

CABAlliance Bioinformatics Workshop  
Université des Sciences et Techniques de Masuku  
Gabon  
July 3<sup>rd</sup> – 8<sup>th</sup>, 2017

## Introduction

DNA sequencing technologies have advanced in such a way that individual researchers and research teams can generate large genomic and transcriptomic datasets quickly and relatively inexpensively. This increased accessibility of genome-scale data has made significant contributions to conservation by revealing cryptic species, identifying at-risk populations, quantifying genetic variation and identifying signatures of selection. However, the most common obstacle to performing genomics studies has become the bioinformatics analysis needed to make sense of these large datasets and diverse computational approaches needed to analyze the data. This workshop will provide background theory behind these topics before moving to a variety of hands-on exercises in conservation genomics and associated bioinformatics pipelines.

The workshop will feature hands-on, practical sessions using datasets generated as part of the CAB-Alliance research program. This research aims to take a genomic approach to mapping patterns of adaptive variation across the Central African landscape and predicting future patterns under projected models of climate change. We will be using datasets generated for two taxa: the forest mouse *Praomys misonnei* and the golden puddle frog *Phrynobatrachus auritus*, both of which have been sampled from sites across Gabon and Cameroon. During the workshop, we will follow through the key steps needed in the pipeline used to analyze data and look at ways of exploring environmentally-associated patterns of genetic variation using a variety of available software.

By the end of this workshop, participants will be expected to:

- Gain experience in the command-line (and understand its integral and integrative role for data processing)
- Gain understanding of the steps involved in the RAD-sequencing approach, knowledge of the type of data the approach generates and the practical applications of the approach.
- Gain experience in processing raw RAD-sequencing data, including quality control and filtering.
- Gain experience with using RAD-sequencing data to identify patterns of geographic population structure and assign individuals to populations.
- Gain experience with detecting signatures of selection within RAD-sequencing data.
- Gain experience of mapping patterns of variation across the Central African landscape.
- Understand importance of genomic data in conservation

## Topics

Overall context:

Overview of conservation genetics/genomics  
Next generation sequencing  
Role of bioinformatics in conservation

### Data types:

Short read DNA sequences  
Single Nucleotide Polymorphisms (SNPs)  
Transcriptomics

### Data analyses and pipelines:

RADseq pipelines, from raw reads to analyzing data  
Visualizing geographic population structure  
Identifying populations (Admixture, hybridization)  
Detecting adaptive variation

### Tentative schedule:

PREPARATION DAY		
Time		Instructors
TBA	Day to sort out logistics and set-up computers with necessary software.	Dr. Nicola Anthony Dr. Katy Morgan Courtney Miller

DAY ONE		
Time	Topic/Exercise	Instructors
9:00-1:00	<ol style="list-style-type: none"><li>1. Brief introduction of participants and welcome from USTM</li><li>2. Brief overview of conservation genomics (NA)</li><li>3. Metabarcoding and its applications (SN)</li><li>4. Introduction to bioinformatics (KM)</li><li>5. Navigating data resources - tutorials<ol style="list-style-type: none"><li>a. Data Dryad (KM)</li><li>b. Galaxy (CM)</li><li>c. NCBI + Blast (NA)</li></ol></li><li>6. "Virtual" computing in Atmosphere (or other super computer/cloud computing resources) and the Discovery Environment (CM/NA)</li></ol>	Dr. Patrick Mickala Dr. Stephan Ntie Dr. Nicola Anthony Dr. Katy Morgan Courtney Miller

<b>1:00-2:00</b>	LUNCH	
<b>2:00-5:00</b>	<ol style="list-style-type: none"> <li>1. Working in the command line (Cygwin) <ol style="list-style-type: none"> <li>a. Navigation</li> <li>b. Manipulating files and file content</li> </ol> </li> </ol>	All

<b>DAY TWO</b>		
<b>Time</b>	<b>Topic/Exercise</b>	<b>Instructors</b>
<b>9:00-1:00</b>	<ol style="list-style-type: none"> <li>1. Introduction to Next Generation Sequencing techniques and applications <ol style="list-style-type: none"> <li>a. File formats (VCF, fasta etc).</li> <li>b. Illumina sequencing (NA)</li> <li>c. RADseq and SNPs as molecular marker (KM)</li> <li>d. Case studies (All)</li> </ol> </li> </ol>	All
<b>1:00-2:00</b>	LUNCH	
<b>2:00-5:00</b>	<ol style="list-style-type: none"> <li>1. RAD-seq data processing (Part 1) <ol style="list-style-type: none"> <li>a. FastQC (KM)</li> <li>b. STACKS – process Raddtags (CM)</li> </ol> </li> </ol>	Dr. Katy Morgan Courtney Miller

<b>DAY THREE</b>		
<b>Time</b>	<b>Topic/Exercise</b>	<b>Instructors</b>
<b>9:00-1:00</b>	<ol style="list-style-type: none"> <li>1. STACKS - Running DENOVO maps and populations</li> <li>2. Visualizing SNP data and answering questions</li> </ol>	Dr. Katy Morgan
<b>1:00-2:00</b>	LUNCH	
<b>2:00-5:00</b>	<ol style="list-style-type: none"> <li>1. RADseq data processing to visualize population structure. <ol style="list-style-type: none"> <li>a. PCA (KM)</li> <li>b. STRUCTURE/fastSTRUCTURE. (CM)</li> </ol> </li> </ol>	Dr. Katy Morgan Courtney Miller

<b>DAY FOUR</b>		
<b>Time</b>	<b>Topic/Exercise</b>	<b>Instructors</b>
<b>9:00-1:00</b>	<ol style="list-style-type: none"> <li>1. Importance of detecting/predicting adaptive variation (NA)</li> <li>2. More RADseq data processing <ol style="list-style-type: none"> <li>a. VCF tools and PGDSpider</li> <li>b. Bayescan (or BayENV/LFMM) (KM)</li> </ol> </li> </ol>	Dr. Katy Morgan Dr. Nicola Anthony
<b>1:00-2:00</b>	LUNCH	

<b>2:00-5:00</b>	1. GDM case studies and analysis (CM)	Courtney Miller
------------------	---------------------------------------	-----------------

<b>DAY FIVE</b>		
<b>Time</b>	<b>Topic/Exercise</b>	<b>Instructors</b>
<b>9:00-1:00</b>	Summary and discussion day. More time to explore specific analyses.	All
<b>1:00-2:00</b>	LUNCH	
<b>2:00-5:00</b>	Could include discussions of literature or grant ideas, or reviews of anything people have further questions about.	<b>All</b>