Genes and Immunity (2009) 10, 112–119
© 2009 Macmillan Publishers Limited All rights reserved 1466-4879/09 $32.00
www.nature.com/gene

ORIGINAL ARTICLE

Integrated analysis of genetic and proteomic data identifies biomarkers associated with adverse events following smallpox vaccination

DM Reif[1], AA Motsinger-Reif[2], BA McKinney[3], MT Rock[4], JE Crowe Jr[4,5,6] and JH Moore[7,8]
[1]National Center for Computational Toxicology, US Environmental Protection Agency, Research Triangle Park, NC, USA; [2]Department of Statistics, Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA; [3]Department of Genetics, University of Alabama School of Medicine, Birmingham, AL, USA; [4]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA; [5]Department of Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA; [6]Program in Vaccine Sciences, Vanderbilt University Medical Center, Nashville, TN, USA; [7]Department of Genetics, Dartmouth Medical School, Lebanon, NH, USA and [8]Computational Genetics Laboratory, Dartmouth Medical School, Lebanon, NH, USA

Complex clinical outcomes, such as adverse reaction to vaccination, arise from the concerted interactions among the myriad components of a biological system. Therefore, comprehensive etiological models can be developed only through the integrated study of multiple types of experimental data. In this study, we apply this paradigm to high-dimensional genetic and proteomic data collected to elucidate the mechanisms underlying the development of adverse events (AEs) in patients after smallpox vaccination. As vaccination was successful in all of the patients under study, the AE outcomes reported likely represent the result of interactions among immune system components that result in excessive or prolonged immune stimulation. In this study, we examined 1442 genetic variables (single nucleotide polymorphisms) and 108 proteomic variables (serum cytokine concentrations) to model AE risk. To accomplish this daunting analytical task, we employed the Random Forests (RF) method to filter the most important attributes, then we used the selected attributes to build a final decision tree model. This strategy is well suited to integrated analysis, as relevant attributes may be selected from categorical or continuous data. Importantly, RF is a natural approach for studying the type of gene–gene, gene–protein and protein–protein interactions we hypothesize to be involved in the development of clinical AEs. RF importance scores for particular attributes take interactions into account, and there may be interactions across data types. Combining information from previous studies on AEs related to smallpox vaccination with the genetic and proteomic attributes identified by RF, we built a comprehensive model of AE development that includes the cytokines intercellular adhesion molecule-1 (ICAM-1 or CD54), interleukin-10 (IL-10), and colony stimulating factor-3 (CSF-3 or G-CSF) and a genetic polymorphism in the cyokine gene interleukin-4 (IL4). The biological factors included in the model support our hypothesized mechanism for the development of AEs involving prolonged stimulation of inflammatory pathways and an imbalance of normal tissue damage repair pathways. This study shows the utility of RF for such analytical tasks, while both enhancing and reinforcing our working model of AE development after smallpox vaccination.

Genes and Immunity (2009) **10,** 112–119; doi:10.1038/gene.2008.80; published online 16 October 2008

Keywords: *smallpox; Random Forests; integrated analysis; genetic; proteomic; interactions*

Introduction

Live attenuated vaccinia virus, delivered intradermally, is the vaccine given to immunize individuals against smallpox. Although vaccination of healthy adults with vaccinia virus induces a protective response in the majority of individuals immunized, vaccinia virus is reactogenic in a significant number of vaccinees.[1] The most common adverse events (AEs) after vaccination include fever, lymphadenopathy (swelling and tenderness of lymph nodes) and a generalized acneiform rash. Collectively, these clinical reactions suggest that individuals suffering AEs have immune responses beyond the necessary magnitude, or sustain the immune response longer than necessary.

To elucidate the complex pathophysiology underlying unwanted responses to vaccination, we gathered high-dimensional genetic and proteomic data in a cohort of subjects in which a portion experienced an AE after primary immunization with Aventis Pasteur smallpox vaccine. Through a comprehensive examination of systemic (serum) cytokine/chemokine changes combined with the characterization of polymorphisms in a large panel of candidate genes, we sought to provide a thorough portrayal of the complex genetic and proteomic interplay behind the development of AEs. Knowledge of how risk factors in a subject's genetic back-

Correspondence: Dr DM Reif, National Center for Computational Toxicology, US Environmental Protection Agency, D343-03, 109 TW Alexander Drive, Research Triangle Park, NC 27711, USA.
E-mail: reif.david@epa.gov
Received 11 June 2008; revised and accepted 27 August 2008; published online 16 October 2008

ground interact with dynamically changing levels of immunological proteins could shed light on important therapeutic targets or pathways to direct vaccine modification and pre-vaccination screening procedures.

It is increasingly gaining acceptance that complex clinical outcomes, such as adverse reaction to vaccination, arise from the concerted interactions among the myriad components of a biological system.[2] Complicating genetic factors, such as multiple contributing loci and/or susceptibility alleles, incomplete penetrance and epistasis, are further convoluted by proteomic, metabolomic and environmental effects.[3] If such a multiscale system is to be understood, then interactions among its many attributes must be considered.[4] Although there is considerable intuitive appeal to the incorporation of multiple types of biological data, simultaneous analysis of information on different scales of measurement (that is, continuous proteomic data and categorical genetic data) creates additional analytical challenges. Therefore, appropriate computational analysis methods must traverse large numbers of input variables and handle diverse data types. For this study, we employed a two-stage analytical strategy. The first step was to filter a list of over 1500 genetic and proteomic attributes, taking interactions within and across data types into account, down to an analytically tractable subset of candidates. The second step involved careful statistical and biological exploration of the filtered subset of candidate attributes, resulting in a final model of AE development.

