Report

SPLASH: Workflow and Ideas

Lumi 12112618

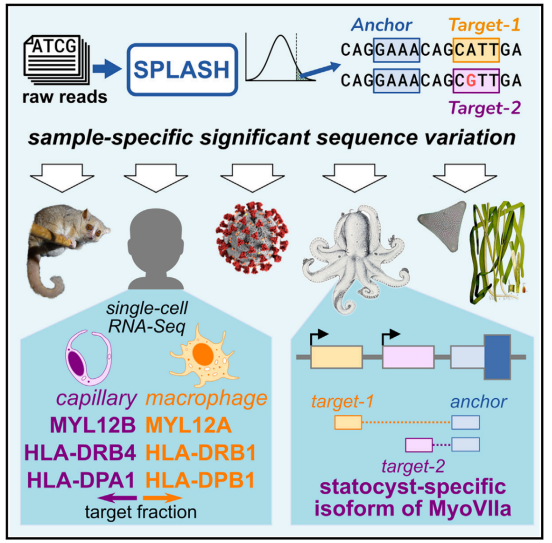
08.10.2024

Cell

Article

SPLASH: A statistical, reference-free genomic algorithm unifies biological discovery

Graphical abstract



Authors

Kaitlin Chaung, Tavor Z. Baharav, George Henderson, Ivan N. Zheludev, Peter L. Wang, Julia Salzman

Correspondence

julia.salzman@stanford.edu

In brief

Genomics workflows typically map reads onto a reference genome as the foundation for downstream analyses. However, this poses severe limitations for biological discovery when references are incomplete or nonexistent, and even for intensely studied genomes with rich population-level diversity. SPLASH is a highly efficient framework for statistics-driven analysis of sequence variation directly from raw sequencing data, overcoming previous limitations.

PNAS RESEARCH ARTICLE | STATISTICS BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

OASIS: An interpretable, finite-sample valid alternative to Pearson's X^2 for scientific discovery

Tavor Z. Baharav[a,b], David Tse[c], and Julia Salzman[d,e,f,1]

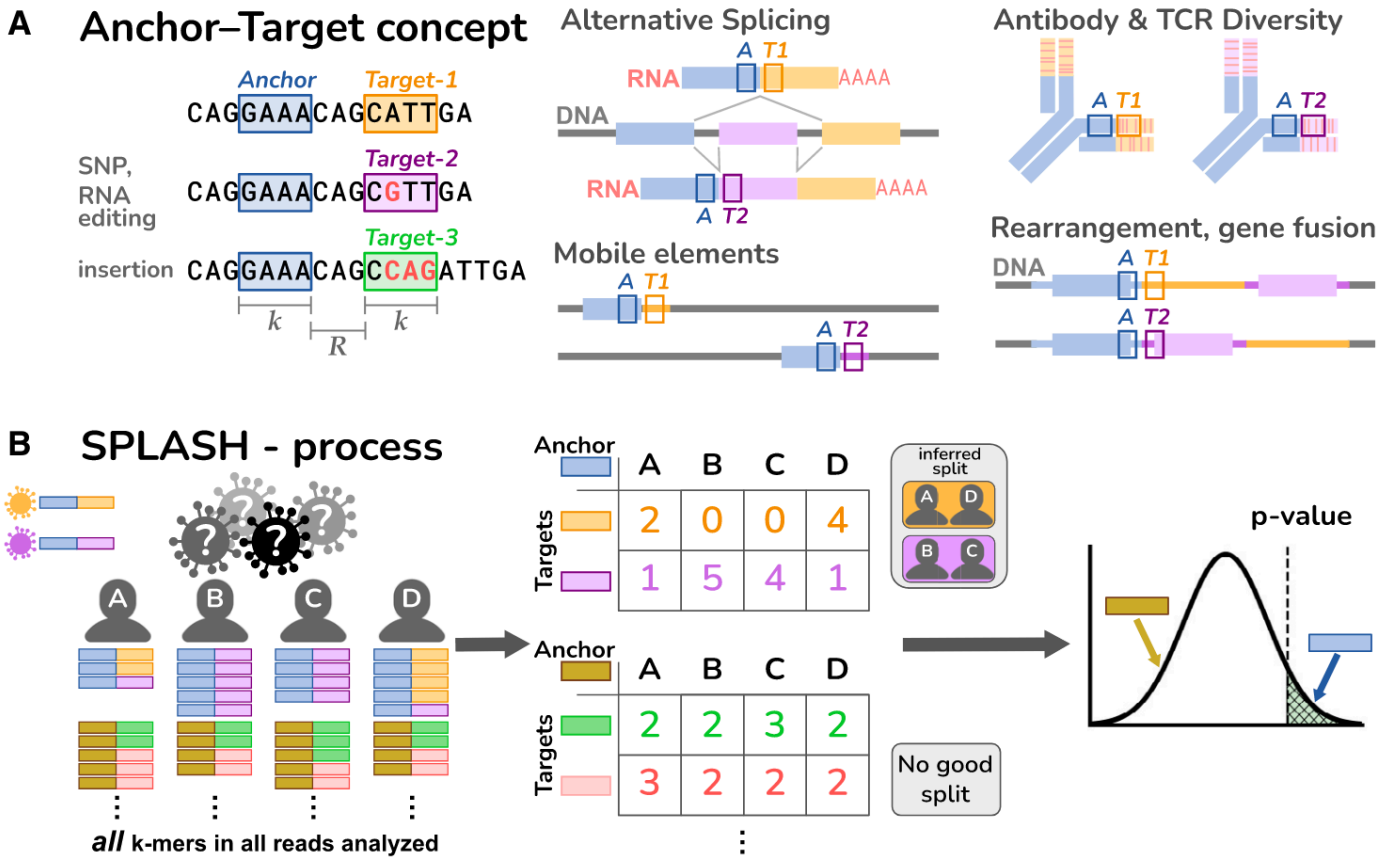
Edited by Kenneth Lange, University of California, Los Angeles, CA; received April 5, 2023; accepted February 8, 2024

