Exploring SPLASH

A new statistically-backed, reference-free approach to genomic analysis

Presentation by Hannah Beakley
Professor Ritambhara Singh's Lab
Brown University

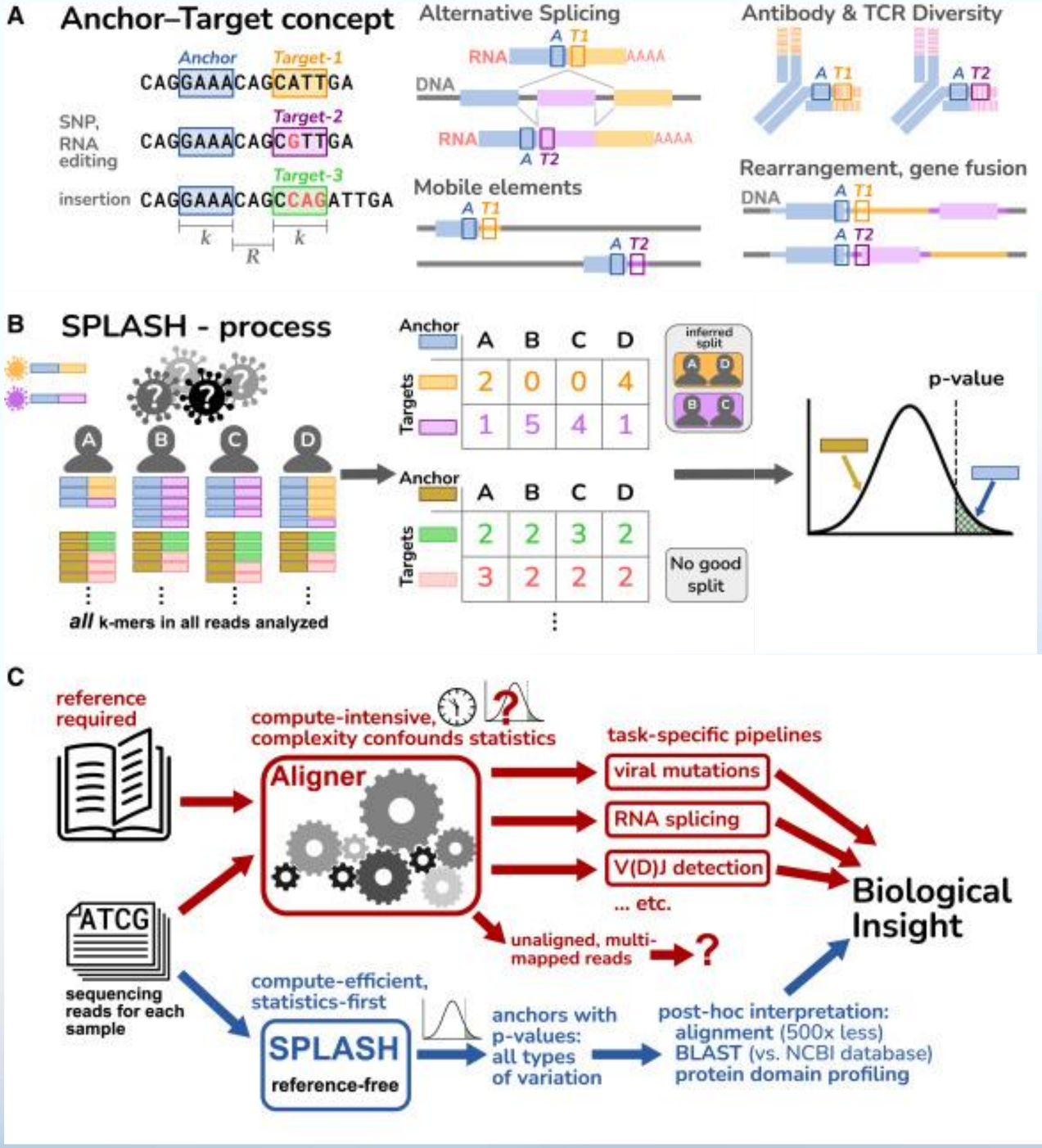
What is SPLASH?

Goal: use SPLASH to analyze IISAGE data
My project: replicate results from SPLASH paper

- Genomic analysis today involves aligning reads to a reference genome, which is problematic
- **Splash: reference free + statistically backed!**
- SPLASH employs a k-mer based algorithm to directly analyze raw sequencing data

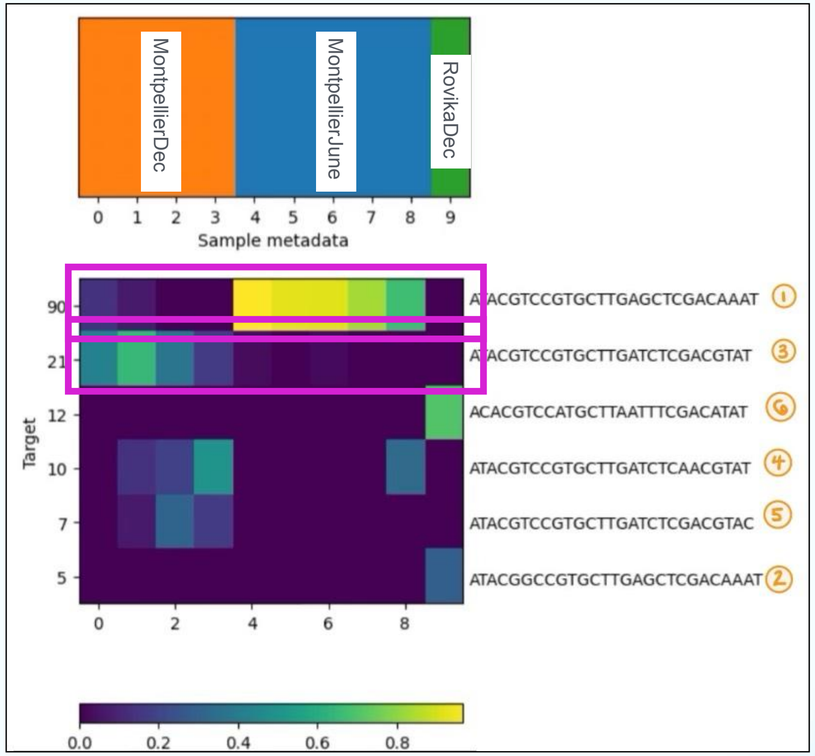
K-mer: subsequence of length k
Anchor: a k-mer that is shared between samples
Target: a k-mer R units from an anchor that varies between samples

