University of Montana

ScholarWorks at University of Montana

Undergraduate Theses and Professional Papers

2015

Developing Microbial Biomarkers to Non-invasively Assess Health in Wild Elk (Cervus canadensis) Populations

Samuel B. Pannoni
University of Montana - Missoula, sam.pannoni@umontana.edu

Recommended Citation

Pannoni, Samuel B., "Developing Microbial Biomarkers to Non-invasively Assess Health in Wild Elk (Cervus canadensis) Populations" (2015). *Undergraduate Theses and Professional Papers*. 66.
https://scholarworks.umt.edu/utpp/66

*Developing Microbial Biomarkers to Non-invasively Assess Health in Wild Elk (*Cervus canadensis*) Populations*

By

SAMUEL B. PANNONI

Undergraduate Thesis

Presented in partial fulfillment of the requirements
for the degree of

Bachelor of Science
in *Wildlife Biology*

The University of Montana
Missoula, MT

May 2015

Approved by:

Professor Mark Hebblewhite
College of Forestry and Conservation

Professor William E. Holben
Division of Biological Sciences

Professor Jeffrey M. Good
Division of Biological Sciences

ABSTRACT

Pannoni, Samuel B., B.A., May 2015
Wildlife Biology

*Developing Microbial Biomarkers to Non-invasively Assess Health in Wild Elk (*Cervus elephus*) Populations*

Faculty Mentor: *Mark Hebblewhite*

Second Faculty Reader: *William Holben*

Third Faculty Reader: *Jeffery Good*

The composition of the intestinal bacterial community (intestinal microbiome) of mammals is associated with changes in diet, stress, disease and physical condition of the animal. The relationship between health and the microbiome has been extensively demonstrated in studies of humans and mice; this provides strong support for its potential utility in wildlife. When managing elk (*Cervus canadensis*), federal and state agencies currently must rely on invasive sampling and coarse demographic data on which to base their decisions. By developing microbiome-based biomarkers that vary as a function of elk body condition and disease (i.e. microbial biomarkers), we hope to provide managers with the ability to monitor direct impacts from environmental stressors on individual animals and the herd. This approach, once established, represents a low cost, non-invasive sampling method based simply on fecal pellet collection in the field and intestinal microbiome analysis in the lab. Montana Fish, Wildlife and Parks collected the scat and linked body condition metrics from four GPS collared populations in Montana in winter 2014, using helicopter teams and invasive sampling methods. We analyzed 111 individual wild elk fecal microbiomes using Illumina MiSeq sequencing of partial 16S-rRNA gene amplicons. Using the QIIME pipeline and a floating search feature selection algorithm (SFFS) with linear discriminate analysis (LDA) and leave-one-out cross validation (CV) we were able to elucidate informative patterns in bacterial taxa presence and abundance by comparing them to various measured body conditions and geographic locations of elk sampled. Microbial biomarkers provide potential for managers to routinely obtain fine scale non-invasive health metrics from scat samples obtained in the field for species of concern.

INTRODUCTION

Western landscapes are home to charismatic game species that are managed under sustainable yield mandates to provide for their public use and enjoyment [4]. When managing game animals (or threatened and endangered species for that matter), federal and state agencies currently must rely on invasive sampling for reliable health data [5]. Managers and biologists often use diverse data sampling methods paired with population models to inform wildlife management actions across large habitat ranges in order to ensure this maximum sustainable yield. These population models traditionally incorporate recruitment, survival, emigration, and immigration estimates for multiple age classes. More recent models have been evolving in complexity to include environmental variations and the stochasticity affecting these parameters, and have therefore become more accurate [6]. However, our limited understanding and ability to measure how local environmental effects influence individual animals (e.g. their survival and reproduction) has caused this source of uncertainty to remain largely undefined in wildlife population modeling, leading to poor predictive power, even when using well informed models.

We have begun to develop a new, non-invasive data source that captures feedback from environmental-host condition, called the fecal microbiome. This has value in itself for informing relative health of populations and proximate health of environments. Although not specifically tested here, this data source could characterize and reduce the uncertainty around stress sources we currently conclude to be random or stochastic, which plagues current population modeling approaches. This “truthing” could be accomplished by directly monitoring individual animal health, including disease presence/absence and effects, and body condition, based on microbial biomarkers within fecal pellets across wide geographic areas. This parameter can then be more generally integrated into better estimates of survival and recruitment parameters that are ubiquitously found in population models. As the relationships between environmental-animal-microbiome feedbacks are established, the microbiome could function as a proxy for assessing species-specific environmental needs and monitoring goal-oriented habitat improvements. Further understanding the connection between environment and animal health, and predicting its effects using microbiome data can help resolve some of the general shortcomings inherent in current population modeling and provide more information to wildlife managers.

