Next Generation Sequencing, Tiling Arrays and Predictive Sequence Analysis for Transcriptome Analysis

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society, Tübingen, Germany

9^{th} Course in Bioinformatics and Systems Biology for Molecular Biologists (March 24, 2009)



Friedrich Miescher Laboratory
of the Max Planck Society



Discovery of the Nuclein
(Friedrich Miescher, 1869)









Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein" (Miescher, 1874)

Discovery of the Nuclein
(Friedrich Miescher, 1869)









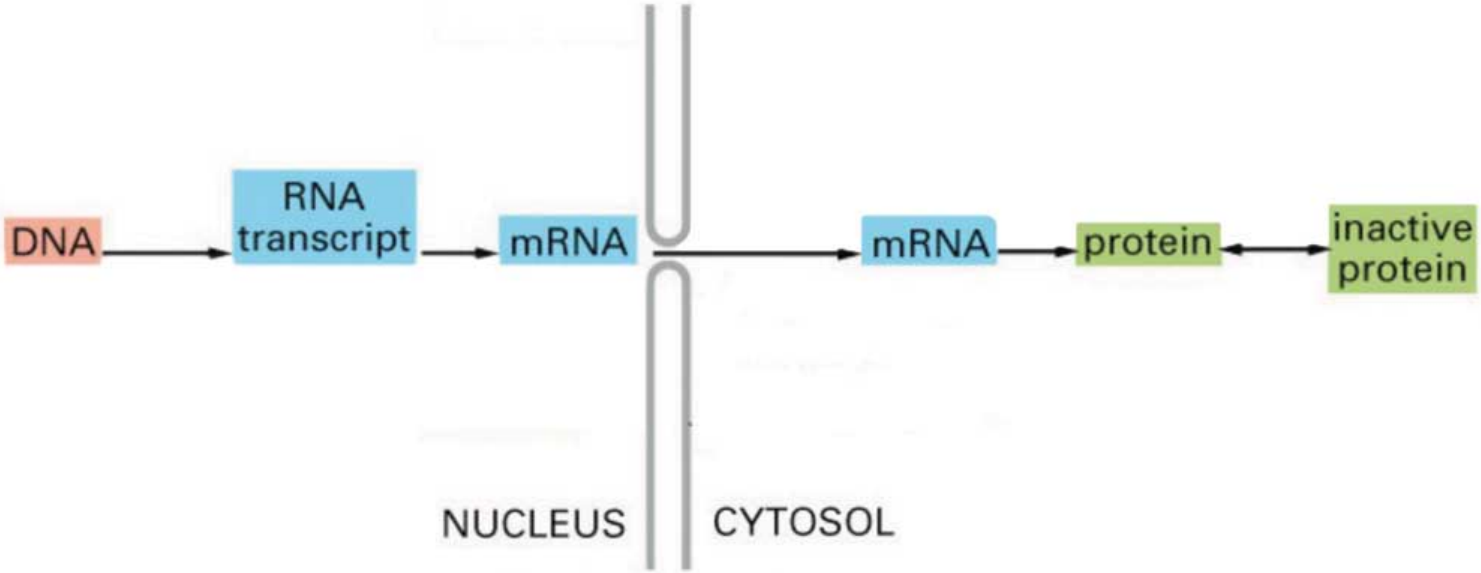
Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein" (Miescher, 1874)

Transcriptome Analysis

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



Then we can (try to) understand:

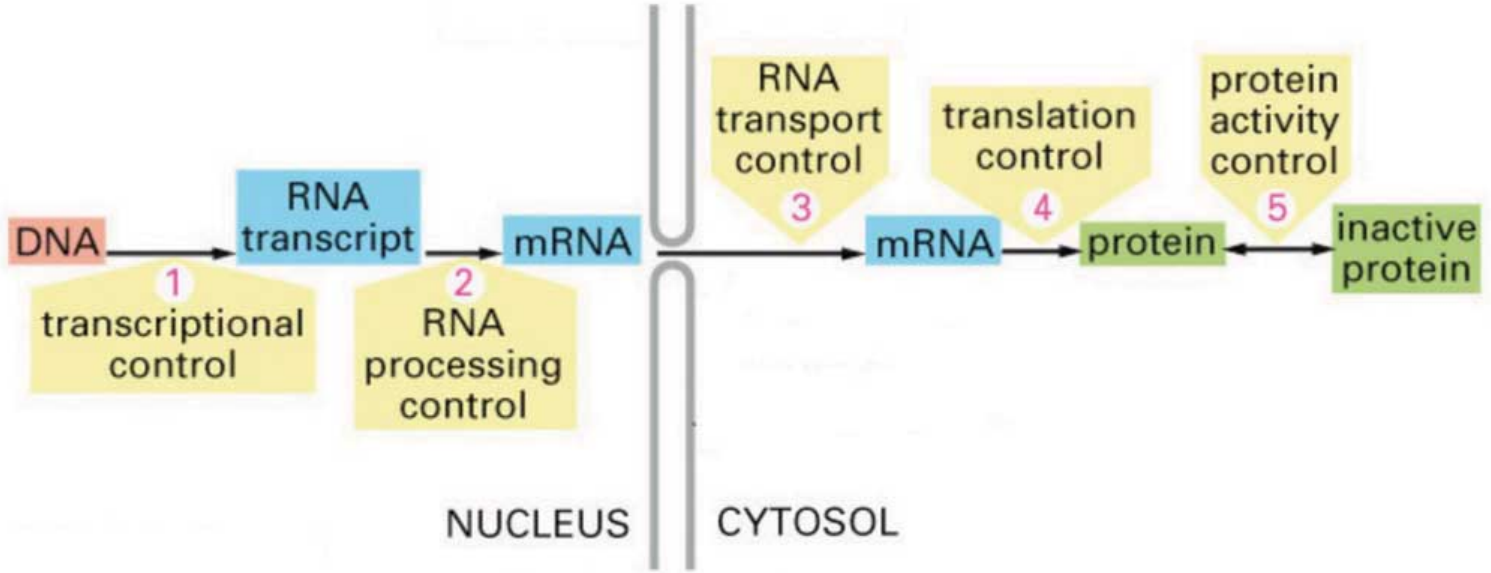
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

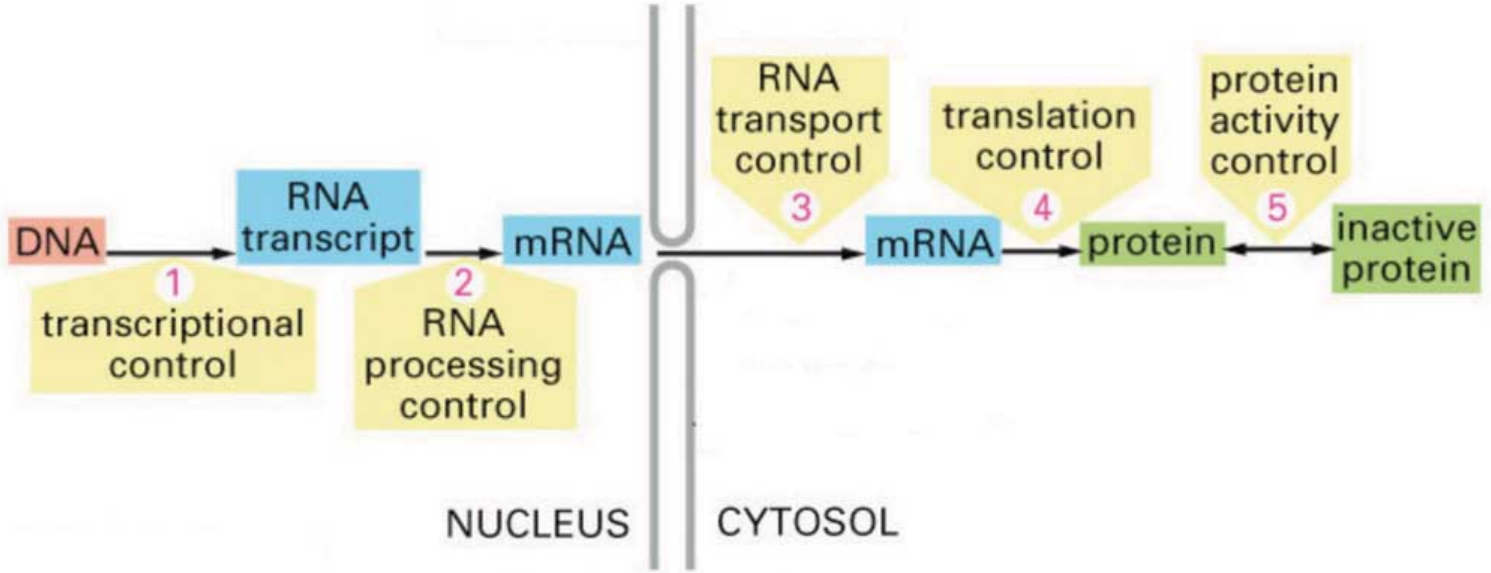
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

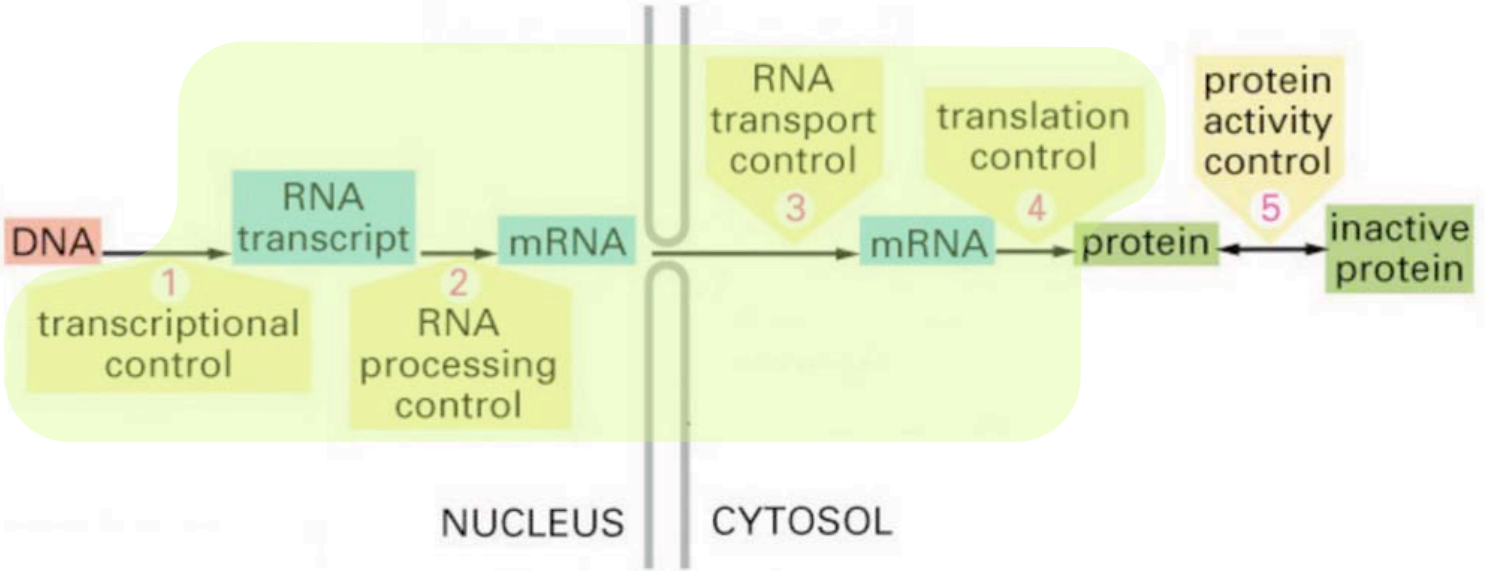
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

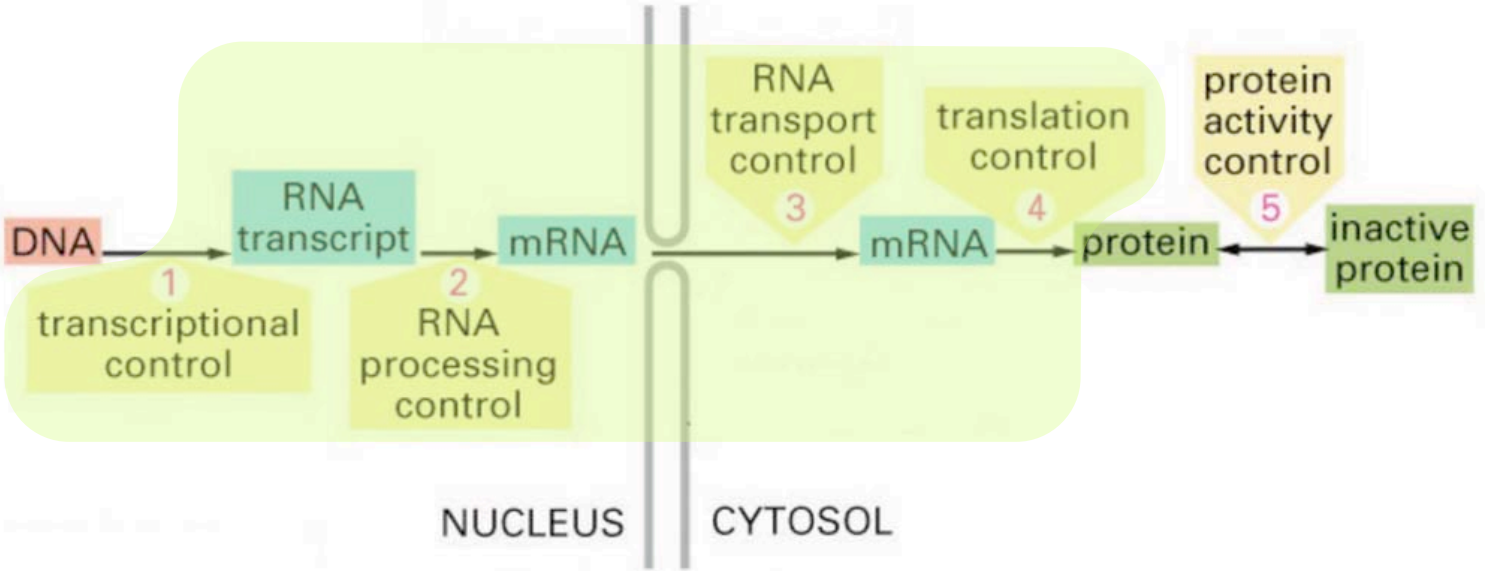
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (try to) understand:

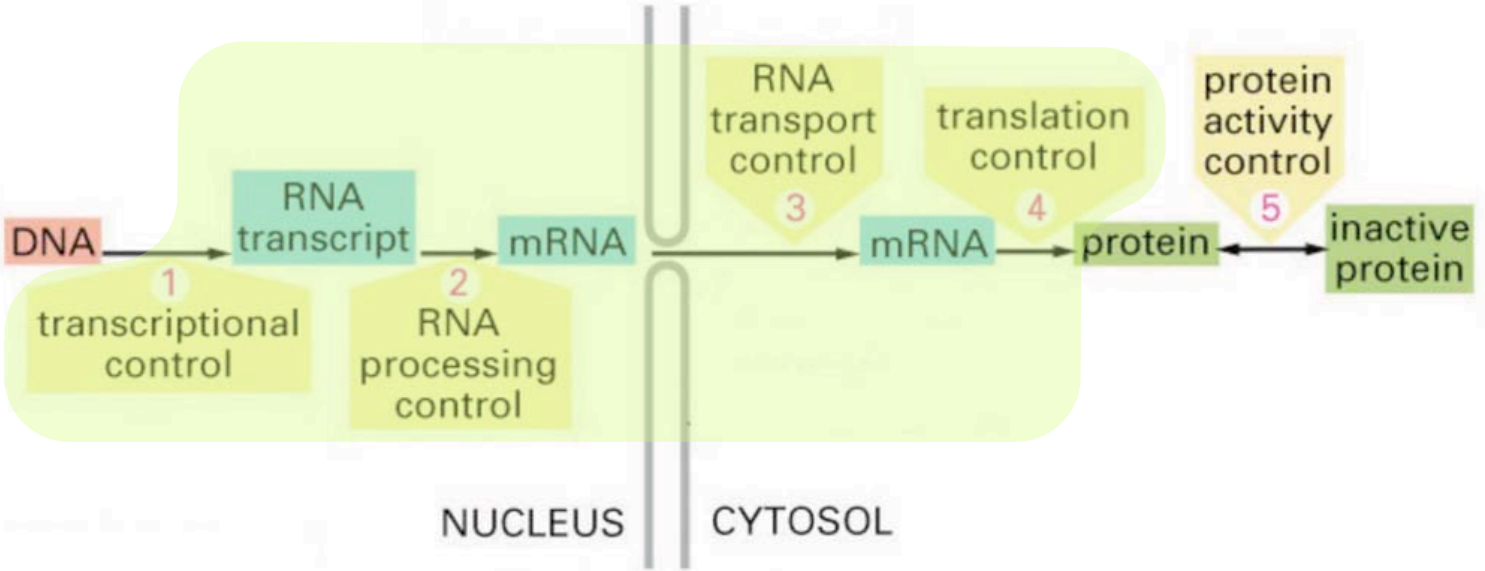
- Differences of active components between conditions/organisms?
- What changes when perturbing the biological system?

How to get the transcriptome?

1. Infer transcriptome from genomic DNA
2. Measure properties of transcriptome
3. Combine predictions with measurements

Transcriptome Analysis

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



Then we can (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
3. Combine predictions with measurements

Transcription & RNA Processing

- Newly synthesized pre-mRNA is capped.

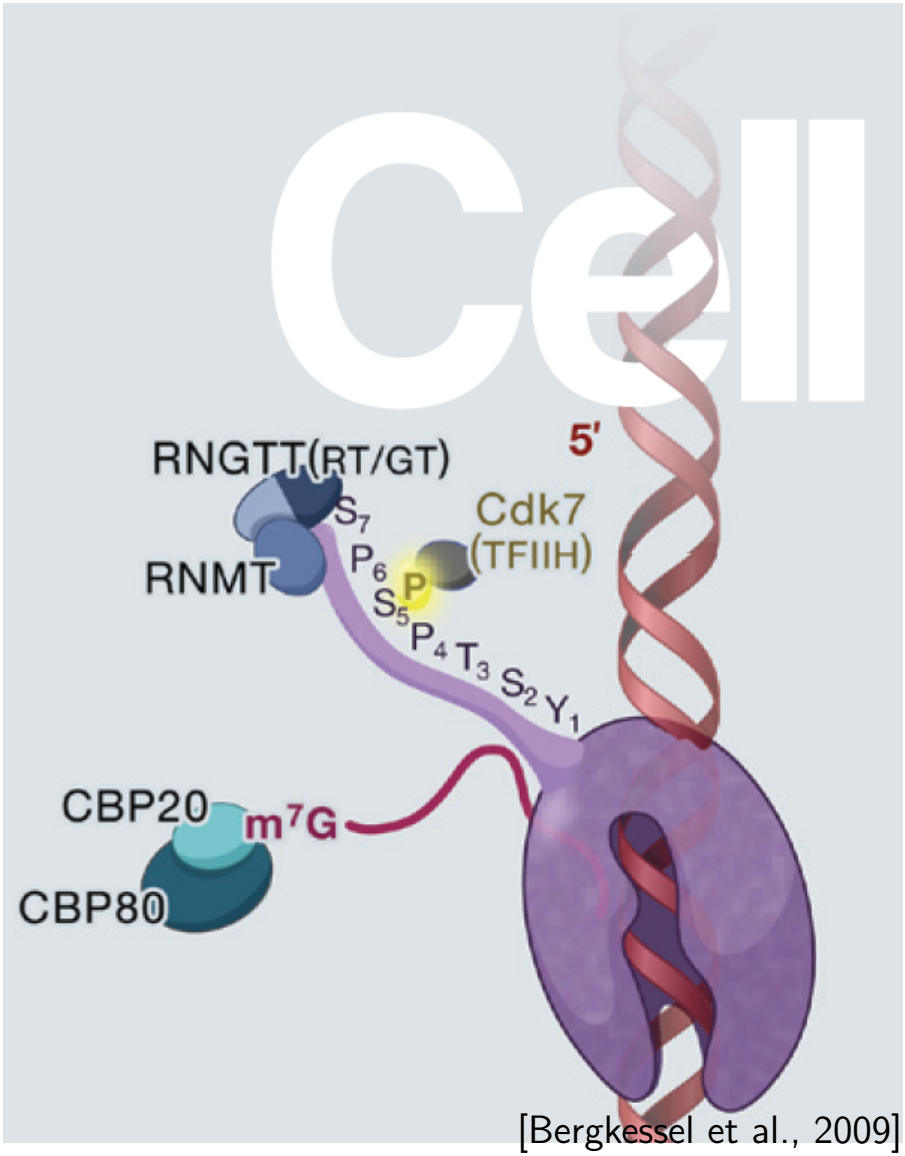
  [CBP20 & CBP80: cap-binding proteins]

- 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.



[Bergkessel et al., 2009]

Transcription & RNA Processing

- Newly synthesized pre-mRNA is capped.

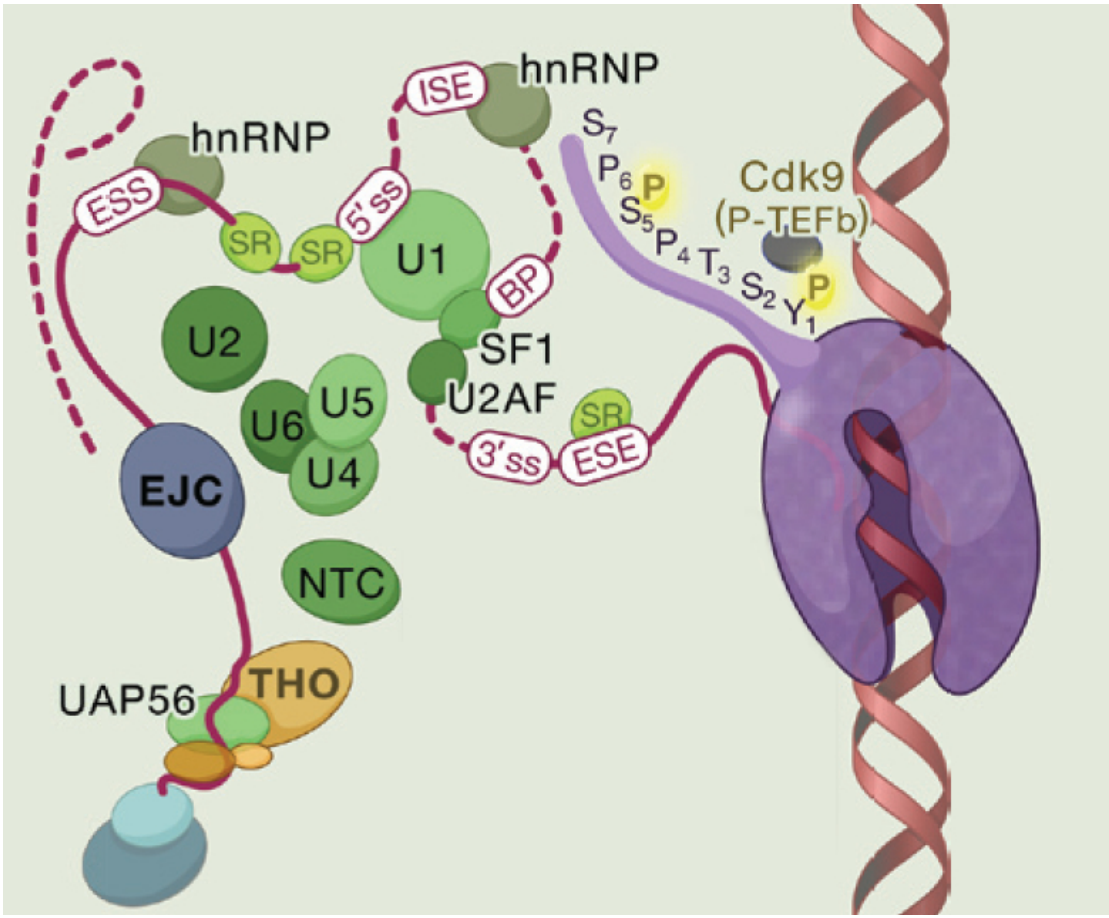
  [CBP20 & CBP80: cap-binding proteins]

- 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.



[Bergkessel et al., 2009]

Transcription & RNA Processing

- Newly synthesized pre-mRNA is capped.

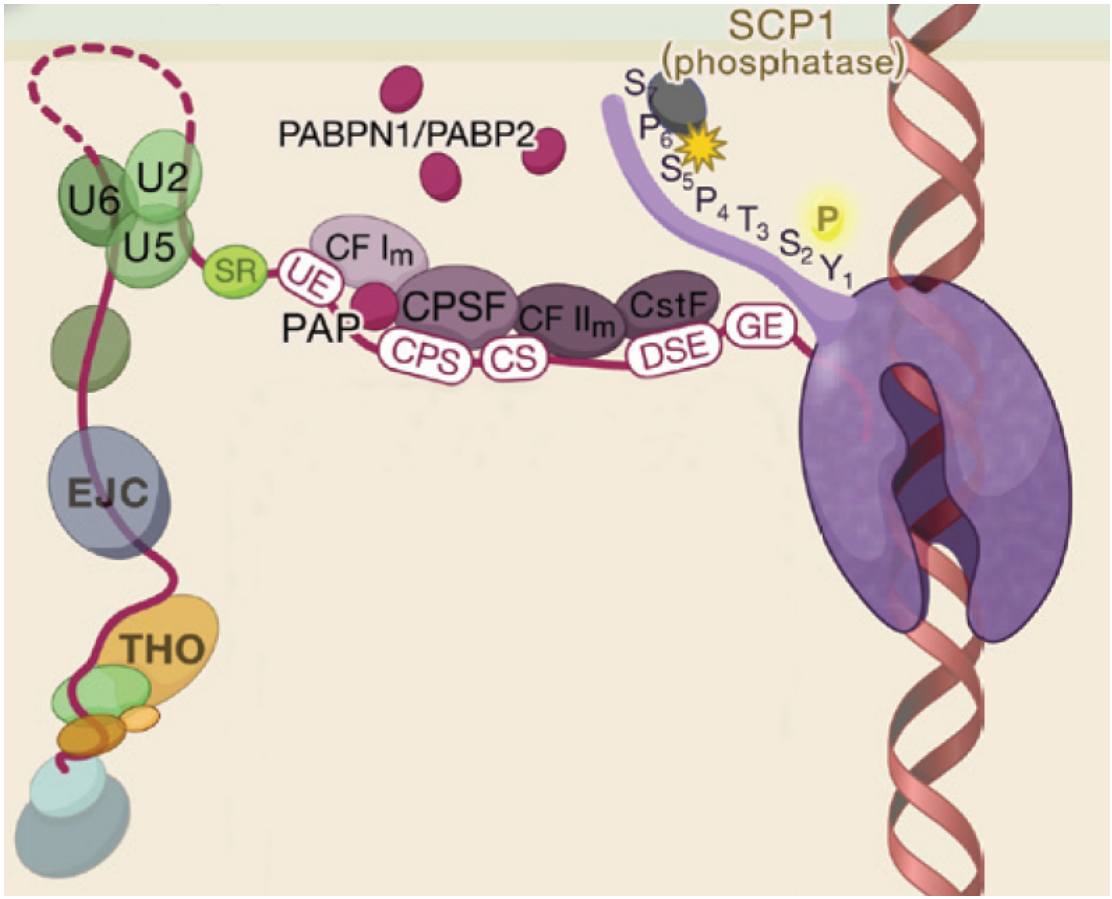
  [CBP20 & CBP80: cap-binding proteins]

- 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.



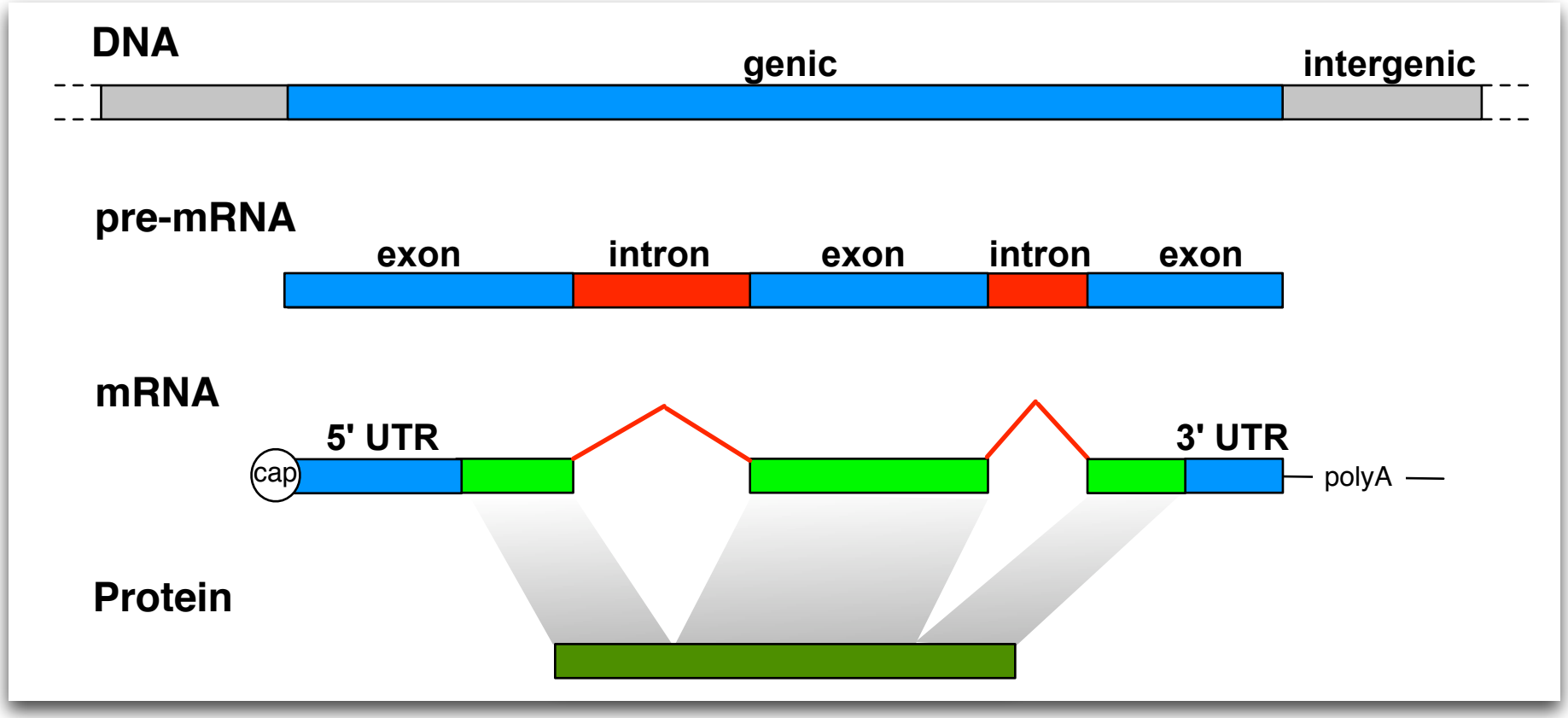
[Bergkessel et al., 2009]

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

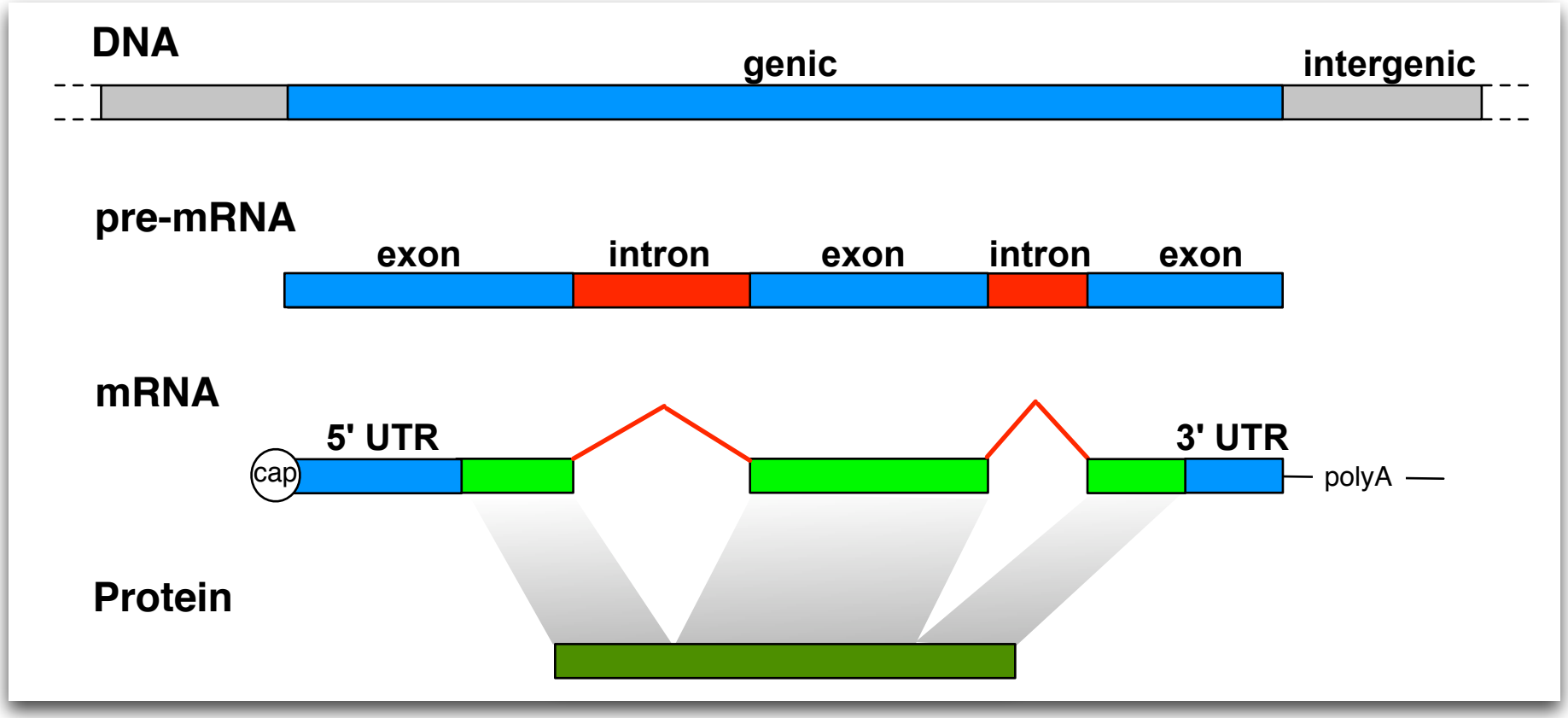
⇝ Labeling the Genome



- Given a piece of DNA sequence
- Predict protein-coding mRNAs
  - Less well developed for non-coding RNAs

Computational Gene Finding

⇝ Labeling the Genome

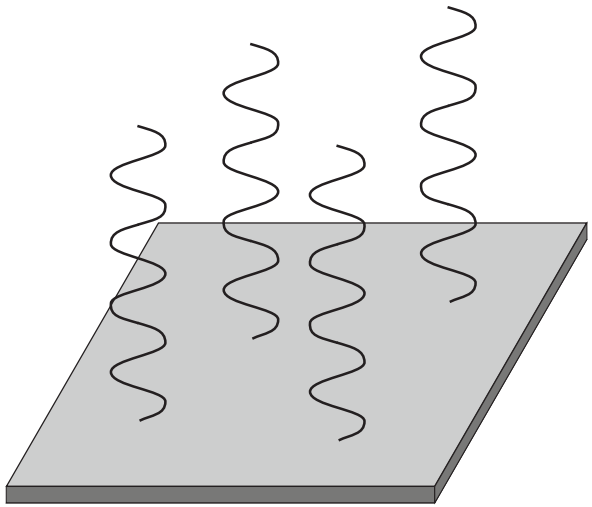


- Given a piece of DNA sequence
- Predict protein-coding mRNAs
  - Less well developed for non-coding RNAs

Experimental Characterization of the Transcriptome

DNA Microarrays

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



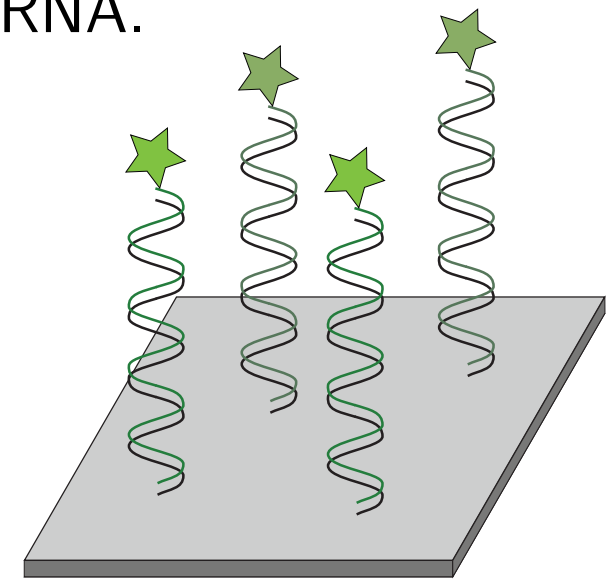
[Wikipedia]

cDNA Sequencing

Experimental Characterization of the Transcriptome

DNA Microarrays

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



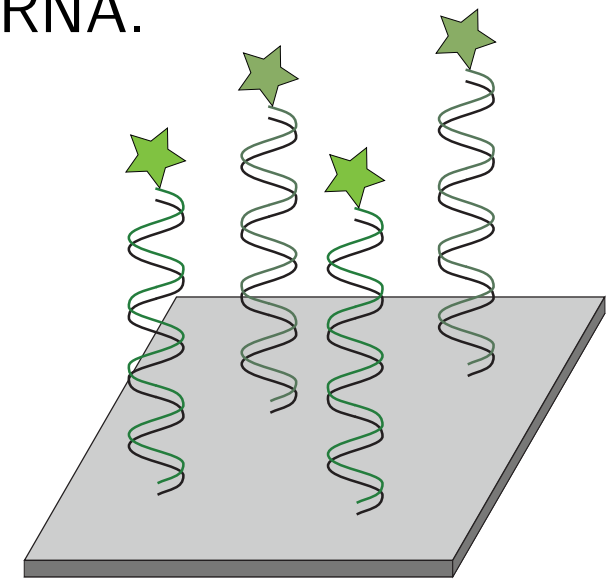
[Wikipedia]

cDNA Sequencing

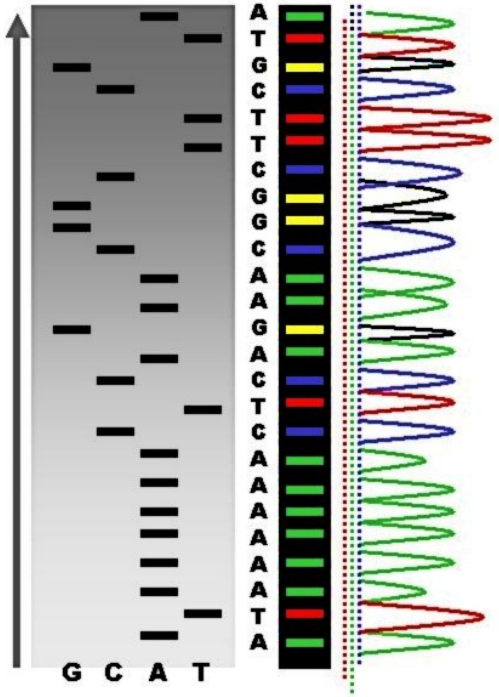
Experimental Characterization of the Transcriptome

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)

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)

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)

Key Research Questions

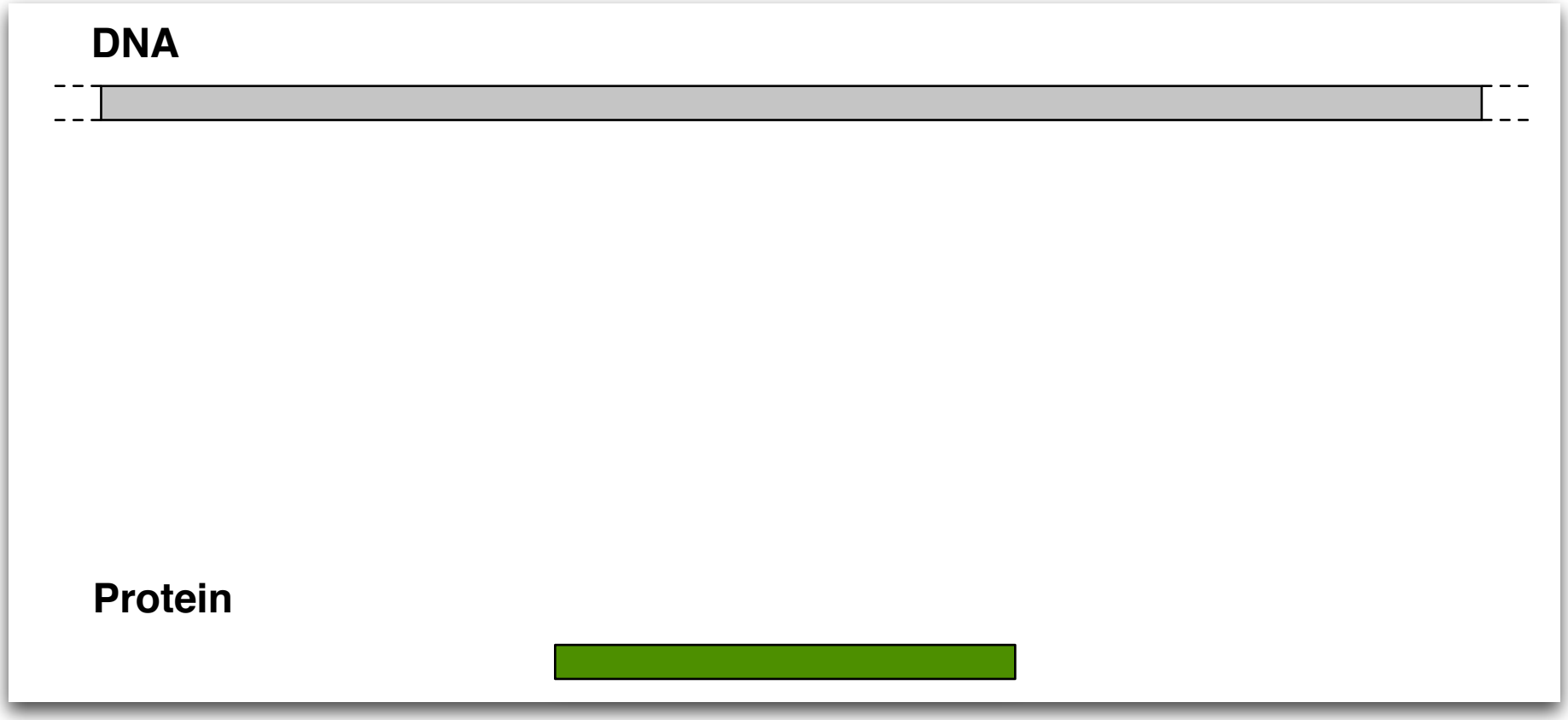
- Characterize an organism's full complement of genes
  - $\Rightarrow$ Find new (possibly noncoding) genes
  - $\Rightarrow$ Compare genes among organisms
- Characterize transcript isoforms
  - $\Rightarrow$ Find new alternative splice forms / transcript ends
- Monitor transcriptome changes between 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)

Roadmap

1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Identification of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology & Limitations
   - Assembly & Read Mapping
4. Extensions
   - Quantification of Transcripts
   - ChIP-on-Chip Studies

Computational Gene Finding

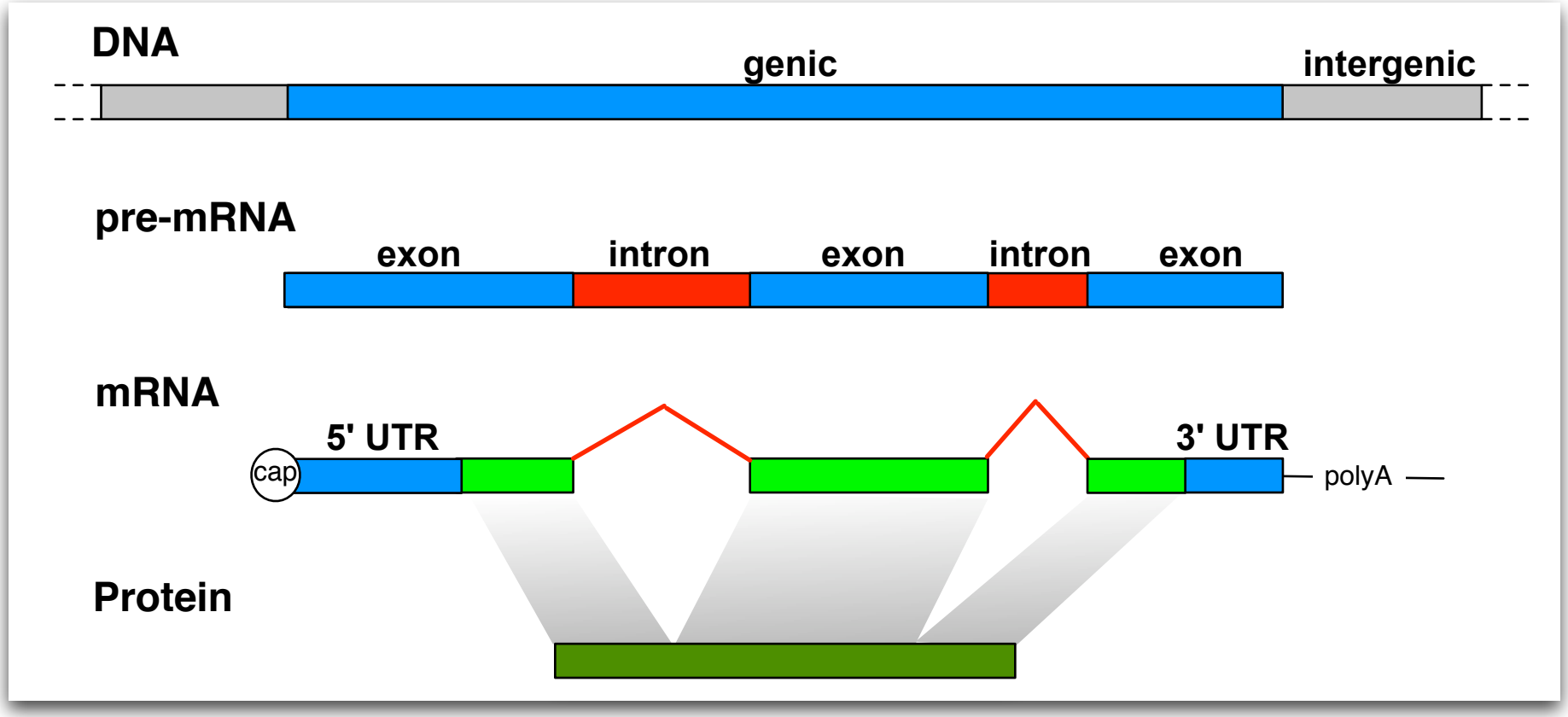
⇝ Labeling the Genome



- 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 "intergenic", "exon", "intron", "5' UTR", etc.

