QPalma - Optimal Spliced Alignments of Short Sequence Reads

Fabio De Bona[1]
Stephan Ossowski[2]
Korbinian Schneeberger[2]
Gunnar Rätsch[1]

(1) Friedrich Miescher Laboratory of the Max Planck Society
(2) Max Planck Institute for Developmental Biology
Tübingen, Germany

September 25, 2008

Introduction

- Next Generation (NG) sequencing technologies produce huge amounts of sequencing data
- Difference to Sanger sequencing:
  - Much cheaper and faster
  - Much more and shorter fragments $\Rightarrow$ "reads"
  - Much more errors

Introduction

- Next Generation (NG) sequencing technologies produce huge amounts of sequencing data
- Difference to Sanger sequencing:
  - Much cheaper and faster
  - Much more and shorter fragments $\Rightarrow$ "reads"
  - Much more errors
- Genome / transcriptome sequencing
- Data can be used via local alignments:
  - Discovery of new genes,
  - Identification of alternative splice forms, ...

Transcriptome Analysis with NG Sequencing



QPalma's aim: Accurately Align transcriptome reads to genomic sequences

- Genomic read mapping already challenging
- Transcriptome read mapping is even more difficult:
  - Spliced reads stem from several genomic regions
  - Short reads getting even shorter due to splicing

  $\Rightarrow$ Alignments more error prone

Transcriptome Analysis with NG Sequencing



QPalma's aim: Accurately Align transcriptome reads to genomic sequences

- Genomic read mapping already challenging
- Transcriptome read mapping is even more difficult:
  - Spliced reads stem from several genomic regions
  - Short reads getting even shorter due to splicing

  $\Rightarrow$ Alignments more error prone
- Improve alignments by using more information:
  - Accurate splice site models
  - Intron length model
  - Quality scores model

Transcriptome Analysis with NG Sequencing



QPalma's aim: Accurately Align transcriptome reads to genomic sequences

- Genomic read mapping already challenging
- Transcriptome read mapping is even more difficult:
  - Spliced reads stem from several genomic regions
  - Short reads getting even shorter due to splicing

  $\Rightarrow$ Alignments more error prone
- Improve alignments by using more information:
  - Accurate splice site models
  - Intron length model
  - Quality scores model

Idea: Use a **machine learning** method to infer optimal parameters

Spliced vs. Unspliced Alignments



...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC

- Find matching region on genome with a few mismatches

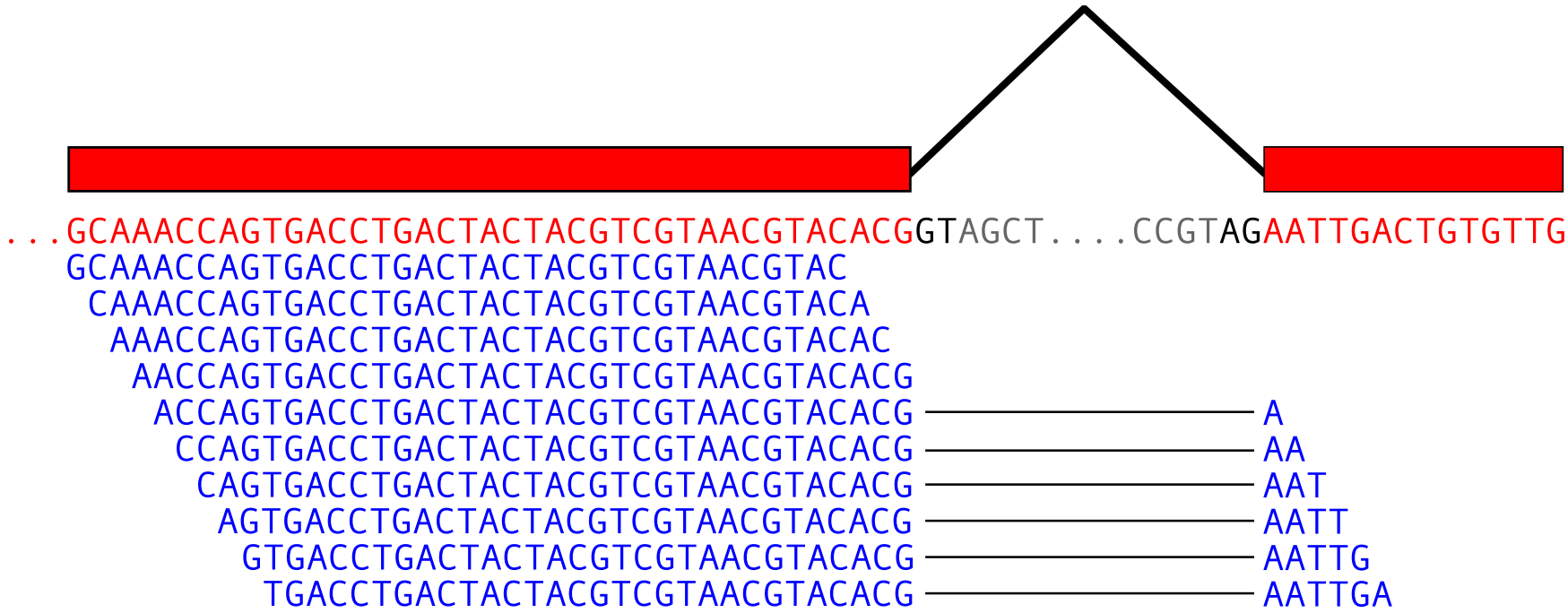
Spliced vs. Unspliced Alignments



```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
```

- Find matching region on genome with a few mismatches
- Efficient data structures for mapping many 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



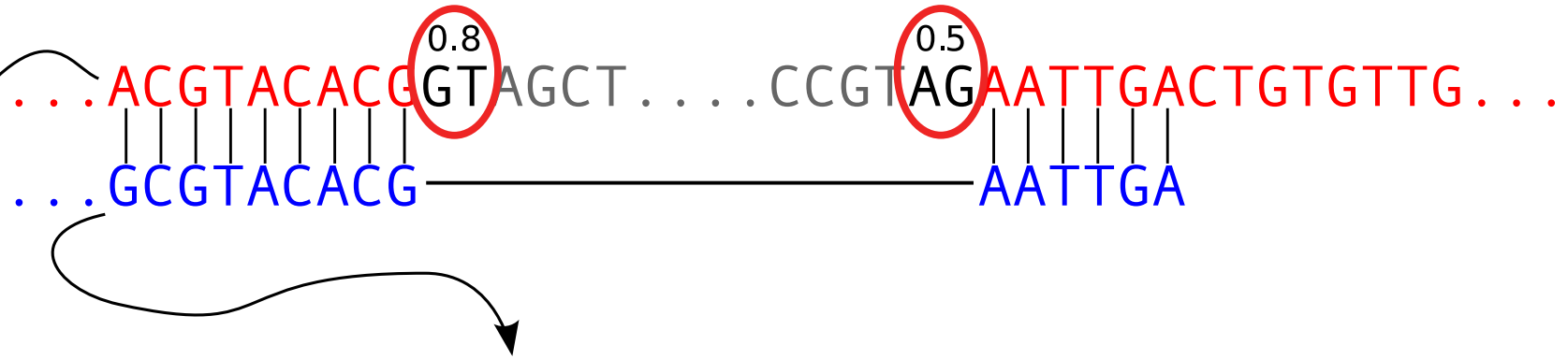
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

