

## Accessing Existing Sequence Resources

## Major Databases

- National Center for Biotechnology Information (NCBI)
- European Molecular Biology Laboratory (EMBL)
- DNA DataBank of Japan (DDBJ)



## NCBI

- Literature - PubMed
- Genomes
- Variation (dbVar, dbGaP, dbSNP)
- Gene expression (GEO, SRA)
- Nucleotides (GenBank, SRA)
- Proteins
- BLAST resource

## Genomes

**Zea mays**  
 Representative genome: **Zea mays (assembly B73 RefGen\_v4)**  
 Download sequences in FASTA format for genome, transcript, protein  
 Download genome annotation in GFF, GenBank or tabular format  
 BLAST against Zea mays genome, transcript, protein  
 All 14 genomes for species:  
 Browse the [list](#)  
 Download sequence and annotation from [RefSeq](#) or [GenBank](#)

Display Settings: Overview Send to: ▾

**Organism Overview** ; [Genome Assembly and Annotation report \[14\]](#) ; [Plasmid Annotation Report](#) ID: 12  
[\[1\]](#) ; [Organelle Annotation Report \[3\]](#)

**Zea mays**  
 Zea mays Organism overview

Lineage: Eukaryota[5016]; Viridiplantae[590]; Streptophyta[518]; Embryophyta[512]; Tracheophyta[567];  
 Spermatophyta[503]; Magnoliopsida[480]; Liliopsida[88]; Poales[59]; Poaceae[57]; PACMAD clade[58];  
 Panicoideae[17]; Andropogonodae[7]; Andropogoneae[7]; Tripsacinae[1]; Zea[1]; Zea mays[1]

Maize is an economically important crop, along with rice and wheat. It is the premier cash crop in the United States. In addition to being used as grain and fodder, it is also used extensively in pharmaceutical production as well as a commodity feedstock for other organic chemical products like rubber, ethanol and plastic. Apart from its economic importance, [More...](#)

**Summary**

**Sequence data:** genome assemblies: 14; sequence reads: 97 (See [Genome Assembly and Annotation report](#))  
**Statistics:** median total length (Mb): 2171.65  
 median protein count: 52470  
 median GC%: 46.7774  
**NCBI Annotation Release:** 102

**Publications (limited to 20 most recent records)**

ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

Genome   [Create alert](#) [Limits](#) [Advanced](#) [Help](#)

## GenBank

- Annotated collection of all publicly available nucleotide and amino acid sequences
- New full release is made every two months, updates are made daily
- Part of the International Nucleotide Sequence Database Collaboration with DDBJ and EMBL
  - Three organizations exchange data on a daily basis
- <http://www.ncbi.nlm.nih.gov/genbank/>

## GeneBank – Sample Record

Explore the sample record at:

<http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>

The screenshot shows a web browser displaying a GenBank sample record. The browser's address bar shows the URL [www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html](http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html). The page title is "Sample GenBank Record". Below the title, there are links for "PubMed", "Entrez", and "BLAST". The main content area is titled "GenBank Flat File Format" and contains a table of metadata for a specific sample. The table includes fields such as LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS, SOURCE, ORGANISM, REFERENCE, and AUTHORS. The sample is identified as "Saccharomyces cerevisiae (baker's yeast)" and the record is for a "TCF1-beta gene, partial cds, and Axl2p (Axl2) and Rev7p (REV7) genes, complete cds.".

LOCUS	SC049845	5028 bp	DNA	PLN	21-JUN-1999
DEFINITION	Saccharomyces cerevisiae TCF1-beta gene, partial cds, and Axl2p (Axl2) and Rev7p (REV7) genes, complete cds.				
ACCESSION	U49845				
VERSION	U49845.1	GI:1293613			
KEYWORDS					
SOURCE	Saccharomyces cerevisiae (baker's yeast)				
ORGANISM	Saccharomyces cerevisiae				
REFERENCE	1 (bases 1 to 5028)				
AUTHORS	Torrey, J. D., Gibbe, P. E., Nelson, J. and Lawrence, C. W.				
TITLE	Cloning and sequence of REV7, a gene whose function is required for DNA damage-induced mutagenesis in Saccharomyces cerevisiae				
JOURNAL	Yeast 10 (11), 1503-1509 (1994)				
PubMed	7871890				
REFERENCE	2 (bases 1 to 5028)				
AUTHORS	Roemer, T., Madden, K., Chang, J. and Snyder, M.				
COMMENT	Relation of axial growth sites in yeast genomes. Axl2p, a novel				

