# An Introduction to Modern Methods for Transcriptome Analysis

**Gunnar Rätsch**

Friedrich Miescher Laboratory

Max Planck Society, Tübingen, Germany

$10^{th}$ Course in Bioinformatics and Systems Biology
for Molecular Biologists (March 16, 2010)



Friedrich Miescher Laboratory
of the Max Planck Society



# Discovery of the Nuclein
## (Friedrich Miescher, 1869)










**Discovery of Nuclein:**

- from lymphocyte & salmon
- "multi-basic acid" ($\geq 4$)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost

# Discovery of the Nuclein
## (Friedrich Miescher, 1869)









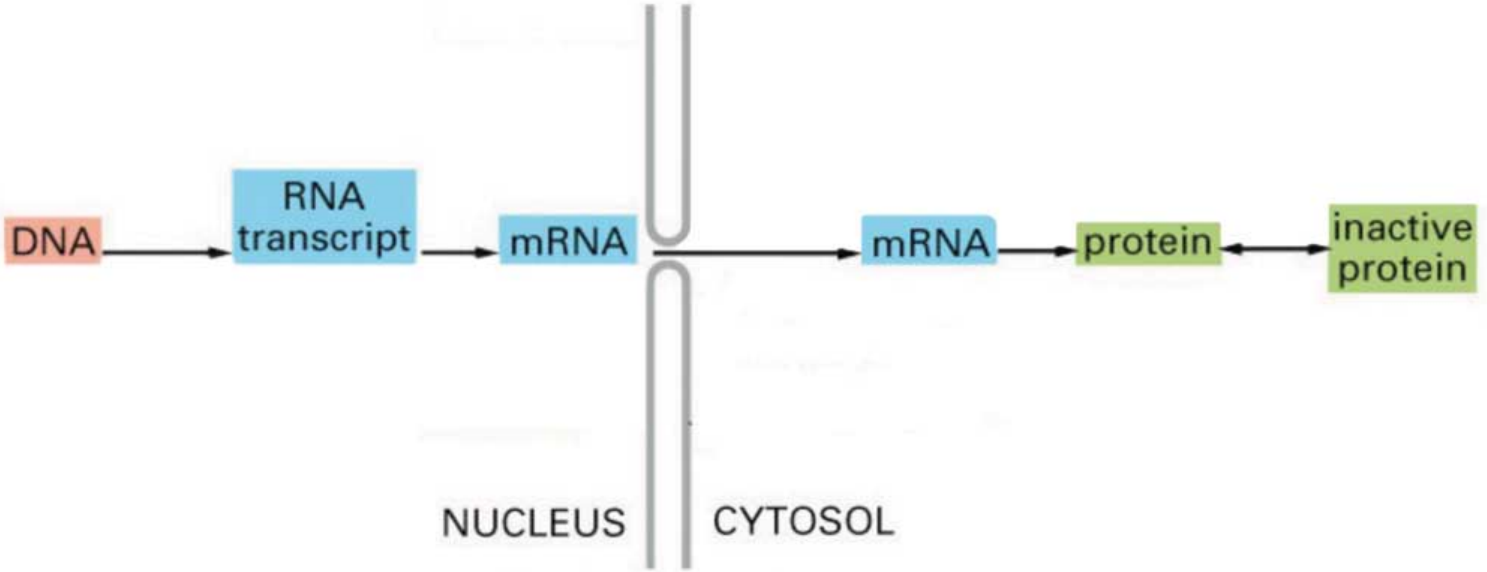

**Discovery of Nuclein:**

- from lymphocyte & salmon
- "multi-basic acid" ($\geq 4$)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost

# Transcriptome Analysis

## ⇝ What is encoded on the genome and how is it processed?




**Try to understand:**

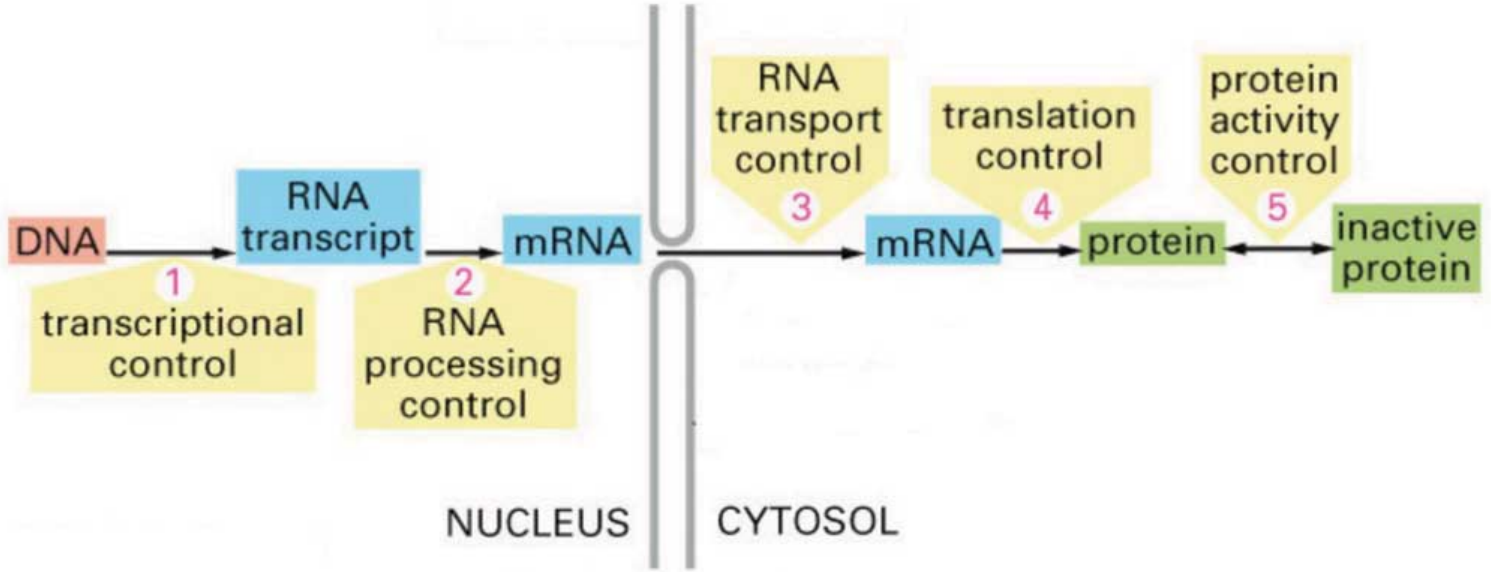
- ⇒ Differences of active components between conditions/organisms?
- ⇒ What changes when perturbing the biological system?

**How to get the transcriptome?**

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome

# Transcriptome Analysis

## ⇝ What is encoded on the genome and how is it processed?




**Try to understand:**

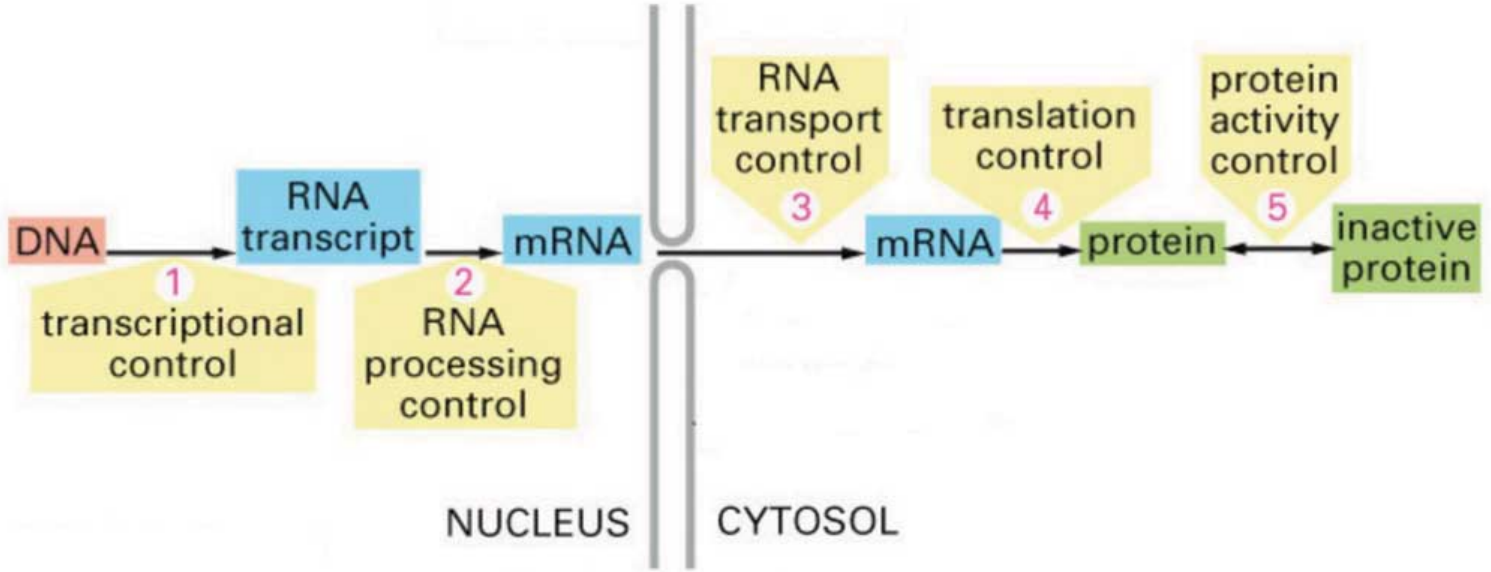
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

**How to get the transcriptome?**

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome

# Transcriptome Analysis

## ⇝ What is encoded on the genome and how is it processed?




**Try to understand:**

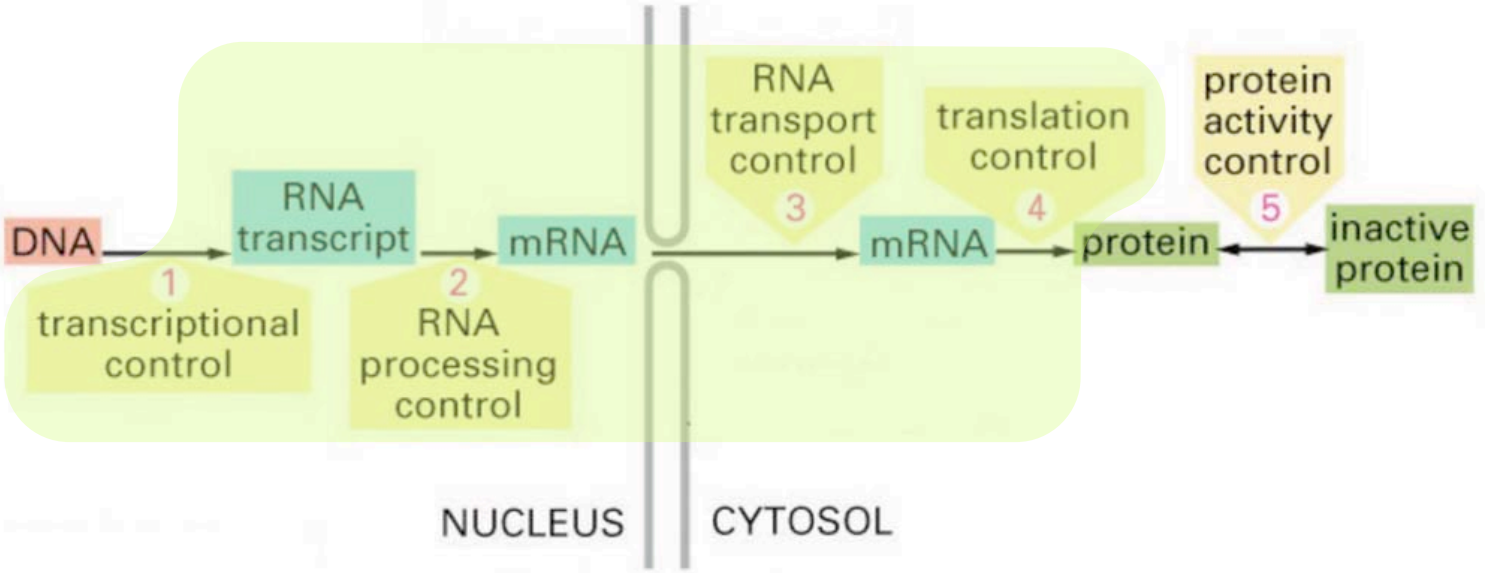
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

**How to get the transcriptome?**

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome

# Transcriptome Analysis

## ⇝ What is encoded on the genome and how is it processed?




**Try to understand:**

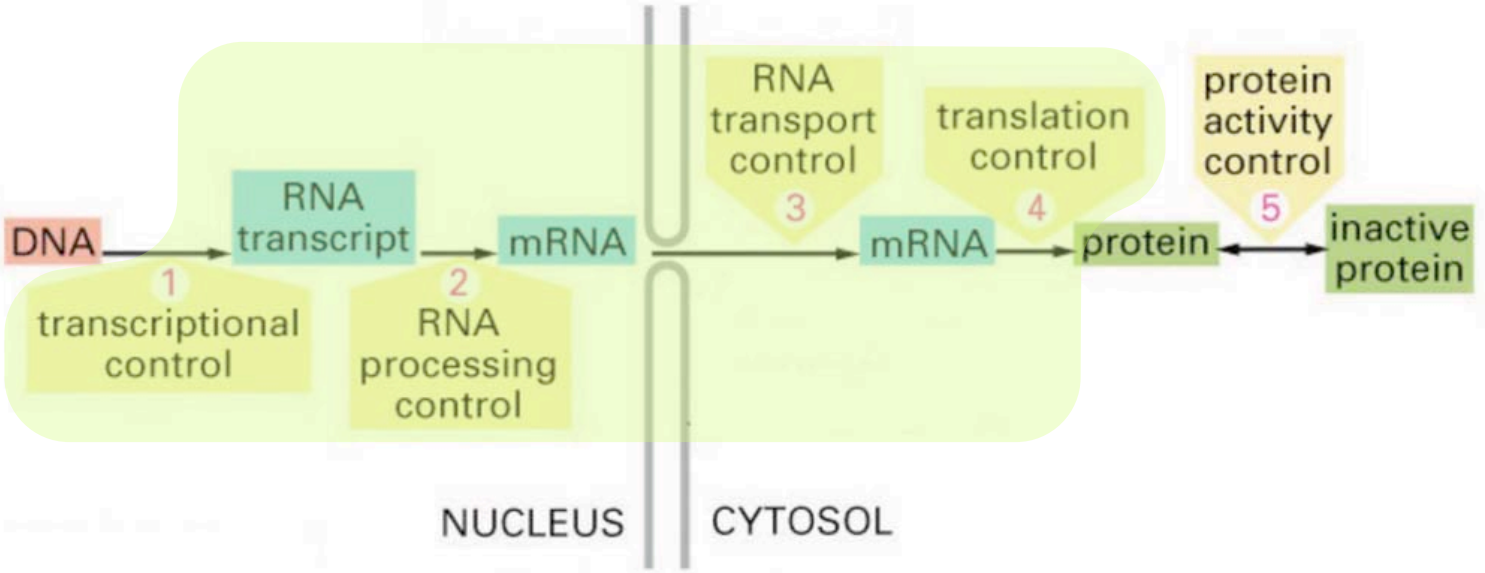
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

**How to get the transcriptome?**

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome

# Transcriptome Analysis

## ⇝ What is encoded on the genome and how is it processed?




**Try to understand:**

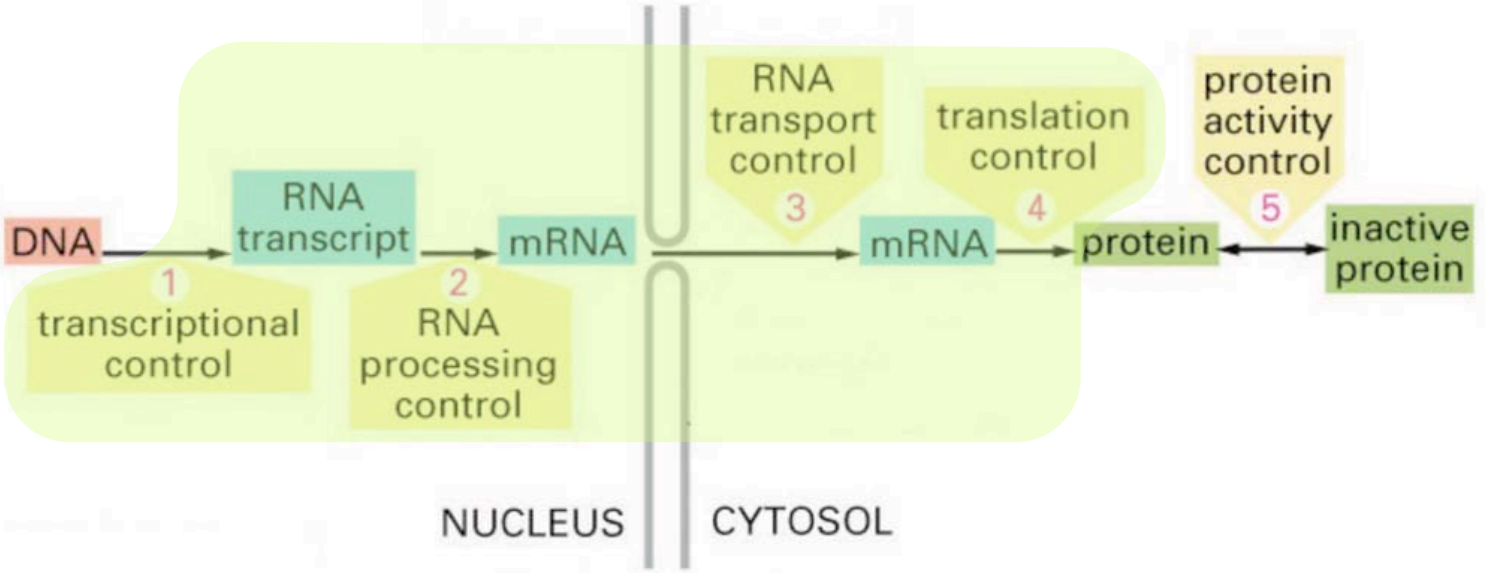
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

**How to get the transcriptome?**

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome

# Transcriptome Analysis

## ⇝ What is encoded on the genome and how is it processed?




**Try to understand:**

- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

**How to get the transcriptome?**

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome

# Transcription & RNA Processing

- Newly synthesized pre-mRNA is capped.

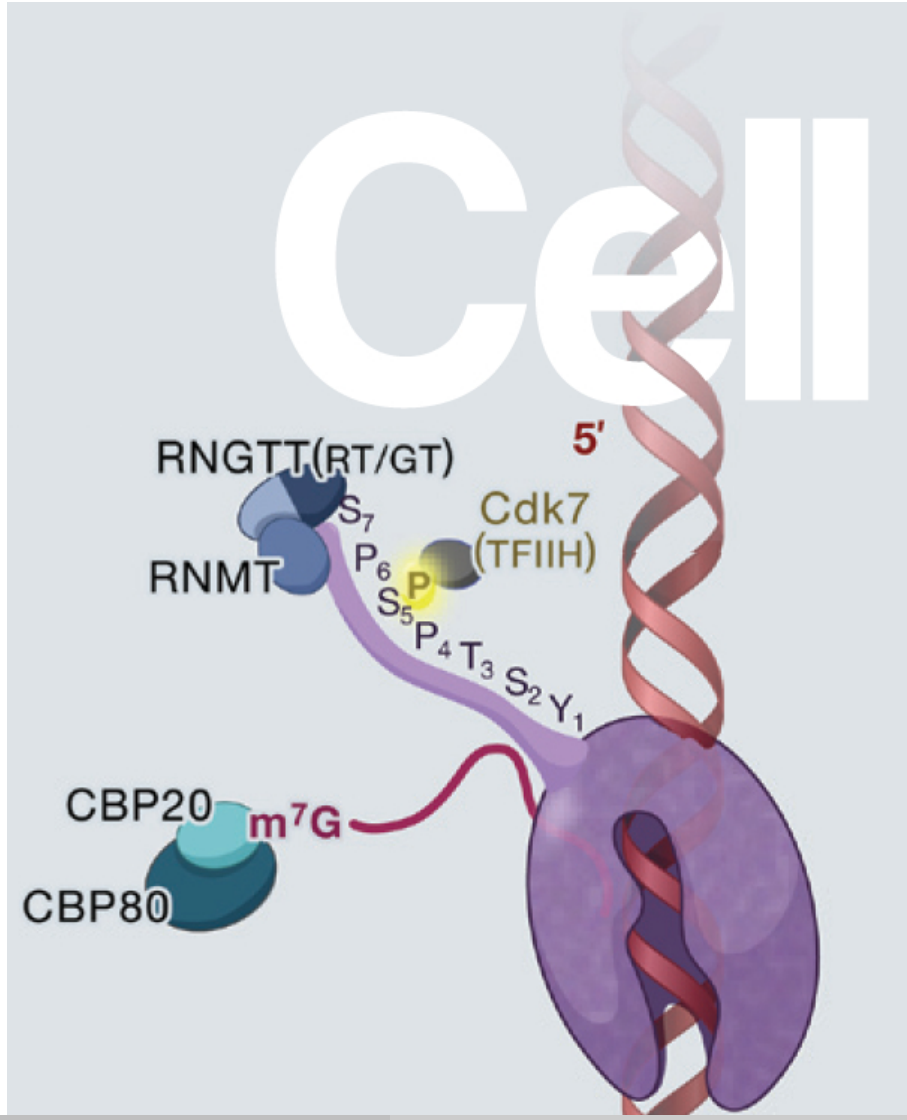
  [CBP20 & CBP80: cap-binding proteins]

- Introns are spliced from pre-mRNA.

  [U1, U2, U4-6: spliceosome

  SF1, U2AF, SR proteins: splicing factors]

- A polyA-tail is added to the 3' terminus of pre-mRNA



# Transcription & RNA Processing

- Newly synthesized pre-mRNA is capped.

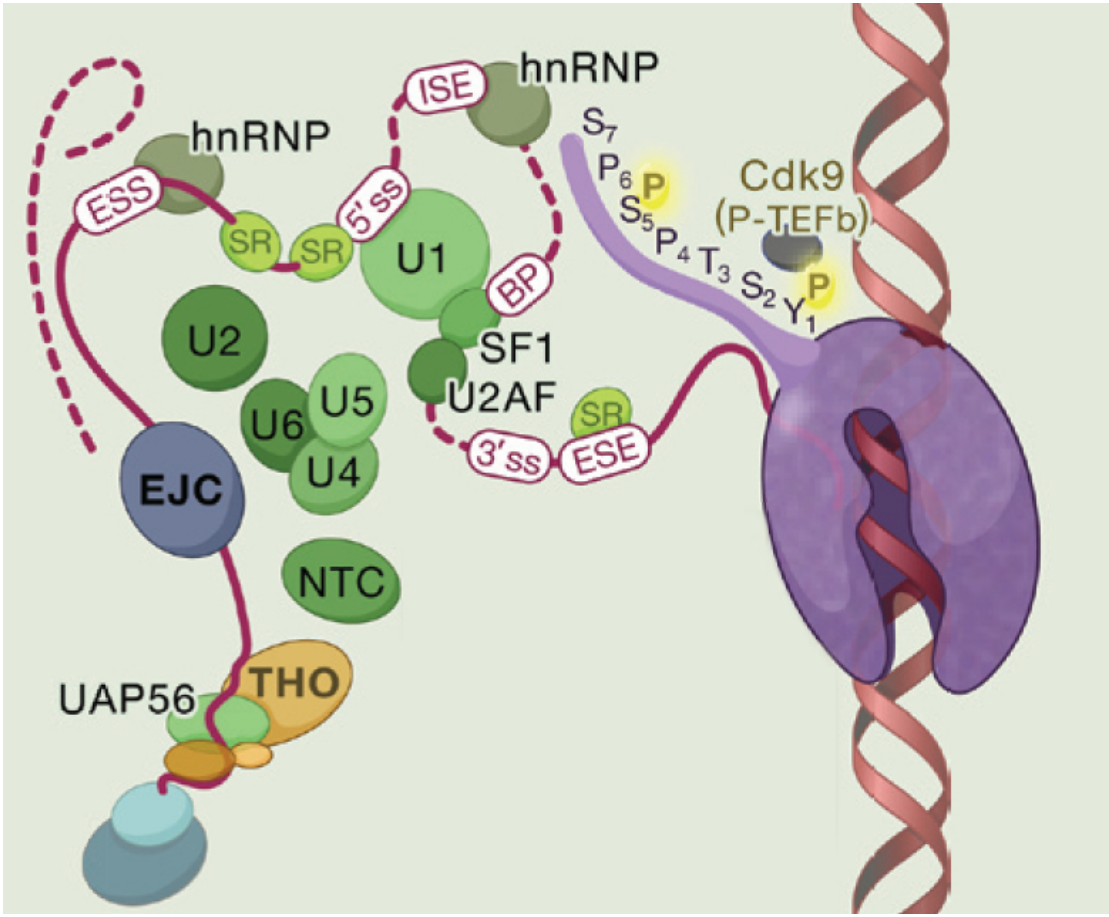
  [CBP20 & CBP80: cap-binding proteins]

- Introns are spliced from pre-mRNA.

  [U1, U2, U4-6: spliceosome

  SF1, U2AF, SR proteins: splicing factors]

- A polyA-tail is added to the 3' terminus of pre-mRNA



# Transcription & RNA Processing

- Newly synthesized pre-mRNA is capped.

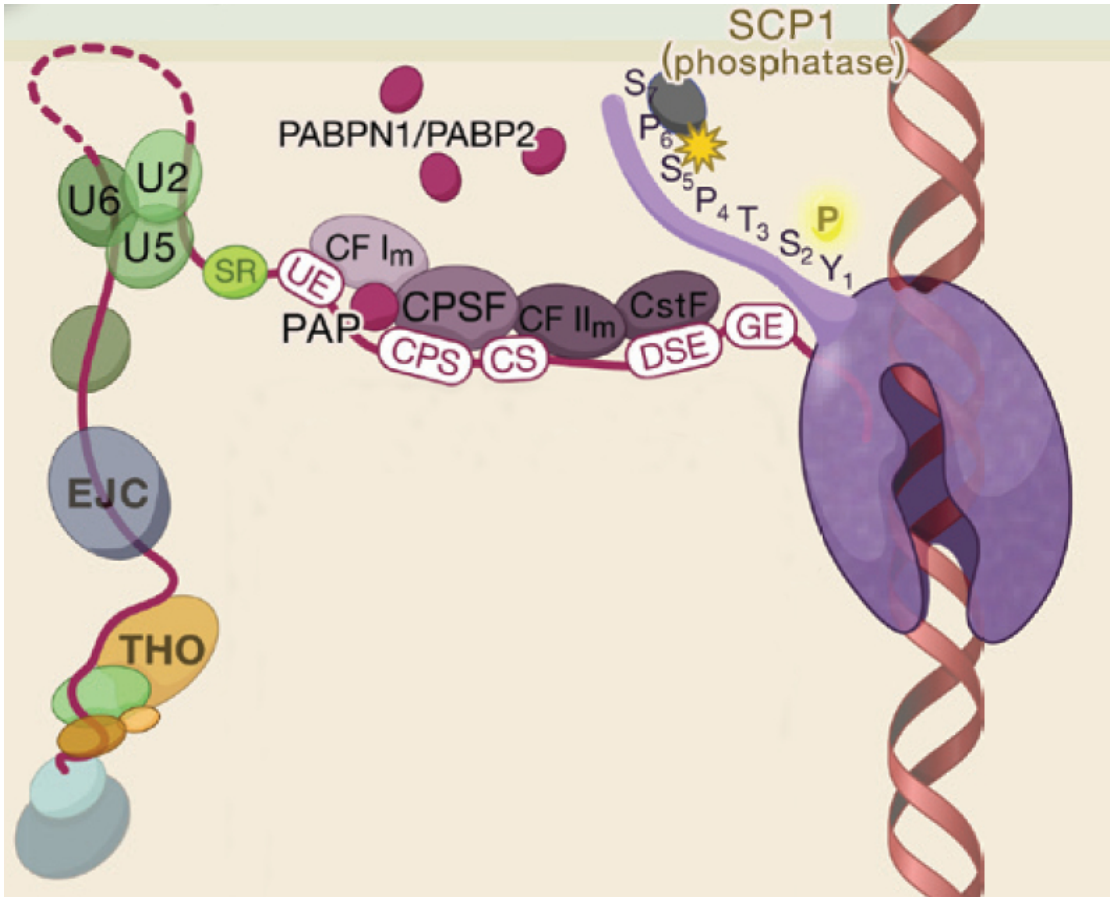
  [CBP20 & CBP80: cap-binding proteins]

- Introns are spliced from pre-mRNA.

  [U1, U2, U4-6: spliceosome

  SF1, U2AF, SR proteins: splicing factors]

- A polyA-tail is added to the 3' terminus of pre-mRNA



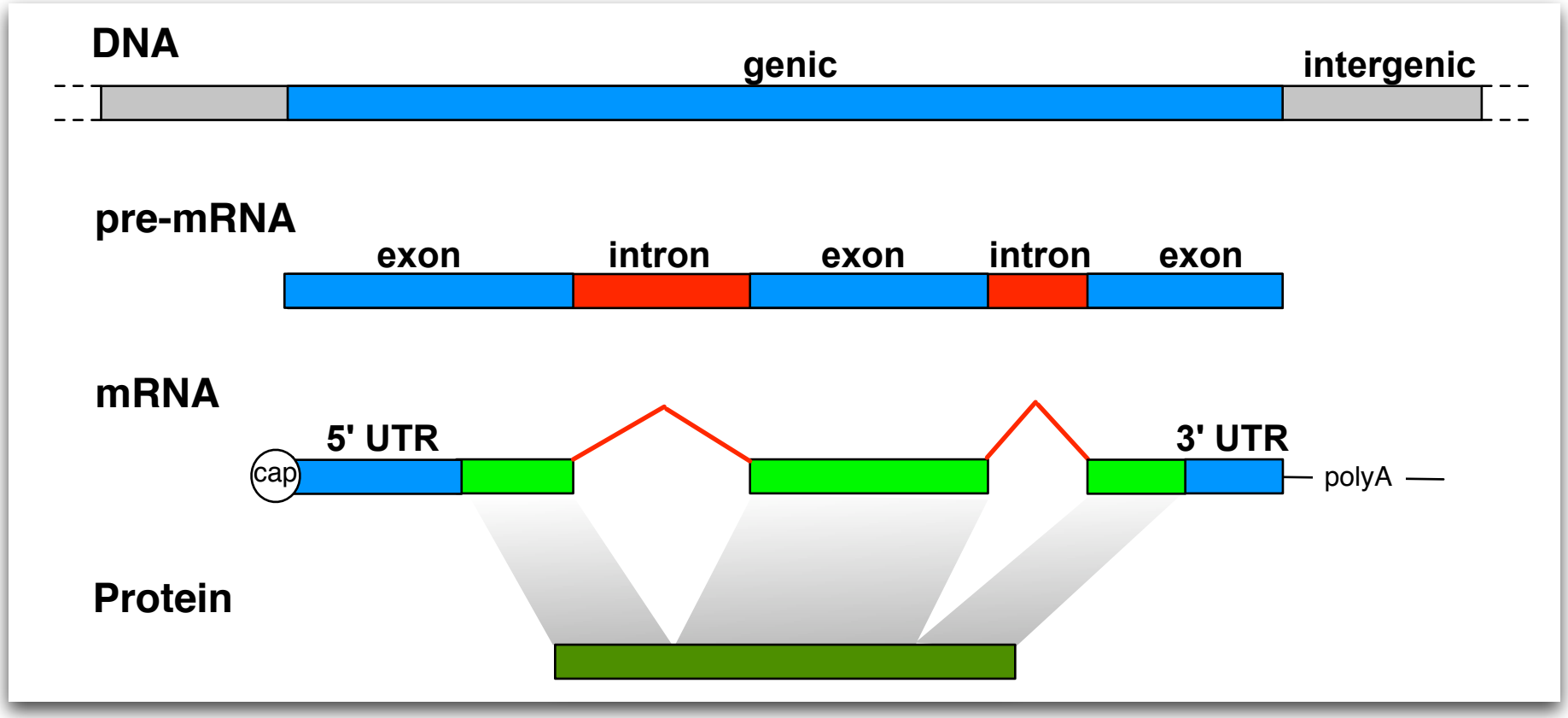
# RNA Transcripts




- Protein-coding mRNAs
- Noncoding RNAs
  - Structural RNAs (e.g. rRNAs, tRNAs, ...)
  - Small RNAs (e.g. miRNAs, endogenous siRNAs, ...)
  - Antisense / promoter-associated transcripts
  - ...

# Computational Gene Finding

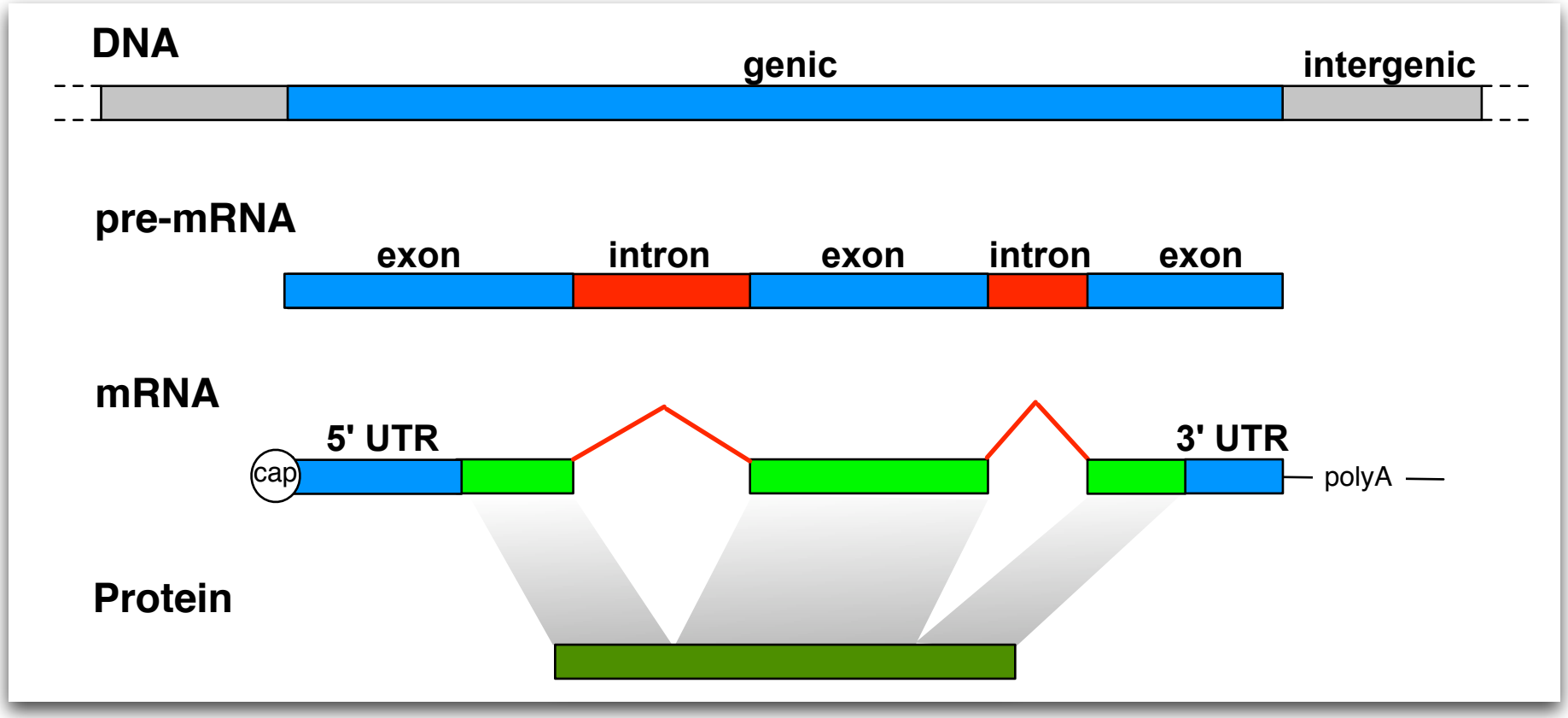
## ⇝ Labeling the Genome




- Given a piece of DNA sequence
- Predict protein-coding mRNAs

# Computational Gene Finding

## ⇝ Labeling the Genome

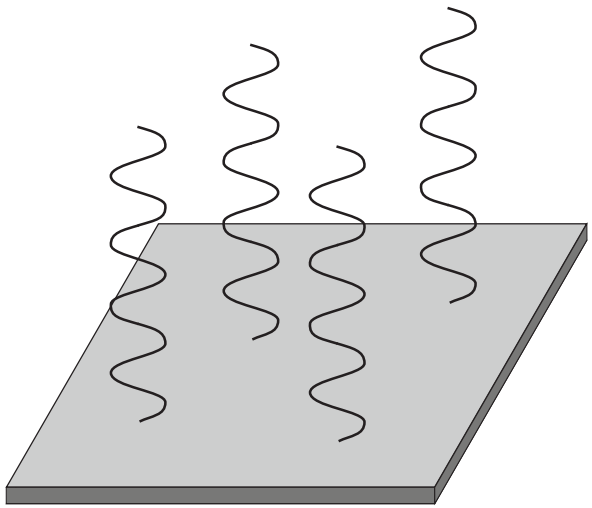



- Given a piece of DNA sequence
- Predict protein-coding mRNAs

# Experimental Characterization of the Transcriptome

## DNA Microarrays

Oligonucleotide probes immobilized on a glass slide hybridize to complementary labeled target RNA.



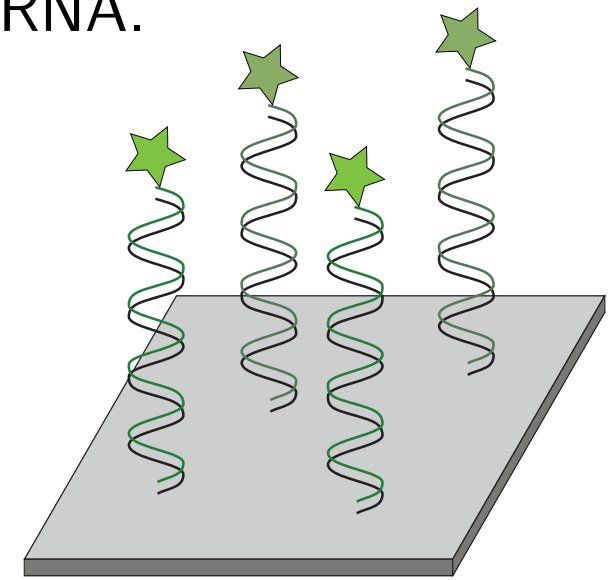
[Wikipedia]

## cDNA Sequencing

# Experimental Characterization of the Transcriptome

## DNA Microarrays

Oligonucleotide probes immobilized on a glass slide hybridize to complementary labeled target RNA.

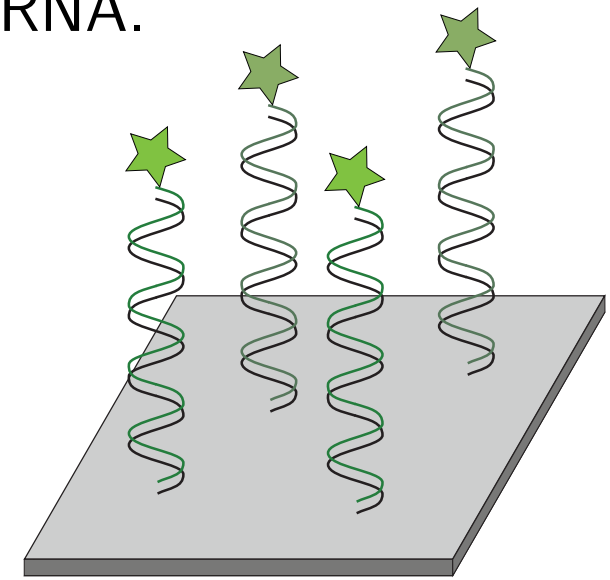


[Wikipedia]

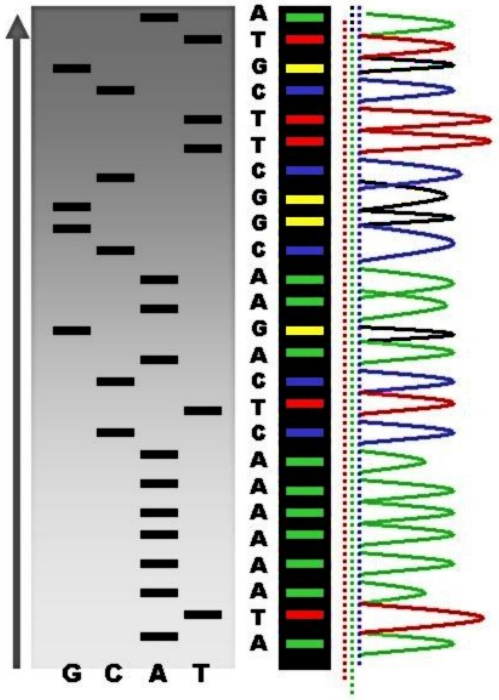
# Experimental Characterization of the Transcriptome

## DNA Microarrays

Oligonucleotide probes immobilized on a glass slide hybridize to complementary labeled target RNA.



## cDNA Sequencing




[Wikipedia]

# Key Research Questions




- Characterize an organism's full complement of genes
  - $\Rightarrow$ Find new (possibly noncoding) genes
  - $\Rightarrow$ Compare genes among organisms
- Characterize transcript isoforms
  - $\Rightarrow$ Find new alternative splice forms / transcript ends
- Monitor transcriptome changes between tissues or in response to environmental changes (e.g. stress)
  - $\Rightarrow$ Identify significant expression changes
- Understand transcriptome regulation
  - $\Rightarrow$ Knock-out / knock-down analysis of regulators
    - $\rightsquigarrow$ Identify regulated targets with significant expression changes
  - $\Rightarrow$ Identify binding sites used in regulation (e.g. ChIP-on-chip)
  - $\Rightarrow$ QTL-Mapping of transcriptome-phenotypes

# Key Research Questions




- Characterize an organism's full complement of genes
  - $\Rightarrow$ Find new (possibly noncoding) genes
  - $\Rightarrow$ Compare genes among organisms
- Characterize transcript isoforms
  - $\Rightarrow$ Find new alternative splice forms / transcript ends
- Monitor transcriptome changes between tissues or in response to environmental changes (e.g. stress)
  - $\Rightarrow$ Identify significant expression changes
- Understand transcriptome regulation
  - $\Rightarrow$ Knock-out / knock-down analysis of regulators
    - $\rightsquigarrow$ Identify regulated targets with significant expression changes
  - $\Rightarrow$ Identify binding sites used in regulation (e.g. ChIP-on-chip)
  - $\Rightarrow$ QTL-Mapping of transcriptome-phenotypes

# Key Research Questions




- Characterize an organism's full complement of genes
  - ⇒ Find new (possibly noncoding) genes
  - ⇒ Compare genes among organisms
- Characterize transcript isoforms
  - ⇒ Find new alternative splice forms / transcript ends
- Monitor transcriptome changes between tissues or in response to environmental changes (e.g. stress)
  - ⇒ Identify significant expression changes
- Understand transcriptome regulation
  - ⇒ Knock-out / knock-down analysis of regulators
    - ⇝ Identify regulated targets with significant expression changes
  - ⇒ Identify binding sites used in regulation (e.g. ChIP-on-chip)
  - ⇒ QTL-Mapping of transcriptome-phenotypes

# Key Research Questions




- Characterize an organism's full complement of genes
  - $\Rightarrow$ Find new (possibly noncoding) genes
  - $\Rightarrow$ Compare genes among organisms
- Characterize transcript isoforms
  - $\Rightarrow$ Find new alternative splice forms / transcript ends
- Monitor transcriptome changes between tissues or in response to environmental changes (e.g. stress)
  - $\Rightarrow$ Identify significant expression changes
- Understand transcriptome regulation
  - $\Rightarrow$ Knock-out / knock-down analysis of regulators
    - $\rightsquigarrow$ Identify regulated targets with significant expression changes
  - $\Rightarrow$ Identify binding sites used in regulation (e.g. ChIP-on-chip)
  - $\Rightarrow$ QTL-Mapping of transcriptome-phenotypes

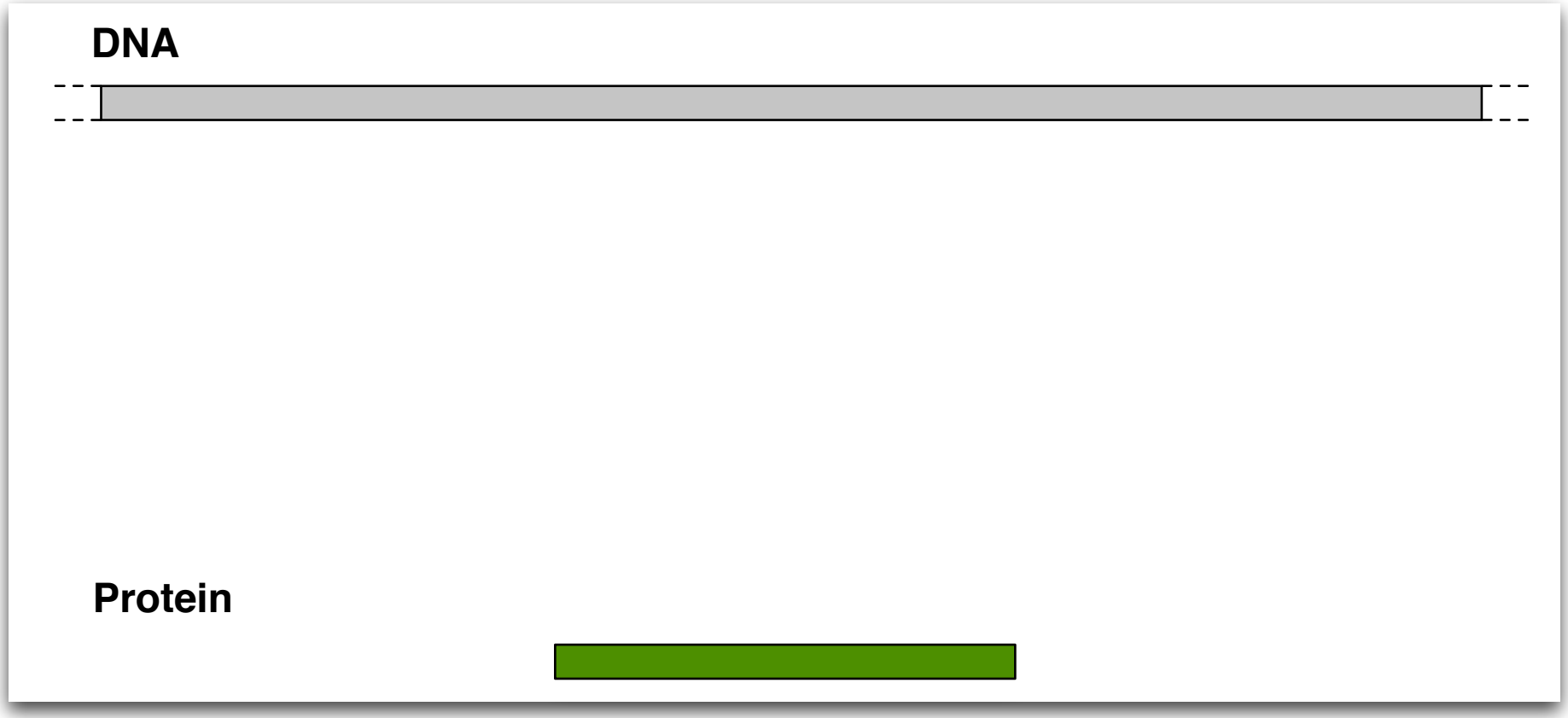
# Roadmap




1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Detection of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology and Limitations
   - Assembly & Read Mapping
   - Transcript Quantitation
   - De novo Transcript Prediction

# Computational Gene Finding

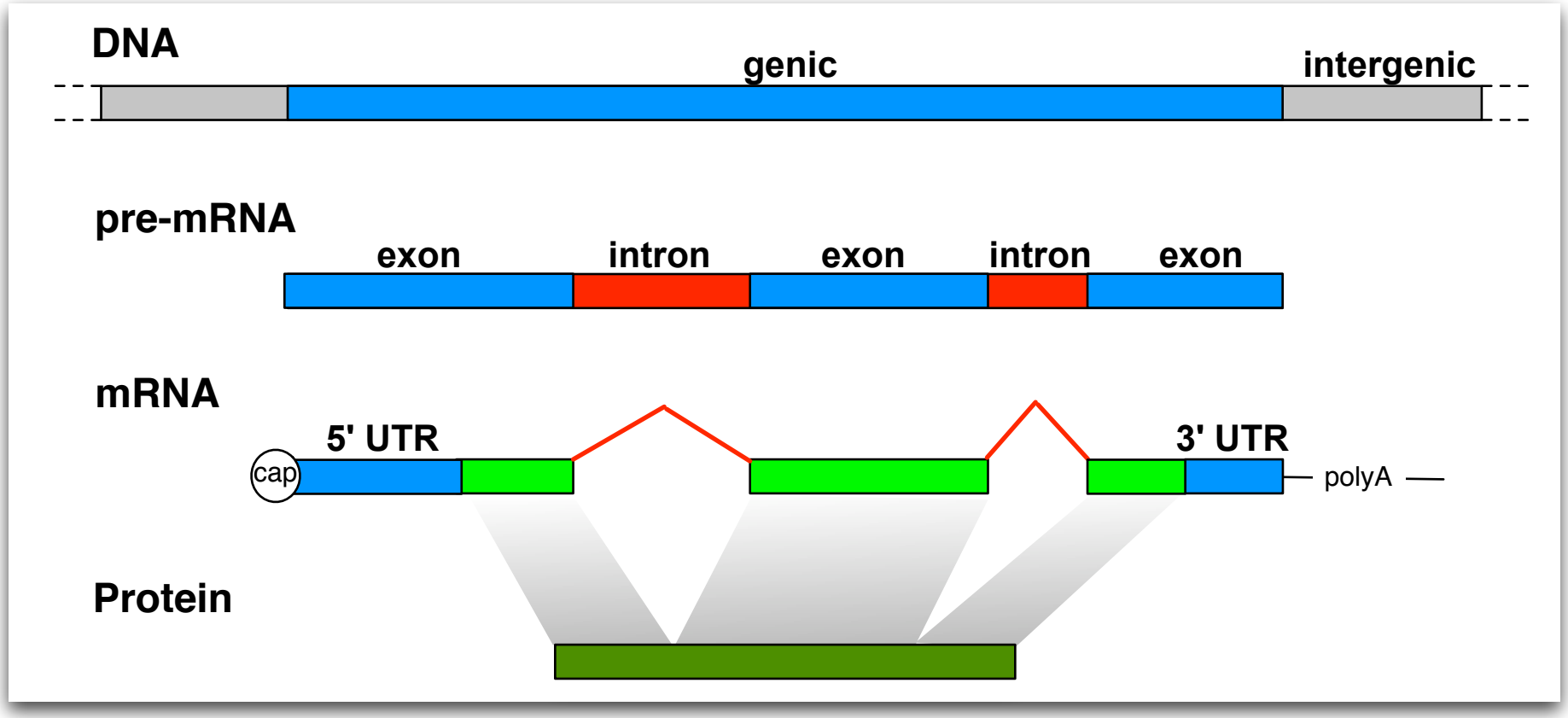
↝ Labeling the Genome




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

# Computational Gene Finding

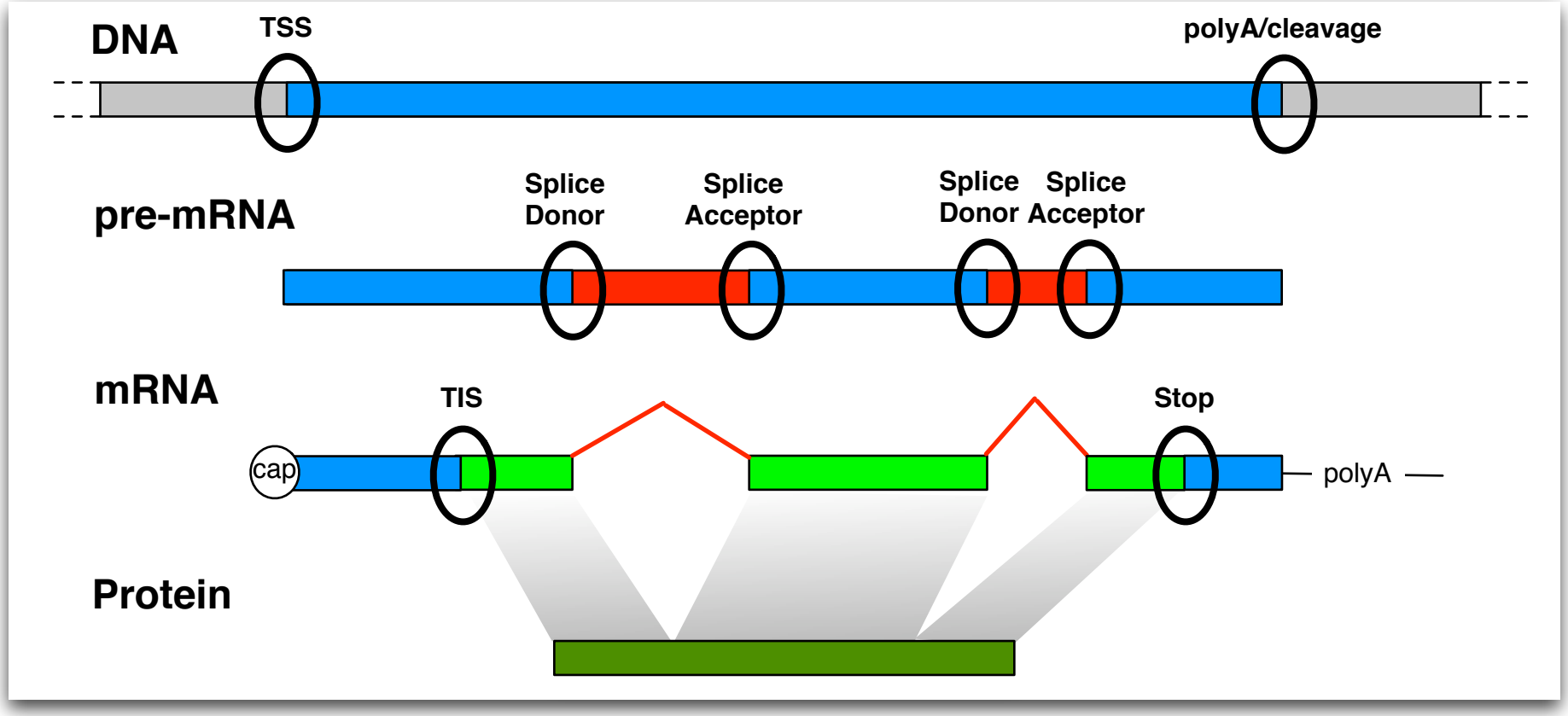
## ⇝ Labeling the Genome




- Given a piece of DNA sequence
- Predict the correct corresponding **label sequence** with labels