Overall Workflow

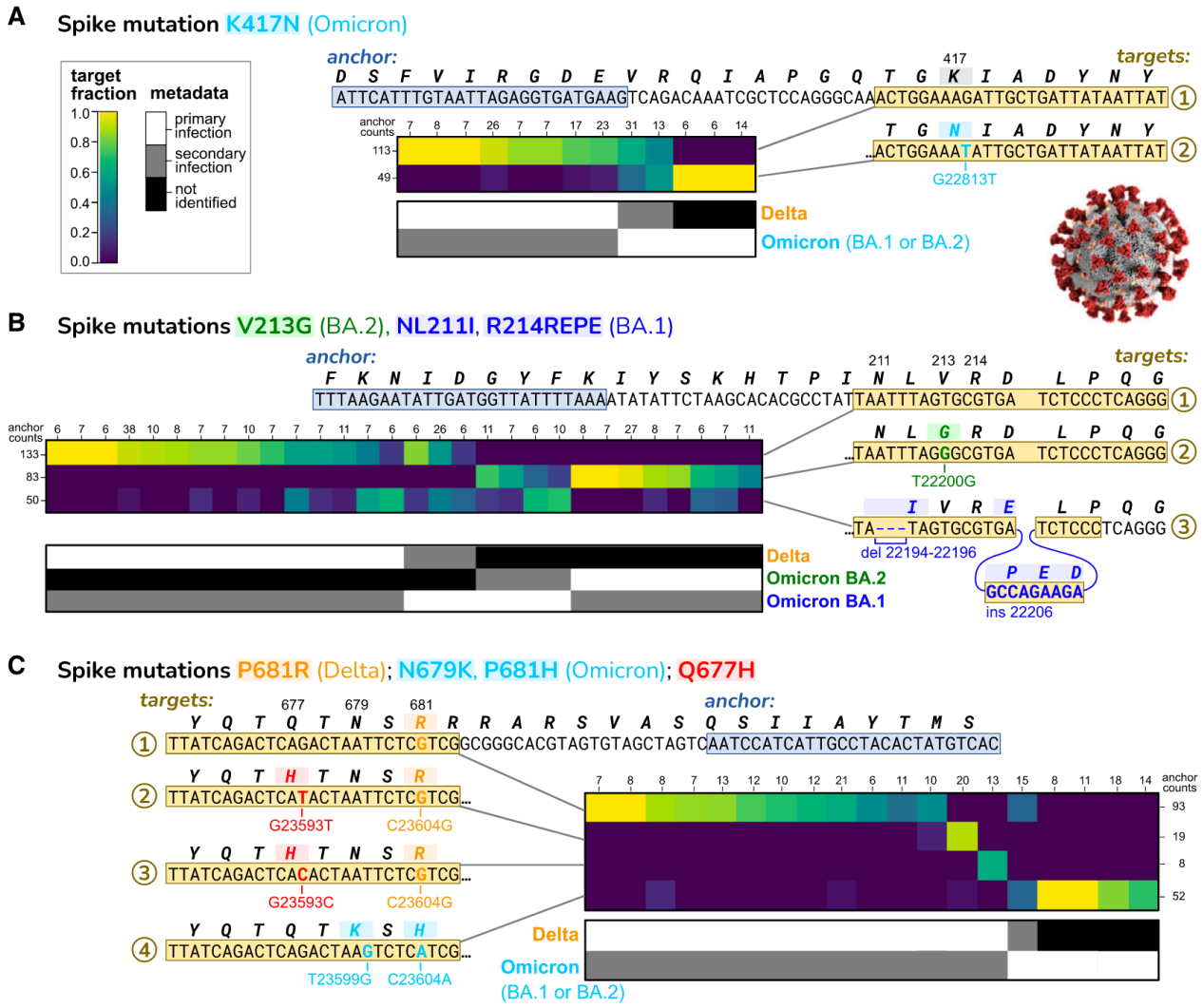
- Input Fastq Files
- Generate Anchor-Target Matrix
- OASIS stastic test
- Differentiate cell type

