Fast and Accurate RNA-Seq alignments with *PALMapper*

Géraldine Jean[1], André Kahles[1], Soeren Sonnenburg[2], Fabio De Bona[1], Korbinian Schneeberger[3], Jörg Hagmann[3], Detlef Weigel[3], Gunnar Rätsch[1]

[1] Friedrich Miescher Laboratory of the Max Planck Society, Spemannstr. 39, 72070 Tübingen, Germany
[2] Machine Learning Group, Berlin Institute of Technology, Franklinstr. 28/29, 10587 Berlin, Germany
[3] Max Planck Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany

MAX-PLANCK-GESELLSCHAFT

fml

Abstract

Next Generation Sequencing (NGS) technologies have revolutionized genome and transcriptome sequencing. RNA-Seq experiments generate huge amounts of mRNA sequence reads which are relatively short, error prone and may span exon-exon junctions. *PALMapper* [1] is a RNA-seq read mapper combining GenomeMapper and an improved version of QPALMA:

- **Aligning spliced and unspliced RNA-seq reads**
- **Benefiting from read quality information and splice site predictions**
- **Not restricted to known splice sites**
- **Allowing non-consensus spliced alignments**
- **Offering a growing pool of features for more accurate alignments**

Information & Contact:
`http://www.fml.mpg.de/raetsch/suppl/palmapper`
palmapper@tuebingen.mpg.de

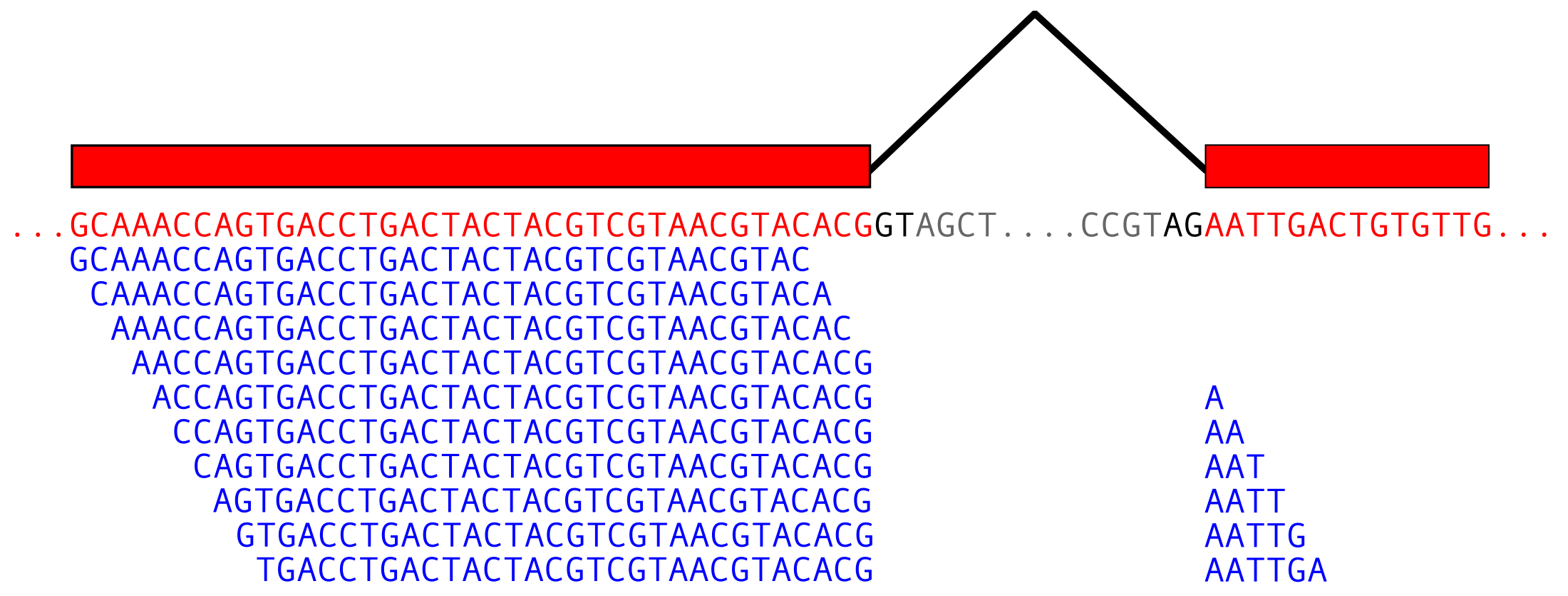
RNA-Seq and Spliced Alignments

RNA-seq produces millions of reads (n-mers typically of fixed size) with n quality values:

```
ACGTACACGCAGTAGTACGACGTGGGTAACGTGGTA
40 40 38 38 32 30 30 28 27 27 18 18 17 27 30 30 30 25 27 30 28 27 27 27 27 27 27 27 15 15 15 14 10 10 11 10
```

