ISVEE 16 Post-Conference Workshop Proposal

Title of the workshop: Introduction to Bioinformatics and Statistics for Metagenomic Research

NOTE: We are offering this workshop in hybrid format. In-person participation is encouraged, but we have done this workshop virtually with good success.

Motivation and description of the objectives, or learning outcomes, of the workshop: Our international multidisciplinary group of scientists and educators is addressing the issues of antimicrobial resistance (AMR) and microbial ecology as they relate to animal health, public health, and ecosystem health through research, education, and outreach. By characterizing risks related to AMR and microbial ecology, our group will identify agricultural production practices that are harmful and can be avoided, while also identifying and promoting production practices and interventions that are beneficial or do no harm to the ecosystem or public health. This will allow society to realize "sustainable intensification" of agriculture.

The Microbial Ecology Group (MEG) workshop lessons' are designed to introduce researchers to several important areas of metagenomic research for assessment of microbiome and resistome composition. Starting with foundational information of metagenomic sequencing, datasets, and research problems, participants will develop hands-on bioinformatics skills and execute post-sequencing tools and software on a high-performance linux computing system. The R programming language will be utilized for statistical analysis of metagenomic sequencing data. Participants will gain valuable experience through discussing and interpreting metagenomic concepts, challenges, and analyses.

Learning Objectives: Upon completion of the workshop, ISVEE 16 participants will be able to:
- Employ basic and complex commands in a linux environment.
  - Navigation
  - Downloading, transferring, and storing files from local directories and remote repositories
  - Modifying command-line parameters and flags
  - Running Nextflow pipeline scripts in the background via a terminal multiplexer
- Understand and interpret different file formats associated with bioinformatic analysis.
  - Fasta and fastq files
  - SAM files
  - Count matrices
  - Qiime2 files
- Describe the unique characteristics of metagenomic data, and how to handle these characteristics during a microbiome-resistome analysis
- Describe each step in the bioinformatic pipeline used for microbiome-resistome analysis (i.e., AMR++)
- Understand the bioinformatic algorithms being used in the analysis, and how these impact interpretation of results
- Use the R package phyloseq to import, store, analyze, and graphically display complex phylogenetic sequencing data. More specifically: Appropriately calculate, interpret, and discuss common statistical tests and summary statistics.
  - Phyloseq object summary statistics
  - Sample metadata summary statistics
  - Alpha diversity, including different alpha diversity measures
  - Wilcoxon test and Generalized linear models
  - Differential abundance testing using the Zero-inflated Gaussian model
  - Ordination and cluster analysis
  - Measures of association between metadata (including batch effects and primary variables) and microbiome-resistome outcomes (including diversity and differential abundance)
  - Create exploratory figures, including:
    - Bar charts of mapped reads
    - Boxplots of diversity values
    - Ordination plots to visualize beta-diversity

Background and skills the workshop attendees are expected to (or must) have: Participants will be expected to have basic knowledge of genomics and epidemiological study design. Previous experience with R and command line coding will be helpful but not strictly necessary.

The following tasks will be assigned prior to the workshop:
- View three pre-recorded videos (Introduction to metagenomic sequencing, Introduction to bioinformatics, Introduction to statistics for metagenomic sequencing data)
- Create a free datacamp account and complete the first chapters of these tutorials:
    - Introduction to Shell: Chapter 1 - Manipulating files and directories
    - Introduction to R: Chapter 1 - Intro to basics
- Install software:
    - R or Rstudio
    - FileZilla
    - For Windows users: Install terminal emulator/SSH client such as MobaXterm or PuTTY
- Join the MEG research discussion group on Slack: MEG Slack group
- Review other shared documents and GitHub repositories
    - Shared drive and Dropbox link(s)
    - AMR++ (https://github.com/EnriqueDoster/amrplusplus_v2)
    - MEG course: Introduction to statistical analysis of metagenomic sequencing data

Information about any previous history of the proposed workshop: *The MEG has led five highly successful previous versions of workshops and courses addressing the material that will be presented in this proposed workshop, and with each iteration the course continues to improve. The previous workshops included a large variety of scientists (including researchers, faculty, post-docs and students) across government, industry and academia, from the U.S., Canada, Norway, and several other international institutions.*

