Computational Gene Finding in *Plasmodium falciparum*

Simon Cawley

U. C. Berkeley

Outline

1 Programs
2 Data Sets
3 Analysis
4 Results

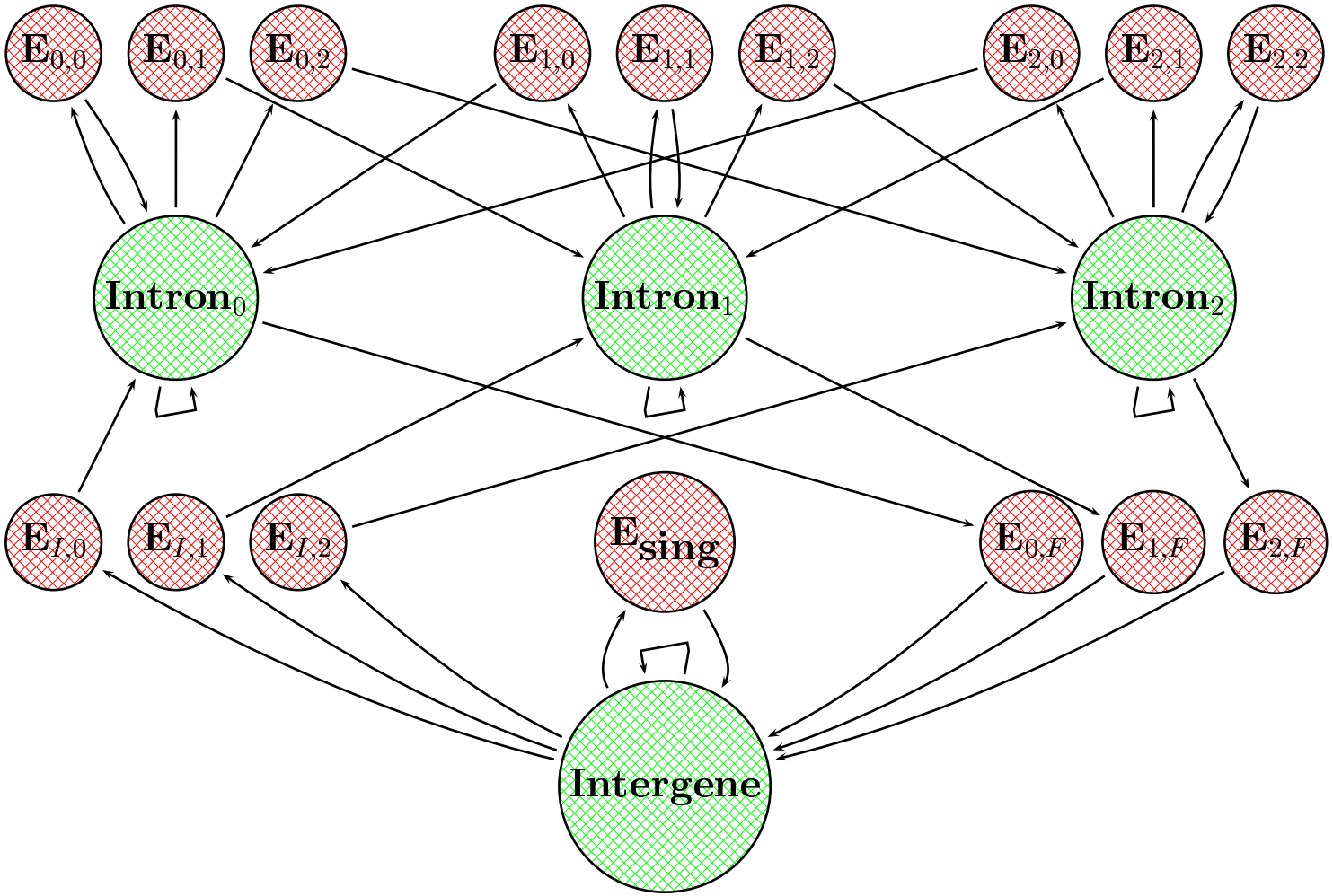
Gene Finding Software for *P. falciparum*

Hexamer	*Richard Durbin*
Genefinder	*Phil Green & Colin Wilson*
GlimmerM	*Steven Salzberg, Mihaela Pertea, Arthur Delcher*
	Malcolm Gardner & Herve Tettelin.
Phat	*Anthony Wirth, Simon Cawley & Terry Speed*

Model Components

Signal Type	Hexamer	Genefinder	GlimmerM	Phat
Coding Exon Content	√	√	√	√
Intron Content		√		√
Intergene Content		√		√
Coding Exon Lengths				√
Intron Lengths				√
Intergene Lengths				√
Intron/Intergene Transitions				√
Splice Sites		√	√	√
Transcription Start Site		√		
Branch Point				
Polyadenylation Signal				
Promoter Signals				
UTR (Non-Coding Exons)				

State space of half Phat generalized HMM



Phat

Some limitations

- No overlapping genes.
- No frameshifts.
- Can't rule out stop codons spanning an intron.

Some convenient features for analysis of draft sequence

- Can search for incomplete genes (lacking initial/terminal exon)
- Can allow for in-frame stopcodons, which are more frequent in error-prone draft sequence (`-okstops` option).

Training Data

Hexamer and Phat were trained on (most) of the GlimmerM training set, all genes confirmed either by homology or by RT-PCR.

