Next Generation Genome Annotation with mGene.ngs

Jonas Behr,[1] Regina Bohnert,[1] Georg Zeller,[1,2] Gabriele Schweikert,[1,2,3] Lisa Hartmann,[1] and Gunnar Rätsch[1]

[1] Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany
[2] Max Planck Institute for Developmental Biology, Tübingen, Germany
[3] Max Planck Institute for Biological Cybernetics, Tübingen, Germany



ISCB-SC, July 9, 2010

What is mGene.ngs doing?



- Task: Identification of protein coding genes
- Why is this Task important:
  - Large number of newly sequenced genomes
  - Automated annotation highly important
- Still not solved:
  Ab initio gene prediction:
  C. elegans 50% Coghlan et al. [2008]
  H. sapiens 20% Guigó et al. [2006]
- What can be done?
  Exploit Next Generation mRNA sequencing (RNA-seq) data

What is mGene.ngs doing?



- Task: Identification of protein coding genes
- Why is this Task important:
  - Large number of newly sequenced genomes
  - Automated annotation highly important
- Still not solved:
  Ab initio gene prediction:
  C. elegans 50% Coghlan et al. [2008]
  H. sapiens 20% Guigó et al. [2006]
- What can be done?
  Exploit Next Generation mRNA sequencing (RNA-seq) data

What is mGene.ngs doing?

- Task: Identification of protein coding genes
- Why is this Task important:
  - Large number of newly sequenced genomes
  - Automated annotation highly important
- Still not solved:
  Ab initio gene prediction:
  C. elegans 50% Coghlan et al. [2008]
  H. sapiens 20% Guigó et al. [2006]
- What can be done?
  Exploit Next Generation mRNA sequencing (RNA-seq) data

What is mGene.ngs doing?

- Task: Identification of protein coding genes
- Why is this Task important:
  - Large number of newly sequenced genomes
  - Automated annotation highly important
- Still not solved:
  Ab initio gene prediction:
  C. elegans 50% Coghlan et al. [2008]
  H. sapiens 20% Guigó et al. [2006]
- What can be done?
  Exploit Next Generation mRNA sequencing (RNA-seq) data

Deep RNA Sequencing (RNA-seq)

RNA-seq allows . . .

- High-throughput transcriptome measurements
- Qualitative studies
- Quantitative studies at high resolution

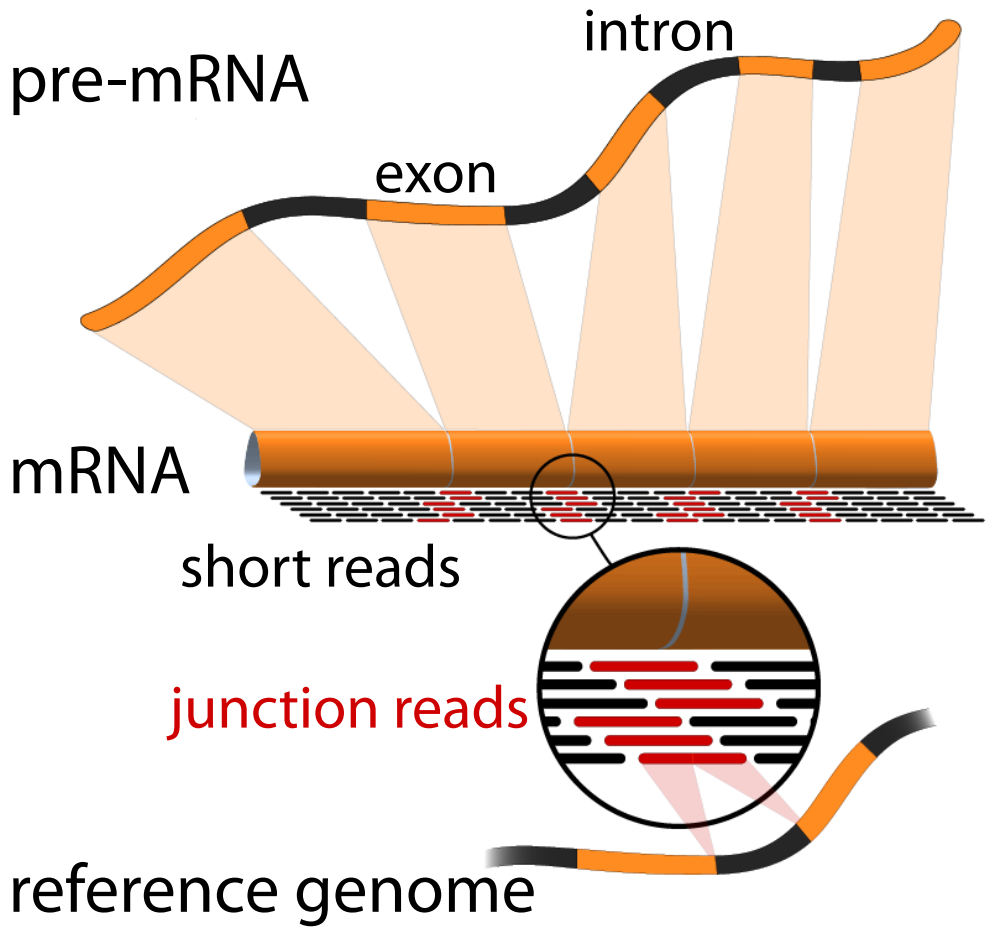


Figure adapted from Wikipedia

Deep RNA Sequencing (RNA-seq)

RNA-seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
- Quantitative studies at high resolution

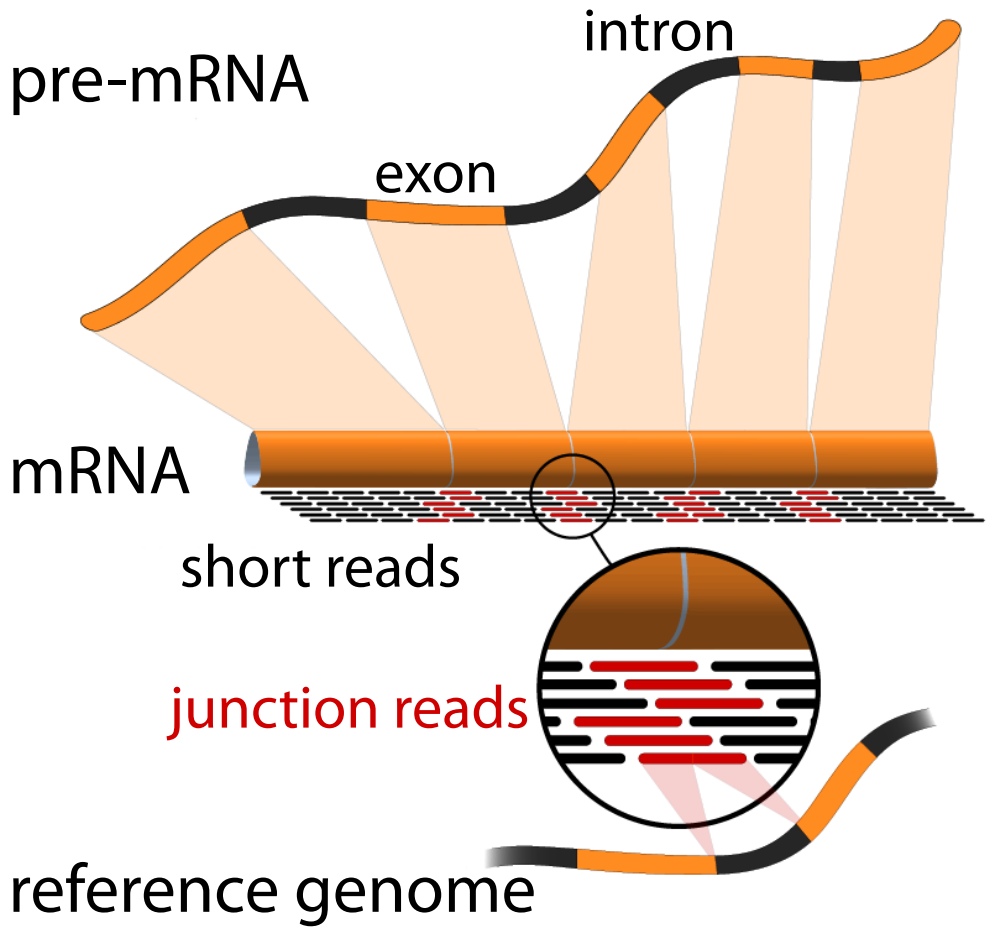


Figure adapted from Wikipedia

Deep RNA Sequencing (RNA-seq)

RNA-seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
- Quantitative studies at high resolution

