# Submission Portal

HOME | MY SUBMISSIONS | GROUPS | TEMPLATES | MY PROFILE

## WGS submission: SUBXXXX

Delete submission

**Your_assembly.1.0 whole genome assembly**

1 SUBMITTER > 2 GENERAL INFO > 3 FILES > 4 GAPS > 5 ASSIGNMENT > 6 REFERENCES > 7 OVERVIEW

## Submitter

Required fields are marked with asterisk *

* First (given) name: Your
Middle name:
* Last (family) name: Name

* E-mail (primary): your@email.com
E-mail (secondary):

At least one e-mail should be from the organization's domain.

**Select group for this submission**

- None (affiliation from my personal profile)
- Your name

* Submitting organization: Your Organization
Submitting organization URL: https://yourgroup.edu

* Department: sequencing consortium

Phone:
Fax:

* Street: Street
* City: City
* State/Province: State
* Postal code: Postal code

* Country: USA

Continue | Update my contact information in profile

# WGS submission: SUBXXXX

Delete submission

## Your_assembly.1.0 whole genome assembly



## General Information

Required fields are marked with asterisk *

**BioProject**

It is strongly suggested you submit a BioSample (below) then submit a BioProject (in that order) before submitting your genome

*** 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

*** BioProject**

PRJNAXXXXX Your Sequence and Assembly Clear field

Organization: Your Organization

The BioProject bundles the data for this research project.

*** 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

*** Sample**

SAMN0XXXXXX Your Sequence and Assembly Clear field

Organism: Zea mays subsp. mays Tax ID: 12345

Submitted: 2016-02-07

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

**Release date**

*** When should this submission be released to the public:**

(•) Release immediately following processing (**recommended**)

( ) Release on specified date or upon publication, whichever is first

**Note:** Release of BioProject or BioSample is also triggered by the release of linked data.

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

**Assembly date**

2016-09

| * Assembly method | * Version or Date program was run | Delete |
|---|---|---|
| DenovoMAGIC | 3.0 | |
| | | |

Add another assembly method

**Assembly name**

Your_assemblyname.1.0

*** Genome coverage**

210.0

| * Sequencing Technology | Delete |
|---|---|
| | |
| 10x Genomics | |
| | |

Add another sequencing technology

*** 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

*** Existing genome accessions**

LWRWXXXXX

**Submission title**

Your_assemblyname.1.0 genome assembly

**Private comments to NCBI staff**

Continue

# WGS submission: SUBXXXX

Delete submission

## Your_assembly.1.0 whole genome assembly



## Files for submission

Required fields are marked with asterisk *

**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

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.

**Upload FASTA**

Browse... No files selected.

| Name | Size | Created | Delete |
|---|---|---|---|
| scaffolds.fasta | 2.0 GB | 12/22/2016 13:23 | |

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

(•) Yes ( ) No

It is strongly suggested you submit scaffold fasta files with an AGP file containing the pseudomolecule coordinates instead of the pseudomolecule fasta files. Submitting only the pseudomolecules will prevent you from updating the assembly in GenBank

**Do you have an AGP file for unplaced scaffolds (these are scaffolds without chromosome or plasmid information, so they have no genomic context)?**

( ) Yes (•) No

**Are there also AGP files that assemble chromosomes, plasmids and/or unlocalized scaffolds?**

(•) Yes ( ) No

Unlocalized scaffolds are assigned to a chromosome, organelle, or plasmid but their location on that chromosome, organelle or plasmid is not known. A single organelle or plasmid sequence that is partial is also unlocalized.

**How are the chromosomes and/or plasmids assembled?**

(•) Directly from contigs in 1 AGP file (with or without scaffold breaking gaps)

( ) Via explicit scaffolds, in 2 AGP files

( ) I have only unlocalized scaffolds

Unlocalized scaffolds have a known chromosome assignment but the location on the chromosome is not known.