Q:

1. Generation of Anchor?
2. OASIS test?
3. Differential analysis with ground truth?



SARS-Covid Example



Found Targets that could differentiate subtypes of the virus.

Anchor-target matrix is treated as a contingency matrix X.

Useful targets are selected for grouping with $\vec{f}$.

Samples are grouped based on a $\vec{c}$.

Statistic S was calculated as:

$$S = \vec{f}^{\top} \tilde{X} \vec{c}$$

Normalization of X

The Matrix X was normalized based on the column counts.

Some Matrices are defined as follows:

$$E = \frac{1}{M} X \vec{1} \vec{1}^\top X$$

$$\tilde{X} = (X - E)\mathrm{diag}(\frac{1}{\sqrt{X^\top \vec{1}}})$$

P-value based on S

In the *Cell* paper, the P-value was calculated as:

$$P = 2\exp\left(-\frac{2(1-\xi)^2 S^2}{\sum_j c_j}\right) + 2\exp\left(-\frac{2\xi^2 M S^2}{(\sum_j c_j \sqrt{n_j})^2}\right)$$

Where:

M is the total number of counts.

n_j is the number of counts in column (sample) j.

c_j is the jth entry of $\vec{c}$.

ξ is specifically chosen to minimize the P-value. $\xi = \left(1 + \sqrt{\frac{M \sum_j c_j}{(\sum_j c_j \sqrt{n_j})^2}}\right)^{-1}$

Finding the Optimal f and c

Algorithm:

1. Randomize $\vec{c}$.
2. Set $\vec{f} = \text{sign}(\tilde{X}\vec{c})$.
3. $\vec{f} = \{\frac{1+\vec{f}}{2}, \frac{1-\vec{f}}{2}\}$
4. $\vec{f} = \text{argmax}_{\vec{f}}(\vec{f}^{\top}\tilde{X}\vec{c})$
5. $\vec{c} \propto \tilde{X}^{\top}\vec{f}$. (With constraint $||\vec{c}|| \leq 1$)
6. Repeat 2-4 until S doesn't change.
7. Output: $\vec{f}, \vec{c}$.