> **Title**: Introduction to Bioinformatics and Statistics for Metagenomic Research
> **Venue**: Virtual Short-course
> **Dates**: January 5-7, 2021
> **Attendance**: 18 participants
>
> **Title**: Introduction to statistical analysis of metagenomic sequencing data
> **Venue**: Virtual course
> **Dates**: May 11th, 2020 to August 12, 2020 (12-weeks)
> **Attendance**: ~40 participants
>
> **Title**: Metagenomics Bioinformatics Workshop
> **Venue**: University of Oslo and Norwegian University of Life Sciences
> **Dates**: September 2019, 2-day workshop
> **Attendance**: ~25 participants
>
> **Title**: Metagenomics Bioinformatics Workshop
> **Venue**: University of Minnesota
> **Dates**: May 2019, 2-day workshop
> **Attendance**: ~45 participants
>
> **Title**: Bioinformatics & Statistics for Shotgun Metagenomic Research
> **Venue**: Conference of Research Workers in Animal Diseases (CRWAD), Chicago Marriott, Downtown Magnificent Mile, Chicago, Illinois
> **Dates**: December 2-3, 2017
> **Attendance**: 51 participants

Workshop specifications:
Duration of the proposed workshop: 3 days

Number of participants in the workshop: **Minimum**: 20 **Maximum**:50

Proposed net registration fees:

Cost category	Number of days	Fee amount	Per-participant Cost
Room fees	3 days	$225 per day	$60
Daily catering fees	3 days	$50/day/participant	$150
AV support fees	3 days	$100/day	$15
General Workshop fees	One-time per participant	$40	$40
Instructor Fees	4 instructors, 3 days each	$40/instructor/day	$480
Instructional Aid Fees	3 days	$5/participant/day	$15
Proposed net registration fees			$760 CAD

Do you prefer a pre- or post-conference workshop: *Post-conference workshop.*

Will you allow non-ISVEE delegates to participate in the workshop: *Yes, we will allow non-ISVEE delegates to participate in the workshop if the registration by ISVEE delegates does not reach the maximal enrollment.*

Working language of the workshop: *English*

Will online participation in the workshop be possible (optional, see "Notes for workshop organizers")? If yes, describe how online participants will be involved in the workshop: *No plans at this time, but if there is strong demand we could offer this option as we have successfully delivered this content in previous virtual short-courses.*

Any specific requirements for the organization of the workshop, for example related to configuration of rooms (see "Notes for workshop organizers" below for issues to consider): *Because of the hands-on and interactive nature of the course, we prefer a setting that allows instructors to easily move among the participants, and for participants to gather and work together in smaller groups. Students need to work on their own computers, so power supply is important for them. They also need wireless internet connections to allow them to access the servers where bioinformatic analysis will be conducted.*

Information about workshop organizers:

Noelle Noyes: Dr. Noyes is an Assistant Professor in the Department of Veterinary Population Medicine Department at the University of Minnesota. Currently, her research focuses on improving the understanding of antibiotic resistance in livestock production, with the ultimate goal of optimizing both public health and food safety and security. Noelle was a USDA NIFA Post-Doctoral Fellow and an NIH T32 Pre-Doctoral Fellow. She was a recipient of the German Chancellor Fellowship from the Alexander von Humboldt Foundation, and received her MA from Osnabrueck University and her BA from Amherst College. Noelle completed a dual-degree PhD-DVM program at Colorado State University before joining the University of Minnesota faculty. Currently, her lab is conducting studies on microbiome, pathogen and antibiotic resistance issues related to livestock production and food safety, with funding from USDA, NIH, NSF, NPB, NCBA and the University of Minnesota.

Paul Morley: Dr. Morley is an epidemiologist and veterinary internal medicine specialist that studies infectious diseases affecting people and animals. Major emphases for his professional activities include investigating the ecology of antimicrobial resistance in animals and food production systems, and using analytical epidemiology to improve our understanding of health and disease. For the past 10 years, his research has emphasized the use of cutting-edge genomic methods to investigate the effects of agriculture production practices on antimicrobial resistance and microbial ecology as these affect human, animal, and ecosystem health. In 2019, Dr. Morley joined the VERO program (Veterinary Education, Research, and Outreach) as the Director of Research. The VERO program is based in the heart of one of the most productive animal agricultural regions in the world, providing unprecedented opportunities for collaboration with industry partners and stakeholders to fulfill the VERO mission. Dr. Morley also

helps to lead the Microbial Ecology Group, a multidisciplinary, multi-institutional research group that is investigating the ecology of antimicrobial resistance and foodborne pathogens, in efforts to address the Grand Challenge of providing nutritional security for the global population. Additionally, he is a recognized authority on infection control in animal populations and has consulted on infection control and biosecurity issues at veterinary hospitals, veterinary colleges, and intensive animal production facilities around the world.   Dr. Morley has authored more than 210 peer-reviewed scientific publications in addition to numerous book chapters and government reports.

