THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

STATISTICAL ASSESSMENT OF ASSOCIATION BETWEEN GENOMIC AND IMMUNE RESPONSE DYNAMICS: CLUES TO LENTIVIRUS PROTECTION

NICHOLAS D. HUSO
SPRING 2011

A thesis
submitted in partial fulfillment
of the requirements for baccalaureate degrees
in Science and Statistics
with honors in Statistics

Reviewed and approved by the following*:

Francesca Chiaromonte
Professor of Statistics and
Professor of Public Health Sciences
Thesis Supervisor

Mary Poss
Professor of Biology and
Professor of Veterinary and Biomedical Sciences
Thesis Supervisor

R. A. Markle
Director of Premedicine and General Science Majors
Professor of Biology
Honors Advisor

*Signatures are on file in the Schreyer Honors College
ABSTRACT

Lentiviruses like the Human Immunodeficiency Virus (HIV) are a well-recognized and important source of disease in both animal and human hosts. They typically induce a progressive decline of the host immune system, which leaves the host susceptible to other infections. Cats, harbor lentiviruses and exhibit symptoms similar to HIV. This allows researchers to use animal models to investigate factors affecting disease progression in a lentiviral infection. Two recent studies have shown that (i) the impact of a Feline Immunodeficiency Virus (FIV) infection can be attenuated if the cats are first infected with a strain of lentivirus derived from a wild cougar, and (ii) the lentiviral populations within cats subject to single (FIV only) and dual (FIV preceded by the wild cougar virus) infections undergo different evolutionary trajectories. The mechanisms underlying these differences are of great interest because of the implications they could have on our understanding and handling of lentiviral infections. In this thesis, a variety of statistical techniques are used to re-analyze some of the data collected in the two above-mentioned studies. Focusing on two time points, (prior and following a bottleneck identified in the viral population) we investigate the association between changes in the host immune parameters and changes in the viral genome and how this association differs between single and dual infections.
# TABLE OF CONTENTS

Abstract .......................................................................................................................... i

List of Figures ............................................................................................................... iii

List of Tables ............................................................................................................... iv

Abbreviations ............................................................................................................... v

Acknowledgements ...................................................................................................... vi

Introduction .................................................................................................................. 1

Materials and Methods ............................................................................................... 11

Results ......................................................................................................................... 19

Conclusion ................................................................................................................... 26

References .................................................................................................................... 28
LIST OF FIGURES

Figure 1: Example of Lentivirus Replication Cycle. ......................................................... 1

Figure 2: Alterations in genomic diversity and population dynamics of FIV-C provided by apathogenic dual infection with PLV................................................................. 9

Figure 3: Host immune changes vs. changes in transitions in the viral genome ............. 21

Figure 4: Host immune changes vs. changes in transversions in the viral genome........... 23
LIST OF TABLES

Table 1: Cat Identification and Assignment .......................................................... 11

Table 2: Immunological Response Parameters. ......................................................... 13

Table 3: First and Second Principal Components of Pre- to Post-Bottleneck Differences in Immune Parameters. ................................................................. 19

Table 4: Output for the fit of model (1A) .................................................................. 21

Table 5: Output for the fit of model (2A) ................................................................. 23

Table 6: Output for the fit of model (1B) ................................................................. 24

Table 7: Output for the fit of model (2B) ................................................................. 24

Table 8: Output for the fit of model (3A) ................................................................. 25
ABBREVIATIONS

HIV- Human Immunodeficiency Virus
FIV- Feline Immunodeficiency Virus
FIV-C-Strain C of FIV
FIVpco-Puma Lentivirus
PLV- Strain of Puma Lentivirus (avirulent in domestic cats)
FIVC- Feline Immunodeficiency Virus subtype C (virulent strain in domestic cats)
PBMC- Peripheral Blood Mononuclear Cells
AID- Activation-induced cytidine deaminase
APOBEC- the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
PPCA-Probabilistic Principal Component Analysis
PCA-Principal Component Analysis
PC-Principal Component
PC1-First Principal Component
PC2-Second Principal Component
LM- Linear Model
ACKNOWLEDGEMENTS

• I would like to thank Dr. Francesca Chiaromonte. I cannot thank her enough for the time, patience, and thought she put in to the process of creating this thesis.

• I would also like to thank Dr. Mary Poss. Without her guidance and support this thesis would not have been possible.

• Thanks also to Guruprasad Ananda for his help with Galaxy.
INTRODUCTION

Lentivirus infection pits viral adaptations aimed at facilitating viral replication and persistence against host defense mechanisms directed toward restriction and clearance of viral organisms.

Figure 1: Example of Lentivirus Replication Cycle, Neil et al [1]. Steps in the replication cycle that occur during infection of target cells. The cycle begins with binding and entry into an uninfected cell, and proceeds through the production of mature progeny virions that leave the infected cell and can spread to additional uninfected cells.

While for many viruses, an understanding of classic virus-host interactions has lead to the development of effective vaccines, and even eradication, similar approaches have not yet yielded a satisfactory control strategy for lentiviruses. This has directed efforts towards
specific features of lentivirus-host interactions, new research perspectives, and emerging tools. Investigating the temporal evolution of viral genomes within hosts and its relationship to both known and hypothesized host anti-viral responses can illuminate novel aspects of lentivirus-host interactions, and may ultimately lead to new strategies to effectively control lentivirus infection; however, it requires a variety of computational and statistical techniques to analyze complex data collected during infection.

