Data Analysis Methods for Microbial Source Tracking of E.coli in the Lamoille River Watershed, Vermont

David Minkoff; Faculty Sponsor: Robert B. Genter, Ph.D.

Department of Environmental and Health Science, Johnson State College

Introduction

Escherichia coli (*E. Coli*) are bacteria found in the digestive tracts of warm blooded animals such as birds and mammals. In surface waters, *E. Coli* are associated with runoff from animal and human fecal matter (EPA, 2005). Although most strains of *E. Coli* are non-pathogenic, a few strains can cause gastrointestinal illnesses, while others are considered opportunistic pathogens, infecting hosts with compromised immune systems (Feng, et al, 2002). It is therefore useful both to monitor overall levels of *E. coli* in surface waters, as well as to gain an understanding of the species sources of these microbes.

E. Coli samples isolated from stream water in tributaries of the Lamoille River were subjected to genotypic comparison analysis through ribotyping. Ribotyping has been used to successfully identify sources of *E. coli* in recreational waters (Carson, et al, 2001) by comparing isolates of unknown origin to a library of isolates whose species origin is documented. A cluster analysis of the results was conducted using a similarity dendrogram. Library and isolate samples that clustered above a given threshold were considered a match. The purpose of this study was to determine the similarity threshold level that would provide the most useful information for determining microbial sources.

Materials and methods

During 10 weeks each summer from 2008-present, colonies of *E. coli* were extracted from water samples taken from nineteen sites on tributaries of the Lamoille River. The isolates were cultured on McConkey II selective media (BD Medical, Franklin, NJ), and Ribotyped on an automated Riboprinter (DuPont-Qualicon, Wilmington, Delaware). The "genetic barcodes" generated by the riboprinter were analyzed using GelComparII software (Applied Maths, Austin, Texas), by creating similarity dendrograms using the Dice association coefficient (Dice, 1945), with software optimization and position tolerance settings of 1.5. Samples that clustered above a given threshold were considered a match, and results using threshold levels of 84%, 86%, 88%, 90%, 92%, 94% and 96% were compared. Considerations for determining the optimal threshold level included the number of single-species matches, the number of multi-species matches and the number of unmatched isolates at each threshold tested. A greater relative number of single-species matches and smaller relative number of multi-species matches was considered preferred.

Results

A total of 528 stream isolate samples and 157 library samples were compared.

The dendrogram cluster analysis revealed that the number of single-species matches increased more than 300% at higher threshold levels, from a low of 28 matches at 84% similarity to a high of 95 matches at 94%, dropping again to 66 matches at 96% (Figure 1). The number of multi-species matches ranged from a high of 397 at 84% to a low of 186 at 96%. There were 205 multi-species matches at 94%, nearly as few as those at 96%. The number of species included in a multi-species cluster ranged from a maximum of ten in a single cluster at 84% (*Cow/ Human/ Fisher/ Calf/ Bear/ Dog/ Goat/ Duck/ ATCC (standard)/ Horse*) to a maximum of four per cluster at 94-96% (both contained the following clusters: *Cow/ Deer/ Fisher/ Pig* and *Calf/ Cow/ Fisher/ Human*). The number of unmatched (and therefore unknown) isolates ranged from a low of 103 at 84% to a high of 276 at 96%. The number of unmatched isolates at 94% was 228, 48 fewer unmatched isolates than at 96% (figure 1).

