Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 4 2012 *Article* 8

Correction for Founder Effects in Host-Viral Association Studies via Principal Components

Karyn L. Reeves, *Murdoch University*
Elizabeth J. McKinnon, *Murdoch University*
Ian R. James, *Murdoch University*

Recommended Citation:
Reeves, Karyn L.; McKinnon, Elizabeth J.; and James, Ian R. (2012) "Correction for Founder Effects in Host-Viral Association Studies via Principal Components," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 4, Article 8.
10.1515/1544-6115.8

Correction for Founder Effects in Host-Viral Association Studies via Principal Components

Karyn L. Reeves, Elizabeth J. McKinnon, and Ian R. James

Abstract

Viruses such as HIV and Hepatitis C (HCV) replicate rapidly and with high transcription error rates, which may facilitate their escape from immune detection through the encoding of mutations at key positions within human leukocyte antigen (HLA)-specific peptides, thus impeding T-cell recognition. Large-scale population-based host-viral association studies are conducted as hypothesis-generating analyses which aim to determine the positions within the viral sequence at which host HLA immune pressure may have led to these viral escape mutations. When transmission of the virus to the host is HLA-associated, however, standard tests of association can be confounded by the viral relatedness of contemporarily circulating viral sequences, as viral sequences descended from a common ancestor may share inherited patterns of polymorphisms, termed 'founder effects'. Recognizing the correspondence between this problem and the confounding of case-control genome-wide association studies by population stratification, we adapt methods taken from that field to the analysis of host-viral associations. In particular, we consider methods based on principal components analysis within a logistic regression framework motivated by alternative formulations in the Frisch-Waugh-Lovell Theorem. We demonstrate via simulation their utility in detecting true host-viral associations whilst minimizing confounding by associations generated by founder effects. The proposed methods incorporate relatively robust, standard statistical procedures which can be easily implemented using widely available software, and provide alternatives to the more complex computer intensive methods often implemented in this area.

KEYWORDS: host-viral association study, founder effect correction, principal components, eigenanalysis, logistic regression, Firth correction, HIV

Author Notes: This work is supported by grant 1011319 from the National Health and Medical Research Council of Australia. KR is supported by an Australian Postgraduate Award. Our thanks to Shay Leary for assistance with data preparation, to Dr Mina John, Prof. Simon Mallal and other colleagues at the Institute for Immunology & Infectious Diseases, and to all participants and study team members of the WA HIV Cohort Study. We thank the reviewers for helpful comments.

INTRODUCTION

Rapid advances in genomics technology have facilitated the routine sequencing of full viral genomes together with the ascertainment of detailed host genotypes, particularly the human leukocyte antigen (HLA) alleles which encode proteins central to the host's immune repertoire in terms of viral recognition and suppression (eg. Goulder and Watkins, 2004, 2008). Briefly, HLA molecules present segments of viral sequence typically 8-11 amino acids long, termed peptide epitopes, at the surface of the infected cell for recognition by T-cells. Viruses such as HIV or Hepatitis C (HCV) which replicate extremely rapidly and with high transcription error rates may possibly escape detection by the immune system if they mutate at key positions within these HLA-specific peptides and hence abrogate either binding of the peptide or subsequent T-cell recognition. Large-scale population-based host-viral association studies are conducted as hypothesis-generating exploratory analyses, and aim to determine those positions within the viral sequence at which host HLA immune pressure may have led to viral escape mutations (eg. Moore et al., 2002). In addition to the intrinsic biological interest of these viral escape mechanisms, identification and assessment of such host-driven mutation patterns may have important implications for drug resistance studies and vaccine design (Goulder and Watkins, 2004, 2008).

Population structure, which may arise as a consequence of viral relatedness within contemporarily circulating viral sequences, has the potential to confound host-viral association studies. As the virus is transmitted from host to host, any sample taken from a population may include multiple recipients infected by a single donor, or both the donor and recipient of a transmission event, and these related sequences may share random patterns of amino acid polymorphisms, termed founder effects, as a consequence of their shared ancestry (Bhattacharya et al., 2007). Where transmission is HLA-associated, these polymorphisms can confound conventional tests of association, such as Fisher's exact test. As both virus subtypes (clades) and HLA alleles have distributions influenced by geography and ethnicity, viral transmission cannot generally be assumed to be independent of HLA type, and so the confounding potential of founder effects must be addressed. A number of model-based methods utilising inferred phylogenetic trees estimated from the observed viral sequences have been proposed and used to correct for this confounding (eg Bhattacharya et al., 2007, Carlson et al., 2007, 2008, Rousseau et al., 2008, Brumme et al., 2009, John et al., 2010). However, the methods are typically very computationally intensive and generally require specialist software for implementation. In this paper we propose an alternative empirical approach based on principal components and study its efficacy by simulation.

