A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples

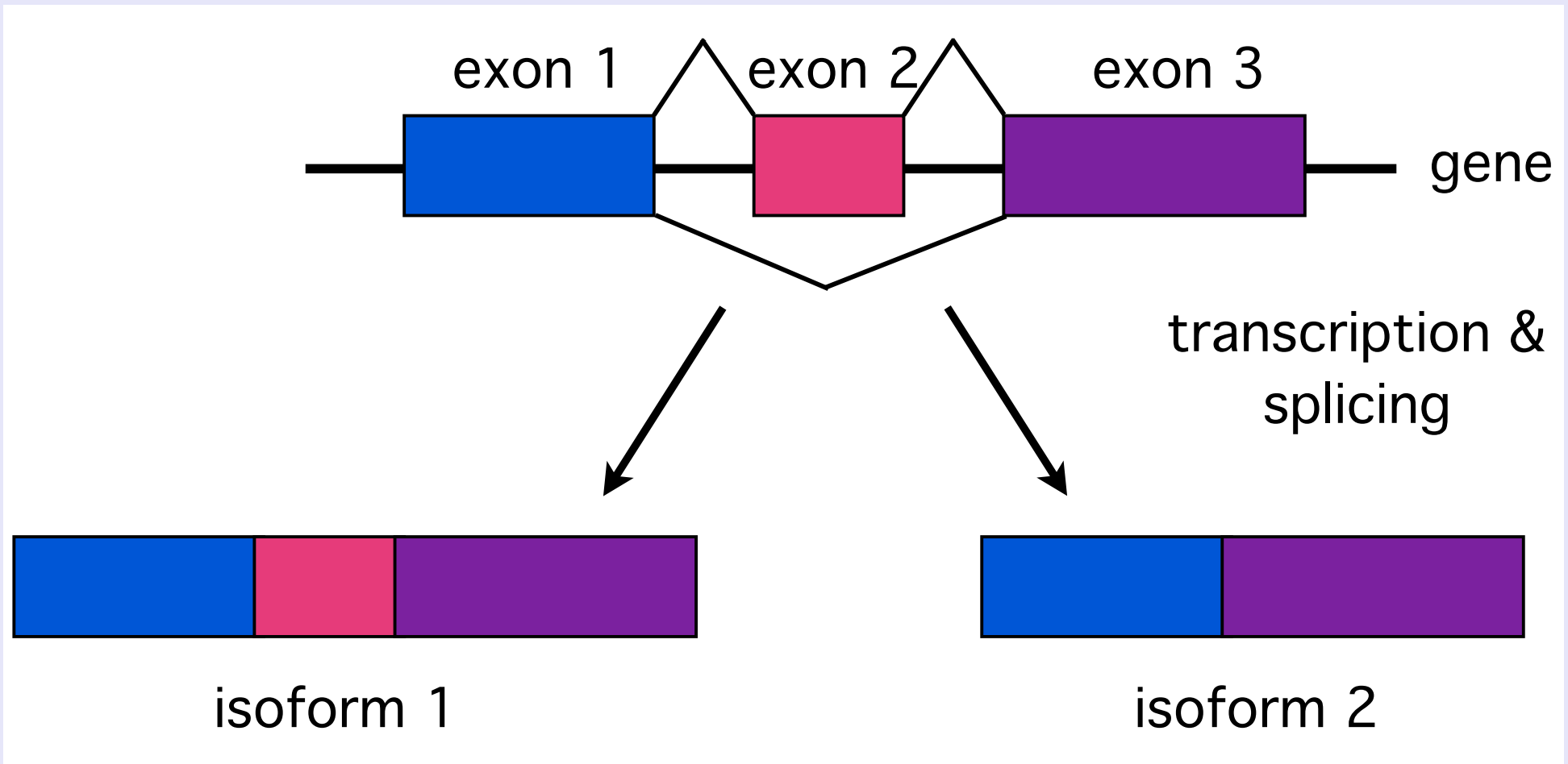
Elsa Bernard[1,2,3], Laurent Jacob[4], Julien Mairal[5], Eric Viara[6], Jean-Philippe Vert[1,2,3]

