

## ORIGINAL ARTICLE

# Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine

NA Davis<sup>1</sup>, JE Crowe Jr<sup>2</sup>, NM Pajewski<sup>3</sup> and BA McKinney<sup>1</sup>

<sup>1</sup>Department of Mathematical and Computer Sciences, University of Tulsa, Tulsa, OK, USA; <sup>2</sup>Departments of Pediatrics, Microbiology and Immunology, Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, TN, USA and <sup>3</sup>Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA

*The variation in antibody response to vaccination likely involves small contributions of numerous genetic variants, such as single-nucleotide polymorphisms (SNPs), which interact in gene networks and pathways. To accumulate the bits of genetic information relevant to the phenotype that are distributed throughout the interaction network, we develop a network eigenvector centrality algorithm (SNPrank) that is sensitive to the weak main effects, gene–gene interactions and small higher-order interactions through hub effects. Analogous to Google PageRank, we interpret the algorithm as the simulation of a random SNP surfer (RSS) that accumulates bits of information in the network through a dynamic probabilistic Markov chain. The transition matrix for the RSS is based on a data-driven genetic association interaction network (GAIN), the nodes of which are SNPs weighted by the main-effect strength and edges weighted by the gene–gene interaction strength. We apply SNPrank to a GAIN analysis of a candidate-gene association study on human immune response to smallpox vaccine. SNPrank implicates a SNP in the retinoid X receptor  $\alpha$  (RXRA) gene through a network interaction effect on antibody response. This vitamin A- and D-signaling mediator has been previously implicated in human immune responses, although it would be neglected in a standard analysis because its significance is unremarkable outside the context of its network centrality. This work suggests SNPrank to be a powerful method for identifying network effects in genetic association data and reveals a potential vitamin regulation network association with antibody response.*

Genes and Immunity (2010) 11, 630–636; doi:10.1038/gene.2010.37; published online 8 July 2010

**Keywords:** genetic association study; gene–gene interaction; single-nucleotide polymorphism; information theory; eigenvector centrality; Markov chain

## Introduction

The availability of high-throughput technology for capturing DNA sequence variation has spawned the growing field of vaccinomics, which attempts to interrogate the role of heritable host genetic variation on vaccine immune response and reactogenicity.<sup>1,2</sup> Although the scope of genotyping continues to grow increasingly comprehensive, most analyses to date in vaccine applications have focused only on the marginal effects of host polymorphisms, that is, the effect of genetic variants in isolation from other genetic and environmental factors. Recent genome-wide association studies have identified hundreds of variants associated with common diseases and traits; however, these genetic variants only account for a small amount of the inherited variation, leaving much of the variation to be discovered.<sup>3</sup>

In a recent study, we developed a more global approach for characterizing single-nucleotide polymorphisms (SNPs) in a genetic association study by treating the system as a phenotype-specific genetic association interaction network (GAIN).<sup>4</sup> Small amounts of the inherited variation are encoded in numerous SNP nodes throughout this network in the form of main effects and gene–gene interactions. In this study, we introduce SNPrank, which we conceptualize as a random SNP surfer (RSS) that circulates through the GAIN, accumulating bits of information from each SNP with regard to the phenotype. SNPrank represents a modified eigenvector centrality algorithm that ranks the importance of each SNP through the complex interactions encoded in the GAIN.

The notion of centrality concerns the importance of a node in a network. For a social network, centrality quantifies how central or influential a person is in the network. Degree centrality simply counts the number of connections that a node has in a network (equation 4). However, not all incoming connections are equally important: some SNPs may exhibit strong interactions that yield considerable information about the phenotype, whereas other SNPs may show weak interactions but

Correspondence: Professor BA McKinney, Department of Mathematical and Computer Sciences, University of Tulsa, Keplinger Hall, U348, 800 South Tucker Drive, Tulsa, OK 74104, USA.  
E-mail: brett.mckinney@gmail.com  
Received 2 March 2010; revised and accepted 05 May 2010; published online 8 July 2010

with numerous connections (a hub) that influence the phenotype. Still other SNPs may not be highly connected but may be highly relevant to the phenotype, analogous to someone being an authority on a given topic. Eigenvector centrality—the basis of Google’s well-known PageRank algorithm for ranking web pages<sup>5</sup>—balances the quantity of connections with the quality. The aim of the SNPrank eigenvector centrality algorithm presented here is to balance the main effect and network interaction effects in the genotype–phenotype relationship encoded in the GAIN. Moreover, by accounting for indirect connections in the recursion process, eigenvector centrality allows for the possibility of including higher-order interactions when ranking the relevance of a particular variant.