For the first (filter) step, we implemented a random forest™ (RF) approach.[5] RF is a machine learning technique that builds a forest of classification trees by sampling, with replacement, from the data and selecting the attribute at each tree node from a random subset of all attributes. The RF method offers many advantages for the analysis of diverse biological data. First, it can handle a large number of input attributes, both discrete (for example, single nucleotide polymorphisms, or SNPs) and continuous (for example, microarray expression levels or data from high-throughput proteomic technologies). Second, RF estimates the relative importance of attributes in discriminating between classes (in this case, AE status), thus providing a metric for feature selection. Third, RF produces a highly accurate classifier with an internal unbiased estimate of generalizability during the forest-building process. Fourth, RF is robust in the presence of etiological heterogeneity and missing data.[6] Finally, learning is fast and computation time is modest even for very large data sets.[7]

In the second (modeling) step, we took advantage of the tractable number of attributes identified by the RF filter to explore thoroughly the statistical and biological relationships among the attributes and AE outcomes. Decision trees were used to derive a descriptive, biologically interpretable model of the functional interactions among the attributes associated with systemic AEs. Our final model justified our multiscale analysis strategy, in that it included the cytokines intercellular adhesion molecule-1 (ICAM-1 or CD54), interleukin-10 (IL-10) and colony stimulating factor-3 (CSF-3 or G-CSF), as well as an SNP in interleukin-4 (IL4). Evaluating our final model from an immunological perspective, we conclude that AEs in response to smallpox vaccination result from the hyperactivation of inflammatory pathways, leading to excess recruitment and stimulation of monocytes in peripheral tissues. This model is consistent with work demonstrating overstimulation of inflammatory and tissue damage repair pathways developed in earlier studies of AEs after smallpox vaccination.[8–11]

Materials and methods

Study subjects

Vaccines, study subjects and clinical vaccine study design have been described in detail.[9] Briefly, 148 (116 with recorded AE information) healthy adults were enrolled at the Vanderbilt University Medical Center as part of a multicenter study of primary immunization against smallpox using the Aventis Pasteur smallpox vaccine at National Institutes of Health (NIH) Vaccine and Treatment Evaluation Units. NIH-DMID Protocol 02-054 was implemented. Volunteers were eligible if they had no smallpox vaccination scar, no history of vaccinia virus immunization, normal renal and hepatic serum chemistry values, no contraindications against immunization (pregnancy, immunosuppression or eczema) and negative serum test results for hepatitis B surface antigen, hepatitis C virus antibody, rapid plasma reagin and HIV-1 ELISA. There were a total of 61 subjects for whom both genetic and proteomic data were gathered. Individuals were asked to self-identify race; white (60) and Asian (1) were the only categories identified in this cohort. To facilitate comparison with earlier studies, and because there was no statistical difference in age, gender or race according to AE status (data not shown), the data were not adjusted for these covariates.

Clinical assessments

Details of the clinical assessments have been described earlier.[9] For all study subjects, a team of trained physicians and nurse providers examined the medical history and clinical symptoms to ensure consistent clinical assessment. Subjects were examined on five visits within the first month after vaccination and were assessed for occurrence of an AE. Collection of serum for cytokine measurements occurred at the evaluation just before vaccination (baseline) and at the evaluation between days 5 and 7 post-vaccination (acute phase). Although all AEs were noted, only *systemic* AEs were considered in this study, as we expected these to be associated more strongly with serum cytokine expression than would an AE displayed only at the site of inoculation. Systemic AEs included fever, generalized rash and lymphadenopathy. Specifically, fever was defined as an oral temperature of $>38.3\,^{\circ}\mathrm{C}$. Generalized rash was defined as skin eruptions on non-contiguous areas in reference to the site of vaccination. Detailed descriptions of the acneiform rashes considered in this study have been described.[12] Lymphadenopathy was defined as enlargement or tenderness of regional lymph nodes attributed to vaccination. For subjects on which both genetic and proteomic data were gathered, 16 subjects experienced a systemic AE and 45 subjects did not experience an AE.

Identification of genetic polymorphisms

The custom SNP panel used in this study was based on the NCI SNP500 Cancer project[13] and has been described earlier.[14] The majority of SNPs included on the panel target soluble factor mediators and signaling pathways, many of which have immunological significance. Genotyping for SNPs was performed using DNA amplified directly from Epstein-Barr virus-transformed B cells generated from peripheral blood samples collected from each subject. Genotyping was performed at the Core Genotyping Facility of the National Cancer Institute (NCI, Gaithersburg, MD, USA). Genotypes were generated using the Illumina GoldenGate assay technology. Of the 1536 SNPs assayed, a total of 1442 genotypes passed standard quality control filters. In Reif *et al.*,[15] the complete list of SNPs analyzed is available.

