Gene Prediction

Preliminary Results

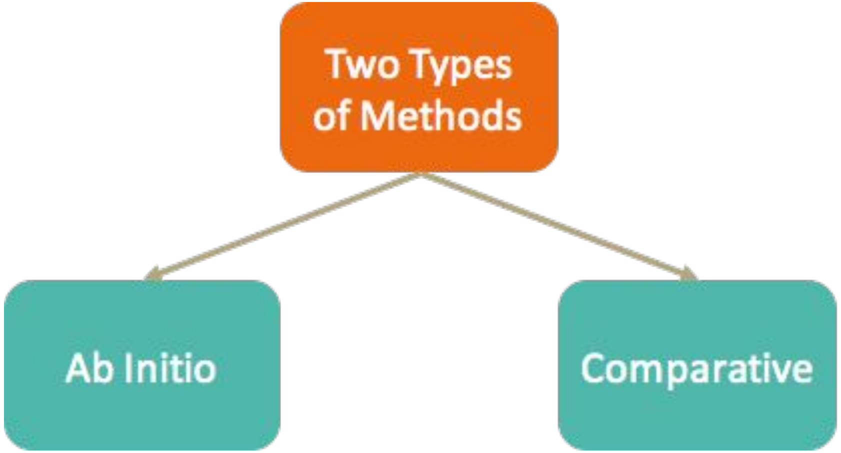
Team II

Beatriz E Saldana, Parisa Y Zowj, Ayush Semwal, Siu Lung Ng, Sini Nagpal,
Sarthak Sharma, Rong Jin, Jiani Long, Qi Zhang

Introduction - Our Plan

- Divide into three groups
  - Comparative / Similarity-Based
  - Ab Initio
  - Non Coding RNA

- Each group task:
  - Find tools
  - Test the tools
  - Compare the tools
  - Choose best tool

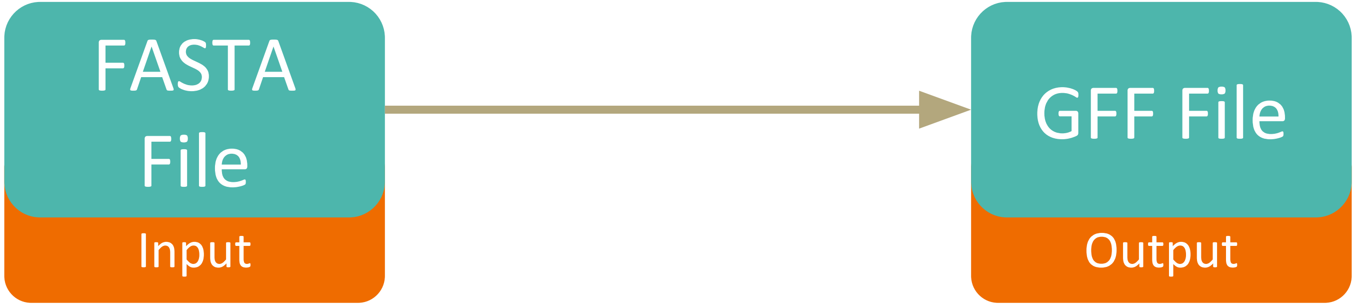


Introduction - Test Files

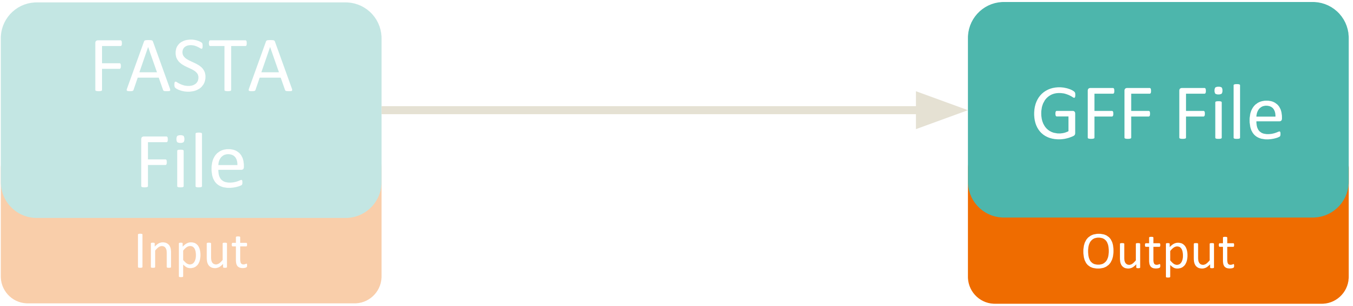
- We ran all of our tools on these files:
    - 4 assembled genomes from Genome Assembly team
    - 1 reference genome

- Why?
    - 4 Assembled genomes
        - Observe how the tools perform with our specific samples
    - Reference genome
        - Compare the output GFF file with the reference annotation file
        - GFF Compare and GFF Intersect

