A Convex Formulation for Joint RNA Isoform Detection and Quantification from Multiple RNA-Seq Samples

Jean-Philippe Vert

Statistics and Genomics Seminar, UC Berkeley, April 15, 2015

Joint work with

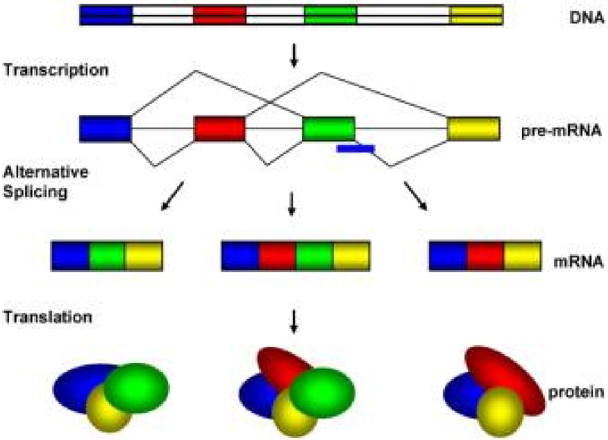
Elsa Bernard

Laurent Jacob

Julien Mairal

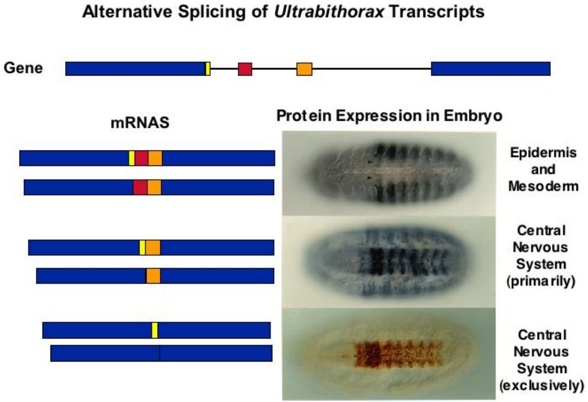
Eric Viara

Alternative splicing: 1 gene = many proteins



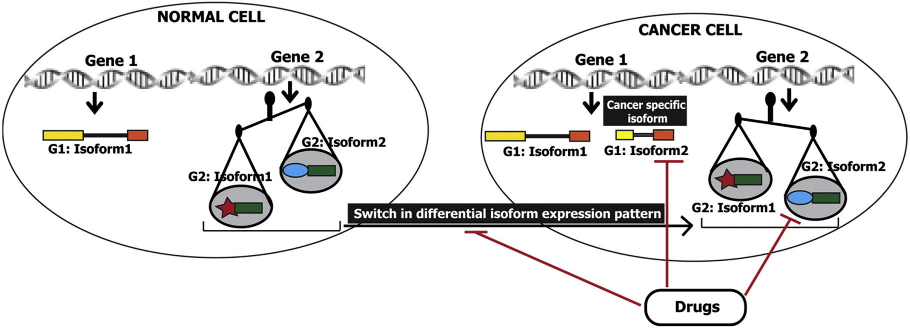
In human, 28k genes give 120k known transcripts (*Pal et al., 2012*)

Alternative splicing matters: developmental regulation in Drosophila


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

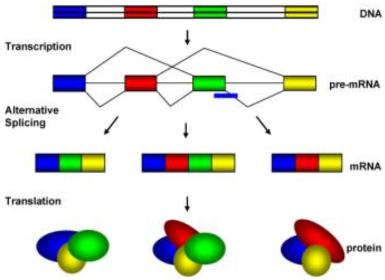
http://orchid.bio.cmu.edu/research.html

Alternative splicing matters: drug targets



(*Pal et al., 2012*)

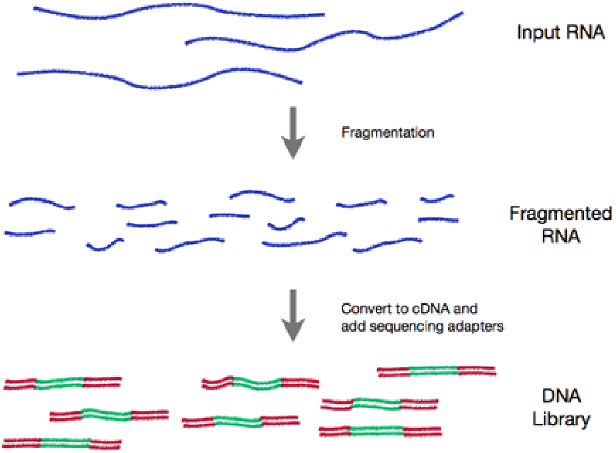
The isoform identification and quantification problem



Given one or several biological samples (e.g., cancer tissues), can we:

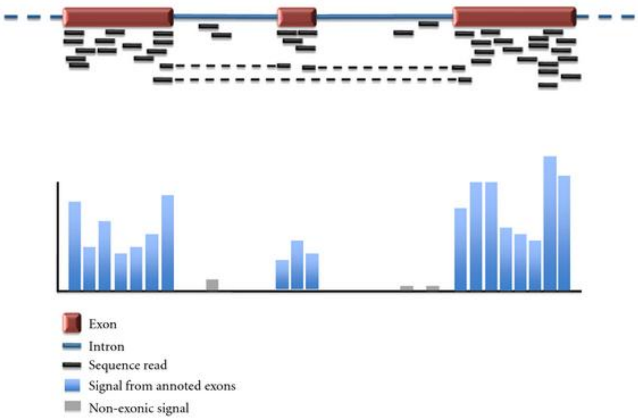
1. identify the isoform(s) of each gene present in the samples?
2. quantify their abundances?

