Evaluation of RGASP.2 Read Alignment Submissions

Andre Kahles, Regina Bohnert, Paolo Ribeca, Jonas Behr, and Gunnar Rätsch



Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

April 6, 2011

RGASP.3 Workshop, Barcelona

RGASP.2 Read Alignment Submissions



Submitters: $*$ = late

ADobin*	Alexander Dobin et al., USA
	STAR
AMortazavi*	Ali Mortazavi et al., USA
	Erange (woldlab.caltech.edu/rnaseq)
CIseli	Christian Iseli et al., CH
	SIBsim4 (sibsim4.sourceforge.net)
GRaetsch	Gunnar Rätsch et al., DE
	PALMapper (Jean et al., 2010)
LPachter	Lior Pachter et al., USA
	TopHat (Trapnell et al., 2009)
MGerstein	Mark Gerstein et al., USA
	TopHat (human only)
MStanke	Mario Stanke, DE
	Blat (Kent, 2002)
SWhite	Simon White et al., UK
	Exonerate (www.ebi.ac.uk/~guy/exonerate)
TAlioto	Tyler Alioto et al., ES
	GEM (gemlibrary.sourceforge.net)
TWu	Thomas Wu, USA
	GSnap (gene.com/share/gmap)

No alignment submissions from V. Solovyev (Fgenesh). Possible data conversion issues in SWhite. 8 submissions for *D. melanogaster*.

RGASP.2 Read Alignment Submissions



Submitters: $* =$ late

ADobin*	Alexander Dobin et al., USA
	STAR
AMortazavi*	Ali Mortazavi et al., USA
	Erange (woldlab.caltech.edu/rnaseq)
CIseli	Christian Iseli et al., CH
	SIBsim4 (sibsim4.sourceforge.net)
GRaetsch	Gunnar Rätsch et al., DE
	PALMapper (Jean et al., 2010)
LPachter	Lior Pachter et al., USA
	TopHat (Trapnell et al., 2009)
MGerstein	Mark Gerstein et al., USA
	TopHat (human only)
MStanke	Mario Stanke, DE
	Blat (Kent, 2002)
SWhite	Simon White et al., UK
	Exonerate (www.ebi.ac.uk/~guy/exonerate)
TAlioto	Tyler Alioto et al., ES
	GEM (gemlibrary.sourceforge.net)
TWu	Thomas Wu, USA
	GSnap (gene.com/share/gmap)

1. *C. elegans*

CIseli	15G	77 M alns.
GRaetsch	12/19G	54/85 M alns.
LPachter	19G	87 M alns.
MStanke	12/47G	56/221 M alns.
SWhite	23G	110 M alns.
TAlioto	31G	148 M alns.
TWu	24G	112 M alns.

No alignment submissions from V. Solovyev (Fgenesh). Possible data conversion issues in SWhite. 8 submissions for *D. melanogaster*.

RGASP.2 Read Alignment Submissions



Submitters: * = late

ADobin*	Alexander Dobin et al., USA
	STAR
AMortazavi*	Ali Mortazavi et al., USA
	Erange (woldlab.caltech.edu/rnaseq)
CIseli	Christian Iseli et al., CH
	SIBsim4 (sibsim4.sourceforge.net)
GRaetsch	Gunnar Rätsch et al., DE
	PALMapper (Jean et al., 2010)
LPachter	Lior Pachter et al., USA
	TopHat (Trapnell et al., 2009)
MGerstein	Mark Gerstein et al., USA
	TopHat (human only)
MStanke	Mario Stanke, DE
	Blat (Kent, 2002)
SWhite	Simon White et al., UK
	Exonerate (www.ebi.ac.uk/~guy/exonerate)
TAlioto	Tyler Alioto et al., ES
	GEM (gemlibrary.sourceforge.net)
TWu	Thomas Wu, USA
	GSnap (gene.com/share/gmap)

1. *C. elegans*

CIseli	15G	77 M alns.
GRaetsch	12/19G	54/85 M alns.
LPachter	19G	87 M alns.
MStanke	12/47G	56/221 M alns.
SWhite	23G	110 M alns.
TAlioto	31G	148 M alns.
TWu	24G	112 M alns.

2. Human

ADobin	37G	91 M alns.
AMortazavi	12G	88 M alns.
CIseli	10G	48 M alns.
GRaetsch	9/21G	43/93 M alns.
LPachter	14G	65 M alns.
MGerstein	16G	74 M alns.
MStanke	9/49G	40/232 M alns.
SWhite	16G	77 M alns.
TAlioto	30G	146 M alns.
TWu	17G	78 M alns.

No alignment submissions from V. Solovyev (Fgenesh). Possible data conversion issues in SWhite. 8 submissions for *D. melanogaster*.

Questions we asked...



- What are specific properties of the different alignments?
  - Mismatch, indel & intron distribution
- How "accurate" are the alignments?
  - Spliced alignments
  - Multiple mappers
- Can the alignments accuracy be improved by filtering?
  - Considered different filtering strategies
- What differences does it make for down-stream analyses?
  - Transcript reconstruction
  - Transcript quantitation

Preliminaries

Some participants submitted two versions (ADobin, MStanke, GRaetsch):

- full set ("*_full*")
- filtered set ("*_filtered*")

$\Rightarrow$ Considered as separate submissions

Preliminaries

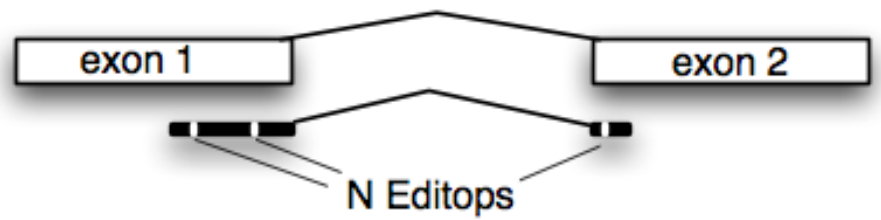
Some participants submitted two versions (ADobin, MStanke, GRaetsch):

- full set ("*_full*")
- filtered set ("*_filtered*")

$\Rightarrow$ Considered as separate submissions

We considered *subsets* of alignments:

1. original submission (sanitized, ...)
2. Filtered by number of edit-operations (e.g., $\leq$ 4 editops)



Preliminaries

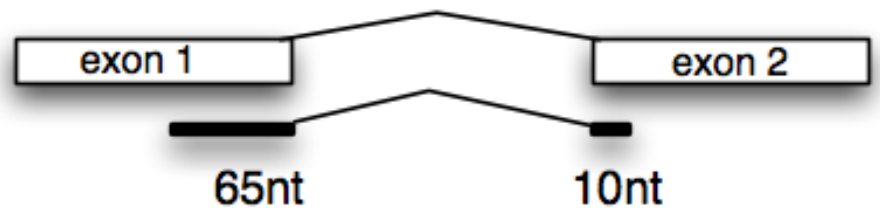
Some participants submitted two versions (ADobin, MStanke, GRaetsch):

- full set ("*_full*")
- filtered set ("*_filtered*")

$\Rightarrow$ Considered as separate submissions

We considered *subsets* of alignments:

1. original submission (sanitized, ...)
2. Filtered by number of edit-operations (e.g., $\leq$ 4 editops)
3. Filtered by number minimal segment length for spliced reads



Preliminaries

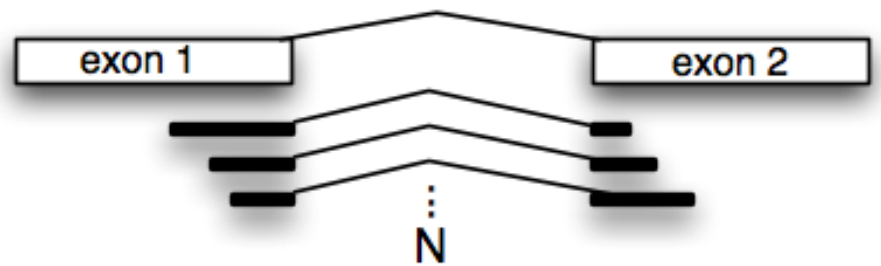
Some participants submitted two versions (ADobin, MStanke, GRaetsch):

- full set ("*_full*")
- filtered set ("*_filtered*")

$\Rightarrow$ Considered as separate submissions

We considered *subsets* of alignments:

1. original submission (sanitized, ...)
2. Filtered by number of edit-operations (e.g., $\leq$ 4 editops)
3. Filtered by number minimal segment length for spliced reads
4. Filtered by number of confirmation of spliced reads



Alignment Accuracy

Problem: We do not know the truth.

Possible solutions:

- Compare different submissions among each-other
  - Pairwise agreement intron predictions
  - Agreement of alignments of multiple mappers
- Compare alignments with information from annotation
  - Predicted vs. annotated introns
- Alignment of artificially generated reads
  - Correct read alignment origin is known

  $\Rightarrow$ RGASP.3

Pairwise Comparison of Intron Predictions



How similar are predicted introns in different submissions?

