Bat ID on the Fly: How to use DNA to Identify Species in the Field
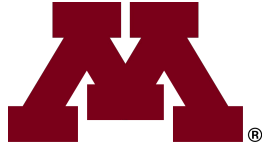
Lexi Frank
GBatNet Student Scholar
PhD Candidate
University of Minnesota

Background

Small mammals, disease, genomics, metagenomics, epigenomics, mobile laboratory

PhD Candidate at the University of Minnesota

GBatNet Student Scholar

Photo: Evan Kipp

Field work in the US and internationally

# Field Research in Guyana, South America

Agenda:
- Travel to our field site in Guyana
- Catch some bats!
- Extract and sequence DNA
- Analyze the sequences
- Identify the species

Field Site

- Kanashen Amerindian Protected Area - Masakenari Village
  - Owned and managed by Indigenous Waiwai
  - Conservation/zoonotic disease monitoring goals
  - Population of ~300 people
  - 625,000 ha protected forest
 - Low levels of human disturbance - subsistence hunting and farming
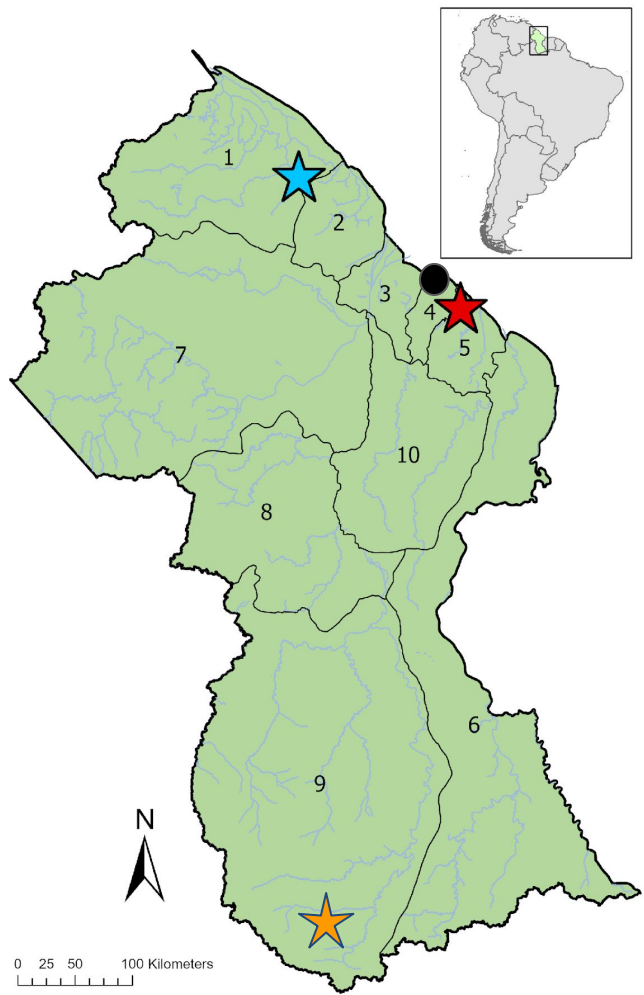
High Species Richness in Guyana

- Over 130 bat species in Guyana
  - 9 families
- Cryptic species – low to no morphological difference, but genetically different
- Species complex – closely related organisms similar in appearance and other features with undetermined relationships between them

Species Richness of Bats

Value

125

2
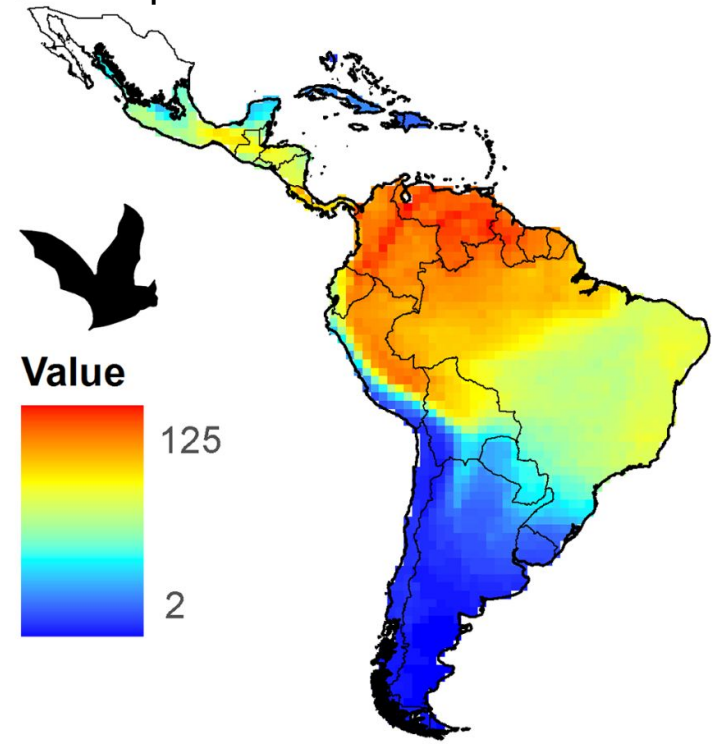
Road to Remote Sites in Guyana

Let's Catch Some Bats!

