**Gene 1**

**FINAL Gene Co-ordinates:** 42 - 374

**Original Auto-Annotation Call:** 42-374

**Glimmer:** 42; **GeneMark:** not called

**Starterator:** (Start: 90 @42 has 55 MA's): Start 90: • Found in 61 of 354 ( 17.2% ) of genes in pham; Manual Annotations of this start: 55 of 319; • Called 100.0% of time when present although does not have most annotated start.

**General:** call has good coding potential; 1:1 in top blast hits; longest ORF;

**Function**? **HNH endonuclease** [supported by all top Blast hits, HHPred, Phamerator synteny, genes in this pham are HNH endonuclease]

--------------------------------------------------------

**Gene 2**

**FINAL Gene Co-ordinates**: 448 - 702

**Original auto-annotation: 4**48 - 702

**Glimmer: 448; GeneMark: 448** (agreement)

**Starterator:** (Start: 7 @445 has 3 MA's), (Start: 8 @448 has 73 MA's) The start number called the most often in the published annotations is 8, it was called in 73 of the 76 non-draft genes in the pham. We called the most annotated start.

**General**: good coding potential; supported by extensive 1:1 top blast hits; not longest ORF but most common call.

**Function**: **terminase small subunit** [supported by all top blast hits; phamerator synteny; all genes in this pham are terminase small subunit]

-------------------------------------------------------------------

**GENE 3**

**FINAL Gene Co-ordinates:** 674 - 2161

**Original auto-annotation:** 662 -2161

**Glimmer: 662; GeneMark: 800**

**Starterator:** Has the "Most Annotated" start but did not call it; we are going with the most annotated start of 674.;: DocMcStuffins\_3 Start: 662, Stop: 2161, Start Num: 135
Candidate Starts for DocMcStuffins\_3: (Start: 135 @662 has 2 MA's), (Start: 143 @674 has 71 MA's), (Start: 215 @800 has 14 MA's),

Start 135:
• Found in 67 of 1285 ( 5.2% ) of genes in pham
• Manual Annotations of this start: 2 of 1137
• Called 4.5% of time when present
• Phage (with cluster) where this start called: BaileyBlu\_2 (FP), CallinAllBarbz\_2 (FP), DocMcStuffins\_3 (F1),

**General:** not the longest ORF but the selected start is the most called and most common start in starterator; also strong blast 1:1 conservation with top hits;

**Function:** **terminase large subunit** [strongly supported by all top blast hits and HHPred; strong synteny in phamerator; genes in this pham also call terminase large subunit

------------------------------------------

**Gene 4**

**FINAL Gene Co-ordinates:** 2202 - 3623

**Original AutoAnnotation**: 2202

**Glimmer and GeneMark** both call this start at 2202;

**Starterator:** This is = Genes that do not have the "Most Annotated" start:

Start 129:
• Found in 69 of 2002 ( 3.4% ) of genes in pham
• Manual Annotations of this start: 57 of 1787
• Called 97.1% of time when present

Gene: DocMcStuffins\_4 Start: 2202, Stop: 3623, Start Num: 129
(Start: 129 @2202 has 57 MA's)

**General:** strong coding potential in GeneMark; also supported by starterator as most called start and most common start in this pham; supported by strong blast hits for all top hits]

**Function:** **portal protein** [function supported by blast results with all top hits as portal protein; all genes in this pham are portal proteins; synteny in phamerator; also supported by HHPred top hits]

--------------------------------------

**Gene 5**

**FINAL Gene Co-ordinates:** **3,607 - 4,332**

**Original auto-annotation:** 3607

**Glimmer and GeneMark** both alled at 3607;

Starterator: The start number called the most often in the published annotations is 14, it was called in 48 of the 164 non-draft genes in the pham.

Start 14:
• Found in 55 of 195 ( 28.2% ) of genes in pham
• Manual Annotations of this start: 48 of 164
• Called 92.7% of time when present

Gene: DocMcStuffins\_5 Start: 3607, Stop: 4332, Start Num: 14
Candidate Starts for DocMcStuffins\_5:
(Start: 14 @3607 has 48 MA's), (27, 3634), (Start: 32 @3655 has 1 MA's),

**General:** strong coding potential in GeneMark; 1:1 blast conservation in all top hits; start most common for this pham in starterator]

**Function:** **capsid maturation protease;** supported by blast, phamerator, synteny

-------------------------------------------

**Gene 6:**

**FINAL Gene Co-ordinates:**4412-4948

**Original auto-annotation: 4412**

**[Glimmer and GeneMark** both call at 4412;

**Starterator:** Is a = Genes that have the "Most Annotated" start but do not call it: keeping 4412 because:

Start 20:
• Found in 68 of 248 ( 27.4% ) of genes in pham
• Manual Annotations of this start: 56 of 228
• Called 94.1% of time when present

Gene: DocMcStuffins\_6 Start: 4412, Stop: 4948, Start Num: 20 Candidate Starts for DocMcStuffins\_6:
(Start: 20 @4412 has 56 MA's),

**General:** this call produces the longest ORF; starterator supports this call - most common start for this pham; strong blast support with 1:1

**Function: scaffolding protein** [scaffolding protein supported by all top blast hits; genes in this pham are also called as scaffolding protein; phamerator synteny; some support in HHPred for scaffolding protein]

-----------------------------------------------------

**GENE 7**

**FINAL Gene Co-ordinates:5011-5919**

**Original auto-annotation: 5011**

**[Glimmer and GeneMark:** both call 5011 which is the longest ORF;

**Starterator:** The start number called the most often in the published annotations is 8, it was called in 382 of the 405 non-draft genes in the pham.

Start 8:
• Found in 424 of 453 ( 93.6% ) of genes in pham • Manual Annotations of this start: 382 of 405
• Called 100.0% of time when present

Gene: DocMcStuffins\_7 Start: 5011, Stop: 5919, Start Num: 8
(Start: 8 @5011 has 382 MA's),

**General:** consistent 1:1 Blast conservation with this start for all top hits; strong starterator conservation of this called start with 380 MA calls for this start]

**Function: major capsid protein** [Blast support with all top hits calling major capsid protein; phamerator synteny and conservation of function as major capsid **protein in this pham with vast majority calling major capsid; support from HHPred which calls major capsid with good values]**

---------------------------------------------------

**GENE 8**

**1a. FINAL Gene Co-ordinates:****5931-6302**

**Original auto-annotation: 5931**

**[Glimmer and GeneMark:** both call this start at 5931;

Starterator: is a = Genes that do not have the "Most Annotated" start:

Start 35:
• Found in 98 of 305 ( 32.1% ) of genes in pham
• Manual Annotations of this start: 72 of 284
• Called 81.6% of time when present

Candidate Starts for DocMcStuffins\_8:
(Start: 35 @5931 has 72 MA's),

**General**: strong conservation of this start with all top hits; 76MAs call this start - convincing evidence from starterator; strong blast 1:1 conservation with this start in majority of genes]

**Function:** **head-to-tail adaptor** [Strong support from blast - all top hits are head-to-tail adaptor; HHPred support for this function with low e values; consistent phamerator synteny; majority of genes in this pham call this function]

**------------------------------**

**GENE 9**

**FINAL Gene Co-ordinates**:**6299-6628**

**Original auto-annotation: 62*99***

**[Glimmer and GeneMark both call 6299;**

**Starterator:** The start number called the most often in the published annotations is 58, it was called in 207 of the 565 non-draft genes in the pham. Is a =Genes that call this "Most Annotated" start:

Start 58:
• Found in 229 of 621 ( 36.9% ) of genes in pham
• Manual Annotations of this start: 207 of 565
• Called 99.6% of time when present

Gene: DocMcStuffins\_9 Start: 6299, Stop: 6628, Start Num: 58
Candidate Starts for DocMcStuffins\_9:
(Start: 58 @6299 has 207 MA's),

**General:** strong GeneMark coding potential; longest ORF; support from starterator with 6299 having 223 MA's; all top blast hits have 1:1 at this start; conservation of this start for majority of gene in this pham]

**Function: head-to-tail stopper** [strong support from blast - all top hits have this call; genes in this pham call this function; phamerator synteny]

---------------------------------------

**GENE 10**

**FINAL Gene Co-ordinates:6635-6961 [went with most called annotation - see below]**

**Original auto-annotation: 6665**

**[Glimmer calls at 6665; GeneMark calls start at 6635;**

**Starterator:** is a = Genes that have the "Most Annotated" start but do not call it:

Start 25:
• Found in 123 of 358 ( 34.4% ) of genes in pham
• Manual Annotations of this start: 21 of 323
• Called 22.8% of time when present

Gene: DocMcStuffins\_10 Start: 6665, Stop: 6961, Start Num: 25
Candidate Starts for DocMcStuffins\_10:
(Start: 16 @6635 has 182 MA's), (Start: 25 @6665 has 21 MA's), (Start: 27 @6680 has 15 MA's),

**General:** strong coding potential in GeneMark plots; Blast support with 1:1 conservation the earlier start; also 6635 lessens gap and is supported by starterator with Start: 16 @6635 having 182 MA's and most common start]

**Function:** **No Known Function** [although some genes present in this pham assign various functions such as head to tail connector, blast and phamerator do not consistently support these calls; HHPred not informative; all genes in this pham currently call NKF]

----------------------------------

**GENE 11**

**FINAL Gene Co-ordinates:** **6951-7355**

**Original auto-annotation: 6984**

**[Glimmer calls 6984; GeneMark** calls 6951;

Start 7:
• Found in 11 of 17 ( 64.7% ) of genes in pham • Manual Annotations of this start: 4 of 9
• Called 54.5% of time when present

**Start: 20 @6951 has 192 MA's** & Start: 23 @6984 has 4 MA's);

this is = Genes that have the "Most Annotated" start but do not call it:

The start number called the most often in the published annotations is 11, it was called in 5 of the 9 non-draft genes in the pham.

**General:** Starterator supports (blast analysis shows conversation with most genes in this pham calling this start]

**function?****tail terminator [majority of genes in this pham call function as tail terminator; also supported by Blast; also supported by several HHPred hits]**

-----------------------------------------

**GENE 12**

**FINAL Gene Co-ordinates:7464 - 8273**

**Original auto-annotation: 7464**

**[Glimmer and GeneMark: both call the 7464 start;**

**Starterator:** The start number called the most often in the published annotations is 17, it was called in 221 of the 238 non-draft genes in the pham.