103 genes from Chromosome 3

111 genes from Chromosome 2, (116 less 5, excluded due to absence of CDS annotation)

118 genes from various chromosomes (Original GlimmerM training set of 110 genbank entries, 3 entries excluded due to non-standard CDS annotations)

Training Data Summary

332	genes (178 single exon, 154 multi exon)
644	exons
697,508	coding bp
63,183	intron bp
874,633	intergene bp

Evaluation Data

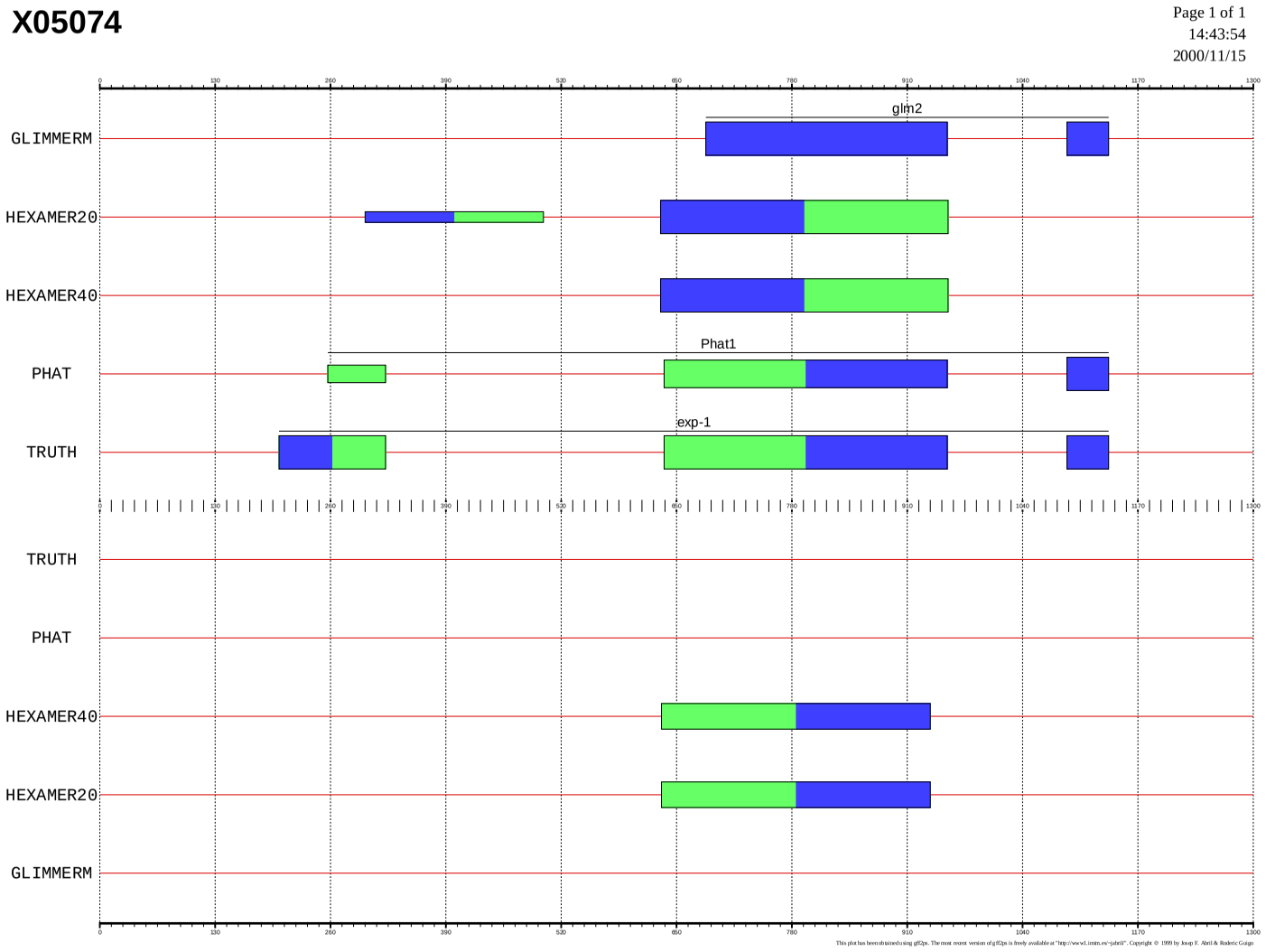
Requests for biologically confirmed gene structures yielded:

- 11 genes courtesy of Neil Hall (Sanger)
- 6 genes courtesy of Sharen Bowman (Sanger)
- 27 genes courtesy of Tony Triglia (WEHI)

