



Nov 20,  
2019

## Steps to Create FASTQ of CCS Overlapping Control SSR - CCS ROI V.6

Gregory Harhay<sup>1</sup>

<sup>1</sup>United States Department of Agriculture

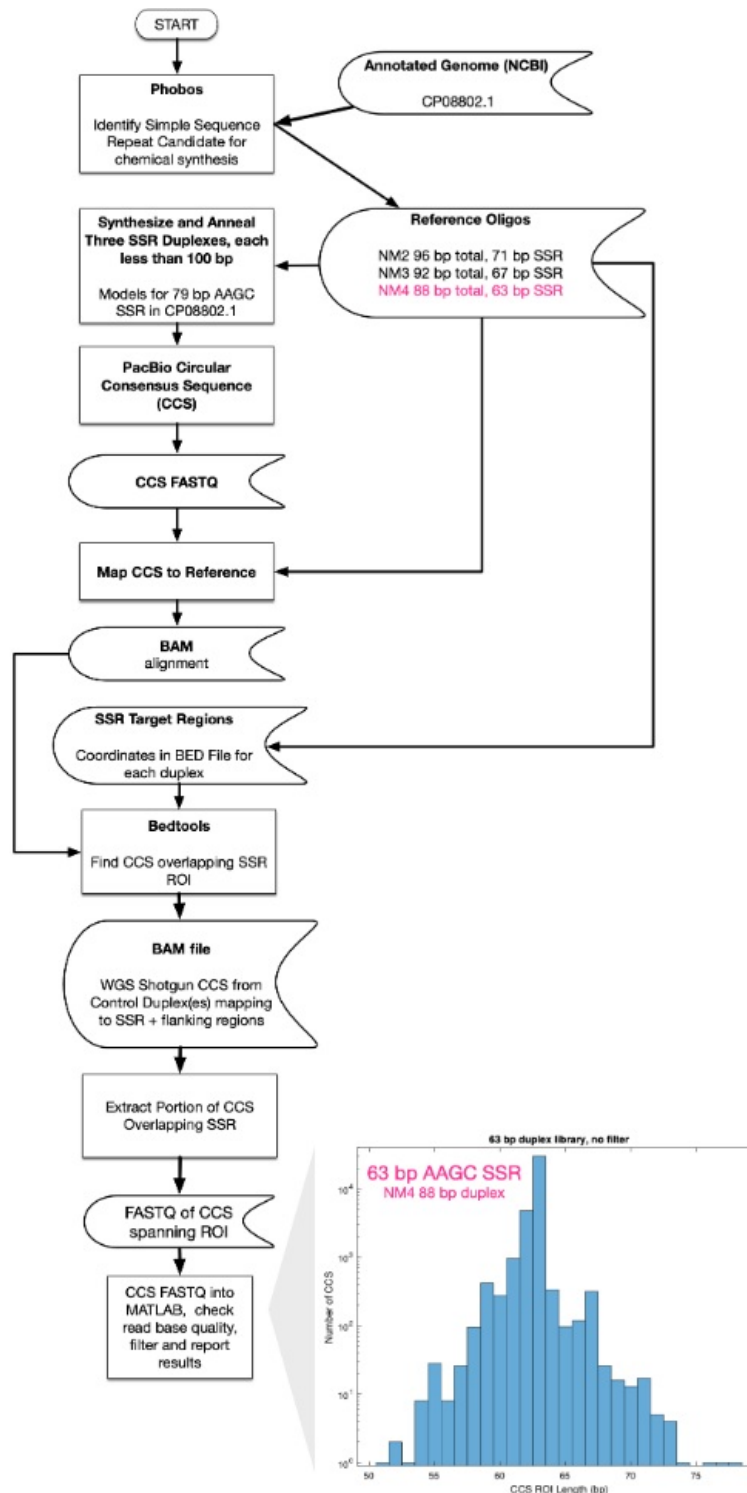
1 Works for me dx.doi.org/10.17504/protocols.io.9i6h4he

 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.


In this protocol we validate the computational approaches presented here by sequencing chemically synthesized oligos and annealed to a form duplexes. These oligos are slightly shorter version of the AAGC SSR found in CP018802.1.