Start 17:
• Found in 266 of 266 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 221 of 238
• Called 93.6% of time when present

Gene: DocMcStuffins\_12 Start: 7464, Stop: 8273, Start Num: 17
Candidate Starts for DocMcStuffins\_12:
(Start: 9 @7392 has 9 MA's), (12, 7407), (Start: 17 @7464 has 221 MA's),

**General:** start also supported by blast with top hits showing 1:1; supported by starterator with Start: 33 @7464 having 220 MA's;

**Function: major tail protein** [this function is supported by blast with all top hits calling this function; HHPred results have strong hits for this function; synteny in phamerator and majority of genes in this pham all call major tail protein]

---------------

**GENE 13: Note Gene 13 and 14 have a ribosomal frame shift/slippage and was annotated accordingly in DNAMaster.**

**FINAL Gene Co-ordinates:****8399-8950**

**[Glimmer and GeneMark both call 8399;**

**Original auto-annotation: 8399**

Starterator: The start number called the most often in the published annotations is 15, it was called in 133 of the 479 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 15:
• Found in 158 of 528 ( 29.9% ) of genes in pham
• Manual Annotations of this start: 133 of 479
• Called 94.9% of time when present

Gene: DocMcStuffins\_13 Start: 8399, Stop: 8950, Start Num: 15
Candidate Starts for DocMcStuffins\_13:
(Start: 15 @8399 has 133 MA's), (Start: 17 @8405 has 64 MA's),

**General:** strong coding potential in GeneMark; longest ORF;

**Function:** **tail assembly chaperone [**strongly supported by blast - all top hits call this function; also supported by phamerator synteny; all genes in this pham call the tail assembly chaperone]

-------------------------------------------------------------------

**GENE 14 [see above] - Note Gene 13 and 14 have a ribosomal frame shift/slippage and was annotated accordingly in DNAMaster.**

**FINAL Gene Co-ordinates**:**8399 -9335**

**Original auto-annotation: 8399**

Function?**Tail assembly chaperone**

------------------------------------------

**GENE 15**

**FINAL Gene Co-ordinates:****9353-12883**

**Auto-annotation :**calls 9353;

**Glimmer and GeneMark** both call 9353;

**Starterator:** The start number called the most often in the published annotations is 2, it was called in 172 of the 206 non-draft genes in the pham. Is a = Genes that call this "Most Annotated" start:

Start 2:
• Found in 228 of 228 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 172 of 206
• Called 85.1% of time when present

Gene: DocMcStuffins\_15 Start: 9353, Stop: 12883, Start Num: 2
Candidate Starts for DocMcStuffins\_15:
(Start: 1 @9332 has 34 MA's), (Start: 2 @9353 has 172 MA's),

**General**: strong GeneMark coding potential; Blast results show strong 1:1 conservation of all top hits; Strong starterator evidence for this start: **Start: 2 @9353 has 172 MA’s]**

**Function: tape measure protein** [all genes in this pham call the tape measure protein function; blast results call tape measure function for all top hits; HHPred also has strong tape measure hits; phamerator shows synteny with similar genes]

--------------------------------------

**GENE 16**

**FINAL Gene Co-ordinates:12884-14593**

**[Auto-annotation calls** 12884;

**Glimmer and GeneMark** both call 12884;

**Starterator:** The start number called the most often in the published annotations is 21, it was called in 241 of the 481 non-draft genes in the pham. Is a =Genes that call this "Most Annotated" start:

Start 21:
• Found in 267 of 531 ( 50.3% ) of genes in pham
• Manual Annotations of this start: 241 of 481
• Called 99.6% of time when present

Gene: DocMcStuffins\_16 Start: 12884, Stop: 14593, Start Num: 21
(Start: 21 @12884 has 241 MA's),

**General:** strong GeneMark coding potential; Blast results show strong 1:1 conservation of all top hits; Strong starterator evidence for this start: **Start: 21 @12884 has 239 MA’s**

**Function:** **minor tail protein** [all genes in this pham call the minor tail protein function; blast results call minor tail protein function for all top hits; HHPred shows some support for a few hits; phamerator shows synteny with similar genes]

-------------------------------------------

**GENE 17**

**FINAL Gene Co-ordinates:****14677-16386**

**[Auto-annotation calls** 14677;

**Glimmer and GeneMark** both call 14677;

**Starterator:** Is a = Genes that have the "Most Annotated" start but do not call it:

Start 132:
• Found in 218 of 2320 ( 9.4% ) of genes in pham
• Manual Annotations of this start: 194 of 2113
• Called 98.2% of time when present

Gene: DocMcStuffins\_17 Start: 14677, Stop: 16386, Start Num: 132

(Start: 132 @14677 has 194 MA's), (Start: 228 @14809 has 2 MA's)

**General:** longest ORF; strong GeneMark coding potential; Blast results show strong 1:1 conservation of all top hits; Strong starterator evidence for this start: **Start 149 @14677 has 193 MA’s**

**Function:****minor tail protein** [all genes in this pham call the minor tail protein function; blast results call minor tail protein function for all top hits; HHPred not informative; phamerator shows synteny with similar genes]

-----------------------------------------------------

**GENE 18**

**FINAL Gene Co-ordinates**:**16444-17274**

**[Auto-annotation calls 16444;**

**Glimmer and GeneMark** both call 16444;

strong GeneMark coding potential; Blast results show strong 1:1 conservation of all top hits; Strong starterator evidence for this start (highly conserved start for this pham): **Start: 10 @16444 has 176 MA’s**

**Starterator:** The start number called the most often in the published annotations is 14, it was called in 178 of the 481 non-draft genes in the pham. Is a = Genes that call this "Most Annotated" start:

Start 14:
• Found in 268 of 531 ( 50.5% ) of genes in pham
• Manual Annotations of this start: 178 of 481
• Called 75.4% of time when present

Gene: DocMcStuffins\_18 Start: 16444, Stop: 17274, Start Num: 14
(Start: 8 @16429 has 27 MA's), (**Start: 14 @16444 has 178 MA's),**

**General:** strong GeneMark coding potential; Blast results show strong 1:1 conservation of all top hits; Strong starterator evidence for this start (highly conserved start for this pham): **Start: 10 @16444 has 176 MA’s**

**Function:minor tail protein** [all genes in this pham call the minor tail protein function; blast results call minor tail protein function for all top hits; HHPred not informative; phamerator shows synteny with similar genes]

------------------

**Gene 19**

**1a. FINAL Gene Co-ordinates:**17271-19796 (FOR)

**[Auto-annotation calls 17271;**

Glimmer and GeneMark both call 17271; strong GeneMark coding potential; longest ORF; Blast results show strong 1:1 conservation of all top hits; Strong starterator evidence for this start (highly conserved start for this pham):*(Start: 18 @17271 has 207 MA's*

**Starterator:** The start number called the most often in the published annotations is 54, it was called in 208 of the 412 non-draft genes in the pham. Is a = Genes that call this "Most Annotated" start:

Start 54:
• Found in 235 of 455 ( 51.6% ) of genes in pham
• Manual Annotations of this start: 208 of 412
• Called 96.2% of time when present

Gene: DocMcStuffins\_19 Start: 17271, Stop: 19796, Start Num: 54
(Start: 54 @17271 has 208 MA's), (64, 17292), (Start: 73 @17313 has 1 MA's)

**General:** strong GeneMark coding potential; longest ORF; Blast results show strong 1:1 conservation of all top hits; Strong starterator evidence for this start (highly conserved start for this pham):*(Start: 18 @17271 has 207 MA's*

**Function?****minor tail protein** [all genes in this pham call the minor tail protein function; blast results call minor tail protein function for all top hits; HHPred not informative; HHPred predicts moderate probability with bacterial enzyme genes, but there is stronger BLAST alignment indicating a function of minor tail proteins. phamerator shows synteny with similar genes]

**------------------**

**Gene 20**

**FINAL Gene Co-ordinates:** 19793 – 21673 (FOR);

[**Original Auto-Annotation Call***:* 19874;

**Glimmer called 19874 and GeneMark** 19793;

**Starterator:** Is a = Genes that have the "Most Annotated" start but do not call it:

Start 21: [NOT convincing! -- called 19793]
• Found in 97 of 246 ( 39.4% ) of genes in pham
• No Manual Annotations of this start.
• Called 1.0% of time when present

Gene: DocMcStuffins\_20 Start: 19874, Stop: 21673, Start Num: 21
(Start: 8 @19793 has 199 MA's),

**General:** strong coding potential for 19793 and 19793 is the longest ORF; Blast results show 1:1 conservation for many hits; Strong starterator evidence for this start (highly conserved start for this pham):There are 197 calls for the start site at bp19793

**Function?****minor tail protein** [many genes in this pham call the minor tail protein function; blast results call minor tail protein function for all top hits; HHPred not informative; HHPred is not informative but but there is stronger BLAST alignment indicating a function of minor tail proteins; the vast majority of genes in this pham are called as a minor tail protein; phamerator shows synteny with similar genes]

**Gene 21**

**1a. FINAL Gene Co-ordinates:** 21676-22848 (FOR);

**Original Auto-Annotation Call***:* 21676;

**Glimmer: 21676 - GeneMark: 21670;**

**Starterator:** It is = Genes that have the "Most Annotated" start but do not call it:

Start 6:
• Found in 201 of 204 ( 98.5% ) of genes in pham
• Manual Annotations of this start: 65 of 185
• Called 40.8% of time when present

Gene: DocMcStuffins\_21 Start: 21676, Stop: 22848, Start Num: 6
Candidate Starts for DocMcStuffins\_21:
(Start: 4 @21634 has 3 MA's), (Start: 5 @21670 has 114 MA's), (Start: 6 @21676 has 65 MA's),

**General:** gene has strong coding potential; it is not the longest ORF however, the start coordinate with the longest ORF is at bp 21634. It’s z- and final scores are more favorable than the starts predicted by Glimmer and GeneMark, however, its -40 bp overlap with the previous gene is less common than the -4 bp overlap and 2 bp gap of the starts called by Glimmer and GeneMark. There is strong synteny with upstream and immediate downstream neighbors

**Function? minor tail protein**; there is poor probability on HHPred and conflicting evidence of strong BLAST alignments; however the majrotiy of hits on blast call this as minor tail protein and the majority of genes in this pham also call it as a minor tail protein. Of the 202 members of Pham 142276, most predicted the function of this gene as a minor tail protein, however a considerable amount were described as NKF.

**------------------**

**Gene 22**

**1a. FINAL Gene Co-ordinates:** 22871 – 23212 (FOR);

**Original Auto-Annotation Call***:* 22871 - 23212;

**Glimmer and GeneMark** both call the 22871 start;

**Starterator:** It is = Genes that do not have the "Most Annotated" start:

Start 104:
• Found in 241 of 708 ( 34.0% ) of genes in pham
• Manual Annotations of this start: 216 of 643
• Called 98.8% of time when present

Gene: DocMcStuffins\_22 Start: 22871, Stop: 23212, Start Num: 104
Candidate Starts for DocMcStuffins\_22:
(40, 22673), (41, 22676), (42, 22682), (43, 22685), (48, 22709), (90, 22823), (Start: 104 @22871 has 216 MA's),

General: This gene has strong coding potential, and its start is not the longest ORF, but it has favorable RBS scores. Not the longest ORF but , the start coordinate with the longest ORF is at base pair 22673, but it has an unlikely -176 bp overlap with gene 21, and it’s z-score and final score are less favorable than those of the start at base pair 22871, which was called by both glimmer and GeneMark. Starterator does not have information on the number of calls for this start site currently.

***Function: NKF;* supported by blast, phamerator, synteny, HHPred**

***-------------------------***

**Gene 23:**

**1a. FINAL Gene Co-ordinates:** 23958 – 24212 (FOR);

**Original Auto-Annotation Call***:* 23958 - 24212;

**Glimmer and GeneMark** both call this start; 23958

**Starterator:** The start number called the most often in the published annotations is 13, it was called in 175 of the 194 non-draft genes in the pham.

It is = Genes that call this "Most Annotated" start:

Start 13:
• Found in 206 of 214 ( 96.3% ) of genes in pham
• Manual Annotations of this start: 175 of 194
• Called 92.7% of time when present

Gene: DocMcStuffins\_23 Start: 23958, Stop: 24212, Start Num: 13
Candidate Starts for DocMcStuffins\_23:
(6, 23886), (7, 23895), (**Start: 13 @23958 has 175 MA's),**

**General:** it has acceptable coding potential and its start site is not the longest ORF, but has favorable RBS scores, and has the most calls within this cluster. Strong evidence from starterator for this start: Start: 13 @23958 has 174 MA's, by far the most called and present start site.

**Function? NKF**; Due to poor probability on HHPred and conflicting evidence of strong BLAST alignments, the function is unknown. Supporting NKF is that all the genes in this Pham have called NKF; blast results split on NKF or minor tail protein; because of inconsistency we have selected NKF.

**---------------**

**Gene 24:**

**1a. FINAL Gene Co-ordinates:** 24223 – 24861(FOR);

**Original Auto-Annotation Call***:* 24223 - 24861;

**Glimmer called** 24223;**GeneMark:** 24220;

**Starterator:** The start number called the most often in the published annotations is 7, it was called in 182 of the 188 non-draft genes in the pham. It is = Genes that call this "Most Annotated" start:

Start 7:
• Found in 208 of 208 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 182 of 188
• Called 97.1% of time when present

Gene: DocMcStuffins\_24 Start: 24223, Stop: 24861, Start Num: 7
Candidate Starts for DocMcStuffins\_24:
(Start: 2 @24163 has 1 MA's), (3, 24166), (Start: 6 @24220 has 1 MA's), (**Start: 7 @24223 has 182 MA's),**

**General**: has GeneMark coding potential; strong evidence in starterator for this call - Start: 7 @24223 has 181; MA's, by far the most common call and present site; also supported by some 1:1 top blast hits.

