Equine genome resources at National Center for Biotechnology Information (NCBI)

Shashikant Pujar, Bhanu Rajput, Francoise Thibaud-Nissen, Terence D. Murphy, Paul Kitts, Kim D. Pruitt

National Center for Biotechnology Information (NCBI)

National Library of Medicine

National Institutes of Health

8600 Rockville Pike Bethesda, MD 20894, USA

Plant and Animal Genome XXIII

January 10-14, 2015, San Diego, CA

Visit Poster **P1115** in Bioinformatics/Databases section (Pujar S. *et al*)

Equine genome resources at NCBI

Outline

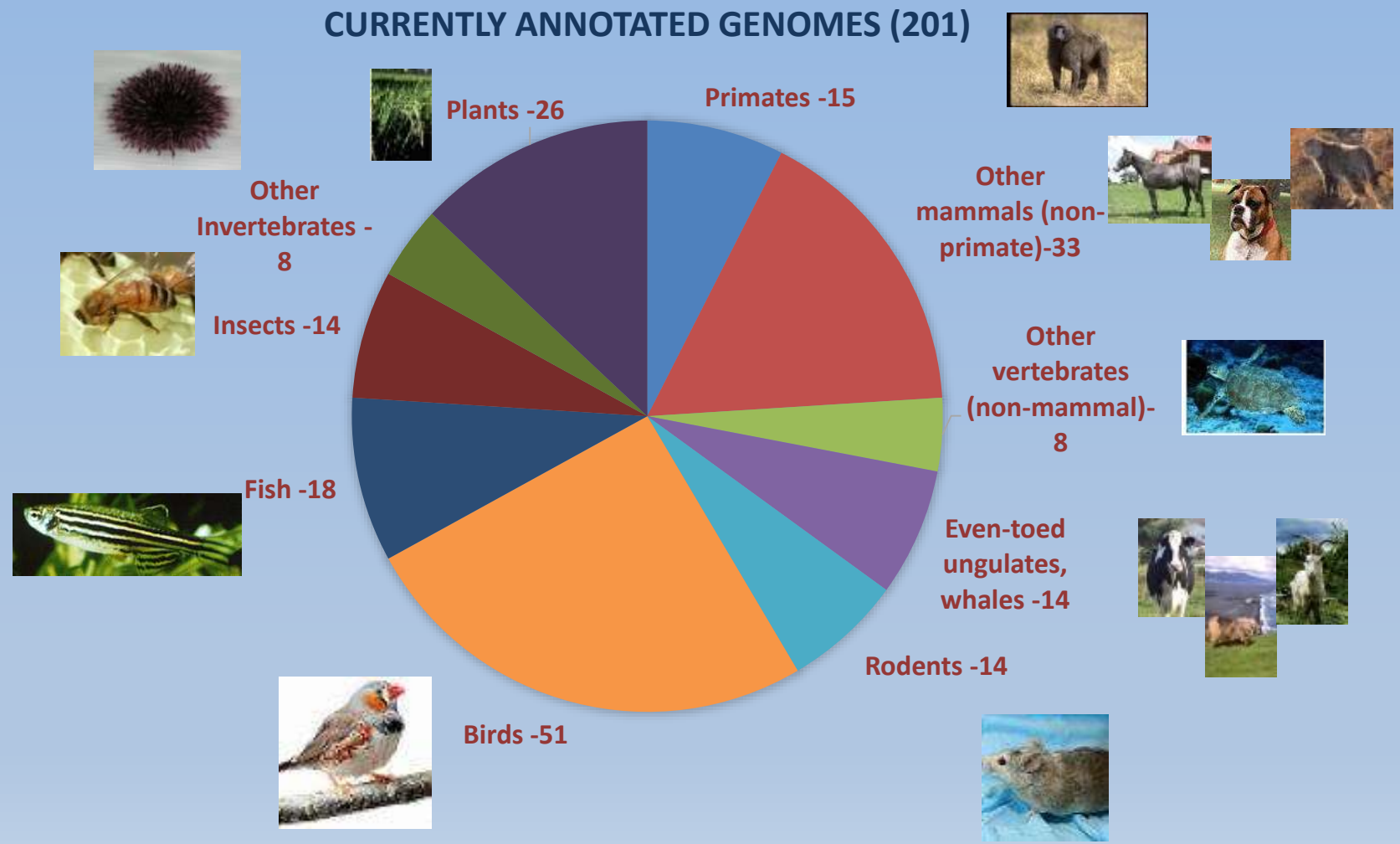
- **Genome annotation at NCBI - highlights**
- **Equine genome resources at NCBI**
- **User feedback: Contact information**

Visit poster P1105: From genes to genomes- new features and data access (Kim D. Pruitt *et al*)

