From Annotation to Visualization: Exploring Genes and Genomes with NCBI Tools

Eric Cox (eric.cox@nih.gov)

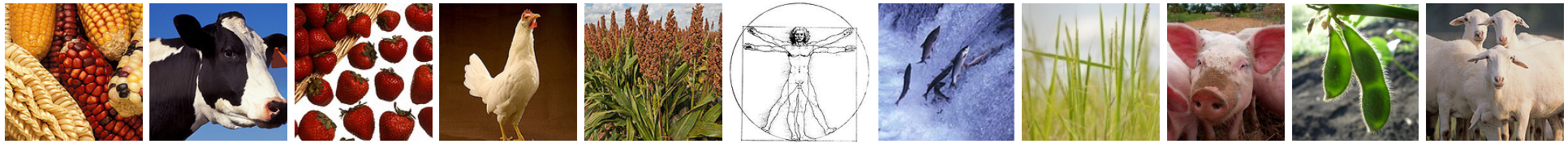
NCBI Genome Resources Workshop

PAG XXVII January 14, 2019

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

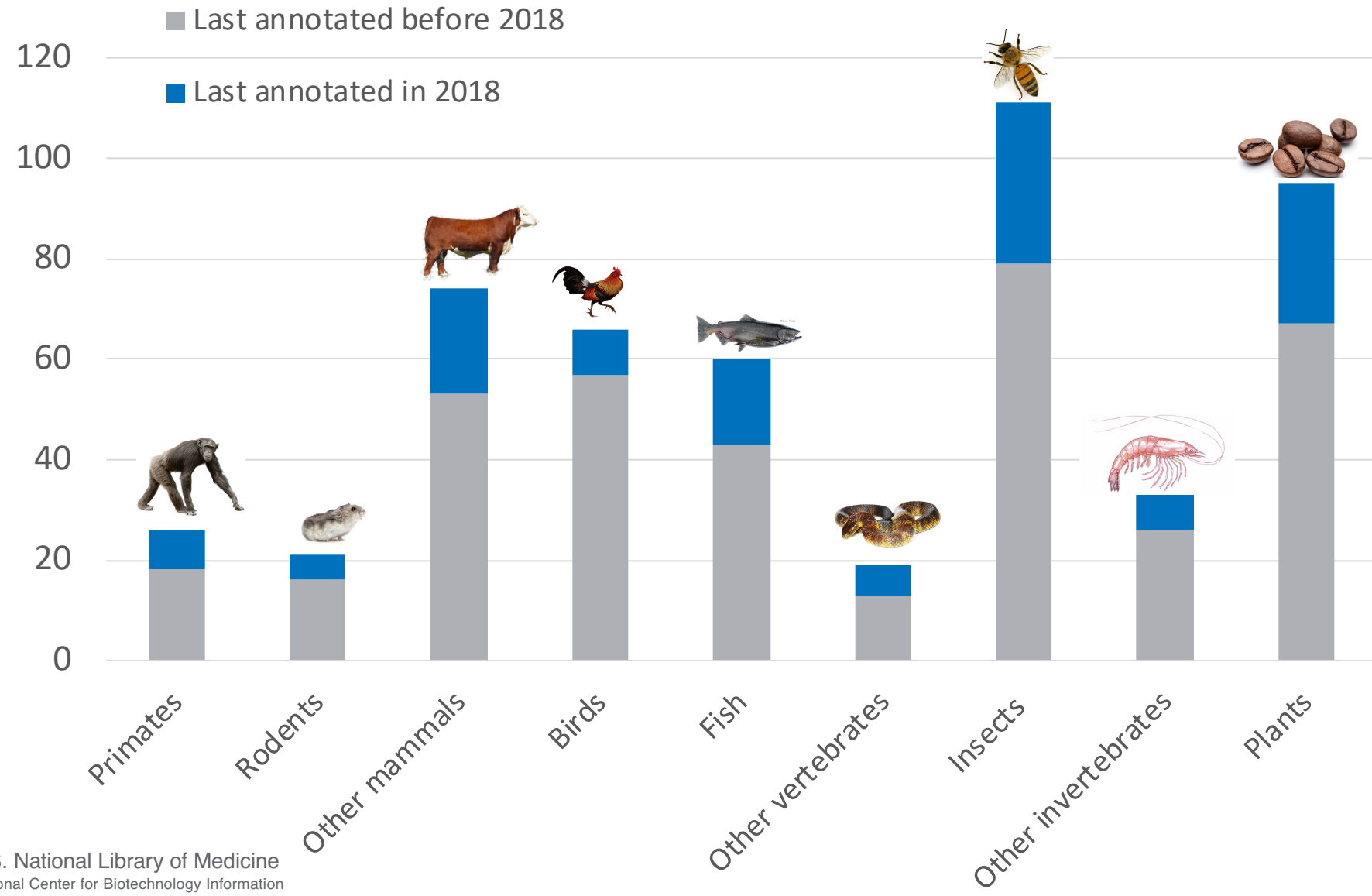
From Annotation to Visualization

How we annotate genomes—including complex assemblies

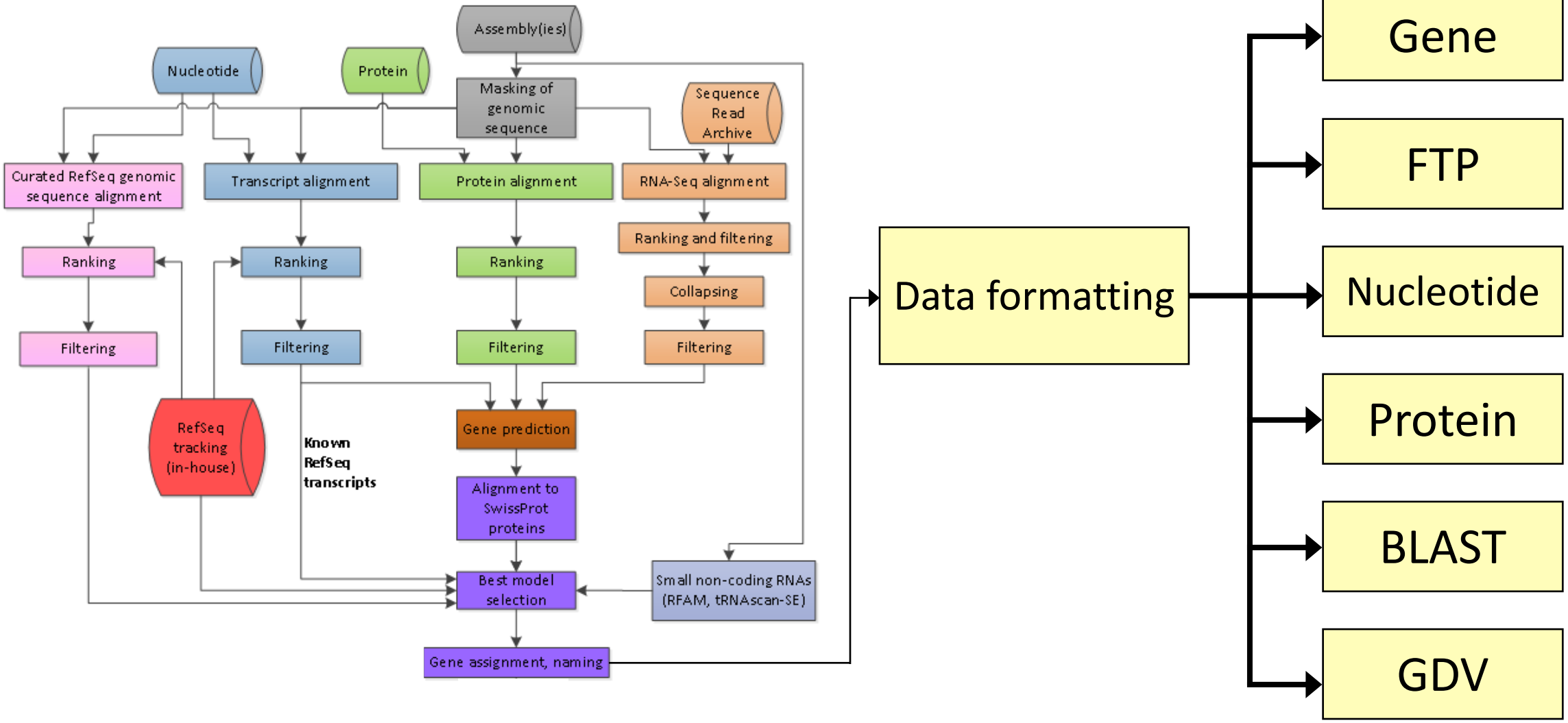


How to visualize annotation data: Gene and Genome Data Viewer

Over 500 species annotated



Eukaryotic Genome Annotation Pipeline



2018 Annotations for Agriculture

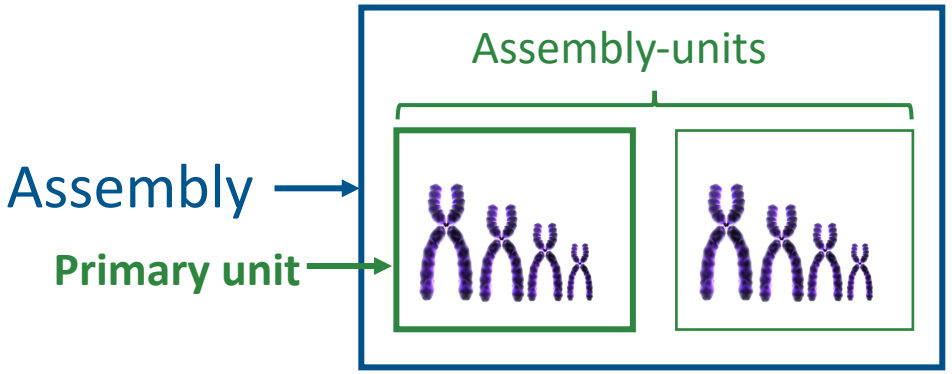
Organism	Assembly	Submitted	Annotation available	RNA-Seq samples	RNA-Seq read count
Penaeus vannamei (Pacific white shrimp)	ASM378908v1	11/16/2018	12/14/2018	133	6.45B
Coffea arabica (coffee)	Cara_1.0	11/08/2018	12/4/2018	66	2.15B
Oreochromis niloticus (Nile tilapia)	O_niloticus_UMD_NMBU	6/29/2018	7/26/2018	138	5.11B
Bubalus bubalis (water buffalo)	UOA_WB_1	5/14/2018	6/25/2018	274	15.57B
Bos taurus (cattle)	ARS-UCD1.2	4/11/2018	5/11/2018	385	12.95B
Gallus gallus (chicken)	GRCg6a	3/27/2018	5/17/2018	129	10.56B
Equus caballus (horse)	EquCab3.0	1/05/2018	1/26/2018	182	10.82B

