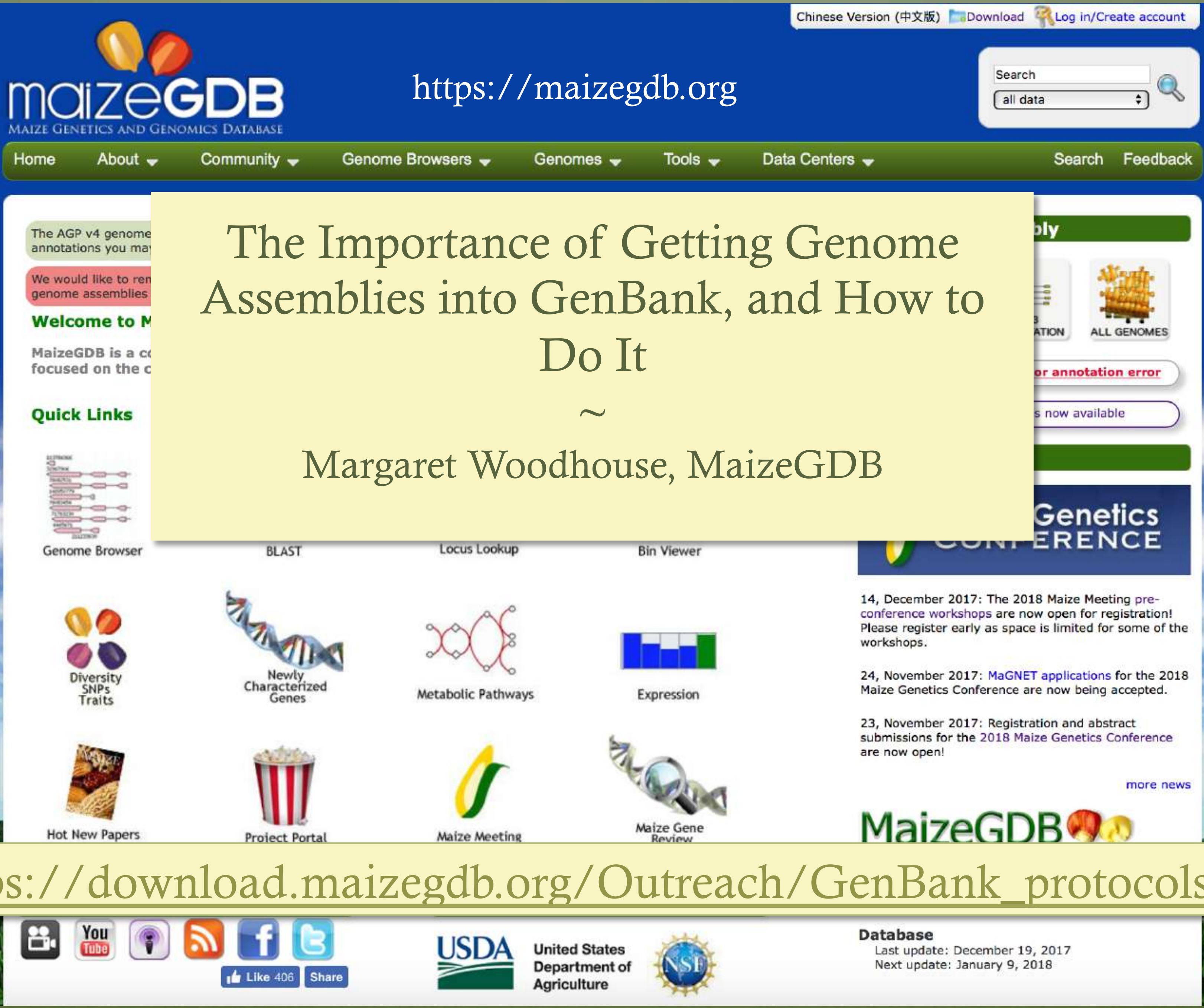
The Importance of Getting Genome Assemblies into GenBank, and How to Do It

~

Margaret Woodhouse, MaizeGDB

https://download.maizegdb.org/Outreach/GenBank_protocols/



The AGP v4 genome annotations you may

We would like to ren genome assemblies

Welcome to M

MaizeGDB is a co focused on the c

14, December 2017: The 2018 Maize Meeting pre-conference workshops are now open for registration! Please register early as space is limited for some of the workshops.

24, November 2017: MaGNET applications for the 2018 Maize Genetics Conference are now being accepted.

23, November 2017: Registration and abstract submissions for the 2018 Maize Genetics Conference are now open!

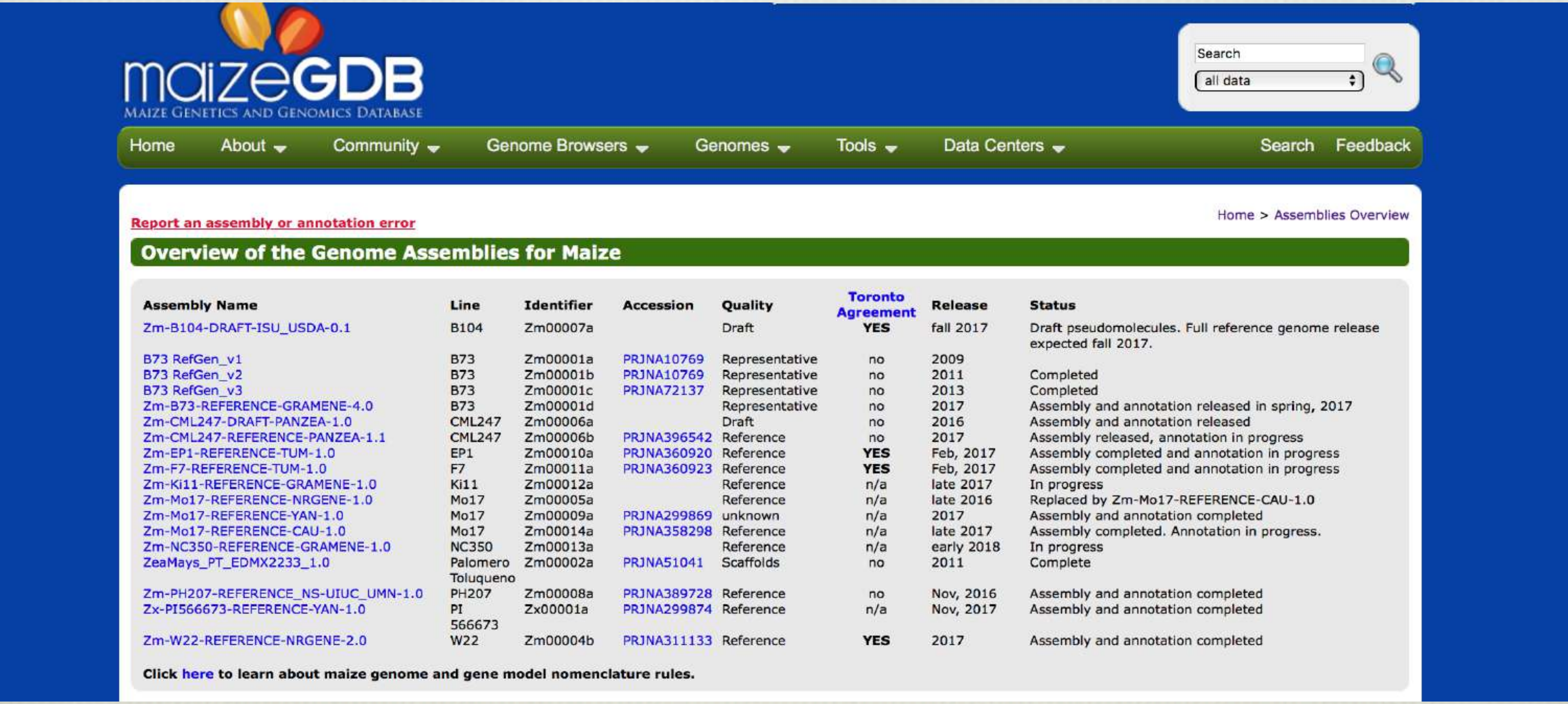
Database
Last update: December 19, 2017
Next update: January 9, 2018

maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE

- MaizeGDB was launched in 2004 to integrate maize genetic, marker, genomic, and germplasm data
- Upgraded in 2015 to meet the needs of the genomics era
- Currently hosts genomes of nine distinct maize lines, and we expect to receive no less than thirty more in the next few years

For more information visit Poster P0840

Ensuring that all these genomes are standardized in quality and have complete metadata is a necessity and a challenge



maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE
Search
all data
Home
About
Community
Genome Browsers
Genomes
Tools
Data Centers
Search
Feedback
Report an assembly or annotation error
Home > Assemblies Overview