**Function? NKF;** HHPred not informative; most genes in this pham call NKF; despite a few blast hits and members of this pham calling minor tail protein, there is too much inconsistency to call this function. Of the 206 members of Pham 388, the majority are categorized as NKF, although some are categorized as minor tail proteins, they are the overwhelming minority. There is synteny with other siilar proteins.

-------------------

**Gene 25**

 **FINAL Gene Co-ordinates:** 24858 – 25682 (FOR);

**Original Auto-Annotation Call***:* 24858 - 25682;

Glimmer and GeneMark both call 24858;

**Starterator:** The start number called the most often in the published annotations is 2, it was called in 159 of the 226 non-draft genes in the pham. It is =Genes that call this "Most Annotated" start:

Start 2:
• Found in 168 of 241 ( 69.7% ) of genes in pham
• Manual Annotations of this start: 159 of 226
• Called 100.0% of time when present

Gene: DocMcStuffins\_25 Start: 24858, Stop: 25682, Start Num: 2
(1, 24798), (Start: 2 @24858 has 159 MA's),

**General:** This gene has moderate coding potential and its not start is not the longest ORF, but it has the most calls for its cluster and favorable RBS scores. This gene has an overlap of four base pairs with the previous genes and more than ten 1:1 query: target genes under BLAST. the start site at 24858 has more favorable start sites than the start site with the longest ORF. Start: 2 @24858 has 150 MA's in starterator.

**Function? NKF;** HHPred is not informative and the vast majority of members in this pham call NKF (although a small minority call minor tail protein). Synteny present.

***----------------------------***

**Gene 26:**

**1a. FINAL Gene Co-ordinates:** 25884 – 25687 (REV);

**Original Auto-Annotation Call***:* 25884 - 25687;

**Glimmer:** 25884 **GeneMark:** 25866

**Starterator:** The start number called the most often in the published annotations is 41, it was called in 270 of the 281 non-draft genes in the pham. It is = Genes that call this "Most Annotated" start:

Gene: DocMcStuffins\_26 Start: 25884, Stop: 25687, Start Num: 41
Candidate Starts for DocMcStuffins\_26:
(**Start: 41 @25884 has 270 MA's),** (Start: 44 @25866 has 3 MA's),

Start 41:
• Found in 312 of 313 ( 99.7% ) of genes in pham
• Manual Annotations of this start: 270 of 281
• Called 95.5% of time when present

The start number called the most often in the published annotations is 41, it was called in 270 of the 281 non-draft genes in the pham.

**Geneeral:** This gene has moderate coding potential and its not start is not the longest ORF, but it has the most calls for its cluster and favorable RBS scores. This gene has an overlap of four base pairs with the previous genes and more than ten 1:1 query: target genes under BLAST. the start site at 24858 has more favorable start sites than the start site with the longest ORF. Start: 2 @24858 has 150 MA's in starterator. ;

**Function? Helix-turn-helix DNA binding domain protein;** HHPred not informative but in phamerator of the 355 members of pham142208, the majority of them are labelled as helix-turn-helix DNA binding domain protein. Others in the minority are labelled as NKF. Synteny is preserved with neighboring genes.

**-----------------**

**Gene 27:**

**FINAL Gene Co-ordinates:** 26033 – 26317 (FOR);

**Original Auto-Annotation Call***:* 26033 - 26317;

**Glimmer:** 26033 **GeneMark:** 26033;

**Starterator:** he start number called the most often in the published annotations is 40, it was called in 523 of the 752 non-draft genes in the pham. It is = Genes that call this "Most Annotated" start:

Start 40:
• Found in 569 of 819 ( 69.5% ) of genes in pham
• Manual Annotations of this start: 523 of 752
• Called 99.6% of time when present

Gene: DocMcStuffins\_27 Start: 26033, Stop: 26317, Start Num: 40 Candidate Starts for DocMcStuffins\_27:
(**Start: 40 @26033 has 523 MA's),**

General: good coding potential; 1:1 conservation with all top blast hits; this start is the longest ORF; strong support from starterator - (Start: 44 @26033 has 514 MA's).

**Function? NKF;** HHPred predicts poor probability of any function, and BLAST predicts great alignment, but there is not enough evidence to assign a function to this gene. Genes in this pham all have NKF. Of the 807 members of Pham 84614, all aside from three have no known function.

**----------------------**

**Gene 28**

**1a. FINAL Gene Co-ordinates:** 26324 – 26470 (FOR);

**Original Auto-Annotation Call***:* 26324 - 26470;

**Glimmer:** 26324; **GeneMark:** 26324;

**Starterator:** has the most calls for its cluster. (Start: 27 @26324 has 376 MA's). Genes that call this "Most Annotated" start:

Gene: DocMcStuffins\_28 Start: 26324, Stop: 26470, Start Num: 39

(Start: 39 @26324 has 381 MA's),

Start 39:
• Found in 418 of 420 ( 99.5% ) of genes in pham
• Manual Annotations of this start: 381 of 384
• Called 99.3% of time when present

The start number called the most often in the published annotations is 39, it was called in 381 of the 384 non-draft genes in the pham.

**General:** This gene has moderate coding potential, and its start site is the longest ORF

**Function?** NKF; Blast, Phamerator and HHPred all support NKF; Of the 413 members of pham 84687, all of them are described as having no known function

**-------------------------**

**Gene 29:**

**1a. FINAL Gene Co-ordinates:** 26467 – 26634 (FOR);

**Original Auto-Annotation Call***:* 26497 - 26634;

**Glimmer:** 26497; **GeneMark:** 26467;

**Starterator:** Start 7:
• Found in 93 of 214 ( 43.5% ) of genes in pham
• Manual Annotations of this start: 9 of 196
• Called 12.9% of time when present

Gene: DocMcStuffins\_29 Start: 26497, Stop: 26634, Start Num: 7 Candidate Starts for DocMcStuffins\_29:
**(Start: 4 @26467 has 74 MA's),** (Start: 7 @26497 has 9 MA's),

**General:** its start site is the longest ORF and has the most calls for its cluster;

**Function? No known function;** Considering minimal favorable predictions from HHPred and BLAST, as well as the weak coding potential of this gene, there is not enough evidence to assign a function. Of the 349 members of pham 142212, all are described as having no known function. It should be noted that this gene has one predicted TMHs.

**----------------------------**

**Gene 30:**

**FINAL Gene Co-ordinates:** 26708 – 26941 (FOR);

**Original Auto-Annotation Call***:* 26708 - 26941;

**Glimmer:** 26708

**GeneMark:** 26708;

**Starterator:** The start number called the most often in the published annotations is 16, it was called in 179 of the 220 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 16:
• Found in 216 of 245 ( 88.2% ) of genes in pham
• Manual Annotations of this start: 179 of 220
• Called 91.7% of time when present

Gene: DocMcStuffins\_30 Start: 26708, Stop: 26941, Start Num: 16
Candidate Starts for DocMcStuffins\_30:
(6, 26555), (Start: 12 @26669 has 8 MA's), (**Start: 16 @26708 has 179 MA's),**

General: its start site is the longest ORF and has the most calls for its cluster. This gene has strong coding potential, and its start site is not the longest ORF, but has the most calls of its cluster and favorable RBS scores. Starterator calls this start in vast majority of genes in this pham - **(Start: 15 @26708 has 187 MA's)**; also strong blast 1:1 conservation for top hits.

**Function? NKF;** Of the 244 members of Pham140960, all are described as having no known function. This is supported by HHPred.

**------------------**

**Gene 31:**

**1a. FINAL Gene Co-ordinates:** 26938 – 28140 (FOR);

**Original Auto-Annotation Call***:* 26938 - 28140;

**Glimmer**: 26938 **GeneMark**: 26938; **:**

**Starterator:** It is = Genes that do not have the "Most Annotated" start:

Start 28:
• Found in 27 of 104 ( 26.0% ) of genes in pham
• Manual Annotations of this start: 23 of 95
• Called 100.0% of time when present

Gene: DocMcStuffins\_31 Start: 26938, Stop: 28140, Start Num: 28
**(Start: 28 @26938 has 23 MA's),**

**General:** This gene has strong coding potential, and its start site is the longest ORF and has the most calls of its cluster; very strong 1:1 conservation with top blast hits; this is by far the most common start called in starterator for this gene - **(Start: 28 @26938 has 23 MA's) - others have no calls.**

**Function? Lysin A;** Of the 104 members of Pham 143989, the overwhelming majority of genes are described as lysin A. Blast also strongly supports Lysin A. Synteny is preserved.

**----------------**

**Gene 32:**

**FINAL Gene Co-ordinates:** 28140 – 29141 (FOR);

**Original Auto-Annotation Call***:* 28140 - 29141;

**Glimmer:** 28140 **GeneMark:** 28140;

**Starterator:** The start number called the most often in the published annotations is 106, it was called in 306 of the 1466 non-draft genes in the pham. It is : Genes that call this "Most Annotated" start:

Start 96:
• Found in 244 of 1589 ( 15.4% ) of genes in pham
• Manual Annotations of this start: 218 of 1466
• Called 98.0% of time when present

Gene: DocMcStuffins\_32 Start: 28140, Stop: 29141, Start Num: 96
(28, 27879), (58, 28041), (**Start: 96 @28140 has 218 MA's),** (Start: 130 @28215 has 2 MA's), (156, 28278), (165, 28299), (Start: 202 @28383 has 1 MA's),

**General:** Strong GeneMark coding potential; its start site is not the longest ORF and has the vast majority of calls of its cluster and favorable RBS scores; Starterator shows **(Start: 53 @28140 has 219 MA's);**  Blast shows strong 1:1 results;

**Function? Lysin B**;**:** Of the 980 members of pham 143688, the vast majority of them are described as lysin B. Blast also supports lysin B as do synteny results.

**--------------------------**

**Gene 33:**

**1a. FINAL Gene Co-ordinates:** 29151 – 29384 (FOR);

**Original Auto-Annotation Call***:* 29151 - 29384;

**Glimmer:** 29151 **GeneMark:** 29151;

**Starterator:** The start number called the most often in the published annotations is 79, it was called in 309 of the 399 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 79:
• Found in 366 of 447 ( 81.9% ) of genes in pham • Manual Annotations of this start: 309 of 399

• Called 95.1% of time when present

Gene: DocMcStuffins\_33 Start: 29151, Stop: 29384, Start Num: 79
**(Start: 79 @29151 has 309 MA's),** (Start: 80 @29154 has 4 MA's), (112, 29265),

**General:** This gene has strong coding potential, and its start site is the longest ORF and has the most calls of its cluster; Start: 73 @29151 has 308 MA's), (Start: 74 @29154 has 3 MA's),strong 1:1 support from all top blast hits.

**Function? Holin;** HHPred predicts strong probability with holin-function genes, and BLAST predicts strong alignment with similarly assigned genes, making holin the most likely function of this gene; also supported by pham calls for this gene as holin. Supported by synteny with neighboring genes in other phages.

.

**----------------------**

**Gene 34:**

**1a. FINAL Gene Co-ordinates:** 29381 – 29755 (FOR);

**Original Auto-Annotation Call***:* 29381 - 29755;

**Glimmer:** 29381 **GeneMark:** 29381;

**Starterator:** The start number called the most often in the published annotations is 25, it was called in 241 of the 717 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 25:
• Found in 278 of 782 ( 35.5% ) of genes in pham
• Manual Annotations of this start: 241 of 717
• Called 95.7% of time when present

Gene: DocMcStuffins\_34 Start: 29381, Stop: 29755, Start Num: 25
Candidate Starts for DocMcStuffins\_34:
(**Start: 25 @29381 has 241 MA's),** (Start: 32 @29396 has 1 MA's), (Start: 33 @29408 has 17 MA's),

**General:** This gene has strong coding potential, and its start site produces the longest ORF and has the most calls of its cluster. Starterator supports this start - *(Start: 38 @29381 has 240 MA's), (Start: 43 @29396 has 1 MA's), (Start: 47 @29408 has 17 MA's*. Start is also supported by 1:1 blast results with top hits;

**function? NKF;** HHPred predicts strong probability with proteins of unknown function, and BLAST predicts strong alignment; most genes in this pham call this as NKF, however ~25% do call function as minor tail protein. Given inconsistency we choose to remain with NKF; There is not enough evidence to assign a known function. Strong synteny as well.

