Transcriptome Analysis with Galaxy

Gunnar Rätsch



Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

May 16, 2010

Galaxy Developer Conference, CSHL

Motivation

- Group focuses on method development for
  - Sequence analysis,
  - Genome annotation,
  - Transcriptome reconstruction.

Motivation

- Group focuses on method development for
  - Sequence analysis,
  - Genome annotation,
  - Transcriptome reconstruction.
- Methods are often based on non-trivial applications of machine learning applications to CompBio problems

  Pro: High accuracy, adaptive to data

  Con: Usage more complex, need more resources

Motivation

- Group focuses on method development for
  - Sequence analysis,
  - Genome annotation,
  - Transcriptome reconstruction.
- Methods are often based on non-trivial applications of machine learning applications to CompBio problems

  Pro: High accuracy, adaptive to data

  Con: Usage more complex, need more resources
- Get many request for performing analyses using these tools, which go beyond the groups capacity

Motivation

- Group focuses on method development for
  - Sequence analysis,
  - Genome annotation,
  - Transcriptome reconstruction.
- Methods are often based on non-trivial applications of machine learning applications to CompBio problems

  Pro: High accuracy, adaptive to data

  Con: Usage more complex, need more resources
- Get many request for performing analyses using these tools, which go beyond the groups capacity
- We use Galaxy as a tool distribution platform to let others
  - benefit from our tools for their analyses ...
  - ... without the need to install or to worry about dependencies.

Tool Development Pipeline

For any new tool,

1. Develop a prototype in C++, Python, or Matlab, illustrate its usefulness, write method paper.

Tool Development Pipeline

For any new tool,

1. Develop a prototype in C++, Python, or Matlab, illustrate its usefulness, write method paper.
2. Integrate prototype into Galaxy:
   - 2.1 Define and refine interfaces to existing data types and tools
   - 2.2 Develop command-line versions of modularized prototype
   - 2.3 Add modules to tool menu
   - 2.4 Documentation, example workflows, example data
   - 2.5 (optional) Write paper, e.g., in NAR Webserver Issue

Tool Development Pipeline

For any new tool,

1. Develop a prototype in C++, Python, or Matlab, illustrate its usefulness, write method paper.
2. Integrate prototype into Galaxy:
   - 2.1 Define and refine interfaces to existing data types and tools
   - 2.2 Develop command-line versions of modularized prototype
   - 2.3 Add modules to tool menu
   - 2.4 Documentation, example workflows, example data
   - 2.5 (optional) Write paper, e.g., in NAR Webserver Issue
3. Software release of modularized tools
   - 3.1 Open source license
   - 3.2 Including Galaxy-binding (sub-directory `./galaxy`)

Tool Development Pipeline

For any new tool,

1. Develop a prototype in C++, Python, or Matlab, illustrate its usefulness, write method paper.
2. Integrate prototype into Galaxy:
   - 2.1 Define and refine interfaces to existing data types and tools
   - 2.2 Develop command-line versions of modularized prototype
   - 2.3 Add modules to tool menu
   - 2.4 Documentation, example workflows, example data
   - 2.5 (optional) Write paper, e.g., in NAR Webserver Issue
3. Software release of modularized tools
   - 3.1 Open source license
   - 3.2 Including Galaxy-binding (sub-directory `./galaxy`)
4. Bug-fixes & support

Tools on http://galaxy.fml.mpg.de



- Machine Learning toolbox "EasySVM" (SVMs, sequence analysis) used in tutorials (Ben-Hur et al., 2008)

Tools on http://galaxy.fml.mpg.de



- Machine Learning toolbox "EasySVM" (SVMs, sequence analysis) used in tutorials (Ben-Hur et al., 2008)
- Predictors for TSS and splice site prediction (Sonnenburg et al., 2006, 2007)

Tools on http://galaxy.fml.mpg.de



- Machine Learning toolbox "EasySVM" (SVMs, sequence analysis) used in tutorials (Ben-Hur et al., 2008)
- Predictors for TSS and splice site prediction (Sonnenburg et al., 2006, 2007)
- Promoter-Analysis "Kirmes" (Schultheiss et al., 2009)

Tools on http://galaxy.fml.mpg.de



- Machine Learning toolbox "EasySVM" (SVMs, sequence analysis) used in tutorials (Ben-Hur et al., 2008)
- Predictors for TSS and splice site prediction (Sonnenburg et al., 2006, 2007)
- Promoter-Analysis "Kirmes" (Schultheiss et al., 2009)
- Gene finding system "mGene" (Schweikert et al., 2009d,a)
  - Train/Predict TSS, TIS, STOP, cleavage, and splice sites
  - Train/Predict gene finder for new genomes from scratch

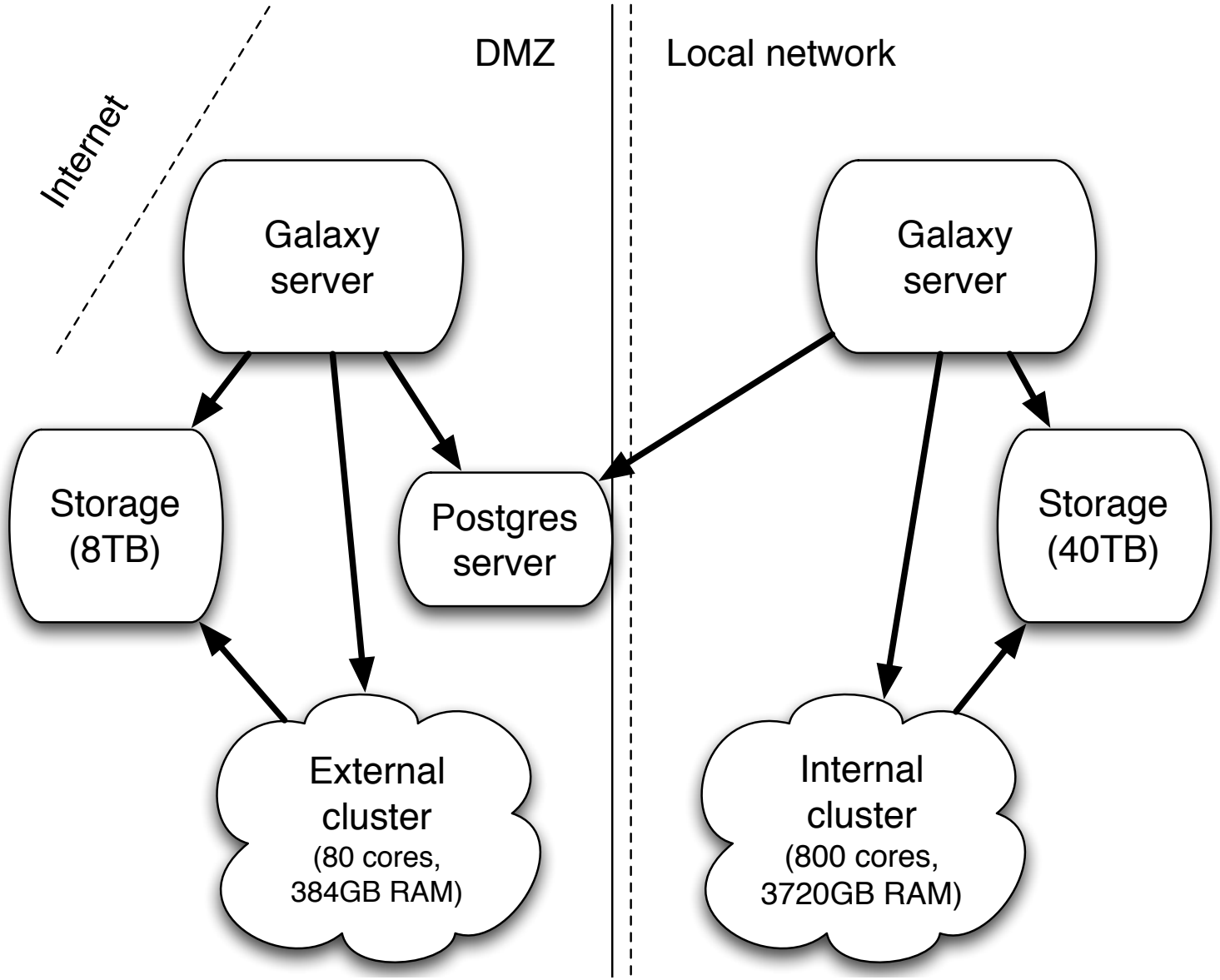
Tools on http://galaxy.fml.mpg.de

- Machine Learning toolbox "EasySVM" (SVMs, sequence analysis) used in tutorials (Ben-Hur et al., 2008)
- Predictors for TSS and splice site prediction (Sonnenburg et al., 2006, 2007)
- Promoter-Analysis "Kirmes" (Schultheiss et al., 2009)
- Gene finding system "mGene" (Schweikert et al., 2009d,a)
  - Train/Predict TSS, TIS, STOP, cleavage, and splice sites
  - Train/Predict gene finder for new genomes from scratch
- RNA-seq Toolbox
  - Spliced read alignment "Palmapper" (Rätsch et al., 2010; Jean et al., 2010)
  - Transcript quantitation "rQuant" (Bohnert et al., 2009; Bohnert and Rätsch, 2010)
  - Transcript reconstruction "mGene.NGS" and "mTiM"
  - Differential expression testing "Isotest" (Stegle et al., 2010)
  - Alignment accuracy evaluation

Data Types

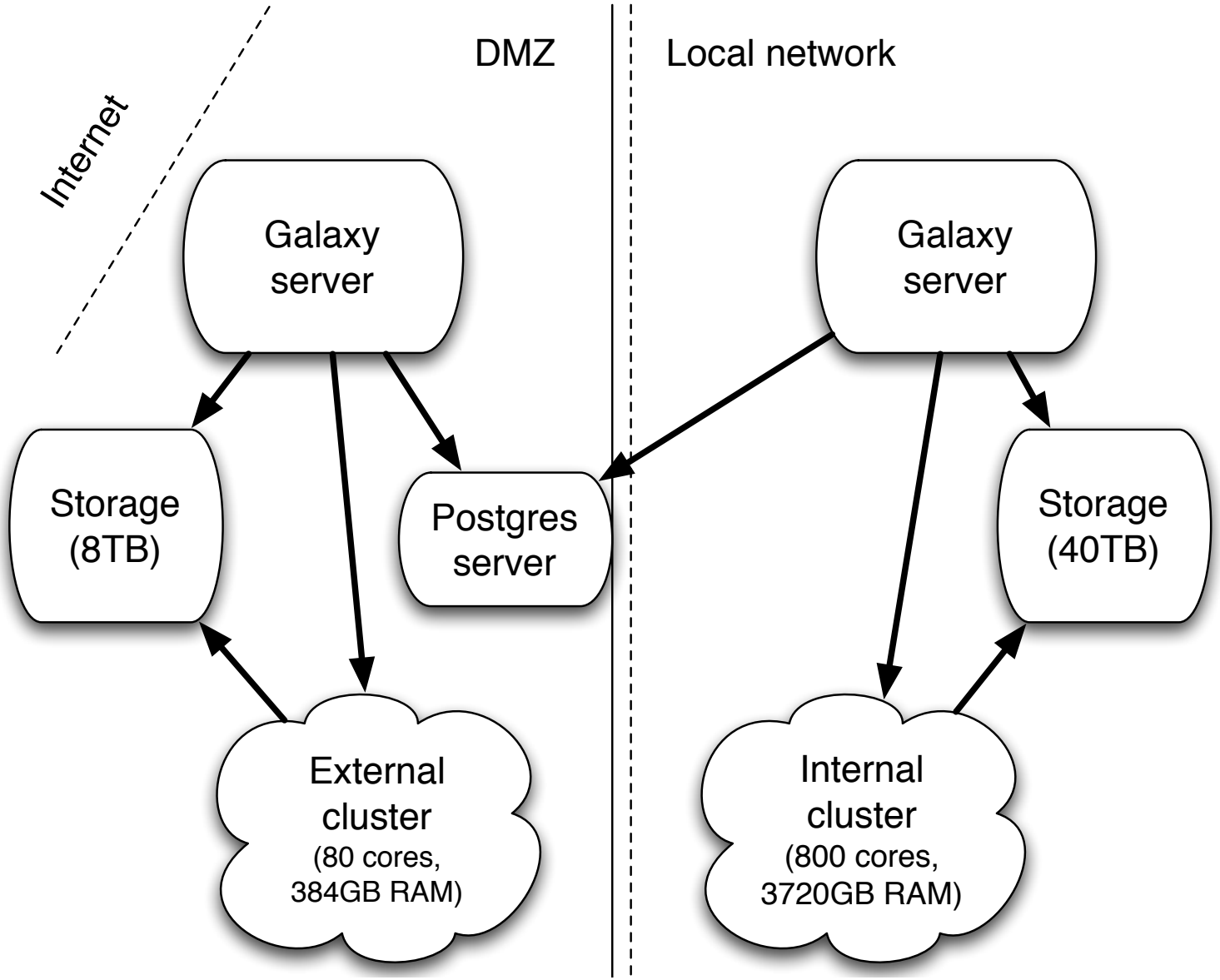
- Standard data types: FASTA, GFF3, FASTQ, SAM, BAM
- Custom data types:
  - Genome Information Object for fast access to FASTA files
  - Genome Annotation Object (binary) for faster access to annotation files (like GFF3)
  - Signal Prediction Format (SPF): bed-like and binary
  - Trained predictors (signals, mGene, ...) in proprietary format
- Composite objects:
  - tool-specific information (`dataset_XXXX.dat`)
  - tool-specific files (`dataset_XXXX_files`)

Galaxy Setup

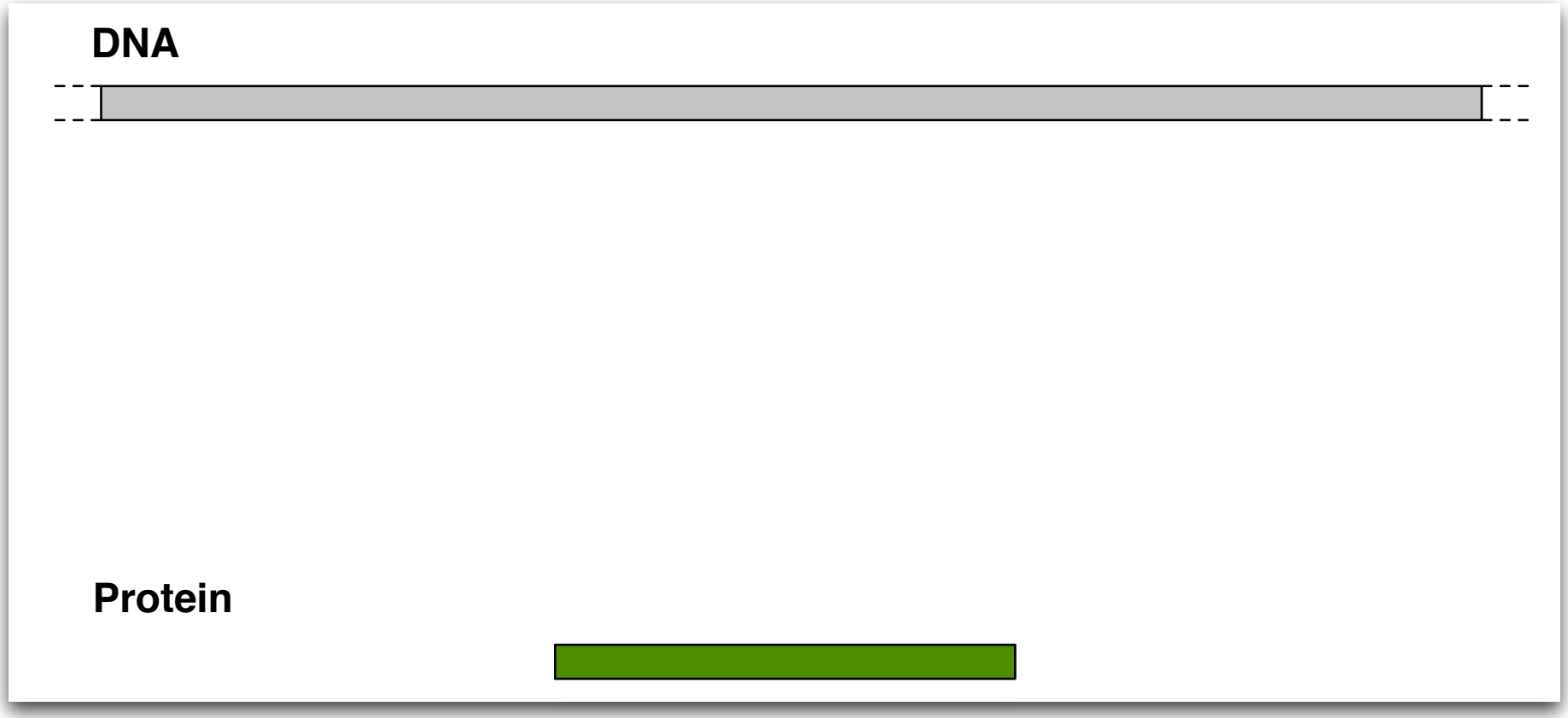


Galaxy Setup

How to synchronize storage (internal/external)?

