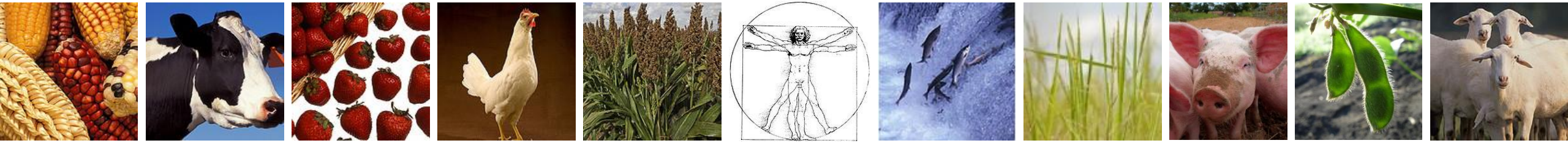


NCBI Annotation of the Water Buffalo Genome

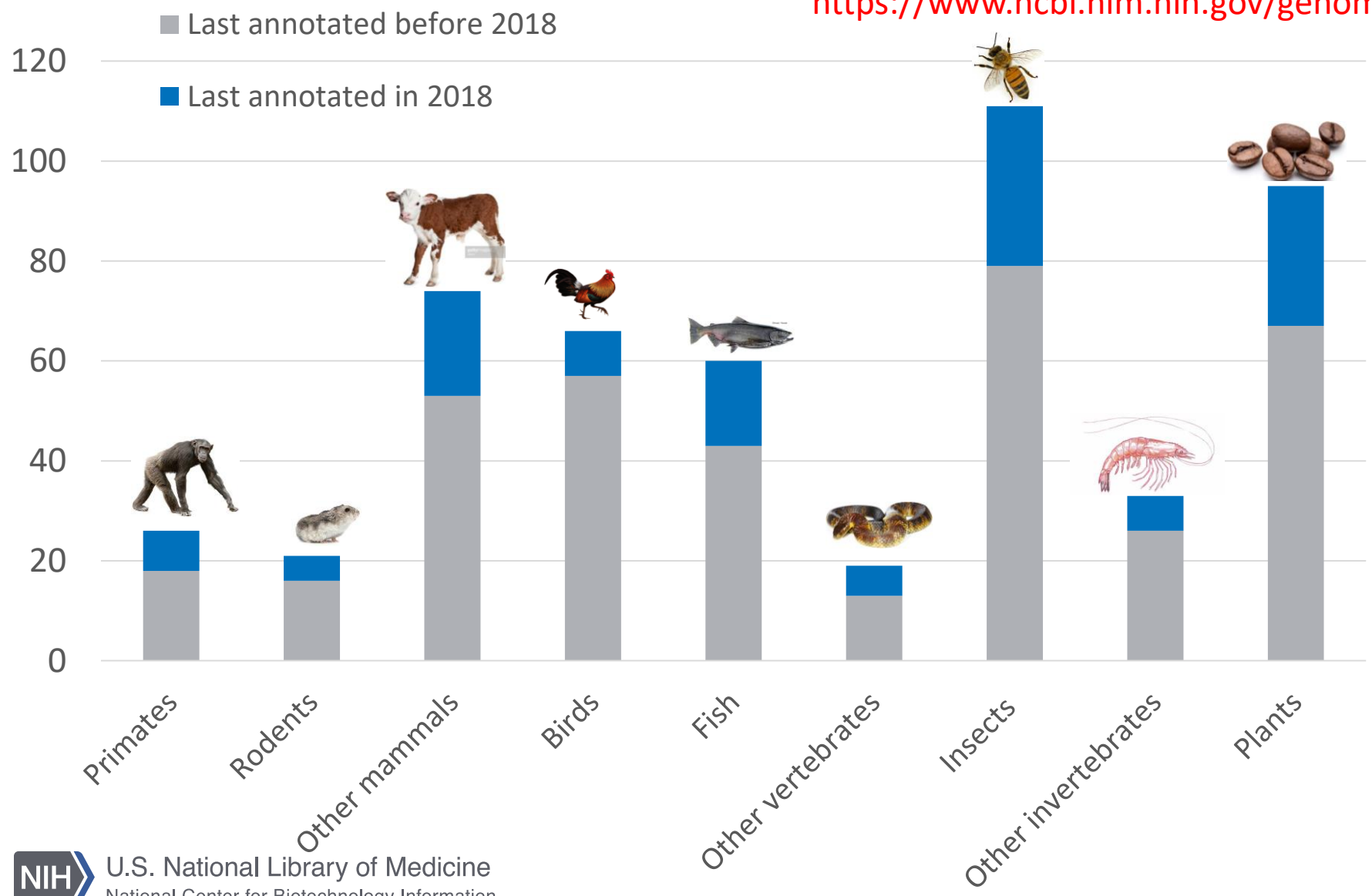
Françoise Thibaud-Nissen (thibaudf@ncbi.nlm.nih.gov)