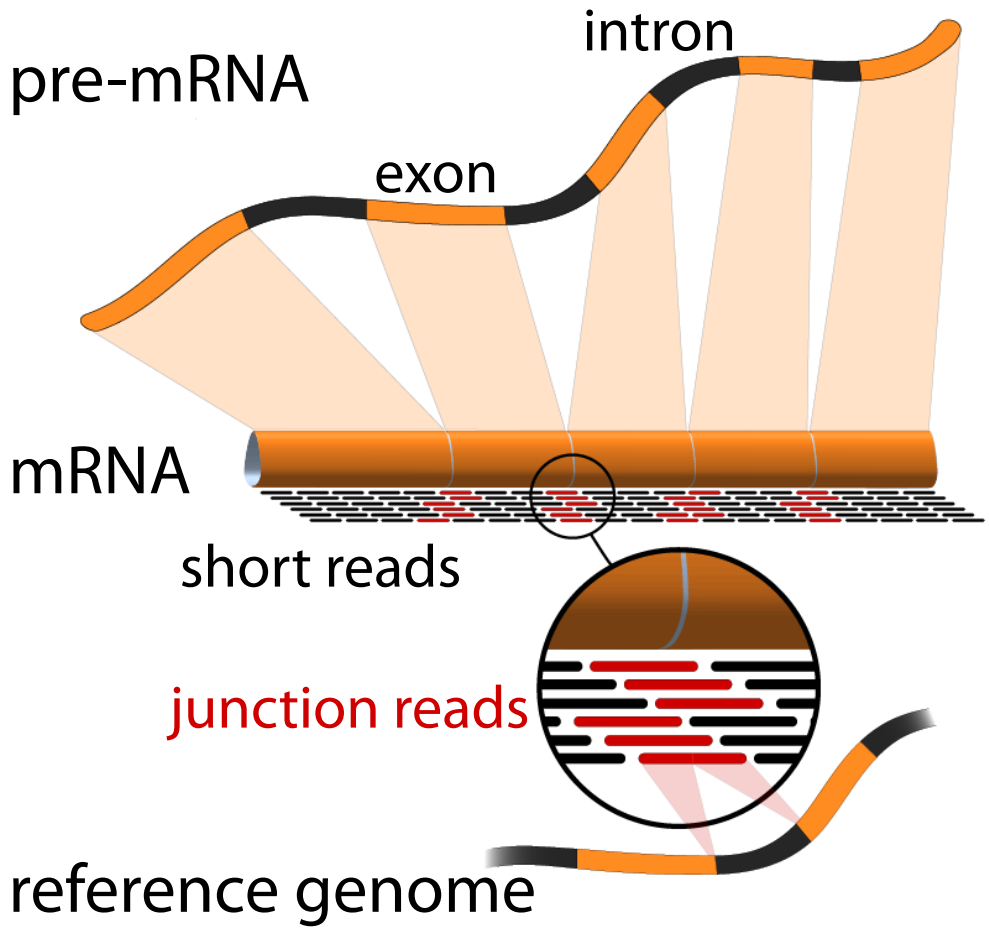
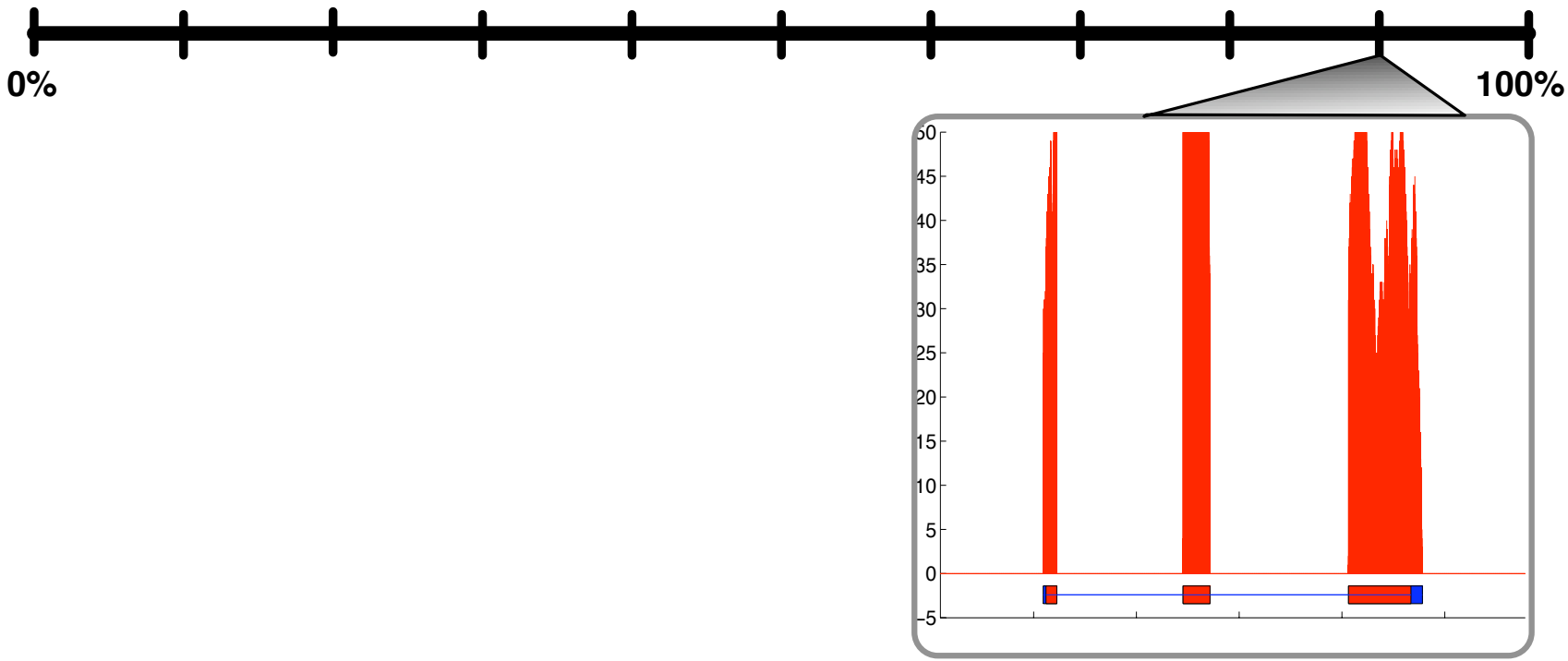


Figure adapted from Wikipedia

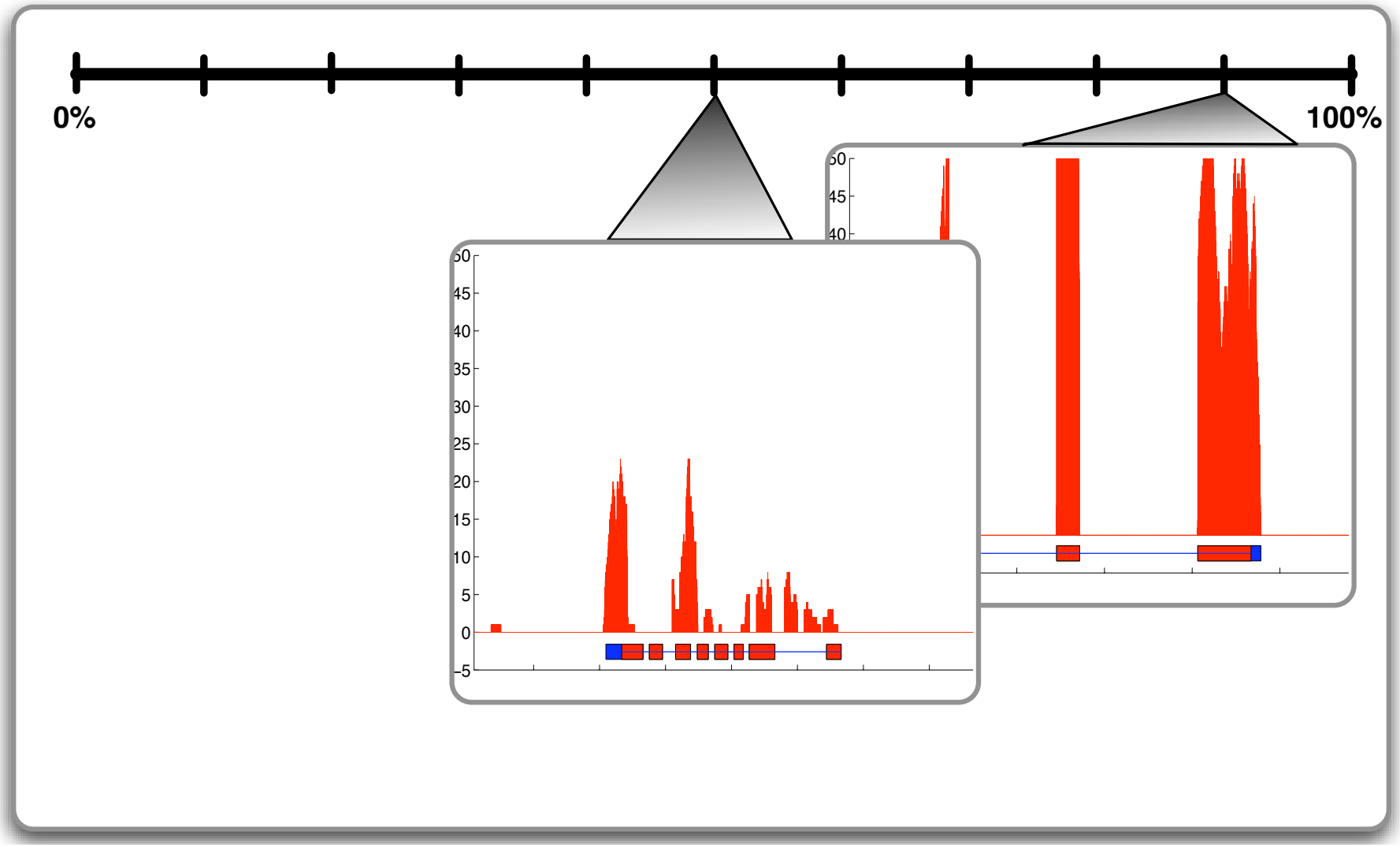
RNA-seq read coverage

Annotated genes sorted by expression level



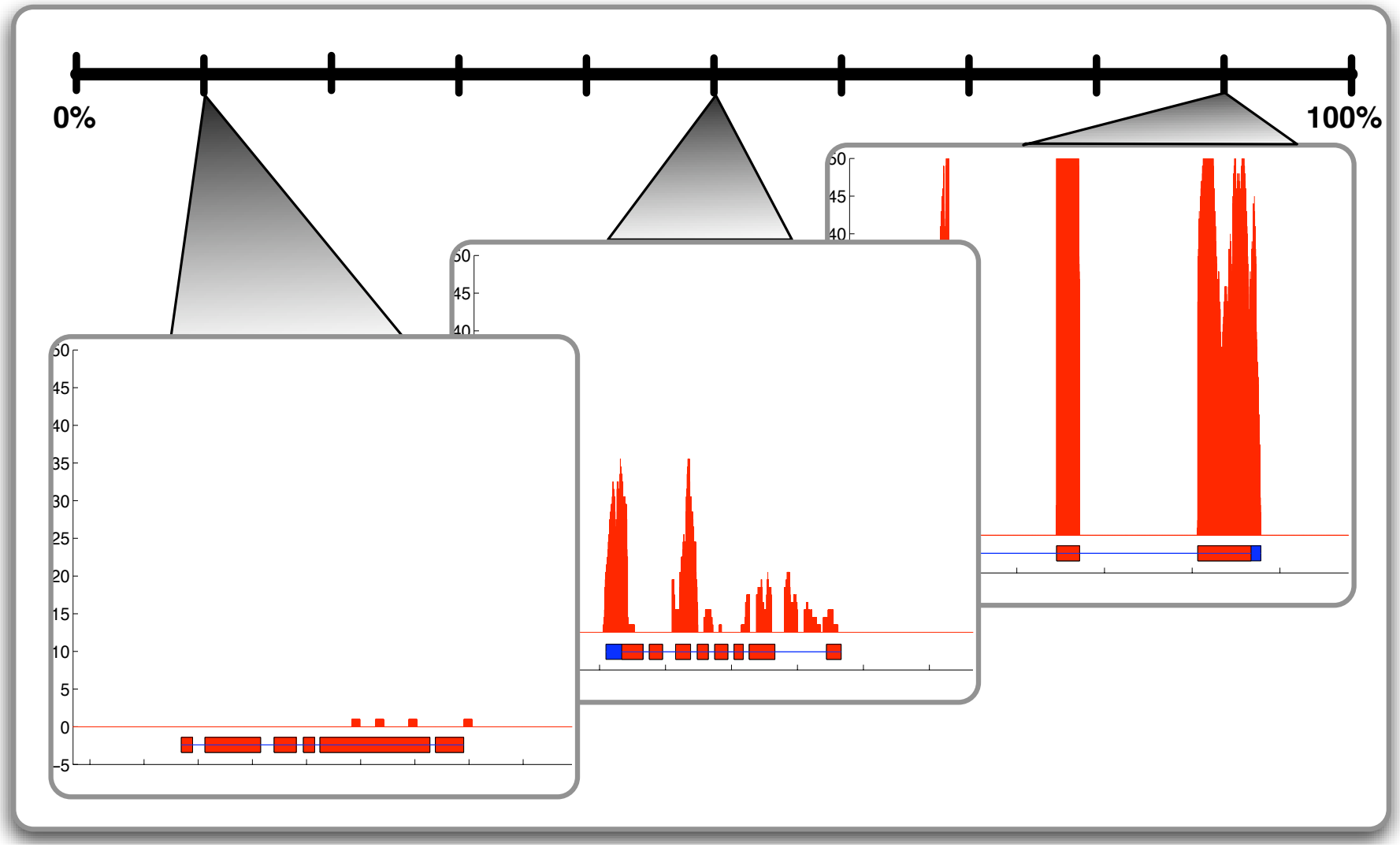
RNA-seq read coverage

Annotated genes sorted by expression level



RNA-seq read coverage

Annotated genes sorted by expression level



Strategy of mGene.ngs

- Integrate RNA-seq and genomic sequence information
  $\Rightarrow$ Hidden semi Markov Support Vector Machines (HsM-SVMs)
- Learn to trade off sources of information during the training
- Address uncertainty in both types of data
- Adapt to error rates of different RNA-seq protocols
- Model long range dependencies e.g ORF constraint