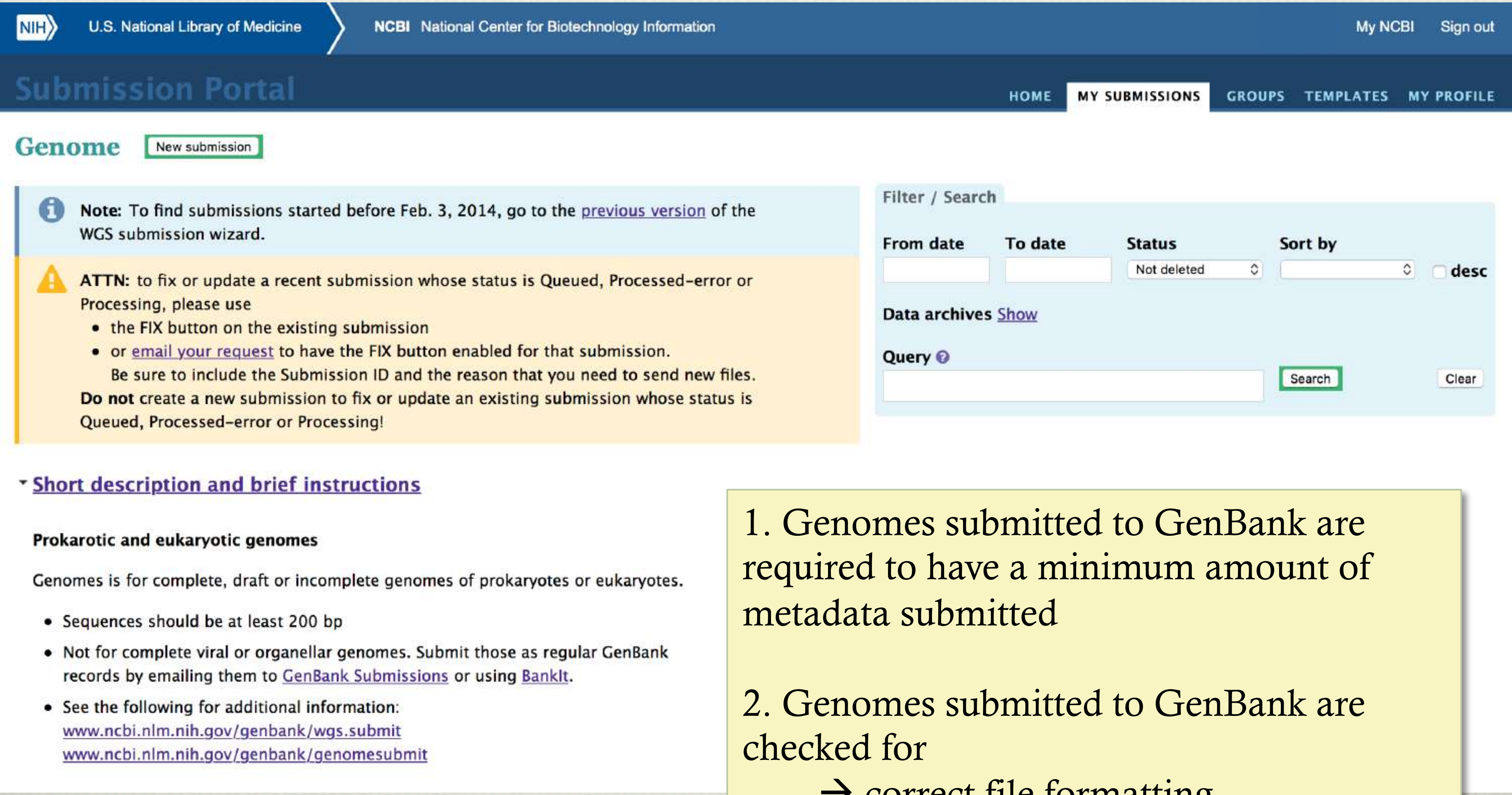
Overview of the Genome Assemblies for Maize

Assembly Name	Line	Identifier	Accession	Quality	Toronto Agreement	Release	Status
Zm-B104-DRAFT-ISU_USDA-0.1	B104	Zm00007a		Draft	**YES**	fall 2017	Draft pseudomolecules. Full reference genome release expected fall 2017.
B73 RefGen_v1	B73	Zm00001a	PRJNA10769	Representative	no	2009	
B73 RefGen_v2	B73	Zm00001b	PRJNA10769	Representative	no	2011	Completed
B73 RefGen_v3	B73	Zm00001c	PRJNA72137	Representative	no	2013	Completed
Zm-B73-REFERENCE-GRAMENE-4.0	B73	Zm00001d		Representative	no	2017	Assembly and annotation released in spring, 2017
Zm-CML247-DRAFT-PANZEA-1.0	CML247	Zm00006a		Draft	no	2016	Assembly and annotation released
Zm-CML247-REFERENCE-PANZEA-1.1	CML247	Zm00006b	PRJNA396542	Reference	no	2017	Assembly released, annotation in progress
Zm-EP1-REFERENCE-TUM-1.0	EP1	Zm00010a	PRJNA360920	Reference	**YES**	Feb, 2017	Assembly completed and annotation in progress
Zm-F7-REFERENCE-TUM-1.0	F7	Zm00011a	PRJNA360923	Reference	**YES**	Feb, 2017	Assembly completed and annotation in progress
Zm-Ki11-REFERENCE-GRAMENE-1.0	Ki11	Zm00012a		Reference	n/a	late 2017	In progress
Zm-Mo17-REFERENCE-NRGENE-1.0	Mo17	Zm00005a		Reference	n/a	late 2016	Replaced by Zm-Mo17-REFERENCE-CAU-1.0
Zm-Mo17-REFERENCE-YAN-1.0	Mo17	Zm00009a	PRJNA299869	unknown	n/a	2017	Assembly and annotation completed
Zm-Mo17-REFERENCE-CAU-1.0	Mo17	Zm00014a	PRJNA358298	Reference	n/a	late 2017	Assembly completed. Annotation in progress.
Zm-NC350-REFERENCE-GRAMENE-1.0	NC350	Zm00013a		Reference	n/a	early 2018	In progress
ZeaMays_PT_EDMX2233_1.0	Palomero Toluqueno	Zm00002a	PRJNA51041	Scaffolds	no	2011	Complete
Zm-PH207-REFERENCE_NS-UIUC_UMN-1.0	PH207	Zm00008a	PRJNA389728	Reference	no	Nov, 2016	Assembly and annotation completed
Zx-PI566673-REFERENCE-YAN-1.0	PI 566673	Zx00001a	PRJNA299874	Reference	n/a	Nov, 2017	Assembly and annotation completed
Zm-W22-REFERENCE-NRGENE-2.0	W22	Zm00004b	PRJNA311133	Reference	**YES**	2017	Assembly and annotation completed

Click here to learn about maize genome and gene model nomenclature rules.

MaizeGDB facilitates this by requiring all our genomes to be submitted to **GenBank**

Why submit to GenBank first?


NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
My NCBI
Sign out
Submission Portal
HOME
MY SUBMISSIONS
GROUPS
TEMPLATES
MY PROFILE

Genome New submission

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: Not deleted
Sort by:
□ desc

Data archives Show

Query:

Search Clear

Short description and brief instructions

Prokarotic and eukaryotic genomes

Genomes is for complete, draft or incomplete genomes of prokaryotes or eukaryotes.

- Sequences should be at least 200 bp
- Not for complete viral or organellar genomes. Submit those as regular GenBank records by emailing them to GenBank Submissions or using BankIt.
- See the following for additional information:
  www.ncbi.nlm.nih.gov/genbank/wgs.submit
  www.ncbi.nlm.nih.gov/genbank/genomesubmit

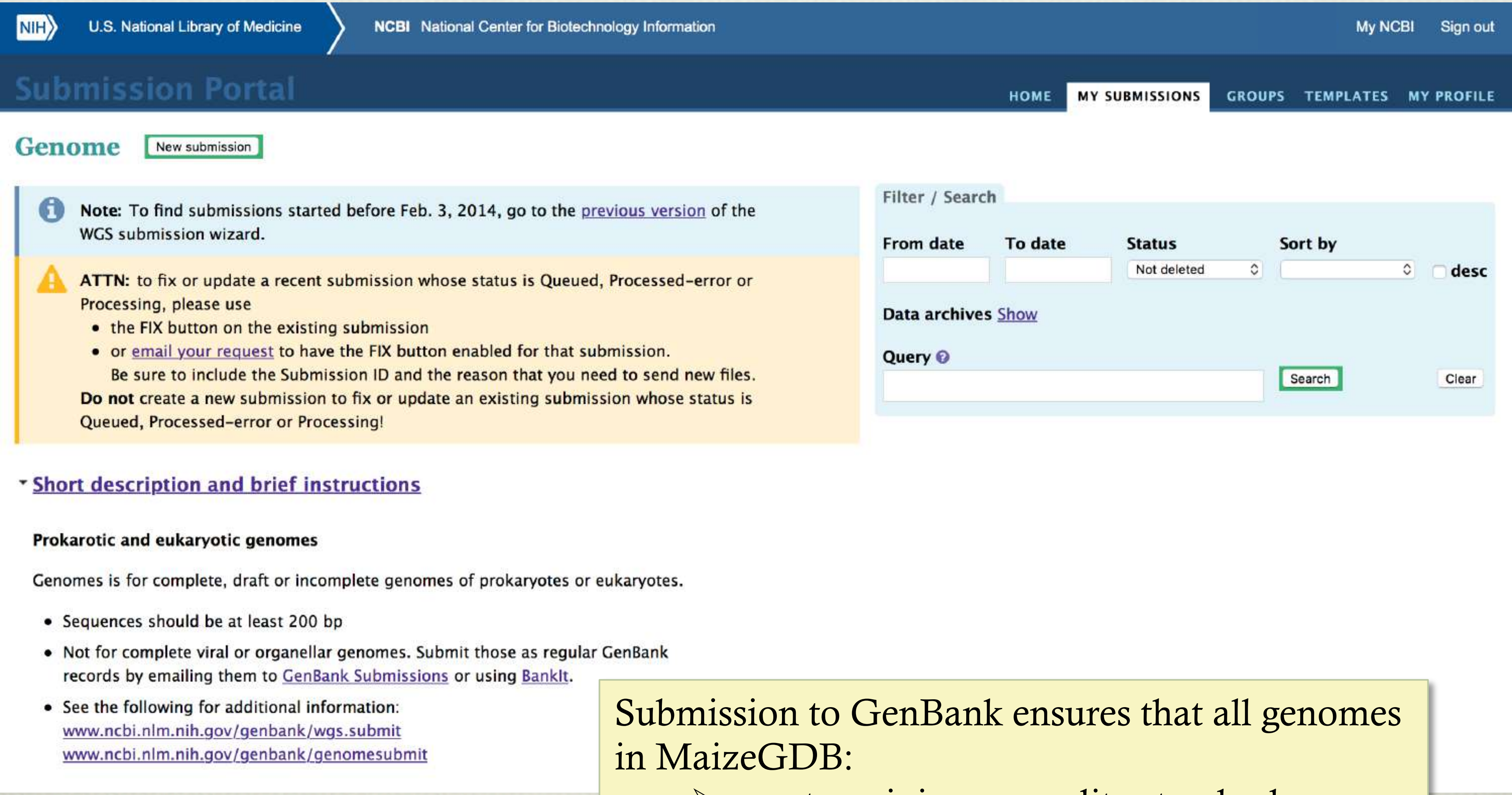
1. Genomes submitted to GenBank are required to have a minimum amount of metadata submitted

2. Genomes submitted to GenBank are checked for

→ correct file formatting

→ contamination from mitochondria, primers, adaptors, or bacteria.

Why submit to GenBank first?



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
My NCBI
Sign out
Submission Portal
HOME
MY SUBMISSIONS
GROUPS
TEMPLATES
MY PROFILE

Genome

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: Not deleted
Sort by:
□ desc

Data archives Show

Query

Short description and brief instructions

Prokarotic and eukaryotic genomes

Genomes is for complete, draft or incomplete genomes of prokaryotes or eukaryotes.

