Genome sequence, assembly, and annotation: from data submission to public resource at the NCBI

Small Skate Genome Annotation Workshop
Kim D. Pruitt
May 25, 2011
Outline

• What resources represent data associated with a genome submission?
• How does NCBI use annotated genome submissions in other resources?
• What is the NCBI eukaryotic annotation pipeline and when is it used?
• How does NCBI manage genome assembly data?
• How should a GenBank submission proceed?
What resources represent data associated with a genome submission

- Genome Assembly & Annotation
- GenBank Processing
- Accessions Sequence DB
- Public Access
Submission -> Public Access (Primary Archives)

• Genome sequence:
  – NCBI nucleotide database
  – Contig browser
  – Assembly database (future)

• Protein annotations:
  – Protein database
  – Blink

• All sequence data on FTP site:

• BLAST databases

• BioProject database (formerly Genome Project db)
**Whole Genome Shotgun sequencing project**

**WGS record:** ABL001

---

**Genome sequence (turquoise killifish)**

**GenBank:** ABL000000000.1

This entry is the master record for the sequence data.

---

**Accession**: ABL000000000.1

**Version**: ABL000000000

**DBLink**: WGS

**Source**: Notobranchius furzeri

**Organism**: Notobranchius furzeri

**Locus**: ABLO10000000

---

**ABLO00000000.1 Notobranchius furzeri strain GRZ**

**Contigs:** 5,299

**Updated:** 17-MAR-2009

---

### Downloads (data files)

- **Genbank**: wgs.ABLO.mstr.gz, 1 kb
- **FASTA**: wgs.ABLO.fsa.gz, 1.5 Mb
- **gb**: wgs.ABLO.gbff.gz, 2.4 Mb

### Downloads (data files)

- **Quality Score**: wgs.ABLO.1.qscore.gz, 3.1 Mb

---

### Accession | GenBank | Contig | Length, bp | FASTA | Quality Score
---|---|---|---|---|---
ABL00000000.1 | gb | GRZ-G-b-01a01.f1 | 660 | >gb|ABLO0000001.1|Notobranchius furzeri strain GRZ clone GRZ-G-b-01a01 GRZ-G-b-01a01.f1, whole genome shotgun sequence
ABL00000000.1 | gb | GRZ-G-b-01a01.r1 | 802 |
ABL00000000.1 | gb | GRZ-G-b-01b01.r1 | 1,406 |
ABL00000000.1 | gb | GRZ-G-b-01c01.f1 | 1,385 |
ABL00000000.1 | gb | GRZ-G-b-01c01.r1 | 1,288 |
ABL00000000.1 | gb | GRZ-G-b-01d01.r1 | 468 |
ABL00000000.1 | gb | GRZ-G-b-01d01.f1 | 979 |
ABL00000000.1 | gb | GRZ-G-b-01e01.f1 | 724 |
ABL00000000.1 | gb | GRZ-G-b-01e01.r1 | 667 |
ABL00000000.1 | gb | GRZ-G-b-01f01.f1 | 294 |
ABL00000000.1 | gb | GRZ-G-b-01f01.r1 | 646 |
ABL00000000.1 | gb | GRZ-G-b-01g01.f1 | 1,199 |
ABL00000000.1 | gb | GRZ-G-b-01g01.r1 | 1,168 |
ABL00000000.1 | gb | GRZ-G-b-01h01.f1 | 1,430 |
ABL00000000.1 | gb | GRZ-G-b-01h01.r1 | 1,780 |
ABL00000000.1 | gb | GRZ-G-b-01i01.r1 | 1,326 |
ABL00000000.1 | gb | GRZ-G-b-01i01.f1 | 555 |
ABL00000000.1 | gb | GRZ-G-b-01j01.f1 | 719 |
ABL00000000.1 | gb | GRZ-G-b-01j01.r1 | 344 |
ABL00000000.1 | gb | GRZ-G-b-01k01.f1 | 824 |

---

**Write to the Help Desk** | **Privacy Notice** | **Disclaimer** | **Accessibility**

---

**Last update:** Thu, 19 May 2011 Rev. 288988
**Name:** Notobranchius furzeri (turquoise killifish)  
**Title:** Low coverage genome sequence of the turquoise killifish, Notobranchius furzeri

Notobranchius furzeri is a vertebrate with a very short life span in captivity, e.g., the GRZ strain, which we have started to sequence and genetically analyze. It lives only up to 12-13 weeks in the lab, even under optimal conditions. With our work we want to contribute to establishing N. furzeri as an alternative vertebrate model organism for aging research.

