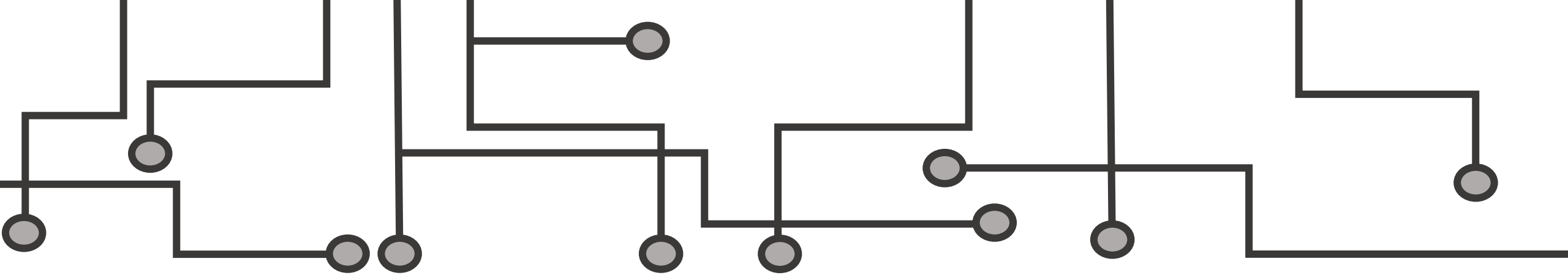


MEDICAL MICROBIOLOGY AND INFECTIOUS DISEASES BIOINFORMATICS WORKSHOP

Presents

Introduction to Genomics and Viral data analysis

INSTRUCTED BY

Grace E. Seo, MSc Student

INFORMATION FOR PARTICIPANTS

All workshops are being recorded and posted to the MMID Bioinformatics Workshop – YouTube

For live Q&A, go to slido.com and use participant code #3661069

2023 MMID Bioinformatics Workshop Schedule

DATE	INSTRUCTOR	TOPIC
March 2	Grace E. Seo	Introduction to the 2023 MMID Bioinformatics Workshop
March 9	Grace E. Seo	Introduction to conda and tool installation
March 16	Grace E. Seo	Introduction to genomics and viral data analysis
March 23	Jill Rumore	Bacterial Genomics
March 30	Jill Rumore	Reference Databases
April 6	Taylor Davedow	Beginner's Guide to Phylogenetic Trees
April 13	Taylor Davedow	Introduction to tree visualization and annotation using ggtree
April 20	-	Bfx workshop: Bring your own dataset!
April 27	-	Bfx workshop: Bring your own dataset!

April 20 and April 27 in-person sessions are open to the public (up to 100 people)!

Work with your colleagues/friends to analyze data together.

SET UP WI-FI (IN-PERSON PARTICIPANTS)

1. ***Connect to UofM-secure (if you are a student or staff)***
   - Use your @myumanitoba.ca or @umanitoba.ca login and password

2. ***Connect to UofM-guest***

To access uofm-guest Wi-Fi:

1. Ensure your wireless card is active and connected to the **uofm-guest** network.
2. Open your web browser (e.g. Google Chrome, Microsoft Edge, Firefox, etc.) and browse to any website. This should redirect you to the **Acceptable Use Agreement** page.
3. Review the Acceptable Use Agreement for the unsecured wireless.
4. Select **I Agree**.

LEARNING OBJECTIVES

1. ***Configure conda channels and review conda tool installation.***
2. ***Describe the methods used for viral genomics data.***
3. ***Provide a background on nanopore long-read sequencing.***
4. ***Provide an overview of King et al., 2020 paper for workshop practice analysis.***
5. ***Provide steps for fastq data analysis.***
6. ***Analyze nanopore sequence data.***

Click to go back to the beginning

LEARNING OBJECTIVES

1. ***Configure conda channels and review conda tool installation.***
2. *Describe the methods used for viral genomics data.*
3. *Provide a background on nanopore long-read sequencing.*
4. *Provide an overview of King et al., 2020 paper for workshop practice analysis.*
5. *Provide steps for fastq data analysis.*
6. *Analyze nanopore sequence data.*

*Configure conda channels

- For newly installed miniconda3, channels need to be configured.

- For anyone experiencing trouble with installing packages i.e. samtools, this is because conda-forge and bioconda channel was not configured.

```
(base) seog@SMARTY:~$ conda config --show channels
channels:
  - bioconda
  - conda-forge
  - defaults
```

Configure conda channels

1. Open terminal and check your conda channels

```
$ conda config --show channels
```

2. Add the following channels in following order if not configured already.

```
$ conda config --add channels conda-forge
$ conda config --add channels bioconda
```

Update conda tools again

3. Check your conda channels again. You should now see three channels in following order.

```
$ conda config --show channels
```

```
(base) seog@SMARTY:~$ conda config --show channels
channels:
  - bioconda
  - conda-forge
  - defaults
```

4. Update conda tools

```
$ conda update -y -c bioconda PACKAGENAME
```

Remove and re-add conda channels

***If your conda channel order is in a wrong order, remove and re-add channels.**

```
$ conda config --remove channels conda-forge
$ conda config --remove channels bioconda

$ conda config --add channels conda-forge
$ conda config --add channels bioconda
```

Conda channel configuration

Demonstration

Install/update conda packages

1. Activate conda environment

```
$ conda activate conda_workshop
```

2. Install the conda package for today's workshop

```
$ conda install -y -c bioconda nanoplot
```

Install/update conda packages

3. Update following conda packages (if you've installed these already from last week's workshop)

```
$ conda update -y fastqc
$ conda update -y minimap2
$ conda update -y vcftools
$ conda update -y samtools
$ conda update -y fastp
$ conda update -y checkm-genome
$ conda update -y kraken2
```

