Deciphering splicing with sparse regression techniques in the era of high-throughput RNA sequencing

Elsa Bernard

CBIO Mines ParisTech, Institut Curie, INSERM U900

Journées MAS 2016

Deciphering splicing with sparse regression techniques in the era of high-throughput RNA sequencing

- **(alternative) splicing.** Functional importance, human diseases, therapies.
- **RNA-seq.** Next generation sequencing of RNA molecules.
- **sparse regression.** Estimating splicing variants.

Deciphering splicing with sparse regression techniques in the era of high-throughput RNA sequencing

- **(alternative) splicing.** Functional importance, human diseases, therapies.
- **RNA-seq.** Next generation sequencing of RNA molecules.
- **sparse regression.** Estimating splicing variants.

Deciphering splicing with sparse regression techniques in the era of high-throughput RNA sequencing

- **(alternative) splicing.** Functional importance, human diseases, therapies.

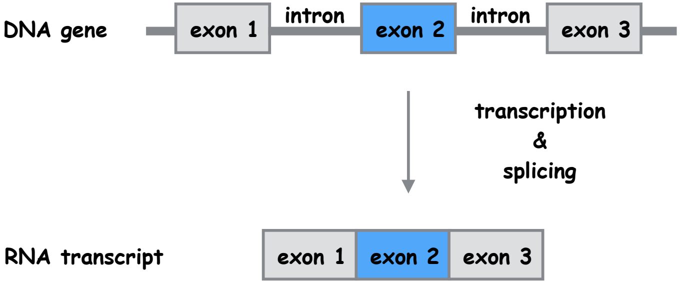
- **RNA-seq.** Next generation sequencing of RNA molecules.

- **sparse regression.** Estimating splicing variants.

Deciphering splicing with sparse regression techniques in the era of high-throughput RNA sequencing

- **(alternative) splicing.** Functional importance, human diseases, therapies.
- **RNA-seq.** Next generation sequencing of RNA molecules.
- **sparse regression.** Estimating splicing variants.

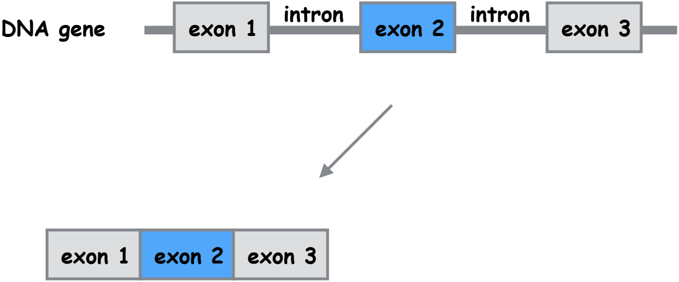
Split genes and splicing of introns



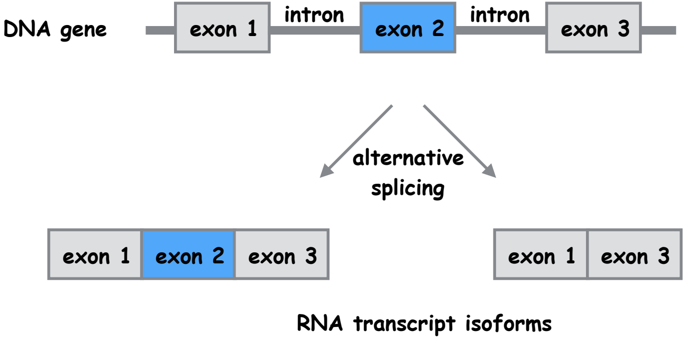
"The discovery of split genes has been of fundamental importance for today's basic research in biology, as well as for more medically oriented research concerning the development of cancer and other diseases"

Nobel Prize Press Release, 1993.

Alternative splicing produces transcript isoforms

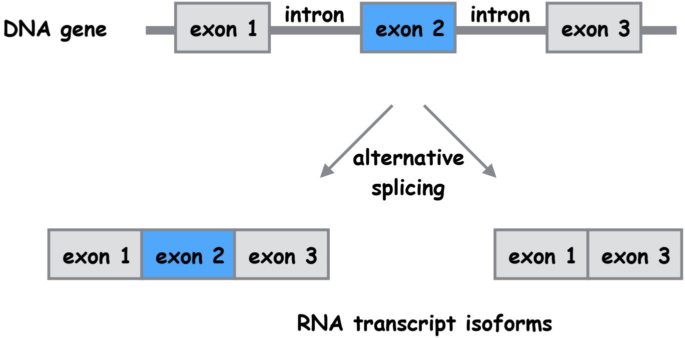


Alternative splicing produces transcript isoforms



- The splicing pattern determines the final genetic message.
- In human, 28k genes give 120k known transcript isoforms (Pal et al., 2012).

The isoform identification and quantification problem

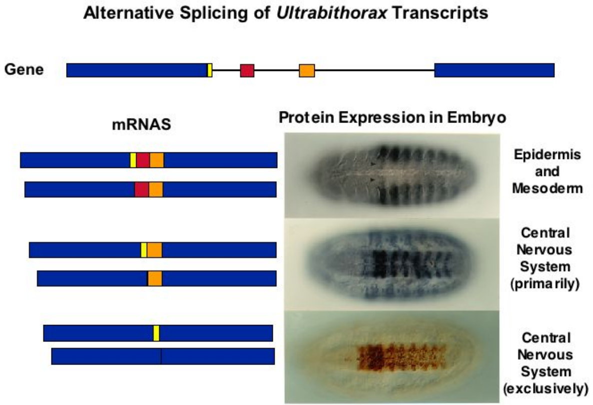


Given a biological sample, can we:

1. identify the isoforms expressed by each gene?
2. quantify their abundances?

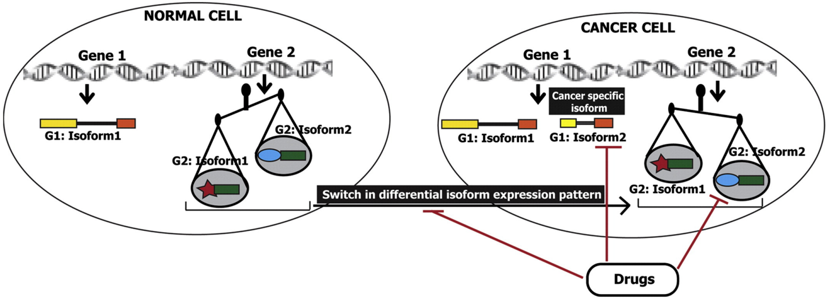
Functional importance of alternative splicing

- Developmental regulation of alternative splicing in Drosophila:


Alternative Splicing of *Ultrabithorax* Transcripts
Gene
mRNAS
Protein Expression in Embryo
Epidermis and Mesoderm
Central Nervous System (primarily)
Central Nervous System (exclusively)

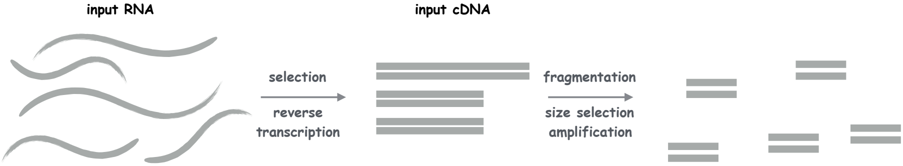
http://orchid.bio.cmu.edu/research.html

Drug targets

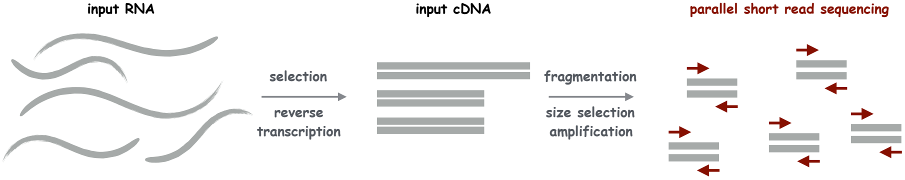


(Pal et al., 2012)

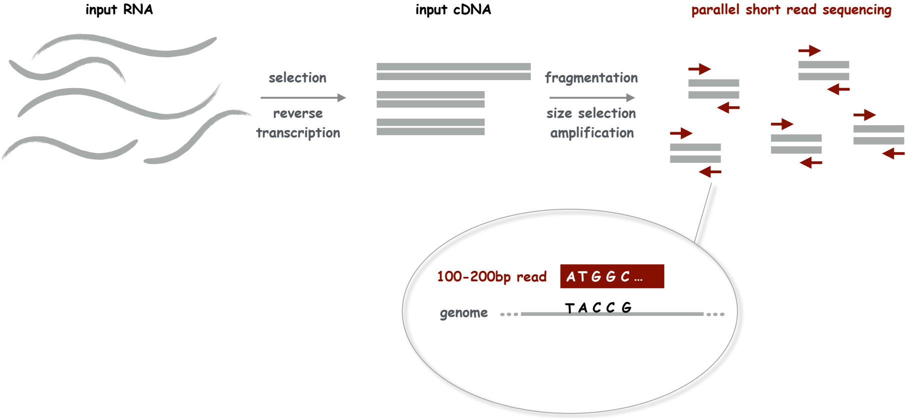
RNA-seq: shear RNA into pieces and sequence



