NCBI Resources for Curation and Maintenance of Genomes

Brian Smith-White, Paul Kitts, Francoise Thibaud-Nissen, Kim Pruitt, Valerie Schneider, Terence Murphy

Two kinds of resources at NCBI

Primary Data Archives

- INSDC
  - GenBank
  - dbEST, dbGSS
- dbVar
- dbSNP
- Probe
- Trace
- SRA
- Geo
- Biosample
- BioProject

Reference Collections

- RefSeq
- Gene
- UniGene
- Clone DB

Difference between the two resources

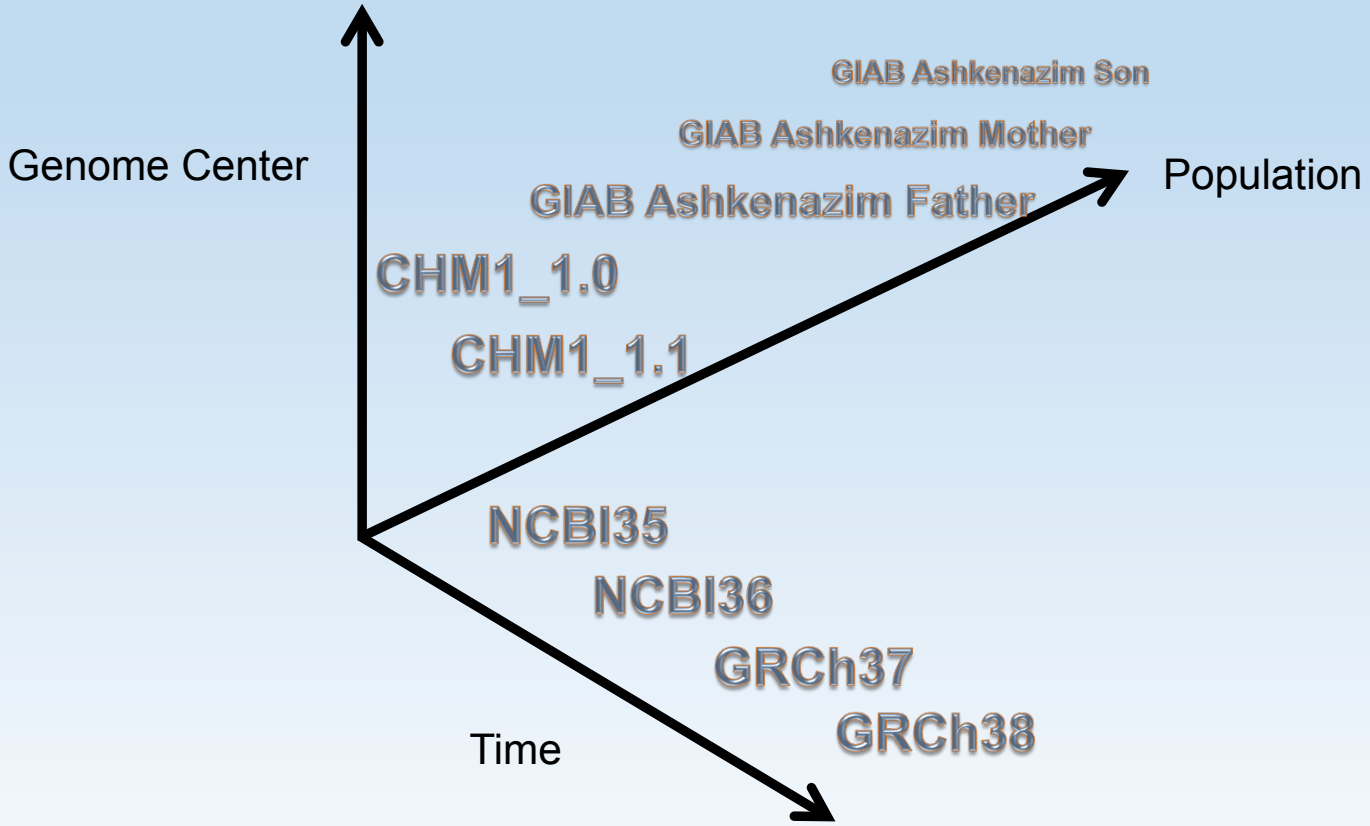
Primary Data Archives

- A database record is derived from submission of work performed outside of NCBI
- The submitter owns the record

Reference Collections

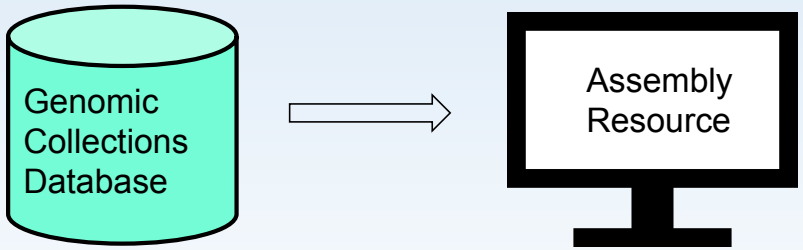
- A database record is derived through computation and/or curation by NCBI upon primary data archive records and other public data
- NCBI owns the record