The problem posed by viral relatedness in the presence of HLA-associated transmission is similar to that of population stratification in case-control association studies, in that the two measured variables are both related in some way to a third unmeasured (and unmeasurable) variable reflecting demographic and geographic factors. Here we consider the utility of adapting methods taken from this field (see for example Price et al., 2010) to the analysis of host-viral associations. In particular, we investigate the use of structured association methods based on principal components analysis which can be considered as analogues of the popular Eigenstrat method (Price et al., 2006) typically utilized in the analysis of case-control association studies. The approaches rely on relatively robust, standard statistical procedures and can be readily implemented using widely available software. A simulation study demonstrates the utility of the methods in detecting true host-viral associations whilst minimizing confounding by associations generated by founder effects. For these simulations we construct a pool of underlying host HLA profiles and a pool of viral amino acid sequences based on data for the HIV gag protein from the Western Australian HIV Cohort (Mallal, 1998). The random reallocation of HLA alleles and viral sequences provides a data set with biologically real HLA profiles and sequences but with no HLA-sequence associations. Known HLA-driven mutations and founder effects are then superimposed for the simulations as described in detail below.

METHODS

PCA adjusted linear regression (Eigenstrat)

Structured association testing is used in case-control association studies to correct for confounding due to underlying population structure. The Eigenstrat method (Price et al., 2006) uses principal component analysis (PCA) to capture genetic structure resulting from population stratification. As genetic variation within a continent tends to vary continuously across a geographic space, Eigenstrat seeks to identify clines which describe the variation. These clines are termed ‘axes of variation’ and it is suggested that they can be represented by linear combinations of unobserved eigenvectors of the population covariance, estimated by the eigenvectors corresponding to larger eigenvalues of the sample covariance matrix. Eigenstrat tests for associations between traits and genotypes after removing correlations assumed to be due to ancestry, and can be reformulated as a regression approach which adjusts for these eigenvectors.

We place this method into our context of host-viral association studies. In these studies the interest centres on whether carriage of individual HLA alleles associates with mutation of the virus at any amino acid position across the whole viral sequence. Here the confounding does not stem from the hidden ancestries of

the individuals included in the sample, but rather from unobservable relationships between the viral sequences circulating within the population from which the sample was taken. Hence we note that, in contrast to applications that typically apply Eigenstrat, the population structure we aim to capture is that obtained from the viral sequences rather than the individuals' genotypes. Suppose we have n individuals each with a corresponding viral sequence consisting of m amino acids and let $\boldsymbol{S}$ represent an $n \times m$ sequence matrix, with s_{ij} an indicator of the presence of a non-consensus variant amino-acid (mutation) at the j^{th} position (residue) of the i^{th} individual sequence. Following Price et al. (2006) (and noting that our matrix is transposed to conform to typical statistical data conventions) we mean correct the columns of $\boldsymbol{S}$ to obtain the matrix $\boldsymbol{S}^*$. If there are $k + 1$ distinct sub-clades within the sequences, or an admixture of $k + 1$ sub-clades, the eigenvectors associated with the k largest eigenvalues of the matrix $\boldsymbol{S}^*\boldsymbol{S}^{*\prime}$ are assumed to form a basis for the structure sub-space within the column space of $\boldsymbol{S}$. Alternatively, the eigenvectors can be determined through a singular value decomposition of $\boldsymbol{S}^*$. The selected eigenvectors are used to form the columns of an $n \times k$ basis matrix $\boldsymbol{E}$, so that $\boldsymbol{M_E} = \boldsymbol{I} - \boldsymbol{E}(\boldsymbol{E}'\boldsymbol{E})^{-1}\boldsymbol{E}'$ is a projection matrix assumed to project onto the sub-space orthogonal to the structure sub-space. As the eigenvectors are orthonormal, $(\boldsymbol{E}'\boldsymbol{E})^{-1} = \boldsymbol{I_k}$ and the projection matrix reduces to $\boldsymbol{I} - \boldsymbol{E}\boldsymbol{E}'$. If $\boldsymbol{H}$ denotes the $n \times r$ HLA matrix with 0/1 entries h_{ij} indicating carriage of the j^{th} HLA type by the i^{th} individual, the matrices $\boldsymbol{M_E}\boldsymbol{S}$ and $\boldsymbol{M_E}\boldsymbol{H}$ are the corrected sequence and HLA matrices from which correlations assumed to be due to viral relatedness have been removed. With $\boldsymbol{y}$ denoting a column of $\boldsymbol{H}$ and $\boldsymbol{x}$ a column of $\boldsymbol{S}$, implementation of the PCA-corrected host-viral association tests can then be achieved by regressing $\boldsymbol{M_E}\boldsymbol{y}$ on $\boldsymbol{M_E}\boldsymbol{x}$ by least-squares for each column of $\boldsymbol{H}$ and each column of $\boldsymbol{S}$. As has been noted by Price et al. (2006) and others, an equivalent analysis is obtained if $\boldsymbol{y}$ is regressed on $\boldsymbol{x}$ with simultaneous adjustment for the columns of $\boldsymbol{E}$. We refer to this approach as **PCA-R** (for regression), noting that the response $\boldsymbol{y}$ is binary and thus will not typically satisfy the normality assumptions inherent in linear least-squares inference.