Just as the connectivity of a person’s social network depends on context—such as family, friends or business relationships—there is growing awareness that genetic effects will exhibit differences in network connectivity that depend on contexts such as the environment, the organism or particular antigen perturbation in a vaccine.<sup>4</sup> Most current network approaches neglect the phenotypic context when defining connections between genes. For example, traditional coexpression networks do not account for the context of the phenotype because connections are based on the correlation between biomarkers across all phenotype states.<sup>6</sup> Recent studies have suggested that improvements in predicting biological connections may be gained when the context is restricted to a more relevant space. For example, when mining gene expression databases for coexpression, Oldham *et al.*<sup>6</sup> found that restricting inputs to biologically relevant sample types yields a more informative network. In their case study, restricting the database to gene expression experiments involving relevant tissue or sample types allowed them to identify known functional genes that are absent from the network generated from all available samples. Similar improvements have been suggested in predicting the biological relationships between diseases from the human phenotype ontology as the resolution of these ontologies become finer grained.<sup>7</sup> The data-driven GAIN network approach used in this study is not biased by restricting correlation to a particular phenotype state, but rather calculates the interaction between SNPs, which is a conditional correlation that conditions all phenotype states in the experiment.

The quality of SNPrank importance scores will depend critically on the quality of the starting network. A familiar analogy with social networks further motivates the need for a data-driven, context-sensitive network. Consider a hypothetical network in which actors represent SNPs and the context of acting in a drama represents immune response to vaccine phenotype. If we considered the context of acting in a comedy instead of in a drama, the network connections and the relative importance of nodes will change somewhat. Similarly, we would expect a different network and SNP importance scores if we considered the context of immunization with influenza instead of smallpox. We would expect an even more divergent social network if we considered the context of acting in a musical; similarly, a SNP network for vaccine immune response would look very different from a network for major depressive disorder. Therefore, an advantage of the data-driven

network constructed by GAIN is that it includes the context of the phenotype when defining gene–gene interactions between SNP nodes (equation 1).

When ranking the importance of a webpage URL  $U$ , Google’s PageRank not only takes into account the number of connections that  $U$  has but also considers the connectivity of the pages connected to  $U$ . This can be understood through the random surfer interpretation of PageRank in which web-surfing behavior is simulated, with the surfer either following the link structure of the Internet or going to a random page. The more frequently the random surfer lands on a page, the higher the page’s rank. Similarly, SNPrank simulates an RSS that walks through the GAIN Markov chain. The more often the RSS lands on a SNP, the higher its SNPrank. To guard against human manipulation of the rank of a web page, PageRank does not reward pages with self-referential links. In contrast, SNPrank includes self-connections that arise from the independent relevance of that SNP to the phenotype. In the connectivity matrix, these self-connections are calculated using the SNP’s information gain  $I_2$  with respect to the phenotype (equation 2). Self-connections represent the probability of the RSS to stay at the current node during a simulation step. GeneRank is an eigenvector centrality method for gene expression data that uses either a coexpression network or a gene ontology (GO) annotation network to define connections between genes.<sup>8</sup> The GeneRank Markov chain incorporates single-marker differential expression information, but the connectivity matrix is not sensitive to the context of the phenotype. SNPrank uses a data-driven, phenotype-sensitive network of main effects and gene–gene interactions inferred using the recently developed GAIN tool.<sup>4</sup>

In summary, the SNPrank transition matrix  $T$  (equation 5) accounts for single-locus main effects through the information gain  $I_2$  (equation 2) along the diagonal of the GAIN matrix  $G$  (equation 3). It considers pairwise interactions through the interaction gain  $I_3$  (equation 1) on the off-diagonal elements of  $G$  and higher-order interactions through the network sampling by the RSS. To continue the social network analogy with the phenotype analogous to acting in a drama, the aim of our importance score is (A) to identify actors who are important in the network because of their ability to connect to other actors in dramas, such as Kevin Bacon, and (B) to identify actors who have strong individual importance to dramatic acting, such as Marlon Brando. SNPrank is designed to identify SNPs that are important to the phenotype because of their ability to connect to other SNPs (Bacon SNPs), as well as SNPs that are important because of their individual importance to the phenotype (Brando SNPs).

