

MEDICAL MICROBIOLOGY AND INFECTIOUS DISEASES BIOINFORMATICS WORKSHOP

Presents

INTRODUCTION TO BACTERIAL GENOMICS: Quality Control and Data Preprocessing

INSTRUCTED BY

Jill Rumore, PhD Candidate

Department of Medical Microbiology and Infectious Diseases

University of Manitoba



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 #[2583403](#)**

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. *Use a publically available dataset to:*
 - I. *Assess general data quality using fastqc*
 - II. *Filter out low quality reads using fastp*
 - III. *Remove host sequence content using Bowtie2 and samtools*
2. *Assess assembled sequence data quality using checkM*

DISCLAIMER

To provide a basic working instruction, all tools will be run with default settings. HOWEVER, careful consideration of analysis parameters in the context of the research question should be taken into account when analyzing your own datasets, as default parameters do not always provide the most optimal result.

GETTING STARTED...

1. Open the terminal and navigate to the `conda_workshop` folder on your desktop

```
cd /mnt/c/Users/Username/Desktop/conda_workshop
```

2. Make a new directory called `Bacterial_Genomics`

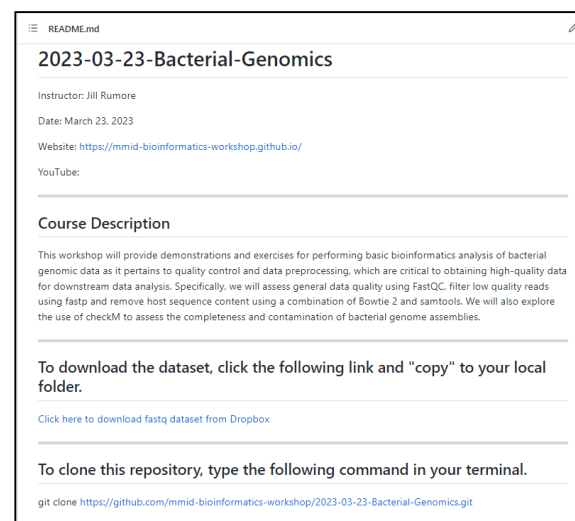
```
mkdir Bacterial_Genomics
```

3. List the contents of the directory to confirm the new directory has been created

```
ls
```

4. Open your internet browser and navigate to the **Bacterial Genomics Repository** hosted on the **MMID Bioinformatics Workshop Github**

```
https://github.com/mmidi-bioinformatics-workshop/2023-03-23-Bacterial-Genomics
```



5. Clone the 2023-03-23-Bacterial-Genomics repository in the Bacterial_Genomics directory

```
git clone https://github.com/mmids-bioinformatics-workshop/2023-03-23-Bacterial-Genomics.git
```

6. Download the fastq dataset to the Bacterial_Genomics directory

```
https://www.dropbox.com/scl/fo/gx9ef004h5l537d58gk13/h?dl=0&rlkey=ouafrzefs7wv9nhbabzhhu9yb
```

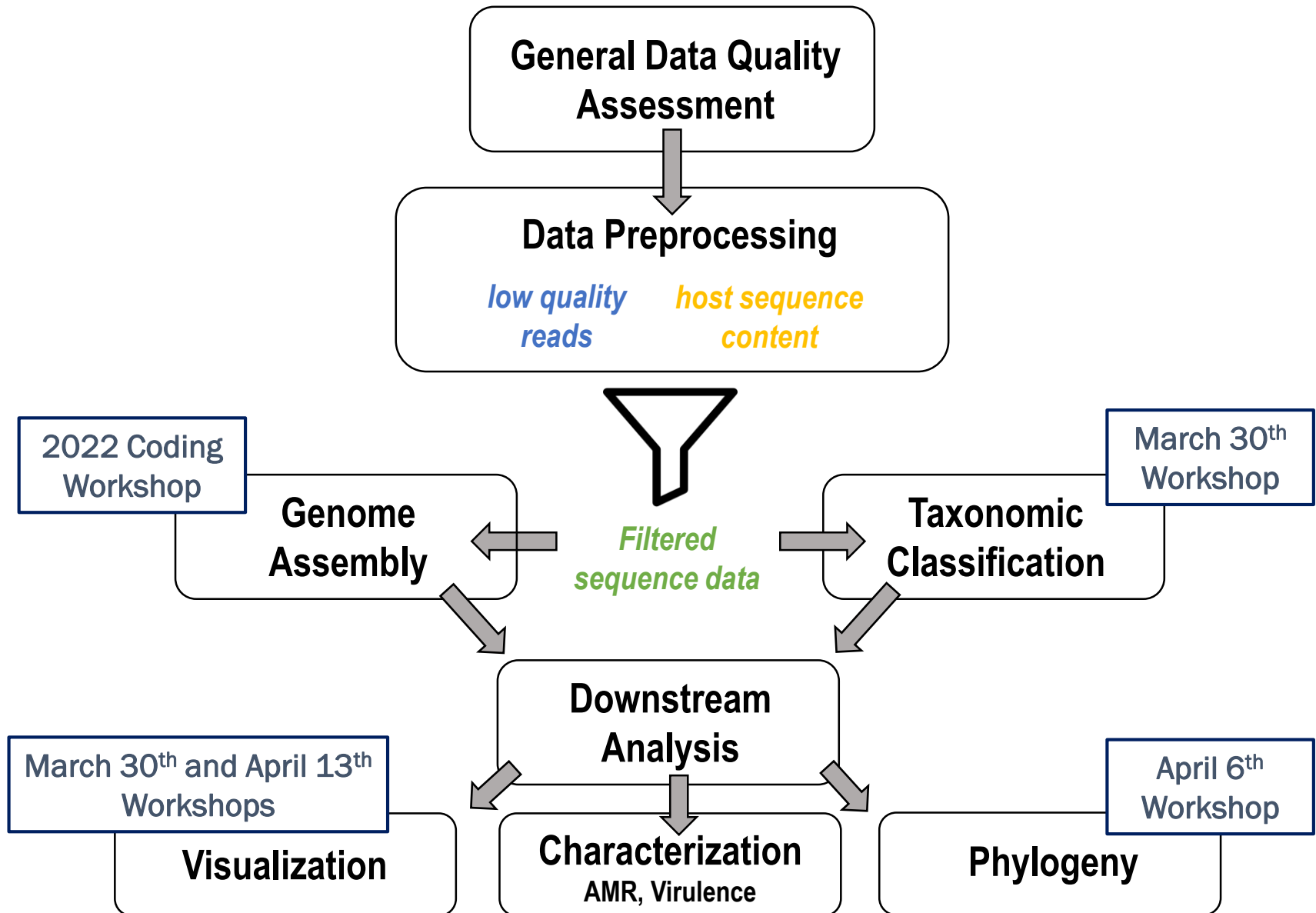
7. Click on the link to download the Bowtie2 index from the MMID Bioinformatics Github repository to the Bacterial_Genomics folder

```
https://www.dropbox.com/s/wrb253faf3i1bfo/Bowtie2Index.tar.gz?dl=0
```

