Comparison of bacteriophage annotation methods

Alicia Salisbury and Philippos Tsourkas
School of Life Sciences, University of Nevada Las Vegas
McNair Summer Research Internship

Abstract:

The rise of antibiotic-resistant bacteria has increased interest in bacteriophages (viruses that kill bacteria) in recent years. Due to the decreasing cost of genome sequencing, the number of sequenced phage genomes is growing at a geometric rate. Sequencing is followed by annotation, in which genes, start codons, and putative protein functions are identified. Most phage genomes are auto-annotated with programs designed for prokaryotes. Accuracy metrics for these programs with regard to phage genomes are not available. The genome of *Escherichia coli* phage Lambda was used to benchmark the accuracy of several genome annotation methods and programs. Discovered in 1951, Lambda is the most well studied phage, with nearly all gene functions and start sites demonstrated experimentally. Eight programs were used to annotate the Lambda genome: Glimmer, BASys, RAST, GeneMark, GeneMark.hmm, GeneMarkS, GeneMarkS2, and GeneMark with Heuristic models. Calls were compared to the reference genome from the literature.

Goal: To determine the accuracy of the eight selected programs in regard to bacteriophage genome annotation.

Hypothesis: Manual curation and compilation of auto-annotation results obtained from several programs will yield more accurate gene feature and start codon prediction than auto-annotation alone.

Methods:

Step 1 → Lambda Annotation

→ Lambda raw nucleotide sequence obtained from GenBank.

→Auto-annotation with each program under investigation.

→ Results tabulated and compared to protein interaction map to find TN, TP, FN, and FP.

Step 2 → Random Sequence Analysis

→Random nucleotide sequence (40,000 bases) generated in R.

→Auto-annotation with each program.

→Results tabulated and compared to obtain a clearer picture of the TN and FP.

Step 3 → Manual Annotation

→Auto-annotation with Glimmer and 5 GeneMark programs.

→Gene calling (features called by >6 programs)

- Evaluated on ntBLAST, coding potential (CPM), calls, length, and overlap.

→Start Codon Calling (all potential starts for each gene)

- Evaluated on pBLAST, program calls, length, codon, overlap, CPM, and likelihood of translational start (RBS).

Step 4 → Pseudogene Analysis

→Pseudogenes called by ≥1 non-Glimmer program were scrutinized (all called starts, 8 pseudogenes and 9 starts total).

- Evaluated on ntBLAST, pBLAST, CPM, program calls, length, codon, and RBS.
- Overlap, operon formation, and synteny were unavailable.

→Results expressed in terms of hypothetical synteny, overlap, and operon formation.

Calculations:

Sensitivity

True Positive Rate (TPR)

Describes the proportion of genes called correctly

$$TPR = \frac{TP}{TP+FN}$$

Precision

Positive Predictive Value (PPV)

Describes the probability that a call is a gene

$$PPV = \frac{TP}{TP+FP}$$

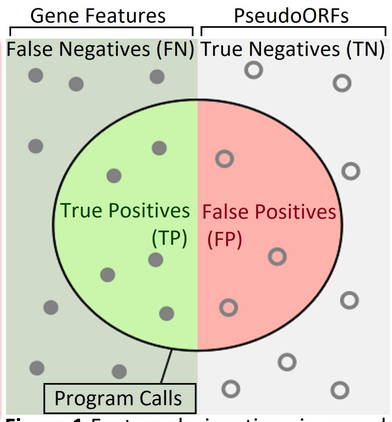


Figure 1 Feature designations in regard to phage genomics. A pseudoORF is a continuous reading frame ≤75bp long.

Source: Image modified from https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Data:

Table 1 Calling programs, coordinates, length,CPM for all PseudoORFs generated by non-Glimmer programs in the randomly generated sequence.

Feature	1	2	3	4	5	6	7		8
Direction	R	F	R	F	R	F	F		R
Stop	6259	7068	7403	22324	27950	31520	35911		39017
Start	6384	6904	7570	22130	28174	31299	35636	35711	39196
GeneMark		called	called	called	called	called		called	
Glimmer								called	called
GeneMark.hmm			called		called		called		
GeneMarkS		called	called	called	called	called		called	
Heuristic GeneMark		called	called	called	called	called		called	
RAST	called			called		called			called
Length	126	165	168	195	225	222	276	201	180
Coding Potential									

Table 2 Positives and negatives. 73 genes, 545 true negatives in the reference.

