Starterator Guide

Marissa Pacey
Last guide edit by Welkin Pope, Dec 2014.

Starterator is a tool designed to help resolve the conundrum of which start to choose for a given gene when there is no clear solution from the evaluation of the guiding principles of gene annotation (see DNA Master Annotation guide). For example, for a given gene, one start may yield the longest gene, but the other start has the best ribosome binding site, and BLAST data shows that similar genes have been called both ways. Which one should you pick?

By aligning the related sequences in a multiple-sequence alignment, it is possible to see which starts are present and conserved in all/most of the related genes, and which are not. Our new program Starterator is designed to automate this process for the phams generated by Phamerator in the most current Mycobacteriophage_Draft database (which uses kClust for pham building).

Starterator is written in Python, runs on an Ubuntu operating system and requires a concurrent installation of Phamerator. Starterator performs multiple sequence alignments of Phamerator phams using ClustalW and then generates the results of the alignment as a graph and a text report in .pdf format. It is possible to select either a single gene within a pham, which yields a single multiple sequence alignment (and is relatively quick) or to select an entire phage genome, in which results in separate multiple sequence alignments for all the phams in the genome concatenated into the same file (takes several hours). For classroom purposes, you may want to run a whole genome and share the pdf. Starterator is automatically updated to the current version each time it is run.

Installation

Starterator Software Requirements

Starterator can only be used in conjunction with Phamerator, so it must be installed onto a computer or virtual machine that has Phamerator. Starterator also requires a few other Unbuntu packages to function correctly. These are automatically installed when the installStarterator.sh script is run (see below on how to do this).
The requirements are:

- The Ubuntu packages ncbi-blast+, pip, and git which can be installed using the command:
  `sudo apt-get install PACKAGE`
- The python packages PyPDF2, BeuatifulSoup4, and requests which can be installed using the command:
  `sudo pip install PACKAGE`

Starterator/Phamerator Virtual Machine Requirements

- Ubuntu 12.04 "Precise Pangolin"
- 1 GHz processor
- 2 GB RAM
- 128 MB video memory
- 1 GB free hard-drive space
- Internet connection
- **FULL sudo PRIVILEGES**

Starterator Installation using the installStarterator.sh script

1. After you have logged in to Ubuntu, launch the terminal application by clicking the Ubuntu button in the top left corner and typing “terminal”.