8. Unzip the fastq.zip folder than gunzip the individual fastq.gz files

```
unzip fastq.zip  
gunzip *fastq.gz
```

BASIC WORKFLOW



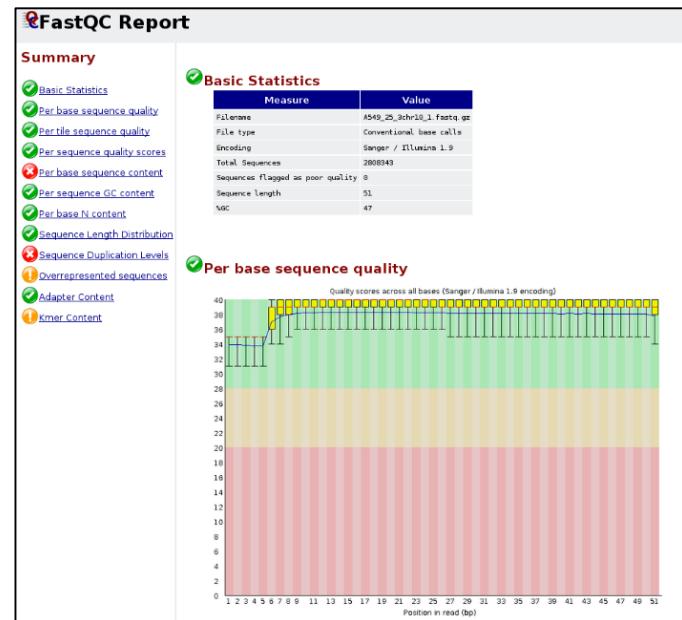
GENERAL DATA QUALITY

Sequence Analysis Viewer (SAV)



