A Pattern Recognition Bioinformatics Alternative System to Rodent Models in Fundamental Research

Brett A. Lidbury[1] *and Alice M. Richardson*[1,2]

[1]Alternatives to Animal Research through Bioinformatics Group, The Department of Genome Biology, John Curtin School of Medical Research, The Australian National University, Canberra, Australia; [2]The Faculty of Information Sciences and Engineering, The University of Canberra, Australia

Summary

Reliance upon rodent models in fundamental biomedical research is increasing due to technologies that allow the genetic engineering of inbred mouse strains. Through these genetic technologies and other established benefits for biomedical research success, there is a rodent-based experimental "system" that serves the researcher's needs for whole organism experimentation. Furthermore, it is accepted that medical advances require an "animal model" as an essential aspect to ultimate success. Direct studies on humans are seen to be problematic due to significant heterogeneity and diversity in human populations. Modern bioinformatics, particularly the pattern recognition field, provides potential answers to the problem of overcoming such outbred diversity. Advanced statistical methods and machine learning algorithms developed by computer scientists, for example recursive partitioning and support vector machines (SVMs), applied to human biomedical data, provide a basis for an alternative system to mouse models through unravelling diversity and providing clues to guide investigations into human disease.

Keywords: bioinformatics, pattern recognition, machine learning, rodent, replacement