PAG XXVII January 12, 2019

U.S. National Library of Medicine
National Center for Biotechnology Information

Over 500 eukaryotic species annotated

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/



U.S. National Library of Medicine
National Center for Biotechnology Information

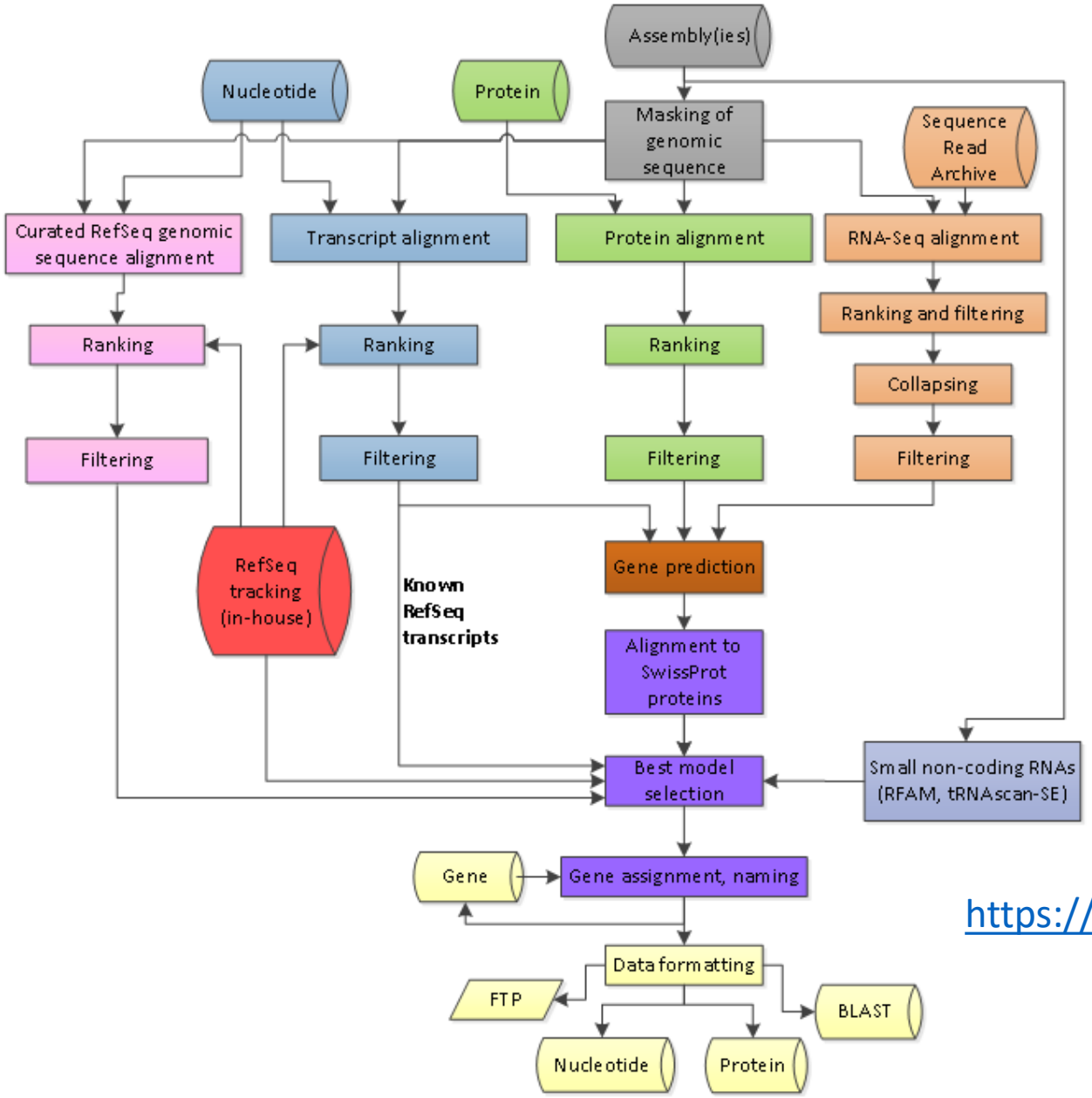
Water buffalo assembly improvements

Assembly	Submission date	Assembly level	Assembly Accession	Contig N50	Scaffold N50	Annotation release
UMD_CASPUR_WB_2.0	09/30/2013	Scaffold	GCA_000471725.1	22 Kb	1.4 Mb	**100**
UOA_WB_1	05/14/2018	Chromosome	GCA_003121395.1	22 Mb	117 Mb	**101**

Inputs for annotation release 101

- 15 billion short reads (RNA-Seq) from 70 tissues
- Same-species proteins, ESTs, and cDNAs
- RefSeq proteins from human and cow

