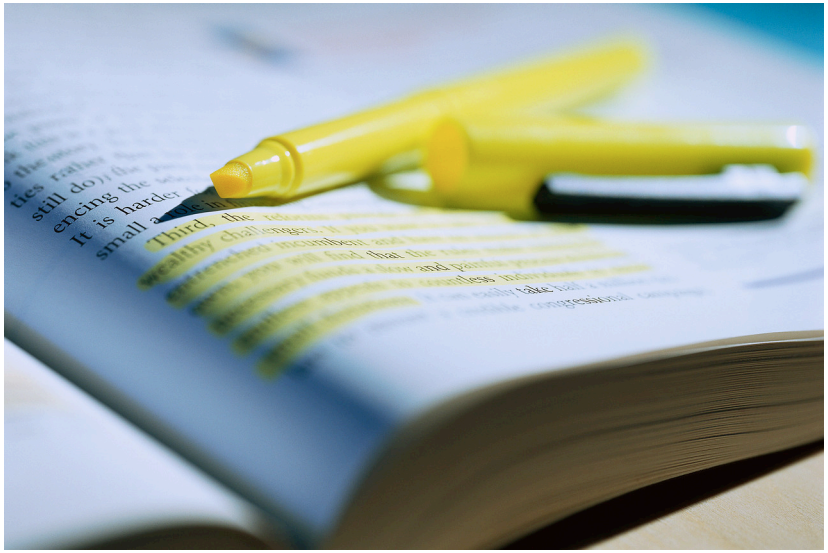
Basics of Genome Annotation

Daniel Standage
Biology Department
Indiana University

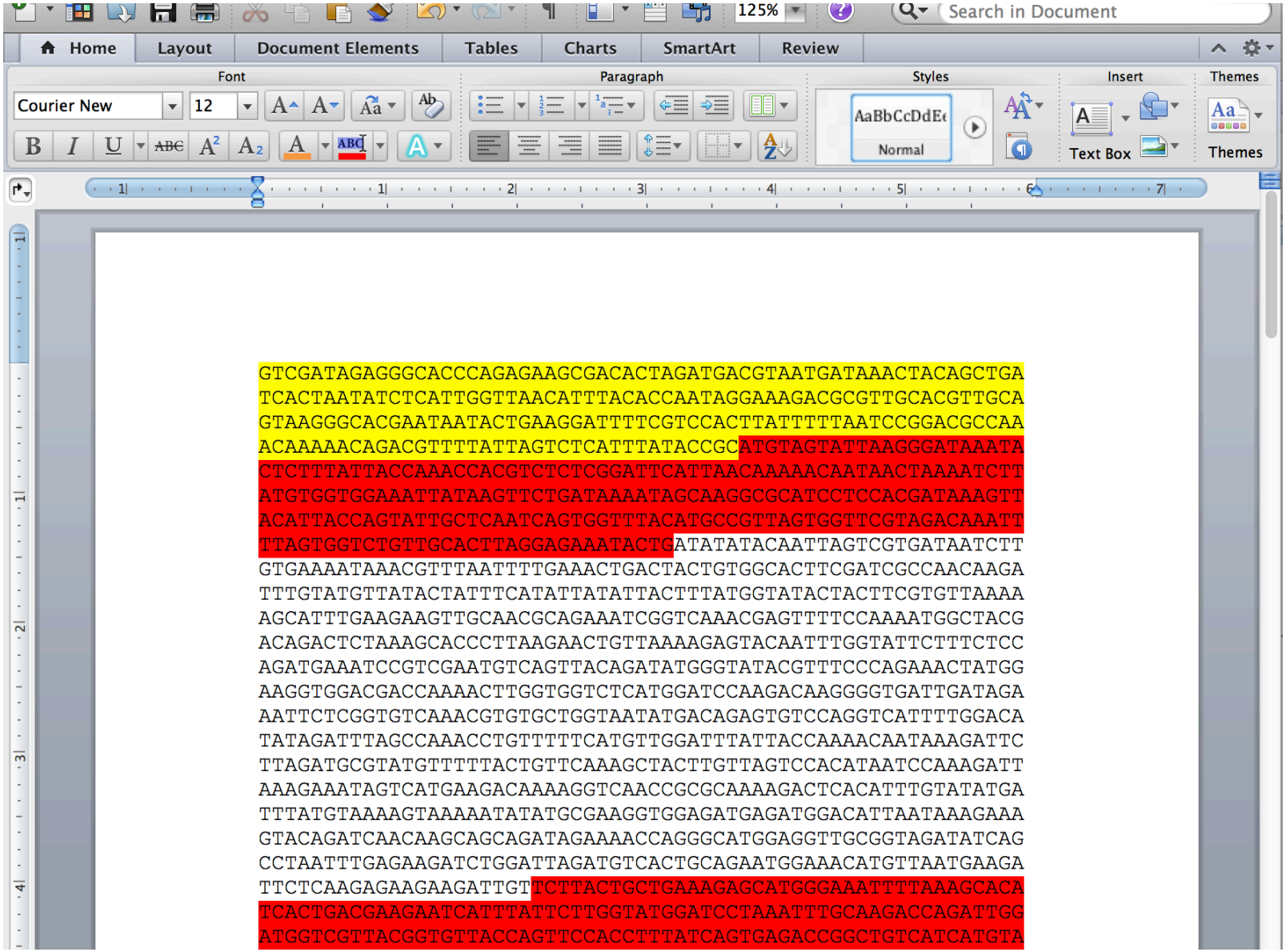
An-no-ta-tion \ˌa-nə-ˈtā-shən\

1. A critical or explanatory note or body of notes added to a text
