

Translation, Splicing, Clinical Data and other Challenges in Biomedicine

Gunnar Rätsch

Biomedical Data Science Group

Computational Biology Center

Memorial Sloan Kettering Cancer Center



@gxr #ML #RNA #Cancer #ClinicalData

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 models. *Need better models!*

Group's research

- **Data Science**

↪ *Machine Learning,*

↪ *Bioinformatics.*

Algorithms, Models & Tools

- **Biology & Medicine**

↪ *RNA processing regulation,*

↪ *Clinical data analysis.*

Problem Setting & Goals

Biomedical Data Sciences Group

Facts

- Cost of collecting data drops, amounts increase exponentially.
- We have more data than accurate models. *Need better models!*

Group's research

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

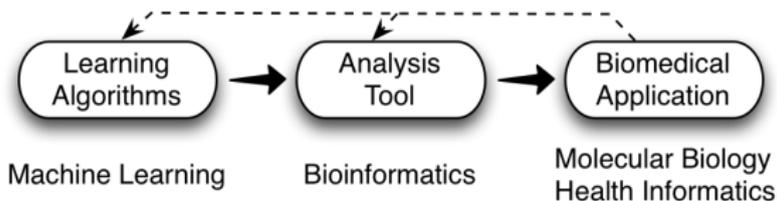
Biomedical Data Sciences Group

Facts

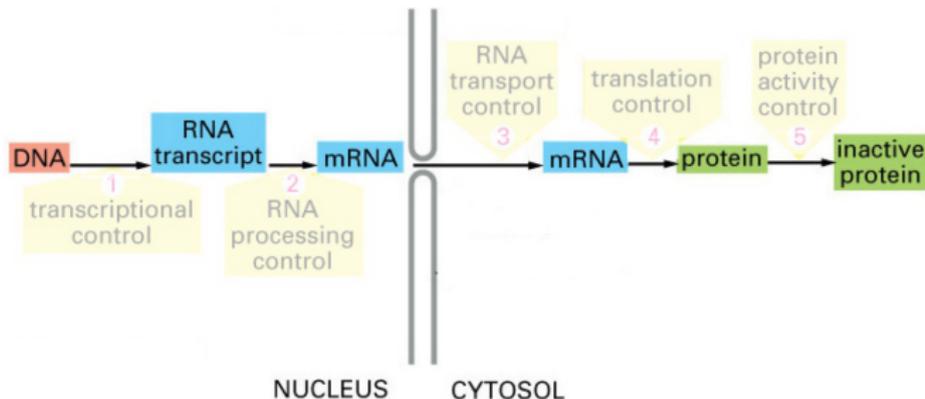
- Cost of collecting data drops, amounts increase exponentially.
- We have more data than accurate models. *Need better models!*

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



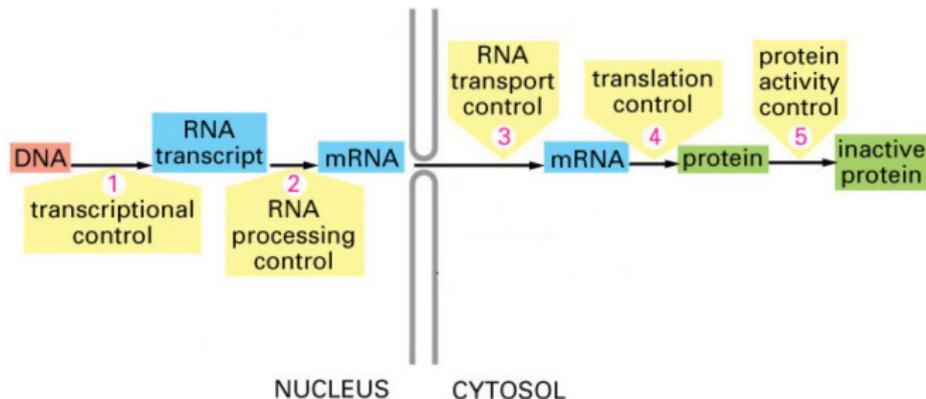
Goal: Learn to predict what these processes accomplish:

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

$$f(\text{DNA}, \dots) = \text{RNA} \quad g(\text{RNA}, \dots) = \text{protein}$$

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

Learning About the Central Dogma



Goal: Learn to predict what these processes accomplish:

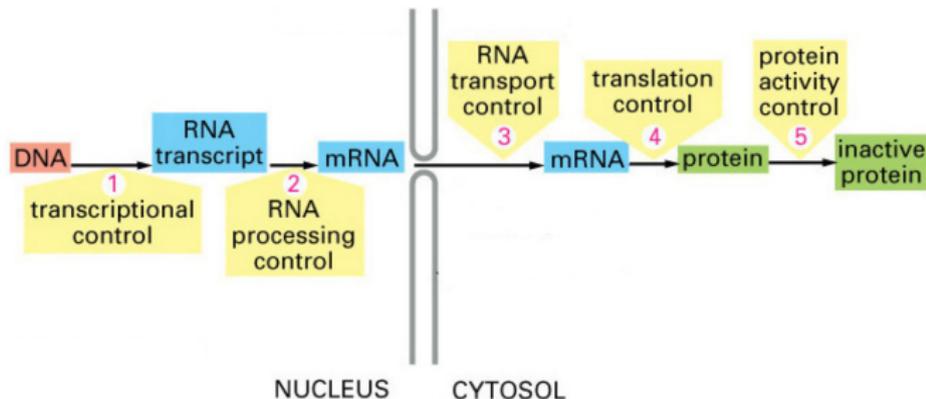
- Given the DNA, . . . , predict all gene products

$$f(\text{DNA}, \boxed{1\ 2\ 3}) = \text{RNA}$$

$$g(\text{RNA}, \boxed{4\ 5}) = \text{protein}$$

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

Learning About the Central Dogma



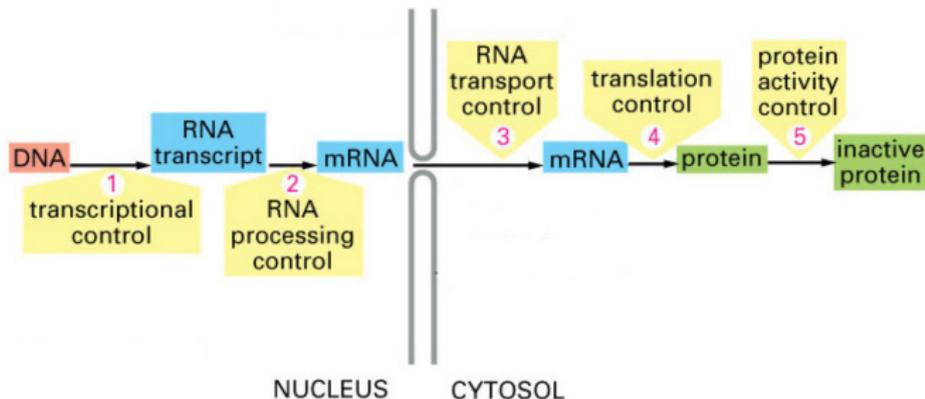
Goal: Learn to predict what these processes accomplish:

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

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

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

Learning About the Central Dogma



Goal: Learn to predict what these processes accomplish:

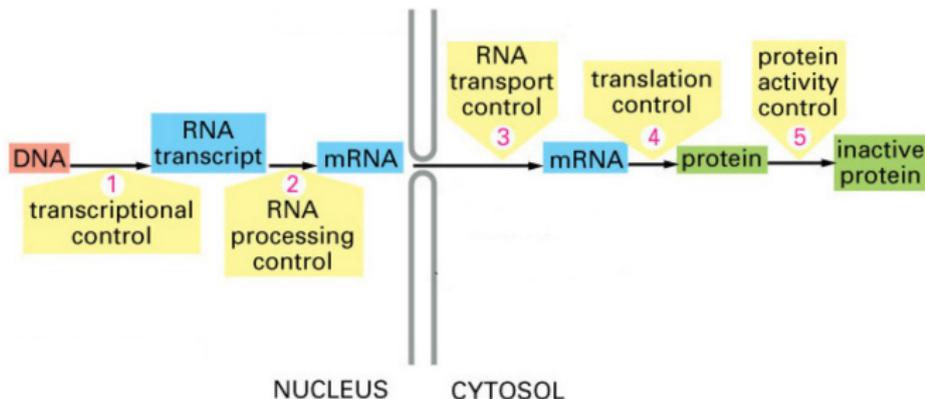
- Given the DNA, . . . , predict all gene products

$$f(\text{DNA}, \{1, 2, 3\}) = \text{RNA}$$

$$g(\text{RNA}, \{4, 5\}) = \text{protein}$$

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

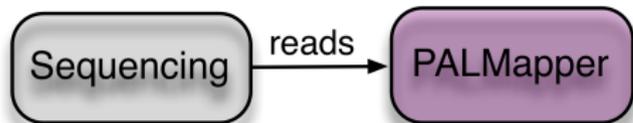
Learning About the Central Dogma



Three things are crucial:

- Biological insights (a.k.a. *prior knowledge*)
- Many observations of the system: $(\text{DNA}, \text{1 2 3}, \text{RNA})_{i=1}^N$
- Learning methods to estimate Θ : $f_{\Theta}(\text{DNA}, \text{1 2 3}) = \text{RNA}$

RNA-seq based Transcriptome Characterization

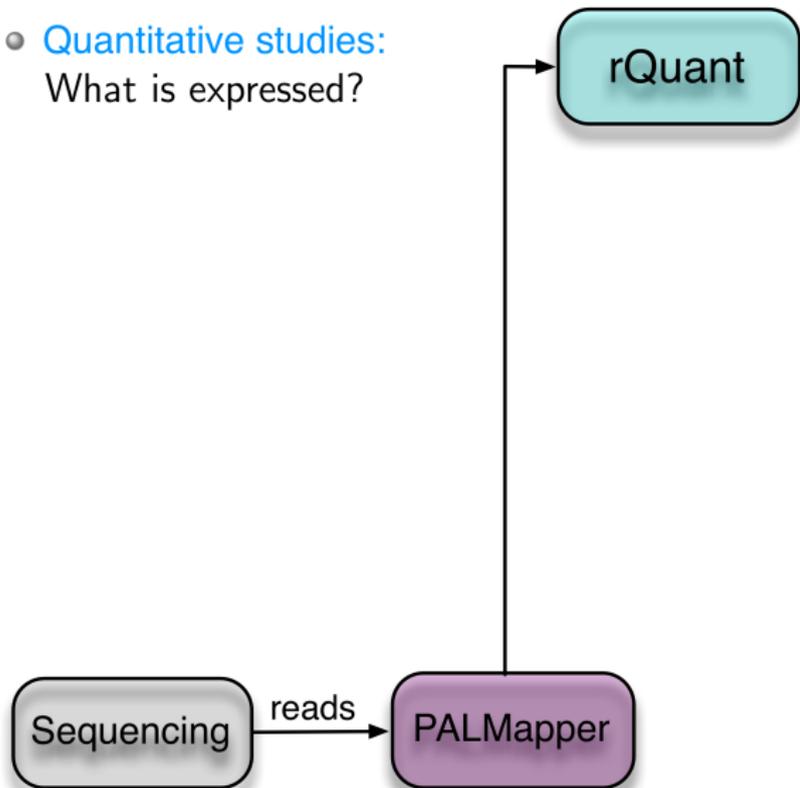


Accurate spliced
alignments

[Bona et al., 2008, Jean et al., 2010]

RNA-seq based Transcriptome Characterization

- Quantitative studies:
What is expressed?

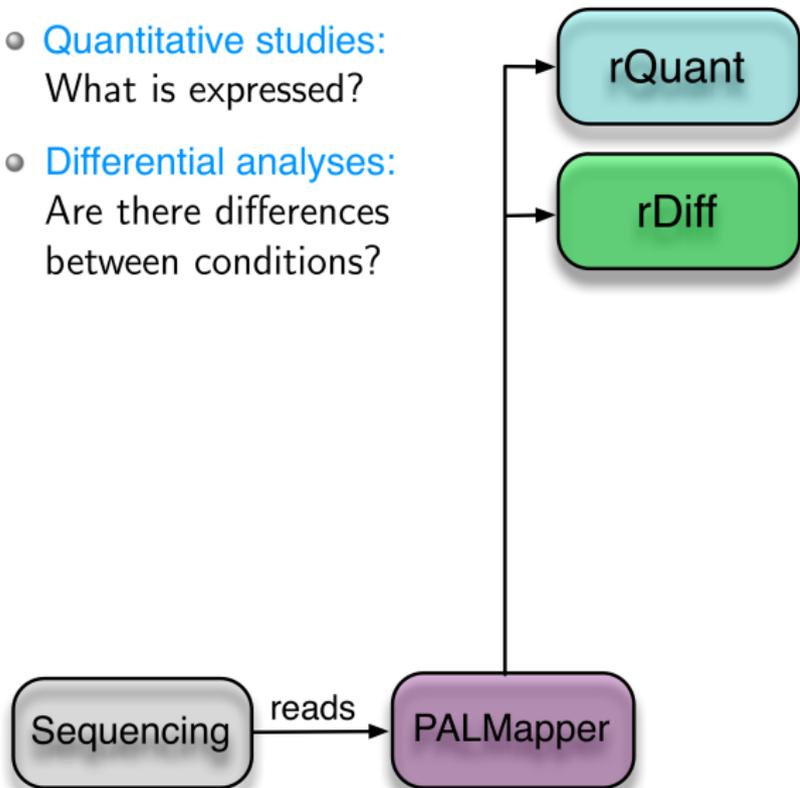


Isoform quantitation
and bias modeling
[\[Bohnert et al., 2009, 2010\]](#)

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?



