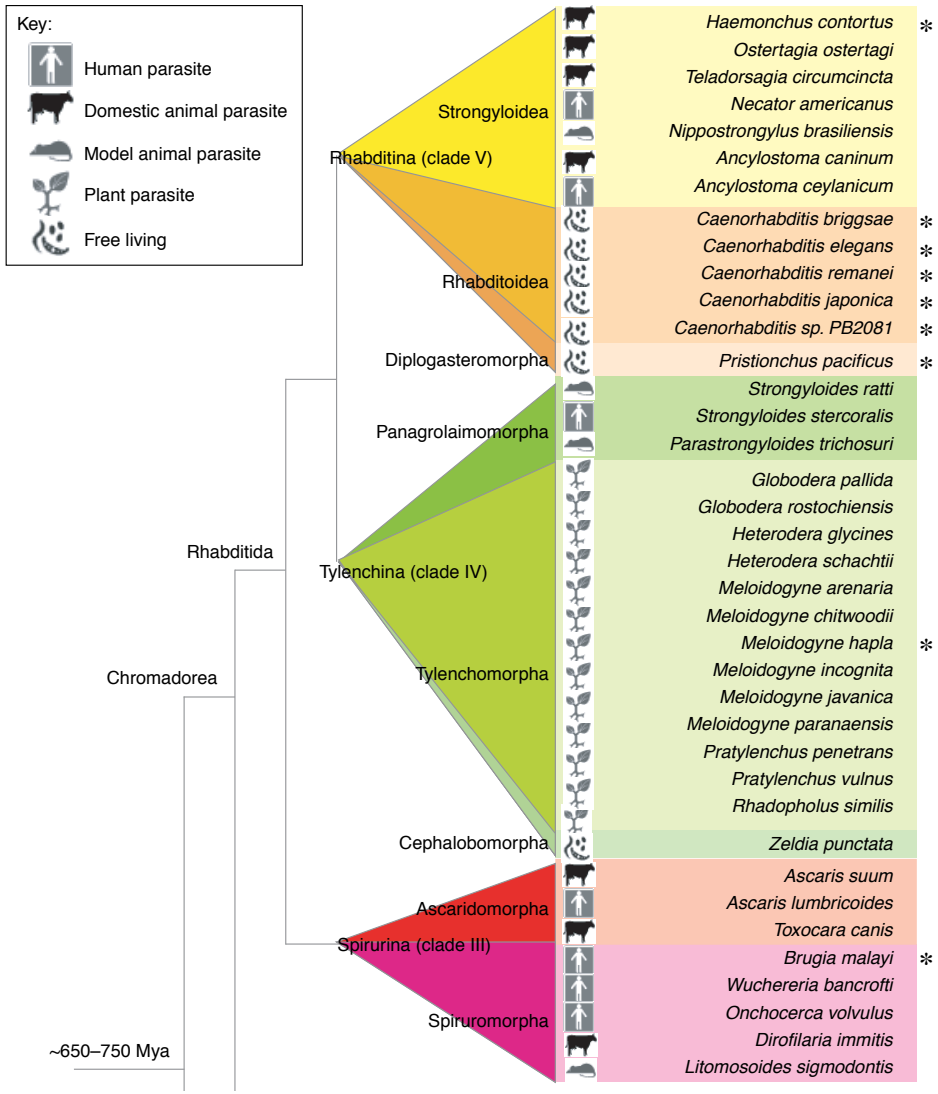
mGene: A Novel Discriminative Gene Finding System

Gabriele Schweikert

Max Planck Institutes
Tübingen, Germany

Worm Genomics and Systems Biology Workshop,
Cambridge, July 24, 2008

Introduction



Mitreva, 2005

- 1998: *C. elegans* genome completed
- Today: several more genome projects finished / underway
- Genomes have to be annotated!

Use of cDNA and EST sequences

- Creation of cDNA libraries; selection of random clones
  - ⇒ High-copy-number mRNAs overrepresented
  - ⇒ Low-copy-number mRNAs missed entirely
  - ⇒ Mostly ESTs (single sequencing reads 500-700 nucleotides)
- Alignment of EST cDNA sequences against genome
  - ⇒ Cis-alignment; mostly corect
  - ⇒ Trans-alignment of homologous genes
- Typical cDNA sequencing project 20-40% of transcripts sequenced incorrectly or not at all

⇒ Use cDNA and EST alignments as labeled training set

⇒ Predict mRNA & gene products

Use of cDNA and EST sequences

- Creation of cDNA libraries; selection of random clones
  - ⇒ High-copy-number mRNAs overrepresented
  - ⇒ Low-copy-number mRNAs missed entirely
  - ⇒ Mostly ESTs (single sequencing reads 500-700 nucleotides)
- Alignment of EST cDNA sequences against genome
  - ⇒ Cis-alignment; mostly corect
  - ⇒ Trans-alignment of homologous genes
- Typical cDNA sequencing project 20-40% of transcripts sequenced incorrectly or not at all

⇒ Use cDNA and EST alignments as labeled training set

⇒ Predict mRNA & gene products

Use of cDNA and EST sequences

- Creation of cDNA libraries; selection of random clones
  - ⇒ High-copy-number mRNAs overrepresented
  - ⇒ Low-copy-number mRNAs missed entirely
  - ⇒ Mostly ESTs (single sequencing reads 500-700 nucleotides)
- Alignment of EST cDNA sequences against genome
  - ⇒ Cis-alignment; mostly corect
  - ⇒ Trans-alignment of homologous genes
- Typical cDNA sequencing project 20-40% of transcripts sequenced incorrectly or not at all

⇒ Use cDNA and EST alignments as labeled training set

⇒ Predict mRNA & gene products

Genome Annotation

Use of cDNA and EST sequences

- Creation of cDNA libraries; selection of random clones
  - ⇒ High-copy-number mRNAs overrepresented
  - ⇒ Low-copy-number mRNAs missed entirely
  - ⇒ Mostly ESTs (single sequencing reads 500-700 nucleotides)
- Alignment of EST cDNA sequences against genome
  - ⇒ Cis-alignment; mostly corect
  - ⇒ Trans-alignment of homologous genes
- Typical cDNA sequencing project 20-40% of transcripts sequenced incorrectly or not at all