Need To Track Different Assemblies

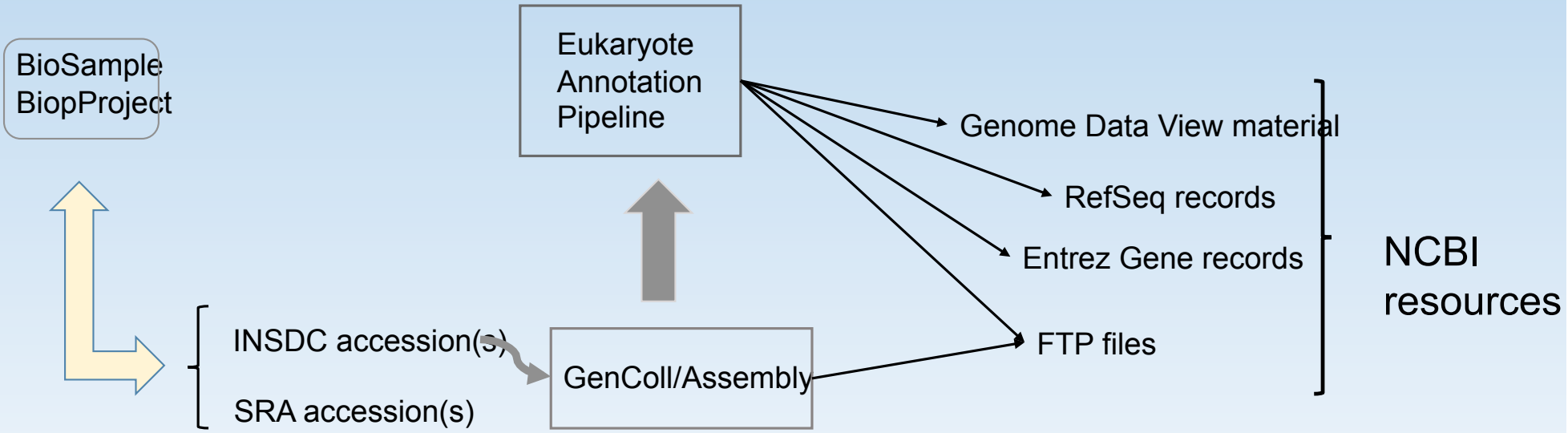


Roles Of The Genomic Collections Database (GenColl)

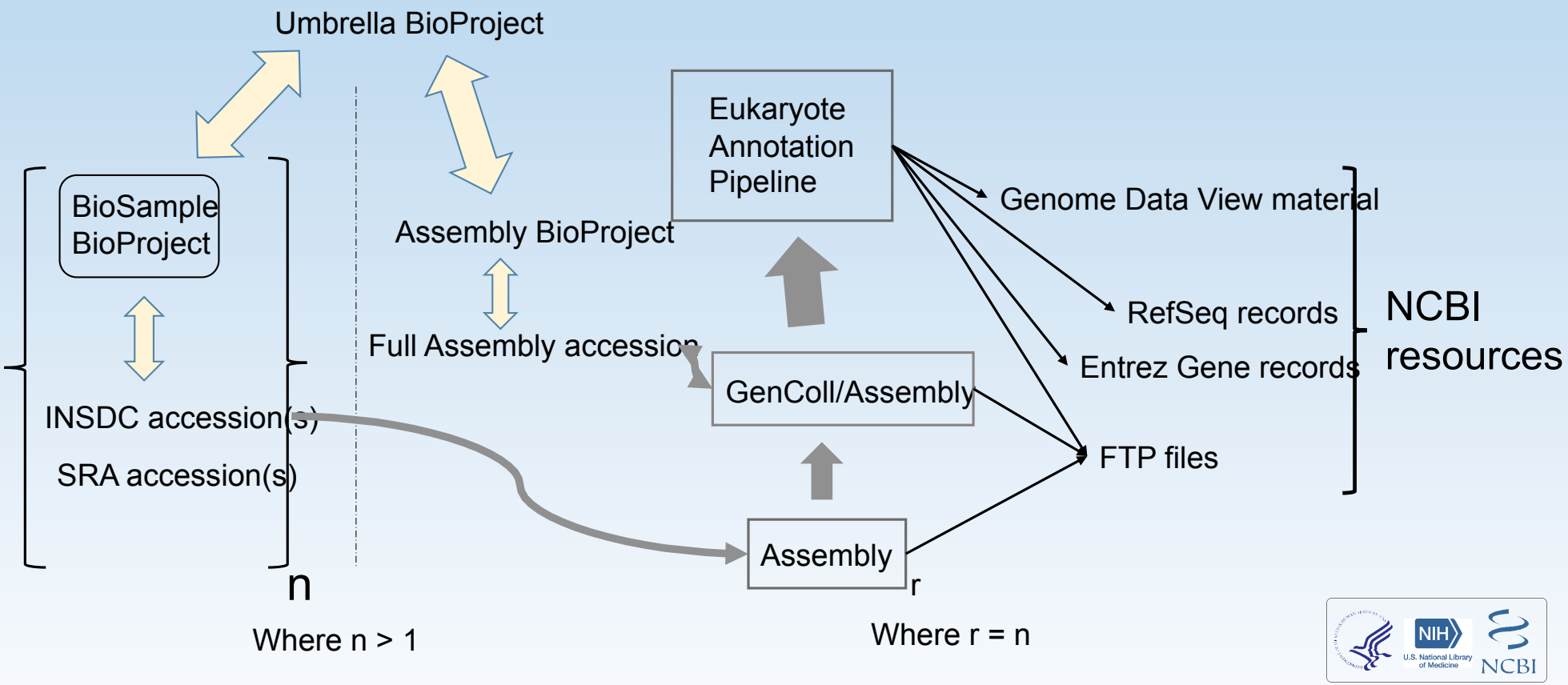
- Records the set of sequences that constitute an assembly
- Assigns an accession & version to the assembly as a whole
  - GCA_#########.# for GenBank assemblies
  - GCF_#########.# for RefSeq assemblies
- Tracks successive versions of an assembly
- Defines the role of each object in the assembly hierarchy
- Organizes & stores assembly metadata
- Calculates and stores numerous statistics for each assembly
- Tracks the relationship between a GenBank assembly and its RefSeq assembly pair



