
604 BMC Bioinformatics. 2016;17(5):S194.
605 14. Akey JM, Biswas S, Leek JT, Storey JD. On the design and analysis of gene expression studies
606 in human populations. Nature Genetics. 2007;39(7):807-8.
607 15. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic
608 microarray data biases. Bioinformatics. 2004;20(1):105-14.
609 16. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using
610 empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
611 17. Trevino V, Falciani F. GALGO: an R package for multivariate variable selection using genetic
612 algorithms. Bioinformatics. 2006;22(9):1154-6.
613 18. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
614 19. Denil M, Matheson D, Freitas ND. Narrowing the Gap: Random Forests In Theory and In
615 Practice. In: Eric PX, Tony J, editors. Proceedings of the 31st International Conference on Machine
616 Learning; Proceedings of Machine Learning Research: PMLR; 2014. p. 665--73.
617 20. Segal M. Machine Learning Benchmarks and Random Forest Regression. Technical Report,
618 Center for Bioinformatics & Molecular Biostatistics, University of California, San Francisco. 2003.
619 21. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in
620 Performance Evaluation. J Mach Learn Res. 2010;11:2079–107.

621 22. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data
622 using random forest. BMC Bioinformatics. 2006;7(1):3.
623 23. Jiang H, Deng Y, Chen H-S, Tao L, Sha Q, Chen J, et al. Joint analysis of two microarray gene-
624 expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics. 2004;5(1):81.
625 24. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics.
626 Bioinformatics. 2007;23(19):2507-17.
627 25. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine
628 learning research. 2003;3(Mar):1157-82.
629 26. Bellman RE. Adaptive control processes: a guided tour: Princeton university press; 2015.
630 27. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists
631 using DAVID bioinformatics resources. Nature protocols. 2009;4(1):44-57.
632 28. de la Fraga LG, Coello Coello CA. A Review of Applications of Evolutionary Algorithms in
633 Pattern Recognition. In: Wang PSP, editor. Pattern Recognition, Machine Intelligence and Biometrics.
634 Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 3-28.
635 29. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning
636 algorithms for disease prediction. BMC Medical Informatics and Decision Making. 2019;19(1):281.
637 30. Newman M. Networks: An Introduction. Oxford University Press. 2010.
638 31. Rui X, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks.
639 2005;16(3):645-78.
640 32. Su G, Kuchinsky A, Morris JH, States DJ, Meng F. GLay: community structure analysis of
641 biological networks. Bioinformatics. 2010;26(24):3135-7.
642 33. Maier M, Luxburg Uv, Hein M. Influence of graph construction on graph-based clustering
643 measures. Proceedings of the 21st International Conference on Neural Information Processing
644 Systems; Vancouver, British Columbia, Canada: Curran Associates Inc.; 2008. p. 1025–32.
645 34. Wang X, Lin Y, Song C, Sibille E, Tseng GC. Detecting disease-associated genes with
646 confounding variable adjustment and the impact on genomic meta-analysis: With application to
647 major depressive disorder. BMC Bioinformatics. 2012;13(1):52.
648 35. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London,
649 Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901;2(11):559-72.
650 36. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High
651 Dimensional Data in C++ and R. 2017. 2017;77(1):17.
652 37. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference,
653 and prediction: Springer Science & Business Media; 2009.
654 38. Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection
655 and classification using random forest. BMC Bioinformatics. 2007;8(1):328.
656 39. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE:
657 an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.
658 BMC bioinformatics. 2006;7 Suppl 1(Suppl 1):S7-S.
659 40. Boucher B, Jenna S. Genetic interaction networks: better understand to better predict.
660 Frontiers in genetics. 2013;4:290.
661 41. Mani R, St.Onge RP, Hartman JL, Giaever G, Roth FP. Defining genetic interaction.
662 Proceedings of the National Academy of Sciences. 2008;105(9):3461-6.
663 42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software
664 environment for integrated models of biomolecular interaction networks. Genome research.
665 2003;13(11):2498-504.
666 43. Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical
667 review E. 2004;69(2):026113.
668 44. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID Gene
669 Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze
670 large gene lists. Genome Biol. 2007;8(9):R183-R.