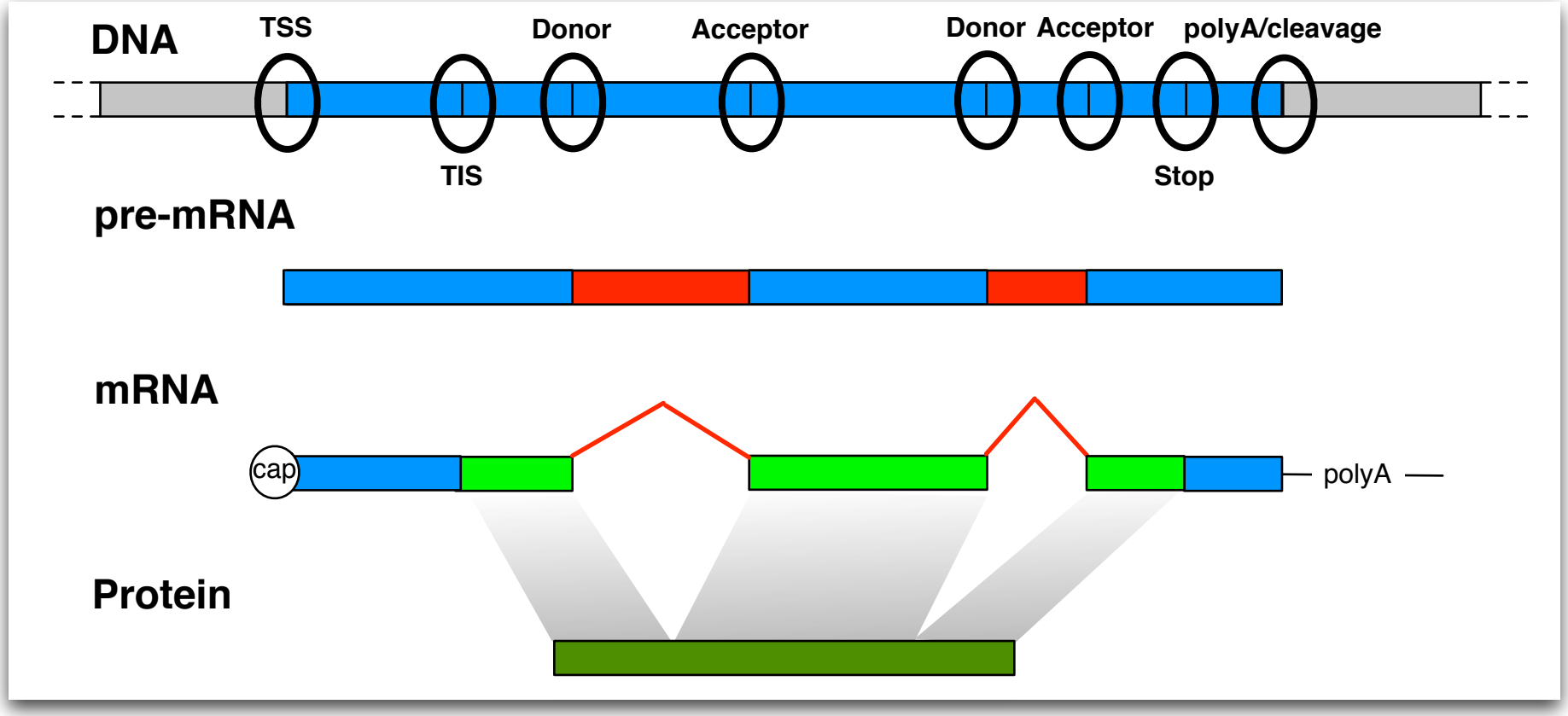
# Computational Gene Finding

## ⇝ Labeling the Genome



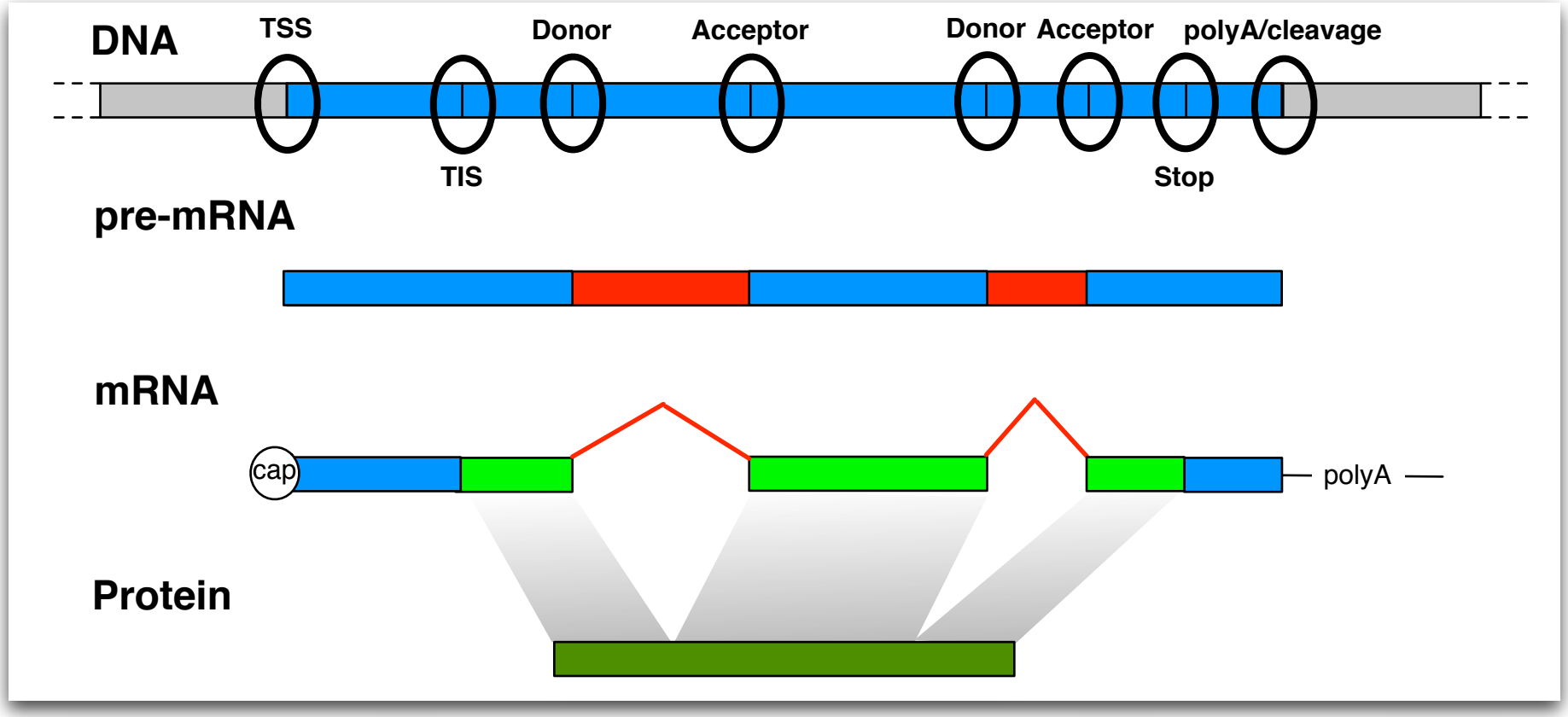
# Computational Gene Finding

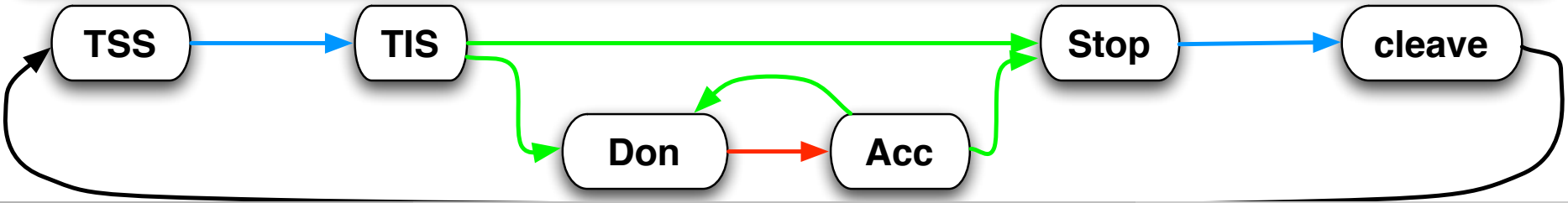
## ⇝ Labeling the Genome



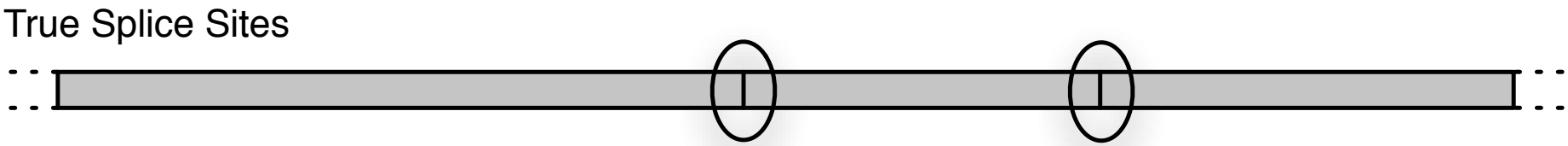
# Computational Gene Finding

## ⇝ Labeling the Genome



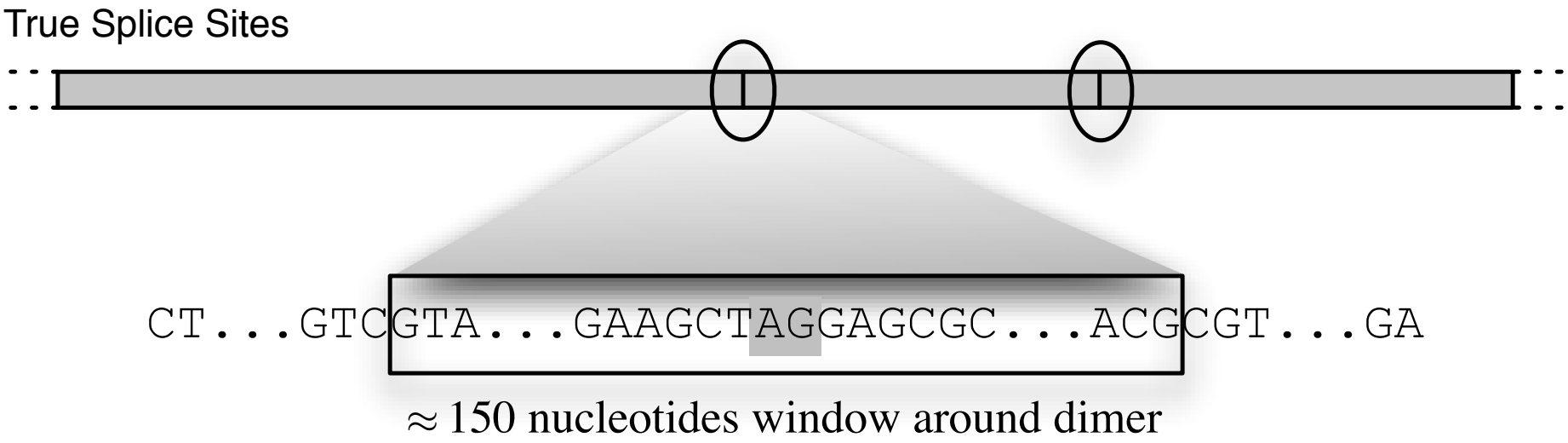
# Example: Splice Site Recognition




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

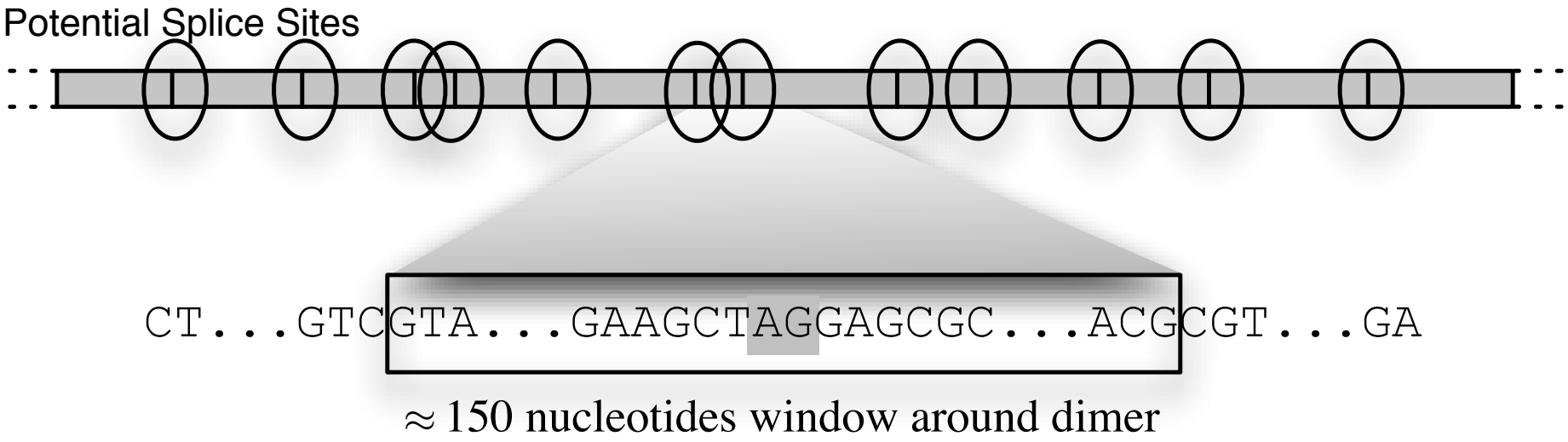
$\Rightarrow$ **Millions of labeled instances from EST databases**

# Example: Splice Site Recognition



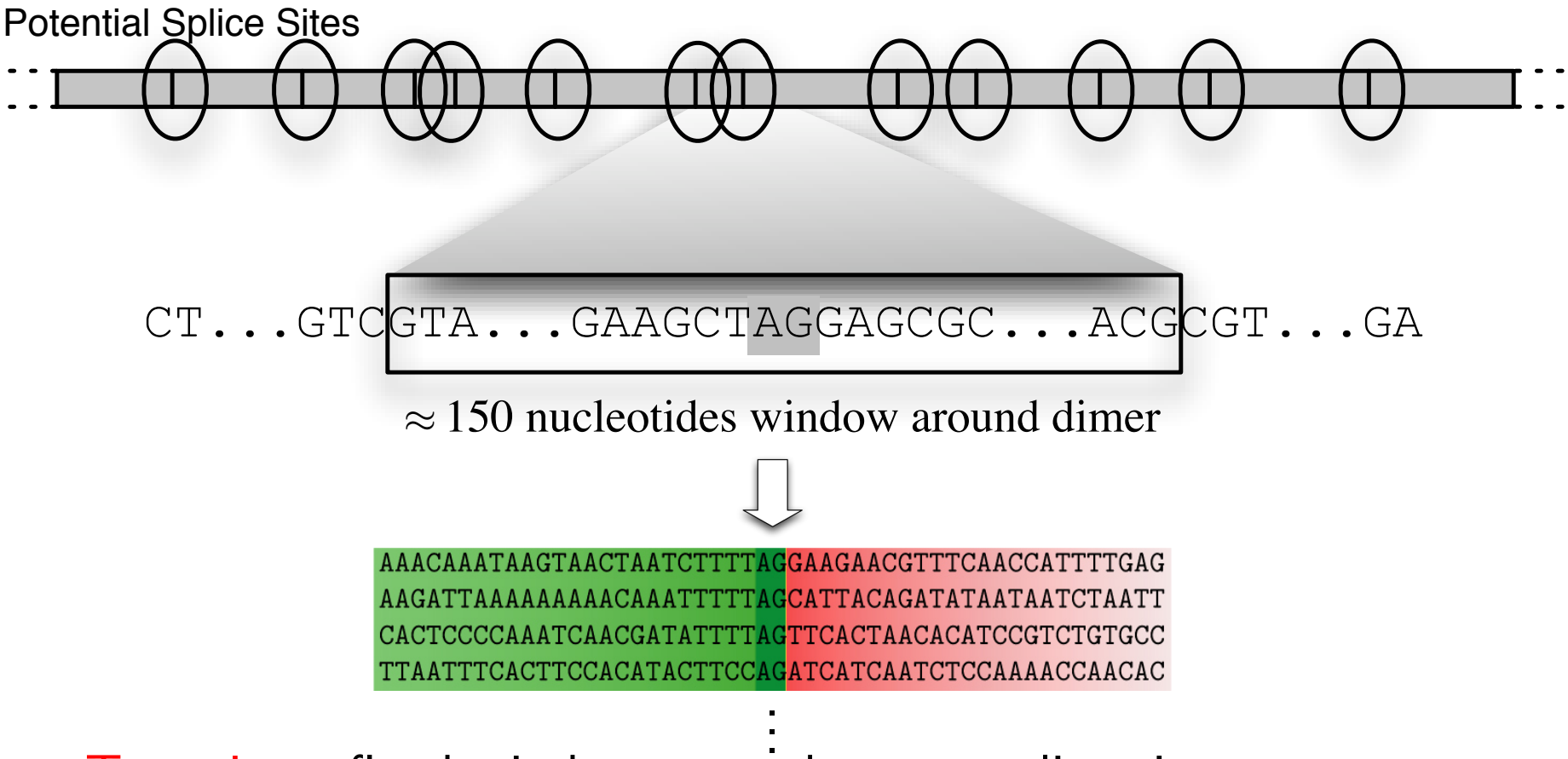

- True sites: fixed window around a true splice site

# Example: Splice Site Recognition




- True sites: fixed window around a true splice site

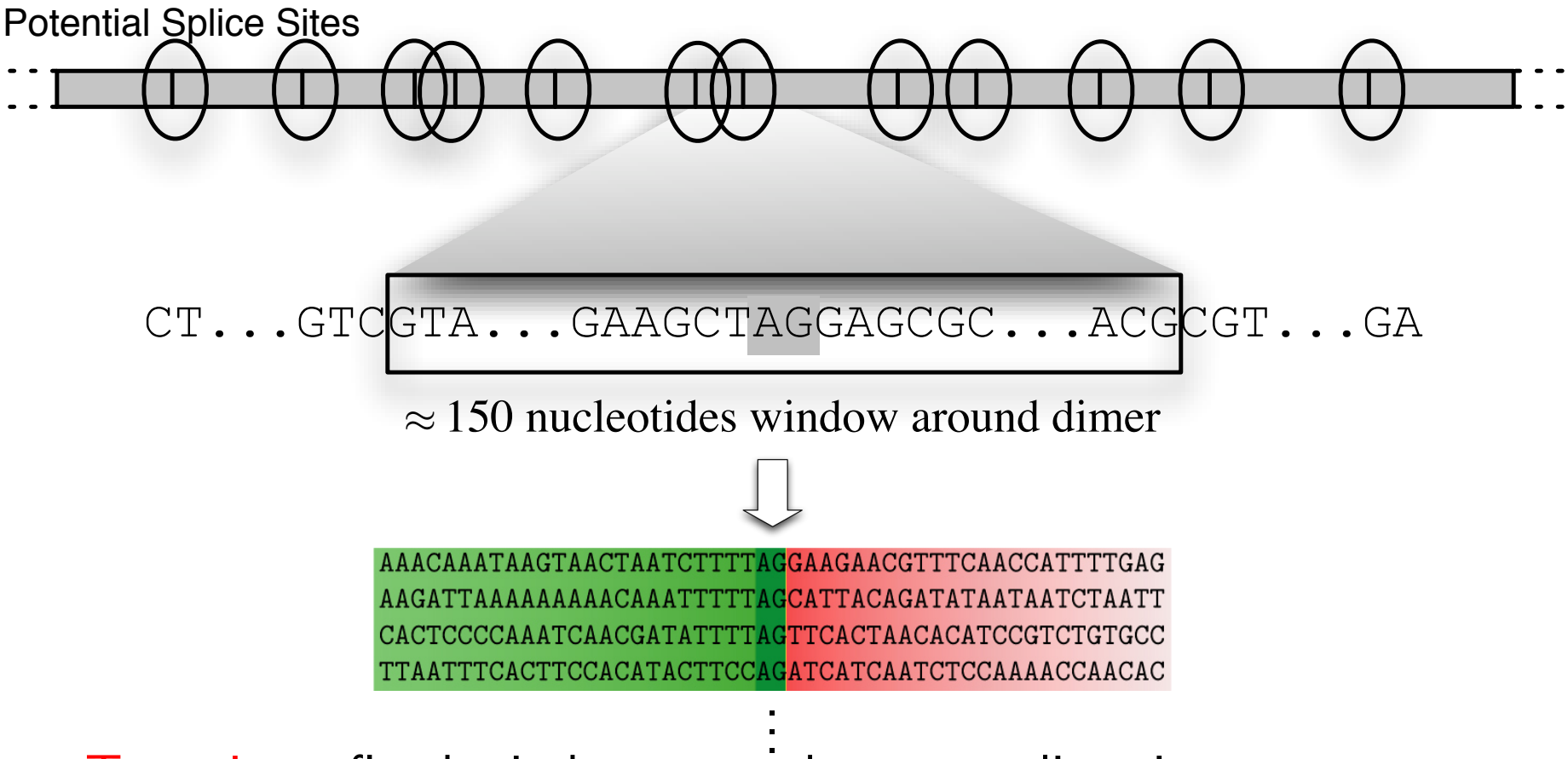
# Example: Splice Site Recognition




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

⇒ Millions of labeled instances from EST databases

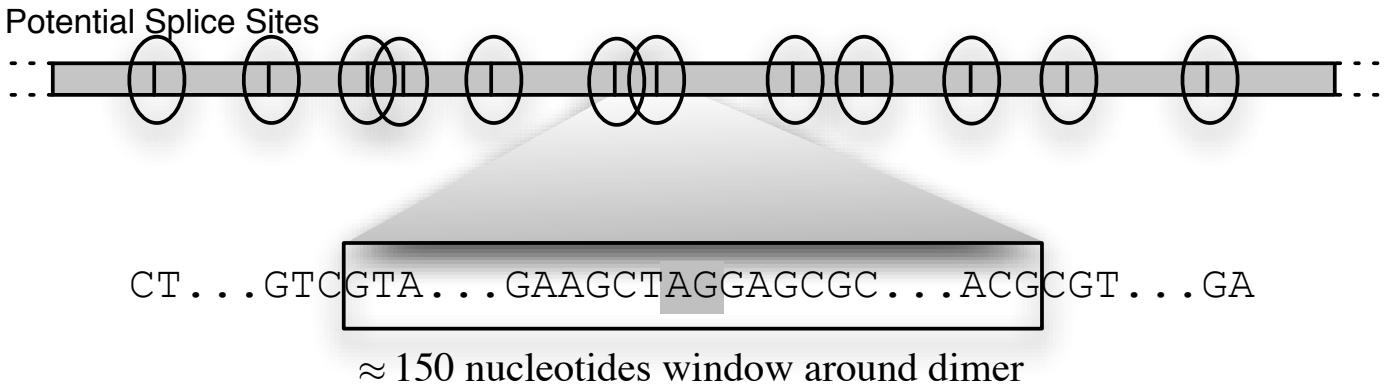
# Example: Splice Site Recognition



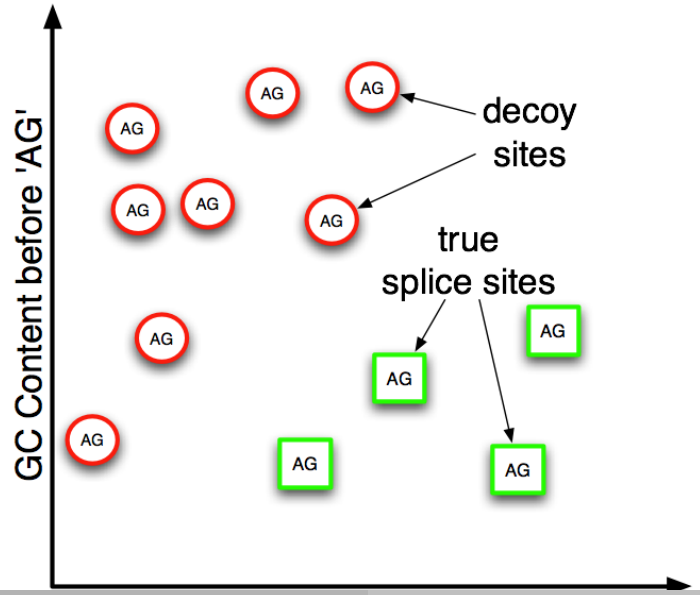

- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

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

# Example: Splice Site Recognition




## Basic idea:



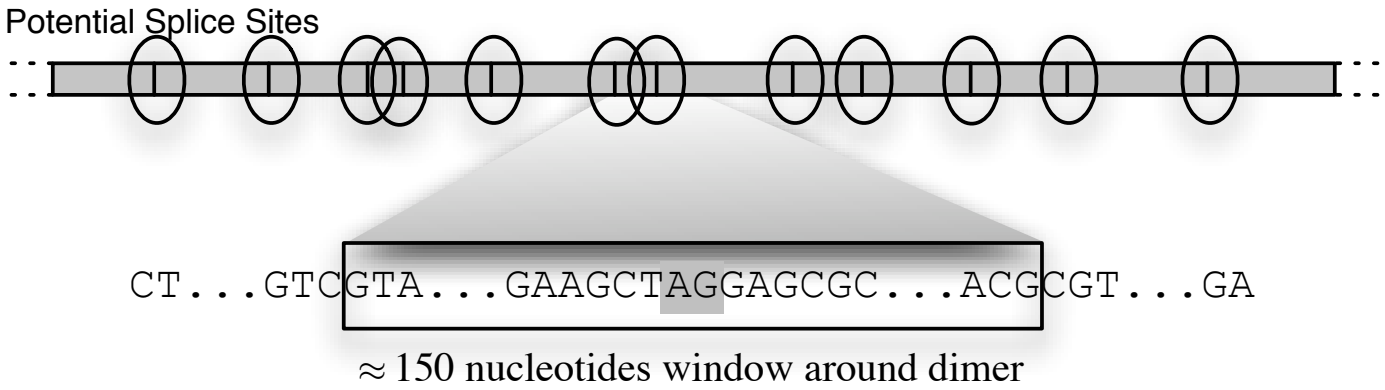

For instance, exploit that exons have higher GC content

or

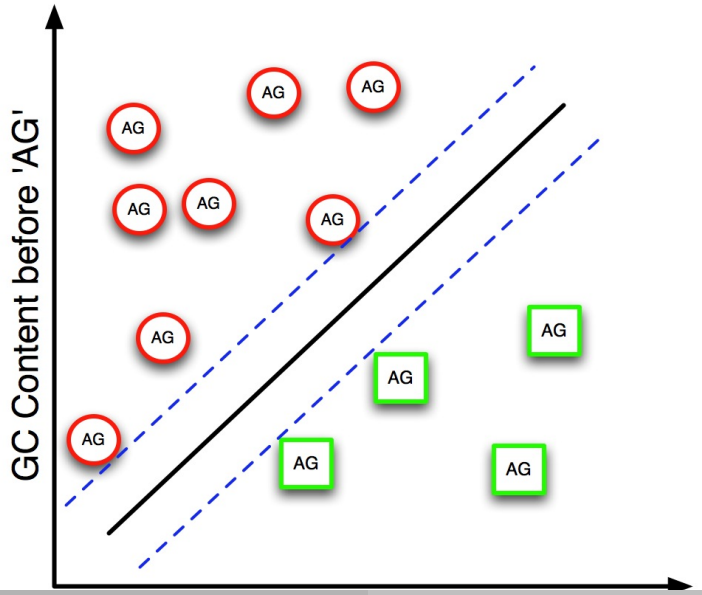
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

# Example: Splice Site Recognition




## Basic idea:



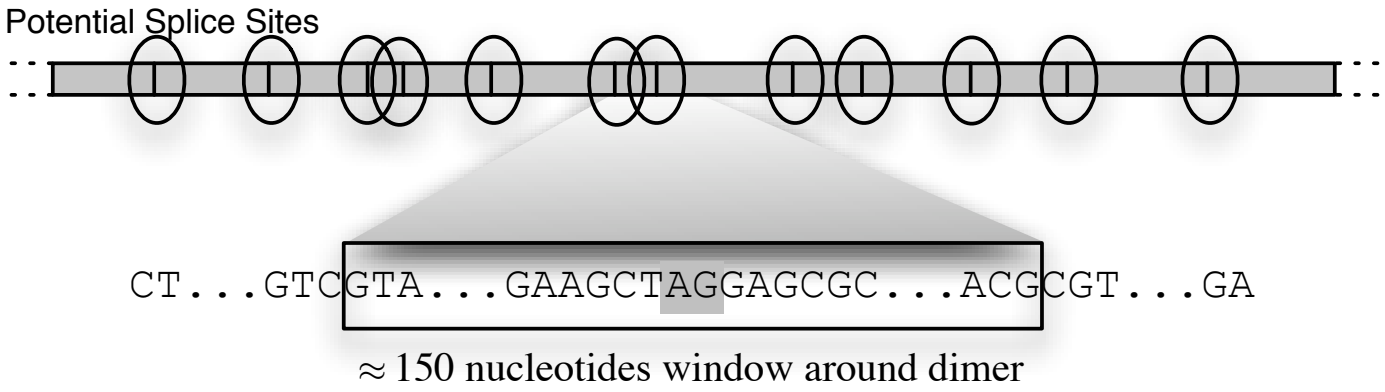

For instance, exploit that exons have higher GC content

or

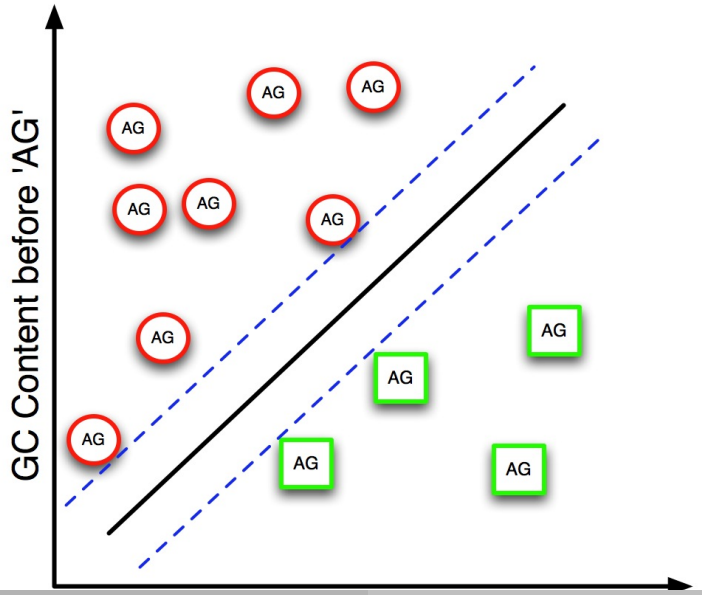
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

# Example: Splice Site Recognition




## Basic idea:



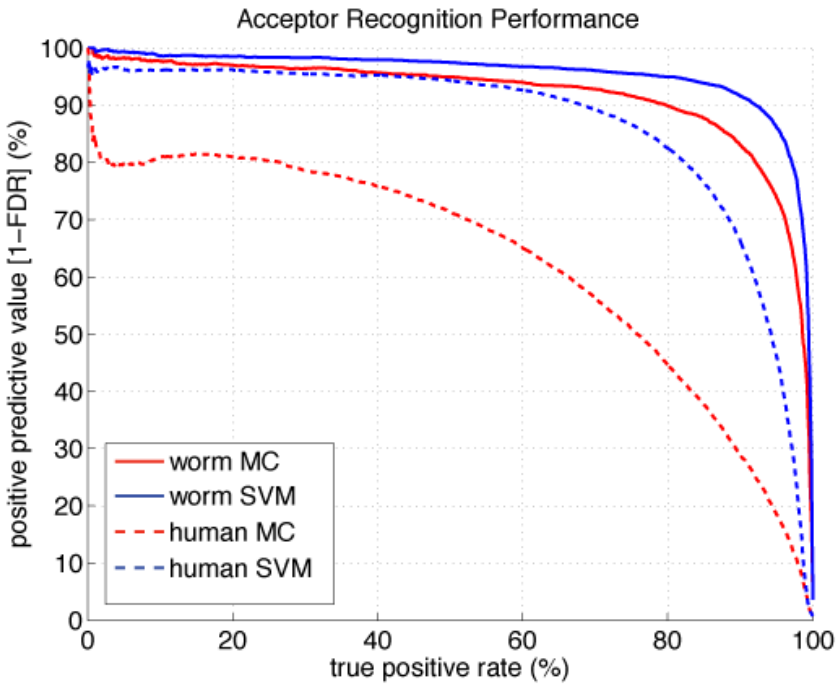

**In practice:** Use one feature per possible substring (e.g. $\leq$20) at all positions

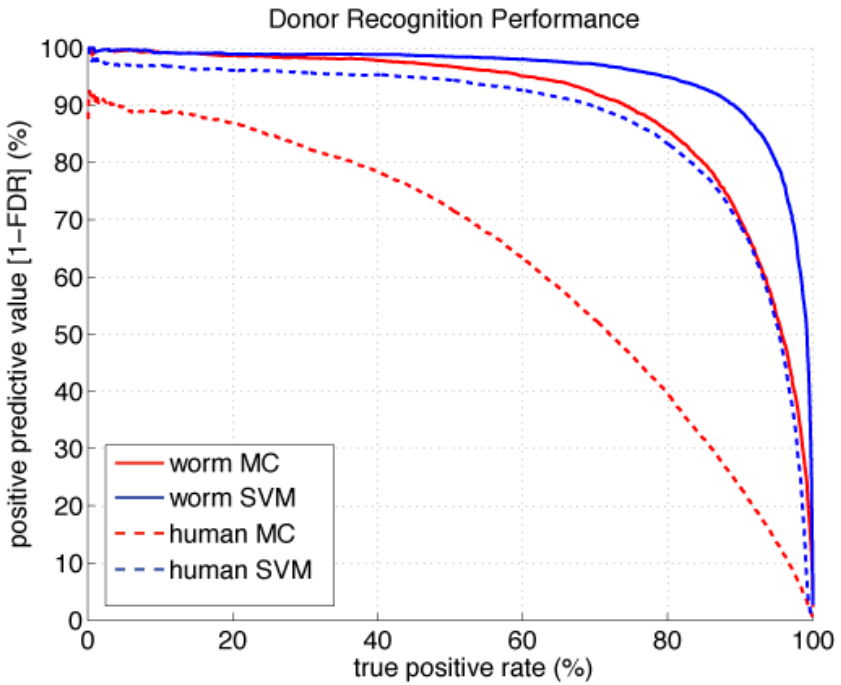
$150 \cdot (4^1 + \ldots + 4^{20}) \approx 2 \cdot 10^{14}$ features

[Sonnenburg et al., 2007]

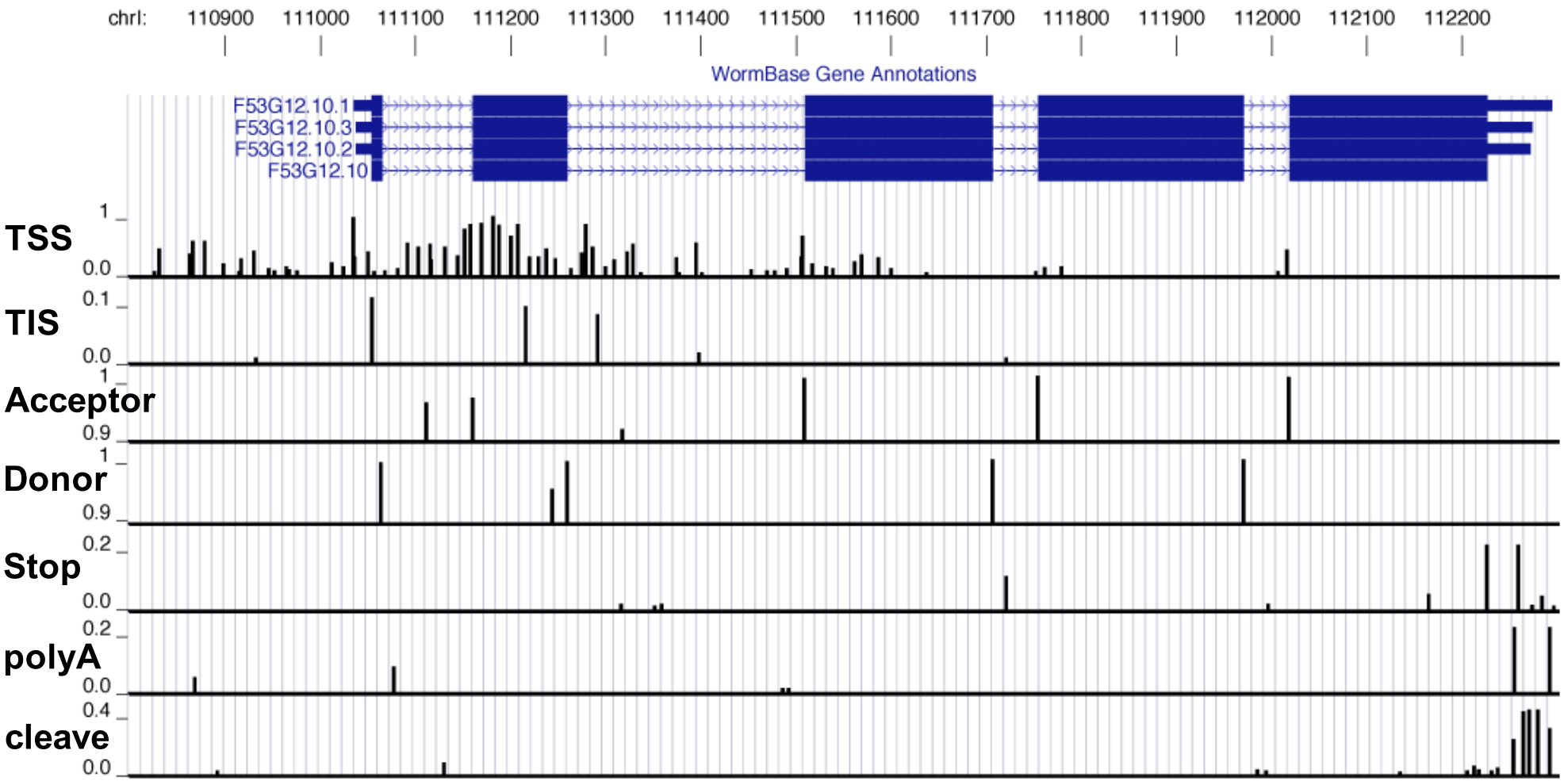
# Results on Splice Site Recognition

| | Worm | | Fly | | Cress | | Fish | | Human | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Don | Acc | Don | Acc | Don | Acc | Don | Acc | Don |
| **Markov Chain** auPRC(%) | 92.1 | 90.0 | 80.3 | 78.5 | 87.4 | 88.2 | 63.6 | 62.9 | 16.2 | 26.0 |
| **SVM** auPRC(%) | 95.9 | 95.3 | 86.7 | 87.5 | 92.2 | 92.9 | 86.6 | 86.9 | 54.4 | 56.9 |



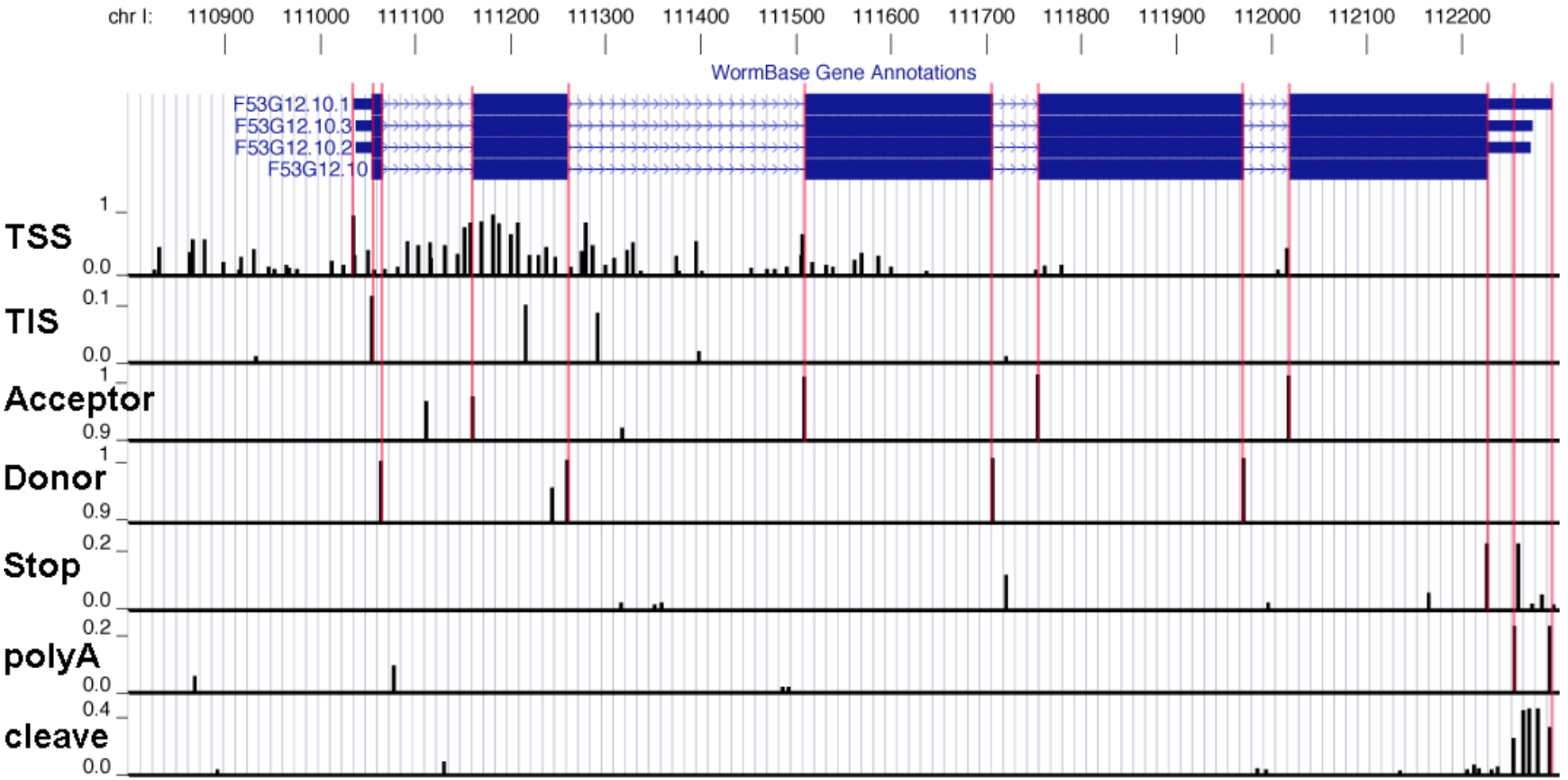



# Example: Predictions in UCSC Browser



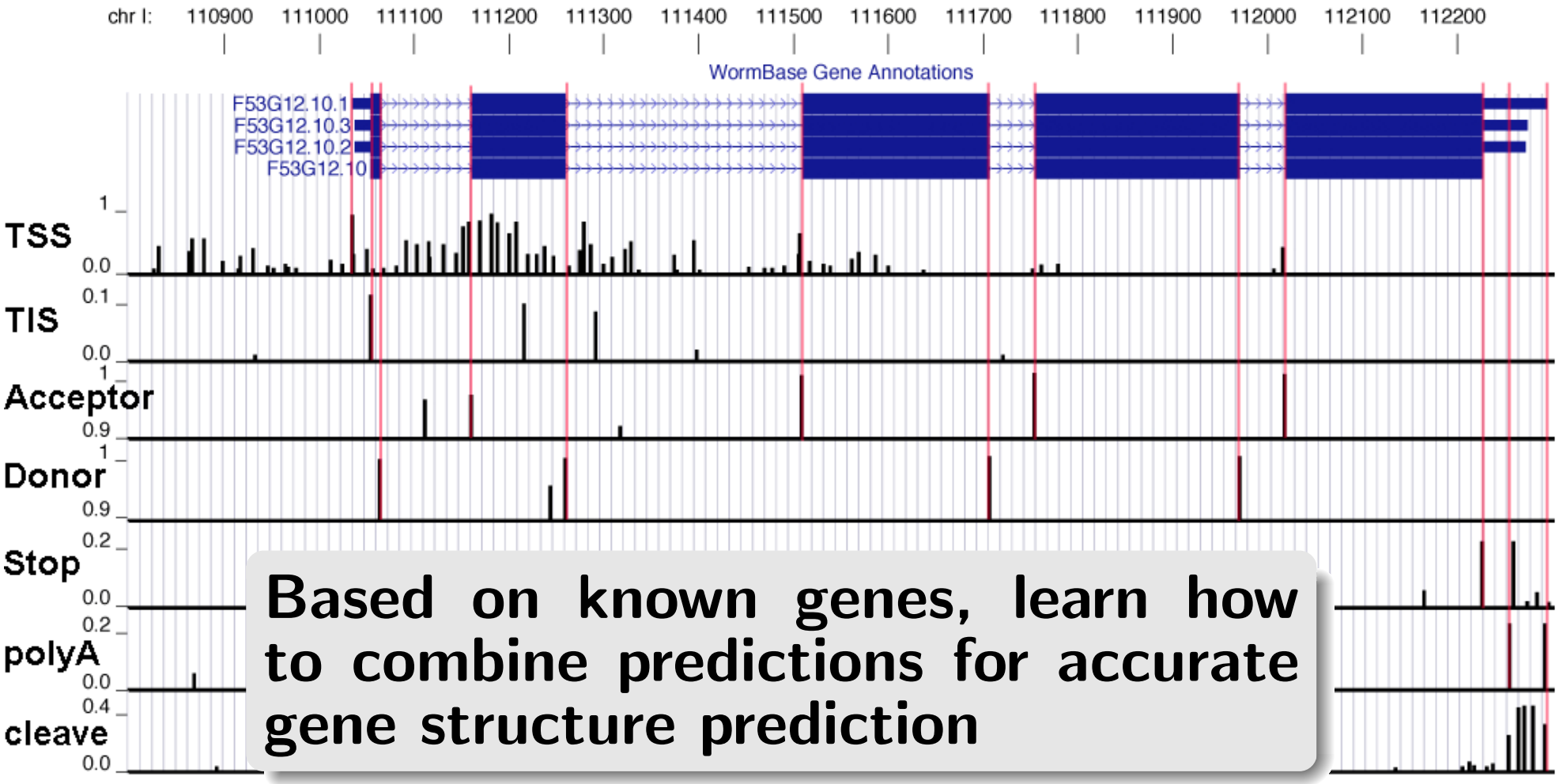

[Schweikert et al., 2009b]

# Example: Predictions in UCSC Browser




[Schweikert et al., 2009b]

# Example: Predictions in UCSC Browser




**Based on known genes, learn how to combine predictions for accurate gene structure prediction**

[Schweikert et al., 2009b]

# Results Using mGene

## (Schweikert et al., Genome Research, 2009)




- One of most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

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

  - New gene predictions will be checked by Wormbase curators for integration in annotation
- Web service for gene prediction [Schweikert et al., 2009a]
  - Gene prediction
  - Prediction of functional elements
  - Adaptable to new genomes

  `http://mgene.org/web`

# Results Using mGene

## (Schweikert et al., Genome Research, 2009)




- One of most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

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

  - New gene predictions will be checked by Wormbase curators for integration in annotation

- Web service for gene prediction [Schweikert et al., 2009a]
  - Gene prediction
  - Prediction of functional elements
  - Adaptable to new genomes

  http://mgene.org/web

# Results Using mGene

## (Schweikert et al., Genome Research, 2009)




- One of most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

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

  - New gene predictions will be checked by Wormbase curators for integration in annotation

- Web service for gene prediction [Schweikert et al., 2009a]
  - Gene prediction
  - Prediction of functional elements
  - Adaptable to new genomes

  http://mgene.org/web

# Results Using mGene

## (Schweikert et al., Genome Research, 2009)

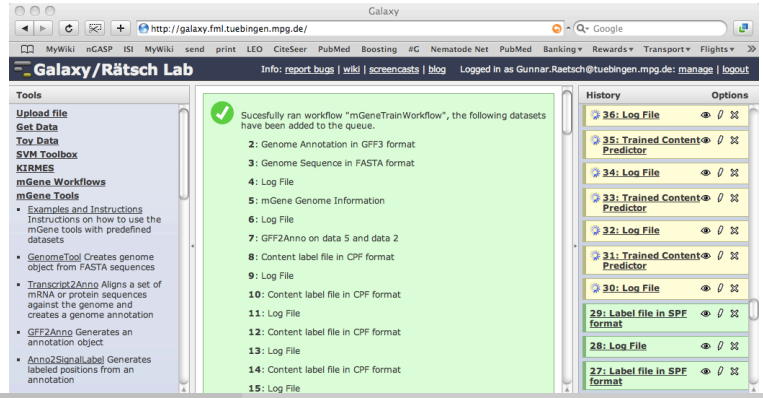
- One of most accurate *ab initio* method in the nGASP genome annotation challenge for *C. elegans* [Coghlan et al., 2008]
- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009b]

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

  - New gene predictions will be checked by Wormbase curators for integration in annotation

- Web service for gene prediction
  - Gene prediction
  - Prediction of functional elements
  - Adaptable to new genomes

    http://mgene.org/web



# Limitations/Extensions




Gene finding accuracy still far from perfect

- Misses genes, predicts incorrect gene models
- Does not predict alternative transcripts
- Cannot predict when transcripts are expressed/modified/degraded...
- Does not take epigenetic and other factors into account

Need experimental data for condition specific transcriptomes.

Then we can learn to predict (at some point).

# Limitations/Extensions




Gene finding accuracy still far from perfect

- Misses genes, predicts incorrect gene models
- Does not predict alternative transcripts
- Cannot predict when transcripts are expressed/modified/degraded...
- Does not take epigenetic and other factors into account

**Need experimental data for condition specific transcriptomes.**

Then we can learn to predict (at some point).

# Limitations/Extensions




Gene finding accuracy still far from perfect

- Misses genes, predicts incorrect gene models
- Does not predict alternative transcripts
- Cannot predict when transcripts are expressed/modified/degraded...
- Does not take epigenetic and other factors into account

**Need experimental data for condition specific transcriptomes.**

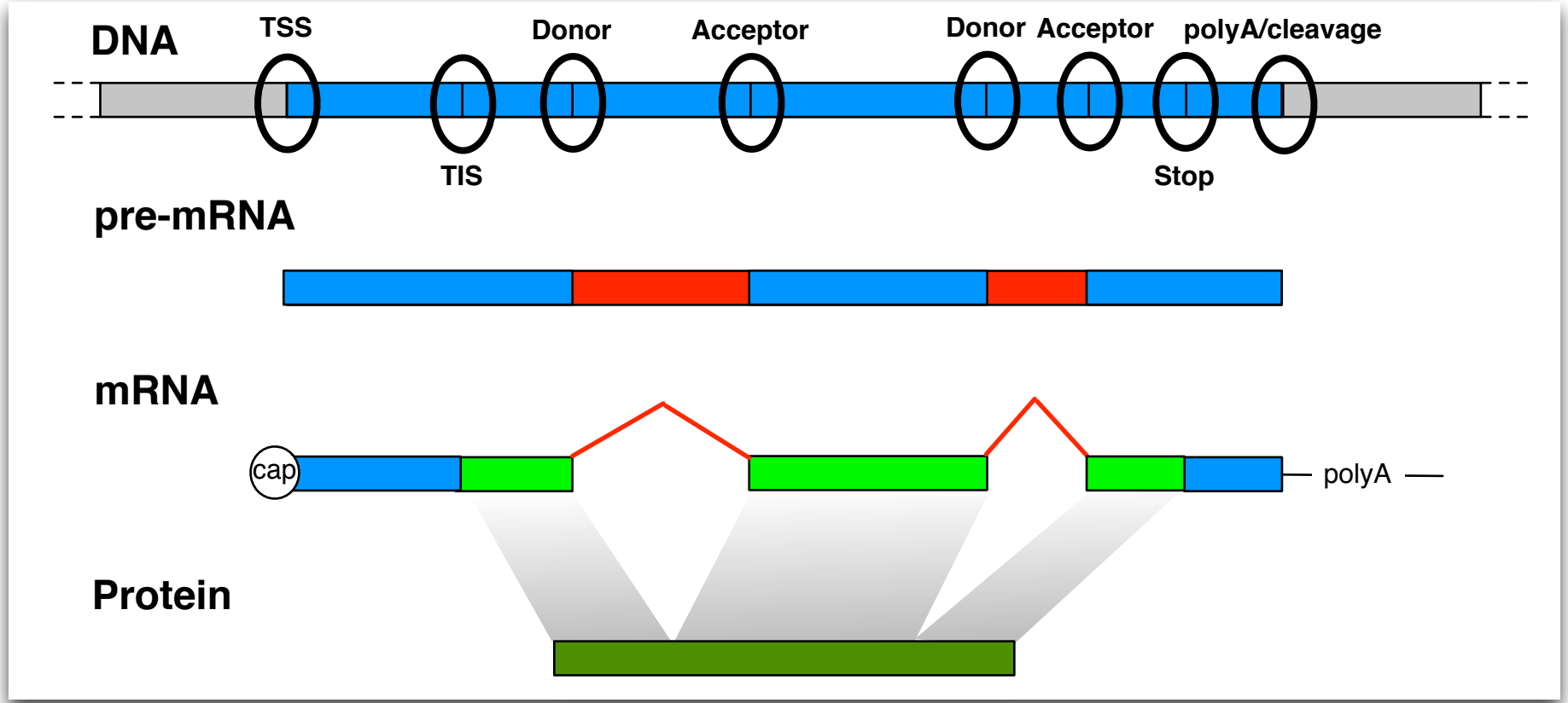
**Then we can learn to predict (at some point).**

# Roadmap




1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Detection of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology and Limitations
   - Assembly & Read Mapping
   - Transcript Quantitation
   - De novo Transcript Prediction

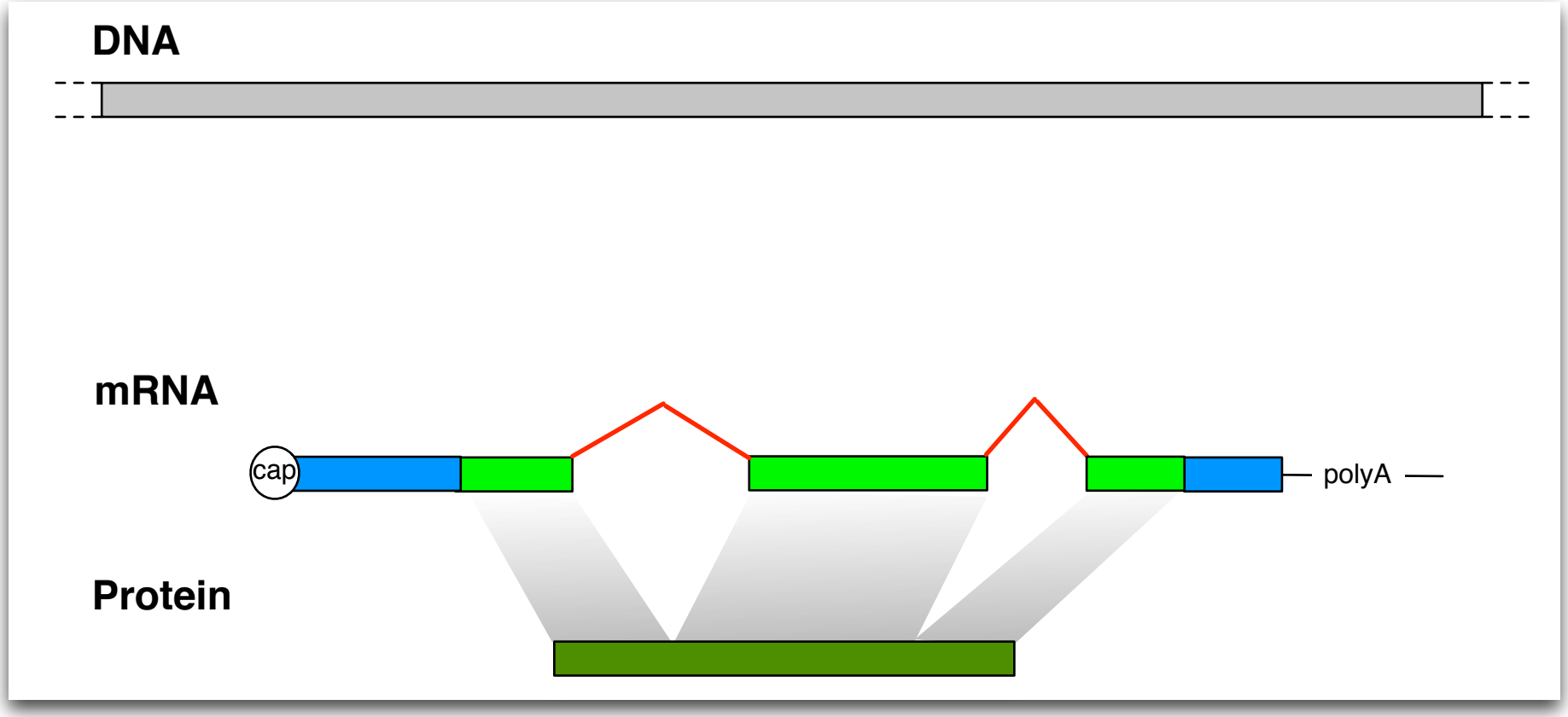
# From Genome to Proteins etc.