RNA-seq measures mRNA abundance by sequencing short fragments



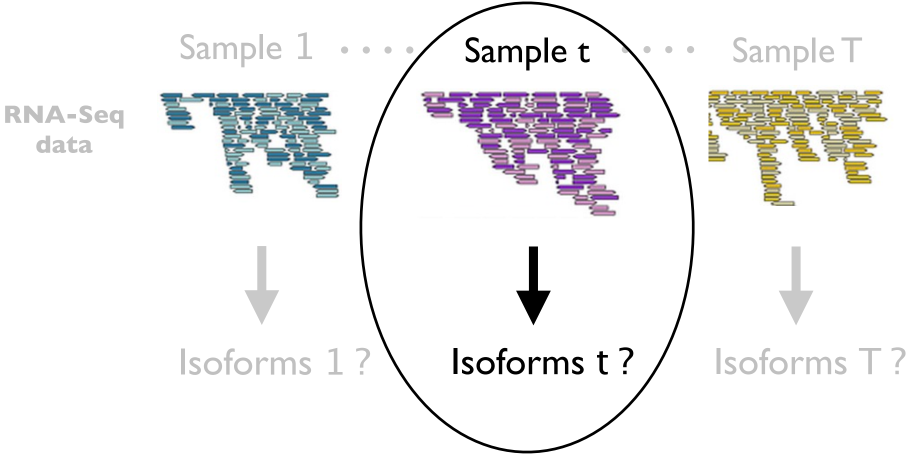
http://rnaseq.uoregon.edu

RNA-seq and alternative splicing



(Costa et al., 2011)

The one-sample case



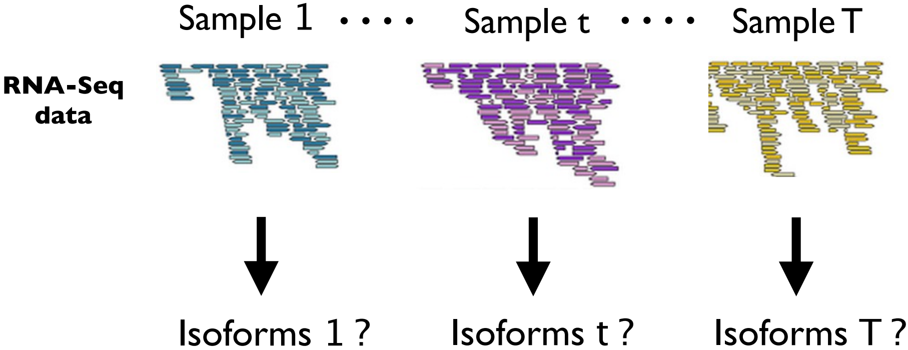
Sample 1 . . . . Sample t . . . . Sample T

RNA-Seq data

Isoforms 1 ? Isoforms t ? Isoforms T ?

Can we perform accurate de novo isoform reconstruction for one given sample?

The multi-sample case



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

RNA-Seq data

Isoforms 1 ? Isoforms t ? Isoforms T ?

Can we improve isoform detection by using several samples simultaneously?

Outline

1. The one-sample case

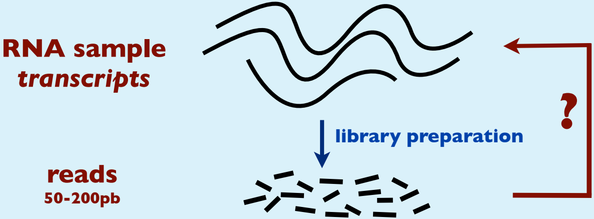
2. The multi-sample case

Outline

1. The one-sample case

2. The multi-sample case

From RNA-Seq reads to isoforms



Transcripts Quantification using annotations

- RQuant (Bohnert et al. 2009)
- FluxCapacitor (Montgomery et al. 2010)
- IsoEM (Nicolae et al. 2011)
- BitSeq (Glaus et al., 2012)
- eXpress (Roberts et al. 2013)

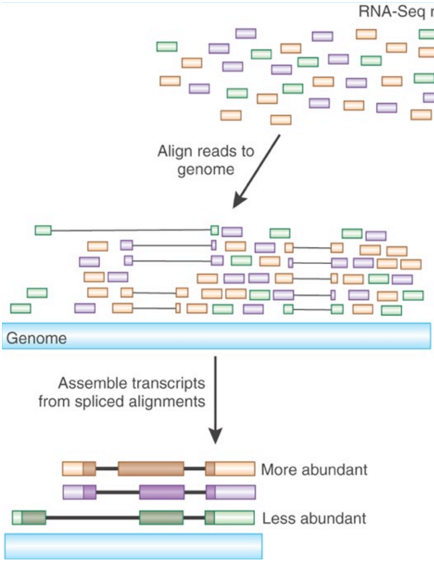
De Novo approaches

- Trinity (Grabherr et al. 2011)
- OASES (Schultz et al. 2012)
- Kissplice (Sacomoto et al. 2012)

Genome-based Transcripts Reconstruction

- Scripture (Guttman et al. 2010)
- Cufflinks (Trapnell et al. 2010)
- IsoLasso (Li et al. 2011a)
- NSMAP (Xia et al. 2011)
- SLIDE (Li et al. 2011b)
- iReckon (Mezlini et al. 2012)
- MiTie (Behr et al. 2013)
- **FlipFlop**

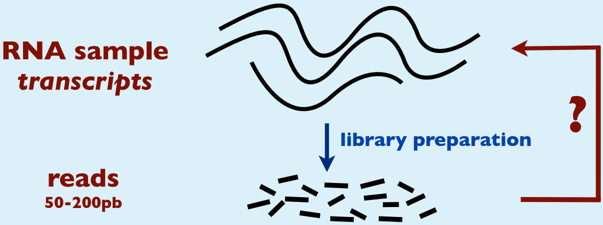
Genome-based isoform reconstruction



Input:
spliced alignment of reads against reference genome

Job:
reconstruct transcripts multi-assembly problem

Contributions?



Genome-based Transcripts Reconstruction

- Scripture (Guttman et al. 2010)
- Cufflinks (Trapnell et al. 2010)
- IsoLasso (Li et al. 2011a)
- NSMAP (Xia et al. 2011)
- SLIDE (Li et al. 2011b)
- iReckon (Mezlini et al. 2012)
- MiTie (Behr et al. 2013)
- **FlipFlop**

What is new ?

Contributions

- **NO NEED for FILTERING of candidate isoforms**
- **FASTER than existing methods that solve the same problem**

}

flow method

- **adapted to LONG READS**
- **R package**

Contributions

- NO NEED for FILTERING of candidate isoforms

- FASTER than existing methods that solve the same problem

- **adapted to LONG READS**

- R package

particular splicing graph

Contributions

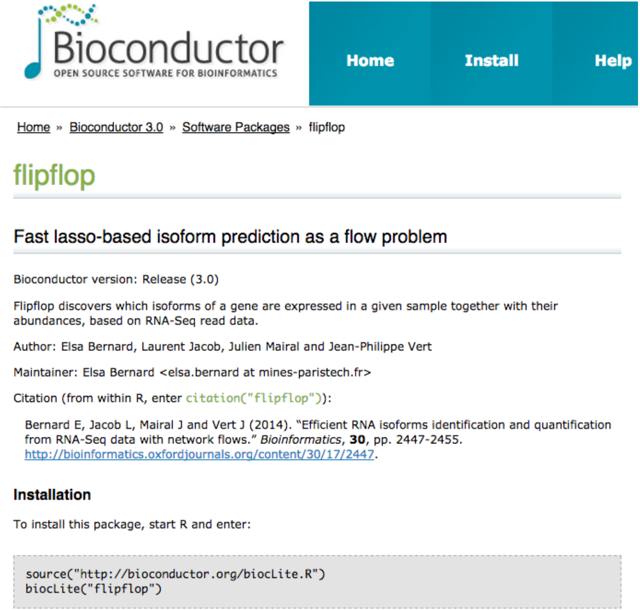
- NO NEED for FILTERING of candidate isoforms

- FASTER than existing methods that solve the same problem

- adapted to long reads

- **R package**

Contributions



Home » Bioconductor 3.0 » Software Packages » flipflop

flipflop

Fast lasso-based isoform prediction as a flow problem

Bioconductor version: Release (3.0)

Flipflop discovers which isoforms of a gene are expressed in a given sample together with their abundances, based on RNA-Seq read data.

