CIS529: Bioinformatics

Spliced Read Alignment

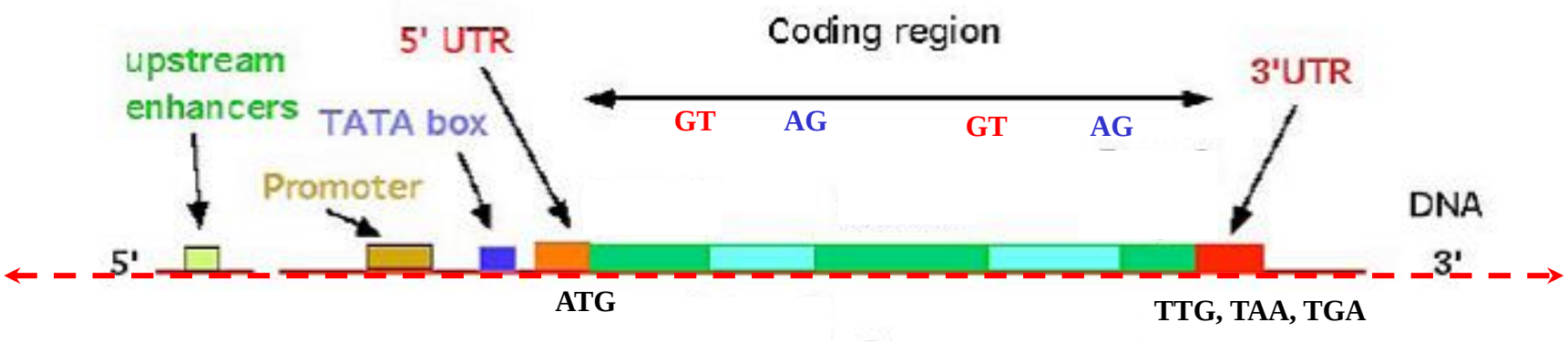
Presented by

Dr. Fayyaz-ul-Amir Afsar Minhas

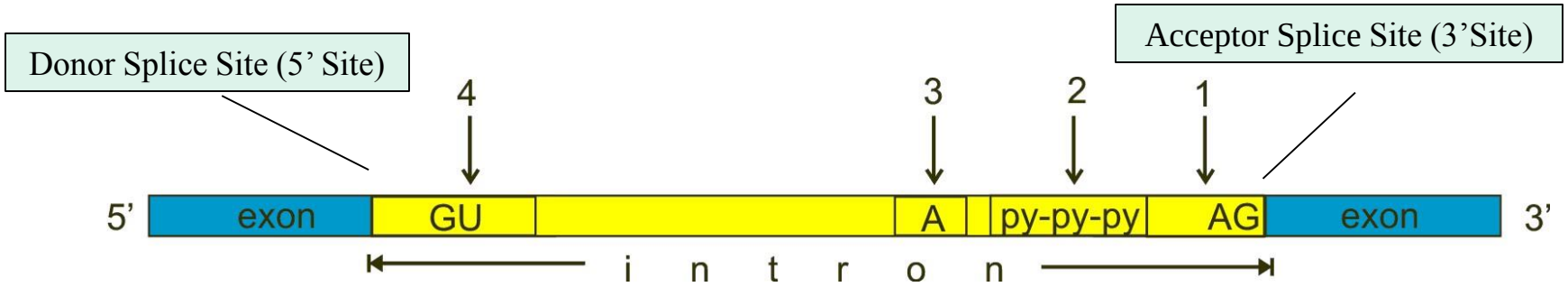
http://faculty.pieas.edu.pk/fayyaz

Department of Computer & Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad 45650
Pakistan