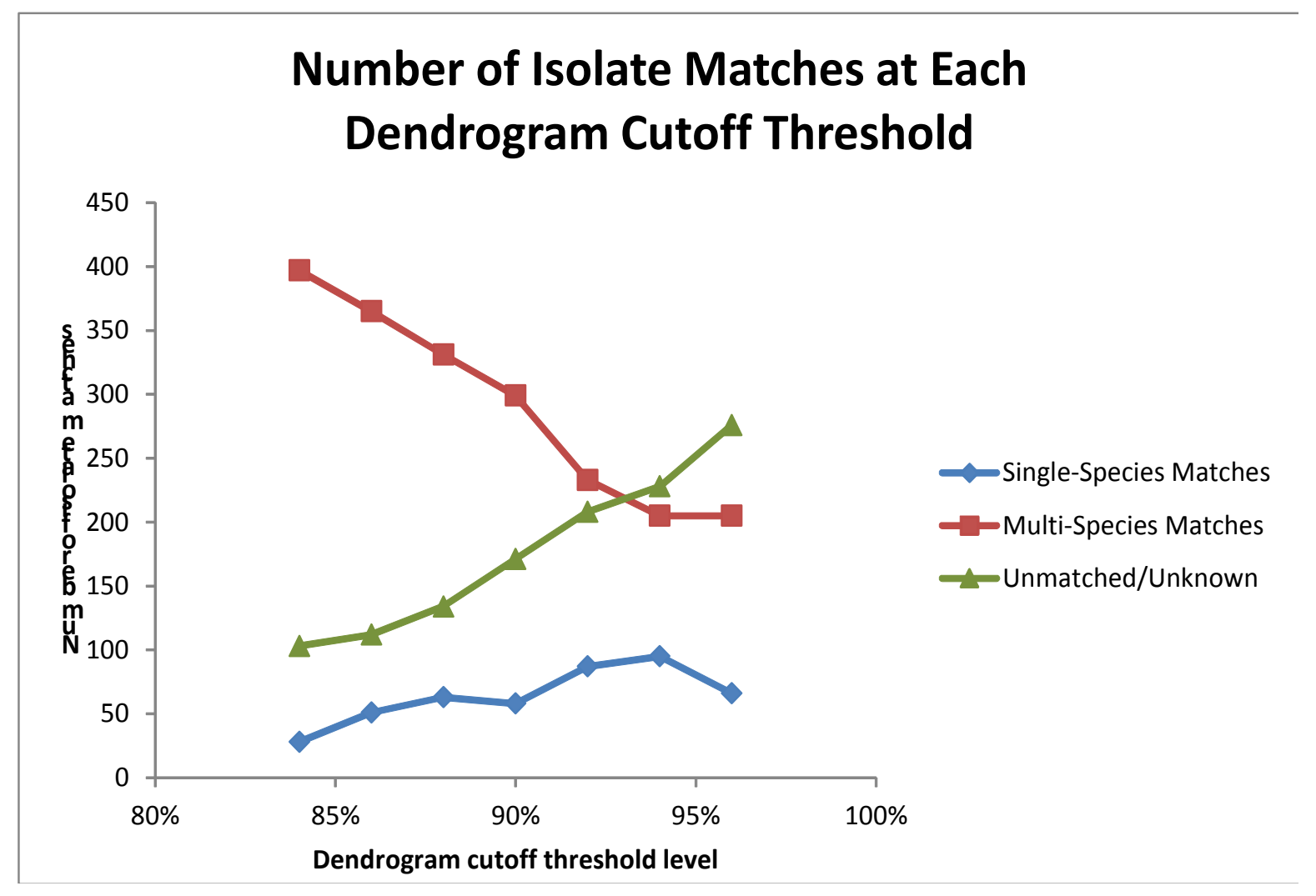
