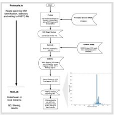


Nov 20, 2019

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

Gregory Harhay[1]

[1]United States Department of Agriculture

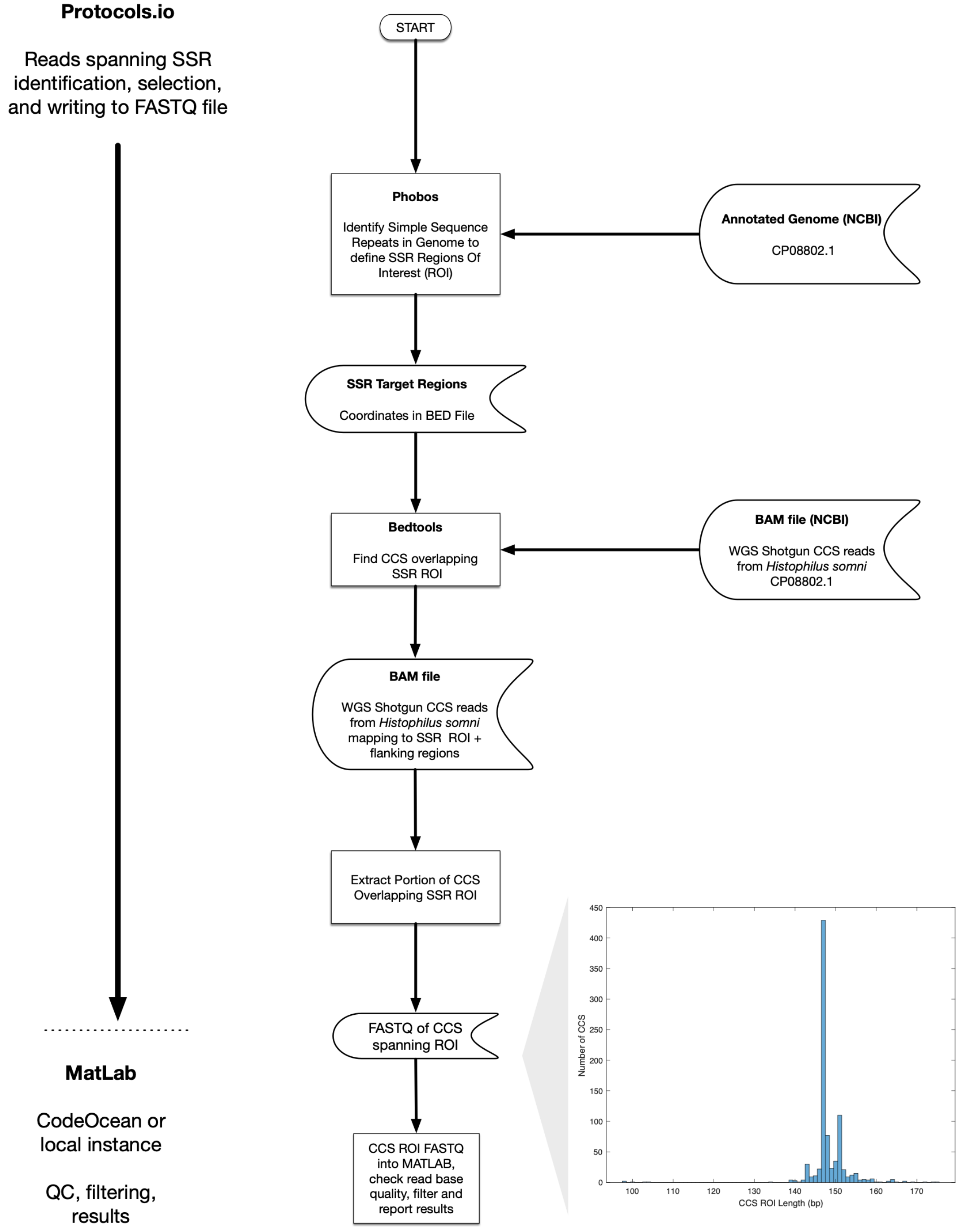
1 Works for me dx.doi.org/10.17504/protocols.io.9i7h4hn

Gregory P. Harhay
United States Department of Agriculture

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 to contigs of non-closed genomes may map to multiple locations, complicating tha analysis. Mapping CCS (circular consensus sequence) wiith repetitve sequence to closed genomes are guaranted to map to a single locus if sufficent unique flanking sequence is used to confim the unique mapping. Consequently, long CCS with high base quality are the most desirable input into this workflow.