Biological Foundations: Transcription



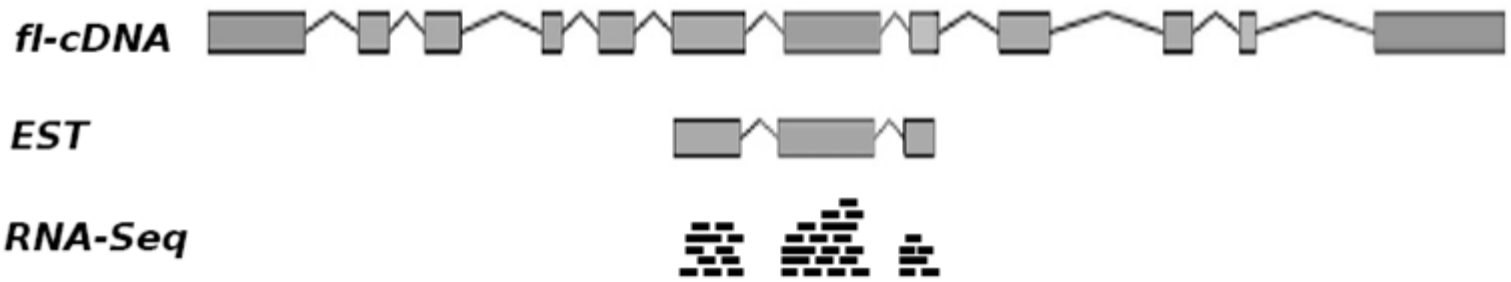
More Terminology: Splice Sites



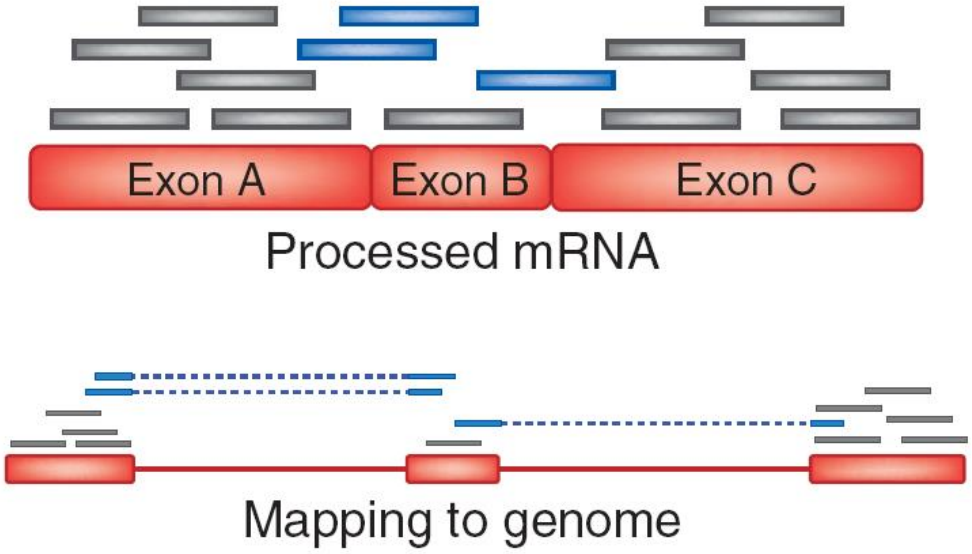
- **Canonical Splice Site Consensus: GT---AG**
- **GT Occurs about 1 billion times in the 3 Billion nucleotide long human DNA**
- **Acceptor Consensus**
  - $A_{64}G_{73}G_{100}T_{100}A_{62}A_{68}G_{84}T_{63}$
- **Donor Consensus**
  - $C_{65}A_{100}G_{100}$

RNA-Seq

- The mRNA is converted to cDNA for sequencing
- A small part of the cDNA is sequenced and is called an Expressed Sequence Tag (EST)
- NGS gives a large number of very small reads which can then be aligned to the genome



Two types of reads



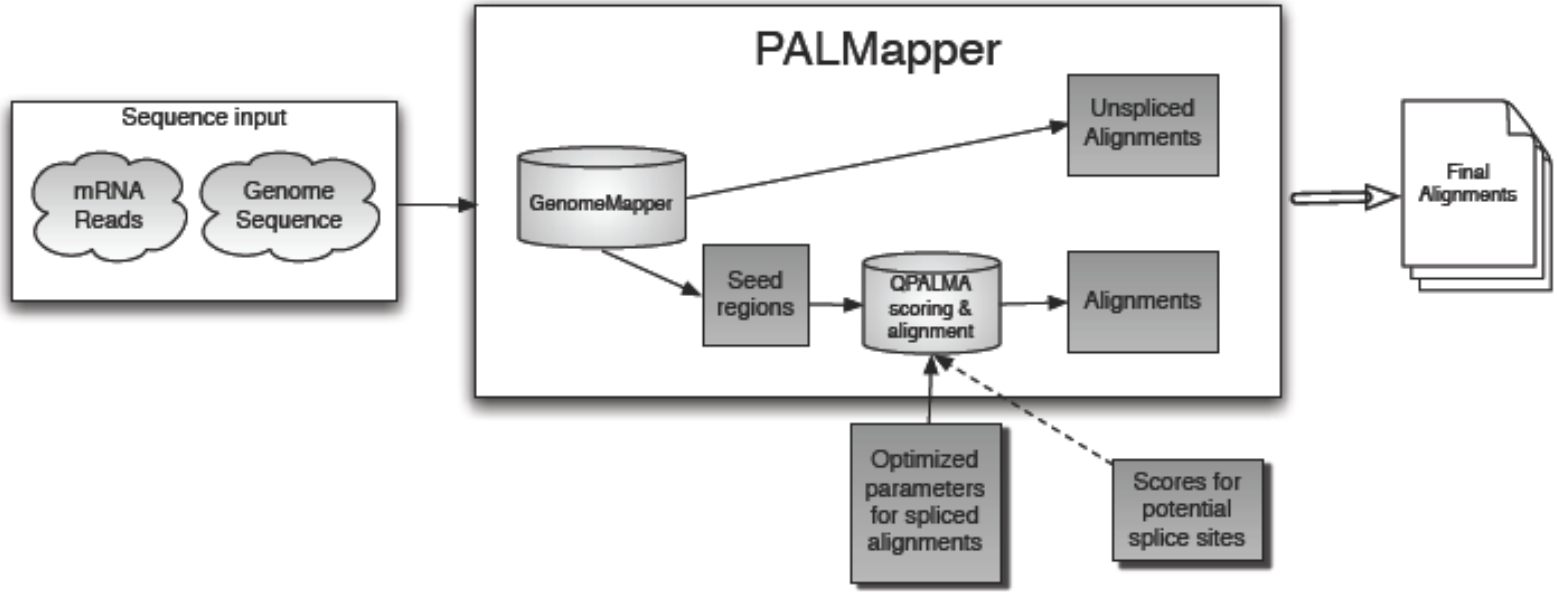
- Unspliced Read (Black)
  - Read lies within exactly one exon
- Spliced Read (Blue)
  - Read overlaps between two or more exons

Tools for Spliced Alignment

- QPALMA
- PALMAPPER
- Map Splice

PALMapper

- Tool for Performing Spliced Alignments of RNA-Seq Reads
  - GenomeMapper identies unspliced alignments and seed regions for spliced reads
  - QPALMA infers spliced alignments from seed regions



