Starterator Guide

© University of Pittsburgh
Marissa Pacey
Last guide edit by Welkin Pope, Dec 2016.

Often, the scientific evidence generated by the bioinformatics tools we use does not agree on a specific start codon for a gene. However, by analyzing that gene as part of a set of related genes in a multiple sequence alignment, a start codon that is common to all genes in the alignment can sometimes be found. Starterator, a companion program of Phamerator, streamlines this analysis for you.

Starterator examines all the genes within a pham and generates the longest possible ORF for each gene within a pham (from stop codon to stop codon) using the phage genome sequence. We use the longest possible genes in the alignment to account for erroneously truncated genes generated during auto-annotations. Starterator aligns the nucleotide sequence of these lengthened ORFs using the multiple sequence alignment program ClustalW. As an output, Starterator produces a graphic that represents the alignment of all the ORFs, with each of the possible starts for all of those ORFs mapped on to each gene in the pham. The starts are consistently colored and numbered across all the ORFs to assist you in determining which starts are present in the majority of the pham members, and therefore are more likely to be the true start of the gene. The analysis is not always conclusive. You, as the human, still needs to interpret the data and decide if the analysis suggests one start choice over another.

As input, it is possible to select either a single gene within a pham, which yields a single multiple sequence alignment of the pham the gene belongs to (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.

When interpreting Starterator data, in general the start that is present in all genes that yields the longest possible gene is the correct one. The underlying rationale for this is that upstream sequence is more likely to vary than protein encoding sequence, and so the most conserved start that yields the longest genes should be selected. As always, there are exceptions to this, and so sometimes the analysis is not informative or not applicable. Examples of this will be described below.

Starterator is written in Python, runs on an Ubuntu operating system and requires a concurrent installation of Phamerator. The program is automatically updated to the current version each time it is run. Starterator was initially written by Marisa Pacey, at the University of Pittsburgh, and has been further developed and is currently maintained by Chris Shaffer at Washington University in St. Louis.

Starterator is pre-installed on the SEA Ubuntu VM, and is preset to use the Actinobacteriophage_Draft database. To install on a new Ubuntu machine, see the New Installation section of this guide.

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. 1)

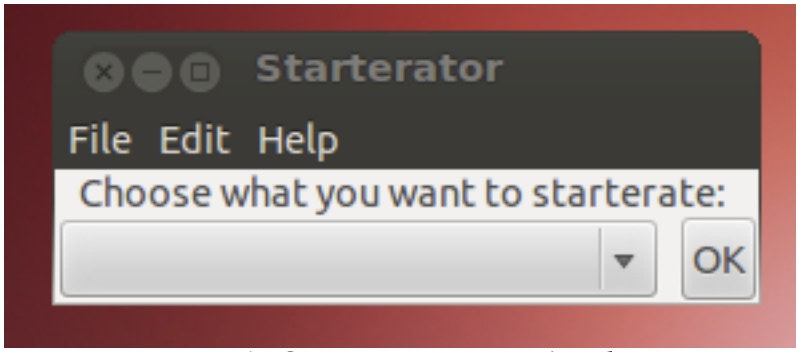

Figure 1: Starterator Home Window

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

1. Whole Phamerated Phage:
This option generates a multiple sequence alignment for each pham found in a phage in the Phamerator Actinobacteriophage_Draft database, and concatenates the results.

2. Whole Unphamerated Phage:
This option generates a multiple sequence alignment for each gene found in a phage that is not in the Actinobacteriophage_Draft database, and concatenates the results. It 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 using the "save ORFs" function. (See DNA Master Annotation Guide).

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

4. One Unphamerated Gene:

This option generates a multiple sequence alignment of the putative pham that the unphamerated gene belongs to. It requires the upload of the nucleotide sequence of the gene in .fasta format.

5. One Pham:
This option generates a multiple sequence alignment of the pham found in Actinobacteriophage_Draft

Location of 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.