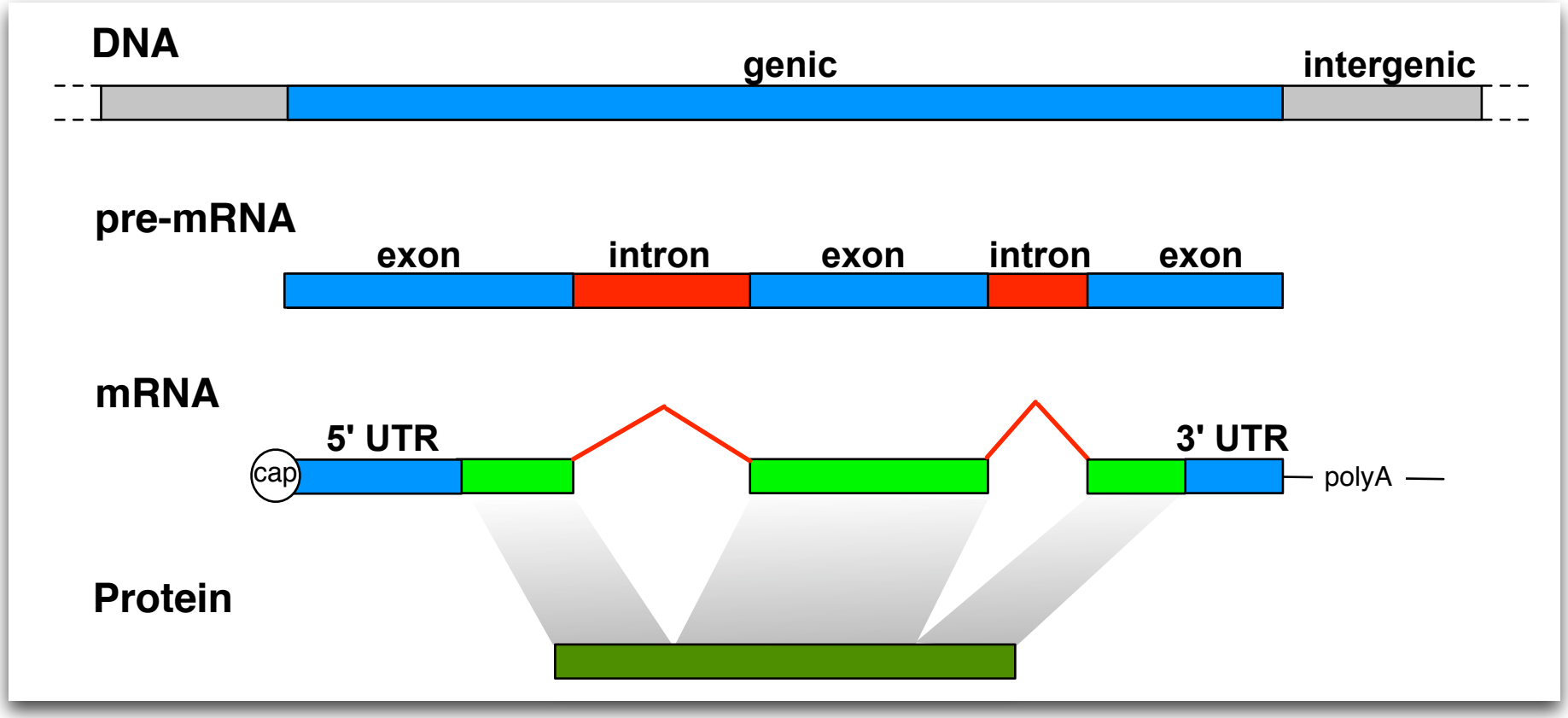


Computational Gene Finding



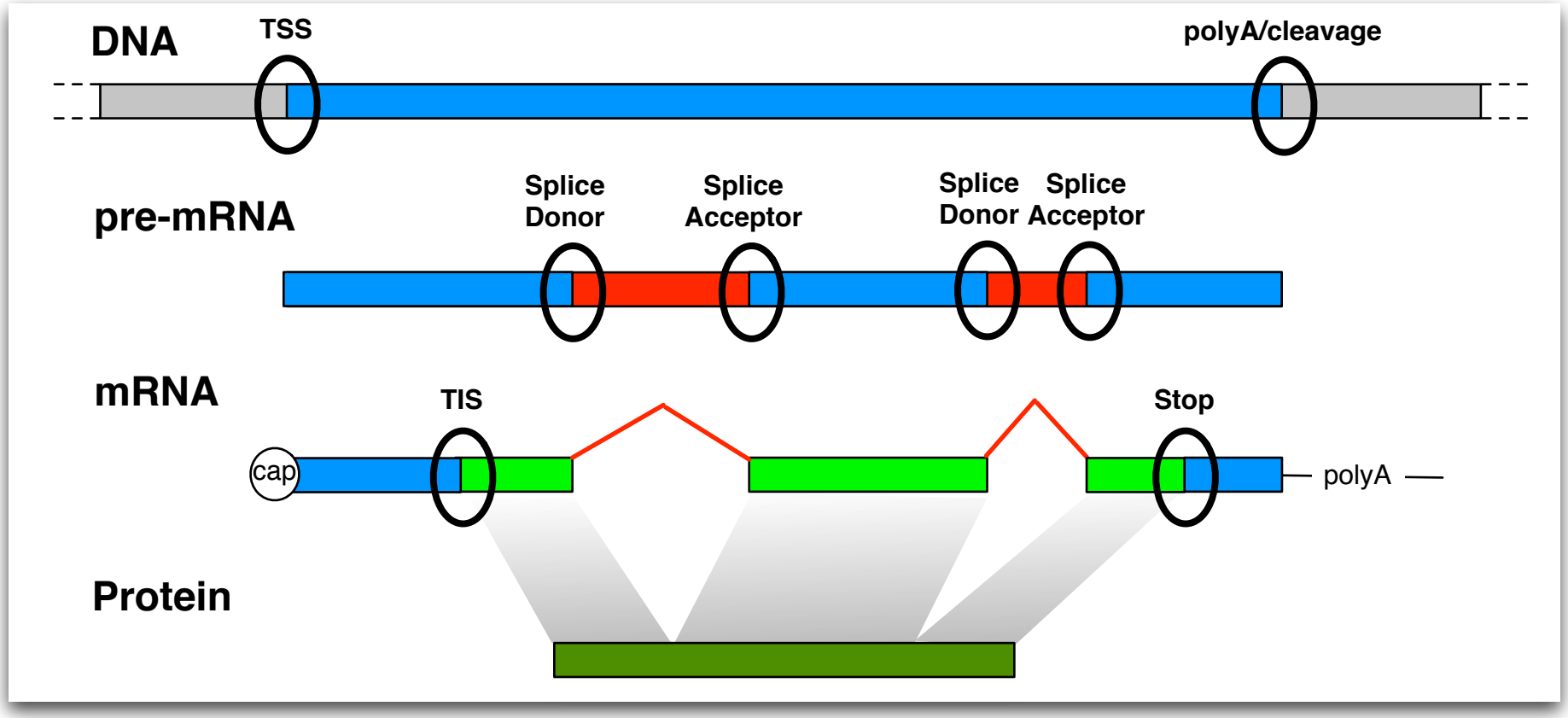
- Given a piece of DNA sequence
- Predict proteins (or non-coding RNAs)

Computational Gene Finding

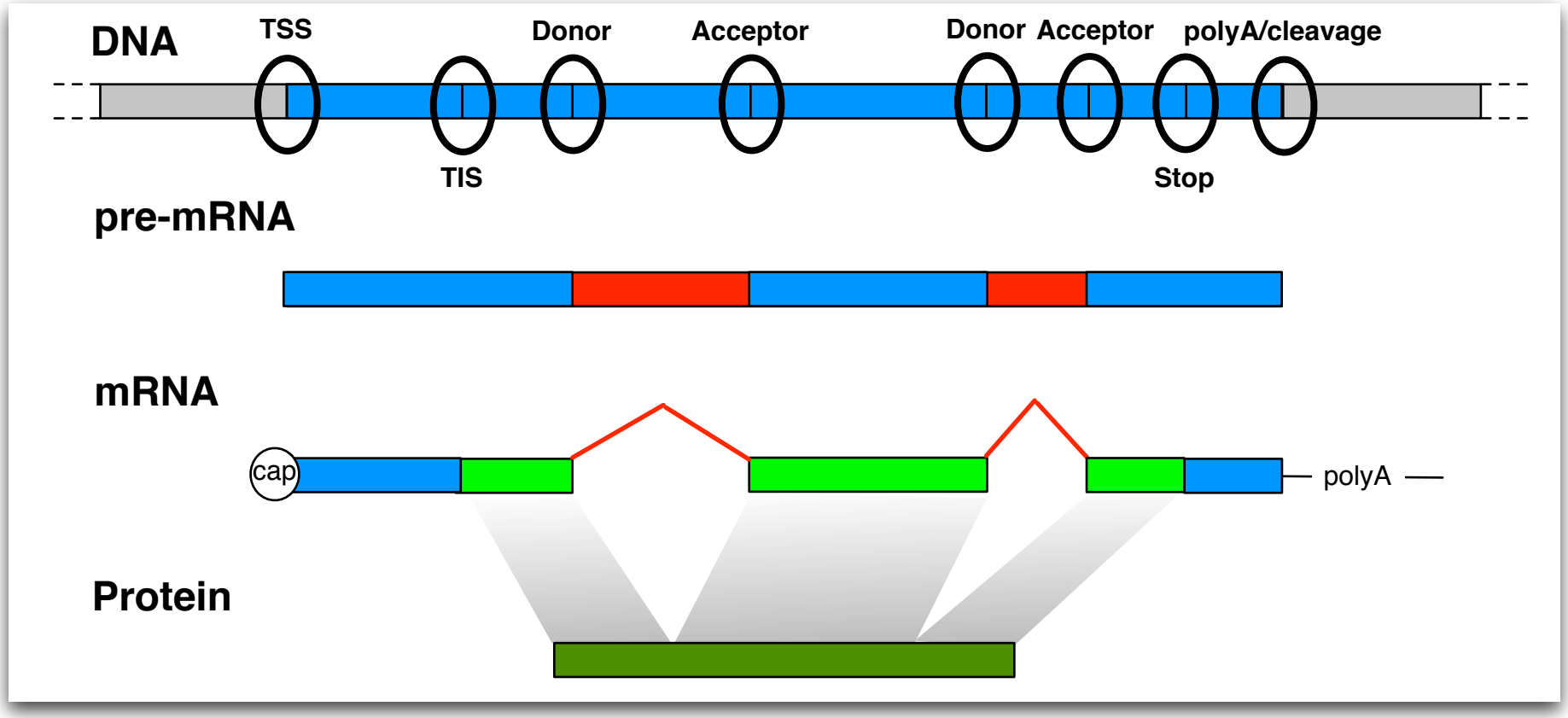


- Given a piece of DNA sequence
- Predict the correct corresponding **label sequence** with labels "intergenic", "exon", "intron", "5' UTR", etc.