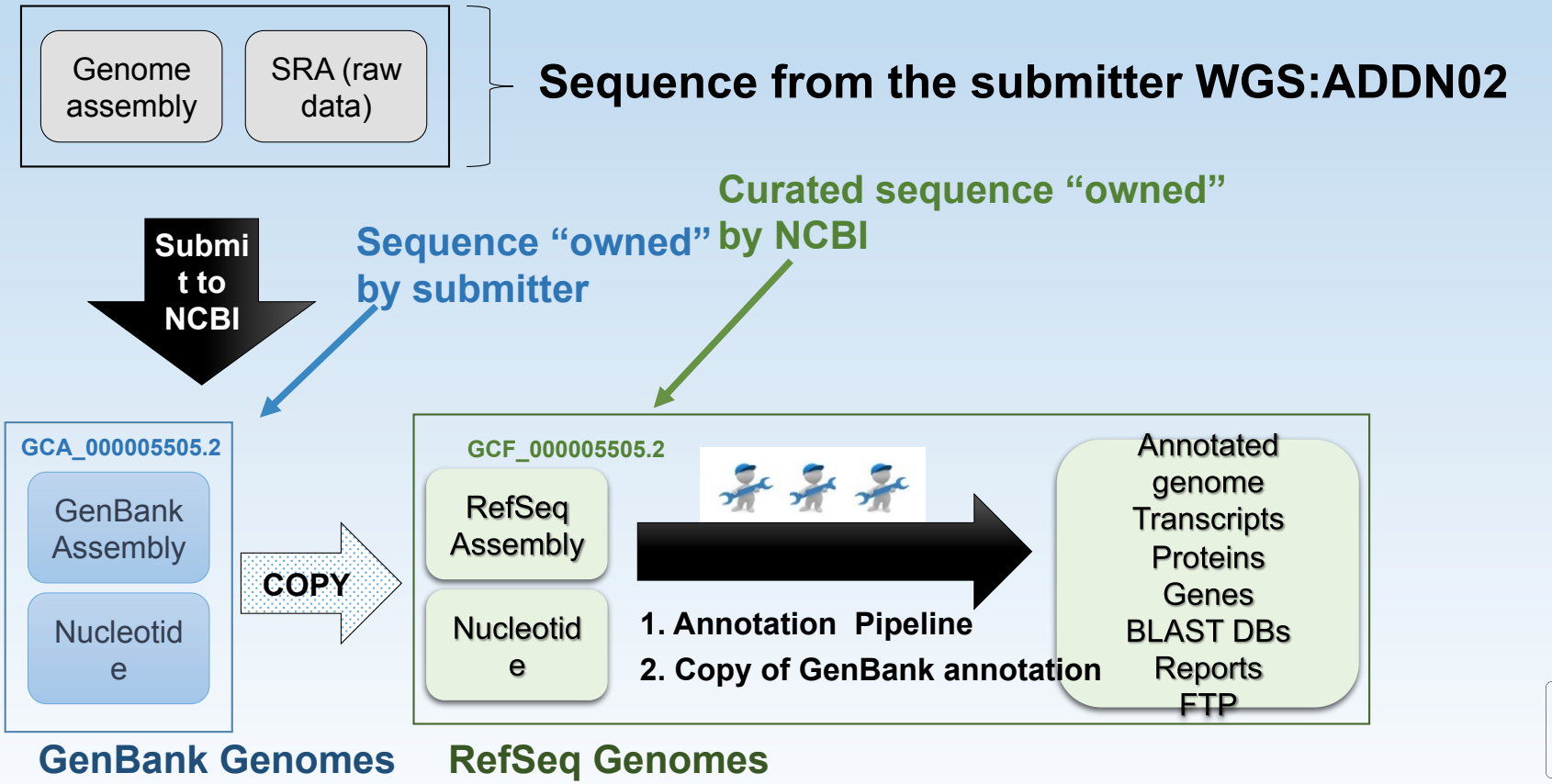
Unified (or synchronous) genome submission