2. The act of annotating



http://dictionary.reference.com/browse/annotation?s=t

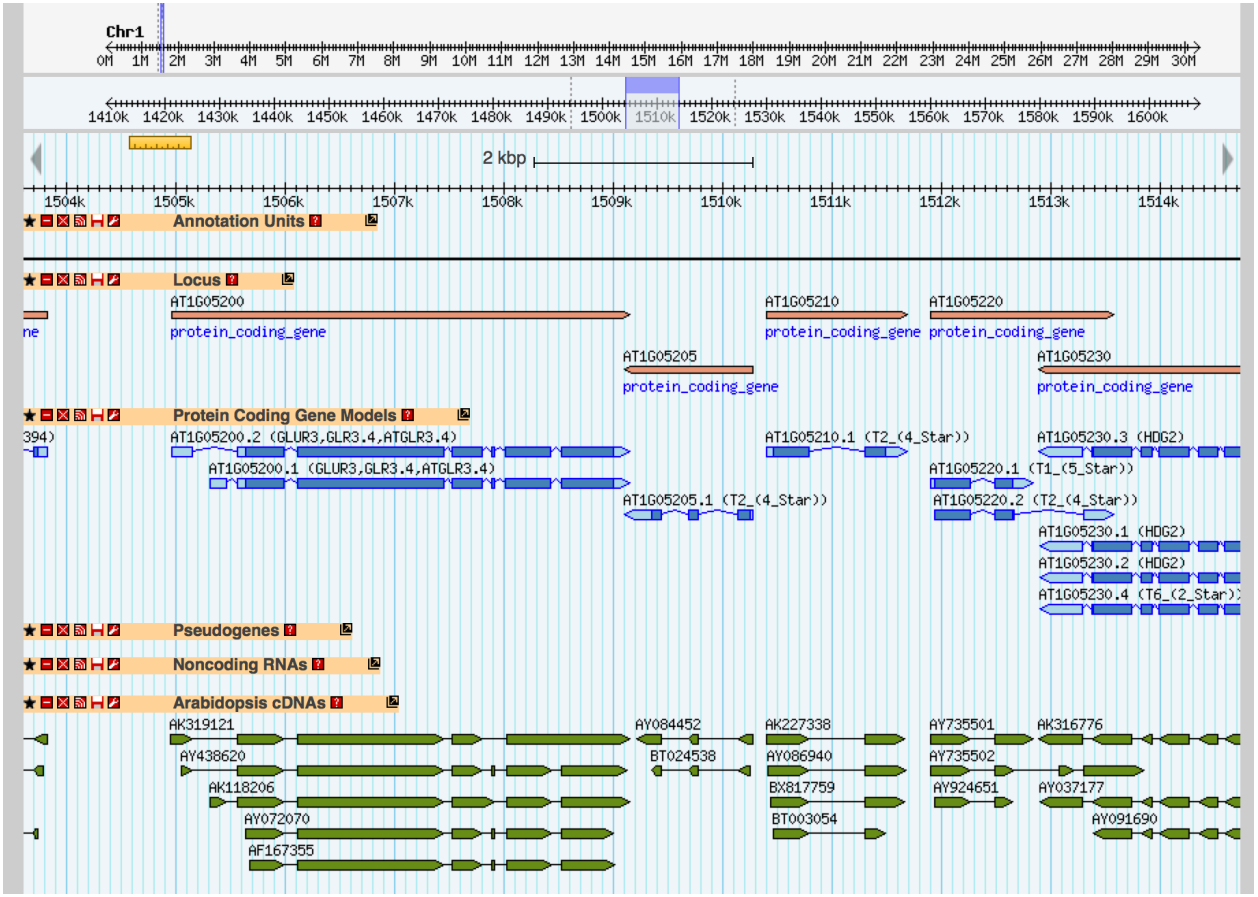
TTCACCTTT
TGGACAACA
ATGCAATC
TCCCAGTAG
CAAGAAACT
AAAGGGTCG
CAGGCCGGT
TTCTGCTAA
TTTCTTTAA
CATTCATTT