The NCBI Eukaryotic Genome Annotation Pipeline



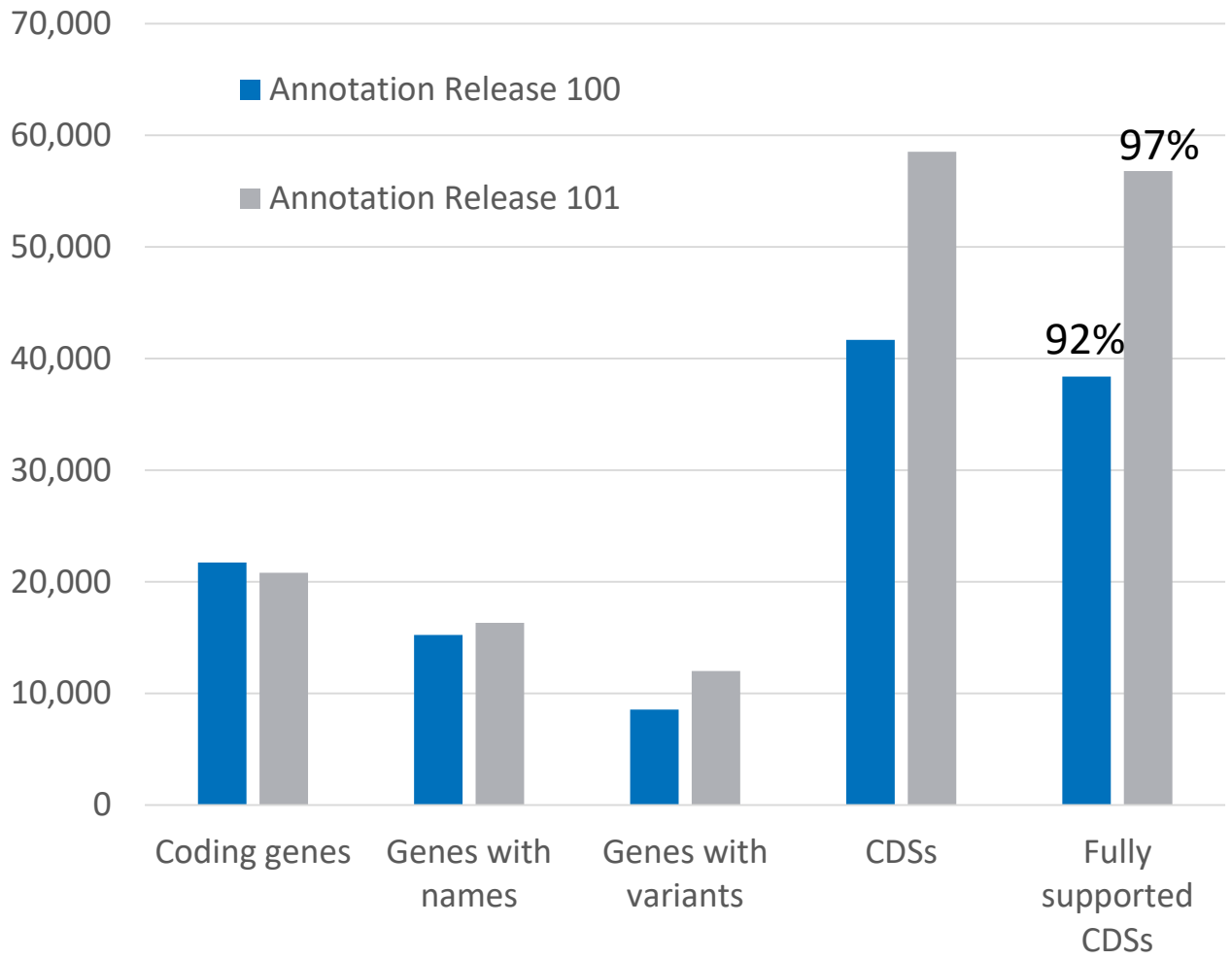
- Automated
- 1 to 2 weeks start to end
- Consumes public data!
- Evidence driven:
  - Same species and close cross-species transcripts, proteins
  - RNA-Seq and IsoSeq
  - TSA (Transcript Shotgun Assemblies)

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

Annotation release 101 facts

- 20,801 protein-coding genes
  - 16,327 genes with names
- 8,443 non-coding genes

- Over 600 of AR100 genes merged



Merged genes: Transferrin Receptor Protein (TFRC)

B: NC_037545.1 (Graphical Sequence View) [GCF_003121395]

Assembly: UOA_WB_1 (GCF_003121395.1)

115,450 K | 115,452 K | 115,454 K | 115,456 K | 115,458 K | 115,460 K | 115,462 K | 115,464 K | 115,466 K | 115,468 K | 115,470 K | 115,472 K | 115,474 K | 115,476 K | 115,478 K | 115,480 K | 115,482 K

