Gordonia Terrae phage Charianelly contains 4 novel genes capable of infecting 2 out of 3 alternate hosts

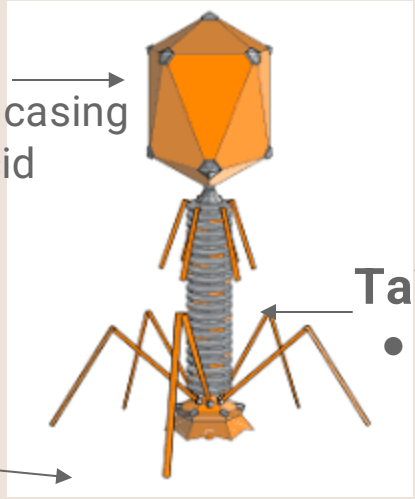
Hamilton, Reagan Khan, Ayaan
Mashru, Simran

BIOSC 0067 Foundations of Biology 2 Lab Fall 2023 semester

University of Pittsburgh Department of Biological Sciences

The study of a bacteriophage

Capsid Head
- Protein encasing nucleic acid genome



Tail Sheath
- Channel for genome delivery

Tail Fibers
- Mediate phage binding

Fig. 1: Bacteriophage

- Viruses that infect bacteria
- Can expand host range by new mutations and genes
- Most abundant entity on Earth and widely diverse

SEA-PHAGES program- undergraduate research that focuses on discovery and retention in the biological field

Gordonia Terrae

- Used by pitt sea phage students
- Useful because it is related to *M. smegmatis*
- Found within the soil

Why studying phages is important

Experimental:
- Exploration in host interactions and genetics
- SEA-PHAGES 1 and SEA-PHAGES 2 (Hatfull lab)

Medical:
- Phage therapy to eliminate bacterial infections
- Help in antibiotic resistance

Technological:
- Genome editing: CRISPR inspired by phage
- Development of biosensors

Evolutionary:
- Provide insight to evolutionary dynamics with bacteriophages and their hosts

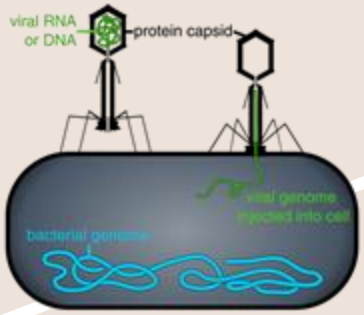


Fig 2: Bacteriophage infection

Genome annotation and its significance

- Genome annotation is is the process of trying to identify functional parts within the sequence of a genome, giving it an identity and more context in order to more accurately understand it
- Very useful in modern medical science as well as evolutionary science
  - Medical science has been able to experience breakthroughs due to us being able to identify key functions within genes quicker and more efficiently
  - Evolutionary science has also made breakthrough now that we can use genome annotation to make more accurate inferences about how these sequences might have evolved