https://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html

FastQC

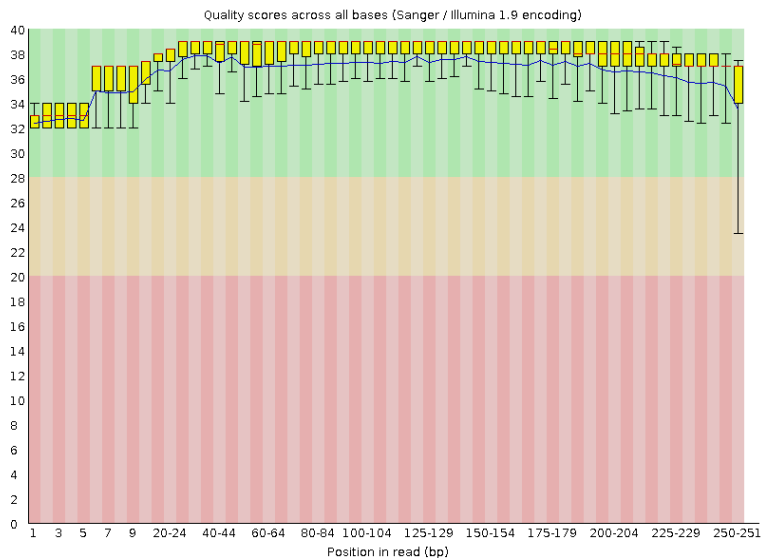


<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

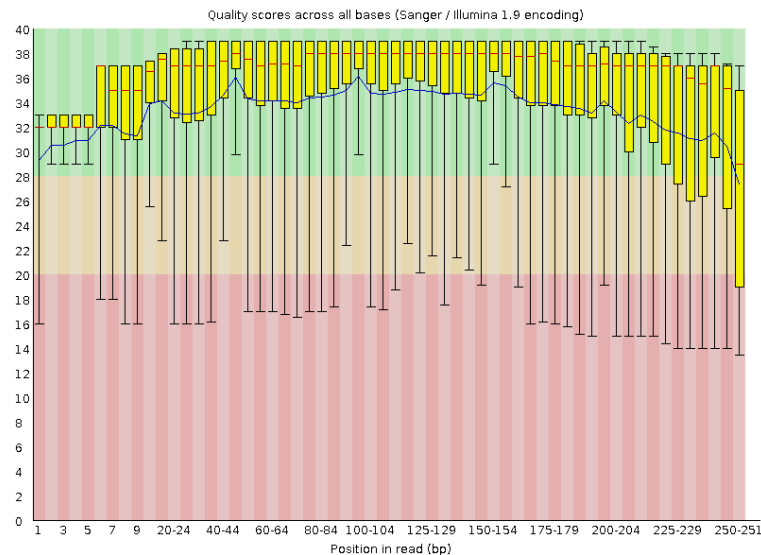
FASTQC

Per Base Sequence Quality

R1



R2

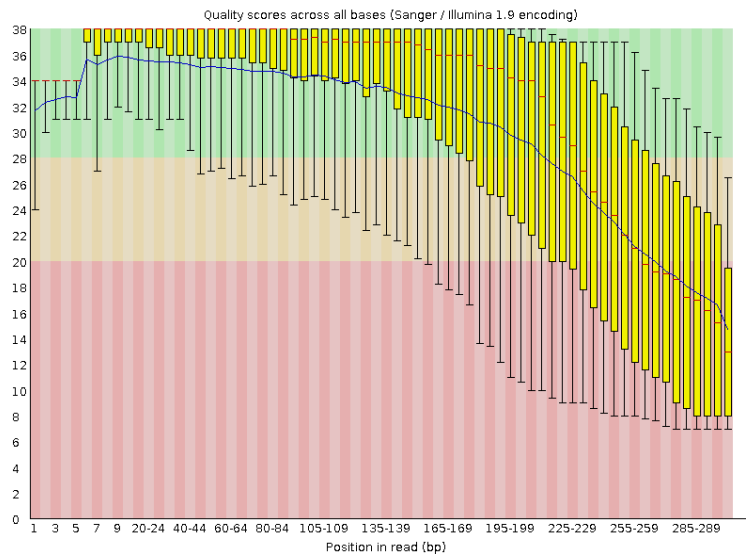


QUALITY SCORE	PROBABILITY OF INCORRECT BASE CALL	INFERRED BASE CALL ACCURACY
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

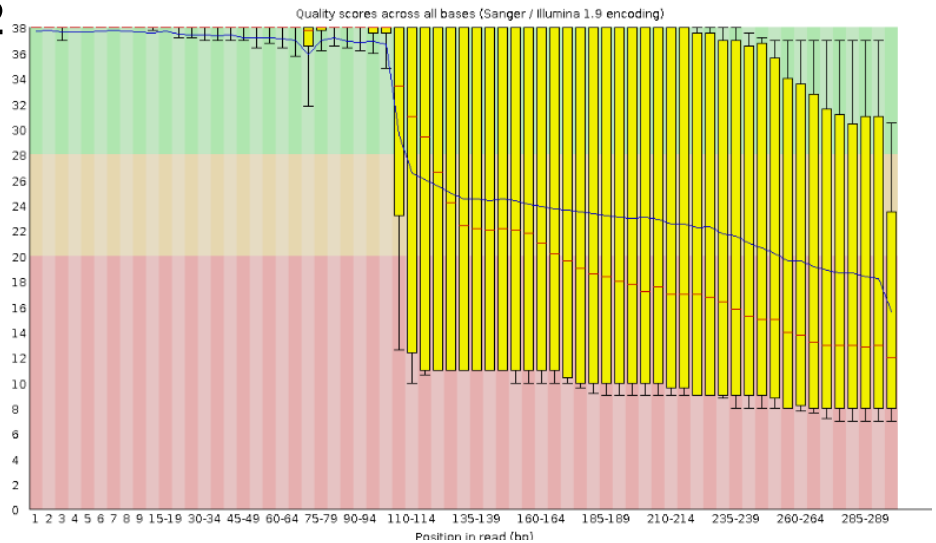
FASTQC

Per Base Sequence Quality

R1



R2



Would this dataset pass the per base sequence quality check?

PUBLICALLY AVAILABLE DATASET

<https://doi.org/10.3389/fpubh.2019.00066>

ORIGINAL RESEARCH article

Front. Public Health, 08 May 2019

Sec. Infectious Diseases: Epidemiology and Prevention

Volume 7 - 2019 | <https://doi.org/10.3389/fpubh.2019.00066>

Clustering of *Vibrio parahaemolyticus* Isolates Using MLST and Whole-Genome Phylogenetics and Protein Motif Fingerprinting

Kelsey J. Jesser^{1*}, Willy Valdivia-Granda², Jessica L. Jones³ and Rachel T. Noble¹

¹ Institute of Marine Sciences, University of North Carolina at Chapel Hill, Morehead City, NC, United States

