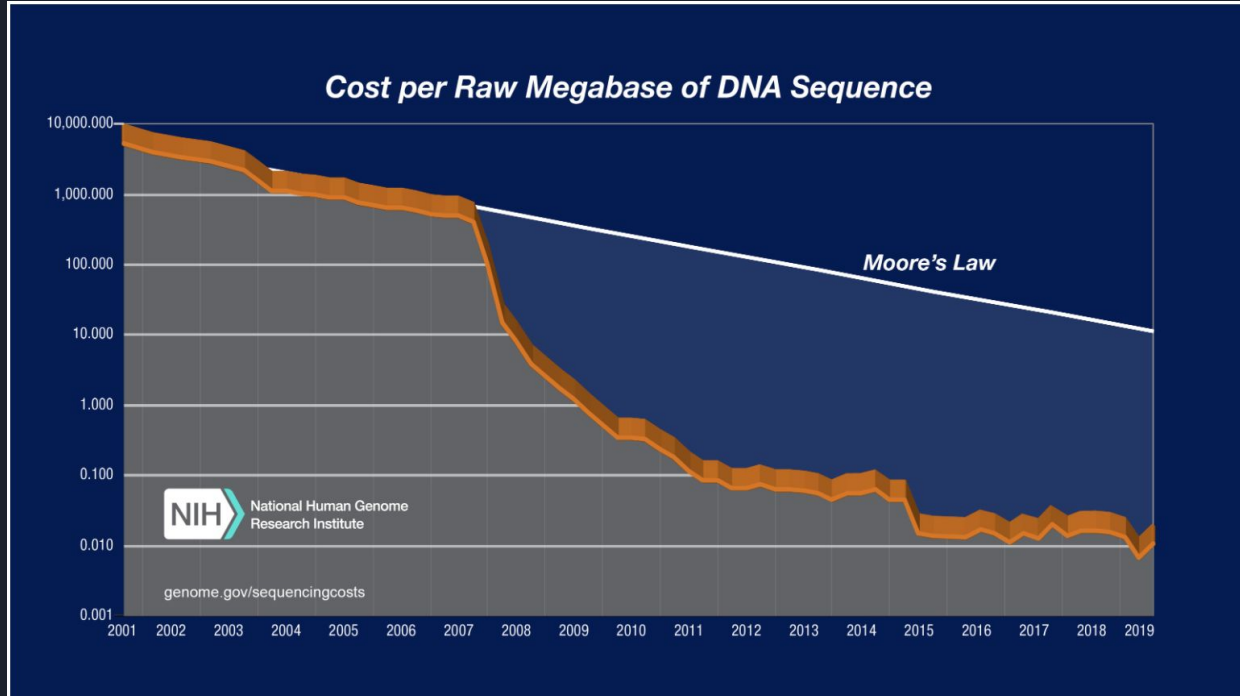




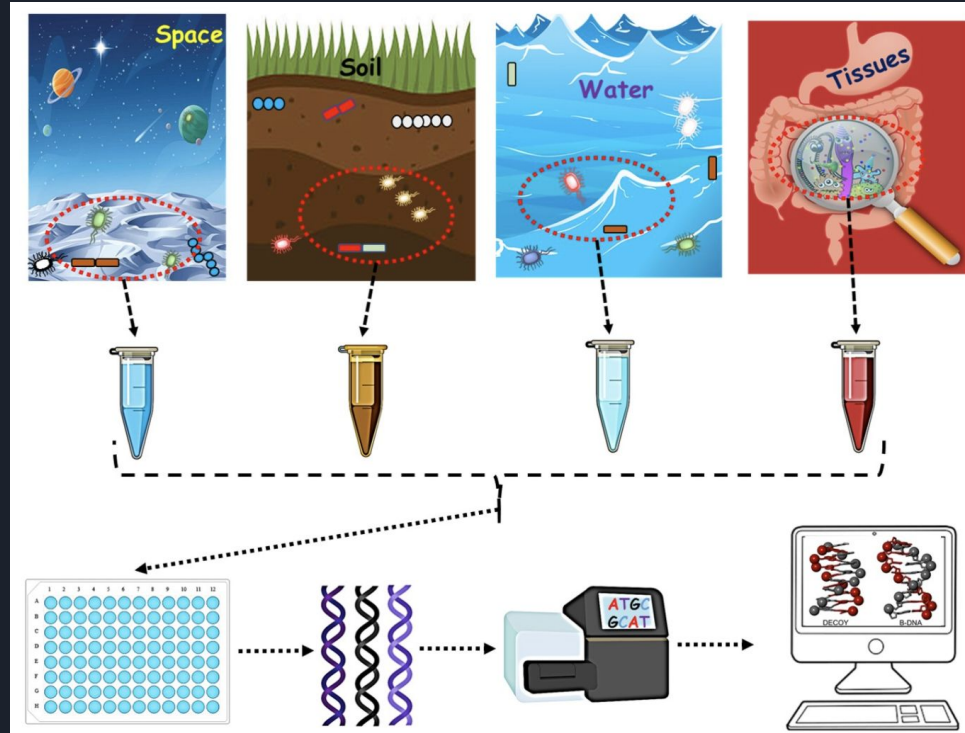
Optimizing Data Processing of Metagenomes