45. Ronnblom L, Eloranta ML. The interferon signature in autoimmune diseases. Curr Opin Rheumatol. 2013;25(2):248-53.
46. Schneider WM, Chevillotte MD, Rice CM. Interferon-stimulated genes: a complex web of host defenses. Annu Rev Immunol. 2014;32:513-45.
47. McNab F, Mayer-Barber K, Sher A, Wack A, O'Garra A. Type I interferons in infectious disease. Nat Rev Immunol. 2015;15(2):87-103.
48. Kyogoku C, Smiljanovic B, Grun JR, Biesen R, Schulte-Wrede U, Haupl T, et al. Cell-specific type I IFN signatures in autoimmunity and viral infection: what makes the difference? PLoS One. 2013;8(12):e83776.
49. Singh H, Samani D, Nambiar N, Ghate MV, Gangakhedkar RR. Prevalence of MMP-8 gene polymorphisms in HIV-infected individuals and its association with HIV-associated neurocognitive disorder. Gene. 2018;646:83-90.
50. Jans J, Unger WWJ, Vissers M, Ahout IML, Schreurs I, Wickenhagen A, et al. Siglec-1 inhibits RSV-induced interferon gamma production by adult T cells in contrast to newborn T cells. Eur J Immunol. 2018;48(4):621-31.
51. Hammonds JE, Beeman N, Ding L, Takushi S, Francis AC, Wang JJ, et al. Siglec-1 initiates formation of the virus-containing compartment and enhances macrophage-to-T cell transmission of HIV-1. PLoS Pathog. 2017;13(1):e1006181.
52. Demaret J, Venet F, Plassais J, Cazalis M-A, Vallin H, Friggeri A, et al. Identification of CD177 as the most dysregulated parameter in a microarray study of purified neutrophils from septic shock patients. Immunology Letters. 2016;178:122-30.
53. Stroncek DF. Neutrophil-specific antigen HNA-2a, NB1 glycoprotein, and CD177. Curr Opin Hematol. 2007;14(6):688-93.
54. Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation. 1998;10(7):1895-923.
55. Jin T, Xu X, Hereld D. Chemotaxis, chemokine receptors and human disease. Cytokine. 2008;44(1):1-8.
56. Das A, Guha P, Sen D, Chaudhuri TK. Role of toll like receptors in bacterial and viral diseases – A systemic approach. Egyptian Journal of Medical Human Genetics. 2017;18(4):373-9.
57. Dinarello CA. Interleukin-1 in the pathogenesis and treatment of inflammatory diseases. Blood. 2011;117(14):3720-32.
58. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic acids research. 2005;33(18):5914-23.
59. Rosebeck S, Leaman DW. Mitochondrial localization and pro-apoptotic effects of the interferon-inducible protein ISG12a. Apoptosis : an international journal on programmed cell death. 2008;13(4):562-72.
60. Liu N, Zuo C, Wang X, Chen T, Yang D, Wang J, et al. miR-942 decreases TRAIL-induced apoptosis through ISG12a downregulation and is regulated by AKT. Oncotarget. 2014;5(13):4959-71.
61. Gytz H, Hansen MF, Skovbjerg S, Kristensen AC, Horlyck S, Jensen MB, et al. Apoptotic properties of the type 1 interferon induced family of human mitochondrial membrane ISG12 proteins. Biology of the cell. 2017;109(2):94-112.
62. Mar KB, Rinkenberger NR, Boys IN, Eitson JL, McDougal MB, Richardson RB, et al. LY6E mediates an evolutionarily conserved enhancement of virus infection by targeting a late entry step. Nat Commun. 2018;9(1):3603.
63. Saitoh S, Kosugi A, Noda S, Yamamoto N, Ogata M, Minami Y, et al. Modulation of TCR-mediated signaling pathway by thymic shared antigen-1 (TSA-1)/stem cell antigen-2 (Sca-2). Journal of immunology (Baltimore, Md : 1950). 1995;155(12):5574-81.
64. Meng F, Lowell CA. Lipopolysaccharide (LPS)-induced macrophage activation and signal transduction in the absence of Src-family kinases Hck, Fgr, and Lyn. The Journal of experimental medicine. 1997;185(9):1661-70.

65. Tang BM, Shojaei M, Parnell GP, Huang S, Nalos M, Teoh S, et al. A novel immune biomarker IFI27 discriminates between influenza and bacteria in patients with suspected respiratory infection. The European respiratory journal. 2017;49(6).
66. Andres-Terre M, McGuire HM, Pouliot Y, Bongen E, Sweeney TE, Tato CM, et al. Integrated, Multi-cohort Analysis Identifies Conserved Transcriptional Signatures across Multiple Respiratory Viruses. Immunity. 2015;43(6):1199-211.
67. Tough DF. Type I interferon as a link between innate and adaptive immunity through dendritic cell stimulation. Leuk Lymphoma. 2004;45(2):257-64.
68. Hertzog PJ, O'Neill LA, Hamilton JA. The interferon in TLR signaling: more than just antiviral. Trends in immunology. 2003;24(10):534-9.
69. Kovarik P, Castiglia V, Ivin M, Ebner F. Type I Interferons in Bacterial Infections: A Balancing Act. Frontiers in immunology. 2016;7(652).

Supporting information

S1 Appendix. Pre-processing.

S1 Fig. Affymetrix Interaction Network. Affymetrix recovered interaction network at first level of clustering. Selected model genes are highlighted.

S1 Table. Model Gene Selection Frequency. Affymetrix and Illumina model selected genes with relative frequency of selection (genes with greater than 5% aggregated inclusion across all search procedures).

S2 Appendix. Biomarker Search.

S2 Fig. Illumina Interaction Network. Illumina recovered interaction network at first level of clustering. Selected model genes are highlighted.

S2 Table. Highly selected gene clusters from Affymetrix and Illumina interaction network. Table containing the genes from the 4 highly model selected Illumina clusters, and 5 highly model selected gens from the Affymetrix clusters.

S3 Appendix. Inferred Interaction Networks.

749 **S3 Table. Out-sample results of gene lists**. The out-sample results from running Affymetrix

750 derived gene lists on the Illumina data, and the Illumina derived gene lists on the Affymetrix

751 data

752