- Sequence matches

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

Extended Smith-Waterman Algorithm



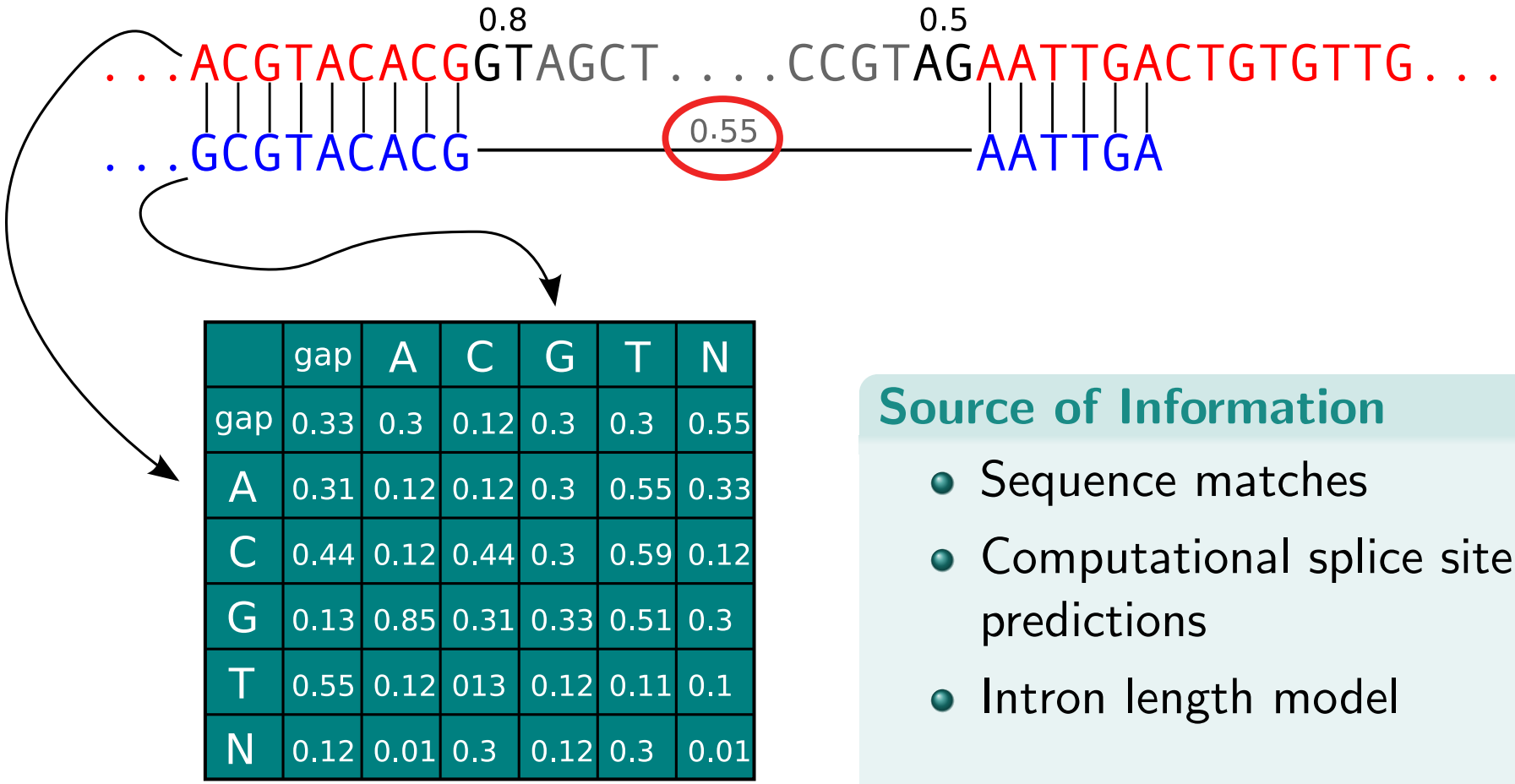
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

- Sequence matches
- Computational splice site predictions

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

Extended Smith-Waterman Algorithm



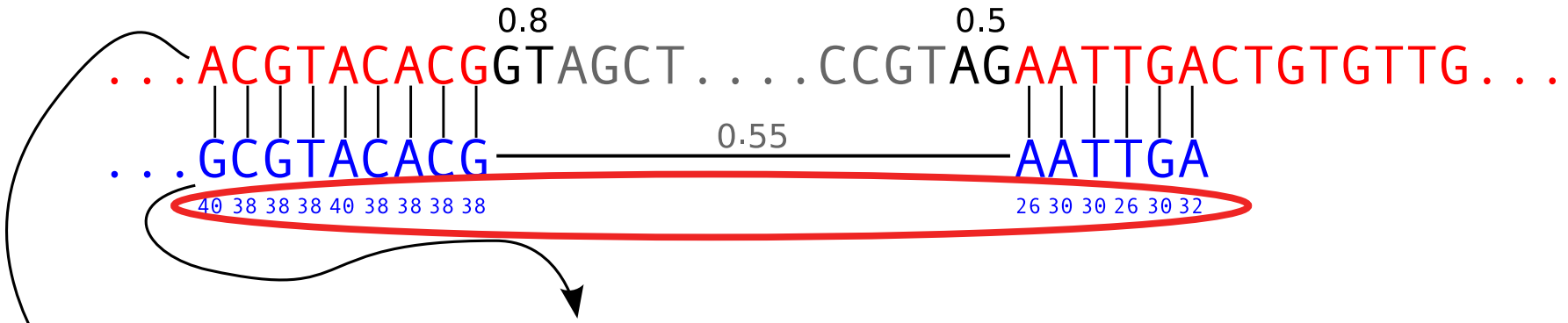
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

Source of Information

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

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

Extended Smith-Waterman Algorithm



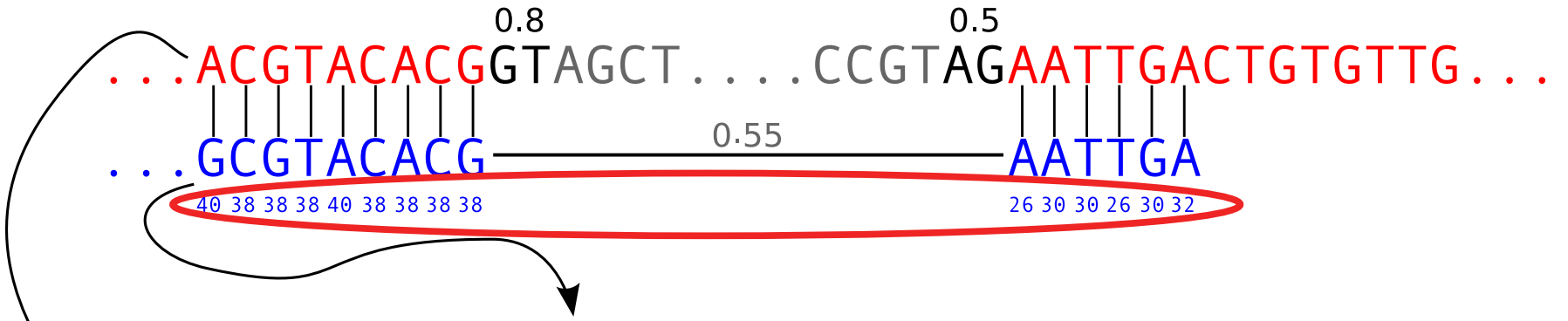
	gap	A	C	G	T	N
gap	0.33	$f_{-A}(\cdot)$	$f_{-C}(\cdot)$	$f_{-G}(\cdot)$	$f_{-T}(\cdot)$	$f_{-N}(\cdot)$
A	0.31	$f_{AA}(\cdot)$	$f_{AC}(\cdot)$	$f_{AG}(\cdot)$	$f_{AT}(\cdot)$	$f_{AN}(\cdot)$
C	0.44	$f_{CA}(\cdot)$	$f_{CC}(\cdot)$	$f_{CG}(\cdot)$	$f_{CT}(\cdot)$	$f_{CN}(\cdot)$
G	0.13	$f_{GA}(\cdot)$	$f_{GC}(\cdot)$	$f_{GG}(\cdot)$	$f_{GT}(\cdot)$	$f_{GN}(\cdot)$
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

