Accurate RNA-seq Alignments with PALMapper

Gunnar Rätsch, André Kahles, Géraldine Jean



Friedrich Miescher Laboratory
of the Max Planck Society

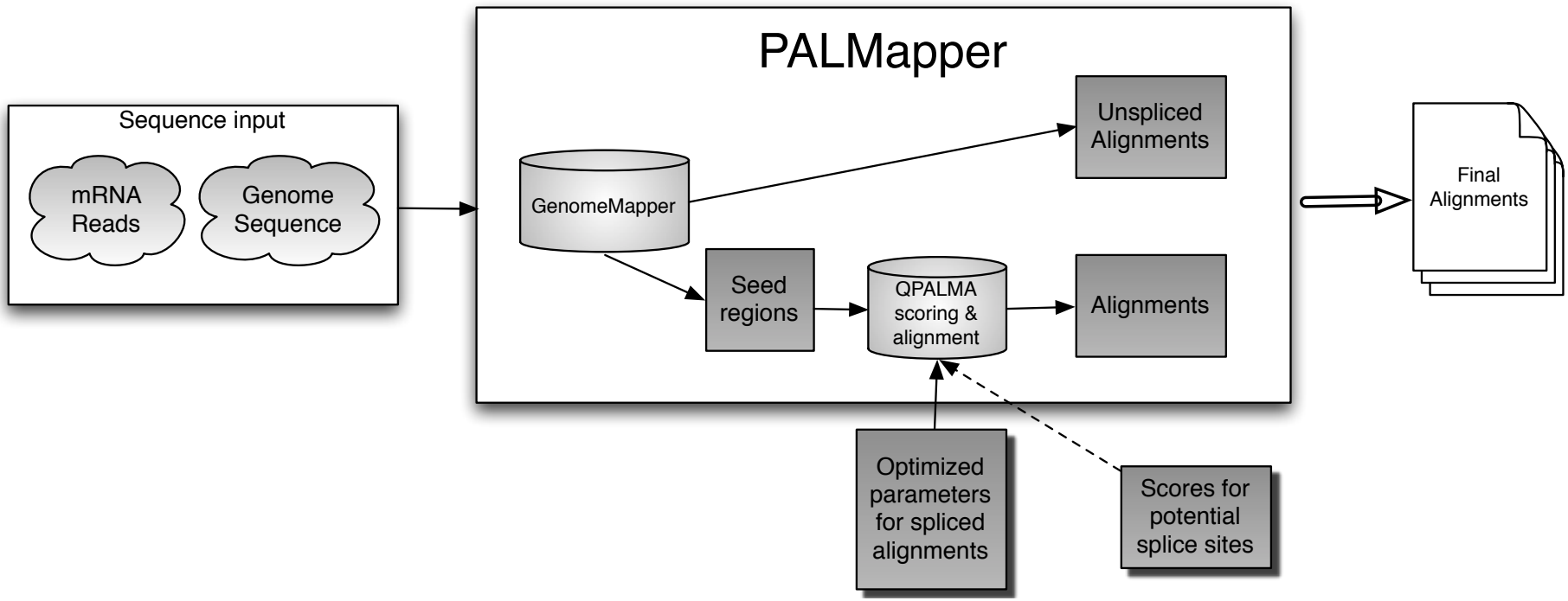
Tübingen, Germany

SMAAP-RGASP3, Barcelona, April 6, 2011

RNA-seq Read Alignments with PALMapper

PALMapper pipeline [Jean et al., 2010]

- GenomeMapper identifies unspliced alignments and *seed regions* for spliced reads [Schneeberger et al., 2009]
- QPALMA infers spliced alignments from *seed regions* [De Bona et al., 2008]



Unspliced read mapping with GenomeMapper



Read mapper developed for the 1001 Plant Project Genome

- Indexing genome with k-mer based index or bwt-based index
- Reporting all extended hits within a range of mismatches/gaps
  - unspliced alignments: mapping of the whole read
  - seed regions: mapping of parts from a plausible splice read

