INTEGRATED ANALYSIS OF GENETIC AND PROTEOMIC DATA

By

David Michael Reif

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

December, 2006

Nashville, Tennessee

Approved:

Professor James E. Crowe

Professor Douglas H. Fisher

Professor Jonathan L. Haines

Professor Jason H. Moore

Professor Scott M. Williams

This work is dedicated to my family—Mom, Dad, and Dan—for teaching me how to work hard and play nice with others.

And to Alison, for making me happier than I have ever been—no matter what else is going on.

ACKNOWLEDGMENTS

My graduate training was supported by the Vanderbilt University Interdisciplinary Graduate Program (1st year), the NIH Human Genetics Training Grant (2nd-3rd years), and my mentor, Jason H. Moore (4th-5th years).

I want to acknowledge the vast contributions and support of the scientists and staff at both Vanderbilt and Dartmouth Medical School. In the Vanderbilt Center for Human Genetics Research, I wish to thank Jackie Bartlett, Kylee Spencer, Tricia Thornton-Wells, Jacob McCauley, Scott Dudek, Jeff Canter, Marylyn Ritchie, Kim Taylor, Alicia Davis, Lynn Roberts, and Maria Comer. I thank Chun Li for his insights into teaching and his philosophy on statistics in science. At Dartmouth, I would like to thank Todd Holden and Nate Barney.

Special thanks go to Bill White at Dartmouth for his friendship and extensive help with computational issues, as well as stimulating discussions on topics relating to science and beyond.

Special thanks also go to Brett McKinney for his support on both scientific and personal levels. Time and again, his inquisitiveness and optimism helped me find solutions to uncooperative problems.

I am greatly indebted to the members of my thesis committee (James Crowe, Jr., Douglas Fisher, Jonathan Haines, Jason Moore, and Scott Williams) for their invaluable time, guidance, friendship, and support. Their ability to effective guide a project involving biological, computational, genetic, and immunological aspects is a testament to their diverse interdisciplinary expertise and commitment to training students.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

°C	Degrees Centigrade
Ab	Antibody
AE	Adverse Event
APSV	Aventis-Pasteur Smallpox Vaccine
bp	Base Pair
cDNA	Complementary Deoxyribonucleic Acid
CGF	Core Genotyping Facility (http://dceg.cancer.gov/genotype.html)
Celera	Private Genome Assembly Database (http://www.celeradiscoverysystem.com/index.cfm)
CEPH	Centre d'Etude du Polymorphism Humain
CHGR	Center for Human Genetics Research
cM	CentiMorgan
CSF-3	Colony Stimulating Factor 3 (Granulocyte)
CTL	Cytotoxic T-lymphocyte
CV	Cross Validation
dbSNP	Public SNP database (http://www.ncbi.nlm.nih.gov/SNP/index.html)
df	Degrees of Freedom
dNTPs	Deoxyribonucleotides
dHPLC	Denaturing High Performance Liquid Chromatography
DMID	Division of Microbiology and Infectious Diseases (http://www.niaid.nih.gov/dmid)
DNA	Deoxyribonucleic Acid

DZ	Dizygotic Twins
ELISA	Enzyme-linked Immunosorbent Assay
Ensemble	Genome Browser (http://www.ensembl.org/)
FBAT	Family Based Association Tests (http://www.biostat.harvard.edu/~fbat/fbat.htm)
g	Gram
GAM	Generalized Additive Model
GASP	Genometric Analysis Simulation Program
G-CSF	Granulocyte Colony Stimulating Factor
Haploview	Java-based Tool for Visualizing LD blocks (http://www.broad.mit.edu/mpg/haploview/index.php)
HIV	Human Immunodeficiency Virus
HLOD	Heterogeneity LOD Score
HPLC	High Performance Liquid Chromatography
htSNP	Haplotype Tag Single Nucleotide Polymorphism
HWE	Hardy-Weinberg Equilibrium
IC	Imprinting Center
ICAM-1	Intercellular Adhesion Molecule-1
IFN-γ	Interferon-γ
Ig	Immunoglobulin
IL-4	Interleukin-4
IL-10	Interleukin-10
IRF-1	Interferon Regulatory Factor-1
Kb	Kilobase

LD	Linkage Disequilibrium
LINE	Long Interspersed Nuclear Element
LOD	Logarithm of the Odds
MALDI	Matrix Assisted Laser Desporption/Ionization
Mb	Megabase
MED	Maternal Expression Domain
MFI	Mean Fluorescence Intensity
MIG	Monokine Induced by Interferon-γ
ml	Milliliter
MLS	Multipoint LOD Score
MMP	Matrix Metalloproteinase
MMR	Measles-Mumps-Rubella
mRNA	Messenger Ribonucleic Acid
MS	Mass spectrometry
MTHFR	Methylenetetrahydrofolate reductase
MZ	Monozygotic twins
μg	Microgram
μL	Microliter
NCI	National Cancer Institute (http://cancer.gov)
NEMC	New England Medical Center
ng	Nanogram
NIAID	National Institute of Allergy and Infectious Diseases (http://www.niaid.nih.gov/)

NIH	National Institutes of Health (http://www.nih.gov/)
NIMH	National Institute of Mental Health (http://www.nimh.nih.gov/)
nL	Nanoliter
NN	Neural Network
NSC	Nearest Shrunken Centroid
OMIM	Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/Omim)
OOB	Out-Of-Bag
OSA	Ordered-Subsets Analysis
PCR	Polymerase Chain Reaction
PDT	Pedigree Disequilibrium Test
QTL	Quantitative Trait Locus
RCAT	Rolling Circle Amplification Technology
RT	Real Time
RF	Random Forest™
SAGE	Serial Analysis of Gene Expression
SCF	Stem Cell Factor
SDA	Symbolic Discriminant Analysis
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
Taq	*Thermus Aquaticus* Polymerase
TH_1	T-Helper type-1
TH_2	T-Helper type-2
TIMP-2	Tissue Inhibitor of Metalloproteinase -2

UTR	Untranslated Region
VISTA	Visualization Tools for Alignment (http://www.gsd.lbl.gov/vista/)
VNTR	Variable Number Tandem Repeat
VV	Vaccinia Virus

CHAPTER I

INTRODUCTION:

INTEGRATED ANALYSIS OF GENETIC AND PROTEOMIC DATA

Biological organisms are complex systems that dynamically integrate inputs from a multitude of physiological and environmental factors. Complex clinical outcomes arise from the concerted interactions among the myriad components of a biological system. Therefore, in addressing questions concerning the etiology of phenotypes as complex as common human diseases or systemic reaction to vaccination, it is essential that the systemic nature of biology is taken into account. Analysis methods must integrate the information provided by each data type in a manner analogous to the operation of the body itself. It is hypothesized that such integrated approaches will provide a more comprehensive portrayal of the mechanisms underlying complex phenotypes and lend confidence to the biological interpretation of analytical conclusions.

This dissertation concerns the development of the paradigm outlined above and applies it to genetic and proteomic data in both simulated and real analysis situations. Chapters two through six are presented as self-contained studies that review our philosophy and its initial applications, describe analysis of real proteomic data alone, describe analysis of real genetic data alone, describe analysis of simulated proteomic and/or genetic data, and apply all the lessons

learned to combined analysis of real genetic and proteomic data. Regarded in its entirety, this dissertation progresses from philosophical underpinnings to successful applications in a real-world analysis setting.

Chapter II lays out the rationale behind integrated analysis strategies, reviews the current state of the art in combined analysis, and details a simulation study that addresses our hypothesis concerning situations in which the analysis of multiple data types is beneficial. The intuitive, intellectual appeal offered by joint analysis of multiple data types includes the integration of information that is insensitive to spatial and temporal flux (*e.g.* stable genetic polymorphisms found throughout the human genome) with information subject to dynamic changes (*e.g.* protein concentrations measured at multiple time points), the amelioration of possible methodological unreliability by the partial redundancy between biological levels, and the improved generalizability of results that are robust to nonsystematic variability in data from any one source. Our review of the initial forays into the joint analysis of multiple data types finds that these studies, while limited in scope, have yielded interesting results that would have been missed had only one type of data been considered. From the simulation studies, we conclude that the analysis of multiple data types is beneficial when the underlying etiological model is complex and functional biomarkers of any particular data type are missing.

Chapter III introduces the smallpox vaccine trial data to which the data-integration philosophy will be applied. In this chapter, the proteomic portion of the analysis is discussed. The proteomic data are measured concentrations of a

panel of immunological cytokines collected from serum samples at pre- and post-vaccination time points. The analysis identified cytokines whose changes in dynamic concentration after vaccination accurately discriminated between subjects who suffered a vaccine-related adverse event (AE) and those who did not. We developed a model of systemic AEs that implicates a cytokine signature characterized by protraction and/or hyper-activation of inflammatory pathways.

Chapter IV describes the analysis of genetic data gathered as part of the smallpox vaccine trials. In this chapter, the same panel of single-nucleotide polymorphisms (SNPs) was analyzed in two independent studies to investigate the relationship between AEs and stable genetic factors. The second study was held out of our original statistical analysis for use as a validation data set. The significant AE-associated genetic factors that replicated in the validation data set complement the conclusions drawn from the proteomic data. The validated SNPs are within genes involved in processes consistent with previously hypothesized mechanisms relating the development of AEs to prolonged stimulation of inflammatory pathways and imbalance of normal tissue damage repair pathways.

Chapter V introduces random forests (RF) as promising solution to the analysis challenge posed by high-dimensional datasets including interactions among biomarkers of multiple data types. This chapter characterizes the performance of RF on a range of simulated datasets when given genetic data alone, proteomic data alone, or a combined dataset of genetic plus proteomic data. The results indicate that utilizing multiple data types is beneficial when the

disease model is complex and the phenotypic outcome-associated data type is unknown. This study also shed light on the nature of effects that could be detected by random forests analysis. The simulation results were used to refine the parameters of RF implemented for analysis of the combined genetic and proteomic vaccine trial data in Chapter VI.

Chapter VI applies the lessons learned in previous chapters to the analysis of high-dimensional, combined genetic and proteomic data collected to elucidate mechanisms underlying development of adverse events (AEs) in patients following smallpox vaccination. In a two-stage analysis strategy, Random Forests were used to identify the most important genetic and proteomic biomarkers from a combined dataset, then the selected attributes were used to build a final decision tree model of AE development. 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 both genetic and proteomic biomarkers. These results demonstrated the utility of the RF for integrated analytical tasks, while both enhancing and reinforcing our working model of AE development following smallpox vaccination.

Chapter VII discusses future directions for integrated analysis strategies that capitalize on the lessons learned in this dissertation. It is hoped that this body of work lends credence to the notion that integration of multiple data types is the only way to truly represent a complex system. Given the rapid expansion of technologies able to generate immense quantities of data, it is anticipated that

the incorporation of multiple data types will become the standard—rather than the exception—for studies of complex human health and disease.

CHAPTER II

INTEGRATED ANALYSIS OF GENETIC, GENOMIC, AND PROTEOMIC DATA

The rapid expansion of methods for measuring biological data ranging from DNA sequence variations through mRNA expression through protein abundance presents the opportunity to utilize multiple types of information jointly in the study of human health and disease. Organisms are complex systems that integrate inputs at myriad levels to arrive at an observable phenotype. Therefore, it is essential that questions concerning the etiology of phenotypes as complex as common human diseases take the systemic nature of biology into account and integrate the information provided by each data type in a manner analogous to the operation of the body itself. While limited in scope, the initial forays into the joint analysis of multiple data types have yielded interesting results that would not have been reached had only one type of data been considered. These early successes, along with the aforementioned theoretical appeal of data integration, provide impetus for the development of methods for the parallel, high-throughput analysis of multiple data types. We present as a working hypothesis the idea that the integrated analysis of multiple data types will improve the identification of biomarkers of clinical endpoints such as disease susceptibility.

A Case for integrated analysis of multiple data types

Technology has advanced to the point that variations in DNA sequence, mRNA expression levels, and a wide spectrum of protein abundance can each be measured with manageable efficiency. The development of single nucleotide polymorphism (SNP) typing technology can identify minute DNA sequence variations between samples [1-5]. Oligonucleotide and cDNA microarrays can simultaneously measure the expression (mRNA) levels of thousands of genes simultaneously [6-8]. Mass spectrometry (MS) techniques can characterize huge swatches of the spectrum of proteins in a given sample [9-13]. Taken together, these technologies provide a veritable flood of information to the researcher. Given the wealth of publications devoted to extending these methods, as well as their becoming less expensive and more accessible, it is expected that the availability of such data will continue to expand [14-16].

Here, we present a working hypothesis that the joint analysis of multiple data types will improve the detection of biomarkers diagnostic of clinical endpoints. The expected benefits offered by joint analysis of multiple data types over singular analysis include provision of surrogate data to fill gaps in data from any one biological level, amelioration of some methodological unreliability via the partial redundancy between stages, integration of information that is *in*sensitive to spatial and temporal flux (e.g. SNPs) with information subject to dynamic changes (mRNA, protein), and recognition that organisms are systems comprising many layers of complexity. We review the state of the art in joint

analysis of multiple data types and then present a preliminary simulation study that addresses our working hypothesis.

Organisms as Complex Systems

The huge bodies of data generated by high-throughput experiments have given rise to the notion that analysis methods for "omic" data are needed [14, 17]. Presently, the analysis methods concentrate on mining data generated by a single type of experiment. Ge *et. al* [14] call for the integration of functional genomic and proteomic techniques with annotation information, signaling a step toward joint analysis—transitioning from traditional, stand-alone biology towards a systemic "modular biology" approach. A modular biology approach studies biological processes of interest (modules) as complex systems of functionally interacting components. Incorporating annotation information provides a more complete picture of the organismal system, complementing and extending the information provided by raw experimental data. While the use of annotation information is attractive, limiting factors include the unreliability of available annotation databases and the wide variability of information provided by such data sources [18, 19].

Initial attempts aimed at developing methods for incorporating multiple types of experimental data into analysis of a biological system have met with some success. For example, Perrin *et al.* have developed an array method to measure a limited collection of nucleic acids and proteins in a single experiment

[20]. Yeger-Lotem and Margalit have integrated information from various cellular networks to detect regulatory circuits in S. *Cerevisiae* [21]. Other groups, using lower animals as experimental models, have made strides toward an integrative analysis of multiple data types on a small scale [22, 23]. However, at present, high-throughput analysis methods for human data have not been put forth, and most studies thus far have concentrated on development of methodological measurement reliability, rather than procedures for the analysis itself.

Biological complexity along the progression from genotype to phenotype

The central dogma of biology states that information progresses from DNA to mRNA to protein [24]. At each stage in this hierarchy, variation is introduced, meaning that inferences made about a later stage based upon measurements taken at an earlier stage will have an inherent amount of uncertainty. Indeed, there have been many studies published regarding the poor correlation of mRNA levels with protein [8]. Additionally, an enormous diversity of RNA transcripts and proteins is encoded by a given DNA sequence [24]. Since living organisms are complex systems, it follows that the study of their inner workings will be replete with emergent properties that are not predictable from the simple sum of parts [25]. The dynamic flux of protein levels is more complex than can be inferred by examining simple mRNA transcripts, and far more complex than can be inferred by examining DNA sequence. The same holds for the prediction of mRNA levels from DNA. While it may be possible to characterize certain SNPs as up- or

down-regulating expression, mRNA levels in the organism as a whole cannot be perfectly predicted at this time. Thus, examination of only a single type of data does not provide a valid description of any biological system.

The sources of introduced variation are myriad (see Figure 1). Between DNA sequence and production of mRNA transcripts, there exists transcriptional control by proteins, proximal and distal control elements, imprinting via methylation, action by enzymes such as histone acetylases/deacetylases, and differences in stability of transcripts [24, 26]. For example, DNA methylation is known to affect gene expression and genomic stability, with major implications in human disease [27]. The initial mRNA transcript is still subject to multiple layers of modification before it is translated into protein. Recent studies have highlighted the prevalence of alternative splicing and A to I editing [28]. Many transcripts are cleaved into multiple bioactive products [29]. Additionally, there are translational controls that determine first if, then the abundance of, translation of a particular messenger RNA into peptide. Upon translation, the polypeptide chain must still fold into its functional conformation. Folding is a complex process requiring the interplay of intracellular conditions, chaperones, and other factors that vary by cell. Once a protein has reached its native conformation (folded state), it has a finite lifetime subject to cellular conditions, targeted degradation by ubiquitination, and enzymatic modification—all of which create wide temporal flux in protein levels. Proteins must be transported to the proper location, then correctly processed by the cell in which they are needed, adding spatial variability to protein levels within multicellular organisms. Thus, the simple

presence of a protein at any one experimental point in time and space may not be representative of the true biology of an organism.

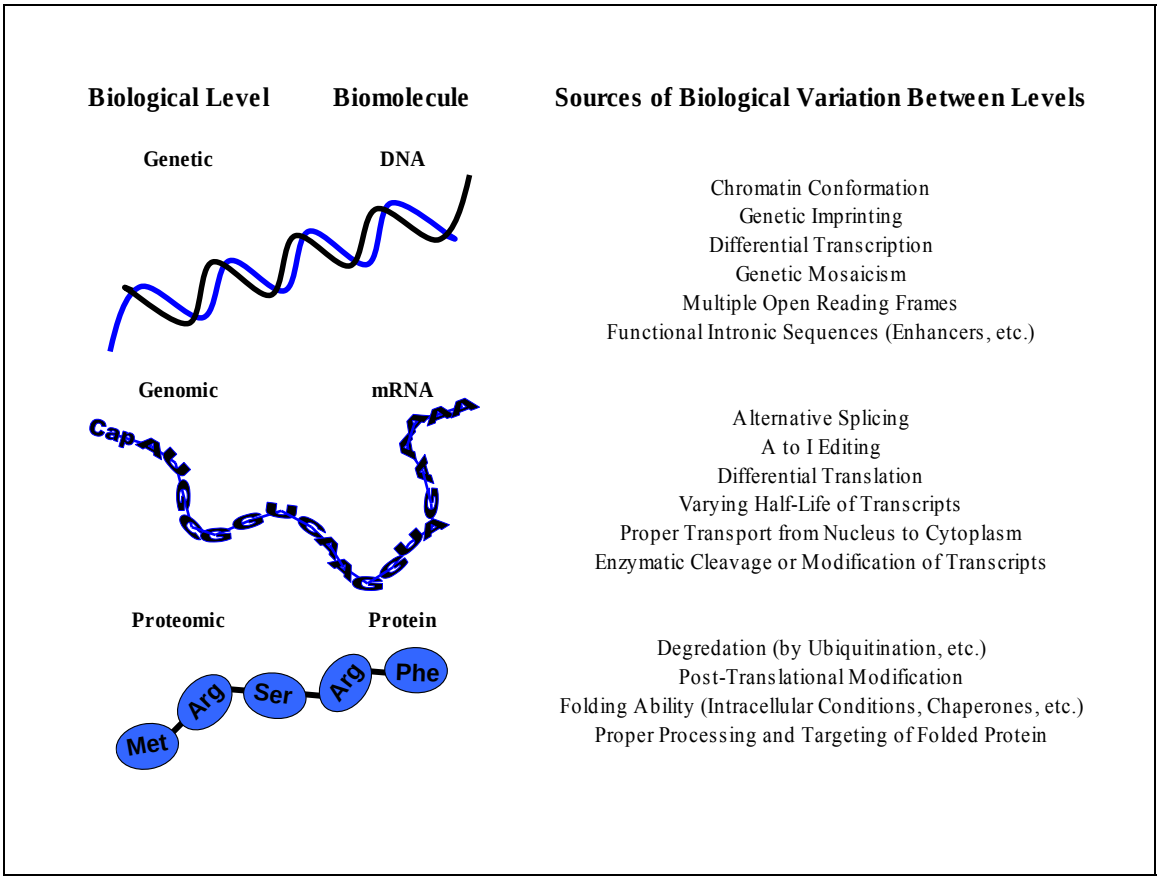


Figure 1. Sources of variation along the biological progression from gene to protein. For each biological level, those phenomena listed introduce variability from 1) DNA sequence through nascent mRNA transcript, 2) immature RNA transcript through nascent polypeptide, and 3) unfolded polypeptide through protein in its native conformation.

A typical simplifying corollary of the central dogma is that phenotype is determined solely by the action of proteins. Adhering to such a model, measuring protein levels alone would be perfectly predictive of disease. In vivo, each step in the progression exerts influence over the other steps, both along the

normal progression and in a feedback manner. The idea of deviations from the central dogma is well documented [24, 30]. Proteins such as transcription factors regulate the expression of genes. Members of the mammalian LINE-1 family encode the necessary products to ensure retrotranscription. Small interfering RNAs mediate post-transcriptional gene silencing via the RNA interference pathway [31]. Proteins regulate other proteins via ubiquitins functioning in degredation. Outside environmental influences may also alter the normal progression. Given these deviations from the central dogma, it is important to obtain information from multiple levels of the hierarchy. It is evident that measuring proteins alone could miss vital information regarding the enormous complexity of biological systems.

Methodology concerns and missing data

Useful data is currently measurable at each of the three main stages along the biological progression from DNA to mRNA to protein. However, the techniques used to gather data at each stage introduce experimental error in excess of the inherent biological variation in measurement. Current SNP typing methods can accurately and rather efficiently identify differences in nucleotide sequences. The primary limitation to gathering SNP data is the cost—in terms of both time and money—of sample acquisition. Although the monetary cost of SNP typing is steadily decreasing, there remain technical issues with such popular methods as Matrix Assisted Laser Desorption/Ionization Mass

Spectrometry (MALDI-MS), where efficiency is limited by the size of DNA products that can be analyzed and the stringent purification necessitated because of adduct formation of alkali ions with the phosphate linkages of DNA [5]. Techniques such as Serial Analysis of Gene Expression (SAGE), RT-PCR, and Oligonucleotide or cDNA microarrays can quantitatively measure gene expression levels. Microarrays and SAGE can measure expression levels for thousands of genes simultaneously. Nonetheless, substantial question marks with these high-throughput methods include the binding behavior of promiscuous probes in a convoluted solution, quantitative reliability, and the fact that gene expression is both temporally and spatially variable—meaning that microarray results only represent conditions at a particular time point in a particular population of cells. The various flavors of mass spectrometry are adept at identifying proteins in a sample. However, there are important sources of unreliability in MS experiments, including the complex physicochemistry of samples with differing ionization tendencies and structural complications, the difficulty in tuning the instrument to accurately measure a broad mass range of samples, and the correct separation of peaks in spectra [32-35]. Of vital importance is the fact that the wide spatial and temporal flux of proteins in an organism means that even a perfect measurement is at best a chance snapshot of proteomic action.

Outside of data that may be missing due to technical errors, it is very probable that important data could be missing because researchers chose not to collect it. For example, SNP typing usually focuses on coding sequences, yet

involvement of distal control elements or other non-coding regions of DNA is commonplace [36]. Additional monetary concerns govern any experiment, limiting the amount and types of data that may be collected. Time is also a factor, as the pressure to publish and the transiency of personnel put effective limits on the duration of a study. These factors may limit the collection of data at a given level. Missing data, whether arising from methodological error or holes in experimental design, can confound any analysis and thus inferences made about molecular etiology. Such a scenario presents an excellent case for the integration of information from multiple biological levels.

Joint analysis simulation study

We anticipate that the collection of suitable data for joint analysis will become commonplace in the near future. As preparation for the availability of such data, we have developed a simulation to test our working hypothesis that the integrated analysis of multiple data types will improve the identification of biomarkers of clinical endpoints. The simulation represents an experiment in which SNP and protein data have been collected for two hypothetical diseases. Certain variables are then selectively deleted from the complete SNP and protein dataset to represent a situation in which relevant information is missing from the data to be analyzed. These “missing data” variants of only SNP, only protein, or partially missing protein data are evaluated to discern whether joint analysis offers benefits in any of these situations. Details of the simulation study—

including the simulation models generated, the datasets analyzed, the analysis method chosen, the software and hardware used, and the results—are presented in subsequent sections of this chapter. A discussion of the relevance of this study and its application to real data is also presented.

Simulation models

Figure 2 illustrates the general modeling strategy. We begin by simulating two unlinked and uncorrelated SNPs with equal allele frequencies and genotypes consistent with Hardy-Weinberg proportions. Each SNP additively explains 30% or 60% of the variation in its respective protein levels. Thus, the mean protein level associated with the heterozygotes is midway between the two homozygotes. Genotypes and protein levels were simulated using the Genometric Analysis Simulation Package or GASP [37]. Each protein level is then categorized as high, medium, or low with frequencies of 0.25, 0.50, and 0.25, respectively. Disease susceptibility is dependent on an interaction between the two proteins. Under disease model A, subjects affected with the disease have medium or high protein levels for both proteins while those that are unaffected have low protein levels for at least one of the proteins. Under disease model B, subjects are affected if they have medium protein levels for the first protein or the second protein but not both. This is based on the nonlinear XOR function that is not linearly separable. The difference between these two models is that the two proteins in model A each have an independent main effect on disease susceptibility in addition to an interaction effect while the proteins in

model B only have an interaction effect. Models similar to A and B have been described previously by Li and Reich [38] and Moore *et al*. [39]. A total of four different simulation models were used in the present study. Each model combined the amount of protein variation explained (30% or 60%) with each disease model (A or B).

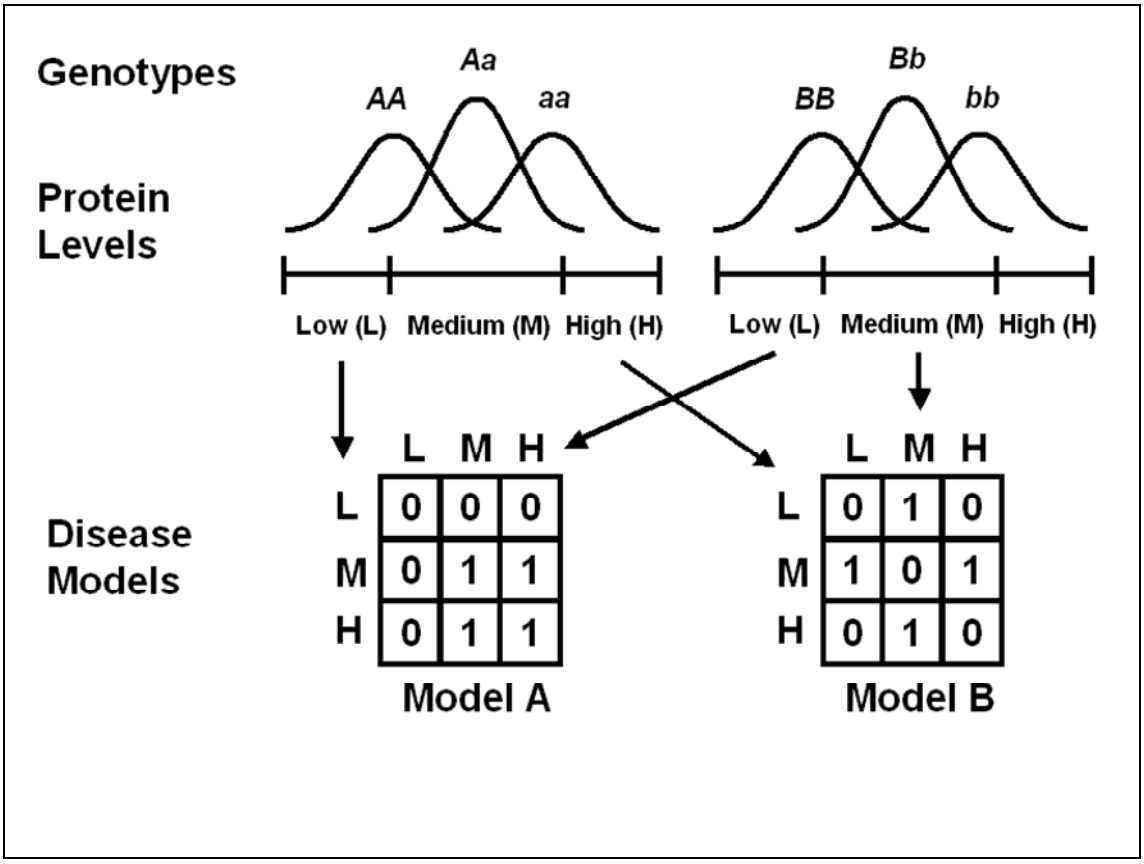


Figure 2. Summary of the simulation models. Protein levels are simulated using an additive genetic model that explains either 30% or 60% of their variation. Protein levels are then discretized into high, medium, and low groups. Under disease models A and B, the probability (P) of disease (D) is dependent on the combination of protein levels (PL) present. Here, P(D|PL) = 0 or 1.

Datasets

Each dataset consisted of a total of 100 subjects that were simulated using each of the four models. Approximately half of the subjects were affected and half unaffected. A total of 100 datasets were simulated using each of the four models. We then took each dataset and created seven new datasets that consisted of 1) all the SNP and protein variables, 2) both SNPs and protein 1, 3) both SNPs and protein 2, 4) just the two SNPs, 5) just the two proteins, 6) just protein 1, and 7) just protein 2 (Figure 3). This study design allows us to evaluate the benefit of having multiple data types when all the functional variables are present in the dataset or only certain subsets of variables are present.

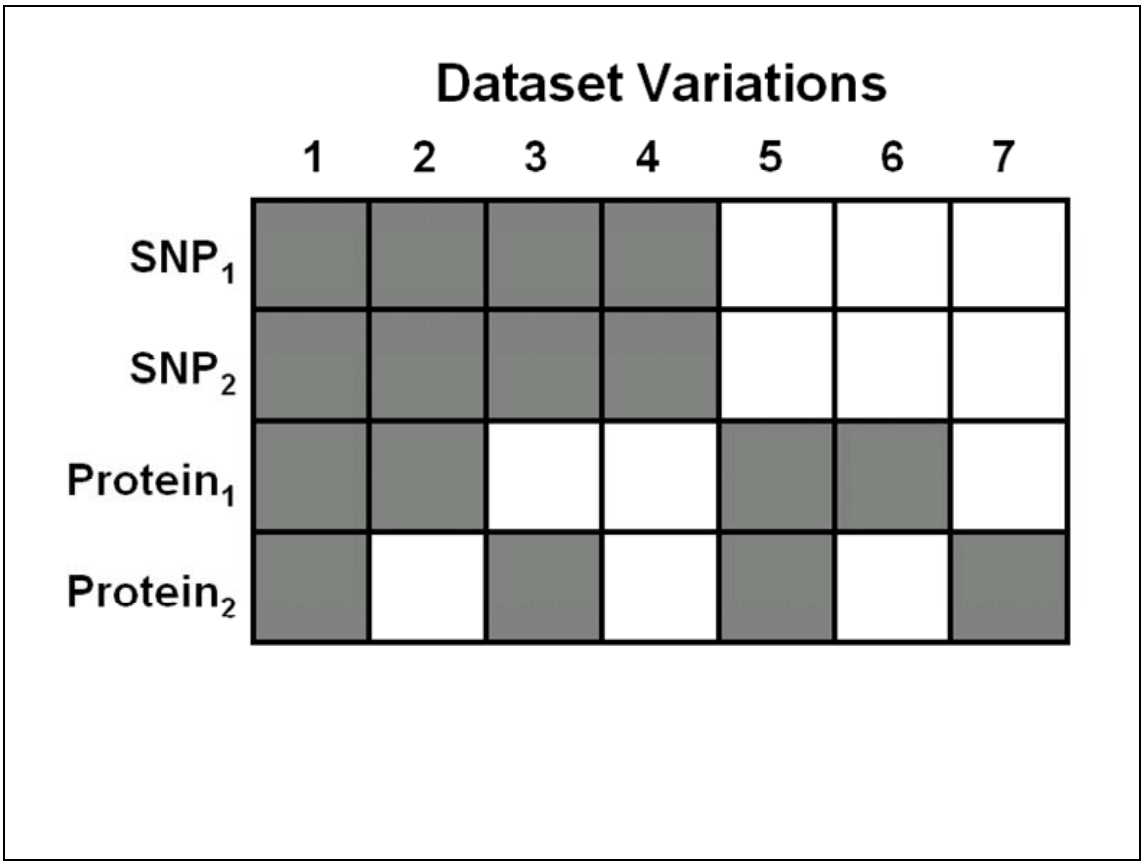


Figure 3. Summary of the dataset variations analyzed. Variables included in each dataset variation are shaded

Data analysis

The test of our working hypothesis that the integrated analysis of multiple data types will improve the identification of biomarkers of clinical endpoints involved two primary analysis goals. The first goal was to model the relationship between each set of genetic and proteomic variables and the clinical endpoint. While many analysis methods, such as Neural Networks, Regression, Generalized Additive Models, and others may prove useful for accomplishing this goal, the symbolic discriminant analysis method, or SDA [40-43], was selected for use here because of its flexibility for modeling different data types. SDA is a

supervised pattern mining approach that carries out variable selection and model selection simultaneously and automatically. Using evolutionary computation as the parallel search strategy, SDA builds discriminant functions from a list of mathematical operators and explanatory variables that can distinguish between disease classes in the data. In this study, we provided the selected genetic and proteomic explanatory variables, plus basic model building blocks consisting of arithmetic functions (e.g. +, -, *, /) and additional mathematical functions (e.g. log, exp, sqrt, abs, sine, cos) as has been suggested by Reif *et al*. [43]. SDA was thus free to construct classification models consisting of any combination of the above mathematical functions and biological variables, without any further *a priori* specification of model structure. Therefore, no assumptions about the relationships among the variables need be pre-specified, and since it has the flexibility to operate on continuous or discrete variables, SDA is a logical choice for handling multiple data types. The goal of the evolutionary search is to identify the combination of variables and functions that minimizes the overlap of the distributions of symbolic discriminant scores among affected and unaffected subjects. A classification error of zero indicates there is no overlap among the symbolic discriminant score distributions. Because we are modeling only functional variables in this study, we applied SDA directly to the entire dataset to get an estimate of the classification error. Thus, overfitting (spurious selection of noise variables) is not a concern here and the cross-validation and permutation-testing methods suggested by Moore [44] are not necessary.

The second goal of the data analysis was to determine whether there are differences in classification error when different subsets of variables are used in the analysis, simulating a situation wherein data are missing. Mean classification errors between dataset types defined in Figure 3 were evaluated using a paired t-test. The resulting p-values for the differences in the mean classification errors between datasets were compared to determine which types of missing data had the most significant effect on disease classification and to identify situations in which integrating multiple data types was beneficial.

Software and hardware

The SDA algorithms are programmed in C and integrated into the lil-gp software package (http://garage.cps.msu.edu/software/software-index.html) that was used to carry out genetic programming. In this study, we carried out the parallel search using grammatical evolution, a variation on genetic programming that utilizes Backus-Naur Form grammars to specify construction of SDA models [45]. The SDA modeling was carried out on the VAnderbilt Multi-Processor Integrated Research Engine or VAMPIRE, a 380-processor Beowulf-style parallel computer system running the Linux operating system. Each population consisted of 100 individuals. We allowed the genetic programs to run a total of 100 iterations. A recombination frequency of 0.6 was used along with a mutation frequency of 0.02. These parameters are standard for evolutionary searches [46].

Simulation results and discussion

Table 1 summarizes the average classification errors across the 100 datasets for each of the four models and each of the combinations of variables analyzed. Figure 4 illustrates the statistical comparison of the mean classification errors resulting from the analysis of each combination of variables for each of the four models. As in real-world data, there was overlap in the distribution of continuous protein values, thus precluding SDA from achieving perfect classification. For model A, SDA achieved the lowest classification error when given both functional proteins alone. This is expected because disease status was assigned based upon protein levels, meaning there is no noise in this dataset. The phenomenon of increased classification error in datasets with both SNP and protein data compared with both proteins alone would have been mitigated using models with higher heritability between genotype and protein level. The mean classification errors for those datasets consisting of only SNPs was significantly higher than those including protein data as the variation explained by SNPs decreased to 30%. Such a result is expected since there was not a deterministic relationship between the SNPs and the protein levels. SNP data was of increasing utility for classification as the percent variation in protein levels explained by the SNPs increased from 30% to 60%. In the case where SNP variation explained 60% of variation in protein levels, the inclusion of SNP data with either protein classified significantly better than either protein alone. The mean classification error associated with either protein (P_1 or P_2) alone was consistently high, indicating that information (in SNP or protein form)

on both functional proteins is necessary for classification of the disease endpoint. The results for model A suggest that having multiple types of data is beneficial when the etiological model is complex and one or more variables may be missing.

Table 1. Summary of the average classification errors across 100 datasets for each model and each type of dataset analyzed.

Model[2]	Mean classification error for each type of dataset[1]						
	$S_1S_2P_1P_2$	$S_1S_2P_1$	$S_1S_2P_2$	S_1S_2	P_1P_2	P_1	P_2
A 30%	0.1438	0.1981	0.1877	0.3098	0.1133	0.1891	0.1855
A 60%	0.1359	0.1584	0.1550	0.2108	0.1008	0.1913	0.1937
B 30%	0.3313	0.3567	0.3573	0.3946	0.2900	0.3532	0.3517
B 60%	0.3068	0.3231	0.3226	0.3673	0.2848	0.3468	0.3518

[1] Combination of variables analyzed where S_1 = SNP_1, S_2 = SNP_2, P_1 = $Protein_1$, P_2 = $Protein_2$

[2] A = disease model A (interaction plus main effect), B = disease model B (just interaction), 30% = 30% of the protein variation explained by additive genetic model, 60% = 60% of the protein variation explained by additive genetic model.

The results for model B mirror the pattern seen with model A. The major difference is that the raw classification errors across model B datasets are higher because the underlying disease model is more complex. The best results were for dataset variants in which both proteins were present because other variants are missing a critical variable for modeling the interaction. This is expected since model B is not linearly separable, and there is thus a deterministic relationship between *both* protein levels and disease risk. As in model A, the inclusion of SNP data was of additional utility as the variation in protein levels explained by

the SNPs increased. These results suggest that the joint analysis of multiple data types—in this case, SNP and protein—improves modeling when one of the functional proteins is absent and the etiological model consists of a nonlinear interaction in the absence of main effects.