QPALMA: Optimal Spliced Alignments of Reads

- **Q = Quality**
  - **Uses a read's quality information**
    - **Raw Illumina Genome Analyzer data comes with per-base intensities for each position**

```
ACGTACGTACGTACACAAGTAGACGAACGTAGCTAA
40 40 40 40 38 38 36 36 36 32 32 32 30 25 18 18 40 40 40 40 38 38 36 36 36 32 32 32 30 25 18 18 25 18 18
```

- **PALMA**
  - **Uses Computational Splice Site Prediction**
- **Information Used in QPALMA**
  - **Quality Information**
  - **Splice Site Predictions**
  - **Intron Length Models**

Splice Site Prediction

- Two SVMs
  - Donor Splice Site vs. Not
  - Acceptor Splice Site vs. Not

- Training
  - Known Acceptor & Donor Sites
    - Obtained by aligning EST and cDNA sequences using BLAT (or any program for spliced alignments for long sequences)
    - High quality alignments can be used as positive examples
    - Other decoy sites used as negative examples
    - Window of length 141 nt around the splice site
    - Weighted degree kernel

Relationship between PALMapper & Smith Waterman Algorithm

- PALMapper is an "Intelligent" variant of SW that changes the scoring matrix in certain ways to incorporate
  - Sequence Quality Scores
  - Donor & Acceptor Splice Site Prediction Scores
  - Intron Length Model
- To perform spliced alignments of short read sequences against a genome

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

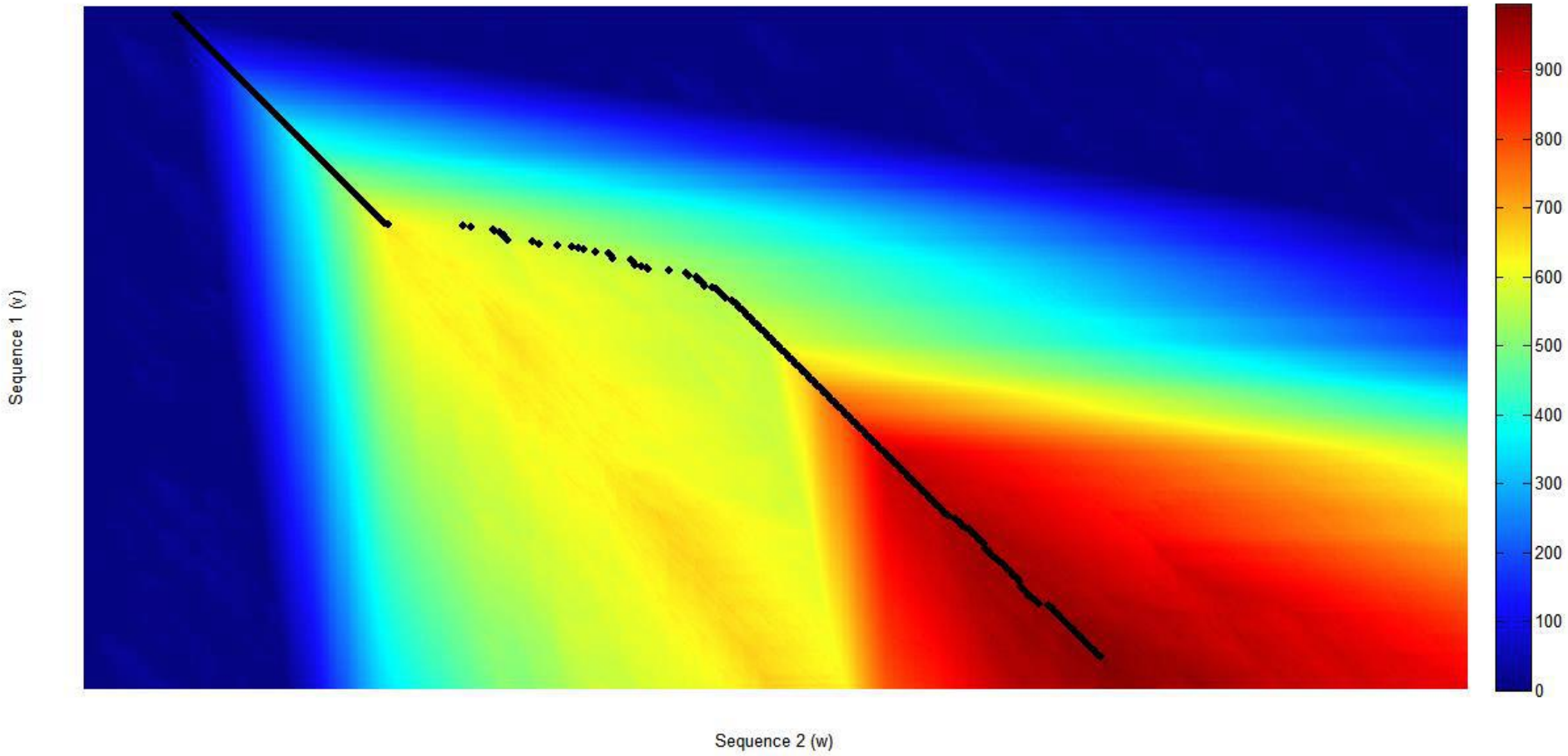
- Sequence matches

$$V(i,j)=\max\begin{cases}0\\ V(i-1,j-1)+M(S_E(i),S_D(j))\\ V(i-1,j)+M(S_E(i),'-')\\ V(i,j-1)+M('-',S_D(j))\end{cases}$$

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

score of alignment between $S_E(1:i)$ and $S_D(1:j)$ based upon a 6x6 substitution matrix