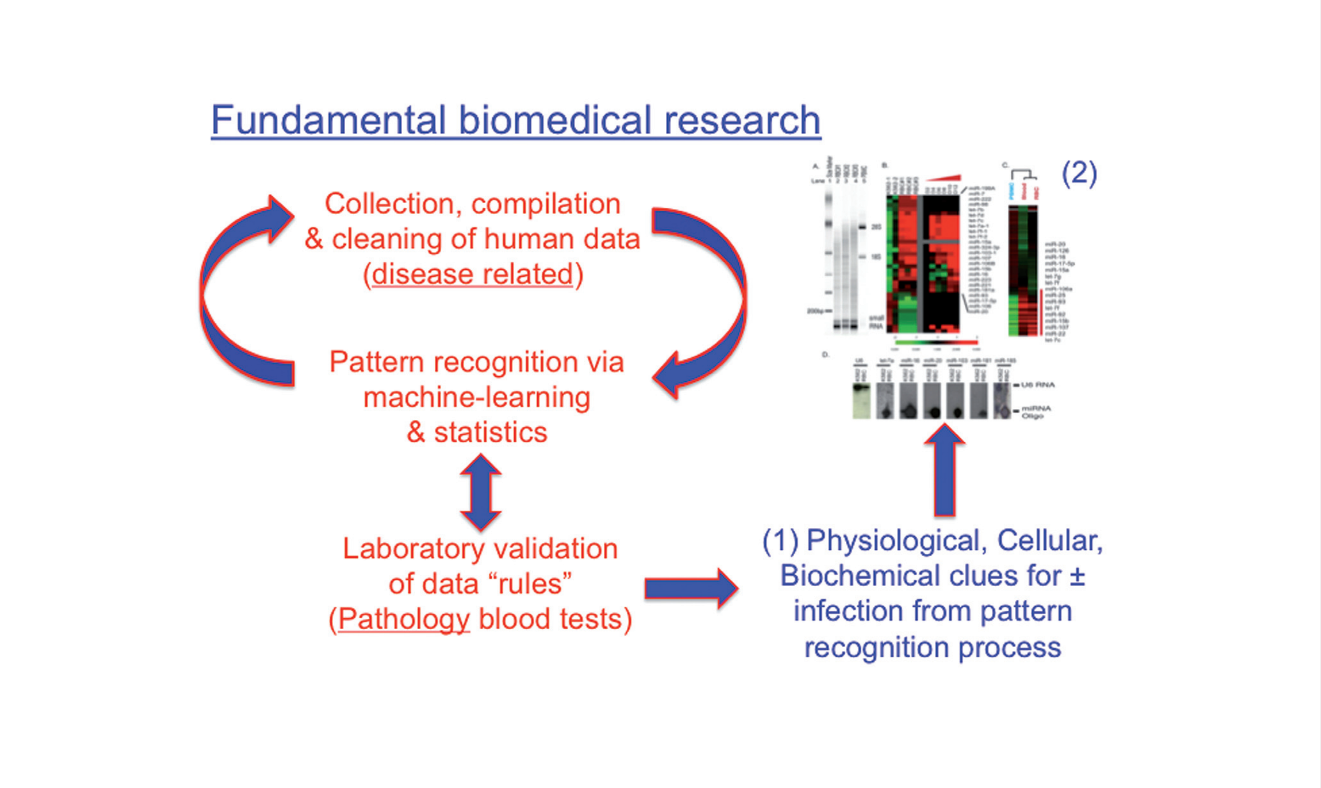
1 Introduction

A common explanation for slow progress in fundamental human disease research is "the lack of a suitable animal model." Traditionally, an animal model of a given human disease is regarded as an essential requirement for progress towards a vaccine, therapy, treatment, or diagnostic marker (Nguyen et al., 2011). Much of this view is predicated on notable human health advances gained via mouse models, for example ectromelia models for smallpox (Fenner, 1947, 2000) and MHC restriction to explain the basis of cell-mediated immunity (Zinkernagel and Doherty, 1974). These examples from virology have understandably inspired continued efforts in mouse models to understand contemporary human virus disease challenges, with some resultant "humanized" mouse models established, despite the fact that viruses such as hepatitis C virus (HCV) (Ploss and Rice, 2009) do not naturally infect rodents. This was also true for *Human Immunodeficiency Virus* (HIV) through the use of SCID (severe combined immunodeficiency) mice transfused with human immune cells to attempt an animal model of human disease (McCune et al., 1991; Goldstein et al., 1996). A very recent development in HCV pathogenesis has achieved a humanized mouse model, not by human cell transfer but via the genetic manipulation and whole mouse expression of humanized viral receptor proteins that allow HCV binding and infection (Dorner et al., 2011).

Given the historical success of some mouse models for providing clues to human disease biology, coupled with practical advantages such as large litters, short generation time, easy housing, and now routine mouse genome manipulation (e.g., gene knock-out mice), the preclinical animal model of choice for any human disease condition is the mouse. These advantages are amplified further when acknowledging the well known difficulties of direct human research. Because of these factors, mice are now being used to model human mental health disorders, such as depression and anxiety (Pryce and Seifritz, 2011), as well as autism (Bangash et al., 2011), in the hope of identifying genetic and biological mechanisms that underpin these conditions.

The 3Rs relevance of this article is the proposal herein of an alternative system to well-entrenched mouse models of human disease utilized routinely in preclinical and fundamental biomedical research, following the principle of absolute animal replacement (Balls, 1994). The system (Fig. 1) uses advanced bioinformatics (pattern recognition through machine learning) to overcome problems of diverse human data and responses, tethered to validation studies on human blood and DNA samples, which combined will provide insights into genetic and biological properties to guide scientific investigations of human disease.

The example used to illustrate this alternative systems approach will focus on human *Mycoplasma pneumoniae* (Mp) infection. Mp and *Chlamydia pneumoniae* have mouse disease models present in the biomedical literature (Chu et al., 2006; Sommer et al., 2009). This replacement alternative uses immunoassay results from Mp laboratory investigations and associ-



Fundamental biomedical research

Collection, compilation & cleaning of human data (disease related)

Pattern recognition via machine-learning & statistics

Laboratory validation of data “rules” (Pathology blood tests)

(1) Physiological, Cellular, Biochemical clues for ± infection from pattern recognition process

Fig. 1: Summary of the total alternative system for investigating fundamental research questions in human disease, which are generally addressed through mouse models
The system comprises the human data acquisition and data-cleaning phase prior to pattern recognition (via machine-learning) analyses to generate quantitative rules from pathology test predictor variables. The *in silico* rules are subjected to biological validation through prospective pathology testing, providing additional data for the optimization of *in silico* models. With a validated model for infection or disease, genetic and/or proteomic investigations can be performed on cohorts of prospective samples divided by predictor variable patterns as associated with a positive or negative disease/infection condition.
The gene array data example shown (2) is on erythrocyte micro-RNA patterns (Chen S-Y et al., 2008) that can be linked to routine pathology test results like MCHC and RDW.