Strategy of mGene.ngs

- Integrate RNA-seq and genomic sequence information
  $\Rightarrow$ Hidden semi Markov Support Vector Machines (HsM-SVMs)
- Learn to trade off sources of information during the training
- Address uncertainty in both types of data
- Adapt to error rates of different RNA-seq protocols
- Model long range dependencies e.g ORF constraint

Strategy of mGene.ngs



- Integrate RNA-seq and genomic sequence information
  $\Rightarrow$ Hidden semi Markov Support Vector Machines (HsM-SVMs)
- Learn to trade off sources of information during the training
- Address uncertainty in both types of data
- Adapt to error rates of different RNA-seq protocols
- Model long range dependencies e.g ORF constraint

Strategy of mGene.ngs

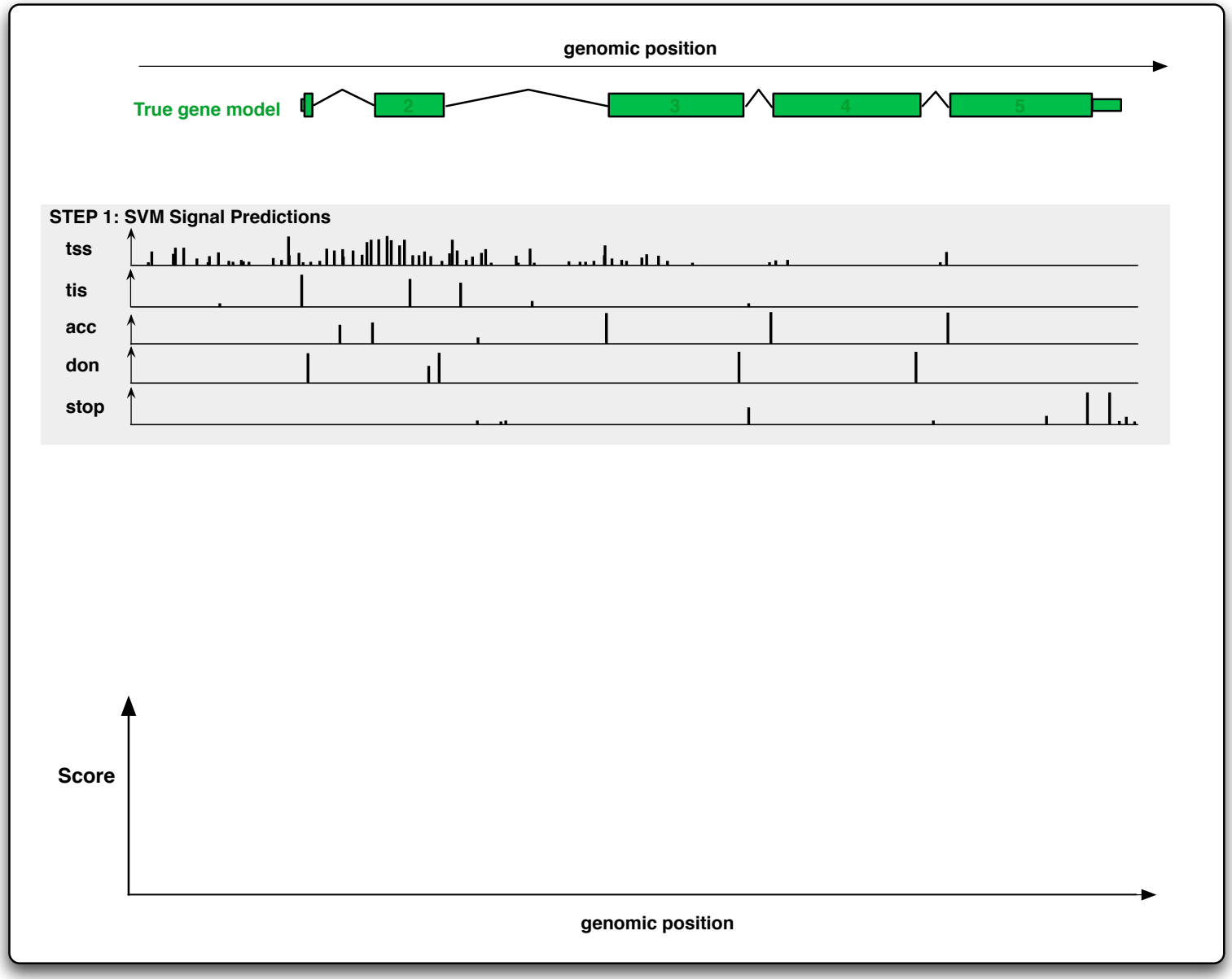


- Integrate RNA-seq and genomic sequence information
  $\Rightarrow$ Hidden semi Markov Support Vector Machines (HsM-SVMs)
- Learn to trade off sources of information during the training
- Address uncertainty in both types of data
- Adapt to error rates of different RNA-seq protocols
- Model long range dependencies e.g ORF constraint

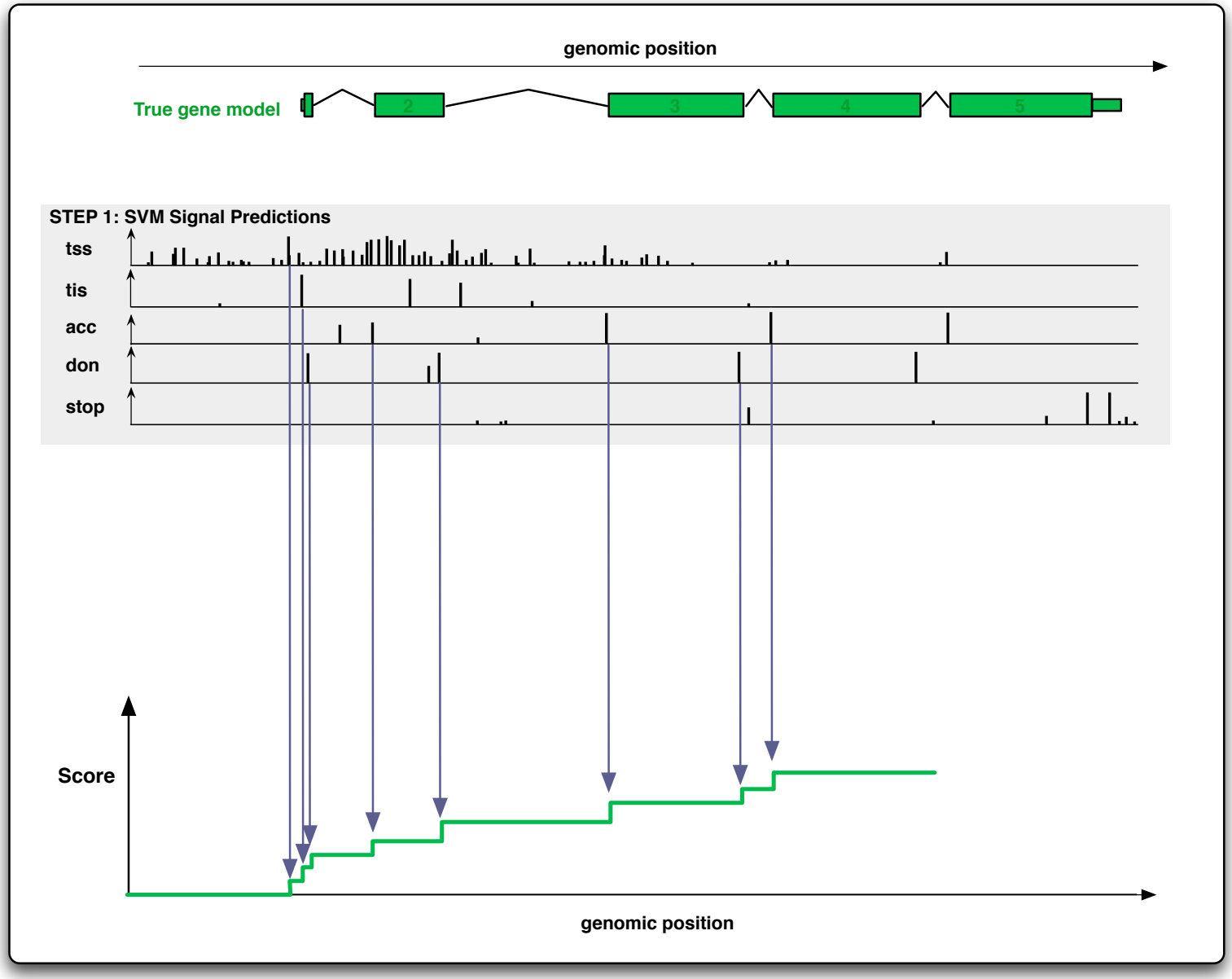
Strategy of mGene.ngs

- Integrate RNA-seq and genomic sequence information
  $\Rightarrow$ Hidden semi Markov Support Vector Machines (HsM-SVMs)
- Learn to trade off sources of information during the training
- Address uncertainty in both types of data
- Adapt to error rates of different RNA-seq protocols
- Model long range dependencies e.g ORF constraint

