RNAseq Deduplication with UMIs

Radha Ganesh & David Ho

Under the Guidance of Dr. Barbara Novak

BPS

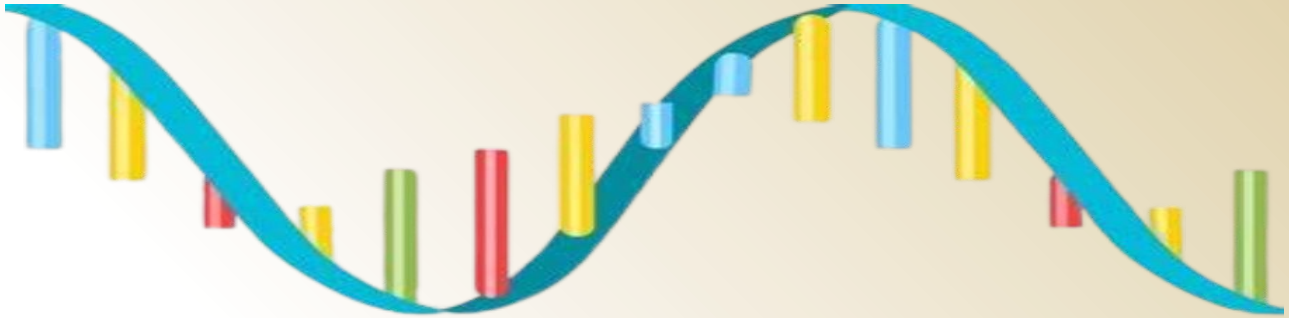
Biological & Physical Sciences

National Aeronautics and Space Administration

NASA

Background

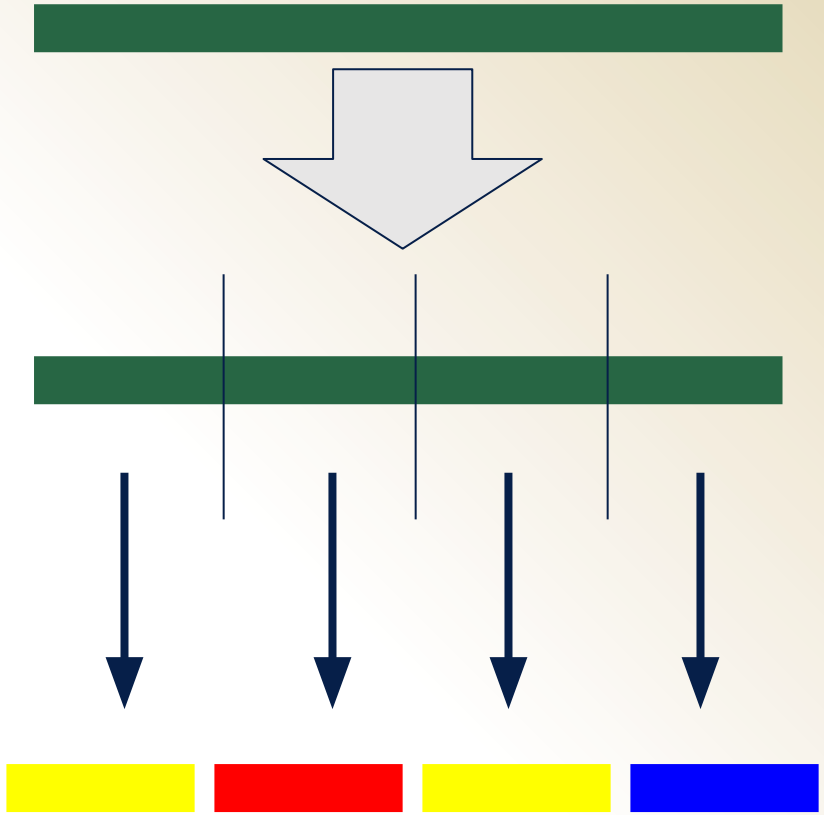
What is RNAseq?