Click to go back to the beginning

LEARNING OBJECTIVES

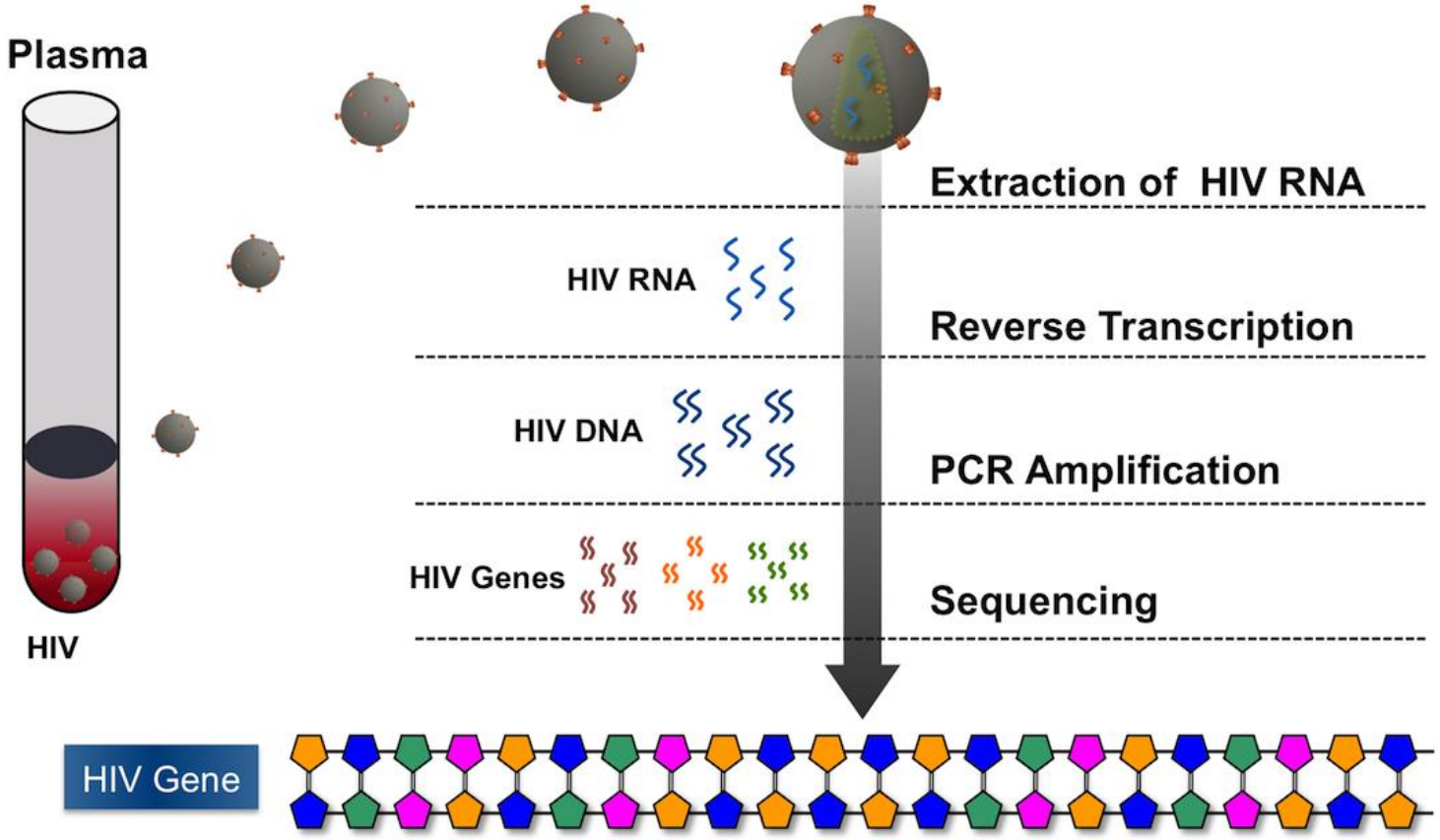
1. *Configure conda channels and review conda tool installation.*
2. ***Describe the methods used for viral genomics data.***
3. *Provide a background on nanopore long-read sequencing.*
4. *Provide an overview of King et al., 2020 paper for workshop practice analysis.*
5. *Provide steps for fastq data analysis.*
6. *Analyze nanopore sequence data.*

Why sequence virus genome?

- Virus genome sequence can be used for:
  - Detecting drug resistance[1]
  - Understanding virus pathogenicity[2]
  - Genomic surveillance (i.e. SARS-CoV-2 lineage profile)[3]