The Frisch-Waugh-Lovell Theorem

The above regressions represent two formulations in the Frisch-Waugh-Lovell (FWL) theorem, highlighted particularly in the econometrics literature (Frisch and Waugh, 1933, Lovell, 1963, Davidson and MacKinnon, 1993). Consider the general linear regression of a response vector $\boldsymbol{y}$ on a set of variables divided into two groups described by a single vector $\boldsymbol{x}$ and a matrix $\boldsymbol{Z}$ with k columns, and where the focus is on assessing the significance of association of $\boldsymbol{y}$ with $\boldsymbol{x}$ by least-squares rather than in determining the regression model itself:

$$\boldsymbol{y} = \beta_0 + \boldsymbol{x}\beta_1 + \boldsymbol{Z}\beta_2 + \boldsymbol{\varepsilon_1} \quad (1)$$

According to FWL, an identical estimate of β_1 is obtained from the model obtained by annihilating $\boldsymbol{Z}$ through pre-multiplication by the projection matrix $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$, namely

$$\boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{1}\beta_0 + \boldsymbol{M}\boldsymbol{x}\beta_1 + \boldsymbol{\varepsilon_1} \quad (2)$$

Here $\boldsymbol{M}$ projects both the dependent and independent variables onto the sub-space orthogonal to the column-space of $\boldsymbol{Z}$. These two formulations correspond to those described in the previous section. However the FWL theorem also notes that there are other additional equivalent formulations, in particular

$$\boldsymbol{M}\boldsymbol{y} = \beta_0 + \boldsymbol{M}\boldsymbol{x}\beta_1 + \boldsymbol{Z}\beta_2 + \boldsymbol{\varepsilon_1} \quad (3)$$

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{1}\beta_0 + \boldsymbol{M}\boldsymbol{x}\beta_1 + \boldsymbol{\varepsilon_2} \quad (4)$$

with the residuals in (1) - (3) numerically identical. As noted above, in the context of our host-viral association application the response $\boldsymbol{y}$ is a binary indicator of HLA carriage and hence the use of least-squares approaches and inferences will be approximate and may lead to biased results, particularly when the response proportions are small. In such situations it is more appropriate to replace linear regression with, for example, logistic regression, as noted by Price et al. (2006), Zeggini et al. (2008), Need et al. (2009) and Wu et al. (2011), with inference based on the analogue of equation (1). While the transformed responses $\boldsymbol{My}$ are no longer binary, as they correspond to the residuals from a regression analysis, inferences arising from a least-squares approach may still not be valid for data of the type we consider here. Although the FWL theorem applies only to linear least-squares regressions, it does suggest alternatives which may be implemented in the logistic regression setting and we consider two of these in the following section.

PCA adjusted logistic regression

PCA-L: Standard logistic regression based on adjustment by the eigenvectors associated with the larger eigenvalues analogous to FWL regression (1) has been considered by Wu et al. (2011) and references therein, and referred to as PCA-L. In this model the components of β_2 are treated as nuisance parameters in the context of the association test and the logistic linear predictor for expected proportion $\boldsymbol{\pi}$ represented by

$$\log(\boldsymbol{\pi}/(\boldsymbol{1} - \boldsymbol{\pi})) = \beta_0 + \boldsymbol{x}\beta_1 + \boldsymbol{E}\beta_2$$

PCA-P: Our second approach is based on FWL regression (4), with the independent variable replaced by its projection and the response variable remaining in its un-projected binary form, namely

$$\log(\boldsymbol{\pi}/(\mathbf{1}-\boldsymbol{\pi})) = \boldsymbol{M_E}\,\mathbf{1}\,\beta_0 + \boldsymbol{M_E}\,\boldsymbol{x}\,\beta_1 = \beta_0 + \boldsymbol{M_E}\,\boldsymbol{x}\,\beta_1$$

where the last equality follows from $\boldsymbol{M_E}\,\mathbf{1} = \mathbf{1}$ due to the column mean centering in $\boldsymbol{S}^*$. We refer to this model as PCA-P (for projection). For both PCA-L and PCA-P we assess the significance of β_1 by likelihood ratio (drop-in-deviance) tests.

