


Submitting assembled genomes to NCBI

This protocol covers the submission of assembled prokaryotic genomes to NCBI as a Whole-Genome Shotgun (WGS) project.

1. If you haven't already done so, download the **fasta** file containing your decontaminated scaffolds (or scaffolds from your raw spades assembly if no decontamination was necessary) to your own computer. The **.fna** file from your prokka annotation directory is the same thing.
2. Open the **fasta** file in Sublime Text and check for any scaffolds less than 200 bp in length (the **length_X** in each scaffold's name will tell you its length). Delete any scaffolds shorter than 200 bp and save the file.
3. Open a web browser and go to <https://submit.ncbi.nlm.nih.gov/subs/genome/>. Click *Login*. If you have a Google account you can log in using that, otherwise you will need to register for an NCBI account.



The screenshot shows the NCBI login interface. At the top, it says "Sign in to NCBI" and "Last signed in from this computer via: Google". Below this, there's a "Sign in with" section with buttons for Google, NIH, and ORCID. An arrow points to the Google button with the text "click here to use a Google account". Below the "Sign in with" section, there's a "Sign in directly to NCBI" section with fields for "NCBI Username" and "Password", a "Keep me signed in" checkbox, and a "Sign In" button. Below the "Sign In" button, there are two links: "Forgot NCBI username or password?" and "Register for an NCBI account". An arrow points to the "Register for an NCBI account" link with the text "click here to make an NCBI account".

4. Click *New submission*.

Submission Portal Home **My submissions** My SRA data Templates My profile

Genome New submission ← click here

Note: To find submissions started before Feb. 3, 2014, go to the [previous version](#) of the WGS submission wizard.

ATTN: to fix or update a recent submission whose status is Queued, Processed-error or Processing, please use

- the [FIX](#) button on the existing submission
- or [email your request](#) to have the [FIX](#) button enabled for that submission.

Be sure to include the Submission ID and the reason that you need to send new files.

Do not create a new submission to fix or update an existing submission whose status is Queued, Processed-error or Processing!

Filter / Search

From date To date Status Sort by ☐ desc

Data archives

Query Search Clear

5. Select *Single Genome* and click *Continue*.

Submission Type

*** How do you want to submit your data?**

☒ **Single genome** ← click here

Manually complete a web form to describe one genome assembly and to upload its sequences.

☐ Batch/multiple genomes

Upload the "Genome Info" file ([download template ahead of time](#) or during submission), a tab-delimited text file that describes each of your genome assemblies and their attributes/metadata, plus the genome sequences (one file per genome).

Information that must be common to all genomes in the batch are:

- BioProject
- (initial) release date
- assembly type (either WGS or non-wgs, not a mix of both types)
- file type (FASTA or SQN)
- gap/Ns details
- publication information (for FASTA submissions only)
- PGAP request status (Yes/No; for prokaryotic genomes only)
- [See more details here](#)

Continue ← click here

6. Enter your personal contact details, AU Biology, and click *Continue*. You need to include an AU email address in at least one of these email fields.

★ First (given) name Middle name ★ Last (family) name
 Ian Marshall
 ★ Email (primary) Email (secondary)
 ipgmarshall@gmail.com ianpgm@bio.au.dk

be sure to use your AU email address

ⓘ At least one email should be from the organization's domain.

Group for this submission
 No group (affiliation from my personal profile)
 Create group ⓘ Allow selected collaborators to read, modify, submit and delete your submissions

★ Submitting organization Submitting organization URL ★ Department
 Aarhus University http://www.au.dk/ Department of Biology
 Phone ⓘ Fax ⓘ
 ★ Street ★ City State/Province ★ Postal code ★ Country
 Ny Munkegade 114-116 Aarhus C 8000 Denmark

7. Select the following options then continue.

General Information

BioProject
 ★ Did you already register a BioProject for this research, eg for the submission of the reads to SRA and/or of the genome to GenBank?
☐ Yes ☒ No

ⓘ The BioProject bundles the data for this research project.

BioSample
 ★ Did you already register a BioSample for this sample, eg for the submission of the reads to SRA and/or of the genome to GenBank?
☐ Yes ☒ No

ⓘ The BioSample stores the detailed metadata of the sample that was sequenced.

Release date
 ⓘ **Note:** Release of BioProject or BioSample is also triggered by the release of linked data.
 ★ When should this submission be released to the public?
☒ Release immediately following processing

☐ Release on specified date or upon publication, whichever is first

8. Enter information about your Genome Assembly.

Genome info

Genome assembly metadata

☐ Genome Assembly structured comment is in the contig .sqn file(s)

Assembly date [?](#)

★ Assembly method [?](#) ★ Version or Date program was run [?](#) Delete

SPAdes 3.12.0

[Add another assembly method](#)

Assembly name [?](#)

AI4-7

[?](#) If you have a meaningful assembly name like UCLA_Agam_2.1 ([see naming recommendations](#)), please provide it here, otherwise we will auto-generate it.

★ Genome coverage [?](#)

80

★ Sequencing Technology [?](#) Delete

Illumina MiSeq

[Add another sequencing technology](#)

9. Tell NCBI that this is a full genome, and name your submission with the full binomial name for your species followed by strain ID (e.g. “*Pseudomonas stutzeri* IC-126”), then click continue. Groups using their own strains from Microbial Physiology use a strain ID “MM2021_X” where X is your group number, otherwise choose the existing strain ID.

★ Did your sample include the full genome?

