**PHAGE – DONKEYMAN CLUSTER DY – ANNOTATION NOTES**

**GENE 1** **hypothetical protein**

Original Glimmer call @bp 76 has strength 12.68

SSC: 76 to 507 (Forward)

CP: yes

SCS: both

ST: SS, Start 3 [76], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 85.7% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_1, function unknown, query 1-143 subject 1-143, E value = 6e-73, Identity = 96%, Similarity = 98%

GAP: na

LO: NA

RBS: Kibler7, spacer 9, Z score 2.481, Final -3.894, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842599.1, query 1-143 subject 1-143, E value = 5e-93, Identity = 97%, Similarity = 98%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 2 terminase**

Original Glimmer call @bp 512 has strength 17.33

SSC: 512 to 2101 (Forward)

CP: yes

SCS: both

ST: NI, Start 114 [512 ], Found in 30 of 1447 ( 2.1% ) of genes in pham, Called 43.3% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_2, terminase, query 1-529 subject 1-529, E value = 0.0, Identity = 99%, Similarity = 99%

GAP: 5bp

LO: NA

RBS: Kibler7, spacer 14, Z score 1.949, Final -5.536, no

F: terminase

SIF-blast: aligns with terminase from Gordonia phage Santhid, ID YP\_010842540.1, query 1-529 subject 1-529, E value = 0.0, Identity = 99%, Similarity = 99%

SIF-HHPred: 6Z6D\_A, Terminase large subunit; genome packaging, bacteriophage, ATPase, nuclease, VIRAL PROTEIN; HET: BR; 2.2A, Enterobacteria, probability = 100%, E value = 1.2e-37

SF-Syn: This gene is pham 229711 [terminase], upstream is pham 8805, downstream is pham 219755[RNA binding protein], as seen in Gordonia phage Tarzan.

**GENE 3 RNA binding protein**

Original Glimmer call @bp 2115 has strength 13.99

SSC: 2115 to 2354 (Forward)

CP: yes

SCS: both

ST: NI, Start 8 [2115], Found in 7 of 60 ( 11.7% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_3, RNA binding protein, query 1-79 subject 1-79, E value = 1e-40, Identity = 98%, Similarity = 100%

GAP: 15 bp

LO: yes

RBS: Kibler7, spacer 11, Z score 2.995, Final -2.922, yes

F: RNA binding protein

SIF-blast: aligns with RNA binding protein from Gordonia phage Santhid, ID YP\_010842541.1, query 1-79 subject 1-79, E value = 1e-44, Identity = 97%, Similarity = 100%

SIF-HHPred: 8RO0\_b, Probable small nuclear ribonucleoprotein-associated protein B; mRNA, splicing, Intorn Lariat spliceosome, ILS, pre-mRNA; HET: GTP, IHP; 2.9A {Caenorhabditis elegans}

Probability: 99.51%, E-value: 3.6e-13

SF-Syn: This gene is pham 219755[RNA binding protein], upstream is pham 229711 [terminase], downstream is pham 230074 [portal protein], as seen in Gordonia phage Tarzan.

**GENE 4 portal protein**

Original Glimmer call @bp 2354 has strength 17.59

SSC: 2354 to 3652 (Forward)

CP: yes

SCS: both

ST: NI, Start 14 [2354], Found in 59 of 90 ( 65.6% ) of genes in pham, Called 98.3% of time when present

BLAST-start: aligns with Gordonia phage Hibiscus\_4, portal protein, query 1-432 subject 1-432, E value = 0.0, Identity = 98%, Similarity = 98%

GAP: TAGTG overlap

LO: NA

RBS: Kibler7, spacer 11, Z score 1.778, Final -5.289, yes

F: portal protein

SIF-blast: aligns with portal protein from Gordonia phage Santhid, ID YP\_010842542.1, query 1-432 subject 1-432, E value = 0.0, Identity = 97%, Similarity = 98%

SIF-HHPred: 8FQL\_B, Portal protein; Prohead I, icosahedral symmetry, HK97, phage, capsid, VIRUS; 3.6A {Escherichia phage HK97}, Probability: 100%, E-value: 1e-39,

SF-Syn: This gene is pham 230074 [portal protein], upstream is pham 219755[RNA binding protein], downstream is pham 227183 [capsid maturation protease], as seen in Gordonia phage Tarzan.

**GENE 5 capsid maturation protease**

Original Glimmer call @bp 3652 has strength 15.88

SSC: 3652 to 4452 (Forward)

CP: yes

SCS: both

ST: NI, Start 41[3652], Found in 154 of 324 ( 47.5% ) of genes in pham, Called 98.7% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_5, capsid maturation protease, query 1-266 subject 1-266, E value = e-149, Identity = 98%, Similarity = 99%

GAP: TAATG overlap

LO: NA

RBS: Kibler7, spacer 12, Z score 3.002, Final -2.906, yes

F: capsid maturation protease

SIF-blast: aligns with capsid maturation protease from Gordonia phage Reyja, ID YP\_010842668.1, query 1-266 subject 1-266, E value = 0.0, Identity = 99%, Similarity = 100%

SIF-HHPred: 5JBL\_D, Prohead core protein protease; protease pentamer, phage T4, prohead, HYDROLASE; 1.943A {Enterobacteria phage T4}, Probability: 97.14%, E-value: 0.048,

SF-Syn: This gene is pham 227183 [capsid maturation protease], upstream is pham 230074 [portal protein], downstream is pham 228212[major capsid protein], as seen in Gordonia phage Tarzan.

**GENE 6 major capsid protein**

Original Glimmer call @bp 4491 has strength 15.19

SSC: 4491 to 5804 (Forward)

CP: yes

SCS: both

ST: NI, Start 29 [4491], Found in 115 of 350 ( 32.9% ) of genes in pham, Called 89.6% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_6, major capsid protein, query 1-437 subject 1-437, E value = 0.0, Identity = 97%, Similarity = 98%

GAP: 40bp

LO: yes

RBS: Kibler7, spacer 11, Z score 2.901, Final -3.030, yes

F: major capsid protein

SIF-blast: aligns with major capsid protein from Gordonia phage Jojo24, ID YP\_010842733.1, query 1-437 subject 1-437, E value = 0.0, Identity = 96%, Similarity = 98%

SIF-HHPred: 7RWZ\_B, Major capsid protein; HK97-like fold, capsid size redirection, major capsid protein, VIRUS; 4.0A {Staphylococcus aureus}, Probability: 100%, E-value: 1e-32,

SF-Syn: This gene is pham 228212[major capsid protein], upstream is pham 227183 [capsid maturation protease], downstream is pham 8238, as seen in Gordonia phage Tarzan.