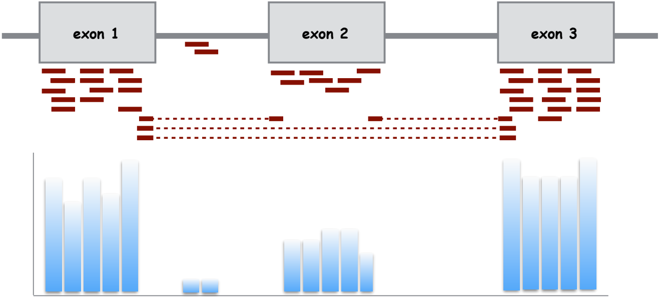
RNA-seq: shear RNA into pieces and sequence



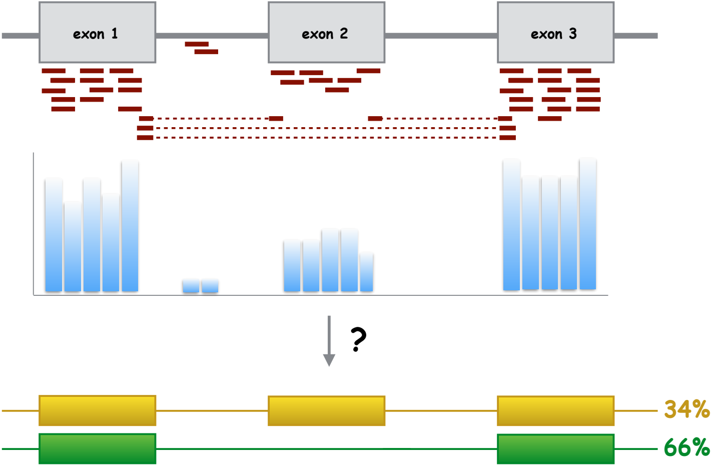
RNA-seq: shear RNA into pieces and sequence



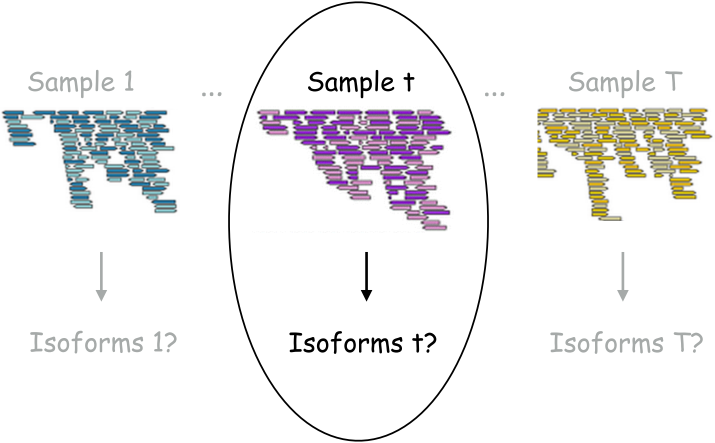
RNA-seq and alternative splicing



The isoform deconvolution problem



The one-sample case

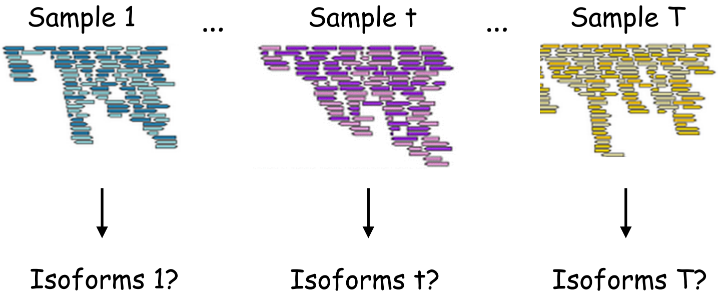


Sample 1 ... Sample t ... Sample T

Isoforms 1? Isoforms t? Isoforms T?

One-sample: can we perform accurate de novo isoform reconstruction for one given RNA-seq sample?

The multi-sample case



Sample 1 ... Sample t ... Sample T

Isoforms 1? Isoforms t? Isoforms T?

Multi-sample: can we improve isoform detection by using several samples simultaneously?

Outline

1) the one-sample case

FlipFlop Fast Lasso based Isoform Prediction as a FLOw Problem

2) the multi-sample case

Isoform detection from multiple RNA-seq samples

3) clinical application

Quantify abnormal splicing from targeted RNA-seq

Outline

1) the one-sample case

FlipFlop Fast Lasso based Isoform Prediction as a FLOw Problem

2) the multi-sample case

Isoform detection from multiple RNA-seq samples

3) clinical application

Quantify abnormal splicing from targeted RNA-seq

Outline

1) the one-sample case

FlipFlop Fast Lasso based Isoform Prediction as a FLOw Problem

2) the multi-sample case

Isoform detection from multiple RNA-seq samples

3) clinical application

Quantify abnormal splicing from targeted RNA-seq

Outline

1) the one-sample case

FlipFlop Fast Lasso based Isoform Prediction as a FLOw Problem

2) the multi-sample case

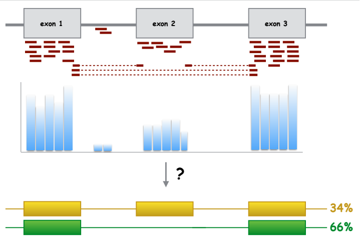
Isoform detection from multiple RNA-seq samples

3) clinical application

Quantify abnormal splicing from targeted RNA-seq

Genome-guided isoform reconstruction

- **Input:** spliced alignment of reads against reference genome
- **Goal:** reconstruct transcripts (multi-assembly problem)



MonteBello

CLASS

IsoLasso

SLIDE

CEM

Cufflinks

NSMAP

CLIIQ

FlipFlop

MiTie

Scripture

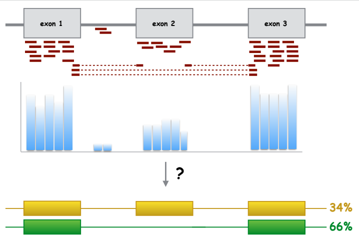
CIDANE

iReckon

StringTie

What's new?

- Input: spliced alignment of reads against reference genome
- Goal: reconstruct transcripts (multi-assembly problem)



MonteBello

CLASS

IsoLasso

SLIDE

CEM

Cufflinks

NSMAP

CLIIQ

FlipFlop

MiTie

Scripture

CIDANE

iReckon

StringTie

Contributions

1. **No need to filter** candidate transcript isoforms
2. **Faster** than existing methods that solve the same problem

} **Flow methods**

3. Adapted to **long reads** } **Particular splicing graph**
4. R package (open-access, maintained, parallelizable) } **Bioconductor**

Contributions

1. **No need to filter** candidate transcript isoforms
2. **Faster** than existing methods that solve the same problem

} **Flow methods**

3. Adapted to **long reads** } **Particular splicing graph**
4. R package (open-access, maintained, parallelizable) } **Bioconductor**

Contributions

1. **No need to filter** candidate transcript isoforms
2. **Faster** than existing methods that solve the same problem

$\Big\}$ **Flow methods**

3. Adapted to **long reads** $\Big\}$ **Particular splicing graph**
4. R package (open-access, maintained, parallelizable) $\Big\}$ **Bioconductor**

Contributions

1. **No need to filter** candidate transcript isoforms
2. **Faster** than existing methods that solve the same problem

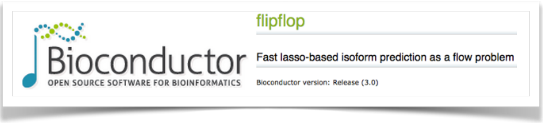
Flow methods

3. Adapted to **long reads**

Particular splicing graph

4. R package (open-access, maintained, parallelizable)

Bioconductor



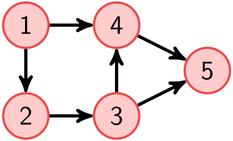
flipflop

Fast lasso-based isoform prediction as a flow problem

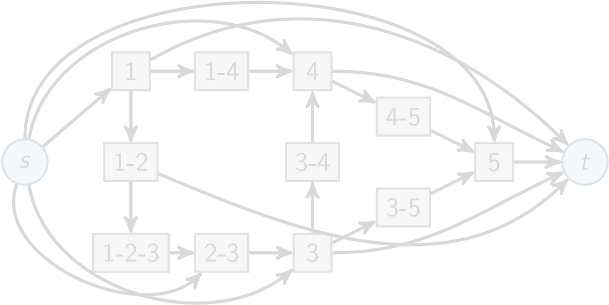
Bioconductor version: Release (3.0)

Isoforms are paths in a graph

- Splicing graph for a gene with 5 exons:

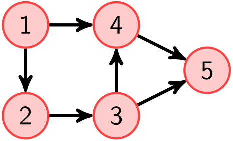


- FlipFlop graph: **1 type of read ↔ 1 node**

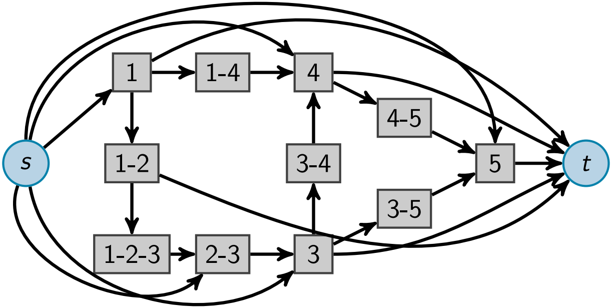


Graph adapted to long reads

- Splicing graph for a gene with 5 exons:

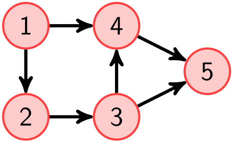


- FlipFlop graph: **1 type of read ↔ 1 node**

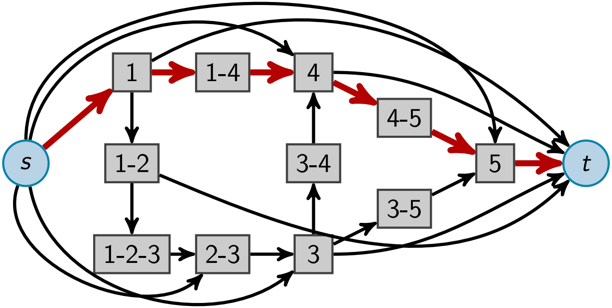


Isoforms are paths in a graph

- Splicing graph for a gene with 5 exons:

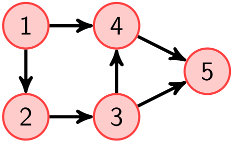


- FlipFlop graph: **one path with abundance θ_1**

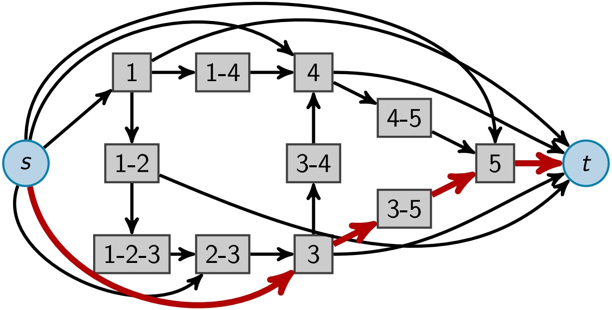


Isoforms are paths in a graph

- Splicing graph for a gene with 5 exons:



- FlipFlop graph: **another path with abundance θ_2 ...**



Select a small number of paths?

n **exons** $\rightarrow$ $\sim 2^n$ **paths/candidate isoforms**

feature selection problem with $\sim 10^3$ candidates for 10 exons
and $\sim 10^6$ for 20 exons

Minimum path cover

- Cufflinks, CLASS

✗ **do not use read counts**

Sparse regression

- IsoLasso, NSMAP, SLIDE, CEM, iReckon, MiTie, **FlipFlop**, CIDANE

✓ **use read counts**

Isoform deconvolution with the ℓ_1-norm penalization

- Estimate θ sparse by solving:

$$\min_{\theta} \underbrace{\mathcal{L}(\theta)}_{\text{fit to the data}} + \underbrace{\lambda\|\theta\|_1}_{\text{sparsity-inducing effect}}$$

big vector!

do you well explain read counts with the selected isoforms?

e.g: minus log-likelihood

you select a few isoforms among many candidates

- **Computationally challenging**
  → IsoLasso: strong filtering
  → NSMAP, SLIDE: number of exons cut-off

- **FlipFlop**
  → no filtering
  → no exon restrictions

Isoform deconvolution with the ℓ_1-norm penalization

- Estimate θ sparse by solving:

$$\min_{\underset{\textbf{big vector!}}{\overset{\uparrow}{\theta}}} \underbrace{\mathcal{L}(\theta)}_{\substack{\textbf{fit to the data}\\ \text{do you well explain}\\ \text{read counts with the}\\ \text{selected isoforms?}\\ \text{e.g: minus log-likelihood}}} + \underbrace{\lambda\|\theta\|_1}_{\substack{\textbf{sparsity-inducing effect}\\ \text{you select a few isoforms}\\ \text{among many candidates}}}$$

- **Computationally challenging**
  → IsoLasso: strong filtering
  → NSMAP, SLIDE: number of exons cut-off

- **FlipFlop**
  → no filtering
  → no exon restrictions

Fast isoform deconvolution

The isoform deconvolution problem

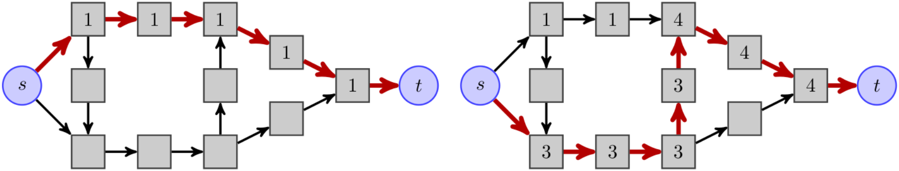
$$\min_{\theta} \; \mathcal{L}(\theta) + \lambda \|\theta\|_1 \; ,$$

is solvable in **polynomial time** with the number of nodes of the splicing graph.

Ideas:

1. the sum of isoform abundances corresponds to a **flow** on the graph
2. reformulation as a **convex cost flow problem** (Mairal and Yu, 2012)
3. recover isoforms by flow decomposition algorithm

Combinations of isoforms are flows



(a) Reads at every node corresponding to one isoform.

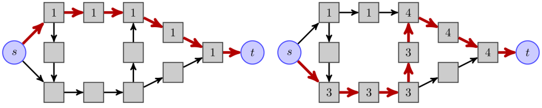
(b) Reads at every node after adding another isoform.

- Linear combinations of isoforms ⇒ Flow value on every edges
- Flow value on every edges ⇒ Paths with given value/abundance

Flow Decomposition
(linear time algorithm)

A Novel Min-Cost Flow Method for Estimating Transcript Expression with RNA-Seq. RECOMB-2013.

Equivalent flow problem (simpler!)



(a) Reads at every node corresponding to one isoform.

(b) Reads at every node after adding another isoform.

- $\mathcal{L}(\theta)$ depends only on the values of the flow on the vertices
- $\|\theta\|_1 = \sum_{\text{path } p} \theta_p = f_t$
- Therefore,

$$\min_{\theta} \mathcal{L}(\theta) + \lambda\|\theta\|_1 \quad \text{is equivalent to} \quad \min_{f \text{ flow}} \tilde{\mathcal{L}}(f) + \lambda f_t$$

FlipFlop Summary

Isoform detection = Path selection problem

$\sim 2^n$ variables (all paths in the splicing graph)

⇕

Equivalent network flow problem

$\sim \frac{n^2}{2}$ variables (all nodes of the splicing graph)

↓

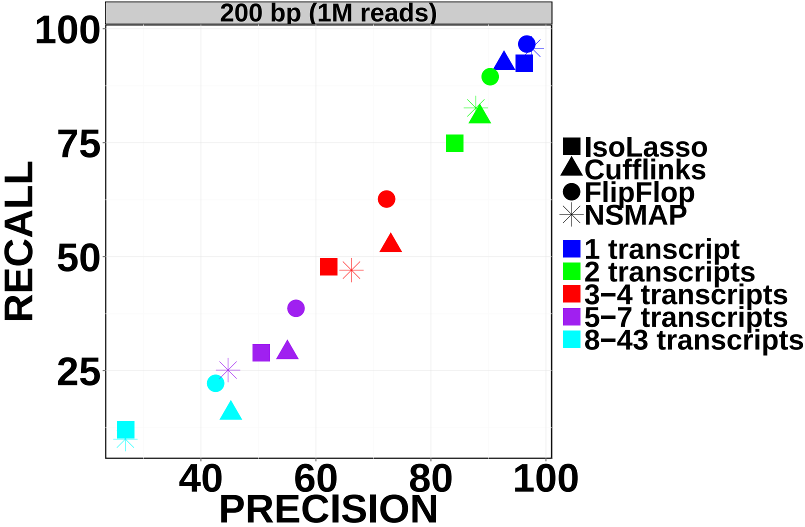
Network flow algorithms

Efficient algorithms. Polynomial time.

Human Simulation: precision / recall

hg19, 1137 genes on chr1, 1million 200 bp single-end reads by transcript levels.

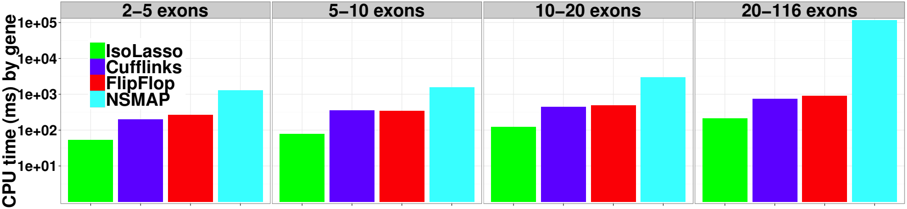
Simulator: `http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html`



Speed Trial

hg19, 1137 genes on chr1, 1million reads by exon levels.

