# RNAseq Deduplication with UMIs

**Radha Ganesh & David Ho**

**Under the Guidance of Dr. Barbara Novak**

BPS

Biological & Physical Sciences

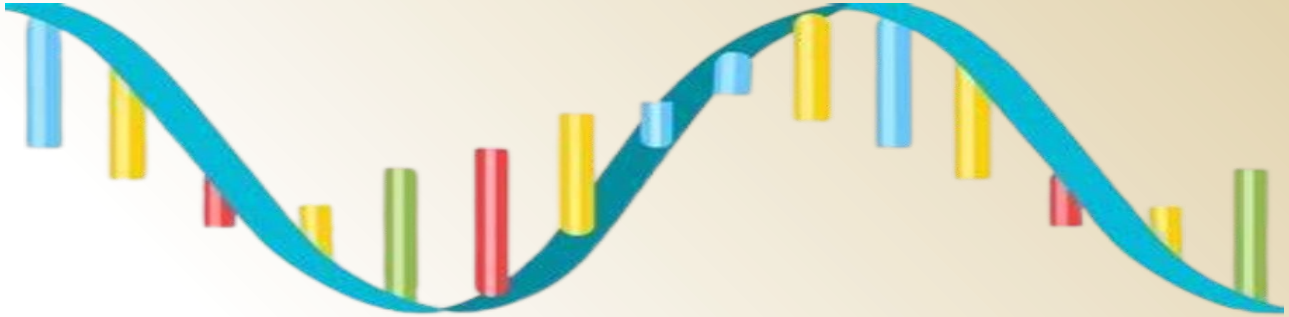
National Aeronautics and Space Administration

NASA

# Table of Contents

# Background

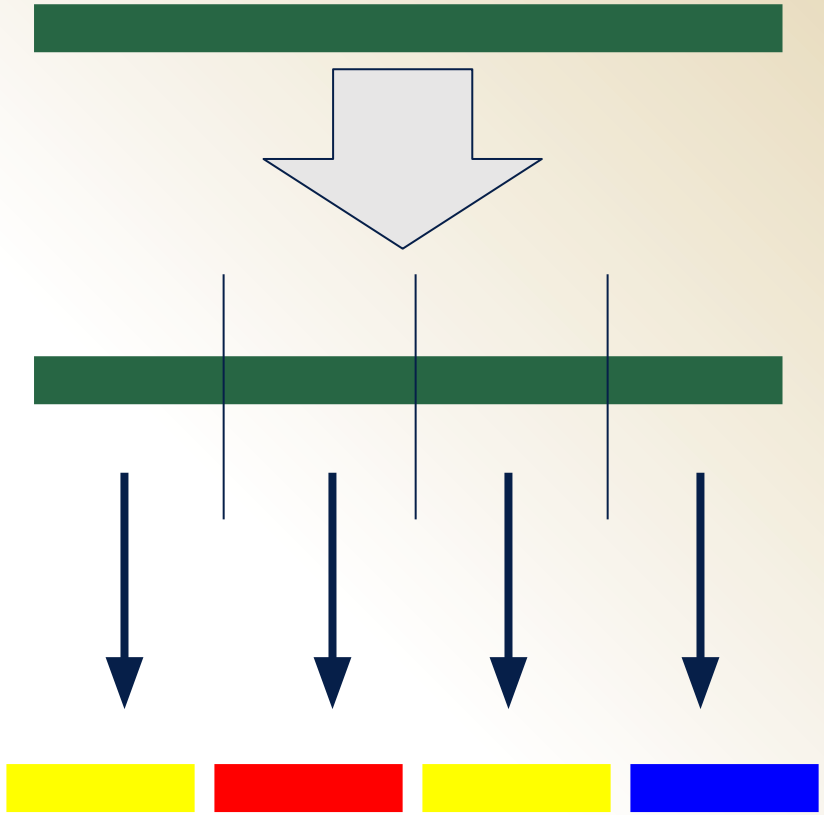
# What is RNAseq?



