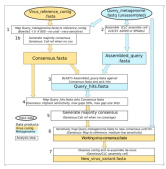


Jul 23, 2020

Gapclosing metagenomic islands

Chris Bellas[1]

[1]University of Innsbruck

1 Works for me dx.doi.org/10.17504/protocols.io.bix8kfrw

Chris Bellas
University of Innsbruck

ABSTRACT

This protocol allows gapclosing of metagenomic islands which occur when mapping metagenomic reads to a phage reference genome. These gaps arise because multiple gene variants co-exist for specific points on the genome, these variants fluctuate in dominance over space and time. This means mapping metagenomic reads from one location onto a phage reference genome from another will often result in multiple metagenomic islands. These gaps can be closed with the following protocol, allowing the other gene variants present in a phage pan-genome to be detected.

The protocol relies on detection of the same phages in similar ecosystems and relies on the phage being sequenced to sufficient depth in each sample to allow metagenomic assembly of large contigs (>10 X coverage).

Prerequisites:
Bowtie2
Geneious software package
SPAdes assembler
Blast+ tools

DOI

dx.doi.org/10.17504/protocols.io.bix8kfrw

PROTOCOL CITATION

Chris Bellas 2020. Gapclosing metagenomic islands. **protocols.io** dx.doi.org/10.17504/protocols.io.bix8kfrw

CREATED

Jul 23, 2020

LAST MODIFIED

Jul 23, 2020

PROTOCOL INTEGER ID

39648

1

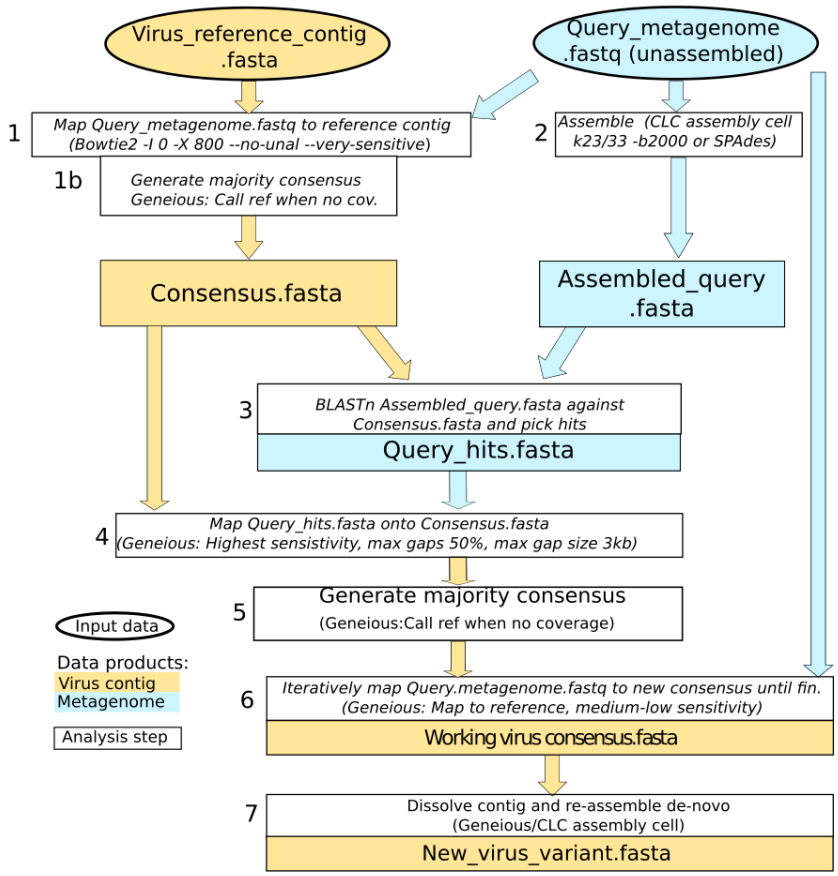


Fig 1 Gapclosing metagenomic islands workflow

1) Map metagenomic reads to virus reference contig

Starting with an assembled phage contig (phage_reference_contig.fasta).