Protocol Workflow for Creating FASTQ of CCS ROI for Control SSR


## Software Requirements


1

**Bedtools 2.27.1** [↗](#)  
[source](#) by <http://quinlanlab.org>

**MatLab R2018b** [↗](#)

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 for running sequence mappers, phobos, and sequence extraction

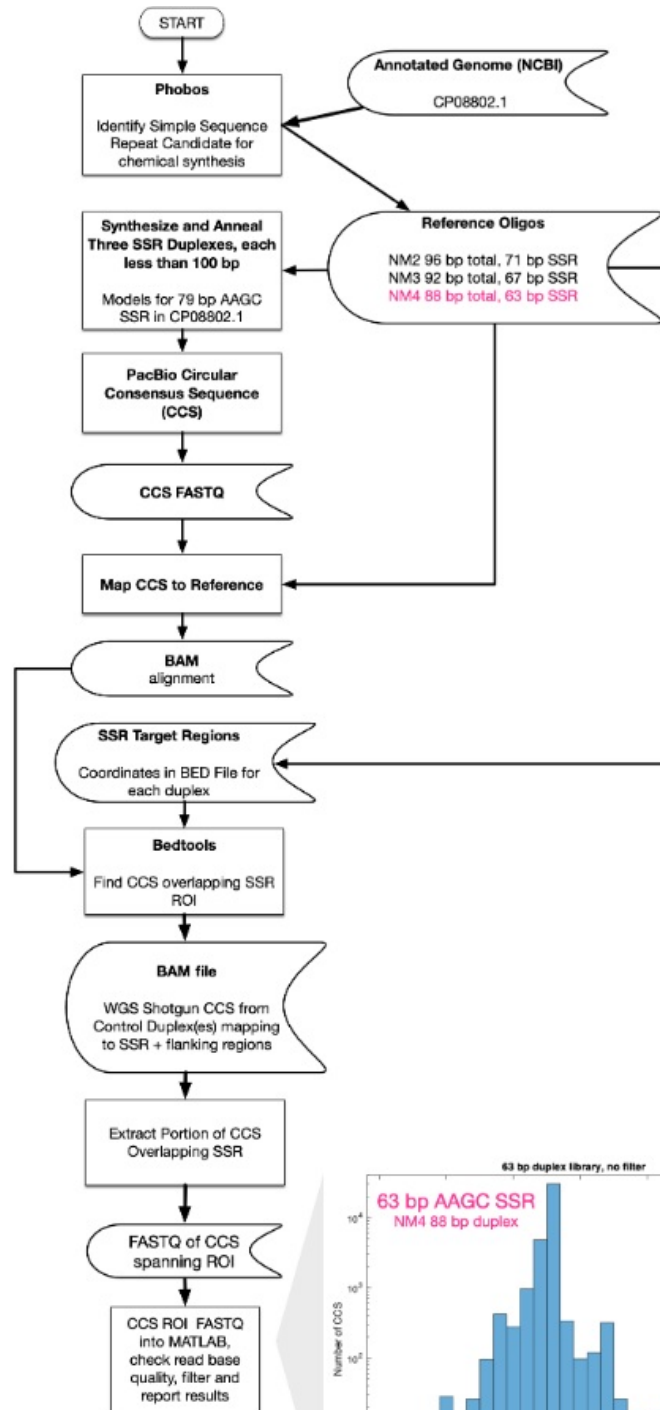
## Workflow Summary

2

**Protocols.io**

CCS Spanning SSR

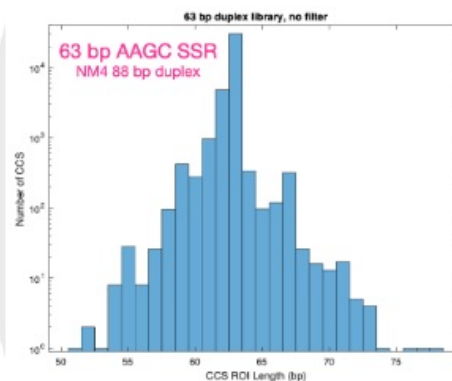
*Identification,  
selection, and writing  
to FASTQ file*



**MatLab**

CodeOcean or local instance

QC, filtering, and results



- Download from GenBank manually, Geneious, or MatLab function

<https://www.ncbi.nlm.nih.gov/nuccore/CP018802>

## Run Phobos to Identify SSR and Define SSR target to Synthesize

- 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](#)*

Repeat Unit	Minimum	Maximum	Length	Percentage Perfection	Repeat Class
AACC	1,792,217	1,792,466	250	100.000%	tetranucleotide
AATC	1,452,562	1,452,715	154	100.000%	tetranucleotide
ACTG	1,501,321	1,501,467	147	100.000%	tetranucleotide
ACTG	1,456,013	1,456,119	107	100.000%	tetranucleotide
AAGC	1,834,016	1,834,094	79	100.000%	tetranucleotide

Because its length was the largest tetranucleotide SSR below the 100 bp, the AAGC SSR locus was selected as the basis to synthesize the SSR control duplex.