Author: Elsa Bernard, Laurent Jacob, Julien Mairal and Jean-Philippe Vert

Maintainer: Elsa Bernard <elsa.bernard at mines-paristech.fr>

Citation (from within R, enter `citation("flipflop")`):

> Bernard E, Jacob L, Mairal J and Vert J (2014). "Efficient RNA isoforms identification and quantification from RNA-Seq data with network flows." *Bioinformatics*, **30**, pp. 2447-2455. http://bioinformatics.oxfordjournals.org/content/30/17/2447.

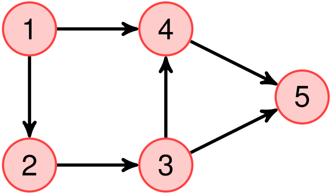
Installation

To install this package, start R and enter:

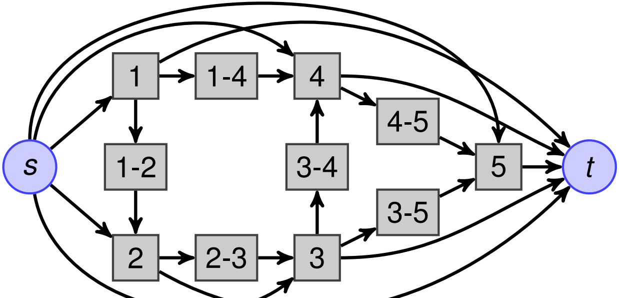
```
source("http://bioconductor.org/biocLite.R")
biocLite("flipflop")
```

Isoforms are Paths in a Graph

- Splicing graph for a gene with 5 exons:

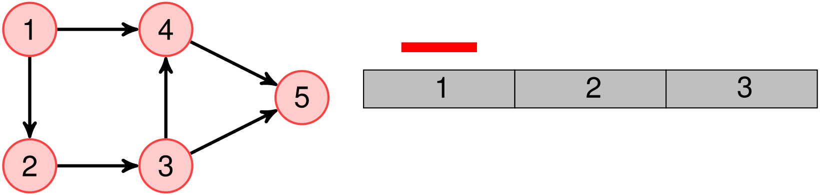


- FlipFlop graph: **1 type of read ↔ 1 node**

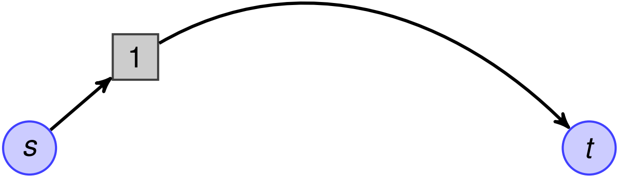


Graph adapted to long reads

- Splicing graph for a gene with 5 exons:

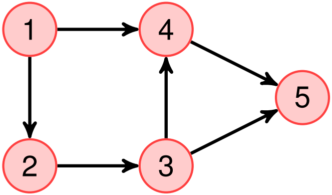


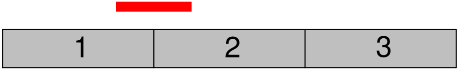
- FlipFlop graph:



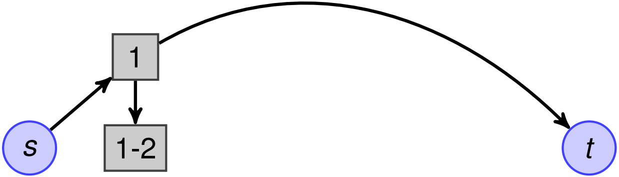
Graph adapted to long reads

- Splicing graph for a gene with 5 exons:



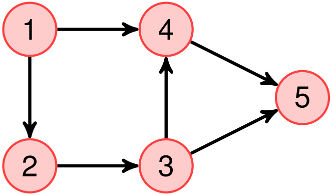


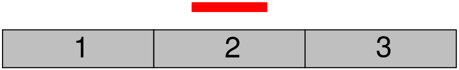
- FlipFlop graph:



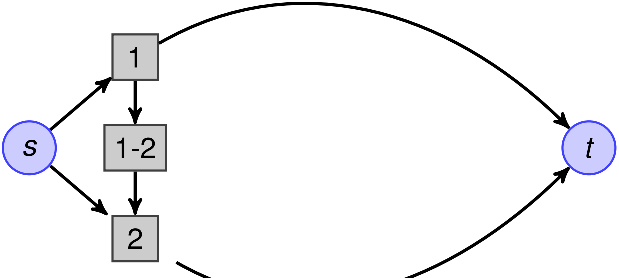
Graph adapted to long reads

- Splicing graph for a gene with 5 exons:



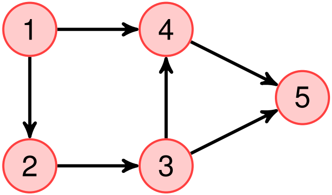


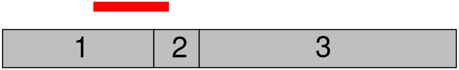
- FlipFlop graph:



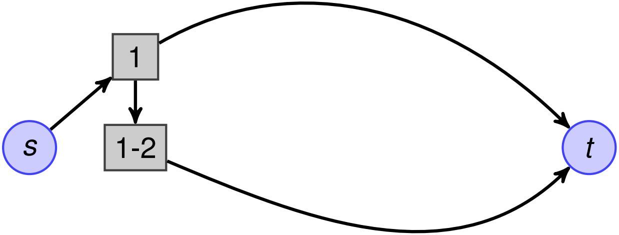
Graph adapted to long reads

- Splicing graph for a gene with 5 exons:



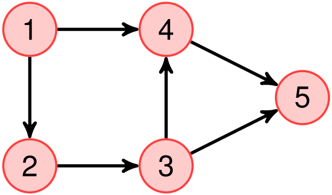


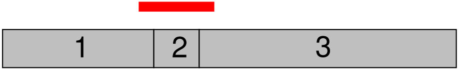
- FlipFlop graph:



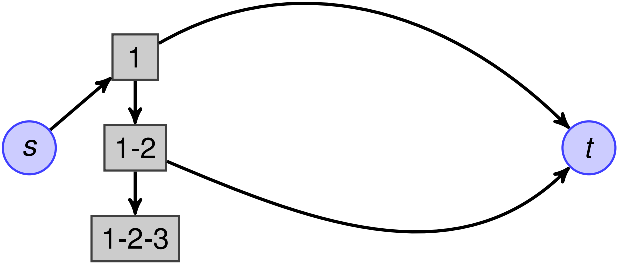
Graph adapted to long reads

- Splicing graph for a gene with 5 exons:



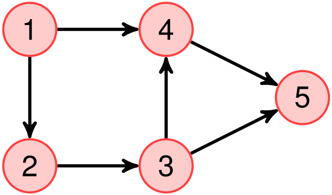


- FlipFlop graph:

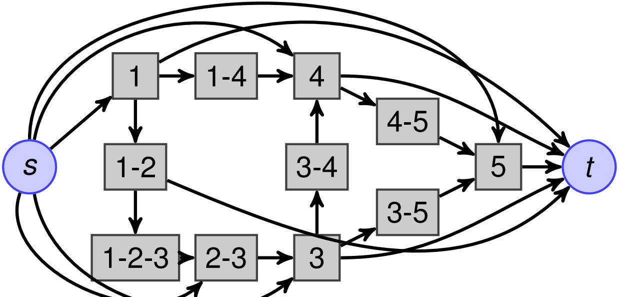


Graph adapted to long reads

- Splicing graph for a gene with 5 exons:

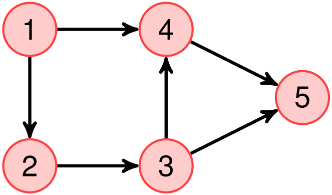


- FlipFlop graph:

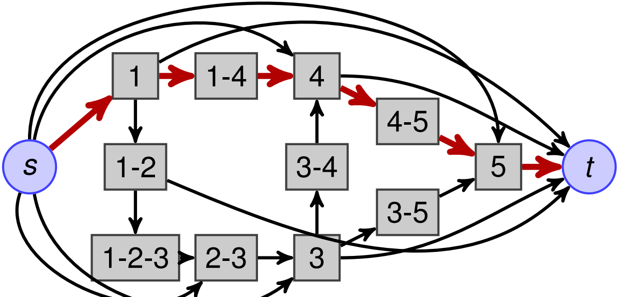


Graph adapted to long reads

- Splicing graph for a gene with 5 exons:

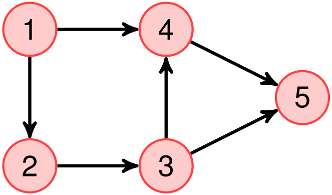


- FlipFlop graph: one path with abundance θ_1

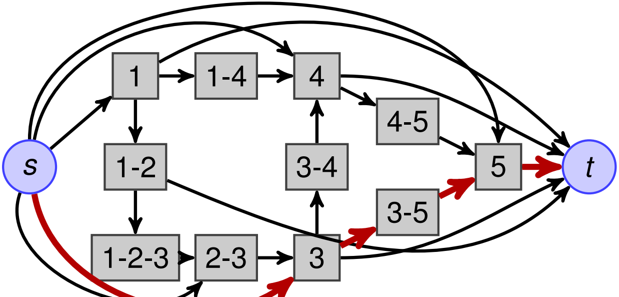


Graph adapted to long reads

- 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

Minimal path cover

- Cufflinks

Regularization approach

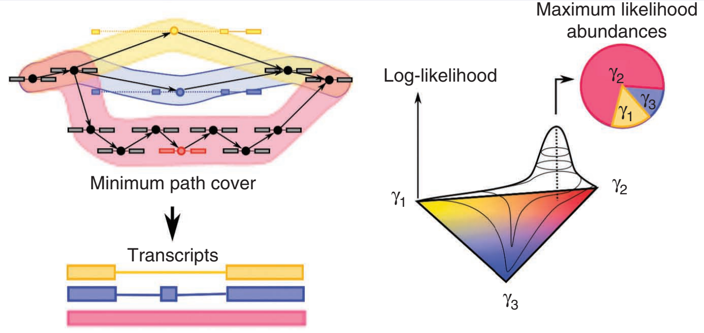
- IsoLasso, NSMAP, SLIDE, iReckon, MiTie, FlipFlop

Select a small number of paths?

Cufflinks strategy

A two-step approach

1. find a set of *minimal paths* to explain read positions (independent from read counts)
2. estimate isoform abundances using read counts



Select a small number of paths?

Regularization approach

1. Suppose there are *c* candidate isoforms (c large)
2. Let θ the unknown c-dimensional vector of abundance
3. Let $\mathcal{L}(\phi)$ quantify whether θ explains the observed read counts
   - e.g., Poisson negative log-likelihood:

$$\mathcal{L}(\theta) = \sum_{\text{node } u} -\log p(X_u) \text{ with } X_u \sim \mathcal{P}(\delta_u) \text{ and } \delta_u \propto l_u \sum_{\text{path } p \ni u} \theta_p$$

4. Regularization-based approaches try to solve:

$$\min_{\theta \in \mathbb{R}_+^c} \mathcal{L}(\theta) \text{ such that } \theta \text{ is sparse}$$

Select a small number of paths?

Regularization approach

1. Suppose there are *c* candidate isoforms (c large)
2. Let θ the unknown c-dimensional vector of abundance
3. Let $\mathcal{L}(\phi)$ quantify whether θ explains the observed read counts
   - e.g., Poisson negative log-likelihood:

$$\mathcal{L}(\theta) = \sum_{\text{node } u} -\log p(X_u) \;\; \text{with} \;\; X_u \sim \mathcal{P}(\delta_u) \;\; \text{and} \;\; \delta_u \propto l_u \sum_{\text{path } p \ni u} \theta_p$$

4. Regularization-based approaches try to solve:

$$\min_{\theta \in \mathbb{R}_+^c} \mathcal{L}(\theta) \text{ such that } \theta \text{ is sparse}$$

Select a small number of paths?