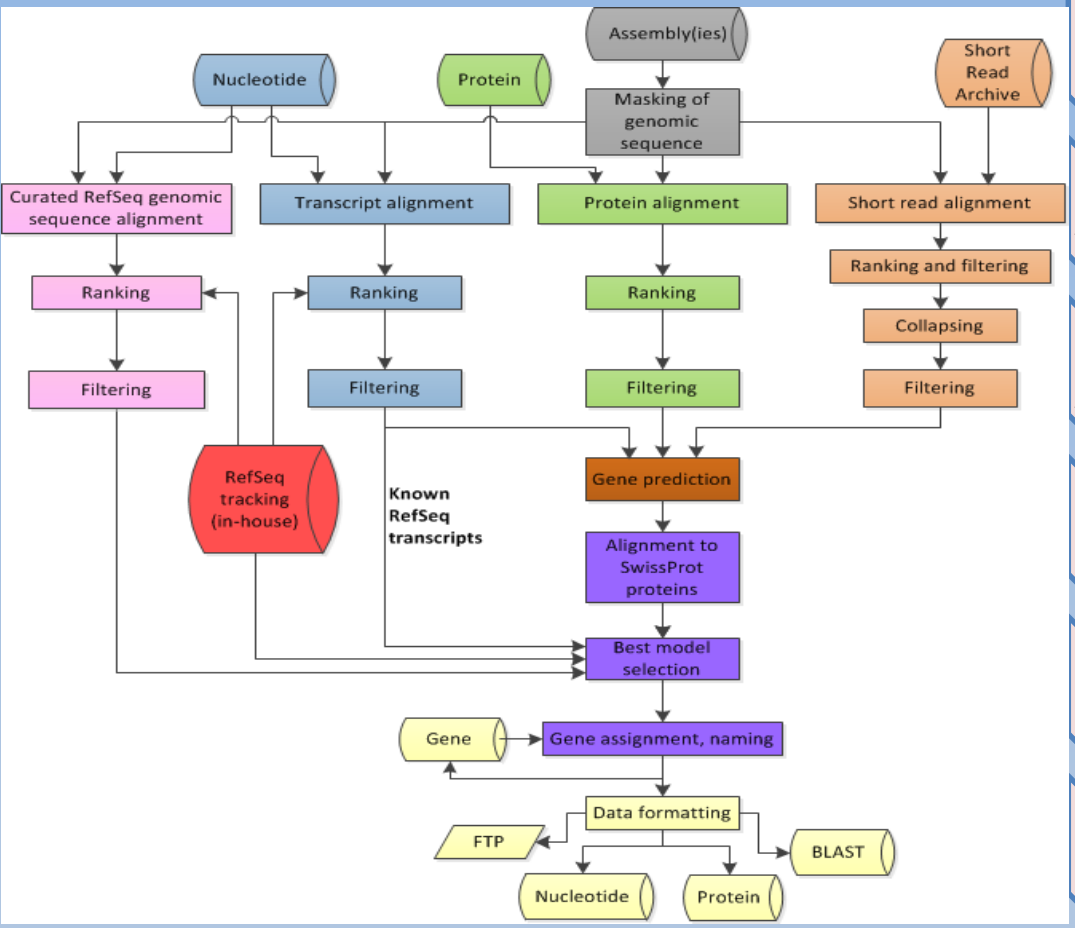
Eukaryotic Genome Annotation

www.ncbi.nlm.nih.gov/genome/annotation_euk/status/all

CURRENTLY ANNOTATED GENOMES (201)



Eukaryotic Genome Annotation



Reagents
- RefSeqs, GenBank transcripts and proteins
- RNA-seq data

Alignments
- Splign and ProSplign

Best Alignments
- Filter to obtain best alignments

Predict Models
- Gnomon predicts gene models, best models added to RefSeq set (XM_/XP, XR_)

Ortholog calls
- Based on reciprocal best hit and synteny calculation

Gene IDs and names
- Orthology used to assign names and locus types

***Speed (5-10 days), *RNA-seq data for gene prediction, *using aligned transcripts for 'gap-filling', *enhanced public reports**

Visit Poster P0051 for more on NCBI's eukaryotic genome annotation

Eukaryotic Genome Annotation

NCBI Resources How To shashikant.pujar@nih.gov My NCBI Sign Out

Genome Genome Search

Limits Advanced

Eukaryotic Annotation Home | Documentation | Annotated Genomes | Annotation Policy | Request Annotation

NCBI Equus caballus Annotation Release 101

The RefSeq genome records for Equus caballus were annotated by the NCBI Eukaryotic Genome Annotation Pipeline, an automated pipeline that annotates genes, transcripts and proteins on draft and finished genome assemblies. This report presents statistics on the annotation products, the input data used in the pipeline and intermediate alignment results.

The annotation products are available in the sequence databases and on the FTP site.

This report provides:

- Annotation Release information: The name of the release, important dates, the software version
- Assemblies: A brief description of the annotated assembly(ies)
- Gene and feature statistics: The counts and characteristics of the annotated features
- Masking of genomic sequence: How much of the genome was masked
- Transcript and protein alignments: The number and type of evidence retrieved from public databases and used for gene prediction

For more information on the annotation process, please visit the NCBI Eukaryotic Genome Annotation Pipeline page.

Annotation Release information

This annotation should be referred to as NCBI Equus caballus Annotation Release 101

Annotation release ID: 101
Date of Entrez queries for transcripts and proteins: Sep 13 2013
Date of submission of annotation to the public databases: Sep 23 2013
Software version: 5.1

Assemblies

The following assemblies were included in this annotation run:

Assembly name	Assembly accession	Submitter	Assembly date	Reference/Alternate	Assembly content
EquCab2.0	GCF_000002305.2	The Genome Assembly Team	10-29-2007	Reference	33 assembled chromosomes; unplaced scaffolds

This detailed report includes:

- **Counts of genes and gene subsets**
- **Counts and statistics of gene features**
- **Assembly masking**
- **Alignment counts/statistics**
- **List of RNA-seq datasets used**