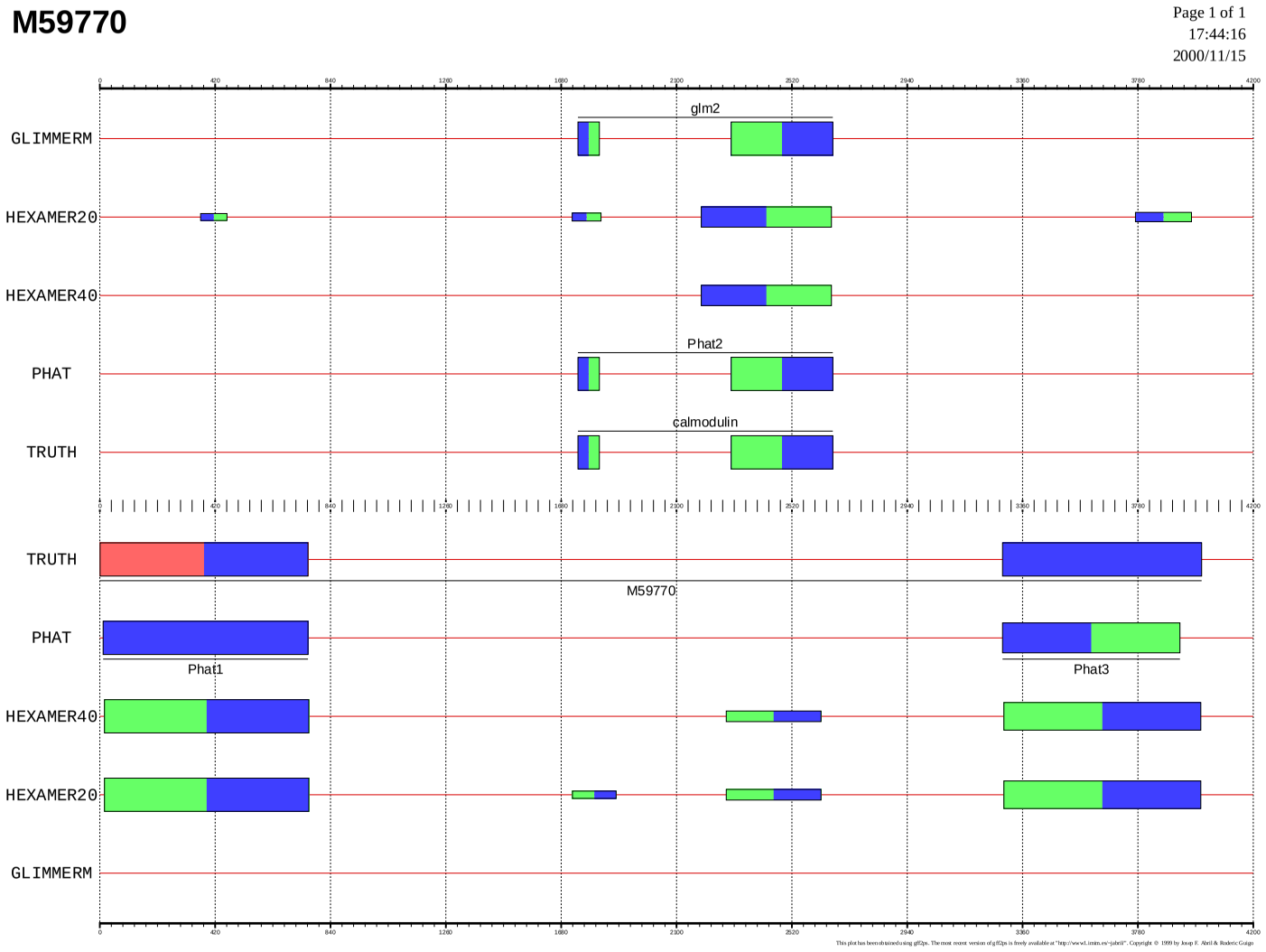
Of these 44 genes,

- 19 are in the GimmerM training set (the Training Set).
- 25 are not (the Test Set).

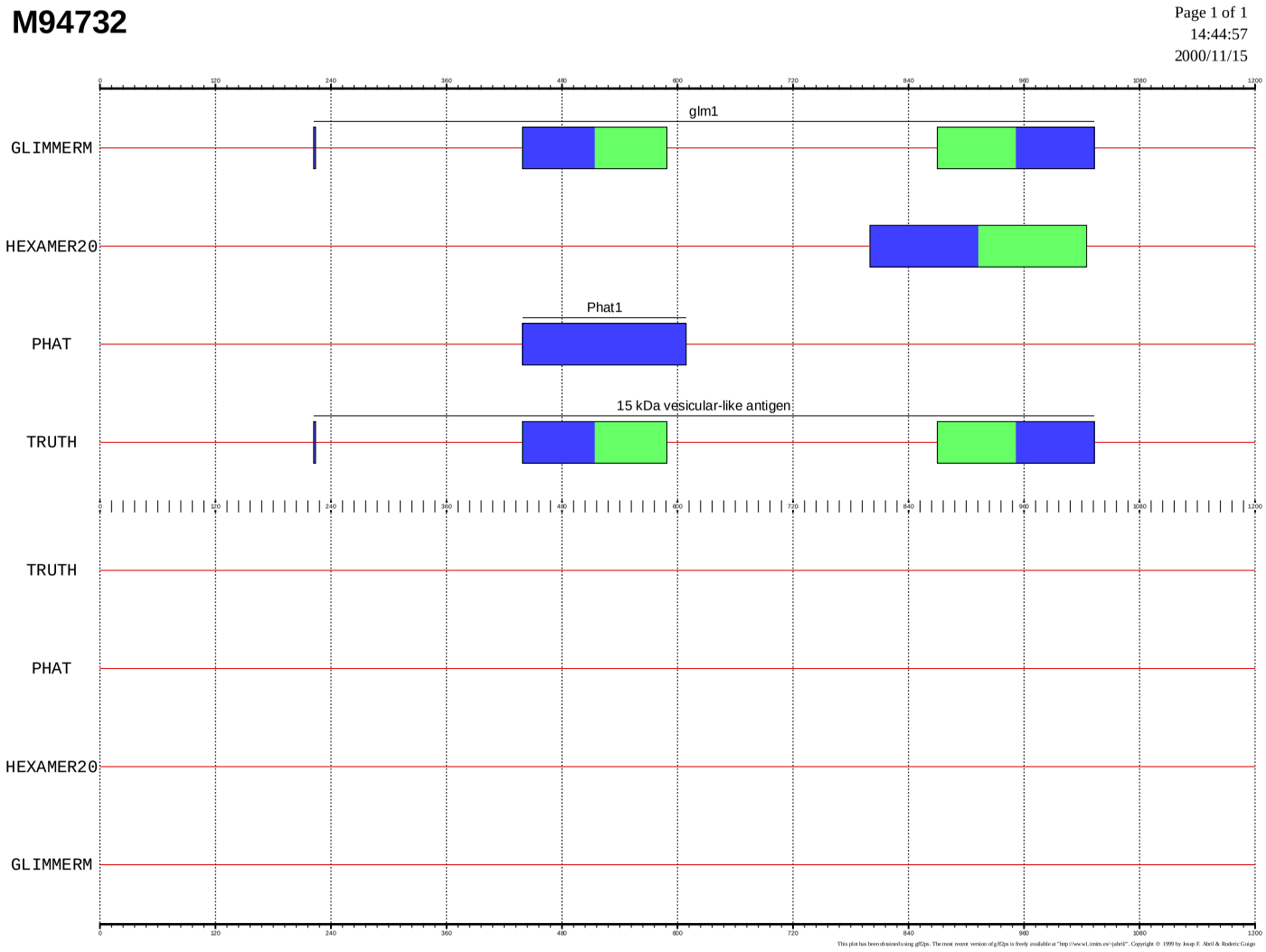
P. falciparum coding sequences often look like coding on the opposite strand too.



