De novo RNA-seq-based Genome Annotation

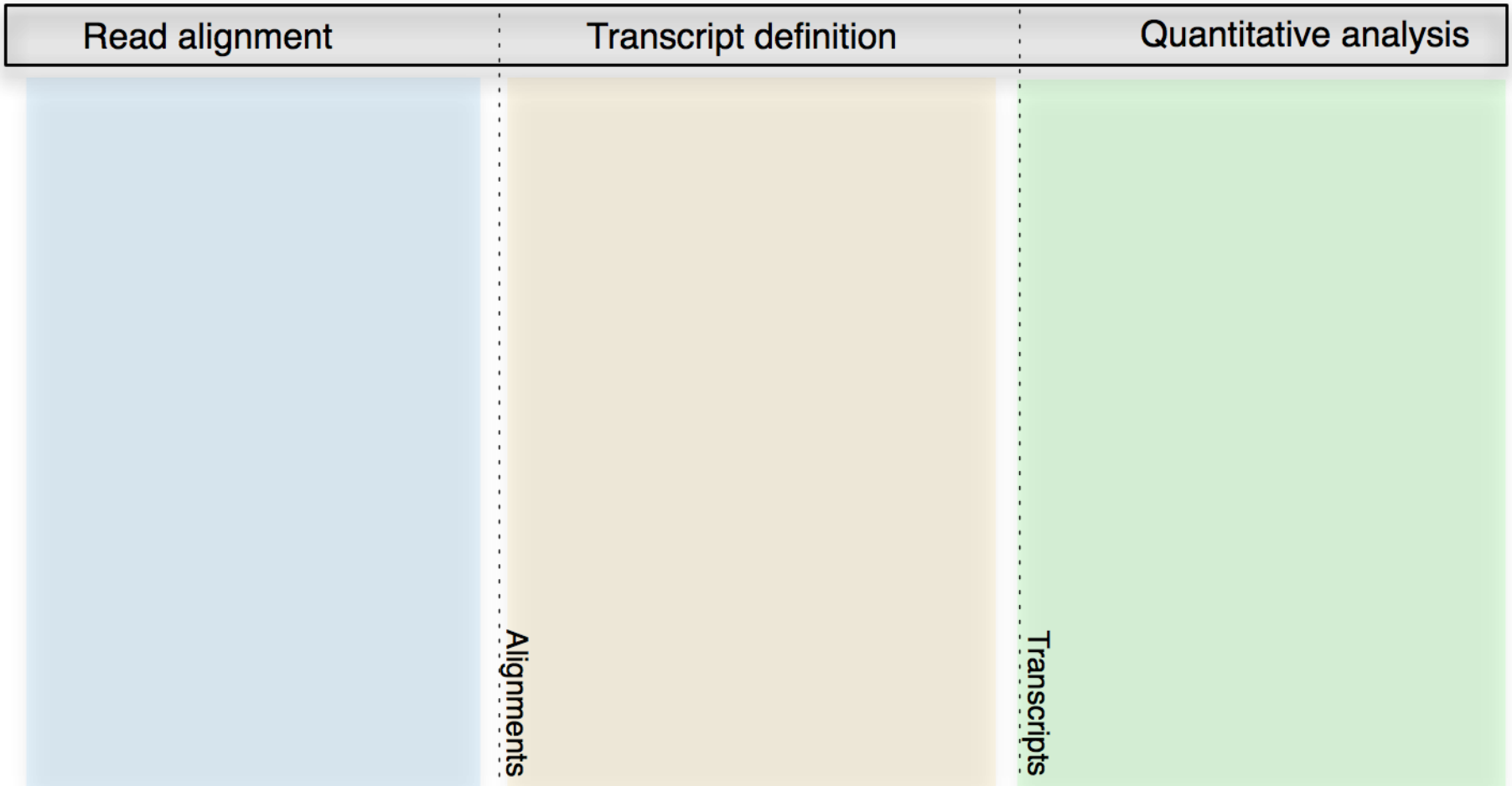
Gunnar Rätsch

Friedrich Miescher Laboratory of the Max Planck Society
Tübingen, Germany



Genome Informatics Meeting, September 19, 2010

RNA-seq Analysis Pipeline(s)

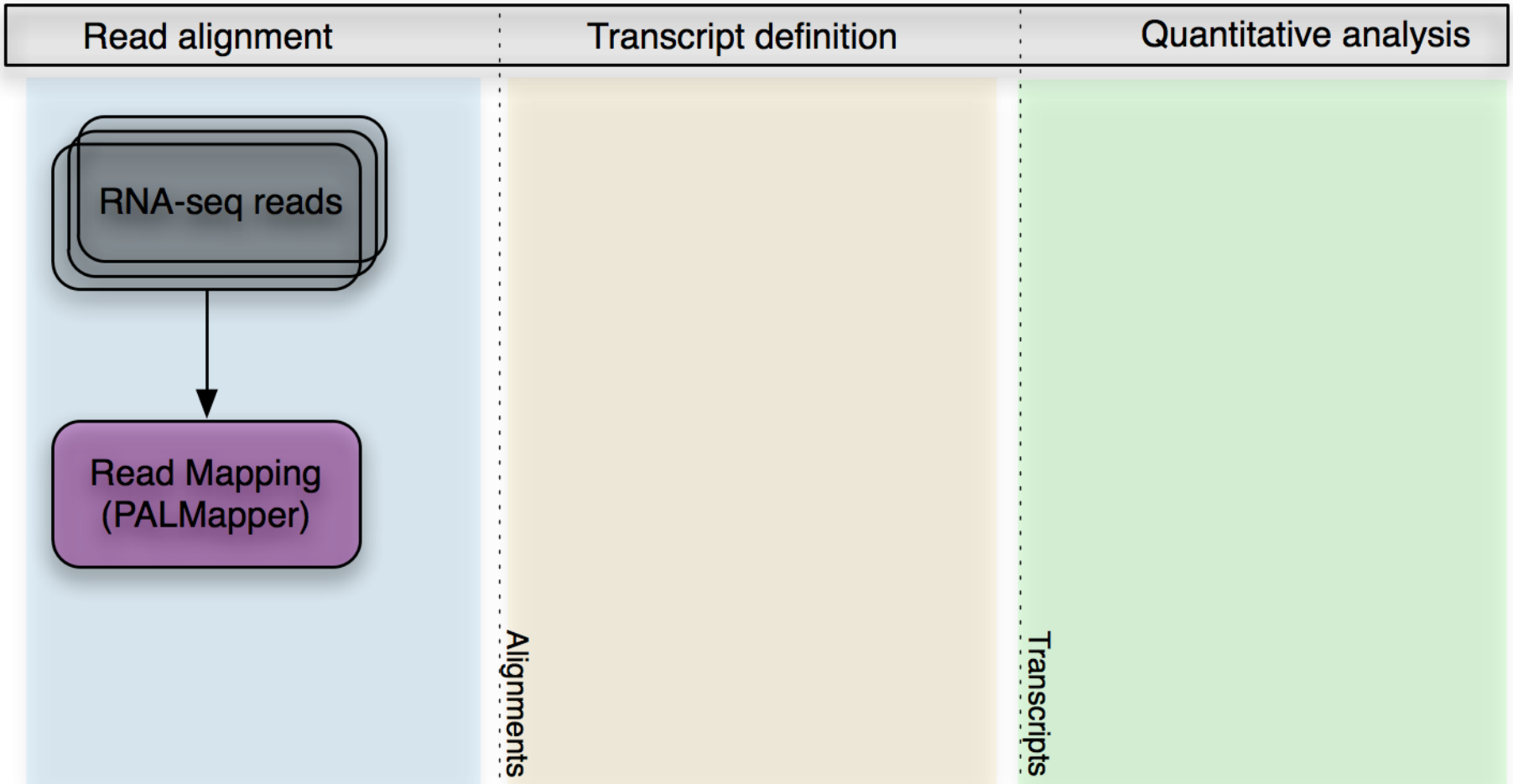


ShortSIG 2008: QPalma (now Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff (see Abstract 128) [Stegle et al., 2010]

RNA-seq Analysis Pipeline(s)

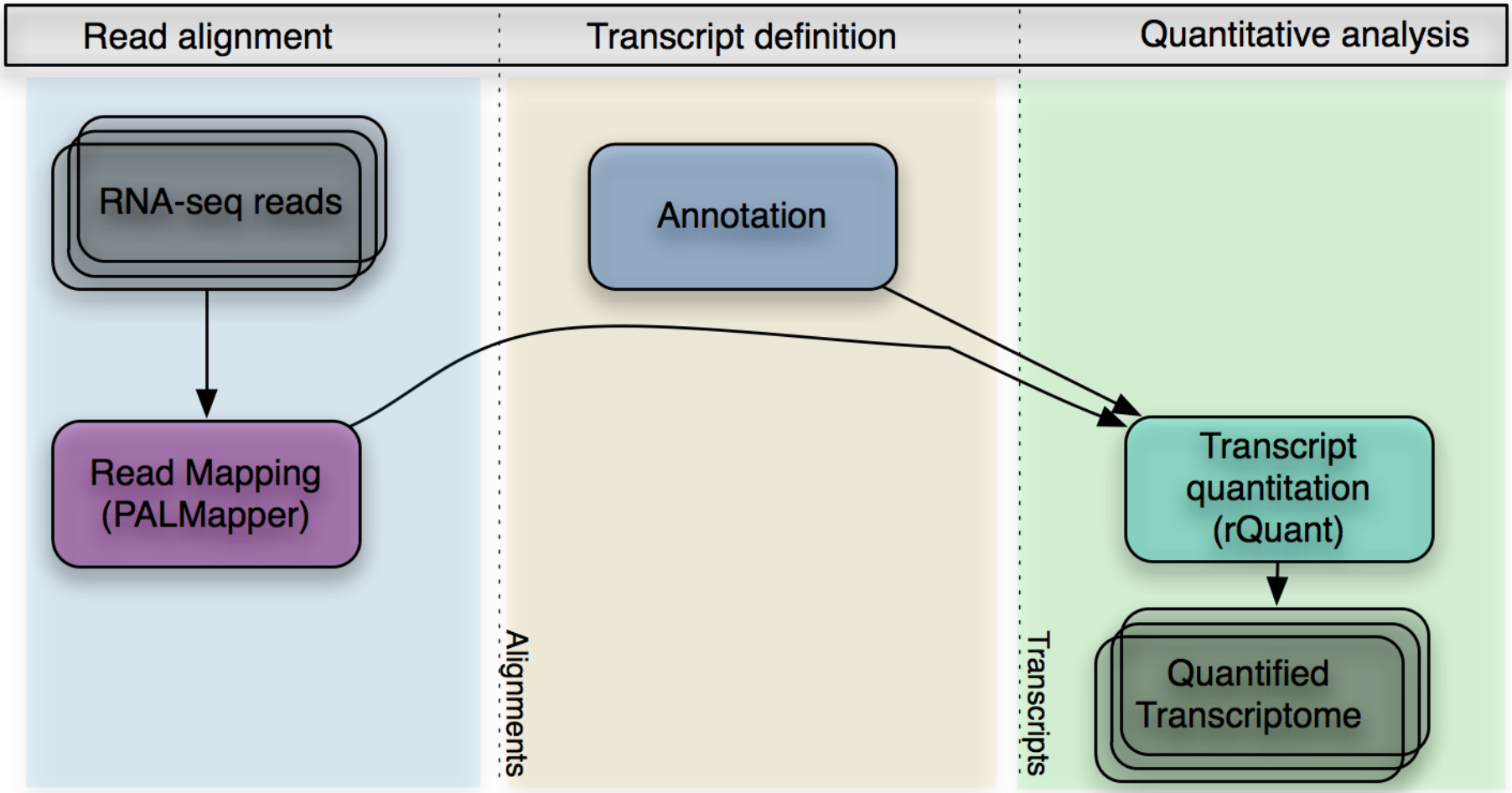


ShortSIG 2008: QPalma (now Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff (see Abstract 128) [Stegle et al., 2010]

RNA-seq Analysis Pipeline(s)

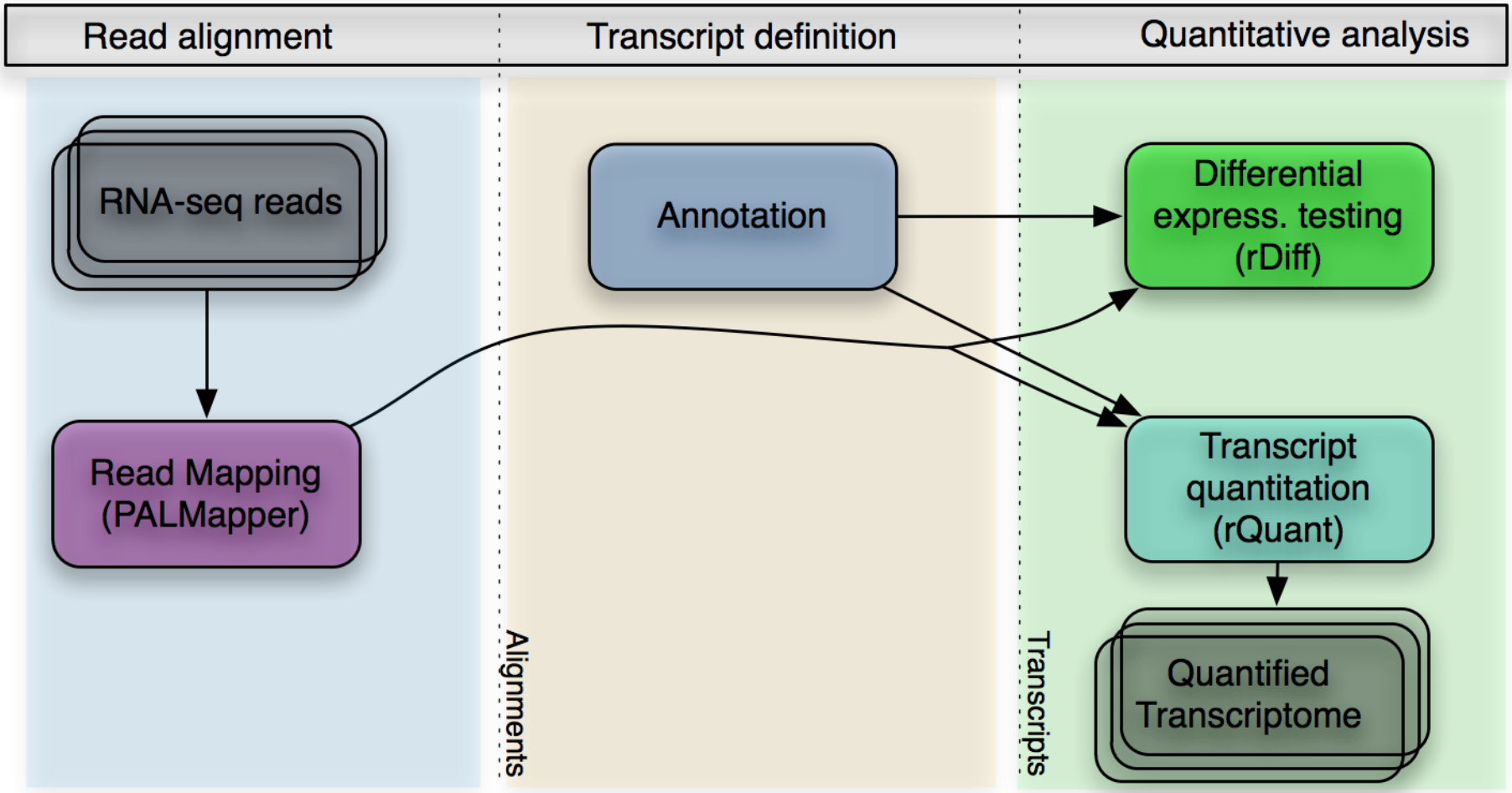


ShortSIG 2008: QPalma (now Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff (see Abstract 128) [Stegle et al., 2010]