The composition of the intestinal microbiome of mammals is associated with changes in diet, stress, disease and physical condition of the animal [1]. These characteristics make microbiomes ideal for generally informing animal condition. As such, the relationship between host health and the microbiome has been extensively demonstrated in studies of humans and mice [2, 3]. This prior wealth of research provides strong support for applying these microbiome characteristics as a monitoring tool for wildlife management.

One recent exploration into elk (*C. canadensis*) rumen bacterial communities provided support for a "core microbiome" shared by elk [7], but this research did not focus on useful patterns of variation in bacterial taxa at scales of taxonomic resolution finer than phylum-level. We utilize the potential of Genera level resolution to inform correlates of environmental stress acting on the host animal. Our development of 16S small subunit rRNA gene-based (hereafter 16S-based) microbial biomarkers in elk shows promise for using an individual's microbiome composition to distinguish between and predict states of health as it does in humans and mice.

HYPOTHESES

H1 - Consistent Health Utility Hypothesis: *Microbiome composition can be used to accurately predict and cluster between different states of elk health (e.g. as described by measured body-fat) in all individuals regardless of source population.*

- If aspects of elk fecal microbiomes are associated with individual body condition at high taxonomic resolution, then consistent presence and abundance (or for that matter absence) of specific bacterial taxa in fecal pellets will reliably predict states of body-fat in individuals regardless of location or population (presence of a strong overriding signal for the health biomarker).

H2 – Population-Specific Health Utility Hypothesis: *Aspects of elk intestinal microbiome diversity that predict health (and other metrics) will be driven by population-specific factors (strong local effect for the health biomarker).*

- If elk biomarkers are driven by population-specific genetic factors, then biomarker signals of health will be strongest for individual populations and when different populations are combined and features are selected, the signal strength will suffer (as measured by CV accuracy and LDA).

H3 – Microbiome Biogeography Hypothesis: *Aspects of the elk intestinal microbiome will be predictive of an individual's location, indicating biogeographic influence.*

- If aspects of elk microbiome composition are driven by local factors such as food type, quality and availability, which will be manifested at the level of location (i.e. host biogeography), the microbiome will be more homogenous within social groups and less homogenous between spatially isolated groups (causing clear LDA clustering and high CV accuracy).

METHODS

Durable Equipment

This project was supported by existing laboratory infrastructure of the Holben Lab in the Health Sciences Building at the University of Montana including all of the durable equipment needed for preparing samples for sequencing.

Sample Collection:

For our study, we received fecal pellet samples from wild Montana elk of known physical condition. Collection of scat samples, body condition metrics and GIS collaring of elk were conducted in February 2014 by an ongoing collaboration developed with Dr. Kelly Proffitt and Montana Fish, Wildlife and Parks (MTFWP). This sampling event used currently available and accepted invasive methods for wildlife immobilization, measurements of digesta-free body fat, sex classification, age, and thyroid screening [5]. Linked invasive health metrics and non-invasive fecal pellet collection (collected outside the body following expulsion by the animal) were gathered from individual elk from four populations across Montana including the Bitterroot Mountains, Sapphire Mountains, hunting district 311(Black's Ford) and the Tobacco Root Mountains. A small subset of these data (including data from all male individuals), contained incomplete additional metadata and therefore were only used in a subset of microbial-host comparisons.

Sample Preparation and Sequencing

We obtained next generation sequencing (NGS) data from 16S rRNA gene amplicons focusing on the V4 & V5 variable regions in the rRNA gene using a generally conserved 16/18S-specific barcoded primer set and PCR to classify the taxa present in fecal samples. The barcoded primer sequences used were 536F for forward and 907R for reverse priming [8]. Once amplified, the samples were gel purified using the QIAGEN Gel Purification kit (QIAGEN, Germantown, MD) following the manufacturer's recommended protocol for downstream direct sequencing. This gel purification step separates any 18S eukaryotic DNA amplicons or potential PCR artifacts produced by the generality of the PCR primer set, isolating and purifying the desired 16S bacterial amplicons for sequencing. Illumina MiSeq 300 bp paired-end sequencing of 16S amplicon libraries was conducted on all sampled individuals from the 4 populations (111 elk).