**Attributes:**  
Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Sequencing;  
Acantophora, Acanthopterygii, Perciformes, Actinopterygii, Cypriodontiformes, Notobranchiidae, Notobranchius; Notobranchius furzeri

**Publications:**  
1. Reichwald K et al., "High tandem repeat content in the genome of the short-lived annual fish Notobranchius furzeri: a new vertebrate model for aging research," Genomics 2009 Feb 11; 99 (2) R15

---

**Umbrella Projects**

<table>
<thead>
<tr>
<th>Umbrella Project Title</th>
<th>Number of Subprojects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notobranchius furzeri (organism overview)</td>
<td>▼ 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BioProject accession</th>
<th>Project type</th>
<th>Organism</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRJNA29535</td>
<td>Genome sequencing</td>
<td>Notobranchius furzeri</td>
<td>Low coverage genome sequence of the turquoise killifish, Notobranchius furzeri (Kathrin Reichwald)</td>
</tr>
<tr>
<td>PRJNA33315</td>
<td>Genome sequencing</td>
<td>Notobranchius furzeri</td>
<td>Notobranchius furzeri (Dept. of Genome Analysis)</td>
</tr>
<tr>
<td>PRJNA33503</td>
<td>RefSeq Genome</td>
<td>Notobranchius furzeri</td>
<td>Notobranchius furzeri (Fritz Lipmann Institute,...)</td>
</tr>
</tbody>
</table>

**Submission:**  
Registration date: 4-May-2008  
Leibniz Institute for Age Research - Fritz Lipmann Institute (FLI)  
- Kathrin Reichwald  
- Dept. of Genome Analysis
What resources represent data associated with a genome submission

- Genome Assembly & Annotation
- GenBank Processing
- Accessions Sequence DB
- Public Access
  - Sequence DBs
  - BLAST, BLink
  - Links
  - FTP
- BioProject (locus_tag)
- BioSample (biological material)
- www
- Links
- FTP
How does NCBI use annotated genome submissions in other resources?

Genome Assembly & Annotation

GenBank Processing

Accessions Sequence DB

Public Access
- www
- BLAST, BLink
- Links
- FTP

RefSeq

Derivative Resources
GenBank to RefSeq

• RefSeq makes a copy of the GenBank sequence when:
  – Submission of a new organism (not in RefSeq), or submission of an alternate assembly (rare)
  – Sequence depth ~>5x
  – Assembly to chromosomes, linkage groups, scaffolds
  – Whole genome approach
  – RefSeq processing forks depending on annotation
Why?

- RefSeq facilitates management of derivative resources.
- RefSeq enables independent annotation updates to the RefSeq record if needed or requested.
How do I recognize RefSeq records?

• All RefSeq accessions include an underscore
• RefSeq instantiates the mRNA as a separate record, GenBank does not (from genome submission)

Right: NM_002111.6
Wrong: NM002111.6, NM 002111.6

Scope and Size (May 7 2011)
Species: 12,000
Genomic: 2,387,395
Protein: 12,625,466
RNA: 2,619,015

More information:
www.ncbi.nlm.nih.gov/books/NBK21091/
## RefSeq versus GenBank (INSDC)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>GenBank</th>
<th>RefSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Member of INSDC</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Source of sequence</td>
<td>Submitter</td>
<td>GenBank (INSDC)</td>
</tr>
<tr>
<td>‘Owner’ of sequence</td>
<td>Submitter</td>
<td>NCBI</td>
</tr>
<tr>
<td>Source of annotation</td>
<td>Submitter</td>
<td>Multiple</td>
</tr>
<tr>
<td>NCBI staff can update based on any request</td>
<td>No; submitter must authorize</td>
<td>Yes</td>
</tr>
<tr>
<td>Redundant in database</td>
<td>Frequently</td>
<td>Infrequently</td>
</tr>
<tr>
<td>Like a review article</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