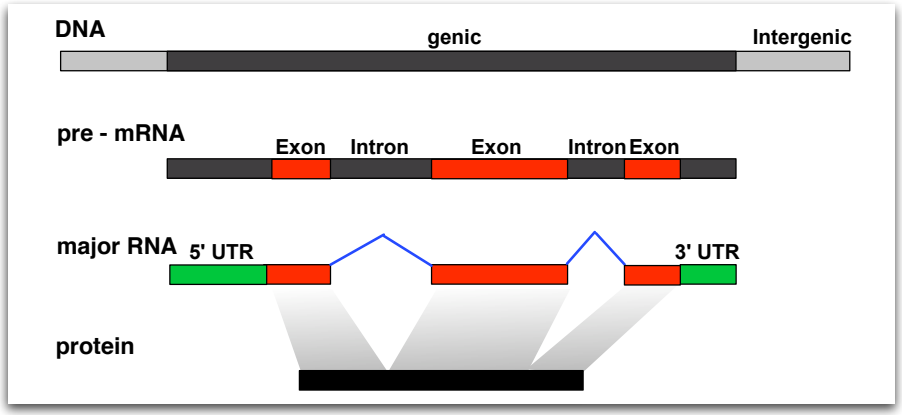
Hidden Markov Models

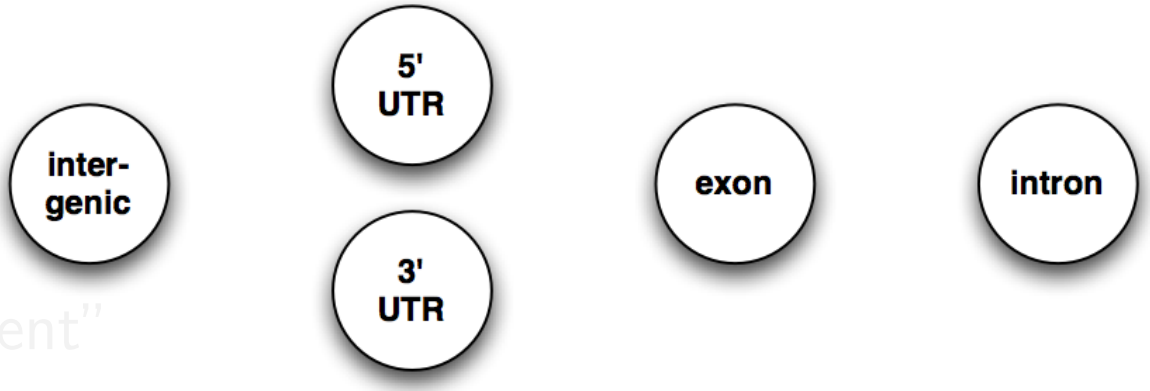
Model sequence content:

- One state per segment type
  - Allow only plausible transitions
  - Content statistics at each state
    - Derived from known genes

Prediction:

- Given DNA, find most likely state sequences
- Focuses on "content"
- Weak models for "signals"

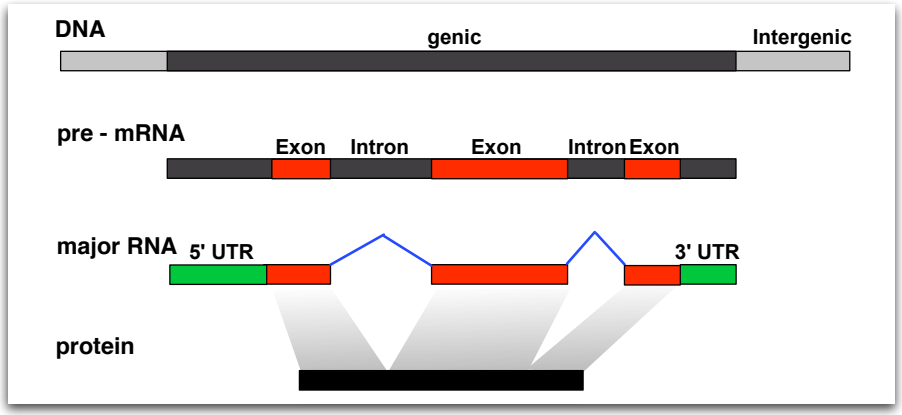




Hidden Markov Models

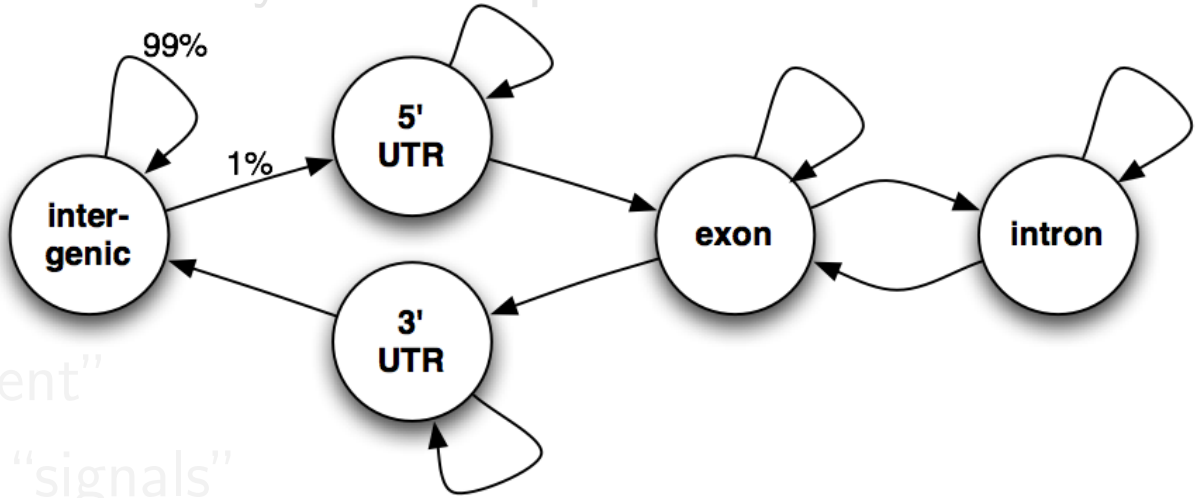
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences

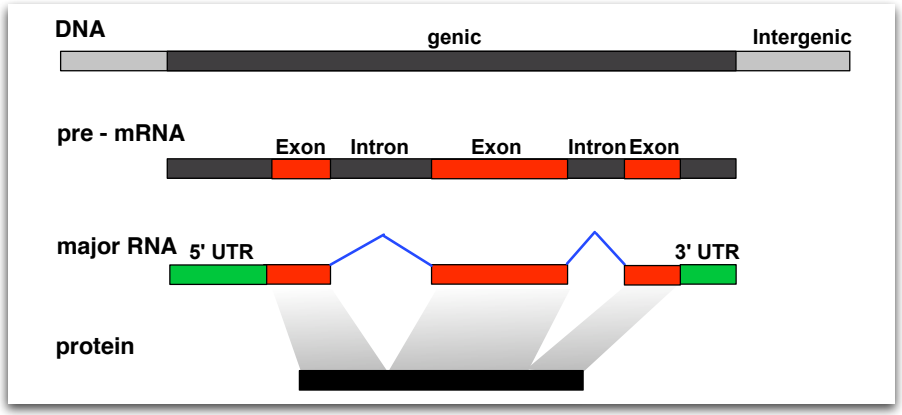


- Focuses on "content"
- Weak models for "signals"

Hidden Markov Models

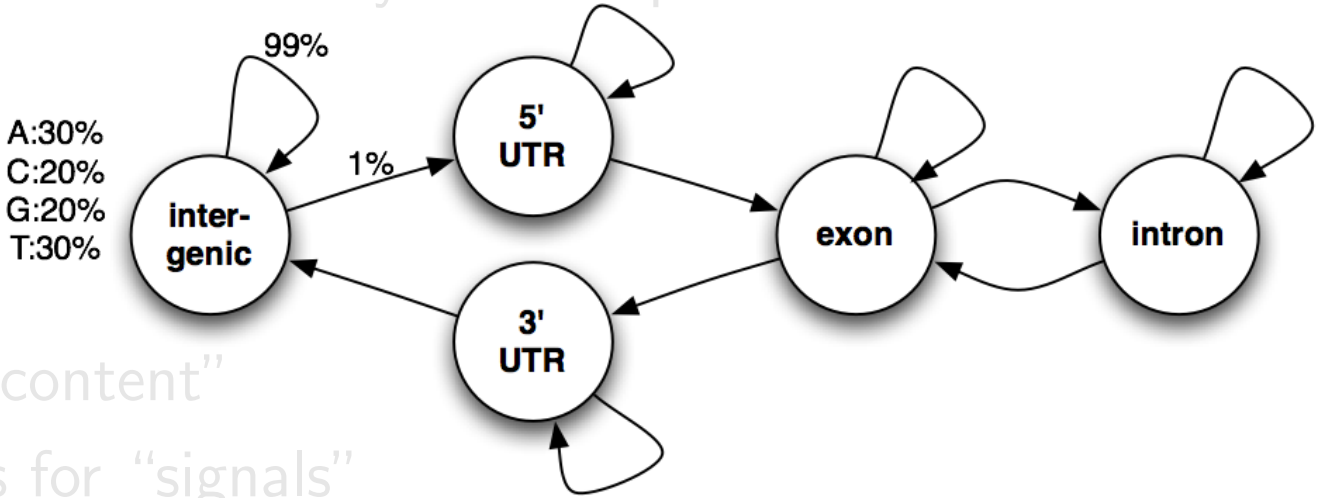
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences

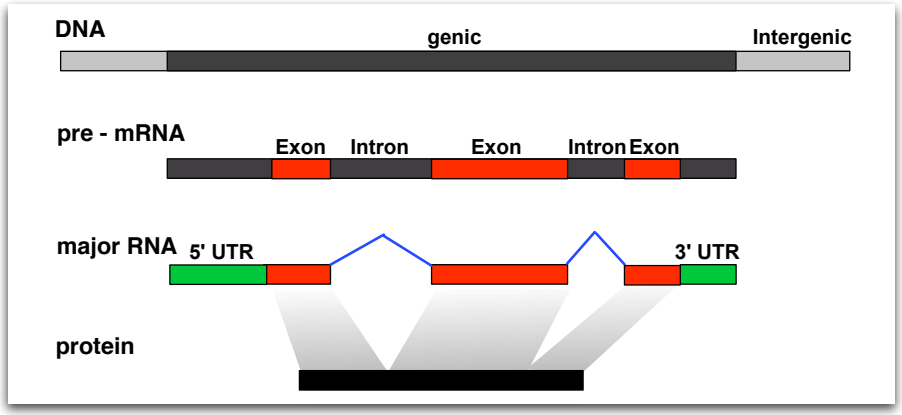


- Focuses on "content"
- Weak models for "signals"

Hidden Markov Models

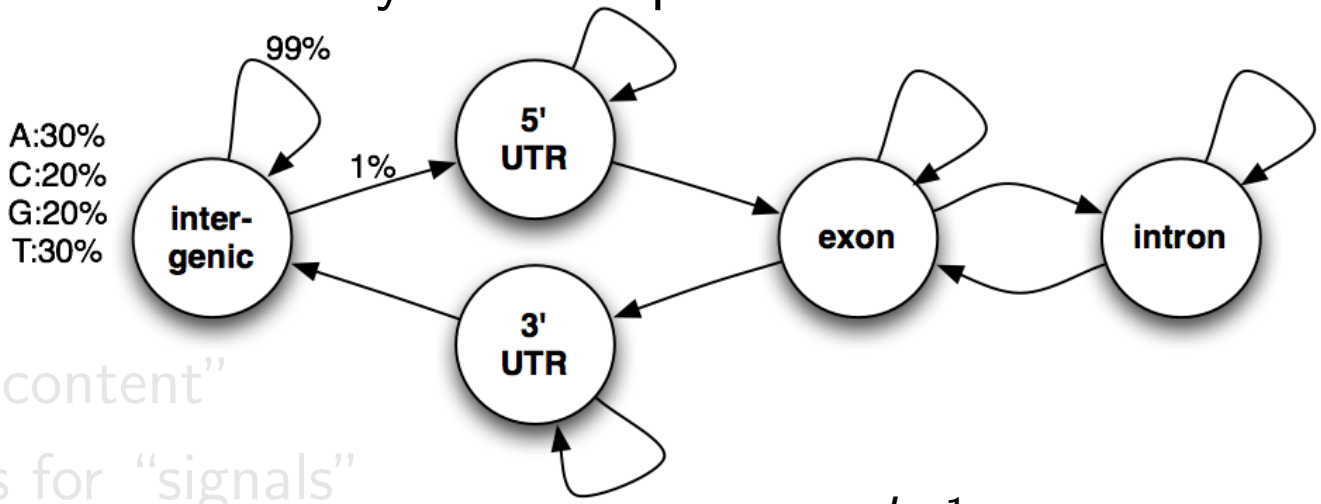
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences



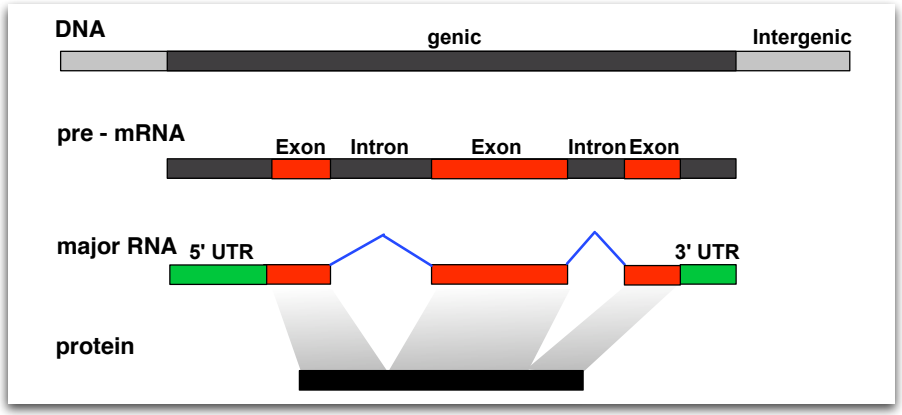
- Focuses on "content"
- Weak models for "signals"

$$p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{L-1} p(x_i|y_i)p(y_{i+1}|y_i)$$

Hidden Markov Models

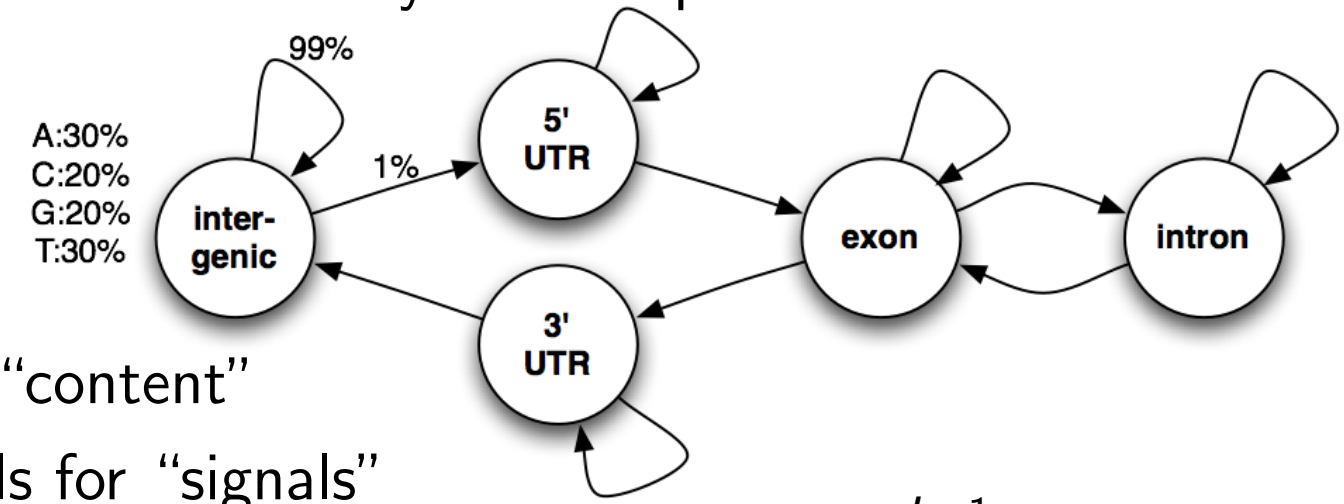
Model sequence content:

- One state per segment type
- Allow only plausible transitions
- Content statistics at each state
  - Derived from known genes



Prediction:

- Given DNA, find most likely state sequences

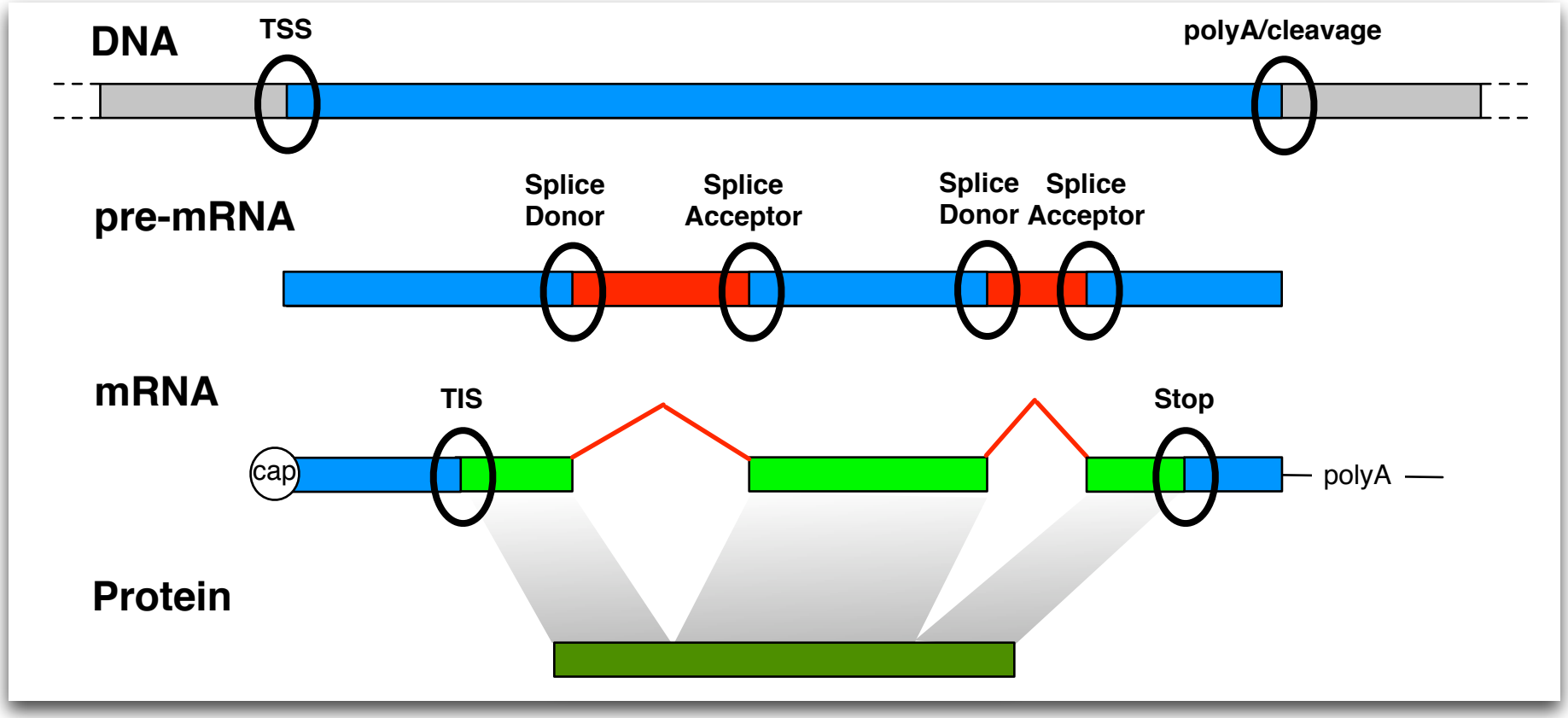


- Focuses on "content"
- Weak models for "signals"

$$p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{L-1} p(x_i|y_i)p(y_{i+1}|y_i)$$

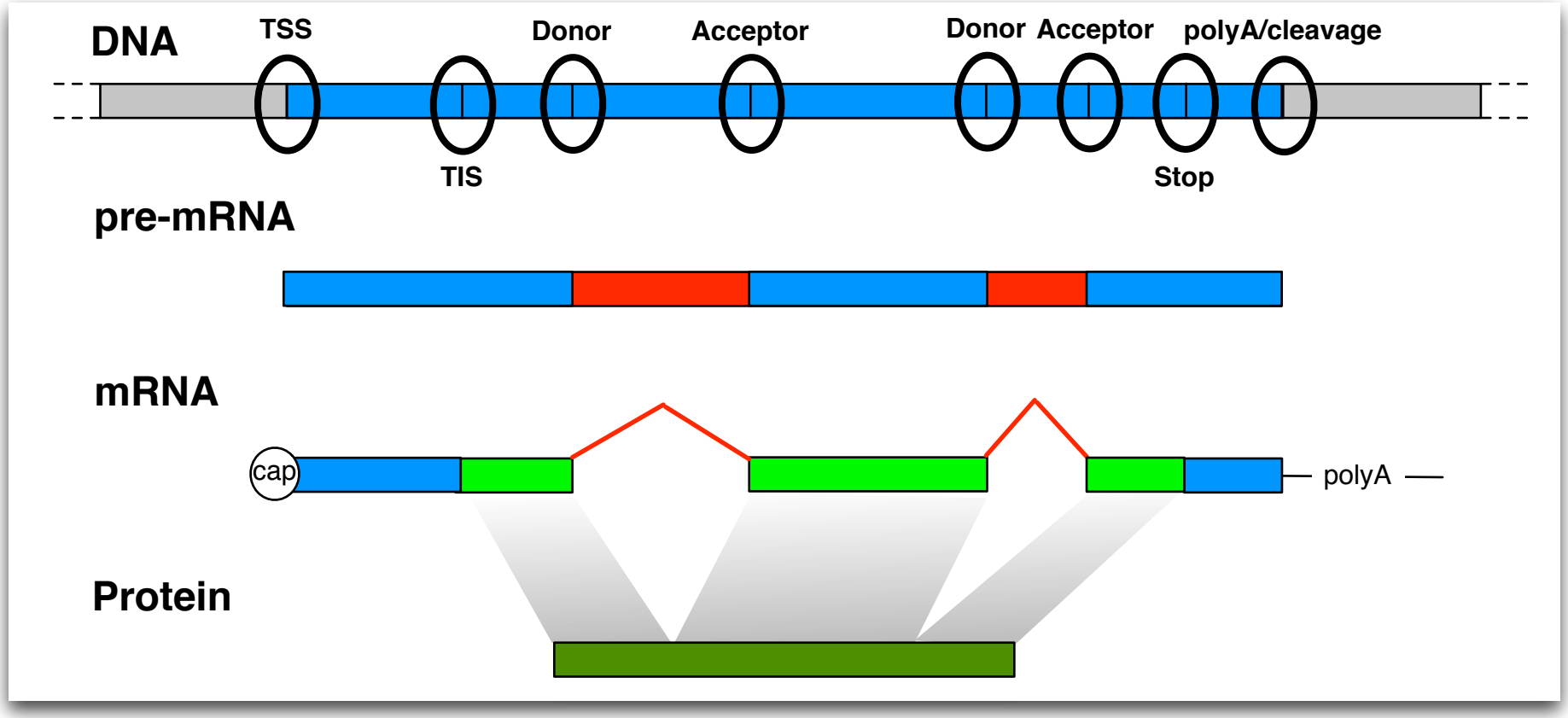
Computational Gene Finding

⇝ Labeling the Genome



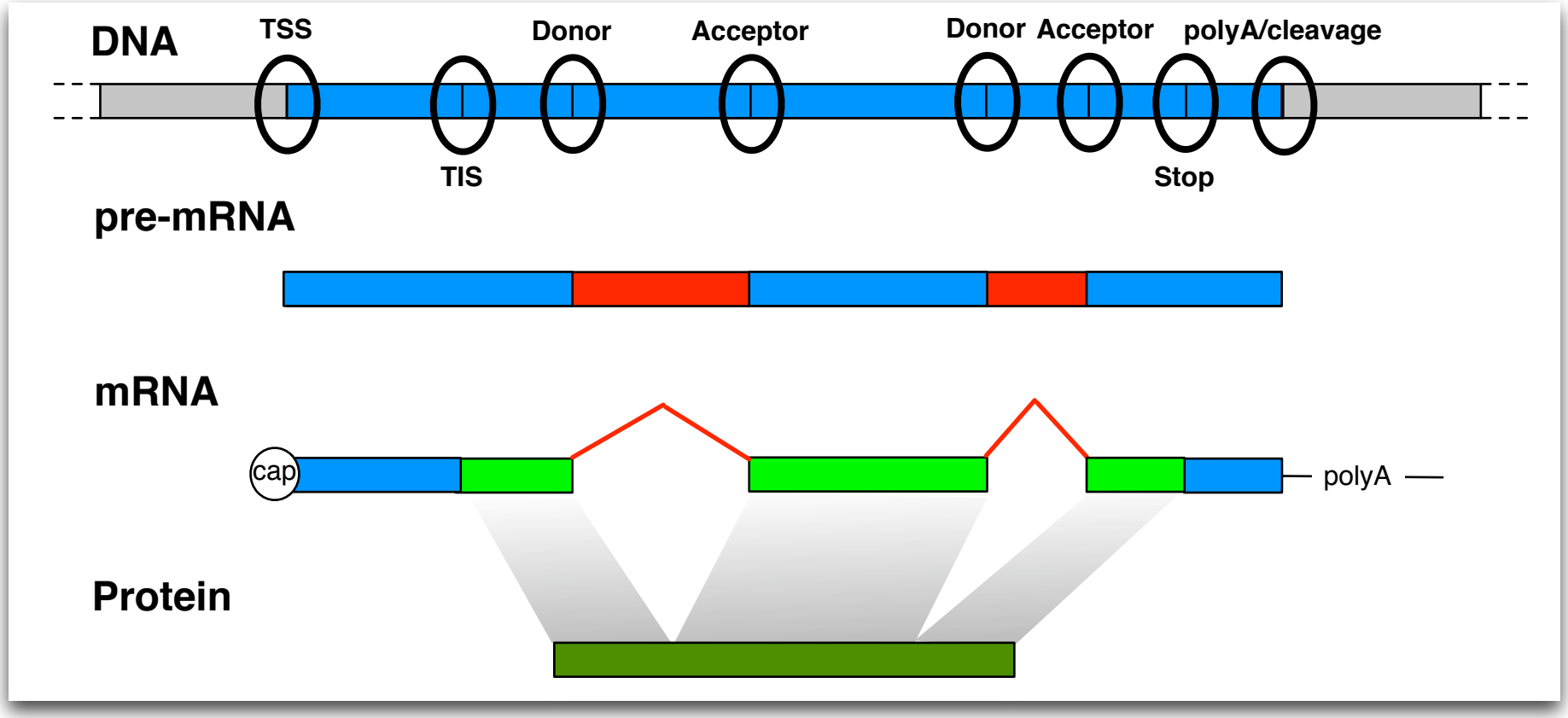
Computational Gene Finding

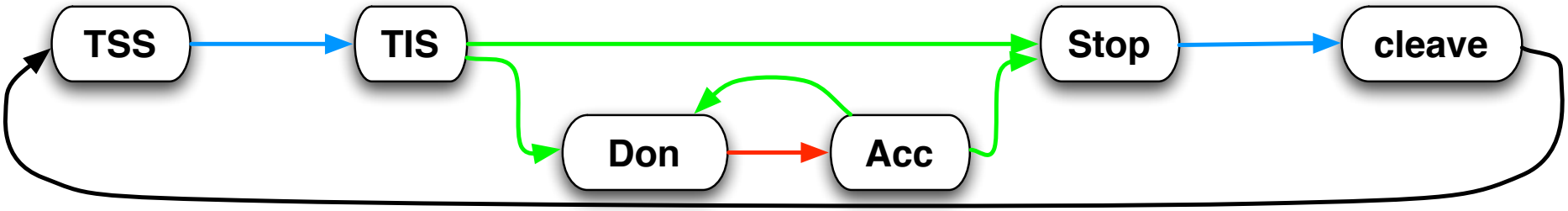
⇝ Labeling the Genome



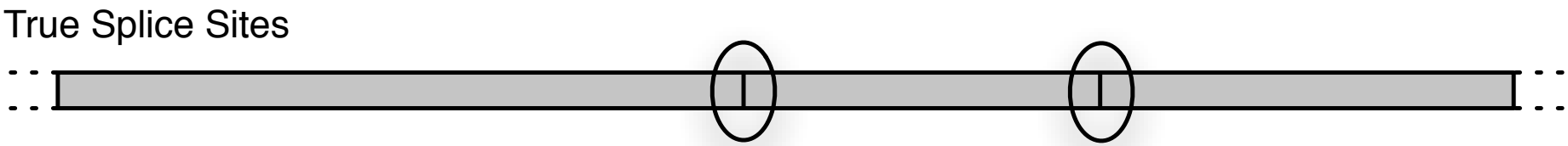
Computational Gene Finding

⇝ Labeling the Genome





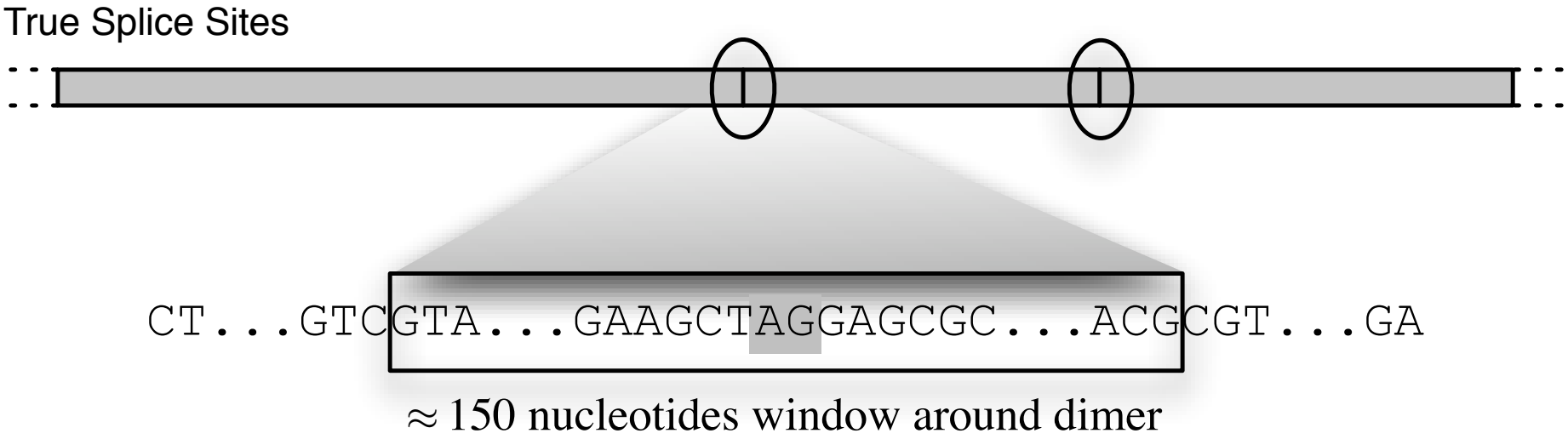
Example: Splice Site Recognition



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

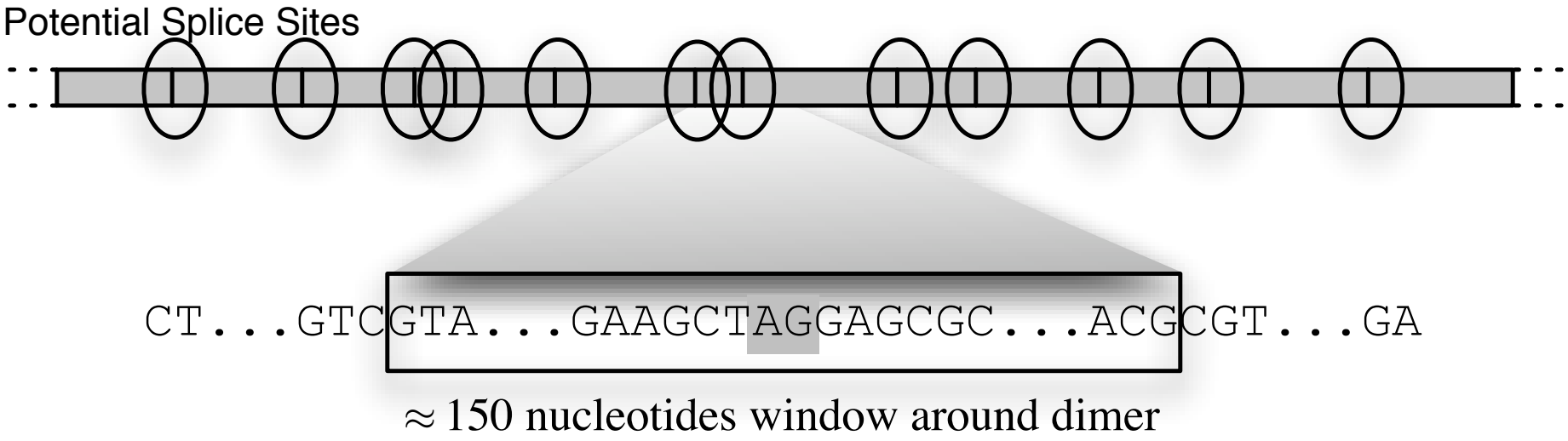
$\Rightarrow$ **Millions of labeled instances from EST databases**

Example: Splice Site Recognition



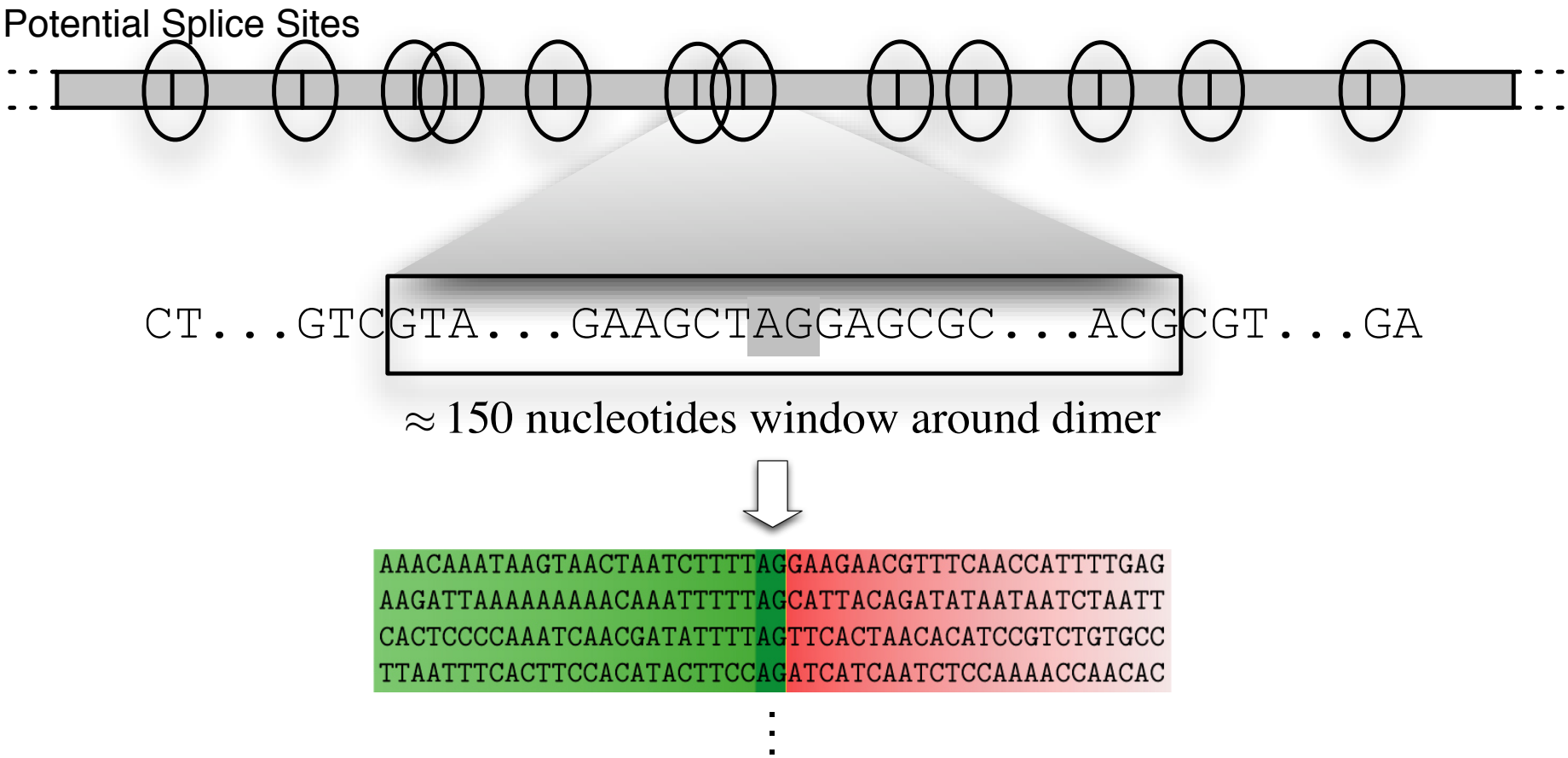
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

Example: Splice Site Recognition



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

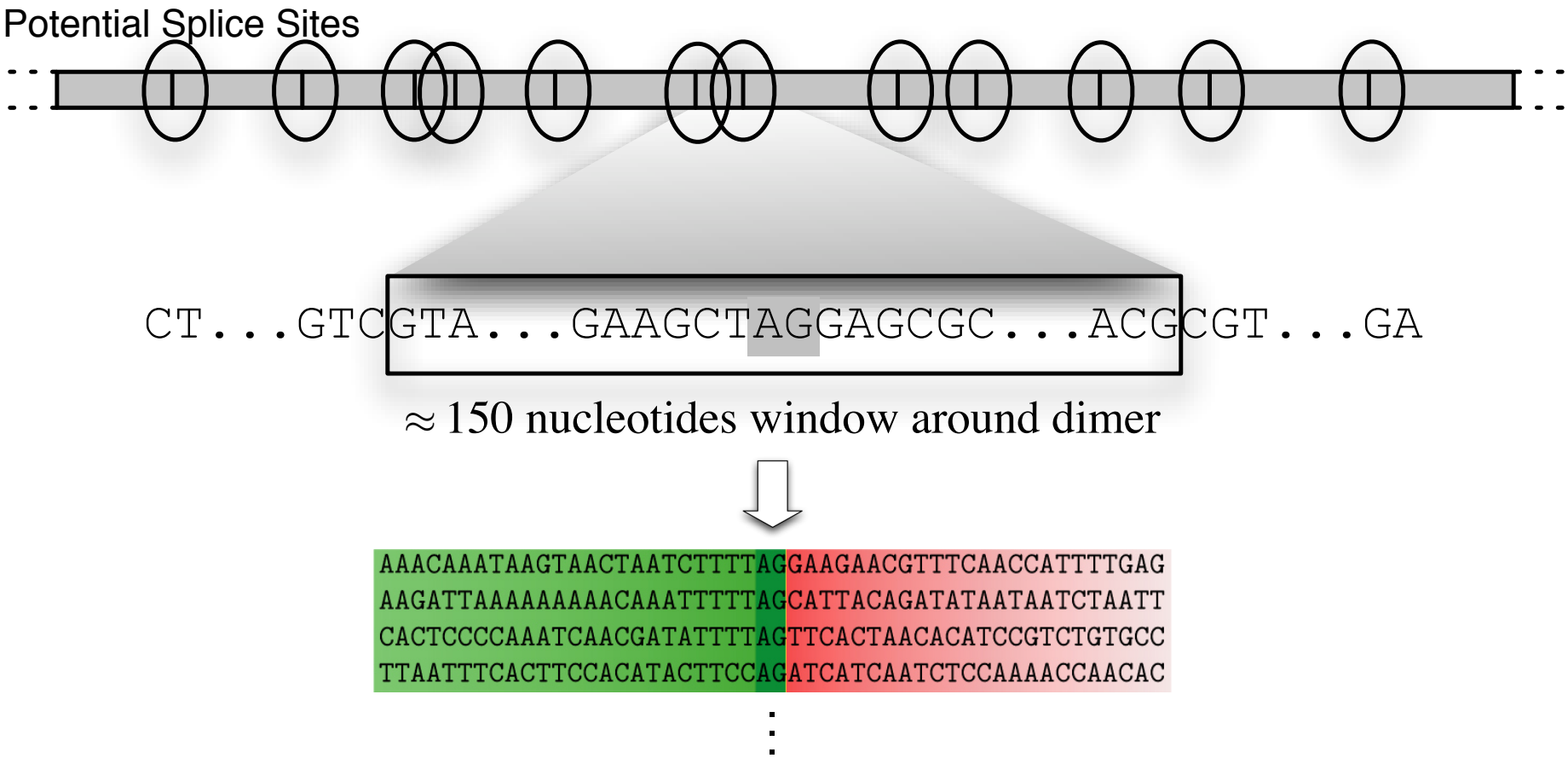
Example: Splice Site Recognition



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

$\Rightarrow$ Millions of labeled instances from EST databases

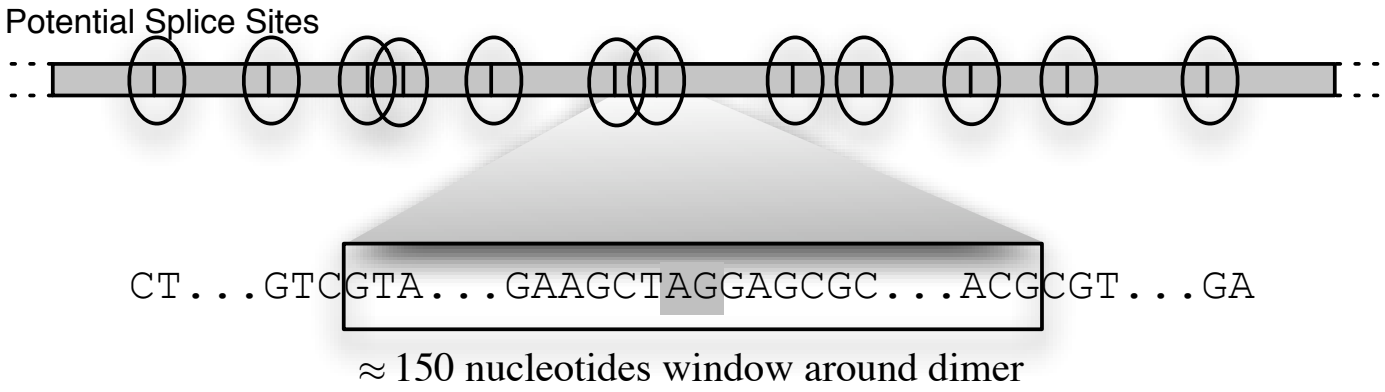
Example: Splice Site Recognition



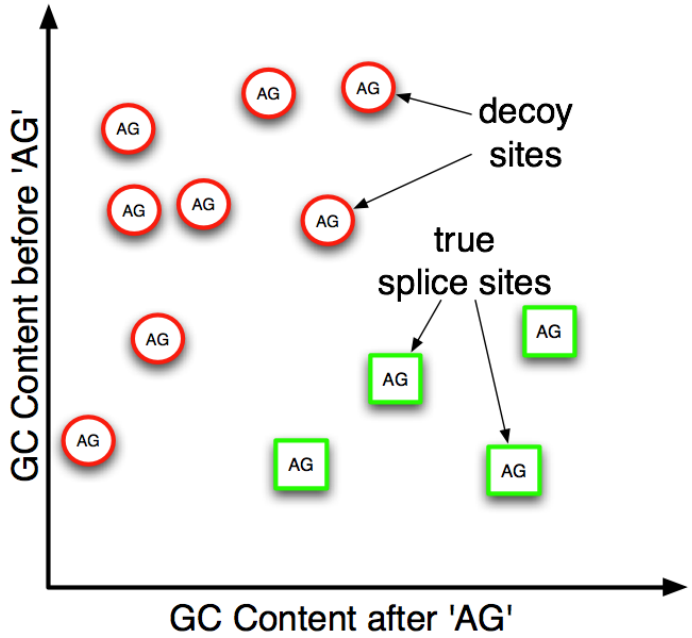
- True sites: fixed window around a true splice site
- Decoy sites: all other consensus sites

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

Example: Splice Site Recognition



Basic idea:



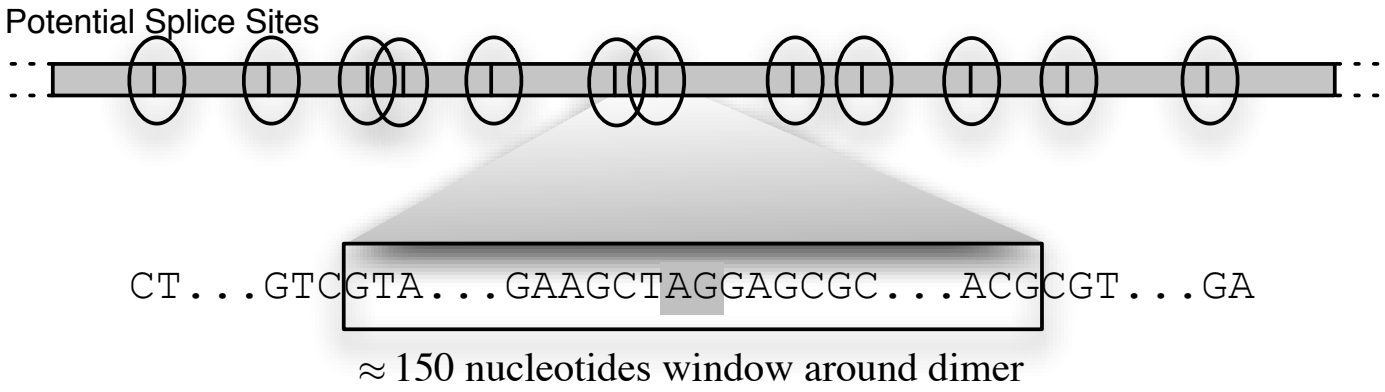
For instance, exploit that exons have higher GC content

or

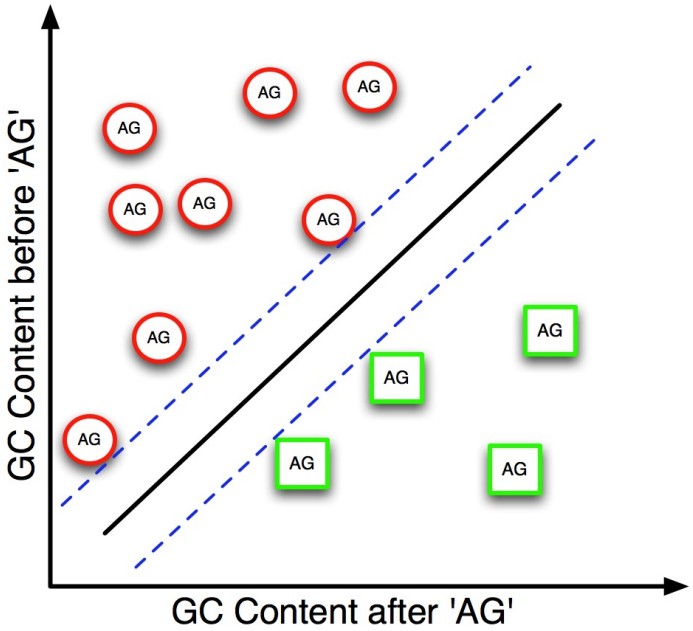
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

Example: Splice Site Recognition



Basic idea:



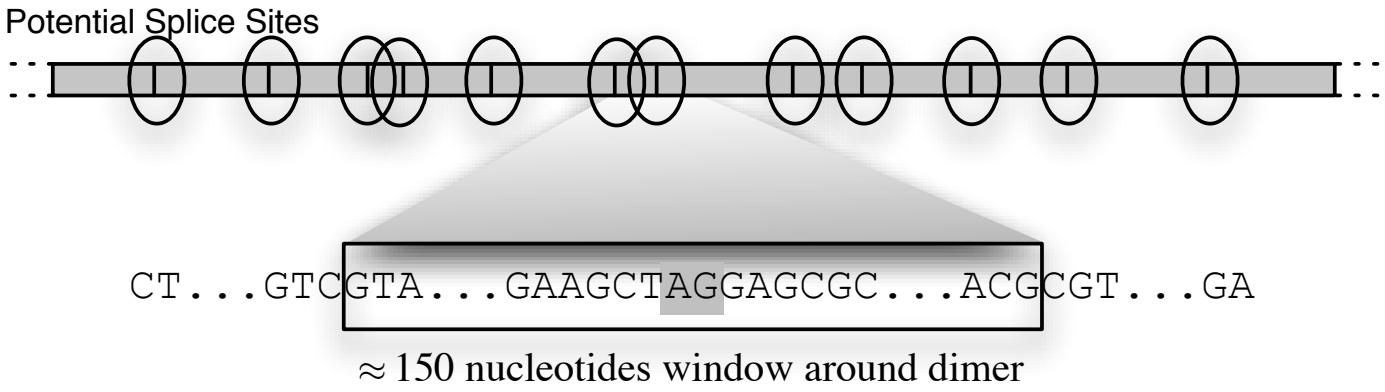
For instance, exploit that exons have higher GC content

or

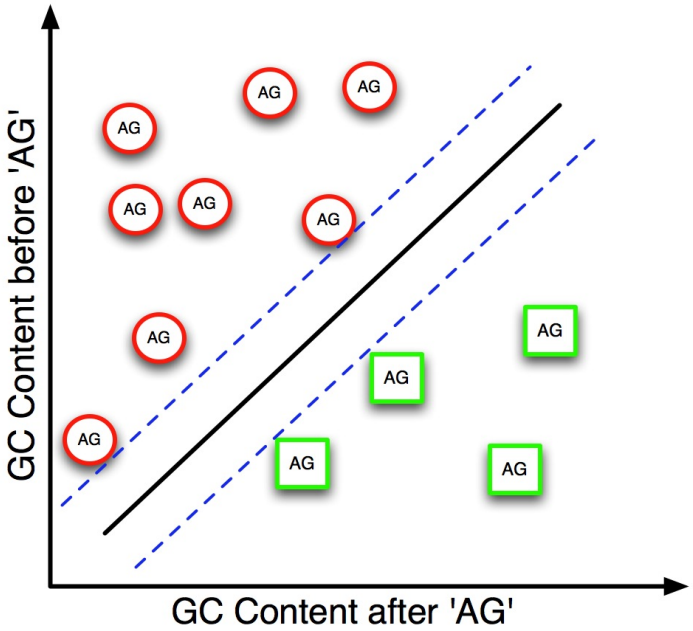
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

Example: Splice Site Recognition



Basic idea:



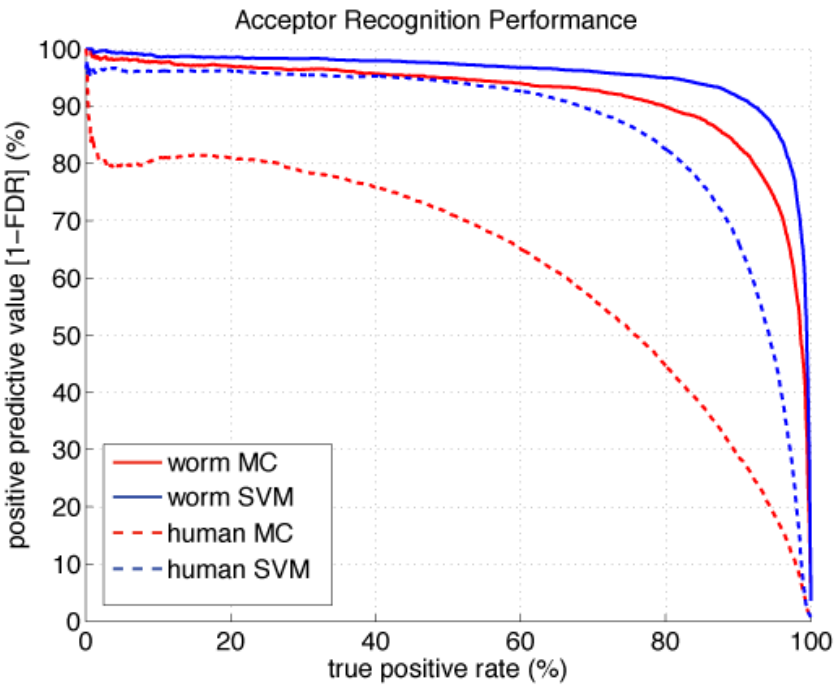
In practice: Use one feature per possible substring (e.g. $\leq$20) at all positions

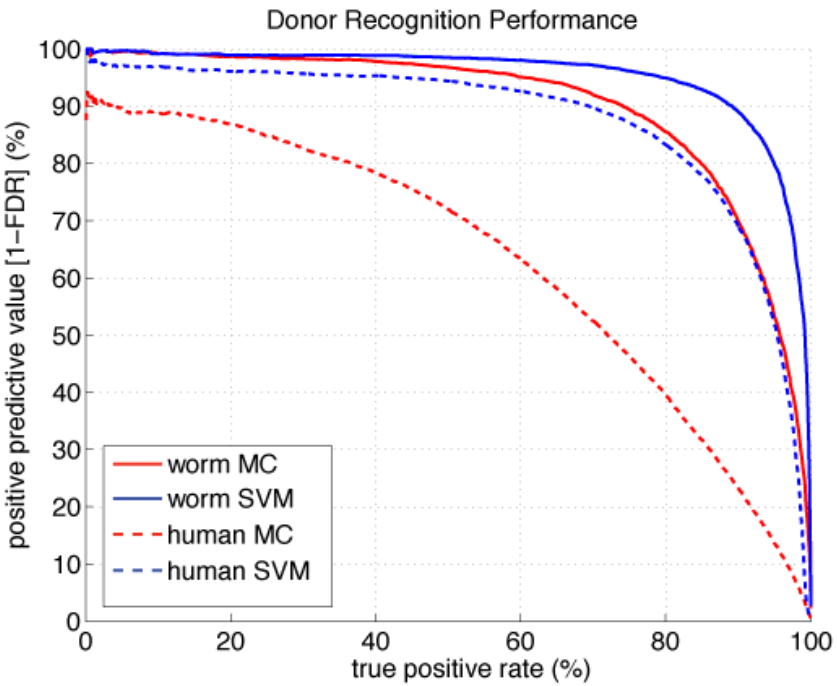
$150 \cdot (4^1 + \ldots + 4^{20}) \approx 2 \cdot 10^{14}$ features