## SNPrank

In the following sections, we describe the mathematical machinery and steps of the SNPrank algorithm. In brief, we first describe the information-theoretic GAIN connectivity matrix. Then, on the basis of GAIN, we construct the SNPrank Markov chain, which the RSS surfs. The construction of the Markov chain is constrained by the need to balance main and interaction effects and by the requirement that the Markov chain be

a stochastic matrix. Finally, having satisfied these constraints, we are able to use the iterative power method to solve for the SNPrank eigenvector. The components of the eigenvector are the SNPrank scores of the corresponding SNPs. The software implementing GAIN and the proposed SNPrank algorithm is available at <http://insilico.utulsa.edu/Software>.

#### GAIN connectivity matrix

The performance of SNPrank depends on the quality and context sensitivity of the underlying network. Thus, we characterize the main effects and interactions among SNPs by inferring a GAIN for a particular genetic association study. GAIN is based on the interaction information,  $I_3$ , between three variables (in this case, between two SNPs  $A$  and  $B$  and the class/phenotype variable  $C$ ):

$$I_3(A; B; C) = I_2(AB; C) - I_2(A; C) - I_2(B; C) \quad (1)$$

The quantities  $I_2(A; C)$  and  $I_2(B; C)$  are the information gained about the class/phenotype  $C$  when locus  $A$  or locus  $B$ , respectively, are measured. The information gain between two discrete random variables  $X$  and  $Y$  is given by

$$I_2(X; Y) = \sum_{x \in X} \sum_{y \in Y} p_{XY}(x, y) \log \left( \frac{p_{XY}(x, y)}{p_X(x)p_Y(y)} \right) \quad (2)$$

The quantity  $AB$  in equation 1 is a joint attribute constructed from attributes  $A$  and  $B$ , with states given by the Cartesian product of the states of  $A$  and  $B$ .  $I_3$  is then the gain in information obtained about the phenotype when considering  $A$  and  $B$  jointly ( $I_2(AB; C)$ ) above the phenotype information that would be gained by considering variables  $A$  and  $B$  independently ( $I_2(A; C) + I_2(B; C)$ ). We use  $I_3$  as the connection strength of each edge in the GAIN. Thus, each edge represents the increase in information about the phenotype achieved by considering the two SNPs jointly, compared with the expected increase in information with the assumption of independence between SNPs. We emphasize that a connection between SNPs in GAIN is specific to the given phenotype because it measures the correlation between two SNPs that influences association with the phenotype. In the current work, we weigh the nodes and edges of the connectivity matrix using information theoretic quantities. However, other data-driven weighting schemes are compatible with SNPrank, such as the main effect and pairwise interaction estimates from regression models. One may also augment or replace the data-driven GAIN connectivity matrix with a matrix based on gene annotation or pathway knowledge.

The GAIN connectivity matrix, with weights  $g_{ij}$ , will be used below to construct the probability for the RSS to make a transition from SNP $_i$  to SNP $_j$  in the network. Off-diagonal weights are defined as the interaction,  $I_3$ , between two distinct SNPs, conditional on the phenotype:

$$g_{ij} = \begin{cases} I_3(\text{SNP}_i; \text{SNP}_j; \text{Phenotype}), & i \neq j \\ I_2(\text{SNP}_i; \text{Phenotype}), & i = j \end{cases} \quad (3)$$

and the diagonal is a vector of information gains,  $I_2$ , or main effect associations of each SNP with the phenotype.

#### SNPrank Markov chain

We now derive a new variant of the eigenvector centrality algorithm to determine the importance of SNPs in a genetic association study by transforming

the GAIN matrix into a Markov chain transition matrix,  $T$ . The construction involves multiple scalings to ensure that the final matrix  $T$  (equation 5 below) is a stochastic matrix (that is,  $\sum_{i=1}^N t_{ij} = 1$ , where  $N$  is the number of SNP nodes) so that the recursion procedure will converge. We begin the construction by considering the elements of the GAIN matrix (equation 3), which are used to weigh the probability of the RSS to make a transition from SNP $_i$  to SNP $_j$  in the network. We scale the elements of the GAIN matrix by column sums, which are the out-degree association fluxes of each SNP:

$$d_j = \sum_{i=1}^N g_{ij}. \quad (4)$$

The PageRank matrix includes a probability  $\gamma$  to follow direct connections,  $\gamma g/d$ , and a corresponding term with the probability  $(1-\gamma)$  that a random page will follow. The latter may be interpreted as a fatigue term that models an Internet surfer that has grown tired of the current thread of links. Similar to GeneRank, we modify this probability to be proportional to the correlation of each SNP with the phenotype. But rather than use differential expression, we use the information gain of each SNP, which is given by the diagonal elements of  $G$ ,  $g_{ii}$ . This requires added scaling, involving the trace of the  $G$  matrix to constrain  $T$  to be a stochastic matrix. Also note that in place of the  $g_{ii}$  term, one could use expert knowledge if one wished to enrich for certain biological pathways. Depending on the interaction gain ( $I_3$ ) cutoff used, some paths in the network may lead to closed cycles or absorbing states. An absorbing state occurs when a jump takes place to a SNP  $j$  that has nowhere to go, that is,  $d_j = 0$ . Again, rather than use a uniform probability when  $d_j = 0$ , we weigh the transition to another SNP by the information gain  $I_2$  of each SNP (the diagonal elements of  $G$ ). The elements of the stochastic transition matrix  $T$  become