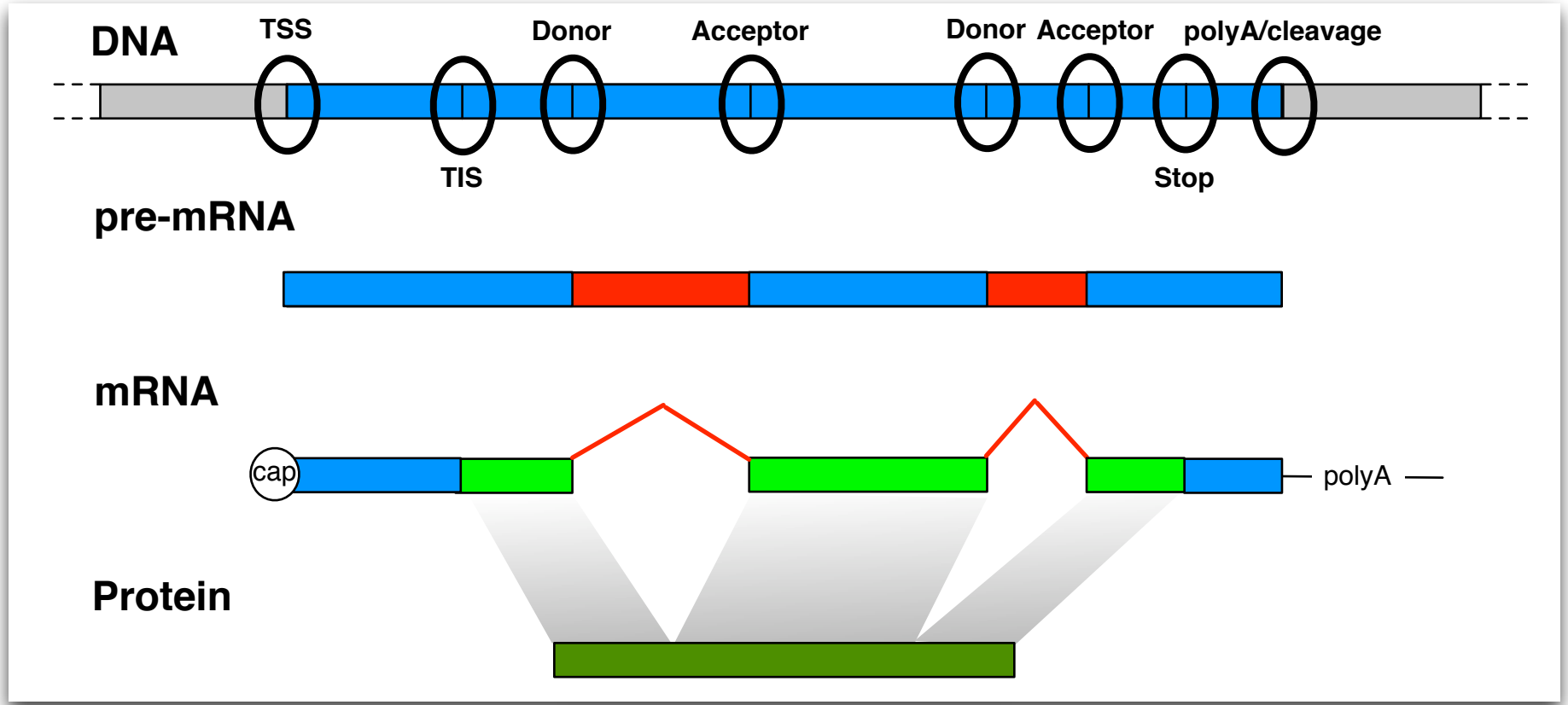
Computational Gene Finding

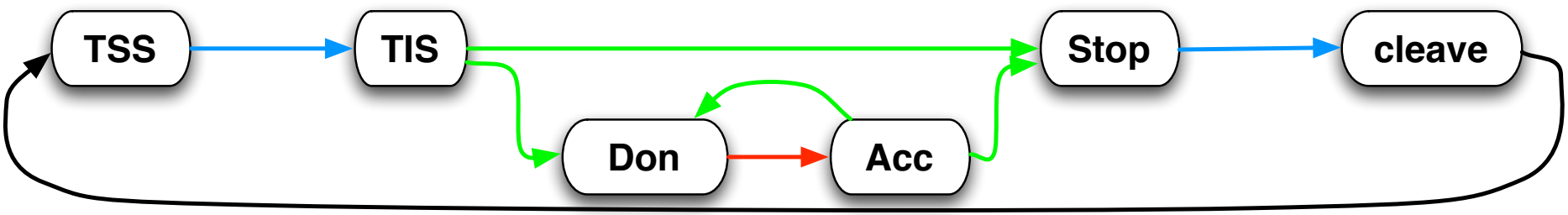


Computational Gene Finding

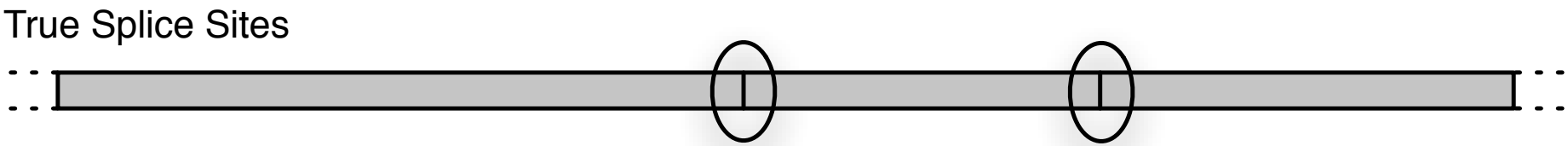


Computational Gene Finding

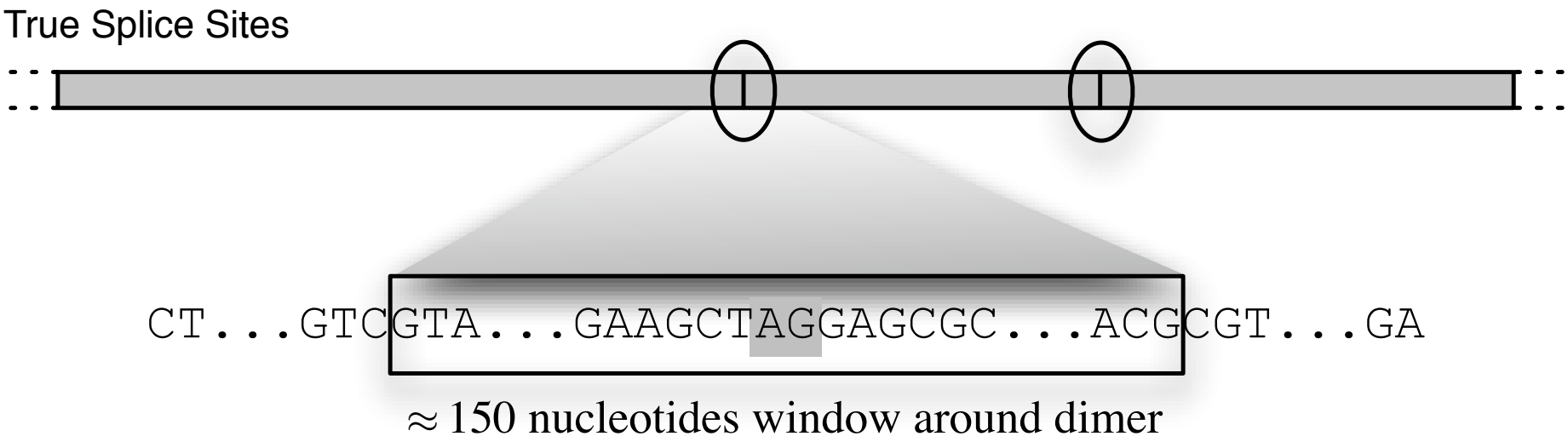




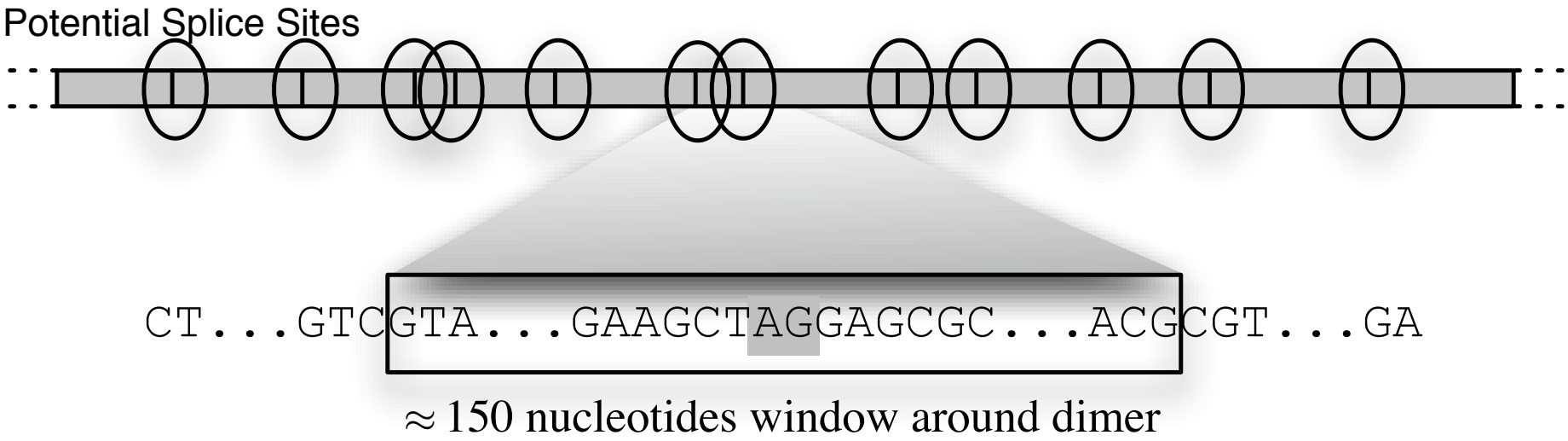
Example: Splice Site Recognition



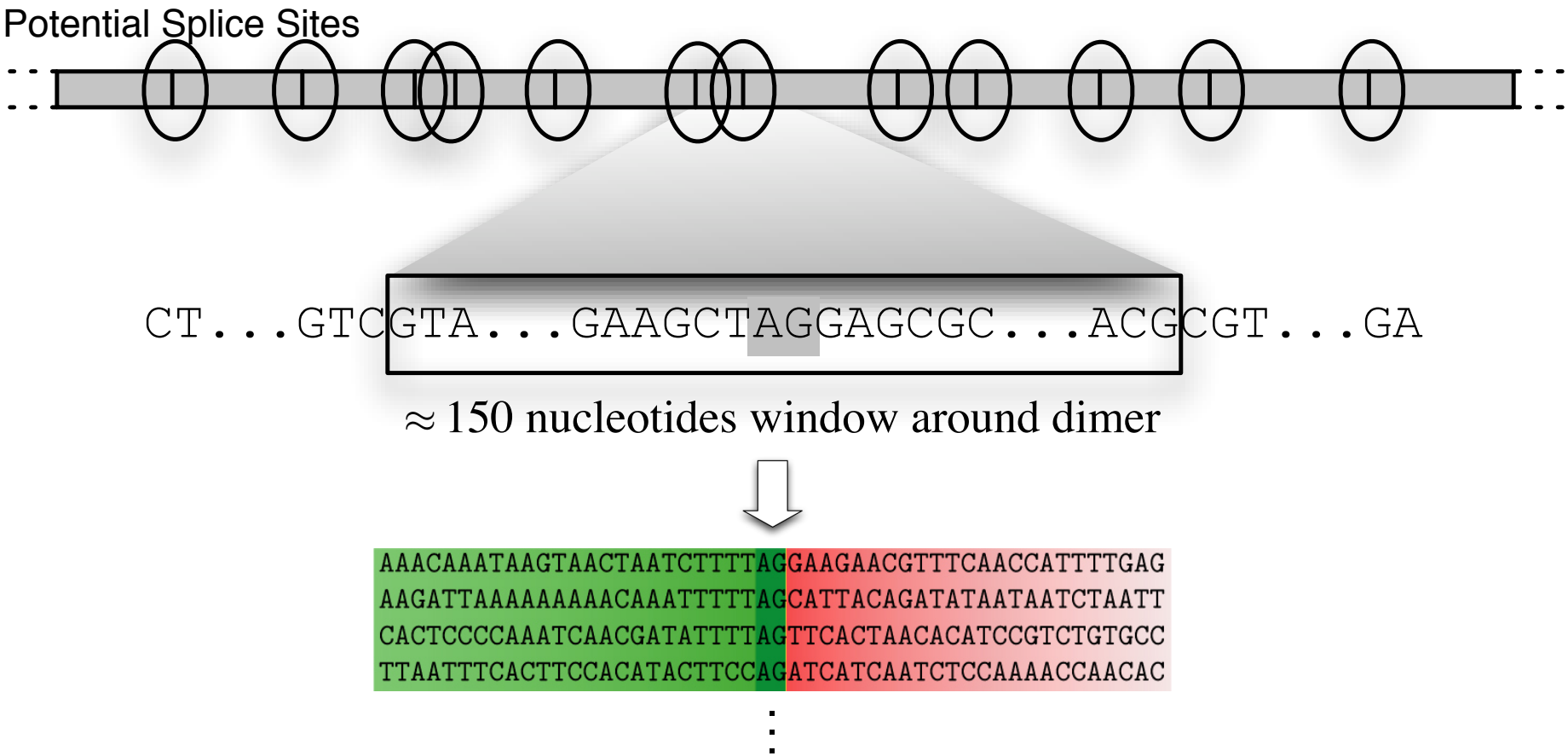
Example: Splice Site Recognition



Example: Splice Site Recognition

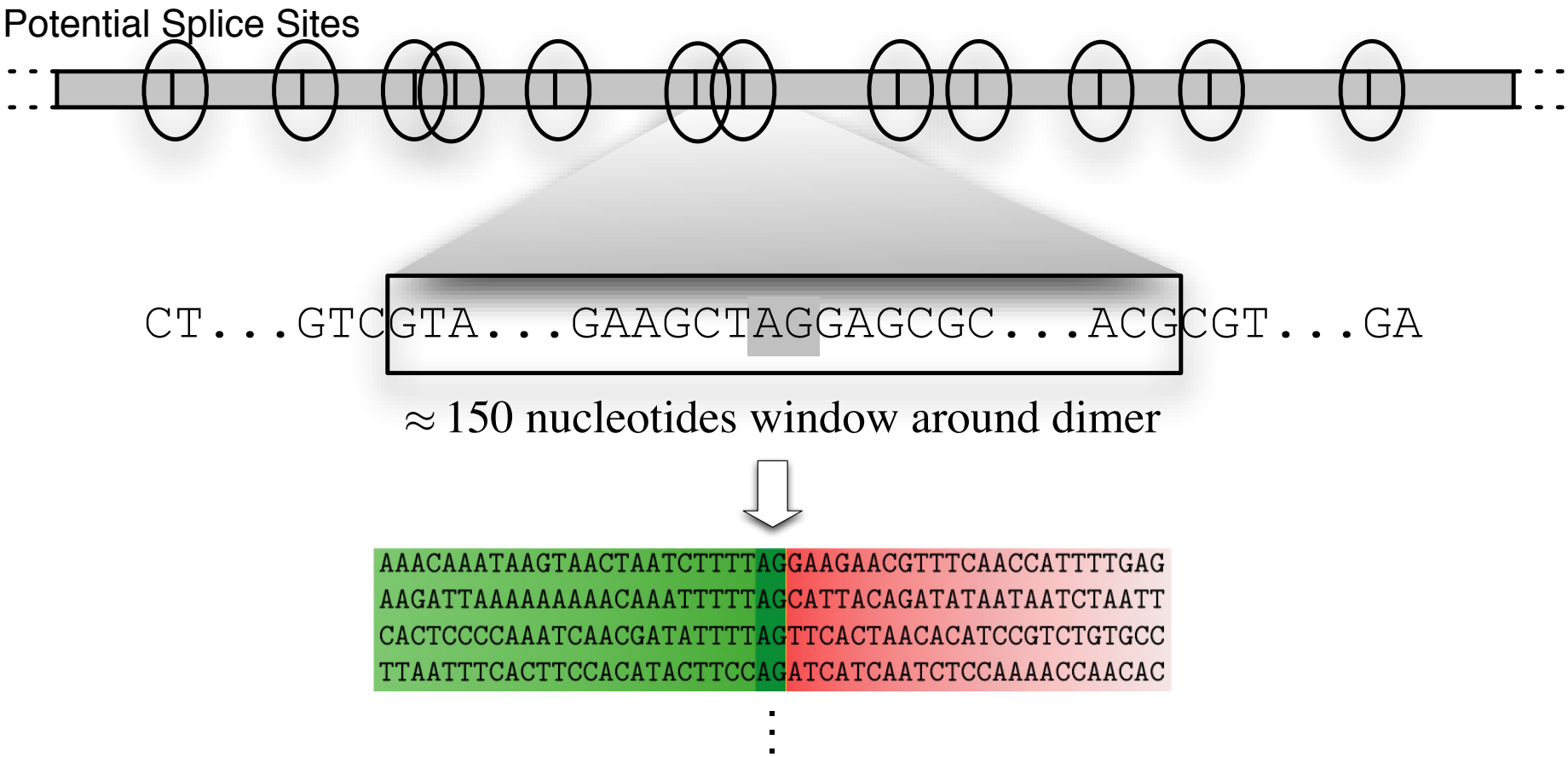


Example: Splice Site Recognition



- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

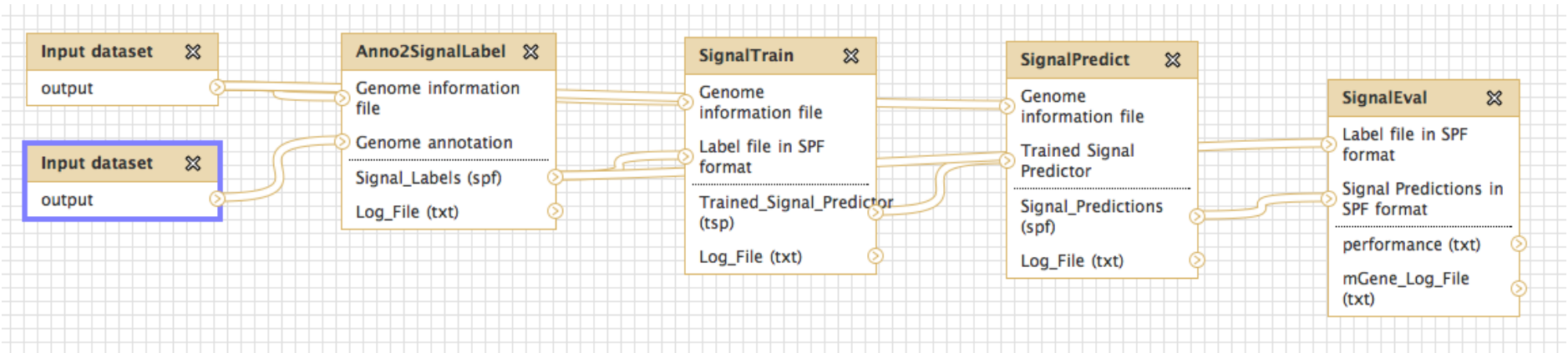
Example: Splice Site Recognition



- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

$\Rightarrow$ **Millions of labeled instances from EST databases**

Workflow for Predicting Signals

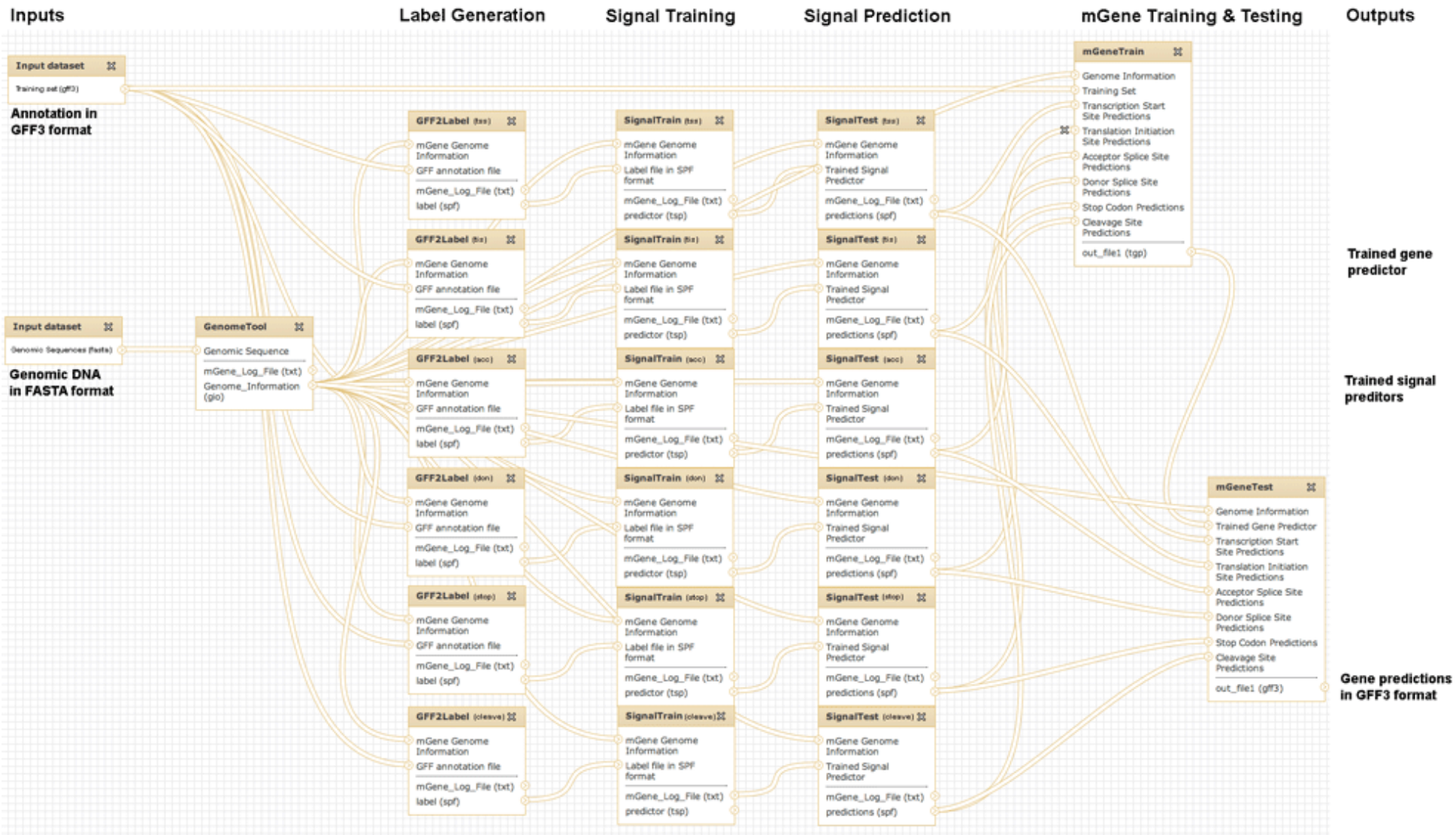


Inputs: Genome & its (incomplete) annotation

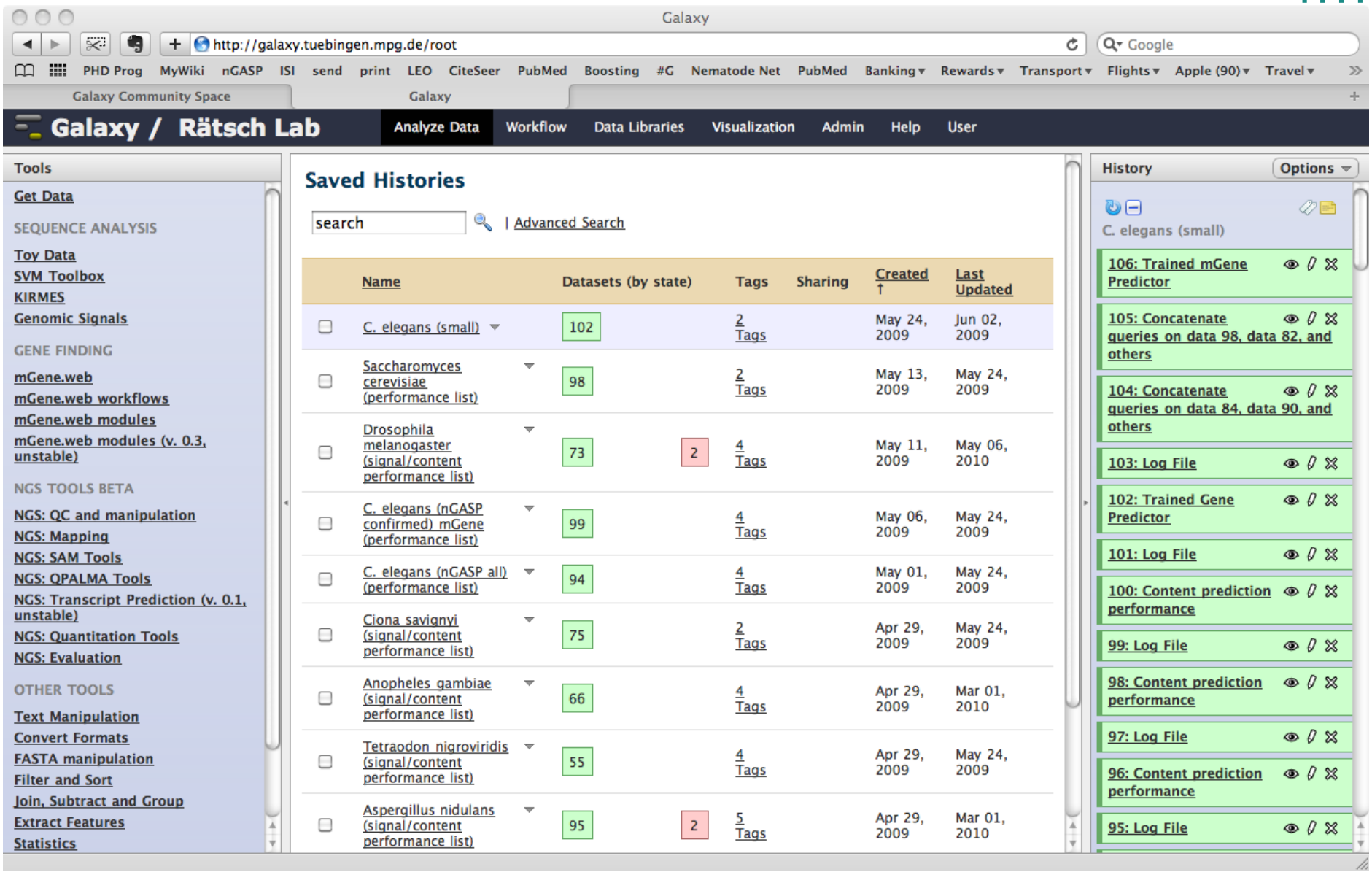
Steps:

1. Generate labels for signal from annotation
2. Train predictor
3. Predict on whole genome
4. Evaluate performance

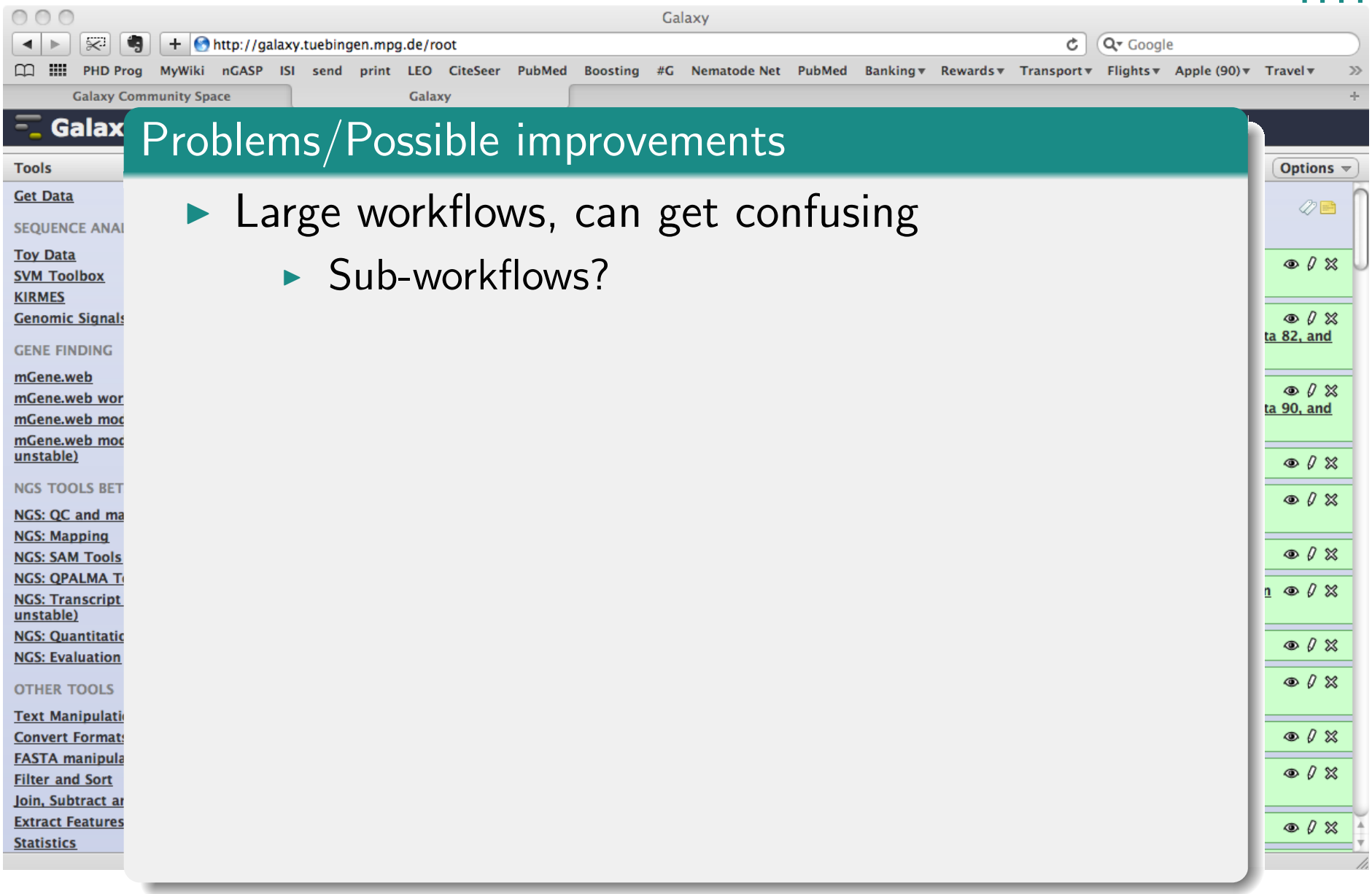
Workflow for Gene Finding



Workflow for Gene Finding



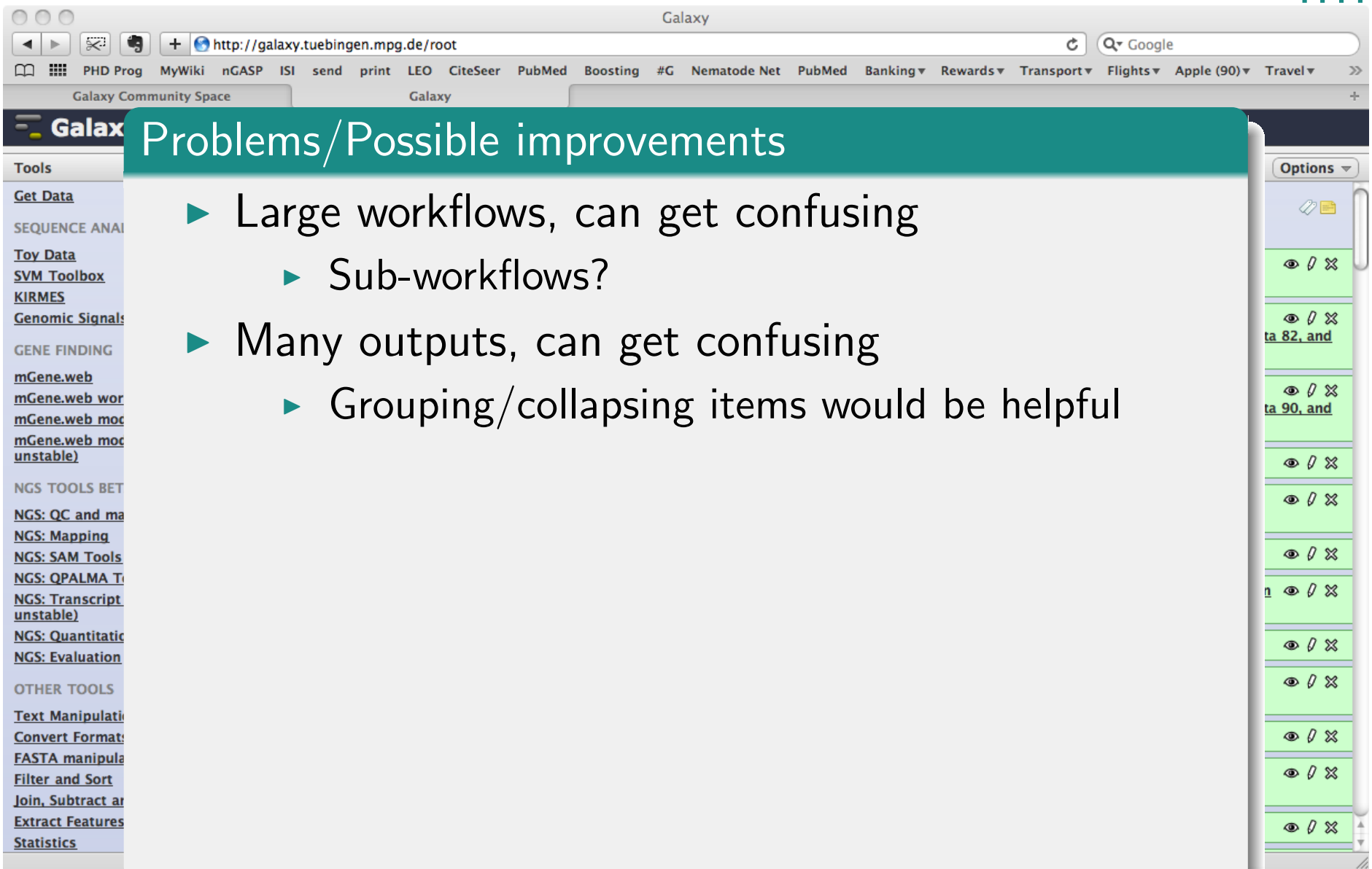
Workflow for Gene Finding



Problems/Possible improvements

- Large workflows, can get confusing
  - Sub-workflows?

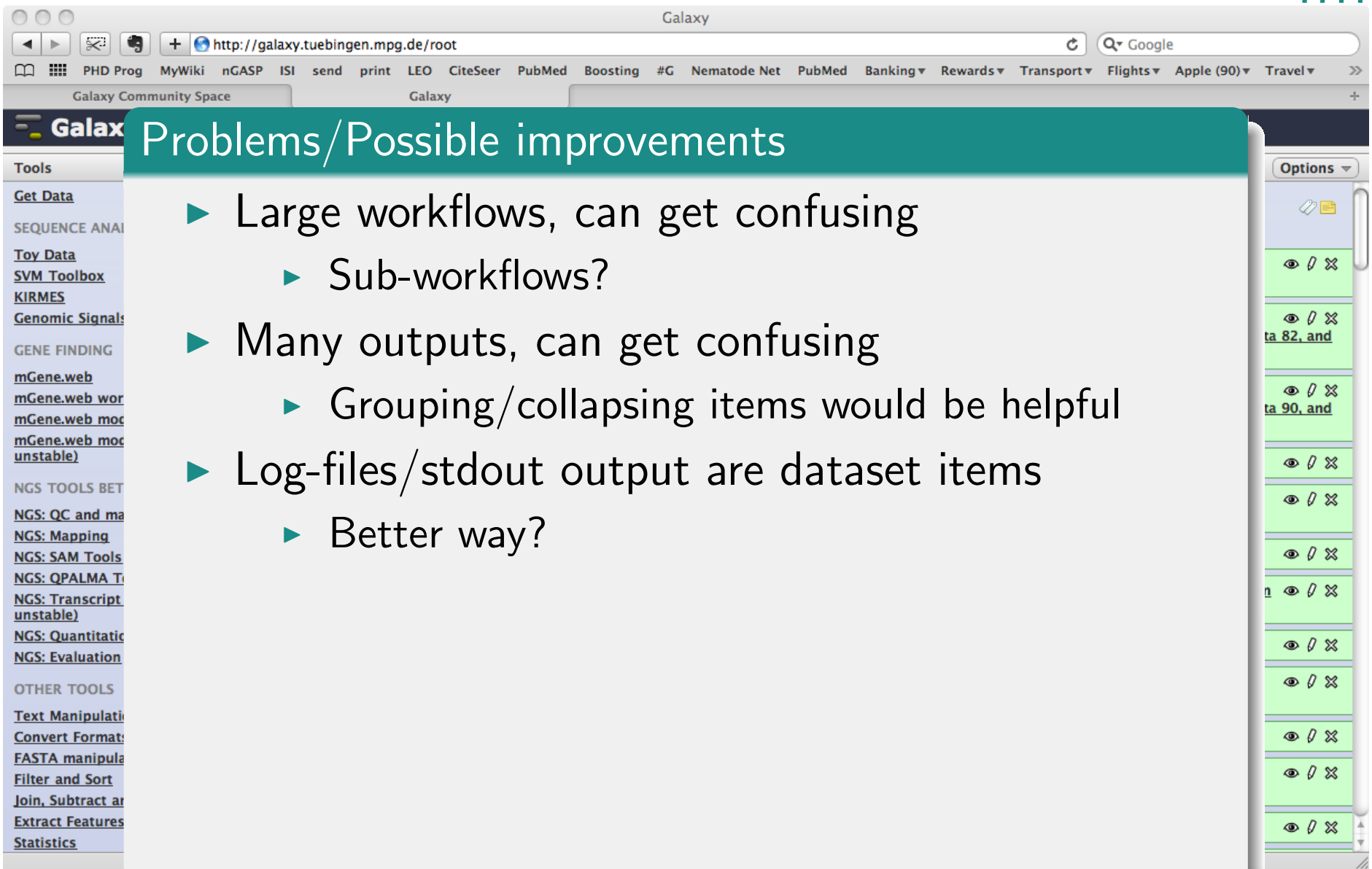
Workflow for Gene Finding



Problems/Possible improvements

- Large workflows, can get confusing
  - Sub-workflows?
- Many outputs, can get confusing
  - Grouping/collapsing items would be helpful

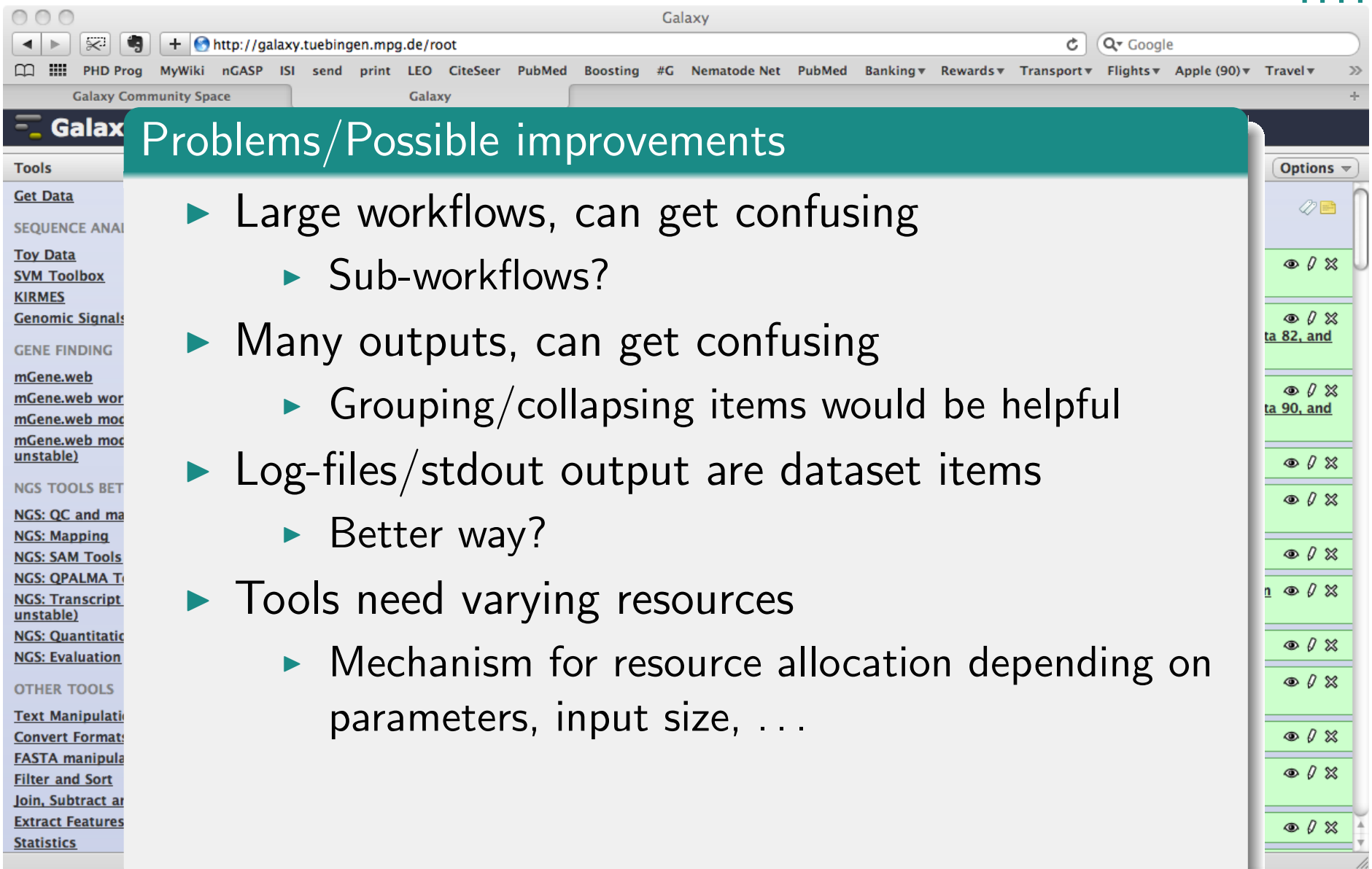
Workflow for Gene Finding



Problems/Possible improvements

- Large workflows, can get confusing
  - Sub-workflows?
- Many outputs, can get confusing
  - Grouping/collapsing items would be helpful
- Log-files/stdout output are dataset items
  - Better way?

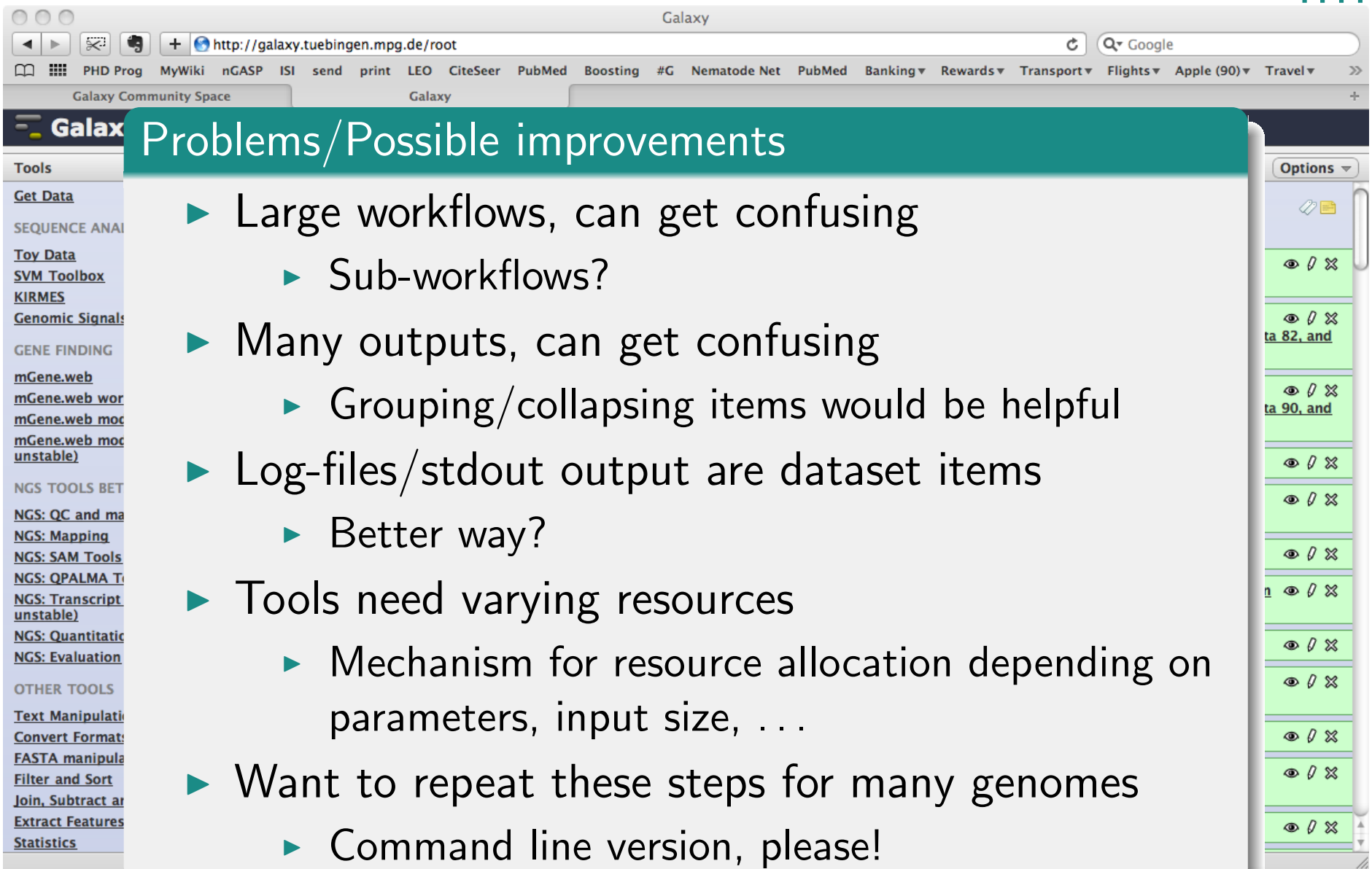
Workflow for Gene Finding



Problems/Possible improvements

- Large workflows, can get confusing
  - Sub-workflows?
- Many outputs, can get confusing
  - Grouping/collapsing items would be helpful
- Log-files/stdout output are dataset items
  - Better way?
- Tools need varying resources
  - Mechanism for resource allocation depending on parameters, input size, . . .

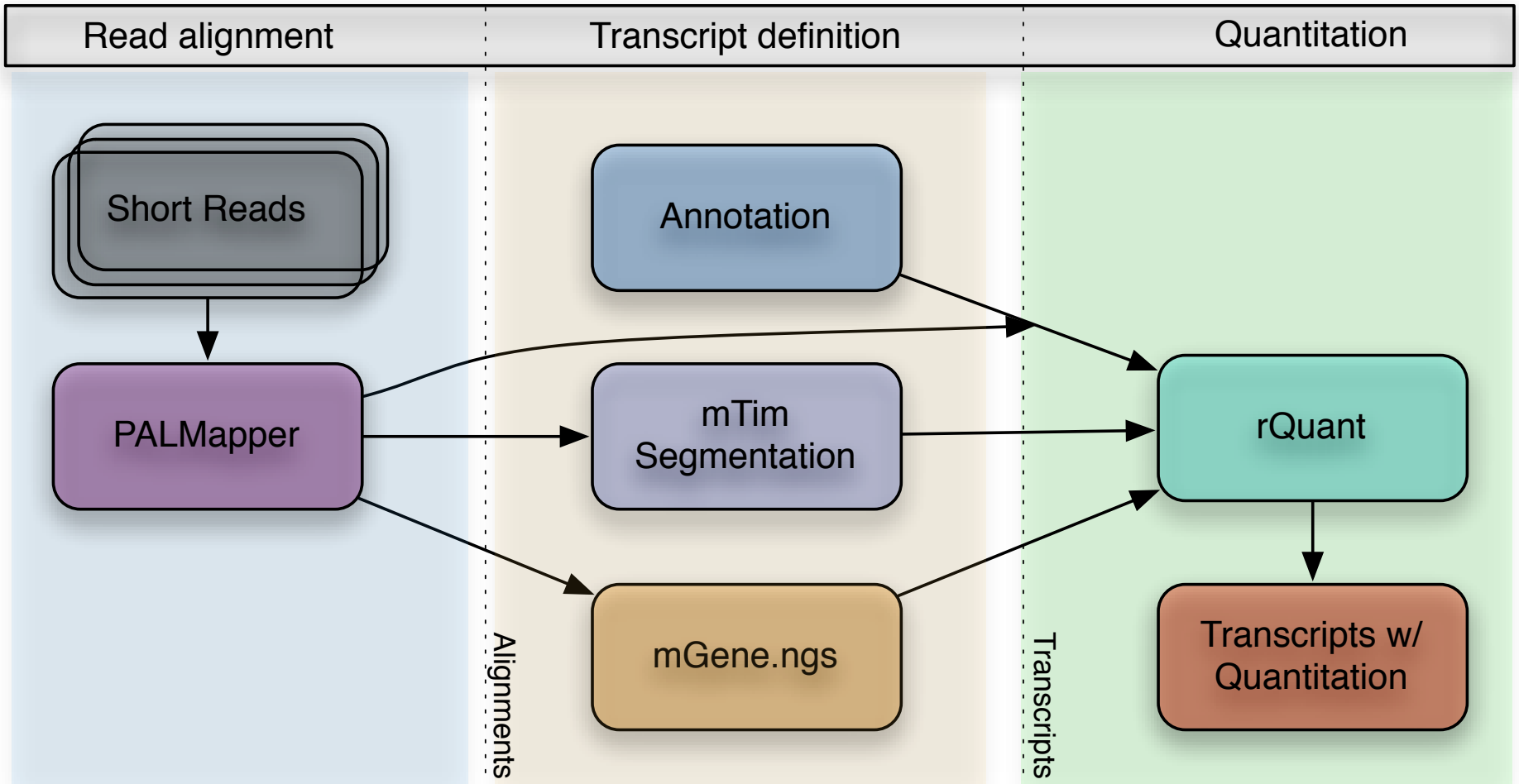
Workflow for Gene Finding



Problems/Possible improvements

- Large workflows, can get confusing
  - Sub-workflows?
- Many outputs, can get confusing
  - Grouping/collapsing items would be helpful
- Log-files/stdout output are dataset items
  - Better way?
- Tools need varying resources
  - Mechanism for resource allocation depending on parameters, input size, ...
- Want to repeat these steps for many genomes
  - Command line version, please!