Directly measure the transcriptome

- Whole genome tiling arrays
- Transcriptome sequencing (Sanger or Next Generation Sequencing)

# From Transcriptome Measurements to Proteins etc.




Directly measure the transcriptome

- Whole genome tiling arrays
- Transcriptome sequencing (Sanger or Next Generation Sequencing)

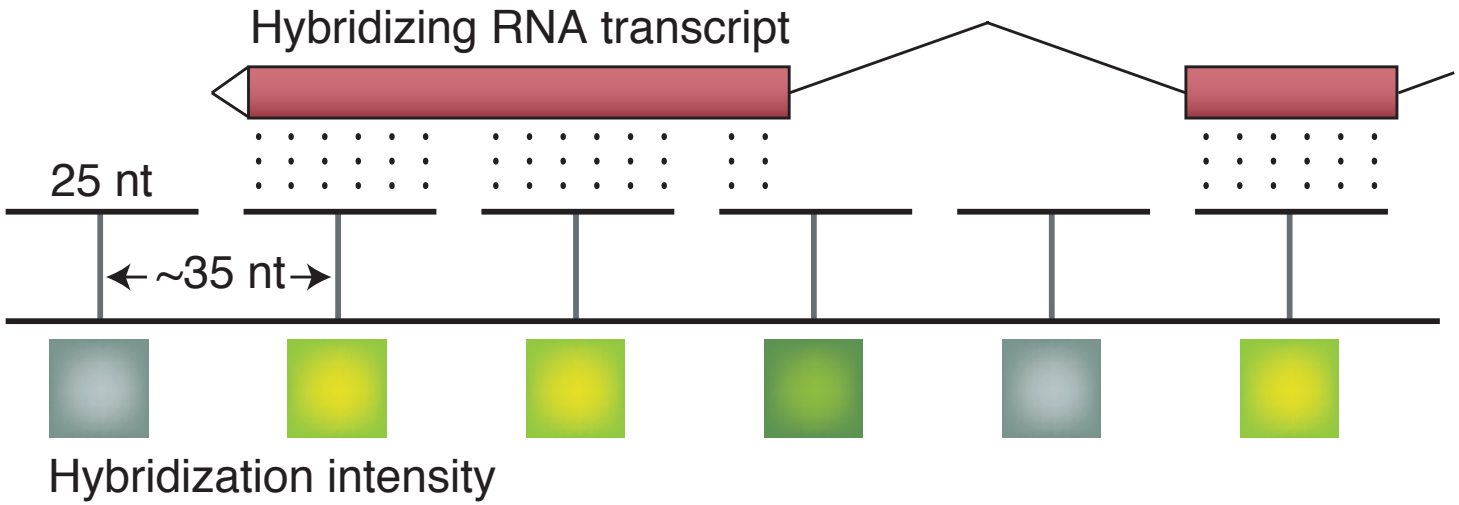
# Whole-genome Tiling Arrays




- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays

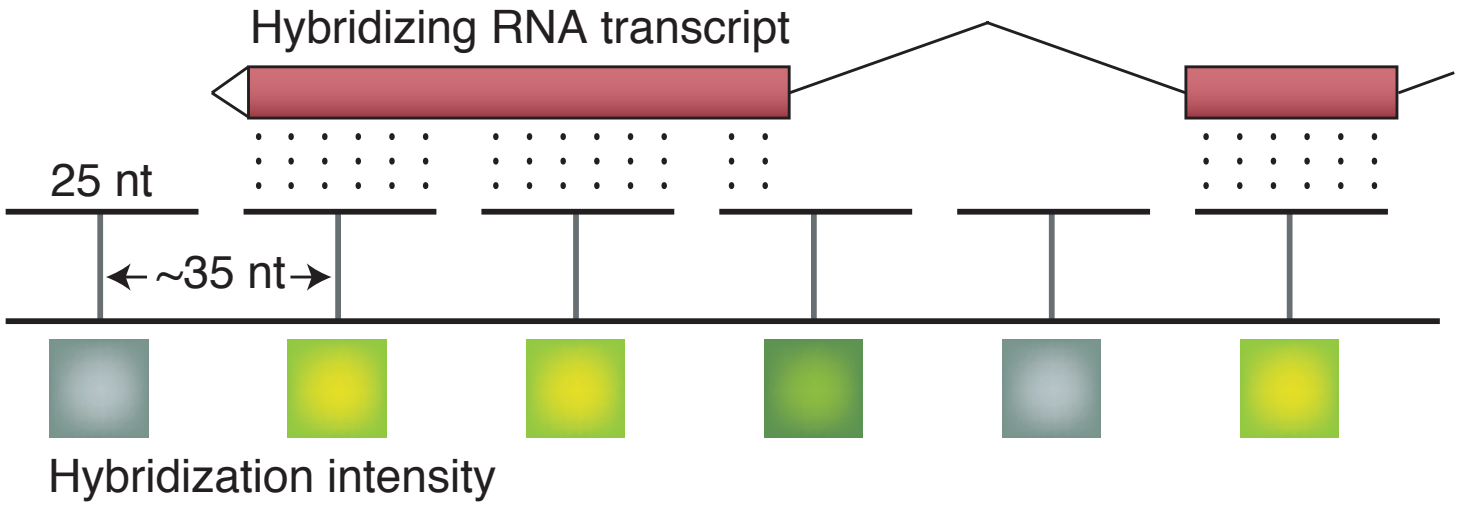



- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays,

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays

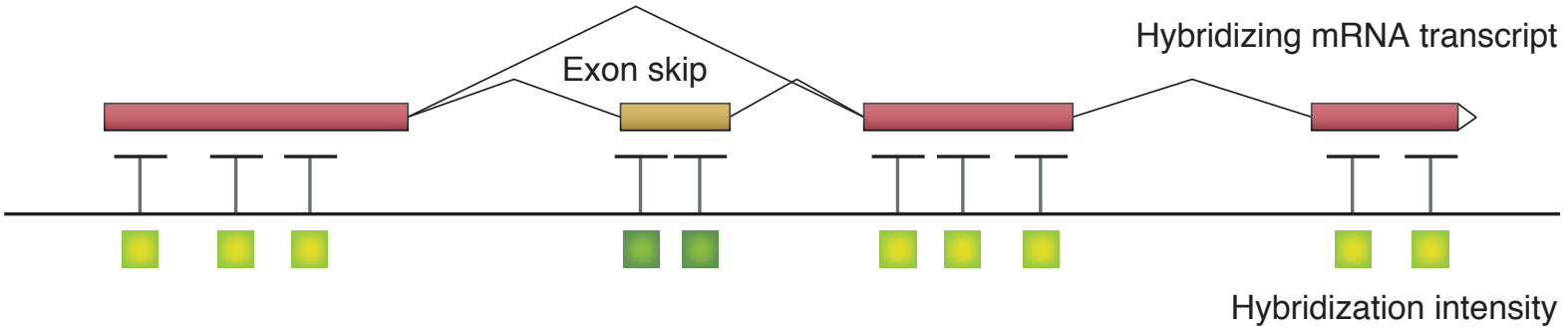



- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays, exon junction arrays

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays

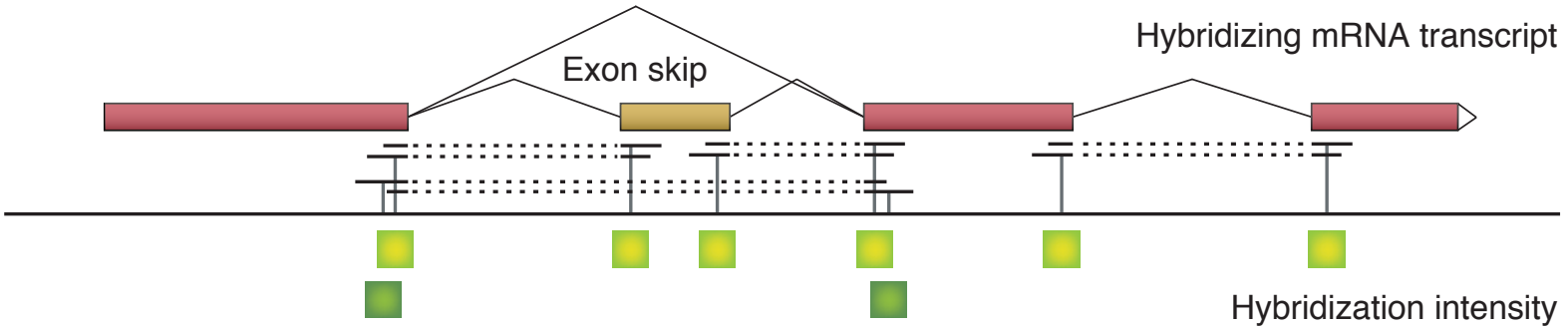



- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays, exon junction arrays

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Whole-genome Tiling Arrays



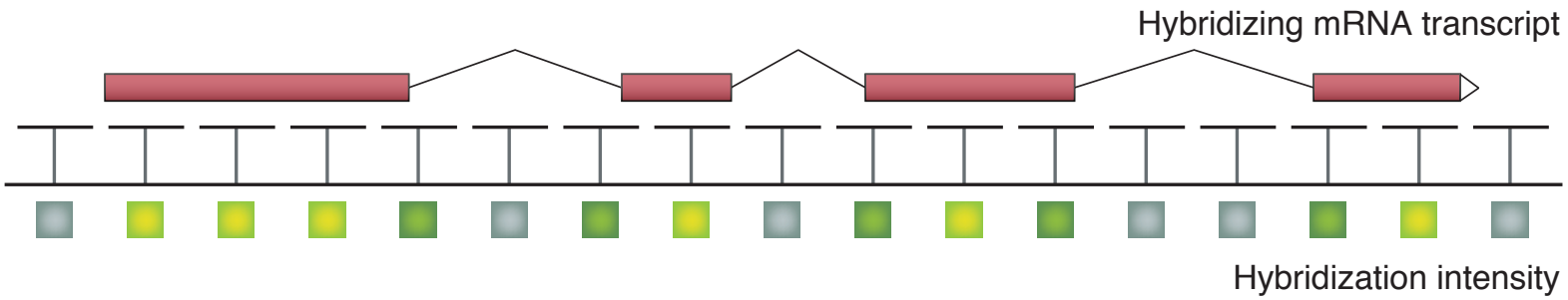

- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays, exon junction arrays

---

see Mockler et al. [2005], Yazaki et al. [2007] for comprehensive reviews

# Tiling Array Analysis Challenges (I)

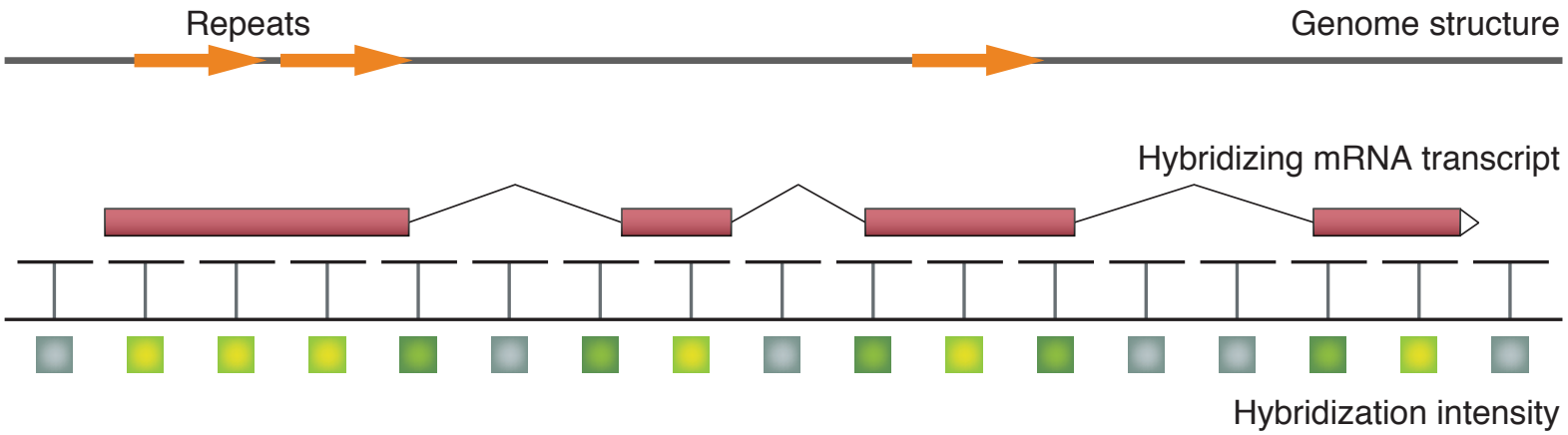
Repeats cause cross-hybridization




⇒ Discard tiling probes with high sequence similarity to >1 location in the genome

# Tiling Array Analysis Challenges (I)

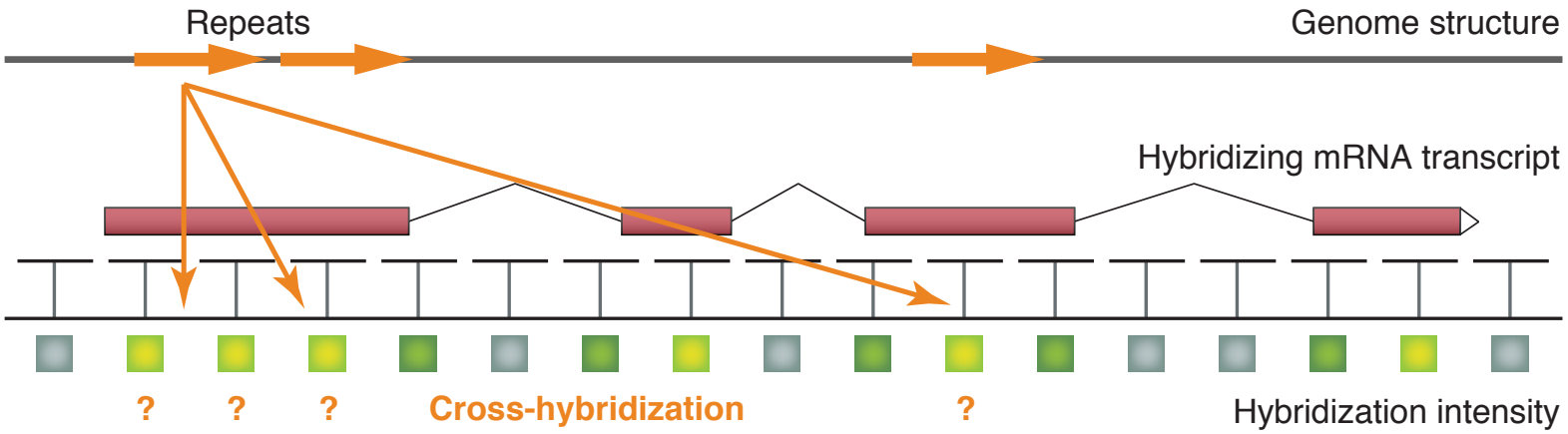
Repeats cause cross-hybridization




⇒ Discard tiling probes with high sequence similarity to >1 location in the genome

# Tiling Array Analysis Challenges (I)

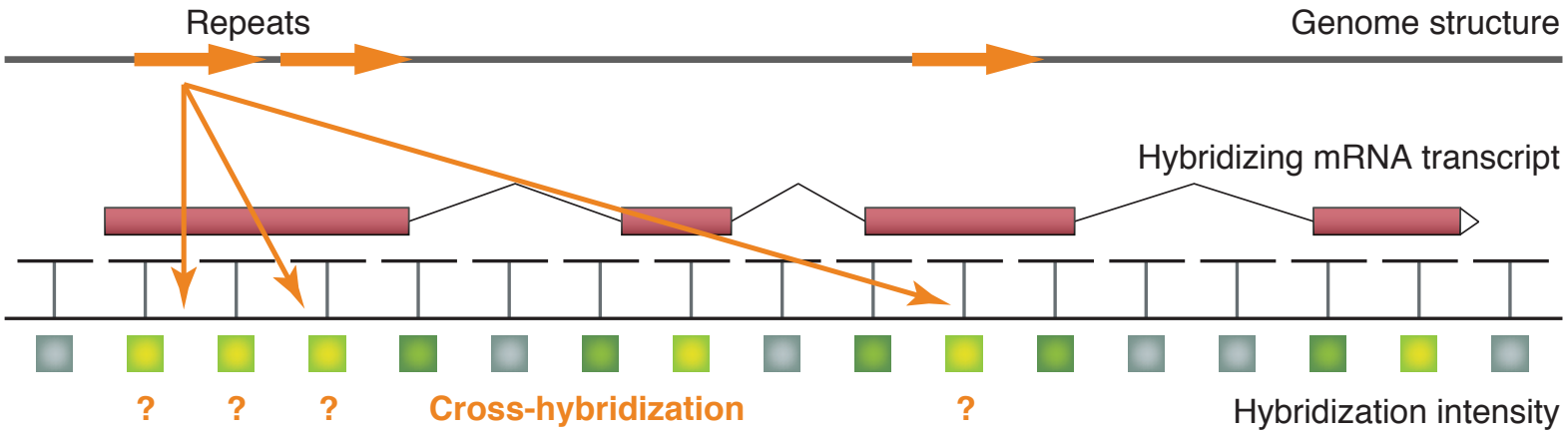
Repeats cause cross-hybridization




⇒ Discard tiling probes with high sequence similarity to >1 location in the genome

# Tiling Array Analysis Challenges (I)

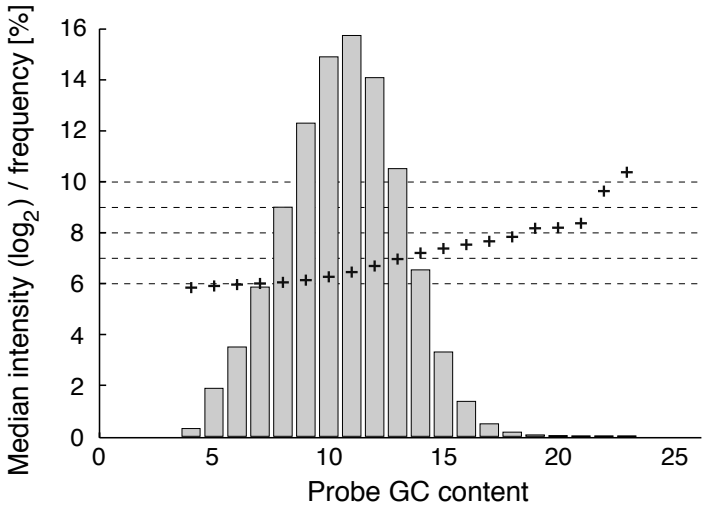
Repeats cause cross-hybridization




⇒ Discard tiling probes with high sequence similarity to >1 location in the genome

# Tiling Array Analysis Challenges (II)

Hybridization intensity exhibits a probe sequence bias




Sequence-normalization approaches:

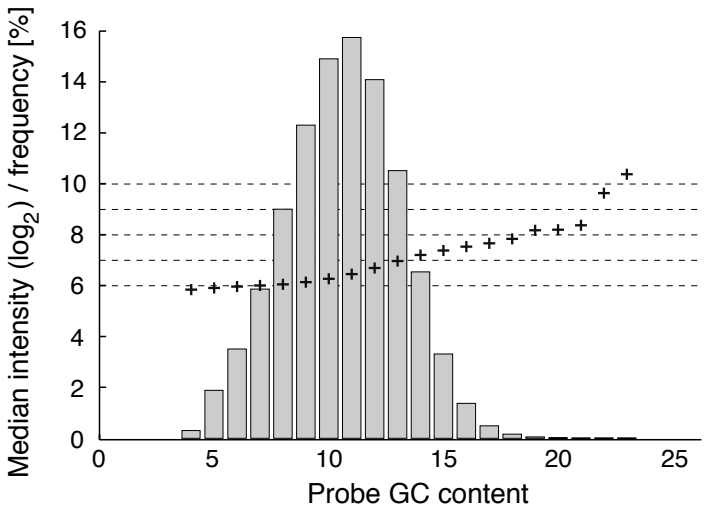
- Rescaling by probe GC content [Samanta et al., 2006]
- Rescaling using genomic DNA hybridization [David et al., 2006].

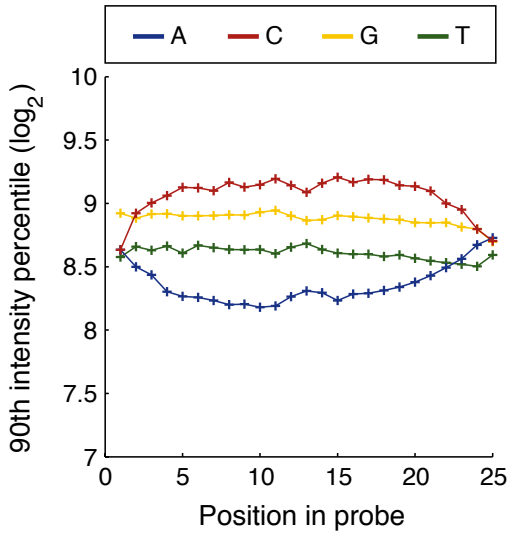
$$n_{ij} = \frac{x_{ij} - b_j(y_i)}{y_i}$$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$ to obtain normalized signal $n_{ij}$; DNA hybridization signal $y_i$ is transformed into RNA background signal by $b_j$ estimated from

# Tiling Array Analysis Challenges (II)

Hybridization intensity exhibits a probe sequence bias




Sequence-normalization approaches:

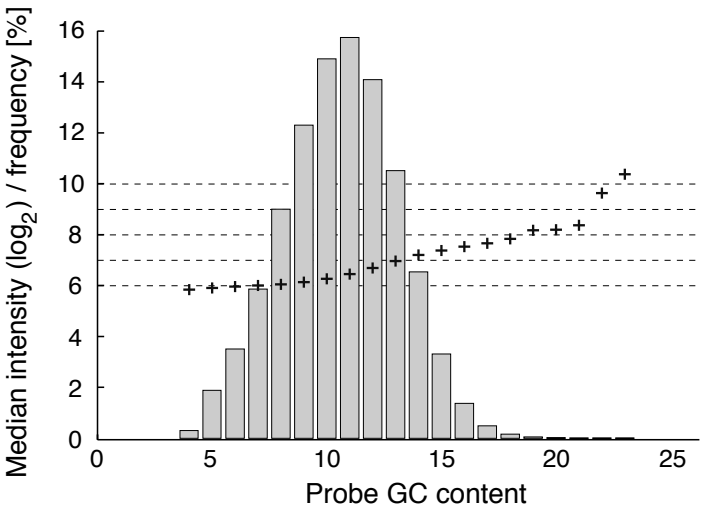
- Rescaling by probe GC content [Samanta et al., 2006]
- Rescaling using genomic DNA hybridization [David et al., 2006].

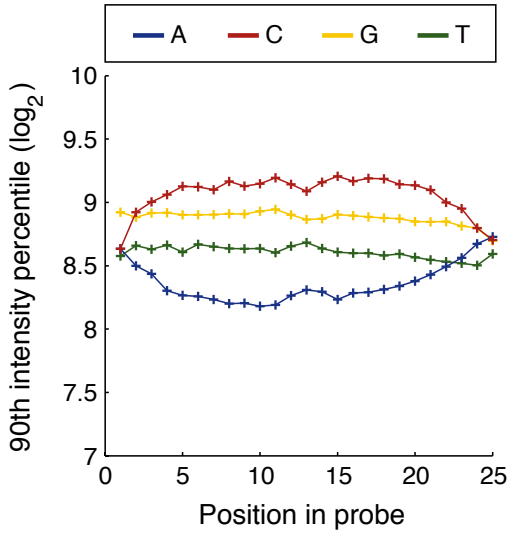
$$n_{ij} = \frac{x_{ij} - b_j(y_i)}{y_i}$$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$ to obtain normalized signal $n_{ij}$; DNA hybridization signal $y_i$ is transformed into RNA background signal by $b_j$ estimated from

# Tiling Array Analysis Challenges (II)

Hybridization intensity exhibits a probe sequence bias




Sequence-normalization approaches:

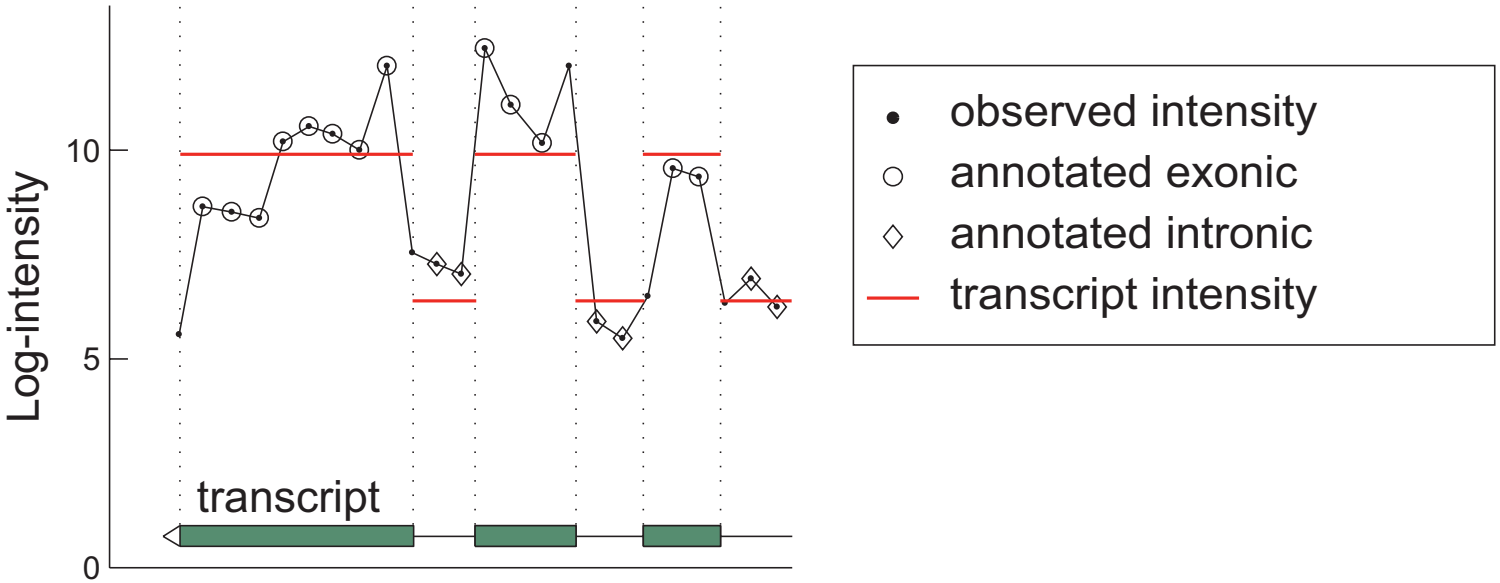
- Rescaling by probe GC content [Samanta et al., 2006]
- Rescaling using genomic DNA hybridization [David et al., 2006].

$$n_{ij} = \frac{x_{ij} - b_j(y_i)}{y_i}$$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$ to obtain normalized signal $n_{ij}$; DNA hybridization signal $y_i$ is transformed into RNA background signal by $b_j$ estimated from

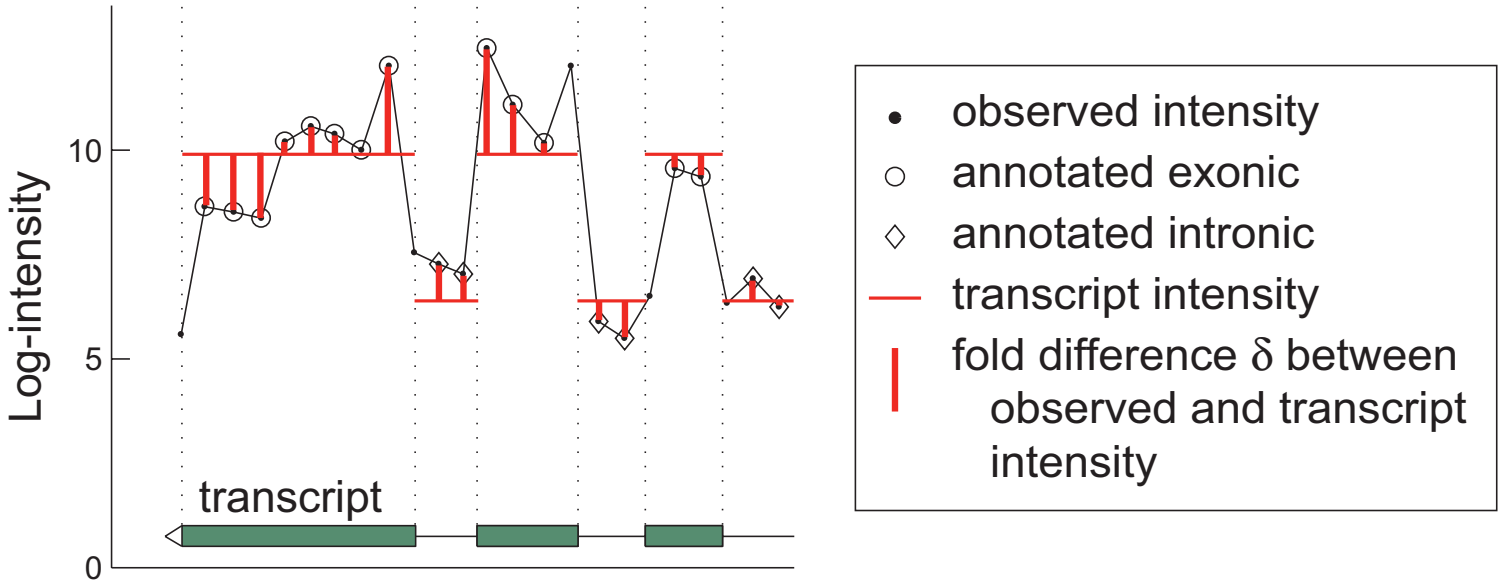
# Tiling Array Analysis Challenges (III)

- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008c]
- Learns intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Takes probe sequence $\mathbf{x}_i$ as input for regression.
  Models probe sequence effect depending on $y_i$: $f(\mathbf{x}_i, y_i) \approx \delta_i$



# Tiling Array Analysis Challenges (III)

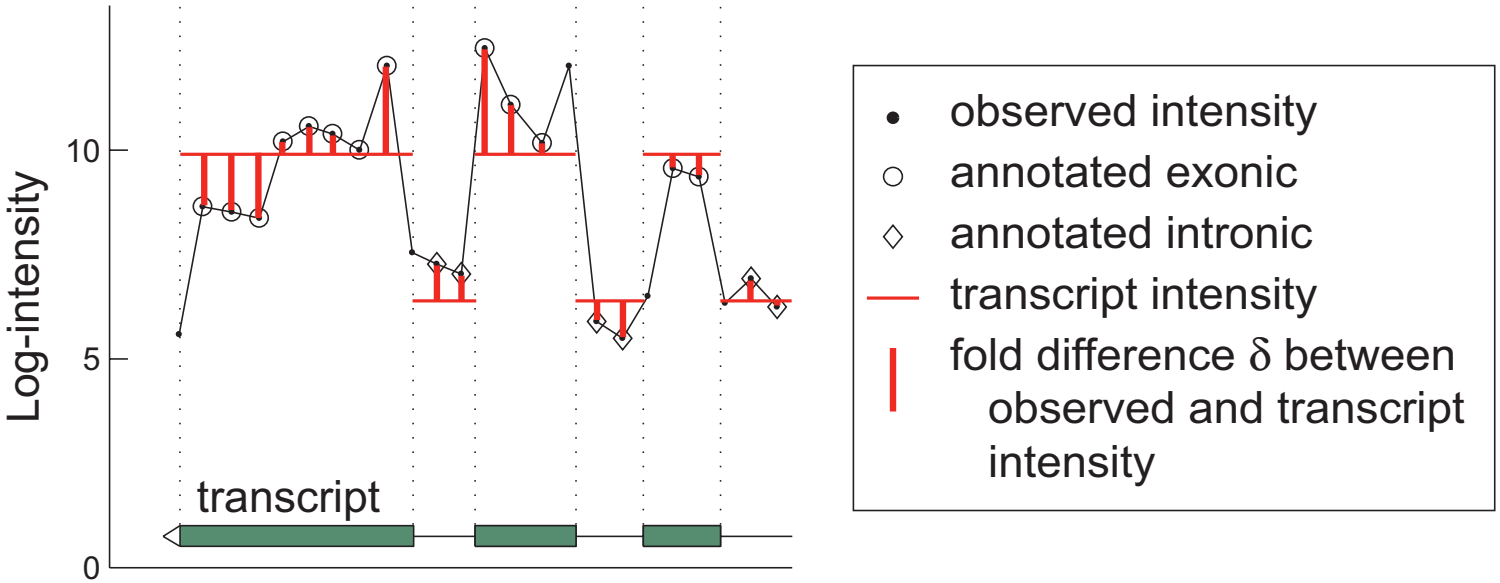
- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008c]
- Learns intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Takes probe sequence $\mathbf{x}_i$ as input for regression.
  Models probe sequence effect depending on $y_i$: $f(\mathbf{x}_i, y_i) \approx \delta_i$



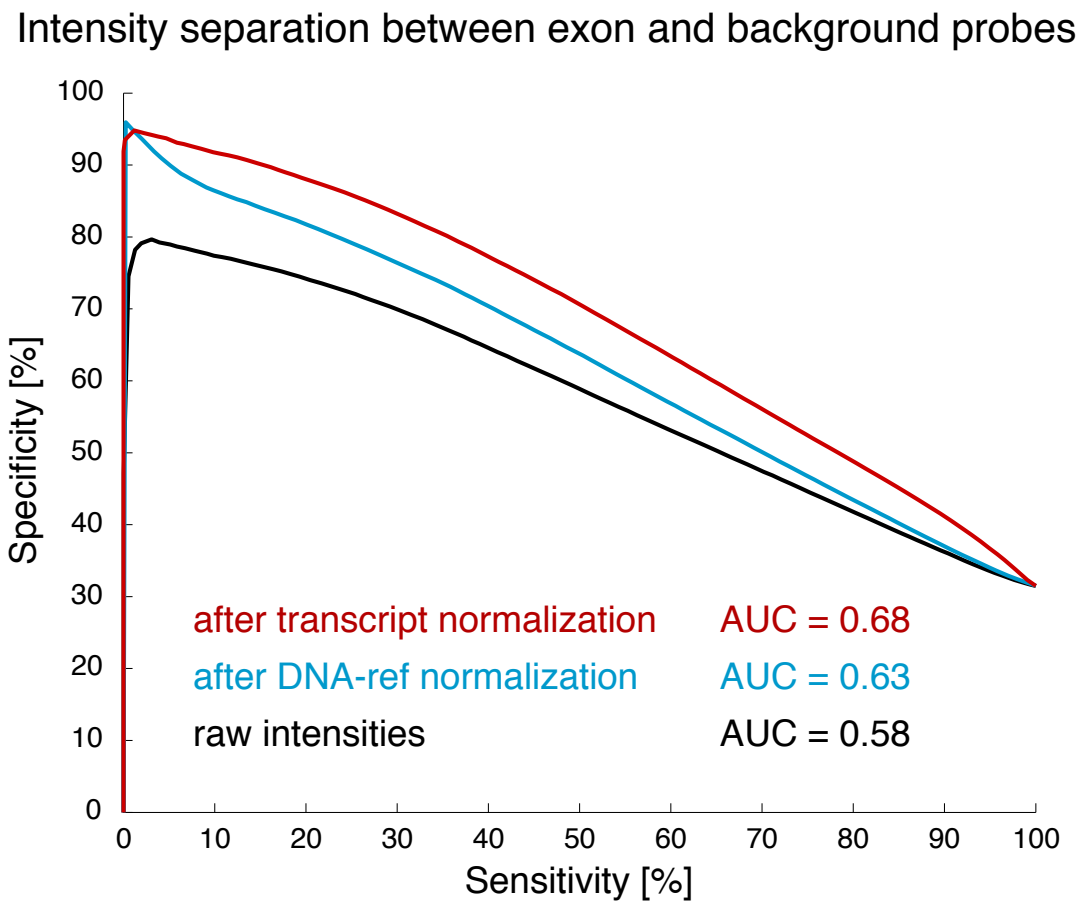
# Tiling Array Analysis Challenges (III)




- Transcript normalization assumes constant transcript intensities $\overline{y}_i$ (median estimate) [Zeller et al., 2008c]
- Learns intensity deviation from transcript intensity $\delta_i := y_i - \overline{y}_i$
- Takes probe sequence $\mathbf{x}_i$ as input for regression.
  Models probe sequence effect depending on $y_i$: $f(\mathbf{x}_i, y_i) \approx \delta_i$



# How good does it work?




Transcript normalization (TN) improves intensity separation

$\longrightarrow$ enhances exon and transcript identification [Zeller et al., 2008b]

$\longrightarrow$ more than DNA reference based normalization

# Detection of Expression Differences




1. Map tiling probes to annotated transcripts (define probe sets)
2. Use standard microarray tools to analyze gene expression

Gene expression values are typically computed using robust "summarization methods" that account for probe noise

[e.g. Irizarry et al., 2003b]

Significant expression changes are typically identified with a statistical test. Results have to be corrected for multiple testing

[e.g. Storey and Tibshirani, 2003]

Advantages of tiling arrays:

- Annotations change, only remapping is needed to obtain expression measurements for the latest annotation.
- Expression can be measured per exon, not only per gene.
- Expression can be measured for introns ($\Rightarrow$ detect retention).

# Detection of Expression Differences




1. Map tiling probes to annotated transcripts (define probe sets)
2. Use standard microarray tools to analyze gene expression

Gene expression values are typically computed using robust "summarization methods" that account for probe noise

[e.g. Irizarry et al., 2003b]

Significant expression changes are typically identified with a statistical test. Results have to be corrected for multiple testing

[e.g. Storey and Tibshirani, 2003]

Advantages of tiling arrays:

- Annotations change, only remapping is needed to obtain expression measurements for the latest annotation.
- Expression can be measured per exon, not only per gene.
- Expression can be measured for introns ($\Rightarrow$ detect retention).

# Example Study

(*C. elegans* gene expression with tiling arrays)




## Methods

- Constitutive exon probe sets for each gene (WS200)
- Quantile Normalization [Bolstad et al., 2003]
- FARMS [Hochreiter et al., 2006] and RMA [Irizarry et al., 2003a] for expression summarization
- RankProduct [Breitling et al., 2004] and LIMMA [Smyth, 2005] ($\approx$ SAM) with empirical Bayes-moderated $t$-statistic [Benjamini and Hochberg, 1995] for significance scoring
- $p$-value cutoffs at 10%

## Results

| neural tissue vs. reference | | |
|---|---|---|
| | FARMS | RMA |
| RankProduct | ↑ 264 ↓ 516 | ↑ 68 ↓ 267 |
| LIMMA | ↑ 411 ↓ 888 | ↑ 654 ↓1196 |

(Joint work with Clay Spencer, David Miller, Stefan R. Henz, Georg Zeller)

# Example Study

(*C. elegans* gene expression with tiling arrays)




## Methods

- Constitutive exon probe sets for each gene (WS200)
- Quantile Normalization [Bolstad et al., 2003]
- FARMS [Hochreiter et al., 2006] and RMA [Irizarry et al., 2003a] for expression summarization
- RankProduct [Breitling et al., 2004] and LIMMA [Smyth, 2005] ($\approx$ SAM) with empirical Bayes-moderated $t$-statistic [Benjamini and Hochberg, 1995] for significance scoring
- $p$-value cutoffs at 10%

## Results

| neural tissue vs. reference | | |
|---|---|---|
| | FARMS | RMA |
| RankProduct | ↑ 264 ↓ 516 | ↑ 68 ↓ 267 |
| LIMMA | ↑ 411 ↓ 888 | ↑ 654 ↓1196 |

(Joint work with Clay Spencer, David Miller, Stefan R. Henz, Georg Zeller)

# Evaluation of Scoring Methods




## RMA-LIMMA detects largest fraction of known neural genes

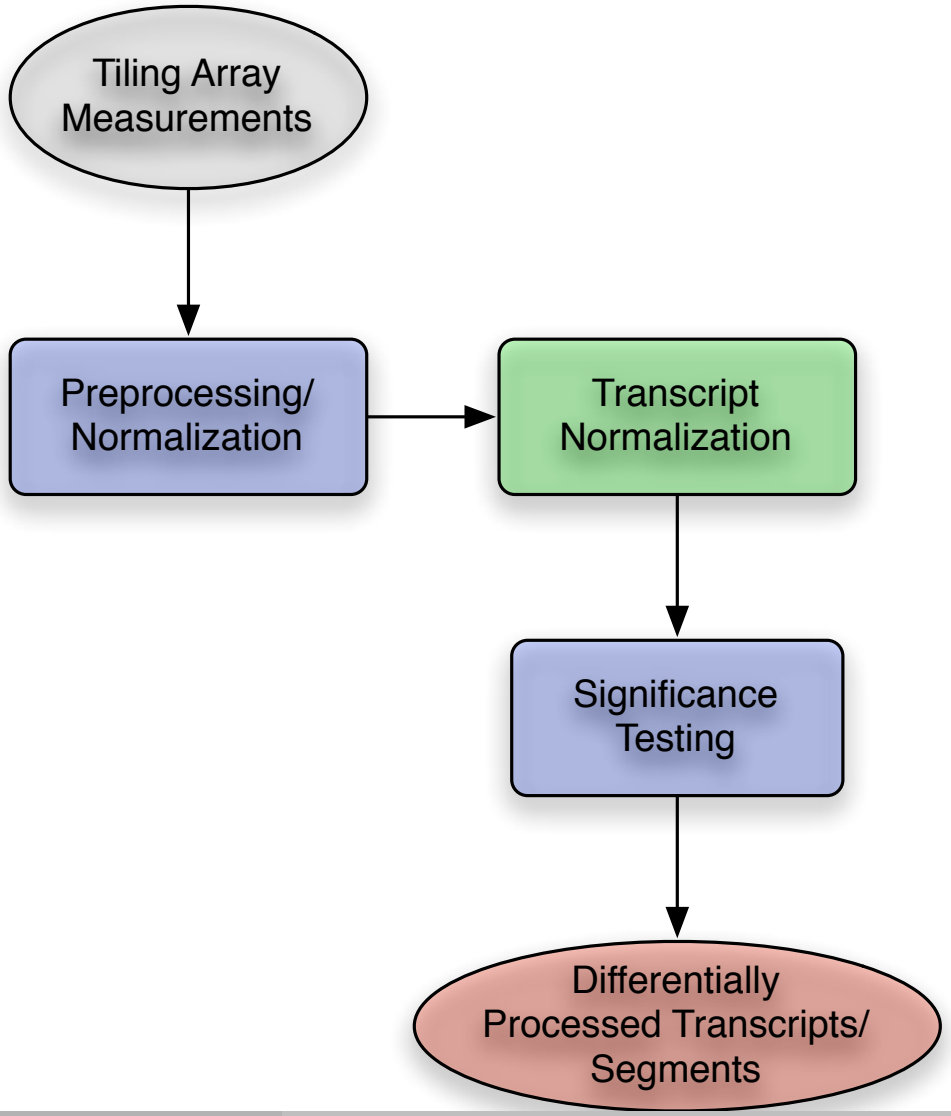
| Method | enriched genes* | enriched neural genes | known neural genes | TPR | muscle genes[1] | FPR | intest. genes[2] | FPR |
|---|---|---|---|---|---|---|---|---|
| RMA-LIMMA | 781 | 12 | 23 | 52% | 19 | 2.4% | 52 | 6.7% |
| RMA-RankProduct | 377 | 8 | 23 | 35% | 8 | 2.1% | 20 | 5.3% |
| FARMS-LIMMA | 65 | 4 | 23 | 17% | 3 | 4.6% | 2 | 3.1% |
| FARMS-RankProduct | 471 | 9 | 23 | 39% | 19 | 4.0% | 46 | 9.8% |

TPR (True Positive Rate) = enriched neural genes / known neural genes
FPR (False Positive Rate) = enriched muscle or intestine genes / enriched genes

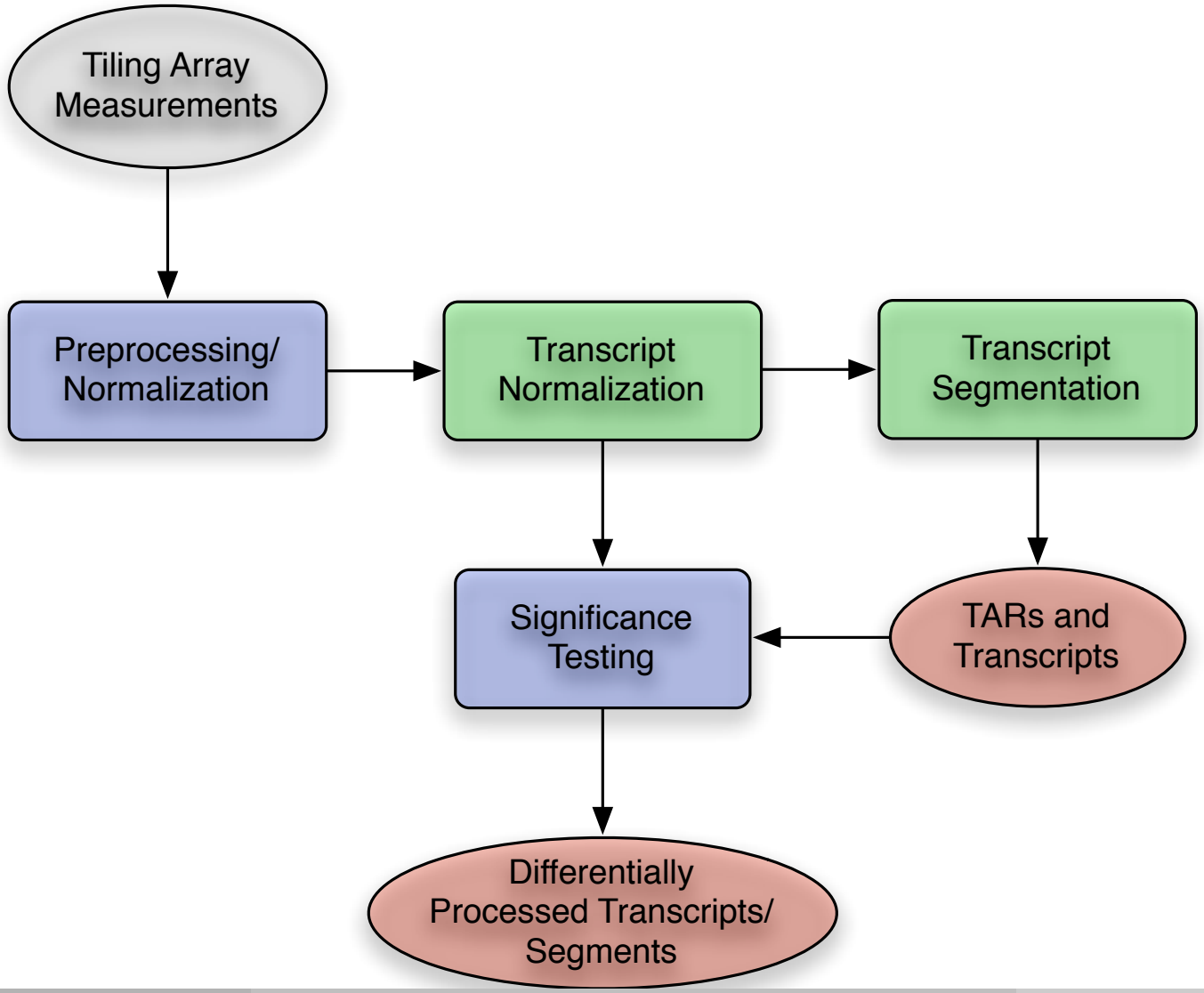
* neural vs. reference
[1] muscle: 937 enriched genes [Fox et al., 2007]
[2] intestine: 1,938 enriched genes [Pauli et al., 2006]

(Analysis done by Clay Spencer and David Miller)

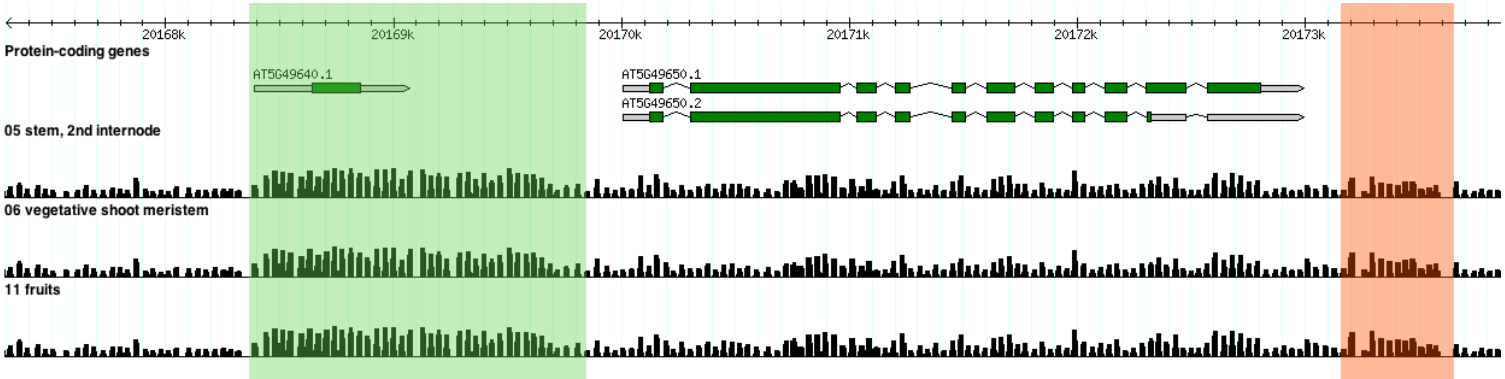
# Analysis Pipeline Summary



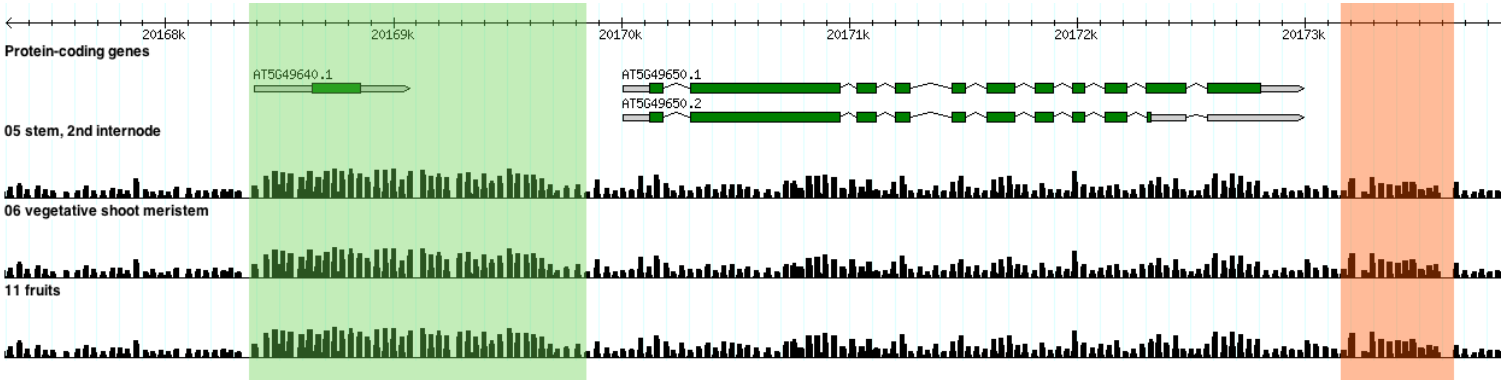
# Analysis Pipeline Summary



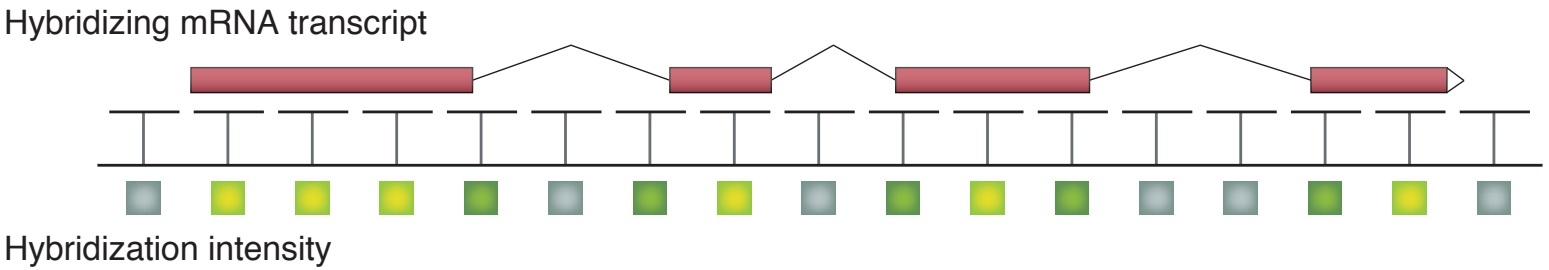
# Discovery of Expressed Transcripts



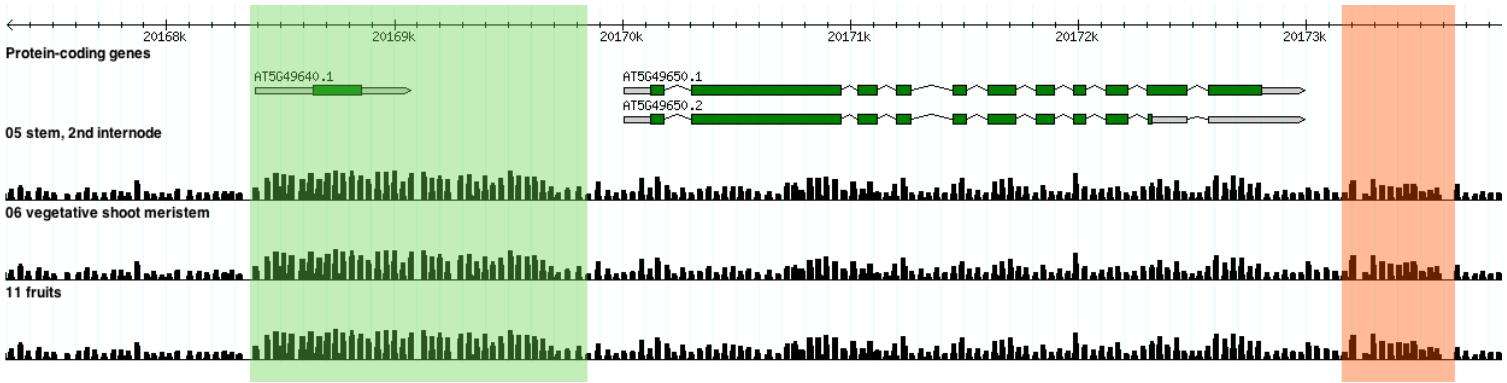
# Discovery of Expressed Transcripts



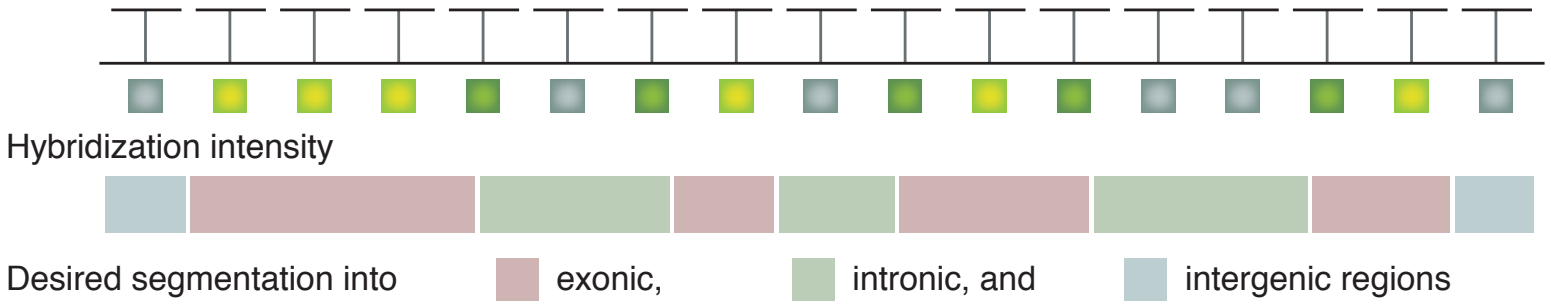

De novo segmentation is needed to re-annotate expressed genes.



# Discovery of Expressed Transcripts




De novo segmentation is needed to re-annotate expressed genes.



# Transfrag Method / Affymetrix TARs




1. Identify "positive probes" in local neighborhood. Smooth data locally, across replicates (Pseudomedian[1]) [Royce et al., 2007a]
   Two approaches:
   - define an *ad hoc* threshold on smoothed signal intensity (e.g. 90th signal percentile) [Kampa et al., 2004]
   - estimate a threshold from negative bacterial control probes to adjust an empirical false discovery rate [He et al., 2007]
2. Combine positive probes into "transfrags" in case of a run of consecutive positive probes (minRun) interrupted by a limited number of negative probes (maxGap) [Bertone et al., 2004, Kampa et al., 2004]

**Problem:** Manual parameter "tuning"

[1] median of all pairwise averages of probe signals within a sliding window

# Transfrag Method / Affymetrix TARs




1. Identify "positive probes" in local neighborhood. Smooth data locally, across replicates (Pseudomedian[1]) [Royce et al., 2007a]
   Two approaches:
   - define an *ad hoc* threshold on smoothed signal intensity (e.g. 90th signal percentile) [Kampa et al., 2004]
   - estimate a threshold from negative bacterial control probes to adjust an empirical false discovery rate [He et al., 2007]
2. Combine positive probes into "transfrags" in case of a run of consecutive positive probes (minRun) interrupted by a limited number of negative probes (maxGap) [Bertone et al., 2004, Kampa et al., 2004]

**Problem:** Manual parameter "tuning"

[1] median of all pairwise averages of probe signals within a sliding window

# Transfrag Method / Affymetrix TARs




1. Identify "positive probes" in local neighborhood. Smooth data locally, across replicates (Pseudomedian[1]) [Royce et al., 2007a]
   Two approaches:
   - define an *ad hoc* threshold on smoothed signal intensity (e.g. 90th signal percentile) [Kampa et al., 2004]
   - estimate a threshold from negative bacterial control probes to adjust an empirical false discovery rate [He et al., 2007]
2. Combine positive probes into "transfrags" in case of a run of consecutive positive probes (minRun) interrupted by a limited number of negative probes (maxGap) [Bertone et al., 2004, Kampa et al., 2004]

**Problem:** Manual parameter "tuning"

[1] median of all pairwise averages of probe signals within a sliding window

# Dynamic Programming Segmentation




Model intensities as piecewise constant function: [Huber et al., 2006b]

$x_{ij} = \mu_s + \epsilon_{ij}$ for $t_s \leq i < t_{s+1}$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$; segment boundaries $t_s$ and $t_{s+1}$. $\mu_s$ is the mean signal of the $s$th segment, $\epsilon_{ij}$ error terms.

Minimize the sum of squared residuals:

$$G(t_1, \ldots, t_S) = \sum_{s=1}^{S} \sum_{j=1}^{R} \sum_{i=t_s}^{t_{s+1}-1} (x_{ij} - \mu_s)^2$$

where $S$ is the number of segments
and $R$ the number of replicates.