⇒ Use cDNA and EST alignments as labeled training set

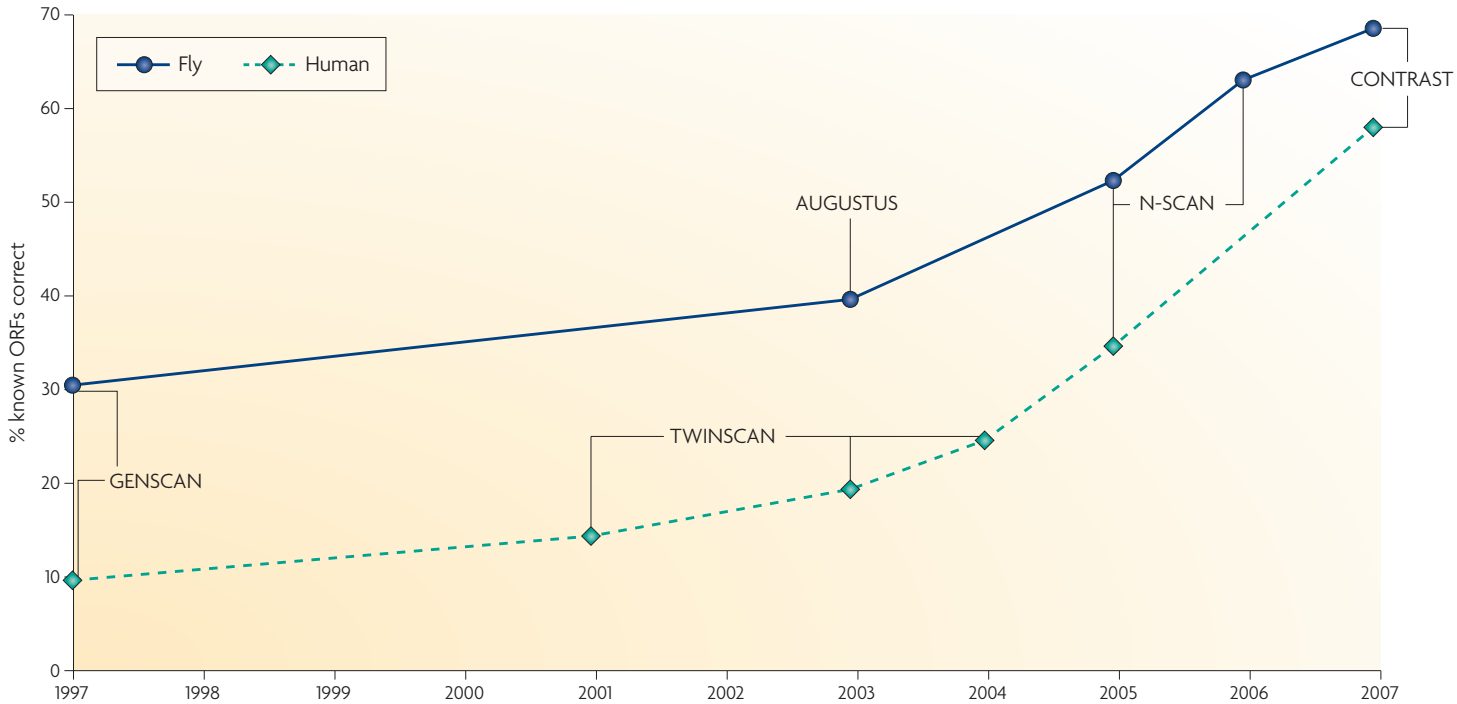
⇒ Predict mRNA & gene products

Why Yet Another Gene Finder?

- GENSCAN, Burge 1997
- Twinscan, Korf 2001
- Augustus, Stanke 2003
- Contrast, Gross 2007
- ...

Why Yet Another Gene Finder?

- GENSCAN, Burge 1997
- Twinscan, Korf 2001
- Augustus, Stanke 2003
- Contrast, Gross 2007
- ...

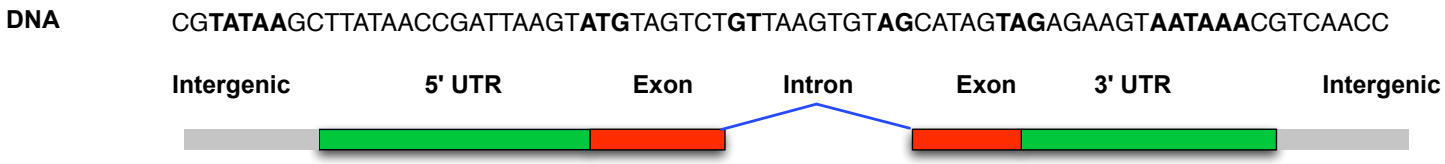


Brent, 2008

What's the Difference?

Traditionally: Generative models (HMM)

- Learn complete generative model



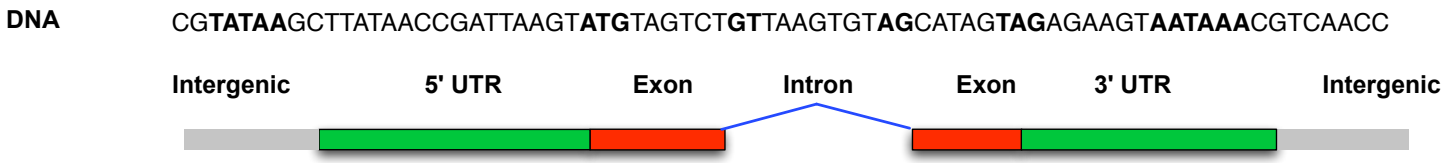
New approach: Discriminative setting

- Vapnik 1998: Never solve a more general problem as an intermediate step
- Solve the classification problem directly. much easier!

What's the Difference?