www.ncbi.nlm.nih.gov/genome/annotation_euk/Equus_caballus/101/

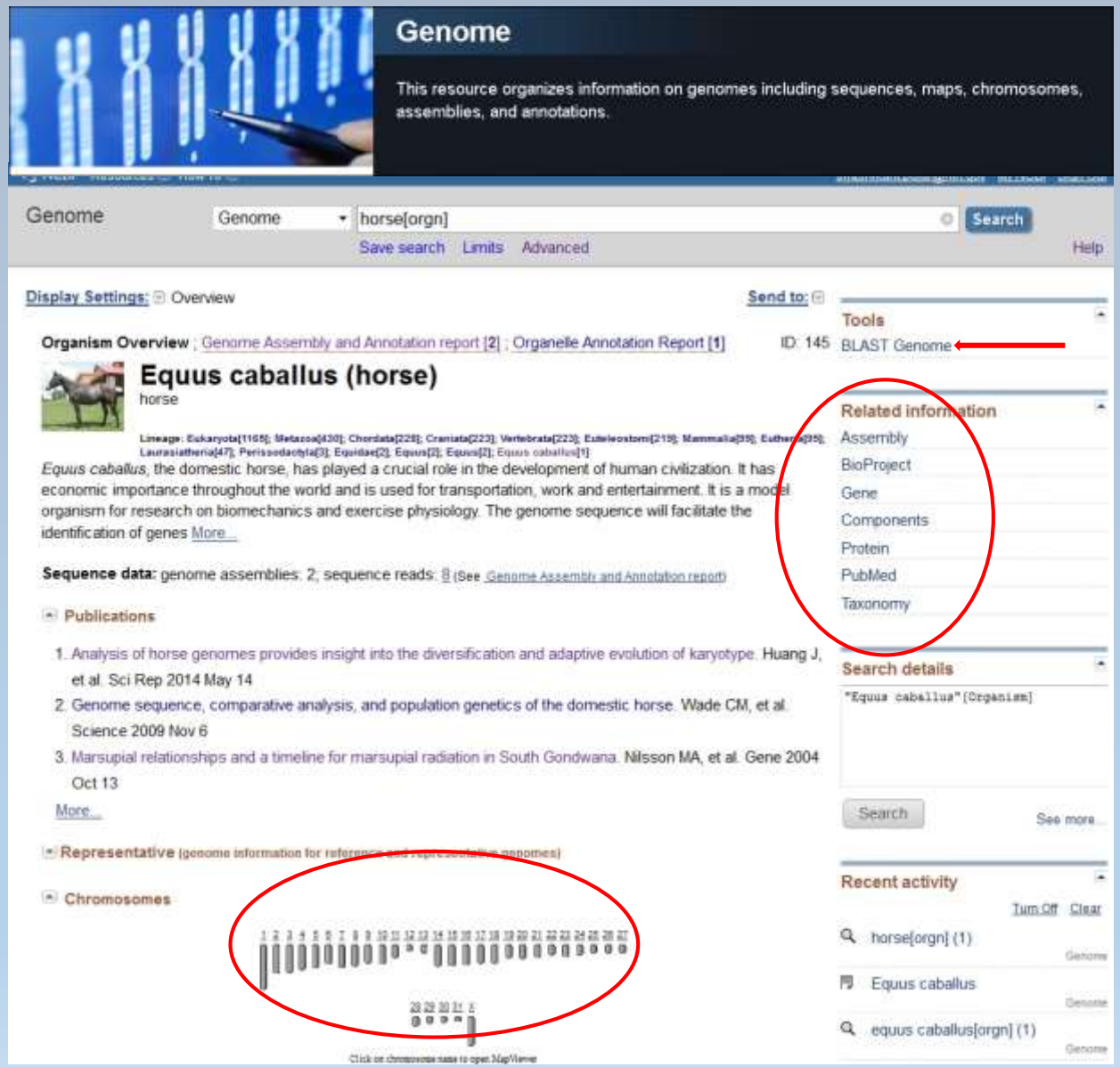
Equine Genome Resources at NCBI

- **Genome**
- **Map Viewer**
- **RefSeq**
- **Gene**
- **Blast**
- **Other resources: ftp, dbSNP**

Equus caballus; taxid: 9796
Assembly: EquCab2.0
Annotation Release 101
Total genes: 25,768
Total RefSeq RNA: 36,435

Equus przewalskii; taxid: 9798
Assembly: Burgud
Annotation Release 100
Total genes: 28,179
Total RefSeq RNA: 45,076

Equine resources at NCBI: Genome



- Every annotated tax ID has a genome page
- Gateway to other resources

www.ncbi.nlm.nih.gov/genome/?term=horse

Equine resources at NCBI: Map Viewer

NCBI

NCBI Map Viewer

PubMed | Nucleotide | Protein | Genome | Gene | Structure | PopSet | Taxonomy | Help

Search for | on chromosome(s) | Find | Advanced Search

Map Viewer

Map Viewer Home
Map Viewer Help
Horse Maps Help
FTP

NCBI Resources

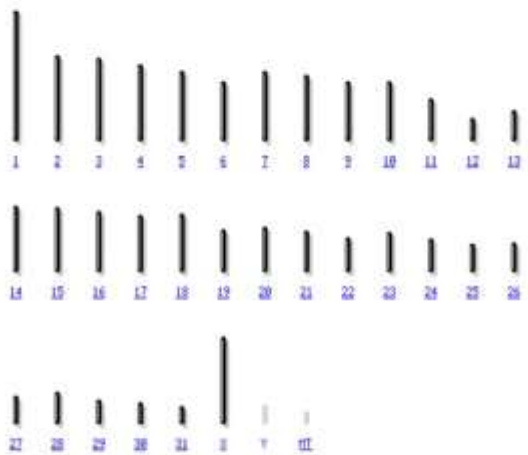
Assembly
Gene
Genome

Organism Data in GenBank

EST
Genomic
mRNA
Protein
WGS

Equus caballus (horse) genome view
Annotation Release 101 statistics

BLAST search *Equus caballus* sequences



Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Perissodactyla; Equidae; Equus; Equus caballus

September 2013, Annotation Release 101

This full annotation run includes the following assembly(ies):