Quantification of serum cytokine levels

Serum samples were obtained just prior to vaccination (baseline) and 6–9 days after vaccination (acute), as described earlier in detail.[9] Serum samples were collected in 5 ml Vacutainer serum separator tubes (Becton Dickinson, San Jose, CA, USA) and were centrifuged at $700 \times g$ for 10 min. The serum then was collected, aliquoted into cryovials (Sarstedt Inc., Numbrecht, Germany) and stored at –80 °C until assayed. Cytokine concentrations were determined using rolling circle amplification technology-enhanced custom dual antibody sandwich immunoassay arrays, as described.[16–19] The expression levels of 108 protein analytes were measured in 100 μl serum aliquots from the patient samples. Glass slides held 12 replicate spots of monoclonal capture antibodies specific for each analyte. Duplicate samples of sera were incubated for 2 h, washed and then incubated with secondary biotinylated polyclonal antibodies. The 'rolling circle' method was then used to amplify signals.[17] Quality control measures were used to optimize antibody pairs, minimize array-to-array variation and standardize procedures of chip manufacturing.[17] A Tecan LS200 unit was used to scan arrays and customized software was used to determine mean fluorescence intensities. In addition, 15 serial dilutions of recombinant analytes at known concentrations (studied in parallel on each slide) were used to develop best-fit equations for each analyte, and the upper and lower limits of quantitation were defined. Changes in serum cytokine concentrations were calculated as percent change from the subject's baseline value because of the broad individual range of systemic cytokine expression before and after immunization.

Random Forests

An RF is a collection of decision tree classifiers, where each tree in the forest has been trained using a bootstrap sample of individuals from the data, and each split attribute in the tree is chosen from among a random subset of attributes. Classification of individuals is on the basis of aggregate voting over all trees in the forest.

Each tree in the forest was constructed as follows from data having $N = 61$ individuals and $M = 1552$ explanatory (genetic plus proteomic) attributes:

(1) The method chose a training sample by selecting N individuals, with replacement, from the entire data set.
(2) At each node in the tree, m attributes were selected randomly from the entire set of M attributes in the data. The absolute magnitude of m was a function of the number of attributes in the data set and remained constant throughout the forest-building process.
(3) The method chose the best split at the current node from among the subset of m attributes selected above.
(4) We iterated the second and third steps until the tree was fully grown (no pruning).

Repetition of this algorithm yielded a forest of trees, each of which had been trained on bootstrap samples of individuals (see Figure 1). Thus, for a given tree, certain individuals were left out during training. Prediction error and attribute importance were estimated from these 'out-of-bag' individuals.

The out-of-bag (unseen) individuals were used to estimate the importance of particular attributes according to the following logic: If randomly permuting values of a particular attribute did *not* affect the predictive ability of trees on out-of-bag samples, then that attribute was assigned a low importance score. If, however, randomly permuting the values of a particular attribute drastically impaired the ability of trees to correctly predict the class of out-of-bag samples, then the importance score of that attribute was high. By running out-of-bag samples down entire trees during the permutation procedure, attribute interactions were taken into account when calculating importance scores, as class was assigned in the context of other attribute nodes in the tree.

The recursive partitioning trees comprising an RF provide an explicit representation of attribute interaction that is readily applicable to the study of interactions among multiple data types.[20,21] These models may uncover interactions among genes, proteins and/or environmental factors that do not exhibit strong marginal effects. In addition, tree methods are suited to dealing with certain types of genetic heterogeneity, as splits near the root node define separate model subsets in the data. RFs capitalize on the solid benefits of decision trees and

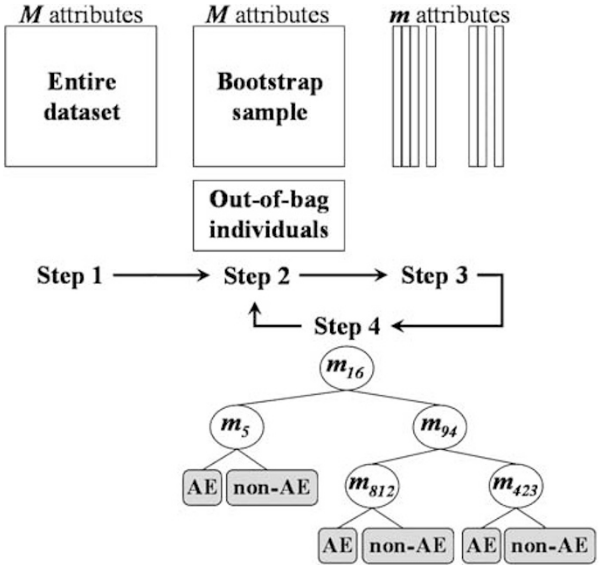

Figure 1 Construction of individual trees using the Random Forest method from a full data set of N individuals and M attributes. Proceeding from the root node, individual subjects were classified into terminal AE status leaves according to the value of that individual's genetic or proteomic attribute at each node. The steps correspond to those described in the text.