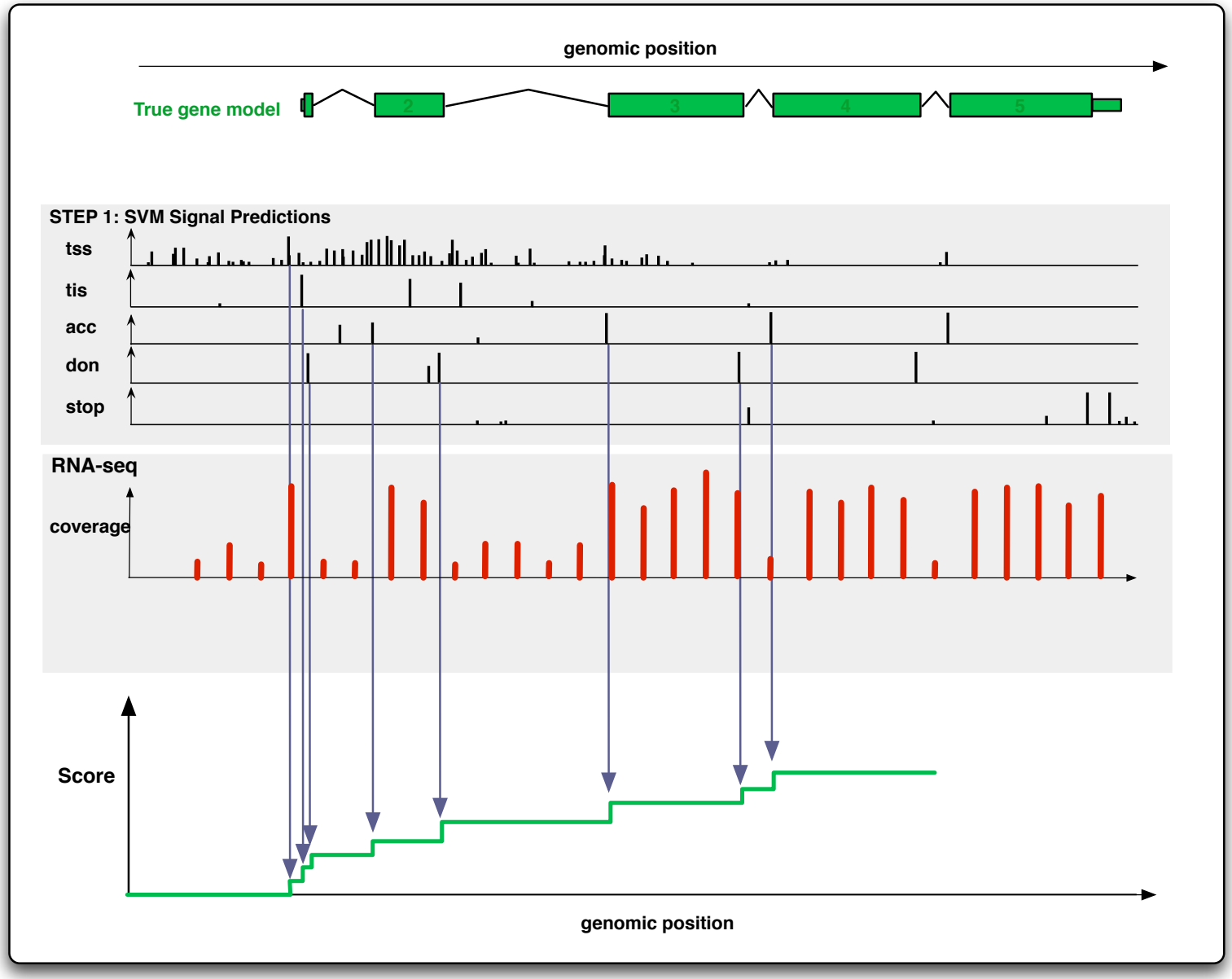
HsM-SVM


genomic position
True gene model
2
3
4
5
STEP 1: SVM Signal Predictions
tss
tis
acc
don
stop
Score
genomic position

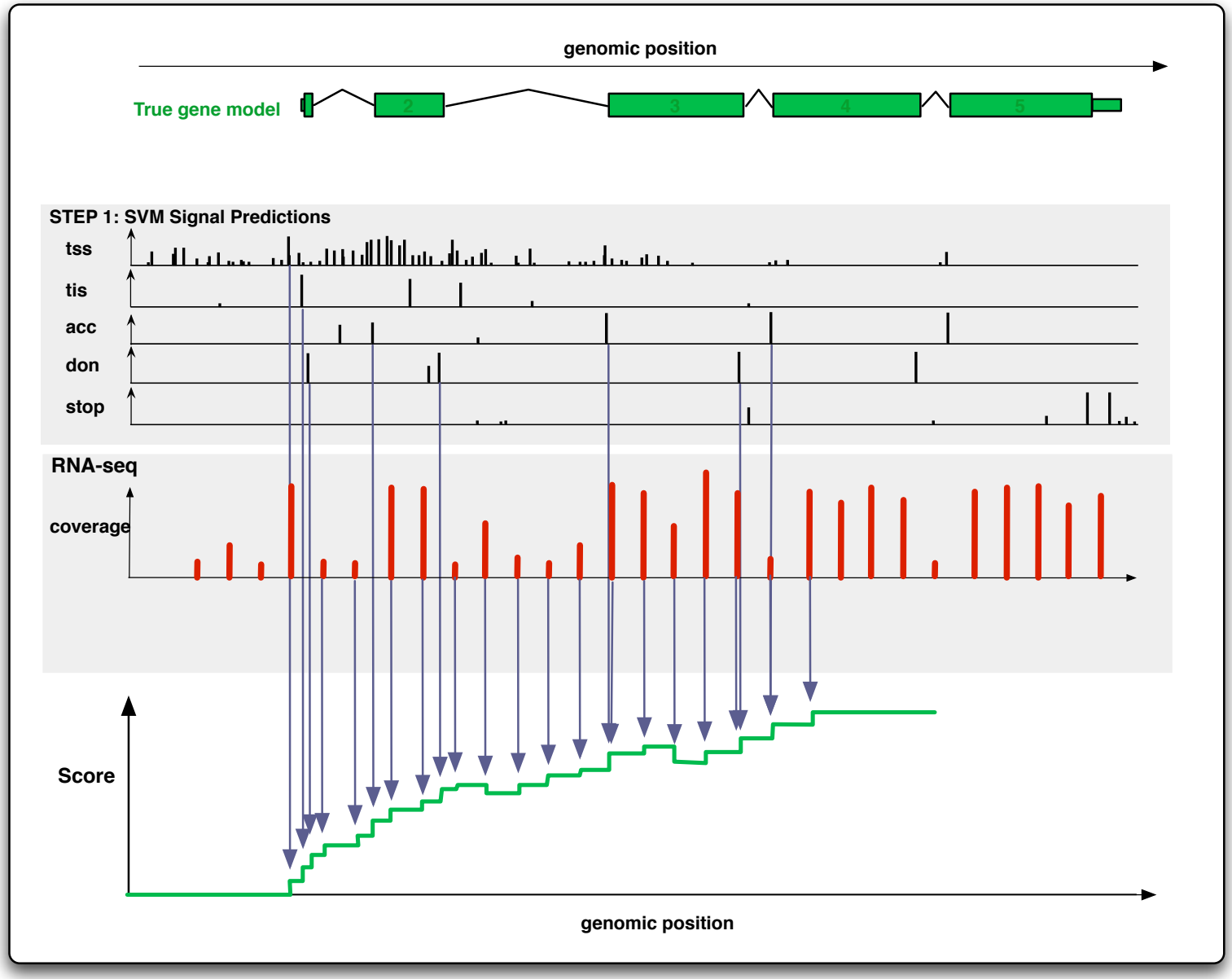
HsM-SVM


genomic position
True gene model
2
3
4
5
STEP 1: SVM Signal Predictions
tss
tis
acc
don
stop
Score
genomic position

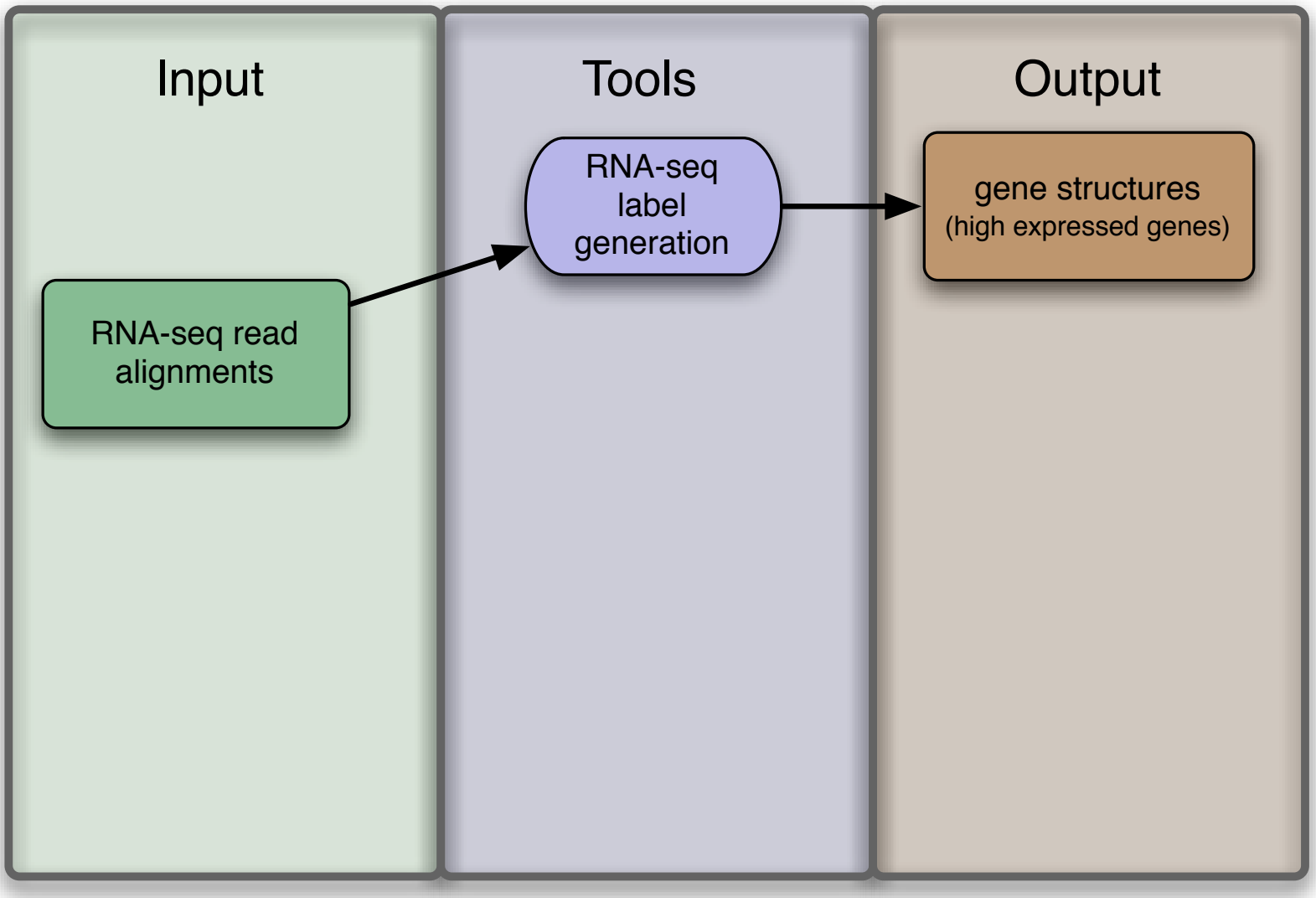
HsM-SVM


genomic position
True gene model
2
3
4
5
STEP 1: SVM Signal Predictions
tss
tis
acc
don
stop
RNA-seq
coverage
Score
genomic position