RNA-seq Analysis



Accurate RNA-seq Alignment with PALMapper

PALMapper is the fusion of *GenomeMapper* (Schneeberger et al., 2009b) for fast read mapping and *QPALMA* (De Bona et al., 2008) for accurate spliced alignment, incorporating

- read sequence and quality
- splice site information

during the alignment.

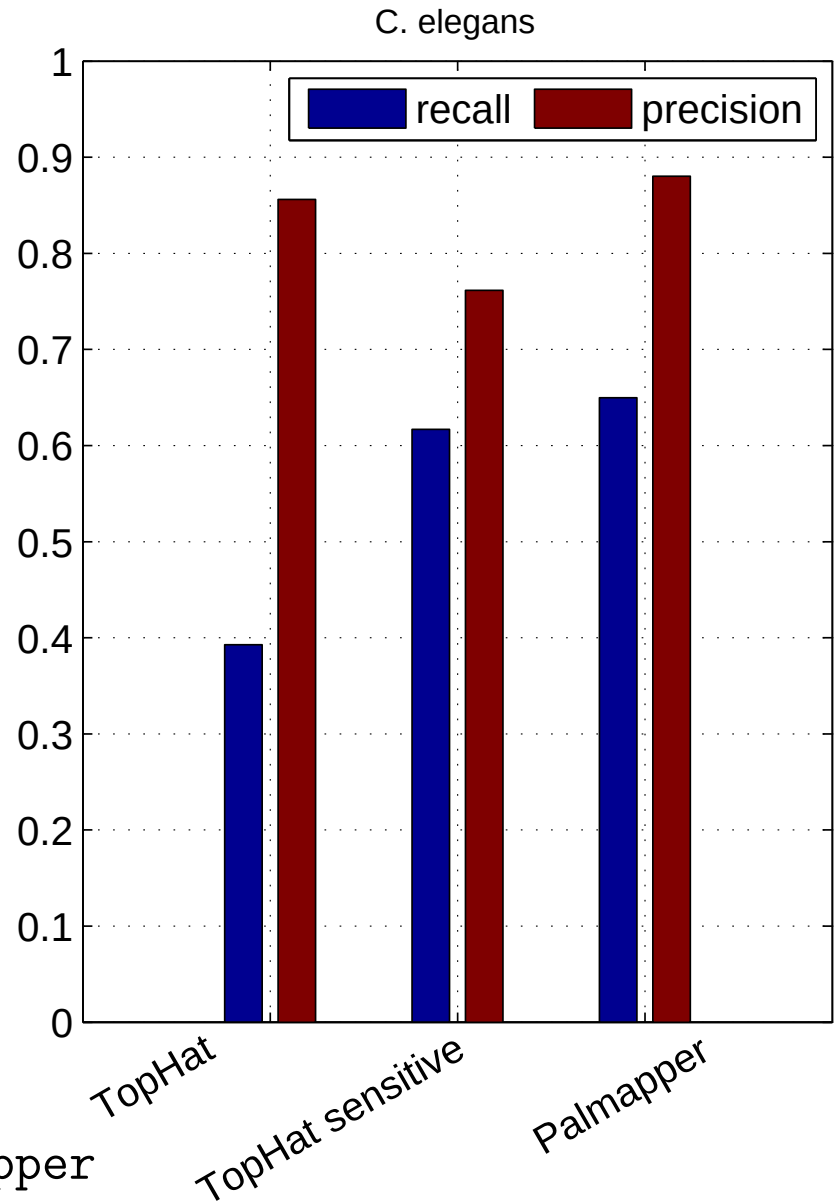
`http://fml.mpg.de/raetsch/suppl/palmapper`

Accurate RNA-seq Alignment with PALMapper

PALMapper is the fusion of *GenomeMapper* (Schneeberger et al., 2009b) for fast read mapping and *QPALMA* (De Bona et al., 2008) for accurate spliced alignment, incorporating

- read sequence and quality
- splice site information

during the alignment.



`http://fml.mpg.de/raetsch/suppl/palmapper`

QPALMA: Adaptive Alignment Scoring



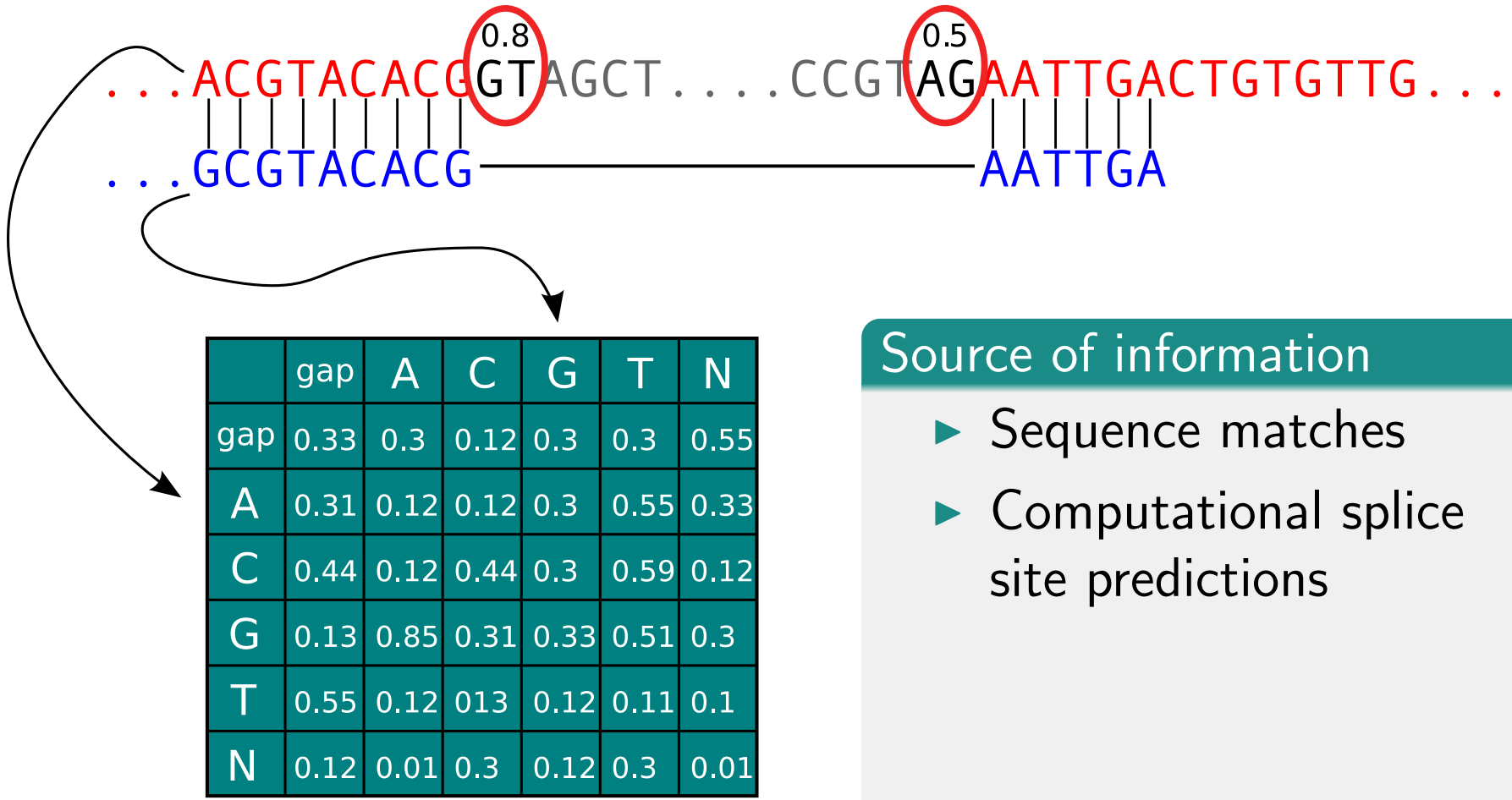
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Adaptive Alignment Scoring



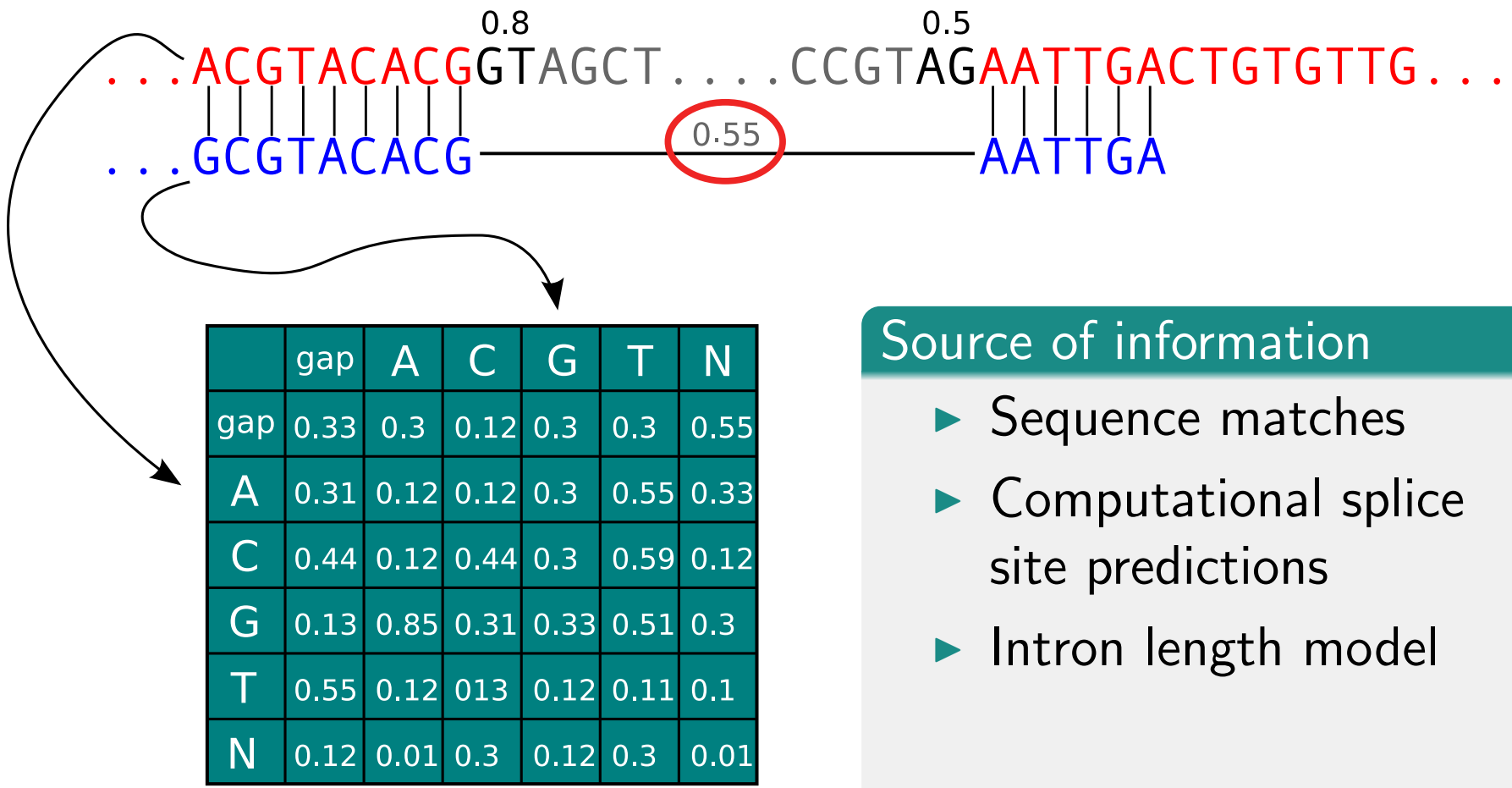
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Adaptive Alignment Scoring



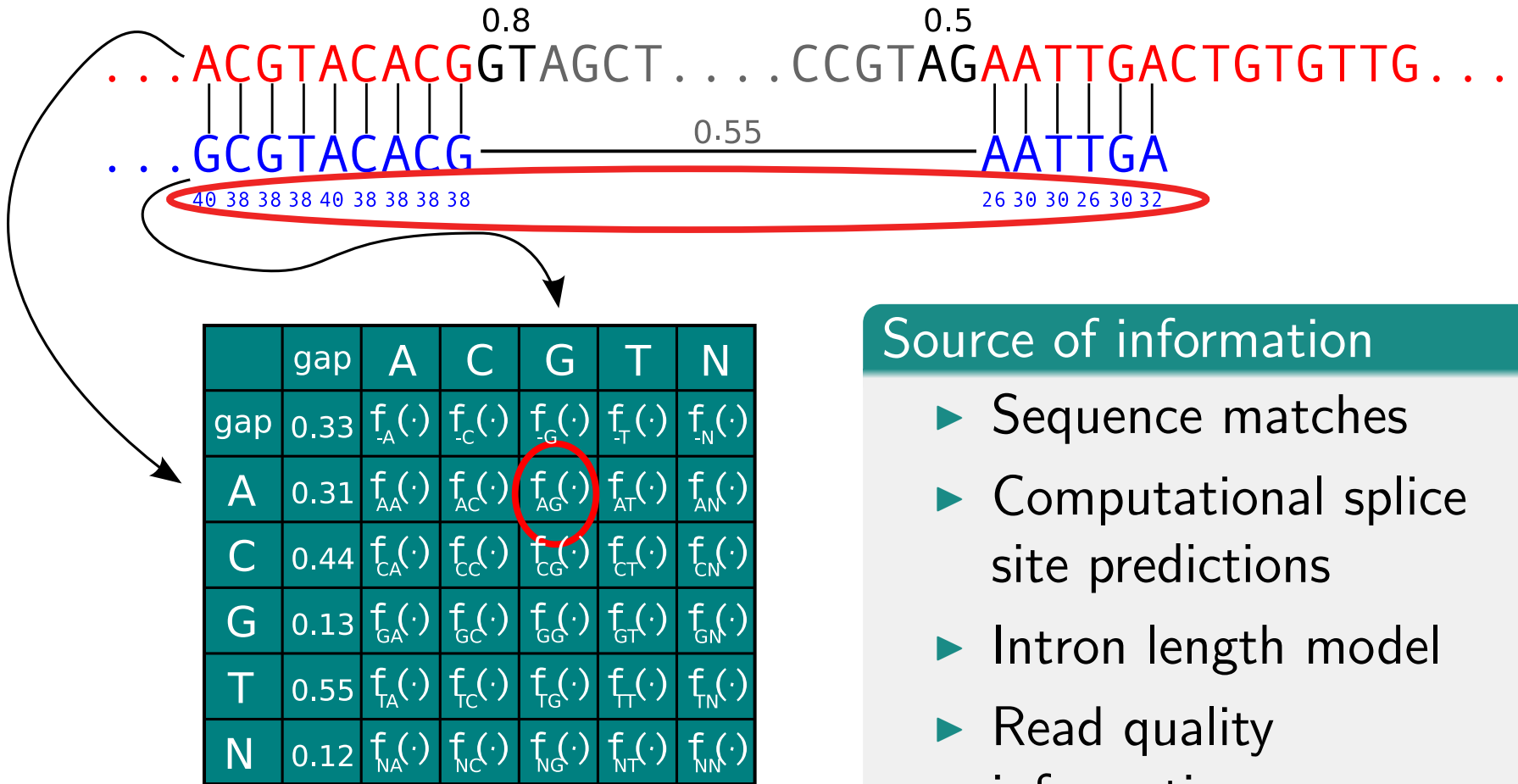
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model

Classical scoring $f : \Sigma \times \Sigma \rightarrow \mathbb{R}$

QPALMA: Adaptive Alignment Scoring



Source of information

- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$ (De Bona et al., 2008)

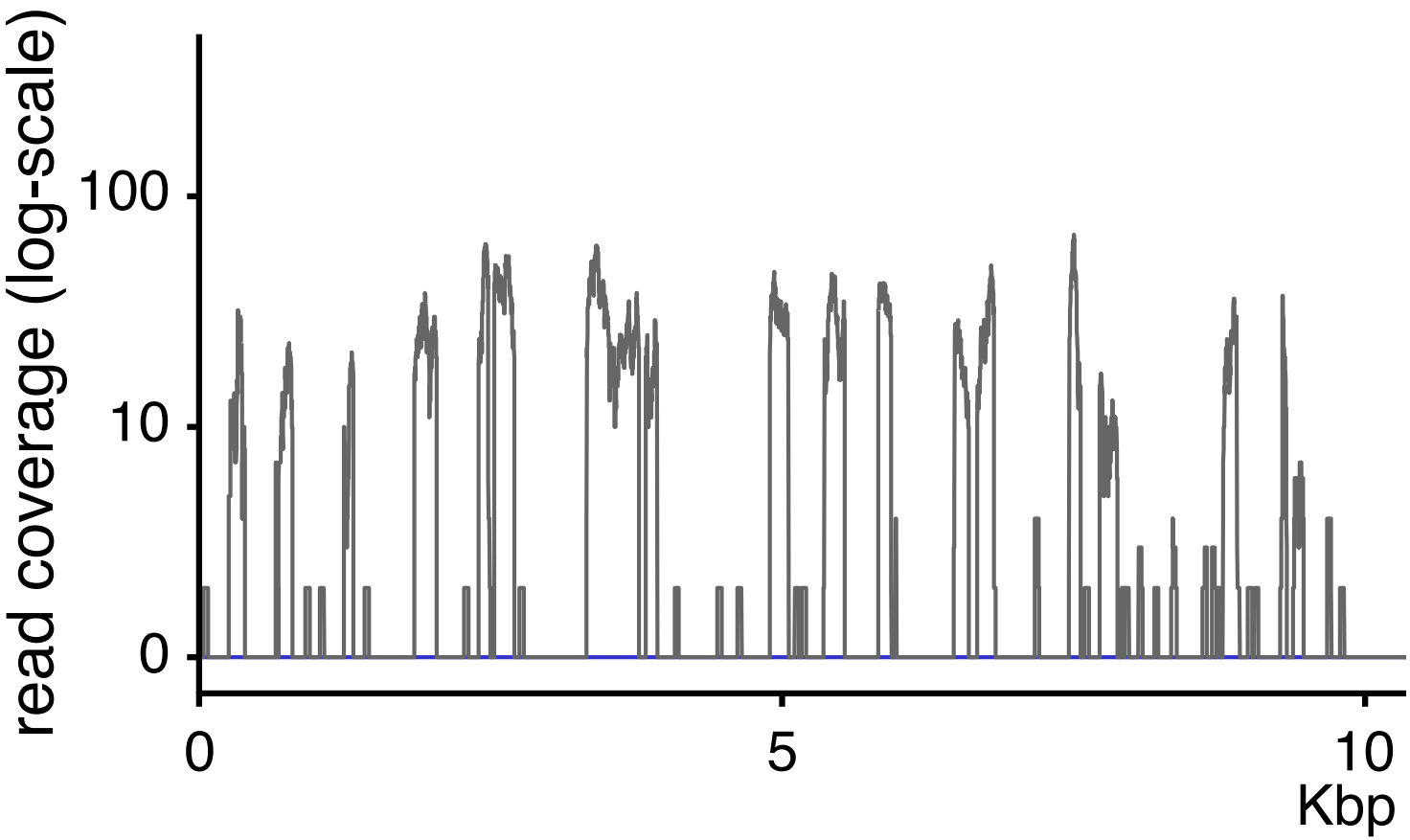
Step 2: Transcript Prediction

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

Input: Genome and BAM file for prediction
(For training they additionally need a set of annotated genes.)