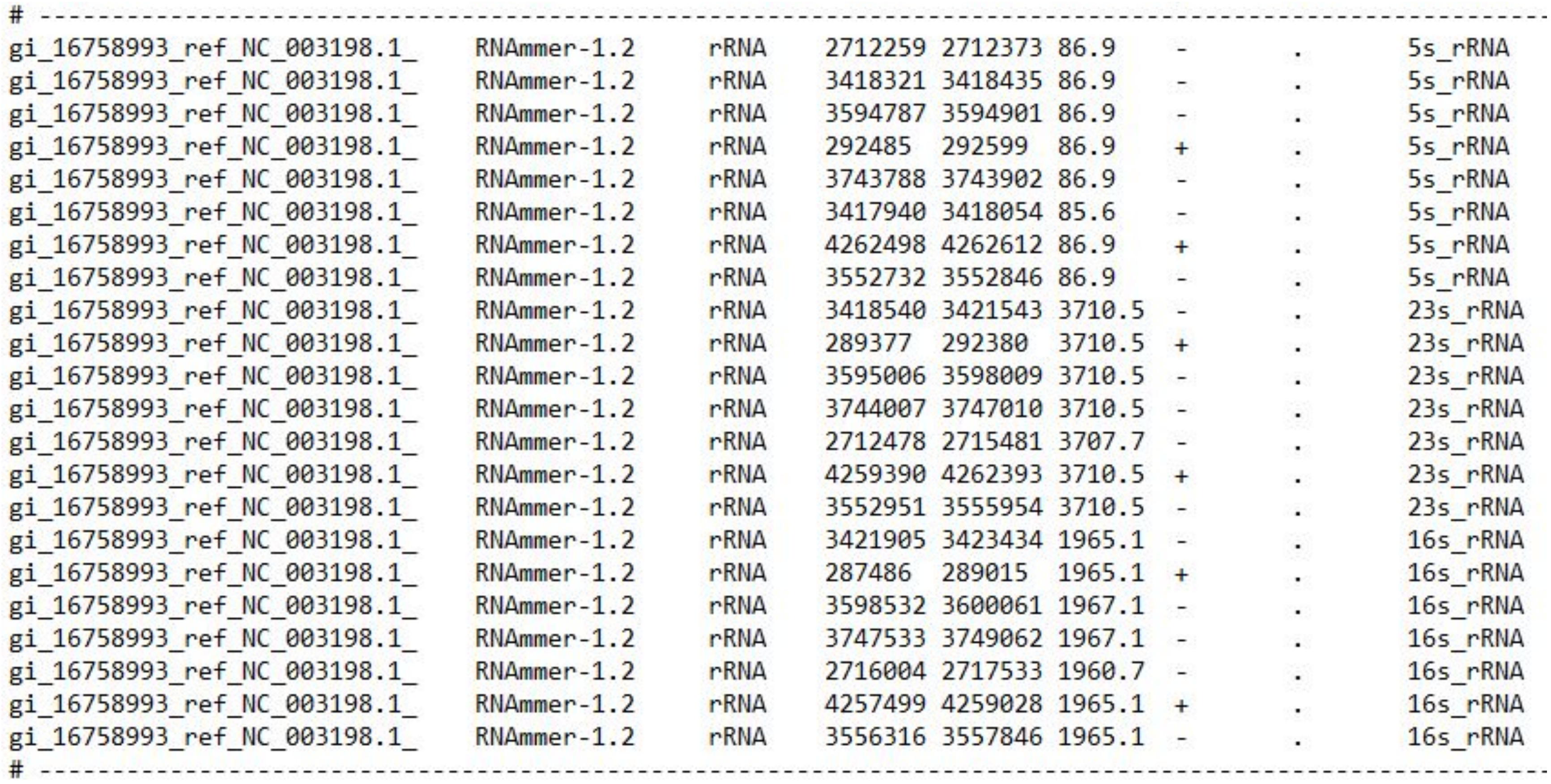
Introduction - Input and Output



Introduction - Input and Output

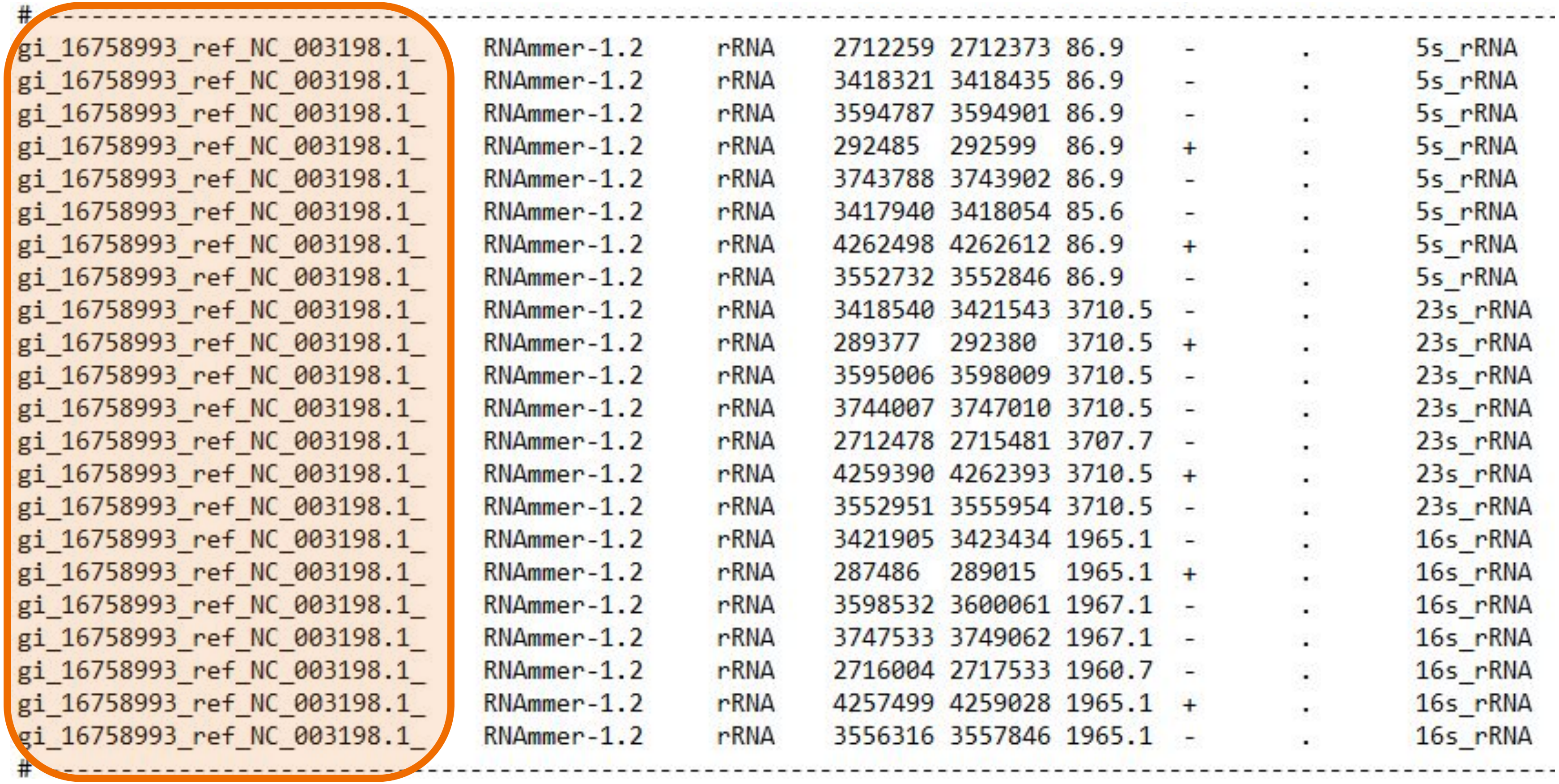


General Feature Format



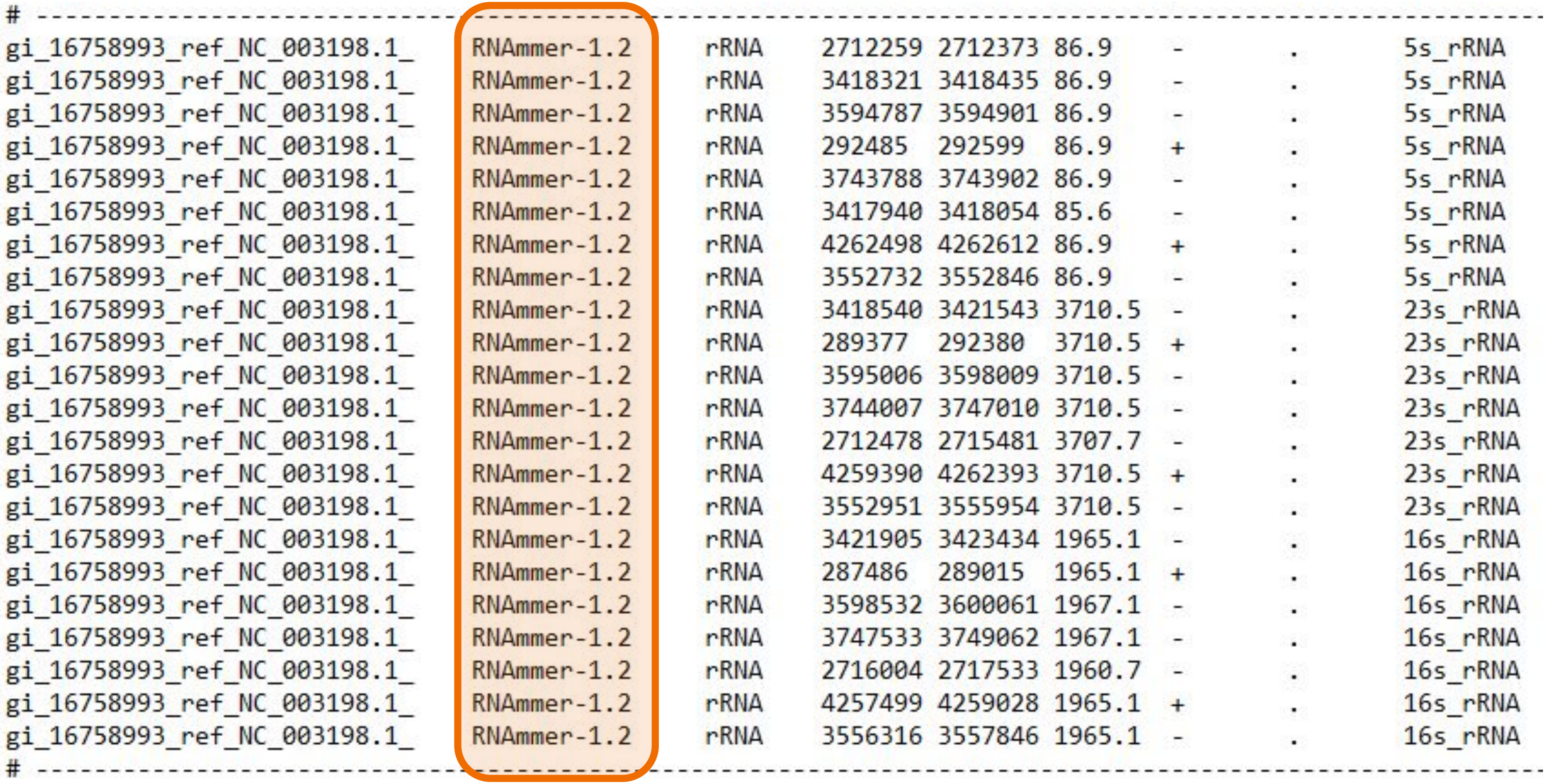
```
# ---------------------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    2712259 2712373 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3418321 3418435 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3594787 3594901 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    292485  292599  86.9    +    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3743788 3743902 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3417940 3418054 85.6    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    4262498 4262612 86.9    +    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3552732 3552846 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3418540 3421543 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    289377  292380  3710.5  +    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3595006 3598009 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3744007 3747010 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    2712478 2715481 3707.7  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    4259390 4262393 3710.5  +    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3552951 3555954 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3421905 3423434 1965.1  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    287486  289015  1965.1  +    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3598532 3600061 1967.1  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3747533 3749062 1967.1  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    2716004 2717533 1960.7  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    4257499 4259028 1965.1  +    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3556316 3557846 1965.1  -    .    16s_rRNA
# ---------------------------------------------------------------------------------------------------------------------
```

Sequence



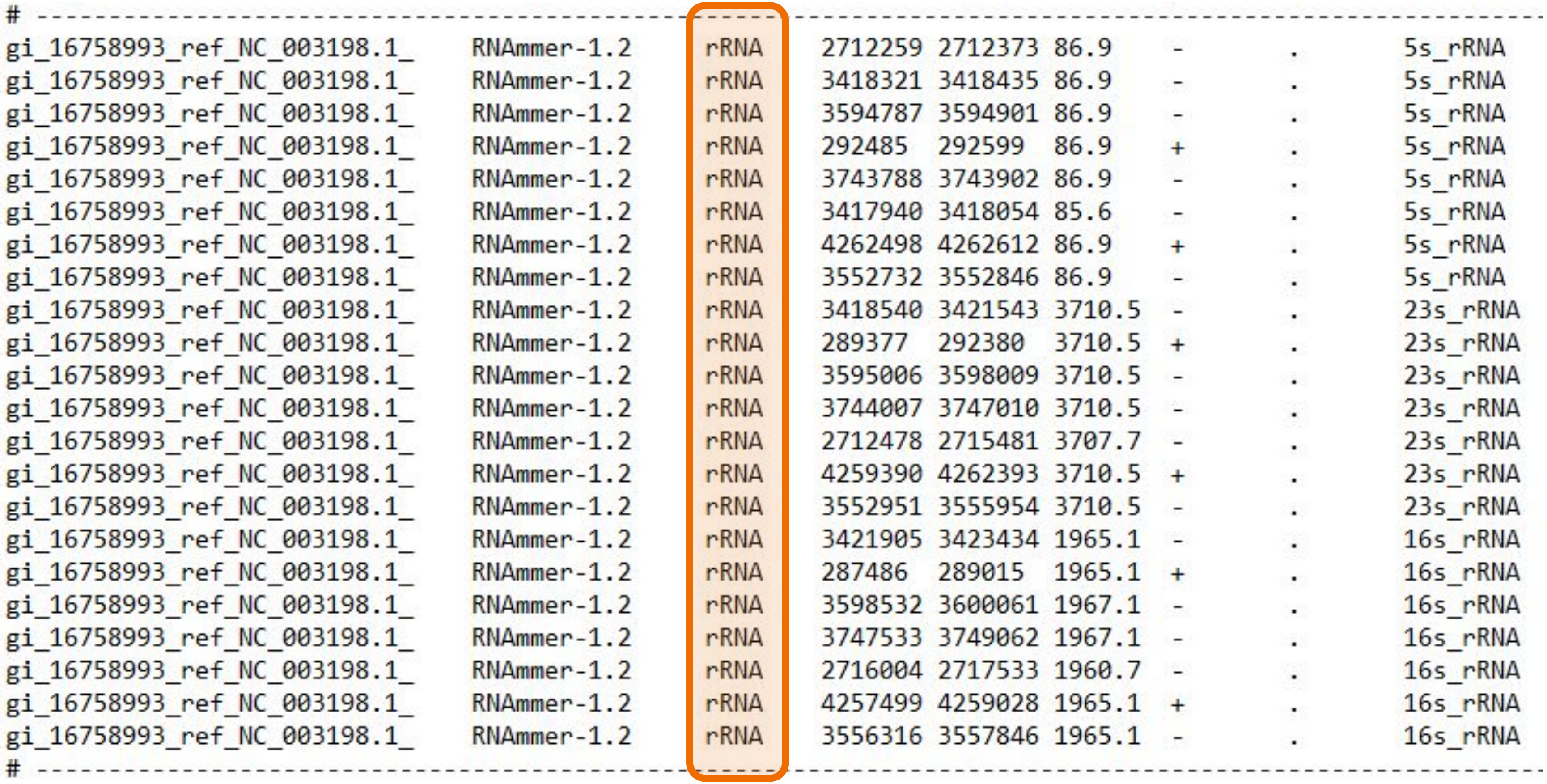
```
# ---------------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712259 2712373 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418321 3418435 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3594787 3594901 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    292485  292599  86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3743788 3743902 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3417940 3418054 85.6    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4262498 4262612 86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552732 3552846 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418540 3421543 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    289377  292380  3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3595006 3598009 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3744007 3747010 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712478 2715481 3707.7  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4259390 4262393 3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552951 3555954 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3421905 3423434 1965.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    287486  289015  1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3598532 3600061 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3747533 3749062 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2716004 2717533 1960.7  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4257499 4259028 1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3556316 3557846 1965.1  -       .       16s_rRNA
# ---------------------------------------------------------------------------------------------------------------
```

Source



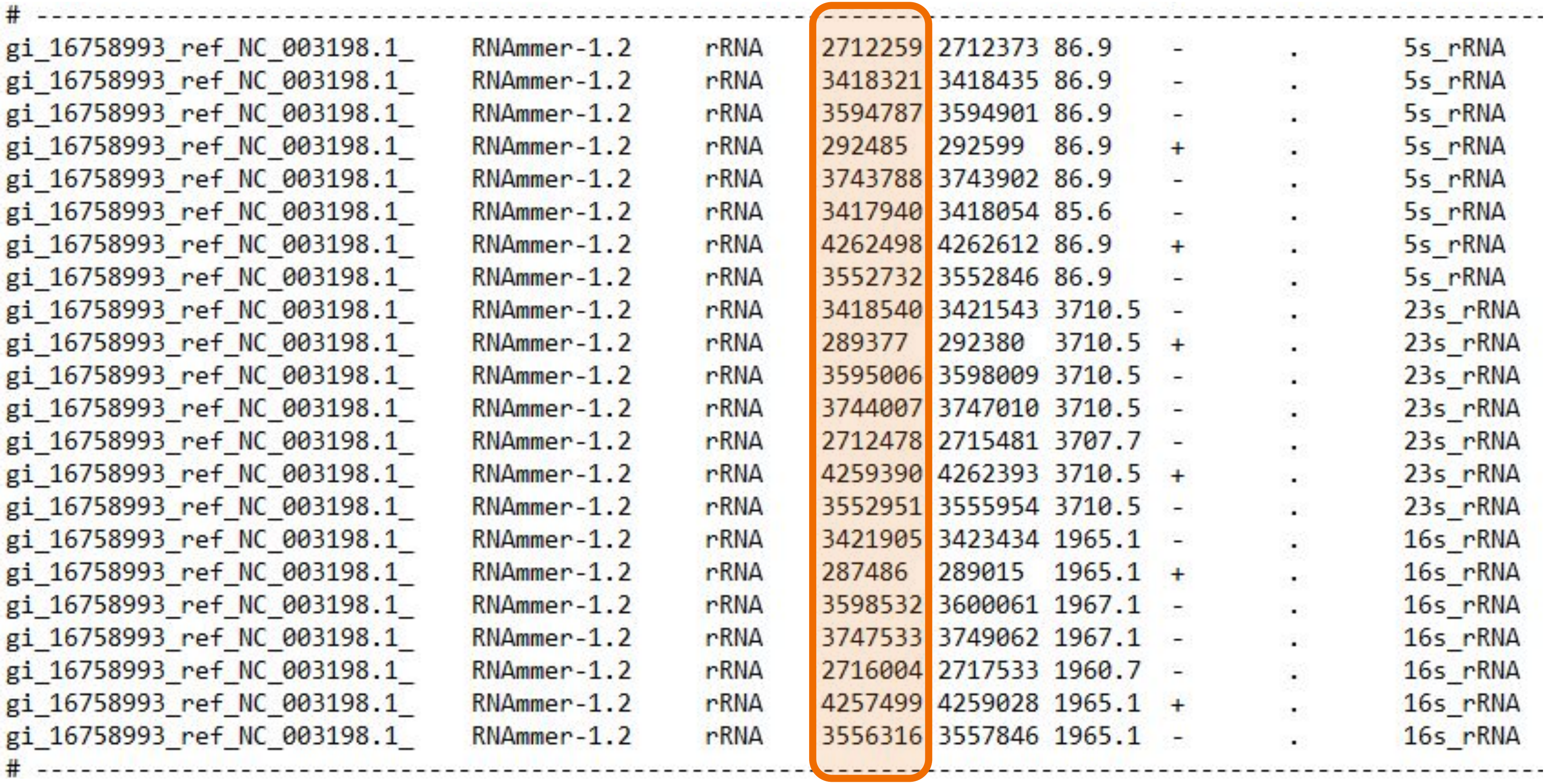
```
# ------------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   2712259 2712373 86.9     -      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3418321 3418435 86.9     -      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3594787 3594901 86.9     -      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   292485  292599  86.9     +      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3743788 3743902 86.9     -      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3417940 3418054 85.6     -      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   4262498 4262612 86.9     +      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3552732 3552846 86.9     -      .      5s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3418540 3421543 3710.5   -      .      23s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   289377  292380  3710.5   +      .      23s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3595006 3598009 3710.5   -      .      23s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3744007 3747010 3710.5   -      .      23s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   2712478 2715481 3707.7   -      .      23s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   4259390 4262393 3710.5   +      .      23s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3552951 3555954 3710.5   -      .      23s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3421905 3423434 1965.1   -      .      16s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   287486  289015  1965.1   +      .      16s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3598532 3600061 1967.1   -      .      16s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3747533 3749062 1967.1   -      .      16s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   2716004 2717533 1960.7   -      .      16s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   4257499 4259028 1965.1   +      .      16s_rRNA
gi_16758993_ref_NC_003198.1_   RNAmmer-1.2   rRNA   3556316 3557846 1965.1   -      .      16s_rRNA
# ------------------------------------------------------------------------------------------------------------
```

