

- This software/database/presentation is a "United States Government Work" under the terms of the United States Copyright Act. It was written as part of the author's official duties as a United States Government employee and thus cannot be copyrighted. This software/database/presentation is freely available to the public for use. The National Library of Medicine and the U.S. Government have not placed any restriction on its use or reproduction.

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

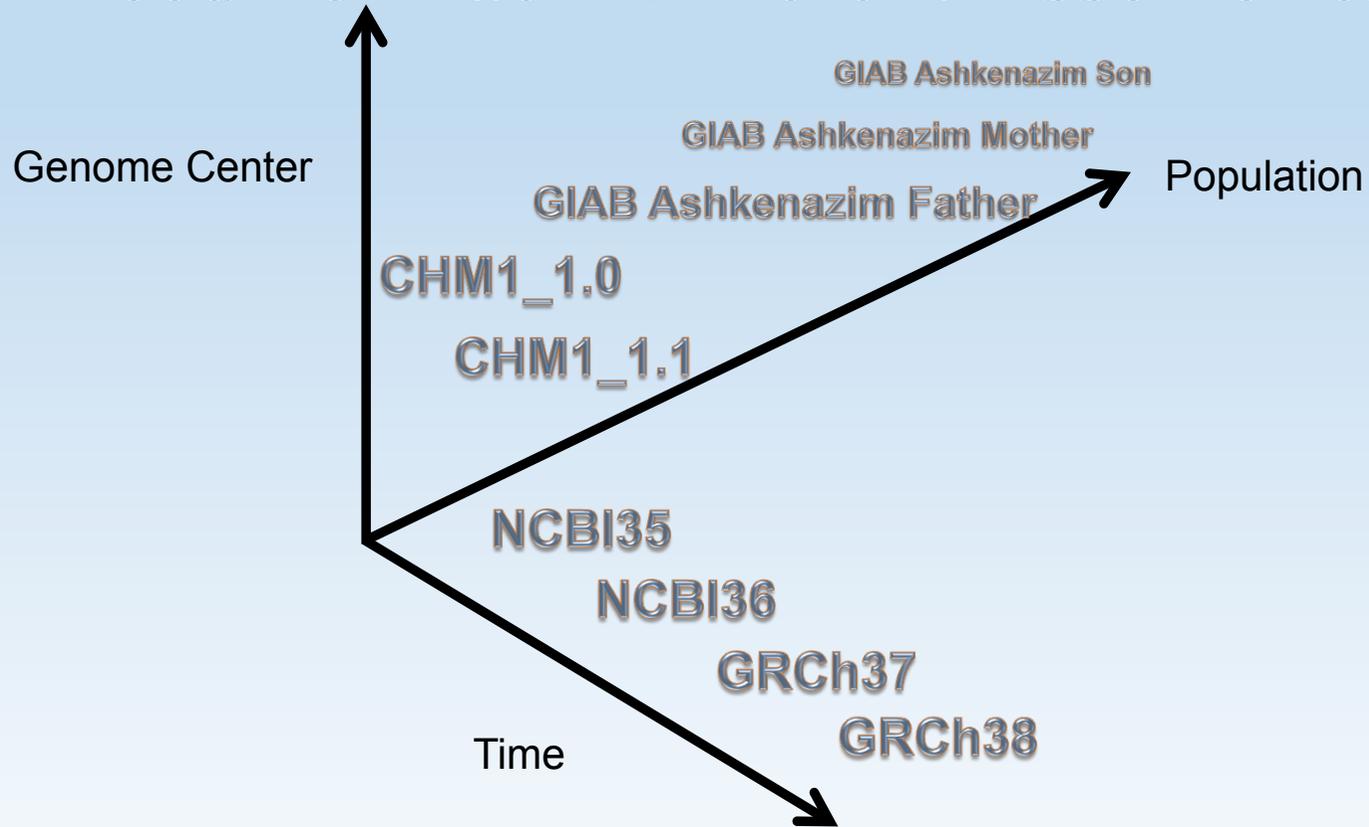
Primary Data Archives

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

Reference Collections

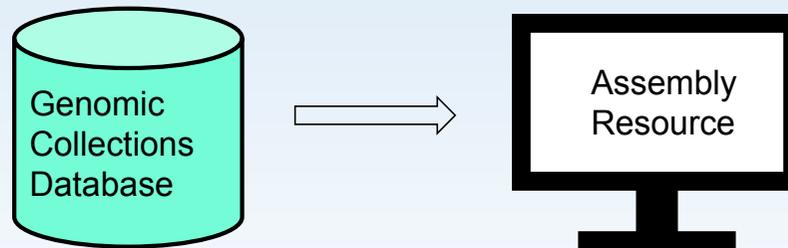
- A database record is derived through computation and/or curation by NCBI upon primary data archive records and other public data
- NCBI owns the record

Need To Track Different Assemblies

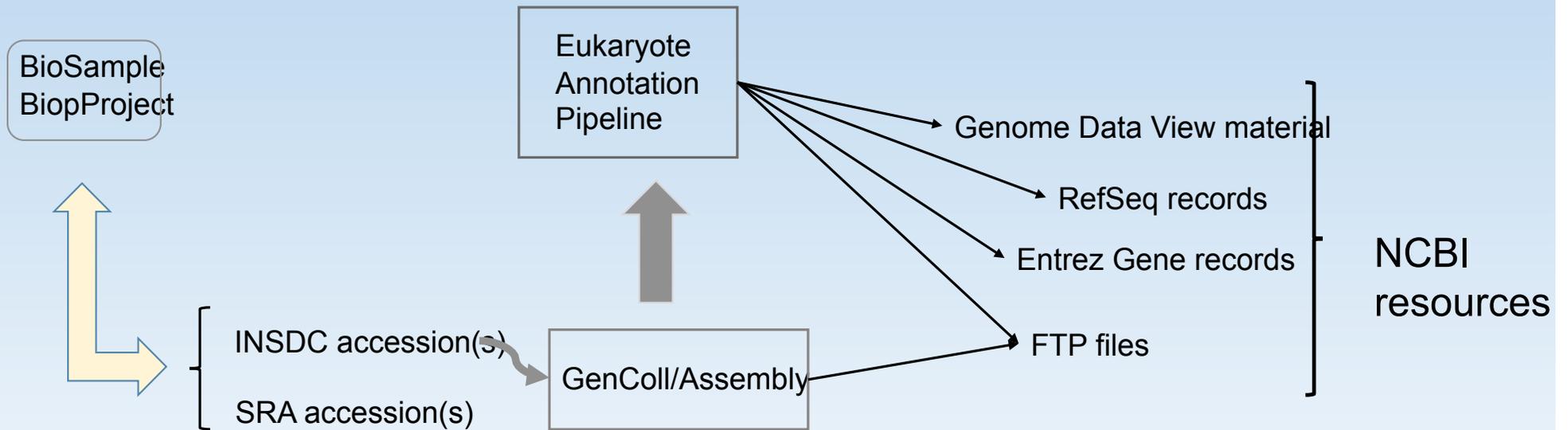


Roles Of The Genomic Collections Database (GenColl)