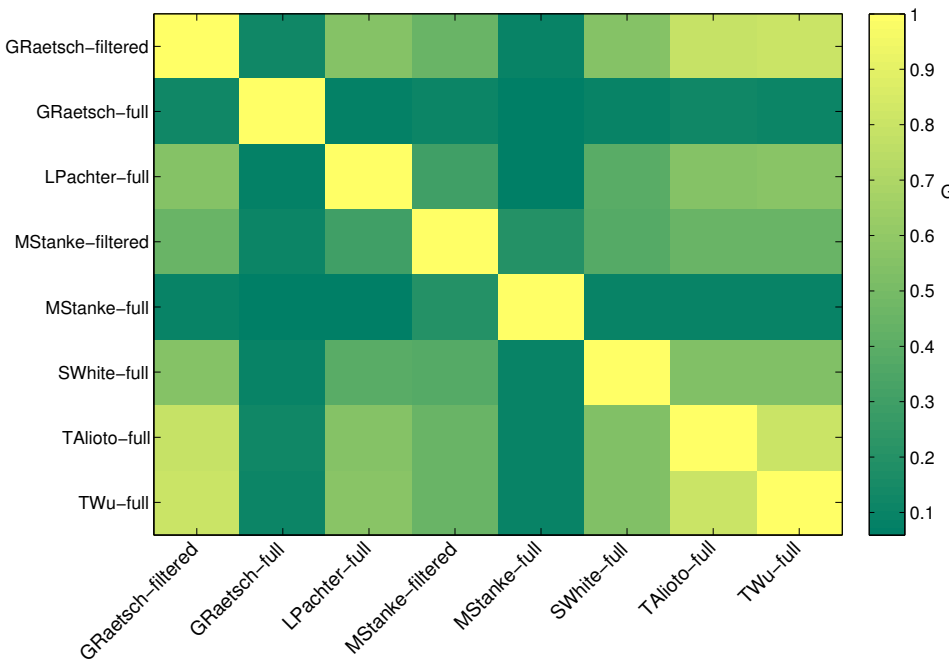
I_A and I_B are the sets of all introns in submission A and B.

We compute an agreement score:

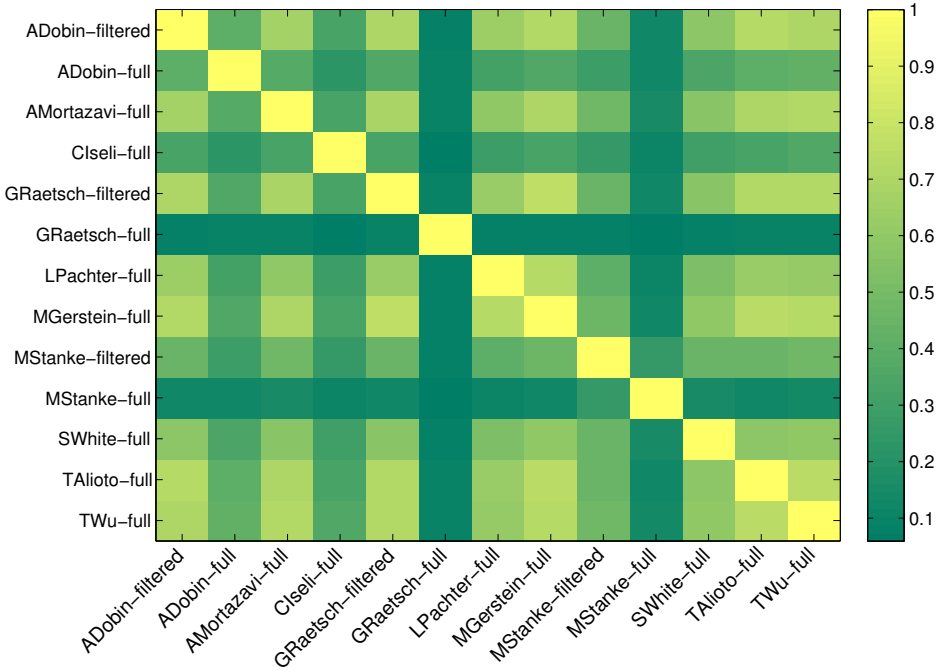
$$S_{A,B} = \frac{|I_A \cap I_B|}{|I_A \cup I_B|}$$

Pairwise Comparison of Intron Predictions

C. elegans



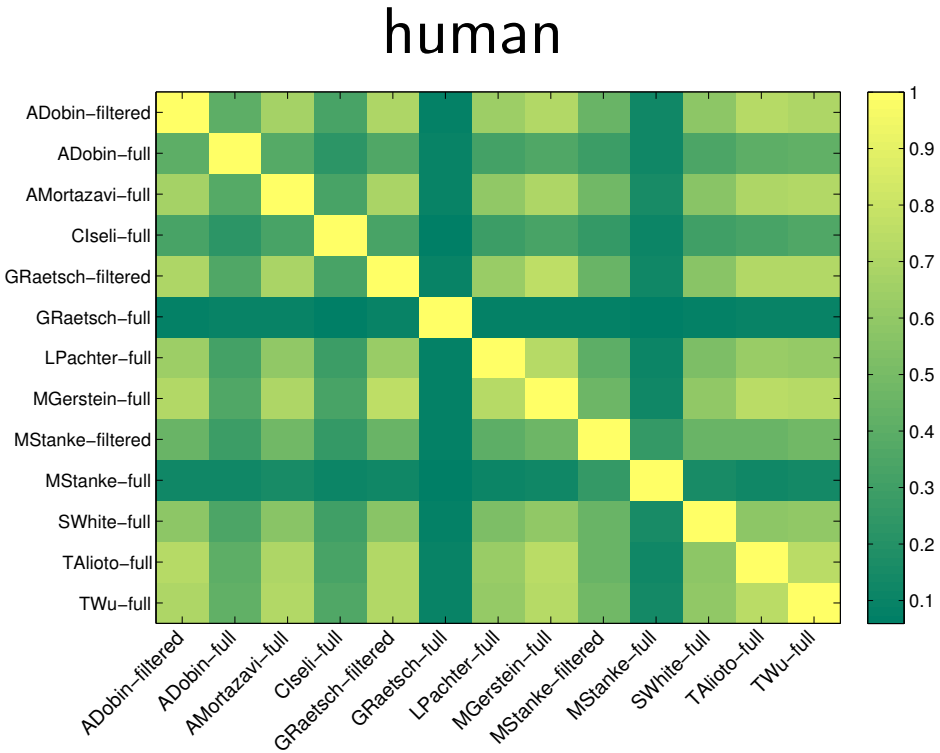
human



Observations:

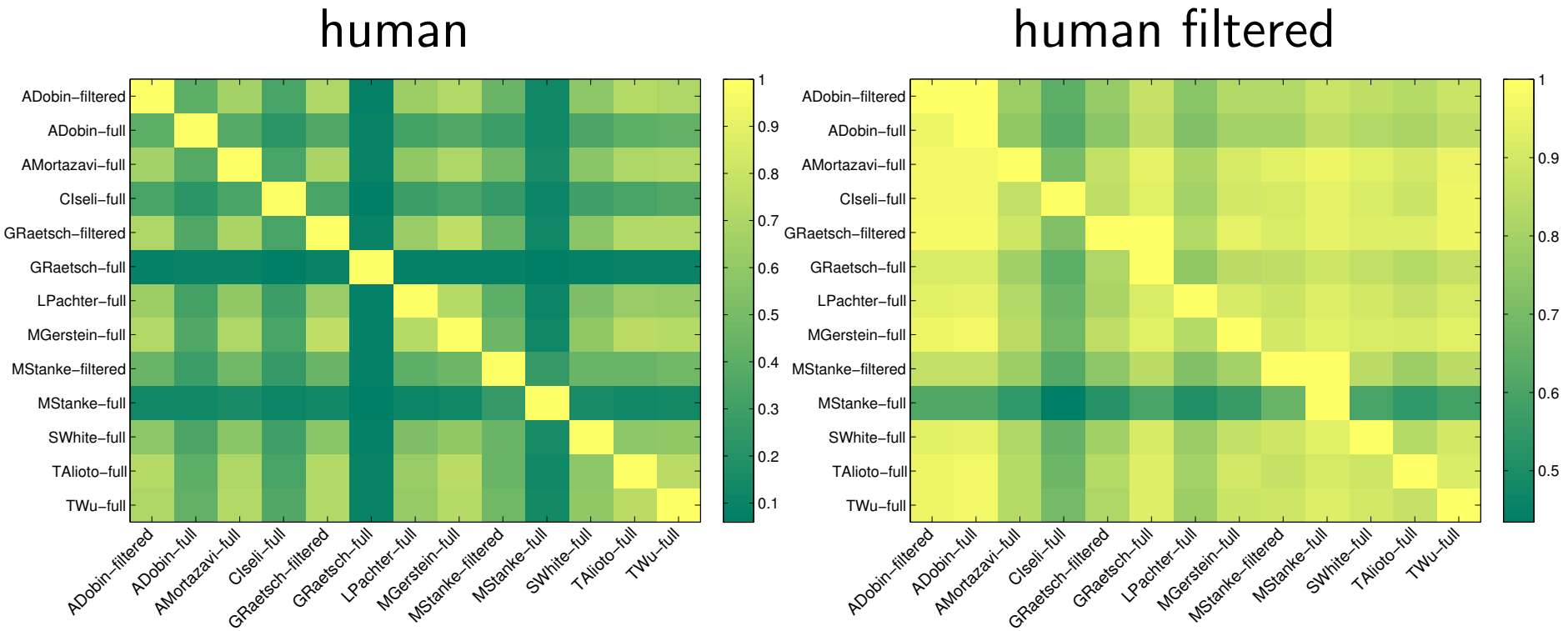
- Unfiltered sets MStanke-full, GRaetsch-full, predict many unique (wrong?) introns.
- Good agreement between a subset of alignment sets.