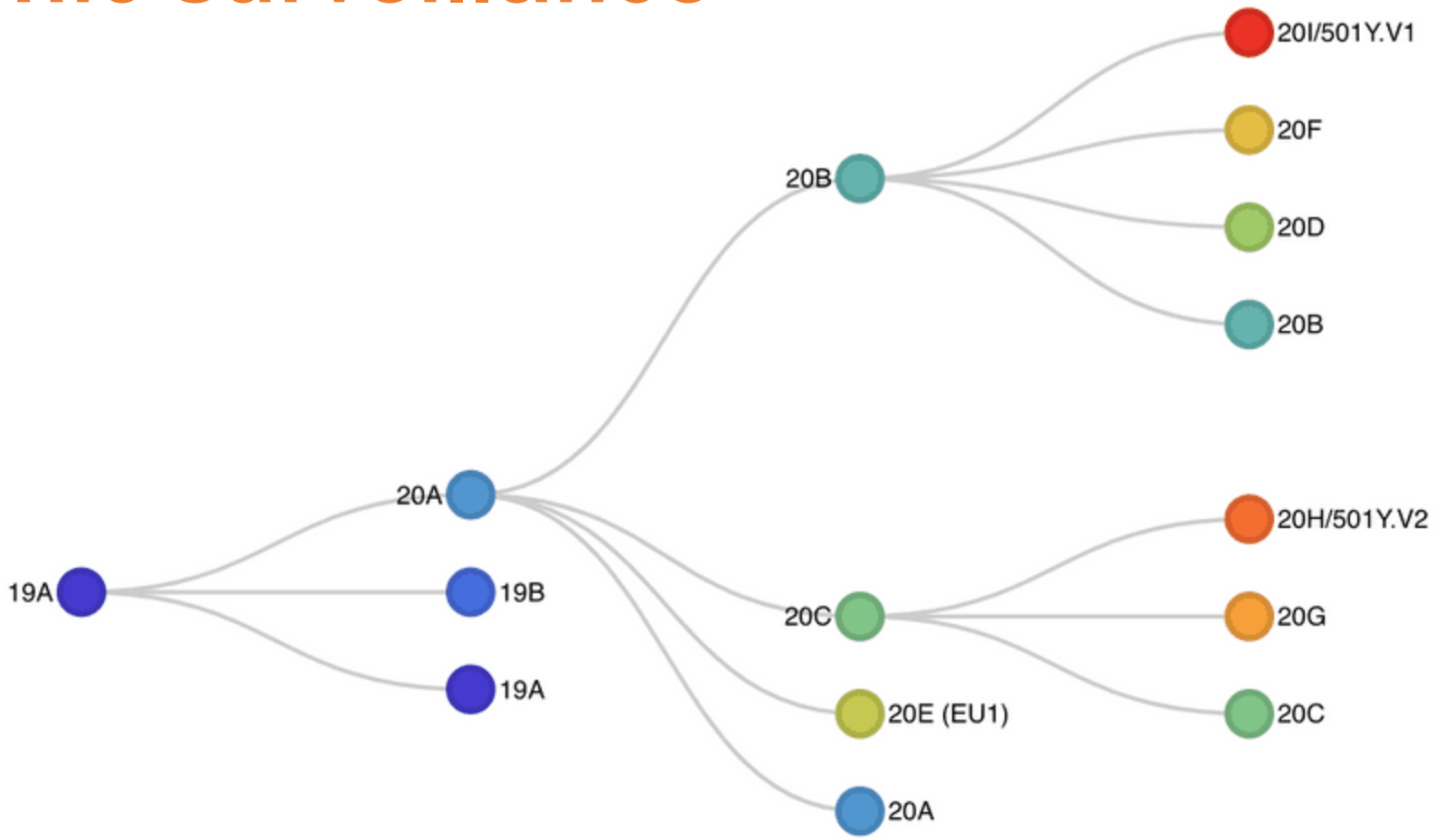
1. Houldcroft et al., 2017. Nat Rev Microbiol. 15: 183-192. doi: https://doi.org/10.1038/nrmicro.2016.182
2. Maurier et al., 2019. Virology. 531: 141-148. doi: https://doi.org/10.1016/j.virol.2019.03.006.
3. Chen et al., 2022. Nat Genet. 54: 499-507. doi: https://doi.org/10.1038/s41588-022-01033-y.

Drug Resistance



Wood, B. R. 2022. https://www.hiv.uw.edu/go/antiretroviral-therapy/evaluation-management-virologic-failure/core-concept/all

Genomic surveillance



Clade naming. https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming

Click to go back to the beginning

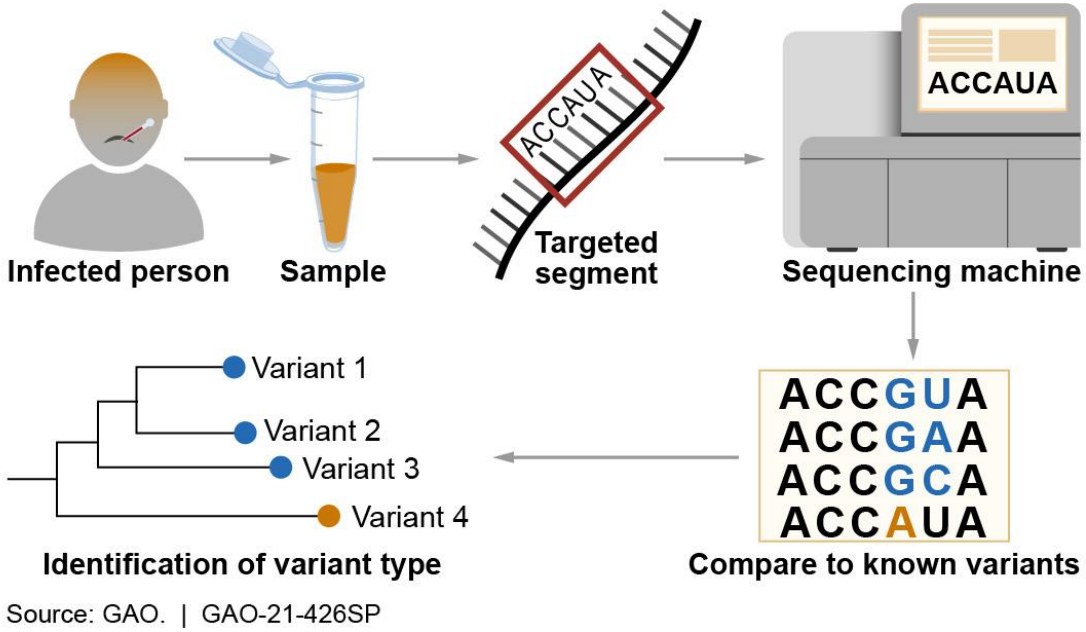
LEARNING OBJECTIVES

1. *Configure conda channels and review conda tool installation.*
2. *Describe the methods used for viral genomics data.*
3. ***Provide a background on nanopore long-read sequencing.***
4. *Provide an overview of King et al., 2020 paper for workshop practice analysis.*
5. *Provide steps for fastq data analysis.*
6. *Analyze nanopore sequence data.*