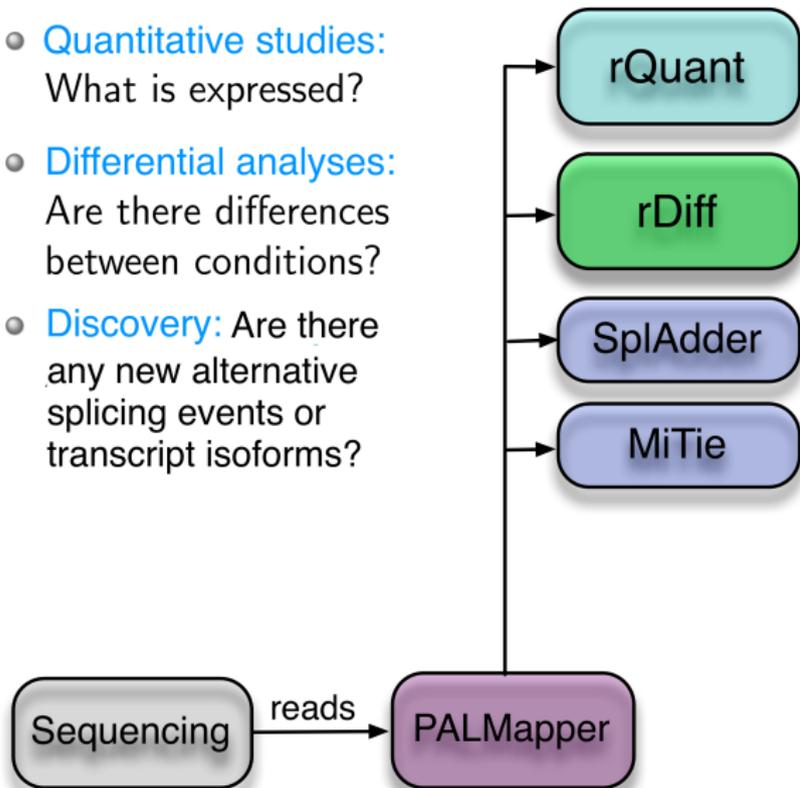
Isoform quantitation
and bias modeling
[Bohnert et al., 2009, 2010]

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 alternative splicing events or transcript isoforms?



Isoform quantitation
and bias modeling

[Bohnert et al., 2009, 2010]

Tests for differential
isoform expression

[Drewe et al., 2013]

Identify novel
alternative splicing

[Kahles et al., i.P., 2014]

Simultaneous transcript
identification & quantitation

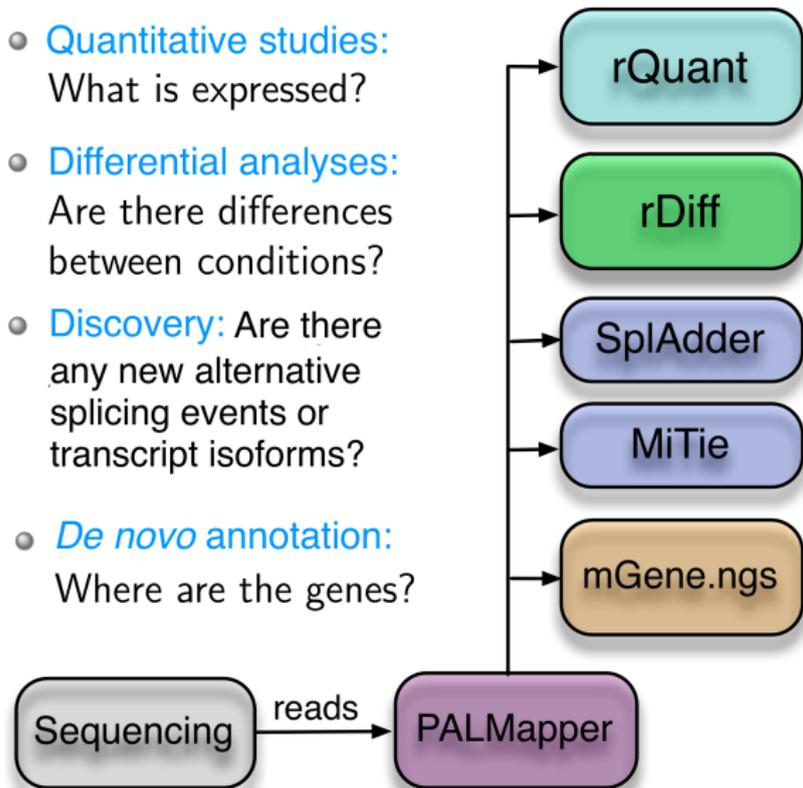
[Behr et al., 2013]

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 alternative splicing events or transcript isoforms?
- **De novo annotation:**
Where are the genes?



Isoform quantitation and bias modeling

[Bohnert et al., 2009, 2010]

Tests for differential isoform expression

[Drewe et al., 2013]

Identify novel alternative splicing

[Kahles et al., i.P., 2014]

Simultaneous transcript identification & quantitation

[Behr et al., 2013]

Gene finding with RNA-seq evidence

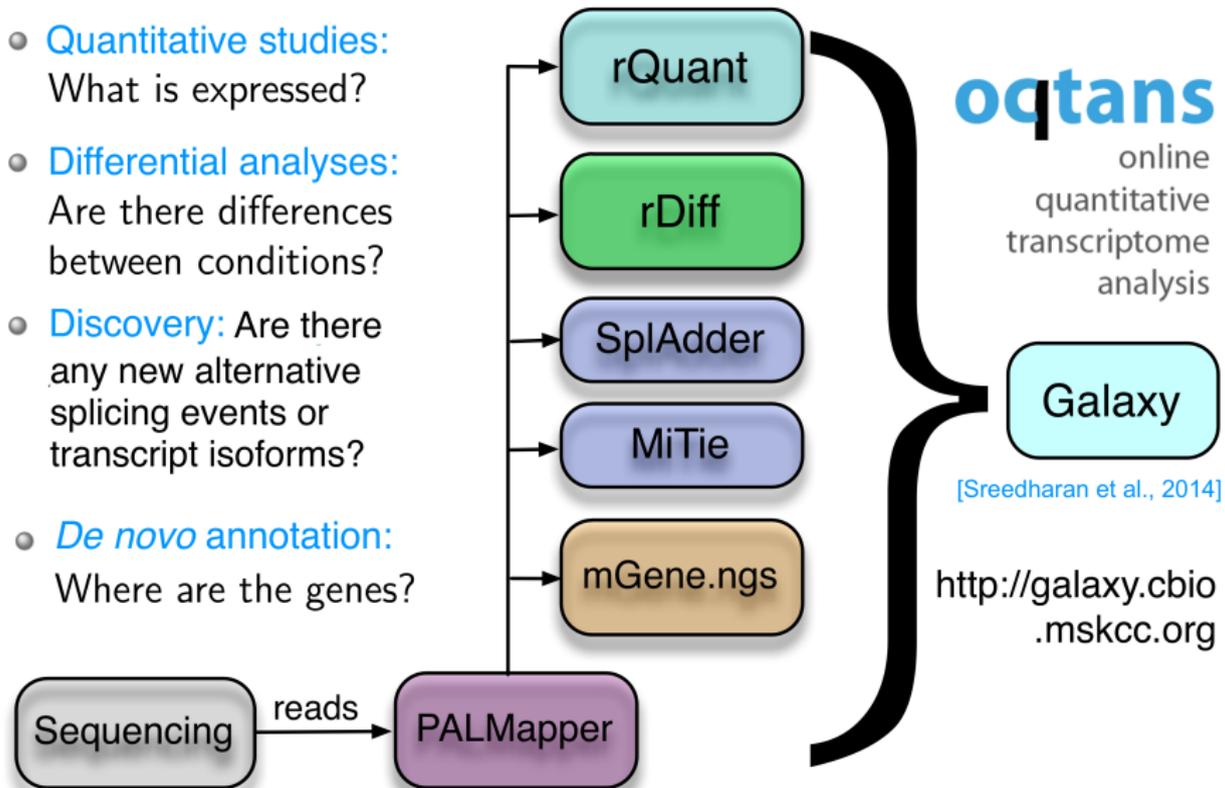
[Behr et al., 2010, 2013, Gan et al., 2011]

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 alternative splicing events or transcript isoforms?
- **De novo annotation:**
Where are the genes?



Accurate detection of differential RNA processing

Philipp Drewe^{1,2,*}, Oliver Stegle^{3,4}, Lisa Hartmann^{2,5}, André Kahles^{1,2}, Regina Bohnert²,
Andreas Wachter⁵, Karsten Borgwardt^{3,4,6} and Gunnar Rätsch^{1,2,*}

¹Computational Biology Center, Sloan-Kettering Institute, 1275 York Avenue, New York, NY 10065, USA,

²Friedrich Miescher Laboratory of the Max-Planck Society, Spemannstrasse 39, 72076 Tübingen, Germany,

³Machine Learning and Computational Biology Research Group, Max Planck Institute for Intelligent Systems, Spemannstrasse 38, 72076 Tübingen, Germany, ⁴Department of Molecular Biology, Max Planck Institute for Developmental Biology, Spemannstrasse 38, 72076 Tübingen, Germany, ⁵Center for Plant Mol. Biology, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany and ⁶Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

rDiff



Novel non-parametric statistical test for differential transcript expression that even works when annotations are incomplete.

<http://bioweb.me/rdiff>

Application & extension of techniques to Ribosome footprinting and analysis of effect of drug Silvestrol on translation (+ biology).

Accurate detection of differential RNA processing

Philipp Drewe^{1,2,*}, Oliver Stegle^{3,4}, Lisa Hartmann^{2,5}, André Kahles^{1,2}, Regina Bohnert², Andreas Wachter⁵, Karsten Borgwardt^{3,4,6} and Gunnar Rätsch^{1,2,*}

rDiff



¹Computational Biology Center, Sloan-Kettering Institute, 1275 York Avenue, New York, NY 10065, USA, ²Friedrich Miescher Laboratory of the Max-Planck Society, Spemannstrasse 39, 72076 Tübingen, Germany, ³Machine Learning and Computational Biology Research Group, Max Planck Institute for Intelligent Systems, Spemannstrasse 38, 72076 Tübingen, Germany, ⁴Department of Molecular Biology, Max Planck Institute for Developmental Biology, Spemannstrasse 38, 72076 Tübingen, Germany, ⁵Center for Plant Mol. Biology, University of Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany and ⁶Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

Novel non-parametric statistical test for differential transcript expression that even works when annotations are incomplete.

<http://bioweb.me/rdiff>

rDiff

Application



ARTICLE

Nature, 2014 Sep 4; 513(7516):65-70

[doi:10.1038/nature13485](https://doi.org/10.1038/nature13485)

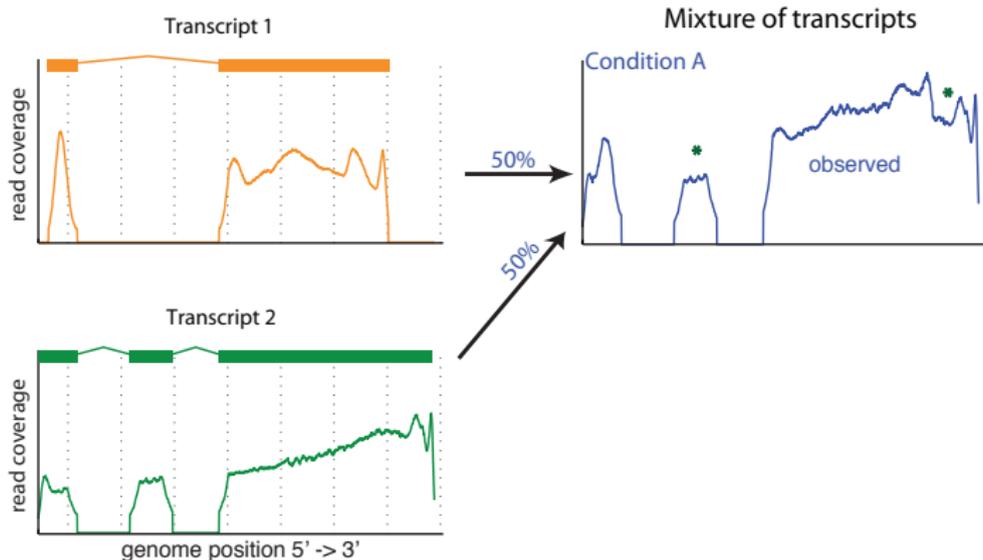
RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer

Andrew L. Wolfe^{1,2*}, Kamini Singh^{1*}, Yi Zhong³, Philipp Drewe³, Vinagolu K. Rajasekhar⁴, Viraj R. Sanghvi¹, Konstantinos J. Mavrikis^{1†}, Man Jiang², Justine E. Roderick³, Joni Van der Meulen^{1,6}, Jonathan H. Schatz^{1,7†}, Christina M. Rodrigo⁸, Chunying Zhao¹, Pieter Rondou⁶, Elisa de Stanchina⁹, Julie Teruya-Feldstein¹⁰, Michelle A. Kelliher⁵, Frank Speleman⁶, John A. Porco Jr⁸, Jerry Pelletier^{11,12,13}, Gunnar Rätsch³ & Hans-Guido Wendel¹

Application & extension of techniques to Ribosome footprinting and analysis of effect of drug Silvestrol on translation (+ biology).