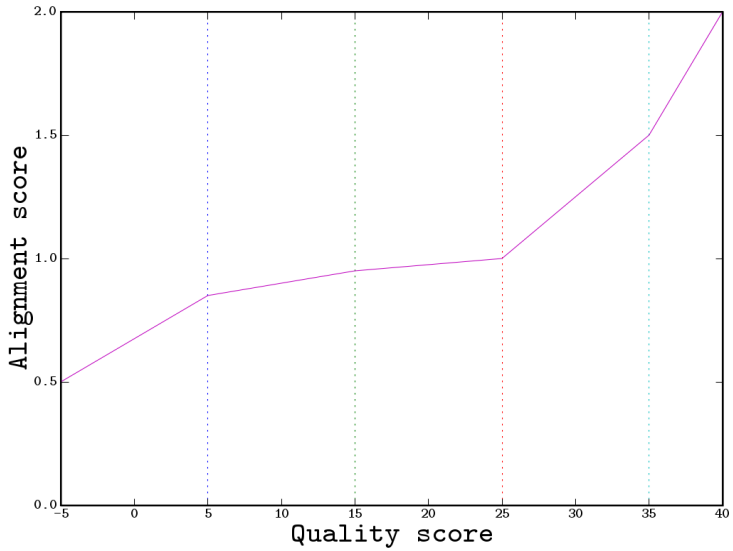
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

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

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	$f_{NA}(\cdot)$	$f_{NC}(\cdot)$	$f_{NG}(\cdot)$	$f_{NT}(\cdot)$	$f_{NN}(\cdot)$



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

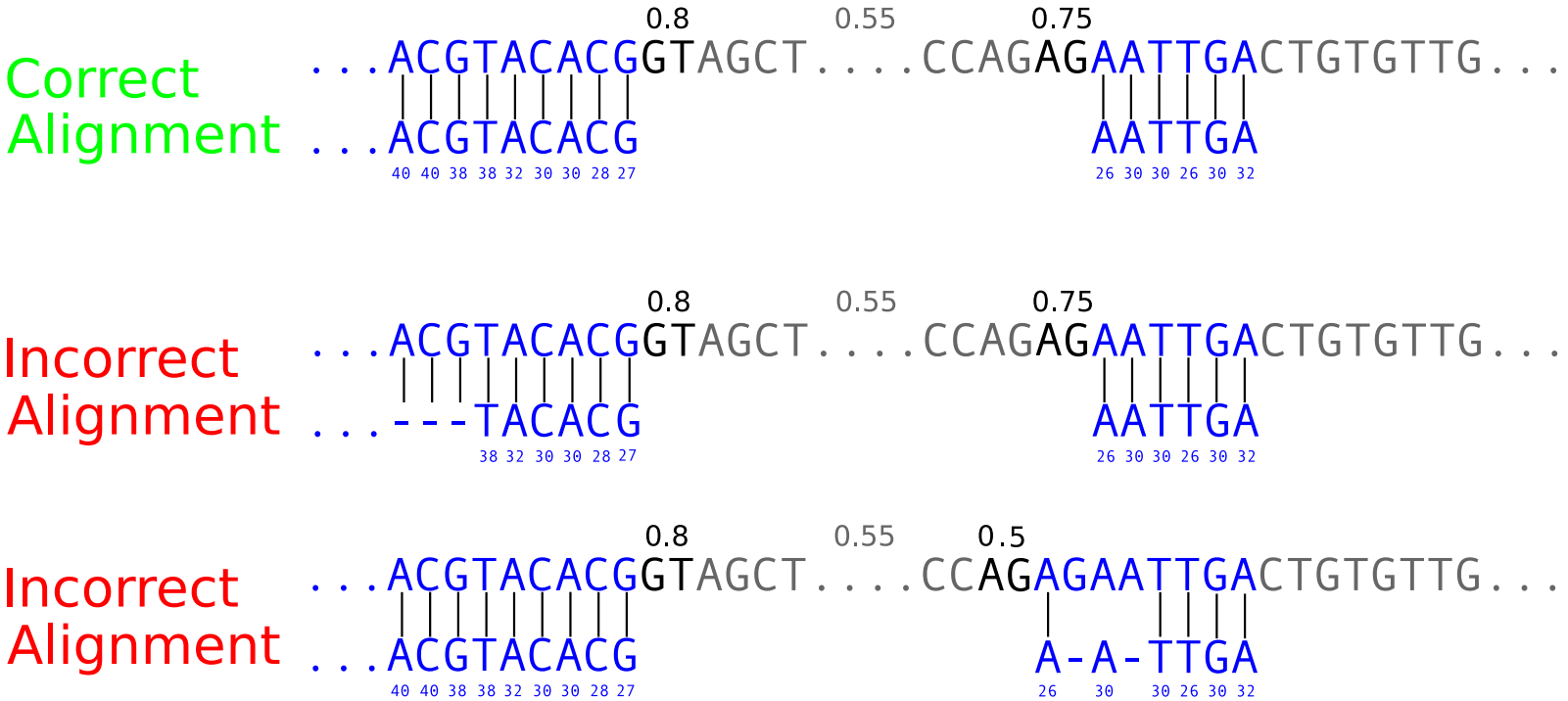
Scoring Parameter Inference



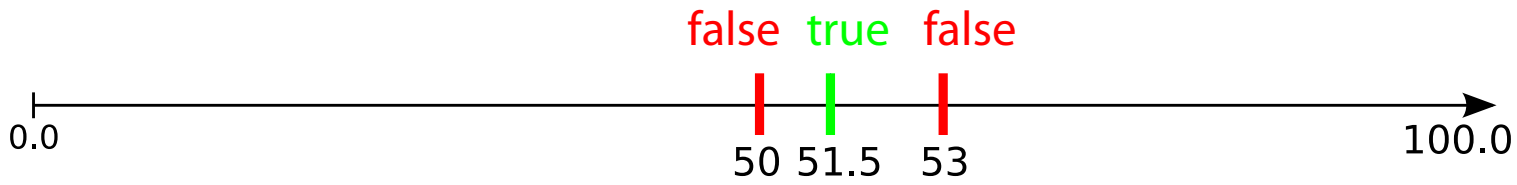
- What are optimal parameters?
- How do we jointly optimize 336 parameters?

Scoring Parameter Inference

- What are optimal parameters?
- How do we jointly optimize 336 parameters?

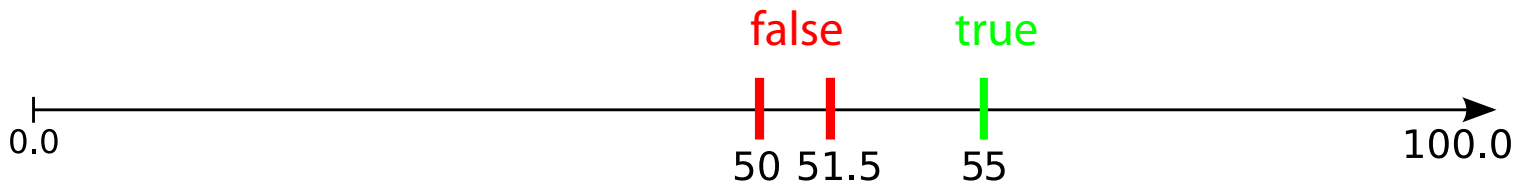


Cartoon: Maximize the Margin



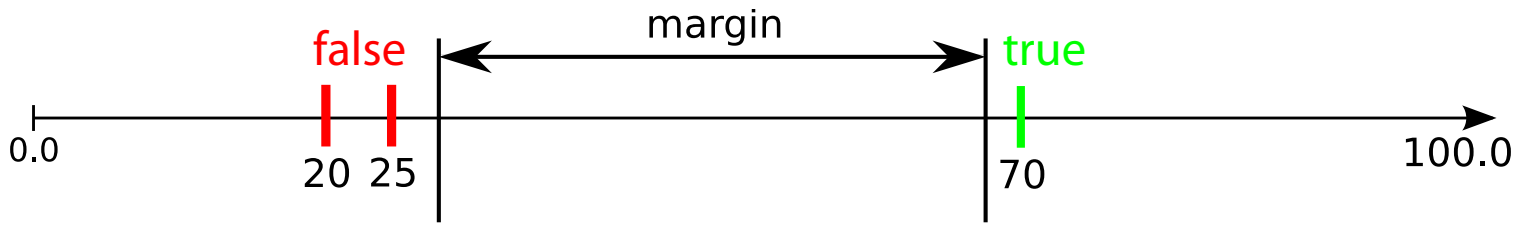
- Correct alignment is **not** highest scoring one

Cartoon: Maximize the Margin



- Correct alignment is highest scoring one
- Can we do better?