Rhynchonycteris naso

- Family: Emballonuridae
- Common Name: Proboscis bat
- Forearm: 36 mm
- Weight: 5.2 g
- Features:
  - Gray-grizzled brown with 2 faint white stripes
  - Evenly-spaced tufts of pale fur along forearms
  - Diet: Insectivore

Photo: Lexi Frank

Phyllostomus hastatus

- Family: Phyllostomidae
- Common Name: Greater spear-nosed bat
- Forearm: 64 mm
- Weight: 35.8 g
- Features:
  - Second largest Neotropical bat species
  - Pointed Ears
  - Diet: Omnivorous

Photo: Lexi Frank

9

Lonchophylla thomasi

- Family: Phyllostomidae
- Common Name: Thomas's nectar bat
- Forearm: 35 mm
- Weight: 8.8 g
- Features:
  - Long, narrow rostrum and tongue
  - Diet: Nectarivore

Photo: Lexi Frank

Lophostoma brasiliense

- Family: Phyllostomidae
- Common Name: Pygmy round-eared bat
- Forearm: 32 mm
- Weight: 7.3 g
- Features:
  - Large, rounded ears
  - Diet: Insectivore

Photo: Lexi Frank

Noctilio leporinus

- Family: Noctilionidae
- Common Name: Greater Bulldog Bat
- Forearm: 75 mm
- Weight: 60 g
- Features:
  - Long hind limbs with large claws
  - Pointed nose
  - Diet: Fish and Insects

Photo: Chris Shaffer

Desmodus rotundus bite

- Family: Phyllostomidae
- Common Name: Common Vampire bat
- Features:
  - Small (3-5 mm) bite mark on ear of a domestic dog
  - Difficult to clot wound

Photo: Lexi Frank

Desmodus rotundus

- Family: Phyllostomidae
- Common Name: Common Vampire bat
- Forearm: 54 mm
- Weight: 23 g
- Features:
  - Compact muzzle
  - Large thumbs
  - Strong back limbs
  - Diet: Sanguivore

Photo: Lexi Frank

Carollia sp.

- Family: Phyllostomidae
- Common Name: Short-tailed Fruit Bat
- Forearm: 16 mm
- Weight: 40 g
- Features:
  - Chin with central, rounded wart surrounded by smaller protuberances
  - *C. perspicillata* and *C. brevicauda* overlap in FA and Tibia lengths
- Good candidate for molecular field ID!

????????????

- Family: ?
- Common Name: ?
- Forearm: 63 mm
- Weight: 24 g
- Features:
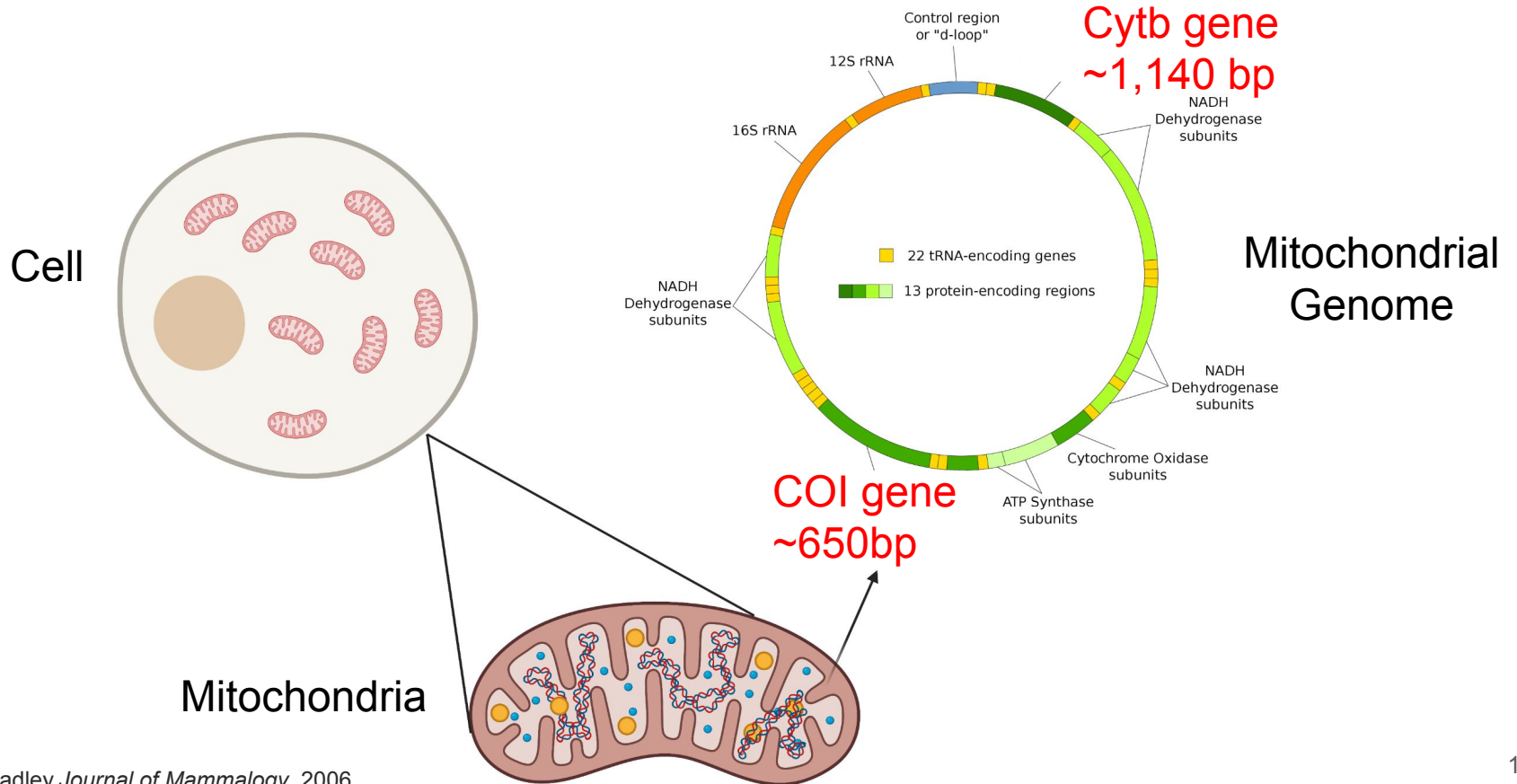 - Captured in the forest near humans
  - Diet: ?
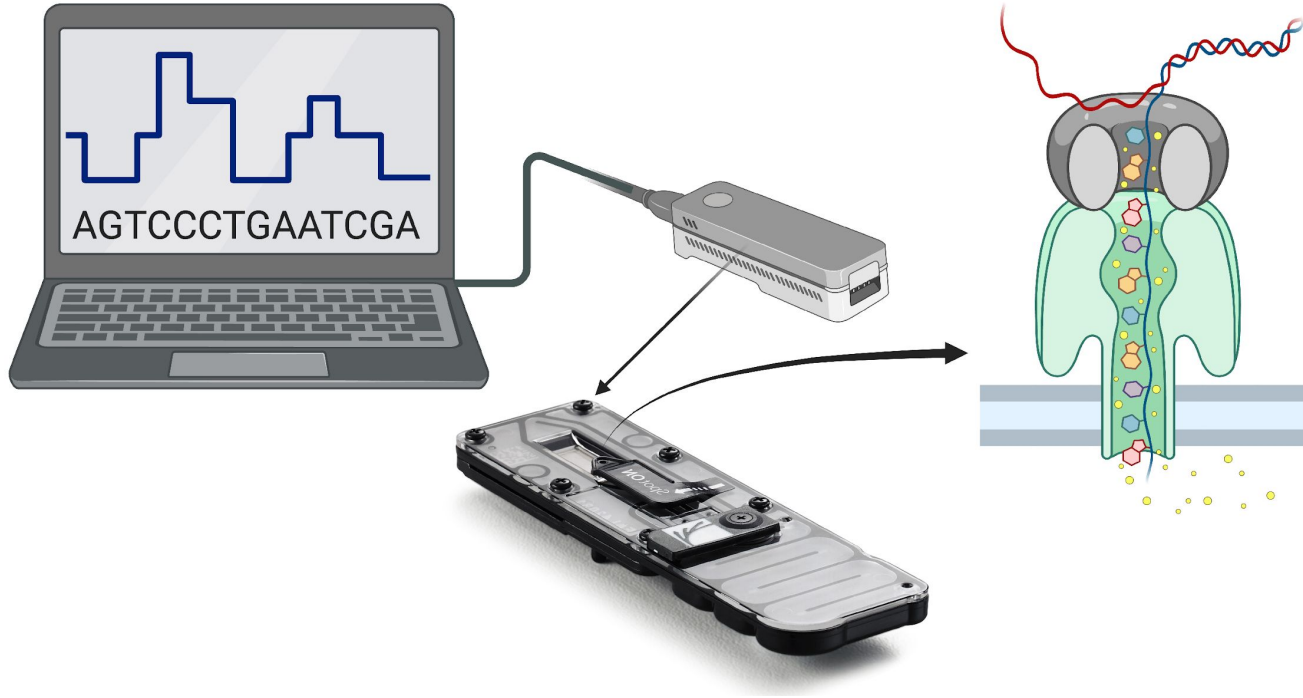
Why Use Molecular ID in the Field?