Base quality: related to probability for an erroneous base call

Aligning a transcriptome read to a genome sequence:


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

- **Unspliced read** falls exactly into one exon
- **Spliced read** is spread over two or more exons

References

[1] G. Jean, A. Kahles, V.T. Sreedharan, F. De Bona, G. Rätsch RNA-Seq Read Alignments with PALMapper *Curr. Protoc. Bioinform.* 32:11.6.1-11.6.38, 2010.

[2] F. De Bona and S. Ossowski and K. Schneeberger and G. Rätsch Optimal Spliced Alignments of Short Sequence Reads *ECCB08/Bioinformatics* 24 (16): i174, 2008.

[3] K. Schneeberger and J. Hagmann and S. Ossowski and N. Warthmann and S. Gesing and O. Kohlbacher and D. Weigel Simultaneous alignment of short reads against multiple genomes *Genome Biol.* 10 (9): R98, 2009.

[4] C. Trapnell and L. Pachter and S. L. Salzberg TopHat: discovering splice junctions with RNA-Seq *Bioinformatics* 25 (9) : 1105-11, 2009.

[5] T.D. Wu, and S. Nacu Fast and SNP-tolerant detection of complex variants and splicing in short reads *Bioinformatics* 26: 873-881, 2010.

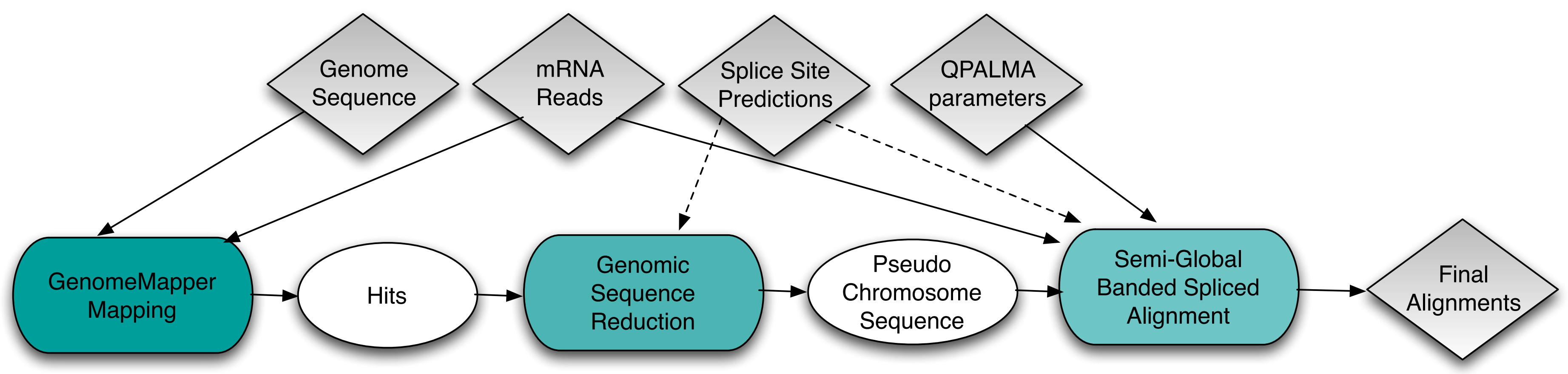
[6] B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W.J. Kent, and A. Nekrutenko Galaxy: a platform for interactive large-scale genome analysis *Genome Research* 15(10):1451-1455, 2005.

[7] G. Zeller, et al. mTiM: margin-based transcript mapping from RNA-seq *BMC Bioinformatics* in preparation, 2011.

[8] R. Bohnert, and G. Rätsch rQuant.web: a tool for RNA-Seq-based transcript quantitation *Nucleic Acids Res.* 38 (suppl 2): W348-W351, 2010.

[9] O. Stegle, et al. Statistical tests for detecting differential RNA-transcript expression from read counts *Nature Precedings* 2010.

PALMapper workflow



Globally aligning transcriptome reads against the whole genome is computationally too expensive. *PALMapper* [1]:

- uses efficient genome indexing to locate **unspliced read** or **parts from a plausible spliced read**,
- reduces the size of genome sequence to map against by identifying **mappable regions** (excluding plausible introns or intergenic regions),
- uses a seed position to guide a fast banded semi-global alignment of the **whole read** to a **portion of pseudo chromosome sequence**.