Protocols.io
Reads spanning SSR identification, selection, and writing to FASTQ file
MatLab
CodeOcean or local instance
QC, filtering, results
START
Phobos
Identify Simple Sequence Repeats in Genome to define SSR Regions Of Interest (ROI)
Annotated Genome (NCBI)
CP08802.1
SSR Target Regions
Coordinates in BED File
Bedtools
Find CCS overlapping SSR ROI
BAM file (NCBI)
WGS Shotgun CCS reads from *Histophilus somni* CP08802.1
BAM file
WGS Shotgun CCS reads from *Histophilus somni* mapping to SSR ROI + flanking regions
Extract Portion of CCS Overlapping SSR ROI
FASTQ of CCS spanning ROI
CCS ROI FASTQ into MATLAB, check read base quality, filter and report results
Number of CCS
0
50
100
150
200
250
300
350
400
450
CCS ROI Length (bp)
100
110
120
130
140
150
160
170

Protocol workflow

Software Requirements

1

Bedtools 2.27.1
source by http://quinlanlab.org

For those without Matlab licences, Matlab code can be run in CodeOcean at this Matlab Compute Capsule

Phobos 3.3.12
by Dr. Christoph Mayer

Geneious 11.1.5
by Biomatters Ltd

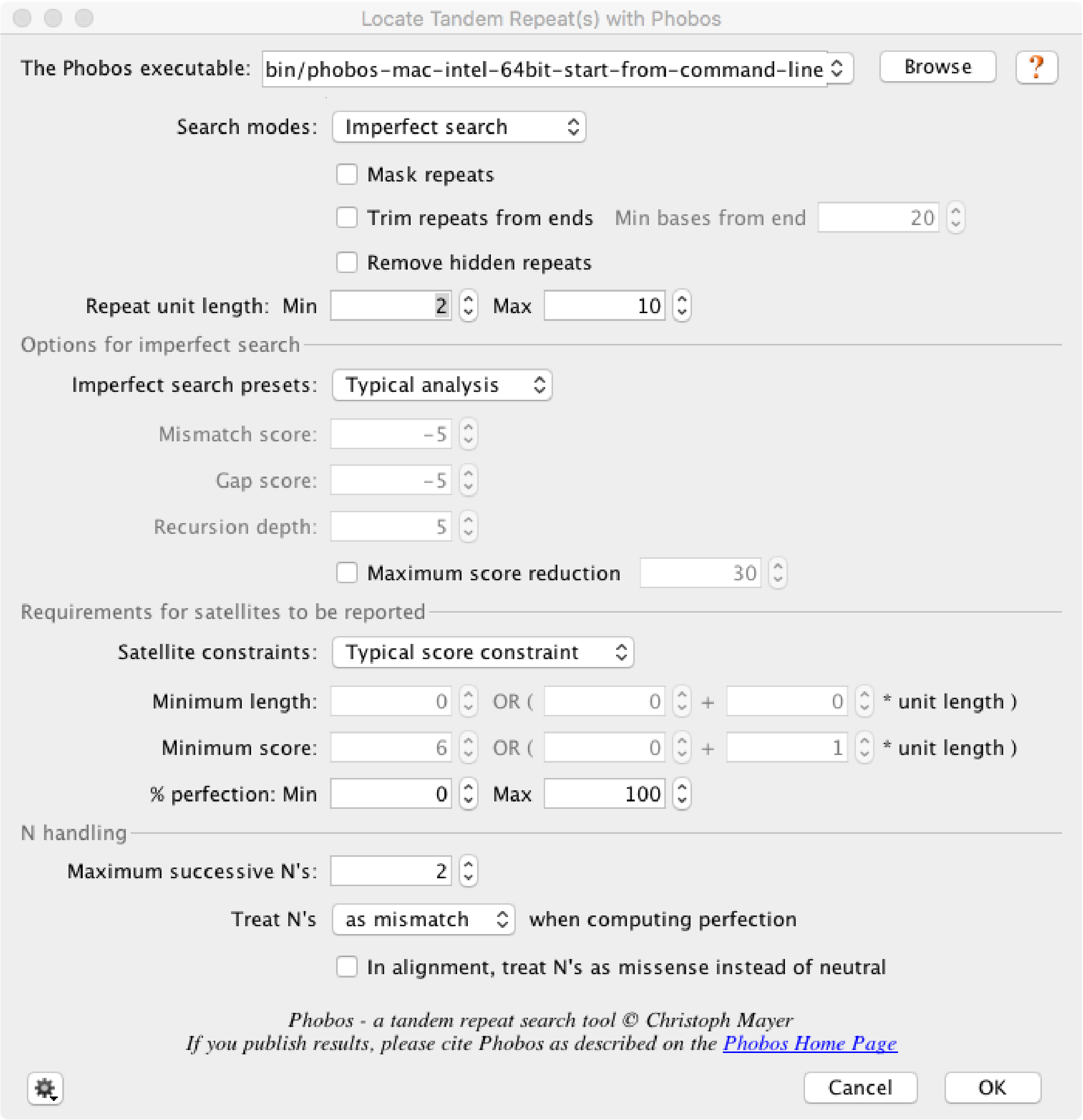
Geneious was used to as wrapper to for running sequence mappers, phobos, and sequence