☒ Yes (even for draft genomes or if a prokaryotic genome assembly may not include plasmids)

☐ No, I deliberately selected a subset of the genome (e.g. only one chromosome of a eukaryote or only the non-repetitive regions of the genome)

★ Is this the final version? [?](#)

☒ Yes ☐ No

★ Is it a *de novo* assembly?

☒ Yes ☐ No

★ Is it an update of existing submission?

☐ Yes ☒ No

Submission title [?](#)

Verminephrobacter AI4-7

Private comments to NCBI staff [?](#)

Continue

10. Give the organism a public name the same as your submission name, and select what “relevance” your study has from the drop-down menu. You should decide on the relevance based on the background of your organism.

BioProject General Info

Required fields are marked with an asterisk (*)

Public description

Verminephrobacter AI4-7

Relevance

Evolution

External Links

Link description	URL	Delete

Add another link

- Select an appropriate submission package for your genome depending on the kind of sample it was isolated from. This will determine the kind of metadata to be included in your submission (see below).

Genome, metagenome or marker sequences (MIXS compliant)

Use for genomes, metagenomes, and marker sequences. These samples include specific attributes that have been defined by the Genome Standards Consortium (GSC) to formally describe and standardize sample metadata for genomes, metagenomes, and marker sequences. The samples are validated for compliance based on the presence of the required core attributes as described in [MIXS](#).

Cultured Bacterial/Archaeal Genomic Sequences [MIGS](#)

No environmental package

air

built

host-associated

human-associated

human-gut

human-oral

human-skin

human-vaginal

microbial mat/biofilm

Attributes

Required fields are marked with an asterisk (*)

Package MIGS: cultured bacteria/archaea, host-associated; version 5.0

Sample Name

Earthworm symbiont Verminephrobacter sp. AI3-7

Organism

Verminephrobacter sp.

strain

AI3-7

the full binomial name of your organism plus strain ID

the full binomial name of your organism

strain ID

► **Environment** 1 empty optional fields collapsed

★ **collection date** ?

2013

★ **broad-scale environmental context** ?

soil

★ **local-scale environmental context** ?

garden soil

★ **environmental medium** ?

soil

★ **geographic location** ?

Denmark: Aarhus

★ **latitude and longitude** ?

56.163 N10.204 E

12. After adding the relevant metadata, you need to specify where the strain is available (enter the name of the relevant P.I., i.e. Kai Finster) and specify the yes, we would like to have the genome annotated.

Source

Required fields are

Prokaryote source

Bacteria and/or source DNA is available from ?

Marie B Lund

★ Annotate this prokaryotic genome in the [NCBI Prokaryotic Annotation Pipeline \(PGAP\)](#) before its release?

☒ Yes ☐ No

Note that annotation is not required, but you may request to have your prokaryotic genome annotated by PGAP before it is released. The pre-release annotated genome will be posted back to this submission for you to view. If no Release date has been requested in this submission, the PGAP-annotated genome will be released upon completion of processing.

Continue

13. Now select some final options and (finally!) upload your **fasta** file from step 2. Be sure to have the instructor check your submission details before you upload!

Which of these 3 options describes this genome submission?

- ☐ 1. Each chromosome is in a single sequence and there are no extra sequences
- There can still be gaps within the sequences.
We will prompt you to provide the information for any Ns that represent gaps.
 - Internal sequences must be arranged in the correct order and orientation.
Sequences concatenated in unknown order are not allowed.
 - Plasmids and organelles can still be in multiple pieces.
 - If the sequences are assembled using an AGP file, choose the next option.
- ☒ 2. One or more chromosomes are still in multiple pieces and/or some sequences are not assembled into chromosomes
- This will be processed as a WGS genome and may include AGP files in the submission
 - There can still be gaps within the sequences.
We will prompt you to provide the information for any Ns that represent gaps.
 - Internal sequences must be arranged in the correct order and orientation.
Sequences concatenated in unknown order are not allowed.
- ☐ 3. We are submitting just the AGP file(s) for a genome assembly; the components of the AGP file are already in GenBank


Select file type for the sequences

- ☐ ASN.1 (.sqn) ☒ FASTA

[Request preload folder](#)


Select upload type

- ☐ I have all files preloaded for this submission
- ☒ I will upload all the files now via HTTP/Aspera

 Current versions of browsers Firefox, Chrome, Safari or Internet Explorer are recommended.
To upload large eukaryotic files (larger than 2GB), please use [Aspera Connect plugin](#).
Please note: in order to use Aspera for file upload with Chrome, you need to update Aspera Connect plugin to version 3.6 or newer. [More details...](#)

Upload FASTA

No file chosen

Name	Size	Created	Delete
Al3-7.fasta	5.0 MB	2/27/2019 20:26	

Do you have AGP files that assemble the individual contigs into scaffolds or chromosomes, OR assemble the submitted gapped sequences into chromosomes?

- ☐ Yes ☒ No

Assignment

★ Is any sequence a complete chromosome?

☐ Yes ☒ No

★ Does any sequence belong to a plasmid?

☐ Yes ☒ No

Continue


14. Finally, specify the authors of your submission (your group members, Ian P.G. Marshall, and Kai Finster). Under *Reference title* since there's no publication associated with this, you should give it a name that makes sense ("Draft genome sequence of *Bacillus magnificus* MM2021_2" or something along those lines).

References

Sequence authors

★ First (given) name

Marie


MI 

B

★ Last (family) name


Lund

Delete



Flatfile preview

Lund, M. B.

 Add another sequence author

Reference

★ Publication status

☒ Unpublished


☐ In-press

☐ Published

★ Reference title

Verminephrobacter genomes

Reference authors



☒ Same as sequence authors

☐ Specify new authors

Continue