FOUNDER EFFECTS AND RELATED ISSUES IN HOST-VIRAL ASSOCIATION STUDIES

Karyn Reeves

This thesis is presented for the degree of Doctor of Philosophy at Murdoch University, 2013

I declare that the work presented in this thesis is my own account of my research and contains as its main content work which has not been previously submitted for a degree at any tertiary educational institution.

.....

Karyn Reeves
17 January 2013

ACKNOWLEDGEMENTS

I would like to thank Professor Ian James for first bringing this research topic to my attention, for his encouragement and patient supervision during the past three years of study, and for his suggestions and guidance during the writing of this thesis. I would also like to thank my co-supervisor Dr. Elizabeth McKinnon for her advice relating to content and structure of the thesis and her unremitting attention to detail.

Thanks also to Dr Mina John, Prof. Simon Mallal and other colleagues at the Institute for Immunology & Infectious Diseases, Shay Leary for assistance with data preparation, and all participants and study team members of the WA HIV Cohort Study.

Finally, I would like to acknowledge the patience and support of my husband and children during my years of study.

ABSTRACT

Viruses such as HIV which replicate rapidly and with high transcription error rates may evade immune detection by mutating at key positions within the viral amino acid sequence. Large-scale host-viral association studies are conducted to identify positions of possible escape mutation in response to host immune pressure, with this pressure predominantly governed by genes within the human leukocyte antigen (HLA) complex. When transmission of the virus is HLA-associated, however, standard tests of association can be confounded by the relatedness of contemporarily circulating viral sequences, as sequences descended from a common ancestor may share inherited patterns of polymorphisms, termed 'founder effects'. A number of model-based methods utilizing inferred phylogenetic trees estimated from the observed viral sequences have been proposed to correct for this confounding, although such methods are typically computationally intensive and require specialist software for their implementation. In this thesis we propose an alternative empirical approach based on principal components analysis (PCA) which can be implemented using widely available software, and which adapts and extends methods currently used to control for population stratification in case-control genome-wide association studies. To accommodate data with small proportions commonly observed in host-viral studies we implement the PCA-based controlling procedure within a logistic regression framework using novel formulations motivated by the Frisch-Waugh-Lovell Theorem and demonstrate their utility in detecting true associations whilst minimizing confounding generated by founder effects via simulation. The approach is then extended to the multivariate setting through the adaptation of well-known techniques which expand the scope of host-viral analyses by accommodating possible linkages within the HLA and viral data.

The thesis concludes with a discussion of issues arising from the application of tail-based rejection regions and false discovery rates in large-scale analyses based on pooled contingency tables with varying margins. We argue that constraints imposed by the margins have implications overlooked in the rigid application of techniques developed for tests based on statistics with continuous distributions, but by leveraging the scale of such analyses it may be possible to consider local deviations between observed and expected p-value distributions to better identify hypotheses of interest.

CONTENTS

INTRODUCTION

Host-viral association studies are exploratory analyses conducted to identify leads for follow up experimental investigation. They aim to determine the positions within a viral sequence at which immune pressure imposed through the actions of host human leukocyte antigens (HLAs) may have induced viral escape mutations. These escape mutations are a characteristic of viruses which replicate rapidly and with high transcriptional error rates, such as human immunodeficiency virus type 1 (HIV-1) and hepatitis C virus (HCV). The transcriptional errors facilitate viral escape from immune detection by enabling the chance encoding of mutations at key positions within HLA-specific peptides. These mutations may offer a survival benefit to the virus by potentially disrupting the binding and display of epitopes which signal a cell as infected, and by impeding killer T-cell recognition of infected cells. 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).

Typically, a host-viral association study proceeds with no *a priori* assumptions about where in the viral sequence such HLA-induced escape mutations may occur. Association between consensus/non-consensus amino acid observations and HLA allele carriage is tested for every combination of viral residue and HLA allele in the sample, subject to certain constraints, using, for example, Fisher's exact tests. When transmission of the virus to the host is HLA-associated, however, standard tests of association such as Fisher's exact test can be confounded by the relatedness of contemporarily circulating viral sequences. As viral transmission is from host to host, any sample taken from a population may include both donors and multiple recipients, 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). The distributions of both viral sub-groups and HLA alleles are influenced by geography and ethnicity, and so viral transmission cannot generally be assumed to be unassociated with host HLA types, and the confounding potential of founder effects should 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, Carlson, Brumme et al., 2008, Rousseau et al., 2008, Brumme et al., 2009, John et al., 2010). However, these methods are typically complex and computationally intensive, and require specialist software for their implementation. This complexity limits their widespread implementation.

Our primary intention here is to develop a simple and robust founder effect-correction procedure based on standard statistical procedures which can be implemented using basic computing facilities and widely available statistical software. In noting that the problems posed by viral-relatedness in the presence of HLA-associated transmission have similarities with those caused by population stratification in case-control genome-wide association studies, we have adapted methods taken from that field to the analysis of host-viral associations. In particular, we have investigated the use of structured association methods based on principal components analysis (PCA) which can be considered as analogues of the popular Eigenstrat method (Price et al., 2006) which seeks to control for the confounding caused by ancestry differences between cases and controls in genome-wide association studies by correcting along eigenvectors estimated from a covariance matrix derived from the genotypes of the cases and the controls, and uses linear regression or simple correlation to test for association. Our application differs from those which typically employ Eigenstrat in that the confounding in a host-viral association study does not stem from the hidden ancestries of the individuals in the

sample, but from the unobservable relationships between viral sequences circulating within the host population from which the sample was drawn. The population structure we seek to capture is that obtained from the viral sequences rather than from the individuals' genotypes.