Fig.3: Genome sequence

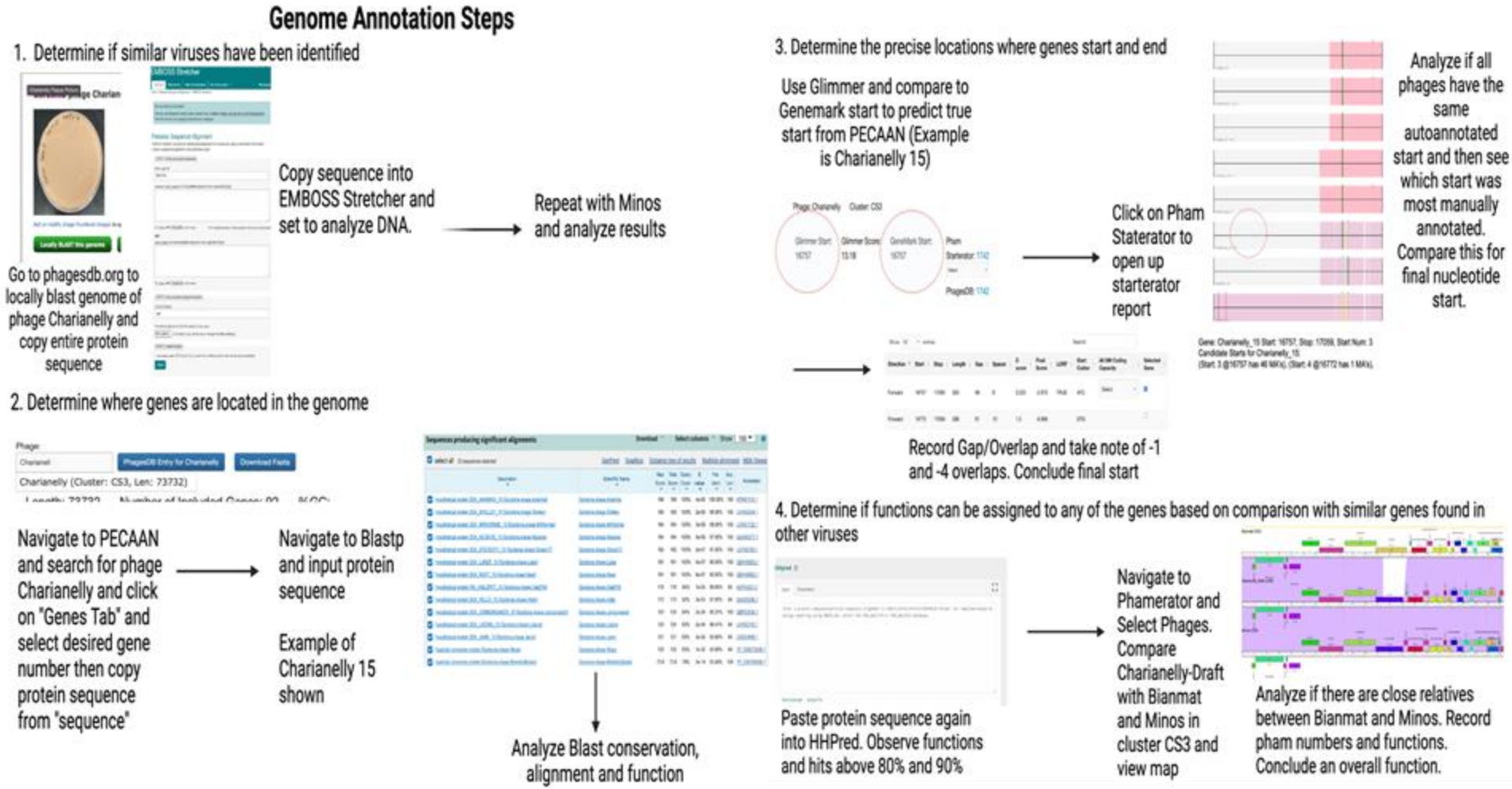


Figure 4: Methods Flow Diagram for Overall Phage Annotation

GeneMark and Glimmer are two programs that are used to predict a gene's start location

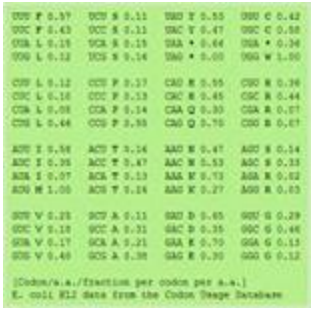
GeneMark:

- GeneMark scans the chosen genome for start/stop codons
  - Examines Open Reading Frames (ORFs) for continuous presence of the host's preferred codon (codon bias)
- Genemark correctly calls about 9 out of 10 times
  - Human input needed for the other 10% of annotation
- Doesn't call TTG starts

Codon Bias:

- Phages tend to have codons they prefer, which show up in higher frequency
  - This known as **Codon Usage Bias**
- Codon usage Bias differs among phages

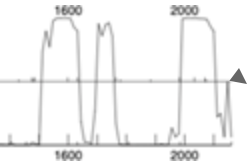
- As shown in the image, different codons have different frequencies within a phage



Glimmer:

- Coding potential program similar to GeneMark
- Does "self modeling"
  - Scans entire sequence for all the possible ORFs in all frames
  - Samples the longest frames for predicting the phage's preferred codons(codon bias)
  - Calls TTG starts
- Tends to be more accurate than GeneMark

How it works:

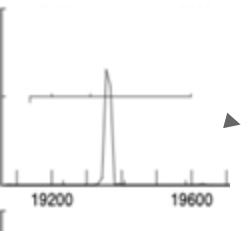


Example of "Strong" coding potential (high intensity waves, more than one)- Charianelly Gene 12

Phage: Charianelly Cluster: CS3

Glimmer Start:	Glimmer Score:	GeneMark Start:
14804	10.79	14804

When GeneMark and Glimmer agree: Strong chance that the start position is correct (Charianelly Gene 12)



Example of "Weak" coding potential (one low intensity wave, nothing else) - Charianelly Gene 12

Phage: Charianelly Cluster: CS3

Glimmer Start:	Glimmer Score:	GeneMark Start:
56266	8.51	56176

When GeneMark and Glimmer disagree: Glimmer predictions are usually more accurate