**GENE 7 hypothetical protein**

Original Glimmer call @bp 5886 has strength 15.96

SSC: 5886 to 6275 (Forward)

CP: yes

SCS: both

ST: SS, Start 3[5886], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 85.7% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_7, function unknown, query 1-129 subject 1-129, E value =2e-67 , Identity = 97%, Similarity = 99%

GAP: 82bp

LO: no

RBS: Kibler7, spacer 13, Z score 3.002, Final -3.116, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Jojo24, ID YP\_010842734.1, query 1-129 subject 1-131, E value = 5e-81, Identity = 97%, Similarity = 97%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 8 head-to-tail adaptor**

Original Glimmer call @bp 6278 has strength 16.19

SSC: 6278 to 6811 (Forward)

CP: yes

SCS: both

ST: NI, Start 15[6278], Found in 9 of 378 ( 2.4% ) of genes in pham, Called 88.9% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_8, head-to-tail adapter, query 1-177 subject 1-177, E value = e-100, Identity = 99%, Similarity = 100%

GAP: 3bp

LO: NA

RBS: Kibler7, spacer 6, Z score 1.813, Final -6.208, yes

F: head-to-tail adaptor

SIF-blast: aligns with head-to-tail adaptor from Gordonia phage Reyja, ID YP\_010842671.1, query 1-177 subject 1-177, E value = 4e-103, Identity = 85%, Similarity = 91%

SIF-HHPred: 8HQO\_R, Head completion protein; Neck, Portal, T5, VIRUS, VIRAL PROTEIN; 3.2A {Escherichia phage DT57C}, Probability: 99.93%, E-value: 3.9e-24

SF-Syn: This gene is pham 229835[head-to-tail adaptor], upstream is pham 8238, downstream is pham 225195[head-to-tail stopper], as seen in Gordonia phage Tarzan.

**GENE 9 head-to-tail stopper**

Original Glimmer call @bp 6832 has strength 15.68; GeneMark calls start at 6808

SSC: 6808 to 7146 (Forward)

CP: yes

SCS: both – GM [ based on starterator, alignment, RBS score the start site was changed to 6808]

ST: SS, Start 4 [6808], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 85.7% of time when present

Start 5[6832], Found in 6 of 7 ( 85.7% ) of genes in pham, Called 16.7% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_9, head-to-tail stopper, query 1-104 subject 1-112, E value = 2e-54, Identity = 98%, Similarity = 99%

GAP: GTGA overlap

LO: NA

RBS: Kibler7, spacer 9, Z score 2.060, Final -4.741, yes

F: head-to-tail stopper

SIF-blast: aligns with head-to-tail stopper from Gordonia phage Santhid, ID YP\_010842547.1, query 1-104 subject 1-112, E value = 3e-60, Identity = 92%, Similarity = 95%

SIF-HHPred: 9D94\_Hb, Head-to-tail stopper; Bacteriophage, portal, VIRAL PROTEIN; {Mycobacterium phage Bxb1}, Probability: 98.89%, E-value: 5.7e-8

SF-Syn: This gene is pham 225195[head-to-tail stopper], upstream is pham 229835[head-to-tail adaptor], downstream is pham 230206, as seen in Gordonia phage Tarzan.

**GENE 10 hypothetical protein**

Original Glimmer call @bp 7146 has strength 15.66

SSC: 7146 to 7430 (Forward)

CP: yes

SCS: both

ST: NI, Start 20[7146], Found in 12 of 45 ( 26.7% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_10, function unknown, query 1-94 subject 1-94, E value = 5e-44, Identity = 98%, Similarity = 98%

GAP: TAGTG overlap

LO: NA

RBS: Kibler7, spacer 11, Z score 1.978, Final -4.888, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842608.1, query 1-94 subject 1-94, E value = 2e-56, Identity = 99%, Similarity = 98%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 11 tail terminator**

Original Glimmer call @bp 7427 has strength 17.89

SSC: 7427 to 7831 (Forward)

CP: yes

SCS: both

ST: NI, Start 29[7427], Found in 35 of 82 ( 42.7% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_11, tail terminator, query 1-134 subject 1-134, E value = 2e-72, Identity = 99%, Similarity = 99%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 12, Z score 2.051, Final -4.820, yes

F: tail terminator

SIF-blast: aligns with tail terminator from Gordonia phage Jojo24, ID YP\_010842738.1, query 1-134 subject 1-134, E value = 3e-79, Identity = 88%, Similarity = 94%

SIF-HHPred: 9D94\_Ib, Tail terminator; Bacteriophage, portal, VIRAL PROTEIN; {Mycobacterium phage Bxb1}, Probability: 99.84%, E-value: 5.4e-19

SF-Syn: This gene is pham 224509[tail terminator], upstream is pham 230206, downstream is pham 229963[major tail protein], as seen in Gordonia phage Tarzan.

**GENE 12 major tail protein**

Original Glimmer call @bp 7926 has strength 15.70

SSC: 7926 to 8492 (Forward)

CP: yes

SCS: both

ST: SS, Start 10[7926], Found in 163 of 163 ( 100.0% ) of genes in pham, Called 98.8% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_12, major tail protein, query 1-188 subject 1-188, E value = e-106, Identity = 100%, Similarity = 100%

GAP: 95bp

LO: yes

RBS: Kibler7, spacer 12, Z score 3.002, Final -2.906, yes

F: major tail protein

SIF-blast: aligns with major tail protein from Gordonia phage Reyja, ID YP\_010842675.1, query 1-188 subject 1-188, E value = 9e-132, Identity = 97%, Similarity = 99%

SIF-HHPred: 9D94\_Jf, Major tail protein; Bacteriophage, portal, VIRAL PROTEIN; {Mycobacterium phage Bxb1}, Probability: 99.92%, E-value: 3.4e-24

SF-Syn: This gene is pham 229963[major tail protein], upstream is pham 224509[tail terminator], downstream is pham 5041, as seen in Gordonia phage Tarzan.

**GENE 13 hypothetical protein**

Original Glimmer call @bp 8605 has strength 17.35

SSC: 8605 to 8931 (Forward)

CP: yes

SCS: both

ST: NI, Start 11[8605], Found in 7 of 14 ( 50.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_13, function unknown, query 1-108 subject 1-108, E value = 1e-56, Identity = 99%, Similarity = 99%

GAP: 113bp

LO: yes

RBS: Kibler7, spacer 11, Z score 3.084, Final -2.664, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Santhid, ID YP\_010842551.1, query 1-108 subject 1-108, E value = , 2e-58 Identity = 84%, Similarity = 90%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 14 tail assembly chaperone**

Original Glimmer call @bp 8928 has strength 12.81

SSC: 8928 to 9542 (Forward)

CP: yes

SCS: both

ST: SS, Start 15[8928], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_14, tail assembly chaperone, query 1-204 subject 1-204, E value = e-119, Identity = 99%, Similarity = 99%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 15, Z score 2.633, Final -4.416, yes

F: tail assembly chaperone

SIF-blast: aligns with tail assembly chaperone from Gordonia phage Tarzan, ID YP\_010842613.1, query 1-204 subject 1-204, E value = 8e-143, Identity = 99%, Similarity = 99%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: This gene is pham 132724[tail assembly chaperone], upstream is pham 5041, downstream is pham 220257 [tape measure protein], as seen in Gordonia phage Tarzan.