Feature

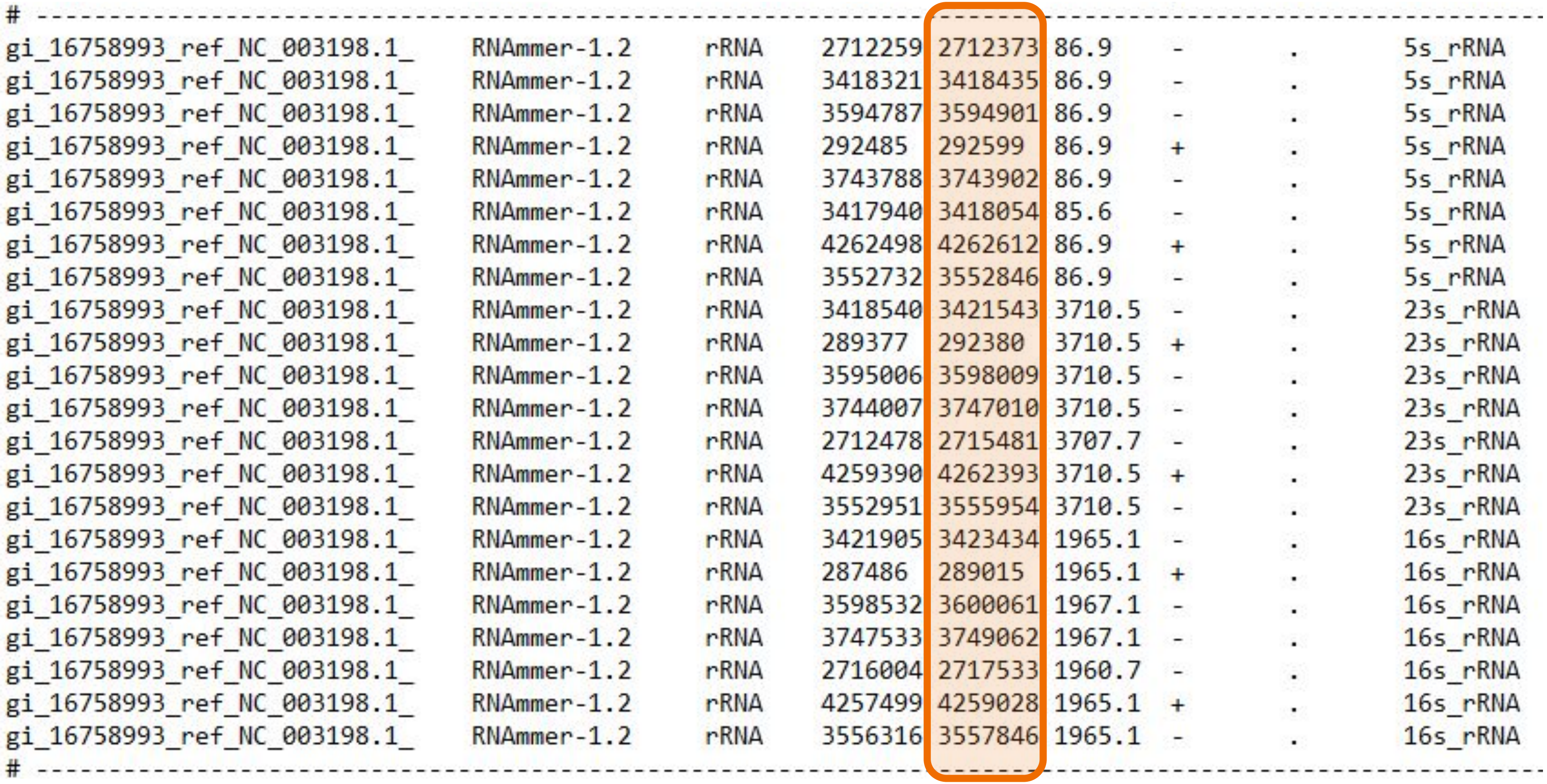


```
# ------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712259 2712373 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418321 3418435 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3594787 3594901 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    292485  292599  86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3743788 3743902 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3417940 3418054 85.6    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4262498 4262612 86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552732 3552846 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418540 3421543 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    289377  292380  3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3595006 3598009 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3744007 3747010 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712478 2715481 3707.7  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4259390 4262393 3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552951 3555954 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3421905 3423434 1965.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    287486  289015  1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3598532 3600061 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3747533 3749062 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2716004 2717533 1960.7  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4257499 4259028 1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3556316 3557846 1965.1  -       .       16s_rRNA
# ------------------------------------------------------------------------------------------
```

Feature Start

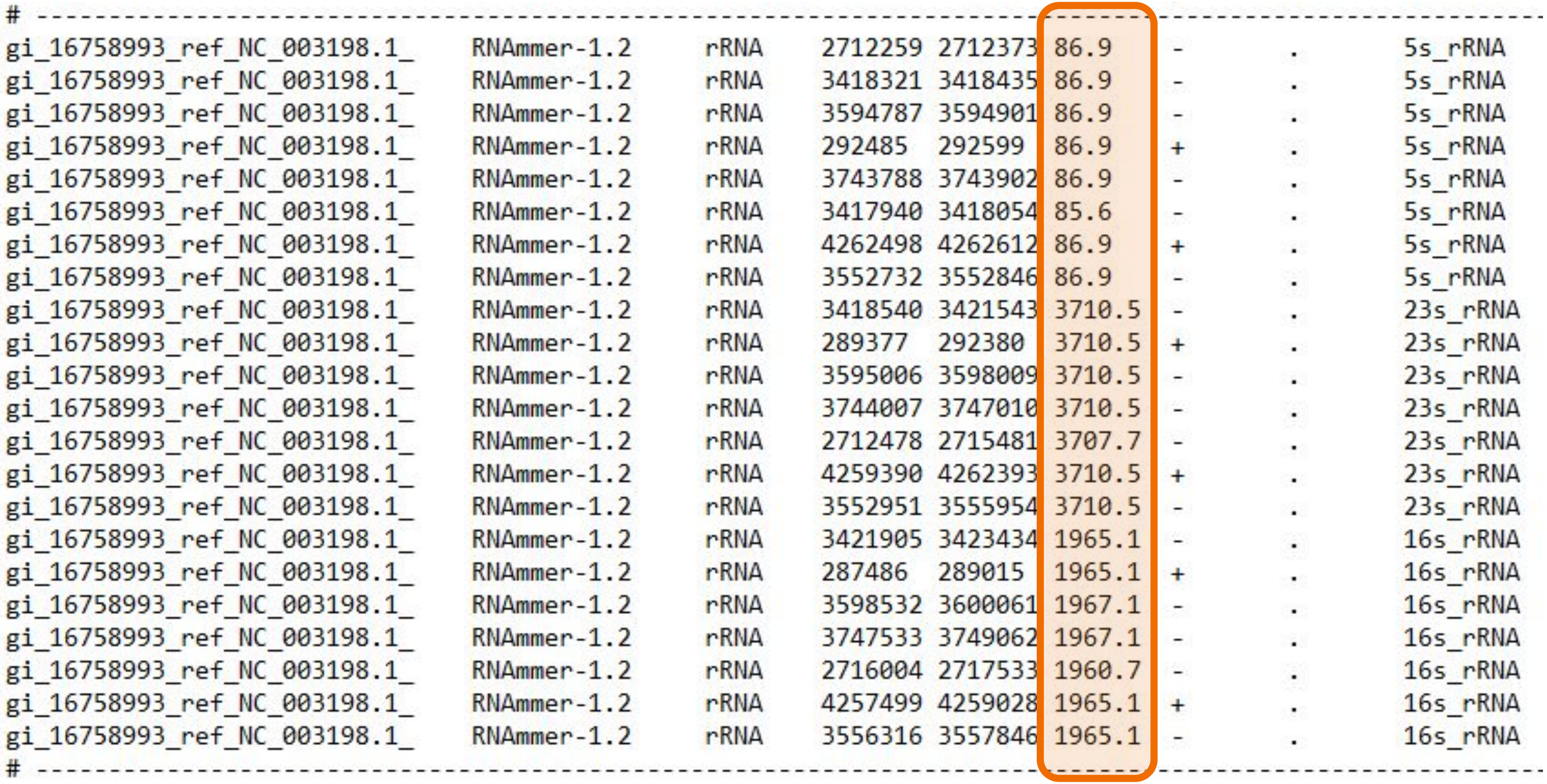


```
# ---------------------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712259 2712373 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418321 3418435 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3594787 3594901 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    292485  292599  86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3743788 3743902 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3417940 3418054 85.6    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4262498 4262612 86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552732 3552846 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418540 3421543 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    289377  292380  3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3595006 3598009 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3744007 3747010 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712478 2715481 3707.7  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4259390 4262393 3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552951 3555954 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3421905 3423434 1965.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    287486  289015  1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3598532 3600061 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3747533 3749062 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2716004 2717533 1960.7  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4257499 4259028 1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3556316 3557846 1965.1  -       .       16s_rRNA
# ---------------------------------------------------------------------------------------------------------------------
```



Feature End

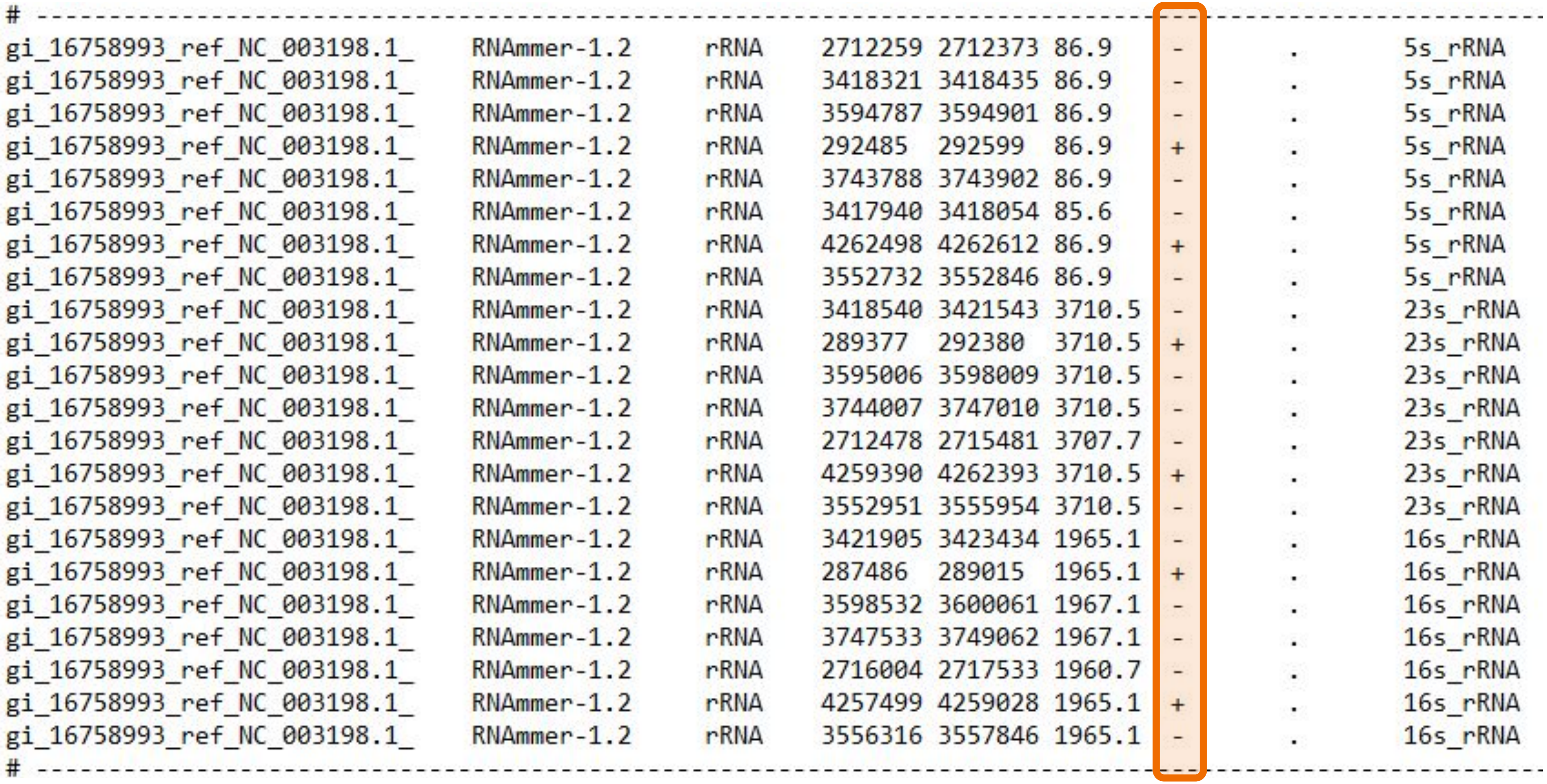
```
# ------------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    2712259 2712373 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3418321 3418435 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3594787 3594901 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    292485  292599  86.9    +    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3743788 3743902 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3417940 3418054 85.6    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    4262498 4262612 86.9    +    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3552732 3552846 86.9    -    .    5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3418540 3421543 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    289377  292380  3710.5  +    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3595006 3598009 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3744007 3747010 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    2712478 2715481 3707.7  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    4259390 4262393 3710.5  +    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3552951 3555954 3710.5  -    .    23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3421905 3423434 1965.1  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    287486  289015  1965.1  +    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3598532 3600061 1967.1  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3747533 3749062 1967.1  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    2716004 2717533 1960.7  -    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    4257499 4259028 1965.1  +    .    16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2    rRNA    3556316 3557846 1965.1  -    .    16s_rRNA
# ------------------------------------------------------------------------------------------------------------
```