RNA-seq Analysis Pipeline(s)

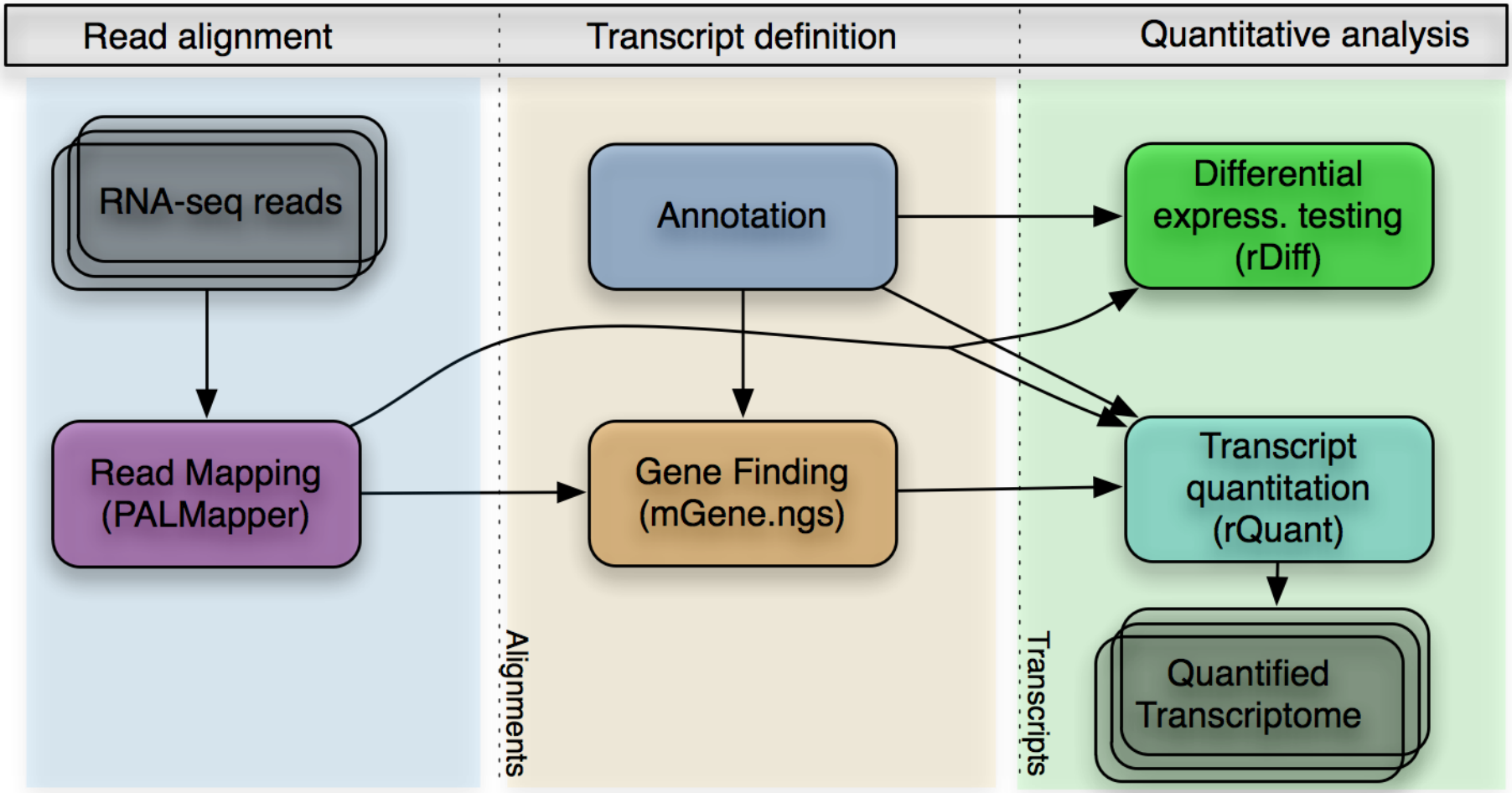


ShortSIG 2008: QPalma (now Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff (see Abstract 128) [Stegle et al., 2010]

RNA-seq Analysis Pipeline(s)

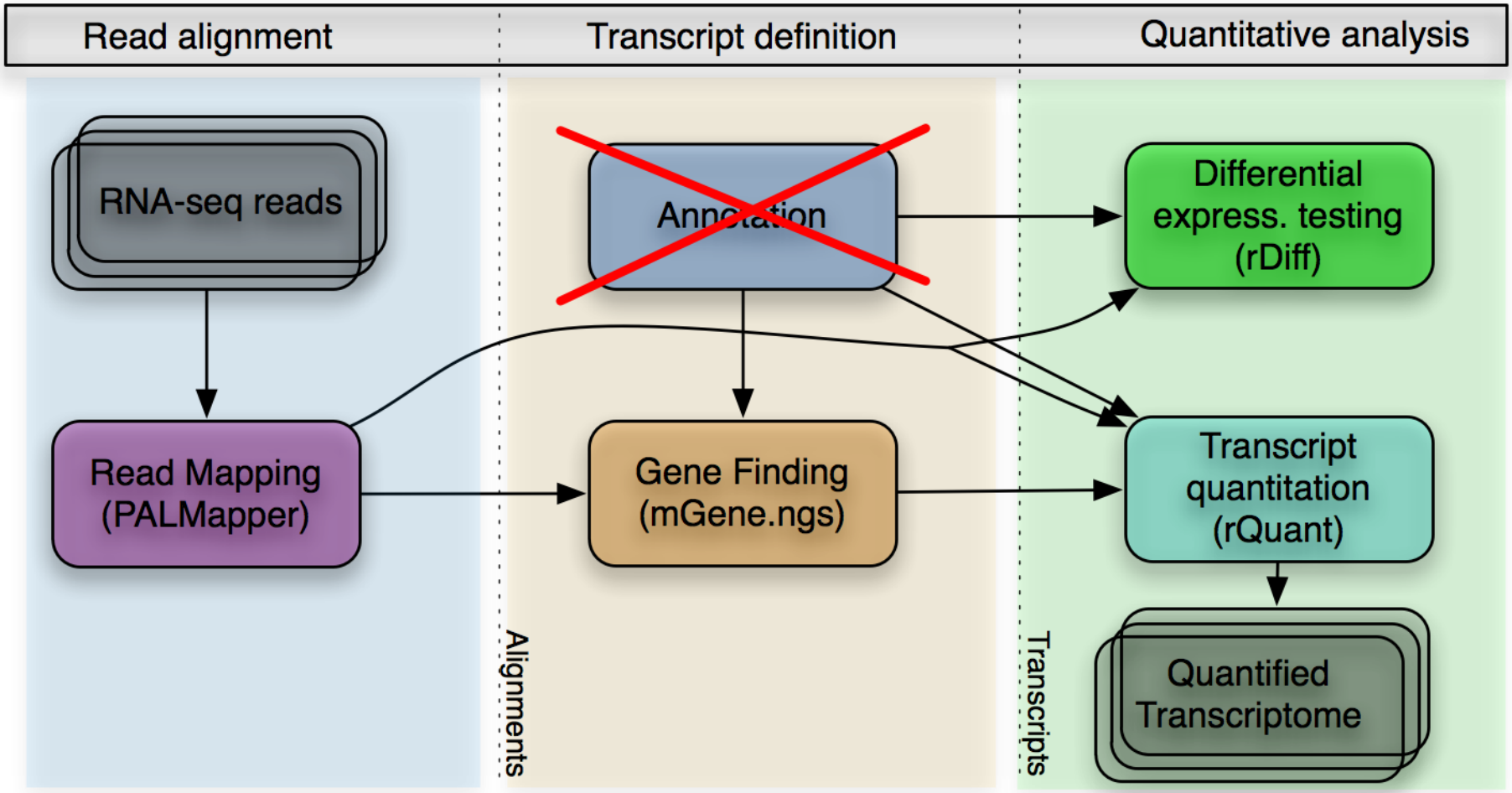


ShortSIG 2008: QPalma (now Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff (see Abstract 128) [Stegle et al., 2010]

RNA-seq Analysis Pipeline(s)

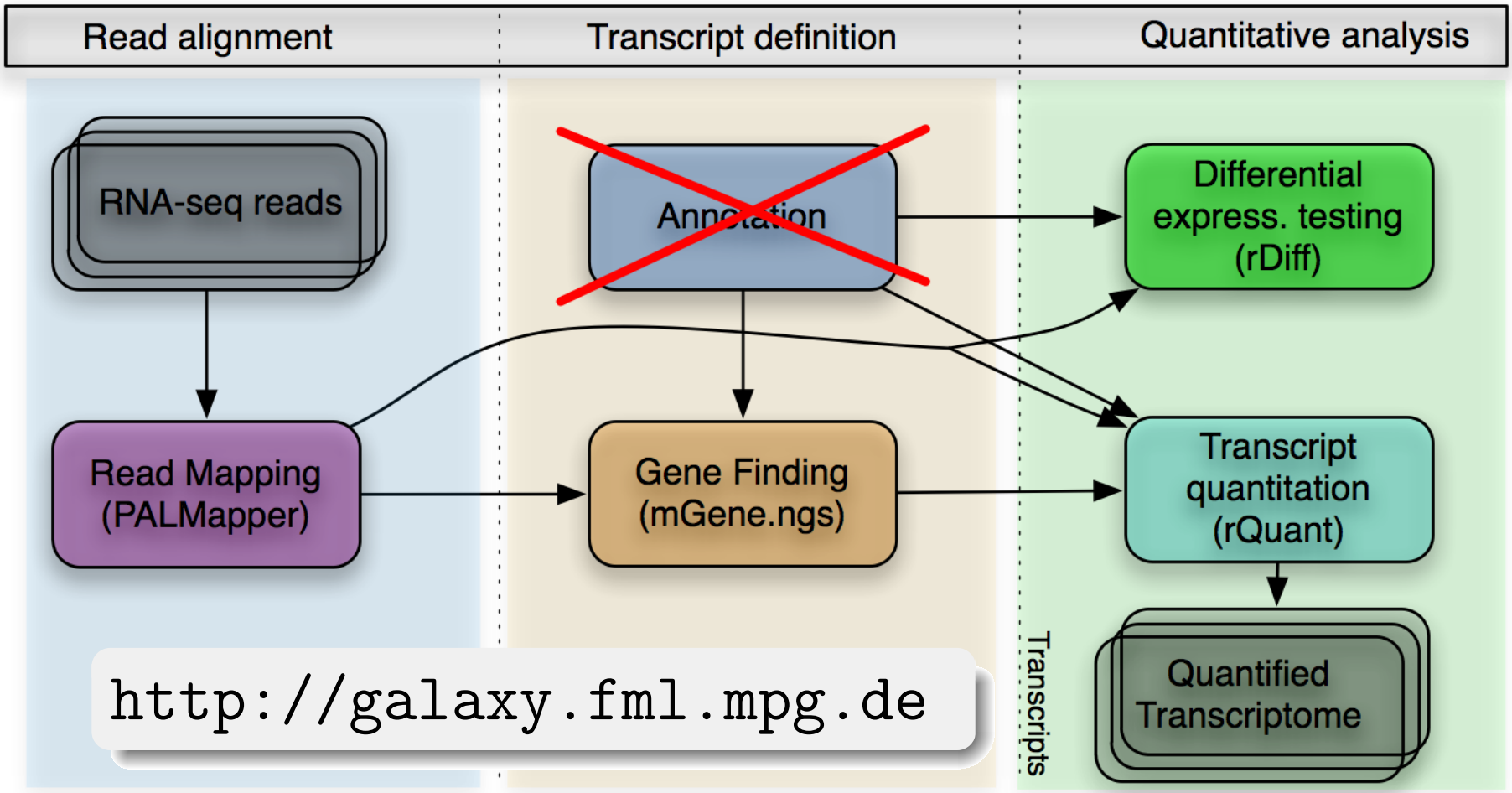


ShortSIG 2008: QPalma (now Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff (see Abstract 128) [Stegle et al., 2010]

RNA-seq Analysis Pipeline(s)



ShortSIG 2008: QPalma (now Palmapper) [De Bona et al., 2008, Jean et al., 2010]

ShortSIG 2009: rQuant [Bohnert and Rätsch, 2010]

Hitseq 2010: rDiff (see Abstract 128) [Stegle et al., 2010]

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{"large margin"})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{"large margin"})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{"large margin"})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

mGene.ngs Overview

- Goal: Integrate RNA-seq and genomic sequence information
- Learn function $f(y|x)$ that scores gene models y based on different sources of information x
  - Choose parameters of scoring function such that

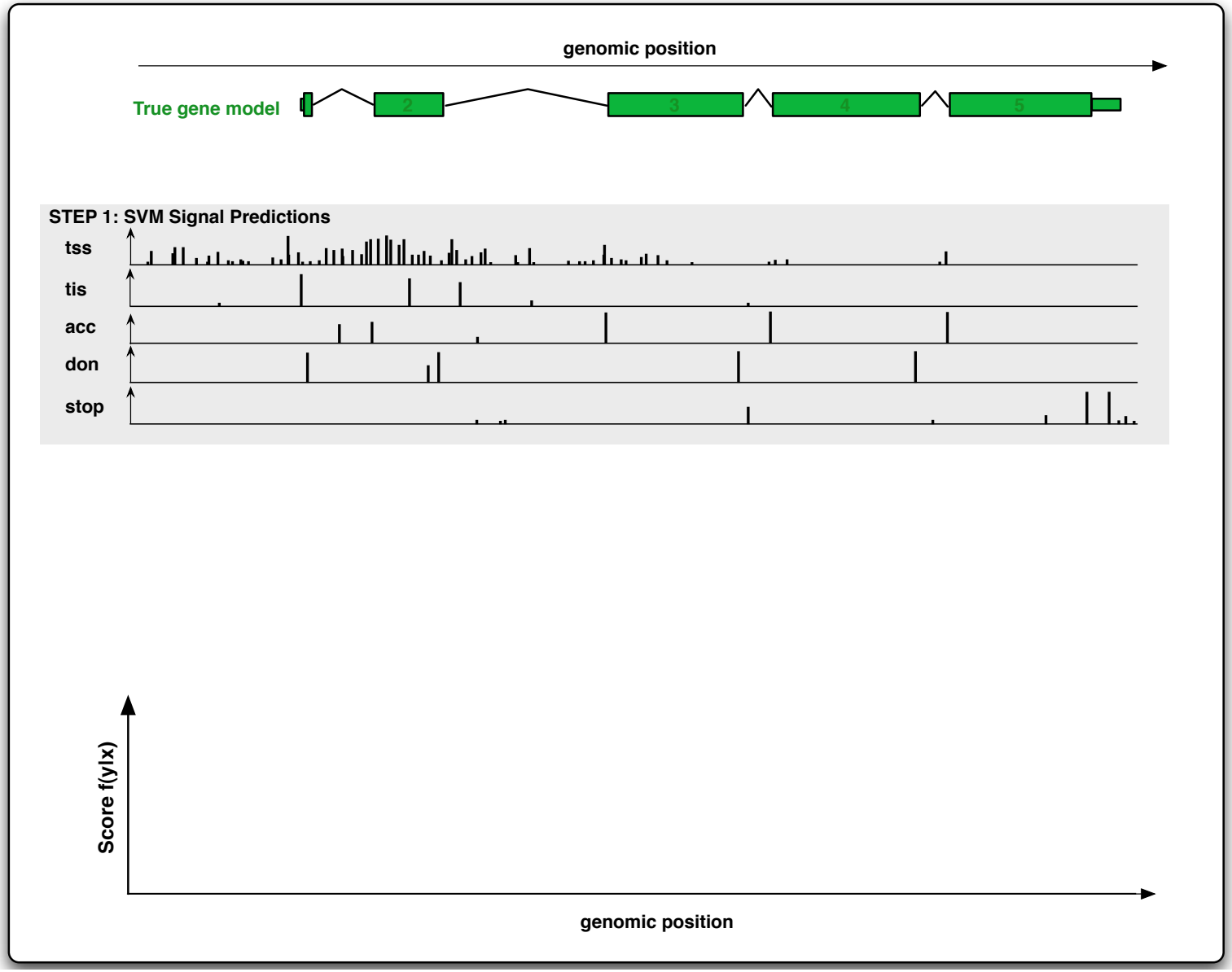
$$f(y|x) \gg f(y'|x) \text{ for all } y' \neq y \qquad (\text{“large margin”})$$

