Online Quantitative Transcriptome Analysis

Gunnar Rätsch



Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

July 9, 2010

Hitseq 2010, Boston

http://galaxy.fml.mpg.de

Online Quantitative Transcriptome Analysis

(Oqtans)

Gunnar Rätsch



Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

July 9, 2010

Hitseq 2010, Boston

http://galaxy.fml.mpg.de

Online Quantitative Transcriptome Analysis

(Oqtans)

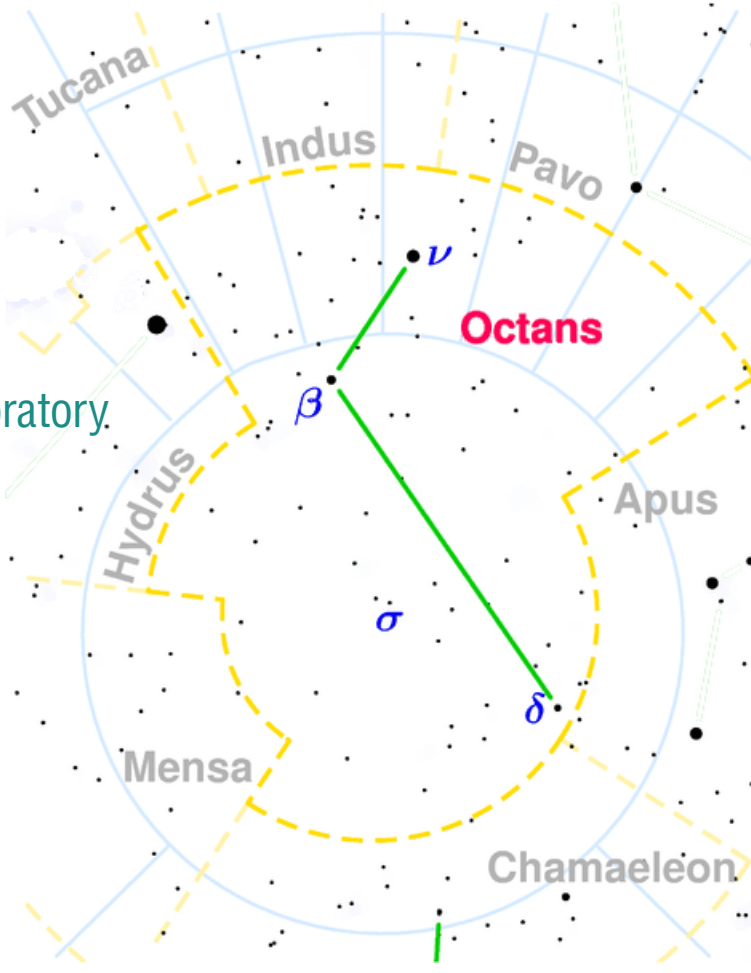
Gunnar Rätsch

Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

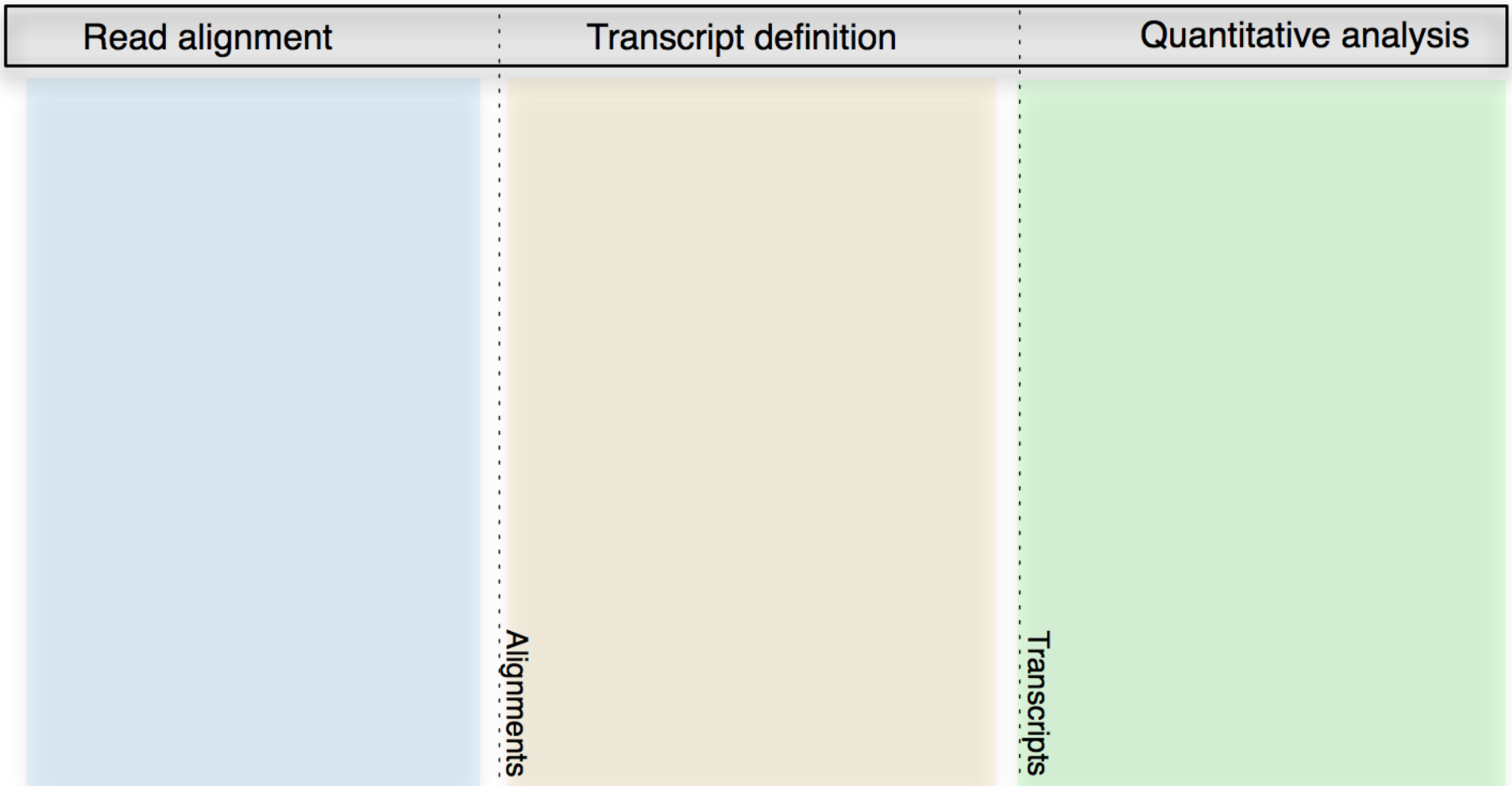
July 9, 2010

Hitseq 2010, Boston

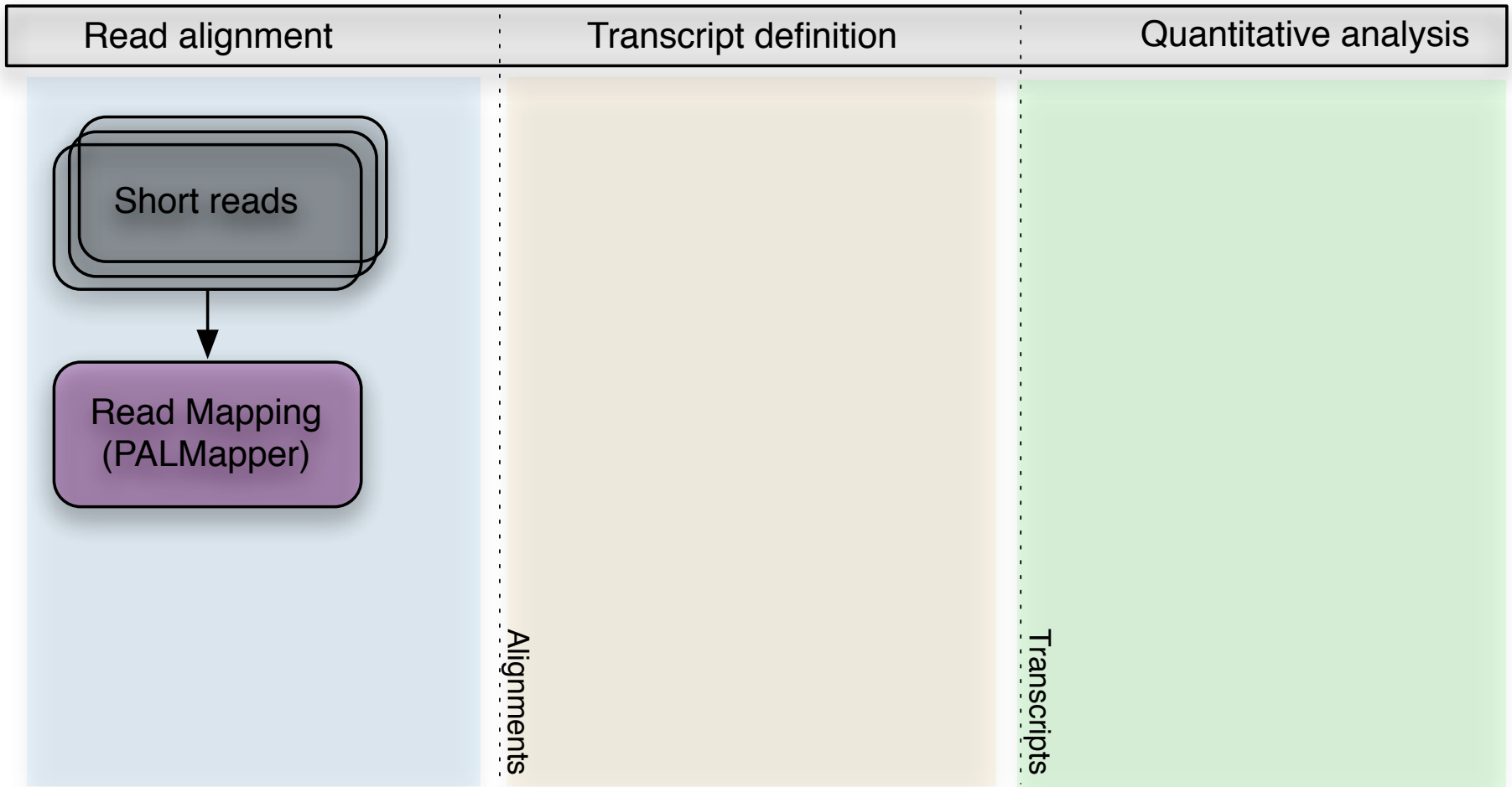


http://galaxy.fml.mpg.de

RNA-seq Analysis Pipeline(s)

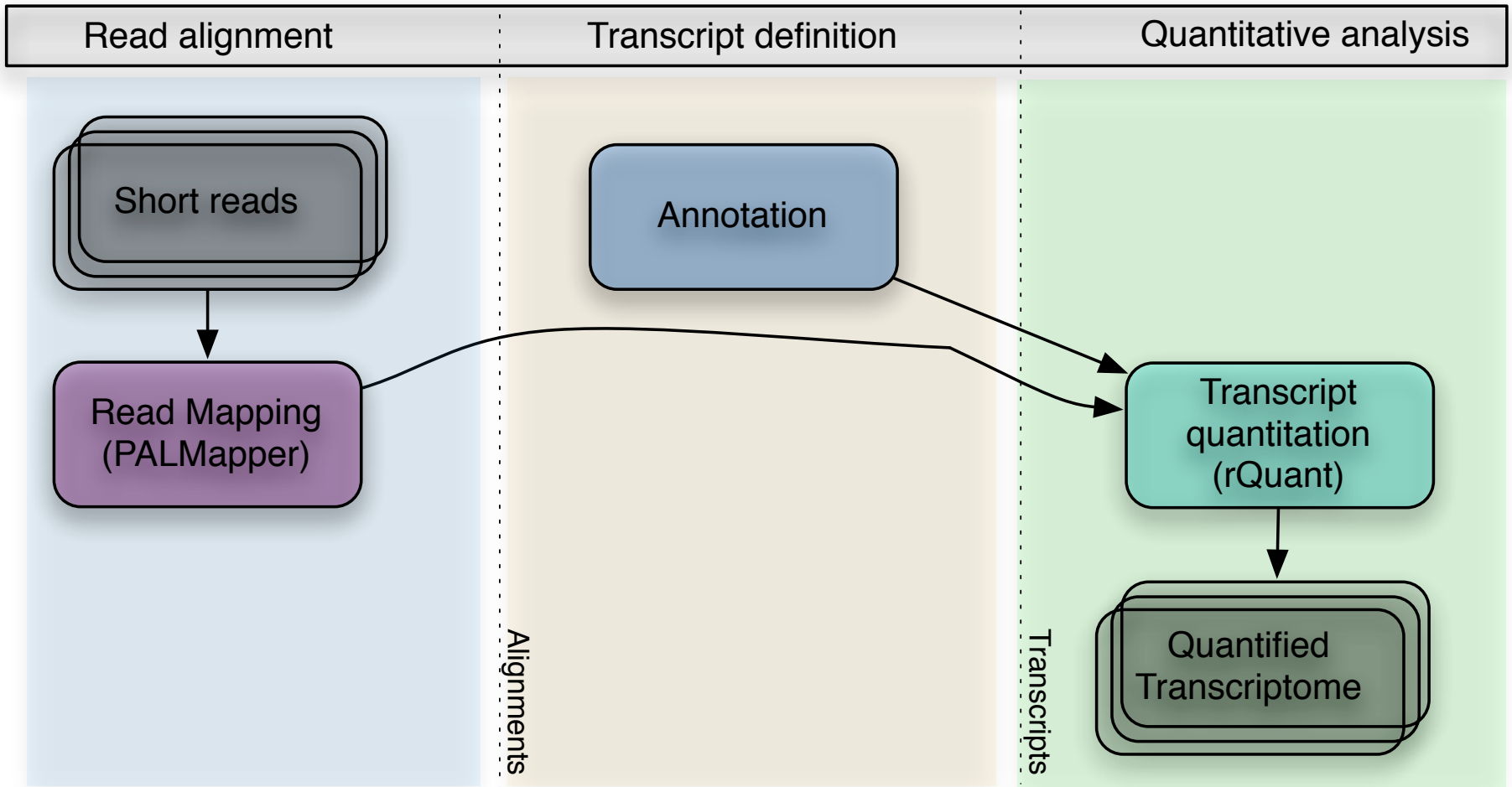


RNA-seq Analysis Pipeline(s)



ShortSIG 2008: Palmapper ($\sim$ QPalma) (De Bona et al., 2008)