Sequence analysis:

The MiSeq sequence reads were filtered for quality and combined using Fastq-join with a minimum overlap of 6 bp [9]. Average pairwise alignments exceeded 80 bp for all forward and reverse combined reads making this minimum redundant. The QIIME pipeline, which combines many bioinformatics tools into a single package, was used to produce a table of OTUs to the genus level for downstream analysis [10]. Within the QIIME pipeline, Uclust [11] was selected for its open-reference OTU picking process where reads are clustered against a reference 16S sequence collection (in this case the Greengenes database) [12], and any reads which do not hit the reference sequence collection will be subsequently clustered de novo. This sequence classification process also uses UCHIME to detect chimeric 16S sequences (which were discarded) before proceeding [13]. An OTU matrix was produced at this step containing counts corresponding to the number of times each OTU was present in each sample. QIIME produces an OTU table in the form of a biological observation matrix file (BIOM), which is an attempt to provide file format consistency across the comparative "omics" realm, and was adopted here to support compatibility with future projects [14]. After the BIOM file was produced, the RDP II Ribosomal Database Project [15] was used to assign taxonomy to the OTU table at the genus level [16]. A new genus-level OTU table was produced at this stage including the RDP II taxonomy values (S_{ab} scores). To obtain β-diversity plots, a multiple sequence alignment was made using Pynast [17] and a phylogenetic tree built using FastTree2 [18]. With these files, MacQIIME [19] was utilized to produce α-diversity plots based on the Chao-1 metric [20] as well as principle coordinate analysis (PCoA) plots of beta-diversity using Emperor [21] from the updated taxonomy table.

Feature Selection and Cross Validation:

Metagenomic and 16S studies produce large amounts of data because of the need to sample microbial communities as deeply and completely as possible, but not all taxa have predictive power during statistical analysis for determining health or disease states of the host. We used a form of the Sequential Forward Floating Search algorithm (SFFS) to select for informative genera from the elk microbiome [22]. This algorithm selects a subset of genera from the total pool of those present using a heuristic or sub-optimal method that maintains (or minimally reduces) the performance of the complete data set. The complete data matrix would have been intractable to analyze and contains "noisy" genera that obscured the biological patterns present. SFFS avoids nesting issues where taxa or features are falsely fixed early in the selection process (an issue with other feature selection methods which results in reduced performance [23]). By allowing all features (genera) to be added or subtracted as the algorithm progresses (essentially "floating" the selections) features are allowed to interact to produce dynamic and unbiased performance results not dependent on starting conditions. The SFFS algorithm employed herein uses J3 scores, a form of scatter matrices that rewards close clustering within groups and rewards increased distance between groups of data points using Euclidean distances in multidimensional space. SFFS was developed in collaboration with colleagues in the Computer Science Department at UM (Spaulding, et al., manuscript in preparation [24]). Using the SFFS algorithm, we selected a feature number that provided the optimal performance in the model while avoiding potential over-fitting by comparing the cross validation (CV) performance differences between multiple numbers of features [24, 26]. We visualized this relationship with box plots and linear discriminant analysis (LDA)[25] (Figure 1) to choose the optimal number of features (genera) for the visualizations presented later in this work.

The LDA was performed with CV, which uses a leave-one-out method of training and testing to reduce over-fitting the model to the training data set as part of the SFFS [23, 25]. This method removes a sample from the training data, builds the model with remaining samples then tries to predict the classification of the removed sample. This leave-one-out method is iterated over all samples or "folds". A performance percentage is then calculated from the CV by summing the number of CV events (usually equal to the number of samples) in the denominator and summing the successful classification events in the numerator (e.g. 25/26 = 96.15% such that # of successful/total # of cross validation attempts = percent correct). The intent is that training the model in this way will allow it to function on future data sets of similar character with very little optimization necessary, potentially producing an optimized model for determining these states blindly from non-invasive scat samples without the current accompanying metadata necessary for this work to develop and validate the approach.