Gene Calls	GeneMark	.hhm	S	Heuristic	S2	RAST	BASys	Glimmer	Manual
False Positives	3	4	3	3	2	2	4	4	4
True Positives	58	62	58	58	58	62	62	63	65
True Negatives	542	541	542	542	543	543	541	541	541
False Negatives	16	11	15	15	15	11	11	10	9
Sensitivity (TPR)	0.784	0.849	0.795	0.795	0.795	0.863	0.849	0.849	0.878
Precision (PPV)	0.951	0.939	0.951	0.951	0.967	0.940	0.969	0.939	0.942

Table 3 Starts called long or short relative to the reference genome.

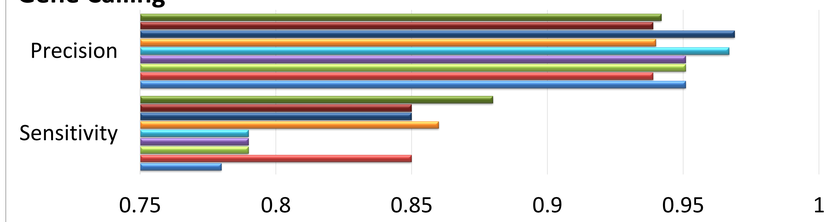
Start Calls	GeneMark	.hhm	S	Heuristic	S2	RAST	BASys	Glimmer	Manual
Called Long	6	1	6	6	3	2	1	4	1
Called Short	5	7	5	5	7	7	5	7	4
Genes Called	58	62	58	58	58	62	62	63	64
% Accurate	81.03%	87.10%	81.03%	81.03%	82.76%	85.48%	90.32%	82.54%	92.19%

Table 4 PseudoORFs generated for the randomly generated sequence, in which 628 true negatives exist. GeneMarkS2 and BASys returned no result.

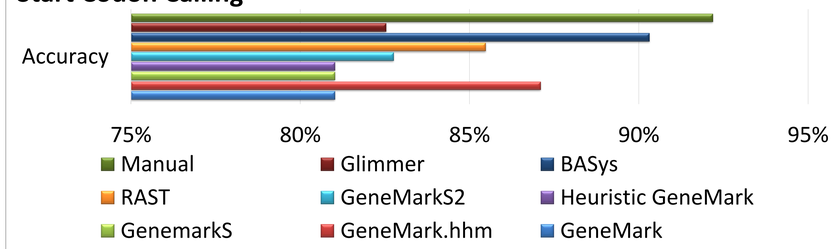
Generated	GeneMark	Glimmer	.hhm	S	Heuristic	RAST
False Positives	6	47	3	6	6	4
True Negatives	622	581	625	622	622	624

Results:

Gene Calling



Start Codon Calling



Conclusions:

- Manual annotation is slightly more accurate, particularly in start calling.
- No gene called by all programs was a false positive.
- Some genes were not detected by any annotation method.
- Glimmer is disproportionately prone to nonsense FN.
- Pseudogenes called by any program other than Glimmer represent borderline features, which, during manual annotation, would be:
  - Deleted in the absence of any additional evidence.
  - Kept if they satisfied one of the following conditions:
    1. Filled a gap completely without generating overlap or direction change.
    2. Created 4bp overlap on one or both ends, suggesting an operon.
- Pseudogenes called in the random sequence have:
  1. No plateaus of coding potential.
  2. No significant ntBLAST result.
  3. No significant pBLAST results.
  4. Coding potential does not align well with start and STOP coordinates.

Future Research:

- Uncalled genes will be scrutinized and compared to pseudogenes called in the random sequence.

Acknowledgements:

This research is supported by the McNair Scholars Summer Research Institute.

Thanks to the Office of Undergraduate Research Summer Undergraduate Research Funding Scholarship (OUR SURF), and to the Mechanisms of Evolution Research Experience for Undergraduates (MOE REU).

UNLV Office of UNDERGRADUATE RESEARCH

Career Exploration · Working like a Scientist · Science Communication

UNLV Mechanisms of Evolution

CSUN UNLVCSUN STUDENT GOVERNMENT

TRIO RONALD E. MCNAIR POST-BACCALAUREATE ACHIEVEMENT PROGRAM

UNLV College of SCIENCES