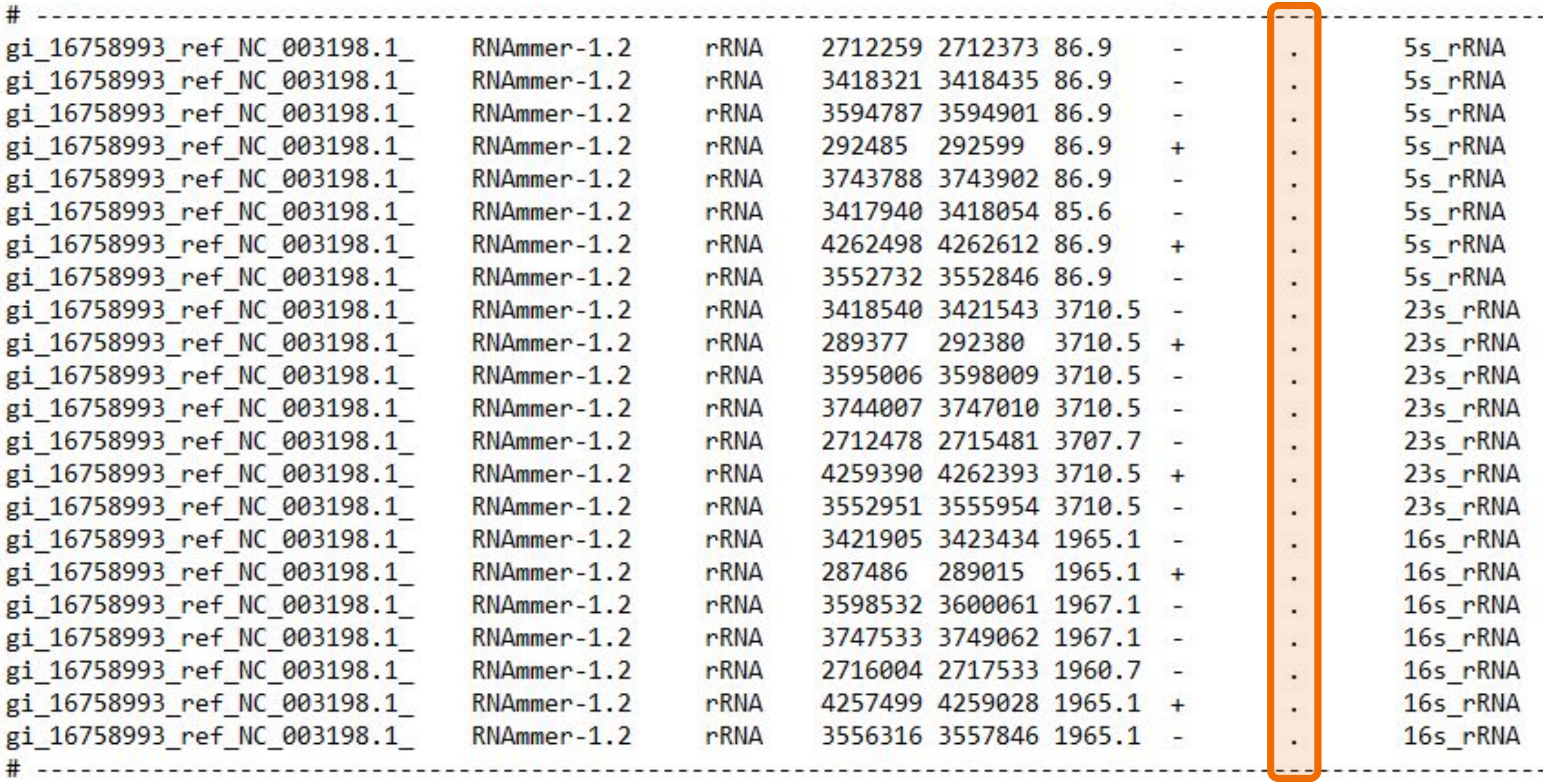
Score

```
# ---------------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712259 2712373 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418321 3418435 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3594787 3594901 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    292485  292599  86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3743788 3743902 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3417940 3418054 85.6    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4262498 4262612 86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552732 3552846 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418540 3421543 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    289377  292380  3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3595006 3598009 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3744007 3747010 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712478 2715481 3707.7  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4259390 4262393 3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552951 3555954 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3421905 3423434 1965.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    287486  289015  1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3598532 3600061 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3747533 3749062 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2716004 2717533 1960.7  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4257499 4259028 1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3556316 3557846 1965.1  -       .       16s_rRNA
# ---------------------------------------------------------------------------------------------------------------
```

Strand



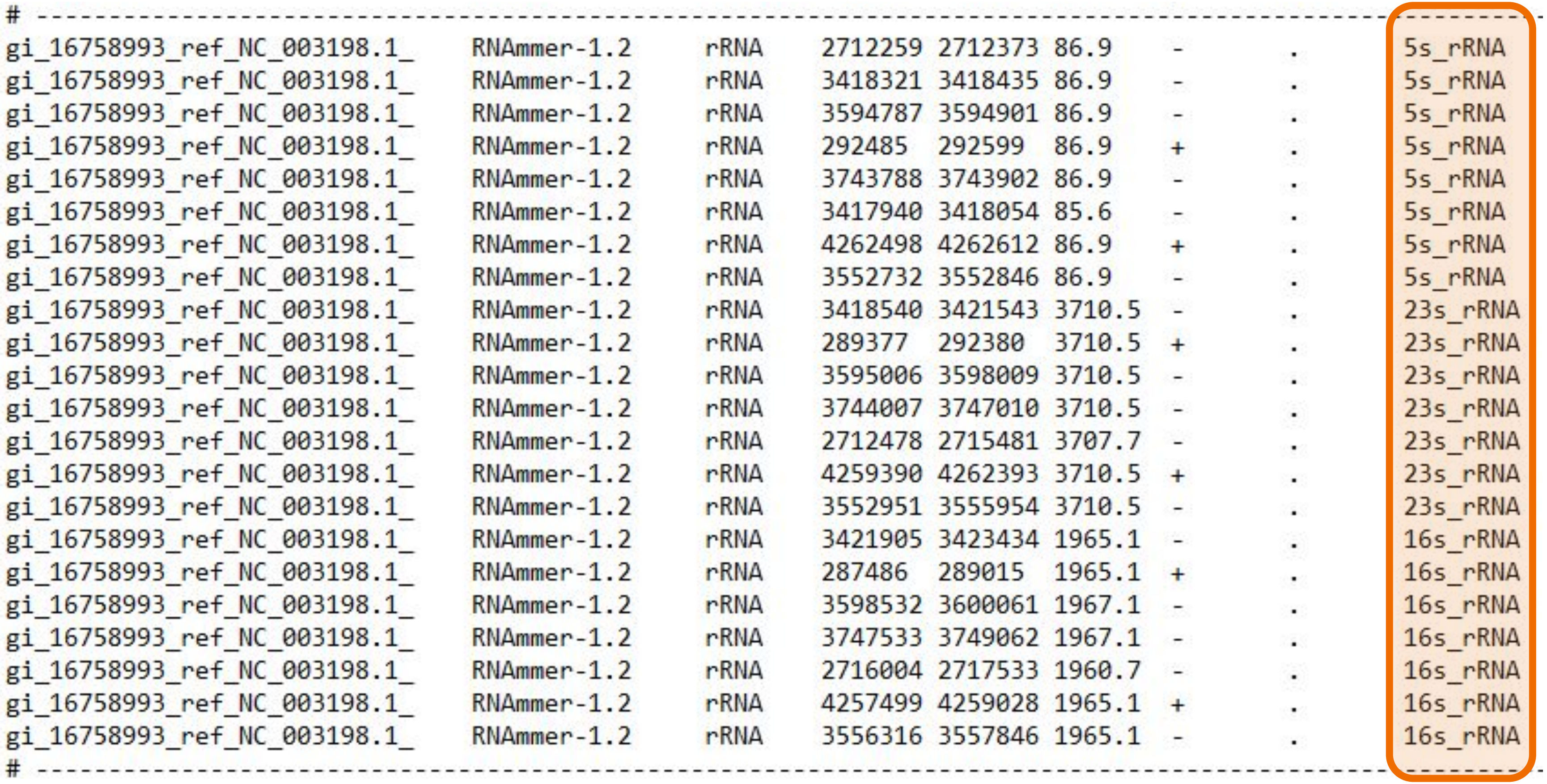
```
# --------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712259 2712373 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418321 3418435 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3594787 3594901 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    292485  292599  86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3743788 3743902 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3417940 3418054 85.6    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4262498 4262612 86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552732 3552846 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418540 3421543 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    289377  292380  3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3595006 3598009 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3744007 3747010 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712478 2715481 3707.7  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4259390 4262393 3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552951 3555954 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3421905 3423434 1965.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    287486  289015  1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3598532 3600061 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3747533 3749062 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2716004 2717533 1960.7  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4257499 4259028 1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3556316 3557846 1965.1  -       .       16s_rRNA
# --------------------------------------------------------------------------------------------------------
```



Frame/Phase

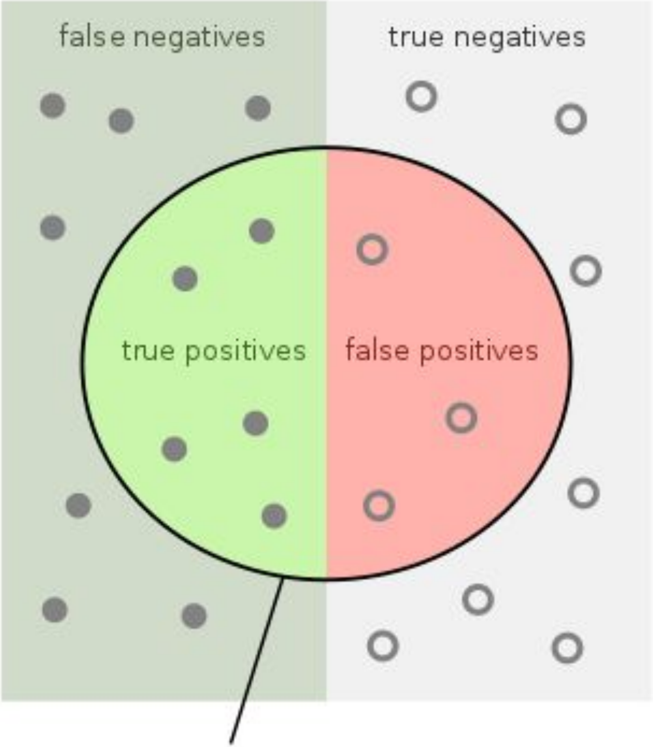
```
# -----------------------------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712259 2712373 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418321 3418435 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3594787 3594901 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    292485  292599  86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3743788 3743902 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3417940 3418054 85.6    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4262498 4262612 86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552732 3552846 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418540 3421543 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    289377  292380  3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3595006 3598009 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3744007 3747010 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712478 2715481 3707.7  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4259390 4262393 3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552951 3555954 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3421905 3423434 1965.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    287486  289015  1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3598532 3600061 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3747533 3749062 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2716004 2717533 1960.7  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4257499 4259028 1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3556316 3557846 1965.1  -       .       16s_rRNA
# -----------------------------------------------------------------------------------------------------------------------
```

Attributes



```
# ------------------------------------------------------------------------------------------------
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712259 2712373 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418321 3418435 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3594787 3594901 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    292485  292599  86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3743788 3743902 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3417940 3418054 85.6    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4262498 4262612 86.9    +       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552732 3552846 86.9    -       .       5s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3418540 3421543 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    289377  292380  3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3595006 3598009 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3744007 3747010 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2712478 2715481 3707.7  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4259390 4262393 3710.5  +       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3552951 3555954 3710.5  -       .       23s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3421905 3423434 1965.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    287486  289015  1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3598532 3600061 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3747533 3749062 1967.1  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    2716004 2717533 1960.7  -       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    4257499 4259028 1965.1  +       .       16s_rRNA
gi_16758993_ref_NC_003198.1_    RNAmmer-1.2     rRNA    3556316 3557846 1965.1  -       .       16s_rRNA
# ------------------------------------------------------------------------------------------------
```