$$t_{ij} = \begin{cases} \gamma \frac{g_{ij}}{d_j} + \frac{(1-\gamma)}{\text{Tr}(G)} g_{ii}, & d_j \neq 0 \\ \frac{g_{ii}}{\text{Tr}(G)}, & d_j = 0 \end{cases} \quad (5)$$

where  $\gamma$  is the probability of the RSS to follow a gene-gene interaction-weighted path in the network and  $1-\gamma$  is the probability of the RSS to remain at a SNP weighted by the main effect strength. We use  $\gamma = 0.85$  in our data analysis.

#### SNPrank eigenvector scores

Beginning with some arbitrary vector  $r^{(0)}$ , the RSS random walk is simulated by the repeated application of the transition matrix  $T$ :

$$r^{(k+1)} = Tr^{(k)} \quad (6)$$

The frequency with which the RSS lands on a SNP node during a simulation through the network is related to the SNPrank importance score, which is obtained by normalizing the frequencies to yield a probability field. The final SNPrank score of SNP $_i$  is given by the  $i$ th element of the stationary probability vector,  $r_s$ , which is the vector whose the elements do not change under application of the transition matrix:

$$r_s = Tr_s \quad (7)$$

Typically,  $r_s$  is obtained in the limit of a large number of transitions  $k$ ; however, SNPrank only requires a few recursion steps to converge for our data set. Finding  $r_s$  reduces the problem of finding the eigenvector with

eigenvalue  $\lambda=1$ , which can be solved by the power method.<sup>9</sup> The Perron–Frobenius theorem ensures that the eigenvector  $r_s$  exists, and that the largest eigenvalue associated with the stochastic matrix is always 1. The power method recursively applies equation 6, with  $T$  defined by equation 5, until the eigenvalue converges to 1 to within some small tolerance. Below is the pseudo-code outline for the power method of calculating the SNPrank eigenvector.

1. Initialize SNPrank eigenvector  $r^{(0)}$ , and tolerance  $\varepsilon$
2. Begin Power Method
  - $k=0$
  - while  $|r^{(k+1)} - r^{(k)}| < \varepsilon$
  - normalize vector  $r^{(k)}$
  - $r^{(k+1)} = Tr^{(k)}$
  - $k=k+1$
3. final eigenvector:  $r_s = r^{(k)}$
4. SNPrank of  $i$ th SNP =  $r_s(i)$

Fast convergence is found for the eigenvalue when the probability elements of the initial eigenvector  $r^{(0)}$  are chosen to be uniform,  $1/N$ ,  $N$  being the number of SNPs. More informed initial guesses, such as the normalized evaporative cooling (EC) feature selection scores<sup>4,10</sup> or the many variants of Relief-F,<sup>11,12</sup> may further speed up convergence.

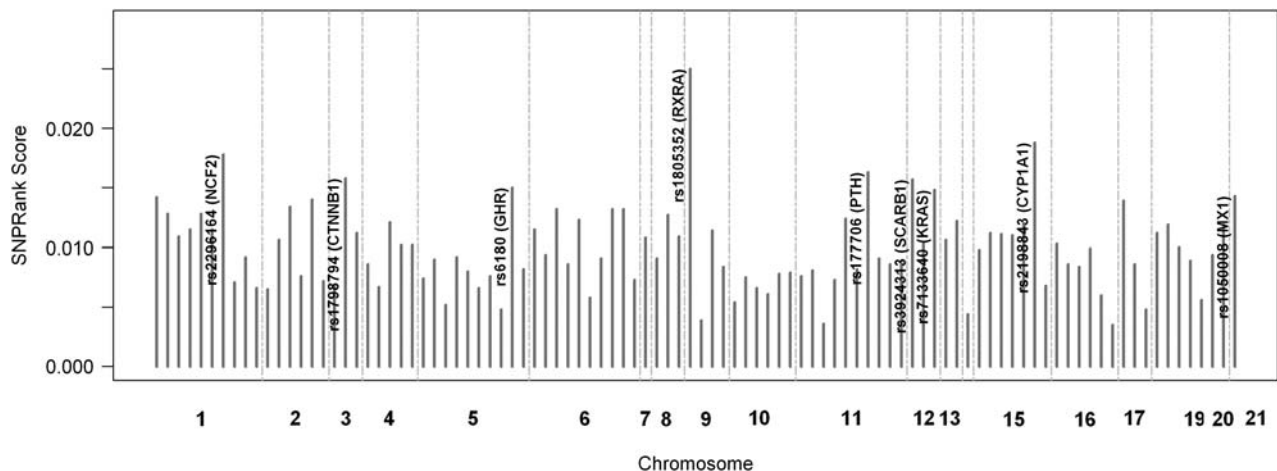
### Application to smallpox vaccine antibody response