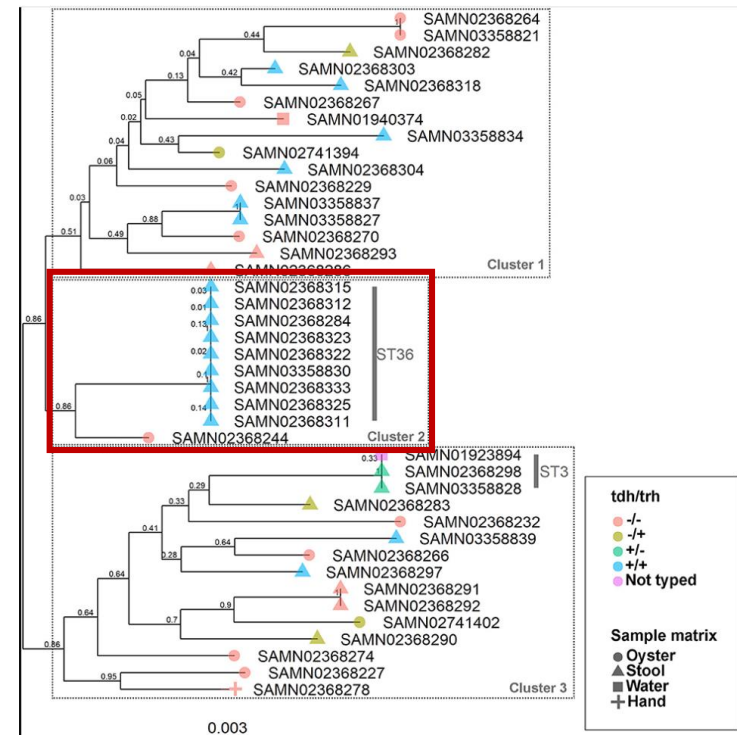
² Orion Integrated Biosciences, New Rochelle, NY, United States

³ Gulf Coast Seafood Laboratory, Division of Seafood Science and Technology, U.S. Food and Drug Administration, Dauphin Island, AL, United States

Uploaded to the MMID Bioinformatics GitHub Repository

<https://github.com/mmidi-bioinformatics-workshop>

Data was sequenced using the Illumina HiSeq 2000



DEMONSTRATION

STEP BY STEP GUIDE

1. Make a new directory called fastq in the Bacterial_Genomics directory

```
mkdir fastq
```

2. Move fastq files into the fastq directory

```
mv *.fastq ./fastq
```

3. Verify the fastq files have been moved

```
cd fastq  
ls
```

4. Activate conda environment that contains the FastQC package

```
conda activate conda_workshop
```

5. Return to the Bacterial_Genomics directory

```
cd ..
```

6. Make a new directory called fastqc_reports

```
mkdir fastqc_reports
```

7. Run fastqc

```
fastqc ./fastq/*.fastq -o ./fastqc_reports/
```

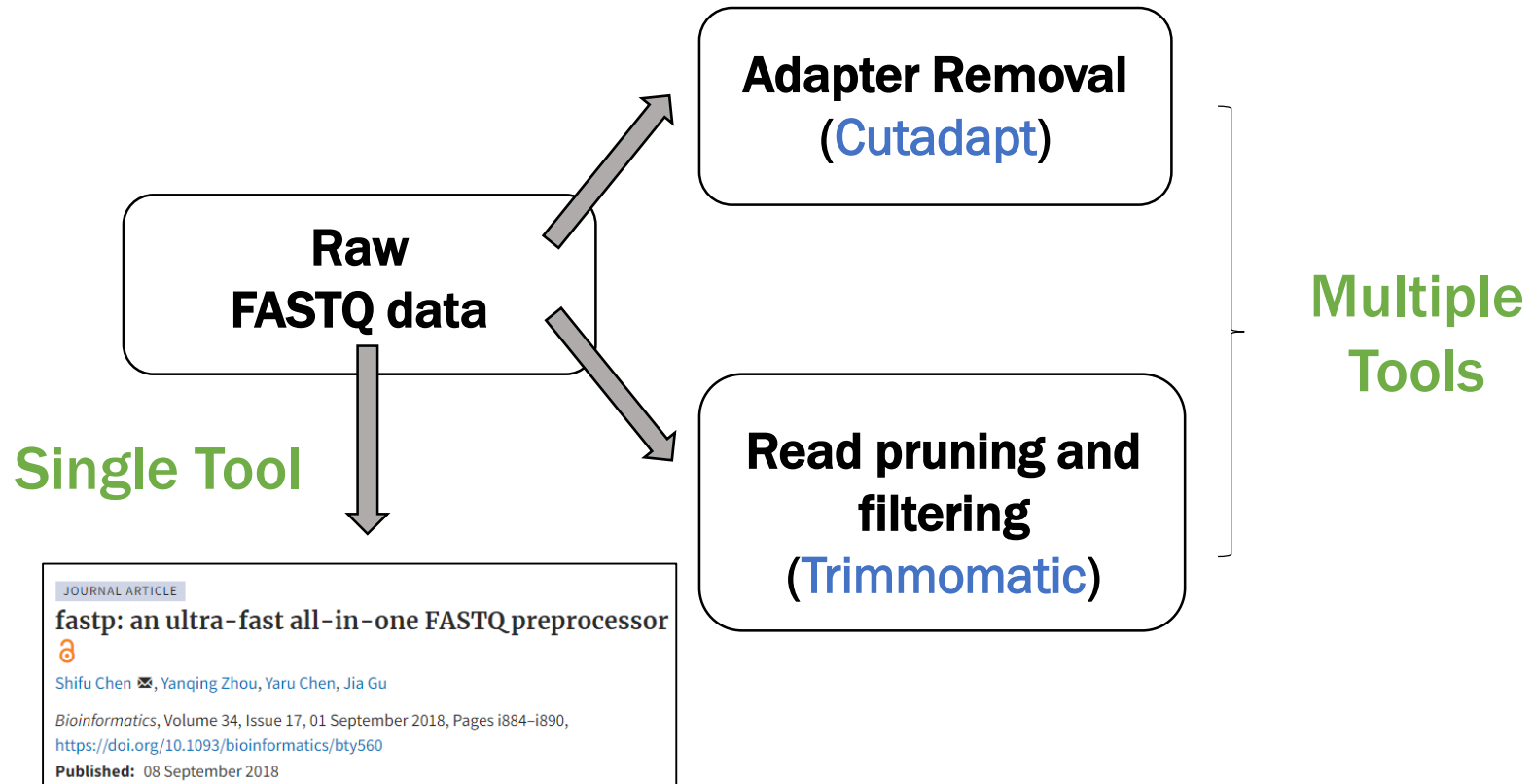
```
sbatch -c 1 --mem 2G -p NMLResearch --wrap="fastqc ./fastq/*.fastq  
-o ./fastqc_reports" (**Waffles Users Only**)
```