have demonstrated excellent predictive performance when the forest is diverse (that is, trees are not highly correlated with each other) and composed of individually strong classifier trees.[5,22] The RF method is a natural approach for studying gene–gene, gene–protein or protein–protein interactions because importance scores for particular attributes take interactions into account without demanding a pre-specified model.[23]

Decision trees

To represent the interactions among genetic and/or proteomic attributes associated with AEs, decision trees were chosen to build the final model because of their ready interpretability and explicit modeling of attribute interactions. The tree classified individual subjects into AE groups by proceeding down a dichotomous tree, where the genetic or proteomic attribute at each node (or split) was selected for the gain in information it provided. Gain in information was attributed when knowledge about the variation in this attribute separated subjects into appropriate AE classes. When interpreting the tree, attributes at each node were taken in the context of attributes at nodes closer to the root—thus allowing an explicit representation of attribute interactions. To augment the generalizability of our final model, we stipulated that at least five subjects must appear in each terminal (status) leaf and used 10-fold cross-validation (CV) to estimate the predictive ability of the final model. Although CV accuracy was reduced by allowing trees with less than five subjects in terminal nodes, CV accuracy proved to be insensitive to changes in other tree parameters for these data. We used the implementation of the C4.5 decision-tree algorithm provided in the Weka machine learning software package to obtain our final model.[24]

Data analysis strategy

RF analysis was performed using the freely available R package randomForest.[25,26] This package is based on the original Fortran code available at the website cited in Breiman and Cutler.[27] RF was used to analyze data sets containing each biological data type separately and in parallel, resulting in two stratified data sets (genetic only; proteomic only) and a combined data set (both genetic and proteomic attributes). Genetic attributes were treated as categorical, whereas proteomic attributes were treated as continuous values. For each genetic, proteomic, or combined data set, forests comprised of 10 000 trees were grown. Attribute importance was calculated using the out-of-bag permutation test described above. The relative importance (rank) of functional genetic attributes and related proteomic attributes was determined from the mean decrease in the Gini index using the out-of-bag permutation testing procedure. The relative importance determined from the mean decrease in classification accuracy produced nearly identical results both here and in extensive simulation studies.[28]

The simulation studies[28] used the current data as the basis for a range of simulated models, providing guidance for the parameters in the analysis discussed here. The relative rank of simulated genetic and proteomic predictors was evaluated for a range of filter cutoffs and on both stratified and combined data sets. Results from these data-based simulation studies demonstrated high confidence that AE-associated attributes having relatively meager effects would be ranked in the top 10% of attributes in RF analysis, and that analysis of the combined (genetic and proteomic) data was generally advantageous. Therefore, we chose the top 10% of attributes as ranked by RF as candidates for inclusion in our final model. To represent the interactions among genetic and/or proteomic attributes associated with AEs, we built a decision tree model.

Biological interpretation of our final model was aided by Chilibot (chip literature robot) knowledge mining software.[29] Chilibot inferred relationship networks among the attributes in the final model based on linguistic analysis of relevant records from public biomedical literature databases. The natural language processing approach used by Chilibot is superior to standard co-occurrence text mining approaches because parsing text into sentences can characterize the type of relationship (for example, inhibition or stimulation) between input terms. The terms given explicitly to Chilibot as input were 'ICAM-1,' 'IL-10,' 'IL-4' and 'CSF-3,' as well as the alternate gene names 'CD54' and 'G-CSF' (for ICAM-1 and CSF-3, respectively). The software automatically adds syntactic synonyms (for example, ''IL 10,' 'IL-10,' 'IL10,' and so on) to the search criteria. Because the goal of this study is hypothesis generation, as opposed to strict hypothesis testing, Chilibot was used to aid in discovery rather than using any pre-defined network relationships.

Results

Filtering of important attributes using RFs

Supplementary Table 1 lists all attributes having an importance rank in the top 10% relative to all attributes in the combined data set. Figure 2 depicts the attribute importance score landscape over the entire data set. This landscape proved robust to changes in RF parameters, provided that a sufficiently large forest (10 000 trees) was grown. RF identified both genetic and proteomic attributes as important discriminators of AE status. Approximately one-third of the attributes identified as important were genetic, with the remaining two-thirds being proteomic. Although this distribution among data

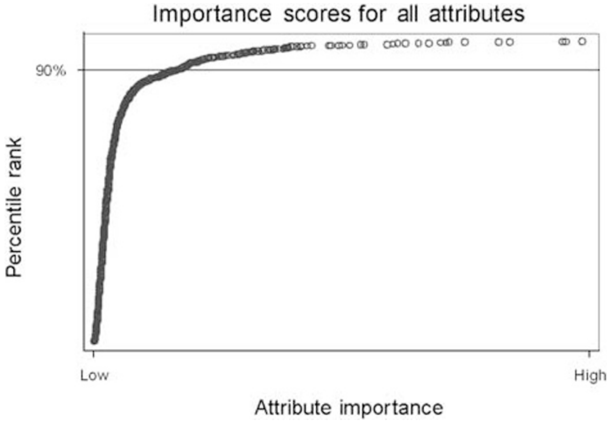

Figure 2 Attribute importance 'landscape' showing the shape of the importance curve ranking all attributes in the combined (genetic plus proteomic) data set. Attributes above the horizontal line indicate a relative importance rank in the top 10% (90th percentile) of all attributes in the data set.