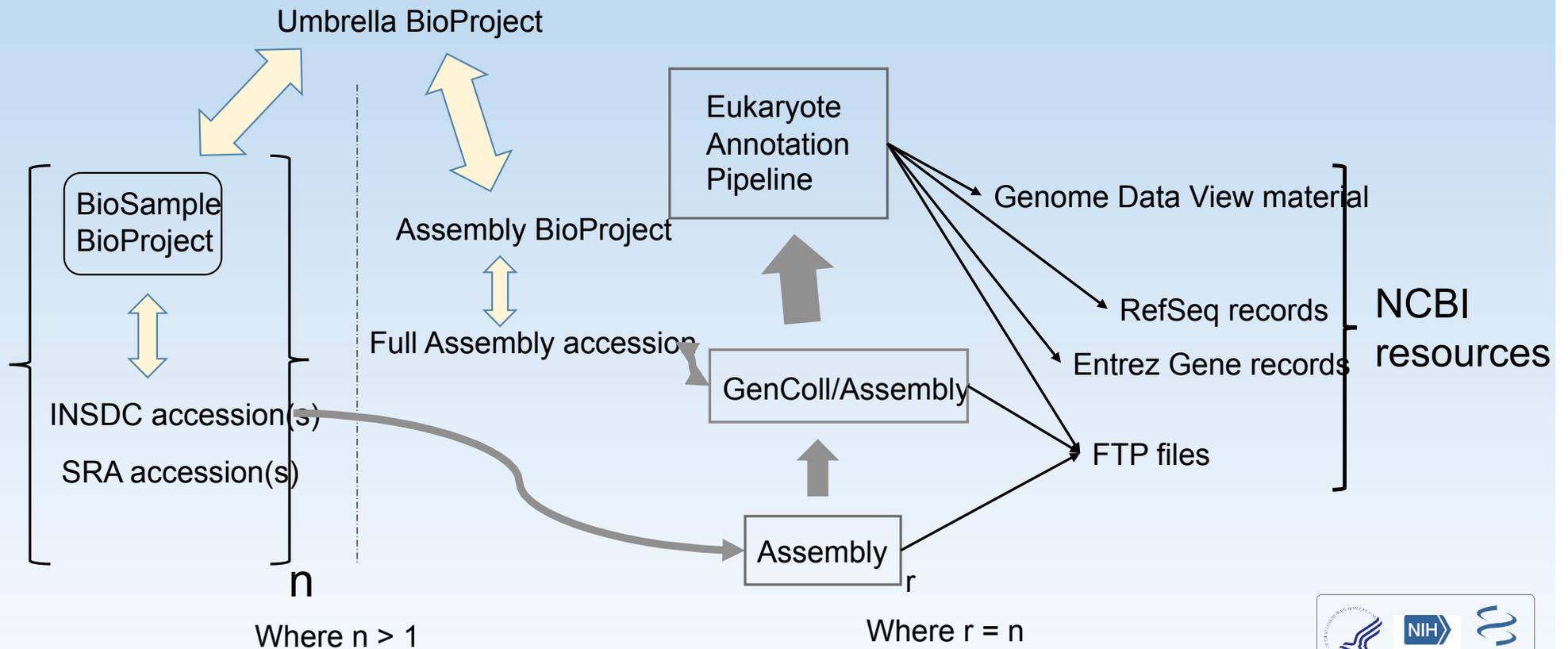
- Records the set of sequences that constitute an assembly
- Assigns an accession & version to the assembly as a whole
 - GCA_#####.# for GenBank assemblies
 - GCF_#####.# for RefSeq assemblies
- Tracks successive versions of an assembly
- Defines the role of each object in the assembly hierarchy
- Organizes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