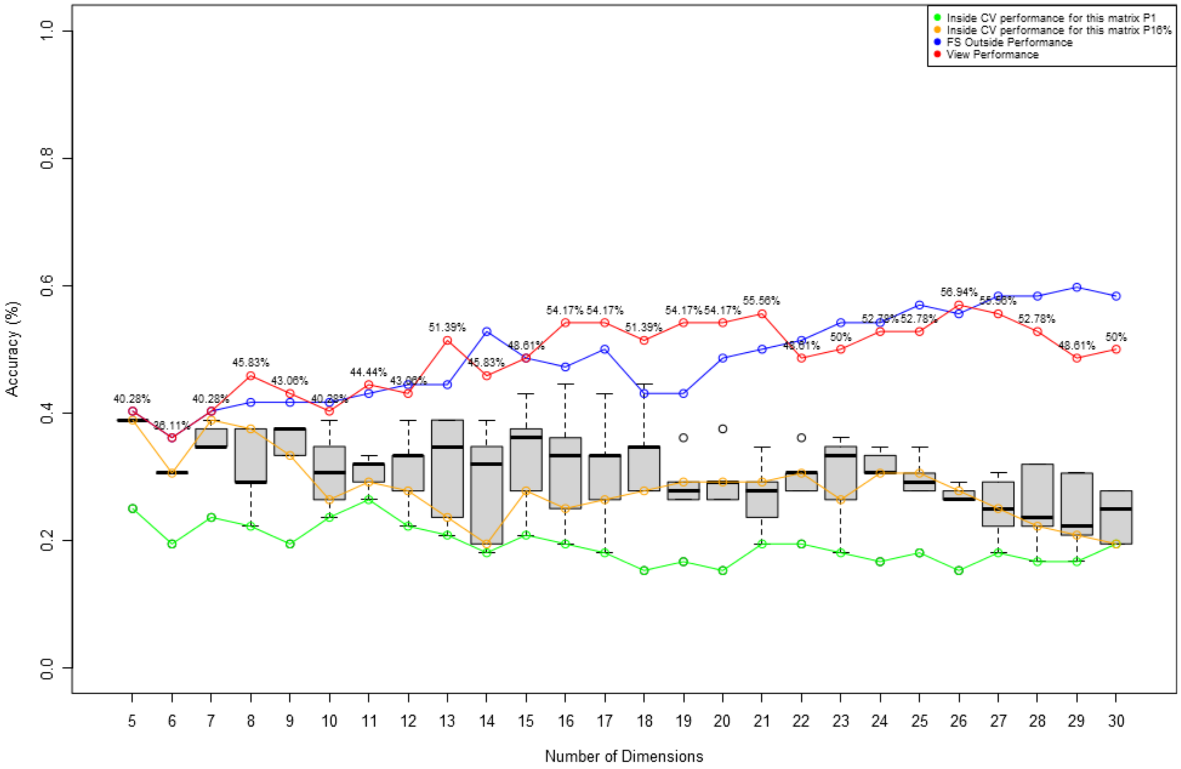

Figure 1. Box plots of cross validation (CV) accuracies (y-axis) with standard error whiskers from 4 types of CV approaches, selecting between 5 and 30 dimensions (x-axis) using female elk microbiome data combined from 3 populations in Montana stratified by body-fat. The green line indicates inside CV performance approach; the yellow line is inside CV performance for a reduced OTU matrix; the blue line is feature selection (SFFS) conducted outside of CV; and the red line is the performance for the LDA plots (one of which is presented later). This plot helps determine the optimal number of features to balance accuracy and reduce over-fitting of the algorithm, which in this case is 21 dimensions. This figure was produced in R using the FSSF package (Spaulding et al., personal communication).

RESULTS & DISCUSSION

Hypothesis 1: Consistent Health Utility

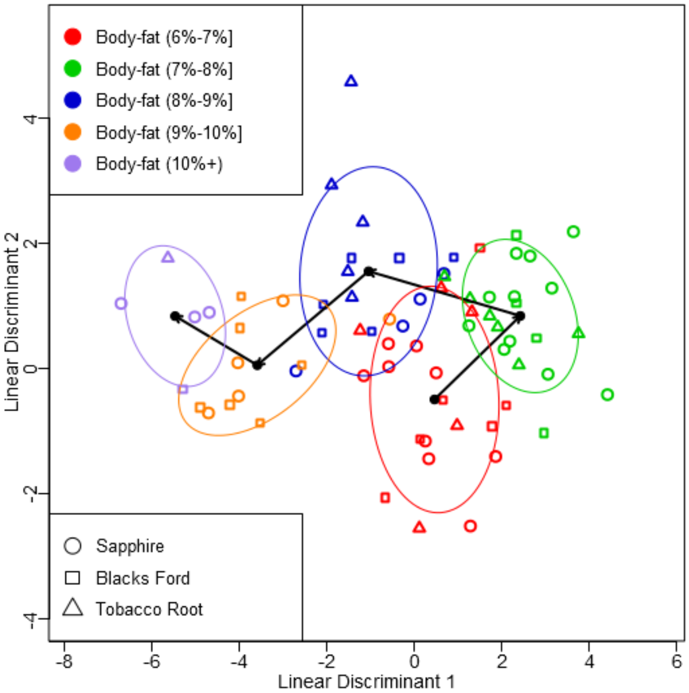

Figure 2. LDA ordination plot of elk microbiome samples from Sapphire (open circles); Black's Ford (squares); and Tobacco Root (triangles) populations as a function of body-fat. Colored circles represent elk of different body-fat percentage categories as indicated. Ellipses depict 1 standard deviation in the data and black circles are the centroid of each cluster. LDA was cross validated with the leave-one-out method producing 55.56% model accuracy. Reduced OTU data performed best at 21 dimensions when clustering between elk body conditions. Bitterroot population omitted from feature selection due to lack of body-fat data. Figure produced in R using the FSSF package (in development).

