Fully Automated Genome Annotation with Deep RNA Sequencing

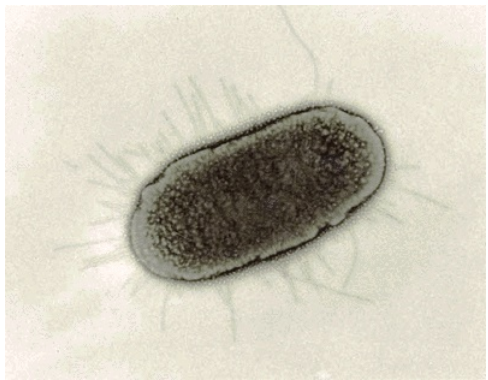
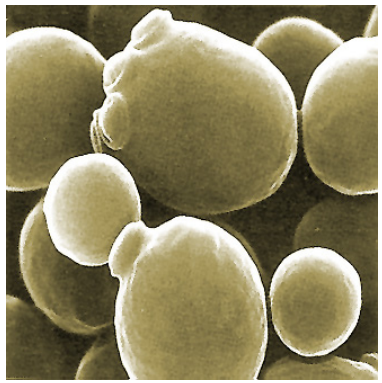
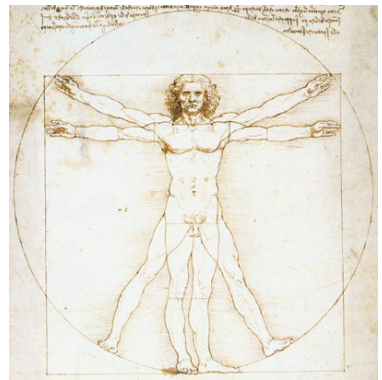
Gunnar Rätsch

Friedrich Miescher Laboratory of the Max Planck Society
Tübingen, Germany

Bio-IT World Europe, October 5, 2010

Sequencing Genomes

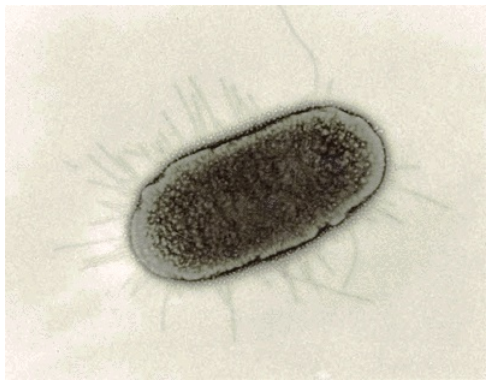
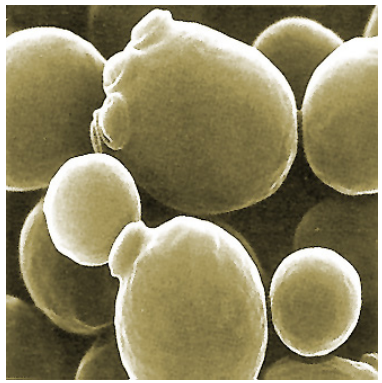
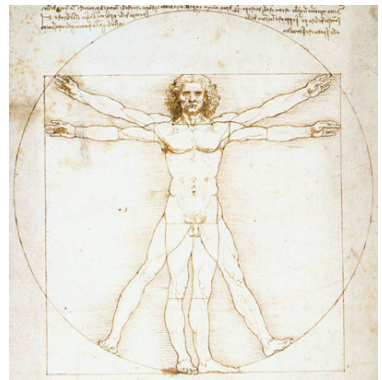
- Large, concerted effort to sequence genomes of model organisms

1997	1997	1998	2000
			
E. coli	*S. cerevisiae*	*C. elegans*	*H. sapiens* **$2.7 billion**

- In 2003, NHGRI committed to develop next-generation sequencing technologies to lower the cost of 30x a human genome (∼100 Gbp):
  - $100,000 genome
  - $1,000 genome

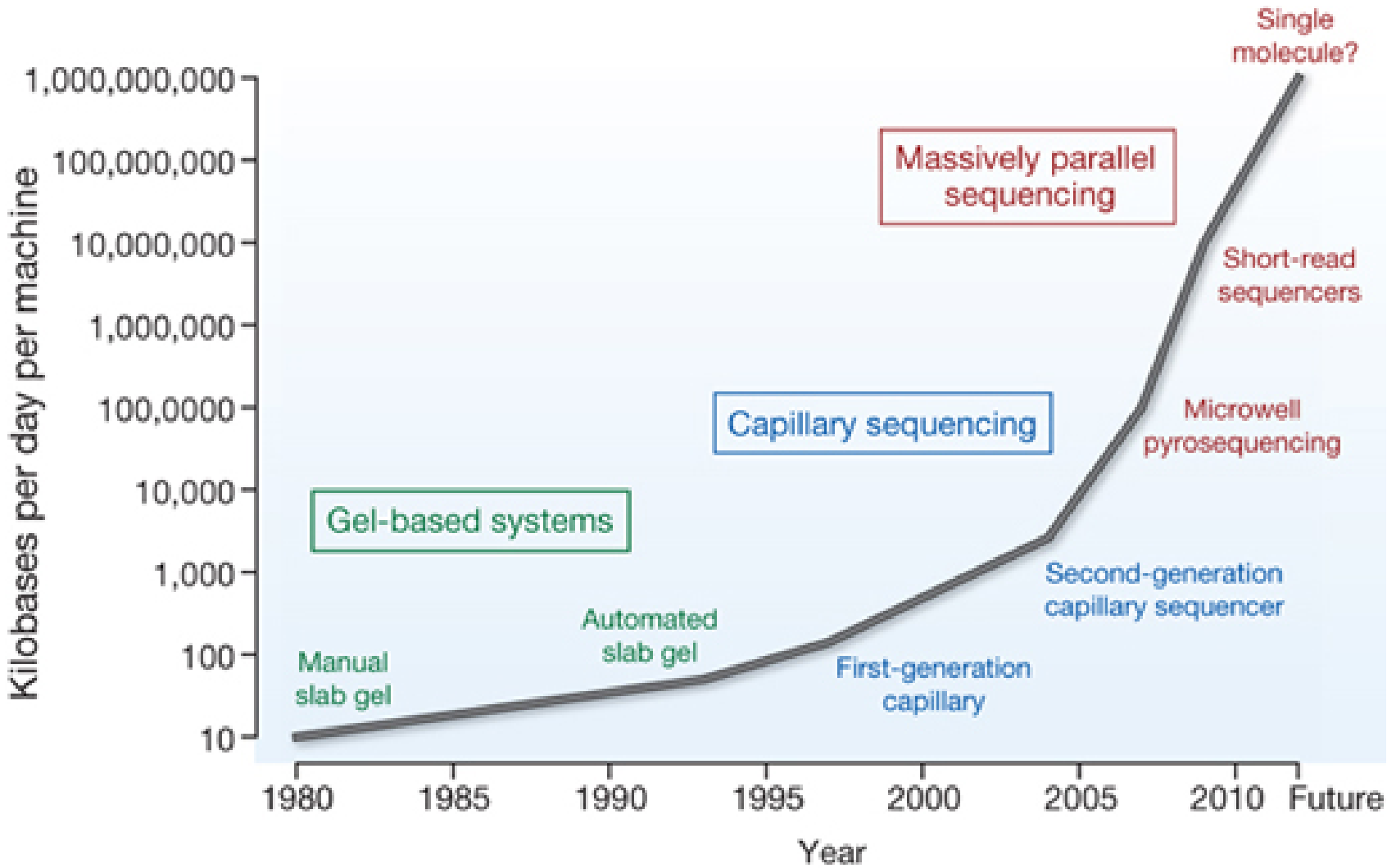
Sequencing Genomes

- Large, concerted effort to sequence genomes of model organisms

1997	1997	1998	2000
			
E. coli	*S. cerevisiae*	*C. elegans*	*H. sapiens* **$2.7 billion**

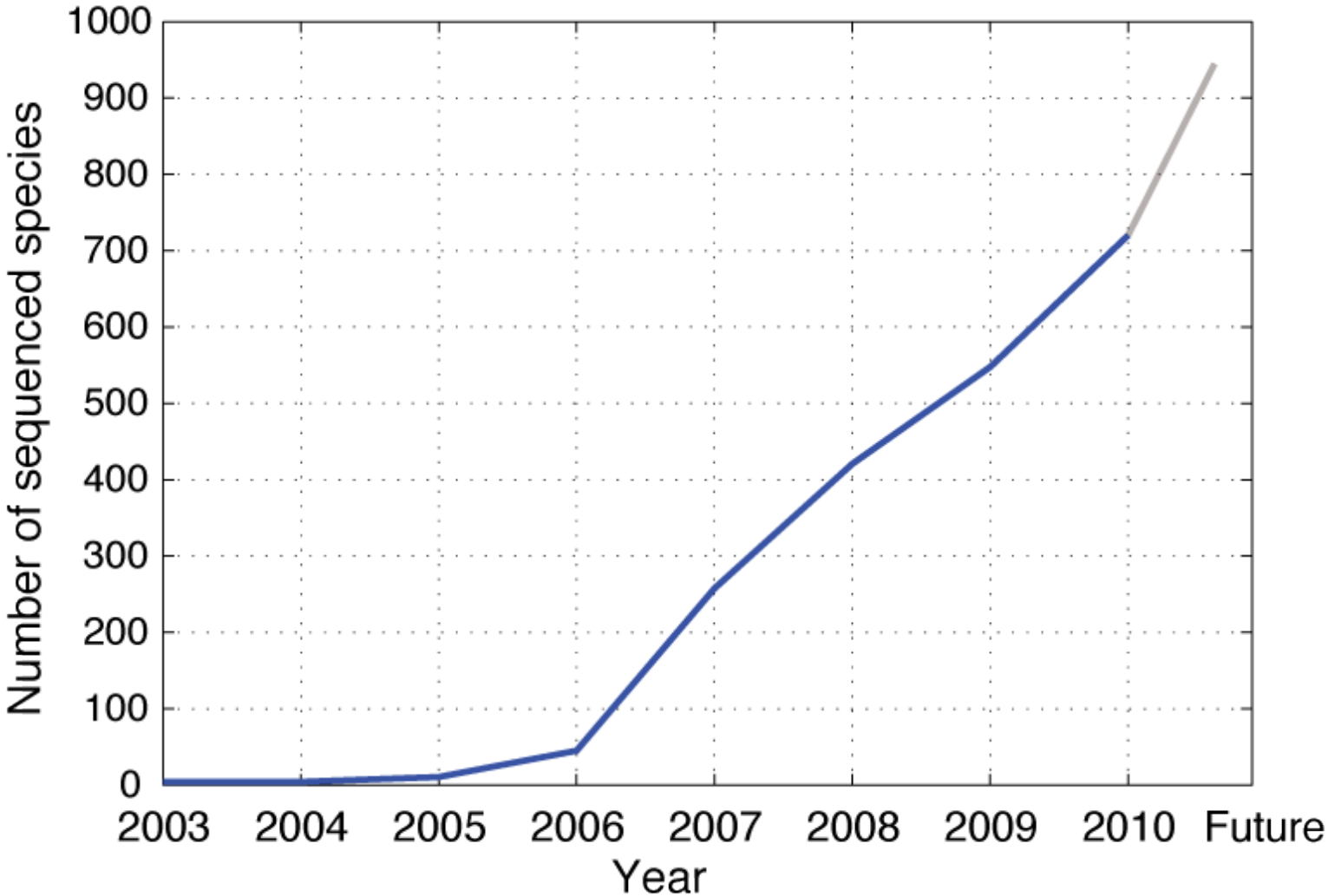
- In 2003, NHGRI committed to develop next-generation sequencing technologies to lower the cost of 30x a human genome (∼100 Gbp):
  - $100,000 genome
  - $1,000 genome

Rate of Sequencing



[Nature 458, 719-724 (2009)]

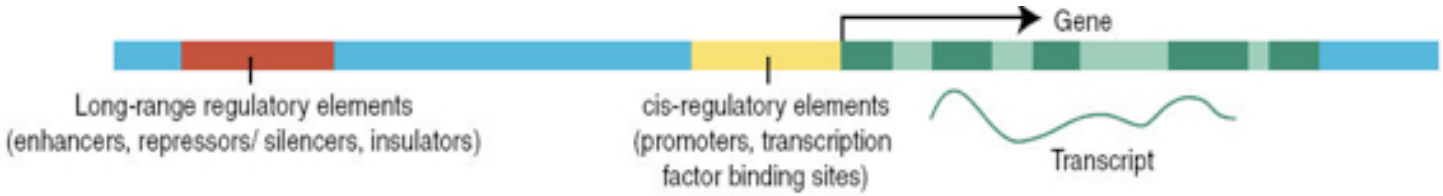
Rate of Sequencing



[Generated from NCBI on Sep 30, 2010]

What does it all mean?

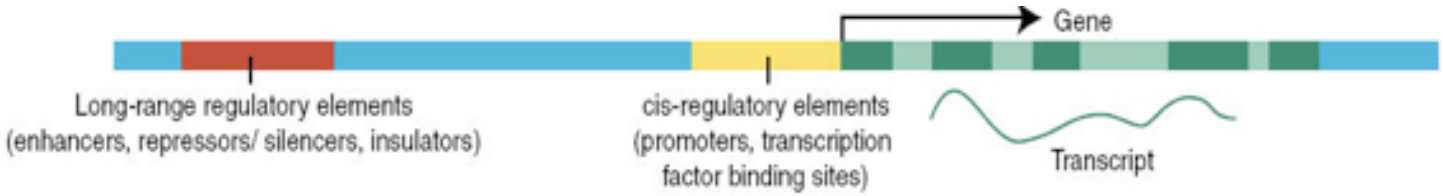
- How to make sense of the genome?
  - Which nucleotides are functional?
  - What is their function?



- Complex process of genome annotation
  - Computational gene prediction & manual curation
  - Based on homology
  - Done by large sequencing centers
- Problems:
  - Most accurate annotations only for model organisms
  - Few gene discoveries, inaccurate annotation for new genes
  - Annotation strategies too inefficient for the many genomes . . .

What does it all mean?

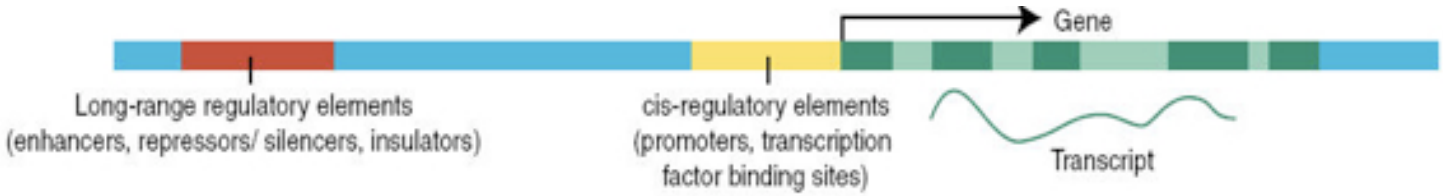
- How to make sense of the genome?
  - Which nucleotides are functional?
  - What is their function?



- Complex process of genome annotation
  - Computational gene prediction & manual curation
  - Based on homology
  - Done by large sequencing centers
- Problems:
  - Most accurate annotations only for model organisms
  - Few gene discoveries, inaccurate annotation for new genes
  - Annotation strategies too inefficient for the many genomes . . .

What does it all mean?

- How to make sense of the genome?
  - Which nucleotides are functional?
  - What is their function?



- Complex process of genome annotation
  - Computational gene prediction & manual curation
  - Based on homology
  - Done by large sequencing centers
- Problems:
  - Most accurate annotations only for model organisms
  - Few gene discoveries, inaccurate annotation for new genes
  - Annotation strategies too inefficient for the many genomes . . .