PCA-FP: A number of correction methods are available to account for potential biases arising in logistic regression when data are sparse, and in particular when separation occurs as a result of estimates of proportions tending to 0 with corresponding logistic parameters tending to infinity (eg. Maiti and Pradhan, 2008). We have therefore also included the Firth (1993) corrected version of PCA-P, denoted PCA-FP, in our simulations. Based on maximization of a penalized likelihood, the Firth correction seeks to correct bias by modifying the score function to

$$g(\beta_r) = \Sigma_i\,[\,y_i - \pi_i + h_i(0.5 - \pi_i)\,]\,x^*_{ir}$$

where the h_i's are the diagonal elements of $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}^*(\boldsymbol{X}^{*\prime}\boldsymbol{W}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{W}^{1/2}$ with $\boldsymbol{W} = \mathrm{diag}\{\pi_i(1-\pi_i)\}$ and here $\boldsymbol{x}^* = \boldsymbol{M_E}\,\boldsymbol{x}$. It is readily programmable with relatively fast convergence for the single explanatory variable; alternatively, implementation is available in some general statistics packages such as SAS or R (brglm package).

SIMULATIONS

A simulation study using re-sampled real data was undertaken to compare the efficacies of PCA-R, PCA-L, PCA-P and PCA-FP, together with uncorrected logistic regression analysis and Fisher exact tests, the last two of which should approximate each other provided cell counts are reasonably large. HLA-alleles and corresponding viral amino acid sequences for the HIV gag protein were extracted from 213 cases in the Western Australian HIV Cohort Study (Mallal, 1998). The observed HLA-B alleles (two per individual) were randomly reassigned to viral sequences (m = 501 amino acids) to remove associations. For the purposes of the simulations, missing amino acids were replaced by the consensus at that position. Two hundred cohorts were simulated according to sampling models which then superimposed viral relatedness in the presence of HLA-associated transmission, while enabling the structural relationships to be identifiable as described below. Up to 10 known HLA-viral mutation associations

were artificially embedded into the datasets prior to analysis at varying proportions approximating in each case the odds ratio estimated from real data and varying between 6.6 and 47. All methods were implemented on each data set and analyses were carried out using TIBCO Spotfire S+ 8.1 (TIBCO Software Inc., Palo Alto, California).

Each cohort was created by firstly sampling with replacement $n = 200$ sets of two HLA-B alleles from the combined pool of HLA alleles. In order to simulate the HLA-related transmission of virus from common donors we then identified HLA alleles carried by at least 15 of these simulated individuals and chose 4 of the alleles (possibly identical) at random. Corresponding to these alleles, four viral sequences were then sampled at random with replacement, with each allocated to a group of cohort members of random size (varying between 8 and 20), at least 50% of whom carried the same randomly chosen HLA allele. Additional sequences were sampled at random with replacement and allocated to the remaining cohort members. Up to 7% of the amino acids across the entire cohort were then allowed to mutate but constrained to maintain the proportion of non-consensus amino acids, so that the sequences preferentially infecting groups of individuals were similar but not identical. The preferentially-infected alleles can be expected to show associations with any polymorphisms characteristic of the infecting related sequences in an association test which fails to correct for such confounding.

Ten HLA-B viral associations based on those found to be significant in previous studies were then embedded as follows: where a simulated individual carried the specified HLA allele, the amino acid at the associated residue was mutated with a probability reflecting the odds ratio of the underlying sampled data. Due to the random nature of HLA and sequence allocation, some cohorts contained fewer than 10 embedded 'true' associations by chance.

Assessing the methods

The six described methods (PCA-R, PCA-L, PCA-P, PCA-FP, uncorrected logistic regression and Fisher tests) were compared by selecting two arbitrary p-value thresholds of 0.001 and 0.005, and declaring in each case any association with a smaller p-value to be significant. Following typical practice we considered only those HLA/amino acid combinations with at least 5 HLA carriers and at least 5 non-consensus amino acids at the residue. As a result of the cohort construction, all associations found will be false apart from the artificially embedded 'true' associations, and the identifiable 'founder effect' associations which are artifacts of the modeled population structure. Accordingly, the positive associations in each simulation were classified as true, false, or where associated with a preferentially infected HLA allele, as being due to founder effects. We note that

the count of founder effect associations may be an overestimate, as an unobservable number of these associations may actually be randomly false. The three quantities of interest in assessing the effectiveness of the proposed methods are the number (or proportion) of true associations correctly identified as significant, the number of founder effect associations incorrectly identified as significant, and therefore not controlled, and the number of false associations also incorrectly identified as significant.