The advent of high-throughput experimental tools that generate large, complex datasets on genomic sequences and transcription levels has revolutionized various areas of the life sciences in the last decade, and has stimulated unprecedented efforts to develop computational resources to handle such data. Galaxy is one such tool, designed to import and process large, complex genomic data through a web-based platform [11-13]. In parallel, the development and application of novel statistical approaches has also become increasingly important to explore and interpret the large amounts of detailed data being generated [2,6-10]. The right combination of computational and statistical tools is especially vital for understanding complex relationships that are not captured by straightforward correlations.

Data collected in the “real world” is rarely ideal. One common issue is missing values, occurring when not all measurements are available for some of the units. If units with missing values are simply removed from the analysis, all information they can contribute is lost. If relatively few values are missing, however, it is possible to retain these units by imputing the missing values. This can be done with a variety of techniques that create “estimates” for the missing entries based on non-missing entries in the data [15-17]. These techniques are designed to handle possible biasing or variance-altering
effects of introducing artificial values in the data, and typically allow us to obtain better results than unit removal.

Another issue that is common whenever experimental replication is expensive and/or difficult, is sparsity or under-sampling; that is, a situation in which the number of available observations (independent statistical units) is small relative to the number of variables being considered (measurements taken on each unit) [4,5]. This complicates the application of most statistical methods. Techniques to augment sparse data are common in computer science (see [5] and references therein), and recently data augmentation by “noising” has been proposed as a means to regularize under-sampled data in statistical genomics [3]. In this approach, artificial replicates are created adding to the existing observations random variables representing small, unstructured perturbations. Results of any statistical procedure applied to the augmented data are more stable because of the larger sample size, and not sizably distorted/biased – provided the perturbations are appropriately defined.

In addition to missing data imputation and augmentation techniques, in this thesis we also use Principal Component Analysis (PCA). PCA seeks to reduce the dimensionality of a multivariate dataset comprising several, co-varying variables. This is accomplished by transforming the original variables into a new set of variables known as their principal components (PC). The PC’s are expressed as linear combinations of the original variables. They are uncorrelated by construction, and ordered so that the first few capture most of the variation in the data cloud [18-19].

Finally, after appropriate dimension reduction, dependences are captured statistically in this study through Linear Models (LM) – a tool with applications in
virtually every discipline. LMs represent a response variable as a function of one or more
predictor variables in a probabilistic way – i.e. allowing for appropriately specified
random errors to affect the response [20]. A variety of least squares and likelihood-based
methods are utilized to fit such a probabilistic relationship; that is, to estimate and make
inferences on the parameters of a LM, and to assess its predictive power [21]. In our
analysis we will exploit the flexibility of LMs by introducing a qualitative predictor, as
well as a random effect used to control for potential undesirable consequences of the data
augmentation.

Lentiviruses are a non-oncogenic retrovirus family whose ability to replicate in
non-dividing cells makes them remarkably efficient at delivering their genetic material to
the host. Most recognize the well-known lentivirus of humans, i.e. the Human
Immunodeficiency Virus (HIV). However, many other lentiviruses exist in other species,
including a lentivirus of cats, i.e. the Feline Immunodeficiency Virus (FIV) [22-23].
It has been observed that cats infected with FIV subtype C develop disease similar to
humans infected with HIV, most notably a decline of CD4 T cells and immunodeficiency
[24-25]. Indeed, cats are the only non-primate hosts that develop an immunodeficiency
syndrome from a natural lentivirus infection. This animal model provides a number of
research opportunities not only towards understanding and developing cures for FIV
itself, but also for eventually gaining insight into HIV.

FIV-C infection typically manifests itself in a manner similar to HIV. Both of
these lentiviruses infect CD4 positive cells including T cells and macrophages and other
peripheral blood mononuclear cells (PBMC). The virus enters the cell by fusion of the
viral envelope with the cell membrane; the viral RNA genome is reverse transcribed to
viral DNA, which integrates into the host genome. This ‘proviral’ DNA provides a template for transcription and viral replication [22]. Infection stimulates both humeral and cell mediated adaptive immune responses against most FIV-C (and HIV-1) proteins; however, these responses fail to curtail viral replication and viral load, and eventually CD4 positive cells begin to decline in parallel with a decline in immune function [26].

An adaptive (acquired) host immune response is specific to a particular foreign agent, takes one to two weeks to develop (delayed), and maintains memory. In contrast, innate responses are a separate arm of the host defense system. They include physical or physiological barriers to entry, and various enzymatic and cellular effectors that inactivate foreign agents in a pattern specific manner. An important feature distinguishing these innate factors from the adaptive response is that they act immediately, as opposed to requiring a delay to develop [27].

One prominent mechanism of the innate response to lentiviruses and other viruses is the apolipoprotein B mRNA editing enzyme, cytidine deaminase. This naturally expressed host DNA deaminase edits viral genomes and induces mutations in their sequence – in particular, transitions from deoxyguanosine to deoxyadenosine (G to A) and inhibits the replication of retroviruses and DNA viruses [28].