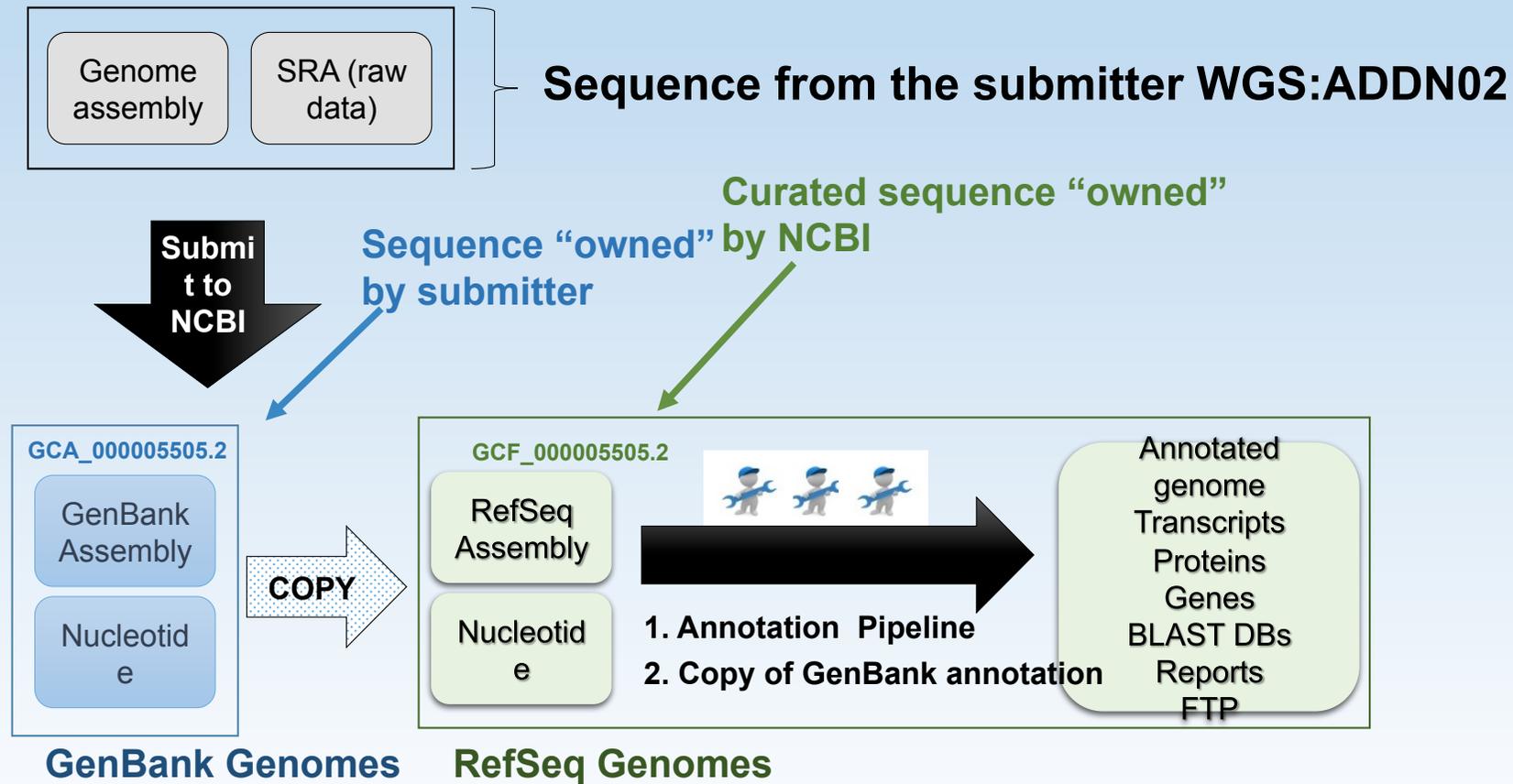
Unified (or synchronous) genome submission



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 distachyon_v2.0
 Organism name: [Brachypodium distachyon \(stiff bromo\)](#)
 Intraspecific name: Strain: Bd21
 BioSample: [SAMN02981254](#)
 Submitter: JGI-PGF
 Date: 2015/10/29
 Assembly level: Chromosome
 Genome representation: full
 RefSeq category: representational
 GenBank assembly accession: [GCF_000005515.2](#)
 RefSeq assembly accession: [GCF_000005515.2](#)
 RefSeq assembly and GenBank assembly: [GCF_000005515.2](#)
 • Only in RefSeq: chromosome
 • Data displayed for RefSeq version: [v2.0](#)
 WGS Project: [ADDN02](#)
 Assembly method: ARACHN
 Genome coverage: 9.43x
 Sequencing technology: ABI
 IDs: 571551 [UID] 2555118 [GenBank]