Regularization approach

1. Suppose there are *c* candidate isoforms (c large)
2. Let θ the unknown c-dimensional vector of abundance
3. Let $\mathcal{L}(\phi)$ quantify whether θ explains the observed read counts
   - e.g., Poisson negative log-likelihood:

$$\mathcal{L}(\theta) = \sum_{\text{node } u} - \log p(X_u) \quad \text{with} \quad X_u \sim \mathcal{P}(\delta_u) \quad \text{and} \quad \delta_u \propto l_u \sum_{\text{path } p \ni u} \theta_p$$

4. Regularization-based approaches try to solve:

$$\min_{\theta \in \mathbb{R}_+^c} \mathcal{L}(\theta) \text{ such that } \theta \text{ is sparse}$$

Isoform Deconvolution with the Lasso

Lasso

Estimate θ sparse by solving:

$$\min_{\theta \in \mathbb{R}_+^c} \mathcal{L}(\theta) + \lambda \|\theta\|_1 \,,$$

with $\mathcal{L}$ a convex loss function.

Computationally challenging:

→ IsoLasso: strong filtering

→ NSMAP, SLIDE: number of exons cut-off

FlipFlop: Fast Lasso-based Isoform Prediction as a FLOw Problem

→ no filtering

→ no exon restrictions

Fast isoform deconvolution with the Lasso (FlipFlop)

Theorem (Bernard, Mairal, Jacob and V., 2014)

The isoform deconvolution problem

$$\min_{\theta \in \mathbb{R}_+^c} \mathcal{L}(\theta) + \lambda \| \theta \|_1$$

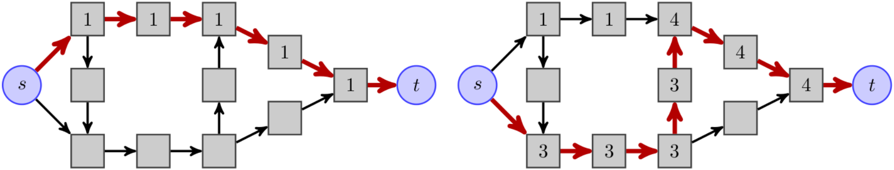
can be solved in polynomial time in the number of exon.

Key ideas

1. Reformulation as a convex cost flow problem (Mairal and Yu, 2012)
2. Recover isoforms by flow decomposition algorithm

"Feature selection on an exponential number of features in polynomial time"

Combinations of isoforms are flows



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

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

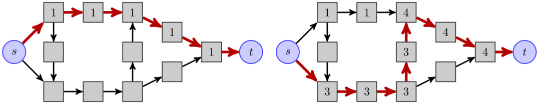
- **Linear combinations of isoforms** ⇒ **Flow value on every edges**
- **Flow value on every edges** ⇒ **Paths with given value/abundance**

Flow Decomposition
(linear time algorithm)

Flux Capacitor. 2008.

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,

$$\min_{\theta\in\mathbb{R}_+^c} \mathcal{L}(\theta) + \lambda\|\theta\|_1 \quad \text{is equivalent to} \quad \min_{f \text{ flow}} \tilde{\mathcal{L}}(f) + \lambda f_t$$

Summary

Isoform Detection=Path Selection Problem

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

$\Updownarrow$

Equivalent Network Flow Problem

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

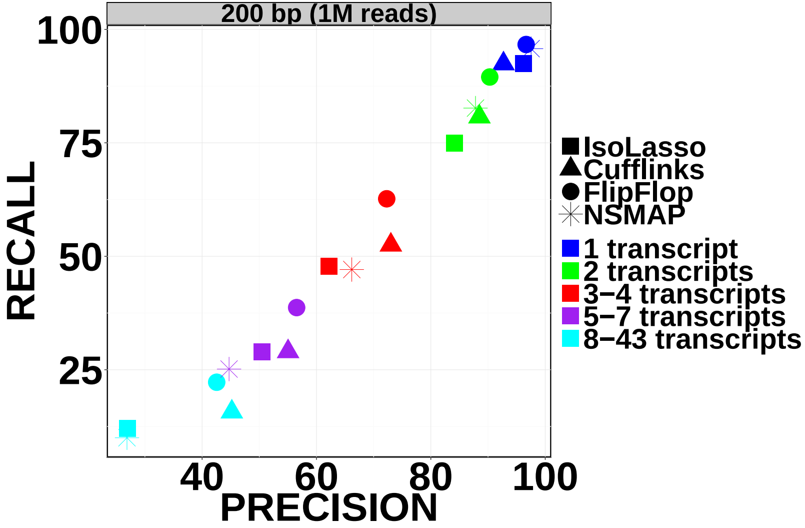
$\downarrow$

Network Flow Algorithms

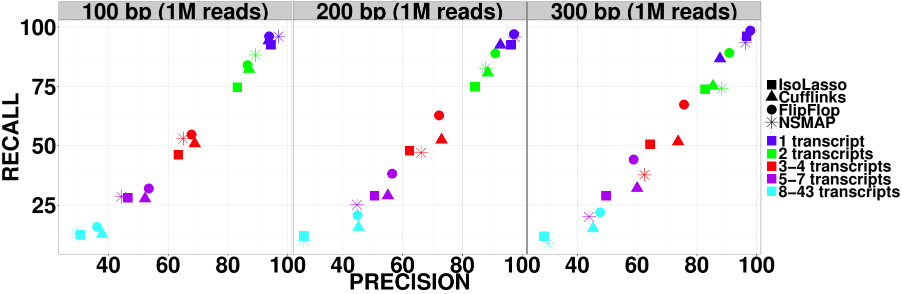
Efficient Algorithms ! Polynomial Time.

Human Simulation: Precision/Recall

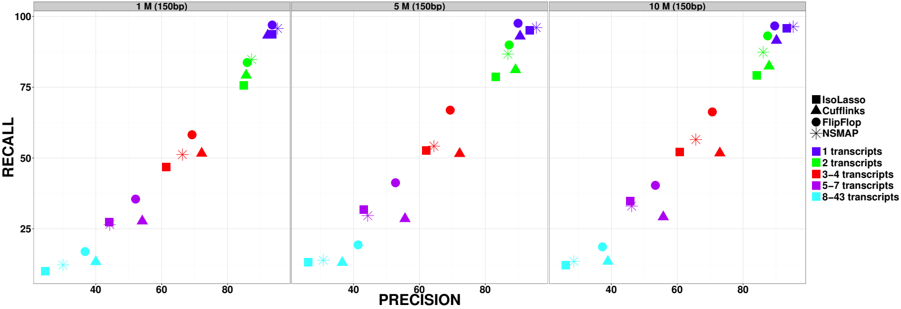
hg19, 1137 genes on chr1, 1million 200 bp single-end reads by transcript levels.
Simulator: `http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html`