A less common example: one gene contained within another

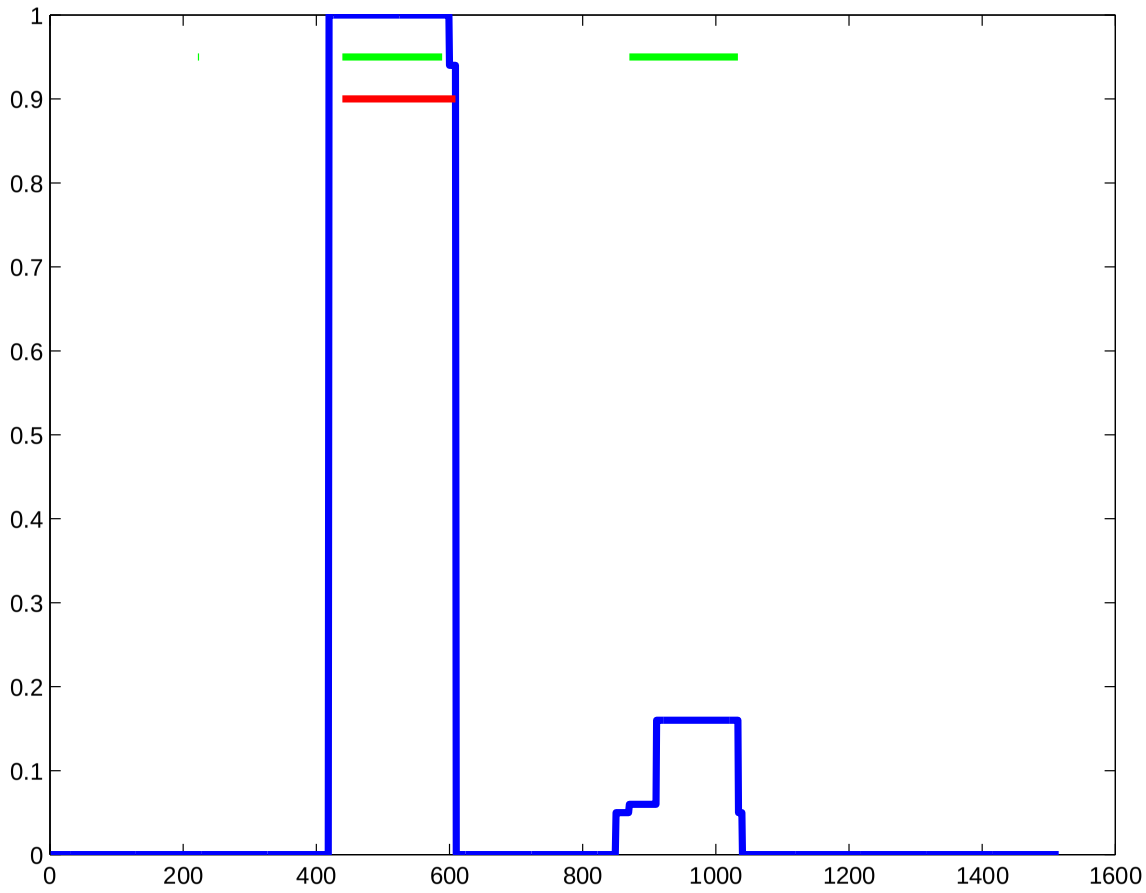


Sometimes GlimmerM does better:

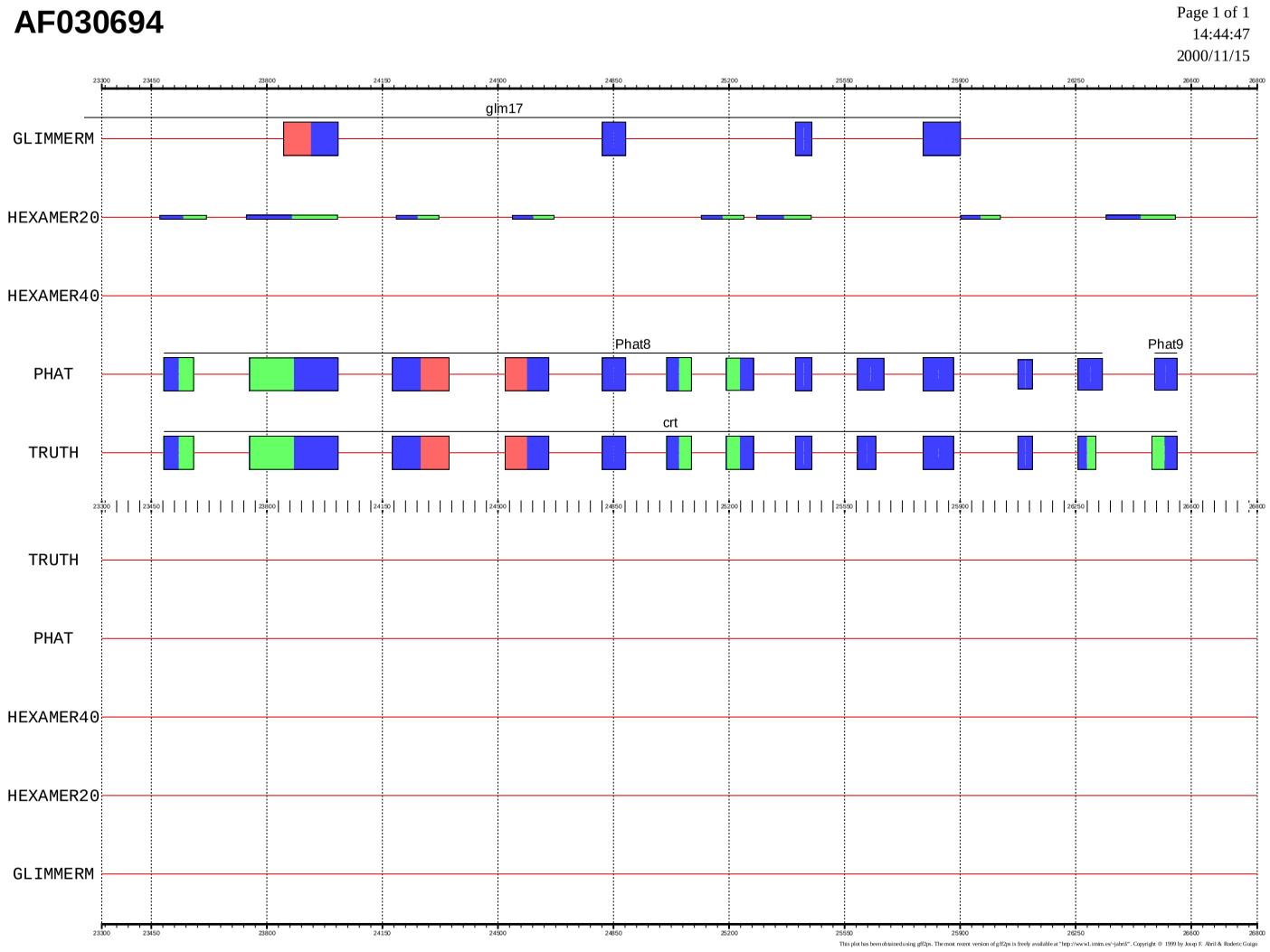


Phat can produce posterior coding probabilities

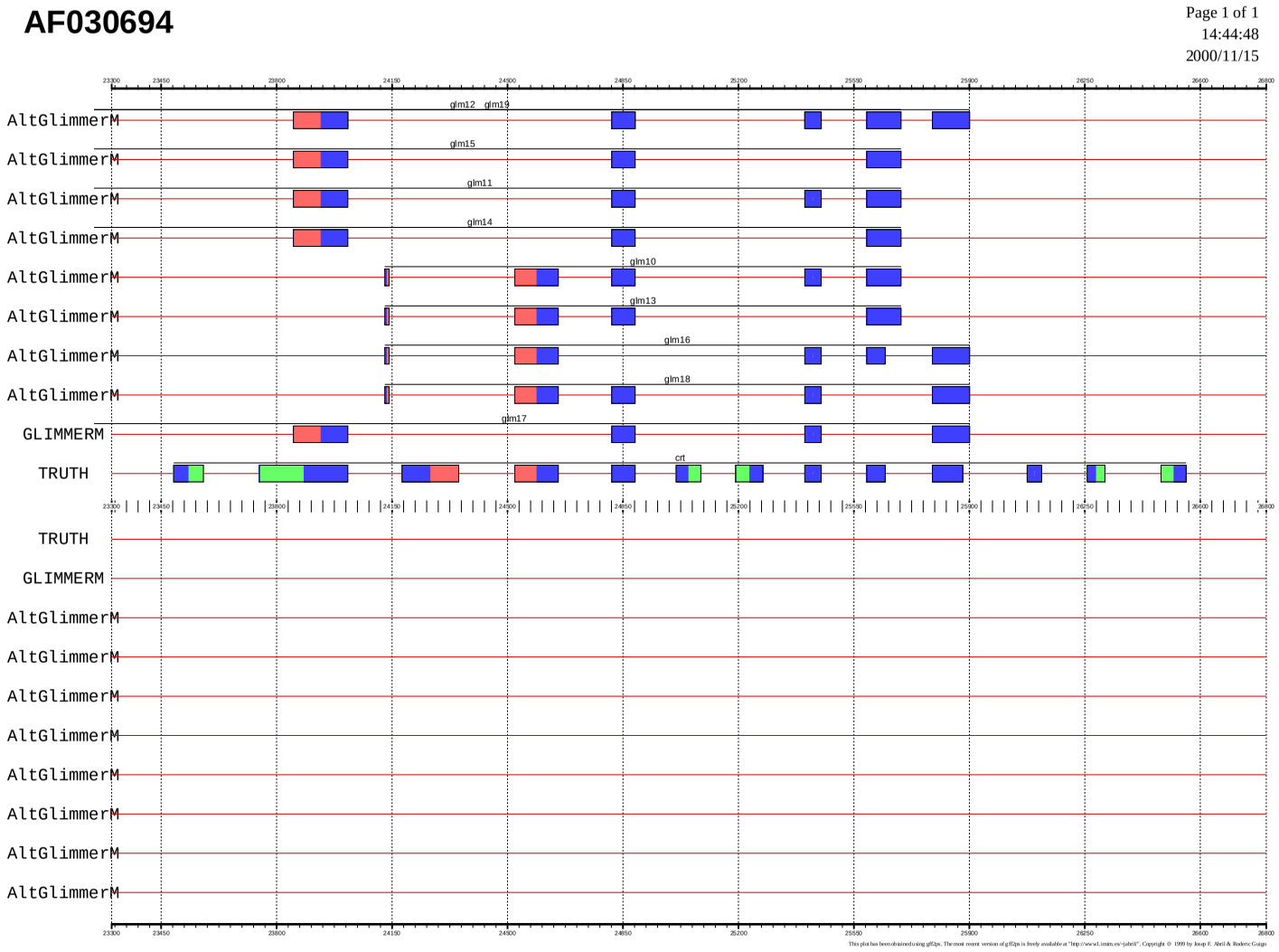
(Green = truth, red = Phat prediction)



Sometimes Phat does better:



GlimmerM provides alternative parses

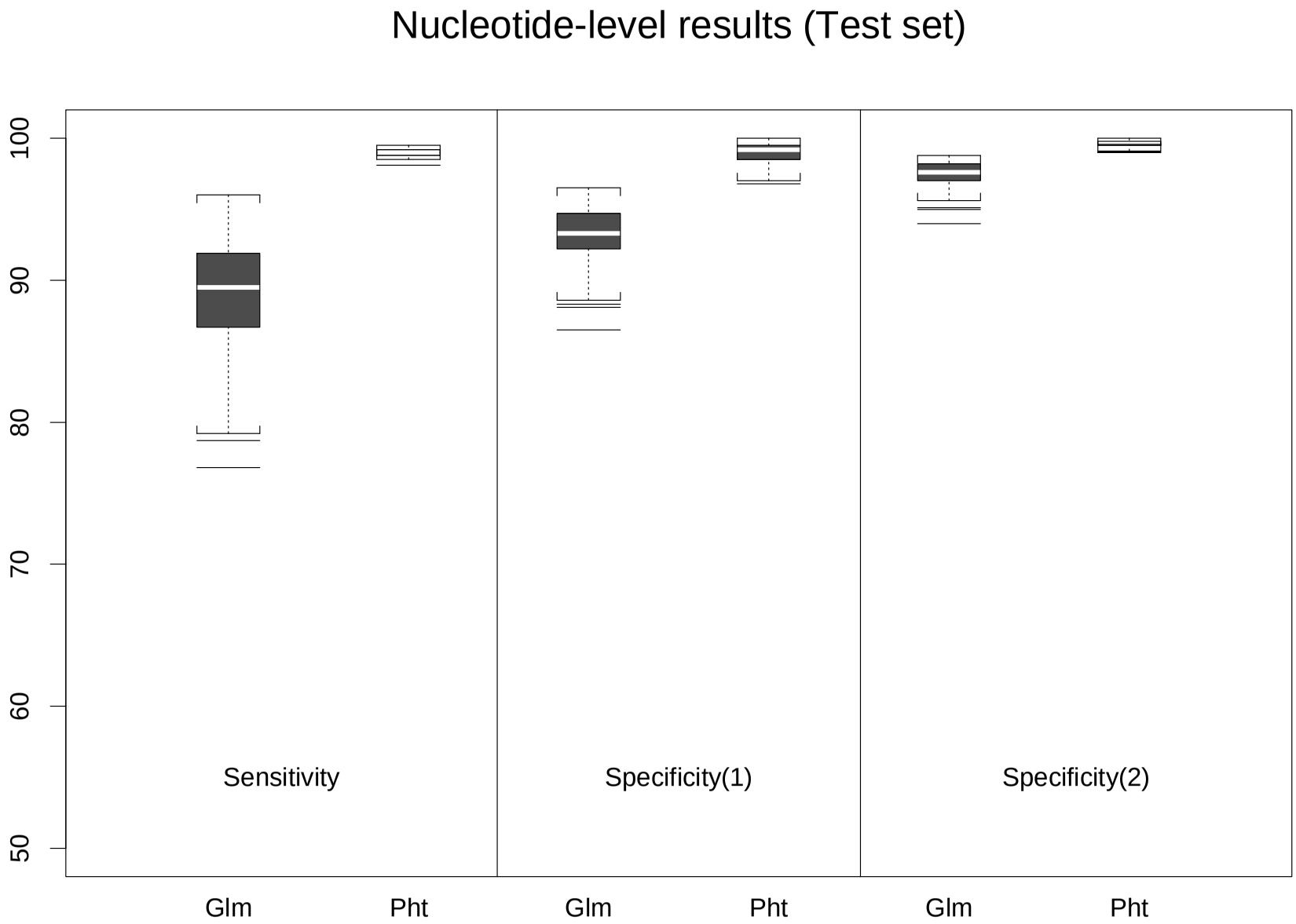


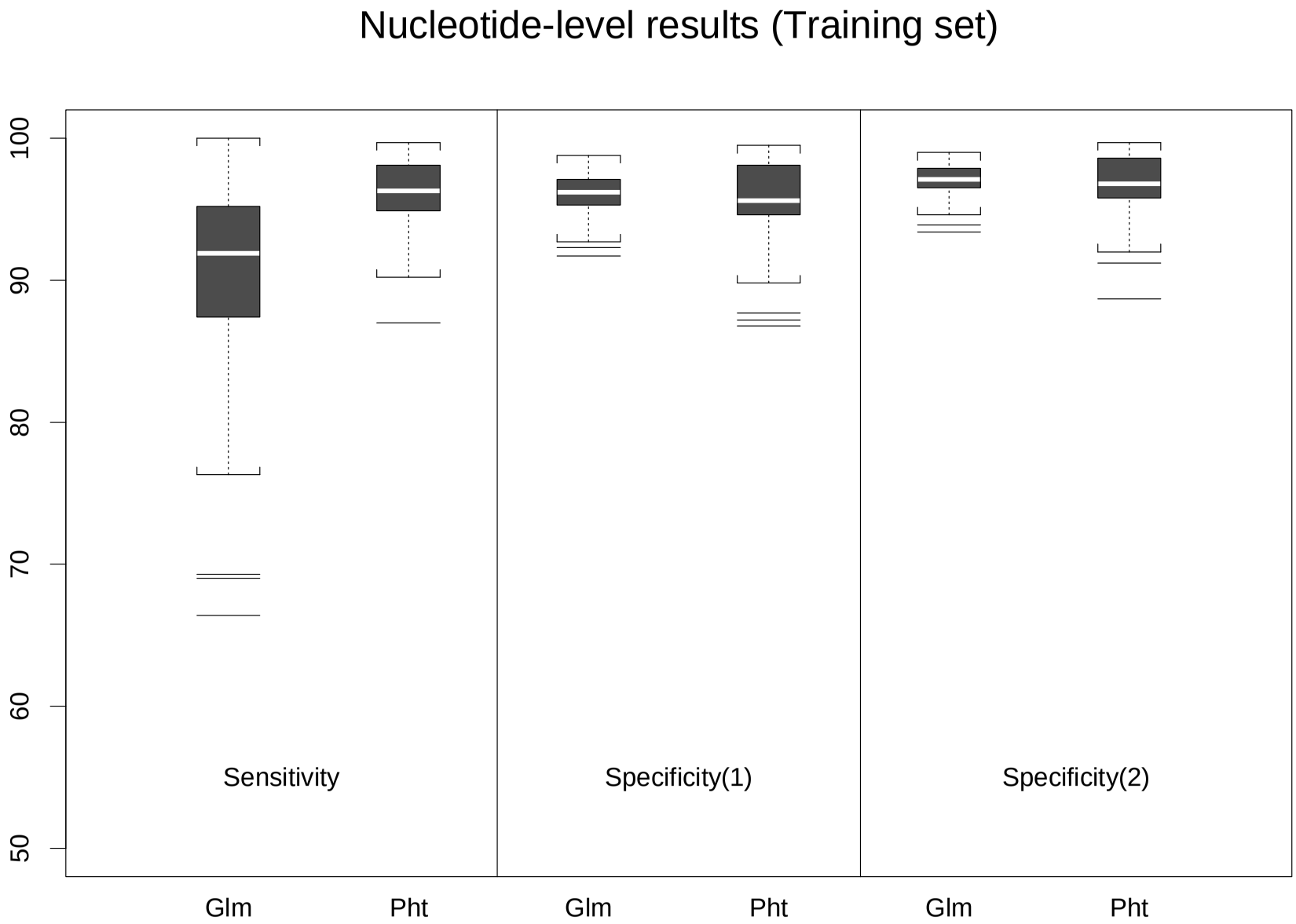
Summary results

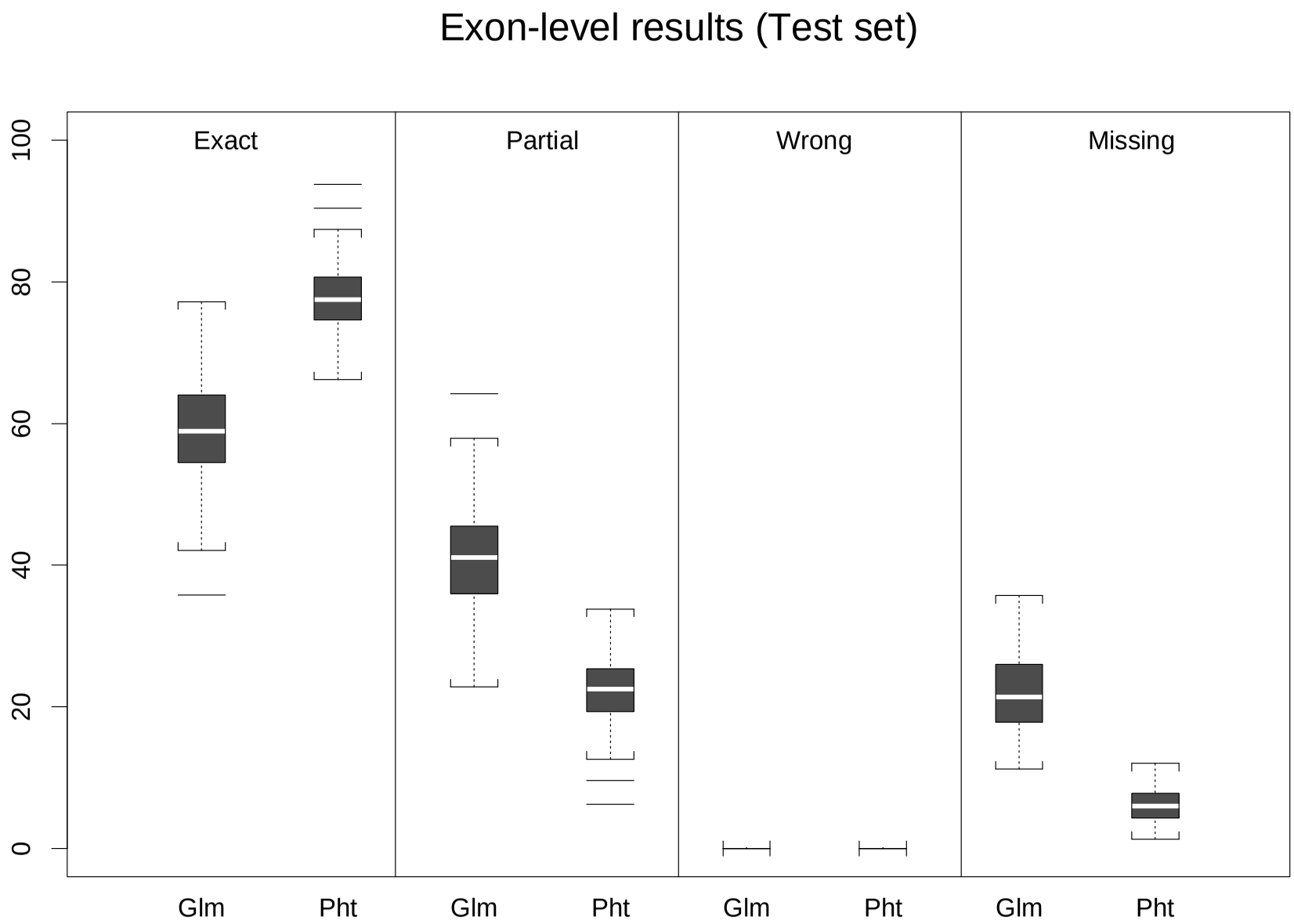
$$Sn = TP/TP + FN \quad \text{(Proportion of positives correctly annotated)}$$
$$Sp1 = TN/TN + FP \quad \text{(Proportion of negatives correctly annotated)}$$
$$Sp2 = TP/TP + FP \quad \text{(Proportion of predictions which are correct)}$$