## Aggregate PacBlo Circular Consensus Sequence (CCS)

- 6 Used circular consensus sequence CCS from PacBio RSII

## Control SSR Libraries

Control SSR Libraries based on the 79 bp AAGC tetranucleotide SSR found spanning 1834016 - 1834094bp in CP018802. Each oligo had 14 bp 5' flanking region and 11 bp 3' flank, with the flanking regions identical to those found in CP018802. In total, six oligos were chemically synthesized and annealed into 3 duplexes, with 3 bp overhands on the 5'-end to facilitate PacBio CCS library creation.

88 bp duplex, **63 bp SSR**, **NM4** ( 4 AAGC repeat units removed from 79 bp genomic)

5'- GACTAAAAATCAGCCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGCAAGTGTAGATCA -3'  
3'- ATTTTAGTCGGGTCGTTCTCGTTCGTTCTCGTTCGTTCTCGTTCGTTCTCGTTCGTTCTCGTTCGTTCTCGTTACATCTAGTGAA -5'

92 bp duplex, **67 bp SSR**, **NM3** (3 AAGC repeat units removed from 79 bp genomic)

[illegible]

96 bp duplex, **71 bp SSR, NM2** (2 AAGC repeat units removed from 79 bp genomic)

[illegible]

### Map CCS to Control SSR

- 7 For Single Control Duplex - 88 bp total length containing a 63 bp AAGC SSR (NM4)

SSR Reference Mapping Parameters: Geneious Assembler, Medium Sensitivity/Fast, iterate up to 5 times, Map multiple best matched randomly

Input File of Raw CCS is

☐ Control\_Single\_Duplex\_63bp\_SSR\_L\_23088\_raw.fastq

Map to Reference

Run : ☒ On my computer ☐ On Geneious Server

Data

Reference Sequence: CP018802\_SSR\_79bp\_AAGC\_Nm4 - Choose... ?

Control\_Single\_Duplex\_63bp\_SSR\_raw will be mapped to CP018802\_SSR\_79bp\_AAGC\_Nm4

☐ Assemble by: 1st part of name, separated by - (Hyphen)

☐ Assemble each sequence list separately

Method

Mapper: Geneious ?

Sensitivity: Medium Sensitivity / Fast ?

☐ Find structural variants, short insertions, and deletions of any size ?

☐ Find short insertions and large deletions up to 20 bp

Fine Tuning: Iterate up to 5 times ?

Memory Required: 84 MB of 51 GB

Note: Paired reads can be set up or changed using Sequence > Set Paired Reads

Trim Before Mapping

☐ Use existing trim regions

☐ Remove existing trim regions from sequences

☐ Trim sequences Options (modified)

☒ Do not trim

Results

Assembly Name: {Reads Name} assembled to {Reference Name} ...

☒ Save assembly report

☒ Save list of unused reads

☐ Save list of used reads ☐ Include mates

☒ Save in sub-folder

☒ Save contigs

☒ Save consensus sequences Options (modified)

Advanced

☐ Minimum mapping quality: 10

Map multiple best matches: Randomly

☒ Trim paired read overhangs

☐ Only map paired reads which map nearby

Minimum support for structural variant discovery: 1 reads ☐ Include insertions in structural variants

To specify any of the settings below, choose the Custom Sensitivity method

☒ Allow Gaps

Maximum Per Read: 15 %

Maximum Gap Size: 50

☐ Minimum Overlap: 25

☐ Minimum Overlap Identity: 80 %

Word Length: 14

Index Word Length: 12

☒ Ignore words repeated more than 20 times

Maximum Mismatches Per Read: 30 %

Maximum Ambiguity: 4

☒ Accurately map reads with errors to repeat regions

☐ Search more thoroughly for near matching reads

Fewer Options

Cancel OK

- Export alignment to BAM file



☐ `Control_Single_Duplex_63bp_SSR_L_23088_raw_map_AAGC_Nm4.bam`  
☐ `Control_Single_Duplex_63bp_SSR_L_23088_raw_map_AAGC_Nm4.bam.bai`