Sergio E. Mares - Oklahoma State University
Humberto G. Ortiz-Zuazaga - University of Puerto
Rico, Rio Piedras

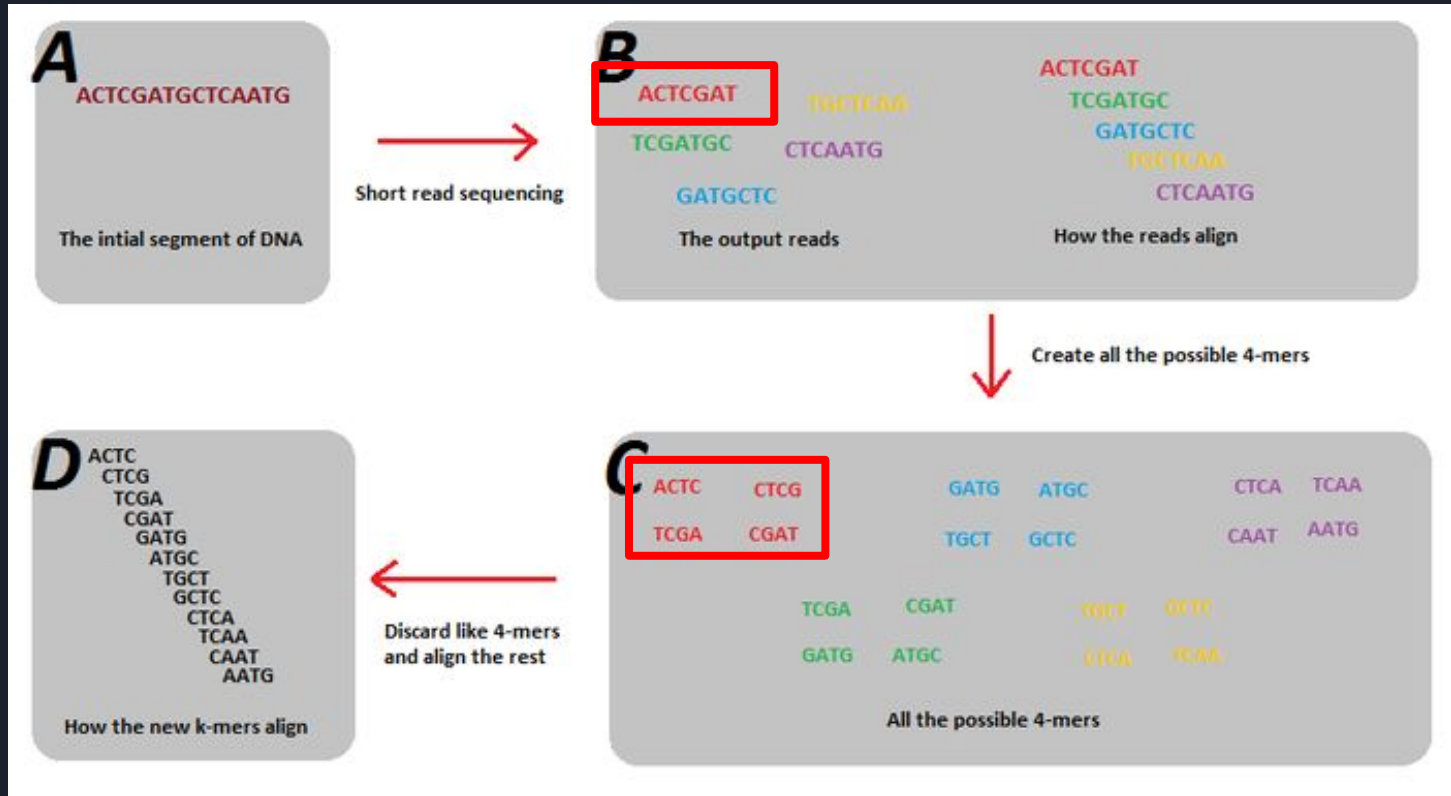