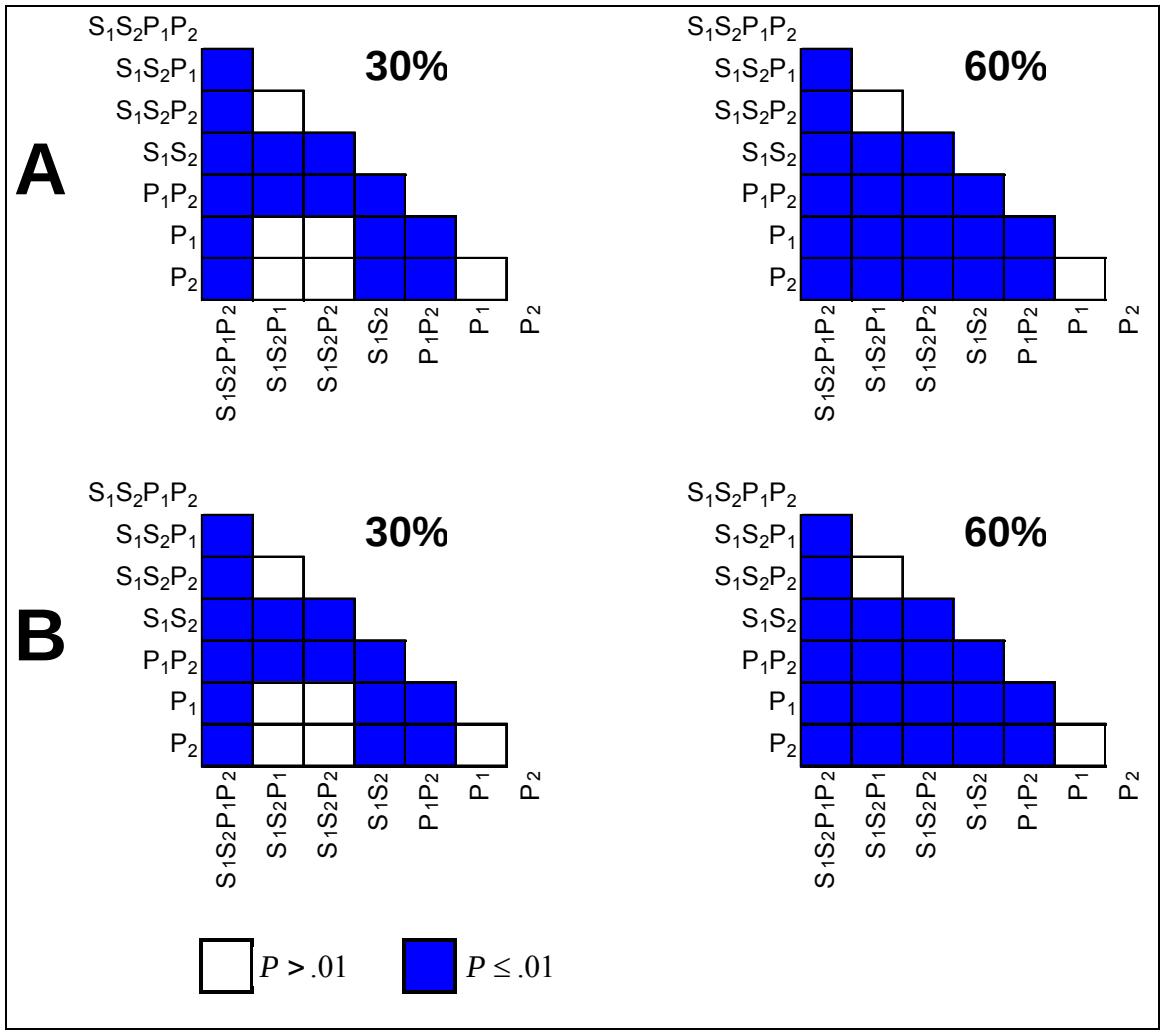


Figure 4. Summary of the statistical comparison of mean classification errors for each type of dataset for a given heritability (30% or 60%) and model (A or B). Mean classification errors between dataset types defined in Figure 3 were evaluated using a paired t-test. The resulting p-values for the differences in the mean classification errors between datasets are shaded according to the level of statistical significance.

Relevance of the joint analysis simulation study and application to real data

How realistic are the disease models?

For this study we assumed that disease risk was determined by an interaction between two proteins. The primary difference between interaction models A and B is that each protein in model A also has an independent main effect on disease risk whereas the proteins in model B only influence disease through a nonlinear interaction. The ultimate utility of this study depends on how realistic these models are. While it is unlikely that any human disease follows either of these models exactly, Moore [47] has made the argument that nonlinear interactions among biomarkers are likely to play a more important role in the etiology of common diseases than the independent main effects of any one biomarker. This argument is based on several key ideas. First, the idea that interactions are important has been around for nearly 100 years [48]. Second, the ubiquity of biomolecular interactions at the transcriptional, translational, and biological network levels suggests that interactions are likely to play a very important role in disease susceptibility. Third, studies of single biomarkers typically don't replicate. Finally, nonlinear interactions are commonly found when properly investigated. Thus, while model B may not be an accurate model for any one disease, it does fall into a category of models that are likely to represent the complexity of the genotype to phenotype mapping relationship.

How realistic is the scenario in which key functional proteins will be missing from the data analyzed?

The models used in this study assume that proteins are the key etiological agents for determining disease susceptibility. How likely is it that one or more of the functional proteins might be missing from a given dataset? Given the technical difficulties in accurately measuring and reliably identifying large numbers of proteins in a single experiment, it is very likely that there may be holes in the protein profile. The current state of the art is to employ some combination of methods such as 2-D gels, HPLC, tryptic digestion, and one of the variety of mass spectrometric methods. Each of these procedures introduces its own set of methodological biases and is ideally honed to precisely identify proteins meeting a narrow range of criteria. Thus obtaining a reliable portrait of a wide range of proteins—both within and across samples—is a vexing problem. In a mass spectrometric analysis, the chemical noise characteristic of the raw data is normalized away—often obscuring or deleting peaks representing proteins in low abundance, which may be important players in protein-protein interactions. The correct identification of a particular protein species' spectral peak in large-scale spectrometric analyses is an active area of research for both academics and instrumentation providers [35]. Aside from the procedural difficulties of proteomics, the dynamic nature of proteins in tissue provides a daunting challenge. Proteins are in continual spatial and temporal flux; thus even an experimentally perfect profile of proteins would represent only a snapshot of protein action in the organism for a given region and a given time slice. Additionally, preserving the native state of protein molecules subject to

denaturation, post-translational modification, and other physico-chemical alterations until they can be processed can confound any analysis.

How realistic is the scenario in which functional SNPs are measured when key functional proteins are not?

SNPs hold great promise as biomarkers of human disease for several reasons. First, more than 10 million SNPs have been described throughout the human genome. Efforts are underway to determine the minimal subset necessary to capture all the common variation in the genome. Second, they are relatively easy to measure using a variety of high-throughput technologies. As these methods become less expensive over the next several years it will be possible to measure hundreds of thousands to millions of SNPs in each of thousands of samples. One can envision a time in the near future when it will be possible to measure a set of non-redundant SNPs in every gene in the genome—although the availability of genome-wide SNP data presents its own set of computational challenges [49]. Third, barring somatic mutations, SNPs do not change in time and space in an individual. This is in contrast to both mRNA and protein expression levels that are highly variable across both time and space. Fourth, SNPs can have functional consequences on both the levels and types of proteins expressed. Given the limitations of proteomic technologies as described above, SNPs hold great promise as biomarkers of human disease. This study indicates that the addition of SNPs to protein information may be beneficial. It is reasonable to assume that it will be easier, and perhaps even less expensive, to measure a comprehensive set of SNPs than a comprehensive

set of expressed proteins due to technological limitations and the enormous variability of protein expression. If this is true, combining SNPs with proteomics data will be a powerful strategy.

Conclusions and future directions

In the present study, we present a working hypothesis that the joint analysis of genetic and proteomic data will provide more information for modeling disease susceptibility than either alone. In the context of the simulations performed, we conclude that the availability of multiple types of data is beneficial when the underlying etiological model is complex and one or more of the functional variables are missing. These results provide a baseline for those planning to collect and/or analyze genetic, genomic, and proteomic data from the same samples.

This study represents a first step towards evaluating the merits of combining genetic, genomic, and proteomic data from the same samples for the detection and characterization of biomarkers of human disease susceptibility. From these initial simulation studies, we make the following recommendations. First, when the underlying etiology of the disease is likely to be complex, measuring multiple types of data is advantageous, especially if it is also likely that the technologies are limited in their ability to measure all biomarkers. Thus, we recommend that SNP data be measured in addition to gene expression and/or protein data. Second, we recommend that the multiple types of data be

analyzed jointly. In the present study, a SNP-protein interaction was found when the etiological model consisted of two interacting proteins and one of the two proteins was missing for technical reasons from the datasets. It is interesting to note that the analysis of each type of data separately may also be beneficial. For example, in the case that the functional SNPs and the functional proteins are all present in their respective datasets, separate analyses may provide a type of cross-validation. That is, confidence in the inferences made about the functional biomarkers could be increased if the SNPs and proteins discovered through statistical modeling are related to the same set of genes. Finally, we recommend that additional simulations be carried out under a wider array of etiological models and dataset variations to fully evaluate the usefulness of the joint analysis of multiple types of data. These types of studies should prove invaluable to those planning to measure genomic and proteomic data from the same samples.

The next five years will see the joint analysis of multiple data types become the standard, rather than the exception, in the study of complex human health and disease. Given the rapid expansion of technologies able to generate huge bodies of data, as well as their increasing acceptance in the biomedical research community, we anticipate real datasets appropriate for joint analysis will become increasingly common in the near future. The burgeoning field of research into high-throughput technologies will lead to continued improvements in cost-efficiency and reliability and make their use even more widespread. With these data in hand, joint analysis of multiple biological levels becomes a viable

option. The notion that integration of multiple data types is the only way to truly represent a complex system flows naturally from the complexity revealed as biologists gain a deeper understanding of common disease etiologies.

Summary of key issues

- Biological organisms are complex systems integrating information at myriad levels to arrive at observable phenotypes.
- Achieving a meaningful understanding of complex phenotypes demands the joint analysis of multiple types of information.
- Development of high-throughput technologies will continue; nonetheless, there will always be issues—whether reflecting biological flux or methodological error—with data collected from any single experiment.
- Benefits offered by joint analysis of multiple data types over singular analysis include provision of surrogate data to fill gaps in data from any one biological level, amelioration of some methodological unreliability via the partial redundancy between stages, integration of information that is *in*sensitive to spatial and temporal flux (e.g. SNPs) with information subject to dynamic changes (mRNA, protein), and evaluation of organisms as systems comprising many layers of complexity.
- Datasets amenable to integrated analysis will become increasingly common in the near future, and the joint analysis of multiple data types will become the norm, rather than the exception.

Acknowledgments

This work was supported by generous funds from the Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation. This work was also supported by National Institutes of Health grants HL-68744, CA-084239, CA-90949, CA-95103, CA-98131, LM-07613, AI-057661, and GM-62758-04.

References

1. Pusch W, Wurmbach JH, Thiele H *et al.* MALDI-TOF mass spectrometry-based SNP genotyping. *Pharmacogenomics*. 3(4), 537-48 (2002).

2. Ye S, Liang X, Yamamoto Y *et al.* Detection of single nucleotide polymorphisms by the combination of nuclease S1 and PNA. *Nucleic Acids Res Suppl*. (2), 235-6 (2002).

3. Bocker S. SNP and mutation discovery using base-specific cleavage and MALDI-TOF mass spectrometry. *Bioinformatics*. 19 Suppl 1, I44-I53 (2003).

4. Iwasaki H, Ezura Y, Ishida R *et al.* Accuracy of genotyping for single nucleotide polymorphisms by a microarray-based single nucleotide polymorphism typing method involving hybridization of short allele-specific oligonucleotides. *DNA Res*. 9(2), 59-62 (2002).

5. Sauer S, Gut IG. Genotyping single-nucleotide polymorphisms by matrix-assisted laser-desorption/ionization time-of-flight mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*. 782(1-2), 73-87 (2002).

6. Liang M, Cowley Jr AW, Greene AS. High throughput gene expression profiling: A molecular approach to integrative physiology. *J Physiol*. [Epub ahead of print] (2003).

7. Huang JX, Mehrens D, Wiese R *et al.* High-throughput genomic and proteomic analysis using microarray technology. *Clin Chem*. 47(10), 1912-6 (2001).

8. Grant GR, Manduchi E, Pizarro A *et al*. Maintaining data integrity in microarray data management. *Biotechnology Bioengineering*. 84 (7), 795-800 (2003).

9. Lion N, Rohner TC, Dayon L *et al.* Microfluidic systems in proteomics. *Electrophoresis*. 24(21), 3533-62 (2003).

10. Marko-Varga G, Nilsson J, Laurell T. New directions of miniaturization within the proteomics research area. *Electrophoresis*. 24(21), 3521-32 (2003).

11. Bodnar WM, Blackburn RK, Krise JM *et al.* Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage. *J Am Soc Mass Spectrom*. 14(9), 971-9 (2003).

12. Zhang S, Van Pelt CK, Henion JD. Automated chip-based nanoelectrospray-mass spectrometry for rapid identification of proteins separated by two-dimensional gel electrophoresis. *Electrophoresis*. 24(21), 3620-32 (2003).

13. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 422(6928), 198-207 (2003).

14. Ge H, Walhout AJ, Vidal M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*. 19(10), 551-60 (2003).

15. Cavalcoli JD. Genomic and proteomic databases: large-scale analysis and integration of data. *Trends Cardiovasc Med*. 11(2), 76-81 (2001).

16. Navarro DJ, Niranjan V, Peri S. From biological databases to platforms for biomedical discovery. *Trends in Biotechnology*. 21(6), 263-268 (2003).

17. Celis JE, Gromov P, Gromova I *et al.* Integrating Proteomic and Functional Genomic Technologies in Discovery-driven Translational Breast Cancer Research. *Mol Cell Proteomics*. 2(6), 369-77 (2003).

18. Camon E, Magrane M, Barrell D *et al.* The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*. 13(4), 662-72 (2003).

19. Mitchell JA, McCray AT, Bodenreider O. From phenotype to genotype: issues in navigating the available information resources. *Methods Inf Med*. 42(5), 557-63 (2003).

20. Perrin A, Duracher D, Perret M *et al.* A combined oligonucleotide and protein microarray for the codetection of nucleic acids and antibodies associated with human immunodeficiency virus, hepatitis B virus, and hepatitis C virus infections. *Anal Biochem*. 322(2), 148-55 (2003).

21. Yeger-Lotem E, Margalit H. Detection of regulatory circuits by integrating the cellular networks of protein-protein interactions and transcription regulation. *Nucleic Acids Res*. 31(20), 6053-61 (2003).

22. Walhout AJ, Reboul J, Shtanko O *et al.* Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr Biol*. 12(22), 1952-8 (2002).

23. Ge H, Liu Z, Church GM *et al.* Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet*. 29(4), 482-6 (2001).

24. Strachan T, Read AP. DNA structure and gene expression. In: *Human Molecular Genetics* (3rd Ed.). Garland Science, New York, USA, 13-31 (2004).

25. Holland JH. *Hidden Order: How Adaptation Builds Complexity*. Perseus Publishing, Cambridge, USA (1996).

26. Luthi-Carter R, Apostol BL, Dunah AW *et al*. Complex alteration of NMDA receptors in transgenic Huntington's disease mouse brain: analysis of mRNA and protein expression, plasma membrane association, interacting proteins, and phosphorylation. *Neurobiol Dis*. 14(3), 624-36 (2003).

27. Novik KL, Nimmrich I, Genc B *et al*. Epigenomics: Genome-Wide Study of Methylation Phenomena. *Current Issues in Molecular Biology*. 4(1), 111-128 (2002).

28. Mass S, Rich A, Nishikura K. A-to-I RNA editing: recent news and residual mysteries. *Journal of Biological Chemistry*. 278(3), 1391-1394 (2003).

29. Tankaka S. Comparative aspects of intracellular proteolytic processing of peptide hormone precursors: studies of proopiomelanocortin processing. *Zoological Science*. 20(10), 1183-1198 (2003)

30. Mattick JS. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*. 25(10), 930-939 (2003).

31. Wall NR, Shi Y. Small RNA: can RNA interference be exploited for therapy? *Lancet*. 362(9393), 1401-1403 (2003).

32. Mamyrin BA. Time of Flight Mass Spectrometry: Concepts, Achievements, and Prospects. *International Journal of Mass Spectrometry*. 206, 251-266 (2001).

33. Gentzel M, Kocher, T, Ponnusamy S *et al.* Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics*. 3(8), 1597-16-10 (2003).

34. Liebler DC. In: *Introduction to Proteomics, Tools for the New Biology*. Human Press, Totowa, NJ, USA, 62 (2002).

35. Wool A, Smilanksy Z. Precalibration of matrix-assisted laser desorption/ionization-time of flight mass spectra for peptide mass fingerprinting. *Proteomics*. 2 (10), 1365-1373 (2002).

36. Xin L, Liu DP, Ling CC. A hypothesis for chromatin domain opening. *Bioessays*. 25 (5), 507-514 (2003).

37. Wilson AF, Bailey-Wilson JE, Pugh EW, *et al*. The Genometric Analysis Simulation Program (G.A.S.P.): A software tool for testing and investigating methods in statistical genetics. *American Journal of Human Genetics*. 59, A193 (1996).

38. Li W, Reich J. A complete enumeration and classification of two-locus disease models. *Hum Heredity*. 50(6), 334-349 (2000).

39. Moore JH, Hahn LW, Ritchie MD, Thornton TA, White BC. Application of genetic algorithms to the discovery of complex genetic models for simulation studies in human genetics. In: *Proceedings of the Genetic and Evolutionary Computation Conference.* Langdon WB *et al*. (Ed.), Morgan Kaufmann Publishers, San Francisco, USA, 1150-1155 (2002).

40. Moore JH, Parker JS, Hahn LW. Symbolic discriminant analysis for mining gene expression patterns. In: *Lecture Notes in Artificial Intelligence* (2167). De Raedt L, Flach P (Ed.), Springer-Verlag, Berlin, DE, 372-381 (2001).

41. Moore JH, Parker JS. Evolutionary computation in microarray data analysis. In: *Methods of Microarray Data Analysis.* Lin S, Johnson K (Ed.), Kluwer Academic Publishers, Boston, USA, 23-35 (2002).

42. Moore JH, Parker JS, Olsen NJ, Aune T. Symbolic discriminant analysis of microarray data in autoimmune disease. *Genetic Epidemiology*. 23, 57-69 (2002).

43. Reif DM, White BC, Olsen NJ, Aune TA, Moore JH. Complex function sets improve symbolic discriminant analysis of microarray data. In: *Lecture Notes in Computer Science* (2724). Cantu-Paz E *et al*. (Ed.), Springer-Verlag, Berlin, DE, 2277-2287 (2003).

44. Moore JH. Cross validation consistency for the assessment of genetic programming results in microarray studies. In: *Lecture Notes in Computer Science* (2611). Raidl, G *et al*. (Ed.), Springer-Verlag, Berlin, 99-106 (2003).

45. O'Neill M, Ryan C. Grammatical evolution. *IEEE Transactions on Evolutionary Computation*. 5, 349-358 (2001).

46. Koza JR *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* The MIT Press, Cambridge, USA (1992).

47. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Heredity*. 56, 73-82 (2003).

48. Bateson, W. *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge, UK (1909).

49. Moore JH, Ritchie MD. The Challenges of Whole-Genome Approaches to Common Diseases. *Journal of the American Medical Association*. 291(13), 1642-1643 (2004).

CHAPTER III

PROTEOMIC BIOMARKERS ASSOCIATED WITH ADVERSE EVENTS FOLLOWING SMALLPOX VACCINATION

The complication rate of smallpox vaccine is higher than any other vaccine currently in widespread use. The live vaccinia virus used is reactogenic in a significant number of vaccinées. While the most common adverse events (AEs) following inoculation are fever, lymphadenopathy, and rash, severe, life-threatening AEs including encephalitis and myopericarditis have been observed. Given that an unacceptably high rate of adverse reactions occurred in limited, pre-screened healthy populations, the complications resulting from a population-wide vaccination program are potentially disruptive on a vast economic and social scale. Studies are needed to elucidate the underlying immunological mechanisms contributing to the development of AEs. It is hypothesized that many systemic AEs, such as fever, lymphadenopathy, and generalized rash, share common etiologies involving the inflammatory response. These systemic AEs likely have a proteomic signature in the serum that involves the action of cytokines and chemokines. Therefore, to capture this signature, we used a protein microarray technique to measure circulating (serum) levels of 108 cytokines and chemokines in vaccinées before and one week after primary immunization with Aventis-Pasteur smallpox vaccine (APSV). Of the 74 individuals with measured proteomic data, 22 suffered a systemic adverse event

and 52 did not. We employed a committee of machine learning and statistical methods to identify proteomic biomarkers whose post-vaccination changes were associated with adverse events [1]. The committee identified a consensus subset of cytokines, which were used to train a final decision-tree model. Our final model included six cytokines: G-CSF (CSF-3), SCF, MIG (CXCL9), ICAM-1 (CD54), eotaxin, and TIMP-2. Changes in dynamic levels of these cytokines after vaccination accurately discriminated between AE status classes. The final model points to a cytokine signature associating adverse events with prolonged or hyper-activated inflammatory pathways. This proteomic signature also indicates a significant impact of cytokine secretion by fibroblasts in the development of adverse events following vaccination.

Introduction

Smallpox is a potentially lethal disease caused by the variola virus. In addition to its high mortality rate, smallpox is highly contagious, and its successful control through vaccination is one of the greatest triumphs of human medicine. Vaccination against smallpox involves inoculation with live vaccinia virus (VV) in the skin. In most healthy adults, vaccination induces a protective response. The protective response induced by VV may even lessen the severity of illness if given within four days after variola virus infection. Studies demonstrate that vaccinia-specific T lymphocytes secrete IFN-γ after immunization, and that these cells may be long-lived [2-4]. In a previous study, we investigated the effect of the Aventis Pasteur smallpox vaccine (APSV) on a limited panel of systemic cytokine concentrations in a cohort of previously vaccinia-naïve individuals [5]. Systemic cytokines representing lymphocyte functional subsets of Th1 cells (IFN-γ, TNF-α, and IL-2) and Th2 cells (IL-4, IL-5, and IL-10) were measured using a sensitive flow cytometric bead array assay that allowed multiple cytokine analyses from a single sample [6]. In the systemic compartment, smallpox immunization induces an IFN-γ-dominant response one week after immunization, with concentrations returning to baseline during convalescence. However, systemic IFN-γ concentration was not discriminatory between AE status groups.

To identify proteomic biomarkers responsible for systemic AEs following smallpox vaccination, we precisely quantitated 108 serum cytokines and

chemokines using rolling-circle amplification technology (RCAT) [7-13] just before (baseline), and one week after (acute phase), immunization with APSV. Of 74 individuals studied following primary vaccination, 22 suffered a systemic AE. We employed an unweighted voting strategy among a committee of machine learning methods and statistical procedures to limit the number of false discoveries while maintaining statistical power. We used support vector machines (SVMs), nearest shrunken centroids (NSCs), and a false discovery rate (FDR) corrected Wilcoxon rank-sum test to select the soluble factors most associated with AEs. We then used a decision tree to model the functional relationship between the selected cytokines and systemic AEs. In this analysis, we find systemic cytokine patterns characteristic of inflammation marked by the prominent induction of IL-17 and IFN-γ related cytokines, as well as patterns characteristic of tissue inflammation and moderate destruction.

Subjects, materials, and methods

Study subjects

Healthy adult subjects 18-32 years of age were enrolled in a multi-center study of primary immunization against smallpox using the APSV in the National Institutes of Health Vaccine and Treatment Evaluation Units. At the Vanderbilt University Medical Center site, 148 volunteers were enrolled in this NIH-sponsored APSV immunization trial (NIH-DMID Protocol 02-054). Vaccines, study subjects, and study design were previously described in detail [14]. All

subjects participating in the main smallpox immunization study at the Vanderbilt University Medical Center were invited to participate in the cytokine substudy. Serum samples for cytokine analysis were obtained following informed consent under approval from the Vanderbilt University Institutional Review Board from 107 of the 148 subjects vaccinated in this study at Vanderbilt. All 22 subjects suffering systemic AEs among the 107 who donated serum were included in this analysis, and 52 subjects who did not experience any AE were used as a control group.

Clinical assessments

Trained physicians and nurse providers examined the subjects by history and physical examination for indications of vaccine take (presence of a vesicle or pustule at the inoculation site) and AEs at five post-immunization visits in the first month (on days 3-5, 6-8, 9-11, 12-15, and 26-30). For the purposes of the current study, we considered the occurrence of three systemic AEs: generalized rash, fever, and lymphadenopathy. Fever was defined as an oral temperature > 38.3 ºC. A generalized rash was defined as skin eruptions in regions not contiguous with the site of vaccination. The frequent acneiform rashes seen in this trial have been described elsewhere [15]. Lymphadenopathy was defined as tenderness or enlargement of regional lymph nodes associated with vaccination.

Sample collection

Pre-vaccination serum samples (baseline) were collected during a screening visit immediately prior to vaccination, and post-vaccination samples were obtained 6-9 days after vaccination (acute phase). Serum samples were collected in 5 ml Vacutainer serum separator tubes (Becton Dickinson, San Jose, CA) and were centrifuged at 700 x g for 10 minutes. The serum then was collected, aliquoted into cryovials (Sarstedt Inc., Numbrecht, Germany) and stored at –80 ºC until assayed for cytokine concentrations. RCAT was used to measure 108 serum cytokines and chemokines for all 22 subjects who experienced a systemic AE and 52 subjects who did not experience an AE. Because we are studying cytokine expression in the serum compartment, we focus on systemic AEs, which we expected to be more strongly associated with serum cytokine expression than would a local AE.

Proteomic assay

The expression levels of 108 protein analytes were measured in 100 µL serum aliquots from the patient samples using custom dual antibody sandwich immunoassay arrays, as described in [7-13]. The list of analytes is shown in Table 1. Briefly, monoclonal capture antibodies specific for each analyte were fixed to glass slides, with 12 replicate spots for each analyte. Duplicate samples of sera were incubated for 2 hours, and then washed. Slides were then incubated with secondary biotinylated polyclonal antibodies, and signals were amplified using a 'rolling circle' method [10]. Quality control measures included

optimization of antibody pairs, the use of internal controls to minimize array-to-array variation, and standardized procedures of chip manufacturing [10]. Arrays were scanned using a Tecan LS200 unit and mean fluorescence intensities (MFIs) were generated with customized software. To ensure a dynamic working range for each assay, 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. Because of the broad individual range of systemic cytokine expression before and after immunization, changes in serum cytokine concentrations during the early post-immunization phase were calculated as the percent of the corresponding individual's baseline expression at the pre-vaccination visit.

Table 1. Gene names and symbols of 108 protein analytes measured in 100 µL serum aliquots from the patient samples using custom dual antibody sandwich immunoassay arrays. (Continued on following pages)

Gene Symbol	Gene Name
CSF3 (G-CSF)	Colony stimulating factor 3 (granulocyte)
IL-10	Interleukin 10
IFNG	Interferon, gamma
ALCAM	Activated leukocyte cell adhesion molecule
ANGPT4	Angiopoietin 4
BDNF	Brain-derived neurotrophic factor
CXCL13 (BLC)	Chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)
CCL28 (MEC)	Chemokine (C-C motif) ligand 28
TNFSF7 (CD27)	Tumor necrosis factor (ligand) superfamily, member 7
TNFSF8 (CD30)	Tumor necrosis factor (ligand) superfamily, member 8
CCL27 (CTACK)	Chemokine (C-C motif) ligand 27
TNFRSF21 (DR6)	Tumor necrosis factor receptor superfamily, member 21
EGF	Epidermal growth factor (beta-urogastrone)
CXCL5 (ENA78)	Chemokine (C-X-C motif) ligand 5
CCL11 (Eot)	Chemokine (C-C motif) ligand 11
CCL26 (Eot3)	Chemokine (C-C motif) ligand 26
CCL24 (Eot2)	Chemokine (C-C motif) ligand 24

FGF4	Fibroblast growth factor 4 (heparin secretory transforming protein 1)
FGF7	Fibroblast growth factor 7 (keratinocyte growth factor)
FGF9	Fibroblast growth factor 9 (glia-activating factor)
FGF2 (FGFB)	Fibroblast growth factor 2 (basic)
FGF1	Fibroblast growth factor 1 (acidic)
FAS	Fas (TNF receptor superfamily, member 6)
FASLG	Fas ligand (TNF superfamily, member 6)
FLT3LG	Fms-related tyrosine kinase 3 ligand
FST	Follistatin
CX3CL1	Chemokine (C-X3-C motif) ligand 1 (fractalkine, neurotactin)
CXCL6 (GCP2)	Chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)
GDNF	Glial cell derived neurotrophic factor
CSF2 (GMCSF)	Colony stimulating factor 2 (granulocyte-macrophage)
CXCL3 (GRO3)	Chemokine (C-X-C motif) ligand 3
CXCL2 (GRO2)	Chemokine (C-X-C motif) ligand 2
CCL14 (HCC1)	Chemokine (C-C motif) ligand 14
CCL16 (HCC4)	Chemokine (C-C motif) ligand 16
HGF	Hepatocyte growth factor (hepapoietin A; scatter factor)
TNFRSF14 (HVEM)	Tumor necrosis factor receptor superfamily 14 (herpesvirus entry mediator)
CCL1 (I309)	Chemokine (C-C motif) ligand 1
CXCL11 (ITAC)	Chemokine (C-X-C motif) ligand 11
ICAM1	Intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
ICAM3	Intercellular adhesion molecule 3
IGF2	Insulin-like growth factor 2 (somatomedin A)
IGF1R	Insulin-like growth factor 1 receptor
IGFBP1	Insulin-like growth factor binding protein 1
IGFBP3	Insulin-like growth factor binding protein 3
IGFBP4	Insulin-like growth factor binding protein 4
IGFBP2	Insulin-like growth factor binding protein 2, 36kDa
IL10RB	Interleukin 10 receptor, beta
IL-13	Interleukin 13
IL-15	Interleukin 15
IL-17	Interleukin 17 (cytotoxic T-lymphocyte-associated serine esterase 8)
IL-1A	Interleukin 1, alpha
IL-1B	Interleukin 1, beta
IL1-RN	Interleukin 1 receptor antagonist
IL1RL2	Interleukin 1 receptor-like 2
IL-2	Interleukin 2
IL2RB	Interleukin 2 receptor, beta
IL2RA	Interleukin 2 receptor, alpha
IL-3	Interleukin 3 (colony-stimulating factor, multiple)
IL-4	Interleukin 4
IL-5	Interleukin 5 (colony-stimulating factor, eosinophil)
IL-6	Interleukin 6 (interferon, beta 2)
IL-7	Interleukin 7
IL-8	Interleukin 8
IL2RG	Interleukin 2 receptor, gamma (severe combined immunodeficiency)
IL5RA	Interleukin 5 receptor, alpha
IL-9	Interleukin 9

SELL	Selectin L (lymphocyte adhesion molecule 1)
CSF1	Colony stimulating factor 1 (macrophage)
CSF1R	Colony stimulating factor 1 receptor
CCL2 (MCP1)	Chemokine (C-C motif) ligand 2
CCL8 (MCP2)	Chemokine (C-C motif) ligand 8
CCL7 (MCP3)	Chemokine (C-C motif) ligand 7
CCL13 (MCP4)	Chemokine (C-C motif) ligand 13
CXCL9 (MIG)	Chemokine (C-X-C motif) ligand 9 (monokine induced by gamma interferon)
CCL3 (MIP1A)	Chemokine (C-C motif) ligand 3
CCL4 (MIP1B)	Chemokine (C-C motif) ligand 4
CCL5 (MIP1D)	Chemokine (C-C motif) ligand 5
CCL20 (MIP3A)	Chemokine (C-C motif) ligand 20
CCL19 (MIP3B)	Chemokine (C-C motif) ligand 19
MMP7	Matrix metalloproteinase 7 (matrilysin, uterine)
MMP9	Matrix metalloproteinase 9 (gelatinase B, 92kDa gelatinase)
CCL23 (MPIF1)	Chemokine (C-C motif) ligand 23
NTF3	Neurotrophin 3
NTF5	Neurotrophin 5 (neurotrophin 4/5)
OSM	Oncostatin M
PARC	P53-associated parkin-like cytoplasmic protein
PDGFRA	Platelet-derived growth factor receptor, alpha polypeptide
PECAM1	Platelet/endothelial cell adhesion molecule (CD31 antigen)
PGF	Placental growth factor, vascular endothelial growth factor-related protein
TNFRSF11A	Tumor necrosis factor receptor superfamily, member 11a, activator of NFKB
CCL5 (RANTES)	Chemokine (C-C motif) ligand 5
KITLG (SCF)	KIT ligand
KIT (SCFR)	V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
CXCL12 (SDF1)	Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
IL1RL1	Interleukin 1 receptor-like 1
CCL17 (TARC)	Chemokine (C-C motif) ligand 17
TGFA	Transforming growth factor, alpha
TIMP-2	Tissue inhibitor of metalloproteinase 2
TIMP1	Tissue inhibitor of metalloproteinase 1
TNFRSF1A	Tumor necrosis factor receptor superfamily, member 1A
TNF	Tumor necrosis factor (TNF superfamily, member 2)
LTA	Lymphotoxin alpha (TNF superfamily, member 1)
TNFRSF10A	Tumor necrosis factor receptor superfamily, member 10a
TNFRSF10D	Tumor necrosis factor receptor superfamily, member 10d
VEGF	Vascular endothelial growth factor
KDR	Kinase insert domain receptor (a type III receptor tyrosine kinase)
BTG2	BTG family, member 2
NM	Neutrophil migration

Statistical analysis methods

Because the types of proteomic effects contributing to systemic AEs after vaccination have not been fully characterized, as well as the fact that changes in the cytokine concentrations measured followed non-standard distributions, we employed a committee of machine learning and statistical methods to identify AE-associated proteomic biomarkers. Since each method in the committee was chosen for its unique analytical perspective, agreement (consensus) between methods on the importance of particular variables indicates that the association of that variable with AEs is more than a single method-specific bias. Consensus cytokines were defined as those identified by at least two of the three committee methods. After using the committee to identify a consensus subset of cytokines whose post-vaccination changes were associated with adverse events, a final decision-tree model was built from these variables. Descriptions of each method in our committee are given below.

Modern high-throughput experimental techniques allow for the simultaneous testing of multitudes of statistical hypotheses. The issue of multiple-testing arises in such situations, with the probability of false-positive results in a raft of tests increasing with the total number of tests (N) performed. For a single statistical hypothesis test in the context of this study, the discrepancy in cytokine levels between the two AE groups is declared significant if the p-value is $< \alpha$. Traditionally, α is set to 0.05, meaning that the probability of making a Type I (false-positive) error is approximately α. A naïve solution to the multiple-testing problem is the Bonferroni correction, which chooses a significance level of

$\alpha^* = \alpha / N$. This method takes the number of hypotheses tested into account, but the significance level required to declare a positive association becomes prohibitively stringent as N grows large.

An alternative statistical procedure that controls the number of type I errors while providing reasonable power when performing multiple hypothesis tests is the false-discovery rate (FDR) method [16]. The FDR procedure returns a significance threshold linked to the distribution of p-values generated by a statistical test, controlling the average fraction of false discoveries made among the multiple hypothesis tests whose null hypotheses were rejected. The q-value measures the proportion of false-positive occurrences (*i.e.* the false-discovery rate) when a particular test is declared significant. We used the non-parametric Wilcoxon rank-sum test to compare means between systemic AE and non-AE groups. Using a false discovery rate $q = 0.3$, we found the significance threshold to be 0.02. The unweighted voting procedure involving SVM and NSC (described below) was used to further filter out spurious associations. Unless otherwise stated, methods were implemented in the MATLAB programming language Version 7.1 (release 13).

The other two methods in our committee were NSC and SVM, both of which are supervised machine learning methods, meaning that the outcome classes are known to the method. In this study, we considered two classes of individuals: those experiencing systemic adverse events (AE) versus those without a reported adverse event (non-AE), although the particular AE subsets

could be treated as multiple classes in multi-class implementations of these algorithms.

For binary classification tasks, a Support Vector Machine (SVM) finds a hyper-plane that maximally separates training data from the two classes. The optimal hyper-plane maximizes the separation (margin) between individuals from each class. Individuals (each representing a vector of measured proteomic variables) closest to the hyper-plane are referred to as support vectors. SVMs create non-linear separations by using a kernel technique to automatically realize a non-linear mapping to a feature space. The hyper-plane found by the SVM in feature space corresponds to a non-linear decision boundary in input space [17]. We implemented SVM using the GEMS (Gene Expression Model Selector Version 2.0.2) [18] analysis software. Parameters included a radial basis function kernel, Markov blanket feature selection [19], and ten-fold cross-validation (CV). Prediction accuracy was calculated on each test set created during the ten stage CV procedure, achieving an average prediction accuracy of 69%. SVM found seven cytokines predictive of AE status, five of which passed our inclusion criterion of consensus with the other committee methods.

Nearest Shrunken Centroids (NSC) [20] was the final statistical learning method used in our committee strategy to identify consensus proteomic biomarkers. NSC is appropriate for such a task because of its ability to perform automatic feature selection. For each cytokine *i*, the *k* components of the class centroids $\overline{x}_{ik}$ (the mean change of cytokine *i* for individuals in class *k*) are shrunk toward the overall centroid $\overline{x}_i$ (the mean change of cytokine *i* across all

individuals). Here, *k* = 2, corresponding to the two AE classes. The centroids are shrunk by a t-statistic-like quantity d'_{ik}, which is a measure of the ability of cytokine *i* to distinguish the class-*k* centroid from the overall centroid. If d'_{ik} is zero, then the cytokine-*i* component of the class-*k* centroid is equal to the component of the overall centroid, and this cytokine does not contribute to classification for class-*k*. We used the same discriminant score as in Tibshirani *et al*. [20]. Ten-fold CV was used to tune the regularization term s_0 as well as the shrinkage Δ. We found an average prediction accuracy of 70% with $\Delta = 2.1$ and $s_0 = 0.002$. NSC required less computational time than SVMs. Four out of five cytokines selected by NSC overlapped with those selected by SVM.

The final step of our analysis strategy was to create an interpretable model from the cytokines found by consensus among the three feature-selection methods (FDR-Wilcoxon, SVM, and NSC). Decision trees were chosen to build the final AE model because of their ready interpretability and explicit modeling of variable interactions. We used the implementation of the C4.5 decision-tree algorithm provided in the Weka machine learning software package [21] to obtain the model in Fig. 1 (see Results section). Individuals are classified into AE or non-AE groups by sorting down a dichotomous tree toward terminal leaves. Starting from the root, the tree splits at a cytokine according to how well the relative change of a given cytokine separates individuals into the appropriate classes. The relative change threshold is calculated by choosing among a set of possible values for each particular split. Using information gain to rank cytokines, we place cytokines at tree nodes with the greatest gain among

attributes not yet considered in the path from the root node. We used a 25% confidence value for pruning branches that do not improve training accuracy—finding the CV accuracy to be insensitive to changes in this value. We optimized the minimum number of instances that must be present from each AE class in the training data for a new leaf to be created to handle those instances. For these data, a minimum of 5 instances resulted in a more parsimonious tree that more readily generalizes to test sets.

Results

This study aimed to shed light on the proteomic mechanisms underlying the high rate of AEs reported in subjects receiving smallpox vaccine [15]. We measured serum concentrations of cytokines in vaccinia-naïve adults at two time points: pre-vaccination (baseline) and one week post-vaccination (acute phase). Since AEs of a systemic nature, such as fever, generalized rash, and lymphadenopathy are likely related to circulating immune mediators, this study sought to determine whether systemic alterations of serum cytokines are associated with these AEs. In our clinical study, systemic AEs were reported in 22 subjects. There were no subjects reporting serious AEs—defined by the need for clinic or emergency visits or for hospitalization related to vaccination. Subjects without a reported AE (n = 52) exhibited significantly different serum cytokine signatures than subjects with a reported systemic AE (n = 22).