- Sequence NC_037545.1: Bubalus bubalis breed Mediterranean chromosome 1, ASM312139v1, whole genome shotgun sequence

- Level 1, Subtracks: 2 on, 3 off

Annotation release 101

- Genes, Subtracks: 1 on, 2 off

-- Genes, NCBI Bubalus bubalis Annotation Release 101, 2018-06-25, 1 gene model shown

TFRC

XM_025285613.1

XP_025141398.1

XM_025285617.1

XP_025141402.1

XM_025285606.1

XP_025141391.1

XM_025285624.1

XP_025141409.1

- Alignments, Subtracks: 3 on, 0 off

Annotation release 100

-- Other alignments, total 5 objects shown

NW_0057845826.1

LOC102414334

XR_329299.1

NW_0057845857.1

NW_005784594.1

LOC102407099

XM_006054243.1

XP_006054305.1

NW_005784594.1

LOC102407404

XM_006054245.1

XP_006054307.1

XM_006054244.1

XP_006054306.1

XM_006054246.1

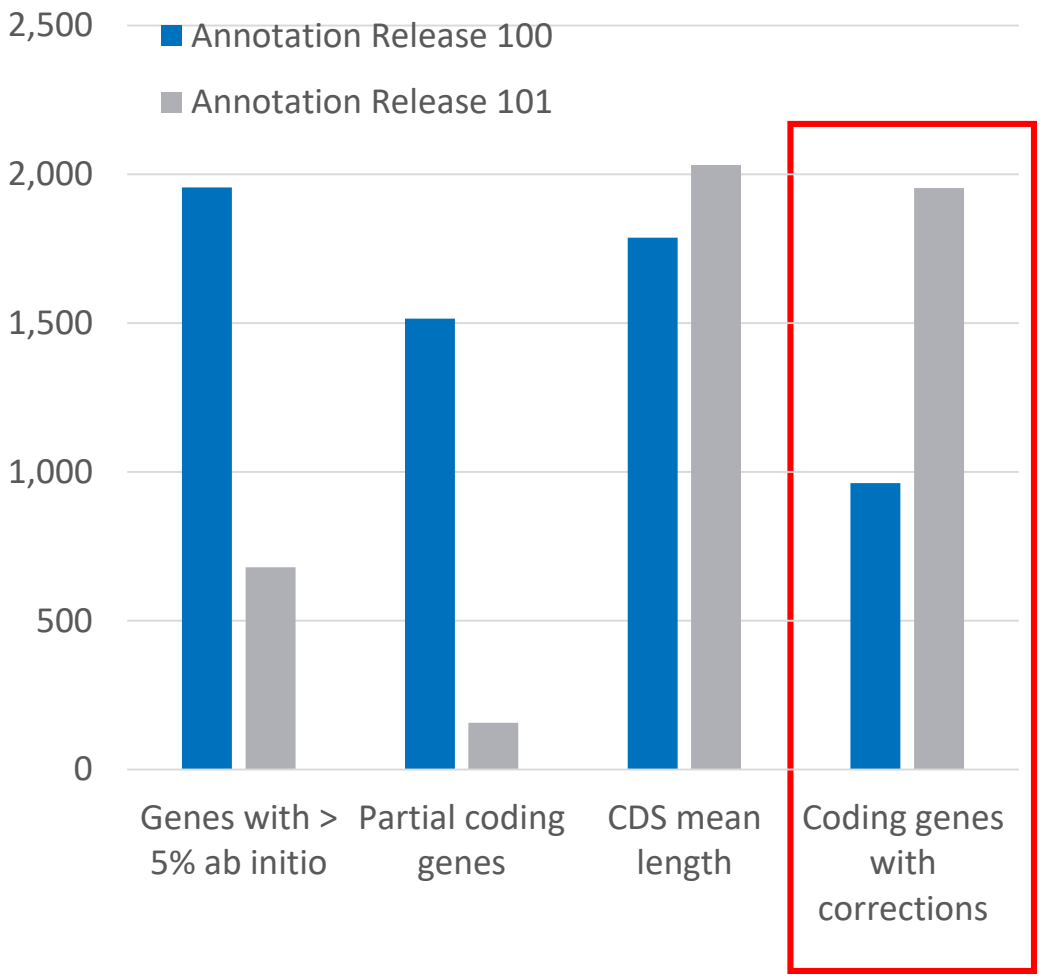
XP_006054308.1

3 genes involving 3 scaffolds

Improved quality metrics

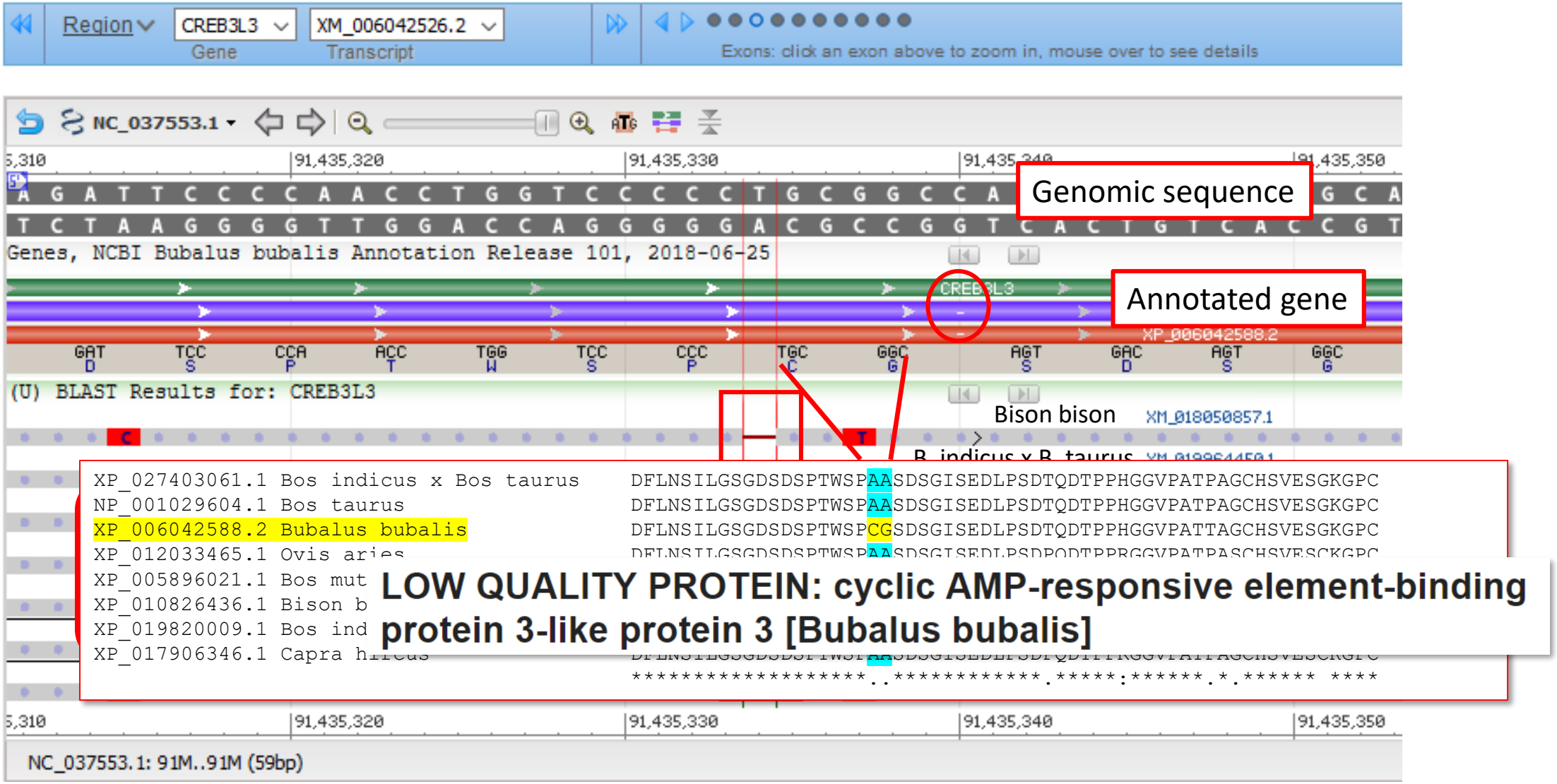
Best SwissProt matches for buffalo coding genes

	Annotation release 101	Annotation release 100
Mean coverage of buffalo coding gene by a SwissProt protein	96	95
Mean coverage of the SwissProt protein by a buffalo coding gene	95	89