types may reflect systematic patterns concerning the etiology of AE outcomes, the bias toward proteomic attributes probably arose out of the fact that the cytokine array was specifically designed to capture variation in important systemic mediators. In contrast, the genetic data include candidate SNPs in and around genes having a variety of immunological functions. In addition, with multiple SNPs per gene, correlation existing among polymorphisms (that is, haplotypes) could drive down RF importance scores for particular SNPs, as RF might select any SNP from within a haplotype at a particular node. Indeed, the *IL4* SNP in our final model was part of a group of four SNPs in *IL4* having nearly identical importance scores, and Haploview analysis showed them to be in high linkage disequilibrium, providing evidence that these genetic polymorphisms are inherited as a haplotype.[30] In this context, linkage disequilibrium has an impact akin to etiological heterogeneity, which is a concern in any association study. The heterogeneity concern is part of the rationale for using RF as a first-stage filter that identifies a handful (the top 10%) of attributes for further consideration. The effect of repeated samplings over many thousands of trees gives all attributes an unbiased opportunity to demonstrate AE association, even if importance scores for groups of SNPs in linkage disequilibrium are slightly tamped down. Thus, attributes whose importance scores may be tamped down by phenomena such as linkage disequilibrium still have a chance to surpass our 10% importance threshold over a sufficiently large forest of resampled trees, whereas slightly down-weighted importance scores may push interesting attributes below an overly strict first-stage threshold in a smaller forest. Considering the RF importance rank of attributes included in our final model relative to all attributes in the combined data set, all three proteomic attributes were ranked in the top 1%, and the *IL4* SNP (rs 2243290) was ranked in the top 5%. Relative to their respective data types, the *IL4* SNP was ranked in the top 1% among all attributes in the genetic data set, and ICAM-1, CSF-3 and IL-10 were ranked in the top 1% among all proteomic attributes.

Modeling the association of genetic and proteomic biomarkers with AEs

Having filtered out the noise using RFs, we used a decision tree representation to explore interactions among the attributes in our filtered list related to AE status. The final decision tree model is shown in Figure 3. Our final model included four variables—three proteomic attributes and one genetic attribute. Change in ICAM-1 concentration comprises the root node of the tree, with subsequent nodes composed of change in IL-10 concentration, a SNP in *IL4*, and change in CSF-3 concentration. Imposing our minimum of five individuals per terminal (AE status) leaf, this tree correctly classified 89% of individuals (with seven misclassifications) in the full data set and achieved a 10-fold CV (prediction) accuracy of 75%.

Figure 4 characterizes the biological relationships among the attributes in the tree using Chilibot. Interactive relationships were characterized into one of three types based on the verbs connecting pairs of attributes in the biomedical literature as follows: (1) Stimulatory relationships were connected by verbs such as ‘activate,’ ‘stimulate’ or ‘enhance.’ (2) Inhibitory relationships were connected by verbs such as ‘decrease,’ ‘attenuate’ or ‘inhibit.’ (3) Neutral relationships were assigned when the nature of the relationship could not be determined contextually. Mining the biomedical literature suggested interactive relationships connecting all of the attribute nodes in our final model. Stimulatory, inhibitory or neutral pair-wise interactive relationships were identified between each of ICAM-1, IL-10, *IL4* and CSF-3.

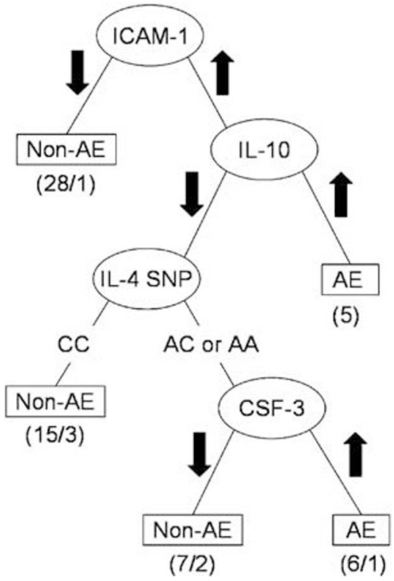

Figure 3 Final model of genetic and proteomic factors contributing to AE development. Each node (oval) constitutes a decision point based on the genotype of genetic attributes (*IL4* SNP) or whether the concentration change from baseline in proteomic attributes (ICAM-1, IL-10 and CSF-3) was above (upward-pointing arrows) or below (downward-facing arrows) a calculated threshold. Starting at the root node (ICAM-1), subjects were classified into AE status leaves (rectangles) by proceeding along the decision points at each attribute node. Given below each terminal leaf is the total number of subjects classified into that AE status group/the number of subjects incorrectly assigned to that AE status group.

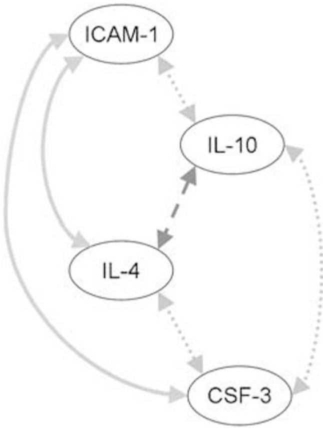

Figure 4 Biological relationships among the attributes in our final model characterized using Chilibot. Connections between each attribute node (oval) are denoted according to the type of interactive relationship they represent: stimulatory (solid), both stimulatory and inhibitory (dotted) or neutral (dashed). Arrowheads indicate that interactions between particular biological attributes are bi-directional.