**GENE 15 frameshift with gene 14 tail assembly chaperone**

To be changed due to being part of a frameshift – see below

GGGGGAAA -1 frameshift from reading frame 3 to reading frame 2 at the slippery sequence 9531 to 9539.

GENE 15 PART 1

8928 - 9536

PART 2

A [9536] GGA [END OF PART 1] IS USED AS START [AAA - K] OF PART 2.

9536 - 9955 [TGA END OF PART 2]

**GENE 16 tape measure protein**

Original Glimmer call @bp 9948 has strength 13.13

SSC: 9948 to 14102 (Forward)

CP: yes

SCS: both

ST: SS, Start 4[9948], Found in 7 of 14 ( 50.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_16, tape measure protein, query 1-1384 subject 1-1387, E value = 0.0, Identity = 99%, Similarity = 99%

GAP: 7bp overlap

LO: NA

RBS: Kibler7, spacer 12, Z score 2.226, Final -4.467, yes

F: tape measure protein

SIF-blast: aligns with tape measure protein from Gordonia phage Tarzan, ID YP\_010842614.1, query 1-1384 subject 1-1387, E value = 0.0, Identity = 99%, Similarity = 99%

SIF-HHPred: 6V8I\_AF, Tape Measure Protein, gp57; phage tail, tail tip, tape measure protein, VIRAL PROTEIN; 3.7A {Staphylococcus virus 80alpha}, Probability: 99.28%, E-value: 0.000053,

SF-Syn: This gene is pham 220257 [tape measure protein], upstream is pham 132724[tail assembly chaperone], downstream is pham 230392 [minor tail protein], as seen in Gordonia phage Tarzan.

SIF-Mem: 5 domains found by TMHMM, 5 domains found by SOSUI

**GENE 17 minor tail protein**

Original Glimmer call @bp 14102 has strength 14.96; GeneMark calls start at 14063

SSC: 14102 to 14944 (Forward)

CP: yes

SCS: both – GL [14063 would give a 40bp overlap, alignment, starterator, etc shows 14102 as start]

ST: SS, Start 5 [14102], Found in 16 of 17 ( 94.1% ) of genes in pham, Called 93.8% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_17, minor tail protein, query 1-280 subject 1-280, E value = e-168, Identity = 100%, Similarity = 100%

GAP: TAGTG overlap

LO: NA

RBS: Kibler7, spacer 8, Z score 2.744, Final -3.812, yes

F: minor tail protein

SIF-blast: aligns with minor tail protein from Gordonia phage Reyja, ID YP\_010842680.1, query 1-280 subject 1-280, E value = 0.0, Identity = 99%, Similarity = 100%

SIF-HHPred: 2X8K\_A, HYPOTHETICAL PROTEIN 19.1; VIRAL PROTEIN, DISTAL TAIL PROTEIN; 2.95A {BACILLUS PHAGE SPP1}, Probability: 99.95%, E-value: 3.8e-26,

SF-Syn: This gene is pham 230392 [minor tail protein], upstream is pham 220257 [tape measure protein], downstream is pham 228138 [minor tail protein], as seen in Gordonia phage Tarzan.

**GENE 18 minor tail protein**

Original Glimmer call @bp 14949 has strength 18.35

SSC: 14949 to 16565 (Forward)

CP: yes

SCS: both

ST: NI, Start 193[14949], Found in 20 of 2500 ( 0.8% ) of genes in pham, Called 70.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_18, minor tail protein, query 1-538 subject 1-538, E value = 0.0, Identity = 99%, Similarity = 99%

GAP: 5bp

LO: NA

RBS: Kibler7, spacer 9, Z score 2.112, Final -4.635, yes

F: minor tail protein

SIF-blast: aligns with minor tail protein from Gordonia phage Reyja, ID YP\_010842681.1, query 1-538 subject 1-538, E value = 0.0, Identity = 99%, Similarity = 99%

SIF-HHPred: 9D93\_Oa, Minor tail protein; Bacteriophage, tail tip, VIRAL PROTEIN; {Mycobacterium phage Bxb1}, Probability: 100%, E-value: 1.3e-66,

SF-Syn: This gene is pham 228138 [minor tail protein], upstream is pham 230392 [minor tail protein], downstream is pham 4920, as seen in Gordonia phage Tarzan.

**GENE 19 hypothetical protein**

Original Glimmer call @bp 16562 has strength 11.62

SSC: 16562 to 16975 (Forward)

CP: yes

SCS: both

ST: SS, Start 2[16562], Found in 14 of 15 ( 93.3% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Reyja\_20, function unknown, query 1-137 subject 1-137, E value = 3e-79, Identity = 100%, Similarity = 100%

GAP: GTGA overlap

LO: NA

RBS: Kibler7, spacer 15, Z score 2.153, Final -5.381, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842617.1 , query 1-137 subject 1-137, E value = 2e-93, Identity = 99%, Similarity = 99%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 20 lysin A, L-Ala-D-Glu peptidase domain**

Original Glimmer call @bp 16981 has strength 15.12

SSC: 16981 to 17565 (Forward)

CP: yes

SCS: both

ST: SS, Start 27[16981], Found in 19 of 211 ( 9.0% ) of genes in pham, called 89.5% of time when present

BLAST-start: aligns with Gordonia phage Reyja\_21, lysin A, L-Ala-D-Glu peptidase domain, query 1-194 subject 1-194, E value = e-116, Identity = 100%, Similarity = 100%

GAP: 6bp

LO: NA

RBS: Kibler7, spacer 11, Z score 1.583, Final -5.682, yes

F: lysin A, L-Ala-D-Glu peptidase domain

SIF-blast: aligns with lysin A, L-Ala-D-Glu peptidase domain from Gordonia phage Tarzan, ID YP\_010842618.1, query 1-194 subject 1-194, E value = 3e-138, Identity = 98%, Similarity = 98%

SIF-HHPred: aligns with the following but only for 54aa’s - 8Q2G\_A, Peptidoglycan L-alanyl-D-glutamate endopeptidase; Endolysins, Bacteriophage lytic enzymes, Gram-negative bacteria, Antibacterial therapy, Antimicrobial agents, ANTIBIOTIC, ANTIMICROBIAL PROTEIN; 2.49A, {Escherichia phage ECD7}, Probability: 98.6%, E-value: 3.2e-7,

SF-Syn: This gene is pham 229916 [lysin A, L-Ala-D-Glu peptidase domain], upstream is pham 4920, downstream is pham 229967[lysin A, glycosyl hydrolase domain], as seen in Gordonia phage Tarzan.