[Sonnenburg et al., 2007]

Results on Splice Site Recognition

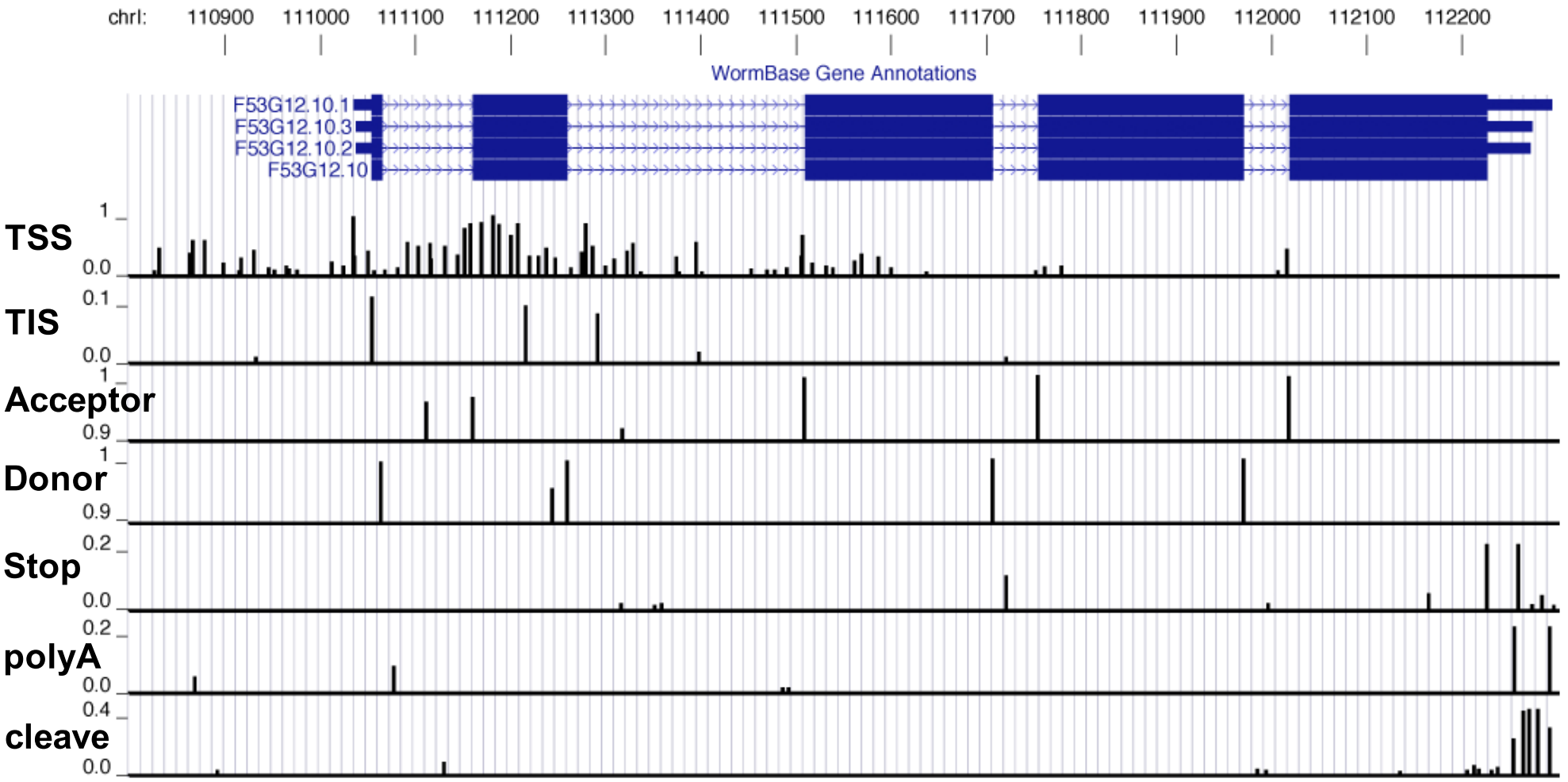
	Worm		Fly		Cress		Fish		Human	
	Acc	Don	Acc	Don	Acc	Don	Acc	Don	Acc	Don
Markov Chain auPRC(%)	92.1	90.0	80.3	78.5	87.4	88.2	63.6	62.9	16.2	26.0
SVM auPRC(%)	95.9	95.3	86.7	87.5	92.2	92.9	86.6	86.9	54.4	56.9





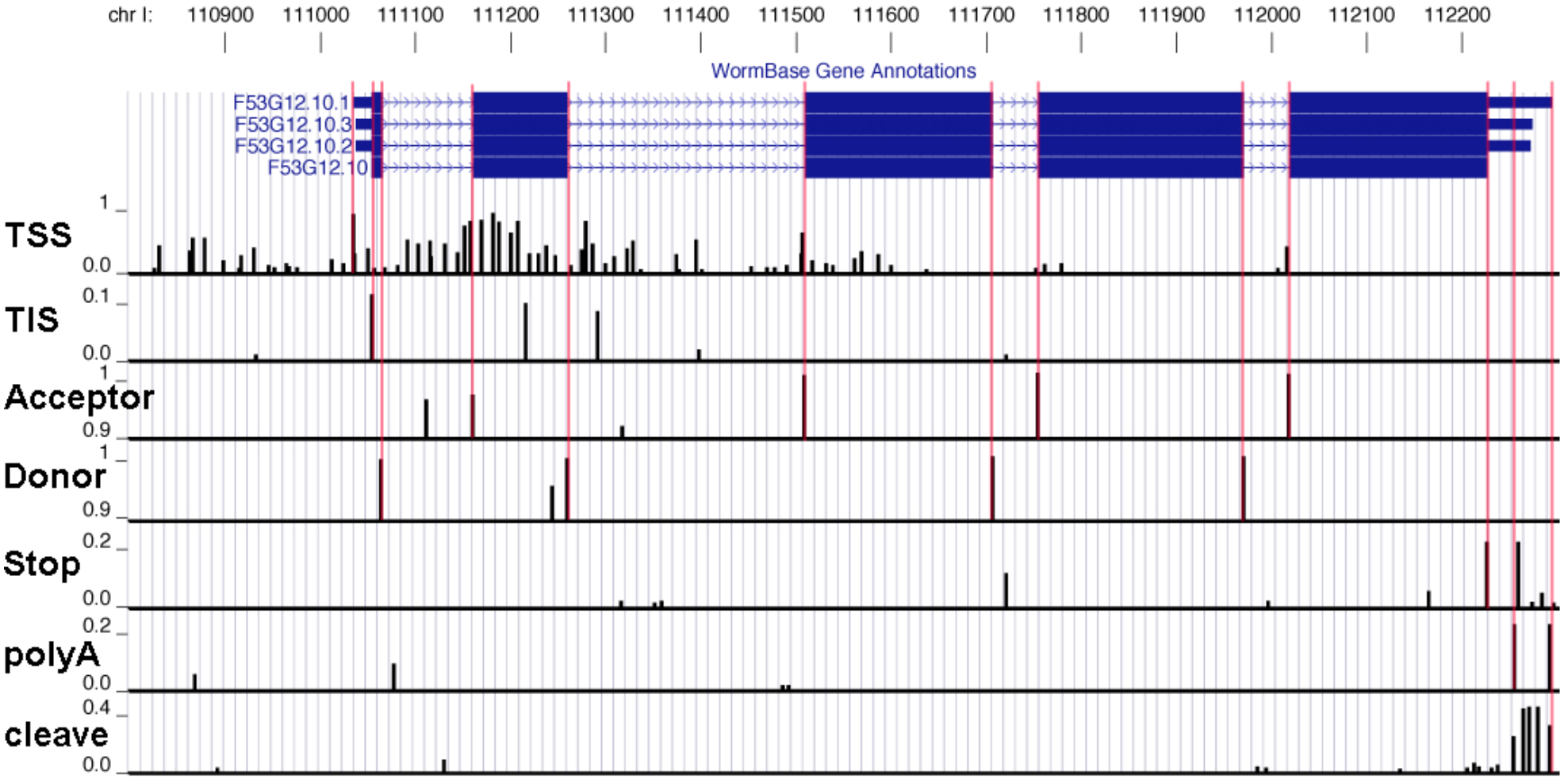
[Sonnenburg, Schweikert, Philips, Behr, Rätsch, 2007]

Example: Predictions in UCSC Browser



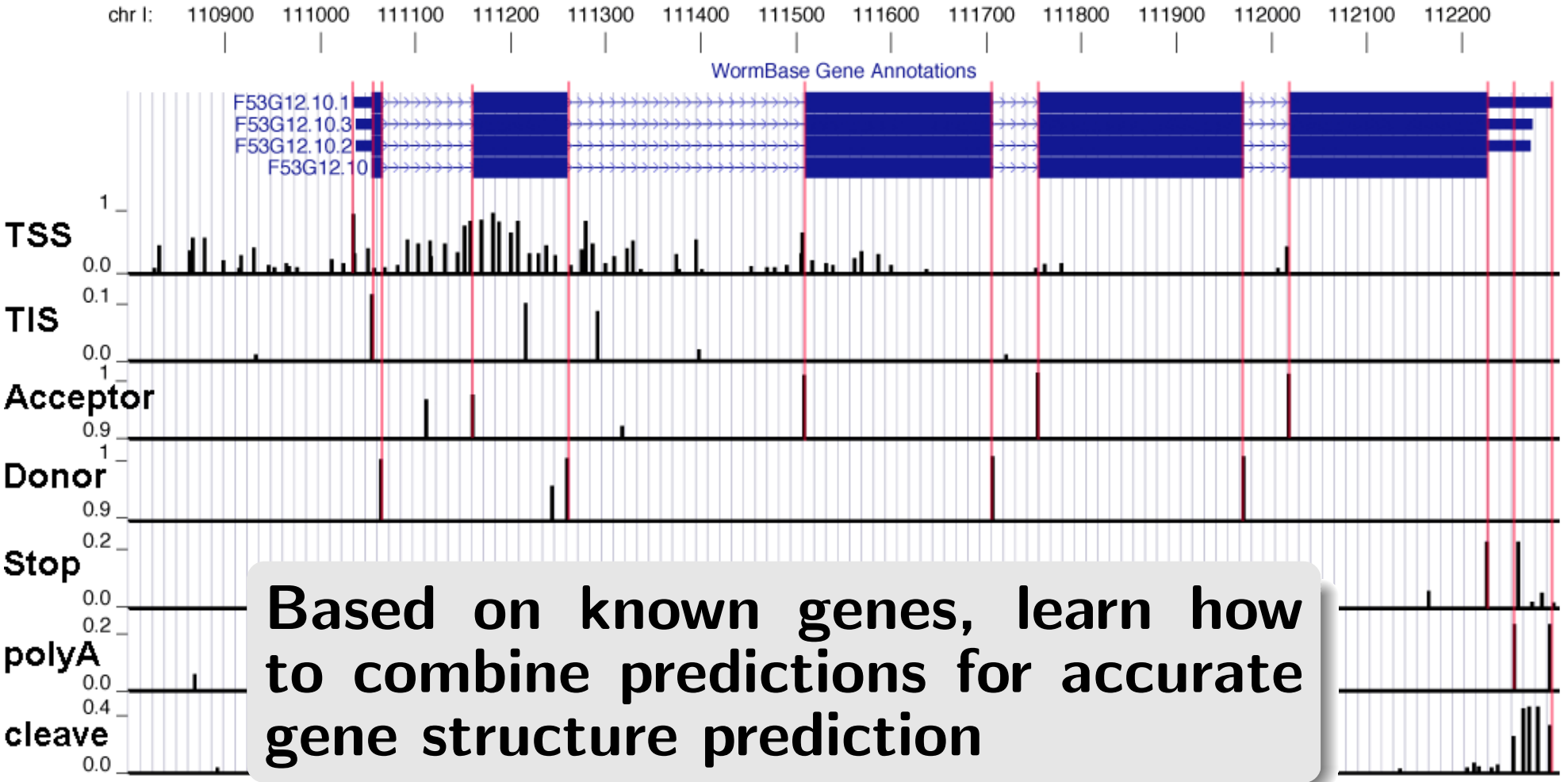
[Schweikert et al., 2009]

Example: Predictions in UCSC Browser



[Schweikert et al., 2009]

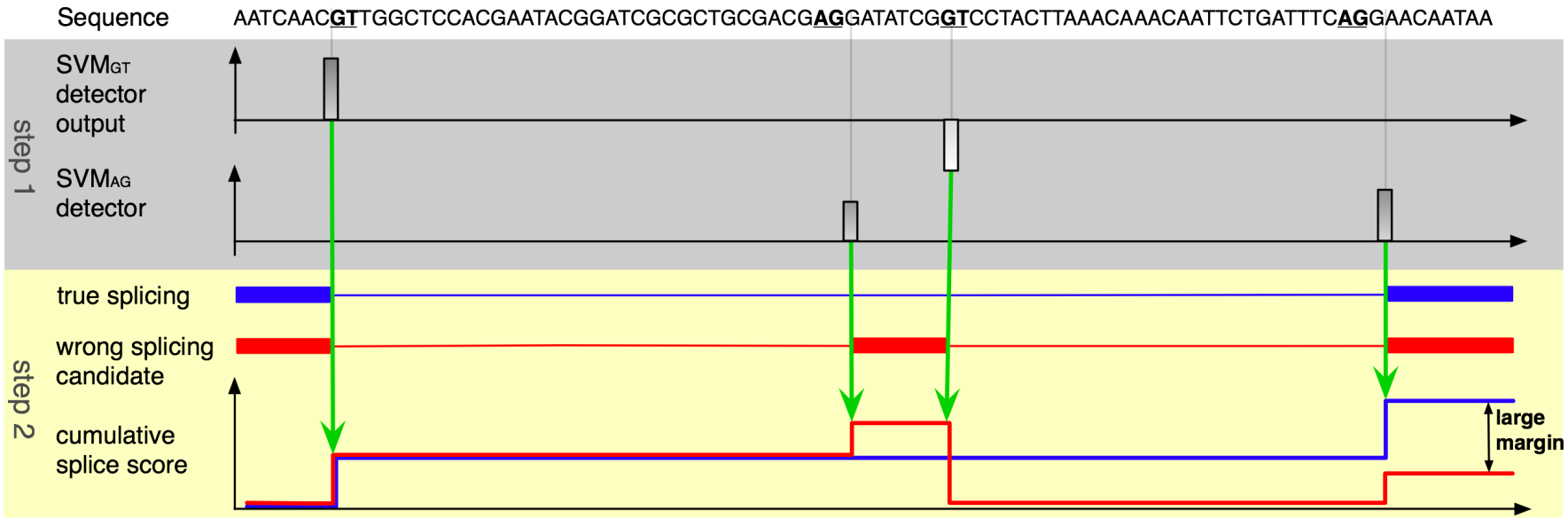
Example: Predictions in UCSC Browser



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

[Schweikert et al., 2009]

Discriminative Gene Prediction (simplified)



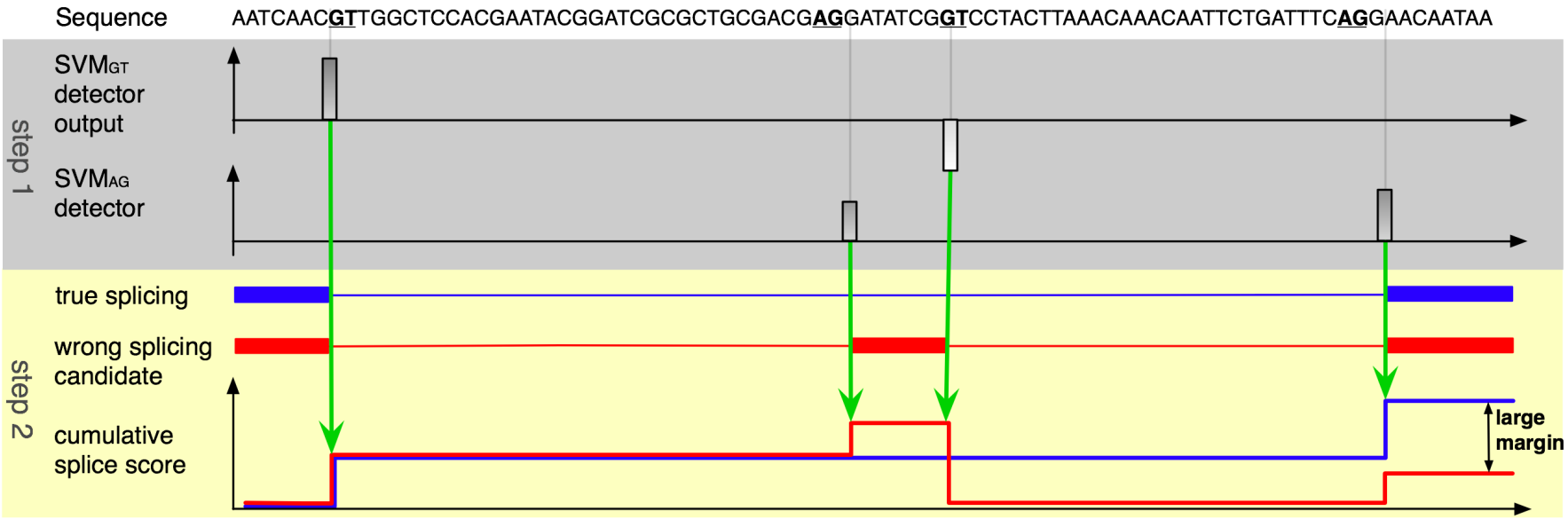
[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

$$F(\mathbf{y}) := \underbrace{\sum_{j=1}^{J-1} S_{GT}(f_j^{GT}) + \sum_{j=2}^{J} S_{AG}(f_j^{AG})}_{\text{Splice signals}} + \underbrace{\sum_{j=1}^{J-1} S_{L_I}(p_{j+1} - q_j) + \sum_{j=1}^{J} S_{L_E}(q_j - p_j)}_{\text{Segment lengths}}$$

Tune free parameters (in functions $S_{GT}, S_{AG}, S_{L_E}, S_{L_I}$) by solving **linear program** using training set with known splice forms

Discriminative Gene Prediction (simplified)



[Rätsch, Sonnenburg, Srinivasan, Witte, Müller, Sommer, Schölkopf, 2007]

Simplified Model: Score for splice form $\mathbf{y} = \{(p_j, q_j)\}_{j=1}^{J}$:

$$F(\mathbf{y}) := \underbrace{\sum_{j=1}^{J-1} S_{GT}(f_j^{GT}) + \sum_{j=2}^{J} S_{AG}(f_j^{AG})}_{\text{Splice signals}} + \underbrace{\sum_{j=1}^{J-1} S_{L_I}(p_{j+1} - q_j) + \sum_{j=1}^{J} S_{L_E}(q_j - p_j)}_{\text{Segment lengths}}$$

Tune free parameters (in functions $S_{GT}, S_{AG}, S_{L_E}, S_{L_I}$) by solving **linear program** using training set with known splice forms

Results using mGene

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

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

- Annotation of other nematode genomes: [Schweikert et al., 2009]

Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge (*C. elegans*) [Coghlan et al., 2008]

- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009]

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

- Annotation of other nematode genomes: [Schweikert et al., 2009]

Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

Results using mGene

- Most accurate *ab initio* method in the nGASP genome annotation challenge (*C. elegans*) [Coghlan et al., 2008]

- Validation of gene predictions for *C. elegans*: [Schweikert et al., 2009]

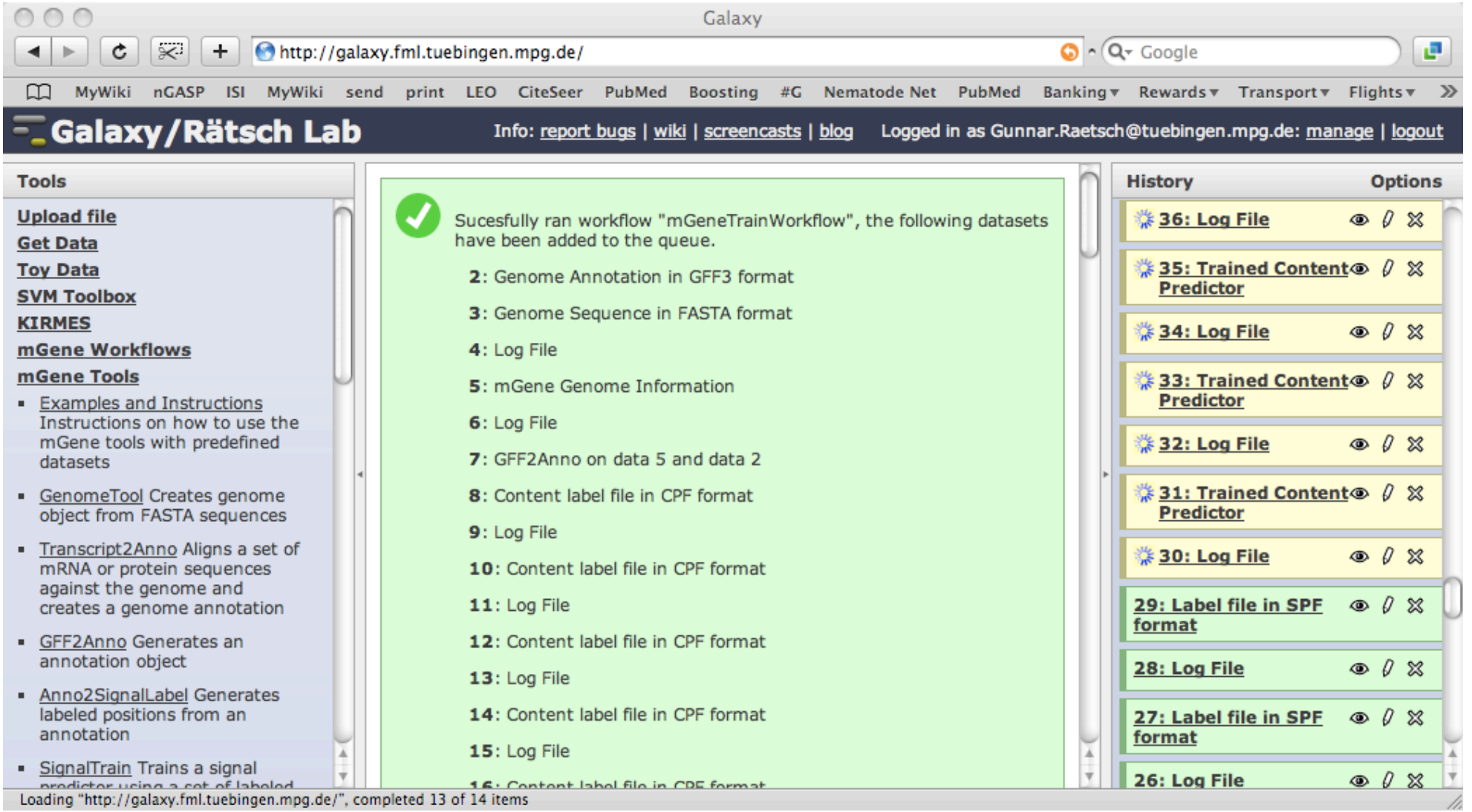
	No. of genes	No. of genes analyzed	Frac. of genes w/ expression
New genes	2,197	57	$\approx 42\%$
Missing unconf. genes	205	24	$\approx 8\%$

- Annotation of other nematode genomes: [Schweikert et al., 2009]

Genome	Genome size [Mbp]	No. of genes	No. exons/gene (mean)	mGene accuracy	best other accuracy
C. remanei	235.94	31503	5.7	**96.6%**	93.8%
C. japonica	266.90	20121	5.3	**93.3%**	88.7%
C. brenneri	453.09	41129	5.4	**93.1%**	87.8%
C. briggsae	108.48	22542	6.0	**87.0%**	82.0%

mGene.web: Gene Finding for Everybody ;-)

(Schweikert et al., 2009)



http://mgene.org/webservice

Limitations/Extensions

Gene finding accuracy still far from perfect

- Misses genes, predicts incorrect gene models
- Does not (yet) predict alternative transcripts
- Cannot predict when transcripts are expressed/modified/degraded...

Need experimental data for condition specific transcriptomes.

Then we can learn to predict (hopefully).

Limitations/Extensions

Gene finding accuracy still far from perfect

- Misses genes, predicts incorrect gene models
- Does not (yet) predict alternative transcripts
- Cannot predict when transcripts are expressed/modified/degraded...

Need experimental data for condition specific transcriptomes.

Then we can learn to predict (hopefully).

Limitations/Extensions

Gene finding accuracy still far from perfect

- Misses genes, predicts incorrect gene models
- Does not (yet) predict alternative transcripts
- Cannot predict when transcripts are expressed/modified/degraded...

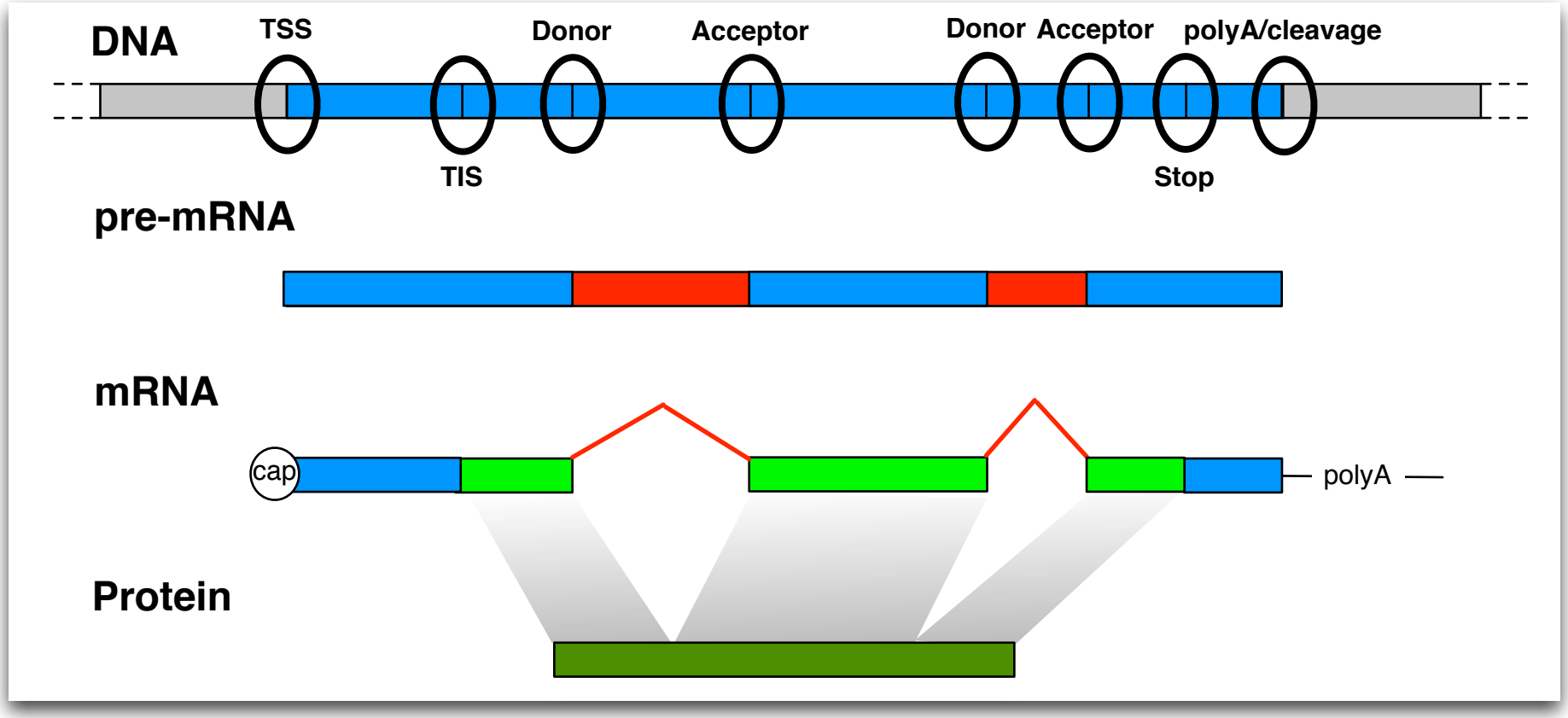
Need experimental data for condition specific transcriptomes.

Then we can learn to predict (hopefully).

Roadmap

1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Identification of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology & Limitations
   - Assembly & Read Mapping
4. Extensions
   - Quantification of Transcripts
   - ChIP-on-Chip Studies

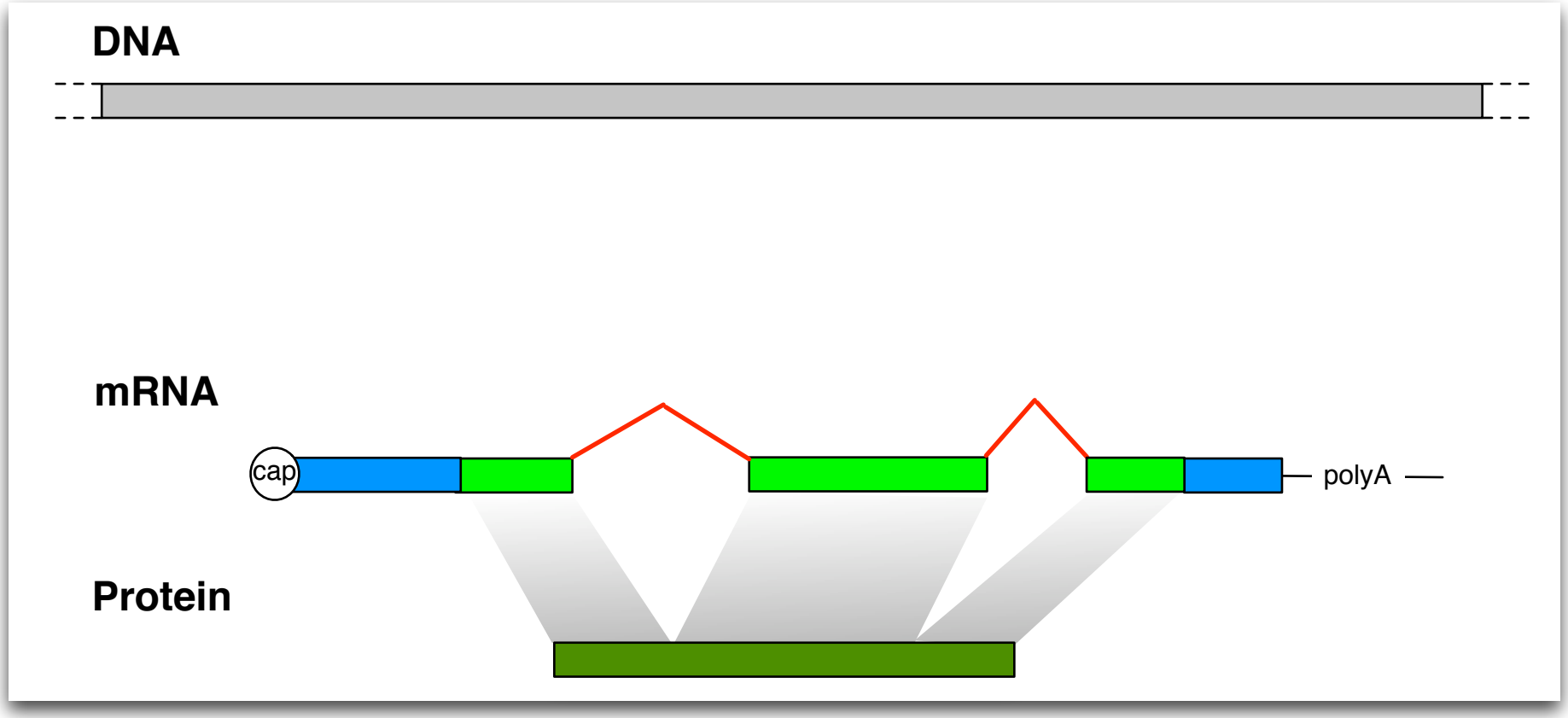
From Genome to Proteins etc.



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)

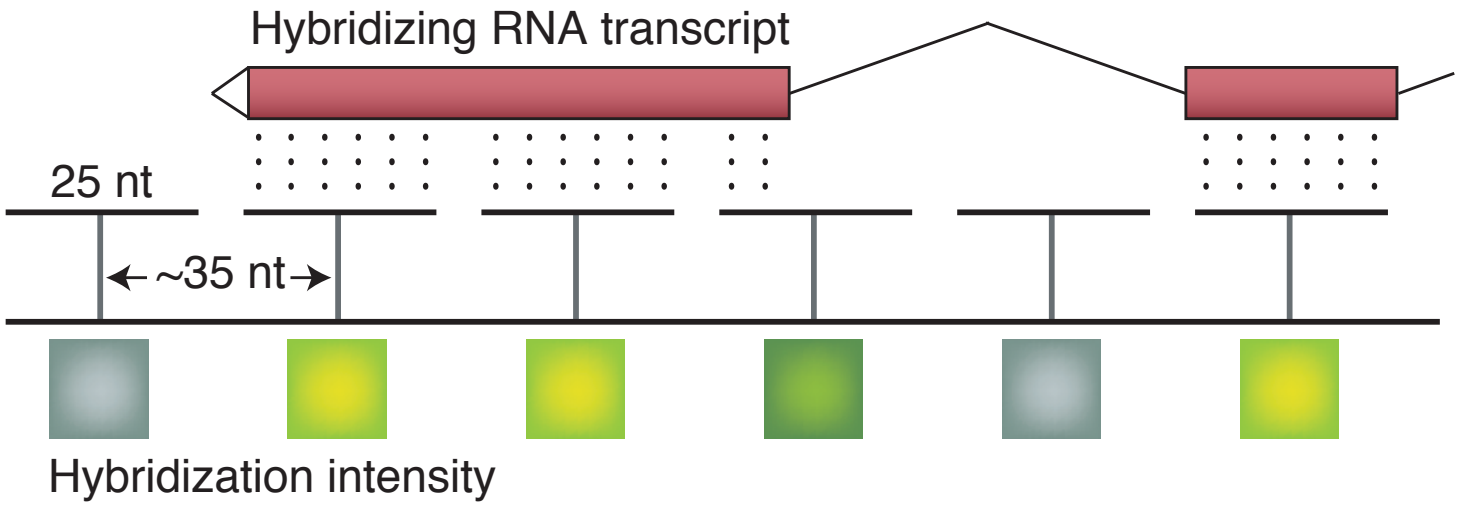
Whole-genome Tiling Arrays



- 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

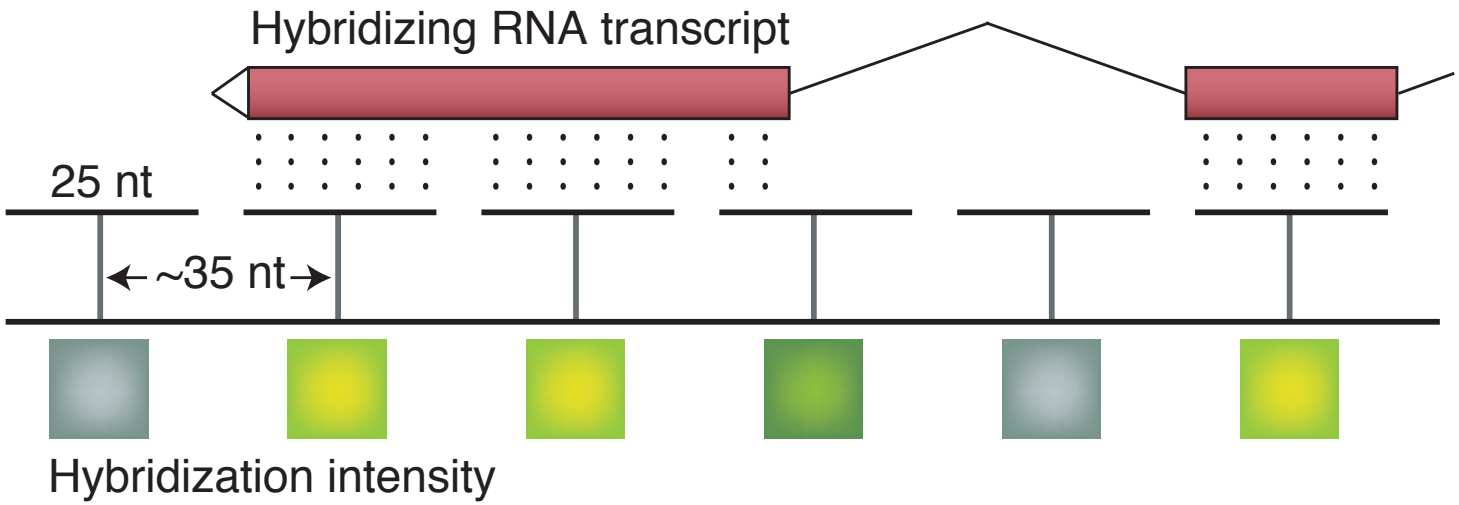


- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays,

---

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

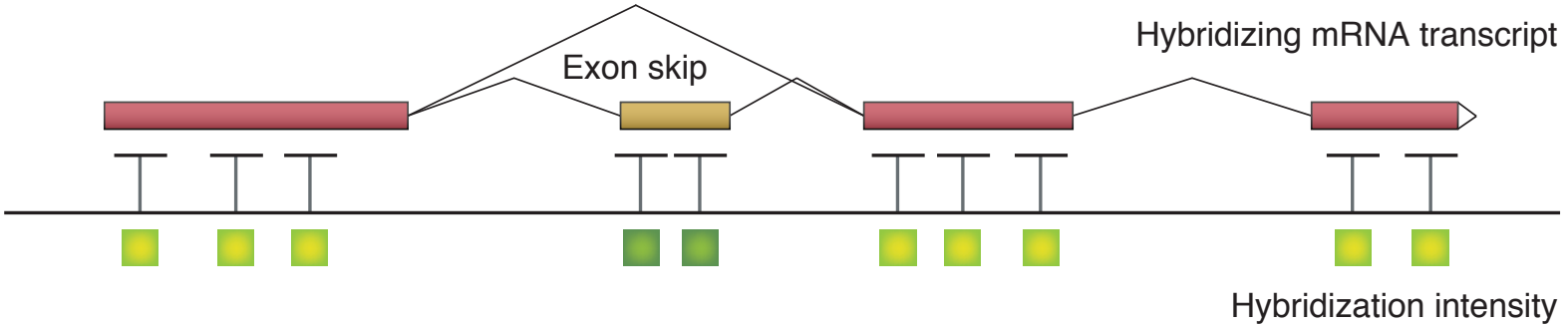
Whole-genome Tiling Arrays



- 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

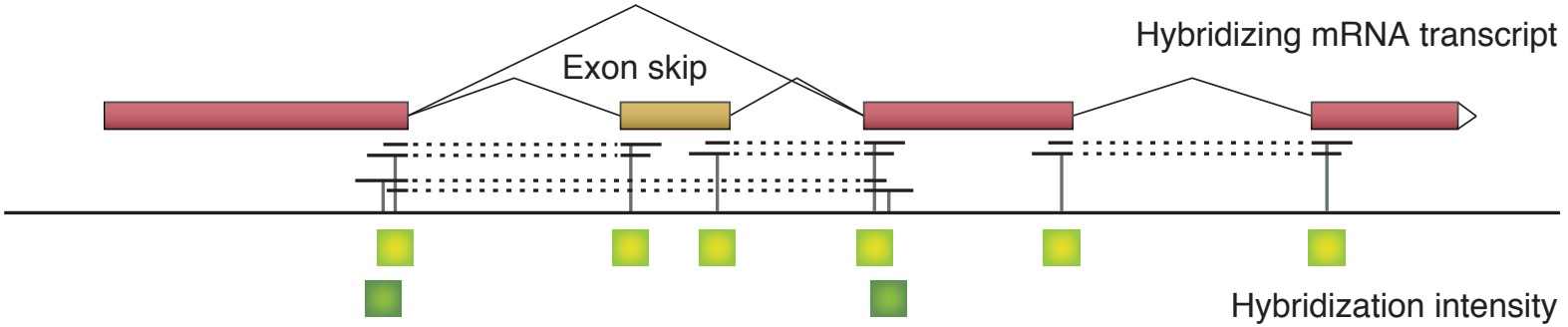


- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays, exon junction arrays

---

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

Whole-genome Tiling Arrays



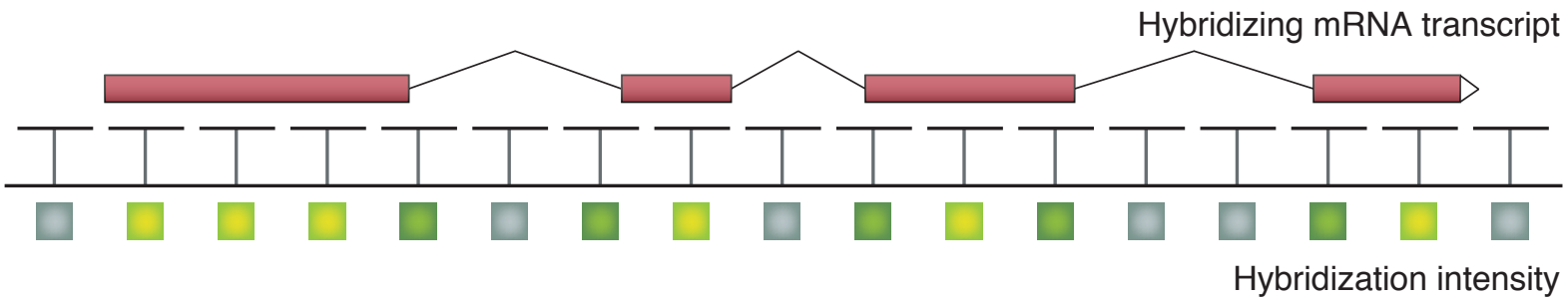
- Whole-genome, quantitative measurements of expression
- Allows to cost-effectively analyze many conditions (replicates)
- Hybridization data is noisy, analysis challenging
- Variants: exon arrays, exon junction arrays

---

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

Tiling Array Analysis Challenges (I)

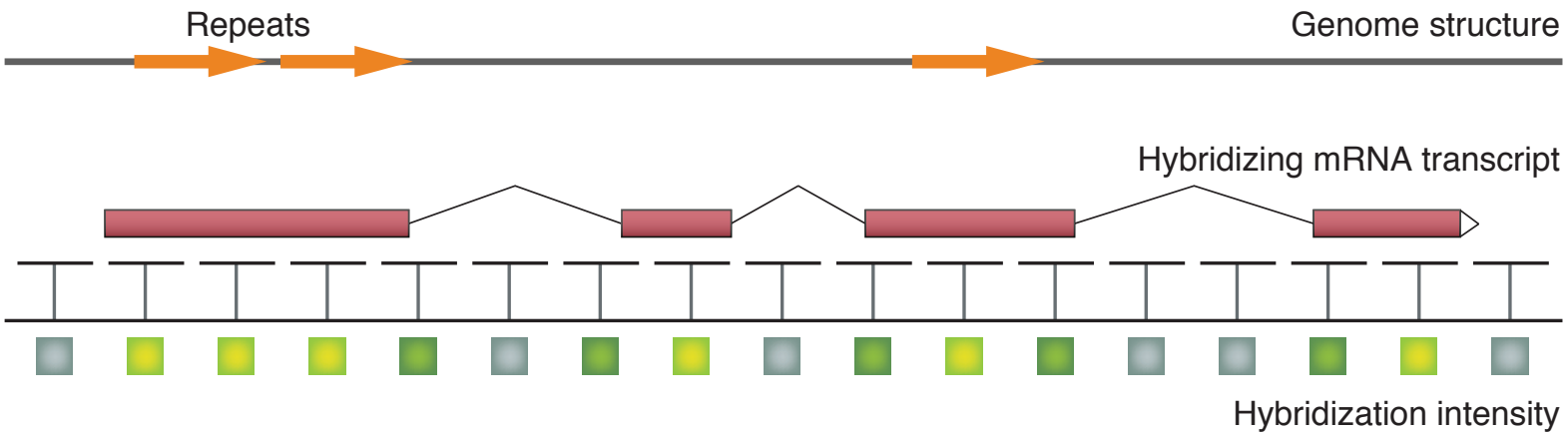
Repeats cause cross-hybridization



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

Tiling Array Analysis Challenges (I)

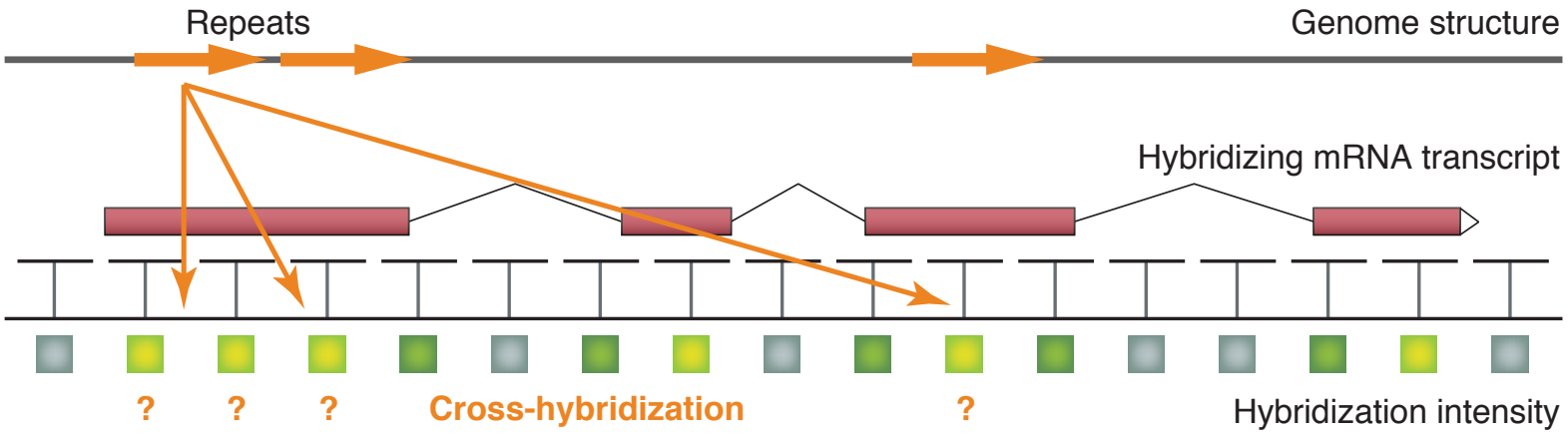
Repeats cause cross-hybridization



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

Tiling Array Analysis Challenges (I)

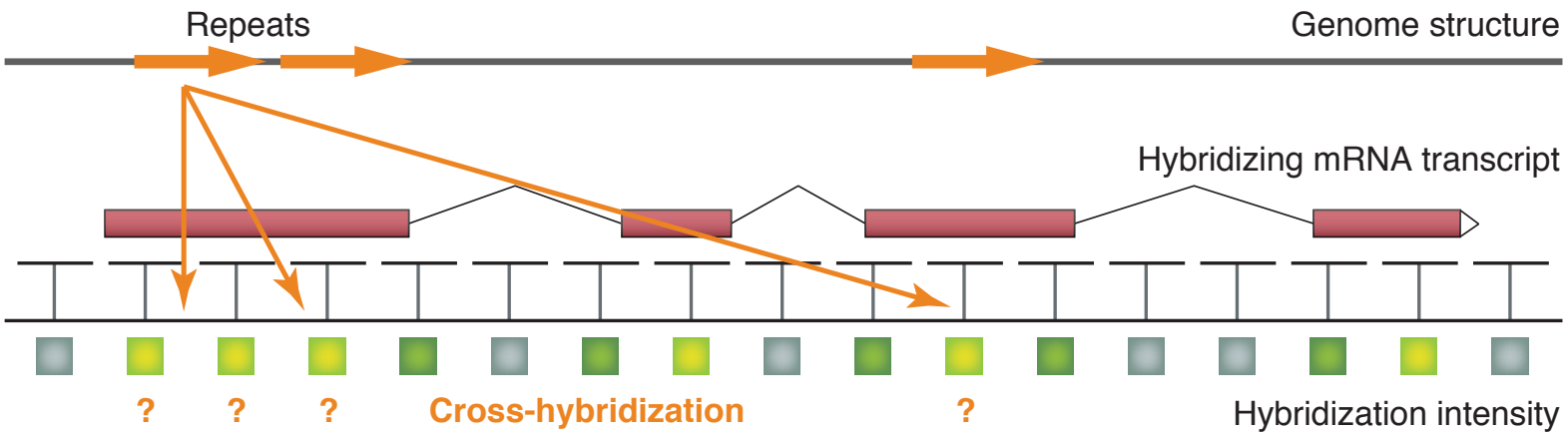
Repeats cause cross-hybridization



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

Tiling Array Analysis Challenges (I)

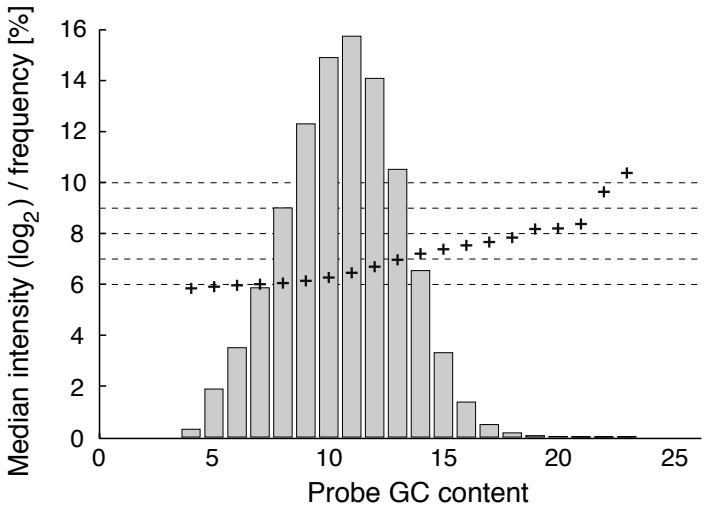
Repeats cause cross-hybridization



⇒ 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].