125%
Search in Document
Home
Layout
Document Elements
Tables
Charts
SmartArt
Review
Font
Courier New
12
Paragraph
Styles
AaBbCcDdEe
Normal
Insert
Text Box
Themes
Themes
GTCGATAGAGGGCACCCAGAGAAGCGACACTAGATGACGTAATGATAAACTACAGCTGA
TCACTAATATCTCATTGGTTAACATTTACACCAATAGGAAAGACGCGTTGCACGTTGCA
GTAAGGGCACGAATAATACTGAAGGATTTTCGTCCACTTATTTTTAATCCGGACGCCAA
ACAAAAACAGACGTTTTATTAGTCTCATTTATACCGCATGTAGTATTAAGGGATAAATA
CTCTTTATTACCAAACCACGTCTCTCGGATTCATTAACAAAAACAATAACTAAAATCTT
ATGTGGTGGAAATTATAAGTTCTGATAAAATAGCAAGGCGCATCCTCCACGATAAAGTT
ACATTACCAGTATTGCTCAATCAGTGGTTTACATGCCGTTAGTGGTTCGTAGACAAATT
TTAGTGGTCTGTTGCACTTAGGAGAAATACTGATATATACAATTAGTCGTGATAATCTT
GTGAAAATAAACGTTTAATTTTGAAACTGACTACTGTGGCACTTCGATCGCCAACAAGA
TTTGTATGTTATACTATTTCATATTATATTACTTTATGGTATACTACTTCGTGTTAAAA
AGCATTTGAAGAAGTTGCAACGCAGAAATCGGTCAAACGAGTTTTCCAAAATGGCTACG
ACAGACTCTAAAGCACCCTTAAGAACTGTTAAAAGAGTACAATTTGGTATTCTTTCTCC
AGATGAAATCCGTCGAATGTCAGTTACAGATATGGGTATACGTTTCCCAGAAACTATGG
AAGGTGGACGACCAAAACTTGGTGGTCTCATGGATCCAAGACAAGGGGTGATTGATAGA
AATTCTCGGTGTCAAACGTGTGCTGGTAATATGACAGAGTGTCCAGGTCATTTTGGACA
TATAGATTTAGCCAAACCTGTTTTTCATGTTGGATTTATTACCAAAACAATAAAGATTC
TTAGATGCGTATGTTTTTACTGTTCAAAGCTACTTGTTAGTCCACATAATCCAAAGATT
AAAGAAATAGTCATGAAGACAAAAGGTCAACCGCGCAAAAGACTCACATTTGTATATGA
TTTATGTAAAAGTAAAAATATATGCGAAGGTGGAGATGAGATGGACATTAATAAAGAAA
GTACAGATCAACAAGCAGCAGATAGAAAACCAGGGCATGGAGGTTGCGGTAGATATCAG
CCTAATTTGAGAAGATCTGGATTAGATGTCACTGCAGAATGGAAACATGTTAATGAAGA
TTCTCAAGAGAAGAAGATTGTTCTTACTGCTGAAAGAGCATGGGAAATTTTAAAGCACA
TCACTGACGAAGAATCATTTATTCTTGGTATGGATCCTAAATTTGCAAGACCAGATTGG
ATGGTCGTTACGGTGTTACCAGTTCCACCTTTATCAGTGAGACCGGCTGTCATCATGTA