All correction methods used the eigenvectors associated with the six largest eigenvalues. In addition, to assess the robustness of the methods to the choice of numbers of eigenvectors for correction, PCA-R, PCA-L, PCA-P and PCA-FP were implemented correcting along increasing (even) numbers of eigenvectors, between 2 and 10.

RESULTS

The counts of the true, false and founder effect associations for the six methods averaged over the 200 simulations are shown in Figure 1 for p-value thresholds of 0.001 and 0.005, respectively. The actual average counts are given in Table 1. These results show clearly that viral relatedness in the presence of HLA associated transmission leads to a large number of significant founder effect associations in the uncorrected Fisher and logistic regression analyses, while these founder effects are largely abrogated by use of the PCA corrections. The methods PCA-R, PCA-L, PCA-P and PCA-FP retain approximately the same numbers of significant "true" associations, and are slightly superior in this respect to the two uncorrected methods (Table 1). The uncorrected Fisher and logistic regression tests show a large number of artifactual associations due to the founder effects, while all the corrected methods reduce these to very low levels, illustrating the efficacy of the PCA corrections in largely eliminating the founder effects. We note that the uncorrected Fisher and logistic regression methods demonstrate many fewer than the typical 100p% "false" associations expected when null p-values follow a uniform distribution. This results from the small marginal totals in many of our tables and the corresponding discreteness of the attainable p-values given the fixed margins which leads to a large skewness of Fisher p-values away from 0 and towards 1.

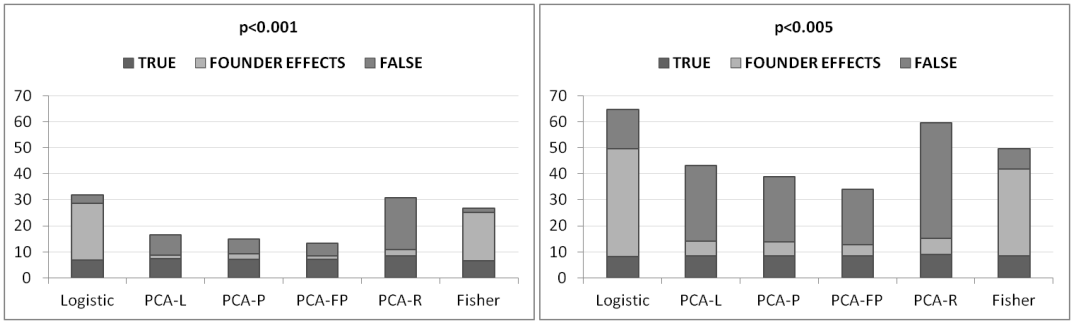

Figure 1: The average counts per cohort of true, founder effect and false associations from an analysis of 200 simulated cohorts using significance thresholds of $p < 0.001$ and $p < 0.005$. Cohorts contained an average of 9.57 embedded true associations, and founder effect associations resulting from modeled population structure. Results obtained from application of standard logistic regression and Fisher tests are compared with those obtained from PCA-corrected linear regression (PCA-R) and logistic regression (PCA-L, PCA-P and PCA-FP). All PCA-corrections are based on the projection matrix derived from the eigenvectors associated with the 6 largest eigenvalues.

		Analysis method					
Threshold	**Association type**	**Logistic**	**PCA-L**	**PCA-P**	**PCA-FP**	**PCA-R**	**Fisher**
$p < 0.001$	**True**	6.72	7.32	7.03	7.02	8.44	6.57
	Founder effect	21.92	1.35	2.14	1.36	2.18	18.55
	False	3.12	7.65	5.69	4.72	19.96	1.48
$p < 0.005$	**True**	8.16	8.32	8.27	8.25	8.96	8.20
	Founder effect	41.28	5.66	5.44	4.26	6.11	33.42
	False	15.24	29.22	25.07	21.45	44.54	7.90

Table 1: The average counts per cohort of true, founder effect and false associations from an analysis of 200 simulated cohorts using significance thresholds of $p < 0.005$ and $p < 0.001$. The PCA methods correct along 6 eigenvectors.

Whilst performing markedly better than the uncorrected methods overall, the PCA-corrected logistic methods do induce slightly more "false" associations. For PCA-R, the linear regression approximation analogous to Eigenstrat, the number of "false" associations was high, suggesting that this approach may be unsuitable for analyses of data such as encountered here without some p-value

correction. This contrasts with the findings of Wu et al. (2011) who found little difference in linear and logistic regression implementations with case-control data. We note that here we are dealing with response proportions which are much closer to zero than those typically found in the case-control situation, and it is in the tails where the linear and logistic regressions tend to deviate. Simulations on cohorts created without embedded population structure result in comparable average per cohort counts of true and false positive associations (data not shown).