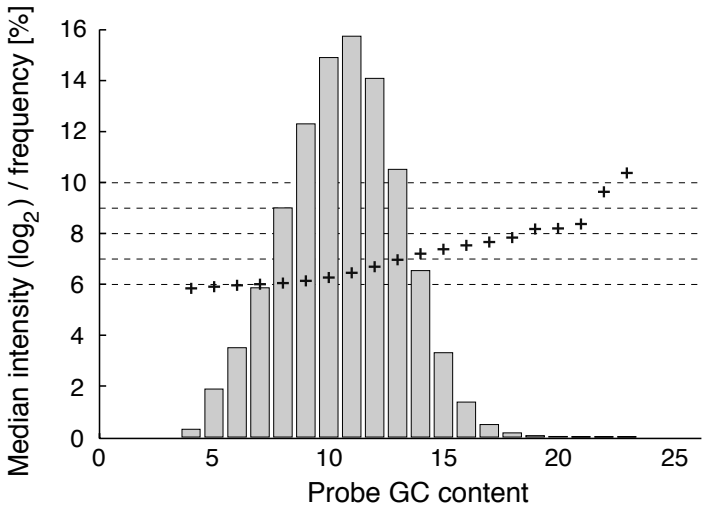
$$n_{ij} = \frac{x_{ij} - b_j(y_i)}{y_i}$$

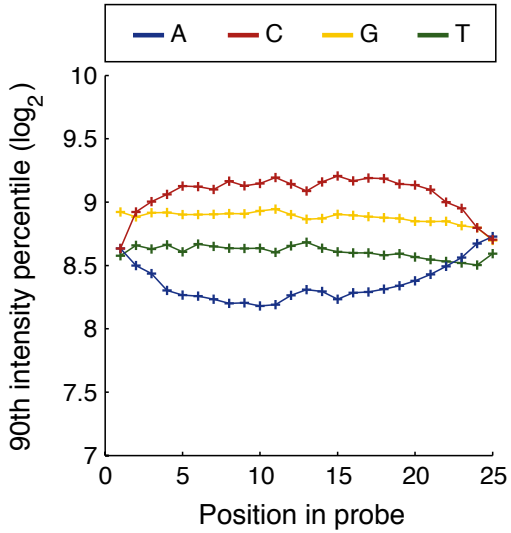
for probe i on replicate array j with RNA hybridization signal x_{ij} to obtain normalized signal n_{ij}; DNA hybridization signal y_i is transformed into RNA background signal by b_j estimated from intergenic probes.

- Regression techniques [Royce et al., 2007b, Zeller et al., 2008c]

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].

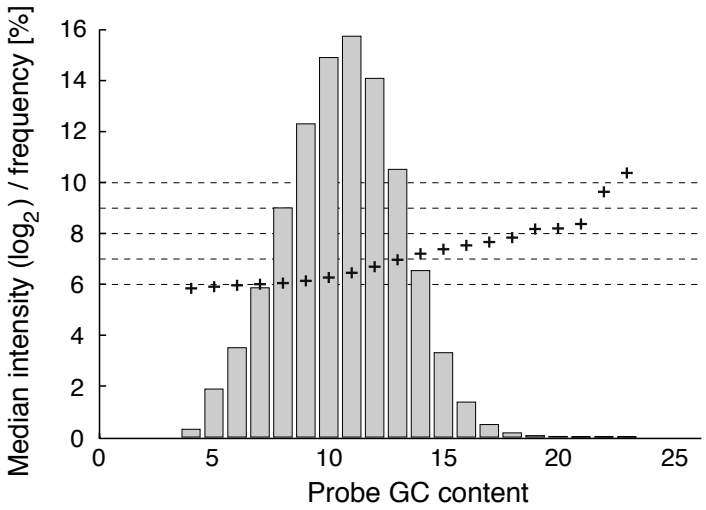
$$n_{ij} = \frac{x_{ij} - b_j(y_i)}{y_i}$$

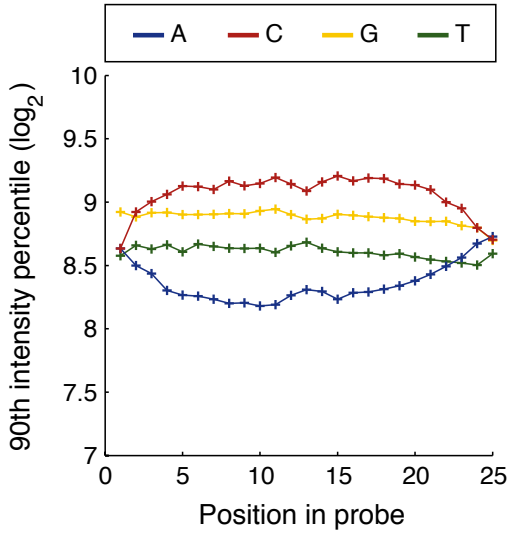
for probe i on replicate array j with RNA hybridization signal x_{ij} to obtain normalized signal n_{ij}; DNA hybridization signal y_i is transformed into RNA background signal by b_j estimated from intergenic probes.

- Regression techniques [Royce et al., 2007b, Zeller et al., 2008c]

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].

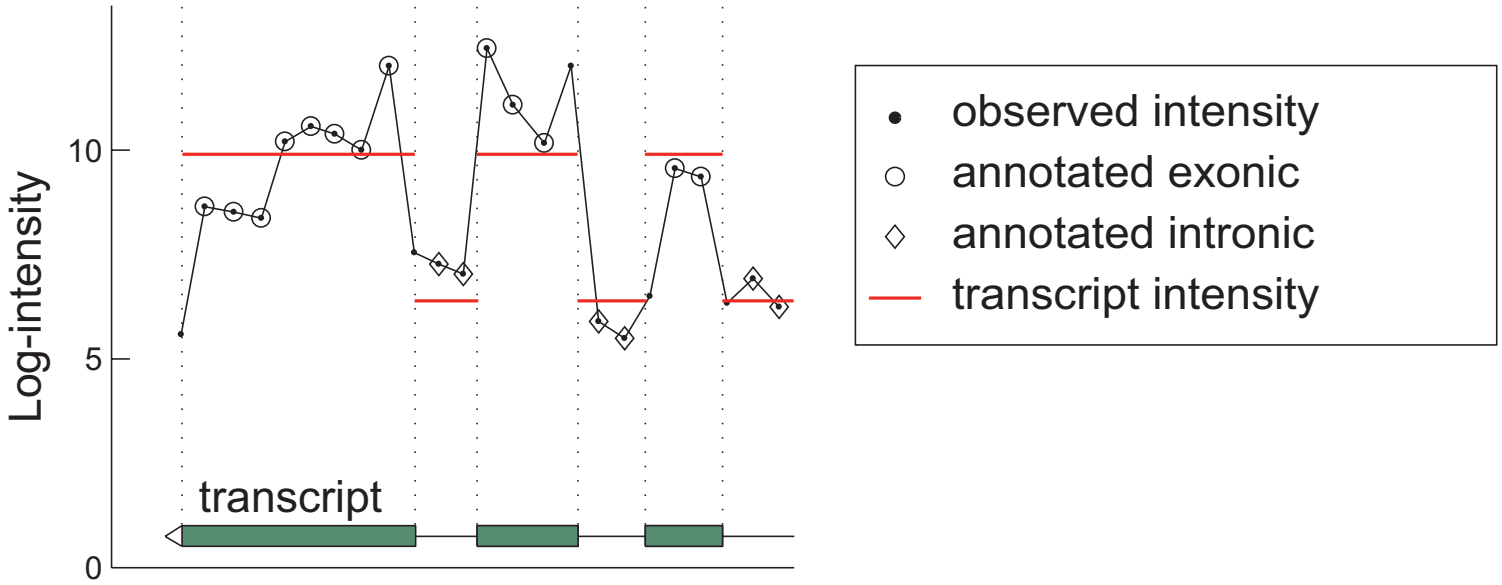
$$n_{ij} = \frac{x_{ij} - b_j(y_i)}{y_i}$$

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 intergenic probes.

- Regression techniques [Royce et al., 2007b, Zeller et al., 2008c]

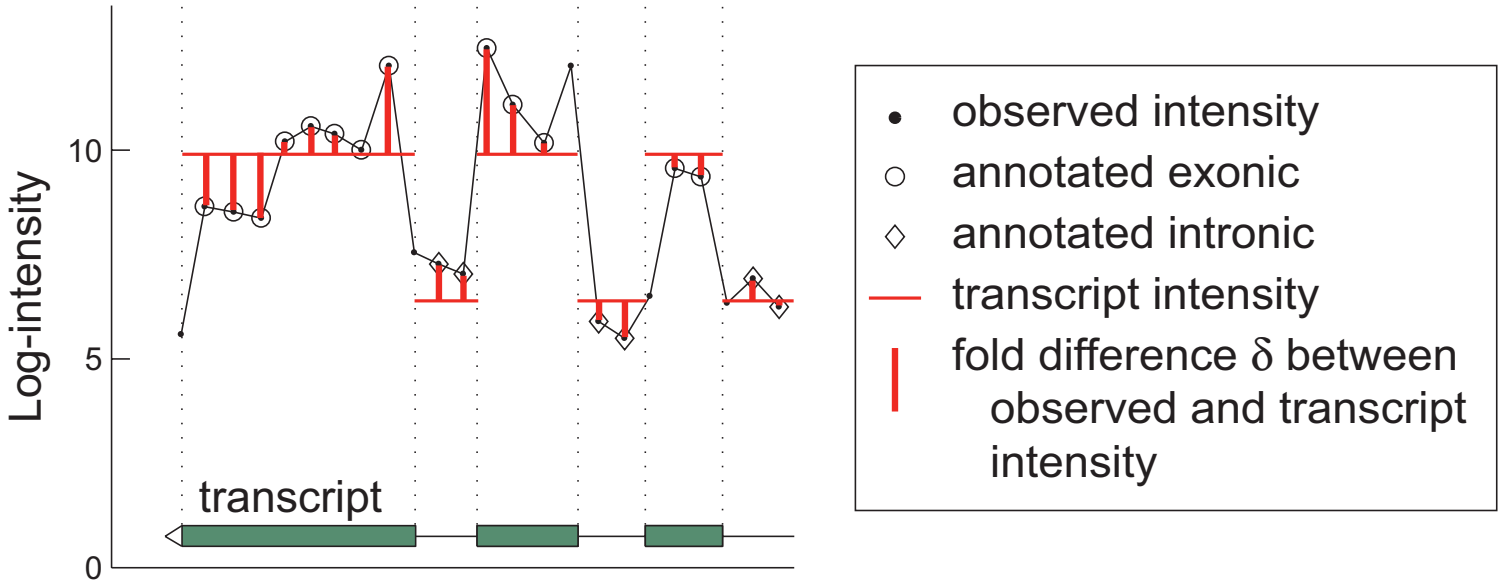
Tiling Array Analysis Challenges (III)

- 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 $\mathbf{x}_i$ (positional information on mono-, di- and tri-mer occurrence) as input for regression.
  Models probe sequence effect depending on y_i: $f(\mathbf{x}_i, y_i) \approx \delta_i$



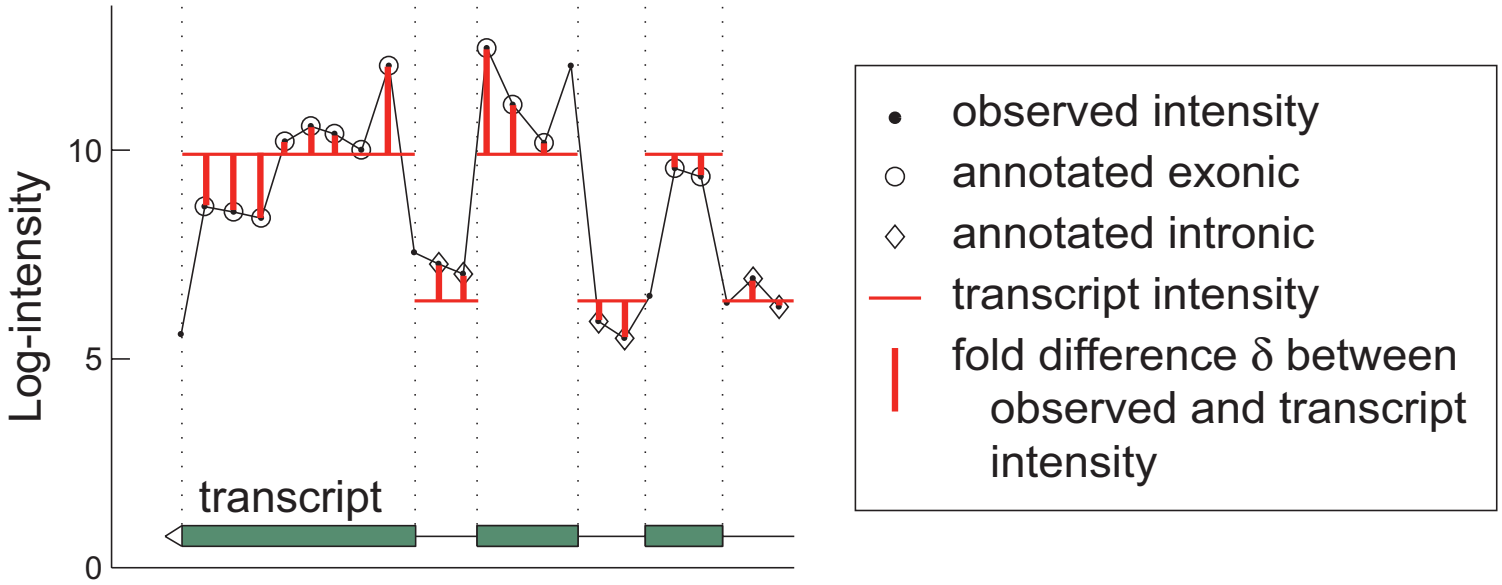
Tiling Array Analysis Challenges (III)

- 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 $\mathbf{x}_i$ (positional information on mono-, di- and tri-mer occurrence) as input for regression.
  Models probe sequence effect depending on y_i: $f(\mathbf{x}_i, y_i) \approx \delta_i$



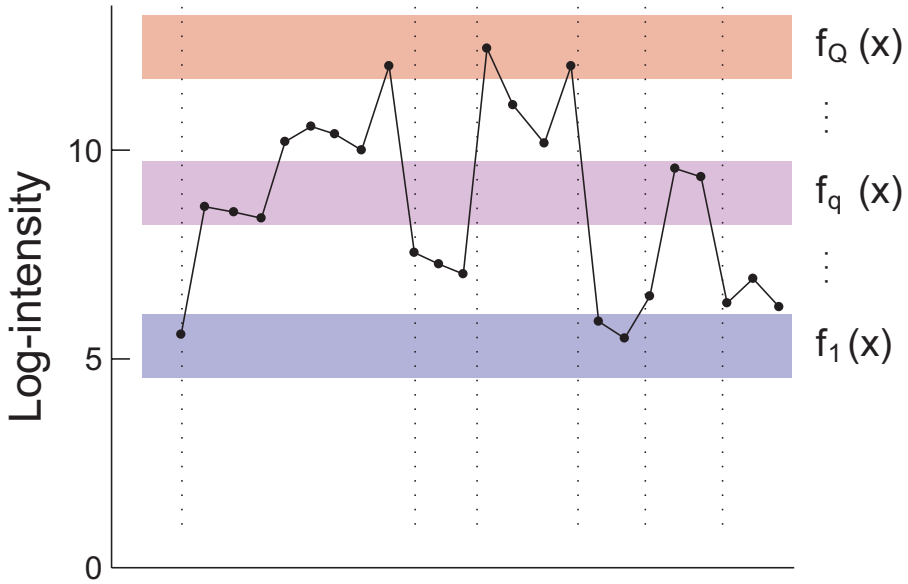
Tiling Array Analysis Challenges (III)

- 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 $\mathbf{x}_i$ (positional information on mono-, di- and tri-mer occurrence) 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 $\mathbf{x}_i$ (positional information on mono-, di- and tri-mer occurrence) as input for regression.
  Models probe sequence effect depending on y_i: $f(\mathbf{x}_i, y_i) \approx \delta_i$



Discretize y into $Q = 20$ quantiles and estimate Q independent functions $f_1(\mathbf{x}), \ldots, f_Q(\mathbf{x})$

Linear regression
$f_q(\mathbf{x}) = \mathbf{w}_q^T \mathbf{x}$

Identification of Expression Changes



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., 2003]

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).

Identification of Expression Changes



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

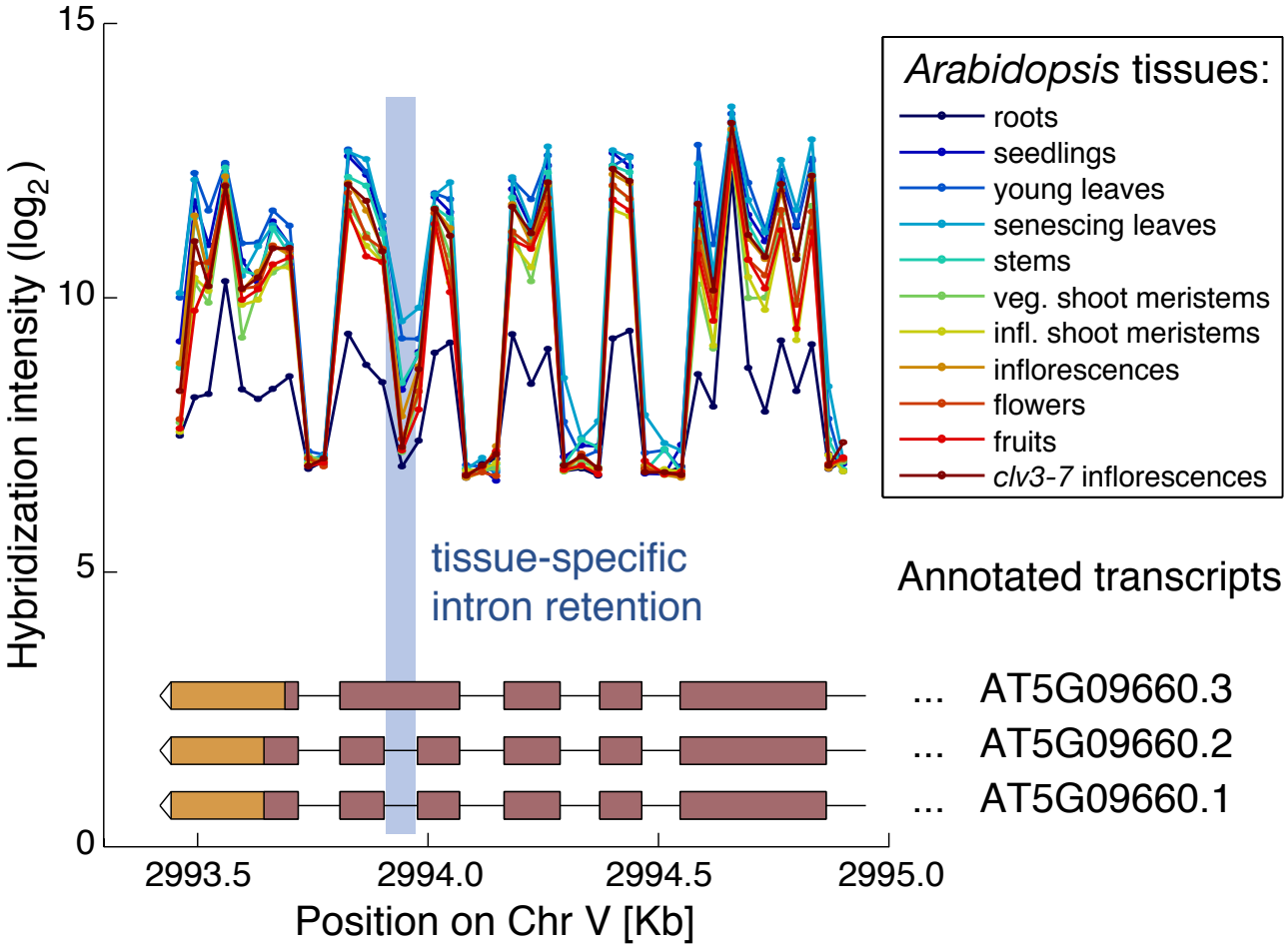
Gene expression values are typically computed using robust "summarization methods" that account for probe noise
[e.g. Irizarry et al., 2003]

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:

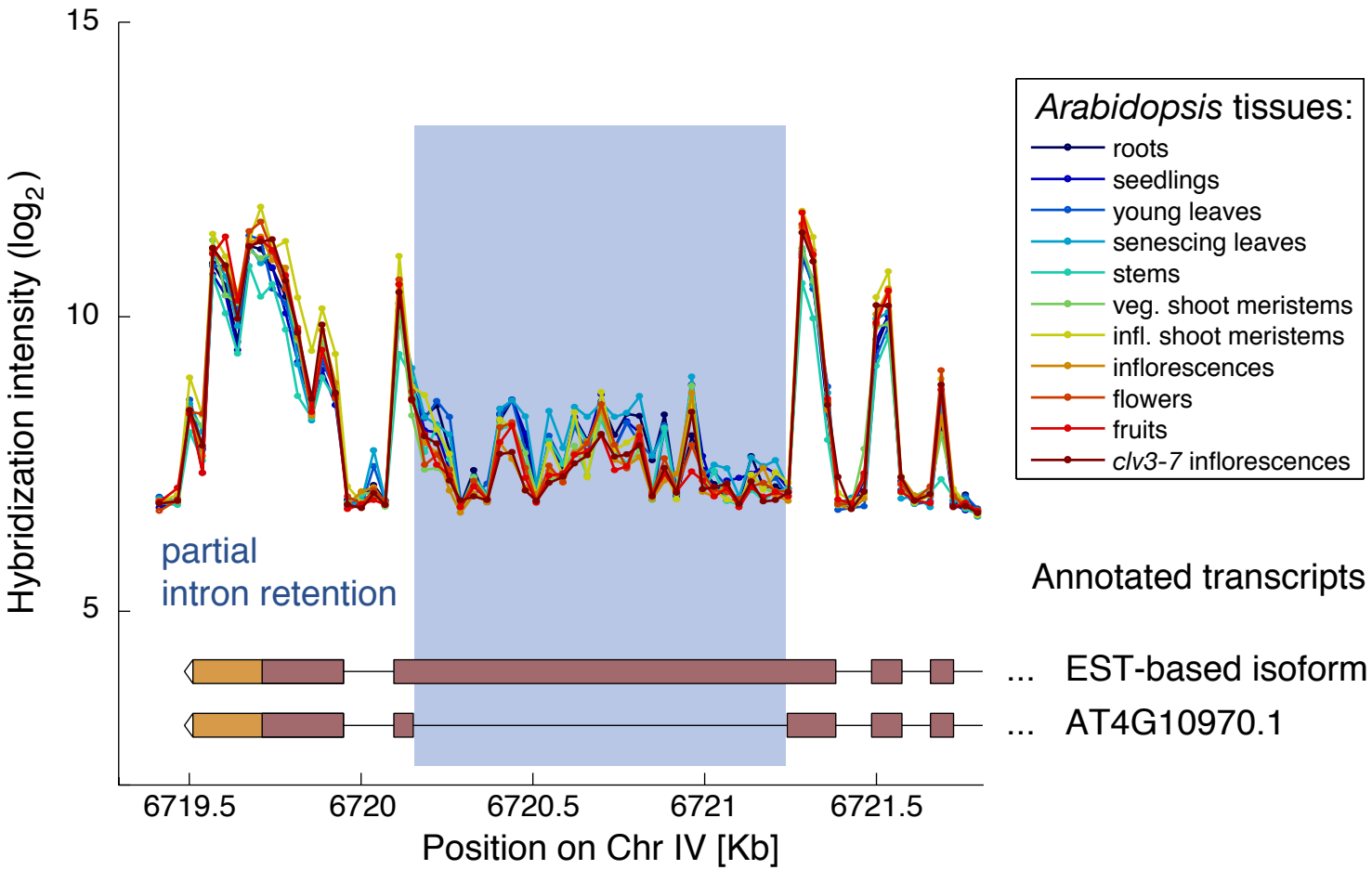
- Annotations change, only remapping is needed to obtain expression measurements for the latest annotation.
- Expression can be measured per exon, not only per gene.
- Expression can be measured for introns ($\Rightarrow$ detect retention).

Detection of Alternative Transcripts (I)



Goal: Identify exon/intron segments that are *differentially* spliced in the analyzed samples.

Detection of Alternative Transcripts (II)



Goal: Identify exon/intron segments that show different intensities than *other exons/introns* in at least one analyzed sample.

Detecting Alternative Exons

Fit a gene expression model to exon array data [Irizarry et al., 2003]:

$$x_{ik} = g_k + p_i + \epsilon_{ik}$$

RNA hybridization signal x_{ij},
g_k gene-wide expression effect in sample k,
p_i effect of probe i, error terms ϵ_{ik}.

Detect alternatively spliced exons as outliers [Purdom et al., 2008] from large residuals $\epsilon_{i'k'}$ for alternative exon probes i' in sample k'.

Test exon junction probes for different transcript isoforms for differential expression using e.g. the Kruskal Wallis test [Sugnet et al., 2006].

More sophisticated methods use classification techniques.

[Eichner, 2008, Eichner et al., 2009]

Detecting Alternative Exons

Fit a gene expression model to exon array data [Irizarry et al., 2003]:

$$x_{ik} = g_k + p_i + \epsilon_{ik}$$

RNA hybridization signal x_{ij},
g_k gene-wide expression effect in sample k,
p_i effect of probe i, error terms ϵ_{ik}.

Detect alternatively spliced exons as outliers [Purdom et al., 2008] from large residuals $\epsilon_{i'k'}$ for alternative exon probes i' in sample k'.

Test exon junction probes for different transcript isoforms for differential expression using e.g. the Kruskal Wallis test [Sugnet et al., 2006].

More sophisticated methods use classification techniques.

[Eichner, 2008, Eichner et al., 2009]

Detecting Alternative Exons

Fit a gene expression model to exon array data [Irizarry et al., 2003]:

$$x_{ik} = g_k + p_i + \epsilon_{ik}$$

RNA hybridization signal x_{ij},
g_k gene-wide expression effect in sample k,
p_i effect of probe i, error terms ϵ_{ik}.

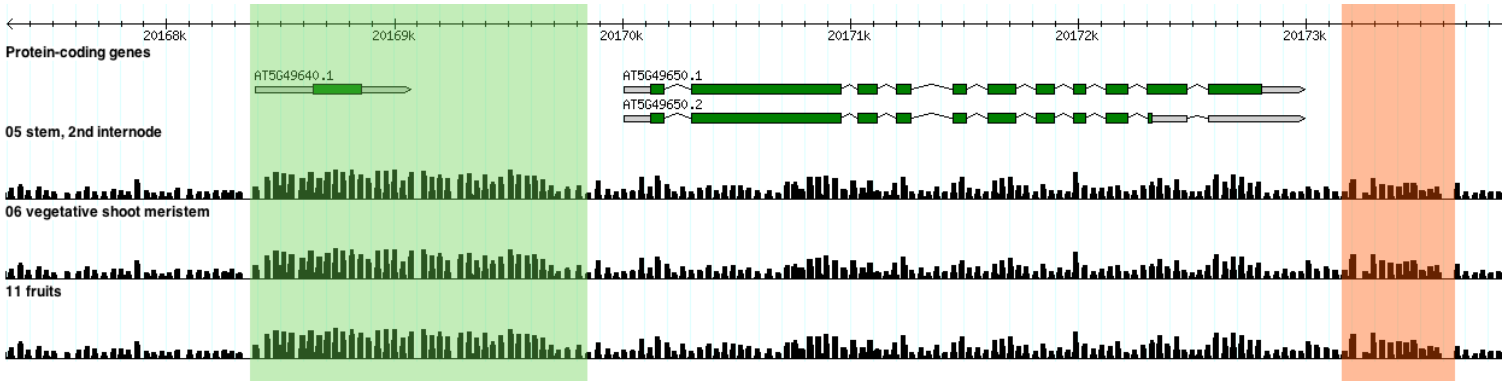
Detect alternatively spliced exons as outliers [Purdom et al., 2008] from large residuals $\epsilon_{i'k'}$ for alternative exon probes i' in sample k'.

Test exon junction probes for different transcript isoforms for differential expression using e.g. the Kruskal Wallis test [Sugnet et al., 2006].

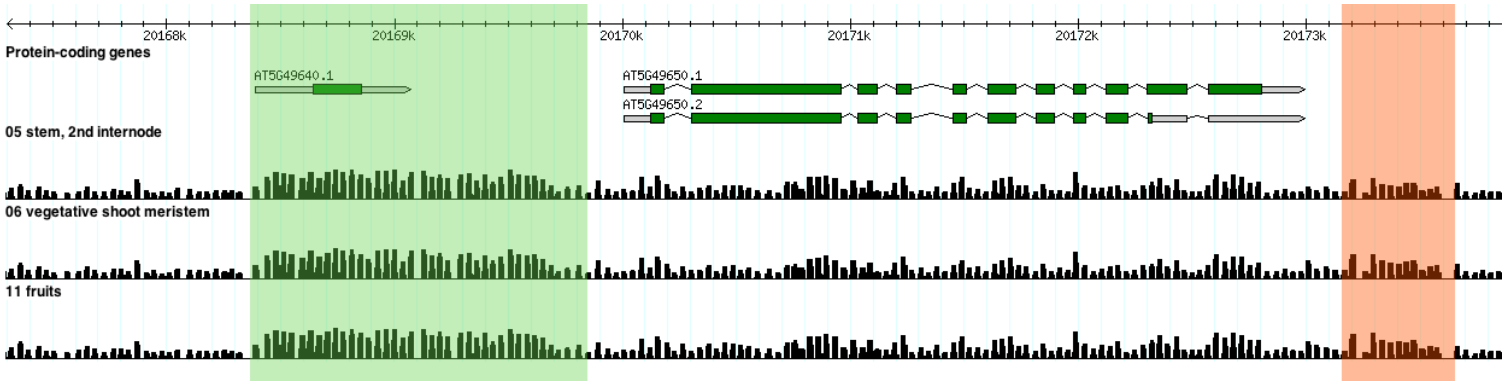
More sophisticated methods use classification techniques.

[Eichner, 2008, Eichner et al., 2009]

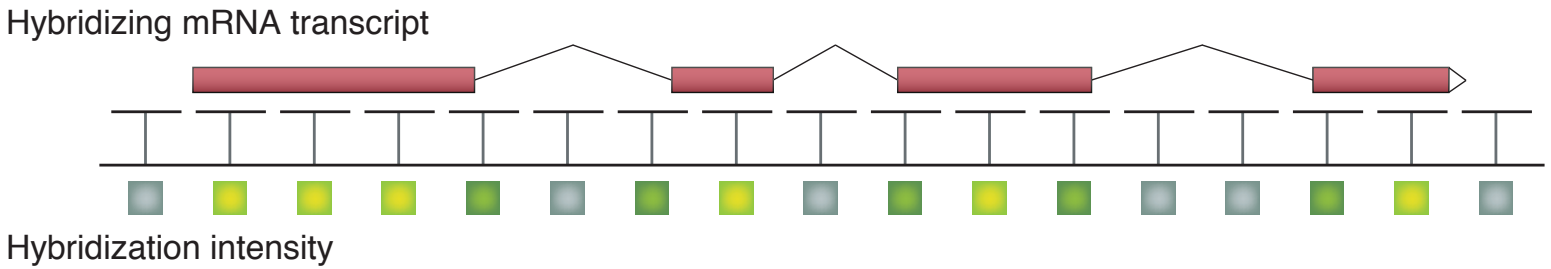
Discovery of Expressed Transcripts



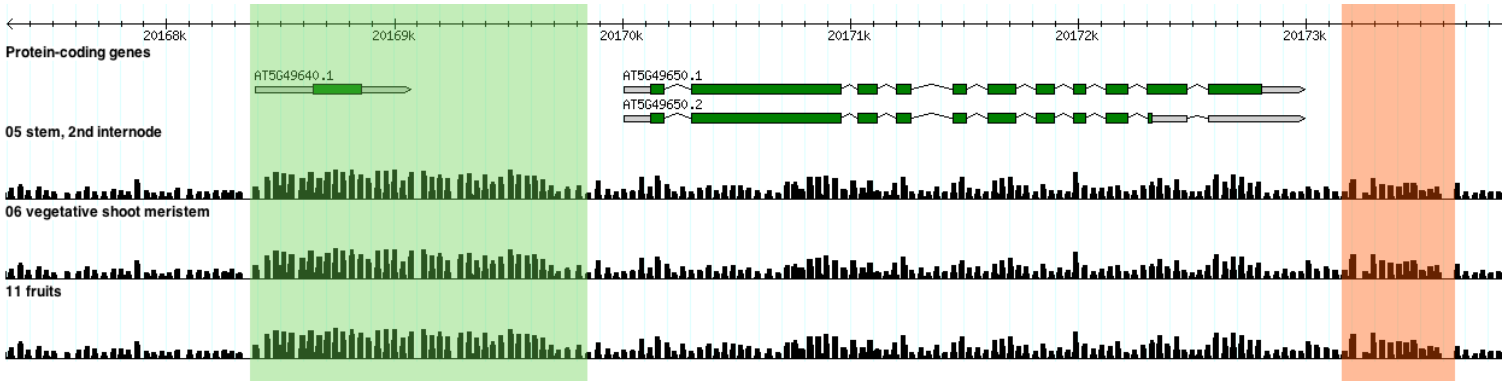
Discovery of Expressed Transcripts



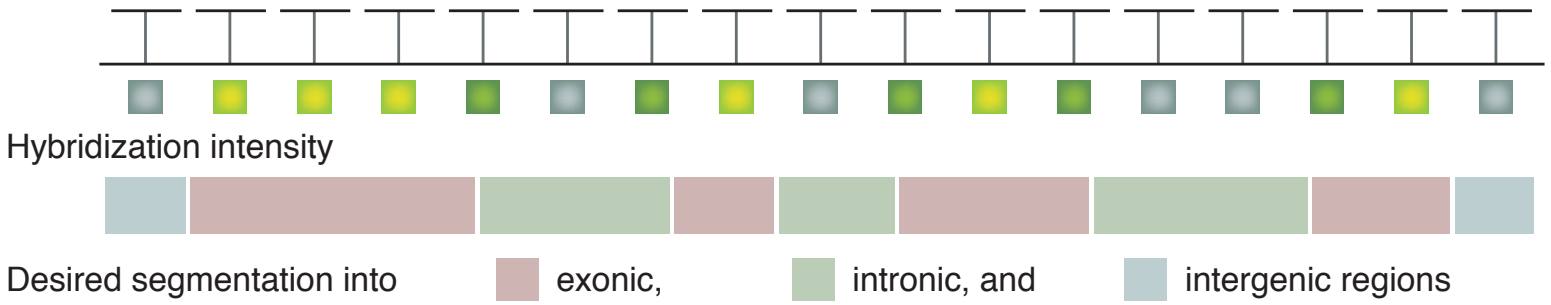
De novo segmentation is needed to re-annotate expressed genes.



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., 2006]:

$x_{ij} = \mu_s + \epsilon_{ij}$ 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}. μ_s is the mean signal of the sth segment, ϵ_{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., 2006].

Problem: S is to be user-specified

Dynamic Programming Segmentation

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

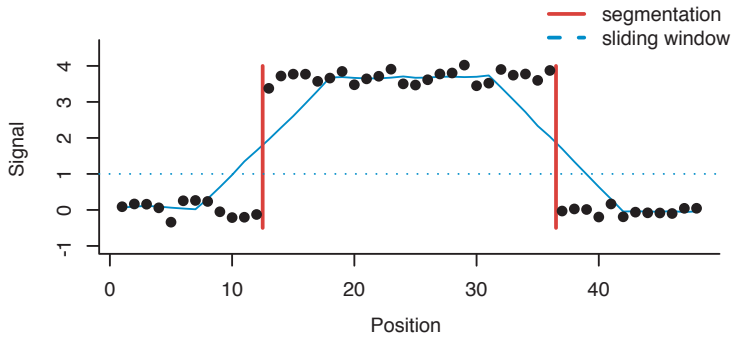
$x_{ij} = \mu_s + \epsilon_{ij}$ for $t_s \leq i < t_{s+1}$

for probe i on replicate array j with RNA hybridization signal x_{ij}; segment boundaries t_s and t_{s+1}. μ_s is the mean signal of the sth segment, ϵ_{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., 2006].

Problem: S is to be user-specified

Dynamic Programming Segmentation

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

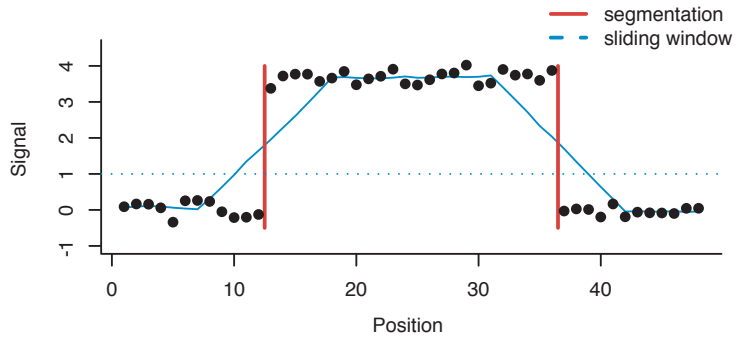
$x_{ij} = \mu_s + \epsilon_{ij}$ for $t_s \leq i < t_{s+1}$

for probe i on replicate array j with RNA hybridization signal x_{ij}; segment boundaries t_s and t_{s+1}. μ_s is the mean signal of the sth segment, ϵ_{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., 2006].

Problem: S is to be user-specified

Dynamic Programming Segmentation

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

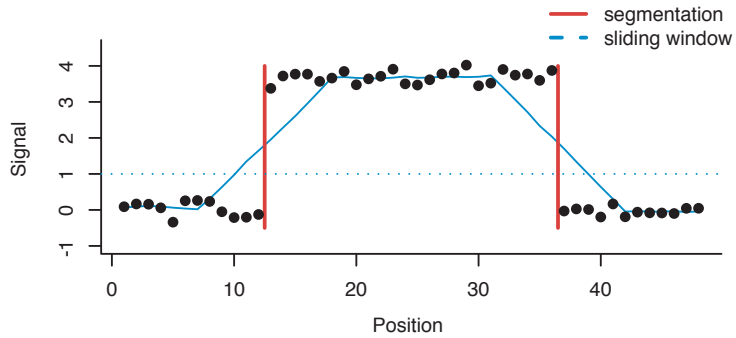
$x_{ij} = \mu_s + \epsilon_{ij}$ for $t_s \leq i < t_{s+1}$

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

Minimize the sum of squared residuals:

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

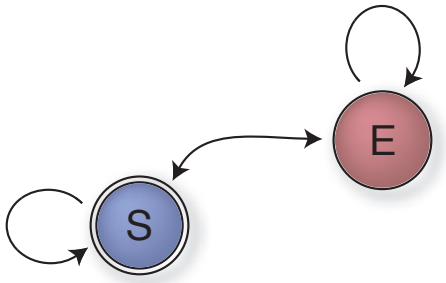
where S is the number of segments and R the number of replicates.



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

Problem: S is to be user-specified

Hidden Markov Models



expressed

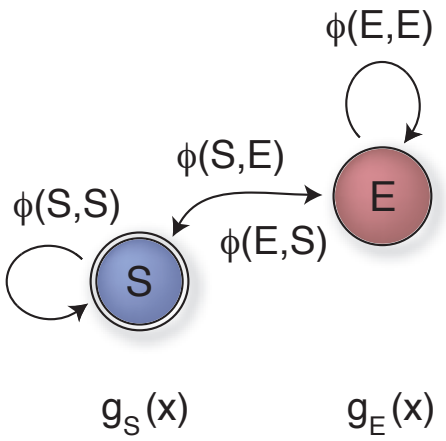
non-expressed

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

  [Ji and Wong, 2005a, 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

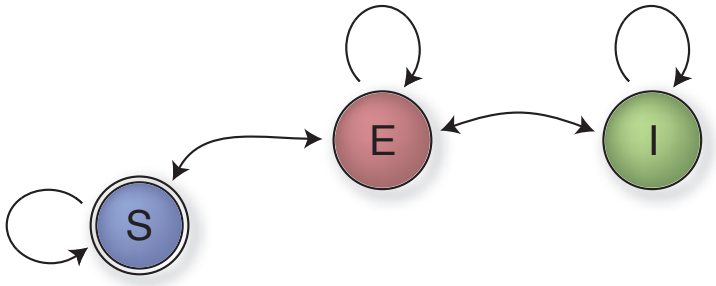
non-expressed

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

  [Ji and Wong, 2005a, 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



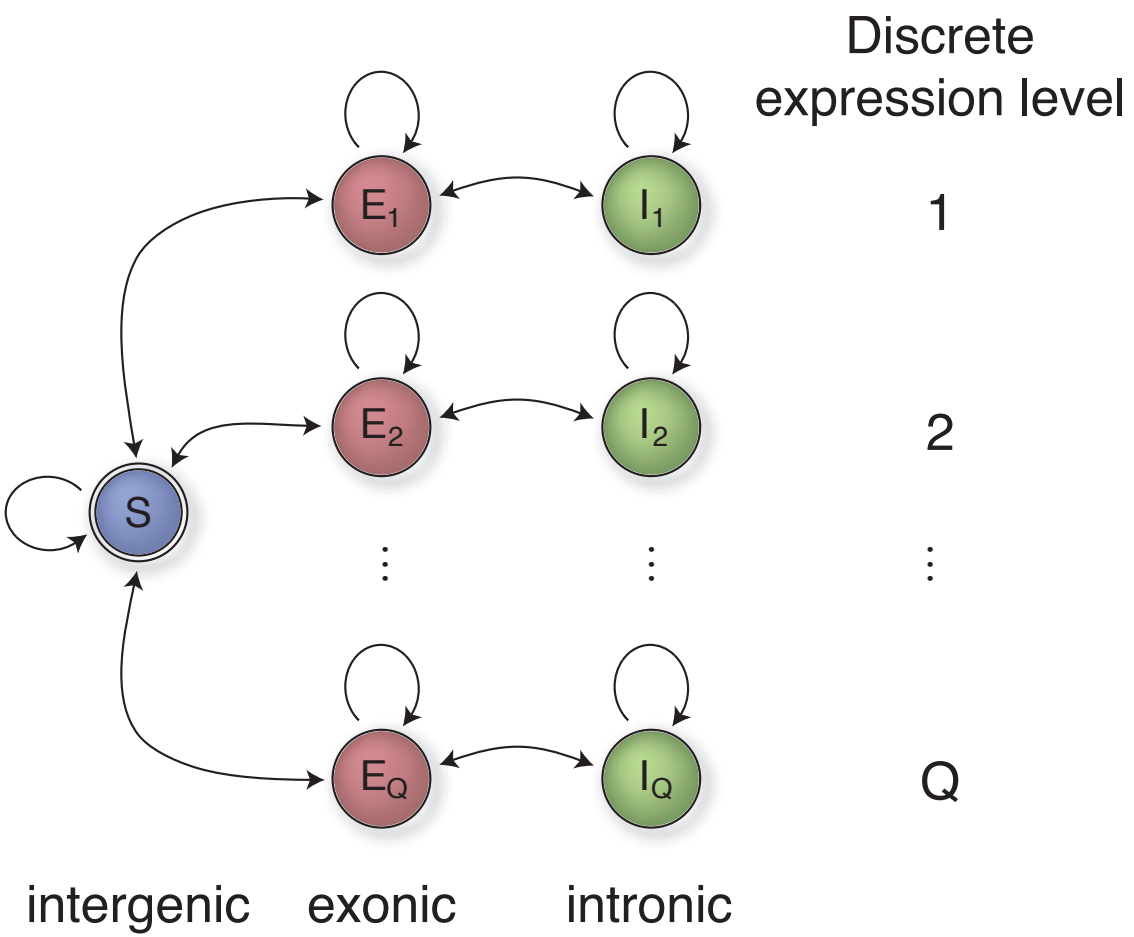
intergenic exonic intronic

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

  [Ji and Wong, 2005a, 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, 2005a, 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 π and parametrization θ:

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

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

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

$$\operatorname*{argmax}_{\pi} F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi})$$ (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 $\boldsymbol{\theta}$:

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

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

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

$$\operatorname*{argmax}_{\boldsymbol{\pi}} F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi})$$ (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 $\boldsymbol{\theta}$:

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

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

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

$$\underset{\boldsymbol{\pi}}{\operatorname{argmax}}\, F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\pi})$$ (Viterbi decoding [Durbin et al., 1998])

Training an HMM

Training sequences: signals $\mathbf{x}^i$ and labels π^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: $A_{k,l} = \sum_{i=1}^{n} \sum_{p=1}^{|\pi^i|} [[\pi_p^i = k \wedge \pi_{p+1}^i = l]]$

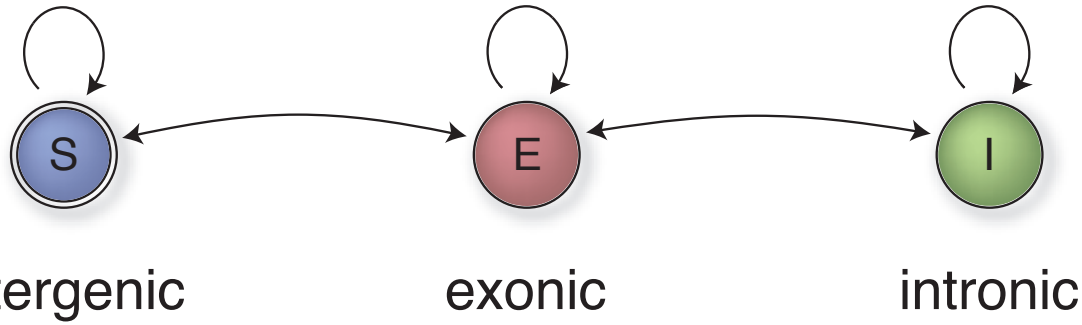
Log emission probabilities [Durbin et al., 1998]
for piece-wise constant g_k with L levels (ranging from t_l to t_{l+1}):
$g_{k,l} = log(\frac{E_l}{\sum_{l'} E_l'})$
counting discrete signal values: $E_l = \sum_{i=1}^{n} \sum_{p=1}^{|\pi^i|} [[\pi_p^i = k \wedge t_l < x_p^i \leq t_{l+1}]]$

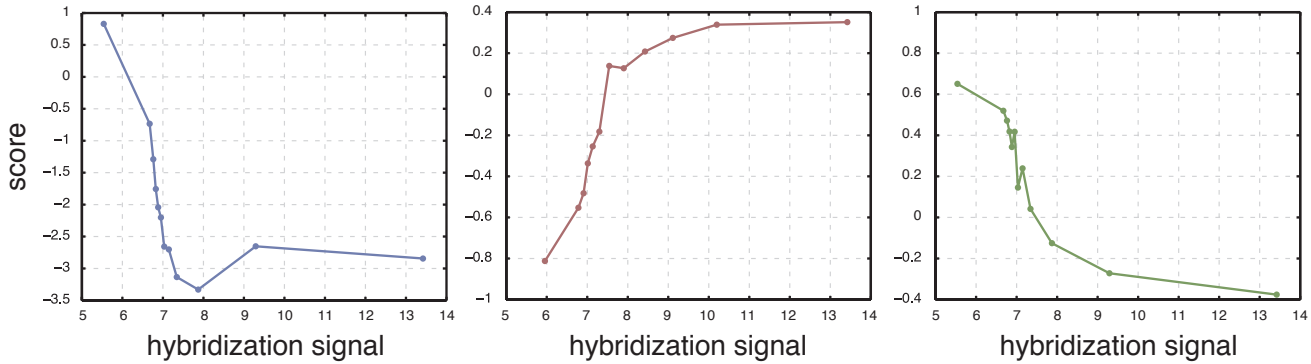
HMMs can also be (re-)trained in an unsupervised fashion
[e.g. Munch et al., 2006]

Hidden Markov SVMs

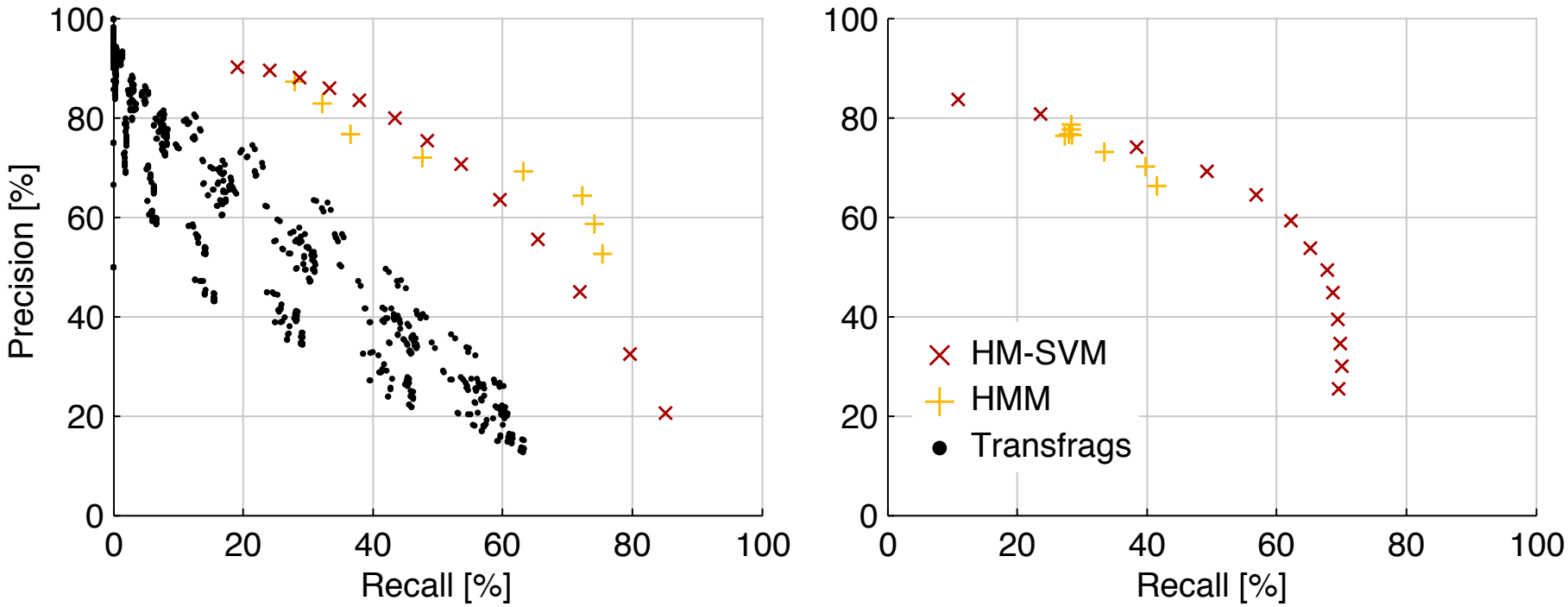
Enforce a large margin (cf. gene finding)
between the correct one $\boldsymbol{\pi}^{(i)}$ and any other labeling $\overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)}$:

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



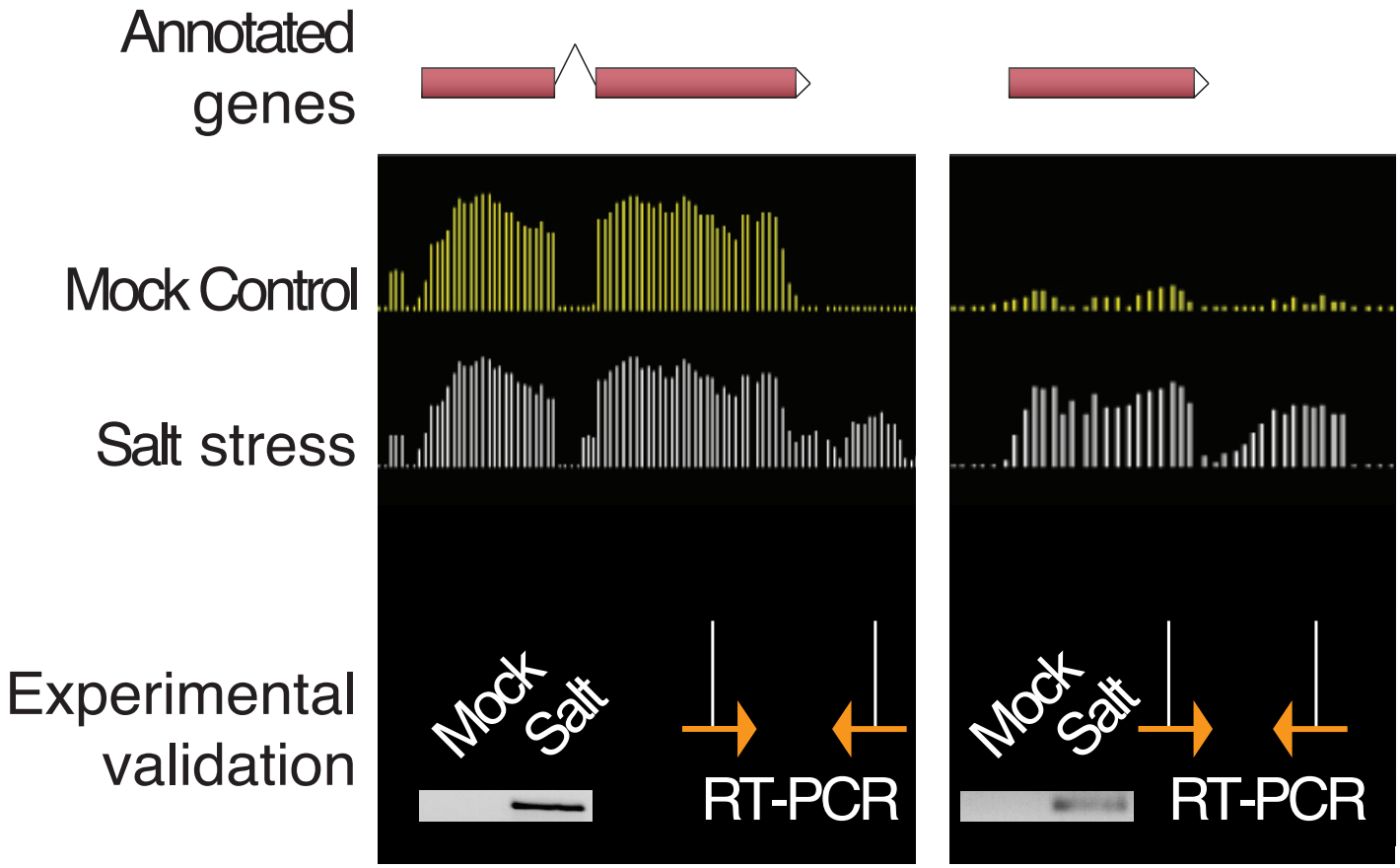


Method Comparison



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

Identification of Differential TARs



Apply statistical test for significant expression change to signal from transcriptionally active regions (TARs) defined by previous segmentation [Zeller et al., 2009].