- Uses next-generation sequencing (NGS) to reveal the presence and quantify RNA in a biological sample.
- Strictly speaking this could be any type of RNA (mRNA, lncRNA, miRNA) from any type of biological sample.
- One goal is to profile gene expression by identifying genes that are differentially expressed (DE) between two or more biological conditions

# Duplicates

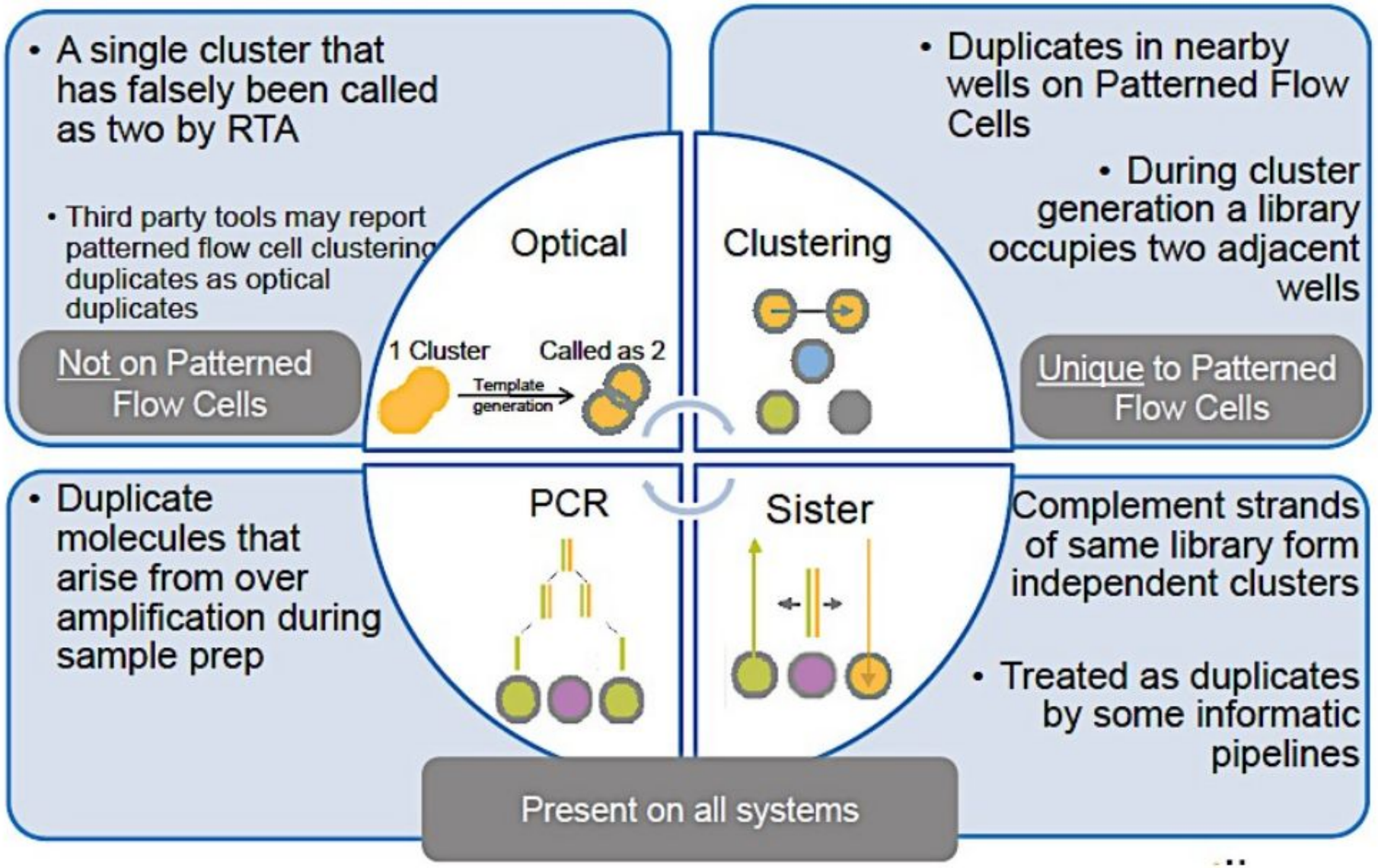
- When you sequence cDNA, you get multiple short sequences, referred to as reads, from each sample
- A duplicate is when you get more than one read with the exact same sequence.