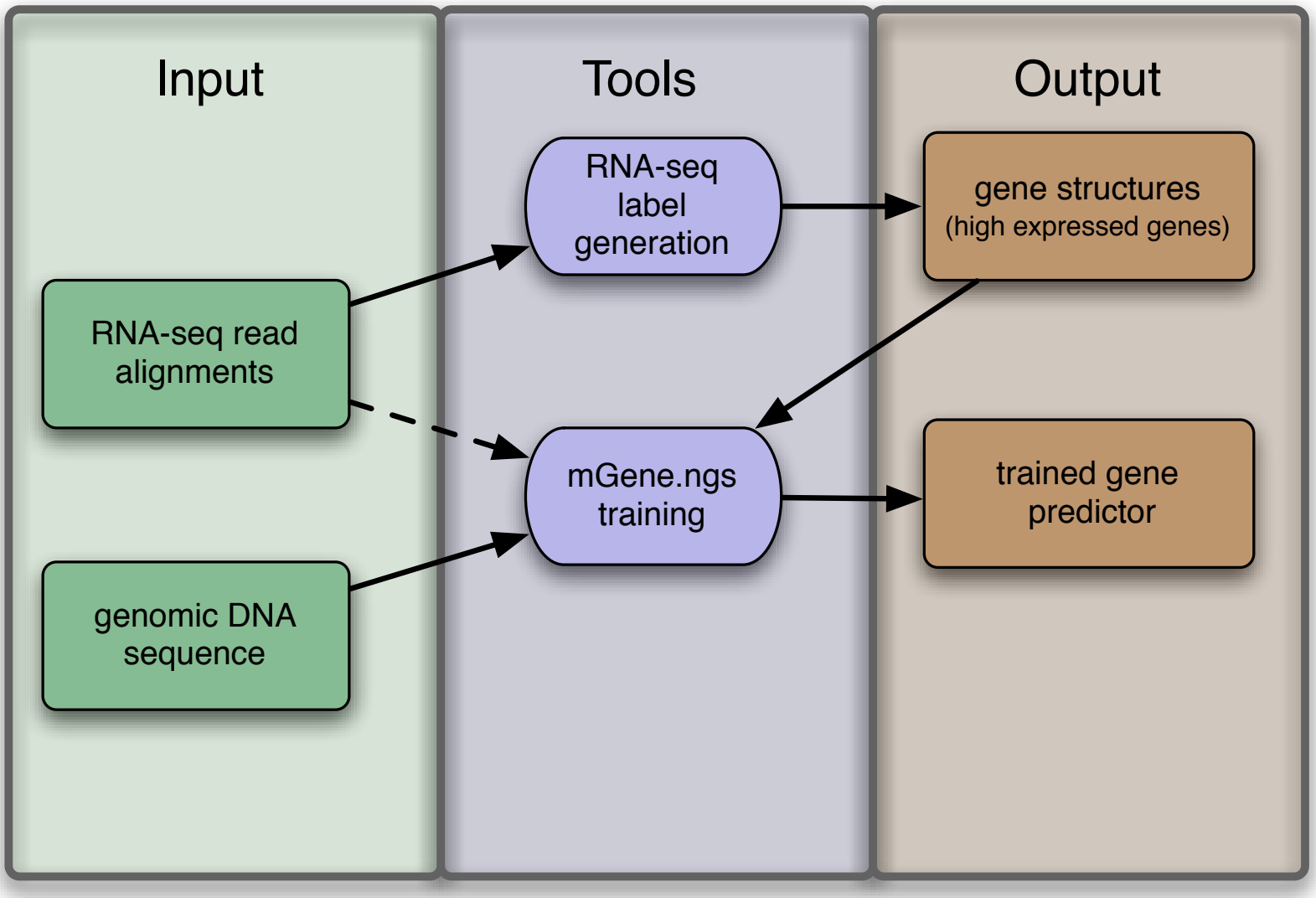
HsM-SVM


genomic position
True gene model
2
3
4
5
STEP 1: SVM Signal Predictions
tss
tis
acc
don
stop
RNA-seq
coverage
Score
genomic position