Automated Genome Annotation

What is needed?

1. Highly accurate annotation of genomes
   - Non-coding RNAs
   - Alternative transcripts
2. With as little as possible prior knowledge
   - Unbiased approach to allow new discovery
3. Fully automated such that everybody can do it him/herself

Steps to get there:

- Deep RNA sequencing
- Improved analysis methods
- Easy to use software/services

Automated Genome Annotation

What is needed?

1. Highly accurate annotation of genomes
   - Non-coding RNAs
   - Alternative transcripts
2. With as little as possible prior knowledge
   - Unbiased approach to allow new discovery
3. Fully automated such that everybody can do it him/herself

Steps to get there:

- Deep RNA sequencing
- Improved analysis methods
- Easy to use software/services

Automated Genome Annotation

What is needed?

1. Highly accurate annotation of genomes
   - Non-coding RNAs
   - Alternative transcripts
2. With as little as possible prior knowledge
   - Unbiased approach to allow new discovery
3. Fully automated such that everybody can do it him/herself

Steps to get there:

- Deep RNA sequencing
- Improved analysis methods
- Easy to use software/services

Automated Genome Annotation

What is needed?

1. Highly accurate annotation of genomes
   - Non-coding RNAs
   - Alternative transcripts
2. With as little as possible prior knowledge
   - Unbiased approach to allow new discovery
3. Fully automated such that everybody can do it him/herself

Steps to get there:

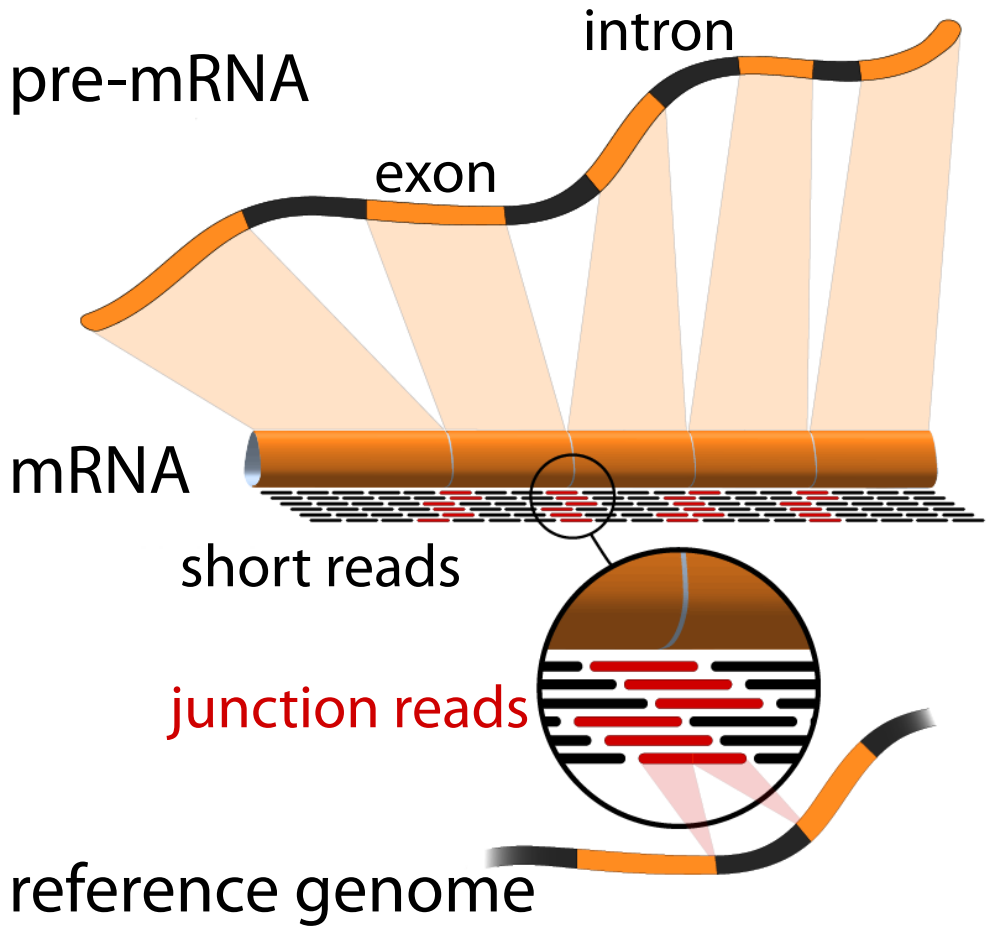
- Deep RNA sequencing
- Improved analysis methods
- Easy to use software/services

Deep RNA Sequencing (RNA-Seq)

fml

RNA-Seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
  - New transcripts
  - Improved gene models
- Quantitative studies at high resolution
  - Differential expression in tissues, conditions, genotypes, etc.



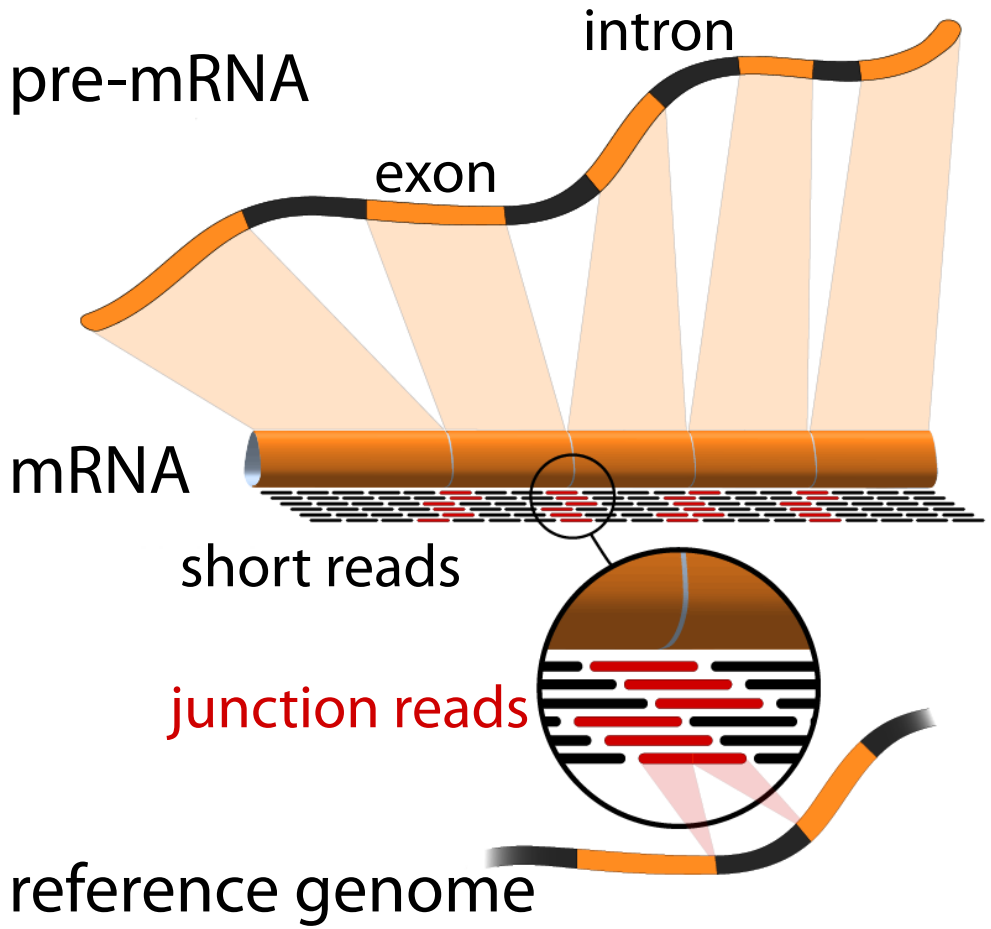
[Adapted from Wikipedia]

Deep RNA Sequencing (RNA-Seq)

fml

RNA-Seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
  - New transcripts
  - Improved gene models
- Quantitative studies at high resolution
  - Differential expression in tissues, conditions, genotypes, etc.

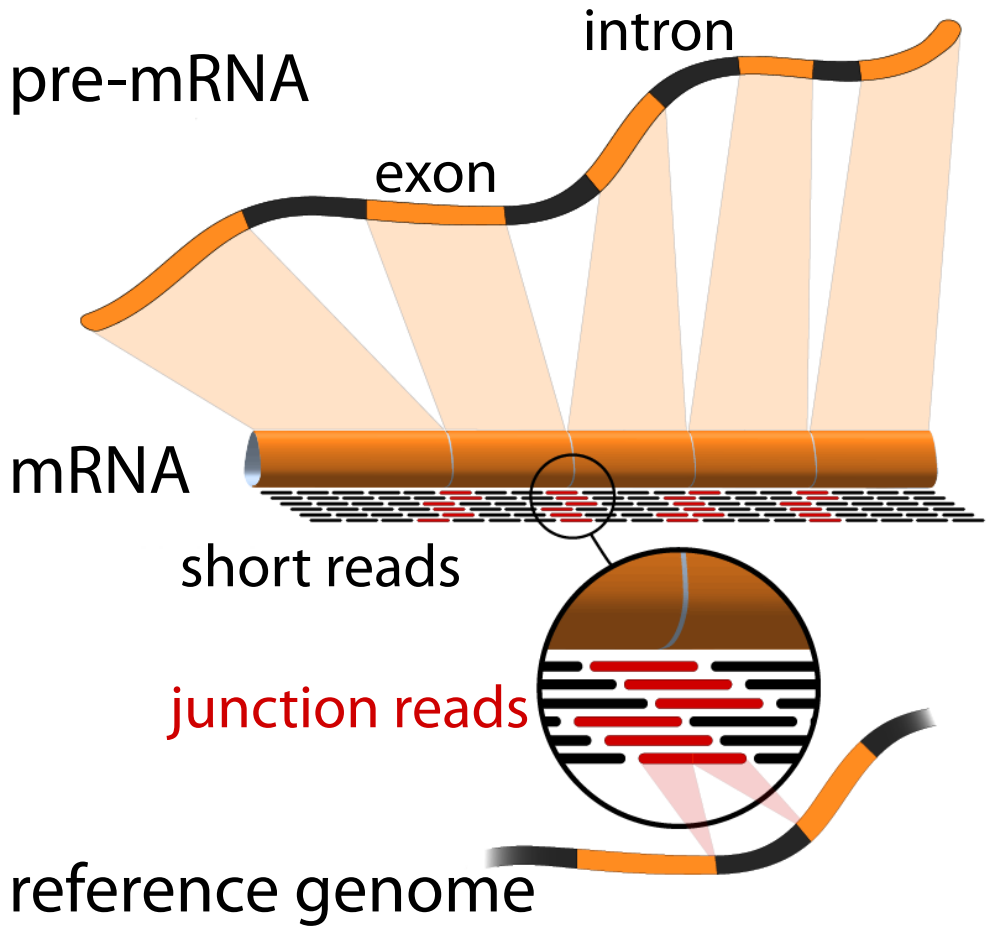


[Adapted from Wikipedia]

Deep RNA Sequencing (RNA-Seq)

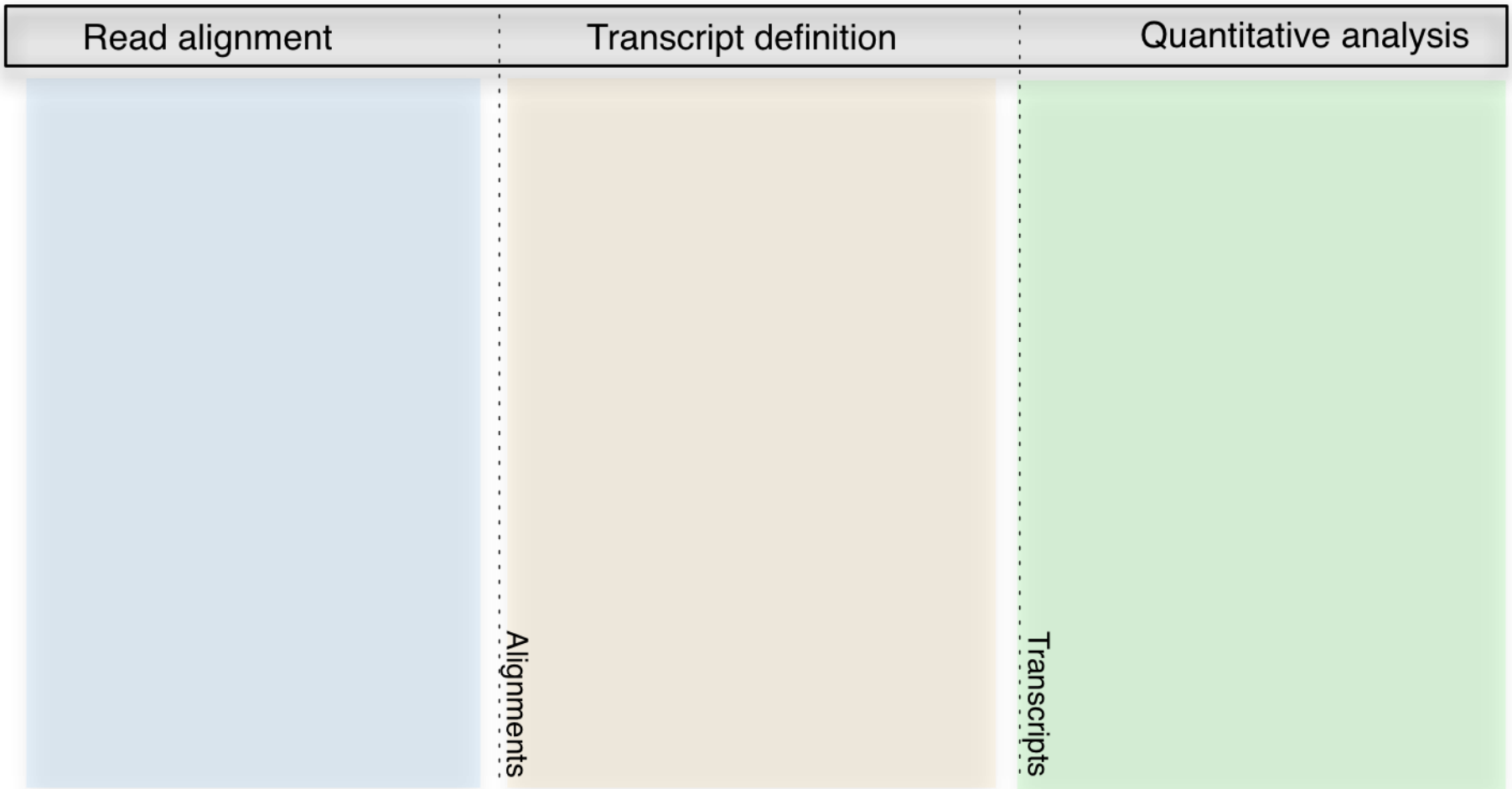
RNA-Seq allows ...

- High-throughput transcriptome measurements
- Qualitative studies
  - New transcripts
  - Improved gene models
- Quantitative studies at high resolution
  - Differential expression in tissues, conditions, genotypes, etc.