Asynchronous genome submission



Flow to Produce a RefSeq Assembly



Data differences between barley and Brachypodium

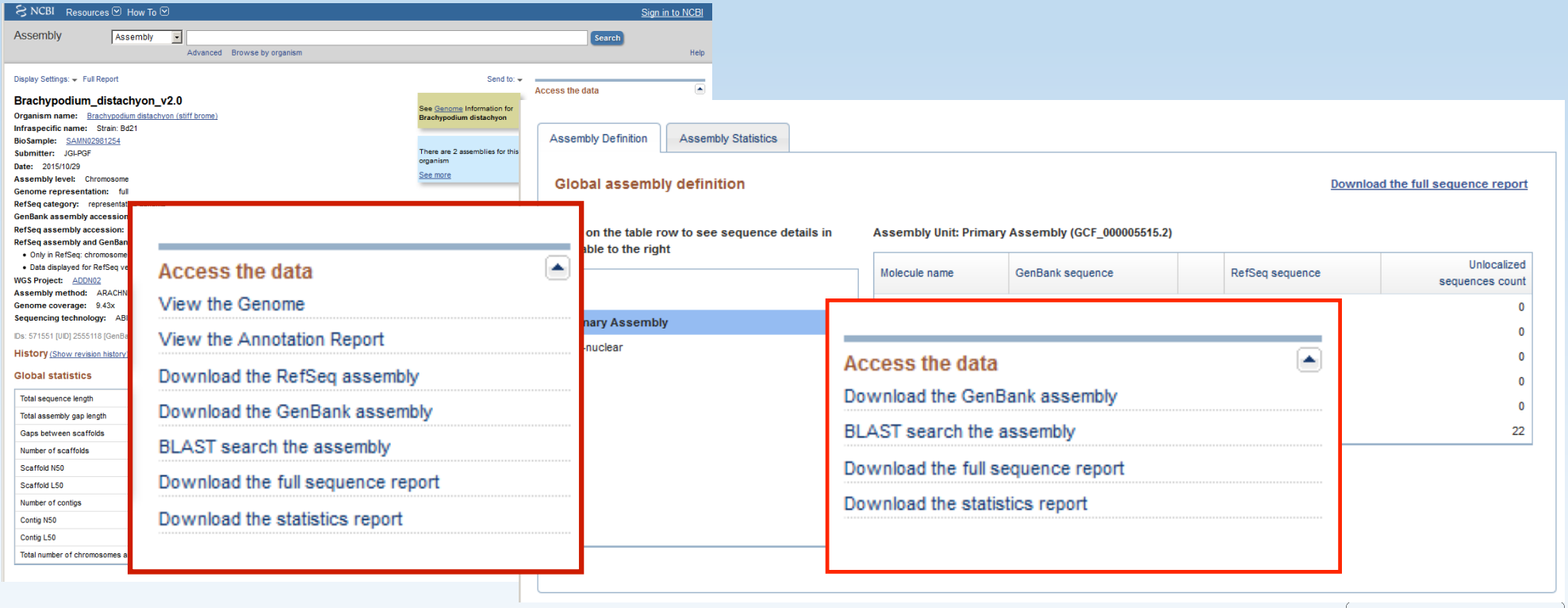
Brachypodium

- Chromosomes submitted to INSDC
- Material in INSDC
  - Scaffold number - 28
  - Scaffold N50 - 59,384,932
  - Scaffold L50 – 3
- There are RefSeq chromosomes which have been processed by Eukaryote Annotation Pipeline