- Cryptic species, species complex
- Changes in taxonomy
- Need skull for morphological ID
- Geographically distance from available data
- Confirm target species for study/client
- Difficulties moving samples from field site to lab
  - Costs, laws/policies, time, etc.

Mitochondrial Genome: Standard for Species Barcoding

Cell

Mitochondria

Mitochondrial Genome

Cytb gene ~1,140 bp

COI gene ~650bp

Control region or "d-loop"

12S rRNA

16S rRNA

NADH Dehydrogenase subunits

NADH Dehydrogenase subunits

NADH Dehydrogenase subunits

Cytochrome Oxidase subunits

ATP Synthase subunits

22 tRNA-encoding genes

13 protein-encoding regions

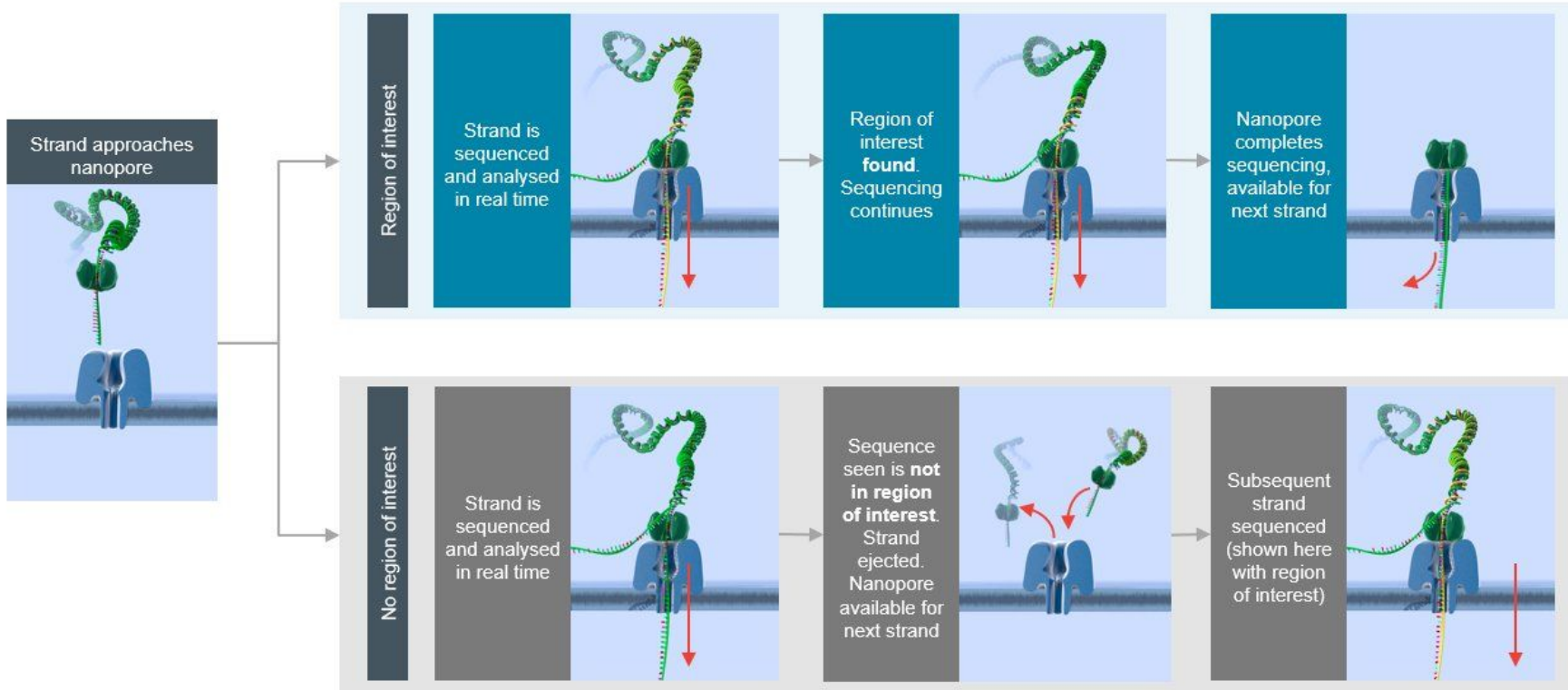# Nanopore Sequencing and Adaptive Sampling