Smith Waterman Algorithm



Smith Waterman Algorithm Output

```
001  HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRP
     ||||||||||||:||||||||||||||||||||||||||||||||||||||||||||||||||||
001  HSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSNGCVSKILGRYYETGSIRP

065  RAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLAS
     ||||||||||||:||||||:|||||||||||||||||:|:|||||||||||||||||||||||:
065  RAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQENVCTNDNIPSVSSINRVLRNLAA

129  EK--Q-----------------------------------------------Q----M--------
     :|  |                                                |    :
129  QKEQQSTGSGSSSTSAGNSISAKVSVSIGGNVSNVASGSRGTLSSSTDLMQTATPLNSSESGGA

134  -----GA--DGMYDK--------------L---R-----------M--------L---N--G--
          |:  :::|:|               |   |           :         |   |  :
193  SNSGEGSEQEAIYEKLRLLNTQHAAGPGPLEPARAAPLVGQSPNHLGTRSSHPQLVHGNHQALQ

148  ----Q-------TGSW----------GTRP---G--WY------------P---------GTS-
         |       :|||          :::|   :  :|            |         |::
257  QHQQQSWPPRHYSGSWYPTSLSEIPISSAPNIASVTAYASGPSLAHSLSPPNDIESLASIGHQR

164  ---VPGQPT----Q-DGCQQQE---GGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQE
        |:::::    : ||:|::|   |:|||:|:::||::::::|:||:|||||||||||||::
321  NCPVATEDIHLKKELDGHQSDETGSGEGENSNGGASNIGNTEDDQARLILKRKLQRNRTSFTND

217  QIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRRQASNTPSH
     ||::||||||||||||||||||:||:||||||||||||||||||||||||||||:::|:::
385  QIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQVWFSNRRAKWRREEKLRNQRRTPNSTGAS

281  IPISSSFSTSVYQPIPQPTTPVSSFTSGSMLG---RTDTAL-T-NTYSALPPMPSFTMANNLP-
     :::||:::|:::::::|:::::::||::|||::|   :|:::| : :|:|:::::|  |::::::
449  ATSSSTSATASLTDSPNSLSACSSLLSGSAGGPSVSTINGLSSPSTLSTNVNAP--TLGAGIDS

339  MQPPVPSQTSSYSCMLPTSPSVNGR-SYD-----TYTPPHMQTHMNSQPMGTSGTTSTGLISPG
     :::|:|:::::::||   ||:::||| |:|     :::|:::::|:|::::::::|::::::::|:
511  SESPTPIPHIRPSC---TSDNDNGRQSEDCRRVCSPCPLGVGGHQNTHHIQSNGHAQGHALVPA

397  VS
     :|
572  IS
```