**--------------------------------------------**

**Gene 35:**

**FINAL Gene Co-ordinates:** 29880 – 30013 (FOR);

**Original Auto-Annotation Call***:* 29885 - 30013;

**Glimmer:** 29885 **GeneMark:** 29780;

**Starterator:** Pham number 135161 has 2 members, 2 are drafts.

Start 2:
• Found in 2 of 2 ( 100.0% ) of genes in pham
• No Manual Annotations of this start.
• Called 100.0% of time when present

Gene: DocMcStuffins\_35 Start: 29885, Stop: 30013, Start Num: 2 ; (1, 29780), (2, 29885), (3, 29924), (4, 29951), (5, 29960),

**General:** This gene has moderate coding potential; We have selected the longest ORF given the lack of other starterator calls. there is only one other member of this cluster.

**Function? NKF;** there is no compelling evidence for a function from HHpred, Blast or phamerator.

**Gene 36:**

**1a. FINAL Gene Co-ordinates:** 30000 – 30821 (FOR);

**Original Auto-Annotation Call***:* 30045 - 30821;

**Glimmer:** 30000 **GeneMark:** 30045

**Starterator:** The start number called the most often in the published annotations is 18, it was called in 197 of the 439 non-draft genes in the pham.

It is: Genes that call this "Most Annotated" start:

Start 18:
• Found in 225 of 484 ( 46.5% ) of genes in pham
• Manual Annotations of this start: 197 of 439
• Called 96.4% of time when present

Gene: DocMcStuffins\_36 Start: 30000, Stop: 30821, Start Num: 18
(**Start: 18 @30000 has 197 MA's),** (Start: 32 @30045 has 4 MA's),

; **General:** This start is supported by starterator - **(Start: 18 @30000 has 196 MA's**), (Start: 32 @30045 has 4 MA's); also has support from blast with 1:1 top hits;

**Function? DnaQ-like DNA polymerase III subunit;.** HHpred predicts strong probability with exonuclease genes, while BLAST predicts great alignment with genes described as DnaQ-like DNA Polymerase. Considering the function assigned to the vast majority of members within this pham, the most likely function is DnaQ-like DNA Polymerase. Synteny conserved on phamerator.

------------------------

Gene 37:

**FINAL Gene Co-ordinates:** 30823, 30918;

**Original Auto-Annotation** - START: 30805, END: 30918, Forward

Glimmer predicts 30805 start, GeneMark predicts no start;

**Starterator:** It is: Genes that have the "Most Annotated" start but do not call it:

Start 11:
• Found in 36 of 62 ( 58.1% ) of genes in pham
• Manual Annotations of this start: 8 of 57
• Called 36.1% of time when present

Gene: DocMcStuffins\_37 Start: 30805, Stop: 30918, Start Num: 11

(Start: 11 @30805 has 8 MA's), (**Start: 13 @30823 has 42 MA's),**

**General:** Longest ORF, different call in Glimmer and GeneMark, relatively weak coding potential, 1:1 query target for top 4 genes. Not keeping original coordinates from original annotation even though there was strong evidence in one of the frames of the GeneMark output. An overlap of the size of 17 bases unlikely, so changing coordinates to (30823, 30918). Has strongest support from starterator with 41 MAs.- Start: 11 @30805 has 8 MA's), (Start: 13 @30823 has 41 MA's),

**Function: NKF;** supported by HHPred, Blast;

---------------

Gene 38:

**FINAL Gene Co-ordinates:** START: 30905, END: 31126, Forward;

AutoAnnotation: 30908

**Glimmer** predicts 30908 start, GeneMark predicts the exact same 30908 start;

**Starterator:** The start number called the most often in the published annotations is 10, it was called in 93 of the 146 non-draft genes in the pham. This is = Genes that have the "Most Annotated" start but do not call it:

Start 11:
• Found in 110 of 167 ( 65.9% ) of genes in pham
• Manual Annotations of this start: 30 of 146
• Called 33.6% of time when present

Gene: DocMcStuffins\_38 Start: 30908, Stop: 31126, Start Num: 11
**(Start: 10 @30905 has 93 MA's),** (Start: 11 @30908 has 30 MA's

**General:** not keeping coordinates from original annotation becuase only called 33% of time when there and this lessens gap; strong blast 1:1 conservation in top hits;

**FunctionL NKF; Hypothetical protein,** call supported by NCBI Blast, Phages DB and HHPred support lack of unknown function.

---------------------

Gene 39: (31119, 31211).

**FINAL Gene Co-ordinates:** 31119, 31211).

Auto annotation: 31116

GLimmer- NA Genemark: 31116:

**Starterator:** It is: Genes that have the "Most Annotated" start but do not call it:

Start 3:
• Found in 98 of 232 ( 42.2% ) of genes in pham
• Manual Annotations of this start: 20 of 210
• Called 23.5% of time when present

Gene: DocMcStuffins\_39 Start: 31116, Stop: 31211, Start Num: 3
Candidate Starts for DocMcStuffins\_39:
(Start: 3 @31116 has 20 MA's), (**Start: 4 @31119 has 186 MA's),**

**General:** Longest ORF, different call in Glimmer and GeneMark, Glimmer predicts no start, GeneMark predicts 31116 start ; 1:1 query target for top 5 genes. Keeping coordinates from original annotation even though GeneMark predicts 31116 start because of support from starterator (185 manual annotations of 31119).

**Function: NKF;** supported by HHPred low scores; NKF in other genes in this pham; and by blast results

-----------------

Gene 40:

**FINAL Gene Co-ordinates:** 31599, 32870):

**Original Autoannotation:** 31599, 32870;

**Glimmer** predicts 31599 start, GeneMark predicts the exact same 31599 start;

**Starterator:** It is: Genes that do not have the "Most Annotated" start:

Start 50:
• Found in 41 of 306 ( 13.4% ) of genes in pham
• Manual Annotations of this start: 21 of 278
• Called 58.5% of time when present

Gene: DocMcStuffins\_40 Start: 31599, Stop: 32870, Start Num: 50
Candidate Starts for DocMcStuffins\_40:
(Start: 50 @31599 has 21 MA's),

**General:** Longest ORF, same call in Glimmer and GeneMark, strong coding potential, 1:1 query target for top 2 genes. Keeping original coordinates from original annotation. (31599, 32870); strong support from 1:1 blast top hits

**Function:** **tyrosine integrase**; strongly supported by HHPred score and BLAST, called by numerous top genes that are perfectly identical; genes in this pham are all tyrosine integrase.

----------------------

Gene 41:

**FINAL Gene Co-ordinates:** START: 32875, END: 33285;

**Original autoannotation:** START: 32875, END: 33285; Reverse;

**Glimmer** predicts 33285 start, GeneMark predicts 33348;

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 16:
• Found in 222 of 222 ( 100.0% ) of genes in pham • Manual Annotations of this start: 1 of 202
• Called 1.8% of time when present

Gene: DocMcStuffins\_41 Start: 33285, Stop: 32875, Start Num: 16
(Start: 10 @33348 has 4 MA's), (11, 33339), (14, 33297), (Start: 16 @33285 has 1 MA's),

General: Keeping coordinates from original annotation even though GeneMark calls differently because starterator shows more confidence in 33285, strong coding potential. Starterator MAs only have this start: (Start: 10 @33348 has 4 MA's) in this small pham.

**Function: NKF:** supported by HHPred, blast; Also no genes in this pham call a function.

--------------------------

Gene 42:

**FINAL Gene Co-ordinates:** START: 34085, END: 33540; [keeping original start]

**Original** - START: 34085, END: 33540;

**Glimmer** predicts 34085 start, GeneMark predicts the exact same 34085 start;

**Startertor:** Genes that do not have the "Most Annotated" start:

Start 22 Found in 30 of 61 ( 49.2% ) of genes in pham
• Manual Annotations of this start: 11 of 57
• Called 36.7% of time when present

Gene: DocMcStuffins\_42 Start: 34085, Stop: 33540, Start Num: 21
(9, 34211), (Start: 14 @34151 has 1 MA's), (Start: 17 @34127 has 3 MA's), (Start: 18 @34109 has 1 MA's), (Start: 20 @34103 has 1 MA's), (21, 34085), (Start: 22 @34073 has 11 MA's),

**General:** Keeping coordinates from original annotation because same call in Glimmer and GeneMark; starterator has many different possible starts called with no consensus, therefore kept the start predicted by glimmer/genemark agreement.

**Function**: **immunity repressor;** phamerator shows same function in compared phage and 3rd top hit (immunity repressor function) in HHPred still has pretty high probability of 98.1%. Function also backed by blast results and virtually all genes in this pham have this function.

--------------

Gene 43

**FINAL Gene Co-ordinates:** START: 34343, END: 34561, Forward;

**Original** START: 34343, END: 34561, Forward;

**Different call** in Glimmer and GeneMark, Glimmer predicts 34343 start, GeneMark predicts 34241.;

**Starterator:** Start 20:
• Found in 11 of 20 ( 55.0% ) of genes in pham
• Manual Annotations of this start: 8 of 16
• Called 90.9% of time when present

Genes that call this "Most Annotated" start:

Gene: DocMcStuffins\_43 Start: 34343, Stop: 34561, Start Num: 20
(9, 34190), (10, 34205), (14, 34241), (Start: 17 @34310 has 2 MA's), (**Start: 20 @34343 has 8 MA's),**

General: 1:1 query target for top 1-2 genes. Keeping coordinates from original annotation because strong support from starterator.- (Start: 17 @34310 has 2 MA's), (Start: 20 @34343 has 8 MA's), (34343, 34561).

**Function: Cro protein** ; function because top hit in PhagesdB function freq with F1 subcluster. Most genes in this pham call this gene as a Cro protein; blast also supports this; HHPred not informative; good synteny.

------------

Gene 44

**FINAL Gene Co-ordinates:** 34616, 35476;

**Original** 34616, 35476;

**Same call** in Glimmer and GeneMark, strong coding potential,

**Starterator:** The start number called the most often in the published annotations is 2, it was called in 4 of the 4 non-draft genes in the pham. It is: Genes that do not have the "Most Annotated" start:

Start 2:
• Found in 6 of 6 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 4 of 4
• Called 100.0% of time when present

Gene: DocMcStuffins\_44 Start: 34616, Stop: 35476, Start Num: 2
(1, 3477), (Start: 2 @34616 has 4 MA's),

**General:** 1:1 query target for top gene. Keeping coordinates from original annotation. (34616, 35476). It is the only start in starterator that has multiple starts at this site; (Start: 2 @34616 has 4 MA's

**Function: antirepressor.** Function supported by HHPred, DB Blast, NCBI Blast; called by one gene in this very small pham. Synteny acceptable.

------------------

Gene 45

**FINAL Gene Co-ordinates:** START: 35479, END: 35691;

**Original** START: 35479, END: 35691;

**Glimmer** predicts 35479 start, GeneMark predicts the exact same 35479 start;

**Starterator:** Genes that call this "Most Annotated" start:

Start 2:
• Found in 6 of 6 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 4 of 4
• Called 100.0% of time when present

Gene: DocMcStuffins\_45 Start: 35479, Stop: 35691, Start Num: 2
(1, 35287), (Start: 2 @35479 has 4 MA's),

**General:** Same call in Glimmer and GeneMark, very strong coding potential, 1:1 query target for top 6 genes, keeping coordinates from original annotation with support from starterator, (35479, 35691) - (Start: 2 @35479 has 4 MA's), the only start called for multiple genes in this small pham.