Introduction - Comparison Metrics

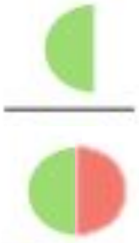


Predicted Genes

Sensitivity

$$\frac{TP}{TP + FN}$$

Precision

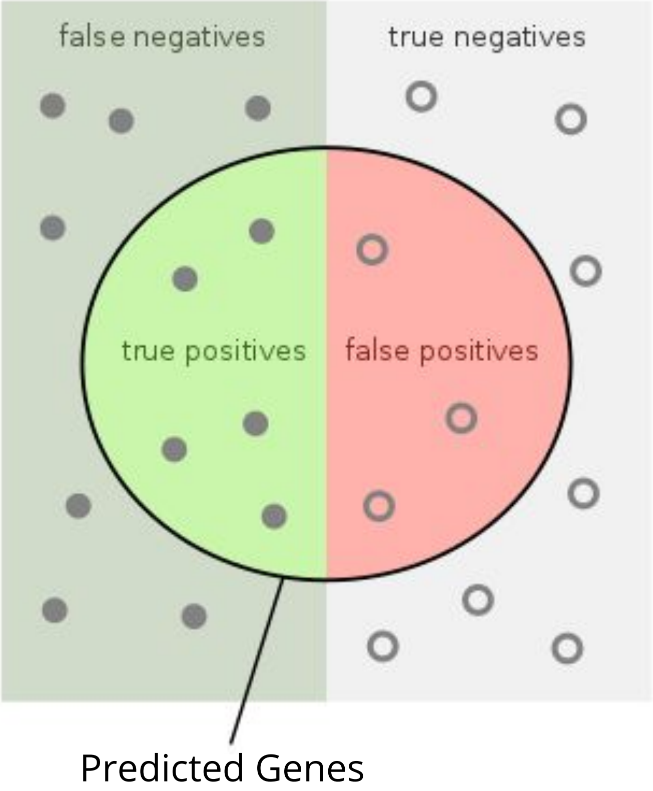


$$\frac{TP}{TP + FP}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

Introduction - Comparison Metrics



Sensitivity

$$\frac{\text{TP}}{\text{TP + FN}}$$

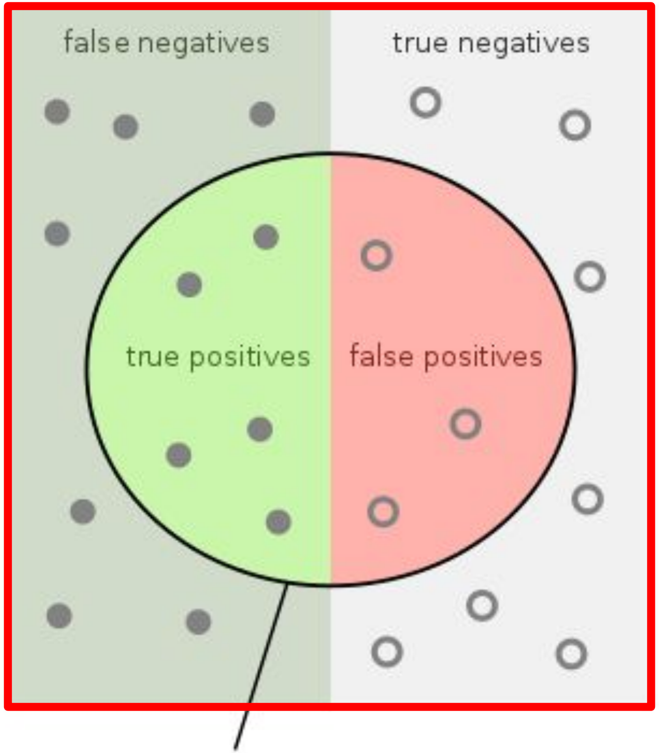
Precision

$$\frac{\text{TP}}{\text{TP + FP}}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

Introduction - Comparison Metrics



Predicted Genes

Sensitivity

$$\frac{TP}{TP + FN}$$

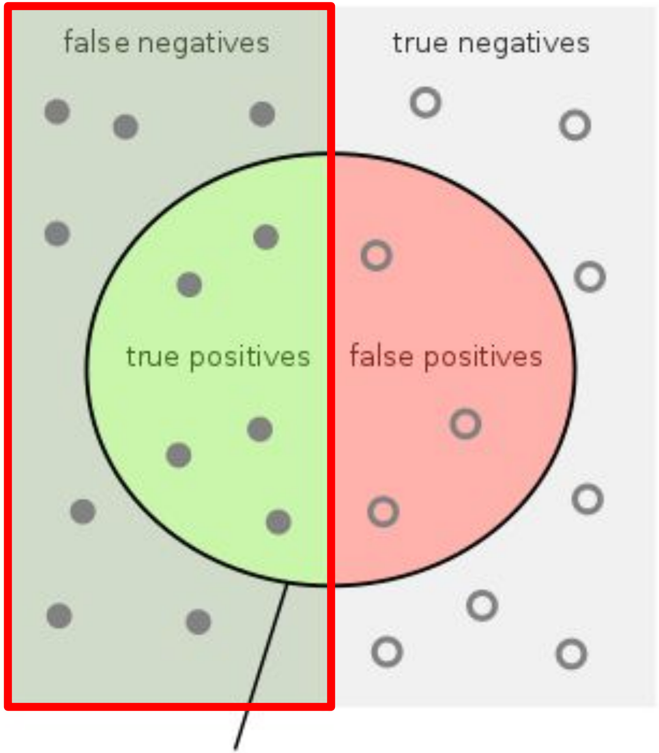
Precision

$$\frac{TP}{TP + FP}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

Introduction - Comparison Metrics



Predicted Genes

Sensitivity

$$\frac{TP}{TP + FN}$$

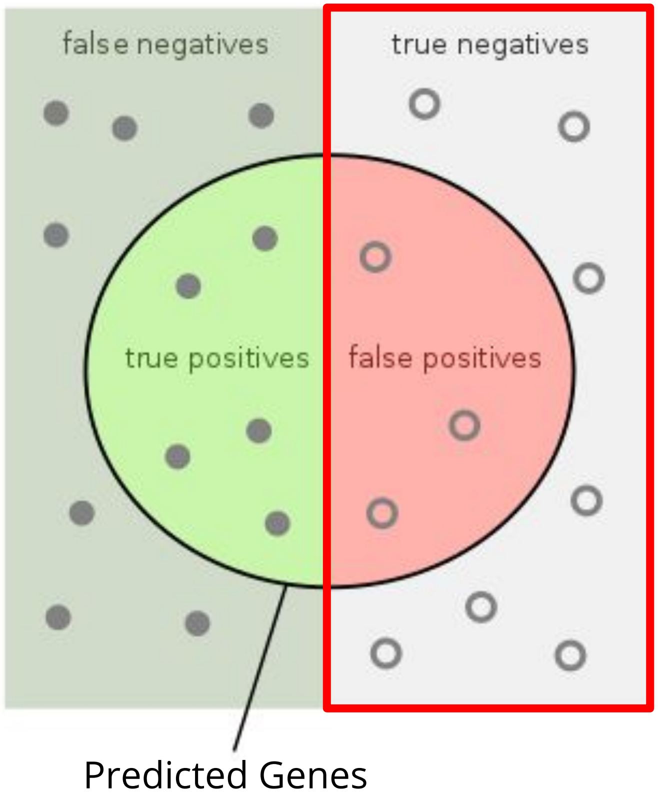
Precision

$$\frac{TP}{TP + FP}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

Introduction - Comparison Metrics



Sensitivity

$$\frac{TP}{TP + FN}$$

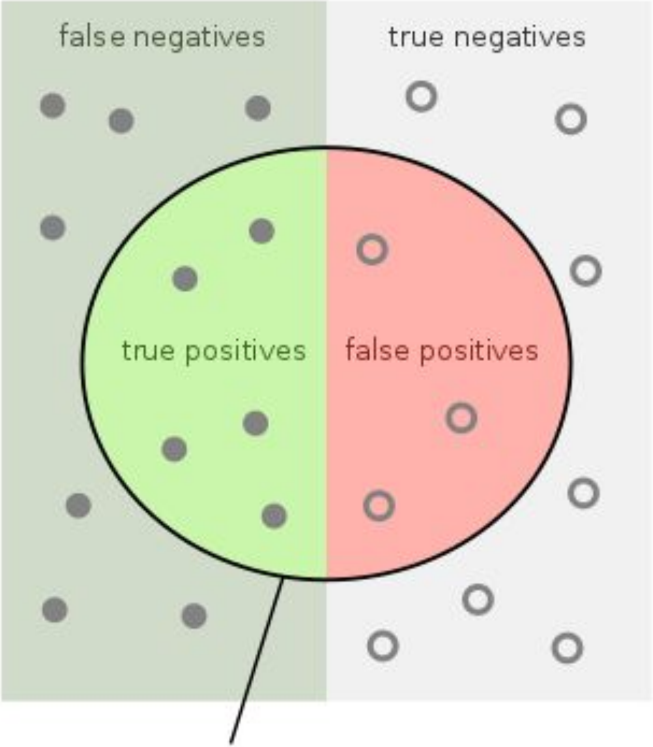
Precision

$$\frac{TP}{TP + FP}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

Introduction - Comparison Metrics



Predicted Genes

Sensitivity

$$\frac{TP}{TP + FN}$$

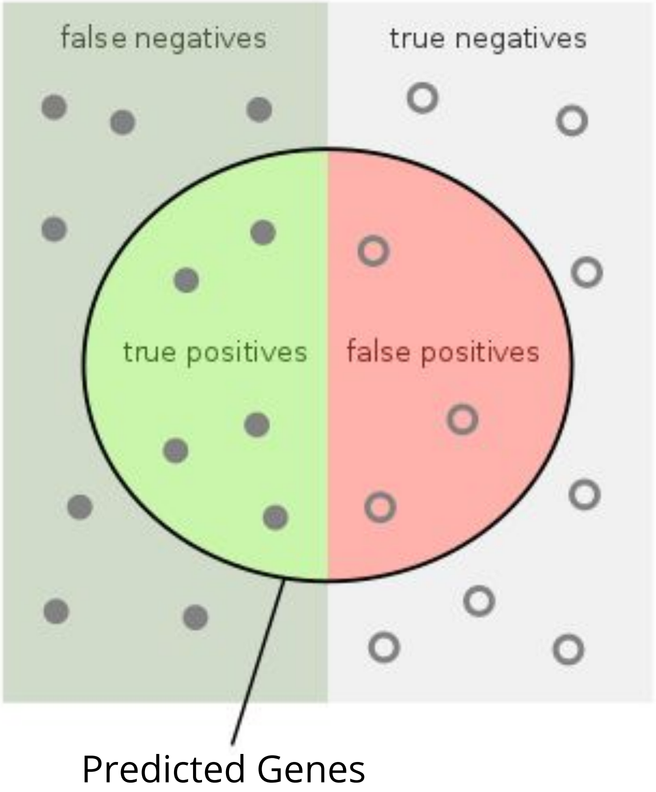
Precision

$$\frac{TP}{TP + FP}$$

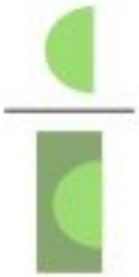
Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

Introduction - Comparison Metrics

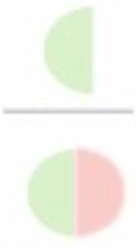


Sensitivity



$$\frac{TP}{TP + FN}$$

Precision

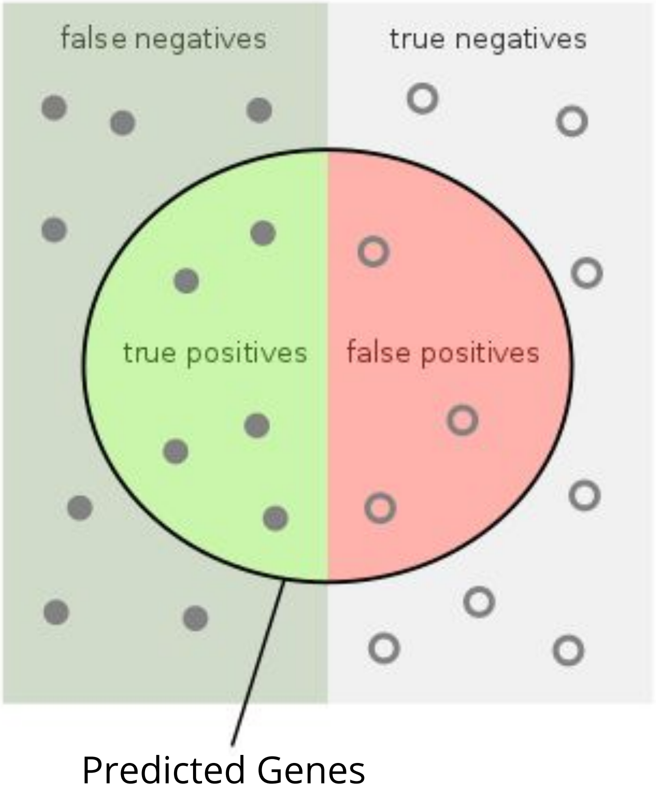


$$\frac{TP}{TP + FP}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

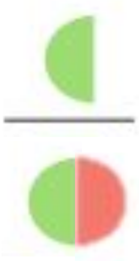
Introduction - Comparison Metrics



Sensitivity

$$\frac{TP}{TP + FN}$$

Precision

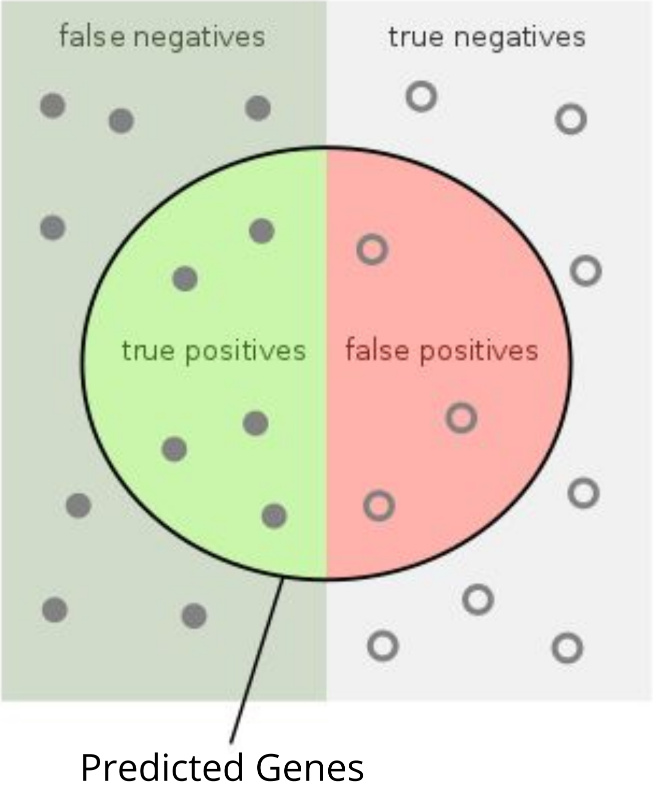


$$\frac{TP}{TP + FP}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)