Global assembly definition

Download the full sequence report

Assembly Unit: Primary Assembly (GCF_000005515.2)

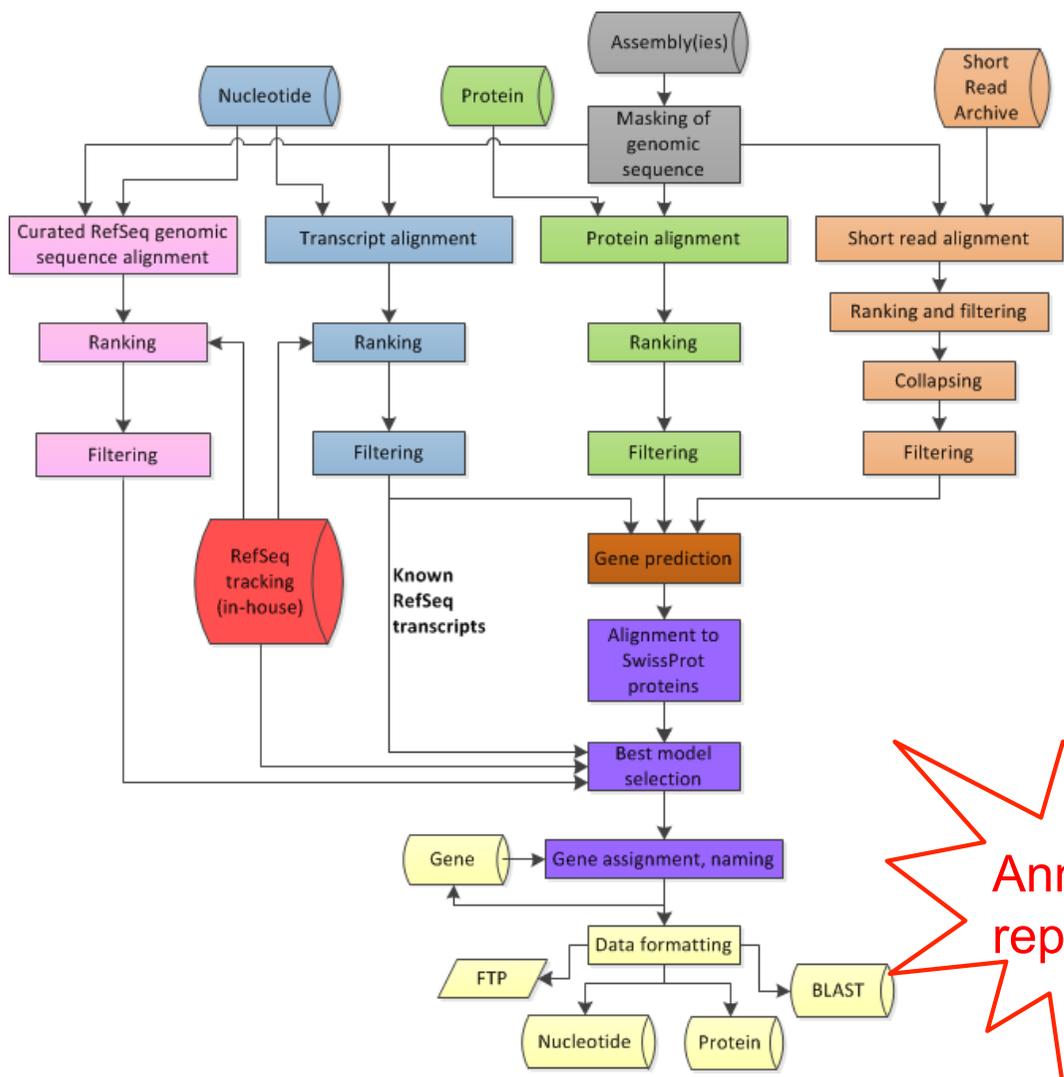
Molecule name	GenBank sequence	RefSeq sequence	Unlocalized sequences count
			0
			0
			0
			0
			22