  - Hidden semi-Markov Support Vector Machines (HsM-SVMs)

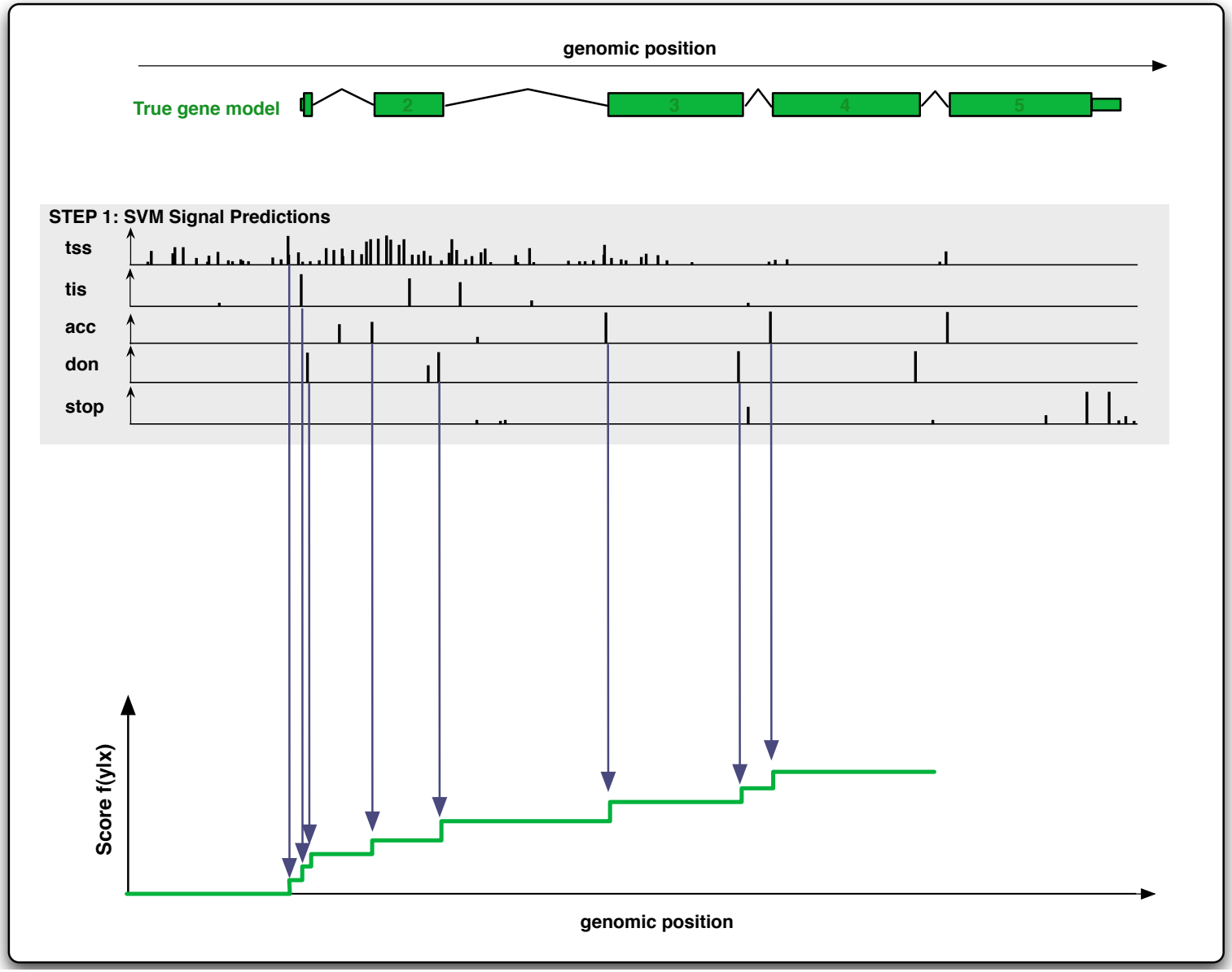
[Altun et al., 2003, Rätsch and Sonnenburg, 2007]

- Adapts to quality of RNA-seq alignments
- Alternative transcripts are predicted by *parsimonious quantitation* of paths through splice graph