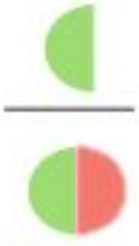
Introduction - Comparison Metrics



Sensitivity

$$\frac{TP}{TP + FN}$$

Precision



$$\frac{TP}{TP + FP}$$

Other Important Metrics:

- Run time per genome
- Installation complexity
- Use of storage space (limited storage)
- Output file format

Comparative Approach

Preliminary Results

Comparative - BLAST+ Command Line

From Genome Assembly Group

- **Reference Genome :** Klebsiella Pneumonia from NCBI
  - NC_016845.1
  - Size 5.33 Mb
  - 57.5 GC %

- **Query Genomes :** 4 Assembled genome FASTA files

Comparative - BLAST+ Command Line

Steps

1. Make Blast database of the reference genome

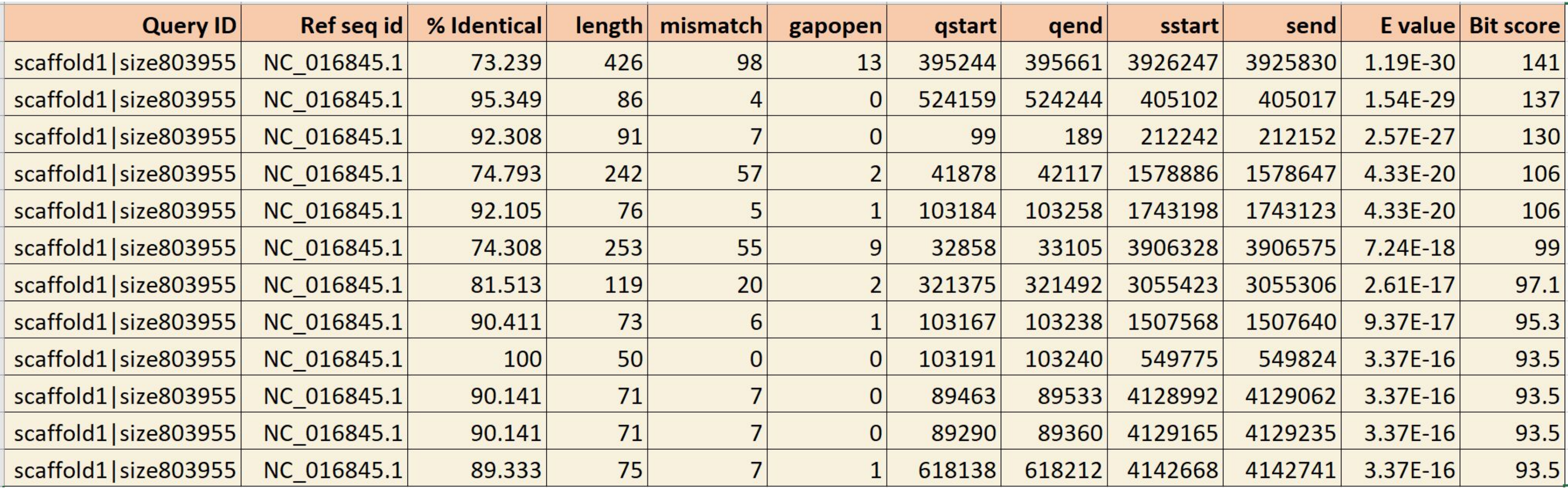
```
makeblastdb -in reference_genomic.fna -dbtype nucl -out
k_pneomoniae_database
```

2. Query assembled genomes against the database

```
blastn -db k_pneomoniae_database -query
assembled_genome.fasta -outfmt 6 -out predicted_genes.gff
```

Comparative - BLAST+ Command Line

Blast Output GFF



Query ID	Ref seq id	% Identical	length	mismatch	gapopen	qstart	qend	sstart	send	E value	Bit score
scaffold1|size803955	NC_016845.1	73.239	426	98	13	395244	395661	3926247	3925830	1.19E-30	141
scaffold1|size803955	NC_016845.1	95.349	86	4	0	524159	524244	405102	405017	1.54E-29	137
scaffold1|size803955	NC_016845.1	92.308	91	7	0	99	189	212242	212152	2.57E-27	130
scaffold1|size803955	NC_016845.1	74.793	242	57	2	41878	42117	1578886	1578647	4.33E-20	106
scaffold1|size803955	NC_016845.1	92.105	76	5	1	103184	103258	1743198	1743123	4.33E-20	106
scaffold1|size803955	NC_016845.1	74.308	253	55	9	32858	33105	3906328	3906575	7.24E-18	99
scaffold1|size803955	NC_016845.1	81.513	119	20	2	321375	321492	3055423	3055306	2.61E-17	97.1
scaffold1|size803955	NC_016845.1	90.411	73	6	1	103167	103238	1507568	1507640	9.37E-17	95.3
scaffold1|size803955	NC_016845.1	100	50	0	0	103191	103240	549775	549824	3.37E-16	93.5
scaffold1|size803955	NC_016845.1	90.141	71	7	0	89463	89533	4128992	4129062	3.37E-16	93.5
scaffold1|size803955	NC_016845.1	90.141	71	7	0	89290	89360	4129165	4129235	3.37E-16	93.5
scaffold1|size803955	NC_016845.1	89.333	75	7	1	618138	618212	4142668	4142741	3.37E-16	93.5

Comparative - BLAST+ Command Line

- **Blast+ command line** is **pretty fast** and **easy to install and use**
- Handles **large files**
- Only **disadvantage is that it is REFERENCE DEPENDENT and so it FINDS the known genes rather than predicting new ones.**

Time	~30 sec (FATSA file size ~5642KB)
Sensitivity	90.6%
Precision	92.3%

Ab Initio

Tools & Preliminary Results

Ab-Initio - Tools

- GLIMMER
- GeneMark-S
- Prodigal
    - Battle-tested
    - For prokaryotes
    - Popular, and known to be accurate

Ab-Initio - Tools: GLIMMER

- Gene Locator and Interpolated Markov ModelER
- Based on Interpolated (variable-order) Markov Model
- 2-step process
  - build ICM (interpolated context model)
  - then analyze sequence, make gene predictions

-- *it gives you .predict .detail files, not GFF files (script !)*

Ab-Initio - Tools: GeneMark-S

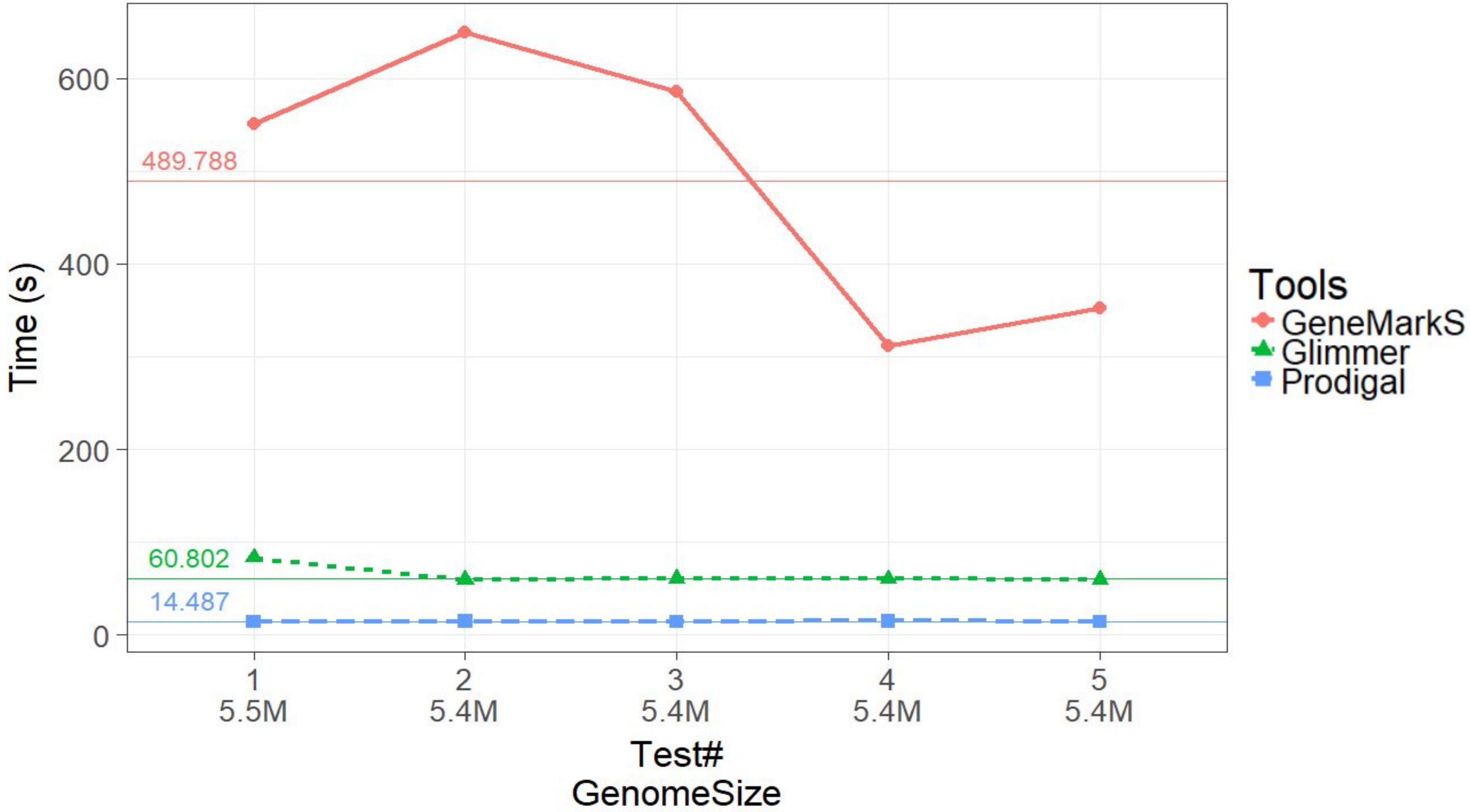
- Based on Inhomogeneous Markov Chain Model
- With heuristic models to predict genes
- Multiple output format (gff, fna,faa, etc.)
- Slow !

Ab-Initio - Tools: Prodigal

- Prokaryotic Dynamic Programming Genefinding Algorithm
- Based on Log-likelihood functions
- Works for high GC-content genomes
- Runs un-supervised
- Very fast
- Output gff, gbk, etc.

