An Evaluation of RGASP Read Alignment Submissions

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



Friedrich Miescher Laboratory
of the Max Planck Society

Tübingen, Germany

July 10, 2010

Hitseq 2010, Boston

The RGASP Project



RGASP = RNA-seq-based Genome Annotation Assessment Project

- eGASP-style competition to annotate genome with RNA-seq
- Question: Who accurate is transcript prediction & quantitation?

RGASP committee: Jen Harrow, Roderic Guigo, Barbara Wold, Tom Gingeras, Ali Mortazavi, Mark Gerstein, Alex Reymond, Sandrine Dudoit

The RGASP Project



RGASP = RNA-seq-based Genome Annotation Assessment Project

- eGASP-style competition to annotate genome with RNA-seq
- Question: Who accurate is transcript prediction & quantitation?
- Human, *C. elegans*, *D. melanogaster*

RGASP committee: Jen Harrow, Roderic Guigo, Barbara Wold, Tom Gingeras, Ali Mortazavi, Mark Gerstein, Alex Reymond, Sandrine Dudoit

The RGASP Project



RGASP = RNA-seq-based Genome Annotation Assessment Project

- eGASP-style competition to annotate genome with RNA-seq
- Question: Who accurate is transcript prediction & quantitation?
- Human, *C. elegans*, *D. melanogaster*
- Phase 1: different sequencing platforms, read lengths
- Phase 2: Illumina sequencing platform with focused studies
  - Evaluation of quantitations
  - Evaluation of read alignments

RGASP committee: Jen Harrow, Roderic Guigo, Barbara Wold, Tom Gingeras, Ali Mortazavi, Mark Gerstein, Alex Reymond, Sandrine Dudoit

The RGASP Project



RGASP = RNA-seq-based Genome Annotation Assessment Project

- eGASP-style competition to annotate genome with RNA-seq
- Question: Who accurate is transcript prediction & quantitation?
- Human, *C. elegans*, *D. melanogaster*
- Phase 1: different sequencing platforms, read lengths
- Phase 2: Illumina sequencing platform with focused studies
  - Evaluation of quantitations
  - Evaluation of read alignments

RGASP summary by Felix Kokocinski, Monday, 2:30 pm,
Late breaking research session.

RGASP committee: Jen Harrow, Roderic Guigo, Barbara Wold, Tom Gingeras, Ali Mortazavi, Mark Gerstein, Alex Reymond, Sandrine Dudoit

RGASP Read Alignment Submissions



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.
TAlioto	31G	148 M alns.
TWu	24G	112 M alns.

2. *D. melanogaster* (not shown)

3. Human

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.
TAlioto	30G	146 M alns.
TWu	17G	78 M alns.

Legend:

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)
TAlioto	Tyler Alioto et al., ES **GEM** (gemlibrary.sourceforge.net)
TWu	Thomas Wu, USA **GSnap** (gene.com/share/gmap)

No alignment submissions from A. Dobin (STAR), V. Solovyev (Fgenesh), A. Mortazavi (Erange).

Remaining data format issues in files from S. White (exonerate).

Preliminaries

Some participants submitted two versions (MStanke, GRaetsch):

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

$\Rightarrow$ Considered as separate submissions

Preliminaries

Some participants submitted two versions (MStanke, GRaetsch):

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

$\Rightarrow$ Considered as separate submissions

We considered *subsets* of alignments:

1. original submission (sanitized, ...)
2. best alignment only (minimal number of edit operations)

Preliminaries

Some participants submitted two versions (MStanke, GRaetsch):

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

$\Rightarrow$ Considered as separate submissions

We considered *subsets* of alignments:

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

Preliminaries

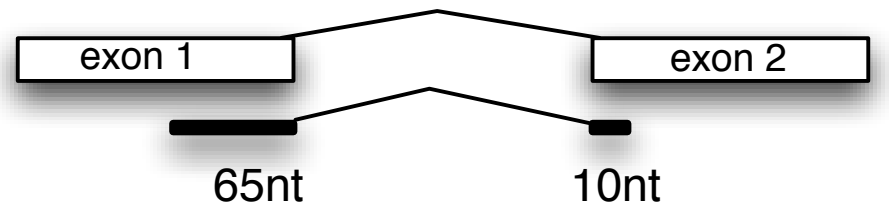
Some participants submitted two versions (MStanke, GRaetsch):

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

$\Rightarrow$ Considered as separate submissions

We considered *subsets* of alignments:

1. original submission (sanitized, ...)
2. best alignment only (minimal number of edit operations)
3. Filtered by number of edit-operations (e.g., ≤ 4 editops)
4. Filtered by number minimal segment length for spliced reads



Preliminaries

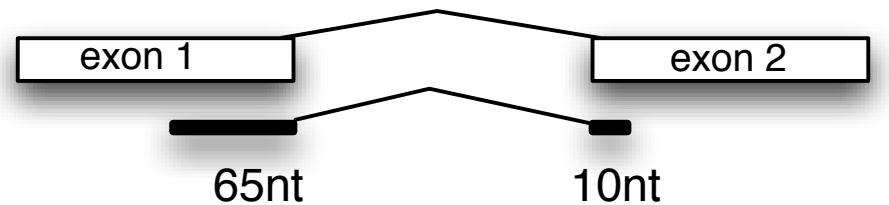
Some participants submitted two versions (MStanke, GRaetsch):

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

$\Rightarrow$ Considered as separate submissions

We considered *subsets* of alignments:

1. original submission (sanitized, ...)
2. best alignment only (minimal number of edit operations)
3. Filtered by number of edit-operations (e.g., ≤ 4 editops)
4. Filtered by number minimal segment length for spliced reads

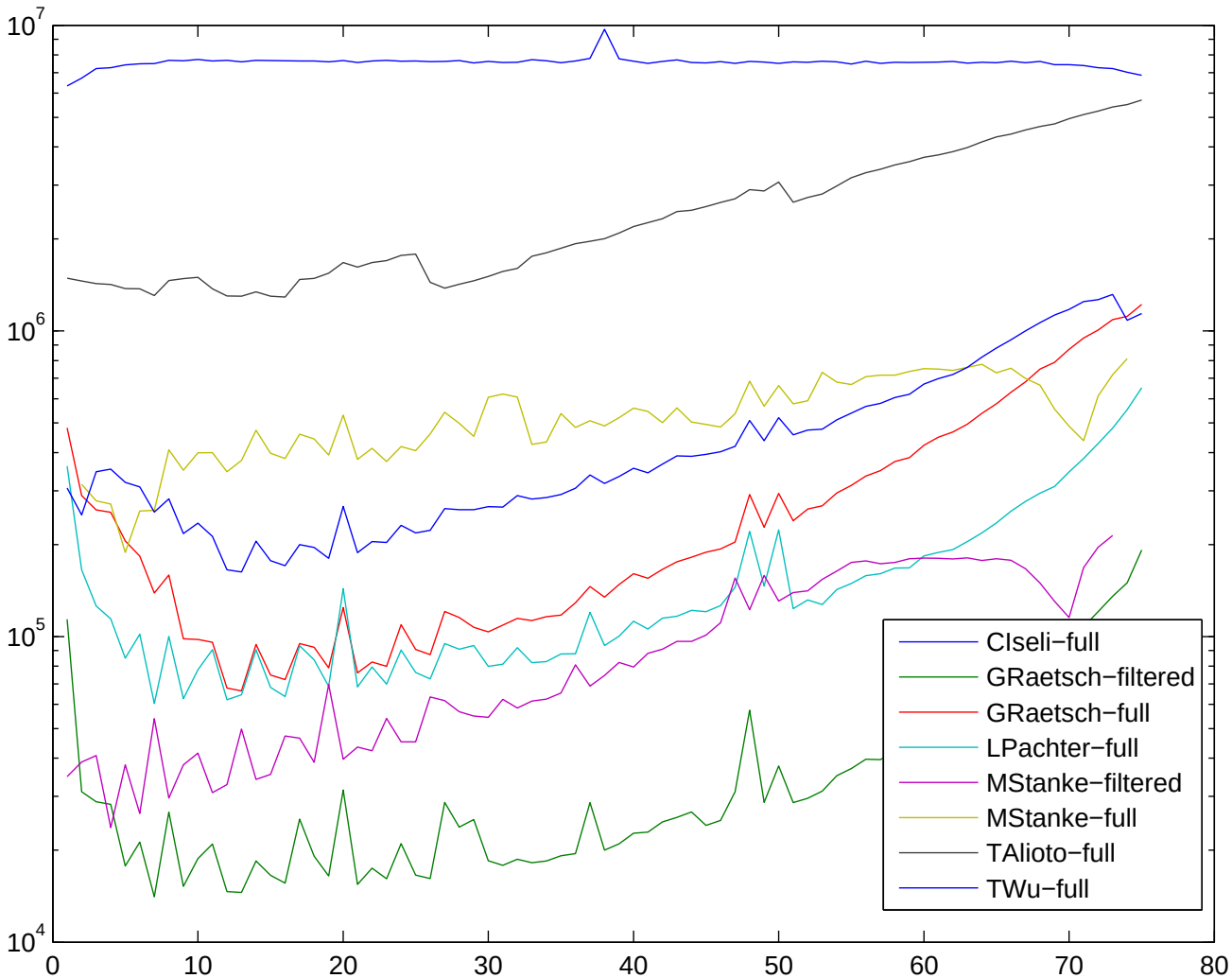


For some analyses we only considered reads from chromosome 1 (when sample size was already large enough).

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 be improve accuracy by filtering?
  - Considered filtering by edit-operations and segment length
- What differences does it make for down-stream analyses?
  - Transcript reconstruction
  - Transcript quantitation

Mismatches per Read Position (*C. elegans*)

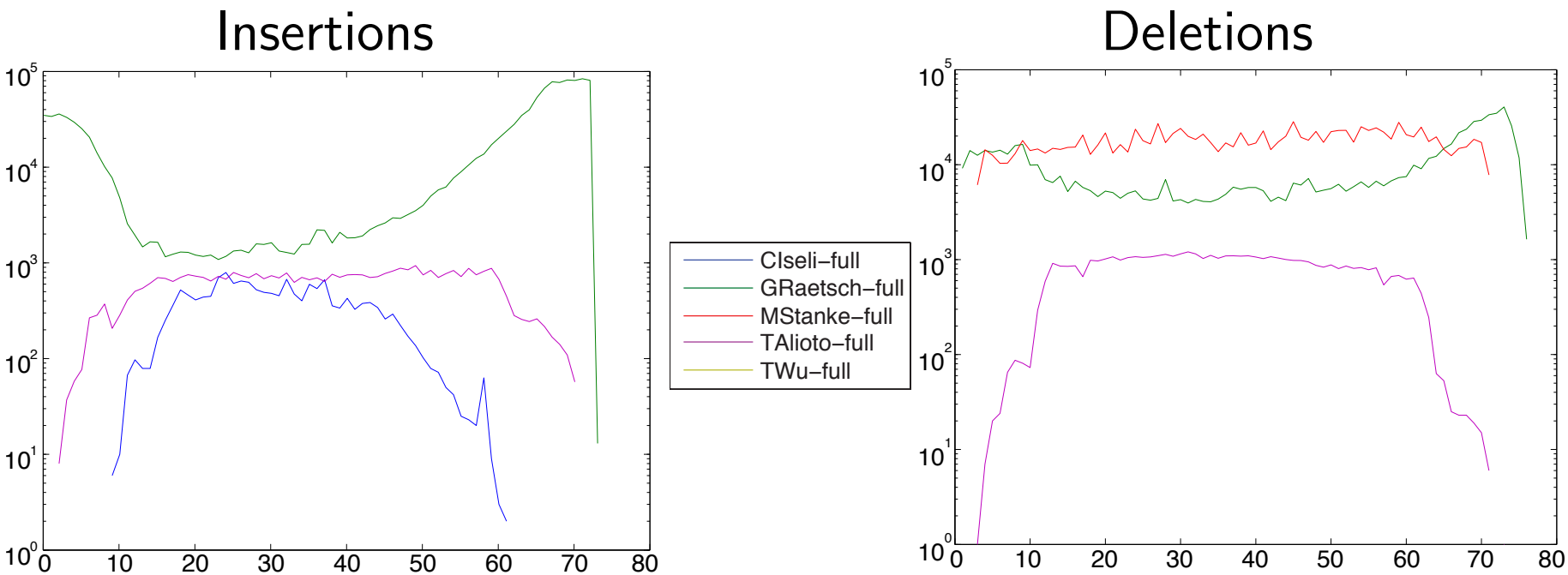


Which methods produced alignments with most mismatches?

Most mismatches for CIseli and MStanke-full. Remaining data conversion issues?

More mismatches near read boundary.