More information:  
**NCBI Reference Sequence**: NC_007136.5

**FASTA**

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>NC_007136</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFINITION</td>
<td>Danio rerio strain Tuebingen chromosome 25, Zv9.</td>
</tr>
<tr>
<td>ACCESSION</td>
<td>NC_007136.5 GI:312144705</td>
</tr>
<tr>
<td>SOURCE</td>
<td>Danio rerio (zebrafish)</td>
</tr>
<tr>
<td>ORGANISM</td>
<td>Danio rerio</td>
</tr>
<tr>
<td>COMMENT</td>
<td>Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio.</td>
</tr>
<tr>
<td>COMMENT</td>
<td>REFSEQ INFORMATION: Features on this sequence have been produced for build 5 version 1 of the NCBI's genome annotation [see documentation]. The reference sequence is identical to CU651628.3. On Mar 24, 2011 this sequence version replaced gi:258456196. Assembly Name: Zv9. The genomic sequence for this RefSeq record is from the genome assembly released by the Wellcome Trust Sanger Institute as Zv9 in July 2010 (see <a href="http://www.sanger.ac.uk/Projects/D_rerio/Zv9_assembly_information.shtml">http://www.sanger.ac.uk/Projects/D_rerio/Zv9_assembly_information.shtml</a>). The Zv9 assembly is a mixed assembly of clones and contigs from the whole genome shotgun project CABZ0000000.1.</td>
</tr>
</tbody>
</table>

**Graphics**
RefSeq: a foundation for derivative resources & process flows

- Gene
- Map Viewer
- HomoloGene
- Protein Clusters
- Entrez Genomes
- NCBI genome annotation pipelines
- Variation annotation pipelines
Gene

• Aggregates information from many sources
• RefSeq is the primary source of genes
• Integrates:
  – Sequence information
  – Name information
  – Bibliography
  – Interactions, GO, pathways, disease
  – Links to NCBI and other resources
• Markers
• Pathways
• GO
• Other names
• RefSeq
• Other Seqs
• More LINKS
Two graphical views

- Graphical sequence display of feature annotation
- Map Viewer – NCBI’s genome browser
Graphics Display – available for ALL accessions
What is the NCBI eukaryotic annotation pipeline and when is it used?

- **What**
  - Calculate alignments and predicted models but RefSeq reflects the submitted annotation (full copy from GenBank)
  - Above, plus full genome annotation of the RefSeq genome
    - *RefSeq genome differs from GenBank in annotation details*
  - Occasionally, calculate annotation results for external group to use in their GenBank submission

- **When:**
  - Reasonable quality genome assembly and/or request from organized group
Process Flow, if RefSeq uses submitted annotation

Input Data:
- Genome Assembly
- Transcripts
- Proteins, from cDNAs

DataFreeze

NCBI Genome Pipeline

ProSplign (Protein)
Gnomon (build models)
Splign (RNA)
BLAST hits
Prepare Map Viewer files
Prepare FTP files
Update Resources:
- Map Viewer
- BLAST dbs
- FTP

Public Release

GenBank Submission
public
**Genome Annotation Process Flow**

**Input Data:**
- Genome Assembly
- Transcripts: RefSeq, GenBank
- Proteins from RefSeq, and cDNAs

**DataFreeze**
- BLAST hits
- ProSplign
- Splign
- Gnomon

**NCBI Genome Annotation Pipeline**
- Model selection, tracking, naming
- Prepare RefSeq ASN

**RefSeq Curation Pipeline**
- Prepare Map Viewer files
- Prepare FTP files

**Public Release**
- Update Resources: RefSeq, Map Viewer Gene, BLAST dbs, FTP

**NCBI/NLM/NIH**
Input Data

• Genome Assembly
  – contamination screening
  – masked with WindowMasker or RepeatMasker

• Transcripts (best data for supporting models)
  – same-species (or closely-related species) mRNAs & ESTs
  – aligned to genome with Splign, constrained to consensus splice sites

• Proteins
  – Proteins annotated on cDNAs (avoid other *ab initio* predictions)
  – Proteins from the same species, and from closely related species
  – RefSeq proteins from highest quality annotated genomes – human, fly, yeast, nematode
  – aligned to genome with ProSplign, constrained to consensus splice sites
Example ProSplign alignment (Bee)

mouse
fly
human
Frog
Zebrafish
Orangutan
Arabidopsis
Fungus
worm

BH1B002219M3
BH1B0082N09M3
BH1B0082H03M3
BH1B0092K08M3
BH1B0067T22M3
BH1B0022A04M3

[Example ProSplign alignment diagram]
Multi-step gene prediction

mRNA
EST

Protein

HMM prediction

model

filtering

Best BLAST hit
Splign

Best BLAST hit
proSplign

Gnomon

filtering

annotation
Some Models are Well Supported
Some Models are Supported by Chained Evidence
Some Models are Extended by *ab initio* Predictions

---

**Ref:** NP_077357.2  
**UG** glycerol kinase [Rattus norvegicus]

**Length:** 524

<table>
<thead>
<tr>
<th>Gene ID: 79223 Gk</th>
<th>glycerol kinase [Rattus norvegicus]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(10 or fewer PubMed links)</td>
<td></td>
</tr>
</tbody>
</table>