There is not a lot to choose between the three logistic regression based correction methods in terms of true associations found and false/artifactual error rates. There is however some improvement in the use of the Firth-corrected PCA-FP method which incorporates bias and separation correction; our results suggest that of the methods considered this has the best overall results. We note that some false associations may still result from failure of the separation correction.

Robustness to the number of eigenvectors

The eigenvector-based correction methods require a decision as to the value of *k*, the number of eigenvectors required to span the structure sub-space. This is equivalent to determining the number of meaningful groups or clusters in the dataset, which remains an unresolved problem in statistics despite many proposed solutions. To assess the robustness of the PCA correction methods to the choice of *k*, 50 of the simulated cohorts were also analyzed using PCA-R, PCA-L, PCA-P and PCA-FP correcting along increasing (even) numbers of eigenvectors between 2 and 10. The true positive rates (TPR) of the true associations and false positive rates (FPR) of the combined false and founder effect associations averaged over all simulations are plotted in Figure 2. These analyses suggest relatively constant true positive rates across dimensions, with false positive rates declining until approximately 6 eigenvectors and relatively constant for PCA-P, PCA-FP, and PCA-R thereafter, while the false positive rate for PCA-L increases with additional eigenvectors. This suggests that for data of this type a minimum of approximately 6 eigenvectors should be included for correction to adequately control the founder effect associations, but beyond this the TPR and FPR rates remain reasonably stable and robust over a range of larger values of *k*. For datasets generated under the sampling models used here, values of *k* at least 6 seem to provide reasonable results.

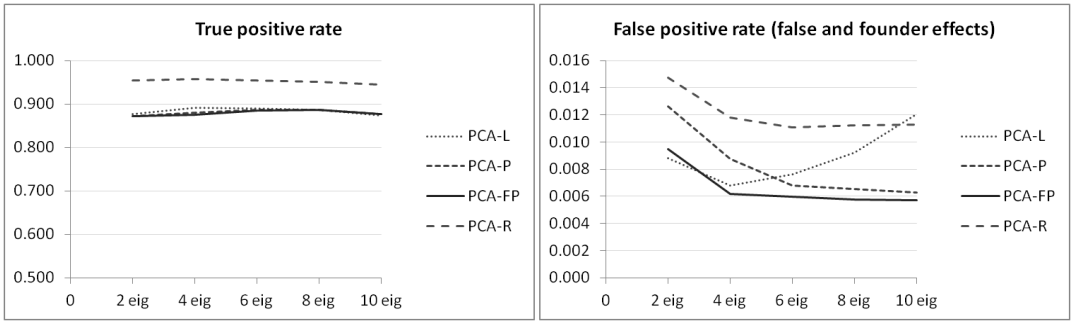

Figure 2: Plots of true and false positive rates correcting along increasing numbers of eigenvectors for PCA-R, PCA-L, PCA-P and Firth-corrected PCA-FP.

A host-viral association study is an exploratory study, focused on finding as many true associations as possible for further experimental verification while at the same time eliminating as many false associations as possible to avoid unnecessary experimentation. In this context our results indicate that the Firth-corrected PCA-FP method is simple to implement and appears to provide the best tradeoff between controlling founder effects and identifying true associations. Although we have carried out simulations for fixed p-values, we note that it would be typical in large scale host-viral association studies to convert the p-values to q-values via false discovery rates analogous to the methods proposed by Storey and Tibshirani (2003), for example, in order to select a suitable limit.

APPLICATION

It is not our purpose here to carry out a comprehensive analysis, but for illustration we applied the methods to HIV gag-protein data from the WA HIV Cohort. Our data consists of the amino acid sequences obtained from 210 cases with predominantly clade-B virus, together with indicators of allele carriage at the class I HLA-A, -B and -C loci. Each individual carries two each of the HLA-A, -B and -C alleles, with a total of fifty-nine different 4-digit alleles represented in frequencies ranging from 0.024 to 0.414. Carriage of each allele was considered separately in analyses. We note that when using uncorrected association analyses one might typically omit sequences displaying as outliers relative to the main clade via phylogenetic or cluster methods, however here we retain all sequences to illustrate the efficacy of the founder-effect correction. Missing values in the viral sequence data have been accommodated by including for analysis at the particular residue only those sequences with a corresponding non-missing amino acid, but imputing the consensus amino acid where missing at other residues in order to estimate the corrections. A separate projection matrix was therefore calculated for each residue, with corrections implemented using the eigenvectors

associated with the six largest eigenvalues calculated from the reduced matrix. Residue/HLA combinations were included only when at least 5 individuals carried the HLA allele and at least 5 had the non-consensus amino acid.