**GENE 21 lysin A, glycosyl hydrolase domain**

Original Glimmer call @bp 17562 has strength 17.14

SSC: 17562 to 18572 (Forward)

CP: yes

SCS: both

ST: NI, Start 47[17562], Found in 16 of 160 ( 10.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_21, lysin A, glycosyl hydrolase domain, query 1-336 subject 1-336, E value = 0.0, Identity = 98%, Similarity = 99%

GAP: TGTGA overlap

LO: NA

RBS: Kibler7, spacer 8, Z score 2.552, Final -4.198, yes

F: lysin A, glycosyl hydrolase domain

SIF-blast: aligns with lysin A, glycosyl hydrolase domain from Gordonia phage Reyja, ID YP\_010842684.1, query 1-336 subject 1-336, E value = 0.0, Identity = 99%, Similarity = 99%

SIF-HHPred: 200aa match - 5JIP\_A, Cortical-lytic enzyme; spore, cortex, peptidoglycan-lysin, hydrolase; HET: MES; 1.8A {Clostridium perfringens}, Probability: 99.63%, E-value: 5.2e-14,

SF-Syn: This gene is pham 229967[lysin A, glycosyl hydrolase domain], upstream is pham 229916 [lysin A, L-Ala-D-Glu peptidase domain], downstream is pham 227727, as seen in Gordonia phage Tarzan.

**GENE 22 hypothetical protein**

Original Glimmer call @bp 18572 has strength 14.86

SSC: 18572 to 18871 (Forward)

CP: yes

SCS: both

ST: NI, Start 15[18572], Found in 7 of 15 ( 46.7% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_22, function unknown, query 1-99 subject 1-99, E value = 8e-52, Identity = 100%, Similarity = 100%

GAP: TGATG overlap

LO: NA

RBS: Kibler7, spacer 9, Z score 2.072, Final -4.717, yes

F: NKF

SIF-blast: aligns with membrane protein from Gordonia phage Tarzan, ID YP\_010842620.1, query 1-99 subject 1-99, E value = 5e-62, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

SIF-Mem: 2 domains found by TMHMM, 2 domains found by SOSUI

Phages db blast shows alignment with proteins that are nkf but ncbi blast shows those same phage proteins as a membrane protein

**GENE 23 hypothetical protein**

Original Glimmer call @bp 18871 has strength 11.11

SSC: 18871 to 19260 (Forward)

CP: yes

SCS: both

ST: NI, Start 34[18871], Found in 43 of 49 ( 87.8% ) of genes in pham, Called 55.8% of time when present

BLAST-start: aligns with Gordonia phage Santhid\_23, function unknown, query 1-129 subject 1-129, E value = 5e-65, Identity = 92%, Similarity = 95%

GAP: TGATG overlap

LO: NA

RBS: Kibler7, spacer 6, Z score 2.562, Final -4.700, yes

F: NKF

SIF-blast: aligns with membrane protein from Gordonia phage Jojo24, ID YP\_010842561.1, query 1-129 subject 1-129, E value = 2e-78, Identity = 92%, Similarity = 95%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

SIF-Mem: 4 domains found by TMHMM, 4 domains found by SOSUI

Phages db blast shows alignment with proteins that are nkf but ncbi blast shows those same phage proteins as a membrane protein

**GENE 24 hypothetical protein**

Original Glimmer call @bp 19236 has strength 19.16; GeneMark calls start at 19257

SSC: 19257 to 19652 (Forward) – based on RBS, starterator, alignment, etc, start changed to 19257

CP: yes

SCS: both – GM

ST: SS, Start 6[19236], Found in 7 of 33 ( 21.2% ) of genes in pham, No Manual Annotations of this start. Start 7[19257], Found in 33 of 33 ( 100.0% ) of genes in pham, Called 97.0% of time when present

BLAST-start: aligns with Gordonia phage Jojo24\_24, function unknown, query 8-138 subject 1-131, E value = 5e-57, Identity = 85%, Similarity = 92% [for start 19236]

GAP: GTGA overlap [for start 19257]

LO: NA

RBS: Kibler7, spacer 10, Z score 2.402, Final -3.972, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842622.1, query 12-138 subject 4-130, E value = 5e-67, Identity = 84%, Similarity = 90%

SIF-HHPred: 8TZL\_E, Peptidase M23; membrane protein, enzyme, TRANSPORT PROTEIN; HET: ADP; {Vibrio cholerae}, Probability: 97.5%, E-value: 0.0044,

SF-Syn: NKF

SIF-Mem: 1 domain found by TMHMM, 1 domain found by SOSUI

TMHMM, SOSUI and HHPred all show as being a membrane protein but none of the other DY phage that have been annotated show this.

**GENE 25 minor tail protein**

Original Glimmer call @bp 19656 has strength 15.35

SSC: 19656 to 21356 (Forward)

CP: yes

SCS: both

ST: SS, Start 82[19656], Found in 1 of 315 ( 0.3% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Malibo\_28, minor tail protein, query 1-566 subject 1-575, E value = 0.0, Identity = 82%, Similarity = 90%

GAP: 3bp

LO: NA

RBS: Kibler7, spacer 18, Z score 2.771, Final -4.836, yes

F: minor tail protein

SIF-blast: aligns with minor tail protein from Gordonia phage Reyja, ID YP\_010842688.1, query 1-564 subject 1-559, E value = 0.0, Identity = 62%, Similarity = 72%

SIF-HHPred: 9D93\_Pc, Tail spike, gp29; Bacteriophage, tail tip, VIRAL PROTEIN; {Mycobacterium phage Bxb1}, Probability: 99.97%, E-value: 9.7e-30,

SF-Syn: shares upstream synteny with DY cluster members but downstream gene is only shared with one other phage of the DW cluster

**GENE 26 minor tail protein**

Original Glimmer call @bp 21367 has strength 12.67

SSC: 21367 to 21951 (Forward)

CP: yes

SCS: both

ST: SS, Start 1[21367], Found in 2 of 2 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_26, minor tail protein, query 1-136 subject 1-136, E value = 5e-67, Identity = 91%, Similarity = 94%

GAP: 12bp

LO: NA

RBS: Kibler7, spacer 10, Z score 3.417, Final -1.931, yes

F: minor tail protein

SIF-blast: aligns with minor tail protein from Gordonia phage Malibo, ID UAJ16206.1, query 1-149 subject 1-149, E value = 1e-58, Identity = 71%, Similarity = 81%

SIF-HHPred: 100 AAs, 9D93\_Ka, Tail collar spacer, gp6; Bacteriophage, tail tip, VIRAL PROTEIN; {Mycobacterium phage Bxb1}, Probability: 99.54%, E-value: 1.2e-13,

SF-Syn: NKF – no major synteny with other DY cluster phage

**GENE 27 minor tail protein**

Original Glimmer call @bp 21955 has strength 12.17

SSC: 21955 to 23616 (Forward)