A clear pattern in microbiome compositional differences as a function of body fat content was observed for individual elk across populations as indicated by the low level of overlap between groupings (Figure 2). This supports the *Consistent Health Utility Hypothesis* and shows that comparative microbiome compositional analysis represents a useful noninvasive monitoring tool that informs individual animal health status. This was supported by leave-one-out CV. The moderate CV accuracy (55.56% reported, where 20% would be random), supports that our model is useful, but also suggests that our CV method may be suboptimal for the continuous (as opposed to discrete) structure of the body-fat data. Unlike biogeography data, which can accurately be represented as discrete variables, the gradient of microbiome characteristics that describe body-fat overlap near the imposed categorical cut-off points in the LDA. The leave-one-out CV approach attempts to fit and test this gradient against discrete body-fat categories, which leads to the appearance of reduced model accuracy because body fat content is a continuous variable. We expect an improvement in CV performance of the health-biomarker method once it is appropriately refined to validate continuous variables.

Hypothesis 2: Population Specific Health Utility

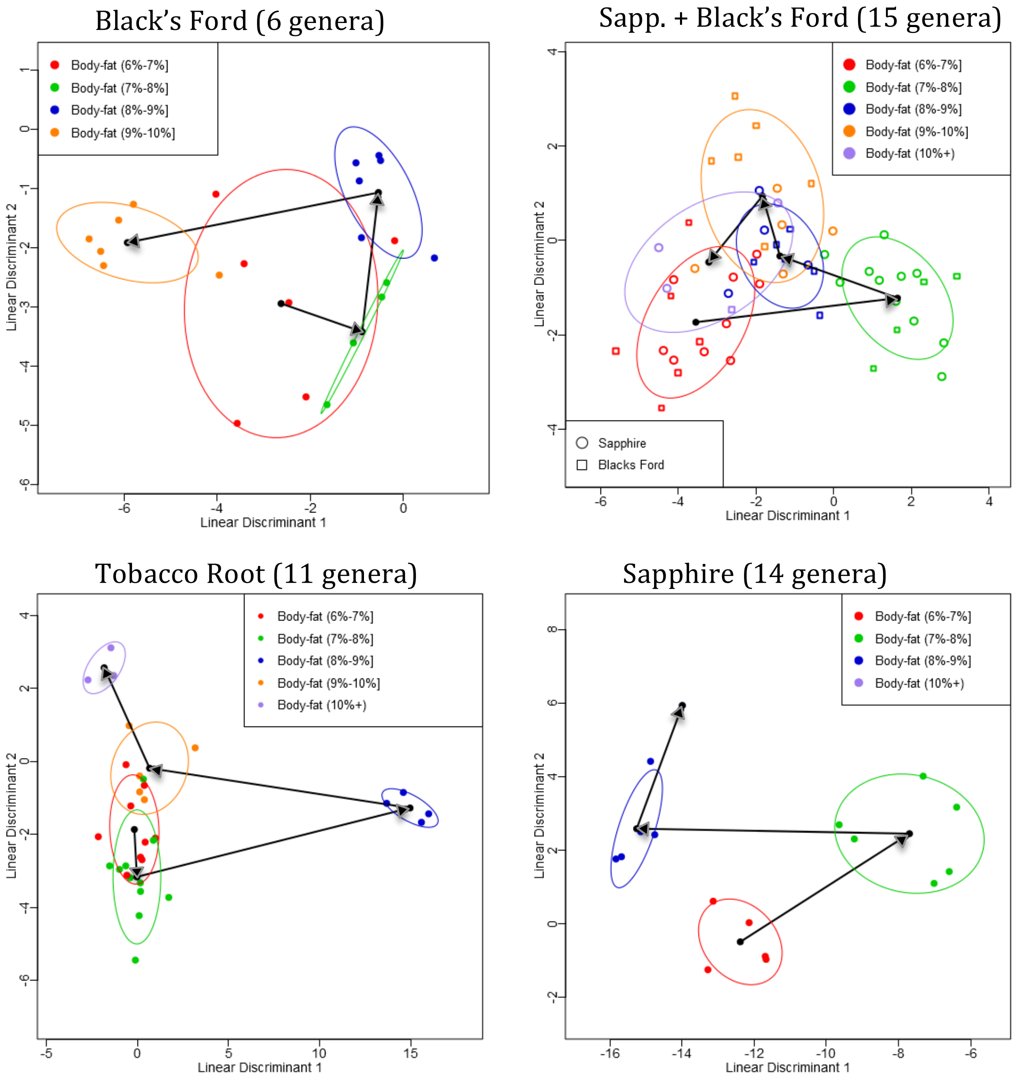

Figure 3. LDA ordination plots of female elk microbiomes clustered by body-fat from 3 separate populations in Montana, Black's Ford (upper left); Tobacco Root (lower left); Sapphire (lower right); and a plot of Sapphire and Black's Ford populations combined (upper right). Colored circles represent different body-fat measurements as indicated; colored ellipses depict 1 standard deviation in the data; and black circles are the centroid of each cluster. Black arrowed lines indicate increasing progression of body fat content. Produced using view cross-validation (CV) with leave-one-out method resulting in 59.09% accuracy for Black's Ford, 23.53% accuracy for Tobacco Root, 43.75% accuracy for Sapphire and 56.36% model accuracy for combined Sapphire and Black's Ford. Figure produced in R using the FSSF package (Spaulding et al. in development). Reduced OTU data performed best at 6, 11, 14 and 15 dimensions when clustering between elk body conditions for the 4 plots. Figure produced in R using the FSSF package (in development).