- 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 technical duplicates

The Use of Unique Molecular Identifiers (UMIs)

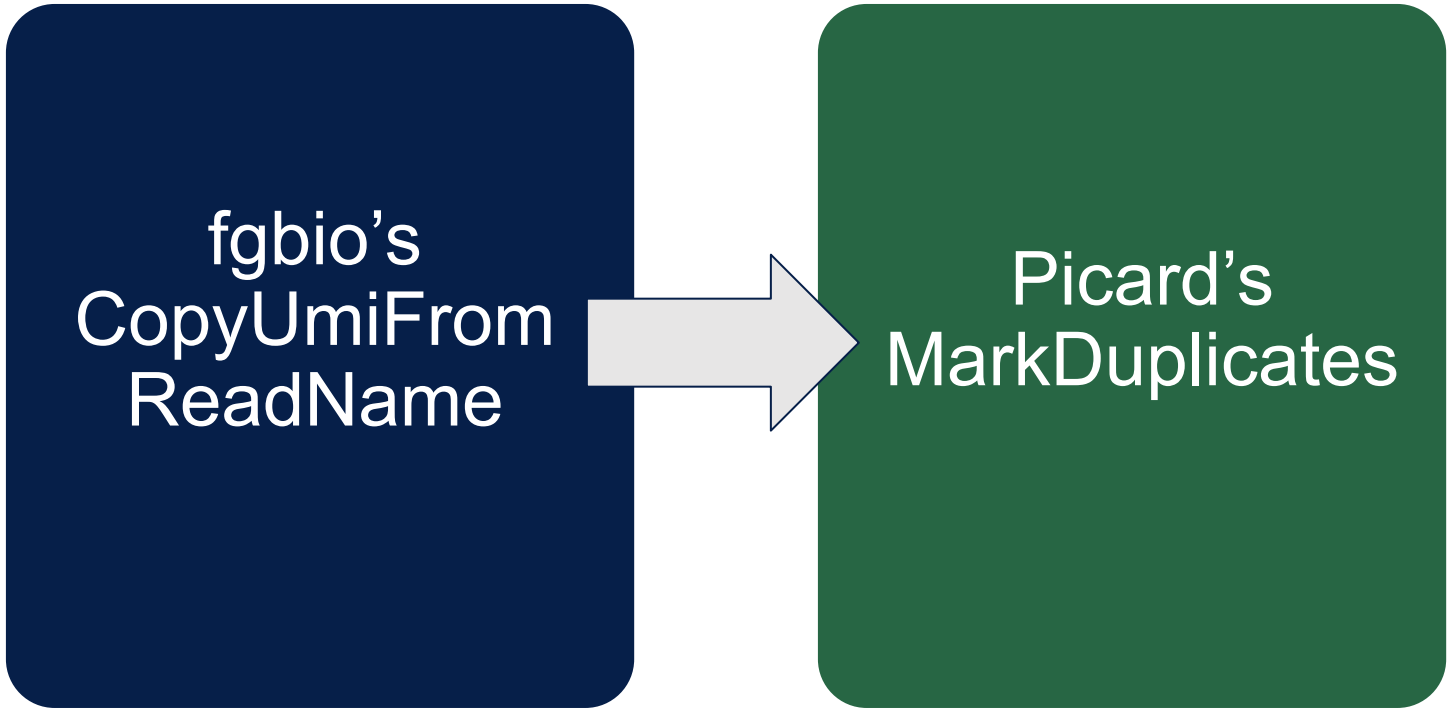
- UMIs help labs keep track of molecules and remove errors during library preparation and sequencing.
- Before the PCR amplification step, every molecule in a sample is uniquely labeled with a UMI, which is typically a random sequence of oligonucleotides.
- When we see two identical tags on two identical 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