Cartoon: Maximize the Margin



- Use a technique motivated by "large-margin" methods
- Idea: Enforce a margin between correct and incorrect examples
- One has to solve a huge quadratic problem

How Can We Generate Data for Training?



- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads

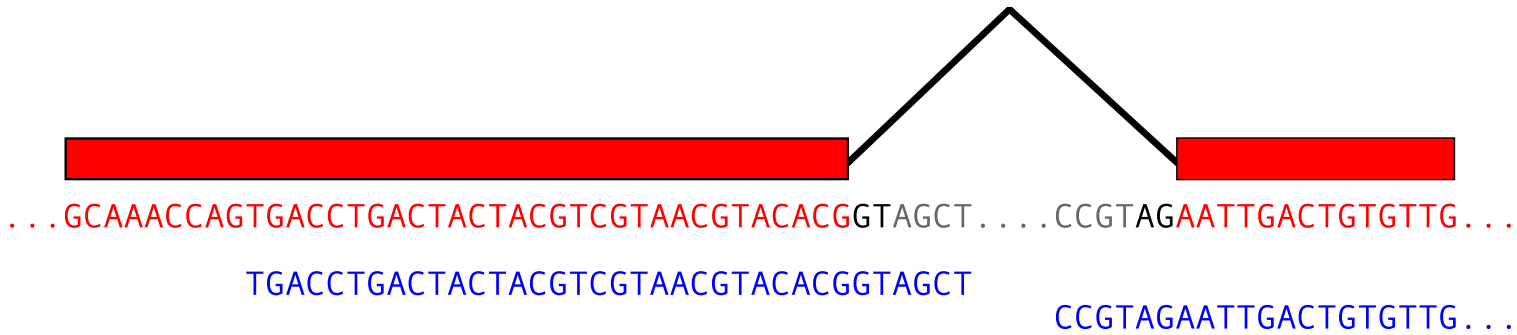
How Can We Generate Data for Training?



- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)

How Can We Generate Data for Training?

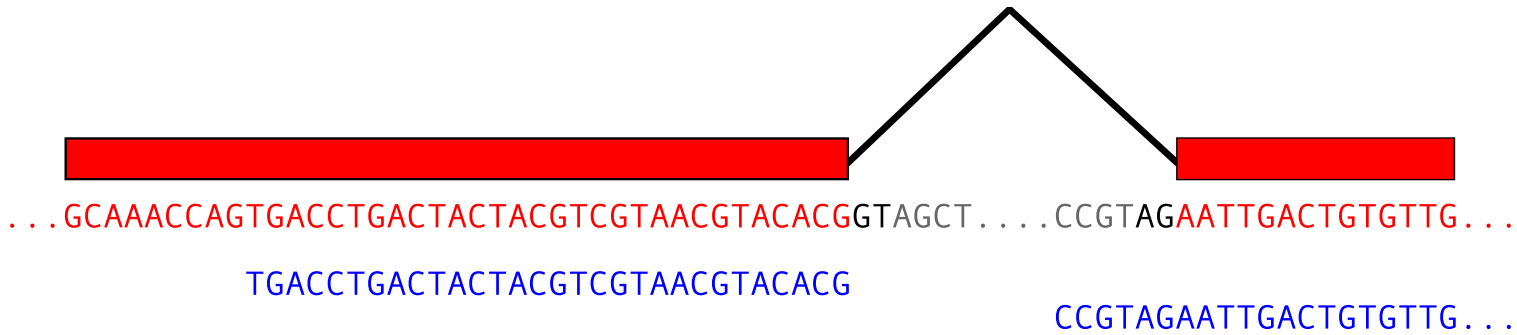
- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:



$\Rightarrow$ Identify reads overlapping into introns

How Can We Generate Data for Training?

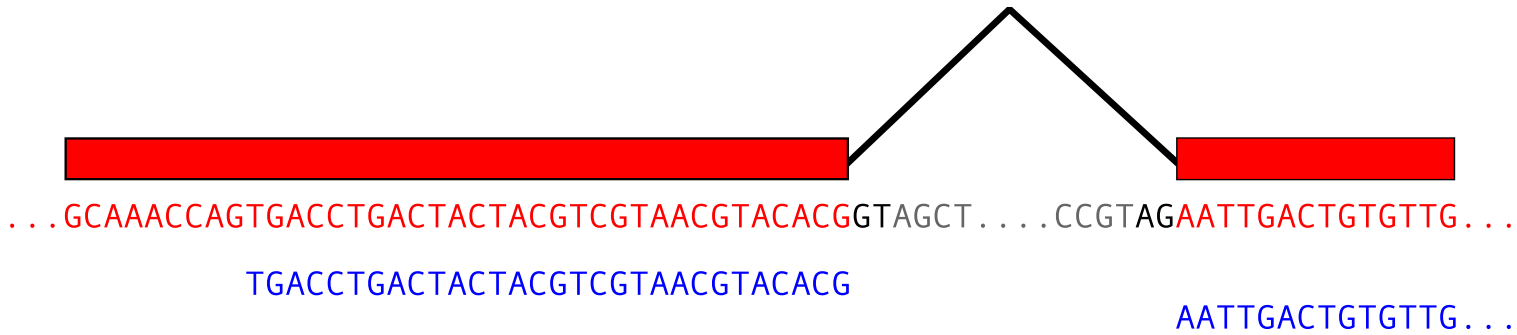
- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:



$\Rightarrow$ Remove overlapping parts

How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:



⇒ Remove overlapping parts

How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:



...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...

TGACCTGACTACTACGTCGTAACGTACACGAATTGACTGTGTTG...

$\Rightarrow$ Combine both exonic parts

How Can We Generate Data for Training?

- How do we obtain true alignments for training QPalma?
- Simulate *realistic* transcriptome reads
  - Consider well-annotated genome: *Arabidopsis thaliana*
  - Use short *genomic* reads (38nt) with natural quality scores
  - Use reference *annotation* (TAIR7)
  - Generate artificially spliced reads:



...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...