New for 2016, all Starterator pham reports for the Actinobacteriophage_Draft database will be available as .pdfs online. See the whole index at:

http://phages.wustl.edu/starterator/

Or use a direct link for the one you want to review. For example, for pham 5221, the direct link to the report looks like:

http://phages.wustl.edu/starterator/Pham5221Report.pdf

Results

Single gene or single pham:

Introduction

Starterator generates a single PDF with 2 elements per Pham. The first is a visual representation – a graph - of the Clustal W alignment of the genes in the pham. Each of the various start sites is overlaid on each gene according to its position in

the alignment. 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 bar—a track-- represents a gene or set of **identical** genes--- i.e. the same nucleotide sequence, candidate start sites, and annotated start site.

-The horizontal tracks are labeled underneath with the name of the phage and gene number that they represent. If the track represents more than one gene, only a single name is listed, and the number of additional phages represented are indicated after a plus sign, e.g. "Hawkeye_5 +2" means that the track represents Hawkeye gene 5 plus two additional genes.

-The length of all of the tracks is scaled to the length of the longest open reading frame in the Pham (from stop codon to stop codon). This means that a number of tracks in any report may begin with a large white gap because those ORFs are not as lengthy as the longest ORF.

-The pink color in each track indicates aligned nucleotide sequences, while white indicates a gap in the alignment.

-All of the possible start codons within the all genes are colored and numbered in order of appearance from left to right, regardless of which track they fall into.

-Colors and numbers of starts are consistent across tracks.

The current annotated start site (according to the Phamerator database) of the gene(s) within each track is yellow or green, while other start sites are colored at random. Yellow starts were selected by Glimmer or GeneMark and are part of an automatic draft annotation, while green starts have been reviewed by a human and are part of a final annotation.