Indels per Read Position (*C. elegans*)

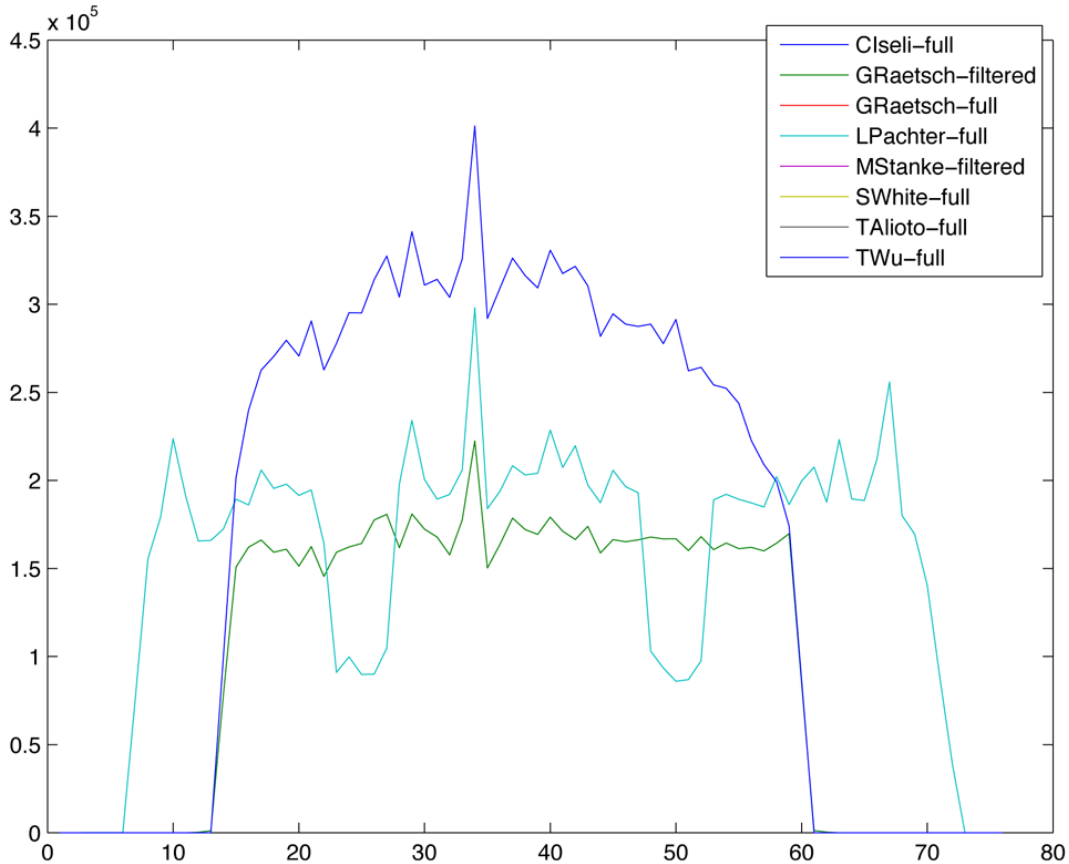


Which methods allow for indels?

No Indel alignments in LPachter, only deletions for MStanke, only insertions in CIseli.

Most indels near read boundary for GRaetsch.

Split Position Distribution (*C. elegans*)

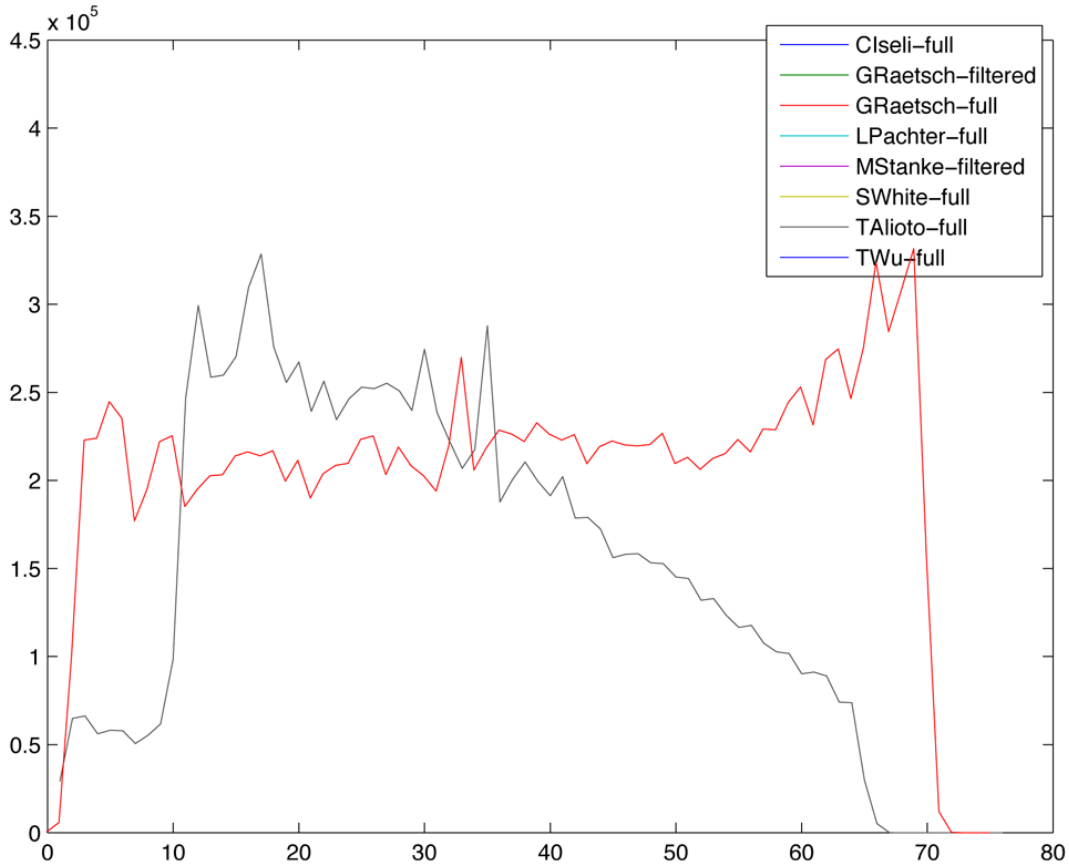


At which read positions are introns predicted?

LPachter predicts less introns in central read positions.

CIseli and GRaetsch do not predict introns at read boundaries

Split Position Distribution (*C. elegans*)



At which read positions are introns predicted?

TAlioto predicts less introns on 3' end.

GRaetsch-full predicts many introns at read boundaries.

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
  - Working on it...

Pairwise Comparison of Intron Predictions



How similar are predicted introns in different submissions?

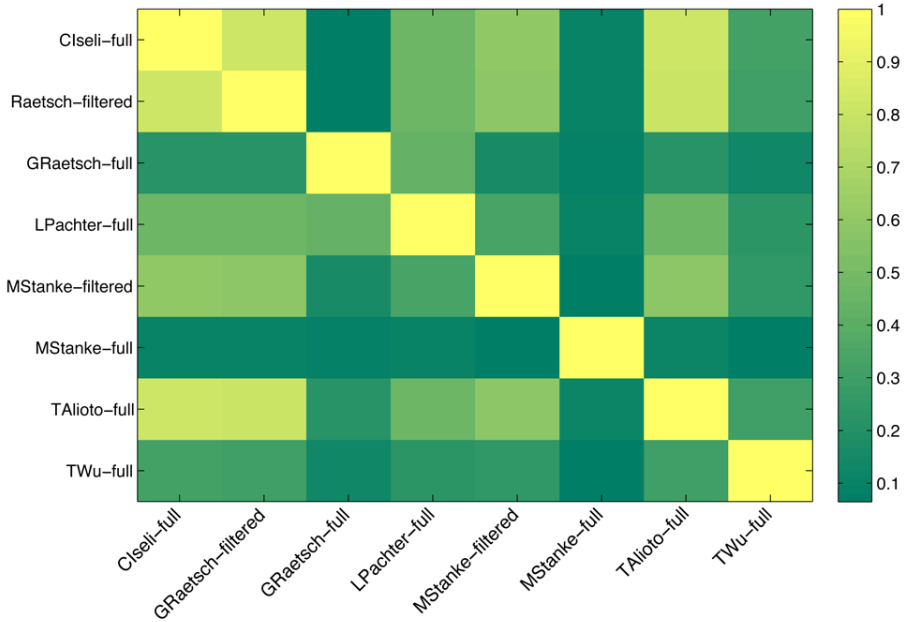
I_A and I_B are the sets of all introns in submission A and B.

We compute an agreement score:

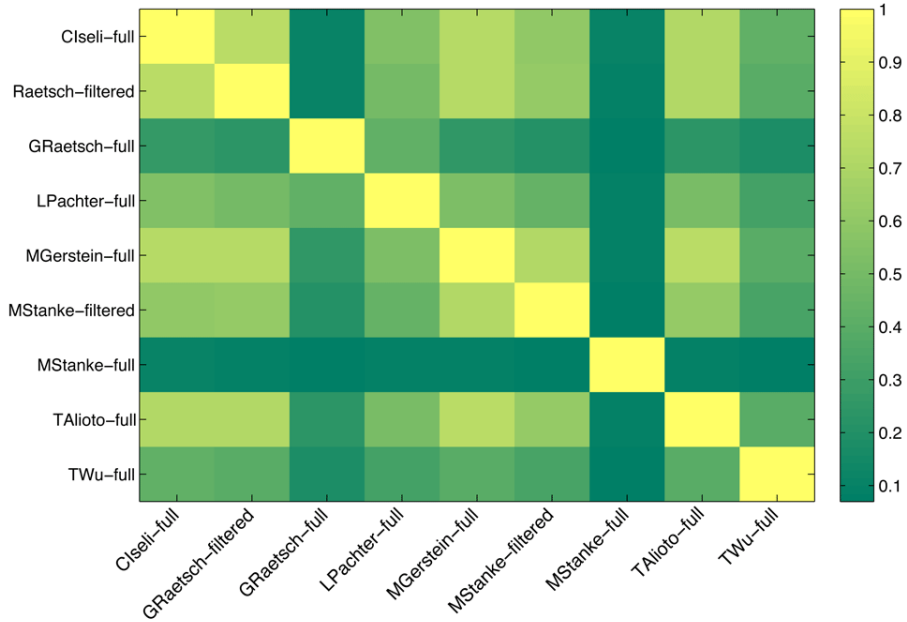
$$S_{A,B} = \frac{|I_A \cap I_B|}{|I_A \cup I_B|}$$

Pairwise Comparison of Intron Predictions

C. elegans



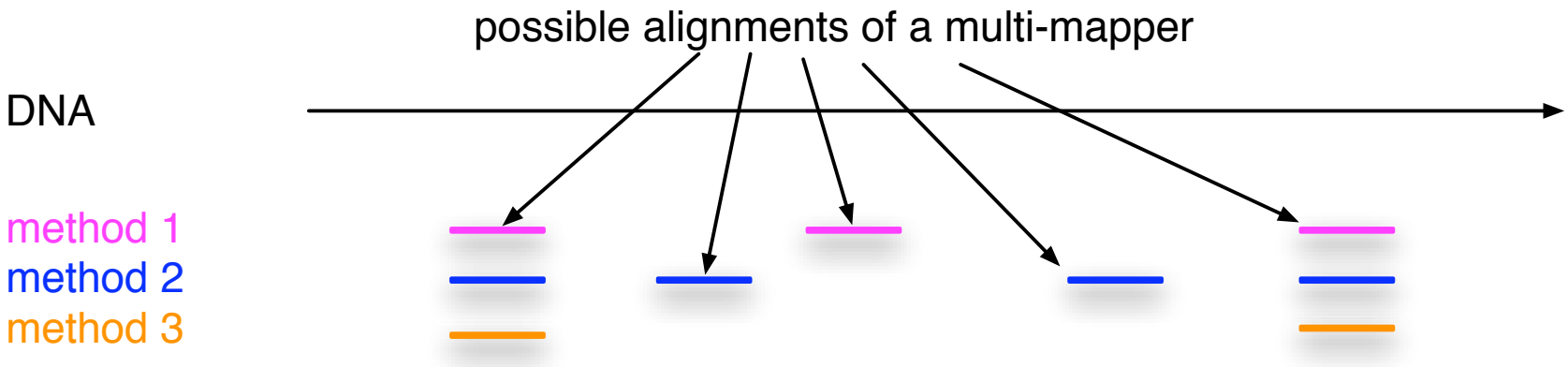
human



Observations:

- Unfiltered sets MStanke-full, GRaetsch-full, predict many unique (wrong?) introns.
- Good agreement between CIseli, GRaetsch-filtered, TAlioto (also MGerstein and MStanke-filtered for human).