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

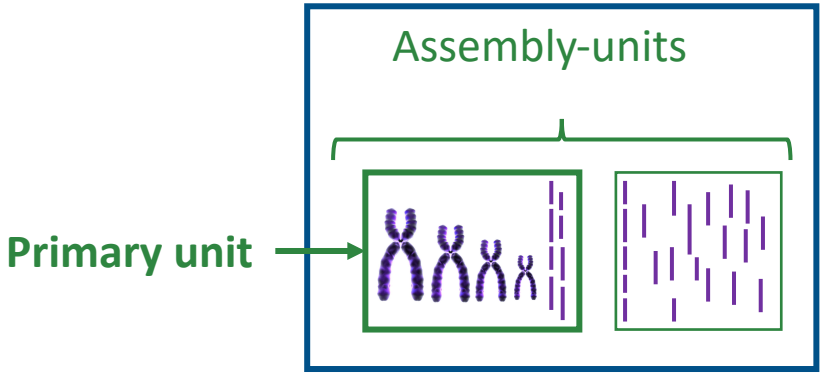
New data model for complex assemblies

Old model

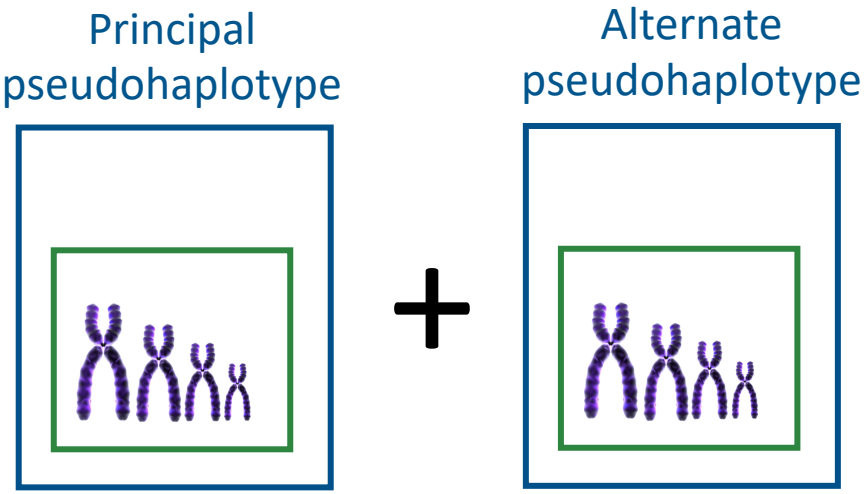
- Diploid assembly

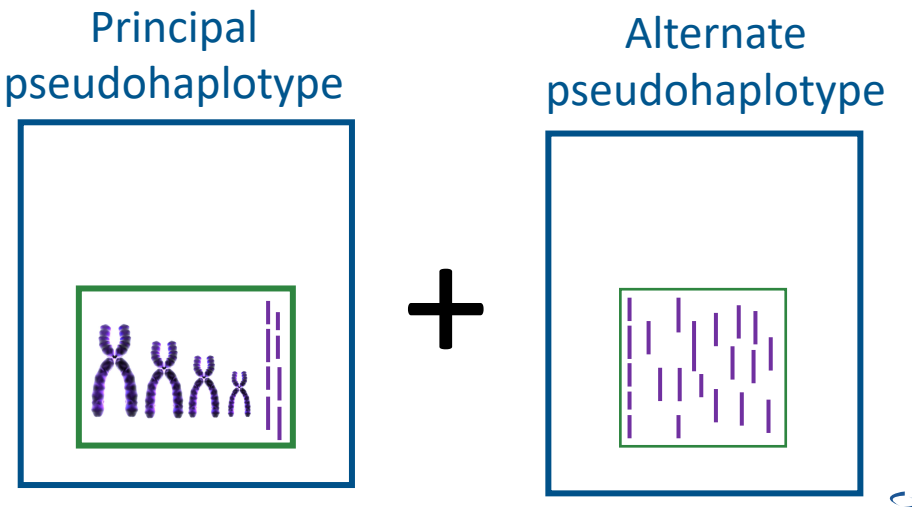


- Haploid with alt loci



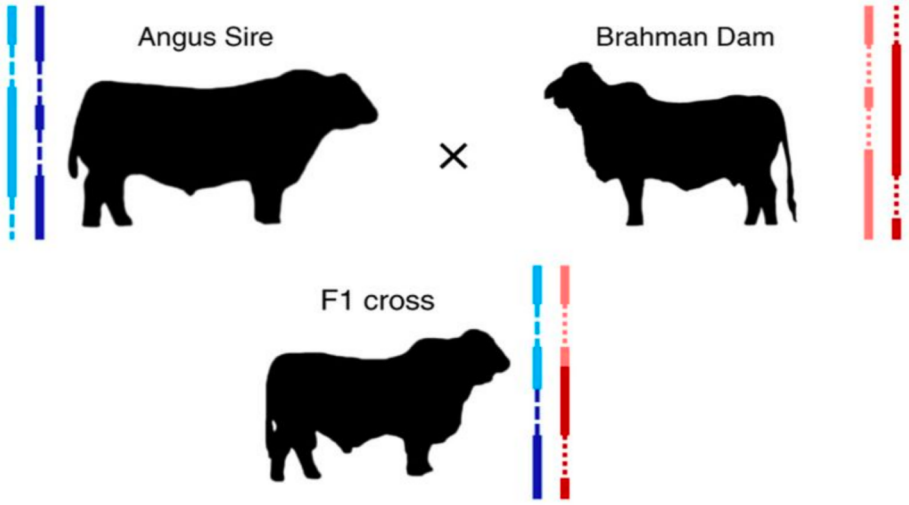
New model: two assemblies



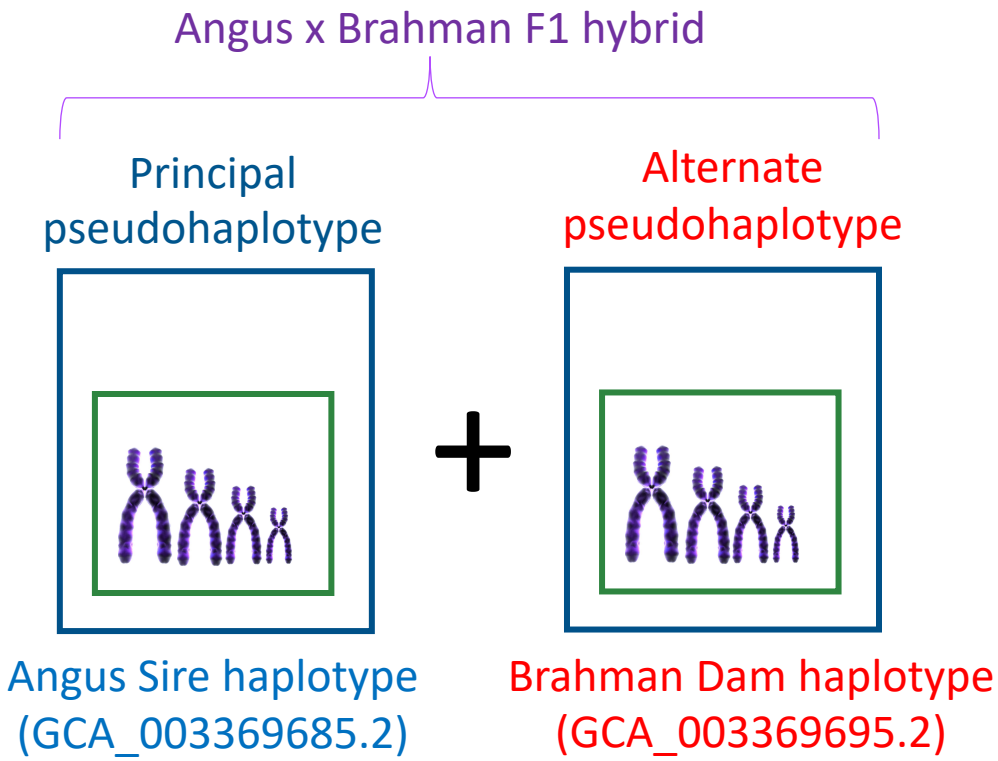


Handout available

New Data Model for Diploid Assemblies



Koren et al.
Nature Biotechnol. 2018



New Data Model for Diploid Assemblies

Angus Sire

Principal pseudohaplotype

UOA_Angus_1

Organism name: Bos indicus x Bos taurus (hybrid cattle)