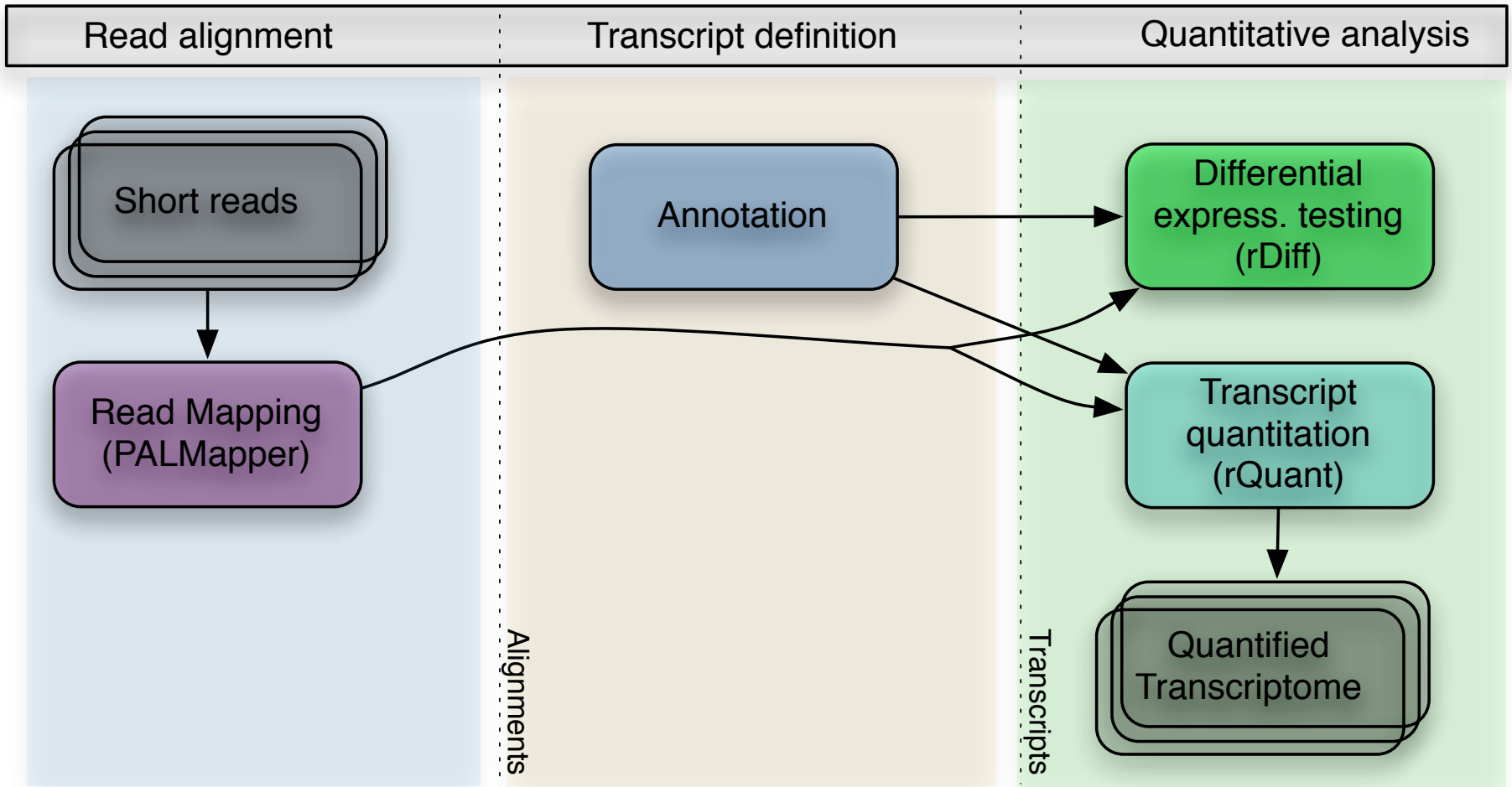
RNA-seq Analysis Pipeline(s)



ShortSIG 2008: Palmapper ($\sim$ QPalma) (De Bona et al., 2008)

ShortSIG 2009: rQuant (Bohnert and Rätsch, 2010)

RNA-seq Analysis Pipeline(s)

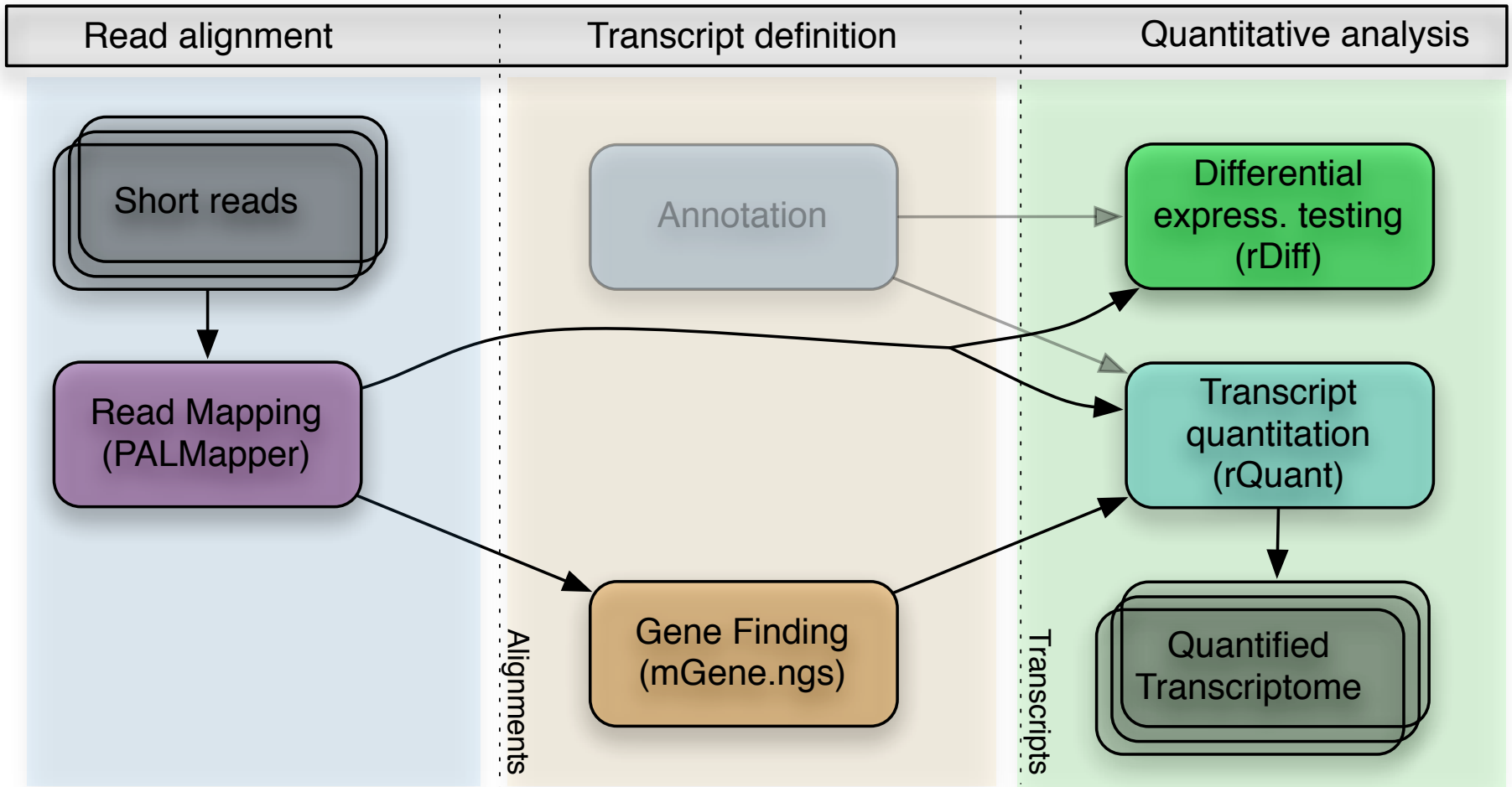


ShortSIG 2008: Palmapper ($\sim$ QPalma) (De Bona et al., 2008)

ShortSIG 2009: rQuant (Bohnert and Rätsch, 2010)

Hitseq 2010: rDiff (Stegle et al., 2010)

RNA-seq Analysis Pipeline(s)

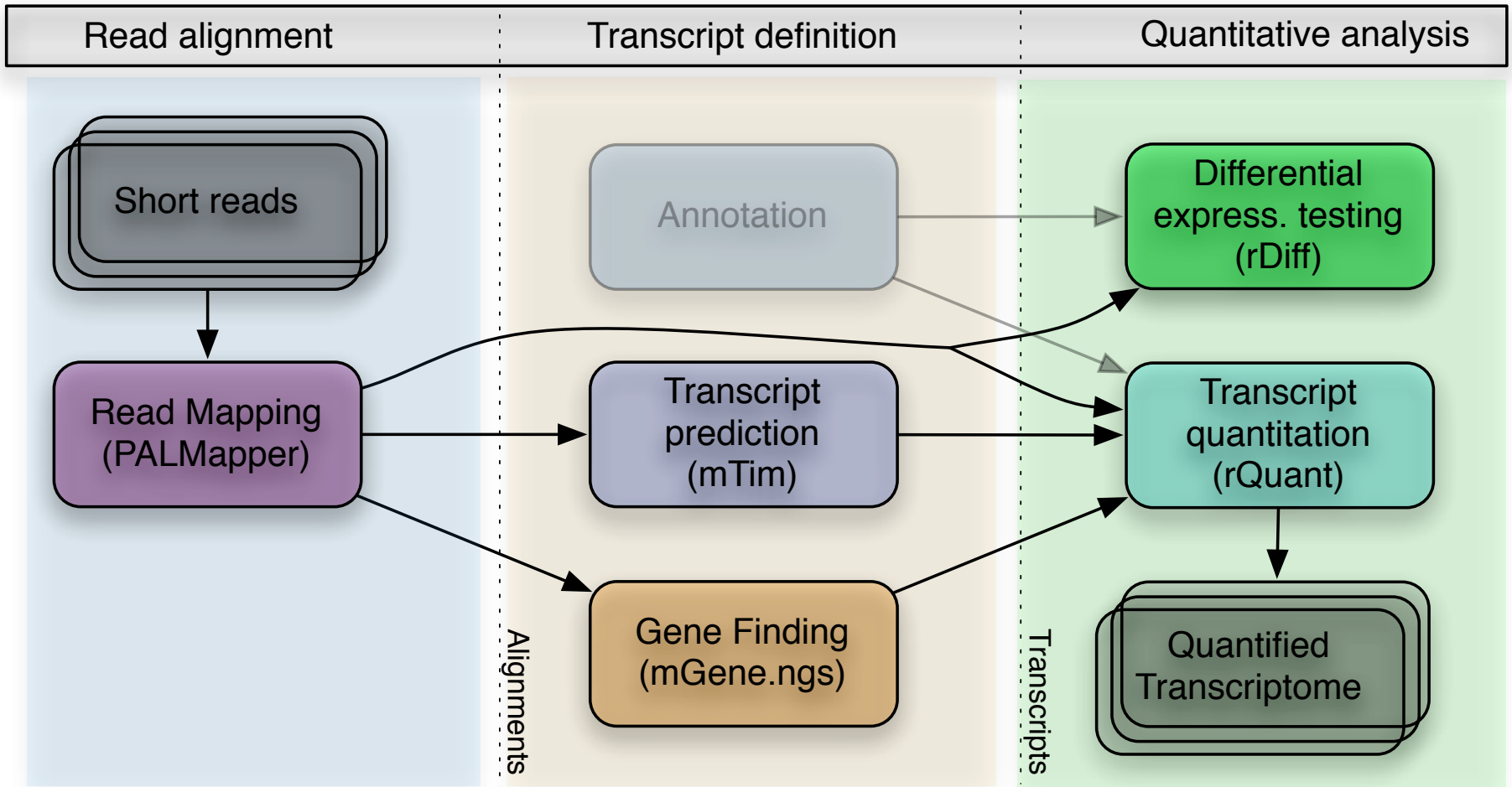


ShortSIG 2008: Palmapper ($\sim$ QPalma) (De Bona et al., 2008)

ShortSIG 2009: rQuant (Bohnert and Rätsch, 2010)

Hitseq 2010: rDiff (Stegle et al., 2010)

RNA-seq Analysis Pipeline(s)

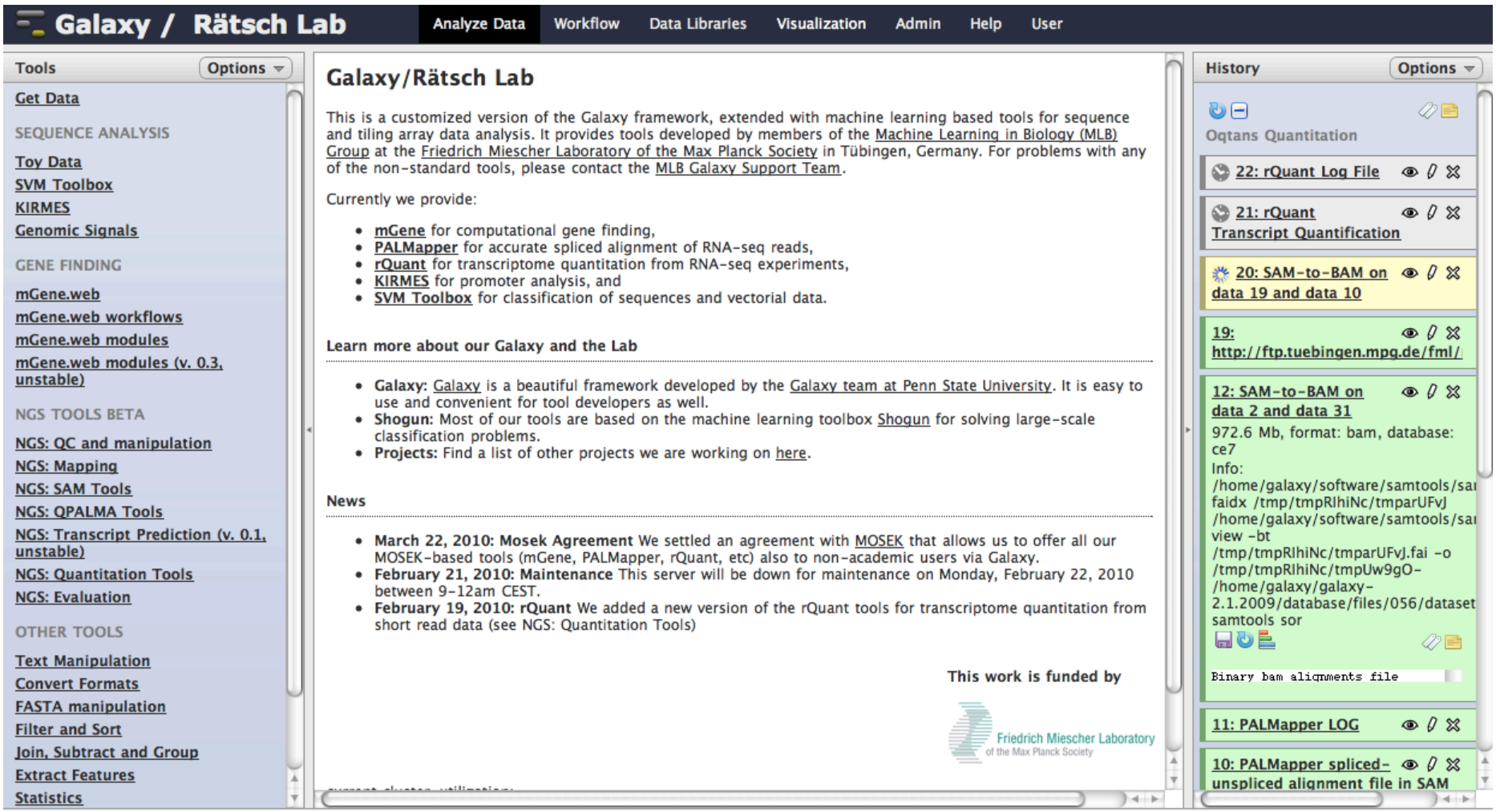


ShortSIG 2008: Palmapper ($\sim$ QPalma) (De Bona et al., 2008)