CP: yes

SCS: both

ST: SS, Start 2[21955], Found in 2 of 3 ( 66.7% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Sampudon\_6, minor tail protein, query 27-550 subject 16-535, E value = 0.0, Identity = 81%, Similarity = 88%

GAP: 3 bp

LO: NA

RBS: Kibler7, spacer 12, Z score 1.896, Final -5.132, yes

F: minor tail protein

SIF-blast: aligns with minor tail protein from Gordonia phage Jalebi, ID UJQ86699.1, query 27-550 subject 16-535, E value = 0.0, Identity = 81%, Similarity = 88%

SIF-HHPred: matches are mainly bacterial hydrolases, no specific phage matches - 1ZMB\_E, Acetylxylan esterase related enzyme; alpha-beta protein, UNKNOWN FUNCTION; HET: MSE; 2.61A {Clostridium acetobutylicum} SCOP: c.23.10.7, l.1.1.1, Probability: 99.93%, E-value: 2.4e-25

SF-Syn: NKF– no major synteny with other DY cluster phage

**GENE 28 membrane protein**

Original Glimmer call @bp 25251 has strength 5.46

SSC: 25251 to 23620 (Reverse) – 25326 has a SLIGHTLY better RBS score

CP: yes

SCS: both

ST: Orpham, no data

BLAST-start: Orpham, no significant alignment

GAP: 77bp

LO: no

RBS: Kibler7, spacer 7, Z score 2.633, Final -4.337, yes

F: NKF

SIF-blast: Orpham, no significant alignment

SIF-HHPred: 9BUR\_A, Vitamin K-dependent gamma-carboxylase; vitamin K cycle, MEMBRANE PROTEIN, LYASE-SUBSTRATE complex; HET: CLR, NAG, POV; 2.95A {Homo sapiens}, Probability: 99.85%, E-value: 3.1e-19

SF-Syn: NKF - Orpham

SIF-Mem: 9 domains found by TMHMM, 10 domains found by SOSUI

**GENE 29 hypothetical protein**

Original Glimmer call @bp 25690 has strength 2.37; GeneMark calls start at 25747

SSC: 25690 to 25328 (Reverse) – 25690 has best RBS score

CP: yes

SCS: both

ST: Orpham, no data

BLAST-start: Orpham, no significant alignment

GAP: 95bp

LO: no

RBS: Kibler7, spacer 17, Z score 2.367, Final -5.349, yes

F: NKF

SIF-blast: Orpham, no significant alignment

SIF-HHPred: NKF, there are matches with a probability above 90% but not phage proteins and only for 60 out of 120 AAs

SF-Syn: NKF- Orpham

**GENE 30 lipid binding protein**

Original Glimmer call @bp 26147 has strength 10.45; GeneMark calls start at 26138

SSC: 26147 to 25785 (Reverse) – only 1 other phage has this gene, Tarzan. while the RBS score is similar between 26147 and 26159, 26159 would give a better alignment with Tarzan

CP: yes

SCS: both

ST: SS, Start 2[26147], found in 2 of 2 ( 100.0% ) of genes in pham, Called 50.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_29, lipid binding protein, query 1-120 subject 5-124, E value = 9e-61, Identity = 97%, Similarity = 100%

GAP: 73bp

LO: no

RBS: Kibler7, spacer 18, Z score 1.999, Final -6.389, yes

F: lipid binding protein

SIF-blast: aligns with lipid binding protein from Gordonia phage Tarzan, ID YP\_010842627.1, query 1-120 subject 5-124, E value = 4e-76, Identity = 98%, Similarity = 100%

SIF-HHPred: 7TXX\_A BA5; Bam complex, outer membrane biogenesis, SSGCID, PROTEIN; NMR {Bartonella henselae str. Houston-1}, Probability: 96.62%, E-value: 0.12

SF-Syn: NKF – only 1 other phage has this gene - Tarzan

**GENE 31 tyrosine integrase**

Original Glimmer call @bp 27041 has strength 13.47; GeneMark calls start at 26954

SSC: 27041 to 26220 (Reverse)

CP: yes

SCS: both- GM [of the potential start sites, 27041 has the lowest RBS score. 27068 is better and 27188, is also good and would give a GTGA overlap – not enough evidence to change from 27041]

ST: NI, Start 55[27041], Found in 30 of 173 ( 17.3% ) of genes in pham, Called 30.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_31, tyrosine integrase, query 1-273 subject 1-273, E value = e-155, Identity = 98%, Similarity = 99%

GAP: 44bp [from 27041]

LO: no

RBS: Kibler7, spacer 8, Z score 0.852, Final -7.618, no

F: tyrosine integrase

SIF-blast: aligns with tyrosine integrase from Gordonia phage Tarzan, ID YP\_010842629.1, query 1-273 subject 1-273, E value = 0.0, Identity = 99%, Similarity = 99%

SIF-HHPred: 5VFZ\_A Gp33; Bacteriophage, Brujita, DNA-binding, Integrase, DNA BINDING PROTEIN; HET: ACT, GOL; 1.847A {Mycobacterium phage Brujita}, Probability: 100%, E-value: 9.7e-36,

SF-Syn: does not share synteny with other DY phages

**GENE 32 immunity repressor**

Original Glimmer call @bp 27496 has strength 4.92

SSC: 27496 change to 27574 to 27185 (Reverse)

CP: yes

SCS: both - cs

ST: NI, Start 62[27496], Found in 3 of 53 ( 5.7% ) of genes in pham, Called 100.0% of time when present – possible starts = 27670, 27649, 27646, 27574, 27517, 27496 – 27649 and 27574 have best RBS scores, and 27574 would give better alignment with Tarzan gene 32.

BLAST-start: aligns with Gordonia phage Tarzan\_32, immunity repressor, query 1-103 subject 27-129, E value = 4e-55, Identity = 99%, Similarity = 99%

GAP: 165bp

LO: no

RBS: Kibler7, spacer 12, Z score 1.433, Final -6.064, yes [for 27574]

F: immunity repressor

SIF-blast: aligns with immunity repressor from Gordonia phage Reyja, ID YP\_010842694.1, query 1-103 subject 1-102, E value = 2e-61, Identity = 97%, Similarity = 98%

SIF-HHPred: for 60 of 103 Aas - 7N1N\_B, ComR; Paratox, bacteriophage, quorum sensing, Streptococcus, ComR, natural competence, VIRAL PROTEIN-TRANSCRIPTION complex; HET: EPE; 1.6A {Streptococcus pyogenes serotype M3 (strain ATCC BAA-595 / MGAS315)}, Probability: 95.94%, E-value: 0.11,

SF-Syn: This gene is pham 231449[immunity repressor], upstream is pham 231124 [tyrosine integrase], downstream is pham 229274 [helix-turn-helix DNA binding domain protein], as seen in Gordonia phage Reyja.