Infraspecific name: Breed: Angus x Brahman F1 hybrid

Sex: male

Linked assembly: GCA_003369695.2 (alternate pseudohaplotype of diploid)

Brahman Dam

Alternate pseudohaplotype

UOA_Brahman_1

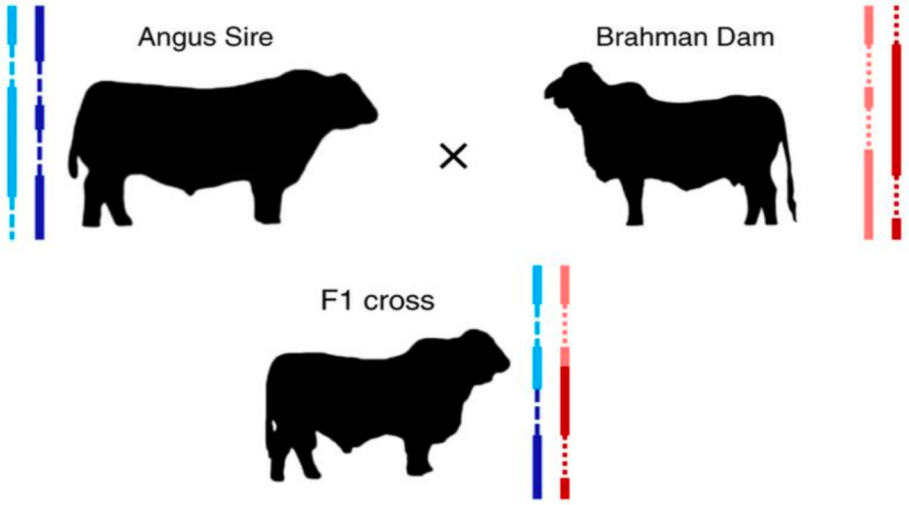
Organism name: Bos indicus x Bos taurus (hybrid cattle)

Infraspecific name: Breed: Angus x Brahman F1 hybrid

Sex: male

Linked assembly: GCA_003369685.2 (principal pseudohaplotype of diploid)

Annotating a diploid assembly with the new data model

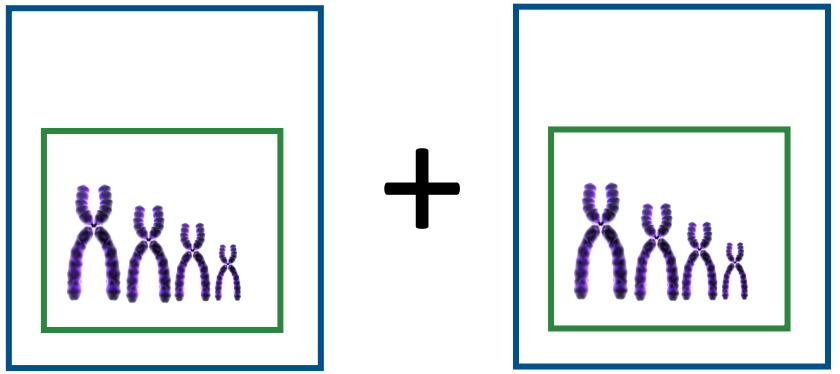


Koren et al.
Nature Biotechnol. 2018

Angus x Brahman F1 hybrid

Principal pseudohaplotype

Alternate pseudohaplotype



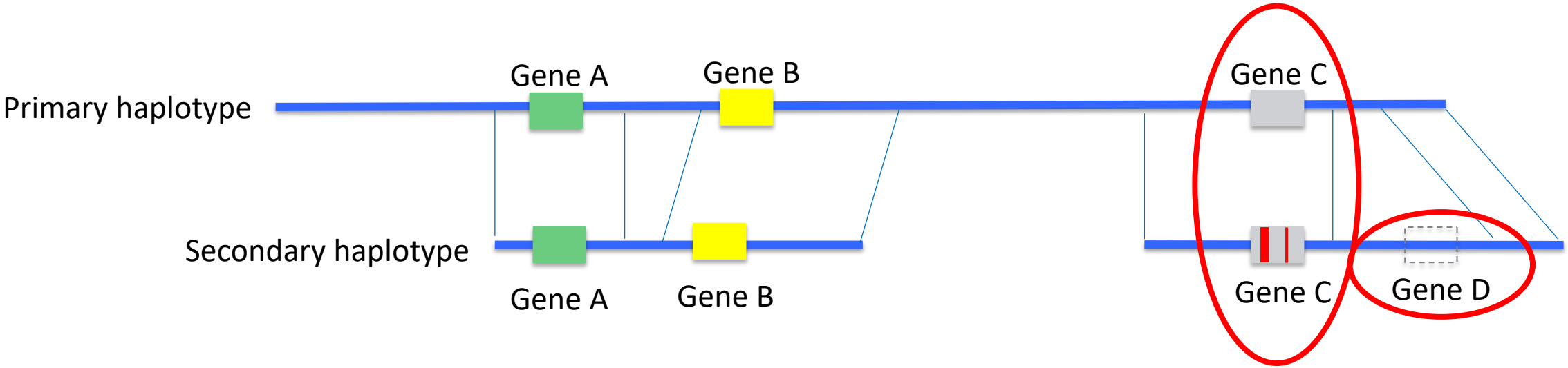
Angus Sire haplotype
(GCA_003369685.2)
not annotated

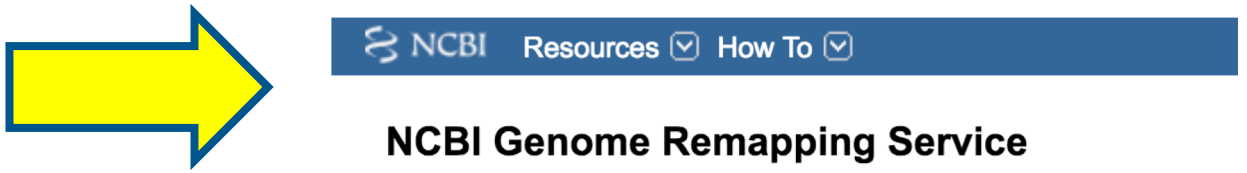
Brahman Dam haplotype
(GCA_003369695.2)
annotated