Pairwise Comparison of Intron Predictions



- Filtering (right): $\leq$1 mismatches, $\geq$12nt segment length, $\geq$2 confirmations
- Much better agreement between alignment sets

Pairwise Comparison of Intron Predictions



- Filtering (right): $\leq$1 mismatches, $\geq$12nt segment length, $\geq$2 confirmations
- Much better agreement between alignment sets

Agreement of Intron Predictions with Annotation

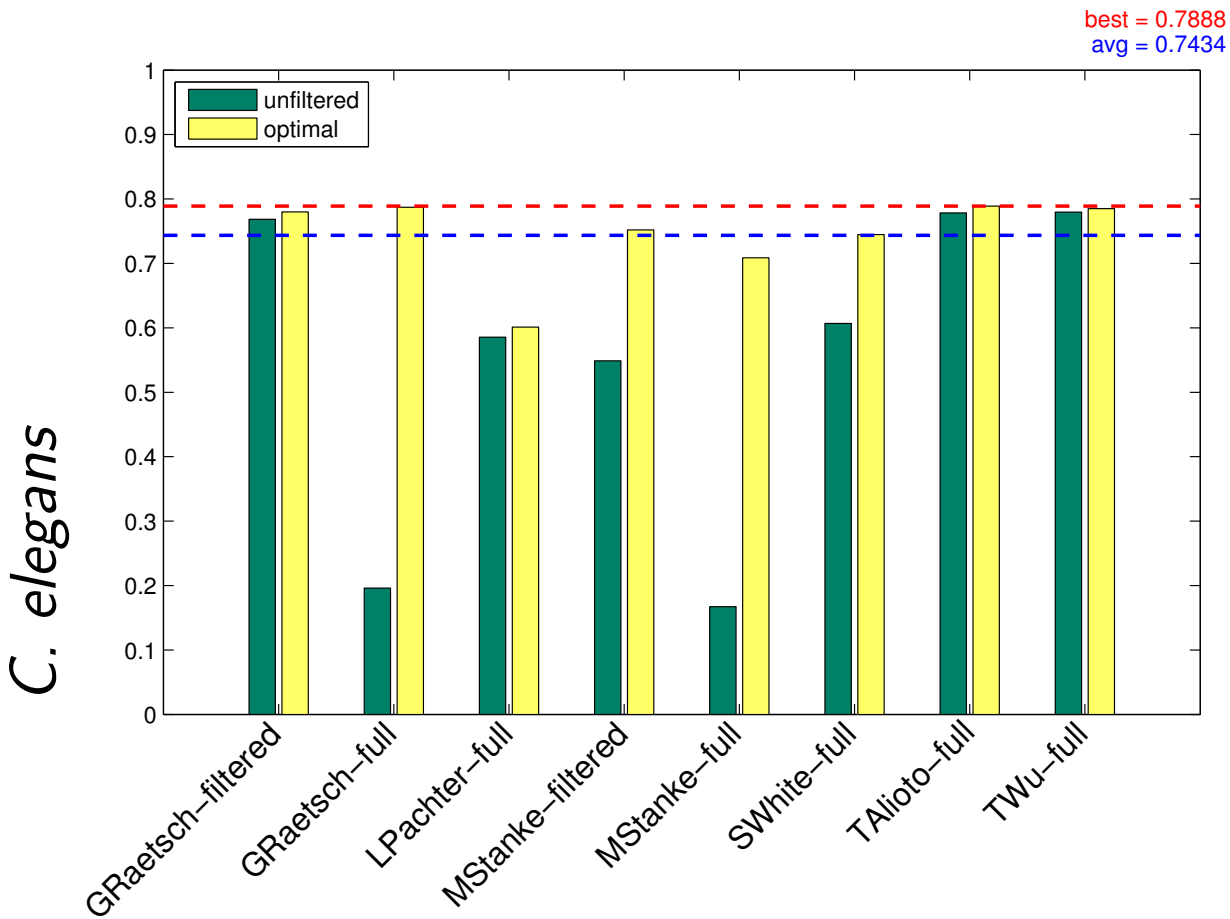
Strategy:

- Consider annotated introns as gold standard
- F-score of agreement of predicted introns with gold standard
- Compare unfiltered with optimally filtered alignment sets

Agreement of Intron Predictions with Annotation

Strategy:

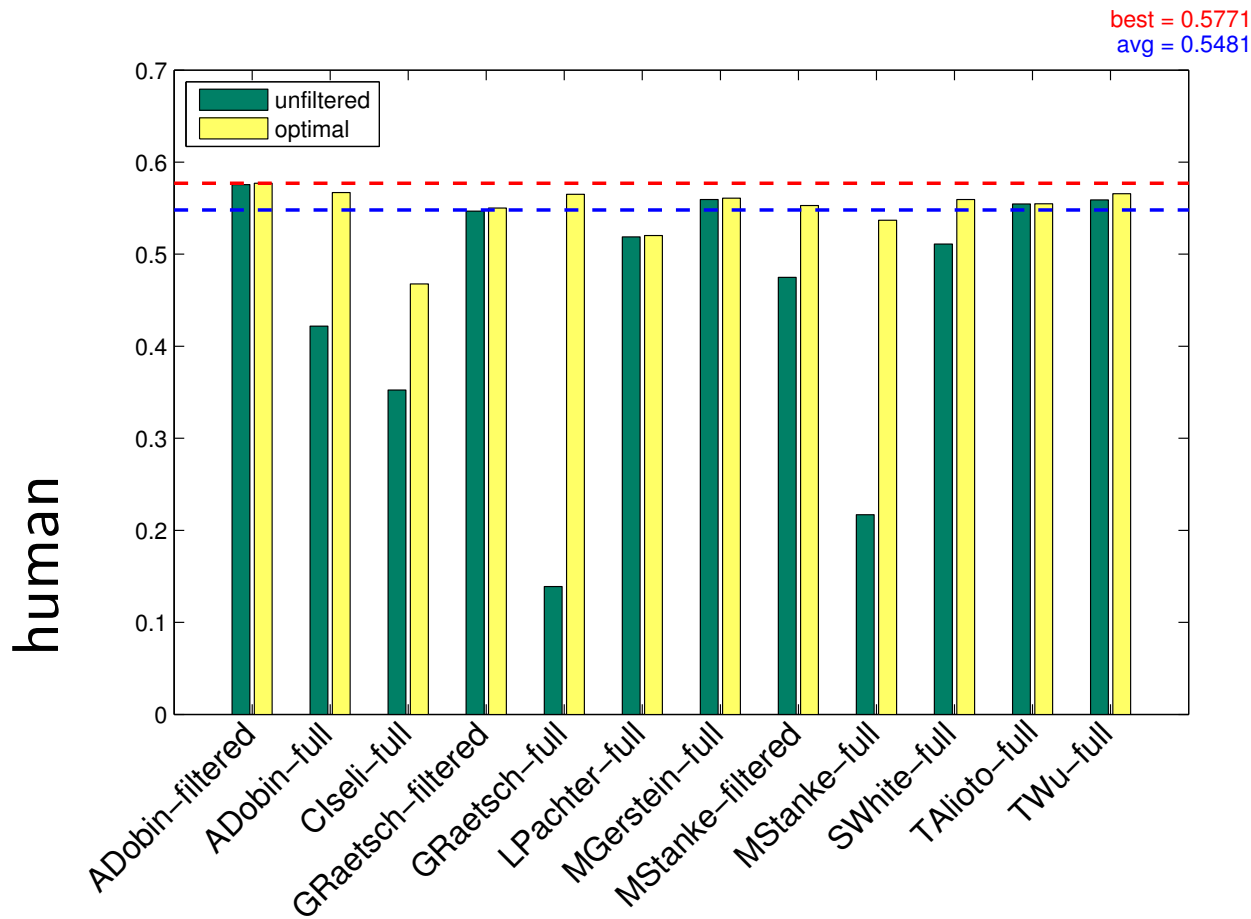
- Consider annotated introns as gold standard
- F-score of agreement of predicted introns with gold standard
- Compare unfiltered with optimally filtered alignment sets