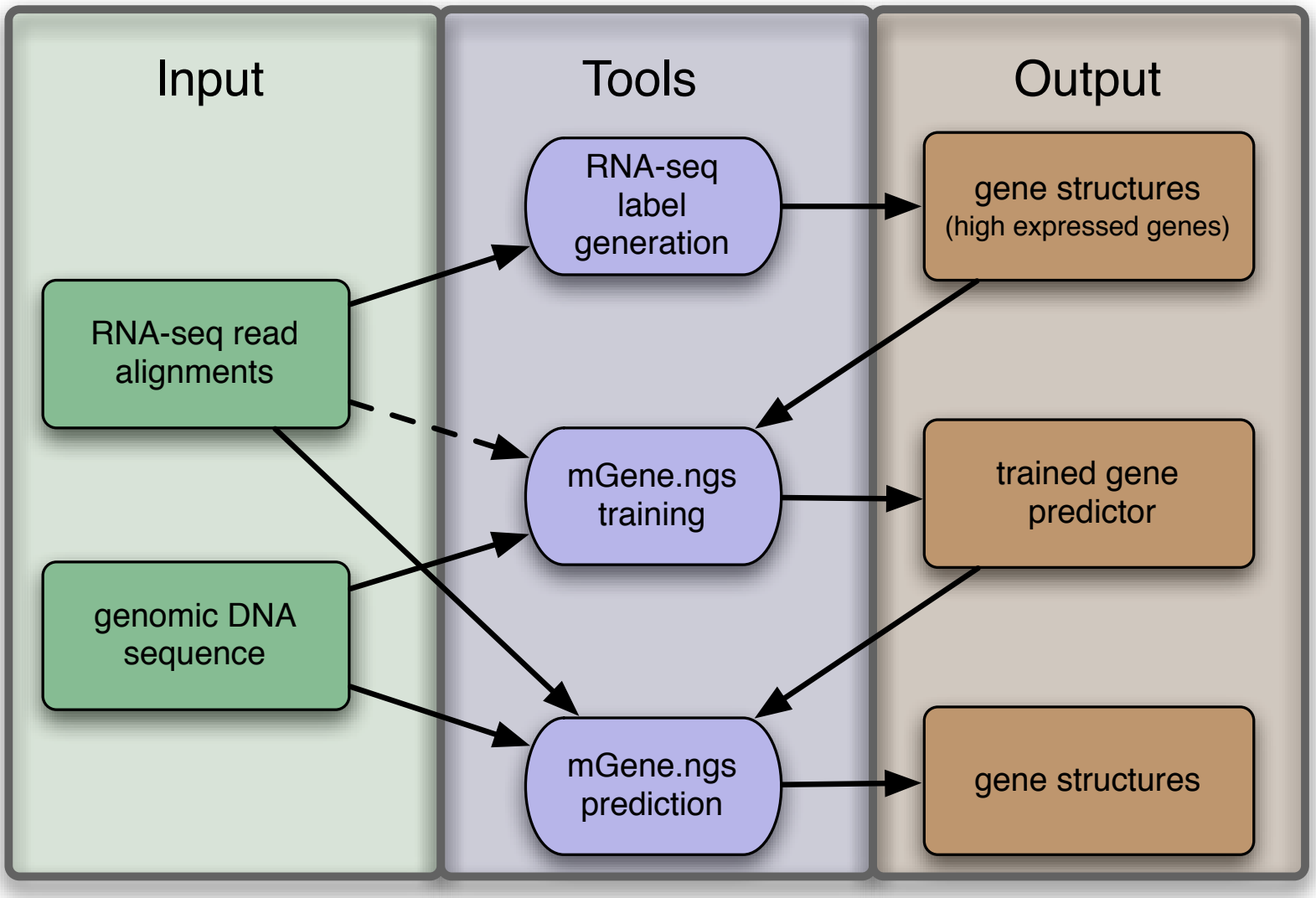
Label generation from RNA-seq data



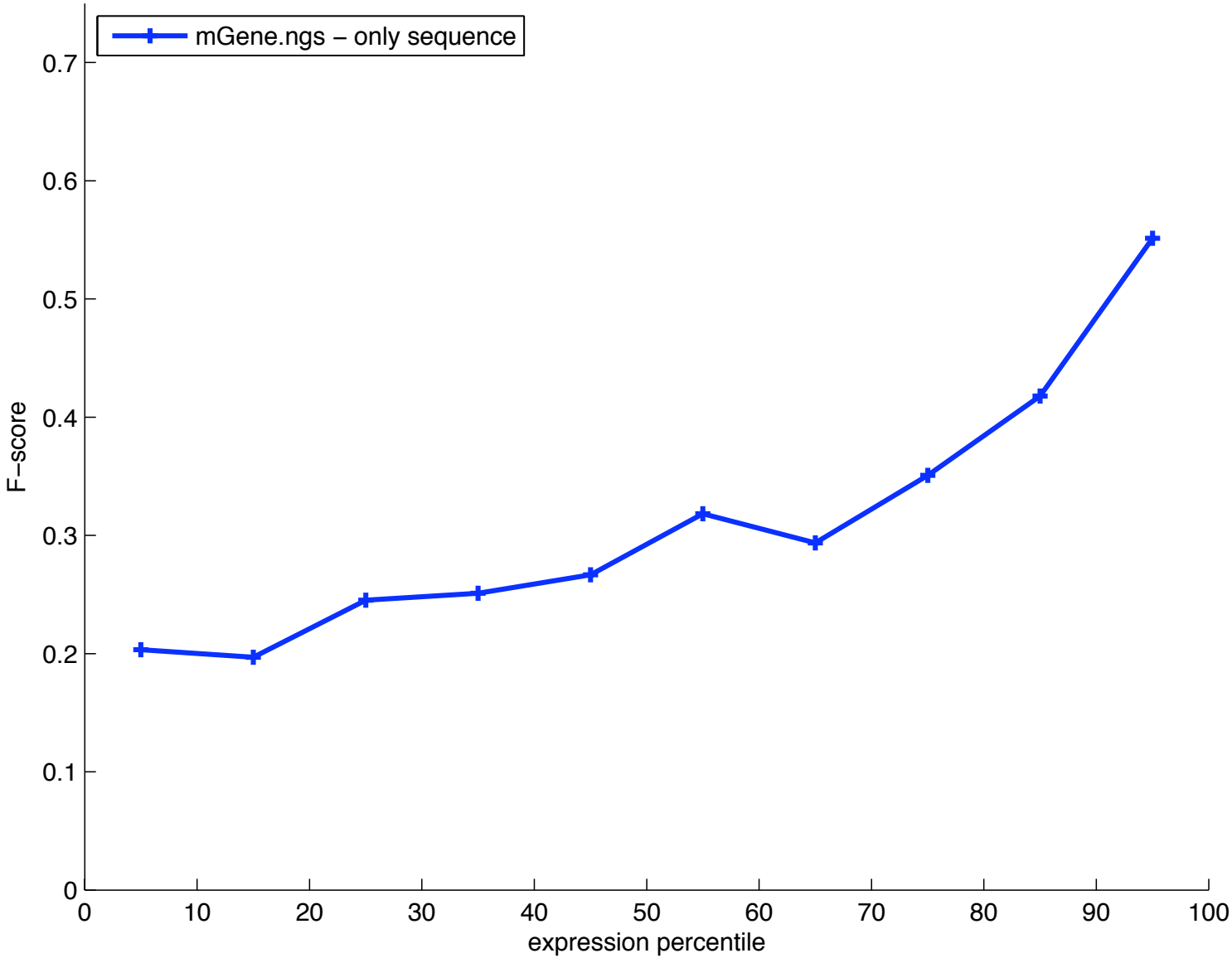
Label generation from RNA-seq data



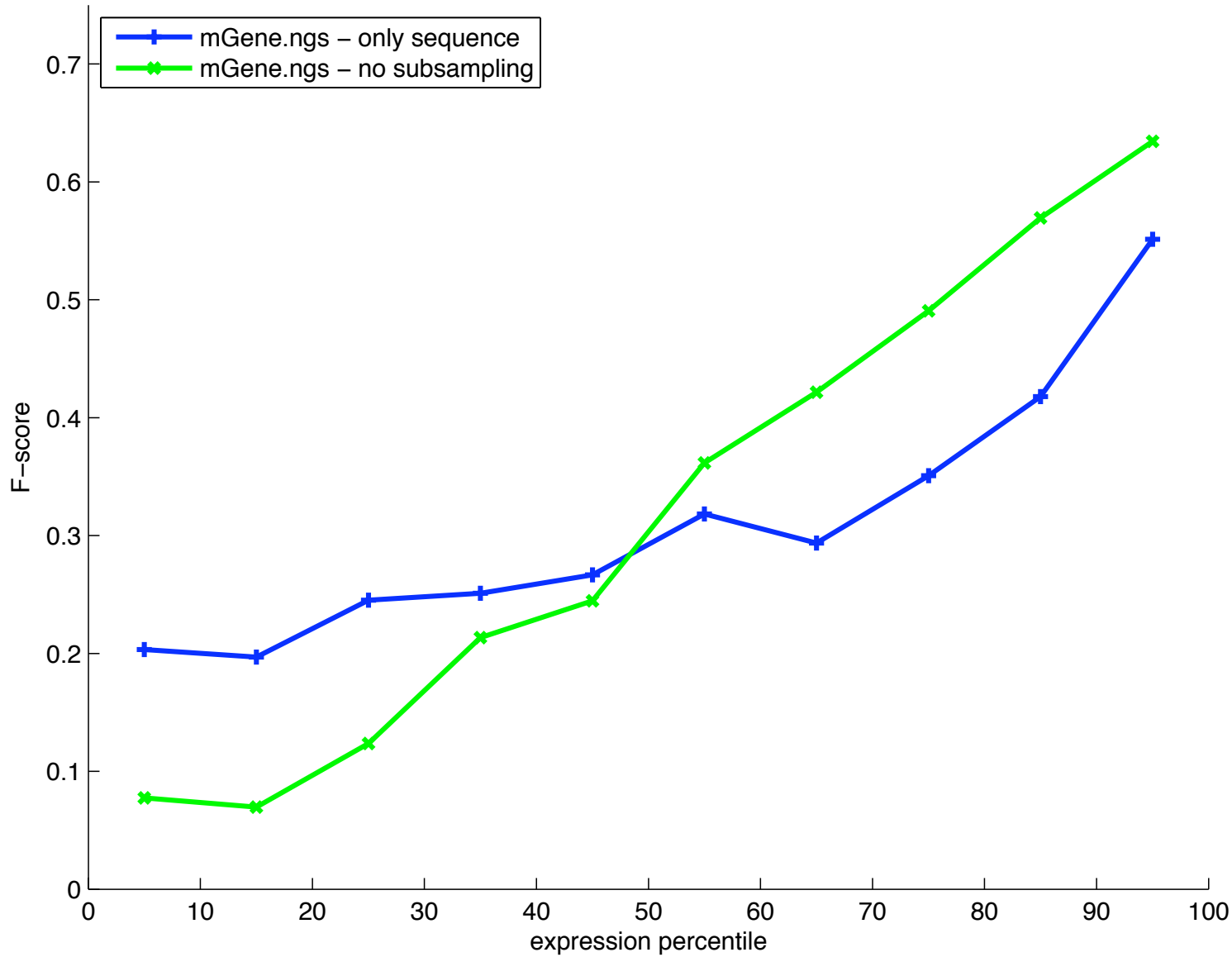
Label generation from RNA-seq data



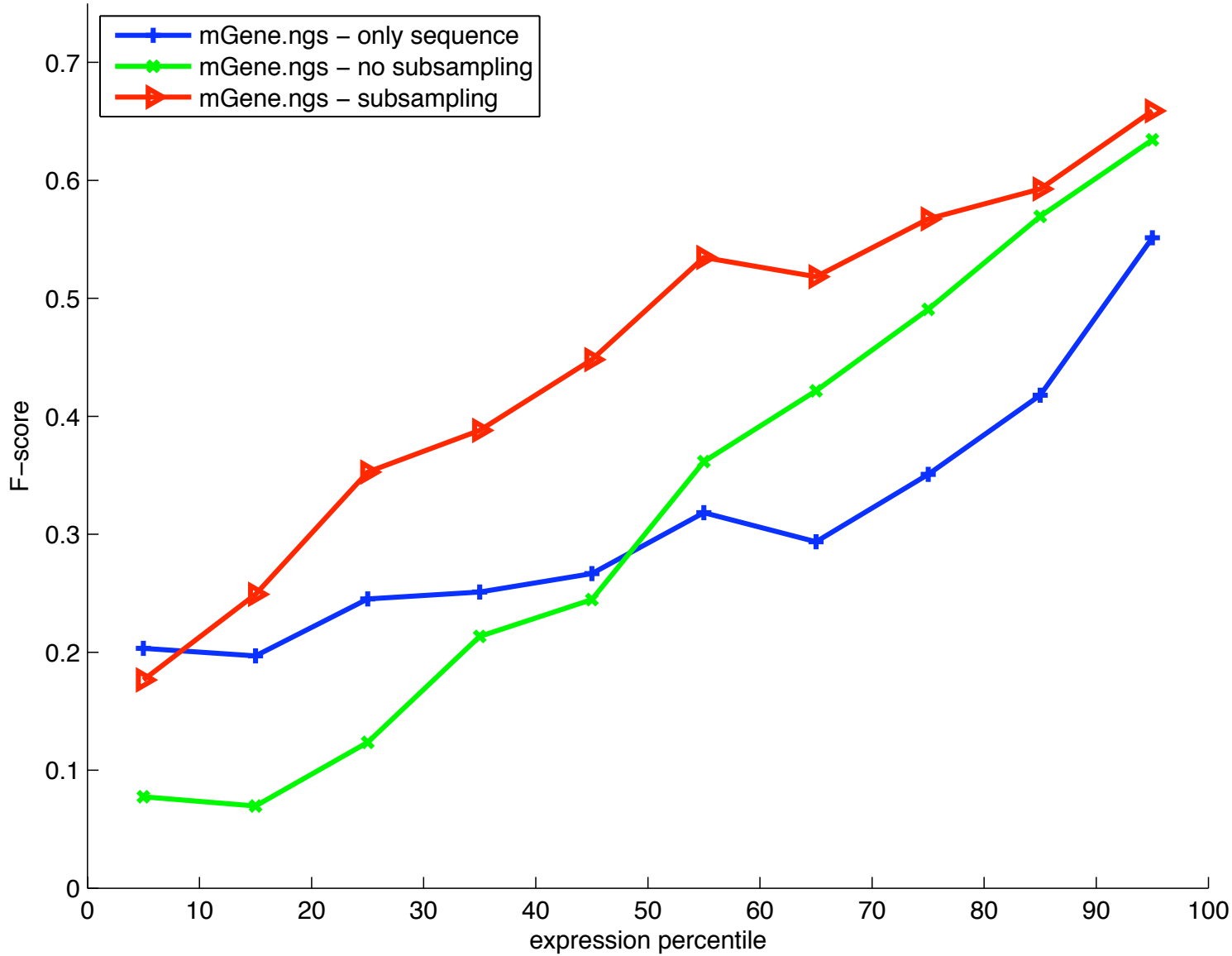
Results on C. elegans



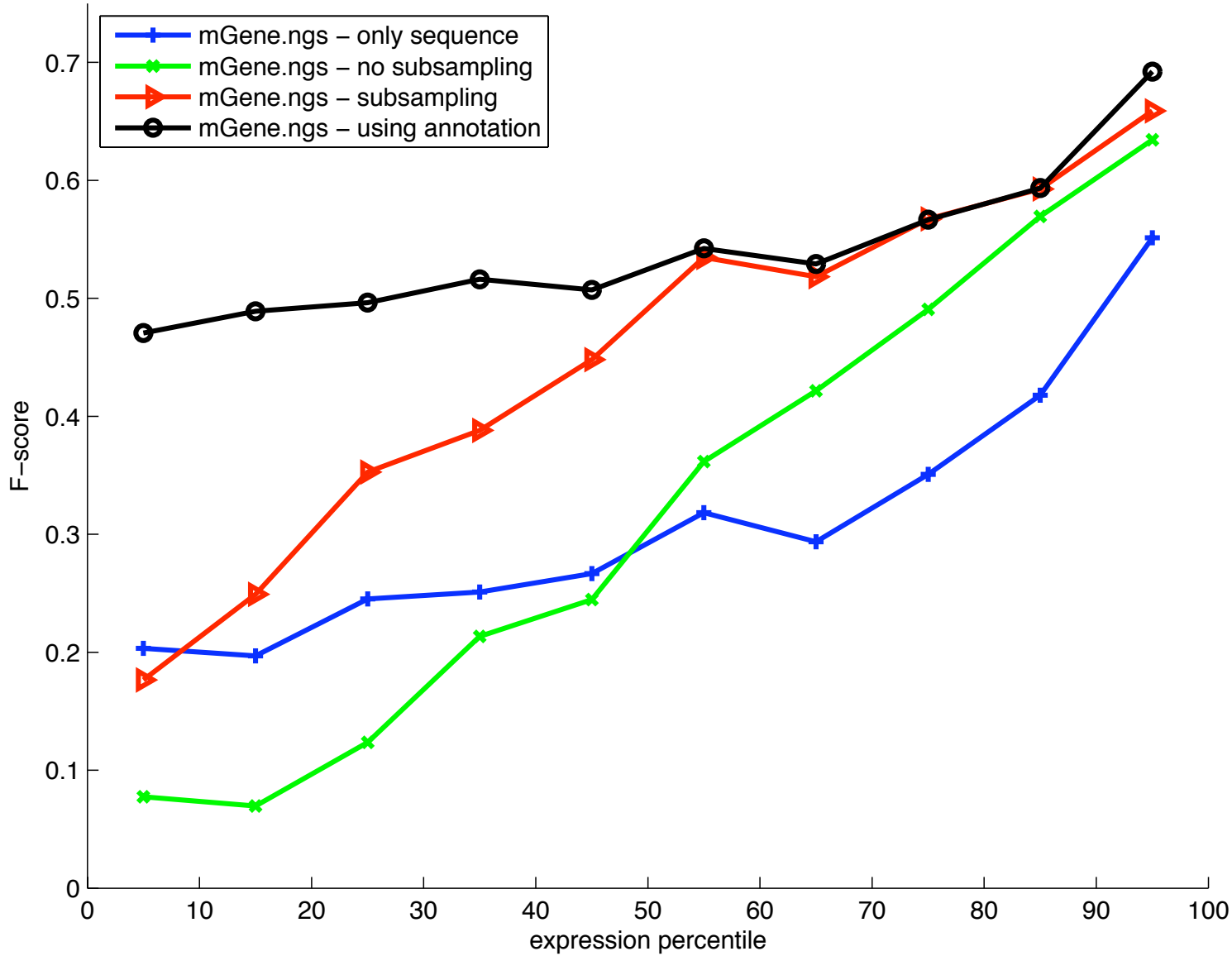
Results on C. elegans



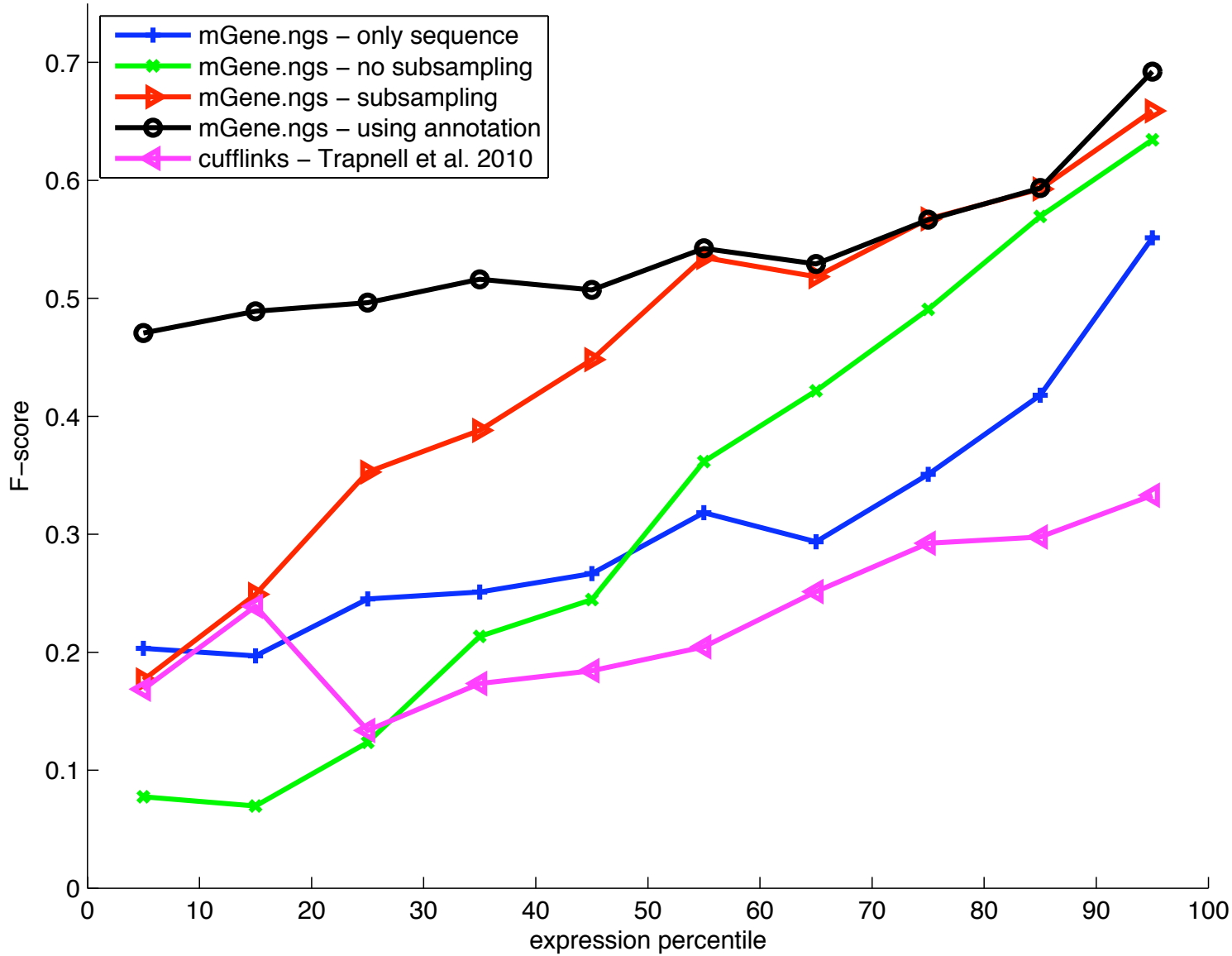
Results on C. elegans



Results on C. elegans



Results on C. elegans



Conclusion

- Transcript identification only from RNA-seq
  $\Rightarrow$ very difficult
- mGene.ngs
  - Highly accurate in a large range of expression levels
  - Can infer alternative isoforms