Comparing Alignments of Multi-mappers

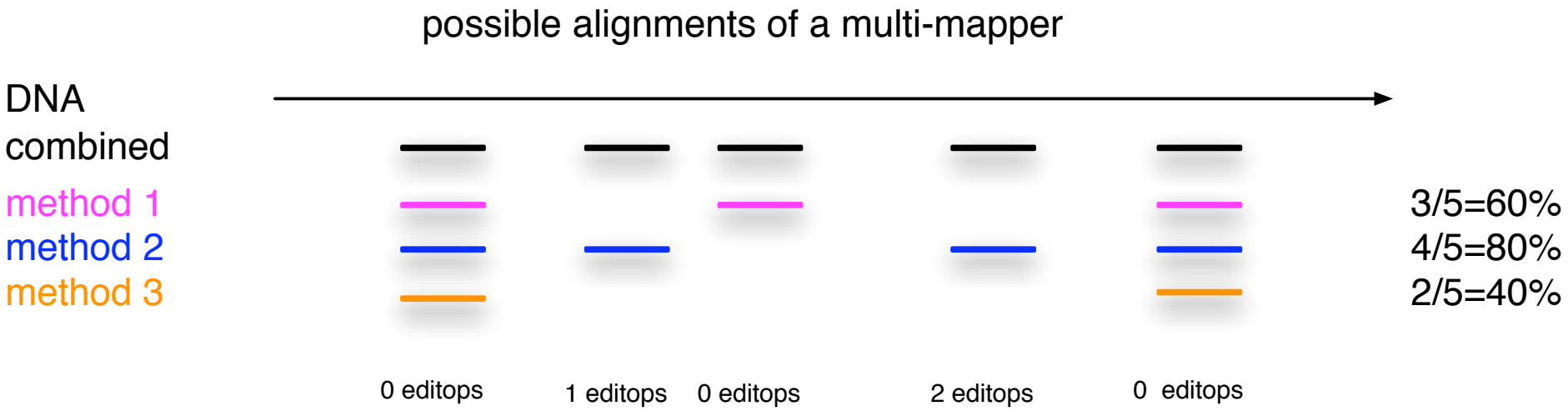


Comparing Alignments of Multi-mappers

possible alignments of a multi-mapper

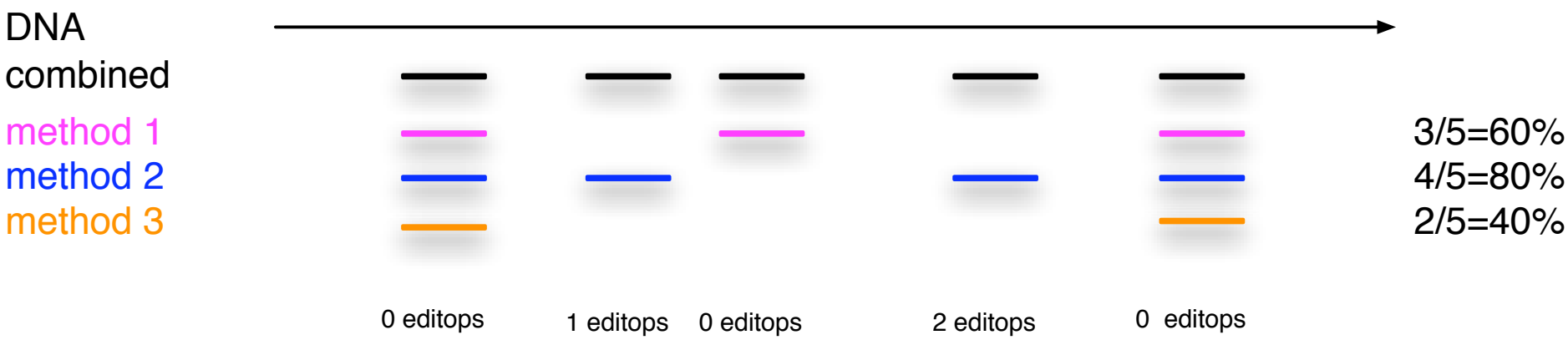


Comparing Alignments of Multi-mappers



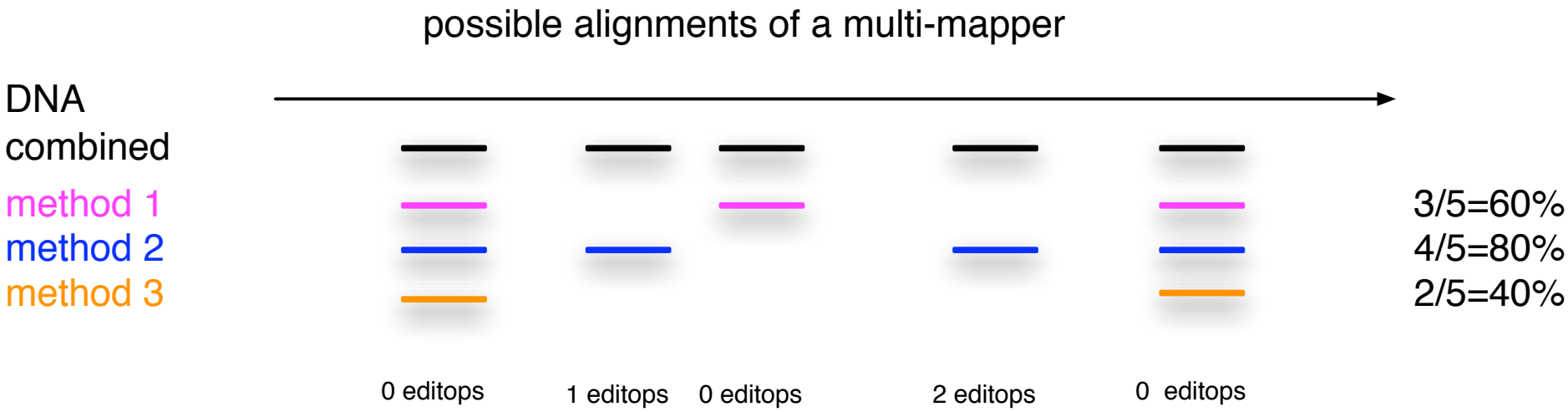
Comparing Alignments of Multi-mappers

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

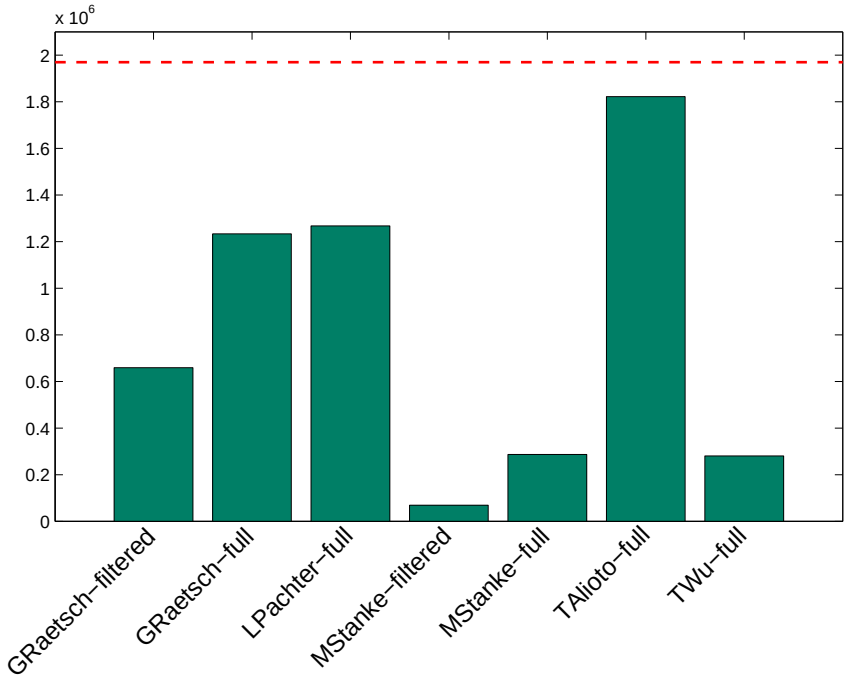


	0 mm	1 mm	2 mm
method 1	3/3	0/1	0/1
method 2	2/3	1/1	1/1
method 3	2/3	0/1	0/1

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

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

0 editops

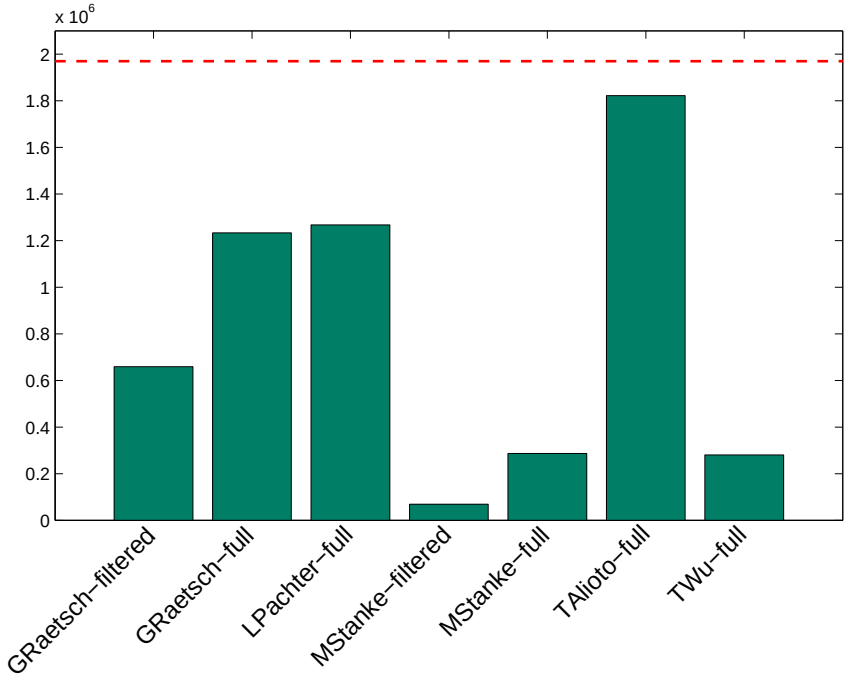


[left] Sensitivity of identifying all alignments with zero mismatches
(scaled by number of multi-mappers that align with 0 editops)

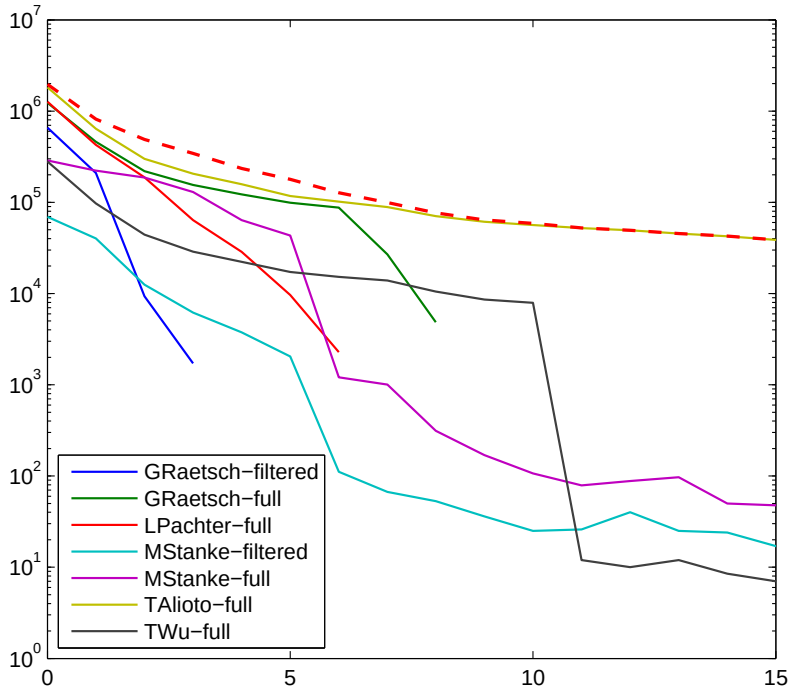
Red dotted line: maximal achievable score

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

0 editops



0-15 editops

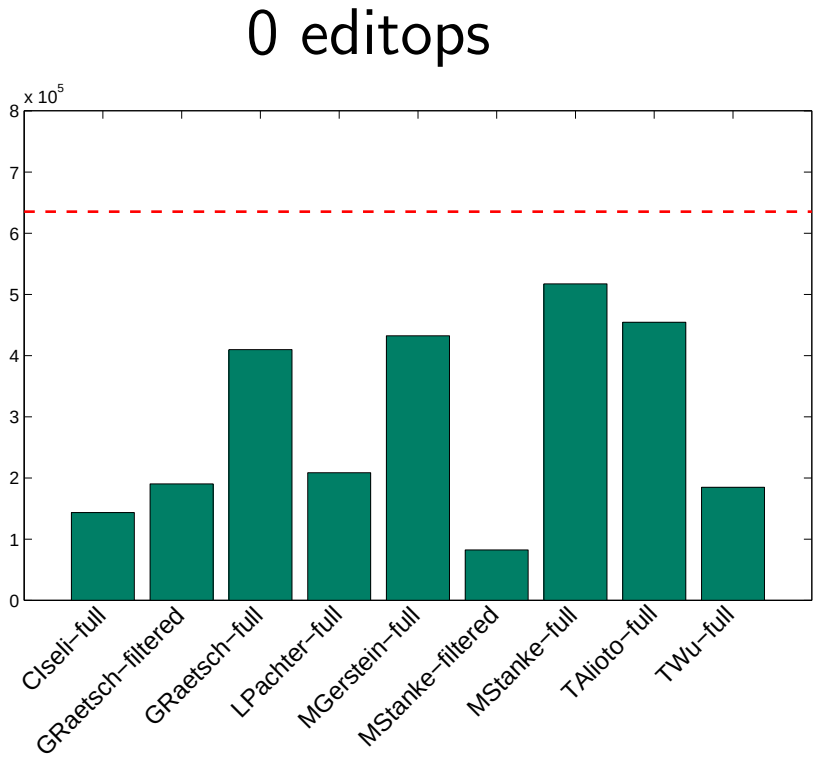


[left] Sensitivity of identifying all alignments with zero mismatches
(scaled by number of multi-mappers that align with 0 editops)

[right] For different number of mismatches.