We expect a combination of genetic main effects and interactions to influence the immune response to vaccine.<sup>13</sup> Thus, we illustrate SNPrank using SNP data from a study of the human immune response after smallpox vaccination. Genotyping was performed using a custom SNP panel based on the NCI SNP500 Cancer project that has been described previously.<sup>14</sup> The majority of SNPs included in the panel target soluble factor mediators and signaling pathways, many of which have immunological significance. Of the 1536 SNPs assayed, a total of 1442 genotypes passed standard quality control filters (minor allele frequency  $>0.01$ , Hardy–Weinberg equilibrium  $P$ -value  $>0.001$ , and SNP call rate  $>0.95$ ).

Previously, we used this panel to investigate the association between SNPs and adverse event phenotypes after smallpox vaccination.<sup>4,15,16</sup> In the current study, we investigate associations in this SNP panel with post-vaccination antibody response. Specimens were collected from 136 smallpox vaccine-naïve volunteers at baseline before vaccination and at 28 days after vaccination. Serum-neutralizing antibody responses were measured by plaque reduction neutralization assay as described previously.<sup>17</sup> We used mixture model density estimation to categorize the neutralizing antibody phenotype into high and low immune response. This Gaussian discretization yielded 76 low and 60 high immune responders.

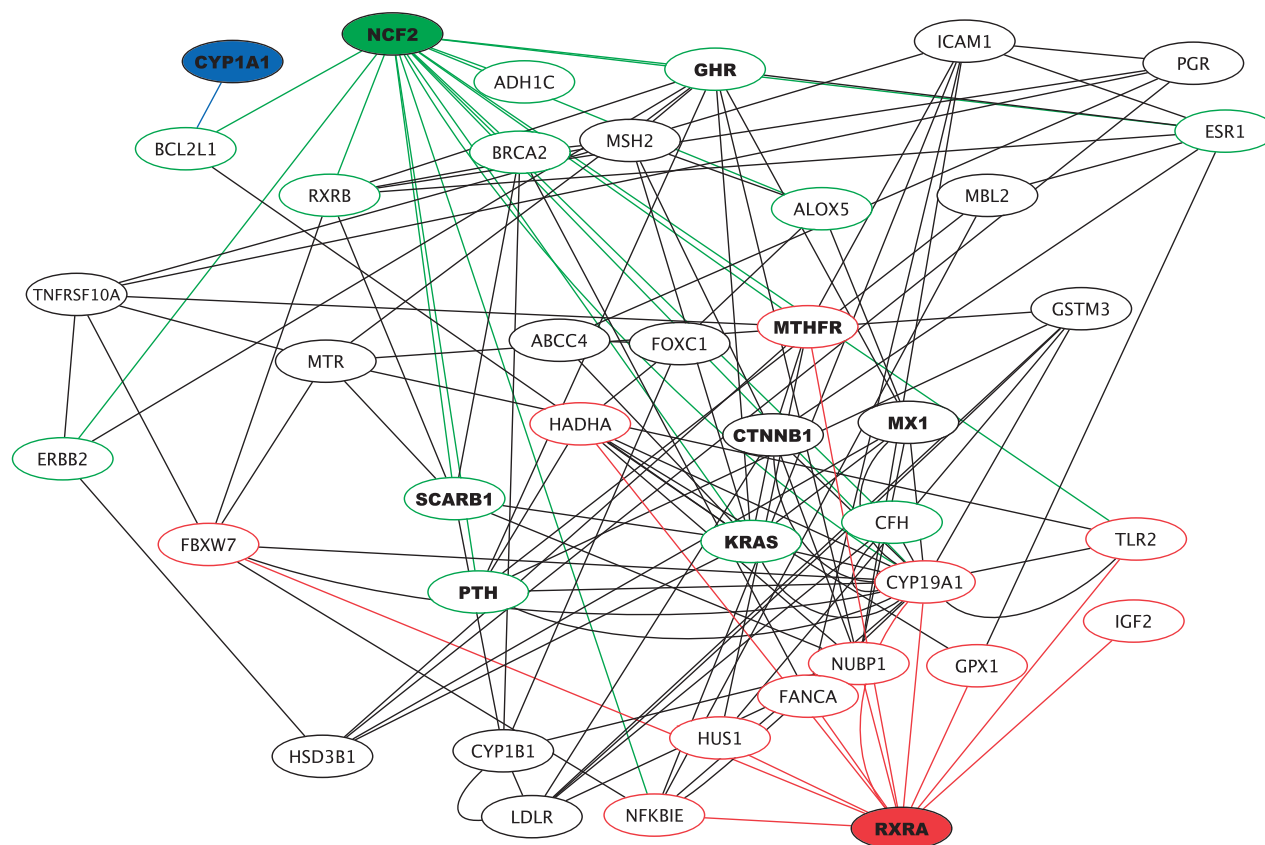
For our primary SNPrank analysis, we filtered the 1442 SNPs to the top 100 using EC feature selection<sup>4</sup> to allow for the combination of main and interaction effects in the initial screen. EC is a machine-learning algorithm that combines the ability of Random Forests to detect main effects with the ability of Relief-F to detect interactions. Our simulation studies have shown that EC greatly boosts the power of Random Forests and Relief-F to detect a combination of effects that may be present in a real data set. More details may be found in the study by McKinney and colleagues.<sup>4,10</sup> Filtering with a univariate approach, such as logistic regression, runs the risk of eliminating interacting SNPs. However, as a secondary analysis to show the effectiveness of SNPrank in conjunction with a conventional genetic association analysis, we also filtered SNPs through marginal hypothesis tests based on logistic regression implemented in PLINK.<sup>18</sup> We assumed an additive model of inheritance in all models, subsequently using the top 100 SNPs ranked according to  $P$ -value to generate the GAIN for SNPrank. The filtered SNPs from both approaches were highly congruent. We present results using the EC filter because of its ability to capture a spectrum of genetic effects. From the filter list, we apply GAIN to construct the data-driven, information theoretic interaction network, and then we apply the SNPrank algorithm.