Thorough examination of the networks inferred facilitated the biological interpretation of the final model discussed below.

Discussion

Our final model provides an immunologically plausible and testable biological mechanism of AE occurrence after smallpox vaccination that includes both genetic and proteomic factors. The analytical strategy used is appropriate for the study of complex phenotypes, as outcomes such as AE development likely result from the interplay of multiple genetic, proteomic and environmental factors.[31,32] The decision tree trained on the attributes passing our RF filter proposes a solid biological model of AE development.

The attributes included in this tree point to an important role of one particular immune cell type: is monocytes. Monocytes are bone marrow-derived circulating blood cells that are precursors of tissue macrophages. Monocytes are recruited actively to the sites of inflammation, where they differentiate into macrophages in tissues. These macrophages play important roles in coordinating both innate and adaptive immune responses. Macrophages are activated by microbial products such as endotoxin and by T-cell cytokines such as interferon-γ. Activated macrophages phagocytose and kill microorganisms, secrete pro-inflammatory cytokines and present antigens to helper T cells.

The root node of the tree we developed is ICAM-1 (CD54), where small changes from baseline concentration ($<11\%$) of ICAM-1 predict a non-AE response to vaccination and high changes from baseline concentration ($>11\%$) point toward AE risk, depending on factors in subsequent nodes. ICAM-1 is mainly expressed on endothelial cells, T cells, B cells and monocytes. It functions in cell–cell adhesion, which plays a crucial role in monocyte differentiation into macrophages, as entry into tissues is necessary. In addition, ICAM-1 expression is upregulated in mature monocytes,[33] aiding in cell adhesion and the eventual differentiation into macrophages. Circulating monocytes are in random contact with endothelial cells, and the adhesion molecule E-selectin slows the monocyte by inducing rolling of the monocyte along the endothelial surface before firm attachment to vascular cell adhesion molecule 1 or ICAM-1, which interact with integrins on the monocyte surface. Once the monocyte is tightly bound, it then migrates between endothelial cells.[34,35] Excessive levels of ICAM-1 might cause an ‘over-recruitment’ of monocytes into tissue, triggering an unnecessarily active innate inflammatory response.

For individuals with large changes in ICAM-1, the next node in the tree is IL-10, where changes from baseline $>85\%$ are associated with AEs. IL-10 is produced by activated macrophages and some helper T cells for which a major function is to inhibit activated macrophages and, therefore, to maintain homeostatic control of innate and cell-mediated immune reactions. Changes in IL-10 levels may indicate an imbalance in this delicate homeostasis, leading to AEs.

For individuals with mild changes in IL-10 concentration, the next node is an SNP in the gene encoding IL-4. In an earlier genetic study of two vaccination cohorts (including a subset of individuals in the present data), this same IL-4 polymorphism was associated with AEs ($P=0.05$ and $P=0.06$ in the first and second cohort, respectively).[15] Interestingly, by including proteomic factors in this study, our model indicates that the AE risk conferred by this SNP is dependent on proteomic context. IL-4 is a cytokine produced mainly by the TH_2 subset of $CD4^+$ helper T cells, whose functions include the induction of the differentiation of TH_2 cells from naive $CD4^+$ precursors, stimulation of IgE production by B cells and suppression of interferon-γ-dependent macrophage functions.[36–38] Although direct functional significance of the SNP is unknown, it is reasonable that the different genotypes could result in functionally different versions of the IL-4 protein or in different bioavailability levels of IL-4. The fact that multiple SNPs in *IL4* achieved nearly identical importance scores indicates that variation within the *IL4* gene region may be related functionally to the development of AEs. Because of the intricate cross-talk between macrophages and the TH_2 response in maintaining homeostasis, it is plausible that the major *IL4* genotype (CC) is associated with calming the activated macrophage response and directing the acquired immune system to progress in response to vaccine presentation, whereas the variant genotypes (AC or AA) fail to calm the innate response, presenting increased AE risk.

For individuals having one of the variant genotypes at *IL4*, the lowest node of the tree is CSF-3 (G-CSF). G-CSF is a cytokine produced by activated T cells, macrophages and endothelial cells at the sites of infection, which acts on the bone marrow to mobilize and increase the production of neutrophils to replace those consumed in inflammatory reactions. In our model, increased levels of CSF-3 after vaccination (change $>78\%$) indicated increased risk of suffering an AE. This finding implies another possible over-recruitment event in the development of AEs, as neutrophils have been associated with host tissue damage and failure to terminate acute inflammatory responses.[39] This reaction is consistent with the types of AE symptoms observed in this study and with the overall proposed biological mechanisms of AE development.

