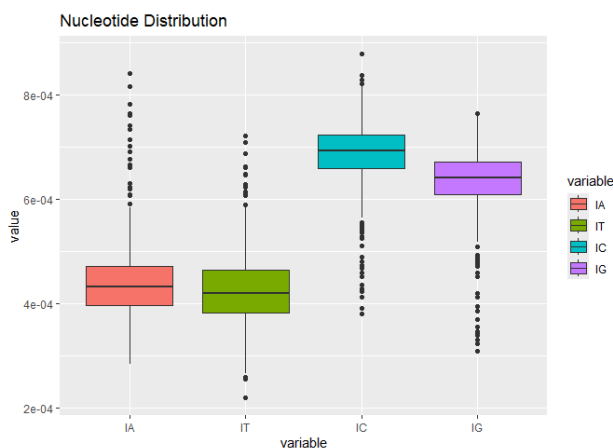# Serratia marcescens strain 3024, 615.1122

BIOL574 Final Report

Marian Lu

1 Initial Short Read

a. Number of sequences: 44
b. Sequence average length: 118107.5
c. Nucleotide composition:
- A: 20.23%
- T: 19.54%
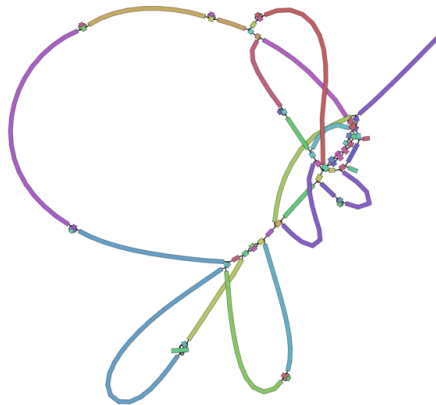- C: 31.31%
- G: 28.92%

2 BVBRC Assembly

a. BVBRC Statistics

Descriptor	Value
Subject ID	NR_074540.1
Query ID	Sim_12_contig_3
% Identity	78.955%
Alignment Length	1454
Mismatches	243
Gap Openings	50
Query Range	(3, 1429)
Subject Range	(1552, 135)
E-value	0.0
Bit score	931

b. N50 Value: 315818
   The N50 Value represents the length of the shortest contig. The higher the N50 value, the better the quality of the assembly. Generally, a N50 value greater than 50% of the total total genome sequence is considered high. However, the N50 of 315818 is only 6% of the total genome sequence, suggesting that this particular assembly is low in quality.
c. Assembly plot

3 Nucleotide Composition Analysis
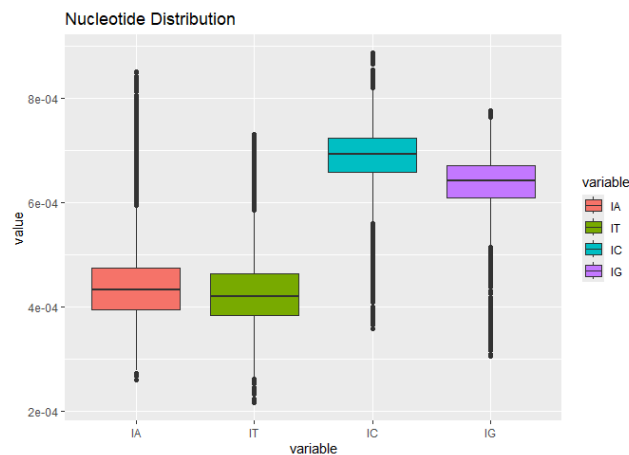
a. Global Stats for Nucleotide Composition
   - A: 20.23%
   - T: 19.54%
   - C: 31.31%
   - G: 28.92%
b. Moving Window Analysis
   Window Size: 1500
   Moving Increment: 1000

This contig seems to be GC rich, indicating that this protein might have higher thermal stability. The box plots show that there are a few outliers that significantly impact the average values of nucleotides. Overall, this closely resembles the distribution from the initial read.