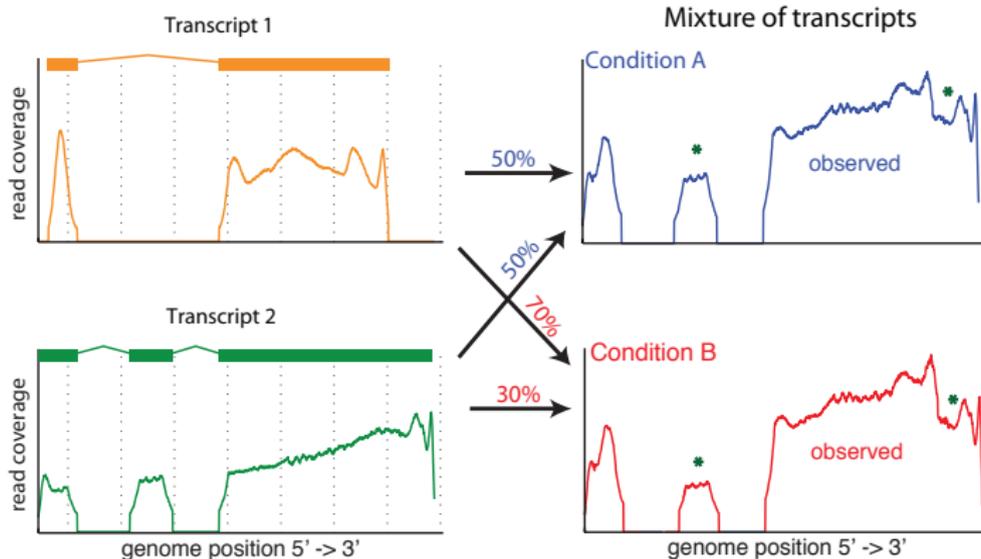
Detecting Differential RNA processing

Compare the read distributions in two conditions



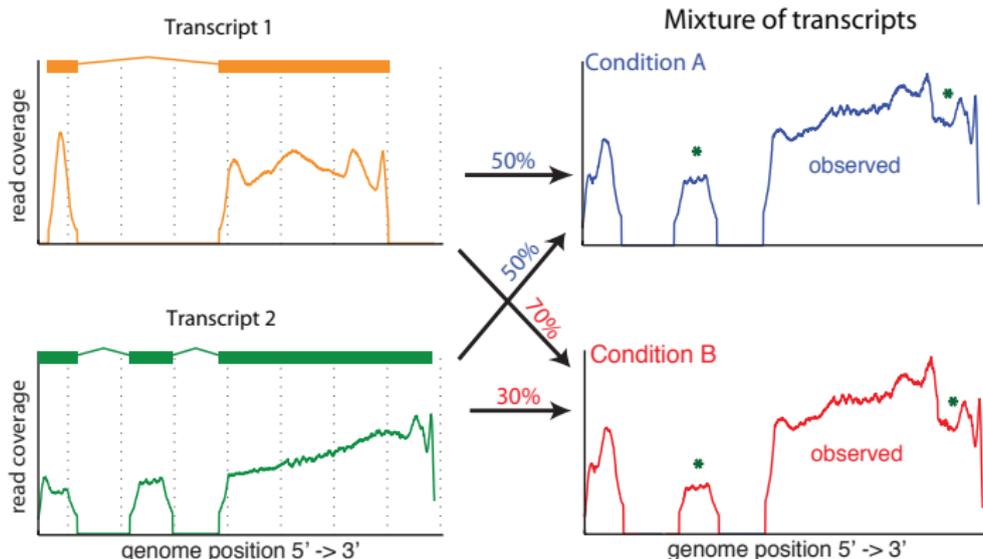
Detecting Differential RNA processing

Compare the read distributions in two conditions



Detecting Differential RNA processing

Compare the read distributions in two conditions



Goal: Design a test to detect differential RNA processing:
(Alternative splicing, promotor usage, NMD, ...)

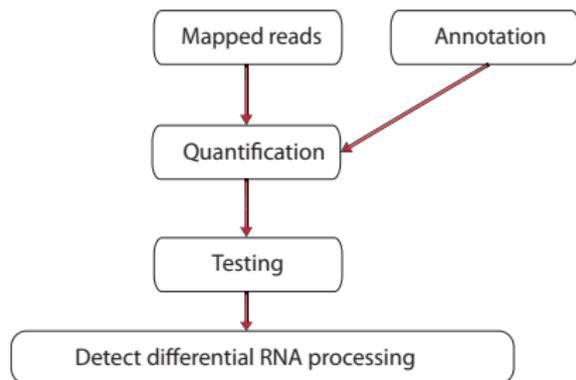
Testing Strategies

Known Transcripts:

- **Two-step** approach
 - 1 Quantification
 - 2 Testing
- Avoid quantification?
 - One-step region testing

Unknown transcripts:

- Include transcript identification in analysis.
 - 1 Detection
 - 2 Region testing
 ⇒ Complex
- **One-step** testing on the read densities!



[Wong et al., Bioinformatics, 2009]
 [Yaspo et al., Nucl. Acids Res., 2010]
 [Bohnert et al., BMC Bioinf., 2010]
 [Stegle et al., Nat. Prec., 2010]
 [Anders et al., Genome Res., 2012]

[Drewe et al., Nuc. Acids Res., 2013]

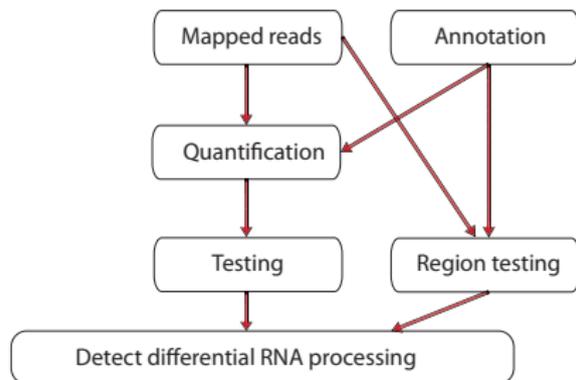
Testing Strategies

Known Transcripts:

- **Two-step** approach
 - ① Quantification
 - ② Testing
- Avoid quantification?
 - **One-step** region testing

Unknown transcripts:

- Include transcript identification in analysis.
 - ① Detection
 - ② Region testing
 ⇒ **Complex**
- **One-step** testing on the read densities!



[Wong et al., Bioinformatics, 2009]
 [Yaspo et al., Nucl. Acids Res., 2010]
 [Bohnert et al., BMC Bioinf., 2010]
 [Stegle et al., Nat. Prec., 2010]
 [Anders et al., Genome Res., 2012]

[Drewe et al., Nuc. Acids Res., 2013]

Testing Strategies

Known Transcripts:

- **Two-step** approach
 - ① Quantification
 - ② Testing
- Avoid quantification?
 - **One-step** region testing

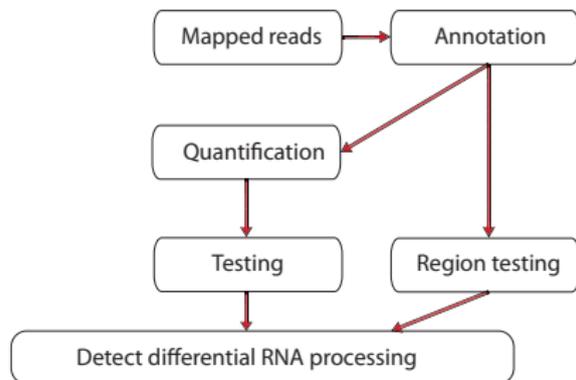
Unknown transcripts:

- Include transcript identification in analysis.

- ① Detection
- ② Region testing

⇒ **Complex!**

- **One-step** testing on the read densities!



[Wong et al., Bioinformatics, 2009]
 [Yaspo et al., Nucl. Acids Res., 2010]
 [Bohnert et al., BMC Bioinf., 2010]
 [Stegle et al., Nat. Prec., 2010]
 [Anders et al., Genome Res., 2012]

[Drewe et al., Nuc. Acids Res., 2013]

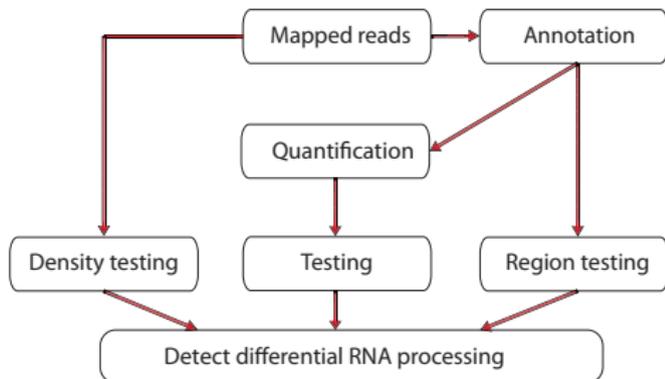
Testing Strategies

Known Transcripts:

- **Two-step** approach
 - 1 Quantification
 - 2 Testing
- Avoid quantification?
 - **One-step** region testing

Unknown transcripts:

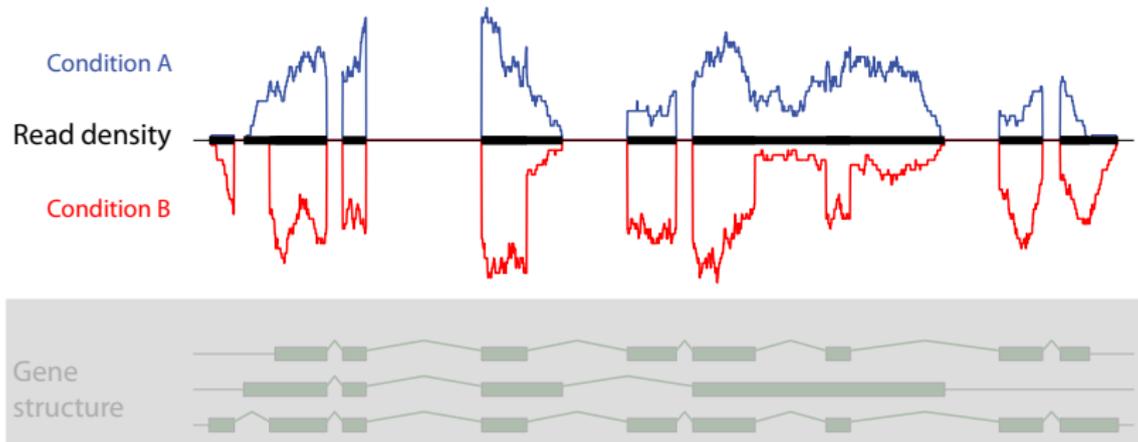
- Include transcript identification in analysis.
 - 1 Detection
 - 2 Region testing
 ⇒ **Complex!**
- **One-step** testing on the read densities!



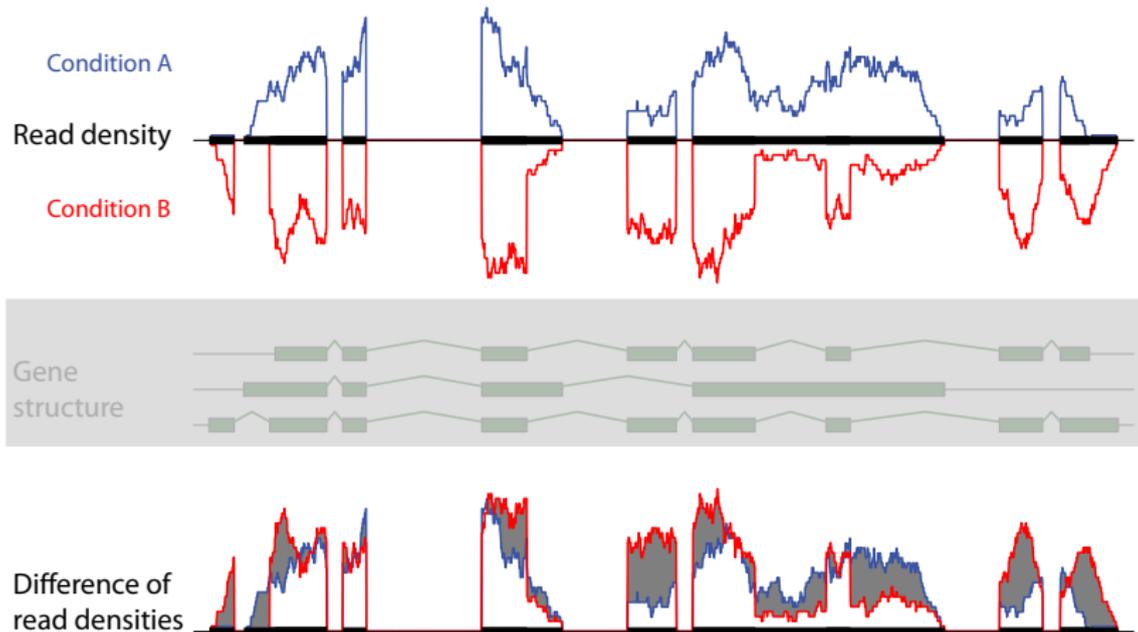
[Wong et al., Bioinformatics, 2009]
 [Yaspo et al., Nucl. Acids Res., 2010]
 [Bohnert et al., BMC Bioinf., 2010]
 [Stegle et al., Nat. Prec., 2010]
 [Anders et al., Genome Res., 2012]

[Drewe et al., Nuc. Acids Res., 2013]

Density Testing Without Gene Structure



Density Testing Without Gene Structure



Non-parametric Test for High-dimensional Data

- Represent the density of reads

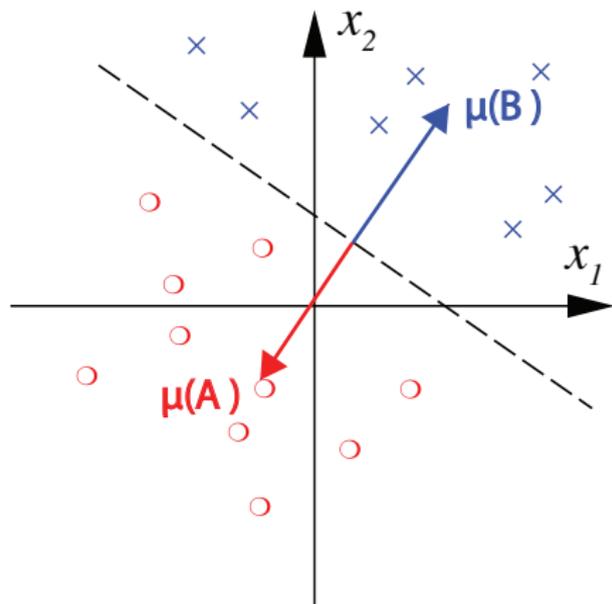
$$\mu^{A/B} = \sum_{i=1}^N \Phi(x_i)$$

- Compute the distance between μ^A and μ^B :

$$D(A, B) = \|\mu^A - \mu^B\|_{\mathcal{H}}$$

- Permute reads between samples to compute p-value

- Trick: Match observed dispersion by subsampling



[Gretton et al., 2008]

Non-parametric Test for High-dimensional Data

- Represent the density of reads

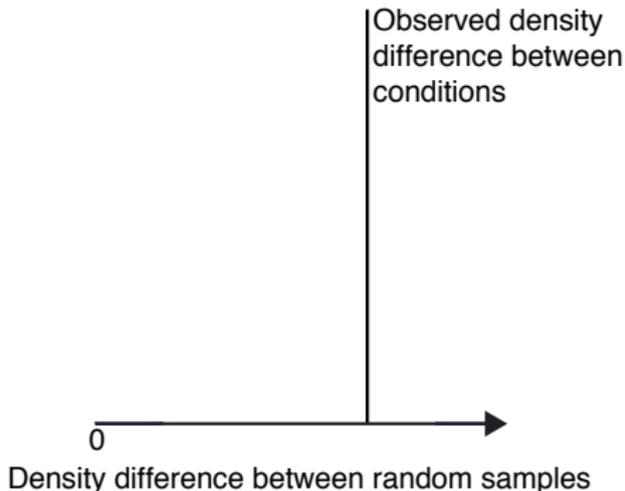
$$\mu^{A/B} = \sum_{i=1}^N \Phi(x_i)$$

- **Compute the distance**

between μ^A and μ^B :

$$D(A, B) = \|\mu^A - \mu^B\|_{\mathcal{H}}$$

- **Permute reads** between samples to compute p-value
- Trick: **Match** observed dispersion by subsampling



[Gretton et al., 2008]

Drewe et al. (2013) contains proper comparison with other state-of-the-art methods (e.g., CuffDiff, Miso).