Barley

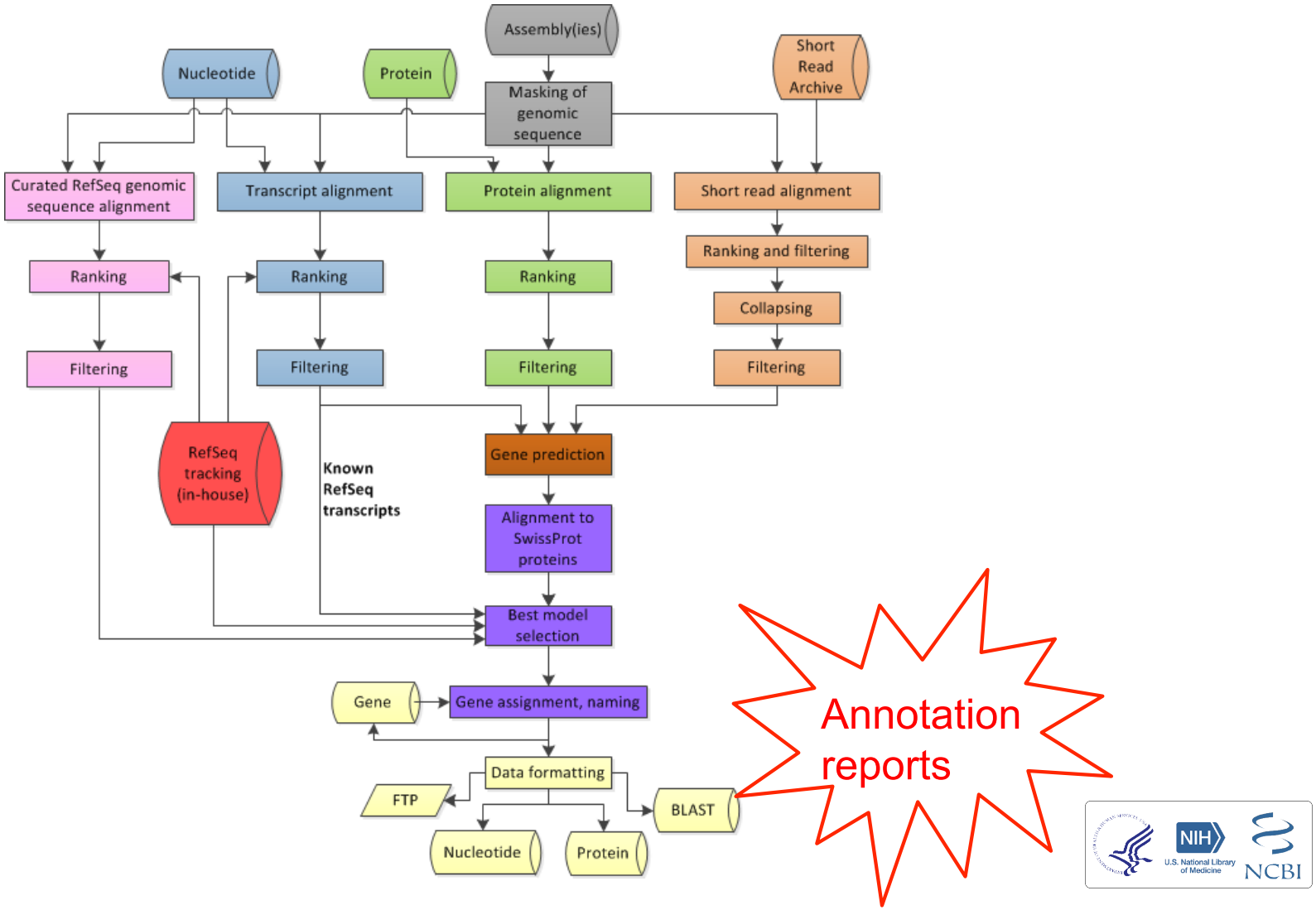
- FASTA sequences labelled as chromosome-specific at EnsemblPlants FTP site
- Nothing with chromosome biomol type submitted to INSDC
- Material in INSDC
  - Scaffold number – 2,280,908
  - Scaffold N50 – 1,469
  - Scaffold L50 – 242,748

Assembly Display



Brachypodium

barley


Assembly(ies)
Nucleotide
Protein
Masking of genomic sequence
Short Read Archive
Curated RefSeq genomic sequence alignment
Transcript alignment
Protein alignment
Short read alignment
Ranking
Ranking
Ranking
Ranking and filtering
Collapsing
Filtering
Filtering
Filtering
Filtering
RefSeq tracking (in-house)
Known RefSeq transcripts
Gene prediction
Alignment to SwissProt proteins
Best model selection
Gene
Gene assignment, naming
Data formatting
FTP
BLAST
Nucleotide
Protein
Annotation reports
NIH
U.S. National Library of Medicine
NCBI

Comparative reports

- Pair features between two sets of annotation based on coordinates

- Use cases:

- Between two annotation releases
- Between two co-annotated assemblies
- Between NCBI and an external annotation (i.e. GenBank)