AGTCCCTGAATCGA

Portable Equipment

Affordable for Individual Labs

Real-time Data Generation

No Length Restriction

Straightforward Library Prep

Oxford Nanopore Technology

Nanopore Sequencing and Adaptive Sampling

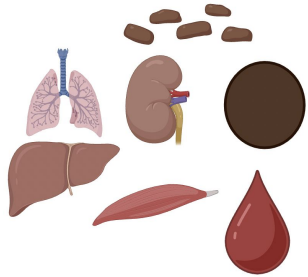Oxford Nanopore Technology

Setup the Mobile Lab
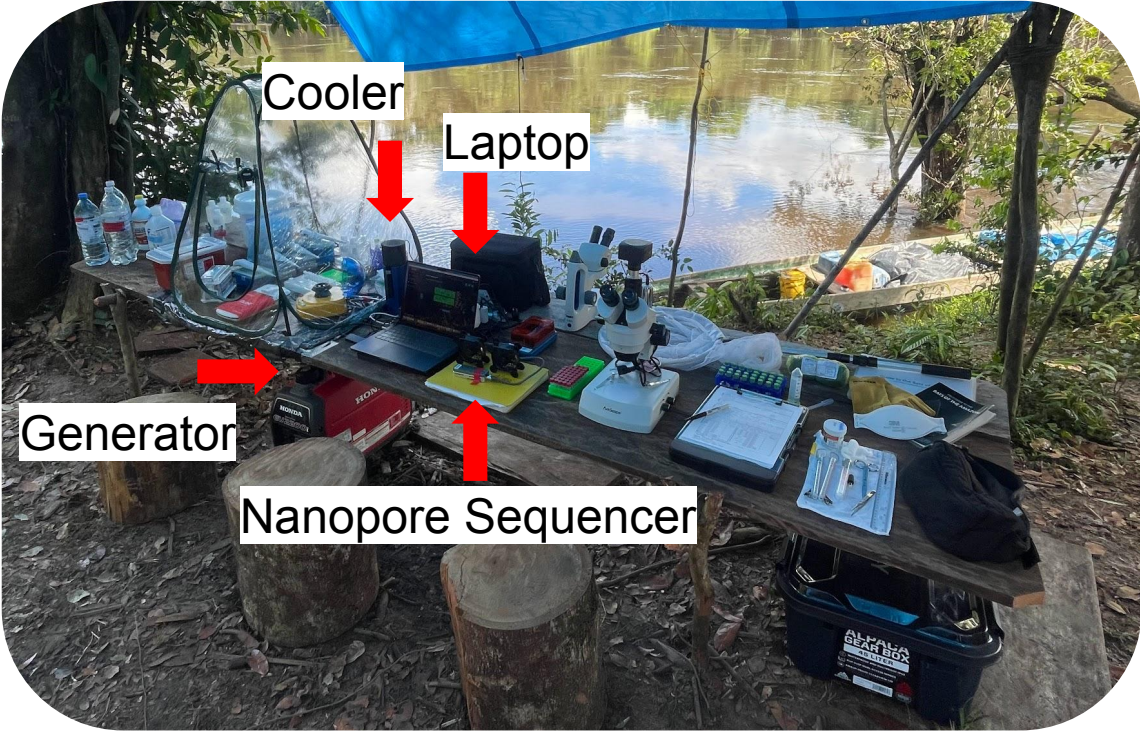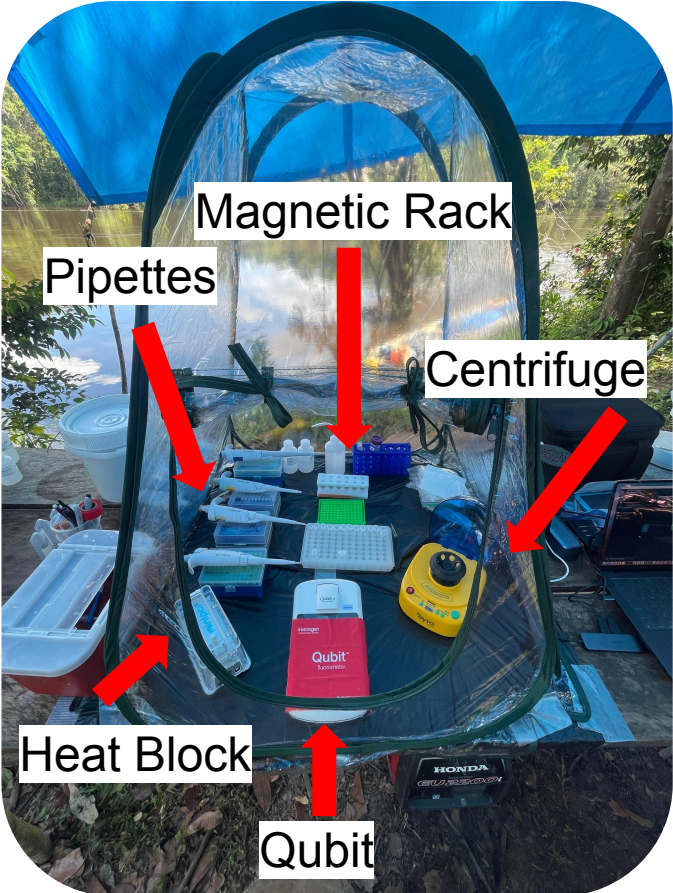
DNA Extraction