ated routine pathology data generated by a hospital pathology department. Assessment of data patterns linked to either a positive or negative Mp result provides guidance to physiological responses that reflect disease or normal physiology in the tested patients. Pathology blood test results reflect organ function, electrolyte, carbohydrate and lipid balance, as well as red cell properties and the broad white blood cell response to allergy or to infectious agents encountered. In addition to routine pathology tests, special tests are available to provide further specific laboratory data on a health condition. Given that blood test results provide a comprehensive laboratory appraisal of health and data is of high quality control standards, *in silico* pattern recognition will provide a viable alternative to mouse models once properly validated prospectively in the pathology laboratory. Furthermore, additional analytical potency will be achieved when linked to modern genomic and proteomic platforms. Well developed human cell culture and cell co-culture systems can also be linked to such a system to assist with the elucidation of cellular, molecular, and biochemical mechanisms.

2 Materials and methods

Data

Data description and analytical methods applied have previously been reported in detail (Richardson et al., 2008). Briefly, longitudinal pathology data (1997-2007) were extracted from the ACT Pathology Laboratory databases (The Canberra Hospital, ACT, Australia) after appropriate permissions had been obtained and human ethics clearance granted, and thereafter de-identified by senior hospital staff prior to analysis. The criteria for inclusion were that the patient had an immunoassay performed for *Mycoplasma pneumoniae* (Mp) within the time interval stated above. Predictor variables also provided for each case included the liver function test results for alanine aminotransferase (ALT) and gamma-glutamyl transpeptidase (GGT), urea, serum sodium (Na^+) and creatinine, haemoglobin (Hb), total red cell (erythrocyte) count (RCC), mean corpuscular volume (MCV), mean corpuscular haemoglobin (MCh), mean corpuscular haemoglobin concentration (MCHC), haematocrit (Hct) and red cell diameter width

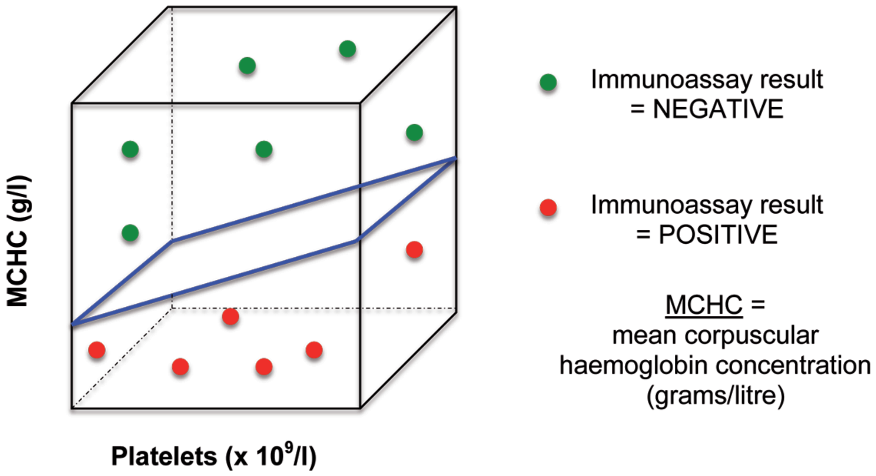

MCHC (g/l)
Platelets (x 10^9/l)
● Immunoassay result = NEGATIVE
● Immunoassay result = POSITIVE
MCHC = mean corpuscular haemoglobin concentration (grams/litre)

Fig. 2: A schematic of the support vector machine (SVM) modeling algorithm
From a two dimensional input space the classification thereafter occurs in a three-dimensional feature space. The wedge in the middle represents the separating hyperplane, defined by "support vectors," which classifies the data as outcome 0 or 1 based on relationships between kernels (see Karatzoglou et al., 2006; Smola and Scholkopf, 2004).

(RDW), platelet (thrombocyte) count, total white cell (leukocyte) count (WCC), and patient age and sex (Richardson et al., 2008).

Data cleaning was required, and cases with missing values were removed from the analysis, as were repeat cases on the same patient and cases with confounding data (e.g., age = 200). Routine preparation and cleaning of data was conducted in Microsoft Excel spreadsheets, prior to data upload into the statistical computing package *R* (versions 2.6.1-2.9.1) for statistical and pattern recognition analyses.