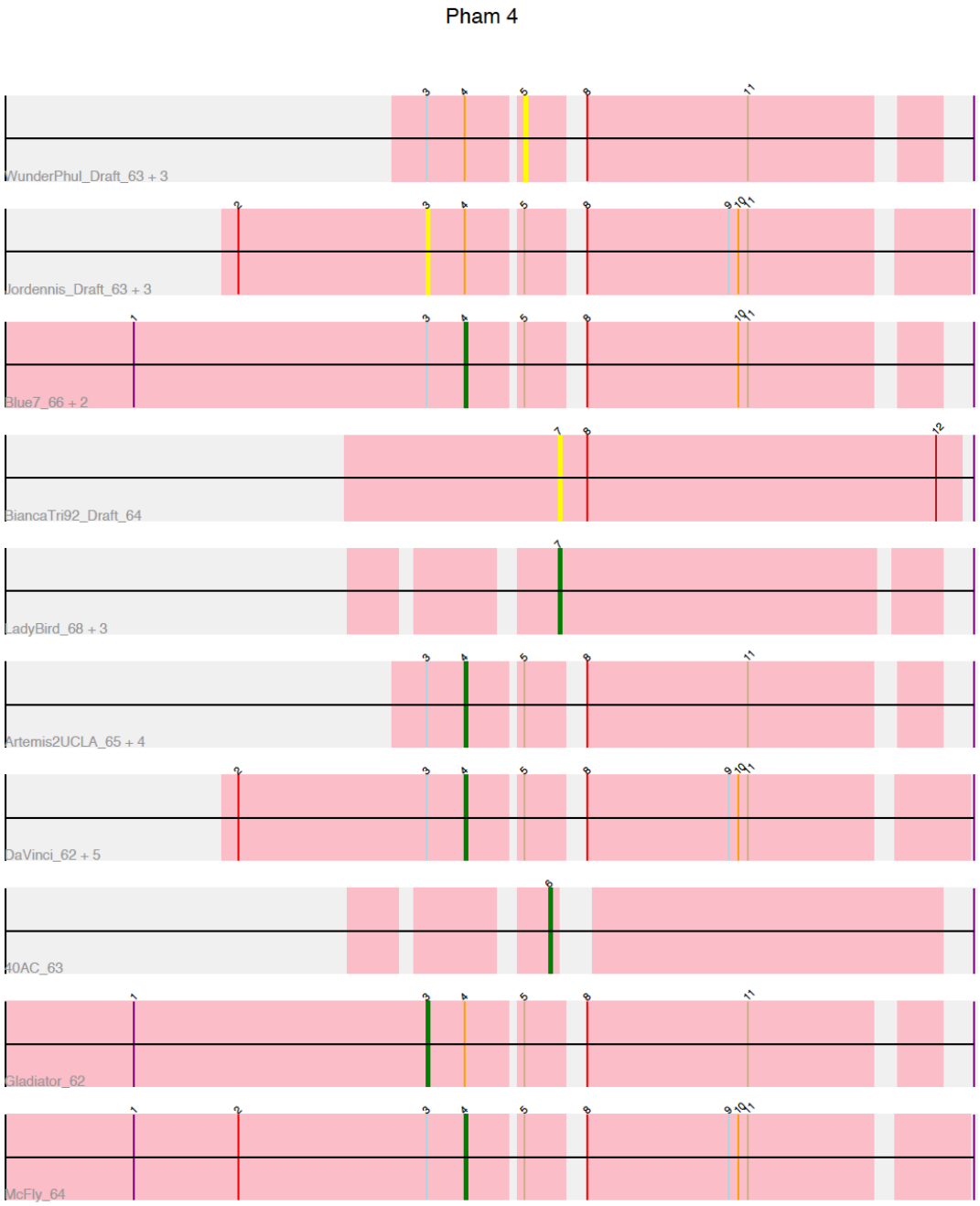

Figure 2: Multiple sequence alignment of Pham 4

- There are ten different gene variants in this pham, as shown by the ten different horizontal tracks. Some of these tracks represent multiple identical genes.
- Tracks 3, 9, and 10 are longer than the others, and so the others show a white gap in the Clustal alignment at the left end.

- Start number 1 is found in tracks 3, 9, and 10; start 2 is found in tracks 2, 6, and 10. The other starts are numbered accordingly from left to right.
- Tracks 1, 2, and 4 have starts chosen by Glimmer or GeneMark without human review (in order, the yellow starts 5, 3, and 7), all the rest of the tracks have starts reviewed by a human—the green starts.
- There are no starts that are in common between all ten tracks. Tracks 1, 2, 3, 6, 7, 9 and 10 share the starts labeled 3, 4, 5, 8 and 11.
- Tracks 4 , 5, and 8 are different from the rest and from each other, as they do not share the same pattern of start codons.

Text Report

The text report contains the legend for the Visual Representation report, and it summarizes all the data from the analysis.

For pham 4, the text is as follows:

Note: In the above figure, yellow indicates the location of called starts comprised solely of computational predictions (i.e. auto-annotations by Glimmer/GeneMark), green indicates the location of called starts with at least 1 manual gene annotation.

Pham 4 Report
This analysis was run 11/27/16.

Pham number 4 has 30 members, 10 are drafts.

Phages represented in each track:
• Track 1 : WunderPhul_Draft_63, Zulu_Draft_63, Wiks_Draft_63, Koko_Draft_65
• Track 2 : Jordennis_Draft_63, Priamo_Draft_64, SuperAwesome_Draft_64, JewelBug_Draft_63
• Track 3 : Blue7_66, Hammer_gp65, Gruunaga_65
• Track 4 : BiancaTri92_Draft_64
• Track 5 : LadyBird_68, First_0065, CRB1_66, 20ES_66
• Track 6 : Artemis2UCLA_65, ToneTone_62, Zaka_65, CloudWang3_66, Jeffabunny_65
• Track 7 : DaVinci_62, GreedyLawyer_Draft_62, EricB_63, Kazan_65, VohminGhazi_64, Isiphiwo_62
• Track 8 : 40AC_63
• Track 9 : Gladiator_62
• Track 10 : McFly_64

Summary of Final Annotations (Info on gene starts based on numbers in diagram):

The start number called the most often in the published annotations is start number 4, it was called in 14 of the 20 non-draft genes in the pham.