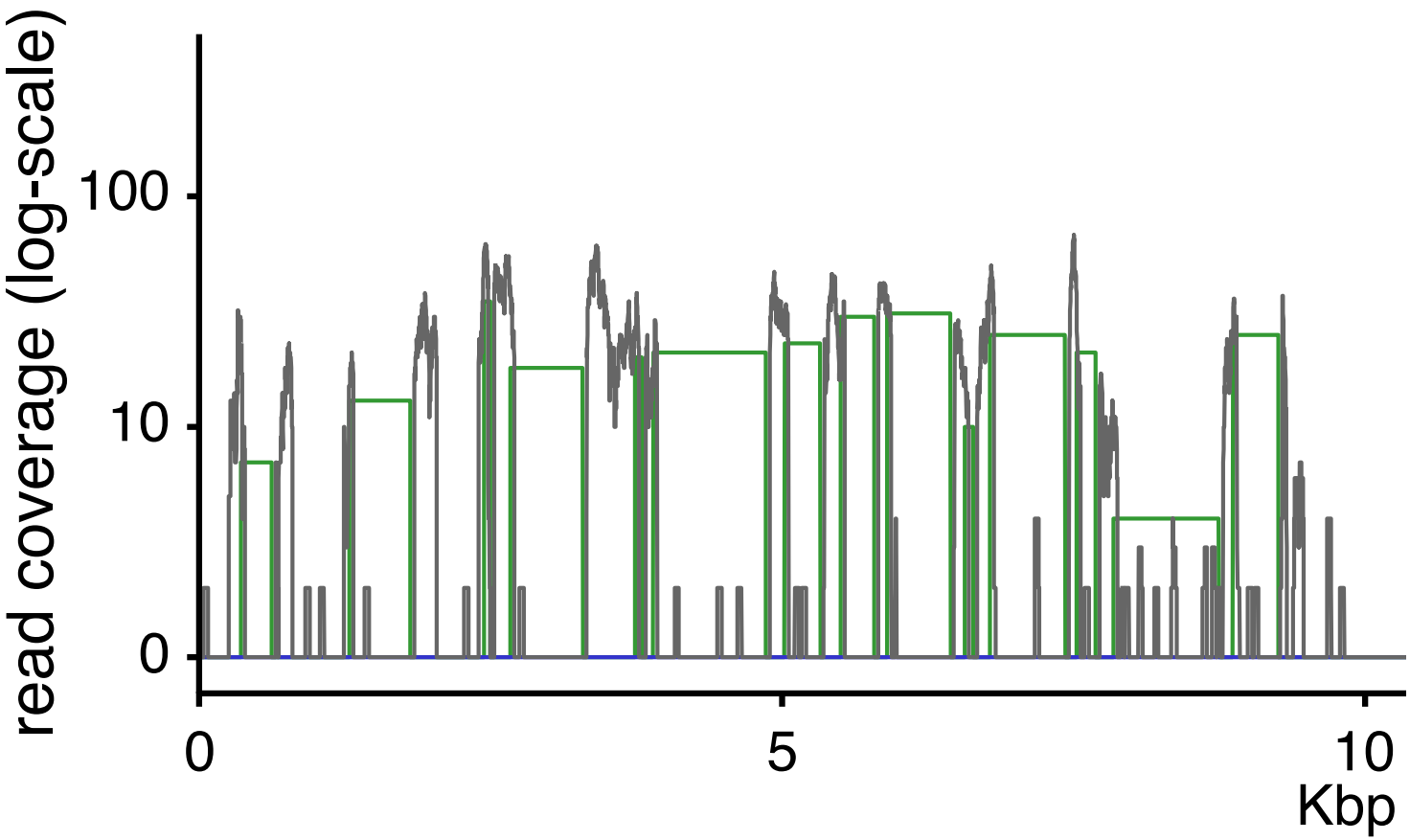
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



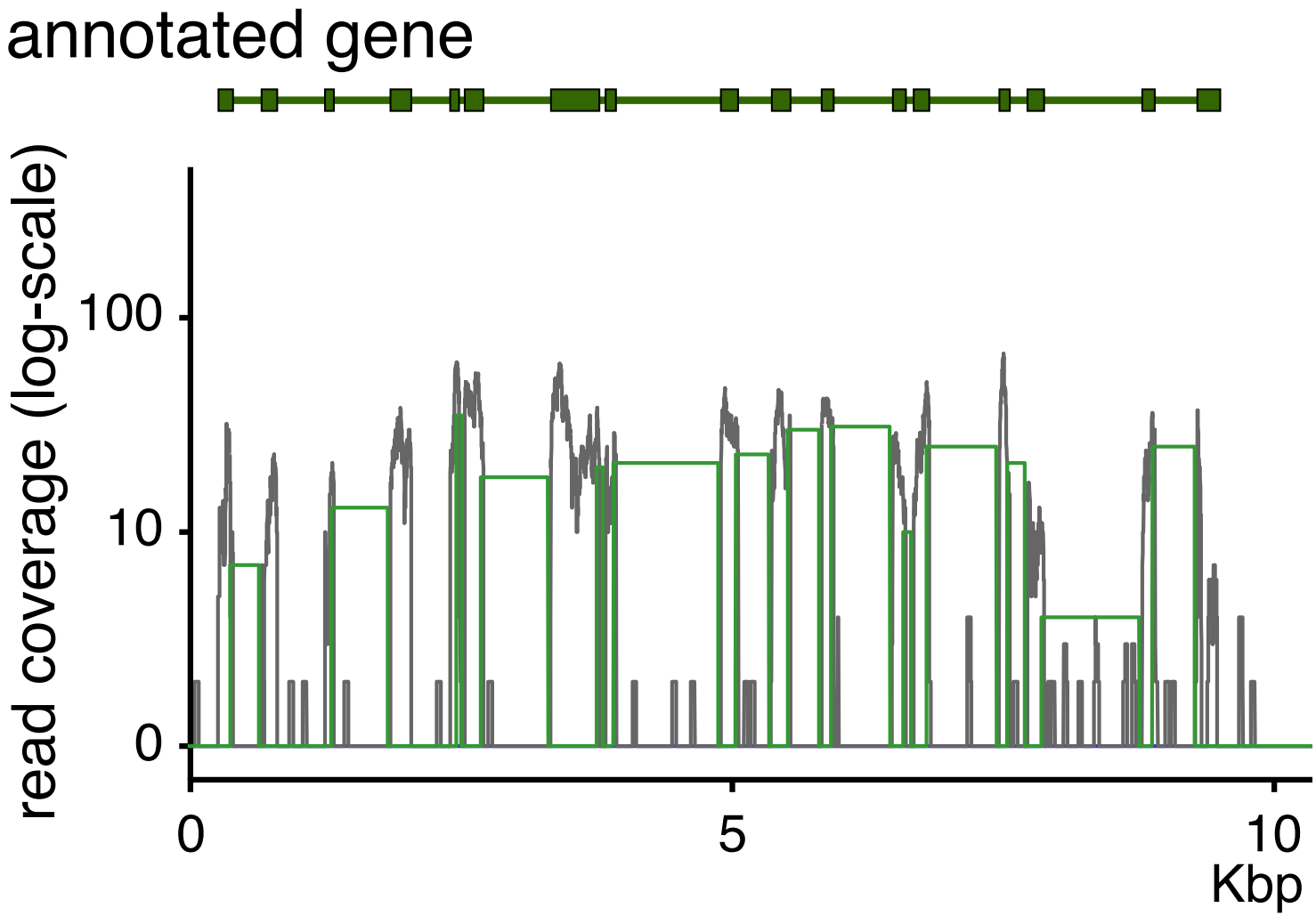
mTiM: Read Coverage Segmentation

Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*

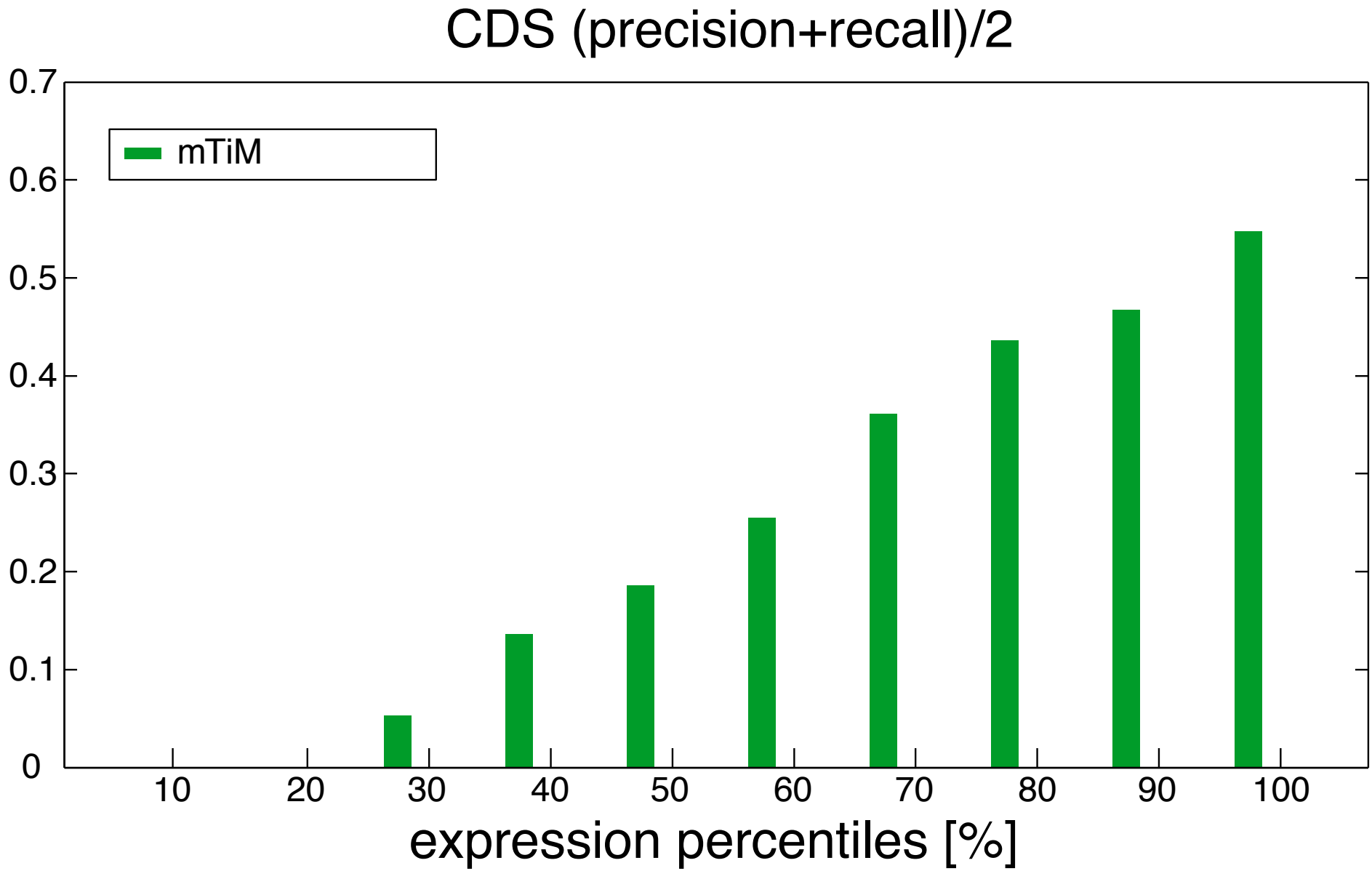


mTiM: Read Coverage Segmentation

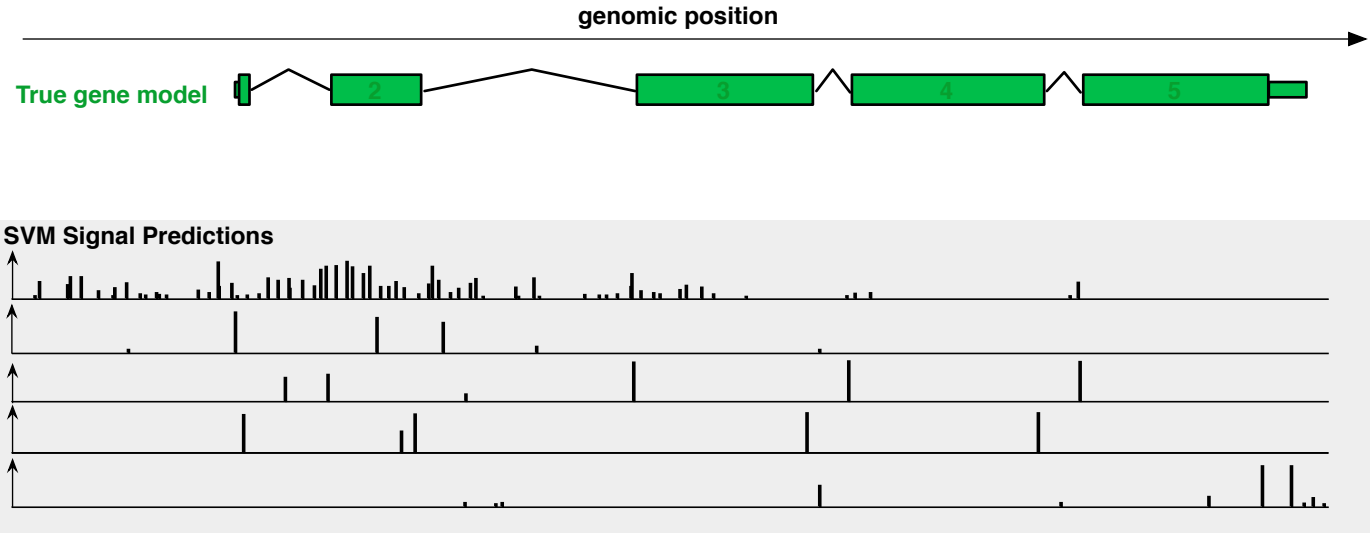
Goal: Characterize each base as *intergenic*, *exonic*, or *intronic*



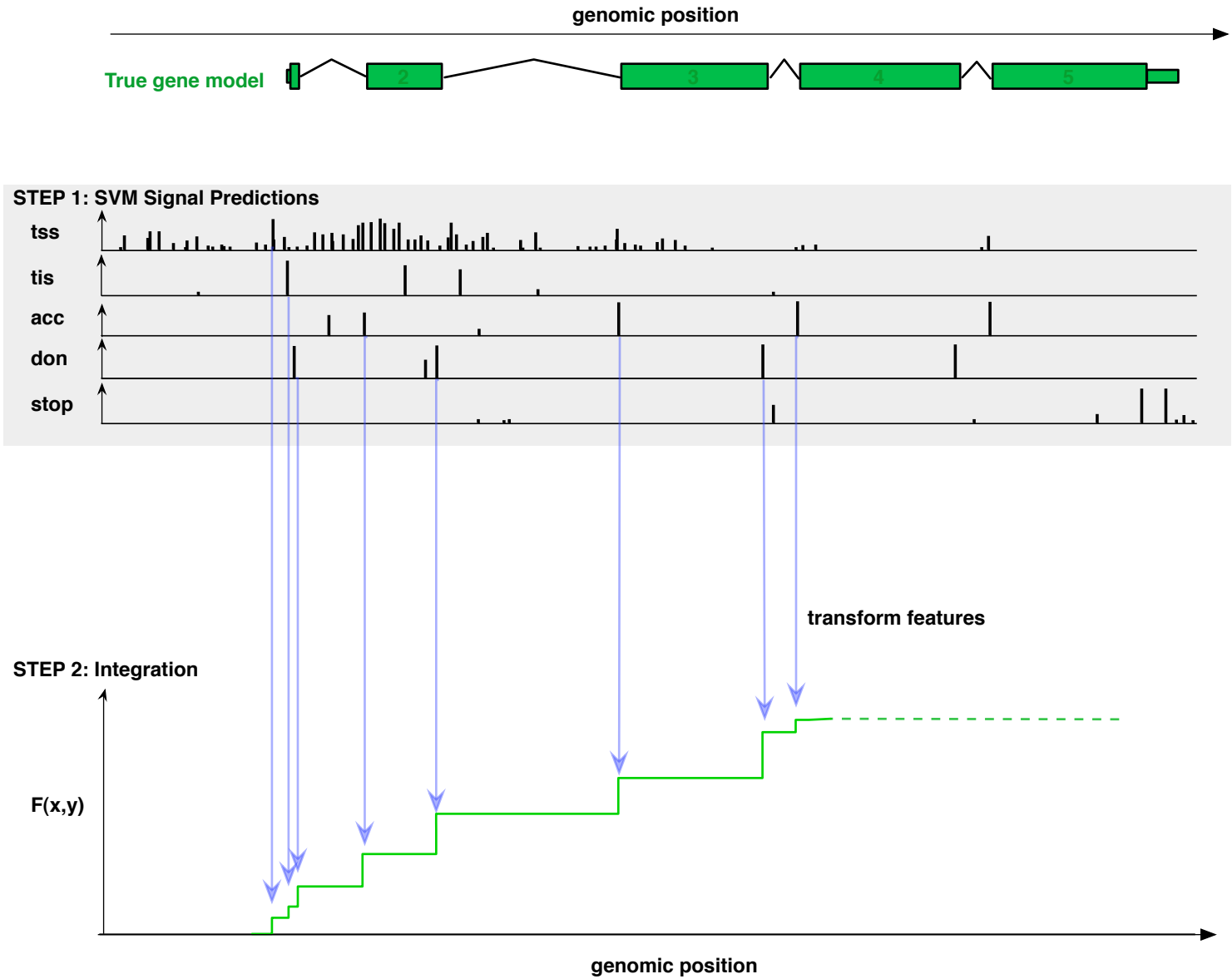
Preliminary Evaluation (C. elegans, RGASP)