Data cannot be processed quick enough



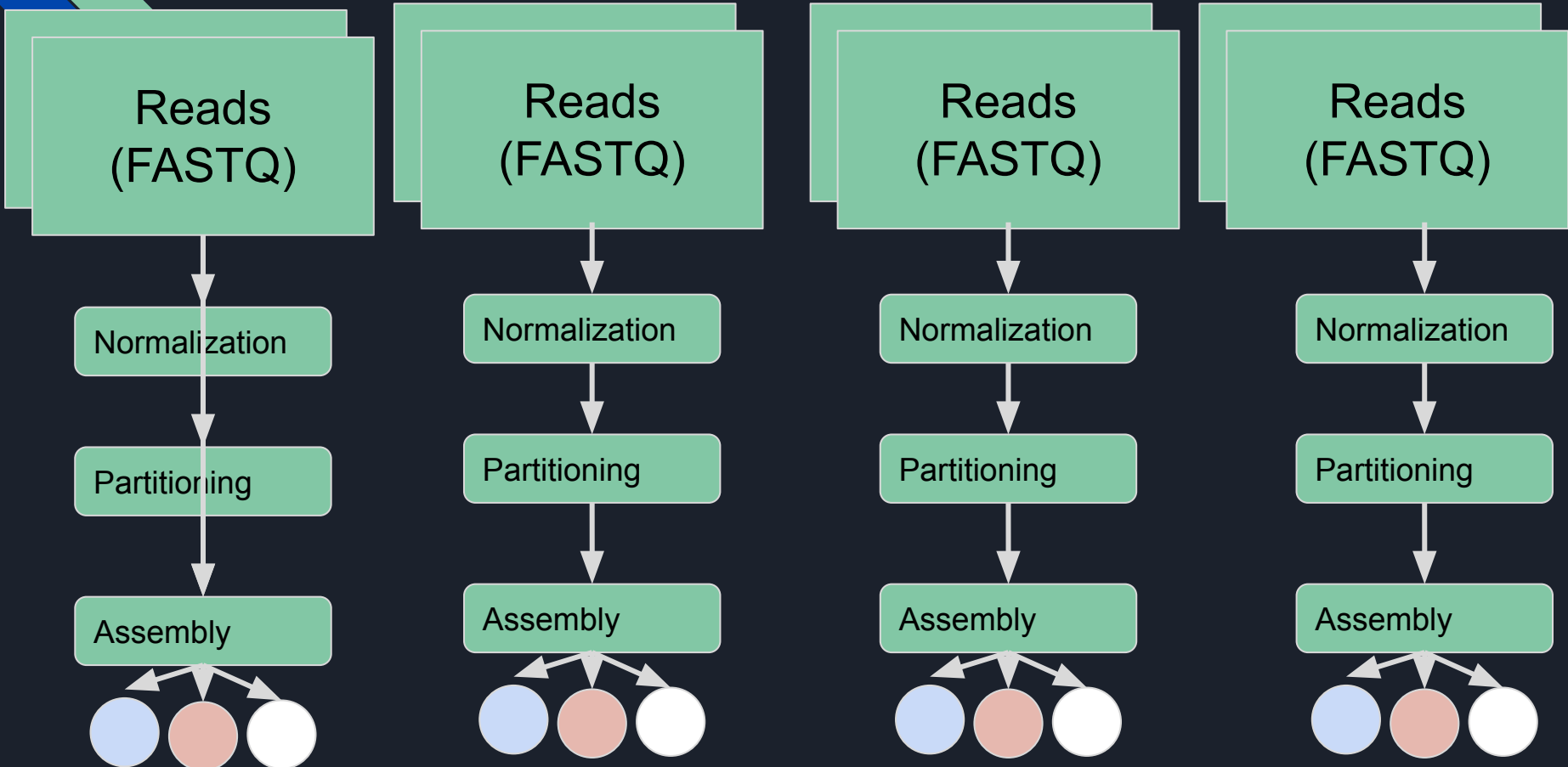
Metagenomics



Structure of data processing (Existing k-mer analysis)