Figure 1 shows a bar plot resulting from the SNPrank importance scores. The SNPs are ranked by their chromosome positions, with the height of each bar



**Figure 1** SNPrank importance scores. SNPs are ordered by basepair position, with chromosome boundaries denoted by dotted lines. The height of each bar corresponds to its SNPrank score. The top SNPrank variant is rs1805352 in the *retinoid X receptor  $\alpha$*  (*RXRA*) gene.





**Figure 2** Genetic association interaction network (GAIN) for the top 40 SNPs selected by SNPrank as being the most relevant to smallpox vaccine-associated immune response. The top 10 SNPs (labeled in Figure 1) have bold labels. The top three SNPrank nodes and their connections are highlighted: *RXRA* (red), *NCF2* (green) and *CYP1A1* (blue).

corresponding to its SNPrank score. The top SNPrank variant (rs1805352) is an intronic SNP in the *retinoid X receptor  $\alpha$*  (*RXRA*) gene, which is known to be a mediator of vitamin D signaling and has recently been shown to be involved in innate immune response.<sup>19,20</sup> Figure 2 shows the smallpox vaccine-specific immune response GAIN. For clarity, we only show the top 40 nodes as ranked by SNPrank. The top three SNPrank nodes and their connections are highlighted: *RXRA* (red); *neutrophil cytosolic factor 2* (*NCF2*, green); and *cytochrome P450, family 1, subfamily a, polypeptide 1* (*CYP1A1*, blue). Being the most interactive SNP in Figure 2, the influence of *RXRA* SNP on the smallpox vaccine antibody response is primarily because of its being a hub in the GAIN network. In addition to *RXRA*, GAIN indicates the importance of other vitamin regulation pathway genes for immune response following smallpox vaccination.