****IF WORKING ON THE CLUSTER (AKA WAFFLES) PLEASE USE THE SLURM WORKLOAD MANAGER WHEN SUBMITTING JOBS****

Detailed instructions can be found here:

https://github.com/MMID-coding-workshop/2022-01-19-Introduction-to-CONDA/blob/main/MMID_Coding_Workshop-IntroToConda_2022-01-19-Supplemental.pdf

LOW QUALITY READ FILTERING



fastp REPORT

fastp report	
Summary	
General	
fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (100 cycles + 100 cycles)
mean length before filtering:	100bp, 100bp
mean length after filtering:	98bp, 98bp
duplication rate:	0.793924%
Insert size peak:	119
Before filtering	
total reads:	7.322614 M
total bases:	732.261400 M
Q20 bases:	699.351353 M (95.505697%)
Q30 bases:	666.183279 M (90.976157%)
GC content:	45.285202%
After filtering	
total reads:	6.890640 M
total bases:	679.597300 M
Q20 bases:	667.849881 M (98.271415%)
Q30 bases:	639.829806 M (94.148374%)
GC content:	45.087958%
Filtering result	
reads passed filters:	6.890640 M (94.100822%)
reads with low quality:	431.518000 K (5.892950%)
reads with too many N:	456 (0.006227%)
reads too short:	0 (0.000000%)

For more information on interpreting fastp output

<https://github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data>

DEMONSTRATION

STEP BY STEP GUIDE

1. Activate conda environment that contains the fastp package

```
conda activate conda_workshop
```

2. Make a new directory called fastp in the Bacterial_Genomics directory

```
mkdir fastp
```

3. Run fastp

```
fastp -i ./fastq/SAMN02368311_R1.fastq -I  
./fastq/SAMN02368311_R2.fastq -o ./fastp/SAMN02368311-fp_R1.fastq  
-O ./fastp/SAMN02368311-fp_R2.fastq -h ./fastp/SAMN02368311.html  
-j ./fastp/SAMN02368311.json
```

HOST SEQUENCE FILTERING

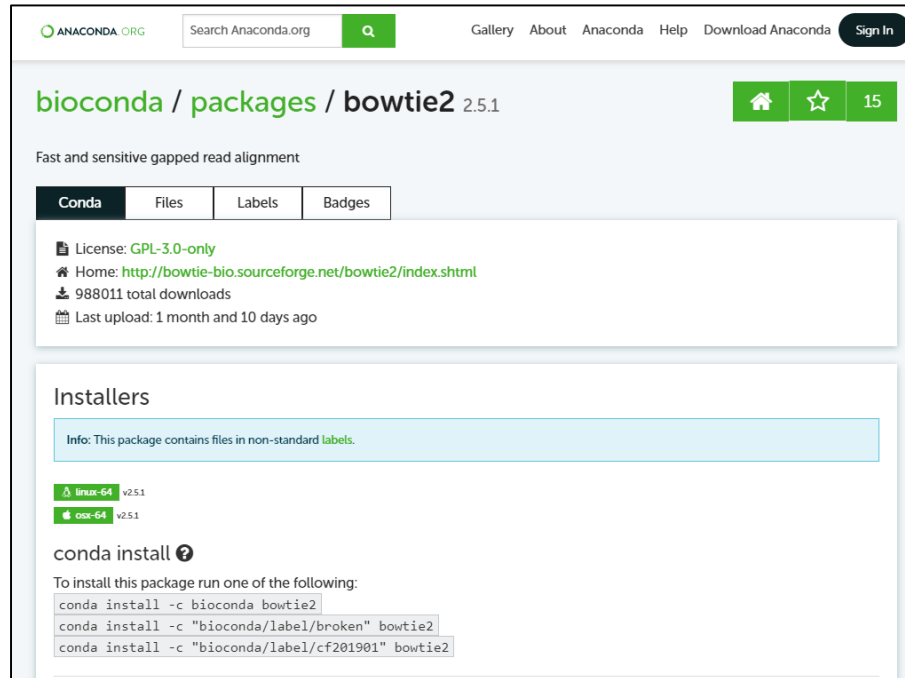
ATGCCAGCATTGCAGTTGCAGCAGCA
AGGTTTCCAGAGCAGTTGCAACAGCA
TCCGATCCAGAGCAGTTAATCCCAGCA
CCCGATCCAGAGCAGGTAATTCCAGCA