Simulator: `http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html`



One-sample case summary

FlipFlop → transcripts reconstruction over an exponential number of candidates in polynomial time

- `http://cbio.ensmp.fr/flipflop/`
- `http://cbio.ensmp.fr/flipflop/experiments.html`
- R package

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("flipflop")
```

E. Bernard, L. Jacob, J. Mairal and J.-P. Vert. **Efficient RNA isoform identification and quantification from RNA-seq data with network flows**. *Bioinformatics*, 2014.

Outline

1) the one-sample case

FlipFlop Fast Lasso based Isoform Prediction as a FLOw Problem

2) the multi-sample case

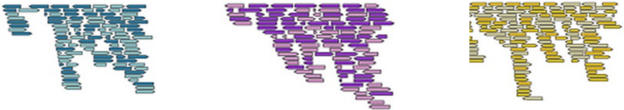
Isoform detection from multiple RNA-seq samples

3) clinical application

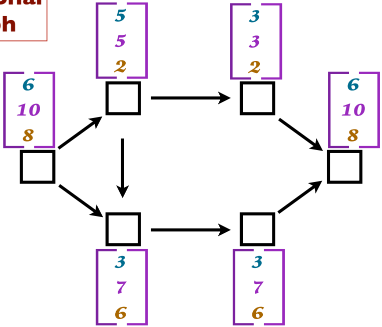
Quantify abnormal splicing from targeted RNA-seq

Multi-dimensional case

Sample 1 • • • • Sample t • • • • Sample T

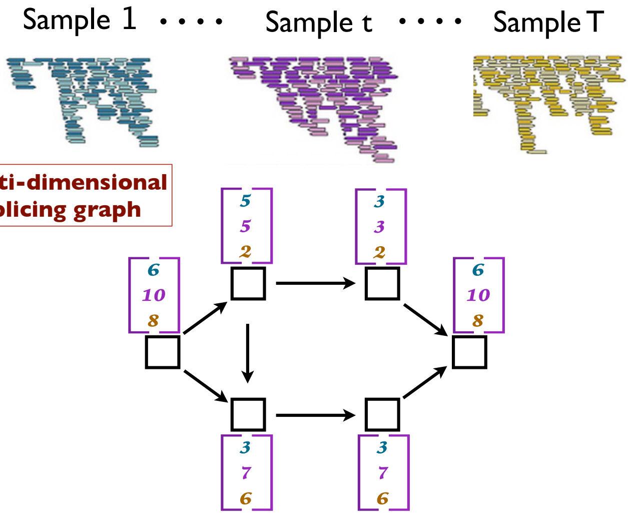


Multi-dimensional splicing graph



Can we find a sparse set of paths that explains the multi-dimensional read counts?

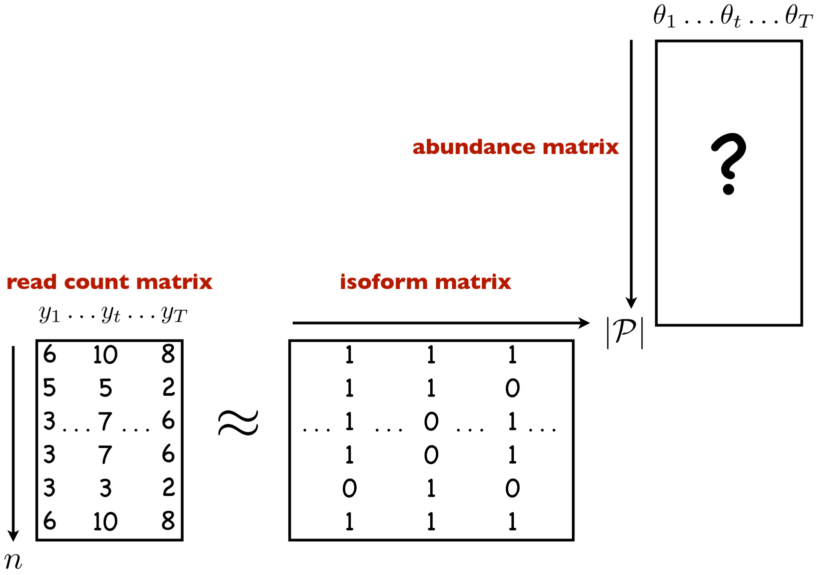
Multi-dimensional case



Multi-dimensional splicing graph

Can we find a sparse set of paths that explains the multi-dimensional read counts?

Group-Lasso strategy



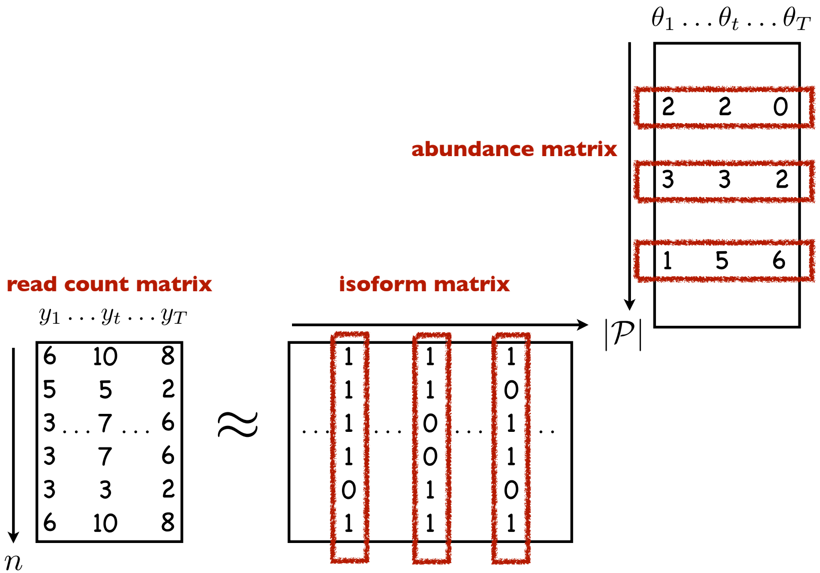
read count matrix

isoform matrix

abundance matrix

$$\begin{array}{c|c} & y_1 \ldots y_t \ldots y_T \\ \hline n \downarrow & \begin{array}{|ccc|} 6 & 10 & 8 \\ 5 & 5 & 2 \\ 3 \ldots & 7 & \ldots 6 \\ 3 & 7 & 6 \\ 3 & 3 & 2 \\ 6 & 10 & 8 \end{array} \end{array} \approx \begin{array}{c} \xrightarrow{\hspace{6em}} |\mathcal{P}| \\ \begin{array}{|ccccccc|} & 1 & & 1 & & 1 & \\ & 1 & & 1 & & 0 & \\ \ldots & 1 & \ldots & 0 & \ldots & 1 & \ldots \\ & 1 & & 0 & & 1 & \\ & 0 & & 1 & & 0 & \\ & 1 & & 1 & & 1 & \end{array} \end{array} \qquad \begin{array}{c|c} & \theta_1 \ldots \theta_t \ldots \theta_T \\ \hline |\mathcal{P}| \downarrow & \boxed{\;?\;} \end{array}$$

Group-Lasso strategy



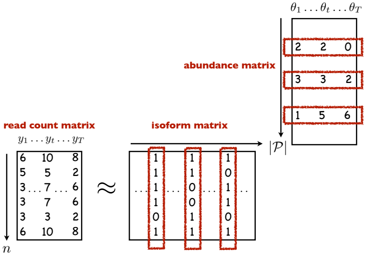
abundance matrix

read count matrix

isoform matrix

$$\begin{array}{c} y_1 \dots y_t \dots y_T \\ \left[\begin{array}{ccc} 6 & 10 & 8 \\ 5 & 5 & 2 \\ 3 \dots & 7 \dots & 6 \\ 3 & 7 & 6 \\ 3 & 3 & 2 \\ 6 & 10 & 8 \end{array}\right] \end{array} \approx \left[\begin{array}{ccccccc} & 1 & & 1 & & 1 & \\ & 1 & & 1 & & 0 & \\ \dots & 1 & \dots & 0 & \dots & 1 & \dots \\ & 1 & & 0 & & 1 & \\ & 0 & & 1 & & 0 & \\ & 1 & & 1 & & 1 & \end{array}\right] \quad \begin{array}{c} \theta_1 \dots \theta_t \dots \theta_T \\ \left[\begin{array}{ccc} 2 & 2 & 0 \\ 3 & 3 & 2 \\ 1 & 5 & 6 \end{array}\right] \end{array}$$

