Analysis Tools for RNA-seq and Isoform Characterization

Slides: `bit.ly/1DeRjGM`

Gunnar Rätsch

Biomedical Data Science Group
Computational Biology Center
Memorial Sloan Kettering Cancer Center

@gxr #RNA #MMR #SplAdder #riboDiff #Cancer

Memorial Sloan-Kettering
Cancer Center

cBio@MSKCC

Biomedical Data Sciences Group

Facts

- Cost of collecting data drops, amounts increase exponentially.
- We have *more data than accurate algorithms*.

Group's research

- **Data Science** *Algorithms, Models & Tools*
  - ⇝ *Machine Learning,*
  - ⇝ *Bioinformatics.*
- **Biology & Medicine** *Problem Setting & Goals*
  - ⇝ *RNA processing regulation,*
  - ⇝ *Clinical data analysis.*

Biomedical Data Sciences Group

Facts

- Cost of collecting data drops, amounts increase exponentially.
- We have *more data than accurate algorithms*.

Group's research

- **Data Science** *Algorithms, Models & Tools*
  - ⇝ *Machine Learning,*
  - ⇝ *Bioinformatics.*
- **Biology & Medicine** *Problem Setting & Goals*
  - ⇝ *RNA processing regulation,*
  - ⇝ *Clinical data analysis.*

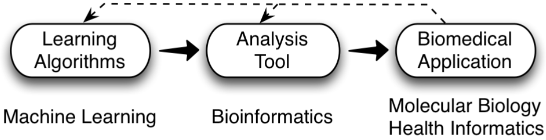
Biomedical Data Sciences Group

Facts

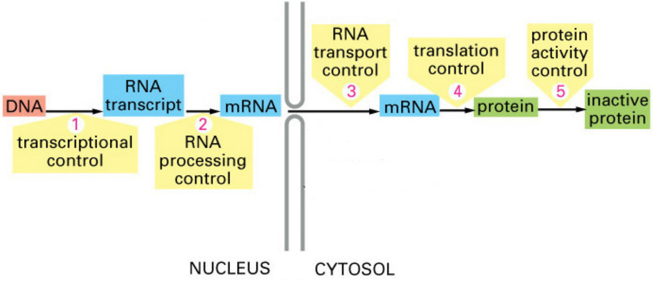
- Cost of collecting data drops, amounts increase exponentially.
- We have *more data than accurate algorithms*.

Group's research

- **Data Science** — *Algorithms, Models & Tools*
  - ⇝ *Machine Learning,*
  - ⇝ *Bioinformatics.*
- **Biology & Medicine** — *Problem Setting & Goals*
  - ⇝ *RNA processing regulation,*
  - ⇝ *Clinical data analysis.*



Learning About the Central Dogma



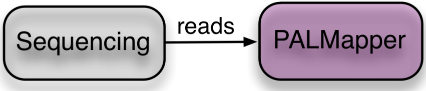
Goal: Learn to predict what these processes accomplish:

- Given the DNA, . . . , predict all gene products

$$f(\text{DNA}, 1\ 2\ 3) = \text{RNA} \qquad g(\text{RNA}, 4\ 5) = \text{protein}$$

- Estimating f, g amounts to cracking the codes of transcription, epigenetics, splicing, . . .

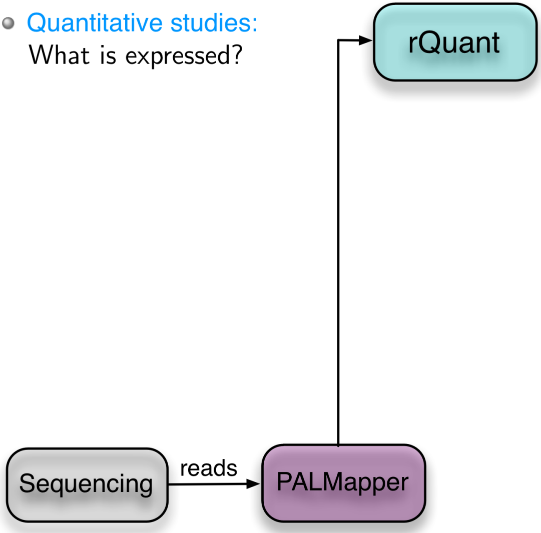
RNA-seq based Transcriptome Characterization



Accurate spliced alignments
[Bona et al., 2008, Jean et al., 2010]

RNA-seq based Transcriptome Characterization

- Quantitative studies:
  What is expressed?

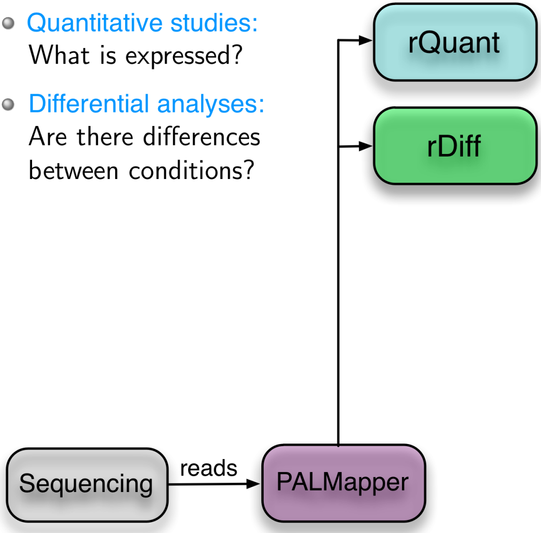


Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

Accurate spliced alignments
[Bona et al., 2008, Jean et al., 2010]

RNA-seq based Transcriptome Characterization

- Quantitative studies:
  What is expressed?
- Differential analyses:
  Are there differences between conditions?



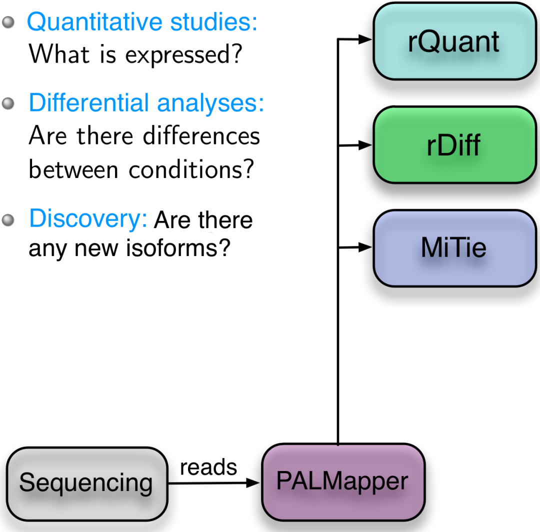
Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

Tests for differential isoform expression
[Drewe et al., 2013]

Accurate spliced alignments
[Bona et al., 2008, Jean et al., 2010]

RNA-seq based Transcriptome Characterization

- Quantitative studies: What is expressed?
- Differential analyses: Are there differences between conditions?
- Discovery: Are there any new isoforms?



Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

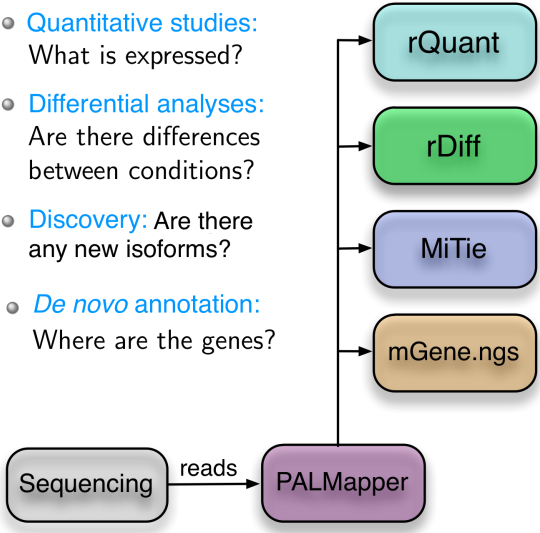
Tests for differential isoform expression
[Drewe et al., 2013]

Simultaneous transcript identification & quantitation
[Behr et al., 2013]

Accurate spliced alignments
[Bona et al., 2008, Jean et al., 2010]

RNA-seq based Transcriptome Characterization

- Quantitative studies:
  What is expressed?
- Differential analyses:
  Are there differences between conditions?
- Discovery: Are there any new isoforms?
- *De novo* annotation:
  Where are the genes?



Sequencing

reads

PALMapper

rQuant

rDiff

MiTie

mGene.ngs

Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

Tests for differential isoform expression
[Drewe et al., 2013]

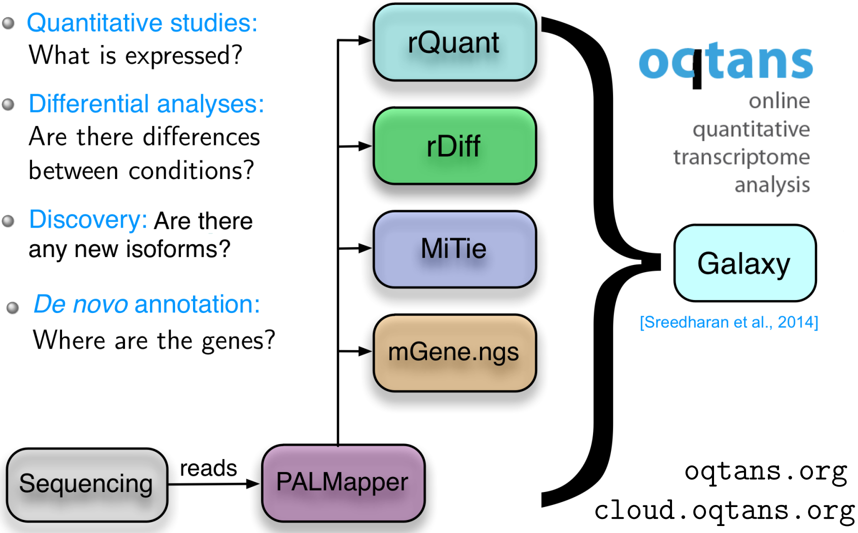
Simultaneous transcript identification & quantitation
[Behr et al., 2013]

Gene finding with RNA-seq evidence
[Behr et al., 2010, 2013, Gan et al., 2011]

Accurate spliced alignments
[Bona et al., 2008, Jean et al., 2010]

RNA-seq based Transcriptome Characterization

- Quantitative studies:
  What is expressed?
- Differential analyses:
  Are there differences between conditions?
- Discovery: Are there any new isoforms?
- *De novo* annotation:
  Where are the genes?



Sequencing

reads

PALMapper

rQuant

rDiff

MiTie

mGene.ngs

oqtans

online quantitative transcriptome analysis

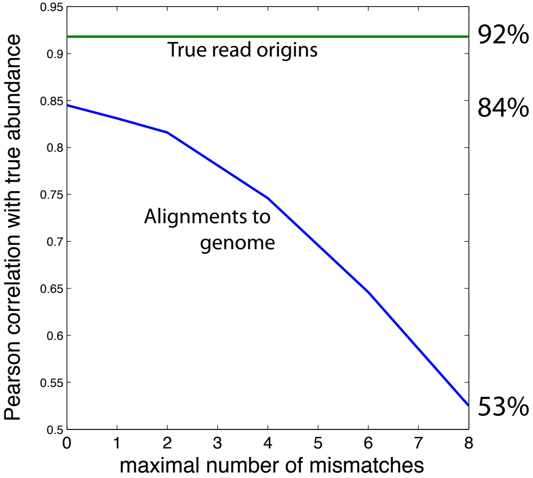
Galaxy

[Sreedharan et al., 2014]

oqtans.org

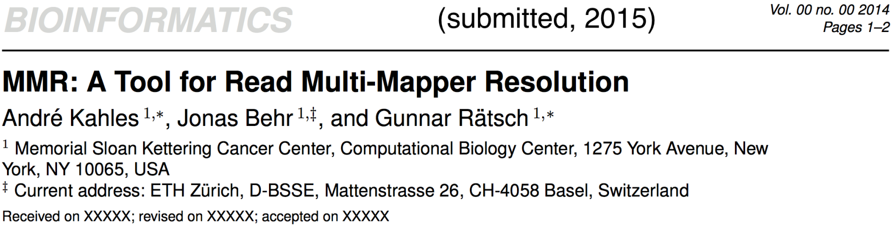
cloud.oqtans.org

Transcript Quantitation and Dependence on Alignments



False alignments, multi-mappers etc. lead to weaker results

Simulated human reads from transcripts of known abundance (Fluxsimulator, [Sammeth, 2009]), 3% error rate, alignment w/ PALMapper [Jean et al., 2010], quantification w/ rQuant [Bohnert et al., 2009], Person correlation over considered transcripts.



BIOINFORMATICS (submitted, 2015) *Vol. 00 no. 00 2014*
Pages 1–2

MMR: A Tool for Read Multi-Mapper Resolution

André Kahles [1,*], Jonas Behr [1,‡], and Gunnar Rätsch [1,*]

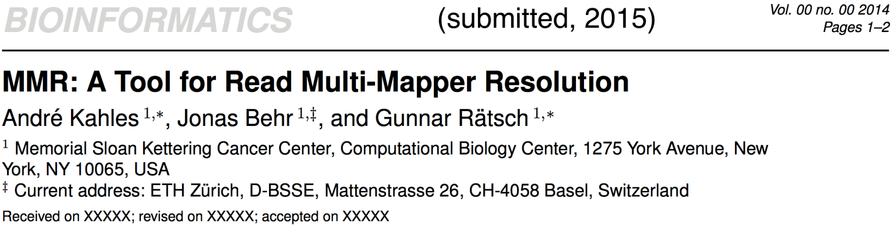
[1] Memorial Sloan Kettering Cancer Center, Computational Biology Center, 1275 York Avenue, New York, NY 10065, USA

[‡] Current address: ETH Zürich, D-BSSE, Mattenstrasse 26, CH-4058 Basel, Switzerland

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv `dx.doi.org/10.1101/017103`

- *Efficient* BAM file postprocessor for RNA- & DNA-seq
  - 100M alignments in 20 minutes (10 threads)
- Suitable for **large-scale projects**
- **Improved accuracy** for transcript quantification and prediction
- Open Source `bioweb.me/mmr` (C++)



BIOINFORMATICS (submitted, 2015) Vol. 00 no. 00 2014
Pages 1–2

MMR: A Tool for Read Multi-Mapper Resolution

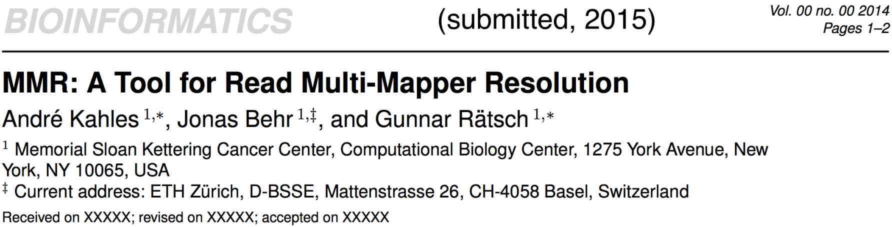
André Kahles [1,*], Jonas Behr [1,‡], and Gunnar Rätsch [1,*]

[1] Memorial Sloan Kettering Cancer Center, Computational Biology Center, 1275 York Avenue, New York, NY 10065, USA
[‡] Current address: ETH Zürich, D-BSSE, Mattenstrasse 26, CH-4058 Basel, Switzerland

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv `dx.doi.org/10.1101/017103`

- *Efficient* BAM file postprocessor for RNA- & DNA-seq
  - 100M alignments in 20 minutes (10 threads)
- Suitable for **large-scale projects**
- **Improved accuracy** for transcript quantification and prediction
- Open Source `bioweb.me/mmr` (C++)



BIOINFORMATICS (submitted, 2015) Vol. 00 no. 00 2014
Pages 1–2

MMR: A Tool for Read Multi-Mapper Resolution

André Kahles [1,*], Jonas Behr [1,‡], and Gunnar Rätsch [1,*]

[1] Memorial Sloan Kettering Cancer Center, Computational Biology Center, 1275 York Avenue, New York, NY 10065, USA
[‡] Current address: ETH Zürich, D-BSSE, Mattenstrasse 26, CH-4058 Basel, Switzerland

Received on XXXXX; revised on XXXXX; accepted on XXXXX

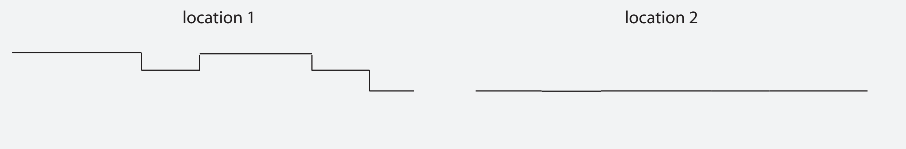
bioRxiv `dx.doi.org/10.1101/017103`

- *Efficient* BAM file postprocessor for RNA- & DNA-seq
  - 100M alignments in 20 minutes (10 threads)
- Suitable for **large-scale projects**
- **Improved accuracy** for transcript quantification and prediction
- Open Source `bioweb.me/mmr` (C++)

Multiple Mapper Resolution

Principle (Iterated over all reads, N times)

- Use the change of local coverage around read mapping ...
- ... and use its smoothness to identify "better" mapping location

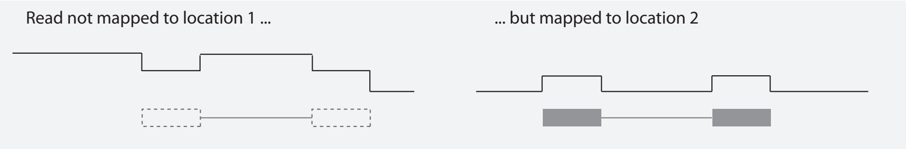


▭ Coverage

Multiple Mapper Resolution

Principle (Iterated over all reads, N times)

- Use the change of local coverage around read mapping ...
- ... and use its smoothness to identify “better” mapping location



Read not mapped to location 1 ...

... but mapped to location 2

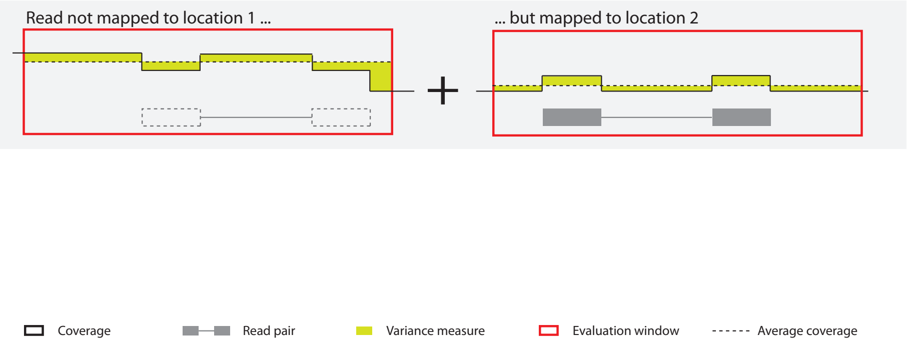
▭ Coverage

▬—▬ Read pair

Multiple Mapper Resolution

Principle (Iterated over all reads, N times)

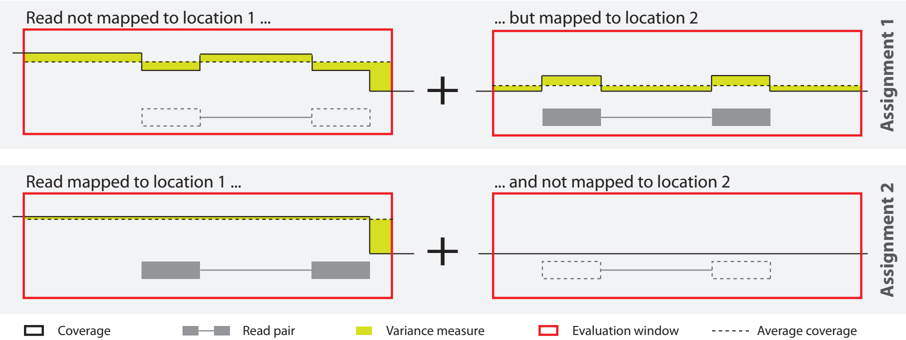
- Use the change of local coverage around read mapping ...
- ... and use its smoothness to identify "better" mapping location



Multiple Mapper Resolution

Principle (Iterated over all reads, N times)

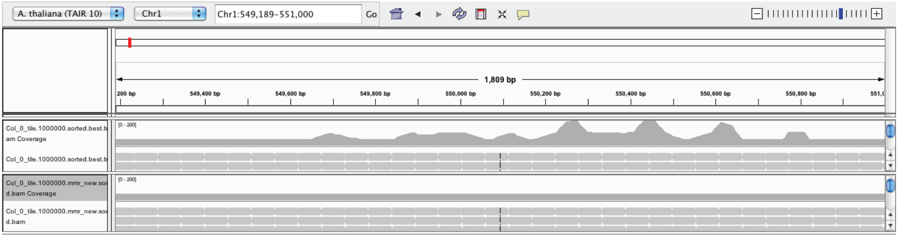
- Use the change of local coverage around read mapping ...
- ... and use its smoothness to identify "better" mapping location


Read not mapped to location 1 ...
... but mapped to location 2
Assignment 1
+
Read mapped to location 1 ...
... and not mapped to location 2
Assignment 2
+
Coverage
Read pair
Variance measure
Evaluation window
Average coverage

Multiple Mapper Resolution

Results for simulated DNA-seq

- Smooths coverage as expected on an artificial dataset



A. thaliana (TAIR 10)

Chr1

Chr1:549,189-551,000

Go

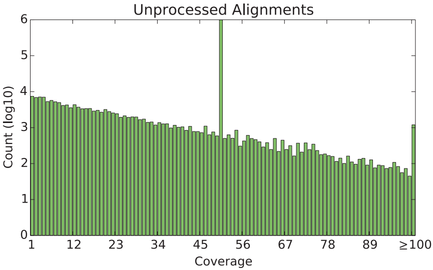
1,809 bp
200 bp
549,400 bp
549,600 bp
549,800 bp
550,000 bp
550,200 bp
550,400 bp
550,600 bp
550,800 bp
551,(
Col_0_tile.1000000.sorted.best.bam Coverage
Col_0_tile.1000000.sorted.best.b
[0 - 200]
Col_0_tile.1000000.mmr_new.sord.bam Coverage
Col_0_tile.1000000.mmr_new.sord.bam
[0 - 200]

Simulated reads from tiling a part of *A. thaliana* genome, alignment w/ PALMapper [Jean et al., 2010] (with -a option), visualization with IGV [Robinson et al., 2011].

Multiple Mapper Resolution

Results for simulated DNA-seq

- Smooths coverage as expected on an artificial dataset

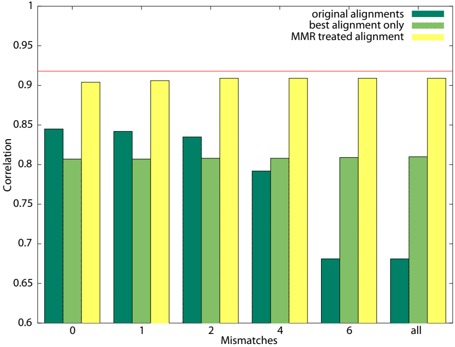


Simulated reads from tiling a part of *A. thaliana* genome, alignment w/ PALMapper [Jean et al., 2010] (with -a option), visualization with IGV [Robinson et al., 2011].

Multiple Mapper Resolution

Results for simulated RNA-seq

- Improves performance of transcript quantification



Simulated reads (75nt) from subset of human annotated transcripts with Fluxsimulator [Sammeth, 2009], PALMapper alignments [Jean et al., 2010], rQuant quantitation Bohnert et al. [2009], Pearson correlation over all considered transcripts.

24 Hills Road, Cambridge CB2 1JP
P 01223 855346
W elifesciences.org
T @elife_sciences

(Accepted, 2015)

DNA methylation variation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation

Manu Dubin (Vienna Biocenter), Pei Zhang (Vienna Biocenter), Dazhe Meng (Vienna Biocenter), Marie-Stanislas Remigereau (University of Southern California), Edward Osborne (University of Utah), Francesco Paolo Casale (Wellcome Trust Genome Campus), Philip Drewe (Max Planck Society), André Kahles (Max Planck Society), Geraldine Jean (Max Planck Society), Bjarni Vilhjálmsson (Vienna Biocenter), Joanna Jagoda (Vienna Biocenter), Selen Irez (Vienna Biocenter), Viktor Voronin (Vienna Biocenter), Qiang Song (University of Southern California), Quan Long (Vienna Biocenter), Gunnar Rätsch (Max Planck Society), Oliver Stegle (Wellcome Trust Genome Campus), Richard Clark (University of Utah), and Magnus Nordborg (Vienna Biocenter)

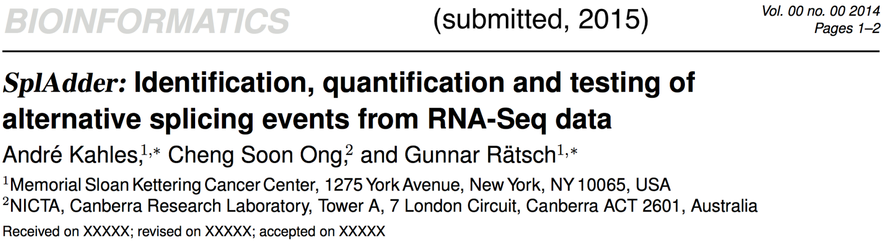
LARGE-SCALE BIOLOGY ARTICLE

Nonsense-Mediated Decay of Alternative Precursor mRNA Splicing Variants Is a Major Determinant of the *Arabidopsis* Steady State Transcriptome[C][W]

Gabriele Drechsel,[a,1] André Kahles,[b,1] Anil K. Kesarwani,[a] Eva Stauffer,[a,2] Jonas Behr,[b] Philipp Drewe,[b] Gunnar Rätsch,[b] and Andreas Wachter[a,3]

[a] Center for Plant Molecular Biology, University of Tübingen, 72076 Tuebingen, Germany

[b] Computational Biology Center, Sloan-Kettering Institute, New York, New York 10065



BIOINFORMATICS (submitted, 2015) *Vol. 00 no. 00 2014*
Pages 1–2

SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data

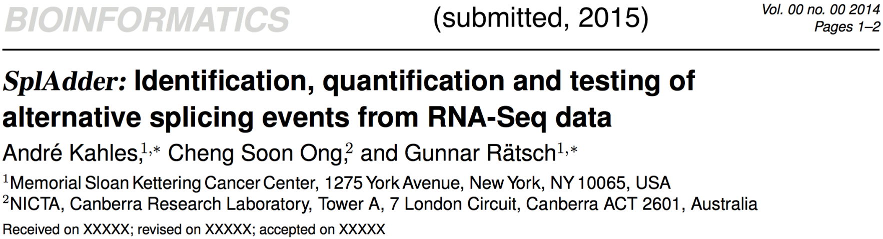
André Kahles,[1,*] Cheng Soon Ong,[2] and Gunnar Rätsch[1,*]

[1]Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA
[2]NICTA, Canberra Research Laboratory, Tower A, 7 London Circuit, Canberra ACT 2601, Australia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv `dx.doi.org/10.1101/017095`

- Analysis of alternative isoforms with RNA-seq data
  - Analyses **known** and identifies **novel** splicing events
  - Quantifies & visualizes splicing-related data
- Suitable for **large-scale projects** (1000's of samples)
- **Improved accuracy** for transcript quantification and prediction
- Open Source `bioweb.me/spladder` (python)



BIOINFORMATICS (submitted, 2015)

Vol. 00 no. 00 2014
Pages 1–2

SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data

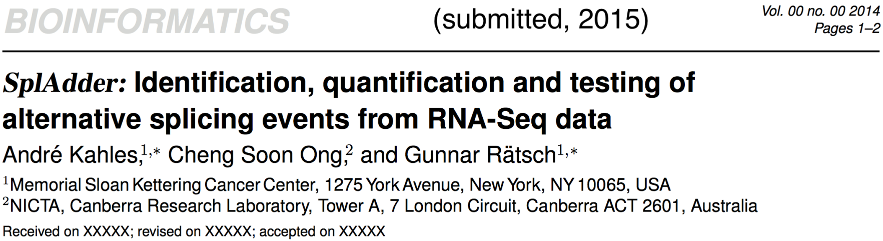
André Kahles,[1,*] Cheng Soon Ong,[2] and Gunnar Rätsch[1,*]

[1]Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA
[2]NICTA, Canberra Research Laboratory, Tower A, 7 London Circuit, Canberra ACT 2601, Australia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv `dx.doi.org/10.1101/017095`

- Analysis of alternative isoforms with RNA-seq data
  - Analyses **known** and identifies **novel** splicing events
  - Quantifies & visualizes splicing-related data
- Suitable for **large-scale projects** (1000's of samples)
- **Improved accuracy** for transcript quantification and prediction
- Open Source `bioweb.me/spladder` (python)



BIOINFORMATICS (submitted, 2015) Vol. 00 no. 00 2014 Pages 1–2

SplAdder: Identification, quantification and testing of alternative splicing events from RNA-Seq data

André Kahles,[1,*] Cheng Soon Ong,[2] and Gunnar Rätsch[1,*]

[1] Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA
[2] NICTA, Canberra Research Laboratory, Tower A, 7 London Circuit, Canberra ACT 2601, Australia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv `dx.doi.org/10.1101/017095`

- Analysis of alternative isoforms with RNA-seq data
  - Analyses **known** and identifies **novel** splicing events
  - Quantifies & visualizes splicing-related data
- Suitable for **large-scale projects** (1000's of samples)
- **Improved accuracy** for transcript quantification and prediction
- Open Source `bioweb.me/spladder` (python)

SplAdder Ideas

Major Problems in Transcriptome Analysis

1. Gene annotations are incomplete and often inaccurate
2. Whole transcript isoforms are difficult to predict/quantify

SplAdder Ideas

Major Problems in Transcriptome Analysis

1. Gene annotations are incomplete and often inaccurate
2. Whole transcript isoforms are difficult to predict/quantify

SplAdder Ideas

Major Problems in Transcriptome Analysis

1. Gene annotations are incomplete and often inaccurate
2. Whole transcript isoforms are difficult to predict/quantify

Solution

- Augment annotation with RNA-Seq evidence
- Use single splicing events instead of full transcripts

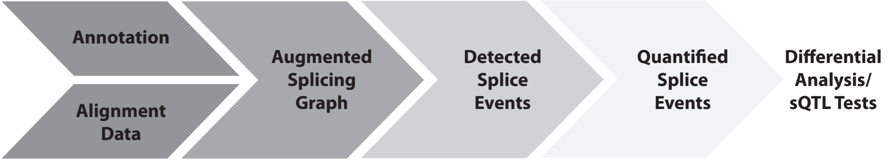
SplAdder Ideas

Major Problems in Transcriptome Analysis

1. Gene annotations are incomplete and often inaccurate
2. Whole transcript isoforms are difficult to predict/quantify

Solution

- Augment annotation with RNA-Seq evidence
- Use single splicing events instead of full transcripts



Kahles et al., bioRxiv, 2015

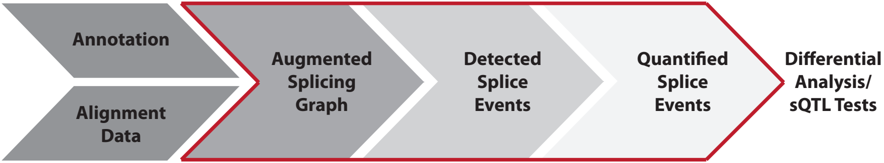
SplAdder Ideas

Major Problems in Transcriptome Analysis

1. Gene annotations are incomplete and often inaccurate
2. Whole transcript isoforms are difficult to predict/quantify

Solution

- Augment annotation with RNA-Seq evidence
- Use single splicing events instead of full transcripts

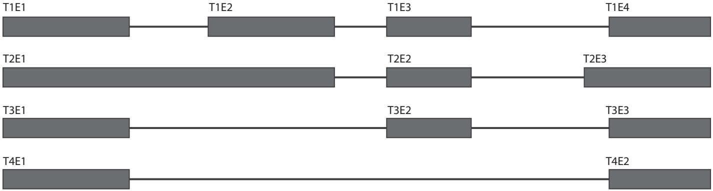


Kahles et al., bioRxiv, 2015

SplAdder Graph Augmentation

Principle

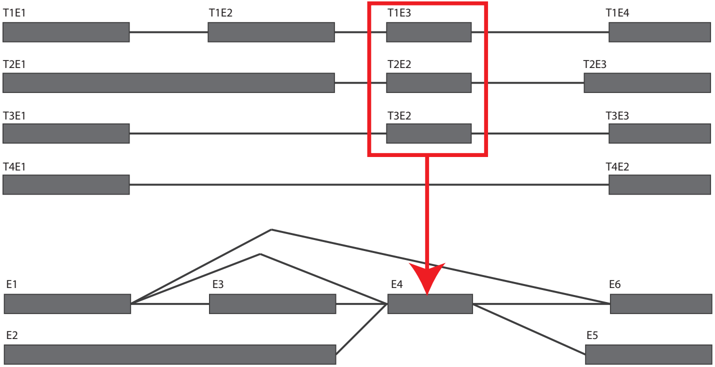
- Collapse annotated transcripts into graph representation
- Use RNA-Seq evidence to add new nodes and edges



SplAdder Graph Augmentation

Principle

- Collapse annotated transcripts into graph representation
- Use RNA-Seq evidence to add new nodes and edges

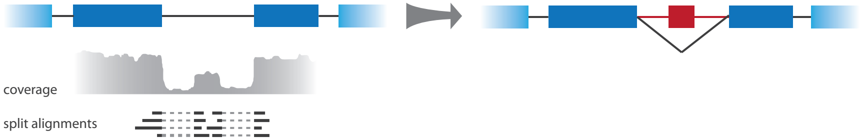


SplAdder Graph Augmentation

Principle

- Collapse annotated transcripts into graph representation
- Use RNA-Seq evidence to add new nodes and edges

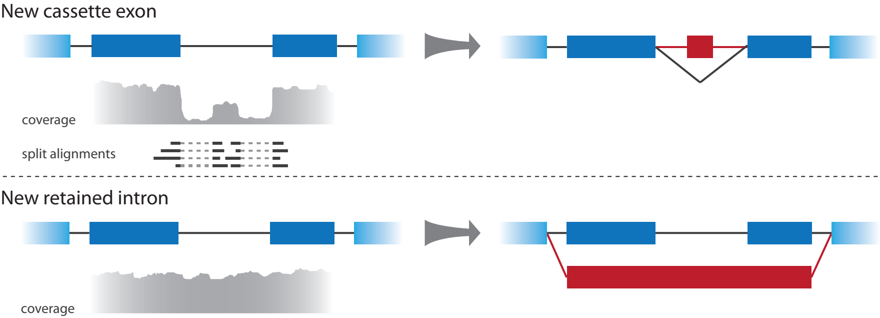
New cassette exon



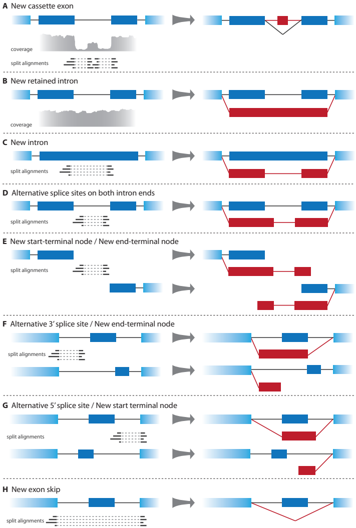
SplAdder Graph Augmentation

Principle

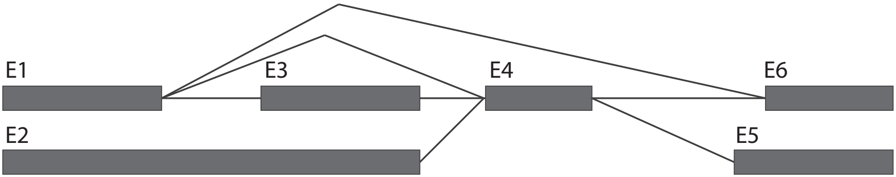
- Collapse annotated transcripts into graph representation
- Use RNA-Seq evidence to add new nodes and edges



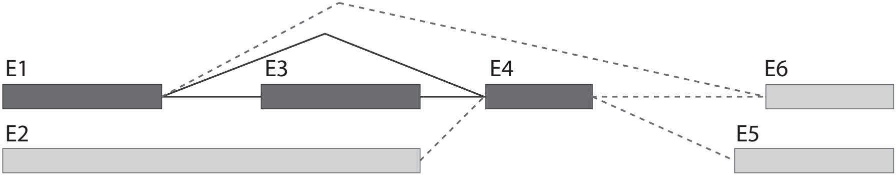
SplAdder Graph Augmentation


A New cassette exon
coverage
split alignments
B New retained intron
coverage
C New intron
split alignments
D Alternative splice sites on both intron ends
split alignments
E New start-terminal node / New end-terminal node
split alignments
F Alternative 3' splice site / New end-terminal node
split alignments
G Alternative 5' splice site / New start terminal node
split alignments
H New exon skip
split alignments

SplAdder Event Extraction

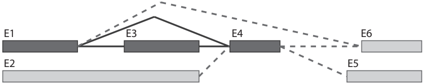


SplAdder Event Extraction

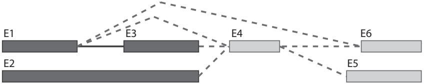


SplAdder Event Extraction

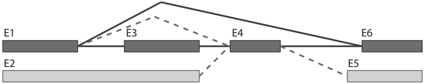
Exon Skip



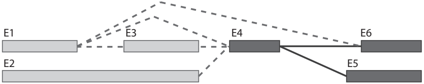
Intron Retention



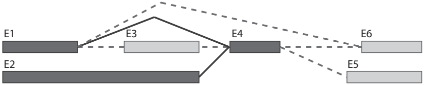
Multiple Exon Skip



Alternative 3' Splice Site

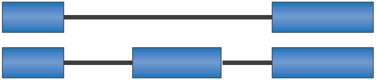


Alternative 5' Splice Site



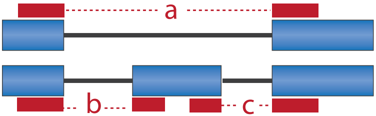
SplAdder Event Quantification and Visualization

Exon Skip



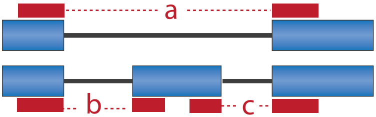
SplAdder Event Quantification and Visualization

Exon Skip



SplAdder Event Quantification and Visualization

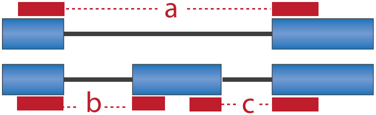
Exon Skip



$$\mathrm{PSI} = \frac{b + c}{2 \cdot a + b + c}$$

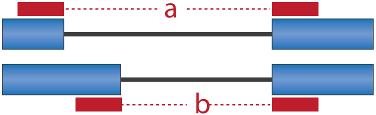
SplAdder Event Quantification and Visualization

Exon Skip



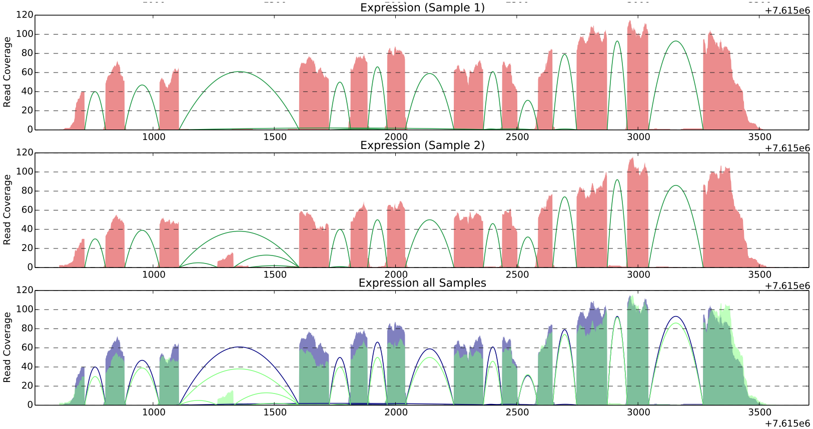
$$\mathrm{PSI} = \frac{b + c}{2 \cdot a + b + c}$$

Alternative 5' Site

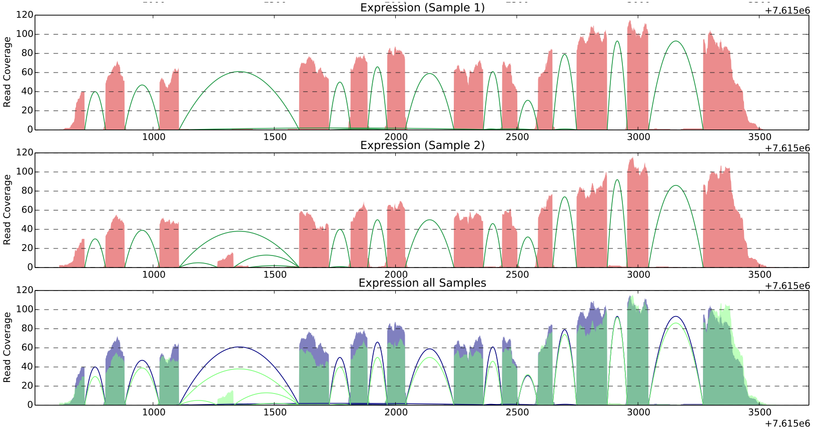


$$\mathrm{PSI} = \frac{b}{a + b}$$

SplAdder Event Quantification and Visualization



SplAdder Event Quantification and Visualization



Summary

- SplAdder effectively augments the annotation
- Enables quantitative analysis of events instead of transcripts

Splicing Analysis Across Multiple Cancer Types

Goals

1. Identify cancer-specific splicing patterns
2. Identify variants regulating splicing in same gene (cis)
3. Identify variants regulating splicing in other cancer genes (trans)

TCGA provides RNA-seq and matching exome data

- RNA-seq ⇝ Find & quantify splicing events
- Exome ⇝ Identify variants in exons & flanking intronic regions

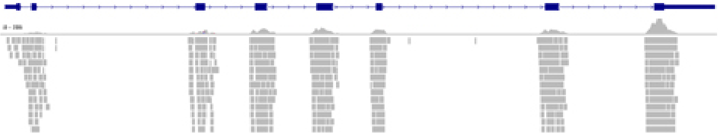
Splicing Analysis Across Multiple Cancer Types

Goals

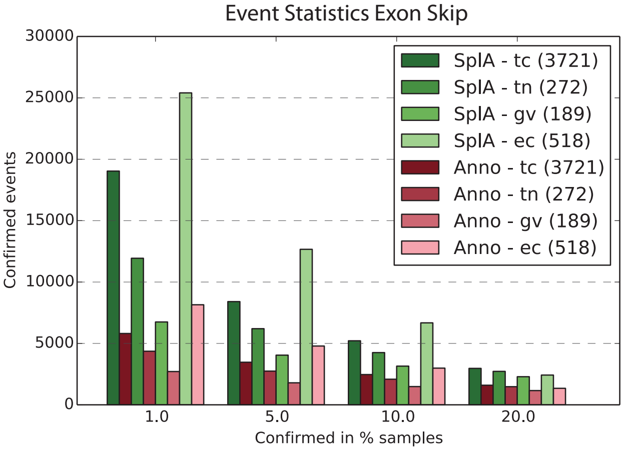
1. Identify cancer-specific splicing patterns
2. Identify variants regulating splicing in same gene (cis)
3. Identify variants regulating splicing in other <u>cancer</u> genes (trans)

TCGA provides RNA-seq and matching exome data

- RNA-seq ⇝ Find & quantify splicing events
- Exome ⇝ Identify variants in exons & flanking intronic regions

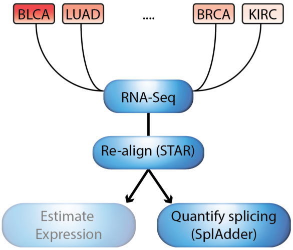


Splicing Variation Across 4,700 Samples



Analysis of a total of 4,700 RNA-seq samples from TCGA normal (tn), TCGA tumors (tc), Encode (ec) and Geuvadis (gv). Alignment w/ STAR [Dobin et al., 2013], analysis w/ SplAdder (SplA) and Gencode annotation (Anno). Figure from [Kahles, 2014].

Uniform analysis of Large-Scale RNA-seq Data



Large-scale Compute

4,700 RNA-seq libraries (≈100 TB)

- ⇒ STAR ≈ 6 CPU years
- ⇒ SplAdder ≈ 0.5 CPU years

[Kahles et al.]

Unified community resources

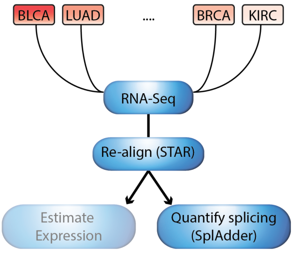
Docker with ICGC RNA-seq alignment SOP

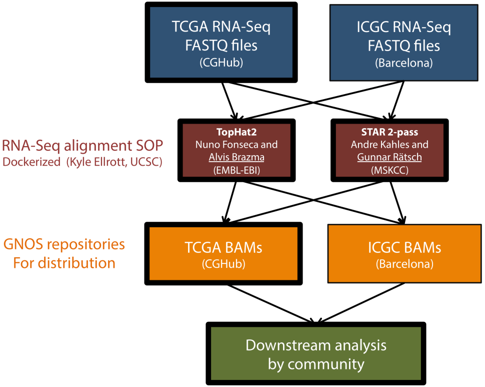
`bioweb.me/ICGC-RNA-SOP`

Syncronize with Encode, gTex, TCGA, . . .

[ICGC PCAWG-3 WG]

Uniform analysis of Large-Scale RNA-seq Data




TCGA RNA-Seq FASTQ files (CGHub)
ICGC RNA-Seq FASTQ files (Barcelona)
RNA-Seq alignment SOP
Dockerized (Kyle Ellrott, UCSC)
TopHat2
Nuno Fonseca and Alvis Brazma (EMBL-EBI)
STAR 2-pass
Andre Kahles and Gunnar Rätsch (MSKCC)
GNOS repositories
For distribution
TCGA BAMs (CGHub)
ICGC BAMs (Barcelona)
Downstream analysis by community

Large-scale Compute

4,700 RNA-seq libraries (≈100 TB)

⇒ STAR ≈ 6 CPU years

⇒ SplAdder ≈ 0.5 CPU years

[Kahles et al.]

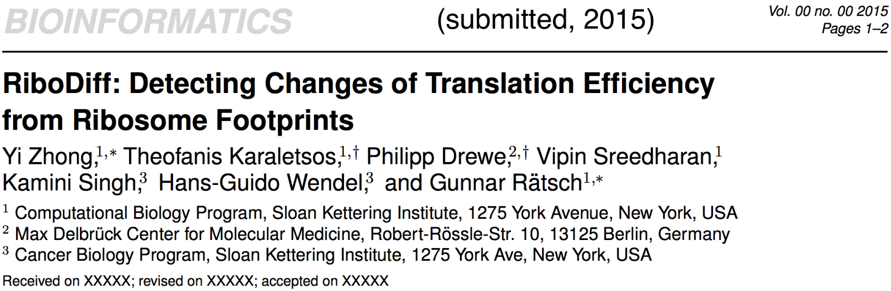
Unified community resources

Docker with ICGC RNA-seq alignment SOP

`bioweb.me/ICGC-RNA-SOP`

Syncronize with Encode, gTex, TCGA, . . .

[ICGC PCAWG-3 WG]



BIOINFORMATICS (submitted, 2015)

Vol. 00 no. 00 2015
Pages 1–2

RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints

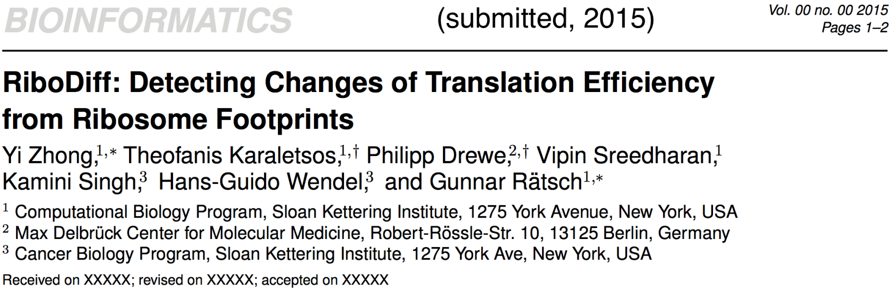
Yi Zhong,[1,*] Theofanis Karaletsos,[1,†] Philipp Drewe,[2,†] Vipin Sreedharan,[1]
Kamini Singh,[3] Hans-Guido Wendel,[3] and Gunnar Rätsch[1,*]

[1] Computational Biology Program, Sloan Kettering Institute, 1275 York Avenue, New York, USA
[2] Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[3] Cancer Biology Program, Sloan Kettering Institute, 1275 York Ave, New York, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv `dx.doi.org/10.1101/017095`

- Analysis of Ribosome profiling and RNA-seq data
  - Study translation efficiency
  - Adjusts for expression differences
- Accurate method based on dispersion estimates and GLMs
- Open Source `bioweb.me/ribodiff` (python)



BIOINFORMATICS (submitted, 2015) Vol. 00 no. 00 2015
Pages 1–2

RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints

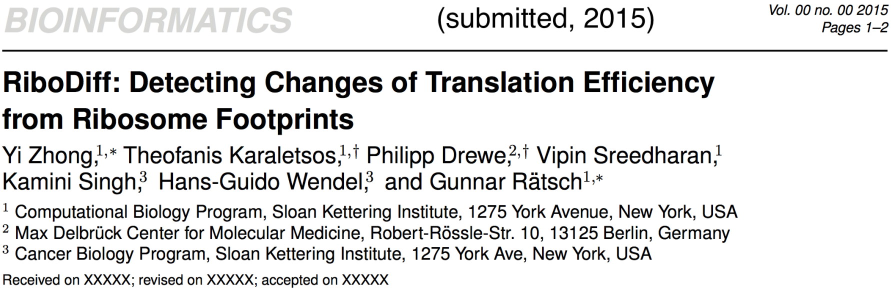
Yi Zhong,[1,*] Theofanis Karaletsos,[1,†] Philipp Drewe,[2,†] Vipin Sreedharan,[1]
Kamini Singh,[3] Hans-Guido Wendel,[3] and Gunnar Rätsch[1,*]

[1] Computational Biology Program, Sloan Kettering Institute, 1275 York Avenue, New York, USA
[2] Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[3] Cancer Biology Program, Sloan Kettering Institute, 1275 York Ave, New York, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv dx.doi.org/10.1101/017095

- Analysis of Ribosome profiling and RNA-seq data
  - Study translation efficiency
  - Adjusts for expression differences
- Accurate method based on dispersion estimates and GLMs
- Open Source bioweb.me/ribodiff (python)



BIOINFORMATICS (submitted, 2015) Vol. 00 no. 00 2015
Pages 1–2

RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints

Yi Zhong,[1,*] Theofanis Karaletsos,[1,†] Philipp Drewe,[2,†] Vipin Sreedharan,[1]
Kamini Singh,[3] Hans-Guido Wendel,[3] and Gunnar Rätsch[1,*]

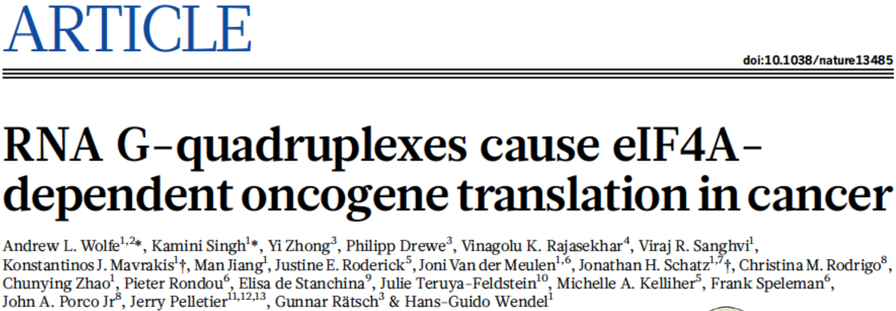
[1] Computational Biology Program, Sloan Kettering Institute, 1275 York Avenue, New York, USA
[2] Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[3] Cancer Biology Program, Sloan Kettering Institute, 1275 York Ave, New York, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

bioRxiv dx.doi.org/10.1101/017095

- Analysis of Ribosome profiling and RNA-seq data
  - Study translation efficiency
  - Adjusts for expression differences

- Accurate method based on dispersion estimates and GLMs
- Open Source bioweb.me/ribodiff (python)

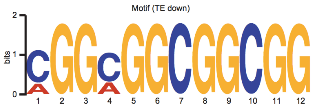


ARTICLE

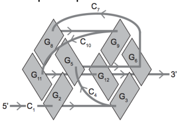
doi:10.1038/nature13485

RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer

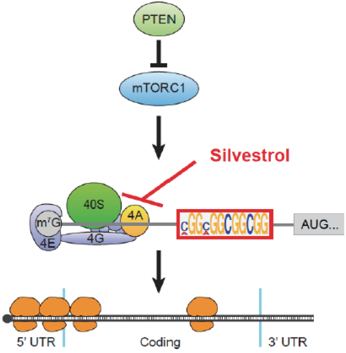
Andrew L. Wolfe[1,2]*, Kamini Singh[1]*, Yi Zhong[3], Philipp Drewe[3], Vinagolu K. Rajasekhar[4], Viraj R. Sanghvi[1], Konstantinos J. Mavrakis[1]†, Man Jiang[1], Justine E. Roderick[5], Joni Van der Meulen[1,6], Jonathan H. Schatz[1,7]†, Christina M. Rodrigo[8], Chunying Zhao[1], Pieter Rondou[6], Elisa de Stanchina[9], Julie Teruya-Feldstein[10], Michelle A. Kelliher[5], Frank Speleman[6], John A. Porco Jr[8], Jerry Pelletier[11,12,13], Gunnar Rätsch[3] & Hans-Guido Wendel[1]



G-quadruplex structures



(Analysis based on related but different strategy [Wolfe et al., 2014].)



5'UTR accumulation and reduction in RF

Summary

- MMR improves alignment choice for multi-mappers
  - ⇒ Helps improving accuracy of tools like Cufflinks
- SplAdder identifies, quantifies & visualizes alternative splicing
  - ⇒ Finds unannotated alternative splicing, tumor/normal splicing differences; splicing reprogramming; sQTLs
- riboDiff accurately detects differential translation efficiency
  - ⇒ Ribosome footprinting revealed RNA G-Quadruplex elements in 5' UTR that interacts with compound via eIF4a
- Tools (+ six other ones) are open source and available
  - ⇒ `ratschlab.org/tools`
  - ⇒ (more) Docker images come soon

Summary

- MMR improves alignment choice for multi-mappers
  - ⇒ Helps improving accuracy of tools like Cufflinks
- SplAdder identifies, quantifies & visualizes alternative splicing
  - ⇒ Finds unannotated alternative splicing, tumor/normal splicing differences; splicing reprogramming; sQTLs
- riboDiff accurately detects differential translation efficiency
  - ⇒ Ribosome footprinting revealed RNA G-Quadruplex elements in 5' UTR that interacts with compound via eIF4a
- Tools (+ six other ones) are open source and available
  - ⇒ `ratschlab.org/tools`
  - ⇒ (more) Docker images come soon . . .

Summary

- MMR improves alignment choice for multi-mappers
  - ⇒ Helps improving accuracy of tools like Cufflinks
- SplAdder identifies, quantifies & visualizes alternative splicing
  - ⇒ Finds unannotated alternative splicing, tumor/normal splicing differences; splicing reprogramming; sQTLs
- riboDiff accurately detects differential translation efficiency
  - ⇒ Ribosome footprinting revealed RNA G-Quadruplex elements in 5' UTR that interacts with compound via eIF4a
- Tools (+ six other ones) are open source and available
  - ⇒ `ratschlab.org/tools`
  - ⇒ (more) Docker images come soon . . .

Summary

- MMR improves alignment choice for multi-mappers
  - ⇒ Helps improving accuracy of tools like Cufflinks
- SplAdder identifies, quantifies & visualizes alternative splicing
  - ⇒ Finds unannotated alternative splicing, tumor/normal splicing differences; splicing reprogramming; sQTLs
- riboDiff accurately detects differential translation efficiency
  - ⇒ Ribosome footprinting revealed RNA G-Quadruplex elements in 5' UTR that interacts with compound via eIF4a
- Tools (+ six other ones) are open source and available
  - ⇒ `ratschlab.org/tools`
  - ⇒ (more) Docker images come soon . . .

Acknowledgements

Rätsch Laboratory

- **Andre Kahles**
- **Yi Zhong**
- **Philipp Drewe** @ MDC Berlin
- **Theofanis Karaletsos**
- **Kjong Van Lehmann**
- Jonas Behr @ ETH Basel
- Regina Bohnert @ Molecular Health
- Geraldine Jean @ University of Nantes

Cancer Biology

- Guido Wendel
- Kamini Singh, . . .

Cancer Genomics Projects

- Angela Brooks, Broad
- Alvis Brazma, EBI
- Matt Wilkerson, UNC
- Niki Schultz, MSKCC
- Chris Sander, MSKCC

Funding from MSKCC, Max Planck Society, European Union, German Research Foundation, Geoffrey Beene Foundation & NIH

MORE
SCIENCE.
LESS
FEAR.

Thank you!

Acknowledgements

Rätsch Laboratory

- **Andre Kahles**
- **Yi Zhong**
- **Philipp Drewe** @ MDC Berlin
- **Theofanis Karaletsos**
- **Kjong Van Lehmann**
- Jonas Behr @ ETH Basel
- Regina Bohnert @ Molecular Health
- Geraldine Jean @ University of Nantes

Cancer Biology

- Guido Wendel
- Kamini Singh, . . .

Cancer Genomics Projects

- Angela Brooks, Broad
- Alvis Brazma, EBI
- Matt Wilkerson, UNC
- Niki Schultz, MSKCC
- Chris Sander, MSKCC

Funding from MSKCC, Max Planck Society, European Union, German Research Foundation, Geoffrey Beene Foundation & NIH

Thank you!

MORE
SCIENCE.
LESS
FEAR.

References I

J. Behr, G. Schweikert, J. Cao, F. De Bona, G. Zeller, S. Laubinger, S. Ossowski, K. Schneeberger, D. Weigel, and G. Rätsch. Rna-seq and tiling arrays for improved gene finding. Oral presentation at the CSHL Genome Informatics Meeting, September 2008. URL `http://www.fml.tuebingen.mpg.de/raetsch/lectures/RaetschGenomeInformatics08.pdf`.

R. Bohnert, J. Behr, and G Rätsch. Transcript quantification with RNA-Seq data. *BMC Bioinformatics*, 10(S13):P5, 2009. URL `http://www.biomedcentral.com/1471-2105/10/S13/P5`.

RM Clark, G Schweikert, C Toomajian, S Ossowski, G Zeller, P Shinn, N Warthmann, TT Hu, G Fu, DA Hinds, H Chen, KA Frazer, DH Huson, B Schölkopf, M Nordborg, G Rätsch, JR Ecker, and D Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–342, 2007. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1138632.

Alexander Dobin, Carrie a Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts635. URL `http://www.ncbi.nlm.nih.gov/pubmed/23104886`.

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.

References II

G Jean, A Kahles, VT Sreedharan, F De Bona, and G Rätsch. Rna-seq read alignments with palmapper. *Curr Protoc Bioinformatics*, Unit 11.6, 2010.

Andre Kahles. *Novel Methods for the Computational Analysis of RNA-Seq Data with Applications to Alternative Splicing*. PhD thesis, University of Tübingen, Tübingen, Germany, September 2014.

G. Rätsch and S. Sonnenburg. Accurate splice site detection for *Caenorhabditis elegans*. In K. Tsuda B. Schoelkopf and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21(Suppl. 1):i369–i377, June 2005.

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29: 24–26, 2011.

M. Sammeth. The Flux Simulator. *Website*, 2009. http://flux.sammeth.net/simulator.html.

Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krüger, Sören Sonnenburg, and Gunnar Rätsch. mgene: Accurate svm-based gene finding with an application to nematode genomes. *Genome Research*, 2009. URL `http://genome.cshlp.org/content/early/2009/06/29/gr.090597.108.full.pdf+html`. Advance access June 29, 2009.

References III

S. Sonnenburg, G. Rätsch, A. Jagota, and K.-R. Müller. New methods for splice-site recognition. In *Proc. International Conference on Artificial Neural Networks*, 2002.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):e472–480, 2006.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotech*, advance online publication, May 2010. doi: 10.1038/nbt.1621. URL http://dx.doi.org/10.1038/nbt.1621.

A Wolfe, K Singh, Y Zhong, P Drewe, others, G Rätsch, and HG Wendel. Rna g-quadruplexes cause eif4a-dependent oncogene translation in cancer. *Nature*, 2014. doi: 10.1038/nature13485.

G Zeller, RM Clark, K Schneeberger, A Bohlen, D Weigel, and G Ratsch. Detecting polymorphic regions in arabidopsis thaliana with resequencing microarrays. *Genome Res*, 18 (6):918–929, 2008. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *BioInformatics*, 16(9): 799–807, September 2000.