Genome sequencing

Genomic sequencing can be performed using

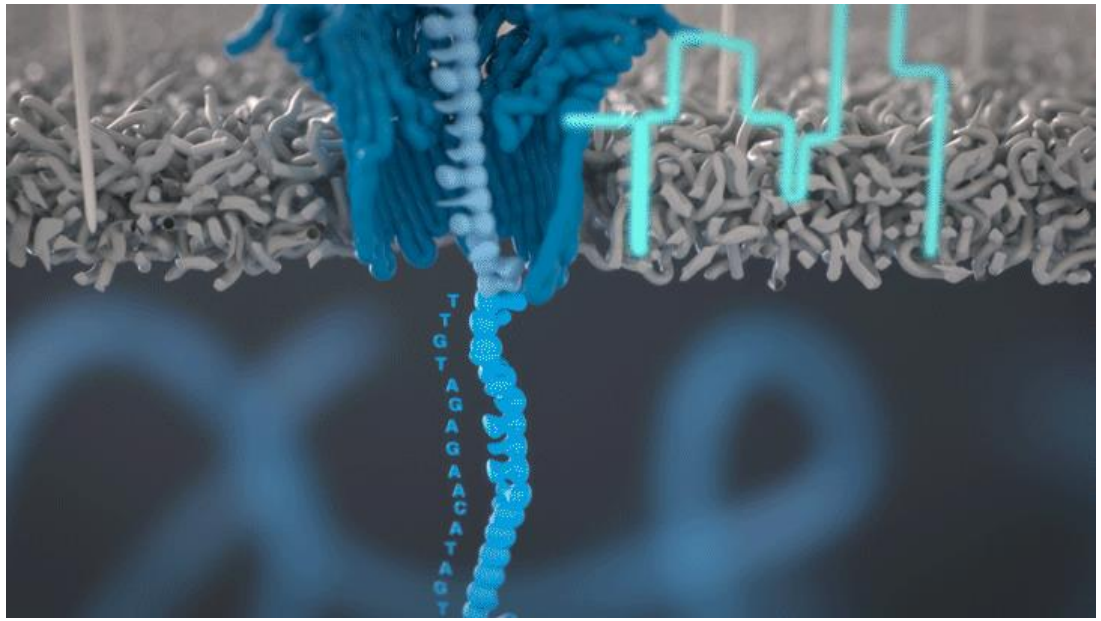
- Sanger
- Illumina (next-generation sequencing)
- Nanopore
- PacBio



https://www.gao.gov/products/gao-21-426sp

Nanopore sequencing technology

Nanopore is one of the long-read sequencing technology by Oxford Nanopore Technologies (ONT).



https://nanoporetech.com/support/how-it-works

Nanopore sequencing technology

An example of data analysis involves:

1. **fast5 (basecalling)**
2. **Fastq**
3. **read-mapping/*de novo***
4. **consensus genome**

https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Click to go back to the beginning

LEARNING OBJECTIVES

1. *Configure conda channels and review conda tool installation.*
2. *Describe the methods used for viral genomics data.*
3. *Provide a background on nanopore long-read sequencing.*
4. ***Provide an overview of King et al., 2020 paper for workshop practice analysis.***
5. *Provide steps for fastq data analysis.*
6. *Analyze nanopore sequence data.*

King et al., 2020 Influenza A viruses

- Authors developed a high-throughput sequencing workflow and speedy screening method for unknown IAV samples.

- Author's data analysis method:
  1. Guppy basecalling to convert fast5 → fastq
  2. Guppy demultiplexing and adapter trimming (separate by barcodes)
  3. NanoPlot read statistics
  4. Geneious (Bowtie2) reference read-mapping

Click to go back to the beginning

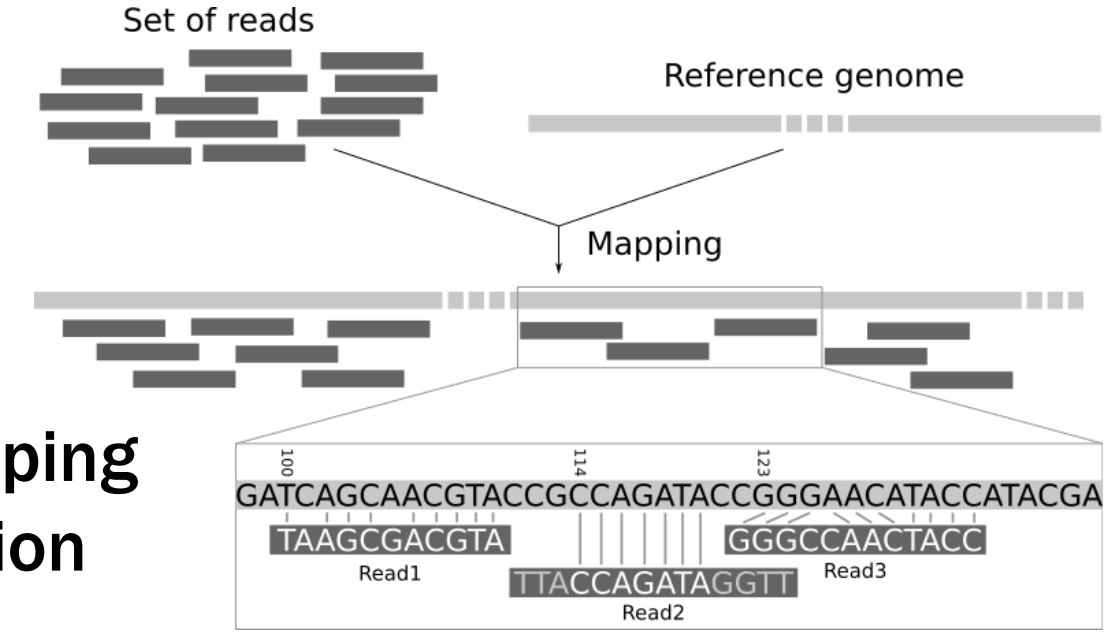
LEARNING OBJECTIVES

1. *Configure conda channels and review conda tool installation.*
2. *Describe the methods used for viral genomics data.*
3. *Provide a background on nanopore long-read sequencing.*
4. *Provide an overview of King et al., 2020 paper for workshop practice analysis.*
5. ***Provide steps for fastq data analysis.***
6. *Analyze nanopore sequence data.*