**Function: NKF:** Call supported by both DB Blast and NCBI Blast; none of six members of this pham show a function.

------------------------------

Gene 46

**FINAL Gene Co-ordinates:** 35702 - 35926;

START: 35645, END: 35926; Glimmer predicts 35645 start, GeneMark predicts 35702; this is = Genes that do not have the "Most Annotated" start:

**Starterator:** Genes that do not have the "Most Annotated" start: starterator not informative - too few genes in this pham;

Start 4:
• Found in 1 of 12 ( 8.3% ) of genes in pham
• No Manual Annotations of this start.
• Called 100.0% of time when present

Gene: DocMcStuffins\_46 Start: 35645, Stop: 35926, Start Num: 4 Candidate Starts for DocMcStuffins\_46:
(4, 35645), (5, 35702), (6, 35729), (10, 35891),

 Different call in Glimmer and GeneMark, strong coding potential. Not keeping coordinates from original annotation because an overlap of the size of 47 base pairs is almost never possible. Changing coordinates to (35702, 35926) because smaller overlap, GeneMark calls the start, and one more annotation from starterator.

**Function: NKF:** no consistency in HHPred; blast supports NKF (although some hits for this pham call it as HTH binding domain protein. Some also call an excise function.

-------------------------

Gene 47

**FINAL Gene Co-ordinates:** START: 35933, END: 36115

**Original:** 35933

**Glimmer** predicts 35933, GeneMark predicts the exact same 35933 start;

**Starterator:** The start number called the most often in the published annotations is 4, it was called in 4 of the 7 non-draft genes in the pham. Genes that call this "Most Annotated" start:

Start 4:
• Found in 10 of 10 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 4 of 7
• Called 70.0% of time when present

Gene: DocMcStuffins\_47 Start: 35933, Stop: 36115, Start Num: 4
(Start: 4 @35933 has 4 MA's), (5, 36014),

**General:** (Start: 4 @35933 has 4 MA's - the most common called start;

1) 1:1 query target for 11 of top genes.

2) START: 35933, END: 36115, Forward (), longest ORF

**Function: NKF;** Supported by HHPred, blast; genes in this pham have NKF

--------------

Gene 48

**FINAL Gene Co-ordinates: S**TART: 36119, END: 36463, Forward;

Original auto annotation: 36119

**Glimmer** predicts 36119, GeneMark predicts the exact same 36119 start.

**Starterator:** This is =Genes that call this "Most Annotated" start:

Start 6:
• Found in 9 of 10 ( 90.0% ) of genes in pham
• Manual Annotations of this start: 6 of 7
• Called 100.0% of time when present

Gene: DocMcStuffins\_48 Start: 36119, Stop: 36463, Start Num: 6
(Start: 6 @36119 has 6 MA's),

**General:** Starterator support - (Start: 6 @36119 has 6 MA's - the only call with multiple MAs.

1) Longest ORF, same call in Glimmer and GeneMark, coding potential there, 1:1 query target for top 4 genes. Keeping coordinates from original annotation with support from starterator? (36119, 36463). Yes, protein coding gene.

**Function: NKF - Hypothetical protein**. HHPred, DB Blast, and NCBI Blast supports this call. Good synteny.

------------------

Gene 49

**FINAL Gene Co-ordinates:** START: 36456, END: 36770, Forward;

**Original** START: 36456, END: 36770; Forward

**Glimmer** predicts 36456 start, GeneMark predicts the exact same 36456 start.

**Starterator:** The start number called the most often in the published annotations is 30, it was called in 59 of the 103 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 30:
• Found in 65 of 114 ( 57.0% ) of genes in pham
• Manual Annotations of this start: 59 of 103
• Called 100.0% of time when present

Gene: DocMcStuffins\_49 Start: 36456, Stop: 36770, Start Num: 30
(12, 36321), (18, 36384), (19, 36402), (22, 36423), (24, 36426), (25, 36435), **(Start: 30 @36456 has 59 MA's),** (33, 36486), (34, 36495), (36, 36516), (38, 36528), (42, 36540), (44, 36561), (47, 36579), (62, 36687),

**General:** Same call in Glimmer and GeneMark, really good coding potential, 1:1 query target for top 7 genes; keeping coordinates from original annotation with additional strong support from phamerator. Support from starterator as well - (Start: 30 @36456 has 58 MA's),

**Function: NKF;** supported by HHPred, blast results; members of this pham have NKF.

------------------------

Gene 50

**FINAL Gene Co-ordinates:** START: 36767, END: 37033;

**Original** ; 36767

**Glimmer** predicts a 36767 start, Glimmer predicts the exact same 36767 start;

**Starterator:**

This is = Genes that call this "Most Annotated" start:

Start 2:
• Found in 8 of 8 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 6 of 6
• Called 100.0% of time when present

Gene: DocMcStuffins\_50 Start: 36767, Stop: 37033, Start Num: 2
Candidate Starts for DocMcStuffins\_50:
(1, 36752), (**Start: 2 @36767 has 6 MA's),**

**General:** support from starterator - Start: 2 @36767 has 6 MA's),- the only start with multiple MAs

**Function: NKF:** NKF supported by HHPred, and blast; also genes in this pham only have NKF.

---------------------------

Gene 51

**FINAL Gene Co-ordinates:** START: 37030, END: 37671, Forward;

**Original** START: 37030, END: 37671, Forward;

**Glimmer** predicts 37030 start, GeneMark predicts the exact same 37030 start.;

**Starterator:** Genes that call this "Most Annotated" start:

Start 2:
• Found in 8 of 8 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 6 of 6
• Called 100.0% of time when present

Gene: DocMcStuffins\_51 Start: 37030, Stop: 37671, Start Num: 2
(1, 36976), (**Start: 2 @37030 has 6 MA's),** (5, 37144), (6, 37207), (7, 37321), (8, 37369), (9, 37420), (10,

**General:** Same call in Glimmer and GeneMark, very strong coding potential, 1:1 query target for top gene. Keeping coordinates from original annotation. (37030, 37671). Supported by starterator - (Start: 2 @37030 has 6 MA's)- only call with any/most MAs.

**Function: NKF:**  Calling hypothetical protein with strong support from NCBI BLAST.

---------------------

Gene 52

**FINAL Gene Co-ordinates:** START: 37729, END: 37992;

Original START: 37729, END: 37992;

Glimmer predicts 37729 start, GeneMark predicts the exact same 37729 start

**Starterator:** Genes that call this "Most Annotated" start:

Start 1:
• Found in 39 of 39 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 32 of 32
• Called 97.4% of time when present

Gene: DocMcStuffins\_52 Start: 37729, Stop: 37992, Start Num: 1
(Start: 1 @37729 has 32 MA's),

**General;** Longest ORF, same call in Glimmer and GeneMark, decent coding potential, 1:1 query for target for top gene. Keeping coordinates from original annotation strongly supported by starterator. (Start: 1 @37729 has 31 MA's), - the only start with any/most MAs(37729, 37992).

**Function: NKF:** hypothetical protein function supported by HHPred and support from NCBI BLAST.

-----------------------

Gene 53

**FINAL Gene Co-ordinates:** START: 37992 END: 38351;

**Original** START: 37992, END: 38351;

**Glimmer** predicts 37992 start, GeneMark predicts the exact same 37992 start;

**Starterator:** Genes that have the "Most Annotated" start but do not call it:

Start 27:
• Found in 8 of 363 ( 2.2% ) of genes in pham
• Manual Annotations of this start: 1 of 333
• Called 37.5% of time when present

Gene: DocMcStuffins\_53 Start: 37992, Stop: 38351, Start Num: 27
Candidate Starts for DocMcStuffins\_53:
(Start: 26 @37989 has 15 MA's), (Start: 27 @37992 has 1 MA's), (42, 38019), (51, 38040), (54, 38055), (57, 38106), (61, 38127), (72, 38193), (95, 38289), (98, 38301), (106, 38340),

**General:** we are keeping the glimmer/genemark start site despite more frequent calls for an earlier call. starterator = Start: 36 @37989 has 14 MA's), (Start: 37 @37992 has 1 MA's),

1) Same call in Glimmer and GeneMark, very good coding potential from GeneMark output, almost equally across the board 1:1 query for target for top gene. Keeping coordinates from original annotation. (37992, 38351).

**Function: NKF** - supported by HHPred poor scores and blast; also members of this pham have NKF.

---------------

Gene 54

**FINAL Gene Co-ordinates:** START: 38348, END: 38545,;

**Original** START: 38348, END: 38545,

**Glimmer** predicts 38348 start, GeneMark predicts the exact same 38348 start.

**Starterator:** Genes that call this "Most Annotated" start:

Start 9:
• Found in 16 of 20 ( 80.0% ) of genes in pham
• Manual Annotations of this start: 13 of 18
• Called 93.8% of time when present

Gene: DocMcStuffins\_54 Start: 38348, Stop: 38545, Start Num: 9
(1, 38195), (**Start: 9 @38348 has 13 MA's),**

**General:** Same call in Glimmer and GeneMark, existent coding potential from GeneMark output, 1:1 query for target for top genes. Keeping coordinates from original annotation. (38348, 38545). Supported by starterator as well - (Start: 9 @38348 has 13 MA's),- the most common only call with multiple genes.

# **Function: NKF**

------------------

# Gene 55:

**1a. FINAL Gene Coordinates:** Start: 38542, Stop: 38679,

**Original Auto-Annotation Call** 38542-: 38679;

**Glimmer** start: 38542- GeneMark Start: 38545;

**Starterator:** Genes that call this "Most Annotated" start:

Start 4:
• Found in 8 of 9 ( 88.9% ) of genes in pham
• Manual Annotations of this start: 5 of 6
• Called 75.0% of time when present

Gene: DocMcStuffins\_55 Start: 38542, Stop: 38679, Start Num: 4
(Start: 4 @38542 has 5 MA's),

**General:** longest ORF; starterator supports  **start 4 in F1 cluster** (Start: 4 @38542 has 5 MA's), (5, 38545); blast support this start with 1:1 of most top hits.

**Function? NKF;** Hypothetical Protein; Function undetermined as HHpred scores include multiple different functions. Function (or lack thereof) is called by top 50 similar genes in BLAST ; synteny strong.

# -------------------

# Gene 56:

**1a. FINAL Gene Coordinates:** Start: 38676, Stop: 38936,

**Original Auto-Annotation Call*:*** 5’ end: 38676 -3’ end: 38936;

**Glimmer & GeneMark Starts***:*Glimmer start: 38676;

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 8:
• Found in 3 of 56 ( 5.4% ) of genes in pham
• Manual Annotations of this start: 1 of 51
• Called 100.0% of time when present

Gene: DocMcStuffins\_56 Start: 38676, Stop: 38936, Start Num: 8
(5, 38646), (Start: 8 @38676 has 1 MA's),

**Function*: NKF* Hypothetical Protein:** Function undetermined as HHpred scores include multiple different functions. Function (or lack thereof) is called by top 50 similar genes in BLAST; genes in this pham have NKF; synteny with other genes conserved.

-------------------

Gene 57:

**1a. FINAL Gene Coordinates:**Start: 38933;3’ end: 39349;

**Original Auto-Annotation Call*:*** 5’ end: 38933;3’ end: 39349;

**Glimmer & GeneMark Starts***:*Glimmer start: 38933- GeneMark Start: 38933;

**Starterator:** Genes that call this "Most Annotated" start:

Start 1:
• Found in 4 of 4 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 1 of 1
• Called 100.0% of time when present

Gene: DocMcStuffins\_57 Start: 38933, Stop: 39349, Start Num: 1
(Start: 1 @38933 has 1 MA's),

General: 38933;in this small pham starterator has only one MA which is at the 38333 start.

**Function: NKF;** HHPred scores support NKF as does blast

# ------------------------

# Gene 58:

**FINAL Gene Coordinates:** Start: 39346, Stop: 39699,

**Original Auto-Annotation Call*:*** 5’ end: 39346; 3’ end: 39699;

**Glimmer & GeneMark Starts***:*Glimmer start: 39346 GeneMark Start: 39346

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 20:
• Found in 9 of 277 ( 3.2% ) of genes in pham
• Manual Annotations of this start: 6 of 250
• Called 100.0% of time when present

Gene: DocMcStuffins\_58 Start: 39346, Stop: 39699, Start Num: 20
(**Start: 20 @39346 has 6 MA's),**

**General:** 1:1 conservation of top blast hits; Starterator supports this start with the most MA calls - ***?*** (Start: 20 @39346 has 6 MA's),

**1c. What is its function?** whib family transcription factor; Function is whib family transcription factor as HHpred scores include this function the most. Function is called by most of the top 50 genes with some of them being called as unknown in BLAST. Strong synteny.