**Score:** 468 bits (1203), Expect = 3e-136, Method: Compositional matrix adjust.

**Identities:** 234/447 (52%), **Positives:** 301/447 (67%), **Gaps:** 11/447 (2%)

**Query** 59

```
KSAAEKFVGVIDEVSSSTRFLVFSSTQLPVASHEISLTHNISGYEGWQDPENLIEV 118
K A VG +D+G SSTRFLVP +S +H++ + EGWVEQDP+ +L V
```

**Sbjct** 5

```
KKAVLGPVLGAVDQGTSSSTFRSVNKTAEALLSHHQVEIKQEPFREGWVEQDFKEILQSV 64
```

**Query** 119

```
KRIVTVTCNKLKNIISAPSAIAAIGVINQDITLVWDKYTGSPLYNAIVMMDVMTIQSIV 178
I TC KL ++NI S I AIGV+NR+IT++VWDK TG FLyna+VW+D + IQS V
```

**Sbjct** 65

```
YECIEKTECQNLNIDSNခAIGVSNQRELNLTVWDL-ELIQSTV 123
```

**Query** 179

```
NRFINKRIPNCNDLDEQKSGSTINPYFSVVKLWLMENIPQVSKAIĐERCMFGTMDSWL 238
+ ++KRIP ++ + K+G ++ YFS VKL WL+NN+ +V +A+E R +FGT+DSWL
```

**Sbjct** 124

```
EK-LSKRIPIGNNVFKNVLGTLPLSTYFSAVKLRLDNVKKVQEAVERNARGLFDTISWL 182
```

**Query** 239

```
INWILIGGINGVHITDVNASRIMLMNHLRSWDEKTLIEFFDPKSLILPDIRSCSEVYG 298
IN+LIGGINGVH TDVNASRA NLHSL WDK L EFF IP ILP++RS SE+YG
```

**Sbjct** 183

```
INSLILGGINGVHITDVNASRMLNHSLWDEKIELCFFGIPME-ILPVNRSSEIYG 241
```

**Query** 299

```
FMAGGPFIKTPISGCIGDHQAGLGQLCFTAGQAKCTFGSNCFLYNTGRKFVISTHGLL 358
M G PISGC+GD AL+GQ+CF GQA+ T+G CFLLL NTG K V S HLSL
```

**Sbjct** 242

```
LMKAGALEGVPISCGLDQSAALVGQCMQDFQDGKNTQGYTGFCFLLCMTGHKVFCSEHGLL 301
```

**Query** 359

```
TTVAYKIGKHSDPIYALEGSVVVAENTREWLDNLWIMDMVYKSEAKEKDDKNI------Y 412
```

**Sbjct** 302

```
TTVAYK+G+ YALESGV +A WL+DL I +S+E EK K + Y
```

**Query** 413

```
FVFPFELYAPFWITGIDGKCSCGVTMETTAEHLREMSTLEGICFQTKMVMQSMQSDTGHVF 472
```

**Sbjct** 360

```
FVPF EL YAPFWITS  +G G +I T + H+ LE +CFQ+E++ +M D G P+
```

**Query** 473

```
MALNVDDGSMNTDFKLKILTNICCLPV 499
L VDGMM++N +++ +I +FP
```

**Sbjct** 420

```
SHLQVDGGMTSNKILMLQADILYIFV 446
```
Factors affecting annotation

• Assembly quality (gaps and indels)

• Same-species transcript data
  – Needed to train *ab initio* component of Gnomon
  – Significantly improves splice site calls and final model

• Evolutionary distance from known proteins
Assembly InDels (and Gaps)

PREDICTED: LOW QUALITY PROTEIN: vacuolar protein sorting-associated protein 13C-like [Macaca mulatta]

NCBI Reference Sequence: XP_002808497.1

Comment Features Sequence

LOCUS XP_002808497 3754 aa linear PRI 01-JUN-2010
DEFINITION PREDICTED: LOW QUALITY PROTEIN: vacuolar protein sorting-associated protein 13C-like [Macaca mulatta].
ACCESSION XP_002808497
VERSION XP_002808497.1 GI:297296580
DBSOURCE REFSEQ: accession XM 002808451.1

gene

1..13377
/gene="LOC709080"
/note="The sequence of the transcript was modified to remove a frameshift represented in this assembly; Derived by automated computational analysis using gene prediction method: Gnomon. Supporting evidence includes similarity to: 9 mRNAs, 163 ESTs, 11 Proteins"
/db xref="GeneID:709080"
How does NCBI manage genome assembly data?