Non-parametric Test for High-dimensional Data

- Represent the density of reads

$$\mu^{A/B} = \sum_{i=1}^N \Phi(x_i)$$

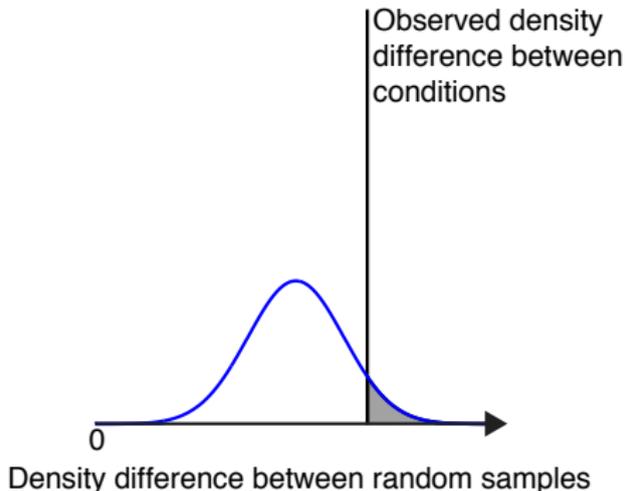
- **Compute the distance**

between μ^A and μ^B :

$$D(A, B) = \|\mu^A - \mu^B\|_{\mathcal{H}}$$

- **Permute reads** between samples to compute p-value

- Trick: **Match** observed dispersion by subsampling



[Gretton et al., 2008]

Drewe et al. (2013) contains proper comparison with other state-of-the-art methods (e.g., CuffDiff, Miso).

Non-parametric Test for High-dimensional Data

- Represent the density of reads

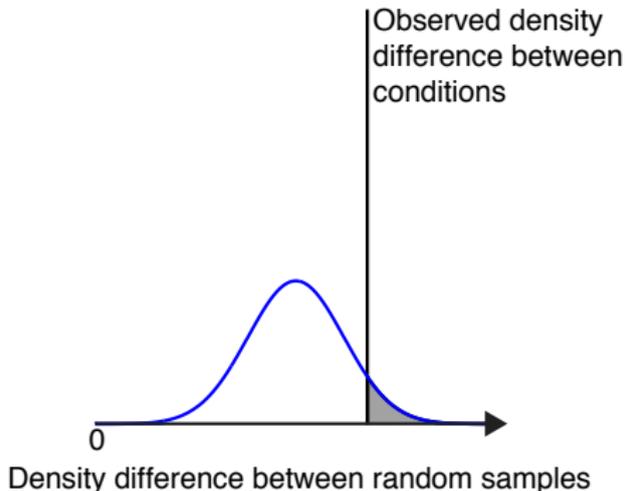
$$\mu^{A/B} = \sum_{i=1}^N \Phi(x_i)$$

- **Compute the distance**

between μ^A and μ^B :

$$D(A, B) = \|\mu^A - \mu^B\|_{\mathcal{H}}$$

- **Permute reads** between samples to compute p-value
- Trick: **Match** observed dispersion by subsampling



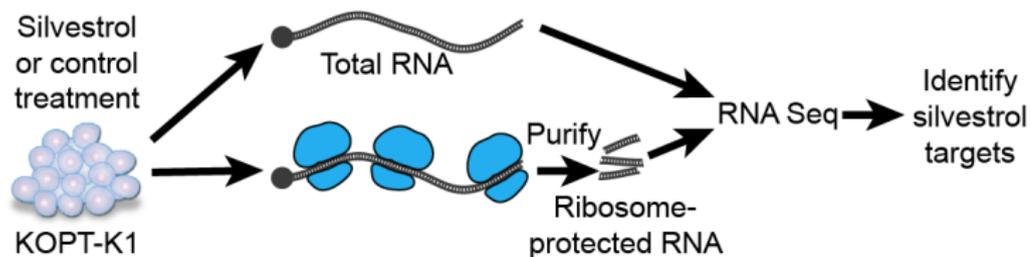
[Gretton et al., 2008]

Drewe et al. (2013) contains proper comparison with other state-of-the-art methods (e.g., CuffDiff, Miso).

Other Applications

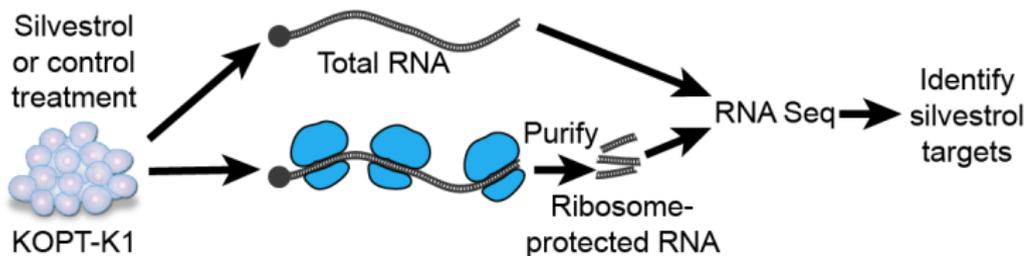
- RNA transcript expression [Drewe et al., NAR, 2013]
- Ribosome footprinting [Wolfe et al., Nature, 2014]
- Protein expression (RPPA, Mass-Spec) [possible/need collaborator]
- ChIP-seq peak analysis [Schweikert et al., BMC Genomics, 2013]
- CLIP-seq peak analysis [possible/need collaborator]
- RNA secondary structure probing [ongoing]
- Protein structure probing (NMR?) [possible/need collaborator]
- Probing of repetitive polymorphisms [Chae et al., Cell, 2014, i.p.]
- ...

Application to Ribosome Profiling



[Wolfe, Sing, Zhong, Drewe et al., Nature, 2014]

Application to Ribosome Profiling



Compound **Silvestrol** extracted from plant has anti-cancer activities:

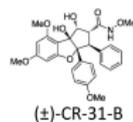
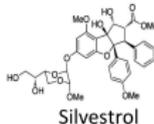
Aglaia silvestris

From Wikipedia, the free encyclopedia

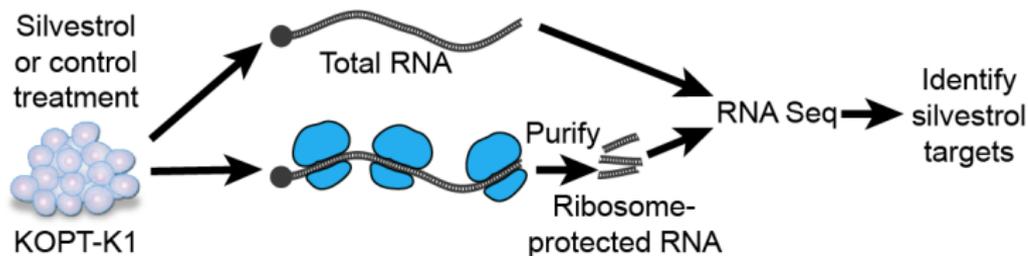
Aglaia silvestris is a species of plant in the [Meliaceae](#) family. It is found in [Cambodia](#), [India](#), [Indonesia](#), [Malaysia](#), [Papua New Guinea](#), the [Philippines](#), the [Solomon Islands](#), [Thailand](#), and [Vietnam](#).

Source [\[edit\]](#)

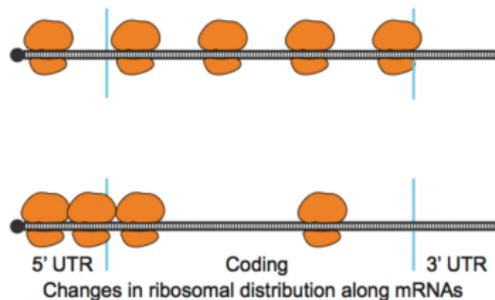
- Pannell, C.M. 1998. *Aglaia silvestris* [\[ref.\]](#). 2006



Application to Ribosome Profiling

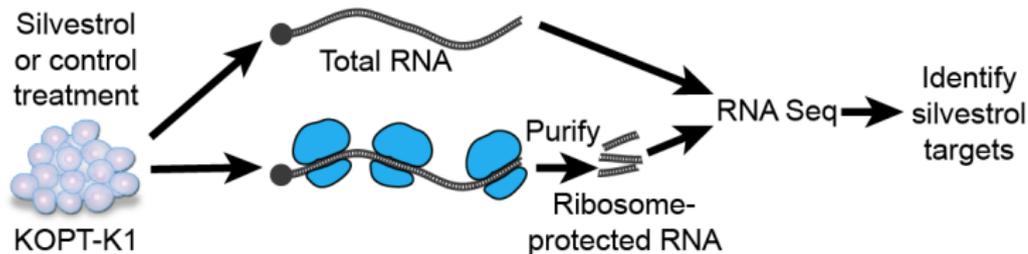


Investigated effect of **Silvestrol** on translation using **rDiff** in T-ALL:

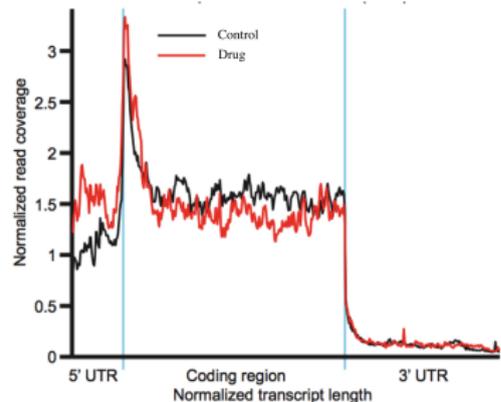
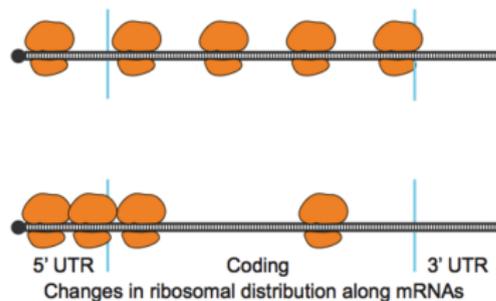


[Wolfe, Sing, Zhong, Drewe et al., Nature, 2014]

Application to Ribosome Profiling



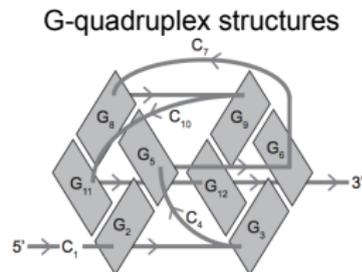
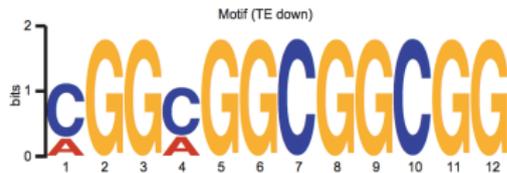
Investigated effect of **Silvestrol** on translation using **rDiff** in T-ALL:



[Wolfe, Sing, Zhong, Drewe et al., Nature, 2014]

Application to Ribosome Profiling

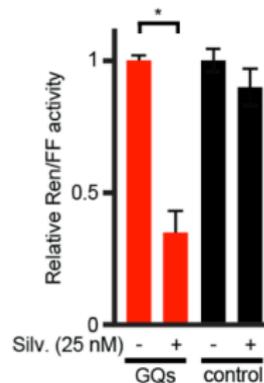
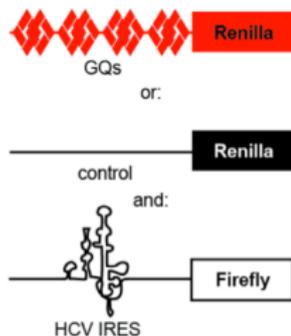
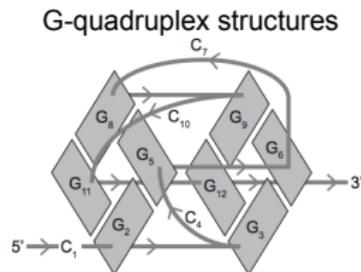
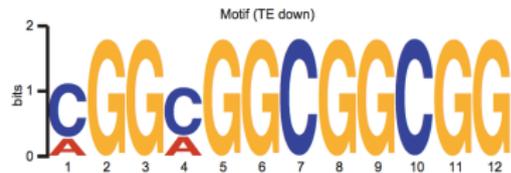
Found a **motif** that was strongly **enriched** in detected genes:



[Wolfe, Sing, Zhong, Drewe et al., Nature, 2014]