More formally



abundance matrix

$$\begin{array}{ccc} \theta_1 & \ldots\theta_t\ldots & \theta_T \\ 2 & 2 & 0 \\ 3 & 3 & 2 \\ 1 & 5 & 6 \end{array}$$

read count matrix

isoform matrix

$$\begin{array}{ccc} y_1 & \ldots y_t \ldots & y_T \\ 6 & 10 & 8 \\ 5 & 5 & 2 \\ 3 & \ldots 7 \ldots & 6 \\ 3 & 7 & 6 \\ 3 & 3 & 2 \\ 6 & 10 & 8 \end{array} \approx \begin{array}{ccccccc} & 1 & & 1 & & 1 & \\ & 1 & & 1 & & 0 & \\ \ldots & 1 & \ldots & 0 & \ldots & 1 & \ldots \\ & 1 & & 0 & & 1 & \\ & 0 & & 1 & & 0 & \\ & 1 & & 1 & & 1 & \end{array}$$

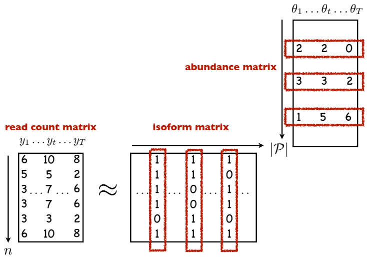
- each isoform defines a **group** $\boldsymbol{\theta}_p = \{\theta_p^t, t \in [\![1, T]\!]\}$
- the multi-sample loss is the sum of the independent losses

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \text{loss}(y_t, \theta_t)$$

- ideally we want to solve the NP-hard ℓ_0 problem

$$\min_{\{\boldsymbol{\theta}_p\}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \sum_{p \in \mathcal{P}} \mathbf{1}_{\{\boldsymbol{\theta}_p \neq \mathbf{0}\}}$$

More formally



abundance matrix
$\theta_1 \dots \theta_t \dots \theta_T$
read count matrix
$y_1 \dots y_t \dots y_T$
isoform matrix
$|\mathcal{P}|$
n

$$\begin{array}{ccc} y_1 & \dots y_t \dots & y_T \\ \hline 6 & 10 & 8 \\ 5 & 5 & 2 \\ 3 & \dots 7 \dots & 6 \\ 3 & 7 & 6 \\ 3 & 3 & 2 \\ 6 & 10 & 8 \end{array} \approx \begin{array}{ccccccc} \dots & 1 & \dots & 1 & \dots & 1 & \dots \\ & 1 & & 1 & & 0 & \\ \dots & 1 & \dots & 0 & \dots & 1 & \dots \\ & 1 & & 0 & & 1 & \\ & 0 & & 1 & & 0 & \\ & 1 & & 1 & & 1 & \end{array} \qquad \begin{array}{ccc} \theta_1 & \dots \theta_t \dots & \theta_T \\ \hline 2 & 2 & 0 \\ 3 & 3 & 2 \\ 1 & 5 & 6 \end{array}$$

- each isoform defines a **group** $\boldsymbol{\theta}_p = \{\theta_p^t, t \in [\![1, T]\!]\}$
- the multi-sample loss is the sum of the independent losses

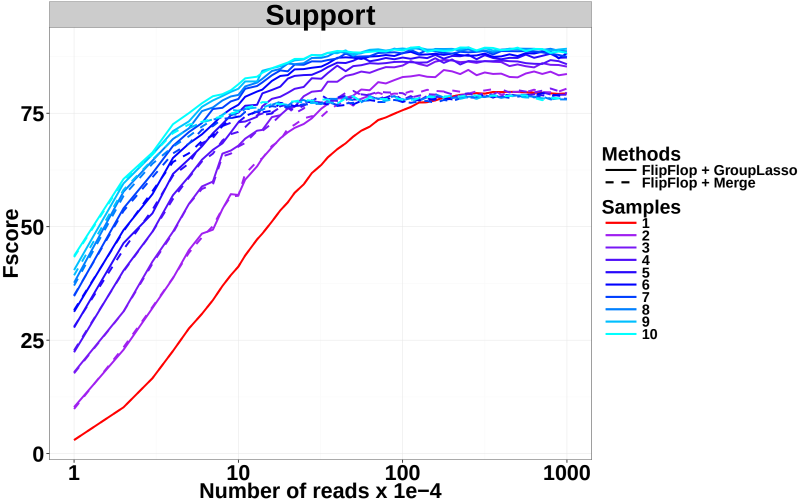
$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \text{loss}(y_t, \theta_t)$$

- instead we solve the **group-lasso convex relaxation**

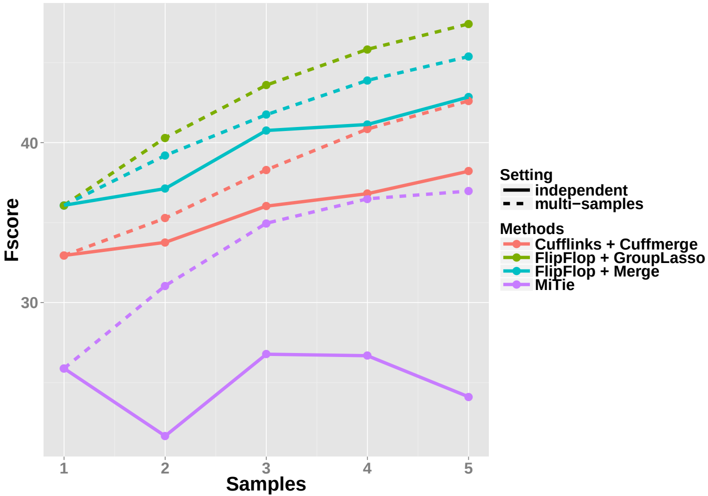
$$\min_{\{\boldsymbol{\theta}_p\}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \sum_{p \in \mathcal{P}} \|\boldsymbol{\theta}_p\|_2$$