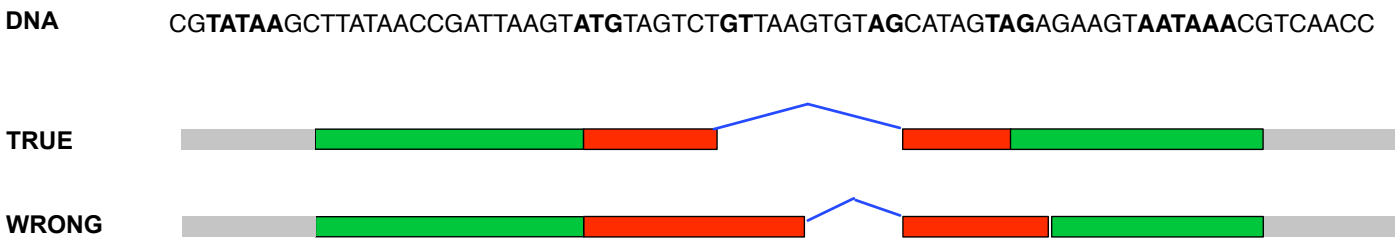
Traditionally: Generative models (HMM)

- Learn complete generative model



New approach: Discriminative setting

- Vapnik 1998: Never solve a more general problem as an intermediate step
- Solve the classification problem directly. much easier!



Our Approach: 2 Discriminative Steps

Step 1: State-of-the-art SVM Classifiers

Predict signals on DNA:

- Transcription start and cleavage, polyA, *trans-splice* sites
- Translation initiation sites and stop codons
- Donor and acceptor splice sites

Recognize segment types:

- Exons
- Introns
- Intergenic

Step 2: Hidden Markov SVMs

Combine Predictions to a valid gene structure

Our Approach: 2 Discriminative Steps

Step 1: State-of-the-art SVM Classifiers

Predict signals on DNA:

- Transcription start and cleavage, polyA, *trans-splice* sites
- Translation initiation sites and stop codons
- Donor and acceptor splice sites

Recognize segment types:

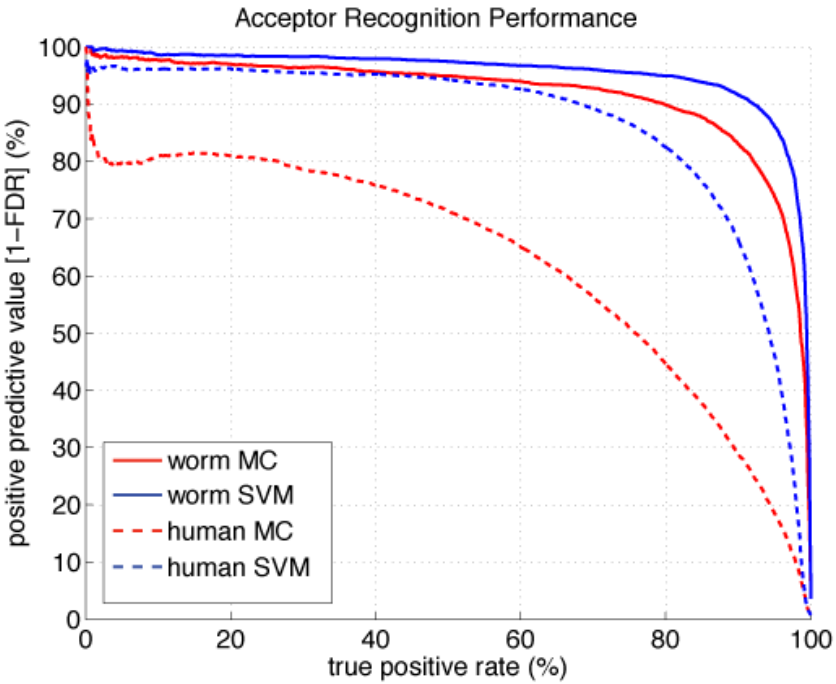
- Exons
- Introns
- Intergenic

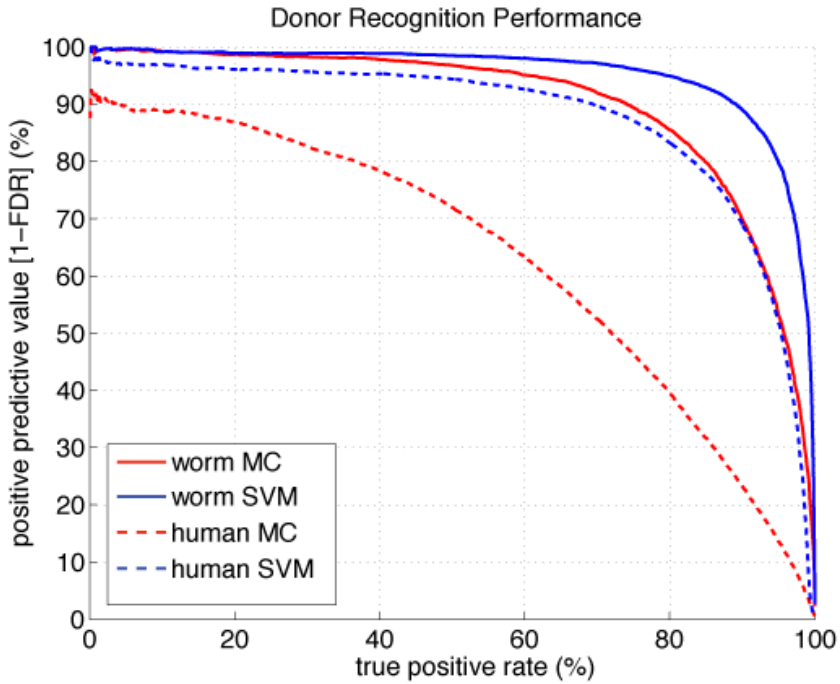
Step 2: Hidden Markov SVMs

Combine Predictions to a valid gene structure

Step 1: Splice Site Recognition

	Worm		Human	
	Acc	**Don**	**Acc**	**Don**
Markov Chain auPRC(%)	92.1	90.0	16.2	26.0
SVM auPRC(%)	95.9	95.3	54.4	56.9