This is the start codon site according to glimmer

How it works:

Phage: Charianelly Cluster: CS3

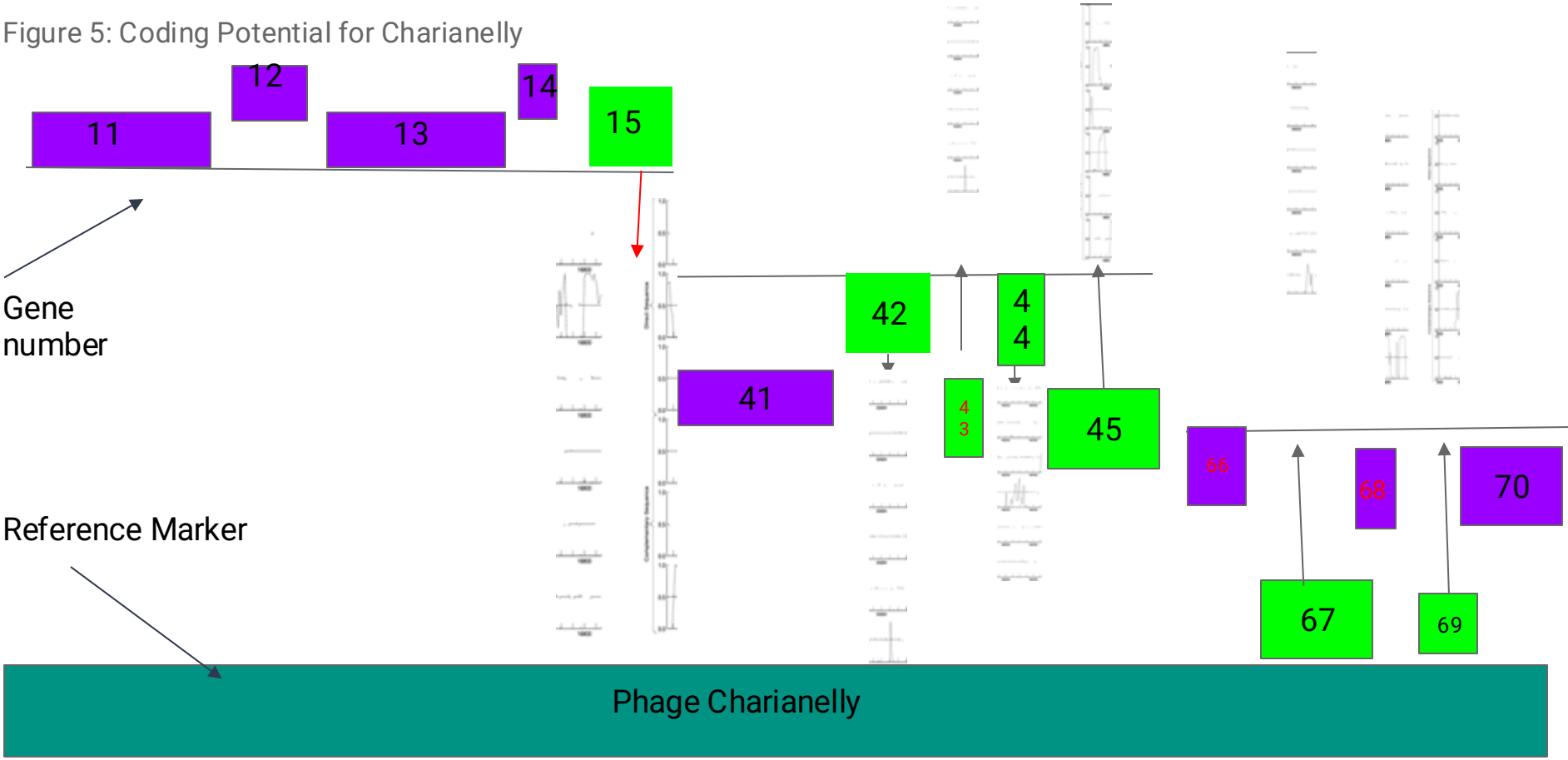
Glimmer Start:	Glimmer Score:	GeneMark Start:
14804	10.79	14804

How to use biological programs to predict a gene's start location

Green: Weak coding potential
Purple: strong coding potential
Red Text: Glimmer and Genemark do not agree

- Below is the size and chromosomal location of 15 genes from Charianelly that our group annotated
- Use of both Glimmer and Genemark