# Biological vs. Technical Duplicates

- Biological duplicates are true duplicates that came from multiple copies of a transcript from the original sample.
- Technical duplicates came from a technical aspect of either library preparation or sequencing.
- When deduplicating, we want to remove the technical duplicates

# Types of Technical Duplicates



# The Use of Unique Molecular Identifiers (UMIs)

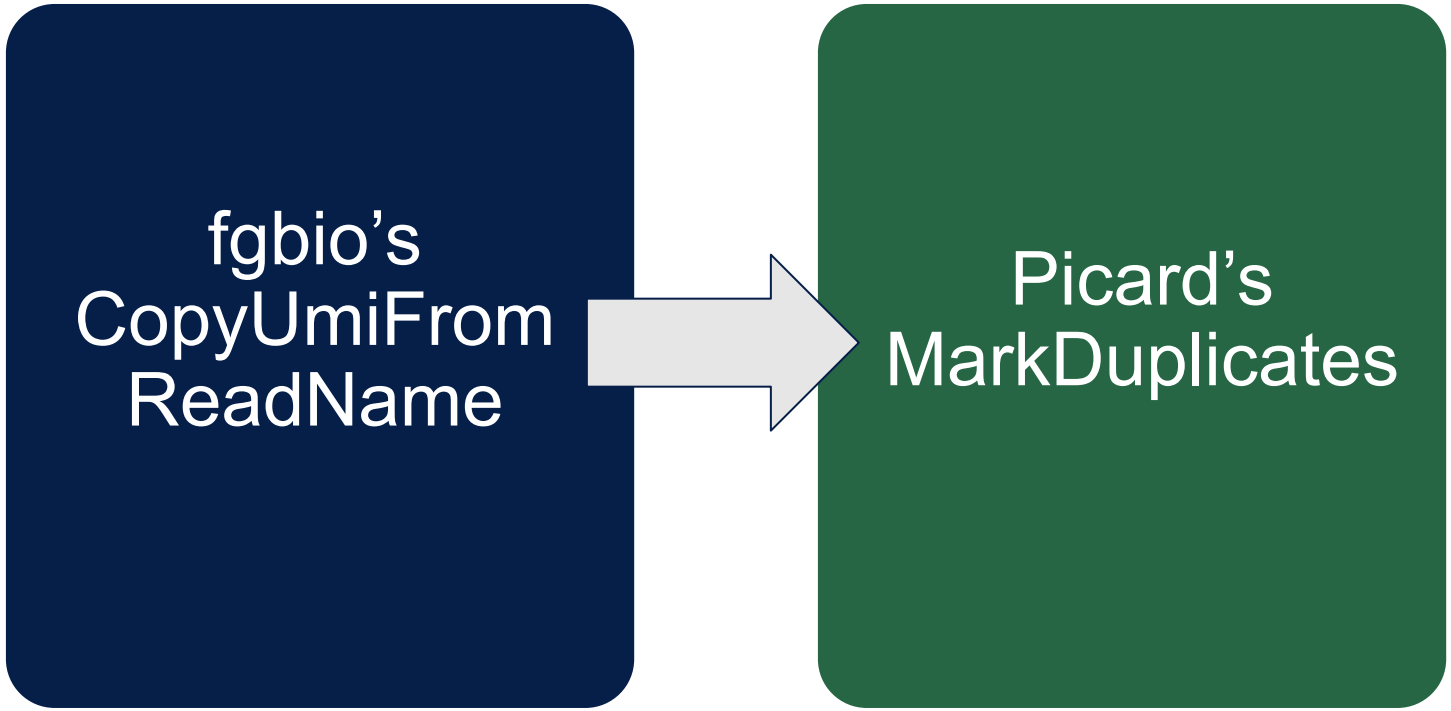
- UMIs help labs keep track of molecules and remove errors during amplification and sequencing
- Before the PCR step, every molecule in a sample is uniquely labeled with a UMI, which is typically a random sequence of oligonucleotides.
- When we see two identical tags on two identical sequences, we can assume that they are from the same original molecule and are therefore technical duplicates. A finding of two different tags on the same sequence means the sequences came from two different original molecules and are therefore biological duplicates.