AGGTTTCCAGAGCAGTTGCAACAGCA
TCCGATCCAGAGCAGTTAATCCCAGCA
CCCGATCCAGAGCAGGTAATTCCAGCA



INSTALL TOOLS

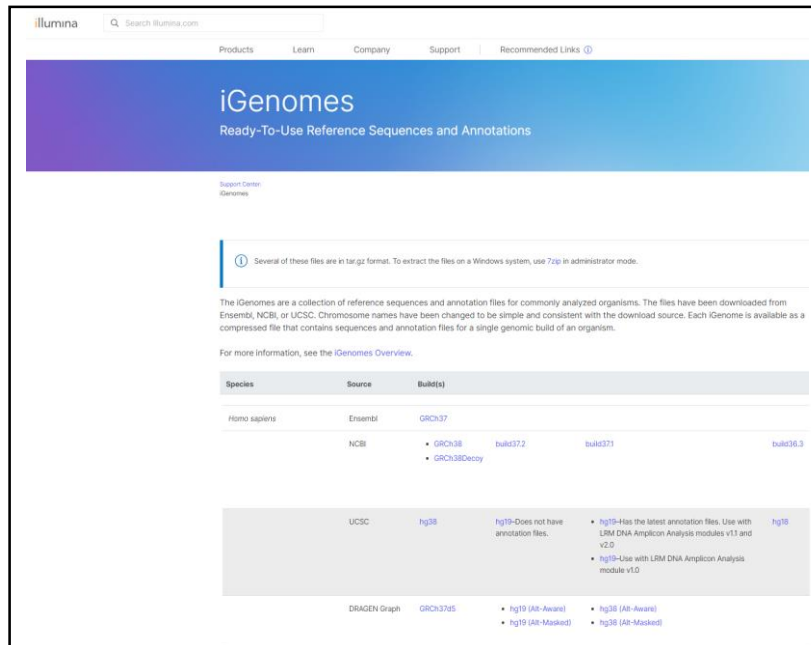


1. Install the bowtie2 package in the conda_workshop environment

```
conda install -y -c bioconda bowtie2
```

BOWTIE2 INDICES

https://support.illumina.com/sequencing/sequencing_software/igenome.html



The screenshot shows the Illumina iGenomes website. The header includes the Illumina logo and navigation links: Products, Learn, Company, Support, and Recommended Links. The main heading is "iGenomes" with the subtitle "Ready-To-Use Reference Sequences and Annotations". Below this, there is a note about file formats: "Several of these files are in tar.gz format. To extract the files on a Windows system, use 7zip in administrator mode." A paragraph explains that iGenomes are a collection of reference sequences and annotation files for commonly analyzed organisms, downloaded from Ensembl, NCBI, or UCSC. A link to the "iGenomes Overview" is provided. The main content is a table with columns: Species, Source, and Build(s).

Species	Source	Build(s)
Homo sapiens	Ensembl	GRCh37
	NCBI	GRCh38 GRCh38Decoy
	UCSC	hg38
	DRAGEN Graph	GRCh37v5

<https://benlangmead.github.io/aws-indexes/bowtie>

Bowtie 2 indexes

Bowtie and Bowtie 2 are read aligners for sequencing reads. Bowtie specializes in short reads, generally about 50bp or shorter. Bowtie 2 specializes in longer reads, up to around hundreds of base pairs. HTTPS URLs allow you to download the files from your web browser or using command-line tools like `wget` or `curl`. The S3 URLs can be used with AWS tools, including the [AWS console](#) and [AWS command-line interface](#).

In the past, Bowtie 1 & 2 had incompatible genome indexes. This changed in July 2019 when Bowtie v1.2.3 gained the ability to use Bowtie 2 formatted genome indexes (ending in `.bt2`). We list only Bowtie 2-format `.bt2` index files here.

You can download all the files for a given assembly as a single `zip` file, or as 6 separate `.bt2` files. For example, if you only need the forward version of the genome index (e.g. for exact matching only), you can download the files individually and omit the `.rev.1.bt2` and `.rev.2.bt2` files. Downloading already-decompressed index files might also be quicker for applications running in the AWS cloud.