4 BLASTn-based Taxonomy

Number of 16s rRNA Sequences identified: **45**

5 Open Reading Frames (ORFs)

    a.   Number of ORFs identified (total): **322**

6 BLASTp-based Annotation of ORFs

    a.   Database was created ✓

    b.   Files modified ✓

    c.   Number of positive hits: **489**

    d.   The initial number of identified ORFs is **492** and the percentage of positive hits is **99%**

8 Multiple Sequence Alignment (MSA)

    a.   Selected protein sequence length: 499

    b.   18 distantly related sequences from BLASTp

    c.   Most closely related protein: Tetrameric PepTSo2 in P4212 space group (6JKC)

    d.   MSA with CLUSTAL and MUSCLE

CLUSTAL O(1.2.4) multiple sequence alignment

```
6RKS_A   ITPVNIEPELKSSYLDYAMSVIYGRALPDVRDGLKPVHRRVLYAMNYHGNDWNKAYKKSA
6RKS_C   ITPVNIEPELKSSYLDYAMSVIYGRALPDVRDGLKPVHRRVLYAMNYHGNDWNKAYKKSA
6RKS_B   ITPVNIEPELKSSYLDYAMSVIYGRALPDVRDGLKPVHRRVLYAMNYHGNDWNKAYKKSA
6RKS_D   ITPVNIEPELKSSYLDYAMSVIYGRALPDVRDGLKPVHRRVLYAMNYHGNDWNKAYKKSA

6RKS_A   RVVGDVIGKYHPHGDSAVYDTIVRMAQPFSLRYMLVDGQGNFGSIDGDSAAAMRYTEIRL
6RKS_C   RVVGDVIGKYHPHGDSAVYDTIVRMAQPFSLRYMLVDGQGNFGSIDGDSAAAMRYTEIRL
6RKS_B   ----GLPGKLAD---------C-QERDPALSELYLVEGDSAGGSAK--------------
6RKS_D   ----GLPGKLAD---------C-QERDPALSELYLVEGDSAGGSAK--------------
         .:  **        .   :  :*   .  **:*:. **   .**

6RKS_A   AVGEVVLNNLYSQTQLQVSFGINMVALHHGQPKISHIRTLLIFFYRQMPEIVERS---
6RKS_C   AVGEVVLNNLYSQTQLQVSFGINMVALHHGQPKISHIRTLLIFFYRQMPEIVERS---
6RKS_B   ---EVVLNNLYSQTQLQVSFGINMVALHHGQ---SHIRTLLIFFYRQMPEIVERSHVYI
6RKS_D   ---EVVLNNLYSQTQLQVSFGINMVALHHGQ---SHIRTLLIFFYRQMPEIVERSHVYI
         .::. **. :*:     *  *.  .     ::: : :: * *:  *:*  *

6RKS_A   ---TIFEDRKARDRAHILEALAVALANIDPIIELIRHAPTPAEAKTALVANPWQLGNVAA
6RKS_C   ---TIFEDRKARDRAHILEALAVALANIDPIIELIRHAPTPAEAKTALVANPWQLGNVAA
6RKS_B   AQPTIFEDRKARDRAHILEALAV-----------------------NPWQLGNVAA
6RKS_D   AQPTIFEDRKARDRAHILEALAV-----------------------NPWQLGNVAA
         :::::*.::. :* * *:            : :*  .  .

6RKS_A   LGEMNPEQLWETTMDGVGLYYLTEQQAQAILDLRLQKLTGLEHEKLLDEYKELLDQIAEL
6RKS_C   LGEMNPEQLWETTMDGVGLYYLTEQQAQAILDLRLQKLTGLEHEKLLDEYKELLDQIAEL
6RKS_B   LGEMNPEQLWETTMD--------PESRRMLRVTV--KDAIAADQL---FT---------
6RKS_D   LGEMNPEQLWETTMD--------PESRRMLRVTV--KDAIAADQL---FT---------
         *:  .*  *          ::: :* : :    .: ::* :.

6RKS_A   LRILGSA
6RKS_C   LRILGSA
6RKS_B   -TLMGDA
6RKS_D   -TLMGDA
         ::*.*
```

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
6RKS_A   ITPVNIEELKSSYLDYAMSVIVGRALPDVRDGLKPVHRRVLYAMNVLGNDWNKAYKKSA
6RKS_C   ITPVNIEELKSSYLDYAMSVIVGRALPDVRDGLKPVHRRVLYAMNVLGNDWNKAYKKSA
6RKS_B   ITPVNIEELKSSYLDYAMSVIVGRGLPG---GLKPYSRR-------------------
6RKS_D   ITPVNIEELKSSYLDYAMSVIVGRGLPG---GLKPYSRR-------------------
         .**.    *    : *

6RKS_A   RVVGDVIGKYHPHGDSAVYDTIVRMAQPFSLRYMLVDGQGNFGSIDGDSAAAMRYTEIRL
6RKS_C   RVVGDVIGKYHPHGDSAVYDTIVRMAQPFSLRYMLVDGQGNFGSIDGDSAAAMRYTEIRL
6RKS_B   --------------DSAVYDTIV------------------IDGDSAA----------
6RKS_D   --------------DSAVYDTIV------------------IDGDSAA----------
         .*:: :                                   ::****.