# Problems with Removing Technical Duplicates Using UMIs

- As of now, there exists no single approach to efficiently remove technical duplicates using UMIs.
- Many programs that exist are very time and memory intensive
- Through this project, we evaluated the efficiency of different deduplication methods and attempted to determine the effect of removing technical duplicates on differential gene expression.
- The tools we used for deduplication were Picard's MarkDuplicates & UMItools' dedup.

# Methods

# Tools Used for Picard's MarkDuplicates



# What is fgbio?

- Java based toolset created by Fulcrum Genomics
- Allow users to manipulate read-level data (SAM, BAM, FASTQ)
  - Perfect tool to extract UMIs from a readname

# What is Picard?

- Java based toolset created by Broad Institute
- Has tools to collect statistics on a given SAM or BAM file
- Can mark any duplicates found in a SAM or BAM file

# How Do They Work Together?

- OSD-511 dataset was preprocessed by UMItools
  - UMIs are placed at the ends of the readnames
- Problematic for Picard
  - The tool will not facilitate duplicate marking using molecular barcodes unless the UMIs are in a SAM tag.
- Therefore, a preprocessing step is required
  - Parse through all of the readnames to strip the UMIs and make them into tags

# fgbio's CopyUmiFromReadName

As the name suggests, this tool will parse through every readname to copy the UMI at the end to create an RX tag.

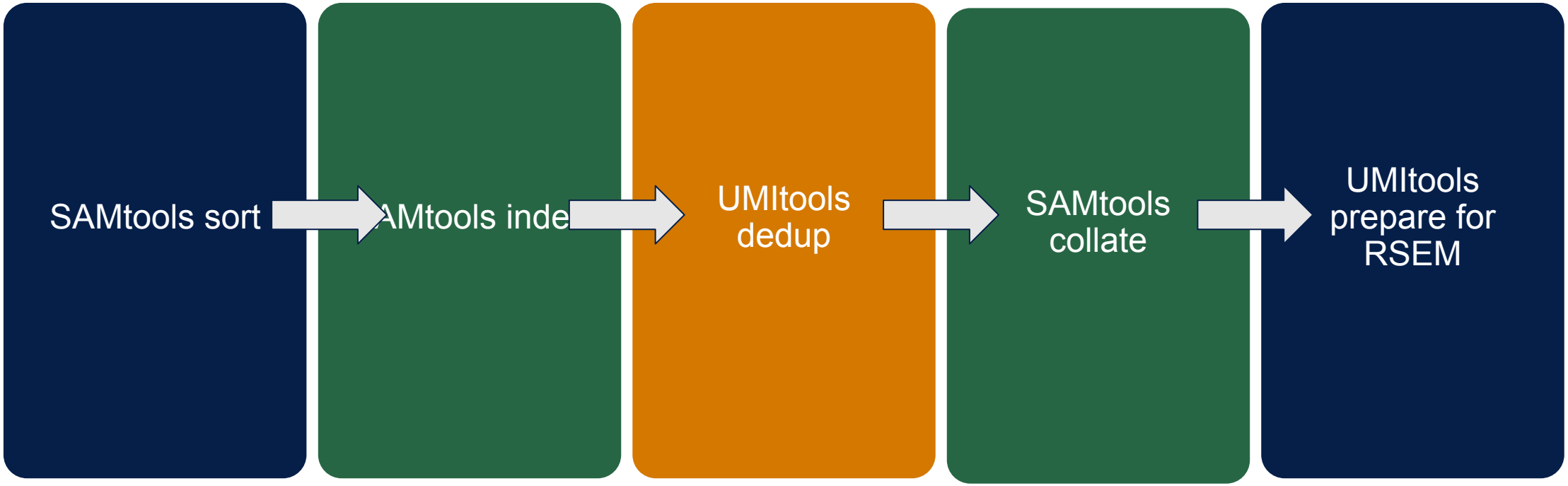
```
fgbio CopyUmiFromReadName \
--input=input.bam \
--output=output.bam \
--remove-umi=true \
--field-delimiter=_
```