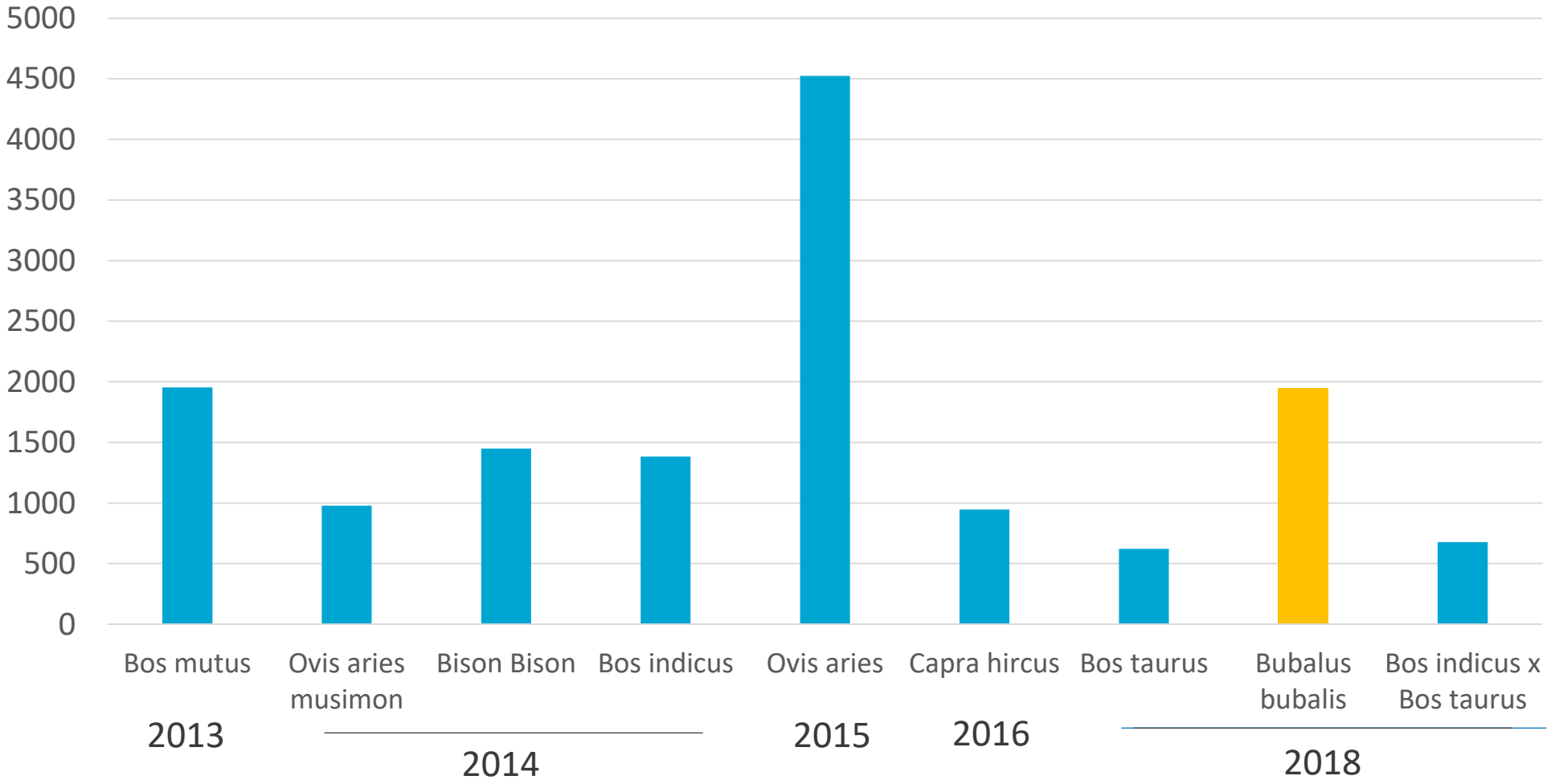


Frequent in PacBio assemblies

Deletion in gene model to compensate for genome error: CREB3L3



Number of corrected protein-coding genes in other bovids



Where is the annotation?

NCBI Resources How To

Assembly

Organism group
Animals (4)
Customize ...

Status clear
Latest (4)
Latest GenBank (4)
Latest RefSeq (1)

Assembly level
Complete genome (0)
Chromosome (1)
Scaffold (3)
Contig (0)

RefSeq category
Reference (0)
Representative (1)

Exclude clear
Exclude partial (0)

GENOME ASSEMBLY

Was this helpful?

UOA_WB_1

Bubalus bubalis (water buffalo)

University of Adelaide (May 2018)

RefSeq assembly: GCF_003121395.1

Genome Browser | BLAST | Download

Assembly statistics

Assembly report

Total sequence length

Assembly level

Contig N50

RefSeq annotation statistic

Annotation Release 101 (June 2018)