6RKS_A   THVINGCLAYIDDEDISIEGLMEFIFGPDFPTAAINGRFGEIEEAYDTGRGLVYIIRARAE
6RKS_C   THVINGCLAYIDDEDISIEGLMEFIFGPDFPTAAINGRFGEIEEAYDTGRGLVYIIRARAE
6RKS_B   THVINGCLAYIDDEDISIEGLMEFIP------NGRFGEIEEAYDTGRGLVY------
6RKS_D   THVINGCLAYIDDEDISIEGLMEFIP------NGRFGEIEEAYDTGRGLVY------
         :***. ::*      *                NGRF

6RKS_A   VEVDARETIIVHEIPYQVNKARLIEKIAELVKEKRVEGISALRDESDKDGMRIVIEVKRD
6RKS_C   VEVDARETIIVHEIPYQVNKARLIEKIAELVKEKRVEGISALRDESDKDGMRIVIEVKRD
6RKS_B   --------------NVEKARF----DKMLSSQEV--------------------
6RKS_D   --------------NVEKARF----DKMLSSQEV--------------------
         *:***:       :::..: *

6RKS_A   AVGEVVLNNLYSQTQLQVSFGINMVALHHGQPKIMNAKDIIAAFVRHRREVVTRRTIFEL
6RKS_C   AVGEVVLNNLYSQTQLQVSFGINMVALHHGQPKIMNAKDIIAAFVRHRREVVTRRTIFEL
6RKS_B   ---ATLITALYSPDKLRYHSIIMTDADVDGSHIRTV--LLTFFYRQMPEIVERGHVY--
6RKS_D   ---ATLITALYSPDKLRYHSIIMTDADVDGSHIRTV--LLTFFYRQMPEIVERGHVY--
         .::.. **. :*.     * *.    ..:* .*   ::: * *:  *:* *  ::

6RKS_A   EWLEFEESGVGLYYLTEQQAQAILDLRLQKLTGLEHEKLLDEYKELLDQIAELLRILGSAD
6RKS_C   EWLEFEESGVGLYYLTEQQAQAILDLRLQKLTGLEHEKLLDEYKELLDQIAELLRILGSAD
6RKS_B   EWLEFEESGVGLYYLTEQQAQALD-RYKGLGEMNPEQLWETTMD--PESRRMLRV-----
6RKS_D   EWLEFEESGVGLYYLTEQQAQALD-RYKGLGEMNPEQLWETTMD--PESRRMLRV-----
         : : *    *:.::*:    : * :* * :: *:*: :    :   :**:

6RKS_A   RLMEVIREELELVREQFGD---KRRTEITAN
6RKS_C   RLMEVIREELELVREQFGD---KRRTEITAN
6RKS_B   RLMEVIREE-QLFTTLMGDAVEPRRA-----
6RKS_D   RLMEVIREE-QLFTTLMGDAVEPRRA-----
         : :.*  :  :*.   :**   **:
```

# 9 Phylogenetic Trees

    a.   Created 2 phylogenetic trees of 20 protein sequences ✓

    b.   Created one distance tree and one bootstrap tree ✓

    c.   How many tree topologies for each of the two trees?

N topologies for Rooted Tree: $N_{unrooted} = \frac{(2n-3)!}{2^{n-3}(n-3)!}$, *where n=number of taxa*

        i.   **Protein Distance Tree:** 3 unrooted topologies (n=4)

        ii.   **Protein Bootstrap Tree**: 3 unrooted topologies (n=4)

    d.   iTOL Visualizations

        i.   **Protein Distance Tree**

(6RKS_C:0.00000,(6RKS_B:0.00000,6RKS_D:0.00000):2.22593,6RKS_A:0.00001);

Tree scale: 0.1 ⊢——⊣
0.00001

6RKS A
6RKS C
6RKS D
6RKS B

2.22593
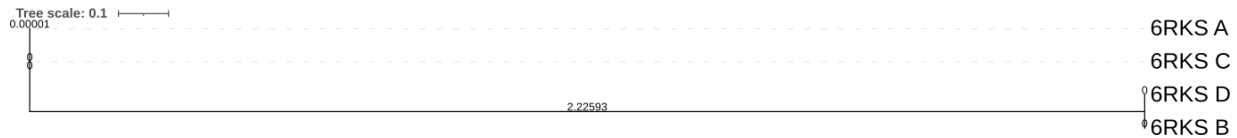
        ii.   **Protein Bootstrap Tree**

(((6RKS_B:100.0,6RKS_D:100.0):100.0,6RKS_A:100.0):100.0,6RKS_C:100.0);

Tree scale: 10 ⊢——⊣

100
100
100
100
100
100

6RKS C
6RKS A
6RKS D
6RKS B