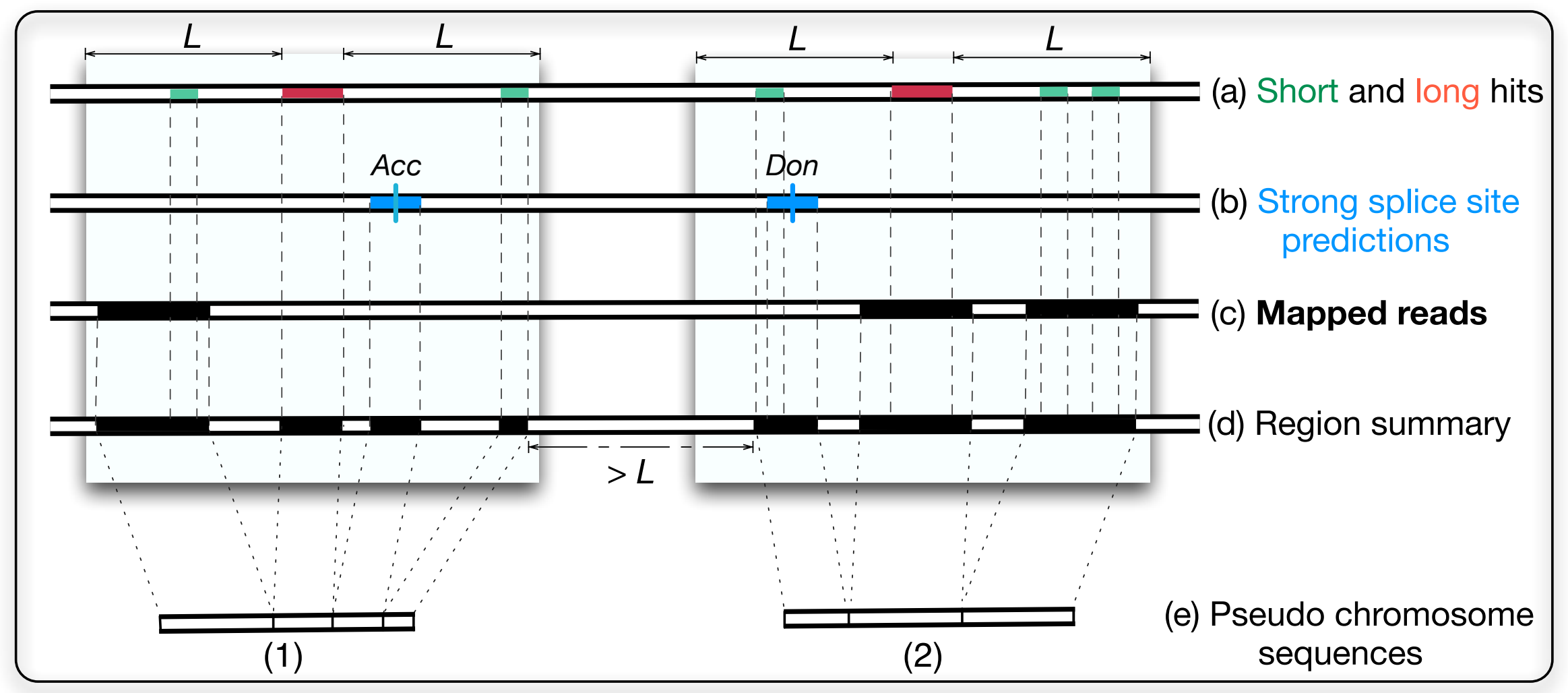
GenomeMapper Mapping

GenomeMapper [3] is a read mapper developed for the 1001 Plant Genomes Project:

- Indexing the genome with k-mer based index or bwt-based index,
- reporting all **extended hits** within the specified range of mismatches and gaps.