Genome annotation

```
##gff-version   3
##sequence-region   chr8 1 100000
chr8    Gaze    gene    10503   11577   7.84    -       .       ID=GSVIVG01033678001;complete=1
chr8    Gaze    mRNA    10503   11577   7.84    -       .       ID=GSVIVT01033678001;Parent=GSVIVG01033678001;complete=1
chr8    Gaze    CDS     10503   10593   -0.4    -       1       ID=GSVIVT01033678001.cds1;Parent=GSVIVT01033678001;complete=1
chr8    Gaze    exon    10503   10593   .       -       1       ID=GSVIVT01033678001.exon1;Parent=GSVIVT01033678001
chr8    .       intron  10594   10900   .       -       .       Parent=GSVIVT01033678001
chr8    Gaze    CDS     10901   11396   5.32    -       2       ID=GSVIVT01033678001.cds1;Parent=GSVIVT01033678001;complete=1
chr8    Gaze    exon    10901   11396   .       -       1       ID=GSVIVT01033678001.exon2;Parent=GSVIVT01033678001
chr8    .       intron  11397   11465   .       -       .       Parent=GSVIVT01033678001
chr8    Gaze    CDS     11466   11577   5.46    -       0       ID=GSVIVT01033678001.cds1;Parent=GSVIVT01033678001;complete=1
chr8    Gaze    exon    11466   11577   .       -       1       ID=GSVIVT01033678001.exon3;Parent=GSVIVT01033678001
###
chr8    Gaze    gene    22057   23119   54.7    +       .       ID=GSVIVG01033677001;complete=1
chr8    Gaze    mRNA    22057   23119   54.7    +       .       ID=GSVIVT01033677001;Parent=GSVIVG01033677001;complete=1
chr8    Gaze    five_prime_UTR  22057   22166   1.22    +       .       ID=GSVIVT01033677001.utr2;Parent=GSVIVT01033677001
chr8    Gaze    exon    22057   22382   .       +       1       ID=GSVIVT01033677001.exon1;Parent=GSVIVT01033677001
chr8    Gaze    CDS     22167   22382   4.64    +       0       ID=GSVIVT01033677001.cds1;Parent=GSVIVT01033677001;complete=1
chr8    .       intron  22383   22496   .       +       .       Parent=GSVIVT01033677001
chr8    Gaze    CDS     22497   22550   12.1    +       0       ID=GSVIVT01033677001.cds1;Parent=GSVIVT01033677001;complete=1
chr8    Gaze    exon    22497   22550   .       +       1       ID=GSVIVT01033677001.exon2;Parent=GSVIVT01033677001
chr8    .       intron  22551   22650   .       +       .       Parent=GSVIVT01033677001
chr8    Gaze    CDS     22651   23022   15.4    +       0       ID=GSVIVT01033677001.cds1;Parent=GSVIVT01033677001;complete=1
chr8    Gaze    exon    22651   23119   .       +       1       ID=GSVIVT01033677001.exon3;Parent=GSVIVT01033677001
chr8    Gaze    three_prime_UTR 23023   23119   0.878   +       .       ID=GSVIVT01033677001.utr1;Parent=GSVIVT01033677001
###
```

Genome annotation



Genome annotation

- Information itself (e.g., this gene encodes a cytochrome P450 protein, with exons at…)
- Annotation process (operational definition)
- Data management
  - formatting
  - storage
  - distribution
  - representation