Figure 5: Coding Potential for Charianelly



Coding potential intensity is varied for different gene predictions

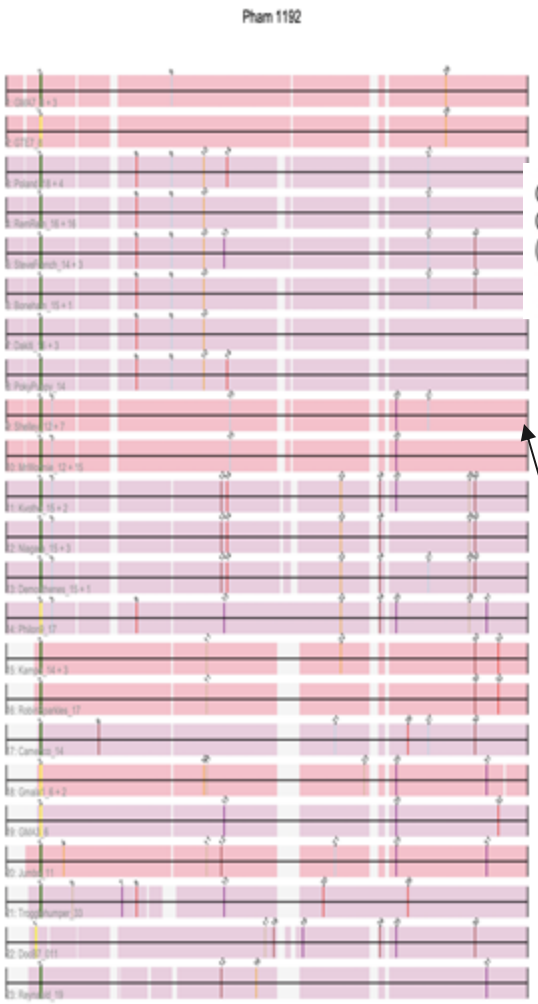


Figure 6:Charianelly Gene 12 Starterator

Manual annotations for 14804 nucleotide

Charianelly 12 on track 9, start indicated in green (2), broad agreement with pham

- Starterator compares each predicted gene within phage's genome with sequence of all members of same pham
- Determines where gene likely starts in consensus
- Able to view how many times nucleotide coordinate has been manually annotated
- Choosing correct start important to differentiate each functional domain

How it works

1. Open starterator report
2. View if phage tracks (right) have same color and same start in green
3. Determine nucleotide coordinate by seeing how many manual annotations there are (More MA = likely start)
4. Observe if there is agreement with cluster and subcluster

Pink = same cluster, Purple = same subcluster, Yellow = called my auto-annotation, Green = Manual Annotations

Start codons are selected using specific annotation programs in relation to similar phages.

- Phage genomes are densely packed with not much overlap measured in base pairs
- Gap of 100+ and Overlap of 30+ rare
- Overlap of 1 or 4 base pairs are best start sites
  - 1 overlap should ALWAYS be selected
- Why it is a strong indicator: Ribosome slide back 1 bp or forward 2bp to translate next gene
- Useful because lower overlap predicts start site better due to packed genome

EXAMPLES

4 bp overlaps:

ATGA

GTGA

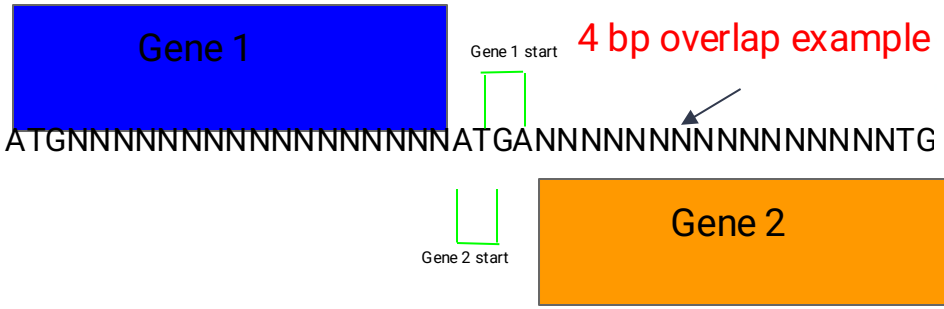
TTGA

1 bp overlaps:

TGATG

TAATG

Charianelly 12 Gap: Only 2 starts below acceptable range of 100, No -1 or -4 overlap



Show 10 entries Search:

Direction	Start	Stop	Length	Gap	Spacer	Z-score	Final Score	LORF	Start Codon	All GM Coding Capacity	Selected Gene
Forward	14804	15274	471	36	12	2.929	-2.786	TRUE	ATG	Select	☑
Forward	14816	15274	459	48	14	1.745	-5.786		GTG		☐
Forward	14987	15274	288	219	10	2.142	-4.299		GTG		☐
Forward	15131	15274	144	363	16	1.303	-7.165		GTG		☐