Ab-Initio - Performance Check: Run Time

Test #	Name	Size	Glimmer	GeneMarkS	Prodigal
1	GCF_000240185.1	5.5M	1m 23.353s	9m 10.761s	14.157s
2	SRR3981086	5.4M	0m 59.885s	10m 49.133s	14.818s
3	SRR3981087	5.4M	1m 0.829s	9m 45.413s	14.445s
4	SRR3982098	5.4M	1m 0.321s	5m 11.310s	14.979s
5	SRR3987120	5.5M	0m 59.622s	5m 52.320s	14.036s
	Average Per Tool		**1m 0.802s**	**8m 9.788s**	**14.487s**

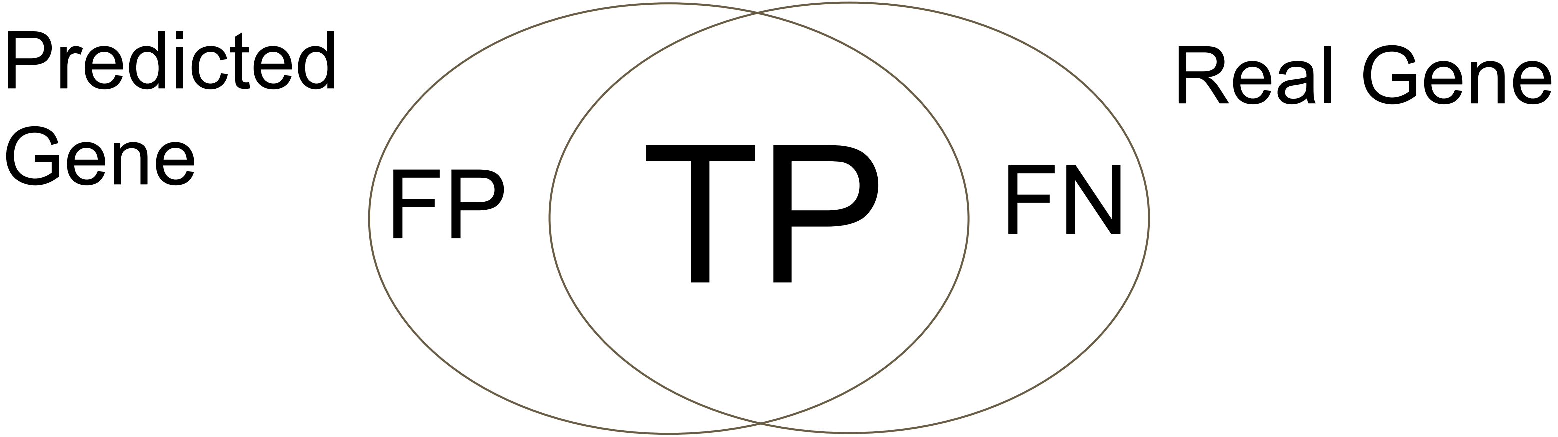

Time (s)
600
400
200
0
489.788
60.802
14.487
1
5.5M
2
5.4M
3
5.4M
4
5.4M
5
5.4M
Test#
GenomeSize
Tools
GeneMarkS
Glimmer
Prodigal

Generic Accuracy Metrics

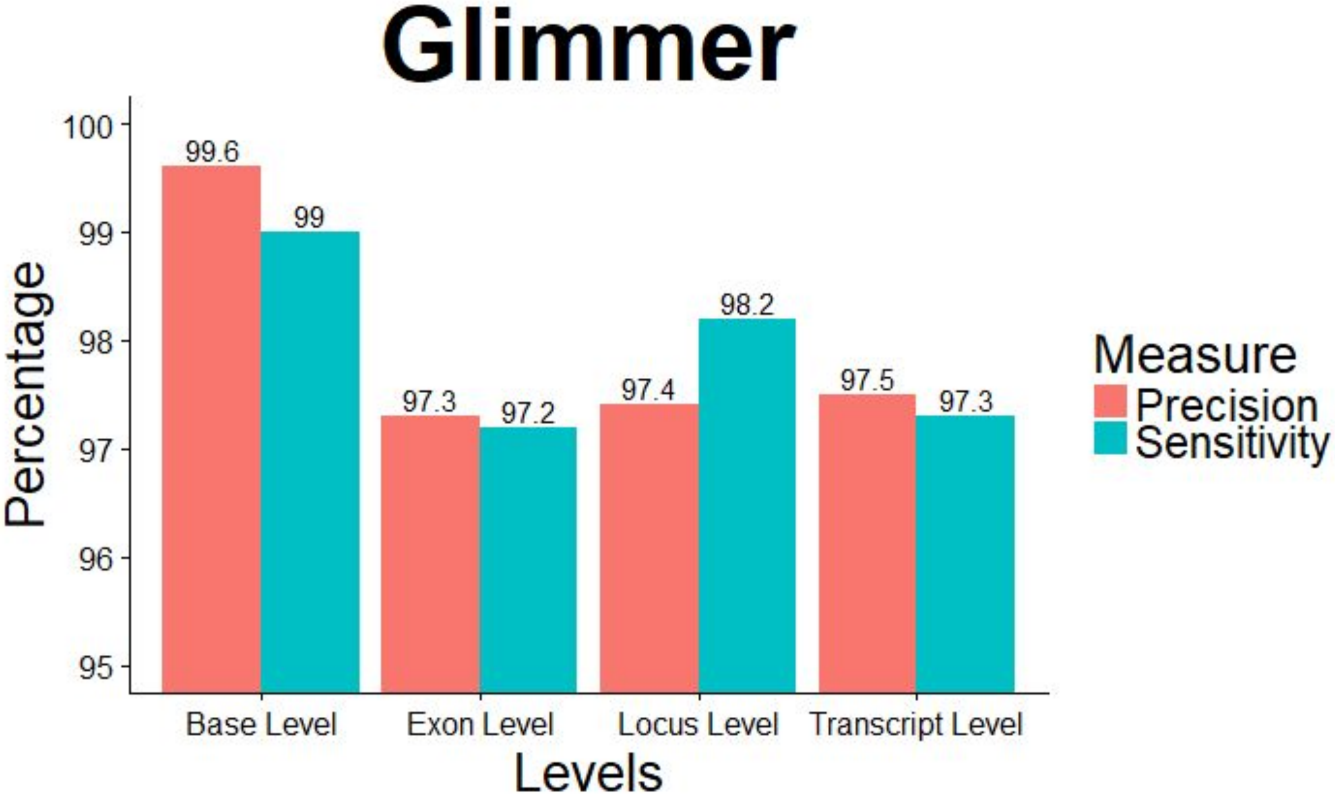
Burset, M., Guigó, R. : **Evaluation of gene structure prediction programs** (1996) Genomics, 34 (3), pp. 353-367

Sensitivity = TP / (TP+FN)

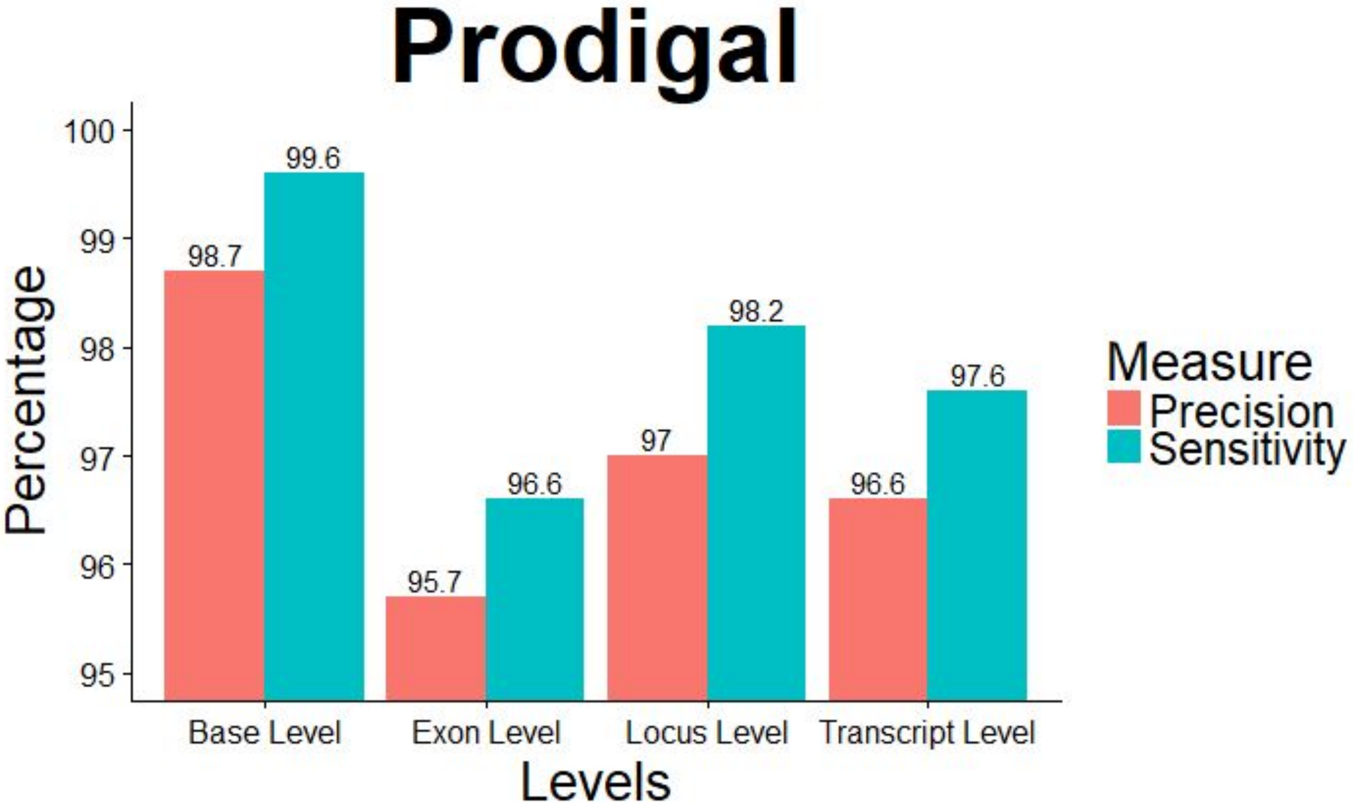
Precision = TP / (TP+FP)



