Sep 02, 2022

De-novo assembly of Xanthomonas genomes from Illumina NovaSeq reads

DOI
https://dx.doi.org/10.17504/protocols.io.kxygxzrqzv8j/v1

David J Studholme[1], Jamie Harrison[1]

[1]University of Exeter

David J Studholme
University of Exeter

DOI: https://dx.doi.org/10.17504/protocols.io.kxygxzrqzv8j/v1

Protocol Citation: David J Studholme, Jamie Harrison 2022. De-novo assembly of Xanthomonas genomes from Illumina NovaSeq reads. **protocols.io** https://dx.doi.org/10.17504/protocols.io.kxygxzrqzv8j/v1

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

Created: August 19, 2022

Last Modified: September 02, 2022

Protocol Integer ID: 68908

Keywords: xanthomonas genome sequence, novo assembly of xanthomona, xanthomona, sequencing data, raw sequence read, quality control of the raw sequence read, read genomic shotgun, novo assembly, alignment of read, illumina novaseq, preliminary assembly, assembly

Funders Acknowledgements:
UKRI
Grant ID: BB/T010916/1

Disclaimer

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to **protocols.io** is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with **protocols.io**, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Abstract

This protocol describes the *de-novo* assembly of *Xanthomonas* genome sequences from short-read genomic shotgun sequencing data. It includes quality control of the raw sequence reads, assembly and finally polishing of the assembly based on alignment of reads against the preliminary assembly.

1 Software pre-requisites.

Note

This protocol assumes that you have already installed fastp, SPAdes, SAMtools, BowTie2 and Pilon. I also assumes that the paired Illumina sequence data comprises two gzipped FASTQ files called *name_r1.fq.gz* and *name_r1.fq.gz*.

2 Perform quality-based filtering and adapter trimming using fastp.

Command

Creates a directory for the fastp QC report files.

new command name

```
mkdir name_fastp_out
```

Command

Generates trimmed and filtered sequence files and QC reports on the Illumina NovaSeq FASTQ sequence files.

new command name

```
fastp -i name_r1.fq.gz -I name_r2.fq.gz -o name_trimmed_r1.fq.gz -O
name_trimmed_r2.fq.gz --unpaired1 name_trimmed_unp.fq.gz --unpaired2
name_trimmed_unp.fq.gz -r --cut_right_window_size 5 --
cut_right_mean_quality 20 -c -l 50 -j
name_fastp_out/name_fastp_report.json -h
name_fastp_out/name_fastp_report.html
```

3 Perform *de-novo* assembly using SPAdes.

Command

Performs the de-novo assembly.

new command name

```
spades.py -1 name_trimmed_r1.fq.gz -2 name_trimmed_r2.fq.gz -s
name_trimmed_unp.fq.gz --careful --cov-cutoff auto -o name_spades_out
```

4 Polishing with Pilon

Note

This step assumes that the SPAdes assembly is contained in a file in the current working directory called *name.fasta*. It is assumed that the two trimmed-and-filtered gzipped FASTQ files are also in the current working directory. If these files are located elsewhere, then you can make symbolic links to them in the current working directory.

Command

Creates BowTie2 index files with 'name' as the prefix for their filenames.

new command name

```
bowtie2-build name.fasta name
```

Command

Performs alignment of the trimmed-and-filtered reads against the genome assembly to generate an alignment in SAM format.

new command name

```
bowtie2 -x name -1 name_trimmed_r1.fq.gz -2 name_trimmed_r2.fq.gz -S
name_vs_name.sam
```

Command

Converts the SAM-formatted file into BAM format.

new command name

```
samtools view -b -T name.fasta name_vs_name.sam -o
name_vs_name.sam.bam
```

Command

Sorts the BAM file.

new command name

```
samtools sort --reference name.fasta name_vs_name.sam.bam -o
name_vs_name.sam.bam.sorted.bam
```

Command

Indexes the sorted BAM file.

new command name

```
samtools index name_vs_name.sam.bam.sorted.bam
```

Command

Removes the intermediate files to save disk space.

new command name

```
rm name_vs_name.sam.bam $name_vs_$name.sam
```

Command

Generates a modified genome assembly based on reconciling discrepancies between assembly and aligned reads.

new command name

```
pilon --genome name.fasta --frags name_vs_name.sam.bam.sorted.bam --
output name.pilon --outdir name_pilon_out
```

Expected result

The polished genome assembly in FASTA format can be found in the Pilon output directory: *./name_pilon_out/name.pilon.fasta*

This file can now be subjected to further quality control and/or submitted to public repositories.

5 Bibliography

Citation

Chen S, Zhou Y, Chen Y, Gu J (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England).

https://doi.org/10.1093/bioinformatics/bty560 LINK

Citation

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA
(2012)
. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.
Journal of computational biology : a journal of computational molecular cell biology.

https://doi.org/10.1089/cmb.2012.0021 LINK

Citation

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H
(2021). Twelve years of SAMtools and BCFtools. GigaScience.

https://doi.org/pii:giab008.10.1093/gigascience/giab008 LINK

Citation

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2.
Nature methods.

https://doi.org/10.1038/nmeth.1923 LINK

Citation

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM
(2014)
. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.
PloS one.

https://doi.org/10.1371/journal.pone.0112963 LINK

Citations

Step 5

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
https://doi.org/10.1093/bioinformatics/bty560

Step 5

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.
https://doi.org/10.1089/cmb.2012.0021

Step 5

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools.
https://doi.org/pii:giab008.10.1093/gigascience/giab008

Step 5

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2.
https://doi.org/10.1038/nmeth.1923

Step 5

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.
https://doi.org/10.1371/journal.pone.0112963