The results of this study provide a viable biological hypothesis of AE occurrence after smallpox vaccination that is experimentally testable. Our model includes both genetic and proteomic biomarkers. Allowing for such an integrative model is an important strength of our analytical strategy. It is increasingly recognized that the pathophysiology of complex clinical outcomes hinges on biological factors acting on multiple levels.[40] Therefore, the formulation of robust etiological models must take this inherent complexity into account and capitalize on the power of modern experimental data-generating techniques.

We conclude that AEs after smallpox vaccination result from hyperactivation of inflammatory signals, leading to excess recruitment and stimulation of monocytes in peripheral tissues. Our analysis identifies a set of interacting genetic and proteomic candidates associated with AEs, such as ICAM-1, IL-10, *IL4* and CSF-3. As the proteomic measurements occurred early in the period after vaccination, before most AEs presented themselves clinically, our model could be used as a diagnostic tool in the prediction of AEs. Of course, the ultimate goal of such a study is the identification and characterization of biological risk factors contributing to the inappropriate immune response to vaccination. We present a hypothesized mechanism of AE development that targets specific elements of systemic inflammatory pathways for further study.

Future studies should further evaluate the reproducibility of the current model, given that the number of vaccinated subjects meeting the criteria for inclusion and having both genetic and proteomic data was relatively small. Ideally, our model would be evaluated for replication in an entirely independent sample. However, the validity of our current model can be assessed through the statistical process of internal CV (where our model achieved a 75% prediction accuracy) and through comparison of these results with our earlier studies of genetic[15] or proteomic[8] data alone. In this study, our RF approach with the combined data identified all attributes highlighted in the earlier proteomic study[8] (ICAM-1, CSF-3, Eotaxin and TIMP-2) and two of the three genes highlighted in the earlier genetic study[15] (*MTHFR* and *IL4* but not *IRF1*). Although the *IRF1* polymorphisms were not ranked in the top 10% of all attribute importance scores in the combined data set, these attributes would have passed the top 10% filter criteria relative to only genetic attributes. Given that the subset of subjects used in this study (that is, those having both genetic and proteomic data) has only partial overlap with subjects in either of the earlier studies, we feel that the current results are remarkably stable.

Finally, our hypothesized model must be tested at the bench. The functional consequences of genetic variability in *IL4* should be characterized fully. Time series studies with dense measurement points are needed to shed light on the dynamic interplay between the signaling of ICAM-1, IL-10 and CSF-3. Additional data are needed on the effects of these cytokines in other physiological compartments. Careful assessment of external factors (such as nutrition, fitness and relevant environmental exposures) influencing protein expression should be considered in future studies. The results from this study suggest that analysis of the molecular and cellular basis of complex clinical phenomena will require an experimental approach that takes into account the broader spatial and temporal physiological context of complex biological systems.

Acknowledgements

This work was supported by the National Institutes of Health (NIH)/National Institute of Allergy and Infectious Diseases (NIAID) Vaccine Trials and Evaluation Unit (contract N01-AI-25462, study DMID 02-054); NIH/NIAID (Grants R21-AI-59365, K25-AI-064625 and R01-AI-59694) and NIH/National Institute of General Medical Sciences (NIGMS) (Grant R01-GM-2758). Cytokine analysis was a kind gift of Stephen Kingsmore, PhD, and Molecular Staging Incorporated. Genotype analysis was a kind gift of Stephen Chanock, MD, and the NCI Center for Cancer Research. Kathryn Edwards, MD, coordinated the original acquisition of the data analyzed for this study. The United States Environmental Protection Agency (EPA), through its Office of Research and Development, collaborated in the research described here. It has been subjected to Agency review and approved for publication, although it does not necessarily represent the views or polices of the US EPA.