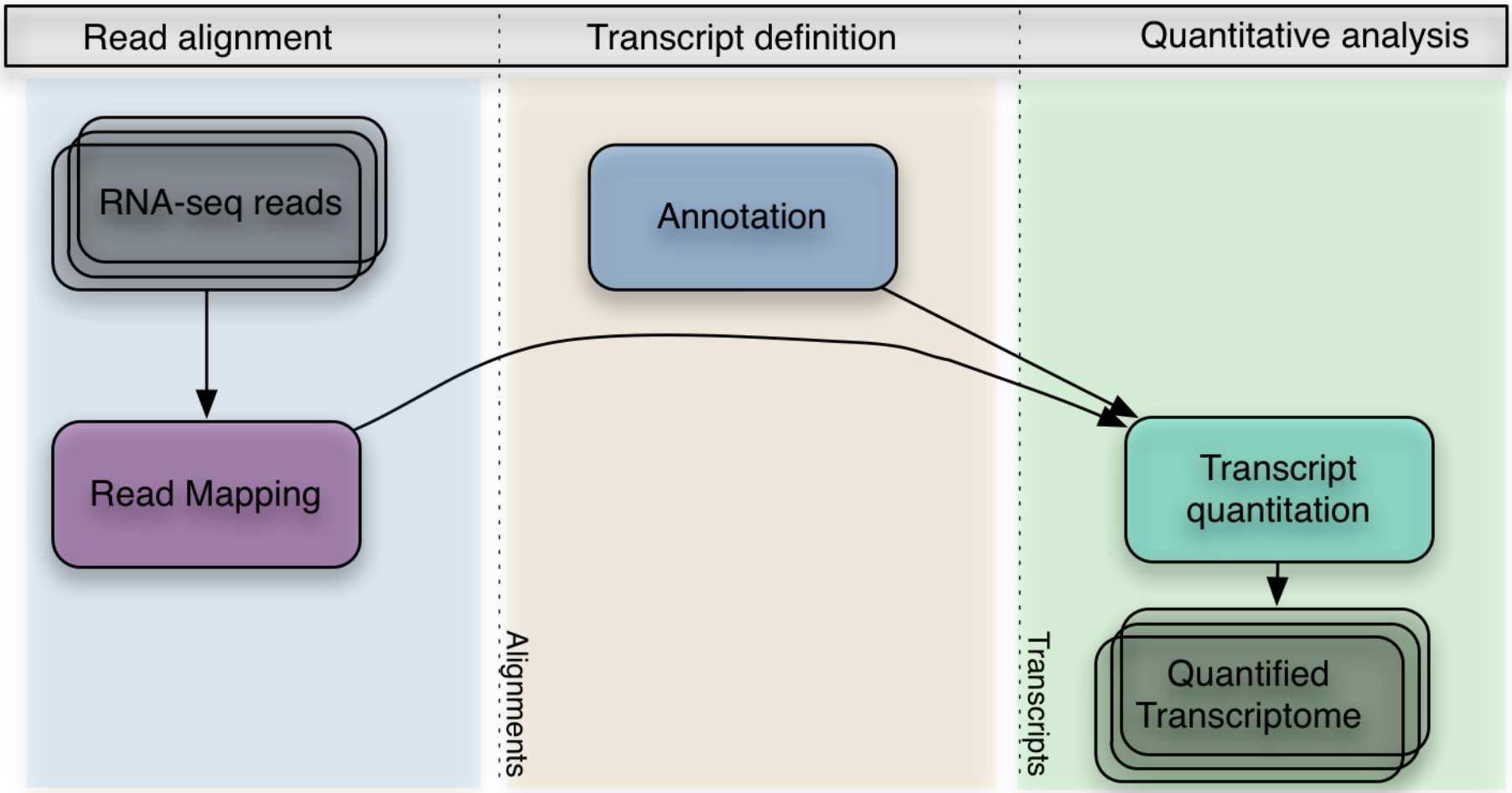


[Adapted from Wikipedia]

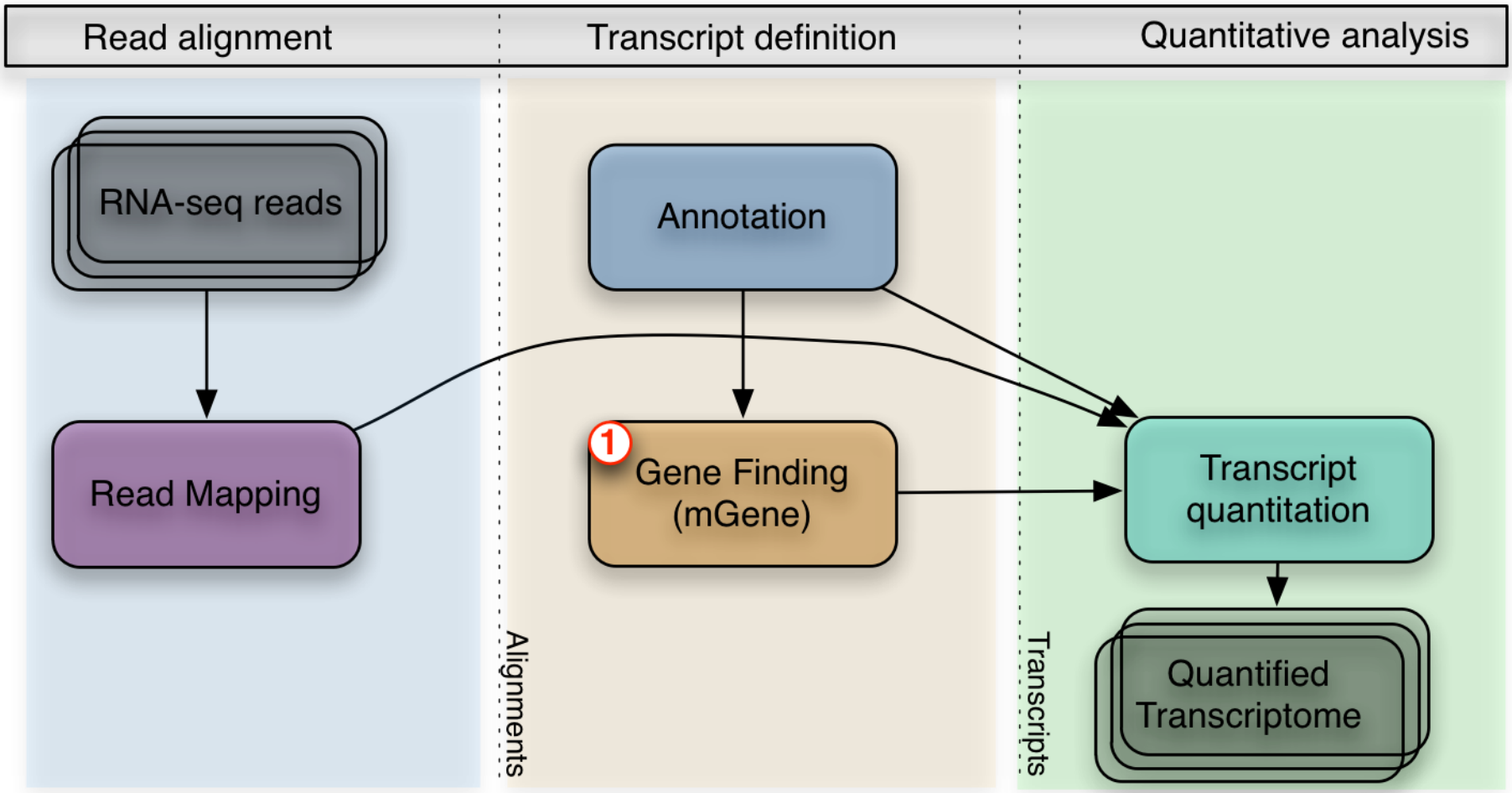
RNA-Seq Analysis Pipeline(s)



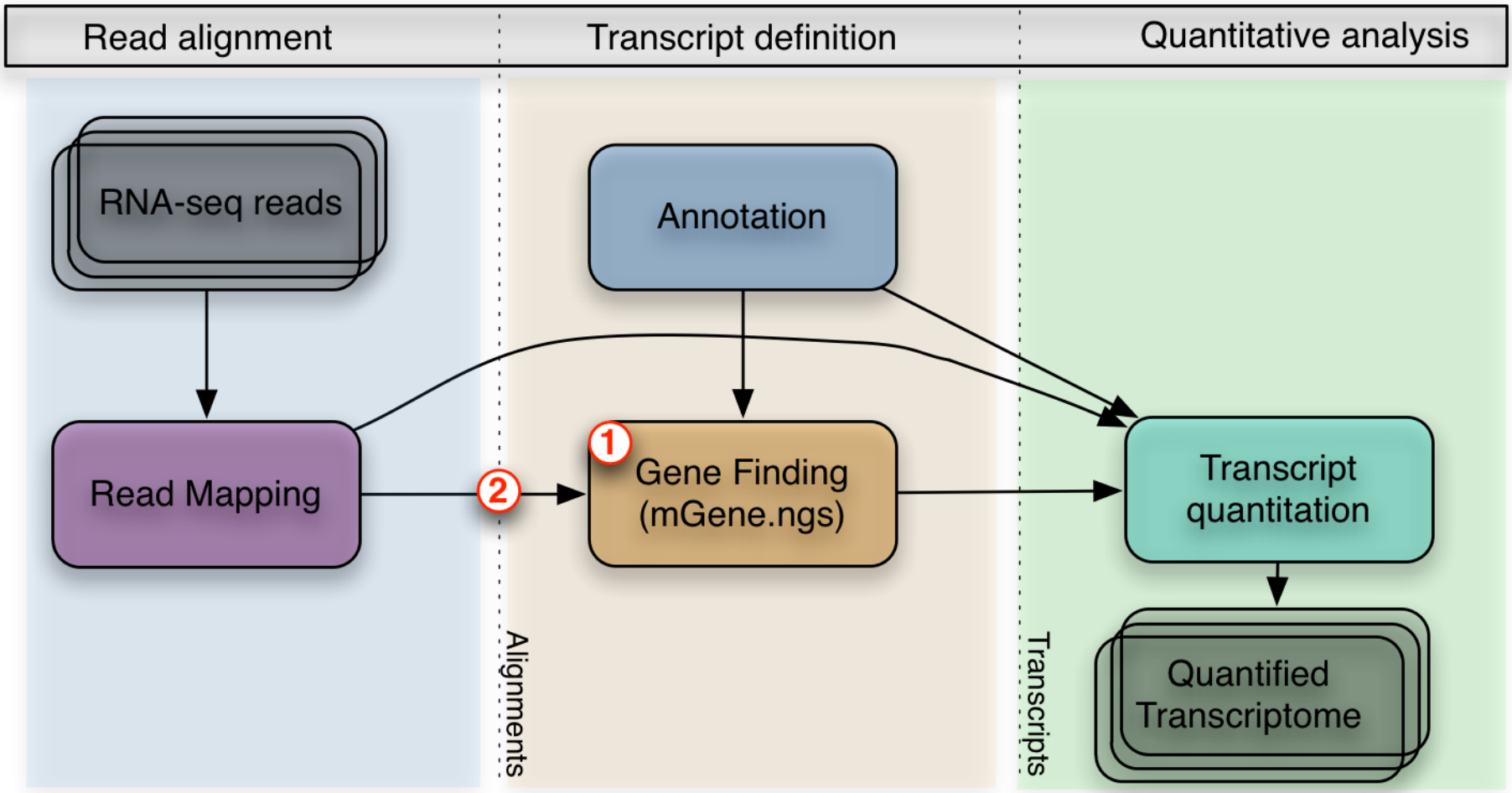
RNA-Seq Analysis Pipeline(s)



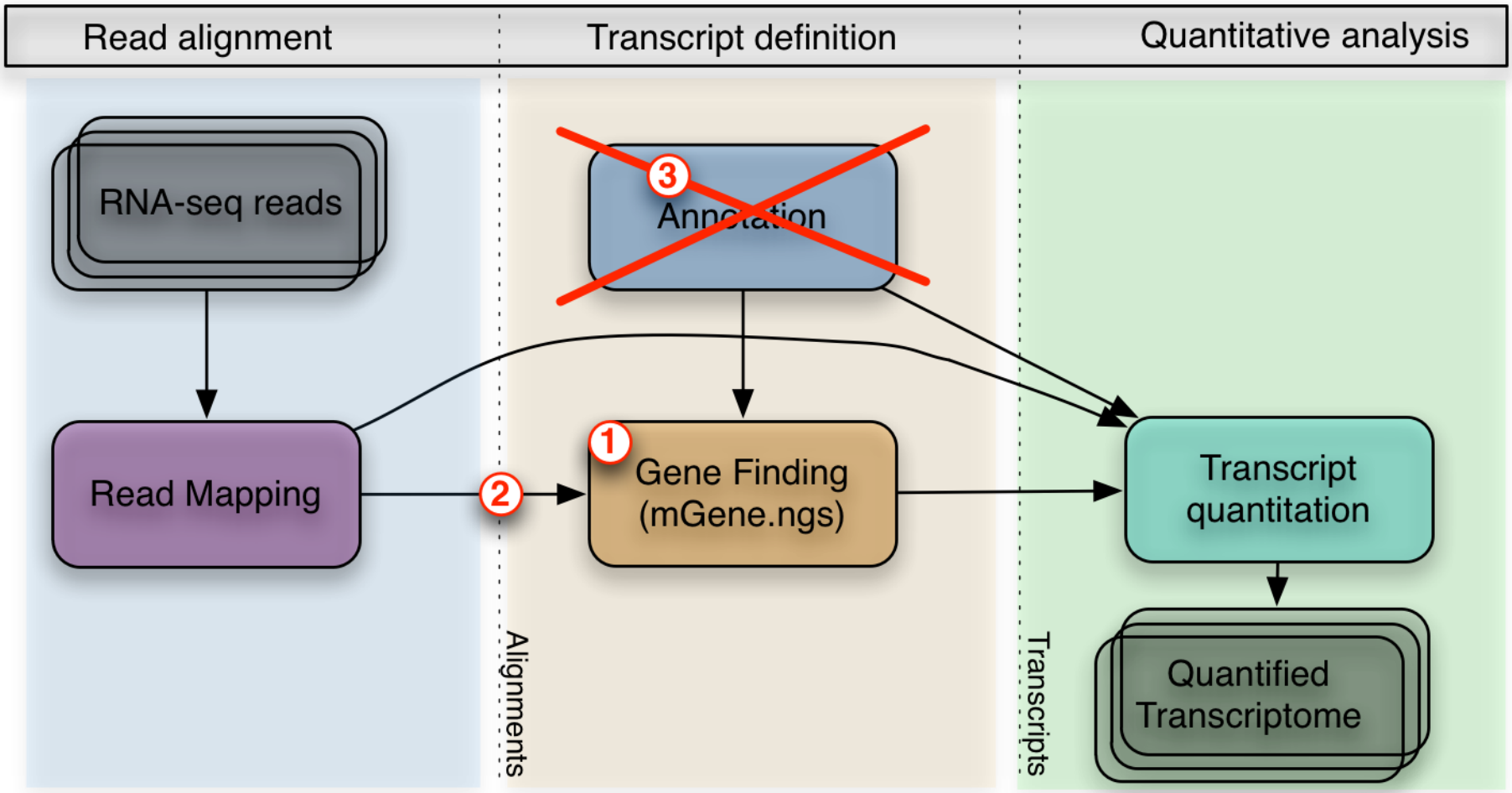
RNA-Seq Analysis Pipeline(s)