Figure 7: Charianelly Gap/Overlap

The gaps in genes also guide start site selection

- No final start calls were changed
- Gene 43, 66 and 70 had an inconclusive start position
- Glimmer and GeneMark for gene 43, 66, 68 do not agree on the start position
- Gene 67 had no geneMark start site data, so the position is based on Glimmer analysis

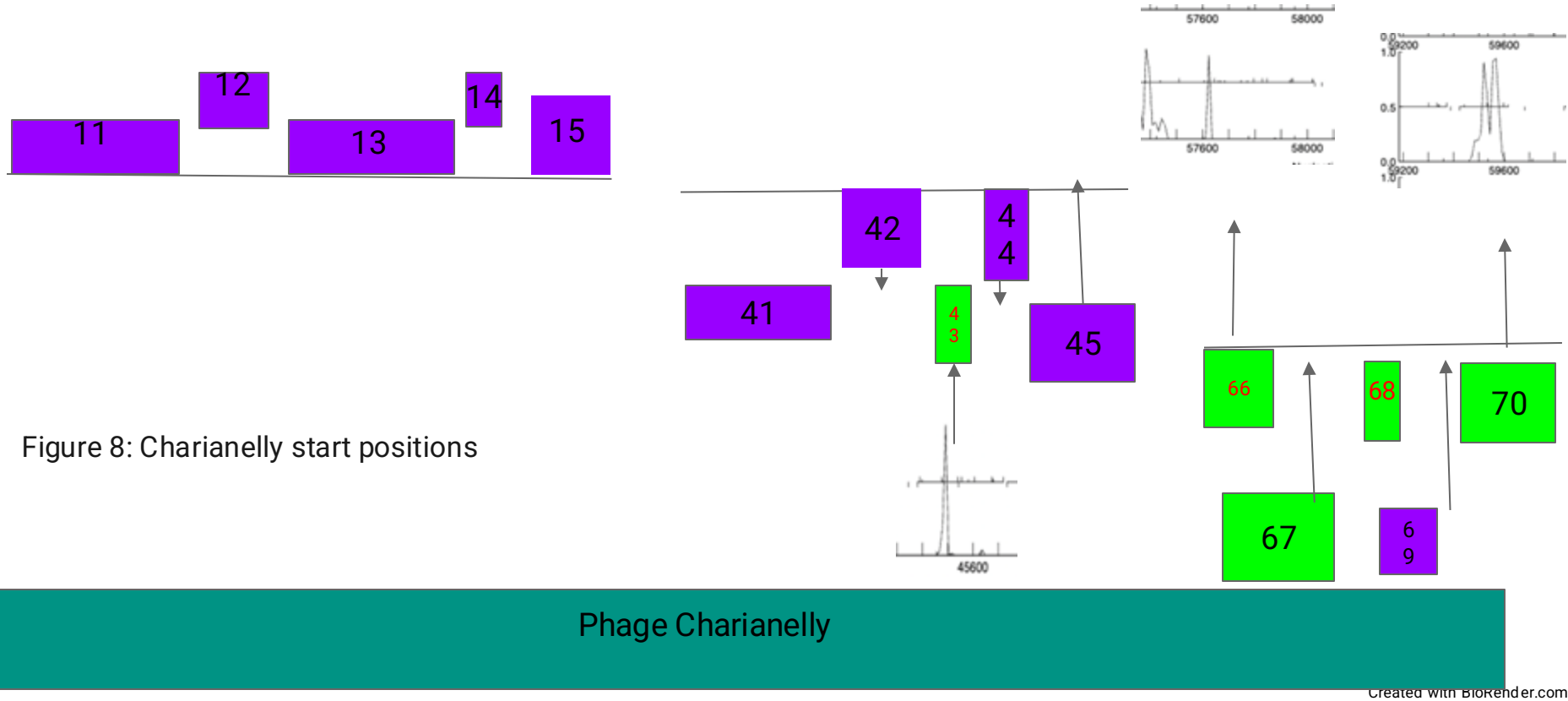


Figure 8: Charianelly start positions

Based on our results the Glimmer and GeneMark start codons were either changed or left as they were originally annotated

HHpred

- Alternate option to BLAST to determine sequence similarity
  - Useful to help us find similarity to phages that have established functions that are most likely present in our phage as well
- Based on structural similarity
- Consists of both similarity probability as well as a visualization of said similarity

