

Jun 11, 2025

Pulling 18S gene from Skimming Data

DOI

<https://dx.doi.org/10.17504/protocols.io.n92ld8xoxv5b/v1>

Dakota Betz¹

¹ucsd

Rouse Lab



Dakota Betz

ucsd

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.n92ld8xoxv5b/v1>

Protocol Citation: Dakota Betz 2025. Pulling 18S gene from Skimming Data. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.n92ld8xoxv5b/v1>

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

Created: September 24, 2024

Last Modified: June 11, 2025

Protocol Integer ID: 108283

Keywords: gene from skimming data, gene, skimming data

Disclaimer

Our protocols are constantly evolving and old versions will be deleted.
The documents here are not intended to be cited in publications

Abstract

Our protocols are constantly evolving and old versions will be deleted.
The documents here are not intended to be cited in publications



- 1 First find your trimmed forward and reverse reads in the cluster. These should be labeled something like "Samplename_P1" and "Sample name_P2"
- 2 First, download BBmap here <https://sourceforge.net/projects/bbmap/>

You'll want to first interleave your trimmed files using BBMap. Leave them in fastq.gz format!

So far I've had better luck doing this step in the terminal, and then logging into the cluster for additional commands.

Drag the downloaded BBMap folder to your applications folder.

>#Using bbmap, create interleaved file of post-trimmomatic cleaned reads

```
/Applications/bbmap/reformat.sh
```

```
in1=/Users/Avery/Dropbox/Polynoidae_Genome_Skimming/A3432_Macellicephalinae_sp2/A3432_P1.fastq.gz
```

```
in2=/Users/Avery/Dropbox/Polynoidae_Genome_Skimming/A3432_Macellicephalinae_sp2/A3432_P2.fastq.gz out=A3432_Interleaved.fastq.gz
```

Basic command structure:

```
/Applications/bbmap/reformat.sh
```

```
in1=/pathto_trimmed_read_P1
```

```
in2=/pathto_trimmed_read_P2
```

```
out=interleaved_file.fastq.gz
```

This will create an interleaved file of your forward and reverse reads.

- 3 Next you'll log in to the TSCC with your given username and password. Minimap2 and Samtools is already installed in the cluster.

#Using minimap2, map out skimming reads that match to your fasta file containing a set of 18S, 28S, or H3 from the most closely related species to your targeted sample.

Preferably, you may have Sanger sequence of the targeted gene from the species, and can directly use this in order to obtain the full length of the gene. Deposit your reference sequence (.fasta) into the folder that your interleaved file is in. It is easiest to have them all in once place.

IMPORTANT: You FIRST need to initialize samtools before doing the commands below, as it won't extract your file if you don't.

It's really important to make sure you're in a familiar folder so the mapped file will be deposited into a place that's accessible. I would recommend creating a folder and the

change directory to that folder so your data is input nicely.

`cd ~/mapping_data`

First:

```
module load cpu/0.17.3
module load samtools/1.13-fd7mbdu
samtools
```

Second:

```
/tscc/nfs/home/nmkoch/Programs/minimap2/minimap2/minimap2 -ax sr
/Users/Avery/Dropbox/Polynoidae_Genome_Skimming/phylogenetic-analyses/minimap-
skims/Macellicephalinae_18S.fasta
/Users/Avery/Dropbox/Polynoidae_Genome_Skimming/A3432_Macellicephalinae_sp2/A3
432_Interleaved.fastq | samtools fastq -n -F 4 - >
A3432_Macellicephalinae_sp2_18Smapped.fq
```

Basic command structure:

It is only one space between the path to reference sequence and path to interleaved file.

```
path/to/minimap2 -ax sr /path_to_reference_sequence /path_to_interleaved_file |
samtools fastq -n -F 4 - > _sample_name_mapped_gene.fq
```

- 4 #You'll need to use Geneious [Align/Assemble→ Map to Reference] to then map the resulting file from samtools (e.g. A3432_Macellicephalinae_sp2_HH3mapped.fq), which will contain a list of skimming reads, to the fasta file you retrieved above containing sequences from the species itself or closely related species of that gene. Then you'll extract the consensus sequence resulting from this. You'll have to figure out the flanking boundaries of the gene etc. by aligning with full-length sequences of the gene from other species to find its correct start and stop point. And also will need to blast your consensus to check for contamination, etc. Check flanking boundaries by aligning your mapped file to other known and CREDIBLE sequences of related species.