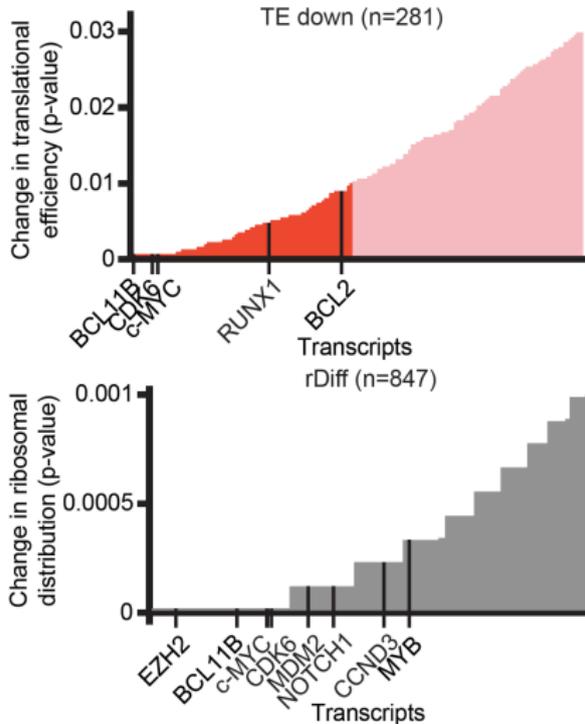
Application to Ribosome Profiling

Found a **motif** that was strongly **enriched** in detected genes:



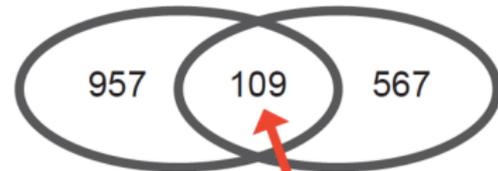
[Wolfe, Sing, Zhong, Drewe et al., Nature, 2014]

Silvestrol's Anti-Cancer Activity



Super-enhancers in T-ALL

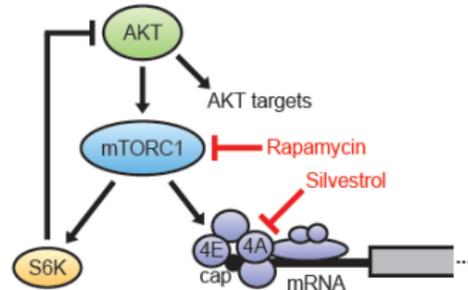
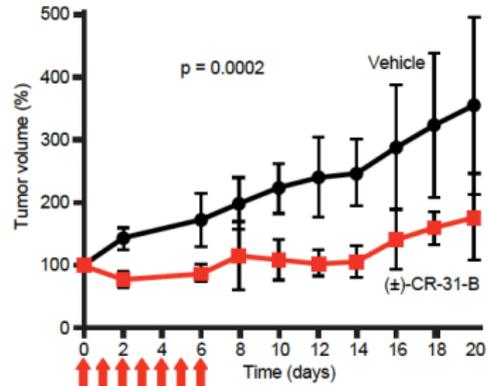
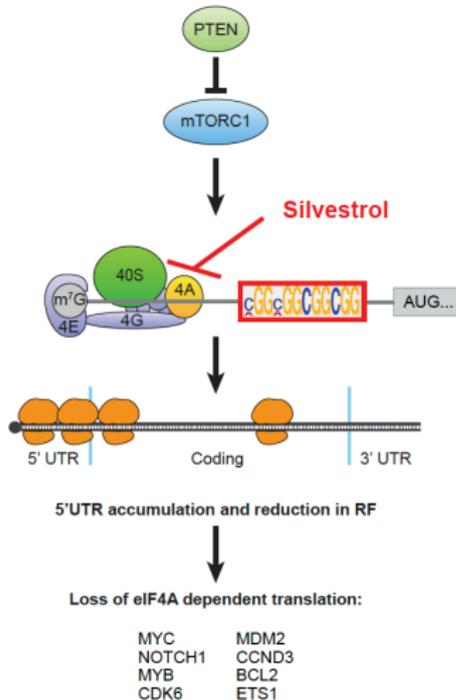
TE down/
rDiff genes



Includes: BCL2
CCND3
CDK6
ETS1
MYB
NOTCH1
RUNX1

[Wolfe, Sing, Zhong, Drewe et al., Nature., 2014]

Silvestrol's Anti-Cancer Activity via eIF4A



[Wolfe, Sing, Zhong, Drewe et al., Nature., 2014]

SplAdder



SplAdder: Integrated Quantification, Visualization and Differential Analysis of Alternative Splicing

André Kahles¹ Cheng Soon Ong,² Kjong-Van Lehmann¹ and Gunnar Rätsch^{1*}

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

²NICTA, Canberra Research Laboratory, Tower A, 7 London Circuit, Canberra ACT 2601, Australia

Novel tool for identifying novel splicing variants, differential analysis and visualization. <http://bioweb.me/spladder>

Integrative splicing analysis of Kidney Renal Clear Cell Carcinoma.

SplAdder



SplAdder: Integrated Quantification, Visualization and Differential Analysis of Alternative Splicing

André Kahles¹ Cheng Soon Ong,² Kjong-Van Lehmann¹ and Gunnar Rätsch^{1*}

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

²NICTA, Canberra Research Laboratory, Tower A, 7 London Circuit, Canberra ACT 2601, Australia

Novel tool for identifying novel splicing variants, differential analysis and visualization. <http://bioweb.me/spladder>

SplAdder

Application



(Submitted to PSB'15)

Integrative Genome-wide Analysis of the Determinants of RNA Splicing in Kidney Renal Clear Cell Carcinoma

Kjong-Van Lehmann,^{1,*†} Andre Kahles,^{1,†} Cyriac Kandath,¹ William Lee,¹ Nikolaus Schultz,¹ and Oliver Stegle,² and Gunnar Rätsch¹

¹ Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, NY 10044, U.S.A

² European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom

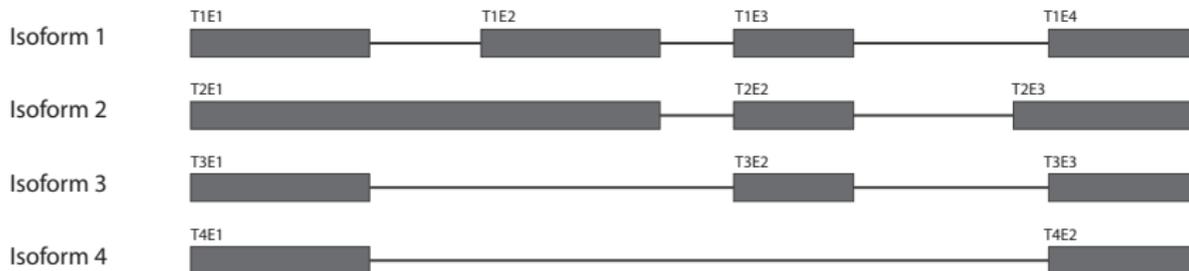
[†] Both authors contributed equally.

Integrative splicing analysis of Kidney Renal Clear Cell Carcinoma.

SplAdder: Detection of Novel Splicing Events

Building the Splicing Graph

- Take all annotated transcript isoforms of a gene
- Resolve redundancies by graph representation



SplAdder: Detection of Novel Splicing Events

Building the Splicing Graph

- Take all annotated transcript isoforms of a gene
- Resolve redundancies by graph representation



SplAdder: Detection of Novel Splicing Events

Building the Splicing Graph

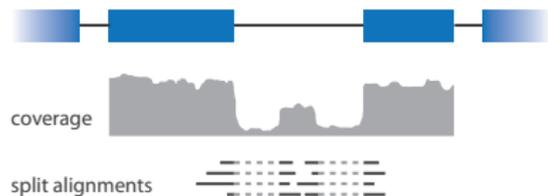
- Take all annotated transcript isoforms of a gene
- Resolve redundancies by graph representation



Splicing Graph Augmentation

Criteria for Augmentation

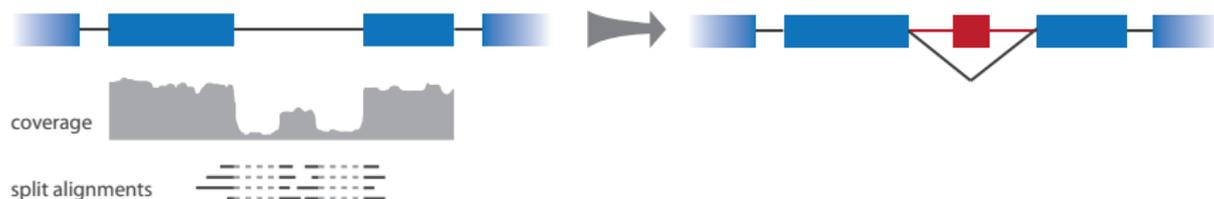
- Support from exonic coverage
- Splice junction evidence from split alignments



Splicing Graph Augmentation

Criteria for Augmentation

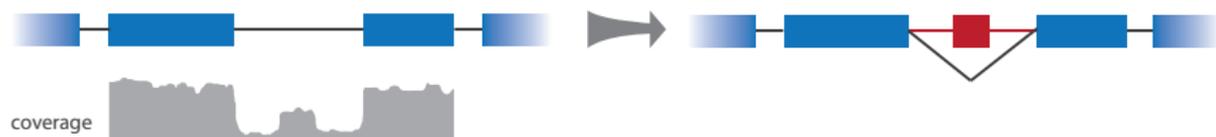
- Support from exonic coverage
- Splice junction evidence from split alignments



Splicing Graph Augmentation

Criteria for Augmentation

- Support from exonic coverage
- Splice junction evidence from split alignments



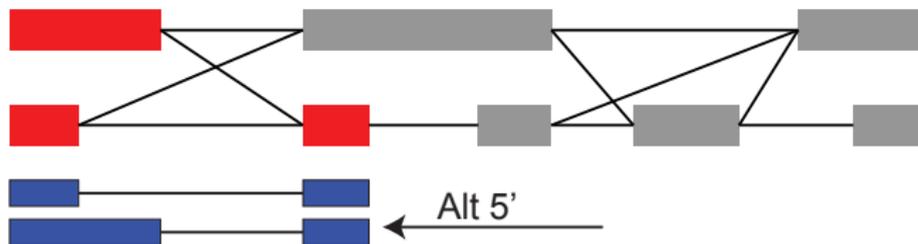
Extract Alternative Splicing Events

- Define alternative splicing events as minimal subsets of nodes in the graph
- Extract one sub-graph per event → genelets



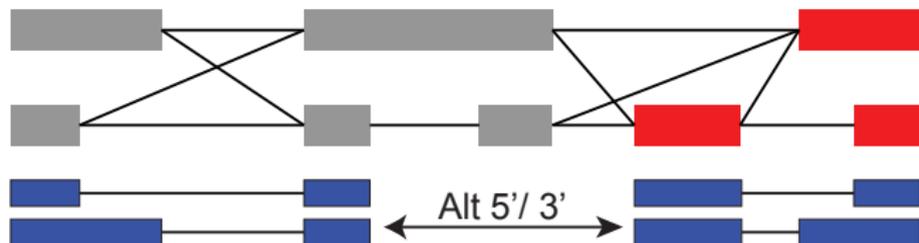
Extract Alternative Splicing Events

- Define alternative splicing events as minimal subsets of nodes in the graph
- Extract one sub-graph per event \rightarrow genelets



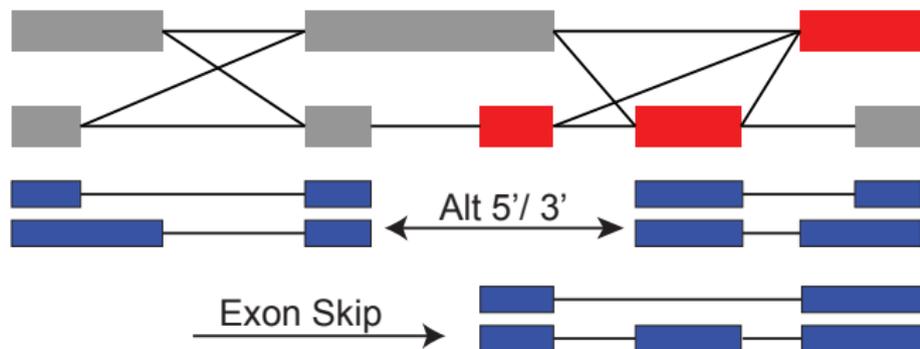
Extract Alternative Splicing Events

- Define alternative splicing events as minimal subsets of nodes in the graph
- Extract one sub-graph per event \rightarrow genelets



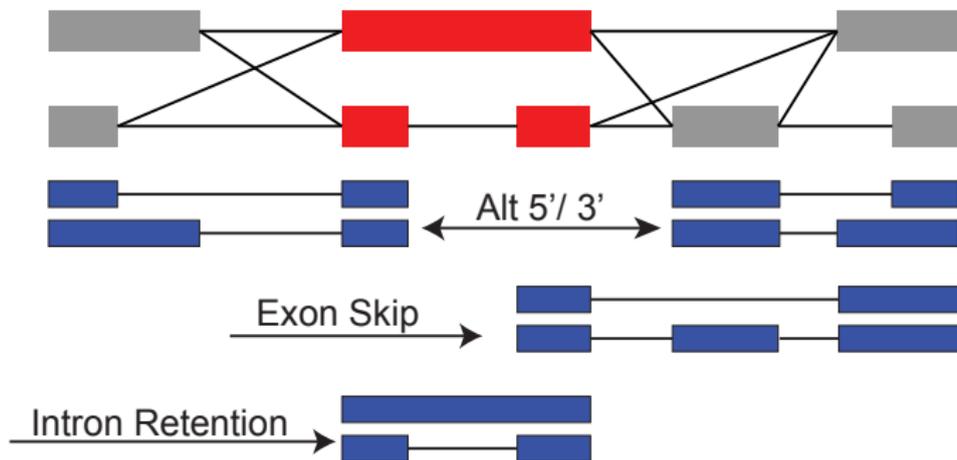
Extract Alternative Splicing Events

- Define alternative splicing events as minimal subsets of nodes in the graph
- Extract one sub-graph per event \rightarrow genelets