**Upload "chromosomes and/or plasmids from contigs" AGP file**

genome.agp 11.1 kB 12/22/2016 13:05 Delete

**Upload "unlocalized scaffolds" AGP file**

Browse... No file selected.

**Did you annotate the scaffolds or chromosomes that are assembled in the AGP files?**

( ) Yes

(•) No, I don't have these files OR I've already uploaded annotated gapped sequences in the first step

Continue

# WGS submission: SUBXXXX

Delete submission

**Your_assembly.1.0 whole genome assembly**



## Gaps

Required fields are marked with asterisk *

The sequences contain one or more N's.

This is automatically generated by NCBI after you submit your files

Overview of the shortest, longest & most frequent runs of Ns in the submission:

| Gap length | # of runs of Ns |
|---|---|
| 10 | 18153 |
| 11 | 93 |
| 12 | 72 |
| 13 | 69 |
| 14 | 60 |
| 15 | 67 |
| 16 | 68 |
| 17 | 61 |
| 18 | 55 |

*** Did you randomly merge the sequences into a single sequence (for example, maybe you just linked the sequences together by size without using an assembler program)?**

○ Yes ◉ No

*** Appropriate minimum number of Ns in a row (0-10) that represents a gap**

10

Note that runs of 10 or more Ns will be identified as gaps when the statistics for this genome are calculated, even if '0' is chosen here. More information about the Assembly resource.

*** Do any of the N's represents gaps of completely unknown size (the gap size was NOT estimated by an assembly program and a single value, eg 100, was used)?**

- (•) Yes
- ( ) No, all gaps are of estimated size (even if a particular size was used for small gaps (eg, 10 N's))

Note that most assembly programs use estimated length gaps.

*** Are all gaps of unknown size represented by the same number of N's, eg 100?**

(•) Yes ( ) No

*** Number of N's in gap of unknown length**

10

*** What type of evidence was used to assert linkage across the assembly gaps?**

- (•) paired-ends: Paired sequences from the two ends of a DNA fragment, including mate-pairs. The most common type for simple de novo assemblies.
- ( ) align-genus: Alignment to a reference genome within the same genus.
- ( ) align-xgenus: Alignment to a reference genome within another genus.
- ( ) strobe: Strobe sequencing (eg, PacBio).
- ( ) map: Linkage asserted using a non-sequence based map such as RH, linkage, fingerprint or optical.
- ( ) align-trnscpt: Alignment to a transcript from the same species.

Much less common:

- ( ) within-clone: Sequence on both sides of the gap is derived from the same clone, but the gap is not spanned by paired-ends. The adjacent sequence contigs have unknown order and orientation.
- ( ) clone-contig: Linkage is provided by a clone contig in the tiling path (TPF). For example, a gap where there is a known clone, but there is not yet sequence for that clone.

Note: if more than one linkage evidence was used, then we cannot convert the runs of Ns appropriately, so you need to split the sequence into the separate contigs and submit a traditional wgs submission with or without an AGP file OR make a .sqn file using MakeGapTable.pl and tbl2asn

Continue

# WGS submission: SUBXXXX

Delete submission

## Your_assembly.1.0 whole genome assembly



## Assignment

Required fields are marked with asterisk *

**Warning:** Some fields on previous steps might be changed that possibly affects data entered on this page.

Reset the form

**Upload a csv file of the chromosome assignments** This step is optional

Browse... No file selected.

You can upload a csv file of the chromosome assignments for the sequences.
If all of the sequences are unlocalized, meaning that they are just part of the chromosome, then upload a 2-column table where the values are:

column 1 = sequence name (seqid)
column 2 = official chromosome name, eg 1 or I or X

Add 'yes' in column 3 to indicate any sequences that represent the full chromosome (even if gaps are present).
Add 'yes' in column 4 when the value of column 3 is 'yes' AND the biological chromosome is circular, as is the case for many prokaryotes.

Note that blank values in columns 3 and 4, and missing columns 3 or 4 all mean 'No'.