Extended Smith Waterman Algorithm

- **Extension-1 (Quality Scores)**

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

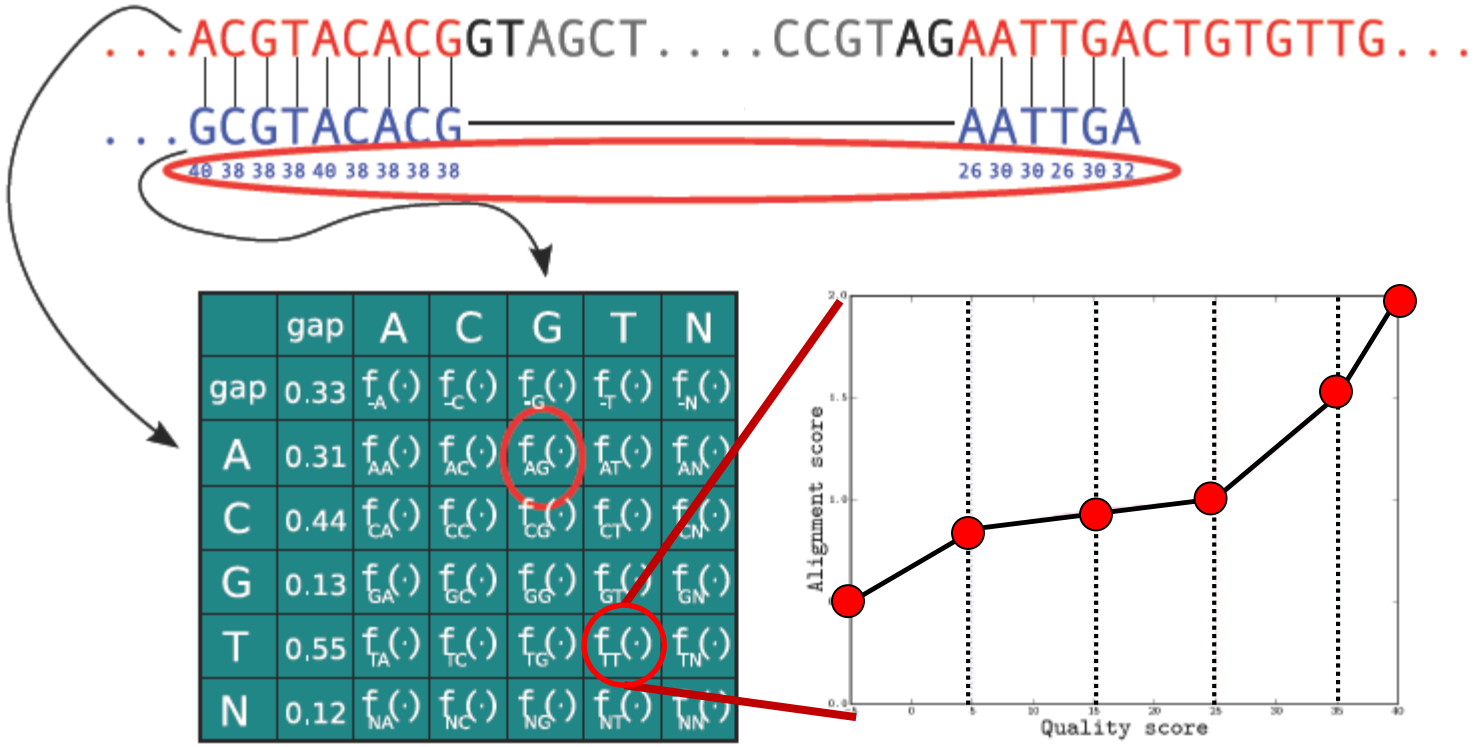
Source of Information

- ✖ 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

- **Extension-1 (Quality scores)**

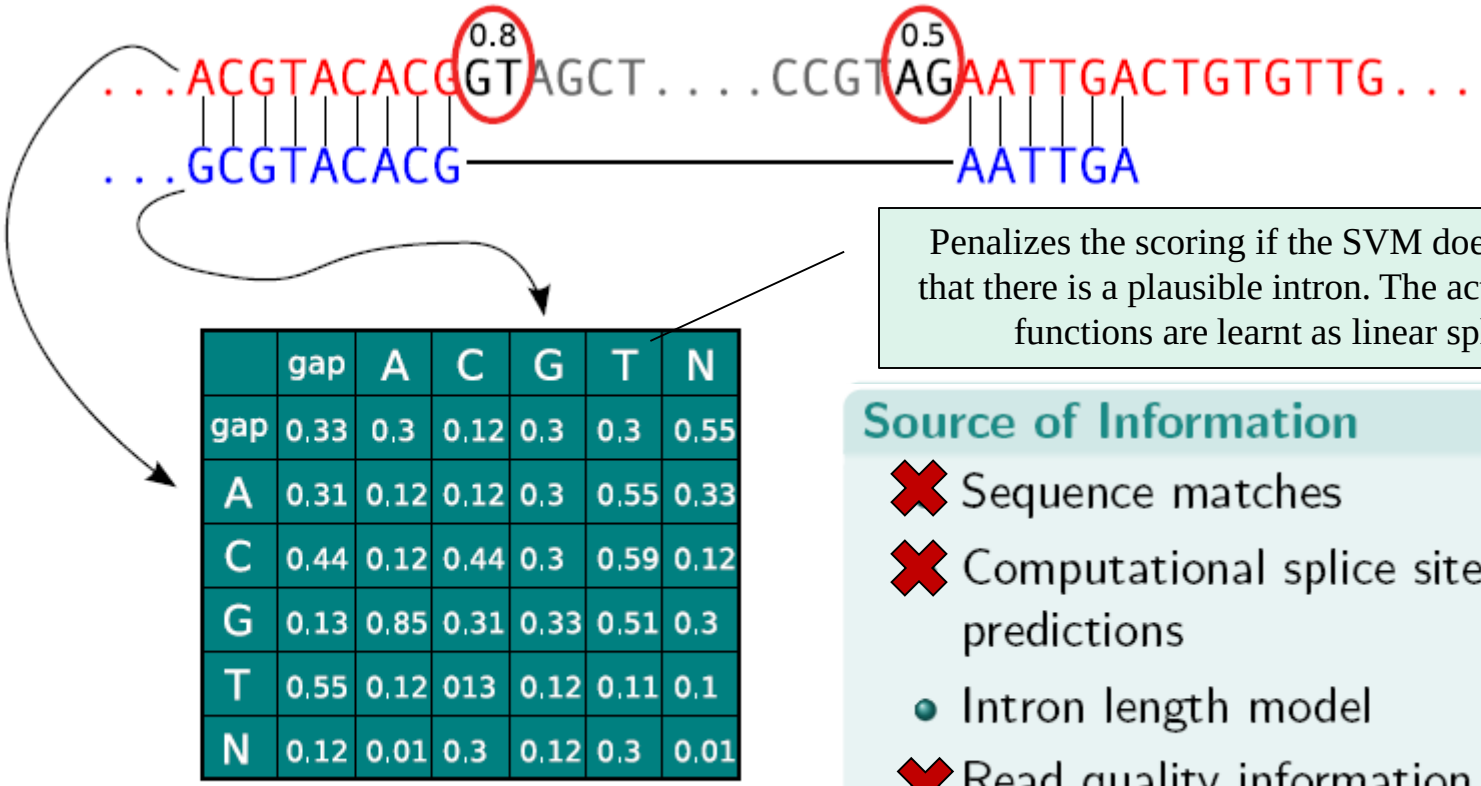


Quality scoring $f : (\Sigma \times \mathbb{R}) \times \Sigma \rightarrow \mathbb{R}$

Regularized linear spline!

Extended Smith Waterman Algorithm

- **Extension-2 (SVM Scores)**



	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

Penalizes the scoring if the SVM does not think that there is a plausible intron. The actual scoring functions are learnt as linear splines.