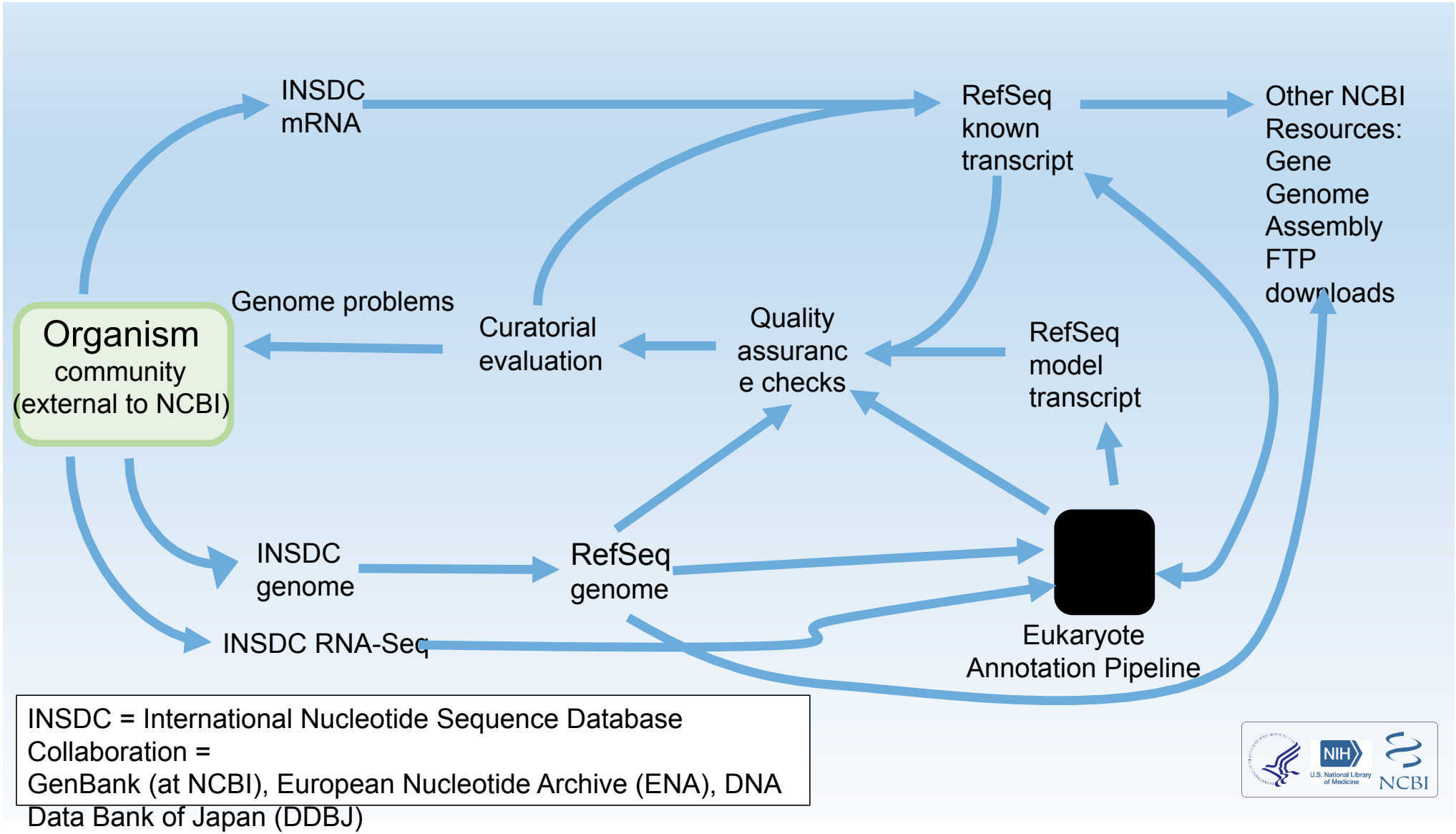
- Report provided as

- summary counts of mapped, new and deprecated features
- tab-delimited file
- Genome WorkBench project

Reference Sequence (RefSeq) collection

- Comprehensive
- Integrated with other NCBI resources
- Non-redundant
- Subject to NCBI curation
- Available through:
  - BLAST
  - Entrez search
  - NCBI FTP site
- Consistency in sequence and annotation
- Up-to-date nomenclature
- Genome annotation using a consistent method
- Expanded feature annotation
- Connected to functional information

PMID: 22121212



INSDC = International Nucleotide Sequence Database Collaboration =
GenBank (at NCBI), European Nucleotide Archive (ENA), DNA Data Bank of Japan (DDBJ)

NCBI Gene – a central source for information related to an individual gene

- Nomenclature
- RefSeq transcripts
- Maps
- Pathways
- Phenotypes
- Gene-specific literature
- Links to external resources

NCBI Gene display

Bibliography

Related articles in PubMed

1. Evolution of AGL6-like MADS box genes in grasses (Poaceae): ovule expression is ancient and palea expression is new.
   Reinheimer R, *et al*. Plant Cell, 2009 Sep. PMID 19749151, Free PMC Article
2. Genome-wide analysis of the MADS-box gene family in Brachypodium distachyon.
   Wei B, *et al*. PLoS One, 2014 Jan 13. PMID 24454749, Free PMC Article

GeneRIFs: Gene References Into Functions What's a GeneRIF?