Extract Alternative Splicing Events

- Define alternative splicing events as minimal subsets of nodes in the graph
- Extract one sub-graph per event \rightarrow genelets



<http://bioweb.me/spladder>

Computing the Splicing Index

Utilize RNA-Seq Evidence

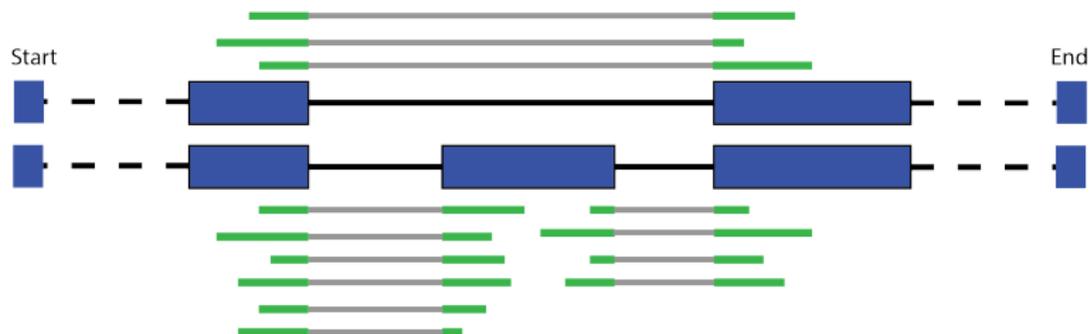
- Count read evidence for each intron edge in the graph
- Compute splicing index as count ratio between the two isoforms
- 62.5% of isoforms have the cassette exon spliced in



Computing the Splicing Index

Utilize RNA-Seq Evidence

- Count read evidence for each intron edge in the graph
- Compute splicing index as count ratio between the two isoforms
- 62.5% of isoforms have the cassette exon spliced in

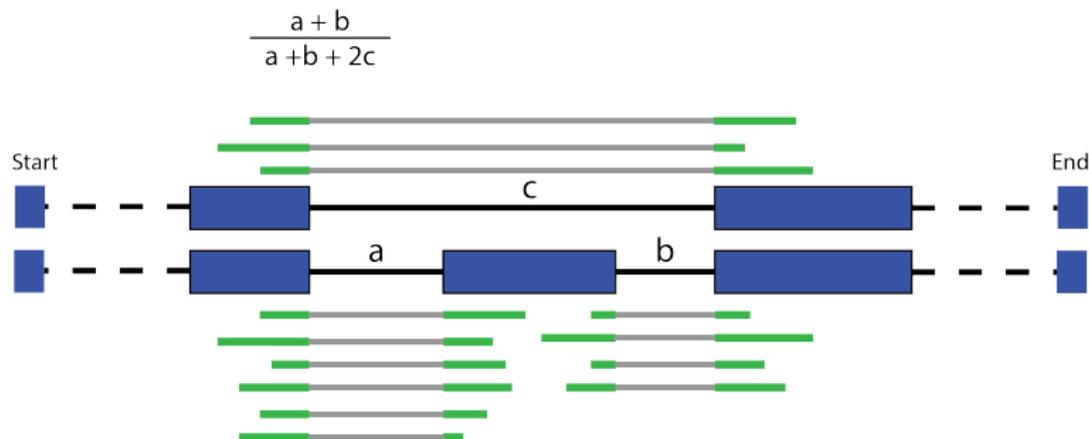


<http://bioweb.me/spladder>

Computing the Splicing Index

Utilize RNA-Seq Evidence

- Count read evidence for each intron edge in the graph
- Compute splicing index as count ratio between the two isoforms
- 62.5% of isoforms have the cassette exon spliced in

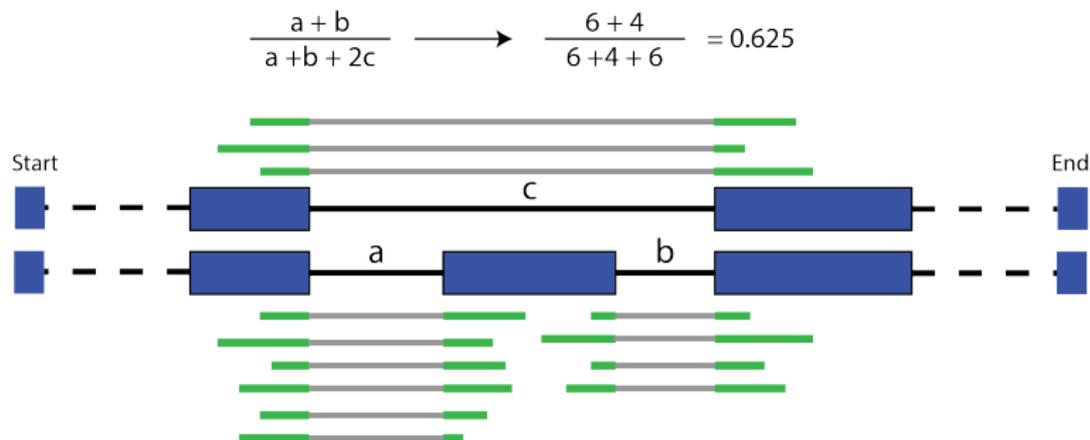


<http://bioweb.me/spladder>

Computing the Splicing Index

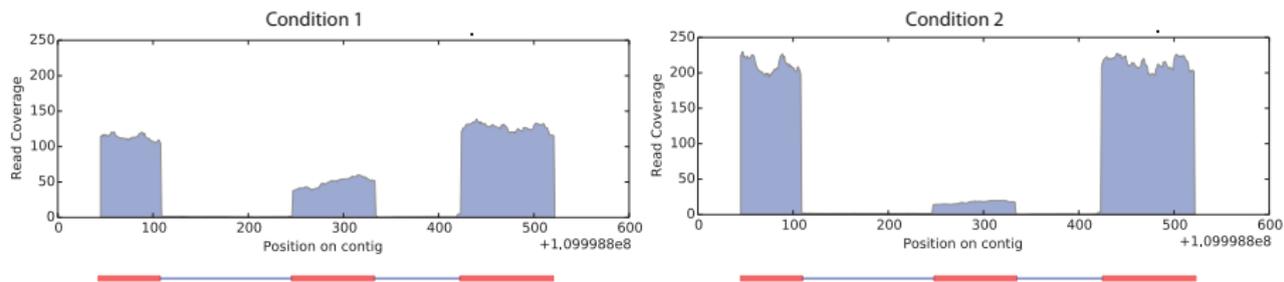
Utilize RNA-Seq Evidence

- Count read evidence for each intron edge in the graph
- Compute splicing index as count ratio between the two isoforms
- 62.5% of isoforms have the cassette exon spliced in

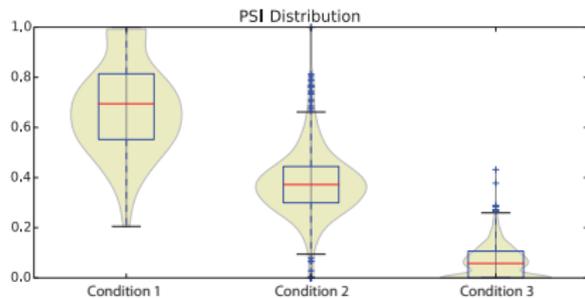


<http://bioweb.me/spladder>

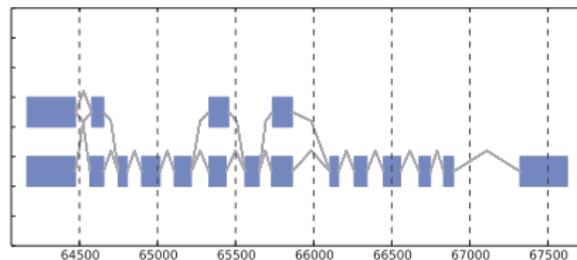
Visualizing Splicing with SplAdder



Aggregated coverage information of over multiple conditions.



Distribution of Splicing Indexes.



Splicing Graph for a gene.

<http://bioweb.me/spladder>

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-Seq data

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

Problem: Non-uniform processing (alignments & variant calling)

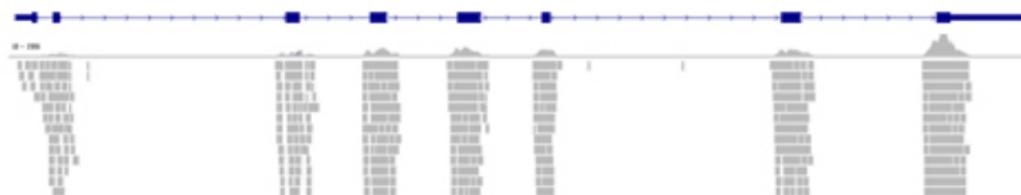
Splicing Analysis Across Multiple Cancer Types

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-Seq data

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



Problem: Non-uniform processing (alignments & variant calling)

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*)

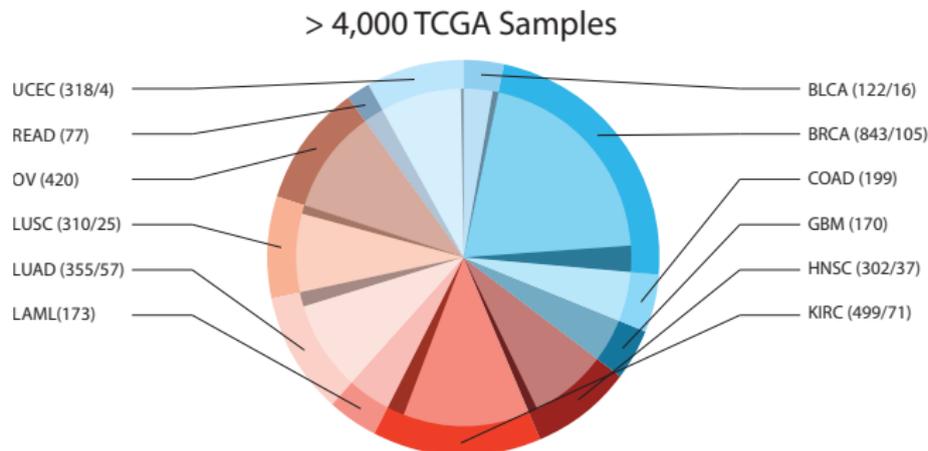
TCGA provides RNA-seq and matching exome-Seq data

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

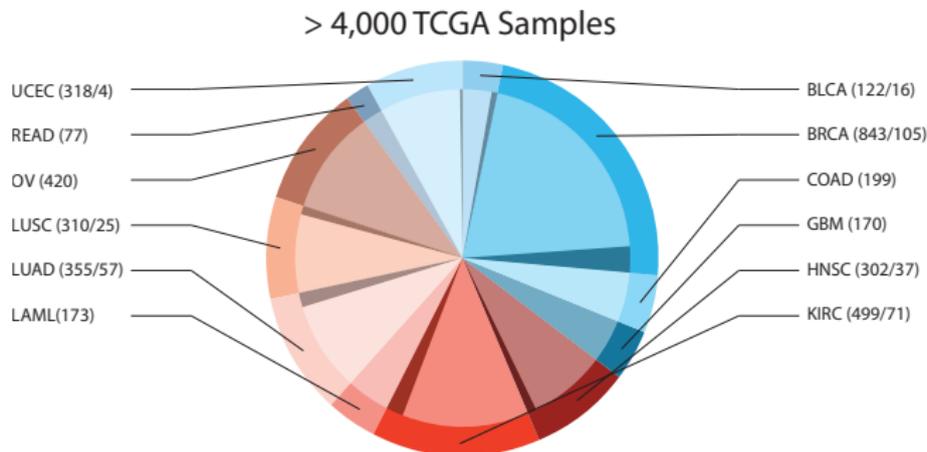


Problem: Non-uniform processing (alignments & variant calling)