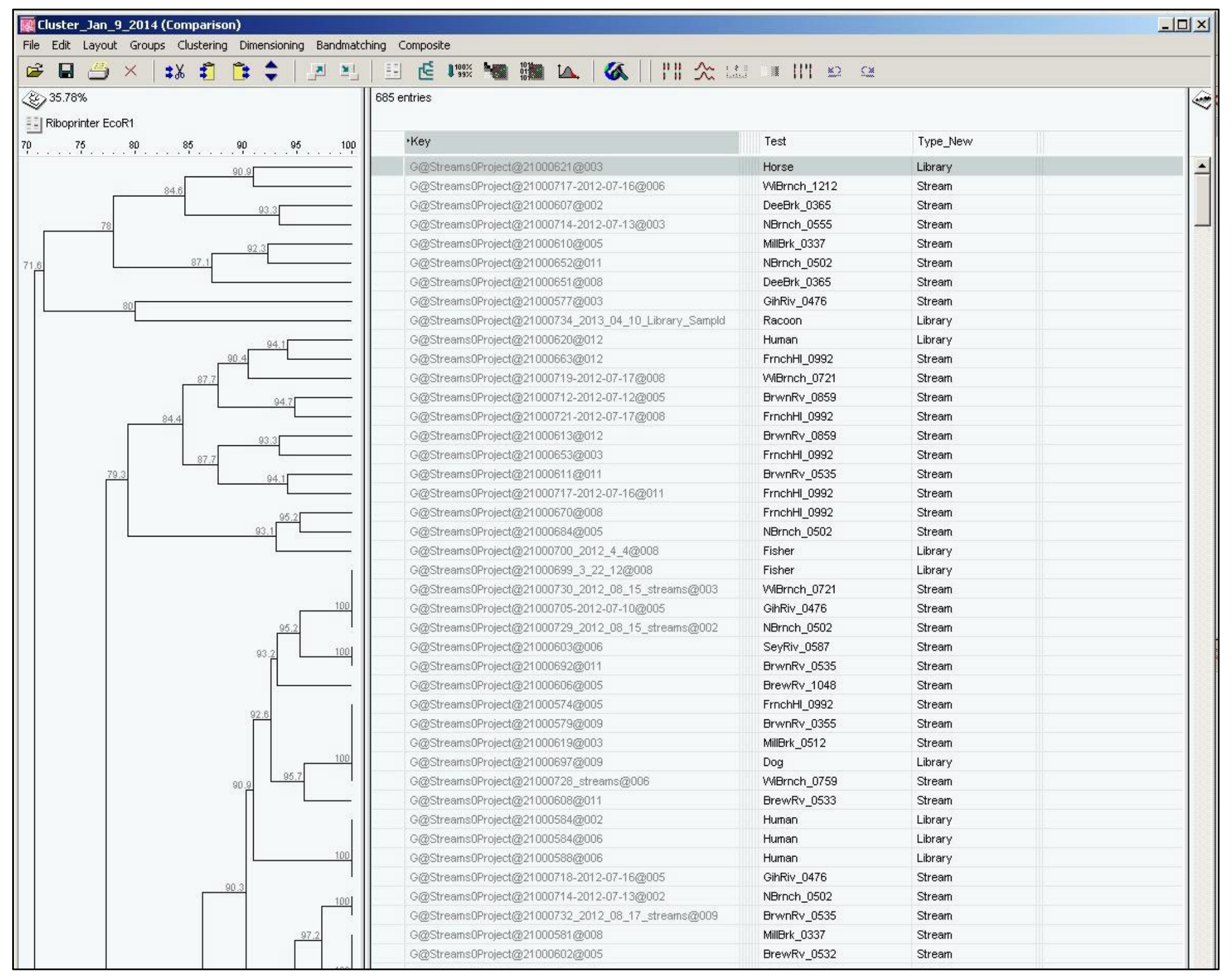