Unspliced read mapping with GenomeMapper



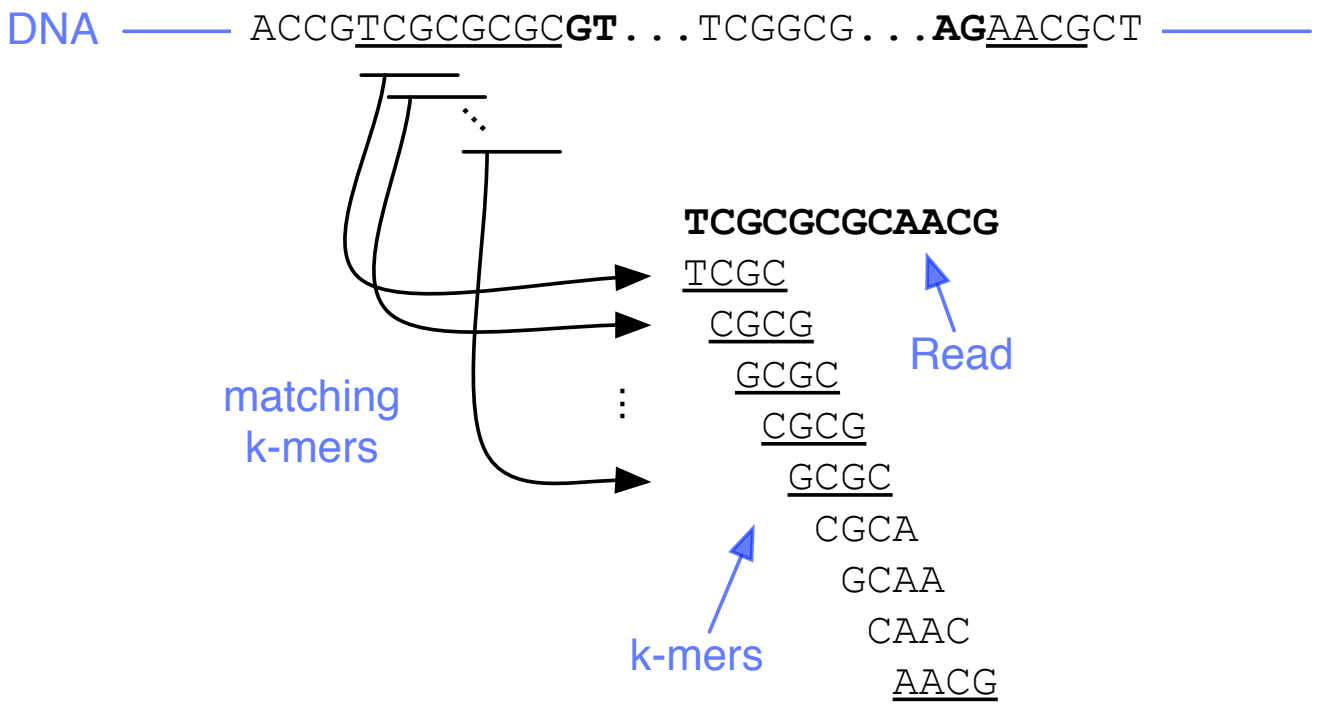
Read mapper developed for the 1001 Plant Project Genome

- Indexing genome with *k*-mer based index or bwt-based index
- Reporting all extended hits within a range of mismatches/gaps
  - unspliced alignments: mapping of the whole read
  - seed regions: mapping of parts from a plausible splice read

Unspliced read mapping with GenomeMapper

Read mapper developed for the 1001 Plant Project Genome

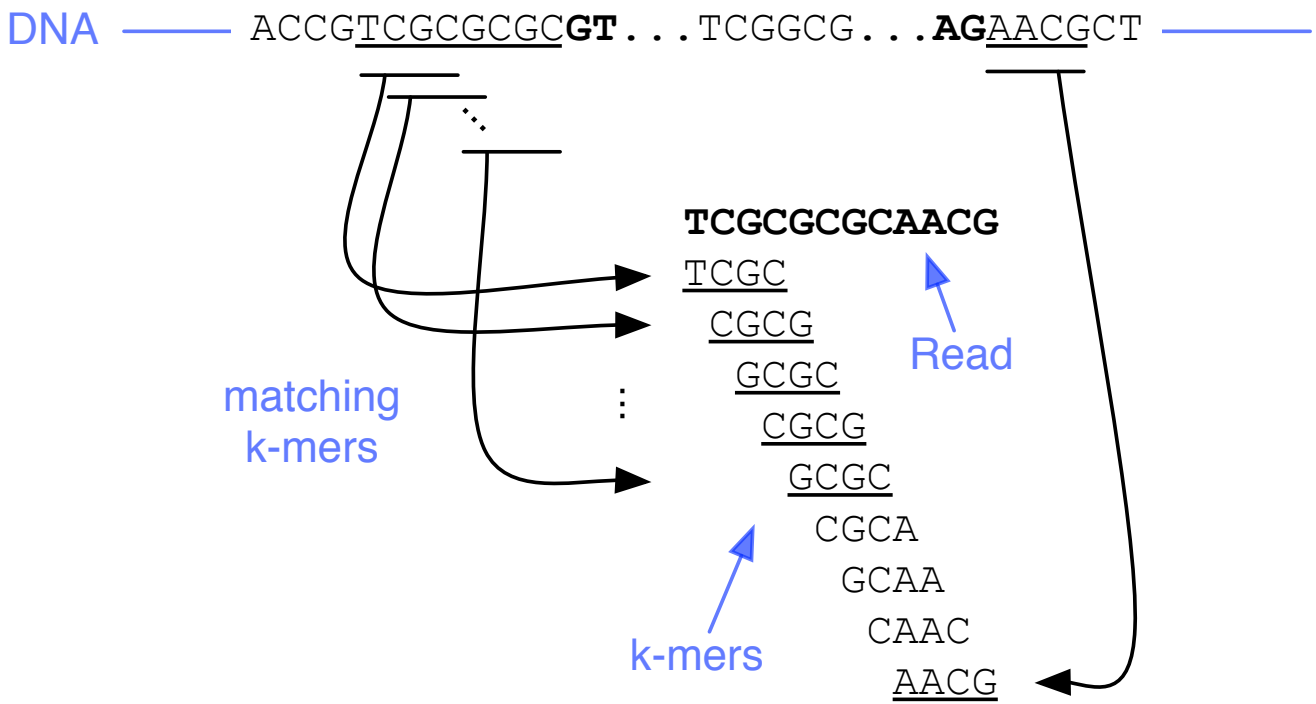
- Indexing genome with *k*-mer based index or bwt-based index
- Reporting all extended hits within a range of mismatches/gaps
  - unspliced alignments: mapping of the whole read
  - seed regions: mapping of parts from a plausible splice read