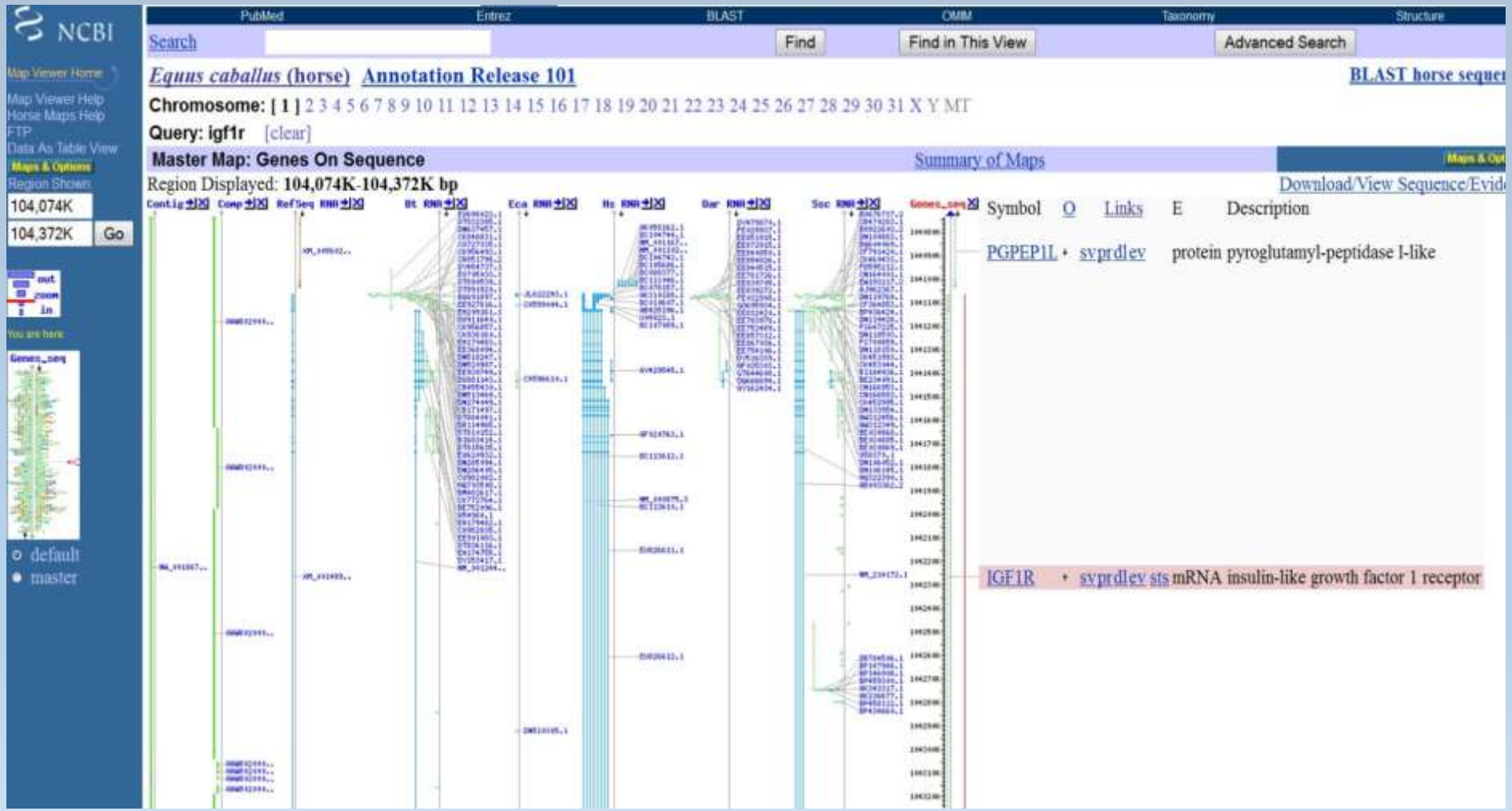
- EquCab2.0 (accession GCF_000002305.2)

Equus caballus, the domestic horse, has played a crucial role in the development of human civilization. It has economic importance throughout the world and is used for transportation, work, entertainment, and even war. It is a model organism for research on biomechanics and exercise physiology. The genome sequence will facilitate the identification of genes that are involved in mammalian evolution and help us to better understand the genetic basis of diseases common to both humans and horses.

www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9796

Equine resources at NCBI: Map Viewer

- Search Options: Gene symbol, Accession, Disease, Entrez queries.
- Maps include genomic assembly components, transcripts, proteins, STS markers, genetic maps.
- To be replaced by a new genome browser in the future.



Equine resources at NCBI: RefSeq

- **Accessions:** Genomic- Chr (AC_ NC_), Genomic (NG_ NT_ NW_)
  Transcripts- Known (NM_ NR_), Model (XM_, XR)
  Proteins - Known (NP_), Model (XP_)
- **Made by:** Automated process, manual curation and collaboration (nomenclature, gene families)
- **Using:** Mainly your data! (GenBank sequences, RNA-seq data)
- **Deliverables:** Gene range, genomic, transcripts, proteins, defined genomic regions (RSG, pseudogenes)
- **Impact:** Reagents in genome annotation, other NCBI DBs (Gene, Blast, Map Viewer), RefSeqGene
- **Access:** Individual accessions from Gene, Blast results, Map Viewer, Entrez search; bulk downloads from FTP

www.ncbi.nlm.nih.gov/refseq/

RefSeq Record: Transcript

NCBI Resources How To shashikant.pujar@nih.gov My NCBI Sign Out

Nucleotide | Nucleotide | Search | Advanced | Help

Display Settings: GenBank | Send:

Equus caballus insulin-like growth factor 1 (somatomedin C) (IGF1), mRNA

NCBI Reference Sequence: NM_001082498.2

FASTA Graphics

Go to:

```
LOCUS       NM_001082498            510 bp    mRNA    linear   MAM 02-MAR-2014
DEFINITION  Equus caballus insulin-like growth factor 1 (somatomedin C) (IGF1),
            mRNA.
ACCESSION   NM_001082498
VERSION     NM_001082498.2  GI:155969720
KEYWORDS    RefSeq.
SOURCE      Equus caballus (horse)
  ORGANISM  Equus caballus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Perissodactyla; Equidae; Equus.
REFERENCE   1  (bases 1 to 510)
  AUTHORS   Burns TA, Watts MR, Weber PS, McCutcheon LJ, Geor RJ and Belknap
            JK.
  TITLE     Distribution of insulin receptor and insulin-like growth factor-1
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the IGF1 gene

Sex-dependent insulin like growth factor-1 expressic [Theriogenology. 2013]

Distribution of insulin receptor and insulin-like growth fa [Equine Vet J. 2013]

Localization of insulin-like growth factor-I (IGF-I) and [Reprod Domest Anim. 2011]

See all

Format

- Summary
- GenBank
- GenBank (full)
- FASTA
- FASTA (text)
- Graphics
- ASN.1
- Revision History
- Accession List
- GI List

Apply

```
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from EU018459.1.
            On Aug 15, 2007 this sequence version replaced gi:126722854.

            ##Evidence-Data-START##
            Transcript exon combination :: EU018459.1, U85272.1 [ECO:0000332]
            RNAseq introns              :: single sample supports all introns
                                           SRS332422, SRS420619 [ECO:0000348]
            ##Evidence-Data-END##
```

```
FEATURES             Location/Qualifiers
     source          1..510
                     /organism="Equus caballus"
                     /mol_type="mRNA"
                     /db_xref="taxon:9796"
                     /chromosome="28"
                     /map="28"
```

Biological Attributes

- Bicistronic transcript CDS
- Uses downstream AUG
- Inferred exon combination
- Gene product(s) localized to mitochondrion
- Non-AUG initiation codon
- Nonsense-mediated mRNA decay
- PolyA required for stop codon
- Protein contains selenocysteine
- Readthrough transcript
- Ribosomal slippage
- Undergoes RNA editing
- Unitary pseudogene
- Regulatory uORF
- Multifuncational protein

RefSeq Record: Protein

Display Settings: GenPept

Send to:

insulin-like growth factor I precursor [Equus caballus]

NCBI Reference Sequence: NP_001075967.2

Identical Proteins FASTA Graphics

NEW!

Go to:

```
LOCUS       NP_001075967            153 aa            linear   MAM 02-MAR-2014
DEFINITION  insulin-like growth factor I precursor [Equus caballus].
ACCESSION   NP_001075967
VERSION     NP_001075967.2  GI:155969721
DBSOURCE    REFSEQ: accession NM_001082498.2
KEYWORDS    RefSeq.
SOURCE      Equus caballus (horse)
  ORGANISM  Equus caballus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Perissodactyla; Equidae; Equus.
REFERENCE   1  (residues 1 to 153)
  AUTHORS   Burns TA, Watts MR, Weber PS, McCutcheon LJ, Geor RJ and Belknap
            JK.
  TITLE     Distribution of insulin receptor and insulin-like growth factor-1
            receptor in the digital laminae of mixed-breed ponies: an
            immunohistochemical study
  JOURNAL   Equine Vet. J. 45 (3), 326-332 (2013)
```

Change region shown

Customize view

Analyze this sequence

Run BLAST

Identify Conserved Domains

Highlight Sequence Features

Find in this Sequence

Articles about the IGF1 gene

Sex-dependent insulin like growth factor-1 expressi [Theriogenology. 2013]

Cloning and sequencing of an equine insulin-lik [Gen Comp Endocrinol. 1996]

Distribution of insulin receptor and insulin-like growth [Equine Vet J. 2013]

See all...

RefSeq Record: Protein

Display Settings: GenPept

Send to:

Change region shown

insulin-like growth fac

NCBI Reference Sequence: NP_001

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_001075967
DEFINITION  insulin-like growth fact
ACCESSION   NP_001075967
VERSION     NP_001075967.2  GI:15596
DBSOURCE    REFSEQ: accession NM 001
KEYWORDS    RefSeq.
SOURCE      Equus caballus (horse)
  ORGANISM  Equus caballus
            Eukaryota; Metazoa; Chor
            Mammalia; Eutheria; Laur
REFERENCE   1  (residues 1 to 153)
  AUTHORS   Burns TA, Watts MR, Webe
            JK.
  TITLE     Distribution of insulin
            receptor in the digital
            immunohistochemical stud
  JOURNAL   Equine Vet. J. 45 (3), 3
```

Display Settings: Identical Protein Report

Send to:

insulin-like growth factor I precursor [Equus caballus]

NCBI Reference Sequence: NP_001075967.2

GenPept FASTA Graphics

RefSeq Selected Product: NP_001075967.2, 153 amino acids
Name: insulin-like growth factor I precursor [Equus caballus]

Source	CDS Region in Nucleotide	Protein	Organism	Superkingdom
RefSeq	NC_009171.2:26184077-26252800 (-)	NP_001075967.2	Equus caballus	Eukaryota
RefSeq	NM_001082498.2:16-477 (+)	NP_001075967.2	Equus caballus	Eukaryota
RefSeq	NW_001867399.1:26184077-26252800 (-)	NP_001075967.2	Equus caballus	Eukaryota
RefSeq	NW_003217384.1:469506-540014 (-)	XP_002915500.1	Ailuropoda melanoleuca	Eukaryota
RefSeq	XM_002915454.1:808-1269 (+)	XP_002915500.1	Ailuropoda melanoleuca	Eukaryota
RefSeq	NW_007673391.1:442979-512084 (-)	XP_008515872.1	Equus przewalskii	Eukaryota
RefSeq	XM_008517650.1:240-701 (+)	XP_008515872.1	Equus przewalskii	Eukaryota
RefSeq	NW_004450971.1:374563-444445 (-)	XP_004413239.1	Odobenus rosmarus divergens	Eukaryota
RefSeq	XM_004413182.1:808-1269 (+)	XP_004413239.1	Odobenus rosmarus divergens	Eukaryota
Swiss-Prot	N/A	Q6JLX1.2	Ailuropoda melanoleuca	Eukaryota
INSDC	AY369779.2:56-517 (+)	AAQ83972.2	Ailuropoda melanoleuca	Eukaryota
INSDC	EU018459.1:16-477 (+)	ABS57177.1	Equus caballus	Eukaryota