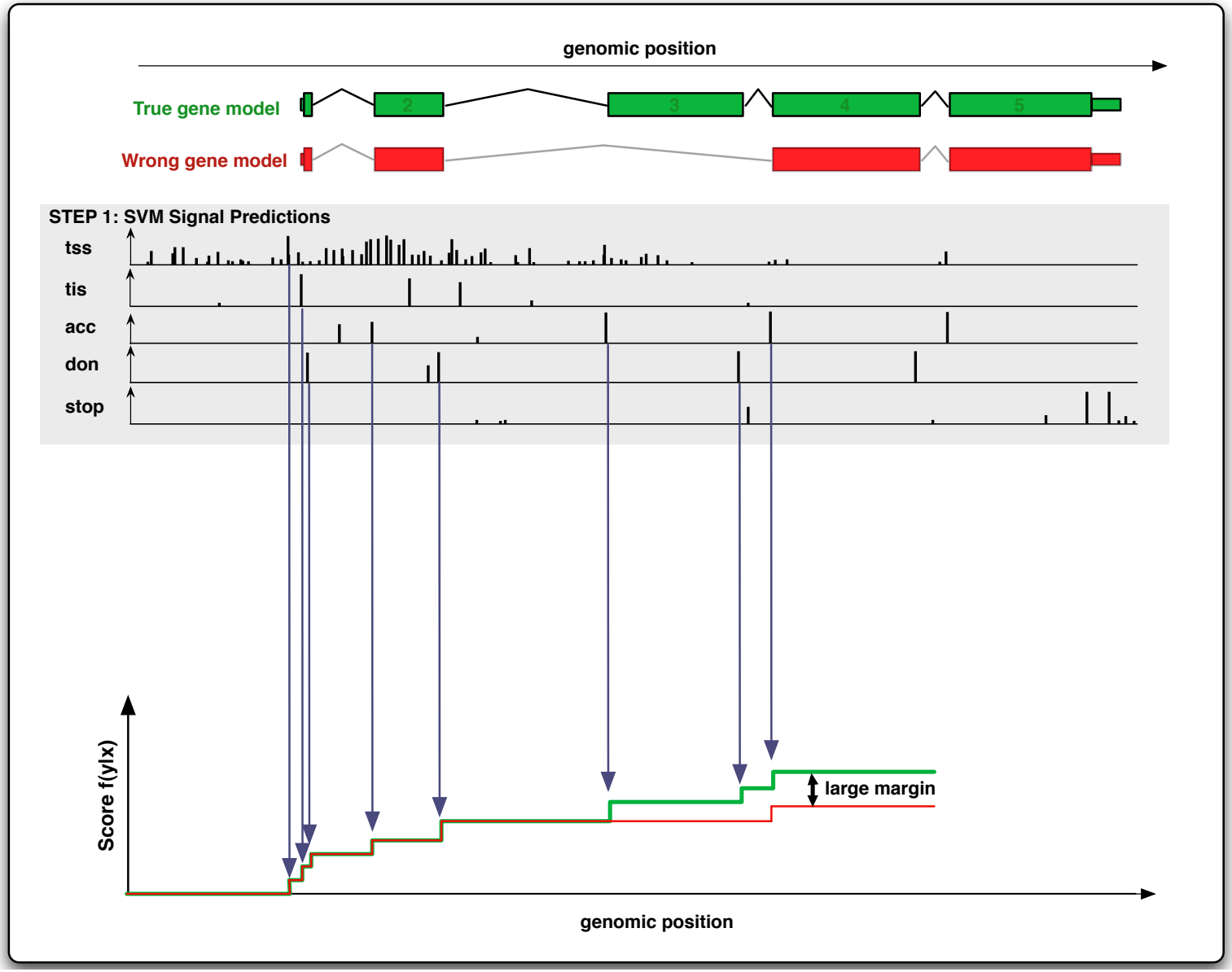
mGene.ngs Parameter Estimation



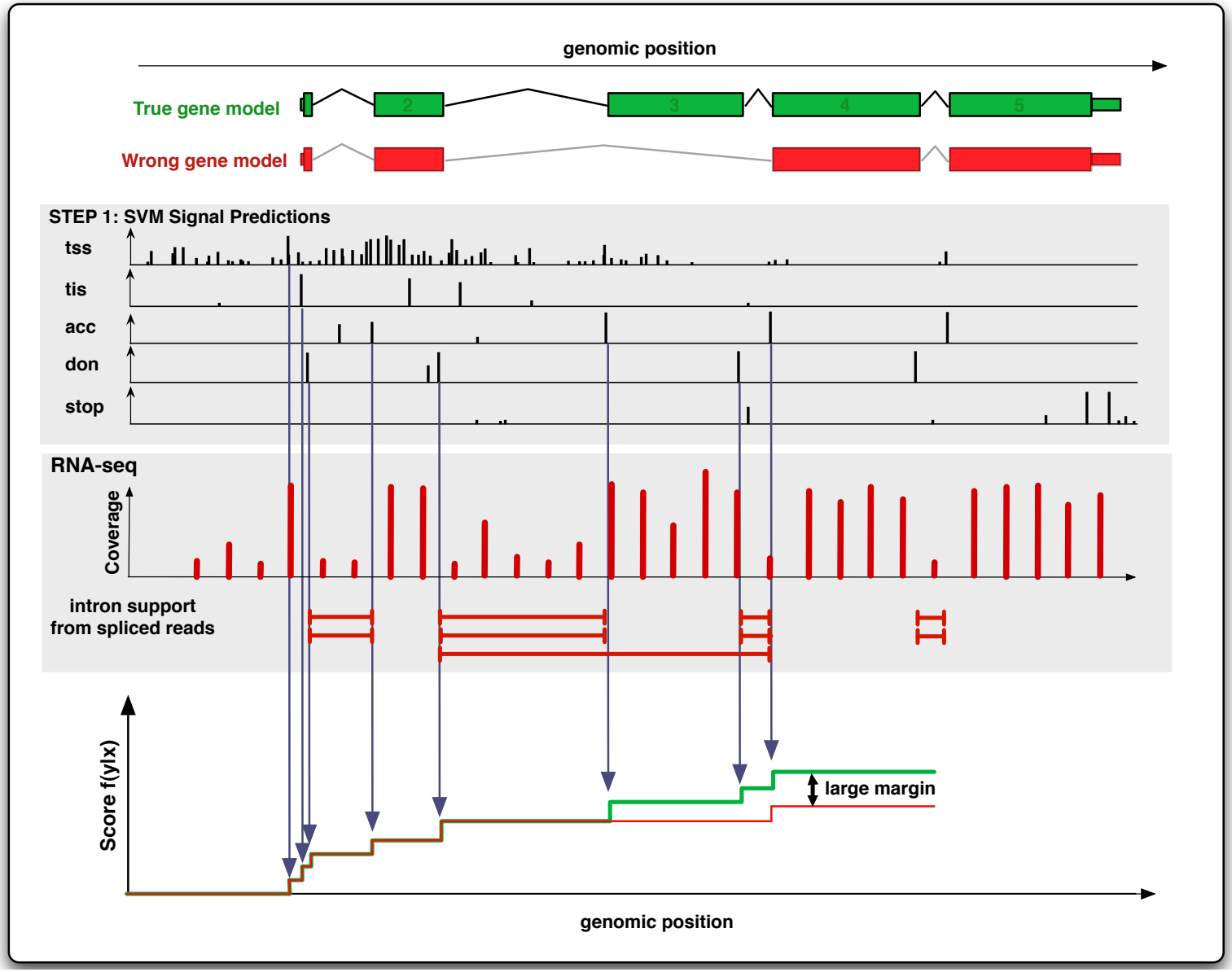
mGene.ngs Parameter Estimation



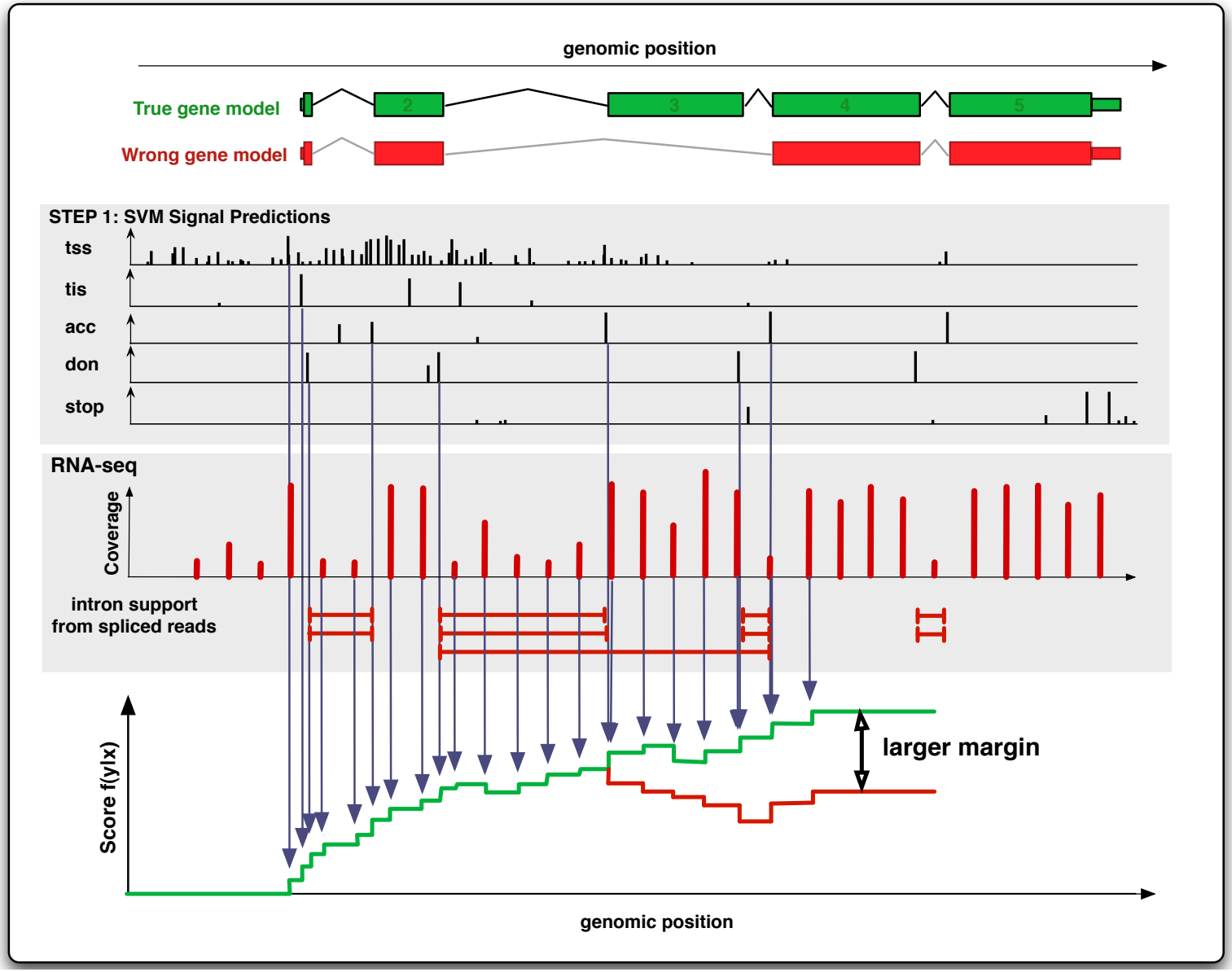
mGene.ngs Parameter Estimation



mGene.ngs Parameter Estimation

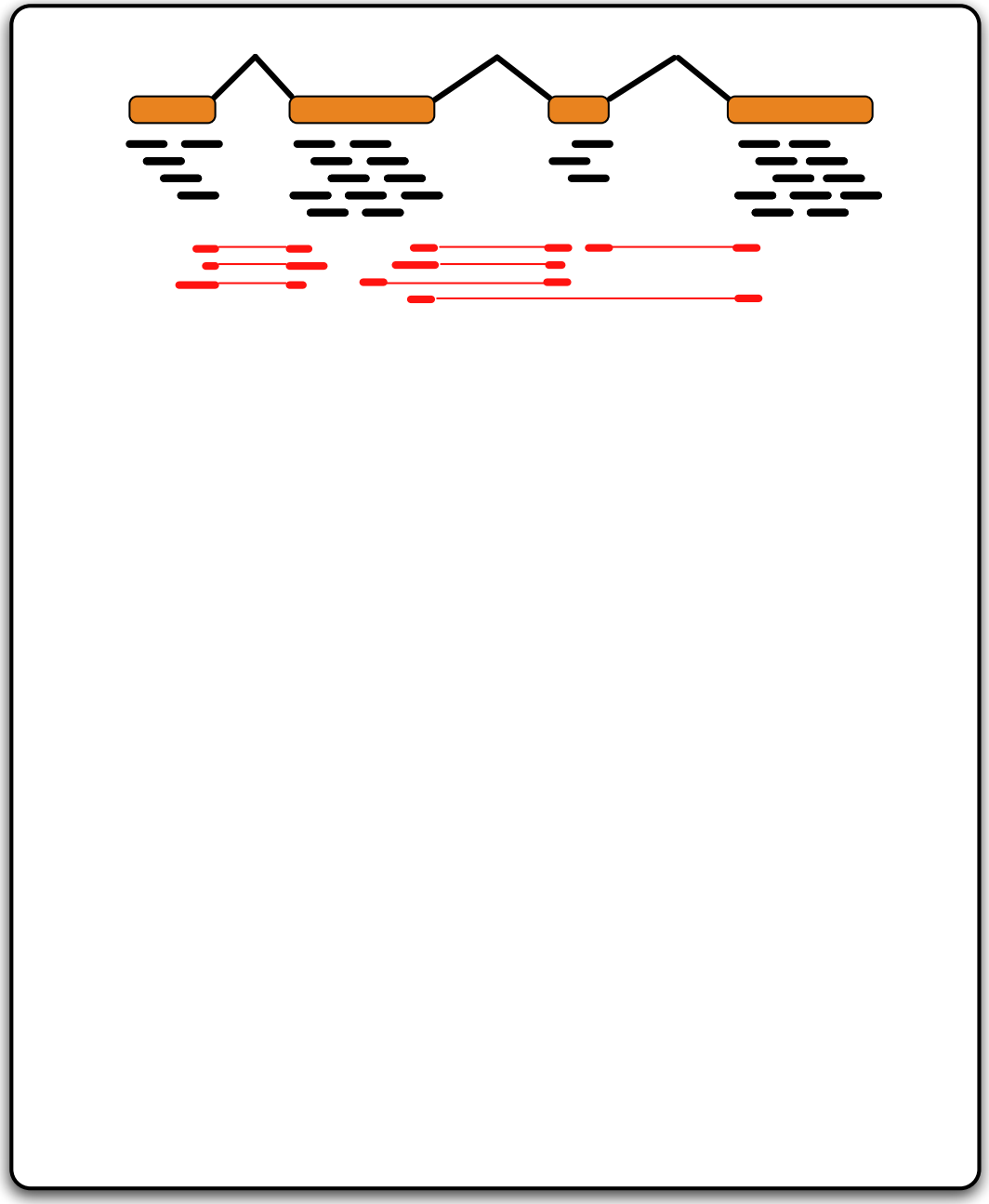


mGene.ngs Parameter Estimation



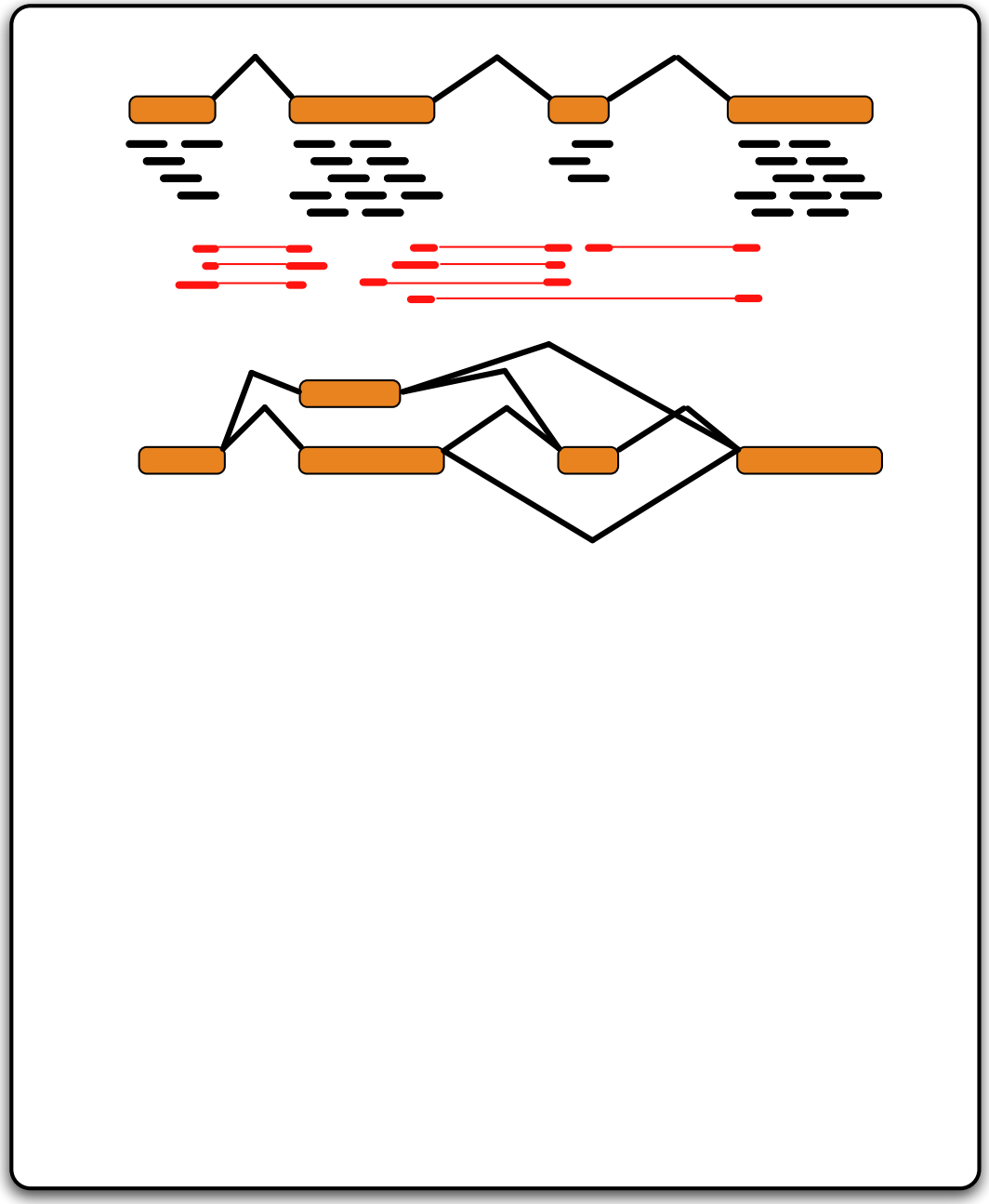
Infer alternative isoforms

- mGene.ngs prediction
- Build splicegraph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



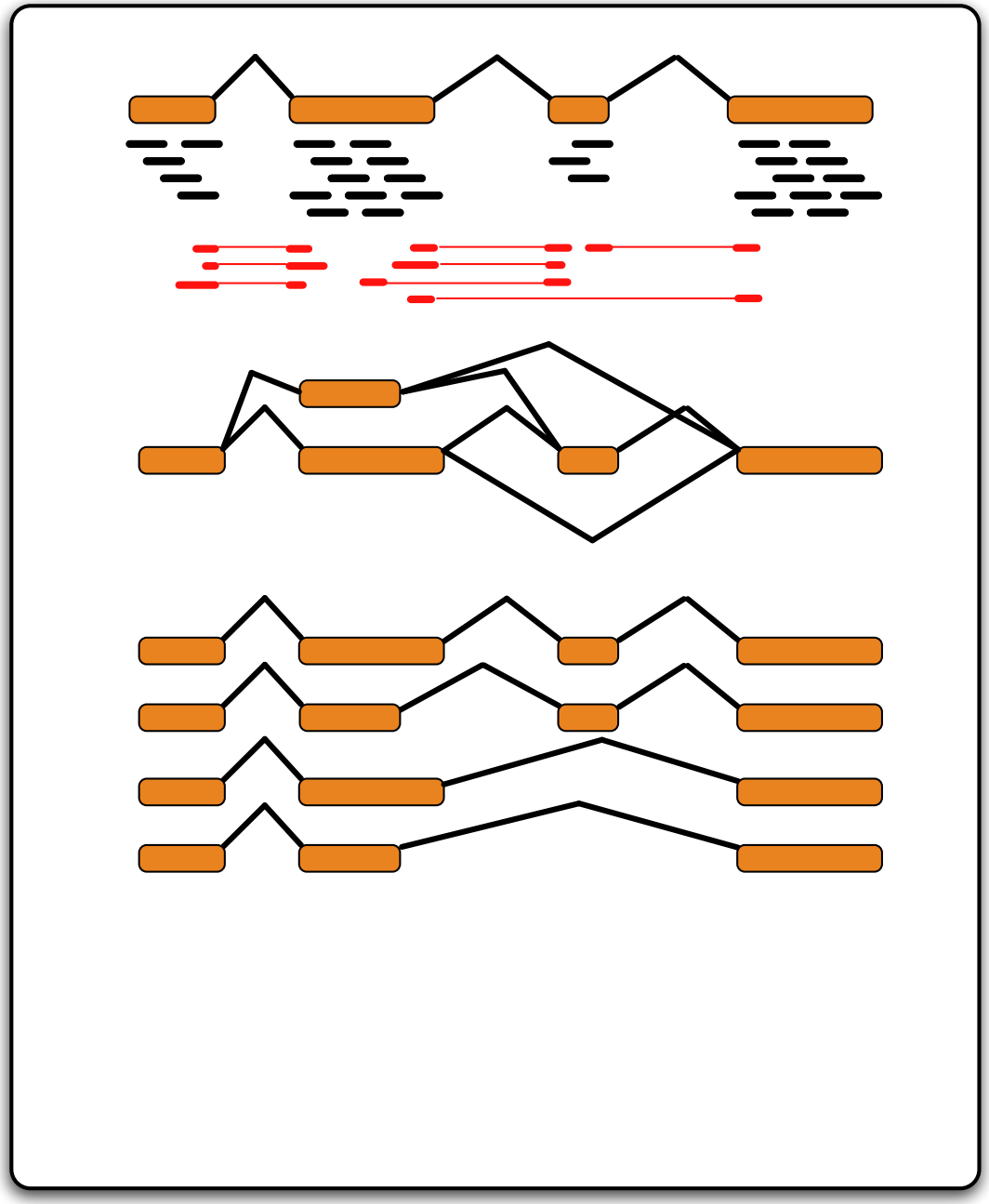
Infer alternative isoforms

- mGene.ngs prediction
- Build splicegraph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



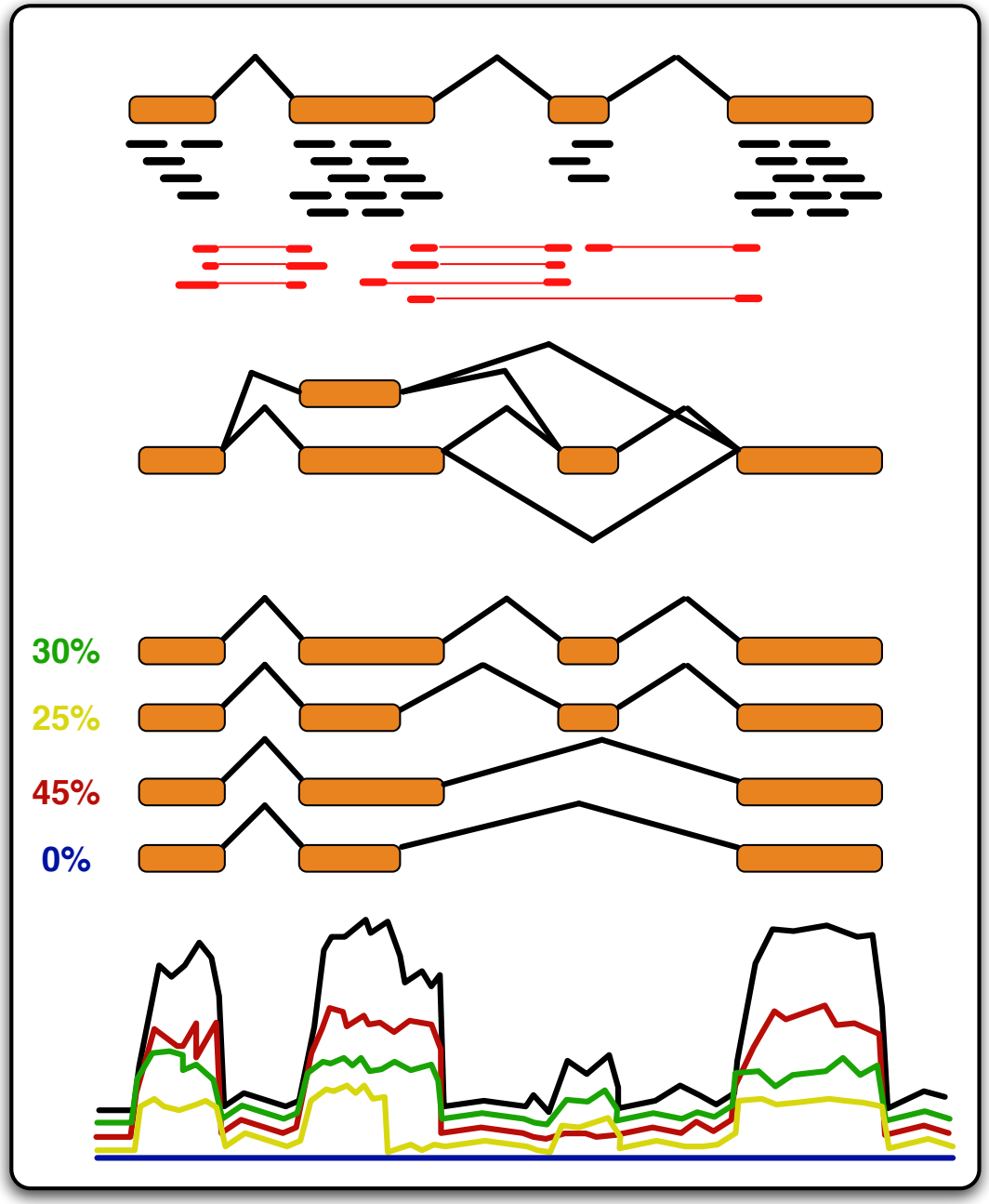
Infer alternative isoforms

- mGene.ngs prediction
- Build splicegraph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