# ----------------

# Gene 59:

**1a. FINAL Gene Coordinates:**Start: 39696, Stop: 39905,

**Original Auto-Annotation Call*:***5’ end: 39696-3’ end: 39905;

**Glimmer & GeneMark Starts***:* Glimmer start: 39696- GeneMark Start: 39696;

**Starterator:** Genes that call this "Most Annotated" start:

Start 16:
• Found in 16 of 16 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 12 of 12
• Called 100.0% of time when present

Gene: DocMcStuffins\_59 Start: 39696, Stop: 39905, Start Num: 16
(Start: 16 @39696 has 12 MA's),

**General:** strong coding potential and strong blast conservation - 1:1 blast results for top hits; Support from starterator - this is start is most common

 (Start: 16 @39696 has 12 MA's),

**Function? NKF;** ; No function listed or known l as high variation in HHpred scores. Function is unknown by top 50 genes in BLAST.

-------------------------

# Gene 60:

**1a. FINAL Gene Coordinates:**Start: 39902 Stop: 40159,

**Original Auto-Annotation Call*:*** 5’ end:

**Glimmer & GeneMark Starts***:*Glimmer start: 39902 genemark 39902

The start number called the most often in the published annotations is 12, it was called in 5 of the 6 non-draft genes in the pham.

**Starterator:** Genes that call this "Most Annotated" start:

Start 12:
• Found in 8 of 8 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 5 of 6
• Called 87.5% of time when present

Gene: DocMcStuffins\_60 Start: 39902, Stop: 40159, Start Num: 12
(1, 39569), (2, 39602), (3, 39710), (4, 39722), (5, 39743), (6, 39764), (7, 39818), (8, 39857), (9, 39872), (10, 39875), (Start: 12 @39902 has 5 MA's),

**General**: longest ORF; Query and target is 1:1 for top 2 genes; starterator has 39902 as most common start;

**Function: NKF;**. No function listed or known as high variation in HHpred scores as well as no data present under function frequency. Function is unknown by top 50 genes in BLAST. Considerable synteny.

--------------------

# Gene 61:

**1a. FINAL Gene Coordinates:** Start: 40244, Stop: 40576,

**Original Auto-Annotation Call*:***  5’ end: 40244- 3’ end: 40576;

**Glimmer & GeneMark Starts***:*  Glimmer start: 40244- GeneMark

**Starterator:** Start: 40244; Moderate support in starterator -  (Start: 7 @40205 has 154 MA's), (Start: 8 @40235 has 1 MA's), (Start: 10 @40244 has 54 MA's),

Genes that have the "Most Annotated" start but do not call it:

Start 10:
• Found in 262 of 262 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 54 of 243
• Called 25.2% of time when present

Gene: DocMcStuffins\_61 Start: 40244, Stop: 40576, Start Num: 10
(Start: 7 @40205 has 155 MA's), (Start: 8 @40235 has 1 MA's), (Start: 10 @40244 has 54 MA's),

**Function? mpme1 protein;** Function listed as mpme 1 protein as it is the most frequent function listed. Function varies between mpme 1 protein, mpme2 protein, and unknown in top 50 genes in BLAST. ;

---------------------

# Gene 62:

**1a. FINAL Gene Coordinates:**Start: 40563, Stop: 40733,

**Original Auto-Annotation Call*:*** 5’ end: 40563-3’ end: 40733;

**Glimmer & GeneMark Starts***:* Glimmer start: 40563; GeneMark Start: 40611;

**Starterator:** Genes that call this "Most Annotated" start:

Start 3:
• Found in 6 of 6 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 3 of 4
• Called 83.3% of time when present

DocMcStuffins\_62 Start: 40563, Stop: 40733, Start Num: 3 Candidate Starts for DocMcStuffins\_62:
(1, 40488), (2, 40509), **(Start: 3 @40563 has 3 MA's),** start 4 @40611 has 1 MA

**General:**  Not the longest ORF. Longest ORF entails a -89 gap. Weak coding potential with strong correlation between start and stop of gene. Different calls for Glimmer and Genemark start.

**Function: NKF;** Function listed as unknown amongst top 50 genes in BLAST. High variation between query and target amongst top 3 genes.

----------------

# Gene 63:

**1a. FINAL Gene Coordinates:**Start: 40730, Stop: 40885,

**Original Auto-Annotation Call*:*** 5’ end: 40730***-***3’ end: 40855;

**Glimmer & GeneMark Starts***:* Glimmer start: 40730*-* GeneMark Start: 40733;

**Starterator:** Genes that call this "Most Annotated" start:

Gene: DocMcStuffins\_63 Start: 40730, Stop: 40885, Start Num: 9
(1, 40577), (**Start: 9 @40730 has 28 MA's),** (Start: 10 @40733 has 23 MA's),

**General:** strong blast 1:1 alignment of top blast hits; support from starterator ), (Start: 9 @40730 has 28 MA's), (Start: 10 @40733 has 22 MA's);

**Function? NKF -** Hypothetical Protein ; NKF; Function listed as unknown amongst top 50 genes in BLAST, however there is data for function frequency (function listed as whib family transcription factor)..

# Gene 64:

**FINAL Gene Coordinates:**Start: 40882, Stop: 41045, Length: 165;

**Original Auto-Annotation Call*:*** 5’ end: 40882-3’ end: 41046;

**Glimmer & GeneMark Starts***:*Glimmer start: 40882-GeneMark Start: 40882; Support from starterator -(Start: 1 @40882 has 5 MA's), (3, 40954),

**Starterator:** Genes that call this "Most Annotated" start:

Start 1:
• Found in 7 of 7 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 5 of 5
• Called 100.0% of time when present

Gene: DocMcStuffins\_64 Start: 40882, Stop: 41046, Start Num: 1
(Start: 1 @40882 has 5 MA's), (3, 40954),

**Function? NKF;** Hypothetical Protein ;Function listed as unknown amongst top 50 genes in BLAST and no data for function under function frequency. HHpred calls for hypothetical protein function among top 3 most common genes. strong synteny

---------------

# Gene 65:

**FINAL Gene Coordinates:**Start: 41043, Stop: 41897, Length: 855; Longest ORF. ;

**Original Auto-Annotation Call*:*** 5’ end: 41043-3’ end: 41897;

**Glimmer & GeneMark Starts***:*Glimmer start: 41043*-*GeneMark Start: 41043; Support from starterator - (Start: 93 @41043 has 6 MA's), - most commonly called start;

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 114:
• Found in 8 of 299 ( 2.7% ) of genes in pham
• Manual Annotations of this start: 6 of 273
• Called 100.0% of time when present

Gene: DocMcStuffins\_65 Start: 41043, Stop: 41897, Start Num: 114
(Start: 114 @41043 has 6 MA's),

**Function? Cas4 family exonuclease.** Function varies among top 50 genes which includes unknown function in BLAST, there is also data for function frequency, with rece-like exonuclease being the most frequent function (20%). Rare start in Starterator.

# -----------------

# Gene 66:

**FINAL Gene Coordinates:**Start: 41891, Stop: 42907, strong coding potential;

**Original Auto-Annotation Call*:*** 5’ end: 41891-3’ end: 42907;

**Original Auto-Annotation Call*:*** 5’ end: 41891-3’ end: 42907;

**Glimmer & GeneMark Starts***:*Glimmer start: 41891-GeneMark Start: 41891; strong blast alignment 1:1 for all top hits; Start: 5 @41891 has 4 MA's);

**Starterator:** Genes that have the "Most Annotated" start but do not call it:

Start 5:
• Found in 9 of 21 ( 42.9% ) of genes in pham
• Manual Annotations of this start: 4 of 17
• Called 77.8% of time when present

Gene: DocMcStuffins\_66 Start: 41891, Stop: 42907, Start Num: 5
(2, 41801), (3, 41864), (Start: 5 @41891 has 4 MA's), (Start: 6 @41894 has 12 MA's),

**function? NKF:** Function is unknown for the top 50 genes in BLAST, strong syntency.

# **----------------------**

# **Gene 67:**

**1a. FINAL Gene Coordinates:**Start: 42904, Stop: 43389,

**Original Auto-Annotation Call*:*** 5’ end: 42904-3’ end: 43389;

**Glimmer & GeneMark Starts***:*Glimmer start: 42904-GeneMark Start: 42904;

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 16:
• Found in 13 of 36 ( 36.1% ) of genes in pham
• Manual Annotations of this start: 10 of 32
• Called 100.0% of time when present

Gene: DocMcStuffins\_67 Start: 42904, Stop: 43389, Start Num: 16 Candidate Starts for DocMcStuffins\_67:

(Start: 16 @42904 has 10 MA's),

***General:*** blast 1:1 conservation for top hits; strong support from starterator - (Start: 16 @42904 has 10 MA's),is the most common and only call with multiple MAs.

**function? NKF;** Hypothetical Protein Function is unknown for the top 50 genes in BLAST, Strong synteny.

# **---------------------**

# **Gene 68:**

**1a. FINAL Gene Coordinates:**Start: 43386, Stop: 43640, Length: 255; .

**Original Auto-Annotation Call*:***5’ end: 43386- 3’ end: 43640;

**Glimmer & GeneMark Starts***:*Glimmer start: 43386*-* GeneMark Start: 43386;

**Starterator:** Genes that call this "Most Annotated" start:

Start 3:
• Found in 87 of 87 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 78 of 78
• Called 100.0% of time when present

Gene: DocMcStuffins\_68 Start: 43386, Stop: 43640, Start Num: 3 Candidate Starts for DocMcStuffins\_68:
(Start: 3 @43386 has 78 MA's), (5, 43431), (10, 43494),

***General:*** strong blast 1:1 alingment; strong support from starterator - (Start: 3 @43386 has 76 MA's),Same call for Glimmer and Genemark start. Query and target is 1:1 for top genes #1 and #2 and different but similar query and target for top gene #3.

**function? NKF;** Hypothetical Protein; Function is unknown for the top 50 genes in BLAST, there is data for function frequency,

# -----------------------

# Gene 69:

**1a. FINAL Gene Coordinates:** Start: 43637, Stop: 44014,

**Original Auto-Annotation Call*:*** 5’ end: 43637***-***3’ end: 44014;

**Glimmer & GeneMark Starts***:*Glimmer start: 43637-GeneMark Start: 43637;

**Starterator:** Genes that have the "Most Annotated" start but do not call it:

Start 84:
• Found in 35 of 243 ( 14.4% ) of genes in pham
• Manual Annotations of this start: 25 of 212
• Called 88.6% of time when present

Gene: DocMcStuffins\_69 Start: 43637, Stop: 44014, Start Num: 84
(Start: 84 @43637 has 25 MA's), (103, 43706), (116, 43742), (127, 43760), (155, 43901), (158, 43925), (174, 43982), (178, 44006),

**General:** longest ORF; Same call for Glimmer and Genemark start. Query and target is 1:1 for top genes; (Strong support from starterator - only called MA - Start: 84 @43637 has 25 MA's), longest ORF; Same call for Glimmer and Genemark start. Query and target is 1:1 for top genes; (Strong support from starterator - only called MA - Start: 84 @43637 has 25 MA's),

**function? RusA-like resolvase; General:**. Function varies between unknown RusA-like resolvases for the top 50 genes in BLAST, there is also data for function frequency, with RusA-like resolvase being the most frequent function (19%). Strong synteny

**--------------------------**

# Gene 70:

**1a. FINAL Gene Coordinates:**Start: 44011, Stop: 44223, Length: 213;

**Original Auto-Annotation Call*:*** 5’ end: 44011***-***3’ end: 44223; **Original Auto-Annotation Call*:*** 5’ end: 44011***-***3’ end: 44223;

**Glimmer & GeneMark Starts***:*Glimmer start: 44011-GeneMark Start: 44011;;

**Starterator:** Genes that call this "Most Annotated" start:

Start 8:
• Found in 12 of 18 ( 66.7% ) of genes in pham
• Manual Annotations of this start: 9 of 14
• Called 100.0% of time when present

Gene: DocMcStuffins\_70 Start: 44011, Stop: 44223, Start Num: 8
(5, 43975), (Start: 8 @44011 has 9 MA's),

**General:** strong blast 1:1 conservation; support from starterator - (Start: 8 @44011 has 9 MA's), - -only MA called start site in starterator

Function: NKF:

# -----------------

# Gene 71:

**FINAL Gene Coordinates:**Start: 44220, Stop: 44615,

**Original Auto-Annotation Call*:*** 5’ end: 44220***-***3’ end: 44615;

**Glimmer & GeneMark Starts***:*Glimmer start: 44220*-*GeneMark Start: 44220Same call for Glimmer and Genemark start.

**Starterator:** The start number called the most often in the published annotations is 17, it was called in 40 of the 42 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 17:
• Found in 47 of 47 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 40 of 42
• Called 95.7% of time when present

Gene: DocMcStuffins\_71 Start: 44220, Stop: 44615, Start Num: 17
(6, 44094), (8, 44118), (9, 44127), (13, 44154), (Start: 17 @44220 has 40 MA's),

**General:** Query and target is 1:1 for top genes. Strog support from starterator - Start: 17 @44220 has 40 MA's),- only start with MAs.