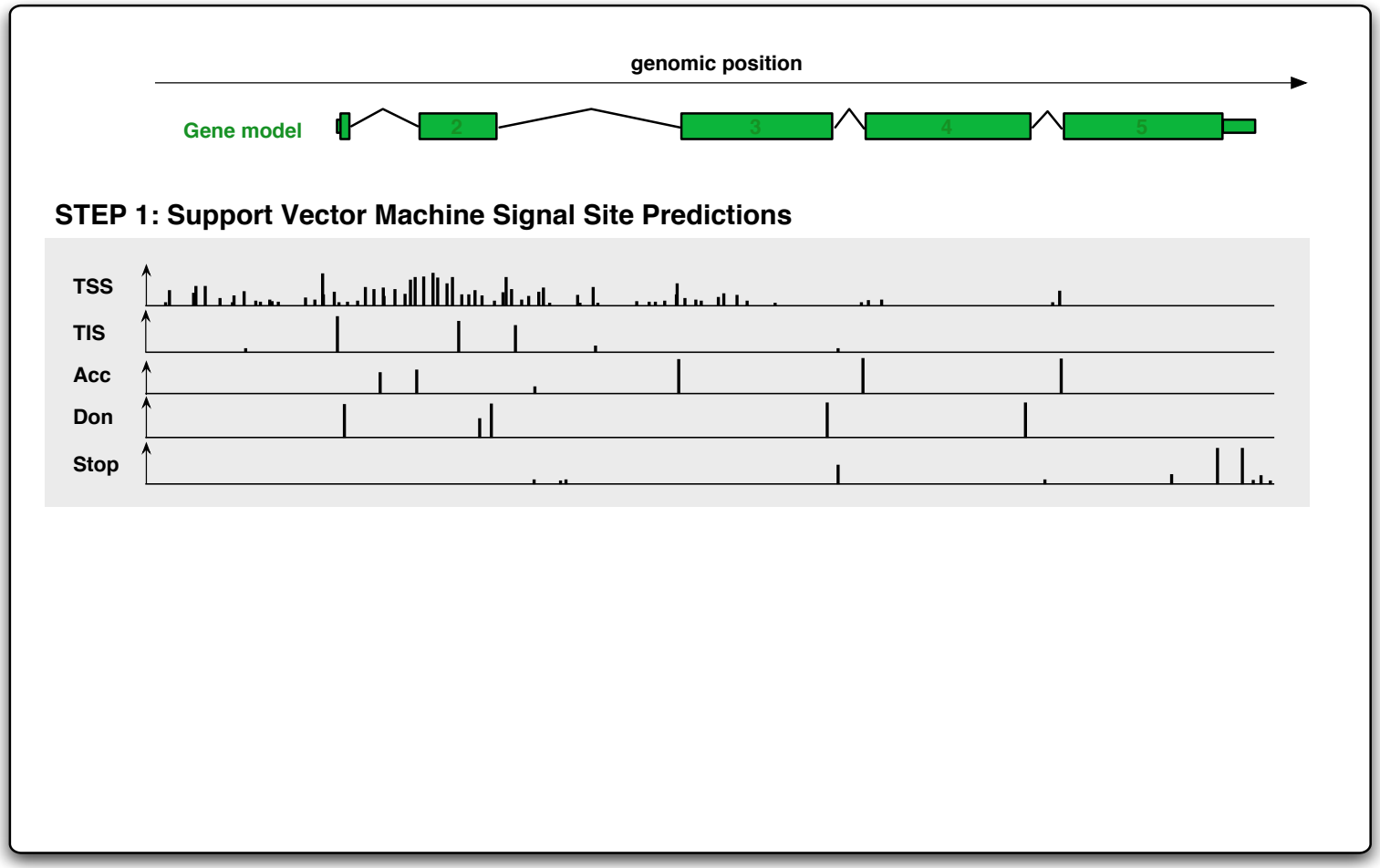
RNA-Seq Analysis Pipeline(s)



RNA-Seq Analysis Pipeline(s)

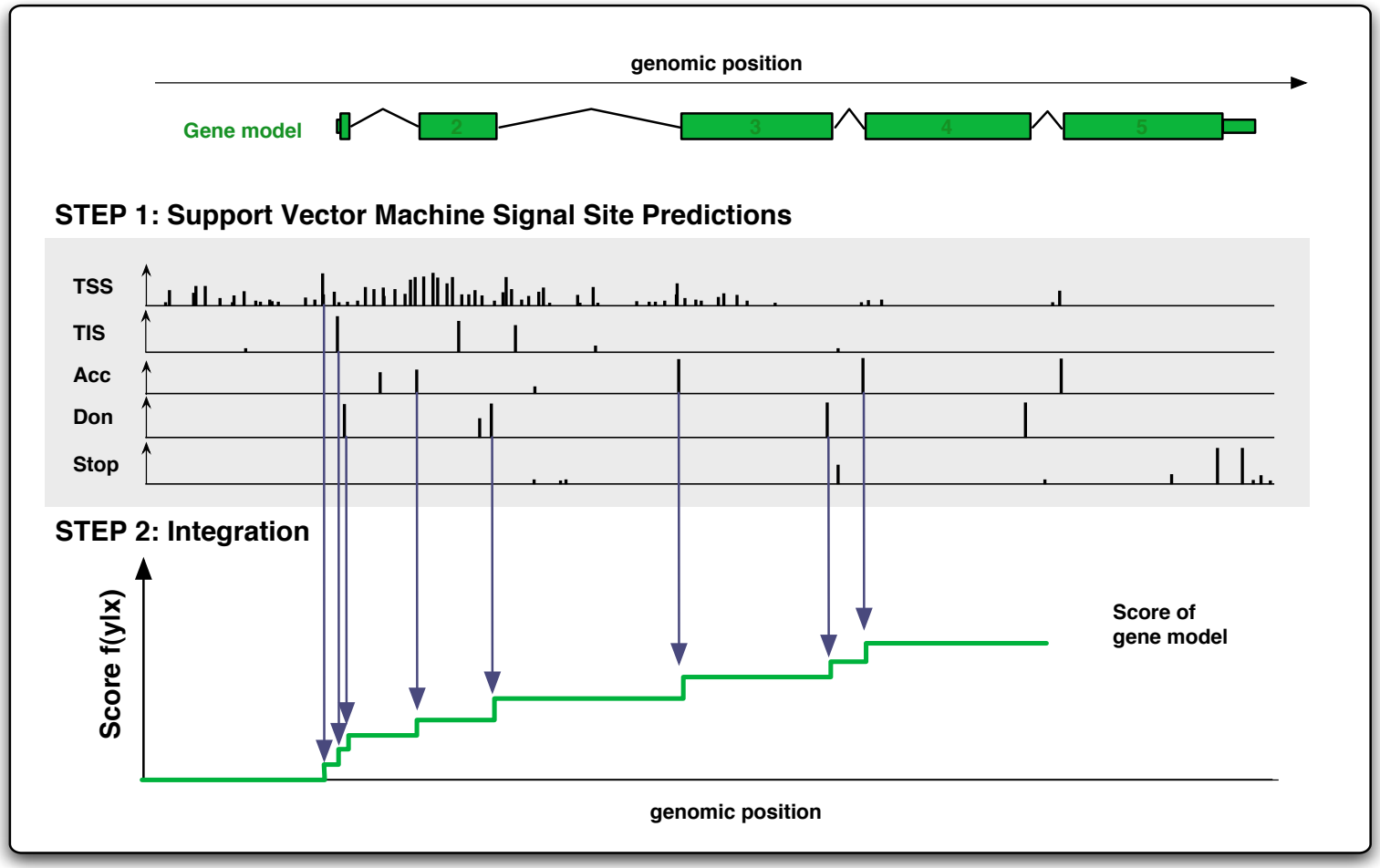


Step 1: Novel Gene Prediction Methods



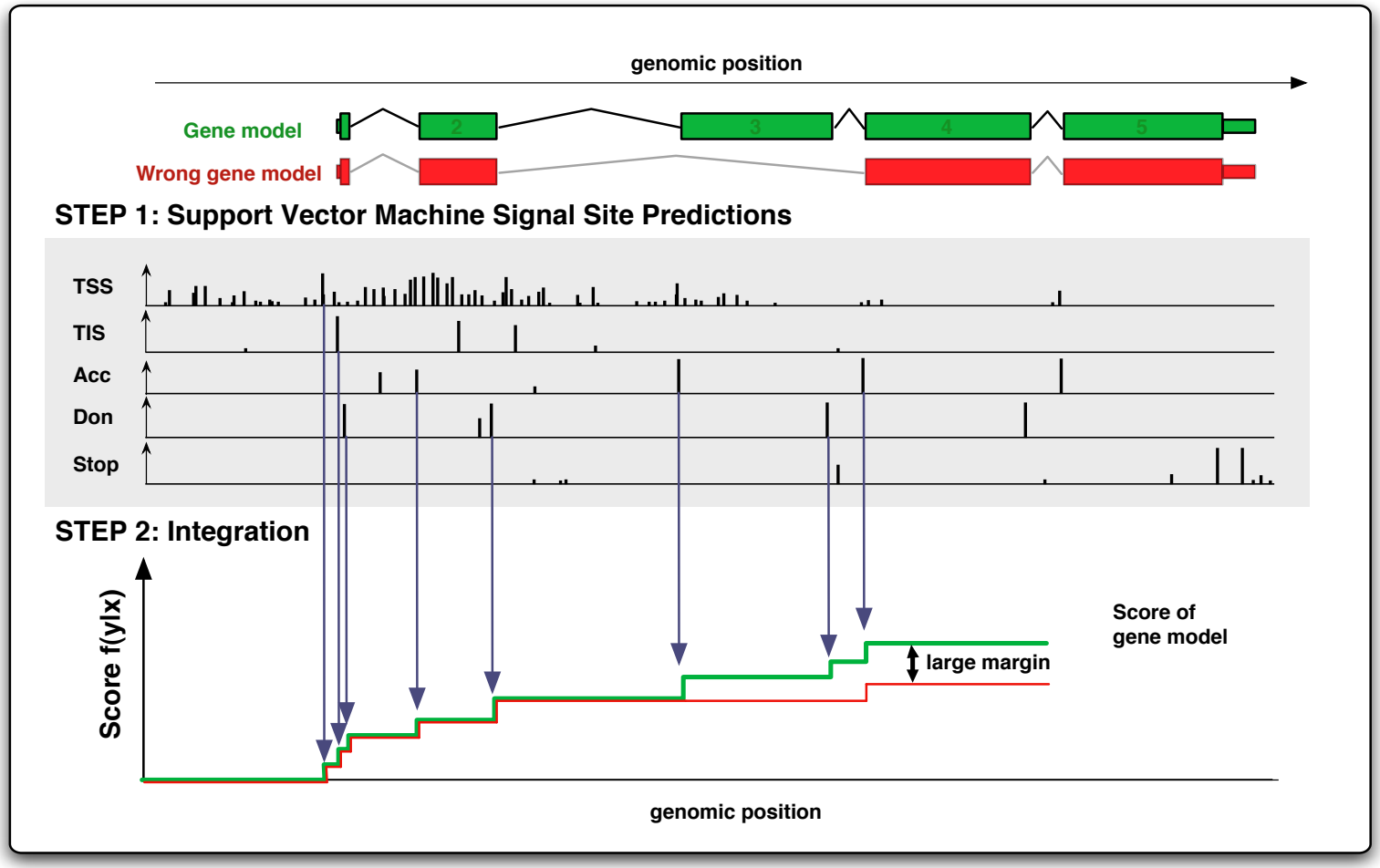
- Accurate signal site predictions using Support Vector Machines
- Novel discriminative learning techniques for data integration

Step 1: Novel Gene Prediction Methods



- Accurate signal site predictions using Support Vector Machines
- Novel discriminative learning techniques for data integration

Step 1: Novel Gene Prediction Methods



- Accurate signal site predictions using Support Vector Machines
- Novel discriminative learning techniques for data integration

Results using **mGene**

(Schweikert et al., Genome Research, 2009)

- Most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [2]
- Validation of novel gene predictions for *C. elegans*: [7]

	No. of genes	No. of genes analyzed	Frac. of genes w/ expression
New genes	2,197	57	$\approx 42\%$
Missing unconf. genes	205	24	$\approx 8\%$

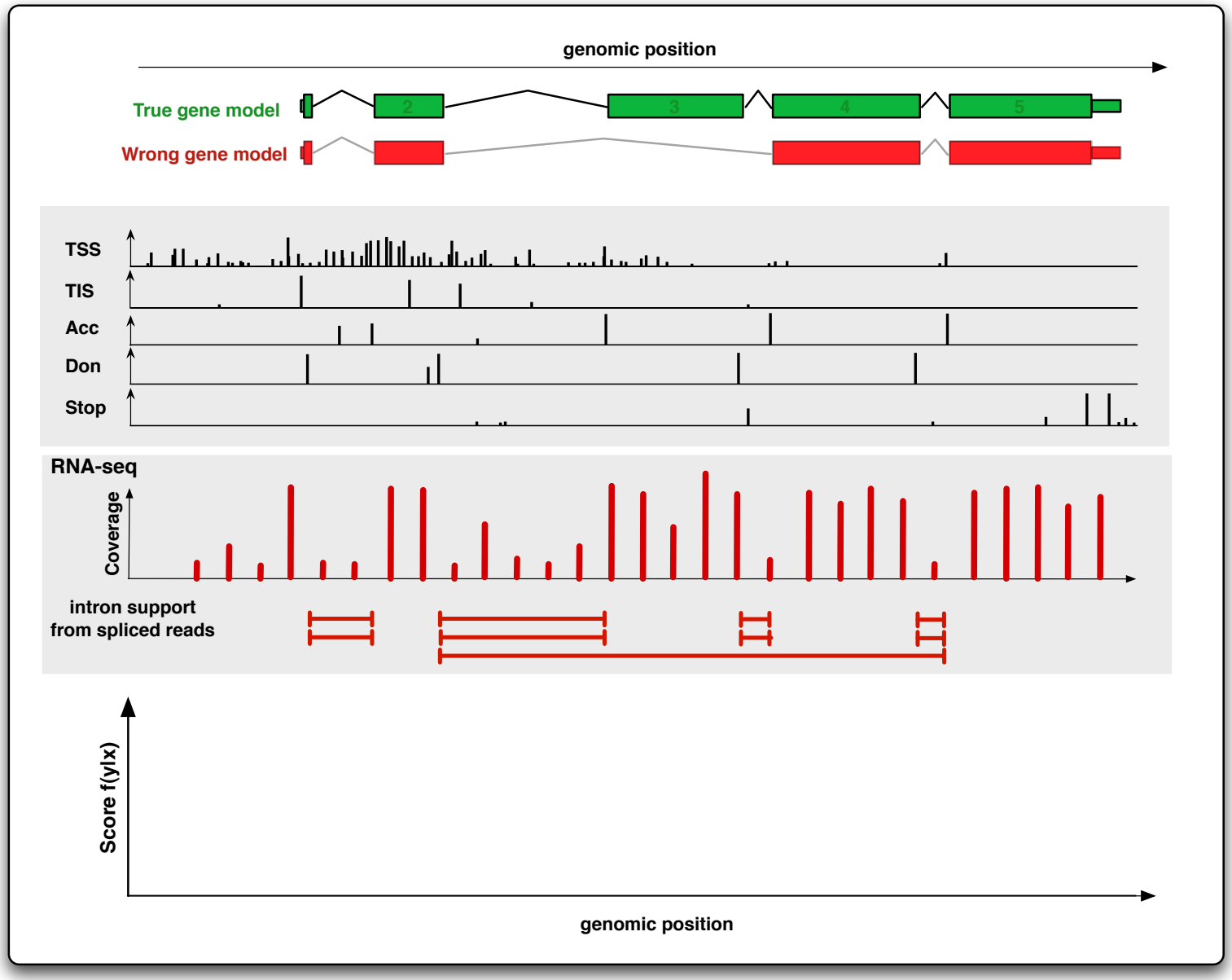
Results using **mGene**

(Schweikert et al., Genome Research, 2009)