Lee Pinnell: As a microbial ecologist, Dr. Pinnell has over 10 years' experience performing molecular and computational research characterizing microbial communities. He received his PhD in marine microbial ecology from Texas A&M University - Corpus Christi and served as a postdoctoral scientist at the Shedd Aquarium in Chicago, Illinois prior to joining the Morley Lab at Texas A&M University. His current work is centered on the computational analysis of metagenomic data from a variety of animal health and agriculture-based research projects. He also has considerable experience designing and performing metagenomic studies in a wide variety of other ecosystems.

Peter Ferm, is a bioinformatic researcher a part of the Noyes Lab in the Veterinary Population Medicine Department at the University of Minnesota. He received his Master's in Bioinformatics and Computational Biology at the University of Minnesota with Dr. Noelle Noyes as a graduate mentor. Peter's current research focuses on comparing different whole-genome sequencing, pipelines, approaches, and datasets. In delving into the many decision areas in a whole-genome sequencing analysis, Peter strives to understand how we can improve transparency and scalability of post-sequencing pipelines for large genomic datasets. When Peter is not on the command-line or in different coding environments, he enjoys the Montana outdoors with his family and friends.

Contact information of the primary applicant(s):
Dr. Noelle Noyes
Food-Centric Corridor, Infectious Disease Laboratory, Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota
nnoyes@umn.edu | cell: 617-953-7837

Dr. Paul Morley
Veterinary Education, Research, and Outreach Program, College of Veterinary Medicine & Biomedical Sciences, Texas A&M University
pmorley@cvm.tamu.edu | cell: 970-219-6089

Dr. Lee Pinnell
Veterinary Education, Research, and Outreach Program, College of Veterinary Medicine & Biomedical Sciences, Texas A&M University
ljpinnell@cvm.tamu.edu | cell: 361-944-3288

Peter Ferm
Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota
fermx014@umn.edu | cell: 406-529-3478

Details about the schedule of the workshop

Day	Time	Activity/Content
1	08:00 - 08:30	Introductions
	08:30 - 08:45	Review of introductory materials
	08:45 - 09:00	Group discussion and questions
	09:00 - 09:15	Intro to research question & dataset
	09:15 - 10:00	Understanding sequencer output
	10:00 - 10:30	*Bio break*
	10:30 - 10:40	Sequence Data QC
	10:40 - 11:00	Intro to AMR and taxonomic DBs
	11:00 - 11:45	Sequence classification methods
	11:45 - 12:00	Group discussion and questions
	12:00 - 13:00	*Lunch*
	13:00 - 13:30	Connect to servers
	13:30 - 13:45	Basic navigation on servers
	13:45 - 14:00	Transferring files to the server
	14:00 - 14:30	Intro to running jobs on server
	14:30 - 15:00	*Bio break*
	15:00 - 16:00	Running bioinformatic pipeline on the server
	16:00 - 16:30	Understanding pipeline output
	16:30 - 17:00	Group discussion and questions
2	08:00 - 08:30	Day 1 review
	08:30 - 09:00	Check day 1 pipeline results
	09:00 - 10:00	Run pipeline with additional flags
	10:00 - 10:30	*Bio break*
	10:30 - 12:00	Group discussion and questions
	12:00 - 13:00	*Lunch*
	13:00 - 13:15	Review of bioinformatics; Intro to statistics
	13:15 - 13:45	Importing pipeline results into R
	13:45 - 14:30	Calculating summary statistics
	14:30 - 15:00	*Bio break*
	15:00 - 15:15	Statistical theory for metagenomic data
	15:15 - 16:30	Normalizing results and creating figures
	16:30 - 17:00	Group discussion and questions
3	08:00 - 08:30	Day 2 review
	08:30 - 10:00	Run statistical tests on previous results
	10:00 - 10:30	*Bio break*
	10:30 - 11:15	Non-metric multidimensional scaling
	11:15 - 12:00	Zero-inflated Gaussian model
	12:00 - 13:00	*Lunch*
	13:00 - 13:30	Intro to breakout group afternoon tasks
	13:30 - 14:00	Getting to know your data
	14:00 - 14:30	Group presentation of results/figures
	14:30 - 14:45	*Bio break*
	14:45 - 15:15	Characterize composition of microbiome/resistome
	15:15 - 15:45	Group presentation of results/figures
	15:45 - 16:15	Hypothesis testing
	16:15 - 16:30	Group presentation of results/figures
	16:30 - 17:00	Final group discussion/questions