Using single elk populations to identify microbiome genera (bacteria) that correlate fecal microbiome composition with animal body-fat percentage yielded LDA results that seemed to be driven by the amount of data available to make such correlations. Data sets with strong representation in all body-fat groups clustered well using this feature selection followed by LDA approach. By contrast, data sets impoverished by low numbers of individuals performed poorly in CV and LDA. When the two populations that performed best (Black's Ford, CV 59.09% and Sapphire, CV 43.75%) were combined (Figure 3, upper right plot), the accuracy seemed to approach an average (CV 56.36%), with Sapphire increasing from 43.75% individually to the combined score of 56.36% and Black's Ford individual score decreasing in accuracy from 59.09%. This could indicate a balancing effect from strong population specific drivers within Black's Ford interacting with weaker Sapphire biomarkers. Or more simply, this could indicate a need for more evenly represented body-fat groups in our data sparse populations when building the algorithm. Further analysis and larger sample sizes will be needed to test this possibility. Although these populations performed variably we are excited to test this method with more data as its potential to distinguish population specific markers is still worthy of development. The population specific biomarker results suggest that larger future cohort data with more evenly represented data categories will be more useful in building the predictive algorithms, an important step before our approach can be expanded to classify individuals in populations with unbalanced data categories.

Hypothesis 3: Microbiome Biogeography

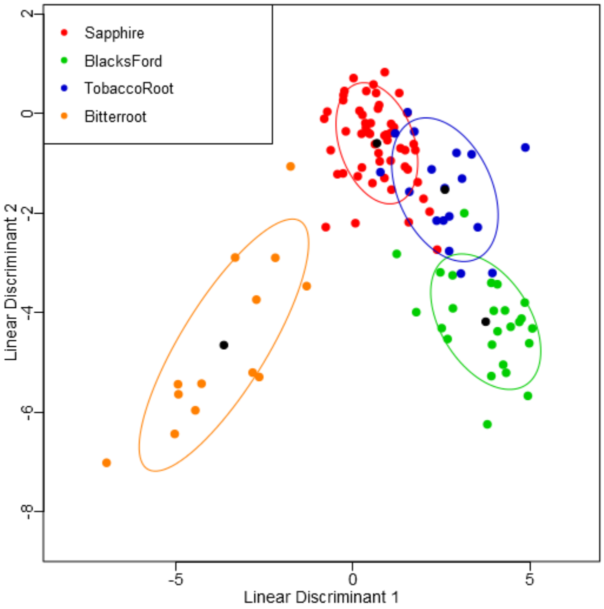

Figure 4. LDA ordination plot of female elk microbiome samples from 4 populations in Montana as a function of geographic location. Colored circles represent different populations as indicated; colored ellipses depict 1 standard deviation of the data in each cluster; and black circles are the centroid of each cluster. This visualization was produced using cross-validation (CV) with leave-one-out method producing 82.73% model accuracy. 20 features were selected. Figure produced in R using the SFFS package (Spaulding et al. in development).