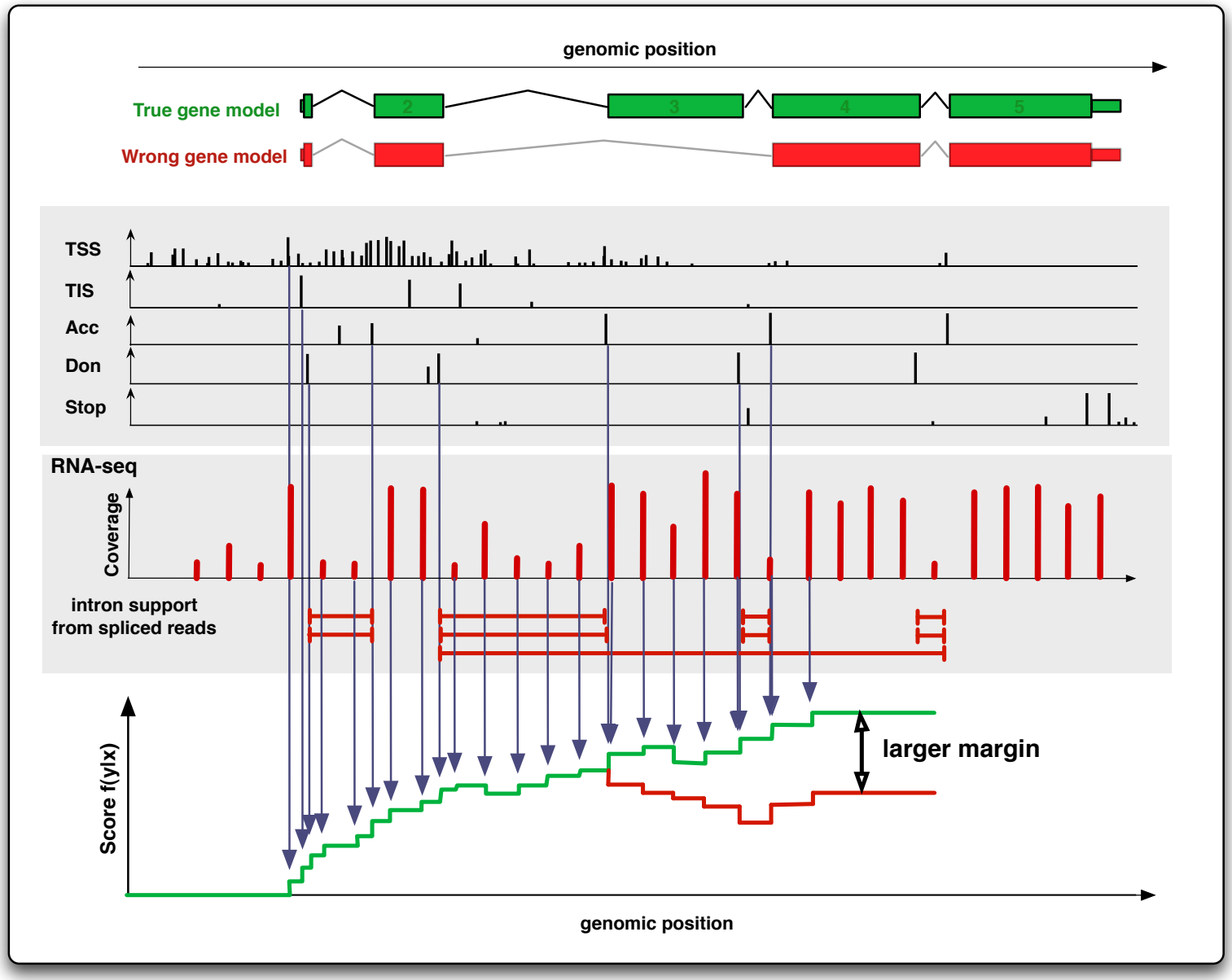
- Most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [2]
- Validation of novel gene predictions for *C. elegans*: [7]

	No. of genes	No. of genes analyzed	Frac. of genes w/ expression
New genes	2,197	57	$\approx 42\%$
Missing unconf. genes	205	24	$\approx 8\%$

Step 2: Integrating RNA-Seq Information



Step 2: Integrating RNA-Seq Information



Results for *C. elegans*

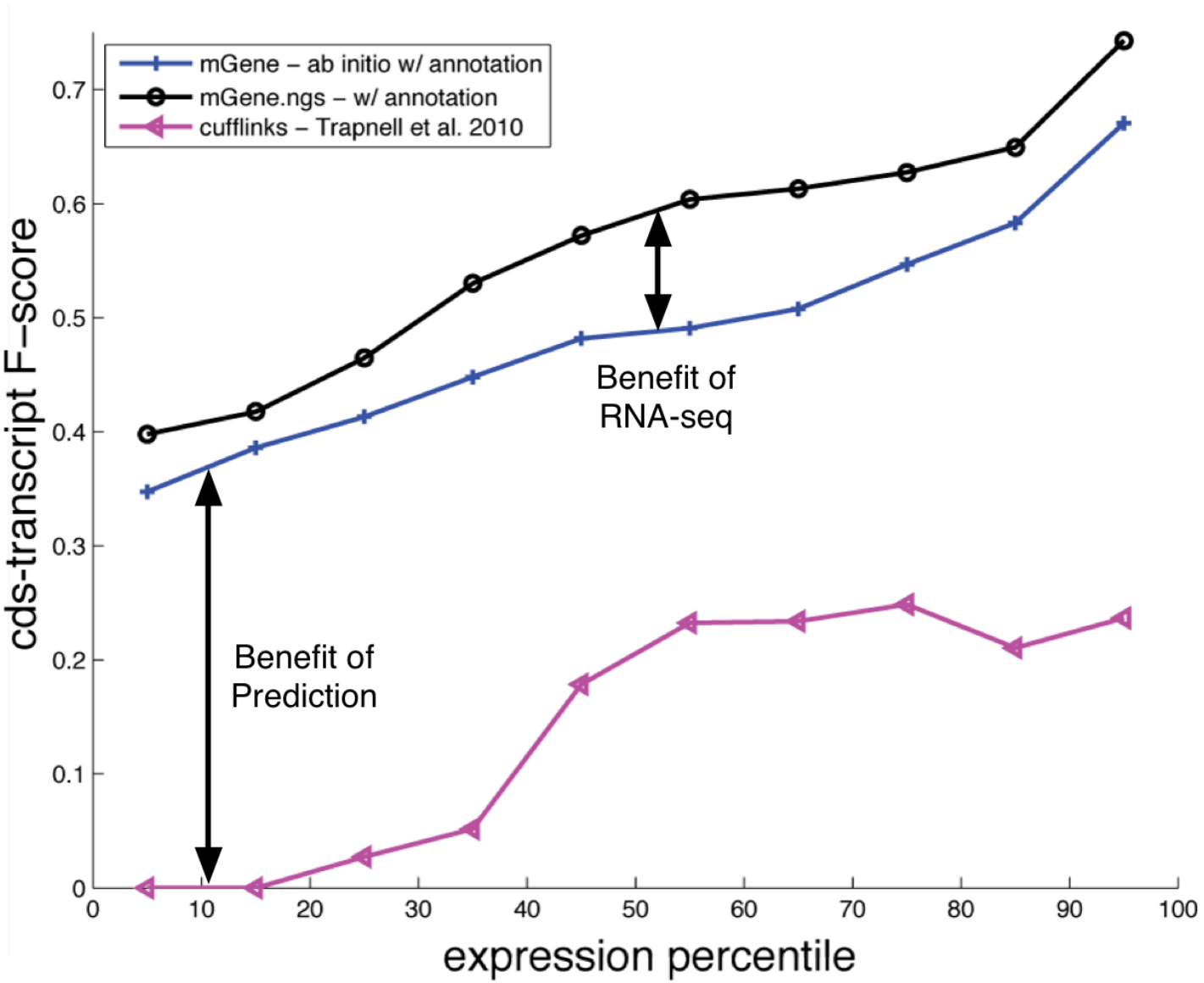
RNA-Seq:

- paired-end, strand-specific RNA-Seq (Illumina)
- 76bp reads, 50 million reads (2 lanes, $\approx$ 2.000 Euro)
- Alignment with Palmapper [4]

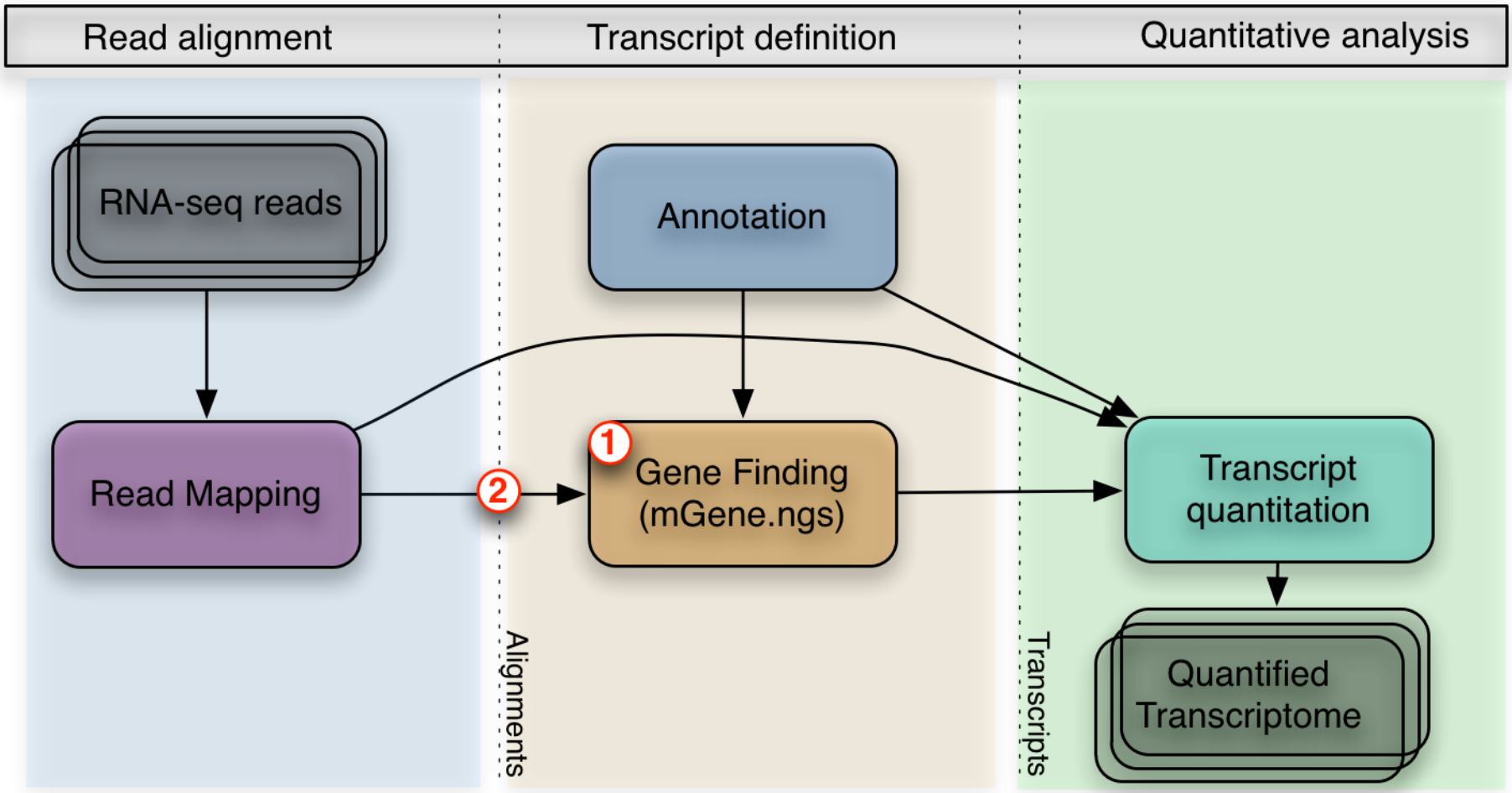
Evaluation:

- Transcript-level F-score of *coding transcripts*
  ... for different expression levels
- Compare
  - mGene (*ab initio*) [6]
  - mGene.ngs (with RNA-Seq)
  - Cufflinks (only based on RNA-Seq) [10]

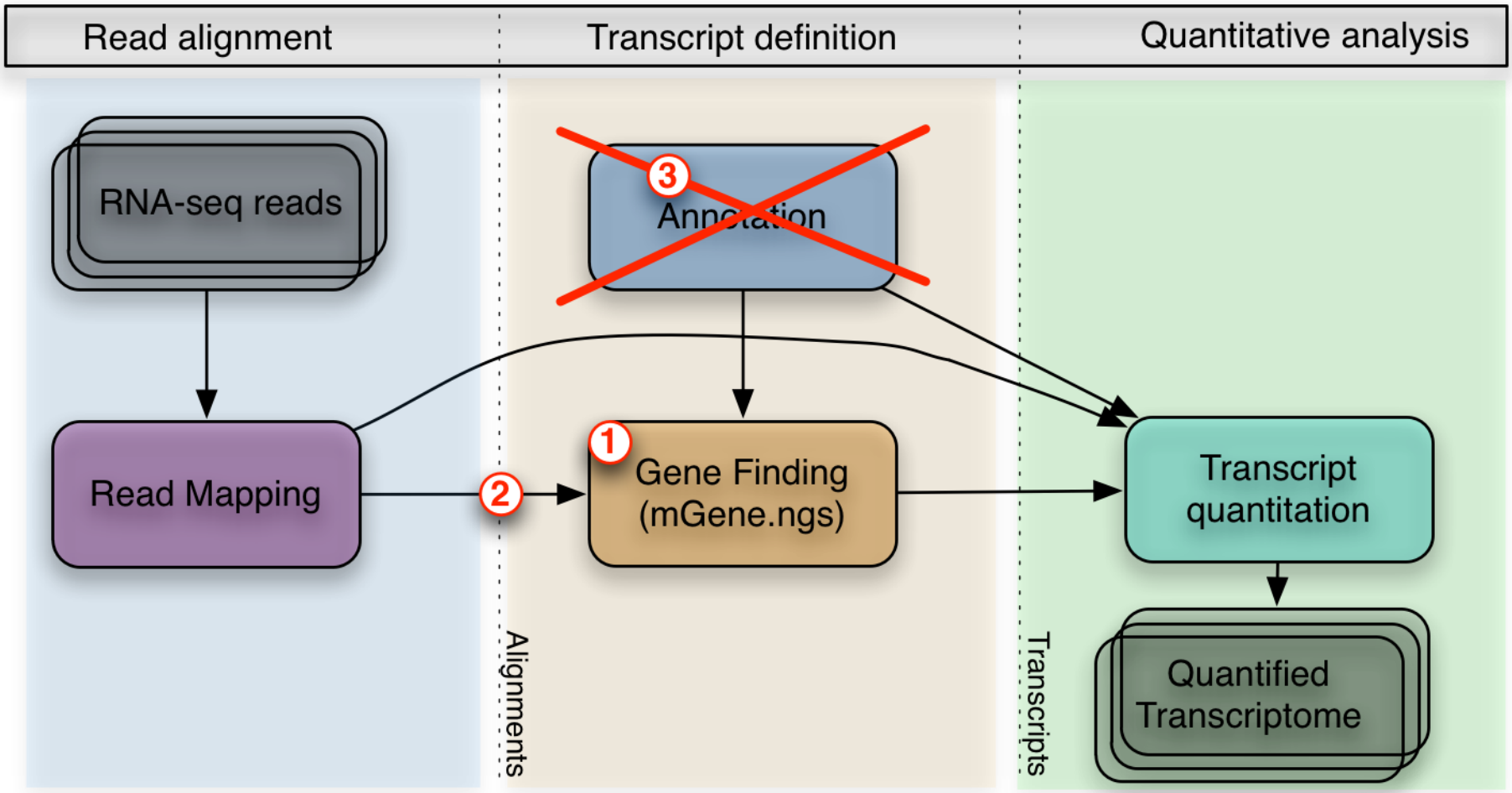
Results for *C. elegans*



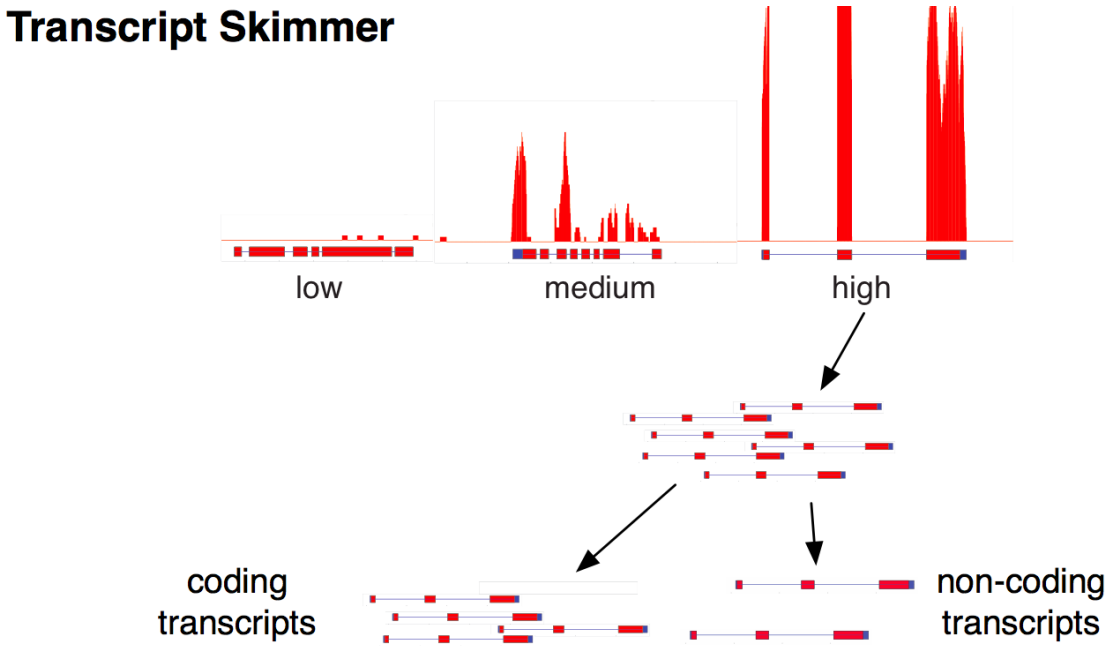
RNA-Seq Analysis Pipeline(s)