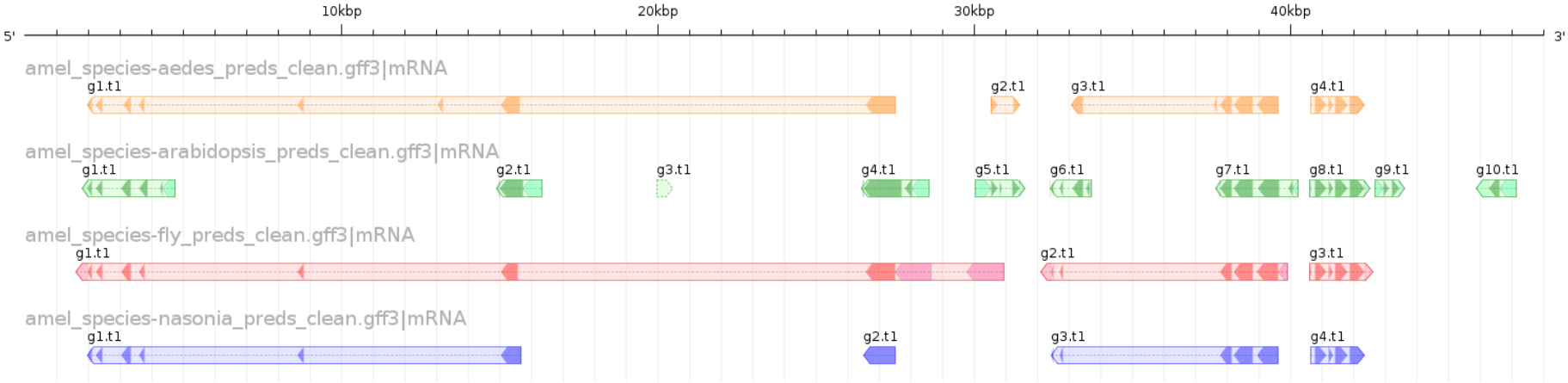
Methods for gene finding

- *Ab initio* gene prediction
- Gene prediction by spliced alignment

Ab initio gene prediction

- *Ab initio*: "from first principles"
- Requires only a genomic sequence
- Uses statistical model of genome composition to identify most probable location of start/stop codons, splice sites
- Popular implementations
  - Augustus
  - GeneMark
  - SNAP

Ab initio gene prediction



Prediction by spliced alignment

- Utilizes experimental (transcript) and/or homology (reference proteins) data
- Spliced alignment of sequences reveals gene structure
  - matches = exons
  - gaps = introns
- Popular implementations
  - GeneSeqer
  - Exonerate
  - GenomeThreader

Comparison of prediction methods

Ab initio	Spliced alignment
Do not require extrinsic evidence	Requires transcript and/or protein sequences
Does not benefit from additional transcript data	Accuracy improves with additional transcript data
More likely to recover complete gene structures	More likely to recover accurate internal exon/intron structure

Issues with gene prediction

- Accuracy (best methods achieve ≈80% at exon level)
- Parameters matter (species-specific codon usage)
- Comparison and assessment

Recurring theme in genomics

- Once I have a result, how to I assess its reliability?
- How do I compare it to alternative results?

Recurring theme in genomics

"Why, when you only had one result, did you think that was the correct one?"


Why don't we have both?

Manual annotation

- Visually inspect gene predictions, spliced alignments
- Determine reliable consensus gene structure
- Available software
  - Apollo: http://apollo.berkeleybop.org
  - yrGATE: http://goblinx.soic.indiana.edu/src/yrGATE

yrGATE-GDB001.
(NEW_ANNOT)

Annotation Class [Select...]

Project [Select...]

Working Group [Select...]

Genome Location

Genome Segment: PdomSCFr1.2-(

start: 13309 end: 21726 Change Location

Strand forward reverse strand Reset mRNA structure

User Defined Exons

20043 21226
(Gene models (Maker))

Portals and Tools

CpGAT

GeneMark GENSCAN

Genome Threader

Manual Entry

start

end add

Clear User-Defined Exons Table

mRNA (6397 nucleotides)

```
GTCGATAGAGGGCACCCAGAGAAGCGACACTAGATGACGTAATGA
TAAACTACAGCTGATCACTAATATCTCATTGGTTAACATTTACAC
CAATAGGAAAGACGCGTTGCACGTTGCAGTAAGGGCACGAATAAT
ACTGAAGGATTTTCGTCCACTTATTTTTAATCCGGACGCCAAACA
AAAACAGACGTTTTATTAGTCTCATTTATACCGCATGTAGTATTA
AGGGATAAATACTCTTTATTACCAAACCACGTCTCTCGGATTCAT
TAACAAAAACAATAACTAAAATCTTATGTGGTGGAAATTATAAGT
TCTGATAAAATAGCAAGGCGCATCCTCCACGATAAAGTTACATTA
CCAGTATTGCTCAATCAGTGGTTTACATGCCGTTAGTGGTTCGTA
GACAAATTTTAGTGGTCTGTTGCACTTAGGAGAAATACTGATATA
```