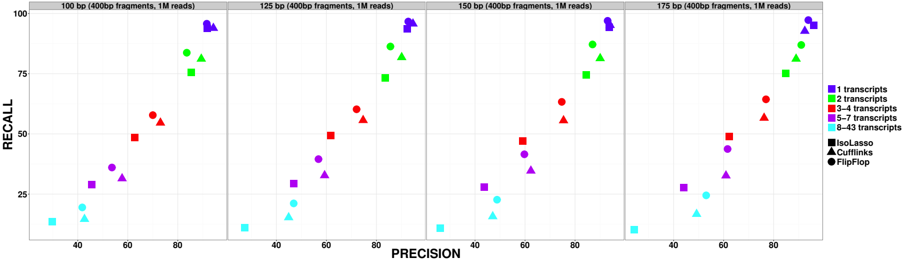
Performance increases with read length



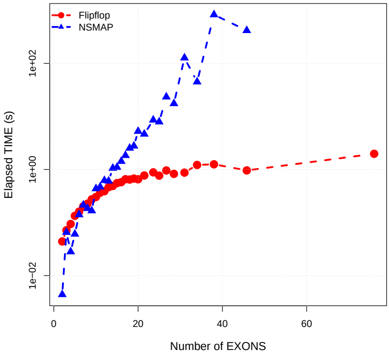
Performance increases with coverage

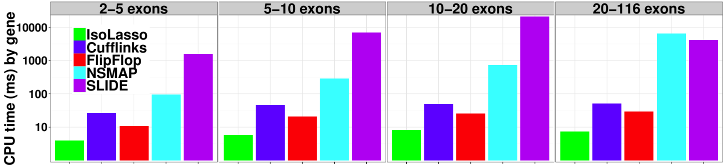


Extension to paired-end reads OK.



Speed trial


Flipflop
NSMAP
Elapsed TIME (s)
1e−02
1e+00
1e+02
0
20
40
60
Number of EXONS


2–5 exons
5–10 exons
10–20 exons
20–116 exons
CPU time (ms) by gene
10000
1000
100
10
IsoLasso
Cufflinks
FlipFlop
NSMAP
SLIDE

One-sample case summary

- FlipFlop: Fast method for exact Lasso-based isoform detection and quantification
- `http://cbio.mines-paristech.fr/flipflop`
- Available as an 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, 30(17):247-55, 2014