The committee strategy identified six consensus cytokines that accurately discriminated the two AE status classes: stem cell factor (SCF), monokine induced by interferon-γ (MIG), tissue inhibitor of metalloproteinases-2, granulocyte colony stimulating factor (G-CSF or CSF-3), intercellular adhesion molecule-1 (ICAM-1 or CD54), and eotaxin. The feature selection results from the committee are summarized in Table 2, with cytokines ordered by Wilcoxon p-values.

The committee consensus strategy aims to reduce spurious associations due to method-specific biases. Our results indicate that such a strategy may be beneficial, even when using corrective procedures such as the FDR. While a q-value of 0.3 can be interpreted to mean that up to 30% of the consensus cytokines are false discoveries, it does not mean that the false discoveries are the 30% with the highest p-values. The unweighted committee voting method eliminated two of the seven cytokines selected by the FDR-Wilcoxon procedure.

Table 2. Cytokines listed in the first column were found to discriminate between AE and non-AE individuals by at least one of the three statistical methods: false discovery rate (FDR) correction to the Wilcoxon rank-sum test, nearest shrunken centroids (NSC), and support vector machines (SVM). For each cytokine row listed, an X in an FDR, NSC, or SVM column indicates that this cytokine was selected by the corresponding method. Highlighted in bold are cytokines identified by consensus of at least two of the three statistical methods. These highlighted cytokines were used to train a final decision-tree model (see Figure 1). The cytokines are ordered by Wilcoxon rank-sum p-value, listed in the fifth column. The last two columns show the mean relative percent change from baseline for the AE and non-AE groups, respectively.

Gene Symbol	FDR	NSC	SVM	Wilcoxon p-value	Change (AE)	Change (non-AE)
ICAM-1	**X**	**X**		**0.0013**	**37.2%**	**17.8%**
G-CSF (CSF3)	**X**		**X**	**0.0029**	**994.5%**	**48.3%**
TIMP-2	**X**	**X**	**X**	**0.0054**	**27.5%**	**10.0%**
IL-10	X			0.0124	378.0%	-2.9%
MIG (CXCL9)	**X**	**X**	**X**	**0.0128**	**53.2%**	**19.4%**
ALCAM	X			0.0151	21.2%	10.1%
SCF	**X**	**X**	**X**	**0.0166**	**19.4%**	**-12.7%**
MPIF1 (CCL23)		X		0.0274	75.8%	36.2%
Eotaxin		**X**	**X**	**0.0463**	**4.0%**	**-6.6%**
IL-4			X	0.0476	21.8%	7.8%
IL-8		X		0.0577	12.4%	-3.7%
NTF3		X		0.0990	17.8%	-1.3%

To obtain a descriptive, interpretive model of the functional relationship between the set of cytokines selected by our committee method and systemic AEs, the final decision-tree in Figure 1 was trained on the full data. This model correctly classifies 92% of individuals in the data. Using ten-fold CV and specifying a minimum of five individuals for creation of new branches, we estimated the prediction accuracy of the final decision-tree model to be 77%. The final model includes the cytokines ICAM-1, G-CSF, eotaxin, and TIMP-2.

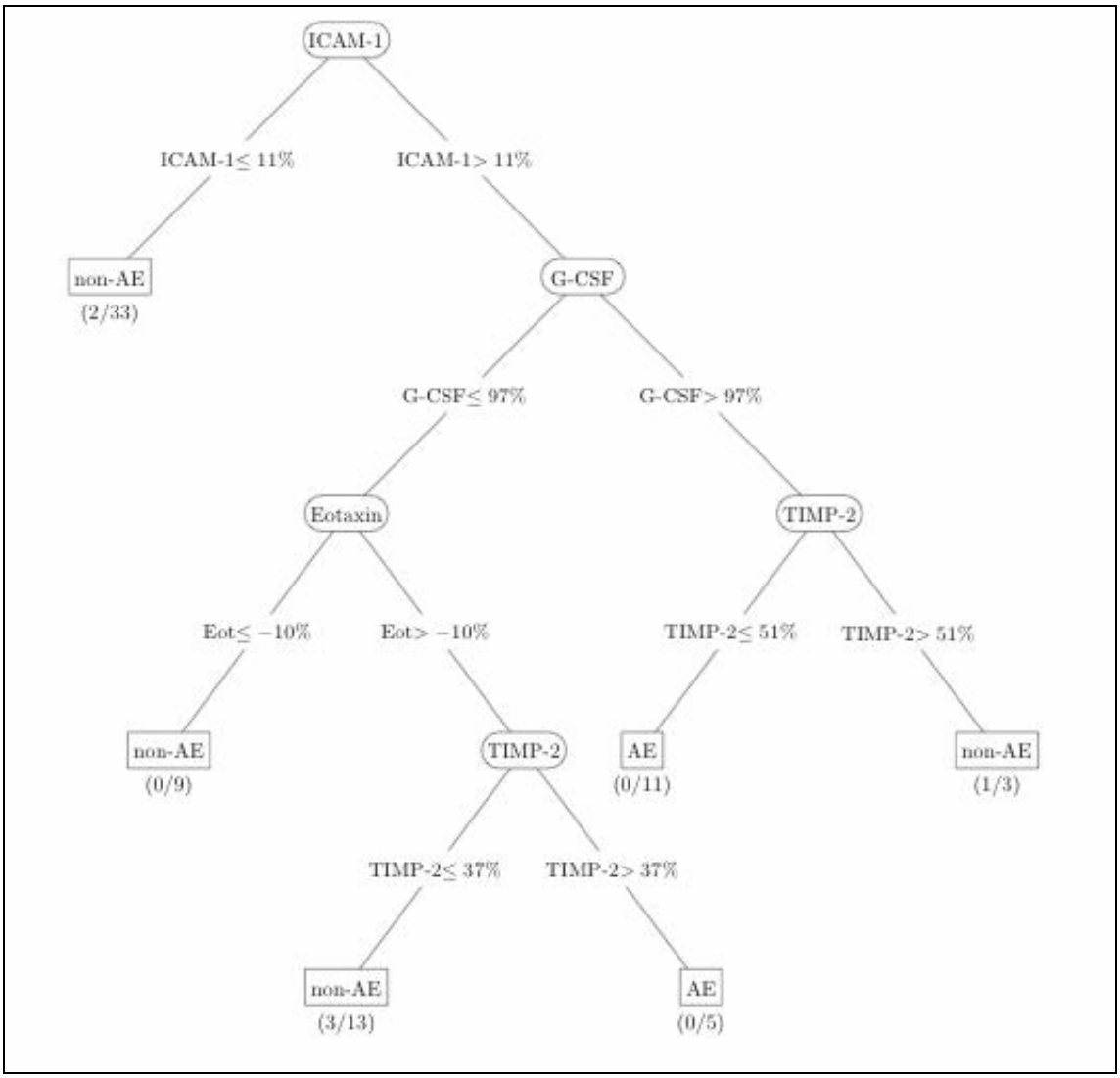


Figure 1. Final pruned decision-tree model for predicting AE status from cytokine expression changes after vaccination. Cytokines identified by the unweighted voting filter (SCF, MIG, TIMP-2, G-CSF, ICAM-1, and eotaxin) were selected to train the decision-tree classifier. Input (ovals) for the if-then rules is the percentage change of the subject's cytokine level during the acute phase relative to the baseline cytokine level. Based on the value of the input, the inequalities guide the decision of which branch to follow. Given an individual's cytokine profile, one follows the decision branches from the root (ICAM-1) downward to one of the six terminal nodes (AE or non-AE boxes). When one of the following decision branches is reached, an individual is predicted to be classified as AE or non-AE depending on which inequality is satisfied: ICAM-1 ≤ 11% (non-AE), Eot ≤ -10% (non-AE), TIMP-2 ≤ 51% (AE), TIMP-2 > 51% (non-AE), TIMP-2 ≤ 37% (non-AE), TIMP-2 > 37% (AE). The misclassification rates are given in parentheses below each terminal node.

Discussion

Immune responses involve an intricate network of both local and systemic signaling proteins. These cytokine and chemokine signals direct both the action and localization of immunological effectors. Dysfunction within any of these communication networks—whether upregulation of activating signals or lack of proper inhibitory signals—can tip the balance of a normally appropriate response towards one in which the immune effectors actually contribute to illness. Adverse events in response to smallpox vaccination may represent such a situation. Since subjects in the present study were successfully vaccinated, it is thought that the development of AEs represents excessive and/or protracted activity of appropriate immune responses.

Since comprehensive data concerning serum cytokine concentrations following smallpox vaccination have not been gathered previously, there is a major knowledge gap in our understanding of the systemic mechanisms contributing to AEs associated with vaccination. Filling this knowledge gap will provide important details regarding the pathophysiology of AEs and the successful control of poxvirus infections. The present study contributes to this learning process by identifying patterns of serum cytokine expression changes associated with systemic AEs after vaccination.

Previous studies have shown that nearly all subjects with vesicle formation exhibit strong VV-specific cytotoxic T lymphocyte (CTL) responses and increased counts of IFN-γ-producing T cells following vaccination with APSV [4]. These

findings suggested that vigorous T-cell and humoral responses are induced if a vesicle forms, independent of vaccine dose. The clinically observable lesions at the inoculation site in subjects receiving APSV suggests the possibility that biologically significant cytokine production occurred locally in our subjects without presenting dramatic increases in the systemic compartment. Since our aim was to identify serum cytokine patterns predictive of systemic AEs, relevant local cytokine dynamics could have been missed. However, serum cytokine expression is more readily and reproducibly measured for rapid clinical diagnostic purposes.

In the present study, we precisely measured systemic concentrations of 108 cytokines and chemokines in serum samples obtained prior to vaccination, and one week post-vaccination, using a sensitive protein microarray technique incorporating RCAT [7-13, 22]. To extract a useful subset of cytokines that discriminates between subjects who suffered at least one systemic AE (fever, lymphadenopathy, or generalized rash) from those who did not experience an AE, we employed three different class comparison methods: FDR-Wilcoxon, SVM, and NSC. A cytokine was selected for building the final decision-tree model if it was identified by at least two of these three methods. Decision trees were used to derive a descriptive, interpretable model of the functional relationships between the six selected cytokines and AE status. It should be noted that these serum cytokine/chemokine expression levels were measured early in the period following vaccination, well before most AEs had occurred. Therefore, the model could be considered *predictive* of subsequent AEs.

Profiling at early time points following immunization may be useful in predicting AE risk and directing at-risk subjects to properly recognize vaccine-related symptoms.

Considering the cytokines selected by our consensus strategy, three are in the pro-inflammatory interleukin-17 (IL-17) signaling pathway. In this pathway, fibroblasts, stimulated by IL-17, are induced to secrete inflammatory and hematopoietic cytokines, including G-CSF, SCF (both identified in our committee method), and IL-8 (also known as CXCL8; identified by the NSC method). These cytokines incite a range of activities that include neutrophil proliferation and differentiation. IL-17 has been shown to enhance cell surface expression of the endothelial cell adhesion molecule ICAM-1 on human fibroblasts [23]. In turn, increased expression of ICAM-1 was shown to aid in T-cell recruitment during contact hypersensitivity (related to delayed-type hypersensitivity) [24]. In the present study, soluble ICAM-1 was a strong discriminator of smallpox vaccine-related AE status. In fact, ICAM-1 was the root node of the decision-tree model in Figure 1, meaning that the effect of other cytokines in our model on AE status was conditionally dependent on changes in serum concentration of ICAM-1. While key cytokines in the IL-17 pathway play an important role in our analysis, changes in circulating levels of IL-17 itself were not found to be differentially expressed between AE status groups. Had different time points been chosen for cytokines measurement, IL-17 may have been selected as important. While the one-week post-immunization time point captures the peak concentration for most cytokines, it may not be representative of all AE-relevant cytokines.

Another cytokine selected was Eotaxin, which is a chemokine ligand for CCR3 (also known as CD193) that activates and recruits eosinophils to the site of inflammation and stimulates macrophage activation. Activated eosinophils can release reactive oxygen species that contribute to host tissue damage during chronic inflammatory responses.

The committee also selected monokine induced by IFN-γ (MIG or CXCL9), a member of the C-X-C subfamily of chemokines and an attractor of activated T cells expressing chemokine receptor CXCR3 [25]. The IFN-γ-induced MIG is produced by macrophages and may play a crucial role in enhancing the recruitment and activation of T cells [26].

The dual function of TIMP-2 in the final model highlights an important property of decision trees in allowing a flexible modeling framework. TIMP-2 was found to be associated with AEs by all three statistical learning methods, and its placement in the decision tree (Fig. 1) points to a complex role in AE development. TIMP-2 appears in two branches of the decision tree, and its effect on the prediction of AE status depends on the context of other cytokines in the tree. Proteins in the TIMP family inhibit the matrix metalloproteinases (MMP), a group of peptidases involved in degradation of extracellular matrix (ECM). Normally, TIMP-2 accelerates wound healing by enhancing the proliferation of epidermal keratinocytes and dermal fibroblasts. Following the branches toward TIMP-2 on the right, when an individual's increase in TIMP-2 expression is less than 51%, then that individual experiences an AE. This is presumably because the balance between the MMP and its inhibitor tips toward the MMP—promoting

excess ECM destruction and further inflammation. However, this situation only occurs when G-CSF expression is substantially increased, relative to baseline. When the increase in G-CSF is less than 97% and the expression of eotaxin does not decrease by more than 10%, then the role of TIMP-2 is qualitatively different. Hence, accurate predictions of AE status based on the expression of TIMP-2 must be taken in the context of other cytokines. This finding demonstrates a salient challenge in complex molecular investigations of clinical populations: statistical interactions between variables must be taken into account when testing association with a phenotype. Although it is possible that the right-hand branch of the decision tree is the result of over-fitting, this type of complex TIMP-2 behavior has been observed previously—where the physiological concentration and proteomic context affects whether TIMP-2 has an inhibitory versus activating effect on MMPs [27].

Application of our consensus analysis strategy to protein microarray data indicates a cytokine signature for the pathogenesis of AEs involving stem cell factor (SCF), monokine induced by IFN-γ (MIG), tissue inhibitor of metalloproteinases 2 (TIMP-2), granulocyte colony stimulating factor (G-CSF or CSF-3), intracellular adhesion molecule 1 (ICAM-1), and eotaxin. This signature suggests that the development of AEs involves excess stimulation of inflammatory pathways and the imbalance of tissue damage repair pathways.

Our model of adverse event development following smallpox vaccination involves interactions among soluble cytokines whose excess local secretion leads to remote diffusion and subsequent detection in circulation. It is

hypothesized that the initial local tissue injury in subjects suffering AEs after vaccination triggers an acute inflammatory response not unlike a delayed-type hypersensitivity (DTH) reaction. During the elicitation phase of DTH, antigen presentation to Th1 cells in the dermis leads to the release of T-cell cytokines such as IFN-γ and IL-17 [28, 29]. A cascade of cytokines and chemokines is then released—enhancing the inflammatory response by inducing the migration of monocytes into the lesion and their maturation into macrophages. This signal cascade attracts additional T cells as well. The dominant cytokine responses in the systemic compartment were characteristic of robust macrophage recruitment and activation. Taken together, the prevalence of inflammatory cytokines in AE development, coupled with previous work demonstrating the importance of T-cell derived factors and the similarities of systemic AEs recorded after smallpox vaccination with the clinical presentation of macrophage activation syndrome (MAS) [30], suggests that systemic AEs following smallpox vaccination may be consistent with low-grade MAS caused by virus replication and hyperactive tissue injury and repair mechanisms.

Acknowledgments

We would like to thank Jennifer Hicks, Karen Adkins (Vanderbilt Pediatric Clinical Research Office) and the Vanderbilt General Clinical Research Center staff for nursing support, and Molecular Staging Inc. for providing RCAT data. This work was supported by the NIH/NIAID Vaccine Trials and Evaluation Unit contract number NO1-AI-25462, by NIH grants K25-AI-64625, AI-59694 and RR-018787, and by infrastructure from the Vanderbilt NIH General Clinical Research Center (RR-00095). Generous support was also provided by the Vanderbilt Program in Biomathematics.

References

1. McKinney BA, Reif DM, Rock MT, Edwards KM, Kingsmore SF, Moore JH, Crowe JE, Jr. Cytokine expression patterns associated with systemic adverse events following smallpox immunization. J Infect Dis 2006;194:444-53

2. Ennis FA, Cruz J, Demkowicz WE, Jr., Rothman AL and McClain DJ. Primary induction of human CD8+ cytotoxic T lymphocytes and interferon-gamma-producing T cells after smallpox vaccination. J Infect Dis 2002;185:1657-9

3. Demkowicz WE, Jr., Littaua RA, Wang J and Ennis FA. Human cytotoxic T-cell memory: long-lived responses to vaccinia virus. J Virol 1996;70:2627-31

4. Rock MT, Yoder SM, Wright PF, Talbot TR, Edwards KM and Crowe JE, Jr. Differential regulation of granzyme and perforin in effector and memory T cells following smallpox immunization. J Immunol 2005;174:3757-64

5. Rock MT, Yoder SM, Talbot TR, Edwards KM and Crowe JE, Jr. Adverse events after smallpox immunizations are associated with alterations in systemic cytokine levels. J Infect Dis 2004;189:1401-10

6. Chen R, Lowe L, Wilson JD, et al. Simultaneous Quantification of Six Human Cytokines in a Single Sample Using Microparticle-based Flow Cytometric Technology. Clin Chem 1999;45:1693-1694

7. Schweitzer B RS, Grimwade B, Shao W, Wang M, Fu Q, Shu Q, Laroche I, Zhou Z, Tchernev VT, Christiansen J, Velleca M, Kingsmore SF. Multiplexed protein profiling on microarrays by rolling-circle amplification. Nature Biotechnology 2002;20:359-365

8. Mor G, Visintin I, Lai Y, et al. Serum protein markers for early detection of ovarian cancer. Proc Natl Acad Sci U S A 2005;102:7677-82

9. Kader HA, Tchernev VT, Satyaraj E, 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;100:414-23

10. Perlee L, Christiansen J, Dondero R, et al. Development and standardization of multiplexed antibody microarrays for use in quantitative proteomics. Proteome Sci 2004;2:9

11. Yang D, Chen Q, Rosenberg HF, et al. Human ribonuclease A superfamily members, eosinophil-derived neurotoxin and pancreatic ribonuclease, induce dendritic cell maturation and activation. J Immunol 2004;173:6134-42

12. Kaukola T, Satyaraj E, Patel DD, et al. Cerebral palsy is characterized by protein mediators in cord serum. Ann Neurol 2004;55:186-94

13. Schweitzer B, Wiltshire S, Lambert J, et al. Inaugural article: immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. Proc Natl Acad Sci U S A 2000;97:10113-9

14. Talbot TR, Stapleton JT, Brady RC, et al. Vaccination success rate and reaction profile with diluted and undiluted smallpox vaccine: a randomized controlled trial. Jama 2004;292:1205-12

15. Talbot TR, Bredenberg HK, Smith M, LaFleur BJ, Boyd A and Edwards KM. Focal and generalized folliculitis following smallpox vaccination among vaccinia-naive recipients. Jama 2003;289:3290-4

16. Benjamini YaH, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 1995;57:289-300

17. Vapnik V. Statistical Learning Theory. New York: Wiley, 1998

18. Statnikov A, Aliferis CF and Tsamardinos I. Methods for multi-category cancer diagnosis from gene expression data: a comprehensive evaluation to inform decision support system development. Medinfo 2004;11:813-7

19. Aliferis CF, Tsamardinos I and Statnikov A. HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA Annu Symp Proc 2003:21-5

20. Tibshirani R, Hastie T, Narasimhan B and Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 2002;99:6567-72

21. Witten IHaF, E. Decision Trees in Data Mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann, 2005

22. Kingsmore SF, Patel DD. Multiplexed protein profiling on antibody-based microarrays by rolling circle amplification. Curr Opin Biotechnol 2003;14:74-81

23. Z. Yao SLP, W.C. Fanslow, D. Ulrich, B.M. Macduff, M.K. Spriggs, R.J. Armitage. Human Il-17: A novel cytokine derived from T-cells. Journal of Immunology 1995;155:5483-

24. McHale JF, Harari OA, Marshall D, Haskard DO. Vascular Endothelial Cell Expression of ICAM-1 and VCAM-1 at the Onset of Eliciting Contact Hypersensitivity in Mice: Evidence for a Dominant Role of TNF-alpha. Journal of Immunology 1999;162:1648 - 1655

25. Weng Y SS, Waldburger KE, et al. Binding and functional properties of recombinant and endogenous CXCR3 chemokine receptors. Journal of Biological Chemistry 1998;273:18288-18291

26. Loetscher M GB, Loetscher P, et al. Chemokine receptor specific for IP10 and mig: structure, function, and expression in activated T-lymphocytes. Journal of Experimental Medicine 1996;184:963-9

27. Lu KV, Jong KA, Rajasekaran AK, Cloughesy TF and Mischel PS. Upregulation of tissue inhibitor of metalloproteinases (TIMP)-2 promotes matrix metalloproteinase (MMP)-2 activation and cell invasion in a human glioblastoma cell line. Lab Invest 2004;84:8-20

28. Fong TA, Mosmann TR. The role of IFN-gamma in delayed-type hypersensitivity mediated by Th1 clones. J Immunol 1989;143:2887-93

29. Susumu Nakae YK, Aya Nambu, Katsuko Sudo, Michiko Iwase, Ikuo Homma, Kenji Sekikawa, Masahide Asano, Yoichiro Iwakura. Antige-specific T cell sensitization is impaired in IL-17-deficient mice, causing suppression of allergic cellular and humoral responses. Immunity 2002;17:375-387

30. Grom AA, Passo M. Macrophage activation syndrome in systemic juvenile rheumatoid arthritis. J Pediatr 1996;129:630-2

CHAPTER IV

GENETIC POLYMORPHISMS ASSOCIATED WITH ADVERSE EVEENTS FOLLOWING SMALLPOX VACCINATION

Immunization with vaccinia virus is highly effective against smallpox, but adverse reaction to vaccination is an unfortunately common occurrence. Population-wide vaccination programs could put many people at risk, given the high reactogenicity of certain vaccines. Identifying stable genetic factors associated with adverse events (AEs) may allow more effective pre-vaccine screening and even direct vaccine development. To investigate the relationship between adverse events following smallpox vaccination and genetic factors, a panel of candidate single-nucleotide polymorphisms (SNPs) was genotyped in two independent study samples. Systemic adverse events, such as fever, prolonged rash, and lymphadenopathy were recorded for all vaccinées. After identifying candidate genetic factors in the first study sample, the statistically significant findings were validated in the second, independent study. We identified multiple AE-associated SNPs in three candidate genes: Interleukin 4 (IL-4), Interferon Regulatory Factor 1 (IRF-1), and Methylenetetrahydrofolate reductase (MTHFR). The odds ratios associating each of these polymorphisms with AEs were consistent across both the original and validation studies. The demographics of both study samples were statistically similar, and the allele frequencies for each significantly associated SNP were comparable between

samples. Confidence in these results is augmented by the fact that they have been validated in an independent study sample. Since all of the patients under study were successfully vaccinated, the AE outcomes reported represent immune reactions either beyond the necessary magnitude or sustained longer than necessary. The candidate genes validated in both studies include a major anti-inflammatory cytokine (IL-4), an immunological transcription factor (IRF-1), and a metabolism gene previously associated with adverse reactions to a variety of pharmacologic agents (MTHFR). Since the outcome of interest is the aggregation of specific symptoms, it is logical that more than one gene may be involved. These genes are all involved in processes that are consistent with previously hypothesized mechanisms for the development of AEs involving prolonged stimulation of inflammatory pathways and imbalance of normal tissue damage repair pathways. While the immune modulatory roles of IL-4 and IRF-1 have been thoroughly studied in a variety of contexts, the association of polymorphisms in these genes with systemic AEs is exciting. The non-synonymous polymorphism in MTHFR is a novel result with promising clinical significance.

Introduction

Adverse reaction following vaccination with vaccinia virus, the live attenuated vaccine for smallpox, is a common occurrence that can have significant health effects. Amid recent geopolitical concerns, there has been renewed interest in protection against infectious agents such as smallpox, which is considered a potential agent of bioterrorism. Population-wide vaccination programs may put many people at risk, given the high reactogenicity of conventional smallpox vaccines [1]. Given that the biological mechanisms underlying such adverse events (AEs) are not well-understood, there is a need to elucidate the molecular and cellular pathways and highlight pharmacological targets for intervention.

Vaccination of healthy adults with vaccinia virus induces a protective response in the majority of individuals who are immunized—indicated by a significant rise in vaccinia virus–neutralizing antibodies in the serum and clinically observable features such as the development and expansion of a papule at the site of vaccination. The papule generally develops into an inflammatory vesicular lesion termed a “pock”, and resolves into a scar, which is a marker of “vaccine take” that correlates with protection. Since all of the patients under study were successfully vaccinated, the AE outcomes reported represent immune reactions either beyond the necessary magnitude or sustained longer than necessary. Poxviruses have evolved multiple mechanisms to evade host immune responses, such as targeting of primary innate immunity mediators (including interferons,

interleukins, chemokines, and others) and manipulating intracellular signal transduction pathways [2]. These mechanisms also may contribute to AE development by creating a state of altered innate or adaptive immune stimulation.

Previously, we have investigated smallpox vaccine with respect to its effects on the humoral and cellular immune response, reactogenicity, and patterns of systemic cytokine expression [3-8]. In the current report, we utilize data collected for two independent studies to identify stable genetic factors associated with adverse events (AEs) in hopes that this may contribute to more effective screening and help direct vaccine development. To investigate the relationship between adverse events following smallpox vaccination and genetics, a panel of candidate single-nucleotide polymorphisms (SNPs) was genotyped in two independent clinical studies of the same vaccine, in which a significant proportion of vaccinees suffered systemic AEs—including fever, lymphadenopathy, and prolonged acneiform rash. For both studies, the data are genotypes at 1536 SNPs across roughly 500 candidate genes. The second study was used to validate the most promising results from the initial study. Most genetic association studies fail to replicate [9,10], thus, independent validation can be seen as the “gold standard” for genetic association studies. Results that successfully replicate are excellent candidates for in-depth follow-up via functional studies and deep resequencing in candidate genomic regions.

Subjects, materials, and methods

Study subjects

Vaccines, study subjects, and study design for the original study have been described in detail [4]. Briefly, the original study enrolled 148 (116 with recorded AE information) healthy adults at the Vanderbilt University Medical Center as part of a multi-center study of primary immunization against smallpox using the Aventis-Pasteur Smallpox Vaccine (APSV) at the Vanderbilt National Institutes of Health (NIH) Vaccine and Treatment Evaluation Units [7]. NIH-DMID Protocol 02-054 was implemented.

The study subjects for the validation sample also were collected at Vanderbilt University Medical Center, as a part of NIH-DMID Protocol 03-044, also using the APSV at the Vanderbilt NIH Vaccine and Treatment Evaluation Units [7]. A total of 102 healthy adults (all with recorded AE information) were enrolled as part of the validation study.

In both studies, individuals were asked to self-identify race, with White (96%), Black (2%), and Asian (2%) as the most common categories reported. Both studies complied with the internal review board policies of Vanderbilt University and the NIH, and written consent was obtained for all individuals participating.

Clinical assessments

For both studies, the same team of trained physicians and nurse providers examined the medical history and clinical symptoms of the subjects to insure consistent clinical assessment across studies. Subjects were examined on 5 visits within the first month post-vaccination and were assessed for occurrence of an adverse event. For all subjects, the first visit occurred during days 3-5 post-vaccination, the second during days 6-8, the third during days 9-11, the fourth during days 12-15, and the final during days 26-30. Systemic adverse events were considered for both studies, including fever, generalized rash, and lymphadenopathy. More specifically, fever was defined as an oral temperature of greater than 38.3° 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 can be found in Talbot *et al*. [11]. Finally, lymphadenopathy was defined as enlargement or tenderness of regional lymph nodes due to vaccination.

Identification of genetic polymorphisms

The custom SNP panel used in this study targets investigation of soluble factor mediators and signaling pathways, many of which have known immunological significance. Genotyping for single nucleotide polymorphisms (SNPs) was performed using DNA amplified directly from EBV-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) in Gaithersburg, Maryland (http://cgf.nci.nih.gov/home.cfm). Genotypes were generated using the Illumina™ GoldenGate assay technology. Of the 1536 SNPs assayed, a total of 1442 genotypes passed quality control filters (genotyping efficiency > 80%) for both the original and validation samples. The list of all 1442 SNPs is given in Table 1.

Table 1. List of all 1442 SNPs analyzed, organized according to location. SNP names are taken from http://snp500cancer.nci.nih.gov.

SNP Name	dbSNP ID (rs#)	SNP Region	SNP Location (Base Pair)
RXRA-03	rs1805352	IVS1-46A>C	15414
RXRA-01	rs1536475	IVS6+70A>G	36621
APOB-01	rs1042034	Ex29+926G>A	41215
CCR2-01	rs1799864	Ex2+241G>A	46295
CCR2-02	rs1799865	Ex2+831C>T	46885
CCR2-06	rs3138042	IVS2+118A>G	48119
CCR5-02	rs2734648	IVS1+151G>T	58934
CCR5-04	rs1799987	IVS1+246A>G	59029
CCR5-07	rs1800024	IVS2+80C>T	59653
APOB-21	rs3791981	IVS18+336T>C	61301
APOB-08	rs1469513	IVS6+410G>A	75496
APOB-04	rs1367117	Ex4+56C>T	79834
APOB-07	rs1800481	-392C>T	83144
ZFPM1-07	rs904797	IVS1+9545A>G	90248
CCND1-02	rs603965	Ex4-1G>A	323109
CCND1-03	rs7177	Ex5+230C>A	326314
CCND1-01	rs678653	Ex5+852C>G	326936
OPRD1-03	rs760589	IVS1-23001G>A	355421
IL15RA-02	rs2296135	Ex8-361A>C	357590
IL15RA-05	rs2296141	IVS6-242A>G	361555
IL15RA-04	rs2228059	Ex5-39A>C	365264
IL15RA-06	rs3136614	IVS4+32C>T	368570
OPRD1-05	rs204076	594bp 3' of STP T>A	383346
IL10-06	rs3024496	Ex5+210C>T	404971
IL10-05	rs3024509	IVS3-58C>T	406404
IL10-13	rs3021094	IVS1-192A>C	408059
IL10-07	rs3024491	IVS1-286G>T	408153
IL10-01	rs1800871	-7334T>C	409741
IL10-03	rs1800896	-1116A>G	410004
IL10-17	rs1800890	-3584T>A	412472

TFRC-01	rs3817672	Ex4-11G>A	420545
AHRR-10	E1518_63	2912bp 3' of STP G>C	427862
AHRR-02	rs10078	3152bp 3' of STP T>G	428102
SFTPD-03	rs2243639	Ex5-13A>G	450238
SFTPD-01	rs721917	Ex2+95T>C	454840
BCL2L1-03	rs1994251	IVS2+22130A>C	483420
CYBB-12	rs6610650	-2820A>G	486282
CYBB-09	rs4422908	IVS2+90A>C	491298
CYP2E1-31	rs8192766	-1514G>T	494964
BCL2L1-02	rs1484994	IVS2+3483C>T	502067
BCL2L1-01	rs3181073	IVS2+2259G>T	503291
CYP2E1-02	rs2070676	IVS7-118G>C	506716
CYBB-11	rs5964125	IVS7+118A>G	508223
CYBB-27	rs5964149	IVS12-350A>G	519466
CYBB-28	rs5964151	Ex13+686G>T	520501
RAD54L-04	rs1048771	Ex18+157C>T	563292
DRD4-15	rs4987059	-870A>G	576433
DRD4-07	rs916457	-290C>T	577014
UGT1A1-24	rs1042640	Ex5-402G>C	614298
OCA2-23	rs1900758	IVS13+113A>G	633086
OCA2-07	rs1800407	Ex13+17G>A	633307
OCA2-03	rs1800404	Ex10+21G>A	638762
TYMS-10	rs1059394	IVS7-68T>C	662792
TYMS-01	rs699517	Ex8+157C>T	663016
TYMS-05	rs2790	Ex7+227A>G	663086
TFF3-02	rs2236705	IVS2-449T>G	727269
IGF1R-05	rs2137680	IVS2+61405G>A	762592
IGF1R-18	rs2175795	IVS2+61518G>A	762705
TFF1-01	rs2839488	IVS1+334G>C	780627
IGF1R-06	rs907806	IVS2-89673G>A	794732
NOS2A-07	rs9282801	IVS16+88T>G	833467
NOS2A-02	rs2297518	Ex16+14C>T	833591
APEX1-09	rs3136814	Ex1+8A>C	843425
RAD52-01	rs11226	Ex11-571C>T	876074
RAD52-07	rs6413436	IVS10-61C>T	876940
IGF1R-26	rs3743259	IVS5+311A>G	893012
IGF1R-27	rs3743260	IVS5+442A>G	893143
IGF1R-01	rs2229765	Ex16-58G>A	928076
IGF2-09	rs2230949	Ex4-233C>T	941429
IGF2-02	rs734351	IVS2+384C>T	943454
IGF2-22	rs3213223	IVS1-171C>T	944171
IGF2-16	rs3213221	IVS1-285C>G	944285
IGF2-03	rs3213216	IVS1+1280A>G	945420
IGF2AS-04	rs3741212	Ex1+112A>G	949099
IGF1R-12	rs9282715	3164bp 3' of STP C>T	953687
IGF2AS-01	rs1003483	Ex2+69T>G	954784
IGF2AS-03	rs3741211	Ex3+563A>G	956351
ABCA7-05	rs3764651	IVS20+166A>G	991751
ABCA7-06	rs3752241	Ex23-7C>G	993524
RET-01	rs1800858	Ex2+62A>G	999281

RET-02	rs1800860	Ex7+33A>G	1010000
ABCA6-05	rs9282553	Ex16-39G>C	1034634
ABCA6-01	rs9282552	IVS13-16G>A	1036166
GPX4-09	rs757228	-2050A>G	1041992
GPX4-06	rs3746165	-1831G>A	1042211
GPX4-08	rs4807542	Ex1-49G>A	1044078
GPX4-12	rs8178977	IVS6+19C>G	1046477
SRA1-04	rs801460	NC_A>G	1094857
SRA1-03	rs801459	NC_A>C	1096550
SRA1-05	rs10463297	NC_G>A	1099166
GC-02	rs7041	Ex11+34T>G	1125344
STK11-03	rs741764	IVS6+145T>C	1161484
ABCA5-01	rs15886	Ex39+497T>A	1169561
CD14-03	rs2569190	-1994T>C	1175843
CD81-04	rs708155	-1784A>G	1184236
CD81-06	rs810225	IVS1+5757A>C	1191843
PIN1-01	rs2233678	IVS1-834C>G	1207981
PIN1-16	rs2233679	IVS1-659C>T	1208156
PIN1-21	rs889162	IVS3+2592T>C	1214718
PIN1-02	rs1985604	IVS3+3419G>A	1215545
PIN1-17	rs2010457	IVS3+62A>G	1221680
SLC6A18-13	E3563_106	IVS8+267C>T	1233245
WRN-01	rs2230009	Ex4-16G>A	1242709
TERT-03	rs2853690	Ex16+203C>T	1243744
WRN-07	rs2725349	Ex6+9C>T	1245331
TERT-02	rs2075786	IVS10+269T>C	1256310
TERT-15	rs13167280	IVS3-24T>C	1270477
TYR-02	rs1393350	IVS3-6895A>G	1272668
TERT-14	rs2853677	IVS2-4455C>T	1277194
TYR-08	rs1800422	Ex4+21G>A	1279583
TERT-08	rs2735940	Ex2T>C	1286486
ARNT-07	rs1889740	IVS12-662A>G	1290110
WRN-08	rs1800392	Ex20-88G>T	1294731
ARNT-01	rs2228099	Ex7+81G>C	1299244
ARNT-01	rs2228099	Ex7+81G>C	1299244
ARNT-10	rs1027699	IVS6+205G>A	1302067
ARNT-06	rs2256355	IVS6+123G>A	1302149
ARNT-06	rs2256355	IVS6+123G>A	1302149
ARNT-05	rs2864873	IVS5+726T>C	1304529
ARNT-05	rs2864873	IVS5+726T>C	1304529
TERT-54	rs3816659	-22715C>T	1307820
TERT-21	rs1801075	-22844G>A	1307949
WRN-03	rs1801195	Ex26-12G>T	1320054
ARNT-23	rs7517566	-991T>C	1340390
WRN-04	rs1346044	Ex34-93T>C	1345428
FANCA-03	rs1061646	IVS39-16C>T	1366594
FANCA-37	rs7195906	IVS39+55T>A	1366964
FANCA-28	rs17227099	IVS32-42A>G	1375834
FANCA-25	rs12931267	IVS30-102G>C	1379349
FANCA-34	rs2159116	IVS27-36G>T	1392127

FANCA-39	rs7203907	IVS26-129G>C	1394391
FANCA-22	rs886951	IVS24+107C>T	1397482
SLC6A3-10	rs6347	Ex9-55A>G	1401412
FANCA-35	rs3785275	IVS21+121C>G	1402646
FANCA-16	rs2016571	IVS20+933C>G	1404893
FANCA-02	rs2239359	Ex16+31C>T	1410097
FANCA-12	rs2239360	IVS15-73G>A	1410200
SLC6A3-14	rs460700	IVS4+2610A>G	1419969
TNFRSF10A-02	rs2235126	IVS7+218G>A	1431494
TNFRSF10A-06	rs4871857	Ex4-4G>C	1433637
SLC6A3-05	rs2652511	-3076C>T	1436389
SLC6A3-03	rs6413429	-3714G>T	1437027
SEPT2-01	rs7568	Ex13+362A>G	1466347
CBS-03	rs12613	Ex18-391A>G	1468132
ERCC6-04	rs2228529	Ex21+176A>G	1471569
CBS-07	rs6586282	IVS15-134G>A	1472938
CBS-01	rs234706	Ex9+33C>T	1479791
ERCC6-12	rs2228527	Ex18-142A>G	1482833
VIL2-03	rs3123109	IVS6+2368G>A	1489523
VIL2-02	rs901369	IVS2-977T>C	1496551
FOXC1-23	rs6928414	-3906C>T	1546773
FOXC1-22	rs2235716	-3564C>T	1547115
FOXC1-02	rs2235718	-3077C>T	1547602
FOXC1-13	rs9405496	-2049A>C	1548630
FOXC1-06	rs984253	1186bp 3' of STP A>T	1553528
FOXC1-07	rs2745599	1343bp 3' of STP A>G	1553685
ERBB2-03	rs1810132	IVS4-61C>T	1590301
ICAM1-19	rs5030390	IVS1+635G>A	1645339
ICAM1-15	rs281432	IVS2-3499C>G	1653460
ICAM1-06	rs5498	Ex6-22A>G	1658485
ICAM1-16	rs3093032	Ex7+546C>T	1659138
CDKN1C-09	rs431222	-1679G>A	1695640
NUBP2-01	rs344357	IVS1-283C>G	1776256
IGFALS-84	rs1065663	Ex7-164A>G	1779024
IGFALS-91	rs344360	505bp 3' of STP C>T	1779222
MTR-01	rs1805087	Ex26-20A>G	1806289
MTR-01	rs1805087	Ex26-20A>G	1806289
MTR-06	rs2275566	IVS26+43G>A	1806351
MTR-05	rs2275565	IVS26+157T>G	1806465
TEP1-03	rs1713449	Ex45+36G>A	1841547
TEP1-02	rs1760904	Ex24+49T>C	1851869
TEP1-10	rs872072	IVS13+84C>T	1858853
TEP1-11	rs872074	Ex13+51G>A	1859045
TEP1-01	rs1760898	Ex4+51C>A	1872721
TEP1-08	rs1760897	Ex1-222T>C	1876093
TNKS-01	E5132_301	Ex1-74G>A	1889400
TNKS-03	E5132_489	IVS1+115C>T	1889588
TNKS-05	E5133_300	IVS1+381C>T	1889854
TNKS-46	E5133_164	IVS1+517C>G	1889990
CSF3-06	rs2227338	IVS3+58A>G	1897238