Choose a Sample   Homogenization   Lysis   Separation   Precipitation

Short Reads   Long Reads

Reference Genome

Assembled Genome

Sequence Gaps
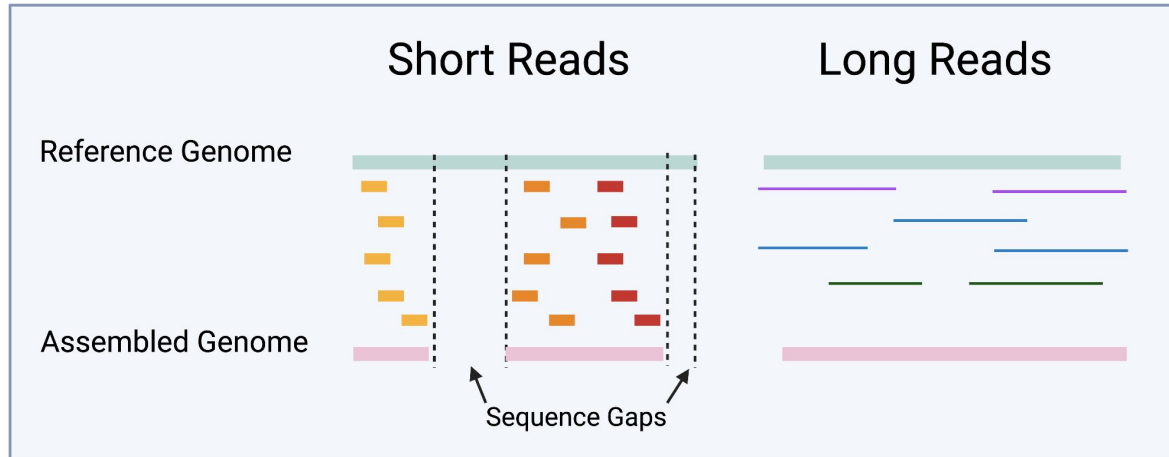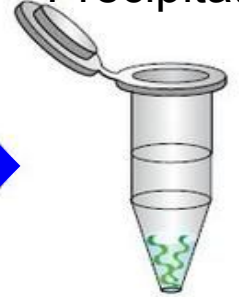
Data Analysis

Tools Needed:

- Access to Command Line (Linux, Mac, Galaxy website)
- Fastq files
- [Reference Mitogenomes from NCBI](#)
- Packages:
  - [Nanofilt](#) - Filter data for length and quality
  - [Minimap2](#) - Align your reads to a reference genome
  - [Samtools](#) - Reading/writing/editing/indexing/viewing SAM, BAM formats
  - [Flye](#) - Assembling sequences
  - [Galaxy: Mitos2](#) - Annotation of genome (labeling genes)
  - [RAxML](#) - Phylogenetic tree construction

Mitogenome Assembly

COI gene
~650bp

Cytb gene
~1,140 bp

Phylogenetic Tree

Carollia sowelli AF511962 N.S.
Carollia perspicillata AF511975 N.S.
Carollia perspicillata TK217513 Guyana: Mahaica
Carollia perspicillata HG003309 French Guiana
Carollia brevicauda AF187019 Peru
Carollia brevicauda TK217563 Guyana: Kwebana
Carollia brevicauda TK217562 Guyana: Kwebana
Carollia castanea ON310505 Ecuador
Sturnira bakeri ON357736 Ecuador
Artibeus lituratus NC016871 Paraguay
Platyrrhinus matapalensis ON357723 Ecuador
Anoura caudifer HG003307 Suriname
Trachops cirrhosus TK217600 Guyana: Kwebana
Trachops cirrhosus DQ233669 French Guiana
Vampyrum spectrum NC022429 French Guiana
Phyllostomus discolor ON357733 Ecuador
Tonatia saurophila NC022428 French Guiana
Tonatia bidens MZ391834 Brazil
Desmodus rotundus HG003310 Venezuela
Rhinophylla pumilio NC022426 French Guiana
Leptonycteris nivalis NC066830 Mexico
Glossophaga soricina ON321893 Ecuador
Pteronotus davyi MW526430 Peru
Pteronotus personatus AF338676 Guyana
Pteronotus personatus KX589869 Guyana
Pteronotus personatus KX589786 French Guiana
Pteronotus parnellii NC023368 N.S.
Pteronotus fuscus KX589870 Venezuela
Pteronotus sp1 alitonus KX589950 Suriname
Pteronotus sp1 alitonus KX589758 French Guiana
Pteronotus alitonus MF547504 Brazil
Pteronotus rubiginosus TK217611 Guyana
Pteronotus rubiginosus KX589948 Suriname
Pteronotus rubiginosus KX589956 Brazil
Pteronotus rubiginosus KX589772 French Guiana
Noctilio leporinus KU743910 French Guiana
Micronycteris megalotis HF947304 French Guiana
Myotis formosus NC015828 South Korea
Myotis simus JX130506 Ecuador
Myotis riparius ON357732 Ecuador
Myotis albescens ON357731 Ecuador
Myotis albescens TK217523 Guyana: Mahaica
Pipistrellus abramus NC005436 Japan
Molossus molossus ON357729 Ecuador
Pteropus dasymallus NC002612 Japan