  - Not relying on previous genome annotation
  - Can benefit from genome annotation if available
  - Scales to mammalian sized genomes



Further information

and the slides of this talk at:

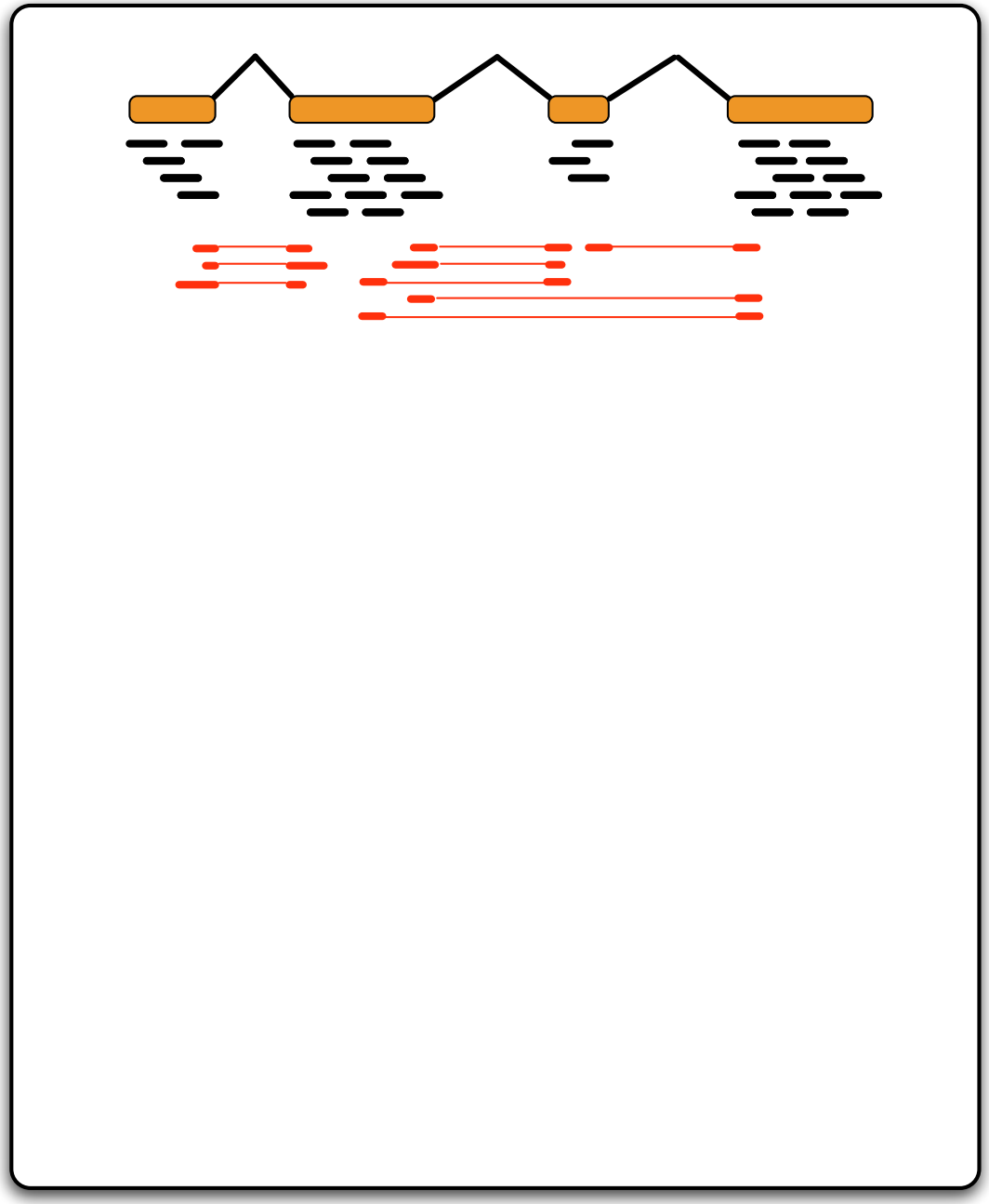
www.mGene.org

Acknowledgments

Quantification:	**Regina Bohnert**
Library preparation:	**Lisa Hartmann, Lisa Smith**
Gene finding:	**Georg Zeller, Gabriele Schweikert, Gunnar Rätsch**
Alignments:	**Andre Kahles, Geraldine Jean, Gunnar Rätsch**
Programming:	**Vipin T Sreedharan**
Shogun:	**Sören Sonnenburg**
Discussions:	**Philipp Drewe, Sebastian Schultheiss, Christian Widmer**
Supervision:	**Gunnar Rätsch**