CSF3-02	rs1042658	Ex4-165C>T	1898198
APEX1-16	rs1760944	-655T>G	1922989
APEX1-03	rs3136820	Ex5+5T>G	1924994
TNKS-33	rs7462102	IVS2-11724C>T	1936719
TNKS-34	rs7464476	IVS2-11659A>C	1936784
TNKS-64	E5135_413	IVS3+238C>T	1948776
TNKS-35	E5153_301	IVS3+11245A>T	1959783
TNKS-36	E5154_301	IVS3+21545G>T	1970083
TNKS-76	rs13276464	IVS3-25352G>T	1987460
TNKS-11	E5137_301	IVS3-25329G>T	1987483
TNKS-12	E5138_301	IVS3-23879C>T	1988933
TNKS-13	rs6985140	IVS3-34A>G	2012778
BPI-01	rs1131847	Ex15+70A>G	2018532
TNKS-38	rs11249938	IVS5+6476A>G	2020137
TNKS-15	rs7006985	Ex8-71A>G	2039788
TNKS-18	rs12542457	IVS8+93C>T	2039951
TNKS-20	rs7462910	IVS12+1931C>T	2055337
TNKS-22	rs7001395	IVS14-34A>T	2066106
TNKS-23	rs9644708	IVS17-4617T>C	2076268
TNKS-26	E5147_301	IVS23-32A>C	2098521
EPHX1-15	rs2854461	-4786A>C	2187838
EPHX1-14	rs2671272	IVS1-1310G>A	2191310
EPHX1-18	rs3738043	IVS1-1127A>G	2191493
EPHX1-17	rs2854456	IVS1-1067C>T	2191553
COL18A1-02	rs2236451	Ex3-8A>G	2193419
EPHX1-06	rs1051740	Ex3-28T>C	2195827
EPHX1-10	rs2260863	IVS3+114C>G	2195968
EPHX1-01	rs2234922	Ex4+52A>G	2202600
EPHX1-12	rs1051741	Ex8+31C>T	2208423
CASP9-27	rs2020898	IVS7-122C>T	2213456
COL18A1-03	rs2236467	Ex13-25C>T	2217746
CASP9-01	rs1052576	Ex5+32A>G	2225381
CASP9-03	rs2020902	IVS3+8T>C	2227198
IL1A-01	rs17561	Ex5+21G>T	2244966
IL1A-04	rs2071374	IVS4-109A>C	2245095
COL18A1-01	rs7499	Ex43+227A>G	2249664
SLC19A1-05	rs1051298	Ex8-198C>T	2252162
SLC19A1-01	rs1051266	Ex4-114T>C	2275130
IL1B-08	rs1071676	Ex7-97C>G	2295176
IL1B-02	rs1143634	Ex5+14C>T	2298133
IL1B-12	rs3136558	IVS3-123C>T	2299018
IL1B-03	rs1143627	-580C>T	2302130
SOD2-01	rs1799725	Ex2+24T>C	2401213
SOD2-06	rs5746081	-1254C>T	2402795
FLJ45983-03	rs1149901	Ex1-425G>A	2457683
FLJ45983-16	rs1269486	-2994T>C	2459095
GATA3-46	rs10905277	-250A>G	2460264
GATA3-10	rs2229359	Ex2-158C>T	2463543
GATA3-76	rs3781093	IVS2+1123T>C	2464823
SLC39A2-10	rs945352	-824T>C	2466621

SLC39A2-07	rs2149666	IVS2-119G>T	2467996
SLC39A2-05	rs2234636	Ex4+46T>C	2468991
GATA3-21	rs570730	IVS3+358T>C	2469355
GATA3-68	rs10752126	IVS3+646C>G	2469643
GATA3-23	rs569421	IVS3+2491C>T	2471488
GATA3-25	rs520236	IVS3-2162C>G	2472170
GATA3-27	rs422628	IVS3-27C>T	2474305
GATA3-28	rs406103	IVS4+60C>T	2474517
GATA3-29	rs528778	IVS4+582C>T	2475039
LDLR-01	rs1003723	IVS9-30C>T	2486983
LDLR-12	rs5930	Ex10+55A>G	2487067
LDLR-18	rs5925	Ex13-29T>C	2493683
MGMT-12	rs16906252	N/A	2499476
LDLR-08	rs2116898	IVS17-147G>A	2504612
LDLR-03	rs14158	Ex18+88A>G	2504846
IL1RN-05	rs419598	Ex5-35T>C	2594950
IL1RN-02	rs454078	IVS6+59A>T	2596536
IL1RN-04	rs380092	IVS6+166A>T	2596643
SCARB1-09	rs989892	IVS7-2428G>T	2697149
SCARB1-08	rs865716	IVS7+1451A>T	2700789
PARP1-14	rs747659	IVS21+59G>A	2726935
PARP1-13	rs747657	IVS20-63G>C	2727118
SCARB1-03	rs4765621	IVS1-18462G>A	2730648
PARP1-01	rs1136410	Ex17+8T>C	2731496
SCARB1-01	rs3924313	IVS1+19766T>C	2738308
MGMT-06	rs12917	Ex2-25C>T	2740214
IGF2R-05	rs1570070	Ex9+5A>G	2741319
PARP1-12	rs1805415	Ex8+45A>G	2747034
SCARB1-02	rs4765181	IVS1+8913T>G	2749161
PARP1-10	rs1805414	Ex7+18T>C	2749558
IGF2R-02	rs998075	Ex16+88A>G	2755619
IGF2R-11	rs998074	IVS16+15T>C	2755724
PARP1-06	rs1805407	IVS2+82A>G	2766027
ALDH2-08	rs2238151	IVS1+6933C>T	2781342
IGF2R-04	rs629849	Ex34-93A>G	2781750
IGF2R-07	rs2282140	IVS34+20C>T	2781862
MGMT-19	rs2296675	IVS3-54A>G	2798929
MGMT-03	rs2308327	Ex4+119A>G	2799101
IGF2R-03	rs1803989	Ex45+11C>T	2804822
KRT23-03	rs2269858	Ex2+338C>T	2817164
LITAF-01	rs7102	1050bp 3' of STP T>C	2955321
LITAF-02	rs4280262	Ex3+54A>G	2960571
BCR-01	rs12233352	IVS8-20A>G	3006370
BCR-02	rs140504	Ex11-20A>G	3017938
TXNRD2-76	rs6518591	IVS2+1485T>C	3076171
MBL2-30	rs2099902	Ex4-710G>A	3077004
MBL2-27	rs10082466	Ex4-1483T>C	3077777
MBL2-09	rs930508	IVS3-28G>C	3079453
TXNRD2-83	rs9306230	IVS1+1202T>C	3080172
MBL2-06	rs1838066	IVS2-250T>C	3080480

MBL2-46	rs10824793	IVS2-405T>C	3080635
TXNRD2-88	rs4646310	IVS1+418T>C	3080956
MBL2-03	rs5030737	Ex1-34C>T	3082397
MBL2-12	rs7096206	-289G>C	3082840
MBL2-44	rs11003124	-495C>A	3083046
MBL2-11	rs11003125	-618G>C	3083169
MBL2-38	rs1031101	-1964T>C	3084515
MBL2-65	rs12264958	-2200T>C	3084751
COMT-29	rs7290221	IVS1-6042C>G	3094830
COMT-16	rs4646312	IVS1-385C>T	3100487
COMT-03	rs6269	IVS2-98A>G	3102102
COMT-01	rs4680	Ex4-12G>A	3103421
IL8-01	rs4073	-351A>T	3113034
IL8-11	rs2227549	IVS1+298A>G	3113747
IL8-05	rs2227306	IVS1-204C>T	3114065
TP53I3-03	rs2303287	IVS4+68G>A	3118179
TP53I3-12	rs4149372	IVS2+243A>C	3121445
ARVCF-05	rs2240716	IVS3-82A>G	3121846
TP53I3-18	rs4149371	-578A>G	3123708
TP53I3-10	rs10170774	-931G>A	3124061
TP53I3-13	rs7603220	-2503T>C	3125633
PLA2G2A-03	rs2236771	Ex4+56G>C	3129304
EDN1-01	rs5369	Ex3-72A>G	3152516
EDN1-02	rs5370	Ex5+61G>T	3154513
NINJ1-03	rs1127857	Ex4-86C>G	3205070
NINJ1-01	rs1127851	Ex4-179A>T	3205163
ALOX5-02	rs4986832	-1699A>G	3271341
ALOX5-06	rs4987105	Ex1+63C>T	3273061
ALOX5-10	rs2029253	IVS3+100A>G	3294797
ALOX5-12	rs1369214	IVS3-6910A>G	3304042
ALOX5-15	rs892691	IVS4-2397A>G	3320405
ALOX5-28	rs1565097	IVS7+44C>T	3327569
ALOX5-26	rs2242332	IVS9-247C>T	3341552
IL12B-04	rs3212227	Ex8+159A>C	3552508
IL12B-11	rs730690	IVS1+1274G>A	3565724
CDC25B-06	rs910656	1887bp 3' of STP G>A	3727496
MSR1-02	rs971594	IVS7+563G>C	3852196
MSR1-01	rs414580	IVS2+93A>T	3880321
UCP3-01	rs2075577	Ex5-14C>T	3938291
UCP3-02	rs1800849	-2077C>T	3942914
ALOX15-02	rs2664593	-11273G>C	4148505
ALOX15-12	rs7220870	-271T>G	4148591
NBN-04	rs1063053	Ex16+304C>T	4165710
CTSB-03	rs1065712	Ex12-296C>G	4177472
NBN-13	rs867185	IVS8+1488T>C	4193323
NBN-01	rs1805794	Ex5-32C>G	4208652
NBN-02	rs1063045	Ex2+65G>A	4213192
CARD15-04	rs2066850	-925A>C>G>T	4344428
CARD15-19	rs2067085	Ex2-7C>G	4348058
CARD15-05	rs2066843	Ex4+731C>T	4359398

CARD15-09	rs748855	IVS6+513A>G	4365597
CARD15-10	rs1077861	IVS10+64T>A	4373746
HSD17B1-06	rs597255	IVS1-42C>T	4429396
HSD17B1-10	rs676387	IVS4-150C>A	4430569
COASY-01	rs598126	Ex3+57A>G	4440816
PGR-18	rs1870019	8373bp 3' of STP C>T	4463889
PGR-23	rs561650	IVS6-3061G>A	4478310
PGR-01	rs1042839	Ex5-48C>T	4484618
PGR-12	rs492457	IVS4-4561T>C	4489276
PGR-07	rs9282823	IVS4+566G>A	4495028
PGR-11	rs1042838	Ex4+72G>T	4495828
PGR-26	rs660541	IVS3-884T>C	4496783
PGR-14	rs516693	IVS2-1136A>G	4526159
PGR-17	rs572483	IVS2-4965A>G	4529988
PGR-16	rs543215	IVS2-11426G>A	4536449
PGR-05	rs613120	IVS2-11671T>C	4536694
PGR-28	rs565186	IVS2+13109G>A	4546045
PGR-15	rs529359	IVS2+2892A>G	4556262
PGR-21	rs481775	IVS2+2063T>C	4557091
PGR-27	rs10895068	Ex1+1042A>G	4562630
PGR-20	rs474320	-14747A>T	4576965
PGR-24	rs568157	-24480T>C	4586698
SLC23A2-02	E1028_301	Ex17-1890G>A	4774894
SLC23A2-01	rs1110277	Ex11+57T>C	4794682
SLC23A2-05	rs4987219	IVS8+453C>G	4804946
SLC23A2-03	rs1776964	Ex6+51C>T	4820308
SLC23A2-25	rs1715364	IVS3-5272G>A	4838896
SLC23A2-33	rs4813725	IVS2-4623G>A	4857985
IL6R-04	rs8192284	Ex9+7A>C	4917325
SLC23A2-48	rs6084957	IVS1+1547G>A	4920505
SLC23A2-31	rs12479919	IVS1+1312G>A	4920740
BRCA1-21	rs1799966	Ex17-150A>G	4947390
BRCA1-32	rs8176212	IVS15-63C>G	4950897
BRCA1-05	rs1060915	Ex14-50T>C	4958766
BRCA1-20	rs4986852	Ex12-978G>A	4968725
BRCA1-01	rs16940	Ex12+1641T>C	4969533
BRCA1-06	rs1799949	Ex12+1412C>T	4969762
BRCA1-18	rs1799950	Ex12+397A>G	4970777
BRCA1-26	rs799923	IVS8-34C>T	4976227
PCNA-10	rs17352	IVS5+140A>C	5037976
PCNA-07	rs17349	IVS2-4C>T	5039516
PCNA-06	rs25406	IVS2-124C>T	5039636
AKR1C3-31	rs10795241	-32346T>C	5044290
LRP6-03	rs2075241	IVS15-11G>C	5050453
AKR1C3-29	rs28943575	-23066G>C	5053570
AKR1C3-33	rs6601899	-18314A>C	5058322
AKR1C3-30	rs17134288	Ex2C>T	5072588
AKR1C3-28	rs28942669	-1632C>T	5075004
AKR1C3-17	rs11252937	-1423T>C	5075213
AKR1C3-19	rs1937845	-488A>G	5076148

AKR1C3-24	rs3763676	-137A>G	5076499
AKR1C3-01	rs12529	Ex1-70C>G	5076651
AKR1C3-08	rs2245191	IVS3+73C>A	5079815
AKR1C3-11	rs2275928	IVS8+40A>G	5087909
AKR1C3-21	rs10904422	1532bp 3' of STP G>C	5091228
AKR1C3-26	rs7070041	1875bp 3' of STP A>G	5091571
AKR1C3-36	rs7921327	9757bp 3' of STP A>G	5099453
AKR1C3-35	rs1937920	12259bp 3' of STP G>A	5101955
AR-12	rs1204038	IVS1+21621G>A	5106213
LRP6-02	rs3782528	IVS3+4071A>T	5111040
AR-15	rs2361634	IVS1-255A>G	5180831
AKR1C4-01	rs3829125	Ex4-14C>G	5187784
AR-13	rs1337080	IVS2+15670A>G	5196907
SLC30A1-01	rs2278651	IVS1+273C>T	5214167
HADHA-05	rs1049987	Ex20-309A>G	5229750
HADHA-01	rs7260	Ex20+348A>G	5229850
HADHA-10	rs2289019	IVS13-163C>G	5236742
AR-14	rs1337082	40331bp 3' of STP A>G	5302003
ERCC4-01	rs1800067	Ex8+31G>A	5342112
ERCC4-15	rs1799800	IVS9-28A>G	5351631
RB1CC1-24	rs17845549	Ex23+32T>C	5390668
PMS2-11	rs6463524	Ex7-24G>C	5394028
PMS2-10	rs2345060	IVS5-223T>C	5396177
PMS2-01	rs3735295	IVS1+72G>A	5405604
RB1CC1-10	rs2305427	Ex18+83T>C	5408421
JTV1-01	rs2009115	IVS1-2221G>A	5409604
RB1CC1-40	rs17337252	Ex7+129T>C	5440058
RB1CC1-50	Poly-0014935	-29373T>C	5480755
CDKN1B-04	rs7330	Ex3-387C>A	5633891
TGM1-01	rs2229463	Ex15-194C>T	5718355
TGM1-02	rs2855006	Ex7-14C>A	5728134
BIRC3-02	rs3758841	IVS6+879C>T	5765267
BIRC3-03	rs3460	Ex9-76C>G	5770806
EPHX2-04	rs1126452	Ex19+4A>C	5776249
BIRC2-01	rs1943781	IVS5+694G>A	5797557
RAC1-03	rs2303364	IVS7+30C>T	5798751
PARP4-01	rs13428	Ex31+172G>C	5989441
PARP4-03	rs6413414	Ex20-19A>T	6013200
PARP4-19	rs750771	IVS17-110A>G	6014384
PARP4-23	rs1807111	IVS15-995C>T	6025158
PARP4-17	rs1539096	Ex3-89G>A	6055859
PHB-02	rs4987082	979bp 3' of STP A>G	6134652
MMP1-03	rs5854	Ex10-224C>T	6223290
MMP1-09	rs2071230	Ex10+294T>C	6223375
MMP1-05	rs5031036	IVS5+19A>G	6228580
MMP1-01	rs10488	Ex2-36G>A	6230438
BLM-02	rs2238335	IVS1-253G>C	6255893
BLM-05	rs2072352	IVS15+105G>A	6299706
BLM-22	rs2072351	IVS15+184T>A	6299785
TNFRSF1A-02	rs887477	IVS1-2572G>T	6300243

BLM-25	rs16944831	IVS16-479C>T	6306468
BLM-03	rs2270132	IVS19-499A>C	6317395
BLM-16	rs389480	IVS19-437G>A	6317457
BLM-06	rs2073919	IVS20+630G>A	6318646
HTR1D-01	rs605367	Ex1-137G>A	6342866
HTR1D-04	rs6300	Ex1-1246T>C	6343975
HTR1D-03	rs676643	-627T>C	6345682
GSTM3-05	rs2234696	Ex8+190A>C	6365717
GSTM3-01	rs7483	Ex8+91G>A	6365816
GSTM3-06	rs1537234	IVS7-30G>T	6365936
MTHFR-07	rs12121543	IVS7-76T>G	6392038
MTHFR-02	rs1801133	Ex5+79C>T	6393745
MTHFR-02	rs1801133	Ex5+79C>T	6393745
MTHFR-03	rs2066470	Ex2-120C>T	6400424
MAOA-01	rs6323	Ex2-65G>T	6440845
ALOX12-02	rs1126667	Ex6-26A>G	6500108
CD4-03	rs3213427	Ex10+283T>C	6783008
SLC2A4-02	rs5435	Ex4-59T>C	6784471
EXO1-02	rs735943	Ex11+20A>G	6787940
EXO1-01	rs4149963	Ex12+49C>T	6793171
INSR-28	rs3745551	2778bp 3' of STP A>G	7054288
INSR-06	rs1051690	Ex22-326T>C	7056963
INSR-01	rs1799817	Ex17-4C>T	7065297
INSR-07	rs2860175	IVS14+88A>G	7072081
INSR-30	rs8110533	IVS13+2860C>A	7078828
MPDU1-01	rs4227	Ex7-334G>T	7088525
INSR-19	rs3815901	IVS7-126C>T	7106541
INSR-05	rs891087	Ex3+131C>T	7124518
SAT2-01	rs13894	Ex6+31C>T	7127251
SAT2-03	rs858520	Ex4-11G>A	7127620
SHBG-05	rs6257	IVS1-17T>C	7131066
SHBG-13	rs858517	IVS3+84C>T	7131620
SHBG-01	rs6259	Ex8+6G>A	7133876
SHBG-12	rs727428	1121bp 3' of STP T>C	7135141
INSR-11	rs1035942	IVS2-15155C>T	7139803
INSR-51	rs1035940	IVS2-15330C>G	7139978
ATP1B2-13	rs1641535	-8703T>C	7143482
ATP1B2-04	rs1624085	Ex2G>C	7151092
INSR-61	rs3745545	IVS2-27193A>G	7151841
INSR-59	E1424_156	IVS2-27322G>A	7151970
ATP1B2-01	rs1641512	Ex7+414G>A	7156811
TP53-14	rs1614984	21226bp 3' of STP C>T	7168801
TP53-11	rs12951053	IVS7+92T>G	7174756
TP53-66	rs2909430	IVS4-91A>G	7175994
TP53-09	rs8079544	IVS1-112G>A	7177401
TP53-69	rs2078486	IVS1-3143C>T	7180432
WDR79-11	rs2287499	Ex1-230C>G	7189517
WDR79-09	rs17885803	IVS1-60C>T	7189831
WDR79-08	rs2287498	Ex2+19C>T	7189909
WDR79-06	rs17886268	IVS2-106C>T	7190151

INSR-13	rs919275	IVS2+5915A>G	7201441
EFNB3-01	rs3744263	Ex5-986T>C	7211057
EFNB3-02	rs3744262	Ex5-929G>A	7211114
MYBL2-19	rs385345	-3962C>T	7344876
MYBL2-30	rs619289	-1310C>T	7347528
MYBL2-31	rs826950	IVS1-75C>T	7355286
MYBL2-06	rs419842	IVS3+316A>T	7363726
MYBL2-36	rs285164	IVS7+92C>T	7381691
MYBL2-03	E3013_458	IVS9+172A>G	7387085
MYBL2-46	rs420755	IVS9-1919C>G	7389599
MYBL2-09	rs285171	IVS9-728C>G	7390790
LPL-01	rs263	IVS5-540C>T	7657740
LPL-04	rs316	Ex8+25C>A	7663364
LPL-08	rs325	IVS8-298C>T	7664256
LPL-03	rs326	IVS8-187A>G	7664367
LPL-09	rs327	IVS8-90G>T	7664464
LPL-05	rs328	Ex9-7C>G	7664652
LPL-06	rs1059507	Ex10-807C>T	7668891
XPA-02	rs1800975	Ex1+62T>C	7780783
SCUBE2-13	rs3751058	IVS18+25A>G	7838647
ARHGDIB-01	rs921	Ex6-198C>T	7854136
SCUBE2-02	rs3751052	IVS13+13G>A	7859381
ARHGDIB-03	rs2075267	-11454C>A	7874075
ARHGDIB-03	rs2075267	-11454C>A	7874075
SCUBE2-03	rs2003906	IVS6-6T>C	7874774
MTRR-22	rs2287779	Ex9+9G>A	7879216
MTRR-05	rs2287780	Ex9-85C>T	7879304
MTRR-10	rs10380	Ex14+14C>T	7887191
MTRR-11	rs1802059	Ex14-42G>A	7887319
MTRR-07	rs9332	Ex15-526G>A	7890712
MTRR-19	rs8659	Ex15-405A>T	7890833
RERG-31	rs1045733	Ex5-277T>C	8019967
RERG-10	rs17834986	Ex5-721A>T	8020411
RERG-41	rs1055151	Ex5+273C>T	8021153
RERG-44	rs2193174	IVS2-30357C>T	8063384
RERG-36	rs3748302	IVS2-30974G>A	8064001
LIG3-08	rs1052536	Ex21-250C>T	8068555
RERG-37	rs715398	IVS2-36011A>C	8069038
RERG-29	E3288_225	IVS2-36221T>G	8069248
RERG-30	rs10160846	IVS2-36286G>A	8069313
RERG-33	rs2216225	IVS2+27493C>T	8101844
RERG-47	rs767201	IVS1-341T>C	8129852
RERG-24	rs6488766	-4206G>A	8133604
RERG-03	E3265_338	-6205A>G	8135603
DIO1-01	rs1883454	-850G>T	8178426
DIO1-05	rs2235544	IVS3-34A>C	8194963
PTEN-10	rs1903858	IVS1-96G>A	8402202
PTEN-01	rs701848	1515bp 3' of STP C>T	8475261
XBP1-02	rs2267131	1062bp 3' of STP G>A	8581040
XBP1-01	rs2097461	IVS4+156G>A	8582448

XBP1-09	rs2239815	IVS3+395G>A	8583239
XBP1-10	rs3788409	-1638C>A	8588720
ABCC4-04	rs3765535	IVS17-65T>C	8905211
CCL5-04	rs2280789	IVS1+231C>T	8943983
CCL5-03	rs2107538	-10155G>A	8944760
ABCC4-07	rs2274406	Ex8+40A>G	8948672
JAK3-01	rs3008	Ex23+291A>G	9200231
JAK3-12	rs3212752	IVS12+9A>G	9211534
JAK3-02	rs3212711	IVS2+22C>T	9217823
TGFBR1-03	rs928180	IVS3-2409A>G	9218937
TGFBR1-04	rs334358	IVS8+547G>T	9231818
TGFBR1-01	rs868	Ex9+195A>G	9232861
FAS-09	rs2234768	-690C>T	9498459
FAS-01	rs1324551	IVS1+853T>C	9500032
FAS-04	rs1468063	Ex9-252T>C	9523807
OGG1-12	rs125701	-1492G>A	9730478
OGG1-13	rs2304277	IVS7+110G>A	9741080
GDF15-01	rs1059519	Ex2+42G>C	9759826
GDF15-02	rs1059369	Ex2-136T>A	9759943
CD40-01	rs1535045	IVS1+1066C>T	9801007
CD40-03	rs3765459	IVS8-114G>A	9810315
CRP-03	rs1205	3431bp 3' of STP G>A	10172588
CRP-02	rs1800947	Ex2+491G>C	10173793
SEC14L2-01	rs1010324	IVS2+182A>G	10186458
SEC14L2-04	rs2267154	IVS5-314A>G	10195431
SEC14L2-05	rs2267155	IVS8-1361C>T	10200956
HSPB8-01	rs11038	Ex3-245A>G	10201816
VDR-12	rs757343	IVS8+443G>A	10382981
VDR-07	rs2239185	IVS6-3968C>T	10387865
CSF1R-05	rs10079250	Ex8+3A>G	10613068
CETP-08	rs820299	IVS2-3014T>G>C>A	10614483
CSF1R-03	rs3829987	IVS6+28C>T	10619747
CSF1R-02	rs2228422	Ex5-4C>T	10620614
CETP-23	rs289717	IVS10+325G>A	10623587
CETP-21	rs1801706	Ex16-95A>G	10631861
SLAMF1-04	rs1061217	Ex7-127C>T	11070370
SLAMF1-03	rs164283	IVS4-1126G>A	11081120
SLAMF1-02	rs2295612	Ex1-44C>A	11107058
NCOA3-04	rs2076546	Ex15-74A>G	11321401
PAK6-43	rs900055	IVS1-216G>A	11323141
NCOA3-01	rs396221	IVS17-1728T>G	11326997
NCOA3-02	rs427967	IVS17-1632A>G	11327093
PAK6-24	rs11636097	Ex3-53A>G	11335995
PAK6-19	rs936216	IVS4+78C>A	11347825
PAK6-16	rs748556	IVS5+48A>G	11349301
PAK6-14	rs2242120	Ex11+558T>A	11359242
PAK6-13	rs2242119	Ex11+696C>A	11359380
GPX3-21	rs2042235	Ex2T>C	11560850
GPX3-04	rs1946234	-1005A>C	11562146
GPX3-25	rs8177404	-631C>T	11562520

GPX3-28	rs8177426	IVS1-1961A>G	11565876
GPX3-18	rs869975	IVS2-89A>G	11569308
GPX3-16	rs8177447	IVS4-14T>C	11570392
TNIP1-02	rs2277940	796bp 3' of STP T>C	11572413
NPAT-01	rs228589	IVS1+19A>T	11655624
ATM-06	rs189037	-4518A>G	11656249
APOA2-04	rs6413453	IVS3-4C>T	11682671
APOA2-06	rs5085	IVS3+197G>C	11682866
APOA2-02	rs5082	-756C>T	11684038
ATM-27	rs664677	IVS21-77C>T	11705598
ATM-02	rs1800889	Ex31-34C>T	11725903
ATM-03	rs1801516	Ex38+61G>A	11737878
RAD51-22	rs2619679	-4719T>A	11776794
RAD51-01	rs1801320	Ex1-96G>C	11778085
RAD51-01	rs1801320	Ex1-96G>C	11778085
RAD51-23	rs2619681	IVS1+1398T>C	11779578
RAD51-15	rs2304579	IVS2+110A>G	11781710
RAD51-17	rs4924496	IVS3+1932T>C	11785892
ATM-01	rs664143	IVS62+60G>A	11788077
ATM-01	rs664143	IVS62+60G>A	11788077
RAD51-16	rs2412546	IVS5-4480G>A	11797080
ATM-37	rs170548	IVS62-973A>C	11797252
ATM-38	rs3092993	IVS62-694C>A	11797531
RAD51-24	rs4144242	IVS5-4016G>A	11797544
MGC33948-02	rs4585	IVS10-12792C>A	11802044
RAD51-21	rs2412547	IVS6+1598C>A	11803252
RAD51-20	rs11852786	1131bp 3' of STP G>C	11815065
GRPR-01	rs4986945	Ex2+40T>C	12135032
GRPR-02	rs4986946	Ex2-103T>C	12135242
P2RX7-10	rs3751144	Ex13+132C>T	12191748
PTH-01	rs6256	Ex3+161C>A	12301294
PTH-04	rs6254	IVS2-50A>G	12301504
PTH-03	rs177706	IVS1-98A>G	12301746
PPARG-11	rs1801282	Ex4-49C>G	12333125
PPARG-06	rs2938392	IVS7+357G>A	12374608
PPARG-07	rs1175541	IVS9+6835A>C	12405488
ABCB1-12	rs1211152	IVS4-118G>T	12448786
ABCB1-01	rs2235074	IVS4+36C>T	12458715
ABCB1-09	rs9282564	Ex3-8G>A	12463109
PCTP-01	rs2114443	-1212A>G	12480530
PCTP-03	rs12948867	-1148A>G	12480594
POLB-05	rs3136717	IVS1-89C>T	12516830
POLB-16	rs2979895	IVS2-2264G>A	12520596
POLB-08	rs2953983	IVS7+171G>A	12533645
BIC-21	E5157_511	NC_G>T	12597384
BIC-33	rs4143370	NC_G>C	12598844
BIC-15	rs12482371	NC_C>T	12600224
BIC-04	rs915860	NC_C>G	12600356
BIC-34	rs4817027	NC_A>G	12600860
BIC-07	rs1893650	NC_C>T	12602686

BIC-01	rs928883	NC_A>G	12605896
BIC-10	rs2829801	NC_G>T	12606176
BIC-11	rs767649	NC_A>T	12606593
BIC-32	rs2829803	NC_A>G	12610181
GSTP1-01	rs947894	Ex5-24A>G	12658484
GSTP1-02	rs1799811	Ex6+5C>T	12659374
SLC2A1-01	rs1770810	IVS2+2784C>T	13378031
LRP5-01	rs312016	IVS1+2130T>C	13388198
LRP5-04	rs491347	IVS7-1263G>A	13475483
LRP5-15	rs608343	IVS16-213T>C	13502625
LRP5-06	rs607887	IVS16-82C>T	13502756
LRP5-07	rs3736228	Ex18-12C>T	13507090
RGS5-01	rs15049	Ex5+416A>C	13525818
BRCA2-25	rs1799943	Ex2+14A>G	13870572
BRCA2-01	rs144848	Ex10+321A>C	13886729
BRCA2-02	rs1801406	Ex11+1487A>G	13891888
BRCA2-03	rs543304	Ex11+1898T>C	13892299
BRCA2-04	rs1799955	Ex14-194A>G	13909232
BRCA2-32	rs206147	IVS24+5507C>T	13939789
BRCA2-06	rs15869	Ex27-336A>C	13953012
CG018-03	rs1207953	IVS6-195G>C	13957532
MGC20255-03	rs2241719	Ex5+1082T>A	14097799
TGFB1-03	rs1800471	Ex1-282C>G	14127094
XPC-01	rs2228001	Ex16+211A>C	14127450
MGC4093-03	rs1800469	308bp 3' of STP C>T	14128514
XPC-08	rs3731151	Ex11+28A>G	14133890
XPC-03	rs2228000	Ex9-377C>T	14139889
APC-09	rs2229992	Ex12+50T>C	14577867
APC-19	rs2909786	IVS14-2583A>G	14583078
APC-03	rs41115	Ex16+2521G>A	14590783
APC-26	rs866006	Ex16+3310T>G	14591572
APC-13	rs459552	Ex16+3507T>A	14591769
ABCA1-17	rs2230808	Ex35-14A>G	14884009
ABCA1-31	rs2297404	IVS33-26C>G	14885671
ABCA1-15	rs2777801	IVS32+30T>G	14888082
ABCA1-12	rs4149313	Ex18-8A>G	14907958
ABCA1-26	rs7031748	Ex15-76C>A	14912477
ABCA1-04	rs2230806	Ex7-65G>A	14942072
MPO-04	rs2071409	IVS11-6A>C	15001508
PLK1-15	rs40076	IVS3+26A>G	15005484
CYP2C19-08	rs4986894	-97C>T	15270891
CYP2C19-03	rs4244285	Ex5+39G>A	15290142
SOD3-05	rs2855262	Ex3-489C>T	15477334
CAV1-29	rs6950798	-3533C>T	15588122
CAV1-19	rs10257125	-2812A>T	15588843
CAV1-23	E5097_419	-2499C>G	15589156
CAV1-02	rs2215448	-1164A>G	15590491
IGFBP6-18	rs7974876	-18227T>C	15622440
CAV1-05	rs8713	Ex3+798A>C	15626340
CAV1-07	rs1049334	Ex3-851A>G	15626923

CAV1-09	rs1049337	Ex3-644C>T	15627130
IGFBP6-17	rs12821902	-3308C>T	15631499
IGFBP6-19	rs822688	IVS1-1109C>T	15636693
SOAT2-21	rs2280698	-73A>G	15640594
SOAT2-01	rs2280699	IVS1-237A>G	15641004
SOAT2-09	rs17123210	IVS1-8C>G	15641233
MET-13	rs11762213	Ex2+158G>A	15765817
MET-26	E4094_67	Ex2+548C>T	15766207
GGH-02	rs1031552	IVS7-3001C>T	15786543
GGH-01	rs719235	-353C>A	15805034
MET-04	rs13223756	Ex7-22A>G	15824125
MET-01	rs41736	Ex20+60C>T	15862321
HTR1B-02	rs6296	Ex1-313G>C	15992431
HTR1B-07	rs130058	-160A>T	15993452
AKR1A1-02	rs2088102	IVS5+282T>C	16004892
HAO2-01	rs1417604	IVS4+707G>A	16014498
HSD3B2-25	rs879332	-17124G>A	16027033
HSD3B2-19	rs4659175	-1569T>C	16042588
HSD3B2-14	rs12411115	IVS2-1665G>T	16046491
HSD3B2-07	rs1361530	Ex4-88C>G	16051679
LOC391073-01	rs1417608	10174G>A	16063948
HSD3B1-23	rs2064902	-31680A>C	16104534
HSD3B1-26	rs6667572	-27428A>G	16108786
HSD3B1-24	rs4659182	-24247G>T	16111967
HSD3B1-22	rs1998182	-12386A>G	16123828
HSD3B1-25	rs6428830	IVS3+485A>G	16140890
HSD3B1-18	rs10754400	713bp 3' of STP G>T	16144097
XRCC1-01	rs25487	Ex10-4A>G	16323944
SLC30A4-01	rs1153829	Ex8-66G>A	16567901
ERCC5-01	rs1047768	Ex2+50T>C	16594193
ERCC5-05	rs2227869	Ex8-369G>C	16604761
ERCC5-02	rs17655	Ex15-344G>C	16617678
AHR-19	rs7796976	Ex1+185A>G	16704367
AHR-17	rs2074113	IVS7+33T>G	16739720
ERCC3-04	rs4150474	IVS10-2790G>T	16741069
AHR-01	rs2066853	Ex10+501G>A	16745061
ERCC3-02	rs4150416	IVS6-108G>T	16754290
ZNF230-01	rs12753	Ex5-284C>A	16783732
ANKK1-01	rs1800497	Ex8-313G>A	16833244
DRD2-03	rs1079597	IVS1-882A>G	16858702
DRD2-01	rs1799978	-50977T>C	16908767
HFE-01	rs1799945	Ex2+111C>G	16949430
HFE-07	rs1572982	IVS5-47A>G	16952618
HFE-08	rs707889	IVS6+462G>A	16954182
SSTR3-01	rs229569	Ex2-807C>T	16993566
SSTR3-03	rs86582	Ex1+453G>A	16993905
FAM82A-01	rs163077	IVS10-8520T>C	17101538
FAM82A-08	rs1367696	IVS10-7211A>G	17102847
FAM82A-02	rs163086	IVS10-1363T>C	17108695
CYP1B1-08	rs10916	Ex3+1284G>T	17113103

CYP1B1-31	rs162562	Ex3+939A>C	17113448
CYP1B1-07	rs1800440	Ex3+315A>G	17114072
CYP1B1-27	rs162556	-3922C>T	17122387
CYP1B1-28	rs162555	-4977A>G	17123442
CYP1B1-18	rs10175368	-5329G>A	17123794
TSG101-30	rs2291752	364bp 3' of STP G>A	17288969
TSG101-40	rs2279900	IVS9+18G>A	17290400
TSG101-28	rs2279902	IVS7-13T>C	17292873
TSG101-33	rs2292176	IVS5+61G>T	17318269
TSG101-07	rs12574333	IVS4+10C>A	17323456
TSG101-36	rs2292179	-182T>C	17335787
CYP7B1-03	rs1451868	9712bp 3' of STP C>T	17352839
CYP7B1-02	rs1376772	9625bp 3' of STP C>T	17352926
CYP7B1-06	E3566_386	IVS4-1678T>C	17372445
CYP7B1-01	rs3779870	IVS4-1752T>C	17372519
RAD23B-02	rs1805335	IVS5-15A>G	17402223
RAD23B-03	rs1805330	IVS6-3C>T	17405466
RAD23B-04	rs1805329	Ex7+65C>T	17405533
RAD23B-05	rs1805334	IVS7-22A>G	17407354
APOE-03	rs440446	IVS1+69C>G	17677385
CYP24A1-08	rs751087	IVS7-1255A>G	17829825
CYP24A1-05	rs2296241	Ex4+9T>C	17839127
CYP24A1-03	rs2259735	IVS2-105T>C	17841222
CYP24A1-01	rs2248359	Ex2G>A	17844426
PLA2G6-08	rs2016755	IVS3-309T>C	17930119
PLA2G6-10	rs84473	IVS2+7899A>G	17947841
PLA2G6-02	rs4376	IVS2+4480G>A	17951260
PLA2G6-12	rs132987	IVS2+1653T>C	17954087
CASP3-09	rs6948	Ex8-280C>A	17961070
CASP3-08	rs1049216	Ex8+567T>C	17962029
CASP3-07	rs1405938	IVS3-46A>G	17968558
CASP3-02	rs3087455	IVS2-1555A>C	17973117
KRAS-12	rs1137196	Ex6-790T>G	18117943
KRAS-05	rs13096	Ex6-1662T>C	18118815
KRAS-22	rs9266	Ex6+629A>G	18121191
KRAS-08	rs712	Ex6+294A>C	18121526
KRAS-04	rs17473423	Ex6+69A>G	18121751
ERCC2-03	rs28365048	Ex23+61A>C	18123137
ERCC2-09	rs1799787	IVS19-70C>T	18124362
KRAS-06	rs4246229	IVS5+702G>A	18126643
KRAS-19	rs6487461	IVS5+287T>C	18127058
KRAS-10	rs11047902	IVS3+375C>T	18138767
KRAS-17	rs17388148	IVS2-1840T>G	18141160
KRAS-15	rs17329025	IVS2-3467A>G	18142787
KRAS-07	rs4623993	IVS2-5082C>T	18144402
KRAS-20	rs7133640	IVS2-7970G>C	18147290
KRAS-16	rs17329424	IVS2+7144A>C	18150038
KRAS-21	rs7973746	IVS2+6969G>C	18150213
PPP1R13L-01	rs6966	Ex6-67A>T	18151180
KRAS-03	rs10505980	IVS2+5765C>T	18151417