TGACCTGACTACTACGTCGTAACGTACACGAATTGA

⇒ Truncate to desired length (38nt)

Results on Artificially Spliced Reads



- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment

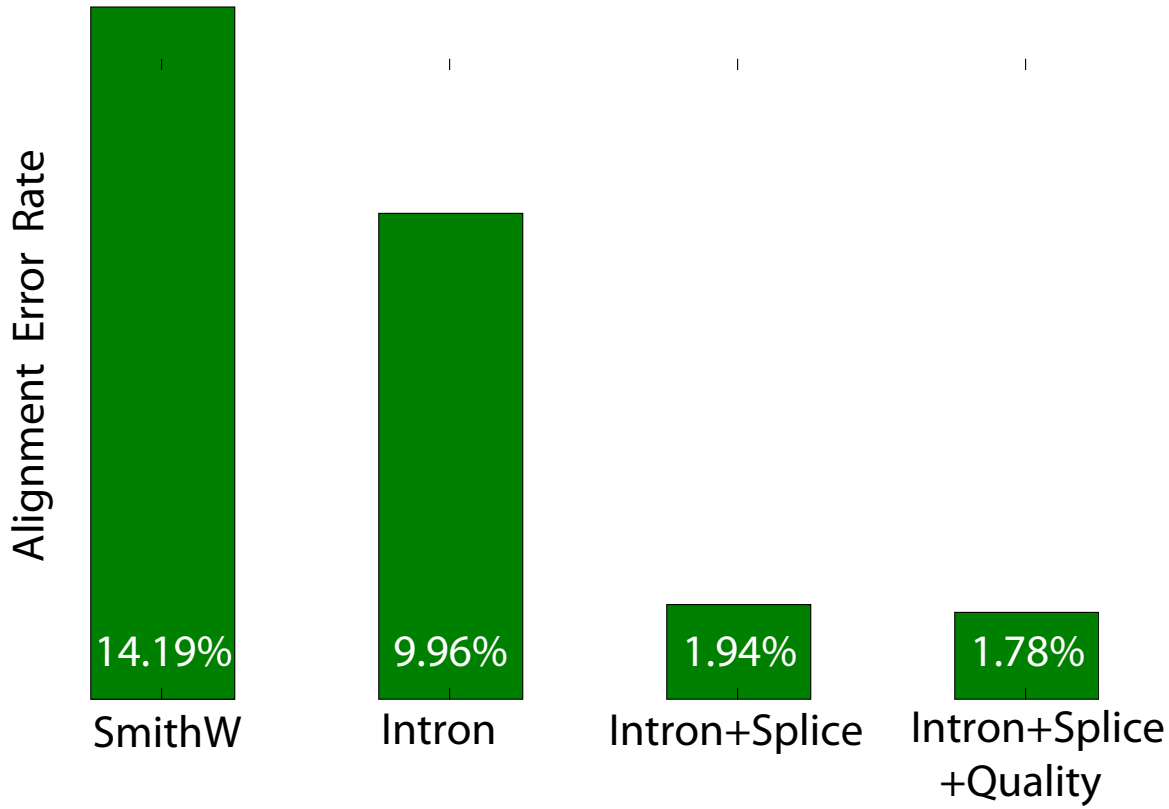
Results on Artificially Spliced Reads



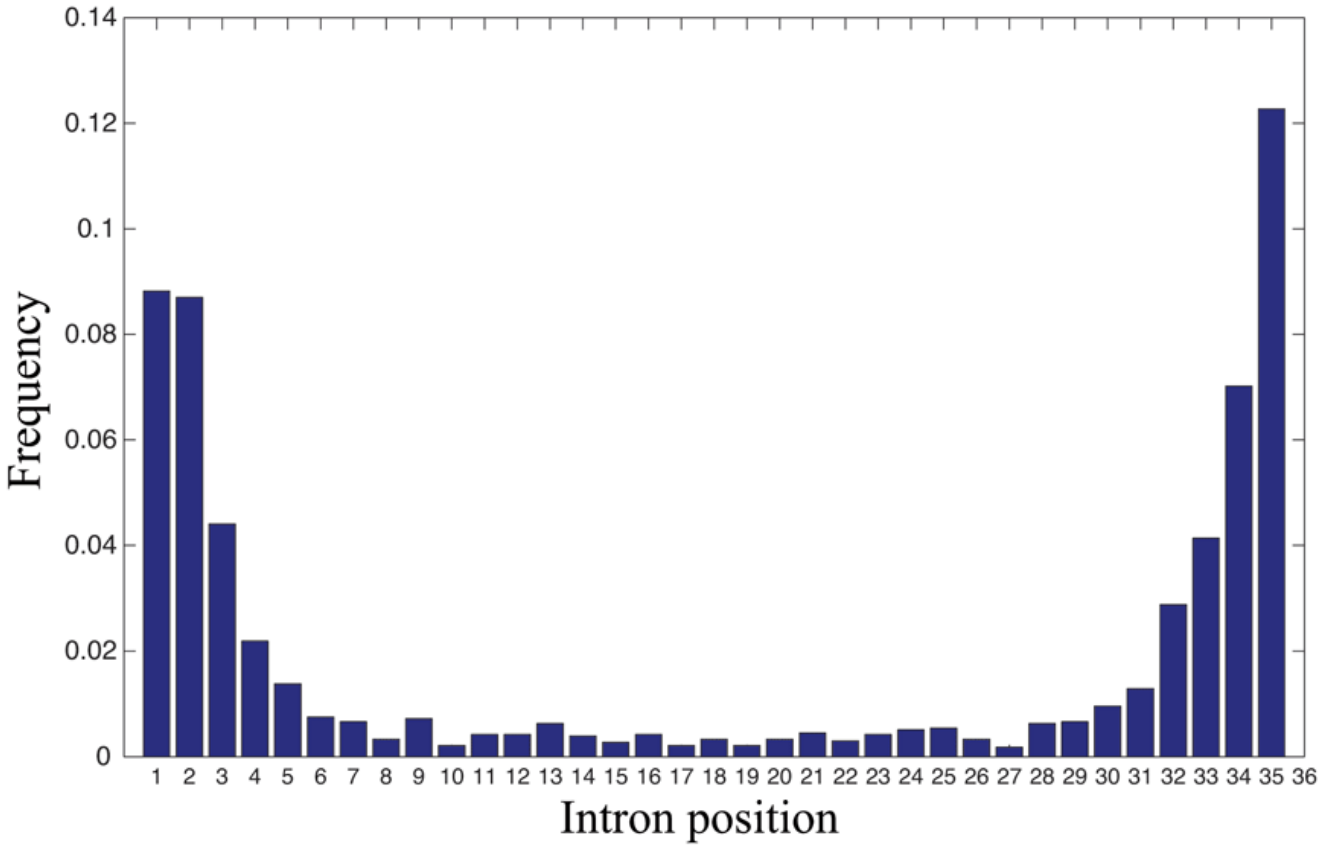
- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
- Trained on $10,000$ artificially spliced genome reads
- Tested on $30,000$ spliced reads