Figure 9: HHpred for Charianelly

BLASTp

- Helps determine sequence similarity, along with HHpred.
- Runs the amino acid sequence of predicted proteins through the NCBI database
  - Our phage's sequence gets measured against every other phage protein sequence in the database
- Generates both a "similarity" value and an "E value"
  - Based on BLOSUM matrix which weighs how likely an amino acid occurs in reality

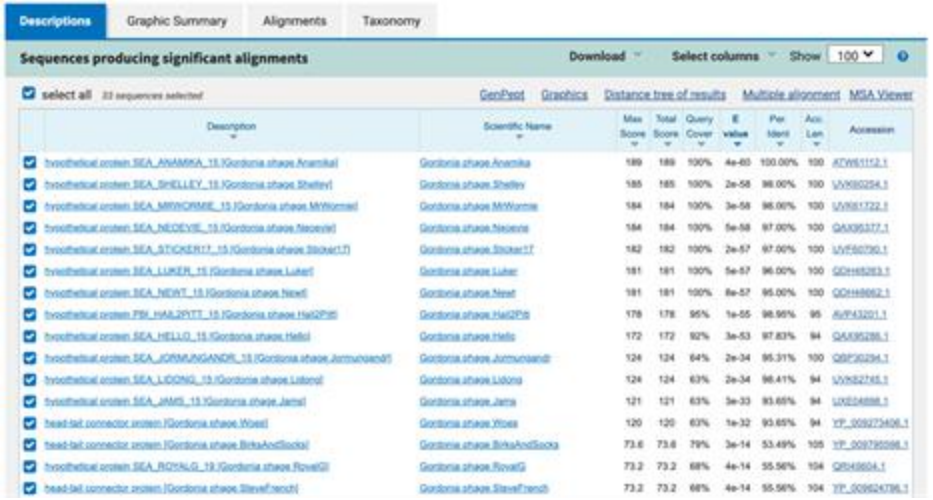
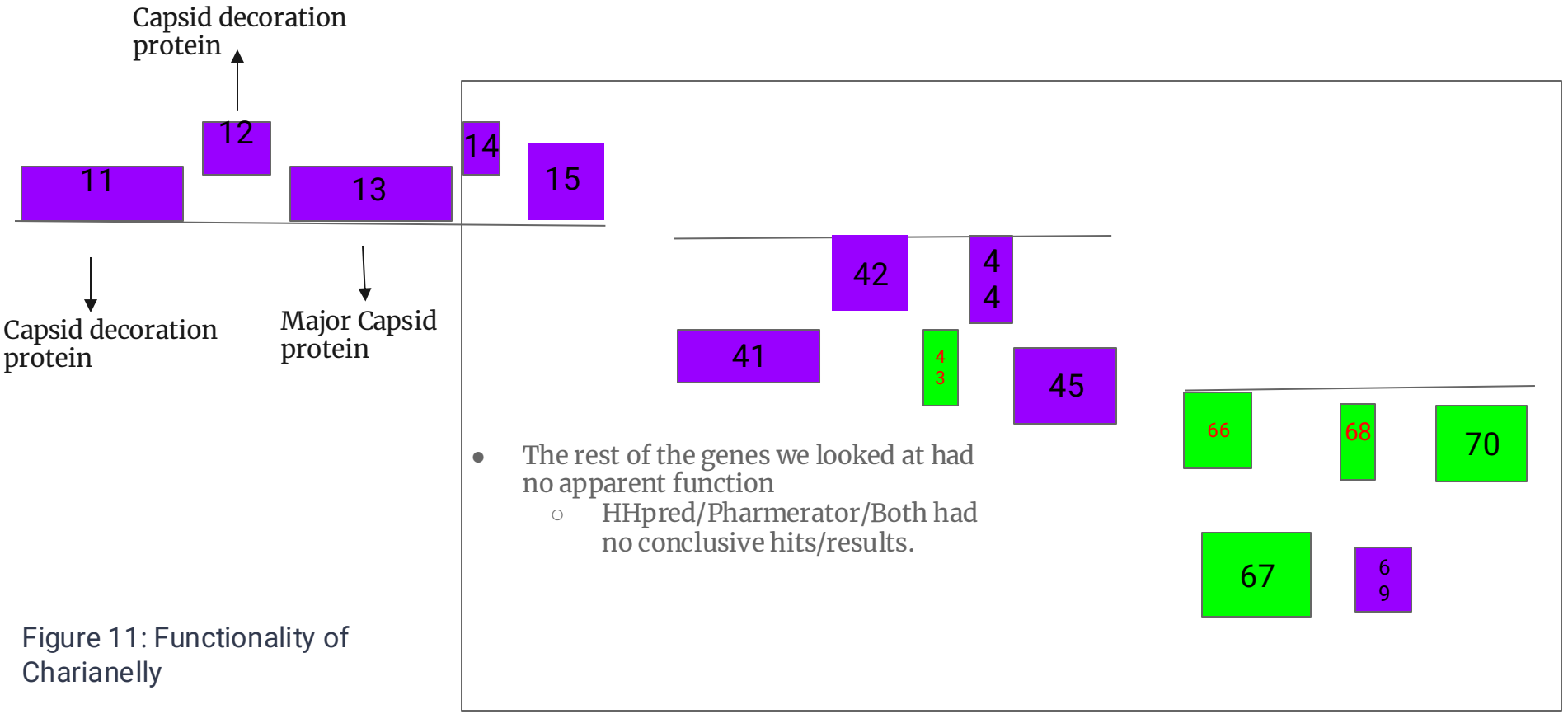