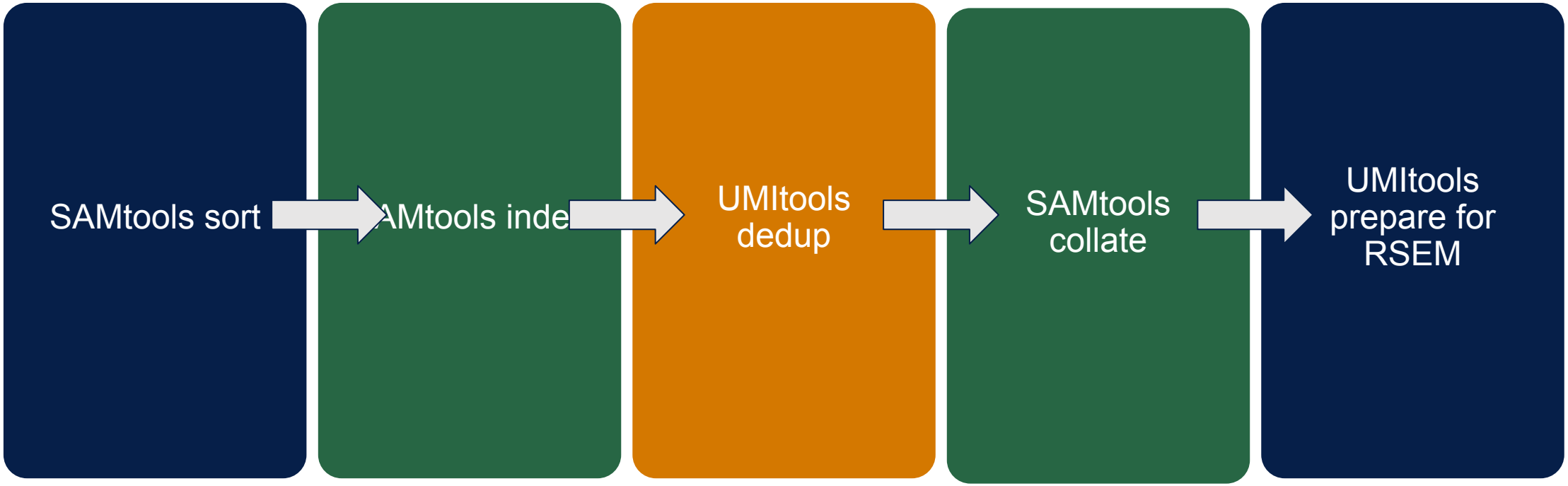
- As of now, there exists no single approach to efficiently remove technical duplicates using UMIs.
- Many programs that exist are very time and memory intensive
- Through this project, we evaluated the efficiency of two different deduplication methods.
- The tools we used for deduplication were Picard's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

Methods

Tools Used for Picard's MarkDuplicates



Tools Used for UMItools dedup



Results

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

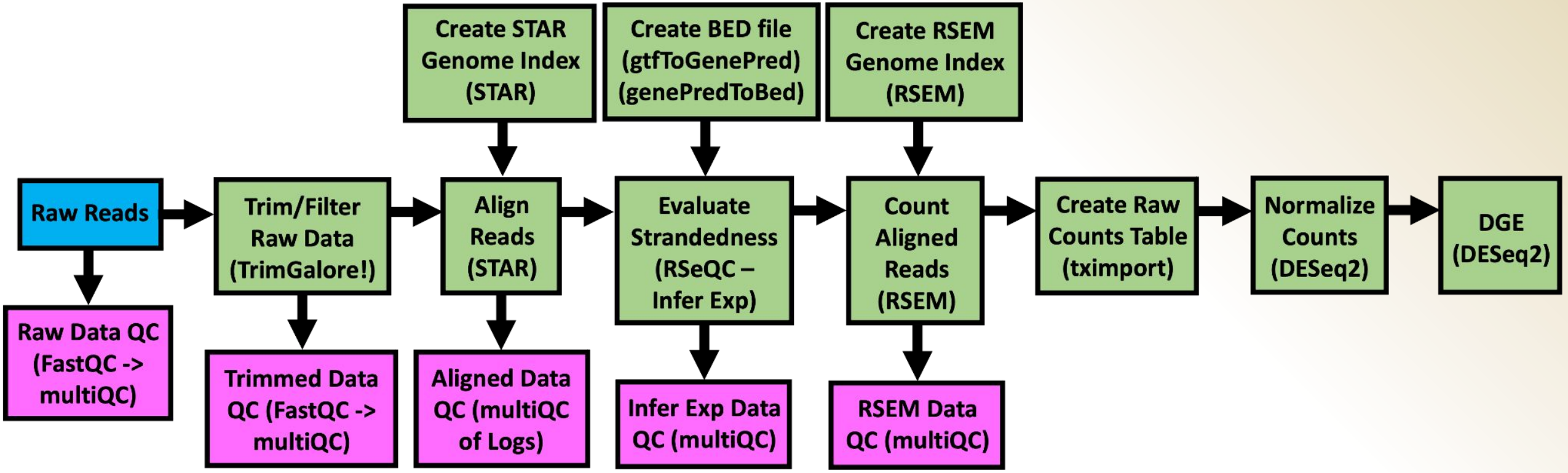
- Our results suggest that UMItools and Picard could work in deduplicating GeneLab RNAseq data
- The difference in duplication rates for UMItools and Picard could be due to how strict the two tools are in determining is two UMIs are the same.
  - Picard requires an exact match, while UMItools allows for one mismatch.