To date, the most effective and widely used interventions against lentivirus infection utilize combinations of highly active pharmacological inhibitors that block multiple steps of the viral life cycle [29]. Since these interventions are expensive and require long term adherence, there is strong motivation to search for an effective vaccine. Modified live vaccines with attenuated strains have proven effective for numerous non-retroviruses and are being explored at various levels in lentivirus models; however,
limited progress has been made towards a vaccine, as well as in understanding immune and other correlates of protection from lentivirus-induced disease. An intriguing addition to this background is that there is evidence that preexisting viral infection reduces severity of disease when an individual becomes dually infected at a later date with a virulent lentivirus. For example, HIV-2 infection (less pathogenic than HIV-1) delays the onset of AIDS in individuals who later become dually infected with HIV-1 [30].

More than 20 species of felines have been shown to sero-react with lentiviral antigens [31]. Importantly, lentiviruses exhibit a significant degree of species specificity. Cougars (Puma concolor) are a large feline species often naturally infected with their own host-adapted feline lentivirus, the Puma Lentivirus (FIVpco) [32-36]. Cross-species infection of domestic cats with FIVpco strain PLV results in a clinically silent, persistent infection [37]. Extensive sequencing suggests that high G to A mutation frequencies due to cytidine deaminase activity contributes to restriction of PLV infection in domestic cats [38,39]. Notably, infection of domestic cats with PLV decreases the pathogenic effects of subsequent dual infection with FIV strain C [40]. Prior PLV infection prevents the FIV-C induced CD4 T-cell depletion which is characteristic of disease in cats infected with only FIV-C [41]. Interestingly, as with HIV-2/HIV-1, this protection does not appear to be due to adaptive immune responses [31]. Alternative explanations are therefore needed for the PLV-induced protection.

Recent studies suggest that changes in the dynamics of viral genomes induced by previous PLV infection correlate closely with protection from disease. Lentiviruses are constantly undergoing changes in their genome sequence due to rapid virus replication
rates, poor fidelity of the viral RNA transcriptase and the RNA pol II genes, and recombination during simultaneous infection with multiple virus strains.

Stochastic sequence changes have been termed genetic drift. Sequence changes that get fixed and lead to a higher than expected frequencies of a specific allele suggest positive (improved viral fitness) or purifying (escape from restriction) selection pressures, and this processes may favor specific mutations at certain sites.

In studying FIV-C mutational profiles at different times during FIV-C infection, clear differences were found between viral populations sampled from cats singly infected with FIV-C and cats initially infected with PLV and subsequently infected with FIV-C. This suggests that infection with the apathogenic PLV alters the population genetics of FIV-C infection. First, The FIVC mutation rate in dually infected cats is higher than the mutation rate in cats singly infected with FIV-C [25]. In addition, the effective virus population size in single and dual infected cats both showed rapid expansion during the first week after FIV-C infection as shown in Fig 2; however, previous PLV infection reduces the effective population size of FIV-C. Note that while the census population size represents a count of the total number of viral particles [38], the effective population size \( N_e (\theta=2N_e\mu) \) [42] estimates the total viral population connecting it directly to the mutation rate (\( \mu \)) and inversely to genetic drift (\( \theta \)). Genetic drift has a more profound effect on populations with small effective size, leading to higher mutation rates. Therefore genetic drift can reduce overall genetic variation and virus population fitness, an effect that becomes amplified as the effective population size decreases.

This concept is predicted by Muller’s Ratchet [43], which states that in a finite and asexual population the average fitness (ability to replicate) may decrease as
deleterious mutations build, since there is no opportunity for recombination [44].
Because Muller’s Ratchet relies on random changes in frequency of an allele in a population, this genetic drift is more pronounced the smaller the population. Muller’s Ratchet predicts that genetic drift and associated higher mutation rate in a decreasing viral population will cause excess deleterious mutations to occur, and eventually lead to a genetic bottleneck. Of particular interest, Padhi et al [38] found dramatic changes in the FIV-C population (see the Bayesian skyline plot in Fig 2), suggesting that Muller’s Ratchet leads to a genetic bottleneck in lentiviruses [29, 45].

A well-known hallmark of lentiviruses is their error prone methods of replication. Poor replication fidelity and absence of proofreading and repair mechanisms can lead to deleterious errors, an enhancement of Muller’s Ratchet, and thus substantial and often negative changes in the fitness of the viral population over relatively short time intervals [46].
Figure 2: Bayesian Skyline plot of effective population size against time (from right to left), from Padhi et al [38]. The time begins at infection with FIV-C, 28 days after infection with PLV. The black lines (solid and dotted) represent FIV-C population in dual (PLV and FIV-C) infected cats, with its 95% confidence interval. The grey lines (solid and dotted) represent FIV-C population in single (FIV-C only) infected cats, with its 95% confidence interval. The viral population in dual infected cats exhibits a bottleneck highlighted by the grey shaded region.

There are several potential explanations that may account for the differences in effective population size between single and dual infected cats. First, the difference may be due to
a diminished availability of susceptible cells in the dual infected cats – however, no evidence was found to suggest this as the basic mechanism at play. Second, although there were no detectable differences in humeral and cell mediated adaptive immune responses between single or dual infected cats, differences below the detection limits of the assays used could be affecting the differences observed in the viral population dynamics. Notwithstanding this important caveat, a third explanation appears more likely; if adaptive immunity is not the main cause for the observed differences, priming of target cells by prior PLV infection may lead to a host environment poised to respond more effectively through *innate* factors, such as the cytidine deaminase editing system. This could have important implications for developing antiviral intervention strategies.