i) Map trimmed, unassembled reads from a different metagenome (Query metagenome) to this reference using Bowtie2

Bowtie2 -I 0 -X 800 --no-unal --very-sensitive -1 Query_metagenome_R1.fastq -2 Querymetagenome_R2.fastq -S samfile.sam

(-I 0 and -X 800 refer to the min and max size of the paired end reads+insert before they are flagged as discordant. --no-unal keeps the files small and discards all reads that do not map to the contig. -1 and -2 are the forward [R1] and reverse [R2] reads from the metagenome)

1b) Import samfile into Geneious

i) *File> Import> From file <samfile.sam>*

Metagenomic islands should be visible as gaps in the read mapping where reads possess <90% identity to the reference. The missing reads are present in the metagenome, we have to find them without a reference.

Citation: Chris Bellas (07/23/2020). Gapclosing metagenomic islands. https://dx.doi.org/10.17504/protocols.io.bix8kfrw

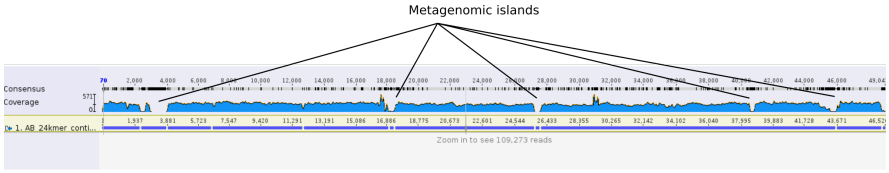


Fig .2 Metagenomic islands in a phage genome

ii) Generate a majority consensus sequence
Right click on the filename > Generate consensus sequence

Settings:
0% Majority (this selects the most frequently occurring base at a specific genome position)
If no coverage call: Ref (the calls the reference genome through the MGI gaps)
Do not trim the sequence.

<Consensus.fasta> file will be generated for this metagenome. MGI gaps will be plugged with the reference genome temporarily.

iii) Export the consensus as a fasta file
File > Export > Selected Documents

<Consensus.fasta> is the exported fasta file.

2 Assemble the query metagenome *<Query_metagenome.fastq>*

Using SPAdes, CLC or another favorite assembler:

i) Assemble the metagenomic reads into contigs.
(This should allow large sections of the target genome to be assembled in the query metagenome. Many of these contigs should cross and include the MGI regions shown in Fig. 2. We next need to find them).

Output: *<Assembled_query.fasta>* (Fasta contigs from the metagenomic assembly)

3 Locate assembled contigs that belong the target virus genome

Using blast we will find virus contigs which might belong to the target virus from Fig. 2.

i) Blast all metagenomic contigs against the virus consensus and pick matching reads.
We will only pick matching reads over 150bp. The following shell script can be copied and pasted into a file Phising.sh

Use the command *./Phising.sh Consensus.fasta Assembled_query.fasta (replaces with your filenames)*
Use the following to make it executable: *chmod +x Phising.sh*

#Phising.sh
#Blasts denovo assembly (Assembled_query.fasta) against virus genome (Consensus.fasta) then picks all results >150bp as fasta files (E-value cut off 10-5).

#Usage ./Phishing.sh <Consensus.fasta> <Assembled_query.fasta>

makeblastdb -dbtype nucl -in $1

blastn -query $2 -db $1 -evalue 0.00001 -num_threads 4 -outfmt 6 -out $2_blastn_$1

Citation: Chris Bellas (07/23/2020). Gapclosing metagenomic islands. https://dx.doi.org/10.17504/protocols.io.bix8kfrw

#Generate results list for matches
awk '$4>=150 {print $1"@"}' $2_blastn_$1 >$2_blastn_$1_phishing_list

#Pick matching fasta entires from Assembled_query.fasta
tr '\n' '@' <$2 | sed 's/>/\n>/g' | grep -f $2_blastn_$1_phishing_list| tr '@' '\n' >$2_blastn_$1_phishes

#Remove ambigous bases (N's) from matching contigs as they hinder future alignment
tr -d "N" <$2_blastn_$1_phishes >$2_blastn_$1_phishes_NoN.fasta

*#*phishes_NoN.fasta is file of matching fasta records*

4 Align assmbled contigs onto the virus reference genome

i) Import <**phishes_NoN.fasta*> into Geneious. *File > Import > From file*

ii) In Geneious, select both *Consensus.fasta* and **phishes_NoN.fasta* then *click "Map to Reference"* and ensure Consensus.fasta is set as reference.