**function? NKF; Hypothetical Protein** ;Function is unknown for the top 50 genes in BLAST, there is also data for function frequency, with dna methylase being the most frequent function (50%). Common start in Starterator.

------------------

# Gene 72:

**FINAL Gene Coordinates:**Start: 44615, Stop: 46129, Length: 1515;

**Original Auto-Annotation Call*:*** 5’ end: 44615***-*** 3’ end: 46129;

**Glimmer & GeneMark Starts***:*Glimmer start: 44615-GeneMark Start: 44615;

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 75:
• Found in 32 of 150 ( 21.3% ) of genes in pham
• Manual Annotations of this start: 27 of 121
• Called 96.9% of time when present

Gene: DocMcStuffins\_72 Start: 44615, Stop: 46129, Start Num: 75
(57, 44531), (68, 44573), (**Start: 75 @44615 has 27 MA's),**

General: strong support from starterator - **(Start: 77 @44615 has 26 MA's),-only start site with multiple calls;** Strong coding potential with strong correlation between start and stop of gene present. Same call for Glimmer and Genemark start. Query and target is 1:1

**function? DNA methylase;** Function varies between DNA methylase and unknown for the top 50 genes in BLAST, there is also data for function frequency, with dna methylase being the most frequent function (25%).

**------------**

**GENE 73**

**1a. FINAL Gene Co-ordinates: *46126 46713;***

***Original 46126 46713;***

***Glimmer*** *Start: 46126 GeneMark Start: 46126;*

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 35:
• Found in 6 of 70 ( 8.6% ) of genes in pham • Manual Annotations of this start: 5 of 61
• Called 100.0% of time when present

Gene: DocMcStuffins\_73 Start: 46126, Stop: 46713, Start Num: 35
(17, 46072), (**Start: 35 @46126 has 5 MA's),** (38, 46144), (47, 46177), (50, 46186), (Start: 58 @46201 has 1 MA's), (79, 46282), (96, 46348), (121, 46420), (133, 46471), (145, 46522), (157, 46573), (161, 46597), (171, 46654),

*General: support from starterator -* Start: 31 @46126 has 10 MA's),;**Glimmer and GeneMark predict the same start site. Strong blast 1:1 support;**

**What is its function? *-> HNH endonuclease;***This gene is a protein-coding gene its function being an HNH endonuclease. NCBI Blast, Blast alignment, HHPRED, and CCD give evidence for function of HNH Endonuclease. There are synteny and conserved domains.

**-----------**

**Gene 74**

**FINAL Gene Co-ordinates: *46780 50322 ;***

***Original- 46780 50322 ;***

***Glimmer*** *Start: 4678; GeneMark Start: 46780***,**

**Starterator:** Start 32:
• Found in 13 of 90 ( 14.4% ) of genes in pham
• Manual Annotations of this start: 4 of 72
• Called 38.5% of time when present

Gene: DocMcStuffins\_74 Start: 46780, Stop: 50322, Start Num: 32
(Start: 32 @46780 has 4 MA's),

Genes that do not have the "Most Annotated" start:

**General:** glimmer and GeneMark and DNA Master predict the same start site. The chosen coordinates are the longest ORF; strong blast 1:1 alignment in all top hits;

**function? *Dna methylase;*** NCBI Blast alignment, and HHPREDshows similarity to multiple DNA methylase genes. 7 of top ten similar genes call DNA methylase. Conserved domains are present, and the downstream gene is the same, showing synteny.

-------------------------------------------------------------------

Gene 75

**FINAL Gene Co-ordinates: *50339 50997;***

***Original 50339 50997;***

***Glimmer*** *Start: 50330*-*GeneMark Start: 50339;*

**Starterator:** The start number called the most often in the published annotations is 4, it was called in 3 of the 3 non-draft genes in the pham. It is : Genes that call this "Most Annotated" start:

Start 4:
• Found in 5 of 5 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 3 of 3
• Called 100.0% of time when present

Gene: DocMcStuffins\_75 Start: 50339, Stop: 50977, Start Num: 4
(3, 50330), (**Start: 4 @50339 has 3 MA's**

**General:** strong support from starterator; Starterator shows this start at 50339 is call in 100% of genes in this pham and by all MA. Strong support from blast - top hits all 1:

**function? *NKF;*** *Unknown function -* NCBI BLAST, and HHPRED call this a gene with no known function. BLAST alignment shows all the top 10 most similar genes are called no known function. Strong synteny.

------------------------

Gene76

**FINAL Gene Co-ordinates: *51114- 51512;***

***Original Call: 51114- 51512;***

***Glimmer*** *Start: 51114; GeneMark Start: 51114*

**Starterator:** This gene = Genes that do not have the "Most Annotated" start:

Start 8:
• Found in 3 of 8 ( 37.5% ) of genes in pham
• Manual Annotations of this start: 1 of 5
• Called 66.7% of time when present

Gene: DocMcStuffins\_76 Start: 51114, Stop: 51512, Start Num: 8
(Start: 2 @51042 has 1 MA's), (3, 51048), (4, 51054), (Start: 8 @51114 has 1 MA's), (10, 51153), (12, 51198), (13, 51219), (14, 51231), (15, 51258), (17, 51288), (18, 51291), (19, 51330), (20, 51333), (21, 51402),

**Function? *NKF:*** Hypothetical protein; supported by blast, HHPred; unspecified in all 8 of this Pham

--------------------------------------

Gene 77

**FINAL Gene Co-ordinates: *51509 - 51712;***

***Original 51509 - 51712;***

***Glimmer*** *Start: 51509 - GeneMark Start: 51509;*

**Starterator:** The start number called the most often in the published annotations is 9, it was called in 17 of the 17 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 9:
• Found in 21 of 22 ( 95.5% ) of genes in pham
• Manual Annotations of this start: 17 of 17
• Called 100.0% of time when present

Gene: DocMcStuffins\_77 Start: 51509, Stop: 51712, Start Num: 9
Candidate Starts for DocMcStuffins\_77:
(**Start: 9 @51509 has 17 MA's**), (12, 51533), (13, 51551), (14, 51581), (16, 51665), (17, 51674), (18, 51695), (19, 51698),

***General:*** Starterator showed this start is called in 95.7% of genes in the pham and by 17/17 MA. These coordinates are the longest ORF.

**Function? NKF:** Support from Blast, HHPred; Final Summary:- NCBI BLAST shows similarity to a hypothetical protein and BLAST alignment shows most similar genes called at proteins of unknown function. Genes in the pham are NKF;

---------------------

**Gene** 78

**FINAL Gene Co-ordinates:***51709 -* **51876; YES >>>>>>>>**

**Original: 51679 - 51876;**

**Glimmer** *Start: 51679; GeneMark Start: 51709;*

**Starterator:** Start 10: • Found in 1 of 2 ( 50.0% ) of genes in pham • No Manual Annotations of this start. • Called 100.0% of time when present •

Start 10:
• Found in 1 of 2 ( 50.0% ) of genes in pham
• No Manual Annotations of this start.
• Called 100.0% of time when present

Gene: DocMcStuffins\_78 Start: 51679, Stop: 51876, Start Num: 10
(3, 51337), (4, 51400), (5, 51448), (6, 51460), (8, 51583), (10, 51679), (11, 51709), (13, 51739), (14, 51808),

**General:** Glimmer calls the start at 51679 and GeneMark calls the start at 51709. Neither of these starts is the longest ORF. 51679 creates an overlap of 34bp and 51709 creates an overlap of 4bp

**Function? *NKF*** Blast split between HNH endonuclease and NKF -> Unspecified in the 2 members of this pham

-----------------------------------------------------

Gene 79

**1a. FINAL Gene Co-ordinates: 51876-52019;**

**Original Auto-Annotation Call*:* 51876-52019;**

***Glimmer*** *Start: 51876-GeneMark Start: 51876;* Start 26: • Found in 27 of 45 ( 60.0% ) of genes in pham • Manual Annotations of this start: 21 of 36 • Called 100.0% of time when present •; **Starterator showed this start is called in 60% of genes in the pham with 21 MA and that it is called 100% of the time when present.**

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 64:
• Found in 32 of 194 ( 16.5% ) of genes in pham
• Manual Annotations of this start: 27 of 175
• Called 100.0% of time when present

Gene: DocMcStuffins\_79 Start: 51876, Stop: 52019, Start Num: 64 Candidate Starts for DocMcStuffins\_79:
(10, 51567), (**Start: 64 @51876 has 27 MA's),** (77, 51972),

**General:**Found in 27 of 45 ( 60.0% ) of genes in pham • Manual Annotations of this start: 21 of 36 • Called 100.0% of time when present •; Starterator showed this start is called in 60% of genes in the pham with 21 MA and that it is called 100% of the time when present.

**Function?NKF *;*** Unknown function; NCBI BLAST and BLAST alignment show similarity to genes encoding proteins of unknown function. -> unspecified of 194 pham ;

-----------------

Gene 80

**FINAL Gene Co-ordinates: *52*016 - 52156*;***

**Original Auto-Annotation Call*: 52016- 5215;***

***Glimmer***Start: 52016-GeneMark Start: 52016; strong blast 1:1 conservation; longest ORF;

**Starterator:** The start number called the most often in the published annotations is 58, it was called in 111 of the 122 non-draft genes in the pham. This is: Genes that call this "Most Annotated" start:

Start 58:
• Found in 132 of 134 ( 98.5% ) of genes in pham
• Manual Annotations of this start: 111 of 122
• Called 93.2% of time when present

Gene: DocMcStuffins\_80 Start: 52016, Stop: 52156, Start Num: 58 (29, 51887), (32, 51896), (Start: 58 @**52016 has 111 MA's),** (61, 52058),

***General:*** Start 58: • Found in 154 of 159 ( 96.9% ) of genes in pham • Manual Annotations of this start: 129 of 143 • Called 93.5% of time when present •

**Function? *NKF;*** Hypothetical protein 🡪 unspecified of 134 pham; BLAST showed conservation with 175 similar phage.

----------------------------------------------

Gene 81

**FINAL Gene Co-ordinates:52185 - 53102;**

**Original 52185 - 53102;**

**Glimmer & GeneMark Start-** *Glimmer Start: 52185GeneMark Start: 52185;*

**Starterator:** This is = Genes that do not have the "Most Annotated" start:

Start 8:
• Found in 6 of 28 ( 21.4% ) of genes in pham
• Manual Annotations of this start: 4 of 26
• Called 100.0% of time when present