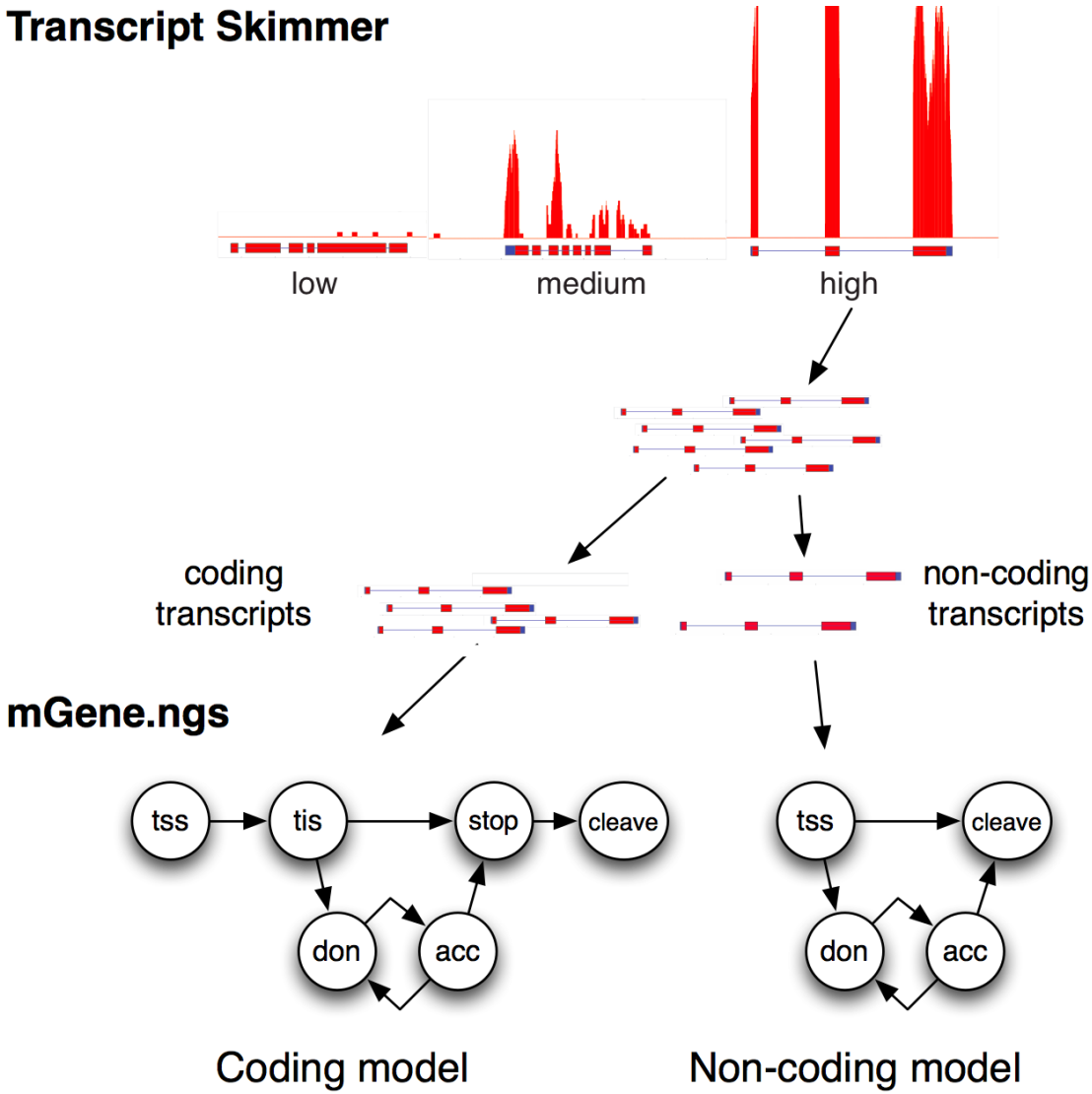
RNA-Seq Analysis Pipeline(s)



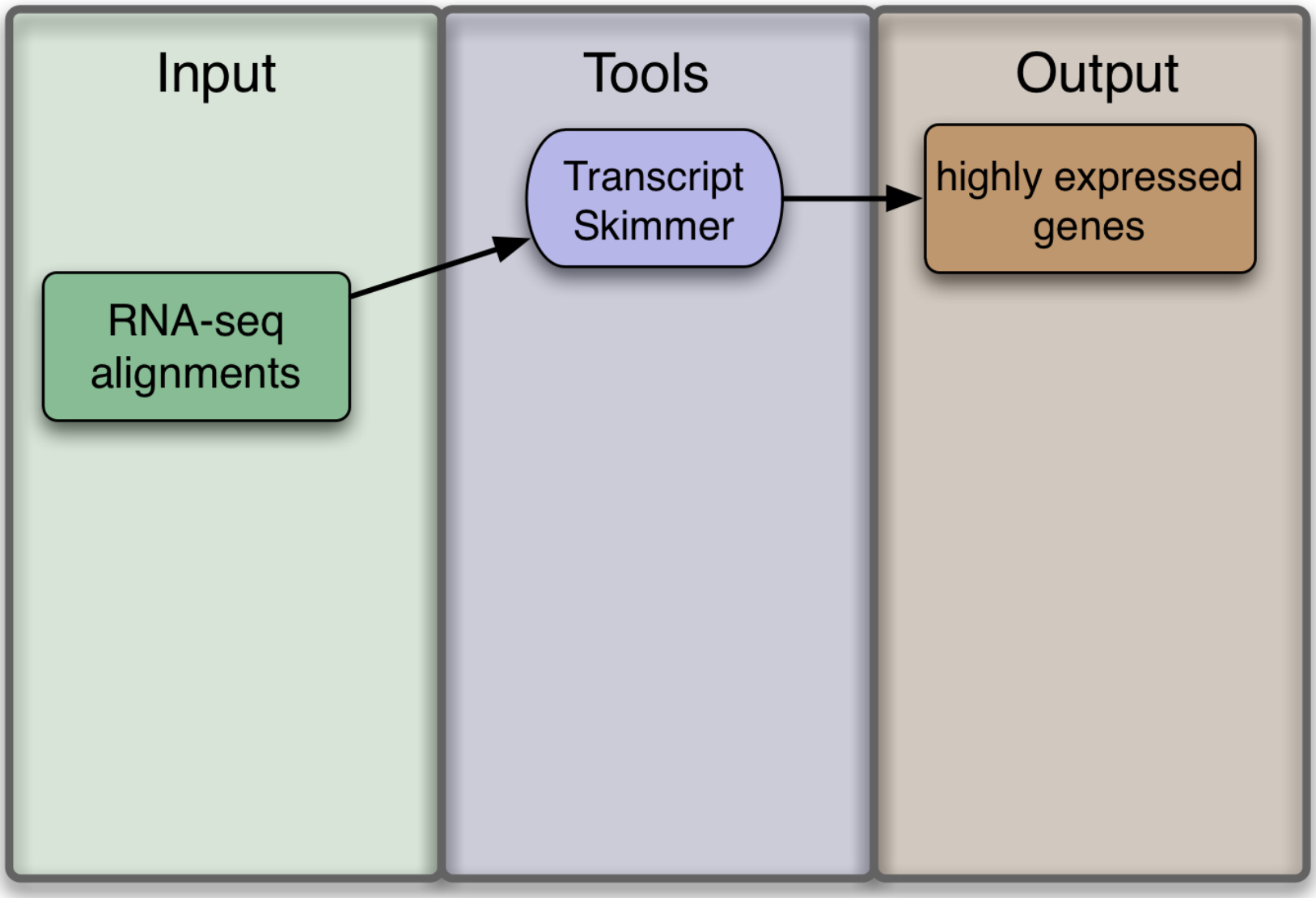
Step 3: Skimming and Non-coding Transcripts



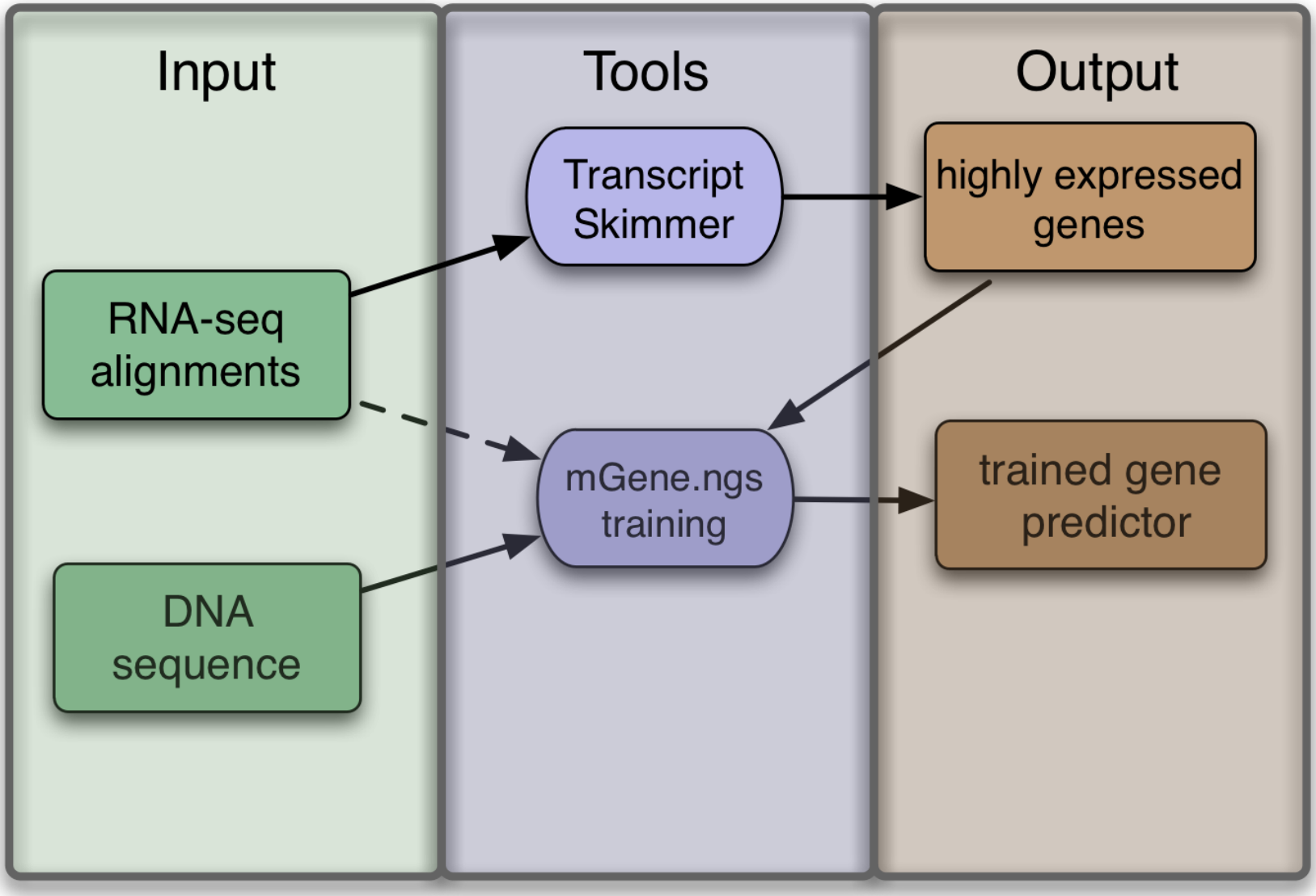
Step 3: Skimming and Non-coding Transcripts