**GENE 33 helix-turn-helix DNA binding domain protein**

Original Glimmer call @bp 27661 has strength 6.33

SSC: 27661 to 27903 (Forward)

CP: yes

SCS: both

ST: SS, Start 3[27761], Found in 2 of 7 ( 28.6% ) of genes in pham, Called 50.0% of time when present

BLAST-start: aligns with Gordonia phage Jojo24\_32, helix-turn-helix DNA binding domain protein, query 7-80 subject 1-74, E value = 4e-35, Identity = 94%, Similarity = 97%

GAP: 165bp

LO: yes

RBS: Kibler7, spacer 10, Z score 1.708, Final -5.368, yes

F: helix-turn-helix DNA binding domain protein

SIF-blast: aligns with helix-turn-helix DNA binding domain protein from Gordonia phage Tarzan, ID YP\_010842631.1, query 14-80 subject 1-67, E value = 6e-40, Identity = 100%, Similarity = 100%

SIF-HHPred: 2AO9\_E, Phage protein; Structural genomics, nine-fold NCS., PSI, UNKNOWN FUNCTION; 1.9A {Bacillus cereus}, SCOP: a.4.1.17, Probability: 96.38%, E-value: 0.018,

SF-Syn: This gene is pham 229274 [helix-turn-helix DNA binding domain protein], upstream is pham 231449[immunity repressor], downstream is pham 8318, as seen in Gordonia phage Reyja.

**GENE 34 hypothetical protein**

Original Glimmer call @bp 27900 has strength 15.04

SSC: 27900 to 28025 (Forward)

CP: yes

SCS: both

ST: SS, Start 2[27900], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_34, function unknown, query 1-41 subject 1-41, E value = 2e-17, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 12, Z score 1.813, Final -5.299, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842632.1, query 1-41 subject 1-41, E value = 2e-19, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

SIF-Mem: 1 domain found by TMHMM, 1 domain found by SOSUI

**GENE 35 hypothetical protein**

Original Glimmer call @bp 28022 has strength 23.67

SSC: 28022 to 28240 (Forward)

CP: yes

SCS: both

ST: SS, Start 3[28022], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_35, function unknown, query 1-72 subject 1-72, E value = 9e-37, Identity = 100%, Similarity = 100%

GAP: GTGA overlap

LO: NA

RBS: Kibler7, spacer 9, Z score 2.072, Final -4.717, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Santhid, ID YP\_010842572.1, query 1-72 subject 1-72, E value = 8e-43, Identity = 99%, Similarity = 98%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 36 WhiB family transcription factor**

Original Glimmer call @bp 28240 has strength 8.86 \*\* not called by GeneMark

SSC: 28240 to 28584 (Forward)

CP: no

SCS: Glimmer only

ST: SS, Start 7 [27240], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_36, WhiB family transcription factor, query 1-114 subject 1-114, E value = 1e-63, Identity = 100%, Similarity = 100%

GAP: TGATG overlap

LO: NA

RBS: Kibler7, spacer 11, Z score 1.978, Final -4.888, yes

F: WhiB family transcription factor

SIF-blast: aligns with WhiB family transcription factor from Gordonia phage Santhid, ID YP\_010842573.1, query 1-114 subject 1-114, E value = 3e-74, Identity = 98%, Similarity = 99%

SIF-HHPred: 7KUF\_A, Probable transcriptional regulator WhiB7; Redox-sensitive, Iron-sulfur cluster, transcription, activator; 2.6A, {Mycobacterium tuberculosis}, Probability: 99.73%, E-value: 1.2e-17

SF-Syn: This gene is pham 8953 [WhiB family transcription factor], upstream is pham 8508, downstream is pham 84809, as seen in Gordonia phage Reyja.

**GENE 37 hypothetical protein**

Original Glimmer call @bp 28581 has strength 15.40

SSC: 28581 to 28874 (Forward)

CP: yes

SCS: both

ST: NI, Start 48[28581], found in 7 of 270 ( 2.6% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_37, function unknown, query 1-197 subject 1-197, E value = 4e-49, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 9, Z score 1.851, Final -5.161, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Santhid, ID YP\_010842574.1, query 1-97 subject 1-97, E value =2e-62, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 38 hypothetical protein**

Original Glimmer call @bp 28874 has strength 11.36

SSC: 28874 to 29059 (Forward)

CP: yes

SCS: both

ST: SS, Start 4[28874], Found in 2 of 3 ( 66.7% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_38, function unknown, query 1-61 subject 1-61, E value = 8e-30, Identity = 100%, Similarity = 100%

GAP: TGATG overlap

LO: NA

RBS: Kibler7, spacer 9, Z score 2.051, Final -4.759, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842636.1, query 1-61 subject 1-61, E value = 1e-35, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 39 Cas4 exonuclease**

Original Glimmer call @bp 29056 has strength 12.42

SSC: 29056 to 29817 (Forward)

CP: yes

SCS: both

ST: NI, Start 55[29056], Found in 6 of 303 ( 2.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_39, Cas4 exonuclease, query 1-253 subject 1-253, E value = e-151, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 8, Z score 1.959, Final -5.391, yes

F: Cas4 exonuclease

SIF-blast: aligns with Cas4 family exonuclease from Gordonia phage Reyja, ID YP\_010842701.1, query 1-253 subject 1-253, E value = 0.0, Identity = 97%, Similarity = 98%

SIF-HHPred: 8D3P\_I, CRISPR-associated exonuclease Cas4; CRISPR Cas adaptation type I-C, HYDROLASE-DNA complex; 4.26A, {Alkalihalobacillus halodurans C-125}, Probability: 99.71%, E-value: 4.6e-15

SF-Syn: NKF

**INSERT GENE 40**

START = ATGA - 29814

STOP = CGTGA – 29999

this corresponds to gene 40 [pham 231298] seen in phage Tarzan

BLAST shows as hypothetical protein – HHPred shows as a DNA binding protein or RNA Polymerase

**GENE 41 hypothetical protein**

Original Glimmer call @bp 29996 has strength 14.53

SSC: 29996 to 30187 (Forward)

CP: yes

SCS: both

ST: SS, Start 2[29996], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_41, function unknown, query 1-63 subject 1-63, E value = 4e-29, Identity = 100%, Similarity = 100%

GAP: GTGA overlap once new gene is inserted

LO: NA

RBS: Kibler7, spacer 9, Z score 2.261, Final -4.337, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842639.1, query 1-63 subject 1-63, E value = 3e-34, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 42 hypothetical protein**

Original Glimmer call @bp 30250 has strength 12.16

SSC: 30250 to 30930 (Forward)

CP: yes

SCS: both

ST: SS, Start 1[30250], Found in 5 of 5 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_42, function unknown, query 1-226 subject 1-226, E value = e-125, Identity = 99%, Similarity = 99%

GAP: 63bp

LO: yes

RBS: Kibler7, spacer 13, Z score 3.305, Final -2.507, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Santhid, ID YP\_010842579.1, query 1-226 subject 1-226, E value = 4e-150, Identity = 96%, Similarity = 96%