# Picard's MarkDuplicates

- Picard's MarkDuplicates goes through 3 main steps:
  - Loads a sorted BAM file and it iterates through each read in order
  - Goes through each input list and builds a duplicate set
  - Marks any duplicate reads and write non-duplicate reads to a resulting output BAM file.

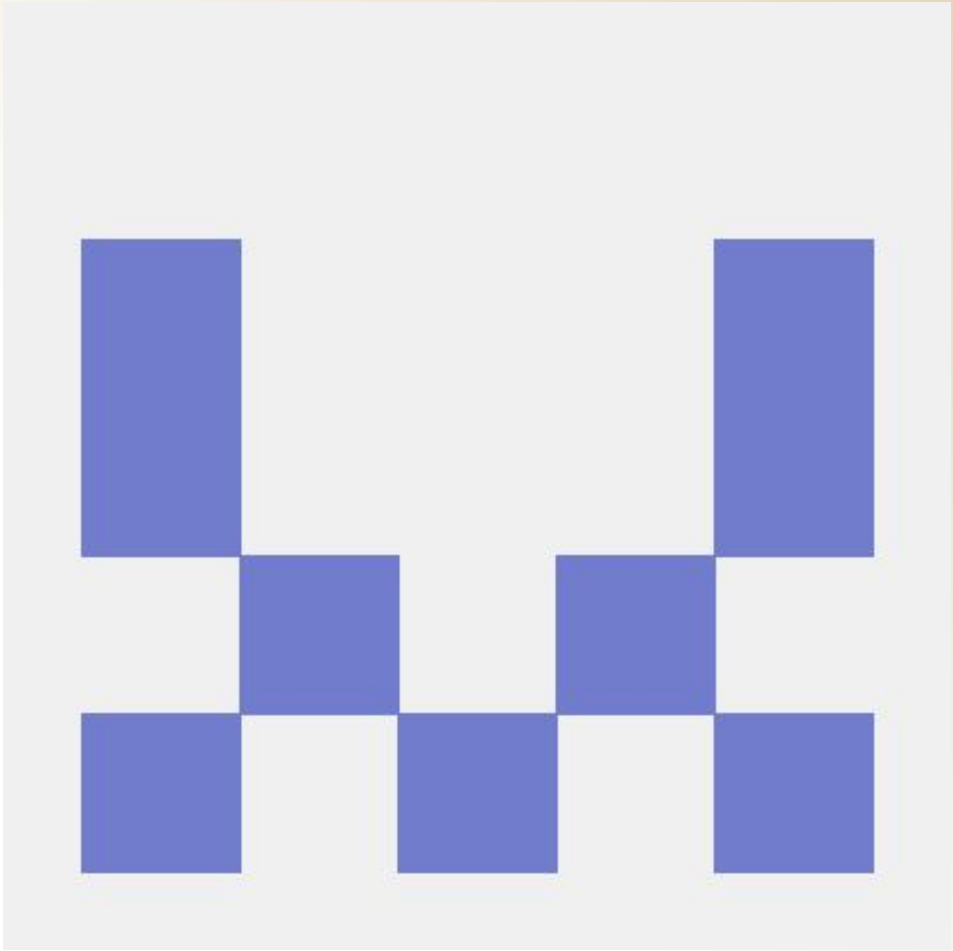
```
picard MarkDuplicates \
I=input.bam \
O=output.bam \
M=metrics_data.txt \
REMOVE_DUPLICATES=true \
OPTICAL_DUPLICATE_PIXEL_DISTANCE=2500 \
ASSUME_SORT_ORDER=queryname \
BARCODE_TAG=RX
```