KRAS-01	rs10842515	IVS2+4685T>C	18152497
KRAS-02	rs2970532	IVS2+2173C>T	18155009
KRAS-11	rs11047918	IVS2+1176G>A	18156006
KRAS-13	rs12226937	IVS2+506C>T	18156676
KRAS-14	rs12228277	IVS2+190T>A	18156992
KRAS-09	rs10842518	IVS1-1877G>T	18159180
KRAS-18	rs4368021	IVS1+1863T>C	18160796
ERCC1-30	rs3212986	196bp 3' of STP G>T	18180954
ERCC1-05	rs11615	Ex4+33A>G	18191871
ERCC1-06	rs3212948	IVS3+74C>G	18192580
CALCR-01	rs1801197	Ex13+149T>C	18286270
CALCR-03	rs2074122	IVS8+245C>A	18303190
BRIP1-05	rs4986763	Ex20+506T>C	18414057
BRIP1-02	rs4986764	Ex19-151T>C	18416408
BRIP1-03	rs4986765	Ex19+62A>G	18416526
BRIP1-09	rs1015771	IVS14+3238T>C	18503585
BRIP1-15	rs4988340	IVS1+12A>G	18593694
BRIP1-01	rs2048718	-1918G>A	18593880
IL4R-24	rs2057768	-29429C>T	18635174
IL4R-27	rs3024544	IVS3-85C>T	18666436
IL4R-02	rs1805011	Ex10+300A>C	18686951
IL4R-03	rs1805012	Ex11+392T>C	18687043
IL4R-05	rs1805015	Ex10+608T>C	18687259
IL4R-07	rs1805016	Ex10-1169T>G	18688006
IL4R-10	rs8832	Ex10-309A>G	18688866
SOD1-01	rs2070424	IVS3-251A>G	18701191
FOXA1-41	E3074_384	N/A	19066999
CDK7-01	rs2972388	Ex2-28C>T	19125611
RPA4-01	rs2642219	Ex1+500G>A	19435714
DRD1-02	rs5326	IVS2-90A>G	19679782
LEPR-08	rs1887285	IVS2+6686G>A	19717140
LEPR-03	rs7602	IVS2+6890A>G	19717344
LEPR-01	rs1137100	Ex4-45A>G	19855834
LEPR-04	rs1137101	Ex6-36A>G	19877906
SULT1A2-09	rs3194168	336bp 3' of STP T>C	19916091
AURKA-08	Poly-0014870	800bp 3' of STP G>C	19997321
AURKA-15	rs8173	Ex11-347G>C	19997699
AURKA-16	rs10485805	IVS9-68T>C	19998691
AURKA-06	rs6024840	IVS7-80T>C	20009615
ABCB11-08	rs853785	IVS19-1123A>G	20012011
AURKA-04	rs2298016	IVS6+30G>C	20012204
AURKA-02	rs1047972	Ex5+127A>G	20014371
AURKA-03	rs2273535	Ex5+49T>A	20014449
CSTF1-21	rs16979877	IVS1+269G>A	20020946
CSTF1-22	rs6064387	IVS1+390A>G	20021067
CSTF1-10	rs6099129	IVS1+870G>T	20021547
CSTF1-08	rs6064389	IVS1+966G>T	20021643
ABCB11-02	rs3770603	IVS1+4517G>A	20092635
SELE-01	rs5361	Ex4+24A>C	20110000
APOA4-07	rs5100	IVS2-97T>C	20255110

APOA4-02	rs5092	Ex2+38G>A	20255880
IFNAR2-01	rs3153	IVS1-4640G>A	20271375
IFNAR2-06	rs7279064	Ex2-28T>G	20276125
IFNAR2-10	rs2236757	IVS6-50A>G	20286787
CDK4-01	rs2072052	-1218T>G	20290025
ABCC2-01	rs717620	Ex1+8C>T	20291104
METTL1-01	rs703842	Ex7+196C>T	20306045
ABCC2-02	rs2273697	Ex10+40G>A	20312341
ABCC2-03	rs3740074	IVS15+169T>C	20320054
ABCC2-10	E3510_102	IVS27-73A>G	20352532
IFNGR2-03	rs1059293	Ex7-128C>T	20471563
LIG1-02	rs13436	Ex26+3G>C	20889226
LIG1-18	rs3729512	IVS25+19A>G	20890565
LIG1-29	rs156641	IVS19-131A>G	20899598
LIG1-01	rs20580	Ex7+44C>A	20922743
LIG1-03	rs20579	Ex2-24C>T	20937020
HIF1AN-02	rs2295780	IVS5+159A>G	21054491
HSD17B4-15	rs2451818	-27855G>T	21175428
HSD17B4-19	rs384346	-18796A>T	21184487
HSD17B4-01	rs28943585	-2124A>T	21201159
HSD17B4-21	rs7737181	IVS8+4959C>G	21234688
HSD17B4-10	rs2546210	IVS9-194C>T	21242614
HSD17B4-17	rs32659	IVS15+428A>G	21258025
HSD17B4-03	rs17145464	IVS22+74G>C	21282186
HSD17B4-18	rs3797372	IVS22-1666G>A	21285465
HSD17B4-08	rs28943596	Ex24-76A>G	21292962
HSD17B4-16	rs246965	13225bp 3' of STP A>G	21305928
IL10RA-08	rs2229114	Ex7+449C>T	21432294
IL10RA-02	rs9610	Ex7-109G>A	21434502
FUT2-05	rs603985	11bp 3' of STP C>T	21475447
LCAT-03	rs5923	Ex6-167C>T	21588152
LCAT-05	rs1109166	IVS1-267A>G	21591581
DHDH-02	rs4987162	IVS2+65C>G	21706623
DHDH-03	rs2270939	Ex4-26T>C	21711123
BAX-03	rs4645887	IVS4+286A>T	21728066
BAX-05	rs905238	490bp 3' of STP A>G	21733574
ROS1-20	rs498251	IVS37+85A>T	21808695
ROS1-18	rs497186	IVS36-4A>G	21808848
ROS1-03	rs581235	IVS32+504A>G	21819417
ROS1-12	rs574664	IVS32+361A>T	21819560
CYP2D6-65	rs2854741	IVS6-56C>G	21834300
ROS1-15	rs1998206	Ex5-6A>C	21894877
ROS1-14	rs2243377	IVS4-31C>T	21895051
ROS1-04	rs2243	IVS3+31C>T	21906819
LIG4-01	rs1805386	Ex2-1349T>C	21951589
CDKN2A-03	rs3088440	Ex4+83C>T	21958159
CDKN2A-09	rs2518719	IVS3+474T>C	21960427
CDKN2A-11	rs3731246	IVS2-682C>G	21961989
CDKN2A-14	rs2811708	IVS2+981C>A	21963422
CDKN2A-12	rs3731239	IVS2+185C>T	21964218

CDKN2A-13	rs2518720	IVS1-3882G>A	21968979
CDKN2A-20	rs3731217	IVS1+9477G>T	21974661
CDKN2A-19	rs3731211	IVS1+7291A>T	21976847
CDKN2A-18	rs3731198	IVS1+4661A>G	21979477
CDKN2A-16	rs3218020	-3418C>T	21987872
IL6-04	rs1800797	-660A>G	22162516
IL6-01	rs1800795	Ex2C>G	22162940
AXIN2-09	rs11868547	2490bp 3' of STP G>C	22252344
AXIN2-12	rs7210356	IVS9+1080A>G	22257691
AXIN2-14	rs4128941	IVS8+413G>A	22260072
AXIN2-11	rs4541111	IVS3-77C>A	22263279
AXIN2-13	rs11867417	IVS2-223T>C	22266639
AXIN2-10	rs3923087	IVS1-3483T>C	22278002
AXIN2-03	rs2240308	Ex1+237G>A	22283332
CYP19A1-08	rs4646	Ex11+410G>T	22293401
CYP19A1-09	rs10046	Ex10+268C>T	22293543
CYP19A1-06	rs1065779	IVS8-53T>G	22295368
CYP19A1-04	rs2304463	IVS7-106T>G	22298677
CYP19A1-01	rs700518	Ex4-57A>G	22319669
CYP19A1-34	rs2414096	IVS3-573T>C	22320336
CYP19A1-40	rs727479	IVS3+418G>T	22325104
CYP19A1-14	rs767199	IVS2-5240T>C	22330944
CYP19A1-29	rs12907866	IVS2-10307T>C	22336011
CYP19A1-39	rs6493494	IVS2-14688T>C	22340392
CYP19A1-41	rs749292	IVS2-23584T>C	22349288
CYP19A1-16	rs730154	IVS2+24809A>G	22381761
CYP19A1-30	rs28566535	IVS2+14872T>G	22391698
LTA-05	rs3093546	Ex1+50A>G	22398393
LTA-01	rs909253	IVS1+90G>A	22398564
TNF-12	rs1799964	-1210C>T	22400559
TNF-09	rs1800630	-1042A>C	22400727
TNF-02	rs1800629	-487A>G	22401282
TNF-13	rs3093661	IVS1+54G>A	22402009
CYP19A1-15	rs1004984	IVS2+2484C>T	22404086
CYP19A1-27	rs1004983	IVS2+2361G>T	22404209
CYP19A1-38	rs2470144	-86615T>C	22412282
CDH1-06	rs9282650	IVS2-25933A>T	22423839
CYP19A1-36	rs2445765	-99788G>C	22425455
IRF3-02	rs7251	Ex8-81G>C	22431099
IRF3-12	rs2304206	IVS1+17C>T	22437061
IRF3-01	rs2304204	-924A>G	22437210
CYP19A1-37	rs2446405	-111683A>T	22437350
CDH1-09	rs1801026	Ex16+264A>C>G>T	22481655
APAF1-03	rs2278361	IVS3-58G>A	22525398
APAF1-04	rs2288729	IVS12+2093G>A	22549781
APAF1-07	rs1007573	IVS16-565G>A	22574812
APAF1-09	rs1866477	IVS25+515T>G	22602024
ABCG8-06	rs9282575	Ex5-20G>A	22895539
ABCG8-01	rs9282572	IVS5+46C>T	22895604
ABCG8-02	rs6544718	Ex13+11T>C	22920858

BZRP-09	rs3937387	IVS1-22C>G	22945707
BZRP-03	rs113515	IVS2-136C>G	22947435
BZRP-05	rs6971	Ex4+118A>G	22949439
TERF2-14	rs251796	IVS7-42T>C	23009633
TERF2-01	rs153045	IVS7-2001A>G	23011592
TERF2-03	E3673_301	IVS6+27G>A	23016451
FASLG-01	rs929087	IVS2-1417A>G	23040996
CBR1-01	rs25678	Ex1-71G>C	23104502
CBR1-10	rs1005695	IVS2+210G>C	23105435
CBR1-11	rs2156406	IVS2+316A>G	23105541
CBR3-01	rs881712	Ex1-11C>T	23169639
POLD1-13	rs1726787	IVS2+21C>T	23170521
CYP17A1-13	rs619824	9170bp 3' of STP G>T	23329814
CYP17A1-08	rs10883782	6526bp 3' of STP T>C	23332458
CYP17A1-11	rs4919682	6128bp 3' of STP G>A	23332856
CYP17A1-10	rs284849	IVS7+83C>A	23339708
CYP17A1-12	rs4919687	IVS1-99T>C	23343774
CYP17A1-01	rs743572	Ex1+27T>C	23345678
NQO1-15	rs10517	Ex6-452T>C	23357959
NQO1-08	rs689453	Ex2+65G>A	23366572
NQO1-07	rs689452	IVS1-27C>G	23366663
MYO5A-01	rs1058219	Ex29-114C>T	23434121
MYO5A-06	rs2290336	IVS20-78G>A	23458292
MYO5A-07	rs2242058	IVS19+38G>A	23462341
ALAD-03	rs1805313	IVS11+66C>T	23472395
ALAD-10	rs8177806	Ex6+17C>T	23474144
ALAD-01	rs1139488	Ex4+4T>C	23475104
POT1-37	E5058_689	22999bp 3' of STP T>C	23862828
POT1-18	rs1034794	22614bp 3' of STP A>T	23863213
POT1-02	rs727506	1988bp 3' of STP C>T	23883839
POT1-11	rs10250202	IVS13-98T>G	23887321
POT1-10	rs10244817	IVS12-111G>A	23889286
POT1-09	rs10263573	IVS12+41T>A	23891083
POT1-07	rs7784168	IVS6-33G>A	23913853
POT1-05	rs6959712	IVS5+8T>A	23920819
POT1-03	E5047_301	-1386G>A	23960442
LOC401398-01	rs6466966	IVS1-7C>T	24002124
RXRB-11	rs2072915	Ex10+525T>A	24020332
RXRB-02	rs2076310	IVS3+51C>T	24024284
BAK1-05	rs210135	Ex6-364T>A	24398942
BAK1-06	rs513349	IVS5-35T>C	24399969
NR1H4-18	E3706_375	IVS4-3518G>A	24404909
BAK1-07	rs210145	IVS1+362G>C	24405690
NR1H4-05	rs35724	IVS9-285G>C	24437569
CYP3A7-01	rs12360	Ex13+125C>T	24565711
CYP3A4-57	Poly-0014748	-17677G>A	24653205
CTH-01	rs663465	-340A>G	24696151
CTH-07	rs6413471	IVS3-66A>C	24706576
CTH-10	rs473334	IVS7-799A>G	24716360
CTH-03	rs663649	IVS7-583G>T	24716576

CTH-14	rs559062	IVS10-430C>T	24723334
CTH-13	rs515064	IVS10-303A>G	24723461
ZNF350-08	rs2278414	Ex5-229T>C	24736012
ZNF350-04	rs2278415	Ex5-610A>T	24736393
ZNF350-06	rs4988334	Ex5+470T>C	24737188
ADH1C-01	rs698	Ex8-56A>G	24755493
ADH1C-16	rs2009181	IVS6-680T>C	24757251
ADH1C-15	rs283411	IVS5+62G>T	24760661
ADH1C-18	rs17526590	IVS1-42C>T	24763749
HMGCR-01	rs2241402	IVS8+56T>A	25240613
HMGCR-02	rs2303151	IVS18+70T>C	25249809
CGA-03	rs4986869	292bp 3' of STP A>G	25615352
CGA-02	rs6631	Ex4-38T>A	25615430
CGA-05	rs6155	Ex2+22A>G	25618075
CGA-06	rs932742	IVS1+46A>G	25624856
TERF1-27	rs10106086	-27187A>G	25747287
CD80-01	rs2228017	Ex3+35G>A	25758826
CD80-04	rs9282638	IVS2-56G>A	25758916
CD80-02	rs1385520	IVS2+851C>T	25770771
TERF1-02	E3663_301	IVS7+82C>T	25796065
TERF1-04	rs2306494	IVS8-124G>A	25804580
TERF1-01	rs2306492	IVS9+448G>A	25805255
TERF1-06	rs3863242	IVS9-163C>T	25811386
GSK3B-37	rs3732361	3337bp 3' of STP A>G	26037443
GSK3B-09	rs2873950	IVS11-1360T>G	26042208
GSK3B-22	rs10934500	IVS10-5923T>C	26063269
GSK3B-14	rs4624596	IVS10-9341G>A	26066687
GSK3B-05	rs1719889	IVS10+3478A>T	26073934
GSK3B-04	rs1719888	IVS10+3386A>G	26074026
GSK3B-08	rs1732170	IVS9-1224A>G	26078822
GSK3B-07	rs1719895	IVS7-148C>T	26090649
GSK3B-35	rs2319398	IVS7+11660C>A	26108088
GSK3B-18	rs7617372	IVS7+5272T>C	26114476
GSK3B-25	rs1574154	IVS6-2548T>C	26122393
GSK3B-38	rs4072520	IVS4-372C>A	26130539
GSK3B-20	rs6438553	IVS4+2191T>A	26135175
GSK3B-41	rs7620750	IVS3-851G>A	26138327
GSK3B-43	rs9873477	IVS3-3646A>G	26141122
GSK3B-19	rs9878473	IVS3-8458G>A	26145934
GSK3B-15	rs4688046	IVS3+2245G>A	26159016
GSK3B-31	rs17810235	IVS2-8853C>T	26170197
GSK3B-03	rs1381841	IVS2-12604T>C	26173948
GSK3B-39	rs4688047	IVS2-16563C>T	26177907
GSK3B-32	rs17810302	IVS2-19279A>G	26180623
GSK3B-23	rs10934503	IVS2-24969T>C	26186313
GSK3B-29	rs17204605	IVS1-12578G>A	26228810
GSK3B-01	rs1154597	IVS1-15747T>C	26231979
GSK3B-40	rs6770314	IVS1-21635G>A	26237867
GSK3B-42	rs9851174	IVS1-28325G>A	26244557
GSK3B-34	rs1870931	IVS1-35927C>G	26252159

GSK3B-21	rs6781942	IVS1-36963T>C	26253195
GSK3B-27	rs16830683	IVS1-39082C>T	26255314
GSK3B-02	rs12630592	IVS1+43948C>A	26263392
GSK3B-28	rs16830689	IVS1+40923G>C	26266417
GSK3B-17	rs6779828	IVS1+37047G>A	26270293
GSK3B-30	rs17204878	IVS1+35314G>T	26272026
IGF1-24	rs5742714	Ex4-177C>G	26272042
GSK3B-33	rs17810676	IVS1+31645T>C	26275695
IGF1-22	rs5742694	IVS3-2892C>A	26281426
IGF1-27	rs978458	IVS3-5895A>G	26284429
GSK3B-36	rs334535	IVS1+19890C>T	26287450
IGF1-16	rs4764883	IVS3+6982G>A	26288495
GSK3B-11	rs334555	IVS1+8058G>C	26299282
GSK3B-12	rs334559	IVS1+2589T>C	26304751
IGF1-44	rs5742667	IVS2-10010C>T	26305668
IGF1-46	rs5742665	IVS2-10082C>G	26305740
IGF1-15	rs2373721	IVS2-13577G>C	26309235
GSK3B-45	rs3755557	N/A	26310103
IGF1-11	rs5742629	IVS2+12158A>G	26339453
IGF1-04	rs2162679	IVS1-1682A>G	26353449
MSH2-08	rs1863332	-432T>G	26445831
MSH2-15	rs4952887	IVS6+3400C>T	26462901
MSH2-06	rs17036577	IVS7-5849T>C	26482771
MSH2-21	rs7607076	IVS7-1122A>G	26487498
MSH2-09	rs1981928	IVS7-212T>A	26488408
MSH2-19	rs7602094	IVS8+719T>C	26489448
MSH2-18	rs7585925	IVS8+1488T>G	26490217
MSH2-12	rs3771281	IVS9-1516C>T	26508214
MSH2-20	rs17036614	IVS11+501A>G	26514635
MSH2-13	rs3821227	IVS11-1207C>T	26516890
MSH2-03	rs2303428	IVS12-6C>T	26519433
MSH2-03	rs2303428	IVS12-6C>T	26519433
MSH2-14	rs4608577	IVS13+274T>G	26519917
MSH2-10	rs2042649	IVS15-214T>C	26525637
MSH2-16	rs6544991	2691bp 3' of STP A>C	26528713
MSH6-01	rs3136228	-556G>T	26825749
MSH6-04	rs1800935	Ex3+83T>C	26839048
LEP-01	rs2167270	Ex1-11A>G	27296893
NFKB1-01	rs3774932	IVS1+1246A>G	27918904
NFKB1-02	rs3774937	IVS1+11306C>T	27928964
NFKB1-33	rs230532	IVS2-826T>A	27944878
NFKB1-09	rs230496	IVS6+199A>G	27983208
PIM1-03	rs262933	-3975A>G	27992626
PIM1-17	rs1757000	-3185A>G	27993416
PIM1-25	rs12197850	Ex6+253C>A	28000212
PIM1-01	rs10507	Ex6+713C>T	28000672
NFKB1-21	rs4648059	IVS12-452C>G	28010316
NFKB1-14	rs230547	IVS23-1330T>C	28030988
CD86-03	rs9282641	Ex2+19A>G	28291914
CD86-02	rs1129055	Ex8+35G>A	28333465

CASR-11	rs4678045	IVS1+20204A>G	28418009
CASR-05	rs1965357	IVS1-4243C>T	28463698
MX1-04	rs458582	IVS5+404G>T	28466376
MX1-28	rs455599	IVS5+577A>G	28466549
CASR-15	rs3749208	IVS3-91C>T	28475430
MX1-07	rs469270	IVS11-198G>A	28479047
MX1-08	rs469390	Ex13+4G>A	28479800
MX1-22	rs2070229	Ex14+50T>C	28482983
MX1-03	rs2280807	IVS14+43A>G	28483135
MX1-10	rs2072683	IVS15-99C>T	28486319
MX1-01	rs1050008	Ex16+114A>G	28486531
MX1-11	rs469304	Ex16-64G>A	28486603
CASR-07	rs2279802	IVS5+52A>G	28490087
CASR-06	rs2270916	IVS6+16C>T	28496245
CASR-09	rs2270917	IVS6+163C>T	28496392
CASR-01	rs1042636	Ex7+1236A>G	28498915
BHMT-02	rs567754	IVS4+52C>T	29010774
BHMT-01	rs585800	Ex8+453A>T	29021566
BHMT-04	rs617219	2654bp 3' of STP A>C	29023952
CHEK1-01	rs558351	-1399T>C	29057680
CHEK1-03	rs491528	IVS2-36G>T	29059882
CHEK1-02	rs506504	Ex13+76A>G	29087611
LIPC-17	rs1077834	-752C>T	29514036
LIPC-01	rs1800588	-556C>T	29514232
LIPC-02	rs3825776	IVS1+22511T>C	29537387
MEST-03	rs2072574	IVS5-85A>G	29555519
LIPC-25	rs1869145	IVS1-33033C>T	29588056
LIPC-04	rs1968687	IVS1-7835G>T	29613254
LIPC-37	rs1968689	IVS1-7747C>T	29613342
LIPC-06	rs6083	Ex5+70A>G	29628567
LIPC-08	rs2242064	IVS5+1098G>T	29629829
LIPC-23	rs2242066	IVS5+1163A>G	29629894
LIPC-09	rs6074	Ex9+49C>A	29651520
DHFR-07	E5043_337	IVS3+2979C>G	30536587
DHFR-11	rs865646	IVS3+2851A>C	30536715
DHFR-18	rs1650697	Ex1-3G>A	30545139
MSH3-02	rs1805355	Ex4-100A>G	30560387
MSH3-29	rs1677649	IVS4+69A>G	30560555
MSH3-03	rs836802	IVS8+8888G>C	30578158
MSH3-07	rs3797896	IVS19+5137C>G	30688158
IFNG-07	rs1861494	IVS3+284G>A	30694715
MSH3-12	rs32983	IVS20+10801C>T	30714719
MSH3-09	rs26279	Ex23+3A>G	30763295
MDM2-01	rs769412	Ex12+162A>G	31376521
ALDH1L1-06	rs9282690	IVS21+46G>A	32321092
ALDH1L1-03	rs1127717	Ex21+31A>G	32321213
ALDH1L1-01	rs2305230	Ex10-40G>T	32351849
PTGS1-02	rs5788	Ex6-40C>A	32464996
LMO2-08	rs3740616	Ex6+226T>A	32668137
LMO2-01	rs3740617	Ex6+106A>G	32668257

LMO2-04	rs3781577	IVS1-3051T>C	32693725
CCND3-01	rs9529	Ex5-337A>G	32761257
CCND3-02	rs2479717	IVS2-42T>A	32763424
RNASEL-01	rs11072	Ex6-560A>G	32952270
RNASEL-02	rs486907	Ex1-96G>A	32963496
XRCC4-05	rs2075685	Ex2G>T	32967023
XRCC4-07	rs2662238	IVS4-64A>G	33093665
MBD2-01	rs7614	Ex8+438A>G	33170346
MBD2-02	rs1145315	IVS6+1938A>G	33178057
MBD2-03	rs609791	IVS3-3743G>C	33208228
XRCC4-10	rs2891980	IVS7-6281C>T	33237021
MBD2-04	rs603097	-2176C>T	33242208
XRCC4-04	rs3777015	IVS7-61G>A	33243241
XRCC4-01	rs1805377	IVS7-1A>G	33243301
CAT-07	rs9282626	-1042C>T	33246759
CAT-02	rs769214	-843G>A	33246958
CAT-15	rs1049982	Ex1+49T>C	33247782
CAT-05	rs769218	IVS1-60A>G	33257920
CAT-03	rs769217	Ex9-29C>T	33270149
CAT-06	rs475043	820bp 3' of STP C>T	33281042
IL3-01	rs40401	Ex1-84C>T	33811491
CSF2-02	rs25882	Ex4+23T>C	33826473
NCF2-05	rs699244	IVS15-87C>A	33934391
NCF2-04	rs2296164	IVS10-21C>T	33943874
NCF2-03	rs2274064	Ex6+41T>C	33951326
AMACR-01	rs2278008	Ex5+90G>A	33962275
AMACR-08	rs6863657	IVS4+4012T>C	33967491
AMACR-09	rs840409	IVS4+3803C>G	33967700
AMACR-02	rs34677	Ex4-23G>T	33971525
AMACR-17	rs10941112	Ex3-29A>G	33977464
AMACR-03	rs34689	IVS1+169G>T	33980466
AMACR-05	rs3195676	Ex1+114A>G	33980857
IRF1-05	rs839	Ex10-347C>T	34234139
IRF1-03	rs9282763	IVS6-68G>A	34237146
IL13-02	rs1881457	-1469A>C	34407422
IL13-03	rs1800925	-1069C>T	34407822
IL13-06	rs1295686	IVS3-24T>C	34410856
IL13-01	rs20541	Ex4+98A>G	34410977
IL4-02	rs2243248	Ex2T>G	34423657
IL4-01	rs2243250	-588C>T	34424167
IL4-03	rs2070874	Ex1-168C>T	34424723
IL4-11	rs2243268	IVS2-1443A>C	34428976
IL4-10	rs2243290	IVS3-9A>C	34433182
VEGF-19	rs1005230	-2487C>T	34594746
VEGF-05	rs25648	Ex1-73C>T	34597227
VEGF-04	rs3025039	236bp 3' of STP C>T	34610786
NFKBIE-03	rs2282151	8321bp 3' of STP A>G	35084445
NFKBIE-02	rs730775	IVS1+645T>C	35090324
NFKBIE-08	rs513688	IVS1-2163C>A	35093940
NFKBIE-01	rs483536	-14107A>T	35094103

EGF-08	rs4444903	Ex1+61A>G	35329240
RAG1-01	rs2227973	Ex2+2473A>G	35384554
EGF-02	rs2237051	Ex14+71G>A	35396328
EGF-04	rs971696	IVS22-1443T>A	35422995
MBD4-02	rs140696	Ex6+2C>T	35647243
HSD17B2-02	rs723012	IVS3-5735C>T	35732971
HSD17B2-01	rs1424151	IVS4-2328A>G	35743551
IL7R-01	rs1494555	Ex4+33G>A	35843947
IL7R-08	rs7737000	Ex4-43C>T	35844030
ENPP1-04	rs1044582	Ex25-243A>T	36316537
MLH1-02	rs1799977	Ex8-23A>G	36993572
MLH1-05	rs2286940	IVS12-169C>T	37010110
PTGS2-33	rs5275	Ex10+837T>C	37051997
PTGS2-44	rs4648276	IVS7+111C>T	37054427
PTGS2-19	rs5277	Ex3-8G>C	37057136
PTGS2-08	rs20417	-898G>C	37059260
PTGS2-05	rs689466	Ex2A>G	37059690
CCNH-01	rs2266690	Ex8+49T>C	37289632
CCNH-04	rs3093816	IVS7+132C>T	37291745
ENG-06	rs1330684	IVS12-117A>G	37900803
CX3CR1-02	rs3732378	Ex2+848C>T	39247166
CX3CR1-01	rs3732379	Ex2+754G>A	39247260
CDC25C-01	rs1042124	Ex1-62G>T	40082358
PMS1-56	rs5742926	Ex2G>T	40858221
PMS1-57	rs5742938	IVS1+639G>A	40859374
PMS1-49	rs1233299	IVS3+3961A>C	40874054
PMS1-15	rs1233302	IVS3-1498C>A	40878296
PMS1-24	rs5743030	IVS4-4198G>A	40887961
PMS1-48	rs1233284	IVS5+6865G>A	40899187
PMS1-27	rs1233288	IVS5+7819C>T	40900141
PMS1-63	rs1233291	IVS5+8045G>C	40900367
PMS1-28	rs1233297	IVS5+11269C>T	40903591
PMS1-60	rs5743072	IVS5-11766A>G	40906340
PMS1-47	rs1233255	IVS5-2598A>C	40915508
PMS1-26	rs1233258	IVS5-1656C>T	40916450
PMS1-50	rs12618262	IVS5-617C>T	40917489
PMS1-61	rs5743112	IVS6+176C>A	40918398
PMS1-62	rs5743116	IVS6-3413T>C	40923384
PMS1-54	rs256567	IVS9-938C>T	40936947
PMS1-31	rs256564	IVS10+1095A>G	40939465
PMS1-53	rs256563	IVS10-693A>G	40941248
PMS1-52	rs256552	211bp 3' of STP A>G	40951790
PMS1-51	rs256550	2575bp 3' of STP T>C	40954154
MATR3-01	rs11738738	3101bp 3' of STP A>T	41083199
SLC23A1-20	rs6596471	IVS14+2088T>C	41120601
SLC23A1-09	rs4257763	IVS10+109T>C	41129172
SLC23A1-05	E3359_310	Ex8+22G>A	41130515
SLC23A1-18	rs10063949	-583G>A	41134539
SEP15-04	rs540049	Ex5-176C>T	41147701
SEP15-02	rs5845	Ex5+450T>C	41148232

DNAJC18-01	rs4315920	4151bp 3' of STP T>C	41160698
CTNNB1-11	rs3864004	-25382A>G	41180181
CTNNB1-01	rs11564437	IVS1+832A>G	41181997
CTNNB1-14	rs4533622	IVS1+1177A>C	41182342
CTNNB1-07	rs2371452	IVS1+1702A>G	41182867
CTNNB1-05	rs1798794	IVS1-10154G>T	41195362
CTNNB1-16	rs9813198	IVS1-7464A>G	41198052
CTNNB1-15	rs5743395	IVS7+309A>T	41209156
CTNNB1-19	rs1880481	IVS7-2751C>A	41212085
CTNNB1-02	rs11564452	IVS7-562A>T	41214274
CTNNB1-03	rs11564465	IVS10-175C>T	41217044
CTNNB1-13	rs4135385	IVS13-67A>G	41219444
CTNNB1-08	rs2953	Ex15-547G>T	41221392
CTNNB1-17	rs11129895	2548bp 3' of STP A>G	41223386
CTNNB1-21	rs9883073	3702bp 3' of STP C>A	41224540
IFNGR1-01	rs11914	Ex7+189T>G	41624017
IFNGR1-05	rs3799488	IVS6-4G>A	41624209
MYC-02	rs3891248	IVS1-355T>A	41968318
STAT1-01	rs2066804	IVS21-8C>T	42051175
GHR-31	rs2972395	-165670C>T	42373063
GHR-28	rs2940930	-160465G>A	42378268
GHR-16	rs7732059	-142504C>G	42396229
GHR-11	Poly-0009029	-142290T>C	42396443
GHR-11	Poly-0009029	-142290T>C	42396443
GHR-214	rs1858136	N/A	42408866
GHR-33	rs2972418	-82283T>C	42456450
GHR-79	rs2972419	Ex2G>A	42456634
GHR-29	rs2940944	IVS1+65085C>A	42461899
GHR-30	rs2972392	Ex2A>C	42468592
GHR-27	rs2940913	-66359G>T	42472374
GHR-50	rs7735889	IVS1-3767A>G	42534956
GHR-47	rs7712701	IVS2+4144A>C	42542947
GHR-21	rs28943882	IVS2+29065C>T	42567868
GHR-45	rs6873545	IVS3+2059C>T	42604021
GHR-90	rs28943889	IVS2+16453T>C	42618415
GHR-77	rs6878512	IVS3-21121C>T	42640628
GHR-46	rs6897530	IVS3-21055C>T	42640694
GHR-01	rs6179	Ex6-61A>G	42672801
GHR-34	rs2972780	IVS8+1229C>T	42687607
GHR-03	rs6180	Ex10+685A>C	42691996
SEPP1-02	rs6413428	Ex5+710T>C	42773481
SEPP1-01	rs7579	Ex5+626C>T	42773565
GSTA4-01	rs405729	Ex7-31A>G	43701012
GSTA4-02	rs367836	Ex7+260C>A	43701362
GSTA4-07	rs543613	IVS6-134A>G	43701755
GSTA4-04	rs4986947	IVS5+32C>T	43707461
IGFBP1-01	rs4619	Ex4+111A>G	45304115
IGFBP3-04	rs2471551	IVS2-17G>C	45328495
ESR2-02	rs4986938	38bp 3' of STP C>T	45699569
ESR2-05	rs3020450	-18598A>G	45768055

CYP1A1-78	rs2198843	11599bp 3' of STP C>G	45791548
CYP1A1-15	rs4646421	IVS1-728C>T	45806510
CYP1A1-14	rs2606345	IVS1+606T>G	45807494
CYP1A1-91	rs17861115	-9893G>A	45815650
CYP1A1-81	rs2472299	-17961T>C	45823718
CCR3-01	rs4987053	Ex3+62T>C	46246704
CCR3-05	rs3091312	754bp 3' of STP A>T	46248476
GPX2-07	Poly-0014684	2680bp 3' of STP T>A	46403278
GPX2-19	rs4902345	2301bp 3' of STP G>A	46403657
GPX2-13	rs10133054	2089bp 3' of STP G>C	46403869
GPX2-14	rs10133290	1763bp 3' of STP T>G	46404195
GPX2-09	rs17880380	1306bp 3' of STP G>A	46404652
GPX2-16	rs12172810	823bp 3' of STP G>A	46405135
GPX2-17	rs2071566	IVS1-444C>T	46406753
GPX2-21	rs2737844	IVS1+714C>T	46408262
GPX2-02	rs1800669	IVS1+19T>A	46408957
GPX2-18	rs2296327	-6793G>A	46410300
RAB15-04	rs3825644	3715bp 3' of STP T>G	46411252
RAB15-03	rs3742599	3306bp 3' of STP C>A	46411661
RAB15-02	rs2277502	2936bp 3' of STP A>G	46412031
CFH-01	rs800292	Ex2-61G>A	47051172
CFH-06	rs1329423	IVS4-219T>C	47055326
CFH-07	rs2300430	IVS7+1346T>C	47064652
CFH-03	rs2274700	Ex10+83G>A	47091886
CFH-05	rs1065489	Ex18+26G>T	47118713
CCNA2-12	rs3217773	IVS7+78C>T	47234252
CCNA2-06	rs1396080	IVS5+73A>C	47235585
CCNA2-01	rs769242	Ex3+30A>G	47237348
HUS1-01	rs1056663	Ex8+74G>A	47376947
HUS1-05	rs2242478	IVS3+25G>A	47389981
IL2-03	rs2069763	Ex2-34G>T	47872613
IL2-01	rs2069762	Ex2T>G	47873111
CDC25A-04	rs936426	IVS9+521T>C	48155257
GPX1-06	rs1800668	Ex1+35C>T	49335761
GPX1-28	rs3448	-39303A>G	49336755
TCTA-04	rs6784820	IVS2+321A>G	49390868
TCTA-02	rs6997	Ex3-75T>C	49393838
NICN1-01	rs8897	Ex7-28G>A	49400411
CTSH-01	rs3129	Ex12-109G>A	50004535
NOS3-34	rs3918226	IVS1-665C>T	50051985
NOS3-01	rs1799983	Ex8-63G>T	50058257
CDK5-08	rs2069456	IVS7+11A>C	50114701
CDK5-16	rs1549760	-903G>A	50117932
SLC4A2-01	rs6464120	IVS1+549A>G	50119491
SLC4A2-02	rs10245199	IVS1-530A>G	50120576
SLC4A2-04	rs13240966	IVS1-194C>G	50120912
CASP10-02	rs3900115	Ex3-171A>G	52260093
LMOD1-03	rs2820312	Ex2+623T>C	52278196
CASP8-06	rs2349070	IVS4-876A>C	52339724
CASP8-07	rs2293554	IVS5+73T>G	52341003

CASP8-22	rs1035142	1760bp 3' of STP G>T	52362494
FZD7-17	rs1207955	-2321G>A	53106465
FZD7-06	E7045_223	-2082G>A	53106704
FZD7-10	rs13034206	Ex1-1926C>T	53110651
FZD7-20	rs4673222	Ex1-1251G>A	53111326
FZD7-15	E7064_389	2569bp 3' of STP G>A	53113081
FZD7-16	rs12474408	2710bp 3' of STP A>G	53113222
MTHFD2-01	rs1667627	IVS1+3323T>C	53245129
RGS6-04	rs3784058	IVS1+12785G>A	53412557
RGS6-02	rs2238284	IVS1-11668C>A	53419574
RGS6-05	rs2238280	IVS1-5967T>C	53425275
EGFR-05	rs2017000	IVS21+96A>G	54635882
EGFR-03	rs1140475	Ex25+8C>T	54659689
EGFR-04	rs2293347	Ex27+36C>T	54662188
CTLA4-16	rs11571315	-1764T>C>G	54940317
CTLA4-19	rs4553808	-1660A>G	54940421
CTLA4-17	rs11571316	-1576G>A	54940505
CTLA4-10	rs11571317	-657C>T	54941424
CTLA4-25	rs5742909	-318C>T	54941763
CTLA4-01	rs231775	Ex1-61A>G	54942130
CTLA4-07	rs3087243	1383bp 3' of STP A>G	54948335
VCAM1-02	rs1041163	-1591T>C	55003218
VCAM1-38	rs2392221	IVS3-7C>T	55009566
VCAM1-05	rs3176879	Ex9+149G>A	55023220
ESR1-31	rs488133	Ex2T>C	56280294
ESR1-14	rs2071454	-2223G>T	56281674
ESR1-34	rs9340770	-945A>C	56282952
ESR1-01	rs2077647	Ex1+392T>C	56283927
ESR1-08	rs1801132	Ex4-122G>C	56420372
ESR1-17	rs2273206	IVS6+52G>T	56537161
ESR1-07	rs2228480	Ex8+229G>A	56574945
ESR1-13	rs3798577	Ex8+1264T>C	56575980
ESR1-30	rs3798758	Ex8+1988C>A	56576704
FOS-02	rs7101	Ex1+96C>T	56745379
FOS-06	rs1063169	IVS2-145G>T	56746871
FOS-08	rs4645856	IVS2-5C>T	56747011
RGS17-01	rs2295231	IVS1-170T>C	57520198
RGS17-03	rs3870366	IVS1+12492C>T	57594617
OPRM1-01	rs1799971	Ex1-173A>G	58515647
OPRM1-02	rs607759	IVS1+11468C>T	58527287
OPRM1-23	rs9282821	IVS3+1768A>C	58569225
OPRM1-03	rs562859	IVS3+1966T>C	58569423
GSTZ1-02	rs7972	Ex5-12G>A	58792990
GSTZ1-03	rs1046428	Ex7+29T>C	58794036
BARD1-18	rs5031011	IVS6+14T>C	65841608
BARD1-04	rs2070094	Ex6-50G>A	65841671
BARD1-11	rs2229571	Ex4-181G>C	65854880
BARD1-22	rs2070096	Ex4-262G>C	65854961
BARD1-02	rs1129804	Ex1+44C>G	65883738
IL12A-09	rs582537	IVS2-701A>C	66205256

IL15-06	rs1493013	Ex3+163C>T	67135593
IL15-01	rs2254514	Ex3-92T>C	67135669
IL15-07	rs2857261	IVS3+8A>G	67135768
IL15-10	rs1057972	Ex9-181A>T	67149563
IL15-02	rs10833	Ex9-66T>C	67149678
XRCC5-14	rs828910	IVS2-711A>G	67186444
XRCC5-17	rs828702	IVS9+1081A>G	67202894
XRCC5-19	rs207916	IVS17+789A>G	67236976
XRCC5-02	rs1051685	Ex22+466A>G	67279792
XRCC5-12	rs2440	Ex22-238G>A	67280182
IGFBP2-26	rs1106037	-13556C>T	67694106
IGFBP2-29	rs2372848	IVS1+5424A>G	67713528
IGFBP2-25	rs2270360	IVS1-294A>C	67734402
IGFBP5-05	rs2241193	IVS1+4949A>G	67763629
IGFBP5-10	rs1978346	-1968C>T	67770883
IL8RA-04	rs2854386	5661bp 3' of STP C>G	69236918
MYNN-01	rs1317082	IVS4+236A>G	75992743
FBXW7-01	rs2676330	IVS4+2575A>G	77760638
FBXW7-05	rs2714804	IVS3+230T>C	77766095
FBXW7-44	rs2676329	IVS1-1417A>G	77770526
FBXW7-04	rs2714805	IVS1-20897G>A	77790006
FBXW7-02	rs2292743	-144T>A	77828231
IRS1-04	rs1366757	IVS1+12345G>C	77856775
IRS1-08	rs9282766	IVS1+4357G>A	77864763
IRS1-03	rs1801278	Ex1-840G>A	77869959
TLR2-06	rs4696480	-16933A>T	79102257
TLR2-04	rs3804099	Ex2+613T>C	79119787
TLR2-05	rs3804100	Ex2-1122T>C	79120540
XRCC3-03	rs1799796	IVS7-14A>G	85165680
XRCC3-04	rs1799794	Ex2+2A>G	85179020
AKT1-15	rs2498799	Ex10+24G>A	86240939
MASP1-21	rs3733001	IVS15-34G>A	93434114
MASP1-42	rs1001073	IVS12-397T>C	93439863
MASP1-01	rs3774268	Ex11+32C>T	93449482
MASP1-48	rs696405	IVS10-1868A>C	93451381
MASP1-53	rs710459	IVS9+790C>T	93455640
MASP1-50	rs698090	IVS8-2891G>A	93459458
MASP1-45	rs3105782	IVS5-193T>C	93466454
MASP1-44	rs1533593	IVS5-1224A>G	93467485
MASP1-46	rs3864099	IVS2+4675A>C	93494096
MASP1-49	rs698079	IVS2+3841T>G	93494930
MASP1-47	rs4376034	IVS2+3257T>C	93495514
MASP1-52	rs698105	IVS2+118A>G	93498653
MASP1-43	rs13094773	IVS1-339A>G	93499341
MASP1-54	rs7609662	IVS1+2718G>A>T	93501856
MASP1-07	rs12635264	IVS1+2650G>A>C	93501924
MASP1-09	rs13089330	-849C>T	93505428
BCL6-07	rs1474326	IVS10+202G>T	93937685
BCL6-09	rs3774309	IVS7-511C>T	93940355
BCL6-06	rs1464645	IVS7-571G>A	93940415

BCL6-11	rs3774306	IVS7-643A>G	93940487
BCL6-05	rs3172469	IVS1+4110A>C	93954246
TP73L-03	rs17514215	IVS5+34T>G	96077399
TP73L-17	rs7653848	IVS7+121C>T	96081010
TP73L-15	rs6789961	IVS8-22A>G	96082249
TP73L-16	rs6790167	IVS9+79A>G	96082432
TP73L-13	rs9840360	IVS10+41A>G	96085983
TP73L-28	rs7613791	IVS10-4859C>T	96094482
TP73L-26	rs1345186	IVS10-23T>C	96099318
TP73L-52	E4057_169	Ex14+342C>T	96107494
TP73L-46	E4064_458	Ex14-559G>A	96109665
TP73L-47	E4065_308	Ex14-430C>T	96109794

Statistical analysis

Demographic characteristics including age, gender, and race were compared between the original and validation studies using Student's T-test (for age) and two-sample tests of proportions (for AE status, gender, and race). Allele frequencies were estimated from the total number of copies of individual alleles divided by the number of all alleles in the sample, and they were compared between the two studies using a two-sample test of proportions. Deviations in frequencies from Hardy-Weinburg Equilibrium were evaluated using the exact test described in Wigginton *et al*. [12].