Settings:
Set Sensitivity to Custom Sensitivity
Fine Tuning: None (fast / read mapping)

In the Advanced settings set:
Allow gaps, Maximum Per Read: 50 %,
Word Length: 10
Maximum Mismatches Per Read: 50%
Accurately map reads with errors to repeat regions: <Select yes>
Map multiple best matches: Randomly
Maximum Gap Size: 4000
Index Word Length: 10
Maximum Ambiguity: 16

Citation: Chris Bellas (07/23/2020). Gapclosing metagenomic islands. https://dx.doi.org/10.17504/protocols.io.bix8kfrw

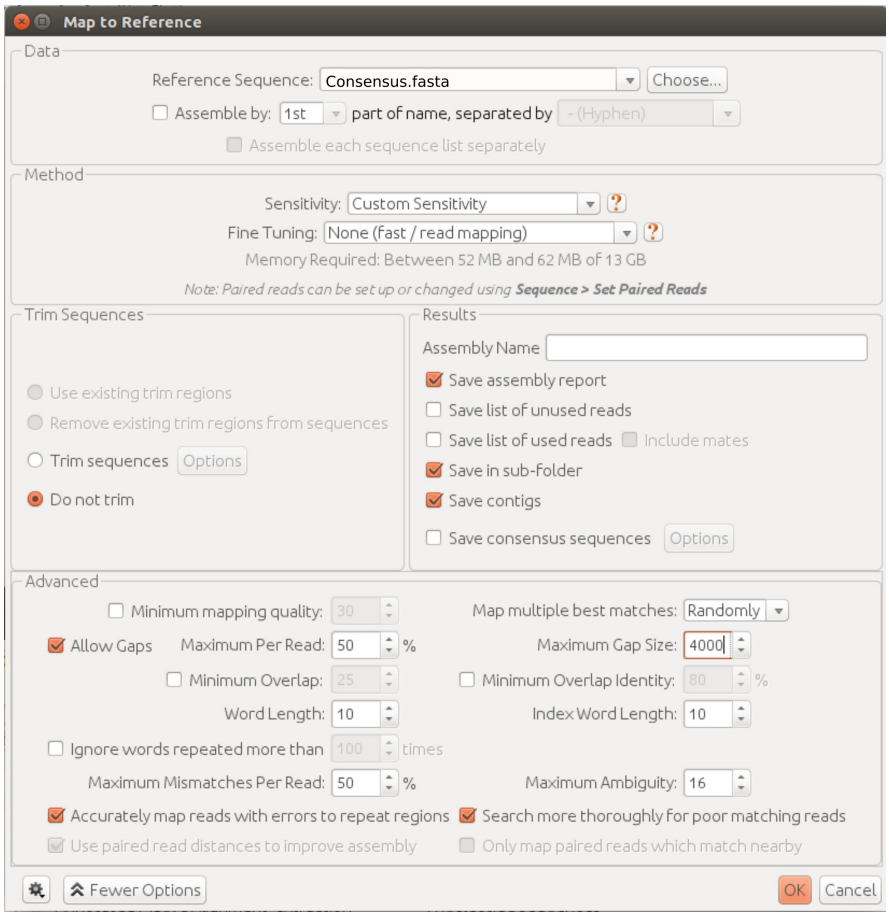


Fig .3

5 Generate majority consensus from resulting mapping file

i) Geneious: *Right click on alignment > General consensus*

Settings:
0% Majority (Calls majority base for each position)
If no coverage call: Reference (calls reference genome if still no coverage)
Do not trim sequence

This will generate the majority consensus (most frequently occurring base for each location in the genome) and cell the original reference if there is still no coverage.