Genome Sequence Reduction

From the seed regions (long hits) found by GenomeMapper, plausible mapping regions are defined for a given read:

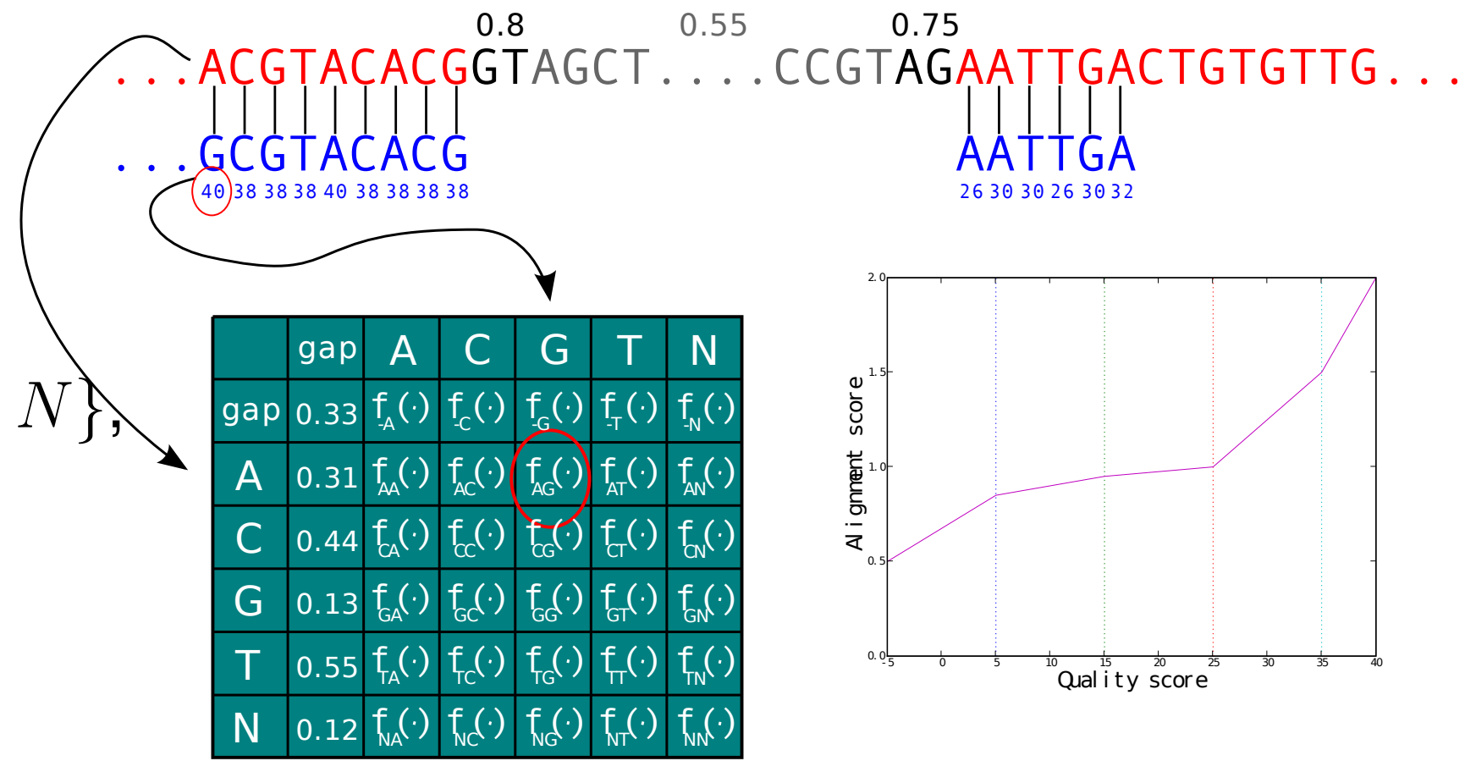


All regions at a distance smaller than the maximal intron size L are concatenated together to give **a pseudo chromosome sequence**.

QPalma Scoring Model

QPalma scoring model is defined by several functions scoring:

- Quality values: $M : \Sigma \times \mathbb{R} \times \Sigma \to \mathbb{R}$ with $\Sigma = \{-, A, C, G, T, N\}$,
- Accurate splice site scores, and
- Intron lengths.