Download Genome and BAM Alignment of nine CCS libraries to CP018802 (H somni) from NCBI

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 potentail 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 chose the the repeat pattern that comes first in the alphabet.



Locate Tandem Repeat(s) with Phobos

The Phobos executable: bin/phobos-mac-intel-64bit-start-from-command-line

Search modes: Imperfect search

□ Mask repeats

□ Trim repeats from ends   Min bases from end 20

□ Remove hidden repeats

Repeat unit length: Min 2 Max 10

Options for imperfect search

Imperfect search presets: Typical analysis

Mismatch score: -5

Gap score: -5

Recursion depth: 5

□ Maximum score reduction 30

Requirements for satellites to be reported

Satellite constraints: Typical score constraint

Minimum length: 0 OR (0 + 0 * unit length)

Minimum score: 6 OR (0 + 1 * unit length)

% perfection: Min 0 Max 100

N handling

Maximum successive N's: 2

Treat N's as mismatch 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 Interest (ROI)

5

Create BAM file of CCS overlapping SSR ROI using coordinates identified in step 4 and transfered to their respective BED file to be used in combination with with the BAM file of all reads mapping to the genome. When specifiying position of SSR, allow for 5 bp on each flank. Please keep in mind the BED file convention, the left coodrinate 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

```
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
```

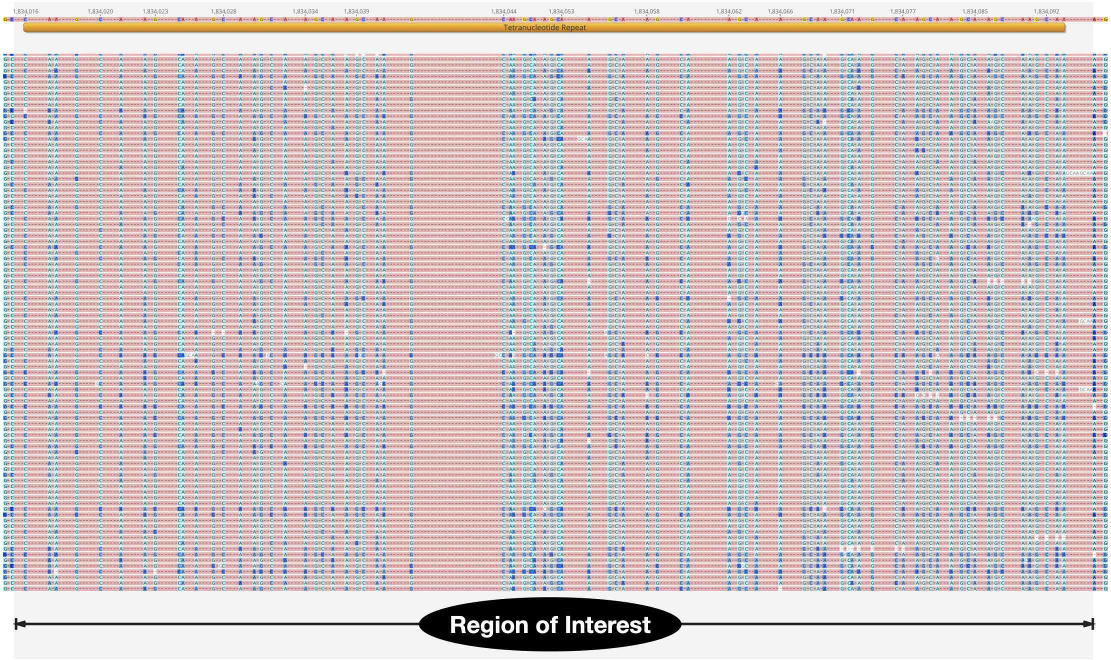
find CCS that completely overlap 79 bp AAGC SSR including 5 bp adjacent non-SSR region on each flank