Red dotted line: maximal achievable score

Comparing Alignments of Multi-mappers (human)



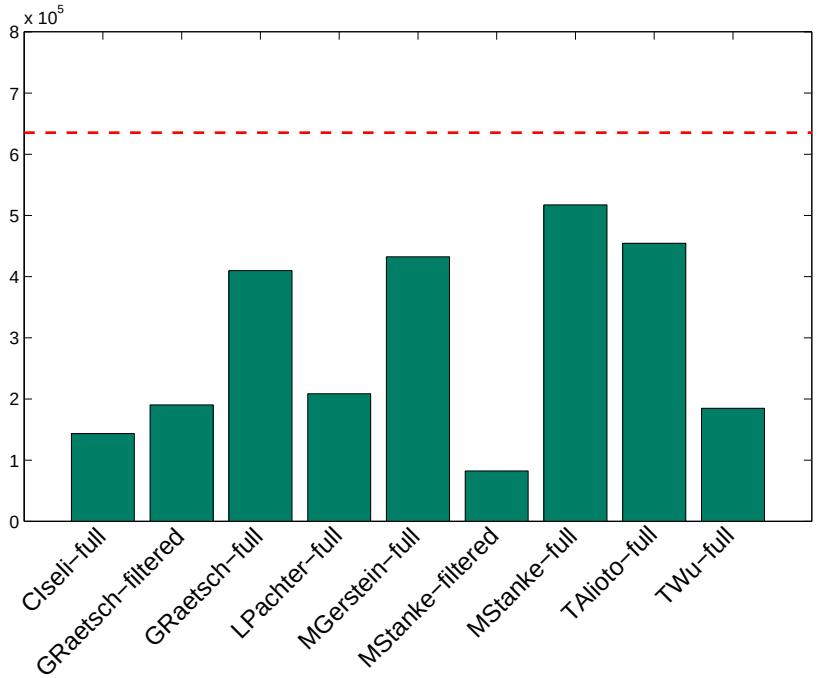
[left] Sensitivity of identifying all alignments with zero mismatches
(scaled by number of multi-mappers that align with 0 editops)

Red dotted line: maximal achievable score

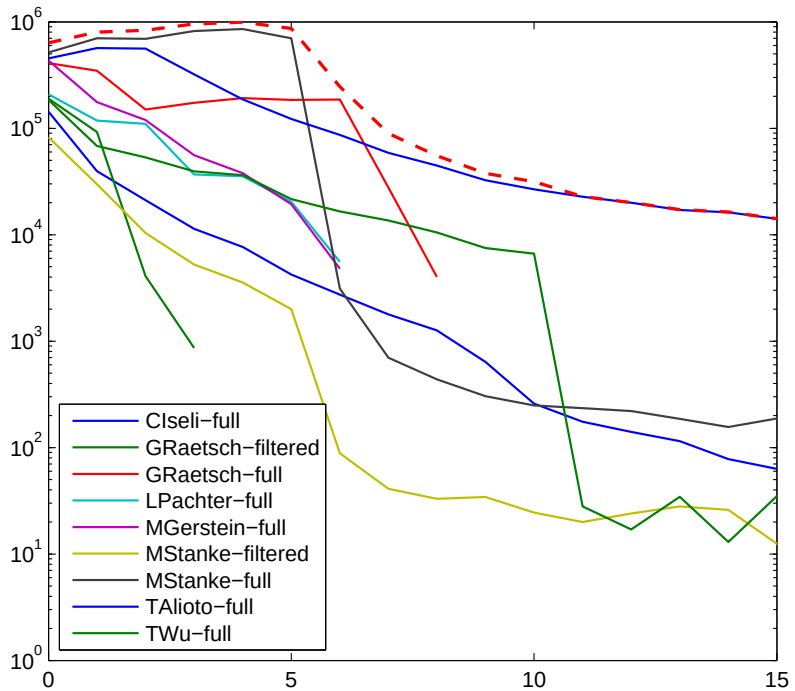
Comparing Alignments of Multi-mappers (human)



0 editops



0-15 editops



[left] Sensitivity of identifying all alignments with zero mismatches

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

[right] For different number of mismatches.

Red dotted line: maximal achievable score

Agreement of Intron Predictions with Annotation (*C. elegans*)

Strategy:

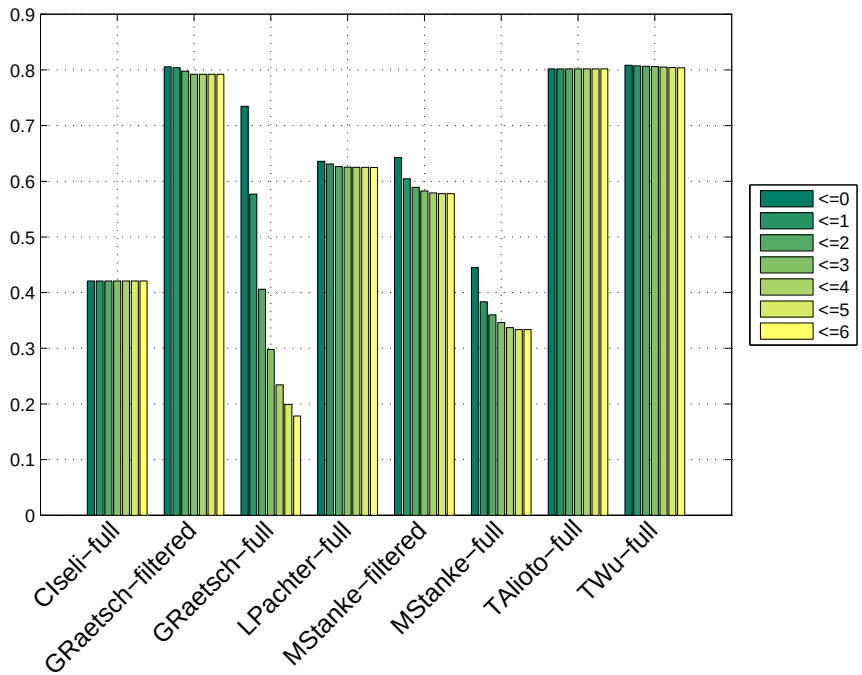
- Consider annotated introns as gold standard.
- Compute F-score of agreement of predicted introns in different submissions with gold standard.

Agreement of Intron Predictions with Annotation (*C. elegans*)

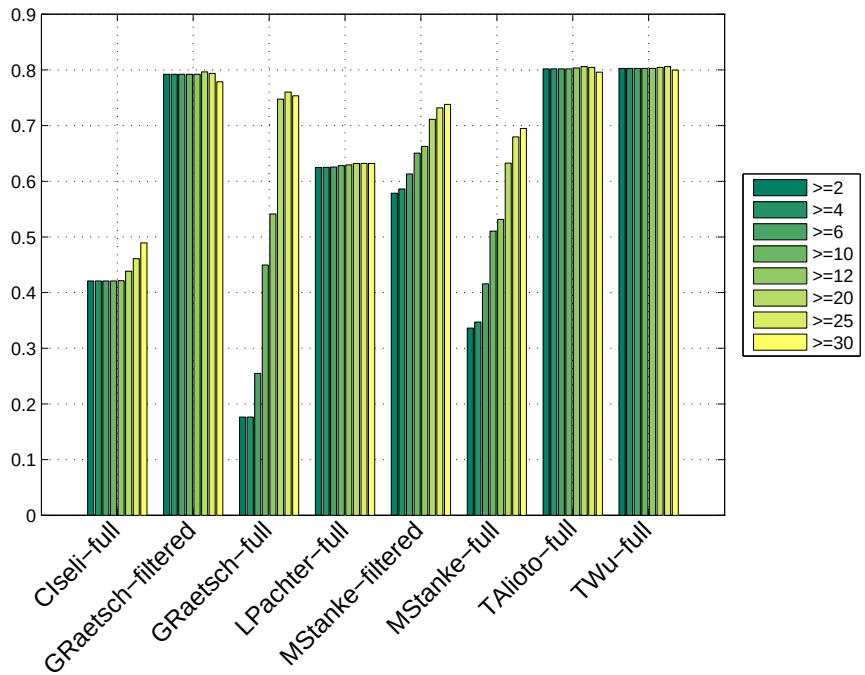
Strategy:

- Consider annotated introns as gold standard.
- Compute F-score of agreement of predicted introns in different submissions with gold standard.

Filtered by mismatches



Filtered by segment length

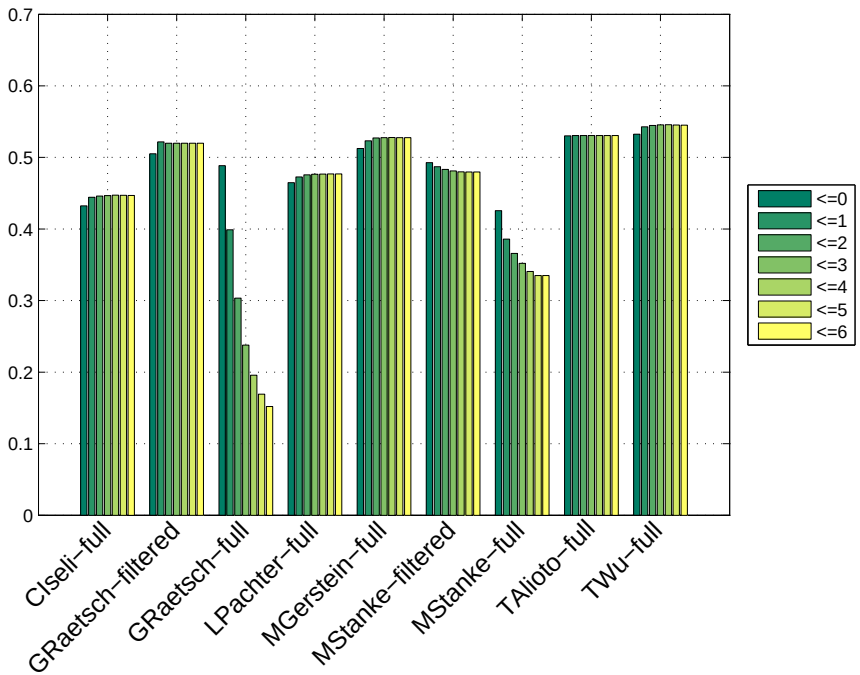


Agreement of Intron Predictions with Annotation (human)

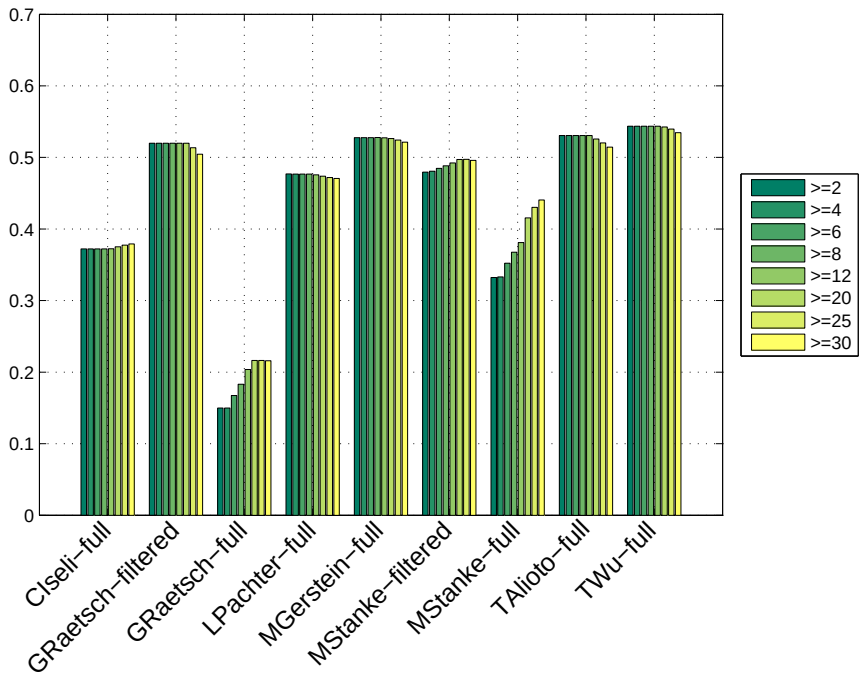
Strategy:

- Consider annotated introns as gold standard.
- Compute F-score of agreement of predicted introns in different submissions with gold standard.