Statistical and machine-learning methods
Prior to analysis, the Mp immunoassay data were classified as positive, negative or indeterminate. A positive result meant that specific antibody titers against Mp were detected above the laboratory immunoassay cut-off value specific to the method used for this assay. These positive or negative (indeterminate was not included) Mp immunoassay results were used as the dependent (response) variables in subsequent analyses, which in addition to machine-learning methods previously included logistic regression (Richardson et al., 2008). The associated pathology blood test results (ALT, Hb, WCC, urea and so on) were included as the independent (predictor) variables for each immunoassay case, and it was the subsequent patterns within the predictor variables that were of interest in terms of relationships to infection status that will guide future genomic and gene expression investigations post validation.

The *R* packages "rpart" and "e1071" were utilized for the recursive partitioning (decision tree) and support vector machines (SVM) machine-learning methods, respectively (Fig. 2). Decision tree analyses (Kingsford and Salzberg, 2008) were run as single trees, or as tree ensembles, to assess the best methods for predictive power and pattern detection (Richardson, Shadabi and Lidbury, unpublished results). The tuning and execution of SVM models for pattern detection in pathology data were based on R code described by Karatzoglou et al. (2006) and also informed by Smola and Scholkopf (2004).

Data modeling and interpretation
For the study described here, decision trees and SVMs were used in tandem, and both as classification algorithms (i.e., positive or negative immunoassay result). The decision tree results were used to inform SVM modeling for the best predictor variables to include. For both methods, the positive or negative Mp results were used as response variables against the range of routine pathology data included in the modeling as predictor variables. The results of analysis highlight a decision tree plus SVM investigation of patterns associated with a positive or negative immunoassay result for Mp.

3 Results

A combination decision tree and SVM machine learning analysis were performed to identify key predictors and predictor thresholds among independent (predictor) variable associations with a positive or negative immunoassay result for Mycoplasma pneumoniae (Mp). Figure 3a shows the result of a single decision tree analysis of 162 Mp positive cases (1) versus 162 Mp negative cases (0). The results show that age (≥50.5 years) and Na^+ (≥130.5 mmol/l) were the most direct predictors of negative Mp immunoassay, with the same age criteria but a serum Na^+ concentration ≤130.5 mmol/l, giving a Mp positive classification. This suggests immediately that when considering the hierarchy in the tree that age and serum sodium (Na^+) levels were of primary importance in terms of Mp infection as assessed by antibody detection via a specific immunoassay. Through the right side of the tree, a more complex pattern was achieved with a platelet count of <201.5 x 10^9/l leading to a Mp negative result (0). However, if the platelet count was >201.5 x 10^9/l and GGT <48.4 U/l (anti-log of 1.685 U/l or $10^{1.685}$) a positive (1) Mp immunoassay was predicted. Beyond this GGT node was an additional node with an age threshold split at 37.5 years showing an additional classification of Mp immunoassay positive or negative. Put simply, these results suggest that in relation to a positive or negative Mp immunoassay result, age, serum Na^+, platelet count and GGT were significant in terms of human respiratory infection by this microbe.

Using the predictor variables identified by decision trees (Fig. 3a), SVM modeling was performed using Mp immunoassay classifications (0 or 1) as the response variable and age, Na^+, platelet count, and GGT as predictor variables. Figure 3b shows the result of one SVM classification plot comparing platelet count (10^9/l) with $\log_{10}$-transformed GGT (U/l). The third dimension, or slice, of this model was provided by Na^+ at 130.5 mmol/l (Fig. 2). Age was also effective in this modeling approach as a visualization variable, but it did not provide as large a positive (pink) classification (data not shown).

Interpreting the SVM plot (Fig. 3b) by extrapolation, at a fixed Na^+ concentration of 130.5 mmol/l, a Mp positive immunoassay result (1) at a serum GGT concentration of 10 U/l (LogGGT = 1.0) is associated with a platelet count range of approximately 350-450 x 10^9/l. Using the same logic, a Mp positive immunoassay result with a GGT of 100 U/l (Log GGT = 2.0) is associated with a platelet count range of approximately 250-700 x 10^9/l, also with Na^+ fixed at 130.5 mmol/l. Platelet and GGT relationships, as defined by serum Na^+ outside of the pink (1) classification area represent negative Mp immunoassay results.