Infer alternative isoforms

- mGene.ngs prediction
- Build splicegraph using spliced reads
- Generate transcripts from graph
- rQuant: explain coverage by sparse, weighted sum of transcripts



Results for *C. elegans*

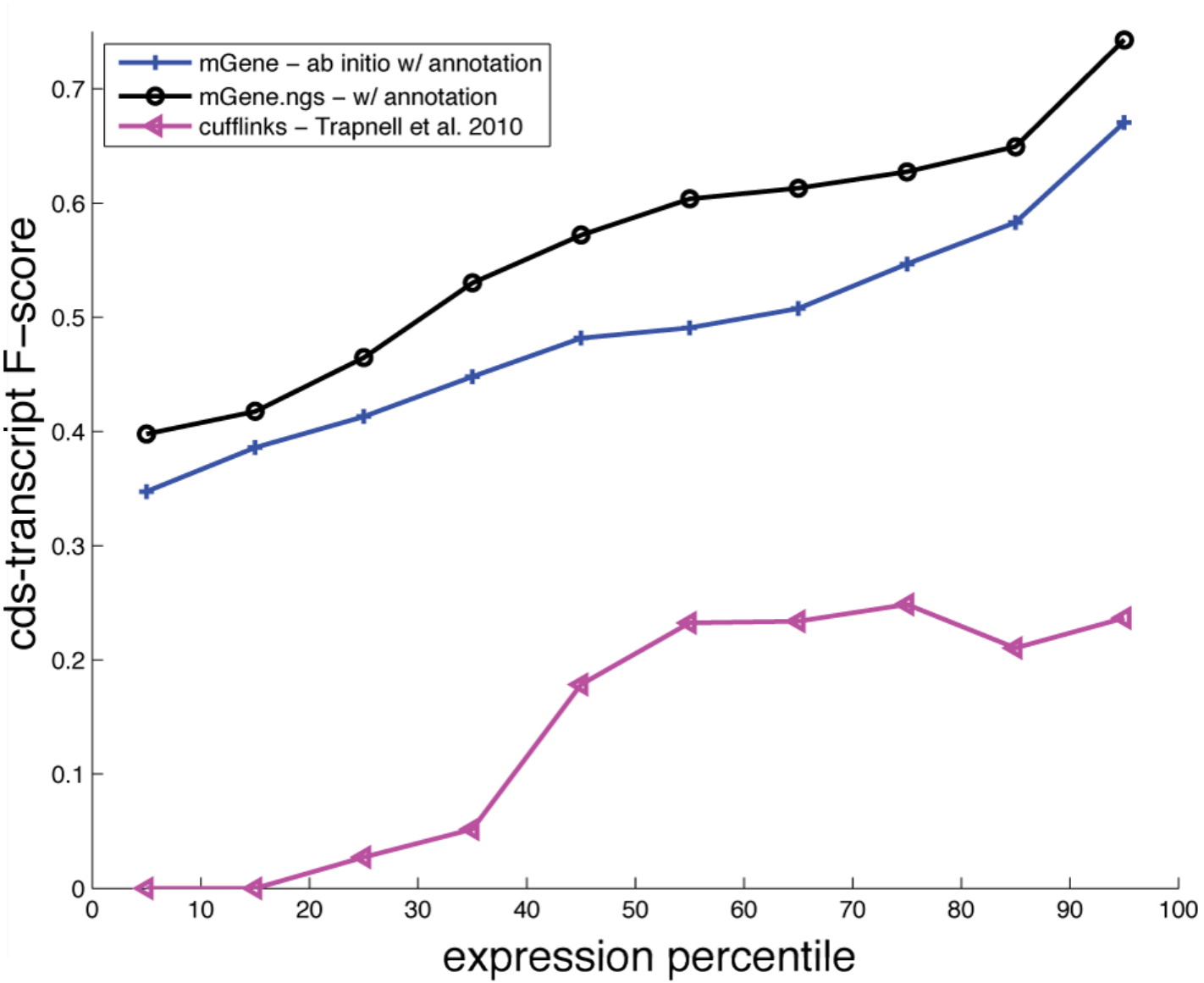
RNA-seq:

- paired-end, strand-specific RNA ligation based protocol
- 76bp reads, 50 million reads
- Alignment with Palmapper

Evaluation:

- Transcript-level F-score of coding transcripts
  ... for different expression levels
- Compare mGene (*ab initio*), mGene.ngs, cufflinks

Results for *C. elegans*



Digestion

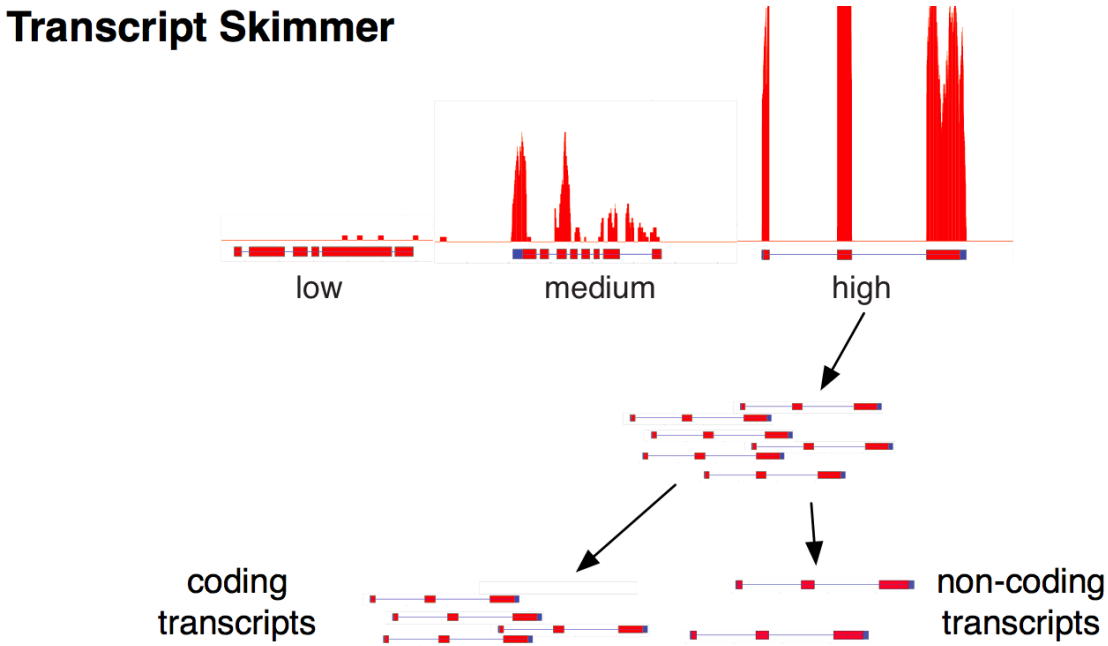
Observations:

- RNA-seq helps to improve performance
- Genomic signals help much (see cufflinks)

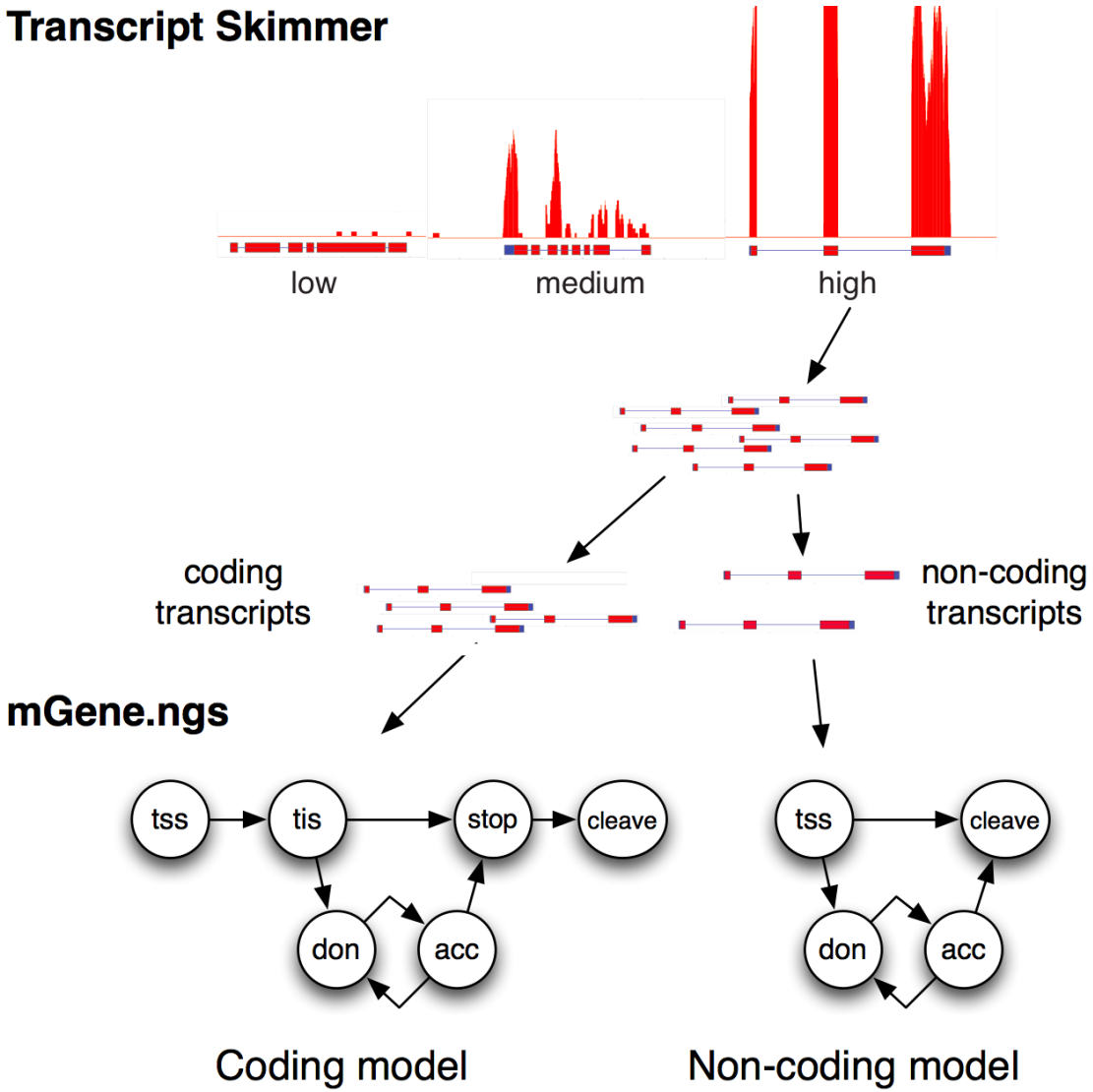
Problems:

- Need existing annotation for training
- Cannot predict non-coding transcripts

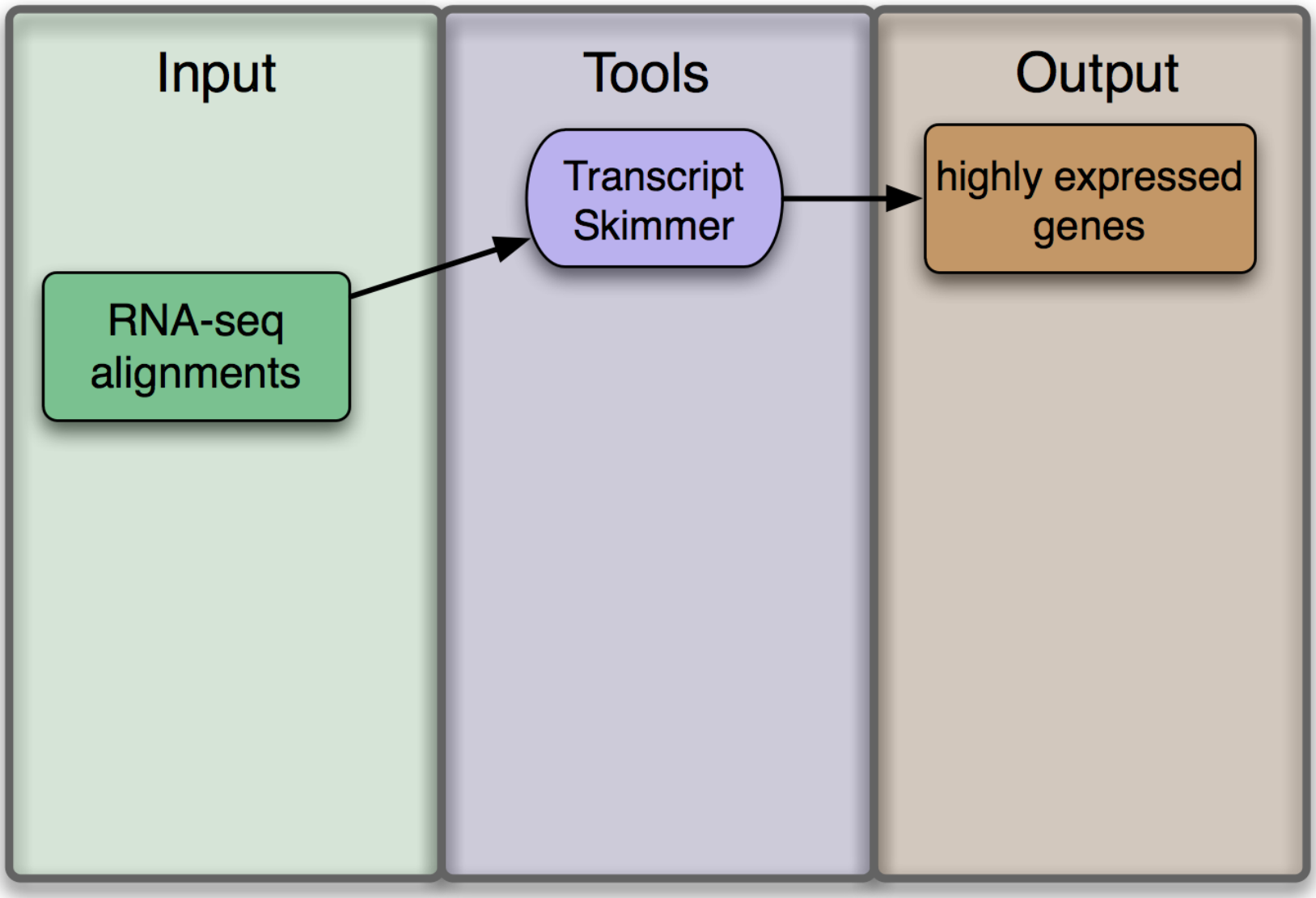
Skimming and Non-coding Transcripts



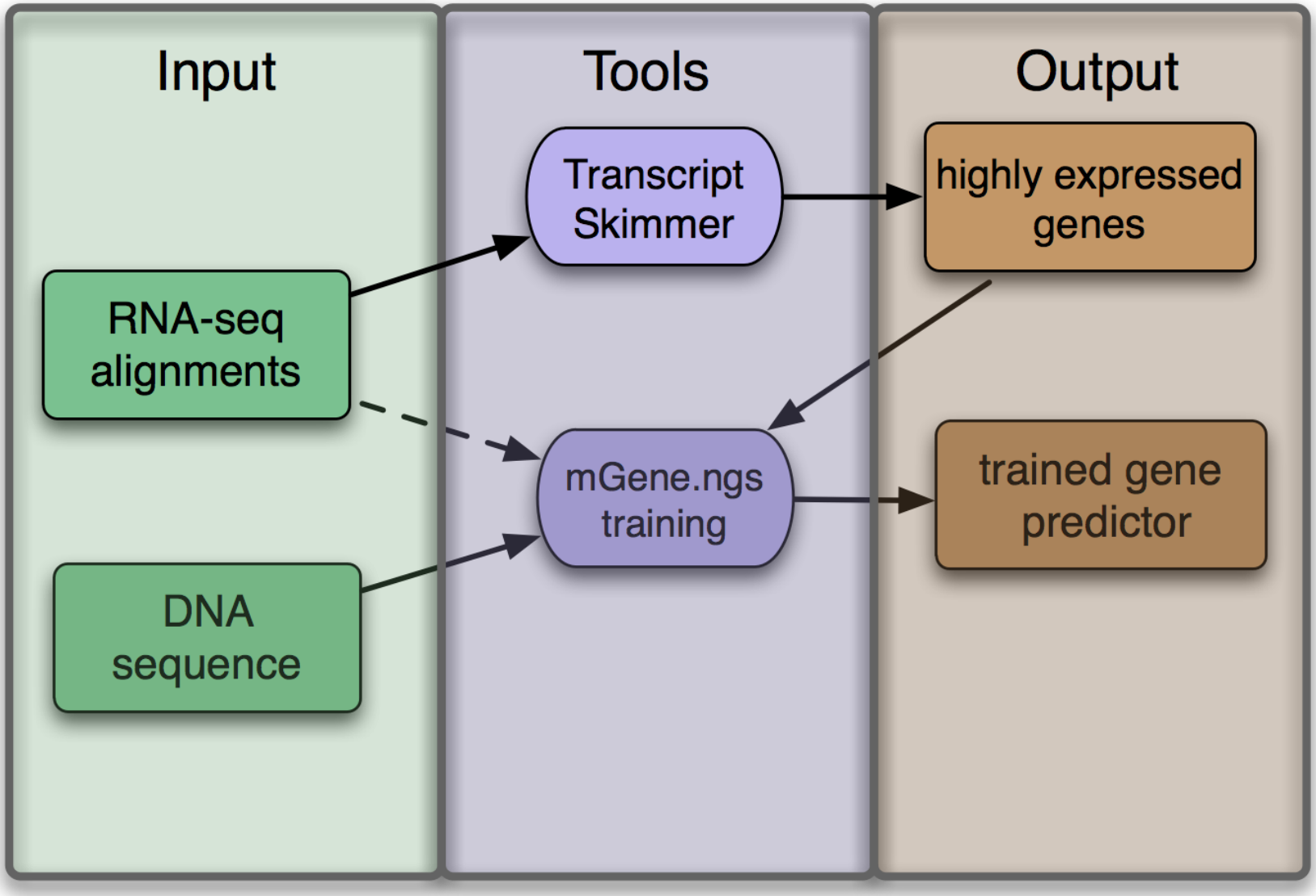
Skimming and Non-coding Transcripts



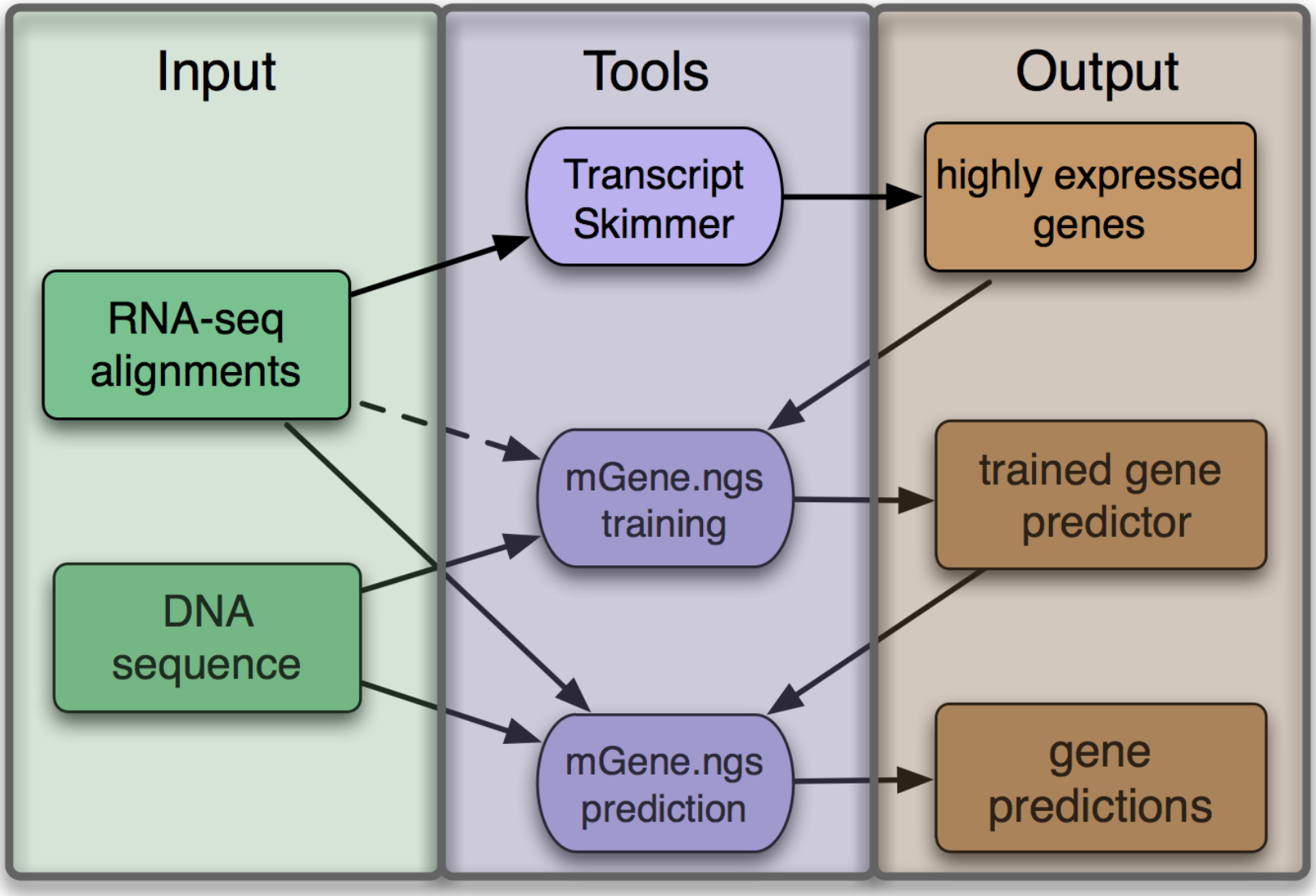
Learning Strategy



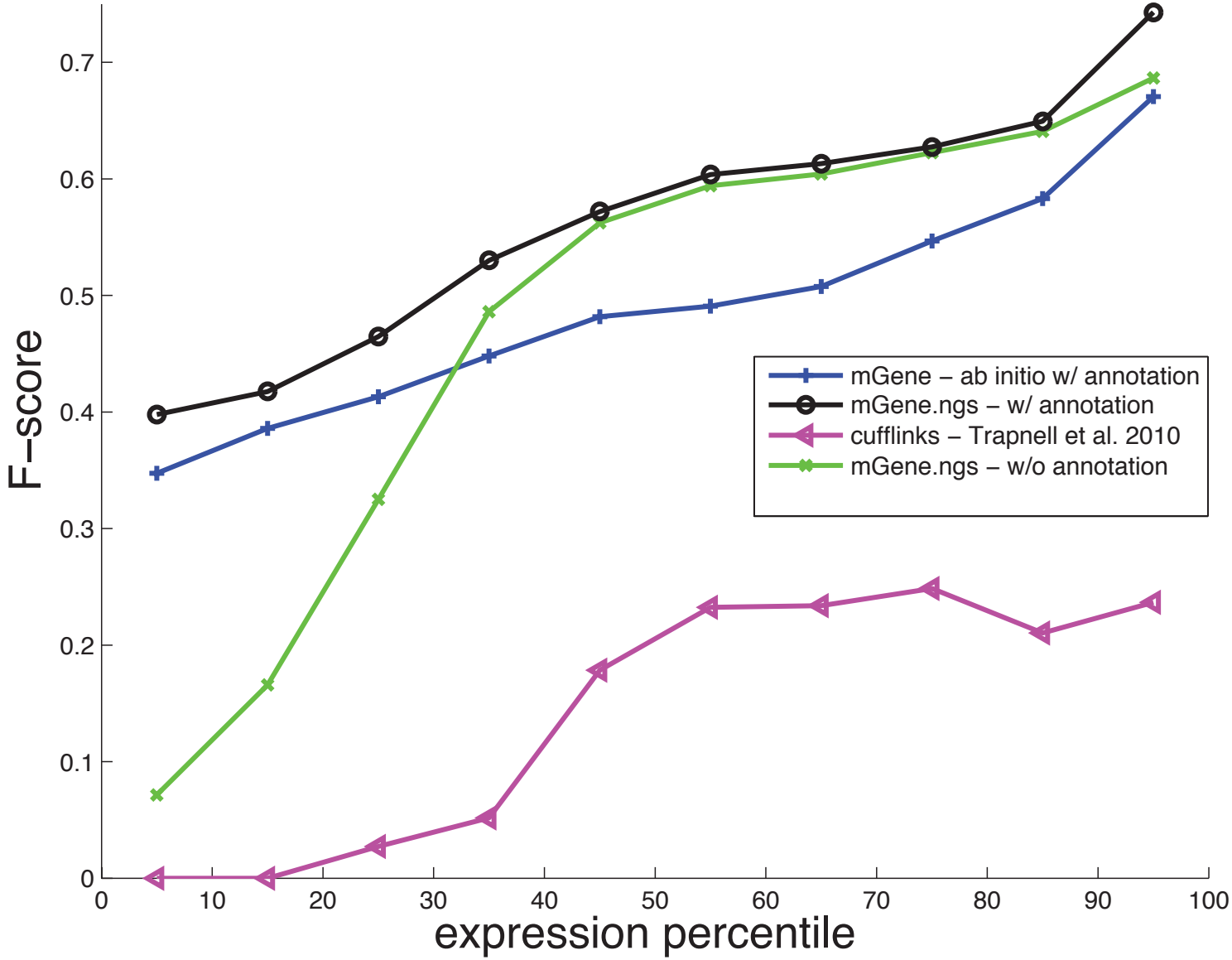
Learning Strategy