RefSeq Records: specialized curation

Display Settings: GenBank

Send:

Equus caballus paternally expressed 10 (PEG10), transcript variant 1, mRNA

NCBI Reference Sequence: NM_001172558.1

```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AAWR02013826.1 and
            AAWR02013827.1.
            On Feb 24, 2010 this sequence version replaced gi:194209639.

            Summary: This gene is thought to have been derived from the
            Ty3/Gypsy family of retrotransposons. It contains two overlapping
            open reading frames, RF1 and RF2, and expresses two proteins: a
            shorter, gag-like protein (with a CCHC-type zinc finger domain)
            from RF1; and a longer, gag/pol-like fusion protein (with an

            Transcript Variant: This variant (1) has two overlapping open
            reading frames (RF1 and RF2) and initiates translation from a
            non-AUG (CUG) start codon. This longer isoform (1) is derived from
            RF1/RF2 by -1 translational frameshifting.

     CDS            join(378..1442,1442..2608)
                    /gene="PEG10"
                    /ribosomal_slippage
                    /note="protein translation is dependent on -1 ribosomal
                    frameshift; isoform 1 is encoded by transcript variant 1;
```

Change region shown

Customize view

Analyze this sequence

PEG10 isoform 1 (NP_001166029.2).

More about the gene PEG10

This gene is thought to have been derived from the Ty3/Gypsy family of retrotransposons. It contains two overlapping open reading frames, RF...

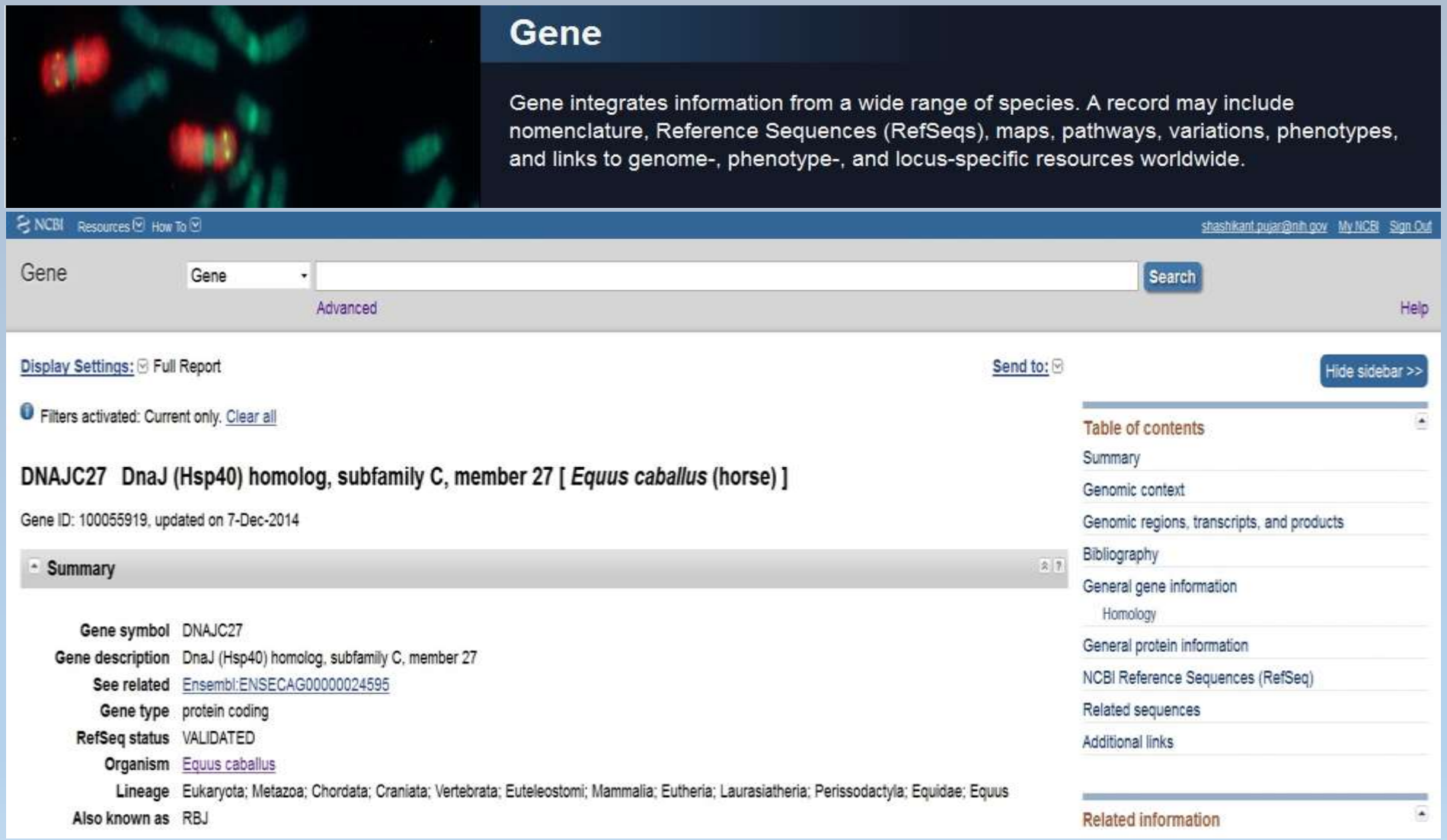
Components (Core)