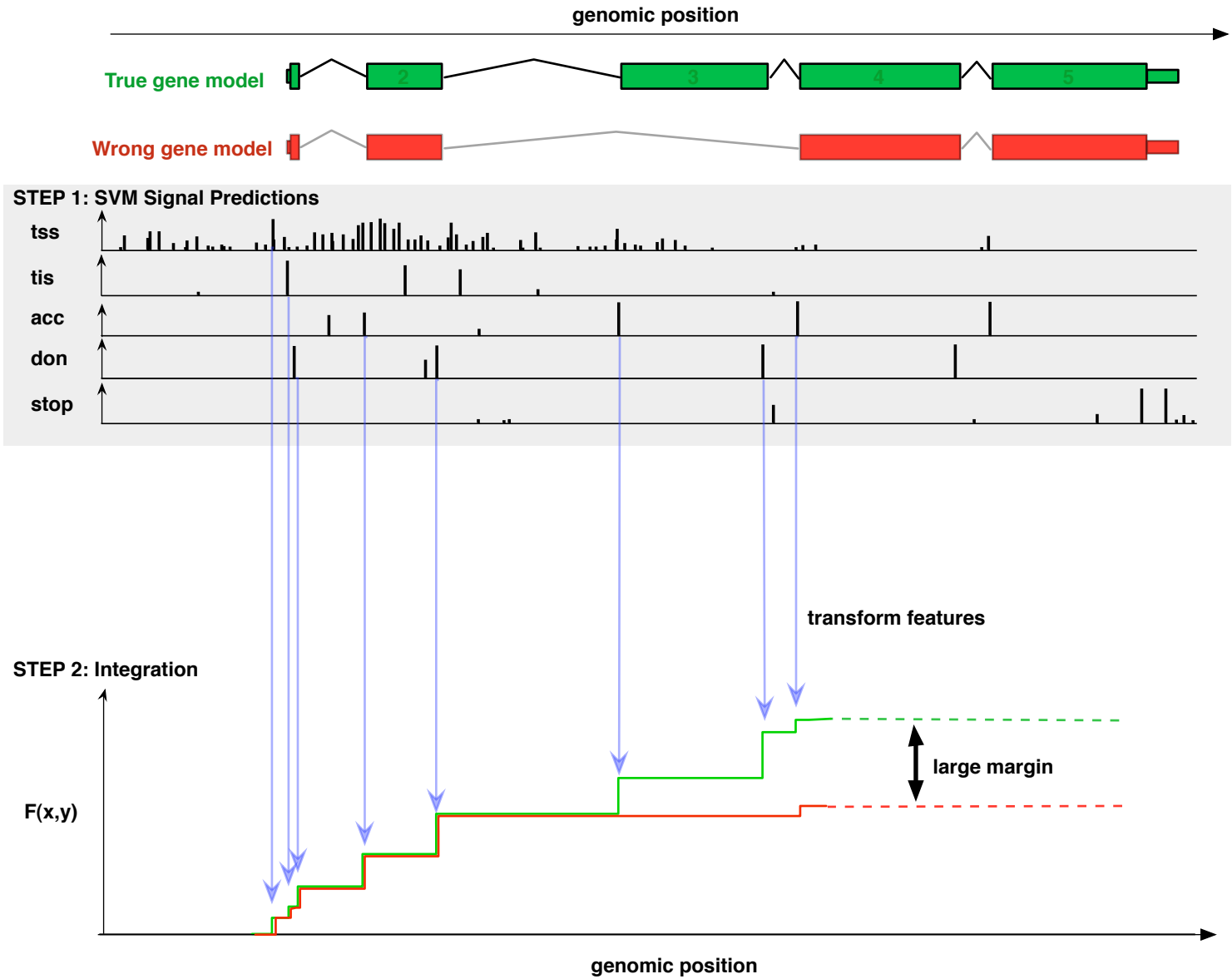
mGene-based Transcript Prediction I



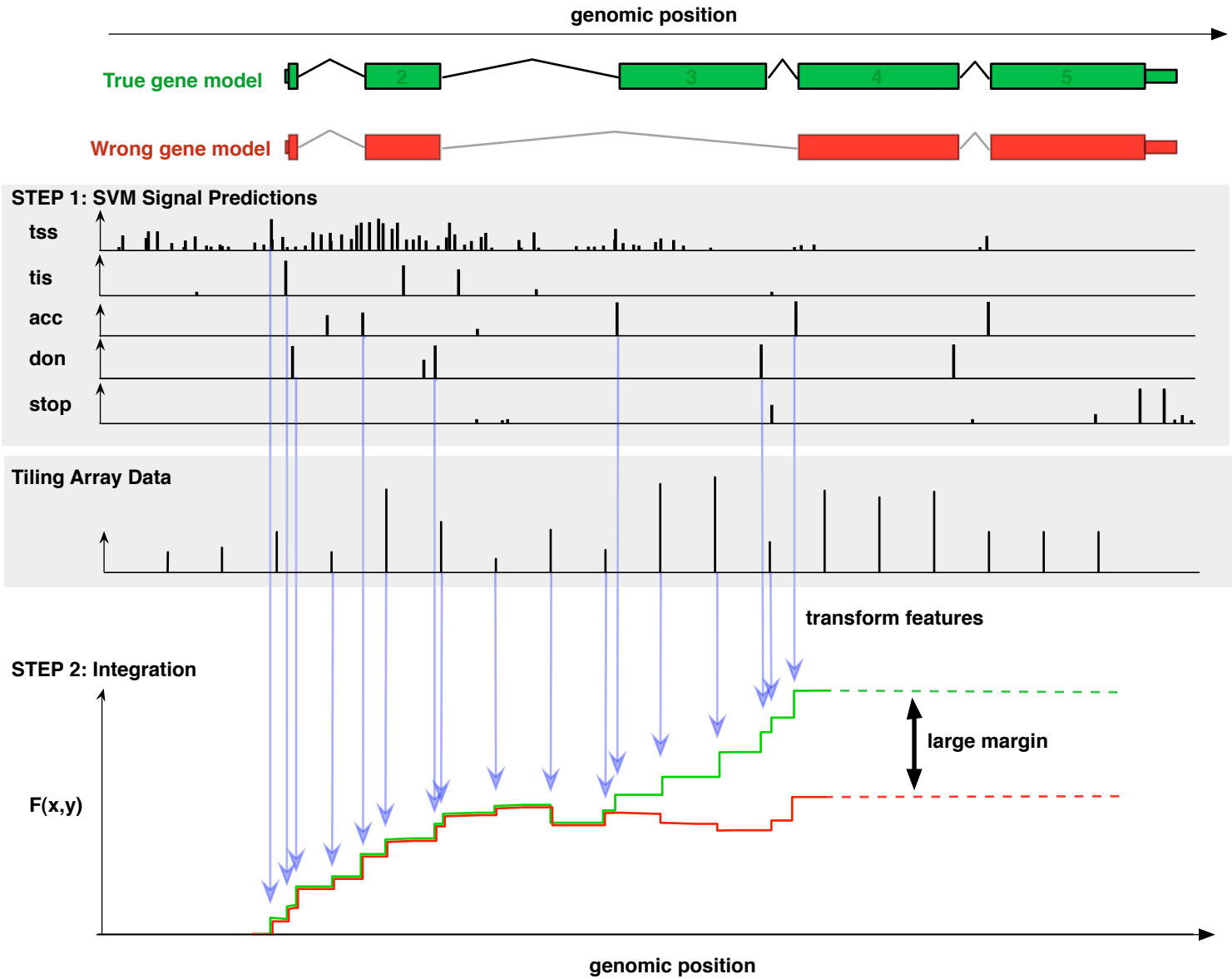
mGene-based Transcript Prediction I



mGene-based Transcript Prediction I



mGene-based Transcript Prediction I



mGene-based Transcript Prediction II

mGene with RNA-Seq (Behr et al., unpublished; Schweikert et al., 2009b,c)

- Use transcriptome measurements to enhance recognition of exonic regions

mGene-based Transcript Prediction II

mGene with RNA-Seq (Behr et al., unpublished; Schweikert et al., 2009b,c)

- Use transcriptome measurements to enhance recognition of exonic regions

Results for A. thaliana: (Comparison with known gene models)

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**

mGene-based Transcript Prediction II

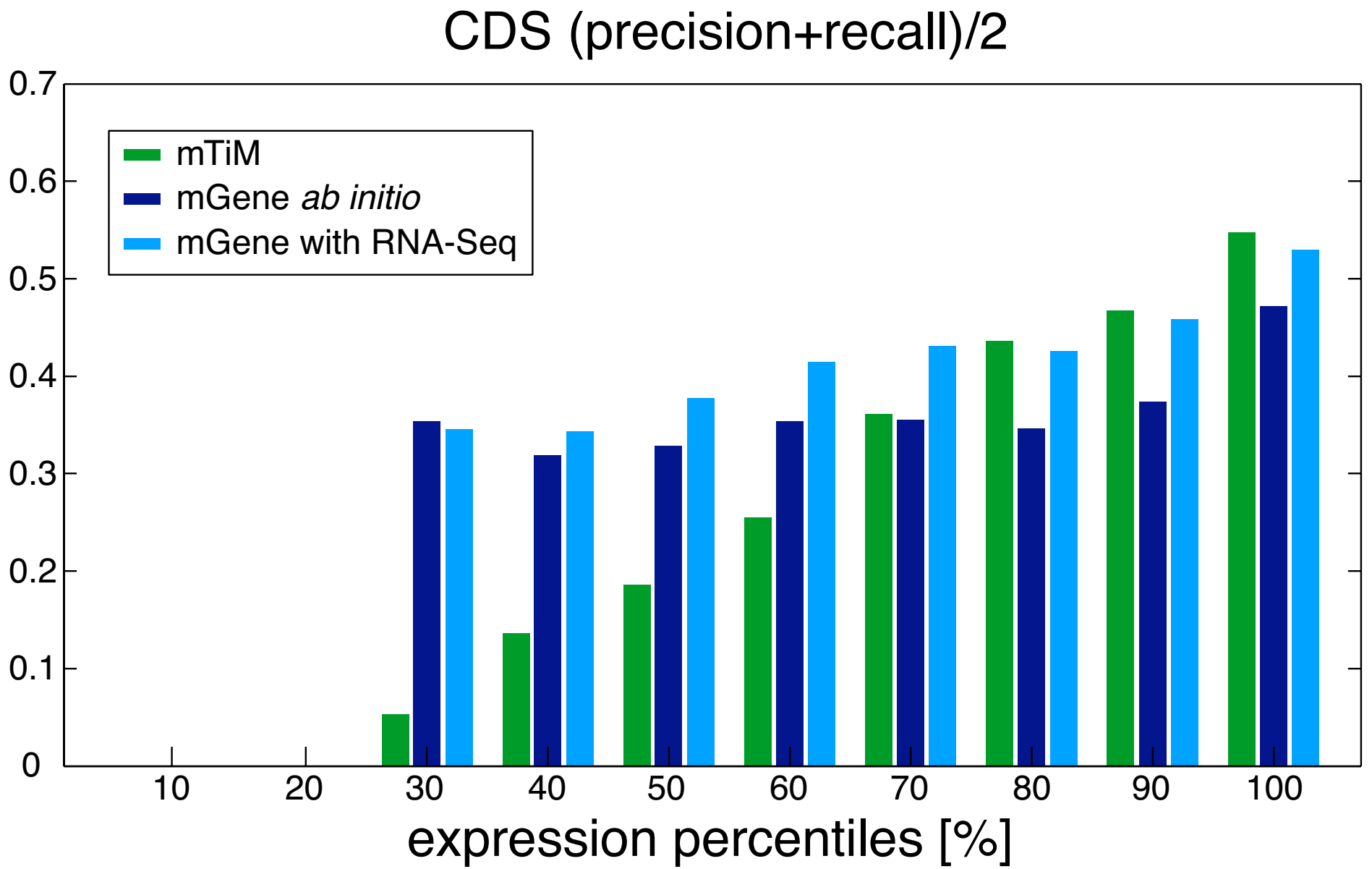
mGene with RNA-Seq (Behr et al., unpublished; Schweikert et al., 2009b,c)

- Use transcriptome measurements to enhance recognition of exonic regions

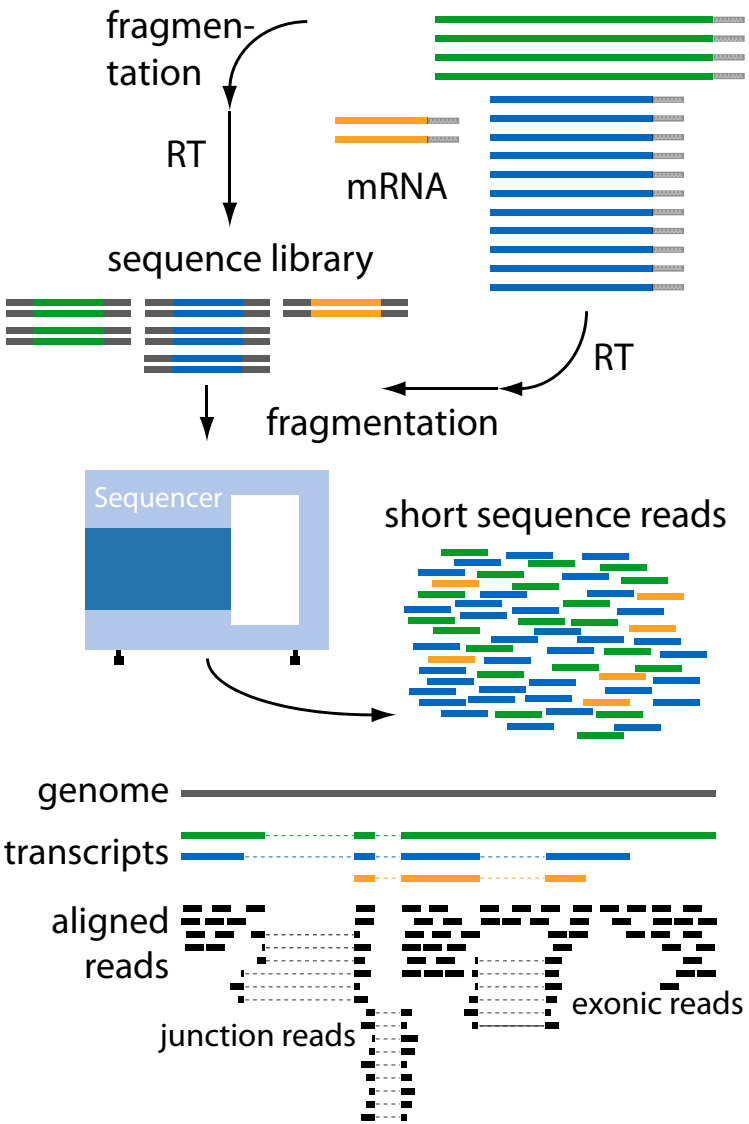
Results for A. thaliana: (Comparison with known gene models)

	transcript level $(SN + SP)/2$
1. mGene (*ab initio*) ...	**73.3%**
2. ...with tiling arrays (11 tissues)	**82.1%**
3. ...with mRNA-seq (1 tissue)	**81.1%**

Preliminary Evaluation (C. elegans, RGASP)



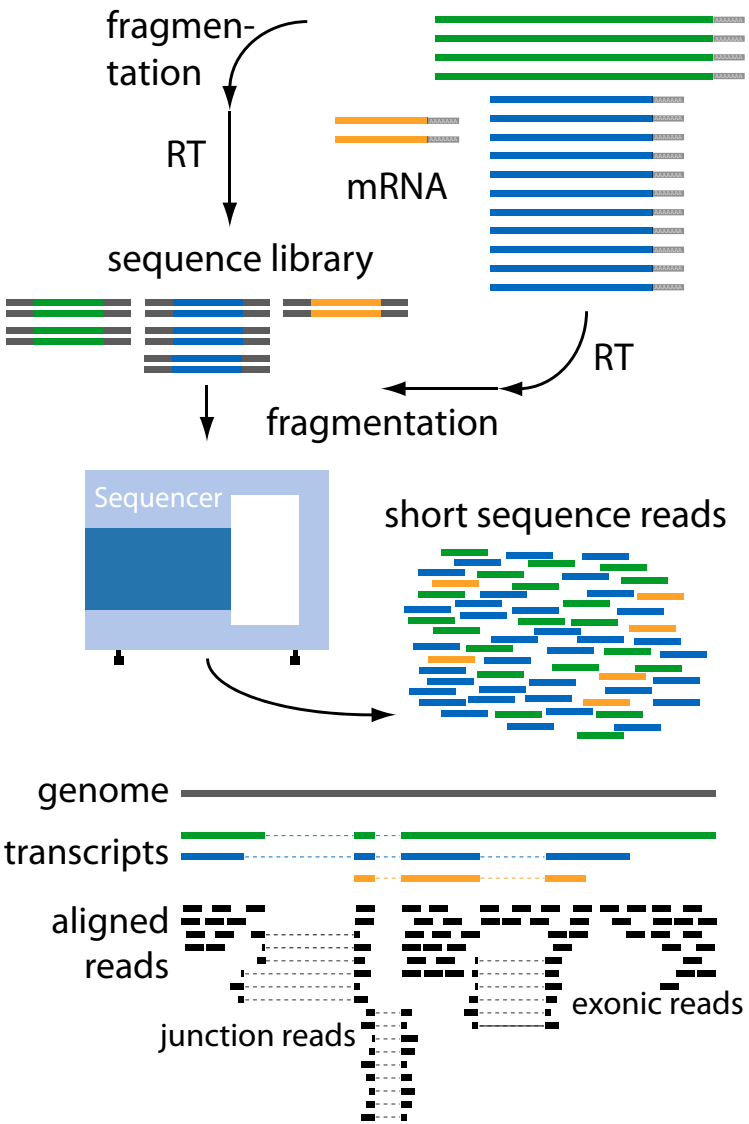
RNA-Seq Biases and Quantitation



Biases due to . . .