Structure of data processing (Existing k-mer analysis)



A Statistical Approach: ANOVA

Reads
(FASTQ)

Normalization

Partitioning

Assembly



Reads
(FASTQ)

Extract the not
changing k-mers

Assembly



A Statistical Approach: ANOVA

S1: Volcanic
Cold:Wet

[...]

S5:
Karst: Hot:
Very Dry

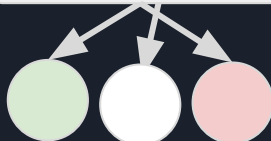
1. Trimming
2. Quality Filter

1. Trimming
2. Quality Filter

Filter the unique
k-mers

1. Modify existing
tools to discard
the k-mers that do
not change
between samples

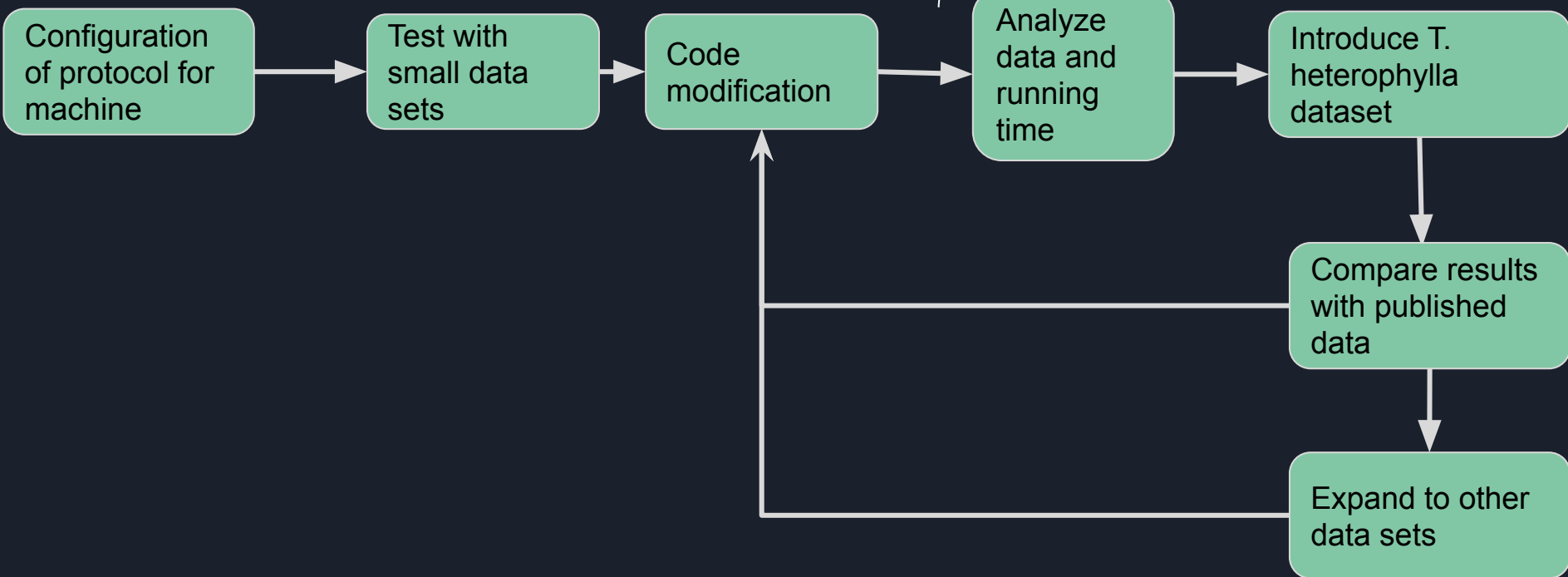
Assembly



Tabebuia heterophylla

Time-line

Approx. 7 weeks





Expected Results

	Size of the file (bytes)
Trim (Step 1)	662,643,774
Normalization (Step 2)	148,972,458
Statistical ANOVA Approach	???

Improving data processing for quicker analysis