The optimal segmentation can be computed in $O(n^2)$ time using dynamic programming. [Huber et al., 2006b]

**Problem:** $S$ is to be user-specified

# Dynamic Programming Segmentation

Model intensities as piecewise constant function: [Huber et al., 2006b]

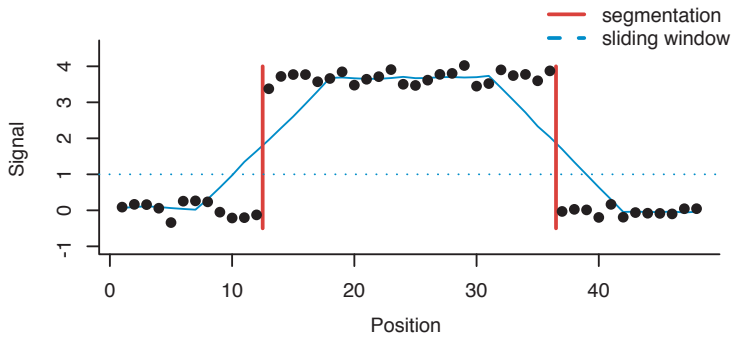
$x_{ij} = \mu_s + \epsilon_{ij}$ for $t_s \leq i < t_{s+1}$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$; segment boundaries $t_s$ and $t_{s+1}$. $\mu_s$ is the mean signal of the $s$th segment, $\epsilon_{ij}$ error terms.

Minimize the sum of squared residuals:

$$G(t_1, \ldots, t_S) = \sum_{s=1}^{S} \sum_{j=1}^{R} \sum_{i=t_s}^{t_{s+1}-1} (x_{ij} - \mu_s)^2$$

where $S$ is the number of segments and $R$ the number of replicates.




The optimal segmentation can be computed in $O(n^2)$ time using dynamic programming. [Huber et al., 2006b]

**Problem:** S is to be user-specified

# Dynamic Programming Segmentation

Model intensities as piecewise constant function: [Huber et al., 2006b]

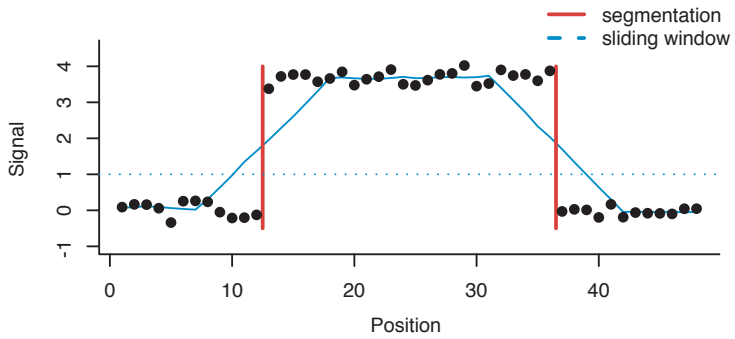
$x_{ij} = \mu_s + \epsilon_{ij}$ for $t_s \leq i < t_{s+1}$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$; segment boundaries $t_s$ and $t_{s+1}$. $\mu_s$ is the mean signal of the $s$th segment, $\epsilon_{ij}$ error terms.

Minimize the sum of squared residuals:

$$G(t_1, \ldots, t_S) = \sum_{s=1}^{S} \sum_{j=1}^{R} \sum_{i=t_s}^{t_{s+1}-1} (x_{ij} - \mu_s)^2$$

where $S$ is the number of segments and $R$ the number of replicates.




The optimal segmentation can be computed in $O(n^2)$ time using dynamic programming. [Huber et al., 2006b]

**Problem:** S is to be user-specified

# Dynamic Programming Segmentation

Model intensities as piecewise constant function: [Huber et al., 2006b]

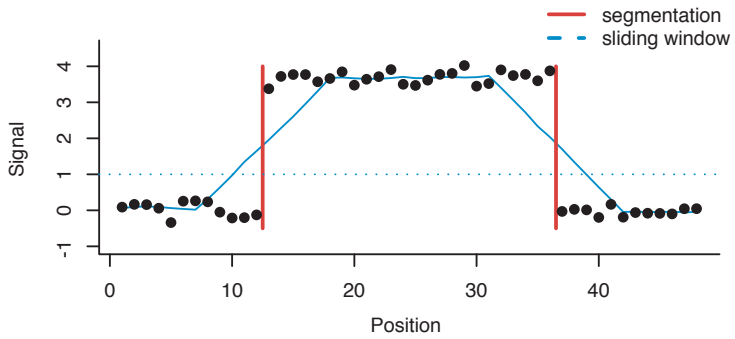
$x_{ij} = \mu_s + \epsilon_{ij}$ for $t_s \leq i < t_{s+1}$

for probe $i$ on replicate array $j$ with RNA hybridization signal $x_{ij}$; segment boundaries $t_s$ and $t_{s+1}$. $\mu_s$ is the mean signal of the $s$th segment, $\epsilon_{ij}$ error terms.

Minimize the sum of squared residuals:

$$G(t_1, \ldots, t_S) = \sum_{s=1}^{S} \sum_{j=1}^{R} \sum_{i=t_s}^{t_{s+1}-1} (x_{ij} - \mu_s)^2$$

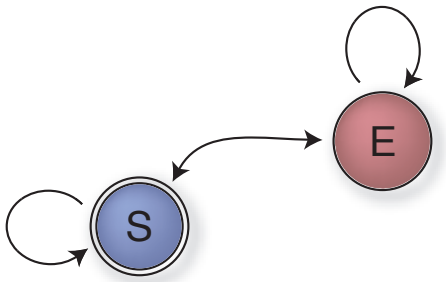
where $S$ is the number of segments and $R$ the number of replicates.




The optimal segmentation can be computed in $O(n^2)$ time using dynamic programming. [Huber et al., 2006b]

**Problem:** S is to be user-specified

# Hidden Markov Models




expressed

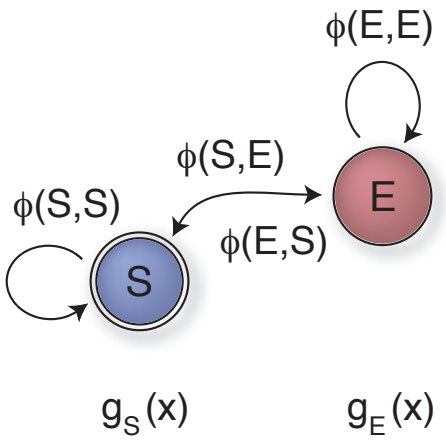
non-expressed

- Learn to label each probe given its hybridization signal and local context

  [Ji and Wong, 2005, Du et al., 2006]
- Train transition and emission probabilities on annotated genes
- Explicit intron model [Zeller et al., 2008c]
- $Q$ discrete expression levels

  [Zeller et al., 2008c]

# Hidden Markov Models




expressed

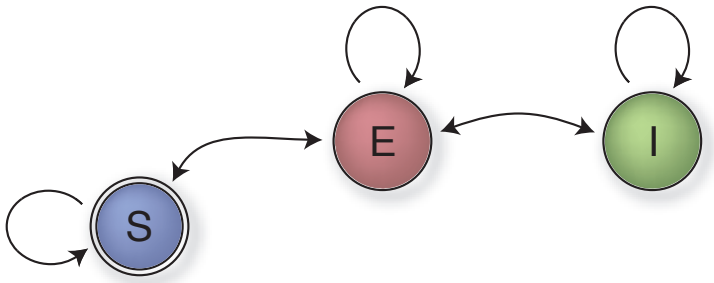
non-expressed

- Learn to label each probe given its hybridization signal and local context

  [Ji and Wong, 2005, Du et al., 2006]
- Train transition and emission probabilities on annotated genes
- Explicit intron model [Zeller et al., 2008c]
- $Q$ discrete expression levels

  [Zeller et al., 2008c]

# Hidden Markov Models



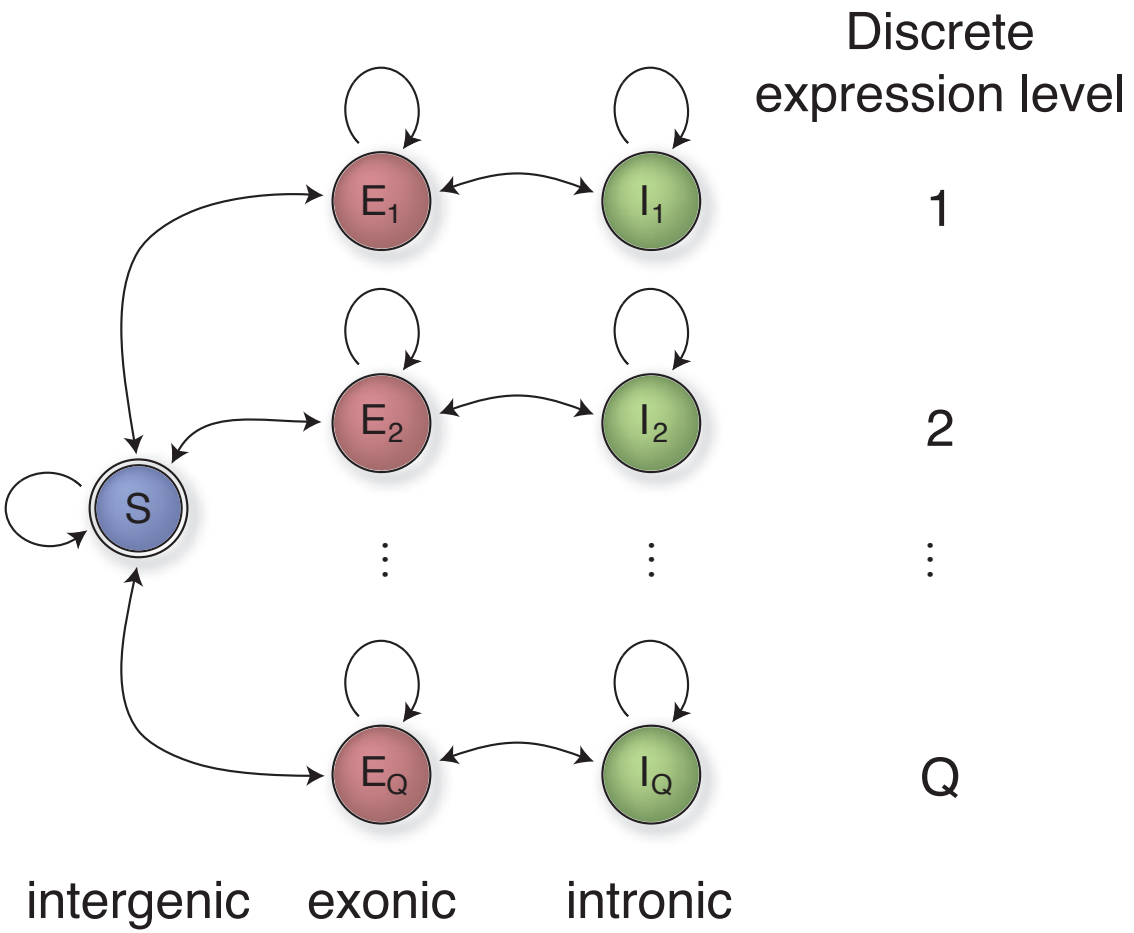

intergenic exonic intronic

- Learn to label each probe given its hybridization signal and local context

  [Ji and Wong, 2005, Du et al., 2006]
- Train transition and emission probabilities on annotated genes
- Explicit intron model [Zeller et al., 2008c]
- $Q$ discrete expression levels

  [Zeller et al., 2008c]

# Hidden Markov Models




- Learn to label each probe given its hybridization signal and local context

  [Ji and Wong, 2005, Du et al., 2006]
- Train transition and emission probabilities on annotated genes
- Explicit intron model [Zeller et al., 2008c]
- $Q$ discrete expression levels

  [Zeller et al., 2008c]

# Parametrization and Decoding




Log transition probabilities $\phi(k, l)$ between states $k$ and $l$

Log emission probabilities $g_k(x)$ in state $k$ for (discretized) hybridization signal $x$

Parametrization $\boldsymbol{\theta}$

Scoring a sequence of hybridization signals $\mathbf{x}$ with a given labeling $\pi$ and parametrization $\theta$:

$$F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi}) = \sum_{p=1}^{|\pi|} \left( g_{\pi_p}(x_p) + \phi(\pi_{p-1}, \pi_p) \right)$$

$\mathcal{S}$ denotes set of states, $|\pi|$ the length of $\pi$

Decoding to obtain the best-scoring labeling for $\mathbf{x}$:

$\operatorname{argmax} F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi})$ (Viterbi decoding)

[Durbin et al., 1998]

# Parametrization and Decoding

Log transition probabilities $\phi(k, l)$ between states $k$ and $l$

Log emission probabilities $g_k(x)$ in state $k$ for (discretized) hybridization signal $x$

Parametrization $\boldsymbol{\theta}$

Scoring a sequence of hybridization signals $\mathbf{x}$ with a given labeling $\boldsymbol{\pi}$ and parametrization $\boldsymbol{\theta}$:

$$F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi}) = \sum_{p=1}^{|\pi|} \left( g_{\pi_p}(x_p) + \phi(\pi_{p-1}, \pi_p) \right)$$

$\mathcal{S}$ denotes set of states, $|\pi|$ the length of $\pi$

Decoding to obtain the best-scoring labeling for $\mathbf{x}$:

$\operatorname{argmax} F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi})$ (Viterbi decoding)

[Durbin et al., 1998]

# Parametrization and Decoding

Log transition probabilities $\phi(k, l)$ between states $k$ and $l$

Log emission probabilities $g_k(x)$ in state $k$ for (discretized) hybridization signal $x$

Parametrization $\boldsymbol{\theta}$

Scoring a sequence of hybridization signals $\mathbf{x}$ with a given labeling $\boldsymbol{\pi}$ and parametrization $\boldsymbol{\theta}$:

$$F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi}) = \sum_{p=1}^{|\pi|} \left( g_{\pi_p}(x_p) + \phi(\pi_{p-1}, \pi_p) \right)$$

$\mathcal{S}$ denotes set of states, $|\pi|$ the length of $\pi$

Decoding to obtain the best-scoring labeling for $\mathbf{x}$:

$\operatorname{argmax} F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi})$ (Viterbi decoding)

[Durbin et al., 1998]

# Training an HMM

Training sequences: signals $\mathbf{x}^i$ and labels $\pi^i$ for $i = 1, \ldots, n$.

Log transition probabilities: [Durbin et al., 1998]

$\phi(k, l) = log(\frac{A_{k,l}}{\sum_{l'} A_{k,l'}})$ for all state pairs $(k, l) \in \mathcal{S}^2$

counting observed transitions: $A_{k,l} = \sum_{i=1}^{n} \sum_{p=1}^{|\pi^i|} [[\pi_p^i = k \wedge \pi_{p+1}^i = l]]$

Log emission probabilities: [Durbin et al., 1998]

for piece-wise constant $g_k$ with $L$ levels (ranging from $t_l$ to $t_{l+1}$):

$g_{k,l} = log(\frac{E_l}{\sum_{l'} E_{l'}})$

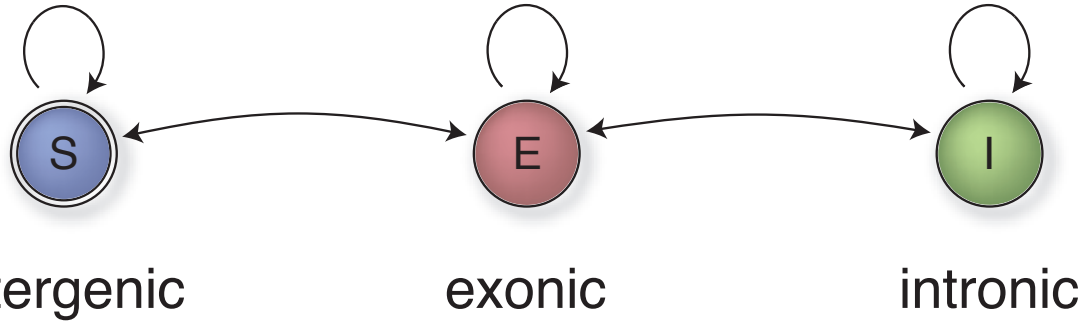
counting discrete signal values: $E_l = \sum_{i=1}^{n} \sum_{p=1}^{|\pi^i|} [[\pi_p^i = k \wedge t_l < x_p^i \leq t_{l+1}]]$

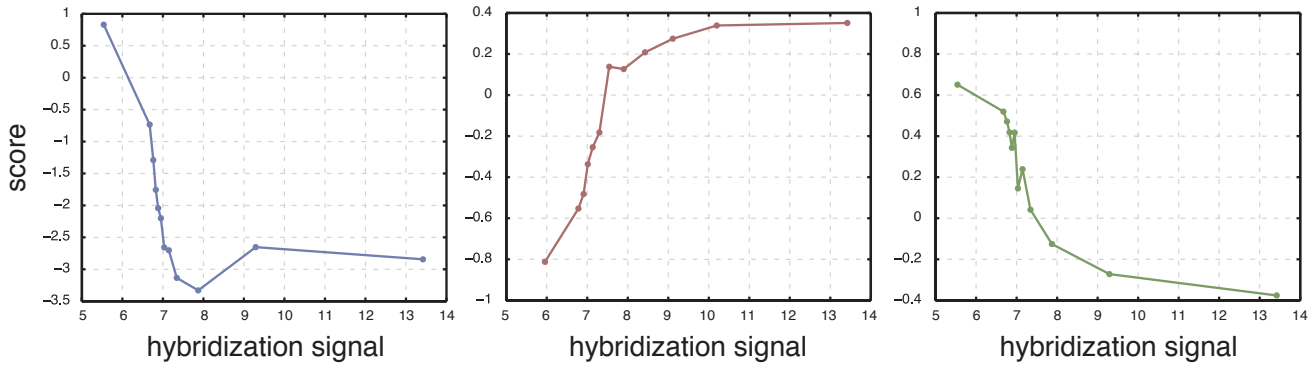
HMMs can also be (re-)trained in an unsupervised fashion [Munch et al., 2006]

# Hidden Markov SVMs

Enforce a large margin
between the correct label $\boldsymbol{\pi}^{(i)}$ and any other labeling $\overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)}$:

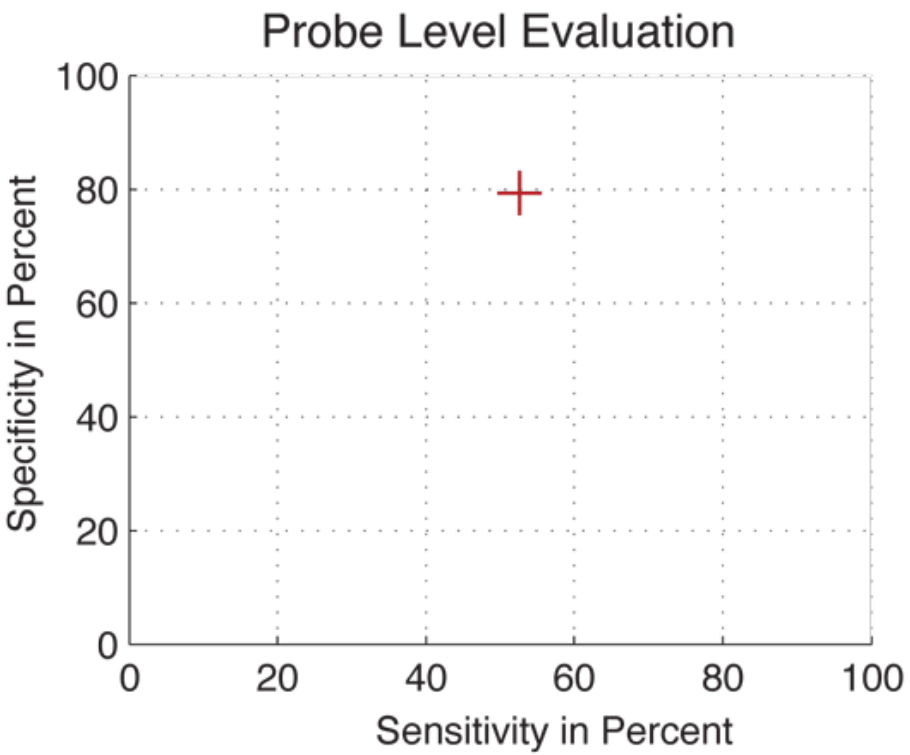
$$F_\theta(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}) - F_\theta(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \gg 0 \qquad \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1, \ldots, n$$

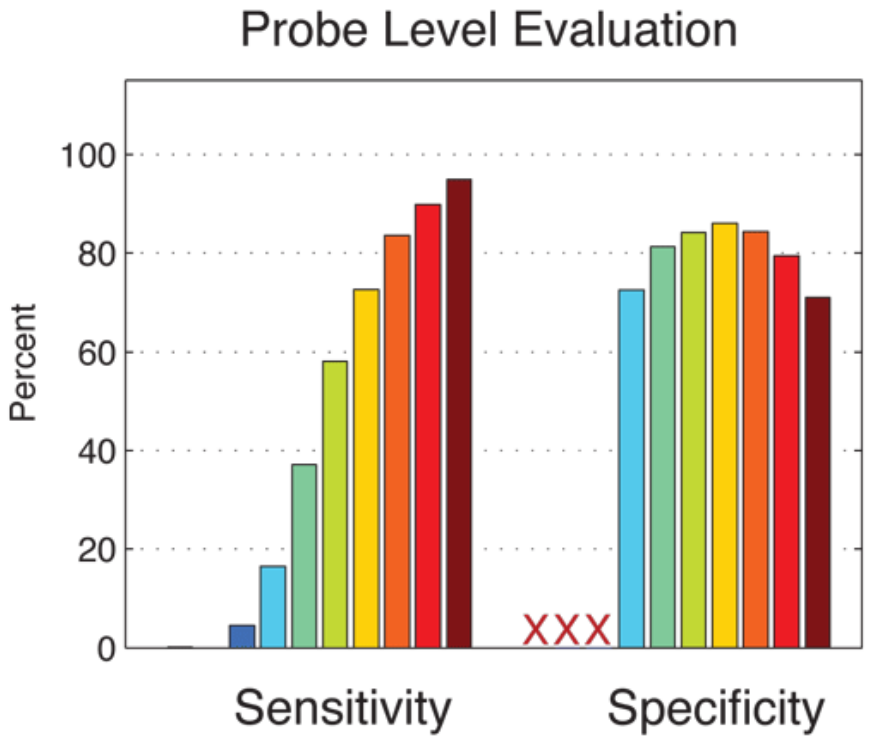


# Accuracy of TAR Predictions (cDNAs)

Higher sensitivity for higher expressed transcripts

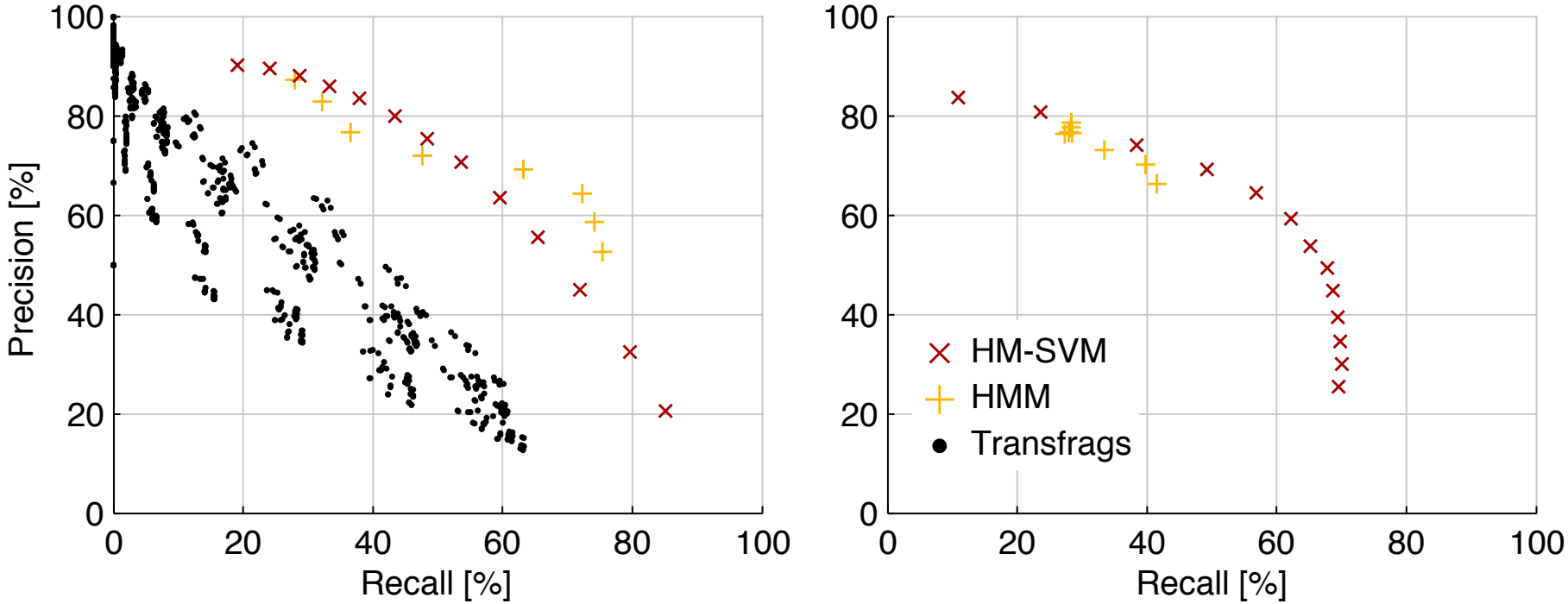



X - Specificity not reported due to very few predictions

(Comparison with genomic regions labeled using full length cDNAs.)

# Method Comparison




Recall: Proportion of annotated exons/introns covered by predictions.
Precision: Proportion of predictions covered by annotated exons/introns.

# Differentially Expressed TARs




Strategy:

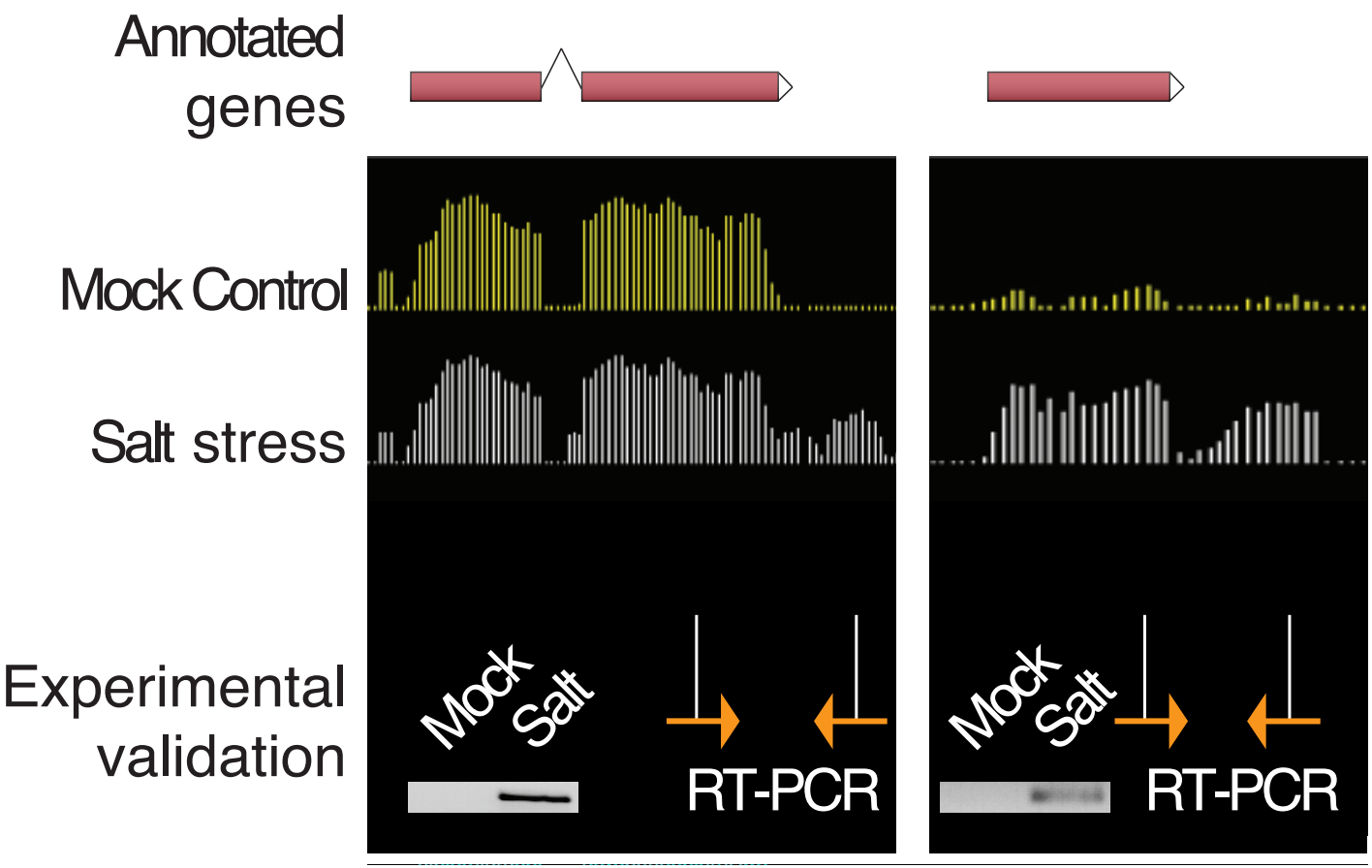
- Define probe sets using TAR predictions (separately for samples)
- To test for

  up-regulation in A vs. B, use probe sets from A,
  down-regulation in A vs. B, use probe sets from B, and

  test for differential expression between A and B.
- Asymmetry takes into account that TAR might not be present or less accurately defined in lower expressed sample.

[Zeller et al., 2009]

# Detection of Differential TARs




Apply statistical test for significant expression change to signal from transcriptionally active regions (TARs)

# Roadmap




1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Detection of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology and Limitations
   - Assembly & Read Mapping
   - Transcript Quantitation
   - De novo Transcript Prediction

# Sequencing Techniques & Applications




Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

Sequencing Technology

- Capillary/Sanger sequencing
- Pyrosequencing (Roche/454)
- SOLiD sequencing (ABI)
- Flow cell sequencing (Illumina)
- Single molecule sequencing (Nano pores, etc.)

} Next Generation Sequencing

# Sequencing Techniques & Applications




Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

Sequencing Technology

- **Capillary/Sanger sequencing**
- Pyrosequencing (Roche/454)
- SOLiD sequencing (ABI)
- Flow cell sequencing (Illumina)
- Single molecule sequencing (Nano pores, etc.)

} Next Generation Sequencing

# Sequencing Techniques & Applications




Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

Sequencing Technology

- **Capillary/Sanger sequencing**
- Pyrosequencing (Roche/454)
- SOLiD sequencing (ABI)
- Flow cell sequencing (Illumina)
- Single molecule sequencing (Nano pores, etc.)

} Next Generation Sequencing

# Sequencing Techniques & Applications




Applications of DNA/RNA sequencing:

- *De novo* genome sequencing
- Genome resequencing
- Transcriptome sequencing
- Methylation analysis

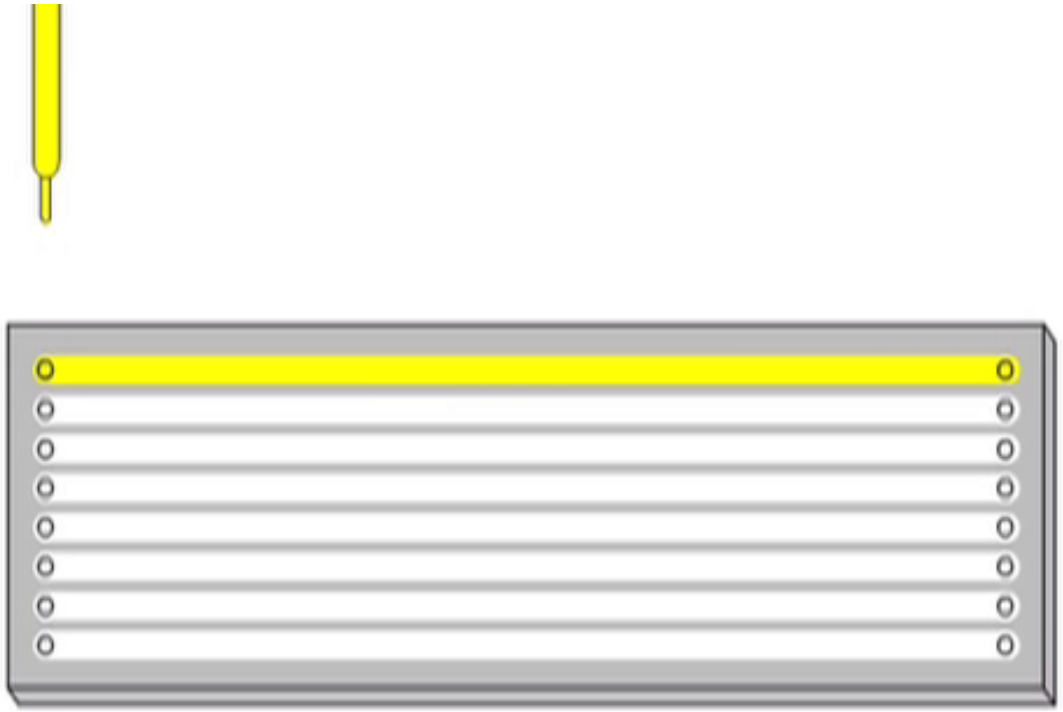
Sequencing Technology

- **Capillary/Sanger sequencing**
- Pyrosequencing (Roche/454)
- SOLiD sequencing (ABI)
- Flow cell sequencing (Illumina)
- Single molecule sequencing (Nano pores, etc.)

} Next Generation Sequencing

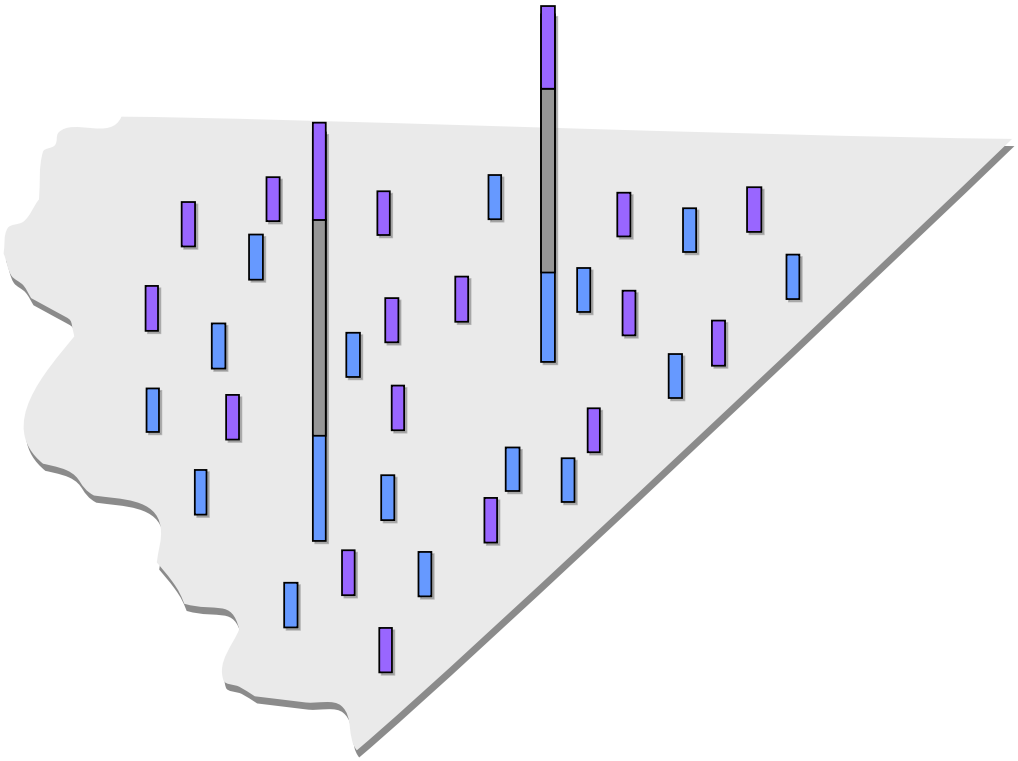
# Illumina Sequencing

- *Solexa* released a sequencing machine in 2006
- Fragment sizes from 28 – 100
- Probes are fixed to a glass plate "flow cell"
- Reagents are directed through flow cell



# Illumina Sequencing

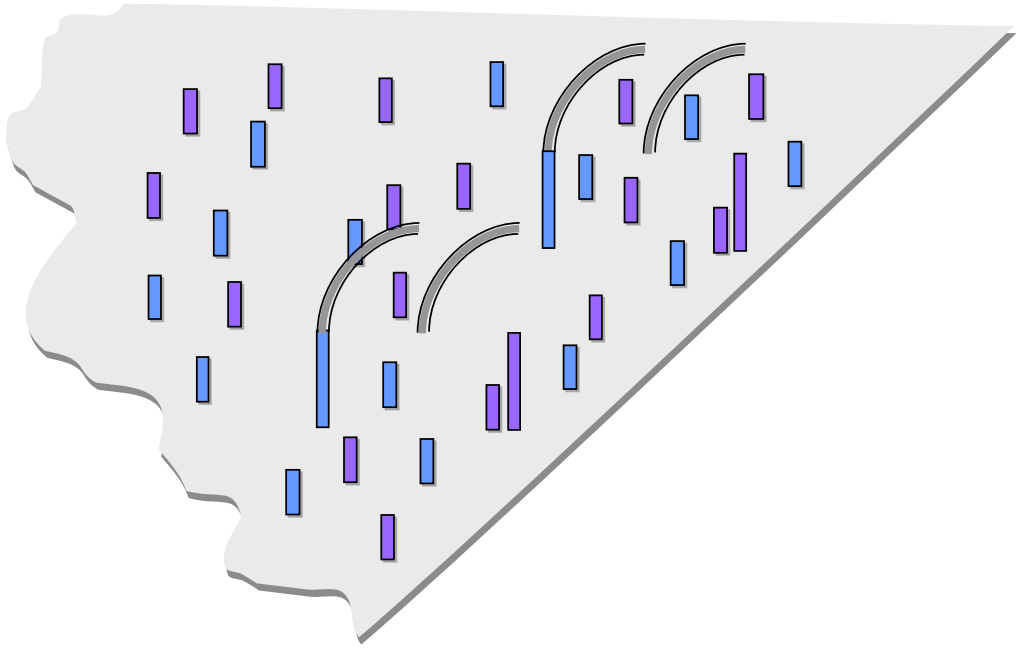
- Flow cell preparation
- Bridge amplification
- Synthesize second strand
- Denaturate to single-stranded samples
- After several cycles clusters are ready for sequencing
- Sequence the fragments



[Ossowski, 2007]

# Illumina Sequencing

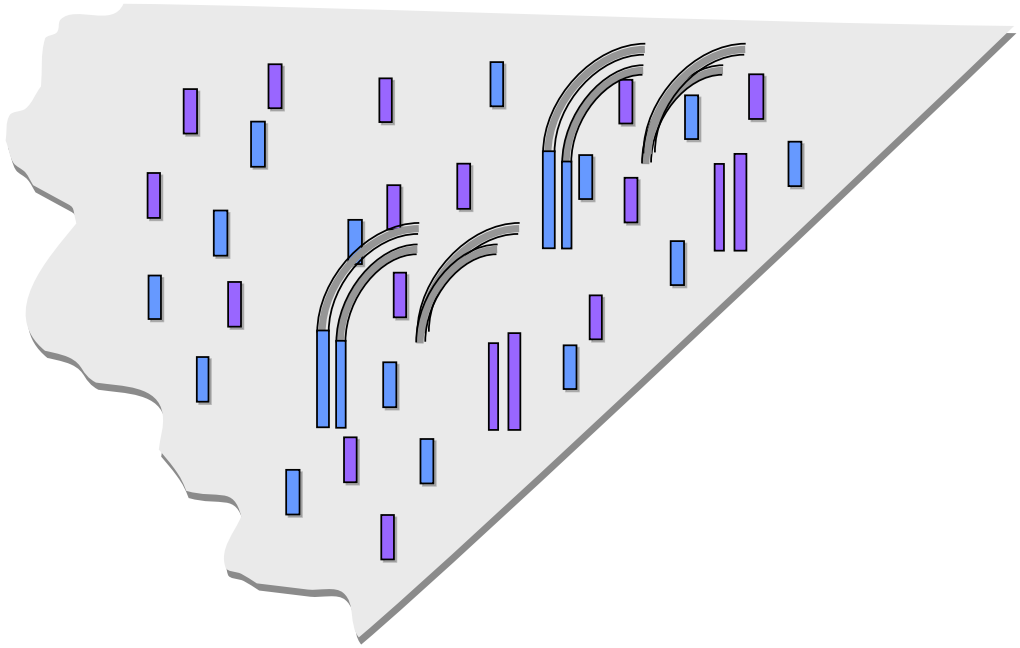
- Flow cell preparation
- Bridge amplification
- Synthesize second strand
- Denaturate to single-stranded samples
- After several cycles clusters are ready for sequencing
- Sequence the fragments



[Ossowski, 2007]

# Illumina Sequencing

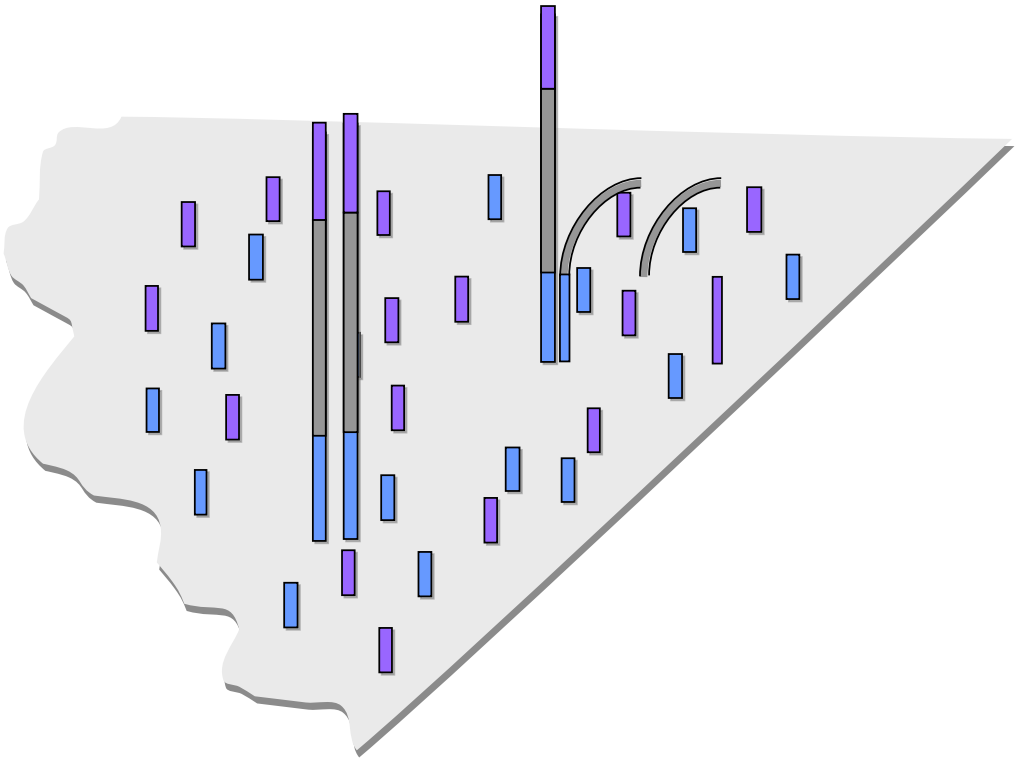
- Flow cell preparation
- Bridge amplification
- Synthesize second strand
- Denaturate to single-stranded samples
- After several cycles clusters are ready for sequencing
- Sequence the fragments



[Ossowski, 2007]

# Illumina Sequencing

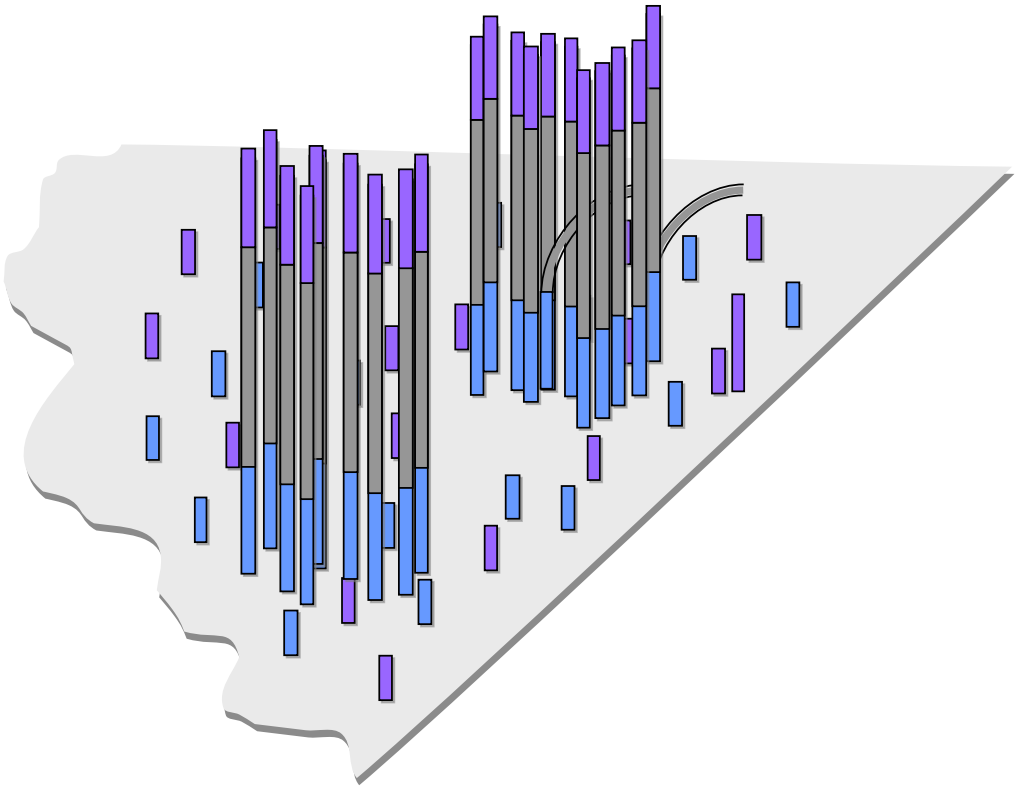
- Flow cell preparation
- Bridge amplification
- Synthesize second strand
- Denaturate to single-stranded samples
- After several cycles clusters are ready for sequencing
- Sequence the fragments



[Ossowski, 2007]

# Illumina Sequencing

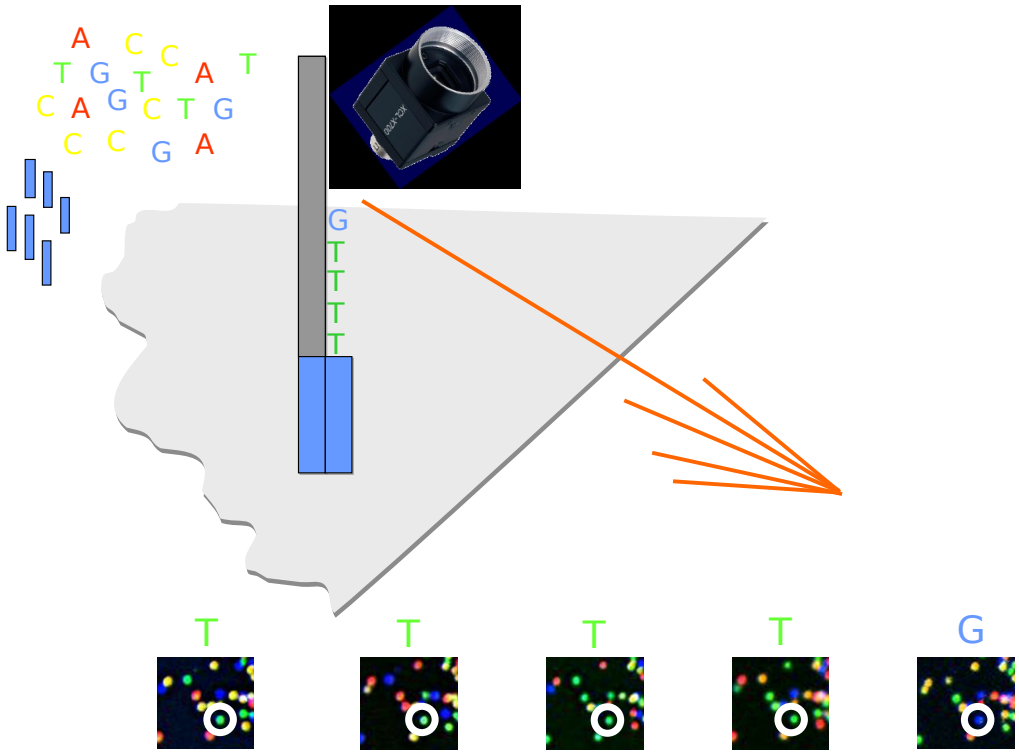
- Flow cell preparation
- Bridge amplification
- Synthesize second strand
- Denaturate to single-stranded samples
- After several cycles clusters are ready for sequencing
- Sequence the fragments



[Ossowski, 2007]

# Illumina Sequencing

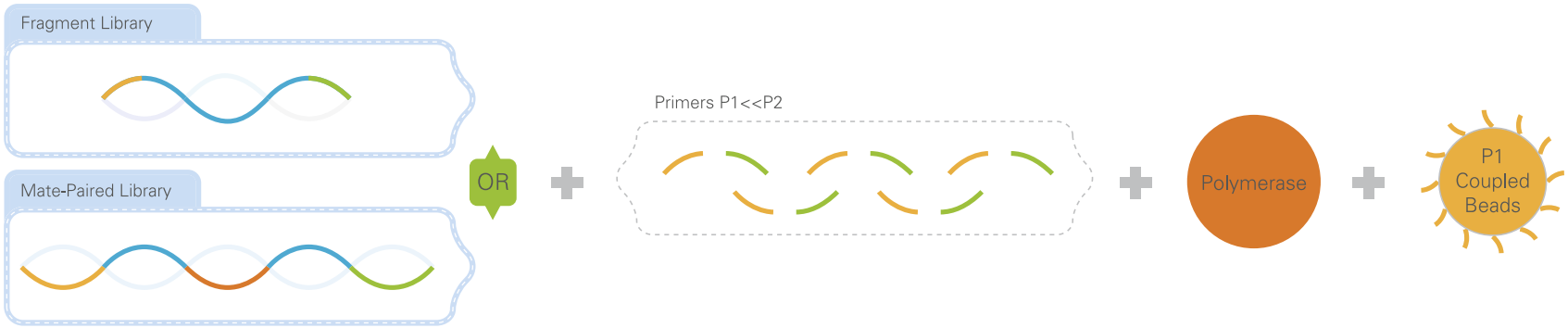
- Flow cell preparation
- Bridge amplification
- Synthesize second strand
- Denaturate to single-stranded samples
- After several cycles clusters are ready for sequencing
- Sequence the fragments




[Ossowski, 2007]

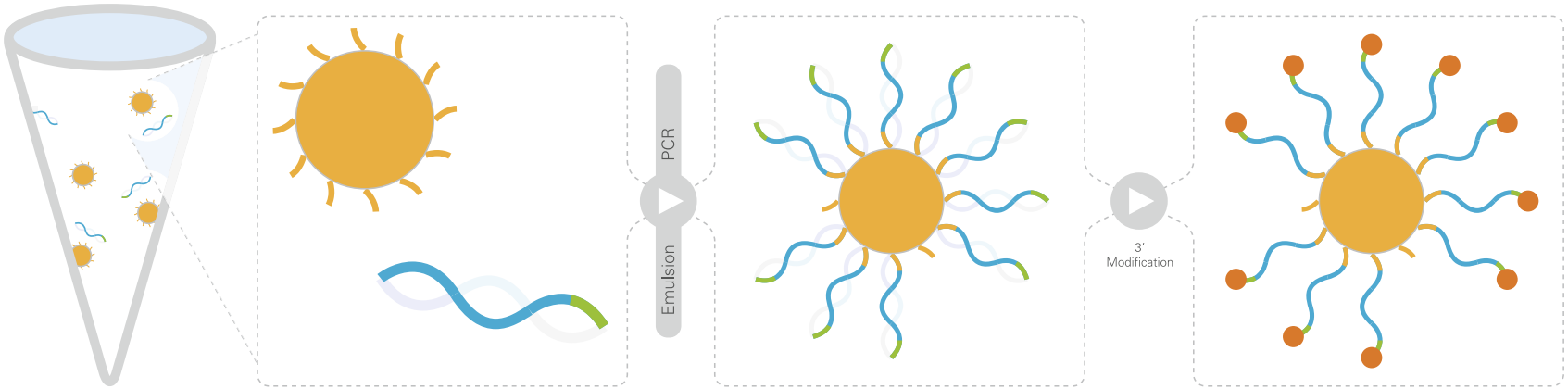
# SOLiD Sequencing

- Sequencing by ligation: Fragments ligated to "beads"
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



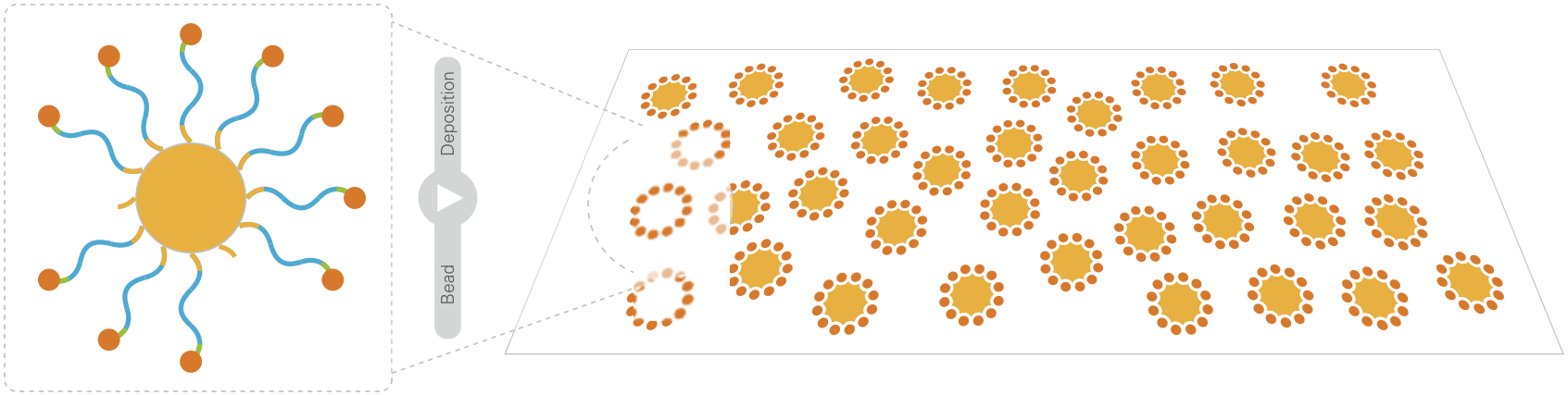
# SOLiD Sequencing

- Sequencing by ligation: Fragments ligated to “beads”
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



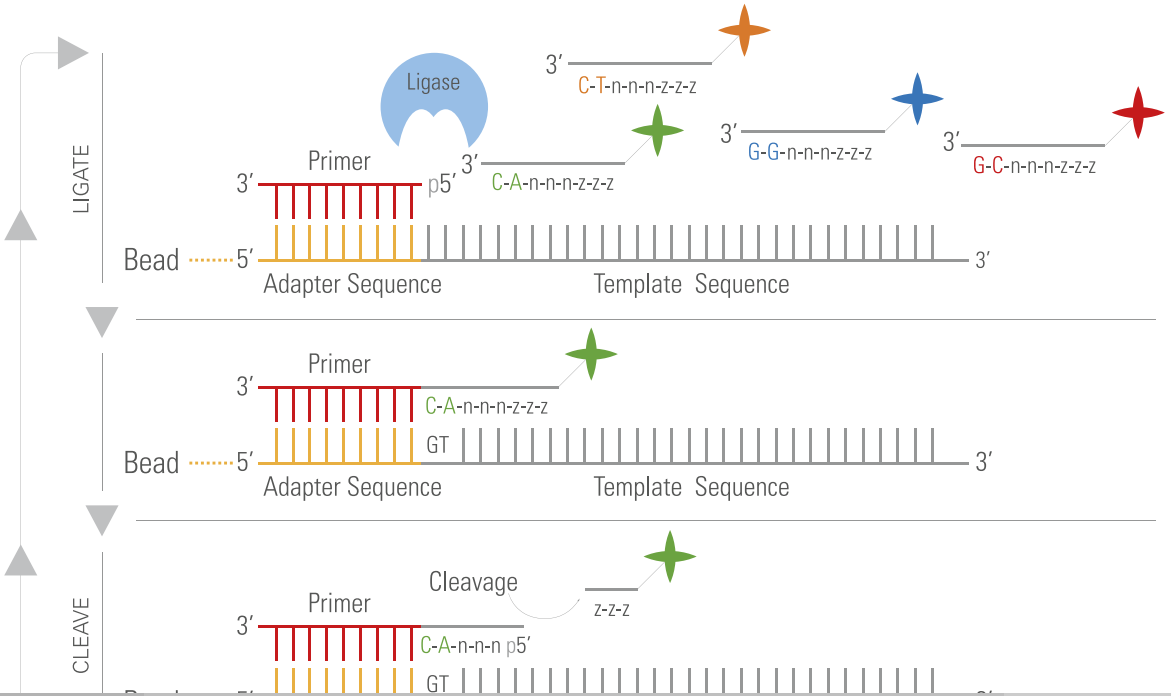
# SOLiD Sequencing

- Sequencing by ligation: Fragments ligated to “beads”
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