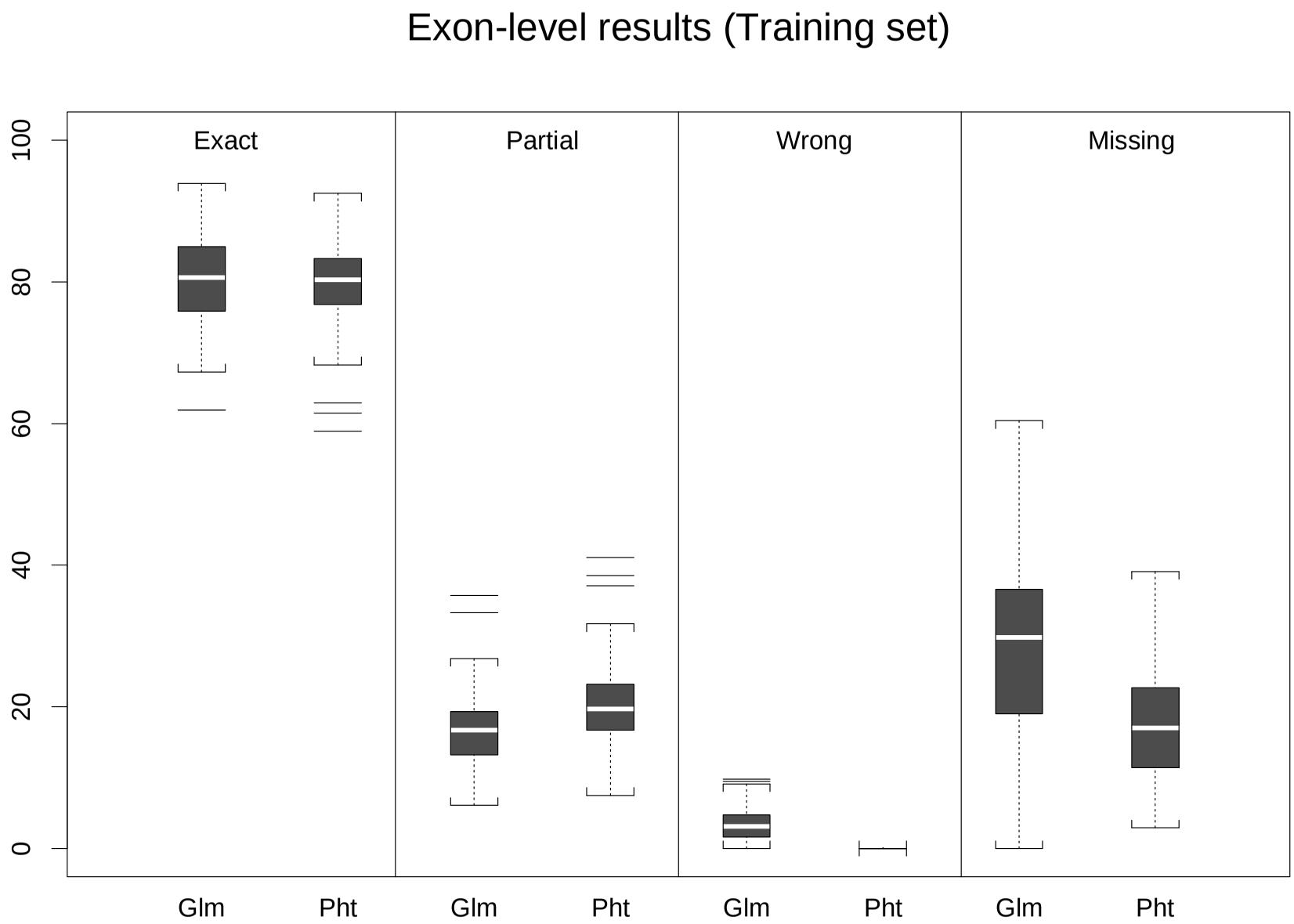
Test Set	Nucleotide-Level			Exon-Level			
Program	Sn	Sp1	Sp2	Correct	Partial	Wrong	Missing
GlimmerM	89.3	93.1	97.6	57.8	42.2	0.0	22.6
Phat	99.0	98.9	99.6	77.2	22.8	0.0	6.0

Train Set	Nucleotide-Level			Exon-Level			
Program	Sn	Sp1	Sp2	Correct	Partial	Wrong	Missing
GlimmerM	90.2	96.2	97.1	79.7	16.9	3.4	30.6
Phat	95.8	95.1	96.5	80.3	19.7	0.0	16.5


Nucleotide-level results (Test set)
100
90
80
70
60
50
Sensitivity
Specificity(1)
Specificity(2)
Glm
Pht
Glm
Pht
Glm
Pht


Nucleotide-level results (Training set)
100
90
80
70
60
50
Sensitivity
Specificity(1)
Specificity(2)
Glm
Pht
Glm
Pht
Glm
Pht


Exon-level results (Test set)
Exact
Partial
Wrong
Missing
0
20
40
60
80
100
Glm
Pht
Glm
Pht
Glm
Pht
Glm
Pht


Exon-level results (Training set)
Exact
Partial
Wrong
Missing
0
20
40
60
80
100
Glm
Pht
Glm
Pht
Glm
Pht
Glm
Pht

Software Details

	Hexamer	Genefinder	GlimmerM	Phat
Availability	Source code freely available to all	Source code free to academia, available through licence for industry	Binary free to academia, available through licence for industry	Source code freely available to all (Gnu GPL)
Re-trainable?	Yes	Yes	In progress	Yes

Acknowledgments

Phat

Simon Cawley
Tony Wirth
Terry Speed

(authors contributed equally)

Data

Steven Salzberg
Mihaela Pertea
Sharen Bowman
Neil Hall
Win Hide
Ralhston Muller
Tony Triglia

For more info:

Hexamer	`ftp://ftp.sanger.ac.uk/pub/pathogens/software/hexamer`
Genefinder	`email colin@u.washington.edu`
GlimmerM	`http://www.tigr.org/softlab/glimmerm`
Phat	`http://www.stat.berkeley.edu/users/scawley/Phat`