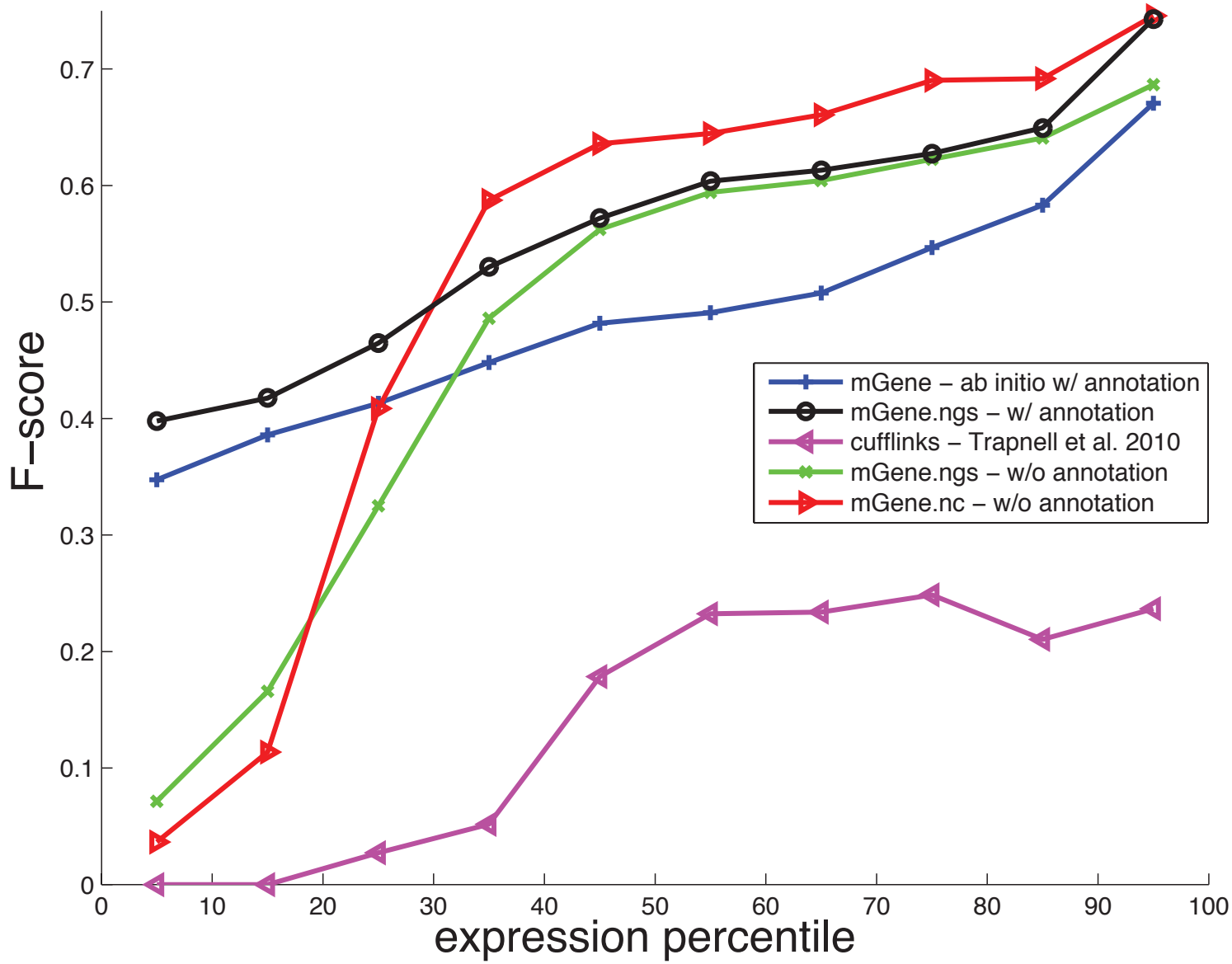
Learning Strategy



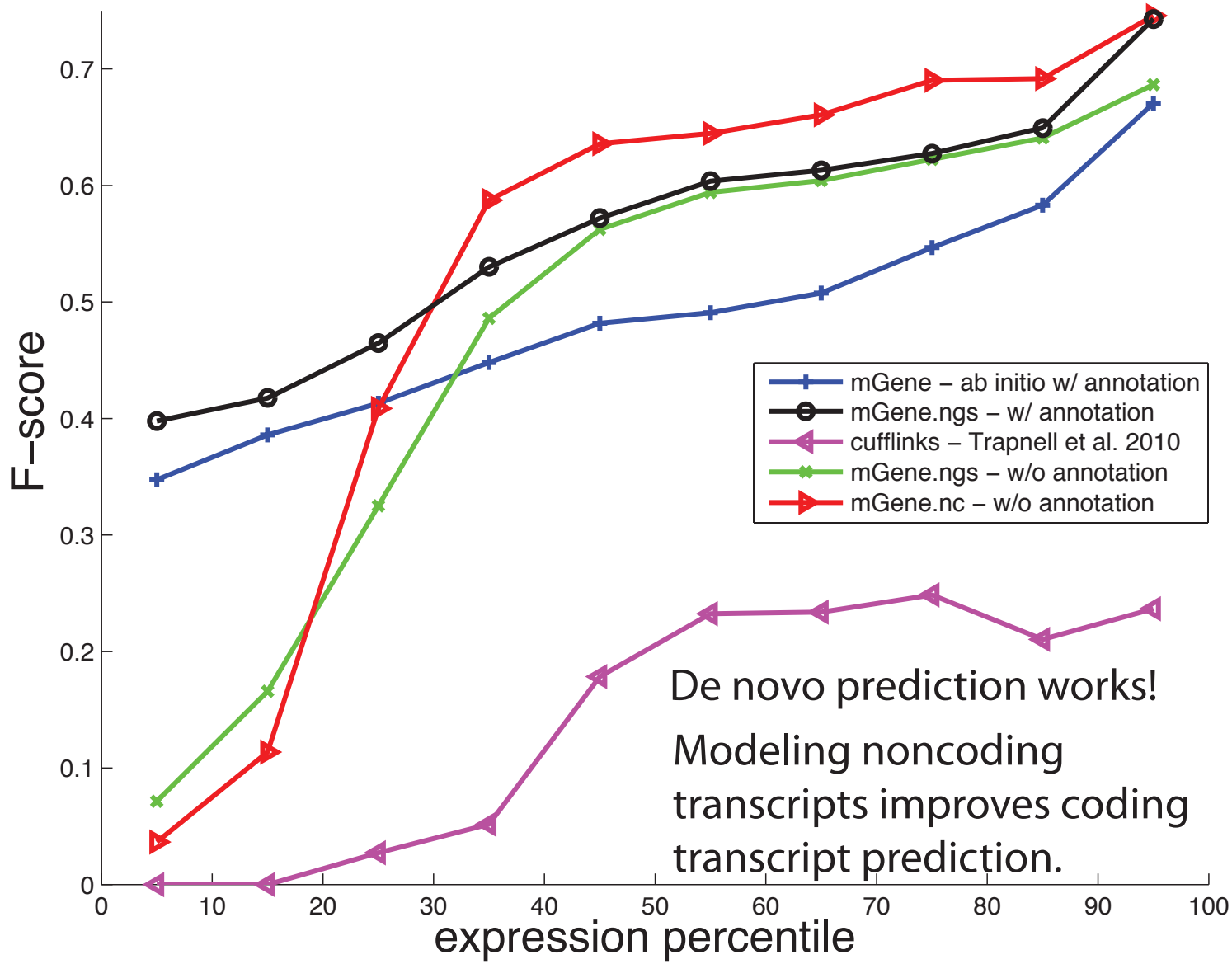
Results for *C. elegans*



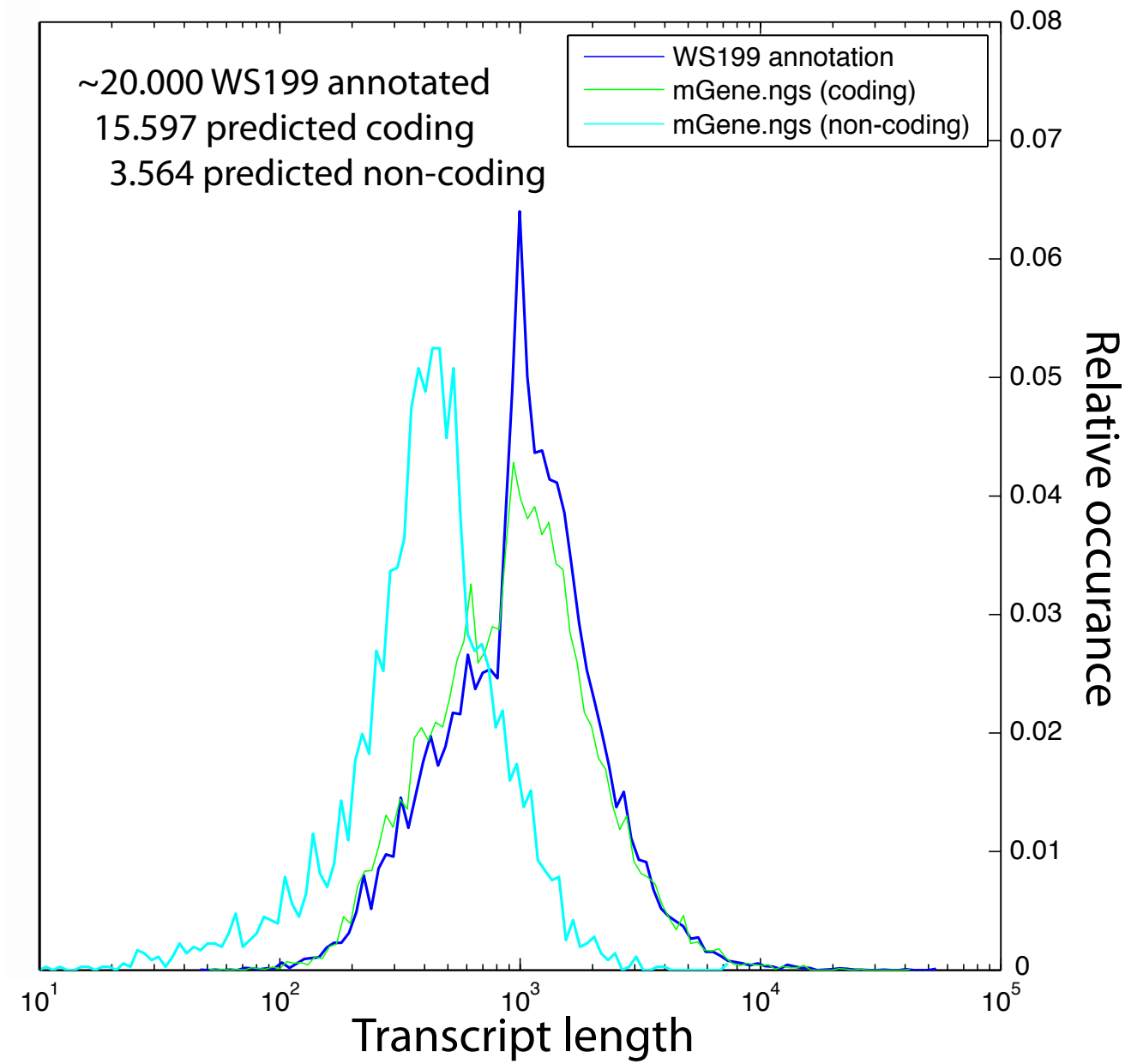
Results for *C. elegans*



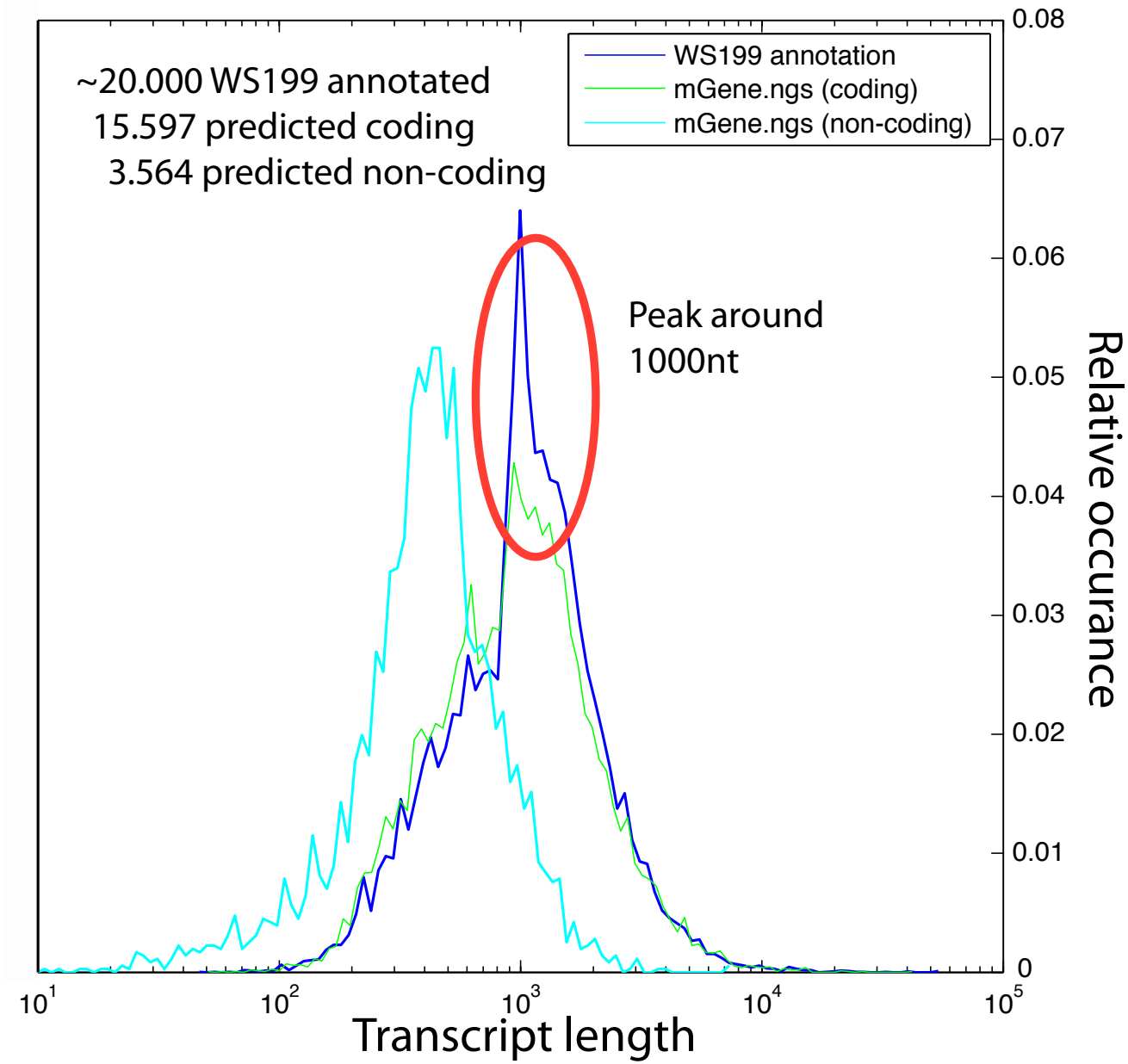
Results for *C. elegans*



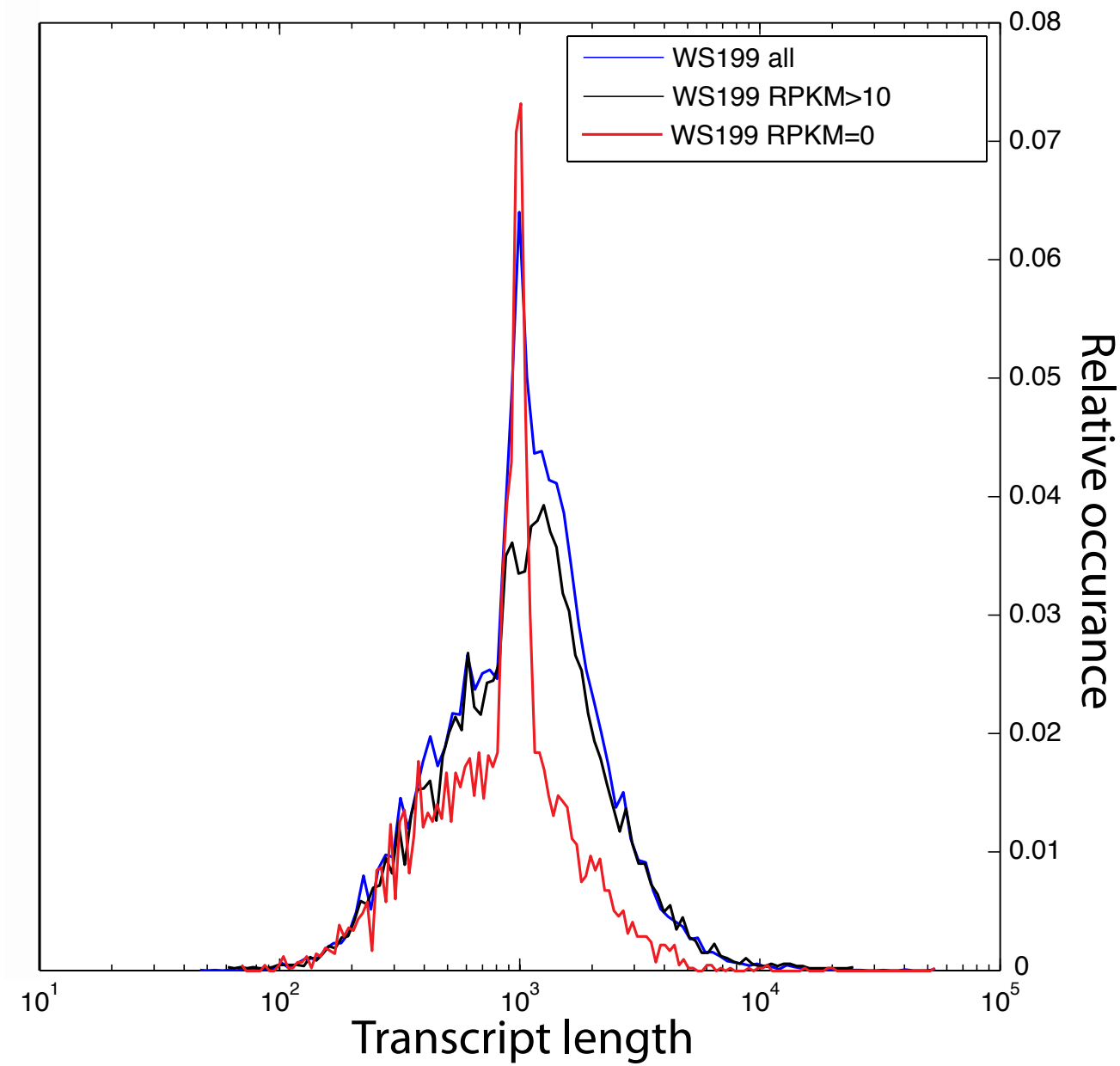
Coding Region Lengths (*C. elegans*)



