

# DIFFERENTIAL GENE EXPRESSION ANALYSIS

## Module 2: Genome (.fasta) and annotation (.gff) file downloads

### DOWNLOADING GENOME AND ANNOTATION FILES FROM PUBLIC DATABASES

```
# The human Genome can be downloaded from several public web portals.  
  
# Gencode is a good option, with a clean and clear webpage:  
# https://www.gencodegenes.org/human/  
# Another is NCBI:  
# https://www.ncbi.nlm.nih.gov/genome/?term=txid9606\[orgn\]  
# To download the genome sequence in FASTA file format  
  
cd /home/$USER/DGE_Virtual/  
  
mkdir human_reference/  
  
cd human_reference/
```

To get the download link to use in the "wget" command, do the following steps,

STEP 1: Go the relevant web page from where the genome fasta should be downloaded

NCBI Resources How To

Genome Genome yeast Search

Create alert Limits Advanced

COVID-19 is an emerging, rapidly evolving situation.  
Get the latest public health information from CDC: <https://www.coronavirus.gov>.  
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.

**Saccharomyces cerevisiae (baker's yeast)**  
Reference genome: [Saccharomyces cerevisiae S288C \(assembly R64\)](#)  
Download sequences in FASTA format for **genome**, transcript, protein  
Download genome annotation in GFF, GenBank or tabular format  
BLAST against Saccharomyces cerevisiae genome, transcript, protein  
All 815 genomes for species:  
Browse the list  
Download sequence and annotation from RefSeq or GenBank  
Try NCBI Datasets - a new way to download genome sequence and annotation we're testing in NCBI Labs

Filters: [Manage Filters](#)

Find related data  
Database: Select  
Find items

Search details  
yeast[All Fields]  
Search

Recent activity

See [UBE2E3 \(YEAST\) ubiquitin conjugating enzyme E2 E3](#) in the Gene database

Display Settings: Summary, 20 per page Send to:

**Search results**  
Items: 1 to 20 of 98

<< First < Prev Page 1 of 5 Next > Last >>

[https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF\\_000146045.2\\_R64/GCF\\_000146045.2\\_R64\\_genomic.fna.gz](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.fna.gz)

NCBI Resources How To Sign in to NCBI

Genome Genome yeast Search Help

Create alert Limits Advanced

COVID-19 is an emerging, rapidly evolving situation.  
Get the latest public health information from CDC: <https://www.coronavirus.gov>.  
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.

**Saccharomyces cerevisiae (baker's yeast)**  
Reference genome: [Saccharomyces cerevisiae S288C \(assembly R64\)](#)  
Download sequences in FASTA format for **genome**  
Download genome annotation in GFF, GenBank  
BLAST against Saccharomyces cerevisiae genome  
All 815 genomes for species:  
Browse the list  
Download sequence and annotation from RefSeq  
Try NCBI Datasets - a new way to download genome

Filters: [Manage Filters](#)

Find related data  
Database: Select  
Find items

Search details  
yeast[All Fields]  
Search See more...

Recent activity

See [UBE2E3 \(YEAST\) ubiquitin conjugating enzyme E2 E3](#) in the Gene database

Display Settings: Summary, 20 per page Send to:

**Search results**  
Items: 1 to 20 of 98

<< First < Prev Page 1 of 5 Next > Last >>

[https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF\\_000146045.2\\_R64/GCF\\_000146045.2\\_R64\\_genomic.fna.gz](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.fna.gz)

Open Link in New Tab  
Open Link in New Window  
Open Link in Incognito Window  
Save Link As...  
Copy Link Address  
Copy  
Search Google for "genome"  
Print...  
Inspect  
Speech  
Services

Right click on the link and click "Copy Link Address"

## STEP 2: Once the link has been copied, return to your Logrus session and execute the wget command

```
# wget command enables you to download files from the web directly to your server.
#A useful command to download genome fasta and annotation files from publicly available databases.

wget ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_29/GRCh38.p12.genome.fa.gz

# gunzip is used to extract the contents within a zipped file (uncompress).

gunzip GRCh38.p12.genome.fa.gz

# The above file will contain the nucleotide sequence of all regions in the GRCh38.p12 assembly.

# To download the genome annotation in a GFF3 file format

wget ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_29/\
gencode.v29.chr_patch_hapl_scaff.annotation.gff3.gz

gunzip gencode.v29.chr_patch_hapl_scaff.annotation.gff3.gz

# The above file will contain comprehensive gene annotation on the reference chromosomes,
# scaffolds, assembly patches and alternate loci (haplotypes) within the genome level assembly.
```

## IN-CLASS EXERCISE TO NAVIGATE THROUGH SOME DATABASES TO DOWNLOAD GENOME AND ANNOTATION FILES

There are several publicly available databases that store genome and annotation files, including ENSEMBL, TAIR, Flybase, etc. While some genomes are resolved to the chromosome level, such as those of model organisms, many others have only a draft assembly available.

### ACTIVITY 1 (download assigned genome and annotation files from NCBI)

In the following activity, you will download genomes and the associated annotation file for the assigned organism from NCBI using the command wget.

```
# Create a folder in your workspace called activity_1

cd /home/$USER/

mkdir activity_1

cd activity_1

# Make a directory called "reference" and download your genome into that folder

mkdir reference

cd reference

wget "paste the link here"
```