Future Plans

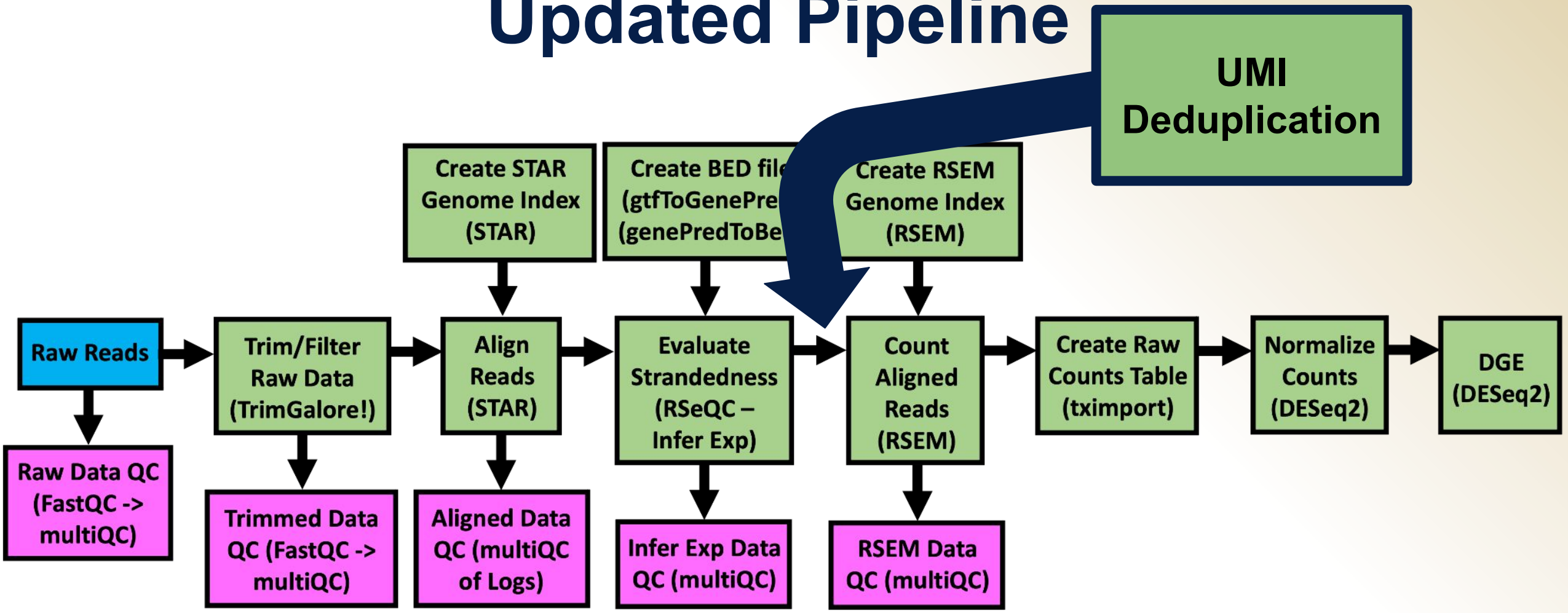
Next Steps

- 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!