RNA-Seq Data Sources



RNA-Seq Data Sources

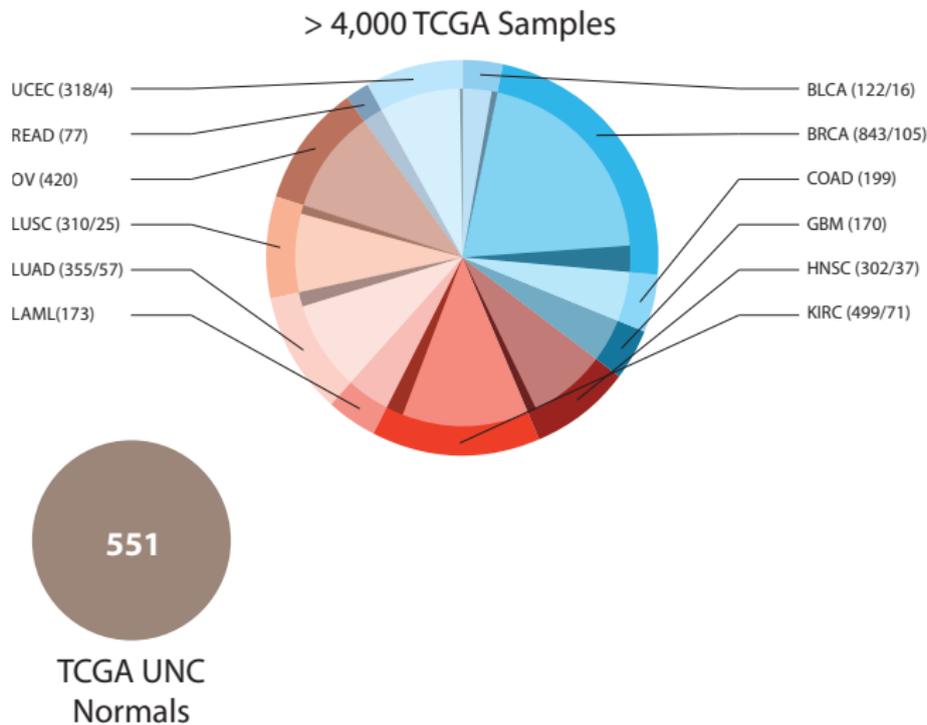


Computing at cluster colocated with CGHub

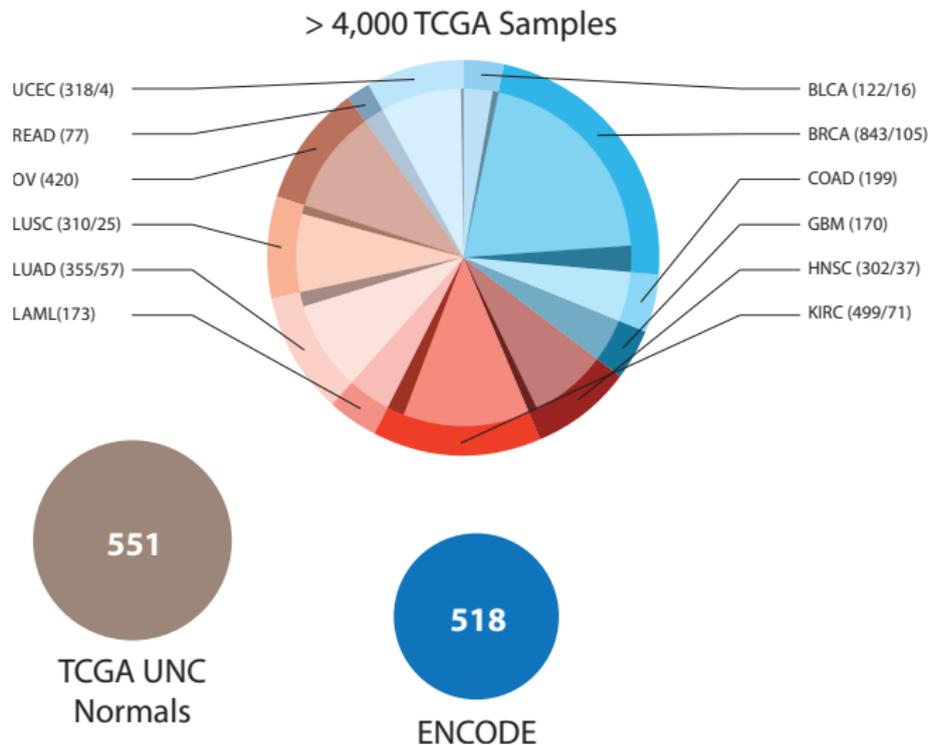
⇒ Re-mapping (STAR): ≈ 6 CPU years

⇒ Splice variant quantification (SplAdder): ≈ 1.5 CPU years

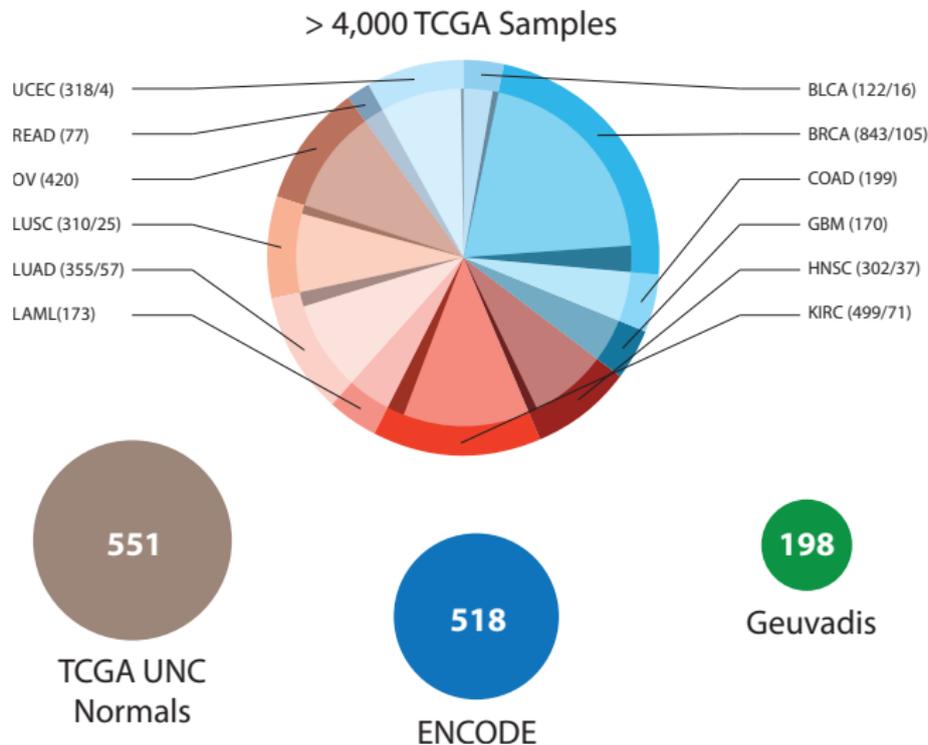
RNA-Seq Data Sources



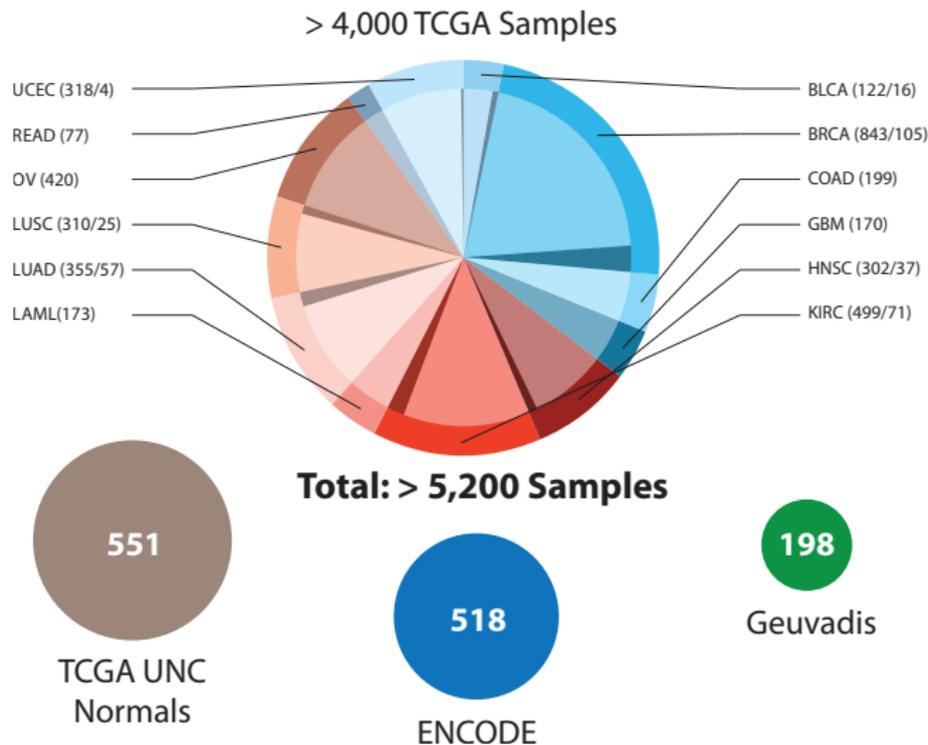
RNA-Seq Data Sources



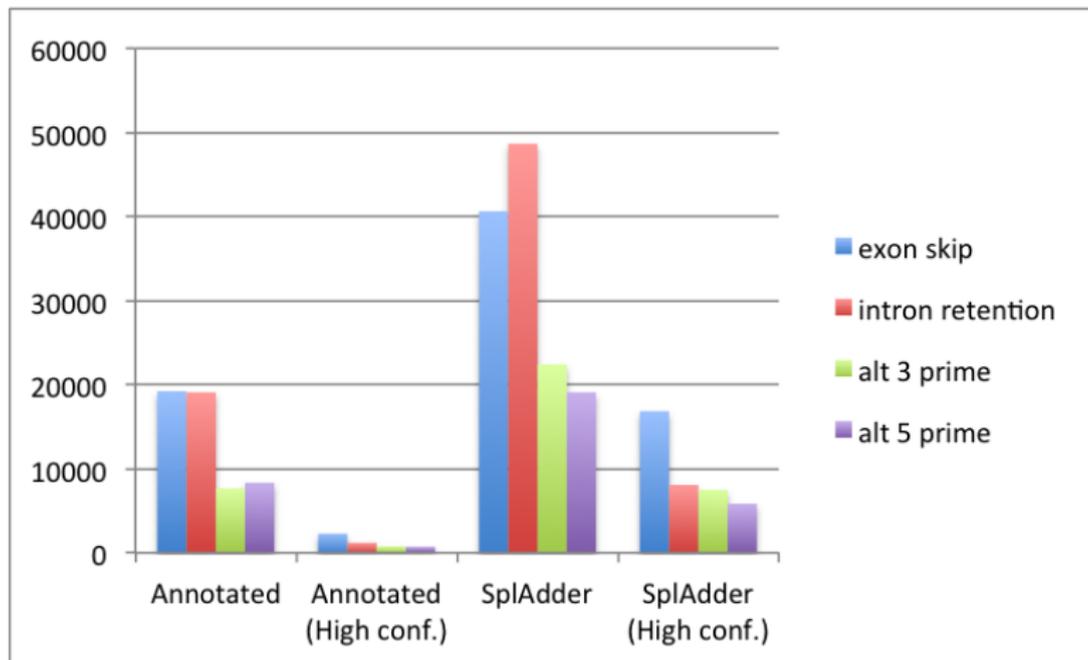
RNA-Seq Data Sources



RNA-Seq Data Sources



Splicing Variation Across $\approx 4,000$ Cancer Samples

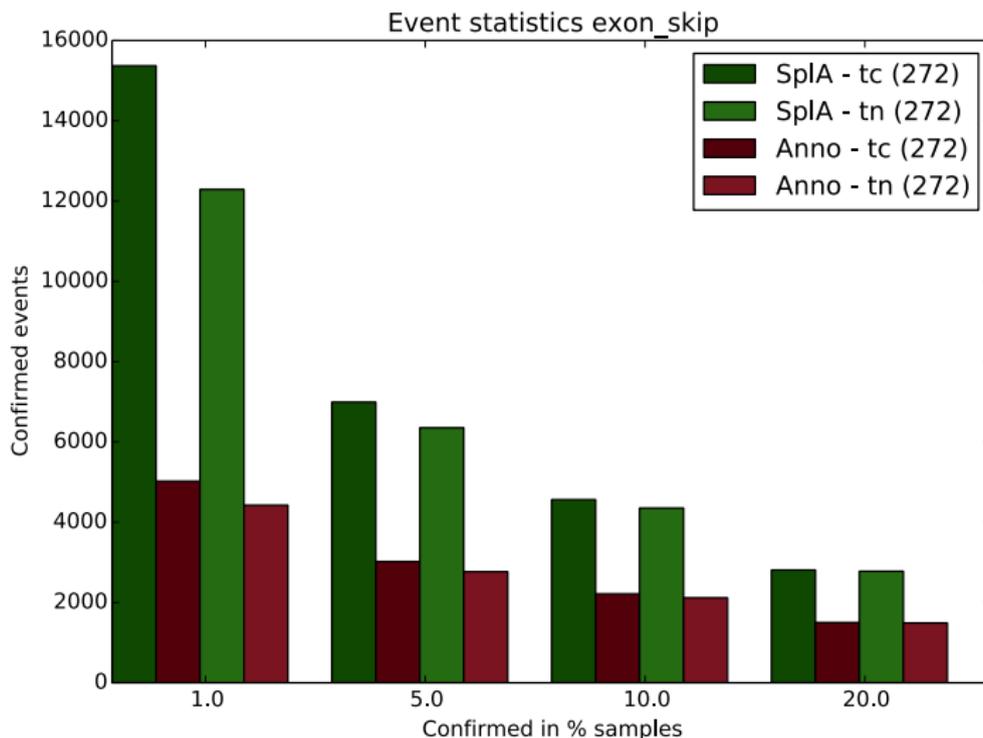


High Confidence:

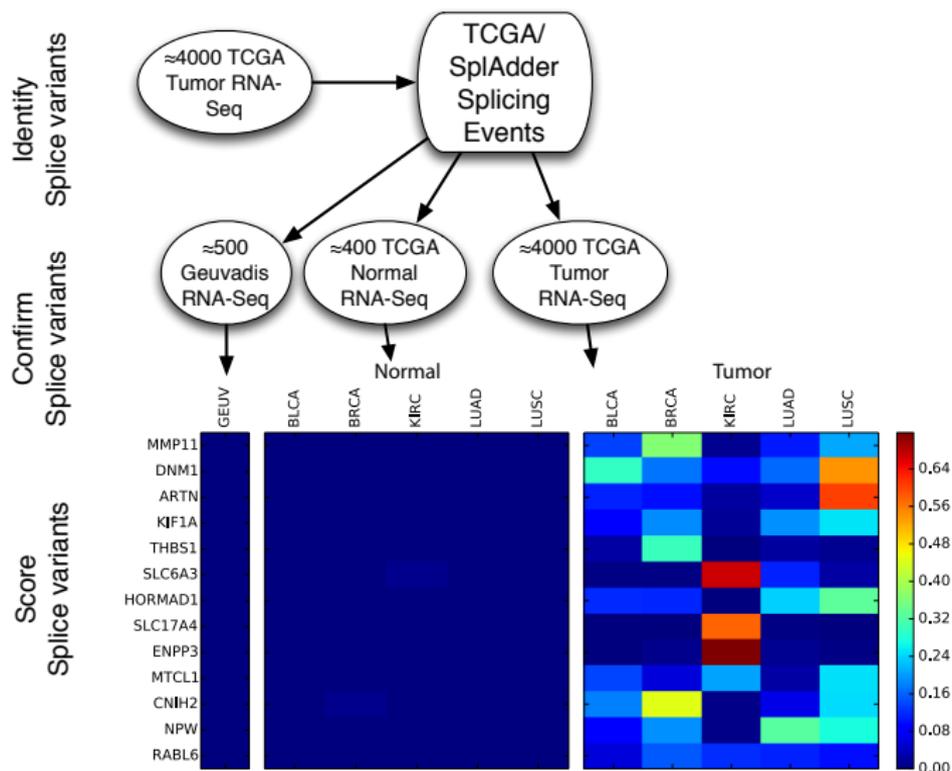
More than 10 spliced reads.