• Assembly database

• Metadata
  – Assembly name, alternate name
  – Submitting center
  – Publications

• Assembly Details – what sequences are part of an assembly
  – AGP file
  – WGS with chromosomes, vs. scaffolds & contigs only, vs. complete bacterial genome

• Track changes over time

• Statistics

New!
Why do we need to manage assembly data? (What’s the problem?)

- Different browsers use different names
- It’s not clear if different browsers are showing the same data
- It’s not clear which sequence versions are part of an assembly
- It’s not clear what changed when an assembly updates
Assembly database

• The assembly database provides a distinct ID for the assembly and each unit of the assembly
• Tracks all changes (over time, & vs. RefSeq)
• Supports QA; public reporting of assembly statistics is under development
  – contamination screening
  – assembly stats
  – Compare assemblies upon update

Zv9    GCA_000002035.2
danRer7
Submitting Data to GenBank

• NCBI’s mission is to serve as a permanent archive of your data

• Why should you submit?
  – Do you really want to commit to hosting your data 20 years from now?
  – Do you really want to commit to maintaining a backup archive and converting it to new archive formats?
  – Submitting genome assemblies, cDNAs, ESTs, RT-PCR or clone-end sequences helps improve the public gene sets and supports future research.
Submitting Genomes

Data → FASTA → Feature table → AGP → GenBank Processing → BioProject → Accessions Sequence DB → Assembly DB


Public Access: www, BLAST, FTP

Assembly Browser → NCBI Annotation

Gene, RefSeq, Map Viewer
Register the project in BioProject DB

Submission Wizard (coming soon)

Submission Portal

Submission: SUB000274 > BioProject

Submitter General info Project type Target Publications Overview

Submitter

- First name
  - Kim

- Last name
  - Pruitt

- E-mail (primary)
  - pruitt@ncbi.nlm.nih.gov

- E-mail (secondary)

- Submitting organization
  - TEST

- Submitting organization URL

Continue
Project Type

- **Project data type**: Genome Sequencing

- **Sample scope**: Monoisolate
- **Material**: Genome
- **Capture**: Whole

- **Methodology**: Sequencing

- **Objective**: Autogenerate locus tag prefix

Sample scope choices:

- **Monoisolate**: a single animal, cultured cell-line, inbred population (or possibly a heterogeneous population when a single genome assembly is generated from the pooled sample; not preferred).
- **Multiisolate**: multiple individuals, a population (representative of a species).
- **Multi-species**: sample represents multiple species.
- **Environment**: the species content of the sample is not known.
- **Synthetic**: the sample is synthetically created by a machine.
- **Other**: specify the sample scope that was used.
WGS Submission Requirements

1. Register the genome sequencing project in BioProject DB
2. Make the genome assembly and annotation data files.
3. Run the command line program tbl2asn to generate the .sqn file for submission and the validation and discrepancy report files.
4. Fix problems that are indicated in the .val and discrep files
5. If you have higher-level assembly information, scaffolds and/or chromosomes, then generate an AGP file to build those objects from the wgs-contigs.
6. Submit with GenomesMacroSend
7. Send a message to genomes@ncbi.nlm.nih.gov with the submission information listed.
Sequin

• Use Sequin to see what the flatfile display will look like:
  – Download Sequin
  – Open the .sqn file
Make the genome assembly data files

• generate a template file with submitter and publication information.

• Represent the sequence data:
  – put the contig sequences into fasta format files with suffix .fsa. Each sequence has a definition line with a unique identifier.
  – quality scores of the sequences (optional, but desired)
  – Assembly information to build scaffolds and chromosomes (AGP file)

• annotation files have the same base names as the .fsa files and use the same sequence identifiers, but have the suffix .tbl
Submitting Genome Annotations

- Submitting annotation isn’t too difficult
  - 5 column feature table, NOT GFF or flatfile
- Eukaryotic Genome Annotation Submission Guide:
Annotation details – watch out for..

• When annotating scaffolds, you need to be careful about crossing gaps.
  – If the gap size is unknown, as defined in the AGP file, then
    • a CDS may not cross the gap.
    • Instead, you could have two partial CDS and partial mRNAs abutting the gap, with a single gene over the whole locus.
  – If the gap size is known, then
    • a CDS can cross the gap.
    • However, a CDS should not cross a gap such that >50% of its translation is X (i.e., in the gap).
Annotation details – watch out for..