Submit: New GeneRIF Correction

General gene information

Homology

The Hierarchical Catalog of Orthologs

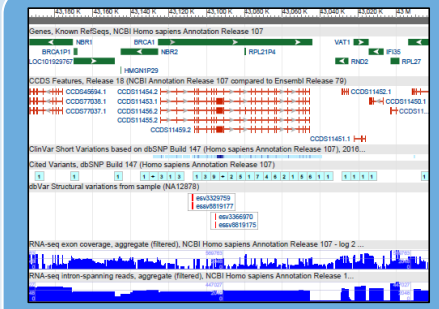
General protein information

Preferred Names
MADS-box transcription factor 6

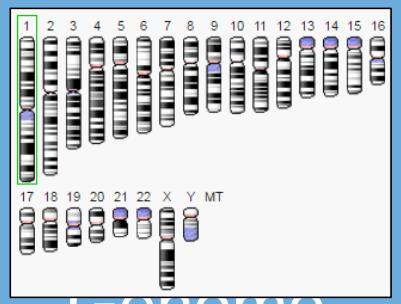
Names
AGL6-like MADS box transcription factor
MADS-box transcription factor 28

Data visualization tools at NCBI

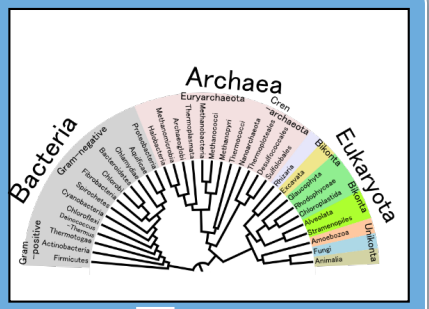
- View gene annotation, variation, expression
- Navigate genome annotation
- Explore taxonomic relationships