Unspliced read mapping with GenomeMapper

Read mapper developed for the 1001 Plant Project Genome

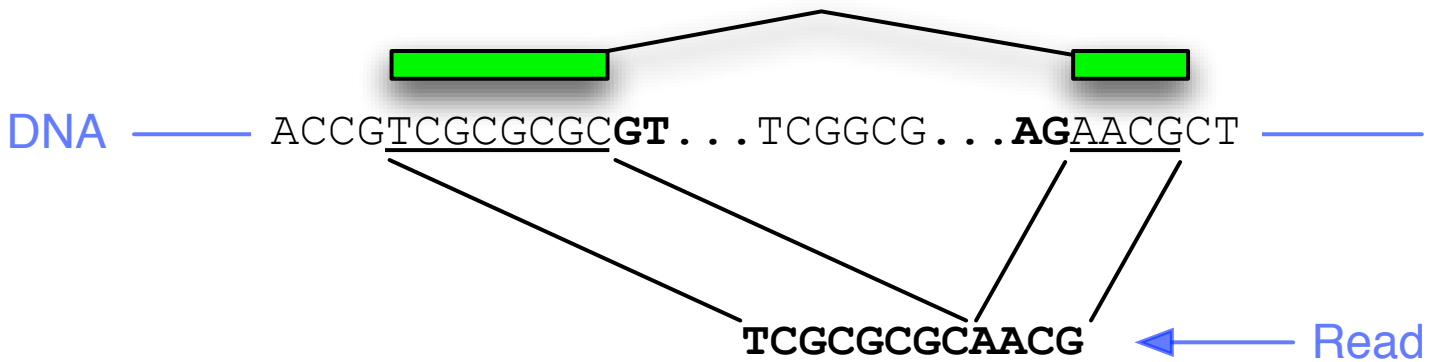
- Indexing genome with *k*-mer based index or bwt-based index
- Reporting all extended hits within a range of mismatches/gaps
  - unspliced alignments: mapping of the whole read
  - seed regions: mapping of parts from a plausible splice read



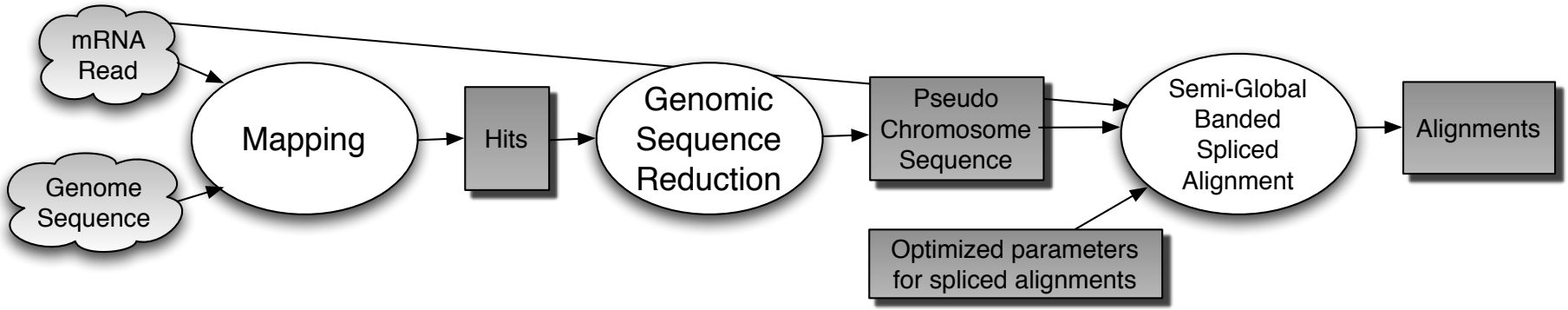
Unspliced read mapping with GenomeMapper

Read mapper developed for the 1001 Plant Project Genome

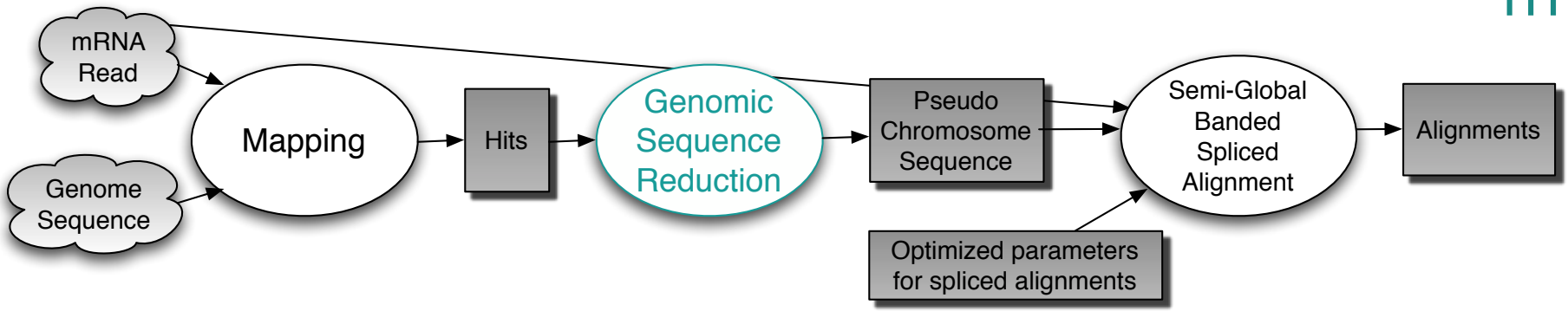
- Indexing genome with *k*-mer based index or bwt-based index
- Reporting all extended hits within a range of mismatches/gaps
  - unspliced alignments: mapping of the whole read
  - seed regions: mapping of parts from a plausible splice read