Results on Artificially Spliced Reads

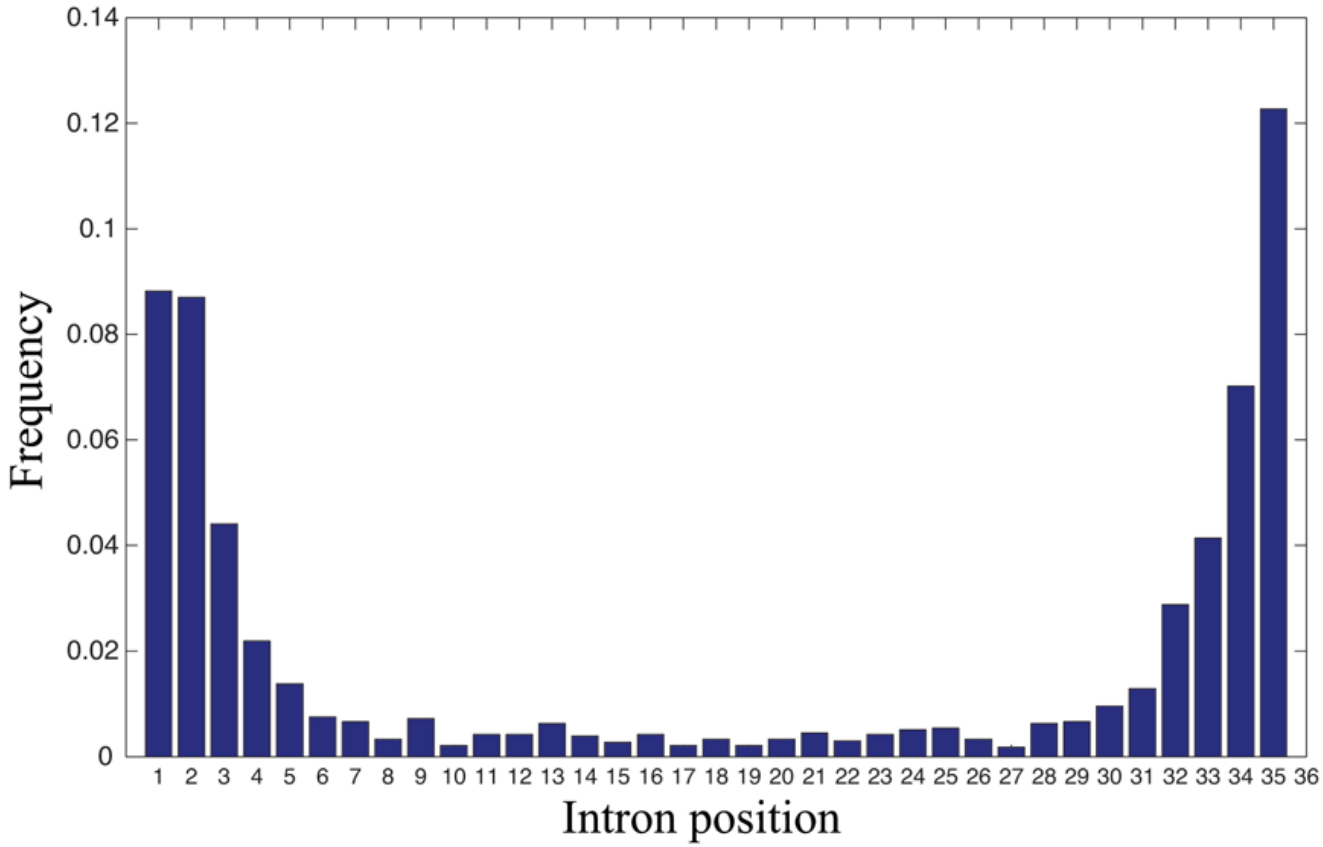
- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment
- Trained on $10,000$ artificially spliced genome reads
- Tested on $30,000$ spliced reads



Error Rate Depends on Intron Position

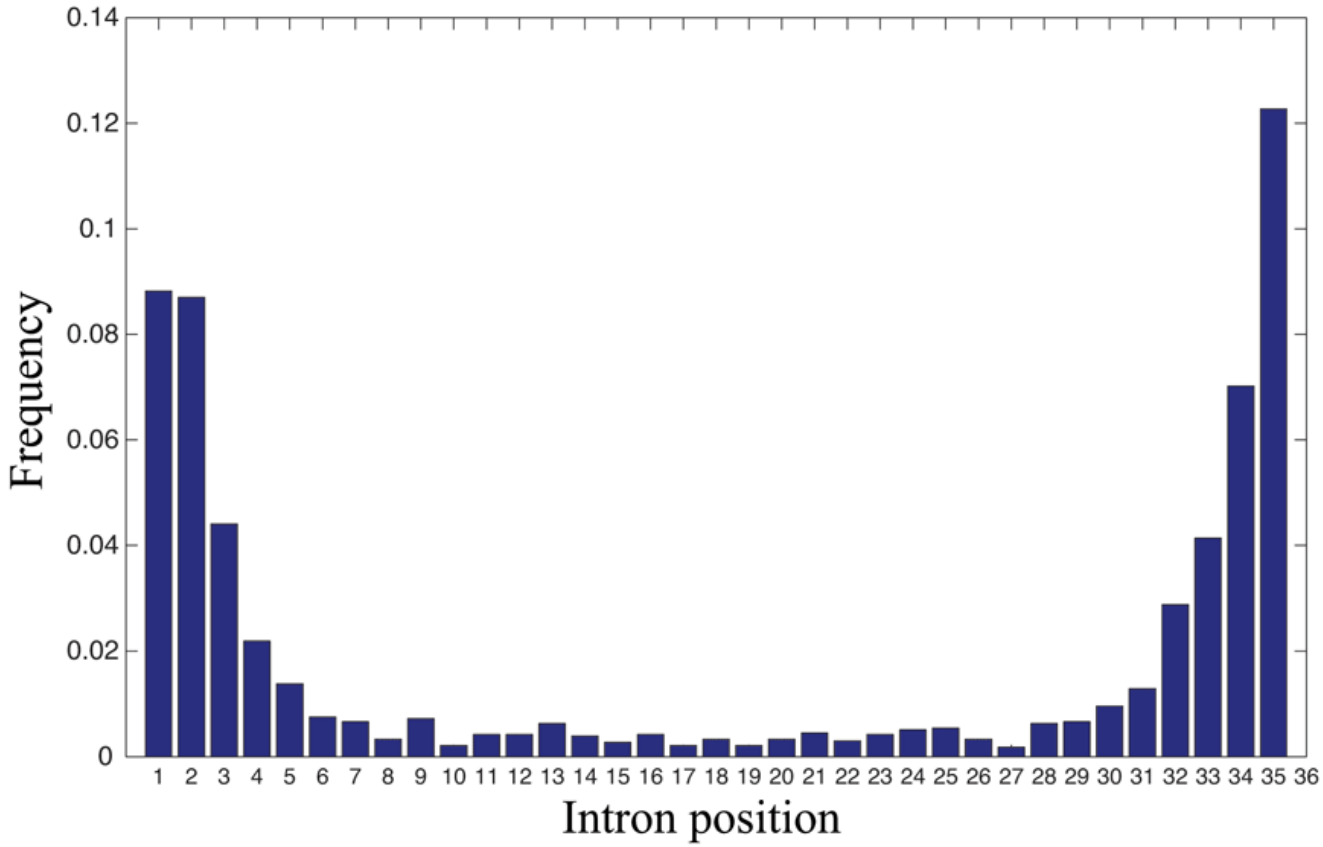


Error Rate Depends on Intron Position



1. Trust introns confirmed by spliced read with $\geq$ 6nt overlap

Error Rate Depends on Intron Position



1. Trust introns confirmed by spliced read with $\geq$ 6nt overlap
2. Pure Smith-Waterman algorithm would need longer overlaps and would still perform worse

Can We Use QPalma for Whole-Genome Alignments?



So far we had two assumptions:

1. All reads are spliced
2. Genomic region is known

Can We Use QPalma for Whole-Genome Alignments?

So far we had two assumptions:

1. All reads are spliced
2. Genomic region is known

1. Many reads will be fully contained in an exon
2. Direct alignment to genome too expensive ($\mathcal{O}(n \cdot m)$)



```
...GCAAACCAGTGACCTGACTACTACGTCGTAACGTACACGGTAGCT....CCGTAGAATTGACTGTGTTG...
   GCAAACCAGTGACCTGACTACTACGTCGTAACGTAC
    CAAACCAGTGACCTGACTACTACGTCGTAACGTACA
     AAACCAGTGACCTGACTACTACGTCGTAACGTACAC
      AACCAGTGACCTGACTACTACGTCGTAACGTACACG
       ACCAGTGACCTGACTACTACGTCGTAACGTACACG ———————————— A
        CCAGTGACCTGACTACTACGTCGTAACGTACACG ———————————— AA
         CAGTGACCTGACTACTACGTCGTAACGTACACG ———————————— AAT
          AGTGACCTGACTACTACGTCGTAACGTACACG ———————————— AATT
           GTGACCTGACTACTACGTCGTAACGTACACG ———————————— AATTG
            TGACCTGACTACTACGTCGTAACGTACACG ———————————— AATTGA
```

A Pipeline for Efficient Large Scale Alignment



Pipeline Workflow

1. Map unspliced reads & find seed regions
   - Use suffix-array based method to find short read match with at most two mismatches (*Vmatch*, Kurtz et al.; *ShoRe*, Ossowski et al. 2008)