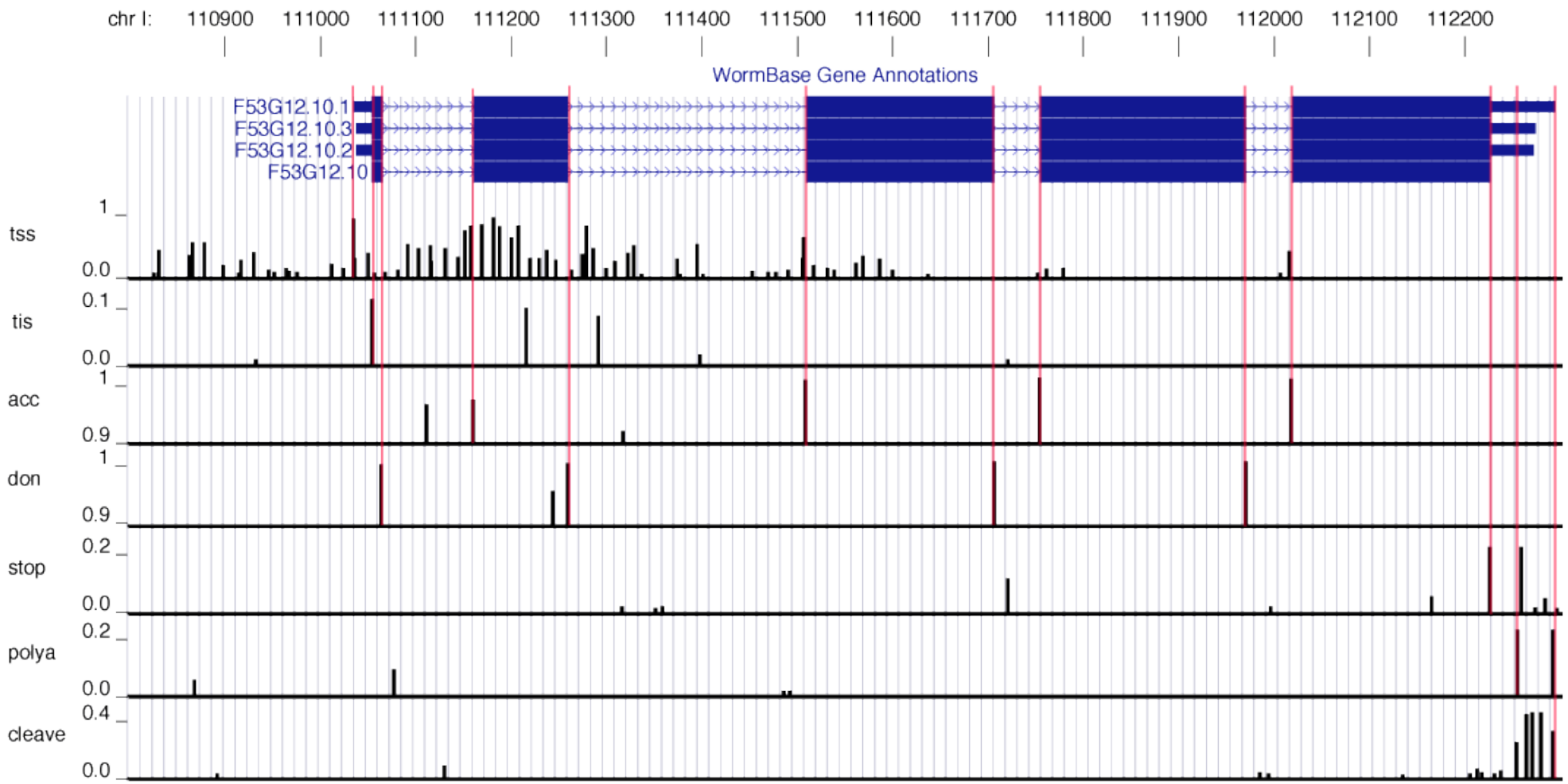




[Sonnenburg, Schweikert, Philips, Behr, Rätsch, 2007]