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
SRR8080935_CP018802_250bp.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)

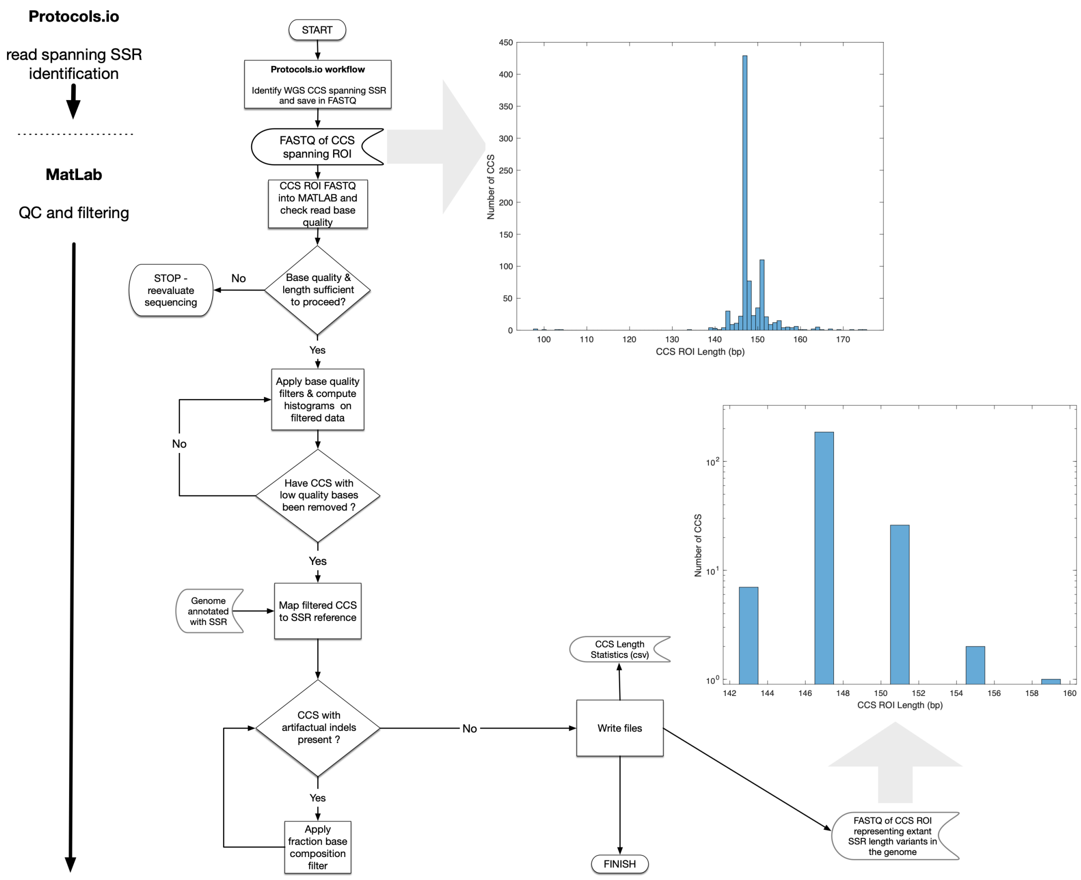
□ **SRR8080935_79bp_AAGC_SSR_ROI_reads.fastq** □ **SRR8080935_107bp_ACTG_SSR_ROI_reads.fastq**

□ **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 reads representing extant molecules

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
START
Protocols.io workflow
Identify WGS CCS spanning SSR and save in FASTQ
FASTQ of CCS spanning ROI
CCS ROI FASTQ into MATLAB and check read base quality
Base quality & length sufficient to proceed?
No
STOP - reevaluate sequencing
Yes
Apply base quality filters & compute histograms on filtered data
Have CCS with low quality bases been removed ?
No
Yes
Genome annotated with SSR
Map filtered CCS to SSR reference
CCS with artifactual indels present ?
Yes
Apply fraction base composition filter
No
Write files
CCS Length Statistics (csv)
FINISH
FASTQ of CCS ROI representing extant SSR length variants in the genome
Number of CCS
0
50
100
150
200
250
300
350
400
450
100
110
120
130
140
150
160
170
CCS ROI Length (bp)
Number of CCS
10^0
10^1
10^2
142
144
146
148
150
152
154
156
158
160
CCS ROI Length (bp)