# Tools Used for UMItools dedup



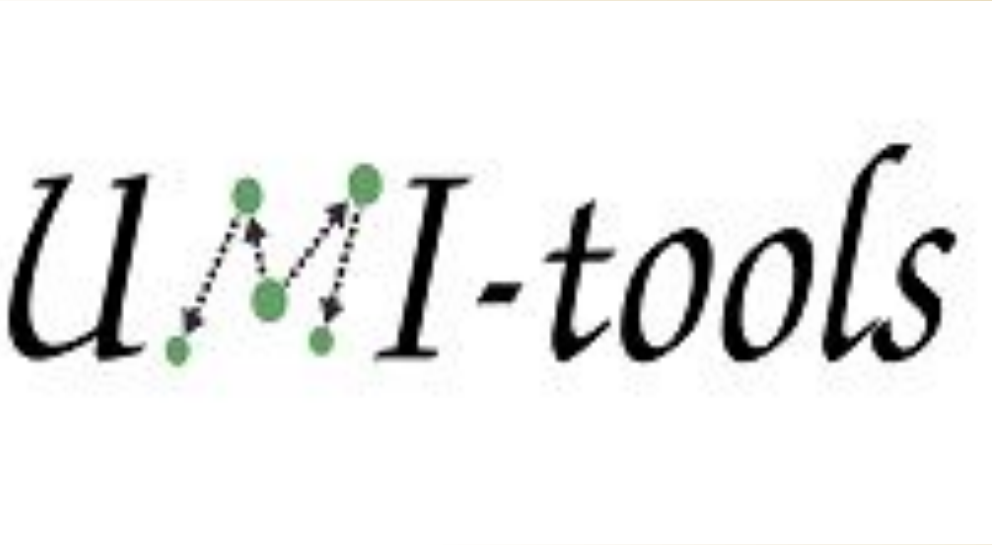
# What is SAMtools?

- A software package that allow users to manipulate alignment files in SAM, BAM, and CRAM formats
  - Good for sorting and indexing alignment files



# What is UMItools?

- A toolset designed for working with UMIs and barcodes
- Has tools to remove PCR duplicates using UMIs



# How Do They Work Together?

- UMItools deduplication requires a sorted (by position) and indexed alignment file as input
  - Files need to be preprocessed using SAMtools' sort and index commands prior to deduplication
- UMItools prepare for rsem requires a deduplicated file sorted (by name) as input
  - Files need to be preprocessed using SAMtools' collate command

# SAMtools' sort and index

- sort command: sorts the input file by position
- index command: indexes the sorted file
  - allows for fast random access

```
samtools sort \
input.bam \
-o output_sorted.bam

samtools index \
input_sorted.bam
```

# UMItools dedup

- Removes duplicate reads based on mapping coordinates and the UMI attached to it

```
umi_tools dedup \
-I input_sorted.bam \
--paired \
-S output_deduplicated.bam \
-L log_file.log \
-E error_file.err
```

# SAMtools collate

- collate command: sorts the input file by name

```
samtools collate \
-o output_sorted.bam \
input.bam
```

# UMItools prepare-for-rsem

- The output from UMItools dedup is not compatible with rsem (the next step in the DP pipeline)
- prepare-for-rsem command: outputs a file compatible with rsem

```
umi_tools prepare-for-rsem \
-I input_deduplicated_sorted.bam \
-S output.bam \
-L log_file.log \
-E error_file.err
```