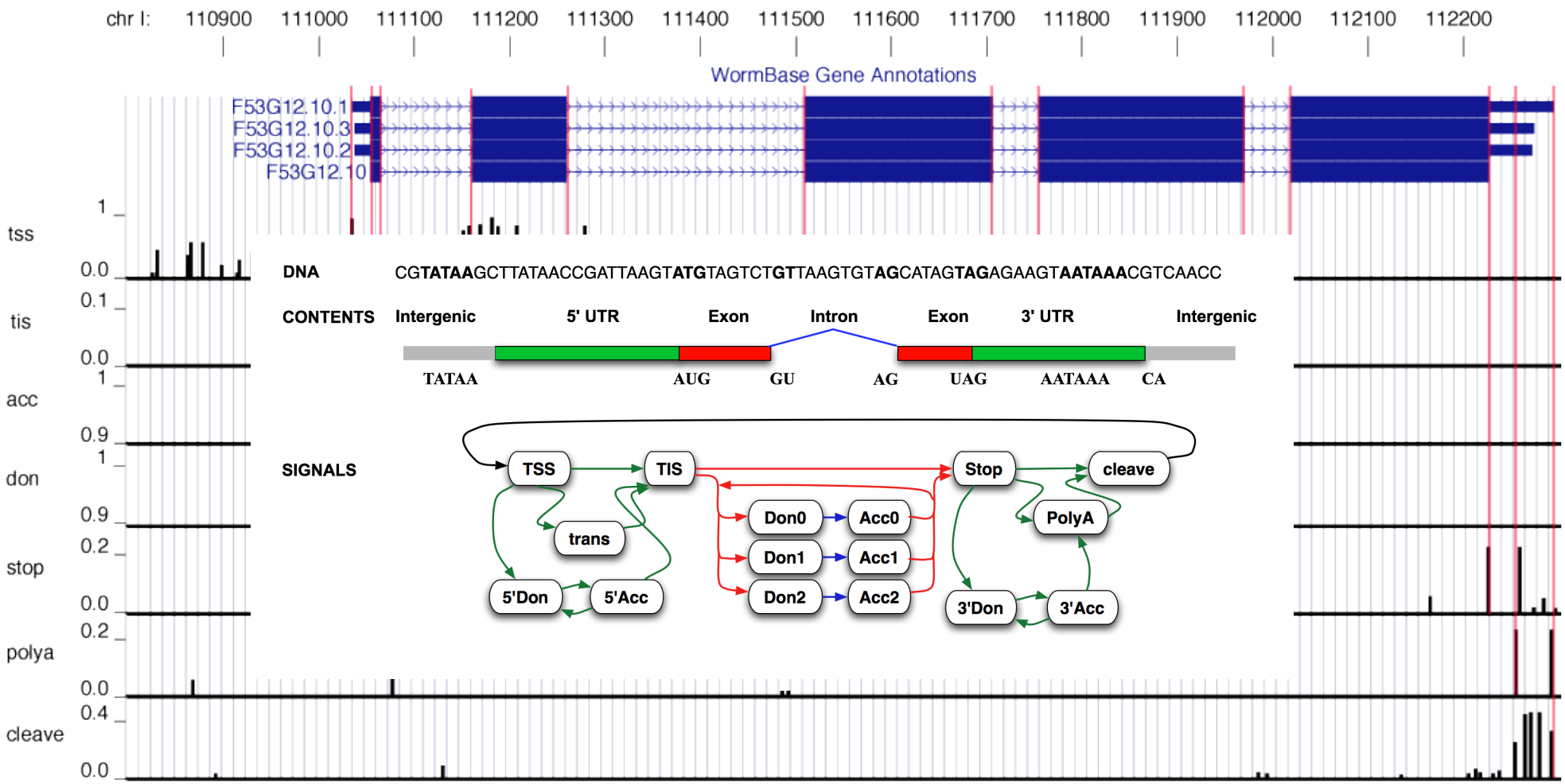
Step 1: Predictions in UCSC Browser

Friedrich Miescher Laboratory
of the Max Planck Society

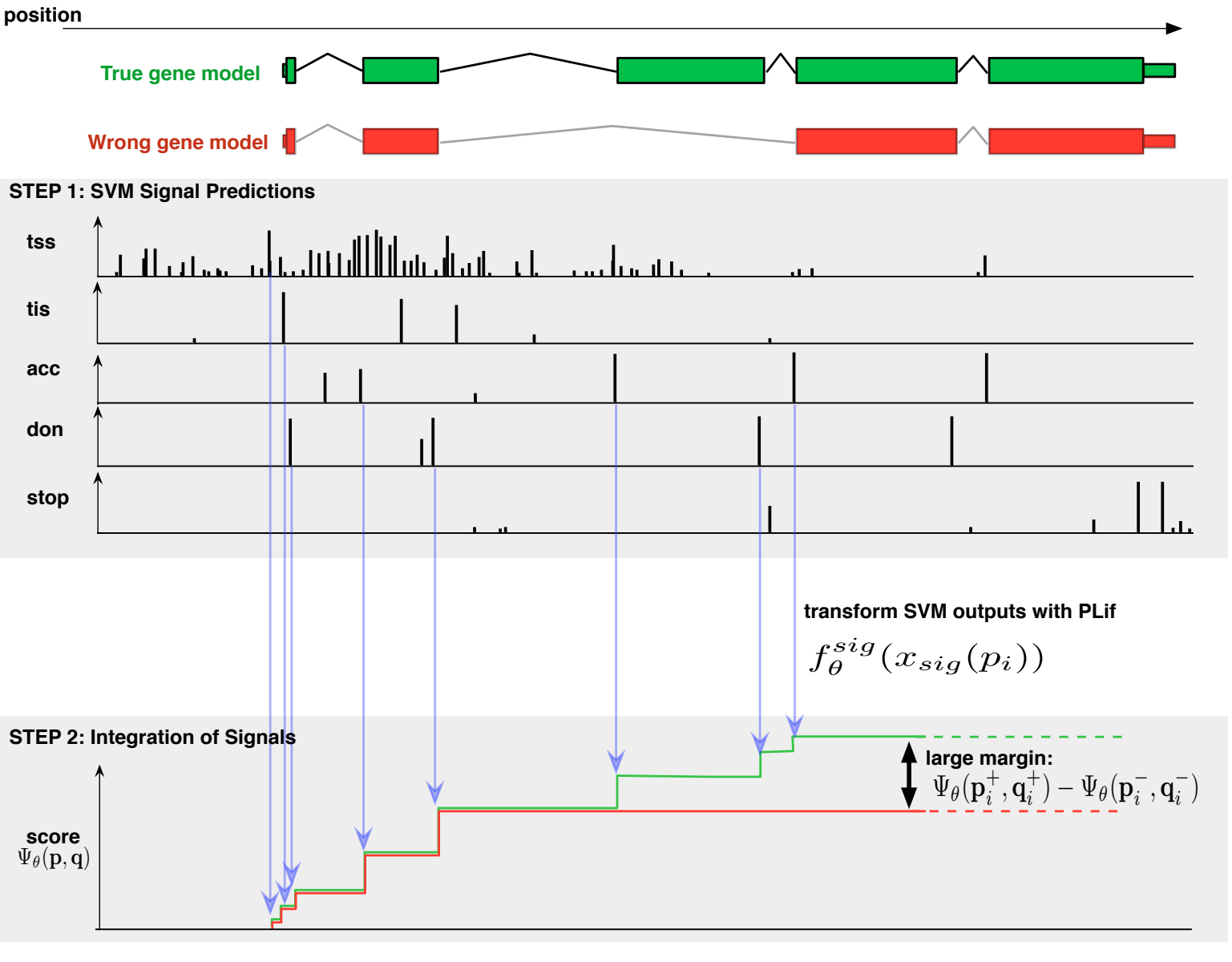


Step 1: Predictions in UCSC Browser

Friedrich Miescher Laboratory
of the Max Planck Society



Step 2: Signal Integration



Results: nGASP Competition

Find the most accurate gene finder for annotation of new nematode genomes:

- Controlled competition conditions:
  - 10% of the genome for training methods
  - 10% of the genome for evaluation
- 4 Categories:
  - Cat 1:*Ab initio* gene finders
  - Cat 2: Dual/Multi-genome gene finders
  - Cat 3: Gene finders that use EST/cDNA alignments
  - Cat 4: Combining algorithms
- Evaluation on WS160 genes

Results: nGASP Competition

Find the most accurate gene finder for annotation of new nematode genomes:

- Controlled competition conditions:
  - 10% of the genome for training methods
  - 10% of the genome for evaluation
- 4 Categories:
  - Cat 1:*Ab initio* gene finders
  - Cat 2: Dual/Multi-genome gene finders
  - Cat 3: Gene finders that use EST/cDNA alignments
  - Cat 4: Combining algorithms
- Evaluation on WS160 genes