Simulation: GroupLasso vs Merging

$$\forall t \in \{1, \dots, T\}, \mathrm{supp}\theta_t = \mathrm{supp}\theta_o$$



modENCODE data
Time course development of D.melanogaster



Multi-sample case summary

FlipFlop → transcript reconstruction using several samples simultaneously leads to more statistical power

- `http://cbio.ensmp.fr/flipflop/details.html`

E. Bernard, L. Jacob, J. Mairal, E. Viara and J.-P. Vert. **A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples.** *BMC Bioinformatics*, 2015.

Outline

1) the one-sample case

FlipFlop Fast Lasso based Isoform Prediction as a FLOw Problem

2) the multi-sample case

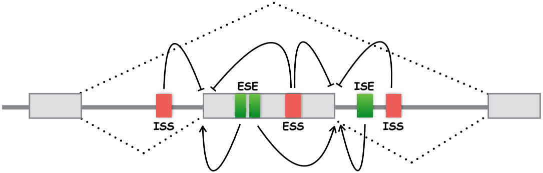
Isoform detection from multiple RNA-seq samples

3) clinical application

Quantify abnormal splicing from targeted RNA-seq

Molecular diagnosis and splicing

- Various splicing enhancing and silencing motifs:

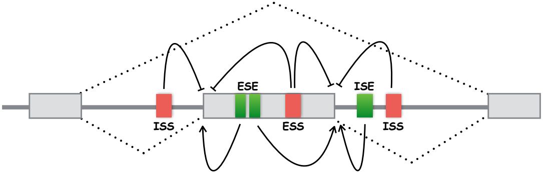


- Variants disrupting/creating these consensus sequences can affect normal splicing

⇒ **molecular diagnosis:** correct interpretation of these variants on splicing is imperative for genetic counseling

Molecular diagnosis and splicing

- Various splicing enhancing and silencing motifs:



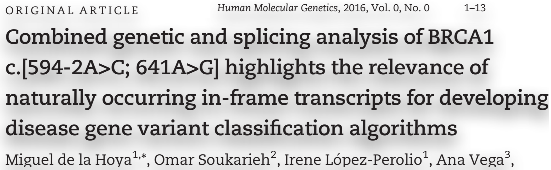
- Variants disrupting/creating these consensus sequences can affect normal splicing

Development of a new diagnostic tool

- time and cost-effective identification and quantification of transcripts using targeted high-throughput RNA-seq
- extension of sparse regression techniques to a new experimental design

Promising results on BRCA1

- BRCA1: Breast Cancer susceptibility gene
- Involved in DNA repair pathway and cell cycle
- High number of splicing events (regulated in a cell-cycle- and cell-type-specific manner)



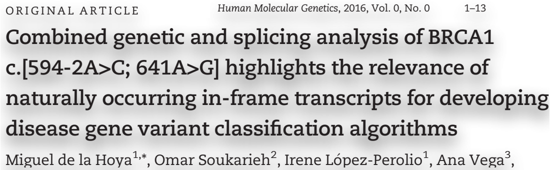
ORIGINAL ARTICLE *Human Molecular Genetics*, 2016, Vol. 0, No. 0 1–13

Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms

Miguel de la Hoya[1,*], Omar Soukarieh[2], Irene López-Perolio[1], Ana Vega[3],

Promising results on BRCA1

- BRCA1: Breast Cancer susceptibility gene
- Involved in DNA repair pathway and cell cycle
- High number of splicing events (regulated in a cell-cycle- and cell-type-specific manner)



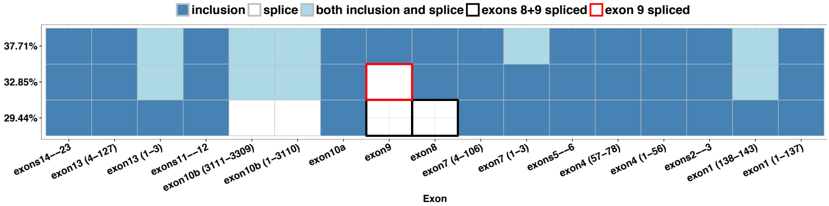
ORIGINAL ARTICLE

Human Molecular Genetics, 2016, Vol. 0, No. 0 1–13

Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms

Miguel de la Hoya[1,*], Omar Soukarieh[2], Irene López-Perolio[1], Ana Vega[3],

Accurate quantification of overlapping splicing events:



Thanks

Laurent Jacob

JP Vert

Julien Mairal

Eric Viara

Elodie Girard

Supplementary Slides

Part 1: one-sample approach

FlipFlop Fast Lasso based Isoform Prediction as a FLOw Problem

Technical details

Poisson Loss:

$$\mathcal{L}(\theta) = \sum_{u \in V} \left[Nl_u \left(\sum_{\text{path } p \ni u} \theta_p \right) - \mathbf{y}_u \log \left(Nl_u \sum_{\text{path } p \ni u} \theta_p \right) \right]$$

Flow Decomposition:

$$f_{uv} = \sum_{\text{path } p \ni (u,v)} \theta_p$$

$$\Rightarrow \; f_v = \sum_{u \in V} f_{uv} = \sum_{\text{path } p \ni v} \theta_p$$

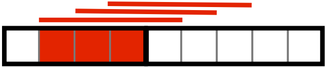
Convex Cost Flow:

$$\min_{f \text{ flow}} \sum_{u \in V} [Nl_u f_u - \mathbf{y}_u \log(f_u)] + \lambda f_t$$

Solved using ϵ-relaxation method (Bertsekas 1998)

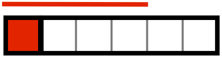
Effective length

1) $l_{\text{left}} \geq L, \quad l_{\text{right}} \geq L$



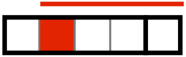
$l_i = L - 1$

2) $l_{\text{left}} < L, \quad l_{\text{right}} \geq L$



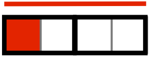
$l_i = l_{\text{left}}$

3) $l_{\text{left}} \geq L, \quad l_{\text{right}} < L$



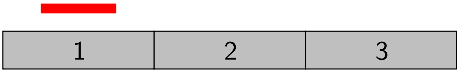
$l_i = l_{\text{right}}$

4) $l_{\text{left}} < L, \quad l_{\text{right}} < L$

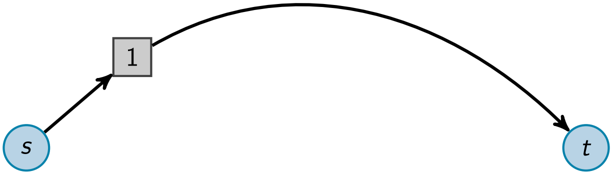


$l_i = l_{\text{left}} + l_{\text{right}} - L + 1$

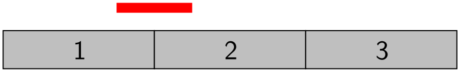
Graph adapted to long reads



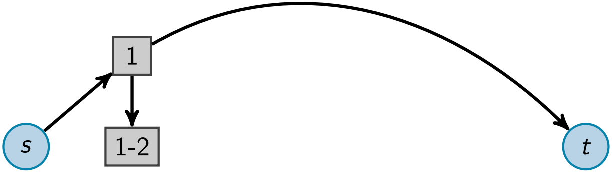
- FlipFlop graph:



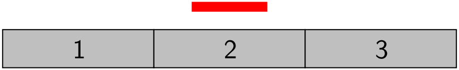
Graph adapted to long reads



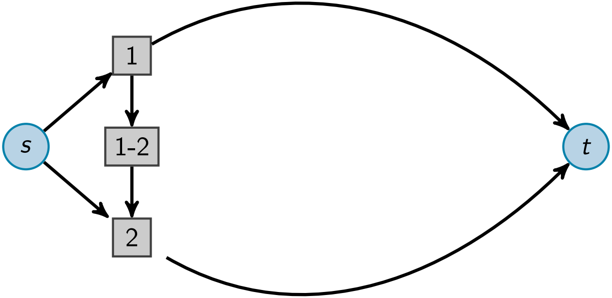
- FlipFlop graph:



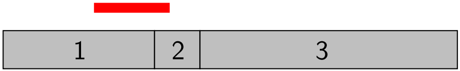
Graph adapted to long reads



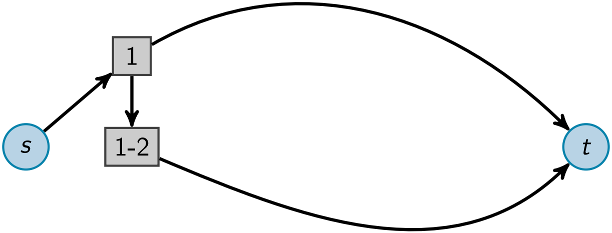
- FlipFlop graph:



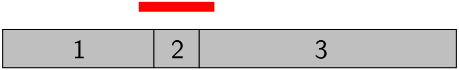
Graph adapted to long reads



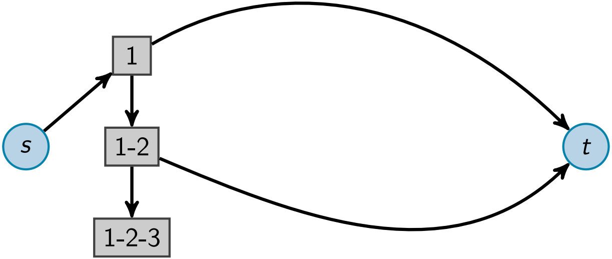
- FlipFlop graph:



Graph adapted to long reads

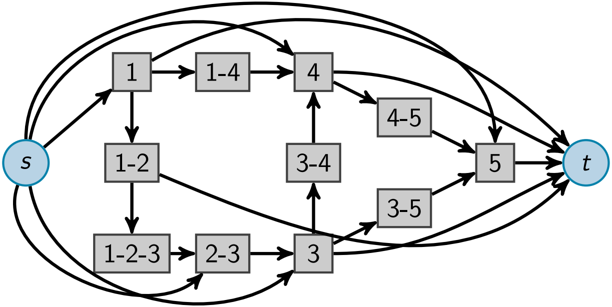


- FlipFlop graph:



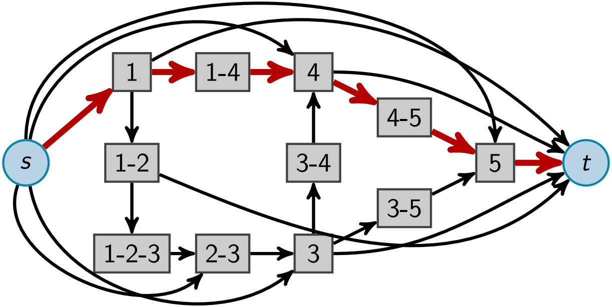
Graph adapted to long reads

- FlipFlop graph:



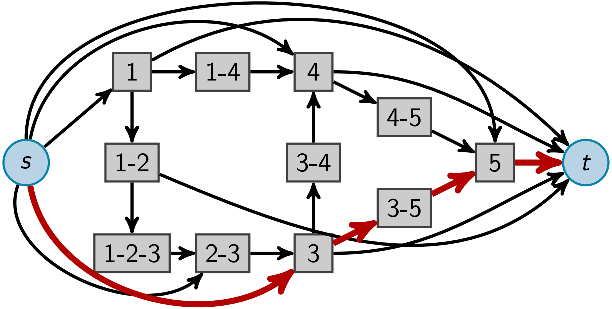
Graph adapted to long reads

- FlipFlop graph:

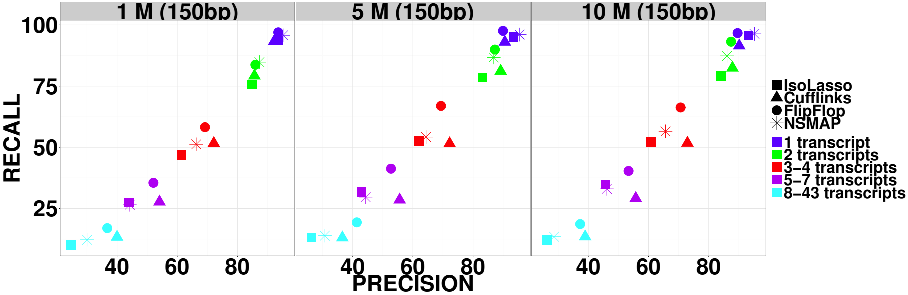


Graph adapted to long reads