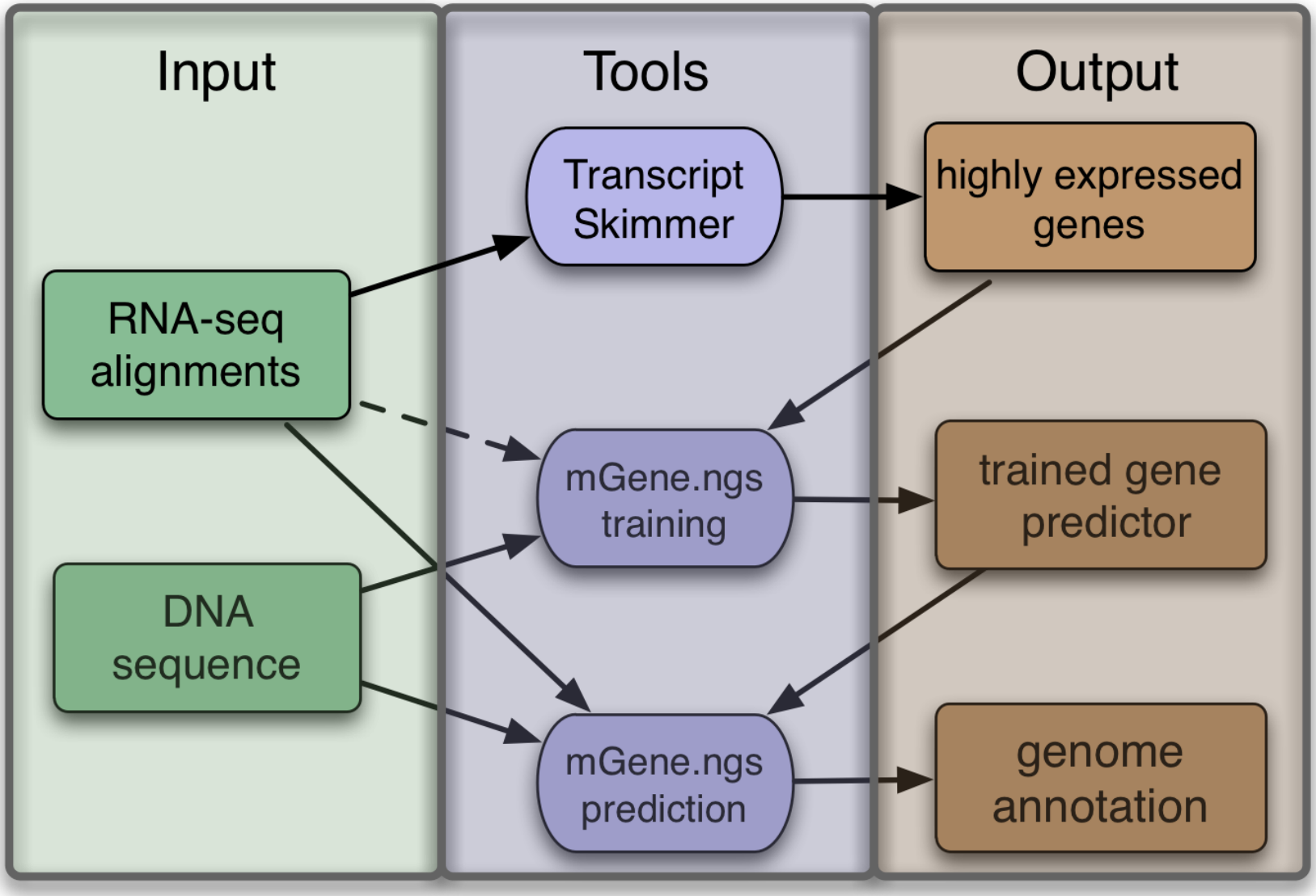
Genome Annotation Workflow



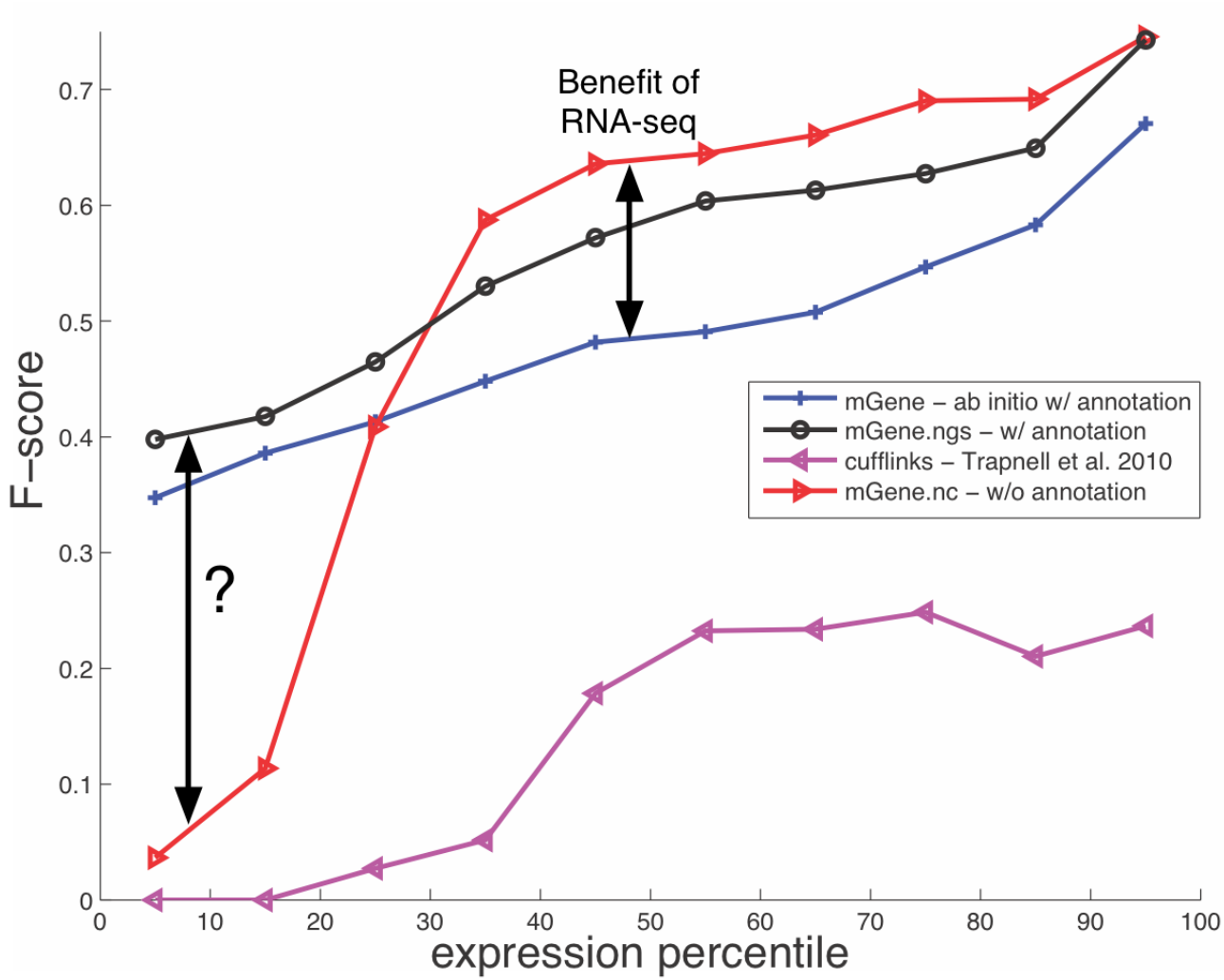
Genome Annotation Workflow



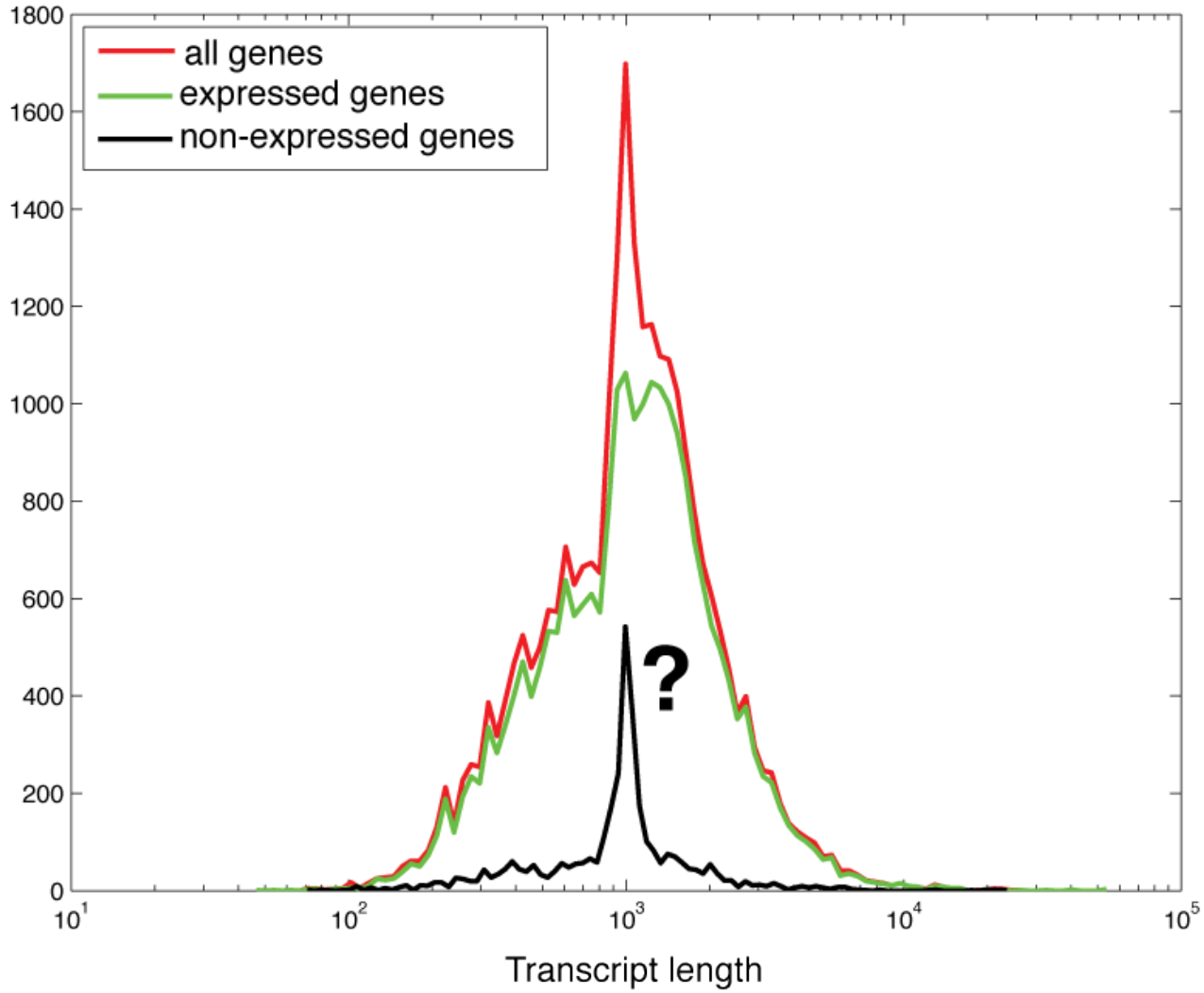
Genome Annotation Workflow



Results for *C. elegans*

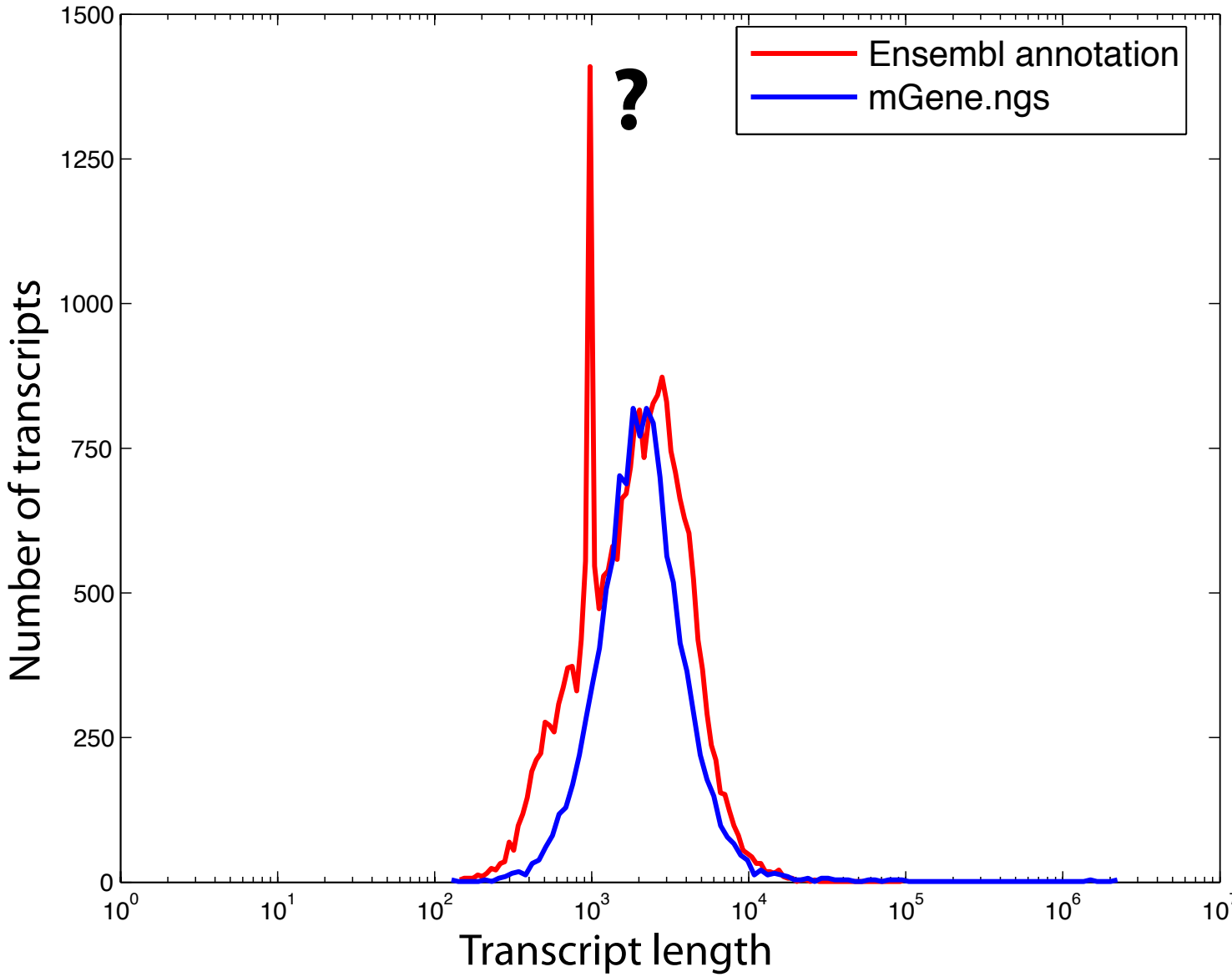


Coding Region Lengths (*C. elegans*)



Artifacts in *C. elegans* annotation?

Coding Region Lengths (mouse)



Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-Seq and genomic information
  - 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
- Currently used to (re-)annotate *mouse*, *Drosophila* species, *A. thaliana* strains, *Capsella*, *maize*, ...
- Web service available `galaxy.fml.mpg.de`
- Cloud integration planned
- Source will be free (GPL and other models) `mgene.org`

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-Seq and genomic information
  - 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
- Currently used to (re-)annotate *mouse*, *Drosophila* species, *A. thaliana* strains, *Capsella*, *maize*, ...
- Web service available `galaxy.fml.mpg.de`
- Cloud integration planned
- Source will be free (GPL and other models) `mgene.org`

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-Seq and genomic information
  - 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
- Currently used to (re-)annotate *mouse*, *Drosophila* species, *A. thaliana* strains, *Capsella*, *maize*, ...
- Web service available `galaxy.fml.mpg.de`
- Cloud integration planned
- Source will be free (GPL and other models) `mgene.org`

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-Seq and genomic information
  - 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
- Currently used to (re-)annotate *mouse*, *Drosophila* species, *A. thaliana* strains, *Capsella*, *maize*, ...
- Web service available `galaxy.fml.mpg.de`
- Cloud integration planned
- Source will be free (GPL and other models) `mgene.org`

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-Seq and genomic information
  - 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
- Currently used to (re-)annotate *mouse*, *Drosophila* species, *A. thaliana* strains, *Capsella*, *maize*, ...
- Web service available `galaxy.fml.mpg.de`
- Cloud integration planned
- Source will be free (GPL and other models) `mgene.org`

Conclusion

- Genome annotation pipeline
  - mGene.ngs integrates RNA-Seq and genomic information
  - 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
- Currently used to (re-)annotate *mouse*, *Drosophila* species, *A. thaliana* strains, *Capsella*, *maize*, ...
- Web service available `galaxy.fml.mpg.de`
- Cloud integration planned
- Source will be free (GPL and other models) `mgene.org`