Results: nGASP Competition

Find the most accurate gene finder for annotation of new nematode genomes:

- Controlled competition conditions:
  - 10% of the genome for training methods
  - 10% of the genome for evaluation
- 4 Categories:
  - Cat 1:*Ab initio* gene finders
  - Cat 2: Dual/Multi-genome gene finders
  - Cat 3: Gene finders that use EST/cDNA alignments
  - Cat 4: Combining algorithms
- Evaluation on WS160 genes

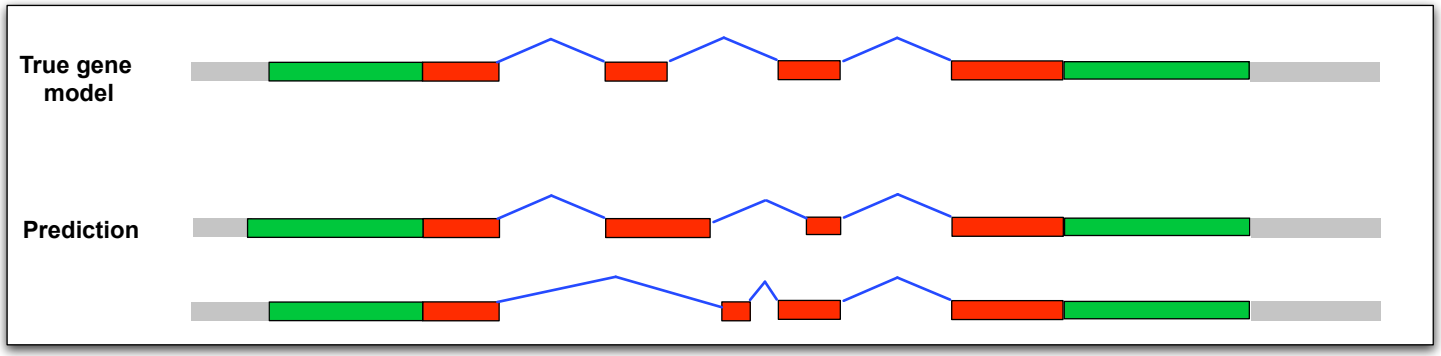
Results: nGASP Competition

Find the most accurate gene finder for annotation of new nematode genomes:

- Controlled competition conditions:
  - 10% of the genome for training methods
  - 10% of the genome for evaluation
- 4 Categories:
  - Cat 1:*Ab initio* gene finders
  - Cat 2: Dual/Multi-genome gene finders
  - Cat 3: Gene finders that use EST/cDNA alignments
  - Cat 4: Combining algorithms
- Evaluation on WS160 genes

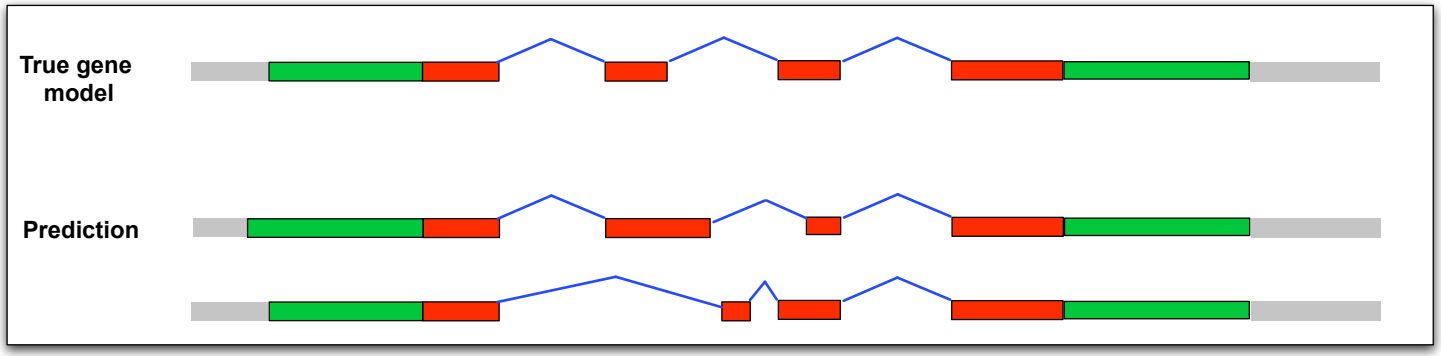
nGASP Evaluation

How to evaluated predictions?

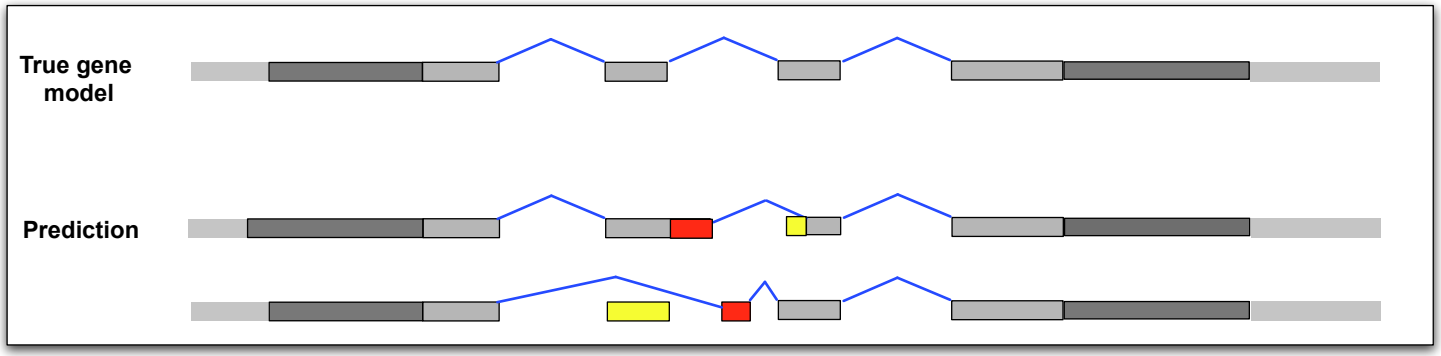


nGASP Evaluation

How to evaluated predictions?