It has been previously shown that the immunological profile is substantially different in single and dual infected cats [24]. Namely single infection with FIV-C elicits an overall increase in the innate immune response parameters IL10, IL4, and IL12 compared to cats dually infected with PLV and FIV-C. Additionally, an overall decline in CD4 cells is observed in single infected cats while dual infection with PLV seems to largely arrest this phenomenon. While these differences have been shown to clearly differentiate between single and dual infected cats, previous studies stopped short of exploring some of the relationships between the immune response and the behavior of the virus itself.

In this study we undertook a statistical analysis of possible relationships between changes in the genome of the FIV-C virus and changes in the host adaptive and innate immune response, especially with regards to the observed bottleneck (Fig 2). Keeping in mind the likely importance of innate defense mechanisms, we considered several
parameters related to the host’s innate immune response in addition to immune
parameters related to its adaptive immune response – giving us a wide spectrum of
information on changes in the host immune system. In particular, we wish to investigate
whether the relationship between change in the viral genome and (innate and adaptive)
host immune responses differs significantly between cats infected with FIV-C only, and
cats infected with PLV prior to FIV-C infection.

MATERIALS AND METHODS

Experimental Design and Data Collection

The experimental design is described in detail by Terwee et al. [41]. Briefly, on day one,
a group of ten cats was inoculated with PLV intravenously and a second group of ten cats
was sham inoculated. At day 28, five randomly selected cats from each of the two groups
were inoculated with FIV-C intravenously. This led to four groups of five cats each: two
single infected groups (one with PLV and one with FIV-C), one dual infected group (both
PLV and FIV-C), and one uninfected control group. The 20 cats were identified
individually by numbers 4189 through 4208. In our analysis, as in Padhi et al [38], we
considered only two of the groups; namely, cats singly infected with FIV-C, and dually
infected cats. Moreover, we considered only four cats in each of the two groups for which
sequencing of FIV-C populations was performed [38].

<table>
<thead>
<tr>
<th>Cat</th>
<th>Infection Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>4189</td>
<td>Dual</td>
</tr>
<tr>
<td>4191</td>
<td>Single</td>
</tr>
<tr>
<td>4196*</td>
<td>Dual</td>
</tr>
<tr>
<td>4197</td>
<td>Single</td>
</tr>
<tr>
<td>4202</td>
<td>Single</td>
</tr>
<tr>
<td>4203</td>
<td>Dual</td>
</tr>
</tbody>
</table>
Table 1: Cat Identification and Assignment: The eight cats shown were randomly assigned to the single (FIV-C only) and dual (FIV-C and PLV) groups. The sequence data (see below) for the cat marked with an asterisk contained one hypermutated clone.

Our analysis therefore comprises eight cats in total. For these eight cats we had viral sequence data, as well as measurements for an array of immune parameters, collected along two different time courses during progression of the FIV-C infection (see Fig 2).

Selection of Time Points

The bottleneck described in [38] is of great interest as it captures a crucial difference between the FIV-C populations inhabiting single and dual infected cats. Working with the time courses used to gather viral sequence and host immune parameters information, we selected times before and after the bottleneck as to achieve a reasonably close match between sequence and immune parameter data. We selected days 45 and 115 post PLV infection for the sequence data, and days 45 and 80 post PVL infection for the immune parameters data. A rationale for these choices is that Fig. 2 shows a fairly stable behavior of the effective population size before and especially after the bottleneck. However, our results are meaningful only as broadly referred to pre- and post-bottleneck time periods (not specific days).

Data on Host Immune Parameters

Using the procedures described in Terwee et al [41] and Roy et al [24], eleven immune response parameters were measured on cats at various time points. Cytokines were
quantitated by real time PCR and Cytokine expression for IL-10, IL-12 and interferon gamma was quantitated relative to that of the gene GAPDH. White blood cell counts were measured using a Coulter Z1. The percentages of lymphocytes positive for CD4 were determined by flow cytometry. Total cell counts for each phenotype were calculated by multiplying the total number of white blood cells by the percentage of lymphocytes in the sample as determined by the differential count and then by the percentage of lymphocytes expressing that particular phenotype. Because of missing data for the time points we selected to represent pre- and post-bottleneck behavior, we eliminated CD8 and CD25 and considered only nine of the original eleven immune parameters.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-4</td>
<td>B-cell growth factor, ‘Th2’ cytokine</td>
</tr>
<tr>
<td>IL-10</td>
<td>B-cell survival and proliferation, ‘Th2’. Generally antagonistic to TNF-α</td>
</tr>
<tr>
<td>IL-12</td>
<td>Stimulates production of N-γ and TNF-α, ‘Th1’</td>
</tr>
<tr>
<td>TNF-α</td>
<td>Stimulates systemic inflammation regulates apoptosis, neutrophil chemoattractant</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>Proinflammatory cytokine, stimulates IL-12 and TNF-α, antagonistic to IL-4, ‘Th1’</td>
</tr>
<tr>
<td>FAS</td>
<td>‘Death receptor’, induces apoptosis</td>
</tr>
<tr>
<td>Lymph</td>
<td>T and B lymphocytes, NK cells and monocytes</td>
</tr>
<tr>
<td>CD4</td>
<td>Cell Surface marker for T helper cells (lymph subset)</td>
</tr>
<tr>
<td>Neutr</td>
<td>Neutrophils: granular leukocytes, phagocytic. (innate immune system)</td>
</tr>
</tbody>
</table>

Table 2: Immunological Response Parameters from Roy et al [24]. The table shows the nine immune parameters used in our study, with a brief description of each one. The Th1 and Th2 cytokines mentioned in the descriptions are soluble factors modulating innate and adaptive immune response.