• No introns < 10 nt
• No really short proteins unless well supported
• Protein names should use UniProt protein naming guidelines
  http://www.uniprot.org/docs/nameprot
  – OK: Uncharacterized protein
  – NOT OK: Hypothetical (or) Possible (or) Conserved protein
  – OK: serine/threonine-protein kinase-like
• Sort names into:
  – Primary gene symbol (gene feature/gene qualifier)
    • Other gene symbols (gene feature/gene_syn qualifier)
  – Primary gene full name (gene feature/gene_desc qualifier)
  – Product name (CDS feature/product qualifier)
    • Other names (additional CDS/product)

• Example:
  – Gene feature/gene qualifier = eve
    • gene/gene_syn = CG2328, l(2)46Ce
  – gene/gene_desc = even skipped
  – CDS/product = segmentation protein even-skipped
    • lethal(2)46Ce, CG2328-PA
• More Rules on names and symbols
  – Don’t start with a species prefix
  – Don’t apply a species-specific arbitrary name or locus_tag to another species
Other Challenges for Annotations

• Non-coding RNAs, pseudogenes
  – Many annotation pipelines are optimized to predict protein-coding gene models
  – It is useful to annotate known functional RNAs or other categories of non-protein-coding loci (tRNAs, ncRNAs, rRNAs, pseudogenes)
  – Most genome annotations underrepresent ncRNAs

• Long term maintenance
  – Most annotations are DRAFT quality, and need to be updated as more data and better algorithms become available
Feature table

- First line
  >Features SeqID table_name

- The SeqID must be the same as the sequence's SeqID in the FASTA file. The table_name is optional.

- Subsequent lines of the table list the features in tab separated columns:
  - Column 1: Start location of feature
  - Column 2: Stop location of feature
  - Column 3: Feature key
  - Column 4: Qualifier key
  - Column 5: Qualifier value

### Table view of gene with both biological name and locus_tag:

<table>
<thead>
<tr>
<th>start</th>
<th>stop</th>
<th>feature</th>
<th>qualifier</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1575</td>
<td>gene</td>
<td>gene</td>
<td>Abc5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>locus_tag</td>
<td>KCS_0001</td>
</tr>
</tbody>
</table>

### Flatfile view:

- gene

  feature

  /gene="Abc5" Qualifier=value

  /locus_tag="KCS_0001"

### Table view of gene with only locus_tag:

<table>
<thead>
<tr>
<th>start</th>
<th>stop</th>
<th>feature</th>
<th>qualifier</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1575</td>
<td>gene</td>
<td>locus_tag</td>
<td>KCS_0001</td>
</tr>
</tbody>
</table>

### Flatfile view:

- gene

  1..1575

  /locus_tag="KCS_0001"
Feature table (cont.)

- Required: all proteins must have a unique protein ID.
  - Column 1: Start location of feature
  - Column 2: Stop location of feature
  - Column 3: Feature key = CDS
  - Column 4: Qualifier key = protein_id
  - Column 5: Qualifier value = gnl|dbname|string

  
  - dbname = a unique version of your lab name
  - string = a unique protein SeqID; preferably the locus_tag

Example:

```
<table>
<thead>
<tr>
<th>start</th>
<th>stop</th>
<th>feature</th>
<th>qualifier</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1</td>
<td>&gt;1575</td>
<td>gene</td>
<td>gene</td>
<td>Abc5</td>
</tr>
<tr>
<td>1</td>
<td>1575</td>
<td>CDS</td>
<td>locus_tag</td>
<td>KCS_0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>product</td>
<td>ABC5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>protein_id</td>
<td>gnl</td>
</tr>
</tbody>
</table>
```
### .tbl example

**.tbl file with feature annotations**

<table>
<thead>
<tr>
<th>locus_tag</th>
<th>feature_type</th>
<th>feature_location</th>
<th>feature_name</th>
<th>feature_products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ngs_3038</td>
<td>gene</td>
<td>102664-100872</td>
<td>troponin isoform B</td>
<td>product isoform B</td>
</tr>
<tr>
<td></td>
<td>mRNA</td>
<td>102664-102502</td>
<td>troponin isoform A</td>
<td>product isoform A</td>
</tr>
<tr>
<td></td>
<td>CDS</td>
<td>102503-102502</td>
<td>troponin isoform B</td>
<td>product isoform B</td>
</tr>
<tr>
<td></td>
<td>mRNA</td>
<td>112616-115107</td>
<td>troponin isoform A</td>
<td>product isoform A</td>
</tr>
</tbody>
</table>