In the original study, potential associations were tested between each of the 1442 SNPs passing quality control filters and the occurrence of adverse events using logistic regression. For each SNP in the initial sample, we recorded the odds ratio estimate and p-value of the likelihood ratio test for a univariate allelic logistic model. No correction for multiple comparisons was made in our initial sample, because we reserved the validation sample for the purpose of weeding out false-positives. In the validation sample, we tested only those SNPs having an AE-associated p-value ≤ 0.05 in the original sample. We considered a significant SNP association in the initial sample to have replicated if it met the following criteria in the validation sample: an odds ratio that consistently associated AE risk with the same genotypes and a p-value ≤ 0.10. The more liberal p-value criterion was chosen to maintain power in the face of the smaller size of the validation sample compared to that of the original sample. While this

approach may increase the chances of false-positive results, the trade-off in favor of power is appropriate given the exploratory scope of the current study.

Potential patterns of linkage disequilibrium (LD) between replicated SNPs on the same chromosome were assessed using Haploview [13]. Haplotypes were estimated for SNPs in high LD ($r^2 > 0.90$) using the iterative approach described in Lake *et al*. [14]. The resulting haplotypes were tested for association with AEs using univariate logistic models. Statistical analyses were performed using R version 2.2.1, Stata version 9, and Haploview version 3.32 [13,15-17].

Results

Demographic characteristics of subjects included in genetic analysis

In both studies, all participants were invited to donate genetic samples. In the original study, of the 148 participants enrolled, a total of 96 individuals consented for the genetic substudy. Of those 96 subjects with genetic data, 27 experienced adverse events relating to immunization. Since *systemic* AEs were the outcome of interest, of the 27 individuals experiencing an AE, those 11 reporting only a localized rash near the inoculation site were left out, and the other 69 reporting no AEs were used as controls. In the validation study, 102 total healthy adults were enrolled and 90 gave consent for genotyping. Of the 90 individuals with genetic data, 46 were vaccine-naïve and 44 were vaccine-

experienced. Of the naïve individuals, 24 experienced systemic AEs, and of the experienced individuals, only 10 suffered systemic AEs.

There was a difference in vaccination history status between the two studies, with the original study including only vaccine-naïve participants and the validation study including both naïve and experienced individuals. Pooling vaccination history status, there was a statistically significant difference in mean age between the two studies ($p < 0.001$). However, when only the vaccine-naïve individuals in the validation study are compared to the original study sample, the mean difference in age was only one year ($p = 0.15$), indicating that the inclusion of vaccine-experienced individuals accounts for the age differential. Because age stratification can have a profound effect on immune function (especially for the inflammatory responses thought to be important in AEs) [18-20], only the vaccine-naïve individuals in the validation study were considered in all subsequent analyses. Table 2 summarizes age, race, gender, and AE status decompositions of both studies. Table 2 also describes the results of the demographic comparisons between the original and validation studies—considering only vaccine-naïve subjects. As the table indicates, there was no statistical difference in age, gender, or race between the two vaccine-naïve study samples. In the original study, 40 (47%) individuals were male, 84 (99%) were Caucasian and 1 (1%) was Asian. In the validation study, 27 (59%) individuals were male, 44 (96%) were Caucasian, 1 (2%) was African American, and 1 (2%) was Asian.

Table 2. Summary of AE status, age, gender, and race for both studies. Only vaccine-naïve subjects are considered.

Dataset/Study	Genetic (SNP) Data			
	AE/nonAE	Age[a]	Gender (M/F)	Race (W/B/A)[b]
Original (*N* = 85)	16/69	23.2 (3.9)	40/45	84/0/1
Validation (*N* = 46)	24/22	24.2 (3.8)	27/19	44/1/1
P-value of Difference[c]	< 0.01	0.15	0.20	0.25

[a] Mean (Standard Deviation)
[b] W = "White", B = "Black", A = "Asian"
[c] Two-sided p-value for t-test (Age) or two-sample test of proportions (AE status, Gender, Race)

Genetic associations with adverse events

Table 3 lists all SNPs with an AE-associated p-value ≤ 0.05 in the original sample. The significant genetic association results from the original study that replicated in the validation study are listed in Table 4. Two SNPs in the IRF-1 gene, three SNPs in the IL-4 gene, and one SNP in the MTHFR gene met our significance criteria for association with the occurrence of systemic adverse events.

Table 3. List of all SNPs with an AE-associated p-value ≤ 0.05 in the original sample, organized according to location. SNP names are taken from http://snp500cancer.nci.nih.gov.

SNP Name	dbSNP ID (rs#)	SNP Region	SNP Location (Base Pair)
IGF1R-26	rs3743259	IVS5+311A>G	893012
SRA1-03	rs801459	NC_A>C	1096550
PIN1-21	rs889162	IVS3+2592T>C	1214718
SLC6A3-14	rs460700	IVS4+2610A>G	1419969
CDKN1C-09	rs431222	-1679G>A	1695640
EPHX1-01	rs2234922	Ex4+52A>G	2202600
GATA3-46	rs10905277	-250A>G	2460264
SLC39A2-07	rs2149666	IVS2-119G>T	2467996
TXNRD2-83	rs9306230	IVS1+1202T>C	3080172
MBL2-03	rs5030737	Ex1-34C>T	3082397
BLM-02	rs2238335	IVS1-253G>C	6255893
BLM-25	rs16944831	IVS16-479C>T	6306468
MTHFR-02	rs1801133	Ex5+79C>T	6393745
MPDU1-01	rs4227	Ex7-334G>T	7088525
SAT2-03	rs858520	Ex4-11G>A	7127620
TP53-14	rs1614984	21226bp 3' of STP C>T	7168801
GDF15-02	rs1059369	Ex2-136T>A	9759943
GGH-01	rs719235	-353C>A	15805034
AHR-19	rs7796976	Ex1+185A>G	16704367
CYP1B1-18	rs10175368	-5329G>A	17123794
TSG101-40	rs2279900	IVS9+18G>A	17290400
TSG101-07	rs12574333	IVS4+10C>A	17323456
TSG101-36	rs2292179	-182T>C	17335787
AURKA-02	rs1047972	Ex5+127A>G	20014371
HSD17B4-19	rs384346	-18796A>T	21184487
HSD17B4-21	rs7737181	IVS8+4959C>G	21234688
LTA-05	rs3093546	Ex1+50A>G	22398393
CDH1-06	rs9282650	IVS2-25933A>T	22423839
CTH-03	rs663649	IVS7-583G>T	24716576
NFKB1-14	rs230547	IVS23-1330T>C	28030988
CASR-06	rs2270916	IVS6+16C>T	28496245
IRF1-05	rs839	Ex10-347C>T	34234139
IRF1-03	rs9282763	IVS6-68G>A	34237146
IL4-03	rs2070874	Ex1-168C>T	34424723
IL4-11	rs2243268	IVS2-1443A>C	34428976
IL4-10	rs2243290	IVS3-9A>C	34433182
CFH-03	rs2274700	Ex10+83G>A	47091886
FZD7-20	rs4673222	Ex1-1251G>A	53111326

Table 4. Significant genetic associations consistent across both studies.

				Original Study		Validation Study	
Gene	**SNP** (rs#)	**SNP Location** (Base Pair)	**Chromosomal Location**	[a]**Odds Ratio** (95% C.I.)	[b]**p-value** (χ^2)	[a]**Odds Ratio** (95% C.I.)	[b]**p-value** (χ^2)
IRF-1	rs9282763	34237146	5q31.1	3.2 (1.1 - 9.8)	0.03	3.0 (1.1 - 8.3)	0.03
	rs839	34234139	5q31.1	3.2 (1.1 - 9.8)	0.03	3.0 (1.1 - 8.3)	0.03
IL-4	rs2070874	34424723	5q31.1	2.4 (1.0 - 5.7)	0.05	3.8 (0.9 - 16.6)	0.06
	rs2243268	34428976	5q31.1	2.6 (1.1 - 6.0)	0.03	3.8 (0.9 - 16.6)	0.06
	rs2243290	34433182	5q31.1	2.4 (1.1 - 5.4)	0.04	3.8 (0.9 - 16.6)	0.06
MTHFR	rs1801133	6393745	1p36.3	2.3 (1.1 - 5.2)	0.04	4.1 (1.4 - 11.4)	< 0.01

[a] Estimated odds ratio (95% Confidence Interval)
[b] Likelihood ratio chi-square test with one degree of freedom

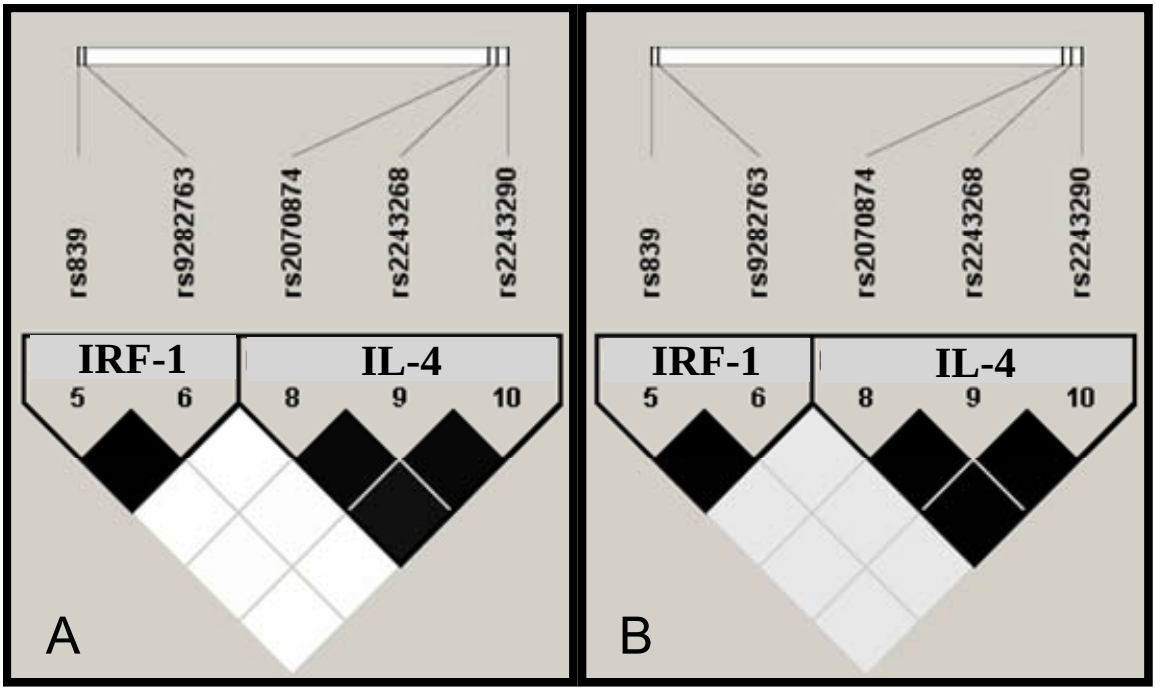
Only those 38 SNPs (within 31 genes) that showed significant associations in the original study were tested for potential associations in the validation study. The statistical results that replicated in the second study are shown alongside those from the original study in Table 4. The SNPs in IRF-1, IL-4, and MTHFR met our statistical significance criterion in the validation sample ($p = 0.03$, $p = 0.06$, and $p < 0.01$, respectively), and maintained an AE risk effect associated with the variant genotypes. While the SNPs in the IL-4 gene were not significant at a strict $p \leq 0.05$ level in the smaller validation set, these SNPs had p-values just beyond the traditional threshold. Considering the reduced size of the validation sample and the fact that the AE risk associated with variant genotypes was consistent across studies, these IL-4 SNPs warrant further study. The distribution of common versus variant genotypes at the replicated candidate SNPs is given in Table 5.

Table 5. Distribution of common versus variant (pooled) genotypes at the replicated candidate SNPs.

Gene	SNP (rs #)	SNP Location (Base Pair)	Genotype	Original Study Sample Count (Percent)	Validation Study Sample Count (Percent)
IRF-1	rs9282763	34237146	AA	39 (46%)	17 (37%)
			AG	43 (51%)	24 (52%)
			GG	3 (4%)	5 (11%)
	rs839	34234139	GG	39 (46%)	17 (37%)
			AG	43 (51%)	24 (52%)
			AA	3 (4%)	5 (11%)
IL-4	rs2070874	34424723	CC	52 (62%)	34 (74%)
			CT	28 (33%)	12 (26%)
			TT	4 (5%)	0 (0%)
	rs2243268	34428976	AA	52 (62%)	34 (74%)
			AC	27 (32%)	12 (26%)
			CC	5 (6%)	0 (0%)
	rs2243290	34433182	CC	53 (62%)	34 (74%)
			AC	26 (31%)	12 (26%)
			AA	6 (7%)	0 (0%)
MTHFR	rs1801133	6393745	CC	36 (42%)	18 (39%)
			CT	39 (46%)	21 (46%)
			TT	10 (12%)	7 (15%)

It is important to note that several of the significant SNPs (those located in the IRF-1 and IL-4 genes) were located in the same chromosomal region (5q31.1), suggesting an indirect association with one or more functional variants in that region. Because of the close physical proximity of the associated variants in those two genes, Haploview [13] software was used to look at the patterns of linkage disequilibrium (LD) among those variants in each sample. Figure 1 shows that the LD plots for SNPs in these two genes follow the same pattern in each study sample. While there is strong LD between SNPs within the two genes, there is no evidence for LD between the two genes, indicating that the associations for each gene are statistically separate signals.

Figure 1. Haploview plot of SNPs at chromosome 5q31.1. Panel A is the plot for the original study, and panel B is the plot for the validation study. Dark squares are indicative of strong evidence for LD between the pairwise markers (r^2 > 0.90), whereas lighter squares indicate no evidence (r^2 < 0.01 for white squares) or very weak evidence (r^2 < 0.10 for light gray squares) for LD. The same two LD blocks, separated by 190 Kb, are apparent in both studies, encompassing SNPs in IRF-1 (rs839 and rs9282763) or IL-4 (rs2070874, rs2243268, and rs2243290).



It has been demonstrated that this region of chromosome 5q31 contains discrete haplotype blocks [21]. Therefore, separate haplotypes were estimated for significant AE-associated SNPs in IRF-1 (rs839 and rs9282763) and IL-4 (rs2070874, rs2243268, rs2243290). In both study samples, two IRF-1 haplotypes accounted for all subjects. The common IRF-1 haplotype listed in Table 6 represented 71% of the original sample and 63% of the validation sample. The rare IRF-1 haplotype was significantly associated with AEs in both samples (p = 0.03). Across both studies, two different three-SNP haplotypes in

IL-4 accounted for 99% of subjects. The common IL-4 haplotype listed in Table 6 represented 78% of the original sample and 87% of the validation sample. The rare IL-4 haplotype was significantly associated with risk of AEs in the original sample (p = 0.05) and marginally associated with risk of AEs in the validation sample (p = 0.06).

Table 6. Haplotypes estimated for significant AE-associated SNPs in IRF-1 and IL-4.

			Original Study		**Validation Study**	
Gene	**Haplotype**		[c]**Odds Ratio** (95% C.I.)	[d]**p-value** (χ^2)	[c]**Odds Ratio** (95% C.I.)	[d]**p-value** (χ^2)
IRF-1	[a]**Baseline**	A - G	3.2 (1.0 - 10.2)	0.03	3.0 (1.0 - 9.0)	0.03
	[b]**Risk**	G - A				
IL-4	[a]**Baseline**	C - A - C	2.4 (1.0 - 5.7)	0.05	3.8 (1.0 - 14.4)	0.06
	[b]**Risk**	T - C - A				

[a] Most common haplotype considering 2 SNPs in IRF-1 (s839-rs9282763) or 3 SNPs in IL-4 (rs2070874-rs2243268-rs2243290)
[b] Rare (variant) haplotype considering 2 SNPs in IRF-1 (s839-rs9282763) or 3 SNPs in IL-4 (rs2070874-rs2243268-rs2243290)
[c] Estimated odds ratio comparing Risk haplotype to Baseline haplotype (95% Confidence Interval)
[d] Likelihood ratio chi-square test with one degree of freedom

Discussion

Biological mechanisms contributing to adverse events

While statistical association in two independent samples is a highly convincing result, it is the biological implications of such findings that are clinically relevant. These statistical results have strong biological plausibility and are in agreement with previous work on the topic of AEs following smallpox vaccination.

The candidate genes validated in both studies include a major anti-inflammatory cytokine (IL-4), an immunological transcription factor (IRF-1), and a metabolism gene previously associated with adverse reactions to a variety of pharmacologic agents (MTHFR). Since the outcome of interest is the aggregation of specific AEs, it is logical that more than one gene may be involved. These genes are all potentially involved in pathways that are in line with our previously hypothesized mechanism of adverse events involving excess stimulation of inflammatory pathways and the imbalance of tissue damage repair pathways. This model was developed from studies of circulating cytokines and relevant immunological effector cells [3-5]. For subjects experiencing adverse events, vaccination appears to trigger an acute inflammatory response akin to a delayed-type hypersensitivity reaction. Antigen presentation to Th1 cells in the dermis leads to the release of T-cell cytokines that trigger a cascade of cytokines and chemokines whose release enhances the inflammatory response by promoting the migration of monocytes into the lesion and their maturation into

macrophages and by further attracting T cells [22,23]. Taken together, these previous findings suggest that systemic adverse events following smallpox vaccination may be consistent with low-grade macrophage activation syndrome caused by virus replication and vigorous tissue injury and repair.

Relationship between genetic results and proposed model of adverse events

The 5,10-methylenetetrahydrofolate reductase (MTHFR) gene is located on chromosome 1. A SNP in MTHFR (rs1801133) is strongly associated with AE risk in both datasets. This non-synonymous SNP in the fifth exonic segment of the gene causes an amino acid change from Alanine to Valine. Functional characterization of this SNP has demonstrated that it is thermolabile and affects both the quantity and activity of the MTHFR enzyme [24].

The gene product catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, which is a cosubstrate for homocysteine remethylation to methionine. Proper MTHFR function provides pools of methyl groups that are crucial for the control of DNA synthesis and repair mechanisms [25]. It is a key enzyme in homocysteine metabolism, which plays a major role in regulating endothelial function.

The MTHFR enzyme has been associated with many phenotypes, including cardiovascular function, transplant health, toxicity of immunosuppressive drugs, and systemic inflammation [26-29]. Elevated plasma homocysteine levels stimulate endothelial inflammatory responses, which could contribute to systemic adverse events. Alternatively, since vaccination elicits

immune responses involving the rapid proliferation of cells, demand for DNA synthesis metabolites would be elevated, and alterations in the level or activity of MTHFR enzyme may exert significant influence over this process.

The Interleukin-4 (IL4) gene is located in a gene cluster on chromosome 5q31 that includes IL-13, IL-5, IRF-1, CSF-2, and IL-3. IL-4 has been found to be coordinately regulated with IL-13 and IL-5 by several long-range regulatory elements on the chromosome [30]. In addition to genetic polymorphisms, two alternatively spliced transcript variants of IL-4 that encode distinct isoforms have been discovered [31,32]. We found three SNPs in the IL-4 gene that are significantly associated with AEs in both studies: rs2070874 is a C>T substitution in the first exonic segment, rs2243268 is a A>C substitution in the second intronic segment, and rs2243290 is a A>C substitution in the third intronic segment.

Interleukin-4 encodes a pleiotropic cytokine produced by a variety of cells, including activated T cells and mast cells. The IL-4 cytokine is normally a major player in the activation of humoral immune responses, isotype switching to IgE, and suppression of Th_1 (CTL) cell functions. IL-4 is considered an anti-inflammatory cytokine for its inhibition of monocyte and dendritic cell migration to inflamed tissue, as well as its promotion of Th_2 effector pathways [33,34]. Naïve CD4+ T-cell differentiation away from the Th_1 pathway renders them unable to activate macrophages. There is also evidence that IL-4 cytokine secretion by T-regulatory cells is associated with the inhibition of many inflammatory T-cell responses [35].

Upon immunological challenge by vaccination, the differentiation of naïve CD4+ T-cells into armed Th_1 versus Th_2 cells plays a vital role in determining whether the adaptive immune response will be dominated by humoral effectors or macrophage activation [35]. Thus, genetic polymorphisms related to inappropriate regulation of IL-4 expression and/or activity of IL-4 cytokine may over-stimulate inflammatory responses—leading to the development of AEs. IL-4 dysregulation may also play a role in AEs resulting from the inappropriate clearance of apoptotic immune effector cells after infection, as this function is normally carried out by macrophages.

The Interferon Regulatory Factor-1 (IRF-1) gene is part of the immunological gene cluster on chromosome 5q31. We found two SNPs in the IRF-1 gene that are significantly associated with AEs in both study samples: rs9282763 is an A>G substitution in the sixth intronic segment and rs839 is a G>A substitution in the tenth exonic segment. The IRF-1 locus was initially mapped as a tumor suppressor, having genetic abnormalities associated with leukemia, myelodysplasia, and other cancers [36].

The IRF-1 gene encodes the transcription factor Interferon Regulatory Factor-1, a member of the interferon regulatory transcription factor (IRF) family. The IRF family regulates interferons and interferon-inducible genes. Many viruses use IRFs to evade host immune responses by binding to cellular IRFs and blocking transcriptional activation of IRF targets [37]. IRF-1 activates transcription of the Type I interferons alpha and beta as well as genes induced by

the Type II interferon gamma [38]. Type I interferon production by virus-infected cells enhances CTL and macrophage activity.

Polymorphisms in the gene coding for a transcription factor with such far-reaching effects as IRF-1 could have profound effects on the proper immune response and clearance of vaccinia virus. Hyperactive IRF-1 may push macrophage activity beyond the threshold of AE development. Hyperactive IRF-1 may also prolong the life of immune cells that should be cleared following infection, protracting the period of inflammation and leading to AEs.

Although the SNPs identified in IRF-1 and IL-4 do not change amino acids in the encoded proteins, recent evidence suggests that synonymous SNPs may exert functional influence over protein abundance [39,40]. Thus, the fact that multiple SNPs in high LD were identified in regions of IRF-1 and IL-4 presents three hypotheses for functional consequences of these SNPs. In one scenario, these SNPs are evidence of indirect association—meaning that the functional variant lies somewhere within the regions of LD defined in these two genes. In another scenario, one of the candidate polymorphisms identified here is the relevant variant; however, for these data, the LD between SNPs is too high to identify which one is functional. Finally, accumulated variation of the haplotypes defined within these genes contributes to alterations in protein levels by altering transcript stability or transcriptional rate.

Summary and future directions

These data present the rare opportunity to study two independent cohorts of smallpox vaccinées relating genetic factors to the occurrence of post-vaccination adverse events. Statistical analysis of the original study revealed potentially interesting associations between SNPs in biologically interesting candidate genes. Of the AE-associated genes identified in the original study, three replicated in an independent validation cohort. Genetic association studies are notorious for their failure to replicate, and validation studies are the epidemiological "gold standard" for reducing the risk of false positive findings. We avoid multiple testing issues by testing only the most promising results in the validation sample. Therefore, while all SNPs were tested in the original study, only those SNPs significantly associated with AEs were tested in the validation cohort. The validation of SNPs in three genes across both studies and their biologically viable connection with AEs lends credence to the reproducibility of these associations.

The results of this study demonstrate the importance and utility of validation in genetic studies of complex phenotypes. As with any statistical association, follow-up studies are needed to identify the particular genetic susceptibility variants and examine the functional consequences of polymorphisms in the AE-associated genes. Since we found multiple AE-associated SNPs in regions of IL-4 and IRF-1, focused studies should be undertaken to characterize the genetic variability in these candidate regions. While the association of AEs with a non-synonymous polymorphism in the gene

for MTHFR points toward functional significance of this SNP, deep resequencing should determine whether this is indeed the case. For all three candidate genes, functional studies are needed to connect genetic polymorphisms to variability in our hypothesized etiological pathways.

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 K25-AI-064625, R21-AI-59365, and R01-AI-59694); and NIH (GM-62758).

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. Seet BT, Johnston JB, Brunetti CR, Barrett JW, Everett H, Cameron C *et al*.: Poxviruses and immune evasion. *Annu Rev Immunol* 2003, 21: 377-423.

3. 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.

4. Rock MT, Yoder SM, Talbot TR, Edwards KM, Crowe JE, Jr.: Adverse events after smallpox immunizations are associated with alterations in systemic cytokine levels. *J Infect Dis* 2004, 189: 1401-1410.

5. Rock MT, Yoder SM, Wright PF, Talbot TR, Edwards KM, Crowe JE, Jr.: Differential regulation of granzyme and perforin in effector and memory T cells following smallpox immunization. *J Immunol* 2005, 174: 3757-3764.

6. Rock MT, Yoder SM, Talbot TR, Edwards KM, Crowe JE, Jr.: Cellular Immune Responses to Diluted and Undiluted Aventis Pasteur Smallpox Vaccine. *J Infect Dis* 2006, 194: 435-443.

7. Shaklee JF, Talbot TR, Muldowney JA, III, Vaughan DE, Butler J, House F *et al*.: Smallpox vaccination does not elevate systemic levels of prothrombotic proteins associated with ischemic cardiac events. *J Infect Dis* 2005, 191: 724-730.

8. 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.

9. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. *Genet Med* 2002, 4: 45-61.

10. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003, 33: 177-182.

11. 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.

12. Wigginton JE, Cutler DJ, Abecasis GR: A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 2005, 76: 887-893.

13. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005, 21: 263-265.

14. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM *et al*.: Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered* 2003, 55: 56-65.

15. Ihaka R, Gentleman R: R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 1996, 5: 299-314.

16. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org. 2006. Vienna, Austria.

17. StataCorp.. Stata Statistical Software. [Release 9]. 2005. College Station, TX, StataCorp LP.

18. Huang H, Patel DD, Manton KG: The immune system in aging: roles of cytokines, T cells and NK cells. *Front Biosci* 2005, 10: 192-215.

19. Larbi A, Douziech N, Fortin C, Linteau A, Dupuis G, Fulop T, Jr.: The role of the MAPK pathway alterations in GM-CSF modulated human neutrophil apoptosis with aging. *Immun Ageing* 2005, 2: 6.

20. Moroni F, Di Paolo ML, Rigo A, Cipriano C, Giacconi R, Recchioni R *et al*.: Interrelationship among neutrophil efficiency, inflammation, antioxidant activity and zinc pool in very old age. *Biogerontology* 2005, 6: 271-281.

21. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001, 29: 229-232.

22. Fong TA, Mosmann TR: The role of IFN-gamma in delayed-type hypersensitivity mediated by Th1 clones. *J Immunol* 1989, 143: 2887-2893.

23. Grom AA, Passo M: Macrophage activation syndrome in systemic juvenile rheumatoid arthritis. *J Pediatr* 1996, 129: 630-632.

24. Martin YN, Salavaggione OE, Eckloff BW, Wieben ED, Schaid DJ, Weinshilboum RM: Human methylenetetrahydrofolate reductase pharmacogenomics: gene resequencing and functional genomics. *Pharmacogenet Genomics* 2006, 16: 265-277.

25. Friso S, Girelli D, Trabetti E, Olivieri O, Guarini P, Pignatti PF *et al*.: The MTHFR 1298A>C polymorphism and genomic DNA methylation in human lymphocytes. *Cancer Epidemiol Biomarkers Prev* 2005, 14: 938-943.

26. Dedoussis GV, Panagiotakos DB, Pitsavos C, Chrysohoou C, Skoumas J, Choumerianou D *et al*.: An association between the methylenetetrahydrofolate reductase (MTHFR) C677T mutation and inflammation markers related to cardiovascular disease. *Int J Cardiol* 2005, 100: 409-414.

27. Lim U, Peng K, Shane B, Stover PJ, Litonjua AA, Weiss ST *et al*.: Polymorphisms in cytoplasmic serine hydroxymethyltransferase and methylenetetrahydrofolate reductase affect the risk of cardiovascular disease in men. *J Nutr* 2005, 135: 1989-1994.

28. Murphy N, Diviney M, Szer J, Bardy P, Grigg A, Hoyt R *et al*.: Donor methylenetetrahydrofolate reductase genotype is associated with graft-versus-host disease in hematopoietic stem cell transplant patients treated with methotrexate. *Bone Marrow Transplant* 2006, 37: 773-779.

29. Urano W, Taniguchi A, Yamanaka H, Tanaka E, Nakajima H, Matsuda Y *et al*.: Polymorphisms in the methylenetetrahydrofolate reductase gene were associated with both the efficacy and the toxicity of methotrexate used for the treatment of rheumatoid arthritis, as evidenced by single locus and haplotype analyses. *Pharmacogenetics* 2002, 12: 183-190.

30. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM *et al*.: Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* 2000, 288: 136-140.

31. Klein SC, Golverdingen JG, van Wichen DF, Bouwens AG, Stuij I, Tilanus MG *et al*.: Expression of two interleukin 4 mRNA isoforms in B lymphoid cells. *Cell Immunol* 1996, 167: 259-268.

32. Sorg RV, Enczmann J, Sorg UR, Schneider EM, Wernet P: Identification of an alternatively spliced transcript of human interleukin-4 lacking the sequence encoded by exon 2. *Exp Hematol* 1993, 21: 560-563.

33. Mangan DF, Robertson B, Wahl SM: IL-4 enhances programmed cell death (apoptosis) in stimulated human monocytes. *J Immunol* 1992, 148: 1812-1816.

34. 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.

35. Janeway CA, Travers P, Walport M, Shlomchik MJ: *Immunobiology: The Immune System in Health and Disease*, 5th edn. New York, New York: Garland Publishing; 2001.

36. Willman CL, Sever CE, Pallavicini MG, Harada H, Tanaka N, Slovak ML *et al*.: Deletion of IRF-1, mapping to chromosome 5q31.1, in human leukemia and preleukemic myelodysplasia. *Science* 1993, 259: 968-971.

37. Goodbourn S, Didcock L, Randall RE: Interferons: cell signalling, immune modulation, antiviral response and virus countermeasures. *J Gen Virol* 2000, 81: 2341-2364.

38. Harada H, Fujita T, Miyamoto M, Kimura Y, Maruyama M, Furia A *et al*.: Structurally similar but functionally distinct factors, IRF-1 and IRF-2, bind to the same regulatory elements of IFN and IFN-inducible genes. *Cell* 1989, 58: 729-739.

39. Crawford DC, Nickerson DA: Definition and clinical importance of haplotypes. *Annu Rev Med* 2005, 56: 303-320.

40. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J *et al*.: Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum Mol Genet* 2003, 12: 205-216.

CHAPTER V

FEATURE SELECTION USING RANDOM FORESTS FOR THE INTEGRATED ANALYSIS OF MULTIPLE SIMULATED DATA TYPES

Complex clinical phenotypes arise from the concerted interactions among the myriad components of a biological system. Therefore, comprehensive models can only be developed through the integrated study of multiple types of experimental data gathered from the system in question. The Random Forests™ (RF) method is adept at identifying relevant features having only slight main effects in high-dimensional data. This method is well-suited to integrated analysis, as relevant attributes may be selected from categorical or continuous data, and there may be interactions across data types. RF is a natural approach for studying gene-gene, gene-protein, or protein-protein interactions because importance scores for particular attributes take interactions into account. Thus, Random Forests is a promising solution to the analysis challenge posed by high-dimensional datasets including interactions among attributes of different types. In this study, we characterize the performance of RF on a range of simulated genetic and/or proteomic datasets. We compare the performance of RF in identifying relevant attributes when given genetic data alone, proteomic data alone, or a combined dataset of genetic plus proteomic data. Our results indicate that utilizing multiple data types is beneficial when the disease model is complex and the phenotypic outcome-associated data type is unknown. The results of

this study also show that RF is adept at identifying relevant features in high-dimensional data with small main effects and low heritability.

Introduction

Adverse drug reaction is one of the leading causes of hospitalizations in the Unites States. For example, in 1994 alone, adverse drug reactions accounted for more than 2.2 million serious hospitalizations [1]. Currently, there is no definitive way to determine how a person will respond to a medication—limiting pharmaceutical development to a "one size fits all" system. This system allows for the development of drugs to which the "typical" patient will respond, but one size does not necessarily fit all, sometimes with dire consequences. The need to screen patients for biomarkers predictive of response *a priori* to prevent adverse reactions has created a subspecialty within the field of human genetics known as pharmacogenomics.

The goal of pharmacogenomics is the identification and characterization of genes that predict drug response [2]. Due to the inherent complexity of the response phenotype, it is hypothesized that patient outcome is largely dependent upon interactions among genes and the environment. These nonlinear genetic interactions, known as epistasis, quickly diminish the applicability of traditional statistical methods. Taken together with the current explosion of genetic information as the field pushes towards genome-wide association studies, epistasis presents analytical challenges of an enormous combinatorial magnitude [3;4]. Traditional parametric analysis methods can be overwhelmed by datasets having huge numbers of attributes yet few samples. In response to the complex nature of current genetic studies, a number of novel statistical and computational

methods have been developed, such as Monte Carlo logic regression, two-stage approaches, Combinatorial Partitioning Method, Multifactor Dimensionality Reduction, and Detection of Informative Combined Effects [5-9].

Even with suitable analytical methodology, considering experimental information gathered from only one type of biological data will not permit the capture of the enormous complexity of systemic response phenotypes. Systems biology seeks to integrate multiple levels of information to understand how biological systems function [10]. By studying the relationships and interactions between various parts of a biological system, a more comprehensive model can be developed. Furthermore, because biology operates through a hierarchy of levels, incorporating data from multiple levels can provide surrogate data to fill gaps from any one biological level, and the partial redundancy between levels can further mitigate methodological unreliability [11].