4 Discussion

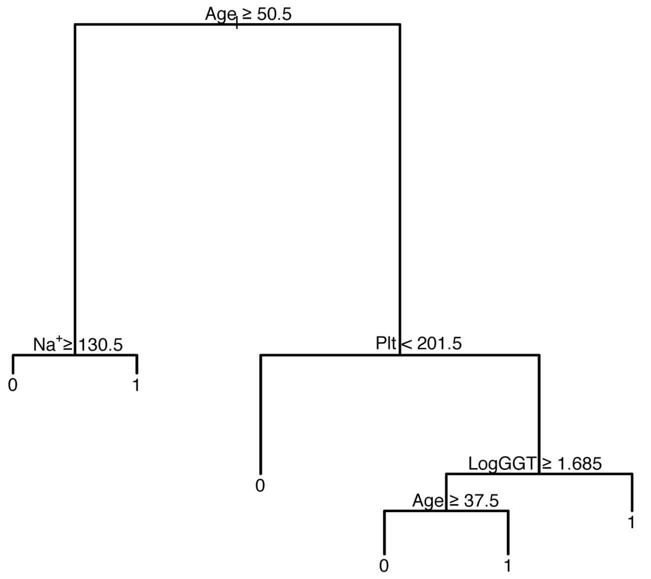
Pattern recognition/knowledge discovery is a branch of bioinformatics concerned with the study of large data sets and the extraction of data patterns associated with a condition of interest (Brusic and Zeleznikow, 1999), in this case microbial infection. Advanced statistical methods can be used, but machine learning algorithms such as support vector machines (SVMs) provide extra analytical power, particularly as these algorithms can be trained on specific data sets and thereafter tested to assess accuracy of prediction. Machine learning methods can be applied individually, or as part of an ensemble where several simpler models are fitted to the same data and their results combined to further enhance predictive accuracy (Kingsford and Salzberg, 2008; Smola and Scholkopf, 2004). These *in silico* methods, however, are not sufficient in isolation and must be linked to biological validation studies (Fig. 1). The focus on pathology data allows a link to high quality human biomedical data to explore the validation of *in silico* rules. For example, the decision tree results (Fig. 3a) can be used to formulate rules through which to predict a positive Mp immunoassay result (e.g., age <50.5 years + platelet count >201 x 10^9/l + GGT <48 U/l). As summarized in Figure 1, once rules are validated and possible biomarkers identified from pathology data as associated with an infection or disease, genomic, gene expression and/or proteomic methods can be used to further explore disease mechanisms without the need for a mouse model.

Furthermore, this combination of sophisticated *in silico* methods, biological validation through high quality human pathology data, and use of these results to guide human genetic studies satisfies the requirements of absolute replacement as first enunciated by Russell and Burch (1959) and further emphasized by Balls (1994). That is, no animal or animal-derived material is involved in any step of this system, as it has been replaced by computer modeling and ethical human sampling and data retrieval.

Mouse models for *Mycoplasma pneumoniae* (Mp) are established in the literature and are used to deduce disease mechanisms and to test drug and vaccine strategies (Chu et al., 2006). The analysis of human data (Fig. 3) indicated a role for liver function through the appearance of GGT. While no direct involvement of liver function could be found in the Mp mouse literature, some studies do point to the Mp-derived CARDS toxin as a basis of lung pathology in the mouse and baboon through the activation of inflammatory cytokines and chemokines (Hardy et al., 2009). Prolonged Mp infection and associated toxin production would inevitably lead to liver involvement. Some human studies suggest liver pathology is associated with community-acquired Mp infection, although this is controversial (Daxboeck et al., 2005; Romero-Gomez et al., 2006). Through this proposed mouse model replacement alternative system and its necessary focus on human biology, a deeper analysis will be possible to clarify human data and examine the relevance of previous results from animal studies.