**Features:**
- **locus_tag (BioProject):** Ngs_3038, Ngs_2945
- **Feature type:** gene, mRNA, CDS
- **Feature location:** 102664-100872, 102502-102234, 102168-101261
- **Feature name:** troponin isoform B, A
- **Feature products:** encoded by transcript variant B, alternatively spliced
### .tbl example

**.tbl file with feature annotations**

<table>
<thead>
<tr>
<th>Locus Tag</th>
<th>Transcript ID</th>
<th>Product</th>
<th>Protein ID</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ngs_3038B</td>
<td>gnl</td>
<td>ncbi</td>
<td>Ngs_mrna3038B</td>
<td>troponin isoform B</td>
</tr>
<tr>
<td>Ngs_3038A</td>
<td>gnl</td>
<td>ncbi</td>
<td>Ngs_mrna3038A</td>
<td>troponin isoform A</td>
</tr>
<tr>
<td>Ngs_2945</td>
<td>gnl</td>
<td>ncbi</td>
<td>Ngs_mrna2945</td>
<td>troponin isoform A</td>
</tr>
</tbody>
</table>

**locus_tag:** feature for the gene

---

**Notes:**
- **CDS:** coding sequence
- **mRNA:** messenger RNA
- **gene:** genetic information
- **protein_id:** sequence of amino acids
- **transcript_id:** specific RNA sequence
- **product:** function or role of the protein
- **note:** additional information about the transcript or protein
tbl2asn: make the .sqn file and validate


- Put all the files in the same directory, and run `tbl2asn`. A common command line, for multiple contigs per .fsa file, is:
  - `tbl2asn -p path_to_files -t template -a s -V v -Z discrep`

**INPUT files:**
- .tbl
- .fsa
- template

**OUTPUT files:**
- .sqn = submission file
- .val = validation report
- descrep = descrepancy report
Fix problems

Evaluating the validation output:

• The results of the validation are in the .val files (generated when -V v is included in the tbl2asn commandline).

• The results of the discrepancy report are in the file named discrep (generated when -Z discrep is included in the tbl2asn commandline).

• Make any necessary fixes to the input files and run tbl2asn again
val file

- Feature validation (integrity checks)
- ERRORS and WARNINGS reported (text file)
  - Review both categories and make corrections
  - Pay attention to counts – high counts may indicate that there is a systematic problem with your data
- Some examples
What type of problems?

- A few examples –
  ERROR: [SEQ_FEAT.InternalStop]
  ERROR: [SEQ_FEAT.NoStop]
  ERROR: [SEQ_FEAT.SeqLocOrder]
  ERROR: [SEQ_FEAT.StartCodon]
  ERROR: [SEQ_INST.BadProteinStart]
  ERROR: [SEQ_INST.StopInProtein]
  NOTE: [SEQ_FEAT.RareSpliceConsensusDonor]
  WARNING: [SEQ_FEAT.CDSmRNArange]
  WARNING: [SEQ_FEAT.CDSwithNoMRNAOverlap]
  WARNING: [SEQ_FEAT.DuplicateFeat]
  WARNING: [SEQ_FEAT.InvalidQualifierValue]
  WARNING: [SEQ_FEAT.NotSpliceConsensusAcceptor]
  WARNING: [SEQ_FEAT.NotSpliceConsensusDonor]
  WARNING: [SEQ_FEAT.PartialProblem]
  WARNING: [SEQ_INST.InternalNsInSeqRaw]
.val ERRORS (bad stuff; must fix)

- grep ERROR test.311.val | more
  - ERROR: valid [SEQ_FEAT.StartCodon] Illegal start codon used. Wrong genetic code [1] or protein should be partial FEATURE: CDS: GME3_g [lcl|scaffold1:c2245-<2233] [lcl|scaffold1: raw, dna len=11904] -> [lcl|GME3_g]

  - ERROR: valid [SEQ_INST.StopInProtein] [2] termination symbols in protein sequence (Gme5_g - GME5_g) BIOSEQ: lcl|GME5_g: raw, aa len=10

.val WARNINGS (some are bad)

• > grep WARNING test.311.val | more
  – WARNING: valid [SEQ_FEAT.CDSwithNoMRNAOverlap] 26 out of 39 CDSs overlapped by 0 mRNAs BIOSEQ: lcl|scaffold10: raw, dna len= 77144
.tbl -> .val connection