ShortSIG 2009: rQuant (Bohnert and Rätsch, 2010)

Hitseq 2010: rDiff (Stegle et al., 2010)

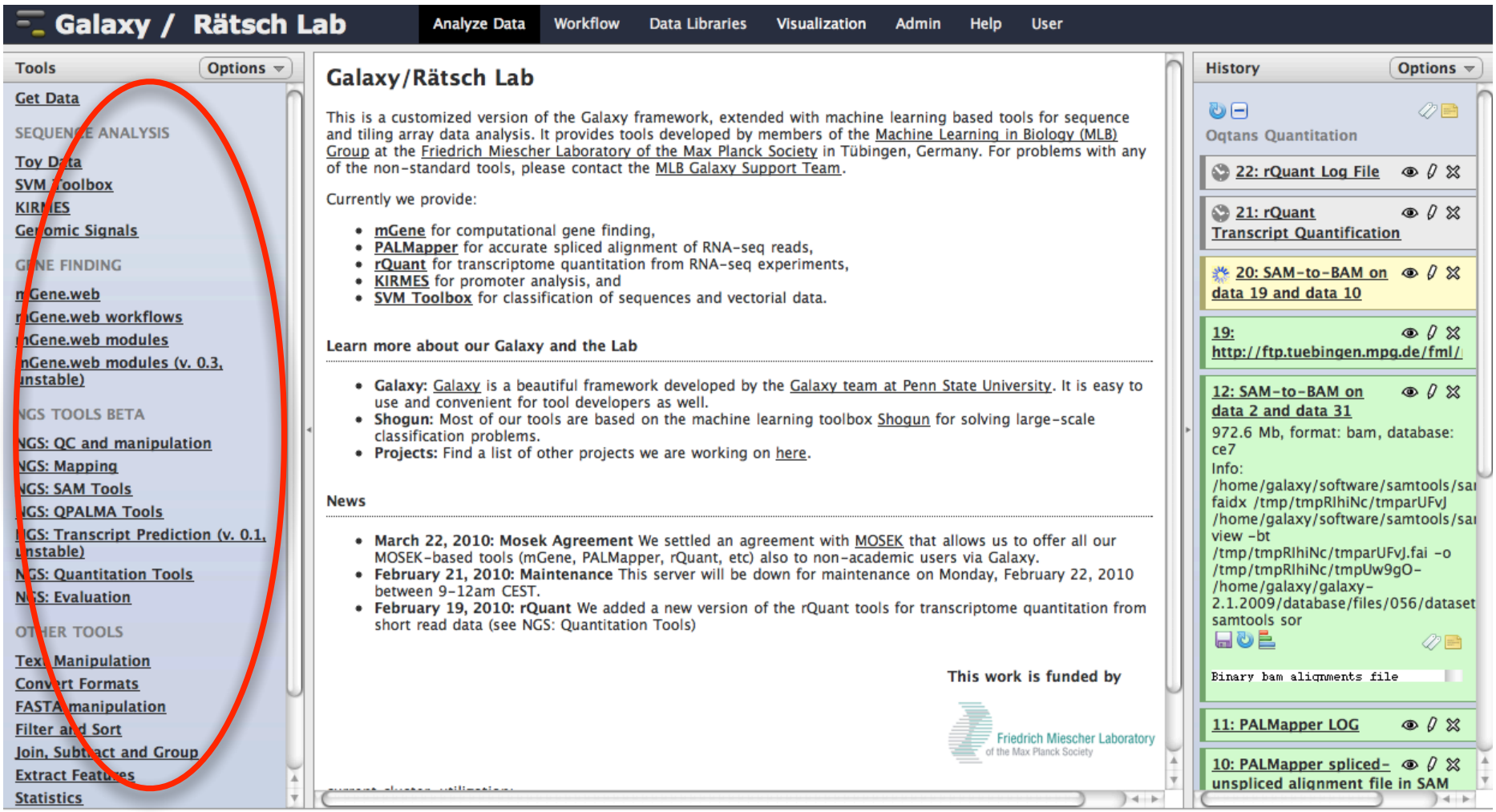
Galaxy is great!



(Taylor et al., 2007; Blankenberg et al., 2010a)

Galaxy is great!

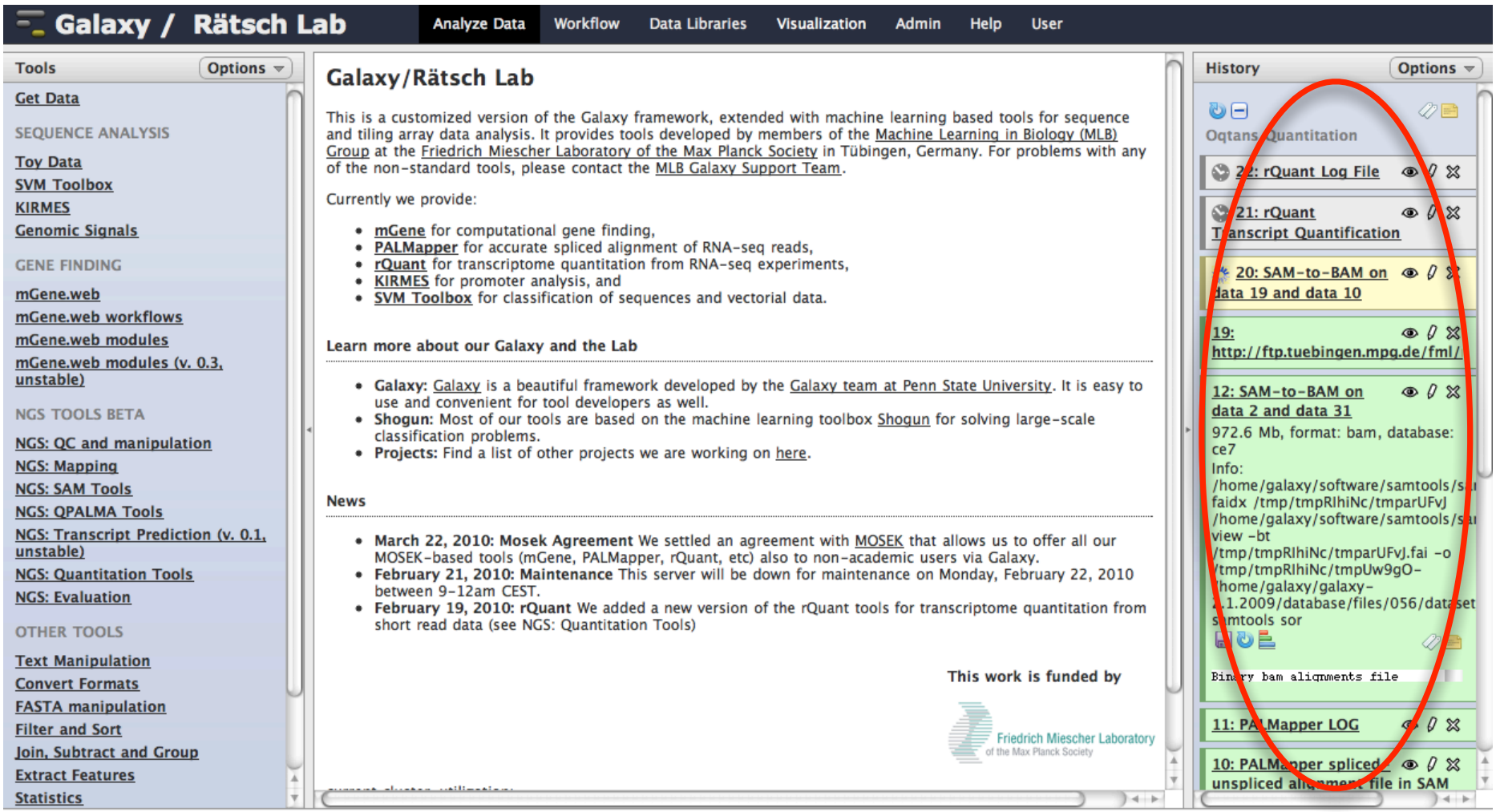
Tools



(Taylor et al., 2007; Blankenberg et al., 2010a)

Galaxy is great!

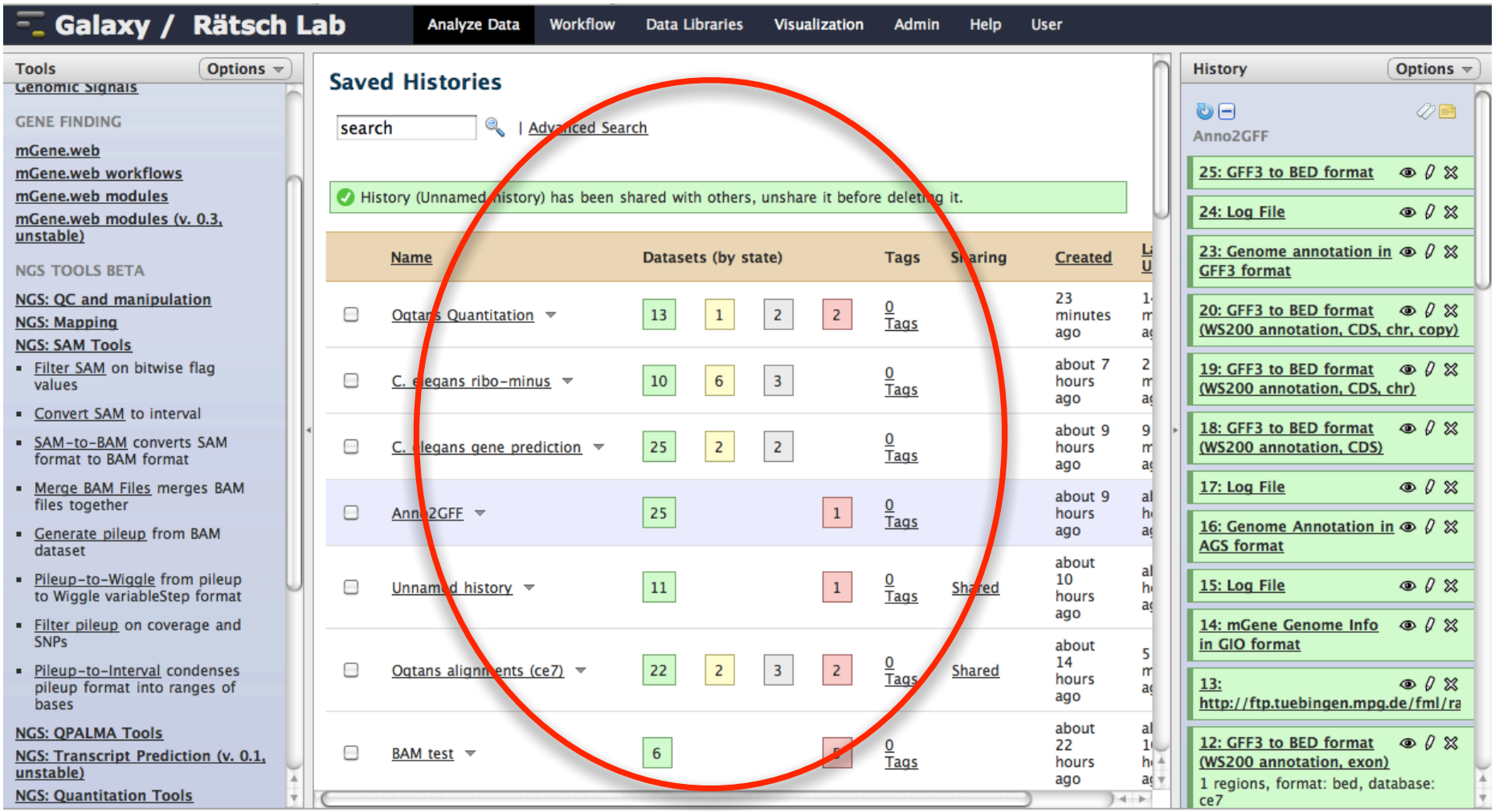
History



(Taylor et al., 2007; Blankenberg et al., 2010a)

Galaxy is great!

Working area



(Taylor et al., 2007; Blankenberg et al., 2010a)

Example 1: rQuant Transcript Quantitation



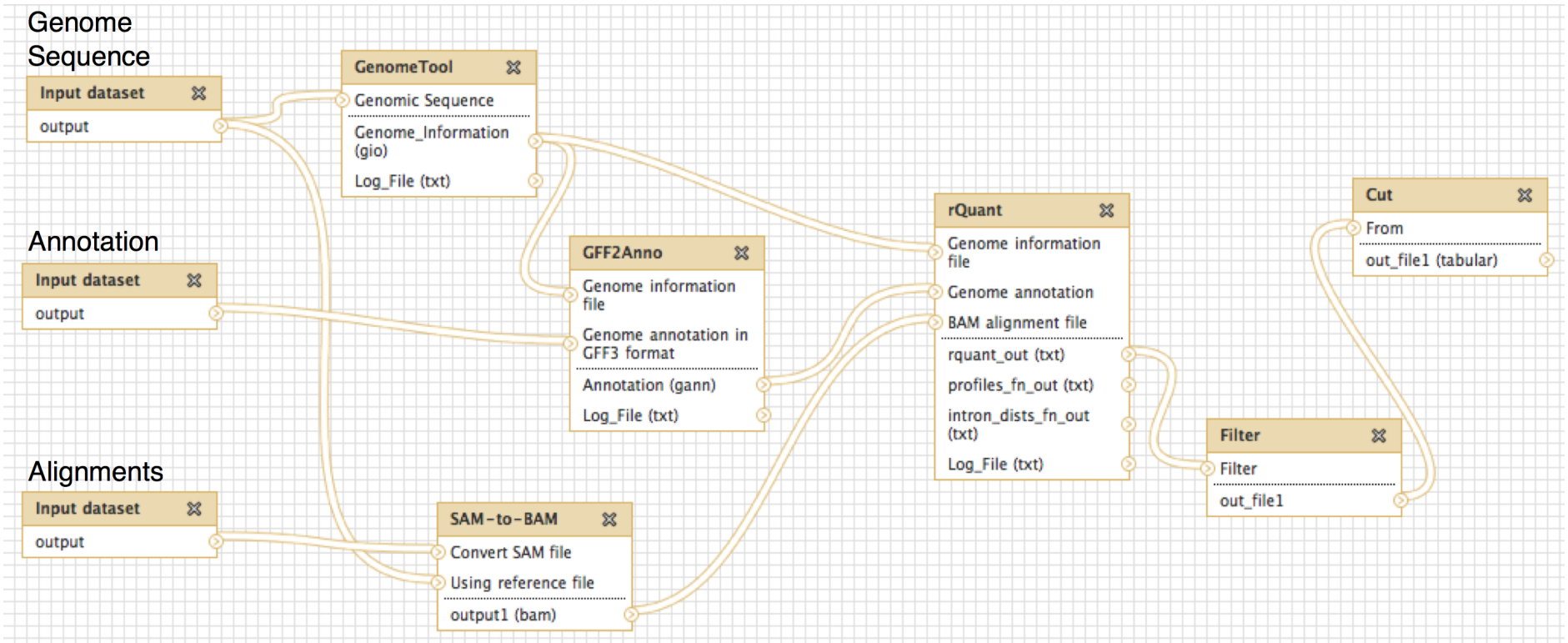
Inputs:

- Genome sequence (fasta)
- Annotation (gff3)
- Alignments (sam)

Example 1: rQuant Transcript Quantitation

Inputs:

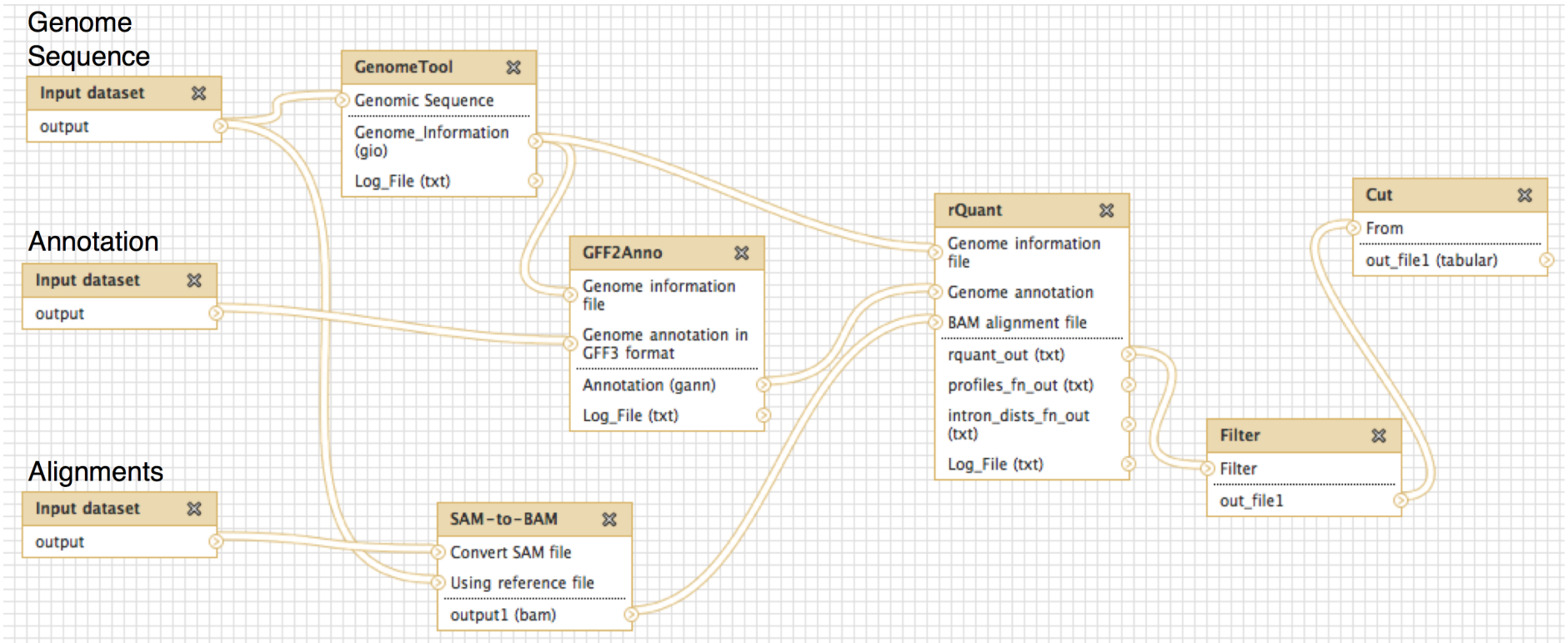
- Genome sequence (fasta)
- Annotation (gff3)
- Alignments (sam)



Example 1: rQuant Transcript Quantitation

Inputs:

- Genome sequence (fasta)
- Annotation (gff3)
- Alignments (sam)

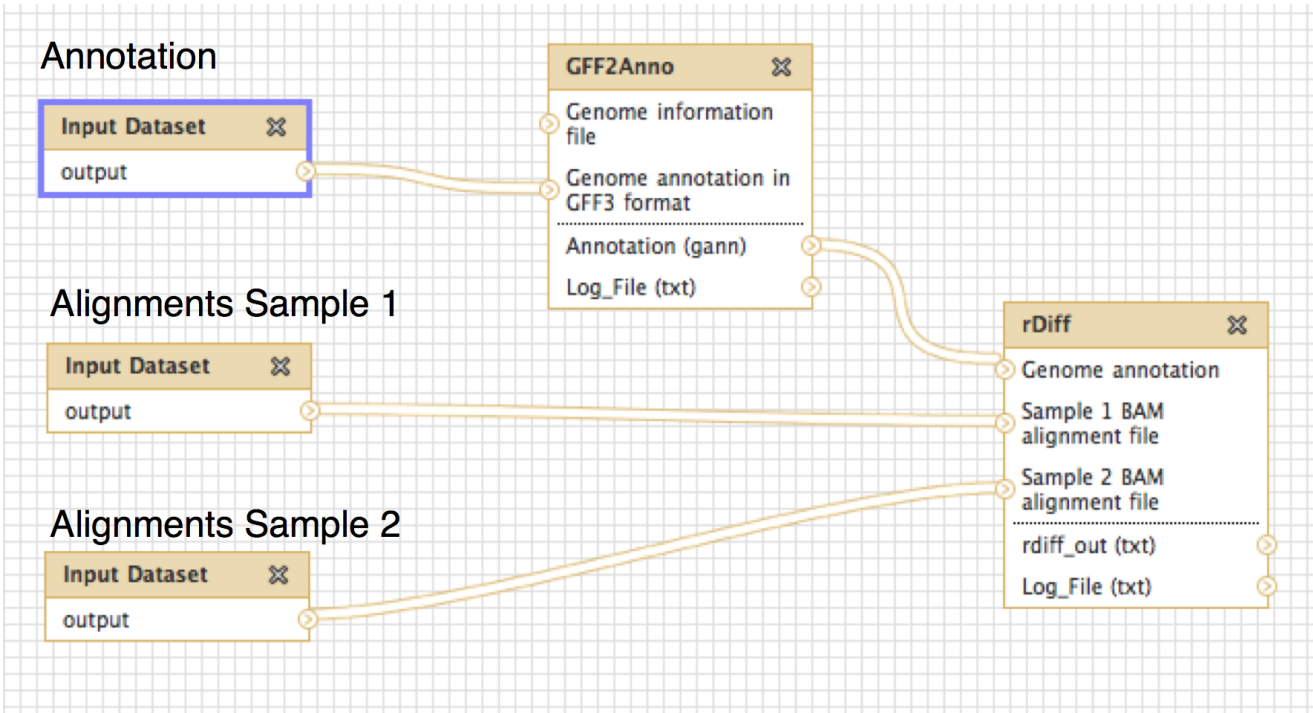


Output: Transcripts with RPKM values

Example 2: rDiff Differential Transcript Expression

Inputs:

- Annotation (gff3)
- Alignments Sample 1 (sam)
- Alignments Sample 2 (sam)



Output: List of transcripts with p-values

Example 3: *De Novo* Genome Annotation



Inputs:

- Genome sequence (contigs)
- RNA-seq sequences

Output:

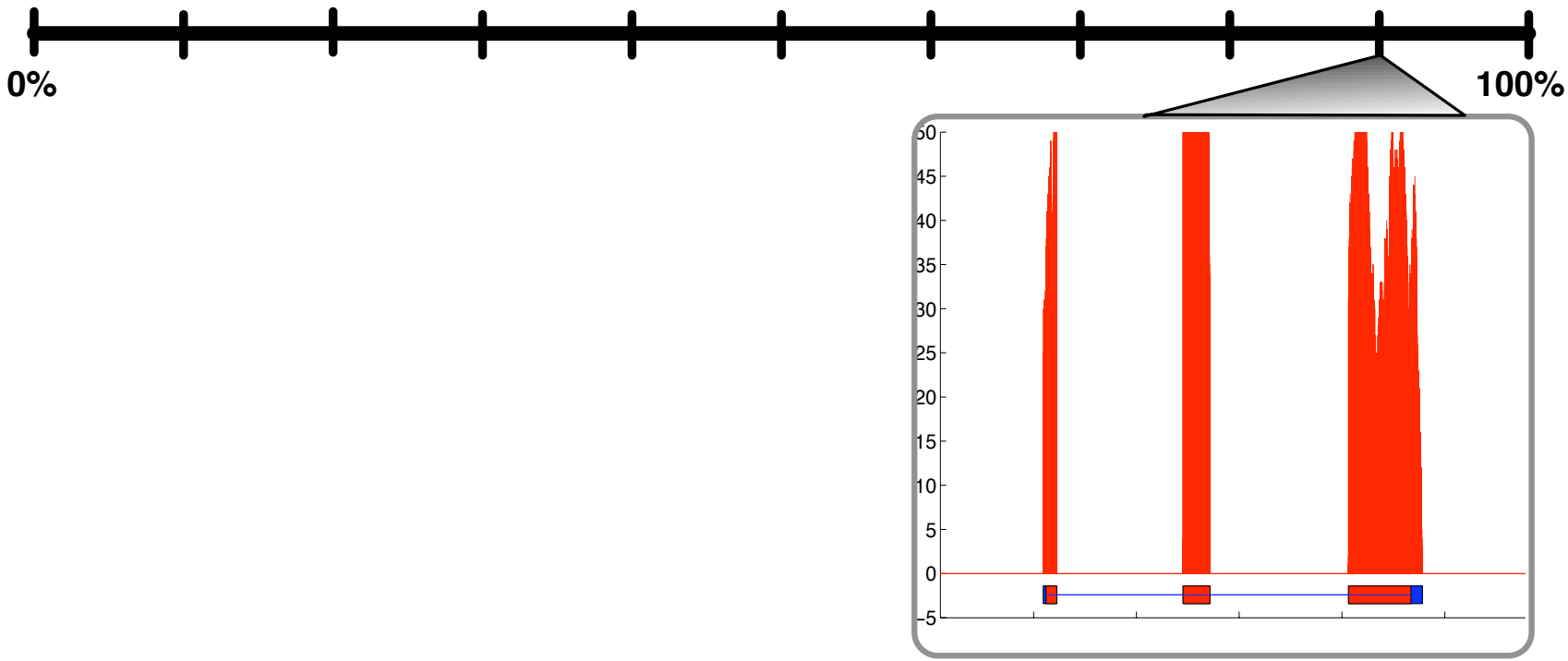
- Annotated genome

NB: No previous annotation needed

(Schweikert et al., 2009d; Behr et al., 2010)