Nucleotide evaluation:

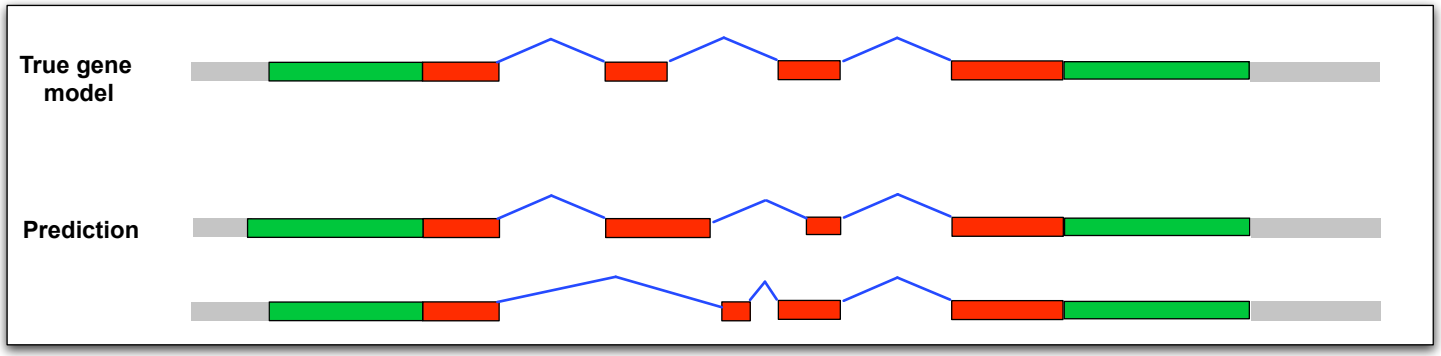


Sensitivity = True Predicted/True

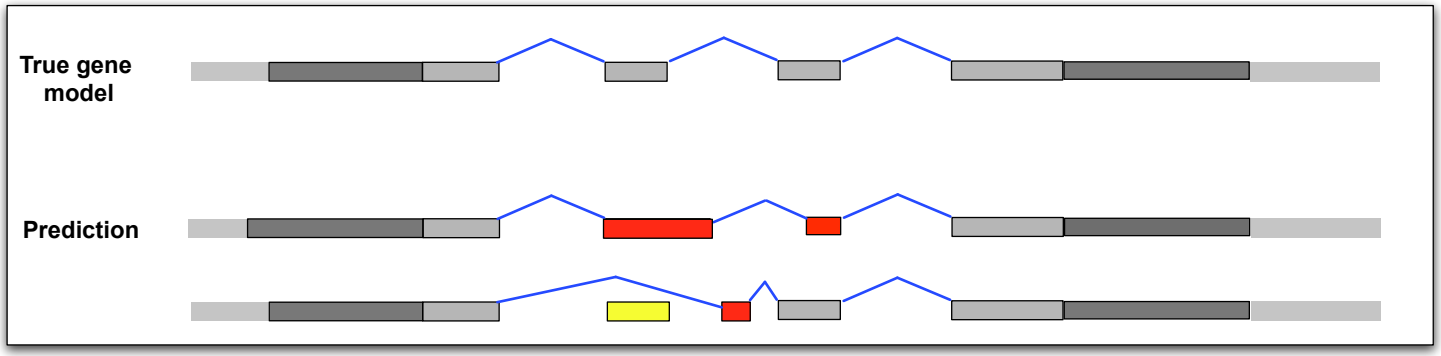
Specificity = True Predicted/Predicted

nGASP Evaluation

How to evaluated predictions?



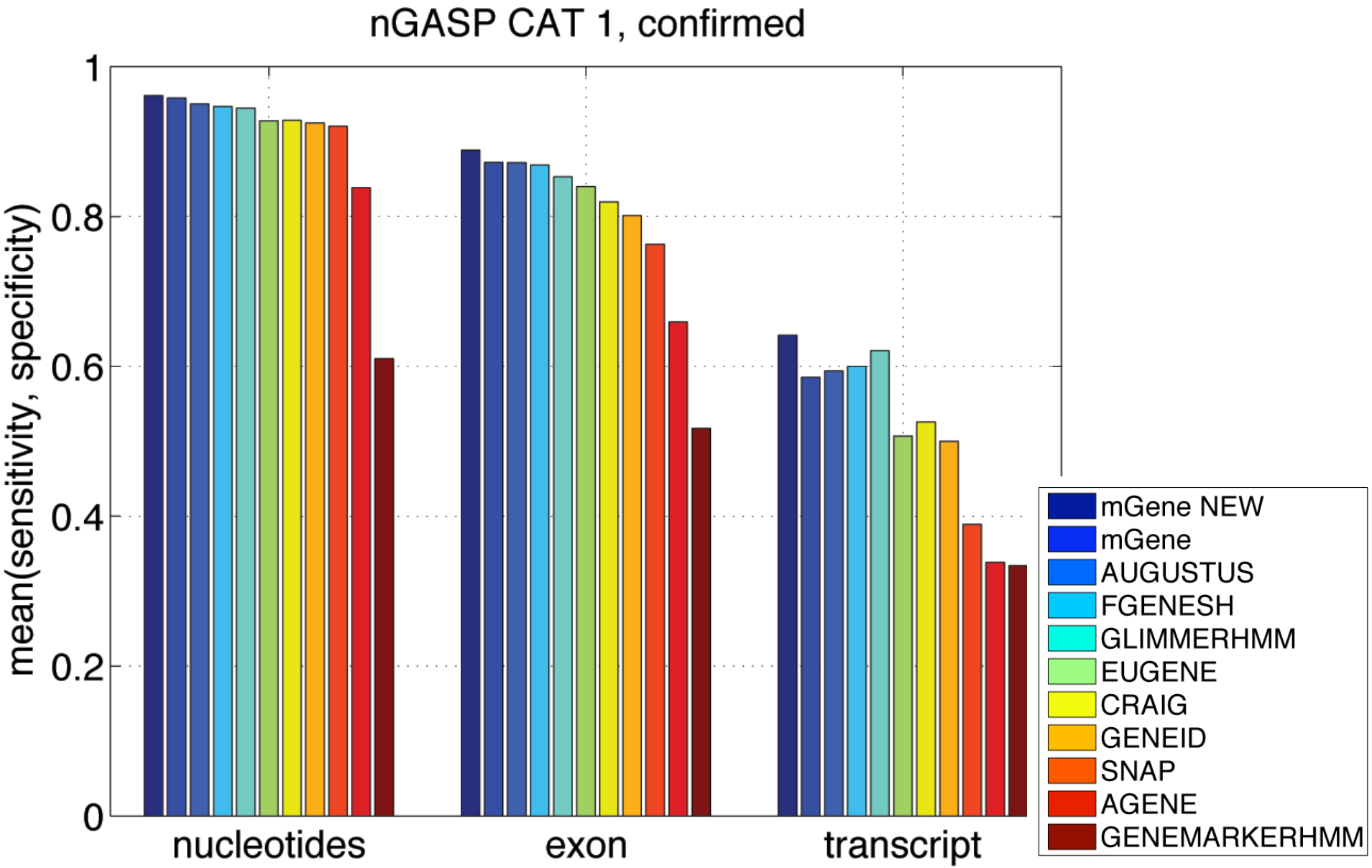
Exon evaluation:



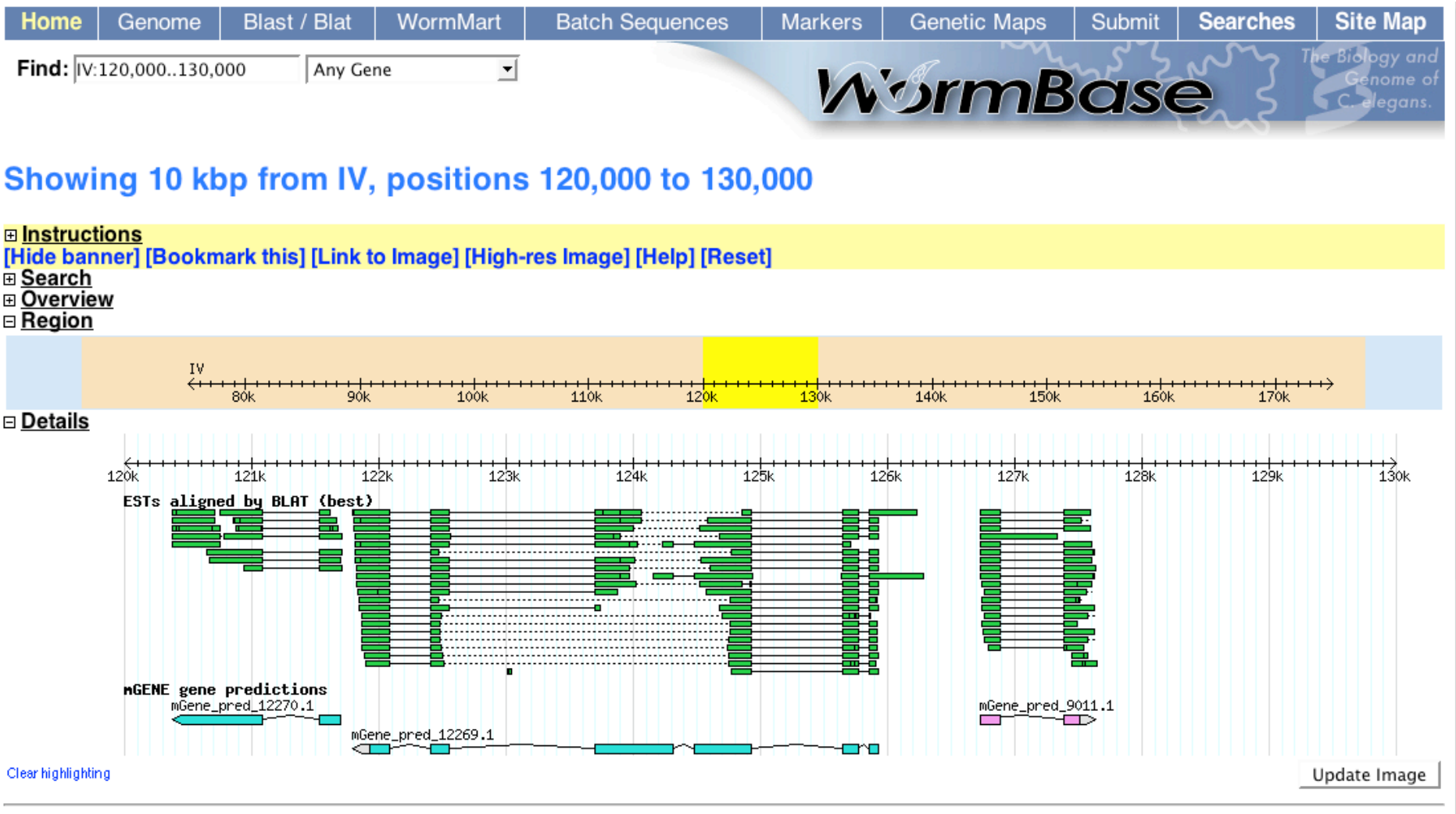
Sensitivity = True Predicted/True

Specificity = True Predicted/Predicted

nGASP Cat. 1 Evaluation (prelim.)



Results: *mGene* on Wormbase



Results

- Very good on nucleotide and exon level
- Substantial improvements on transcript level
- Re-annotation of *C. elegans* in official wormbase annotation
  - total: 19532 genes predicted
  - 635 *new* genes
  - 350 unconfirmed genes are not predicted
- Wetlab confirmations (preliminary results)
  - Confirmed 55 of 103 newly predicted genes
  - Confirmed only 4 of 50 annotated genes that were not predicted

Summary and Future Work

- High accuracy on constitutively spliced genes for nematodes
- Wet-lab experiments confirm good performance of mGene
- Re-annotation of other nematode genomes genomes
- Preliminary results on new information transfer methods improve performance on related organisms
- Integrate new experimental data
  - New sequencing technologies
  - Tiling arrays
- Computational challenge: predict alternative splicing

Acknowledgments

- Gunnar Rätsch, Bernhard Schölkopf, Detlef Weigel
- Georg Zeller, Alexander Zien, Cheng Soon Ong, Sören Sonnenburg, Petra Philips, Jonas Behr, Christian Widmer