Roadmap

1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Identification of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology & Limitations
   - Assembly & Read Mapping
4. Extensions
   - Quantification of Transcripts
   - ChIP-on-Chip Studies

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 (Nanopores, 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

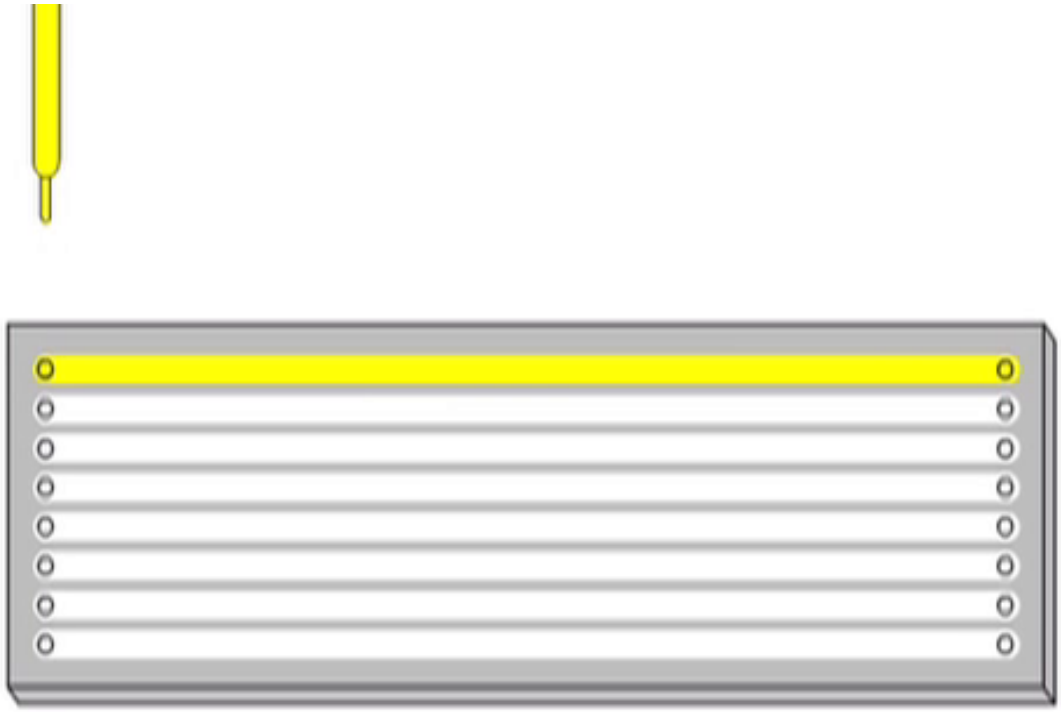
Sequencing Technology

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

} Next Generation Sequencing

Illumina Sequencing

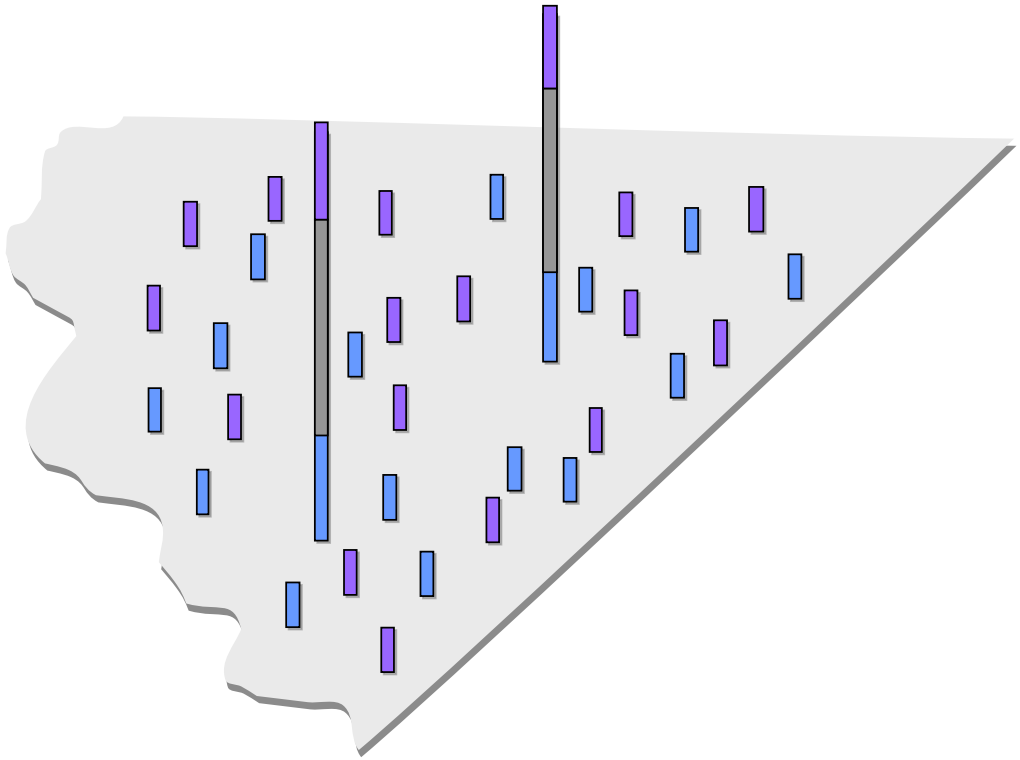
- *Solexa* released a sequencing machine in 2006
- Fragment sizes from $28 - 75$
- Probes are fixed to a glass plate "flow cell"
- Reagents are directed through flow cell



(see Movie)

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

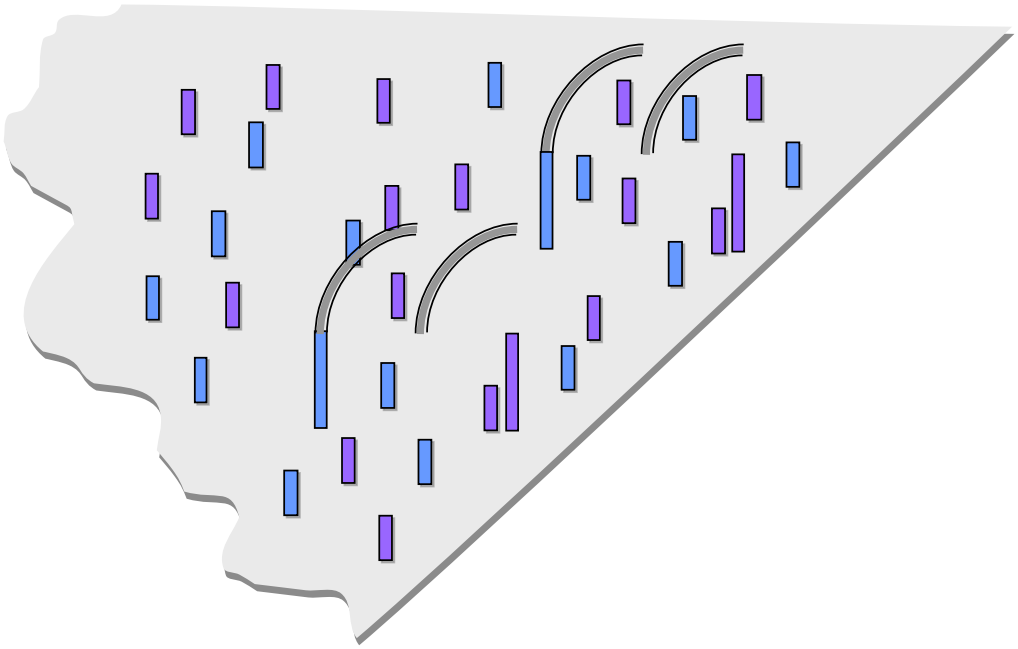


(see Movie)

[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

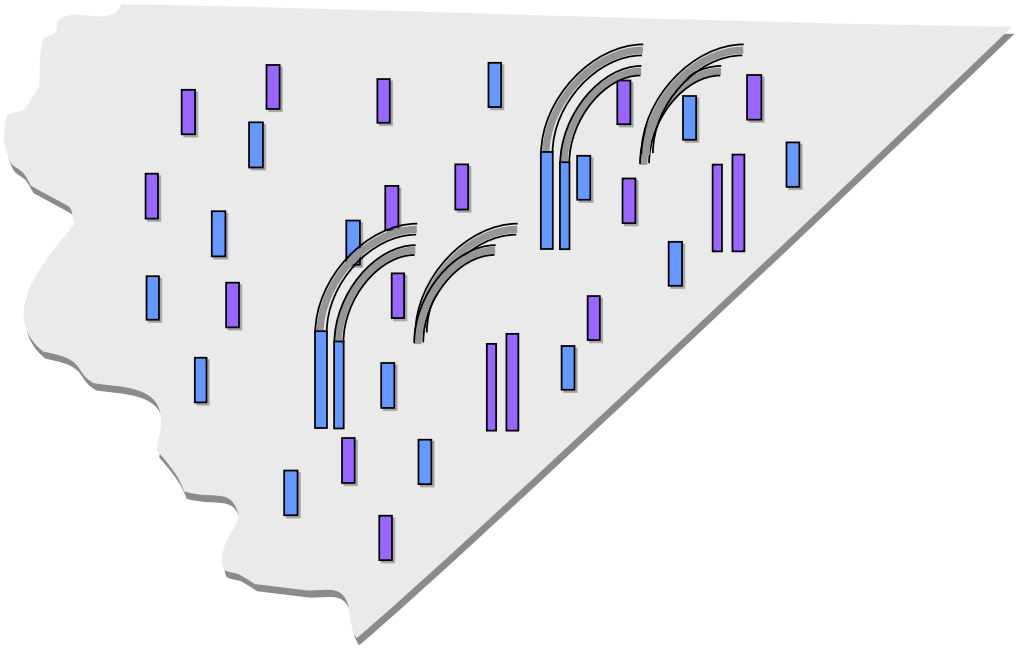


(see Movie)

[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

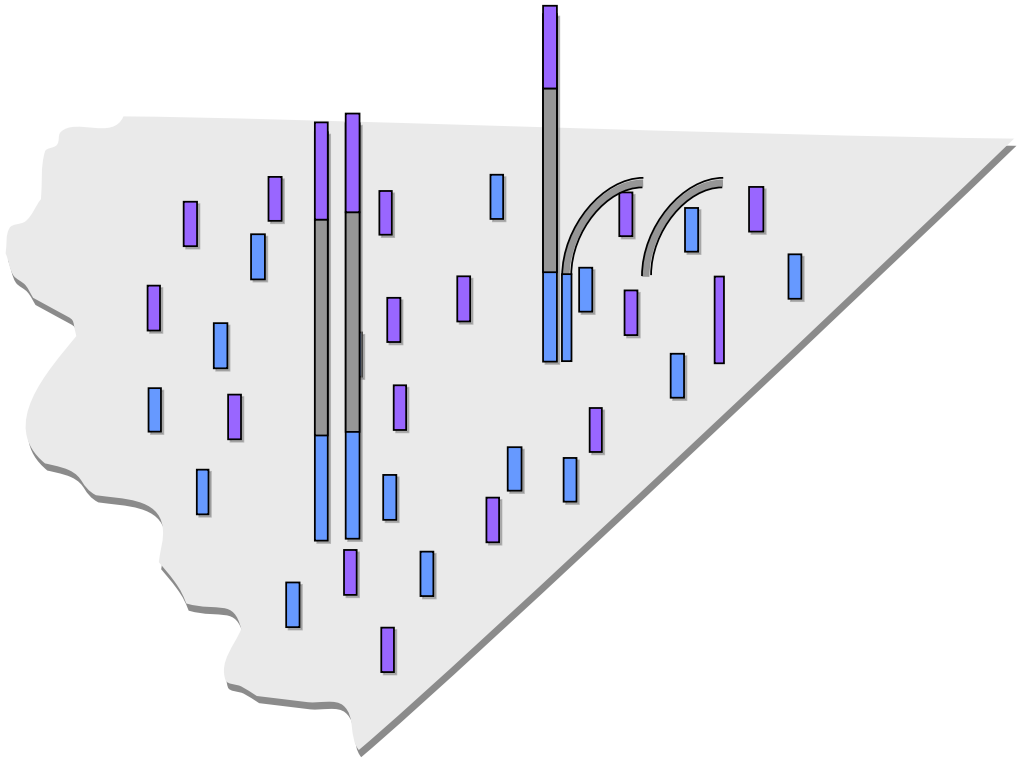


(see Movie)

[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

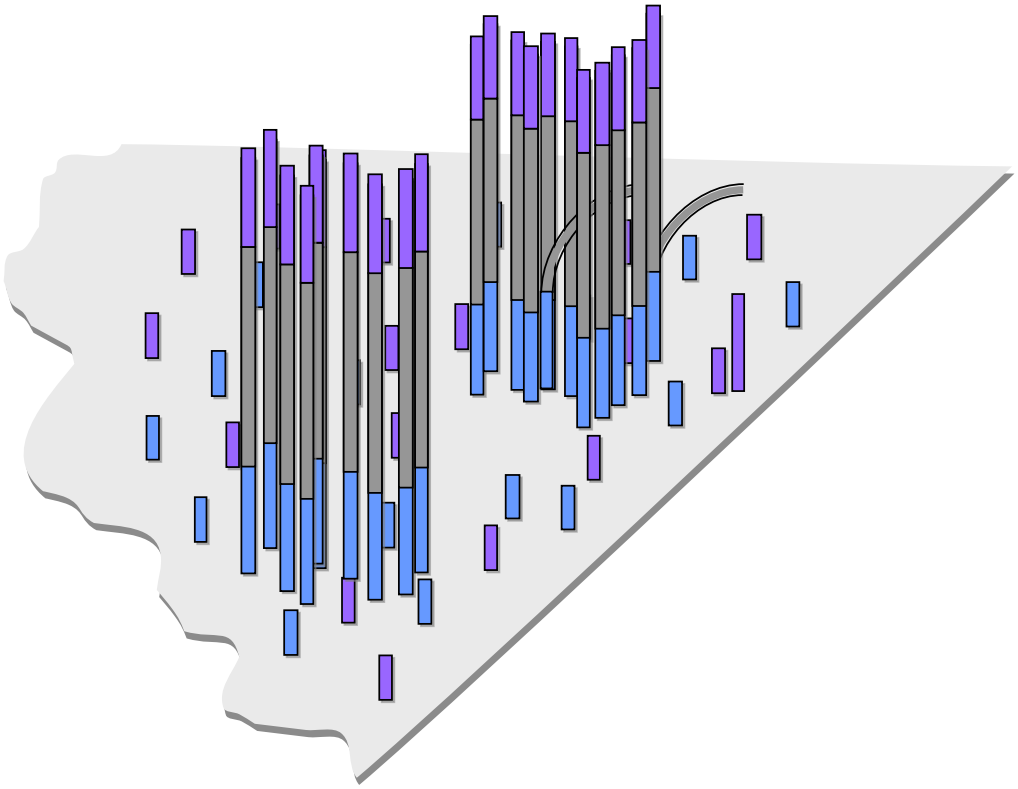


(see Movie)

[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

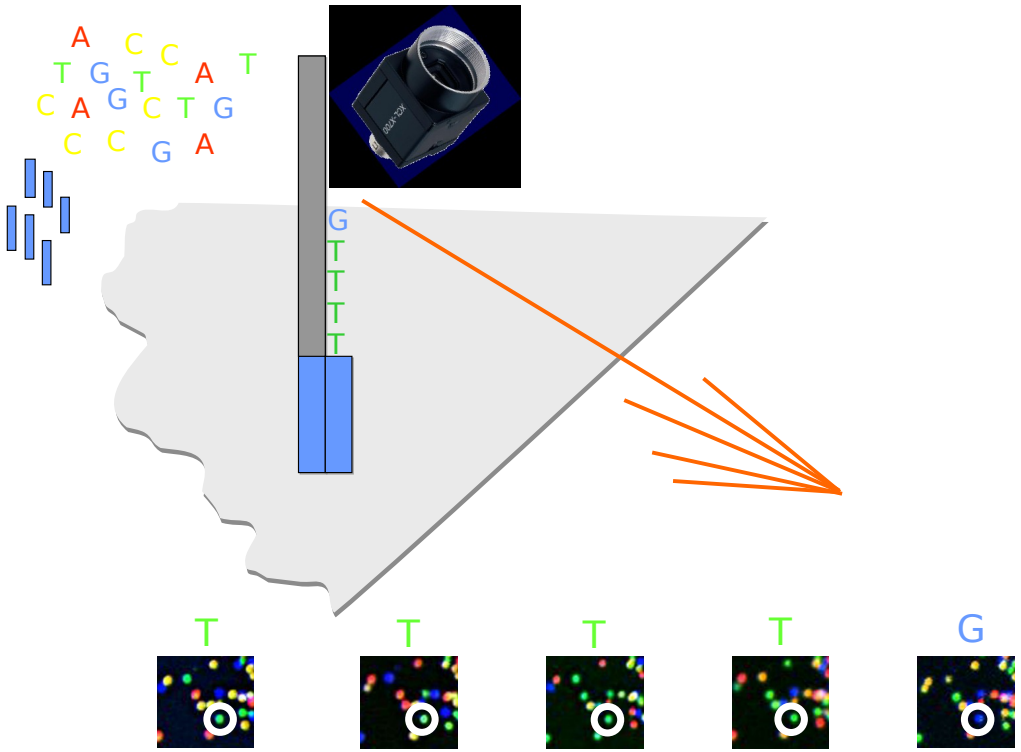


(see Movie)

[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

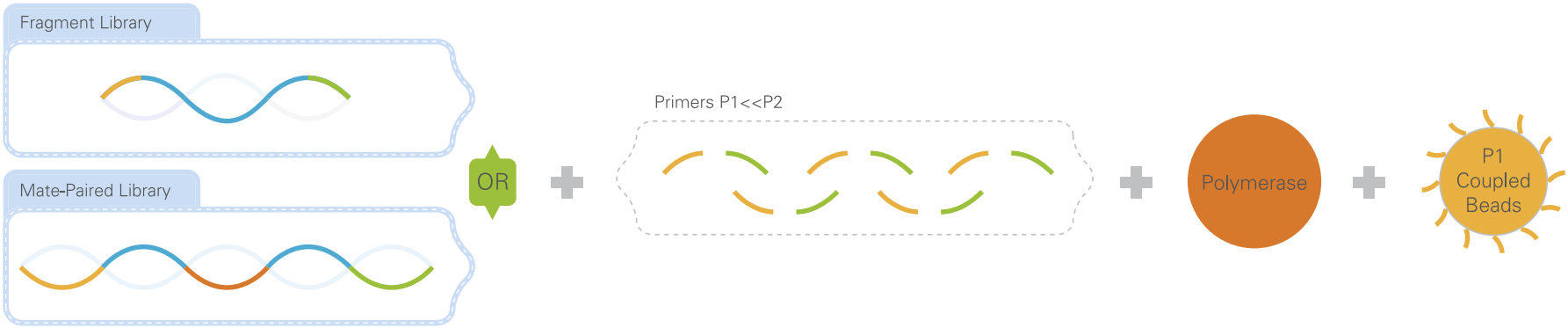


(see Movie)

[Ossowski, 2007]

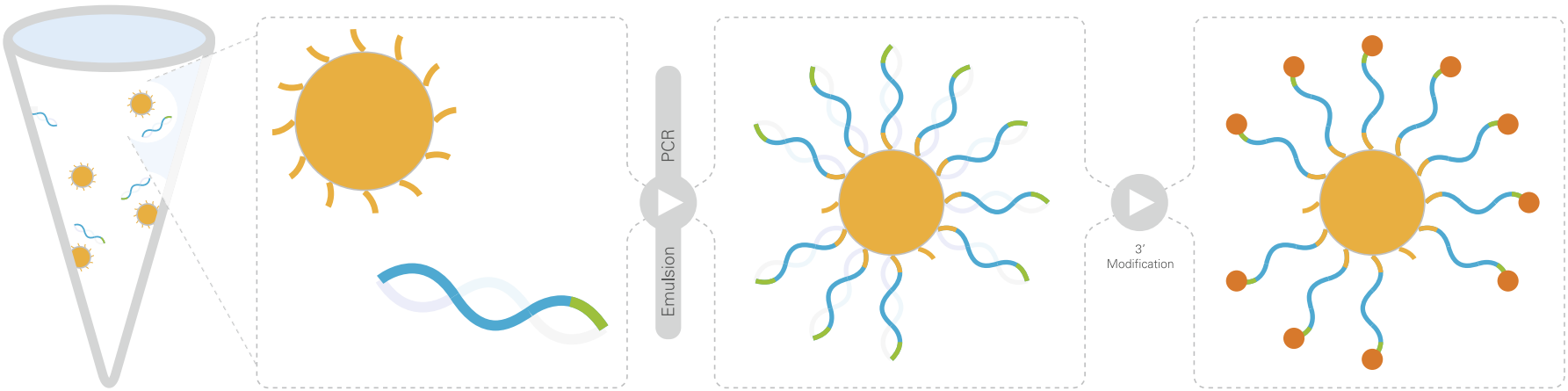
SOLiD Sequencing

- 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



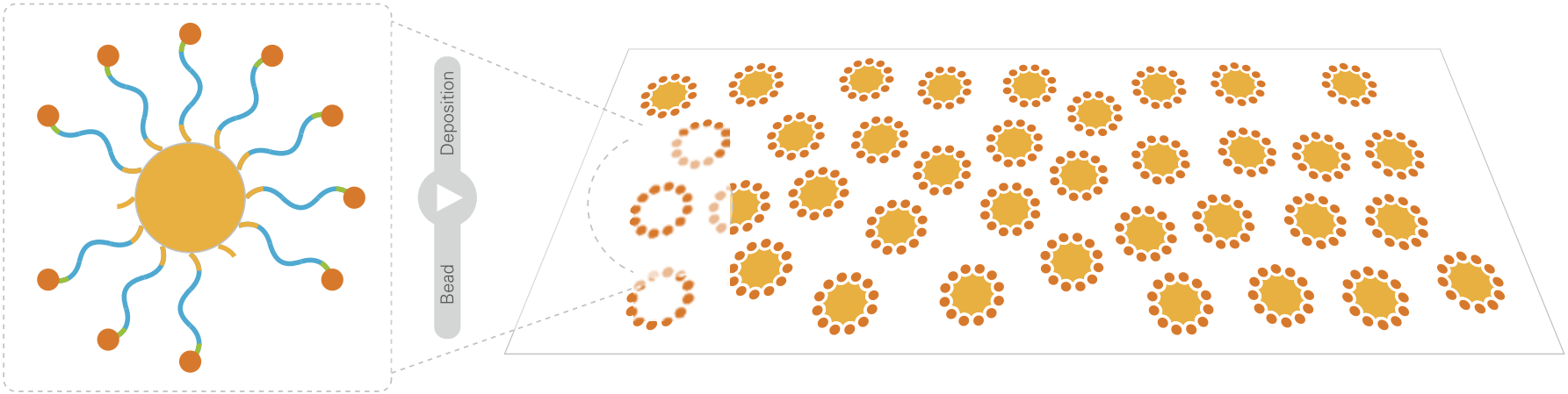
SOLiD Sequencing

- 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



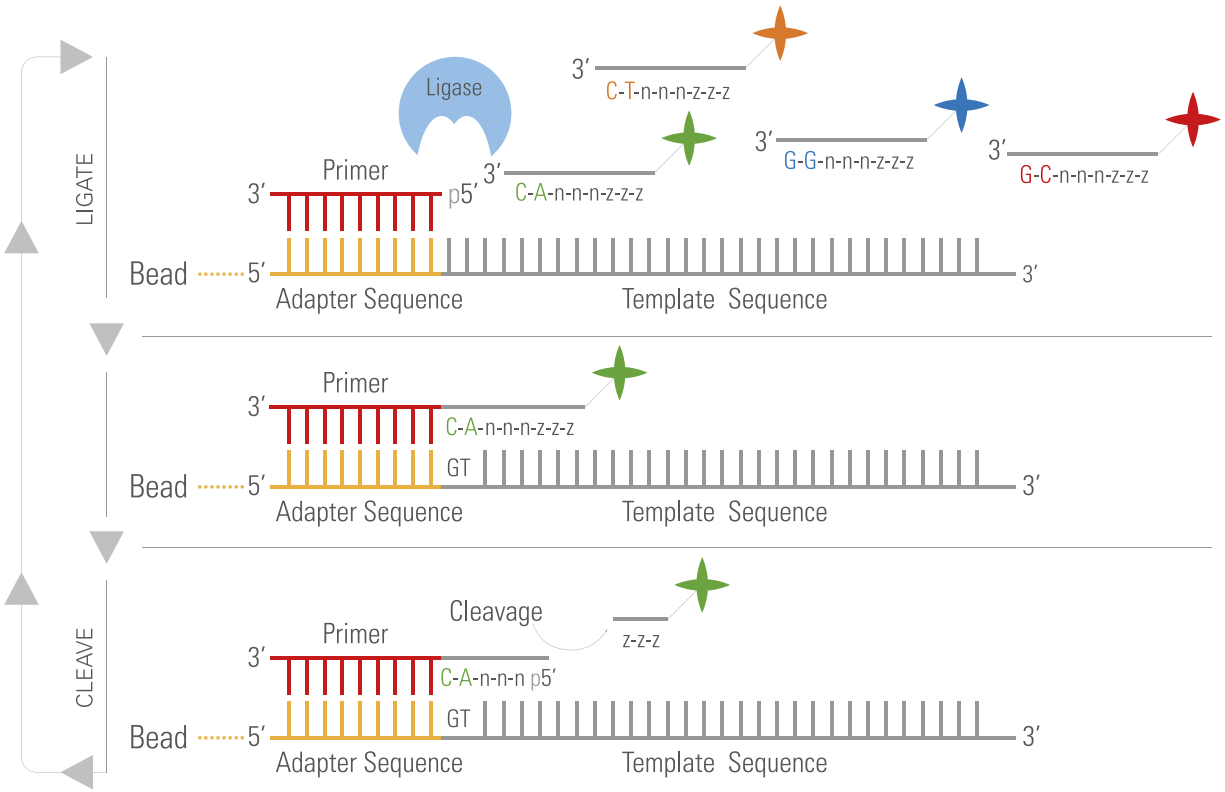
SOLiD Sequencing

- 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



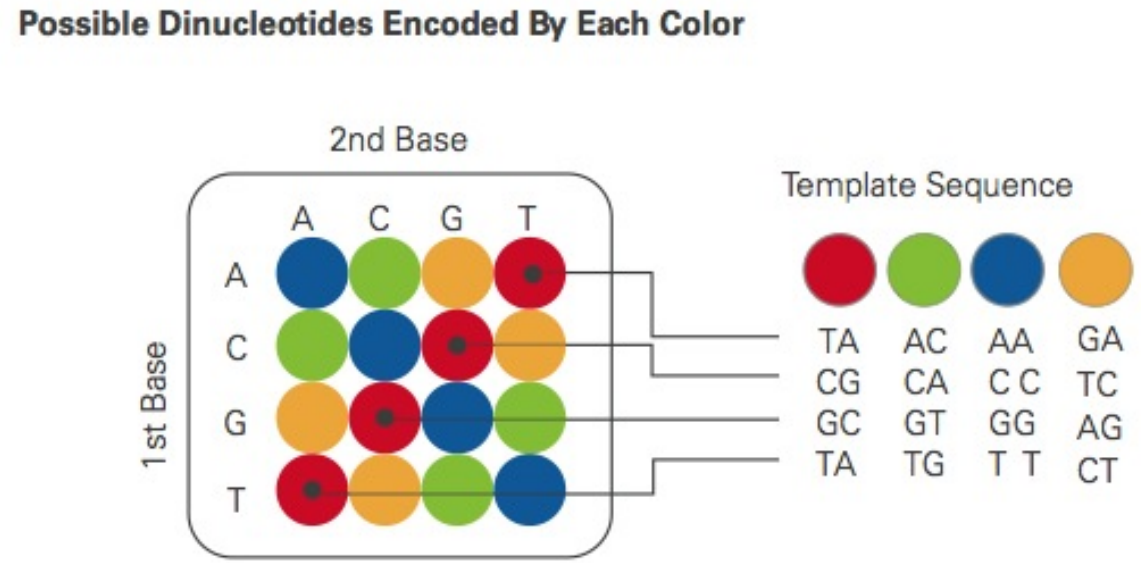
SOLiD Sequencing

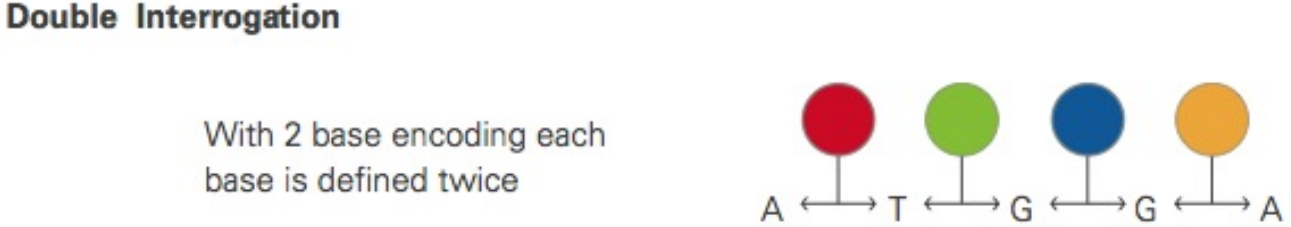
- Sequencing by ligation: Fragments ligated to “beads”
- PCR, beads enriched with fragments, ends of the templates modified to allow for an attachment to the slide
- Beads are deposited onto a glass slide
- Di-base probes compete for ligation to the sequencing primer



SOLiD Sequencing - Color Space

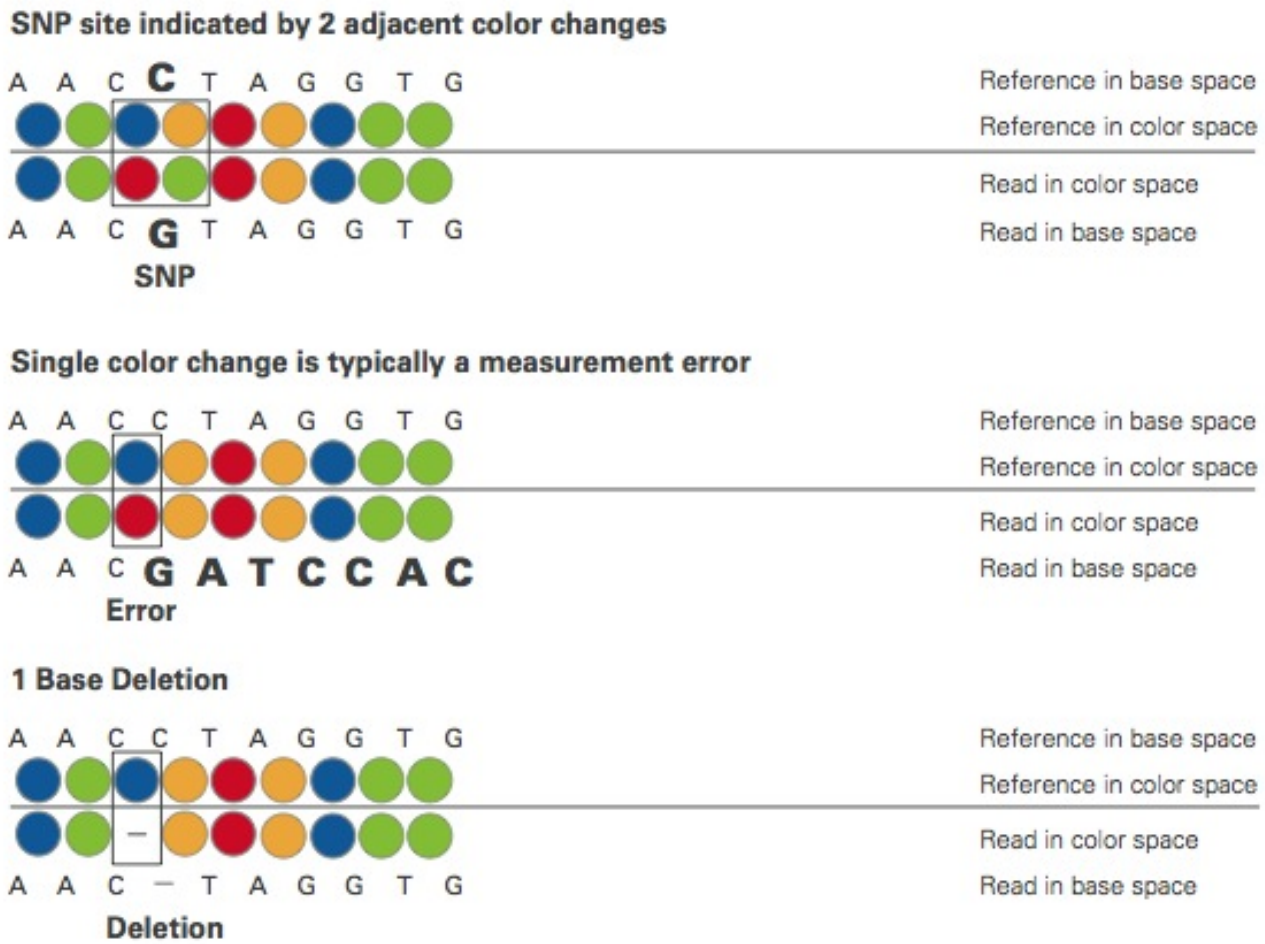
- 4 fluorescent dyes for 16 possible 2-mers
- Reverse, complement and reverse complement are always of same color





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*	$40 - 75$ bp	≈6-20 Gbp	$5 - 8d$
ABI SOLiD	$35 - 50$ bp	≈6-15 Gbp	$6 - 7d$
Roche/454	$300 - 500$ bp	≈100 Mbp	$7h$
Sanger	1000 bp	≈67 kbp	$1h$

There are several extensions available:

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

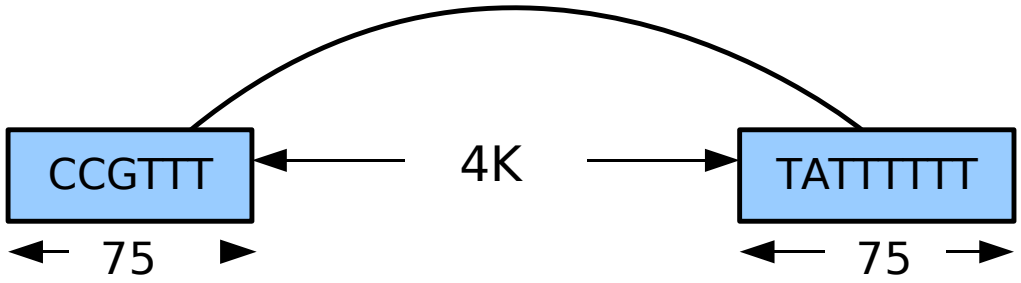
[Sutskever, 2008]

Overview / Extensions

Technology	Read length	Output/run	Run time
Illumina GA *II*	40 – 75 bp	≈6-20 Gbp	$5 - 8d$
ABI SOLiD	35 – 50 bp	≈6-15 Gbp	$6 - 7d$
Roche/454	300 – 500 bp	≈100 Mbp	$7h$
Sanger	1000 bp	≈67 kbp	$1h$

There are several extensions available:

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



[Sutskever, 2008]

Overview / Extensions

Technology	Read length	Output/run	Run time
Illumina GA *II*	$40 - 75$ bp	$\approx$6-20 Gbp	$5 - 8d$
ABI SOLiD	$35 - 50$ bp	$\approx$6-15 Gbp	$6 - 7d$
Roche/454	$300 - 500$ bp	$\approx$100 Mbp	$7h$
Sanger	1000 bp	$\approx$67 kbp	$1h$

There are several extensions available:

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

[Sutskever, 2008]

Overview / Extensions

Technology	Read length	Output/run	Run time
Illumina GA *II*	$40 - 75$ bp	$\approx$6-20 Gbp	$5 - 8d$
ABI SOLiD	$35 - 50$ bp	$\approx$6-15 Gbp	$6 - 7d$
Roche/454	$300 - 500$ bp	$\approx$100 Mbp	$7h$
Sanger	1000 bp	$\approx$67 kbp	$1h$

There are several extensions available:

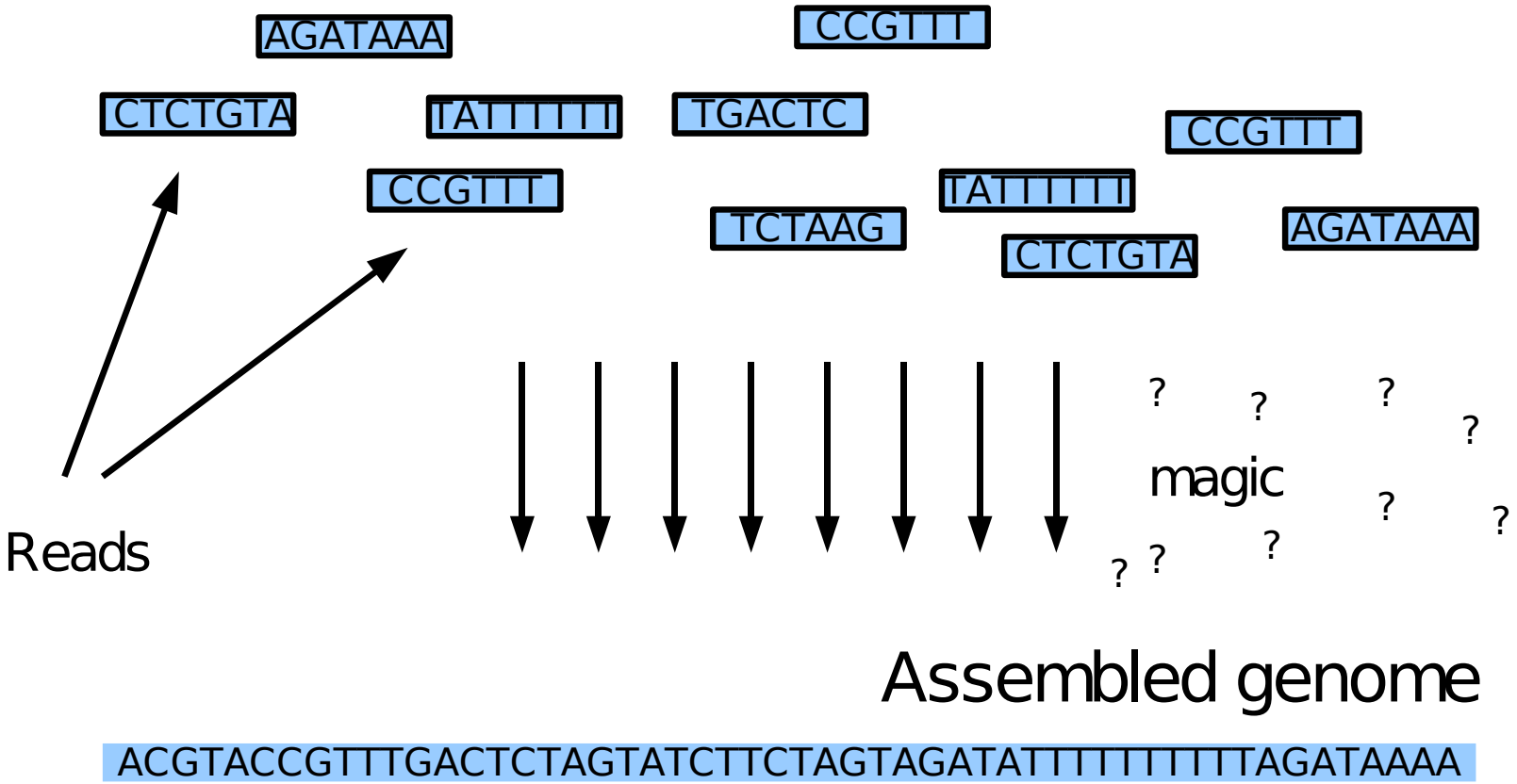
- mate-pair / paired-end
- bisulfite treatment
- multiplexing 8×12
  $\rightsquigarrow$ 96 samples per flow cell

[Sutskever, 2008]

Short Reads Analysis - Methods

Given read data the following analysis steps are possible:

- Assembly
- Mapping/Alignments



[Sutskever, 2008]

Short Reads Analysis - Methods



Given read data the following analysis steps are possible:

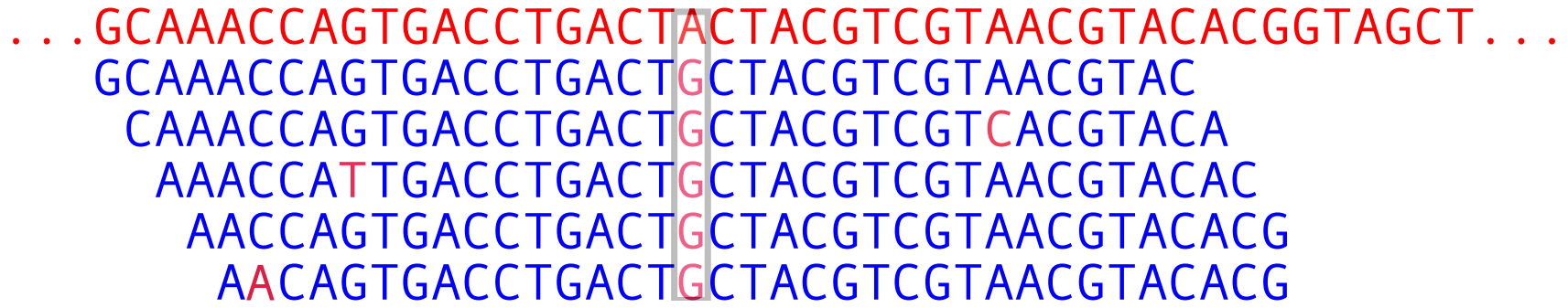
- Assembly
- Mapping/Alignments

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

Short Reads Analysis - Methods

Given read data the following analysis steps are possible:

- Assembly
- Mapping/Alignments



Short Reads Analysis - Methods

Given read data the following analysis steps are possible:

- Assembly
- Mapping/Alignments

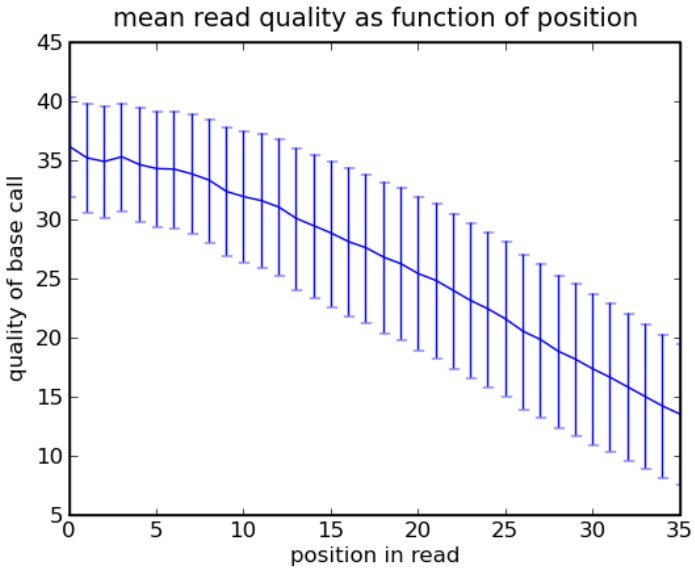
```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT...
   GCAAACCAGTGACCTGACTGCTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTGCTACGTCGTCACGTACA
     AAACCATTGACCTGACTGCTACGTCGTAACGTACAC
      AACCAGTGACCTGACTGCTACGTCGTAACGTACACG
       AACAGTGACCTGACTGCTACGTCGTAACGTACACG
```

Problem: 100 million reads of short length

$\Rightarrow$ Big computational challenge

Short Reads Analysis - Problems

- Experiment leads to millions of reads $36 - 75$nt
- 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 the quality is 30

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: Plethora of 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: Plethora of 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: Plethora of 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**

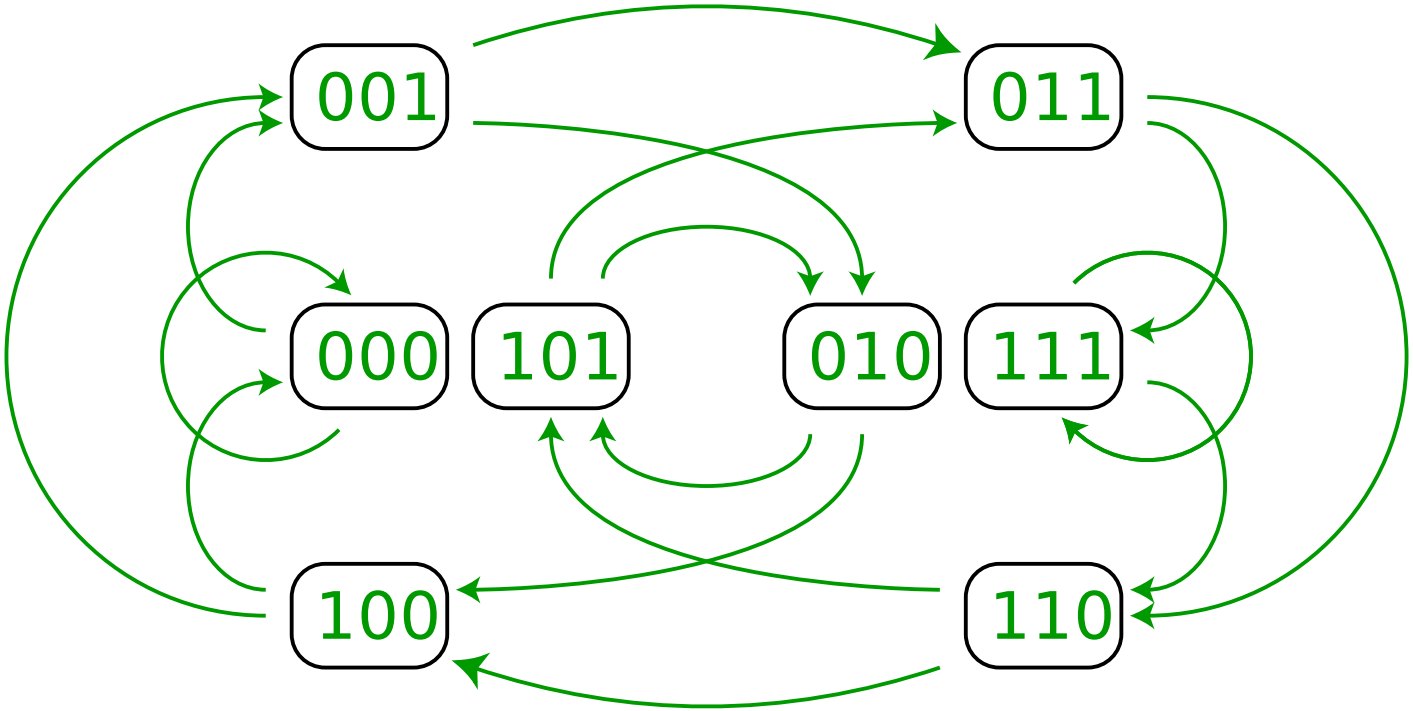
1. *Overlap phase*: Every read is compared with every other read and the overlap graph is computed
2. *Layout phase*: Pairs are determined that position every read in the assembly
3. *Consensus phase*: Multi-alignment of all the placed reads is produced to obtain the final sequence

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

Idea: de Bruijn Graphs

Short Reads Assembly - de Bruijn Graphs

Example:

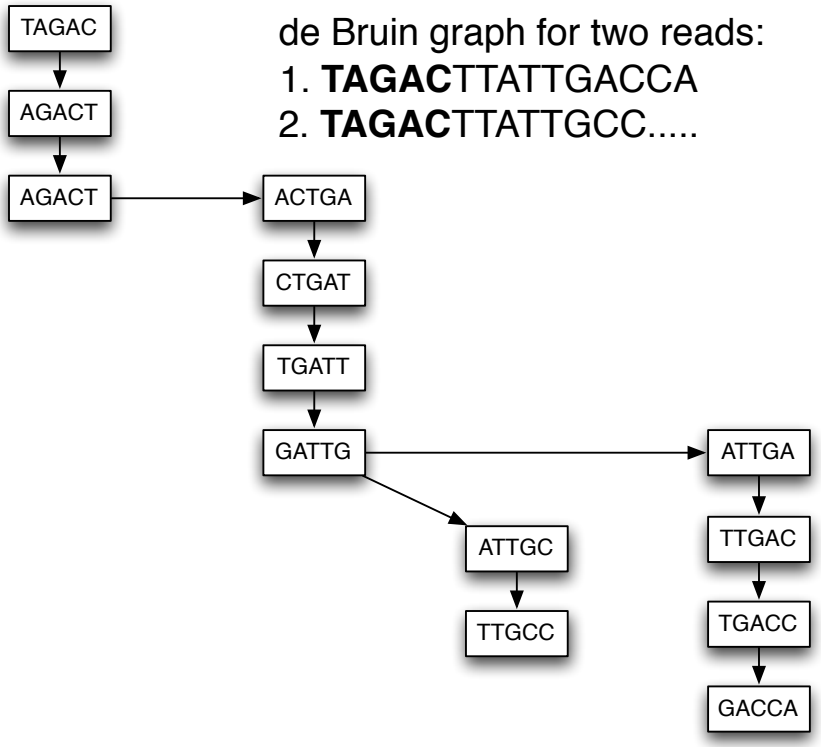


- Reads are mapped as a path in the graph
- Number of reads does not influence number of nodes
  $\Rightarrow$ Use de-Bruijn graphs to solve the problem

[Wikipedia]

de Bruijn Graphs

Example:



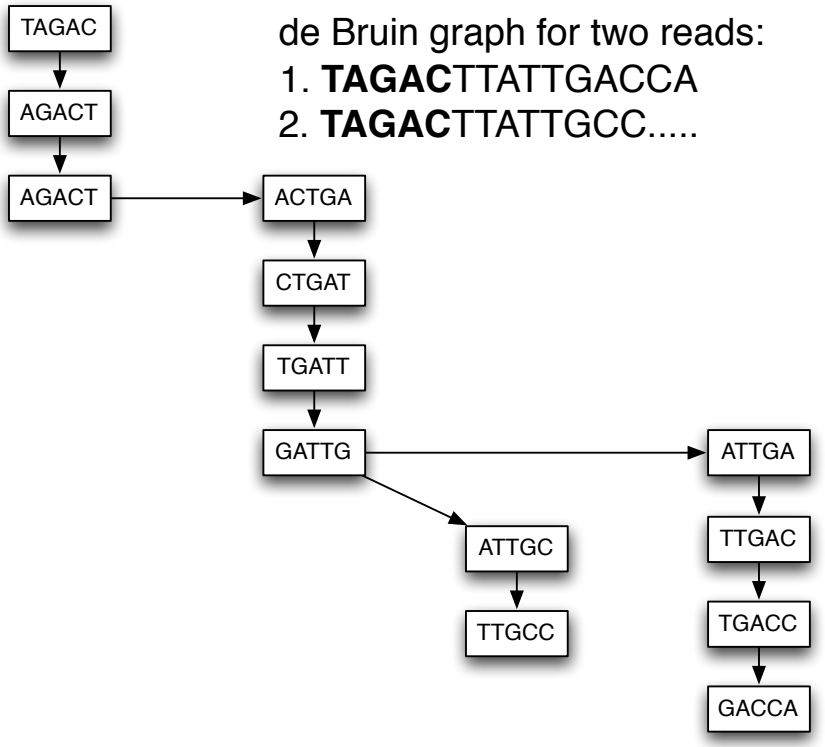
de Bruin graph for two reads:
1. **TAGAC**TTATTGACCA
2. **TAGAC**TTATTGCC.....