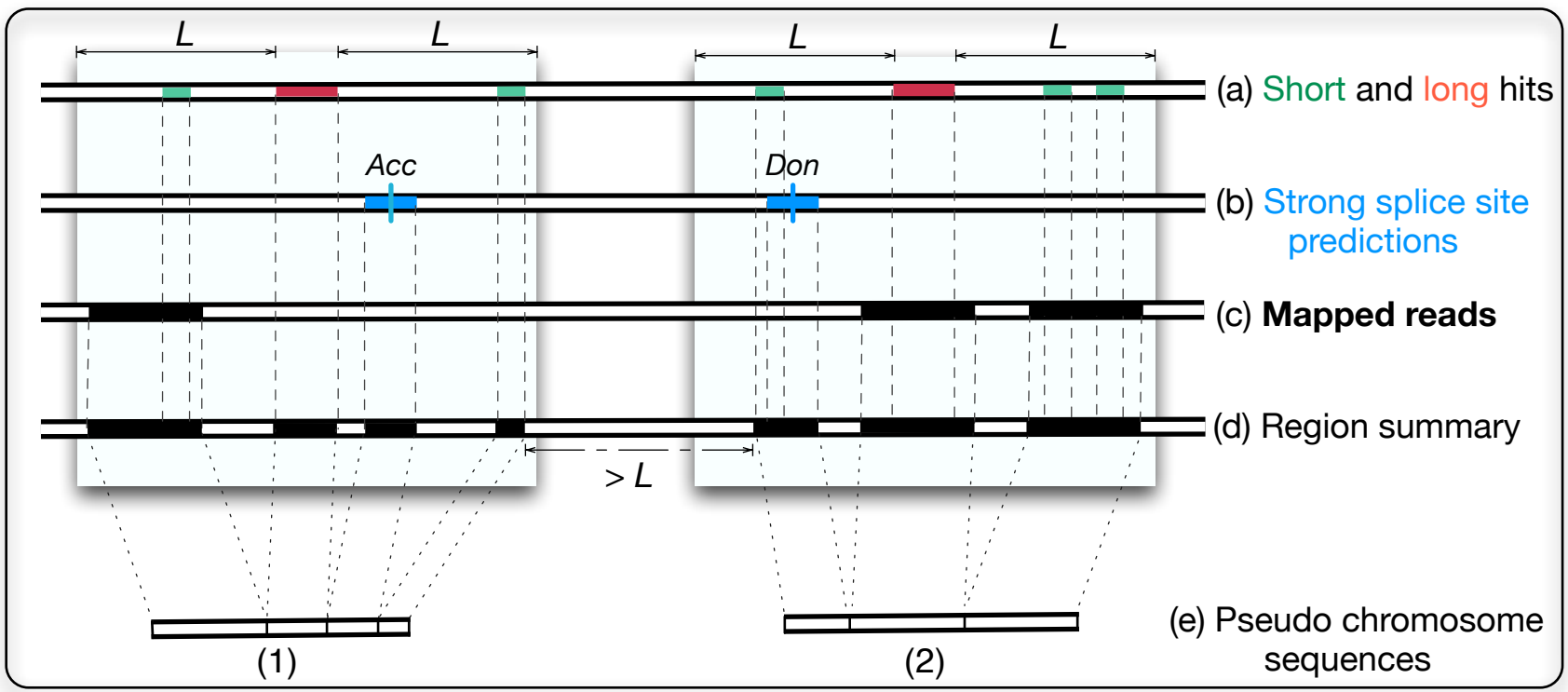
Fast and Accurate spliced alignments with PALMapper



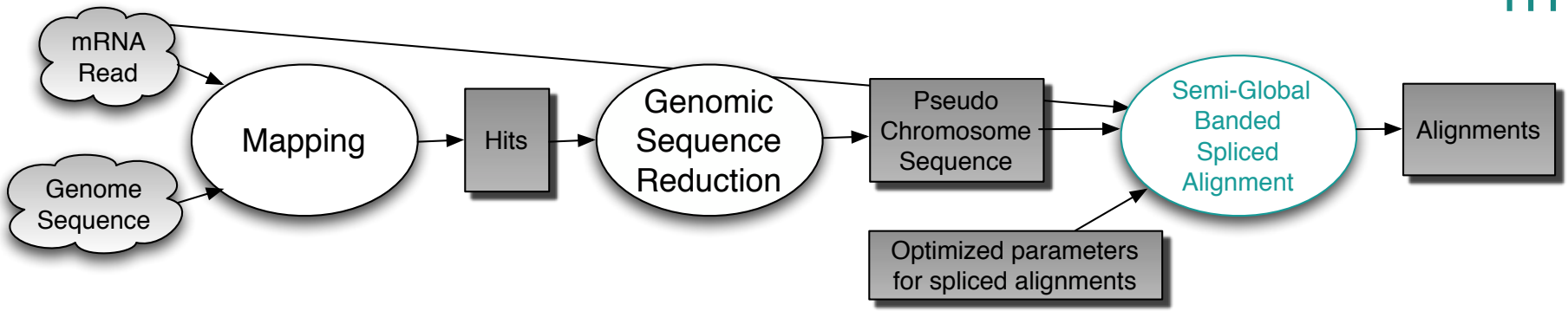
Fast and Accurate spliced alignments with PALMapper



Pseudo Chromosome Sequence from seed regions (long hits)



Fast and Accurate spliced alignments with PALMapper



The *seed position* inferred from the seed region guides a dynamic-programming-based alignment algorithm:

Semi-Global

Align the whole read with a portion of the pseudo chromosome sequence

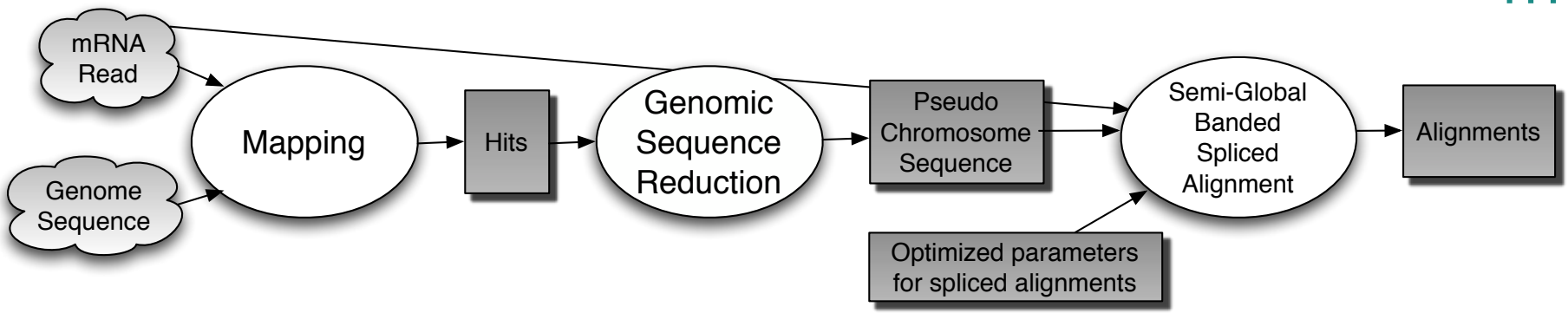
Banded

Limit the number of gaps from the perfect alignment

Spliced

Allow long gaps corresponding to introns

Fast and Accurate spliced alignments with PALMapper



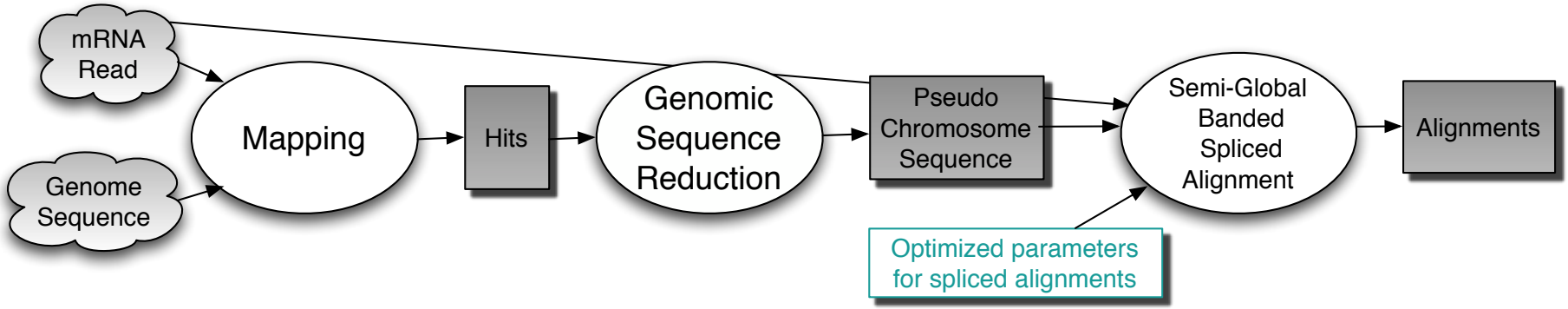
Significant Speed Gain: QPalma *vs.* Palmapper

- Full sequencing run of *C. elegans* RNA-Seq data of 24×10^6 reads of 36-nucleotide length
- Evaluation of predicted introns

	TopHat	TopHat sen.	**QPALMA**	**PALMapper**
Recall	0.39	0.62	**0.65**	**0.65**
Precision	0.86	0.76	**0.88**	**0.88**
Running Time (min)	216	544	12000[1]	**186**

TopHat [Trapnell et al., 2009]. [1]The running time for QPALMA was extrapolated.

Fast and Accurate spliced alignments with PALMapper





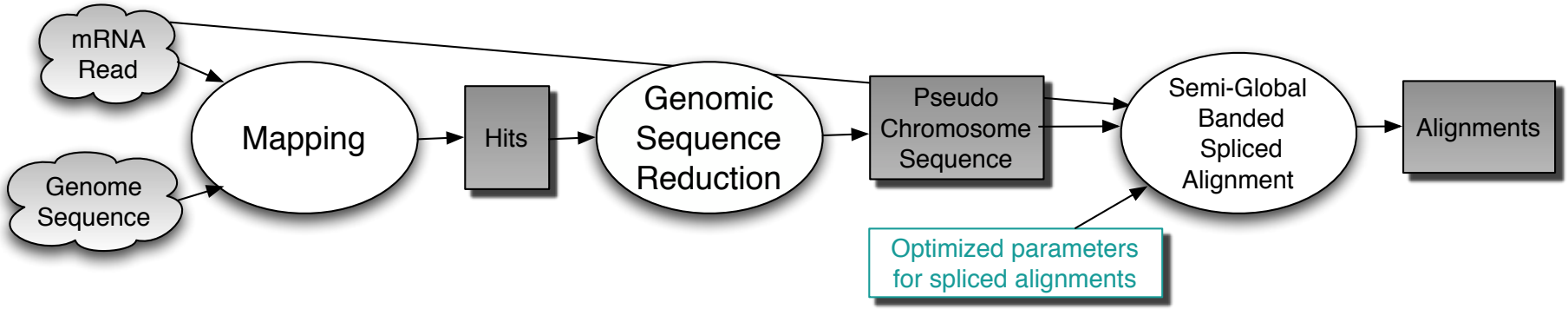
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