# Results

# Picard's Time & Resources

- From what we were able to run:
  - Process for Picard + fgbio was fairly quick, totalling an **average of 70 minutes** for each sample run.
  - Memory wise, Picard required an **average of 65GB of RAM** to properly run.
- In short, Picard was fairly quick to run but a memory hog

# UMItools' Time & Resources

- From what we were able to run:
  - Process for UMItools took an entire day
    - Sorting takes about 1 hour per sample
    - Deduplication took around 5.5 hours per sample (one sample took over 21 hours to deduplicate)
  - Needed 60GB of RAM for most of the samples tested but some need more

# Duplication Rates

| Sample Name | Original % Duplication | UMItools %Duplication | Picard % Duplication |
|---|---|---|---|
| FLT_LAR_OLD_FL4 | 50.86 | 31.08 | 17.42 |
| GC_LAR_YNG_GL2 | 49.68 | 20.31 | 9.61 |
| VIV_LAR_OLD_VL12 | 57.92 | 44.46 | 28.65 |
| VIV_LAR_OLD_VL15 | 41.92 | 21.66 | 12.65 |
| VIV_LAR_YNG_VL17 | 45.18 | 17.49 | 8.84 |

# Conclusions

# Conclusions

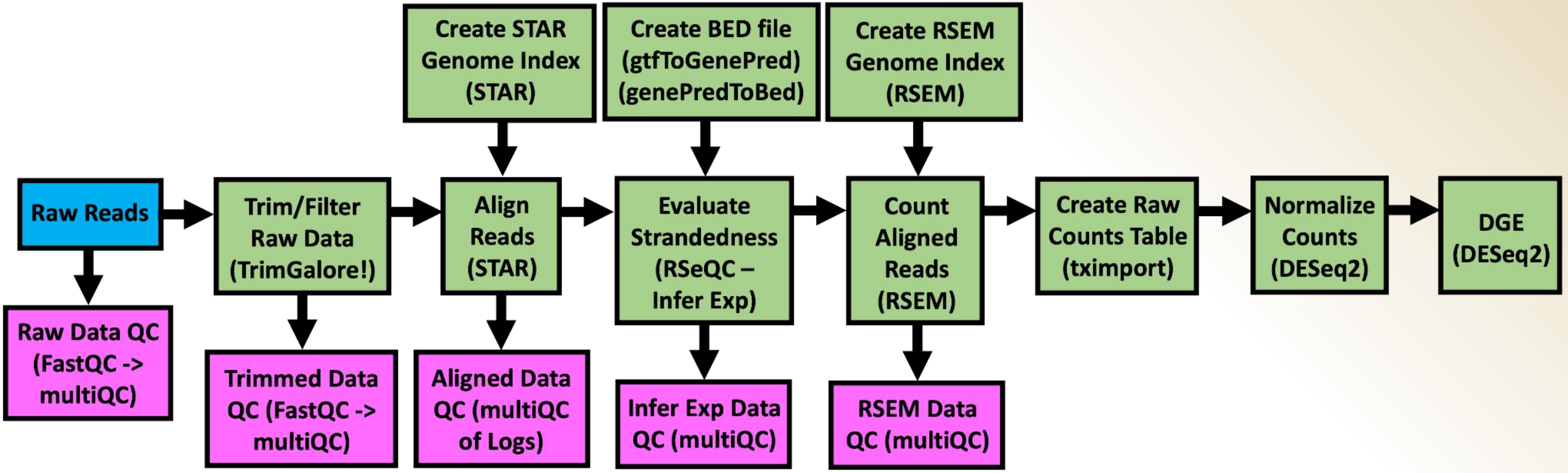
- Our results suggest that UMItools and Picard could work in deduplicating GeneLab RNAseq data
- The difference in duplication rates for UMItools and Picard could be due to the different deduplication methods used by both tools
  - Picard deduplicated using query name sorted files
  - UMItools uses mapping coordinates and UMIs to deduplicate