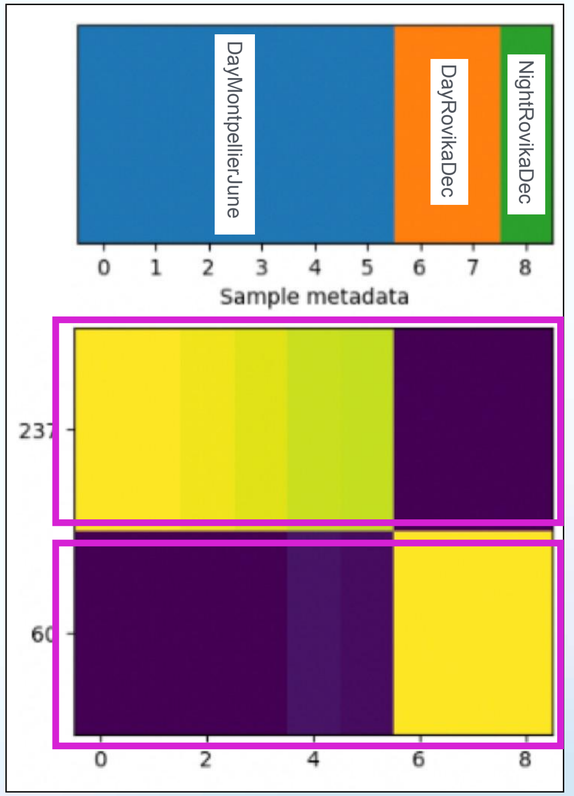
- SPLASH generates a contingency table for each anchor
- Performs statistical test to determine whether target frequencies differ significantly across samples
- Outputs significant anchors, representing genomic regions that vary between samples — promising for further analysis

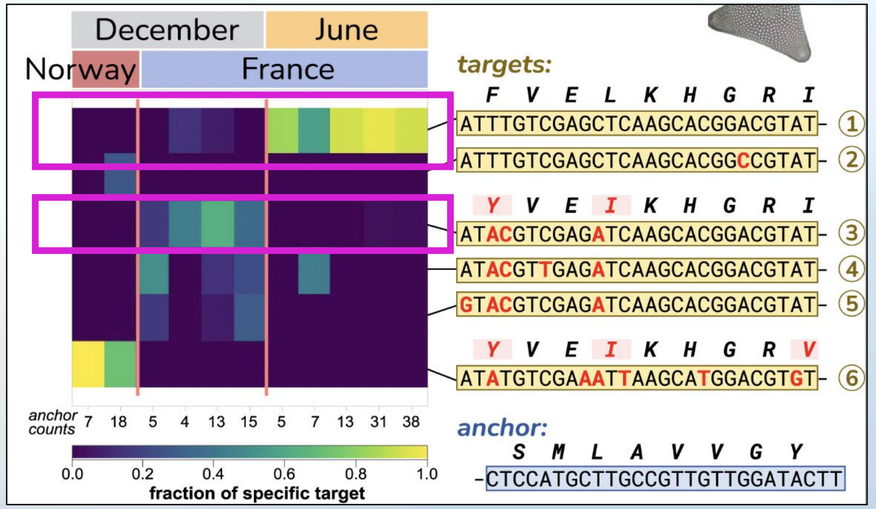


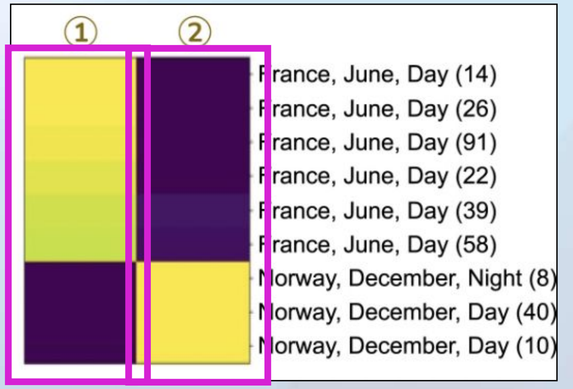
My Methods + Results

- Connected to OSCAR (Brown's computer cluster)
- Installed SPLASH + make sense of scripts/functionalities
- Downloaded 47 eelgrass RNA-seq samples from NCBI
- Wrote scripts to pre-process fastq files
- Ran SPLASH on eelgrass data
- Generated plots for specific anchors referenced in the paper to see if I got similar results
- Results are effectively identical!









Concluding Thoughts

- First time doing comp bio work in real-world setting — working with a super computer, writing scripts, managing disk space, troubleshooting errors
- My results give me confidence that I have a valid SPLASH workflow
- Ultimate goal: run SPLASH on IISAGE data to extract relevant sequences across conditions and use ML to draw insights