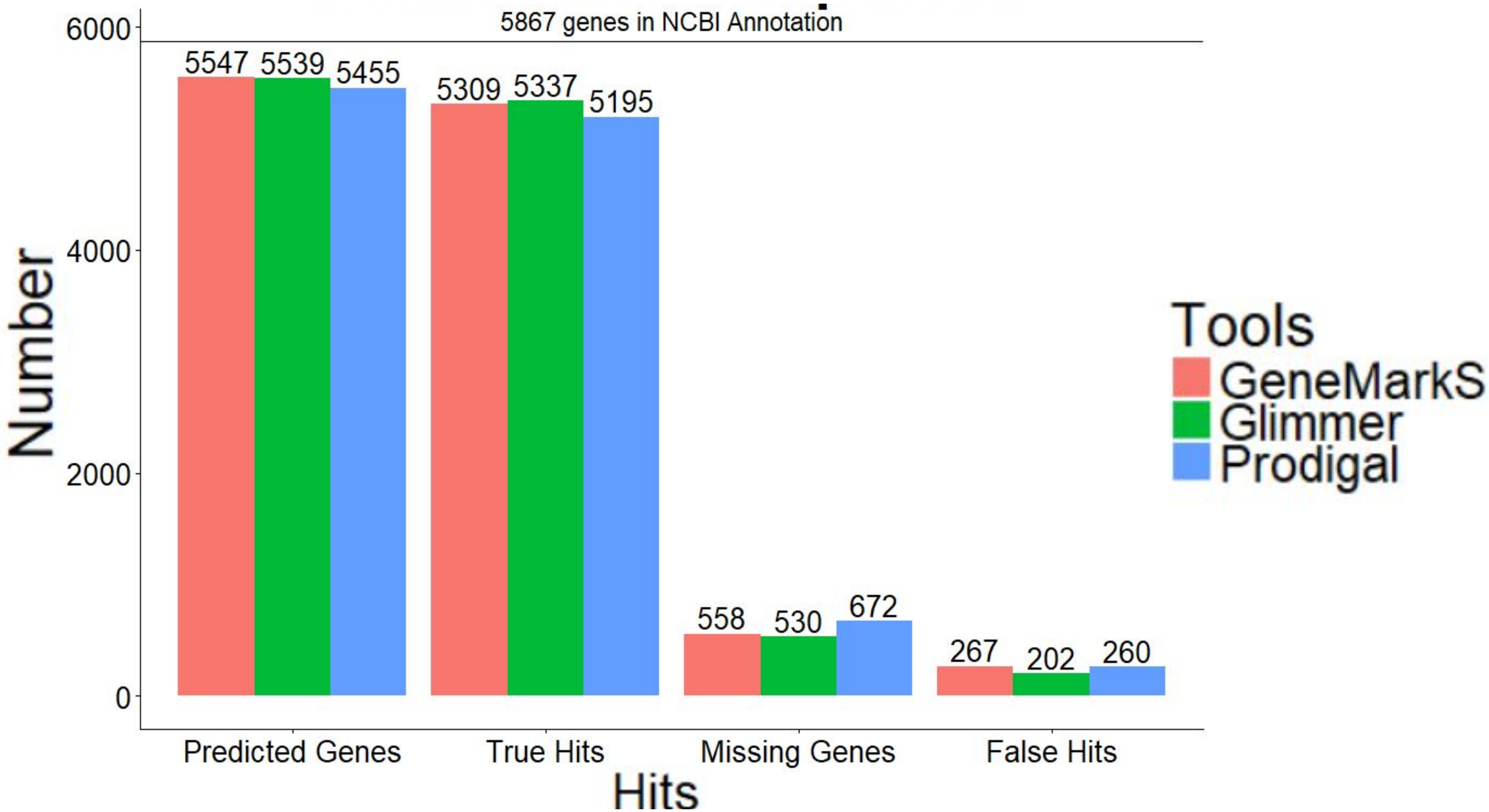
Ab-Initio - Tool Comparison: GFFcompare



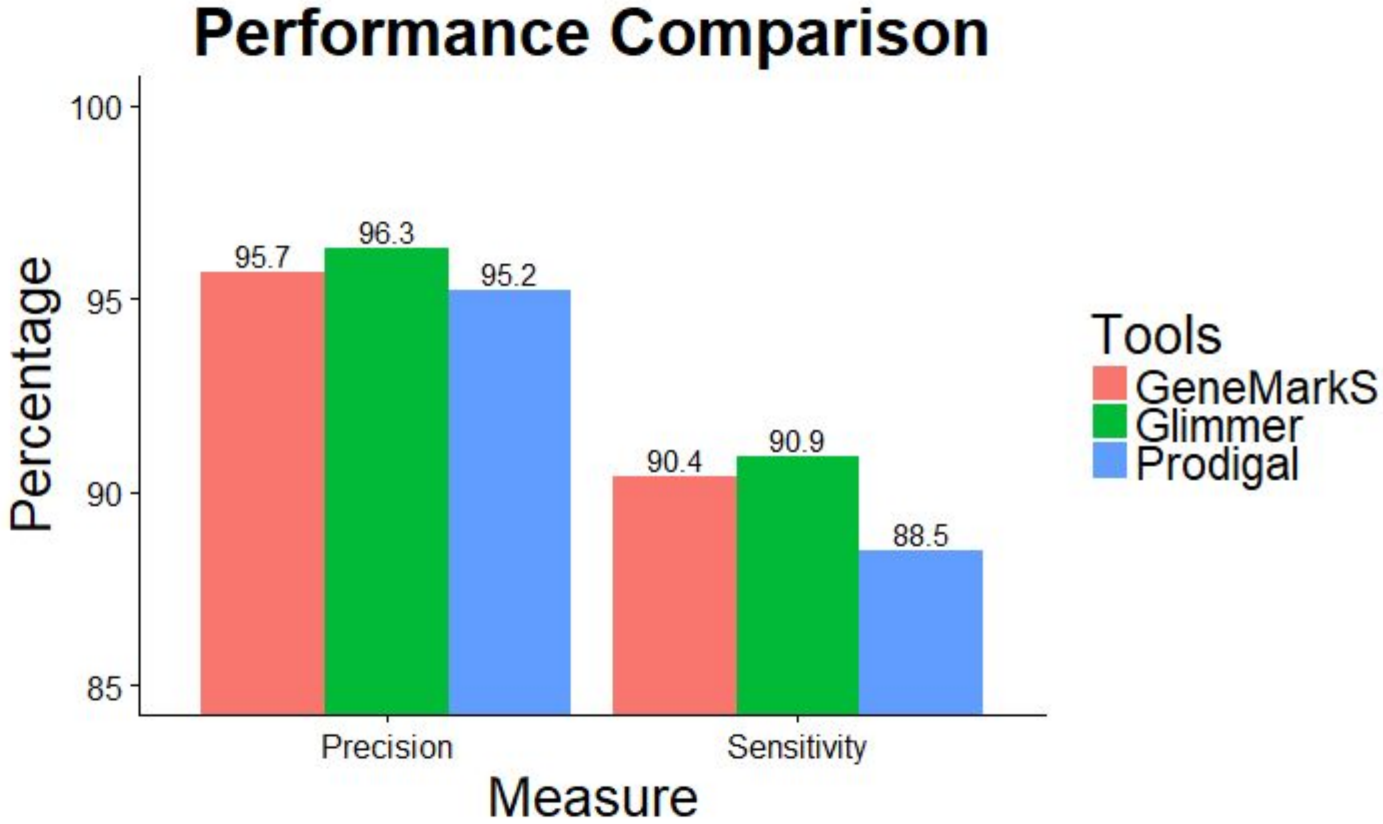
Ab-Initio - Tool Comparison: GFFcompare



Ab-Initio - Tool Comparison: GFFintersect



Ab-Initio - Tool Comparison: GFFintersect



Ab-initio - Results

We have a winner

Prodigal

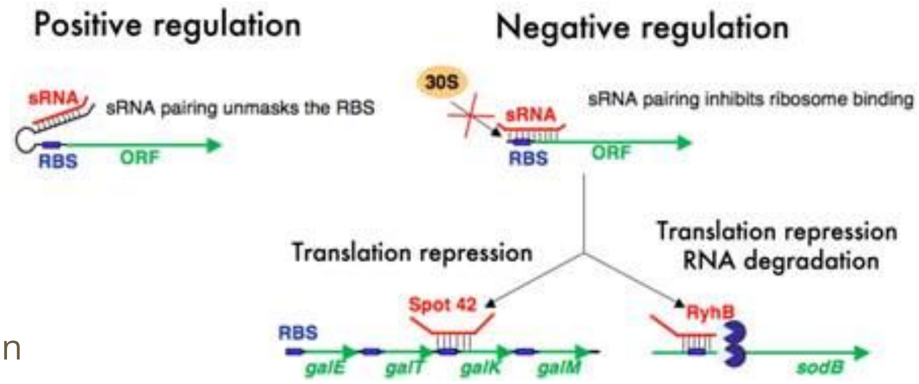
- Improvements
  - tweak the parameters
  - use iterative methods (converge results)

Non-Coding RNA

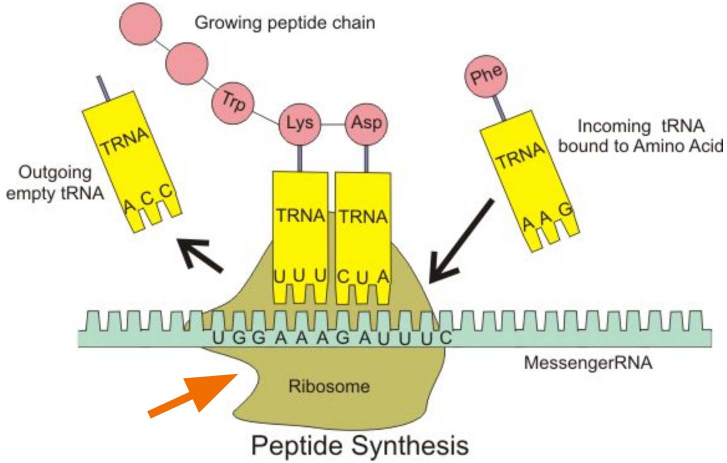
Description, Tools, and Strategy

ncRNA - Molecular Diversity and Tools

- **rRNA** (expected genes: 25)
  - RNAmmer
    - Using data from rRNA database
    - <1 min/genome
  - BarRNAp
    - Similar to RNAmmer
    - Multithreading is supported
  - Silva - Not working
- **tRNA** (expected genes: 62)
  - tRNAscan-SE 2.0
    - Better at finding weird tRNAs
    - Accurate, low error rate and ~1.8 mb/min
  - Aragorn
    - tRNA and tmRNA
    - Error and speed are GC content dependent
    - 5X faster with 40-60% GC
- **sRNA** (expected genes: 1)
  - Rfam
    - Troubleshooting!



ncRNA - rRNA Tools



rRNA genes in reference GFF: **25**

- **RNAmmer**
  - Predicted genes: 25
  - Run-time: 1m8s
  - GFF-version2
  - Predicted all of the same genes as the reference (same start/stop feature positions)
- **BarRNAp**
  - Predicted genes: 25
  - Run-time: 8s
  - GFF-version3
  - Start/Stop feature positions are offset from the reference

All of them predicted 8 16s rRNA, 8 23s rRNA, and 9 5s rRNA

ncRNA - tRNA Tools

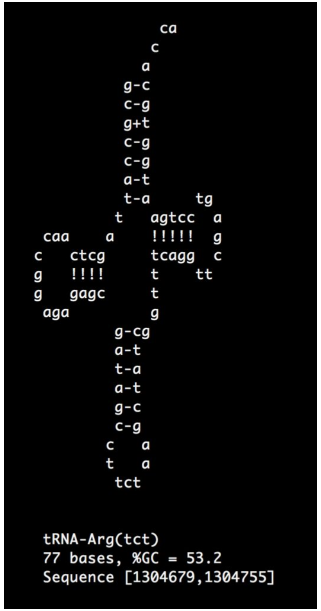
tRNA genes in reference GFF: **62**

- **Aragorn**
  - Predicted genes: 88
  - Run-time: 1s
  - Sensitivity: 98.4
  - Precision: 70.5

  Can also predict tmRNA

  - Query: 1 gene; Reference: 1 gene
  - Sensitivity:100; Precision:100
  - Run-time: 2s

- **tRNAscan-SE 2.0**
  - Predicted genes: 91
  - Run-time: 1m1s
  - Sensitivity: 98.4
  - Precision: 68.1



```
              ca
             c
            a
          g-c
          c-g
          g+t
          c-g
          c-g
          a-t
          t-a     tg
         t   agtcc  a
 caa    a    !!!!!  g
c   ctcg     tcagg  c
g   !!!!     t    tt
g   gagc     t
 aga         g
         g-cg
         a-t
         t-a
         a-t
         g-c
         c-g
        c   a
        t   a
         tct


 tRNA-Arg(tct)
 77 bases, %GC = 53.2
 Sequence [1304679,1304755]
```

ncRNA - Conclusion

rRNA

- BarRNAp
- Fast
- Results are similar to the one in ref. gff
- RNAmmer may be used to see if the results agree
- Output: gff2

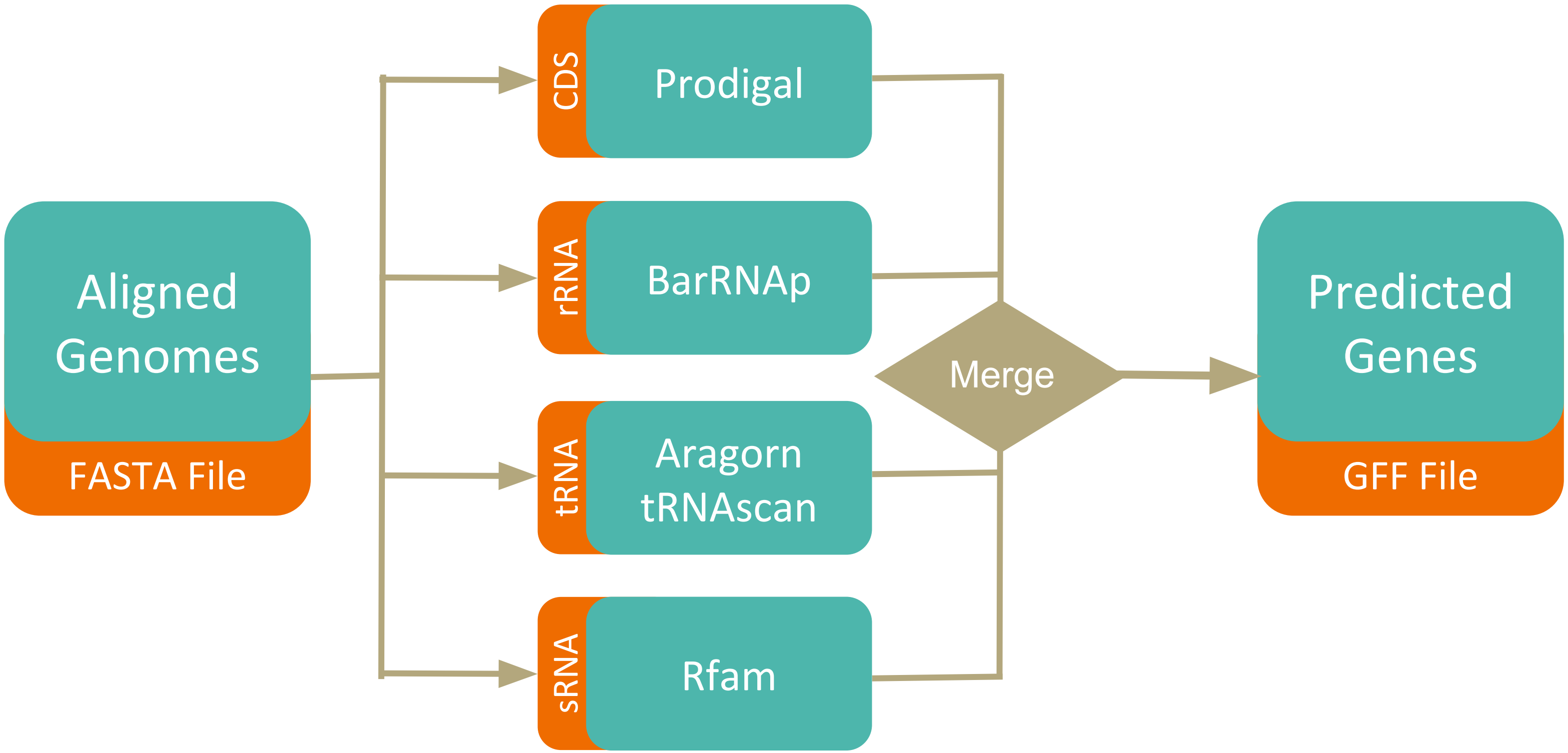
tRNA

- Using both to find consensus tRNA
- Aragorn finds tRNA and tmRNA
- tRNAscan finds more potential tRNA
- They are FAST
- Output: gff3

Final Pipeline

Preliminary Pipeline

Proposed Pipeline



Questions?