- Sequences should be at least 200 bp
- Not for complete viral or organellar genomes. Submit those as regular GenBank records by emailing them to GenBank Submissions or using BankIt.
- See the following for additional information:
  www.ncbi.nlm.nih.gov/genbank/wgs.submit
  www.ncbi.nlm.nih.gov/genbank/genomesubmit

Submission to GenBank ensures that all genomes in MaizeGDB:

- meet a minimum quality standard
- have a minimum amount of metadata reported
- are similarly formatted

Three stages in submitting an assembled genome to GenBank:

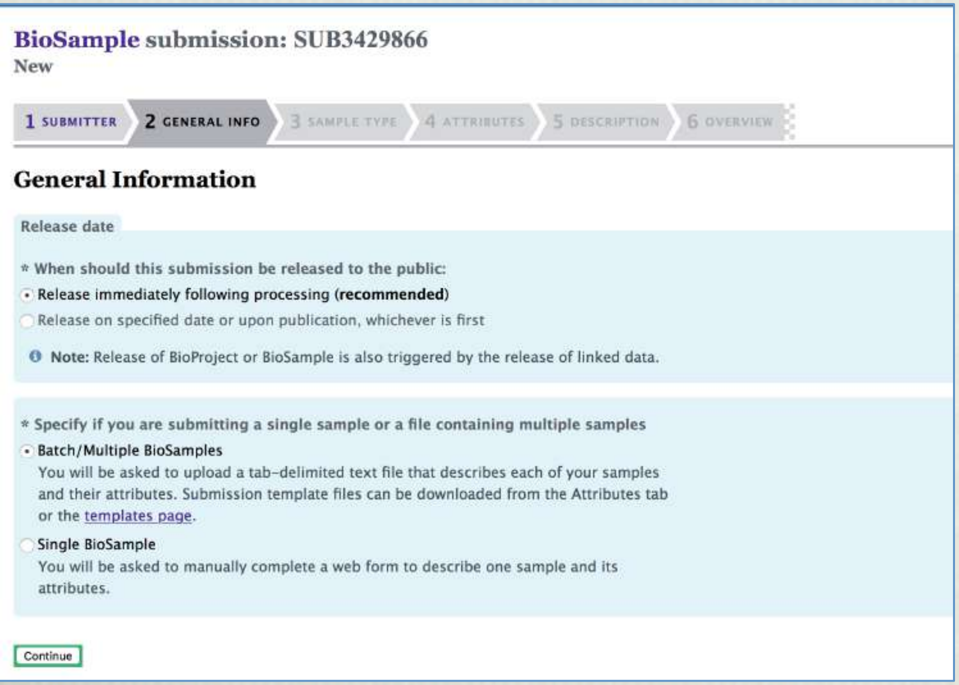
1. **Submit a BioSample**: descriptive information about the physical biological specimen from which your experimental data are derived (tissues, species, etc)

2. **Submit a BioProject**: a collection of biological data related to a single initiative originating from a single organization or from a consortium; provides users a single place to find links to the diverse data generated for that project

3. **Submit your genome!**

BioSample

https://submit.ncbi.nlm.nih.gov/subs/biosample/



BioSample submission: SUB3429866

New

General Information

Release date

* When should this submission be released to the public:

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

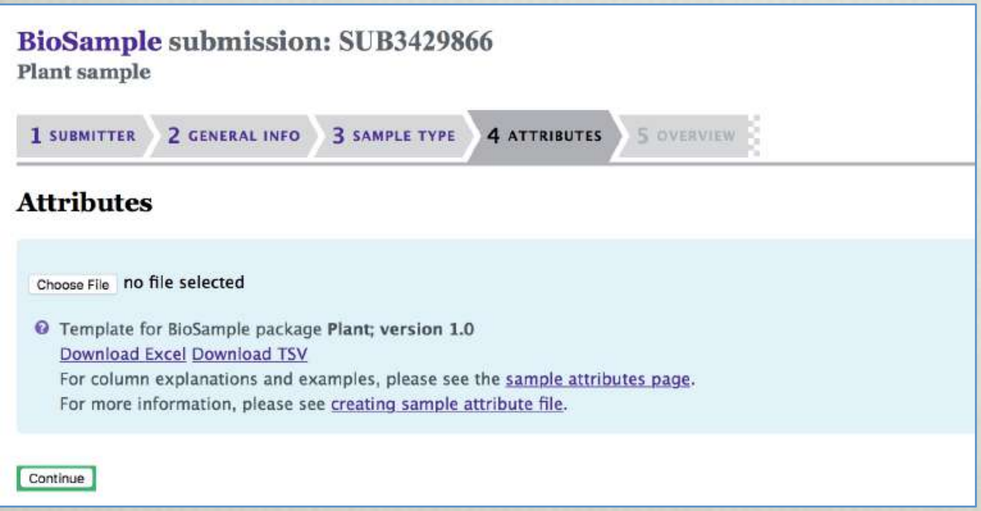
○ Release on specified date or upon publication, whichever is first

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

* Specify if you are submitting a single sample or a file containing multiple samples

● Batch/Multiple BioSamples
You will be asked to upload a tab-delimited text file that describes each of your samples and their attributes. Submission template files can be downloaded from the Attributes tab or the templates page.

○ Single BioSample
You will be asked to manually complete a web form to describe one sample and its attributes.



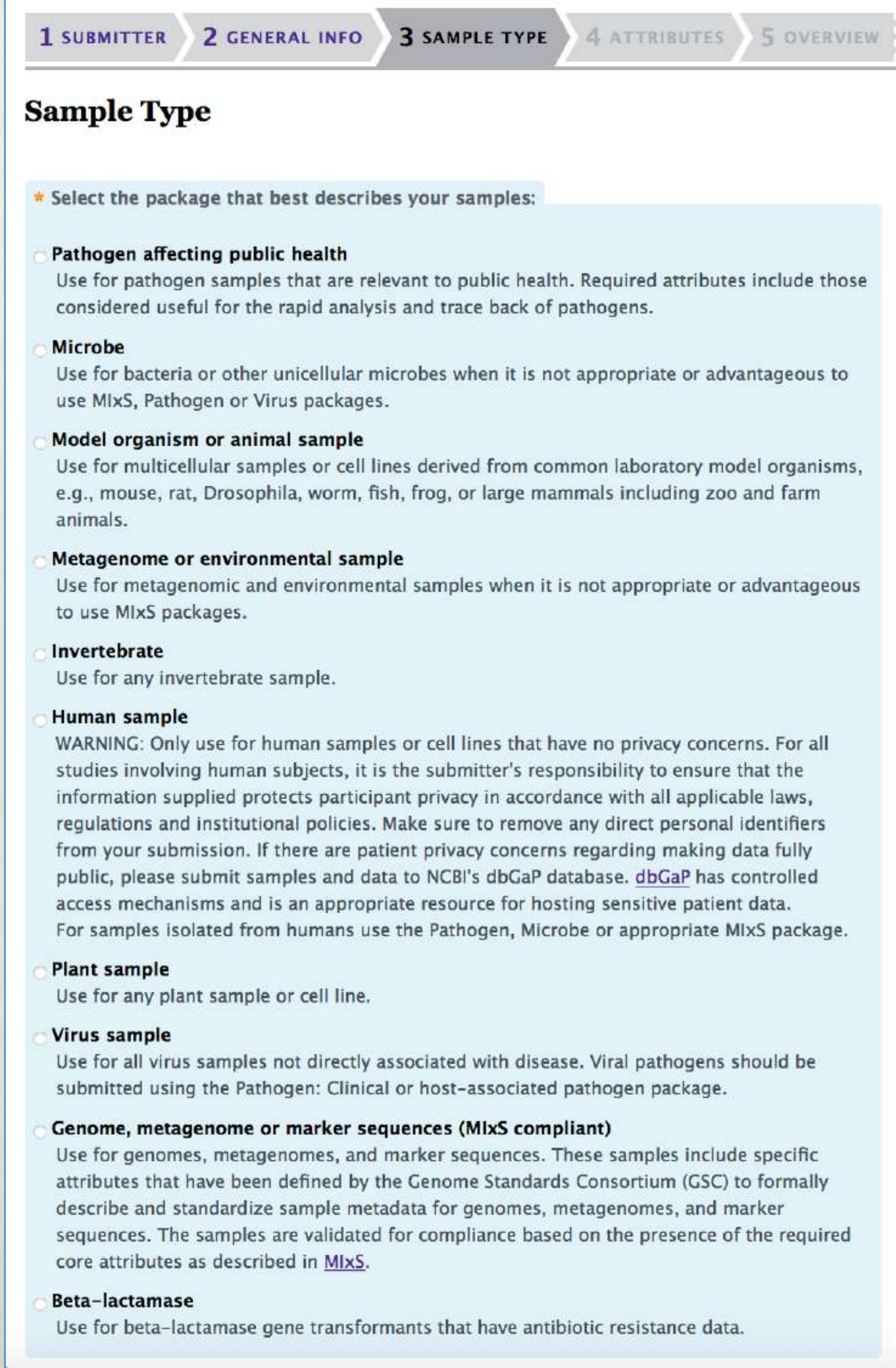
BioSample submission: SUB3429866

Plant sample

Attributes

no file selected

Template for BioSample package **Plant; version 1.0**
Download Excel Download TSV
For column explanations and examples, please see the sample attributes page.
For more information, please see creating sample attribute file.



Sample Type

* Select the package that best describes your samples:

○ **Pathogen affecting public health**
Use for pathogen samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of pathogens.

○ **Microbe**
Use for bacteria or other unicellular microbes when it is not appropriate or advantageous to use MIxS, Pathogen or Virus packages.

○ **Model organism or animal sample**
Use for multicellular samples or cell lines derived from common laboratory model organisms, e.g., mouse, rat, Drosophila, worm, fish, frog, or large mammals including zoo and farm animals.

○ **Metagenome or environmental sample**
Use for metagenomic and environmental samples when it is not appropriate or advantageous to use MIxS packages.

○ **Invertebrate**
Use for any invertebrate sample.

○ **Human sample**
WARNING: Only use for human samples or cell lines that have no privacy concerns. For all studies involving human subjects, it is the submitter's responsibility to ensure that the information supplied protects participant privacy in accordance with all applicable laws, regulations and institutional policies. Make sure to remove any direct personal identifiers from your submission. If there are patient privacy concerns regarding making data fully public, please submit samples and data to NCBI's dbGaP database. dbGaP has controlled access mechanisms and is an appropriate resource for hosting sensitive patient data.
For samples isolated from humans use the Pathogen, Microbe or appropriate MIxS package.

○ **Plant sample**
Use for any plant sample or cell line.