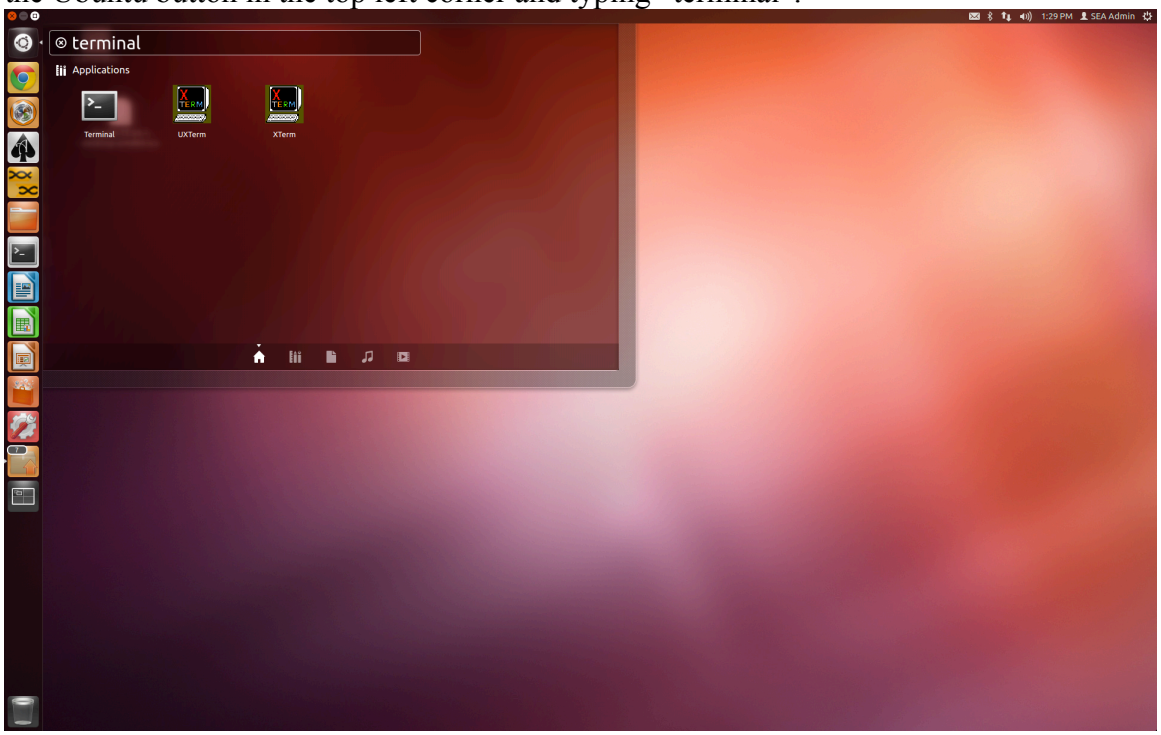

Figure 1: Launching the Terminal Application.

2. Download the script installStarterator.sh from:
   http://phamerator.webfactional.com/installStarterator.sh

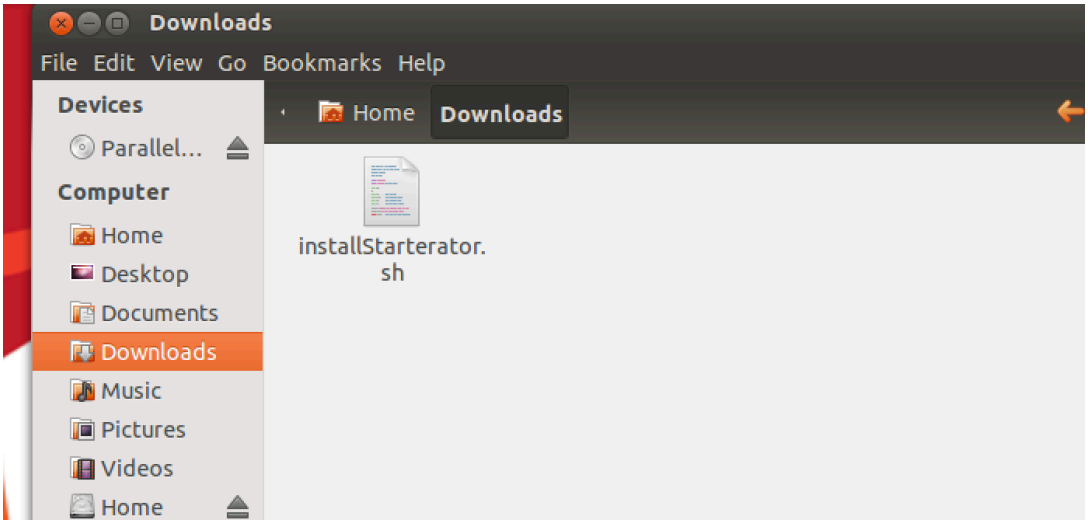

Figure 2: Starterator Install Script in the Downloads Folder.

3. Move the script to your downloads folder (if it isn't there already)
4. Open the Terminal application (Figure 1).
5. In Terminal, navigate to the folder where you saved installStarterator.sh.
   - `cd ~/Downloads`
6. After navigating to this folder, run the script by typing, in Terminal:
   - `bash installStarterator.sh`

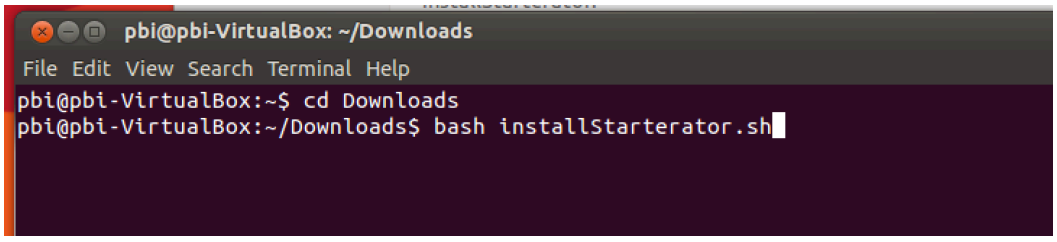

Figure 3: Terminal Commands. Options before the ":" may appear different on your VM.

7. You will then be asked to type your password, which differs depending on if the program is installed under the administrator or faculty account. Once this is done, Starterator and its requirements will be installed, a shortcut created, and the program will launch.

