VERSION 5

NOV 20, 2019

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocol.s.io.9i7h4hn

Protocol Citation: Gregory P Harhay 2019. Steps to Create FASTQ of CCS Overlapping Genomic SSR - CCS ROI . **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.9i7h4hn>

MANUSCRIPT CITATION:
 Harhay, G.P., Harhay, D.M., Bono, J.L. *et al.*A Computational Method to Quantify the Effects of Slipped Strand Mispairing on Bacterial Tetranucleotide Repeats. *Sci Rep*9,18087 (2019).
<https://doi.org/10.1038/s41598-019-53866-z>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
 We use this protocol and it's working

Steps to Create FASTQ of CCS Overlapping Genomic SSR - CCS ROI V.5

Gregory P Harhay¹

¹United States Department of Agriculture (USDA) Agricultural Research Service (ARS), US Meat Animal Research Center (USMARC), State Spur 18D, Clay Center, NE 68933, USA

USMARC Bacterial Systems Biology Working Group

Gregory P Harhay

United States Department of Agriculture (USDA) Agricultural ...

ABSTRACT

The virulence and pathogenicity of bacterial pathogens are related to their adaptability to changing environments. One process enabling adaptation is based on minor changes in genome sequence, as small as a few base pairs, within segments of genome called simple sequence repeats (SSRs) that consist of multiple copies of a short sequence (from one to several nucleotides), repeated in series. SSRs are found in eukaryotes as well as prokaryotes, and variation in them occurs at frequencies up to a million-fold higher than the average bacterial mutation rate through a process of slipped stranded mispairing (SSM) by DNA polymerase during replication. The characterization of SSR length by standard sequencing methods is complicated by the appearance of length variation introduced during the sequencing process that does not accurately quantify lower-abundance repeat number variants in a population. Here we report a computational approach to correct for process-induced artifacts, validated for tetranucleotide repeats by use of synthetic constructs of fixed, known length. We apply this method to a laboratory culture of *Histophilus somni*, prepared from a single colony, and demonstrate that the culture consists of populations of distinct sequence phase and read length variants at individual tetranucleotide SSR loci.

Input requirements: Closed Genome - It is recommended that only organisms with closed genomes be the subject of the analyses described here. Mapping repetitive reads to contigs of non-closed genomes may map to multiple locations, complicating the analysis. Mapping CCS (circular consensus sequence) with repetitive sequence to closed genomes are guaranteed to map to a single locus if sufficient unique flanking sequence is used to confirm the unique mapping. Consequently, long CCS with high base quality are the most desirable input into this workflow.

Created: Nov 20, 2019

Last Modified: Nov 20, 2019

PROTOCOL integer ID:
30015

Protocols.io

Reads spanning SSR
identification, selection,
and writing to FASTQ file

MatLab

CodeOcean or
local instance

QC, filtering,
results

Protocol workflow

Software Requirements

1

Software

Bedtools

NAME

<http://quinlanlab.org>

DEVELOPER

<https://github.com/arq5x/bedtools2>

SOURCE LINK

For those without Matlab licences, Matlab code can be run in CodeOcean at this [Matlab Compute Capsule](#)

Software

Phobos

NAME

Dr. Christoph Mayer

DEVELOPER

Software

Geneious

NAME

Biomatters Ltd

DEVELOPER

Geneious was used to as wrapper to for running sequence mappers, phobos, and sequence

Download Genome and BAM Alignment of nine CCS libraries..

- 2 1. Genome available at at Genbank [CP018802.1](#)

2. BAM file available at <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?run=SRR8080935> , the download BAM file is named SRR8080935_CP018802.bam

Identify SSR in Genome

- 3 Run Phobos to identify simple sequence repeats: search for repeats 2-mers to 10-mers in CP018802.1 genome. This Geneious plugin does not provide access to all potential running modes and defaults to providing repeat unit naming using "normalised alphabetical mode," where the repeat unit reported is independent of strand and phase enabling Phobos to choose the repeat pattern that comes first in the alphabet.