blastn blastx tblastx miRBASE

Protein Coding Region *Use the ORF finder!*

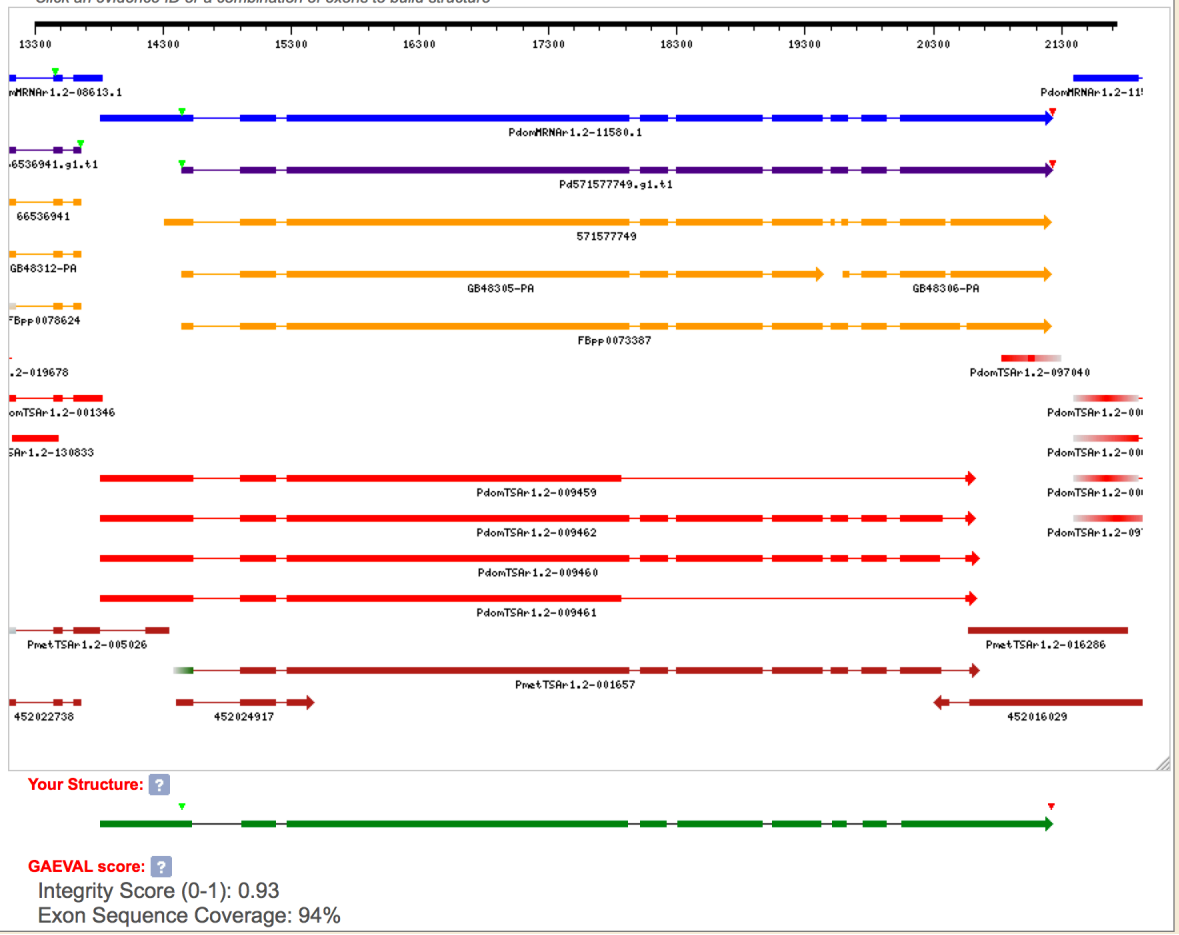
Start 14449 End 21226 ORF Finder

Protein (1918 amino acids)

```
MATTDSKAPLRTVKRVQFGILSPDEIRRMSVTDMGIRFPETMEGG
RPKLGGLMDPRQGVIDRNSRCQTCAGNMTECPGHFGHIDLAKPVF
```

Evidence Plot Change plot width to 850 pixels *Track Color Codes*

Click an evidence ID or a combination of exons to build structure



Your Structure:

GAEVAL score:

Integrity Score (0-1): 0.93

Exon Sequence Coverage: 94%

“Combiner” tools

- Maker: http://www.yandell-lab.org/software/maker.html

- EVidenceModeler: http://evidencemodeler.sourceforge.net

Evaluating annotations

- Comparison
  - ParsEval[1]: http://standage.github.io/AEGeAn
- Quality assessment
  - Annotation Edit Distance[2] (Maker)
  - GAEVAL (PlantGDB)

[1]Standage and Brendel (2012) *BMC Bioinformatics*, **13**:187.
[2]Eilbeck *et al* (2009) *BMC Bioinformatics,* **10**:67.

Recommendations / Considerations

- Automated annotation
- Manual refinement
- Assessment and filtering for particular analyses
- Be very skeptical
- Remember: no “one true” assembly / annotation

xGDBvm

- Pre-installed on iPlant cloud (free for academics!)
  - Search for xGDBvm image
- Includes an EVM pipeline for automated annotation
- Includes yrGATE for manual annotation
- Visualization, search, access control
- More info: http://goblinx.soic.indiana.edu

xGDBvm demo

Polistes dominula example