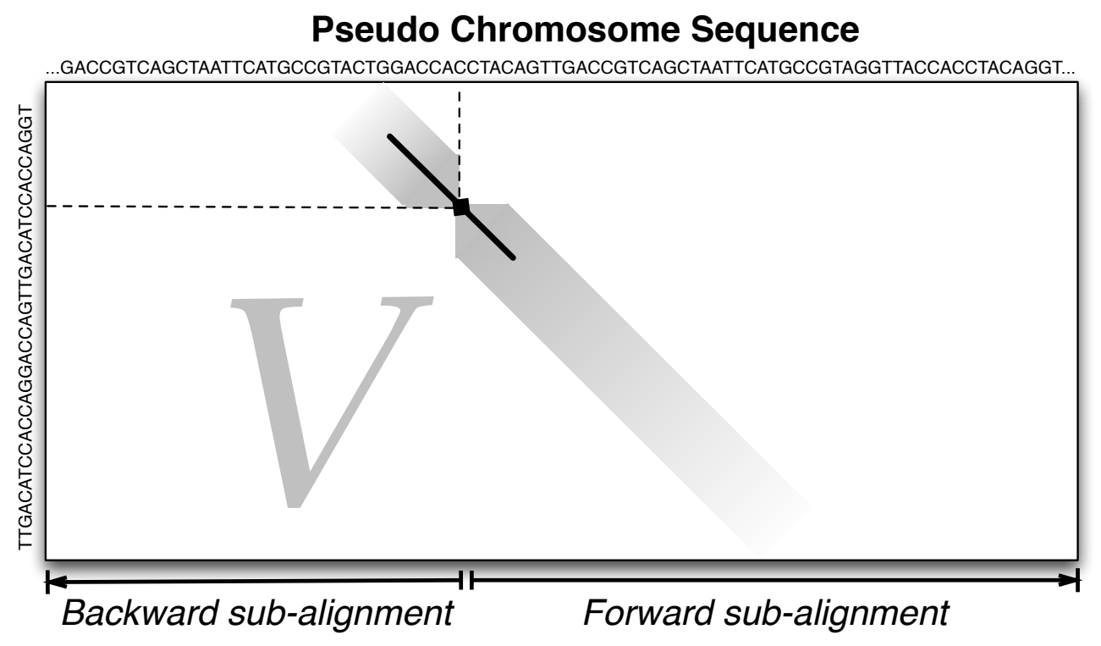


	gap	A	C	G	T	N
gap	0.33	$f_{-A}()$	$f_{-C}()$	$f_{-G}()$	$f_{-T}()$	$f_{-N}()$
A	0.31	$f_{AA}()$	$f_{AC}()$	$f_{AG}()$	$f_{AT}()$	$f_{AN}()$
C	0.44	$f_{CA}()$	$f_{CC}()$	$f_{CG}()$	$f_{CT}()$	$f_{CN}()$
G	0.13	$f_{GA}()$	$f_{GC}()$	$f_{GG}()$	$f_{GT}()$	$f_{GN}()$
T	0.55	$f_{TA}()$	$f_{TC}()$	$f_{TG}()$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