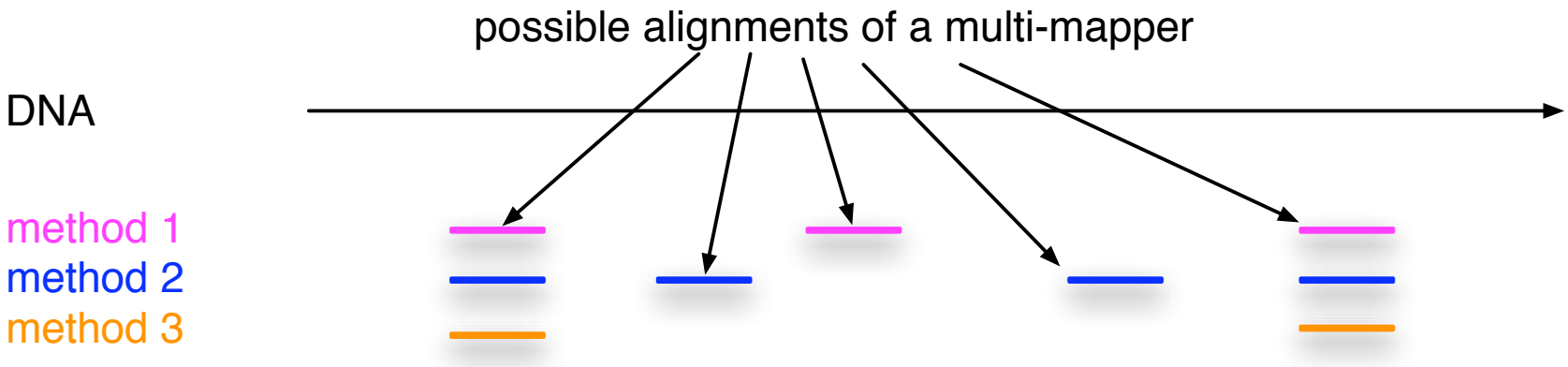
Agreement of Intron Predictions with Annotation

Strategy:

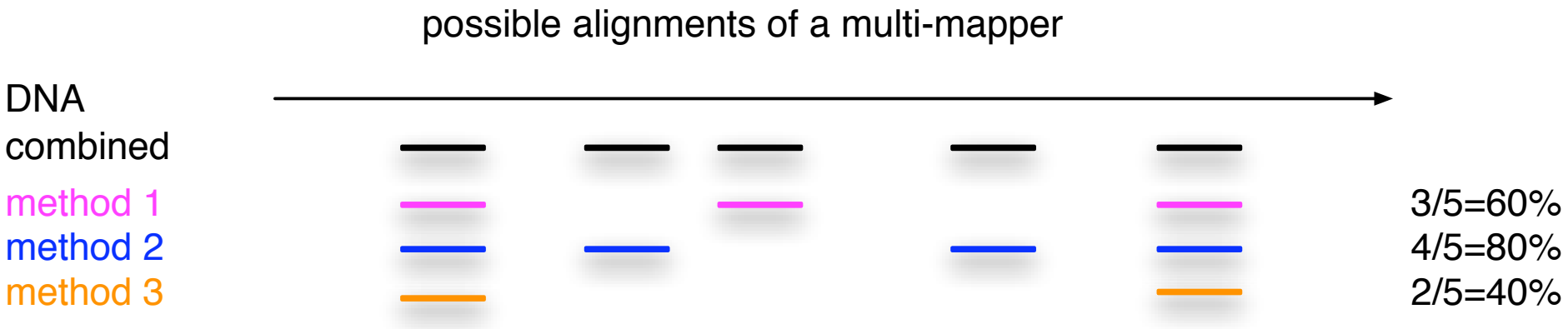
- Consider annotated introns as gold standard
- F-score of agreement of predicted introns with gold standard
- Compare unfiltered with optimally filtered alignment sets