Example where two sequences belong to chromosome I and one sequence IS chromosome IV, which is a linear chromosome:

```
contig51,I
contig52,I
contig53,IV,yes
```

| * Sequence ID | Length | * Chromosome name | Circular | Delete |
|---|---|---|---|---|
| chr1 | 310925244 | 1 | ☐ | |
| chr2 | 244237062 | 2 | ☐ | |
| chr3 | 241278614 | 3 | ☐ | |
| chr4 | 254269898 | 4 | ☐ | |
| chr5 | 222590201 | 5 | ☐ | |
| chr6 | 171602414 | 6 | ☐ | |
| chr7 | 181422836 | 7 | ☐ | |
| chr8 | 182570339 | 8 | ☐ | |
| chr9 | 163066665 | 9 | ☐ | |
| chr10 | 149450367 | 10 | ☐ | |
| | | | ☐ | |

Add another chromosome

Delete all chromosomes

**Upload a csv file of the organelle assignments**

Browse... No file selected.

You can upload a csv file of the organelle assignments for the sequences.
If all of the sequences are unlocalized, meaning that they are just part of the chromosome, then upload a 2-column table where the values are:

column 1 = sequence name (seqid)
column 2 = organelle type (allowed names are in the 'Type' pulldown list)

Add 'yes' in column 3 to indicate any sequences that represent the full chromosome (even if gaps are present).
Add 'yes' in column 4 when the value of column 3 is 'yes' AND the biological chromosome is circular, as is the case for many mitochondrial and plastid chromosomes.

Note that blank values in columns 3 and 4, and missing columns 3 or 4 all mean 'No'.

Example where one sequence IS the circular mitochondrial chromosome and two sequences belong to the chloroplast chromosome:

```
contig501,mitochondrion,yes,yes
contig502,chloroplast
contig503,chloroplast
```

| * Sequence ID | Length | * Type | Complete | Circular | Delete |
|---|---|---|---|---|---|
| | | | ☐ | ☐ | |

Add another organelle

Delete all organelles

**Upload a csv file of the plasmid assignments**

Browse... No file selected.

You can upload a csv file of the plasmid assignments for the sequences.
If all of the sequences are unlocalized, meaning that they are just part of the plasmid, then upload a 2-column table where the values are:

column 1 = sequence name (seqid)
column 2 = plasmid name. Use 'unnamed' if the plasmid name is not determined. Use 'unnamed1' and 'unnamed2', etc if there are multiple plasmids whose names are not determined

Add 'yes' in column 3 to indicate any sequences that represent the full plasmid (even if gaps are present).
Add 'yes' in column 4 when the value of column 3 is 'yes' AND the plasmid is circular.

Note that blank values in columns 3 and 4, and missing columns 3 or 4 all mean 'No'.

Example where one sequence IS the circular plasmid named pMBC123, and two sequences belong to the plasmid named pMBC124:

```
contig11,pMBC123,yes,yes
contig12,pMBC124
contig13,pMBC124
```

| * Sequence ID | Length | * Plasmid name | Complete | Circular | Delete |
|---|---|---|---|---|---|
| | | | ☐ | ☐ | |

Add another plasmid  Delete all plasmids

Continue

# WGS submission: SUBXXXX

Delete submission

**Your_assembly.1.0 whole genome assembly**



## References

Required fields are marked with asterisk *

| * First (given) name | MI | * Last (family) name | Delete |
|---|---|---|---|
| Researcher | | One | |
| Researcher | | Two | |
| Researcher | | Three | |

Add another sequence author

**Flatfile preview**

```
One, R., Two, R., and Three, R.
```

*** Publication status**

(•) Unpublished ( ) In-press ( ) Published

*** Reference title**

Title of Your Paper

*** Reference authors**

Middle initials. When including more than one initial, please follow each initial with a period (for example: F.L.). Only letters are allowed as middle initials.

( ) Same as sequence

| * First (given) name | MI | * Last (family) name | Delete |
|---|---|---|---|
| Researcher | | One | |
| Researcher | | Two | |
| Researcher | | Three | |