GCF_003369695.1
RefSeq annotated assembly

- We only annotate one haplotype

Getting Annotation Data for a Secondary Haplotype

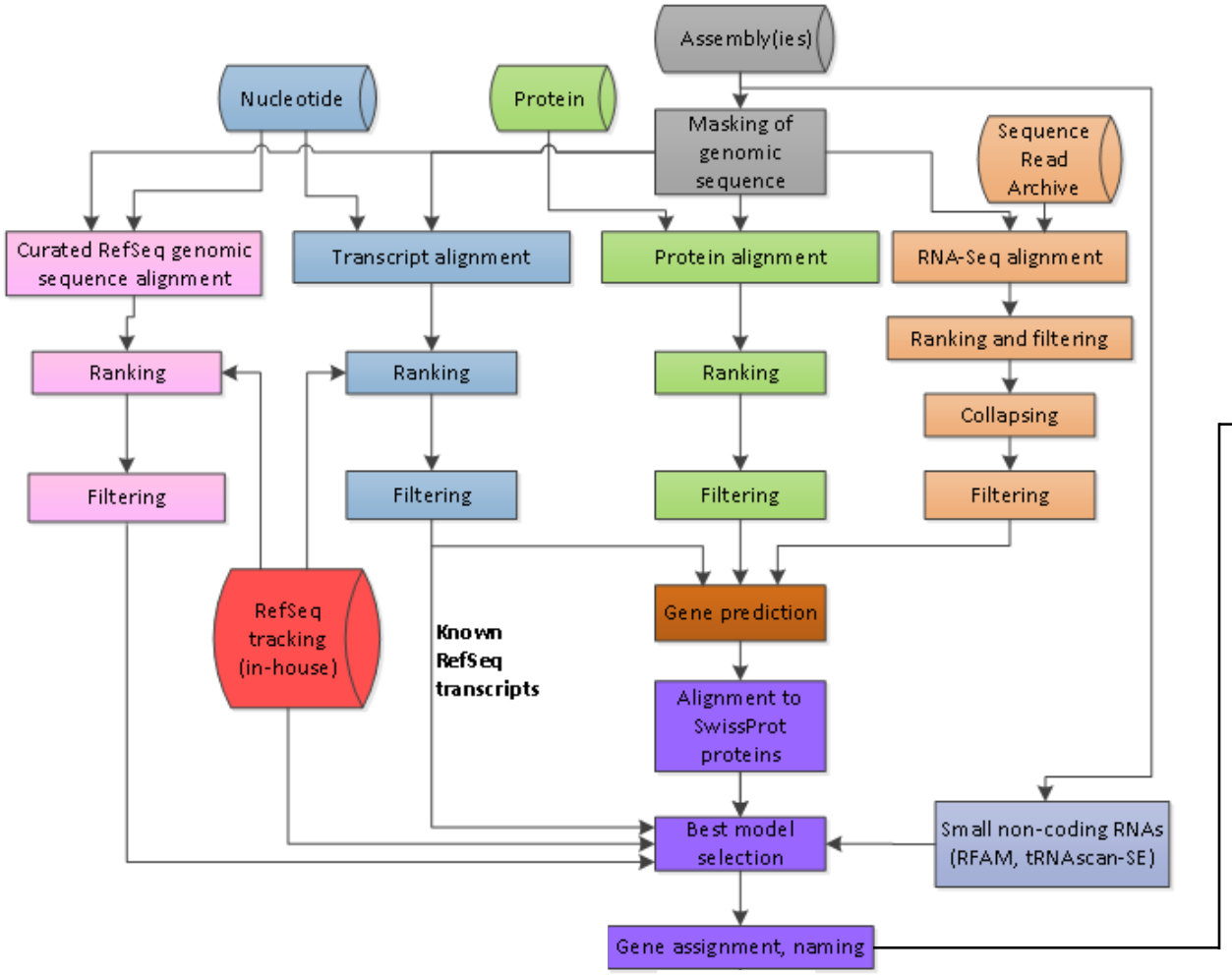




https://www.ncbi.nlm.nih.gov/genome/tools/remap

Handout available

The Products of Annotation

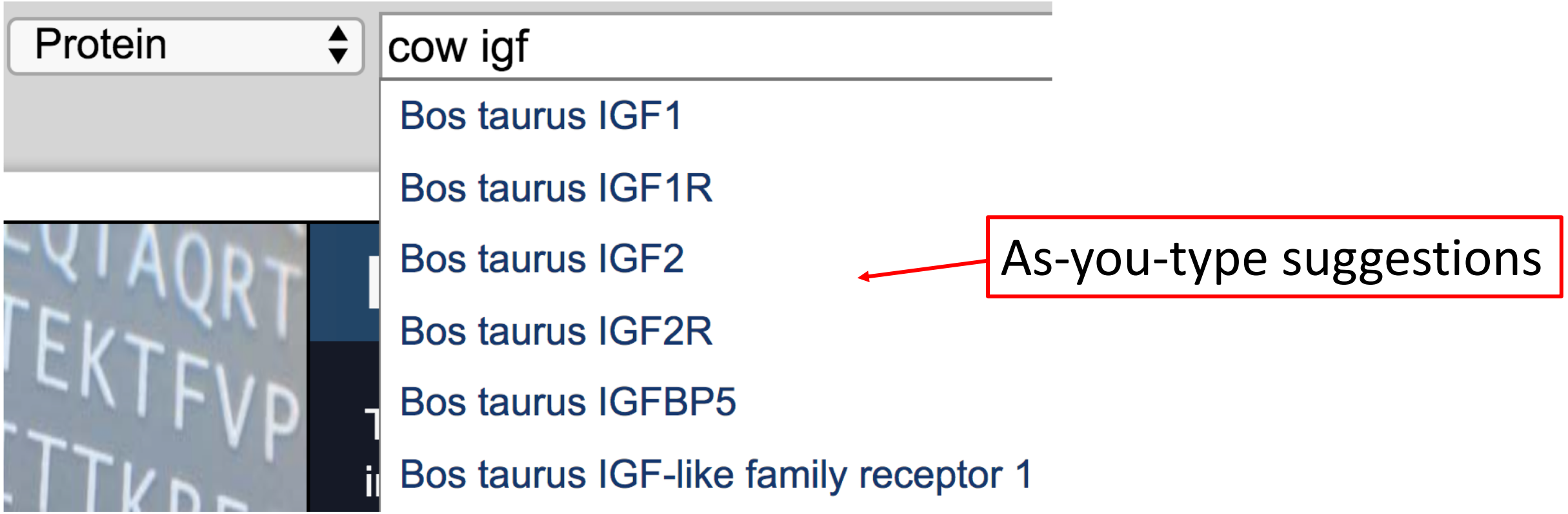


Data formatting

- Gene
- FTP
- Nucleotide
- Protein
- BLAST
- GDV

How do I find my gene?

Search NCBI for organism + gene symbol or name



Searching for annotation products

cow igf2 ✖ Search

GENE

Was this helpful?

IGF2 – insulin like growth factor 2

Bos taurus (cattle)

Processed peptides: Insulin-like growth factor II, Preptin

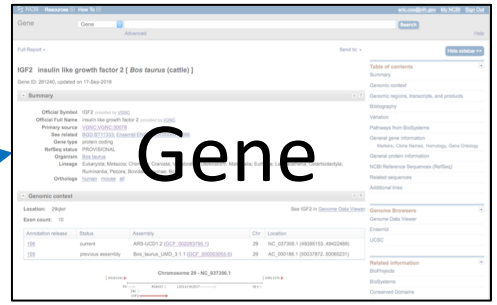
GeneID: 281240

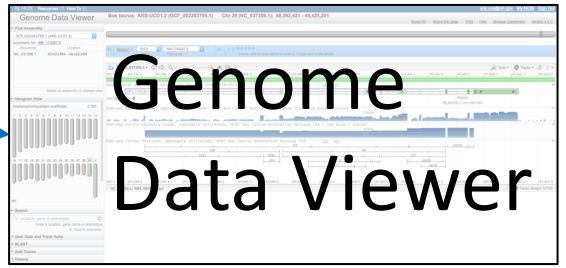
RefSeq transcripts (9) RefSeq proteins (9) PubMed (42)

Genome Browser BLAST Download

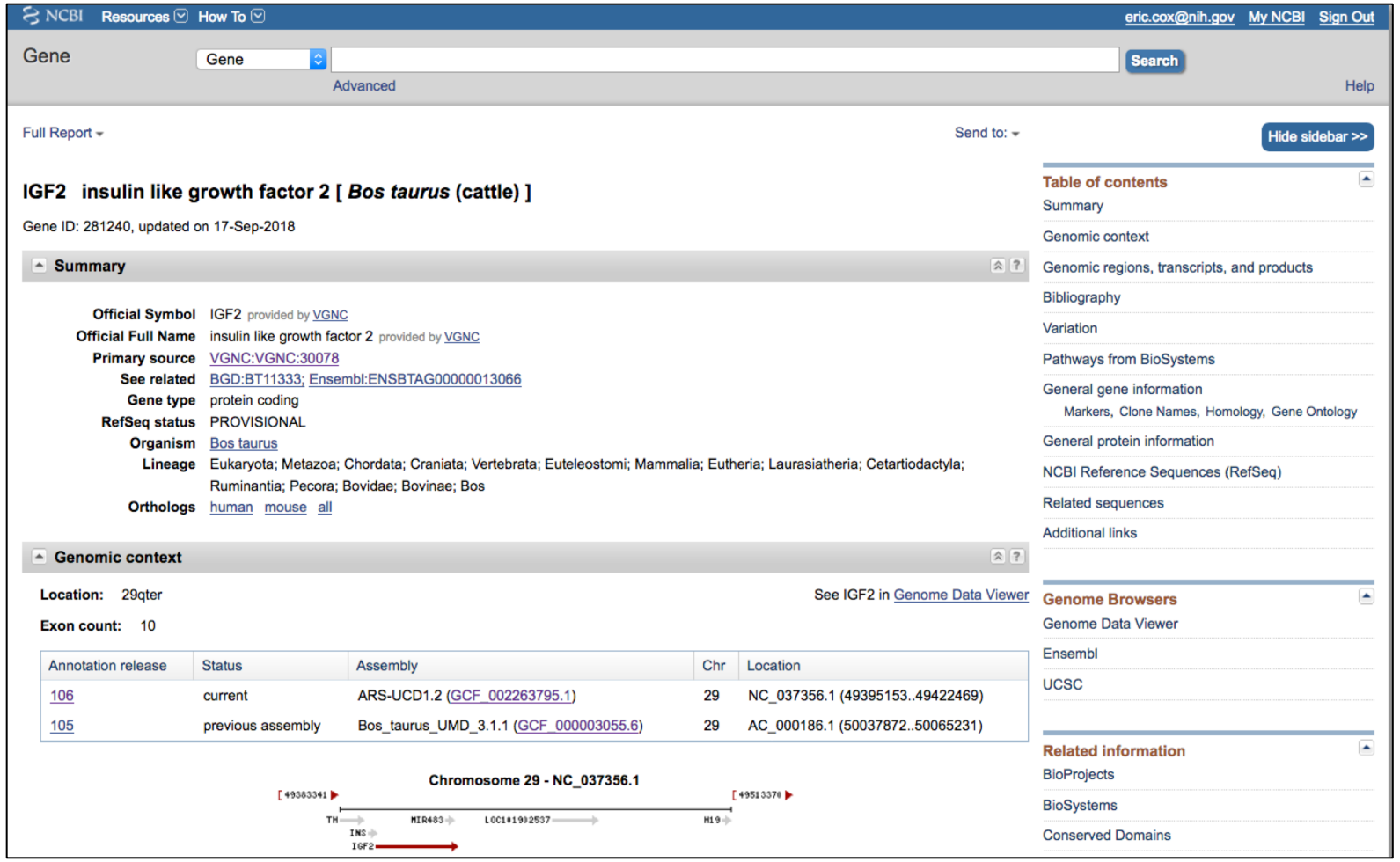
RefSeq transcripts +

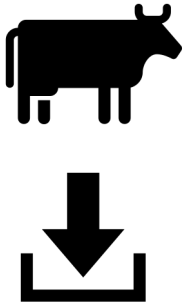
RefSeq proteins +





Gene: portal to NCBI annotation data





Handout available

Nomenclature and Orthology Data in Gene

Full Report

IGF2 insulin like growth factor 2 [*Bos*

Gene ID: 281240, updated on 17-Sep-2018

Summary

Official Symbol	IGF2 provided by VGNC
Official Full Name	insulin like growth factor 2
Primary source	VGNC:VGNC:30078
See related	BGD:BT11333; Ensembl:EN
Gene type	protein coding
RefSeq status	PROVISIONAL
Organism	Bos taurus
Lineage	Eukaryota; Metazoa; Chorda
	Ruminantia; Pecora; Bovida
Orthologs	human mouse all

Send to:

Hide sidebar >>

ducts

Gene Ontology

Igf2 insulin-like growth factor 2 [*Mus musculus* (house mouse)]

