Accessing Existing Sequence Resources

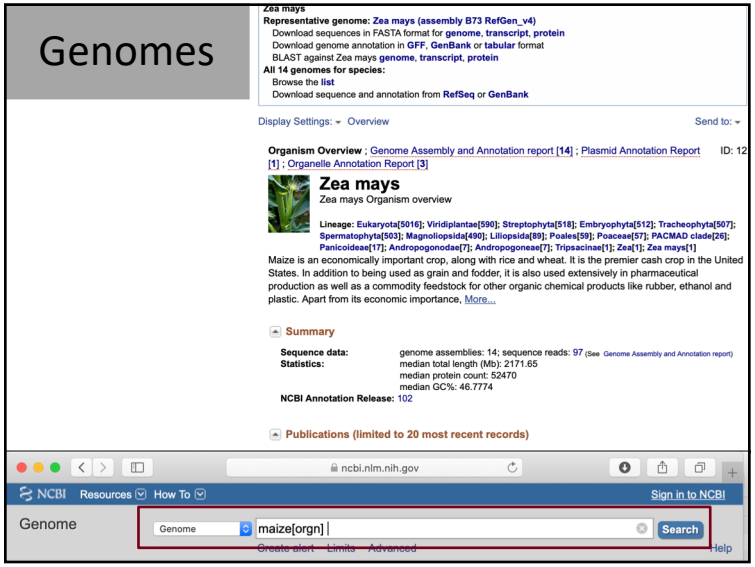
Major Databases

- 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



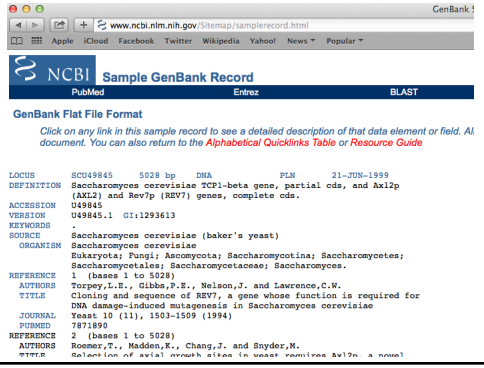
GenBank

- 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:

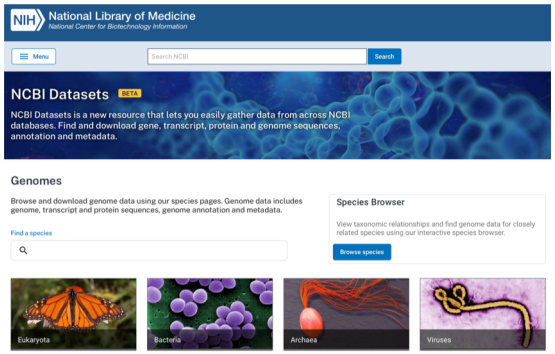
http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html



NCBI Datasets

Retrieval system designed for searching several linked databases

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

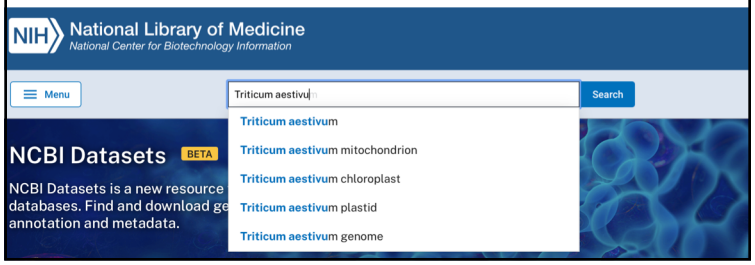


Searching on NCBI

- 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?

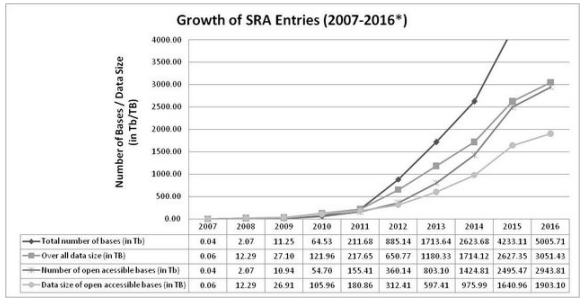


Searching on NCBI

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

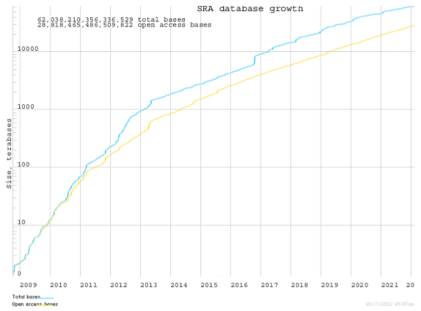
- SRA stores raw data from NGS platforms
- As with GeneBank, information is shared with EMBL and DDBJ
- http://www.ncbi.nlm.nih.gov/sra



	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Total number of bases (in Tb)	0.04	2.07	11.25	64.53	211.68	885.14	1713.64	2623.68	4233.11	5005.71
Over all data size (in TB)	0.06	12.29	27.10	121.96	237.65	650.77	1180.33	1714.12	2627.35	3051.43
Number of open accessible bases (in Tb)	0.04	2.07	10.94	54.70	155.41	360.14	803.10	1424.81	2495.47	2943.81
Data size of open accessible bases (in TB)	0.06	12.29	26.91	105.96	180.86	312.41	597.41	975.99	1640.96	1903.10