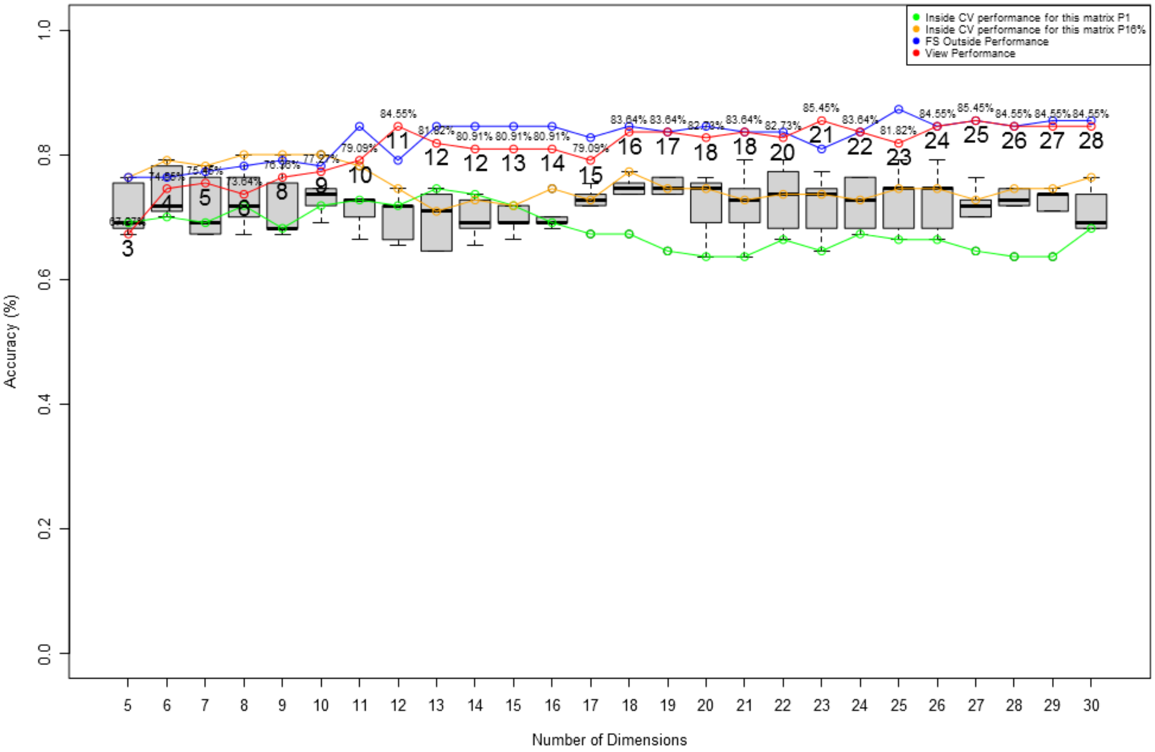

Figure 5. Box plots of cross validation (CV) accuracies (y-axis) with standard error whiskers from 4 types of CV approaches, selecting between 5 and 30 dimensions (x-axis) using female elk microbiome data combined from 4 populations in Montana to correlate with geographical origin. The green line indicates inside CV performance approach; the yellow line is inside CV performance for a reduced OTU matrix; the blue line is feature selection (SFFS) conducted outside of CV; and the red line is the performance for the LDA plots. This multiple plot helps determine the optimal number of features to balance accuracy and reduce over-fitting of the algorithm (the 20 feature LDA model was selected from the report above). Figure produced in R using the SFFS package (Spaulding et al. in development).

The bacterial genera selected in the biogeographical analysis show strong support for a population-specific microbiome based on geographical location, as suggested by the location-based clustering analysis in the LDA and high CV accuracy (Figure 4). This result is congruent with patterns of biogeographically mediated microbiomes seen in a prior study of the wild European house mouse and has implications for future biomarker use [27]. The strength of the relationship between biogeography and the microbiome is also supported by the box and whisker plots (Figure 5), since accuracy values remain high across all dimensions suggesting that all or most predictive genera selected vary according to geographic location.

How the strong effect of physical location interacts with more variable and transient dynamics like health needs further testing. We believe the microbiome is complex enough to contain microbes that respond in complex ways to correlates both strong and weak allowing us to separate through feature selection those genera of importance.

CONCLUSION/ SIGNIFICANCE

Elk intestinal microbiome composition analysis has been shown to represent microbial biomarkers for health and biogeography. If this general approach can be further developed and extrapolated across populations and landscapes with high precision, it will provide a powerful tool to non-invasively monitor disturbance on the landscape (both observable and cryptic) via observable effects on the health of indicator animal populations. This would allow managers to use this technique as an early warning system for demographic responses to environmental pressures in elk. With this approach, we will begin to fill important gaps in our knowledge of elk ecology by providing difficult to measure impacts of environment and disease acting on individuals along with general insight into microbial populations within ungulates. In the future, integrating the use of population genetics and microbial biomarkers from the same sample source can produce a holistic management solution for current and long-term trends while maintaining a low sampling effort and minimal animal handling.

The overarching goal of this research is to establish this approach for identifying microbial biomarkers within the fecal microbiome and the bioinformatics techniques used for their analysis and more broadly apply it to the management and conservation of other wildlife species (including non-mammals) which will allow federally designated threatened and endangered species to be studied with no perturbation. Microbial biomarkers represent a cheaper, less invasive alternative for acquiring information on wildlife populations. This research provides the foundation for expanded microbiome biomarker research and development across a diverse range of wildlife species for deep monitoring and conservation, potentially providing insights and novel solutions to current wildlife management issues.

AKNOWLEDGEMENTS

This research was supported by the Montana Institute on Ecosystems award from the NSF EPSCoR Track-1 program under Grant # EPS-1101342, also the University of Montana, College of Forestry and Conservation Irene Evers Competitive Undergraduate Research Scholarship, and the Davidson Honors College Watkins Scholarship. The elk fecal samples were provided for this research through a valued collaboration with MFWP, particularly Dr. Kelly Proffitt. The Metagenomics Working Group at UM provided collaboration and valuable feedback during the many stages of SFFS, LDA and CV algorithm development. Eric M. Spaulding conducted the majority of bioinformatics programming under the guidance of Professor Douglas W. Raiford. I appreciate the contributions and commitment of Professor William Holben and the Holben Lab members to this project. All of the above parties have been integral to the success of this project and I look forward to continued collaboration in future projects.