Data on Viral Genome Sequences

As described in Padhi et al [38], 4604 base pairs from the 3’ end of the FIV-C genome were sequenced from samples extracted from blood of four single infected cats (FIV-C) and four dual infected cats (FIV-C and PLV). Nested PCR amplifications were performed using 3FIVC5F and 3FIVC6R. Conditions for the amplification were 35 cycles of 95C
denature (30 sec), 52C anneal (30 sec), 72C extend (5:00 min). Second round primers contained NotI and BamHI restriction sites and conditions were 3FIVC7NotF and #FIVC8BamR. An annealing temperature of 54C was used for second round PCR. Products were gel purified, digested with the appropriate enzyme, and cloned into a NotI, BamH1 digested pBS vector. Up to 8 clones were produced for each cat. The DNASTAR package was used to assemble and edit the sequences. One clone from a dual infected cat (4196) at day 45 appeared to be hypermutated, but was still considered in this study.

Data Preprocessing

Preprocessing of Immune Parameter Data: Missing entries in the immune parameter data were imputed using Probabilistic Principal Components Analysis (PPCA), as implemented in the R package pcaMethods (see http://bomserv.univ-lyon1.fr/~dray/files/software/ppca.R) [47]. PPCA is particularly effective in this case as relatively few components are needed for a reasonable degree of accuracy in missing data imputation. The preprocessed immune parameter data thus consisted of two paired vectors (pre- and post-bottleneck) in 9 dimensions (the immune parameters) for each of the 8 cats – with all 16 vectors are complete because missing entries have been imputed.

Preprocessing Sequence Data: For each cat, we started with two sets of clone sequences, one pre- and one post bottleneck, and identified mutations in the viral genome by comparing all clones to a reference viral sequence. For each clone, counts of transitions and transversions from the reference were produced using the Galaxy tool diffseq which, similar to a local alignment algorithm, is designed to report differences between two nearly identical sequences (see http://main.g2.bx.psu.edu/) [48-49]. For our analysis we
used the default settings of *diffseq*; namely, a “wordsize” of 20 and no reporting of differences at the ends of the sequences being compared. We then averaged the number of transitions and the number of transversions across clones, pre- and post bottleneck. The preprocessed sequence data thus consisted of a pair of average transitions counts (pre- and post-bottleneck), and similarly a pair of average transversions counts, for each of the 8 cats.

**Data Augmentation**

Since our analysis involves a large number of sequence and immune parameter features, and sample size available to us is small (8 cats), we decided to perform data augmentation following the approach described in [3].

**Augmentation of Immune Parameter Data:** For each cat, we generated 10 “artificial cats” whose pre and post bottleneck vectors of immune parameters are those of the original cat, perturbed by adding draws from a 9-dimensional Gaussian. The Gaussian had independent components with means 0, and variances equal to 1/100 of the variances observed across cats for the 9 immune parameters. The augmentation thus resulted in 88 paired pre- and post- bottleneck vectors (8 for the actual cats, plus 80 for the artificial cats). We then formed differences between these pairs for each (actual and artificial) cat, obtaining 88 difference vectors in 9 dimensions.

**Augmentation of Sequence Data:** For each cat, we generated 10 “artificial cats” whose pre and post bottleneck transitions (and separately transversions) counts are those of the clones for the original cat, perturbed by adding draws from a Poisson. The Poisson parameter, representing mean and variance of the distribution, was fixed at $\lambda=.5$ for transitions and at $\lambda=.25$ for transversions. These values were selected to be a quarter of
the smallest observed transition or transversion counts observed across clones and cats.

Next, as was done for the actual cats, we averaged the number of transitions and the number of transversions across clones, pre- and post bottleneck. The augmentation thus resulted in 88 pairs of average transition counts (pre- and post-bottleneck), and similarly 88 pairs of average transversion counts (8 for the actual cats, plus 80 for the artificial cats). We then formed differences between average transitions counts pairs and average transversions counts pairs for each (actual and artificial) cat, obtaining 88 transitions differences and 88 transversions differences.

In summary, augmentation provided us with 88 “observations” of immune responses and viral sequence changes pre and post bottleneck, quantified as 9-dimensional immune parameter difference vectors, and transitions and transversions differences. Note that while for the 8 actual cats immune parameter and sequences differences are naturally linked, for artificial cats the link is “random” conditional on the actual cat they are generated from (i.e. any instantiation of perturbed immune parameter information generated from an actual cat can get paired with any instantiation of perturbed sequence information generated from the same actual cat).

Principal Components Analysis

We use Principal Components Analysis to reduce the dimension in which we represent immune parameters changes. In particular, we run PCA on the 9-dimensional data cloud of 88 immune parameter difference vectors using the R function *prcomp* (http://stat.ethz.ch/R-manual/R-patched/library/stats/.../prcomp.html) [50], and extract the first component; PC1. This is the linear combination of the 9 original differences along which the data cloud presents the largest spread. Its coefficients (the loadings)
allow us to interpret it in terms of the original immune parameters. We also extract and interpret the second principal components (PC2), but we focus on PC1 when fitting linear models (see below, and Results section). Importantly, PCA was applied to the correlation matrix, i.e. to the centered and standardized coordinates, using the logical operators within \textit{prcomp}. This rescaling is crucial because the immune parameters, and thus their differences, are expressed in different units and vary on different magnitudes [24].