References

1 Kemper AR, Davis MM, Freed GL. Expected adverse events in a mass smallpox vaccination campaign. *Eff Clin Pract* 2002; **5**: 84–90.
2 Reif DM, White BC, Moore JH. Integrated analysis of genetic, genomic and proteomic data. *Expert Rev Proteomics* 2004; **1**: 67–75.
3 Maniolo TA, Collins FS. Genes, environment, health, and disease: facing up to complexity. *Hum Hered* 2007; **63**: 63–66.
4 Nicholson JK. Global systems biology, personalized medicine and molecular epidemiology. *Mol Syst Biol* 2006; **3**: 1–6.
5 Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
6 Lunetta KL, Hayward LB, Segal J, Van EP. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet* 2004; **5**: 32.
7 Robnik-Sikonja M. Improving random forests. *Proc Eur Conf Mach Learn* 2004; **3201**: 359–370.
8 McKinney BA, Reif DM, Rock MT, Edwards KM, Kingsmore SF, Moore JH *et al*. Cytokine expression patterns associated with systemic adverse events following smallpox immunization. *J Infect Dis* 2006; **194**: 444–453.
9 Rock MT, Yoder SM, Talbot TR, Edwards KM, Crowe Jr JE. Adverse events after smallpox immunizations are associated with alterations in systemic cytokine levels. *J Infect Dis* 2004; **189**: 1401–1410.
10 Rock MT, Yoder SM, Talbot TR, Edwards KM, Crowe Jr JE. Cellular immune responses to diluted and undiluted Aventis Pasteur smallpox vaccine. *J Infect Dis* 2006; **194**: 435–443.
11 Talbot TR, Stapleton JT, Brady RC, Winokur PL, Bernstein DI, Germanson T *et al*. Vaccination success rate and reaction profile with diluted and undiluted smallpox vaccine: a randomized controlled trial. *JAMA* 2004; **292**: 1205–1212.
12 Talbot TR, Bredenberg HK, Smith M, LaFleur BJ, Boyd A, Edwards KM. Focal and generalized folliculitis following smallpox vaccination among vaccinia-naive recipients. *JAMA* 2003; **289**: 3290–3294.
13 Garcia-Closas M, Malats N, Real FX, Yeager M, Welch R, Silverman D *et al*. Large-scale evaluation of candidate genes identifies associations between VEGF polymorphisms and bladder cancer risk. *PLoS Genet* 2007; **3**: e29.
14 Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L *et al*. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res* 2006; **34**: D617–D621.
15 Reif DM, McKinney BA, Motsinger-Reif AA, Chanock SJ, Edwards KM, Rock MT *et al*. Genetic basis for adverse events after smallpox vaccination. *J Infect Dis* 2008; **198**: 16–22.
16 Kader HA, Tchernev VT, Satyaraj E, Lejnine S, Kotler G, Kingsmore SF *et al*. Protein microarray analysis of disease activity in pediatric inflammatory bowel disease demonstrates elevated serum PLGF, IL-7, TGF-beta1, and IL-12p40 levels in Crohn's disease and ulcerative colitis patients in remission versus active disease. *Am J Gastroenterol* 2005; **100**: 414–423.
17 Perlee L, Christiansen J, Dondero R, Grimwade B, Lejnine S, Mullenix M *et al*. Development and standardization of multiplexed antibody microarrays for use in quantitative proteomics. *Proteome Sci* 2004; **2**: 9.
18 Schweitzer B, Wiltshire S, Lambert J, O'Malley S, Kukanskis K, Zhu Z *et al*. Inaugural article: immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc Natl Acad Sci USA* 2000; **97**: 10113–10119.
19 Schweitzer B, Roberts S, Grimwade B, Shao W, Wang M, Fu Q *et al*. Multiplexed protein profiling on microarrays by rolling-circle amplification. *Nat Biotechnol* 2002; **20**: 359–365.

20 Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Chapman & Hall: New York, 1984.
21 Province MA, Shannon WD, Rao DC. Classification methods for confronting heterogeneity. *Adv Genet* 2001; **42**: 273–286.
22 Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP *et al.* Identifying SNPs predictive of phenotype using random forests. *Genet Epidemiol* 2005; **28**: 171–182.
23 McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene–gene interactions: a review. *Appl Bioinformatics* 2006; **5**: 77–88.
24 Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann: San Francisco, 2005.
25 Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat* 1996; **5**: 299–314.
26 R Development Core Team. R: a language and environment for statistical computing. R foundation for statistical computing. Available at http://www.R-project.org, 2006.
27 Breiman L, Cutler A. Random forests. Available at http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, 2004.
28 Reif DM, Motsinger AA, McKinney BA, Crowe Jr JE, Moore JH. Feature selection using a random forests classifier for the integrated analysis of multiple data types. In: *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2006, pp 171–178.
29 Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 2004; **8**: 5–147.
30 Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**: 263–265.
31 Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003; **56**: 73–82.
32 Wilke RA, Reif DM, Moore JH. Combinatorial pharmacogenetics. *Nat Rev Drug Discov* 2005; **4**: 911–918.
33 Most J, Schwaeble W, Drach J, Sommerauer A, Dierich MP. Regulation of the expression of ICAM-1 on human monocytes and monocytic tumor cell lines. *J Immunol* 1992; **148**: 1635–1642.
34 Peters W, Charo IF. Involvement of chemokine receptor 2 and its ligand, monocyte chemoattractant protein-1, in the development of atherosclerosis: lessons from knockout mice. *Curr Opin Lipidol* 2001; **12**: 175–180.
35 Zittermann SI, Issekutz AC. Basic fibroblast growth factor (bFGF, FGF-2) potentiates leukocyte recruitment to inflammation by enhancing endothelial adhesion molecule expression. *Am J Pathol* 2006; **168**: 835–846.
36 Eslick J, Scatizzi JC, Albee L, Bickel E, Bradley K, Perlman H. IL-4 and IL-10 inhibition of spontaneous monocyte apoptosis is associated with Flip upregulation. *Inflammation* 2004; **28**: 139–145.
37 Mangan DF, Robertson B, Wahl SM. IL-4 enhances programmed cell death (apoptosis) in stimulated human monocytes. *J Immunol* 1992; **148**: 1812–1816.
38 Soruri A, Kiafard Z, Dettmer C, Riggert J, Kohl J, Zwirner J. IL-4 down-regulates anaphylatoxin receptors in monocytes and dendritic cells and impairs anaphylatoxin-induced migration *in vivo*. *J Immunol* 2003; **170**: 3306–3314.
39 Serhan CN, Savill J. Resolution of inflammation: the beginning programs the end. *Nat Immunol* 2005; **6**: 1191–1197.
40 Hood L. Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev* 2003; **124**: 9–16.

Supplementary Information accompanies the paper on Genes and Immunity website (http://www.nature.com/gene)