Access the data
 View the Genome
 View the Annotation Report
 Download the RefSeq assembly
 Download the GenBank assembly
 BLAST search the assembly
 Download the full sequence report
 Download the statistics report

Access the data
 Download the GenBank assembly
 BLAST search the assembly
 Download the full sequence report
 Download the statistics report

Brachypodium

barley



Annotation reports

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

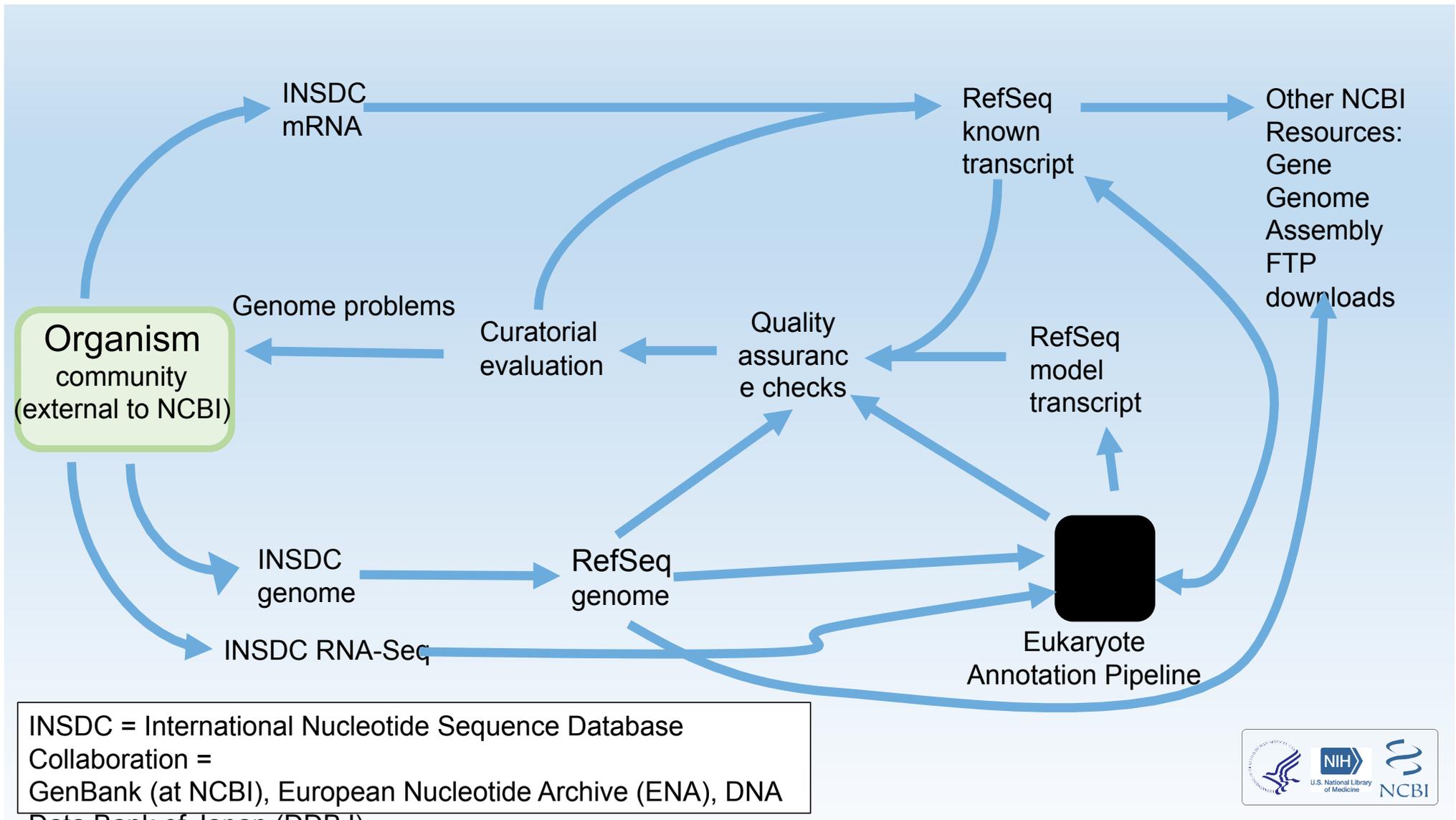


Reference Sequence (RefSeq) collection

- 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





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

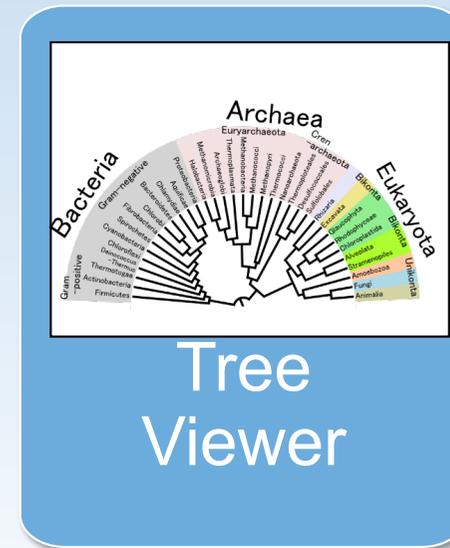