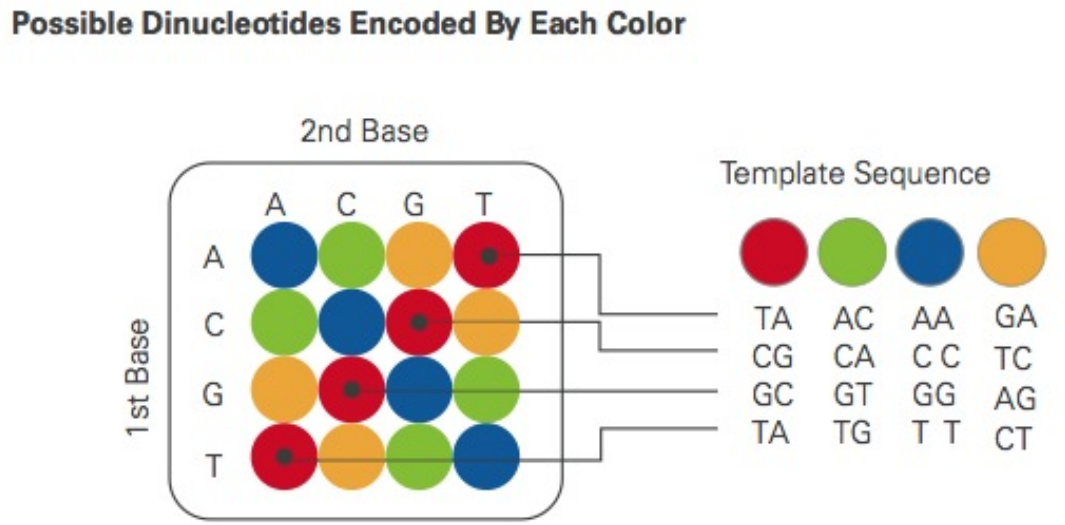
# SOLiD Sequencing

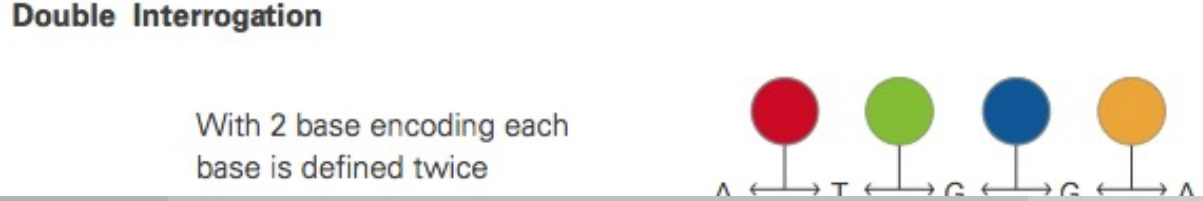
- Sequencing by ligation: Fragments ligated to “beads”
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



# SOLiD Sequencing - Color Space

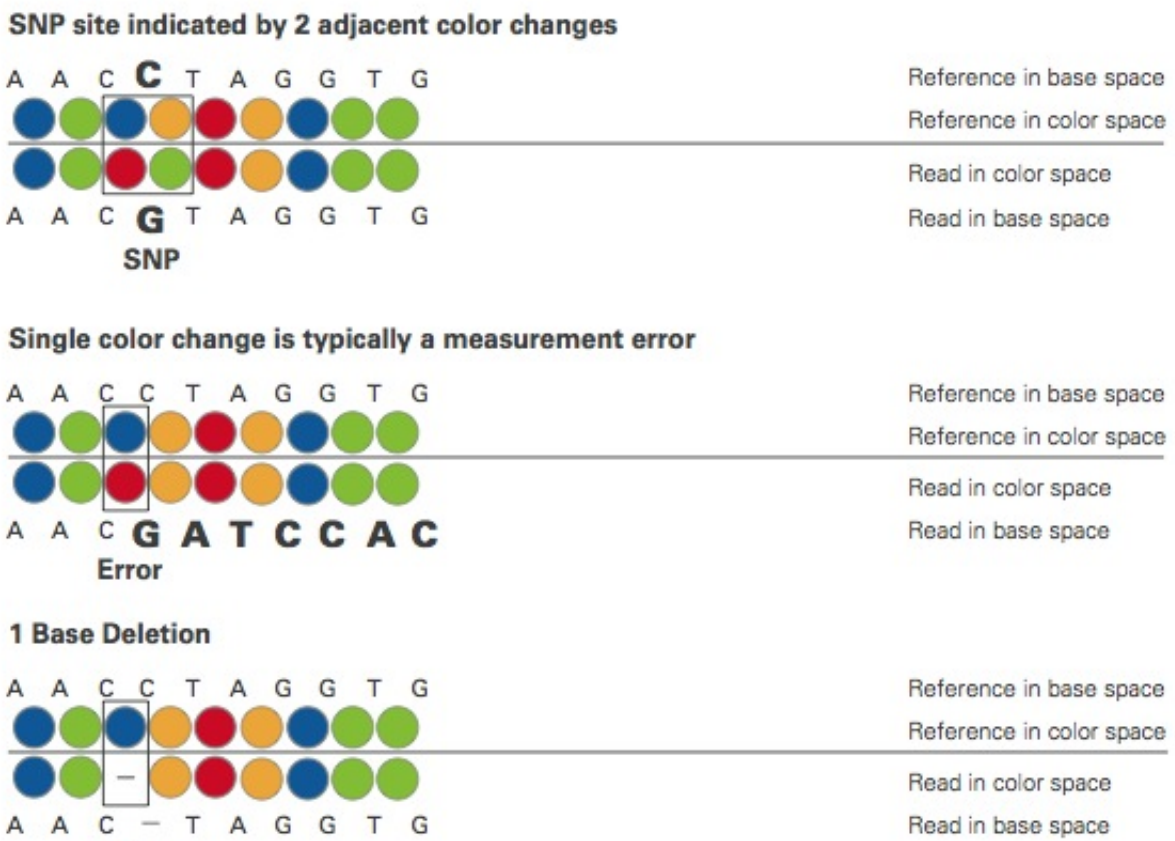
- 4 fluorescent dyes for 16 possible 2-mers
- Reverse, complement and reverse complement are always of same color



# SOLiD Sequencing - Color Space

- 4 fluorescent dyes for 16 possible 2-mers
- Reverse, complement and reverse complement are always of same color



# Overview / Extensions




| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | 40 – 100 bp | ≈6-20 Gbp | 5 – 14*d* |
| ABI SOLiD | 35 – 50 bp | ≈6-15 Gbp | 6 – 7*d* |
| Roche/454 | 300 – 500 bp | ≈100 Mbp | 7*h* |
| Sanger | 1000 bp | ≈67 kbp | 1*h* |

There are several extensions available:

- mate-pair / paired-end
- bisulfite treatment
- multiplexing $8 \times 12$
  $\rightsquigarrow$ 96 samples per flow cell

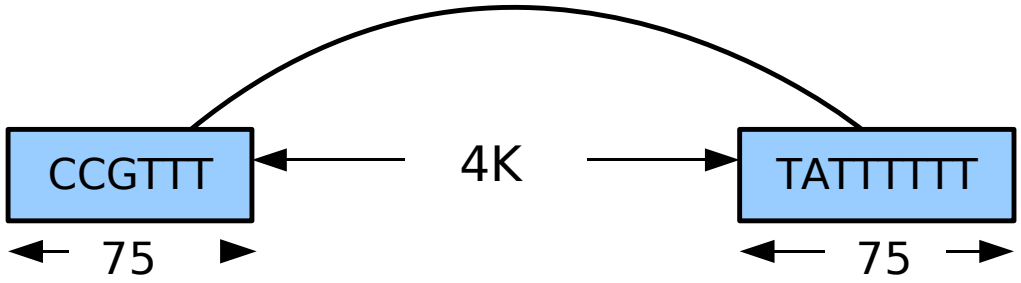
[Sutskever, 2008]

# Overview / Extensions

| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | 40 – 100 bp | ≈6-20 Gbp | 5 – 14*d* |
| ABI SOLiD | 35 – 50 bp | ≈6-15 Gbp | 6 – 7*d* |
| Roche/454 | 300 – 500 bp | ≈100 Mbp | 7*h* |
| Sanger | 1000 bp | ≈67 kbp | 1*h* |

There are several extensions available:

- mate-pair / paired-end
- bisulfite treatment
- multiplexing $8 \times 12$
  $\rightsquigarrow$ 96 samples per flow cell




[Sutskever, 2008]

# Overview / Extensions




| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | 40 – 100 bp | ≈6-20 Gbp | 5 – 14*d* |
| ABI SOLiD | 35 – 50 bp | ≈6-15 Gbp | 6 – 7*d* |
| Roche/454 | 300 – 500 bp | ≈100 Mbp | 7*h* |
| Sanger | 1000 bp | ≈67 kbp | 1*h* |

There are several extensions available:

- mate-pair / paired-end
- bisulfite treatment
- multiplexing $8 \times 12$
  $\rightsquigarrow$ 96 samples per flow cell

[Sutskever, 2008]

# Overview / Extensions

| Technology | Read length | Output/run | Run time |
|---|---|---|---|
| Illumina GA *II* | 40 – 100 bp | ≈6-20 Gbp | 5 – 14*d* |
| ABI SOLiD | 35 – 50 bp | ≈6-15 Gbp | 6 – 7*d* |
| Roche/454 | 300 – 500 bp | ≈100 Mbp | 7*h* |
| Sanger | 1000 bp | ≈67 kbp | 1*h* |

There are several extensions available:

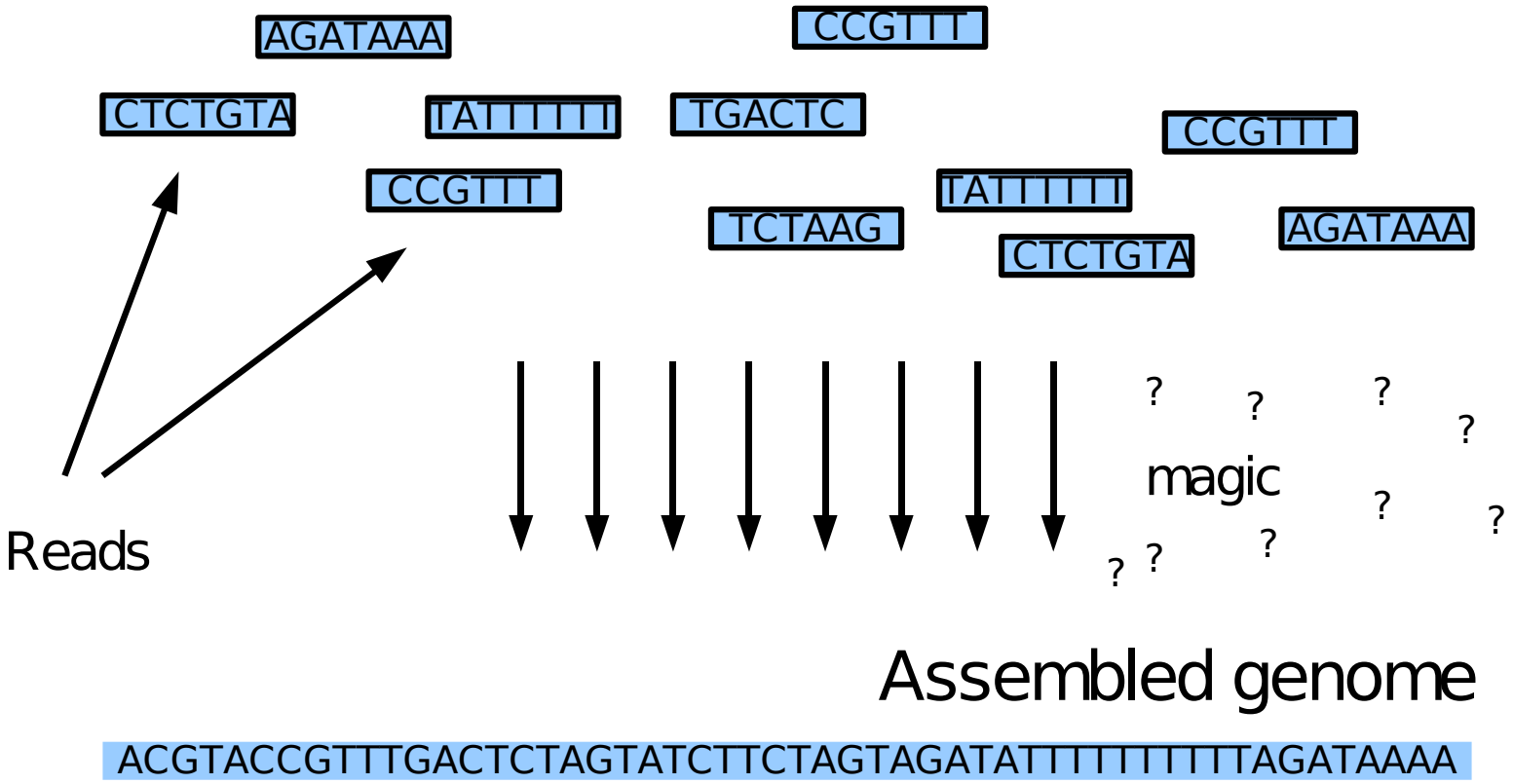
- mate-pair / paired-end
- bisulfite treatment
- multiplexing $8 \times 12$
  $\rightsquigarrow$ 96 samples per flow cell

[Sutskever, 2008]

# Short Reads Analysis - Methods

Given read data the following analysis steps are possible:

- Assembly
- Mapping/Alignments



# Short Reads Analysis - Methods




Given read data the following analysis steps are possible:

- Assembly
- Mapping/Alignments

```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
       ACCAGTGACCTGACTACTACGTCGTAACGTACACG
```

# Short Reads Analysis - Methods




Given read data the following analysis steps are possible:

- Assembly
- Mapping/Alignments

```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTGCTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTGCTACGTCGTCACGTACA
     AAACCATTGACCTGACTGCTACGTCGTAACGTACAC
      AACCAGTGACCTGACTGCTACGTCGTAACGTACACG
       AACAGTGACCTGACTGCTACGTCGTAACGTACACG
```

# Short Reads Analysis - Methods




Given read data the following analysis steps are possible:

- Assembly
- Mapping/Alignments

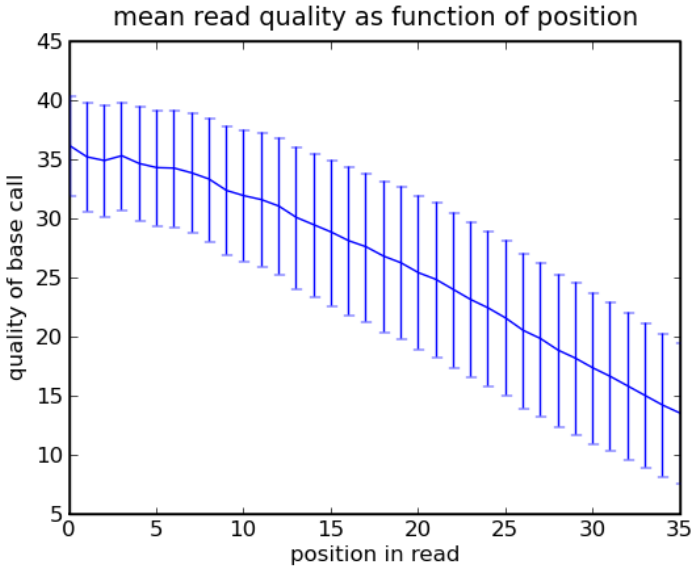
```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTGCTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTGCTACGTCGTCACGTACA
     AAACCATTGACCTGACTGCTACGTCGTAACGTACAC
      AACCAGTGACCTGACTGCTACGTCGTAACGTACACG
       AACAGTGACCTGACTGCTACGTCGTAACGTACACG
```

**Problem:** 100 million reads of short length

$\Rightarrow$ Big computational challenge

# Short Reads Analysis - Problems

- Experiment leads to millions of reads 36 – 100nt
- Reads may have a position-wise varying quality

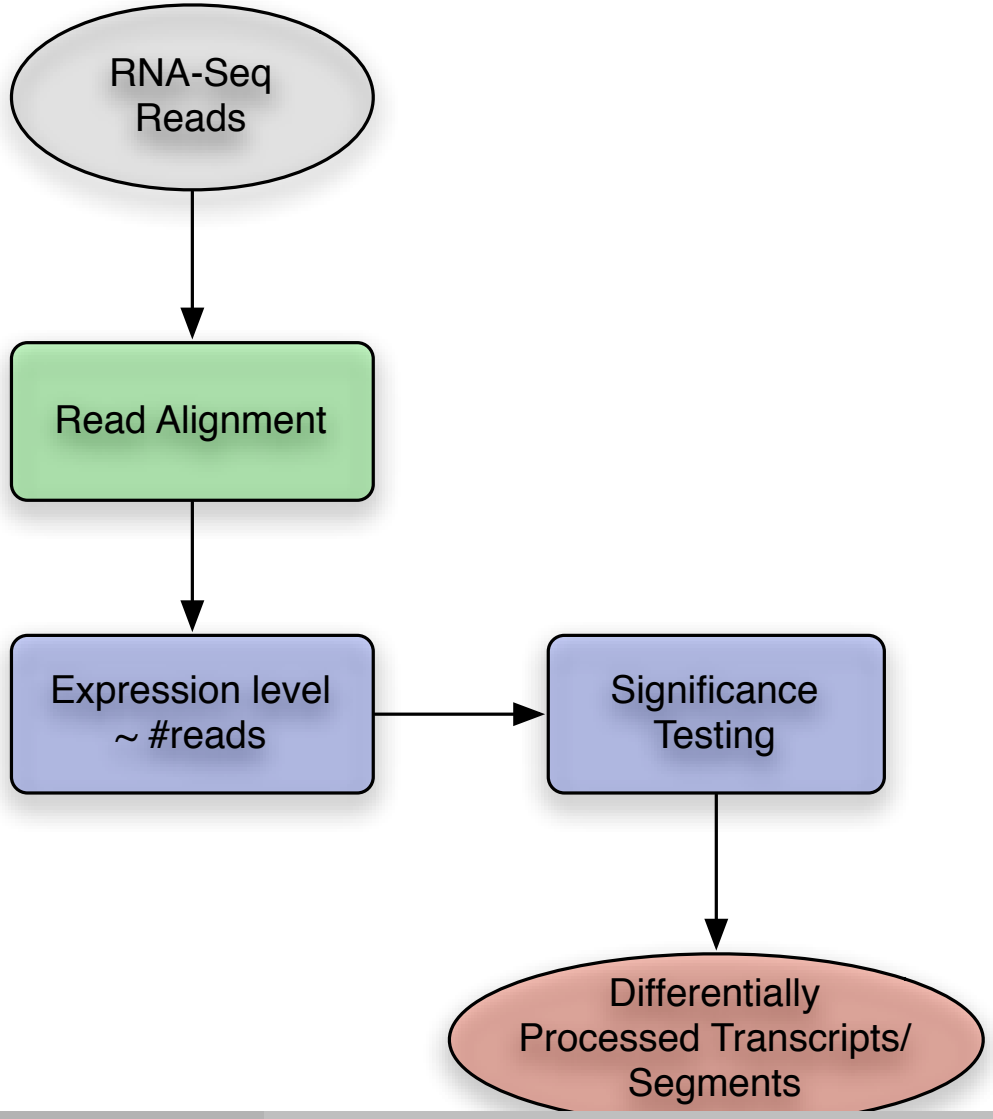



- Quality corresponds to error probability:

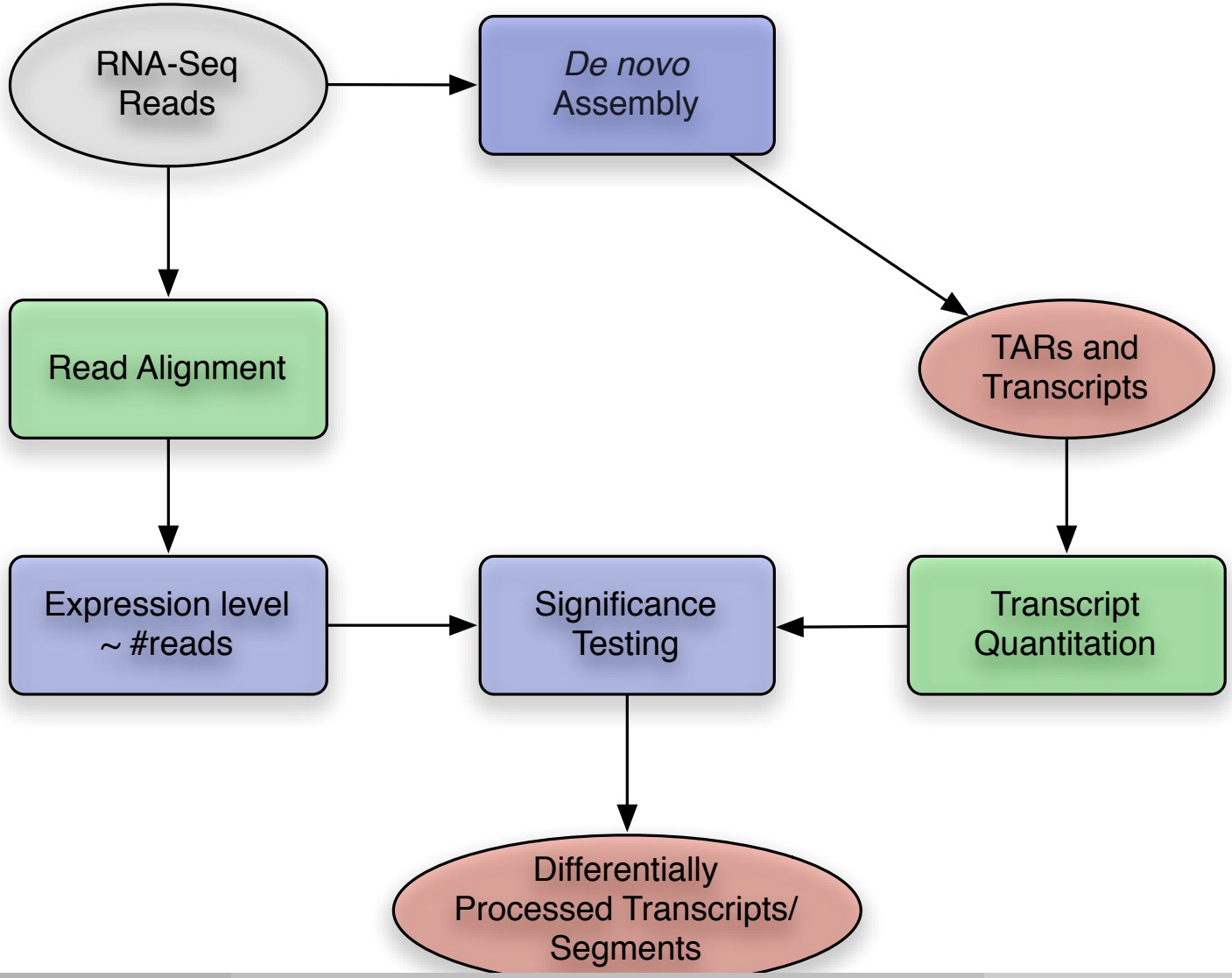
$$q = -10 \cdot log_{10}(\frac{p}{1-p})$$

- Example: If we have an error probability of $10^{-3}$ per base then

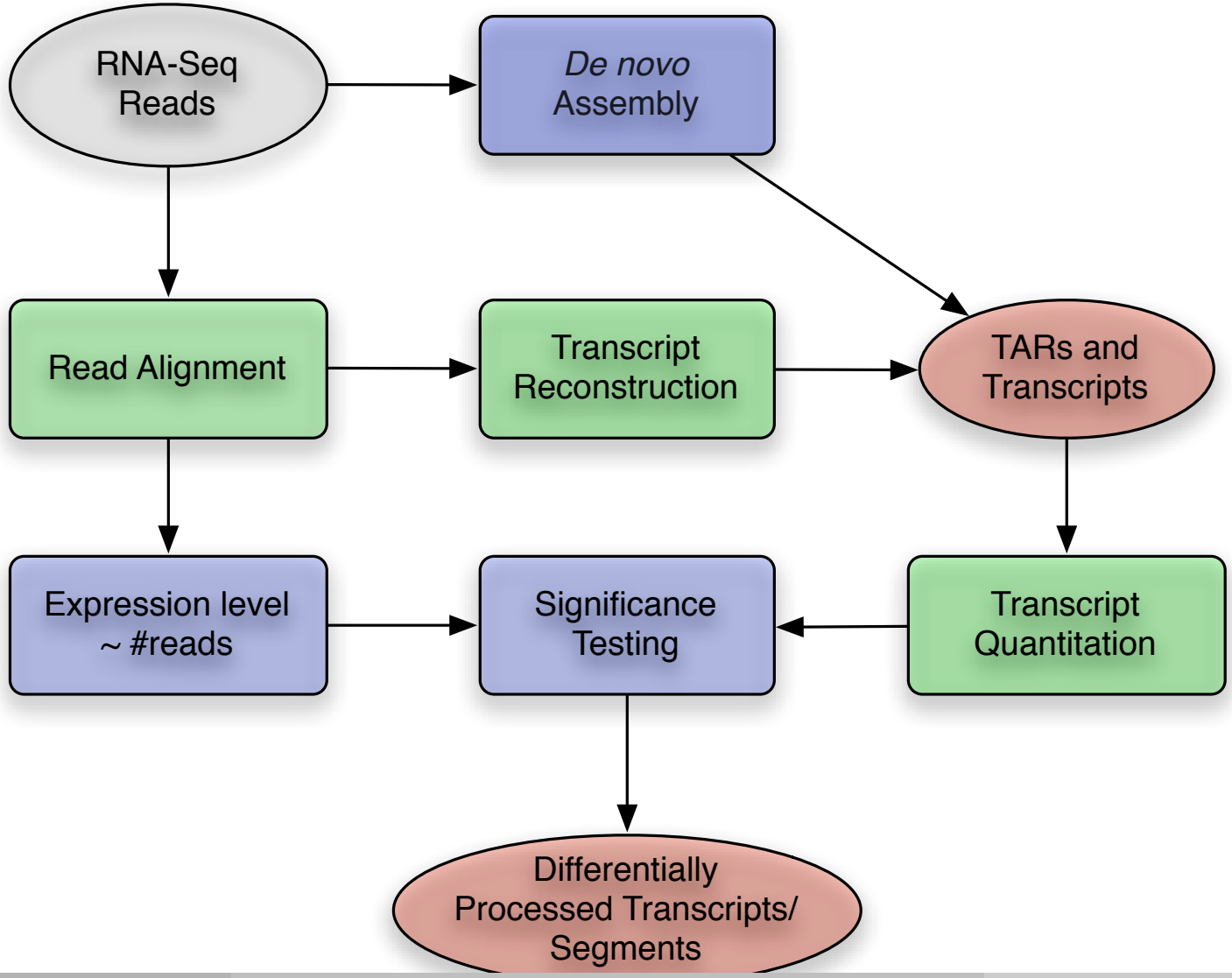
# Common RNA-Seq Analysis Steps



# Common RNA-Seq Analysis Steps



# Common RNA-Seq Analysis Steps



# Short Reads Assembly




## Read assembly problem

For a set of reads stemming from a reference genome find maximally overlapping parts in order to reconstruct the genomic sequence

Classical assembly: $\Rightarrow$ **Too inefficient for short reads**

1. *Overlap phase*: Every read is compared with every other read and the overlap graph is computed
2. *Layout phase*: Pairs are determined that position every read in the assembly
3. *Consensus phase*: Multi-alignment of all the placed reads is produced to obtain the final sequence

**New techniques:** Many new tools available (EULER, *VELVET*, SHARCGS, SSAKE/VCAKE, . . . )

**Idea:** de Bruijn Graphs

# Short Reads Assembly




## Read assembly problem

For a set of reads stemming from a reference genome find maximally overlapping parts in order to reconstruct the genomic sequence

Classical assembly: $\Rightarrow$ **Too inefficient for short reads**

1. *Overlap phase*: Every read is compared with every other read and the overlap graph is computed
2. *Layout phase*: Pairs are determined that position every read in the assembly
3. *Consensus phase*: Multi-alignment of all the placed reads is produced to obtain the final sequence

**New techniques:** Many new tools available (EULER, *VELVET*, SHARCGS, SSAKE/VCAKE, . . . )

**Idea:** de Bruijn Graphs

# Short Reads Assembly




## Read assembly problem

For a set of reads stemming from a reference genome find maximally overlapping parts in order to reconstruct the genomic sequence

Classical assembly: $\Rightarrow$ **Too inefficient for short reads**

1. *Overlap phase*: Every read is compared with every other read and the overlap graph is computed
2. *Layout phase*: Pairs are determined that position every read in the assembly
3. *Consensus phase*: Multi-alignment of all the placed reads is produced to obtain the final sequence

**New techniques:** Many new tools available (EULER, *VELVET*, SHARCGS, SSAKE/VCAKE, . . . )

**Idea:** de Bruijn Graphs

# Short Reads Assembly




## Read assembly problem

For a set of reads stemming from a reference genome find maximally overlapping parts in order to reconstruct the genomic sequence

Classical assembly: $\Rightarrow$ **Too inefficient for short reads**

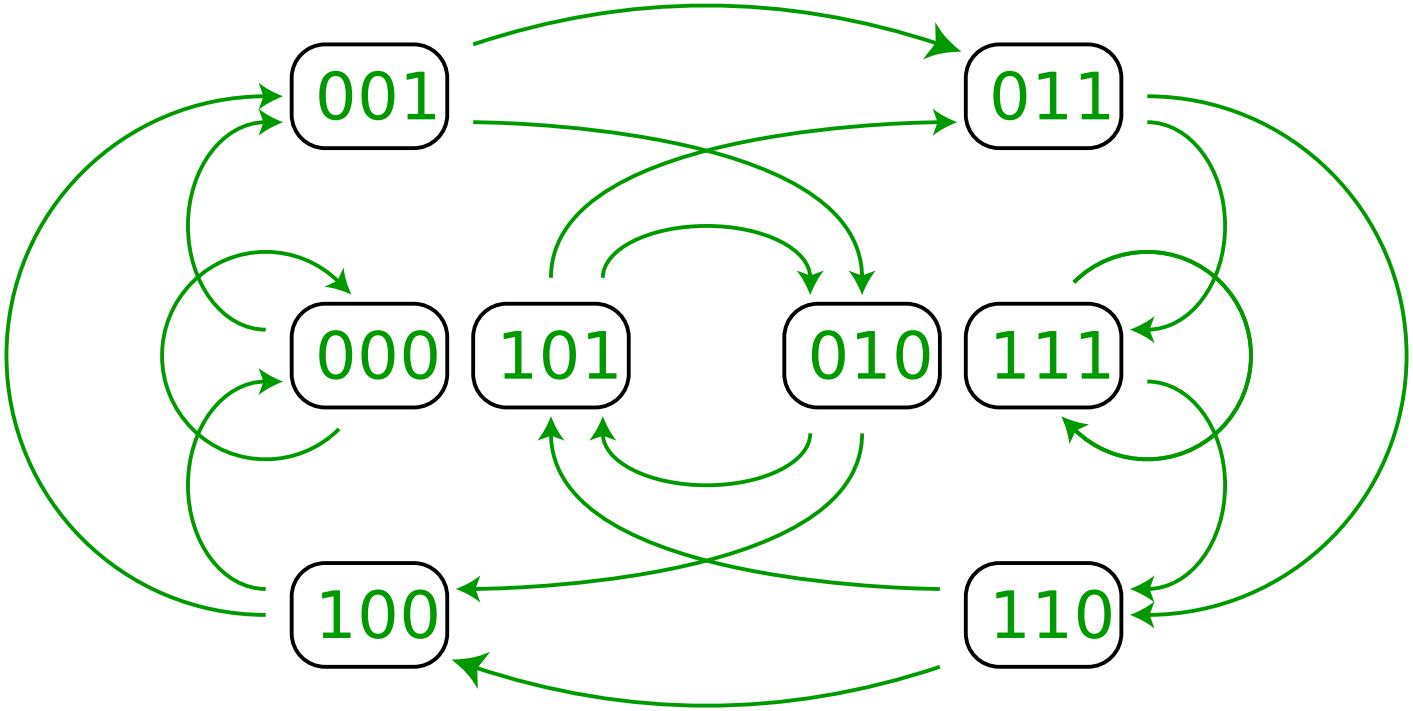
1. *Overlap phase*: Every read is compared with every other read and the overlap graph is computed
2. *Layout phase*: Pairs are determined that position every read in the assembly
3. *Consensus phase*: Multi-alignment of all the placed reads is produced to obtain the final sequence

**New techniques:** Many new tools available (EULER, *VELVET*, SHARCGS, SSAKE/VCAKE, . . . )

**Idea:** de Bruijn Graphs

# Short Reads Assembly - de Bruijn Graphs

Example:

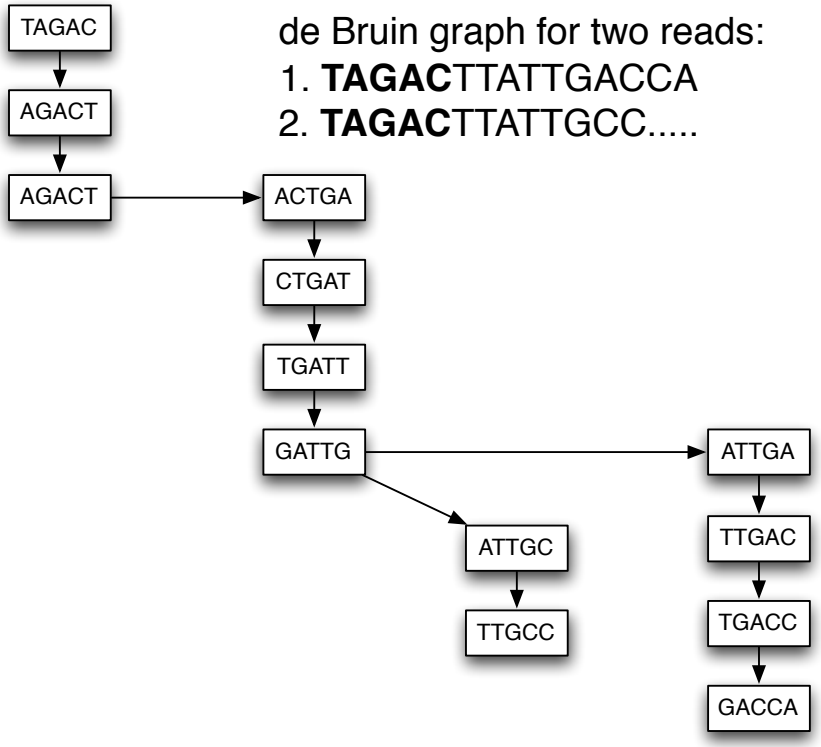



- Reads are mapped as a path in the graph
- Number of reads does not influence number of nodes
  $\Rightarrow$ Use de-Bruijn graphs to solve the problem

[Wikipedia]

# de Bruijn Graphs

Example:

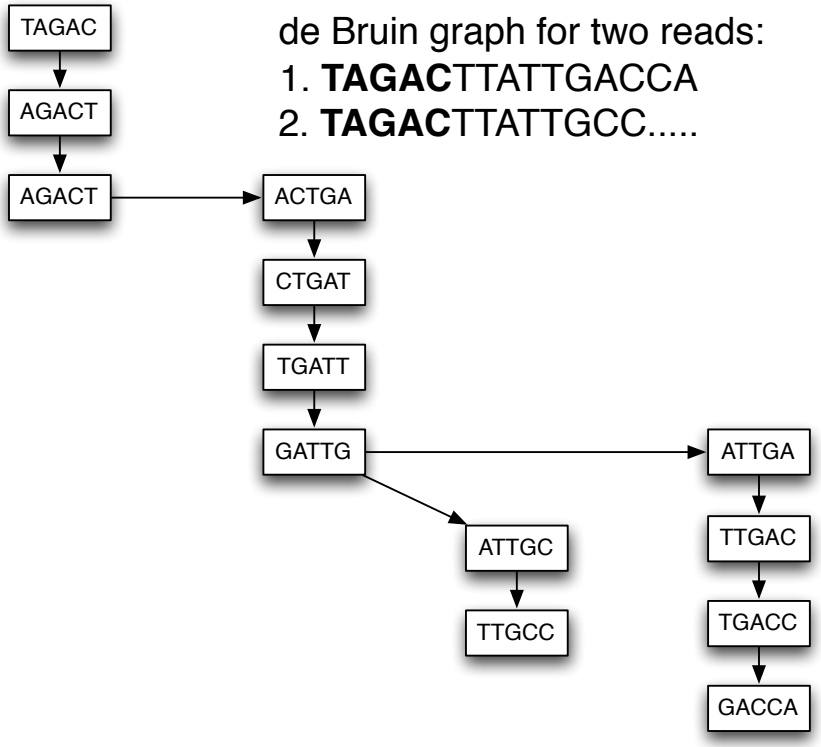



[Zerbino and Birney, 2008]

- Nodes represent *k*-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths

# de Bruijn Graphs

Example:

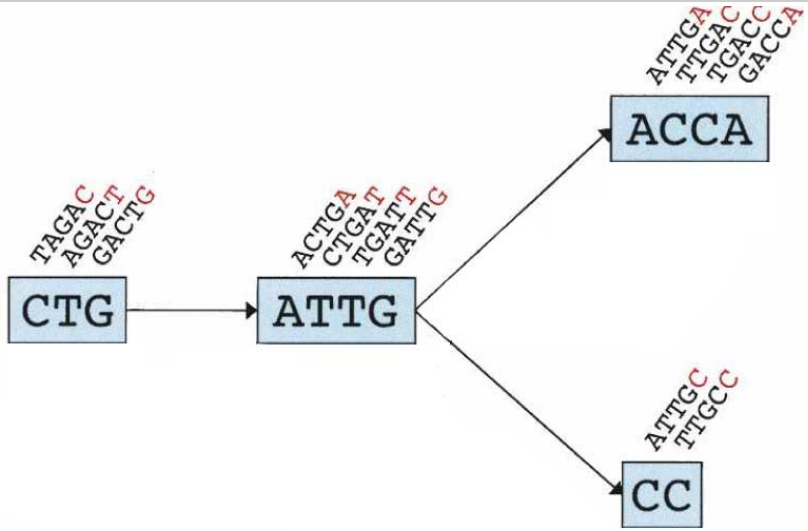



[Zerbino and Birney, 2008]

- Nodes represent *k*-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths

# de Bruijn Graphs

Example:

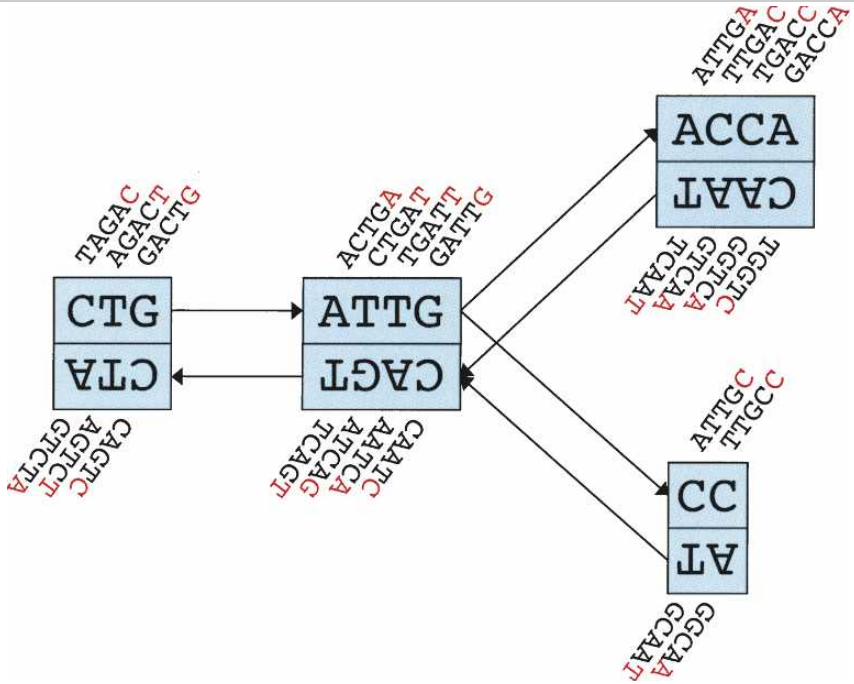



[Zerbino and Birney, 2008]

- Nodes represent $k$-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths
- Each node also stores the reverse complement

# de Bruijn Graphs

Example:



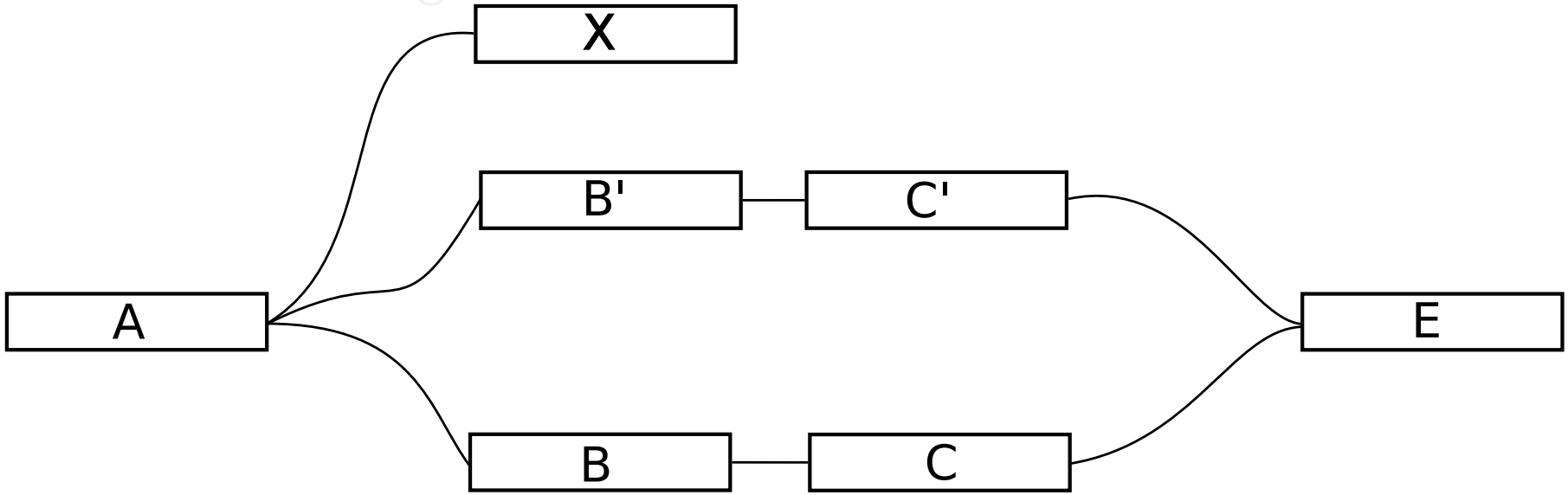

[Zerbino and Birney, 2008]

- Nodes represent *k*-mers smaller than read length
- A k-mer can refer to thousands of reads containing it
- Read errors or ambiguities lead to branching of paths
- Each node also stores the reverse complement

# Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



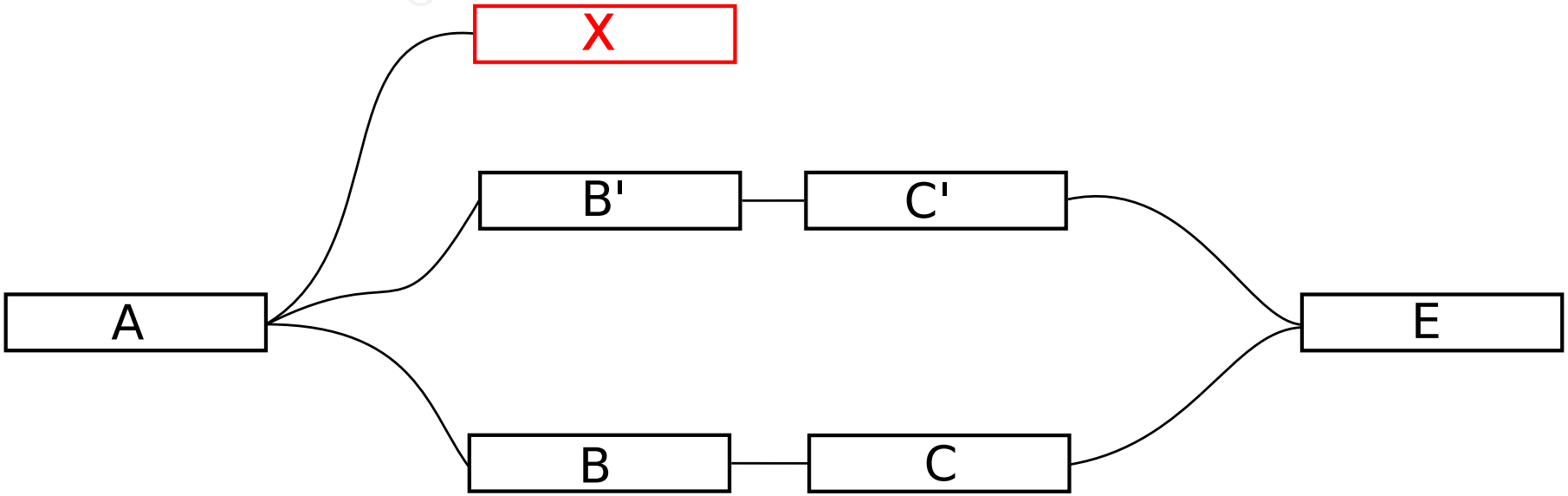

Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

# Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



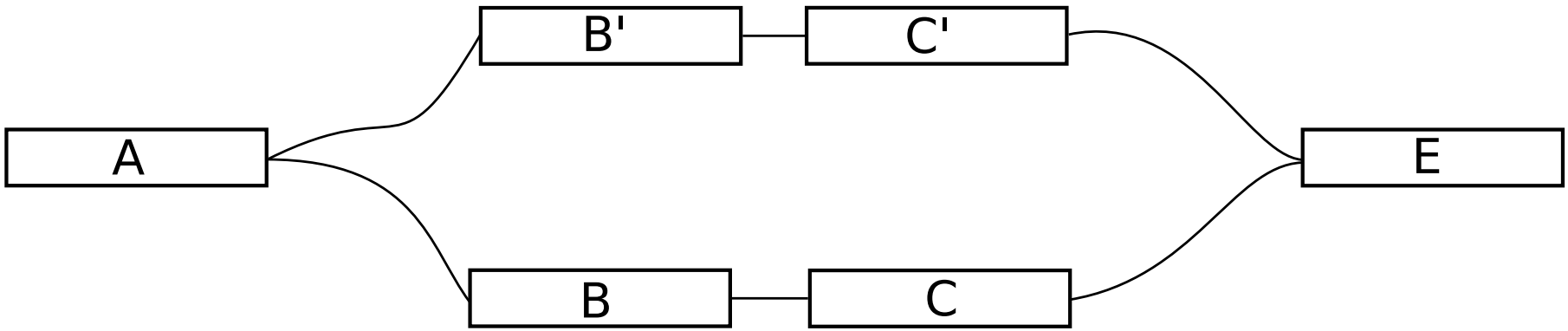

Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

# Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



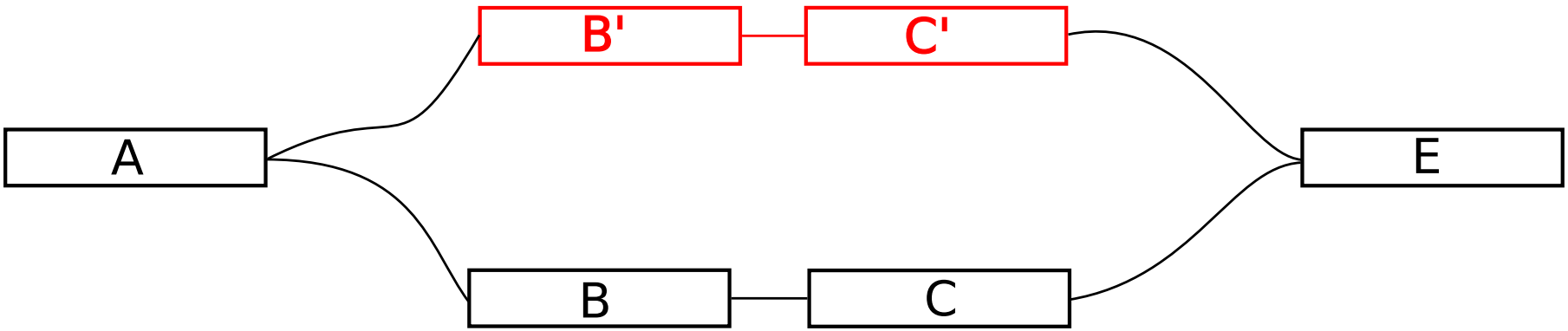

Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

# Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



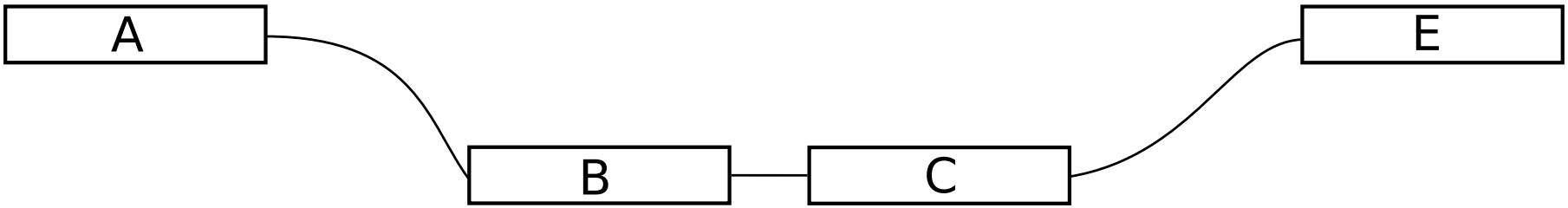

Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

# Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges




Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

# Results for Velvet

**Table 3.** Comparison of short read assemblers on experimental *Streptococcus suis* Solexa reads

| Assembler | No. of contigs | N50 | Average error rate | Memory | Time | Seq. Cov. |
|---|---|---|---|---|---|---|
| Velvet 0.3 | 470 | 8661 bp | 0.02% | 2.0G | 2 min 57 sec | 97% |
| SSAKE 2.0 | 265 | 1727 bp | 0.20% | 1.7G | 1 h 47 min | 16% |
| VCAKE 1.0 | 7675 | 1137 bp | 0.64% | 1.8G | 4 h 25 min | 134% |

Considerably shorter fragments for larger genomes

[Zerbino and Birney, 2008]

# Results for Velvet

**Table 3.** Comparison of short read assemblers on experimental *Streptococcus suis* Solexa reads

| Assembler | No. of contigs | N50 | Average error rate | Memory | Time | Seq. Cov. |
|---|---|---|---|---|---|---|
| Velvet 0.3 | 470 | 8661 bp | 0.02% | 2.0G | 2 min 57 sec | 97% |
| SSAKE 2.0 | 265 | 1727 bp | 0.20% | 1.7G | 1 h 47 min | 16% |
| VCAKE 1.0 | 7675 | 1137 bp | 0.64% | 1.8G | 4 h 25 min | 134% |

**Considerably shorter fragments for larger genomes**

[Zerbino and Birney, 2008]

# Short Reads Analysis - Mapping




## Reads mapping problem

For each read find its target regions on the reference genome such that are at most $k$ mismatches between read and target.

- Global/local alignment of all reads prohibitive
- A read stems from a certain small region
- Find this region and then do an alignment
  - spaced seeds
  - suffix trees/arrays
- Common tools: *GenomeMapper*, *Shrimp*, *SOAP*, *VMATCH*, ...

# Short Reads Analysis - Mapping



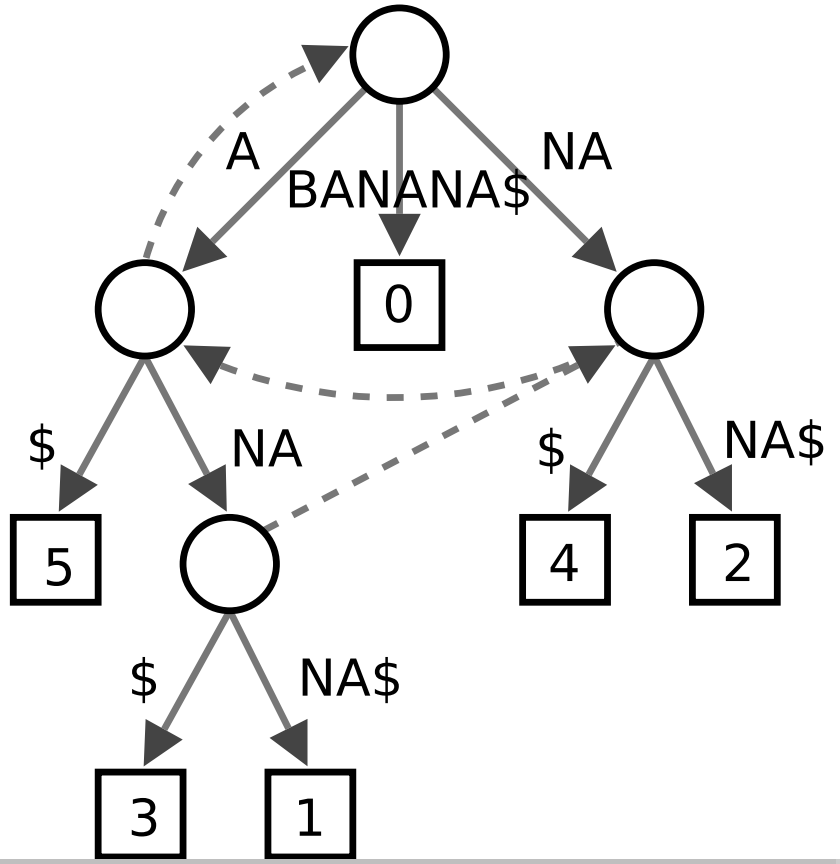

## Reads mapping problem

For each read find its target regions on the reference genome such that are at most $k$ mismatches between read and target.

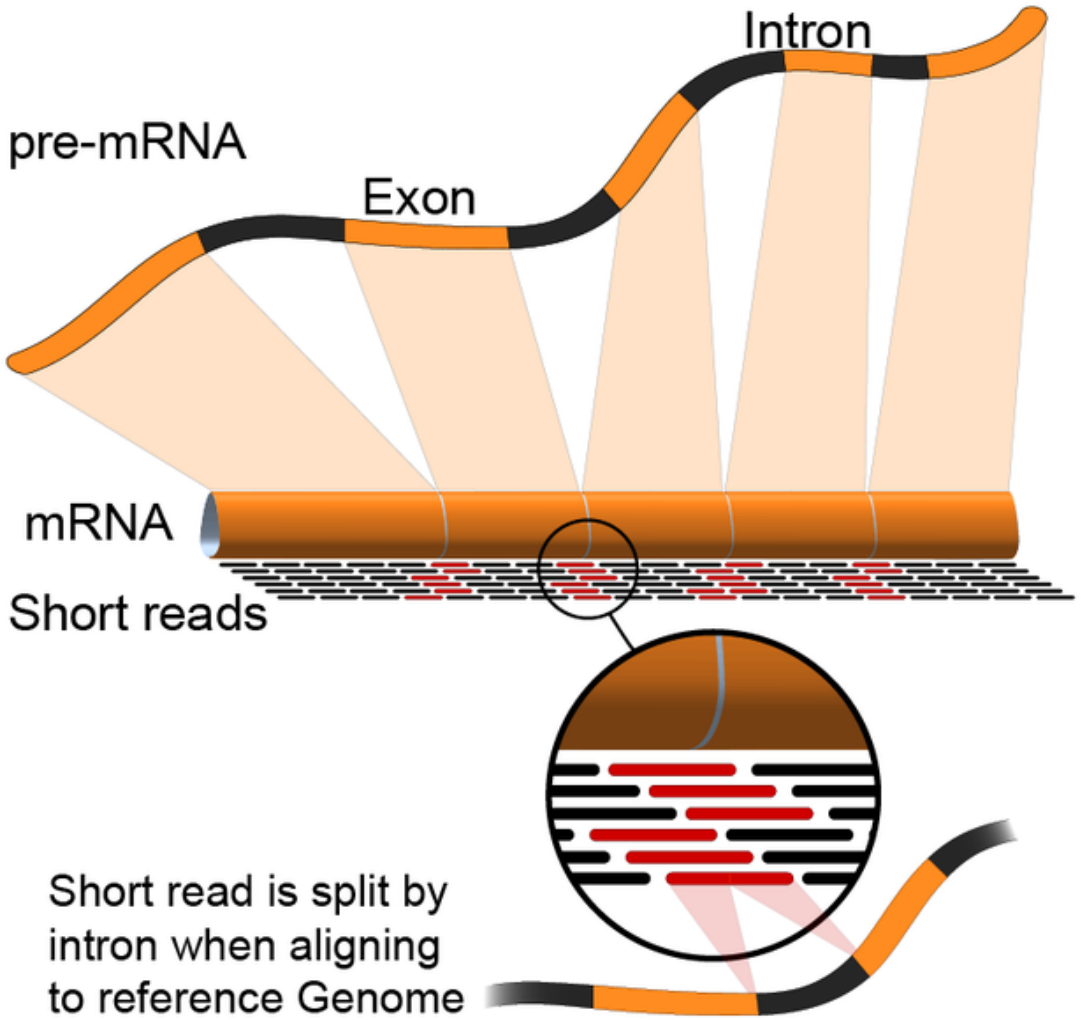
- Global/local alignment of all reads prohibitive
- A read stems from a certain small region
- Find this region and then do an alignment
  - spaced seeds
  - suffix trees/arrays
- Common tools: *GenomeMapper*, *Shrimp*, *SOAP*, *VMATCH*, ...

