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
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

- Primates
- Rodents
- Other mammals
- Birds
- Fish
- Other vertebrates
- Insects
- Other invertebrates
- Plants

- Last annotated before 2018
- Last annotated in 2018
Eukaryotic Genome Annotation Pipeline

Data formatting

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

[Diagram showing the process of genome annotation, including stages such as assembly, alignment, gene prediction, and data formatting.]

## 2018 Annotations for Agriculture

<table>
<thead>
<tr>
<th>Organism</th>
<th>Assembly</th>
<th>Submitted</th>
<th>Annotation available</th>
<th>RNA-Seq samples</th>
<th>RNA-Seq read count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penaeus vannamei (Pacific white shrimp)</td>
<td>ASM378908v1</td>
<td>11/16/2018</td>
<td>12/14/2018</td>
<td>133</td>
<td>6.45B</td>
</tr>
<tr>
<td>Coffea arabica (coffee)</td>
<td>Cara_1.0</td>
<td>11/08/2018</td>
<td>12/4/2018</td>
<td>66</td>
<td>2.15B</td>
</tr>
<tr>
<td>Oreochromis niloticus (Nile tilapia)</td>
<td>O_niloticus_U MD_NMBU</td>
<td>6/29/2018</td>
<td>7/26/2018</td>
<td>138</td>
<td>5.11B</td>
</tr>
<tr>
<td>Bubalus bubalis (water buffalo)</td>
<td>UOA_WB_1</td>
<td>5/14/2018</td>
<td>6/25/2018</td>
<td>274</td>
<td>15.57B</td>
</tr>
<tr>
<td>Gallus gallus (chicken)</td>
<td>GRCg6a</td>
<td>3/27/2018</td>
<td>5/17/2018</td>
<td>129</td>
<td>10.56B</td>
</tr>
<tr>
<td>Equus caballus (horse)</td>
<td>EquCab3.0</td>
<td>1/05/2018</td>
<td>1/26/2018</td>
<td>182</td>
<td>10.82B</td>
</tr>
</tbody>
</table>

New data model for complex assemblies

**Old model**

- Diploid assembly
  - Assembly
    - Assembly-units
      - Primary unit

- Haploid with alt loci
  - Assembly
    - Assembly-units
      - Primary unit

**New model: two assemblies**

- Principal pseudohaplotype
  - Assembly-units
    - Primary unit

- Alternate pseudohaplotype

Handout available
New Data Model for Diploid Assemblies

Koren et al.
Nature Biotechnol. 2018

Angus x Brahman F1 hybrid

Principal pseudohaplotype

Angus Sire haplotype (GCA_003369685.2)

Alternate pseudohaplotype

Brahman Dam haplotype (GCA_003369695.2)
New Data Model for Diploid Assemblies

Principal pseudohaplotype

Angus Sire
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)

Alternate pseudohaplotype

Brahman Dam
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

- We only annotate one haplotype

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
Getting Annotation Data for a Secondary Haplotype

Primary haplotype
- Gene A
- Gene B
- Gene C
- Gene D

Secondary haplotype
- Gene A
- Gene B
- Gene C

NCBI Genome Remapping Service
The Products of Annotation

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

Data formatting

- Assemblies
- mRNA-seq alignment
- RNA-seq alignment
- Protein alignment
- Transcript alignment
- Curated RefSeq genomic sequence alignment

- Ranking
- Filtering
- Gene prediction
- Alignment to SwissProt proteins
- Best model selection
- Small non-coding RNAs (Rfam, RNA scan SE)

- Gene assignment, naming
How do I find my gene?
Search NCBI for organism + gene symbol or name

- Bos taurus IGF1
- Bos taurus IGF1R
- Bos taurus IGF2
- Bos taurus IGF2R
- Bos taurus IGFBP5
- Bos taurus IGF-like family receptor 1

As-you-type suggestions
Searching for annotation products

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)

Gene

Genome Data Viewer
Gene: portal to NCBI annotation data

Handout available
Using orthology to get clues to function
Genomic Context Data in Gene

Location: 29qter
Exon count: 10

<table>
<thead>
<tr>
<th>Annotation release</th>
<th>Status</th>
<th>Assembly</th>
<th>Chr</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>106</td>
<td>current</td>
<td>ARS-UCD1.2 (GCF_002263795.1)</td>
<td>29</td>
<td>NC_037356.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(49395153..49422469)</td>
</tr>
<tr>
<td>105</td>
<td>previous assembly</td>
<td>Bos_taurus_UMD_3.1.1 (GCF_000003055.6)</td>
<td>29</td>
<td>AC_000186.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(50037872..50065231)</td>
</tr>
</tbody>
</table>

Chromosome 29 - NC_037356.1
What is Gene? Bibliography

Related articles in PubMed

1. SNP and haplotype analysis reveal IGF2 variants associated with growth traits in Chinese Qinchuan cattle.

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.

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
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

Genome Data Viewer: Control the Display with Widgets
Genome Data Viewer: Expression Tracks
Genome Data Viewer: Expression Tracks

Liver

Kidney
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

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

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

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!
https://www.youtube.com/user/NCBINLM
Subscribe to the NCBI YouTube Channel
# NCBI Genome Resources Workshop

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Presenters</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:50 – 1:10</td>
<td>Submission of Genomes to GenBank</td>
<td>Karen Clark</td>
</tr>
<tr>
<td>1:10 – 1:30</td>
<td>GEO Submissions and Usage</td>
<td>Steve Wilhite</td>
</tr>
<tr>
<td>1:30 – 1:55</td>
<td>From Annotation to Visualization: Exploring Genes and Genomes with NCBI Tools</td>
<td>Eric Cox</td>
</tr>
<tr>
<td>2:15 – 2:35</td>
<td>NCBI Resources for Phyletically-Defined Next Generation Analysis in and out of the Cloud (a.k.a. Cool New Stuff!)</td>
<td>Ben Busby</td>
</tr>
<tr>
<td>2:35 – 3:00</td>
<td>Q &amp; A session</td>
<td></td>
</tr>
</tbody>
</table>

Handouts available