These analyses found 31 HLA/residue associations with a PCA-FP p-value less than 0.001 (Table 2). Nineteen of the associations were located within epitopes previously reported for the HLA and listed in the HIV molecular immunology database of the Los Alamos National Laboratory (LANL) [http://www.hiv.lanl.gov/content/immunology/tables/tables.html], two are with HLA alleles which are in strong linkage disequilibrium with an allele included in the list of reported associations together with the same residue, and a further two have been identified as significant by other correction methods including those using clustering or phylogenetic trees (Bhattacharya et al., 2007, Carlson et al., 2007). Additionally, we observed 4 associations with PCA-FP p-values <0.001 but large Fisher p-values. As unusually large parameter estimates were also observed for these instances, we judged them to be false associations resulting from the occasional inability of the method to provide reliable estimates in the presence of separation even after correction.

Associations	**Total (N)**	**Falling within reported epitope[#] (N)**	**Identifiable as founder effect* (N)**
Concordant			
PCA-FP < 0.001; Fisher p < 0.001	15	12[‡,†]	0
Discordant			
PCA-FP < 0.001; Fisher p ≥0.001	16	7[†]	0
PCA-FP ≥ 0.001; Fisher p <0.001	83	1[††]	74

[#]*As listed in the Los Alamos National Laboratory (LANL) HIV immunology database.*
**Based on clustering of low-frequency HLA alleles with a minority viral subtype*
[‡]*An additional 2 associations have been reported in other studies.*
[†]*The HLA allele of an additional association is in linkage disequilibrium with an allele reported in the LANL list of associations together with the same viral residue.*

Table 2: Numbers of HLA/residue associations identified as significant (p < 0.001) in real data implementing a Firth-corrected PCA-FP along 6 eigenvectors and/or an uncorrected Fisher test.

The uncorrected association analysis using Fisher exact tests returned 98 associations with a p-value below 0.001. Of these, 74 were assessed as known founder effect associations resulting from carriage of the alleles A*02:07, B*46:01, A*30:01, B*42:01, B*15:03 and C*17:01 and infection with a differing viral subtype as indicated by clustering of the sequences. Associations between these HLA alleles and the polymorphisms distinguishing the differing clades can be expected to return small p-values in an analysis which fails to correct for founder effects. All recognizable founder effect associations were removed by the PCA-FP correction at this significance level.

DISCUSSION

In this study we have demonstrated that eigenvector-corrected logistic regression procedures can provide effective methods to control for founder effects in host-viral association studies in which one is interested in whether carriage of particular HLA alleles is associated with mutation of viral amino acids across the whole viral genome. These eigenvector-corrected methods are modifications of the Eigenstrat procedure widely used in human genome-wide association studies to control for confounding from population stratification. We differ from the typical GWAS approach in that the projections here are based on the viral sequence data because this captures the primary structure. Our genotype data is obtained from only a small number of multi-allelic loci and the consequent sparseness in these data does not enable definition of relevant strata. Four alternative implementations were considered with simulations suggesting the preferred approach was a logistic regression analysis of the unprojected HLA vector on the projected viral consensus/non-consensus vector, combined with a Firth correction to control for spurious p-values which can be caused by the separation problems inherent in large scale logistic regression analyses with small samples. The method can be readily implemented using widely available statistical software.

The founder effect correction is implemented by projecting the viral vector into the subspace orthogonal to that assumed to contain the structure confounding the tests. In a host-viral study this structure subspace is estimated from the sequences, and assumed to be spanned by the eigenvectors associated with the largest eigenvalues of the SSP matrix of the column mean-centered viral consensus/non-consensus matrix. This approach is underpinned by the observation that HIV variability follows a hierarchical pattern in that the variability between the HIV clades exceeds that within the clades, which in turn exceeds that circulating within an individual host. We therefore expect that the eigenvectors associated with the larger eigenvalues from the SSP matrix will mostly reflect between-clade or between-sub-clade variability, while the

eigenvectors associated with eigenvalues of a lower rank may reflect smaller levels of intra-sub-clade variability, the footprinting effect of the HLA-induced escape mutations we seek to identify, or noise. In common with other founder effect correction methods it is possible that some of the footprinting may be confounded with the structure correction.