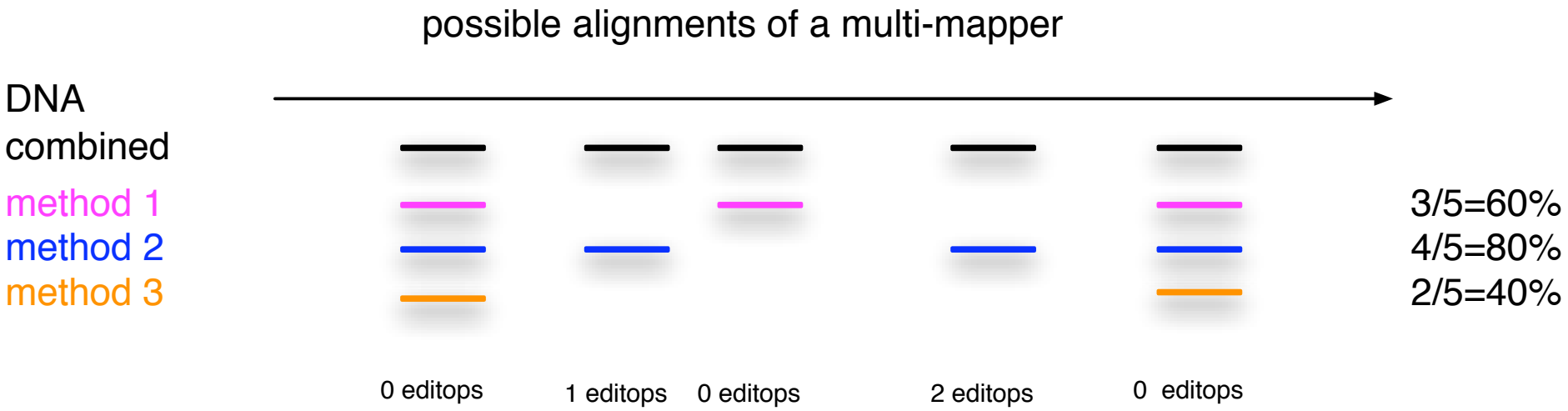
Comparing Alignments of Multi-mappers



Comparing Alignments of Multi-mappers

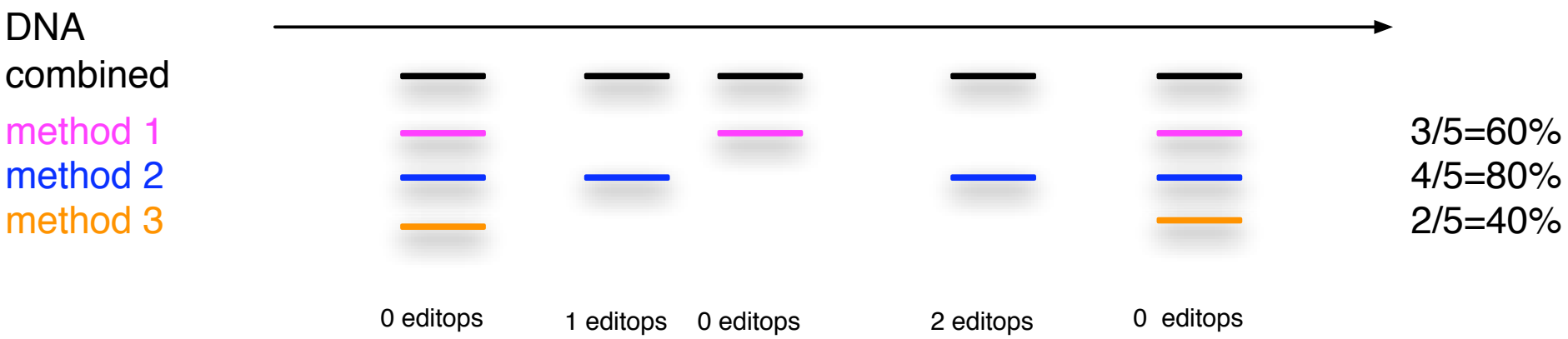


Comparing Alignments of Multi-mappers



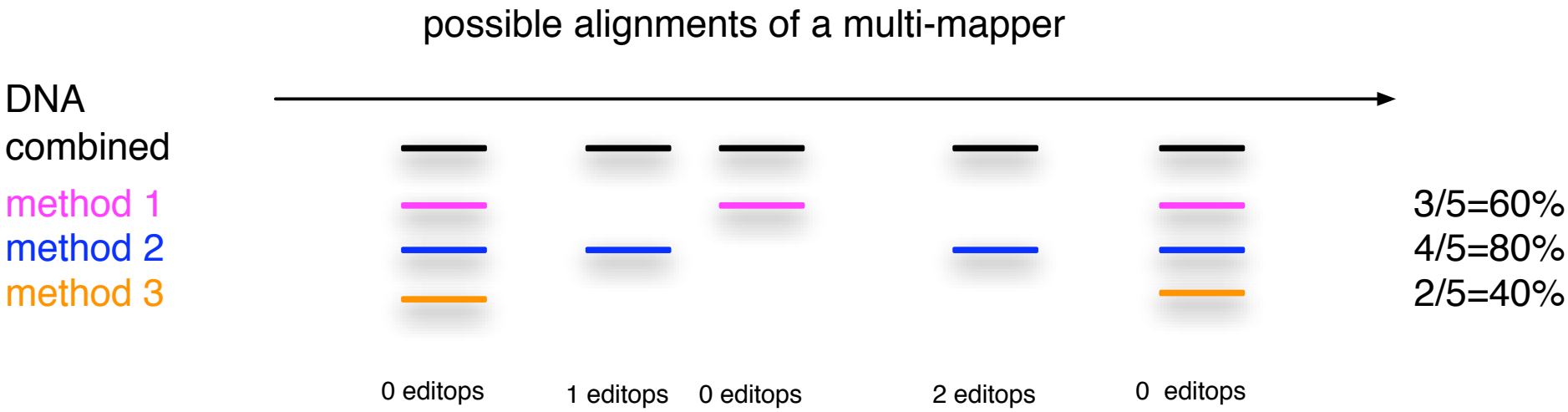
Comparing Alignments of Multi-mappers

possible alignments of a multi-mapper



	0 mm	1 mm	2 mm
method 1	3/3	0/1	0/1
method 2	2/3	1/1	1/1
method 3	2/3	0/1	0/1

Comparing Alignments of Multi-mappers

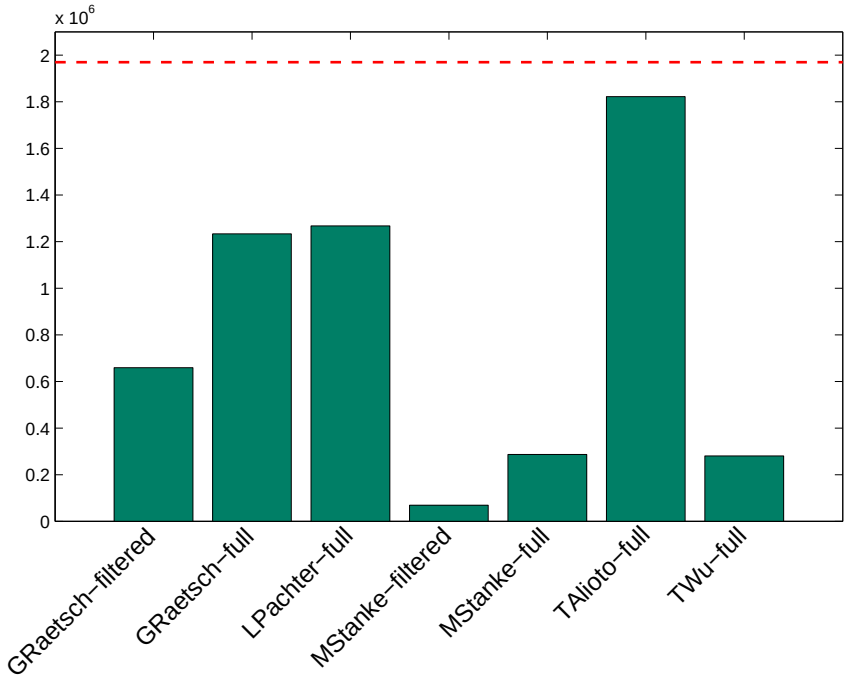


	0 mm	1 mm	2 mm
method 1	3/3	0/1	0/1
method 2	2/3	1/1	1/1
method 3	2/3	0/1	0/1

Alignments are also counted as found when 90% of exonic positions overlap.

Comparing Alignments of Multi-mappers (*C. elegans*)

0 editops



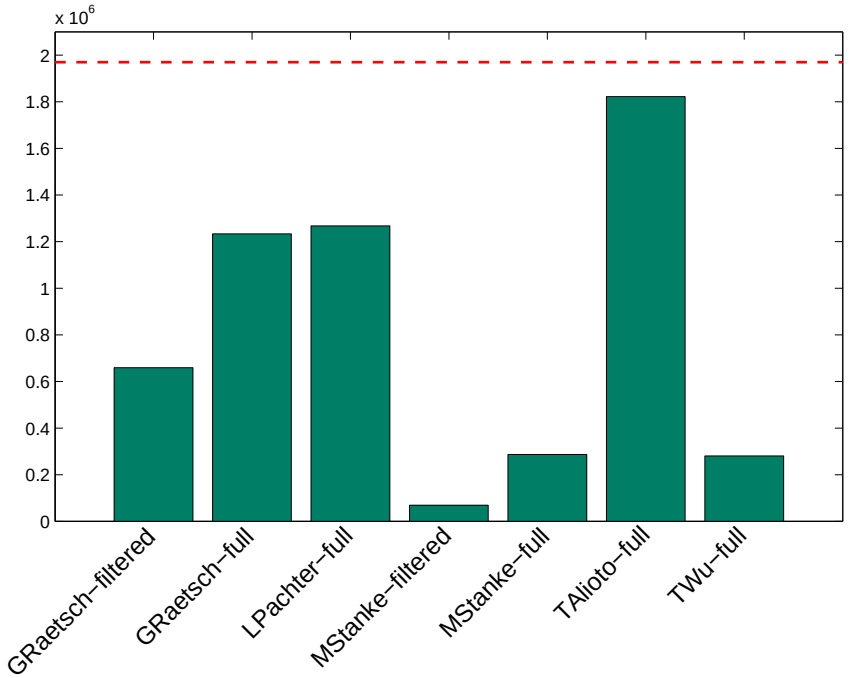
[left] Sensitivity of identifying all alignments with zero editops

(scaled by number of multi-mappers that align with 0 editops)

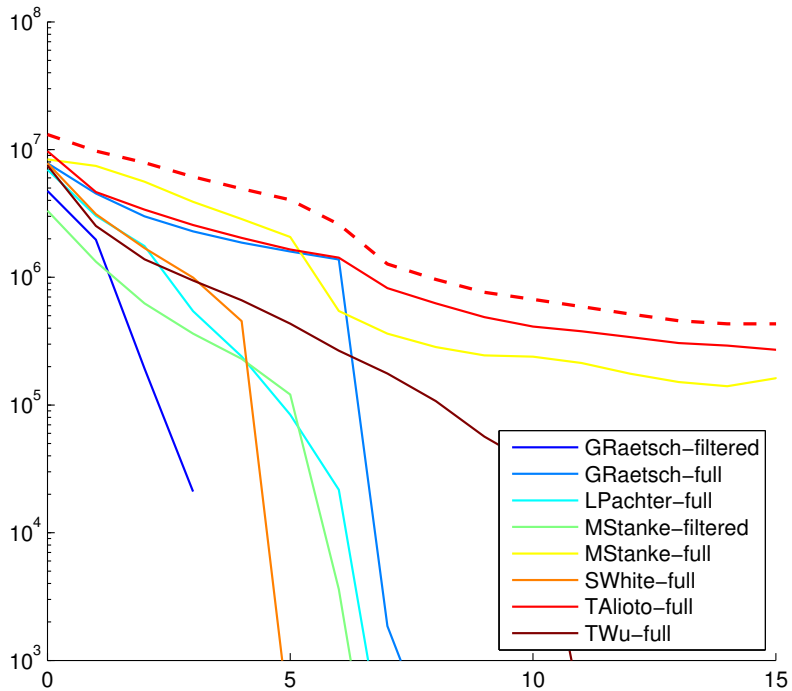
Red dotted line: maximal achievable score

Comparing Alignments of Multi-mappers (*C. elegans*)

0 editops



0-15 editops



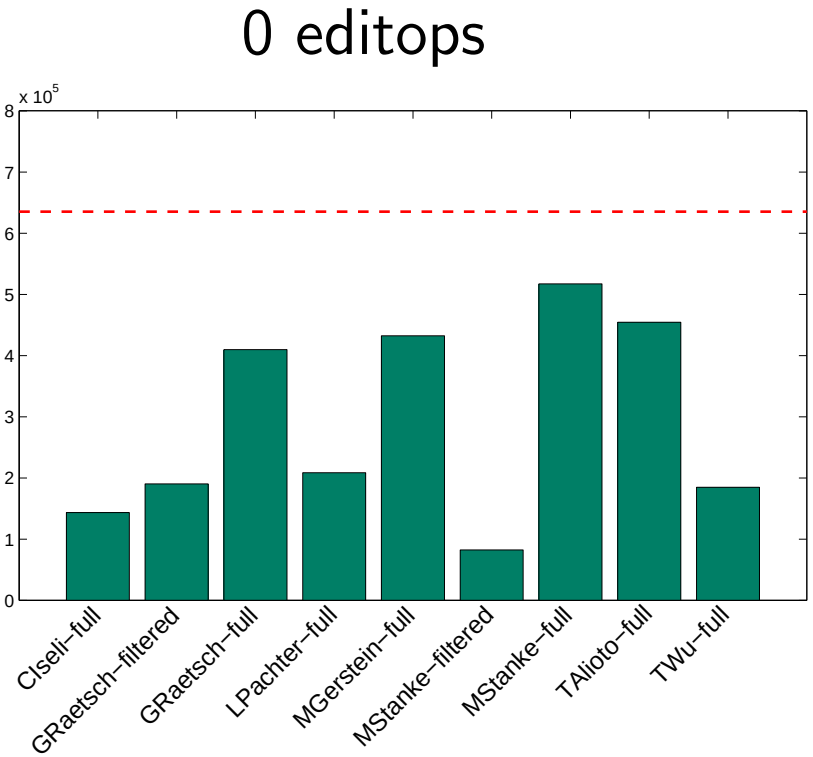
[left] Sensitivity of identifying all alignments with zero editops

(scaled by number of multi-mappers that align with 0 editops)

[right] For different number of editops.