Figure 1: Number of isolate matches for each analyzed dendrogram cutoff threshold level.



685 entries

Key	Test	Type_New
G@Streams0Project@21000621@003	Horse	Library
G@Streams0Project@21000717-2012-07-16@006	WiBrnch_1212	Stream
G@Streams0Project@21000607@002	DeeBrk_0365	Stream
G@Streams0Project@21000714-2012-07-13@003	NBrnch_0555	Stream
G@Streams0Project@21000610@005	MillBrk_0337	Stream
G@Streams0Project@21000652@011	NBrnch_0502	Stream
G@Streams0Project@21000651@008	DeeBrk_0365	Stream
G@Streams0Project@21000577@003	GihRiv_0476	Stream
G@Streams0Project@21000734_2013_04_10_Library_Sampld	Racoon	Library
G@Streams0Project@21000620@012	Human	Library
G@Streams0Project@21000663@012	FrnchHl_0992	Stream
G@Streams0Project@21000719-2012-07-17@008	WiBrnch_0721	Stream
G@Streams0Project@21000712-2012-07-12@005	BrwnRv_0859	Stream
G@Streams0Project@21000721-2012-07-17@008	FrnchHl_0992	Stream
G@Streams0Project@21000613@012	BrwnRv_0859	Stream
G@Streams0Project@21000653@003	FrnchHl_0992	Stream
G@Streams0Project@21000611@011	BrwnRv_0535	Stream
G@Streams0Project@21000717-2012-07-16@011	FrnchHl_0992	Stream
G@Streams0Project@21000670@008	FrnchHl_0992	Stream
G@Streams0Project@21000684@005	NBrnch_0502	Stream
G@Streams0Project@21000700_2012_4_4@008	Fisher	Library
G@Streams0Project@21000699_3_22_12@008	Fisher	Library
G@Streams0Project@21000730_2012_08_15_streams@003	WiBrnch_0721	Stream
G@Streams0Project@21000705-2012-07-10@005	GihRiv_0476	Stream
G@Streams0Project@21000729_2012_08_15_streams@002	NBrnch_0502	Stream
G@Streams0Project@21000603@006	SeyRiv_0587	Stream
G@Streams0Project@21000692@011	BrwnRv_0535	Stream
G@Streams0Project@21000606@005	BrewRv_1048	Stream
G@Streams0Project@21000574@005	FrnchHl_0992	Stream
G@Streams0Project@21000579@009	BrwnRv_0355	Stream
G@Streams0Project@21000619@003	MillBrk_0512	Stream
G@Streams0Project@21000697@009	Dog	Library
G@Streams0Project@21000728_streams@006	WiBrnch_0759	Stream
G@Streams0Project@21000608@011	BrewRv_0533	Stream
G@Streams0Project@21000584@002	Human	Library
G@Streams0Project@21000584@006	Human	Library
G@Streams0Project@21000588@006	Human	Library
G@Streams0Project@21000718-2012-07-16@005	GihRiv_0476	Stream
G@Streams0Project@21000714-2012-07-13@002	NBrnch_0502	Stream
G@Streams0Project@21000732_2012_08_17_streams@009	BrwnRv_0535	Stream
G@Streams0Project@21000581@008	MillBrk_0337	Stream
G@Streams0Project@21000602@005	BrewRv_0532	Stream

Figure 2: Portion of a dendrogram generated from cluster analysis in GelComparII software.

An examination of the single-species matched isolates shows that the average number of matches for each species represented ranged from a low of 1.75 species represented at 84% to a high of 5.94 at 94%, dropping to 4.13 at 96% (figure 3). The median number of matches per species represented more than tripled from 1 match per species at 84% to 3.5 matches per species at 94%, dropping again to 1 match per species at 96%. The number of species represented in single-species matches rose from 9 species at 84% to 15 species represented at 94%, dropping to 12 species represented at 96% (figure 3).

