Introduction to Bioinformatics and Statistics for Metagenomic Research

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 how metagenomic data are generated, including the most important pre-sequencing workflow considerations
- Describe the unique characteristics of metagenomic data, and how to handle these characteristics during a microbiome-resistome analysis
- Explain the strengths and weaknesses of existing databases for microbial genomes and genes
- Describe each step in the bioinformatic pipeline used for microbiome-resistome analysis (i.e., AMR++)
- Understand the bioinformatic algorithms being used for analysis, and how they impact interpretation of results
- Execute the microbiome and resistome bioinformatic pipeline AMR++, and appropriately interpret the structure and content of the output folders and files
- 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

Pre-workshop Tasks and Requirements (content will be provided prior to the workshop)

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.

- 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++ v2
    - MEG course: Introduction to statistical analysis of metagenomic sequencing data

History of MEG Workshops

The MEG has led five highly successful previous versions of workshops and courses addressing the material that will be presented in this 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.

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: Peter 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

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

Workshop Schedule

Day	**Activity/Content**
1	Review of introductory materials Understanding sequencer output Sequence Data QC Intro to AMR and taxonomic DBs Sequence classification methods Transferring files to the server Intro to running jobs on server Running bioinformatic pipeline on the server Understanding pipeline output
2	Check day 1 pipeline results Review of bioinformatics; Intro to statistics Importing pipeline results into R Calculating summary statistics Statistical theory for metagenomic data Normalizing results and creating figures
3	Non-metric multidimensional scaling Zero-inflated Gaussian model Characterize composition of microbiome/resistome

	Hypothesis testing Final group discussion/questions