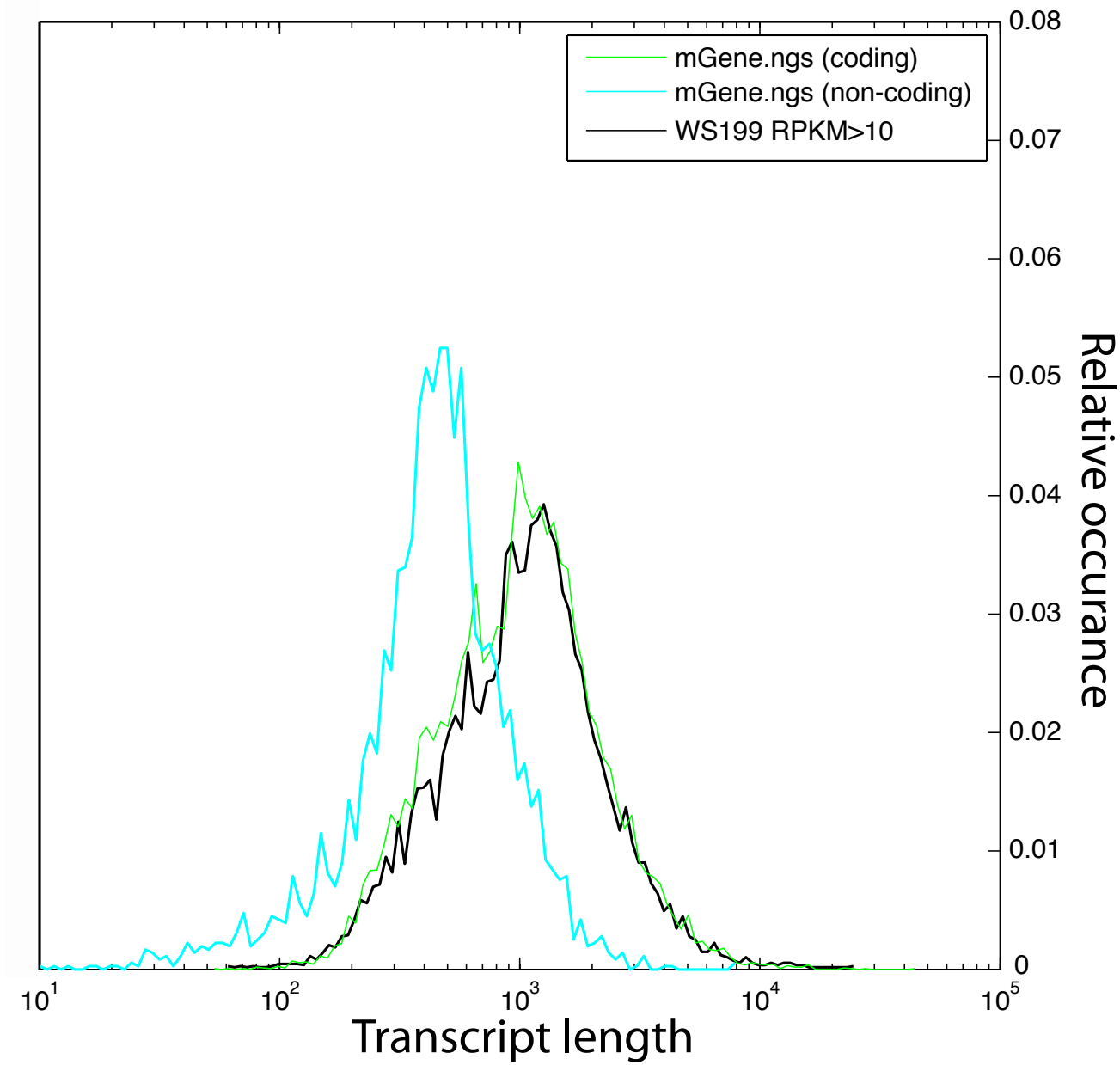
Coding Region Lengths (*C. elegans*)



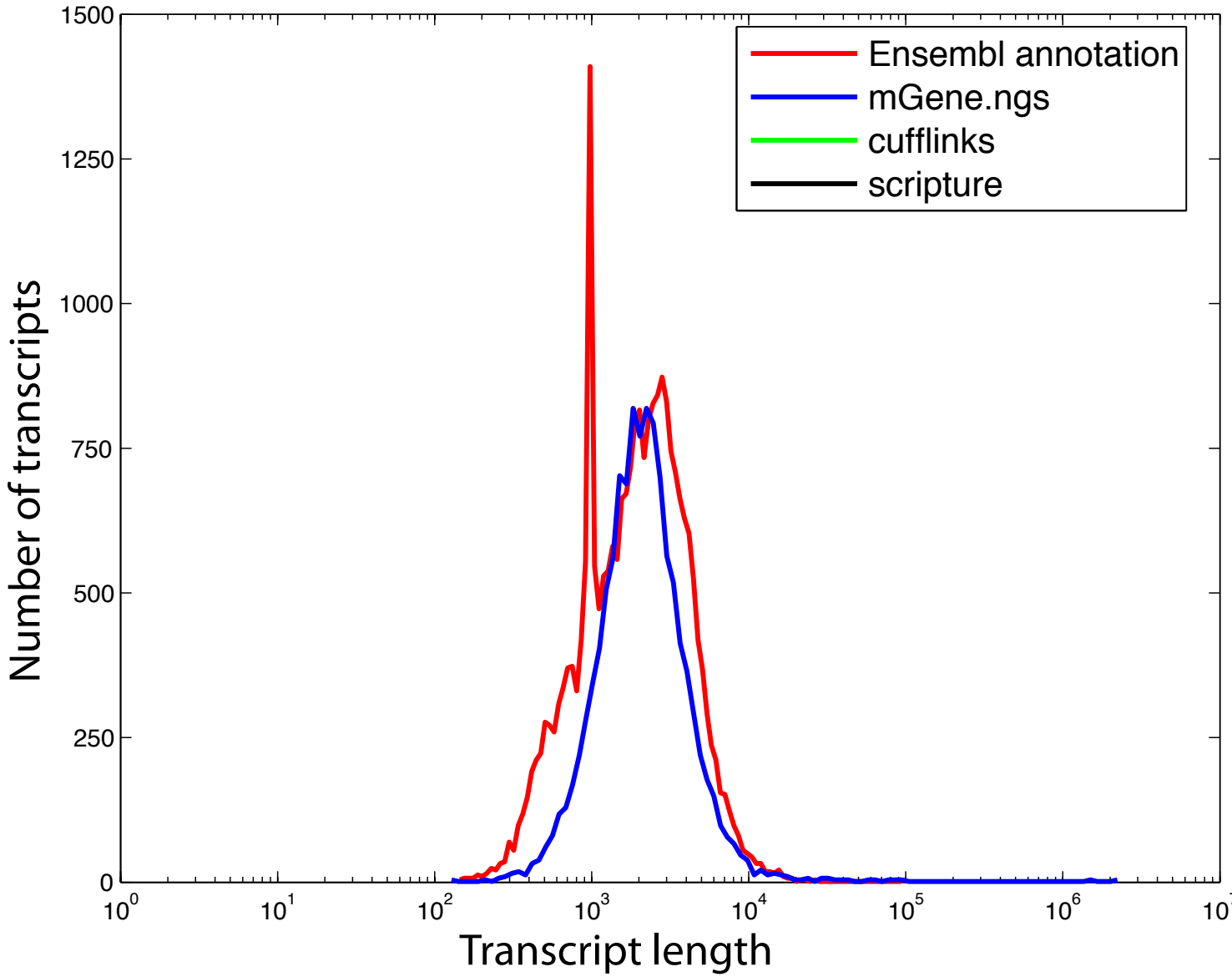
Coding Region Lengths (*C. elegans*)



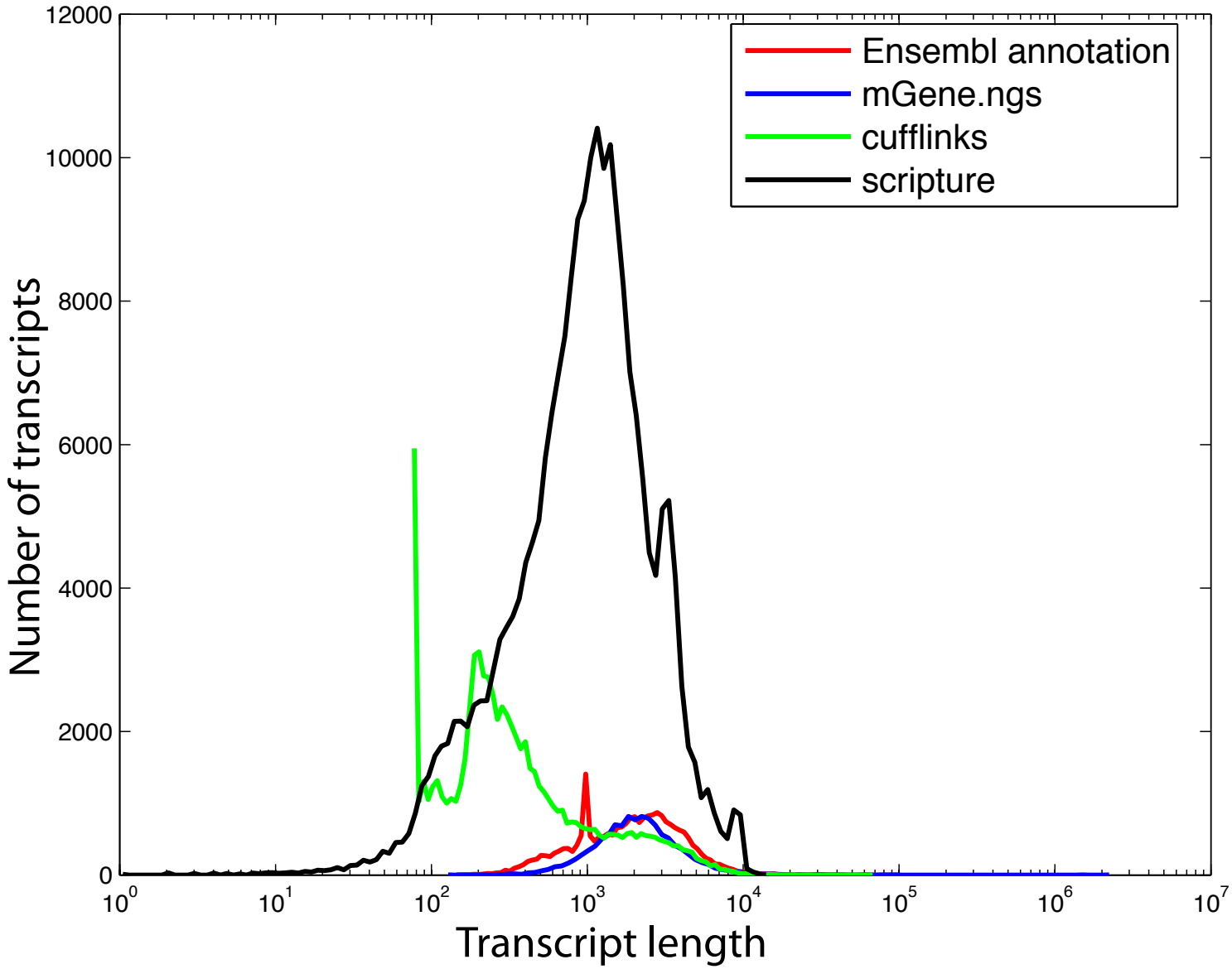
Coding Region Lengths (*C. elegans*)



Coding Region Lengths (mouse)



Coding Region Lengths (mouse)



Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-seq and genomic information
  - rQuant finds parsimonious quantification explaining observed read coverage
  - Transcript Skimmer identifies highly expressed genes for training
  - Accurate prediction of *coding and non-coding transcripts*
- Fully automated training requiring only
  - Genome sequence
  - RNA-seq alignments
- Good for annotating new genomes or improving existing ones
- Identified possible problems in genome annotations of *C. elegans* and mouse
- Source will be free, functionality partially available in Galaxy

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-seq and genomic information
  - rQuant finds parsimonious quantification explaining observed read coverage
  - Transcript Skimmer identifies highly expressed genes for training
  - Accurate prediction of *coding and non-coding transcripts*
- Fully automated training requiring only
  - Genome sequence
  - RNA-seq alignments
- Good for annotating new genomes or improving existing ones
- Identified possible problems in genome annotations of *C. elegans* and mouse
- Source will be free, functionality partially available in Galaxy

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-seq and genomic information
  - rQuant finds parsimonious quantification explaining observed read coverage
  - Transcript Skimmer identifies highly expressed genes for training
  - Accurate prediction of *coding and non-coding transcripts*
- Fully automated training requiring only
  - Genome sequence
  - RNA-seq alignments
- Good for annotating new genomes or improving existing ones
- Identified possible problems in genome annotations of *C. elegans* and mouse
- Source will be free, functionality partially available in Galaxy

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-seq and genomic information
  - rQuant finds parsimonious quantification explaining observed read coverage
  - Transcript Skimmer identifies highly expressed genes for training
  - Accurate prediction of *coding and non-coding transcripts*
- Fully automated training requiring only
  - Genome sequence
  - RNA-seq alignments
- Good for annotating new genomes or improving existing ones
- Identified possible problems in genome annotations of *C. elegans* and mouse
- Source will be free, functionality partially available in Galaxy

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-seq and genomic information
  - rQuant finds parsimonious quantification explaining observed read coverage
  - Transcript Skimmer identifies highly expressed genes for training
  - Accurate prediction of *coding and non-coding transcripts*
- Fully automated training requiring only
  - Genome sequence
  - RNA-seq alignments
- Good for annotating new genomes or improving existing ones
- Identified possible problems in genome annotations of *C. elegans* and mouse
- Source will be free, functionality partially available in Galaxy

Further information

and the slides of this talk at:

fml.mpg.de/raetsch/lectures



Jonas Behr

Acknowledgments

Gene finding:	**Jonas Behr, Georg Zeller, Gabriele Schweikert**
Quantification:	**Regina Bohnert**
Library preparation:	**Lisa Hartmann, Lisa Smith**
Alignments:	**Andre Kahles, Geraldine Jean, Jonas Behr**
Shogun support:	**Sören Sonnenburg**
Discussions:	**Philipp Drewe, Sebastian Schultheiss, Christian Widmer**

Funding: Max Planck Society and German Research Foundation

References

I

Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov Support Vector Machines. In *Proc. 20th Int. Conf. Mach. Learn.*, pages 3–10, 2003.

R. Bohnert and G. Rätsch. rQuant.web: a tool for RNA-seq-based transcript quantitation. *NAR Webserver Issue*, 38(Suppl): W348–51, 2010.

Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics (Oxford, England)*, 24(16):i174–180, August 2008.

References II

G Jean, A Kahles, VT Sreedharan, F De Bona, and G Rätsch. Rna-seq read alignments with palmapper. *Current Protocols in Bioinformatics*, 2010.

G Rätsch and S Sonnenburg. Large scale hidden semi-markov svms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NIPS'06)*, volume 19, pages 1161–1168, Cambridge, MA, 2007. MIT Press. URL `http://www.fml.tuebingen.mpg.de/raetsch/projects/HSMSVM`.

O. Stegle, P. Drewe, R. Bohnert, K. Borgwardt, and G. Rätsch. Statistical tests for detecting differential rna-transcript expression from read counts. Technical report, Nature Preceedings, 2010.