Bootstrap support value
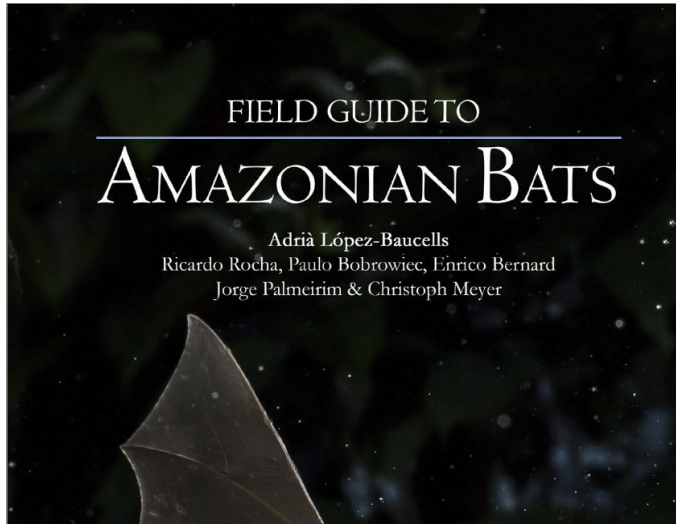- 60
- 70
- 80
- 90
- 100

Tree scale: 0.1

Photo: Chris Shaffer

25

Final Species ID

Pteronotus rubiginosus

- Family: Mormoopidae
- Common Name: Wagner's Common Mustached Bat
- Features:
  - Formerly considered a subspecies of *Pteronotus parnellii*
  - Diet: Insectivore

FIELD GUIDE TO AMAZONIAN BATS

Adrià López-Baucells
Ricardo Rocha, Paulo Bobrowiec, Enrico Bernard
Jorge Palmeirim & Christoph Meyer

Field Guide to Amazonian Bats

Mormoopidae *(Moustached bats)*

Pteronotus (Moustached bats)

1a. Bare back; wings attached on the middle of dorsum.

   2a. FA < 49 mm.

                Pteronotus davyi (40-49 mm)

   2b. FA > 49 mm.

                Pteronotus gymnonotus (50-55 mm)

1b. Furry back; wings attached to the side of body.

   2a. FA > 50 mm.

                Pteronotus cf parnellii * (50-63 mm)

   2b. FA < 50 mm.

                Pteronotus cf personatus (40-48 mm)

* Cryptic species complex. We recommend using acoustic and/or genetic data for species identification (see also echolocation keys at the end of this guide).

1a   1b

* Cryptic species complex. We recommend using acoustic and/or genetic data for species identification (see also echolocation keys at the end of this guide).

84

27

Resources For Bioinformatics/Analysis

Bioinformatics

- [Practical Computing for Biologists by Casey Dunn and Steven Haddock](#) - Book that shows how to use general computing tools to work more effectively
- [Data Carpentry](#) - Lesson program teaching universal data skills
- [Galaxy](#) - Open source, web-based platform for data intensive biomedical research
- [Linux Journey](#) - Basics of Linux tutorials for free
- [Github](#) - Download software packages, access manuals and issues reports
- Command Line Cheat Sheet - Easy to find with Google

Analysis

- [NCBI Blast](#) - Search your sequences
- [Barcode of Life](#) - Database of COI sequences, search your sequences
- [NCBI Organelle Refseq](#) - Reference quality mitochondrial genomes
- [Frank et al](#) - Paper outlining mitogenome species ID, see supplementary data
- [Anaconda](#) - Used to easily download some software packages

Acknowledgements

Guyana Research Team
Dr. Chris Shaffer
Dr. Marissa Milstein

Guyana Collaborators - Waiwai community, Kwebana community, Mahaica community

Dr. Peter Larsen and Larsen Lab
Dr. Evan Kipp

Drs. Chris Faulk and Tiffany Wolf

Funding provided by University of Minnesota

Disclaimer: Situations/data were altered for the purpose of this workshop

Questions?

Email: fran1464@umn.edu
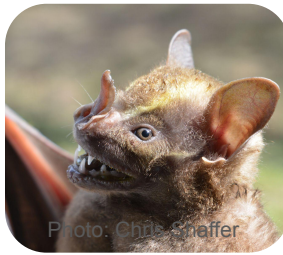
Nanopore Sequencing and Adaptive Sampling

- Nanopore Website:
  https://nanoporetech.com/documentation
- Adaptive Sampling Documentation:
  https://nanoporetech.com/document/adaptive-sampling
- Adaptive Sampling Video:
  https://nanoporetech.com/resource-centre/adaptive-sampling-oxford-nanopore

Command Line Cheat Sheet

DIRECTORIES

Display path of current working directory
`$ pwd`

Change directory to <directory>
`$ cd <directory>`

Navigate to parent directory
`$ cd ..`

List directory contents
`$ ls`

List detailed directory contents, including hidden files
`$ ls -la`

Create new directory named <directory>
`$ mkdir <directory>`

OUTPUT

Output the contents of <file>
`$ cat <file>`

Output the contents of <file> using the less command (which supports pagination etc.)
`$ less <file>`

Output the first 10 lines of <file>
`$ head <file>`

Direct the output of <cmd> into <file>
`$ <cmd> > <file>`

Append the output of <cmd> to <file>
`$ <cmd> >> <file>`

Direct the output of <cmd1> to <cmd2>
`$ <cmd1> | <cmd2>`

Clear the command line window
`$ clear`

From Tower.com

FILES

Delete <file>
`$ rm <file>`

Delete <directory>
`$ rm -r <directory>`

Force-delete <file> (add -r to force-delete a directory)
`$ rm -f <file>`

Rename <file-old> to <file-new>
`$ mv <file-old> <file-new>`

Move <file> to <directory> (possibly overwriting an existing file)
`$ mv <file> <directory>`

Copy <file> to <directory> (possibly overwriting an existing file)
`$ cp <file> <directory>`

Copy <directory1> and its contents to <directory2> (possibly overwriting files in an existing directory)
`$ cp -r <directory1> <directory2>`

Update file access & modification time (and create <file> if it doesn't exis
`$ touch <file>`

PERMISSIONS

Change permissions of <file> to 755
`$ chmod 755 <file>`

Change permissions of <directory> (and its contents) to 600
`$ chmod -R 600 <directory>`