Genes that call this "Most Annotated" start:
• Blue7_66, Gruunaga_65, Artemis2UCLA_65, DaVinci_62, GreedyLawyer_Draft_62, EricB_63, ToneTone_62, McFly_64, Zaka_65, Kazan_65, VohminGhazi_64, CloudWang3_66, Isiphiwo_62, Hammer_gp65, Jeffabunny_65,

Genes that have the "Most Annotated" start but do not call it:
• WunderPhul_Draft_63, Jordennis_Draft_63, Priamo_Draft_64, Zulu_Draft_63, SuperAwesome_Draft_64, Gladiator_62, Wiks_Draft_63, JewelBug_Draft_63, Koko_Draft_65,

Genes that do not have the "Most Annotated" start:
•BiancaTri92_Draft_64, LadyBird_68, 40AC_63, First_0065, CRB1_66, 20ES_66,

Summary by start number:

• Start number 3 is called in: Jordennis_Draft_63, Priamo_Draft_64, SuperAwesome_Draft_64, Gladiator_62, JewelBug_Draft_63,
Percent with start 3 called: 16.7%

• Start number 4 is called in: Blue7_66, Gruunaga_65, Artemis2UCLA_65, DaVinci_62, GreedyLawyer_Draft_62, EricB_63, ToneTone_62, McFly_64, Zaka_65, Kazan_65, VohminGhazi_64, CloudWang3_66, Isiphiwo_62, Hammer_gp65, Jeffabunny_65,
Percent with start 4 called: 50.0%

• Start number 5 is called in: WunderPhul_Draft_63, Zulu_Draft_63, Wiks_Draft_63, Koko_Draft_65,
Percent with start 5 called: 13.3%

• Start number 6 is called in: 40AC_63,
Percent with start 6 called: 3.3%

• Start number 7 is called in: BiancaTri92_Draft_64, LadyBird_68, First_0065, CRB1_66, 20ES_66,
Percent with start 7 called: 16.7%

The text of the report lists all of the genes and phages represented in the pham, and summarizes the analysis of the alignment. It highlights:

-which starts were present most often in the set of the genes
-which starts were chosen by a human or computer program most often in the set of the genes.

Final Interpretation:

Based on the above alignment data alone, the most likely start for pham 4 is start 3 (for the tracks that have it), as it is present in most of the genes and yields the longest possible. The rationale for this is that upstream sequence is more likely to vary than protein encoding sequence, and so the most conserved start that yields the longest genes should be selected.
However, most of the time in published annotations, a human has chosen start 4, which suggests that additional factors are in play. Starterator alone should not overrule the rest of the evidence (perhaps start 4 yields a 4bp overlap with the upstream gene in every phage? In that case, start 3 would still be part of a protein-encoding sequence, but just in the upstream gene). The Starterator evidence does not rule out start 4, it just gives better support for start 3 when taken out of context with all the other bioinformatics programs.

The starts in the outlying tracks are more straightforward: start 7 is the obvious choice for tracks 4 and 5, and start 6 is the only choice for track 8.

Correct Documentation for Starterator in your annotation file:

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.

Ladybird 68 would be noted NA (as there is only one start possible for the gene)

"SS" for "Suggested Start"

BiancaTri92 gene 64 would be would be noted "SS" as Starterator aids in the choice of start 7

"NI" for "Not Informative"--This notation indicates that Starterator was run for the gene, but the output didn't assist in the start decision.

The rest of the tracks in the report could either start at start 3 or start 4, and Startertator does not give you a clear answer. Writing a sentence to that effect in your annotation notes for this gene would be appropriate.