Figure 10: BLASTp for Charianelly

What did we use to determine sequence similarity?

- Genes 11,12, and 13 had a function identified.
  - Both HHpred and Pharmerator both showed positive results for them to be identified as such



- The rest of the genes we looked at had no apparent function
  - HHpred/Pharmerator/Both had no conclusive hits/results.

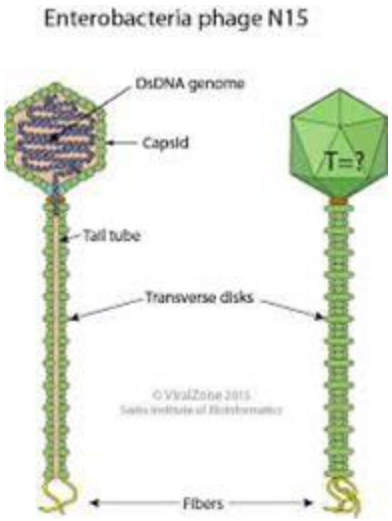
Figure 11: Functionality of Charianelly

Functionality of our various genes for Charianelly varied between various proteins and the majority having no function.

Host range is described as the range of cell types a host species is able to infect
The importance of knowing the host range

The background research of annotating the charianelly phage, and understanding what classifies as a gene laid the groundwork for the host range project

Minor tail proteins are a crucial part in the assembly of the tail in the bacteriophage that makes it easier for viruses to infect their hosts



The host range project conducted consisted of 6 different phages each tested on three different strains of G. *terrae*. Our group used the phage Capybara and tested the infection patterns on G. *lacunae,* G. *westfalia*, and G. *rubripertincta*. We hypothesized that all the phages would have similar infection patterns on the different hosts

Importance

- The minor tail protein aids the phage identify in the correct host, and then break the surface so the DNA can be injected.
- Host range can show the dynamic of the survival of pathogens which can be a leading factor in understanding evolution patterns