Better version of P value bound:

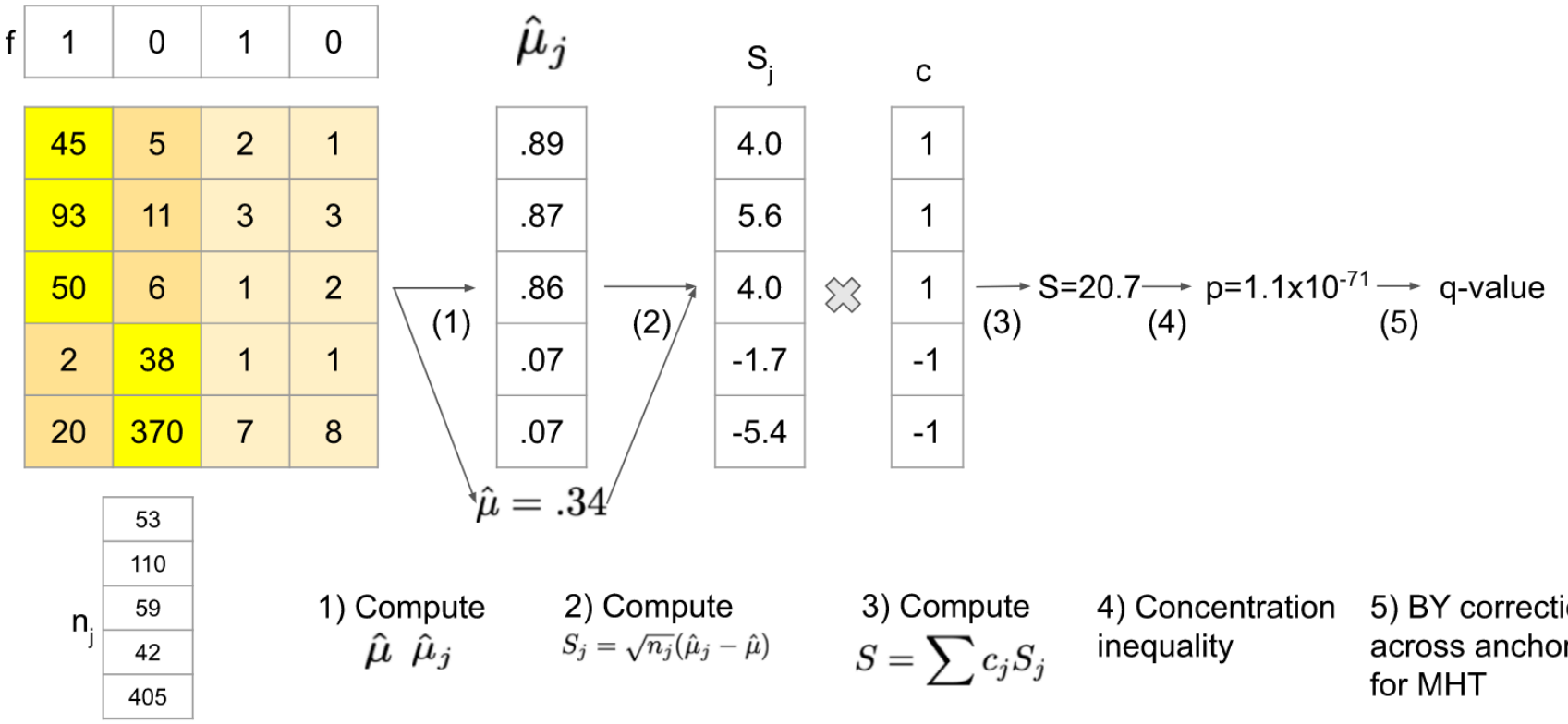
$$P \leq 2\exp(-\frac{2s^2}{1-\gamma})$$

Want larger s:

$$\arg\max_{0\leq\vec{f}\leq 1, ||\vec{c}||_2\leq 1}\vec{f}^{\top}\tilde{X}\vec{c}$$

Visual Example

A p-value computation



1) Compute $\hat{\mu}$ $\hat{\mu}_j$

2) Compute $S_j = \sqrt{n_j}(\hat{\mu}_j - \hat{\mu})$

3) Compute $S = \sum c_j S_j$

4) Concentration inequality

5) BY correction across anchors for MHT

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} \vec{f}_{Z_{k,j}}$$

Where k indicates the k th observation.

Interesting Properties of c