# Mapping via Suffix Arrays

- Given a long fixed string of length $n$ and smaller patterns of lengths $m$ to be searched for
- Construction in $\mathcal{O}(n)$, Patterns can be detected in $\mathcal{O}(m)$

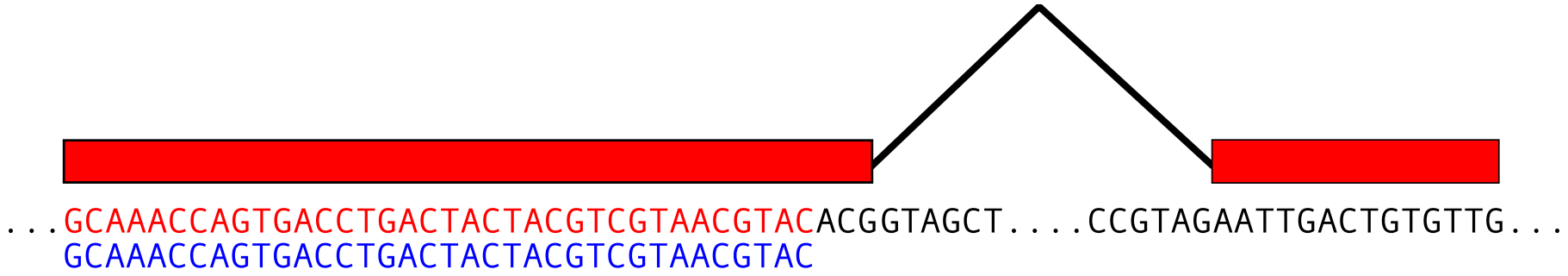


# Spliced vs. Unspliced Alignments



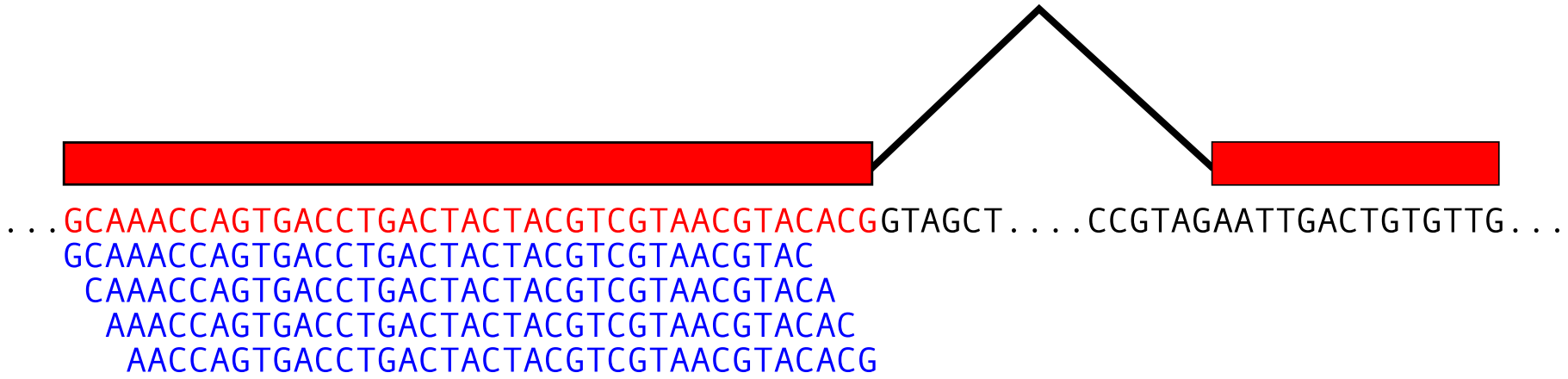

[Wikipedia]

# Spliced vs. Unspliced Alignments



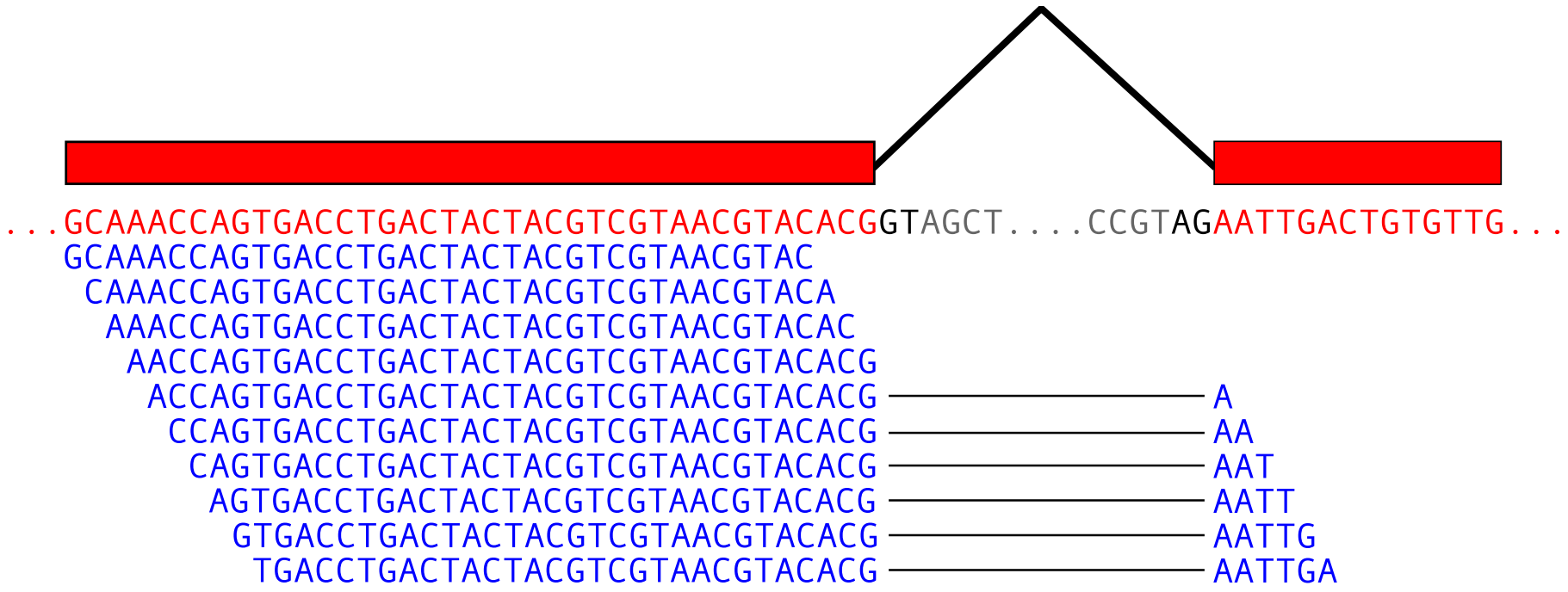

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

# Spliced vs. Unspliced Alignments




- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

# Spliced vs. Unspliced Alignments




- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

# Read Alignment (GenomeMapper/QPALMA)



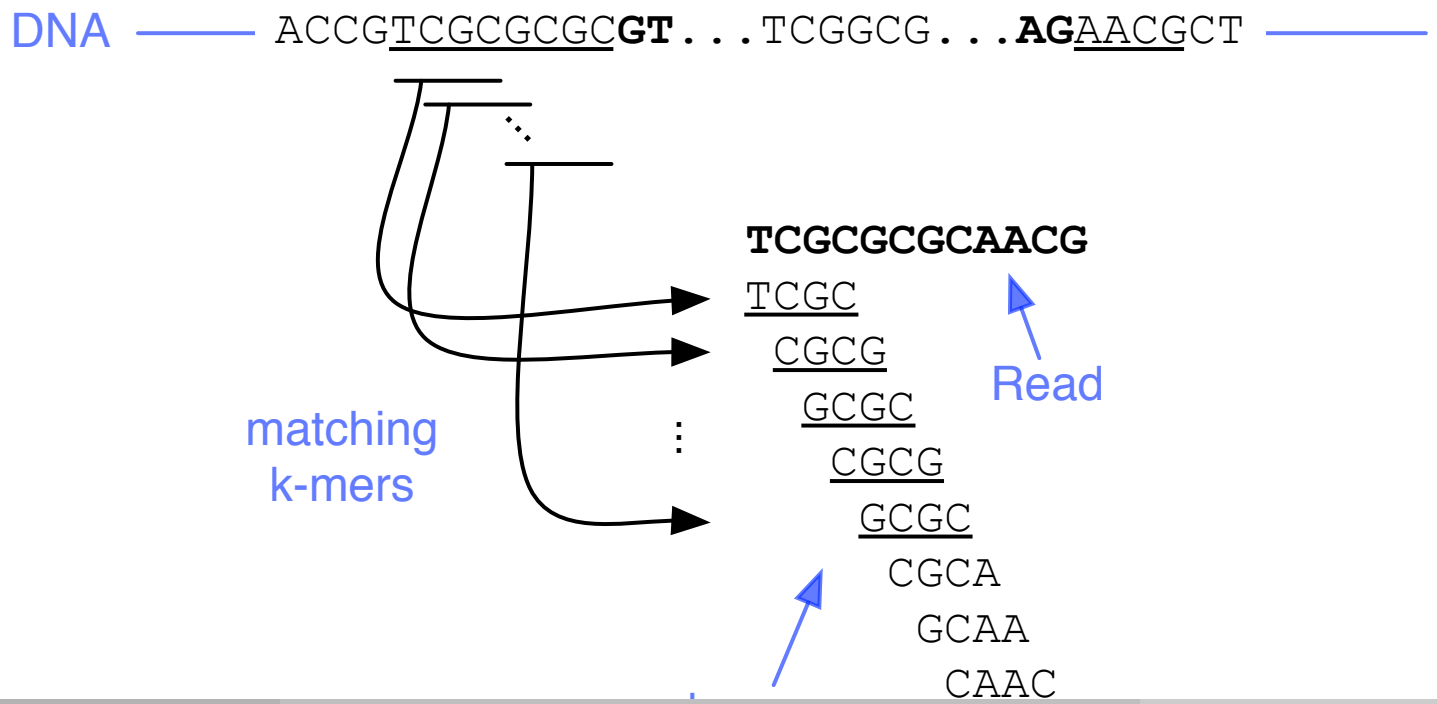

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- $k$-mer based index, well suited for smaller genomes

# Read Alignment (GenomeMapper/QPALMA)

GenomeMapper for (unspliced) read mapping:

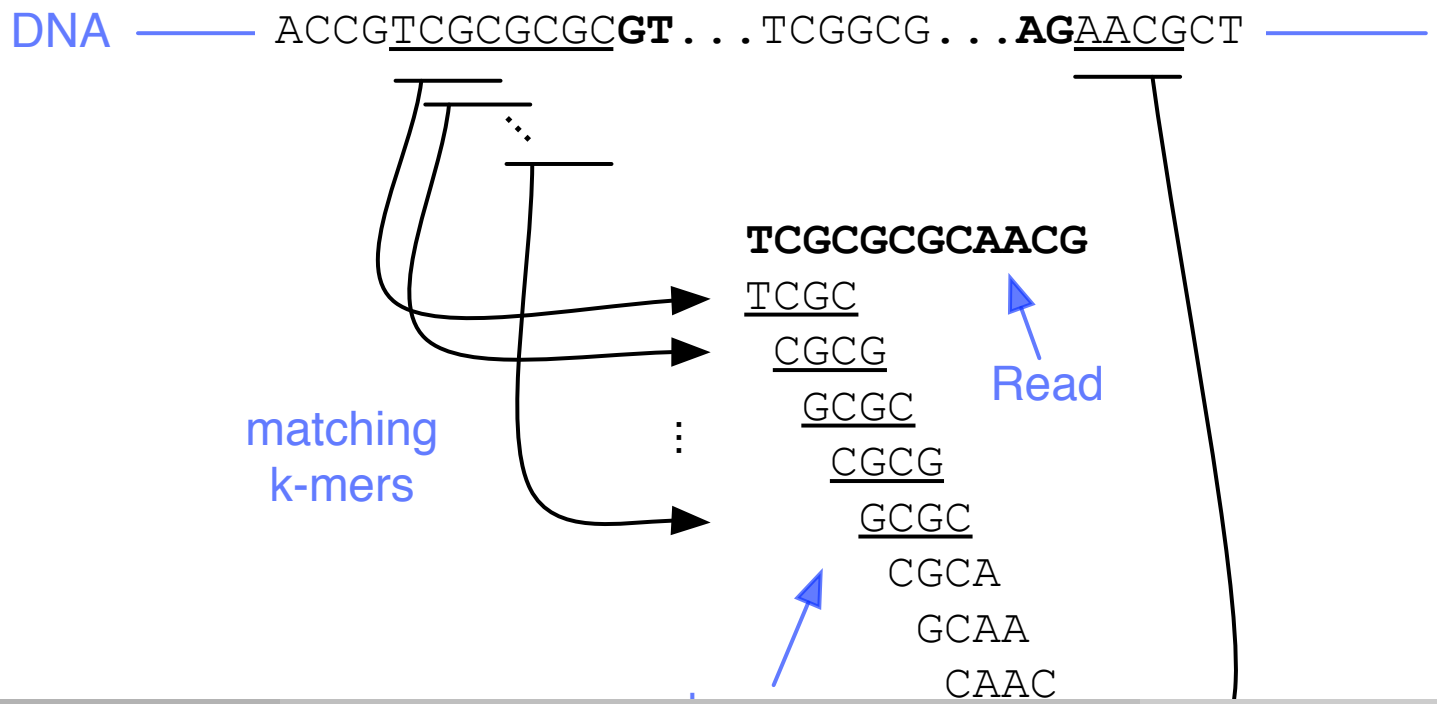
- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes



# Read Alignment (GenomeMapper/QPALMA)

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes



# Read Alignment (GenomeMapper/QPALMA)




GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:

- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA [De Bona et al., 2008]

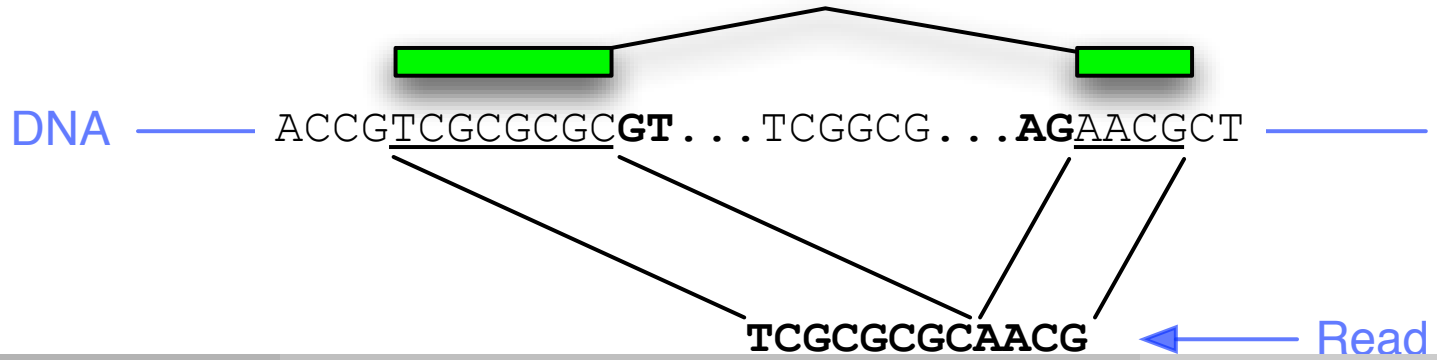
# Read Alignment (GenomeMapper/QPALMA)

GenomeMapper for (unspliced) read mapping:

- Alignments based on GenomeMapper developed in Tübingen for the 1001 plant genome project [Schneeberger et al., 2009]
- *k*-mer based index, well suited for smaller genomes

QPALMA for spliced read alignments:

- GenomeMapper identifies *seed regions*
- *Spliced alignments* by QPALMA [De Bona et al., 2008]



# QPALMA: Extended Smith-Waterman Alg.



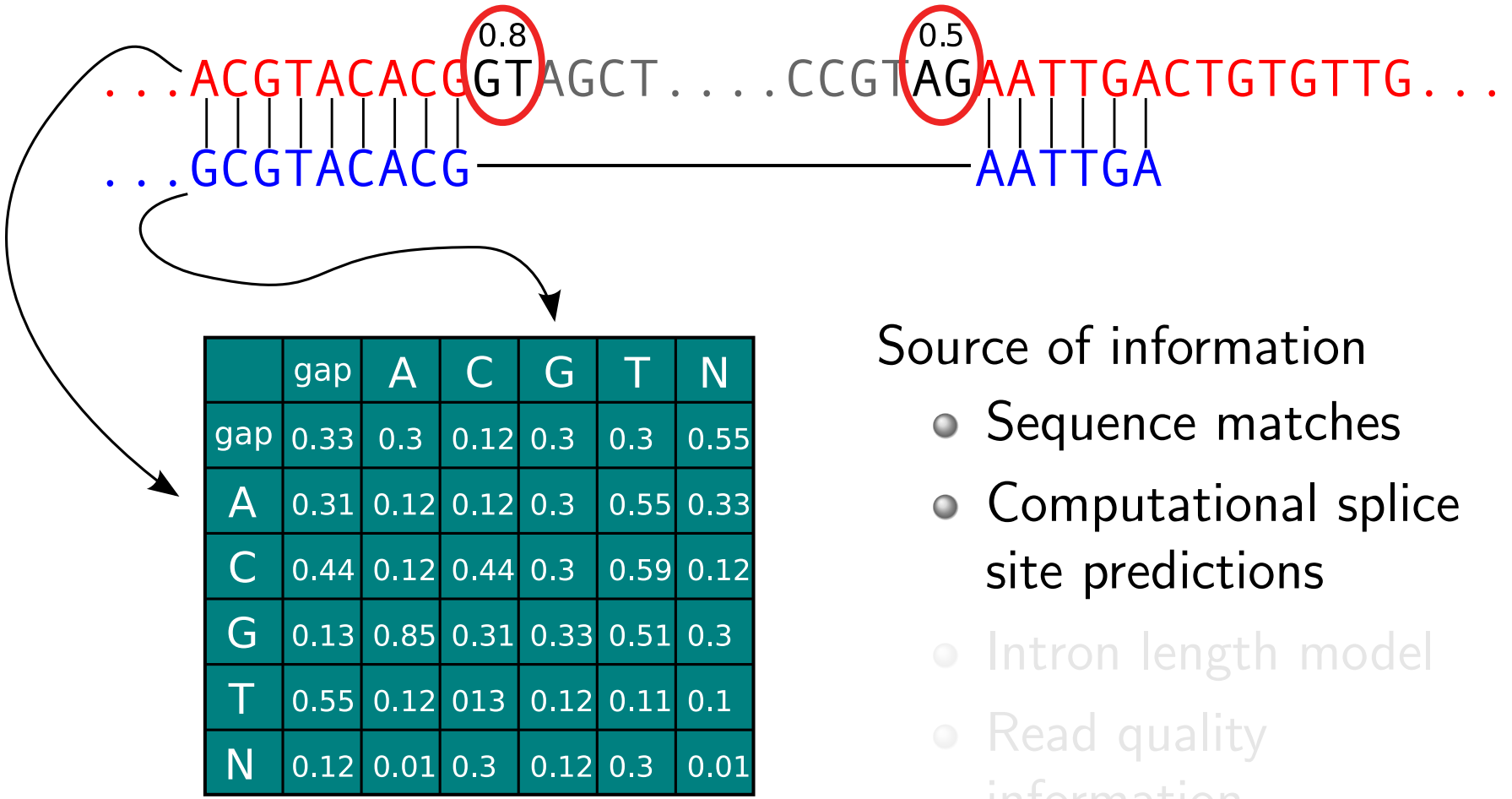

| | gap | A | C | G | T | N |
|---|---|---|---|---|---|---|
| gap | 0.33 | 0.3 | 0.12 | 0.3 | 0.3 | 0.55 |
| A | 0.31 | 0.12 | 0.12 | 0.3 | 0.55 | 0.33 |
| C | 0.44 | 0.12 | 0.44 | 0.3 | 0.59 | 0.12 |
| G | 0.13 | 0.85 | 0.31 | 0.33 | 0.51 | 0.3 |
| T | 0.55 | 0.12 | 013 | 0.12 | 0.11 | 0.1 |
| N | 0.12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

Source of information

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

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

# QPALMA: Extended Smith-Waterman Alg.



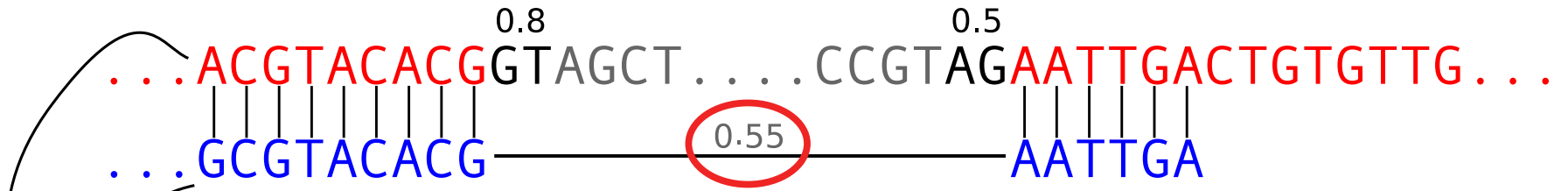

| | gap | A | C | G | T | N |
|---|---|---|---|---|---|---|
| gap | 0.33 | 0.3 | 0.12 | 0.3 | 0.3 | 0.55 |
| A | 0.31 | 0.12 | 0.12 | 0.3 | 0.55 | 0.33 |
| C | 0.44 | 0.12 | 0.44 | 0.3 | 0.59 | 0.12 |
| G | 0.13 | 0.85 | 0.31 | 0.33 | 0.51 | 0.3 |
| T | 0.55 | 0.12 | 013 | 0.12 | 0.11 | 0.1 |
| N | 0.12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

Source of information

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

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

# QPALMA: Extended Smith-Waterman Alg.



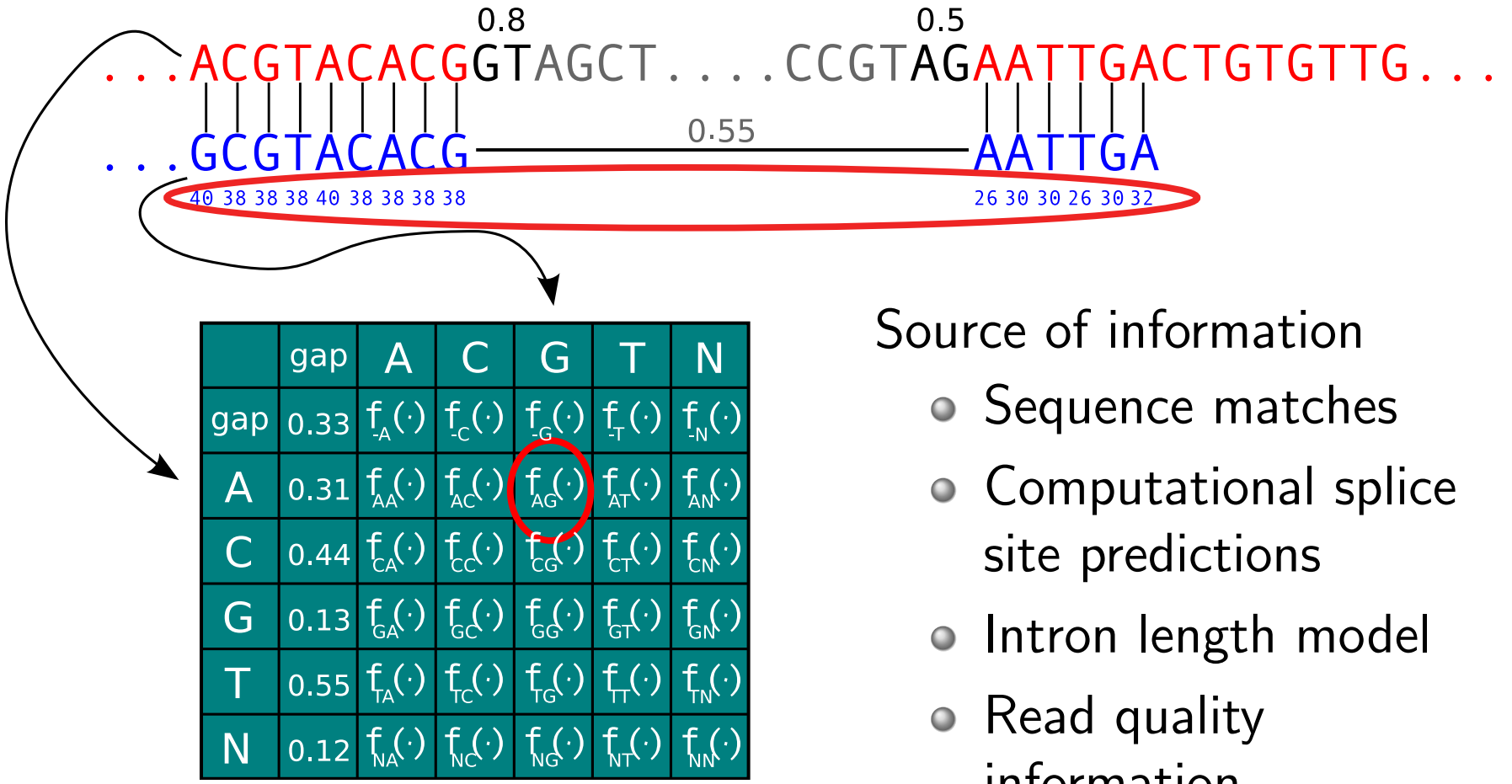

| | gap | A | C | G | T | N |
|---|---|---|---|---|---|---|
| gap | 0.33 | 0.3 | 0.12 | 0.3 | 0.3 | 0.55 |
| A | 0.31 | 0.12 | 0.12 | 0.3 | 0.55 | 0.33 |
| C | 0.44 | 0.12 | 0.44 | 0.3 | 0.59 | 0.12 |
| G | 0.13 | 0.85 | 0.31 | 0.33 | 0.51 | 0.3 |
| T | 0.55 | 0.12 | 013 | 0.12 | 0.11 | 0.1 |
| N | 0.12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

Source of information

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

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

# QPALMA: Extended Smith-Waterman Alg.




Source of information

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

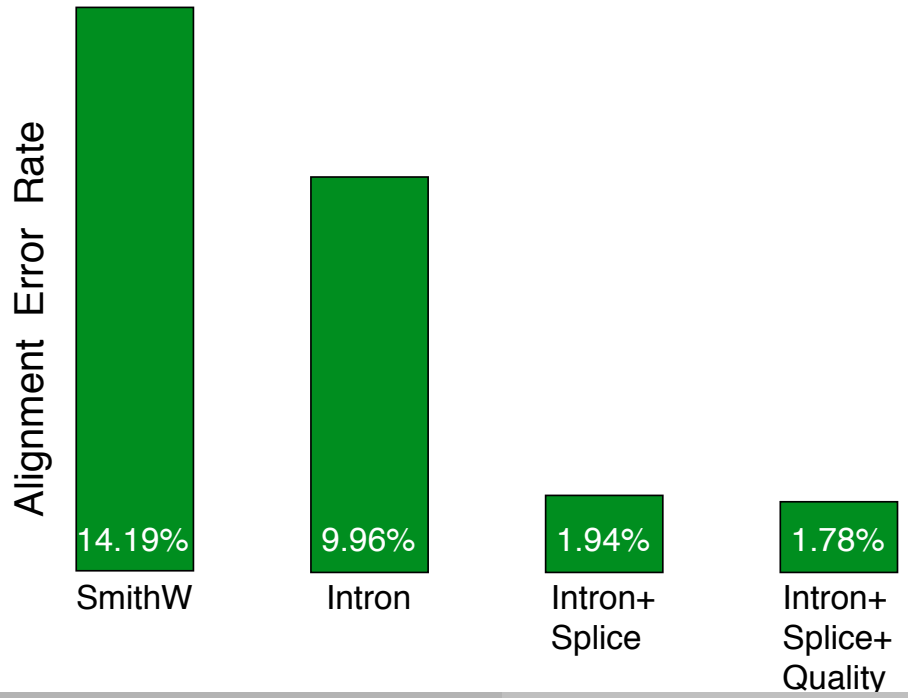
Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

[De Bona et al., 2008]

# RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing

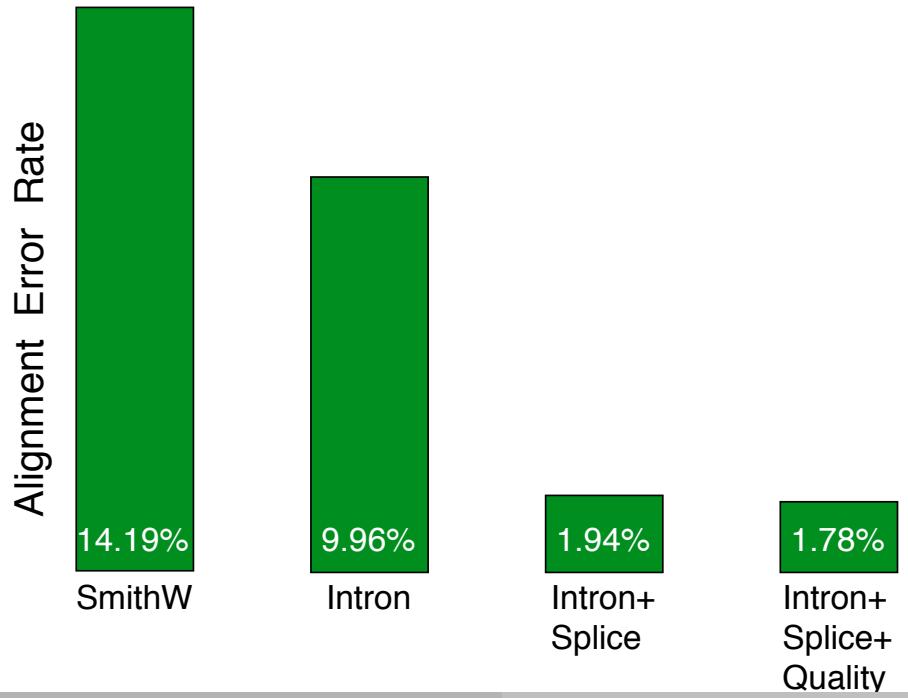



[De Bona et al., 2008]

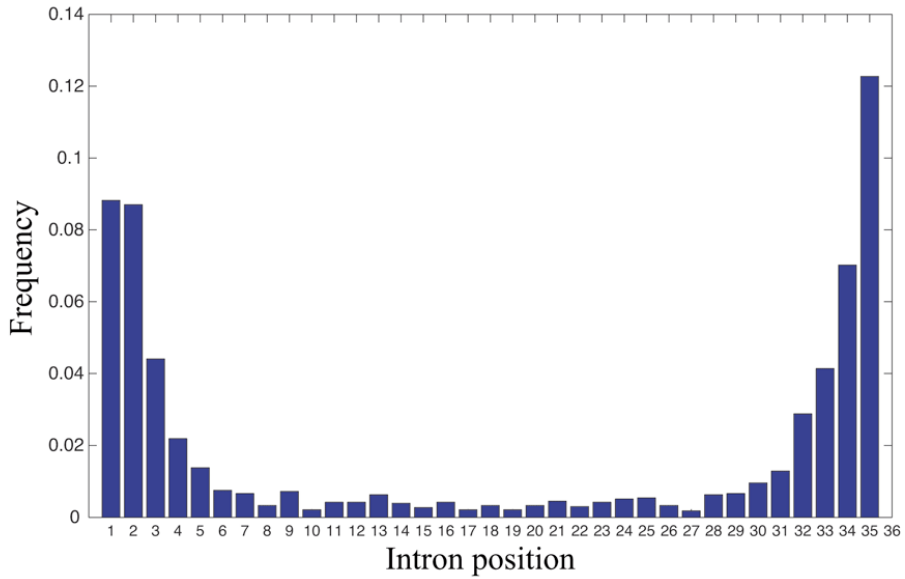
# RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing




Error vs. intron position

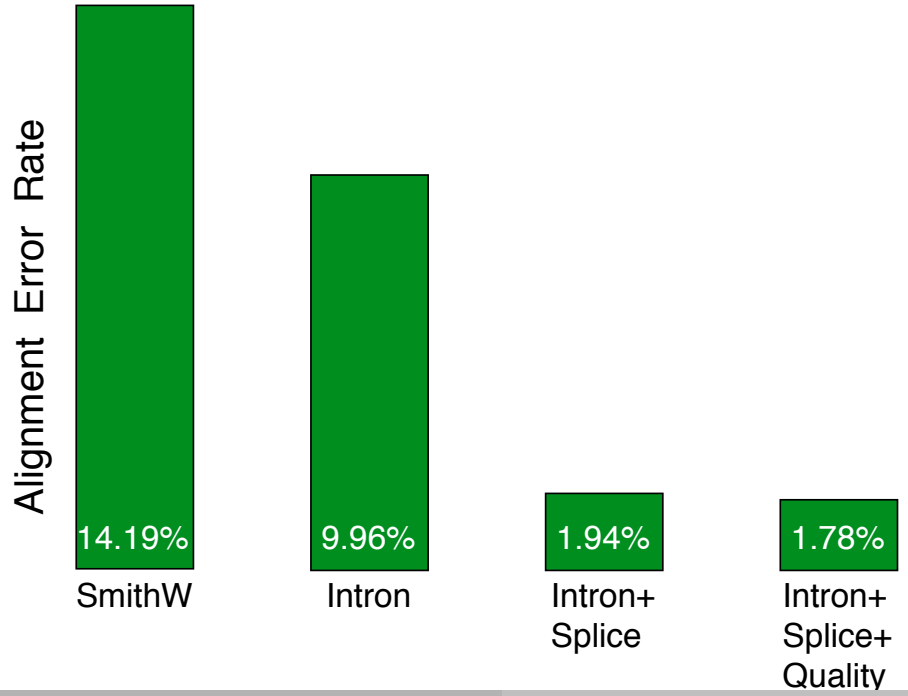


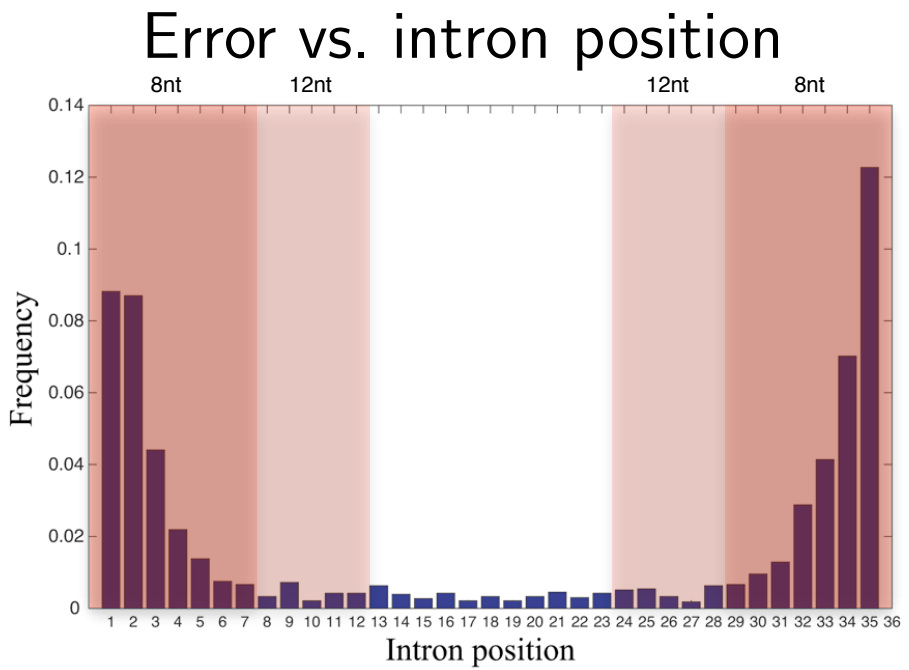

[De Bona et al., 2008]

# RNA-Seq Read Alignment with QPALMA

Generate set of artificially spliced reads

- Genomic reads with quality information
- Genome annotation for artificially splicing the reads
- Use 10,000 reads for training and 30,000 for testing




[De Bona et al., 2008]

# Transcriptome Studies in Human

nature Vol 456|27 November 2008|doi:10.1038/nature07509

ARTICLES

## Alternative isoform regulation in human tissue transcriptomes

Eric T. Wang[1,2]*, Rickard Sandberg[1,3]*, Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4], Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4] & Christopher B. Burge[1]

[Wang et al., 2008]

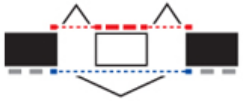
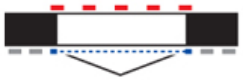
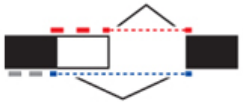
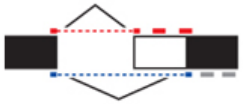
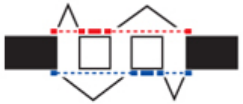
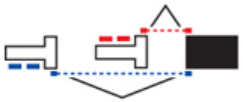
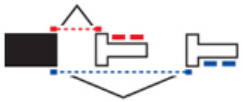
15 AUGUST 2008 VOL 321 SCIENCE www.sciencemag.org

## A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome

Marc Sultan,[1]* Marcel H. Schulz,[2,3]* Hugues Richard,[2]* Alon Magen,[1] Andreas Klingenhoff,[4] Matthias Scherf,[4] Martin Seifert,[4] Tatjana Borodina,[1]

tan et al., 2008]

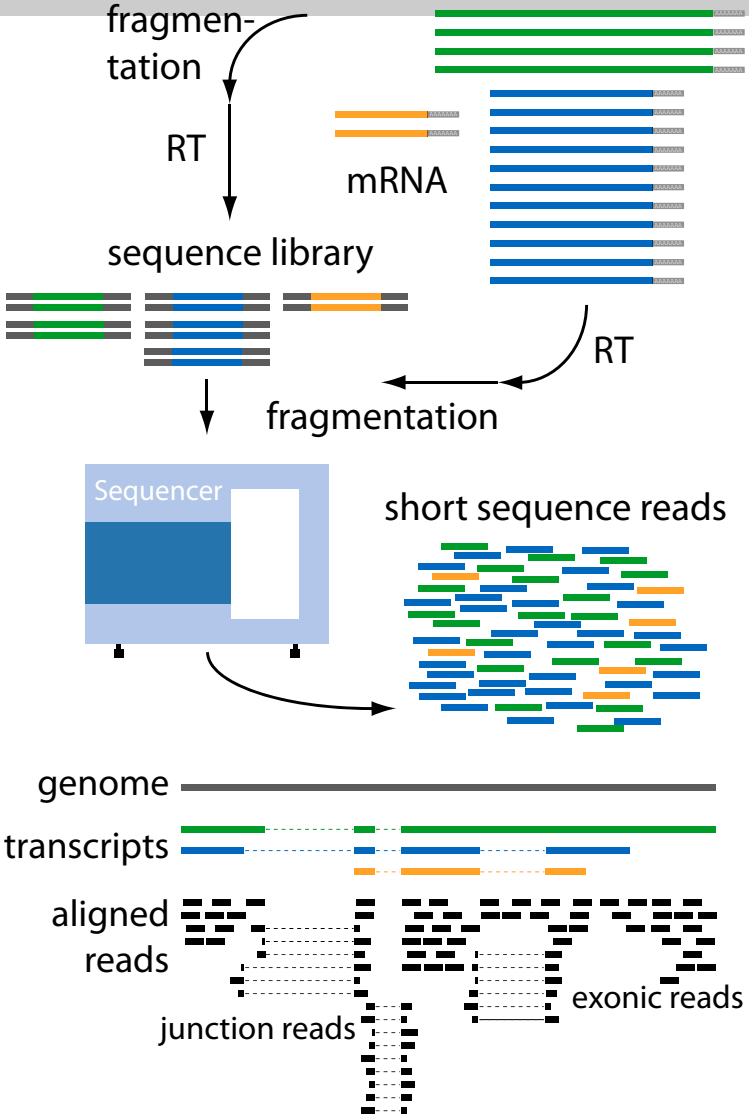
# Transcriptome Studies in Human

| Alternative transcript events | | Total events ($\times 10^3$) | Number detected ($\times 10^3$) | Both isoforms detected | Number tissue-regulated | % Tissue-regulated (observed) | % Tissue-regulated (estimated) |
|---|---|---|---|---|---|---|---|
| Skipped exon |  | 37 | 35 | 10,436 | 6,822 | 65 | 72 |
| Retained intron |  | 1 | 1 | 167 | 96 | 57 | 71 |
| Alternative 5′ splice site (A5SS) |  | 15 | 15 | 2,168 | 1,386 | 64 | 72 |
| Alternative 3′ splice site (A3SS) |  | 17 | 16 | 4,181 | 2,655 | 64 | 74 |
| Mutually exclusive exon (MXE) |  | 4 | 4 | 167 | 95 | 57 | 66 |
| Alternative first exon (AFE) |  | 14 | 13 | 10,281 | 5,311 | 52 | 63 |
| Alternative last exon (ALE) |  | 9 | 8 | 5,246 | 2,491 | 47 | 52 |
| Tandem 3′ UTRs |   | 7 | 7 | 5,136 | 3,801 | 74 | 80 |
| Total | | 105 | 100 | 37,782 | 22,657 | 60 | 68 |

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

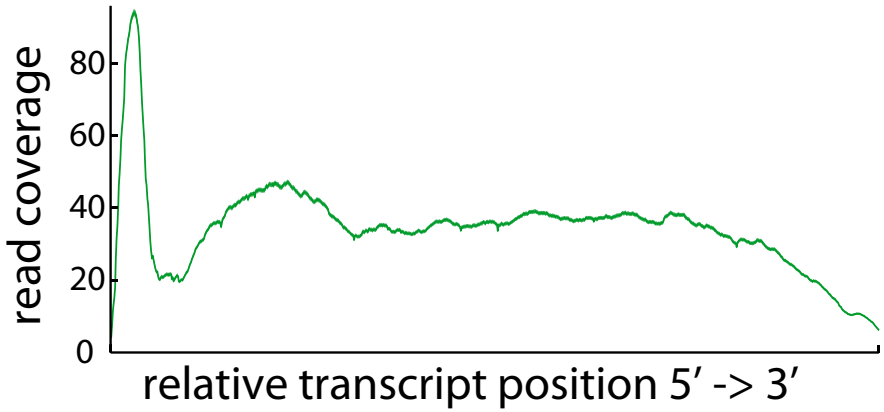
[Wang et al., 2008]

# RNA-Seq Biases and Quantitation



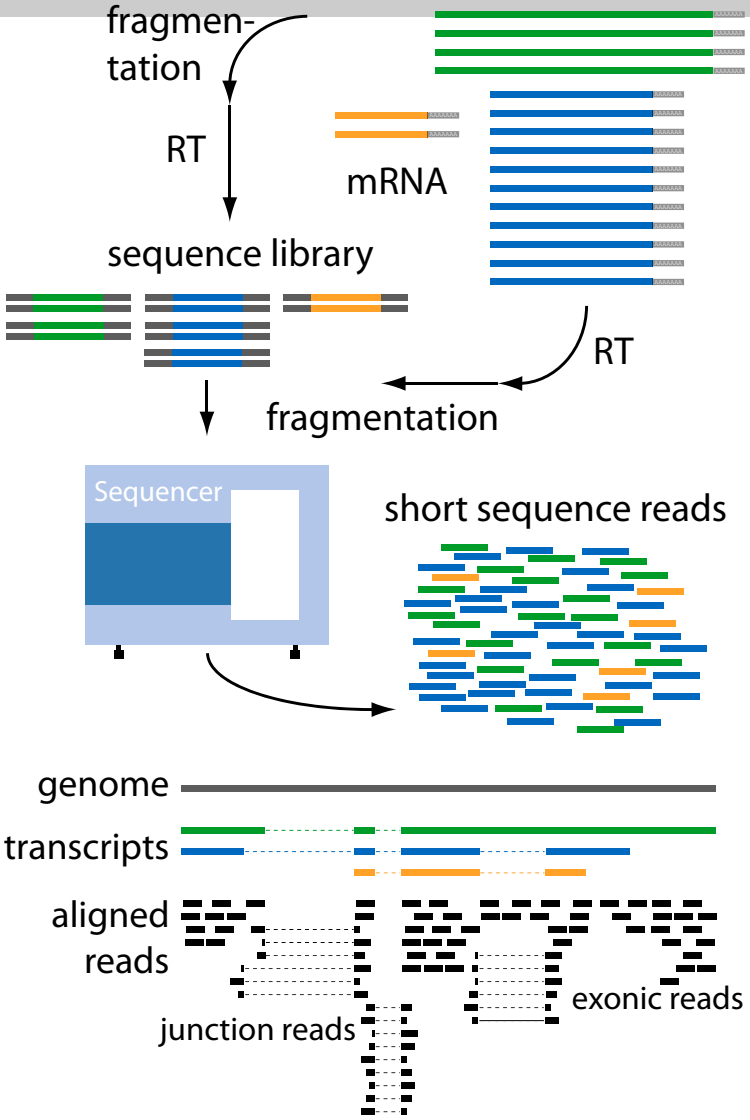

Biases due to ...

- cDNA library construction
- Sequencing
- Read mapping



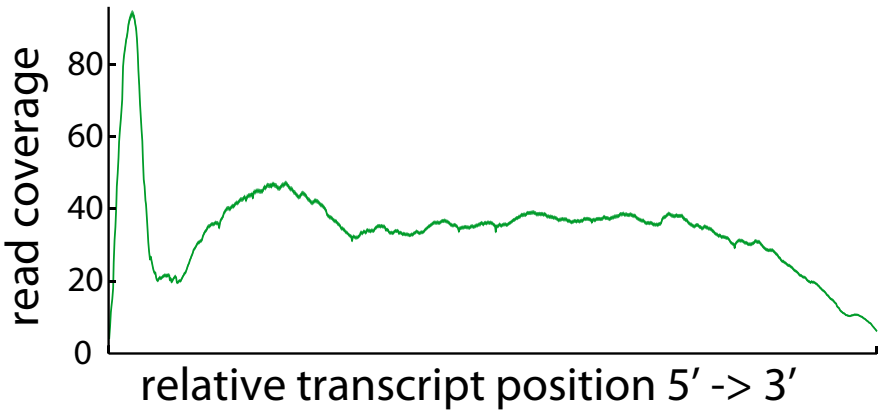

(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

# RNA-Seq Biases and Quantitation



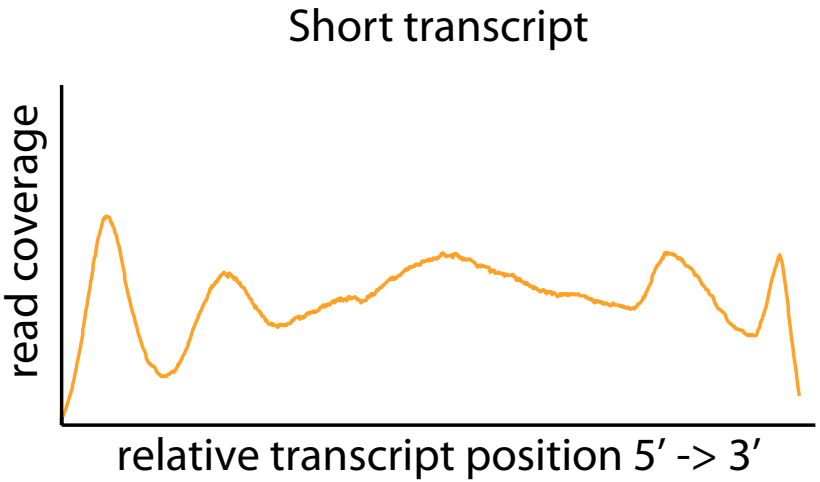

Biases due to ...

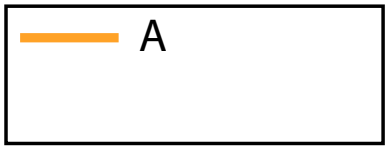
- cDNA library construction
- Sequencing
- Read mapping



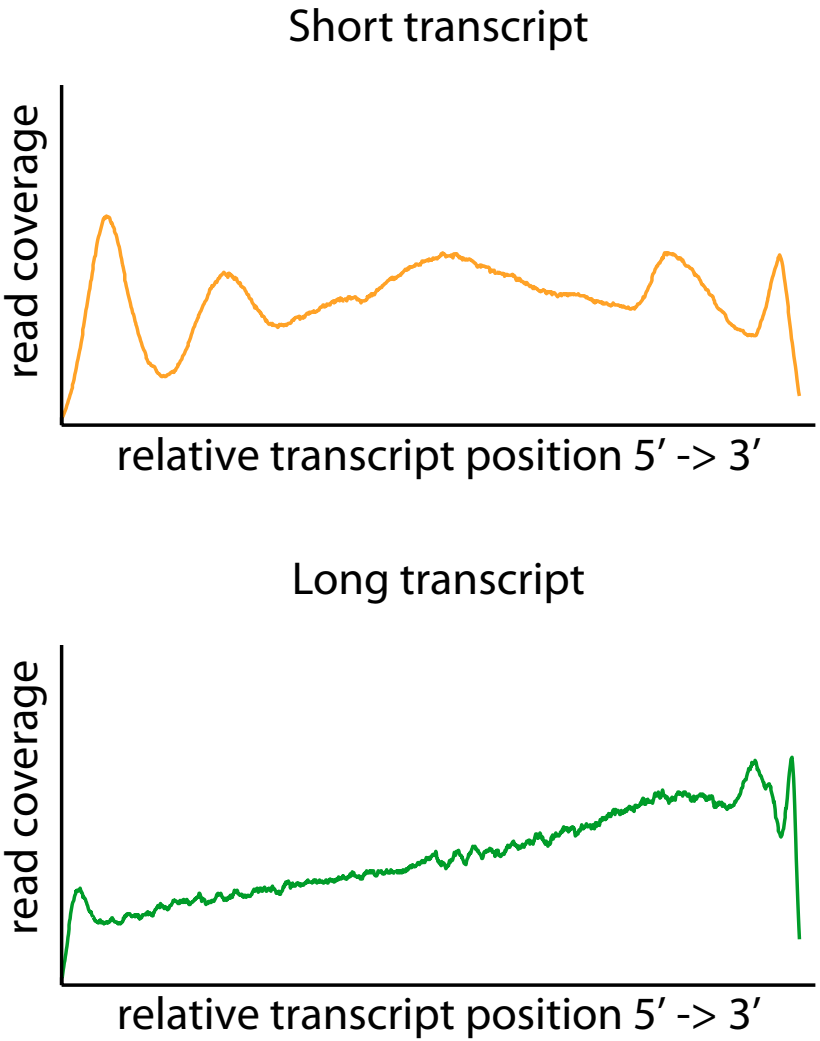

(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

# rQuant – Basic Idea

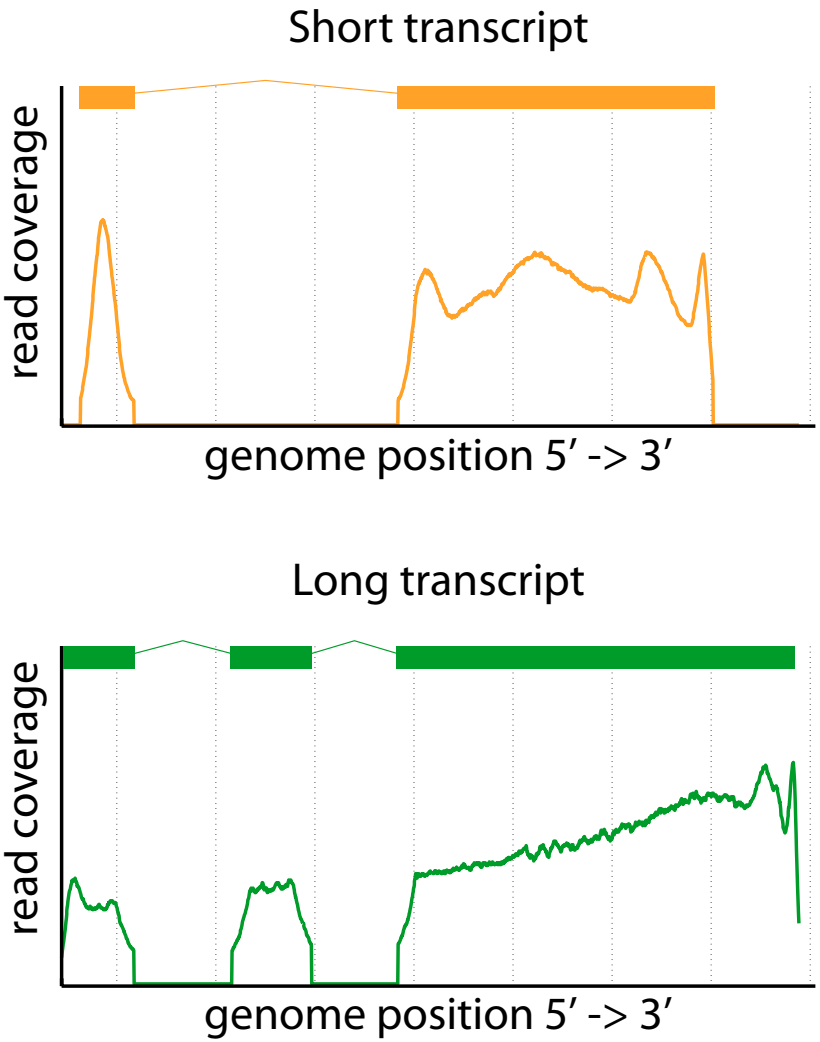


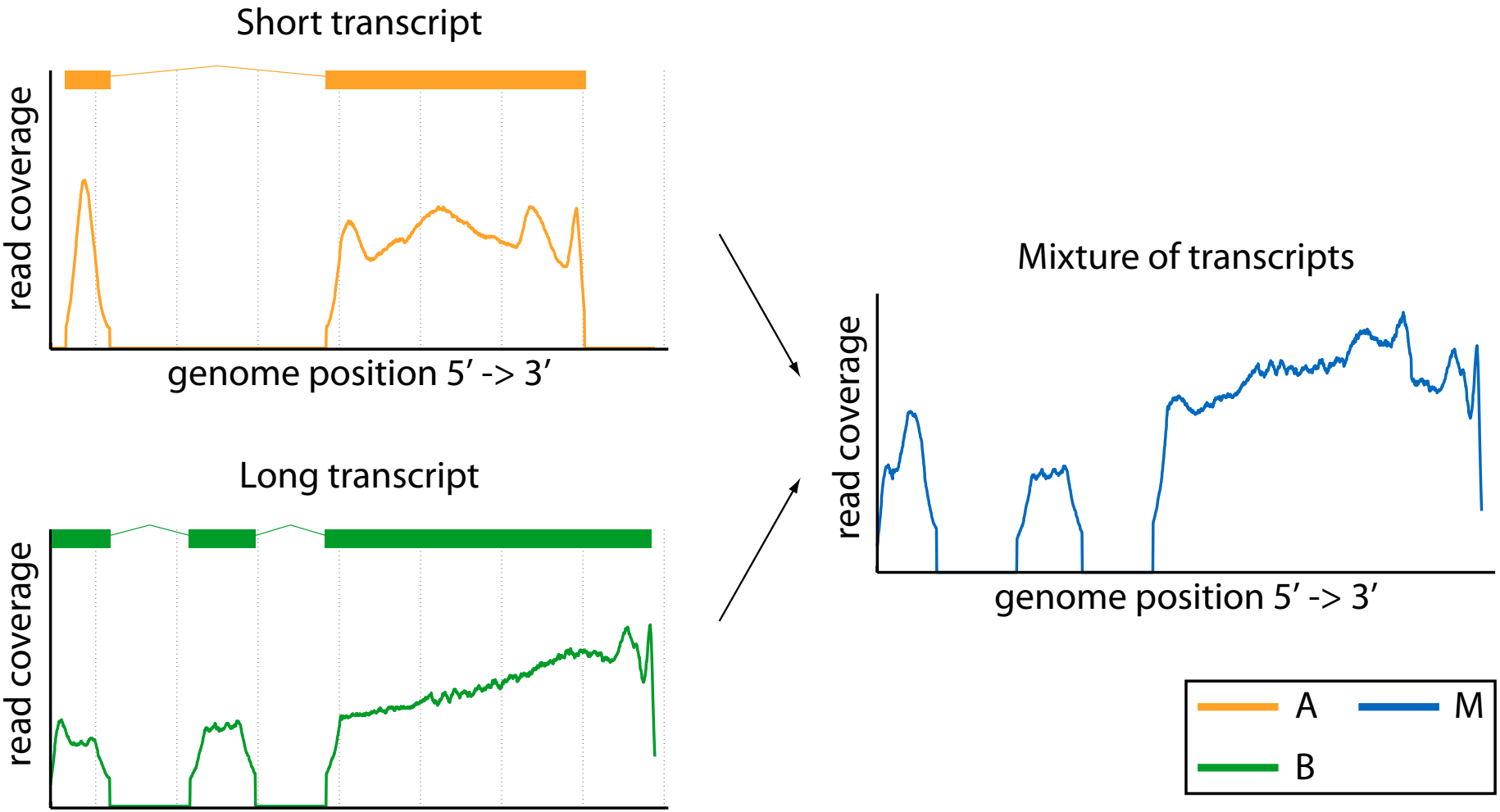
# rQuant – Basic Idea



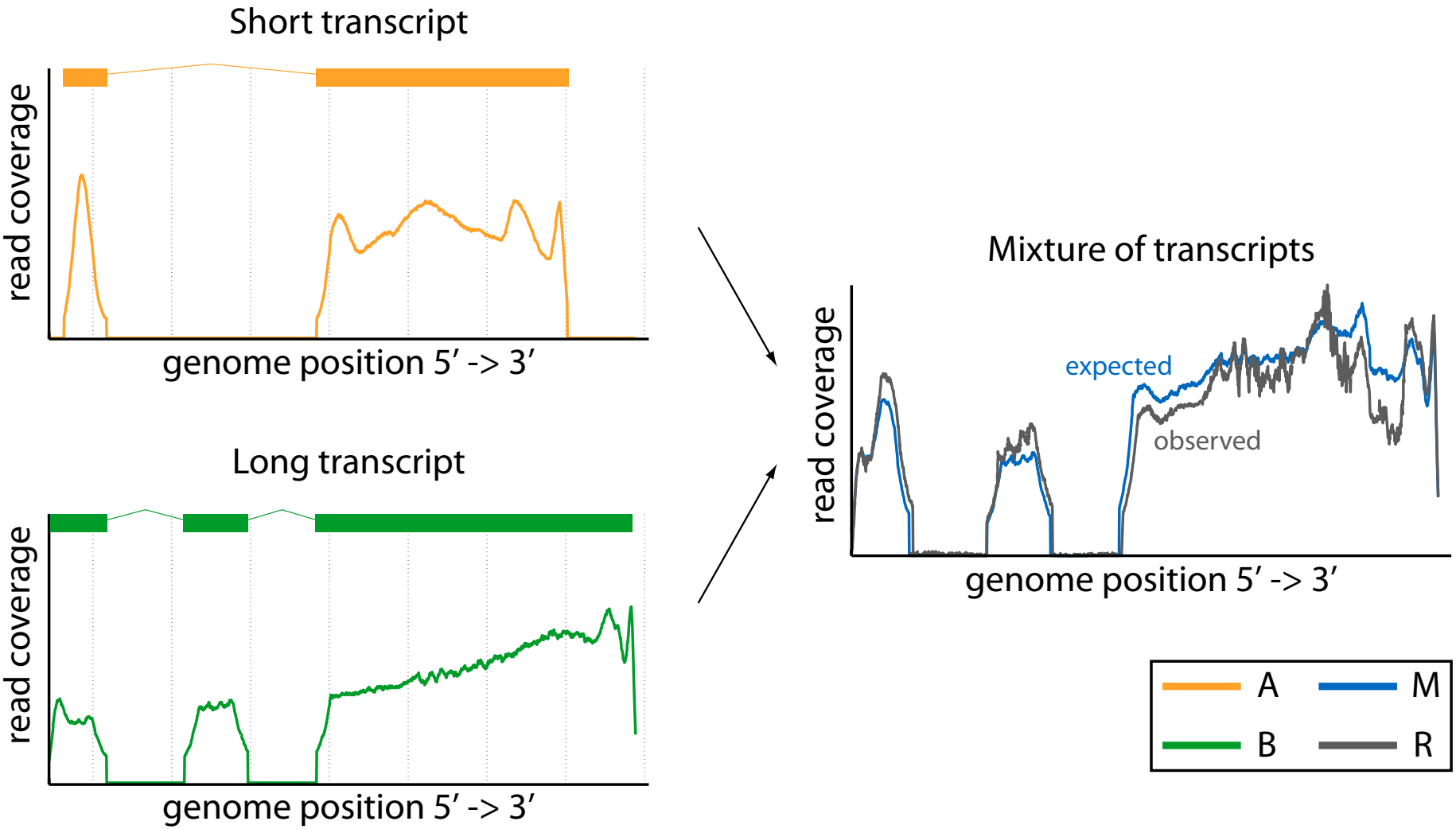
# rQuant – Basic Idea



# rQuant – Basic Idea



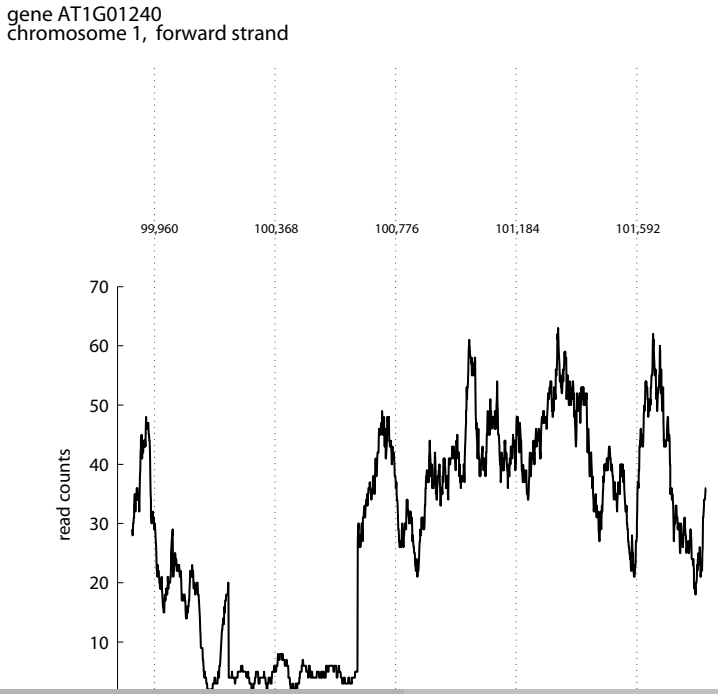
# rQuant – Basic Idea




$$M = w_A A + w_B B \qquad \min \sum \ell(M, R)$$

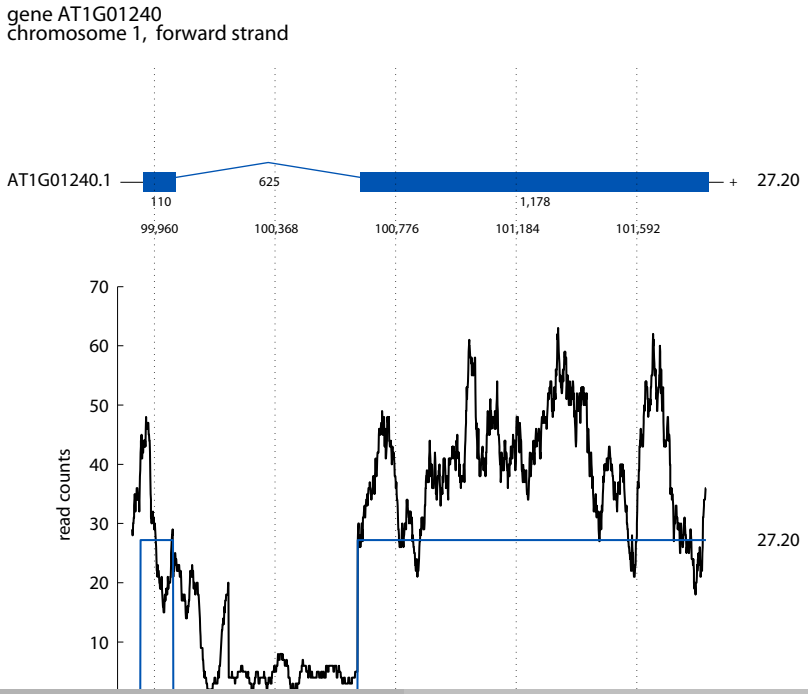
# rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