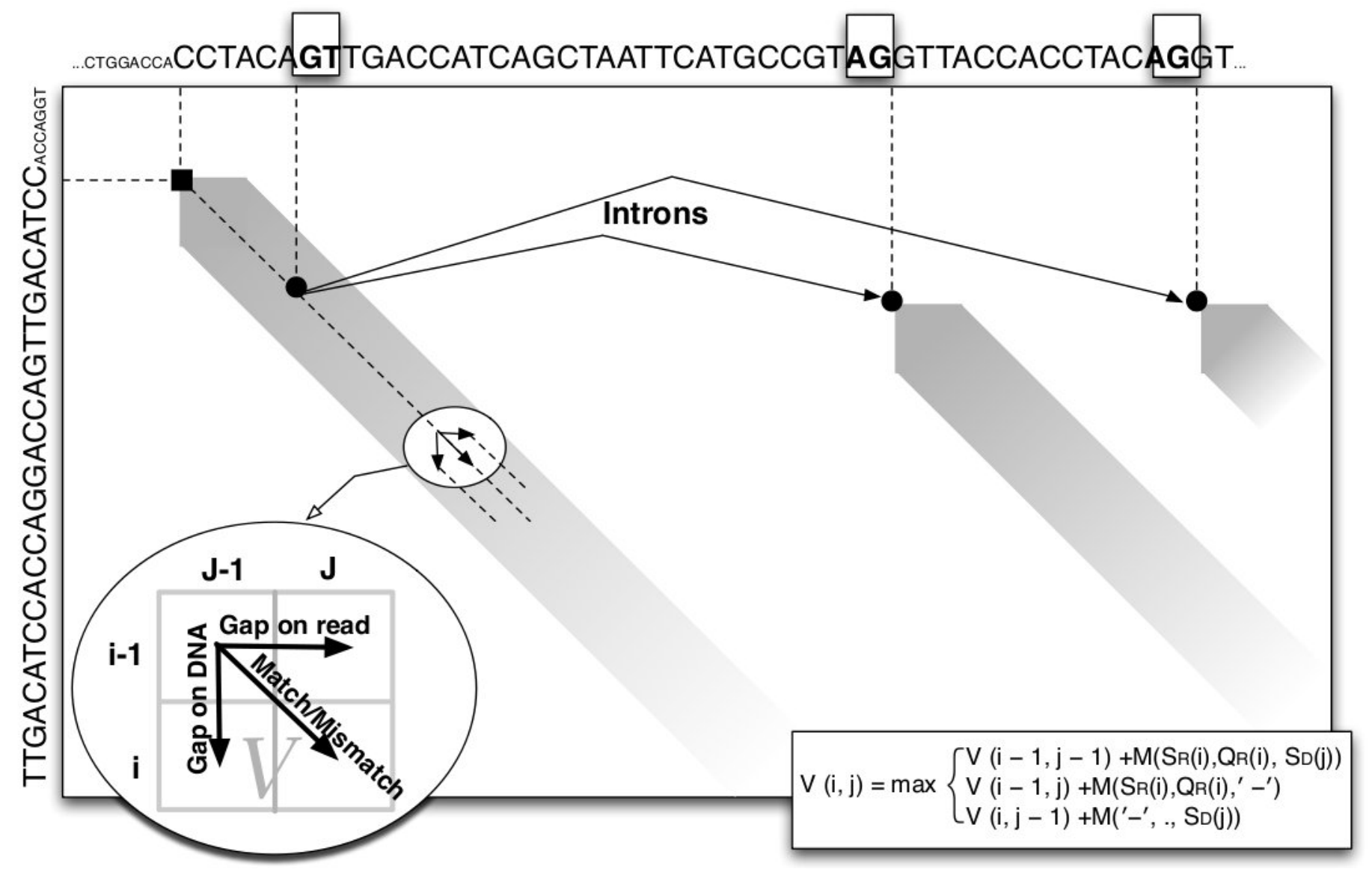
Semi-Global Alignment Algorithm

General Algorithm:

- **Seed position**: best match within the first seed region
- **2 sub-alignments in both directions** from the seed position
- **Final alignment**: merge of the best 2 sub-alignments


Pseudo Chromosome Sequence
...GACCGTCAGCTAATTCATGCCGTACTGGACCACCTACAGTTGACCGTCAGCTAATTCATGCCGTAGGTTACCACCTACAGGT...
Read
TTGACATCCACCAGGACCAGTTGACATCCACCAGGT
V
Backward sub-alignment
Forward sub-alignment

Forward sub-alignment algorithm:

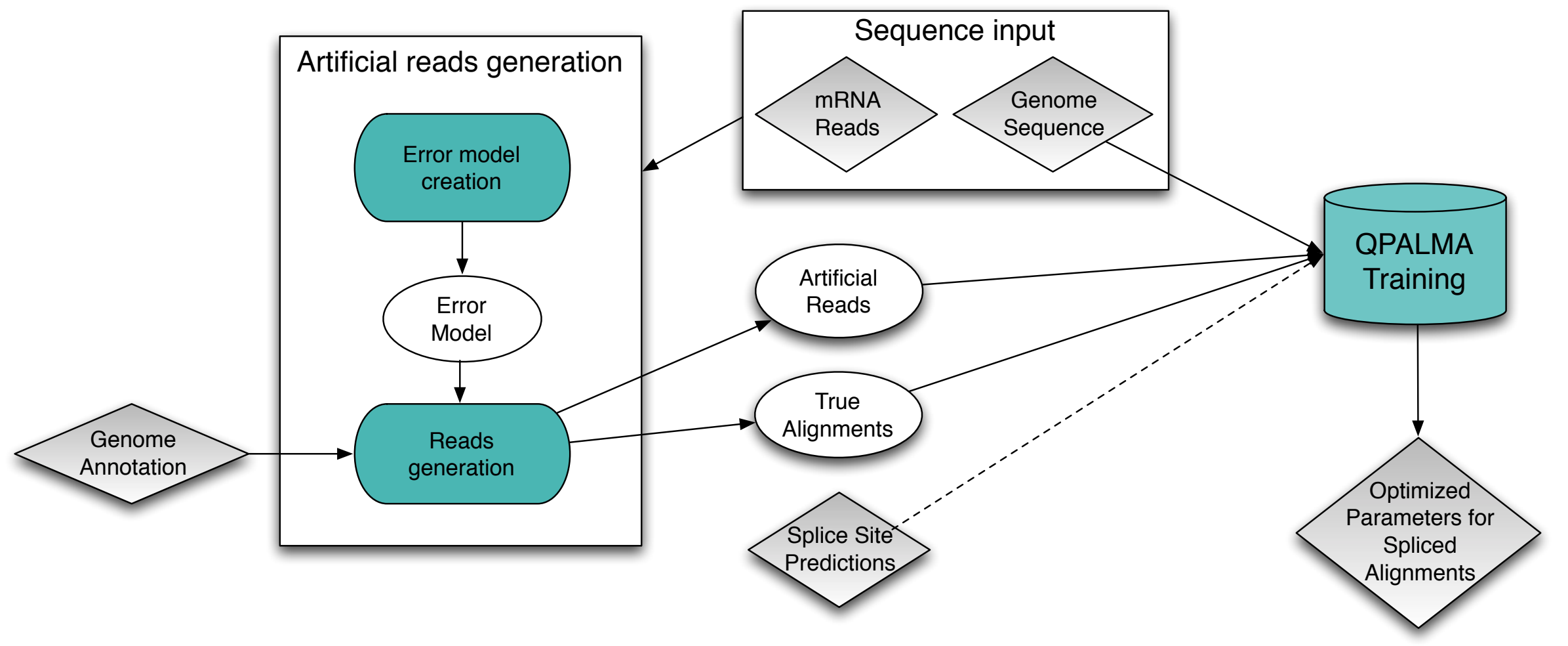

...CTGGACCACCTACAGTTGACCATCAGCTAATTCATGCCGTAGGTTACCACCTACAGGT...
TTGACATCCACCAGGACCAGTTGACATCCACCAGGT
Introns
J-1
J
i-1
i
Gap on read
Gap on DNA
Match/Mismatch
V

$$V(i,j) = \max \begin{cases} V(i-1, j-1) + M(S_R(i), Q_R(i), S_D(j)) \\ V(i-1, j) + M(S_R(i), Q_R(i), '-') \\ V(i, j-1) + M('-', ., S_D(j)) \end{cases}$$

- **Banded**: limits the number of gaps from the perfect alignment
- **Spliced**: Allows long gaps corresponding to introns via recursive calls of the sub-alignment algorithm from novel seed positions deduced from plausible splice site positions