Species/Build	Source	HTTPS URLs	S3 URLs
Human / GRCh38 no-alt analysis set ¹	NCBI	full zip, .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, .rev.2.bt2	full zip, .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, .rev.2.bt2
Human / GRCh38 no-alt +decoy set ¹	NCBI	full zip, .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, .rev.2.bt2	full zip, .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, .rev.2.bt2
Human / GRCh38 + major SNVs	NCBI+1KG ²	full zip, .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, .rev.2.bt2	full zip, .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, .rev.2.bt2

DEMONSTRATION

STEP BY STEP GUIDE

1. Make a new directory in Bacterial_Genomics directory called host_filtered

```
mkdir host_filtered
```

2. Decompress the Bowtie2Index.tar.gz file

```
tar -xvzf Bowtie2Index.tar.gz
```

3. Filter host reads using the iGenomes *Homo sapiens* bowtie2 index

```
bowtie2 -x ./Bowtie2Index/genome -1 ./fastp/SAMN02368311-fp_R1.fastq  
-2 ./fastp/SAMN02368311-fp_R2.fastq -S  
./host_filtered/SAMN02368311.sam
```

4. Move into the host_filtered directory and convert sam to bam

```
cd host_filtered  
samtools view -bS SAMN02368311.sam > SAMN02368311.bam
```


5. Make a new directory called unmapped

```
mkdir unmapped
```

6. Extract unmapped reads for both pairs

```
samtools view -b -f 12 -F 256 SAMN02368311.bam >  
./unmapped/SAMN02368311_unmapped.bam
```

7. Move into the unmapped directory and make a new directory called sorted

```
cd unmapped/  
mkdir sorted
```

8. Sort the bam file

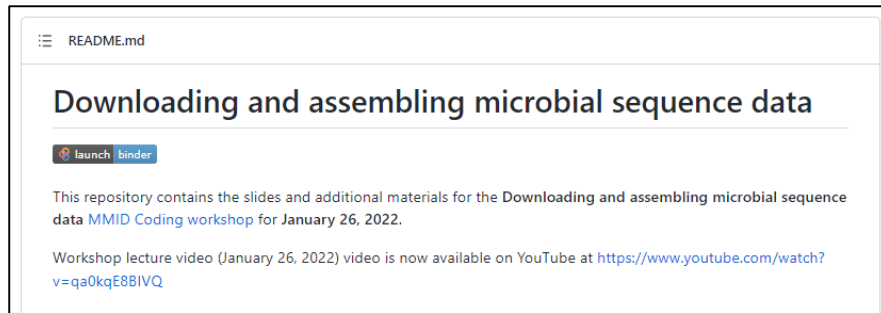
```
samtools sort -n -o ./sorted/SAMN02368311_sorted.bam --output-fmt  
BAM SAMN02368311_unmapped.bam
```

9. Move into the sorted directory and convert BAM file to fastq

```
cd sorted/  
samtools fastq -@ 2 -1 SAMN02368311-HR_R1.fastq -2 SAMN02368311-  
HR_R2.fastq SAMN02368311_sorted.bam
```

BACTERIAL GENOME ASSEMBLY

<https://github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data>



README.md

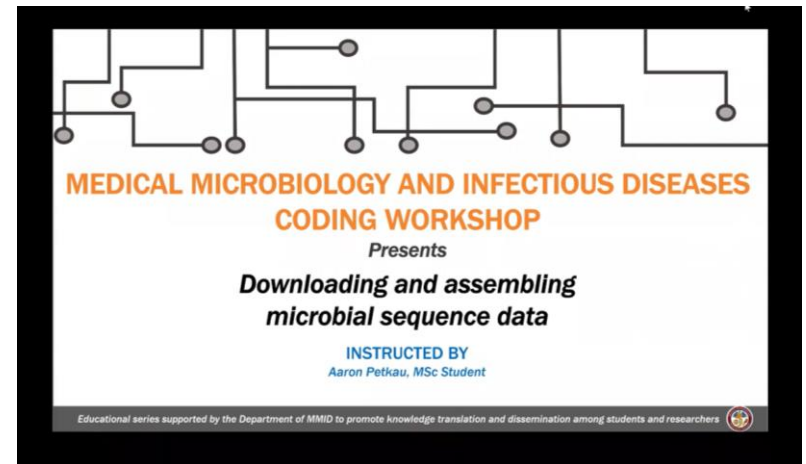
Downloading and assembling microbial sequence data

[launch](#) [binder](#)

This repository contains the slides and additional materials for the Downloading and assembling microbial sequence data MMID Coding workshop for January 26, 2022.

Workshop lecture video (January 26, 2022) video is now available on YouTube at <https://www.youtube.com/watch?v=qa0kqE8BIVQ>

<https://www.youtube.com/watch?v=qa0kqE8BIVQ>



**MEDICAL MICROBIOLOGY AND INFECTIOUS DISEASES
CODING WORKSHOP**

Presents

**Downloading and assembling
microbial sequence data**

INSTRUCTED BY
Aaron Petkau, MSc Student

Educational series supported by the Department of MMID to promote knowledge translation and dissemination among students and researchers

ASSEMBLY QUALITY - checkM



<https://github.com/Ecogenomics/CheckM/wiki>

► Genome Res. 2015 Jul;25(7):1043-55. doi: 10.1101/gr.186072.114. Epub 2015 May 14.

CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H Parks¹, Michael Imelfort¹, Connor T Skennerton¹, Philip Hugenholtz², Gene W Tyson³

Affiliations + expand

PMID: 25977477 PMCID: PMC4484387 DOI: 10.1101/gr.186072.114

[Free PMC article](#)

Completeness

Estimated completeness of genomes as determined from presence/absence of marker genes

Contamination

Estimated contamination of genome as determined by the presence of multi-copy marker genes

Workflows

Lineage-specific – use this workflow to assess genomes from different taxonomic groups
Taxonomy-specific – use this workflow to assess genomes from the same taxonomic group

DEMONSTRATION

STEP BY STEP GUIDE

1. Make a new directory in Bacterial_Genomics directory called checkM

```
mkdir checkM
```

2. Navigate to the downloaded assemblies folder and decompress the fasta files

```
cd 2023-03-23-Bacterial-Genomics  
cd assemblies_skesa  
gunzip *.gz
```

3. Return to the Bacterial_Genomics folder

```
cd ../../
```

4. Run checkM taxonomy_wf on assembled data

```
checkm taxonomy_wf genus Vibrio ./2023-03-23-Bacterial-Genomics/assemblies_skesa/ ./checkM/ -t 2 -x fasta
```

5. Run qa workflow

```
checkm qa -o 2 -f ./checkM/checkM_quality.tsv --tab_table  
./checkM/Vibrio.ms ./checkM/
```

checkM RESULTS

<https://github.com/Ecogenomics/CheckM/wiki/Reported-Statistics>

Reported Statistics

Donovan Parks edited this page on May 2, 2019 · 5 revisions

qa

- bin id:** unique identifier of genome bin (derived from input fasta file)
- marker lineage:** indicates the taxonomic rank of the lineage-specific marker set used to estimate genome completeness, contamination, and strain heterogeneity. More detailed information about the placement of a genome within the reference genome tree can be obtained with the `tree_qa` command. The UID indicates the branch within the reference tree used to infer the marker set applied to estimate the bins quality.
- # genomes:** number of reference genomes used to infer the lineage-specific marker set
- markers:** number of marker genes within the inferred lineage-specific marker set
- marker sets:** number of co-located marker sets within the inferred lineage-specific marker set
- 0-5+:** number of times each marker gene is identified
- completeness:** estimated completeness of genome as determined from the presence/absence of marker genes and the expected colocalization of these genes (see Methods in the [PeerJ preprint](#) for details)
- contamination:** estimated contamination of genome as determined by the presence of multi-copy marker genes and the expected colocalization of these genes (see Methods in the [PeerJ preprint](#) for details)
- strain heterogeneity:** estimated strain heterogeneity as determined from the number of multi-copy marker pairs which exceed a specified amino acid identity threshold (default = 90%). High strain heterogeneity suggests the majority of reported contamination is from one or more closely related organisms (i.e. potentially the same species), while low strain heterogeneity suggests the majority of contamination is from more phylogenetically diverse sources (see Methods in the [CheckM manuscript](#) for more details).

Pages 12

- Home
- Bin Exploration and Modification
- Bugs and Feature Requests
- Genome Quality Commands
- Installation
- Introduction
- Overview
- Plots
- Quick Start
- Reported Statistics

qa

A	B	C	D	E	F	G	H	I
Bin Id	Marker lineage	# genomes	# markers	# marker sets	Completeness	Contamination	Strain heterogeneity	Genome size (bp)
SAMN02368311	Vibrio (5)	70	1084	381	100	0.13	0	5079511
SAMN02368315	Vibrio (5)	70	1084	381	100	0.13	0	5079810

HELPFUL RESOURCES

How to change your password for sudo

<https://askubuntu.com/questions/931940/unable-to-change-the-root-password-in-windows-10-wsl>

Interpreting Quality Scores

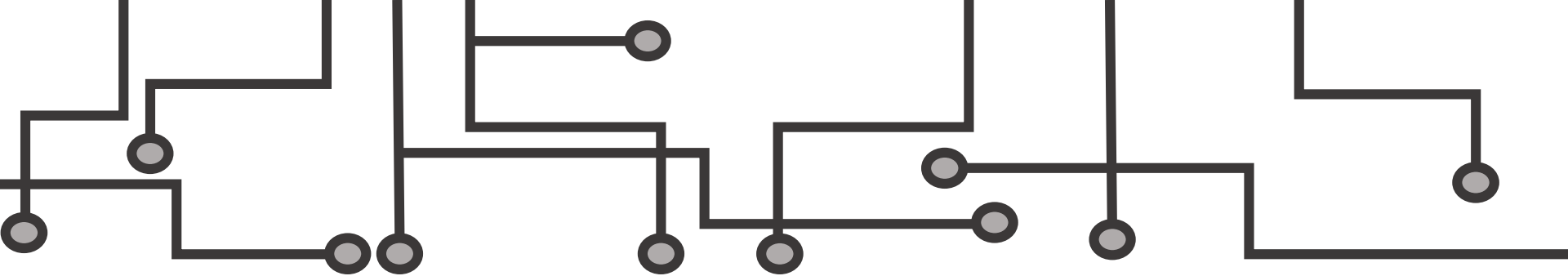
<https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html>

SAM Flags

<https://broadinstitute.github.io/picard/explain-flags.html>

Bowtie2 documentation

<https://github.com/BenLangmead/bowtie2>



THANK YOU FOR ATTENDING!

***Please make sure to fill out the [Exit Survey](https://docs.google.com/forms/d/e/1FAIpQLScNW1XUrn16psbmW8yP3JTxlhWnVxp8n7ThwG3pBqdYbEXyQ/viewform?usp=sharing) at
<https://docs.google.com/forms/d/e/1FAIpQLScNW1XUrn16psbmW8yP3JTxlhWnVxp8n7ThwG3pBqdYbEXyQ/viewform?usp=sharing>
We value your feedback!***

***More questions? Please email us at
mmid.bioinformatics.workshop@gmail.com or post them to the workshop [slack channel](#)***