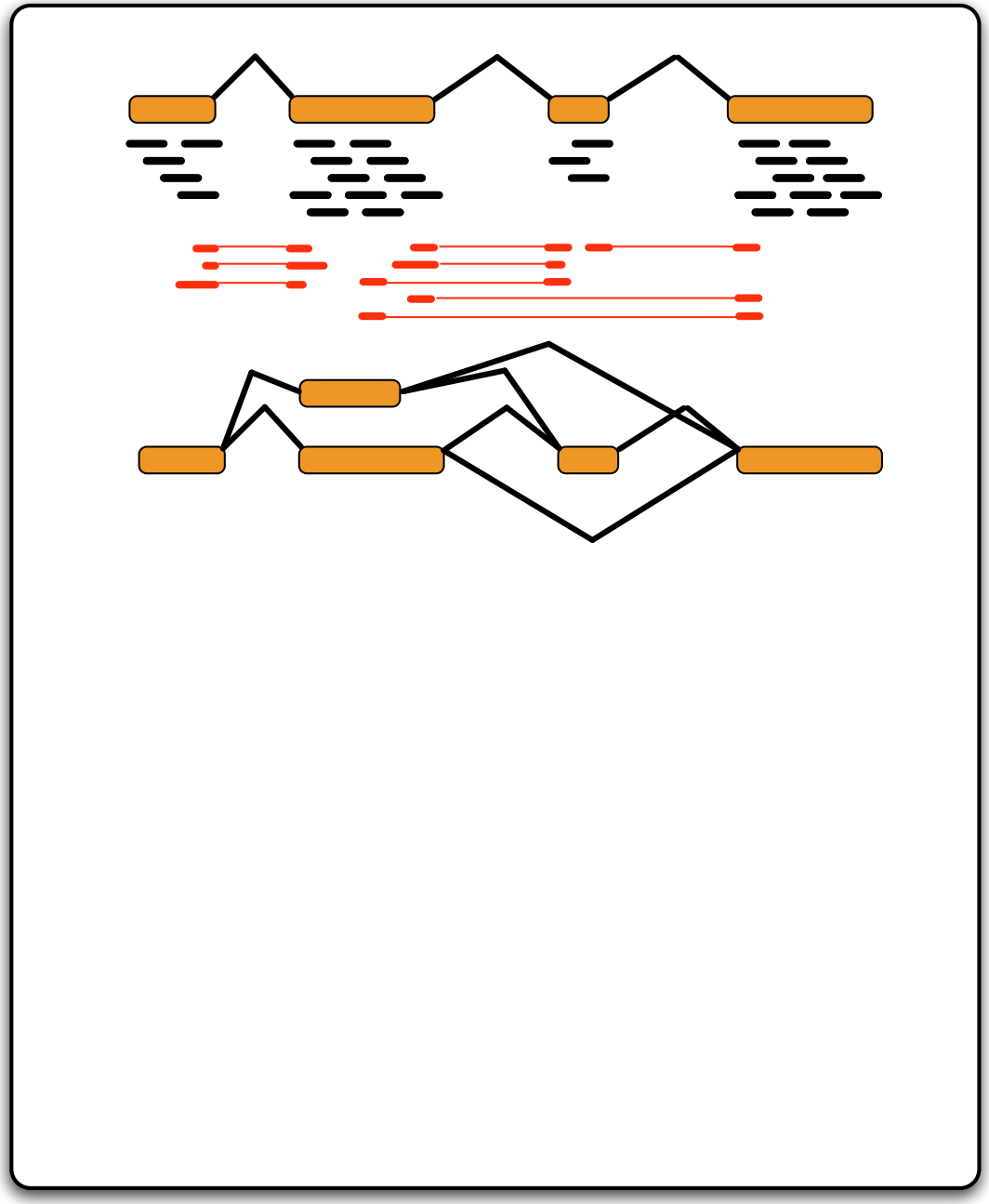
Infer alternative isoforms

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



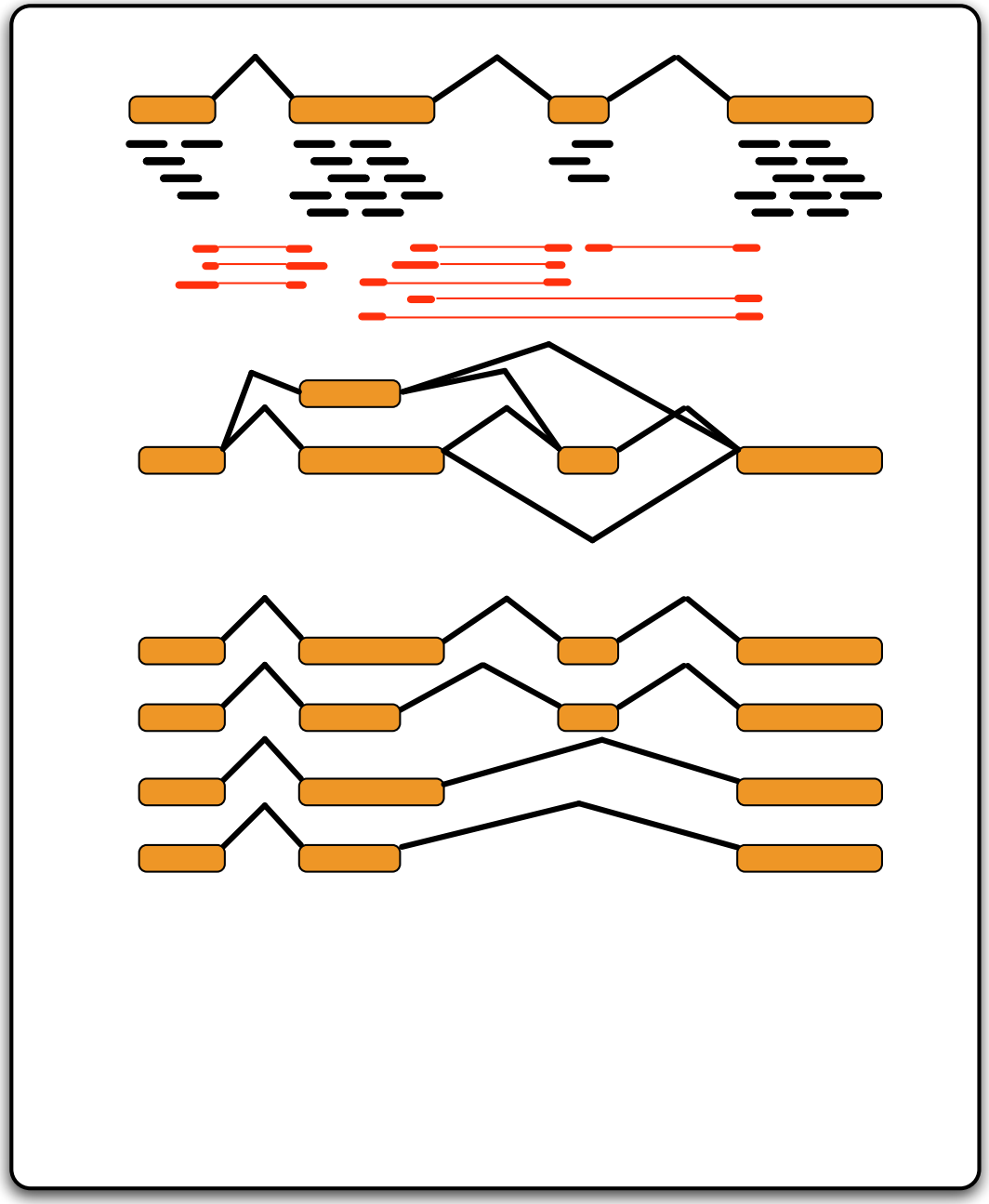
Infer alternative isoforms

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



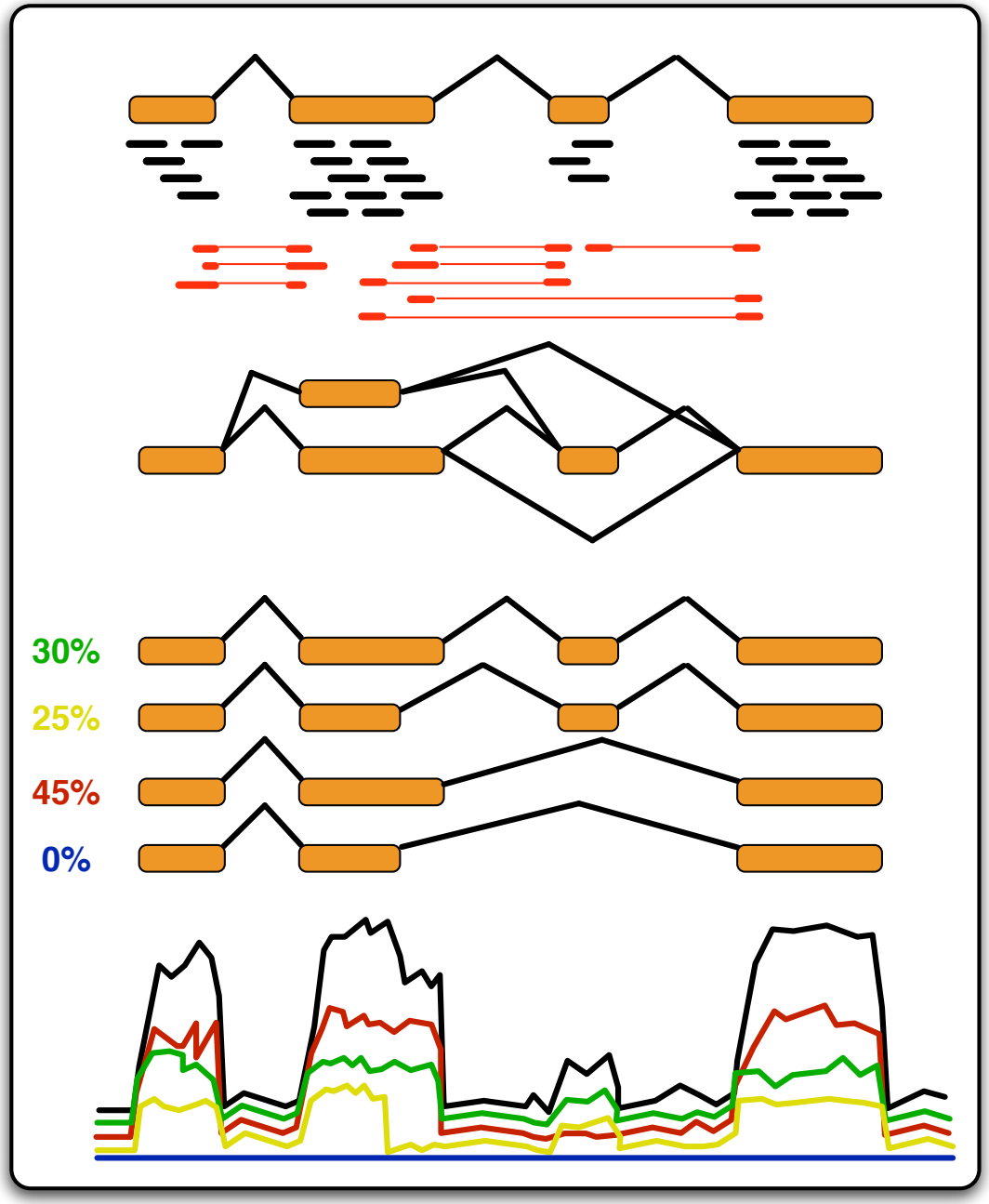
Infer alternative isoforms

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



Infer alternative isoforms

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



References

I

A. Coghlan, T.J. Fiedler, S.J. McKay, P. Flicek, T.W. Harris, D. Blasiar, The nGASP Consortium, and L.D. Stein. nGASP: the nematode genome annotation assessment project. *BMC Bioinformatics*, 2008.

R. Guigó, J.F. Flicek, P. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V.B. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T.R. Gingeras, J. Harrow, T. Hubbard, S.E. Lewis, and M.G. Reese. EGASP: The human ENCODE genome annotation assessment project. *Genome Biology*, 7(S2), 2006.