For pharmacogenomic studies, an initial systems biology approach might measure variation in both genes and proteins in a patient to identify biomarkers that predict response to a given drug. While there is intuitive appeal to such a strategy, adding pieces of information on different scales of measurement (*i.e.* continuous proteomic data as well as categorical genetic data) creates additional analytical challenges. Therefore, appropriate computational analysis methods must not only traverse large numbers of input variables, but will also need to handle diverse data types.

One such computational method is the Random Forests (RF) approach [12]. RF is a machine learning technique that builds a forest of classification

trees wherein each tree is grown on a bootstrap sample of the data, and the attribute at each tree node is selected from a random subset of all attributes. The final classification of an individual is determined by voting over all trees in the forest. There are many advantages of the RF method that make it an ideal approach for the analysis of diverse biological data in pharmacogenomic studies. First, it can handle a large number of input attributes—both qualitative (*e.g.* Single Nucleotide Polymorphisms, or "SNPs") and quantitative (*e.g.* microarray expression levels or data from high-throughput proteomic technologies). Second, it estimates the relative importance of attributes in determining classification, 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 fairly robust in the presence of etiological heterogeneity and relatively high amounts of missing data [13]. Finally, and of increasing importance as the number of input variables increases, learning is fast and computation time is modest even for very large datasets [14].

In the current study, we use simulated data to investigate the potential of using a RF approach for the combined analysis of both genetic and proteomic data gathered in a study of adverse events associated with trials of a new smallpox vaccine [15;16]. The simulations are based on data collected from recent clinical trials of the Aventis-Pasteur Smallpox Vaccine (APSV), in which a significant proportion of vaccinees suffered systemic adverse events (AEs)—including fever, lymphadenopathy, and generalized rash. The data include genotypes at 1442 SNPs and measured circulating levels of 108 immunological

proteins. This dataset was chosen for its complex phenotype, the large number of attributes, and the multiple types of data collected. By using the actual data collected as the basis for our simulations, we reduce the number of over-simplifying assumptions and hope to better model the complexity inherent in real data. Because adverse reaction to vaccination is a complex phenotype, it is likely due to the coordinated action of multiple biological factors. Therefore, our simulated outcome (adverse event) models involve attribute interactions with only slight main effects.

In this study, we evaluate the ability of RF to detect outcome-associated simulated attributes by analyzing genetic data alone, proteomic data alone, or combined genetic and proteomic data. We address several questions with this study. First, to address the unresolved issue of where to set the importance cutoff for relevant attributes [13], can an appropriate threshold be set for the calculated RF importance relative to all attributes in the particular dataset analyzed that includes our simulated functional attributes? Second, how does RF perform when given different types of simulated biological data as input? Third, is there a relationship between the degree of informational redundancy and the ability of RF to select proteomic attributes related to the functional genetic attributes? Fourth, are there situations in which the analysis of multiple data types proves beneficial? In brief, our results indicate that utilizing multiple data types is beneficial when the disease model is complex and the outcome-associated data type is unknown. Importantly, using RF, we do not observe any significant *dis*advantage to an analysis strategy integrating both data types.

Methods

Random Forests

A Random Forest 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 based upon aggregate voting over all trees in the forest.

Each tree in the forest is constructed as follows from data having *N* individuals and *M* explanatory attributes:

1. Choose a training sample by selecting *N* individuals, with replacement, from the entire dataset.
2. At each node in the tree, randomly select *m* attributes from the entire set of *M* attributes in the data. The absolute magnitude of *m* is a function of the number of attributes in the dataset ($m = \sqrt{M}$) and remains constant throughout the forest building process.
3. Choose the best split at the current node from among the subset of *m* attributes selected above.
4. Iterate the second and third steps until the tree is fully grown (no pruning of lower branches with lesser predictive value).

Repetition of this algorithm yields a forest of trees, each of which have been trained on bootstrap samples of individuals (see Figure 1). Thus, for a given

tree, certain individuals will have been left out during training. Prediction error and attribute importance is estimated from these “out-of-bag” individuals.

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

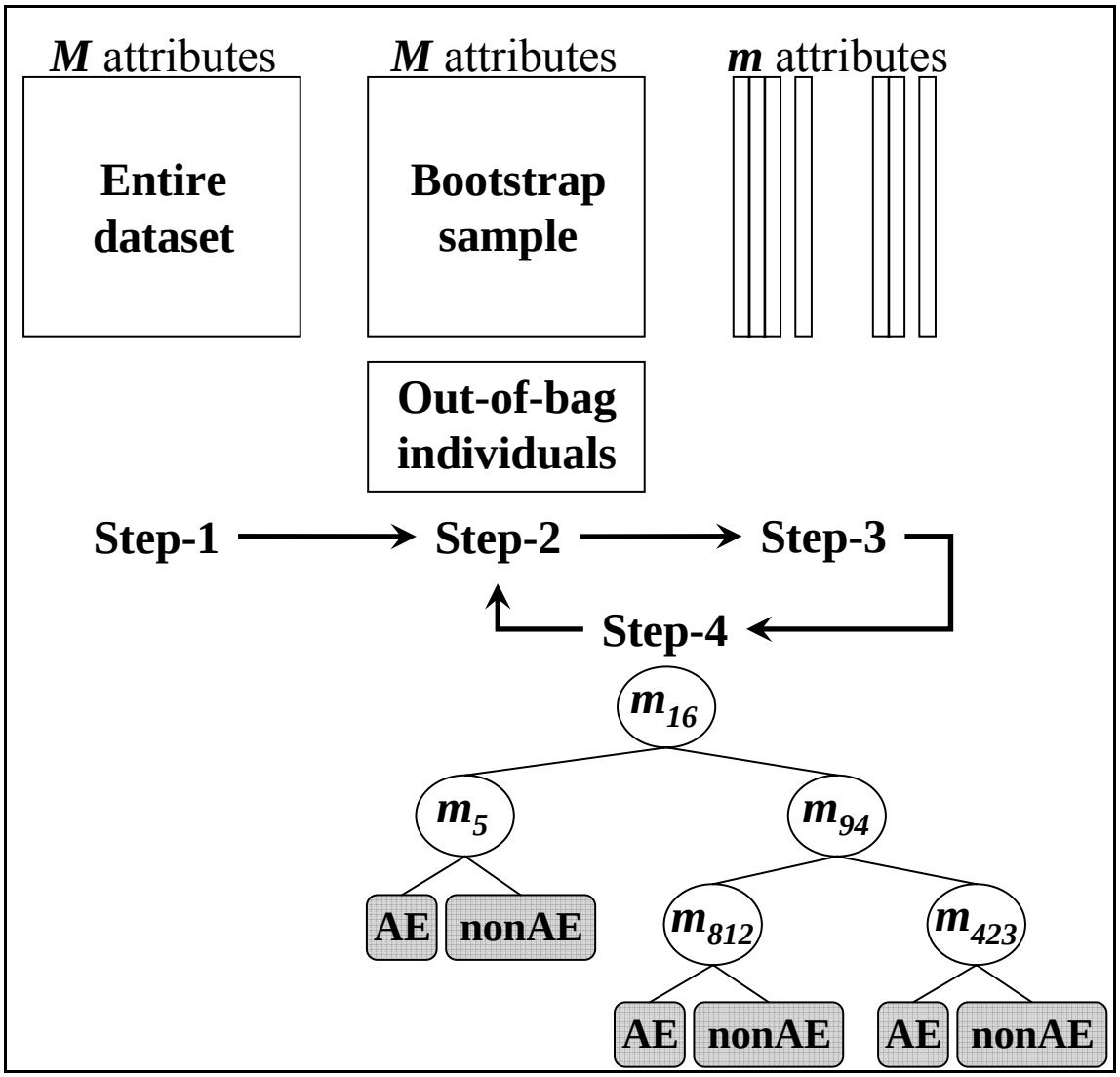


Figure 1. Construction of individual trees using the Random Forest method from a full dataset of *N* individuals and *M* attributes. The steps correspond to those described in the Methods section.

The recursive partitioning trees comprising a RF provide an explicit representation of attribute interaction that is readily applicable to the study of interactions among multiple data types [17;18]. These models may uncover interactions among genes, proteins, and/or environmental factors that do not exhibit strong marginal effects. Additionally, tree methods are suited to dealing with certain types of genetic heterogeneity, since splits near the root node define

separate model subsets in the data. Random Forests capitalize on the solid benefits of decision trees and have demonstrated excellent predictive performance when the forest is diverse (*i.e.* trees are not highly correlated with each other) and composed of individually strong classifier trees [12;19]. Diversity is achieved by finding an optimal *m* (the number of attributes tried at each node) that is considerably less than *M* (the total number of attributes in the data), which introduces variation into the forest building process; the optimal *m* will also build strong classifier trees by providing a sufficiently complete search through attributes in the data. 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 [20].

Data simulation

Simulation studies were designed to assess whether a Random Forests classifier is able to select the appropriate (outcome-associated) attributes from datasets consisting of categorical genetic (SNP) attributes, continuous proteomic (cytokine) attributes, or both. The results of this study will be used to develop an analysis strategy that effectively combines information gathered on diverse biological data types for the vaccine trial described below.

As mentioned previously, the simulations are based on data collected from recent clinical trials of the Aventis-Pasteur Smallpox Vaccine (APSV), where a high proportion of vaccinees suffered systemic adverse events (AEs). These

AEs included fever, lymphadenopathy, and generalized rash. The data collected include genotypes at 1442 SNPs (selected from genomic regions within or near candidate genes) and circulating levels of 108 immunological proteins (cytokines). For the APSV data, some proteomic attributes are also represented by genetic data in the corresponding gene. Thus, there is biological overlap between the two data types. Following the protocol described below, the simulated datasets mirrored the actual (APSV) trial data in terms of allele frequencies, SNP distribution across proteins, case (AE)/control (non-AE) ratio, potential patterns of linkage disequilibrium between SNPs, covariance structure across protein levels, etc.

To create simulated data reflecting the complex properties of that collected for the APSV study, those data were used as the basis for the simulations. First, the AE status was stripped from the APSV data. Next, a new AE status was assigned according to genetic attributes in the data consistent with our simulated genetic models and maintaining the overall case/control (AE/nonAE) ratio. Then, to represent the biological transfer of information between genes and proteins, proteomic attributes related to the functional genetic attributes were added. The related proteomic attributes simulate a range of gene→protein information transfer proportions. For example, to simulate a functional (outcome-associated) genetic attribute that is represented by the corresponding protein in the proteomic data, a related proteomic attribute is added to the proteomic data. However, to account for biological variation between genotype and protein level, the functional genetic attribute is only

responsible for a portion of the variation in protein level for the related attribute (see Figure 2). Thus, information is not transferred between related attributes with perfect fidelity.

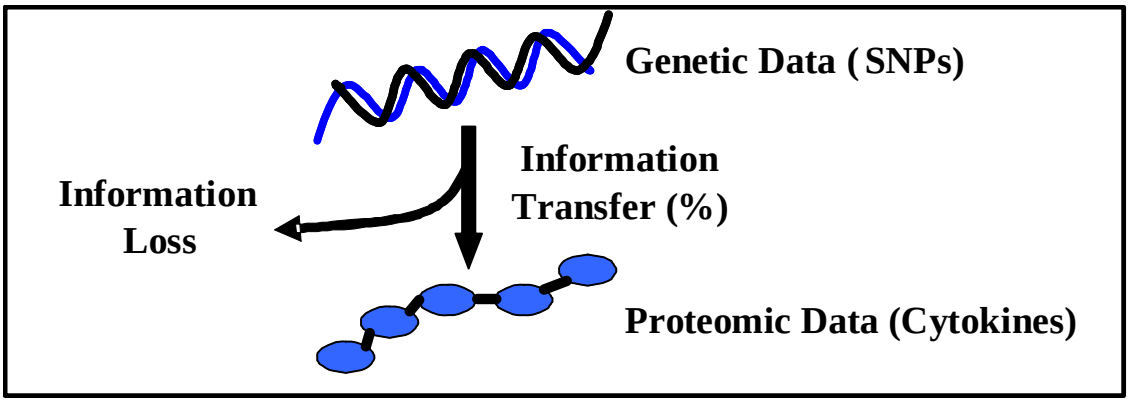


Figure 2. Information transfer between simulated genetic and proteomic attributes. For a particular attribute, the information transfer is the proportion (%) of variation in the simulated proteomic attribute explained by the corresponding genetic attribute.

Penetrance functions are used to represent our partially epistatic genetic models. As in Table 1, penetrance defines the probability of experiencing an adverse event given a particular genotype combination. For these models, two genetic attributes ($Genetic_A$ + $Genetic_B$) have a joint (epistatic) effect upon outcome class, and each attribute also has a very slight marginal effect (*M*) above the population prevalence (*K*). For a particular combination of genotypes at *i*=$Genetic_A$ and *j*=$Genetic_B$, the probability of belonging to the outcome class AE = f_{ij} in Table 1. A range of heritability values was selected for our simulations, including 10%, 20%, and 40%. Roughly, heritability is the proportion of the total variation in outcome that is due to genetic effects. Although the heritability

values used here translate to weak signals in the data, these values would classify as low to moderate genetic effects. Since there is scant data relating adverse events after vaccination with APSV to serum proteomic data or SNP data, heritability values in the low- to mid-range of those estimated for common complex phenotypes were used in these simulations. For a more thorough explanation of the heritability calculations used in this study, see [21]. An example of the penetrance functions used for the models generated in this study is given in Table 2.

Datasets with a range of genetic→proteomic information transfer (see Figure 2) were created for each genetic model. For each combination defined in Table 3 by a genetic model heritability (10%, 20%, 40%), a proportion of genetic→proteomic information transfer (15%, 30%, 45%, 60%, 75%, 90%), and a data type (Genetic, Proteomic, Genetic+Proteomic), 100 datasets were simulated for analysis, resulting in 5400 total datasets. The data simulation strategy is summarized in Figure 3.

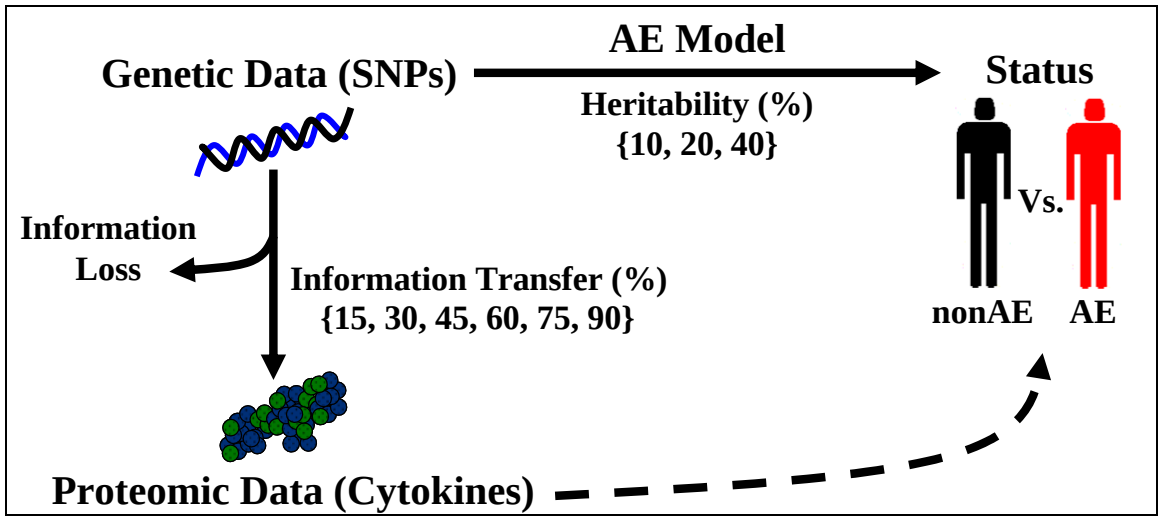


Figure 3. Summary of the data simulation strategy. First, the AE status was stripped from the APSV data. Next, a new AE status was assigned according to simulated genetic models with a range of heritability. Then, proteomic attributes related to the functional genetic attributes were added with a range of information transfer percentages, resulting in proteomic attributes that are indirectly related to AE status (represented by the dashed line).

Table 1. Penetrance function for a model of AE status associated with two functional genetic attributes: *A* and *B*.

		Genetic Attribute B			
		BB	Bb	bb	
Genetic Attribute A	AA	f_{11}	f_{12}	f_{13}	M_{A1}
	Aa	f_{21}	f_{22}	f_{23}	M_{A2}
	aa	f_{31}	f_{32}	f_{33}	M_{A3}
		M_{B1}	M_{B2}	M_{B3}	$\boldsymbol{K}$

Table 2. Example penetrance function for a simulated genetic AE model with 10% heritability. Allele frequencies for each attribute are equal ($p = q = 0.5$).

		Genetic Attribute B			
		BB	Bb	bb	
Genetic Attribute A	AA	0	0	0	M_{A1}
	Aa	0	0.2	0.2	M_{A2}
	aa	0	0.2	0.2	M_{A3}
		M_{B1}	M_{B2}	M_{B3}	$\boldsymbol{K}$

Table 3. Overview of simulated datasets. For each combination of genetic heritability and genetic→proteomic information transfer, 100 datasets were simulated, each containing one of the following data types: Genetic data alone = G; Proteomic data alone = P; Genetic + Proteomic data combined = GP.

		Genetic-Proteomic Information Transfer					
		15%	30%	45%	60%	75%	90%
Genetic Heritability	10%	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP
	20%	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP
	40%	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP	G,P,GP

Data analysis

The analysis was performed using the freely available R package randomForest [22;23]. This package is based on the original Fortran code available at [24].

Given a dataset having a particular combination of genetic model heritability and genetic→proteomic information transfer (see Table 3), RF was used to analyze datasets containing each simulated biological data type separately and in parallel. Genetic attributes were treated as categorical while proteomic attributes were treated as continuous values. For each of the 100 genetic, proteomic, or combined datasets, forests comprised of 10,000 trees were grown. Attribute importance was calculated using the out-of-bag permutation test. The relative importance (rank) of functional genetic attributes and related proteomic attributes was determined from the mean decrease in Gini index using the out-of-bag permutation testing procedure [17]. The Gini diversity index, *i*, at a tree node, *t*, has the form $\sum_{j \neq i} p(j \mid t) p(i \mid t)$, where $p(j|t)$ and $p(i|t)$ are the probabilities of assigning a subject to classes *j* or *i*, respectively [17]. The relative importance determined from the mean decrease in classification accuracy produced statistically similar results.

Results

Figure 4 shows the relative importance rank (expressed as a percentile) of the two functional genetic attributes calculated by the RF over all datasets. Each data point on the graph represents the mean relative importance rank calculated over 100 datasets, with the bars representing 95% confidence intervals about the mean. This figure demonstrates several important trends regarding the relative importance of the functional genetic variables with the three possible combinations of data types analyzed (Genetic alone = G, Proteomic alone = P, Genetic + Proteomic combined = GP). Analyzing the genetic data alone consistently demonstrated the highest relative importance for the functional genetic attributes. Analyzing the combined genetic + proteomic data demonstrated relative importance that was very near to that of the genetic alone. This slight discrepancy may be due to the increased number of noise attributes (the combined dataset has 1550 attributes while the genetic data alone has only 1442). It is interesting to note that as the heritability of the model increases, the gap in functional attribute importance between the genetic and combined analyses narrows. Of course, regardless of heritability or genetic→proteomic information transfer, analyzing the proteomic data alone makes it impossible to identify the correct genetic attributes since they are not present in the proteomic datasets. Also, as expected, the relative importance of the genetic attributes is not influenced by the amount of information transfer between genetic and proteomic data.

Additionally, it is clear from Figure 4 that as the heritability of the model increases (across the panels from 10-20-40%), the relative rank of the functional genetic attributes increases. This is expected, since increased heritability increases the signal strength in the data. It is also important to note that even at the lowest heritability simulated (10%), RF successfully identifies the functional variables as relatively important (above the 80th percentile for all models). The fact that the functional variables are not always ranked at the top of all importance scores means that RF is also finding chance AE associations in datasets with weak simulated genetic signals.

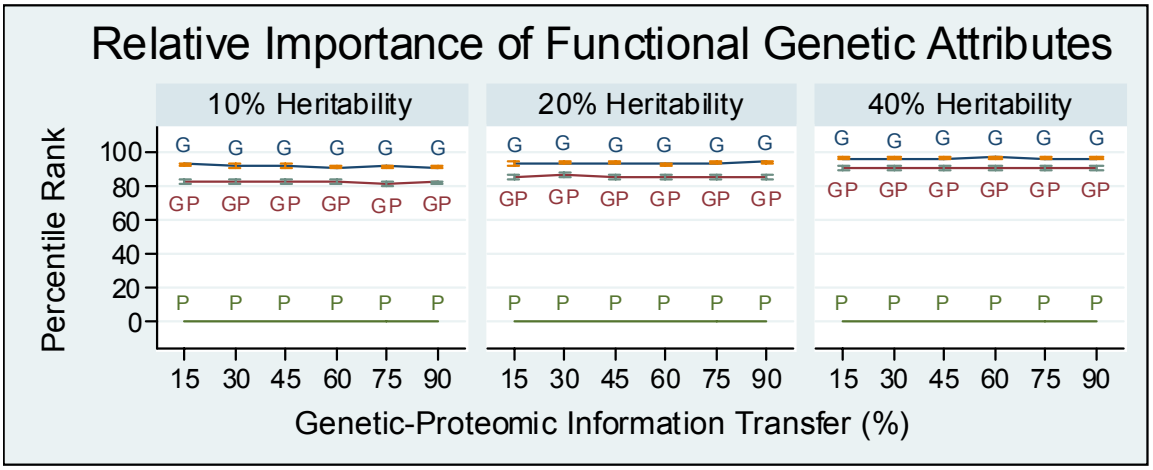


Figure 4. Relative importance of functional genetic outcome-associated attributes for each data type analyzed: Genetic data alone = G; Proteomic data alone = P; Genetic data + Proteomic data combined = GP. Each labeled point represents the mean (plus 95% confidence interval) importance over 100 datasets. Note: the functional genetic attributes are not present in datasets comprised of only proteomic data (P).

Figure 5 shows the RF relative importance rank of the proteomic variables related to the functional genetic variables (by the % information transfer given along the horizontal axis). Again, each data point represents the mean relative importance rank of the related proteomic attributes calculated over 100 datasets, with the bars representing 95% confidence intervals about the mean. The results are shown for all models, and several significant trends are clear. As expected, when just the genetic datasets are analyzed, it is impossible to identify any proteomic variables as important since they are excluded from those data. Also apparent from Figure 5 are the wider confidence intervals associated with analysis of the proteomic datasets alone.

As in Figure 4, increased heritability of the underlying genetic models generally increases the relative importance of outcome-associated attributes

(which are the related proteomic attributes in Figure 5). Unlike the relative importance of genetic attributes considered in Figure 4, where the results were unaffected by the amount of information transfer between the genomic and proteomic data, when considering the related proteomic attributes in Figure 5, it is clear that the degree of relatedness between the functional genetic attributes and the related proteomic attributes (information transfer) exerts significant influence over the relative importance. This trend is very pronounced in the analysis of the proteomic data alone. As the information transfer increases, the relative importance of the related proteomic attributes increases. The same is true, although to a lesser degree, for the combined genetic + proteomic analyses. Since the disease models are genetic, it is intuitive that as the amount of information transfer between genetic and proteomic attributes increases, the stronger the signal in the proteomic data.

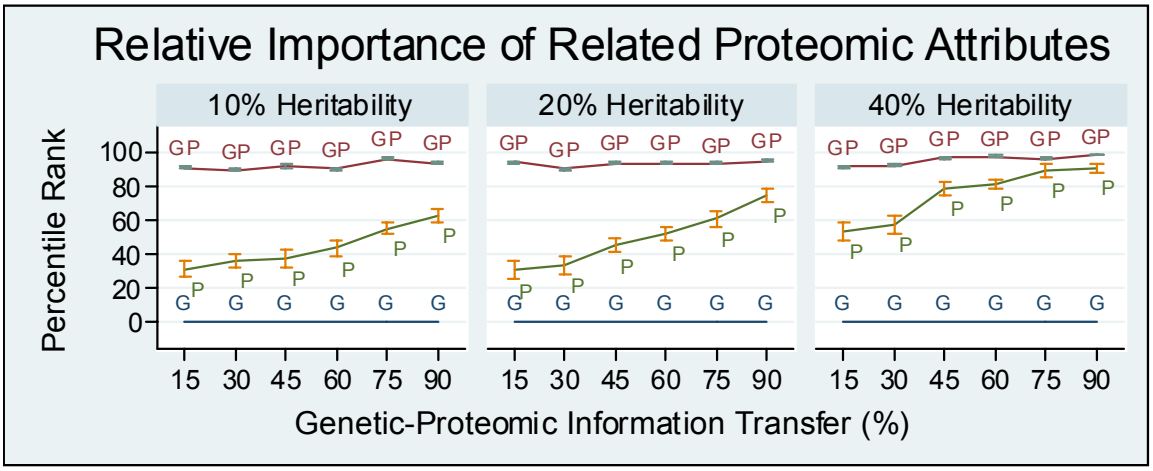


Figure 5. Relative importance of proteomic attributes related (according to the amount of genetic-proteomic information transfer along the horizontal axis) to functional genetic attributes for each data type analyzed: Genetic alone = G, Proteomic alone = P, Genetic + Proteomic combined = GP. Each labeled point represents the mean (plus 95% confidence interval) importance over 100 datasets. Note: functional proteomic attributes are not present in datasets comprised of only genetic data (G).

The most striking trend shown in Figure 5 is the large difference between the proteomic and the combined genetic + proteomic analysis strategies. The combined genetic + proteomic analysis strategy is substantially more successful at identifying the related proteomic attributes as important than analysis of the proteomic data alone, especially for models with lower heritability and information transfer. This performance gap may arise out of the partially epistatic nature of the models and the stochastic nature of the RF methodology. Considering models with only slight marginal effects, for RF to assign high attribute importance scores, trees must consistently contain both of the relevant interacting attributes. For the combined dataset (containing two functional genetic attributes and two related proteomic attributes), there are more opportunities to choose one of the interacting relevant attributes nearer the root

of the tree and then choose the complementary attribute at subsequent splits than for the proteomic data alone (containing only two related proteomic attributes). The performance gap between genetic versus combined datasets in identifying relevant proteomic attributes narrows as both information transfer and heritability increase.

Discussion

The results of this study demonstrate that there is a marked advantage to an integrated analysis approach incorporating multiple data types. While the genetic analysis was appropriate for identifying the functional genetic features, the combined strategy analyzing both genetic and proteomic data performed nearly as well at identifying functional genetic attributes and provides another distinct advantage—the identification of important related proteomic variables. This property would be beneficial in situations where the functional outcome-associated data type is unknown or not appropriately measured. For example, our simulated models are not determined by protein *abundance*, as is often measured experimentally. Instead, our simulations represent a situation wherein genotype codes for some unmeasured proteomic aspect (*e.g.* enzymatic activity) that determines phenotype. Still, if protein abundance is also related to genotype, even with some loss of information, the proteomic data can be analytically useful. The convergence of genetic and related proteomic attributes receiving high importance scores could serve as a strategy for limiting false

positive results. Also, including multiple data types has the intangible advantage of allowing for better biological interpretation of a resulting model. These results show no substantial *dis*advantage to the joint analysis of multiple data types.

With respect to setting an appropriate cutoff for selection of relevant features using RF, our results indicate that the choice of threshold depends upon the strength of the signal in the data. From Figures 4 and 5, it appears that the importance threshold may need to be relaxed to identify relevant attributes in datasets with low signal and a low degree of information transfer between related data types. However, RF seems largely robust to the addition of noise variables in the larger datasets—so long as relevant attributes are present in the data. The results of this study also show that RF is adept at identifying relevant features in high-dimensional data containing attributes on multiple scales of measurement. RF identifies features with small marginal effects and low heritability. Relevant attributes may be selected from either data type, and there may be interactions across data types. RF is thus well-suited to the study of phenotypes with complex underlying etiologies, where the biological features of interest have yet to be elucidated.

While the results of this study are promising, there are questions yet to be addressed. The combined RF approach needs to be applied to a real dataset (and the results tested at the lab bench) to confirm the conclusions of the simulation study. Currently, the dataset used as the template for the simulations is being analyzed using the integrated RF approach found to be successful with these simulations. Additionally, work must be continued on modifications to RF

that allow for the discovery of purely epistatic genetic models [19]. Because RF chooses only one attribute at each tree split during construction, strictly epistatic (*i.e.* absence of even miniscule main effects) attributes will not be selected. Finally, strategies for automatically translating the features selected by RF into meaningful biological hypothesis need to be developed.

Acknowledgments

This work was supported by NIH grants AI-064625, AI-59694, AI-057661, and GM-62758.

References

1. J. Lazarou, B. H. Pomeranz, and P. N. Corey, "Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies," *JAMA*, vol. 279, no. 15, pp. 1200-1205, 1998.

2. R. A. Wilke, D. M. Reif, and J. H. Moore, "Combinatorial pharmacogenetics," *Nature Reviews Drug Discovery*, vol. 4, no. 11, pp. 911-918, 2005.

3. J. H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Human Heredity*, vol. 56, no. 1-3, pp. 73-82, 2003.

4. J. H. Moore and S. M. Williams, "Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis," *Bioessays*, vol. 27, no. 6, pp. 637-646, 2005.

5. C. Kooperberg and I. Ruczinski, "Identifying interacting SNPs using Monte Carlo logic regression," *Genetic Epidemiology*, vol. 28, no. 2, pp. 157-170, 2005.

6. J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature Genetics*, vol. 37, no. 4, pp. 413-417, 2005.

7. M. R. Nelson, S. L. Kardia, R. E. Ferrell, and C. F. Sing, "A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation," *Genome Res.*, vol. 11, no. 3, pp. 458-470, 2001.

8. M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *American J Human Genetics*, vol. 69, no. 1, pp. 138-147, 2001.

9. N. Tahri-Daizadeh, D. A. Tregouet, V. Nicaud, N. Manuel, F. Cambien, and L. Tiret, "Automated detection of informative combined effects in genetic association studies of complex traits," *Genome Res.*, vol. 13, no. 8, pp. 1952-1960, 2003.

10. L. Hood, "Systems biology: integrating technology, biology, and computation," *Mech. Ageing Dev.*, vol. 124, no. 1, pp. 9-16, 2003.

11. D. M. Reif, B. C. White, and J. H. Moore, "Integrated analysis of genetic, genomic, and proteomic data," *Expert Reviews in Proteomics*, vol. 1, no. 1, pp. 67-75, 2004.

12. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

13. K. L. Lunetta, L. B. Hayward, J. Segal, and E. P. Van, "Screening large-scale association study data: exploiting interactions using Random Forests," *BMC Genet*, vol. 5, no. 1, p. 32, 2004.

14. M. Robnik-Sikonja, "Improving Random Forests," *Machine Learning: Ecml 2004, Proceedings*, vol. 3201, pp. 359-370, 2004.

15. B. A. McKinney, D. M. Reif, M. T. Rock, K. M. Edwards, S. F. Kingsmore, J. H. Moore, and J. E. Crowe, Jr., "Cytokine expression patterns associated with systemic adverse events following smallpox immunization," *J Infect. Dis.*, vol. 194, no. 4, pp. 444-453, 2006.

16. M. T. Rock, S. M. Yoder, T. R. Talbot, K. M. Edwards, and J. E. Crowe, Jr., "Adverse events after smallpox immunizations are associated with alterations in systemic cytokine levels," *J Infect. Dis.*, vol. 189, no. 8, pp. 1401-1410, 2004.

17. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York: Chapman & Hall, 1984.

18. M. A. Province, W. D. Shannon, and D. C. Rao, "Classification methods for confronting heterogeneity," *Adv Genetics*, vol. 42, pp. 273-286, 2001.

19. A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and E. P. Van, "Identifying SNPs predictive of phenotype using Random Forests," *Genet Epidemiology*, vol. 28, no. 2, pp. 171-182, 2005.

20. B. A. McKinney, D. M. Reif, M. D. Ritchie, and J. H. Moore, "Machine Learning for Detecting Gene-Gene Interactions: A Review," *Appl. Bioinformatics*, vol. 5, no. 2, pp. 77-88, 2006.

21. R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich, "A perspective on epistasis: limits of models displaying no main effect," *American J Human Genetics*, vol. 70, no. 2, pp. 461-471, 2002.

22. R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, vol. 5, pp. 299-314, 1996.

23. R Development Core Team, "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, 2006.

24. L. Breiman and A. Cutler, "Random Forests," www.stat.berkeley.edu/users/breiman/RandomForests, 2004.

CHAPTER VI

INTEGRATED ANALYSIS OF GENETIC AND PROTEOMIC DATA IDENTIFIES BIOMARKERS ASSOCIATED WITH ADVERSE EVENTS FOLLOWING SMALLPOX VACCINATION

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 only be developed 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 development of adverse events (AEs) in patients following smallpox vaccination. Since vaccination was successful in the patients under study, the AE outcomes reported likely represent interactions among immune system components that either push immune responses beyond the necessary magnitude or sustain responses longer than necessary. In the current study, we examined 1442 genetic variables (SNPs) and 108 proteomic variables (cytokine levels) to model AE risk. To accomplish this daunting analytical task, we employed the Random Forests^{TM} (RF) method to filter out the most important attributes, then 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 AE

development because 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 build a comprehensive model of AE development that includes the cytokines ICAM-1 (CD54), IL-10, and CSF-3 (G-CSF), as well as a genetic polymorphism in IL-4. The biological factors included in the model support our hypothesized mechanism for the development of AEs involving prolonged stimulation of inflammatory pathways and the imbalance of normal tissue damage repair pathways. This study demonstrates the utility of the RF for such analytical tasks, and both enhances and reinforces our working model of AE development following smallpox vaccination.

Introduction

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

To elucidate the complex pathophysiology underlying inappropriate response to vaccination, we gathered high-dimensional genetic and proteomic data in a cohort of subjects in which an unacceptably high proportion experienced an AE following primary immunization with Aventis Pasteur smallpox vaccine (APSV). Through a comprehensive examination of systemic (serum) cytokine/chemokine changes combined with characterization of polymorphisms in a panel of candidate genes, we aim to provide a thorough portrayal of the complex genetic and proteomic interplay behind the development of adverse events. Knowledge of how risk factors in a subject's genetic background 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.

Although there is considerable intuitive appeal to incorporation of multiple types of biological data, simultaneous analysis of information on different scales of measurement (*i.e.* continuous proteomic data as well as categorical genetic data) creates additional analytical challenges. Therefore, appropriate computational analysis methods must not only traverse large numbers of input variables, but will also need to handle diverse data types. For this study, we employed a two-stage analysis strategy. The first step was to effectively 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 Forests™ (RF) approach [2]. 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 (*e.g.* Single Nucleotide Polymorphisms, or "SNPs") and continuous (*e.g.* microarray expression levels or data from high-throughput proteomic technologies). Second, it estimates the relative importance of attributes in discriminating between classes (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 fairly robust in the presence of etiological heterogeneity and missing data [3]. Finally, learning is fast and computation time is modest even for very large datasets [4].

In the second (modeling) step, we took advantage of the tractable number of attributes identified by the RF filter to thoroughly explore 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 justifies our multi-scale analysis strategy, in that it includes the cytokines ICAM-1, IL-10, and CSF-3 (G-CSF), as well as a SNP in IL-4. Evaluating our final model from an immunological perspective, we conclude that AEs in response to smallpox vaccination result from hyperactivation of innate inflammatory pathways leading to excess recruitment and stimulation of monocytes in peripheral tissues. This model is consistent with work demonstrating over-stimulation of inflammatory and tissue damage repair pathways developed in previous studies of AEs following smallpox vaccination [5-8].

Methods

Study subjects

Vaccines, study subjects, and study design have been described in detail in [6]. Briefly, 148 (116 with recorded AE information) healthy adults were

enrolled at the Vanderbilt University Medical Center as part of a multi-center study of primary immunization against smallpox using the APSV 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 was gathered. Individuals were asked to self-identify race, with White (60) and Asian (1) as the only categories. There was no statistical difference in age, gender, or race according to AE status.

Clinical assessments

For all study subjects, a team of trained physicians and nurse providers examined the medical history and clinical symptoms to insure consistent clinical assessment. Subjects were examined on 5 visits within the first month post-vaccination and were assessed for occurrence of an adverse event. Collection of serum for cytokine measurements occurred at the evaluation just before vaccination (baseline) and at the evaluation between days 6-9 post-vaccination (acute phase). While all adverse events were noted, only systemic AEs were considered in this study, since 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 greater than 38.3°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 can be found in [9]. Lymphadenopathy was defined as enlargement or tenderness of regional lymph nodes attributed to vaccination. For subjects on which both genetic and proteomic data was 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 originally developed for genetic studies of human cancers. Thus, the SNPs were chosen for genotyping based on their oncological relevance. As such, the majority of SNPs included on the panel were involved in signaling pathways, many of which had immunological components. Genotyping for single nucleotide polymorphisms (SNPs) was performed using DNA amplified directly from blood samples collected from each subject. Genotyping was performed at the Core Genotyping Facility of the National Cancer Institute (NCI) in Gaithersburg, Maryland (http://cgf.nci.nih.gov/home.cfm) [10]. Genotypes were generated using the Illumina™ GoldenGate assay technology. Of the 1536 SNPs assayed, a total of 1442 genotypes passed quality control filters.

Quantification of serum cytokine levels

Serum samples were obtained just prior to vaccination (baseline) and 6-9 days after vaccination (acute). Serum samples were collected in 5 ml Vacutainer serum separator tubes (Becton Dickinson, San Jose, CA) and were centrifuged at 700 x g for 10 minutes. The serum then was collected, aliquoted into cryovials (Sarstedt Inc., Numbrecht, Germany) and stored at –80 ºC until assayed for cytokine concentrations using Rolling circle amplification technology (RCAT).

A custom dual antibody sandwich immunoassay array, as described in [11-14] was used to measure the expression levels of 108 protein analytes in 100 µL serum aliquots from the patient samples. Briefly, glass slides held 12 replicate spots of monoclonal capture antibodies specific for each analyte. Duplicate samples of sera were incubated for 2 hours, washed, and then incubated with secondary biotinylated polyclonal antibodies. The 'rolling circle' method was then used to amplify signals [12]. Quality control measures were used to optimize antibody pairs, minimize array-to-array variation, and standardize procedures of chip manufacturing [12]. A Tecan LS200 unit was used to scan arrays and customized software was used to determine mean fluorescence intensities (MFIs). Additionally, 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 baseline due to the broad individual range of systemic cytokine expression before and after immunization.

Random forests

A Random Forest 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 based upon aggregate voting over all trees in the forest.