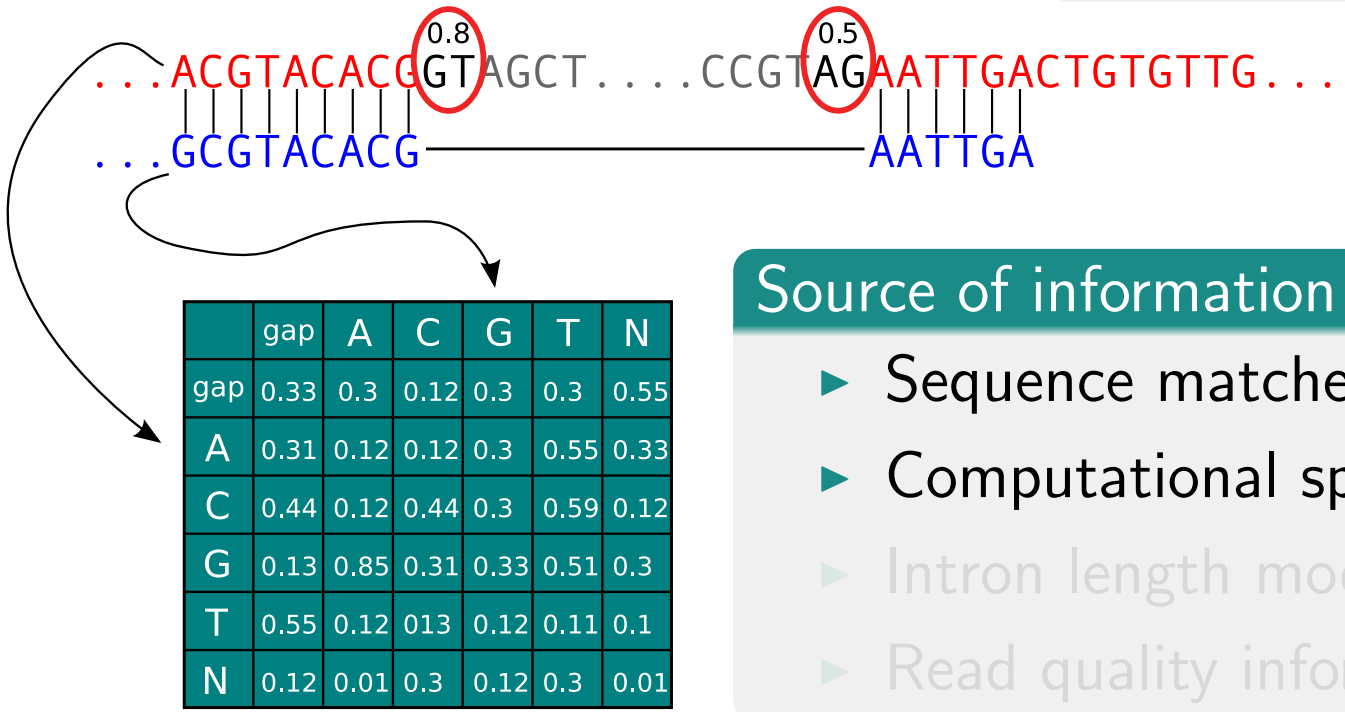
Source of information

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

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

Fast and Accurate spliced alignments with PALMapper





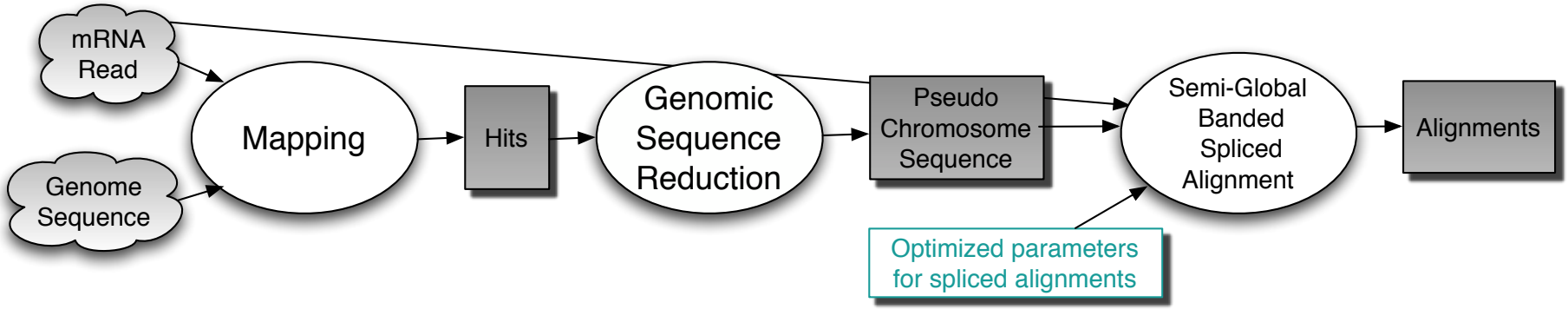
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