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.

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, followed by a pham report 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.

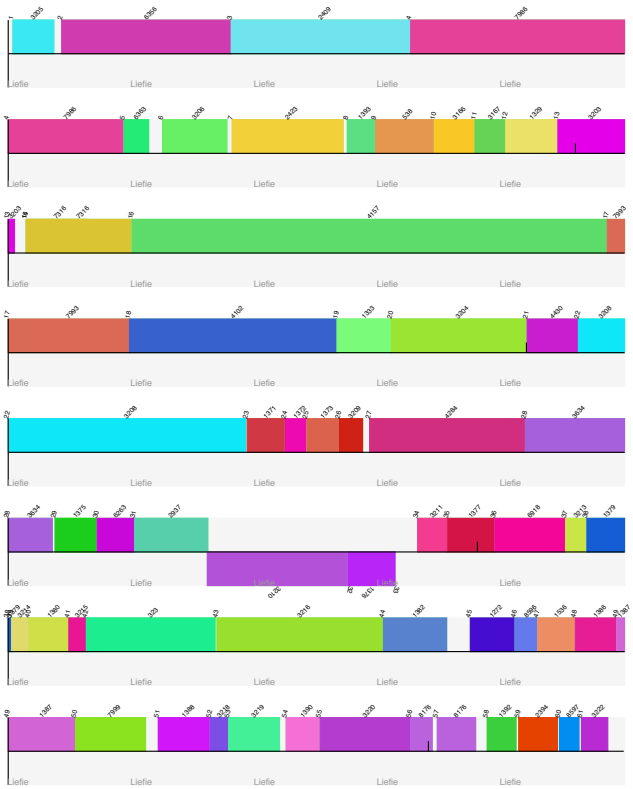

Figure 3: Whole genome Display for Mycobacteriophage Liefie

New Installation

Starterator Software Requirements

You can install Starterator on any Ubuntu machine using the instructions below. 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”.