Filtered by mismatches



Filtered by segment length

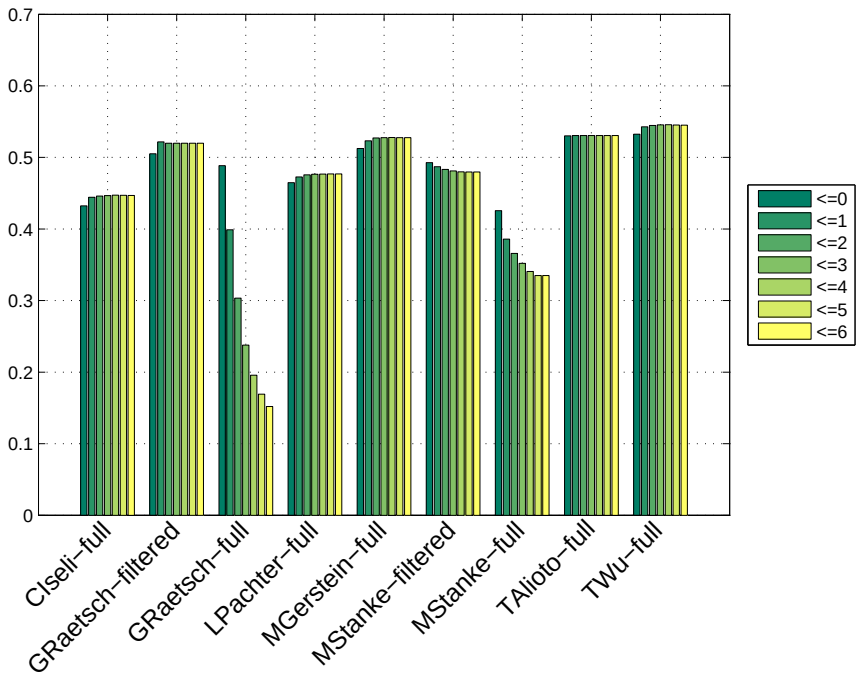


Agreement of Intron Predictions with Annotation (human)

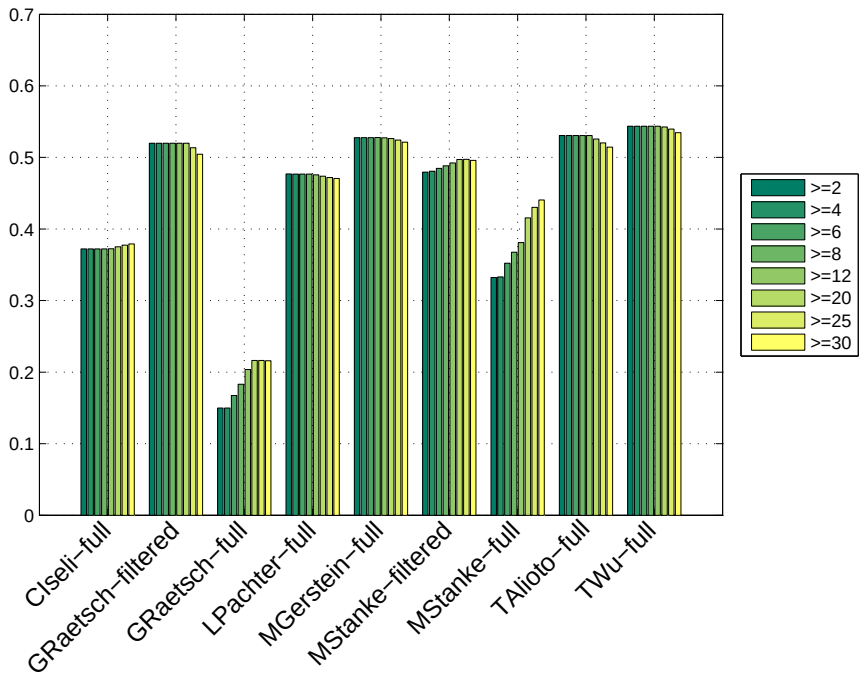
Strategy:

- Consider annotated introns as gold standard.
- Compute F-score of agreement of predicted introns in different submissions with gold standard.

Filtered by mismatches



Filtered by segment length



Filtering can make a big difference for intron prediction accuracy

Effect on Transcript Prediction



Which influence have alignments on transcript prediction accuracy?

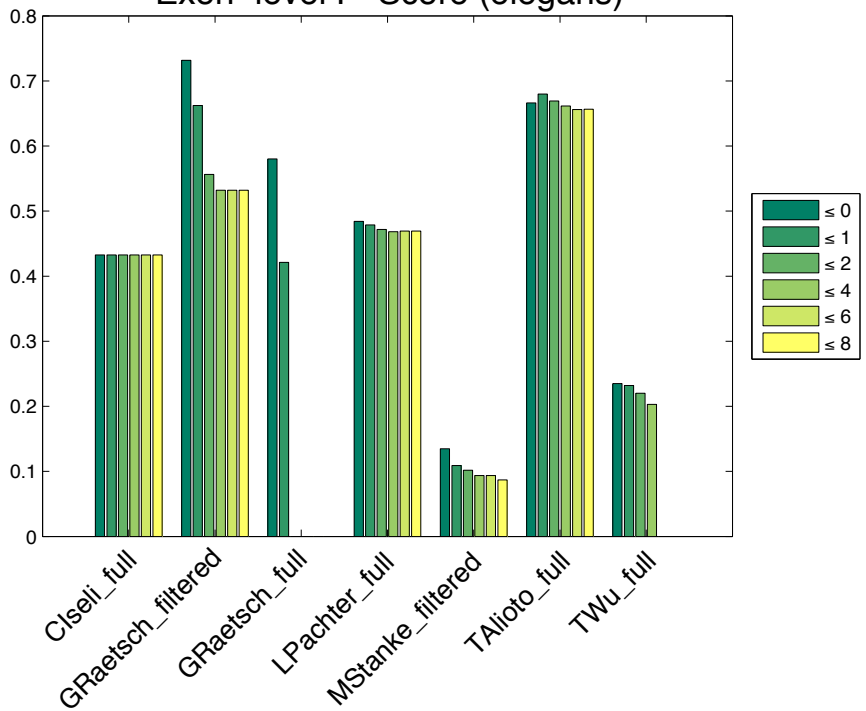
Strategy:

- Use a published transcript reconstruction technique, here Scripture (Broad), to predict transcripts. (Guttman et al., 2010)
- 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 scripture output
  - Only one transcript per locus (e.g., with maximal length or estimated abundance, see appendix)

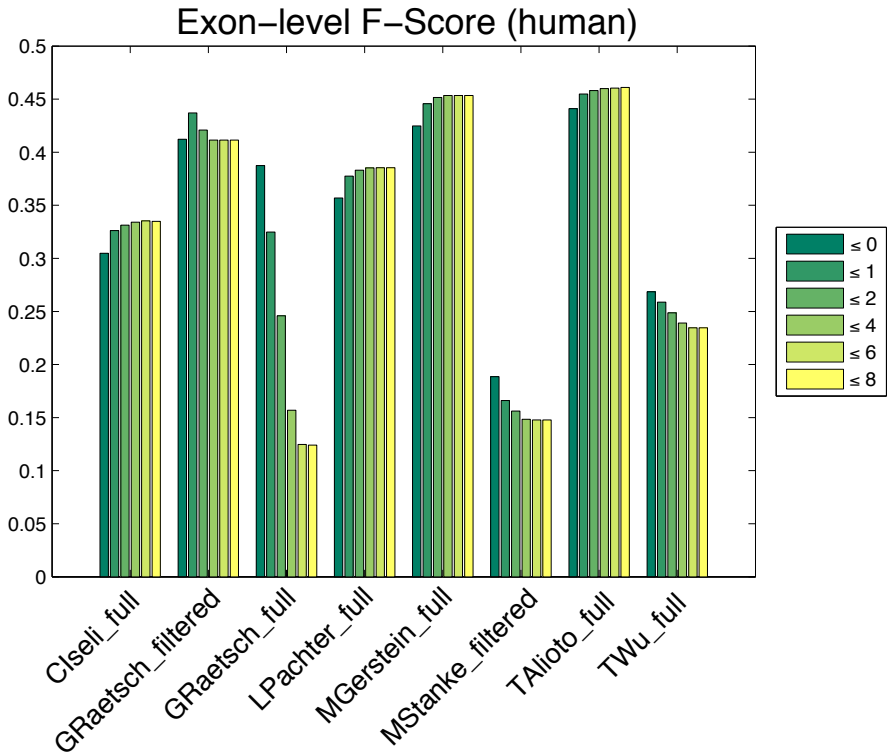
Transcript Prediction with Scripture

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

C. elegans



human



Filtering can significantly increase accuracy.

Scripture on less-filtered alignments may not terminate (GRaetsch_full, TAlioto, TWu).

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?

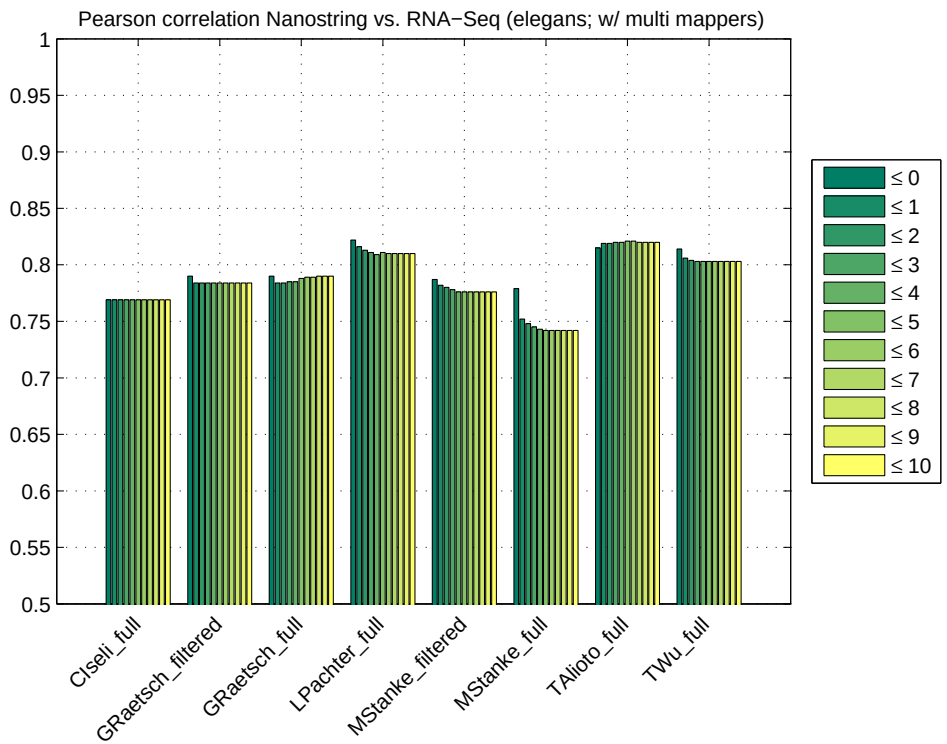
Strategy:

- 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)

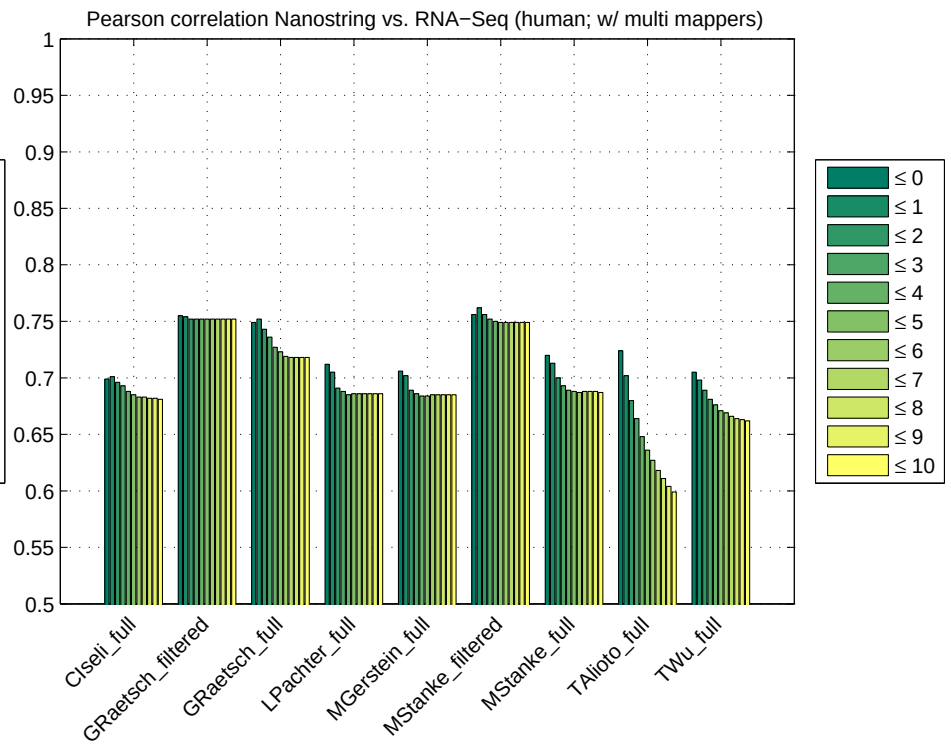
Correlation NanoString vs. RNA-seq

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

C. elegans



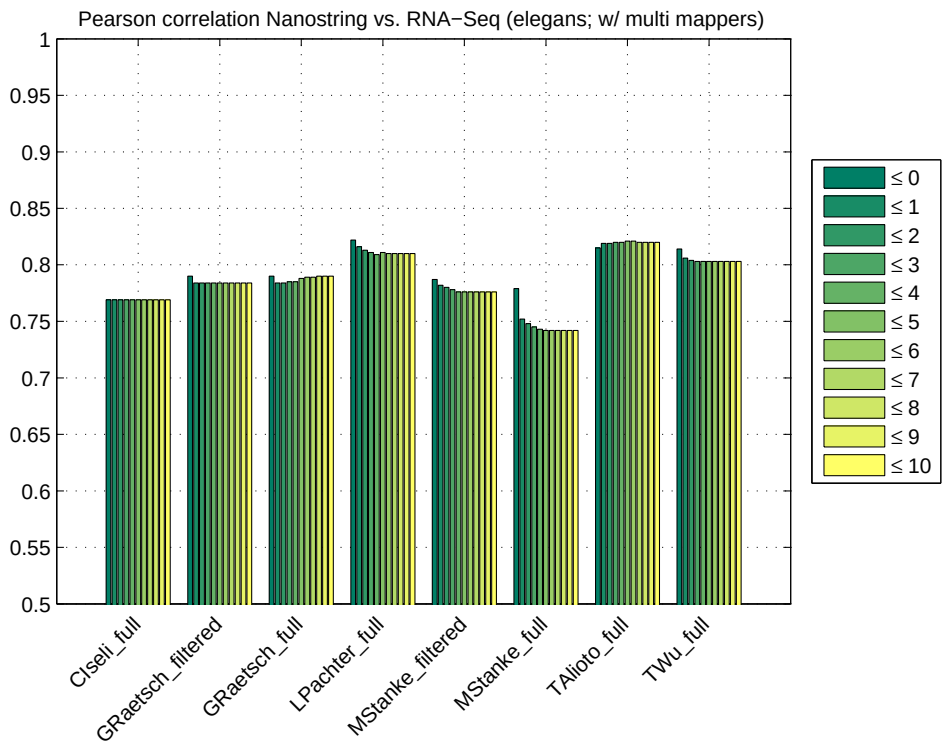
human



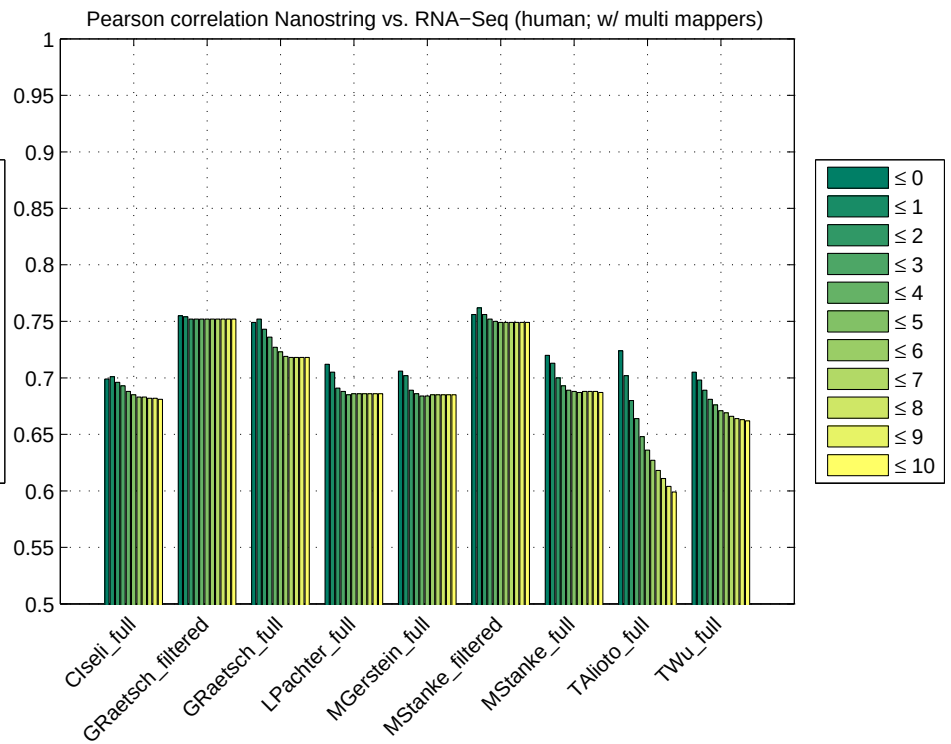
Correlation NanoString vs. RNA-seq

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

C. elegans



human



Mismatch filtering helps to improve the Pearson correlation.

Discussion/Conclusion



- Alignments are surprisingly different!
- Almost every alignment method has benefits in one of the tasks:
  - prediction of introns
  - identification of multiple mapping locations
  - transcript quantitation
  - exon/transcript prediction

Discussion/Conclusion

- Alignments are surprisingly different!
- Almost every alignment method has benefits in one of the tasks:
  - prediction of introns
  - identification of multiple mapping locations
  - transcript quantitation
  - exon/transcript prediction
- 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!
- Almost every alignment method has benefits in one of the tasks:
  - prediction of introns
  - identification of multiple mapping locations
  - transcript quantitation
  - exon/transcript prediction
- 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
- Alignment problem not solved and deserves further investigation
  - new RGASP round for alignment evaluations (end of 2010)
    - `http://tinyurl.com/RGASP2010`

Discussion/Conclusion



- Alignments are surprisingly different!
- Almost every alignment method has benefits in one of the tasks:
  - prediction of introns
  - identification of multiple mapping locations
  - transcript quantitation
  - exon/transcript prediction
- 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
- Alignment problem not solved and deserves further investigation
  - new RGASP round for alignment evaluations (end of 2010)
    - `http://tinyurl.com/RGASP2010`
- Slides online `http://tinyurl.com/talks-raetschlab`

Acknowledgements



Andre Kahles

Almost everything



Regina Bohnert

Quantitation



Paolo Ribeca@CNAG

Multi mappers



Jonas Behr

Support



Acknowledgements



Andre Kahles

Almost everything



Regina Bohnert

Quantitation



Paolo Ribeca@CNAG

Multi mappers



Jonas Behr

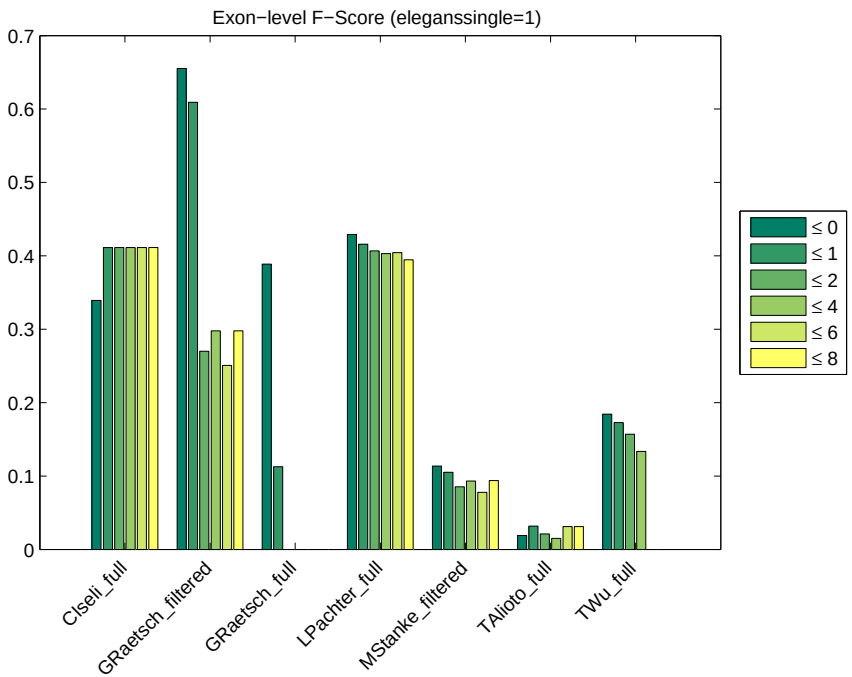
Support



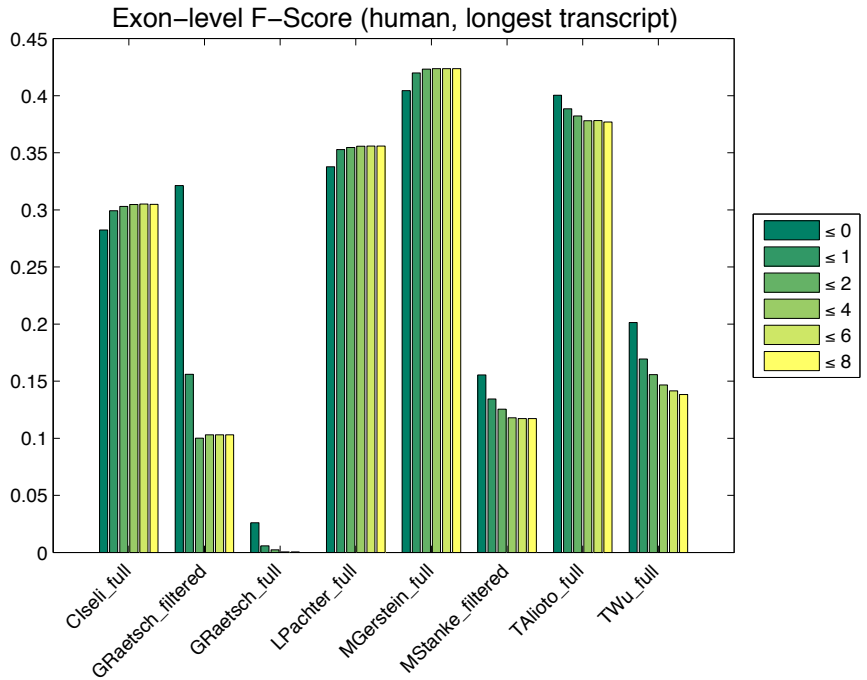
Transcript Prediction with Scripture

Exon-level F-Score of Scripture transcript predictions (longest transcripts only) relative to annotation.

C. elegans



human



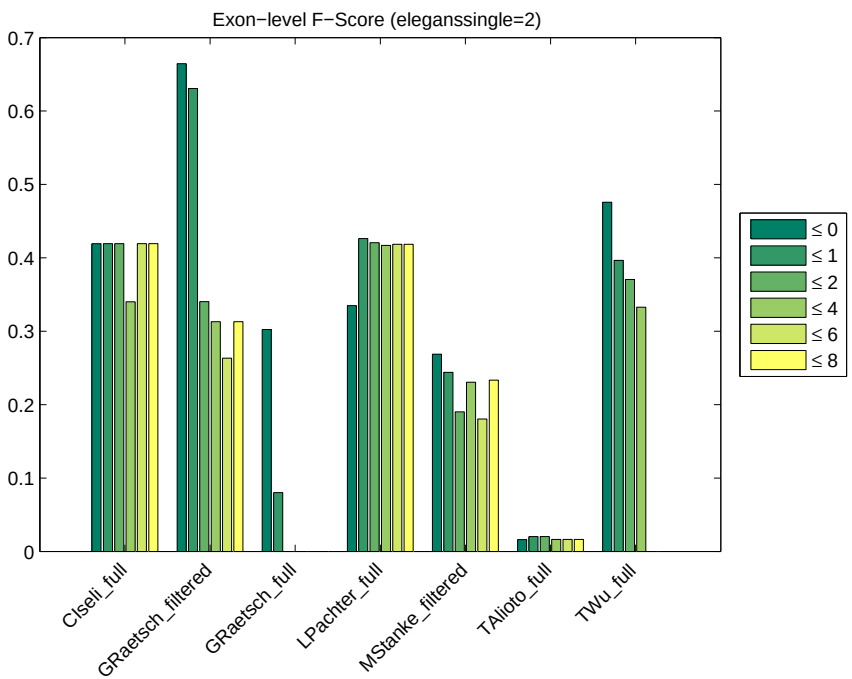
Filtering can significantly increase accuracy.

Scripture on less-filtered alignments may not terminate (GRaetsch-full, TAlioto, TWu).

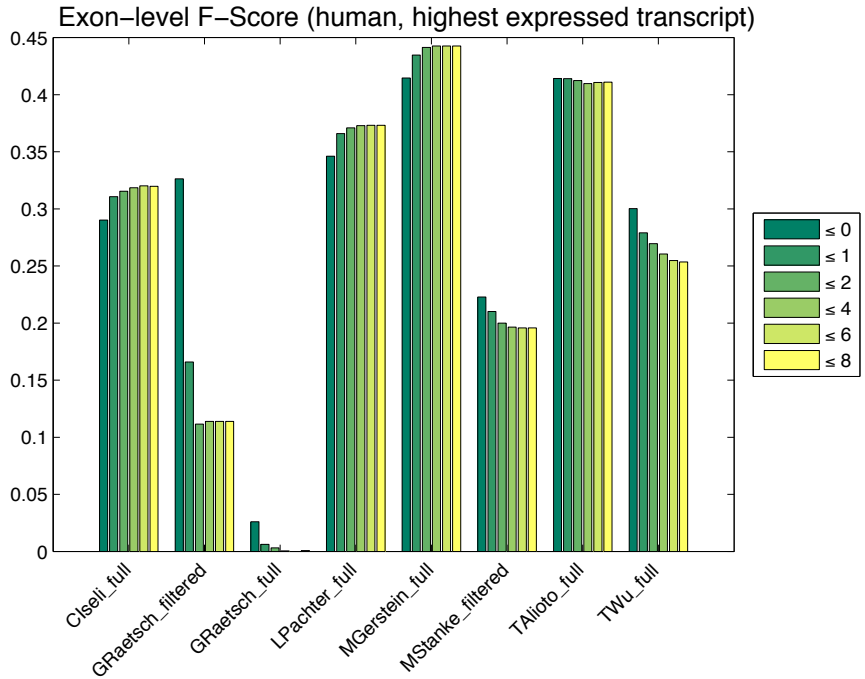
Transcript Prediction with Scripture

Exon-level F-Score of Scripture transcript predictions (highest expressed transcripts only) relative to annotation.

C. elegans



human



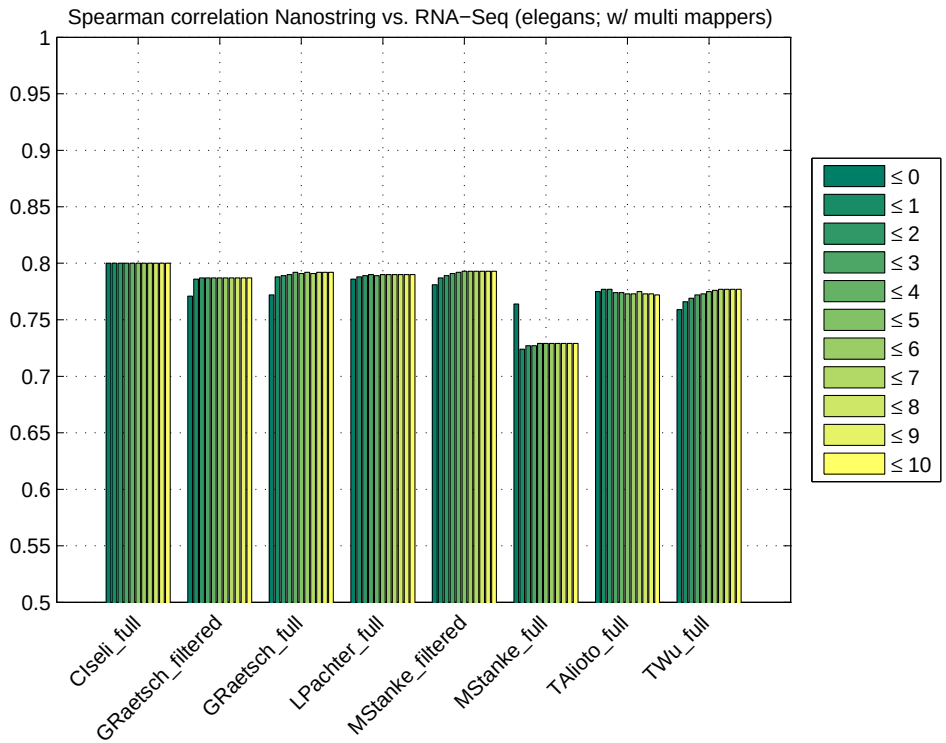
Filtering can significantly increase accuracy.

Scripture on less-filtered alignments may not terminate (GRaetsch-full, TAlioto, TWu).

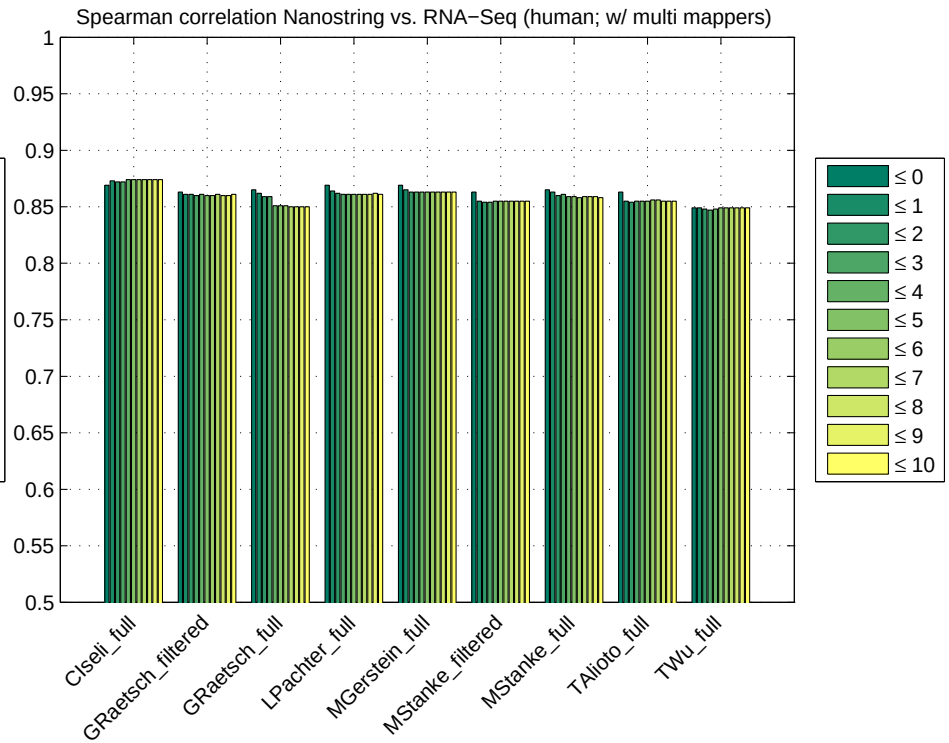
Correlation NanoString vs. RNA-seq

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

C. elegans



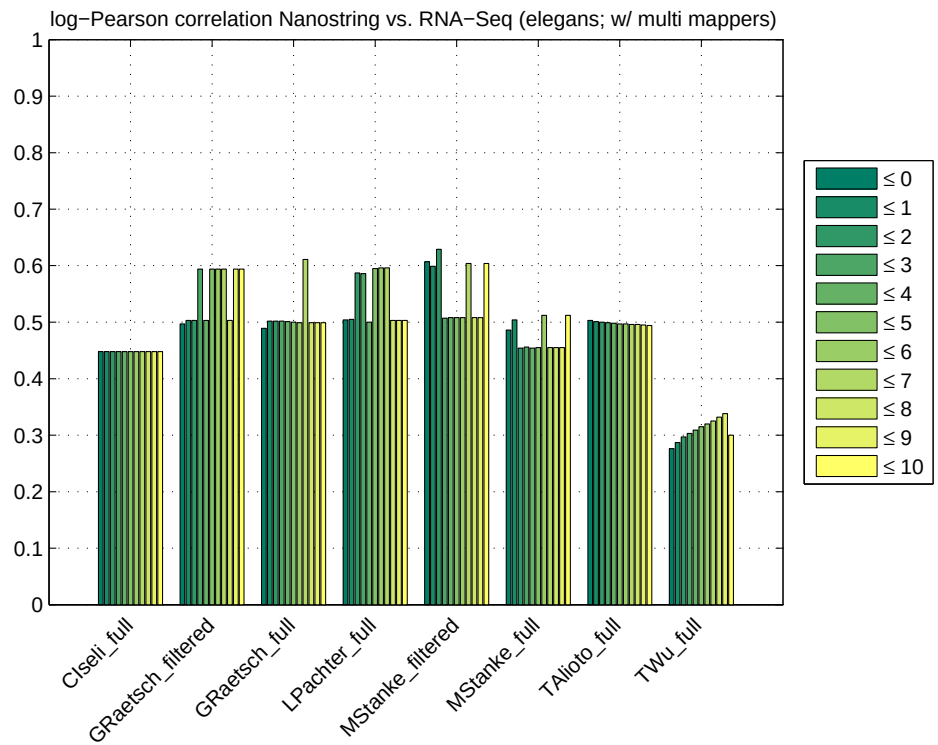
human



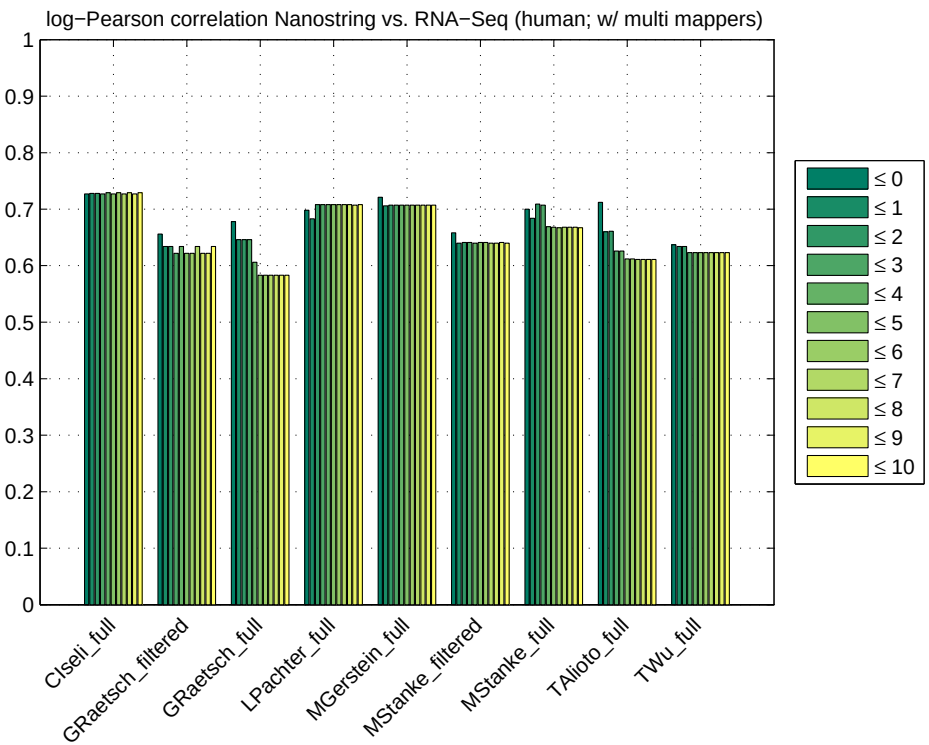
Correlation NanoString vs. RNA-seq

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

C. elegans



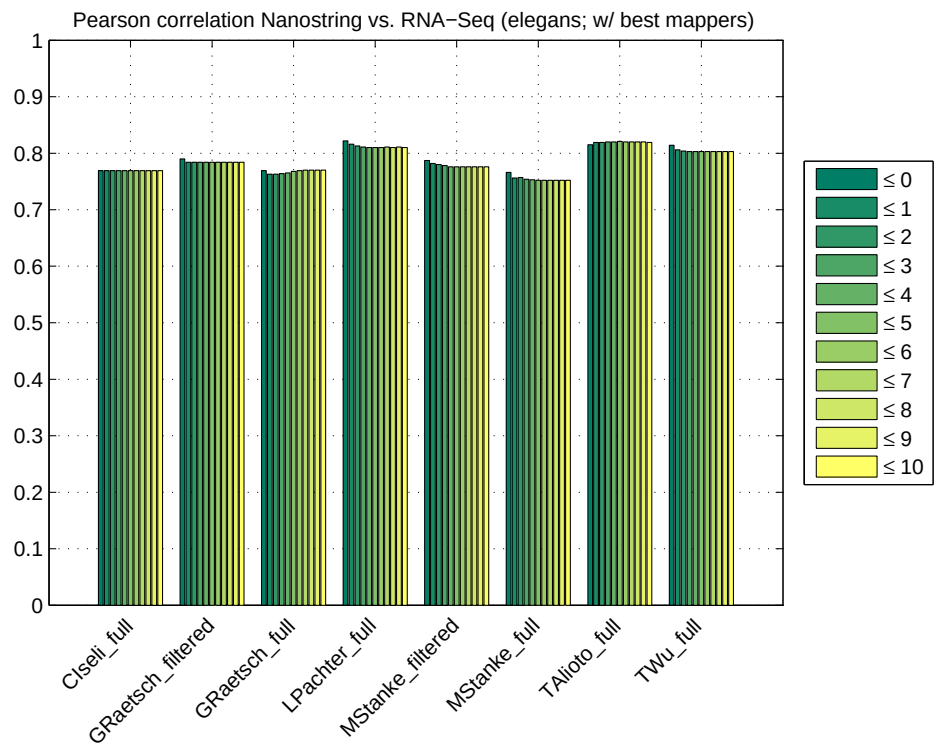
human



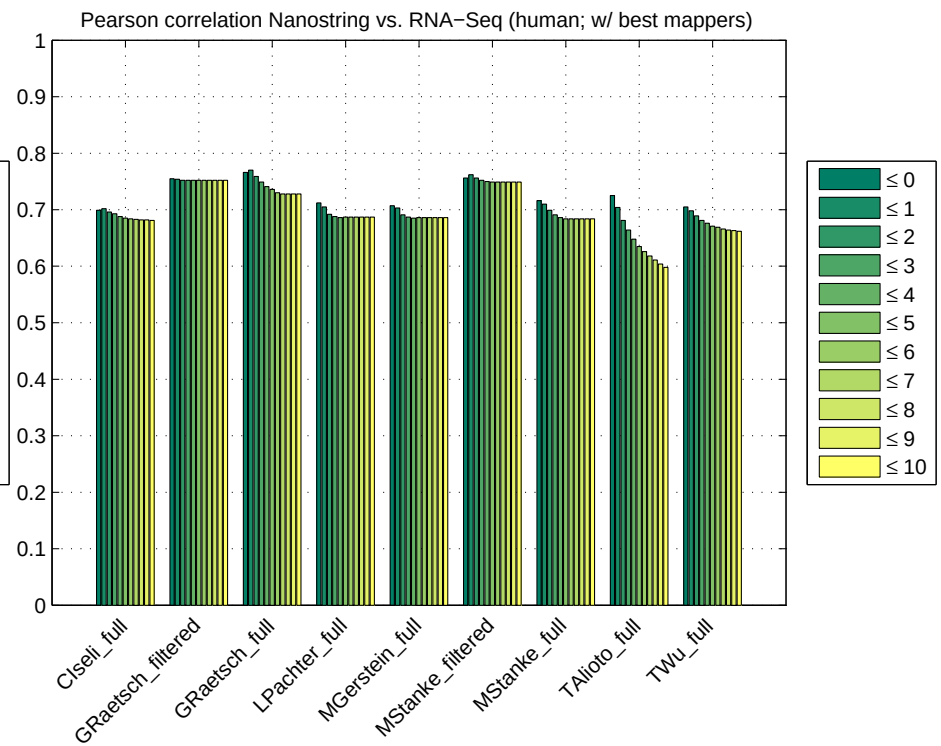
Correlation NanoString vs. RNA-seq

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

C. elegans



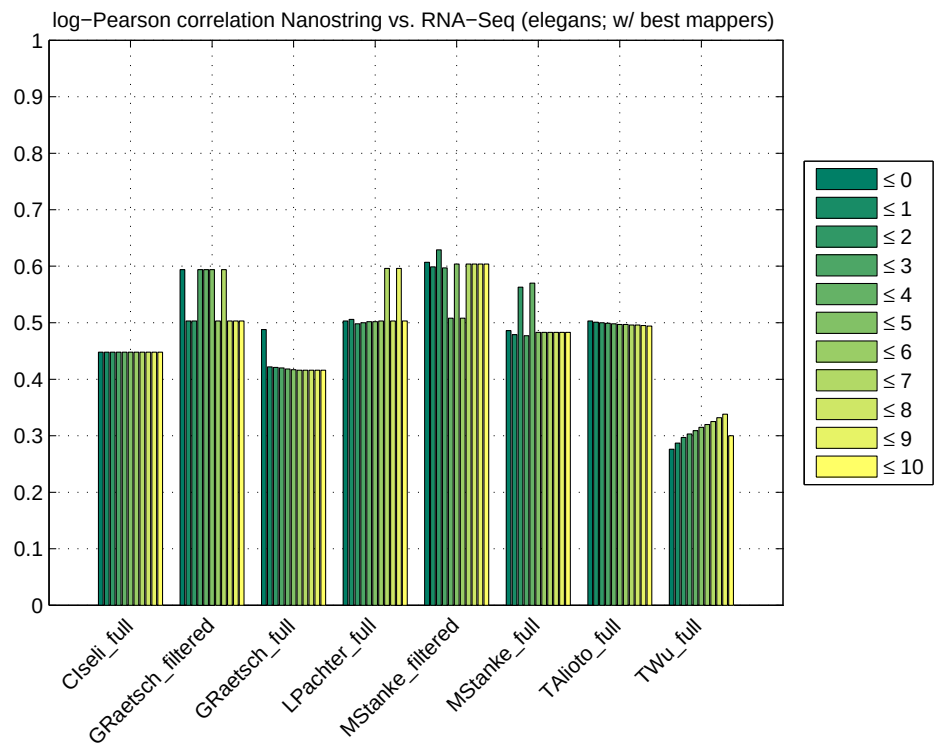
human



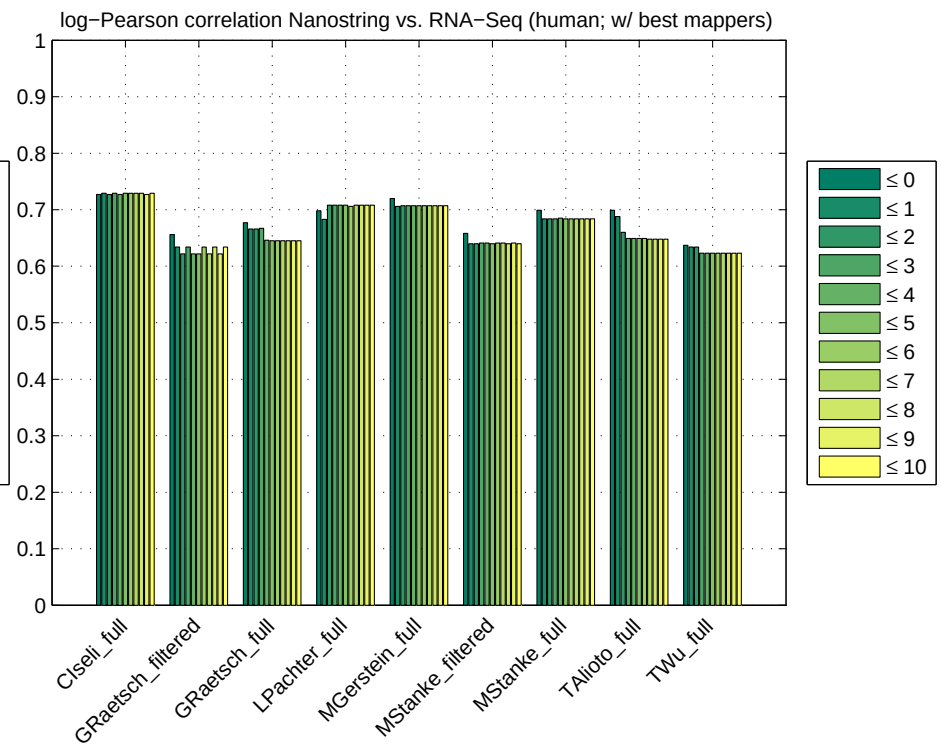
Correlation NanoString vs. RNA-seq

log-Pearson correlation between NanoString and RNA-seq quantitation for different mismatch filters and best mappers only.

C. elegans



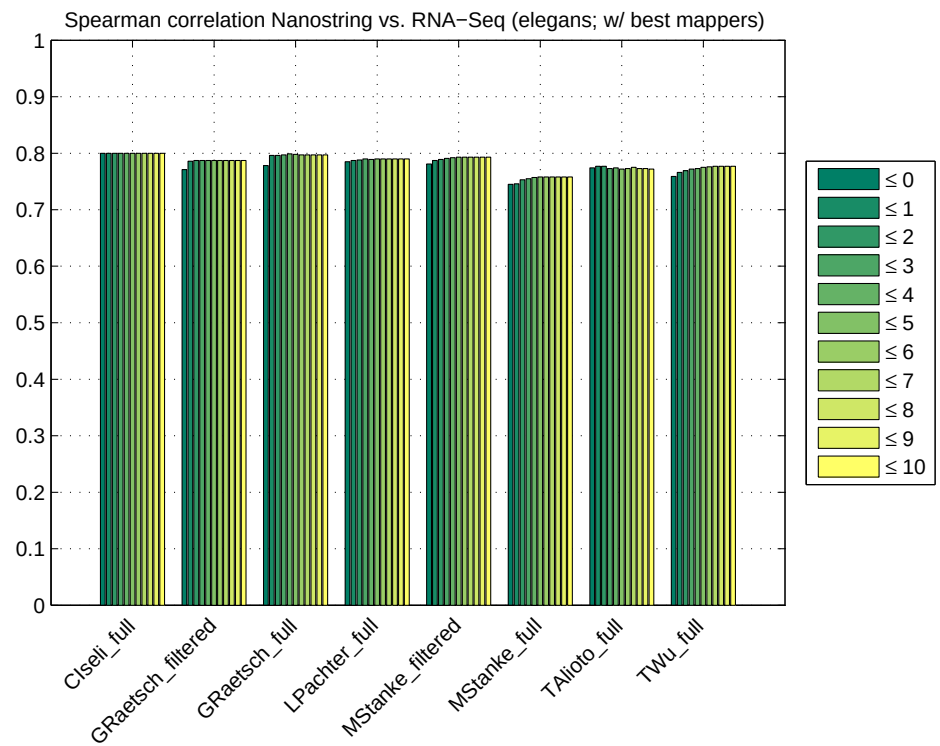
human



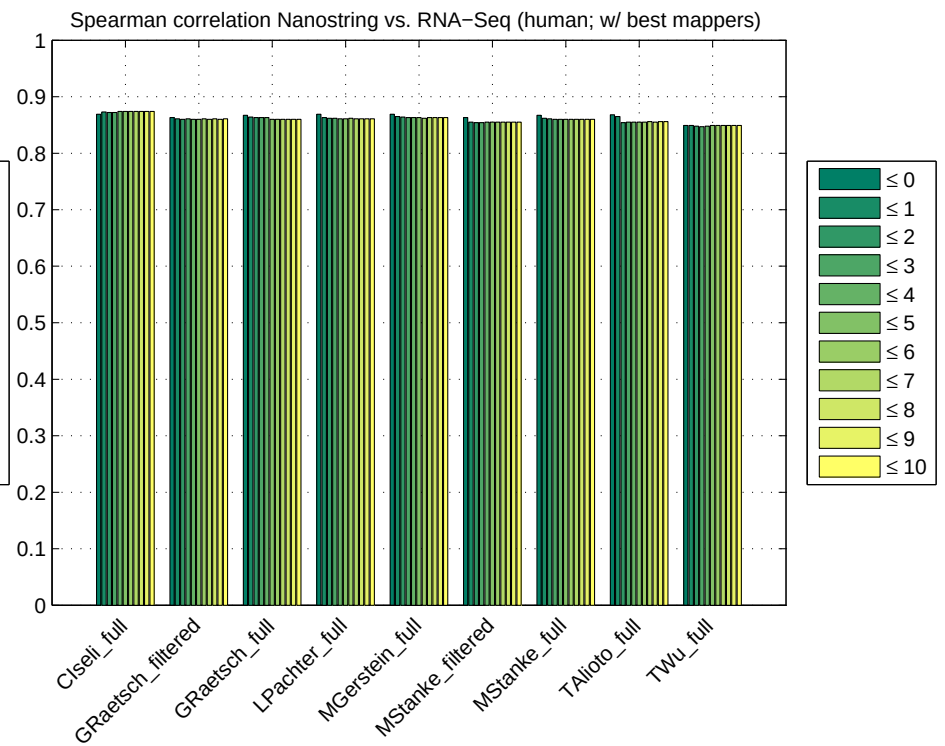
Correlation NanoString vs. RNA-seq

Spearman correlation between NanoString and RNA-seq quantitation for different mismatch filters and best mappers only.

C. elegans



human



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.

Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol*, 28(5): 503–10, May 2010. doi: 10.1038/nbt.1633.

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.