\textbf{Linear Models}

After PCA, we have 88 “observations” each comprising a 1-dimensional summary of immune parameter differences (PC1), a transitions difference (TSD), and a transversions difference (TVD). In addition, for each observation we have a binary variable expressing the group to which the cat belongs (singly infected \(G=0\); doubly infected \(G=1\)) – for artificial cats, the binary label is the same as that of the actual cat they are generated from.

Our objective is to study how differences in immune parameters are related to differences in sequence, and whether this relationship is affected by the single or dual infection status. Technically, we therefore fit linear models for PC1 (response) as a function of TSD and/or TVD (quantitative predictors), as well as \(G\) (categorical predictor in binary encoding). This was performed using the \textit{lm} function from the R base package [50]. In particular, we used the models:

\begin{align}
PC1_{ij} &= \beta_0 + \beta_1 \text{TSD}_{ij} + \delta_0 G_{ij} + \delta_1 \text{TSD}_{ij} G_{ij} + \epsilon_{ij} \quad (1A) \\
PC1_{ij} &= \beta_0 + \beta_1 \text{TVD}_{ij} + \delta_0 G_{ij} + \delta_1 \text{TVD}_{ij} G_{ij} + \epsilon_{ij} \quad (2A) \\
PC1_{ij} &= \beta_0 + \beta_1 \text{TSD}_{ij} + \beta_2 \text{TVD}_{ij} + \delta_0 G_{ij} + \delta_1 \text{TSD}_{ij} G_{ij} + \delta_2 \text{TVD}_{ij} G_{ij} + \epsilon_{ij} \quad (3A)
\end{align}

where \(i=1, 2, \ldots, 8\) indexes the original cats, \(j=0, 1, 2, \ldots, 10\) the augmentation (so \((i,0)\) is the original \(i\)-th cat, and \((i,1)\) through \((i,10)\) the “artificial cats” generated from it), and
\( \varepsilon_{ij} \sim N(0, \sigma_e^2) \) are independent and identically distributed Gaussian errors. The first two models above represent a pair of lines: \( \beta_0 \) and \( \beta_1 \) are intercept and slope of the line for PC1 on TSD (TVS) for single infected cats, and \( \delta_0 \) and \( \delta_1 \) the changes in intercept and slope for dual infected cats relative to single infected ones – in other words, the intercept and slope for the dual case are \((\beta_0 + \delta_0)\) and \((\beta_1 + \delta_1)\). The third model represents a pair of planes: \( \beta_0 \), \( \beta_1 \) and \( \beta_2 \) are intercept and two slopes of the plane for PC1 on YSD and TVD for single infected cats, and \( \delta_0 \), \( \delta_1 \) and \( \delta_2 \) the changes in intercept and slopes for dual infected cats relative to single infected ones.

Notably, the augmentation scheme we implemented introduces an additional structure in the variability of our data – because the 88 “observations” at our disposal tend to cluster by construction into eight subsets corresponding to the actual cats. This should not affect the results of our analysis, and accounting for this structure was not suggested in [3]. However, to verify that this is the case, we also fit linear mixed effects models parallel to the first two of the linear models described above, but comprising an additional random effect for a “cat factor” \( (C) \) – which groups the observations derived from each actual cat during augmentation. This was performed using the \textit{lme} function from the R package \textit{nlme} (http://cran.r-project.org/web/packages/nlme/index.html) [51].

For this exercise, the models used were:

\[
\text{PC1}_{ij} = \beta_0 + \beta_1 \text{TSD}_{ij} + \delta_0 G_{ij} + \delta_1 \text{TSD}_{ij} \times G_{ij} + C_i + \varepsilon_{ij} \quad (1B)
\]

\[
\text{PC1}_{ij} = \beta_0 + \beta_1 \text{TVD}_{ij} + \delta_0 G_{ij} + \delta_1 \text{TVD}_{ij} \times G_{ij} + C_i + \varepsilon_{ij} \quad (2B)
\]

where again \( i = 1, 2, \ldots, 8 \) indexes the original cats and \( j = 0,1, 2, \ldots, 10 \) the augmentation, \( \varepsilon_{ij} \sim N(0, \sigma_e^2) \) are independent and identically distributed Gaussian errors, and \( C_i \sim N(0, \sigma_C^2) \).
are independent and identically distributed Gaussian random effects associated with the 8 original cats.

RESULTS

First, we consider the first as well as the and second principal components of the pre- to post-bottleneck differences in immune parameters – computed on the data set comprising the original eight cats, plus the “artificial cats” generated with augmentation. Our aim is to consider the original immune parameters and their changes.