Output *<working_virus_consensus.fasta>*

6 Iteratively map unassembled reads back to *working_virus_consensus.fasta* to close any leftover gaps

i) Import unassembled forward and reverse reads into Geneious
File > Import > Query_metagenome_R1.fastq
File > Import > Query_metagenome_R2.fastq

ii) Link the paired reads in Geneious
Sequence > Set paired reads

Citation: Chris Bellas (07/23/2020). Gapclosing metagenomic islands. https://dx.doi.org/10.17504/protocols.io.bix8kfrw

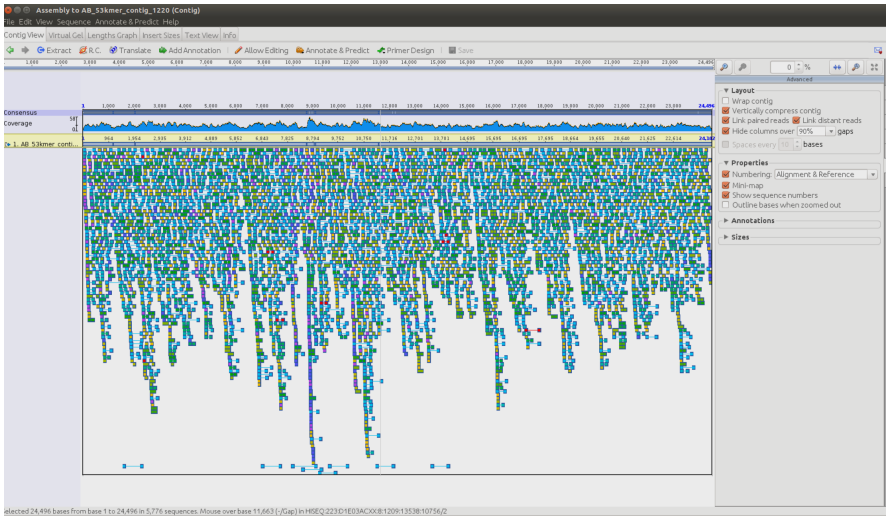
Select the imported paired metagenomic reads and *working_virus_consensus.fasta* then:
Align/Assemble > Map to reference

Settings:
Reference: working_virus_consensus.fasta
Sensitivity: Medium-Low Sensitivity / Fast
Fine Tuning: Iterate up to 25 times (or more until gaps are closed)

This will map all metagenomic reads to the contig, extending any ambiguous regions, then iterate the process.

7 Extract reads and reassemble the contig *de novo*

i) View the read mapping file in Geneious and click in the mapped reads to highlight one.
Press *Ctrl A* to select all reads



Fig,4 Export mapped reads. MGI gaps have been filled by iterative read mapping but broken read pairs and a few framshifts remian. Performing a *de novo* assembly with the extracted reads corrects this and is a check on the accuracy of the reconstruction.

ii) *Right click* (on the read mappings) > *Extract Regions*
Confirm you want to extract them as a list of sequences.
This generates a file with all mapped reads, paired and unpaired, and two consensus genomes.

iii) Delete the consensus and reference genomes to leave only the unassembled reads

ii) *De novo* assemble the contigs agian (Either in Geneious or export the reads to a favorite assembler)

Align/Assemble > De novo Assemble:

Settings:
Assembler: Geneious
Sensitivity: High Sensitivity

The virus genome should be reassembled into 1 or more large contigs which should read accross the MGI regions and allow the new gene variants to be detected.

8 Align whole genomes using Mauve/MUSCLE to compare MGI regions.

Or extract MGI regions manually and align.

Citation: Chris Bellas (07/23/2020). Gapclosing metagenomic islands. https://dx.doi.org/10.17504/protocols.io.bix8kfrw