- [SEQ_FEAT.CDSmRNARange] mRNA contains CDS but internal intron-exon boundaries do not match

FEATURE: CDS: GME10_g <- protein ID from .tbl file
Discrep file

- Summarizes input files (name, sequence length)
  test.311:GME29_g (length 473)
  test.311:GME30_g (length 36)

- Feature annotation statistics
  DiscRep_ALL:DISC_FEATURE_COUNT::gene: 43 present
  DiscRep_ALL:DISC_FEATURE_COUNT::CDS: 43 present
  DiscRep_ALL:DISC_FEATURE_COUNT::mRNA: 43 present
  DiscRep_ALL:DISC_FEATURE_COUNT::tRNA: 2 present
  DiscRep_ALL:DISC_FEATURE_COUNT::rRNA: 1 present

- Reports other errors, some not reported in .val
  - MISSING_PROTEIN_ID: 43 proteins have invalid IDs.
  - MISSING_GENES: 3 features have no genes.
  - SUSPECT_PRODUCT_NAMES: 43 product_names contain suspect phrase or characters
    - 43 product names contain '_'
  - MISSING_GENOMEASSEMBLY_COMMENTS: 2 bioseqs are missing GenomeAssembly structured comments
  - DUPLICATE_LOCUS_TAGS: 2 genes have duplicate locus tags
Descrep example: DUPLICATE_LOCUS_TAGS

- The discrepancy report tells you explicitly about the duplicate locus_tags:

- Summary
  - DUPLICATE_LOCUS_TAGS: 2 genes have duplicate locus tags.

- After the Summary you get the details:
  - DiscRep_SUB:DUPLICATE_LOCUS_TAGS:: 2 genes have locus tag Gme_016.
  - testing:Gene Gme_016 lcl|scaffold1:c>475<395 Gme_0016
  - testing:Gene Gme_016 lcl|scaffold1:<1333>2891 Gme_0016
AGP file

- AGP files provide the ordering and orientation information to construct supercontigs or scaffolds from contigs, or to construct chromosomes from supercontigs and/or contigs.

AGP example

- Columns
  1. Object ID
  2. Object ‘From’
  3. Object ‘To’
  4. Part number
  5. Component type (sequence status)
  6. Component ID or gap length
  7. Component ‘from’ or gap type
  8. Component ‘to’ or linkage
  9. Orientation

```plaintext
# ORGANISM: Homo sapiens
# TAX_ID: 9606
# ASSEMBLY NAME: EG1
# ASSEMBLY DATE: 06-September-2006
# GENOME CENTER: NCBI
# DESCRIPTION: Example AGP specifying the assembly of scaffolds from WGS contigs

EG1_scaffold1 1 3043 1 W AADB02037551.1 1 3043 +
EG1_scaffold2 1 40448 1 W AADB02037552.1 1 40448 +
EG1_scaffold2 40449 40548 2 N 100 fragment yes
EG1_scaffold2 40549 117529 3 W AADB02037553.1 1 76981 +
EG1_scaffold2 117530 117629 4 N 100 fragment yes
EG1_scaffold2 117630 145298 5 W AADB02037554.1 1 27669 +
EG1_scaffold2 145299 145398 6 N 100 fragment yes
```
Submit with GenomesMacroSend

- Upload the .sqn, .tbl and .agp files to GenBank with [GenomesMacroSend](#).
Notify genomes@ncbi.nlm.nih.gov

• After a successful upload, write to genomes with:
  – GDSub number, provided from GenomesMacroSend
  – BioProject Accession or ID
  – Release date (immediate vs mm/dd/yyyy)
  – Inclusion (or not) of annotation
  – Assembly metadata (name, version, method, sequence coverage, technology, brief description of base level quality determination, linkage quality method and threshold for AGP files)
How does NCBI use annotated genome submissions in other resources?

Gene Map View
More..

Gene
Map
View
More..

RefSeq

Genome Assembly & Annotation

GenBank Processing

Accessions Sequence DB
Public Access
www
Outline

• What resources represent data associated with a genome submission?
• How does NCBI use annotated genome submissions in other resources?
• What is the NCBI eukaryotic annotation pipeline and when is it used?
• How does NCBI manage genome assembly data?
• How should a GenBank submission proceed?
Acknowledgements

• Ilene Mizrachi, Karen Clark (GenBank Genomes submissions)
• Avi Kimchi (Assembly and AGP database)
• Paul Kitts, Mike DiCuccio, many more (Annotation pipeline)
• Ilene Mizrachi, Tatiana Tatusova, and more (BioProject database)