Today's data analysis workflow

What we will do today:

1. fastq (download)
2. NanoPlot read statistics
3. minimap2 reference read-mapping
4. Samtools sam to bam conversion
5. View bam files on IGV



https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Click to go back to the beginning

LEARNING OBJECTIVES

1. *Configure conda channels and review conda tool installation.*
2. *Describe the methods used for viral genomics data.*
3. *Provide a background on nanopore long-read sequencing.*
4. *Provide an overview of King et al., 2020 paper for workshop practice analysis.*
5. *Provide steps for fastq data analysis.*
6. ***Analyze nanopore sequence data.***

Download fastq data from GitHub

1. Open terminal and change directory to your preferred destination.

```
$ cd /mnt/c/Users/gsezz/Desktop
```

2. Clone GitHub repository

```
$ git clone https://github.com/mmid-bioinformatics-workshop/2023-03-16-Intro-to-genomics-and-viral-data-analysis.git
```

Create folders

3. Change directory into cloned git repository.

```
$ cd 2023-03-16-Intro-to-genomics-and-viral-data-analysis/King_et_al_2020/data_influenza
```

4. Create an analysis folder.

```
$ DATE=$(date +%Y%m%d_%H%M)
$ mkdir -p analysis_$DATE && cd analysis_$DATE
$ mkdir -p 001_nanoplot && cd 001_nanoplot
$ mkdir -p ERR3822171 ERR3822172
```

Run NanoPlot on fastq files

5A. Run NanoPlot on sample ERR3822171.

```
$ cd ERR3822171 && NanoPlot -t 2 --fastq
../../../data_influenza/ERR3822171_1.fastq.gz --
maxlength 4000 --tsv_stats && cd ..
```

5B. Run NanoPlot on sample ERR3822172.

```
$ cd ERR3822172 && NanoPlot -t 2 --fastq
../../../data_influenza/ERR3822172_1.fastq.gz --
maxlength 4000 --tsv_stats && cd ..
```

https://github.com/wdecoster/NanoPlot

Create minimap2 fasta reference index

6. Create an reference file index.

```
$ cd $PATH/2023-03-16-Intro-to-genomics-and-viral-data-analysis/King_et_al_2020
```

```
$ minimap2 -x map-ont –d influenza_references.mmi influenza_references.fasta
```

https://github.com/lh3/minimap2

Run minimap2 on fastq files

7. Create minimap2 folder.

```
$ cd analysis_$DATE
$ mkdir -p 002_minimap2 && cd 002_minimap2
```

Run minimap2 on fastq files

8A. Run minimap2 on sample ERR3822171.

```
$ minimap2 -ax map-ont
../../influenza_references.fasta
../../data_influenza/ERR3822171_1.fastq.gz >
ERR3822171.sam
```

8B. Run minimap2 on sample ERR3822172.

```
$ minimap2 -ax map-ont
../../influenza_references.fasta
../../data_influenza/ERR3822172_1.fastq.gz >
ERR3822172.sam
```

https://github.com/lh3/minimap2

Convert .sam to .bam output

9A. Convert .sam output to .bam on sample ERR3822171.

```
$ samtools view -S -b ERR3822171.sam > ERR3822171.bam
```

9B. Convert .sam output to .bam on sample ERR3822172.

```
$ samtools view -S -b ERR3822172.sam > ERR3822172.bam
```

Sort .bam output

10A. Sort .bam output on sample ERR3822171.

```
$ samtools sort ERR3822171.bam -o
ERR3822171.sorted.bam
```

10B. Sort .bam output on sample ERR3822172.

```
$ samtools sort ERR3822172.bam -o
ERR3822172.sorted.bam
```

Create an index

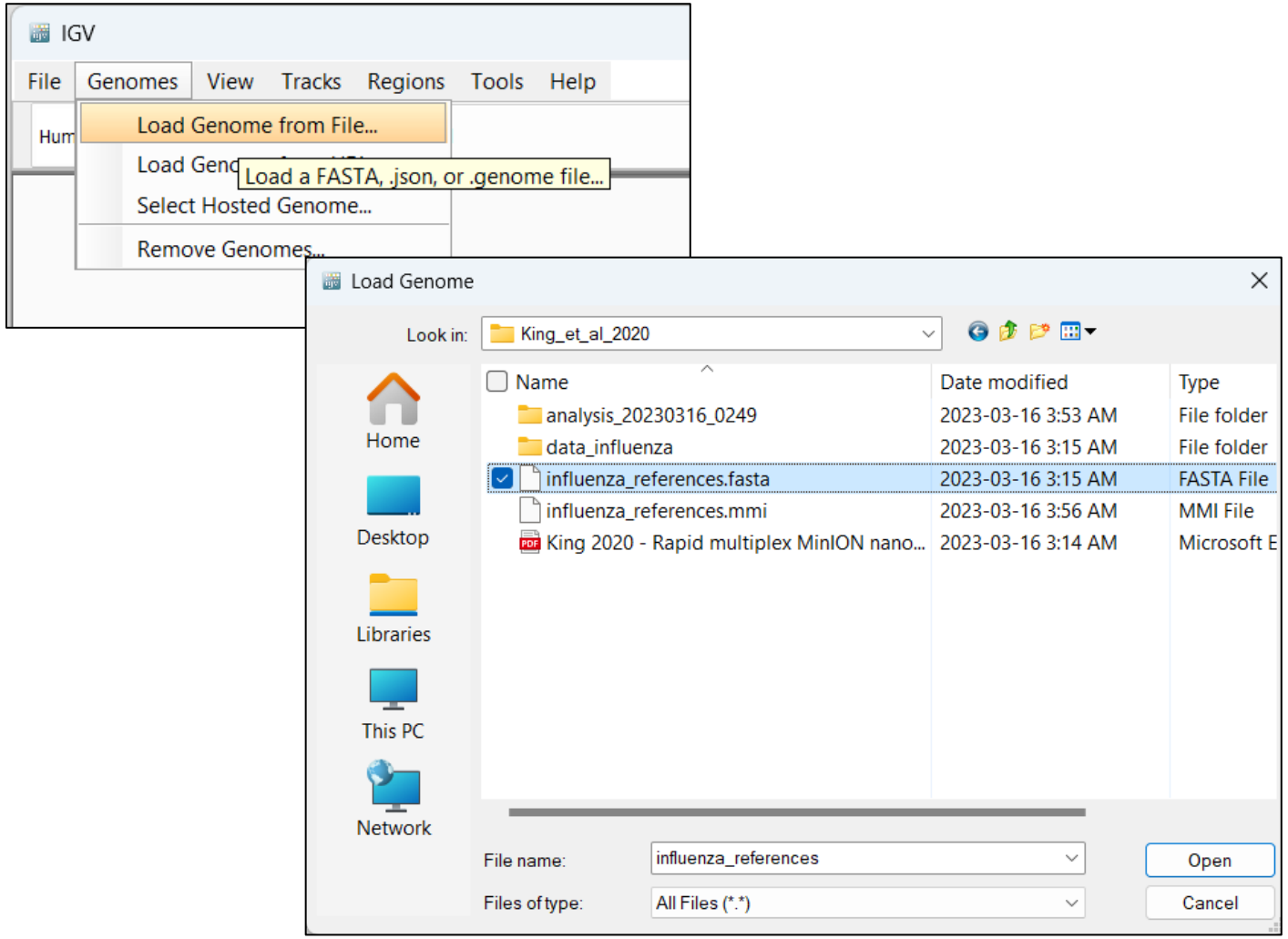
11A. Create an index of .bam on sample ERR3822171.