## Discussion

The important role of *RXRA* in our SNPrank network analysis of smallpox antibody response is noteworthy, given the findings in recent studies of the influence of this vitamin A- and D-signaling mediator on human immune responses. Results from a recent study of Rubella vaccination suggested that an intronic SNP in *RXRA* influences the magnitude and type of cytokine response following vaccination.<sup>20</sup> Another recent study of CpG-activated human B cells showed that nanomolar

concentrations of *RXRA* and peroxisome proliferator-activated receptor (*PPAR*) $\gamma$  ligands increase antibody production.<sup>19</sup> Taken together, these studies suggest that variation in *RXRA* function may explain in part the variability of human adaptive immune responses following vaccination. Furthermore, the findings suggest that *RXRA*, *PPAR* $\gamma$  and pathways related to these molecules could be exploited for development of new adjuvants that enhance antibody responses.

The smallpox vaccine-specific immune response network (Figure 2) reveals a consistent relationship between vitamin regulation and immune response genes. In the GAIN in Figure 2, *RXRA* has a direct connection with *toll-like receptor 2* (*TLR2*). This further supports the evidence in Ovsyannikova *et al.*<sup>20</sup> for Rubella vaccine that polymorphisms in toll-like and vitamin A and D receptors influence adaptive immune response to vaccines more generally. In Figure 2, *RXRA* has an intermediate connection to *RXR* $\beta$ , which is also a mediator of vitamin D signaling. The second-ranked variant on the SNPrank importance list is *CYP1A1*, which is in the pathway of chemical reactions resulting in the formation of 9-*cis*-retinoic acid, a metabolically active vitamin A derivative. The direct GAIN connection of a variant in *5,10-methylenetetrahydrofolate reductase* (*MTHFR*) to *RXRA* (ranked first by SNPrank) and to *NCF2* (ranked third by SNPrank) suggests a role for folic acid metabolism in antibody responses. Understanding the interactions in GAIN may inform the development of new vaccines and immunotherapies, and these interactions

may explain the variability of antibody response to other vaccines and infections.

Using gene set enrichment, we obtain additional weight for *RXRA* as a candidate biomarker for vaccine immune response. Specifically, we use GOrilla, which is a web-based tool that identifies enriched GO terms in a ranked list of genes.<sup>21</sup> We use the EC rank list to compare how the genes would be prioritized by GO enrichment versus the data-driven SNPrank approach. The most significantly enriched GO category has *RXRA* along with *CTNNB1*, *GPX1* and *KRAS* annotated with the term, muscle cell differentiation (GO:0042692,  $P = 6.5e-4$ ). The role of muscle cell differentiation in smallpox vaccine immune response is unclear because of the intradermal route of smallpox vaccine administration. Nevertheless, multiple lines of evidence suggest a role for *RXRA* in vaccine antibody response.

In a conventional analysis alone, the *RXRA* variant found by SNPrank would have been neglected because of its marginally significant  $P$ -value of 0.03. In our primary analysis, we filtered the data according to EC, which maintains main effects and interactions.<sup>4</sup> However, we also performed a SNPrank analysis with a logistic regression filter to determine whether SNPrank could recover important interaction effects. Indeed, with the main effect filter, the rank of the *RXRA* variant was boosted by SNPrank to the top position because of multiple network connections based on gene–gene interactions encoded in the GAIN. The initial ranking of the SNPs by logistic regression and EC differed considerably; however, the top list was nearly the same after SNPrank was applied to both filters. This suggests that SNPrank was able to boost the main effect and interaction effects regardless of filter for this data. Certainly, if an association represents a pure interaction with no marginal effect, then a filter based on main effects will lead to a loss in power. For example, *NCF2* was ranked third in our final analysis, but it did not meet the logistic regression  $P$ -value threshold. Similarly, *RXRB* did not pass the main effect filter, whereas it was found to interact with *RXRA* through an intermediate gene in our primary analysis. Thus, SNPrank with a main effect filter is able to generate novel biological knowledge from genetic association studies through network interactions, suggesting it is a reasonable alternative to more computationally intense filters coupled with SNPrank.

The SNPrank algorithm uses a Markov recursion matrix that couples the individual importance of SNPs (main effects) and their interactions with other SNPs on the basis of the data-driven GAIN connectivity matrix. The GAIN matrix is inferred using information theory, but currently does not include covariate corrections. The SNPrank algorithm itself does not depend on the technique used to construct the network. Thus, SNPrank may use a covariate-corrected GAIN or a network based on other statistical weights, such as linear or logistic regression. Fitted coefficients for interaction and main effect terms in the regression model may be used to construct the Markov chain, and important covariates can be included in the model.