Gene

Map Viewer

Exceptional biology annotation: Summary, transcript variant text, CDS note (curator: Bhanu Rajput)

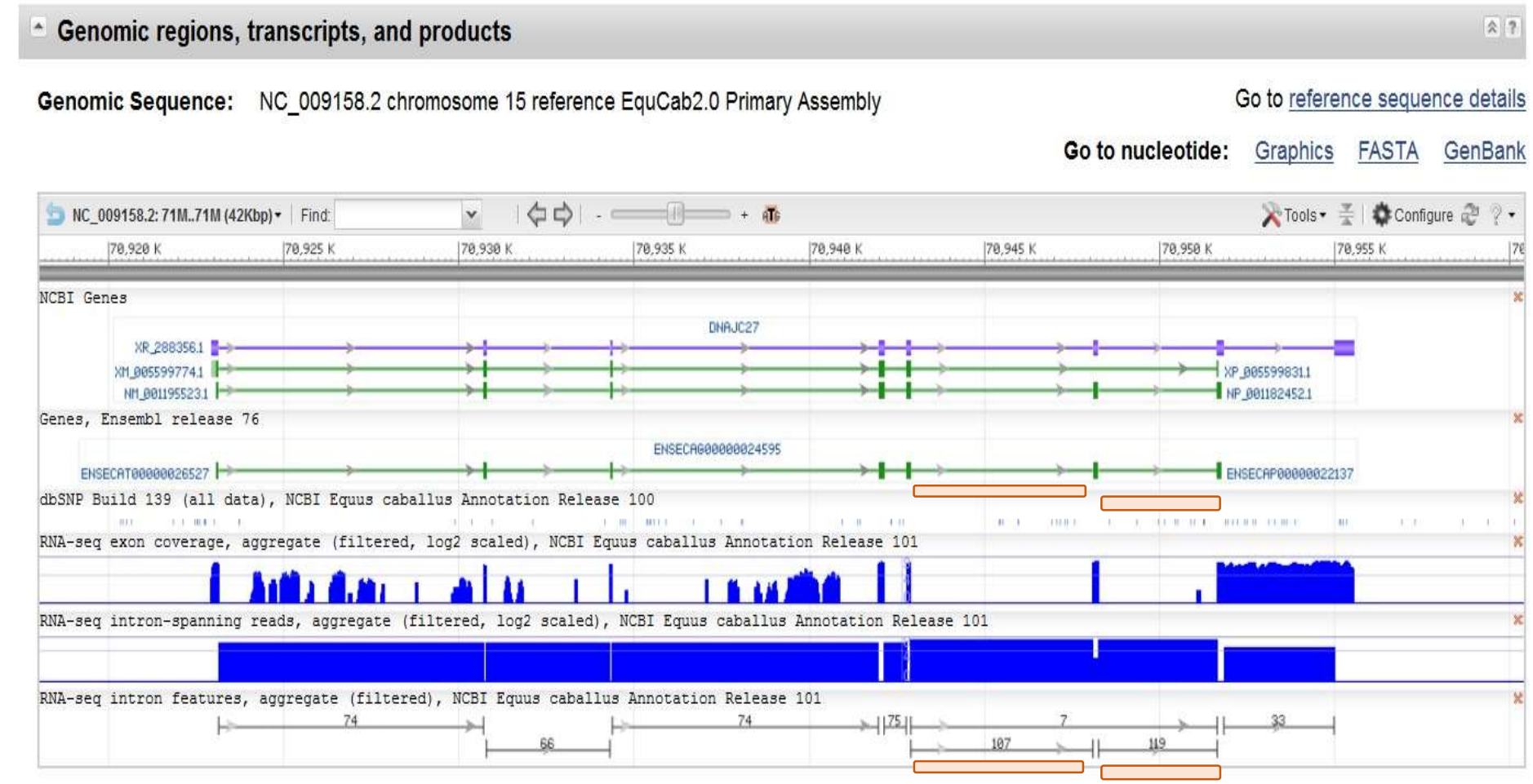
Equine resources at NCBI: Gene



www.ncbi.nlm.nih.gov/gene/100055919

Equine resources at NCBI: Gene

Sequence Viewer-embedded browser



RNA-seq data supports the splice variation seen among the RefSeq transcripts

Sequence Viewer can also be opened from a RefSeq record ("Graphics" link)

Visit Poster P1142 for more on Sequence Viewer

Equine resources at NCBI: Gene

Sequence Viewer-embedded browser

Genomic regions, transcripts, and products

Genomic Sequence: NC_009158.2 chromosome 15 reference EquCab2.0 Primary Assembly

Go to reference sequence details

Go to nucleotide: Graphics FASTA GenBank

Configure Page

Tracks | Custom Data

Active Tracks

Search Tracks

Sequence

Genes

Variation

Expression

Click on track to display settings. To re-order tracks, drag and drop track names.

- Sequence
- NCBI Genes
- Genes, Ensembl release 76
- dbSNP Build 139 (all data), NCBI Equus caballus Annotation Release 100
- RNA-seq exon coverage, aggregate (filtered, log2 scaled), NCBI Equus caballus Annotation Release 101
- RNA-seq intron-spanning reads, aggregate (filtered, log2 scaled), NCBI Equus caballus Annotation Relea...
- RNA-seq intron features, aggregate (filtered), NCBI Equus caballus Annotation Release 101

Track Settings: Sequence

A horizontal bar representing the currently viewed top sequence in a zoomed-out view, and overlaid with sequence letters in a zoomed-in view at sequence level. For nucleotide sequences, both the original sequence and the complementary sequence may be shown with two horizontal bars.