Red dotted line: maximal achievable score

Comparing Alignments of Multi-mappers (human)



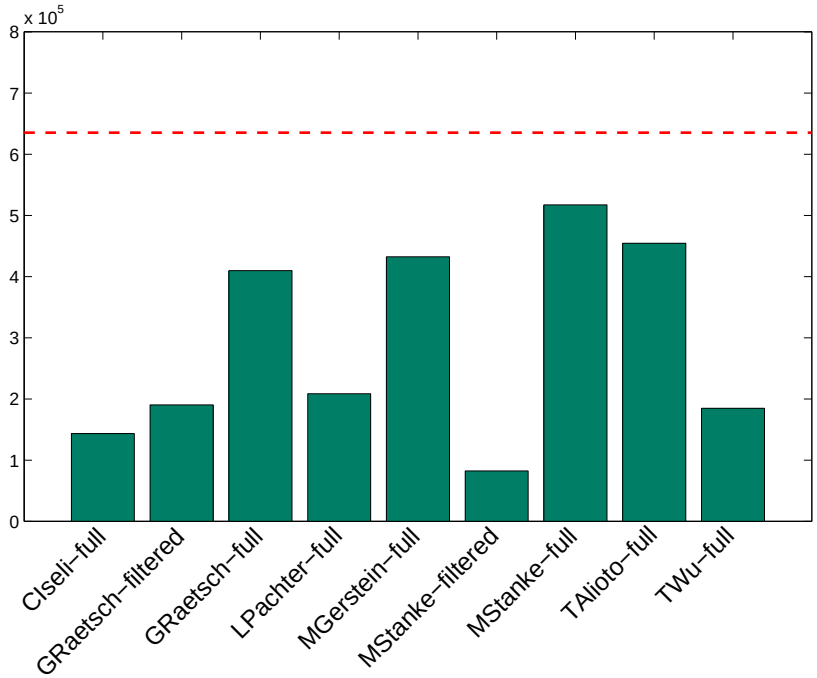
[left] Sensitivity of identifying all alignments with zero editops
(scaled by number of multi-mappers that align with 0 editops)

Red dotted line: maximal achievable score

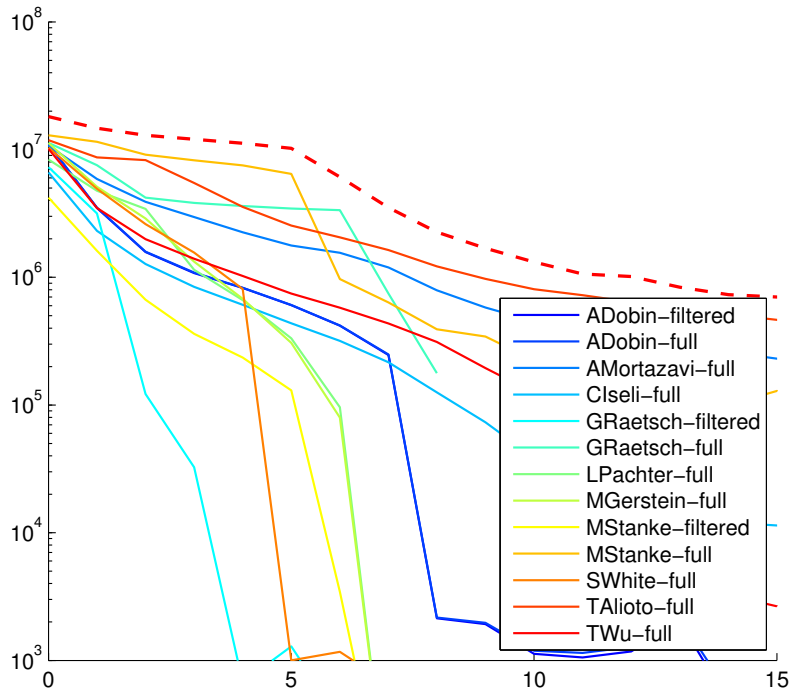
Comparing Alignments of Multi-mappers (human)



0 editops



0-15 editops



[left] Sensitivity of identifying all alignments with zero editops

(scaled by number of multi-mappers that align with 0 editops)

[right] For different number of editops.

Red dotted line: maximal achievable score

Effect on Transcript Prediction



Which influence have alignments on transcript prediction accuracy?

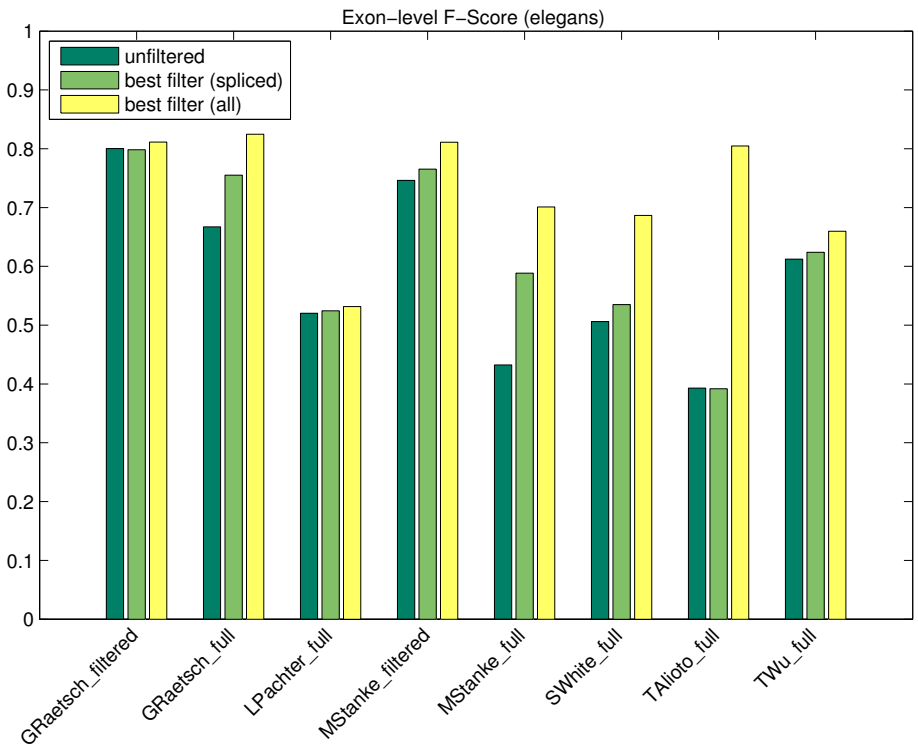
Strategy:

- Use published transcript reconstruction techniques, here Cufflinks & Scripture, to predict transcripts.
- Evaluate prediction accuracy on exon level using annotation as gold-standard:
  - internal exons must match exactly
  - terminal exons have to match on one end
- Report F-Score for
  - Unfiltered alignments
  - Best filtered alignments (spliced reads only)
  - Best filtered alignments (all reads)

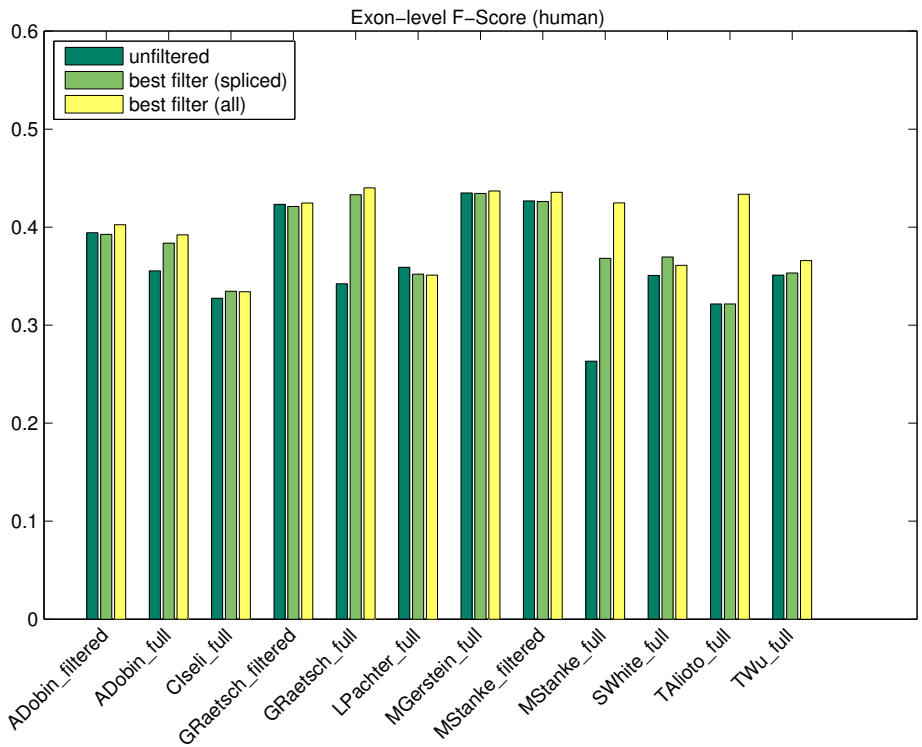
Transcript Prediction with Cufflinks

Exon-level F-Score of Cufflinks transcript predictions relative to annotation.