Figure 12: Phage N15

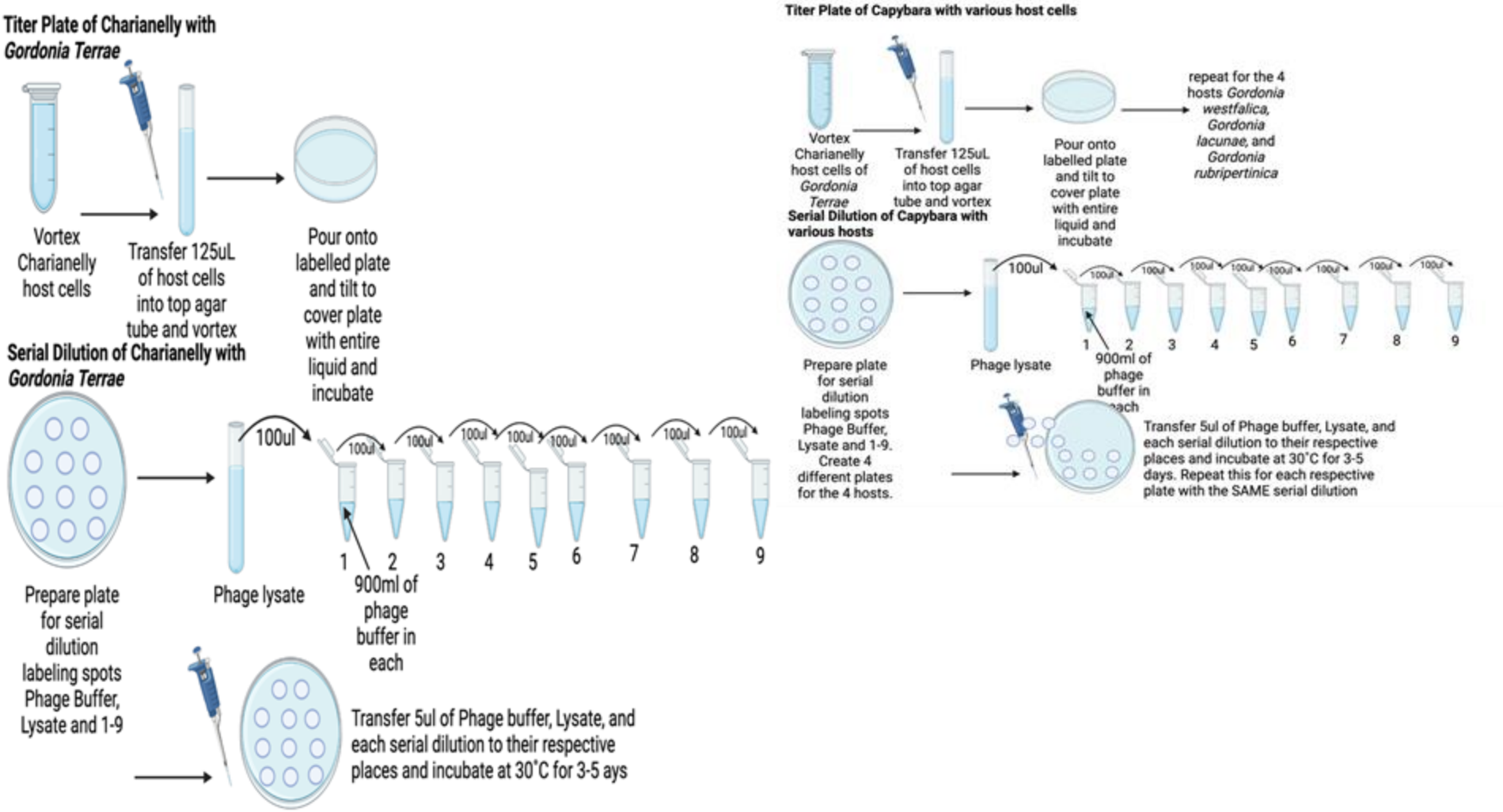
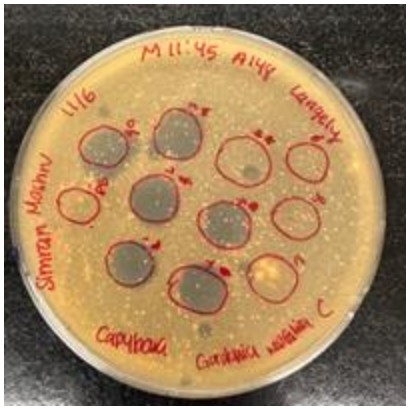


Figure 16: Methods Flow Diagram for Phages Charianelly and Capybara

Phage capybara was able to be infected by all host other than G. *lacunae*

- Tested the Capybara phage in the DJ cluster
- Phage capybara infected all hosts other than G. *lacunae,* with roughly the same efficiency between the other three
- Similar infection pattern with all other phages except belphegor where G. *lacunae* was not able to infect
- Phage capybara shares no other minor pham tails with any other Pitt F23 pages

G. *terrae*



G. *westfalia*

G. *lacunae*

G. *rubripertincta*

The significance of these findings is that there is no similarity between the other Pitt F23 phages and that 3 out of the 4 hosts were able to infect

Figure 17: Serial dilution with alternate hosts using phage Capybara

Conclusions

- Charianelly cluster: CS
- Subcluster: CS3, 24 phages in this subcluster (large)
- GC content: 59.1%
- Total Genes Predicted: 93 (Bianmat and Minos 94)

- **Similar with other Gordonia phages because minor tail phams match up**

Mechanism: If there is not an appropriate host identity, then the minor tail protein cannot break the surface and no DNA can be injected leading to no injection.

- Annotated 15 genes
  - Started changed:
  - Functions assigned: 3
- 2 minor tail proteins on 30 and 31 for Capybara but did not match up with Charianelly so cannot compare infection patterns
- Charianelly had a broader host range; Capybara infection pattern different
- **Minor Tail proteins:** aids the phage identify in the correct host, and then break the surface so the DNA can be injected.

Author Contributions

Simran: completed slides 4, 6, 9, 10, 15, 18
Ayaan: completed slides 3, 7, 8, 12, 13, 18
Reagan: completed slides 5, 11, 14, 16, 17, 18

We all practice our slides and reviewed the groups slides together