Getting Started Using Starterator

Once installed, a Starterator launch button will appear on your Desktop (found in the task bar in the 2015 SEA VM). Click to launch the program's home window (Fig. 3)

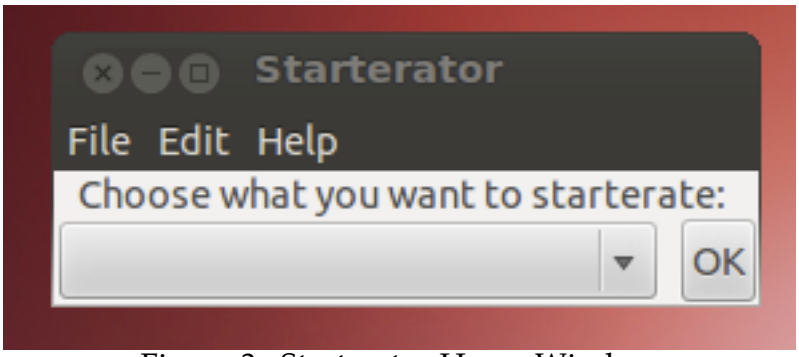

Figure 3: Starterator Home Window

In the drop-down window, there are multiple choices for Starterator inputs:

1. Whole Phamerated Phage:
Generates a multiple sequence alignment for each pham found in a phage in the Phamerator Mycobacteriophage_Draft database, and concatenates the results.

2. Whole Unphamerated Phage:
Generates a multiple sequence alignment for each gene found in a phage that is not in the Mycobacteriophage_Draft database , and concatenates the results. Requires the additional input of a .fasta file of the predicted nucleotide sequence of the genes from a finished phage sequence. This can be generated from a DNA Master auto-annotated sequence (See DNA Master Annotation Guide).

3. One Phamerated Gene:
Generates a multiple sequence alignment of the pham that the phamerated gene belongs to.

4. One Unphamerated Gene:
Generates a multiple sequence alignment of the putative pham that the unphamerated gene belongs to. Requires the upload of the nucleotide sequence of the gene in .fasta format.

5. One Pham:
Generates a multiple sequence alignment of the pham found in Mycobacteriophage_Draft

Outputs

Results from selecting a single gene or single pham:

Introduction

Starterator outputs a single PDF with 2 elements per Pham. The first is a visual representation – a graph - of the various start sites set in a display of the ClustalW alignments. This is followed by a text report that that matches the visual display and highlights the specific coordinates of start codons.

Graphic Representation

The first element of the report for a Pham is the graph representing the genes in the Pham and their candidate start sites. Each horizontal track represents a gene or set of identical genes within the Pham that have the same nucleotide sequence, candidate start sites, and annotated start site. The length of all of the tracks is determined by the gene of that Pham with the longest Open Reading Frame. (from stop codon to stop codon). The pink color indicates alignment of the nucleotide sequences (pink = aligned sequence, white = no alignment, aka gaps). All of the possible starts are colored and numbered in order of appearance within the all genes from left to right. Colors and numbers of starts are consistent across tracks. The annotated start site of the gene(s) within each track is blue, while other start sites are colored at random.

Figure 4: Multiple sequence alignment of Pham 3205

Interpretation:

- All 4 members of Pham 3205 have the same ORF length.
- The first three tracks show a gap in the Clustal alignment at the right end.
- All Tracks have 5 identical start choices. Track 3 has one addition start choice (#6).
- Tracks 1, 3, and 4 have called the same start.
- Investigation is needed to interpret the start prediction of Track 2. The most likely interpretation is that the start of the genes listed in Track 2 is incorrect and should be reassigned.

Text Report

The text report contains the legend for the Visual Representation report. It is divided into seven sections:

1. Track 'Members': The first section in the text report identifies the gene or genes that comprise each track.
2. Most Called Start: The number displayed here represents the most common start choice as identified by the visual representation. All genes of Pham 3205 are included here, except the genes in Track 2.
3. Percent with start called: This is followed by the % of genes (of that Pham) that have that start as the annotated start. In this case, 71.4%.
4. Genes that use the most-frequently annotated start: The next item lists all the genes use the most commonly-called start (Genes of Tracks 1, 3, and 4).
5. Genes that have the most-frequently annotated start but do not call it: The list of genes that have the most commonly-annotated start but don't call it (Genes in Track 2).
6. Genes that do not have the most-frequently annotated start: The list of genes in which the most-commonly called start is not present. In this example, there were not any genes in this category.
7. Other starts called: Here, the list contains the genes from Track 2.