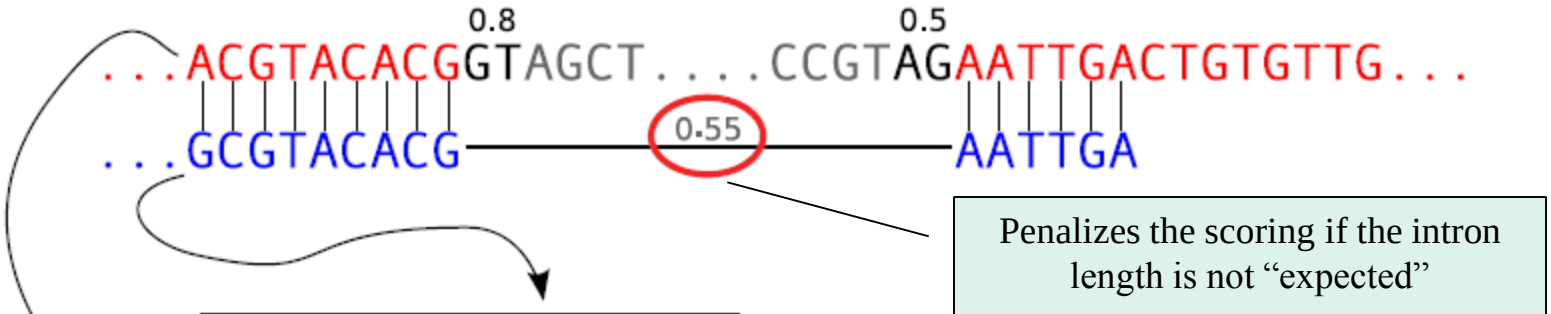
Source of Information

- ✗ 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

- **Extension-3 (Intron length model)**



Penalizes the scoring if the intron length is not "expected"

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

Max-Margin Optimization

- The parameters Θ of the functions are optimized during training
- Training
  - Given
    - N true alignments A_i^+, i=1…N
    - Wrong alignments A_i^-
- Learning problem

Regularization term for the linear splines

Exponential number of constraints! (Working set type solution!)

$$\min_{\xi \geq \mathbf{0}, \theta} \quad \frac{1}{N} \sum_{i=1}^{N} \xi_i + C\mathbf{P}(\boldsymbol{\theta})$$

$$\text{s.t.} \quad f_\theta(\mathcal{A}_i^+) - f_\theta(\mathcal{A}^-) \geq 1 - \xi_i \quad \forall i \text{ and } \mathcal{A}^- \neq \mathcal{A}_i^+.$$

Training

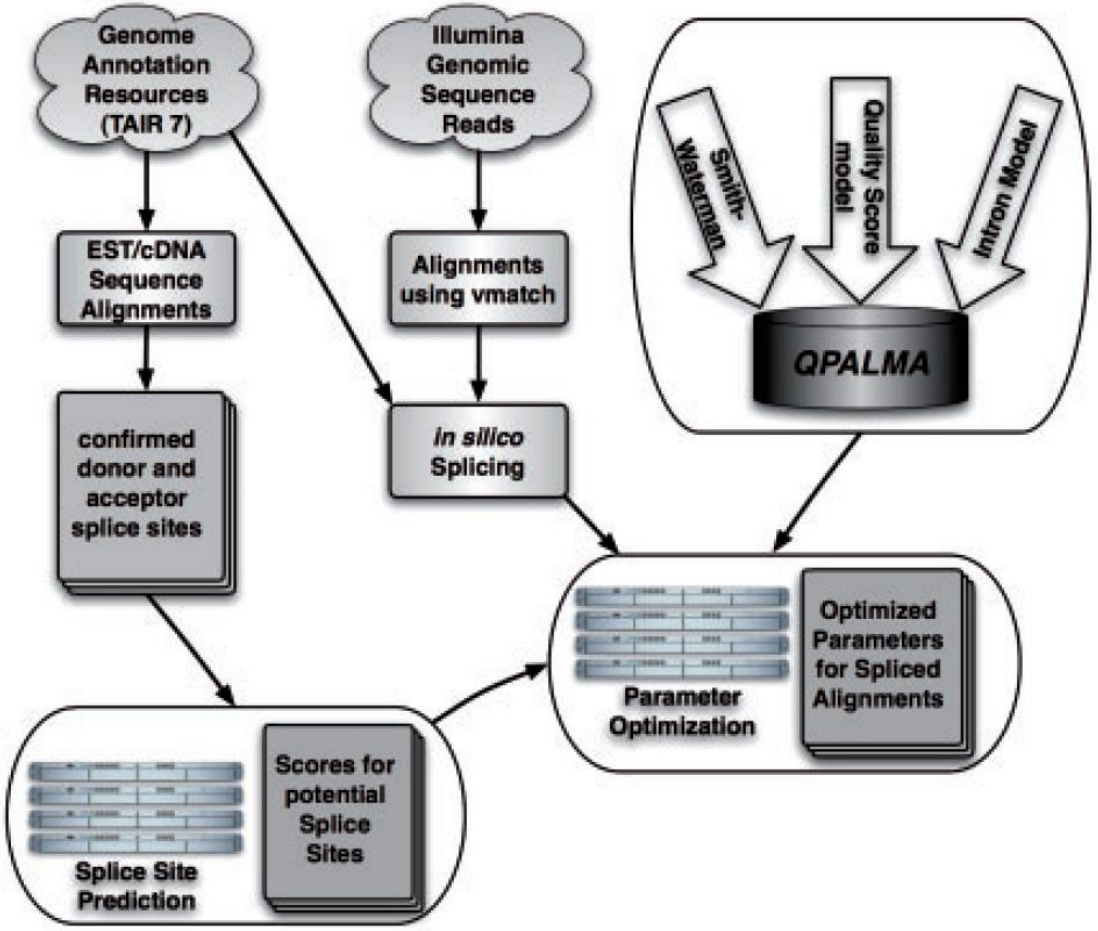


Fig. 1. Work-flow for training *QPALMA*: Confirmed donor and acceptor sites are used to train a SVM-based splice site predictor (Sonnenburg *et al.*, 2007). Short sequence reads (36 nt) obtained from an Illumina 1G sequencing machine and the *A.thaliana* genome annotation (TAIR 7) were used to derive a training and evaluation set. *QPALMA* generalizes the Smith–Waterman algorithm by including sequence quality information and an intron model considering splice site predictions as well as intron length information to learn how to produce optimally spliced alignments.