SIF-HHPred: NKF, no matches with a probability above 90% [however, several matches in the 70-90% match that states - Single-stranded DNA-binding protein]

SF-Syn: NKF

**GENE 43 hypothetical protein**

Original Glimmer call @bp 30930 has strength 7.75

SSC: 30930 to 31166 (Forward)

CP: yes

SCS: both

ST: SS, Start 5[30930], Found in 5 of 5 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_43, function unknown, query 1-78 subject 1-78, E value = 8e-38, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 11, Z score 1.930, Final -4.985, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Santhid, ID YP\_010842580.1, query 1-78 subject 1-78, E value = 3e-48, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 44 hypothetical protein**

Original GeneMark call @bp 31159

SSC: 31159 to 31341 (Forward)

CP: yes

SCS: both

ST: SS, Start 4[31159], Found in 7 of 13 ( 53.8% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_44, function unknown, query 1-60 subject 1-60, E value = 9e-29, Identity = 100%, Similarity = 100%

GAP: 7bp

LO: NA

RBS: Kibler7, spacer 14, Z score 3.002, Final -3.418, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842642.1, query 1-60 subject 1-60, E value = 2e-34, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 45 hypothetical protein**

Original Glimmer call @bp 31711 has strength 15.45; GeneMark calls start at 31696

SSC: 31711 to 31923 (Forward) – RBS favors 31969, CP favors 31711

CP: yes

SCS: both - GL

ST: Orpham, no data

BLAST-start: Orpham, no data

GAP: 31711 start gives GTGA overlap

LO: NA

RBS: Kibler7, spacer 13, Z score 1.787, Final -5.561, no

F: NKF

SIF-blast: Orpham, no data

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 46 hypothetical protein**

Original Glimmer call @bp 32021 has strength 13.01

SSC: 32021 to 32503 (Forward)

CP: yes

SCS: both

ST: SS, Start 3[32021], found in 7 of 7 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_47, function unknown, query 1-160 subject 1-160, E value = 1e-86, Identity = 100%, Similarity = 100%

GAP: 98bp

LO: yes

RBS: Kibler7, spacer 11, Z score 2.661, Final -3.514, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842645.1, query 1-160 subject 1-160, E value = 5e-112, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 47 hypothetical protein**

Original Glimmer call @bp 32500 has strength 10.05

SSC: 32500 to 32829 (Forward)

CP: yes

SCS: both

ST: SS, Start 1[32500], Found in 2 of 2 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_48, function unknown, query 1-109 subject 1-109, E value = 4e-62, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 11, Z score 1.813, Final -5.220, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842646.1, query 1-109 subject 1-109, E value = 2e-74, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 48 hypothetical protein**

Original Glimmer call @bp 32826 has strength 10.72

SSC: 32826 to 33284 (Forward)

CP: yes

SCS: both

ST: SS, Start 5[32826], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_49, function unknown, query 1-152 subject 1-152, E value = 2e-90, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer , Z score , Final -, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842647.1, query 1-152 subject 1-152, E value = 2e-103, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 49 lysin B**

Original Glimmer call @bp 33281 has strength 13.56

SSC: 33281 to 34132 (Forward)

CP: yes

SCS: both

ST: NI, Start 195[33281], Found in 9 of 646 ( 1.4% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_50, lysin B, query 1-283 subject 1-283, E value = e-164, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 15, Z score 3.084, Final -3.509, yes

F: lysin B

SIF-blast: aligns with lysin B from Gordonia phage, ID YP\_010842713.1, query 1-283 subject 1-284, E value = 0.0, Identity = 98%, Similarity = 97%

SIF-HHPred: 3HC7\_A, alpha/beta sandwich, CELL ADHESION; 2.0A {Mycobacterium phage D29}, Probability: 99.94%, E-value: 4.9e-25,

SF-Syn: This gene is pham 230934 [lysin B], upstream is pham 8760, downstream is pham223812, as seen in Gordonia phage Tarzan.

**GENE 50 hypothetical protein**

Original Glimmer call @bp 34129 has strength 7.44

SSC: 34129 to 34614 (Forward)

CP: yes

SCS: both

ST: SS, Start 4[34129], Found in 3 of 5 ( 60.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_51, function unknown, query 1-161 subject 1-161, E value = 2e-90, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 9, Z score 2.552, Final -3.751, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842649.1, query 1-161 subject 1-161, E value = 1e-109, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 51 hypothetical protein**

Original Glimmer call @bp 34611 has strength 16.27

SSC: 34611 to 34760 (Forward)

CP: yes

SCS: both

ST: SS, Start 51[34611], Found in 145 of 182 ( 79.7% ) of genes in pham, Called 99.3% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_52, function unknown, query 1-49 subject 1-49, E value = 2e-22, Identity = 100%, Similarity = 100%

GAP: GTGA overlap

LO: NA

RBS: Kibler7, spacer 11, Z score 2.159, Final -4.523, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Reyja, IDYP\_010842716.1 , query 1-49 subject 1-49, E value = 3e-25, Identity = 94%, Similarity = 97%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

SIF-Mem: 1 domain found by TMHMM, 1 domain found by SOSUI

**GENE 52 helix-turn-helix DNA binding domain protein**

Original Glimmer call @bp 34757 has strength 9.35

SSC: 34757 to 34960 (Forward)

CP: yes

SCS: both

ST: NI, Start 23[34757], Found in 6 of 9 ( 66.7% ) of genes in pham, Called 66.7% of time when present

BLAST-start: aligns with Gordonia phage Reyja\_55, helix-turn-helix DNA binding domain protein, query 1-67 subject 1-68, E value = 4e-31, Identity = 97%, Similarity = 97%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer , Z score , Final -, yes

F: helix-turn-helix DNA binding domain protein

SIF-blast: aligns with helix-turn-helix DNA binding domain protein from Gordonia phage Reyja, ID YP\_010842717.1, query 1-67 subject 1-68, E value = 2e-35, Identity = 97%, Similarity = 97%

SIF-HHPred: 3UJ3\_X, DNA-invertase; helix-turn-helix, site-specific recombinase, 3.51A, {Enterobacteria phage Mu}, Probability: 96.39%, E-value: 0.027,

SF-Syn: This gene is pham 232023 [helix-turn-helix DNA binding domain protein], upstream is pham 233375, downstream is pham 86286 [DNA primase/polymerase], as seen in Gordonia phage Jojo24.