## NCBI Datasets

Retrieval system designed for searching several linked databases

- Books, PubMed, Genbank (Core Nt, EST, GSS), SRA, GEO, etc

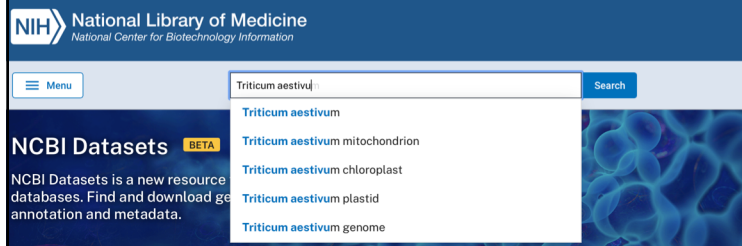
The screenshot shows the NCBI Datasets homepage. At the top, there is a search bar with the text "Search NCBI" and a "Search" button. Below the search bar, there is a section titled "NCBI Datasets" with a brief description: "NCBI Datasets is a new resource that lets you easily gather data from across NCBI databases. Find and download gene, transcript, protein and genome sequences, annotation and metadata." Below this, there is a "Genomes" section with a link to "Browse and download genome data using our species pages. Genome data includes genome, transcript and protein sequences, genome annotation and metadata." To the right of the "Genomes" section is a "Species Browser" section with a link to "View taxonomic relationships and find genome data for closely related species using our interactive species browser." Below these sections, there are four small images representing different biological groups: Eukaryota (a butterfly), Bacteria (a cluster of purple bacteria), Archaea (a red, elongated microorganism), and Viruses (a yellow, elongated virus particle).

## Searching on NCBI

- Can use Boolean operators to query
  - **AND:** to 'AND' two search terms together instructs Entrez to find all documents that contain BOTH terms
  - **OR:** To 'OR' two search terms together instructs Entrez to find all documents that contain EITHER term.
  - **NOT:** To 'NOT' two search terms together instructs Entrez to find all documents that contain search term 1 BUT NOT search term 2

## Searching on NCBI

- <https://www.ncbi.nlm.nih.gov/datasets/>
- How many sequences are there for wheat (*Triticum aestivum*) in the Nucleotide Core Collection, the SRA, GEO?

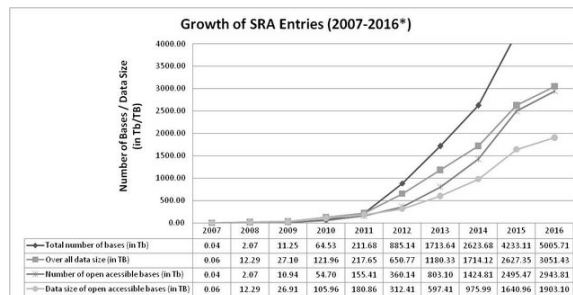


## Searching on NCBI

1. How many sequences are there for wheat (*Triticum aestivum*) from Chinese Spring?
2. How many sequences for wheat (*Triticum aestivum*) are not from Chinese Spring?

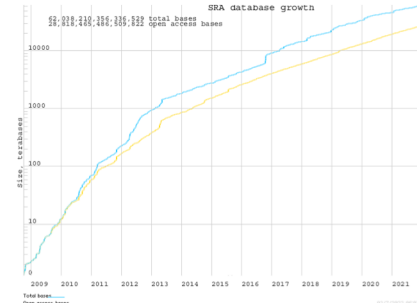
## Downloading Reads from the SRA

- SRA stores raw data from NGS platforms
- As with GeneBank, information is shared with EMBL and DDBJ
- <http://www.ncbi.nlm.nih.gov/sra>



## Downloading Reads from the SRA

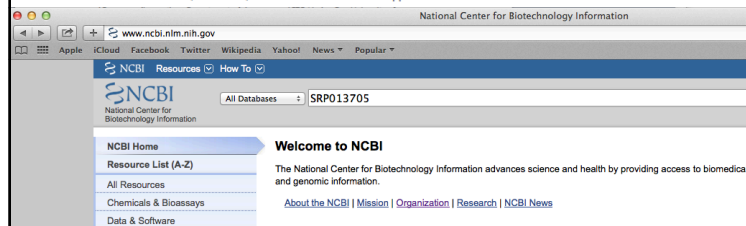
- SRA stores raw data from NGS platforms
- As with GeneBank, information is shared with EMBL and DDBJ
- <http://www.ncbi.nlm.nih.gov/sra>