Learnt Match/Mismatch/IndelScoring

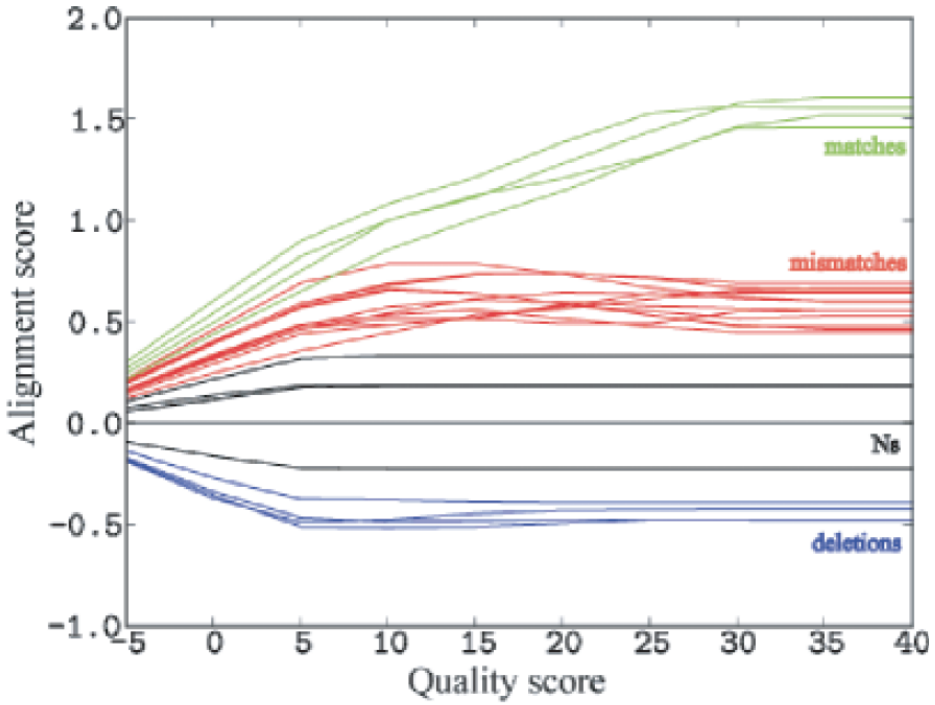


Fig. 5. Learned piece-wise linear functions to score matches (4 × green), mismatches (12 × red), N's (5 × black) and deletions (5 × blue). Larger quality scores for matches contribute stronger to the alignment score. Mismatches with larger quality score receive a lower alignment score than with medium sized quality scores.

Learnt Functions for Scoring SVM Probabilities

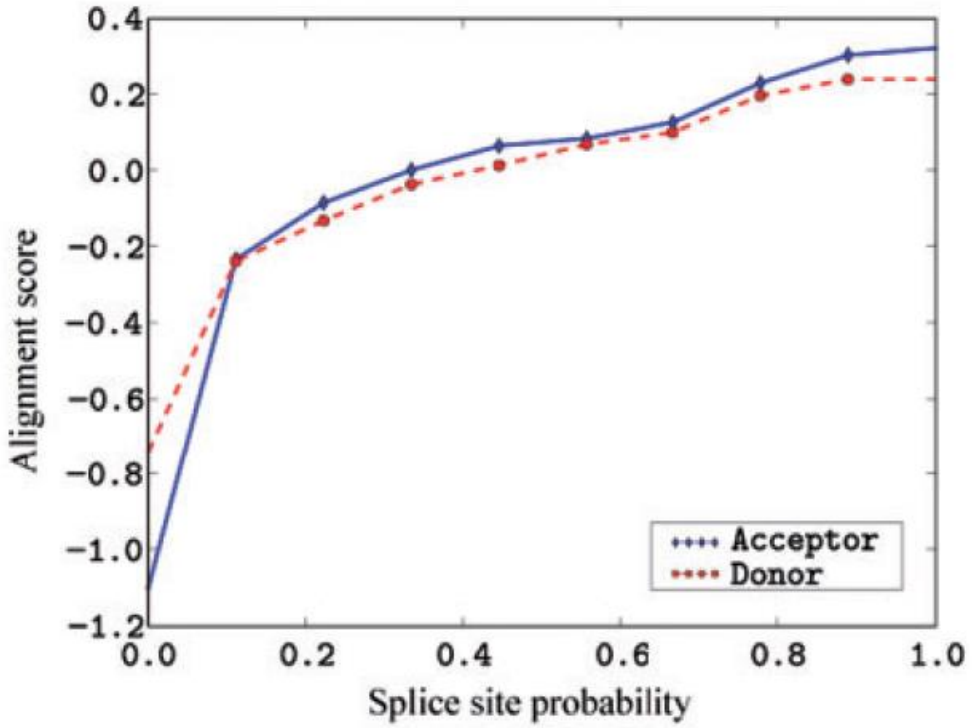


Fig. 4. Learned piece-wise linear functions to score acceptor and donor splice sites: High probabilities for splice sites contribute strongly to the alignment score.