[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:



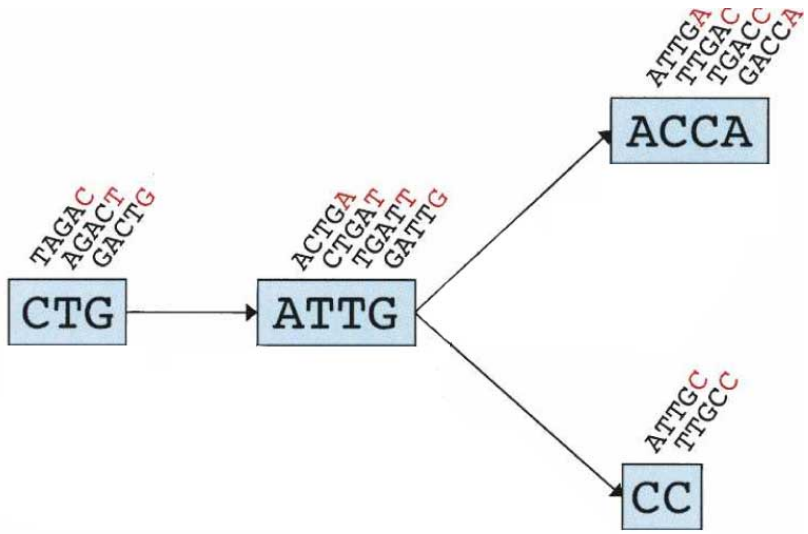
de Bruin graph for two reads:
1. **TAGAC**TTATTGACCA
2. **TAGAC**TTATTGCC.....

[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:

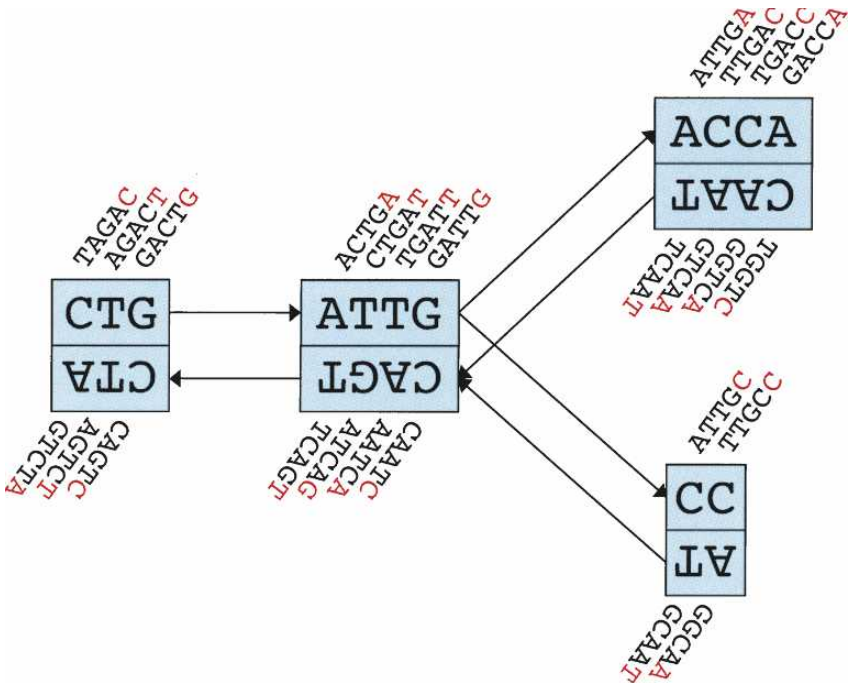


[Zerbino and Birney, 2008]

- 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

de Bruijn Graphs

Example:



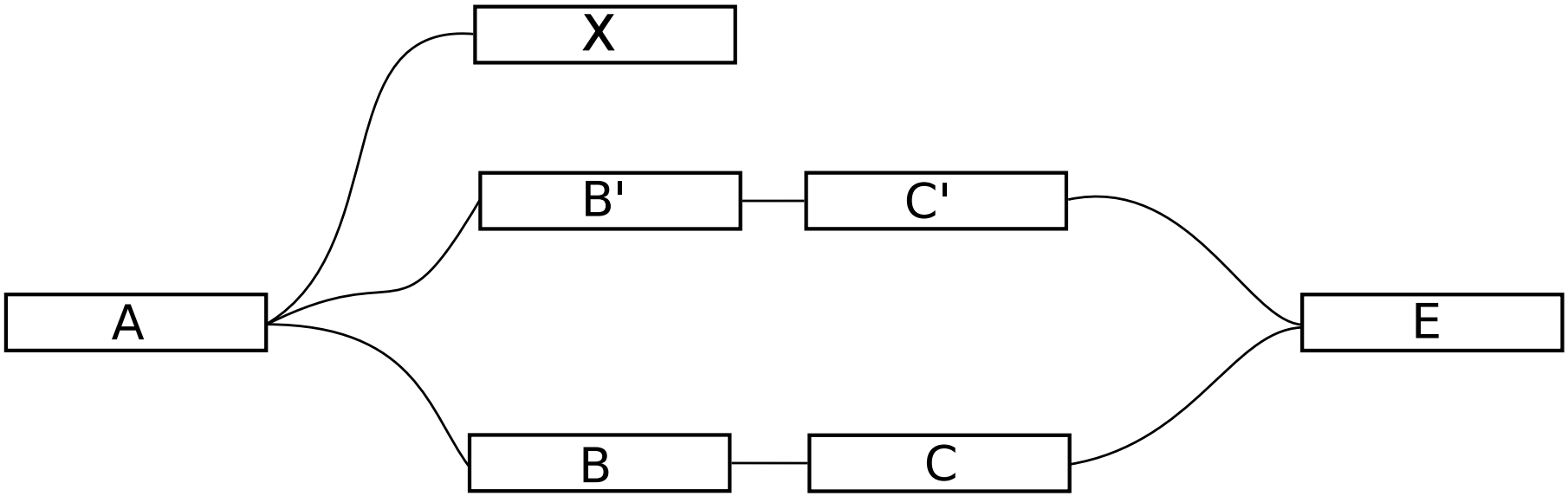
[Zerbino and Birney, 2008]

- 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



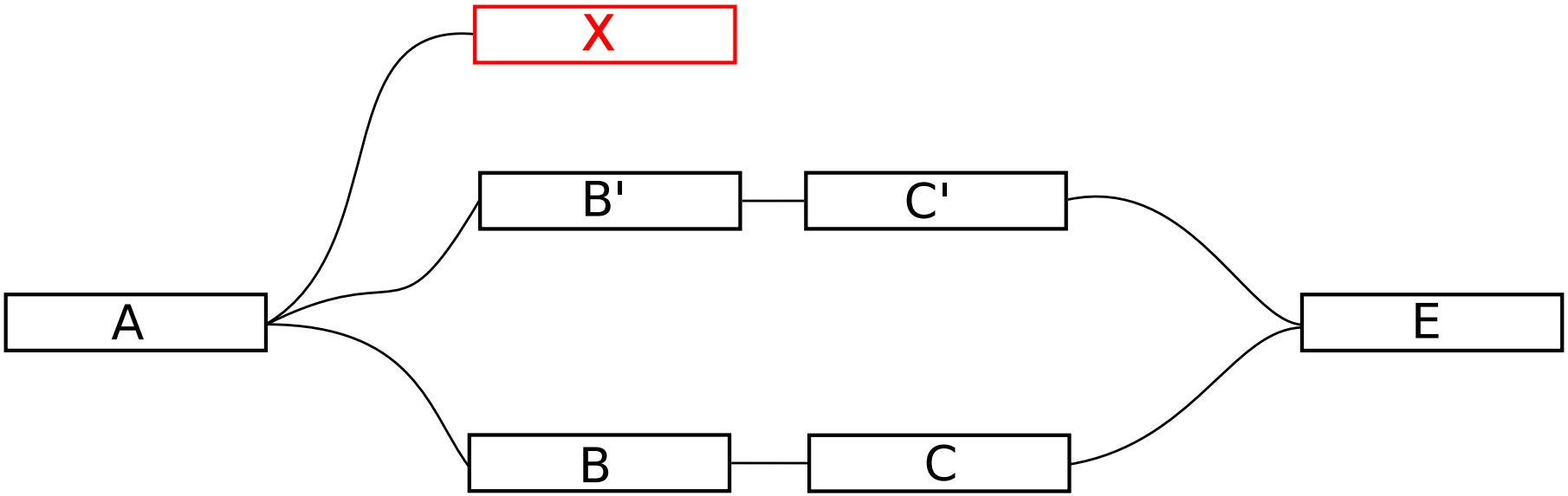
Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



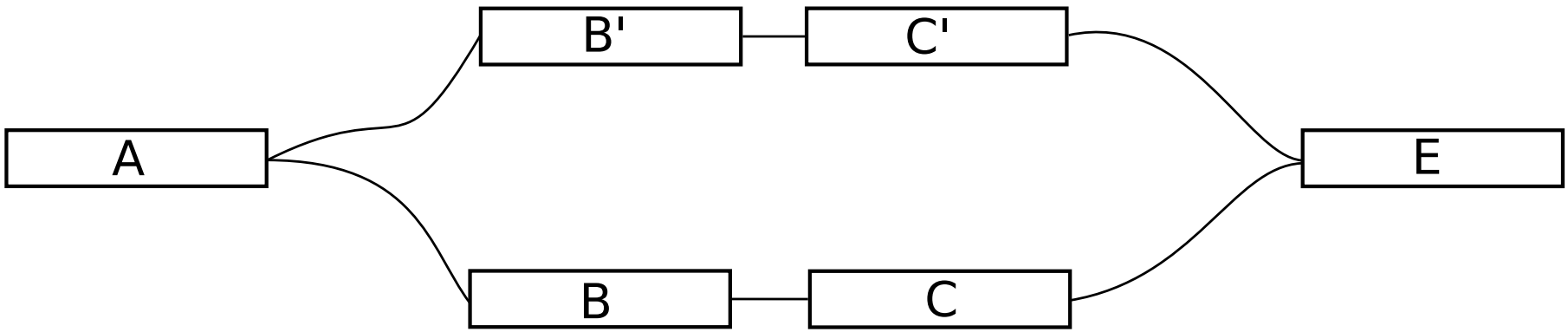
Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



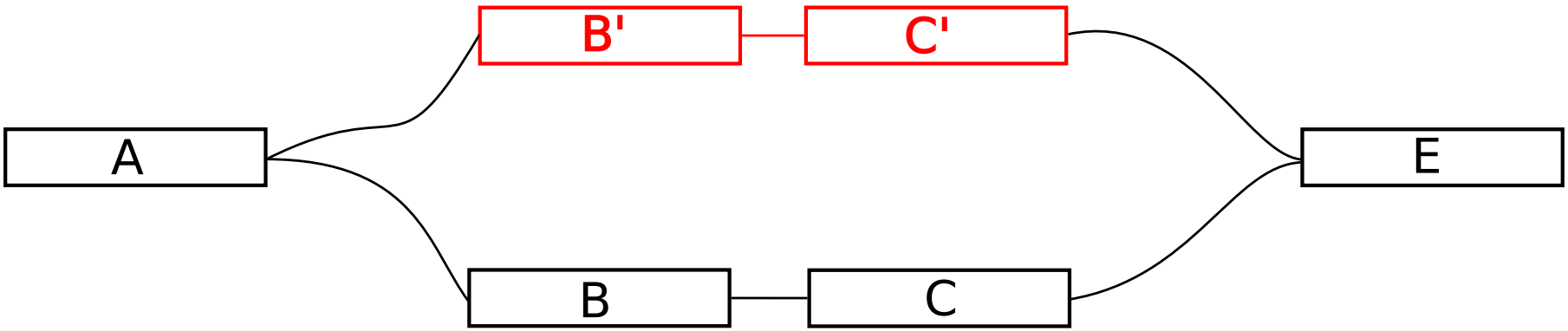
Read off genome sequence (if everything goes well ;-)

[Zerbino and Birney, 2008]

Short Reads Assembly - Example

Consider a constructed de Bruijn graph

- Unconnected nodes
- Ambiguous paths
- Erroneous edges



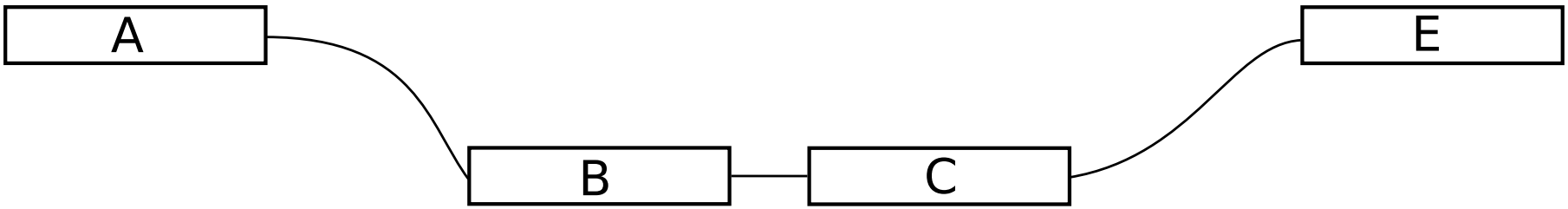
Read off genome sequence (if everything goes well ;-)

[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



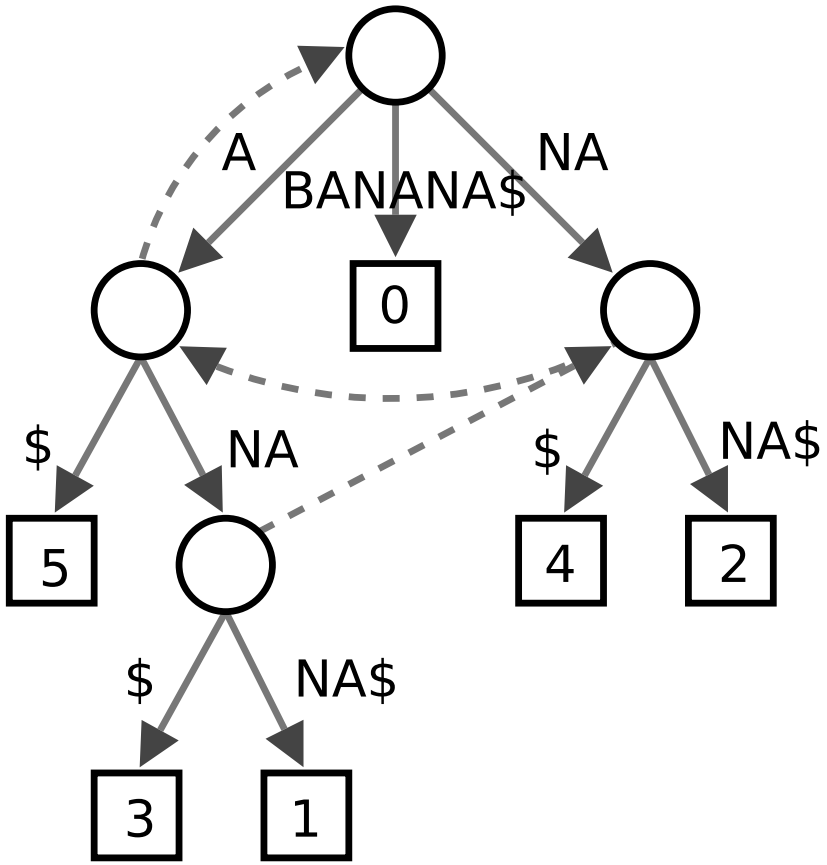
Reads mapping problem

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*, ...

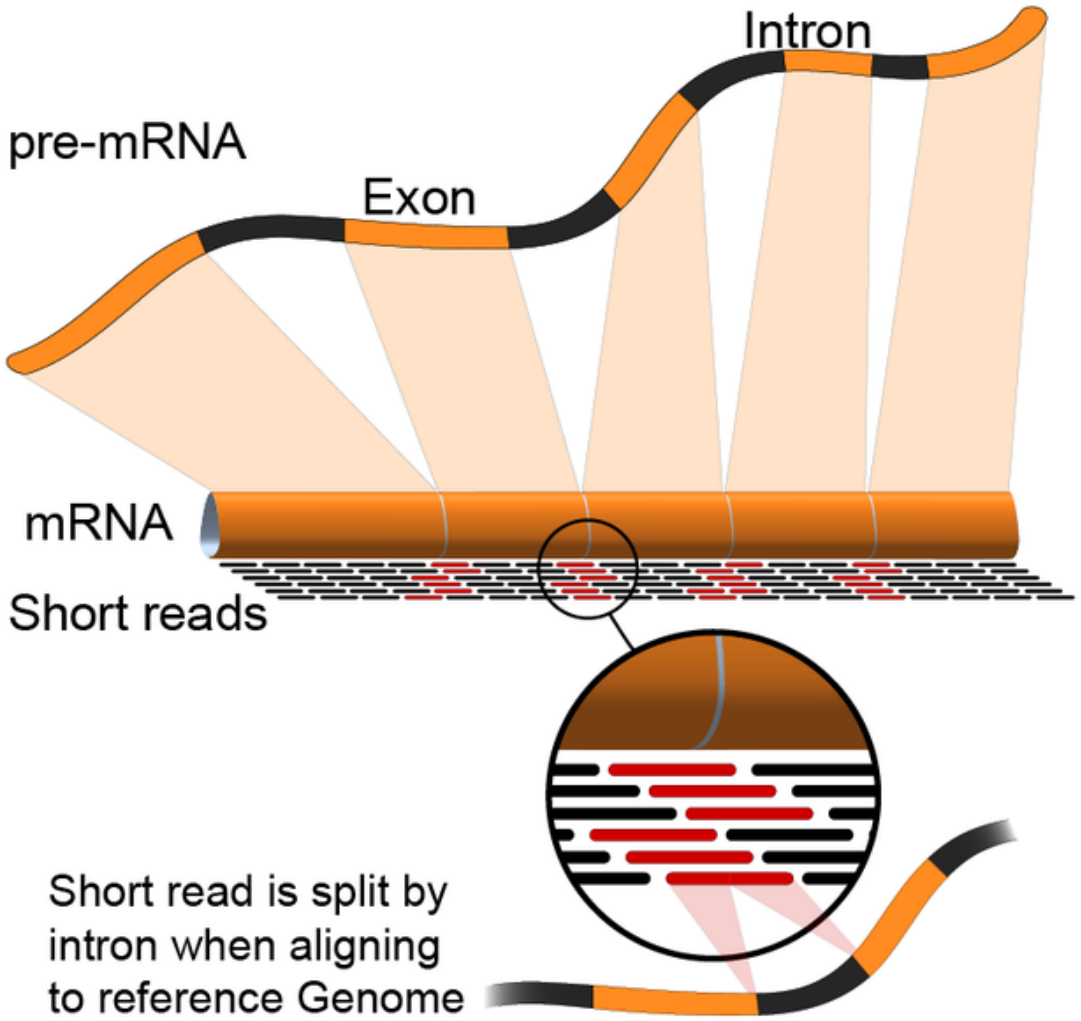
Mapping via Suffix Arrays

- 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 $\mathcal{O}(m)$



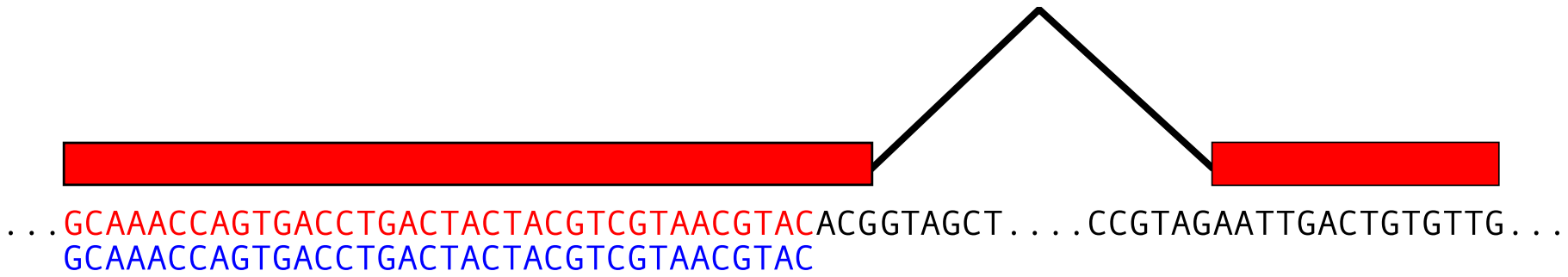
[Wikipedia]

Spliced vs. Unspliced Alignments



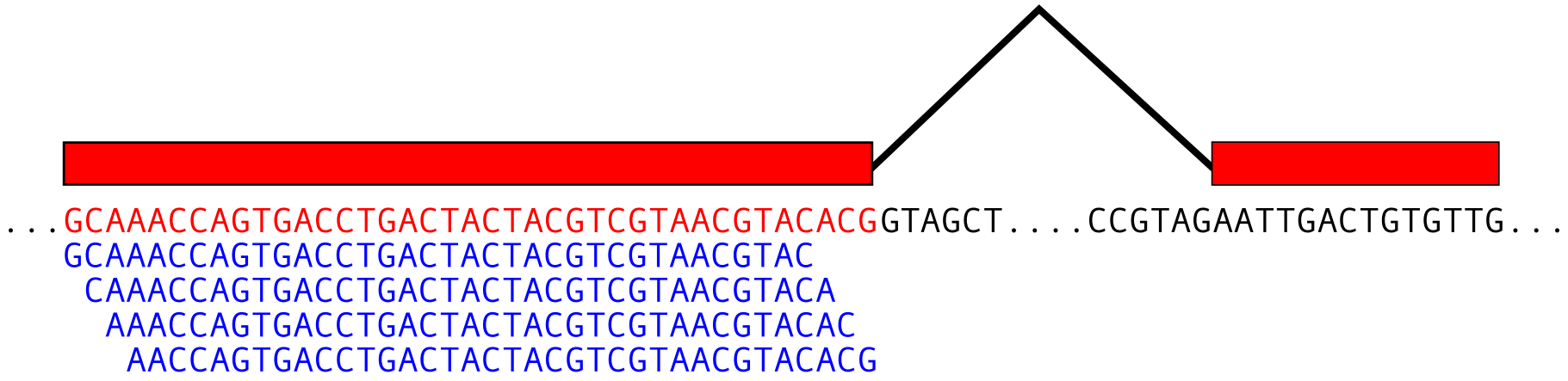
[Wikipedia]

Spliced vs. Unspliced Alignments



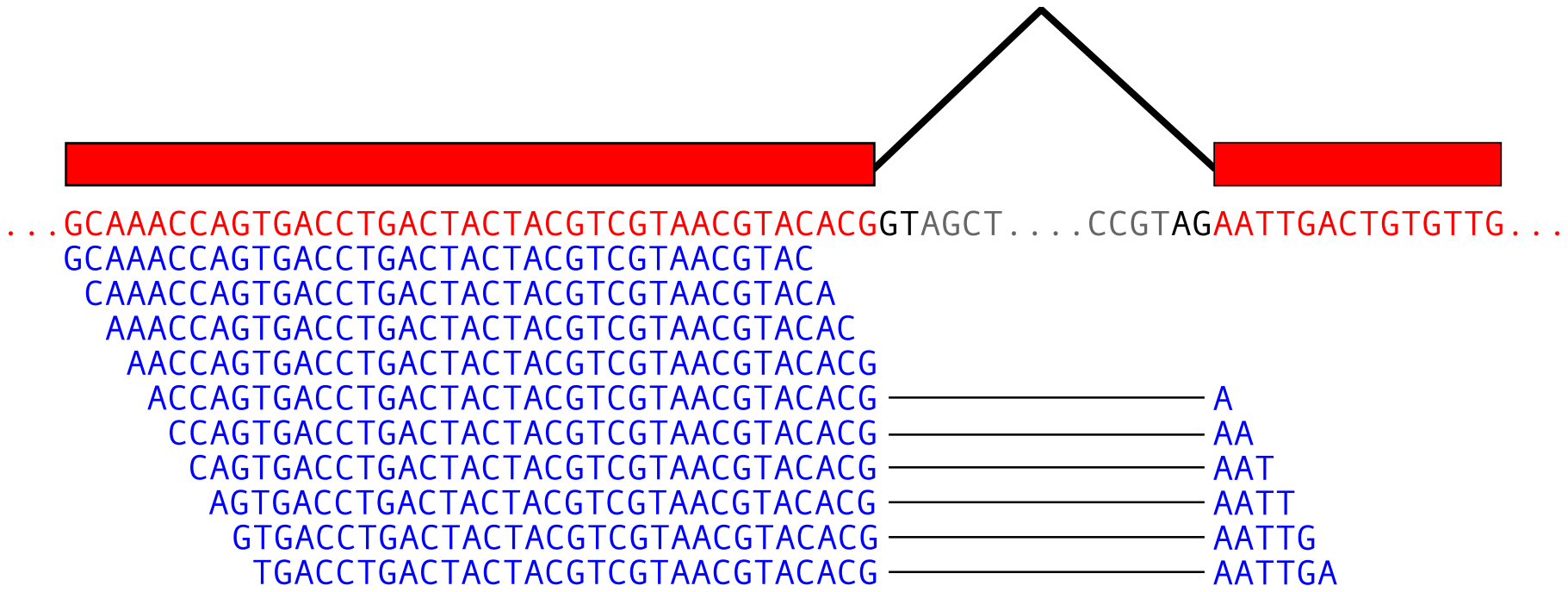
- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many reads
- Most current mapping techniques are limited to unspliced reads

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

Extended Smith-Waterman Algorithm



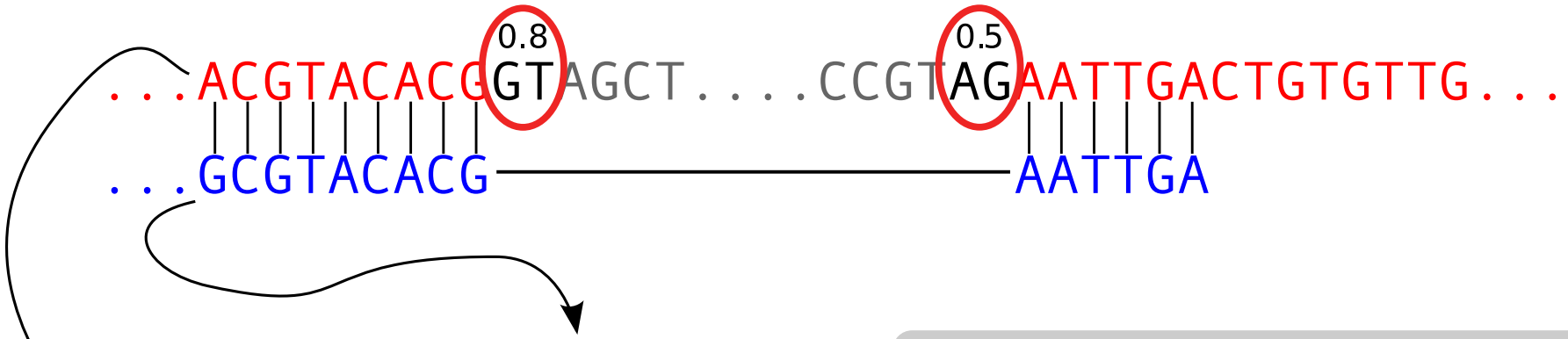
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Extended Smith-Waterman Algorithm



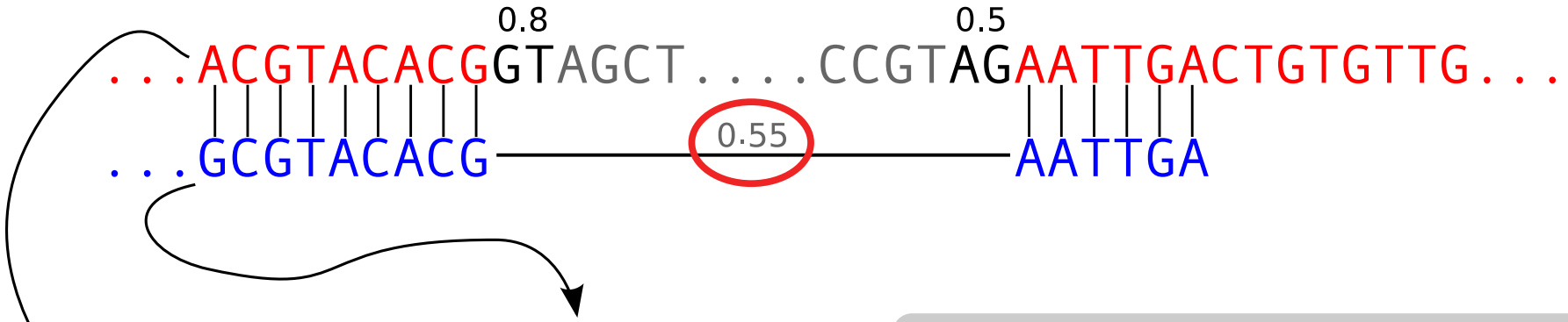
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Extended Smith-Waterman Algorithm



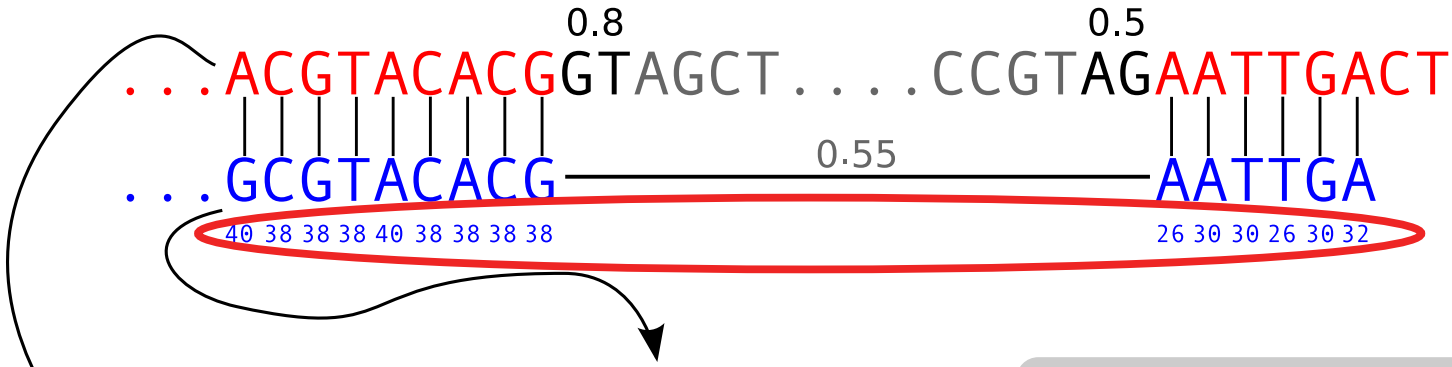
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Extended Smith-Waterman Algorithm



	gap	A	C	G	T	N
gap	0.33	$f_{-A}(\cdot)$	$f_{-C}(\cdot)$	$f_{-G}(\cdot)$	$f_{-T}(\cdot)$	$f_{-N}(\cdot)$
A	0.31	$f_{AA}(\cdot)$	$f_{AC}(\cdot)$	$f_{AG}(\cdot)$	$f_{AT}(\cdot)$	$f_{AN}(\cdot)$
C	0.44	$f_{CA}(\cdot)$	$f_{CC}(\cdot)$	$f_{CG}(\cdot)$	$f_{CT}(\cdot)$	$f_{CN}(\cdot)$
G	0.13	$f_{GA}(\cdot)$	$f_{GC}(\cdot)$	$f_{GG}(\cdot)$	$f_{GT}(\cdot)$	$f_{GN}(\cdot)$
T	0.55	$f_{TA}(\cdot)$	$f_{TC}(\cdot)$	$f_{TG}(\cdot)$	$f_{TT}(\cdot)$	$f_{TN}(\cdot)$
N	0.12	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$

Source of Information

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

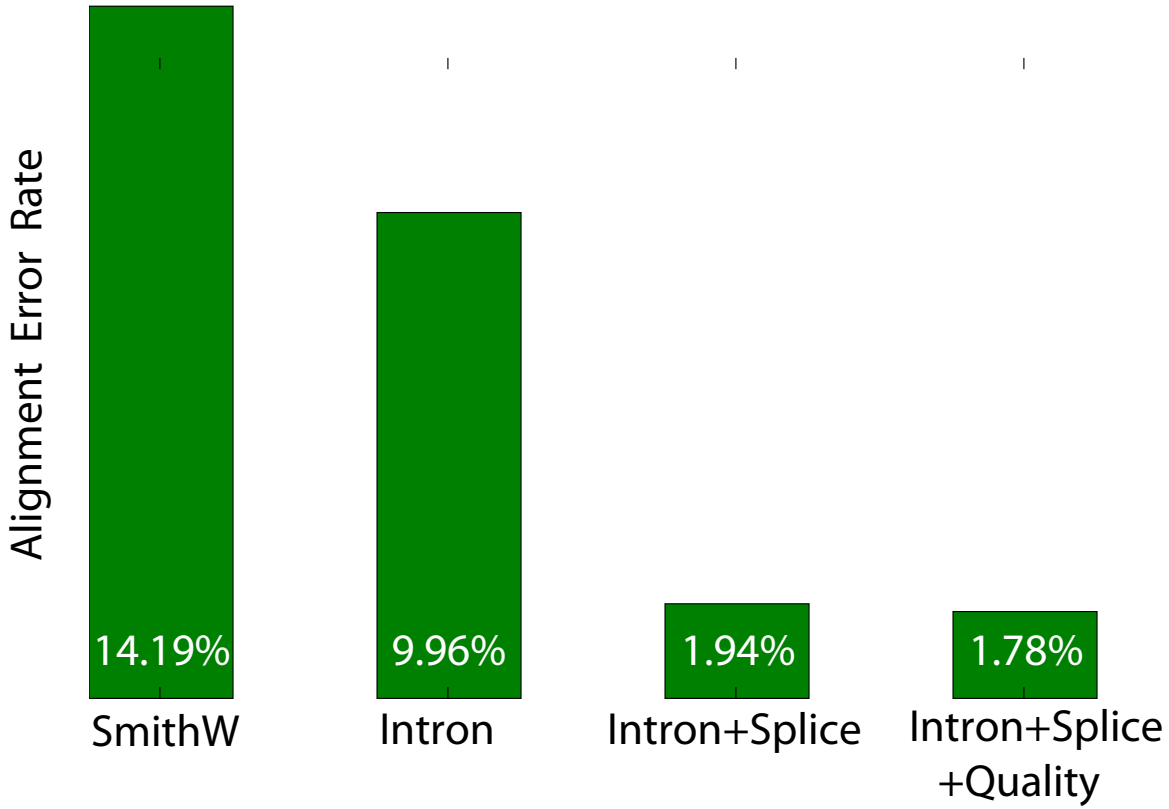
Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

[De Bona et al., 2008]

QPalma's Accurate Alignments

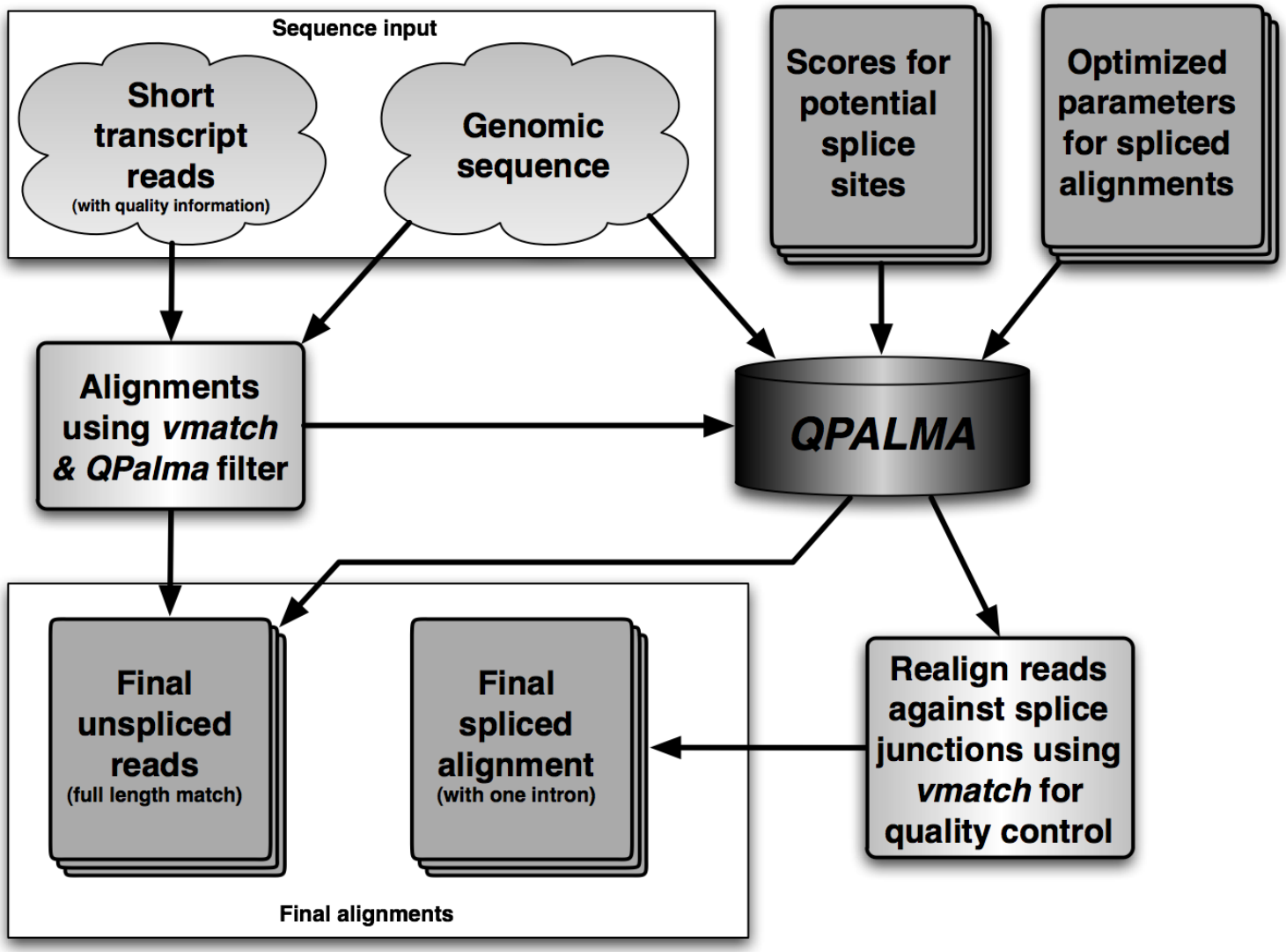
Generate set of artificially spliced reads

- 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]

An Alignment Pipeline



[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]

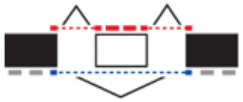
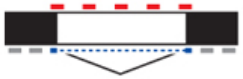
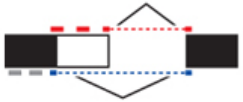
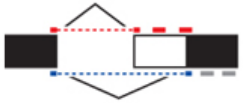
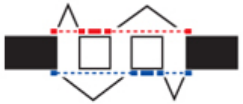
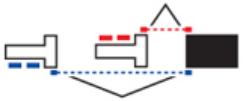
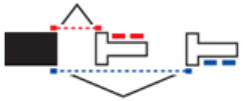
15 AUGUST 2008 VOL 321 SCIENCE www.sciencemag.org

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] Aleksey Soldatov,[1] Dmitri Parkhomchuk,[1] Dominic Schmidt,[1] Sean O'Keeffe,[2] Stefan Haas,[2] Martin Vingron,[2] Hans Lehrach,[1] Marie-Laure Yaspo[1]†

[Sultan 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

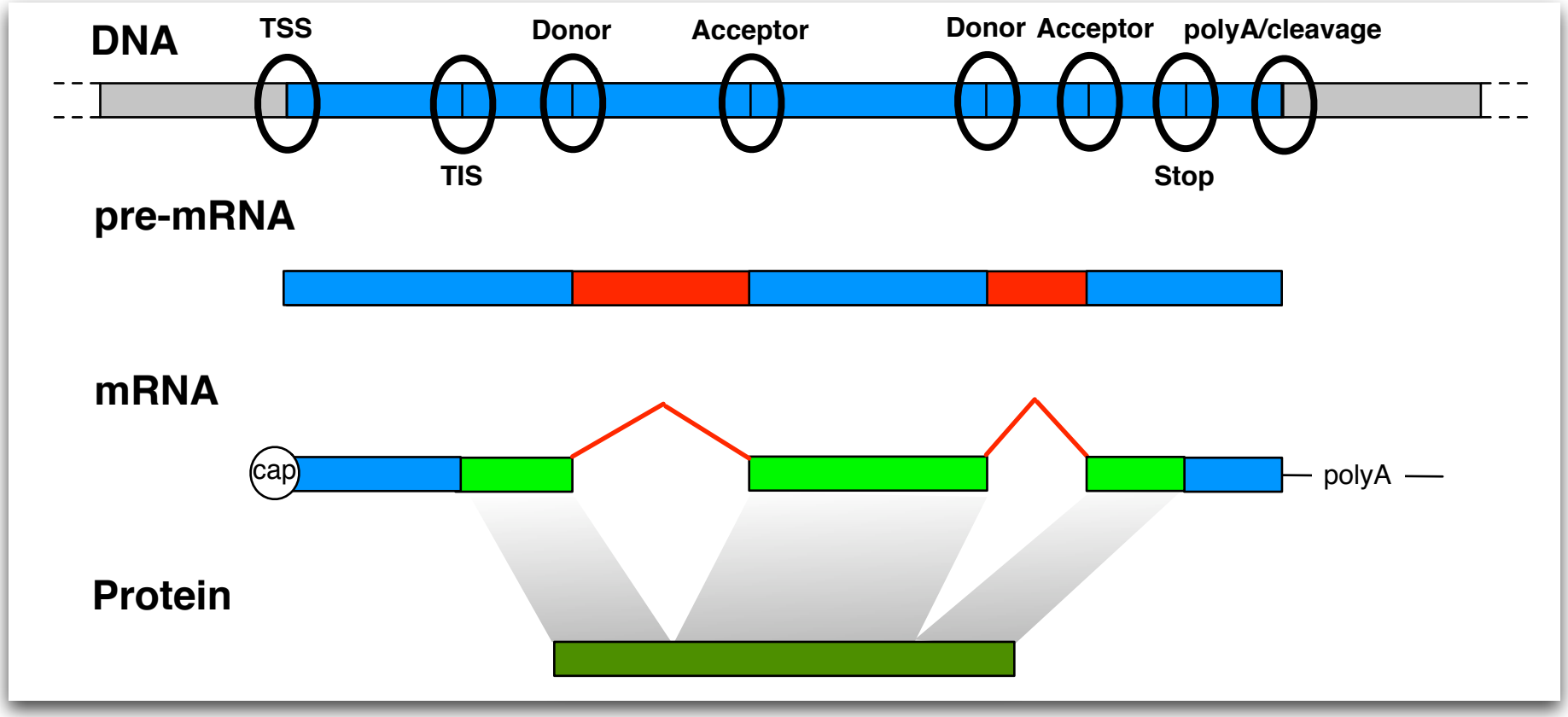
Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

[Wang et al., 2008]

Roadmap

1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Identification of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology & Limitations
   - Assembly & Read Mapping
4. Extensions
   - Quantification of Transcripts
   - ChIP-on-Chip Studies

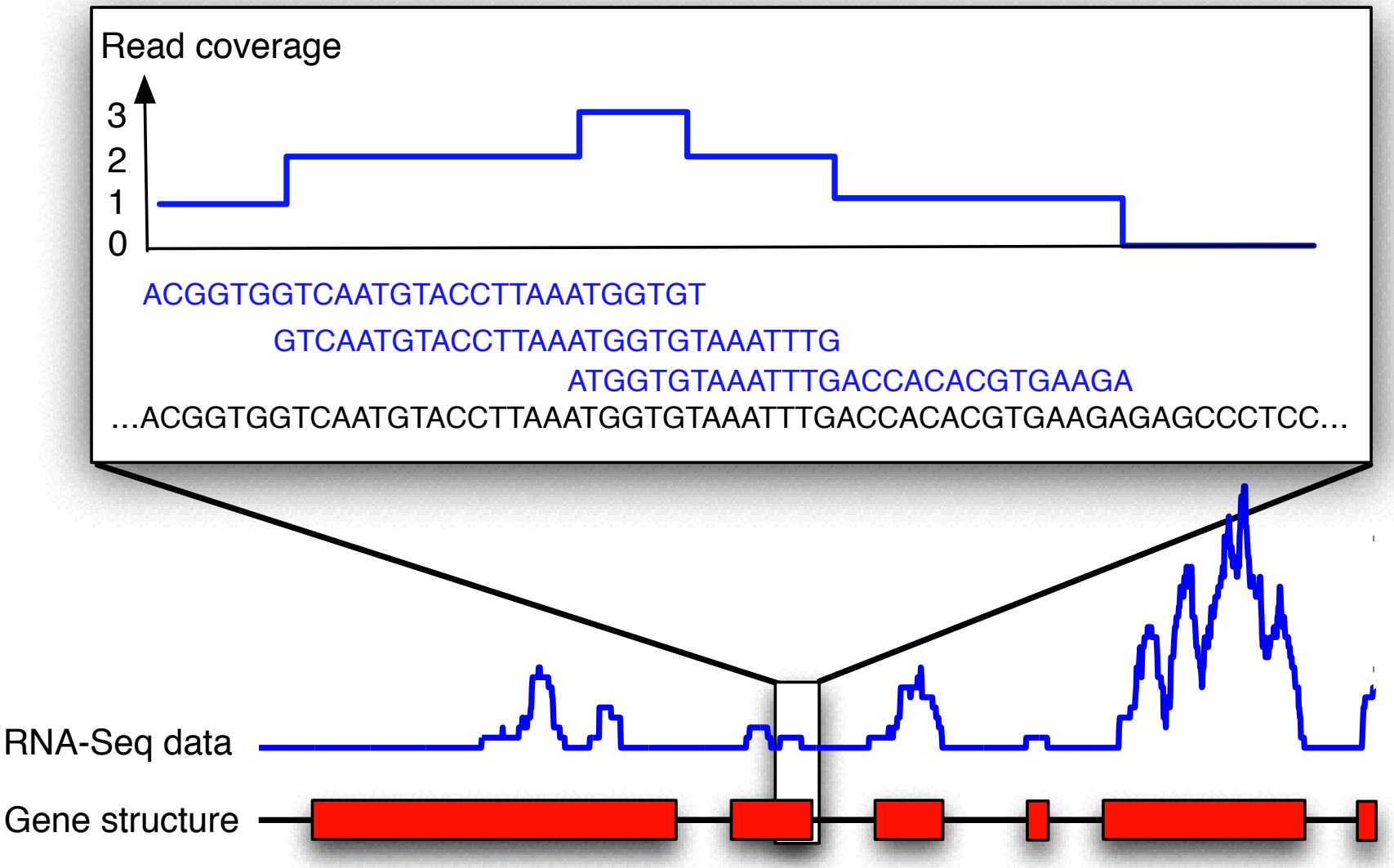
From Genome & Measurements to Proteins



Combine *ab initio* predictions with transcriptome measurements

- Higher accuracy
- Condition/tissue specific predictions

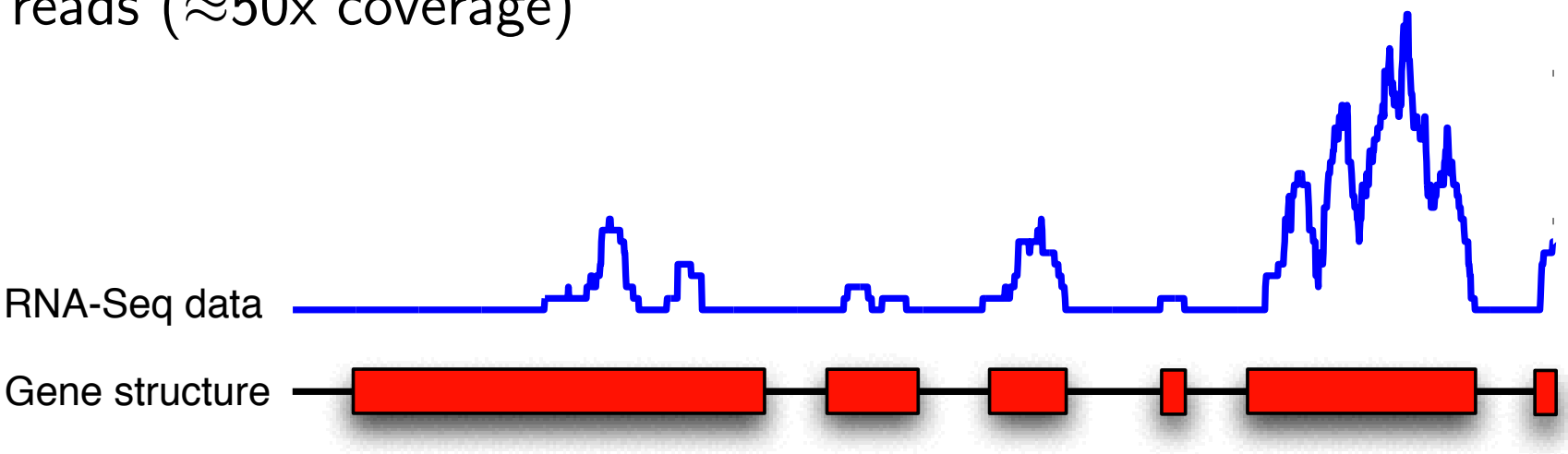
RNA-seq Data


Read coverage
3
2
1
0
ACGGTGGTCAATGTACCTTAAATGGTGT
GTCAATGTACCTTAAATGGTGTAAATTTG
ATGGTGTAAATTTGACCACACGTGAAGA
...ACGGTGGTCAATGTACCTTAAATGGTGTAAATTTGACCACACGTGAAGAGAGCCCTCC...
RNA-Seq data
Gene structure