The Frisch-Waugh-Lovell theorem provides an insight into the effectiveness of this approach. The eigenvectors form a basis which spans a sub-space, with the test of association conducted after allowing for all possible structures contained within this sub-space. It is not necessary in this approach to estimate the true viral interrelationships; it is sufficient to find a basis spanning the space in which the viral structure lies. It also suggests that a founder effect-correcting procedure could be implemented using any set of linearly independent vectors spanning the relevant space and estimated through procedures other than eigen-analysis. This includes the founder-effect controlling method suggested in Rauch et al. (2009), in which the algorithm Partitioning Around Medoids (Kaufman and Rousseeuw, 1990) was used to infer broad groups of viral relatedness so that the sequences could be separated into strata, with a Mantel-Haenszel statistic then used to test for consistent association in the same direction across all strata. This approach approximates an analogue of PCA-L with the substitution of factor variables constructed from the clustering algorithm in place of the eigenvectors.

The eigenvector-corrected regression procedures discussed here lie within the broader framework of the general/generalized linear model. This suggests that eigenvector-based founder effect correction procedures could be integrated into a number of well known and widely-used linear-based statistical techniques, including multiple regression and canonical correlation, all of which would allow for techniques which could accommodate linkage disequilibrium within the viral residues and the non-random groupings of HLA alleles in haplotypes.

REFERENCES

Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C, Carlson J, Yusim K, McMahon B, Gaschen B and others. 2007. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315**:1583-1586.

Brumme ZL, John M, Carlson JM, Brumme CJ, Chan D, Brockman MA, Swenson LC, Tao I, Szetos S, Rosato P and others. 2009. HLA-associated immune escape pathways in HIV-1 Subtype B Gag, Pol and Nef proteins. *PLoS ONE* **4(8)**: e6687.

Carlson JM, Kadie C, Mallal S, Heckerman D. 2007. Leveraging hierarchical population structure in discrete association studies. *PLoS ONE*, **2(7)**:e591.

Carlson JM, Brumme ZL, Rousseau C, Brumme C, Matthews P, Kadie C, Mullins J, Walker BD, Harrigan PR, Goulder PJR, Heckerman D. 2008. Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 gag. *PLoS Comput Biol* **4:**e1000225.

Davidson R, MacKinnon JG. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press

Firth D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* **80**:27-38.

Frisch R, Waugh FV. 1933. Partial time regressions as compared with individual trends. *Econometrica* **1**:387-401.

Goulder PJR, Watkins DI. 2004. HIV and SIV CTL escape: implications for vaccine design. *Nat Rev Immunol* **4**:630-640.

Goulder PJR, Watkins DI. 2008. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. *Nat Rev Immunol* **8**:619-630.

John M, Heckerman D, James I, Park LP, Carlson JM, Chopra A, Gaudieri S, Nolan D, Haas DW, Riddler SA and others. 2010. Adaptive interactions between HLA and HIV-1: Highly divergent selection imposed by HLA Class I molecules with common supertype motifs. *J Immunol* **184**:4368-4377.

Kaufman L, Rousseeuw PJ. 1990. *Finding Groups in Data*. New York: Wiley.

Lovell MC. 1963. Seasonal adjustment of economic time series and multiple regression analysis. *JASA* **58**: 993-1010.

Maiti T, Pradhan V. 2008. A comparative study of the bias corrected estimates in logistic regression. *Stat Methods Med Res* **17**:621-634.

Mallal SA. 1998. The Western Australian HIV cohort study, Perth, Australia. *J Acquir Immune Defic Syndr and Hum Retrovirol* **17** Suppl 1:S23-7.

Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. 2002. Evidence of HIV-1 adaptation to HLA-restricted immune response at a population level. *Science* **296**:1439-1443

Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, Shianna KV, Yoon W, Kasperaviciute D, Gennarelli M and others. 2009. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet* **5(2):** e1000373.

Price A, Patterson N, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:**904-909.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**:459-463.

Rauch A, James I, Pfafferott P, Nolan D, Klenerman P, Cheng W, Mollison L, McCaughan G, Shackel N, Jeffrey GP and others. 2009. Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen – restricted immune pressure. *Hepatology* **50**:1017-1029.

Rousseau CM, Daniels MG, Carlson JM, Kadie C, Crawford H, Prendergast A, Matthews P, Payne R, Rolland M, Raugi DN and others. 2008. HLA class-1 driven evolution of Human Immunodeficiency Virus type 1 subtype C proteome: immune escape and viral load. *J Virol* **82:**6434-6446.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**:9440-9445.

Wu C, DeWan A, Hoh J, Wang Z. 2011. A comparison of association methods correcting for population stratification in case-control studies. *Ann Hum Genet* **75**:418-27.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G and others. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**:638-645.