Further information

Slides available at:

fml.mpg.de/raetsch/lectures

Jonas Behr

Gabriele Schweikert

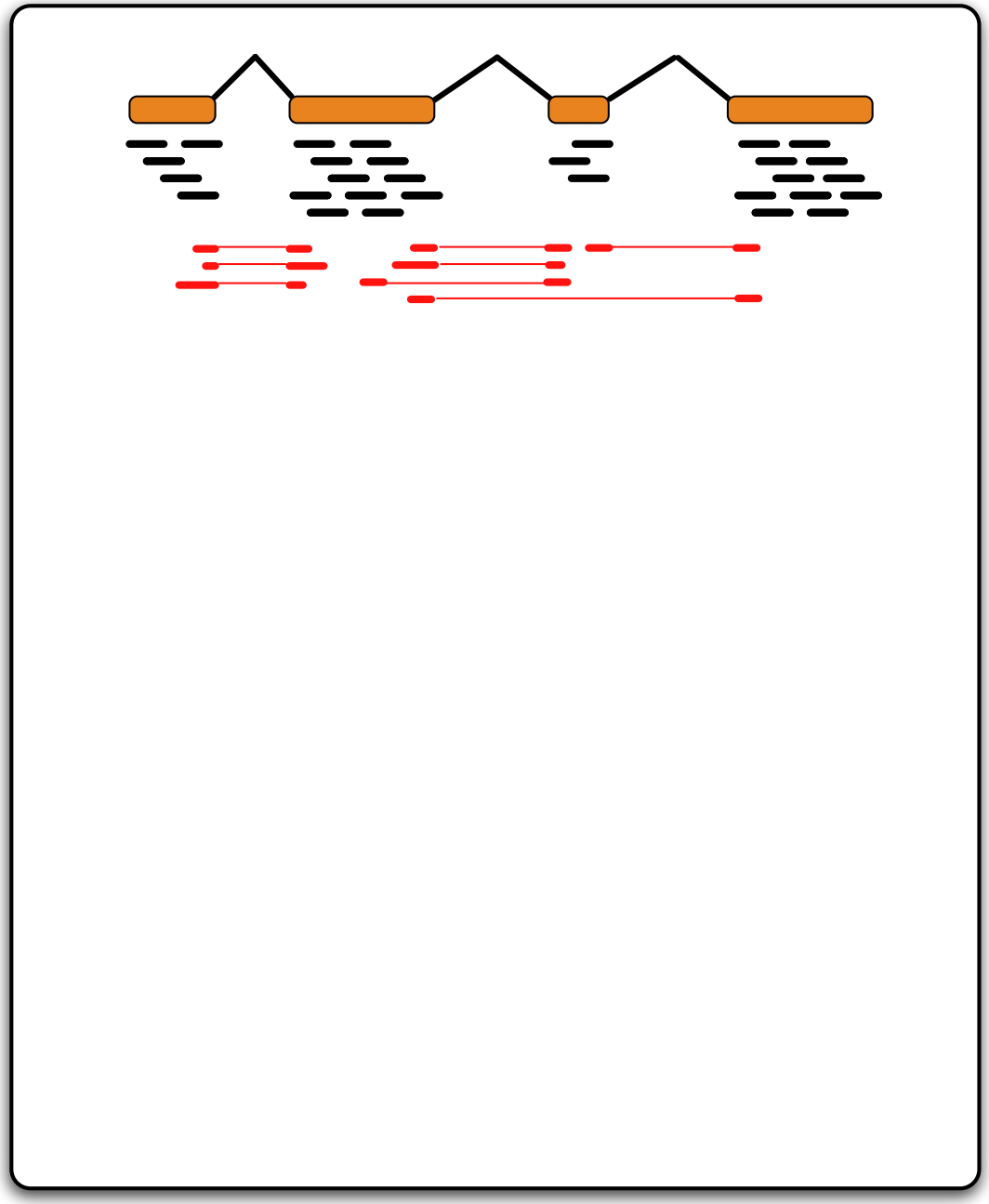
Acknowledgments

Gene finding:	**Jonas Behr, Georg Zeller, Gabriele Schweikert**
Quantification:	**Regina Bohnert**
Library preparation:	**Lisa Hartmann, Lisa M. 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

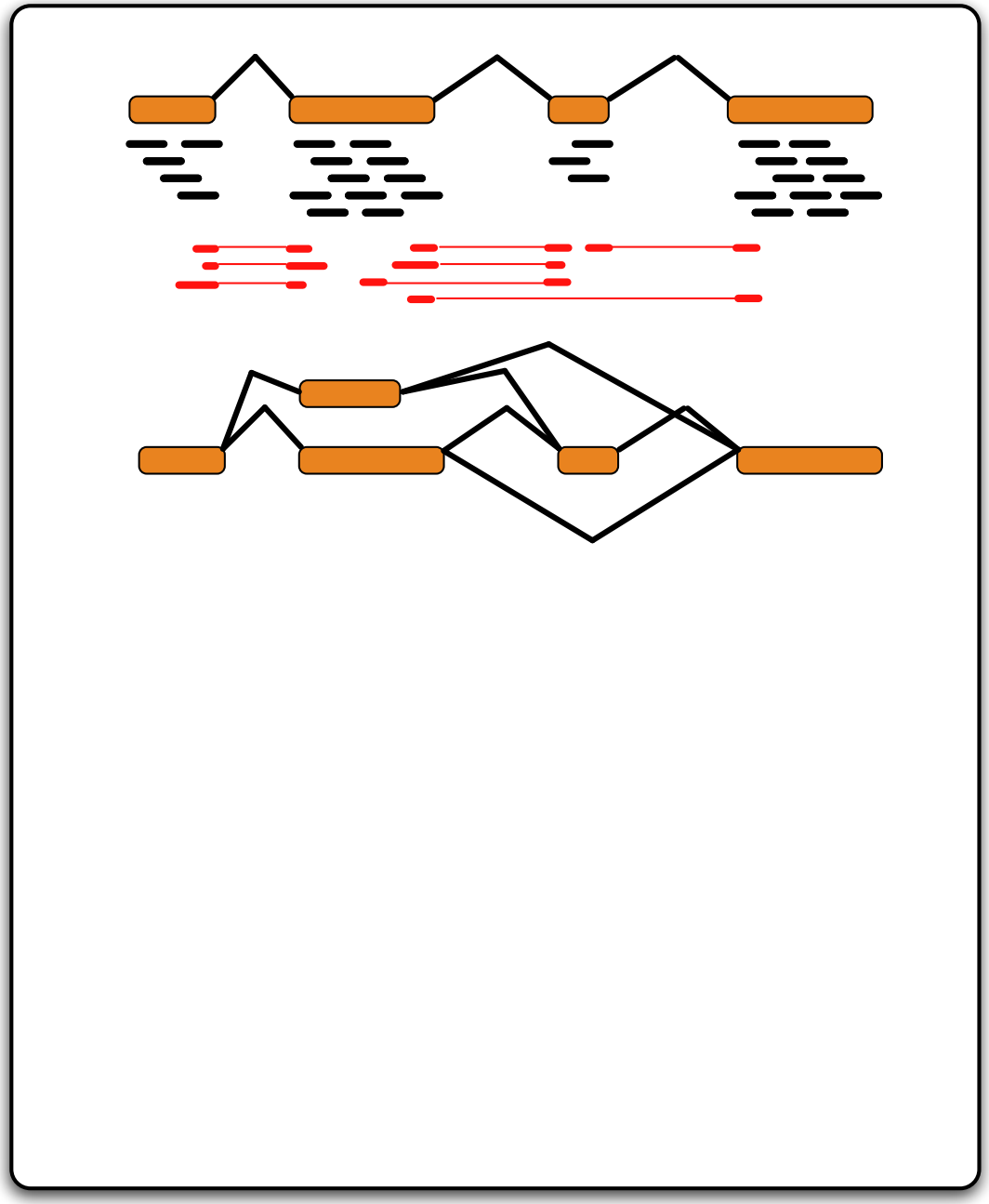
Infer alternative isoforms

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



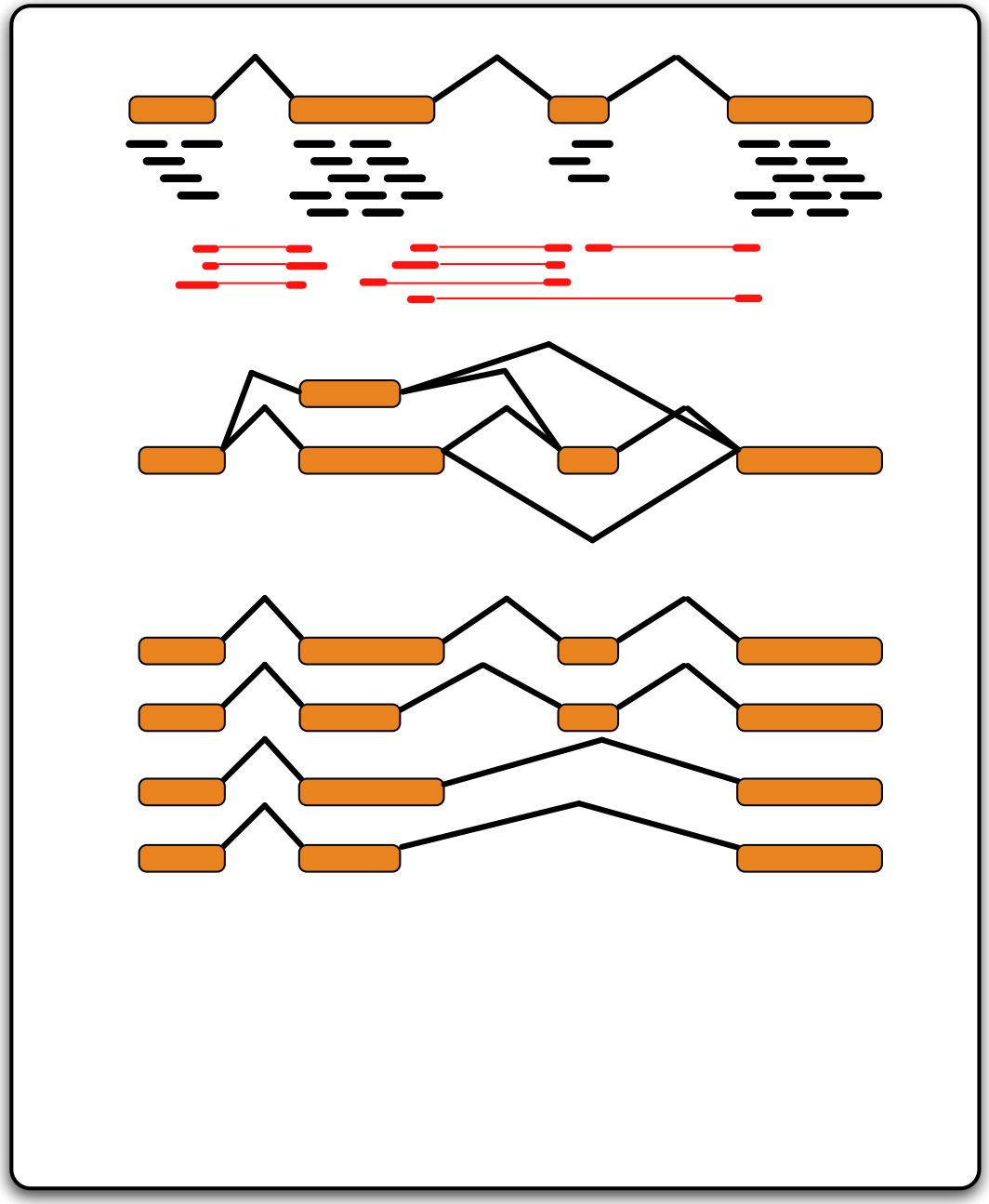
Infer alternative isoforms

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



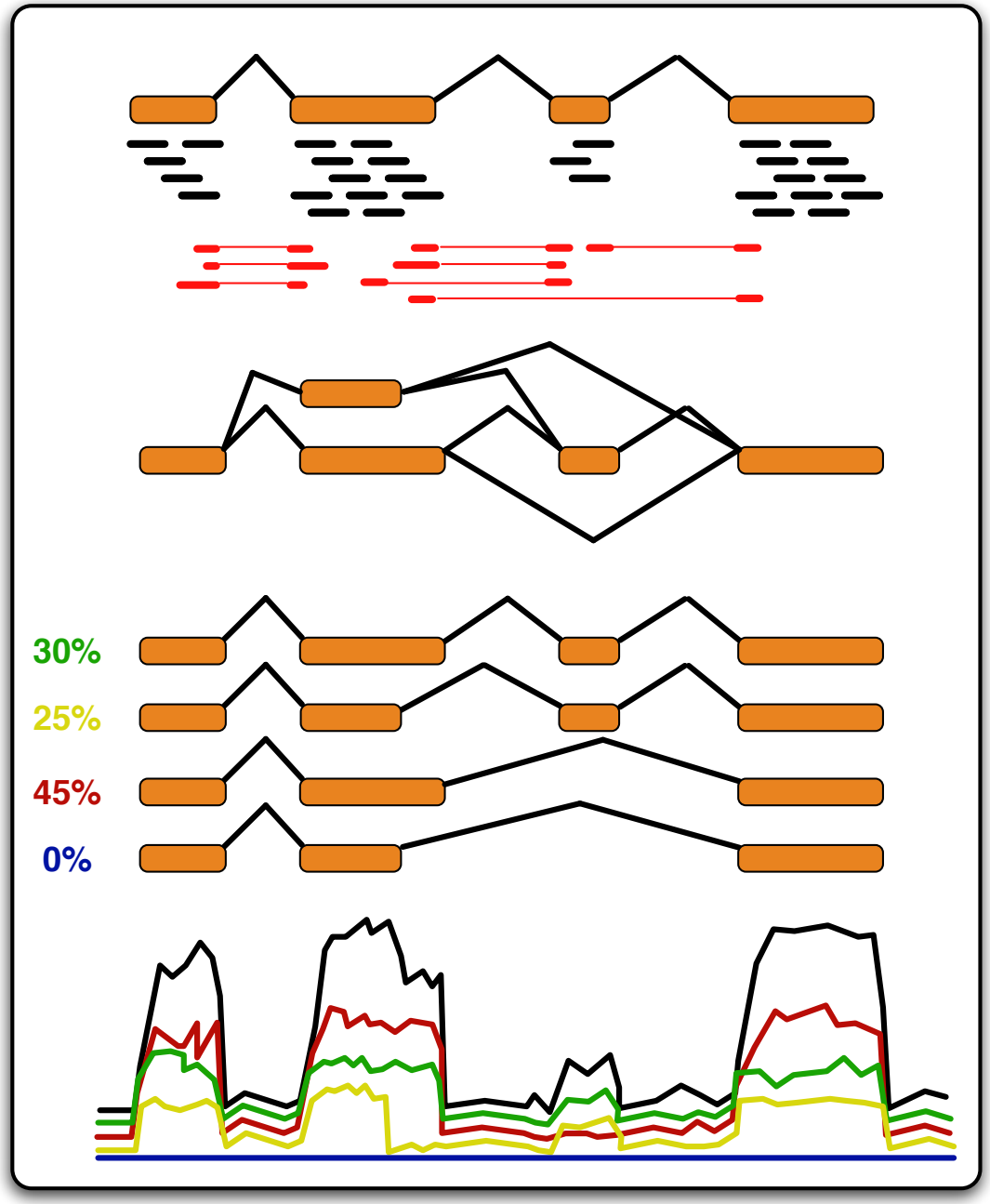
Infer alternative isoforms

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



Infer alternative isoforms

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



References I

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

[2] 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*, 9(549), 2008.

[3] 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

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

[5] Gabriele Schweikert, Jonas Behr, Alexander Zien, Georg Zeller, Cheng Soon Ong, Sören Sonnenburg, and Gunnar Rätsch. mgene.web: a web service for accurate computational gene finding. *Nucleic Acids Research*, Web Server Issue, 2009. URL `http://mgene.org/web`. Advance Access published on June 3, 2009.

References III

[6] Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krüger, Sören Sonnenburg, and Gunnar Rätsch. mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, September 2009.

[7] Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krüger, Sören Sonnenburg, and Gunnar Rätsch. mgene: Accurate svm-based gene finding with an application to nematode genomes. *Genome Research*, 2009. URL `http://genome.cshlp.org/content/early/2009/06/29/gr.090597.108.full.pdf+html`. Advance access June 29, 2009.

References IV

[8] S Sonnenburg, G Schweikert, P Philips, J Behr, and G Rätsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S7, 2007. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-8-S10-S7.

[9] Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):e472–480, 2006.

[10] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28 (5):511–5, May 2010. doi: 10.1038/nbt.1621.