**For Three Control Duplexes - NM2 96 bp total, 71 bp SSR; NM3 92 bp total, 67 bp SSR; NM4 88 bp total, 63 bp SSR**

Input: ☐ `Control_Three_Duplexes_SSR_L_23089_raw.fastq`

Run Geneious Assembler with same parameters as with single duplex mapping job, using the NM3 sequence as the reference, to create the following BAM alignment files



☐ `Control_Three_Duplexes_L_23089_raw_map_AAGC_Nm3.bam.bai`  
☐ `Control_Three_Duplexes_L_23089_raw_map_AAGC_Nm3.bam`

Run bedtools to identify CCS spanning the control SSR

8

**For CCS from sequencing library consisting of single control duplex, 63 bp AAGC SSR (NM4)**

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. 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_SSR_79bp_AAGC_Nm4.bed` and `Control_Single_Duplex_63bp_SSR_L_23088_raw_map_AAGC_Nm4.bam`  
 (generated in previous step)



**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 63 bp AAGC SSR (NM4) including 5 bp adjacent non-SSR region on each flank



☐ `Control_Single_Duplex_63bp_SSR_L_23088_raw_map_intersect_AAGC_Nm4.bam`

**For sequencing library consisting of three control duplexes, 63 bp AAGC SSR in (NM4) + 67 bp AAGC SSR (NM3) + 71 bp AAGC SSR (NM2)**



Test to see if approach can recover the different length constituents from mixture of SSR lengths present in sample. For selecting CCS completely overlapping the NM3 67 bp SSR + flanking region, use bed to define coordinates.

Input

and

Control\_Three\_Duplexes\_L\_23089\_raw\_map\_AAGC\_Nm3.bam (generated in previous step)



```
bedtools intersect -a Control_Three_Duplexes_L_23089_raw_map_AAGC_Nm3.bam  
-b CP018802_SSR_79bp_AAGC_Nm3.bed -F 1.0 -wa >
```

```
Control_Three_Duplexes_L_23089_raw_map_intersect_AAGC_Nm3.bam
```

Find CCS that completely overlap 67 bp AAGC SSR (NM3) including 5 bp adjacent non-SSR region on each flank

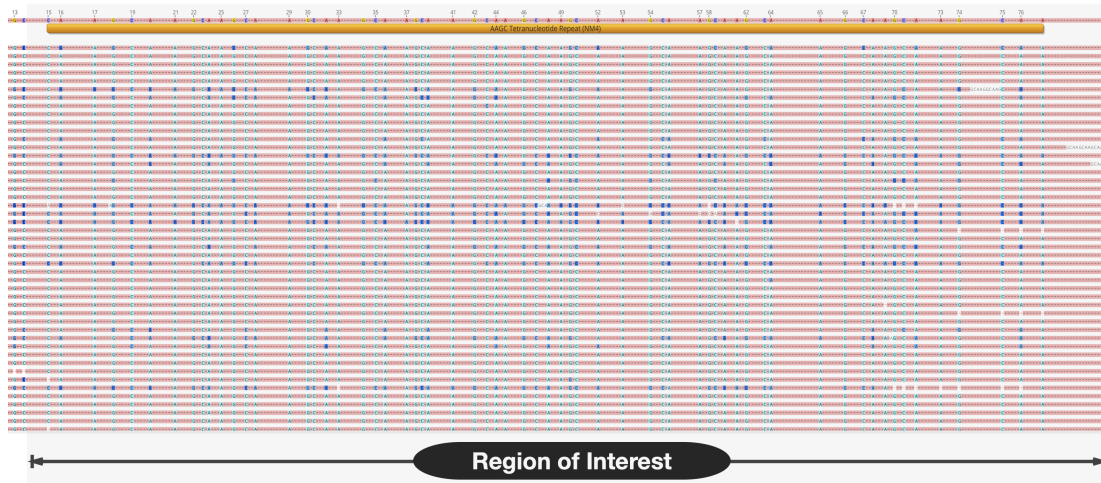


#### Analyze CCS Spanning AAGC SSR and Remove Flanking Sequence

9

#### For CCS sequencing library derived from single spanning 63 bp AAGC SSR (NM4)

- 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 CCS 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.