Each tree in the forest is constructed as follows from data having *N* individuals and *M* explanatory attributes:

1. Choose a training sample by selecting *N* individuals, with replacement, from the entire dataset.
2. At each node in the tree, randomly select *m* attributes from the entire set of *M* attributes in the data. The absolute magnitude of *m* is a function of the number of attributes in the dataset ($m = \sqrt{M}$) and remains constant throughout the forest building process.
3. Choose the best split at the current node from among the subset of *m* attributes selected above.
4. Iterate the second and third steps until the tree is fully grown (lower branches are not trimmed in the interest of generalizability).

Repetition of this algorithm yields a forest of trees, each of which have been trained on bootstrap samples of individuals (see Figure-1). Thus, for a given tree, certain individuals will have been left out during training. Prediction error and attribute importance is estimated from these "out-of-bag" individuals according to the procedure described in Chapter V.

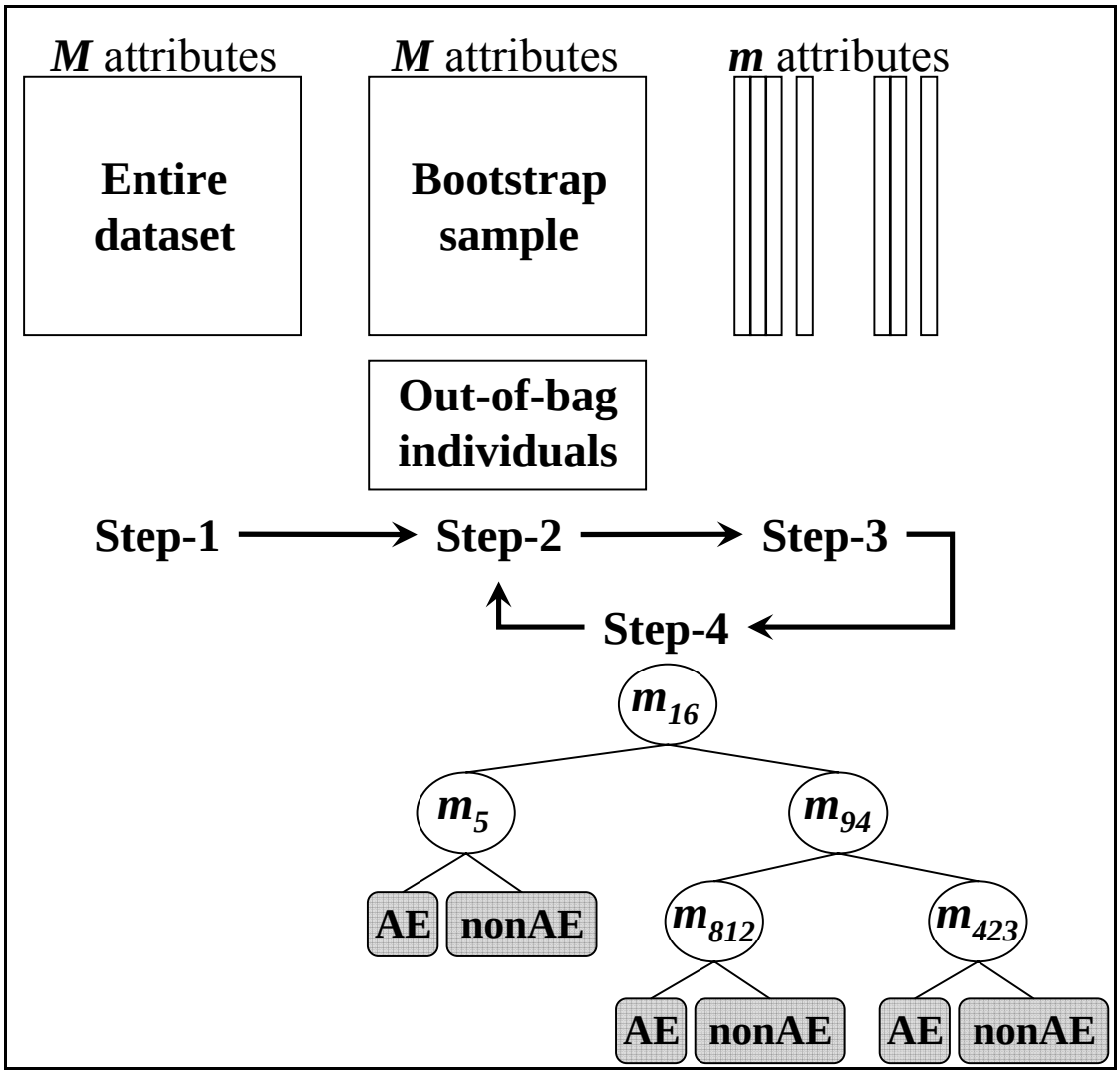


Figure 1: Construction of individual trees using the random forest method from a full dataset of *N* individuals and *M* attributes. Proceeding from the root node, individual subjects are classified into terminal AE status leaves according to the value of that individual's genetic or proteomic attribute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

The recursive partitioning trees comprising a RF provide an explicit representation of attribute interaction that is readily applicable to the study of interactions among multiple data types [15,16]. As discussed in Chapter V, RF have demonstrated excellent predictive performance when the forest is diverse

(*i.e.* trees are not highly correlated with each other) and composed of individually strong classifier trees [17,18]. 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 *a priori* model specification [19].

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 classifies individual subjects into AE groups by proceeding down a dichotomous tree, where the genetic or proteomic attribute at each node (or split) is selected for the gain in information it provides (Essentially: how well knowledge about the variation in this attribute separates subjects into appropriate AE classes). When interpreting the tree, attributes at each node are 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. While cross-validation accuracy was reduced by allowing trees with less than five subjects in terminal nodes, cross-validation 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 [20].

Data analysis strategy

Random Forest analysis was performed using the freely available R package randomForest [21,22]. This package is based on the original Fortran code available at [23]. RF was used to analyze datasets containing each biological data type separately and in parallel. Genetic attributes were treated as categorical while proteomic attributes were treated as continuous values. For each genetic, proteomic, or combined dataset, forests comprised of 10,000 trees were grown. This forest size gave stable estimates of attribute importance. Attribute importance was calculated using the out-of-bag permutation test described in Chapter V. The relative importance (rank) of functional genetic attributes and related proteomic attributes was determined from the mean decrease in Gini index (see Chapter V) 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 [24].

Results from simulation studies based on these data demonstrate high confidence that AE-associated attributes having low to moderate effects will be ranked in the top 10% of attributes in RF analysis [24]. Therefore, we chose the top 10% of attributes as ranked by RF as candidates for inclusion in our final model. While this threshold may have missed attributes with very weak effects, it is unlikely that such effects would have been detectable given our sample size of 61 subjects. To represent the interactions among genetic and/or proteomic

attributes associated with AEs, we built a decision tree model, as previously described.

Biological interpretation of our final model was aided by the Chilibot (chip literature robot) knowledge mining software, as described in [25]. Chilibot inferred relationship networks among the attributes in the final model based upon 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 (*e.g.* inhibition or stimulation) between input terms.

Results

Filtering of important attributes using random forests

Table-1 lists all attributes having an importance rank in the top 10% relative to all attributes in the combined dataset. Figure-2 depicts the attribute importance score landscape over the entire dataset. This landscape proved robust to changes in RF parameters (such as attributes importance metrics and AE class-weighting schemes), provided that a sufficiently large forest 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. While this distribution among data types may reflect systematic patterns concerning the etiology of AE outcomes, the bias toward proteomic attributes probably arises 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. Also, with multiple SNPs per gene, correlation (*i.e.* haplotypes) existing among polymorphisms 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 IL-4 SNP in our final model was part of a group of four SNPs in IL-4 having nearly identical importance scores, and Haploview analysis showed them to be in high linkage disequilibrium (LD), providing evidence that these genetic polymorphisms are inherited as a haplotype [26].

Considering the attributes included in our final model, all three proteomic attributes were ranked in the top 1% relative to all attributes in the combined dataset, and the IL-4 SNP (rs#2243290) was ranked in the top 5% relative to all attributes in the combined dataset. Relative to its respective data type, the IL-4 SNP was ranked in the top 1% among all attributes in the genetic dataset.

Table 1: List of all attributes having a Random Forest importance rank in the top 10% relative to all attributes in the combined dataset. The list is organized by the attribute symbol given in the first column. (Continued on subsequent pages)

Symbol [rs# for SNP]	Data Type	Attribute Name
AHR [rs7796976]	Genetic	Aryl hydrocarbon receptor
ALCAM	Proteomic	Activated leukocyte cell adhesion molecule
ANGPT4	Proteomic	Angiopoietin 4
APAF1 [rs2288729]	Genetic	Apoptotic peptidase activating factor
APOA4 [rs1042034]	Genetic	Apolipoprotein A-IV
BDNF	Proteomic	Brain-derived neurotrophic factor
BLC (CXCL13)	Proteomic	Chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)
BLM [rs235768]	Genetic	Bloom syndrome
BRCA1 [rs144848]	Genetic	Breast cancer 1, early onset
BTC	Proteomic	Betacellulin
BTG2	Proteomic	BTG family, member 2
CASR [rs1001179]	Genetic	Calcium-sensing receptor (hypocalciuric hypercalcemia 1)
CBR3 [rs881712]	Genetic	Carbonyl reductase 3
CCL1	Proteomic	Chemokine (C-C motif) ligand 1
CCL14	Proteomic	Chemokine (C-C motif) ligand 14
CCL16	Proteomic	Chemokine (C-C motif) ligand 16
CCR2 [rs1799865]	Genetic	Chemokine (C-C motif) receptor 2
CCR2 [rs4987053]	Genetic	Chemokine (C-C motif) receptor 2
CDKN1C [rs3731249]	Genetic	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)
CSF1 (MCSF)	Proteomic	Colony stimulating factor 1 (macrophage)
CSF1R	Proteomic	Colony stimulating factor 1 receptor
CSF2 (GMCSF)	Proteomic	Colony stimulating factor 2 (granulocyte-macrophage)
CSF3 (GCSF)	Proteomic	Colony stimulating factor 3 (granulocyte)
CTACK (CCL27)	Proteomic	Chemokine (C-C motif) ligand 27
CTH [rs473334]	Genetic	Cystathionase (cystathionine gamma-lyase)
CTH [rs515064]	Genetic	Cystathionase (cystathionine gamma-lyase)
CTH [rs663649]	Genetic	Cystathionase (cystathionine gamma-lyase)
CX3CL1	Proteomic	Chemokine (C-X3-C motif) ligand 1
CYP1A1 [rs2472299]	Genetic	Cytochrome P450, family 1, subfamily A, polypeptide 1
EGF	Proteomic	Epidermal growth factor (beta-urogastrone)
EOT (CCL11)	Proteomic	Chemokine (C-C motif) ligand 11
EOT2 (CCL24)	Proteomic	Chemokine (C-C motif) ligand 24
EOT3 (CCL26)	Proteomic	Chemokine (C-C motif) ligand 26

ERCC5 [rs1047768]	Genetic	Excision repair cross-complementing rodent repair deficiency, complementation group 5
FAS	Proteomic	Fas (TNF receptor superfamily, member 6)
FASLG (TNFSF6)	Proteomic	Fas ligand (TNF superfamily, member 6)
FASLG [rs929087]	Genetic	Fas ligand (TNF superfamily, member 6)
FGF1	Proteomic	Fibroblast growth factor 1 (acidic)
FGF2 (FGFB)	Proteomic	Fibroblast growth factor 2 (basic)
FGF4	Proteomic	Fibroblast growth factor 4 (Kaposi sarcoma oncogene)
FGF7	Proteomic	Fibroblast growth factor 7 (keratinocyte growth factor)
FST	Proteomic	Follistatin
GATA3 [rs10905277]	Genetic	GATA binding protein 3
GCP2 (CXCL6)	Proteomic	Chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)
GDNF	Proteomic	Glial cell derived neurotrophic factor
GRO2 (CXCL2)	Proteomic	Chemokine (C-X-C motif) ligand 2
GRO3 (CXCL3)	Proteomic	Chemokine (C-X-C motif) ligand 3
HGF	Proteomic	Hepatocyte growth factor (hepapoietin A; scatter factor)
HSD17B4 [rs384346]	Genetic	Hydroxysteroid (17-beta) dehydrogenase 4
HSD17B4 [rs7737181]	Genetic	Hydroxysteroid (17-beta) dehydrogenase 4
ICAM1	Proteomic	Intercellular adhesion molecule 1 (CD54), human rhinovirus receptor
ICAM3	Proteomic	Intercellular adhesion molecule 3
IFNG	Proteomic	Interferon, gamma
IGF1R	Proteomic	Insulin-like growth factor 1 receptor
IGF2	Proteomic	Insulin-like growth factor 2 (somatomedin A)
IGFBP1	Proteomic	Insulin-like growth factor binding protein 1
IGFBP2	Proteomic	Insulin-like growth factor binding protein 2, 36kDa
IGFBP3	Proteomic	Insulin-like growth factor binding protein 3
IGFBP4	Proteomic	Insulin-like growth factor binding protein 4
IL10	Proteomic	Interleukin 10
IL10 [rs1800871]	Genetic	Interleukin 10
IL13	Proteomic	Interleukin 13
IL15	Proteomic	Interleukin 15
IL15RA [rs859]	Genetic	Interleukin 15 receptor, alpha
IL17	Proteomic	Interleukin 17
IL1A	Proteomic	Interleukin 1, alpha
IL1B	Proteomic	Interleukin 1, beta
IL1RL1	Proteomic	Interleukin 1 receptor-like 1
IL1RN	Proteomic	Interleukin 1 receptor antagonist
IL2	Proteomic	Interleukin 2
IL2 [rs2069762]	Genetic	Interleukin 2
IL2 [rs2069763]	Genetic	Interleukin 2
IL2RA	Proteomic	Interleukin 2 receptor, alpha
IL2RB	Proteomic	Interleukin 2 receptor, beta
IL2RG	Proteomic	Interleukin 2 receptor, gamma (severe combined immunodeficiency)
IL3	Proteomic	Interleukin 3 (colony-stimulating factor, multiple)
IL4	Proteomic	Interleukin 4
IL4 [rs2070874]	Genetic	Interleukin 4
IL4 [rs2243250]	Genetic	Interleukin 4
IL4 [rs2243268]	Genetic	Interleukin 4
IL4 [rs2243290]	Genetic	Interleukin 4
IL5RA	Proteomic	Interleukin 5 receptor, alpha
IL6	Proteomic	Interleukin 6 (interferon, beta 2)
IL7	Proteomic	Interleukin 7
IL8	Proteomic	Interleukin 8
IL9	Proteomic	Interleukin 9

ITAC (CXCL11)	Proteomic	Chemokine (C-X-C motif) ligand 11
KDR	Proteomic	Kinase insert domain receptor (a type III receptor tyrosine kinase)
KIT (SCFR)	Proteomic	V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
KITLG (SCF)	Proteomic	KIT ligand
LEP	Proteomic	Leptin (obesity homolog, mouse)
LMO2 [rs2273797]	Genetic	LIM domain only 2 (rhombotin-like 1)
LTA	Proteomic	Lymphotoxin alpha (TNF superfamily, member 1)
LTN (XCL1)	Proteomic	Chemokine (C motif) ligand 1
MBL2 [rs11003125]	Genetic	Mannose-binding lectin (protein C) 2, soluble (opsonic defect)
MBL2 [rs1838066]	Genetic	Mannose-binding lectin (protein C) 2, soluble (opsonic defect)
MBL2 [rs5030737]	Genetic	Mannose-binding lectin (protein C) 2, soluble (opsonic defect)
MCP1 (CCL2)	Proteomic	Chemokine (C-C motif) ligand 2
MCP2 (CCL8)	Proteomic	Chemokine (C-C motif) ligand 8
MCP3 (CCL7)	Proteomic	Chemokine (C-C motif) ligand 7
MCP4 (CCL13)	Proteomic	Chemokine (C-C motif) ligand 13
MEC (CCL28)	Proteomic	Chemokine (C-C motif) ligand 28
MIG (CXCL9)	Proteomic	Chemokine (C-X-C motif) ligand 9
MIP1A (CCL3)	Proteomic	Chemokine (C-C motif) ligand 3
MIP1B (CCL4)	Proteomic	Chemokine (C-C motif) ligand 4
MIP1D (MAPKAP1)	Proteomic	Mitogen-activated protein kinase associated protein 1
MIP3A (CCL20)	Proteomic	Chemokine (C-C motif) ligand 20
MIP3B (CCL19)	Proteomic	Chemokine (C-C motif) ligand 19
MMP7	Proteomic	Matrix metallopeptidase 7 (matrilysin, uterine)
MMP9	Proteomic	Matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)
MPDU1 [rs2333227]	Genetic	Mannose-P-dolichol utilization defect 1
MPIF1 (CCL23)	Proteomic	Chemokine (C-C motif) ligand 23
MSH3 [rs3136228]	Genetic	MutS homolog 3 (E. coli)
MSH3 [rs32950]	Genetic	MutS homolog 3 (E. coli)
MTHFD2 [rs1667627]	Genetic	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2
MTHFR [rs1801133]	Genetic	5,10-methylenetetrahydrofolate reductase (NADPH)
MTR [rs1801394]	Genetic	5-methyltetrahydrofolate-homocysteine methyltransferase
MTRR [rs1802059]	Genetic	5-methyltetrahydrofolate-homocysteine methyltransferase reductase
NM	Proteomic	Neutrophil migration
NTF3	Proteomic	Neurotrophin 3
NTF5	Proteomic	Neurotrophin 5 (neurotrophin 4/5)
OSM	Proteomic	Oncostatin M
PAK6 [rs1136410]	Genetic	P21(CDKN1A)-activated kinase 6
PARC	Proteomic	P53-associated parkin-like cytoplasmic protein
PECAM1	Proteomic	Platelet/endothelial cell adhesion molecule (CD31 antigen)
PGF	Proteomic	Placental growth factor, vascular endothelial growth factor-related protein
PIN1 [rs4744]	Genetic	Protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting 1
RANTES (CCL5)	Proteomic	Chemokine (C-C motif) ligand 5
RERG [rs6488766]	Genetic	RAS-like, estrogen-regulated, growth inhibitor
RERG [rs767201]	Genetic	RAS-like, estrogen-regulated, growth inhibitor
SAT2 [rs3924313]	Genetic	Spermidine/spermine N1-acetyltransferase 2
SCUBE2 [rs1010324]	Genetic	Signal peptide, CUB domain, EGF-like 2
SDF1 (CXCL12)	Proteomic	Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)
SELL	Proteomic	Selectin L (lymphocyte adhesion molecule 1)
SLC39A2 [rs2234636]	Genetic	Solute carrier family 39 (zinc transporter), member 2
SLC6A3 [rs2070424]	Genetic	Solute carrier family 6 (neurotransmitter transporter, dopamine), member 3
SLC6A3 [rs6347]	Genetic	Solute carrier family 6 (neurotransmitter transporter, dopamine), member 3
TARC (CCL17)	Proteomic	Chemokine (C-C motif) ligand 17
TEP1 [rs1760898]	Genetic	Telomerase-associated protein 1

TGFA	Proteomic	Transforming growth factor, alpha
TIMP1	Proteomic	TIMP metallopeptidase inhibitor 1
TIMP2	Proteomic	TIMP metallopeptidase inhibitor 2
TNF	Proteomic	Tumor necrosis factor (TNF superfamily, member 2)
TNFRSF10A	Proteomic	Tumor necrosis factor receptor superfamily, member 10a
TNFRSF10D	Proteomic	Tumor necrosis factor receptor superfamily, member 10d
TNFRSF11A	Proteomic	Tumor necrosis factor receptor superfamily, member 11a, NFKB activator
TNFRSF14 (HVEM)	Proteomic	Tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)
TNFRSF1A	Proteomic	Tumor necrosis factor receptor superfamily, member 1A
TNFRSF21 (DR6)	Proteomic	Tumor necrosis factor receptor superfamily, member 21
TNFSF7 (CD27)	Proteomic	Tumor necrosis factor (ligand) superfamily, member 7
TNFSF8 (CD30)	Proteomic	Tumor necrosis factor (ligand) superfamily, member 8
TSG101 [rs2045224]	Genetic	Tumor susceptibility gene 101
TSG101 [rs2045224]	Genetic	Tumor susceptibility gene 101
TSG101 [rs2045224]	Genetic	Tumor susceptibility gene 101
VEGF	Proteomic	Vascular endothelial growth factor

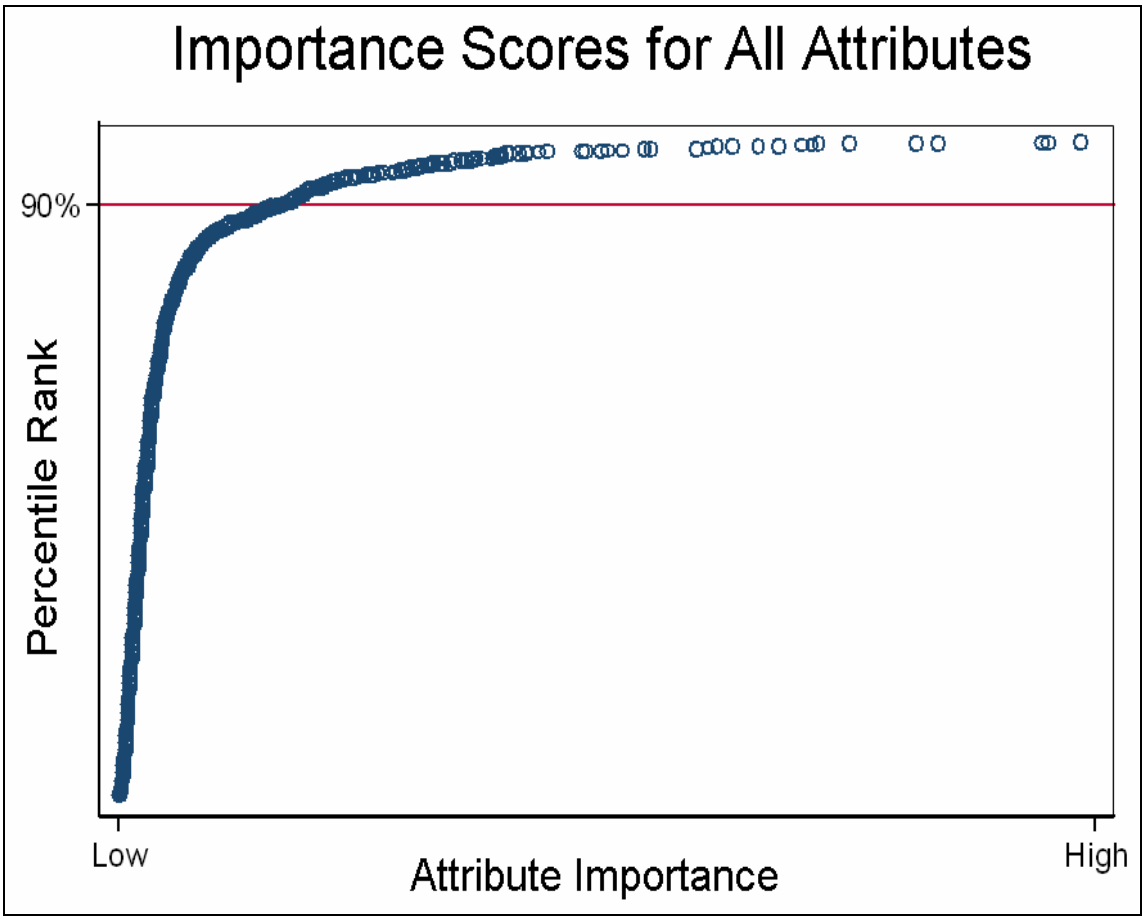


Figure 2: Attribute importance "landscape" showing the shape of the importance curve ranking all attributes in the combined (genetic plus proteomic) dataset. Attributes above the horizontal line indicate a relative importance rank in the top 10% (90th percentile) of all attributes in the dataset. Attributes of high importance resulted in greater reduction of impurity (Gini) than attributes of low importance, as measured by the out-of-bag importance procedure.

Modeling the association of genetic and proteomic biomarkers with adverse events

Having filtered out the noise using Random Forests, 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 attributes. 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 IL-4, and change in CSF-3 concentration. Imposing our minimum of five individuals per terminal (AE status) leaf, this tree correctly classifies 89% of individuals (with 5 AE misclassifications and 2 non-AE misclassifications).

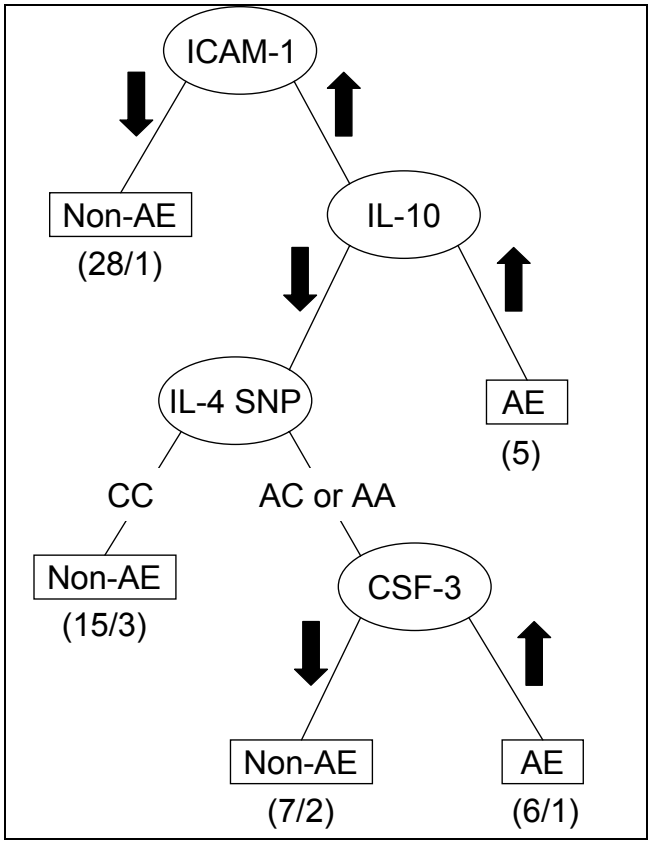


Figure 3: Final model of genetic and proteomic factors contributing to AE development. Each node (oval) constitutes a decision point based upon the genotype of genetic attributes (IL-4 SNP) or whether the concentration change from baseline in proteomic attributes (ICAM-1, IL-10, CSF-3) is above (upward-pointing arrows) or below (downward-facing arrows) a threshold (calculated by choosing the most informative value from among a set of possible values generated for each particular split). Starting at the root node (ICAM-1), subjects are 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 characterizes the biological relationships among the attributes in the tree using Chilibot. Interactive relationships are characterized into one of three types based upon the verbs connecting pairs of attributes in the biomedical literature: 1. stimulatory relationships are connected by verbs such as “activate”,

“stimulate”, or “enhance”, 2. inhibitory relationships are connected by verbs such as “decrease”, “attenuate”, or “inhibit”, and relationships are characterized as neutral when the nature of the relationship cannot be contextually determined. Mining the biomedical literature suggested interactive relationships connecting all of the attribute nodes in our final model. Stimulatory, inhibitory, and neutral pair-wise interactive relationships were identified between each of ICAM-1, IL-10, IL-4, and CSF-3. Thorough examination of the networks inferred facilitated the biological interpretation of the final model discussed below.

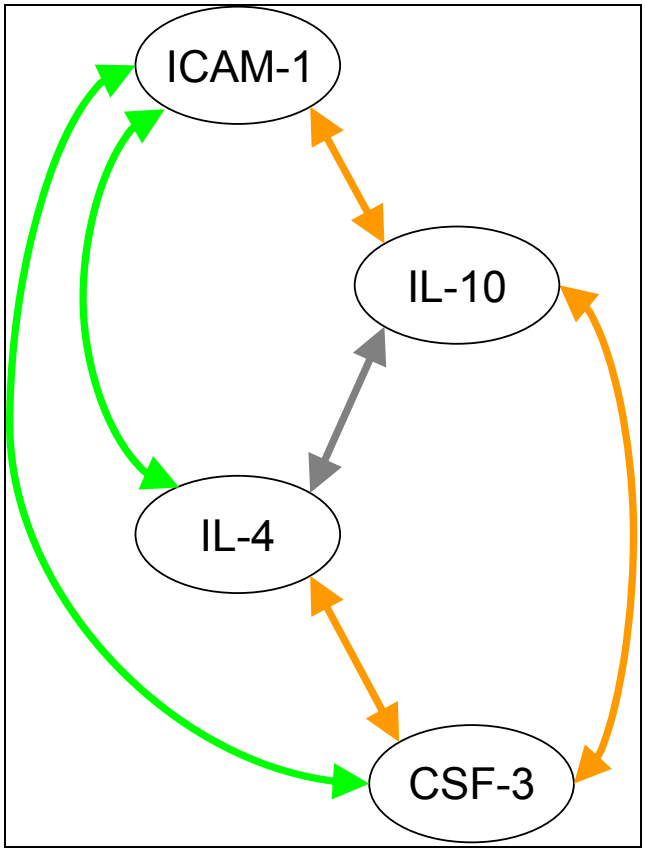


Figure 4: Biological relationships among the attributes in our final model characterized using Chilibot. Connections between each attribute node (oval) are colored according to the type of interactive relationship they represent: stimulatory (green), both stimulatory and inhibitory (orange), or neutral (gray). Arrows indicate that interactions between particular biological attributes are bi-directional. For each connection, 50 abstracts containing both terms were processed to determine the nature of interactive relationship.

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, since outcomes such as AE development likely result from the interplay of multiple genetic, proteomic, and environmental factors [27,28]. The decision tree trained on the attributes passing our RF filter proposes a solid biological model of adverse event development.

The attributes included in this tree point to an important role of one particular immune cell type: monocytes. Monocytes are bone marrow-derived circulating blood cells that are precursors of tissue macrophages. Monocytes are actively recruited to sites of inflammation, where they differentiate into macrophages in tissues. These macrophages play important roles in both innate and adaptive immune responses. Macrophages are activated by microbial products such as endotoxin and by T cell cytokines such as IFN-γ. Activated macrophages phagocytose and kill microorganisms, secrete pro-inflammatory cytokines, and present antigen to helper T cells. Macrophages assume different morphologic forms in different tissues, which might have an important impact in system-wide responses such as the AEs studied here.

The root node of the tree 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. Additionally, ICAM-1 expression is upregulated in mature monocytes [29], 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 (VCAM-1) or intercellular adhesion molecule 1 (ICAM-1), which interact with integrins on the monocyte surface. Once the monocyte is tightly bound, it then migrates between endothelial cells [30,31]. High levels of ICAM-1 might indicate an "overrecruitment" of monocytes into tissue, triggering an unnecessarily active innate inflammatory response.

For individuals with large positive changes in ICAM-1, the next node in the tree is IL-10, where changes from baseline greater than 85% are associated with AEs. IL-10 is produced by activated macrophages and some helper T cells whose major function is to inhibit activated macrophages and therefore 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. Since our cytokine levels are measured within one week of immunization, the high levels of IL-10 secreted into the systemic compartment (serum) might indicate an overabundance of activated macrophages during the acute phase contributing to AE development. Eventually, sufficiently high levels of IL-10

should “calm” the macrophage response, so if cytokines were measured at a later time point (*e.g.* two weeks post-immunization), it is probable that IL-10 levels would return toward baseline. Additionally, high levels of IL-10 have been shown to inhibit the production of other cytokines by monocytes [32], implying that monocytes may not be recruiting proper T helper cell response to balance the acquired and innate reactions.

For individuals with mild changes in IL-10 concentration, the next node is a SNP in the gene encoding IL-4. IL-4 is a cytokine produced mainly by the TH_2 subset of $CD4^+$ helper T cells whose functions include induction of differentiation of TH_2 cells from naïve $CD4^+$ precursors, stimulation of IgE production by B cells, and suppression of IFN-γ-dependent macrophage functions [33,34]. While 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 IL-4 achieved nearly identical importance scores indicates that there may be LD blocks of variation within the IL-4 gene region associated with AE development (see Chapter IV). Because of the intricate cross-talk between macrophages and the TH_2 response in maintaining homeostasis, it is plausible that the major IL-4 genotype (CC) is associated with calming the activated macrophage response and directing the acquired immune system to progress in response to vaccine presentation, while 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 IL-4, the lowest node of the tree is CSF-3 (GCSF). GCSF is a cytokine produced by activated T cells, macrophages, and endothelial cells at sites of infection that acts on bone marrow to increase production of and mobilize 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 implies another over-recruitment in the development of AEs, as neutrophils have been associated with host tissue damage and failure to terminate acute inflammatory responses [35]. This over-reaction is consistent with the types of AE symptoms observed in the current study and with the overall proposed biological mechanisms of AE development.

The results of this study provide a viable biological mechanism 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 [36]. 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.

Together with previous studies on immunological response to smallpox vaccination, we conclude that AEs 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: ICAM-1, IL-10, IL-4, and CSF-3. Since 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 adverse events. 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 mechanism of AE development that targets specific players within the systemic inflammatory pathway for further study.

Future studies will test our hypothesis at the bench. The functional consequences of genetic variability in IL-4 related to bioavailability and overall concentration must be fully characterized. 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 is needed on the effects of these cytokines in other physiological compartments. It is hoped that this study will convince all future work on this subject to adopt an experimental approach that rightfully takes the broader spatial and temporal physiological context of complex biological systems into account.

Acknowledgments

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).

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. Breiman L: Random forests. *Machine Learning* 2001, 45: 5-32.

3. Lunetta KL, Hayward LB, Segal J, Van EP: Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet* 2004, 5: 32.

4. Robnik-Sikonja M: Improving random forests. *Machine Learning: Ecml 2004, Proceedings* 2004, 3201: 359-370.

5. 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.

6. Rock MT, Yoder SM, Talbot TR, Edwards KM, Crowe JE, Jr.: Adverse events after smallpox immunizations are associated with alterations in systemic cytokine levels. *J Infect Dis* 2004, 189: 1401-1410.

7. Rock MT, Yoder SM, Talbot TR, Edwards KM, Crowe JE, Jr.: Cellular Immune Responses to Diluted and Undiluted Aventis Pasteur Smallpox Vaccine. *J Infect Dis* 2006, 194: 435-443.

8. 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.

9. 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.

10. Chanock SJ. Core Genotyping Facility. National Cancer Institute. http://cgf.nci.nih.gov/home.cfm 2004. Gaithersburg, MD, USA.

11. 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.

12. 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.

13. 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 U S A* 2000, 97: 10113-10119.

14. 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.

15. Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and Regression Trees*. New York: Chapman & Hall; 1984.

16. Province MA, Shannon WD, Rao DC: Classification methods for confronting heterogeneity. *Adv Genet* 2001, 42: 273-286.

17. Breiman L: Random forests. *Machine Learning* 2001, 45: 5-32.

18. 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.

19. McKinney BA, Reif DM, Ritchie MD, Moore JH: Machine Learning for Detecting Gene-Gene Interactions: A Review. *Appl Bioinformatics* 2006, 5: 77-88.

20. Witten IH, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. San Francisco, California: Morgan Kaufmann; 2005.

21. Ihaka R, Gentleman R: R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 1996, 5: 299-314.

22. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org 2006. Vienna, Austria.

23. Breiman L, Cutler A. Random Forests. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm 2004.

24. Reif DM, Motsinger AA, McKinney BA, Crowe JE, Jr., Moore JH: Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* 2006, in press.

25. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 2004, 5: 147-160.

26. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005, 21: 263-265.

27. Moore JH: The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003, 56: 73-82.

28. Wilke RA, Reif DM, Moore JH: Combinatorial pharmacogenetics. *Nat Rev Drug Discov* 2005, 4: 911-918.

29. 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.

30. 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.

31. 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.

32. 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.

33. Mangan DF, Robertson B, Wahl SM: IL-4 enhances programmed cell death (apoptosis) in stimulated human monocytes. *J Immunol* 1992, 148: 1812-1816.

34. 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.

35. Serhan CN, Savill J: Resolution of inflammation: the beginning programs the end. *Nat Immunol* 2005, 6: 1191-1197.

36. Hood L: Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev* 2003, 124: 9-16.

CHAPTER VII

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation describes the development and implementation of a comprehensive analysis paradigm wherein experimental data of multiple types is integrated for the study of complex phenotypes. This strategy was applied to genetic and proteomic data in both simulated and real analysis situations. The successful application to combined genetic and proteomic data from smallpox vaccine studies supports the hypothesis that such integrated approaches provide a comprehensive portrayal of the mechanisms underlying complex phenotypes and lend confidence to the biological interpretation of analytical conclusions.

The next steps in elucidating the development of adverse events after smallpox vaccination will involve testing the candidate biomarkers at the bench—taking into account their respective genetic or proteomic context. The functional consequences of genetic variability in IL-4, IRF-1, and MTHFR must be characterized with respect to bioavailability, activity, and overall concentration. Functional genetic studies should be carried out in experimental conditions that stimulate the inflammatory pathways highlighted by the work presented here. Focused studies should be undertaken to describe the variability in all of the AE-associated genomic regions, especially for those in which multiple SNPs were identified.

Dense time-series studies are needed to clarify the dynamic interplay between the signaling of ICAM-1, IL-10, CSF-3, eotaxin, MIG, TIMP-2, and SCF, as well as the protein products of IL-4, IRF-1, and MTHFR. Proteomic studies should be performed in environments wherein the relevant genetic background has been established. Additional data is needed on the effects of these cytokines in other physiological compartments outside the serum.

The ultimate test of these results will be their assessment in large-scale, independent cohorts. From an epidemiological perspective, the studies discussed here involve relatively small samples and represent a very narrow slice of demographic characteristics such as race and age. Future studies will need to evaluate whether these conclusions generalize to populations at-large, or if they only apply to certain subsets.

Perhaps the most important aspect of future studies would be the collection of additional types of information on study subjects. Besides the other types of data that could be collected on the genetic and proteomic levels (*e.g.* genomic methylation status and enzymatic activity, respectively), information on circulating mRNA concentrations, immunological effector cell morphologies, and spatial bioactivity may prove useful. Outside of biological data, information on subjects' lifestyles, dietary intakes, and any other plausible environmental factors should be gathered. For any type of data collected, careful consideration must be given to the particular variables measured—these results are suggestive of particular immunological pathways, and variables should be selected to provide comprehensive coverage of variation in these pathways.

Once these additional data are available, extension and refinement of analytical methods can proceed. Simulation studies should assess alternative variable selection strategies within the RF framework, such as joint variable permutation methods. Other promising analytical methods, such as those employing evolutionary computation or prior domain knowledge, should also be explored as techniques for integrating multiple data types.

As the number of data types increases, an important issue arises with respect to effectively integrating massive amounts of disparate information. Statistical modeling techniques aside, meaningfully interpreting the results of multifaceted models demands expertise in each of the experimental domains considered. While this will foster the evolution of interdisciplinary research, tools will be needed that allow communication in a standardized manner. The development of such tools depends on the adoption of consistent language so that databases can present information that is standardized across both experiments and disciplines.

Considering the rapid progress in experimental technologies able to reliably generate vast quantities of data, as well as continual improvements in cost efficiency, it is expected that comprehensive datasets—including multiple types of experimental information—will become commonplace in the near future. It is hoped that the positive conclusions from this dissertation will help spur the adoption of an experimental approach that rightfully takes the broader spatial and temporal physiological context of complex biological systems into account.