Sequence Viewer



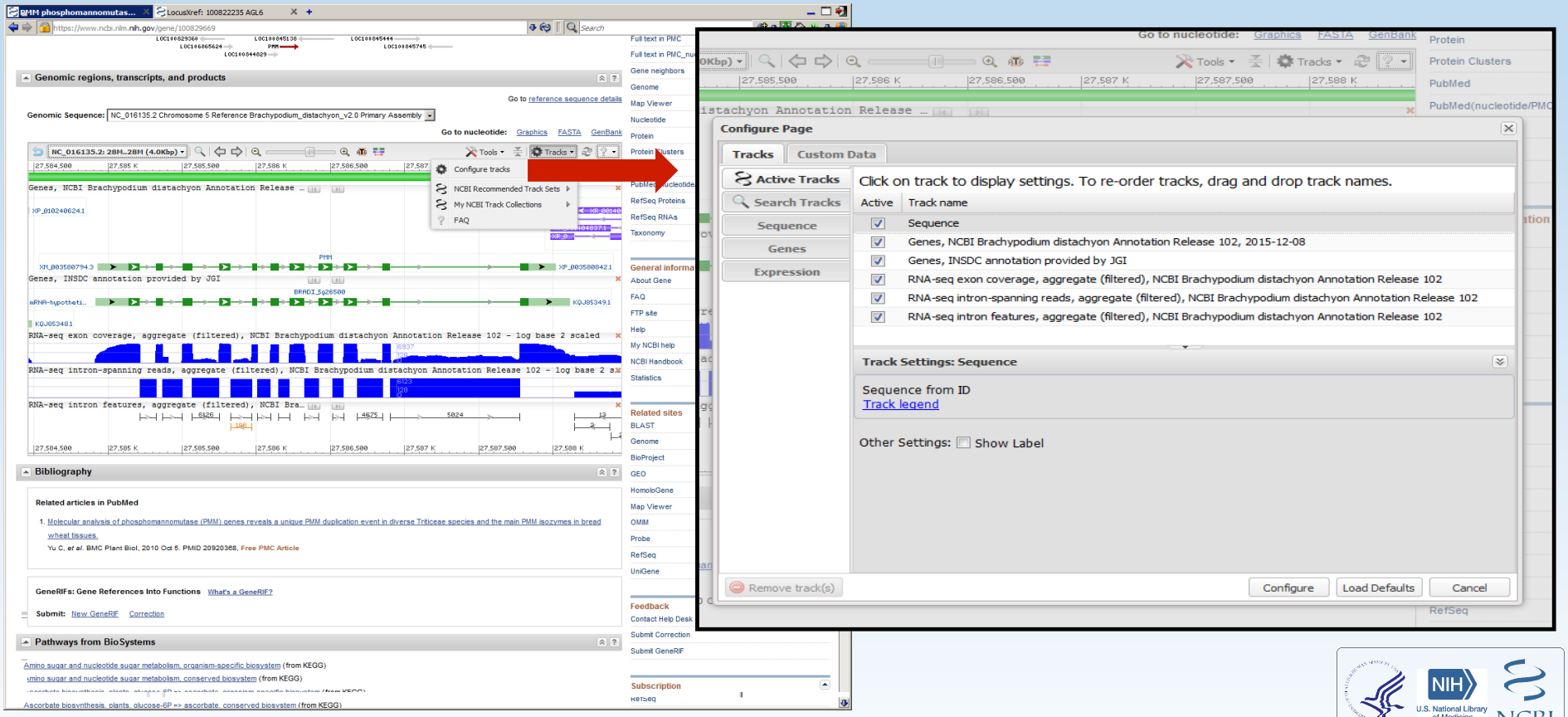
Genome Data Viewer



Tree Viewer

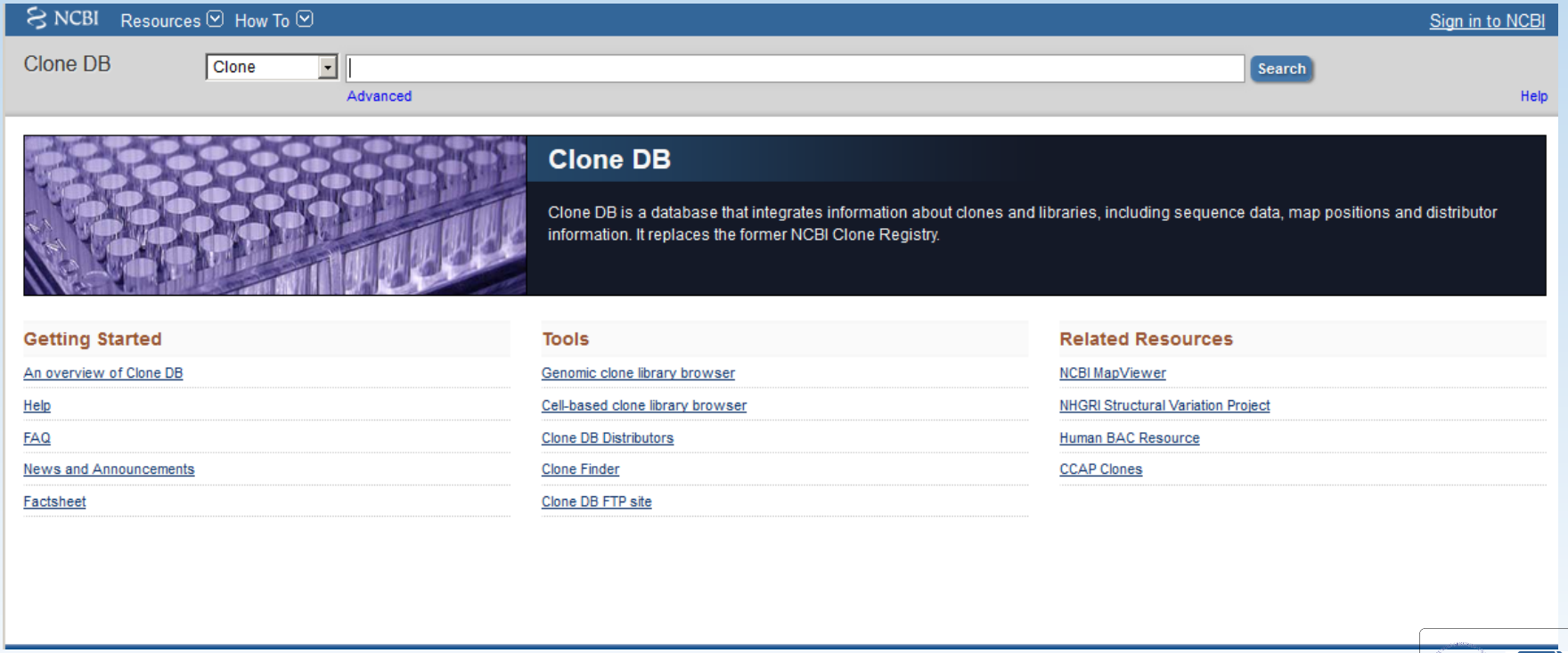
NIH U.S. National Library of Medicine NCBI

Sequence viewer



Genome Data View

Clone DB (formerly Clone Registry)



NIH U.S. National Library of Medicine NCBI

Clone DB in Genome Data View

Now and the future

Initial genome submission

Assembly – with FTP files

- gi deprecated – accession.version still live

Eukaryote Annotation Pipeline

Gene – with FTP files

RefSeq – with FTP files

Genome Data View

Clone DB – with FTP files

Subsequent genome submission

All of initial resources

Remapping Service