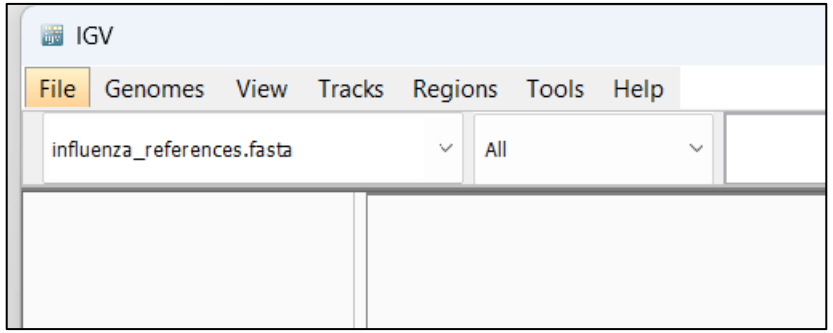
```
$ samtools index ERR3822171.sorted.bam
```

11B. Create an index of .bam on sample ERR3822172.

```
$ samtools index ERR3822172.sorted.bam
```

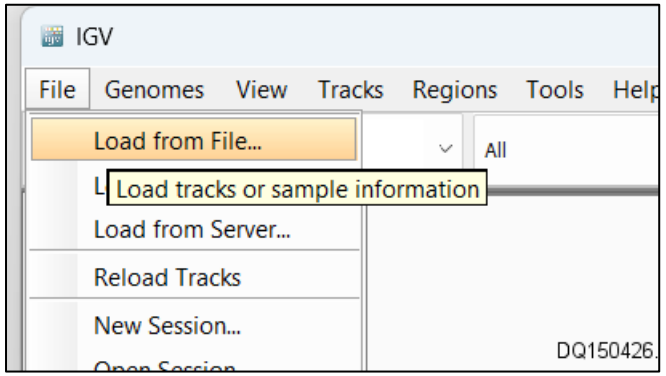
View bam output on IGV

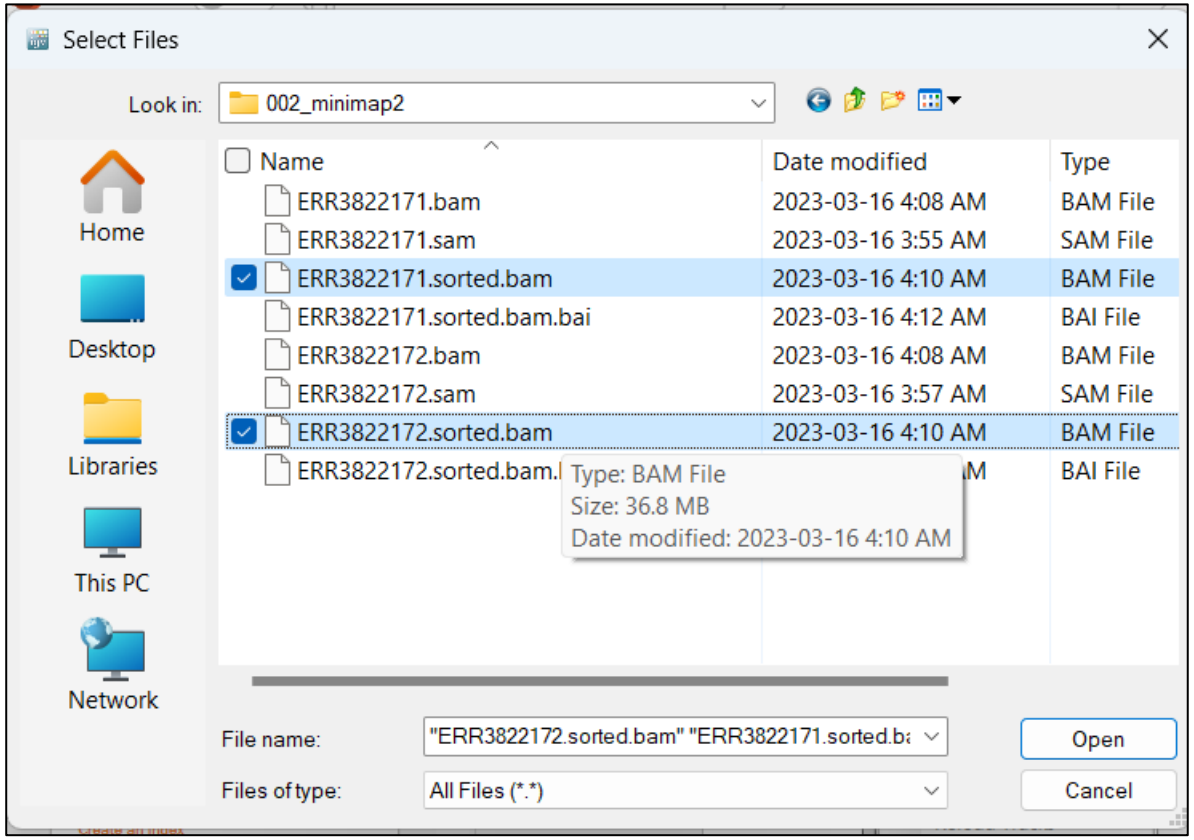




https://software.broadinstitute.org/software/igv/userguide

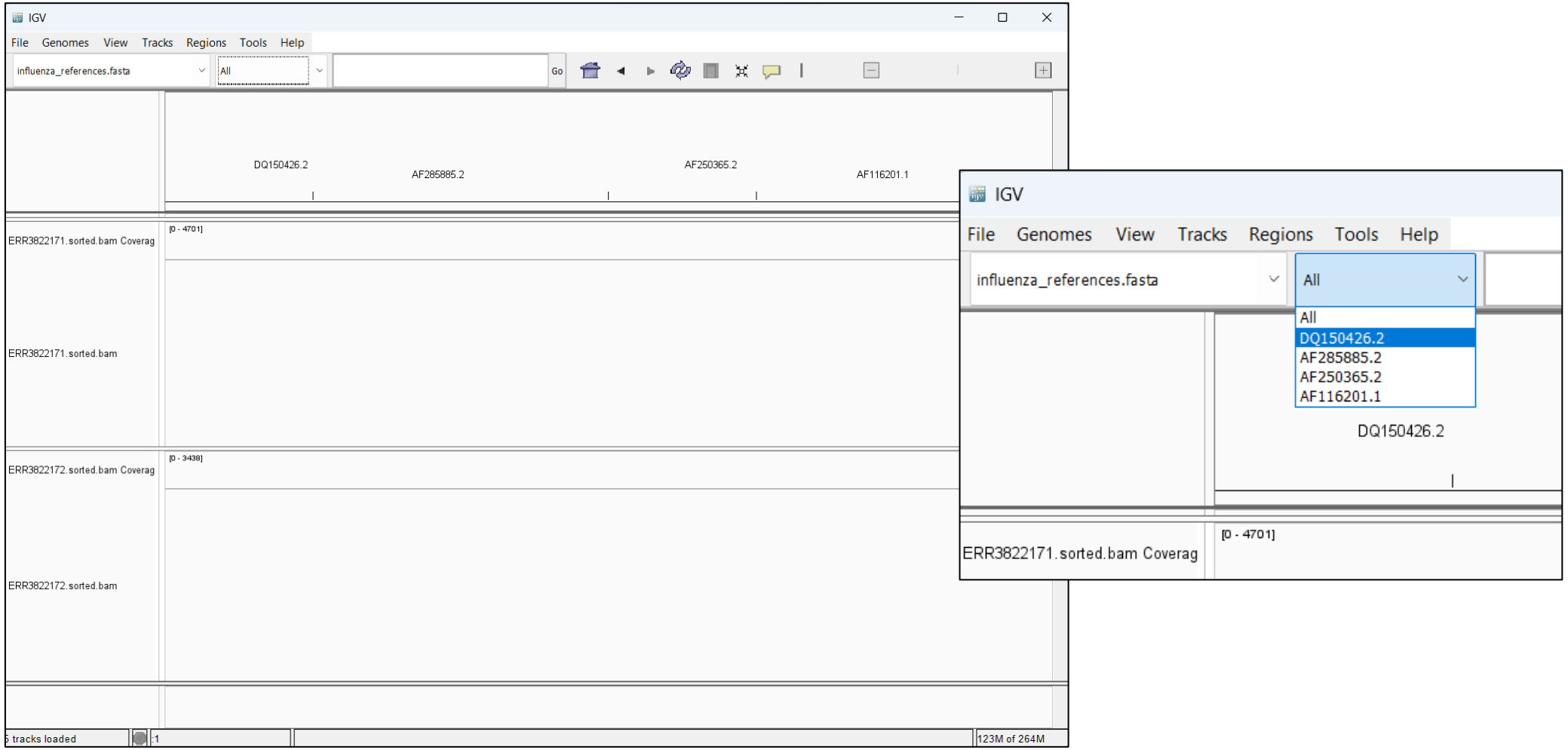
View bam output on IGV





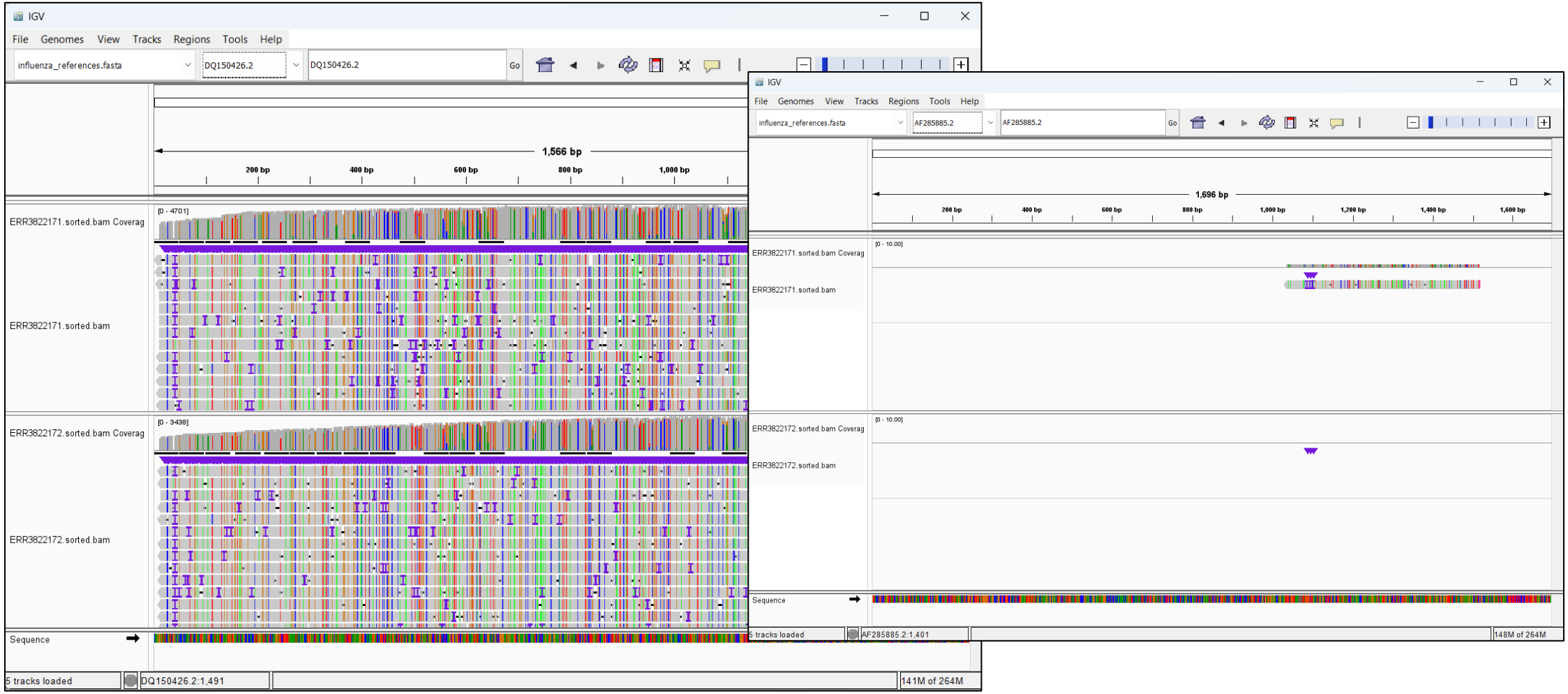
https://software.broadinstitute.org/software/igv/userguide

View bam output on IGV



https://software.broadinstitute.org/software/igv/userguide

View bam output on IGV



https://software.broadinstitute.org/software/igv/userguide