C. elegans



human

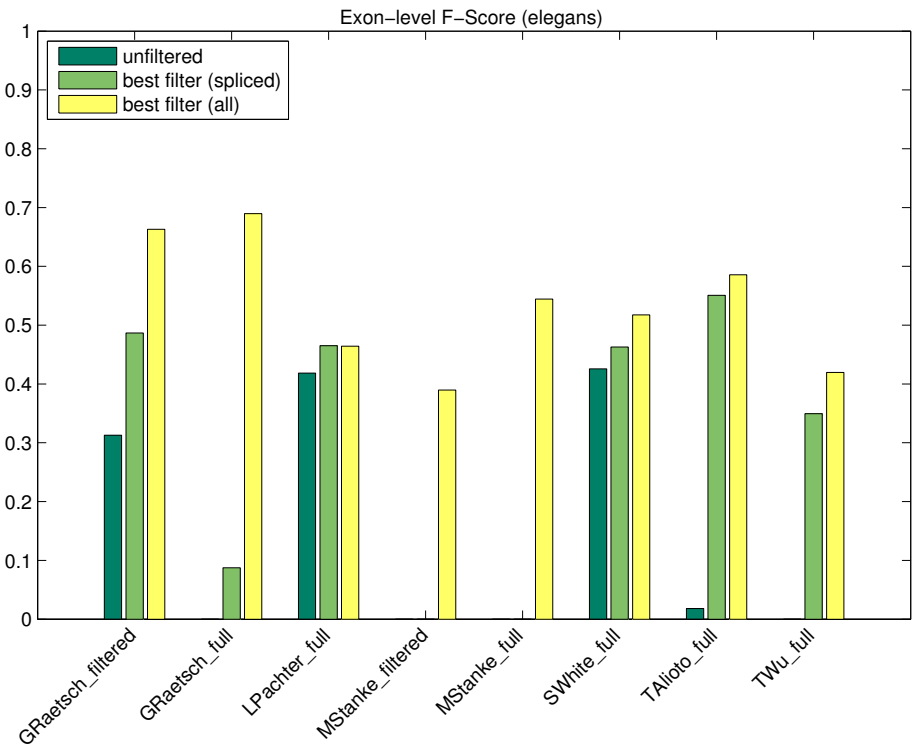


Filtering can significantly increase accuracy.

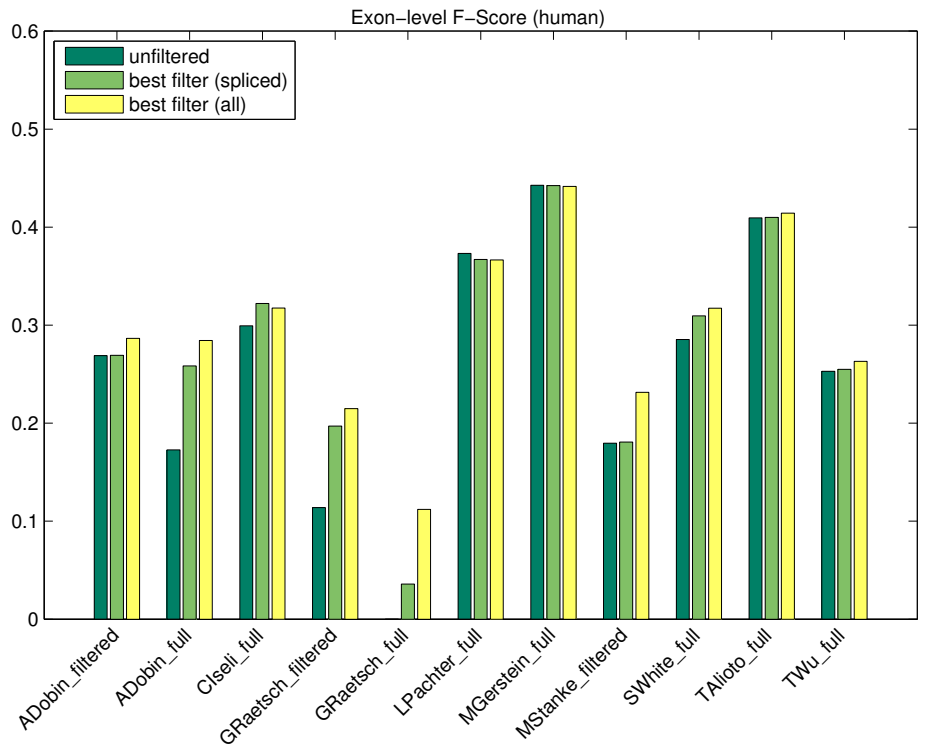
Transcript Prediction with Scripture

Exon-level F-Score of Scripture transcript predictions relative to annotation.

C. elegans



human



Filtering can significantly increase accuracy.

Correlation NanoString vs. RNA-seq

How do the alignments influence the quantitation results?

Strategy:

- Use one quantitation method for all alignment sets
  - Simplified version of *rQuant* (QP-based, without bias-modelling)
  - Original alignments with mismatch filter (best algnmts. in appendix)

Correlation NanoString vs. RNA-seq

How do the alignments influence the quantitation results?

Strategy:

- Use one quantitation method for all alignment sets
  - Simplified version of *rQuant* (QP-based, without bias-modelling)
  - Original alignments with mismatch filter (best algnmts. in appendix)
- Compare quantitation results with NanoString measurements (Geiss et al., 2008; Fortina and Surrey, 2008)
  - *C. elegans* (145 probes, 45 targeting alternative transcripts)
  - human (106 probes, 38 targeting alternative transcripts)

Correlation NanoString vs. RNA-seq

How do the alignments influence the quantitation results?

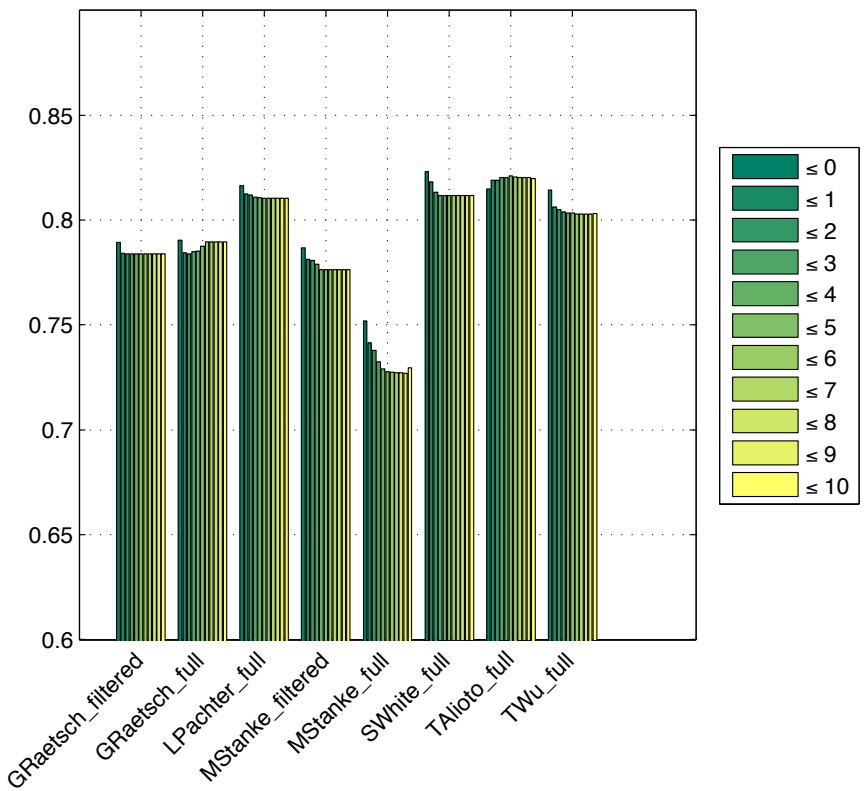
Strategy:

- Use one quantitation method for all alignment sets
  - Simplified version of *rQuant* (QP-based, without bias-modelling)
  - Original alignments with mismatch filter (best algnmts. in appendix)
- Compare quantitation results with NanoString measurements (Geiss et al., 2008; Fortina and Surrey, 2008)
  - *C. elegans* (145 probes, 45 targeting alternative transcripts)
  - human (106 probes, 38 targeting alternative transcripts)
- Evaluation using different measures
  - Pearson correlation (assumes linearity)
  - Spearman correlation (rank correlation, in appendix)
  - Log-Pearson correlation (linear in log-log-plot, in appendix)

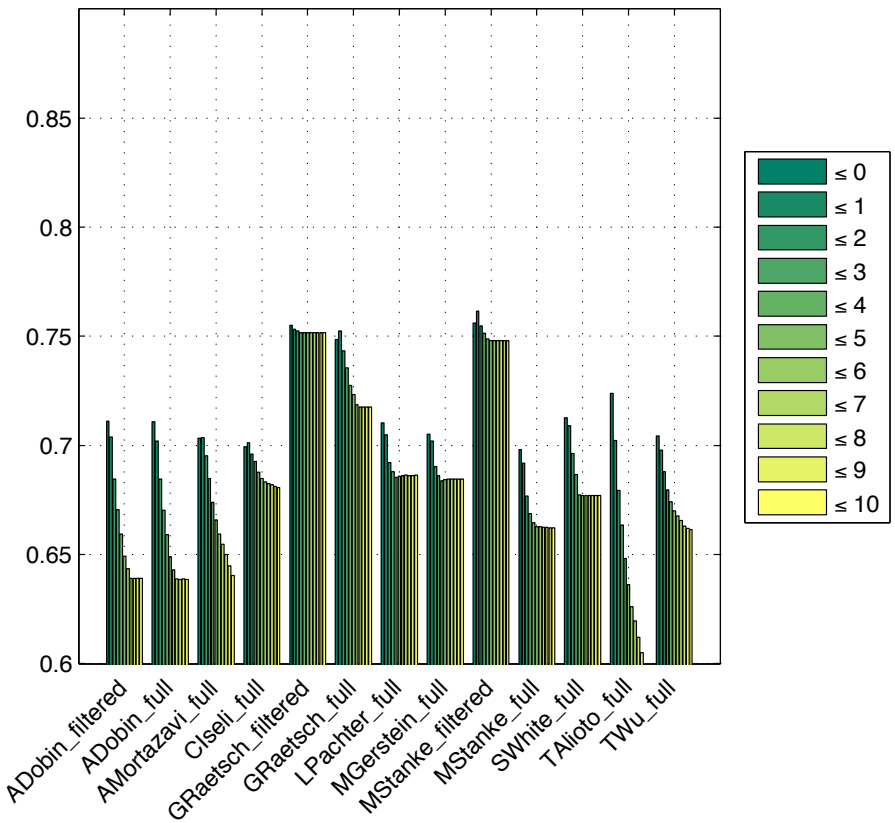
Correlation NanoString vs. RNA-seq

Pearson correlation between NanoString and RNA-seq quantitation for different mismatch filters.