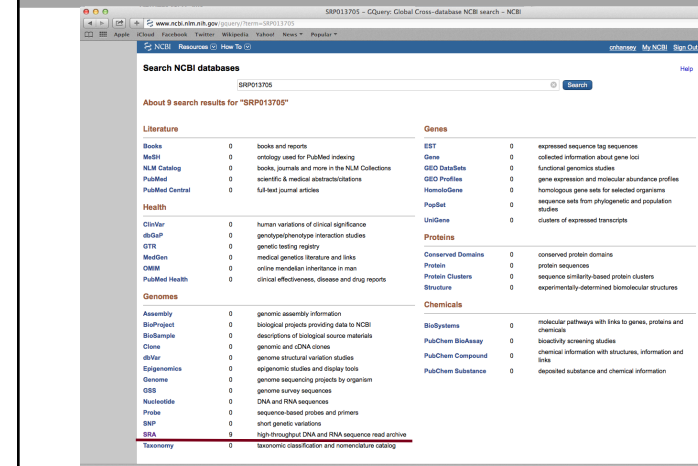
## Downloading Reads from the SRA

### Example from a published paper

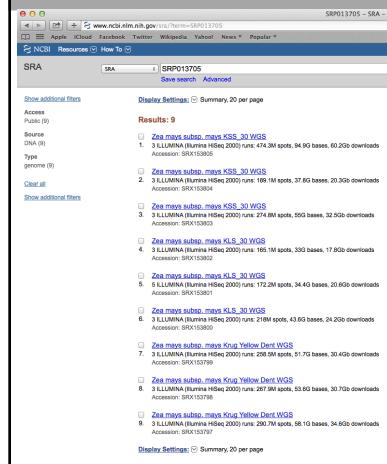
Copyright © 2014 by the Genetics Society of America  
doi: 10.1534/genetics.114.167155  
Manuscript received June 12, 2014; accepted for publication July 8, 2014; published Early Online July 17, 2014.  
Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167155/-DC1>.  
Sequence data from this article have been deposited with the Sequence Read Archive at the National Center for Biotechnology Information study under accession no. SRP013705.  
Corresponding author: Department of Agronomy, 1575 Linden Dr., University of Wisconsin, Madison, WI 53706. E-mail: [smkaepp@wisc.edu](mailto:smkaepp@wisc.edu)



## Downloading Reads from the SRA



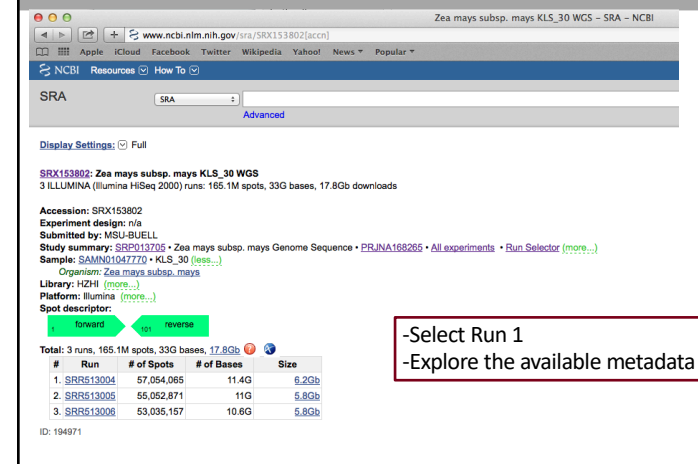
## Downloading Reads from the SRA



-Select accession SRX153802  
-Number 4 in this list  
-Explore the available metadata

KSS\_30  
KLS\_30  
Krug Yellow Dent

## Downloading Reads from the SRA



-Select Run 1  
-Explore the available metadata

## Downloading Reads from the SRA

NCBI Sequence Read Archive

Run: SRR513013, Spots: 160.4M, Bases: 32.1Gbp, Size: 21.8G, GC content: 45.6%, Published: 2014-07-09, Access Type: public

This run has 2 reads per spot:

Legend: L=100, 100%

Experiment: SRX153805, Library: IACA, Illumina, WGS, GENOMIC, RANDOM, PAIRED

Biosample: SAMN01047771 (SRS345374), Sample Description: Zea mays subsp. mays, Organism: Zea mays subsp. mays

Bioproject: PRJNA168265, SRA Study: SRP013705, Title: Zea mays subsp. mays Genome sequencing

Abstract: Zea mays subsp. mays Genome Sequence

You can also click on the "Reads" tab to see individual reads within this run before downloading

## Downloading Reads from the SRA

Go to the 'Data access' tab

Copy the file link and use wget to download the file

```
> wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos1/sra-pub-run-2/SRR513004/SRR513004.1
```

Still need to convert from SRA format to .fastq format (see next slide)

NCBI Sequence Read Archive

COVID-19 is an emerging, rapidly evolving situation. Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (PHE)

Zea mays subsp. mays KLS\_30 WGS (SRR513004)

Metadata | Analysis | Reads | Data access

SRA archive data

SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Size	Location	Name	Free Express	Access Type
run	6,488,700 kb	GCP	gs://sra-pub-run-2/SRR513004/SRR513004.1	?	Use Cloud Data Delivery
run	6,488,700 kb	AWS	s3://sra-pub-run-2/SRR513004/SRR513004.1	worldwide	anonymous
run	6,488,700 kb	AWS	s3://sra-pub-run-2/SRR513004/SRR513004.1	us-east-1	aws identity

Express and Access: what does it mean?

Why is SRA data in the cloud?

## Downloading Reads from the SRA

- Download the SRA toolkit at

<https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit>

```
cnhirsch@ln1002 [~/summer_institute] % module load sratoolkit/2.8.2
cnhirsch@ln1002 [~/summer_institute] % fastq-dump --split-files -F -Q 33 SRR19443996
```

```
cnhirsch@ln1002 [~/summer_institute] % ls
SRR19443996_1.fastq SRR19443996_2.fastq
```

## Accessing Existing Resources Takeaways

- There is a ton of available data with so much untapped potential
- Use it!!
- When you deposit data – provide all metadata to make it useful