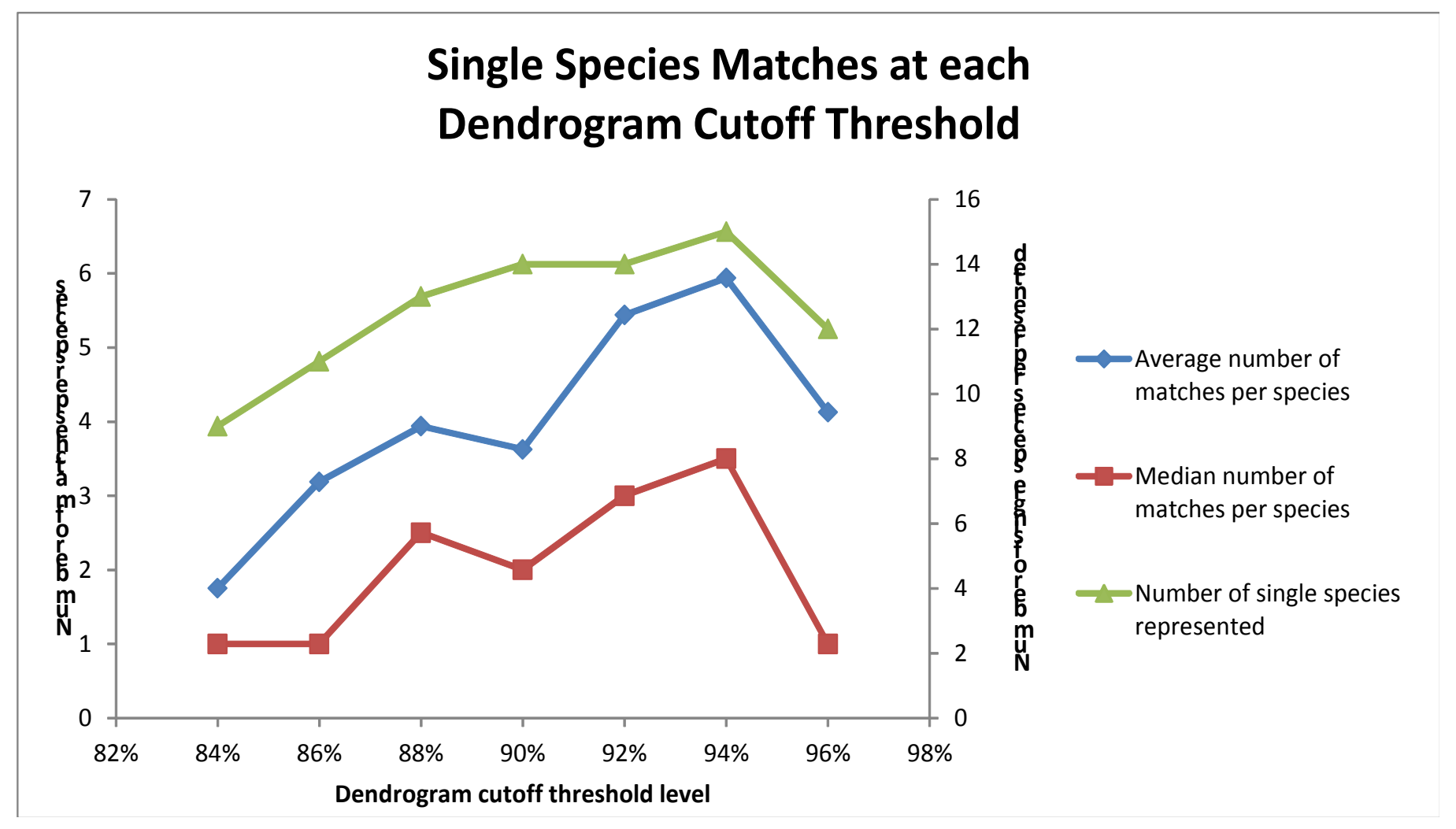


Figure 3: Average and median number of matches per species represented amongst single-species matched isolates, as well as the number of species represented by single-species matches.

Figure 4: Comparing printouts of full dendrograms at different threshold levels.

Conclusions

The overall purpose of this ongoing study is to track the source of each *E. coli* isolate extracted from the water samples. Despite the fact that the number of unmatched isolates is larger with successively higher threshold levels, the observed increase in single-species matches at higher threshold levels up to 94% is more important in terms of yielding the most useful data for the study. The available evidence would therefore strongly suggest that a dendrogram cutoff threshold of 94% is optimal for this data set. The decrease in multi-species matches at higher threshold levels further supports this conclusion. A large number of multiple-species matches were nonetheless found at *every* threshold level, possibly as a result of an insufficient number of library samples. It is also possible that certain strains of *E. coli* are found in more than one species of warm-blooded animal living in the Lamoille River watershed, however the presence of a match such as *Cow/ Human/ Fisher/ Calf/ Bear/ Dog/ Goat/ Duck/ Horse* (at 84%) or even *Cow/ Deer/ Fisher/ Pig* (at 94-96%) was unexpected. The number of library samples will continue to increase every summer, and it will be interesting to monitor the changes in matching clusters as the study progresses. A sensitivity analysis of the existing data could provide more insight into the reliability of our library sample size.

Literature cited

Carson CA, Shear BL, Ellersieck MR, Asfaw A. 2001. Identification of fecal *Escherichia coli* from humans and animals by ribotyping. Applied Environmental Microbiology 67:1503–1507

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297-302

Environmental Protection Agency (EPA). 2005. *Microbial Source Tracking Guide Document* (EPA600-R-05-064). U.S.E.P.A. Office of Research and Development, Cincinnati, OH

Feng, P., Weagant, S. D., & Grant, M. A. 2002. Enumeration of Escherichia coli and the coliform bacteria. *Bacteriological Analytical Manual*, *8*, 102-135

Acknowledgments

Thanks to Robert Genter, Ph.D., Saul Blocher, Keith Kirchner, John Pellerin, Elizabeth Dolci, Ph.D., the VT EPScOR Streams Team, Johnson State College, Kristina Marcussen, Sylvie Minkoff and Lucy Minkoff

Funding provided by NSF Grant EPS-1101317