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



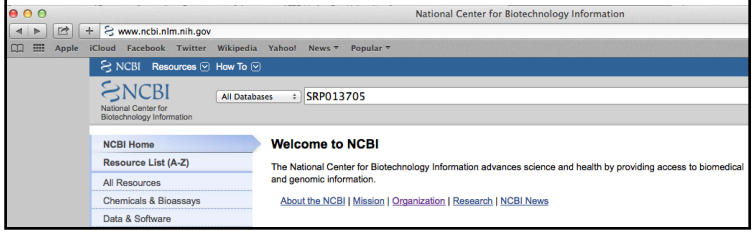
Downloading Reads from the SRA

Example from a published paper

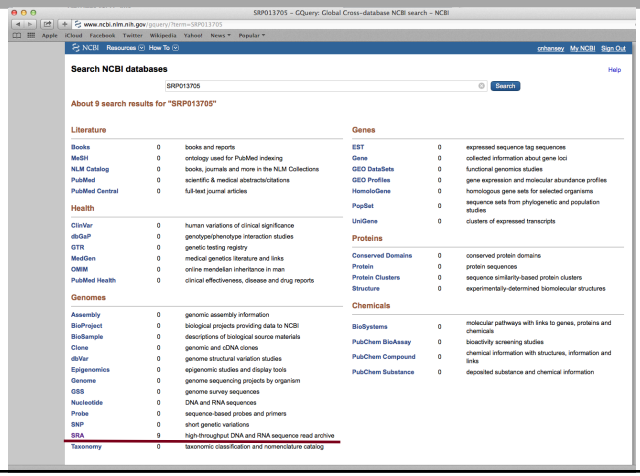
doi: 10.1534/genetics.114.167155
Manuscript received June 12, 2014; accepted for publication July 8, 2014; published Early Online July 17, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167155/-/DC1.

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.

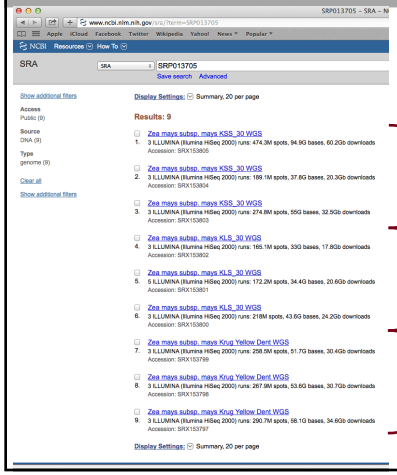
[1]Corresponding author: Department of Agronomy, 1575 Linden Dr., University of Wisconsin, Madison, WI 53706. E-mail: smkaeppl@wisc.edu



Downloading Reads from the SRA



Downloading Reads from the SRA



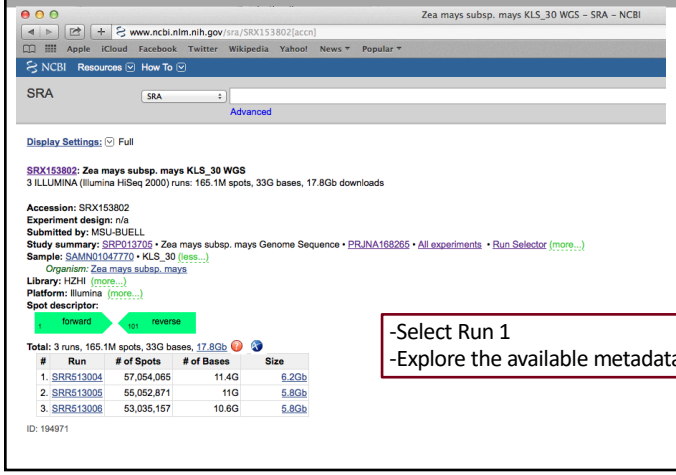
-Select accession SRX153802
-Number 4 in this list
-Explore the available metadata

KSS_30

KLS_30

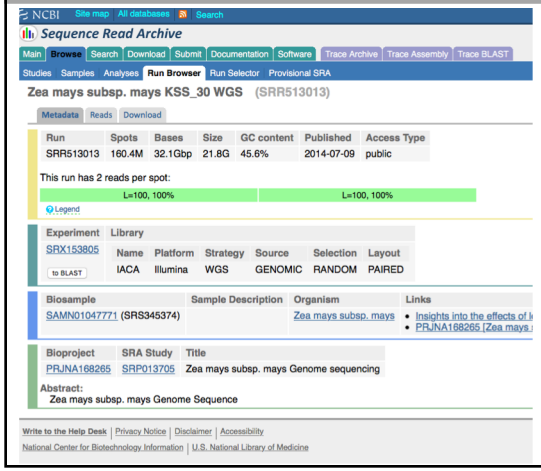
Krug Yellow Dent

Downloading Reads from the SRA



-Select Run 1
-Explore the available metadata

Downloading Reads from the SRA



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

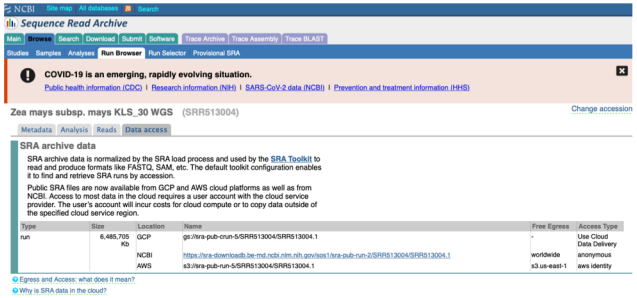
Downloading Reads from the SRA

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)



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