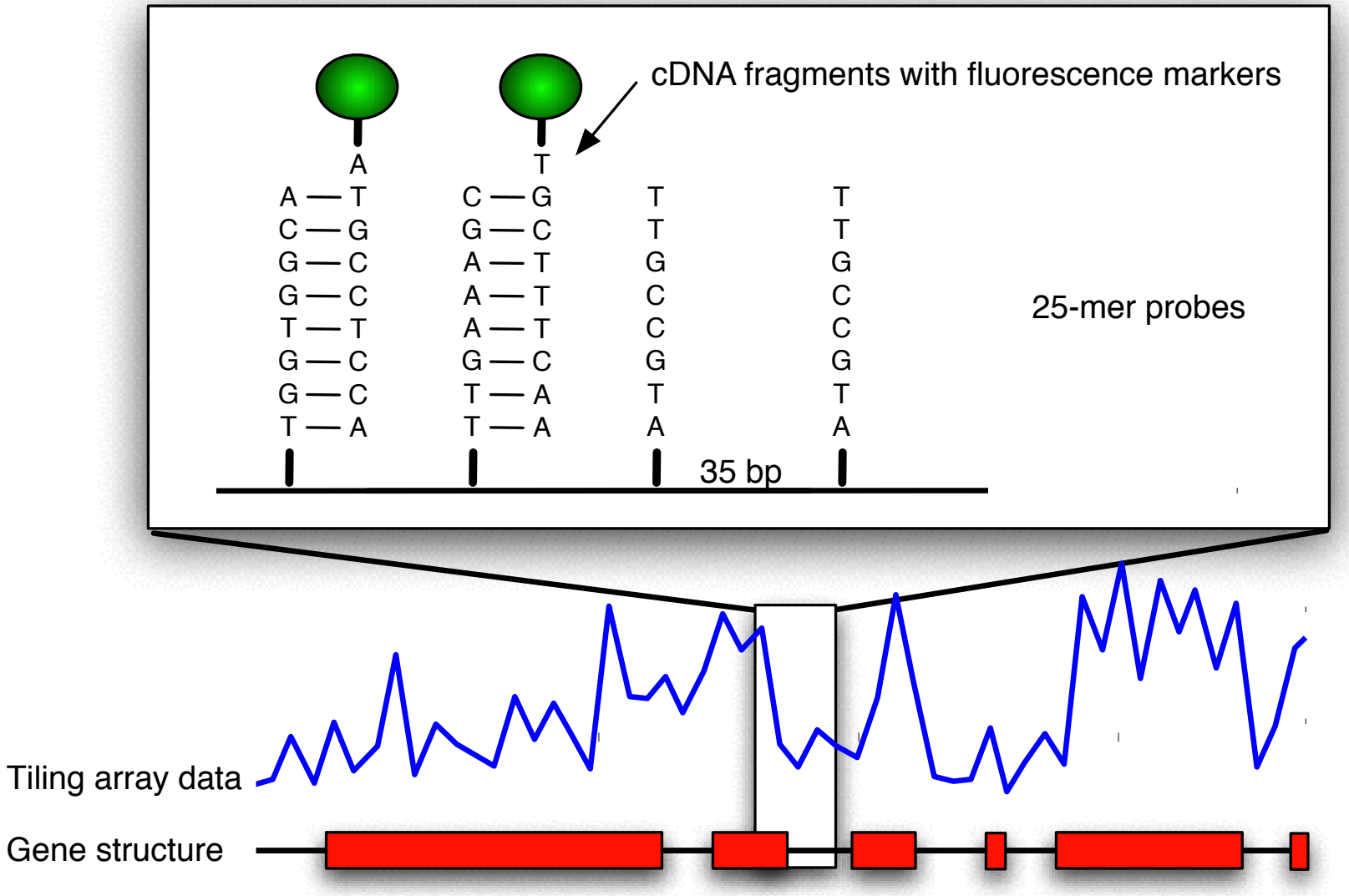
RNA-seq Data

RNA-seq reads for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

- 4 lanes from Illumina Genome Analyzer
- 38nt reads, polyA enriched
- Strand unspecific, young leaves
- Read mapping using **ShoRe** [Ossowski et al., 2008]
- Spliced alignments using **QPalma** [De Bona et al., 2008]
- ≈30 million unspliced and ≈1 million spliced reads (≈50x coverage)



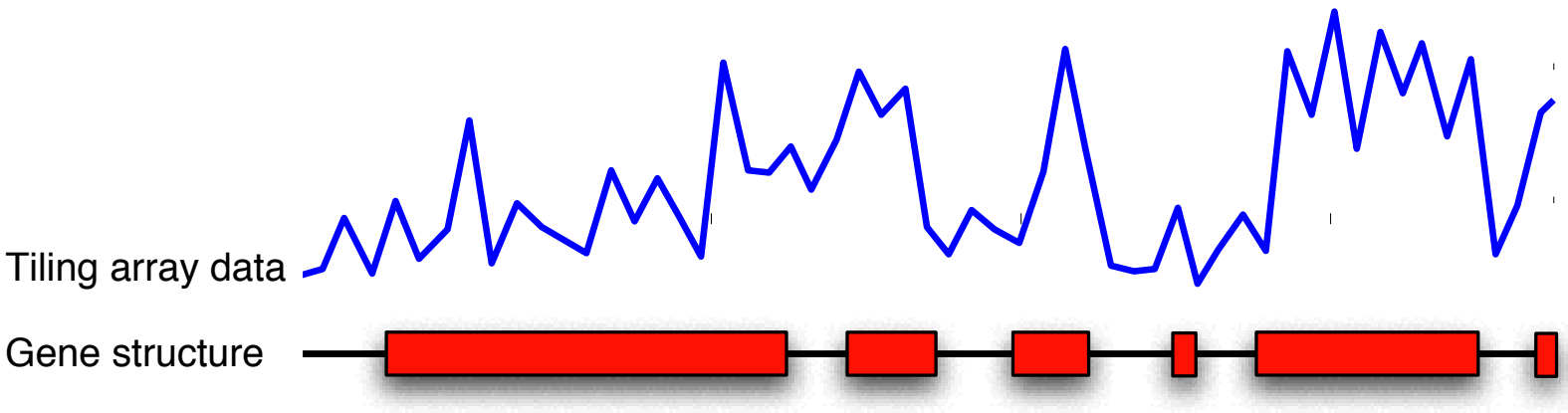
Tiling Array Data



Tiling Array Data

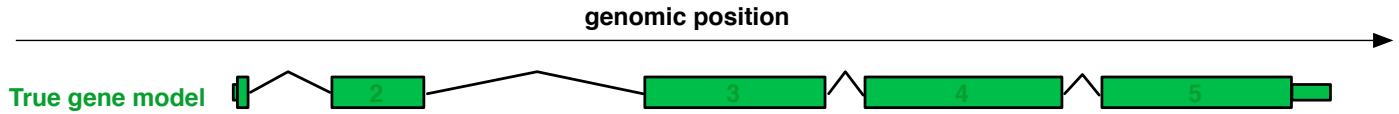
Tiling arrays for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

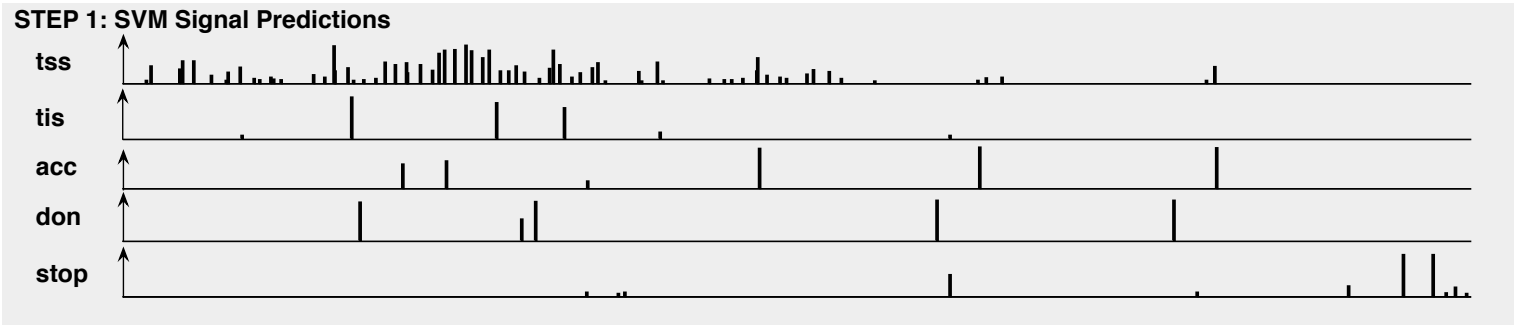
- 25nt probes, 35nt spacing, 3 replicates
- Strand unspecific, polyA enriched
- 12 different tissues (young leaves, root, etc.)
- 12 conditions/mutants (e.g. abiotic stresses)
- Quantile and sequence dependent normalization [Zeller et al., 2008b]



Learning to Integrate Data Sources

[Behr et al., 2008]

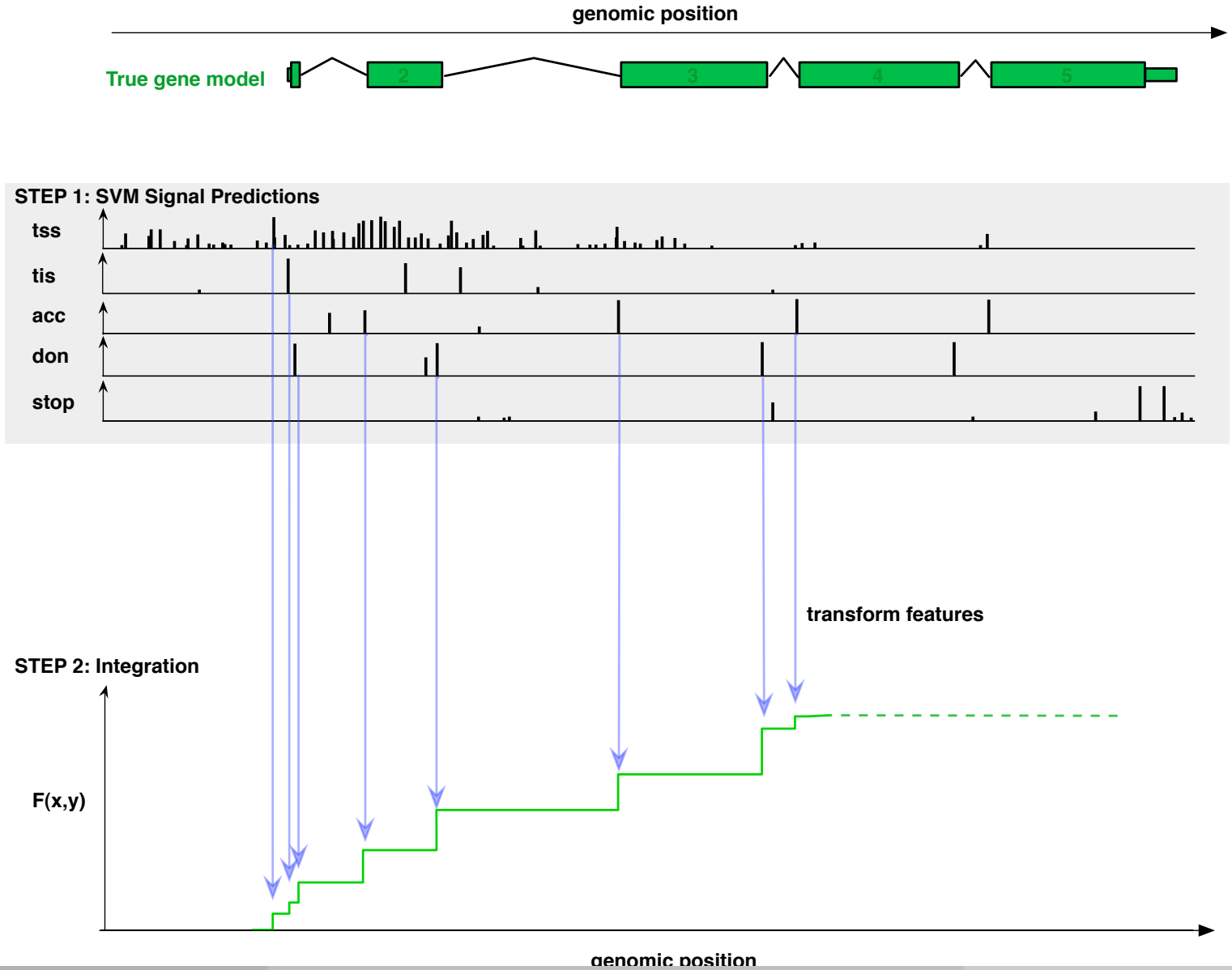






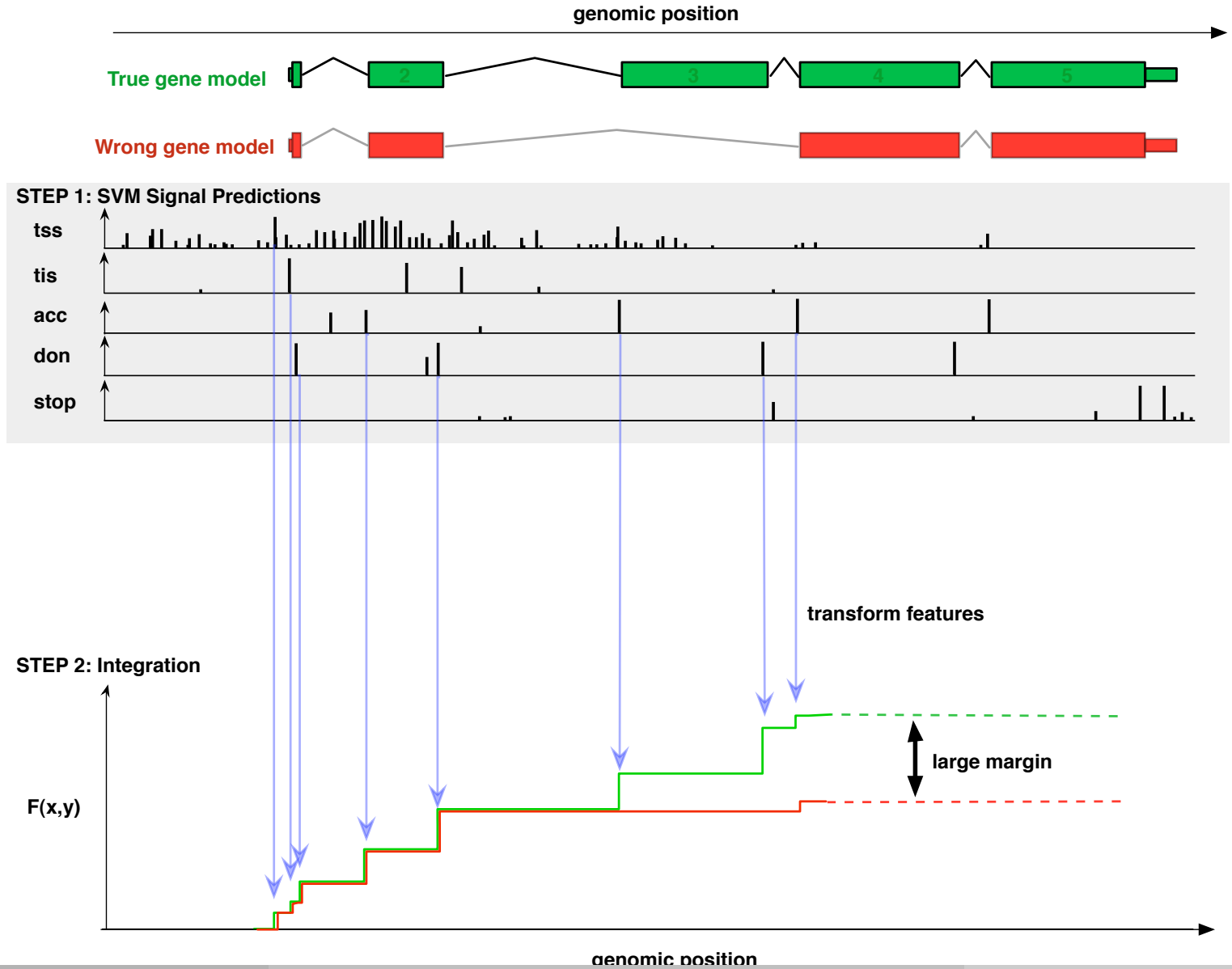
Learning to Integrate Data Sources

[Behr et al., 2008]



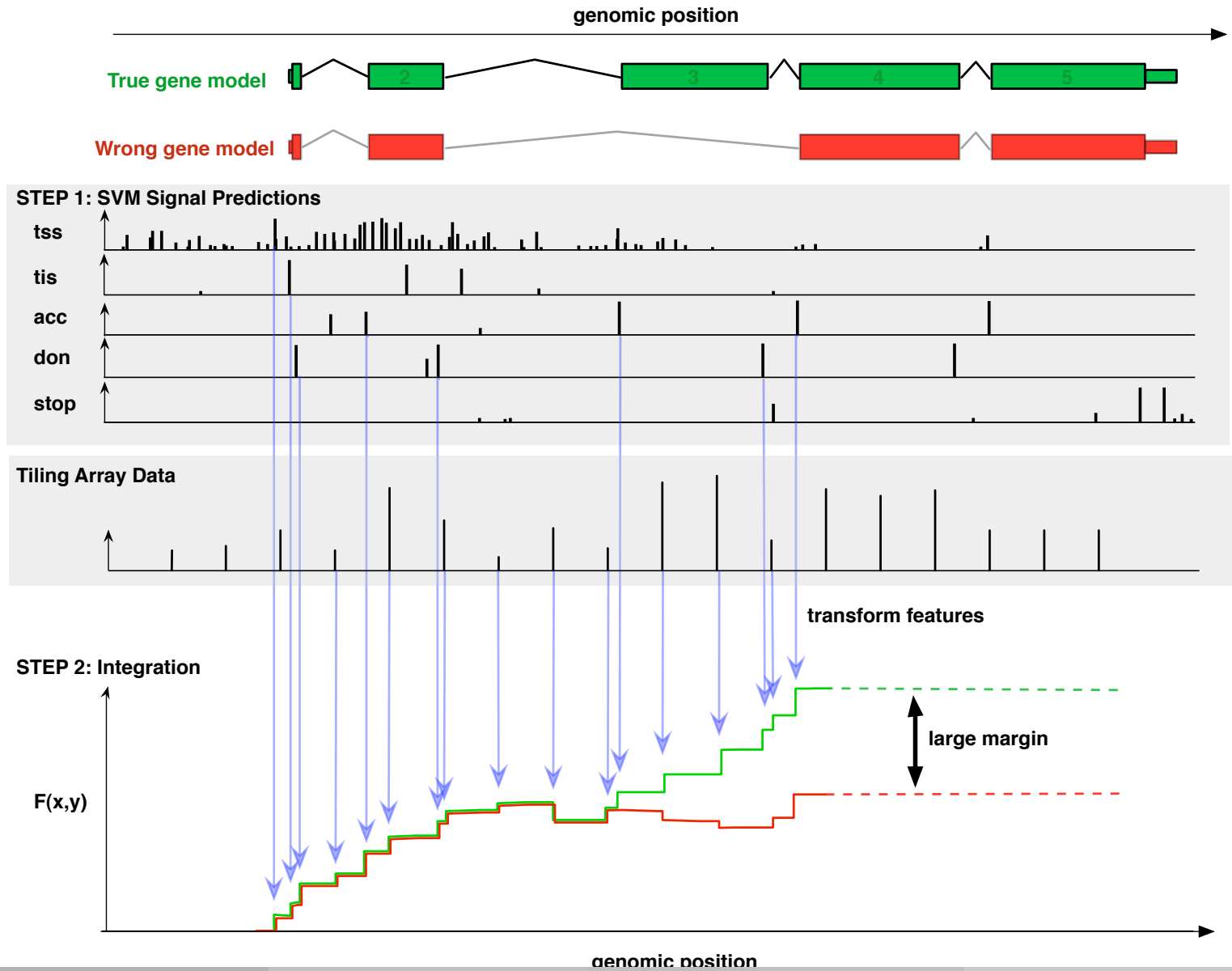
Learning to Integrate Data Sources

[Behr et al., 2008]



Learning to Integrate Data Sources

[Behr et al., 2008]



Results: How much does the data help?

[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of combination of data
- 300 regions around cDNA confirmed genes for evaluation

Results: How much does the data help?

[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of combination of data
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
Tiling arrays (young leaves)	78.0	82.9	**80.4**
Tiling arrays (inflorescence)	77.1	81.6	**79.3**
Tiling arrays (root)	76.2	81.5	**78.9**
Tiling arrays (combined)	77.4	81.4	**79.4**
RNA-seq (w/o spliced reads)	76.8	80.8	**78.7**
RNA-seq (with spliced reads)	79.6	82.1	**80.8**

Results: How much does the data help?

[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of combination of data
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
Tiling arrays (young leaves)	78.0	82.9	**80.4**
Tiling arrays (inflorescence)	77.1	81.6	**79.3**
Tiling arrays (root)	76.2	81.5	**78.9**
Tiling arrays (combined)	77.4	81.4	**79.4**
RNA-seq (w/o spliced reads)	76.8	80.8	**78.7**
RNA-seq (with spliced reads)	79.6	82.1	**80.8**

Results: How much does the data help?

[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of combination of data
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
Tiling arrays (young leaves)	78.0	82.9	**80.4**
Tiling arrays (inflorescence)	77.1	81.6	**79.3**
Tiling arrays (root)	76.2	81.5	**78.9**
Tiling arrays (combined)	77.4	81.4	**79.4**
RNA-seq (w/o spliced reads)	76.8	80.8	**78.7**
RNA-seq (with spliced reads)	79.6	82.1	**80.8**

Results: How much data is needed?

[Behr et al., 2008]

Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of combination of data
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
RNA-seq 1/128	75.5	78.7	**77.1**
RNA-seq 1/64	76.8	79.5	**78.1**
RNA-seq 1/32	76.2	79.6	**77.9**
RNA-seq 1/16	77.1	79.1	**78.1**
RNA-seq 1/8	78.6	80.1	**79.4**
RNA-seq 1/4	77.4	79.6	**78.5**
RNA-seq 1/2	79.3	82.1	**80.6**
RNA-seq	79.6	82.1	**80.8**

Results: How much data is needed?

[Behr et al., 2008]

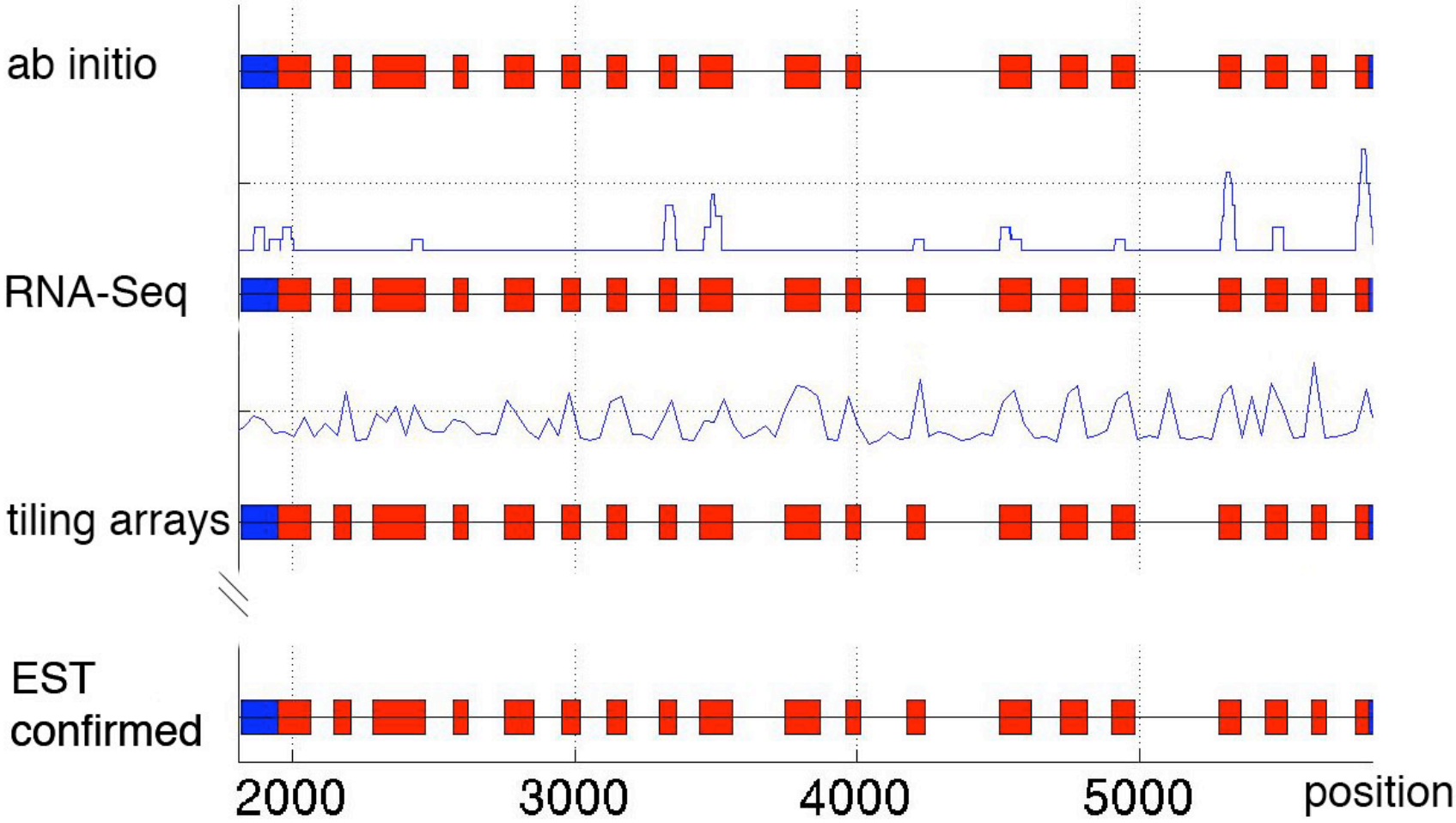
Experimental setup (*Arabidopsis thaliana*):

- 60% of known genes for training signals in step 1
- 400 genes for training of combination of data
- 300 regions around cDNA confirmed genes for evaluation

	transcript level [%]		
	SN	SP	F
Ab initio	73.7	78.6	**76.0**
RNA-seq 1/128	75.5	78.7	**77.1**
RNA-seq 1/64	76.8	79.5	**78.1**
RNA-seq 1/32	76.2	79.6	**77.9**
RNA-seq 1/16	77.1	79.1	**78.1**
RNA-seq 1/8	78.6	80.1	**79.4**
RNA-seq 1/4	77.4	79.6	**78.5**
RNA-seq 1/2	79.3	82.1	**80.6**
RNA-seq	79.6	82.1	**80.8**

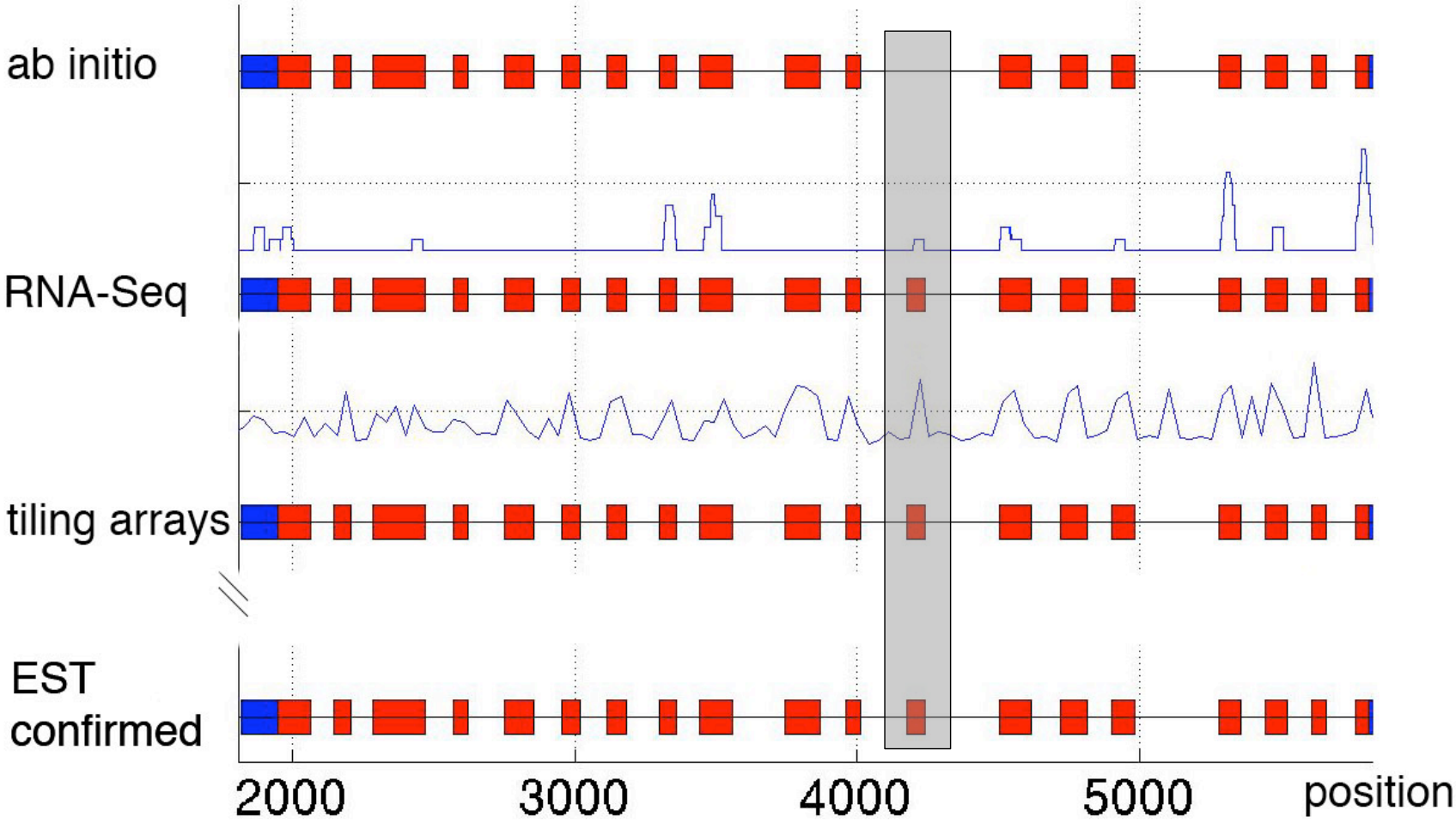
Results: Combining helps!

[Behr et al., 2008]



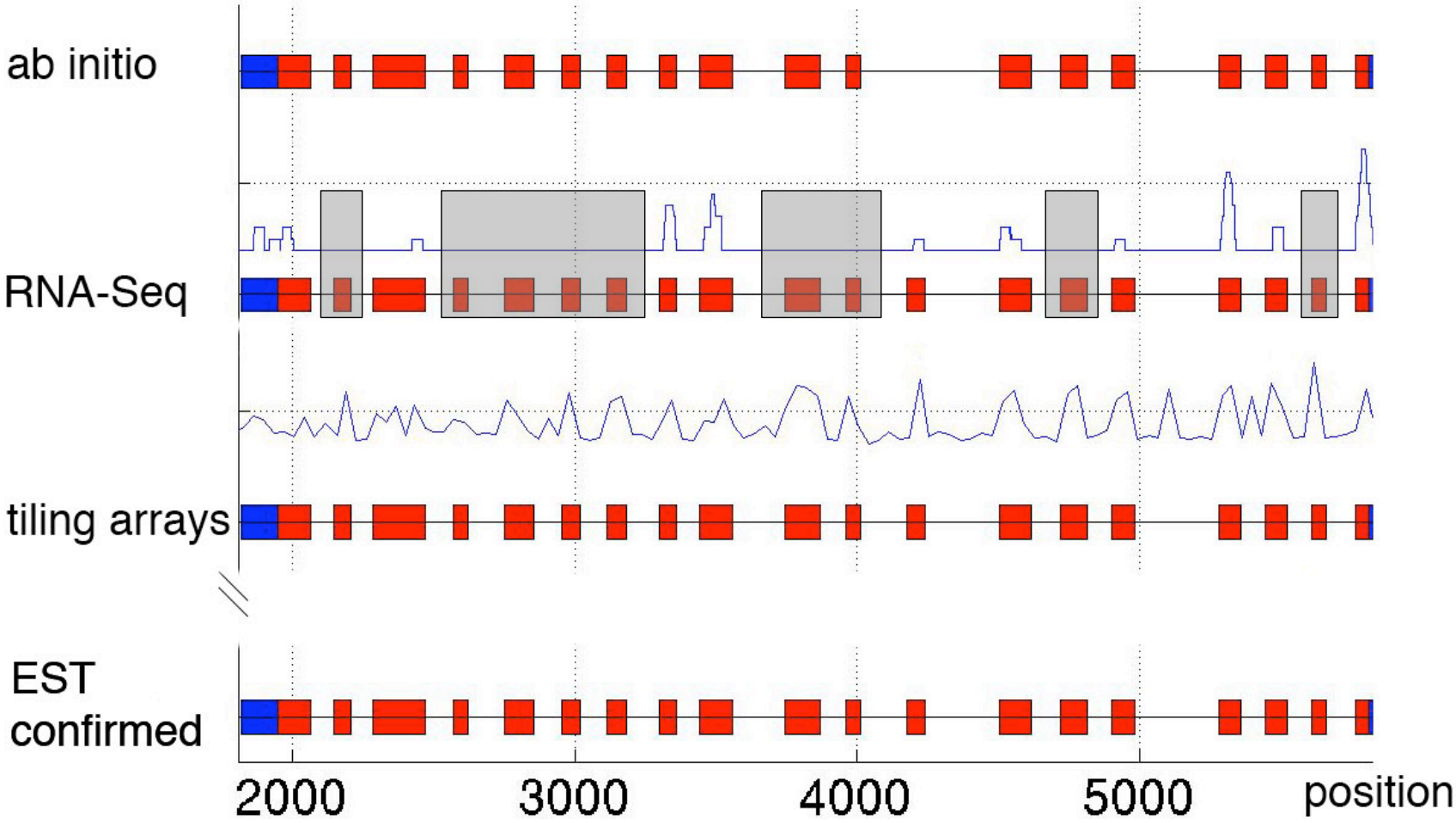
Results: Combining helps!

[Behr et al., 2008]



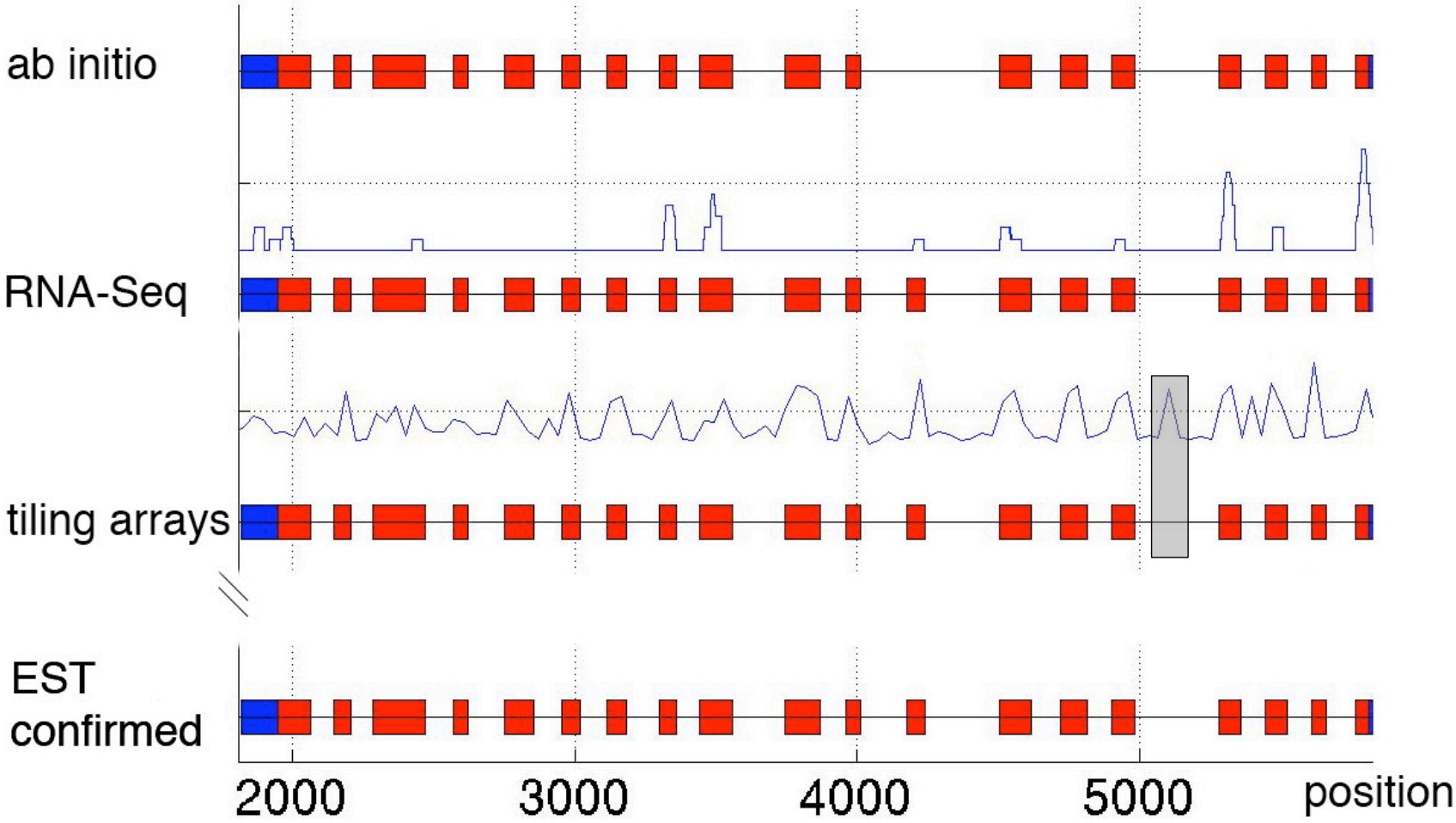
Results: Combining helps!

[Behr et al., 2008]



Results: Combining helps!

[Behr et al., 2008]



Quantification of Transcripts

Given: Accurate short reads alignments
We can use exon/intron read coverage to:

1. Improve gene finder predictions?
2. Predict transcript abundances?

First Step: Given a set of known transcripts, we predict transcript abundances by solving a **linear programming** problem:

- **Optimizes the weights** for each transcript
- Exploits additive nature of the read coverage
- Minimizing the residual error

Quantification of Transcripts

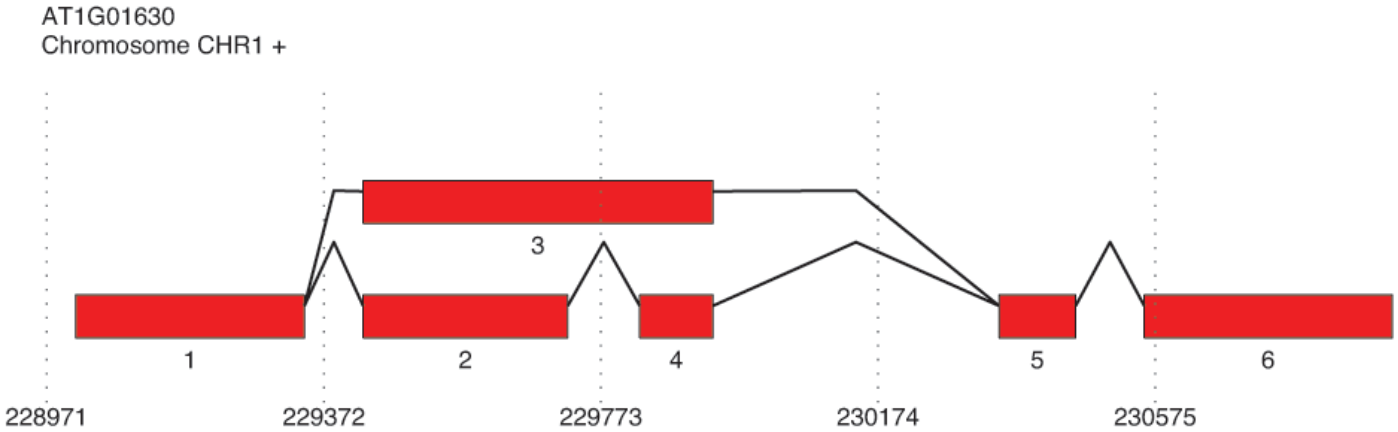
Given: Accurate short reads alignments
We can use exon/intron read coverage to:

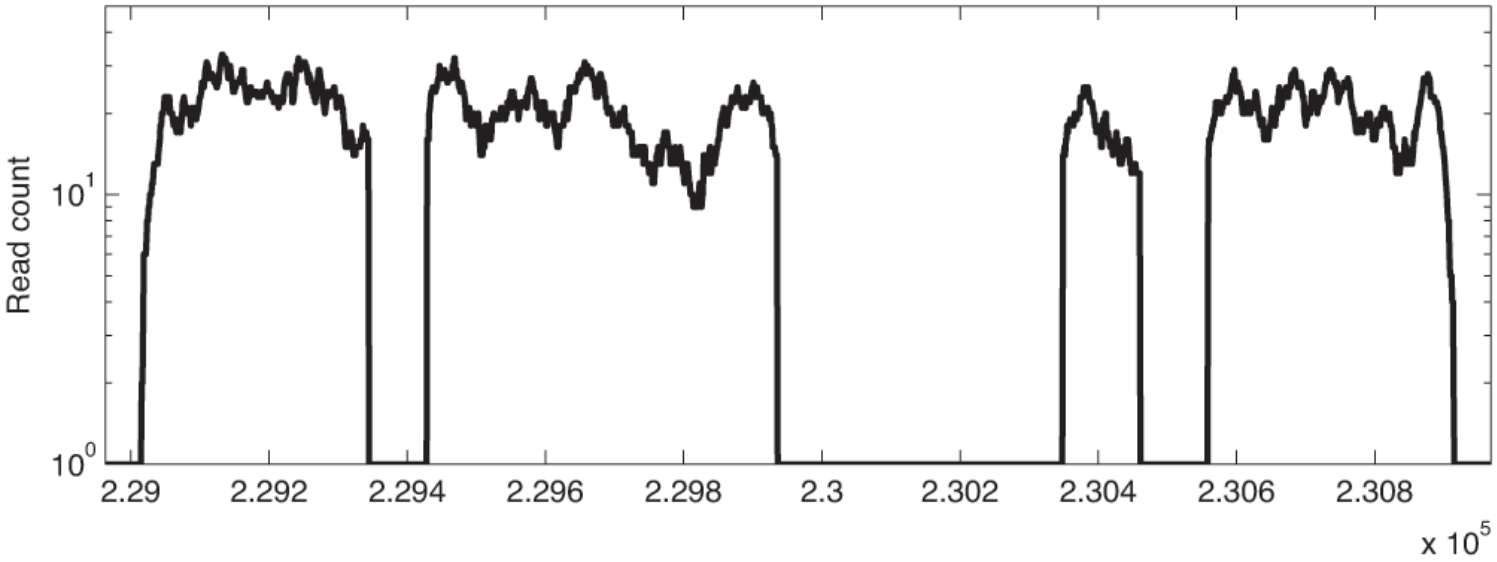
1. Improve gene finder predictions?
2. Predict transcript abundances?

First Step: Given a set of known transcripts, we predict transcript abundances by solving a **linear programming** problem:

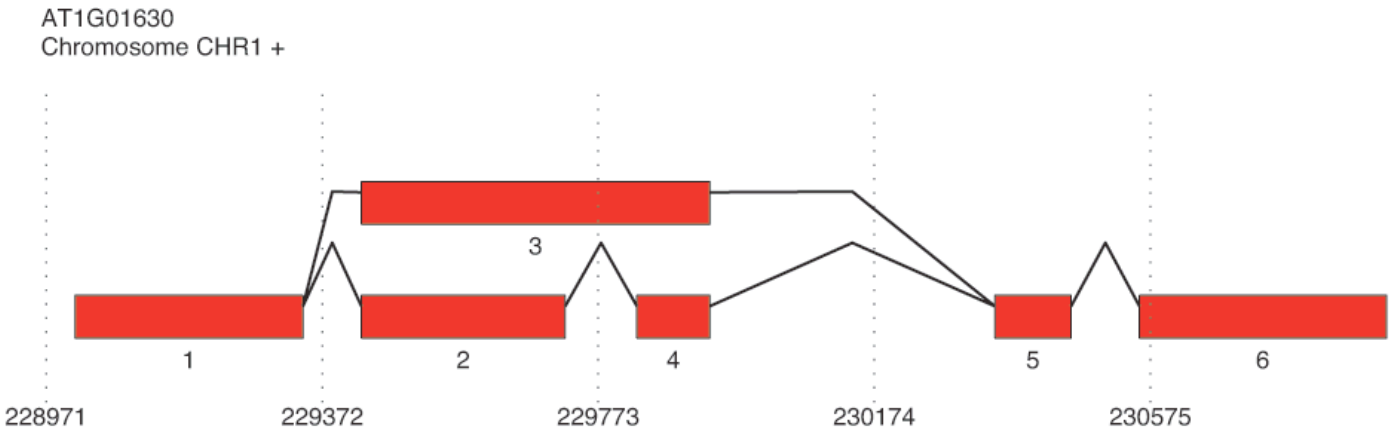
- **Optimizes the weights** for each transcript
- Exploits additive nature of the read coverage
- Minimizing the residual error

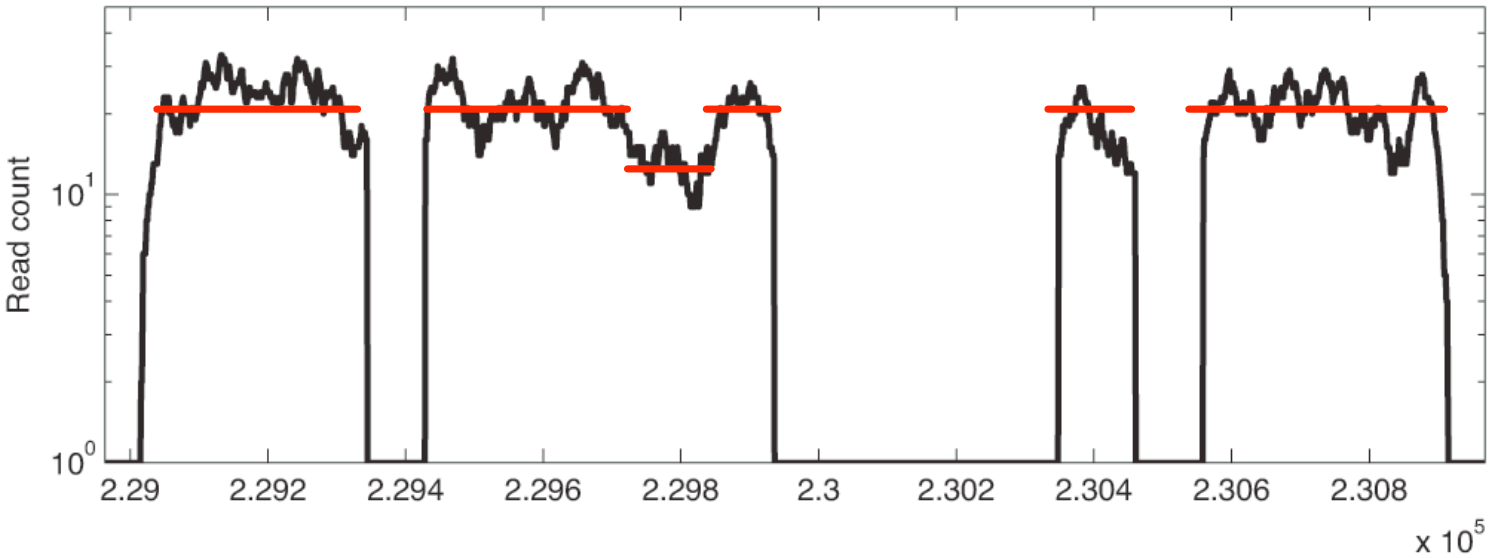
Quantification of Transcripts
(Preliminary)





Quantification of Transcripts
(Preliminary)





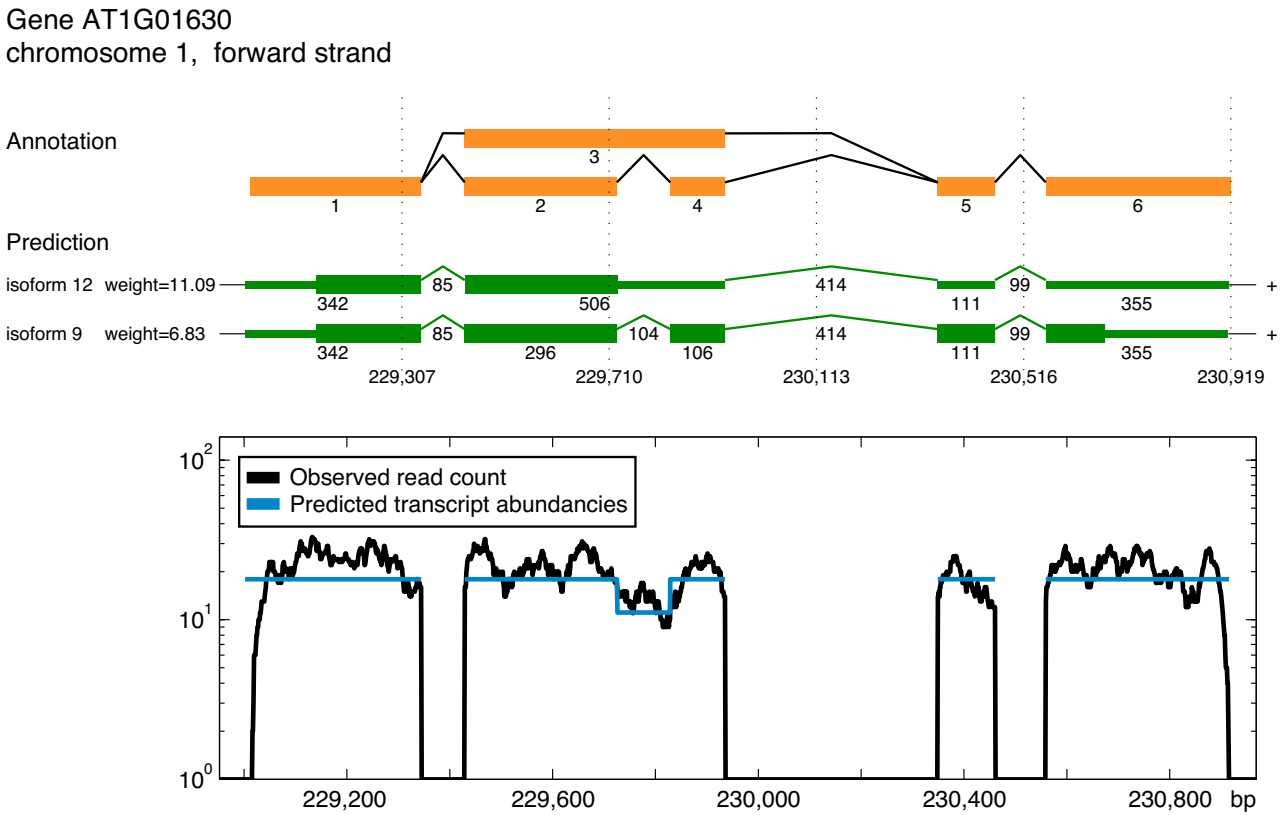
De Novo Transcript Quantification
(Preliminary)

Combination of gene finding and transcript quantification:

Detect alternative transcripts including their abundance *without* relying on a genome annotation.