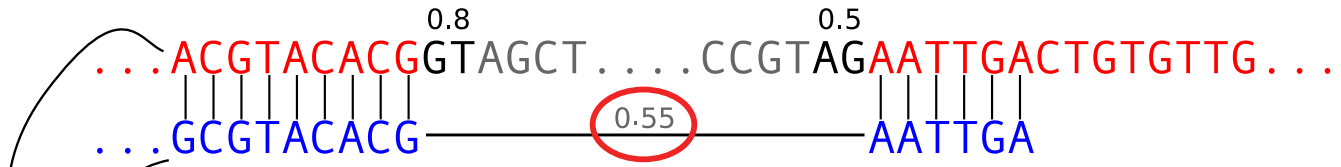
Source of information

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

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

Fast and Accurate spliced alignments with PALMapper





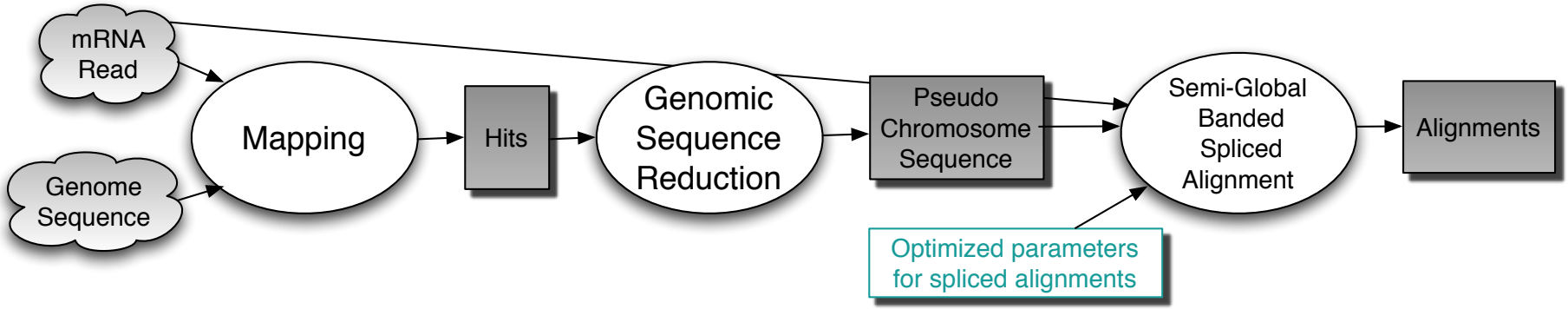
	gap	A	C	G	T	N
gap	0.33	0.3	0.12	0.3	0.3	0.55
A	0.31	0.12	0.12	0.3	0.55	0.33
C	0.44	0.12	0.44	0.3	0.59	0.12
G	0.13	0.85	0.31	0.33	0.51	0.3
T	0.55	0.12	013	0.12	0.11	0.1
N	0.12	0.01	0.3	0.12	0.3	0.01

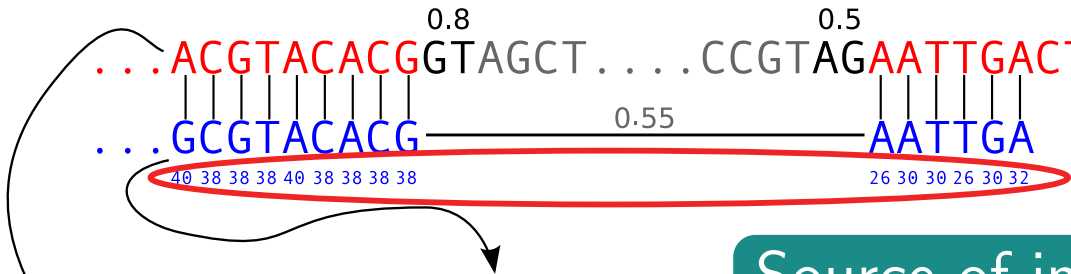
Source of information

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

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

Fast and Accurate spliced alignments with PALMapper





	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

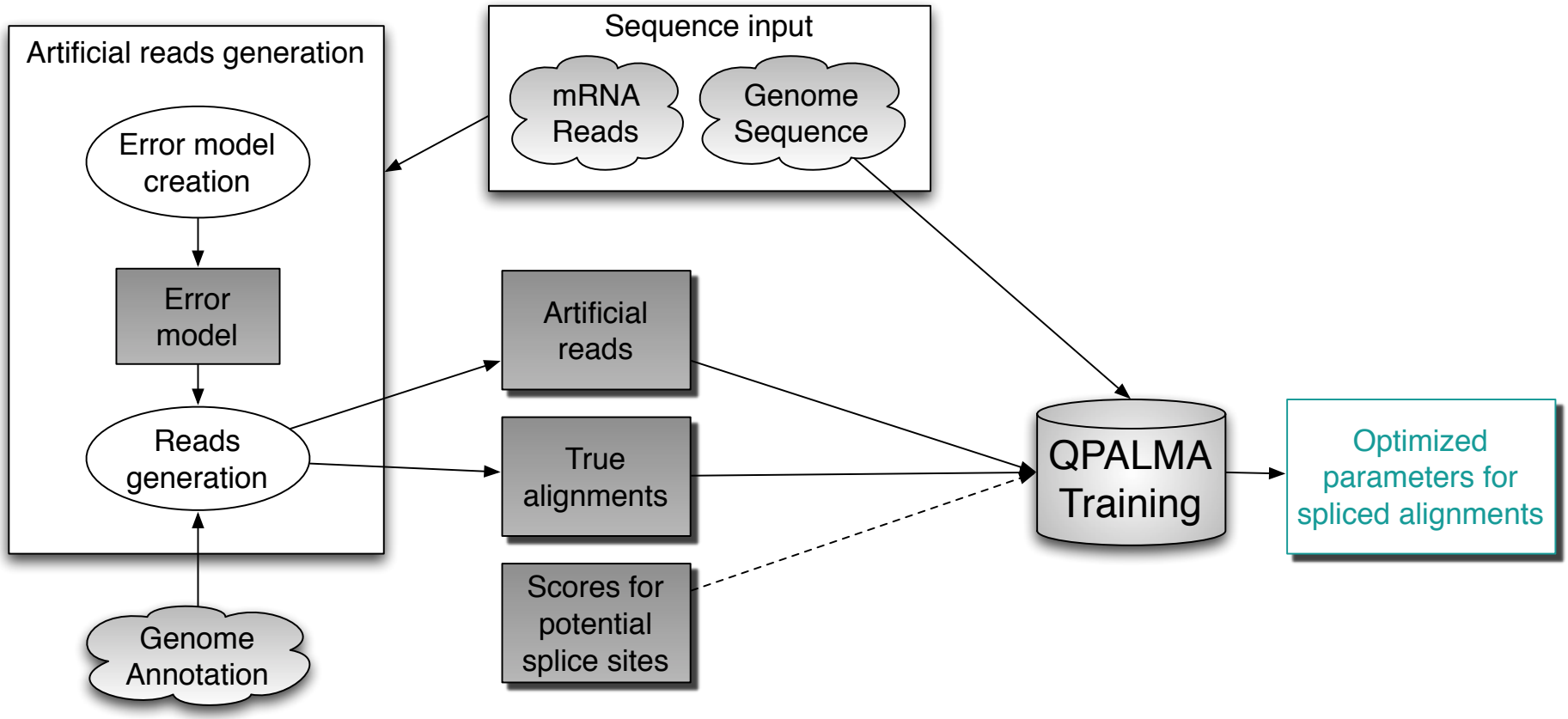
- Sequence matches
- Computational splice site predictions
- Intron length model
- Read quality information