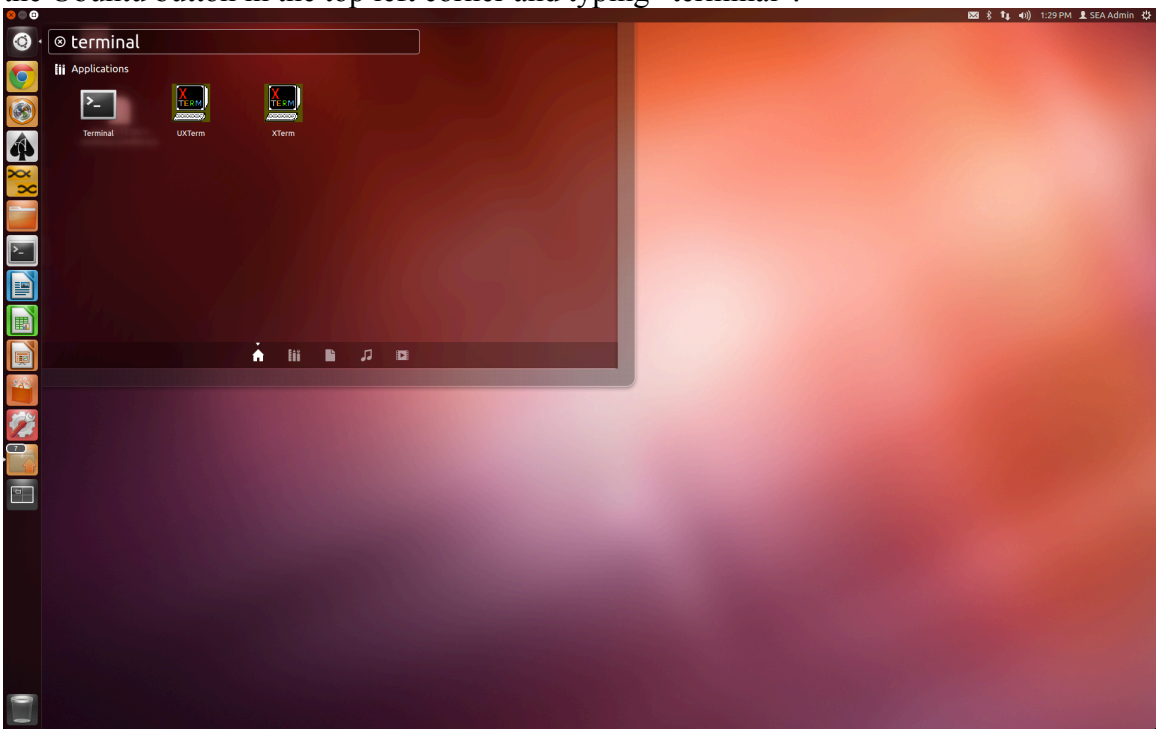

Figure 1: Launching the Terminal Application.

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

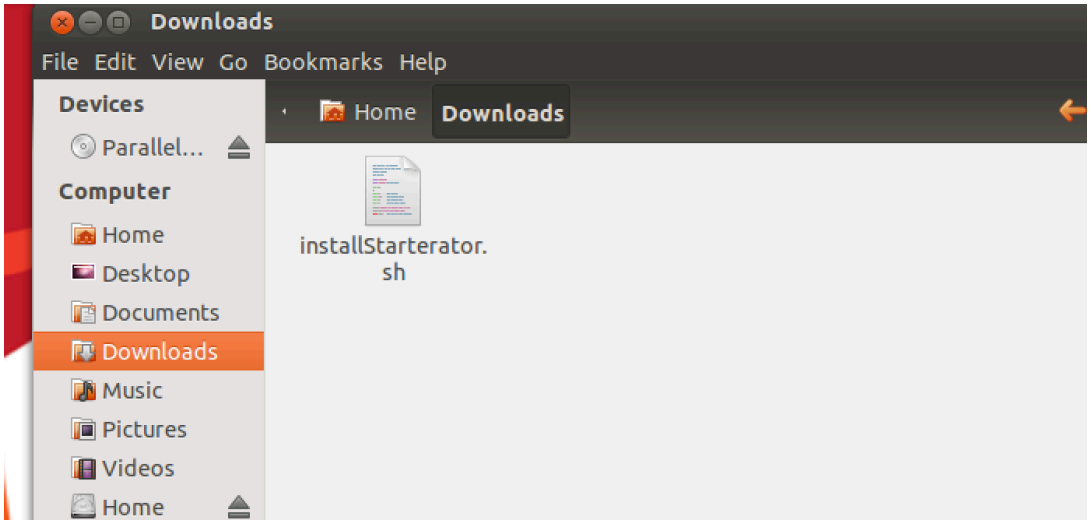

Figure 2: Starterator Install Script in the Downloads Folder.

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`

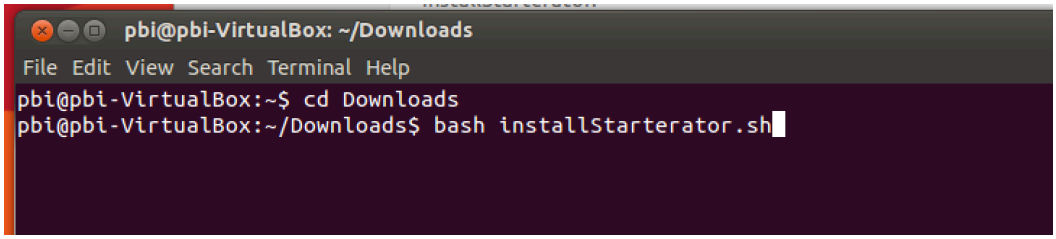

Figure 3: Terminal Commands. Options before the ":" may appear different on your VM.

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.