**GENE 53 DNA primase/polymerase**

Original Glimmer call @bp 34957 has strength 16.42

SSC: 34957 to 37281 (Forward)

CP: yes

SCS: both

ST: SS, Start 15 [34957], Found in 8 of 40 ( 20.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_54, DNA primase/polymerase, query 1-774 subject 1-774, E value = 0.0, Identity = 99%, Similarity = 99%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 10, Z score 2.233, Final -4.313, yes

F: DNA primase/polymerase

SIF-blast: aligns with DNA primase/polymerase from Gordonia phage Tarzan, ID YP\_010842652.1, query 1-774 subject 1-774, E value = 0.0, Identity = 99%, Similarity = 99%

SIF-HHPred: 3M1M\_A ORF904; primase, polymerase, Replication; 1.85A, Sulfolobus islandicus,

Probability: 99.19%, E-value: 4.8e-10

SF-Syn: This gene is pham 86286 [DNA primase/polymerase], upstream is pham 232023 [helix-turn-helix DNA binding domain protein], as seen in Gordonia phage Tarzan

**INSERT GENE 54 – CORRESPONDS WITH GENE 55 FROM TARZAN**

**37278 – ATGA**

**37550 - ATGA**

**GENE 55 hypothetical protein**

Original Glimmer call @bp 37547 has strength 9.79

SSC: 37547 to 37960 (Forward)

CP: yes

SCS: both

ST: NI, Start 16[37574], Found in 10 of 36 ( 27.8% ) of genes in pham, Called 70.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_56, function unknown, query 1-137 subject 1-137, E value = 2e-79, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 10, Z score 1.265, Final -6.261, no

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Santhid, ID YP\_010842591.1, query 1-137 subject 1-137, E value = 5e-87, Identity = 93%, Similarity = 95%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 56 holliday junction resolvase**

Original Glimmer call @bp 37973 has strength 16.69

SSC: 37973 to 38404 (Forward)

CP: yes

SCS: both

ST: NI, Start 37 [37973], Found in 13 of 266 ( 4.9% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_57, holliday junction resolvase, query 1-143 subject 1-143, E value = 7e-77, Identity = 100%, Similarity = 100%

GAP: 13bp

LO: yes

RBS: Kibler7, spacer 10, Z score 1.627, Final -5.532, yes

F: holliday junction resolvase

SIF-blast: aligns with holliday junction resolvase from Gordonia phage Tarzan, ID YP\_010842655.1, query 1-143 subject 1-143, E value = 4e-96, Identity = 100%, Similarity = 100%

SIF-HHPred: 7BGS\_A, Holliday junction resolvase; archeal holliday junction resolvase helicase DNA binding enzyme phage 15-6 thermus thermophilus, RECOMBINATION; HET: SO4, MSE; 2.5A {Thermus thermophilus phage 15-6}, Probability: 99.61%, E-value: 3.3e-14,

SF-Syn: This gene is pham 217571 [holliday junction resolvase], upstream is pham 224741, downstream is pham 235049, as seen in Gordonia phage Tarzan

**GENE 57 hypothetical protein**

Original Glimmer call @bp 38401 has strength 10.30

SSC: 38401 to 38625 (Forward)

CP: yes

SCS: both

ST: NI, Start 58[38401], Found in 6 of 321 ( 1.9% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_58, function unknown, query 1-74 subject 1-74, E value = 7e-37, Identity = 100%, Similarity = 100%

GAP: ATGA overlap

LO: NA

RBS: Kibler7, spacer 12, Z score 1.642, Final -5.642, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842656.1, query 1-74 subject 1-74, E value = 3e-44, Identity = 100%, Similarity = 100%

SIF-HHPred: NKF, no matches with a probability above 90%

SF-Syn: NKF

**GENE 58 helix-turn-helix DNA binding domain protein**

Original Glimmer call @bp 38622 has strength 12.09

SSC: 38622 to 39494 (Forward)

CP: yes

SCS: both

ST: NI, Start 83[38622], Found in 7 of 111 ( 6.3% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_59, helix-turn-helix DNA binding domain protein, query 1-290 subject 1-290, E value = e-173, Identity = 100%, Similarity = 100%

GAP: GTGA overlap

LO: NA

RBS: Kibler7, spacer 6, Z score 3.156, Final -3.506, yes

F: helix-turn-helix DNA binding domain protein

SIF-blast: aligns with helix-turn-helix DNA binding domain protein from Gordonia phage Jojo24, ID YP\_010842785.1, query 1-290 subject 1-290, E value = 0.0, Identity = 99%, Similarity = 99%

SIF-HHPred: there are matches >90% but only for 50-70 amino acids out of a possible 290 - 8DGL\_B, Recombination Directionality Factor RdfS; Excisionase, Recombination Directionality Factor, winged helix-turn-helix, superhelix, DNA BINDING PROTEIN; HET: GOL; 2.45A, {Mesorhizobium japonicum R7A}, Probability: 97.23%, E-value: 0.0023

SF-Syn: This gene is pham 235246 [helix-turn-helix DNA binding domain protein], upstream is pham 235049, downstream is pham 80526, as seen in Gordonia phage Tarzan

**GENE 59 hypothetical protein**

Original GeneMark call @bp 39487

SSC: 39487 to 39594 (Forward)

CP: yes

SCS: both

ST: SS, Start 2[39487], Found in 2 of 2 ( 100.0% ) of genes in pham, Called 100.0% of time when present

BLAST-start: aligns with Gordonia phage Tarzan\_60, function unknown, query 1-35 subject 1-35, E value = 1e-10, Identity = 94%, Similarity = 94%

GAP: 7bp overlap

LO: NA

RBS: Kibler7, spacer 10, Z score 3.084, Final -2.601, yes

F: NKF

SIF-blast: aligns with hypothetical protein from Gordonia phage Tarzan, ID YP\_010842658.1, query 1-35 subject 1-35, E value = 2e-10, Identity = 94%, Similarity = 94%

SIF-HHPred: NKF, no matches with a probability above 90% [but proteins of 5-60% match show as membrane proteins]

SF-Syn: NKF

SIF-Mem: 1 domain found by TMHMM, 1 domain found by SOSUI

**GENE 60 HNH endonuclease**

Original Glimmer call @bp 39766 has strength 11.50; GeneMark calls start at 39820

SSC: 39766 to 40047 (Forward)

CP: yes

SCS: both -GL [starterator, BLAST, RBS favor 39766]

ST: SS, Start 5[39766], Found in 7 of 7 ( 100.0% ) of genes in pham, Called 57.1% of time when present

BLAST-start: aligns with Gordonia phage Jojo24\_63, HNH endonuclease, query 1-93 subject 25-117, E value = 6e-51, Identity = 94%, Similarity = 95%

GAP: 172bp

LO: no

RBS: Kibler7, spacer 9, Z score 1.848, Final -5.166, yes

F: HNH endonuclease

SIF-blast: aligns with HNH endonuclease from Gordonia phage Tarzan, ID YP\_010842661.1, query 1-93 subject 1-94, E value = 1e-42, Identity = 77%, Similarity = 81%

SIF-HHPred: 4H9D\_A, HNH endonuclease; nicking endonuclease, HYDROLASE; HET: MSE; 2.599A {Geobacter metallireducens}, Probability: 97.96%, E-value: 0.00019 – this protein also seems to match CRISPR-associated endonuclease Cas9

SF-Syn: NKF