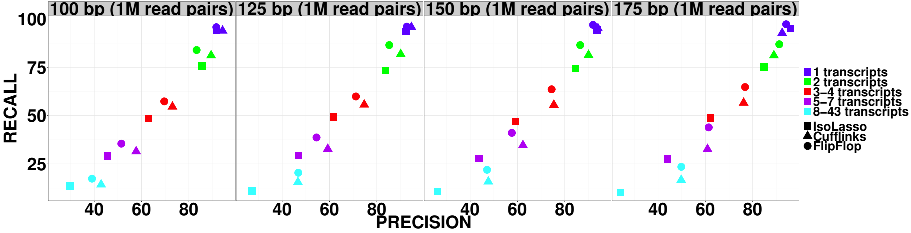
- FlipFlop graph:



Performance increases with coverage

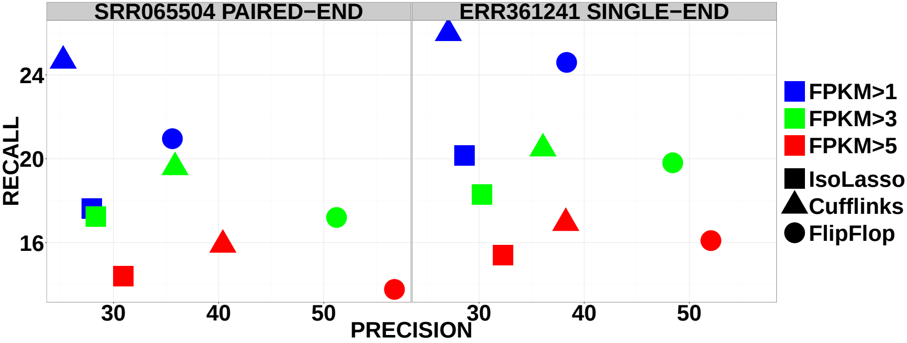


Extension to paired-end reads OK

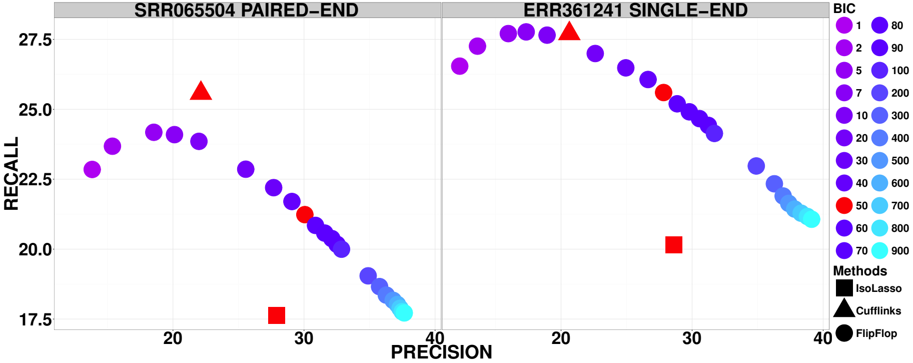


Real Data

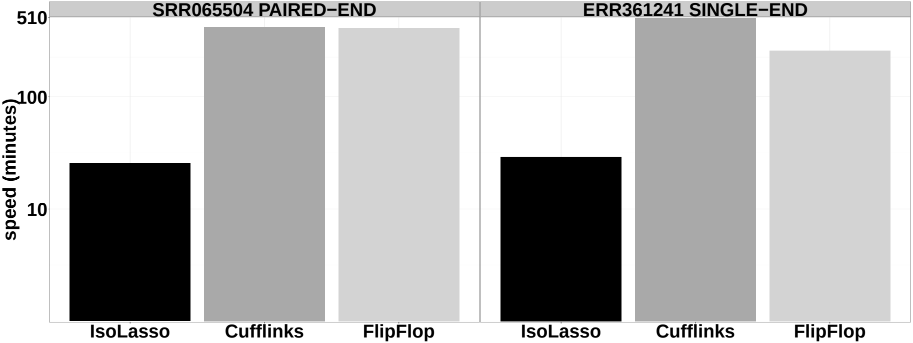
Human: 50 million 75bp reads.



Precision-Recall curves on real data



Speed comparison on real data

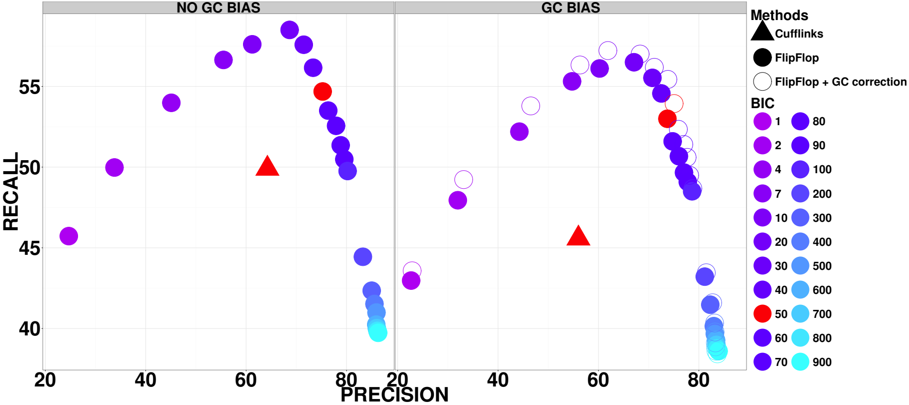


GC bias - Precision-Recall curve

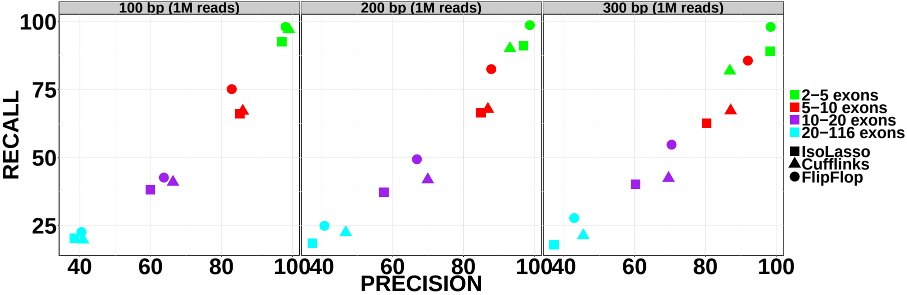
hg19, chr1, 4140 transcripts, 2million 150bp single-end reads

Simulator: FluxSimulator http://sammeth.net/confluence/display/SIM/Home

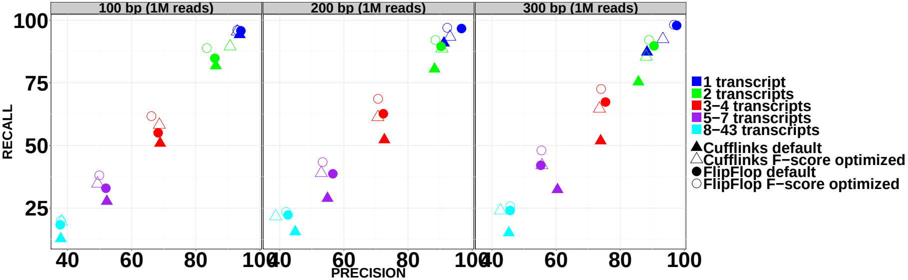
Model selection: set of solutions minimizing $\mathcal{L}(\theta) + \lambda\|\theta\|_1$ for different values of $\lambda \rightarrow$ BIC criteria



Exon stratification

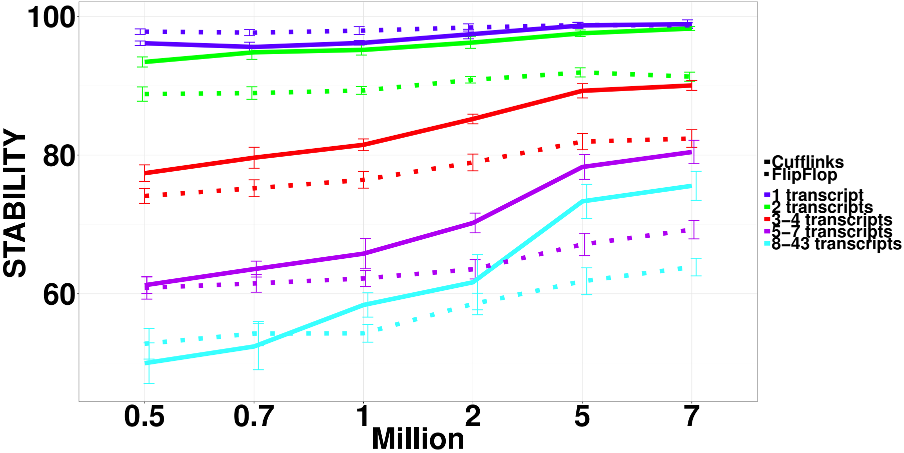


Tuning



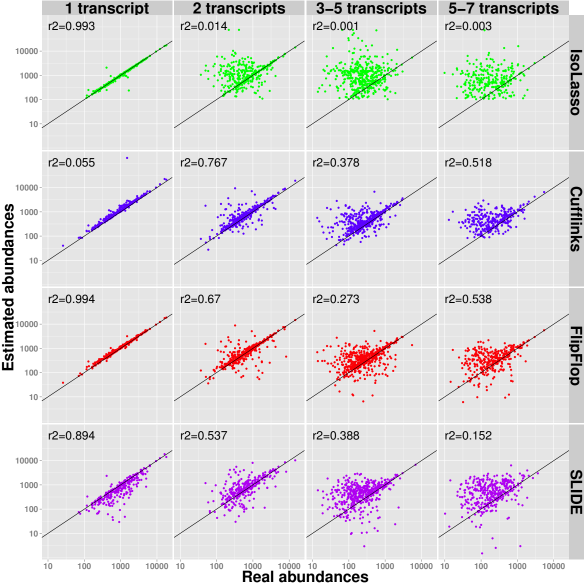
100 bp (1M reads)
200 bp (1M reads)
300 bp (1M reads)
RECALL
100
75
50
25
40
60
80
100
PRECISION
🟦 1 transcript
🟩 2 transcripts
🟥 3−4 transcripts
🟪 5−7 transcripts
🟦 8−43 transcripts
▲ Cufflinks default
△ Cufflinks F−score optimized
● FlipFlop default
○ FlipFlop F−score optimized

Stability study



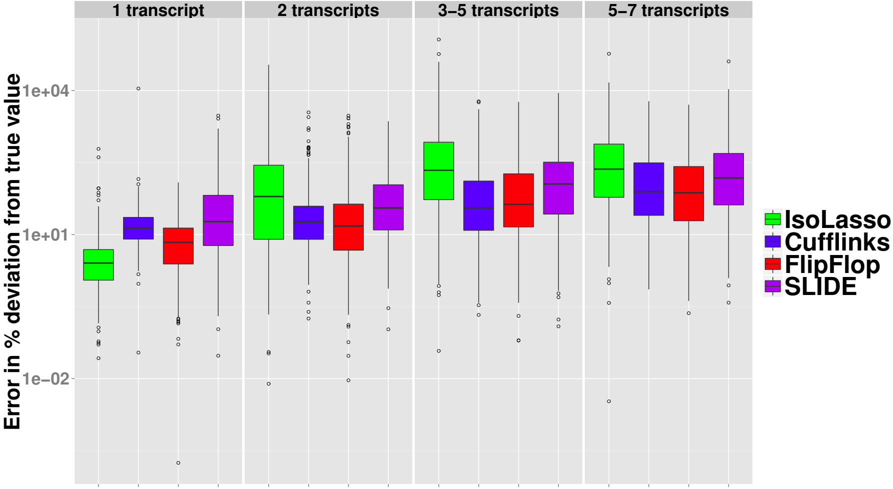
Human Simulation: Abundances

hg19, 1137 genes on chr1, 1million 75 bp single-end reads by transcript levels.



Simulation: Deviation

hg19, 1137 genes on chr1, 1million 75 bp single-end reads by transcript levels.

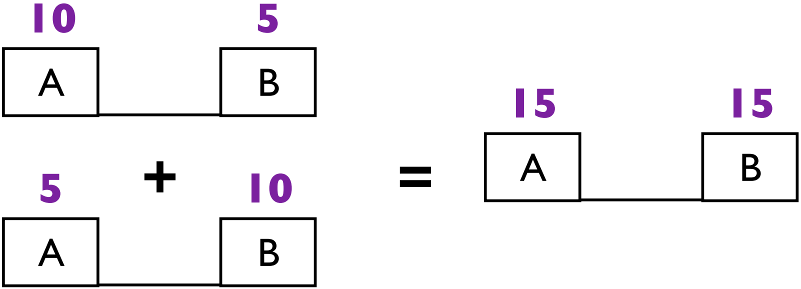


Supplementary Slides

Part 2: multi-sample approach

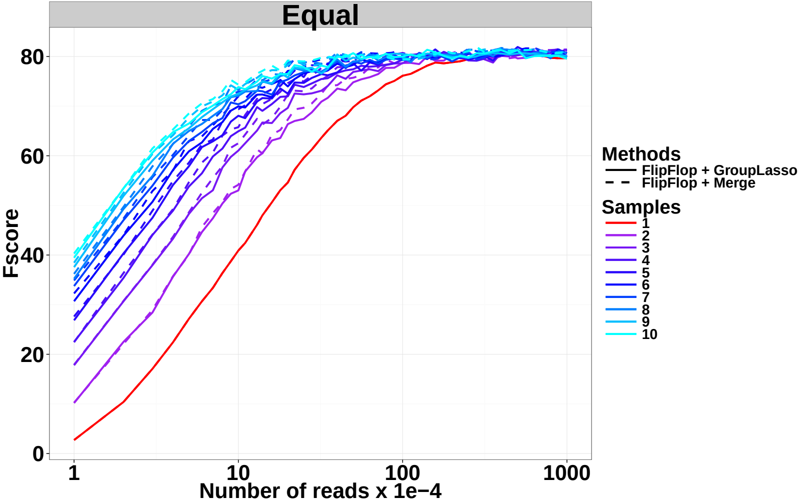
Isoform detection from multiple RNA-seq sample

Why Aggregating can be bad



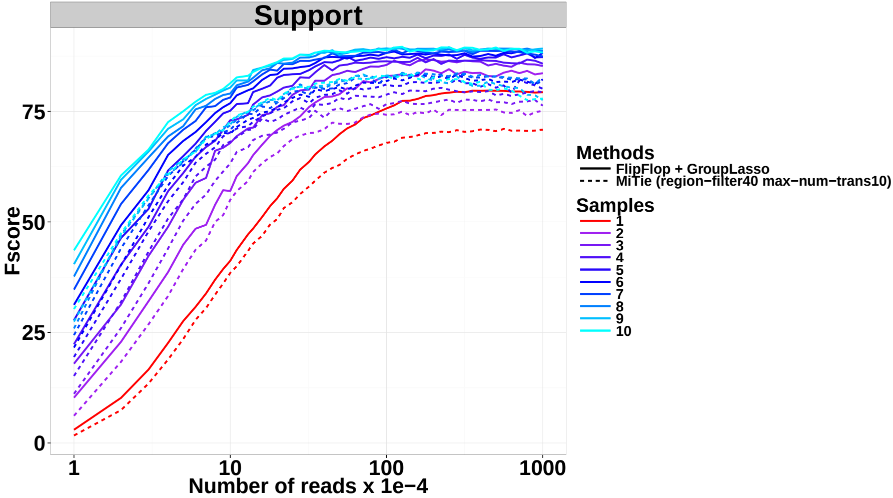
Toy simulation

$$\forall t \in \{1, \ldots, T\}, \theta_t = \theta_o$$



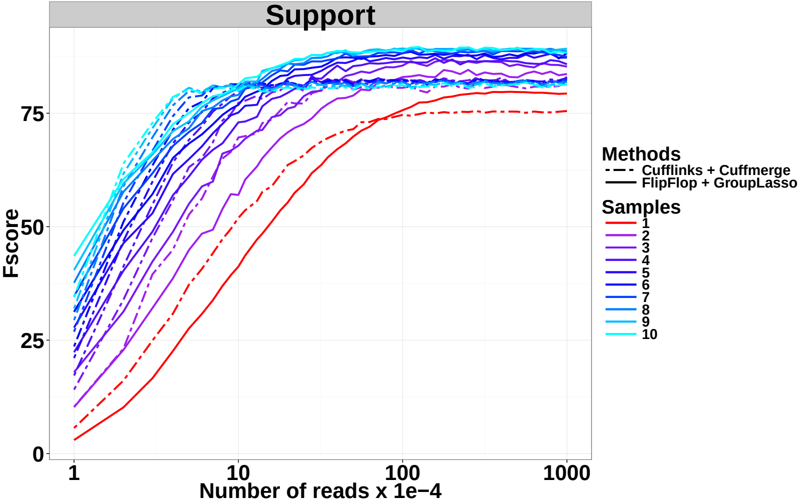
GroupLasso vs State-of-Art 1

$$\forall t \in \{1, \dots, T\}, \text{supp}\theta_t = \text{supp}\theta_o$$

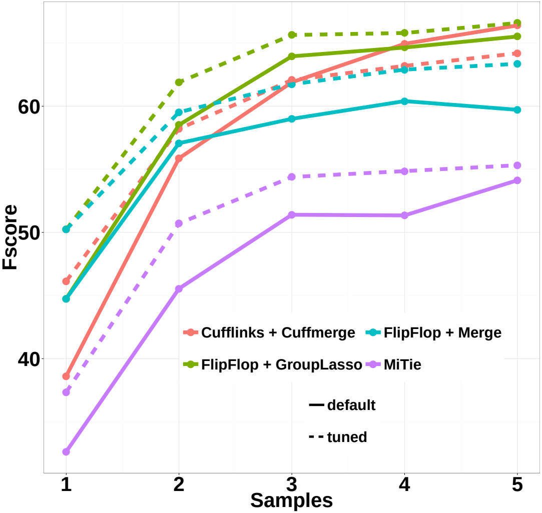


GroupLasso vs State-of-Art 2

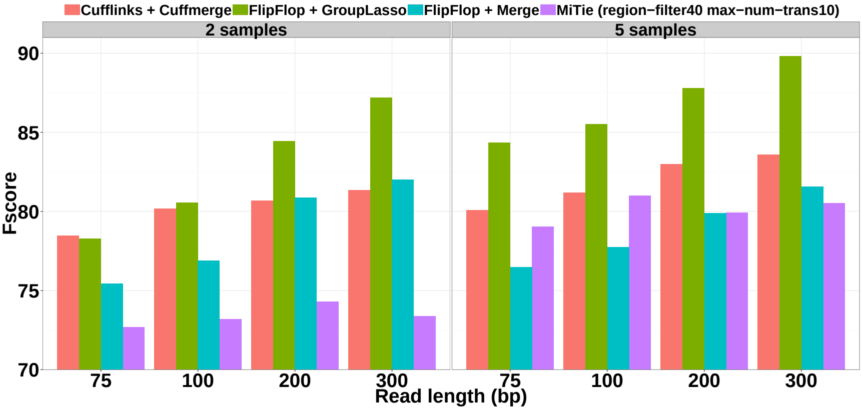
$$\forall t \in \{1, \ldots, T\}, \mathrm{supp}\theta_t = \mathrm{supp}\theta_o$$



Multi-samples simulation
Simulator: FluxSimulator http://sammeth.net/confluence/display/SIM/Home



Simulation: read length


Cufflinks + Cuffmerge
FlipFlop + GroupLasso
FlipFlop + Merge
MiTie (region−filter40 max−num−trans10)
2 samples
5 samples
Fscore
Read length (bp)