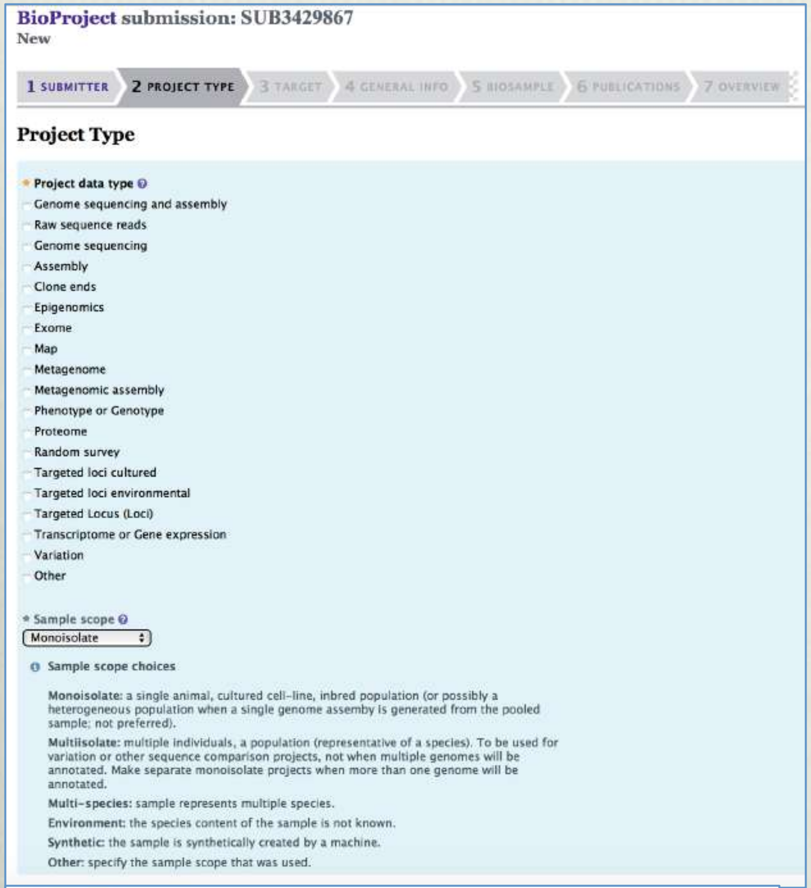
○ **Virus sample**
Use for all virus samples not directly associated with disease. Viral pathogens should be submitted using the Pathogen: Clinical or host-associated pathogen package.

○ **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.

○ **Beta-lactamase**
Use for beta-lactamase gene transformants that have antibiotic resistance data.

BioProject

https://submit.ncbi.nlm.nih.gov/subs/bioproject/



BioProject submission: SUB3429867
New

Project Type

* **Project data type**

- Genome sequencing and assembly
- Raw sequence reads
- Genome sequencing
- Assembly
- Clone ends
- Epigenomics
- Exome
- Map
- Metagenome
- Metagenomic assembly
- Phenotype or Genotype
- Proteome
- Random survey
- Targeted loci cultured
- Targeted loci environmental
- Targeted Locus (Loci)
- Transcriptome or Gene expression
- Variation
- Other

* Sample scope
Monoisolate

Sample scope choices

Monoisolate: a single animal, cultured cell-line, inbred population (or possibly a heterogeneous population when a single genome assemby is generated from the pooled sample; not preferred).

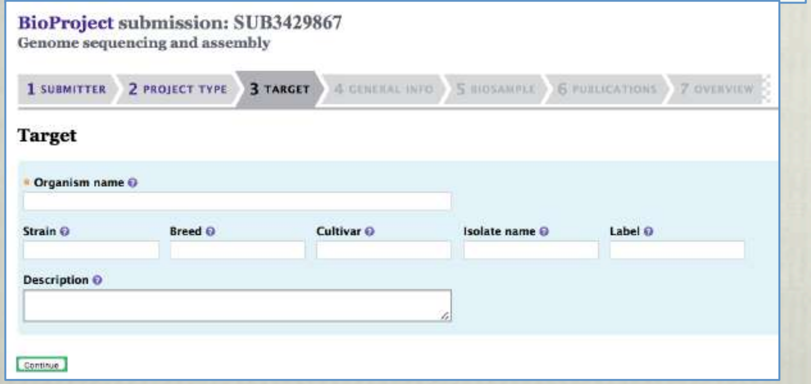
Multiisolate: multiple individuals, a population (representative of a species). To be used for variation or other sequence comparison projects, not when multiple genomes will be annotated. Make separate monoisolate projects when more than one genome will be annotated.

Multi-species: sample represents multiple species.

Environment: the species content of the sample is not known.

Synthetic: the sample is synthetically created by a machine.

Other: specify the sample scope that was used.



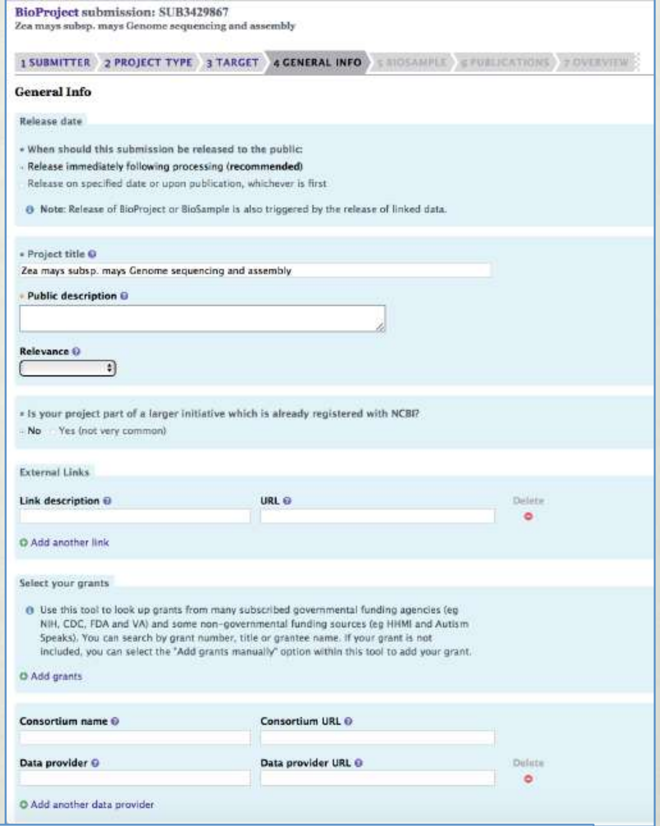
BioProject submission: SUB3429867
Genome sequencing and assembly

Target

* **Organism name**

Strain | **Breed** | **Cultivar** | **Isolate name** | **Label**

Description



BioProject submission: SUB3429867
Zea mays subsp. mays Genome sequencing and assembly

General Info

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.

* Project title
Zea mays subsp. mays Genome sequencing and assembly

* **Public description**

Relevance

* Is your project part of a larger initiative which is already registered with NCBI?
No Yes (not very common)

External Links

Link description | **URL**

Add another link

Select your grants

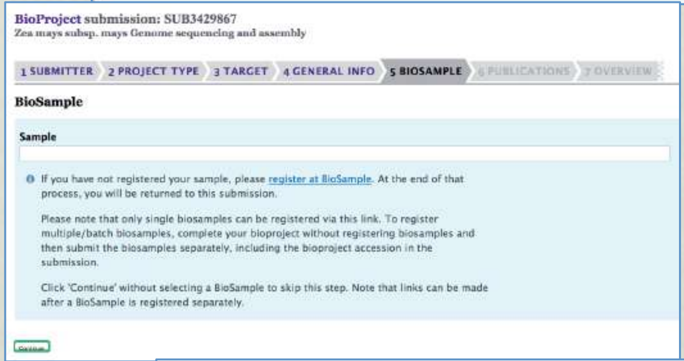
Use this tool to look up grants from many subscribed governmental funding agencies (eg NIH, CDC, FDA and VA) and some non-governmental funding sources (eg HHMI and Autism Speaks). You can search by grant number, title or grantee name. If your grant is not included, you can select the "Add grants manually" option within this tool to add your grant.

Add grants

Consortium name | **Consortium URL**

Data provider | **Data provider URL**

Add another data provider



BioProject submission: SUB3429867
Zea mays subsp. mays Genome sequencing and assembly

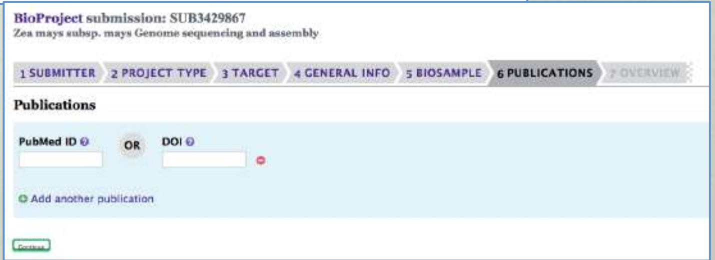
BioSample

Sample

If you have not registered your sample, please register at BioSample. At the end of that process, you will be returned to this submission.

Please note that only single biosamples can be registered via this link. To register multiple/batch biosamples, complete your bioproject without registering biosamples and then submit the biosamples separately, including the bioproject accession in the submission.

Click 'Continue' without selecting a BioSample to skip this step. Note that links can be made after a BioSample is registered separately.



BioProject submission: SUB3429867
Zea mays subsp. mays Genome sequencing and assembly

Publications

PubMed ID OR **DOI**

Add another publication

Only two files needed to submit a genome

1. **FASTA file of the scaffolds** that make up the pseudomolecule assembly

2. **AGP (A Golden Path) file**: a file that orients how the scaffolds are to be assembled into pseudomolecules; often generated by pseudomolecule assembly software https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/

Optional: FASTA file of unplaced scaffolds (AGP file optional for these)→ if you do submit these, they cannot be lumped into one giant pseudo-pseudomolecule but must remain unplaced

❖ Note: whole pseudomolecule fasta files (chromosomes) can be submitted, but you will not be able to update your genome in GenBank

Before submitting your genome files:

Check if your AGP file is correct:

https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Validation/

AGP validation standalone program:

ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/agp_validate/

1. **Make sure scaffold names in FASTA file match scaffold names in AGP**

2. **Make sure all scaffolds in the AGP file are also in the FASTA file, and vice versa (1:1)**

❖ If your fasta or AGP files are not correctly formatted, NCBI will not let you finish your submission until you re-format them.

Before submitting your genome files:

It is helpful to screen for vector contamination before submission:

https://www.ncbi.nlm.nih.gov/tools/vecscreen/

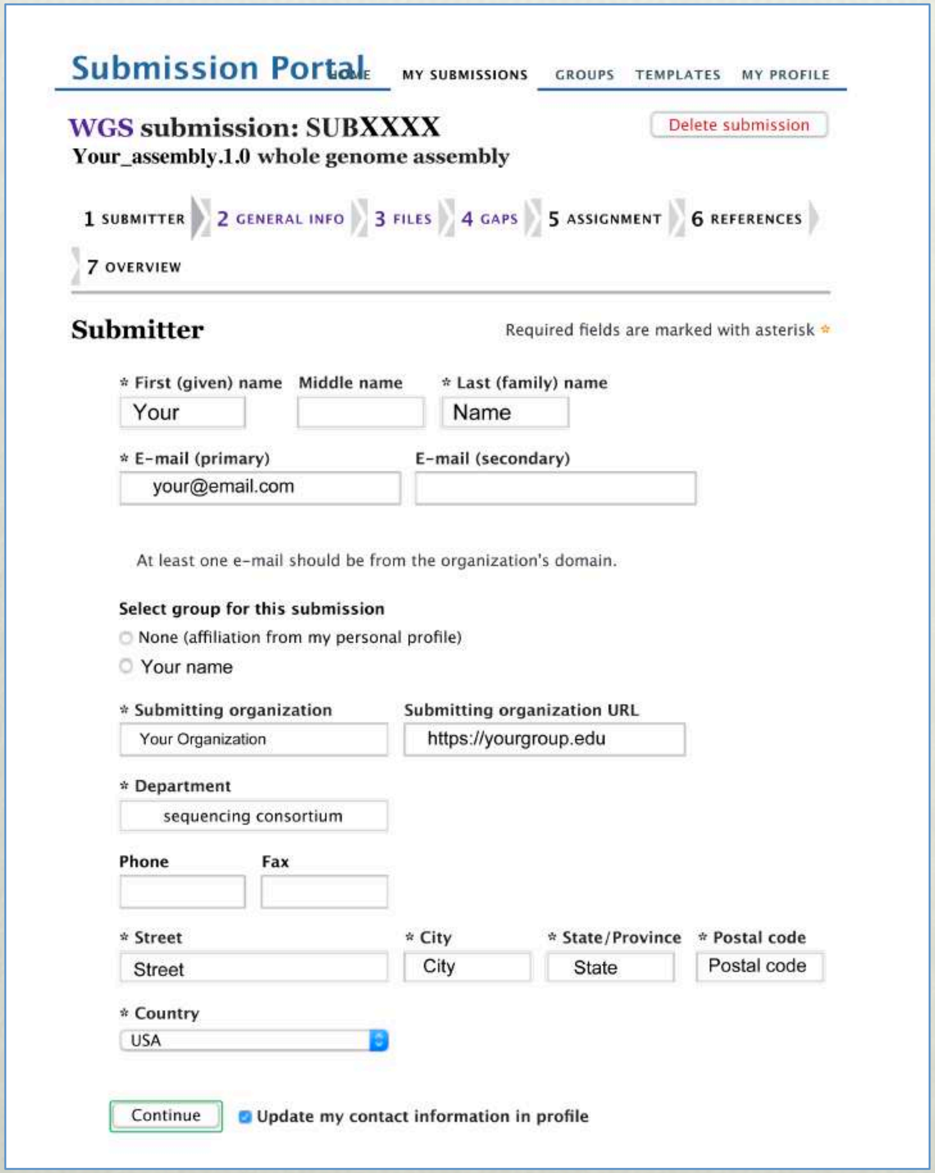
Useful links to formatting guidelines:

https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/

https://www.ncbi.nlm.nih.gov/sites/genbank/genome_validation

GenBank Genome Submission Portal

https://submit.ncbi.nlm.nih.gov/subs/genome/



WGS submission: SUBXXXX

Your_assembly.1.0 whole genome assembly

Submitter

Required fields are marked with asterisk *

* First (given) name	Middle name	* Last (family) name
Your		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	* City	* State/Province	* Postal code
Street	City	State	Postal code

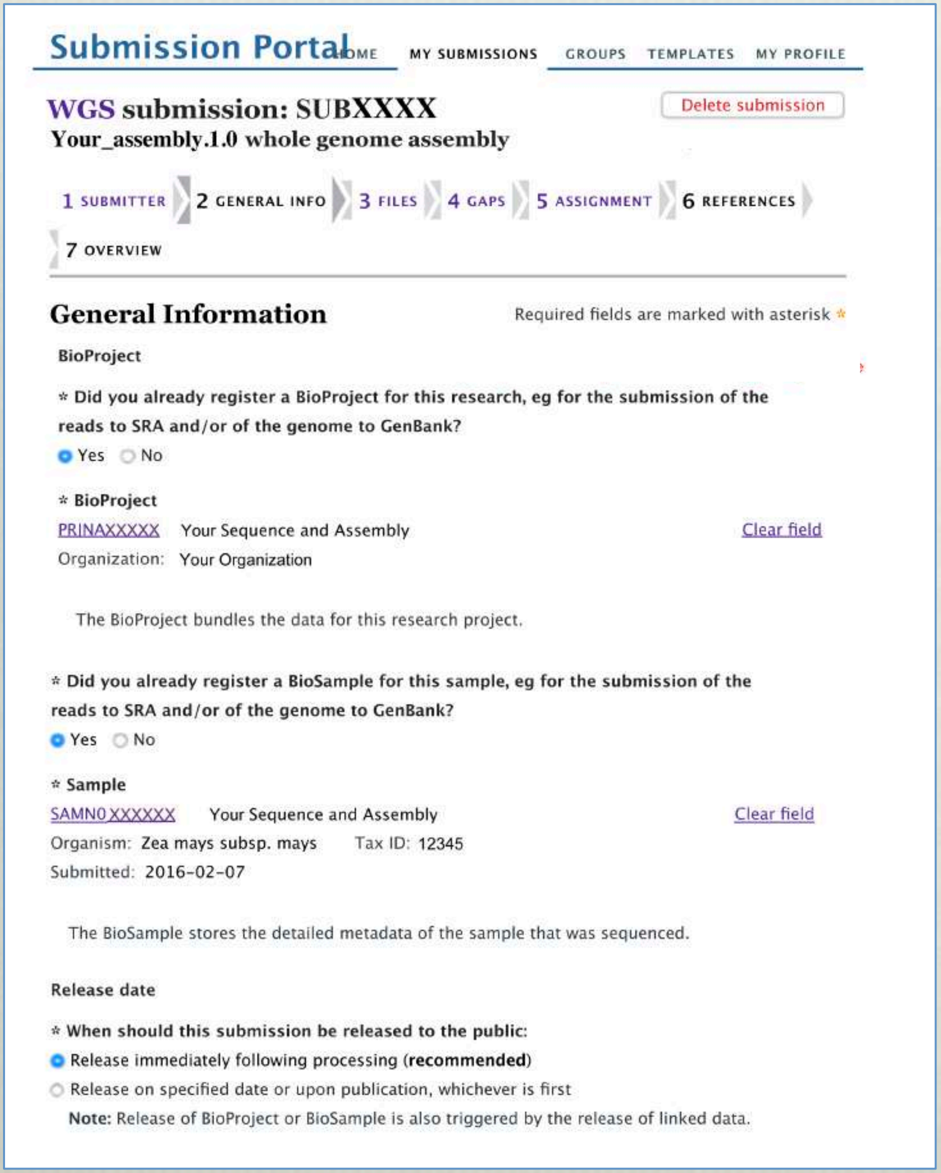
* Country: USA

☑ Update my contact information in profile

First stage: enter your submitter information

(it helps to have all this info ready to go before submission)

Second stage: General info (metadata, BioSample/Project IDs)



WGS submission: SUBXXXX

Your_assembly.1.0 whole genome assembly

General Information

Required fields are marked with asterisk *

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

*** 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.

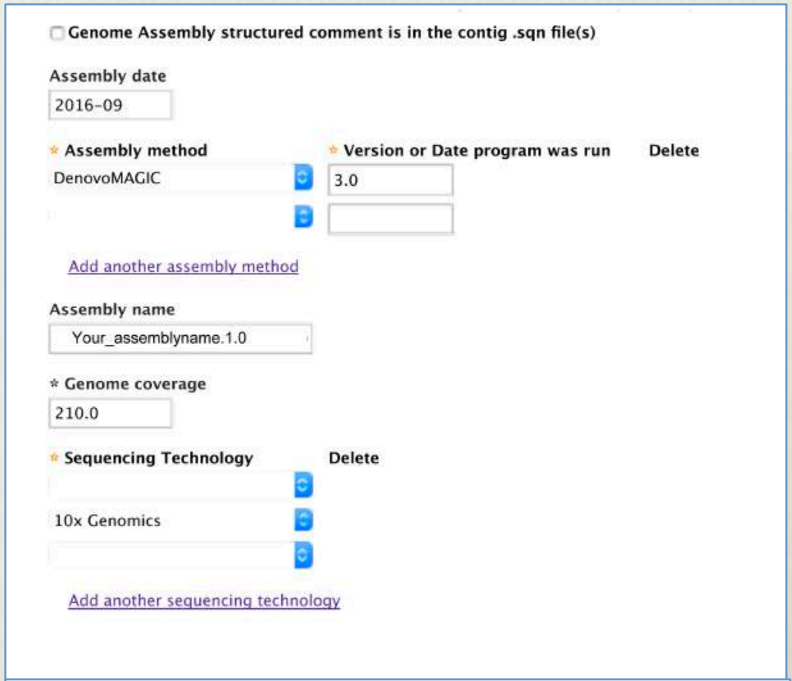
Release date

*** 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

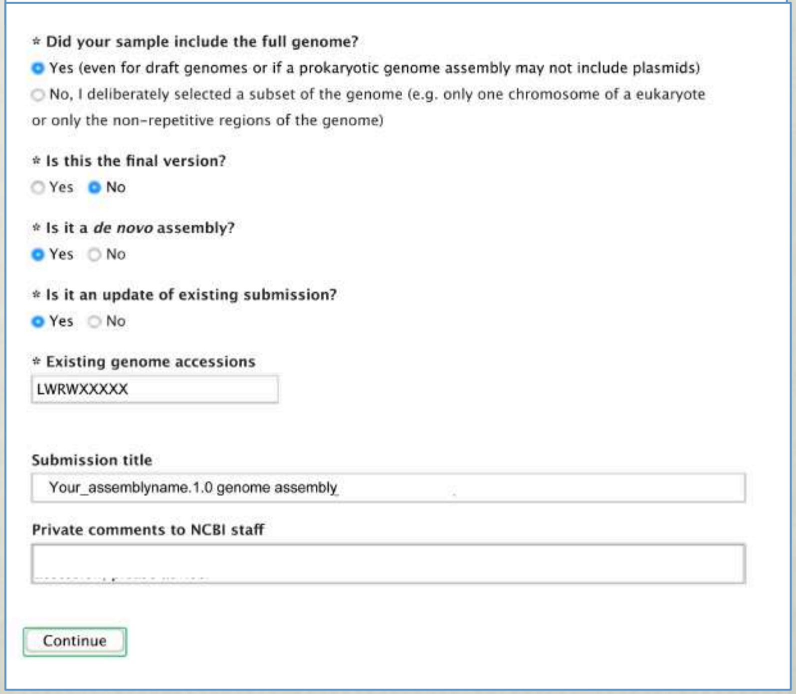
Your_assemblyname.1.0

*** 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

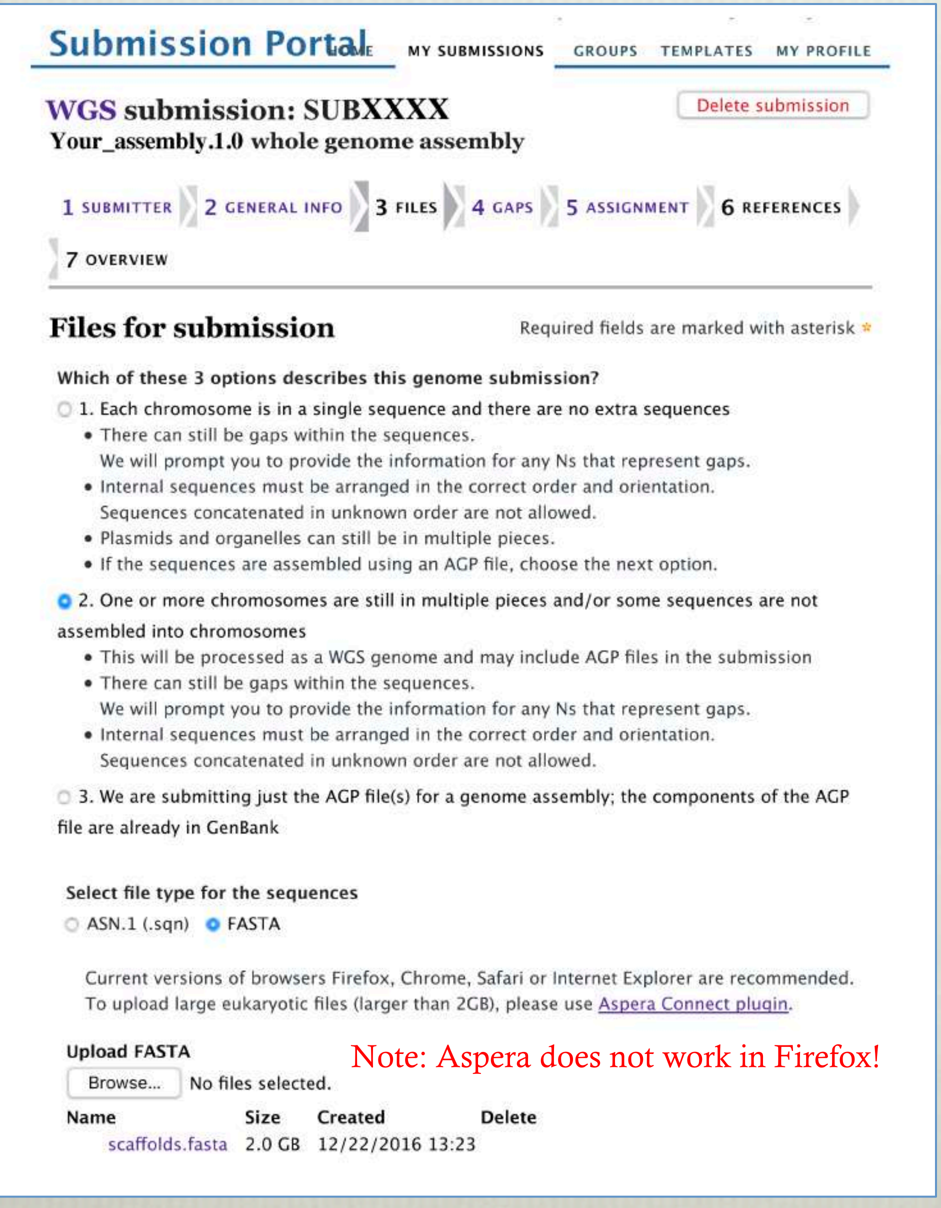
Submission title

Your_assemblyname.1.0 genome assembly

Private comments to NCBI staff

Third stage: file submission

Note: files must be formatted correctly!



WGS submission: SUBXXXX

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

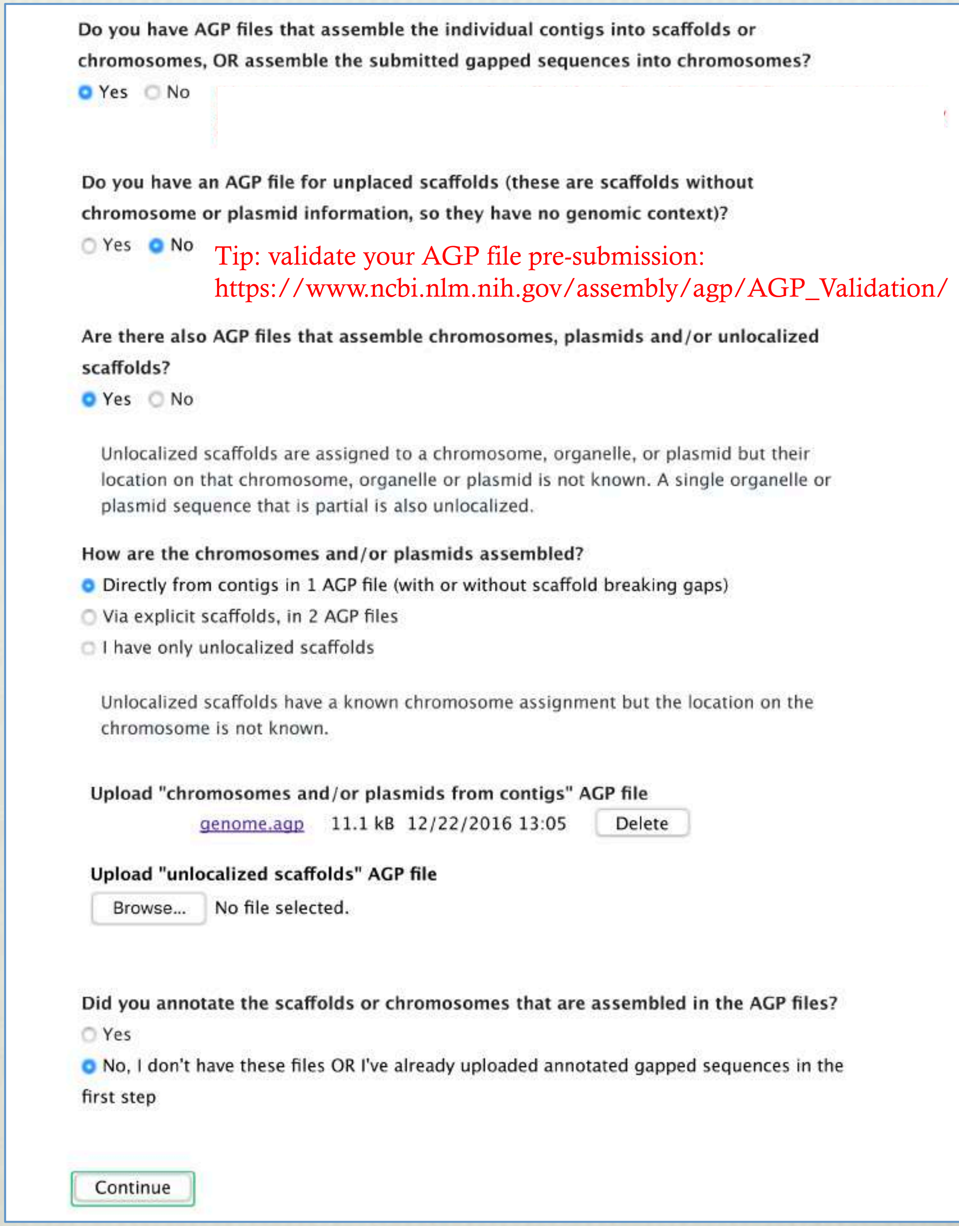
Current versions of browsers Firefox, Chrome, Safari or Internet Explorer are recommended. To upload large eukaryotic files (larger than 2GB), please use Aspera Connect plugin.

Upload FASTA

Browse... No files selected.

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

Note: Aspera does not work in Firefox!



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

● Yes ○ No

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

Tip: validate your AGP file pre-submission: https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Validation/

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?

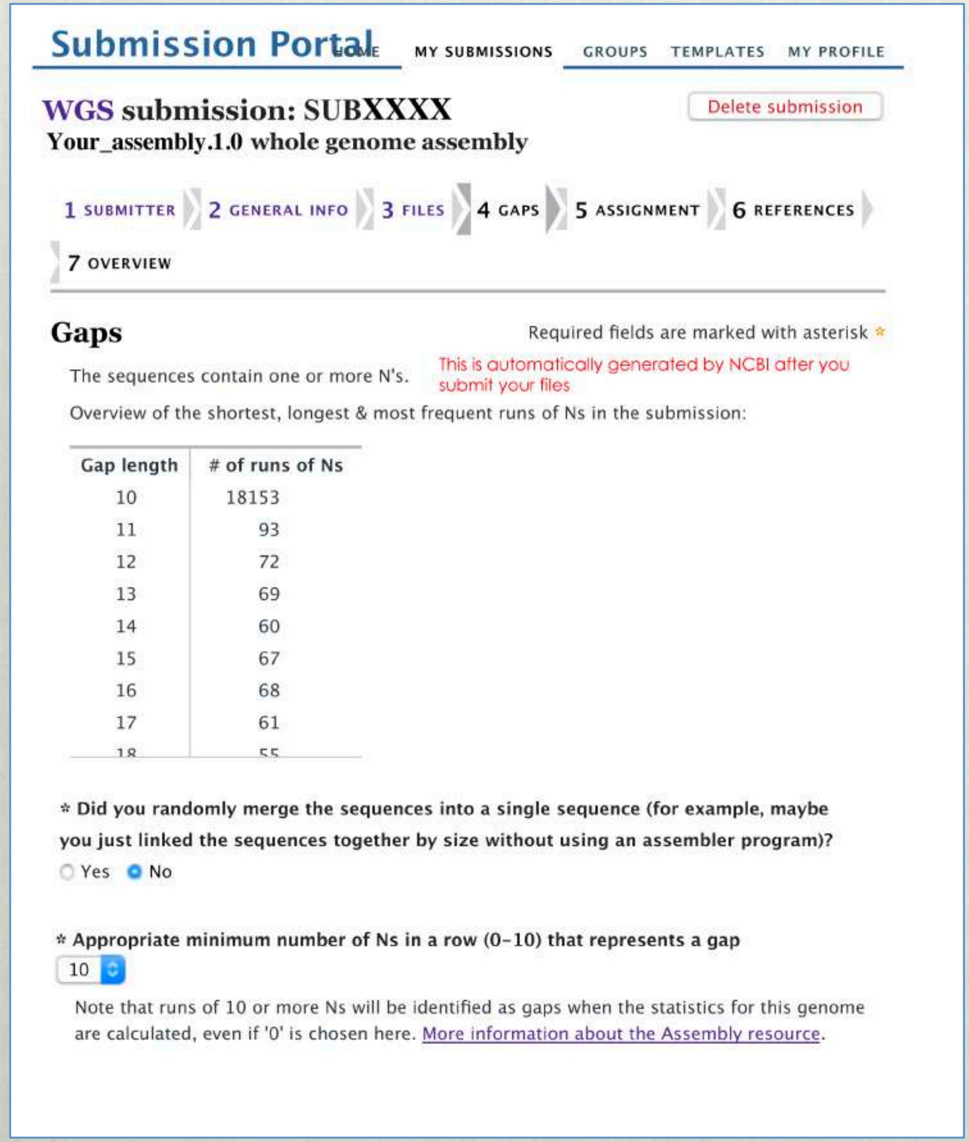
○ Yes

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

Continue

If your fasta or AGP files are not correctly formatted, NCBI will not let you continue to the next step until you re-format them

Fourth stage: assembly gap information



WGS submission: SUBXXXX

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

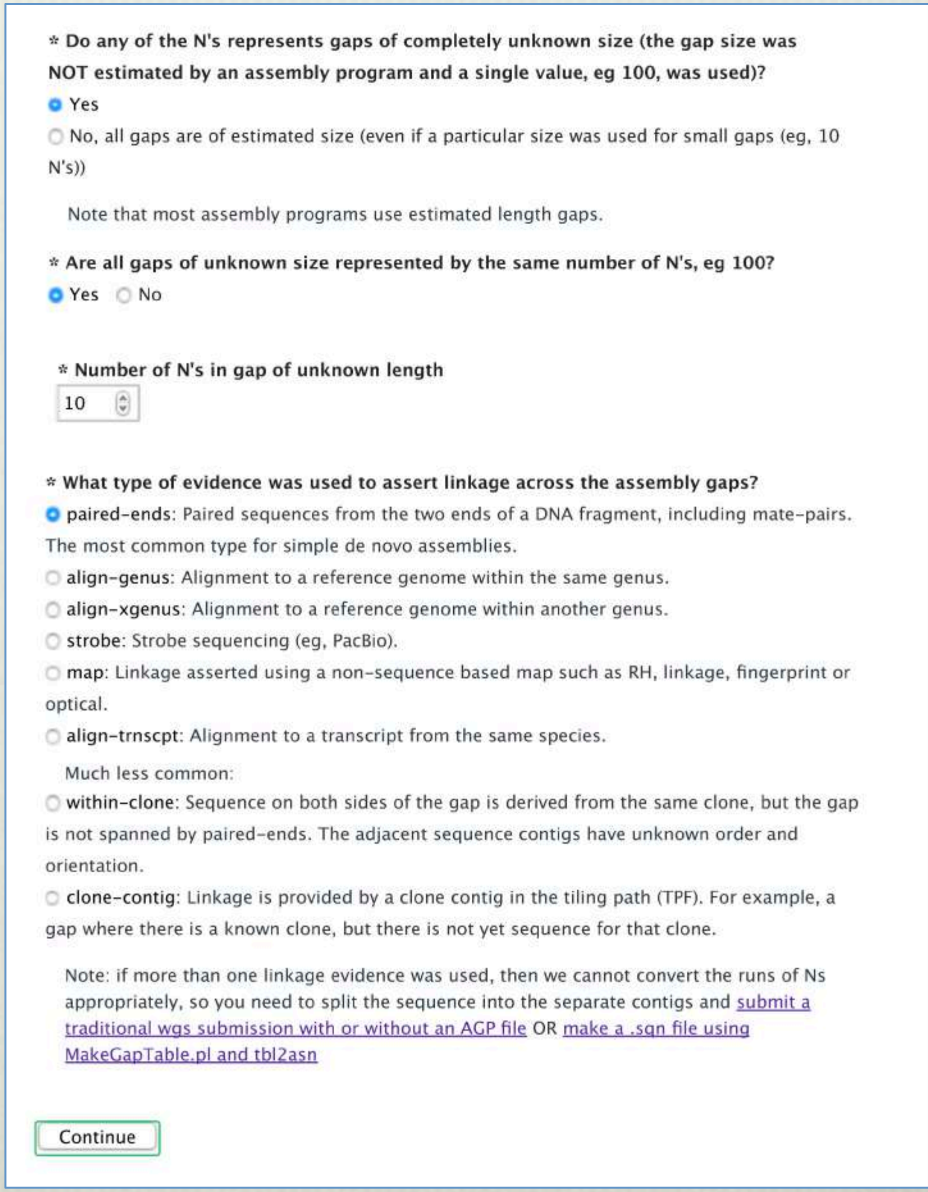
* **Did you randomly merge the sequences into a single sequence (for example, maybe you just linked the sequences together by size without using an assembler program)?**

○ 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.

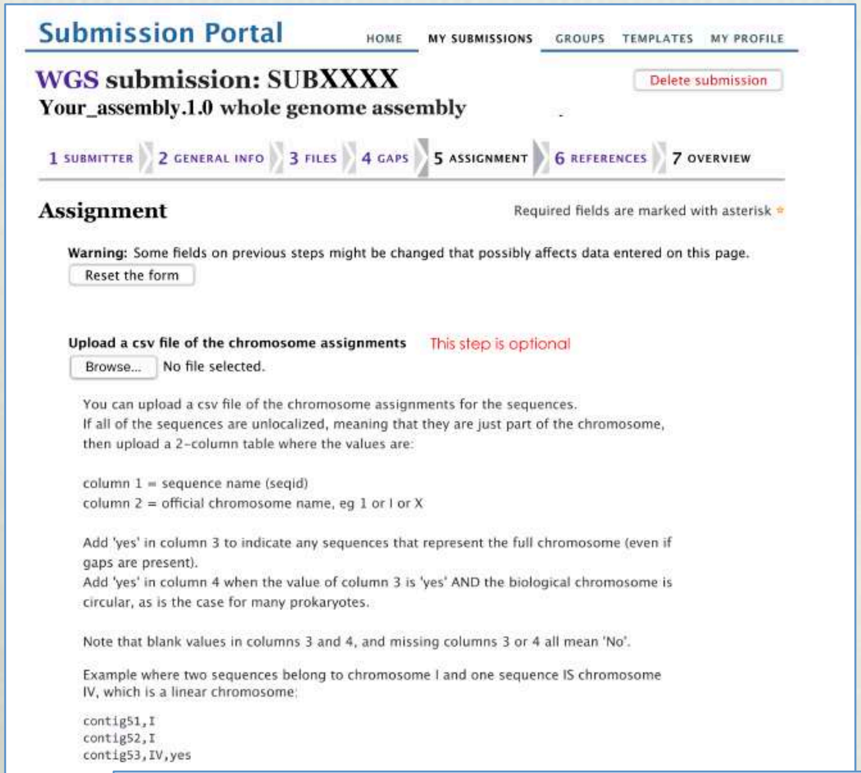
Much less common:

○ **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

Fifth stage: chromosome assignment



WGS submission: SUBXXXX

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.

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

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:

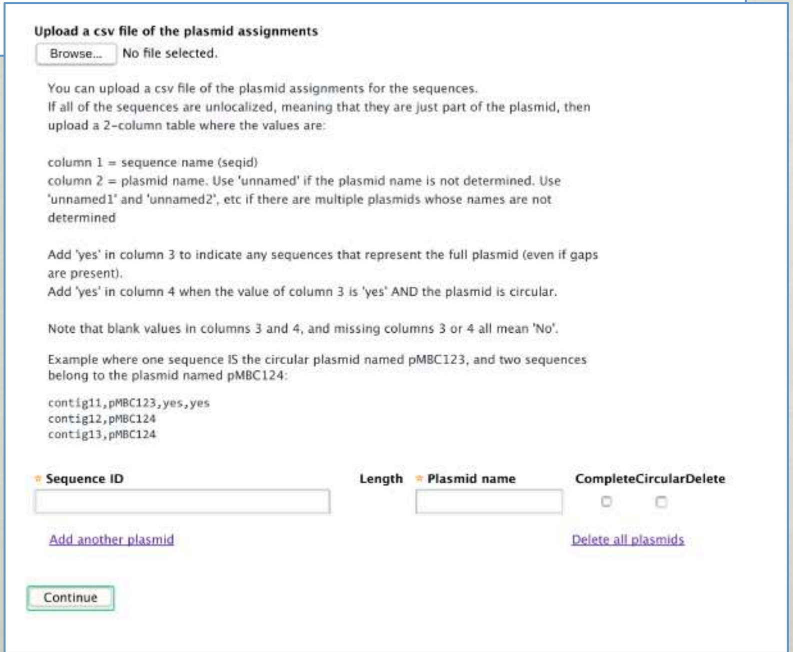
column 1 = sequence name (seqid)
column 2 = official chromosome name, eg 1 or I or X

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
```



Upload a csv file of the plasmid assignments

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

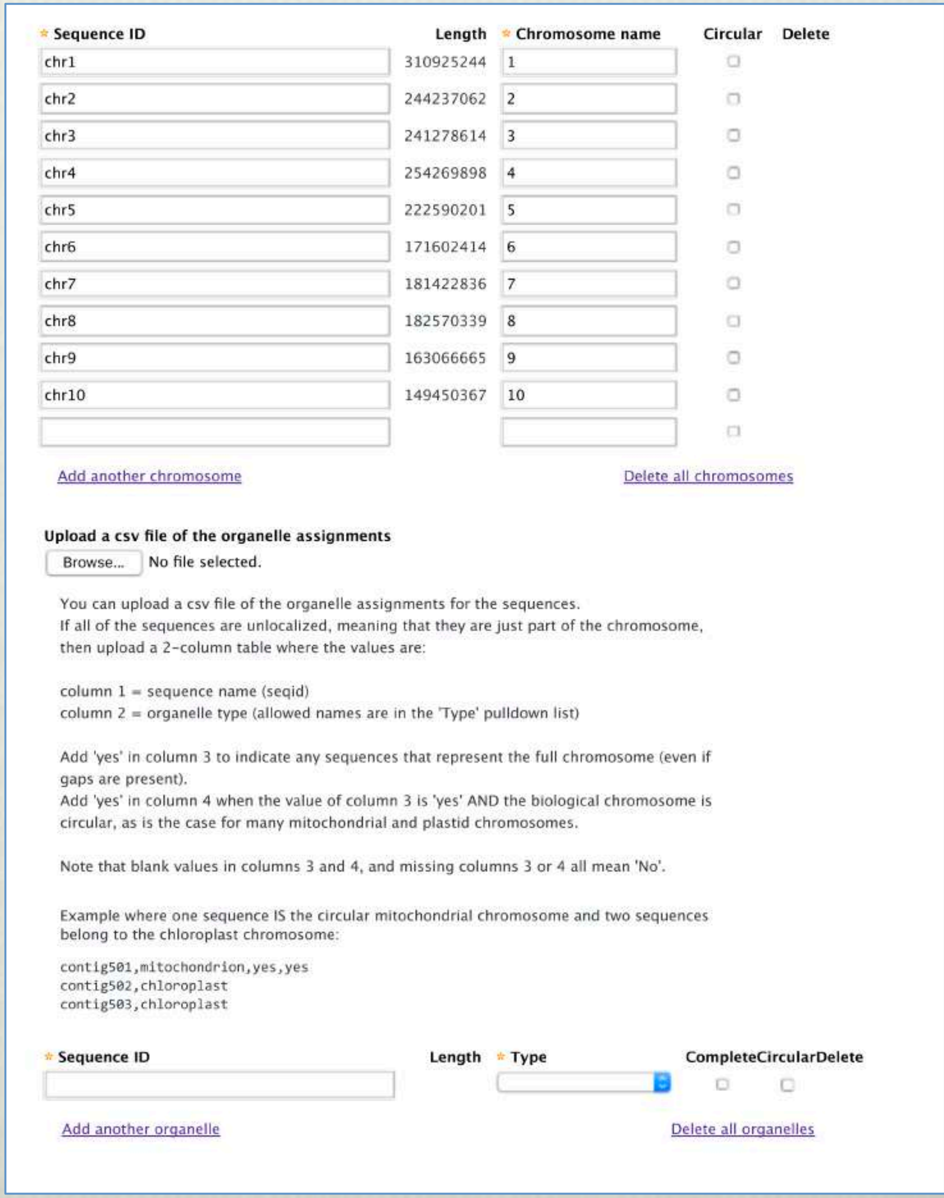
Add 'yes' in column 3 to indicate any sequences that represent the full plasmid (even if gaps are present).
Add 'yes' in column 4 when the value of column 3 is 'yes' AND the plasmid is circular.

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
			☐	☐	



* 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	☐	
			☐	

Upload a csv file of the organelle assignments

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.

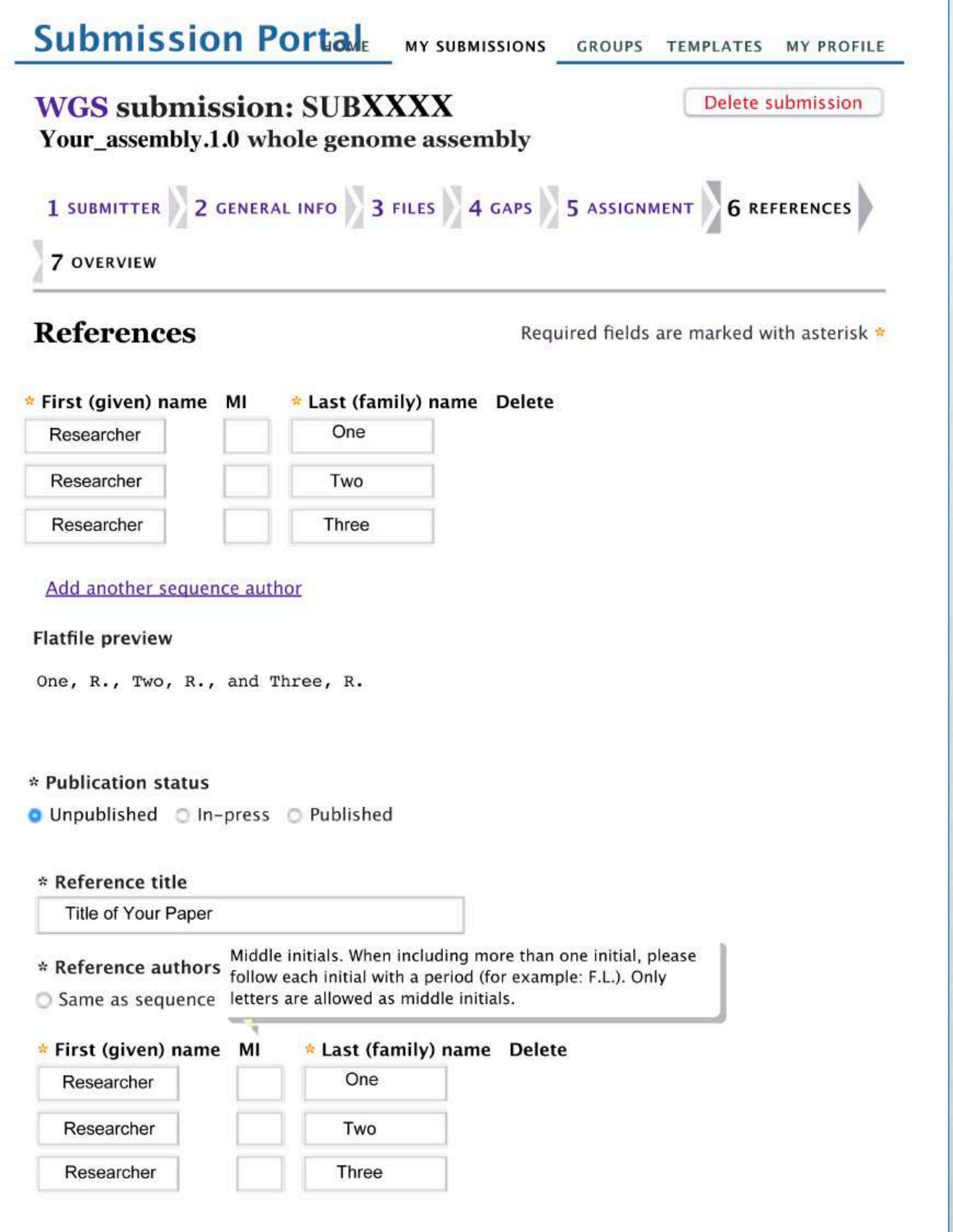
Note that blank values in columns 3 and 4, and missing columns 3 or 4 all mean 'No'.

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
			☐	☐	

Last stage: references



WGS submission: SUBXXXX

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	

Flatfile preview

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

* Publication status

● Unpublished ○ In-press ○ Published

* Reference title

Title of Your Paper

* Reference authors

○ Same as sequence

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.

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

After you submit

Contamination:

You might receive a report from NCBI that your fasta sequences contain contamination from mitochondria*, primers, adaptors, or bacteria.

→You can either mask or trim these contaminants, then resubmit. *Trimmed scaffolds will need a new AGP file*.

→If you mask, NCBI will not accept terminal Ns on scaffolds; therefore, terminal contaminants will still need to be trimmed, and a new AGP file generated.

→Terminal N's can also generate the errors SEQ_INST.TerminalGap or SEQ_INST.HighNContentPercent

*Mitochondrial contamination

Mitochondrial sequence is normally part of many eukaryotic nuclear genomes and may not be considered contamination in your genome. If so, it should not be trimmed or masked, since that sequence is considered part of the actual, biological maize genome; instead, the submitter should deem them “NUMT” in an email to genomes@ncbi.nlm.nih.gov, and keep them in situ.

Below are references that discuss how Mt sequence is found throughout nuclear genomes:

maize:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4632043/

https://link.springer.com/chapter/10.1007%2F978-3-540-74250-0_9

insects:

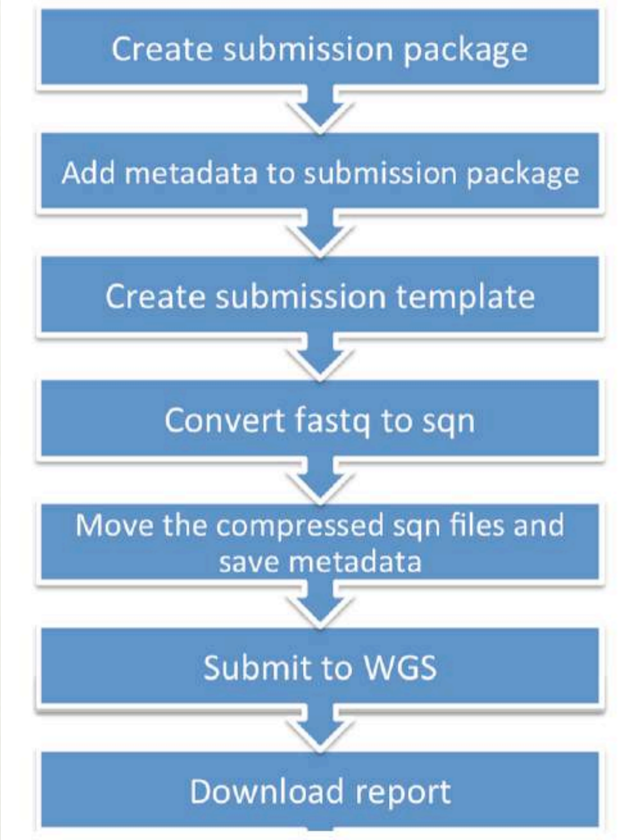
https://www.ncbi.nlm.nih.gov/pubmed/20608164

general article:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4204883/

Coming soon: CyVerse NCBI WGS Submission Portal

https://de.cyverse.org/de/



Create submission package
Add metadata to submission package
Create submission template
Convert fastq to sqn
Move the compressed sqn files and save metadata
Submit to WGS
Download report

This talk and other helpful documents can be downloaded from

https://download.maizegdb.org/Outreach/GenBank_protocols/

THANKS!

maizeGDB

MAIZE GENETICS AND GENOMICS DATABASE

I STATE

USDA