# Future Plans

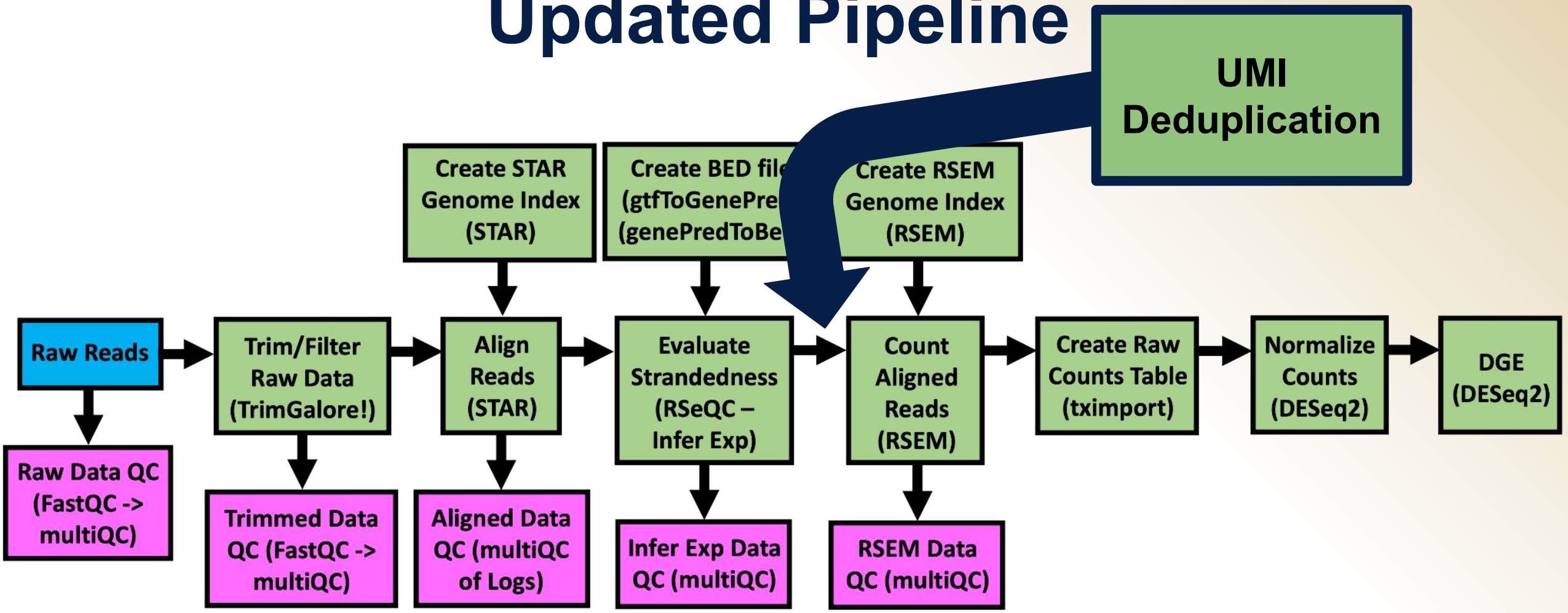
# Next Steps

- Run differential gene expression analysis on deduplicated samples
- Look more into why the UMItools and Picard duplication rates are different
- Optimize the tools used for running on the cluster
  - Time and resource efficiency
- If applicable, update the RNAseq pipeline to include deduplication

# Current RNAseq Pipeline



# Updated Pipeline



# Acknowledgements


- Funding for this project was provided by the NASA/BMSIS YSP
- Special thanks to
  - Dr. Barbara Novak
  - Dr. Amanda Saravia-Butler
  - Alexis Torres
  - NASA GeneLab Data Processing Team

# Thank You!