Gene ID: 16002, updated on 18-Nov-2018

Summary

Official Symbol	Igf2 provided by MGI
Official Full Name	insulin-like growth factor 2 provided by MGI
Primary source	MGI:MGI:96434
See related	Ensembl:ENSMUSG00000048583 Vega:OTTMUSG00000035324
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Mus musculus
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus
Also known as	Mpr; M6pr; Peg2; Igf-2; Igf-II; AL033362
Summary	This gene encodes a member of the insulin-like growth factor (IGF) family of proteins that promote growth and development during fetal and postnatal life. It is an imprinted gene that is expressed only from the paternal allele. The encoded protein undergoes proteolytic processing to generate a mature peptide. The transgenic overexpression of this gene in mice results in prenatal overgrowth, polyhydramnios, fetal and neonatal lethality, disproportionate organ overgrowth including tongue enlargement, and skeletal abnormalities. Mice lacking the encoded protein exhibit growth deficiency. Alternative splicing results in multiple transcript variants encoding different isoforms that may undergo similar processing to generate mature protein. [provided by RefSeq, Oct 2015]
Expression	Biased expression in limb E14.5 (RPKM 601.7), liver E18 (RPKM 468.8) and 7 other tissues See more
Orthologs	human all

Using orthology to get clues to function

Genomic Context Data in Gene

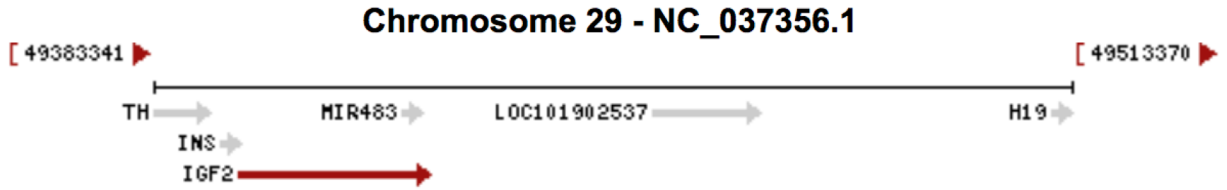
Genomic context

See IGF2 in Genome Data Viewer

Location: 29qter

Exon count: 10

Annotation release	Status	Assembly	Chr	Location
106	current	ARS-UCD1.2 (GCF_002263795.1)	29	NC_037356.1 (49395153..49422469)
105	previous assembly	Bos_taurus_UMD_3.1.1 (GCF_000003055.6)	29	AC_000186.1 (50037872..50065231)



What is Gene? **Bibliography**

Bibliography

Related articles in PubMed

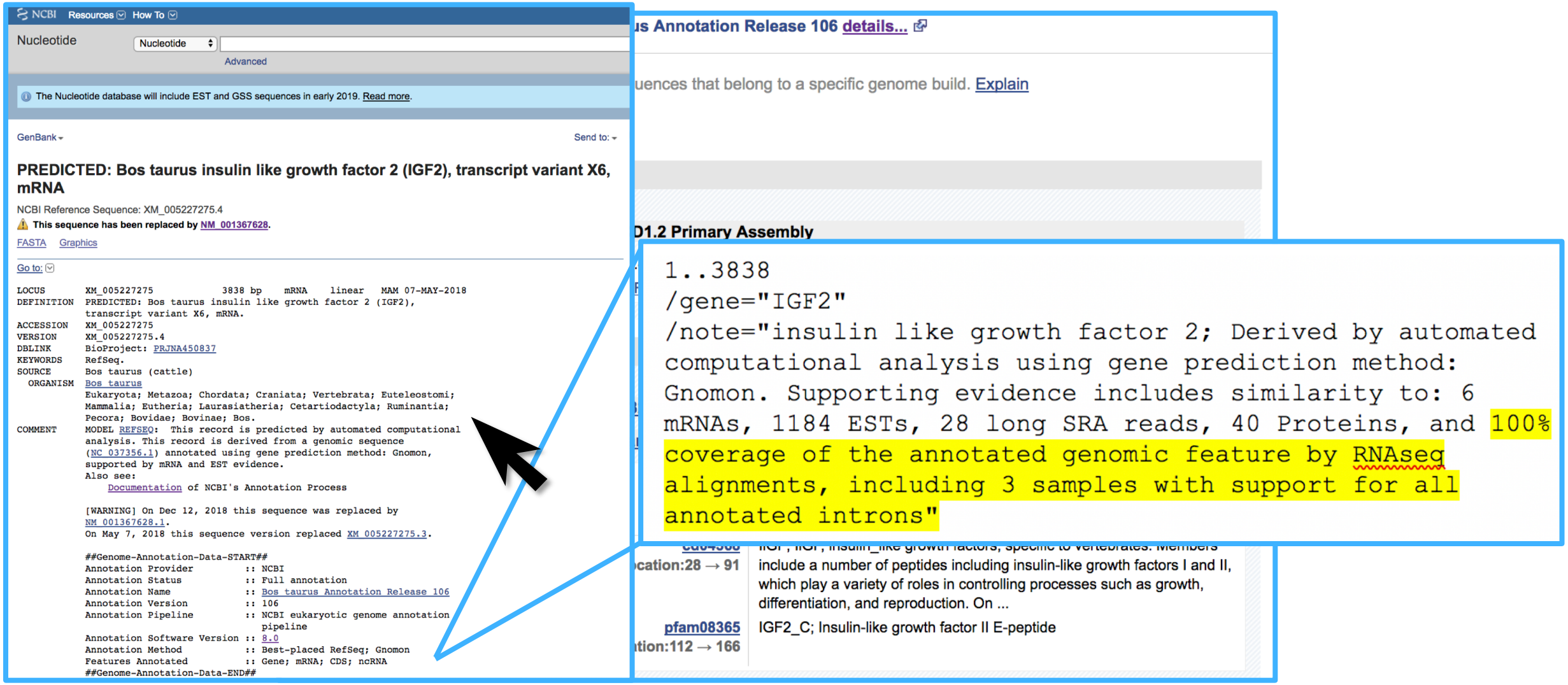
1. SNP and haplotype analysis reveal IGF2 variants associated with growth traits in Chinese Qinchuan cattle.
   Huang YZ, *et al*. Mol Biol Rep, 2014 Feb. PMID 24374893
2. Characterisation of the methylation pattern in the intragenic CpG island of the IGF2 gene in Bos taurus indicus cumulus cells during in vitro maturation.
   Franco MM, *et al*. J Assist Reprod Genet, 2014 Jan. PMID 24174298, **Free PMC Article**

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

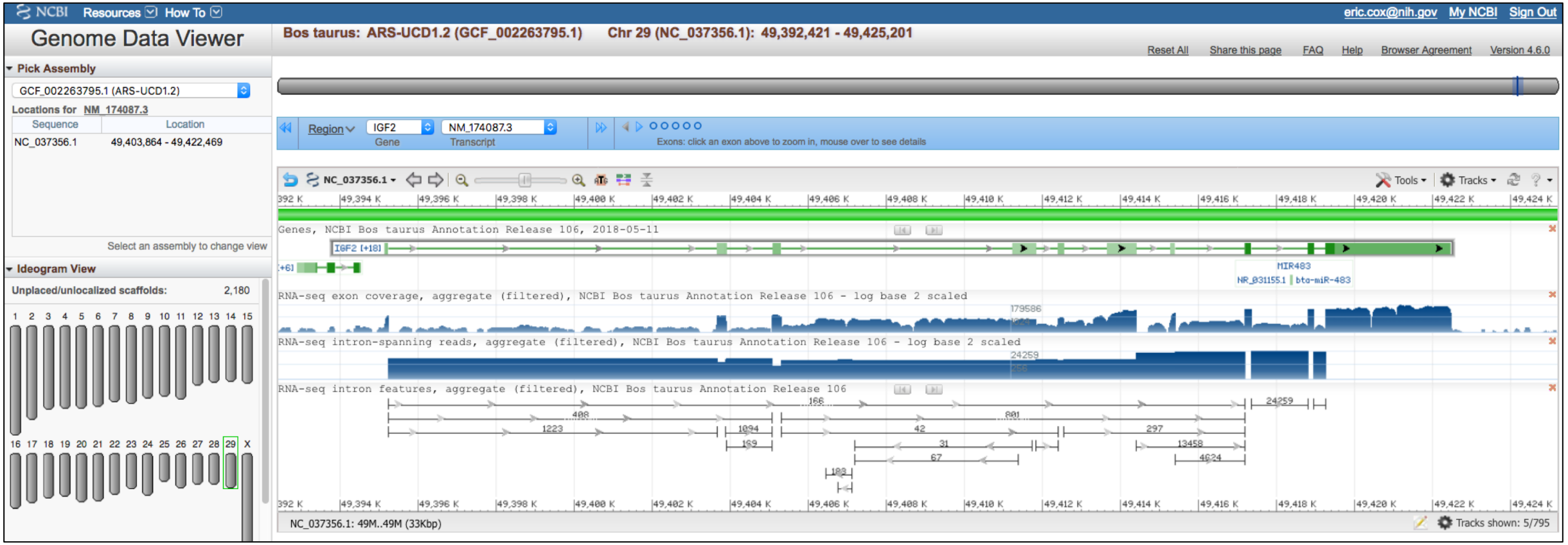
1. Methylation pattern in a CpG island of the IGF2 gene in cumulus cells from 1-3 mm and >/= 8.0 mm follicles and the effects of in vitro maturation on this pattern.
2. Our results provide evidence that polymorphisms in the IGF2 gene are associated with growth traits, and may be used for marker-assisted selection in beef cattle breeding program

Submit: New GeneRIF Correction See all GeneRIFs (21)