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

[De Bona et al., 2008]

QPALMA extended scoring model

Estimation of QPALMA scoring model via a large margin approach similar to SVMs



RGASP.3 pipeline

For each dataset (artificial, mouse, cytosol, nucleus and whole_cell)

- Specific training of QPALMA
- Aligning reads with PALMapper
- Filtering alignments
- Evaluating

Filtering Strategy

Lessons learnt from RGASP2:

- Thorough filtering increases alignment accuracy
- Strategy adapted to the alignment algorithm

Simple alignment filter based on intrinsic alignment features as:

- Maximal number of edit operations
- Minimal segment length
- Minimal read support for splice junctions

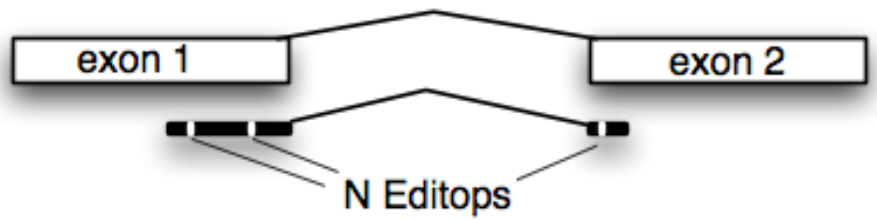
Filtering Strategy

Lessons learnt from RGASP2:

- Thorough filtering increases alignment accuracy
- Strategy adapted to the alignment algorithm

Simple alignment filter based on intrinsic alignment features as:

- Maximal number of edit operations
- Minimal segment length
- Minimal read support for splice junctions



Filtering Strategy

Lessons learnt from RGASP2:

- Thorough filtering increases alignment accuracy
- Strategy adapted to the alignment algorithm

Simple alignment filter based on intrinsic alignment features as:

- Maximal number of edit operations
- Minimal segment length
- Minimal read support for splice junctions



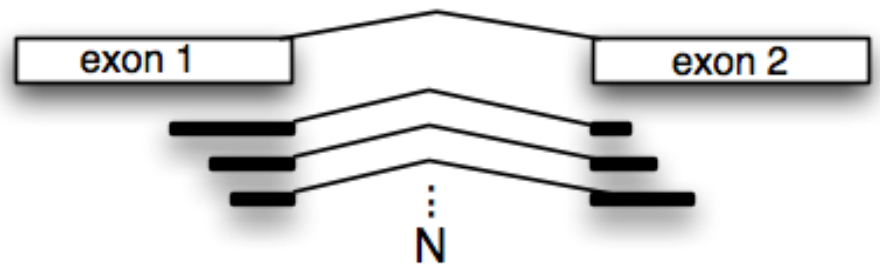
Filtering Strategy

Lessons learnt from RGASP2:

- Thorough filtering increases alignment accuracy
- Strategy adapted to the alignment algorithm

Simple alignment filter based on intrinsic alignment features as:

- Maximal number of edit operations
- Minimal segment length
- Minimal read support for splice junctions



Other Features

- Fully parallelized (used 48 core machines)
- Read trimming: 3' end and polyA-tail
- Handles strand-specific reads
- Built-in filtering
- Allows to report arbitrary many multi-mappers
- Supports non-canonical splice sites
- Can align over several introns

Conclusion

Availability

- Command-line
  `http://ftp.tuebingen.mpg.de/pub/fml/raetsch-lab/software/`
- Galaxy web-interface
  `http://galaxy.fml.mpg.de/`
- Tutorial paper: G. Jean et al., Current Protocols in Bioinformatics, 2010.

Funding by DFG & Max Planck Society.

Thank you for your attention.

References I

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

Géraldine Jean, André Kahles, Vipin T Sreedharan, Fabio De Bona, and Gunnar Rätsch. Rna-seq read alignments with palmapper. *Curr. Protoc. Bioinform.*, 32:11.6.1– 11.6.37, 2010.

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

Cole Trapnell, Lior Pachter, and Steven L. Salzberg. Tophat. *Bioinformatics*, 25:1105–1111, May 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp120. URL `http://portal.acm.org/citation.cfm?id=1536354.1536358`.