Report

SPLASH: Workflow and Ideas

Lumi 12112618 08.10.2024

Cell

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 statisticsdriven analysis of sequence variation directly from raw sequencing data, overcoming previous limitations.

PNAS

STATISTICS RESEARCH ARTICLE BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Check for OPEN ACCESS

Article

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

Tavor Z. Baharav^{a,b}^(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:

- . Generation of Anchor?
- . OASIS test?
- . Differential analysis with ground truth?

SARS-Covid Example

А **Spike mutation K417N (Omicron)** anchor: targets: 417 D S F V I R G D E V R Q I A P G Q T G <mark>K</mark> I A D Y N Y target fraction metadata 1.0 primary
nfection 7 26 7 7 17 23 31 13 6 6 14 T G <mark>N</mark> I A D Y N Y $0.8 -$ <u>LACTGGAAATATTGCTGATTATAATTAT</u> 113 secondary infection $0.6 -$ G22813T L not
Lidentified $0.4 -$ Delta $0.2 -$ **Omicron (BA.1 or BA.2)** в Spike mutations V213G (BA.2), NL211I, R214REPE (BA.1) anchor: targets: 211 213 214 F K N I D G Y F K I Y S K H T P I N L V R D L P Q G TTTAAGAATATTGATGGTTATTTTAAAATATATTCTAAGCACACGCCTATTAATTTAGTGCGTGA 6 26 6 11 7 6 10 8 7 27 8 7 6 7 11 anchor N L G R D LP Q G counts TAATTTAGGGCGTGA TCTCCCTCAGGG(2) $133 -$ 83 T22200G I V R E LP Q G TA---TAGTGCGTGA TCTCCCTCAGGG (3) del 22194-22196 $P E D$ **Omicron BA.2 SCCAGAAGA-Omicron BA.1** ins 22206

С Spike mutations P681R (Delta); N679K, P681H (Omicron); Q677H

> targets: anchor: 677 Y O T O T N S <mark>R</mark> R R A R S V A S O S I I A Y T M S (1) TTATCAGACTCAGACTAATTCTCGTCGCCGGCCACGTAGTGTAGCTAGTCAATCCATCATTGCCTACACTATGTCAC Y Q T H T N S R 14 anchor 7 13 12 10 12 21 6 11 10 20 13 15 8 11 18 2 TTATCAGACTCATACTAATTCTCGTCG. 93 G23593T C23604G 19 Y Q T H T N S R 3) TTATCAGACTCACACTAATTCTCGTCG. 52 $G23593C$ C23604G Y Q T Q T <mark>K</mark> S H **Delta** 4 TTATCAGACTCAGACTAAGTCTCATCG Omicrol T23599G C23604A (BA.1 or BA.2)

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 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:

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

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

P -value based on S

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

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

Where:

 M is the total number of counts.

 n_j is the number of counts in column (sample) j . c_j is the *j*th entry of \vec{c} .

 ζ is specifically chosen to minimize the P -value. $\zeta = (1+\sqrt{\frac{M\sum_j c_j}{(\sum_i c_j\sqrt{n_j})^2}})^{-1}$

Finding the Optimal f and c

Algorithm:

1. Randomize \vec{c} .

2. Set
$$
\vec{f} = sign(\tilde{X}\vec{c})
$$
.
\n3. $\vec{f} = \{\frac{1+\vec{f}}{2}, \frac{1-\vec{f}}{2}\}$
\n4. $\vec{f} = argmax_{\vec{f}}(\vec{f}^\top \tilde{X}\vec{c})$
\n5. $\vec{c} \propto \tilde{X}^\top \vec{f}$. (With constraint $||\vec{c}|| \le 1$)
\n6. Repeat 2-4 until *S* doesn't change.
\n7. Output: \vec{f}, \vec{c} .

Better version of P value bound:

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

Want larger s:

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

Visual Example

p-value computation A

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

Where k indicates the k th observation.

Interesting Properties of c