What is Gene? **NCBI Reference Sequences (RefSeq)**



NCBI's Genome Browser: **Genome Data Viewer**



Handout available

NCBI's Genome Browser:
Genome Data Viewer

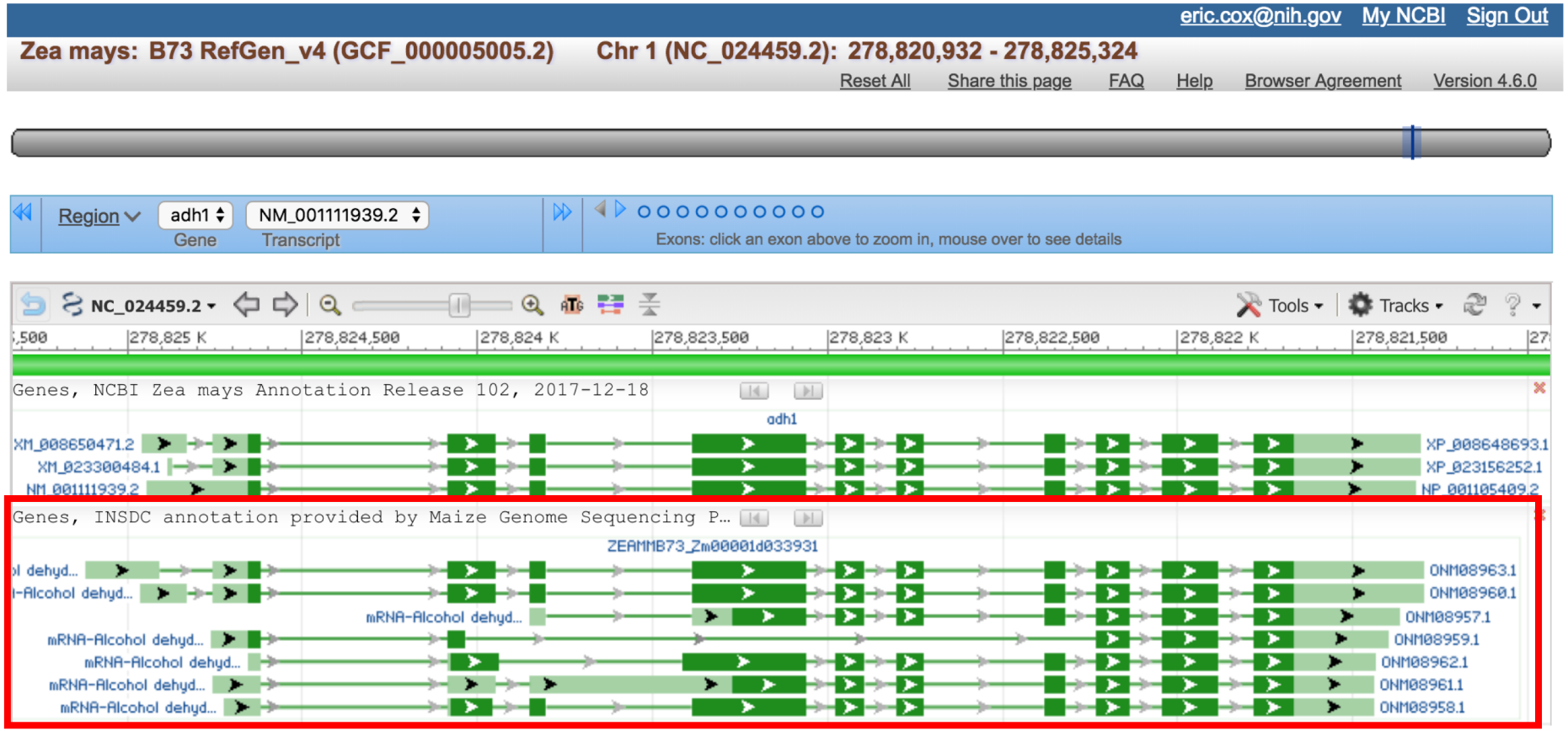
NCBI Annotation

Ensembl Annotation

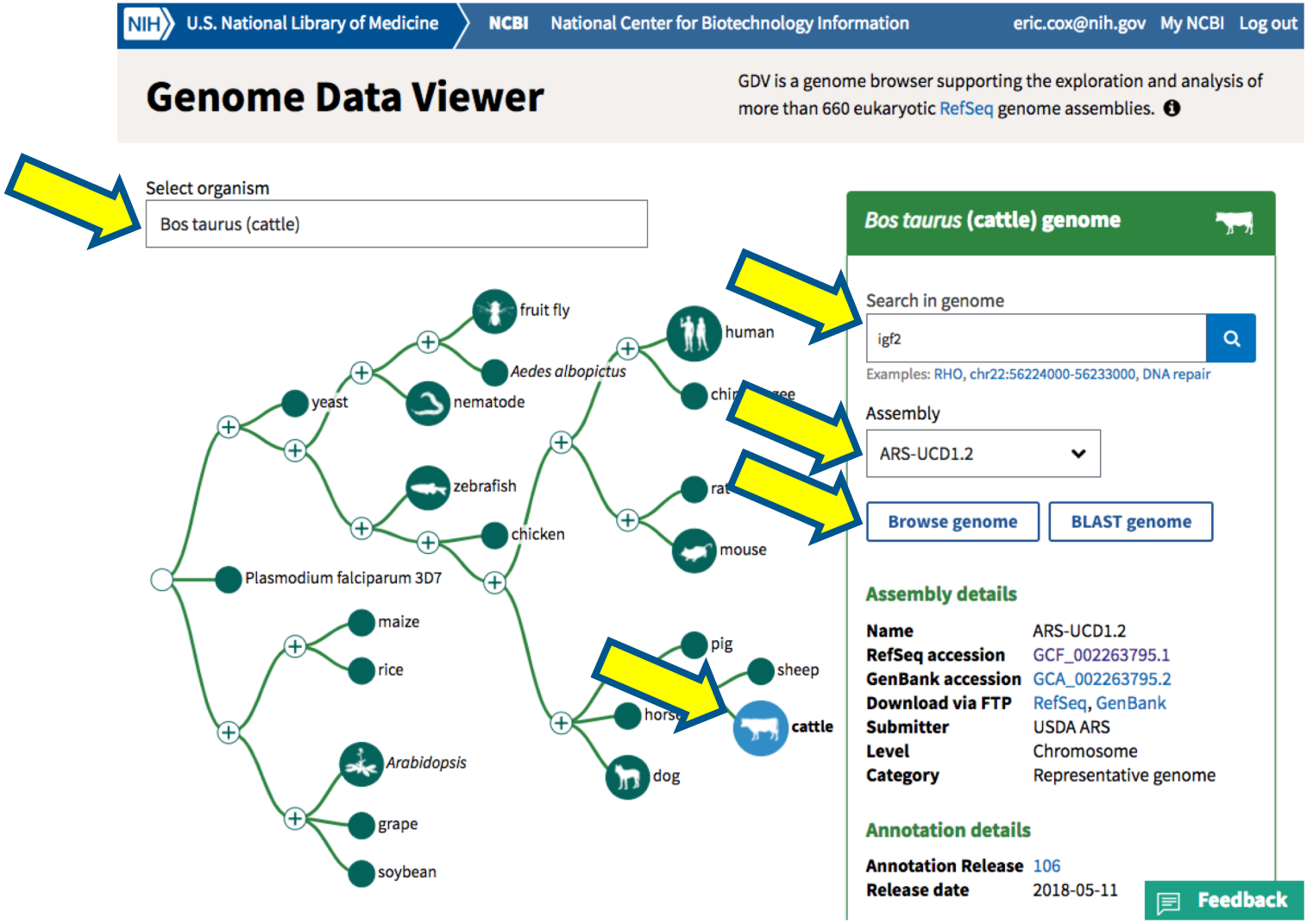
NCBI's Genome Browser:
Genome Data Viewer