The HLA genes are the most polymorphic in the human genome, and so our HLA data consists of observations on a large number of low frequency variables. It may be more appropriate in a host-viral context, therefore, to implement a PCA-based correction within a logistic regression format. We have compared the conventional implementation of a PCA-based correction in a logistic model with an alternative projection-based approach motivated by consideration of the Frisch-Waugh-Lovell theorem. In addition, we have trialled a modification of our proposed approach which accommodates problems with separation not previously considered in regard to PCA-based correction procedures in large-scale association studies. These suggestions may have wider applicability beyond the context of host-viral analysis.

We demonstrate the utility of our proposed approaches in detecting true associations whilst minimizing confounding by founder effect-generated associations via a simulation study which utilises host-viral data taken from the Western Australian HIV Cohort (Mallal, 1998). For these simulations we construct a pool of underlying host HLA profiles and a pool of viral amino acid sequences, randomly reallocating them to construct data sets with biologically real HLA profiles and sequences but with no HLA-sequence associations. Known HLA-driven mutations and founder effects are then superimposed.

Such univariate analyses may be inadequate to fully identify the complex relationships which may exist between human immune mechanisms and viral escape through amino acid mutation, however, and we therefore also consider incorporating the proposed PCA

correction within well known linear-based multivariable statistical techniques including multiple regression and canonical correlation. This extension to multivariate techniques offers host-viral association studies the potential to identify associations between groupings of HLA alleles and small sections of viral sequence, increasing the scope of the questions which can be considered. We canvas a range of issues of interest which can be addressed through the application of such PCA-corrected multivariate approaches, and in particular, consider the use of PCA-corrected canonical correlation analysis to assess the proportion of overall variation in escape mutations attributable to HLA immune pressure.

In the final chapter of this thesis we discuss problems inherent in the application of false discovery rates and q-values for the determination of significance thresholds in large-scale host-viral analyses based on pooled contingency tables with varying margins. As these procedures implicitly assume continuity of the underlying p-value distributions, Pounds and Cheng (2006) questioned their use in analyses based on discrete statistics, suggesting modifications to accommodate the discreteness, particularly in relation to the estimation of the null proportion. However, constraints imposed by the margins in analyses based on pooled contingency tables introduce additional complexities which Pounds and Cheng did not address, and the necessarily skewed distribution of Fisher p-values, for example, further complicates the estimation of this proportion. Here we propose additional modifications to accommodate issues specific to analyses based on pooled Fisher p-values. We also consider the application of q-values in such analyses, and suggest that their large scale could be leveraged to allow for local inference. Deviations between observed and expected p-value distributions may provide an alternative method of identifying hypotheses of interest in host-viral analyses.

This thesis has been structured as follows: in Chapter 1 we discuss the issues motivating the research in greater detail, describing the process by which HLA pressure induces

escape mutations in the HIV viral sequence, and the importance of research into viral immune escape processes in underpinning the development of a vaccine to control the worldwide pandemic. We describe the nature of host-viral data, and further describe the problems posed by founder effects in confounding its analysis. In Chapter 2 we develop the PCA-R founder effect correction approach for host-viral analyses by adapting Eigenstrat to apply within the context of host-viral association studies and consider its properties through a discussion of the implications of the Frisch-Waugh-Lovell theorem. We outline the procedures employed to simulate data sets to assess the efficacies of the proposed methods, and trial the PCA-R procedure on our simulated data sets. In Chapter 3 we implement the PCA-based founder effect correction within a logistic regression model, comparing the conventional approach (termed here PCA-L) with our proposed projection-based correction methods (termed PCA-P and PCA-FP). This work incorporates and extends that presented in Reeves et al. (2012). One implication drawn from the FWL theorem is that a founder-effect correcting procedure could be implemented using any set of linearly independent vectors spanning the sub-space hypothesised to describe the population structure of the viral sequences, and so we also consider implementing the correction using factor variables estimated from the clustering algorithm Partitioning Around Medoids (PAM) (Kauffman and Rousseeuw, 1990) in place of eigenvectors, and we compare these results with those from implementing the PCA-corrected methods and the PAM-partitioned Mantel Haenszel procedure suggested in Rauch et al. (2009). In Chapter 4 we extend the PCA-correction by integrating it within well-known multivariate procedures to broaden the scope of problems which can be investigated in host-viral analyses, particularly by incorporating simultaneous carriage of groups of HLA alleles and windows of viral sequence. In Chapter 5 we discuss issues relating to the application of false discovery rate procedures in host-viral analyses based on pooled contingency tables with varying margins. In the

Conclusion we present a general discussion of the results of this research and suggest directions for further work.