Change ownership of <file> to <user> and <group> (add -R to include a directory's contents)
`$ chown <user>:<group> <file>`

SEARCH

Find all files named <file> inside <dir> (use wildcards [*] to search for parts of filenames, e.g. "file.*")
`$ find <dir> -name "<file>"`

Output all occurrences of <text> inside <file> (add -i for case-insensitivity)
`$ grep "<text>" <file>`

Search for all files containing <text> inside <dir>
`$ grep -rl "<text>" <dir>`

NETWORK

Ping <host> and display status
`$ ping <host>`

Output whois information for <domain>
`$ whois <domain>`

Download <file> (via HTTP[S] or FTP)
`$ curl -O <url/to/file>`

Establish an SSH connection to <host> with user <username>
`$ ssh <username>@<host>`

Copy <file> to a remote <host>
`$ scp <file> <user>@<host>:/remote/path`

PROCESSES

Output currently running processes
`$ ps ax`

Display live information about currently running processes
`$ top`

Quit process with ID <pid>
`$ kill <pid>`

Bioinformatics

Reads and Depth - Reads refers to the sequence of DNA bases of a single DNA fragment (contained in a fastq). Depth is the number of reads that align to or 'cover' a particular section of the genome/gene. Greater depth = greater confidence in the identity of a base. Coverage is the percentage of the whole genome that has been sequenced.

5x depth for this section of the genome

A 'Read' (blue bars) is the sequence of bases representing a single DNA fragment

AGGTCAACGTCAGCA

Read depth

5x
4x
3x
2x
1x

Genome
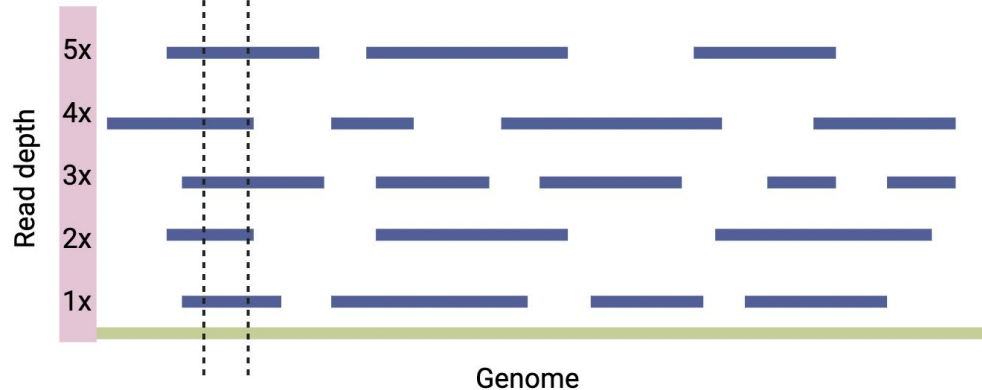
Bioinformatics

Phred Quality Score - measure of the quality of the identification of a base generated by sequencing. Example: if the Phred score of a read is 30, then the chances of a base being incorrect are 1 in 1000. Higher Phred score = Higher accuracy

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Table from wikipedia

Bioinformatics

Fastq file format – text-based format for storing both a biological sequence and its corresponding quality scores

https://biocorecrg.github.io/RNAseq_course_2019/fastq.html
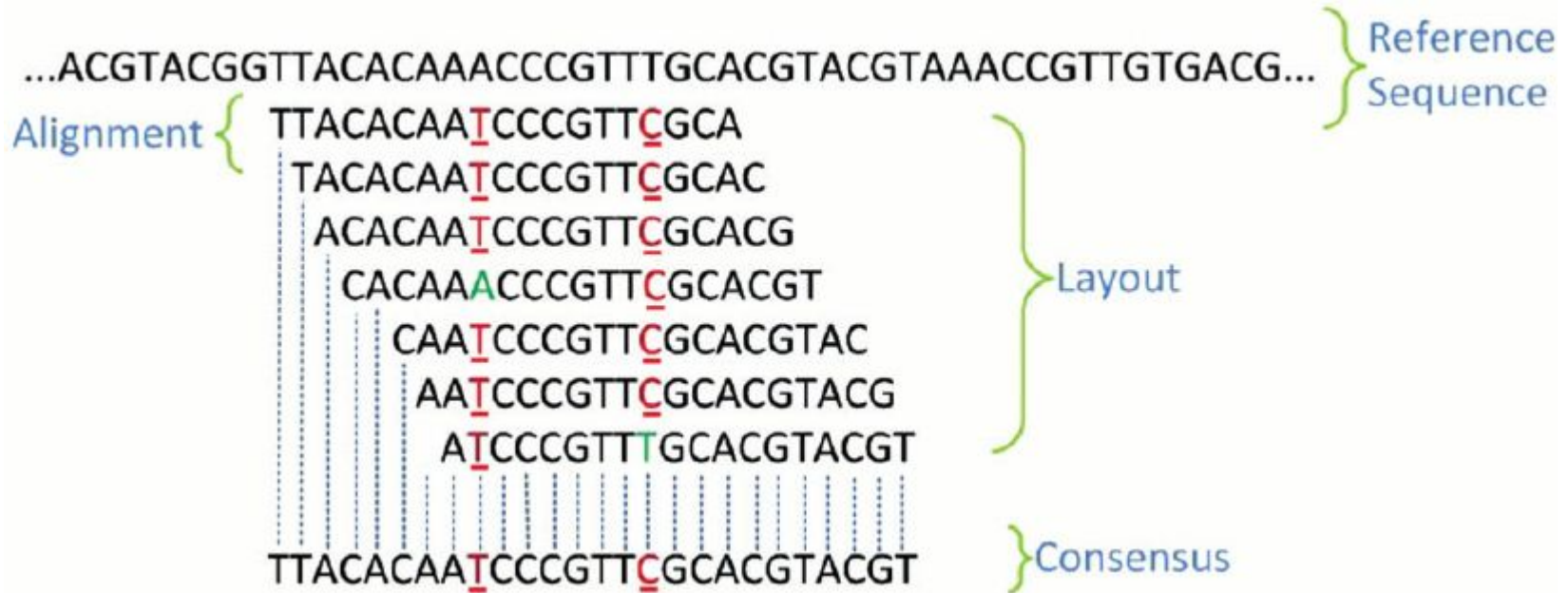
Bioinformatics

SAM and BAM file formats –
SAM files are a type of text file format that contains the alignment information of various sequences that are mapped against reference sequences. These files can also contain unmapped sequences. Since SAM files are a text file format, they are more readable by humans.