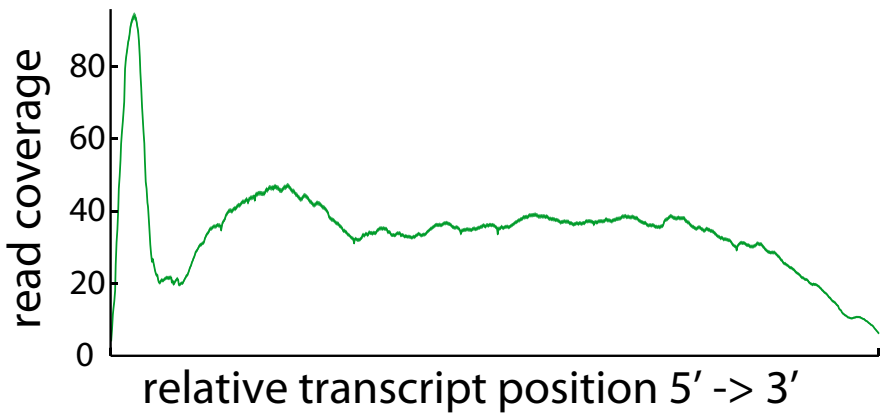
- cDNA library construction
- Sequencing
- Read mapping

RNA-Seq Biases and Quantitation



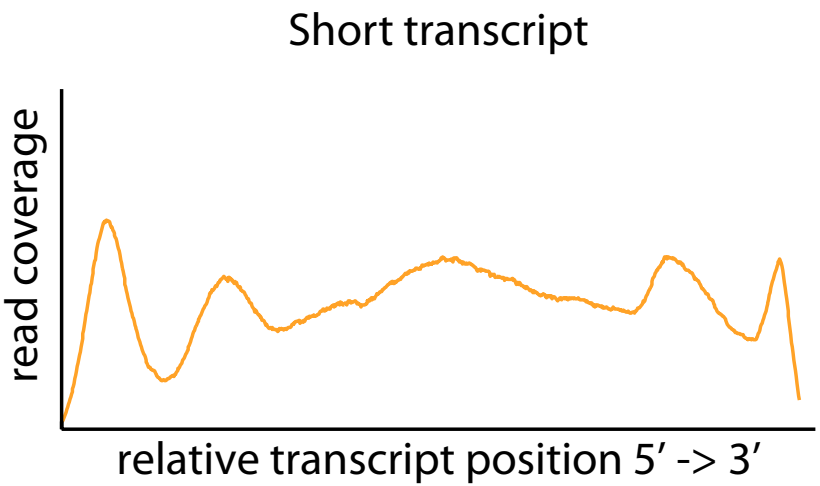
Biases due to . . .

- cDNA library construction
- Sequencing
- Read mapping



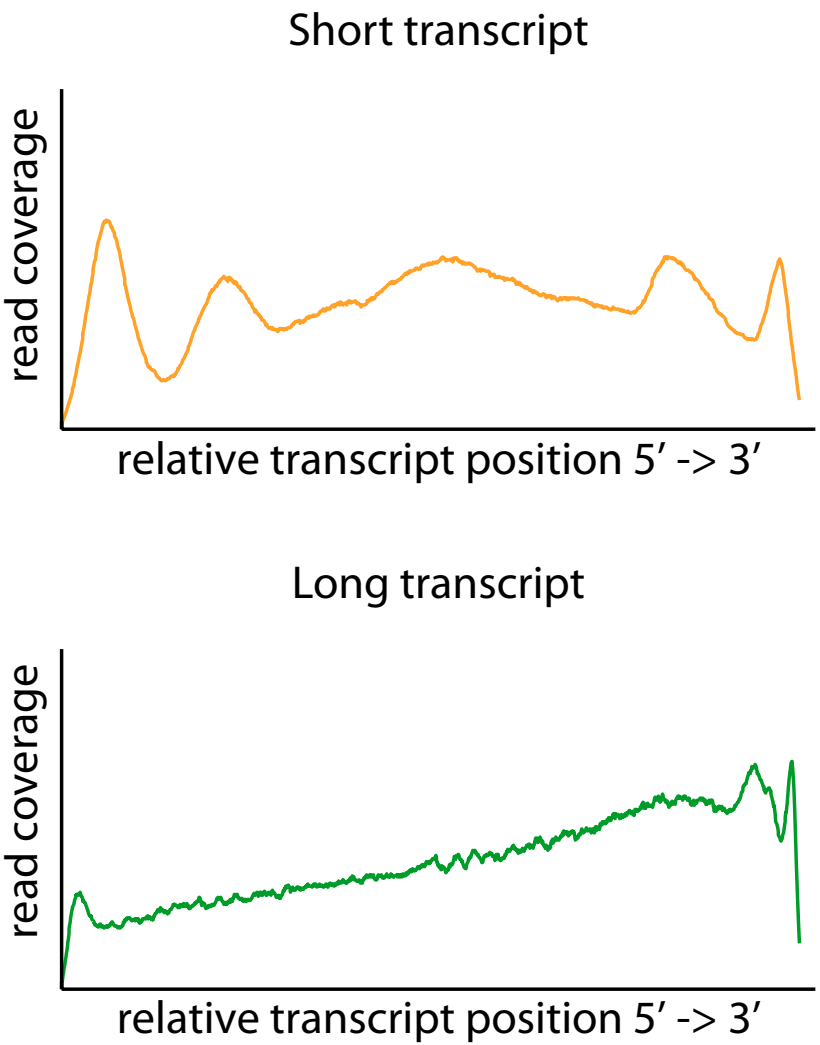
(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

rQuant – Basic Idea

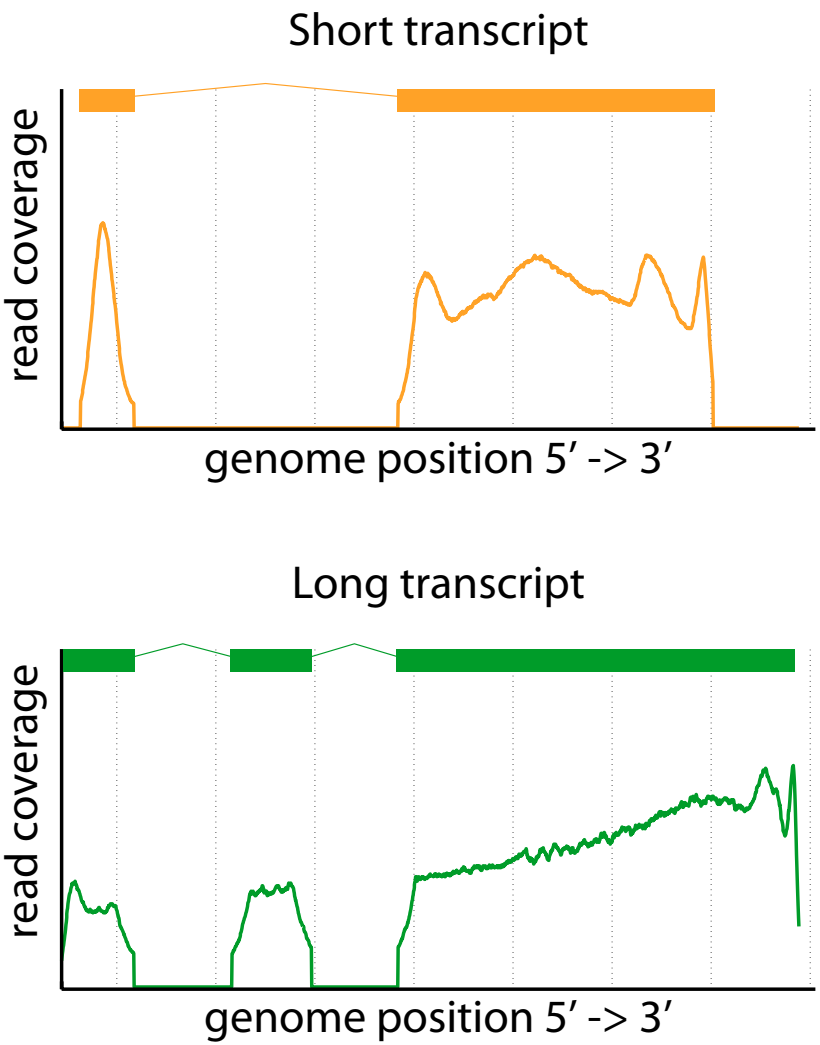




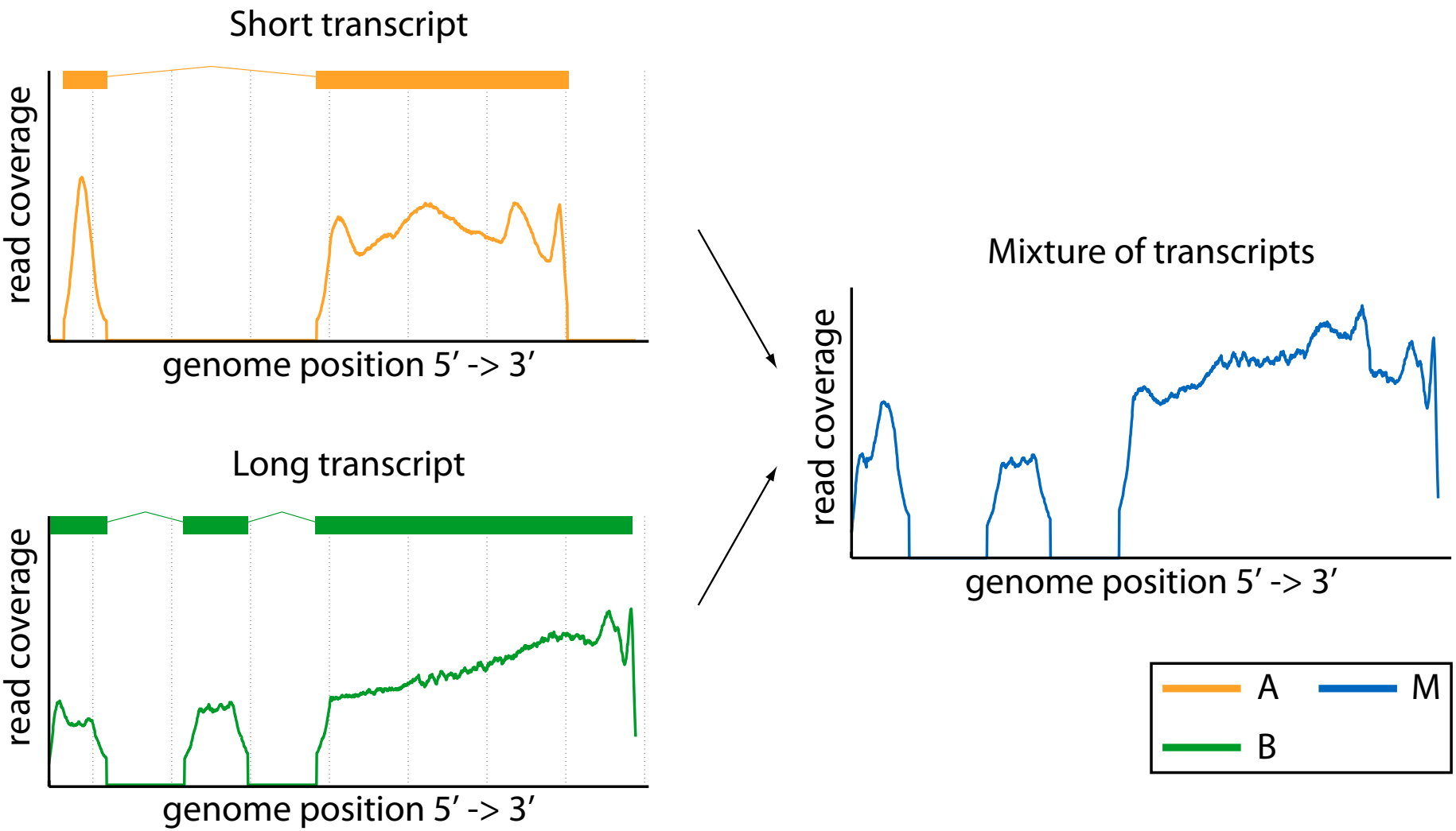
rQuant – Basic Idea



rQuant – Basic Idea

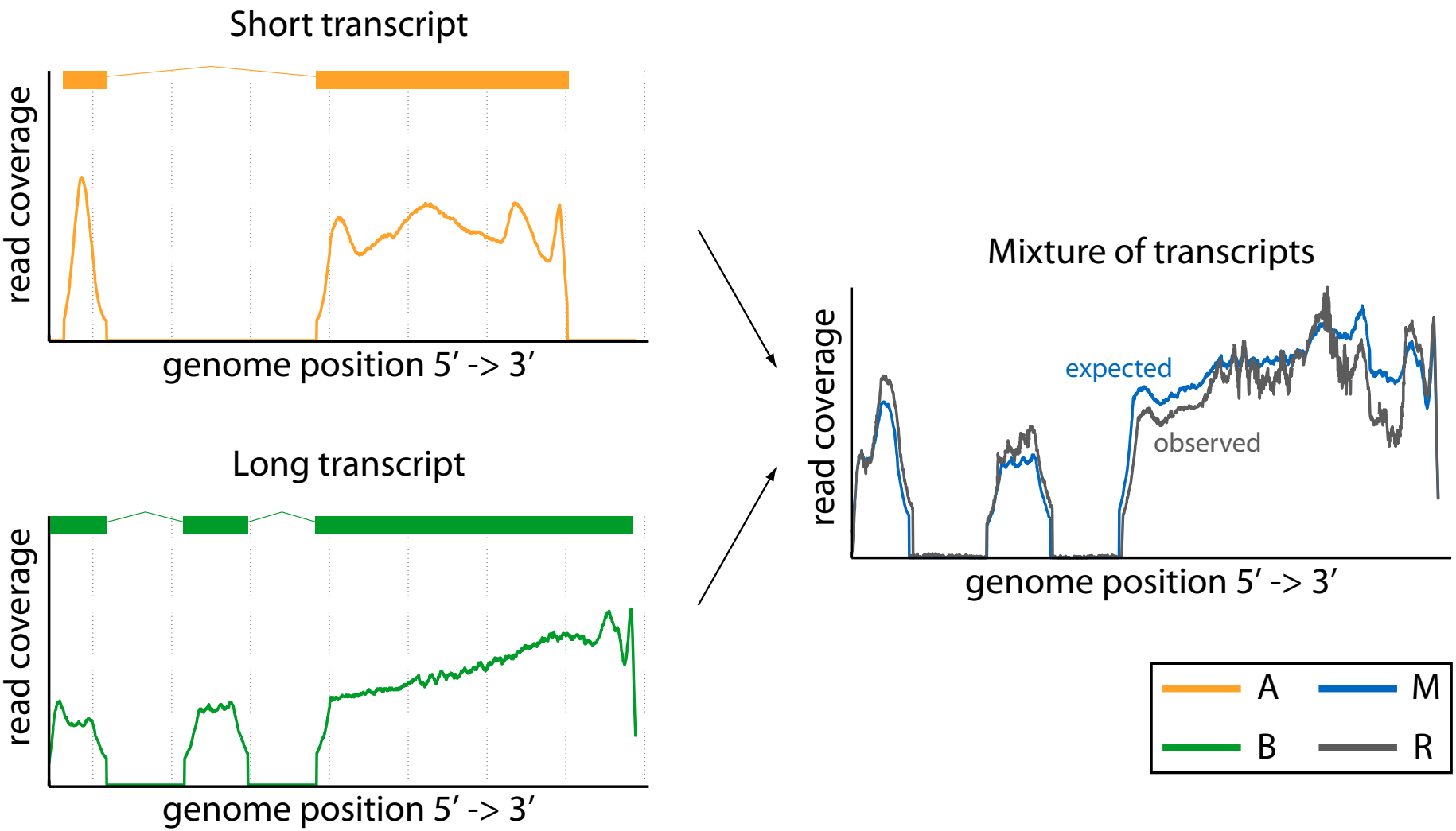


rQuant – Basic Idea



$$M_i = w_A A_i + w_B B_i$$

rQuant – Basic Idea



$$M_i = w_A A_i + w_B B_i \quad \Rightarrow \quad \min_{w_A, w_B} \sum_i \ell(M_i, R_i)$$

Conclusions

- Galaxy is great!

Conclusions

- Galaxy is great!
  - Use galaxy ourselves to perform experiments.

Conclusions

- Galaxy is great!
  - Use galaxy ourselves to perform experiments.
  - Trackster looks great and we have started using it.

Conclusions

- Galaxy is great!
  - Use galaxy ourselves to perform experiments.
  - Trackster looks great and we have started using it.
  - Intend to try cloud-connection again, will also make virtualBox images available soon.

Conclusions

- Galaxy is great!
  - Use galaxy ourselves to perform experiments.
  - Trackster looks great and we have started using it.
  - Intend to try cloud-connection again, will also make virtualBox images available soon.
  - Will suggest NGS sample tracking system to sequencing facility.

Conclusions

- Galaxy is great!
  - Use galaxy ourselves to perform experiments.
  - Trackster looks great and we have started using it.
  - Intend to try cloud-connection again, will also make virtualBox images available soon.
  - Will suggest NGS sample tracking system to sequencing facility.
- Galaxy still needs improvements!
  1. Command line support
  2. Structuring/grouping elements (histories, workflows, . . .)
  3. Package management
  4. Quotas for storage and computing time
  5. Better data library cleanup
  6. Collect list of public servers (with available tools, automatic?)

References I

Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS Comput Biol*, 4(10):e1000173, Oct 2008. doi: 10.1371/journal.pcbi.1000173.

R. Bohnert and G. Rätsch. rQuant.web: a tool for RNA-seq-based transcript quantitation. *NAR Webserver Issue*, 2010.

R. Bohnert, J. Behr, and G Rätsch. Transcript quantification with RNA-Seq data. *BMC Bioinformatics*, 10(S13):P5, 2009. URL `http://www.biomedcentral.com/1471-2105/10/S13/P5`.

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.

G Jean, A Kahles, VT Sreedharan, F F De Bona, and G Rätsch. Rna-seq read alignments with palmapper. submitted, 2010.

Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, April 2009.

Sam E V Linsen, Elzo de Wit, Georges Janssens, Sheila Heater, Laura Chapman, Rachael K Parkin, Brian Fritz, Stacia K Wyman, Ewart de Bruijn, Emile E Voest, Scott Kuersten, Muneesh Tewari, and Edwin Cuppen. Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods*, 6(7):474–476, July 2009.

G Rätsch, G Jean, A Kahles, S Sonnenburg, F De Bona, K Schneeberger, J Hagmann, and D Weigel. PALMapper: Fast and accurate alignment of RNA-seq reads. in preparation, 2010.

M. Sammeth. The Flux Capacitor. *Website*, 2009a. http://flux.sammeth.net/capacitor.html.

M. Sammeth. The Flux Simulator. *Website*, 2009b. http://flux.sammeth.net/simulator.html.

Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biol*, 10(9):R98, Jan 2009a. doi: 10.1186/gb-2009-10-9-r98. URL `http://genomebiology.com/2009/10/9/R98`.

Korbinian Schneeberger, Jörg Hagmann, Stephan Ossowski, Norman Warthmann, Sandra Gesing, Oliver Kohlbacher, and Detlef Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biology*, 10(9):R98, 2009b.

Sebastian J Schultheiss, Wolfgang Busch, Jan U Lohmann, Oliver Kohlbacher, and Gunnar Rätsch. Kirmes: kernel-based identification of regulatory modules in euchromatic sequences. *Bioinformatics*, 25(16):2126–33, Aug 2009. doi: 10.1093/bioinformatics/btp278.

References II

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, 2009a. URL `http://mgene.org/web`. Advance Access published on June 3, 2009.

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*, 37(Web Server issue): W312–W316, July 2009b.

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 2009c.

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*, 2009d. URL `http://genome.cshlp.org/content/early/2009/06/29/gr.090597.108.full.pdf+html`. Advance access June 29, 2009.

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.

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

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.

G. Zeller, S.R. Henz, S. Laubinger, D. Weigel, and G Rätsch. Transcript normalization and segmentation of tiling array data. In *Proceedings Pac. Symp. on Biocomputing*, pages 527–538, 2008a.

Georg Zeller, Stefan R. Henz, Sascha Laubinger, Detlef Weigel, and Gunnar Rätsch. Transcript normalization and segmentation of tiling array data. In *Proceedings Pac. Symp. on Biocomputing*, pages 527–538, 2008b.