Locate Tandem Repeat(s) with Phobos

The Phobos executable:

Search modes:

☐ Mask repeats

☐ Trim repeats from ends Min bases from end

☐ Remove hidden repeats

Repeat unit length: Min Max

Options for imperfect search

Imperfect search presets:

Mismatch score:

Gap score:

Recursion depth:

☐ Maximum score reduction

Requirements for satellites to be reported

Satellite constraints:

Minimum length: OR (+ * unit length)

Minimum score: OR (+ * unit length)

% perfection: Min Max

N handling

Maximum successive N's:

Treat N's when computing perfection

☐ In alignment, treat N's as missense instead of neutral

Phobos - a tandem repeat search tool © Christoph Mayer
If you publish results, please cite Phobos as described on the [Phobos Home Page](#)

Save repeats to CP018802.1 genome annotation

Select SSRs For Further Analysis

- 4 These tetranucleotide repeats were selected for further analysis.

Repeat Unit	Name	Minimum	Maximum	Length	Percentage Perfection

AACC	Tetranucleotide Repeat	1792217	1792466	250	100.00%
AATC	Tetranucleotide Repeat	1452562	1452715	154	100.00%
ACTG	Tetranucleotide Repeat	1501321	1501467	147	100.00%
ACTG	Tetranucleotide Repeat	1456013	1456119	107	100.00%
AAGC	Tetranucleotide Repeat	1834016	1834094	79	100.00%

For each SSR, extract CCS mapping SSR Regions of Interes...

5

Create BAM file of CCS overlapping SSR ROI using coordinates identified in step 4 and transferred to their respective BED file to be used in combination with the BAM file of all reads mapping to the genome. When specifying position of SSR, allow for 5 bp on each flank. Please keep in mind the BED file convention, the left coordinate is 0-based while the right coordinate is 1-based.

For selecting CCS mapping to SSR, use BED to define coordinates

☐ CP018802_79_bp_SSR.bed
 ☐ CP018802_107_bp_SSR.bed
☐ CP018802_147_bp_SSR.bed
 ☐ CP018802_154_bp_SSR.bed
☐ CP018802_250_bp_SSR.bed

Command

find CCS that completely overlap 79 bp AAGC SSR including 5 bp adjacent non-SSR region on each flank

```
bedtools intersect -a Control_Single_Duplex_63bp_SSR_L_23088_raw_map_AAGC_Nm4.bam -b
CP018802_SSR_79bp_AAGC_Nm4.bed -F 1.0 -wa >
Control_Single_Duplex_63bp_SSR_L_23088_raw_map_intersect_AAGC_Nm4.bam
```

Perform this operation for all 5 SSR to create the following BAM files

☐ SRR8080935_CP018802_79bp.bam
 ☐ SRR8080935_CP018802_107bp.bam
☐ SRR8080935_CP018802_147bp.bam
 ☐ SRR8080935_CP018802_154bp.bam

Analyze CCS mapping to ROI on reference

- 6
 - Use Geneious to view CCS mapping to reference
 - Inspect alignment of CCS mapping to duplex. Note that gap regions between end of SSR region and first adjacent base of both flanking regions defined region of interest (ROI).
 - Some mappers such as BowTie2 tend to place "extra" repeat units in the gap region between the SSR and the first adjacent base to the left of the SSR, while Genious mapper tends to place "extra" repeat units to the right of the SSR, in the gap between the SSR and the first adjacent base
 - For each read the Geneious "Extract" function was used to select bases within the ROI to create a new FASTQ file of CCS with bases spanning the ROI.

Portion of CCS spanning the region of interest

Extract region of interest from each read into FASTQ file.

- 7 Create the following FASTQ reads for downstream analysis in MATLAB (link to CodeOcean capsule)

☐ SRR8080935_79bp_AAGC_SSR_ROI_reads.fastq

☐ SRR8080935_107bp_ACTG_SSR_ROI_reads.fastq

☐ SRR8080935_147bp_ACTG_SSR_ROI_reads.fastq

☐ SRR8080935_154bp_AATC_SSR_ROI_reads.fastq

☐ SRR8080935_250bp_AACC_SSR_ROI_reads.fastq

Computed processes in Matlab for pruning reads to reveal r..

- 8 Matlab scripts for processing ROI CCS (above step) are available for download and processing at [Matlab Compute Capsule](#)

Protocols.io
read spanning SSR
identification

MatLab
QC and filtering