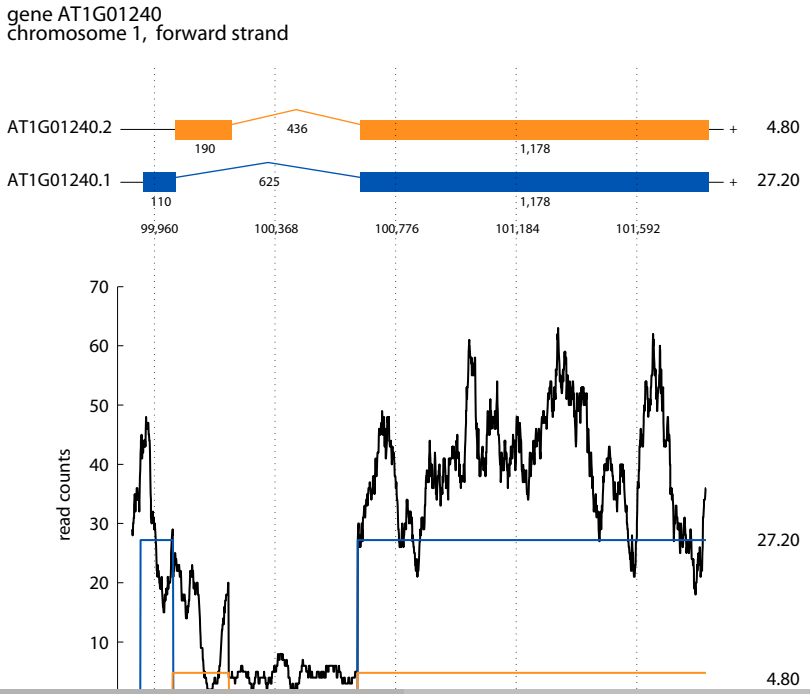
# rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



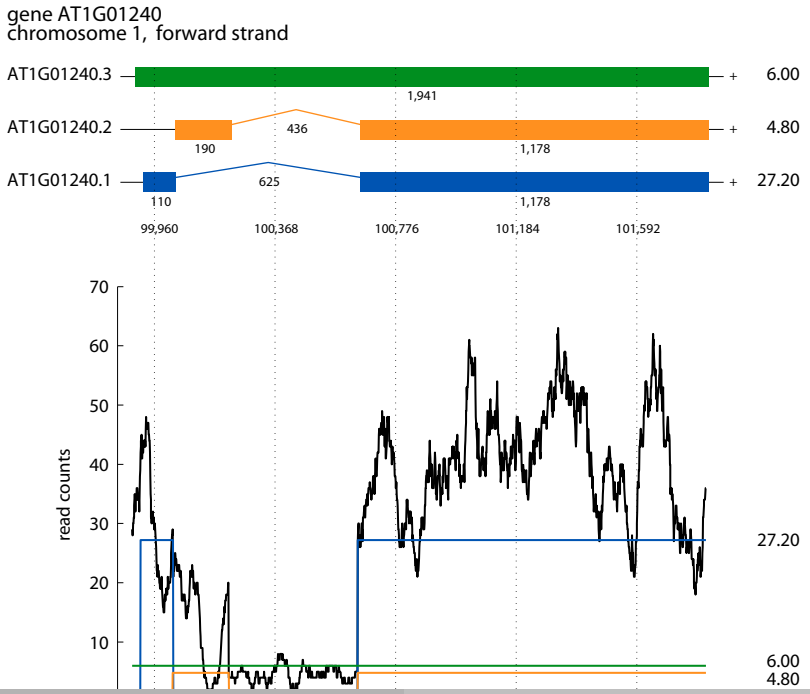
# rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



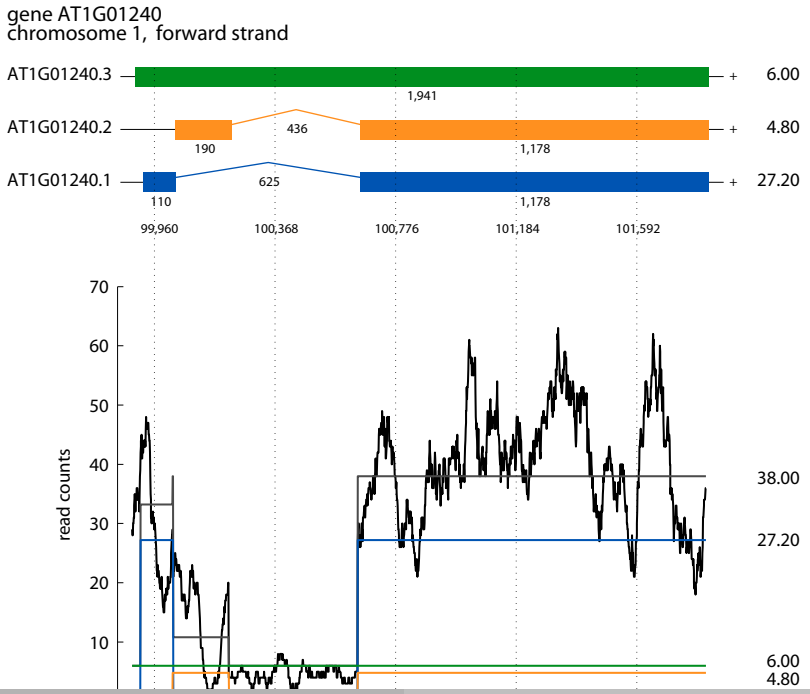
# rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



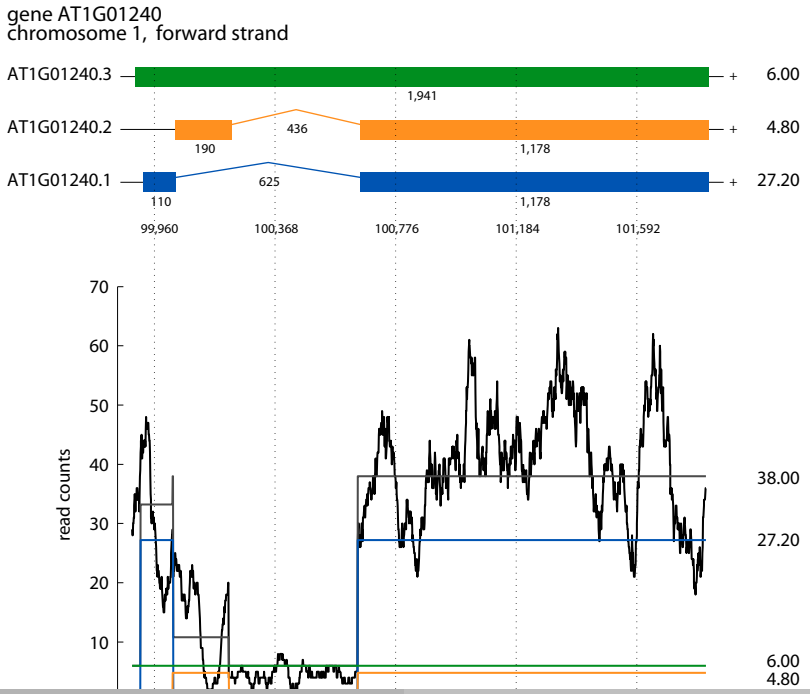
# rQuant – Iterative Algorithm

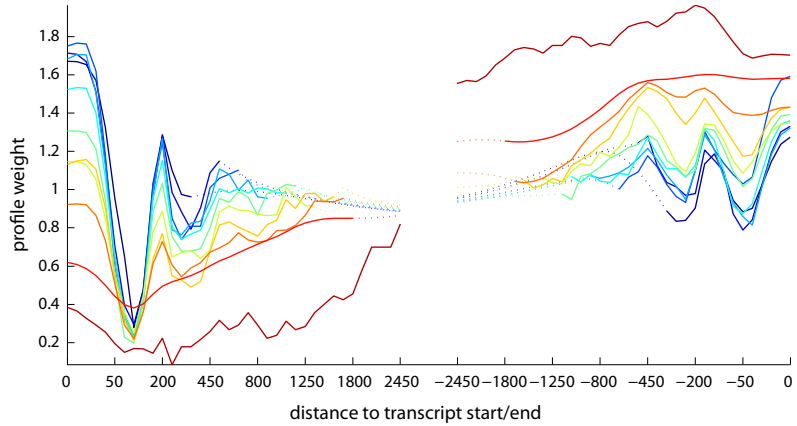
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



# rQuant – Iterative Algorithm

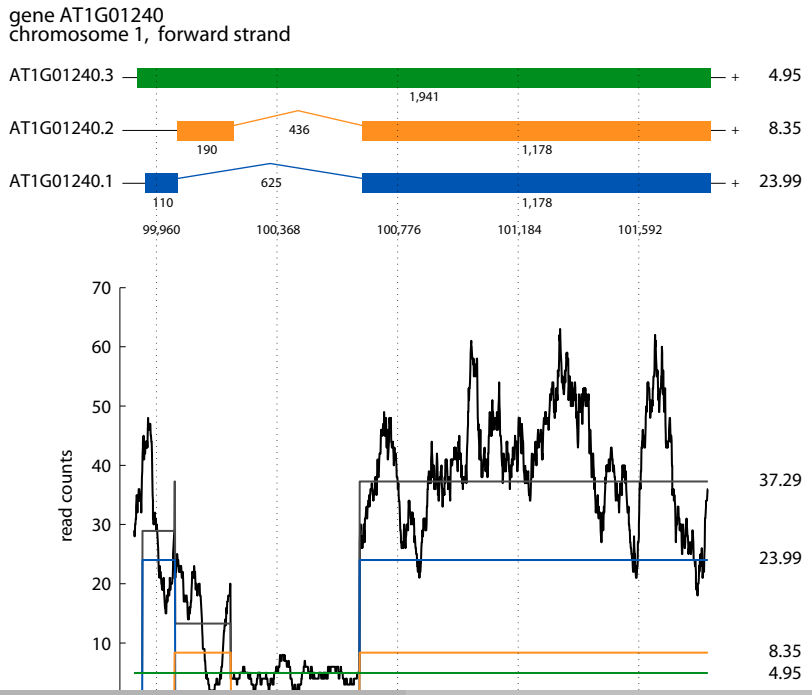
1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.

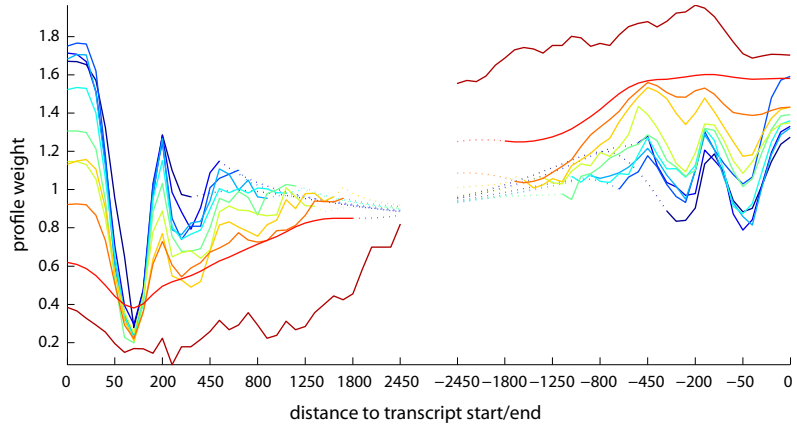


# rQuant – Iterative Algorithm

1. Optimise transcript weights: $\min_{\mathbf{w}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
2. Optimise profile weights: $\min_{\mathbf{p}} \sum_i \ell\left(\sum_t w^{(t)} p_i^{(t)}, R_i\right)$
3. Repeat 1. and 2. until convergence.



# RQuant Evaluation I




## Mmethod rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009] )
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

# RQuant Evaluation I




Mmethod rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009] )
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

# RQuant Evaluation I




Mmethod rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009] )
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

# RQuant Evaluation I




Mmethod rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009] )
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

# RQuant Evaluation I




Mmethod rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009] )
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

- Simulated RNA expression level and
- Predicted transcript weights

# RQuant Evaluation I




Mmethod rQuant: Position-wise with profiles

(estimating library and mapping bias)

compared to

- Position-wise, without profiles
- Segment-wise, without profiles (e.g., Jiang and Wong [2009] )
- Segment-wise, with profiles (e.g., Flux Capacitor [Sammeth, 2009a])

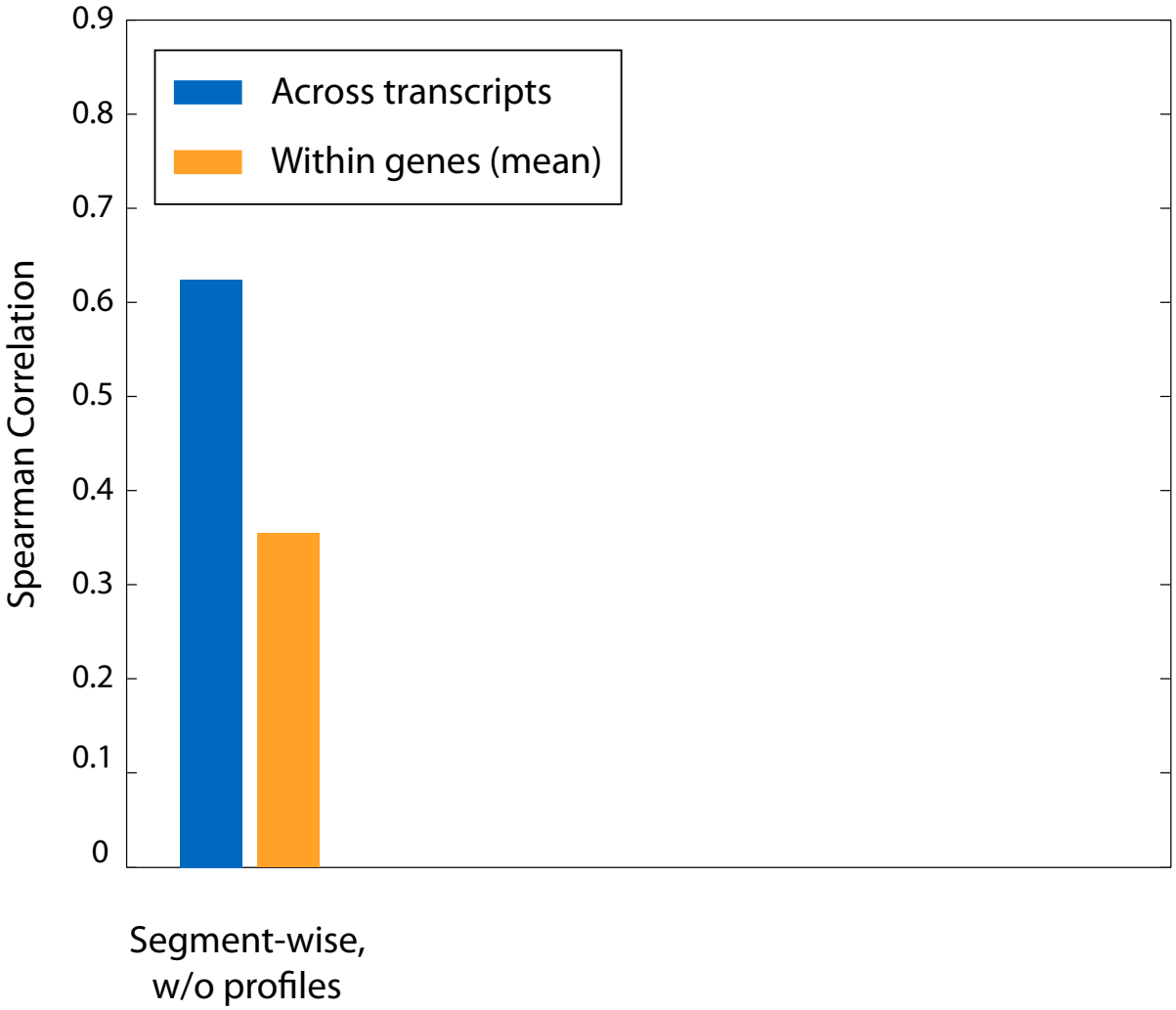
Estimate transcript abundances

- Using simulated data for *A. thaliana* (Flux Simulator [Sammeth, 2009b])
- Subset of alternatively spliced genes

Evaluation: Spearman correlation between

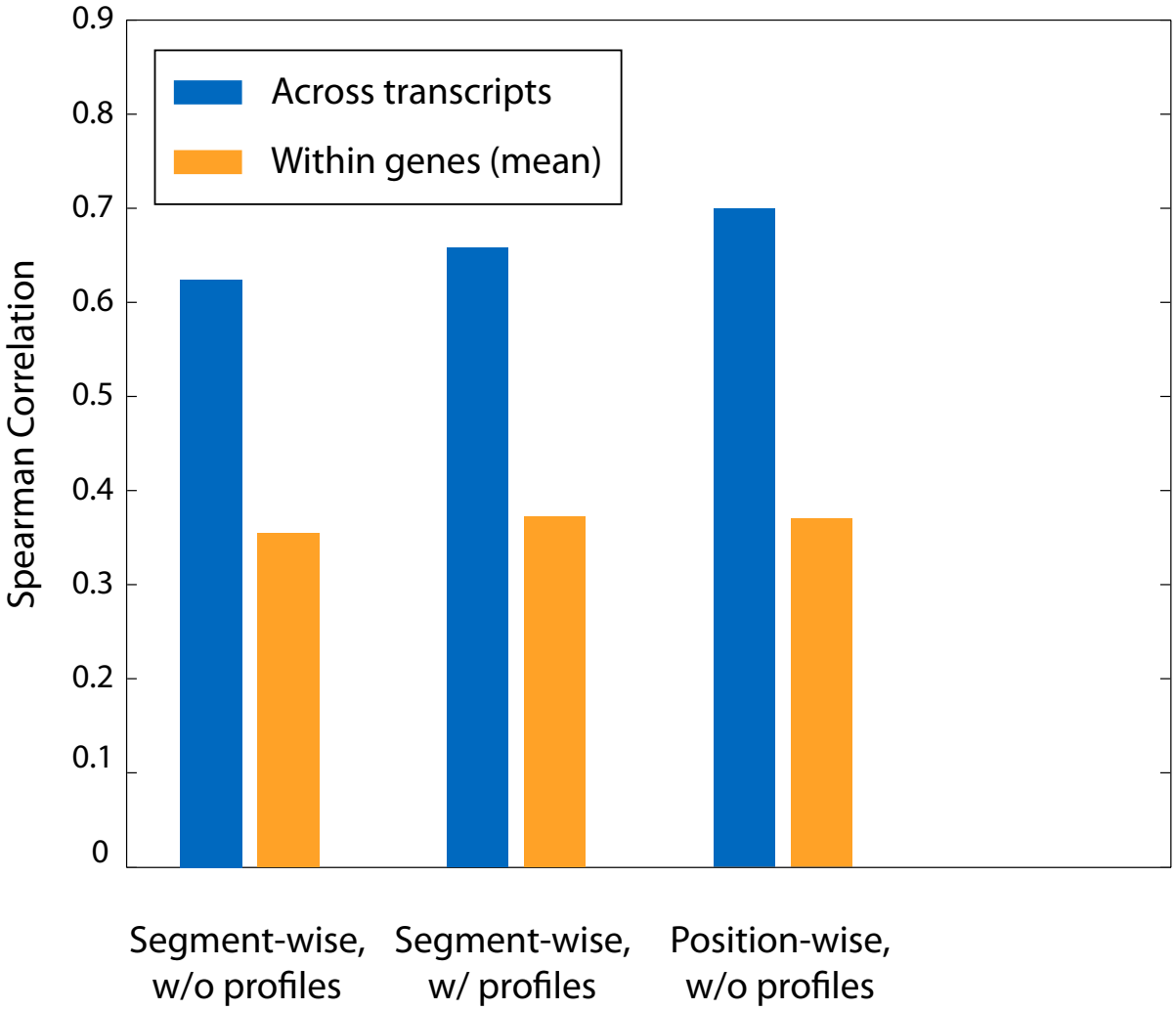
- Simulated RNA expression level and
- Predicted transcript weights

# RQuant Evaluation II



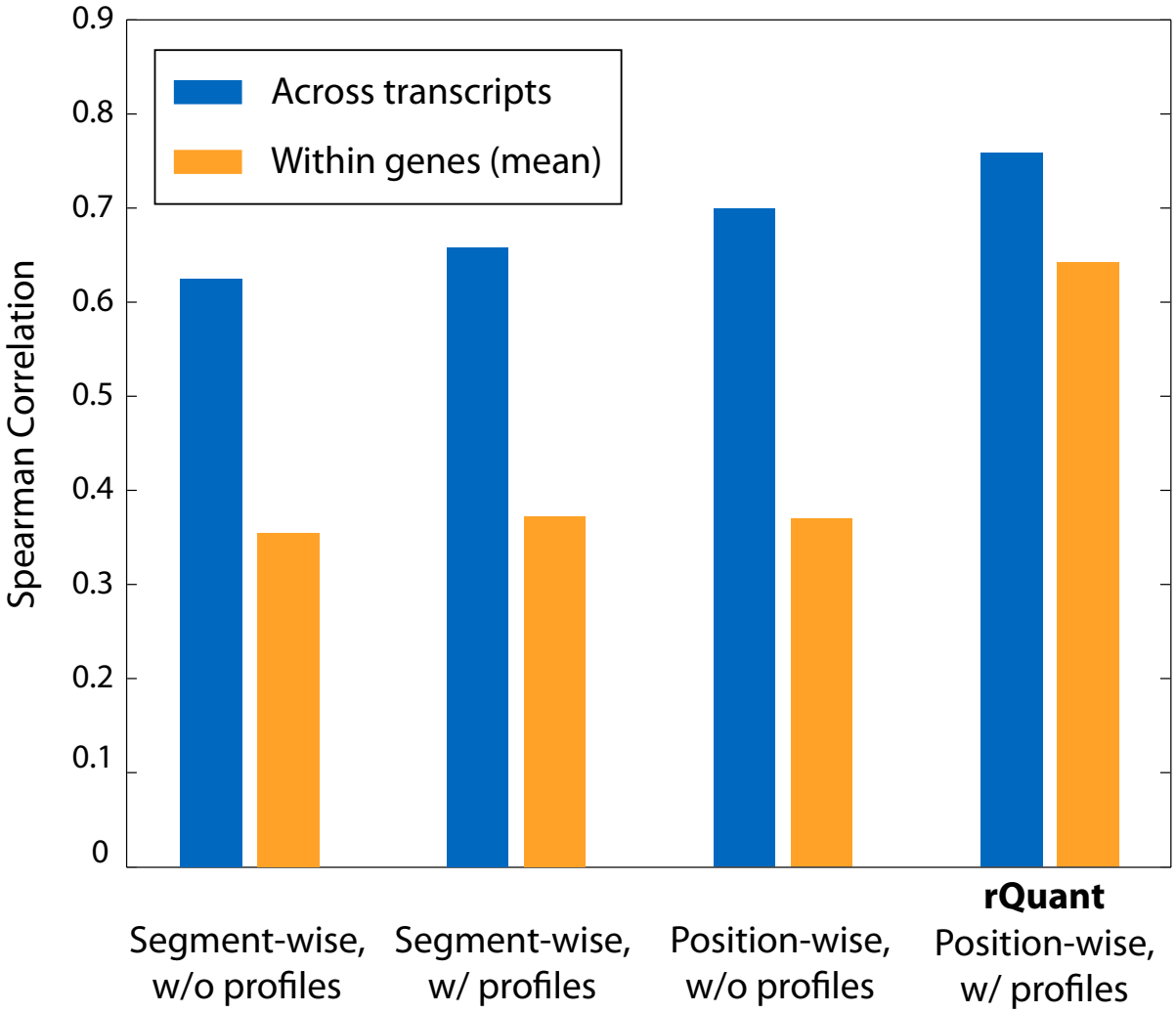

(Bohnert et al., in preparation, 2010)

# RQuant Evaluation II



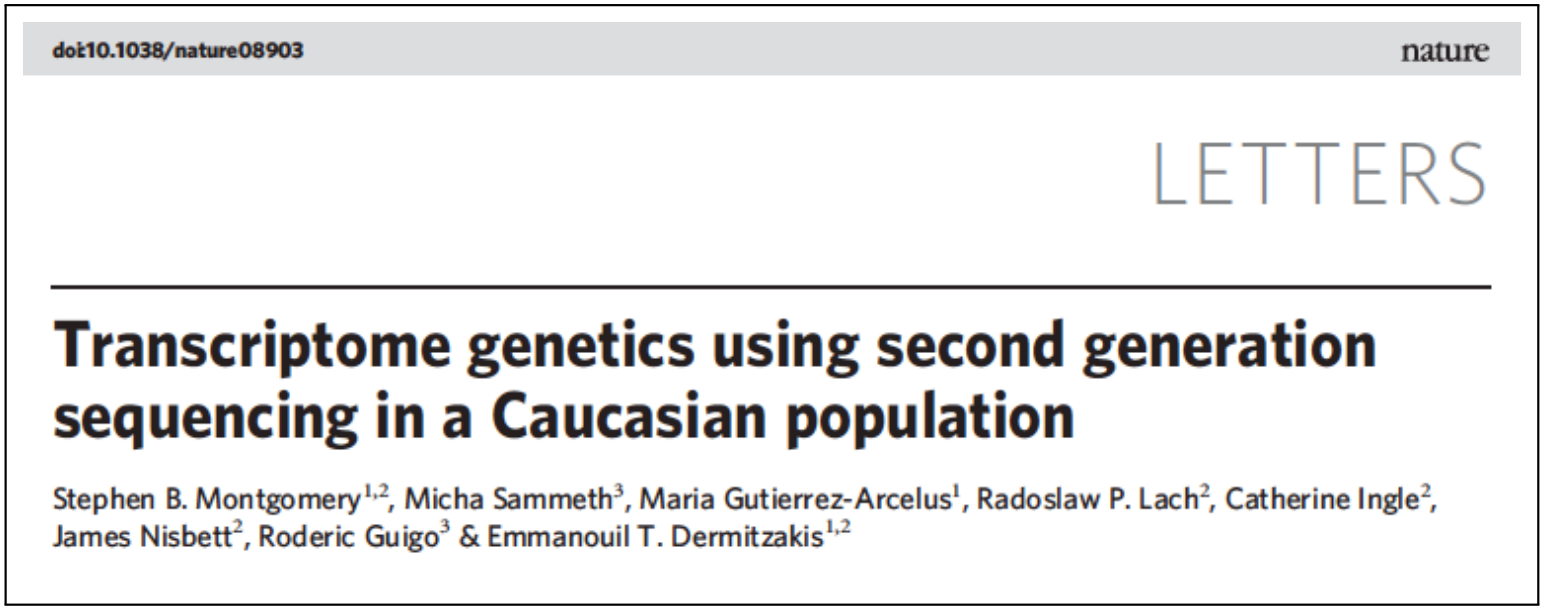

(Bohnert et al., in preparation, 2010)

# RQuant Evaluation II




(Bohnert et al., in preparation, 2010)

# Transcriptome Variation Studies



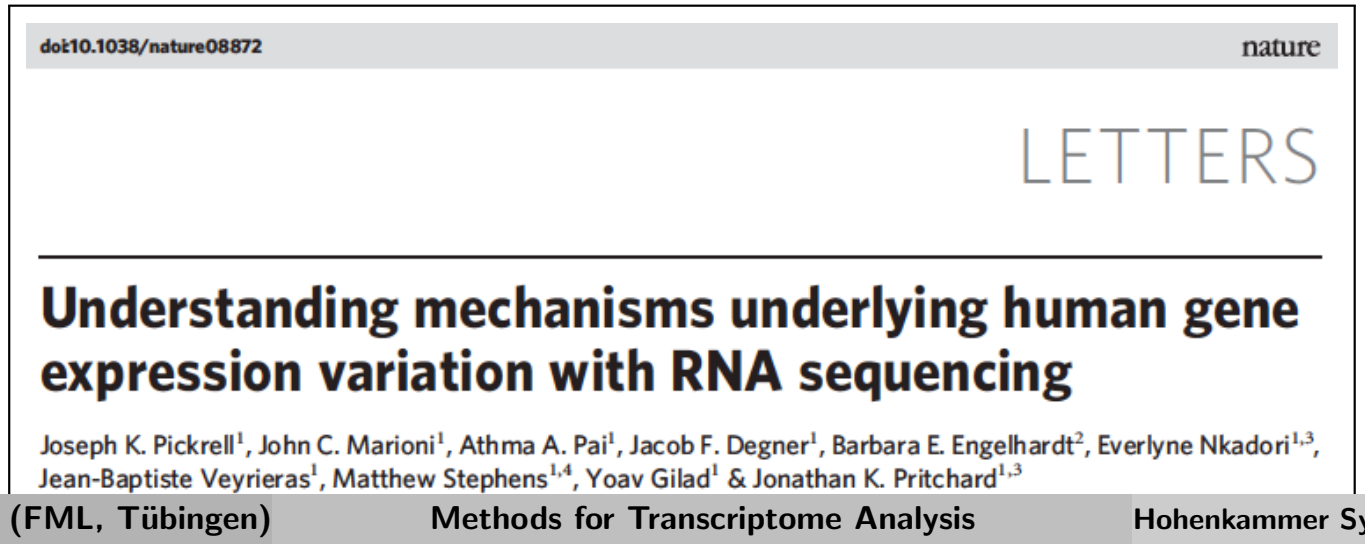
doi:10.1038/nature08903 nature

LETTERS

## Transcriptome genetics using second generation sequencing in a Caucasian population

Stephen B. Montgomery[1,2], Micha Sammeth[3], Maria Gutierrez-Arcelus[1], Radoslaw P. Lach[2], Catherine Ingle[2], James Nisbett[2], Roderic Guigo[3] & Emmanouil T. Dermitzakis[1,2]

[Montgomery et al., 2010]



doi:10.1038/nature08872 nature

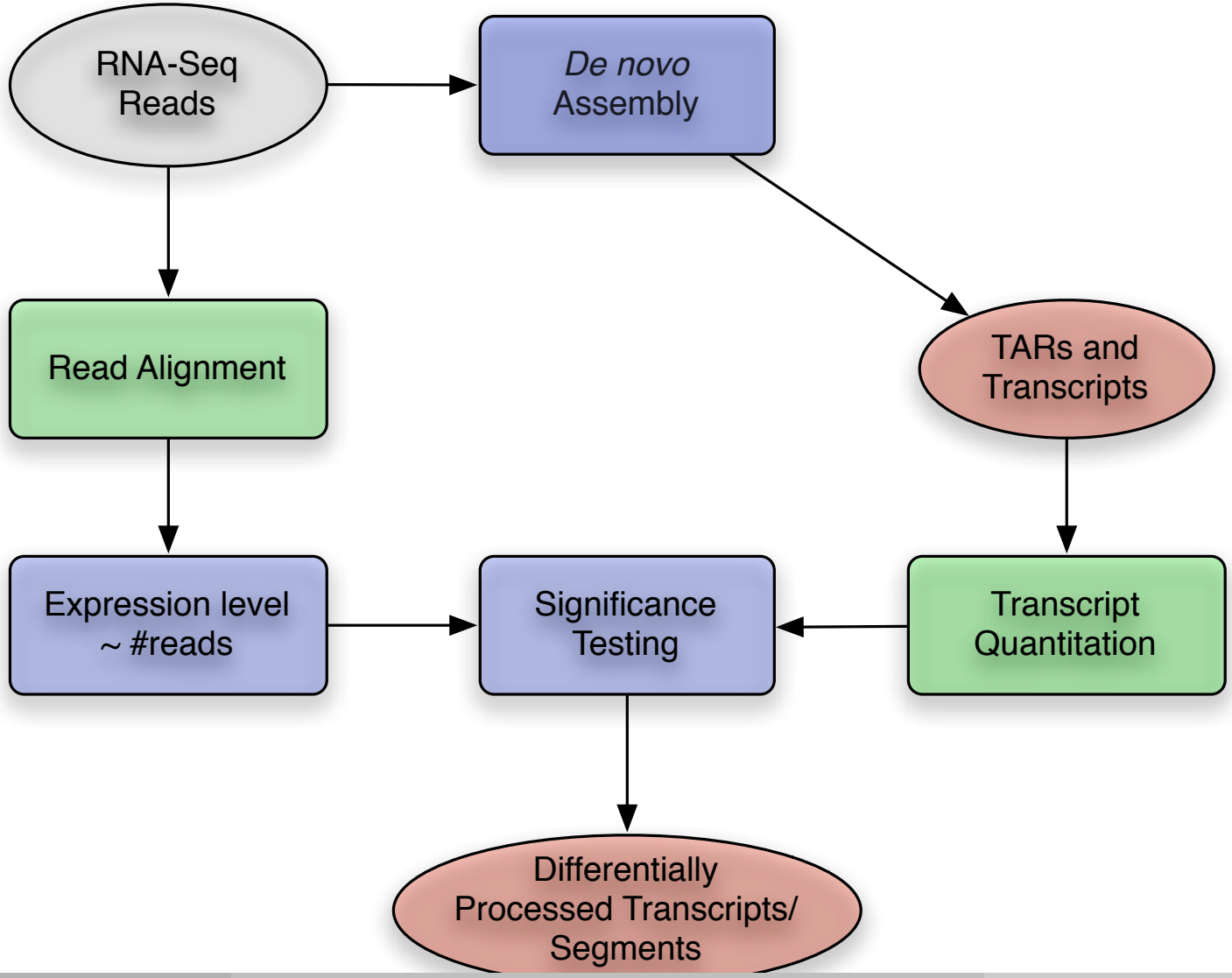
LETTERS

## Understanding mechanisms underlying human gene expression variation with RNA sequencing

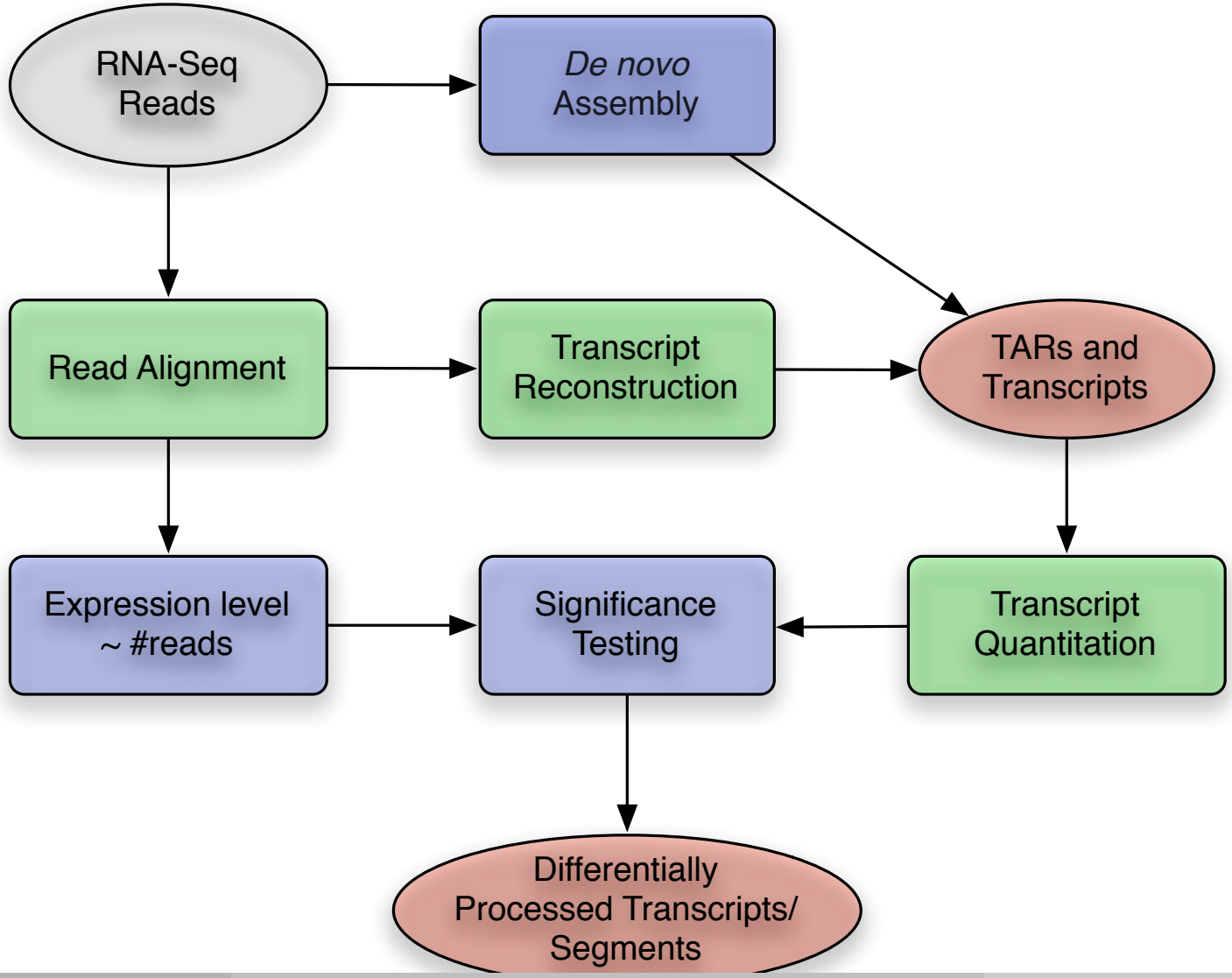
Joseph K. Pickrell[1], John C. Marioni[1], Athma A. Pai[1], Jacob F. Degner[1], Barbara E. Engelhardt[2], Everlyne Nkadori[1,3], Jean-Baptiste Veyrieras[1], Matthew Stephens[1,4], Yoav Gilad[1] & Jonathan K. Pritchard[1,3]

[Pickrell et al., 2010]

# Common RNA-Seq Analysis Steps



# Common RNA-Seq Analysis Steps



# Next Step: Transcript Prediction

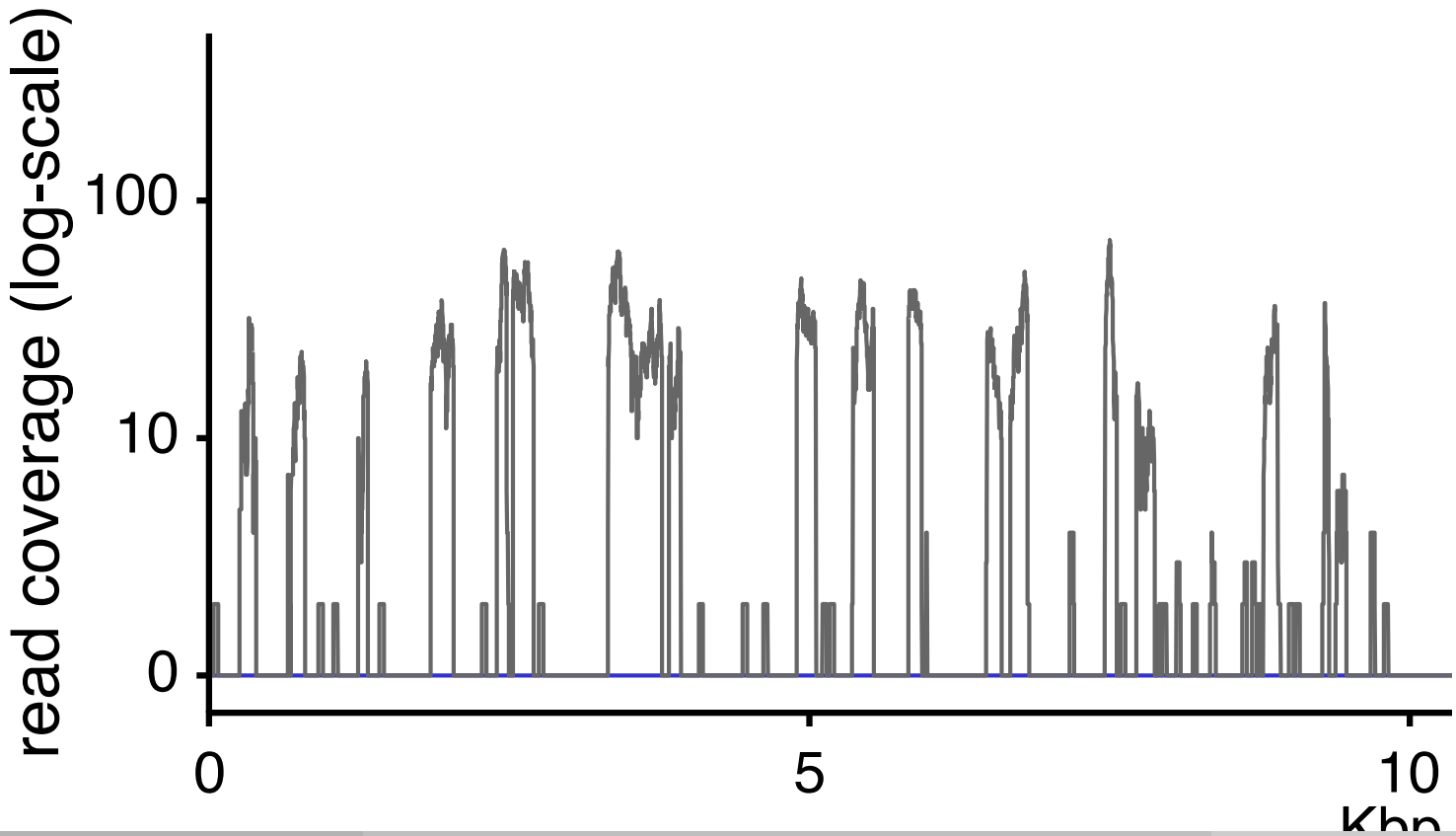



A. Coverage segmentation algorithm **mTiM** for general transcripts (no coding bias/assumption)

B. Extension of the mGene gene finding system to use NGS data for protein coding transcript prediction (**mGene.ngs**)

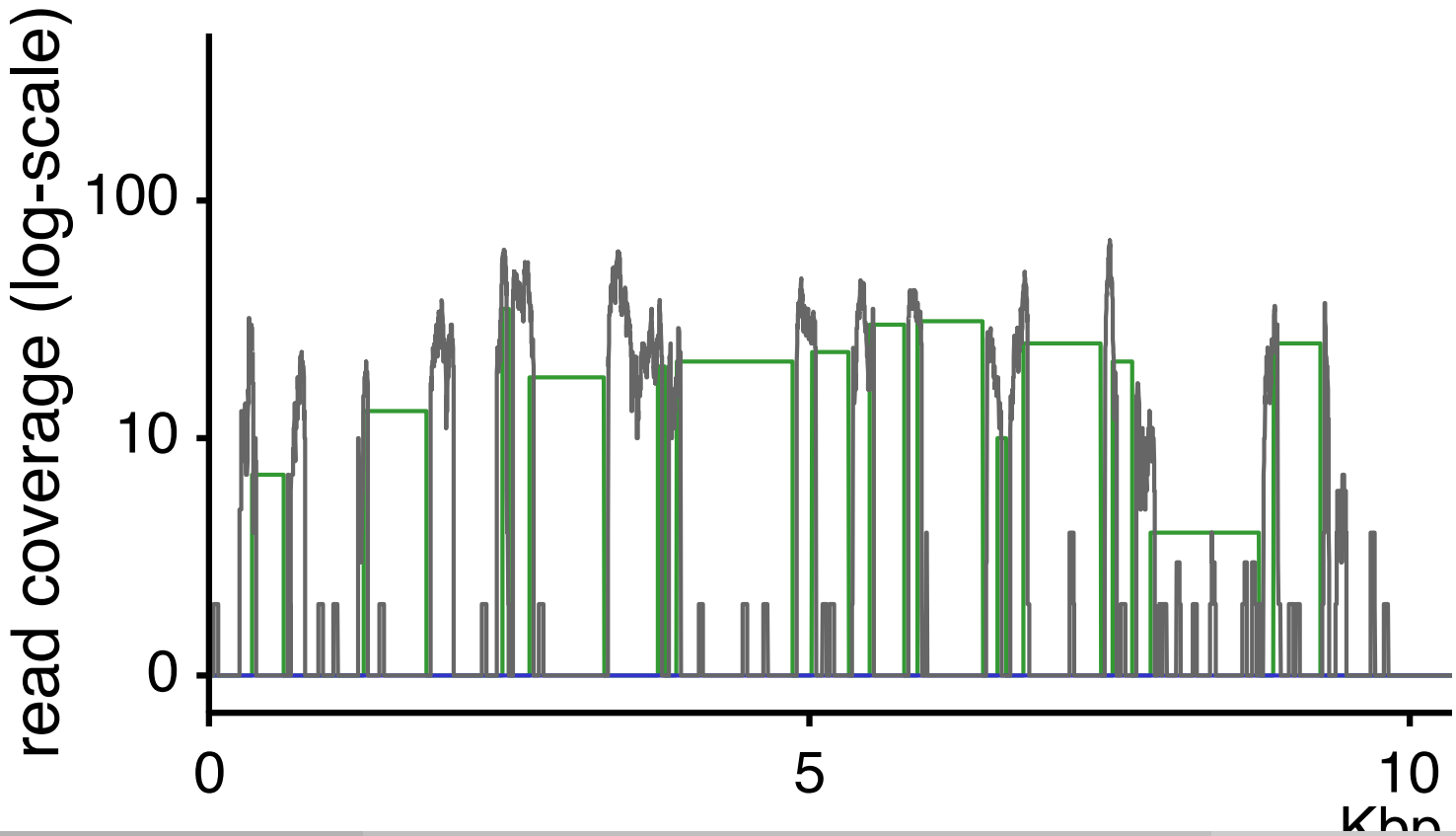
# mTiM: Read Coverage Segmentation

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



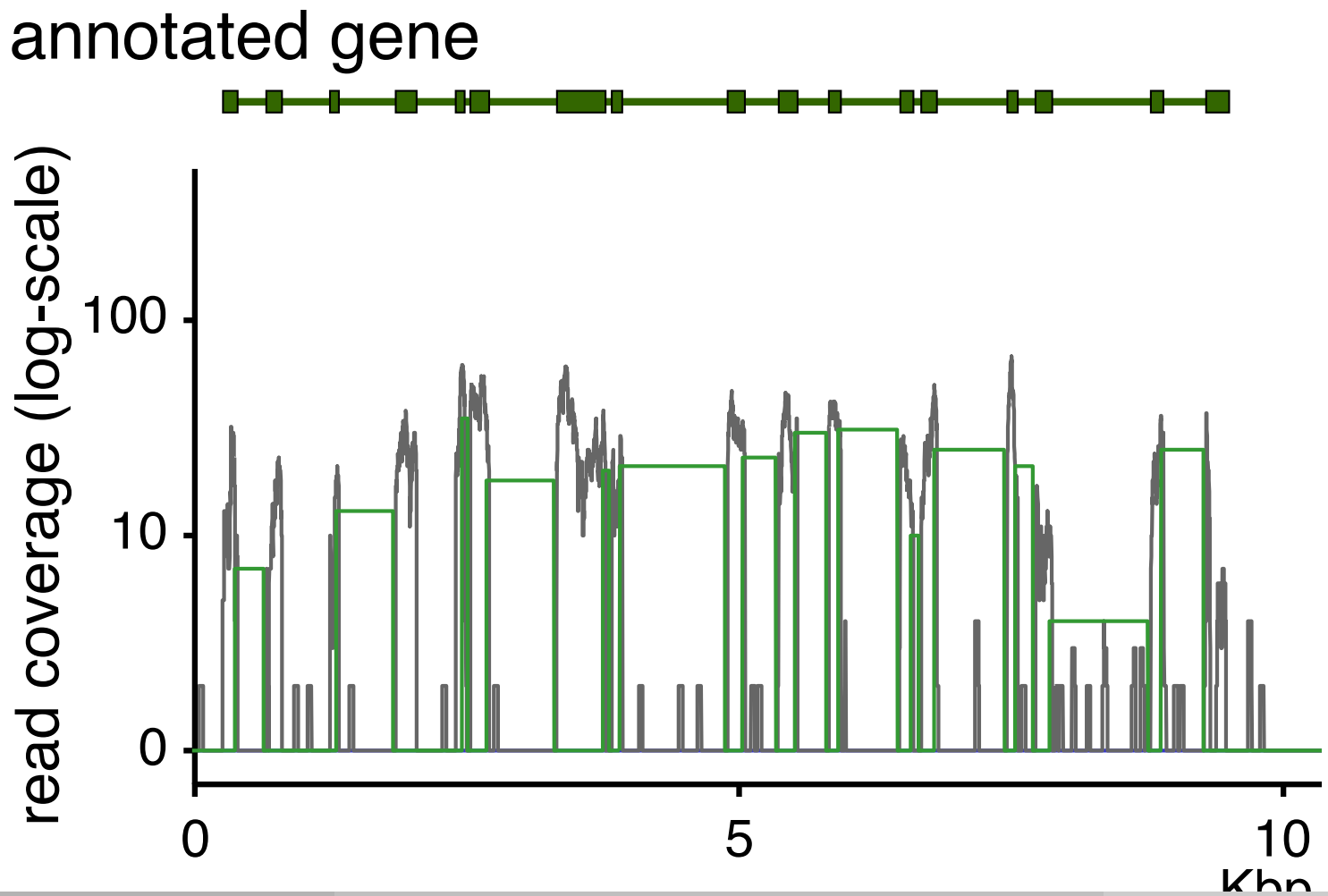
# mTiM: Read Coverage Segmentation

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

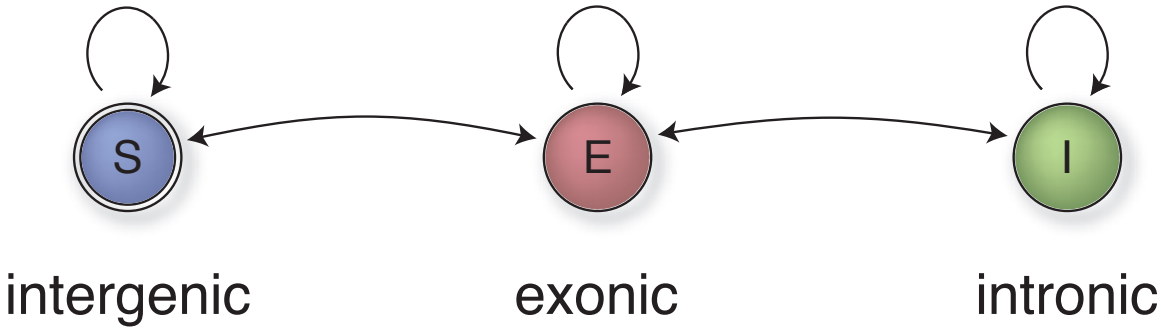


# mTiM: Read Coverage Segmentation

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

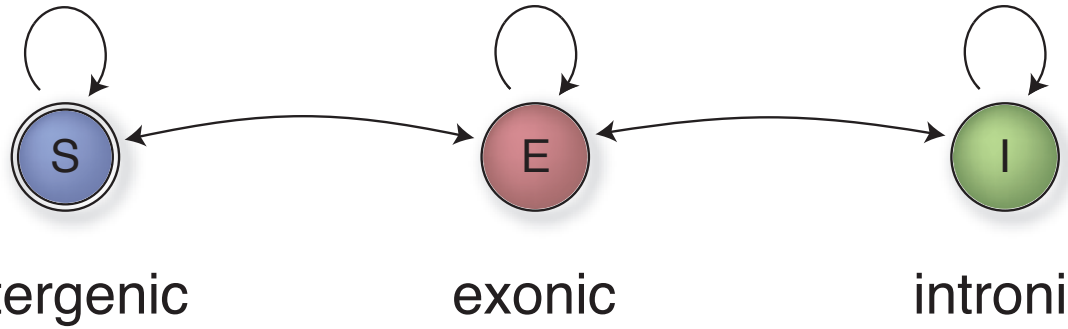


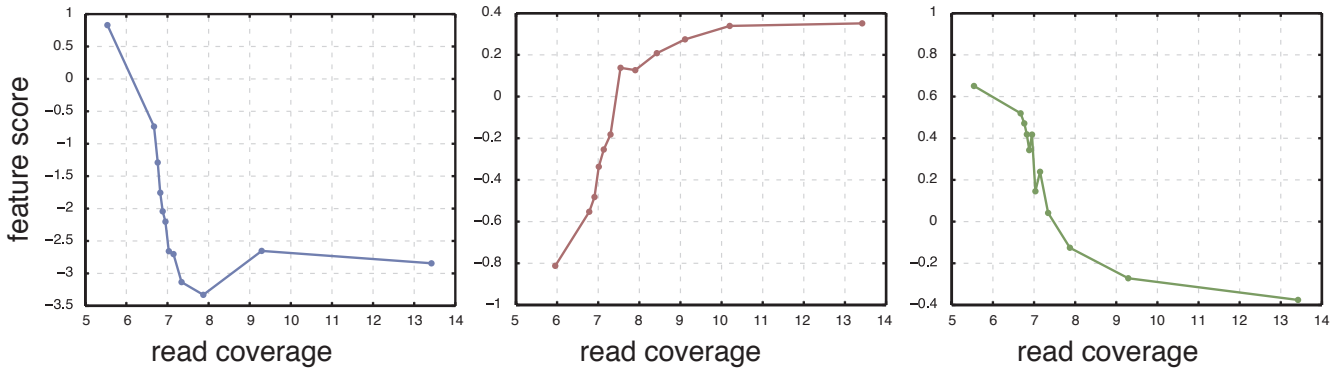
# The mTiM Segmentation Approach




- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal → score
- Extension: Score spliced reads and splice sites