REFERENCES

1. Hooper, LV, Littman DR, and Macpherson AJ: **Interactions between the microbiota and the immune system.** *Science* 2012, **336**(6086):1268-1273.

2. Cho, I, and Blaser MJ: **The human microbiome: at the interface of health and disease**. *Nature Reviews Genetics* 2012, **13**(4):260-270.

3. Segata, N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, and Huttenhower C: **Metagenomic biomarker discovery and explanation.** *Genome Biol* 2011, **12**(6):R60.

4. Heffelfinger, JR, Geist V and Wishart D: **The role of hunting in North American wildlife conservation**. *International Journal of Environmental Studies* 2013, **70**:399-413.

5. Cook, RC, Cook JG, Murray DL, Zager P, Johnson BK, and Gratson MW. **Development of predictive models of nutritional condition for Rocky Mountain elk**. *The Journal of Wildlife Management* 2001, **65**:973-987.

6. Schaub, M, and Abadi F: **Integrated population models: a novel analysis framework for deeper insights into population dynamics**. *Journal of Ornithology* 2011, **152**:227-237.

7. Gruninger, RJ, Sensen CW, McAllister TA, and Forster RJ: **Diversity of rumen bacteria in Canadian cervids.** *PloS one* 2014, **9**(2):e89682.

8. Holben, WE, Feris KP, Kettunen A, and Apajalahti JHA: **GC fractionation enhances microbial community diversity assessment and detection of minority populations of bacteria by denaturing gradient gel electrophoresis.** *Applied Environmental Microbiology* 2004, **70**:2263-2270.

9. Aronesty, E: **Comparison of sequencing utility programs**. *TOBioiJ* 2013*,* DOI:10.2174/1875036201307010001.

10. Caporaso, JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, *et al.*: **QIIME allows analysis of high-throughput community sequencing data**. *Nature methods* 2010, **7**(5):335-336.

11. Edgar, RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics* 2010*,* **26**(19):2460-2461.

12. DeSantis, TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al.: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.** *Appl Environ Microb* 2006*,* **72**(7): 5069-5072.

13. Edgar, RC, Haas BJ, Clemente JC, Quince C, and Knight R: **UCHIME improves sensitivity and speed of chimera detection.** *Bioinformatics* 2011*,* btr381.

14. McDonald, D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Caporaso JG et al.: **The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome.** *GigaScience* 2012, **1**(1), 7.

15. Wang, Q, Garrity GM, Tiedje JM, and Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.** *Appl Environ Microb* 2007*,* **73**(16): 5261-5267.

16. Cole, JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, et al.: **The Ribosomal Database Project: improved alignments and new tools for rRNA analysis**. *Nucleic acids research* 2009, **37**(suppl 1):D141-D145.

17. Caporaso, JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, and Knight R: **PyNAST: a flexible tool for aligning sequences to a template alignment.** *Bioinformatics* 2010*,* **26**:266-267.

18. Price, MN, Dehal PS, and Arkin AP: **FastTree 2 -- approximately maximum-likelihood trees for large alignments**. *PLoS ONE* 2010, **5**(3):e9490.

19. McDonald, D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, and Hugenholtz P: **An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.** *ISME J* 2012*,* **6**(3): 610–618.

20. Chao, A, Gotelli NJ, Hsieh TC, Sander EL, Ma KH, Colwell RK, and Ellison AM: **Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies.** *Ecological Monographs* 2013*, online early.*

21. Vazquez-Baeza, Y, Pirrung M, Gonzalez A, and Knight R: **Emperor: A tool for visualizing high-throughput microbial community data.** *Gigascience* 2013*,* **2**(1):16.

22. Pudil P, Novovičová J, and Kittler J: **Floating search methods in feature selection.** *Pattern recognition letters* 1994, **15**(11), 1119-1125.

23. Saeys Y, Inza I, and Larrañaga P: **A review of feature selection techniques in bioinformatics.** *bioinformatics* 2007, **23**(19):2507-2517.

24. Spaulding, E, et al.: **Feature Selection, Floating Search**. manuscript in progress. 2014.

25. Liu Z, Chen D, Sheng L, and Liu AY: **Class prediction and feature selection with linear optimization for metagenomic count data.** *PloS one* 2013, **8**(3):e53253.

26. Braga-Neto, UM, and Dougherty ER: **Is cross-validation valid for small-sample microarray classification?**. *Bioinformatics* 2004, ***20***(3), 374-380.

27. Linnenbrink, MJ, Wang E, Hardouin A, Künzel S, Metzler D, and Baines JF: **The role of biogeography in shaping diversity of the intestinal microbiota in house mice.** *Molecular Ecology* 2013, **22**, 1904-1916.