Outline

1. The one-sample case

2. The multi-sample case

Strategy for 1 sample

Sample t

mapping & counting

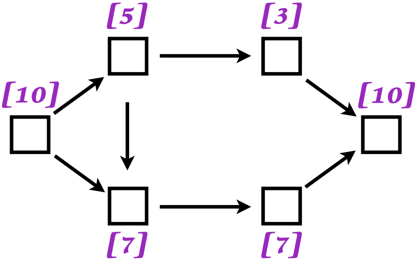
Splicing graph

FlipFlop

fast lasso-based isoform prediction as a flow problem

Isoforms t

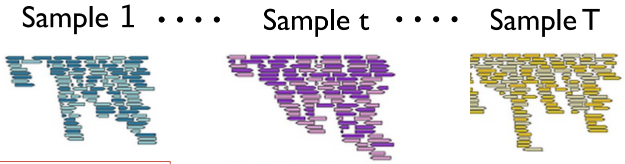
Unidimensional splicing graph



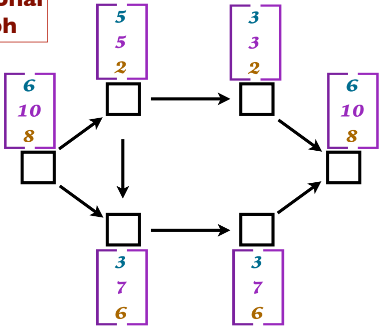
□ : exon or junction

[10]: read counts

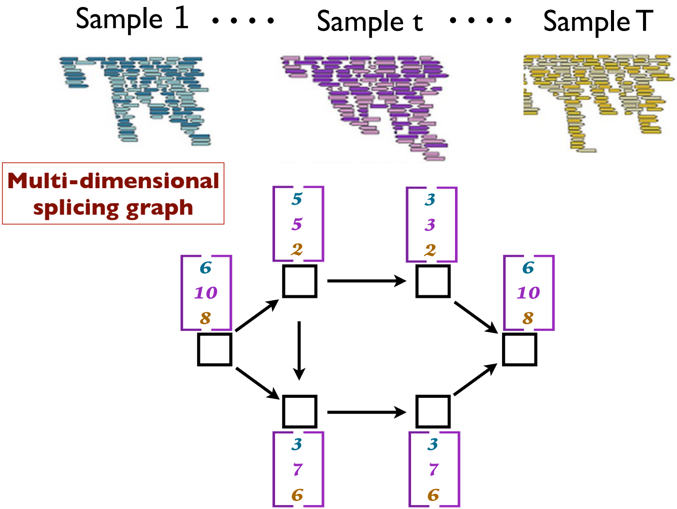
Multi-dimensional case



Multi-dimensional splicing graph



Multi-dimensional case

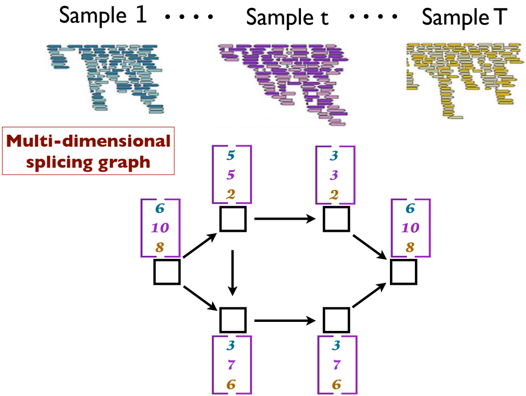


Sample 1 . . . . Sample t . . . . Sample T

Multi-dimensional splicing graph

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

Notations

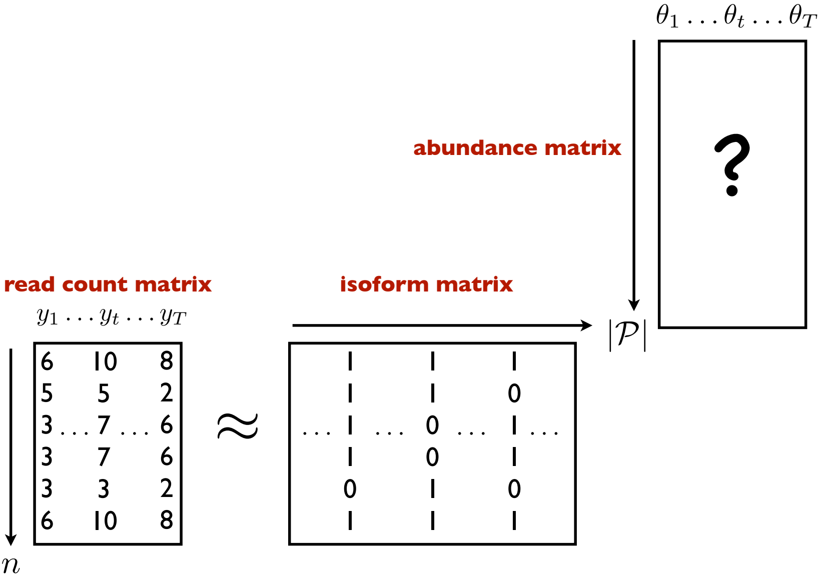

Sample 1 · · · · Sample t · · · · Sample T
Multi-dimensional splicing graph
6
10
8
5
5
2
3
3
2
6
10
8
3
7
6
3
7
6

- n nodes, T samples
- $\mathcal{P}$ paths in the splicing graph
- $y_t \in \mathbb{R}_+^n$ vector of counts for sample t

  $y_1 \dots y_t \dots y_T$
- $\theta_t \in \mathbb{R}_+^{|\mathcal{P}|}$ vector of isoform abundances for sample t

  $\theta_1 \dots \theta_t \dots \theta_T$

Group-Lasso strategy



abundance matrix

$$\begin{array}{c} \theta_1 \ldots \theta_t \ldots \theta_T \\ \left[\begin{array}{c} \\ ? \\ \\ \end{array}\right] \end{array}$$

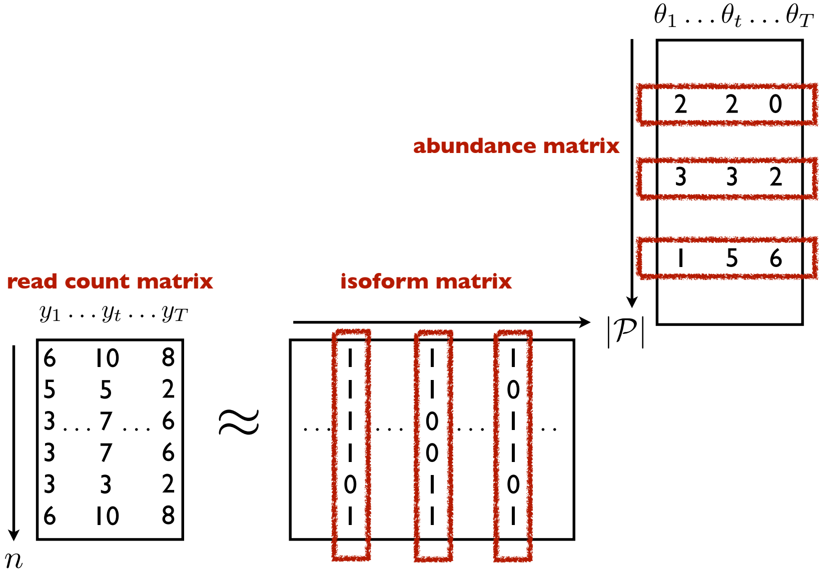
$|\mathcal{P}|$

read count matrix

isoform matrix

$$\begin{array}{c} y_1 \ldots y_t \ldots y_T \\ \left[\begin{array}{ccccc} 6 & & 10 & & 8 \\ 5 & & 5 & & 2 \\ 3 & \ldots & 7 & \ldots & 6 \\ 3 & & 7 & & 6 \\ 3 & & 3 & & 2 \\ 6 & & 10 & & 8 \end{array}\right] \\ n \end{array} \approx \left[\begin{array}{ccccccc} & 1 & & 1 & & 1 & \\ & 1 & & 1 & & 0 & \\ \ldots & 1 & \ldots & 0 & \ldots & 1 & \ldots \\ & 1 & & 0 & & 1 & \\ & 0 & & 1 & & 0 & \\ & 1 & & 1 & & 1 & \end{array}\right]$$

Group-Lasso strategy



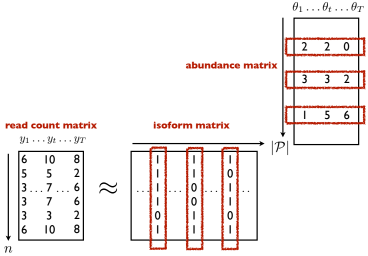
read count matrix

isoform matrix

abundance matrix

$$\begin{array}{c} y_1 \dots y_t \dots y_T \\ \begin{bmatrix} 6 & 10 & 8 \\ 5 & 5 & 2 \\ 3 \dots & 7 \dots & 6 \\ 3 & 7 & 6 \\ 3 & 3 & 2 \\ 6 & 10 & 8 \end{bmatrix} \end{array} \; n \quad \approx \quad \begin{bmatrix} & 1 & & 1 & & 1 & \\ & 1 & & 1 & & 0 & \\ \dots & 1 & \dots & 0 & \dots & 1 & \dots \\ & 1 & & 0 & & 1 & \\ & 0 & & 1 & & 0 & \\ & 1 & & 1 & & 1 & \end{bmatrix} \; |\mathcal{P}| \quad \begin{array}{c} \theta_1 \dots \theta_t \dots \theta_T \\ \begin{bmatrix} \\ 2 & 2 & 0 \\ \\ 3 & 3 & 2 \\ \\ 1 & 5 & 6 \\ \\ \end{bmatrix} \end{array}$$

More formally



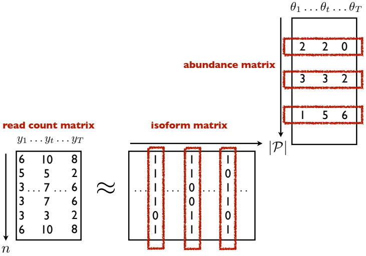
- each isoform defines a **group** $\boldsymbol{\theta}_p = \{\theta_p^t, t \in [\![1, T]\!]\}$
- the multi-samples loss is the sum of the independent losses