Comparison with SW Extensions

- **(% of correct alignments)**

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

More Results

- Data
  - A. Thaliana
  - Chromosomes I-V
  - 2.98 M Reads
  - 10% spliced (excluding the ones used in training) = 285,530
- Spliced Reads
  - With significant overlap (>4nt) into the exon
    - 0.5 %
  - With 1-2nt into the next or previous exon
    - 12%

Summary

- As read length becomes larger more reads require spliced alignment

- There are many tools for this task:

❖ TopHat (uses Bowtie for ungapped mapping)
❖ MapSplice (uses Bowtie)
❖ RUM (uses Bowtie and BLAT)
❖ PASS

Comparison

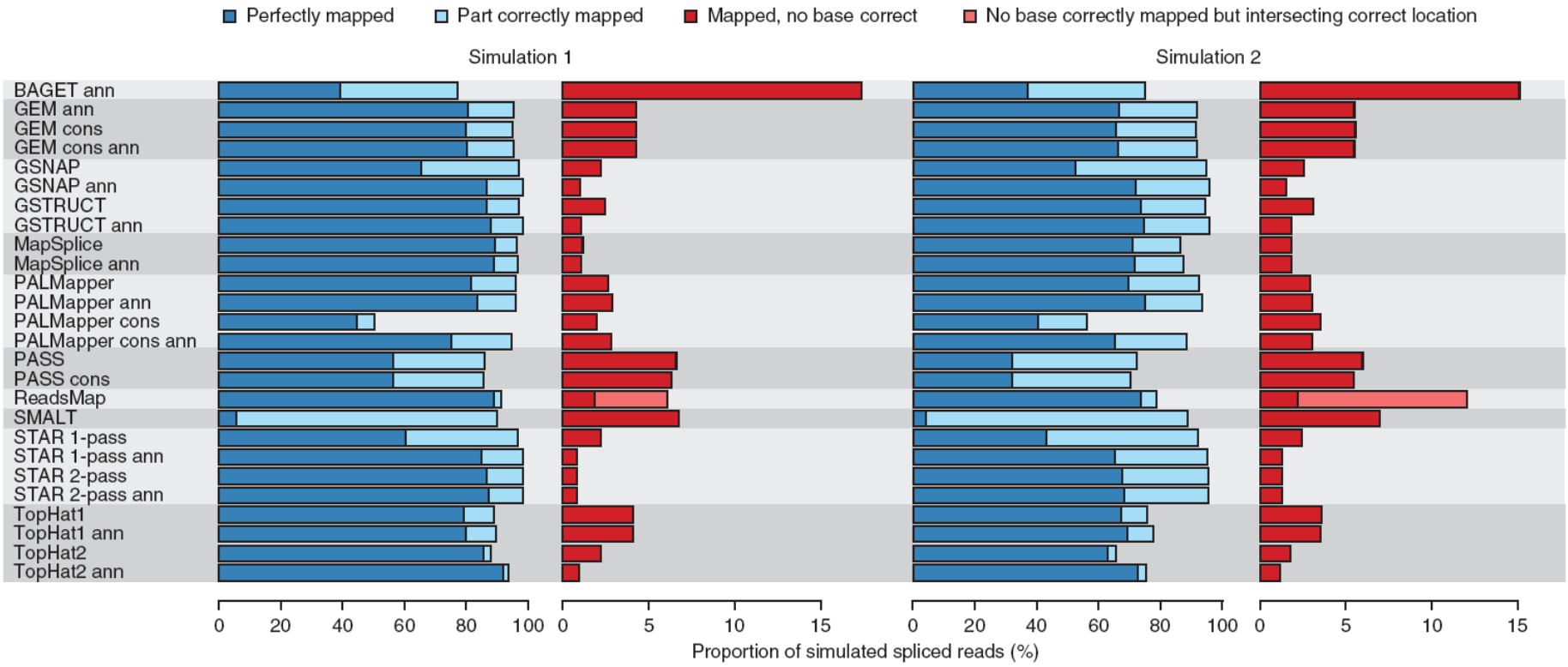


Figure 3 | Read placement accuracy for simulated spliced reads.

Figure from:
Systematic evaluation of spliced alignment programs for RNA-seq data
Nature Methods (2013)
http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.2722.html

References

- [1] G. Jean, A. Kahles, V. T. Sreedharan, F. De Bona, and G. Rätsch, "RNA-Seq read alignments with PALMapper," *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al*, vol. Chapter 11, p. Unit 11.6, Dec. 2010.
- [2] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch, "Accurate splice site prediction using support vector machines," *BMC Bioinformatics*, vol. 8, no. Suppl 10, p. S7, 2007.
- [3] F. D. Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, "Optimal spliced alignments of short sequence reads," *Bioinformatics*, vol. 24, no. 16, pp. i174–i180, Aug. 2008.
- [4] U. Schulze, B. Hepp, C. S. Ong, and G. Rätsch, "PALMA: mRNA to genome alignments using large margin algorithms," *Bioinforma. Oxf. Engl.*, vol. 23, no. 15, pp. 1892–1900, Aug. 2007.
- [5] P. G. Engström, T. Steijger, B. Sipos, G. R. Grant, A. Kahles, The RGASP Consortium, G. Rätsch, N. Goldman, T. J. Hubbard, J. Harrow, R. Guigó, and P. Bertone, "Systematic evaluation of spliced alignment programs for RNA-seq data," *Nat. Methods*, vol. advance online publication, Nov. 2013.

That'

at'x a

's ail

all fo

l folk

olks!