The next part of the text report contains specific information for the gene of interest. There are several possible items which may appear in this list:

- The Suggested Start coordinates (which Starterator will suggest is the most frequently-annotated start found across all the genes in the pham).
- If there is no suggested start (that is, the most-frequently annotated start within the Pham is not present in the gene of interest), then a list containing the coordinates of all the candidate starts for that gene appears.
- If the most-frequently annotated start within the Pham is a candidate start of the gene, but it is not currently called, that start appears in brackets. The first number in parentheses refers to the label of the start on the graph, the second number refers to the coordinates of the gene in the DNA sequence of the phage. (i.e. [(1, 42)] means "Start 1, bp coordinate 42")
- If the most-frequently annotated start of the Pham is the currently-called start of the gene of interest, then it appears without brackets: (1, 42).
- Listed next on the report are the gene number, the current start coordinate, and the stop coordinate.
- Lastly, all the candidate starts of the gene are listed.

Whole Phage Output

Introduction

The results from Starterator for a whole phage genome is similar to that of a single gene or pham, however, it contains one additional component: a visual representation of the genome as a whole. Each gene is a color (the color is of no significance) and is labeled with a pham and gene number. It is followed by a list of suggested starts for each gene. The third set component is the same as the output of the Output of Pham for each gene in the genome. (See previous

section.) This is a large file. The Whole Phage Output for Liefie (Cluster G) is 12.6Mb.

Whole Genome Display

The first component of the Whole Phage Output report is a genome map that identifies all genes called labeled with gene and pham numbers.

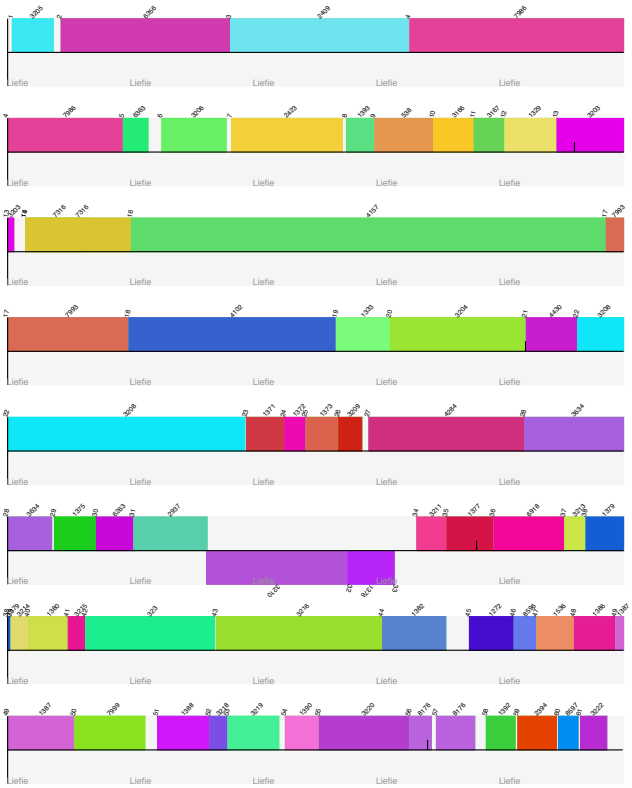

Figure 5: Whole genome Display for Mycobacteriophage Liefie

Text Report

The second component of the Whole Phage Output is a text display of suggested starts for each gene.

Output of Pham

The third component of the Whole Phage Output is the Output of Pham for each gene in the genome. (See Output of Pham section.)

Starterator Reports

By default, all the reports generated by Starterator are saved as .pdf files in the Starterator folder “Report files”. This folder is found within the hidden folder .Starterator within your Home directory.
To view this folder, open your Home folder by clicking on the folder icon in the task bar on the left side of your Ubuntu window. In the window that pops up, make sure “Home” is highlighted in the bar on the left hand side under the heading “Computer”. Then mouse to the top of the Ubuntu window, such that the headings “File”, “Edit”, “View”; etc, appear. Under the menu “View” select “Show Hidden Files”.
The folder .Starterator should appear in the window. Within this folder, you will find the folder “Report Files”. Right-click on this folder, and select “Make link”. This will generate a shortcut icon to this folder that can be dragged to the Desktop.

For more help with Starterator, see the Help files within the Starterator program.