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

- Ideally we want to solve the NP-hard L0 problem

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

More formally



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

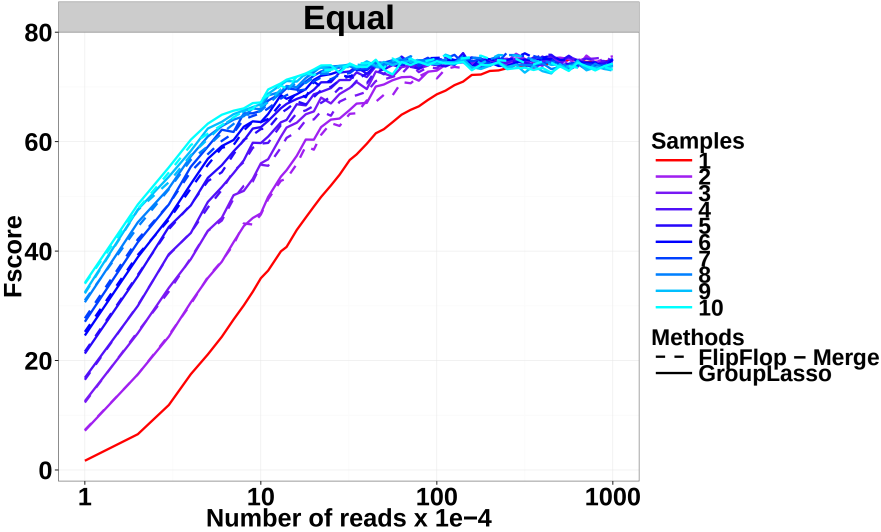
$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \mathrm{loss}(y_t, \theta_t)$$

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

$$\min_{\{\boldsymbol{\theta}_p\}_{p \in 1,\dots,|\mathcal{P}|}} \mathcal{L}(\boldsymbol{\theta}) + \lambda \sum_{p \in \mathcal{P}} \|\boldsymbol{\theta}_p\|_2$$

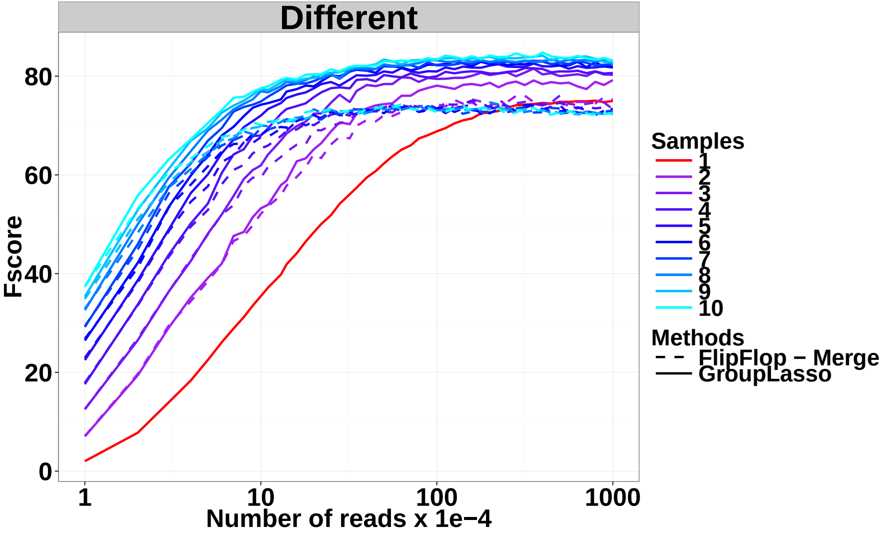
Toy simulation

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



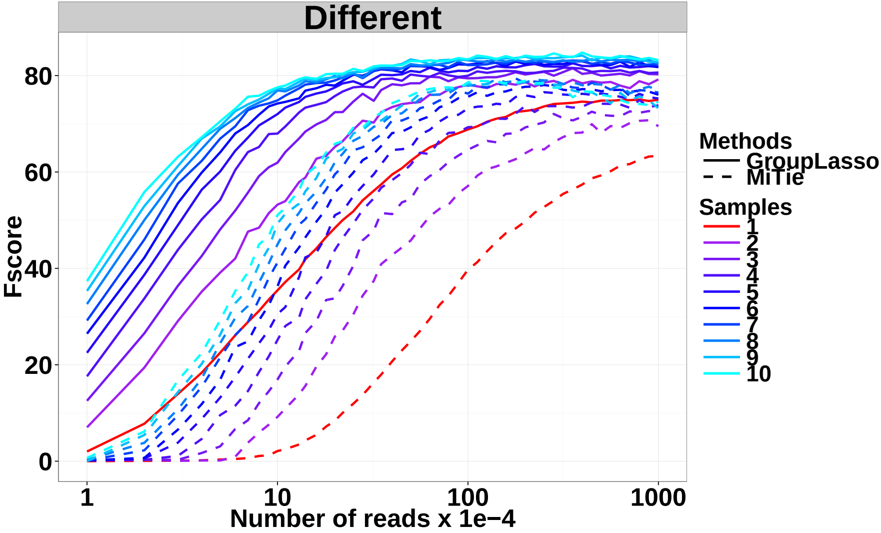
More realistic simulation

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

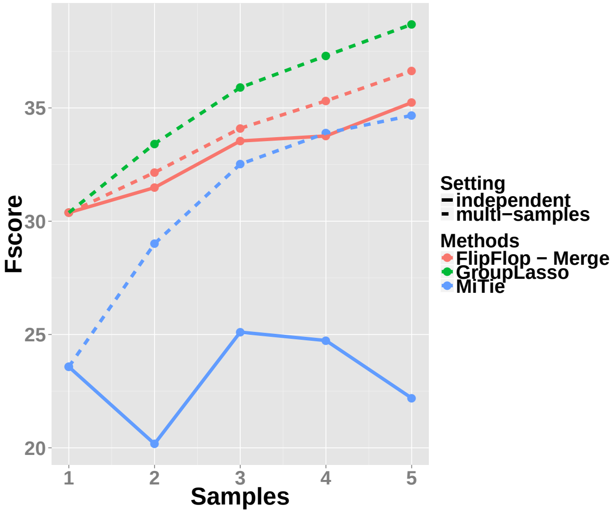


GroupLasso vs State-of-Art

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



modENCODE data
Time course development of D.melanogaster



Multi-sample case summary

- Extension of FlipFlop to multiple samples (with group Lasso formulation)
- No more flow trick
- `http://cbio.mines-paristech.fr/flipflop`
- Available as an R package

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

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. Technical report HAL-01123141, March 2015.

Thanks

erc

SEVENTH FRAMEWORK PROGRAMME

INVESTISSEMENTS D'AVENIR

AGENCE NATIONALE DE LA RECHERCHE ANR

BIOMÉRIEUX

MINES ParisTech

institutCurie
Together, let's beat cancer.

Inserm
Institut national
de la santé et de la recherche médicale

PSL RESEARCH UNIVERSITY