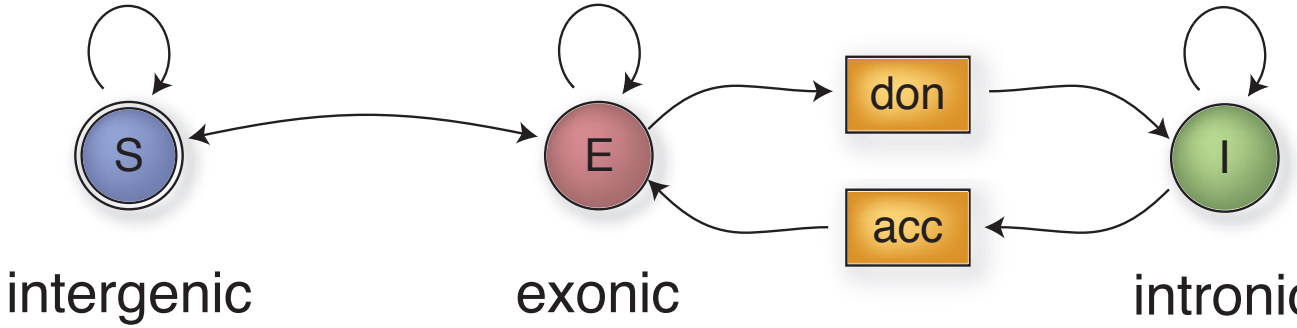
# The mTiM Segmentation Approach

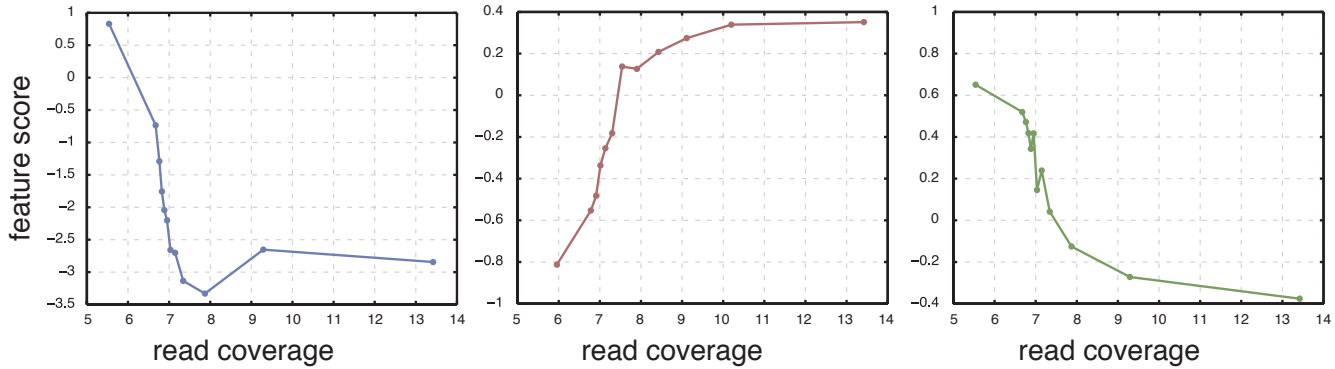



- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal $\rightarrow$ score
- Extension: Score spliced reads and splice sites

# The mTiM Segmentation Approach

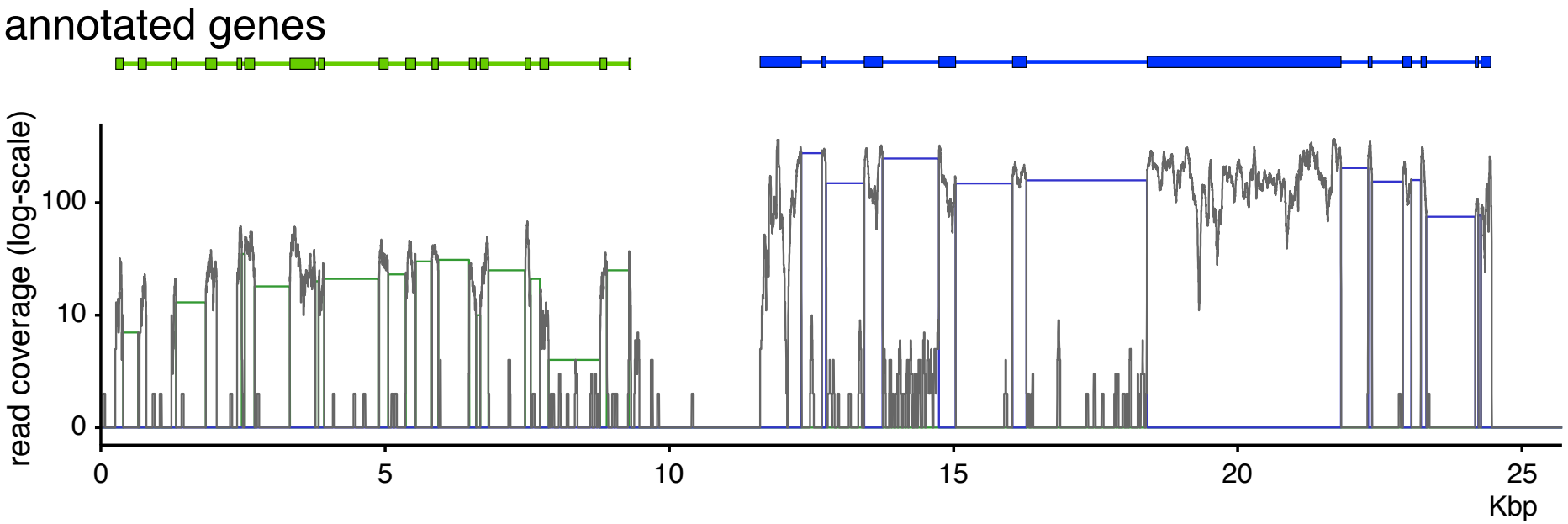


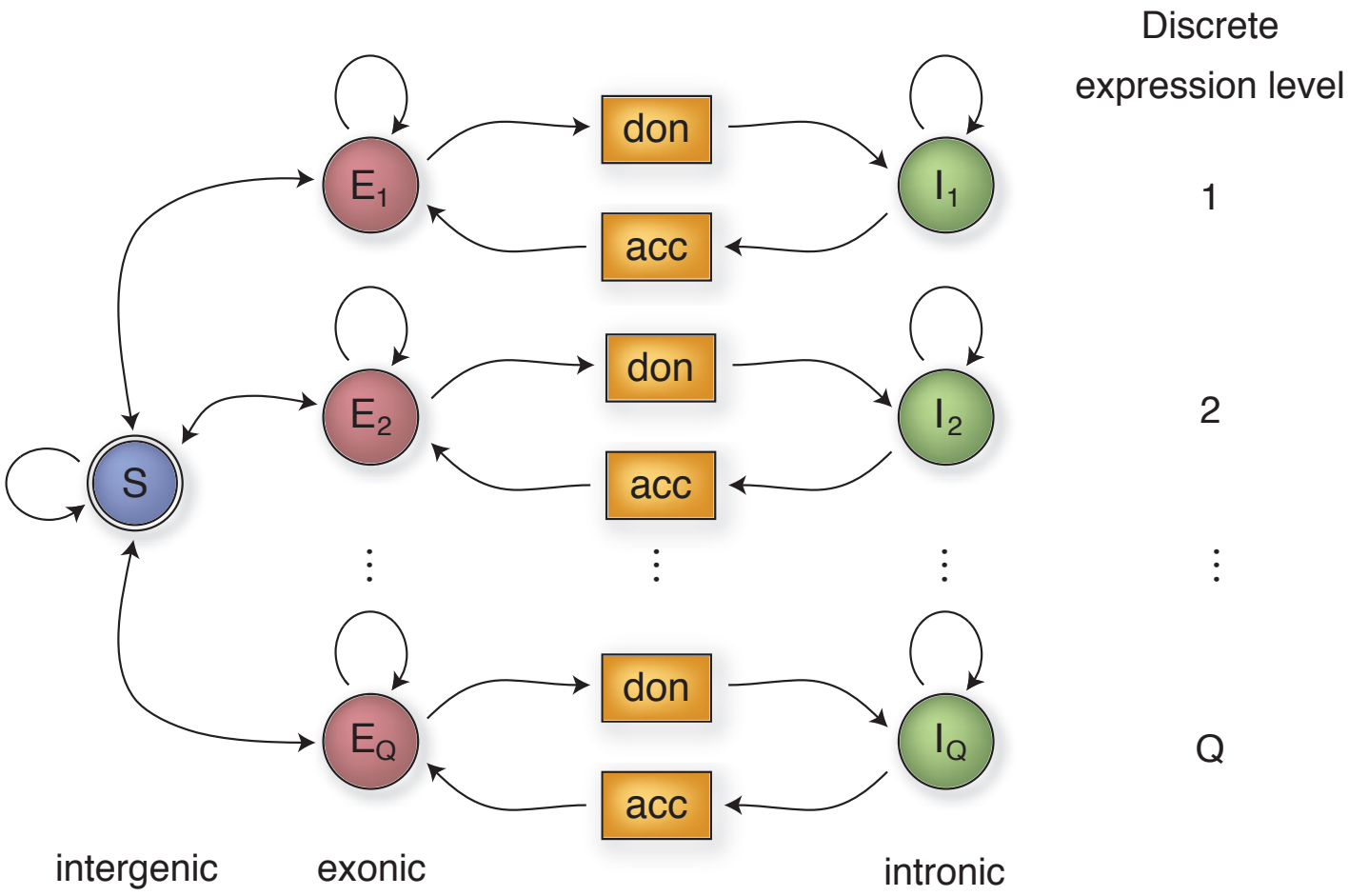

- Learn to associate a state with each position given its read coverage and local context
- HM-SVM training: Optimize transformations: signal $\rightarrow$ score
- Extension: Score spliced reads and splice sites

# The mTiM Segmentation Approach

Idea: Assume uniform read coverage within exons of same transcript

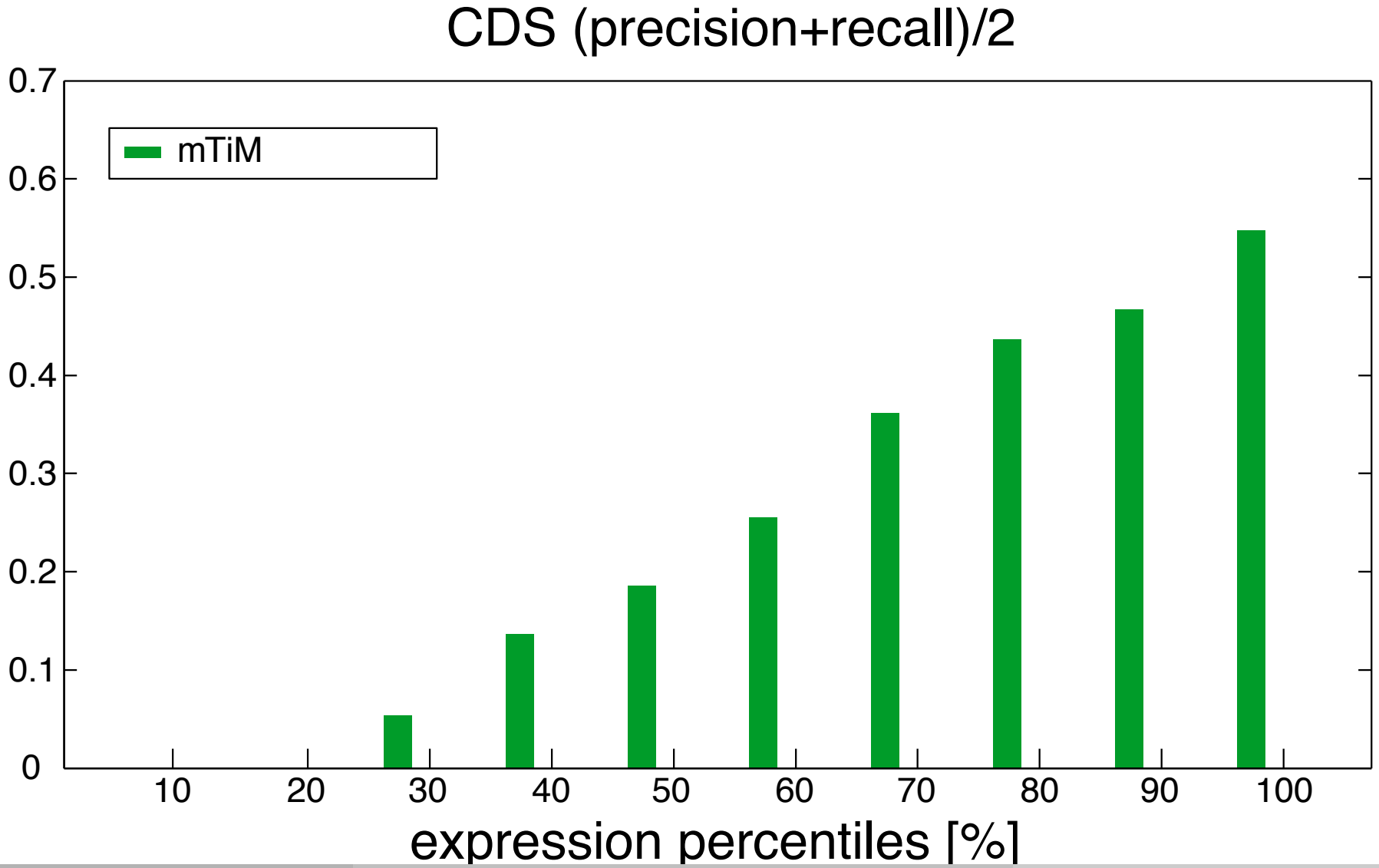


# The mTiM Segmentation Approach

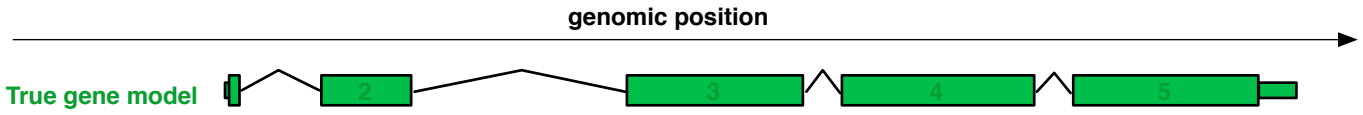


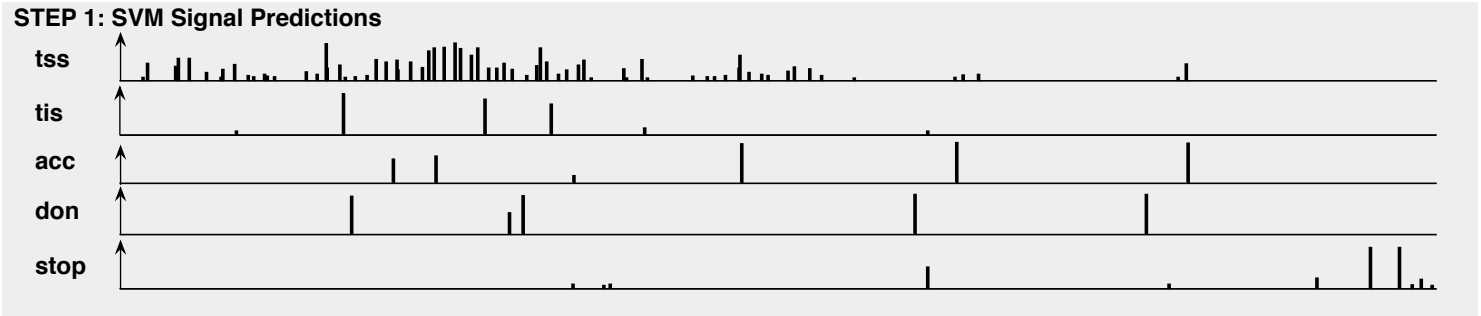

Carry "expression level" information between exons of same transcript

# Preliminary Evaluation (*C. elegans*)

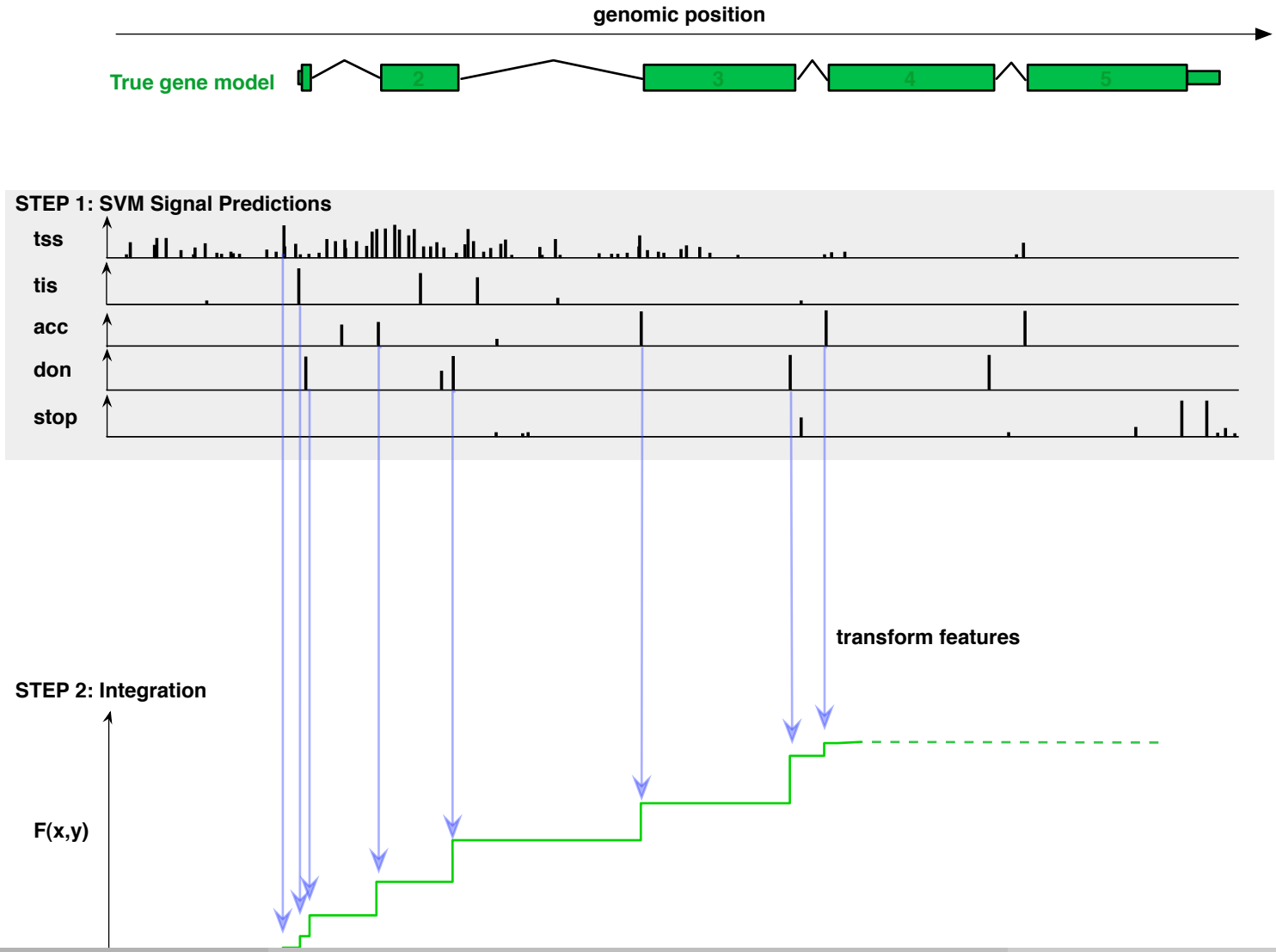


# mGene-based Transcript Prediction I

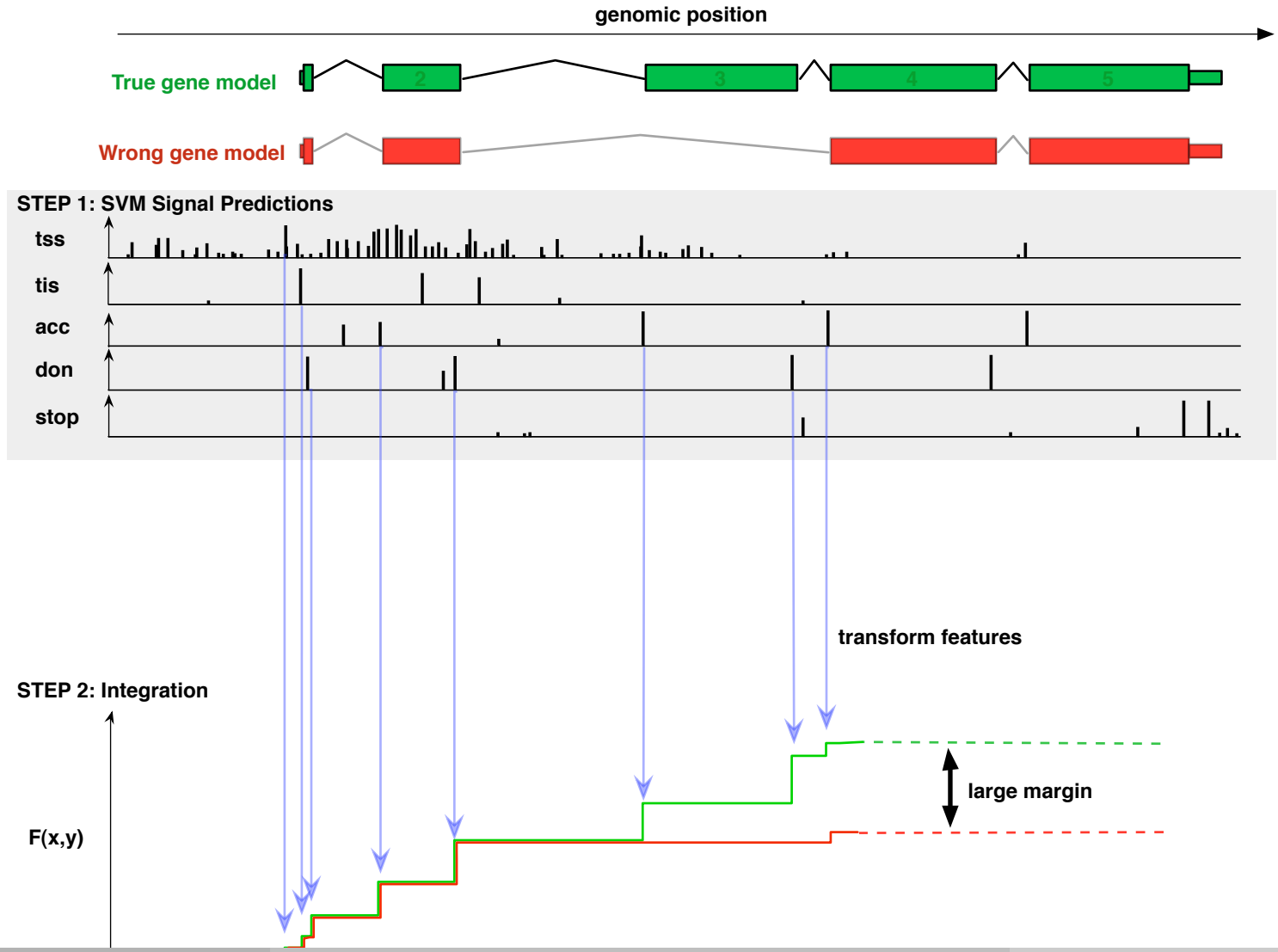


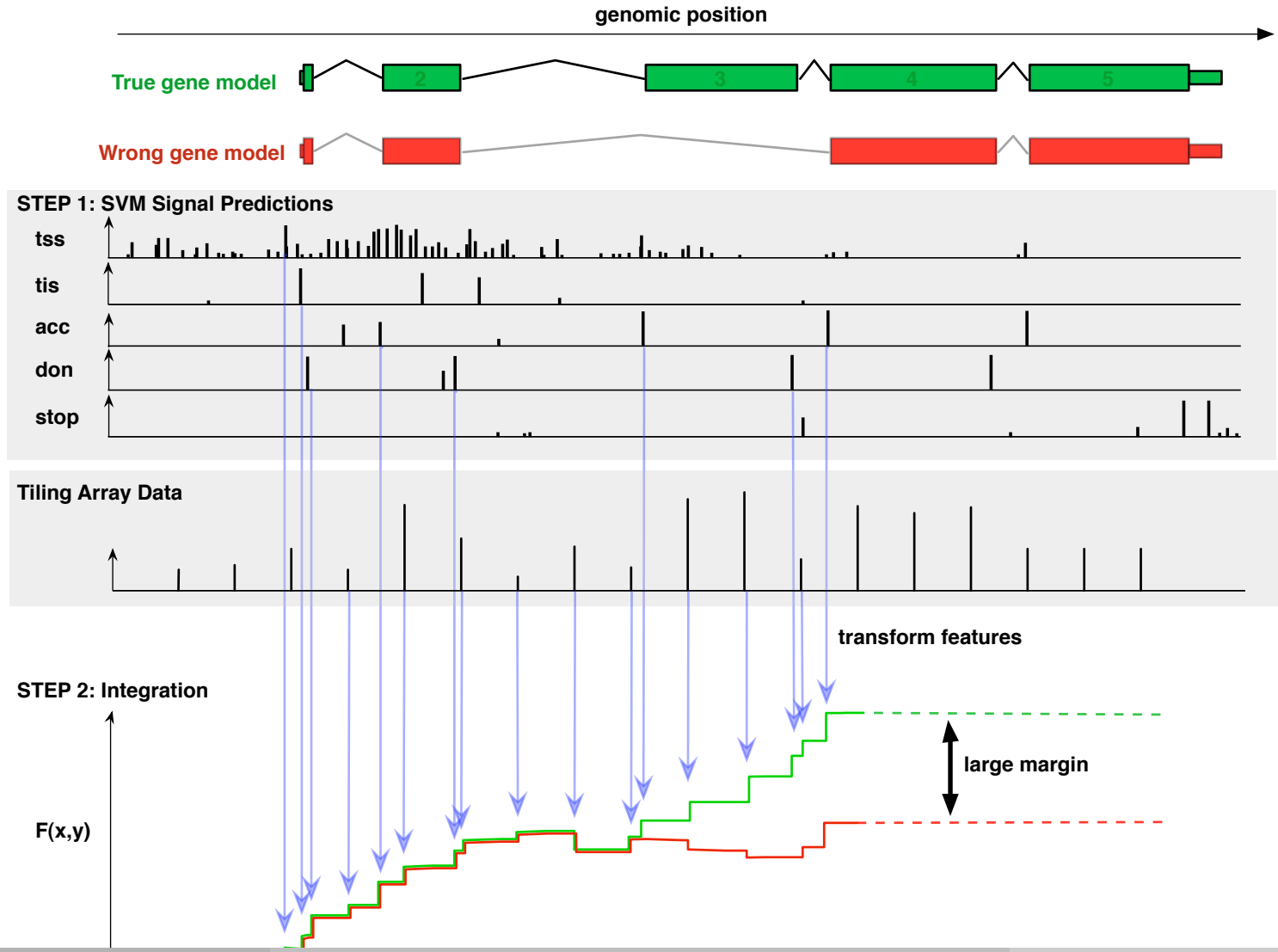
# mGene-based Transcript Prediction I



# mGene-based Transcript Prediction I



# mGene-based Transcript Prediction I



# mGene-based Transcript Prediction II




**mGene.ngs** (Behr et al., unpublished; Schweikert et al., 2009a,b)

- 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 RNA-Seq (1 tissue) | **81.1%** |

# mGene-based Transcript Prediction II




**mGene.ngs** (Behr et al., unpublished; Schweikert et al., 2009a,b)

- 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 RNA-Seq (1 tissue) | 81.1% |

# mGene-based Transcript Prediction II



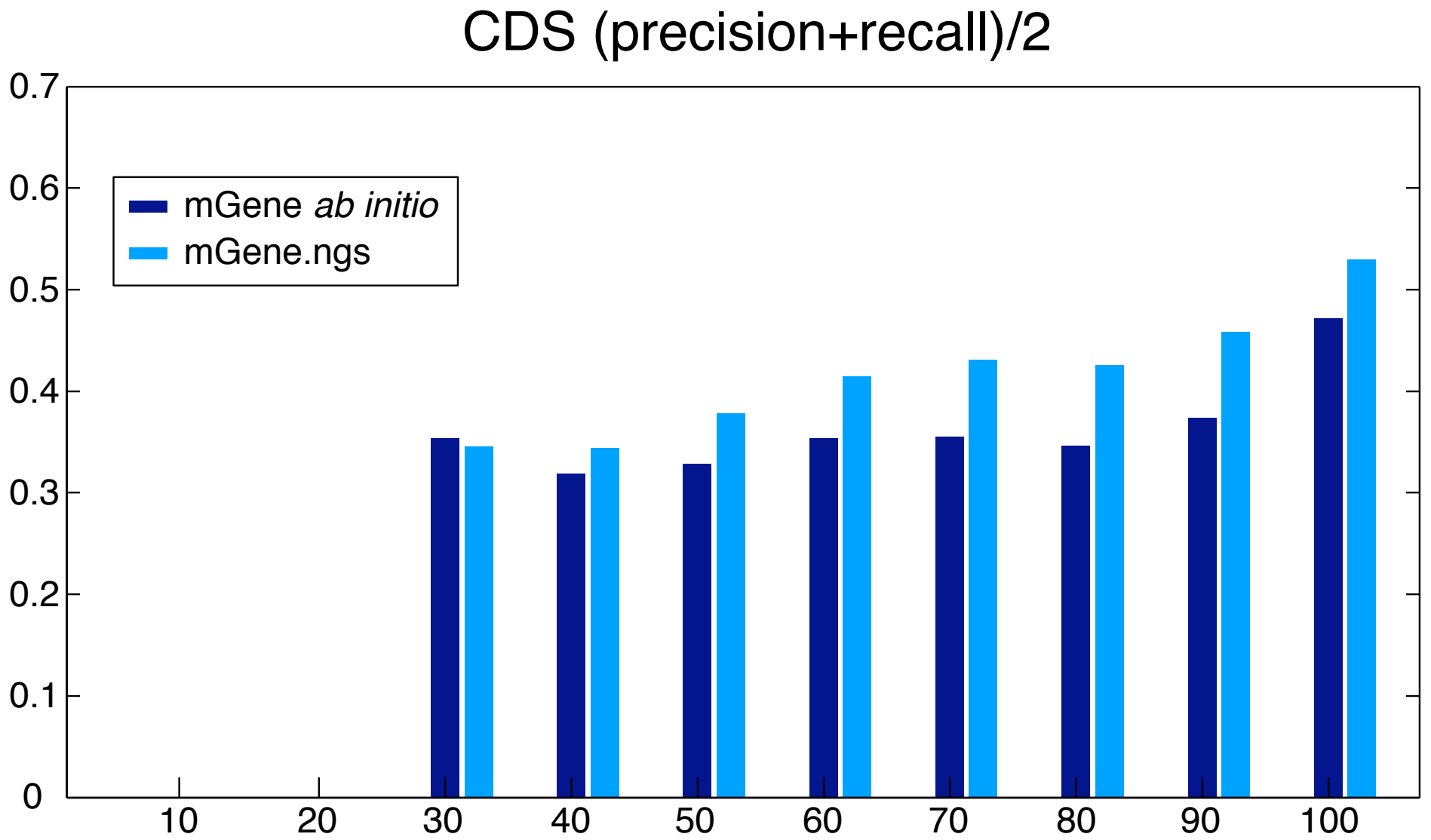

**mGene.ngs** (Behr et al., unpublished; Schweikert et al., 2009a,b)

- Use transcriptome measurements to enhance recognition of exonic regions

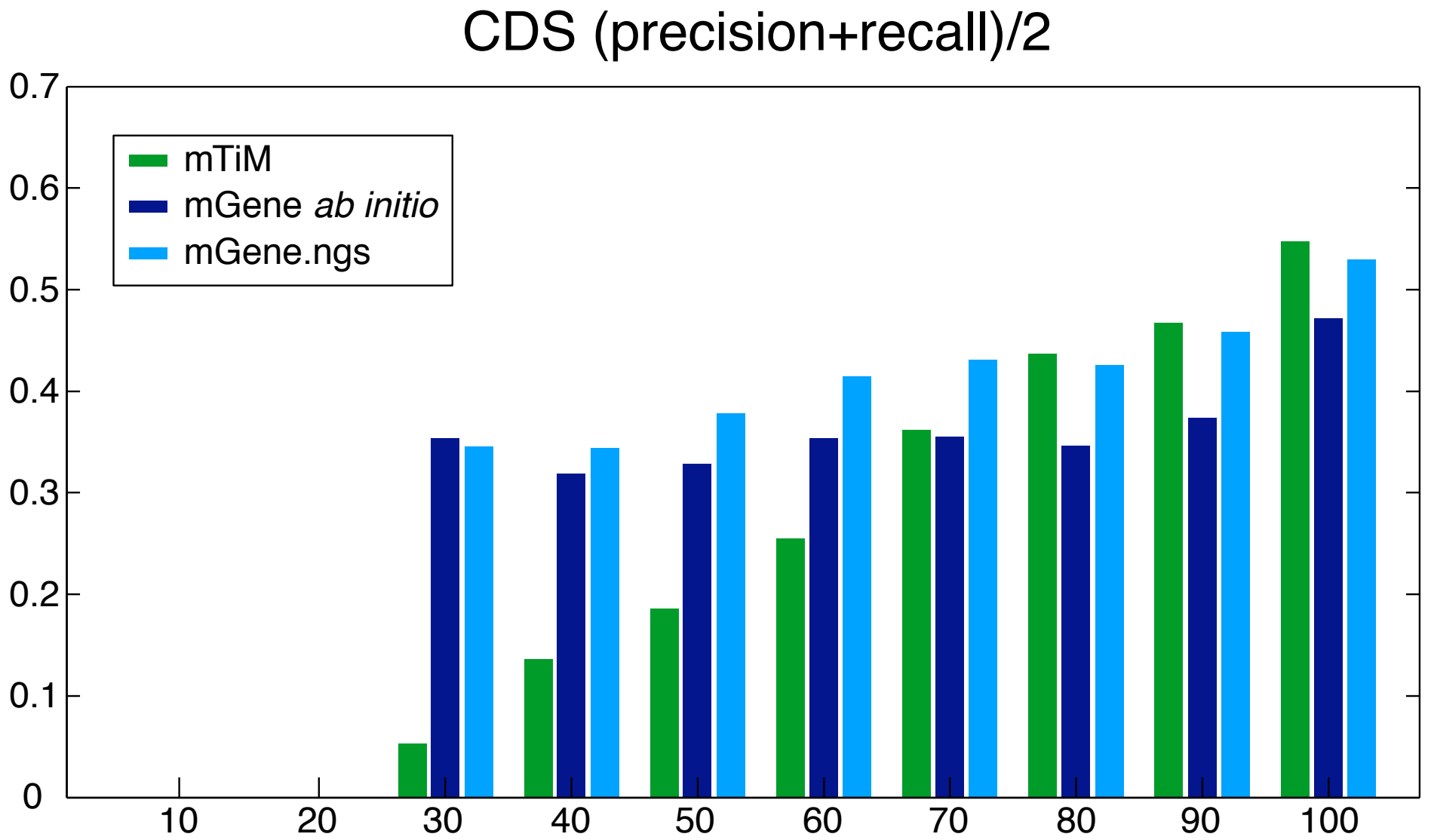
**Results for A. thaliana:** (Comparison with known gene models)

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

# Preliminary Evaluation (*C. elegans*)



# Preliminary Evaluation (*C. elegans*)



# Digestion




- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

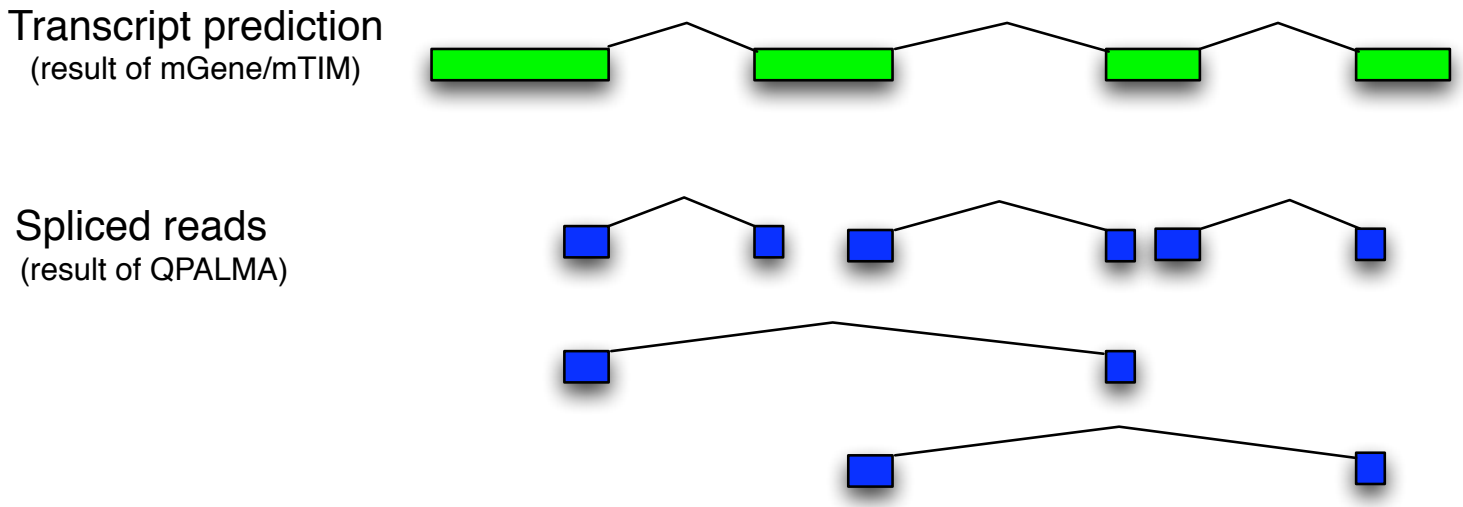
# Digestion




- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:
- Paths through splicing graph define *alternative transcripts*

# Digestion

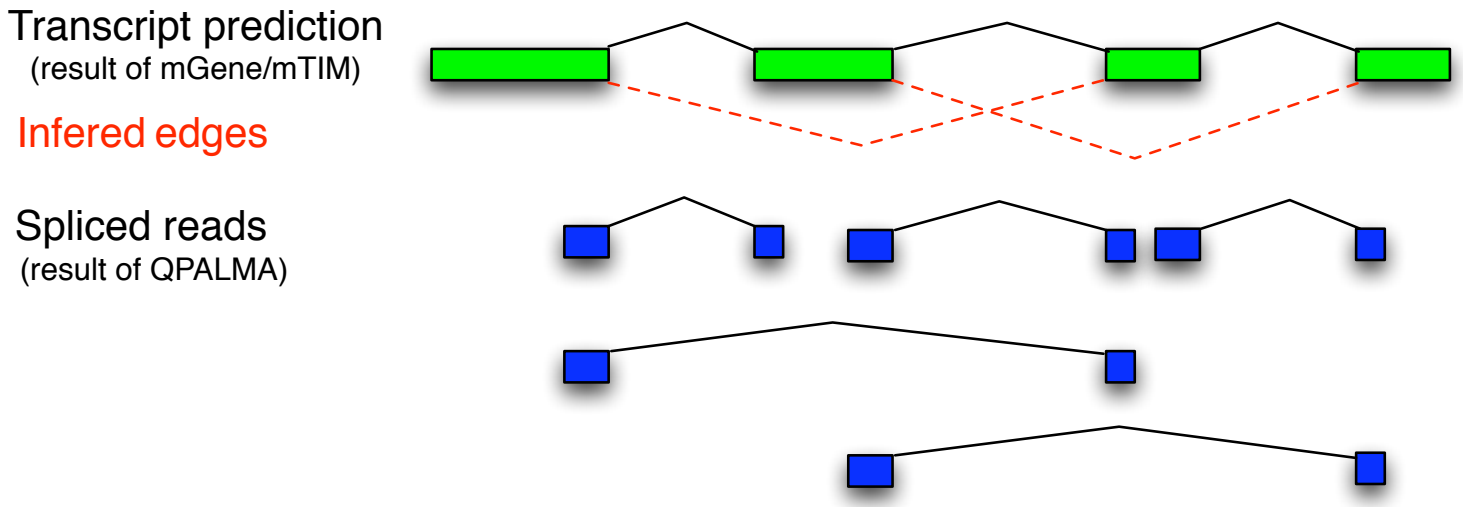
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:




- Paths through splicing graph define *alternative transcripts*

# Digestion

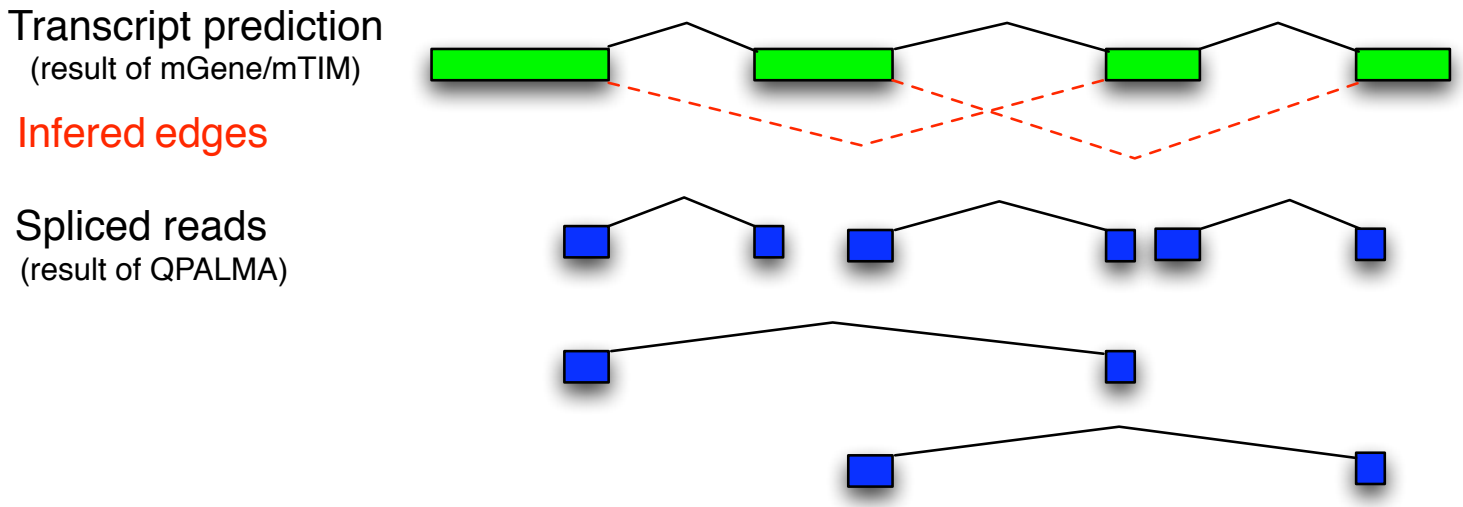
- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:




- Paths through splicing graph define *alternative transcripts*

# Digestion

- **mGene.ngs** and **mTiM** predict single transcripts
- **mGene.ngs** uses more assumptions on structure of transcripts
- **mTiM** exploits "uniformity" of read coverage among exons of same transcript
- **Alt. Transcripts:** Spliced reads for splicing graph completion:




- Paths through splicing graph define *alternative transcripts*

# What's next?




- Improvements of genome annotations
  - Find UTRs, missing exons, ...
  - Identify non-coding RNAs
- Transcript expression quantitation
  - In different conditions, tissues, ...
  - In different genotypes for QTL mapping
- Now we can look at new genomes
  - Model organisms
  - Cancer genomes

# Summary & Conclusions




- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Predict protein coding transcripts well
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- Combination of predictions and transcriptome measurements
  - Lead to improved genome annotations
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

# Summary & Conclusions




- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Predict protein coding transcripts well
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- Combination of predictions and transcriptome measurements
  - Lead to improved genome annotations
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

# Summary & Conclusions




- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Predict protein coding transcripts well
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- Combination of predictions and transcriptome measurements
  - Lead to improved genome annotations
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

# Summary & Conclusions




- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Predict protein coding transcripts well
  - Limitations: no alternative transcripts or expression information
- Analysis of tiling array data & short reads
  - Identification of alternative and differential splicing
  - Segmentation of tiling array data to identify transcribed regions
  - Read alignments difficult, but very promising data
- Combination of predictions and transcriptome measurements
  - Lead to improved genome annotations
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

# Acknowledgments


## Gene Finding

- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Georg Zeller (FML & MPI)

## Tiling Arrays

- **Georg Zeller** (FML & MPI)
- *Sascha Laubinger* (MPI)
- Detlef Weigel (MPI)

## Short Read Analysis

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)
- **Regina Bohnert** (FML)
- Andre Kahles (FML)
- Geraldine Jean (FML)


## More Information

- http://www.fml.mpg.de/raetsch
- Slides with references are available online

# Acknowledgments


## Gene Finding

- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Georg Zeller (FML & MPI)

## Tiling Arrays

- **Georg Zeller** (FML & MPI)
- *Sascha Laubinger* (MPI)
- Detlef Weigel (MPI)

## Short Read Analysis

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)
- **Regina Bohnert** (FML)
- Andre Kahles (FML)
- Geraldine Jean (FML)


## More Information

- `http://www.fml.mpg.de/raetsch`
- Slides with references are available online

# Acknowledgments

## Thank you!





### Gene Finding

- **Gabi Schweikert** (FML)
- **Jonas Behr** (FML)
- Georg Zeller (FML & MPI)

### Tiling Arrays

- **Georg Zeller** (FML & MPI)
- *Sascha Laubinger* (MPI)
- Detlef Weigel (MPI)

### Short Read Analysis

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)
- **Regina Bohnert** (FML)
- Andre Kahles (FML)
- Geraldine Jean (FML)


### More Information

- `http://www.fml.mpg.de/raetsch`
- Slides with references are available online

# References I

J. Behr, G. Schweikert, J. Cao, F. De Bona, G. Zeller, S. Laubinger, S. Ossowski, K. Schneeberger, D. Weigel, and G. Rätsch. Rna-seq and tiling arrays for improved gene finding. URL `http://www.fml.tuebingen.mpg.de/raetsch/lectures/RaetschGenomeInformatics08.pdf`. Oral presentation at the CSHL Genome Informatics Meeting, September 2008.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a pratical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

M. Bergkessel, G. Wilmes, and C. Guthrie. Snapshot: Formation of mrnps. *Cell*, 136, January 2009.

Paul Bertone, Viktor Stolc, Thomas E Royce, Joel S Rozowsky, Alexander E Urban, Xiaowei Zhu, John L Rinn, Waraporn Tongprasit, Manoj Samanta, Sherman Weissman, Mark Gerstein, and Michael Snyder. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–6, Dec 2004. doi: 10.1126/science.1103388.

B.M. Bolstad, R.I. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods of high density oligonucleotide array data based on variance bias. *Bioinformatics*, 2 (19):185–193, 2003.

R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FFEBS Lett.*, 573(1-3):83–92, 2004.

# References II

RM Clark, G Schweikert, C Toomajian, S Ossowski, G Zeller, P Shinn, N Warthmann, TT Hu, G Fu, DA Hinds, H Chen, KA Frazer, DH Huson, B Schölkopf, M Nordborg, G Rätsch, JR Ecker, and D Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–342, 2007. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1138632.

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. submitted.

Lior David, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis, and Lars M Steinmetz. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA*, 103(14):5320–5, Apr 2006. doi: 10.1073/pnas.0601091103.

F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Qpalma: Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24:i174–i180, 2008.

Jiang Du, Joel S Rozowsky, Jan O Korbel, Zhengdong D Zhang, Thomas E Royce, Martin H Schultz, Michael Snyder, and Mark Gerstein. A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chip-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, 22(24):3016–24, Dec 2006. doi: 10.1093/bioinformatics/btl515.

# References III

R Durbin, S Eddy, A Krogh, and G Mitchison. *Biological Sequence Analysis: Probabilistic models of protein and nucleic acids*. Cambridge University Press, 1998.

R.M. Fox, J.D. Watson, S.E. Von Stetina, J. McDermott, T.M. Brodigan, T. Fukushige, M. Krause, and D.M. Miller. The embryonic muscle transcriptome of caenorhabditis elegans. *Genome Biology*, 8(9):R188, 2007.

Housheng He, Jie Wang, Tao Liu, X Shirley Liu, Tiantian Li, Yunfei Wang, Zuwei Qian, Haixia Zheng, Xiaopeng Zhu, Tao Wu, Baochen Shi, Wei Deng, Wei Zhou, Geir Skogerbø, and Runsheng Chen. Mapping the c. elegans noncoding transcriptome with a whole-genome tiling microarray. *Genome Research*, 17(10):1471–7, Oct 2007. doi: 10.1101/gr.6611807.

S. Hochreiter, D.-A. Clevert, and K. Obermayer. A new summerization method for affymetrix probe level data. *Bioinformatics*, 22(8):934–949, 2006.

W. Huber, J. Toedling, and L. M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(6):1963–1970, 2006a.

Wolfgang Huber, Joern Toedling, and Lars M Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(16):1963–70, Aug 2006b. doi: 10.1093/bioinformatics/btl289. URL `http://bioinformatics.oxfordjournals.org/cgi/content/full/22/16/1963`.

R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):e15, 2003a.

# References IV

Rafael A Irizarry, Benjamin M Bolstad, Francois Collin, Leslie M Cope, Bridget Hobbs, and Terence P Speed. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31(4):e15, Feb 2003b.

Hongkai Ji and Wing Hung Wong. Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–36, Sep 2005. doi: 10.1093/bioinformatics/bti593. URL `http://bioinformatics.oxfordjournals.org/cgi/content/full/21/18/3629`.

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

Dione Kampa, Jill Cheng, Philipp Kapranov, Mark Yamanaka, Shane Brubaker, Simon Cawley, Jorg Drenkow, Antonio Piccolboni, Stefan Bekiranov, Gregg Helt, Hari Tammana, and Thomas R Gingeras. Novel rnas identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research*, 14(3):331–42, Mar 2004. doi: 10.1101/gr.2094104. URL `http://genome.cshlp.org/cgi/content/full/14/3/331`.

Todd C Mockler, Simon Chan, Ambika Sundaresan, Huaming Chen, Steven E Jacobsen, and Joseph R Ecker. Applications of dna tiling arrays for whole-genome analysis. *Genomics*, 85 (1):1–15, Jan 2005. doi: 10.1016/j.ygeno.2004.10.005.

S.B. Montgomery, M. Sammeth, and et al. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 2010.

# References V

Kasper Munch, Paul P Gardner, Peter Arctander, and Anders Krogh. A hidden markov model approach for determining expression from genomic tiling micro arrays. *BMC Bioinformatics*, 7:239, Jan 2006. doi: 10.1186/1471-2105-7-239.

Ossowski. Next generation sequencing. Oral presentation at the PhD Symposium in Tübingen, Germany, November 2007.

F. Pauli, Y. Liu, Y.A. Kim, P.J. Chen, and S.K. Kim. Chromosomal clustering and gata transcriptional regulation of intestine-expressed genes in c. elegans. *Development*, 133: 287–295, 2006.

J.K. Pickrell, J.C. Marioni, and et al. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 2010.

G. Rätsch and S. Sonnenburg. Accurate splice site detection for *Caenorhabditis elegans*. In K. Tsuda B. Schoelkopf and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21(Suppl. 1):i369–i377, June 2005.

Thomas E Royce, Nicholas J Carriero, and Mark B Gerstein. An efficient pseudomedian filter for tiling microrrays. *BMC Bioinformatics*, 8:186, Jan 2007a. doi: 10.1186/1471-2105-8-186.

# References VI

Thomas E Royce, Joel S Rozowsky, and Mark B Gerstein. Assessing the need for sequence-based normalization in tiling microarray experiments. *Bioinformatics*, 23(8): 988–97, Apr 2007b. doi: 10.1093/bioinformatics/btm052. URL http://bioinformatics.oxfordjournals.org/cgi/content/full/23/8/988.

Manoj Pratim Samanta, Waraporn Tongprasit, Himanshu Sethi, Chen-Shan Chin, and Viktor Stolc. Global identification of noncoding rnas in saccharomyces cerevisiae by modulating an essential rna processing pathway. *Proc Natl Acad Sci USA*, 103(11):4192–7, Mar 2006. doi: 10.1073/pnas.0507669103.

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 2009. doi: 10.1186/gb-2009-10-9-r98. URL http://genomebiology.com/2009/10/9/R98.

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.

# References VII

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

G.K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

S. Sonnenburg, G. Rätsch, A. Jagota, and K.-R. Müller. New methods for splice-site recognition. In *Proc. International Conference on Artificial Neural Networks*, 2002.

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.

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, 100(16):9440–5, Aug 2003. doi: 10.1073/pnas.1530509100.

# References VIII

Marc Sultan, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, 2008. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1160342.

I. Sutskever. Arachne: A whole genome shotgun assembler. oral presentation, 2008.

E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008. ISSN 1476-4687 (Electronic). doi: 10.1038/nature07509.

Junshi Yazaki, Brian D Gregory, and Joseph R Ecker. Mapping the genome landscape using tiling array technology. *Current Opinion in Plant Biology*, 10(5):534–42, Oct 2007. doi: 10.1016/j.pbi.2007.07.006. URL `http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VS4-4PG2S31-1&_user=29041&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_acct=C000003178&_version=1&_urlVersion=0&_userid=29041&md5=80ec1d3e091fd96a9f662289f6584c05`.

G Zeller, RM Clark, K Schneeberger, A Bohlen, D Weigel, and G Ratsch. Detecting polymorphic regions in arabidopsis thaliana with resequencing microarrays. *Genome Res*, 18 (6):918–929, 2008a. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

# References IX

G. Zeller, S. Henz, S. Laubinger, D. Weigel, and G. Rätsch. Transcript normalization and segmentation of tiling array data. In *Proc. PSB 2008*. World Scientific, 2008b.

G Zeller, S Henz, C Widmer, T Sachsenberg, G Rätsch, D Weigel, and S Laubinger. Stress-induced changes in the arabidopsis thaliana transcriptome analyzed using whole genome tiling arrays. *Plant J*, Feb 2009. doi: 10.1111/j.1365-313X.2009.03835.x.

Georg Zeller, Stefan R Henz, Sascha Laubinger, Detlef Weigel, and Gunnar Rätsch. Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 527–38, Jan 2008c.

D.R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*, 18:828–829, 2008.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *BioInformatics*, 16(9): 799–807, September 2000.