QPalma Training

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



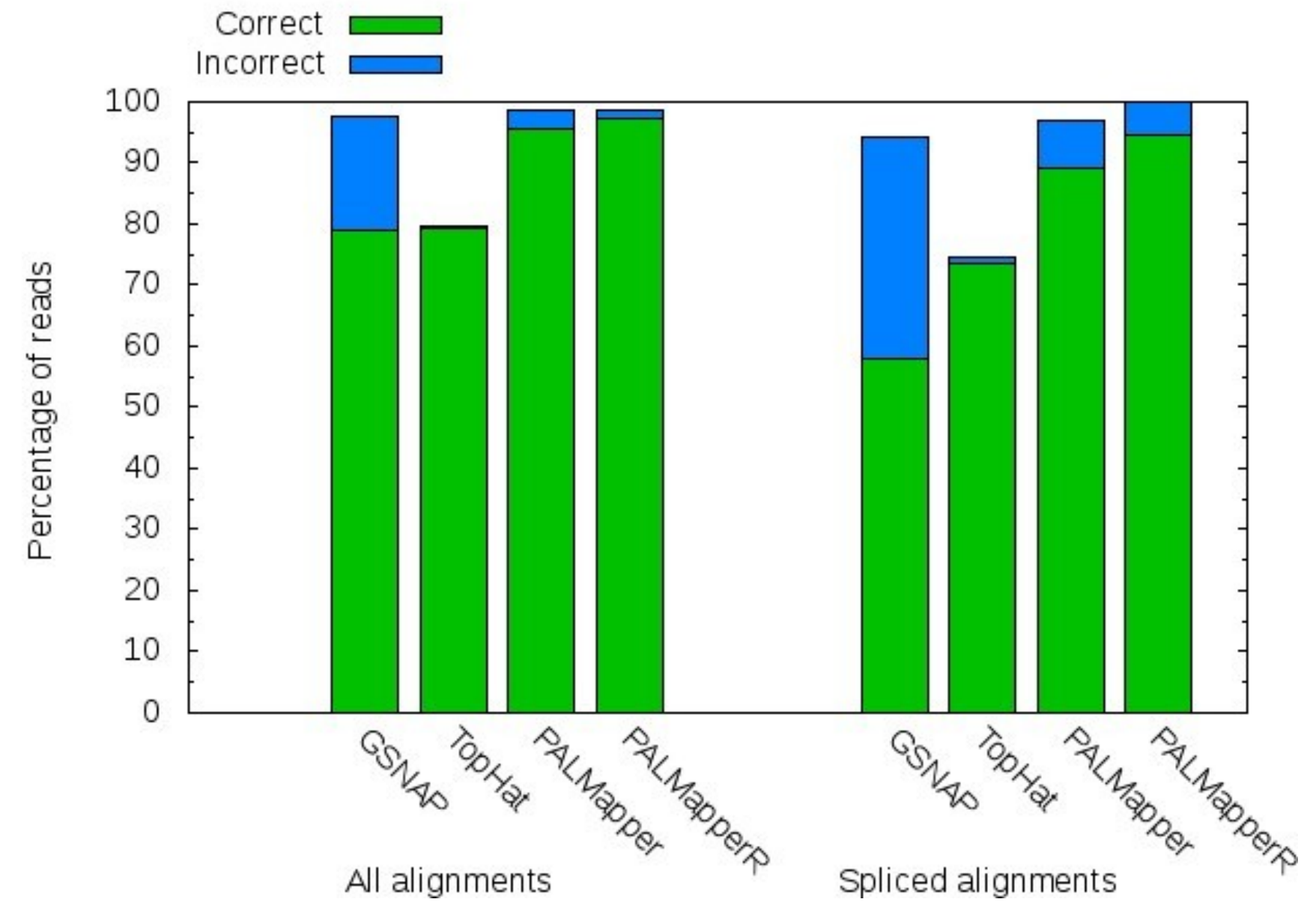
PALMapper Features

- Fully parallelized
- Read trimming: 3' end and polyA-tails
- Handles strand-specific reads
- Allows mismatches and indels
- Built-in filtering
- Able to report sub-optimal alignments
- Supports non-canonical splice sites
- Can align over several introns
- Built-in intron junction library allowing a remapping strategy

Results

Comparison of *PALMapper* with TopHat [4] (v1.0.12) and GSNAP [5] (2010-07-27)

- Simulated RNA-seq reads from *C. elegans*
- 30,439,758 reads of which 8,437,297 are spliced
- Evaluation of alignments according to true alignments



PALMapperR results are obtained by running *PALMapper* with the remapping of reads against the intron junction database obtained from a first round.

Availability

- Galaxy web-interface:
  `http://galaxy.fml.mpg.de/`
- Open-source packaged releases for Unix or Mac OS X:
  `http://fml.mpg.de/raetsch/suppl/palmapper/`

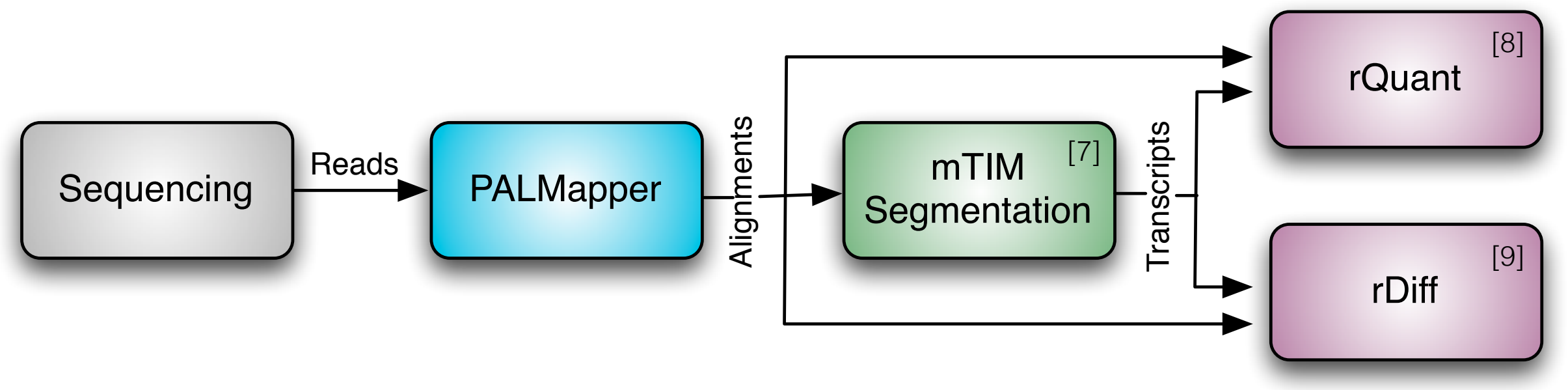
Galaxy - Mozilla Firefox

Galaxy / Rätsch Lab

oqtans
online quantitative transcript analysis

- Included in our Galaxy-integrated workflow for Quantitative Transcriptome Analysis from NGS data

`http://oqtans.org`


Sequencing
Reads
PALMapper
Alignments
mTIM Segmentation [7]
Transcripts
rQuant [8]
rDiff [9]