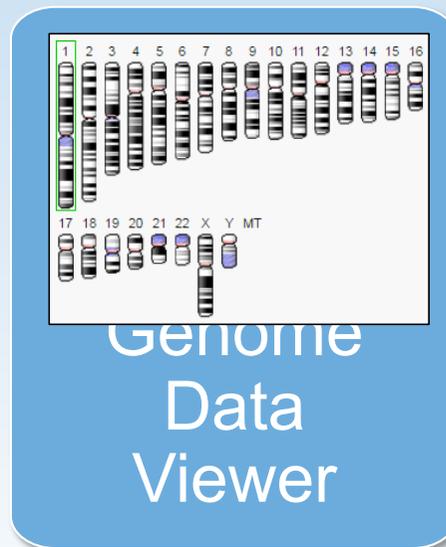
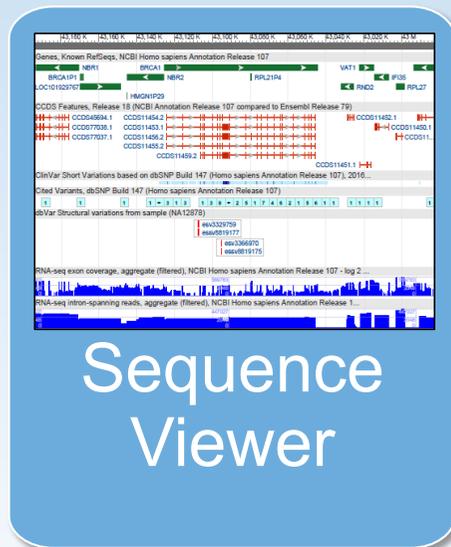
AGL6-like MADS box transcription factor

MADS-box transcription factor 28



Data visualization tools at NCBI

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



Sequence viewer

The screenshot displays the NCBI Sequence Viewer interface for the gene *phosphomannomutase-2* (PMM2) on chromosome 5. The main view shows several tracks: Genes, NCBI Brachypodium distachyon Annotation Release 102, 2015-12-08; RNA-seq exon coverage; RNA-seq intron-spanning reads; and RNA-seq intron features. A red arrow points to the 'Tracks' button in the top right corner, which has opened a 'Configure Page' dialog box.