- Write the portion of each CCS within the ROI to FASTQ file

**Similar analysis performed for sequencing library consisting of three control duplexes, 63 bp AAGC SSR in (NM4) + 67 bp AAGC SSR (NM3) + 71 bp AAGC SSR (NM2) mapping to NM3**

- Write the portion of each CCS within the ROI (NM3) to FASTQ file

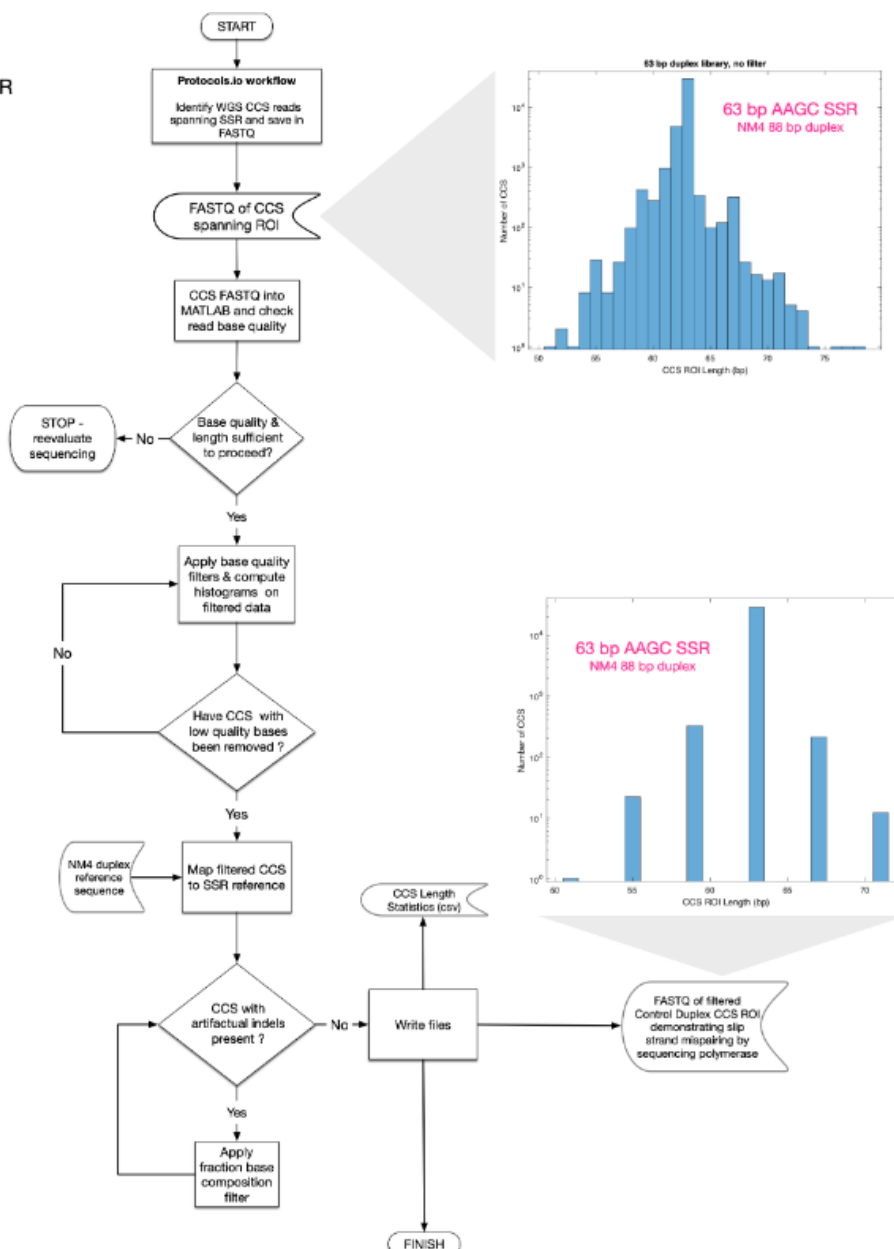
Move CCS within ROI FASTQ to Matlab Compute environment

- Use either Matlab Compute Capsule at Code Ocean or use local environment. The steps outline below for the single NM4 control duplex with a 63 bp SSR.

**Protocols.io**  
read spanning SSR  
identification



**MatLab**  
QC and filtering



This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited