Downloading input files for OrthoFinder

Let’s imagine we want to perform a phylogenomic analysis across a set of species: Dingo (Canis lupus dingo), Great spotted kiwi (Apteryx haastii), Kakapo (Strigops habroptila), Platypus (Ornithorhynchus anatinus), Tammar wallaby (Notamacropus eugenii), and Common wombat (Vombatus ursinus).

OrthoFinder requires as input the amino acid sequences for all the protein coding genes in your species of interest. In this section of the tutorial we will discuss how to get these files for the species that you want to analyse. We will cover three major websites for getting this data, Ensembl, NCBI, and Phytozome.

We recommend **Ensembl** if you want to analyse eukaryotic species, especially vertebrates and model organisms.

We recommend **NCBI** if you want to analyse prokaryotic

species. We recommend **Phytozome** if you want to analyse

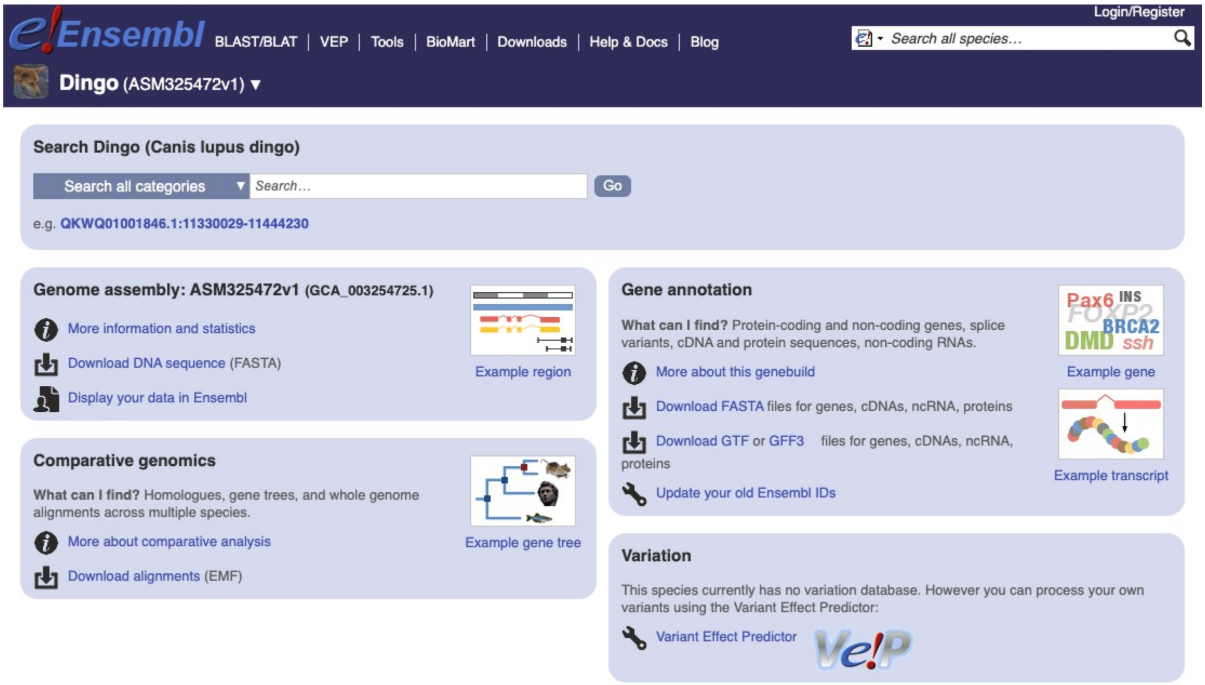
plants.

A key consideration when getting input data for OrthoFinder is gene transcripts. When you download the amino acid sequences for a genome, you might have multiple sequences for the same gene. This is because a single gene can produce multiple transcripts, which each might produce a different protein. You can see an example of transcripts in this table of the Actin gene http://www.ensembl.org/Homo_sapiens/Gene/Splice?db=core;g=ENSG00000075624;r=7:55 27151-5563784.
If we ran OrthoFinder on these raw files it would take much longer than necessary, and could lower the accuracy. We therefore want to extract just the longest transcript variant for each gene. OrthoFinder provides scripts to do this, which we will learn how to use later.

Getting data from Ensembl

First, we go to the ensembl webpage with the list of species
https://www.ensembl.org/info/about/species.html
You can also use the new beta website, which is updated more regularly
https://beta.ensembl.org/species-selector

We can then search for our first species (Canis lupus dingo), and click on the link This will take us to the ensembl page for that species



Search Dingo (Canis lupus dingo)

e.g. QKWQ01001846.1:11330029-11444230

Genome assembly: ASM325472v1 (GCA_003254725.1)

- More information and statistics
- Download DNA sequence (FASTA)
- Display your data in Ensembl

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

- More about comparative analysis
- Download alignments (EMF)

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

- More about this genebuild
- Download FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

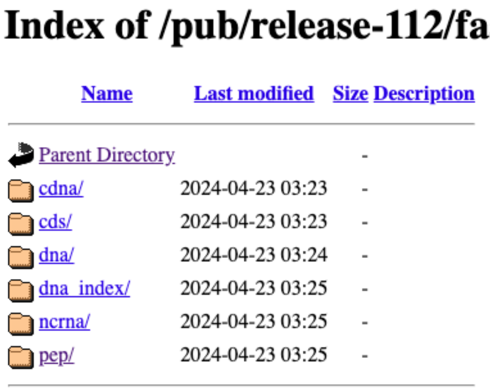
Variation

This species currently has no variation database. However you can process your own variants using the Variant Effect Predictor:

- Variant Effect Predictor

On the right hand side under the heading ‘Gene annotation’ we can press the link to ‘Download FASTA’ files for the genome.

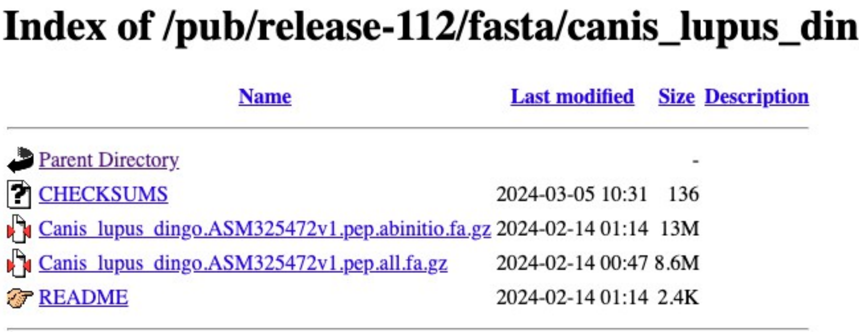
This will take us to a webpage with some folders.



Index of /pub/release-112/fasta/canis_lupus_dingo

Name	Last modified	Size	Description
Parent Directory		-	
cdna/	2024-04-23 03:23	-	
cds/	2024-04-23 03:23	-	
dna/	2024-04-23 03:24	-	
dna_index/	2024-04-23 03:25	-	
ncrna/	2024-04-23 03:25	-	
pep/	2024-04-23 03:25	-	

We need to click on the ‘pep/’ folder, and then click on the file that ends in .pep.all.fa.gz



Index of /pub/release-112/fasta/canis_lupus_dingo/pep

Name	Last modified	Size	Description
Parent Directory		-	
CHECKSUMS	2024-03-05 10:31	136	
Canis_lupus_dingo.ASM325472v1.pep.abinitio.fa.gz	2024-02-14 01:14	13M	
Canis_lupus_dingo.ASM325472v1.pep.all.fa.gz	2024-02-14 00:47	8.6M	
README	2024-02-14 01:14	2.4K	

We can then repeat this step for our other species, and place the files that are downloaded into a folder on our computer. I have named my folder ‘Proteomes’.

The files are stored as .gz compressed files to save space, but we now need to expand the files so that we can access the sequences. You can either double-click on them all, or use the command line

```
gunzip *.gz
```

We’ll use a script provided with OrthoFinder to extract just the longest transcript variant per gene and run OrthoFinder on these files:

You can find the script here
https://github.com/OrthoFinder/OrthoFinder/blob/master/tools/primary_transcript.py

To run the script, first place it in the Proteomes folder. Then, open the command line and navigate to the Proteomes folder (using cd). You can then use the following command to run the script

```
for f in *fa ; do python primary_transcript.py $f ; done
```

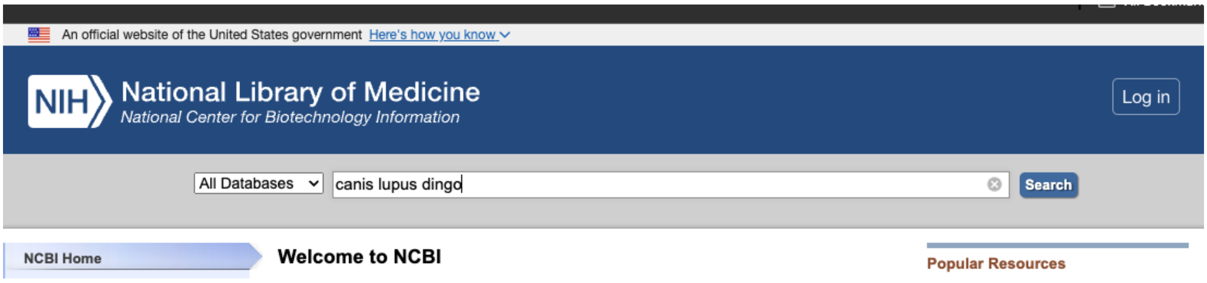
The script will generate a new folder called ‘primary_transcripts’, which contains our files.

Shortening the filenames is a good idea as it keeps the results tidy as the filenames are used to refer to the species, e.g. I shortened it to Canis_lupus_dingo.fa.

Our data is now ready for OrthoFinder. You can skip to Section 3 – Running OrthoFinder, or check out the below guides on getting input data from other sources

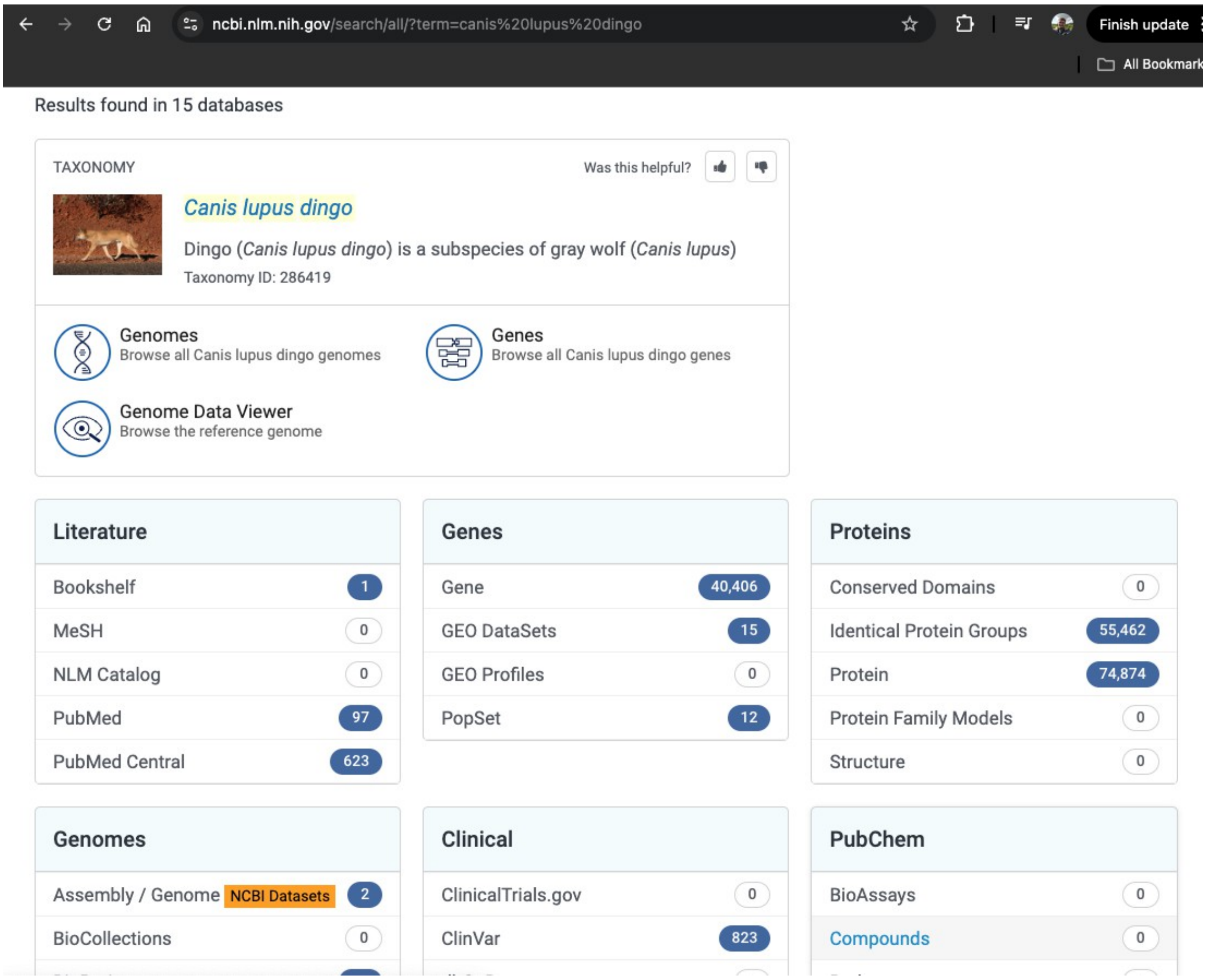
Getting data from NCBI

We start by going to https://www.ncbi.nlm.nih.gov/ and searching for our first species


An official website of the United States government Here’s how you know
NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
All Databases
Search
NCBI Home
Welcome to NCBI
Popular Resources

canis lupus dingo

We want to find the genome of this species, so scroll down to the ‘Genomes’ section, and click on ‘Assembly / Genome’



ncbi.nlm.nih.gov/search/all/?term=canis%20lupus%20dingo

Results found in 15 databases

TAXONOMY

Was this helpful?

Canis lupus dingo

Dingo (*Canis lupus dingo*) is a subspecies of gray wolf (*Canis lupus*)

Taxonomy ID: 286419

Genomes
Browse all Canis lupus dingo genomes

Genes
Browse all Canis lupus dingo genes

Genome Data Viewer
Browse the reference genome

Literature	
Bookshelf	1
MeSH	0
NLM Catalog	0
PubMed	97
PubMed Central	623

Genes	
Gene	40,406
GEO DataSets	15
GEO Profiles	0
PopSet	12

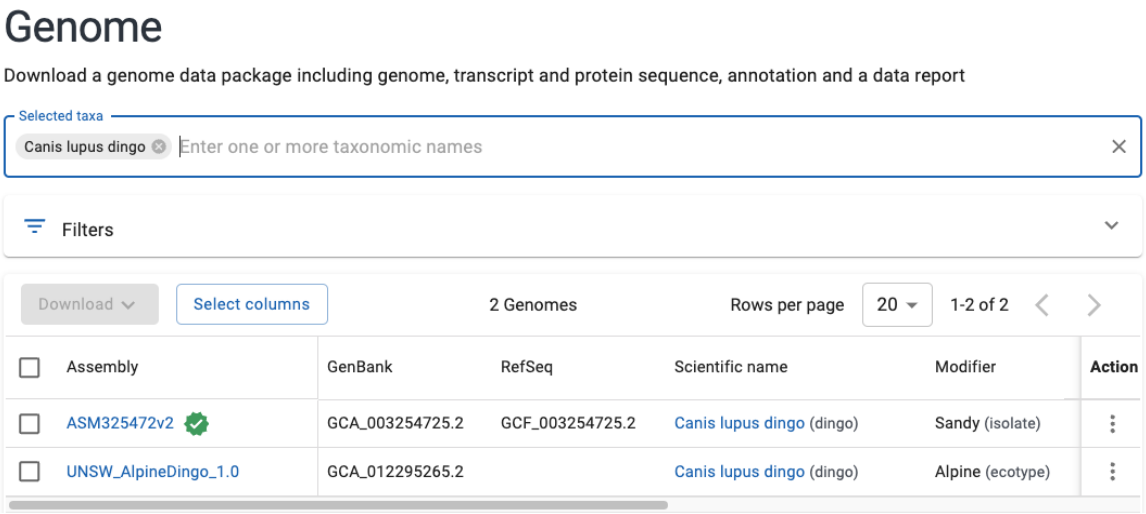
Proteins	
Conserved Domains	0
Identical Protein Groups	55,462
Protein	74,874
Protein Family Models	0
Structure	0

Genomes	
Assembly / Genome NCBI Datasets	2
BioCollections	0

Clinical	
ClinicalTrials.gov	0
ClinVar	823

PubChem	
BioAssays	0
Compounds	0

There might be several genomes listed on this page. We are going to click on the one with the green tick, which shows us the reference genome.



Genome

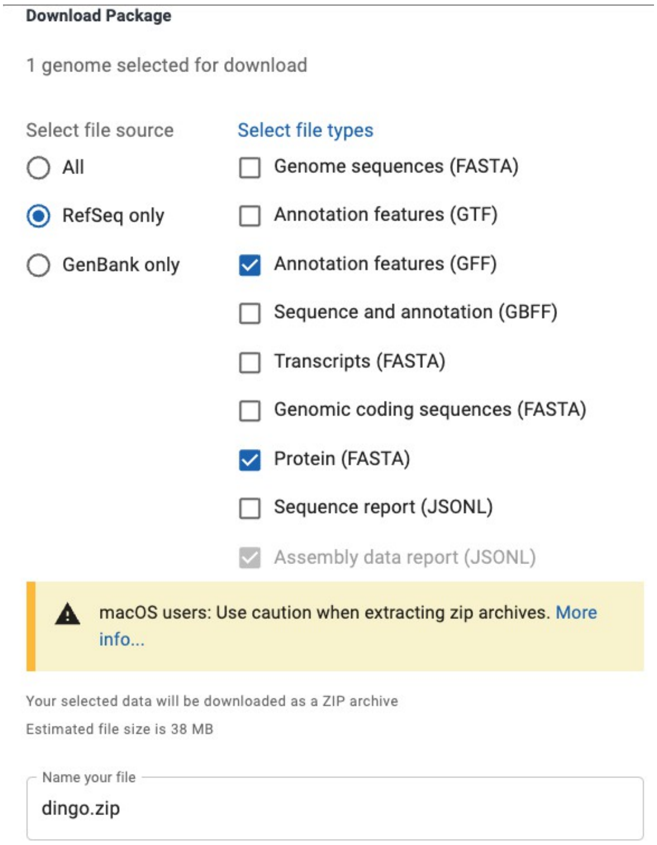
Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa: Canis lupus dingo ⊗ Enter one or more taxonomic names

2 Genomes   Rows per page 20   1-2 of 2

□ Assembly	GenBank	RefSeq	Scientific name	Modifier	Action
□ ASM325472v2 ✓	GCA_003254725.2	GCF_003254725.2	Canis lupus dingo (dingo)	Sandy (isolate)	⋮
□ UNSW_AlpineDingo_1.0	GCA_012295265.2		Canis lupus dingo (dingo)	Alpine (ecotype)	⋮

We can then click the blue ‘Download’ button, and click the boxes to select the curated ‘RefSeq only’ annotation, and that we want the GFF and protein FASTA files. We can also give our download a useful name (such as the name of the species).



Download Package

1 genome selected for download

Select file source

○ All

● RefSeq only

○ GenBank only

Select file types

□ Genome sequences (FASTA)

□ Annotation features (GTF)

☑ Annotation features (GFF)

□ Sequence and annotation (GBFF)

□ Transcripts (FASTA)

□ Genomic coding sequences (FASTA)

☑ Protein (FASTA)

□ Sequence report (JSONL)

☑ Assembly data report (JSONL)

⚠ macOS users: Use caution when extracting zip archives. More info...

Your selected data will be downloaded as a ZIP archive

Estimated file size is 38 MB

Name your file

dingo.zip

We can then click the ‘Download’ button, which downloads a zip file.

Repeat this step for all of the species that you want to use. For some species in our example list, like the Kiwi, there is no RefSeq genome with protein fasta available. For these species, it is recommended to use Ensembl.

Now, we need to deal with the issue of multiple transcripts per gene. For our first species, the Dingo, there are about 20,000 genes. However, if we look at the ‘protein.faa’ file that we have downloaded, it has about 75,000 sequences.

We can use the script ```ncbi_primary_transcript.py``` to extract the longest transcript per gene.

Place the script in the folder that has the zip folders, and run the following line of code in the command line

```
python ncbi_primary_transcripts.py
```

This will then make a folder named ‘primary_transcripts’, which is ready to run Orthofinder on.

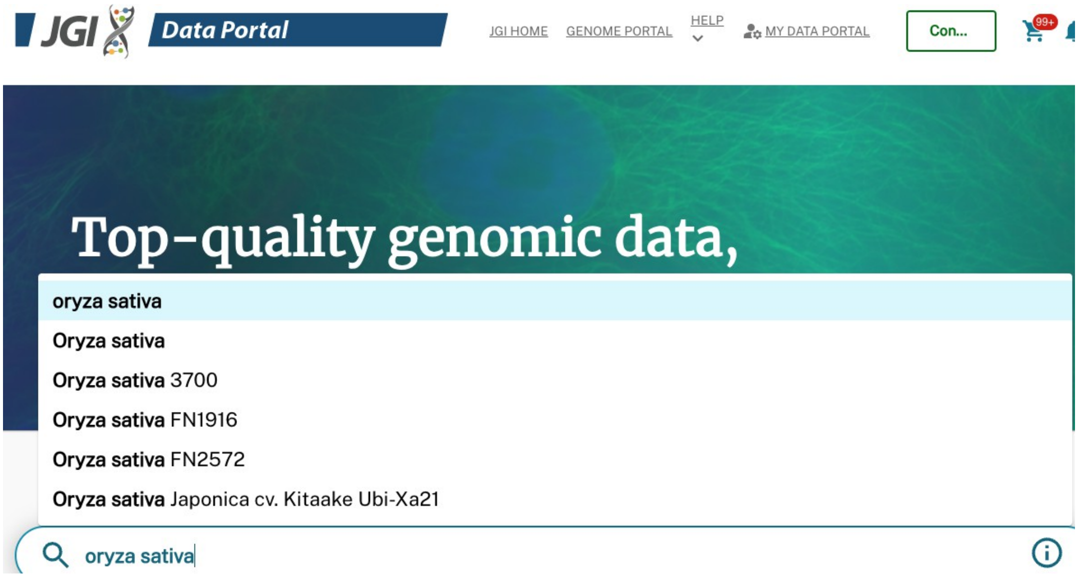
If you want to download data in bulk, you can also use the NCBI Datasets tool, which can run on the command line. For more info, see here https://www.ncbi.nlm.nih.gov/datasets/

Getting data from Phytozome

First, go to the Phytozome data portal website

https://data.jgi.doe.gov/ Here, you can search for a species, such

as Oryza sativa

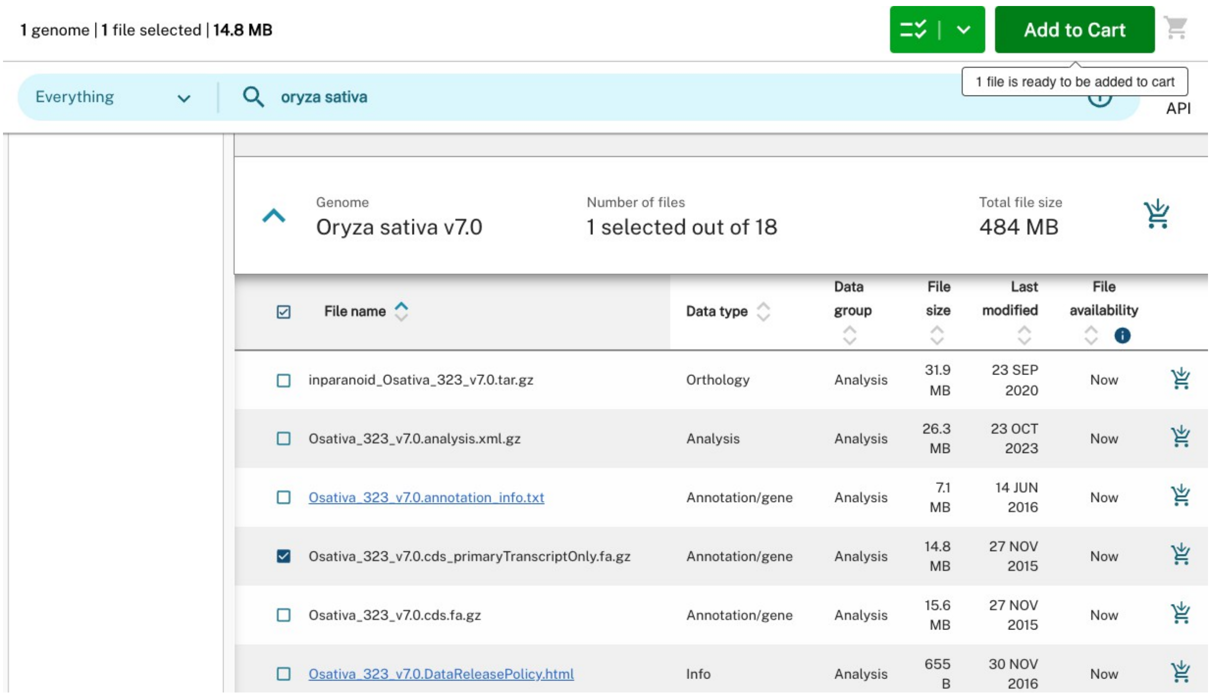


oryza sativa
Oryza sativa
Oryza sativa 3700
Oryza sativa FN1916
Oryza sativa FN2572
Oryza sativa Japonica cv. Kitaake Ubi-Xa21

oryza sativa

You can then find the genome that want to download, and click on it

The file that we want ends in ‘primaryTranscriptOnly.fa.gz’. Luckily for us, Phytozome already deals with the problem of multiple transcripts. We can select that file, and click ‘Add to Cart’.



1 genome | **1** file selected | **14.8 MB**

Genome: Oryza sativa v7.0

Number of files: 1 selected out of 18

Total file size: 484 MB

☑	File name	Data type	Data group	File size	Last modified	File availability
□	inparanoid_Osativa_323_v7.0.tar.gz	Orthology	Analysis	31.9 MB	23 SEP 2020	Now
□	Osativa_323_v7.0.analysis.xml.gz	Analysis	Analysis	26.3 MB	23 OCT 2023	Now
□	Osativa_323_v7.0.annotation_info.txt	Annotation/gene	Analysis	7.1 MB	14 JUN 2016	Now
☑	Osativa_323_v7.0.cds_primaryTranscriptOnly.fa.gz	Annotation/gene	Analysis	14.8 MB	27 NOV 2015	Now
□	Osativa_323_v7.0.cds.fa.gz	Annotation/gene	Analysis	15.6 MB	27 NOV 2015	Now
□	Osativa_323_v7.0.DataReleasePolicy.html	Info	Analysis	655 B	30 NOV 2016	Now

We can then repeat this step for all of the species that we want to download. When we are ready, we then click the shopping cart icon, and press ‘Download’ on the next page. You will have to sign-in or register to download data.

These files can be placed in a folder, where they are now ready to run OrthoFinder.