NCBI Annotation

User-submitted Annotation

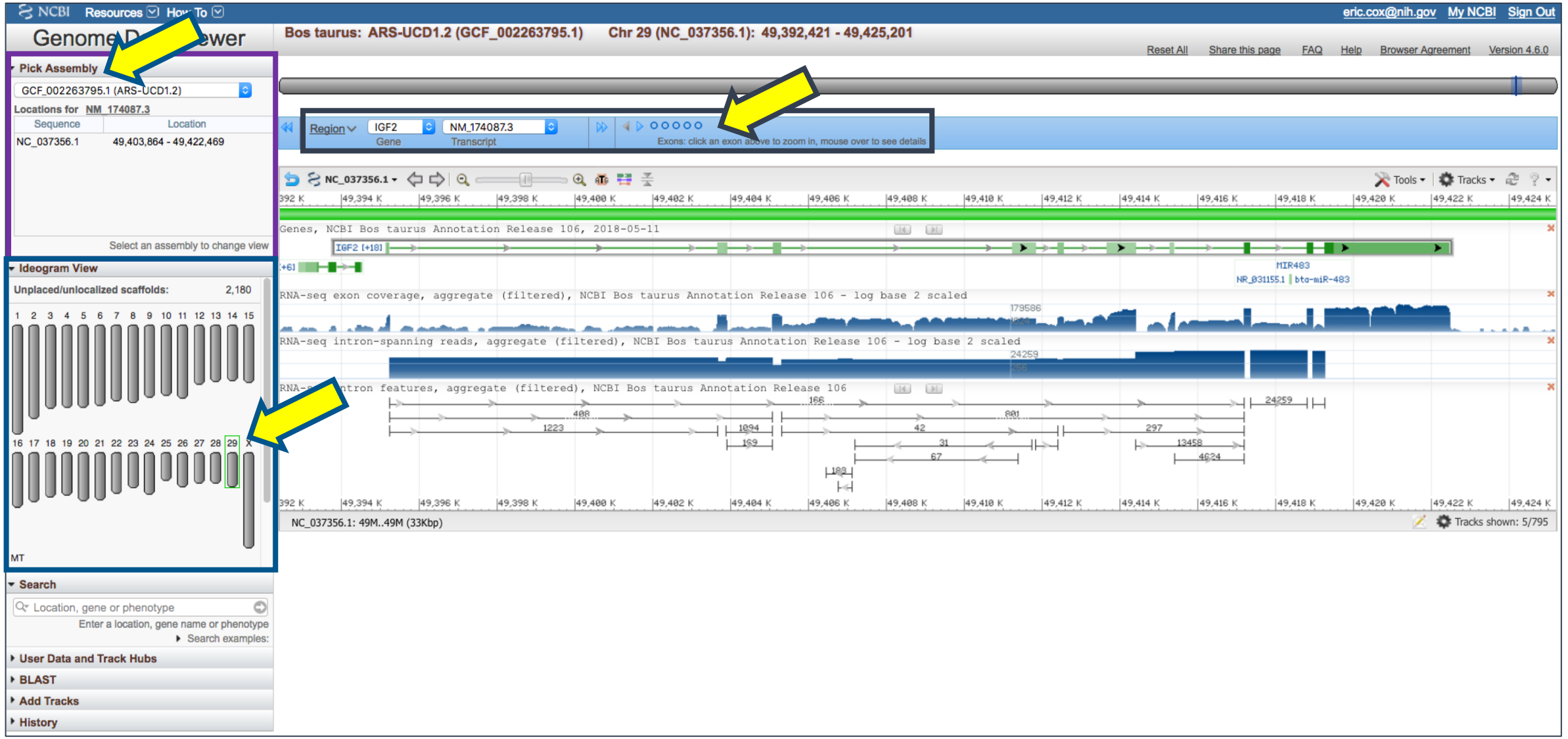


Genome Data Viewer **Homepage**

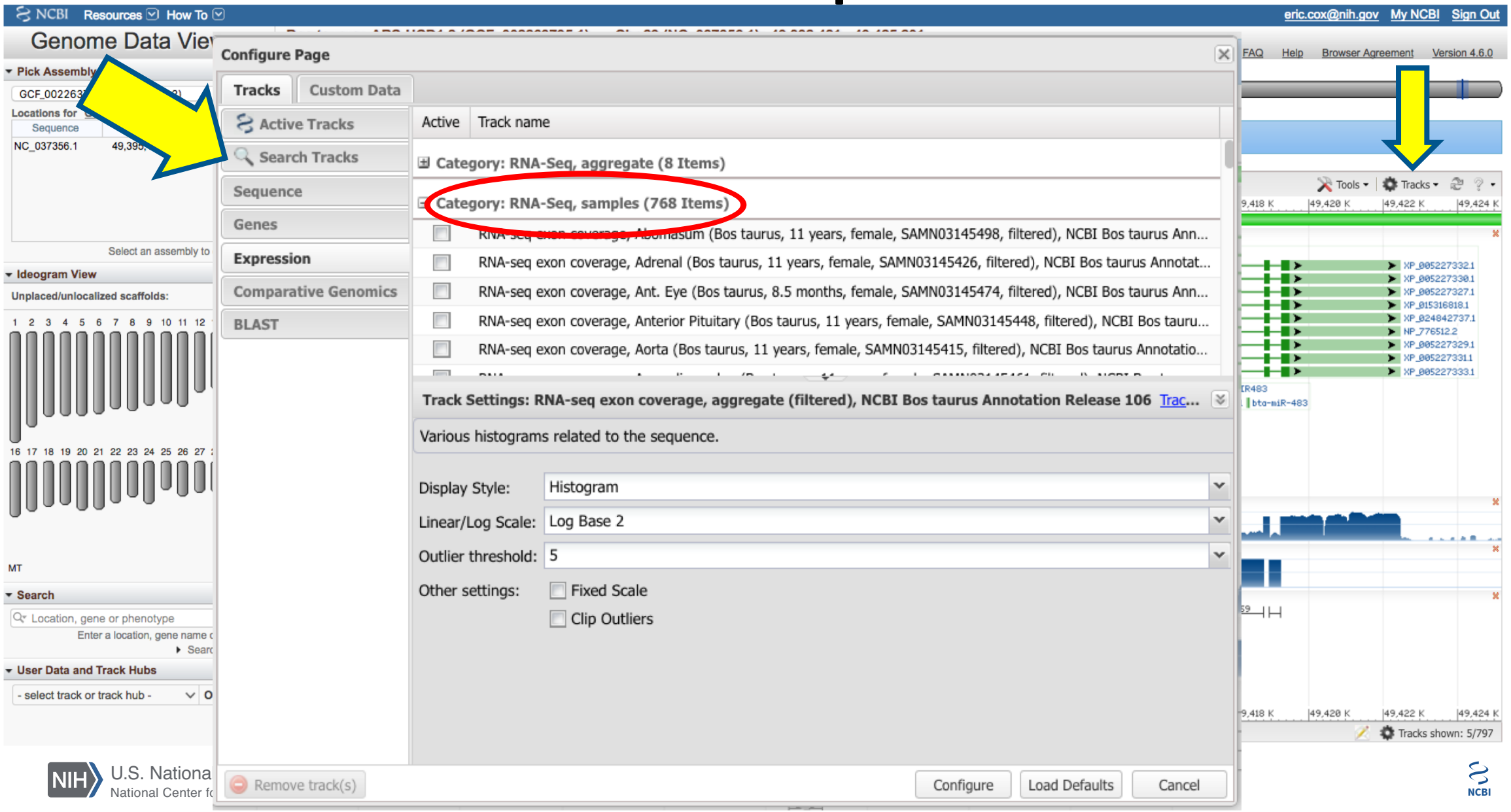


https://www.ncbi.nlm.nih.gov/genome/gdv/

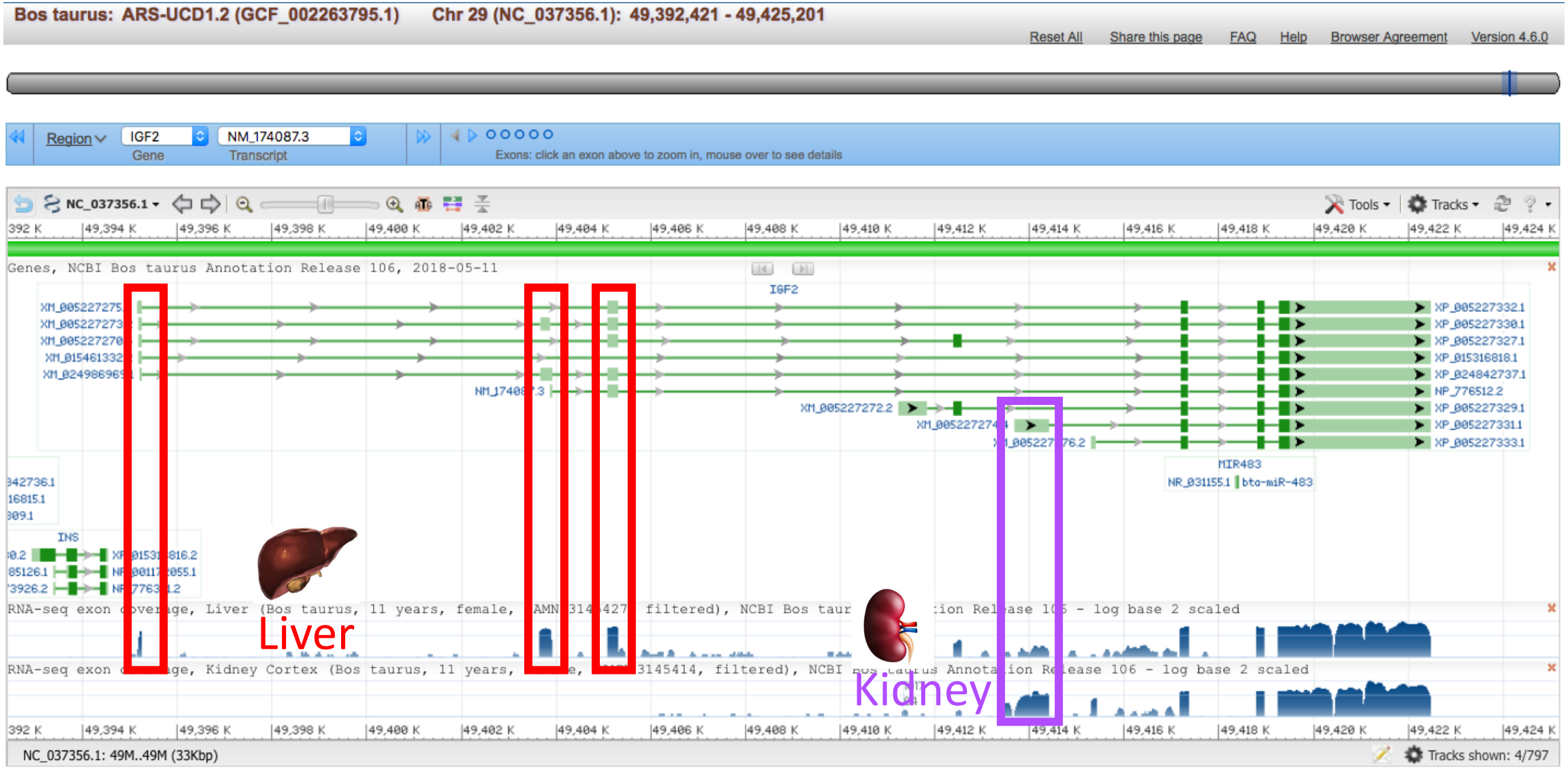
Genome Data Viewer: **Control the Display with Widgets**