Other Settings: Show Label

Configure | Load Defaults | Cancel

Tools | Configure

70,940 K | 70,945 K | 70,950 K | 70,955 K

XP_005599831.1

NP_001182452.1

ENSECAP00000022137

otation Release 101

llus Annotation Release 101

RNA-seq intron features, aggregate (filtered), NCBI Equus caballus Annotation Release 101

74 | 66 | 74 | 75 | 7 | 107 | 119 | 33

RNA-seq data supports the splice variation seen among the RefSeq transcripts

Sequence Viewer can also be opened from a RefSeq record ("Graphics" link)

Visit Poster P1142 for more on Sequence Viewer

Equine resources at NCBI: Gene

Sequence Viewer-embedded browser

Genomic regions, transcripts, and products

Genomic Sequence: NC_009158.2 chromosome 15 reference EquCab2.0 Primary Assembly

Go to reference sequence details

Configure Page

Tracks | Custom Data

Active Tracks

Search Tracks

Sequence

Genes

Variation

Expression

Click on track to display settings. To re-order tracks, drag and drop track names.

- Sequence
- NCBI Genes
- Genes, Ensembl release 76
- dbSNP Build 139 (all data), NCBI Equus caballus Annotation Release 100
- RNA-seq exon coverage, aggregate (filtered, log2 scaled), NCBI Equus caballus Annotation Release 101
- RNA-seq intron-spanning reads, aggregate (filtered, log2 scaled), NCBI Equus caballus Annotation Relea...
- RNA-seq intron features, aggregate (filtered), NCBI Equus caballus Annotation Release 101

Track Settings: Sequence

A horizontal bar representing the currently viewed top sequence in a zoomed-out view, and overlaid with sequence letters in a zoomed-in view at sequence level. For nucleotide sequences, both the original sequence and the complementary sequence may be shown with two horizontal bars.

Other Settings: Show Label

Configure | Load Defaults | Cancel

Configure Page

Tracks | Custom Data

Data Source

- BLAST Results
- Data File
- URL
- Text

Please specify or drop an input file then press Upload to add new track(s).

File to upload: Browse... No file selected.

File format: auto detect

auto detect, asn binary, asn text, bed, bed 15, fasta, gff2, gff3, gtf/gff, gvf, vcf, wiggle, xml

Track Name:

Upload

Configure | Load Defaults | Cancel

RNA-seq intron features, aggregate (filtered), NCBI Equus caballus Annotation Release 101

74 | 66 | 74 | 75 | 7 | 107 | 119 | 33

RNA-seq data supports the splice variation seen among the RefSeq transcripts

Sequence Viewer can also be opened from a RefSeq record ("Graphics" link)

Visit Poster P1142 for more on Sequence Viewer

Equine resources at NCBI: Blast

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Your Recent Results New!

All Recent results...

News

PDB and Swiss-Prot on the FTP site

The NCBI is now distributing the BLAST databases for protein PDB (pdbaa) and Swiss-Prot (swissprot) as stand-alone BLAST databases, rather than as subsets of the non-redundant (nr) database.

Tue, 25 Nov 2014 16:00:00 EST

More BLAST news...

Tip of the Day

How to save custom search pages.

So you have made a few BLAST searches and after adjusting the database, organism limits and maybe a few Algorithm Parameters you arrive at what you think is a good search strategy.

More tips...

www.ncbi.nlm.nih.gov/blast/

Equine resources at NCBI: Blast

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite

Equus przewalskii (Przewalski's horse) Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Or, upload file

Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

Choose Search Set

Database: Genome (Burgud reference assembly top-level) 53097 sequences

- Genome (Burgud reference assembly top-level)
- RefSeq Genomic
- RefSeq RNA
- Annotated RNAs
- Gnomon RNAs

Exclude
Optional

Entrez Query
Optional

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Other ways to access: Genome page, Map Viewer page

Other Genomic Resources

- **Genomes ftp:**

ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/

  Snapshot of genome annotation

- **RefSeq ftp:** ftp://ftp.ncbi.nlm.nih.gov/refseq/

  RefSeqs across all taxa

  Bi-monthly releases, Daily updates

- **dbSNP:** www.ncbi.nlm.nih.gov/SNP/

Need help? Have comments/suggestions?

Contact Us!

- **Every NCBI page** Write to the Help Desk

- info@ncbi.nlm.nih.gov

  for general information

- pujars@ncbi.nlm.nih.gov **(Horse Genome Champion)**

  for genome/annotation issues

- refseq-admin@ncbi.nlm.nih.gov

  for RefSeq-related issues

Acknowledgements

- **Eukaryotic annotation group**
  - Paul Kitts
  - Terence Murphy
  - Francoise Thibaud-Nissen
  - Mike DiCuccio
  - Alexander Souvorov
  - Robert Smith
  - And many more

- **RefSeq group**
  - Kim Pruitt
  - Anjana Vatsan
  - Brian Smith-White
  - Bhanu Rajput
  - Catherine Farrell
  - Diana Haddad
  - David Webb
  - Eric Cox
  - Kelly McGarvey
  - Lilian Riddick
  - Matt Wright
  - Mike Murphy
  - Nuala O'Leary
  - Sanjida Rangwala
  - Shashi Pujar
  - Tripti Gupta
  - Vamsi Kodali
  - Vinita Joardar
  - Wendy Wu

- **Collaborators/others**
  - Nomenclature groups
  - Area experts

AND

- **YOU!**

(keep the data submissions coming!)

- **Other NCBI groups**
  - Who provide the various database resources

NCBI at PAGXXIII
NCBI Booth (#618)
All are welcome!

NCBI at PAGXXIII
Workshop: NCBI Genome Resources
Date: Monday, January 12, 2015
Time: 12:50 PM-3:00 PM
Room: Sunset
All are welcome!