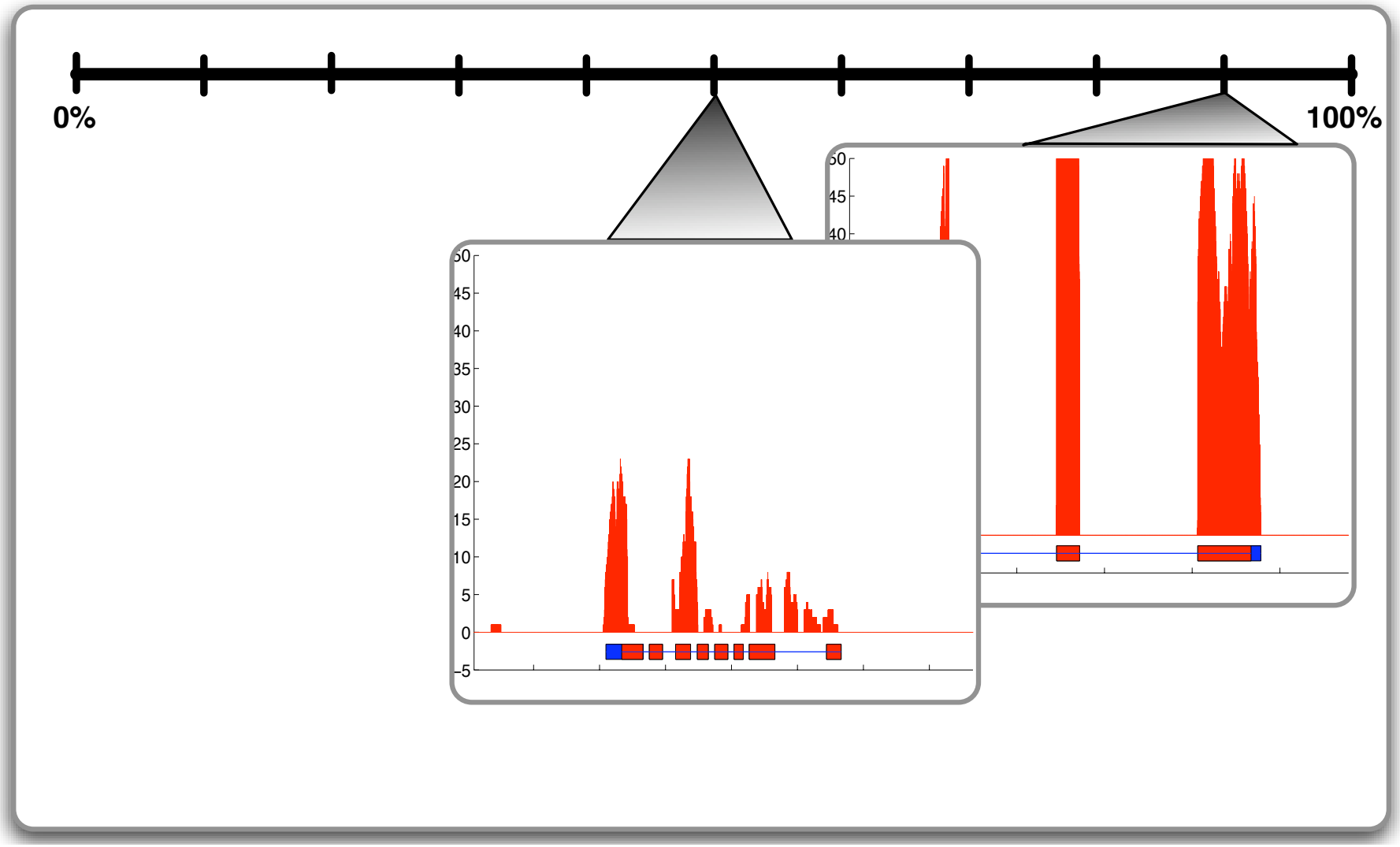
Transcript Prediction vs. Read Coverage

Annotated genes sorted by expression level



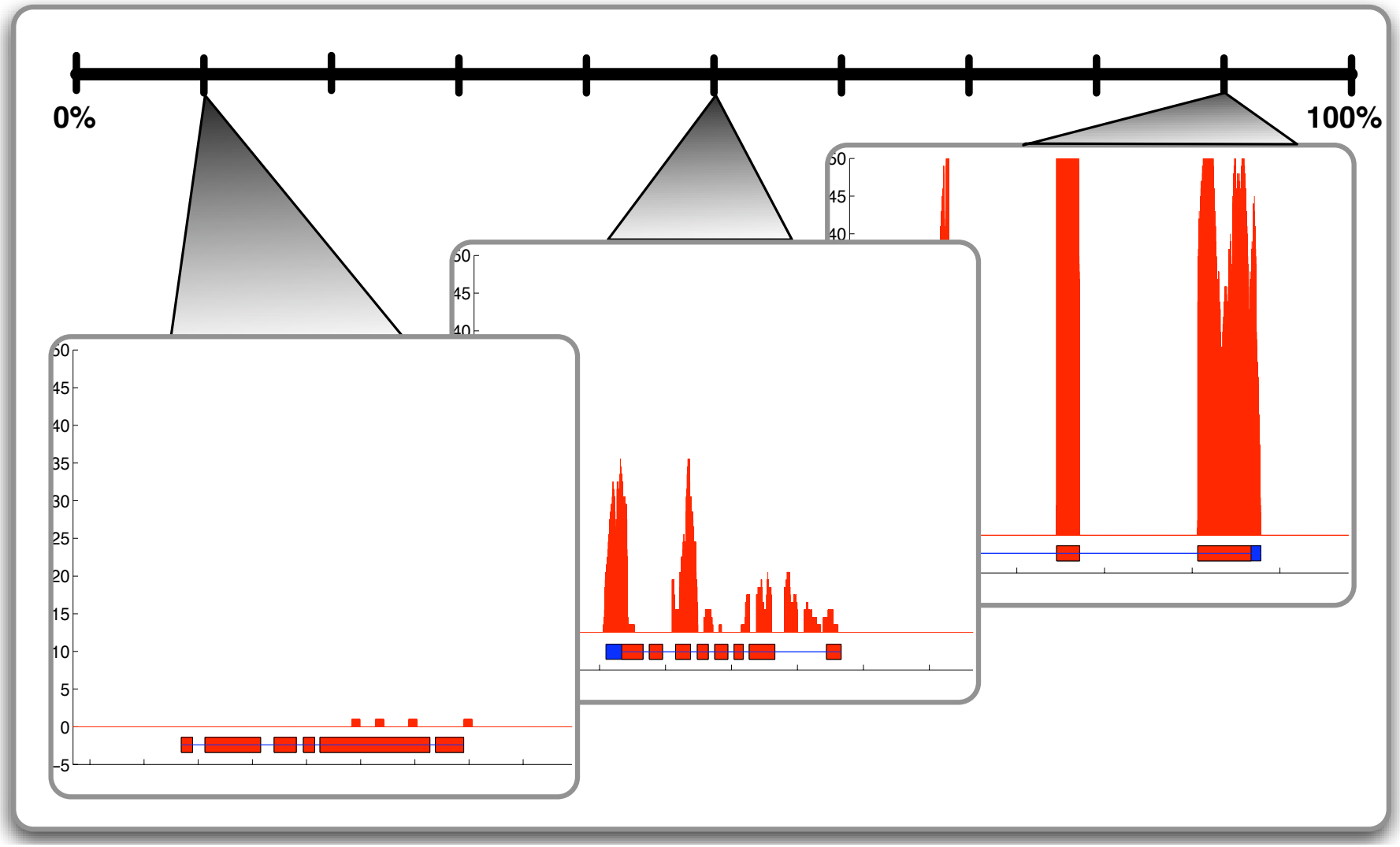
Transcript Prediction vs. Read Coverage

Annotated genes sorted by expression level

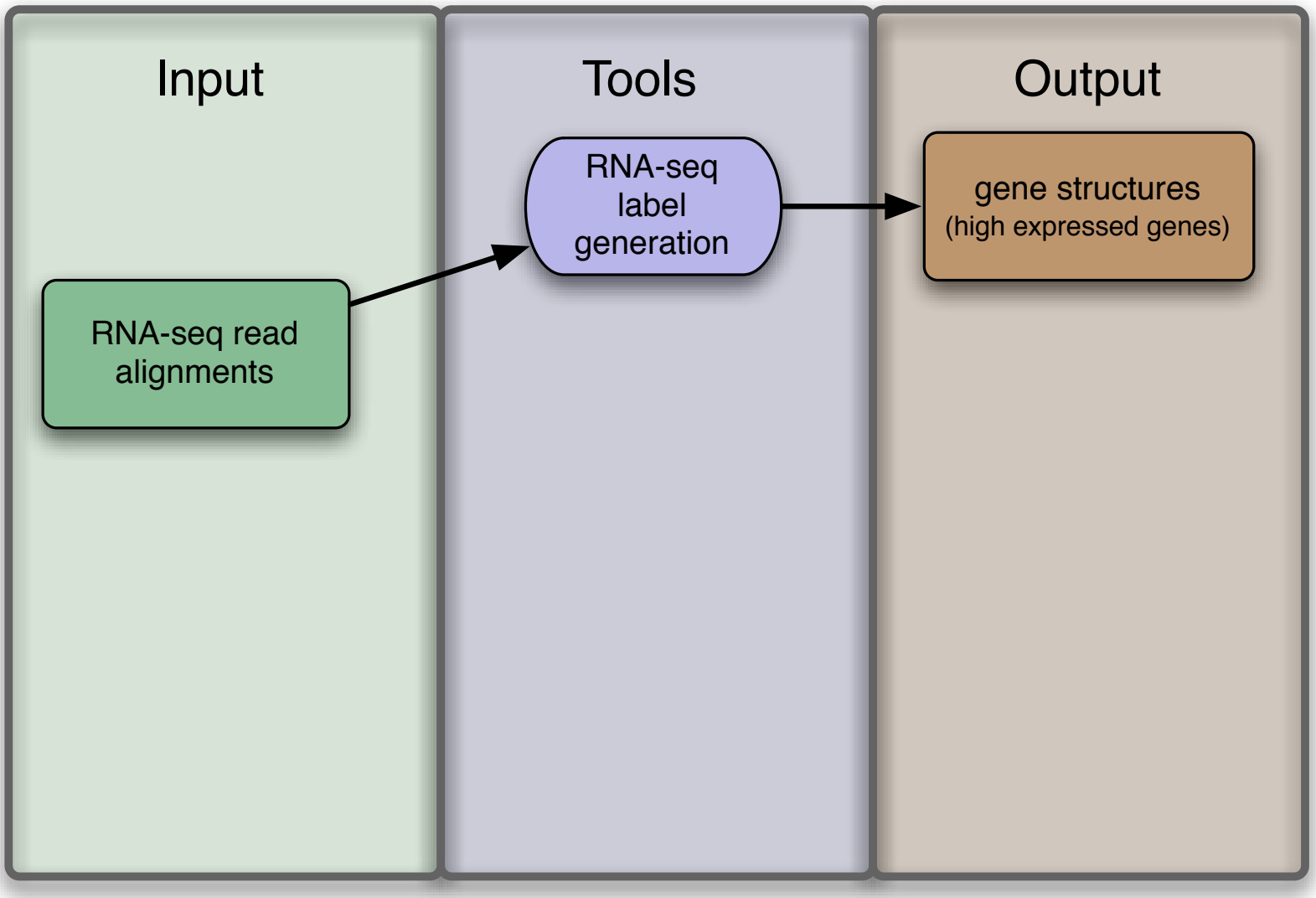


Transcript Prediction vs. Read Coverage

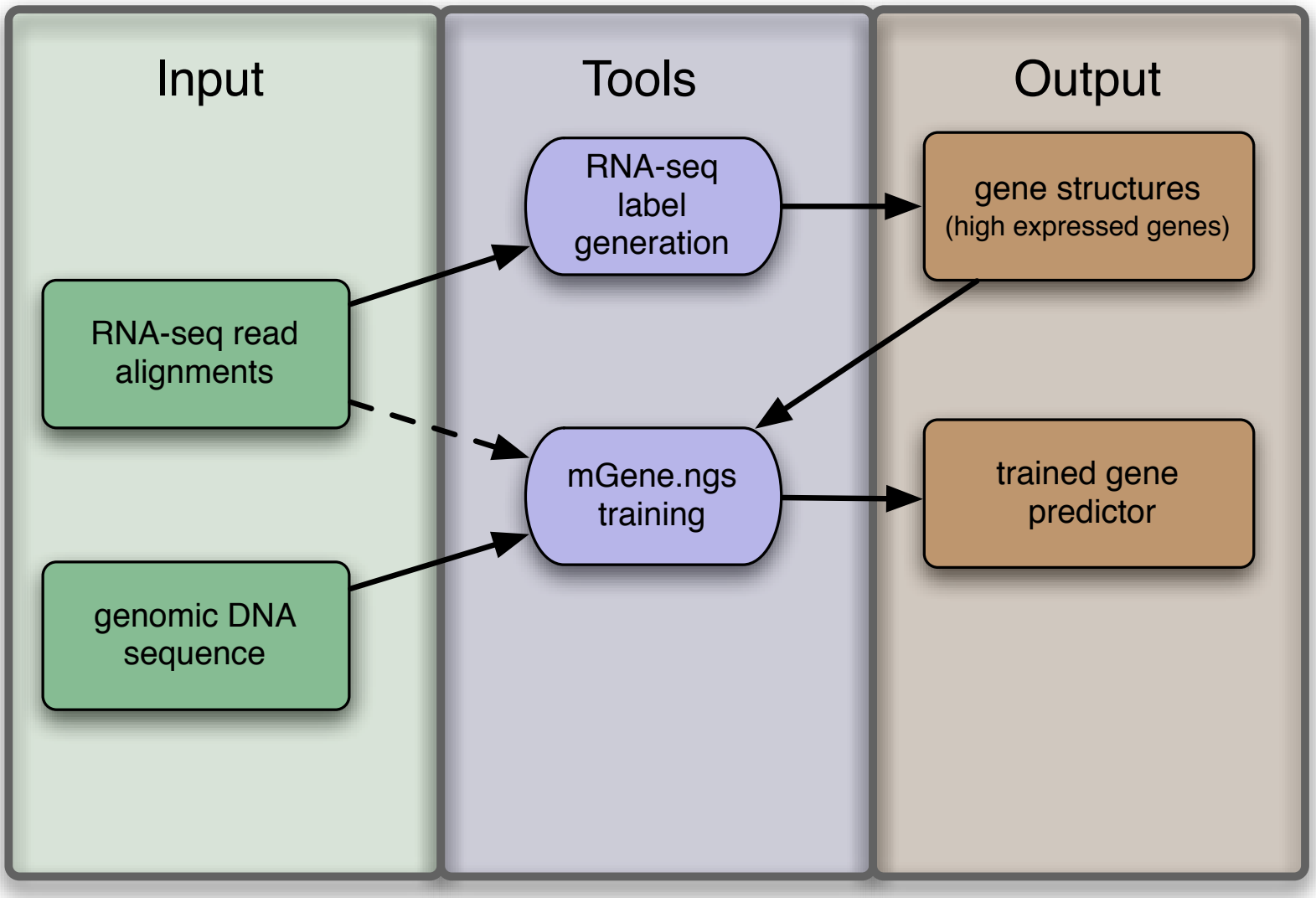
Annotated genes sorted by expression level



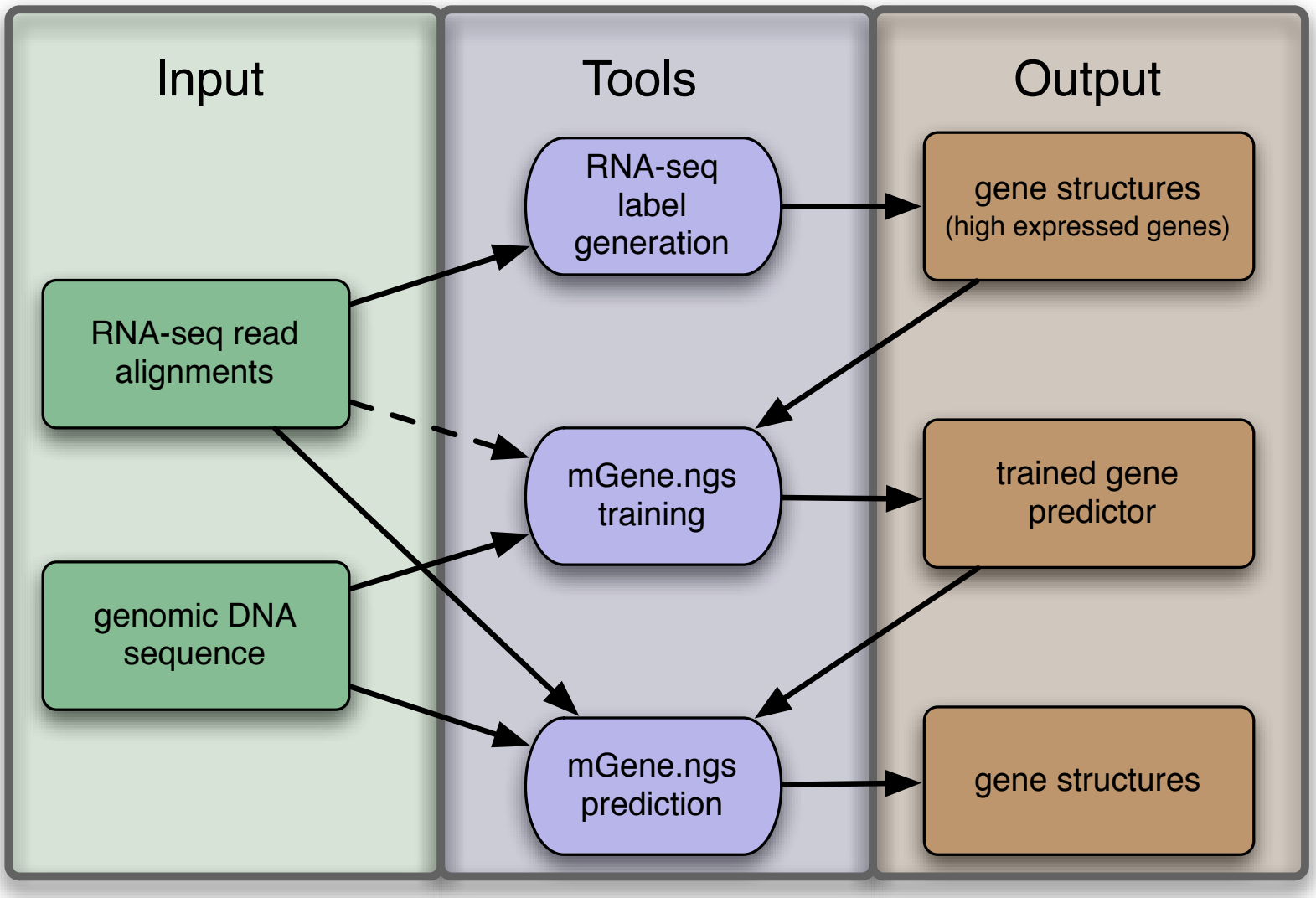
Label generation from RNA-seq data



Label generation from RNA-seq data



Label generation from RNA-seq data



Transcript Prediction

Two approaches:

1. Coverage segmentation algorithm *mTIM* for general transcripts (no coding bias/assumption)
2. Extension of *mGene* gene finding system to use RNA-seq data for protein coding transcript prediction

Transcript Prediction

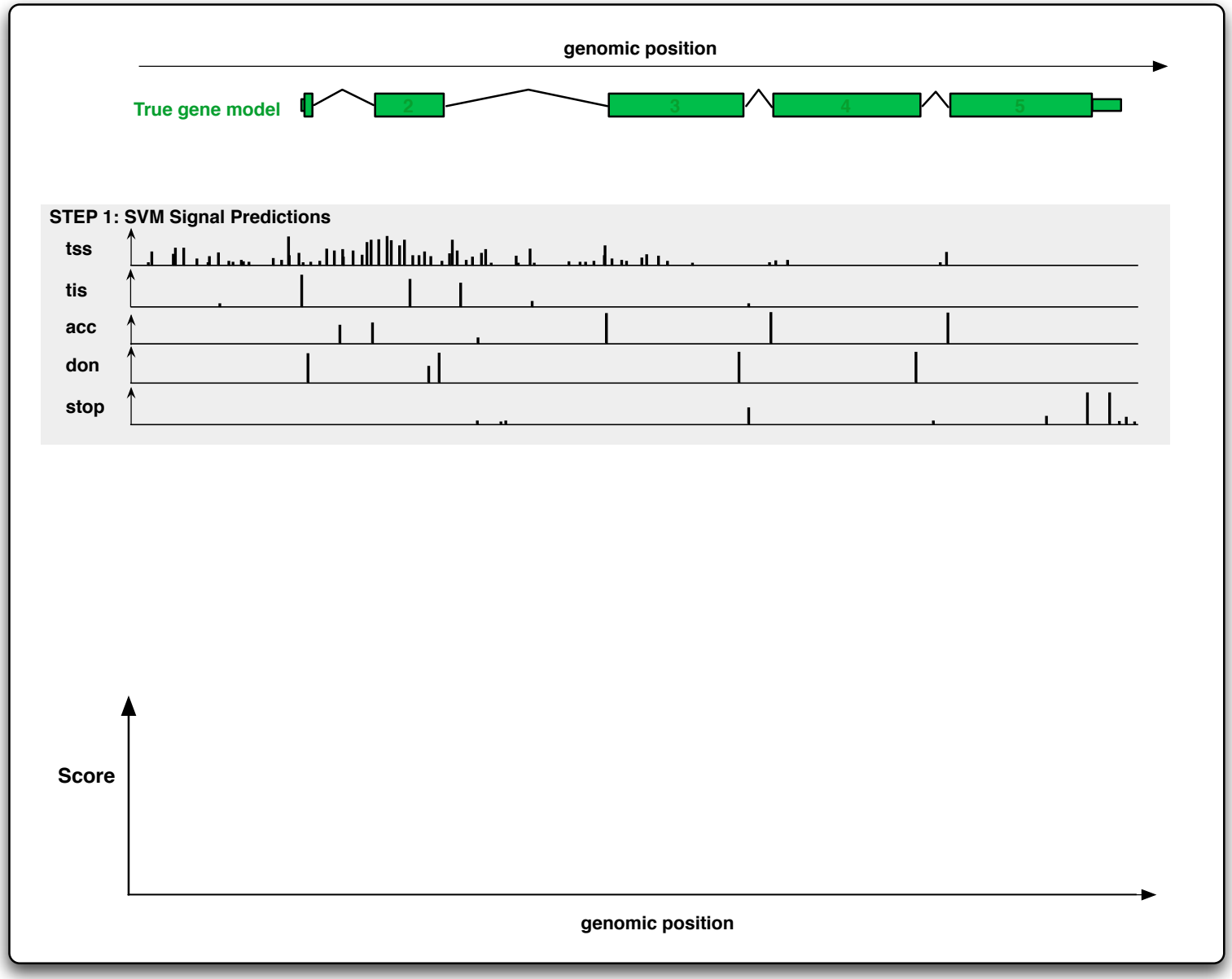
Two approaches:

1. Coverage segmentation algorithm *mTIM* for general transcripts (no coding bias/assumption)
2. Extension of *mGene* gene finding system to use RNA-seq data for protein coding transcript prediction

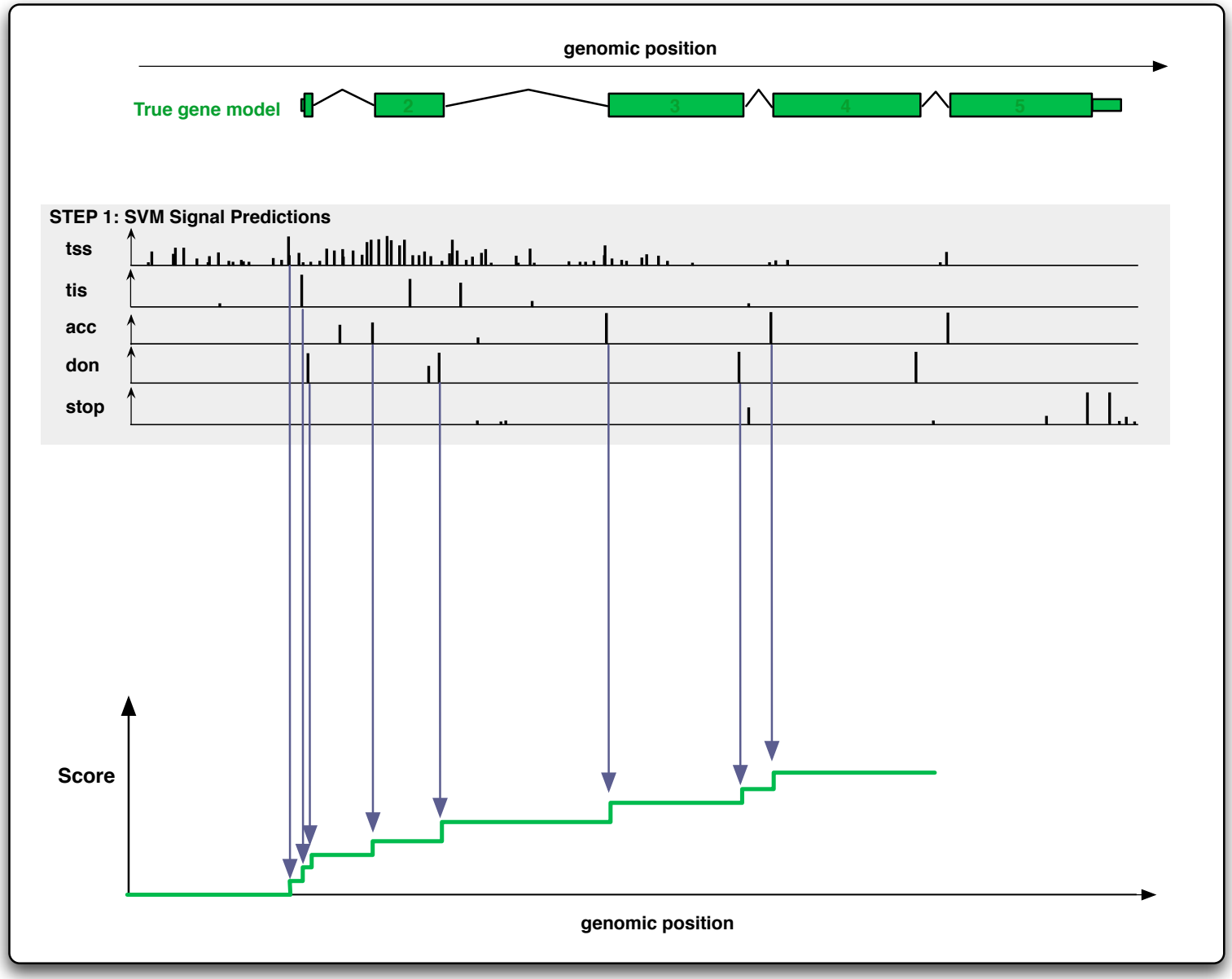
Inputs:

- Genome sequence
- Alignments in BAM format
- Preliminary annotation (training only)