The 'Configure Page' dialog box has two tabs: 'Tracks' and 'Custom Data'. The 'Tracks' tab is active, showing a list of tracks with checkboxes for their visibility. The 'Active' column shows that the following tracks are currently visible:

- Sequence
- Genes, NCBI Brachypodium distachyon Annotation Release 102, 2015-12-08
- Genes, INSDC annotation provided by JGI
- RNA-seq exon coverage, aggregate (filtered), NCBI Brachypodium distachyon Annotation Release 102
- RNA-seq intron-spanning reads, aggregate (filtered), NCBI Brachypodium distachyon Annotation Release 102
- RNA-seq intron features, aggregate (filtered), NCBI Brachypodium distachyon Annotation Release 102

Below the track list, there are sections for 'Track Settings: Sequence' and 'Other Settings'. The 'Track Settings: Sequence' section shows 'Sequence from ID' and a link to the 'Track legend'. The 'Other Settings' section has a checkbox for 'Show Label' which is currently unchecked. At the bottom of the dialog are buttons for 'Remove track(s)', 'Configure', 'Load Defaults', and 'Cancel'.

The background interface includes a search bar at the top, a 'Go to nucleotide' field, and various tool icons. The 'Bibliography' section at the bottom left lists related articles in PubMed, and the 'Pathways from BioSystems' section at the bottom right lists related metabolic pathways.

Genome Data View

The screenshot displays the NCBI Genome Data Viewer interface for *Brachypodium distachyon*. The main view shows a genomic track for chromosome 5 (NC_016135.2) at approximately 27.56M. The track includes gene annotations (e.g., LOC100828458, LOC100829063, LOC100844829), RNA-seq exon coverage, RNA-seq intron-spanning reads, and INSDC annotations (e.g., BRADI_5g26450, BRADI_5g26510). A search bar on the left contains the query 'pmm'. The interface also features a sidebar with assembly information, a search bar, and a 'Global statistics' table.

Assembly Information:

- Organism name: *Brachypodium distachyon* (stiff br.)
- Infraspecific name: Strain: Bd21
- BioSample: SAMN02861254
- Submitter: JGI-PGF
- Date: 2015/10/29
- Assembly level: Chromosome
- Genome representation: full
- RefSeq category: representative genome
- GenBank assembly accession: GCA_000005505.2
- RefSeq assembly accession: GCF_000005505.2
- RefSeq assembly and GenBank assembly identifiers:
 - Only in RefSeq: chromosome Ptd.
 - Data displayed for RefSeq version
- WGS Project: [ADDN02](#)
- Assembly method: ARACHNE v. 20071016_modif
- Genome coverage: 9.43x
- Sequencing technology: ABI 3739

Global statistics:

Total sequence length
Total assembly gap length
Gaps between scaffolds
Number of scaffolds
Scaffold N50
Scaffold L50
Number of contigs
Contig N50
Contig L50
Total number of chromosomes and plasmids

Clone DB (formerly Clone Registry)

The screenshot shows the NCBI Clone DB website. At the top, there is a navigation bar with the NCBI logo, "Resources" and "How To" dropdown menus, and a "Sign in to NCBI" link. Below this is a search bar with a "Clone" dropdown menu, a search input field, and a "Search" button. A "Help" link is also present. The main content area features a large image of a microplate on the left and a dark blue header for "Clone DB" on the right. The header text states: "Clone DB is a database that integrates information about clones and libraries, including sequence data, map positions and distributor information. It replaces the former NCBI Clone Registry." Below the header, there are three columns of links: "Getting Started" (An overview of Clone DB, Help, FAQ, News and Announcements, Factsheet), "Tools" (Genomic clone library browser, Cell-based clone library browser, Clone DB Distributors, Clone Finder, Clone DB FTP site), and "Related Resources" (NCBI MapViewer, NHGRI Structural Variation Project, Human BAC Resource, CCAP Clones). At the bottom right, there are logos for the NIH (U.S. National Library of Medicine) and 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