<table>
<thead>
<tr>
<th>Immune Parameter</th>
<th>PC1</th>
<th>PC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutrophils</td>
<td>-0.329</td>
<td>0.583683934</td>
</tr>
<tr>
<td>IL12</td>
<td>-0.06141720</td>
<td>0.434308739</td>
</tr>
<tr>
<td>IL10</td>
<td>-0.52237936</td>
<td>0.008006037</td>
</tr>
<tr>
<td>IL4</td>
<td>0.41522350</td>
<td>-0.251766020</td>
</tr>
<tr>
<td>IFNγ</td>
<td>-0.37024059</td>
<td>0.403710816</td>
</tr>
<tr>
<td>TNFα</td>
<td>-0.32476426</td>
<td>-0.345036020</td>
</tr>
<tr>
<td>Lymphocytes</td>
<td>0.38662672</td>
<td>0.343135980</td>
</tr>
<tr>
<td>CD4</td>
<td>0.39328706</td>
<td>-0.084193748</td>
</tr>
<tr>
<td>FAS</td>
<td>0.05424741</td>
<td>-0.019380244</td>
</tr>
</tbody>
</table>

Table 3: First and Second Principal Components of Pre- to Post-Bottleneck Differences in Immune Parameters. The table shows the first and second principal component for the nine immune parameters considered. CD4, and Lymphocytes (partially), capture adaptive response mechanisms. Neutrophils, IL12, IL10, IL4, IFNγ and TNFα capture innate response mechanisms. FAS is a cell surface marker on many cells. Sizeable positive and negative loadings are highlighted in blue and green, respectively.
Looking at Table 3, we see that PC1 explains almost a third of the overall variance of the data cloud, while the second principal component, PC2, explains a little over a quarter. PC1 captures a trade-off between changes in IL10, IFNγ and TNFα on one hand, and changes in IL4, Lymphocytes and CD4 on the other. The first have negative signs, signifying that the parameters increased after the bottleneck, and the second have positive signs, signifying that the parameters decreased (recall differences are computed pre-minus post-bottleneck). Notably, along this component all innate response parameters except for IL4 increase, and the two response parameters linked to adaptive defenses decrease.

Next, we focus on PC1 (pre- to post-bottleneck increase of IL10, IFNγ, and TNFα, and decrease of IL4, Lymphocytes and CD4) and consider the linear models used to express it as a function of changes in the viral genome.
**Figure 3: Host immune changes vs. changes in transitions in viral genome.** The first principal component of the pre- to post-bottleneck differences in immune parameters is plotted against the pre-to post-bottleneck difference in transitions for the augmented data (8 original cats, plus 80 “artificial” cats). Cats infected with FIV-C only are represented by gray filled squares, and cats infected with PLV prior to FIV-C by black circles. The gray (single) and black (dual) lines represent the fit of model (1A).

Fig. 3 indicates that, for cats infected only with FIV-C, there is a strong upward trend for PC1 as a function of TSD – spanning values from negative (for both PC1 and TSD) to positive (for both PC1 and TSD). However, the trend appears much weakened for dually infected cats – which also appear to concentrate on positive ranges on both vertical and horizontal axes.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.05089</td>
<td>0.19364</td>
<td>-5.427</td>
<td>0.0000</td>
</tr>
<tr>
<td>Transitions(TSD)</td>
<td>0.47294</td>
<td>0.08267</td>
<td>5.721</td>
<td>0.0000</td>
</tr>
<tr>
<td>Group(G)</td>
<td>1.70244</td>
<td>0.34520</td>
<td>4.932</td>
<td>0.0000</td>
</tr>
<tr>
<td>---------------</td>
<td>---------</td>
<td>---------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>Transitions*Group</td>
<td>-0.37189</td>
<td>0.11987</td>
<td>-3.102</td>
<td>0.0026</td>
</tr>
</tbody>
</table>

Adjusted R-squared: 0.4868  
F-statistic: 25.98 on 3 and 76 DF, p-value: 1.135e-11

### Table 4: Output for the fit of model (1A); PC1 vs TSD (continuous) and G (binary).

These observations are supported by the numerical output of the fit of model (1A), presented in Table 4. The relationship between PC1 and TSD has a significantly negative intercept and a significantly positive slope for single infected cats. For dual infected cats, the relationship erodes; the intercept has a significant increase, leading this group to start out at PC1 values around 0, and the slope has a significant decrease, so that PC1 does not increase nearly as much as TSD increases. Notably, the model explains almost 50% of the variability observed in PC1, according to the adjusted $R^2$ value.
**Figure 4: Host immune changes vs. changes in transversions in viral genome.** The first principal component of the pre- to post-bottleneck differences in immune parameters is plotted against the pre-to post-bottleneck difference in transversions for the augmented data (8 original cats, plus 80 “artificial” cats). Cats infected with FIV-C only are represented by gray filled squares, and cats infected with PLV prior to FIV-C by black circles. The gray (single) and black (dual) lines represent the fit of model (2A).

*IL10, IFNγ, TNFα higher post- than pre-bottlenk.*

*IL4, Lymp, CD4 lower pre- than post-bottlenk.*

Fig. 4 indicates that, for cats infected only with FIV-C, there is a downward trend for PC1 as a function of TVD – spanning values from positive PC1 and negative TSD to negative PC1 and positive TSD. Once again though, the trend appears much weakened for dually infected cats – which still concentrate on positive values of PC1, but show more negative values for TVD than for TSD.
These observations are again supported by the numerical output of the fit of model (1B), presented in Table 5. The relationship between PC1 and TSD has a significantly negative intercept and a significantly negative slope for single infected cats. For dual infected cats, the relationship erodes; both the intercept and the slope have a significant increase (although significance of the latter is somewhat marginal). In all, this model explains less of the variability in PC1 than the one in TSD – only approximately 35% according to the adjusted R² value.