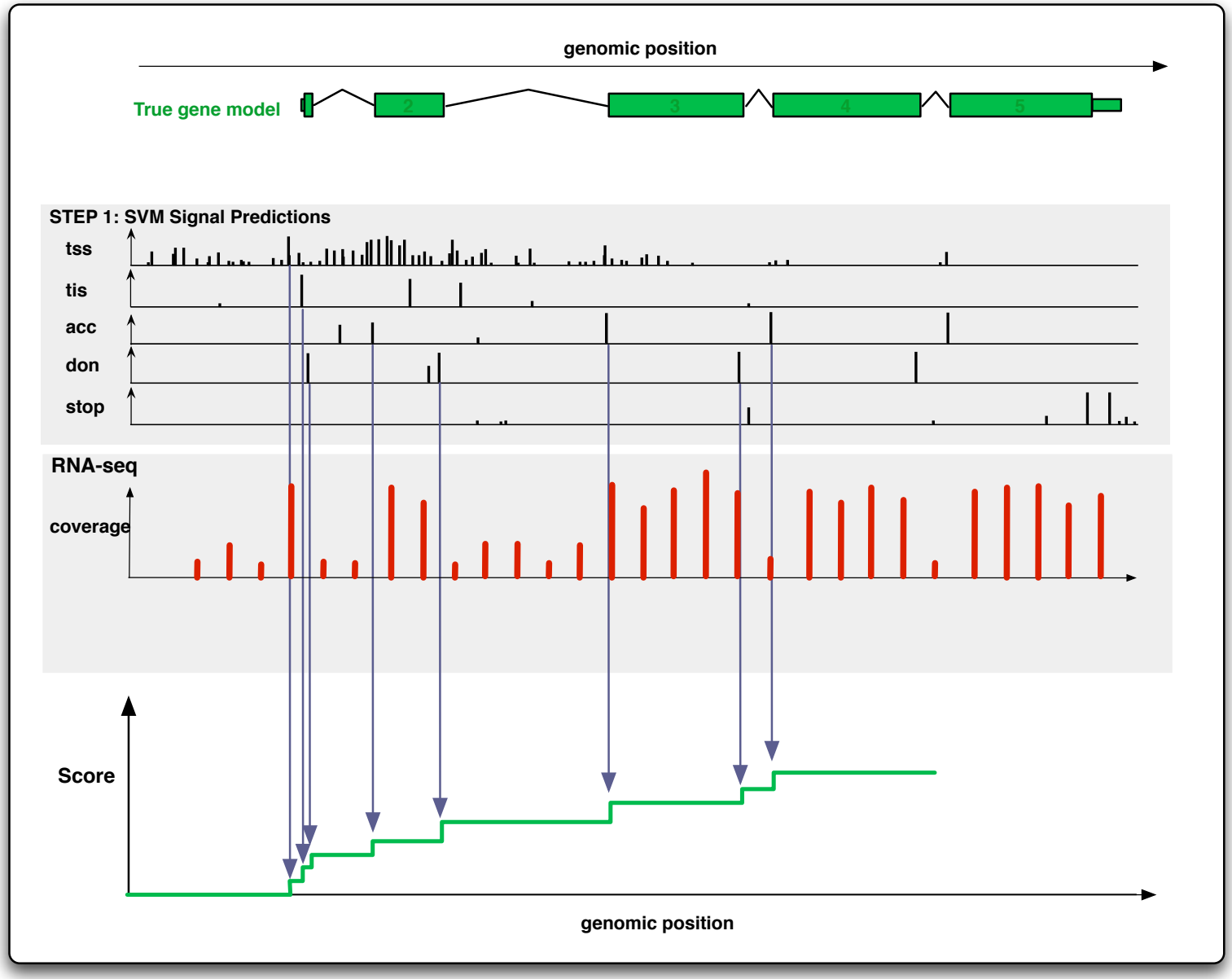
Strategy of mGene.ngs



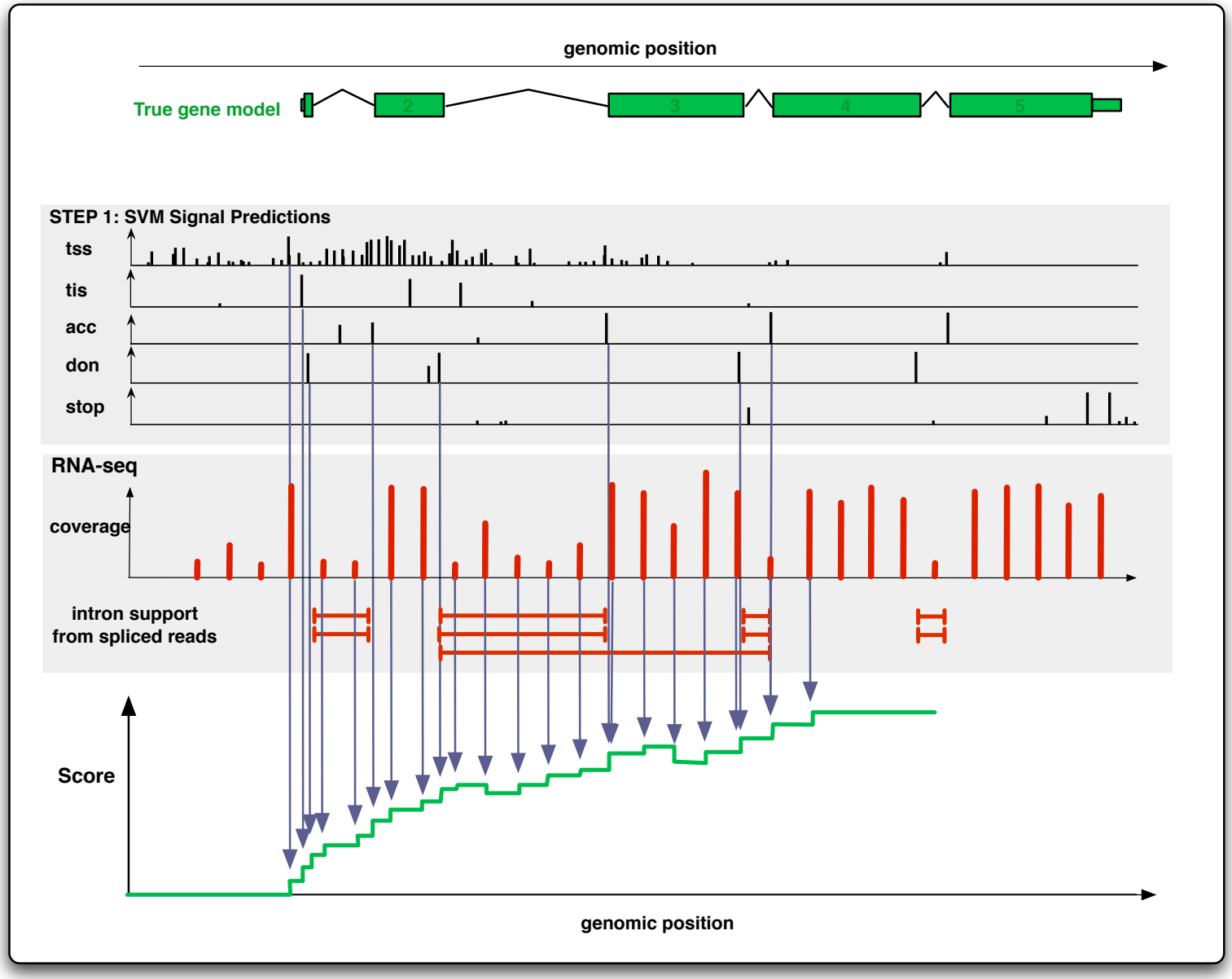
Strategy of mGene.ngs



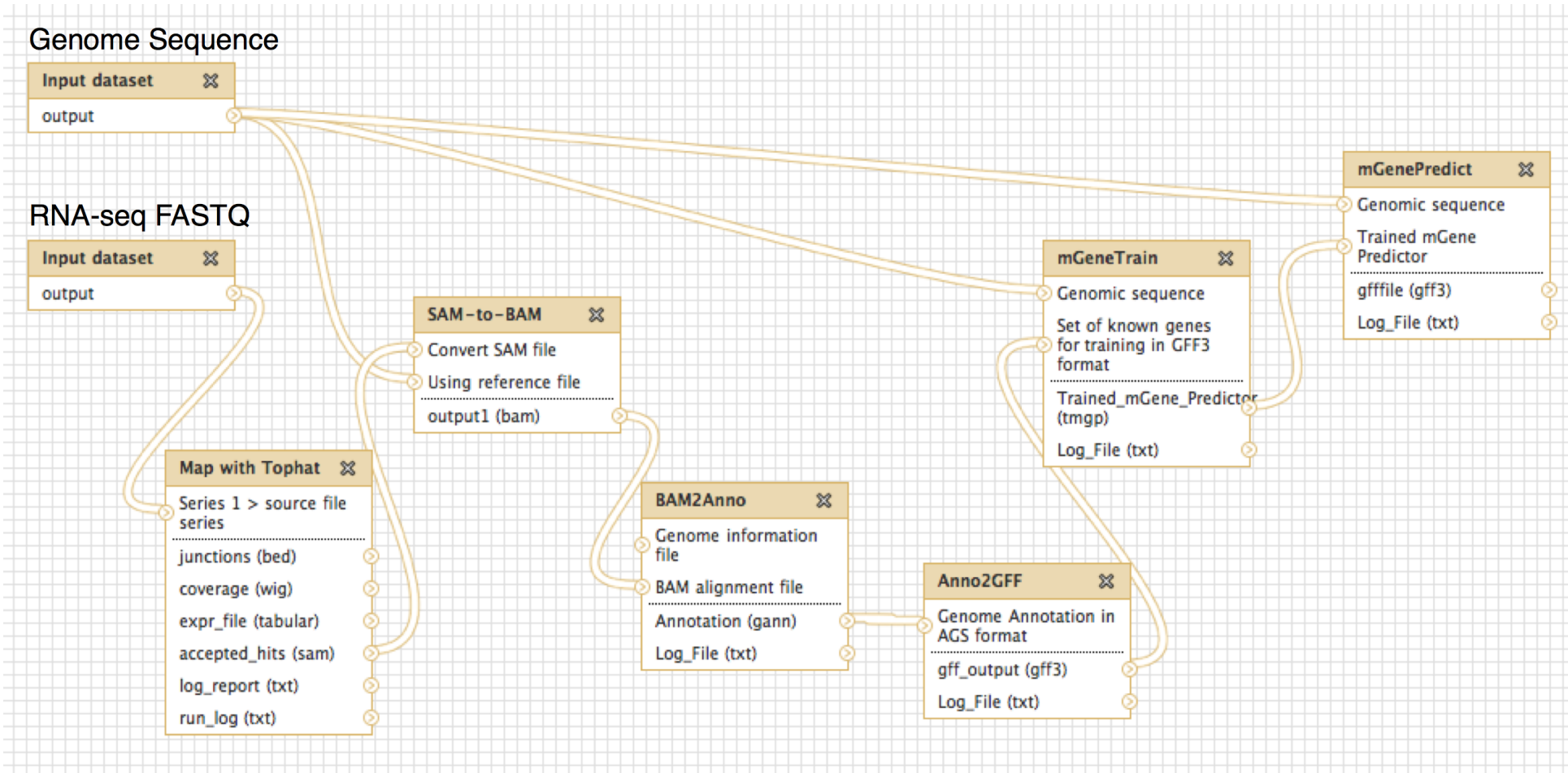
Strategy of mGene.ngs



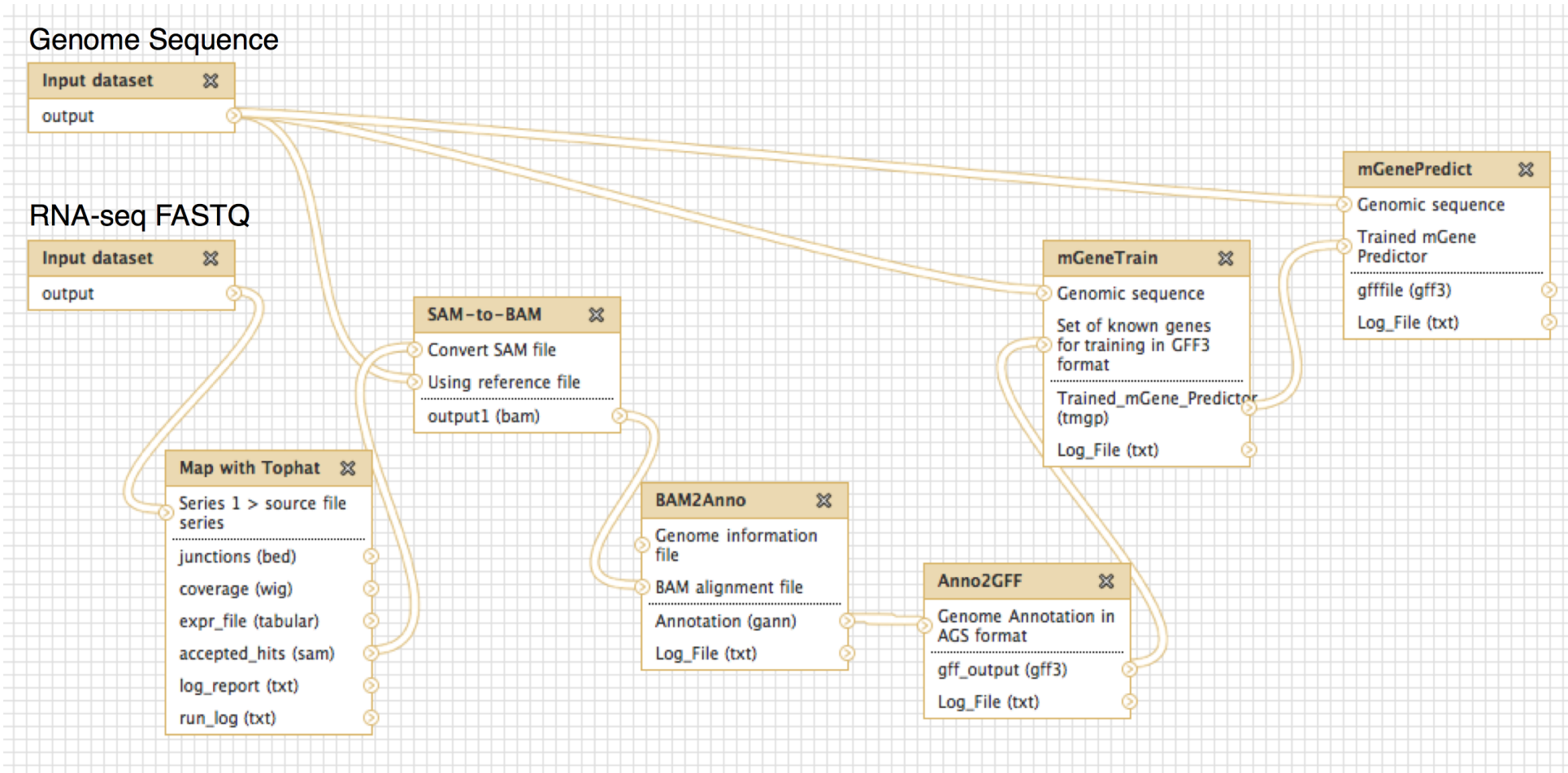
Strategy of mGene.ngs



De Novo Annotation Workflow

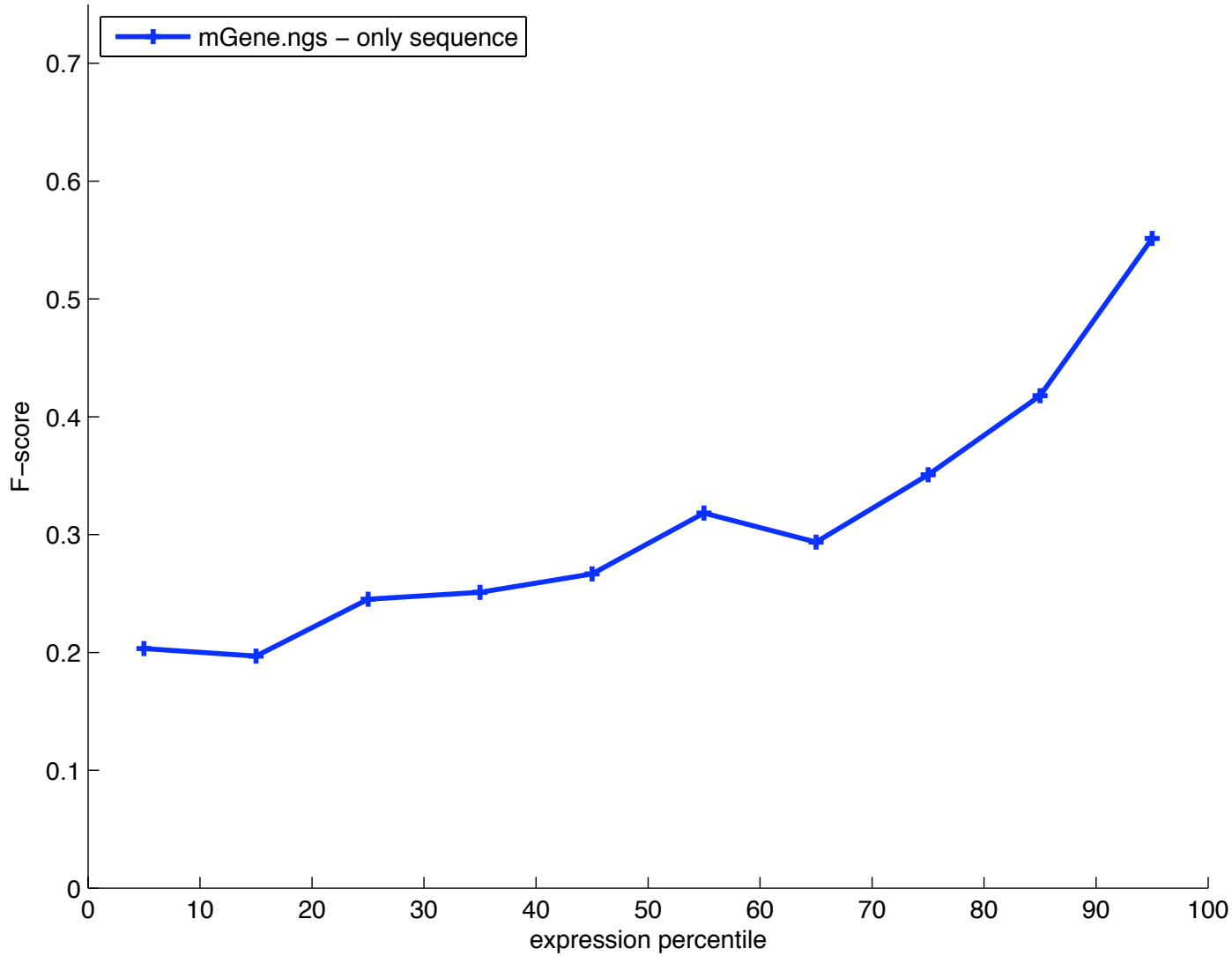


De Novo Annotation Workflow



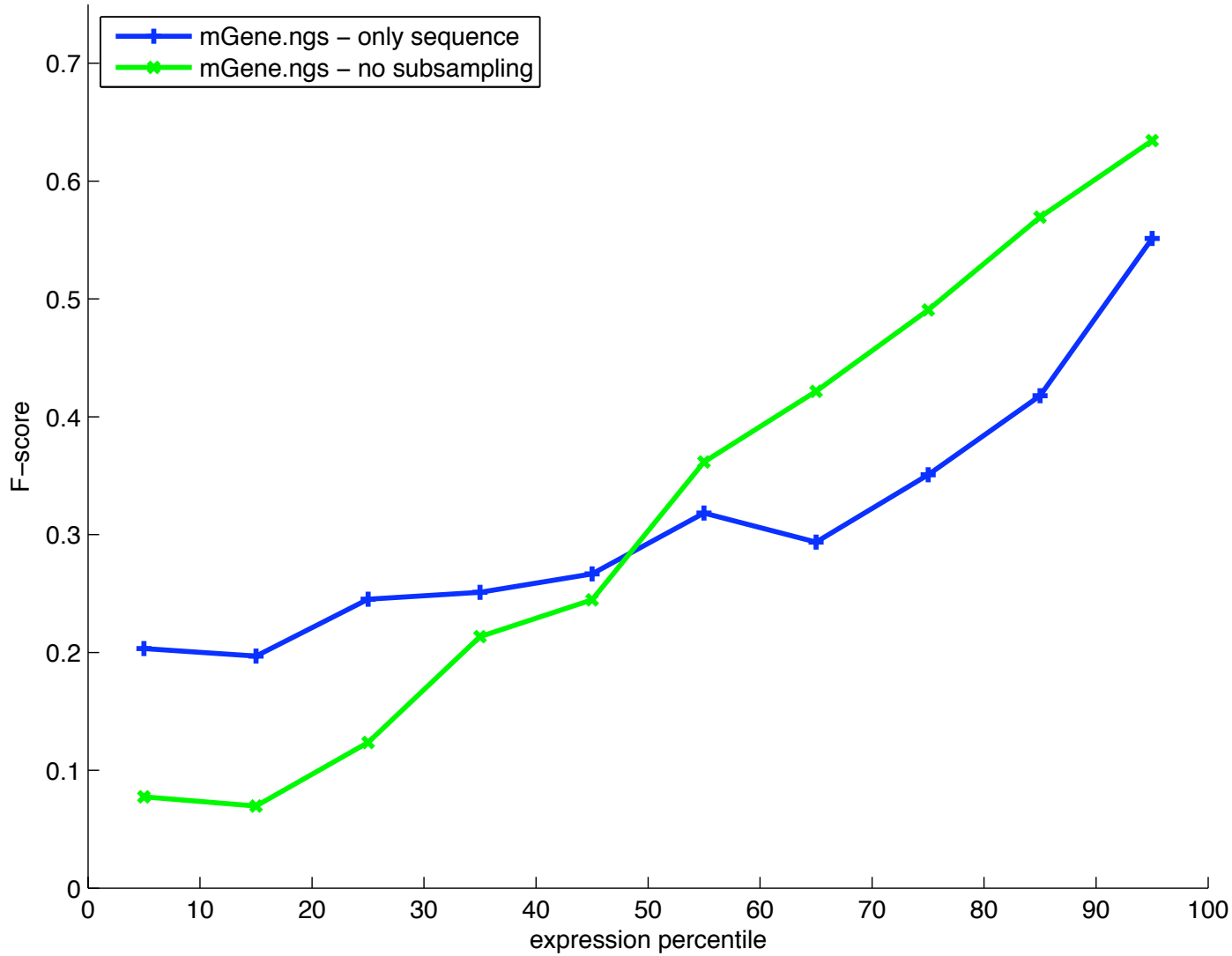
Next step: visualize using Galaxy's Trackster or a genome browser