Download assembly data

File type

Genomic GFF

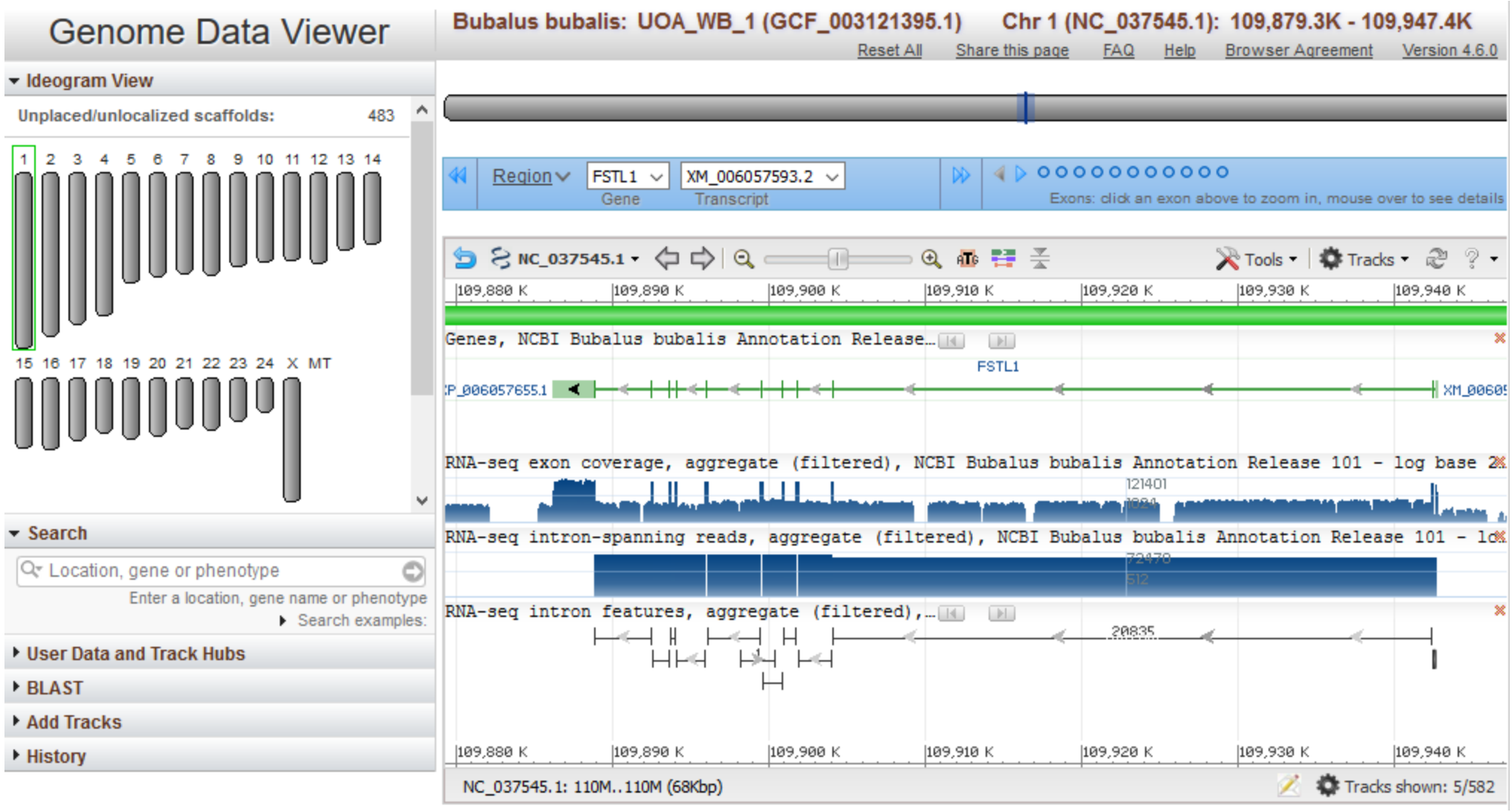
Source database

RefSeq

Estimated size is 0 byte

Download

Water Buffalo in GDV



Water Buffalo in Gene

Gene | Gene | Advanced

Full Report | Send to:

FSTL1 follistatin like 1 [*Bubalus bubalis* (water buffalo)]

Gene ID: 102400106, updated on 27-Jun-2018

Summary

Gene symbol	FSTL1
Gene description	follistatin like 1
Gene type	protein coding
RefSeq status	MODEL
Organism	Bubalus bubalis
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bubalus
Orthologs	human mouse all

Genomic context

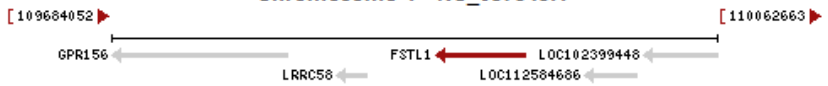
See FSTL1 in Genome Data Viewer

Location: chromosome: 1

Exon count: 11

Annotation release	Status	Assembly	Chr	Location
101	current	ASM312139v1 (GCF_003121395.1)	1	NC_037545.1 (109886222..109942817, complement)
100	previous assembly	UMD_CASPUR_WB_2.0 (GCF_000471725.1)	Unplaced Scaffold	NW_005784826.1 (581670..638211)

Chromosome 1 - NC_037545.1



Bibliography

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

Submit: New GeneRIF Correction

Annotation report

Genome | Genome | Search

Limits Advanced

Transcript alignments

Source	Number of sequences retrieved from Entrez	Number (%) of sequences aligned by Splign	Number (%) of sequences passed to Gnomon	Average % identity	Average % coverage
Same-species known RefSeq (NM_/NR_)	166	166 (100.00%)	160 (96.39%)	99.08%	99.62%
Same-species Genbank					
Same-species EST					

	ASM312139v1 (Current) to UMD_CASPUR_WB_2.0 (Previous)
Identical	3%
Minor changes	47%
Major changes	19%
New	26%
Deprecated	13%
Other	6%
Download the report	tabular, Genome Workbench