Figure 6: Help Menu display for Starterator outputs

Using Starterator to inform your start selection

The next figure was generated by running Starterator on Sisi gene 5:

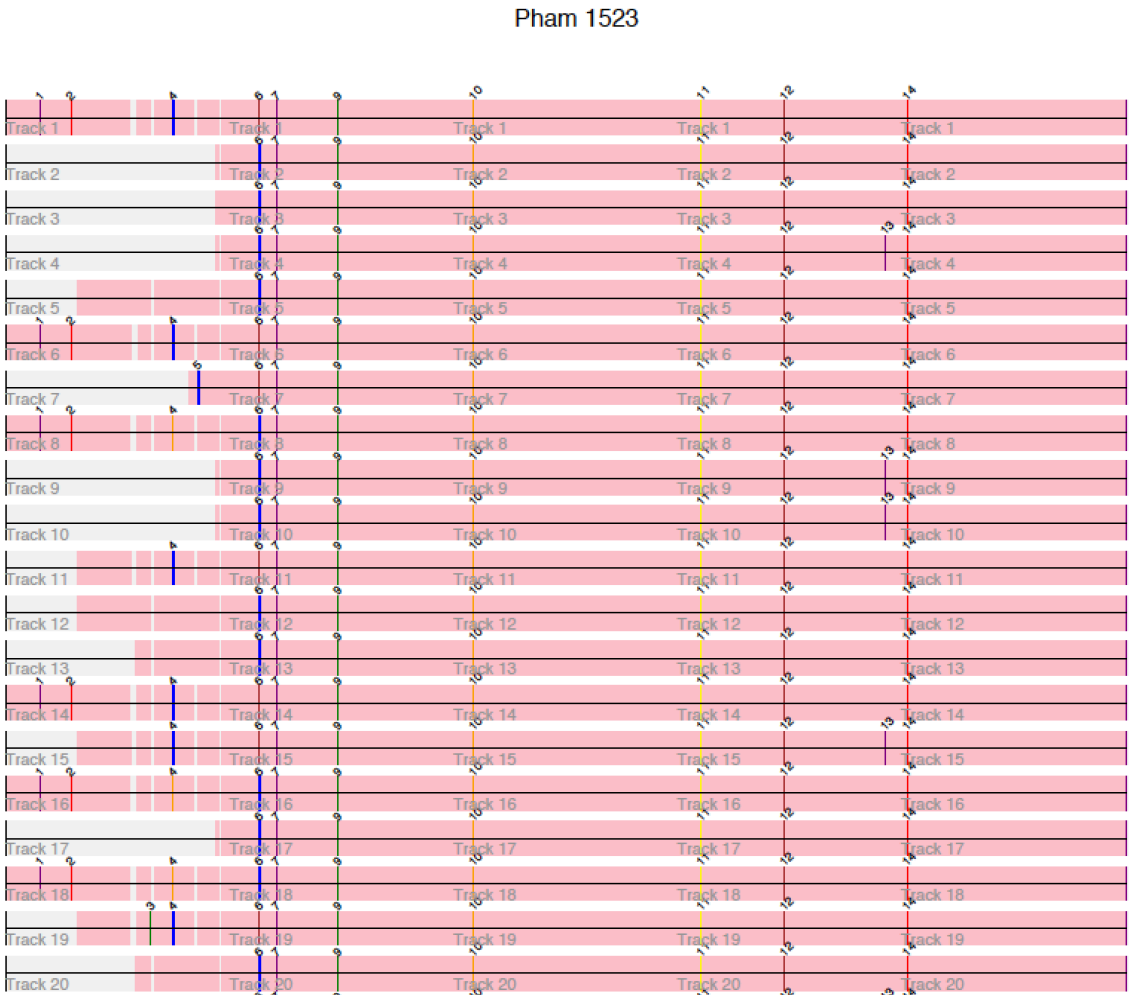

Figure 7: Sisi gene 5

While the genes in the pham above are very similar to each other—as shown by the perfect alignment of starts from start “6” through the ends of the ORFs--- the region upstream of start 6 is not. All of the genes in this pham have start “6” in common. However, start “6” has not been selected as the start in the annotation in Phamerator in a number of genes. Many of these genes are the lengthier genes in the pham, and these starts may have been chosen by annotators who selected the longest ORF possible.