Results for *C. elegans*



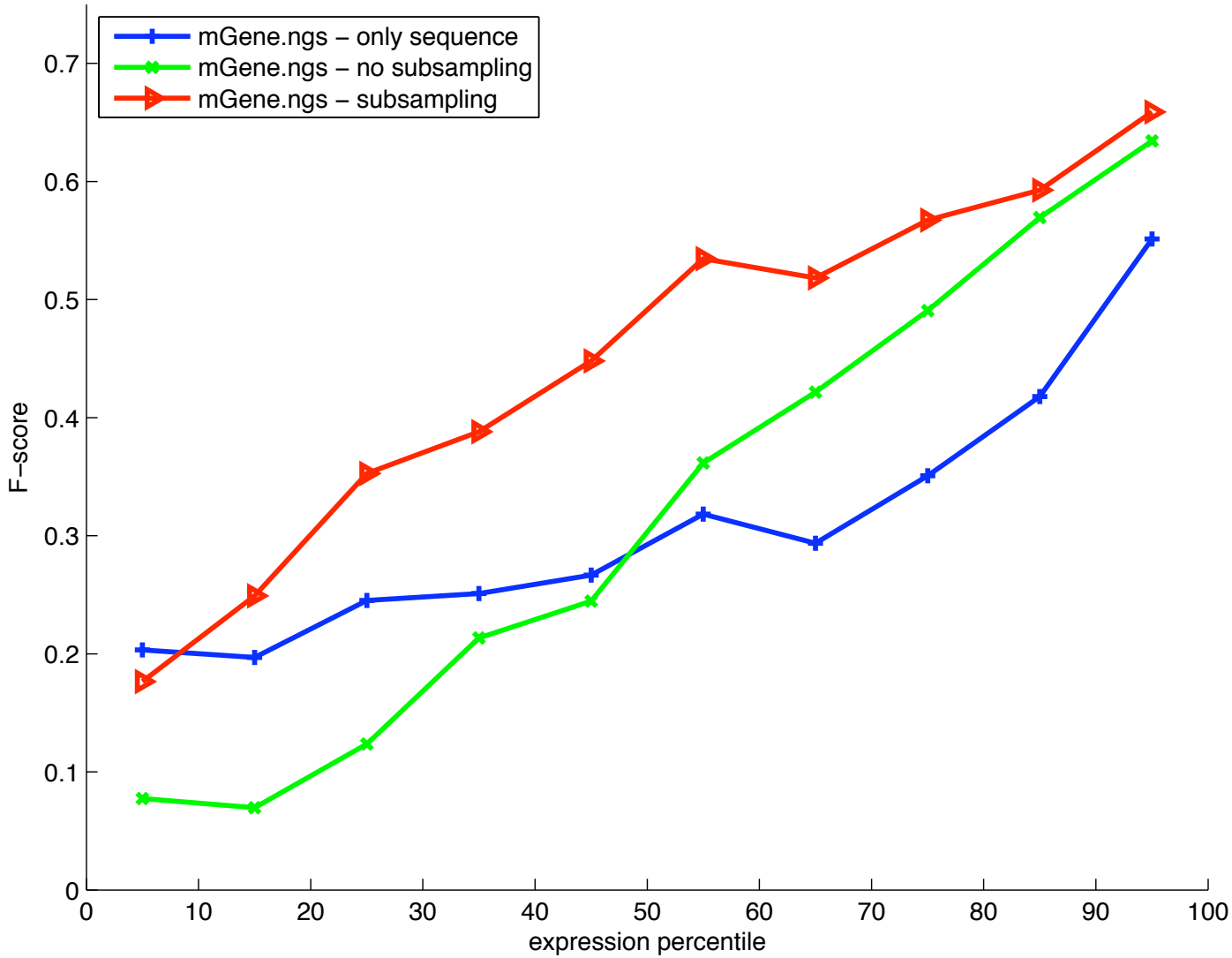
(Behr et al., 2010)

Results for *C. elegans*



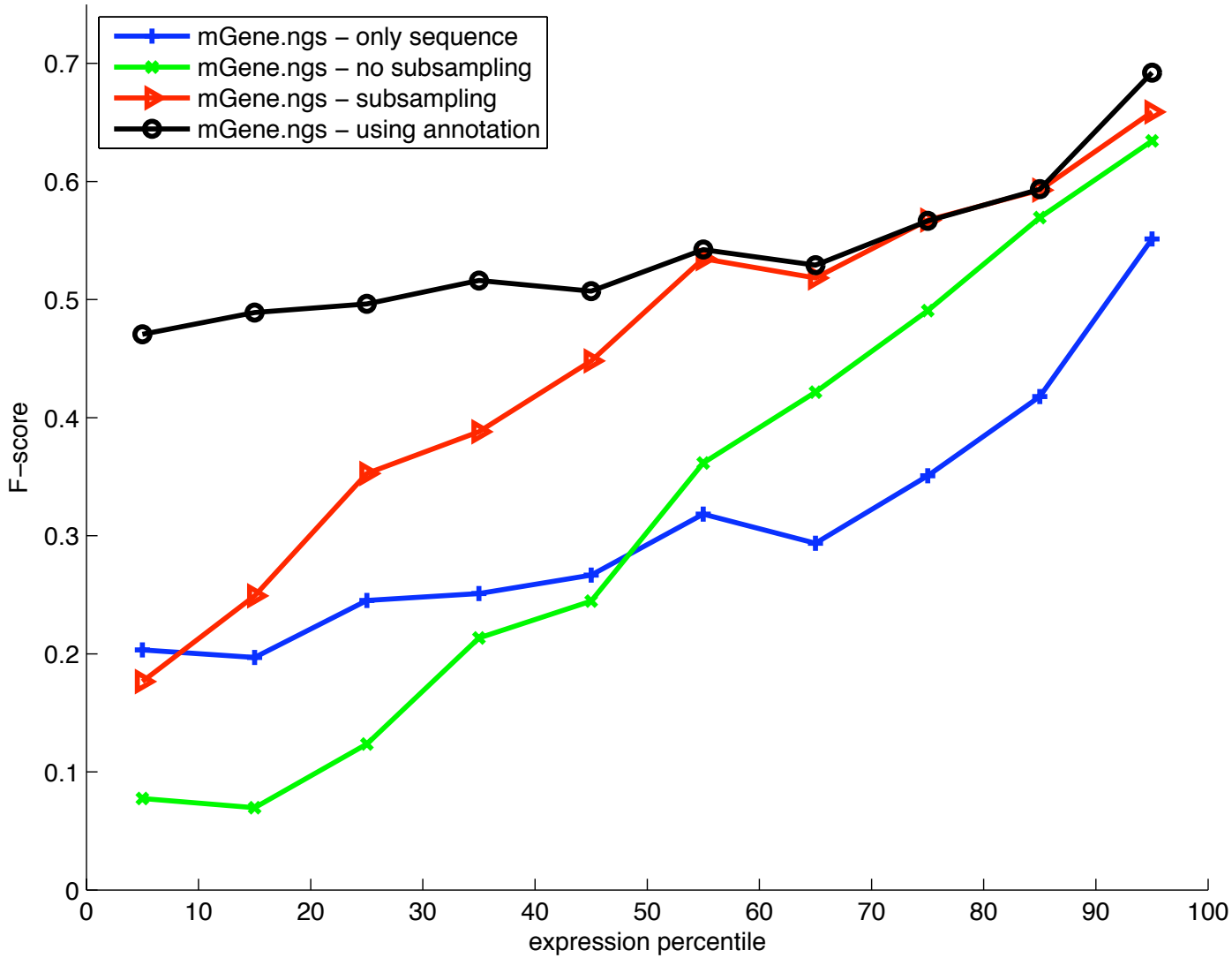
(Behr et al., 2010)

Results for *C. elegans*



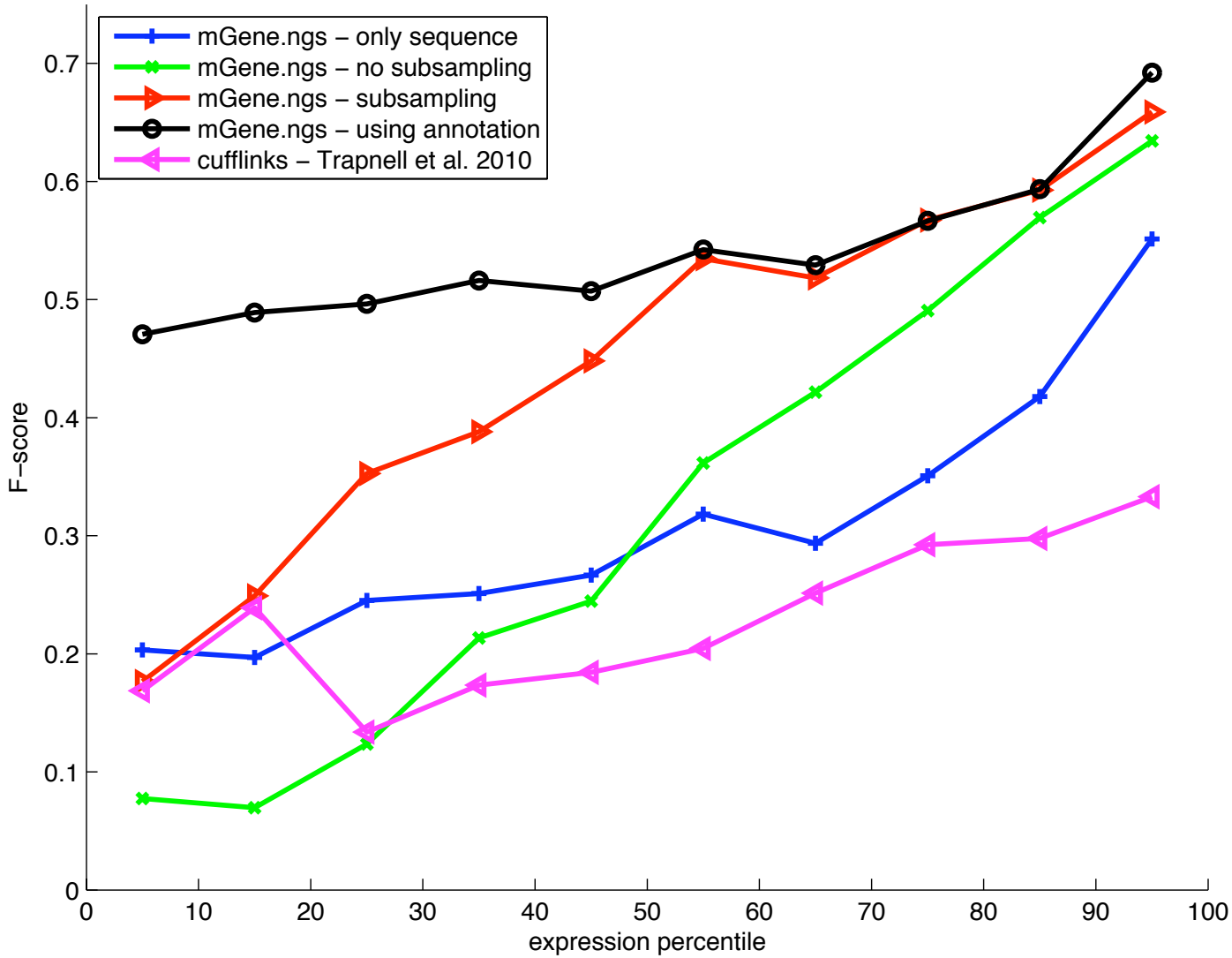
(Behr et al., 2010)

Results for *C. elegans*



(Behr et al., 2010)

Results for *C. elegans*



(Behr et al., 2010)

Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, . . .)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)

Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, . . .)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)
- Genome annotation with RNA-seq works well also for newly sequenced genomes

Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, . . .)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)
- Genome annotation with RNA-seq works well also for newly sequenced genomes
- Limitations for some tools w.r.t. genome size (RAM)
- Works well for few users in parallel, but has limits (CPU, disks)

Summary/Conclusions

- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, . . .)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)
- Genome annotation with RNA-seq works well also for newly sequenced genomes
- Limitations for some tools w.r.t. genome size (RAM)
- Works well for few users in parallel, but has limits (CPU, disks)
- Deployment on Amazon-Cloud using Galaxy's new cloud features

Summary/Conclusions



- Galaxy-Oqtans system allows
  - Read alignment (Palmapper, Tophat, . . .)
  - Quantitative transcriptome analysis (rQuant, rDiff)
  - Genome annotation (mGene, mTIM in prep.)
- Genome annotation with RNA-seq works well also for newly sequenced genomes
- Limitations for some tools w.r.t. genome size (RAM)
- Works well for few users in parallel, but has limits (CPU, disks)
- Deployment on Amazon-Cloud using Galaxy's new cloud features
- Software packages and Galaxy bindings will be made available

Acknowledgements



Fabio De Bona

Alignments



Jonas Behr

Gene finding



Regina Bohnert

Quantitation



Philipp Drewe

Diff. Testing



Andre Kahles

Testing



Vipin Sreedharan

Bug fixes

Acknowledgements



Fabio De Bona

Alignments



Jonas Behr

Gene finding



Regina Bohnert

Quantitation



Philipp Drewe

Diff. Testing



Andre Kahles

Testing



Vipin Sreedharan

Bug fixes

Funding by DFG &
Max Planck Society.

Thank you for your attention.

Poster J50, Monday