Example for A. thaliana.

The two isoforms were correctly determined (upper panel) and the transcript abundances are estimated well (lower panel).



Roadmap

1. Computational Gene Finding
   - Identification of Genomic Signals
   - Learning to Predict mRNA Transcripts
2. Whole-genome Tiling Arrays
   - Technology and Limitations
   - Identification of Expression Differences
   - De Novo Transcript Discovery
3. Next-generation Sequencing
   - Technology & Limitations
   - Assembly & Read Mapping
4. Extensions
   - Quantification of Transcripts
   - ChIP-on-Chip Studies

ChIP-on-chip and ChIP-seq

- ChIP = Chromatin Immunoprecipitation
- Established technique, now used for genome-wide screens on a chip (ChIP-on-chip) or through Next-Generation Sequencing (ChIP-seq)
- Analyze binding of a single transcription factor (TF)
- **Goal:** Identify parts of the chromatin that this TF binds to

RNA Immunoprecipitation to understand RNA processing.

ChIP-on-chip and ChIP-seq

- ChIP = Chromatin Immunoprecipitation
- Established technique, now used for genome-wide screens on a chip (ChIP-on-chip) or through Next-Generation Sequencing (ChIP-seq)
- Analyze binding of a single transcription factor (TF)
- **Goal:** Identify parts of the chromatin that this TF binds to

RNA Immunoprecipitation to understand RNA processing.

ChIP Protocol: Preparations

ChIP: Chromatin Immunoprecipitation

Create antibody against a certain TF:

- Identify the TF-coding gene
- Transfer gene sequence to a cloning vector
- Get cells (*E. coli*, yeast, ...) to express the protein
- Extract (correctly folded) protein from cells, purify, then purify again
- Inject in animal, extract, purify, ...
  $\Rightarrow$ Obtain (poly-clonal) antibody

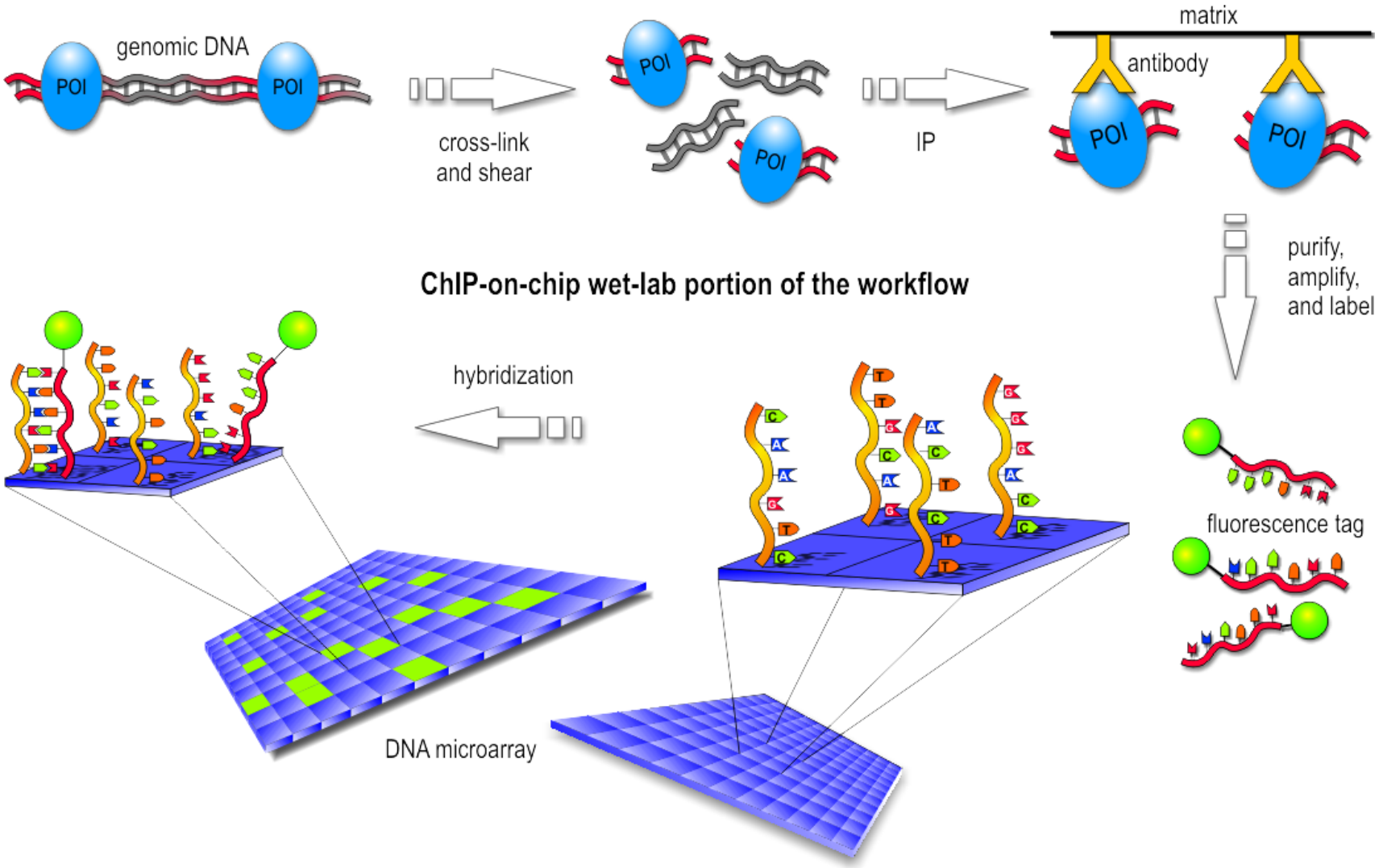
ChIP Protocol: Preparations

ChIP: Chromatin Immunoprecipitation

Create antibody against a certain TF:

- Identify the TF-coding gene
- Transfer gene sequence to a cloning vector
- Get cells (*E. coli*, yeast, ...) to express the protein
- Extract (correctly folded) protein from cells, purify, then purify again
- Inject in animal, extract, purify, ...
  $\Rightarrow$ Obtain (poly-clonal) antibody

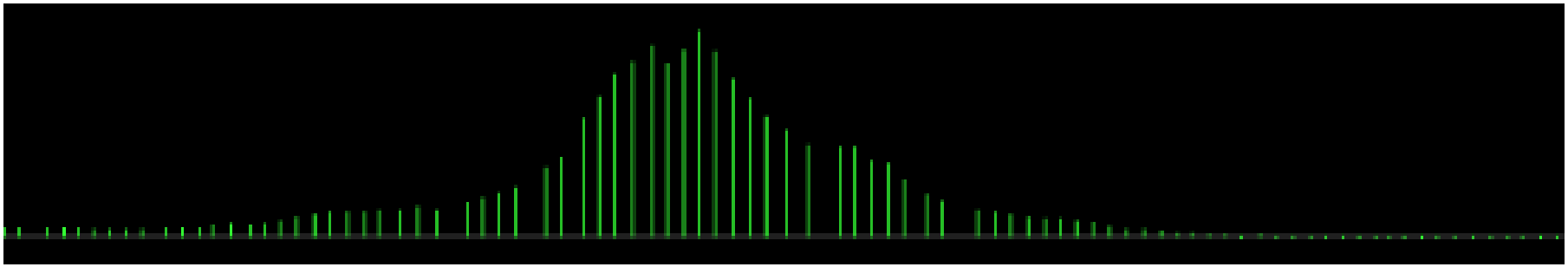
ChIP Protocol: Overview



[Wikipedia]

ChIP-on-chip: Detection

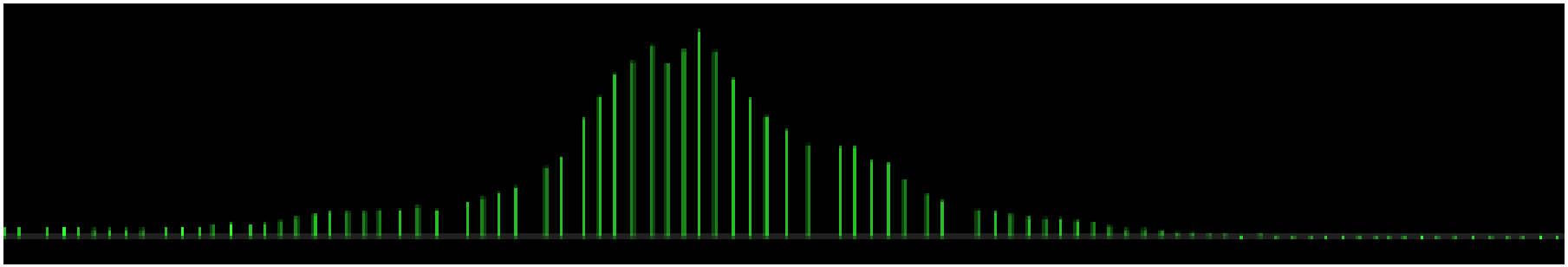
- Tiling arrays or Sequencing analysis pipelines can be used
  $\Rightarrow$ *Similar problems as in transcriptome analysis*
- Compare with control experiment, e.g. LOF of the TF, calculate p-values of binding probability using a smoothing window
- ChIP arrays often analyzed with *specialized methods*:
  Model-based Analysis of Tiling-array (MAT), TileMap; TiMAT
- ML approaches: Learn expected distribution from regulatory region, where binding peak is to be expected
  (often difficult as labeled data is scarse)



[Provided by Sebastian Schultheiss]

ChIP-on-chip: Detection

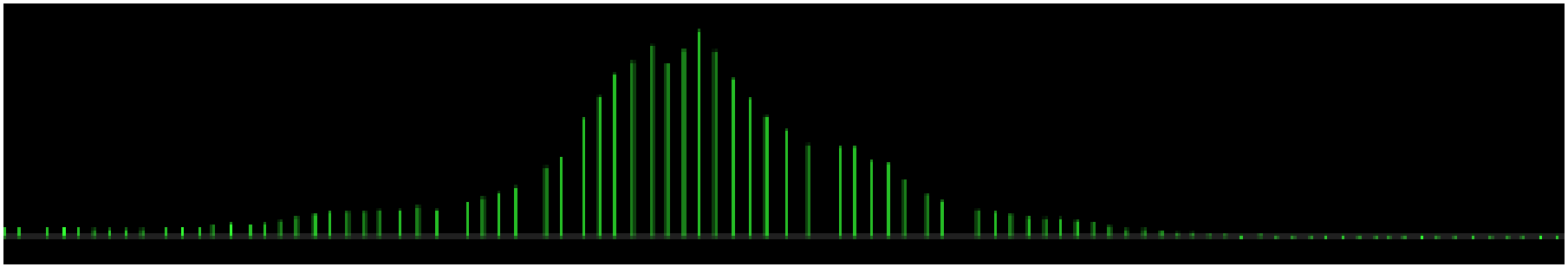
- Tiling arrays or Sequencing analysis pipelines can be used
  $\Rightarrow$ *Similar problems as in transcriptome analysis*
- Compare with control experiment, e.g. LOF of the TF, calculate p-values of binding probability using a smoothing window
- ChIP arrays often analyzed with *specialized methods*:
  Model-based Analysis of Tiling-array (MAT), TileMap; TiMAT
- ML approaches: Learn expected distribution from regulatory region, where binding peak is to be expected
  (often difficult as labeled data is scarse)



[Provided by Sebastian Schultheiss]

ChIP-on-chip: Detection

- Tiling arrays or Sequencing analysis pipelines can be used
  $\Rightarrow$ *Similar problems as in transcriptome analysis*
- Compare with control experiment, e.g. LOF of the TF, calculate p-values of binding probability using a smoothing window
- ChIP arrays often analyzed with *specialized methods*:
  Model-based Analysis of Tiling-array (MAT), TileMap; TiMAT
- ML approaches: Learn expected distribution from regulatory region, where binding peak is to be expected
  (often difficult as labeled data is scarse)



[Provided by Sebastian Schultheiss]

ChIP-on-chip: MAT

- **MAT**: Model-based analysis of tiling arrays for ChIP-chip
- **Idea:** Majority of signal is due to **non-specific binding**, there are strong **probe sequence effects**
  $\Rightarrow$ Formulate an **array probe affinity** model

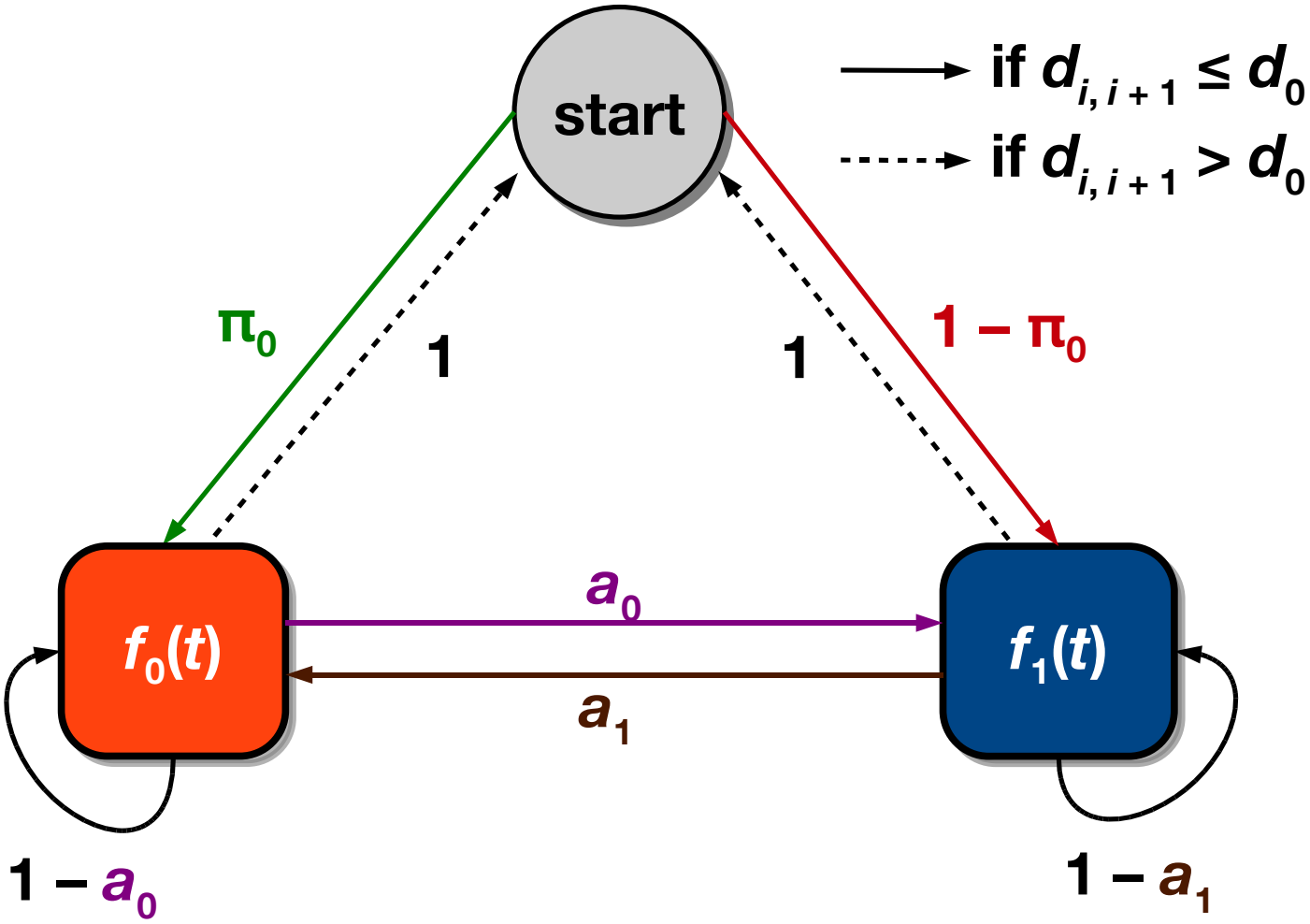
[Johnson et al., 2006]

ChIP-on-chip: MAT

Algorithm:

- Divide probes into **affinity bins**
- Sample probe signal **variance** for every probe per bin (more numbers makes this more stable than just comparing replicates)
- Calculate a **t-value** $t(k)$ per probe k
- Compute **trimmed mean TM** in sliding window
- The **central probe** of the window from probes i to j will be assigned a $\text{MATscore}_{i,j} = \sqrt{j-i} \cdot \text{TM}_{k=i}^{j} t(k)$
- Define a MATscore **threshold** above which a probe is classified as enriched, subtract control experiments if available
- Threshold can be found by MAT using a **p-value** from a non-enriched null sample or a user-supplied **FDR**
- MAT **merges** all enriched regions within 300 bp and assigns them the highest MATscore of the region

ChIP-on-chip: TileMap HMM



[Ji and Wong, 2005b]

ChIP-on-chip: TileMap

- Algorithm:
  - Compute **test-statistic** for each normalized, log-transformed probe X_{ijk}, which is the hybridization intensity for probe *i* under condition *j* in replicate *k*
  - $X_{ijk}|\mu_i^2 \sim N(\mu_{ij}, \sigma_i^2)$ estimate every σ_i^2 to approximate posterior distribution of μ_{ij}
  - Use a formula akin to a t-statistic, which uses not only information for probe i to estimate standard deviation but pools information from **all probes** for higher sensitivity
  - Combine information from **neighboring probes** (moving average/sliding window or an HMM)

[Ji and Wong, 2005b]

ChIP-on-chip: What do we learn?



ChIP-on-chip experiment on known TF (plant stem cell regulator)

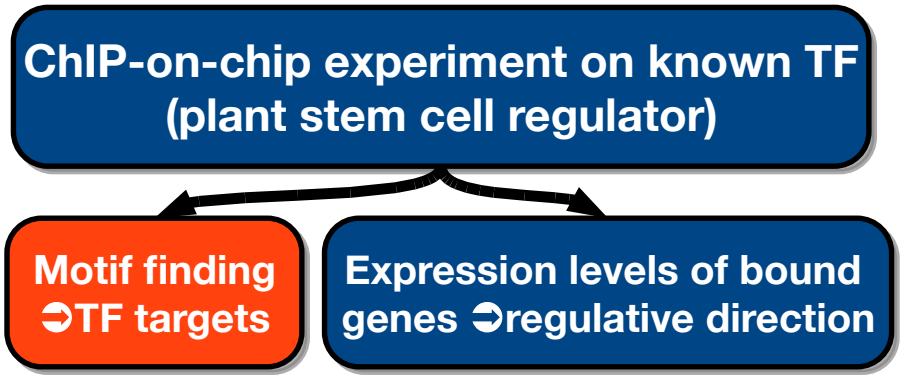
WUSCHEL

Wetlab

Key: Transcription Factor

Models should be based on transcripts not genes!

ChIP-on-chip: What do we learn?



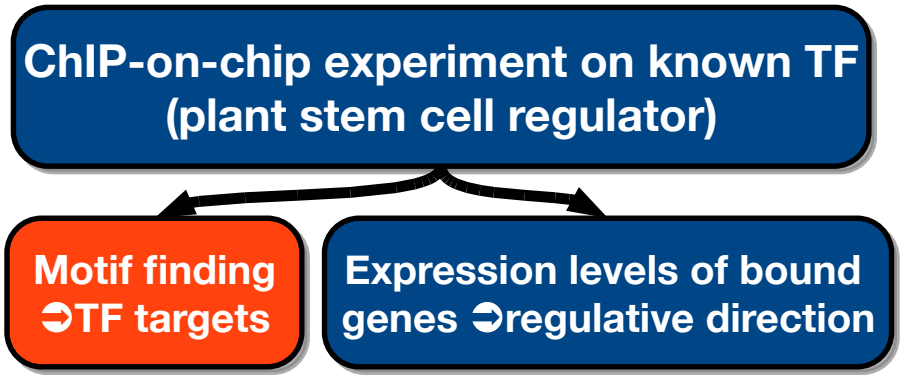
WUSCHEL

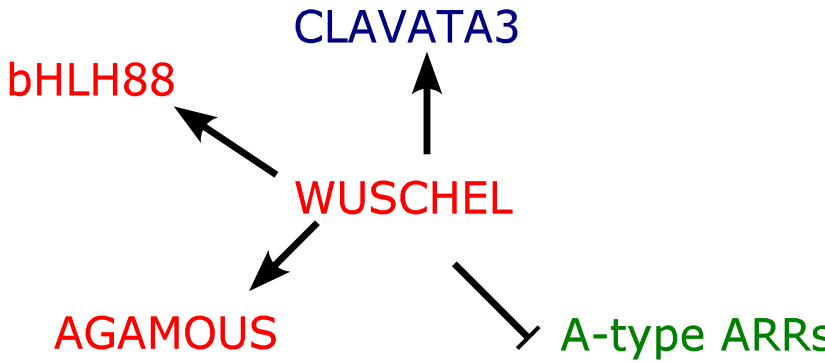
Key: Transcription Factor

in silico Wetlab

Models should be based on transcripts not genes!

ChIP-on-chip: What do we learn?



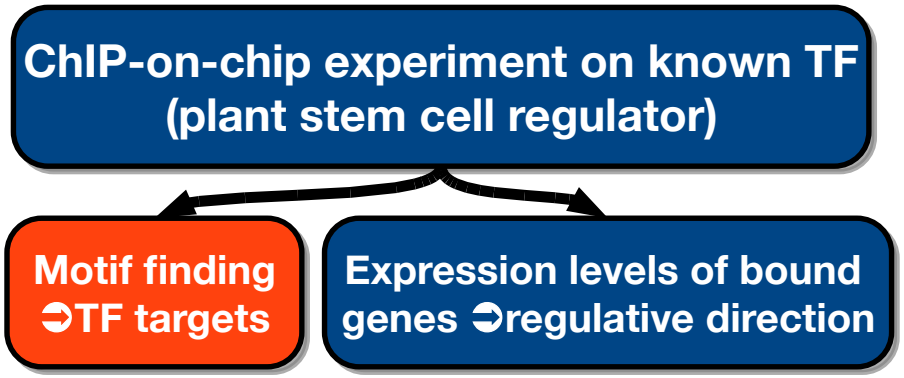


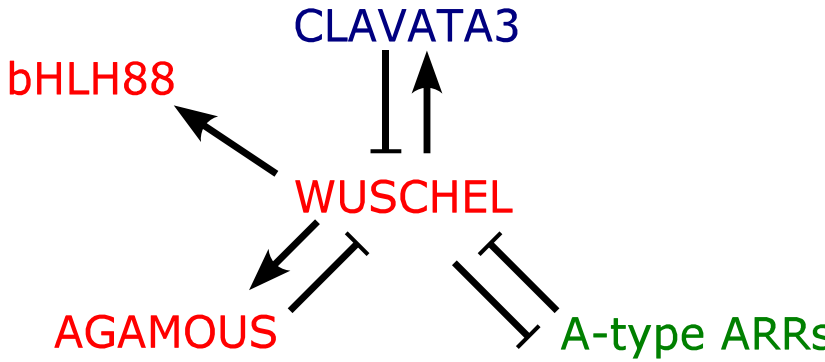
in silico Wetlab

Key:
activates
represses
Transcription Factor
Response Regulator
Peptide, Ligand

Models should be based on transcripts not genes!

ChIP-on-chip: What do we learn?



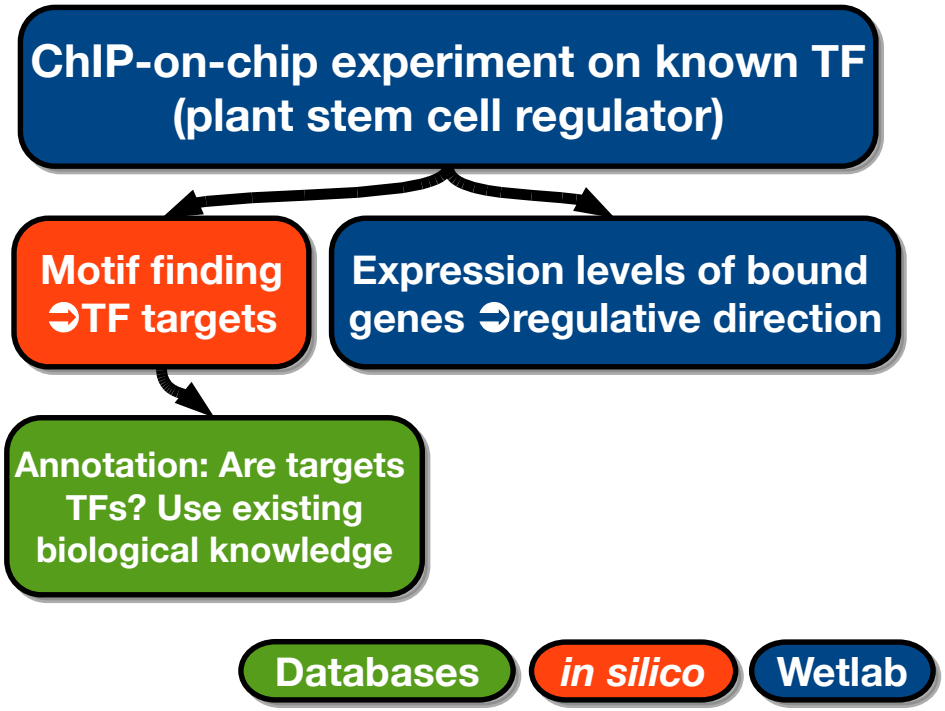


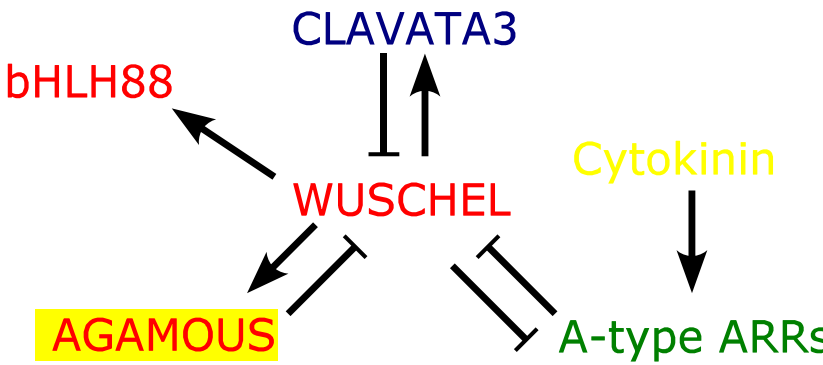
in silico Wetlab

Key:
Transcription Factor
Response Regulator
activates
Peptide, Ligand
represses
Hormone

Models should be based on transcripts not genes!

ChIP-on-chip: What do we learn?





Key:

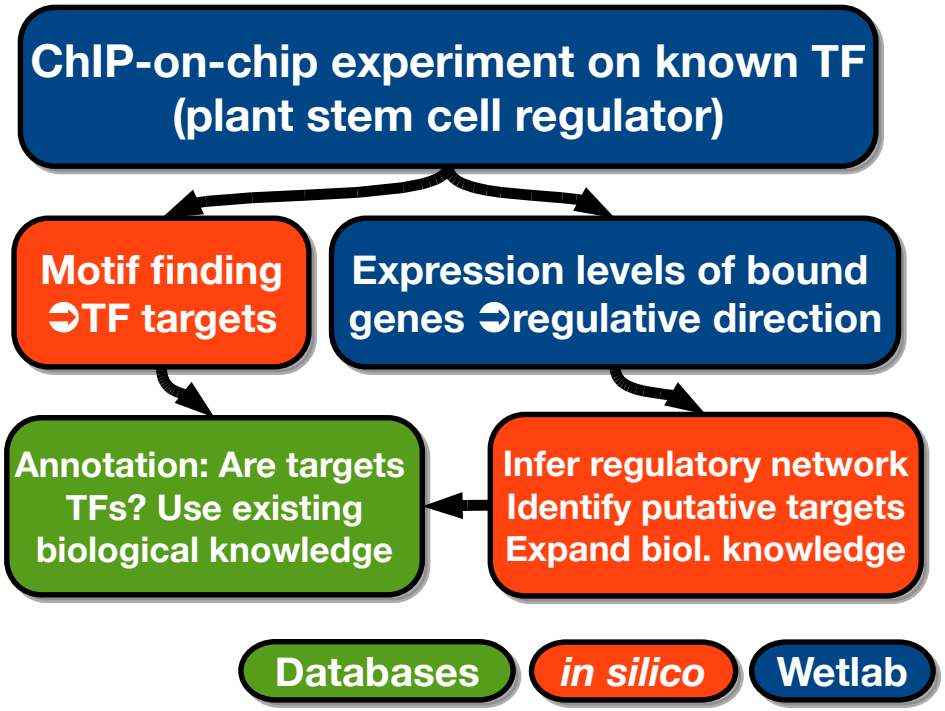
activates →

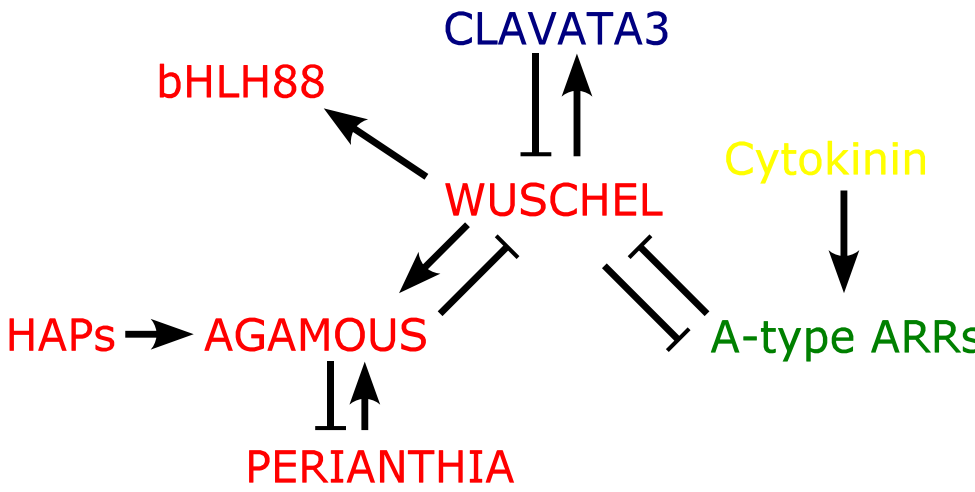
represses ⊣

Transcription Factor
Response Regulator
Peptide, Ligand
Hormone

Models should be based on transcripts not genes!

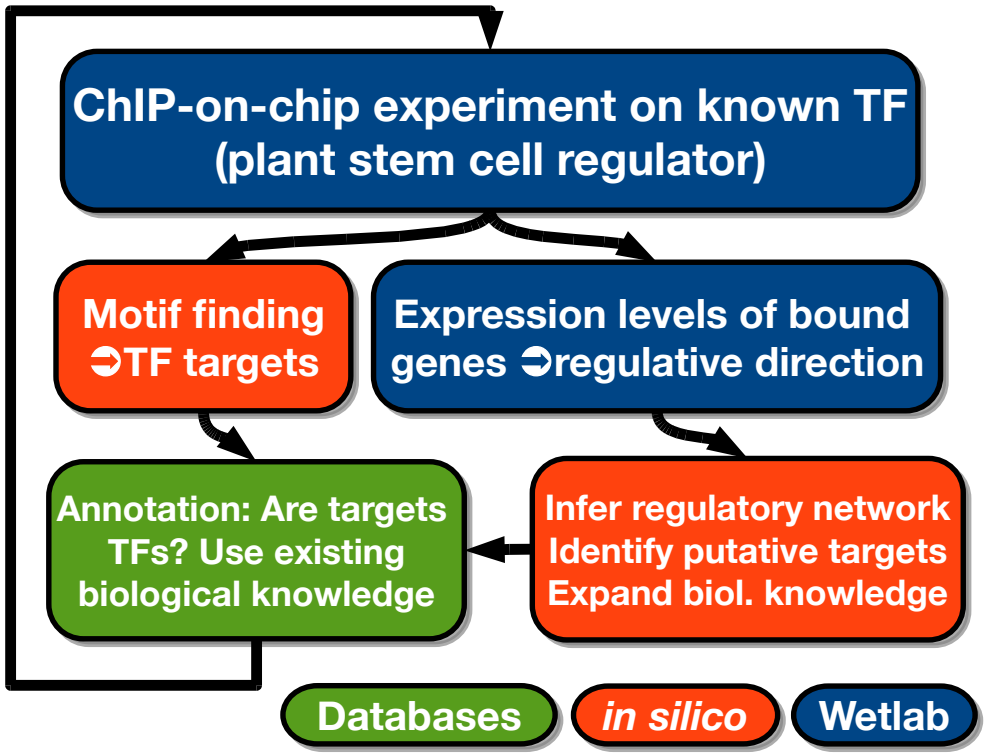
ChIP-on-chip: What do we learn?

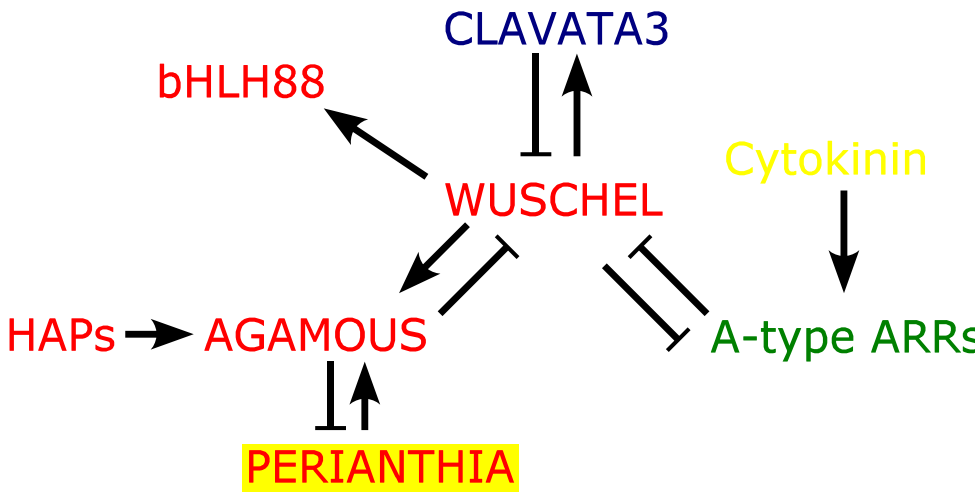




Models should be based on transcripts not genes!

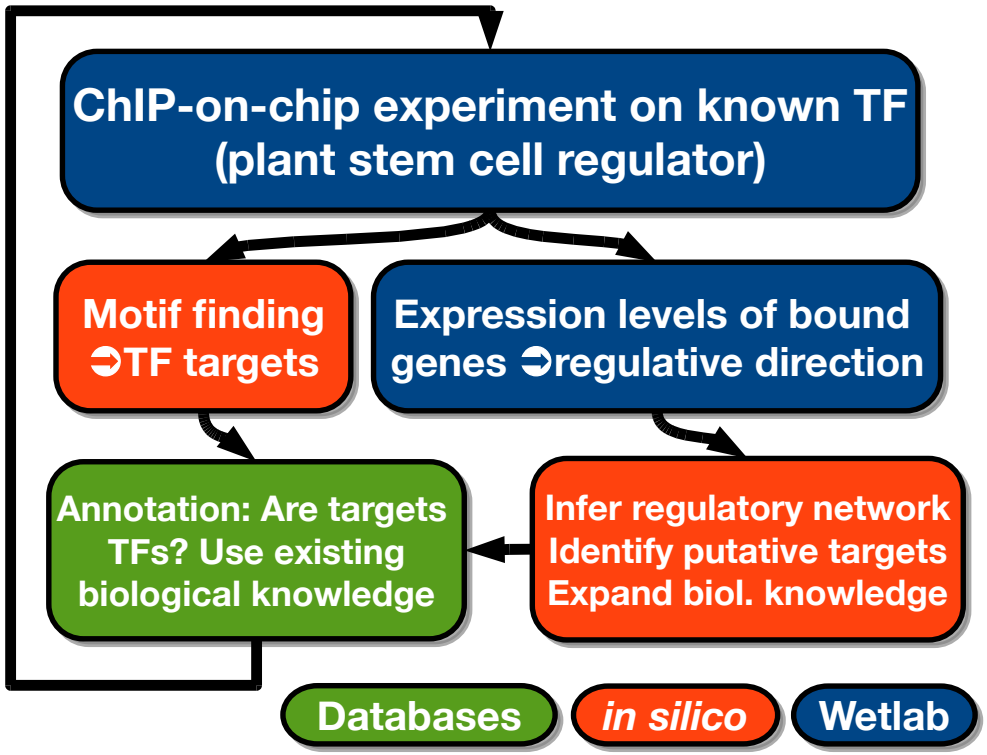
ChIP-on-chip: What do we learn?

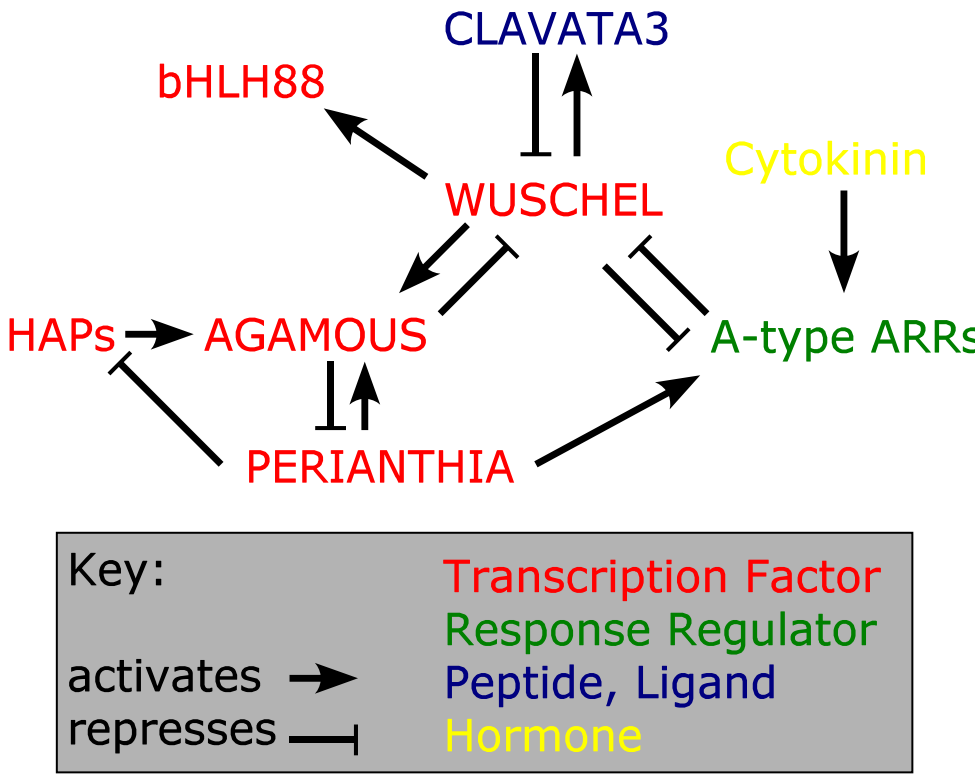




Models should be based on transcripts not genes!

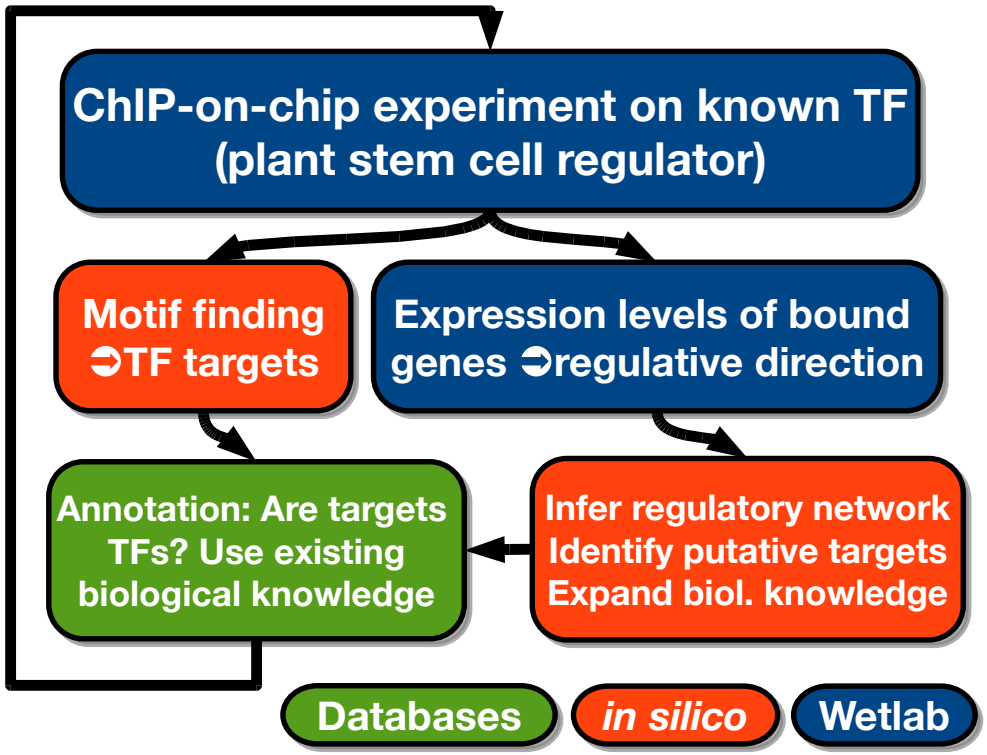
ChIP-on-chip: What do we learn?

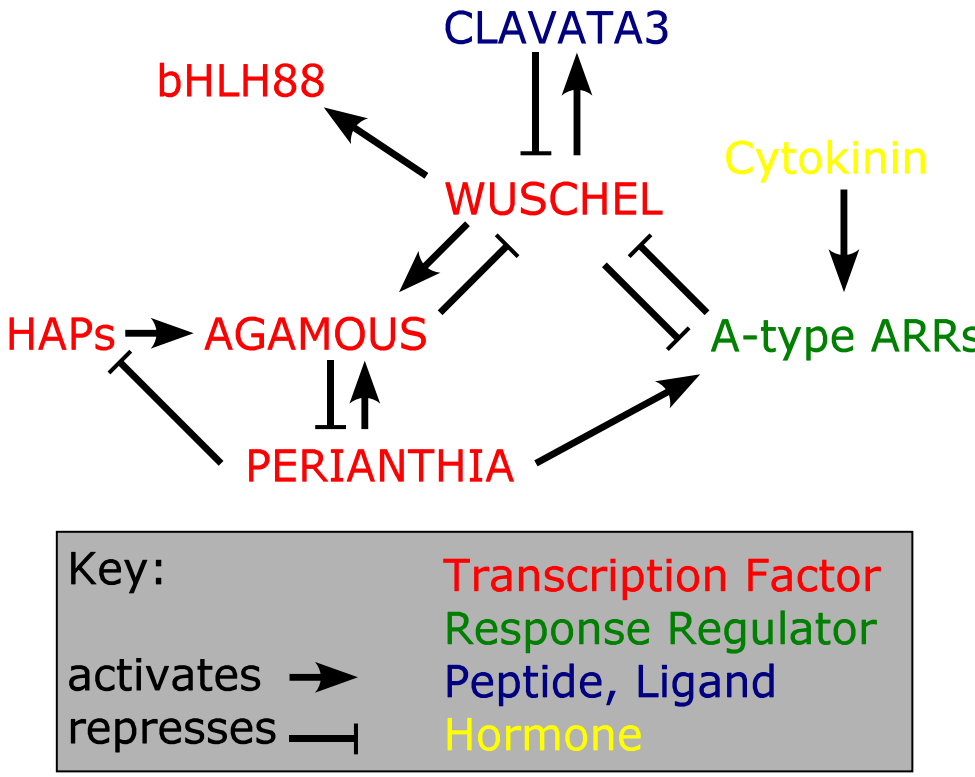




Models should be based on transcripts not genes!

ChIP-on-chip: What do we learn?





Models should be based on transcripts not genes!

Summary & Conclusions

- Methods for characterizing transcriptomes
  - in different organisms
  - under different conditions (development/environment)
- Gene finding methods
  - Improved accuracy due to novel inference methods
  - 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 gene finding
  - 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
  - Improved accuracy due to novel inference methods
  - 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 gene finding
  - 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
  - Improved accuracy due to novel inference methods
  - 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 gene finding
  - 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
  - Improved accuracy due to novel inference methods
  - 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 gene finding
  - Condition specific transcriptome predictions
  - Help to uncover the full complexity of transcriptomes

Acknowledgments

Gene Finding

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

Tiling Arrays

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

Short Read Analysis

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

CHiP-on-Chip

- **Sebastian Schultheiss** (FML & Uni HD)
- Jan Lohmann (Uni HD)

More Information

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

Acknowledgments

Gene Finding

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

Tiling Arrays

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

Short Read Analysis

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

CHiP-on-Chip

- **Sebastian Schultheiss** (FML & Uni HD)
- Jan Lohmann (Uni HD)

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)
- Alex Zien (FML & FIRST)
- Georg Zeller (FML & MPI)

Tiling Arrays

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

Short Read Analysis

- **Fabio De Bona** (FML)
- Stephan Ossowski (MPI)
- Korbinian Schneeberger (MPI)

CHiP-on-Chip

- **Sebastian Schultheiss** (FML & Uni HD)
- Jan Lohmann (Uni HD)

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.

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.

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.

References II

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.

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

J. Eichner. Analysis of alternative transcripts in arabidopsis thaliana with whole genome arrays. Master's thesis, University of Tübingen, Sand 13, 72076 Tübingen, Germany, June 2008.

J. Eichner, G. Zeller, S. Laubinger, D. Weigel, and G. Rätsch. Analysis of alternative transcripts in arabidopsis thaliana with whole genome arrays. forthcoming, March 2009.

References III

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.

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

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 2003.

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

Hongkai Ji and Wing Hung Wong. Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–3636, Sep 2005b. ISSN 1367-4803 (Print). doi: 10.1093/bioinformatics/bti593.

References IV

W Evan Johnson, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown, and X Shirley Liu. Model-based analysis of tiling-arrays for chip-chip. *Proc Natl Acad Sci U S A*, 103(33):12457–12462, Aug 2006. ISSN 0027-8424 (Print). doi: 10.1073/pnas.0601180103.

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.

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.

S. Ossowski, K. Schneeberger, R. Clark, C. Lanz, N. Warthmann, and D. Weigel. Sequencing of natural strains of arabidopsis thaliana with short reads. *Genome Research*, 18(2024–2033), 2008.

References V

E Purdom, K M Simpson, M D Robinson, J G Conboy, A V Lapuk, and T P Speed. Firma: a method for detection of alternative splicing from exon array data. *Bioinformatics*, 24(15): 1707–14, Aug 2008. doi: 10.1093/bioinformatics/btn284. URL http://bioinformatics.oxfordjournals.org/cgi/content/full/24/15/1707.

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.

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.

References VI

G. Schweikert, G. Zeller, A. Zien, J. Behr, C.S. Ong, P. Philips, A. Bohlen, R. Bohnert, F. De Bona, S. Sonnenburg, and G. Rätsch. mGene: Accurate computational gene finding with application to nematode genomes. under revision for Genome Research, March 2009.

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.

Charles W Sugnet, Karpagam Srinivasan, Tyson A Clark, Georgeann O'brien, Melissa S Cline, Hui Wang, Alan Williams, David Kulp, John E Blume, David Haussler, and Manuel Ares. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol*, 2(1):e4, Jan 2006. doi: 10.1371/journal.pcbi.0020004.

References VII

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 VIII

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.