As the foundation for an alternative system to mouse models and as a means to explore fundamental human disease processes, *in silico* pattern recognition methods offer a powerful approach to overcome the diversity of human genetic and biological responses to environmental challenges, including microbial infection. Through hospitals and other health agencies, abundant human data is available for modeling in this way, and the application of powerful machine learning and other data mining techniques offers an approach to counter the often-suggested concerns of fundamental biomedical researchers that human da-

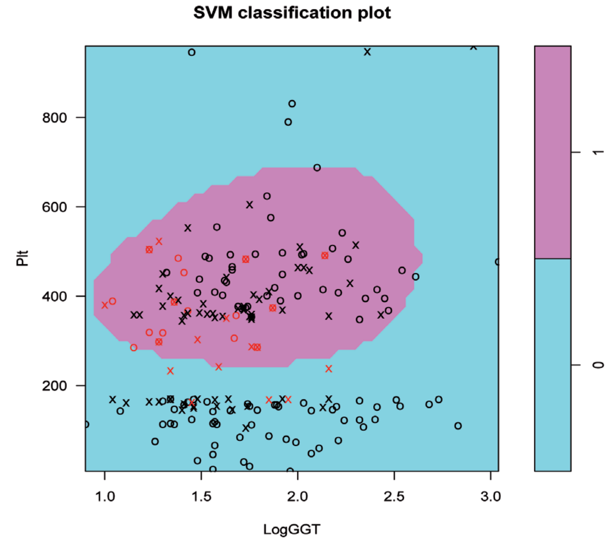
Fig. 3a



***Mycoplasma pneumoniae* (Mp) immunoassay result**	*Examples of pathology test predictor variables associated with an Mp positive (1) or negative (0) immunoassay result*
Antibody positive for Mp (1)	Patients ≤ 50.5 years old with a platelet count > 201.5 x 10^9/l and LogGGT ≤ 1.685 U/l (linear value 48 U/l)
Antibody negative for Mp (0)	Patients ≥ 50.5 years old with a serum Na^+ ≥ 130.5 mmol/l, or, Patients ≤ 50.5 years old with a platelet count < 201.5 x 10^9/l

Fig. 3: Machine-learning analysis of pathology data classified as positive or negative for *Mycoplasma pneumoniae* (Mp) antibody by optimized immunoassay
The presence of specific anti-Mp antibody indicates past Mp infection (1).
(a) A recursive partitioning decision tree output using the *rpart* package in R, with a summary table providing examples of pathology test predictor variables used to classify an Mp positive versus an Mp negative immunoassay result. The tree output is hierarchical, suggesting that variables higher on the tree have stronger interactions with the response variable under test (i.e., Mp positive or negative classification). The tree results suggest the involvement of age, electrolyte balance, blood clotting, and liver function.
(b) Re-analysis of results from (a) by a SVM (e1071 package in R). Three dimensions were taken from the decision tree in (a) for the SVM modeling; platelet count (x 10^9/l), GGT U/l ($\log_{10}$ transformed), and the unseen slice of Na^+ (130.5 mmol/l). The SVM model also included age. On examination of the Mp positive (1 = pink) area in the plot, three predictor variables can be assessed in relation to the Mp positive result. Na^+ as the slice is constant at 130.5 mmol/l, but both platelet count and GGT are continuous scales where different values can be extrapolated through the x-y axes relationship.
Plt = platelet; Na^+ = sodium (serum);
GGT = gamma-glutamyl transpeptidase

Fig. 3b



ta and samples are too heterogeneous to adequately analyze for meaningful whole-organism disease processes and mechanisms. To be a suitable alternative to mouse and other animal models, however, requires careful biological validation of *in silico* rules generated by these methods, which will be performed as a future aspect of this study.

This paper proposes a novel system combining bioinformatics and machine learning with validation studies on human pathology data, and ultimately advanced genomic and proteomic methods, to provide an absolute replacement alternative to widespread mouse models of human disease. While satisfying Russell and Burch's ultimate goal of the 3Rs to see all animal experiments eventually replaced, this system will also provide scientific insights into human disease, which through a human focus will be more readily translated into direct medical advances.

References

Balls, M. (1994). Replacement of animal procedures – alternatives in research, education and testing. *Lab. Anim. 28*, 193-211.

Bangash, M. A., Park, J. M., Melnikova, T., et al. (2011). Enhanced polyubiquitination of Shank3 and NMDA receptor in a mouse model of autism. *Cell 145*, 758-772.

Brusic, V. and Zeleznikow, J. (1999). Knowledge discovery and data mining in biological databases. *Knowl. Eng. Rev. 14*, 257-277.

Chen, S. Y., Wang, Y., Telen, M. J., et al. (2008). The genomic analysis of erythrocyte microRNA expression in sickle cell diseases. *PLoS One 3*, e2360.