BAM files contain the same information as SAM files, except they are in binary file format which is not readable by humans. BAM files are smaller and more efficient for software to work with than SAM files, saving time and reducing costs of computation and storage. Alignment data is almost always stored in BAM files and most software that analyzes aligned reads expects to ingest data in BAM format

Aligned reads
```
TGAAGTCCTACAGTCATAGTC
 AAGTCCTACAGTCATAGTCGA
   GTCCTACAGTCATAGTCGATA
     CCTACAGTCATAGTCGATATT
       TACAGTCATAGTCGATATTT
```

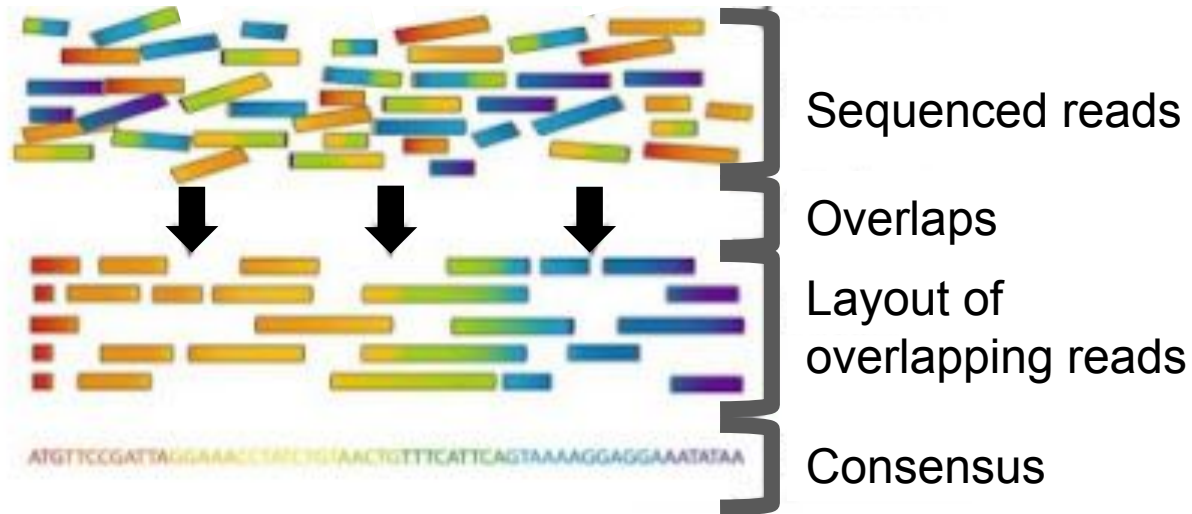Consensus contig TGAAGTCCTACAGTCATAGTCGATATTT

Bioinformatics

Sequence Alignment to a Reference (using minimap2) – Reads are aligned to a reference sequence. The alignment process may allow one or more mismatches between each individual read and the reference sequence.
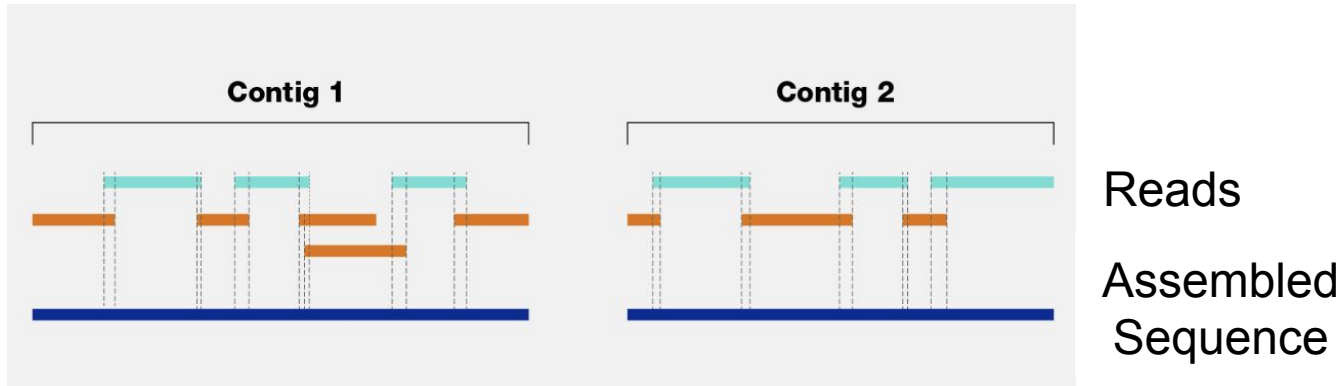
Bioinformatics

De novo assembly (using Flye) – assembling sequencing reads to create full-length (sometimes novel) sequences, without using a reference.

Sequenced reads

Overlaps

Layout of overlapping reads

Consensus

ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

Bioinformatics

Contig – A contig (as related to genomic studies; derived from the word "contiguous") is a set of DNA segments or sequences that overlap in a way that provides a contiguous representation of a genomic region.

https://www.genome.gov/genetics-glossary/Contig

Bioinformatics

- How to find fastq data to test out bioinformatic workflows?
  - NCBI SRA Data – Raw data is published to NCBI Sequence Read Archive Database and can be downloaded
  - European Nucleotide Archive – The European version of NCBI SRA
  - Zenodo – Raw data can also be published on Zenodo and available for download
  - Github – Github is mainly used for storing code, but also has datasets available for download
  - Journal Articles – Papers will often include how to access their raw data