To verify whether the results presented above are affected by the “clustered” structure introduced by the augmentation, we fitted linear mixed effect paralleling (1A) and (2A), including a random cat effect term.

Comparison of Table 6 with Table 4 shows the similarities between the two fits; results are consistent in sign and magnitude and again all terms are significant.
Table 7: Output for the fit of the mixed effect model (2B); PC1 vs TVD (continuous) and G (binary), with a random effect for cats added.

Comparing Table 7 to Table 5, we again see some similarities between the fits. However, significance overall appears to decrease. In particular, the negative slope for TVD is non-significant, suggesting no trend between PC1 and TVD also for the singly infected cats, and the change in slope between dually and singly infected cats is non-significant. This, along with the lower adjusted $R^2$ value for model (2A) (Table 5) than for model (1A) (Table 4), is evidence that the effects of TVD are less marked than those of TSD.

Finally, going back to standard linear models, we consider a fit comprising linear terms in both TSD and TVD, along with all terms required to separate single and dual groups.

Table 8: Output for the fit of model (3A); PC1 vs TSD and TVD (continuous), and G (binary)
As to be expected based on the above discussion, when both TSD and TVD are considered in the same model, TVD contributes very little. The TVD slope for singly infected cats and the change in TVD slope for dually infected cats, are both non-significant. Also the adjusted $R^2$ value dropped as compared to the one for model (1A) (see Table 4) – indicating that the explanatory contribution of the additional terms in model (3A) is negligible. In fact, TSD and TVD are inversely correlated for singly infected cats (correlation = -0.6513649, p-value = 5.318e-06), so it is quite possible that the negative relationship detected when fitting model (2A) is just an indirect consequence of the stronger, positive relationship detected when fitting model (1A).

In summary, we find strong evidence for a positive relationship between PC1 and TSD for cats infected with FIV-C only, and for the fact that a prior infection with PLV erodes much of this relationship.

**CONCLUSIONS**

We interpret our results as follows. For cats infected only with FIV-C, whose FIV-C populations do not undergo a bottle neck, transitions (relative to the viral reference genome) late in the course of infection can exceed or fall short of transitions early in the course of infection (TSD spans negative and positive values) – where late and early are defined as past or prior to the bottleneck mark in Fig. 2. For these cats, declining viral transitions (larger values of TSD) are associated with increases in the levels of IL10, IFN$\gamma$, and TNF$\alpha$, and decreases in the levels of IL4, Lymphocytes and CD4 (larger values of PC1). Changes in viral transversions early versus late after infection, on the other hand, do not seem to have sizeable effects on the host immune parameters.
For cats infected with PLV prior to FIV-C, whose FIV-C populations do undergo a bottleneck, transitions (relative to the viral reference genome) post-bottleneck tend to be fewer than transitions pre-bottleneck (positive TSD values). Consistently with the relationship between changes in the viral genome and changes in the host immune parameters detected for singly infected cats, for dually infected cats IL10, IFNγ, and TNFα are higher, and IL4, Lymphocytes and CD4 are lower, post- than pre-bottleneck (positive values of PC1). However, the relationship is statistically much weakened because TSD and PC1 don’t span as much of a range. In other words, after the bottleneck in their FIV-C populations, most of the dually infected cats are where only some of the singly infected cats arrive at the corresponding time after infection; that is, in the upper-right quadrant of the plane represented in Fig. 3. In terms of the variables considered in our study, this means that after an initial flurry, the number of transitions in the viral genome (relative to the reference) has decreased, and the host has a response profile characterized by increased levels of innate factors such as IL10, IFNγ, and TNFα, and decreased levels of other innate factors such as IL4 and adaptive factors such as CD4 and (partially) lymphocytes.
REFERENCES


Academic Vita

Nicholas D. Huso

Campus: 11 Atherton
University Park, PA  16802
Cell: 443-824-6346

Permanent: 16908 Prettyboy Dam Rd
Parkton, MD 21120

ndh142@psu.edu

Education:
Pennsylvania State University Schreyer Honors College
Double Major
  Bachelor of Science Degree in Statistics with Honors in Statistics
  Bachelor of Science Degree in Science

Thesis Title: Statistical Assessment of Association Between Genomic and Immune Response Dynamics: Clues to Lentivirus Protection

Thesis Advisors:
  Dr. Francesca Chiaromonte,
  Dept of Statistics and Dept of Public Health Sciences
  Dr. Mary Poss,
  Dept of Biology and Dept of Veterinary and Biomedical Sciences

Work Experience:
  Internship at Becton Dickinson in the Bioinformatics and Non-Clinical Statistics group, 2010
  The Cornerstone Professional Group in Web services, 2008
  Nittany Success Center, Tutor in Biology and Chemistry, 2008, 2009
  Construction and Carpentry, Lou Packett Home Construction, 2008-2009

Awards:
  Schwartz Family Scholarship, 2008
  Dean’s List, all semesters 2007-2011
  Penn State York Honors Award, 2009
  Member Phi Beta Kappa Honors Society

Presentations:
  Penn State York Honors Symposium, Analysis of Data using Healthy People 2010 and DNA Microarrays, 2009

Activities:
  Penn State York Varsity Basketball, 2007-2009
  Penn State Habitat for Humanity, 2009-2011
  Sandtown Habitat for Humanity, Baltimore, MD, 2008-2009
  Preschool Teacher, Grace Fellowship Church 2005-2010