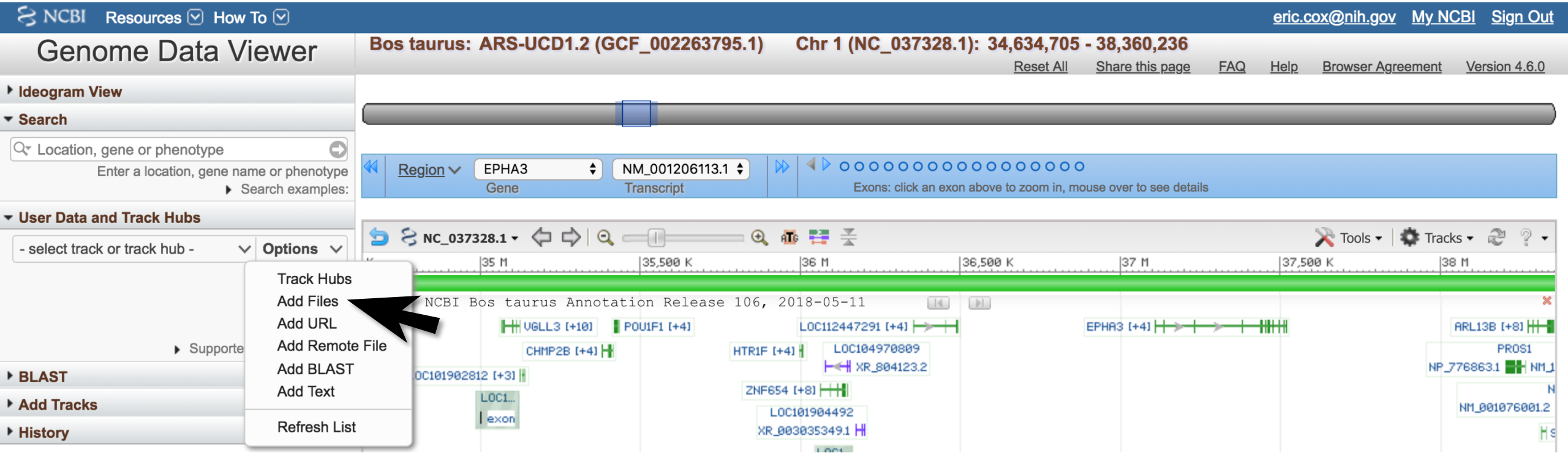
Genome Data Viewer: **Expression Tracks**



Genome Data Viewer: **Expression Tracks**

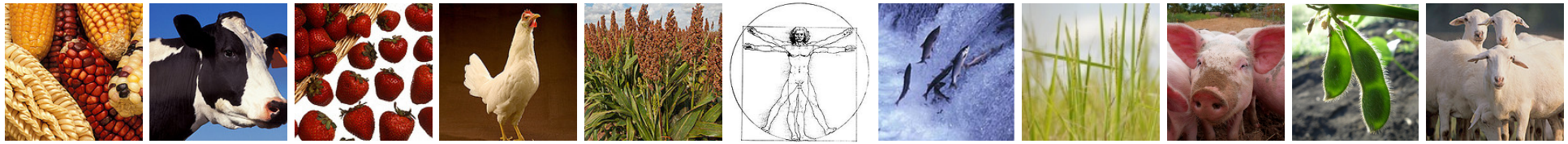


Genome Data Viewer: **Uploading Files**



From Annotation to Visualization

We annotate lots of genomes—and we will annotate your genome on request



The best place to get this annotation data is at NCBI—Gene, GDV, BLAST, etc.

Please submit your data!

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

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

Jinna Choi
Patrick Masterson
Eyal Mozes
Robert Smith
Alexandre Souvorov

RefSeq/Gene

Terence Murphy
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

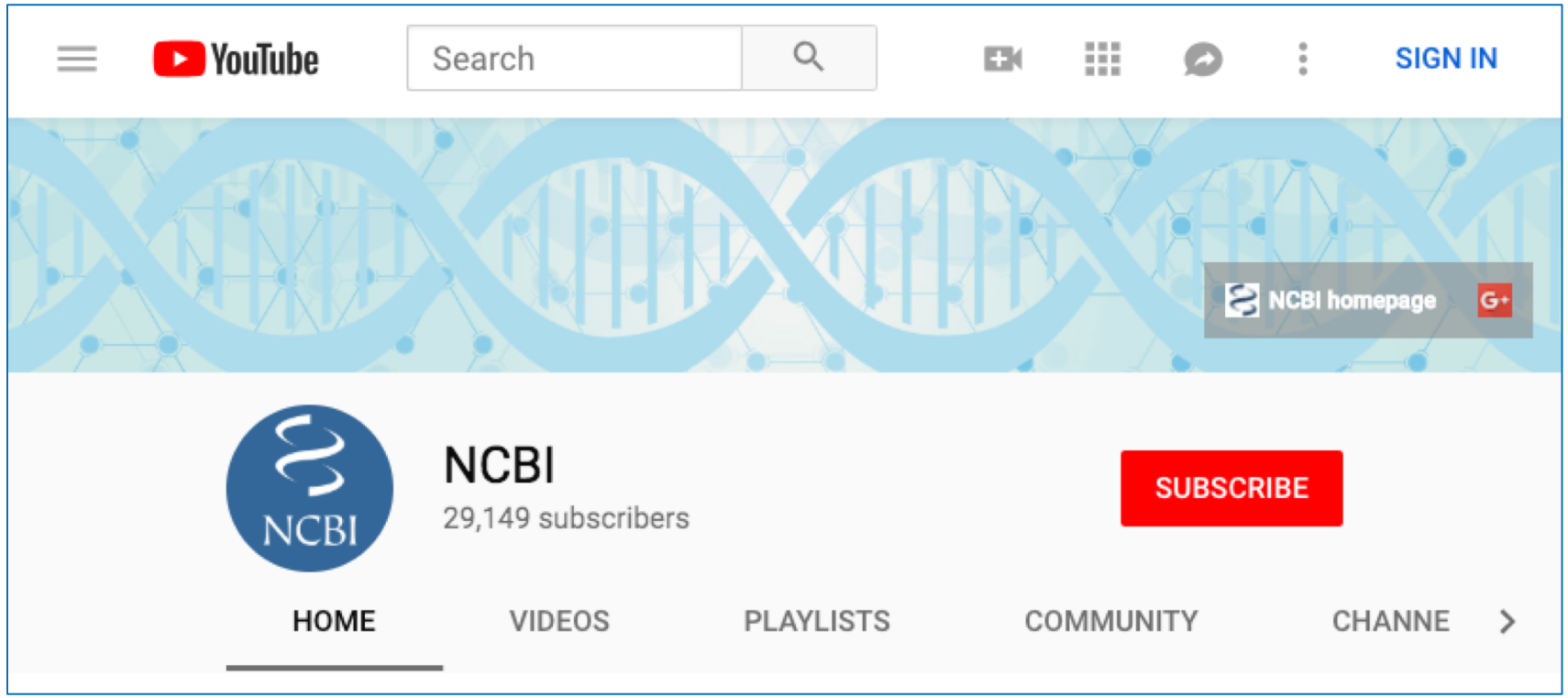
A cast of thousands

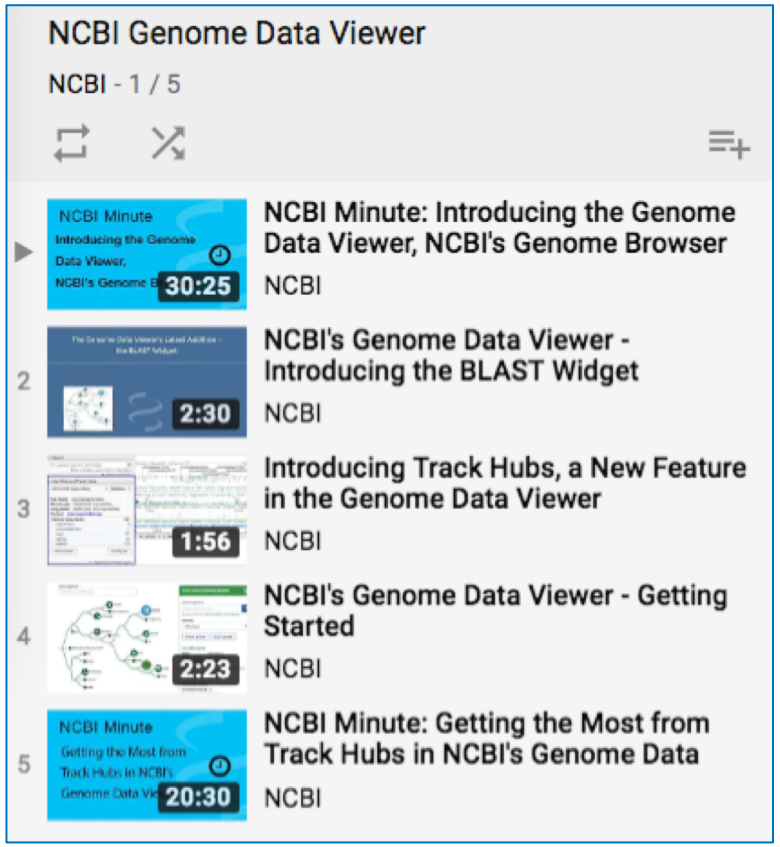
Ken Katz
Michael Ovetsky
Lukas Wagner
Andrei Shkeda
Donna Maglott
Kim Pruitt
Jim Ostell

Watch NCBI News for updates!

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

Subscribe to the NCBI YouTube Channel





NCBI

YouTube

NCBI Genome Resources Workshop

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