A Pipeline for Efficient Large Scale Alignment



Pipeline Workflow

1. Map unspliced reads & find seed regions
   - Use suffix-array based method to find short read match with at most two mismatches (*Vmatch*, Kurtz et al.; *ShoRe*, Ossowski et al. 2008)
2. Recover potentially spliced reads from first mapping
   - Use fast approximation of QPalma to decide which reads may be spliced (even if they can be mapped well)

A Pipeline for Efficient Large Scale Alignment



Pipeline Workflow

1. Map unspliced reads & find seed regions
   - Use suffix-array based method to find short read match with at most two mismatches (*Vmatch*, Kurtz et al.; *ShoRe*, Ossowski et al. 2008)
2. Recover potentially spliced reads from first mapping
   - Use fast approximation of QPalma to decide which reads may be spliced (even if they can be mapped well)
3. Identify accurate alignments for the candidate spliced reads
   - Use full QPalma model to align remaining reads

A Pipeline for Efficient Large Scale Alignment



Pipeline Workflow

1. Map unspliced reads & find seed regions
   - Use suffix-array based method to find short read match with at most two mismatches (*Vmatch*, Kurtz et al.; *ShoRe*, Ossowski et al. 2008)
   - $\approx$ 4**h for** $2,6 \times 10^6$ **reads**
2. Recover potentially spliced reads from first mapping
   - Use fast approximation of QPalma to decide which reads may be spliced (even if they can be mapped well)
   - $\approx$ 17**min for** $2,2 \times 10^6$ **reads**
3. Identify accurate alignments for the candidate spliced reads
   - Use full QPalma model to align remaining reads
   - $\approx$ 8**h for** $\sim 442,000$ **reads**

Application to Natural Transcriptome Reads



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

- 4 lanes from Illumina Genome Analyzer 1G ($\approx 50\times$ coverage)
- 38nt reads, polyA enriched
- Strand unspecific, young leaves

Application to Natural Transcriptome Reads



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

- 4 lanes from Illumina Genome Analyzer 1G ($\approx 50\times$ coverage)
- 38nt reads, polyA enriched
- Strand unspecific, young leaves

- Initial Read mapping by *ShoRe* (`http://1001genomes.org`)
- Spliced alignment by *QPalma* (`http://fml.mpg.de/raetsch/projects/qpalma`)

Application to Natural Transcriptome Reads

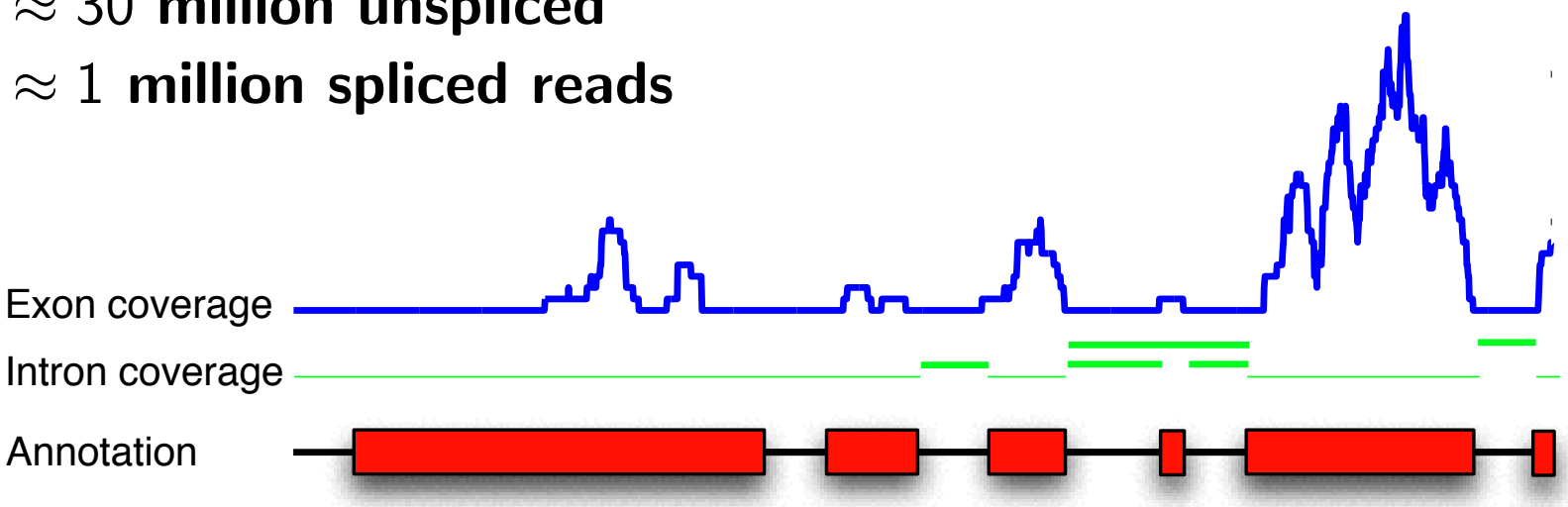
RNA-seq reads for *A. thaliana* (provided by Weigel lab, MPI Devel. Biology)

- 4 lanes from Illumina Genome Analyzer 1G ($\approx 50\times$ coverage)
- 38nt reads, polyA enriched
- Strand unspecific, young leaves

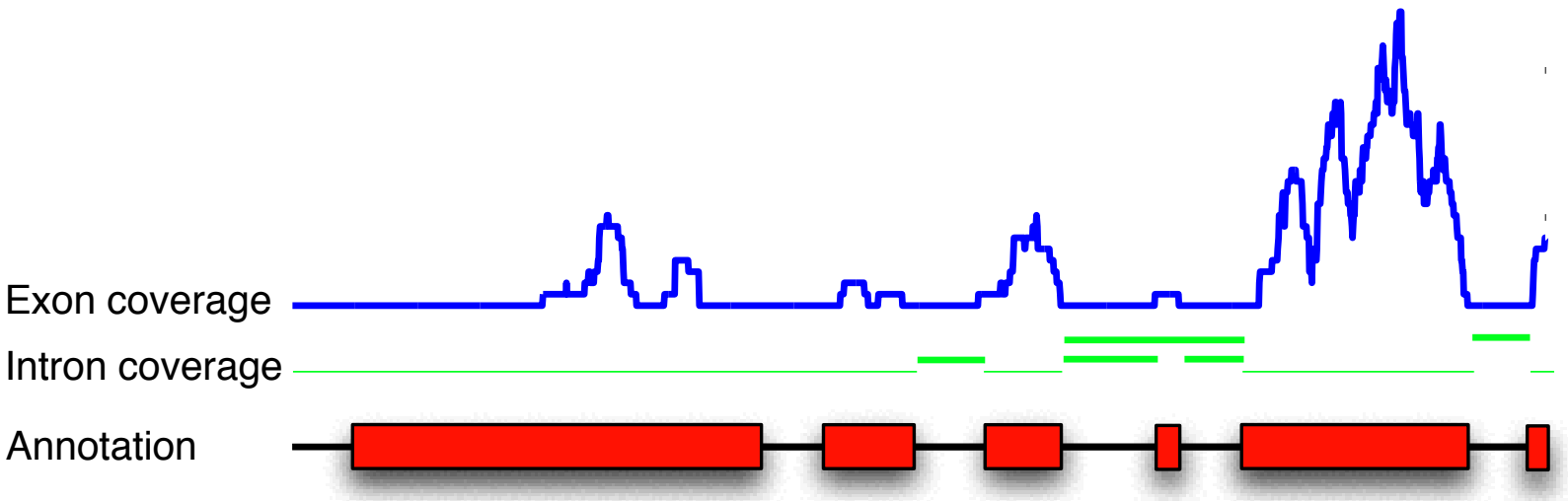
- Initial Read mapping by *ShoRe* (`http://1001genomes.org`)
- Spliced alignment by *QPalma* (`http://fml.mpg.de/raetsch/projects/qpalma`)