Gene: DocMcStuffins\_81 Start: 52185, Stop: 53102, Start Num: 8
**(Start: 8 @52185 has 4 MA's),** (11, 52230), (12, 52233), (15, 52278), (18, 52323), (22, 52368), (23, 52377), (24, 52380), (27, 52443), (32, 52593), (34, 52641), (39, 52824), (42, 52860),

**General:** *strong blast conservation;*

**Function**? NKF; Unknown Function -> no strong support from blast or HHPred; 3/28 RepA-like replication initiator

***-------------------------***

Gene 82

**1a. FINAL Gene Co-ordinates: *53099 to 53470***

**Original Auto-Annotation Call*:* 53099-53470;**

**Glimmer & GeneMark Starts***: Glimmer Start: 53009-GeneMark Start: 53099;*

**Starterator:** The start number called the most often in the published annotations is 2, it was called in 72 of the 72 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 2:
• Found in 82 of 82 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 72 of 72
• Called 100.0% of time when present

Gene: DocMcStuffins\_82 Start: 53099, Stop: 53470, Start Num: 2
(1, 52910), (Start: 2 @53099 has 72 MA's), (4, 53213), (8, 53255), (10, 53291),

**General:****Startarator** Start 2: • Found in 82 of 82 ( 100.0% ) of genes in pham • Manual Annotations of this start: 71 of 71 • Called 100.0% of time when present

**Function? NKF;** Unknown function; supported by HHPrred, blast and phamerator; 90% of genes in this pham have NKF;-> 7/82 DnaC-like helicase loader

----------------------------------

**Gene 83**

**FINAL Gene Co-ordinates:53513 - 54418**

**Original Auto-Annotation Call*:* 53513 - 54418**

**Glimmer & GeneMark Starts***: Glimmer Start: 53513 - GeneMark Start: 53513*

**Starterator:**

Genes that do not have the "Most Annotated" start:

Start 46:
• Found in 64 of 257 ( 24.9% ) of genes in pham
• Manual Annotations of this start: 58 of 242
• Called 98.4% of time when present

Gene: DocMcStuffins\_83 Start: 53513, Stop: 54418, Start Num: 46
(**Start: 46 @53513 has 58 MA's),** (72, 53591), (128, 53876), (140, 53927), (147, 53975), (151, 54005), (184, 54119), (196, 54185), (221, 54323), (224, 54329), (230, 54353), (235, 54377), (239, 54386),

Blast: strong 1:1 conservation in top hits

***Function:*** helix-turn-helix DNA binding domain protein: strong synteny;

 **-----------------------------------------**

**Gene 84**

**FINAL Gene Co-ordinates:**54595- ***54999 YES.........***

**Glimmer & GeneMark Starts***: Glimmer Start:* 54595 - *GeneMark Start:* 54595

**f**

Starterator: The start number called the most often in the published annotations is 2, it was called in 3 of the 4 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 2:
• Found in 6 of 6 ( 100.0% ) of genes in pham
• Manual Annotations of this start: 3 of 4
• Called 83.3% of time when present

Gene: DocMcStuffins\_84 Start: 54595, Stop: 54999, Start Num: 2

(Start: 1 @54502 has 1 MA's), **(Start: 2 @54595 has 3 MA's),** (3, 54676), (4, 54766), (5, 54820),

***General:*** The co-coordinates choice is 54502 to 54999. This gene has strong coding potential according to Host-Trained GeneMark. GeneMark/ Glimmer call a different start at 54595. Starterator showed this start is called in 97.6% of genes in the pham however most of phage this gene is called in are in the P1 cluster and this start does not minimize the gap. Start 54502 minimizes the gap from 176 to 83 and Starterator shows it is called in similar phage Awesomesauce. This is a hypothetical protein because NCBI BLAST and BLAST alignment show similarity to genes that have this function, meaning it is a protein coding gene we just don't know its function yet. Finally, there is syteny and phamerator.

**Function?NKF;** Supported by blast and HHPred;Unspecified of 6 member of this pham

**--------------------------------------------------------**

**Gene 85**

**FINAL Gene Co-ordinates: *54996 - 55400***

**Glimmer & GeneMark Starts***: Glimmer Start: 54996; GeneMark Start: 54996*

**Starterator:** This is = Genes that do not have the "Most Annotated" start:

Start 59:
• Found in 14 of 304 ( 4.6% ) of genes in pham
• Manual Annotations of this start: 11 of 273
• Called 100.0% of time when present

Gene: DocMcStuffins\_85 Start: 54996, Stop: 55400, Start Num: 59 Candidate Starts for DocMcStuffins\_85:
(25, 54798), (27, 54813), (**Start: 59 @54996 has 11 MA's),**

***General:***This gene has strong coding potential, and NCBI BLAST and BLAST alignment show related genes have an unknown function. Glimmer start, GeneMark, and DNA Master call the same start. NCBI Blast and alignment with similar genes shows that there are many similar genes called as hypothetical proteins. Starterator showed this start is called in 30.7% of genes in the pham with 84 MA and it is called 94.5% of the time when present. There is syteny, and conserved domains.

**Function? *NKF;*** Hypothetical protein - unspecified in 304 member of this pham

------------------------------------------------------------------

Gene 86

**FINAL Gene Co-ordinates: *55397 - 55639***

***Original: 55397 - 55639***

**Glimmer & GeneMark Starts***: Glimmer Start: 55397 ; GeneMark Start: 55397*

**Starterator:** The start number called the most often in the published annotations is 11, it was called in 203 of the 267 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 11:
• Found in 223 of 298 ( 74.8% ) of genes in pham
• Manual Annotations of this start: 203 of 267
• Called 100.0% of time when present

Gene: DocMcStuffins\_86 Start: 55397, Stop: 55639, Start Num: 11
Candidate Starts for DocMcStuffins\_86:
(**Start: 11 @55397 has 203 MA's),**

**General:** Co-ordinates choice is 55397 to 55639. This is a protein coding gene that has an unknown function. It has strong coding potential according to the graph in Host-Trained GeneMark. Glimmer and GeneMark are the same and is the longest ORF. Starterator showed this start is called in 74.8% of genes in the pham it has 202 MA and is called 100% of the time when present. Syteny observed upstream, no transmembrane domains, and there are conserved domains. NCBI BLAST, HHPRED and alignment with similar genes give evidence for the function hypothetical protein.

**Function?NKF;** Hypothetical protein -> blast support: unspecified of 298 members of this pham

 **------------------------------------------**

**Gene 87**

**FINAL Gene Co-ordinates: *55632- 55793***

***Original: 55632- 55793***

**Glimmer & GeneMark Starts***: Glimmer Start: 55632 - GeneMark Start: 55632*

**Starterator:** The start number called the most often in the published annotations is 18, it was called in 89 of the 123 non-draft genes in the pham.

Genes that call this "Most Annotated" start:

Start 18:
• Found in 102 of 140 ( 72.9% ) of genes in pham
• Manual Annotations of this start: 89 of 123
• Called 100.0% of time when present

Gene: DocMcStuffins\_87 Start: 55632, Stop: 55793, Start Num: 18 Candidate Starts for DocMcStuffins\_87:
(6, 55551), (Start: 18 @55632 has 89 MA's),

**General:** Co-ordinates choice is 55632 to 55793. BLAST alignment, NCBI BLAST alignment, and HHPRED show this is a protein coding gene and similar genes have hypothetical protein as it functions. Starterator showed this start is called in 100% of genes in the pham with 164 MA and it is called 100% of the time when present. Glimmer start and GeneMark are the same. Syteny is observed upstream, no transmembrane domains, and no conserved domains found.

**Function? *Hypothetical protein -> unspecified in140 members of this pham***

--------------------------------------

Gene 88:

**FINAL Gene Co-ordinates: *55810 - 57246***

***Original auotannotation: 55810 - 57246***

**Glimmer & GeneMark Starts***: Glimmer Start: 55810; GeneMark Start: 55810*

**Starterator:** The start number called the most often in the published annotations is 23, it was called in 157 of the 201 non-draft genes in the pham. This is = Genes that call this "Most Annotated" start:

Start 23:
• Found in 217 of 221 ( 98.2% ) of genes in pham
• Manual Annotations of this start: 157 of 201
• Called 78.8% of time when present

Gene: DocMcStuffins\_88 Start: 55810, Stop: 57246, Start Num: 23
(1, 55051), (2, 55213), (3, 55357), (4, 55360), (5, 55450), (6, 55507), (7, 55612), (Start: 8 @55636 has 35 MA's), (**Start: 23 @55810 has 157 MA's),** (Start: 26 @55822 has 1 MA's), (Start: 45 @56161 has 1 MA's), (48, 56263), (54, 56398), (56, 56428), (58, 56446), (64, 56503), (68, 56548), (69, 56566), (70, 56569), (71, 56620), (79, 56683), (89, 56815), (92, 56845), (94, 56869), (97, 56911), (99, 56926), (103, 57007), (105, 57031), (110, 57049), (116, 57124), (120, 57187),

**General**: Co-ordinates choices are 55810 to 57246. This is a protein coding gene with the function is Galactosyltransferase. NCBI BLAST, alignment with similar phage, HHpred, and conserved domains indicate the function of galactosyltransferase. This gene has strong coding potential and glimmer, Genemark, DNA master have the same start. Starterator showed this start is called in 98.2% of genes in the pham with 157 MA and is called 79.6% of the time when present. Syteny is observed downstream, no transmembrane domains detected, there are conserved domains

**Function: galactosyltransferase** -> 95% of 221 member of this pham call this;

-------------------------------------------

**Gene** 89

**FINAL Gene Co-ordinates: *57265 - 57459***

**Auto annotation*: 57265 - 57459***

**Glimmer & GeneMark Starts***: Glimmer Start: 57265 - GeneMark Start: 57265*

**Starterator:** This is = Genes that do not have the "Most Annotated" start:

Start 31:
• Found in 93 of 250 ( 37.2% ) of genes in pham
• Manual Annotations of this start: 16 of 228
• Called 23.7% of time when present

Gene: DocMcStuffins\_89 Start: 57265, Stop: 57459, Start Num: 31 Candidate Starts for DocMcStuffins\_89:
(Start: 31 @57265 has 16 MA's), (41, 57430),

**Function:NKF *-* Hypothetical protein -** supported by blast and HHPred

* Co-ordinates are 57265 to 57459 and is a protein coding gene that has an unknown function because it lines up with other genes with hypothetical genes. BLAST alignment and NCBI BLAST show similarity to genes encoding proteins of unknown function. The gene has strong coding potential and glimmer start and GeneMark is the same. Starterator showed this start is called in 37.2% of genes in the pham with 16 MA and it is called 22.8% of the time when present. Other similar F1 phage call the same start. Synteny is observed downstream, no transmembrane domains are found and there are no conserved domains.

---------------------------------------

Gene 90

**FINAL Gene Co-ordinates:57456 - 58067**

**Original Auto-annotation: 57456 - 58067**

**Glimmer & GeneMark Starts***: Glimmer Start: 57456 ; GeneMark Start: 57456*

**Starterator:** Genes that do not have the "Most Annotated" start:

Start 48:
• Found in 169 of 498 ( 33.9% ) of genes in pham
• Manual Annotations of this start: 151 of 448
• Called 9 Gene: DocMcStuffins\_90 Start: 57456, Stop: 58067,: Start Num: 48(35, 57345), (37, 57354), **(Start: 48 @57456 has 151 MA's),** (64, 57477), (76, 57516), (77, 57525), (Start: 78 @57540 has 1 MA's), (84, 57588), (89, 57627), (98, 57693), (101, 57708), (127, 57846), (128, 57855),

Called 98.8% of time when present •

**General:** Co-ordinates chosen are 57456 to 58067 and it is a protein-coding gene with the function that matches glycosyltransferase. NCBI BLAST, BLAST alignment, conserved domains, and HHPRED show similarity to genes encoding glycosyltransferase. The gene has strong coding potential and Glimmer GeneMark, DNA master call the same start. Starterator showed this start is called in 33.9% of genes in the pham with 150 MA and it is called 98.8% of the time when present. Synteny is observed upstream, and no transmembrane domains are founded and has Conserved domains.

**Function?** glycosyltransferase -> 97% of 498 Pham