C. elegans



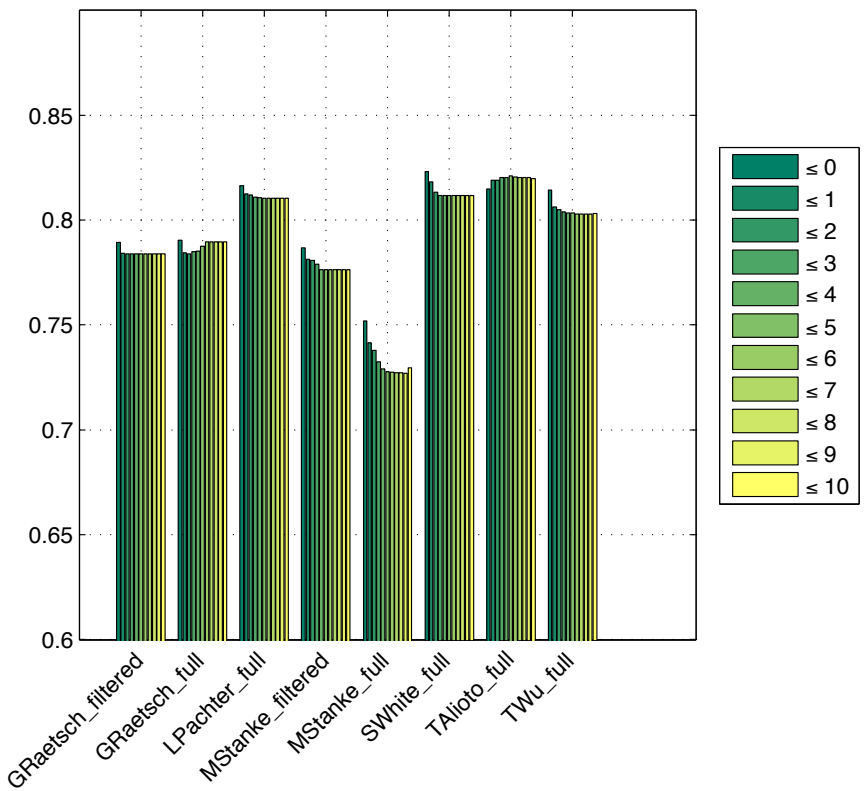
human



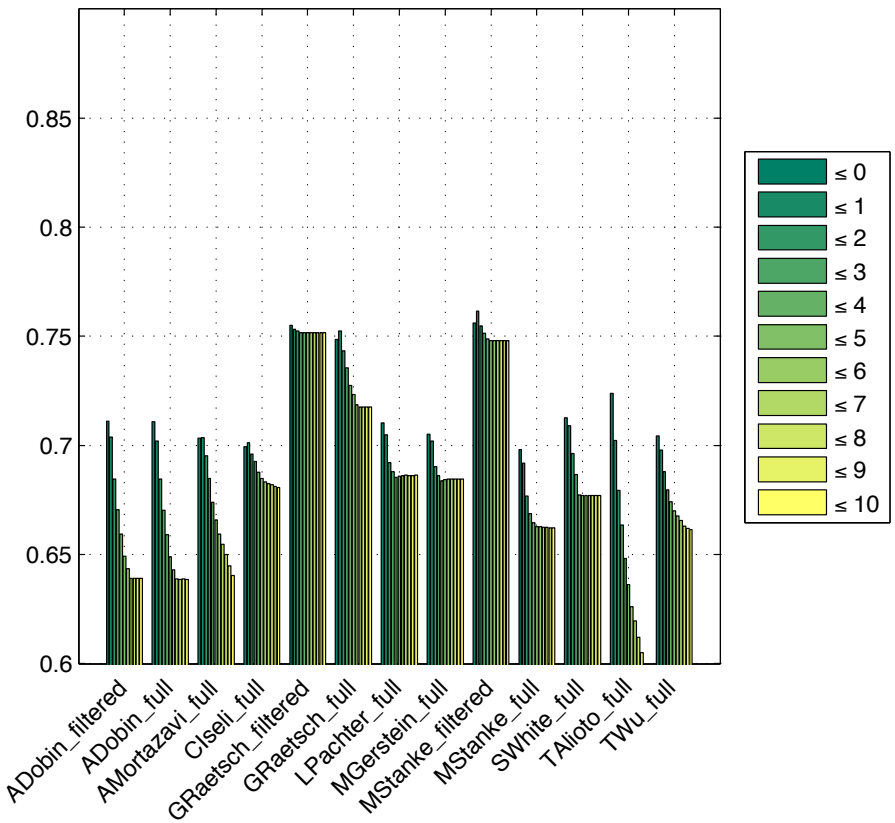
Correlation NanoString vs. RNA-seq

Pearson correlation between NanoString and RNA-seq quantitation for different mismatch filters.

C. elegans



human



Mismatch filtering helps to improve the Pearson correlation.

Summary

Method	Introns	Completeness	Cufflinks	Scripture	Nanostring	# ≤ 3
C. elegans						
GRaetsch	2	3	1	1		4
LPachter		2			3	2
MStanke			3	3		2
SWhite					1	1
TAlioto	1	1	2	2	2	**5**
TWu	3					1
D. melanogaster						
GRaetsch		2	2	2		**3**
LPachter	3		1			2
MStanke	2	1	3			**3**
TAlioto		3		1		2
TWu	1			3		2
Human						
ADobin	1					1
GRaetsch	3		1		2	**3**
LPachter				3		1
MGerstein		2	2	1		**3**
MStanke		1	3		1	**3**
TAlioto		3		2	3	**3**
TWu	2					*1*

Discussion/Conclusion

- Alignments are surprisingly different!
- Filtering is very important
  - optimal strategy depends on task and alignment methods
  - stringent filtering often observed to be beneficial
  - there are more sophisticated filtering methods than used ones

Discussion/Conclusion

- Alignments are surprisingly different!
- Filtering is very important
  - optimal strategy depends on task and alignment methods
  - stringent filtering often observed to be beneficial
  - there are more sophisticated filtering methods than used ones
- Almost every alignment method has benefits in one of the tasks:
  - prediction of introns
  - identification of multiple mapping locations
  - exon/transcript prediction
  - transcript quantitation

Discussion/Conclusion

- Alignments are surprisingly different!
- Filtering is very important
  - optimal strategy depends on task and alignment methods
  - stringent filtering often observed to be beneficial
  - there are more sophisticated filtering methods than used ones
- Almost every alignment method has benefits in one of the tasks:
  - prediction of introns
  - identification of multiple mapping locations
  - exon/transcript prediction
  - transcript quantitation
- Alignment problem not solved and deserves further investigation

Discussion/Conclusion



- Alignments are surprisingly different!
- Filtering is very important
  - optimal strategy depends on task and alignment methods
  - stringent filtering often observed to be beneficial
  - there are more sophisticated filtering methods than used ones
- Almost every alignment method has benefits in one of the tasks:
  - prediction of introns
  - identification of multiple mapping locations
  - exon/transcript prediction
  - transcript quantitation
- Alignment problem not solved and deserves further investigation
- Slides online `http://tinyurl.com/talks-raetschlab`

Acknowledgements



Andre Kahles

Almost everything



Regina Bohnert

Quantitation



Paolo Ribeca@CNAG

Multi mappers



Jonas Behr

Support



References I

Paolo Fortina and Saul Surrey. Digital mrna profiling. *Nat Biotechnol*, 26(3):293–4, Mar 2008. doi: 10.1038/nbt0308-293.

Gary K Geiss, Roger E Bumgarner, Brian Birditt, Timothy Dahl, Naeem Dowidar, Dwayne L Dunaway, H Perry Fell, Sean Ferree, Renee D George, Tammy Grogan, Jeffrey J James, Malini Maysuria, Jeffrey D Mitton, Paola Oliveri, Jennifer L Osborn, Tao Peng, Amber L Ratcliffe, Philippa J Webster, Eric H Davidson, Leroy Hood, and Krassen Dimitrov. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol*, 26(3):317–25, Mar 2008. doi: 10.1038/nbt1385.

G. Jean, A. Kahles, V.T. Sreedharan, F. De Bona, and G. Rätsch. Rna-seq read alignments with palmapper. *Current Protocols in Bioinformatics*, July 2010. accepted.

W James Kent. Blat–the blast-like alignment tool. *Genome Res*, 12(4):656–64, Apr 2002. doi: 10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002.

Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9): 1105–11, May 2009. doi: 10.1093/bioinformatics/btp120.