$\Rightarrow$ ≈ 30 **million unspliced**

$\Rightarrow$ ≈ 1 **million spliced reads**



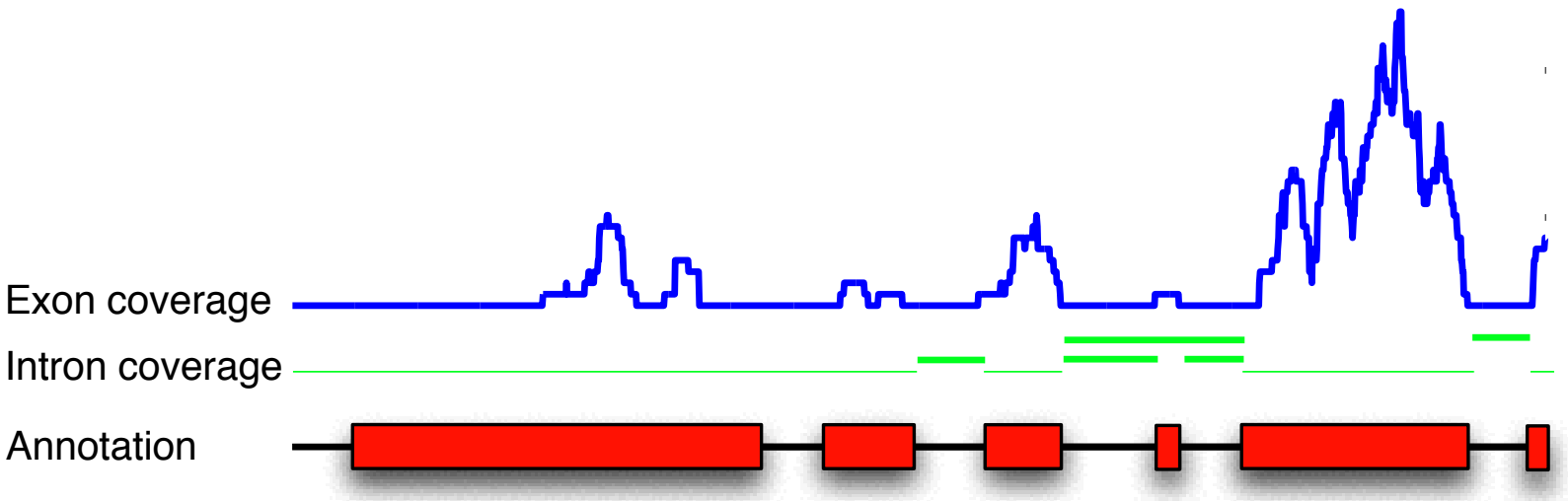
What Can We Do With It?



For instance:

- Assemble alignments to obtain splice graphs (describing alternative splicing)
  - Only works well for highly covered genes

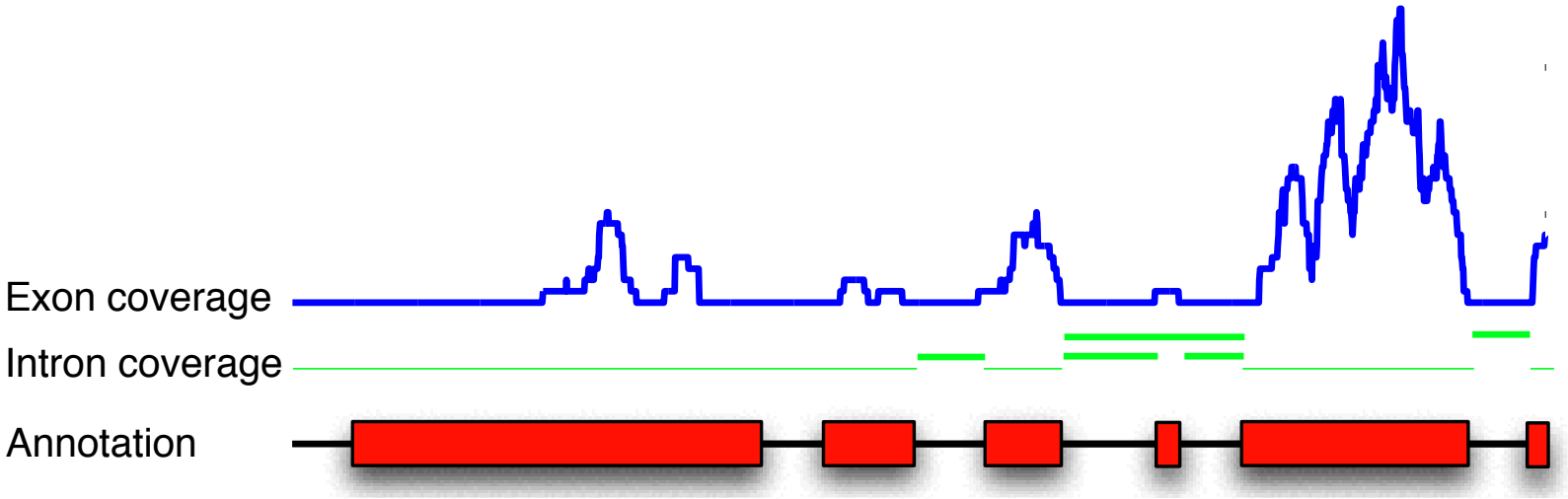
What Can We Do With It?



For instance:

- Assemble alignments to obtain splice graphs (describing alternative splicing)
  - Only works well for highly covered genes
- Use exon/intron coverage tracks to improve gene finding
  - Considerable improvements appear possible

What Can We Do With It?



For instance:

- Assemble alignments to obtain splice graphs (describing alternative splicing)
  - Only works well for highly covered genes
- Use exon/intron coverage tracks to improve gene finding
  - Considerable improvements appear possible
- Estimate relative abundances of alternative transcripts
  - Spliced reads help as they can connect exons

Summary



- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads

Summary



- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads
- Proposed algorithm for parameter estimation
- Successfully combined many information sources

Summary



- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads
- Proposed algorithm for parameter estimation
- Successfully combined many information sources
- Can be applied to data from other sequencing platforms (ABI, ...)
- Fast enough for whole genome/transcriptome sequencing
  - Currently, very good for smaller genomes
  - When excluding indels, then also fast for mouse/human (in progress)

Summary

- New method for accurate *de novo* spliced alignments of short reads
- Challenging data: many short & error prone reads
- Proposed algorithm for parameter estimation
- Successfully combined many information sources
- Can be applied to data from other sequencing platforms (ABI, ...)
- Fast enough for whole genome/transcriptome sequencing
  - Currently, very good for smaller genomes
  - When excluding indels, then also fast for mouse/human (in progress)

Potential improvements

- Speed-up dynamic program
- Integrated seed region finding and spliced alignments

Acknowledgements

Coauthors

- Stephan Ossowski
- Korbinian Schneeberger
- Gunnar Rätsch

Weigel Lab – DNA- and RNA-seq Data

- Jun Cao
- Richard Clark
- Christa Lanz
- Detlef Weigel

Stipend

- Friedrich Miescher Laboratory of the Max Planck Society

Thank you!

fabio@tuebingen.mpg.de

http://www.fml.tuebingen.mpg.de/raetsch/projects/qpalma

Poster D19



Results on Artificially Spliced Reads



- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment

Results on Artificially Spliced Reads



- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment

- Trained on $10,000$ artificially spliced genome reads
- Tested on $30,000$ spliced reads

Results on Artificially Spliced Reads

- Given: Short reads and corresponding genomic regions
- Task: Find the correct spliced alignment

- Trained on $10,000$ artificially spliced genome reads
- Tested on $30,000$ spliced reads

Quality information	Splice site pred.	Intron length	Error rate
-	-	-	14.19 %
+	-	-	13.49 %
-	-	+	9.96 %
+	-	+	9.68 %
-	+	-	3.16 %
+	+	-	2.81 %
-	+	+	1.94 %
+	+	+	**1.78%**