Chu, H. W., Breed, R., Rino, J. G., et al. (2006). Repeated respiratory Mycoplasma pneumoniae infections in mice: effect of host genetic background. *Microbes. Infect. 8*, 1764-1772.

Daxboeck, F., Gattringer, R., Mustafa, S., et al. (2005). Elevated serum alanine aminotransferase (ALT) levels in patients with serologically verified Mycoplasma pneumoniae pneumonia. *Clin. Microbiol. Infect. 11*, 507-510.

Dorner, M., Horwitz, J. A., Robbins, J. B., et al. (2011). A genetically humanized mouse model for hepatitis C virus infection. *Nature 474*, 208-211.

Fenner, F. (1947). Studies in infectious ectromelia of mice; immunization of mice against ectromelia with living vaccinia virus. *Aust. J. Exp. Biol. Med. Sci. 25*, 257-274.

Fenner, F. (2000). Adventures with poxviruses of vertebrates. *FEMS Microbiol. Rev. 24*, 123-133.

Goldstein, H., Pettoello-Mantovani, M., Katopodis, N. F., et al. (1996). SCID-hu mice: a model for studying disseminated HIV infection. *Semin. Immunol. 8*, 223-231.

Hardy, R. D., Coalson, J. J., Peters, J., et al. (2009). Analysis of pulmonary inflammation and function in the mouse and baboon after exposure to Mycoplasma pneumoniae CARDS toxin. *PLoS One 4,* e7562.

Karatzoglou, A., Meyer, D., and Hornik, K. (2006). Support Vector Machines in *R. J. Stat. Softw. 15*, 1-28.

Kingsford, C. and Salzberg, S. L. (2008). What are decision trees? *Nat. Biotechnol. 26*, 1011-1013.

McCune, J., Kaneshima, H., Krowka, J., et al. (1991). The SCID-hu mouse: a small animal model for HIV infection and pathogenesis. *Annu. Rev. Immunol. 9,* 399-429.

Nguyen, C. Q., Lavoie, T. N., and Lee, B. H. (2011). Current concepts: Mouse models of Sjogren's Syndrome. *J. Biomed. Biotechnol.*, Epub before print.

Ploss, A. and Rice, C. M. (2009). Towards a small animal model for hepatitis C. *EMBO Rep. 10*, 1220-1227.

Pryce, C. R. and Seifritz, E. (2011). A translational research framework for enhanced validity of mouse models of psychopathological states in depression. *Psychoneuroendocrinology 36*, 308-329.

Richardson, A., Hawkins, S., Shadabi, F., et al. (2008). Enhanced laboratory diagnosis of human Chlamydia pneumoniae through pattern recognition derived from pathology database analysis. In M. Chetty, S. Ahmand, A. Ngom, and S. W. Teng (eds.), *Supplementary Proceedings of PRIB 2008*, Melbourne, November 14-16, 2008 (227). Berlin: Springer.

Romero-Gomez, M., Otero, M. A., Sanchez-Munoz, D., et al. (2006). Acute hepatitis due to Mycoplasma pneumoniae infection without lung involvement in adult patients. *J. Hepatol. 44*, 827-828.

Russell, W. M. S. and Burch, R. L. (1959). *The principles of humane experimental technique*. London: Methuen.

Smola, A. J. and Scholkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput. 14*, 199-222.

Sommer, K., Njau, F., Wittkop, U., et al. (2009). Identification of high- and low-virulent strains of Chlamydia pneumoniae by their characterization in a mouse pneumonia model. *FEMS Immunol. Med. Microbiol. 55*, 206-214.

Zinkernagel, R. M. and Doherty, P. C. (1974). Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature 248*, 701-702.

Acknowledgments

The authors acknowledge the support of the Medical Advances Without Animals Trust (http://www.mawa-trust.org.au/) and ACT Pathology, The Canberra Hospital, ACT Australia. Particular thanks go to Dr Fariba Shadabi, Michael de Souza and Gus Koerbin for their intellectual and technical advice on aspects of this project.

Correspondence to

Brett A. Lidbury, PhD
Alternatives to Animal Research through
Bioinformatics group
The Department of Genome Biology
John Curtin School of Medical Research
The Australian National University
Canberra 0200
Australia
Phone: +61 2 6125 9429
e-mail: brett.lidbury@anu.edu.au