Each isoform is observed in at least 10 samples.

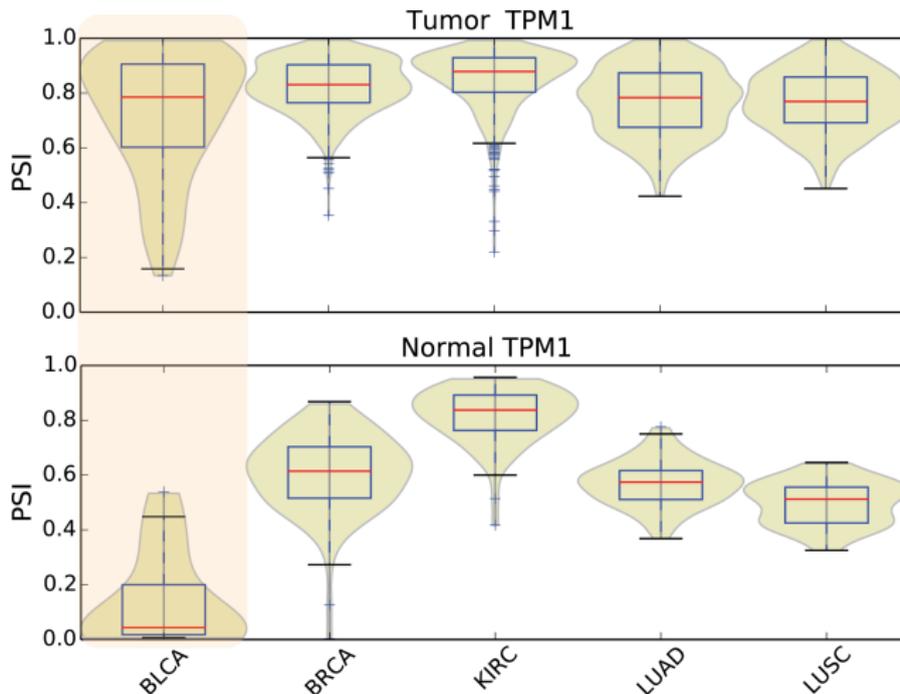
Annotation vs. SplAdder & Tumor vs. Normal



Qualitative Differences: Cancer-specific Splicing



Quantitative Differences: Shift in Abundances

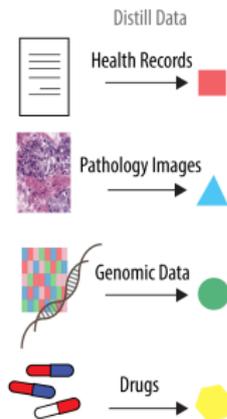


Example: TPM1 - Tropomyosin 1

Comprehensive Clinical Decision Support Systems

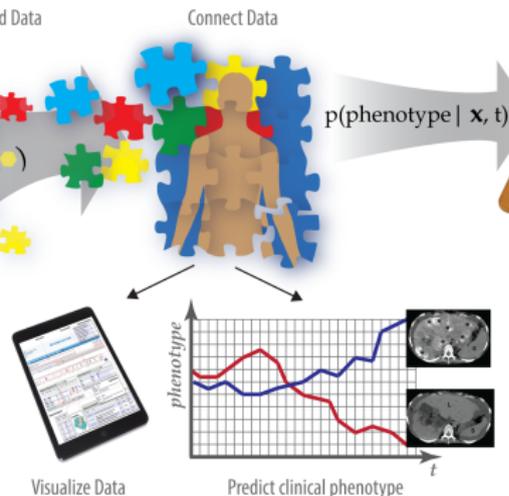
Aim 1: Data

Computable patient representation



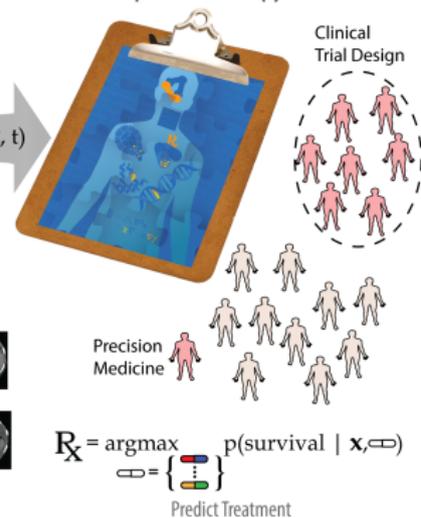
Aim 2: Knowledge

Predict clinical phenotypes

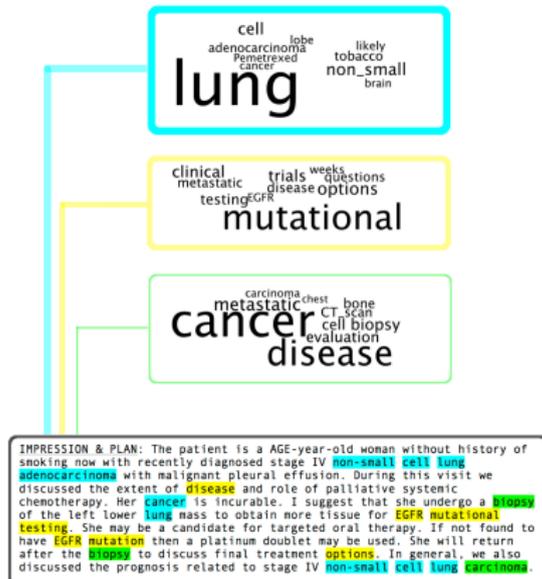


Aim 3: Action

Select optimal therapy



1st Steps: Topic Models & Molecular Pathology



Text summary in terms of **topic presence**

Statistical Association

TOP MUTATION-TOPIC CORRELATIONS BETWEEN POSITIVE MUTATION TESTS AND IMPRESSIONS & PLAN 20 TOPIC GROUP.

Mutation	Topic words	r	p -value
NRAS-Q61	melanoma, trials, options	0.31	3.8E-05
BRAF-V600	melanoma, trials, options	0.29	1.6E-07
EGFR-EXON-19	mutational, lung, testing	0.27	4.4E-05
BRAF-V600	thyroid, disease, PET	0.21	1.6E-07
EGFR-L858	mutational, lung, testing	0.16	1.8E-19
NRAS-Q61	thyroid, disease, PET	0.16	3.8E-05
EGFR-T790	mutational, lung, testing	0.16	4.8E-25
PIK3CA-H1047	breast, cancer, positive	0.14	5.0E-20
EGFR-EXON-20	mutational, lung, testing	0.11	4.5E-07

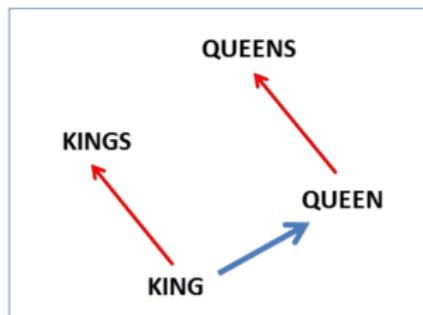
Chan et al., ..., Gardos, Artz, Rättsch, 2013
Karaletsos et al., ..., Rättsch, 2013

Patient De-ID	Mutation X
#####0	Negative
#####1	Negative
#####1	Positive
#####2	
#####3	Positive
#####4	Negative
#####4	Negative

Data: $\approx 200k$ text notes (≈ 6500 patients) + small mutation panel

Next Steps: Language Model & Normalization

- Summarization works for standardized vocabulary
- Challenging for diverse texts from many MDs
- Google: Nonlinear embedding to vector space (word2vec, 2013)



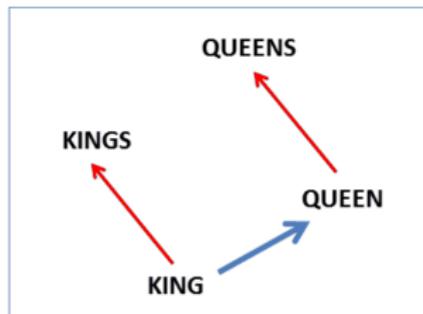
<i>Expression</i>	<i>Nearest token</i>
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

- Use all of MSKCCs text data to learn cancer language model
- “Normalize” documents for subsequent analysis

Data: $\approx 2M$ text notes ($\approx 290k$ patients)

Next Steps: Language Model & Normalization

- Summarization works for standardized vocabulary
- Challenging for diverse texts from many MDs
- Google: Nonlinear embedding to vector space (word2vec, 2013)



<i>Expression</i>	<i>Nearest token</i>
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

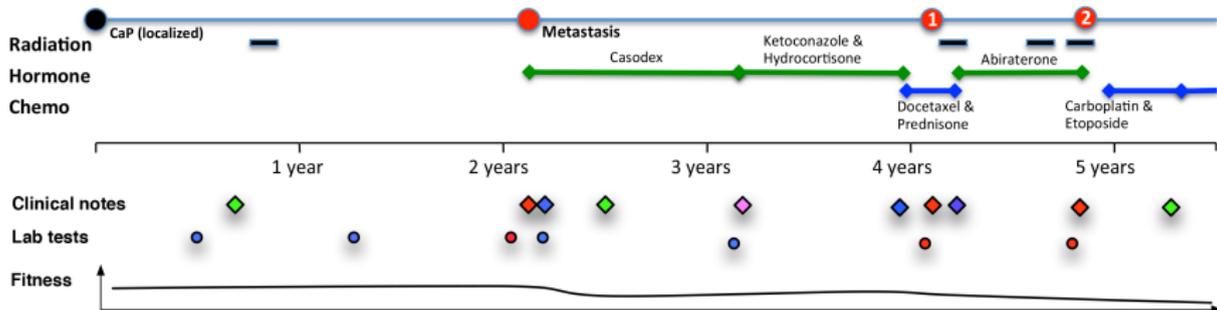
- Use all of MSKCCs text data to learn cancer language model
- “Normalize” documents for subsequent analysis

Data: $\approx 2M$ text notes ($\approx 290k$ patients)

Patient Time Lines with EHR Summaries

Summarized text can be visualized and computed on:

- Text notes can be categorized (e.g., good news/bad news)
- Data integration w/ lab tests for summary of patient 'fitness'
- Predictive models of patient/disease progression



Data: $\approx 2\text{M}$ text notes ($\approx 290\text{k}$ patients) + $\approx 15\text{k}$ genomic panels per year

[Future of] <http://cbioportal.org>

Summary

- rDiff detects differentially covered regions & has many applications
- Application: Ribosome footprinting revealed RNA G-Quadruplex elements in 5' UTR that interacts with compound via eIF4a
- SplAdder identifies and characterizes alternative splicing events
- Application: Characterize tumor/normal splicing differences; major splicing reprogramming; transmembrane proteins
- Topic models & word embeddings allow document summarization to abstract knowledge of patients
- Application: Association study between patient characteristics and somatic/germline variants of patients

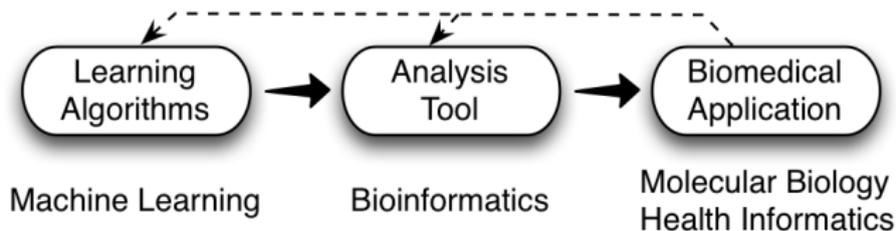
Summary

- rDiff detects differentially covered regions & has many applications
- Application: Ribosome footprinting revealed RNA G-Quadruplex elements in 5' UTR that interacts with compound via eIF4a
- SplAdder identifies and characterizes alternative splicing events
- Application: Characterize tumor/normal splicing differences; major splicing reprogramming; transmembrane proteins
- Topic models & word embeddings allow document summarization to abstract knowledge of patients
- Application: Association study between patient characteristics and somatic/germline variants of patients

Summary

- rDiff detects differentially covered regions & has many applications
- Application: Ribosome footprinting revealed RNA G-Quadruplex elements in 5' UTR that interacts with compound via eIF4a
- SplAdder identifies and characterizes alternative splicing events
- Application: Characterize tumor/normal splicing differences; major splicing reprogramming; transmembrane proteins
- Topic models & word embeddings allow document summarization to abstract knowledge of patients
- Application: Association study between patient characteristics and somatic/germline variants of patients

From Methods to Biomedical Data and Back!



Strategies:

- 1 Know your methods well. Develop & Extend. Publish.
- 2 Develop usable tools. Publish.
- 3 Identify challenging, relevant applications. Collaborate, publish.
- 4 Changed problem formulations, limitations, new analysis approaches, new data types or ideas? Go back to 1.

Acknowledgments

We gratefully acknowledge collaboration and help with the material:

- **Yi Zhong**, Sloan Kettering Institute
- **Philipp Drewe**, MDC Berlin
- **André Kahles**, Sloan Kettering Institute
- Kjong Lehmann, Sloan Kettering Institute
- **Theofanis Karaletsos**, Sloan Kettering Institute
- Oliver Stegle, EMBL-EBI
- **Guido Wendel lab**, Sloan Kettering Institute
- Nikolas Schultz, Sloan Kettering Institute
- Chris Sander, Sloan Kettering Institute

Funding by MSKCC, Max Planck Society, DFG, Geoffrey Beene Foundation, EU, NIH.

Thank You!

References I

- J. Behr, G. Schweikert, J. Cao, F. De Bona, G. Zeller, S. Laubinger, S. Ossowski, K. Schneeberger, D. Weigel, and G. Rättsch. 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>.
- 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 Rättsch, 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.
- 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. Rättsch 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. Rättsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21(Suppl. 1):i369–i377, June 2005.

References II

- 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 Rättsch. 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.
- S. Sonnenburg, G. Rättsch, 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 Rättsch. 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>.
- 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. Rättsch, 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.