RNA-Seq alignments

The following RNA-Seq rea

Hide alignments statist

Sample Id	Pu	ercent of ned reads th introns	Number of introns
All	NA	27%	314,997
SAMEA103886311	NA	21%	178,621
SAMEA103886313	NA	28%	139,059
SAMEA103886315	NA	16%	174,839
SAMEA103886316	NA	13%	162,836

model RefSeq (XM_)	58,026
non-coding RNAs	13,346

More annotated organisms

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/

Eukaryotic genomes annotated at NCBI

Hundreds of eukaryotic genomes have been annotated by the NCBI Eukaryotic Genome Annotation Pipeline (see graphs). The latest annotation release available for each genome is shown in the tables below. The tables are organized by taxonomic group and provide links to the annotation report, FTP site, genome BLAST page, and Genome Data Viewer page.

Only completed annotations are shown here. Please browse the annotation runs currently in progress to see what will become available in a few days.

Show/Hide All

Featured (6)

FTP - FTP Download **B** - Organism-specific BLAST **AR** - Annotation Report **GDV** - Genome Data Viewer

Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links			
Homo sapiens (human)	GRCh38.p12 (GCF_000001405.38)	109	2018-02-13	2018-03-26	FTP	B	AR	GDV
Mus musculus (house mouse)	GRCm38.p4 (GCF_000001635.24)	106	2016-06-09	2016-06-22	FTP	B	AR	GDV
Rattus norvegicus (Norway rat)	Rnor_6.0 (GCF_000001895.5) Rn_Celera (GCF_000002265.2)	106	2016-07-07	2016-07-27	FTP	B	AR	GDV
Danio rerio (zebrafish)	GRCz11 (GCF_000002035.6)	106	2017-06-02	2017-06-26	FTP	B	AR	GDV
Apis mellifera (honey bee)	Amel_HAv3.1 (GCF_003254395.2)	104	2018-09-13	2018-09-19	FTP	B	AR	GDV
Zea mays (maize)	B73 RefGen_v4 (GCF_000005005.2)	102	2017-12-07	2017-12-18	FTP	B	AR	GDV

Primates (26)

Rodents (21)

Thank you.

– This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

GenBank

Shelby Bidwell
Larissa Brown
Jianli Dai
Scott Durkin
Michel Eschenbrenner
Linda Frisse
Leigh Riley

BioProject / Biosample

John Anderson
Carol Scott

GEO

Emily Clough
Carlos Evangelista
Irene Kim
Pierre Ledoux
Hyeseung Lee
Kimberly Marshall
Katherine Phillippy
Patti Sherman
Stephen Wilhite
Tanya Barrett

GEO developers

Alexandra Soboleva
Maxim Tomashevsky
Nadezhda Serova
Naigong Zhang

Annotation Pipeline

Françoise Thibaud-Nissen
Paul Kitts
Mike Dicuccio
Wratko Hlavina
Avi Kimchi

Jinna Choi
Patrick Masterson
Eyal Mozes
Robert Smith
Alexandre Souvorov

RefSeq/Gene

Eric Cox
Catherine Farrell
Tamara Goldfarb
Diana Haddad
John Jackson
Vinita Joardar
Kelly McGarvey
Michael Murphy
Nuala O'Leary

RefSeq Developers

Alex Astashyn
Olga Ermolaeva
Vamsi Kodali
Craig Wallin

Shashi Pujar
Bhanu Rajput
Sanjida Rangwala
Lillian Riddick
Barbara Robbertse
Brian Smith-White
Pooja Strope
Anjana Vatsan
David Webb

GDV/Remap/GBench

Valerie Schneider
Peter Meric
Nathan Bouk
Hsiu-Chuan Chen
Cliff Clausen
Anatoliy Kuznetsov

A cast of thousands

Ken Katz
Michael Ovetsky
Lukas Wagner
Andrei Shkeda
Donna Maglott
Kim Pruitt
Jim Ostell

Watch NCBI News for updates!

http://www.ncbi.nlm.nih.gov/news/
https://www.youtube.com/user/NCBINLM

NCBI Genome Resources Workshop
Monday Jan 14 12:50 – 3:00 Pacific Salon 2

Time	Topic
12:50 – 1:10	Submission of Genomes to GenBank *Karen Clark*
1:10 – 1:30	GEO Submissions and Usage *Steve Wilhite*
1:30 – 1:55	From Annotation to Visualization: Exploring Genes and Genomes with NCBI Tools *Eric Cox*
1:55 – 2:15	Programmatic Access to Genomic Data: E-Utilities and FTP *Vamsi K. Kodali*
2:15 – 2:35	NCBI Resources for Phyletically-Defined Next Generation Analysis in and out of the Cloud (a.k.a. Cool New Stuff!) *Ben Busby*
2:35 – 3:00	Q & A session