Following the graph is the report that lists all the genes that were included in the alignment and the track each is represented by:

Pham 1523 Report

- Track 1 : Ardmore_5, Taj_5, Tweety_gp5, Shauna1_5, Mutaforma13_5, Wee_gp5, SG4_5
- Track 2 : Florinda_5
- Track 3 : Ruby_Draft_4, MisterCuddles_Draft_4, Girr_Draft_4
- Track 4 : Brocalys_Draft_6, Saal_5
- Track 5 : Cabrinians_Draft_5
- Track 6 : Spartacus_5, Hades_Draft_5
- Track 7 : Che8_5
- Track 8 : SuperGrey_Draft_5, Bipolar_Draft_5, Batiatus_Draft_5, Ovechkin_Draft_5
- Track 9 : GUmbie_5, Llij_5, Mantra_Draft_5, PMC_5, Dante_Draft_5
- Track 10 : ShiLan_5
- Track 11 : Dorothy_5, Inventum_Draft_5, Daenerys_5, Pacc40_5
- Track 12 : Bubbles123_Draft_5
- Track 13 : Llama_5
- Track 14 : DotProduct_5
- Track 15 : Empress_Draft_5
- Track 16 : SiSi_5
- Track 17 : OlympiaSaint_Draft_6
- Track 18 : Hamulus_5
- Track 19 : Fruitloop_5
- Track 20 : Ibhubesi_5

Sisi gene 5 is Track 16 of the graph.

The final component of the report lists a "recommended start"; based on the most frequently annotated start in Phamerator, along with all the possible starts for the Starterated gene(s).

Suggested Starts:

SiSi_5, (6, 4435)
Gene Information:
Gene: SiSi_5 Start: 4435, Stop: 5028
Candidate Starts for SiSi_5:
[(1, 4292), (2, 4313), (4, 4379), (6, 4436), (7, 4448), (9, 4490), (10, 4583), (11, 4739), (12, 4796), (14, 4880)]

Starterator suggests that Sisi gene five should use start "6", at Sisi genome coordinate 4435.

It is important to remember that the Phamerator Mycobacteriophage_Draft database contains draft annotations, and therefore a commonly selected start found by Starterator may be an artifact of unrefined auto-annotations generated by the computer gene-calling programs. Draft annotations are easy to identify because the word "_Draft" is added to the end of the Phage name in the report.

Starterator is not always informative.

Below, the results for Sheen gene 5:

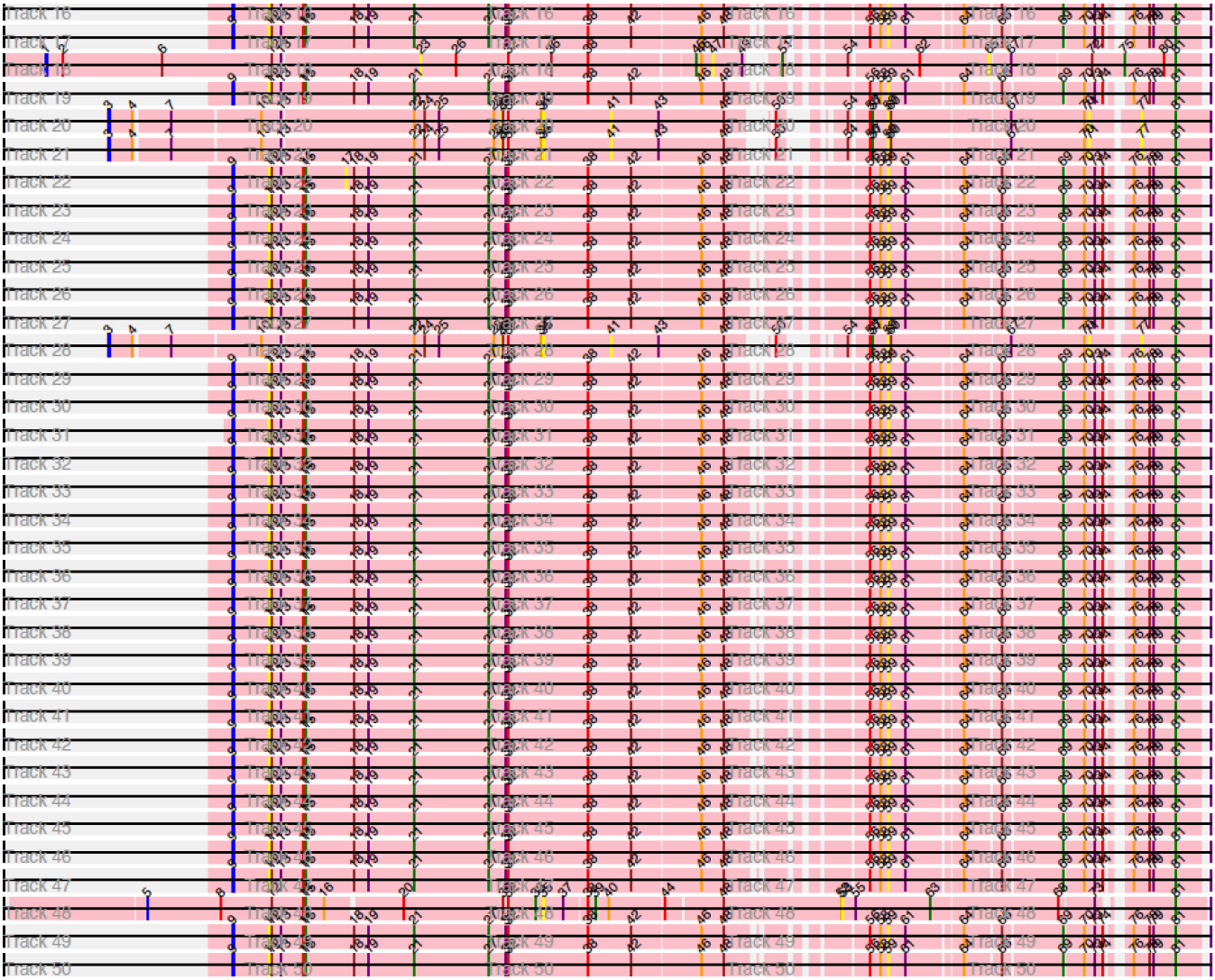

Figure 8: Sheen gene 5

Sheen gene 5 is in Track 48 in the above graph. The majority of genes in this pham appear pretty similar to each other, with the outliers being in tracks 18, 20, 21, 28, and 48 (tracks 1 through 16 look fairly similar to each other so they were omitted from this figure). The blue bars represent the start that is currently annotated in each gene in Phamerator.

Suggested Starts:

Sheen_Draft_5, [(5, 2348), (8, 2435), (12, 2495), (14, 2534), (15, 2537), (16, 2558), (20, 2645), (30, 2765), (32, 2771), (33, 2804), (35, 2813), (37, 2837), (38, 2864), (39, 2873), (40, 2888), (44, 2954), (48, 3020), (52, 3161), (53, 3164), (55, 3179), (63, 3269), (68, 3419), (73, 3461), (81, 3536)]
Gene Information:
Gene: Sheen_Draft_5 Start: 2347, Stop: 3570
Candidate Starts for Sheen_Draft_5:
[(5, 2348), (8, 2435), (12, 2495), (14, 2534), (15, 2537), (16, 2558), (20, 2645), (30, 2765), (32, 2771), (33, 2804), (35, 2813), (37, 2837), (38, 2864), (39, 2873), (40, 2888), (44, 2954), (48, 3020), (52, 3161), (53, 3164), (55, 3179), (63, 3269), (68, 3419), (73, 3461), (81, 3536)]

Unlike the Sisi gene 5 example, Starterator is not able to offer a suggested start for Sheen because the start that is annotated for most of the genes in the pham is not present in Sheen. In fact, Starterator is not particularly informative for Sheen 5, as the two most likely start candidates in Sheen 5 are not present in any of the other genes in the pham; nor do the rest of the starts throughout the gene line up particularly well with the other genes. The sequences have diverged to the point where a single start codon is not present for all pham members. So both starts "5" and "8" are still in the running for Sheen gene 5.

When you are using Starterator to inform your annotation start sites, it should be documented in the Notes for that gene in your DNA Master file. Starterator notations are
"NA" for "Not Applicable"—for orphams, or for genes in which the evidence overwhelmingly supports a single start choice and Starterator was not necessary
"SS" for "Suggested Start".
"NI" for "Not Informative"--This notation indicates that Starterator was run for the gene, but the output didn't assist in the start decision.

Sheen gene 5 would be noted "NI", while Sisi gene 5 would be noted "SS".

Bottom line: While Starterator may help resolve some start issues for particularly well-conserved genes, or genes found in multiple clusters, it should not be relied on to select the correct start every time. The guiding principles of annotation are still the best way to determine a gene start.