The main effects and interaction effects of GAIN are coupled in SNPrank by the probability  $\gamma$  in equation 5. When  $\gamma = 1$ , the SNPrank score depends only on connectivity. In this limit, the RSS agent samples SNPs

by following connections based on gene–gene interactions ( $I_3$ ) that influence the phenotype, but independent effects of SNPs have little influence on the RSS. When  $\gamma = 0$ , the SNPrank scores only depend on the main effect of SNPs, which is based on the information gain ( $I_2$ ) between each SNP and the phenotype. In this limit, the SNPrank scores become correlated with a main effect score such as those obtained from logistic regression. Clearly, these two extremes place too much emphasis on main effects and gene–gene interactions, respectively, in the GAIN connection structure. The developers of PageRank prescribed the value of  $\gamma = 0.85$ , which makes a simulated random web surfer more likely to follow the link structure of the web instead of navigating to a random web page. For SNPrank, we find that the rank order of the top SNPs is robust to values of  $\gamma$  that are not too extreme (close to 0 or 1). As our aim was to discover new biological knowledge, we selected  $\gamma = 0.85$  to enrich for weak higher-order network interactions without completely neglecting main effects. SNPrank gives very similar rankings for  $\gamma$  in the range (0.5, 0.85). The optimum value of  $\gamma$  is likely data dependent and may be estimated through context-dependent GO networks, previous data or other expert knowledge.

## Conflict of interest

The authors declare no conflict of interest.

## Acknowledgements

This work was supported by NIH Grants no. K25 AI-64625 (PI: BA McKinney), R56 AI-80932 (PI: BA McKinney) and T32 HL072757 (PI: DB Allison). We thank JC Diaz for helpful conversations about eigenvector centrality.

## References

- 1 Poland GA, Ovsyannikova IG, Jacobson RM. Vaccine immunogenetics: bedside to bench to population. *Vaccine* 2008; **26**: 6183–6188.
- 2 Poland GA, Ovsyannikova IG, Jacobson RM, Smith DI. Heterogeneity in vaccine immune response: the role of immunogenetics and the emerging field of vaccinomics. *Clin Pharmacol Ther* 2007; **82**: 653–664.
- 3 Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science* 2008; **322**: 881–888.
- 4 McKinney BA, Crowe JE, Guo J, Tian D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genet* 2009; **5**: e1000432.
- 5 Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. *Tech. rep.* 1999-66, *Stanford Digital Library Technologies Project* 1998.
- 6 Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci USA* 2006; **103**: 17973–17978.
- 7 Oti M, Huynen MA, Brunner HG. The biological coherence of human phenome databases. *Am J Hum Genet* 2009; **85**: 801–808.
- 8 Morrison JL, Breitling R, Higham DJ, Gilbert DR. GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* 2005; **6**: 233.

- 9 Strang G. *Linear Algebra and Its Applications*. Saunders College Publishing; New York, 1988.
- 10 McKinney BA, Reif DM, White BC, Crowe Jr JE, Moore JH. Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics* 2007; **23**: 2113–2120.
- 11 Greene CS, Penrod NM, Kiralis J, Moore JH. Spatially Uniform Relief (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min* 2009; **2**: 5.
- 12 Kononenko I. Estimating attributes: analysis and extensions of RELIEF. in *European Conference on Machine Learning* 1994: 171–182.
- 13 McNicholl JM, Downer MV, Udhayakumar V, Alper CA, Swerdlow DL. Host-pathogen interactions in emerging and re-emerging infectious diseases: a genomic perspective of tuberculosis, malaria, human immunodeficiency virus infection, hepatitis B, and cholera. *Annu Rev Public Health* 2000; **21**: 15–46.
- 14 Packer BR *et al*. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res* 2006; **34**: D617–D621.
- 15 Reif DM *et al*. Genetic basis for adverse events after smallpox vaccination. *J Infect Dis* 2008; **198**: 16–22.
- 16 Reif DM *et al*. Integrated analysis of genetic and proteomic data identifies biomarkers associated with adverse events following smallpox vaccination. *Genes Immun* 2009; **10**: 112–119.
- 17 Talbot TR *et al*. Vaccination success rate and reaction profile with diluted and undiluted smallpox vaccine: a randomized controlled trial. *JAMA* 2004; **292**: 1205–1212.
- 18 Purcell S *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
- 19 Garcia-Bates TM *et al*. Peroxisome proliferator-activated receptor gamma ligands enhance human B cell antibody production and differentiation. *J Immunol* 2009; **183**: 6903–6912.
- 20 Ovsyannikova IG *et al*. Rubella vaccine-induced cellular immunity: evidence of associations with polymorphisms in the Toll-like, vitamin A and D receptors, and innate immune response genes. *Hum Genet* 2010; **127**: 207–221.
- 21 Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 2009; **10**: 48.



This work is licensed under the Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>