Integrated Genomics Viewer (IGV)

Demonstration

Click to go back to the beginning

LEARNING OBJECTIVES

1. *Configure conda channels and review conda tool installation.*
2. *Describe the methods used for viral genomics data.*
3. *Provide a background on nanopore long-read sequencing.*
4. *Provide an overview of King et al., 2020 paper for workshop practice analysis.*
5. *Provide steps for fastq data analysis.*
6. *Analyze nanopore sequence data.*

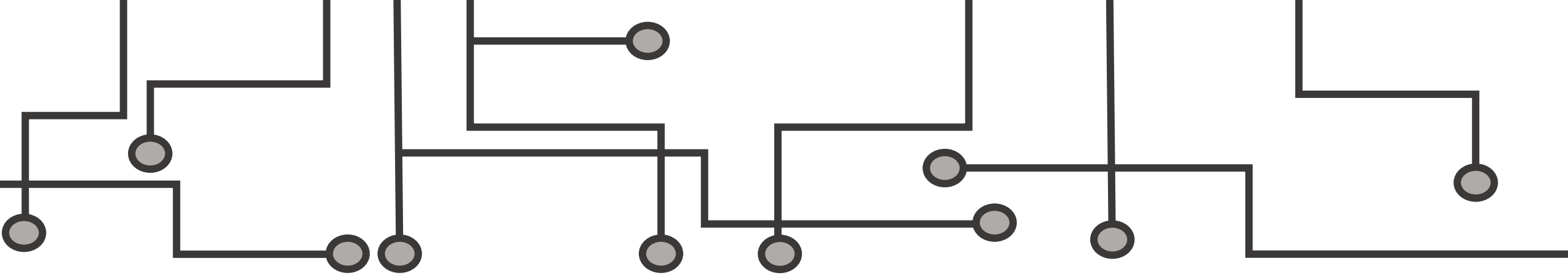
HELPFUL RESOURCES

1. ***Sequence file format:*** ***https://timkahlke.github.io/LongRead_tutorials/APP_FORM.html***
2. ***FastQC:*** ***https://timkahlke.github.io/LongRead_tutorials/QC_F.html***
3. ***NanoPlot:*** *https://github.com/wdecoster/NanoPlot*
4. ***Minimap2:*** ***https://github.com/lh3/minimap2***
5. ***Samtools:*** ***http://www.htslib.org/doc/#manual-pages***

6. ***Samtools guide:*** ***http://quinlanlab.org/tutorials/samtools/samtools.html***

7. ***IGV Guide:*** ***https://software.broadinstitute.org/software/igv/userguide***

8. ***Nanopore bioinformatics guide:*** ***https://www.hadriengourle.com/tutorials/nanopore/***



THANK YOU FOR ATTENDING!

Please make sure to fill out the Exit Survey at

https://forms.gle/WD2WoyUApA75PpWY9

We value your feedback!

More questions? Please email us at

mmid.bioinformatics.workshop@gmail.com or post them to the workshop slack channel