[1]: Center For Computational Biology, Mines ParisTech, Fontainbleau, France;
[2]: INSERM U900, Paris, France;
[3]: Institut Curie, Paris, France;
[4]: CNRS - LBBE Laboratory, Lyon, France;
[5]: LEAR Project-Team, INRIA Grenoble - Rhône Alpes, France
[6]: Sysra, Yerres, France

ABSTRACT: We propose a new method for solving the isoform deconvolution problem jointly across several samples, by penalizing a convex objective function with a group-lasso penalty. We show that the method outperforms simple pooling strategies and other methods based on mixed integer programming.

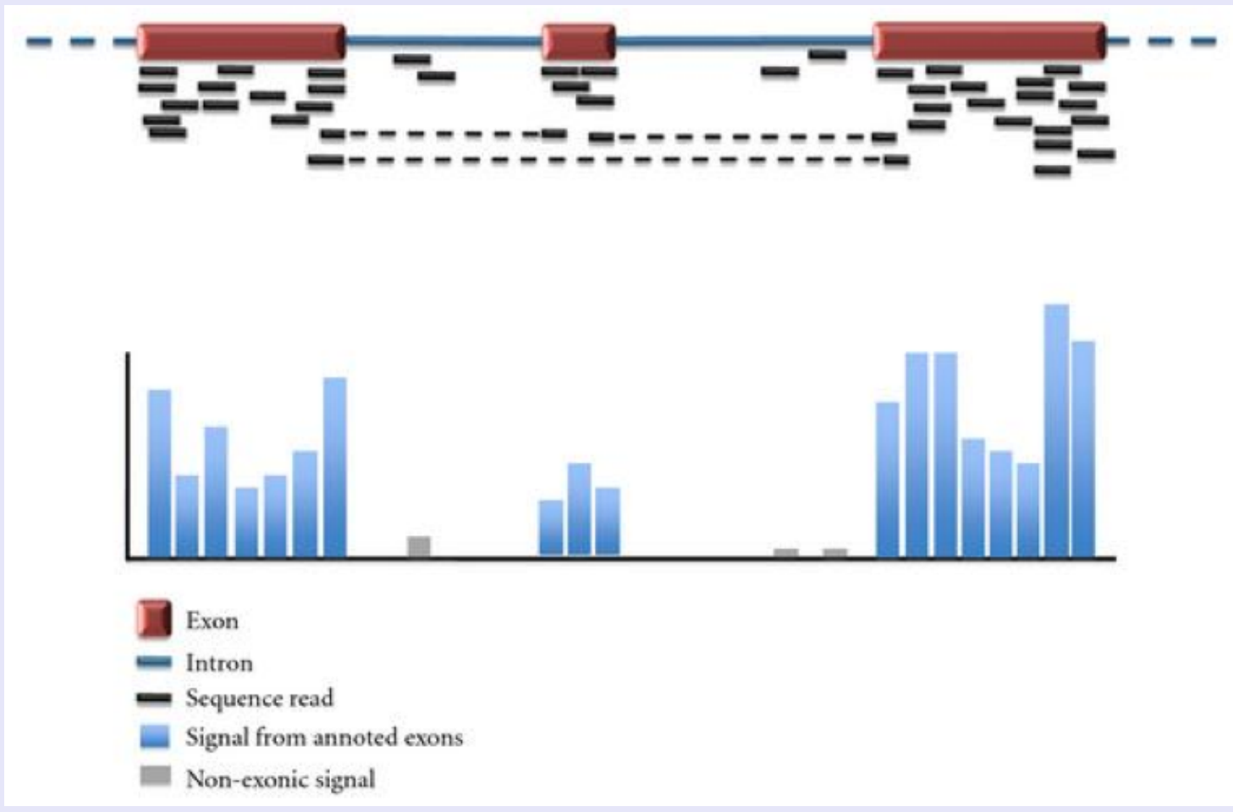
Background

Alternative Splicing



- During transcription of eukaryotic genes, exons and introns are alternatively spliced, producing different isoforms.

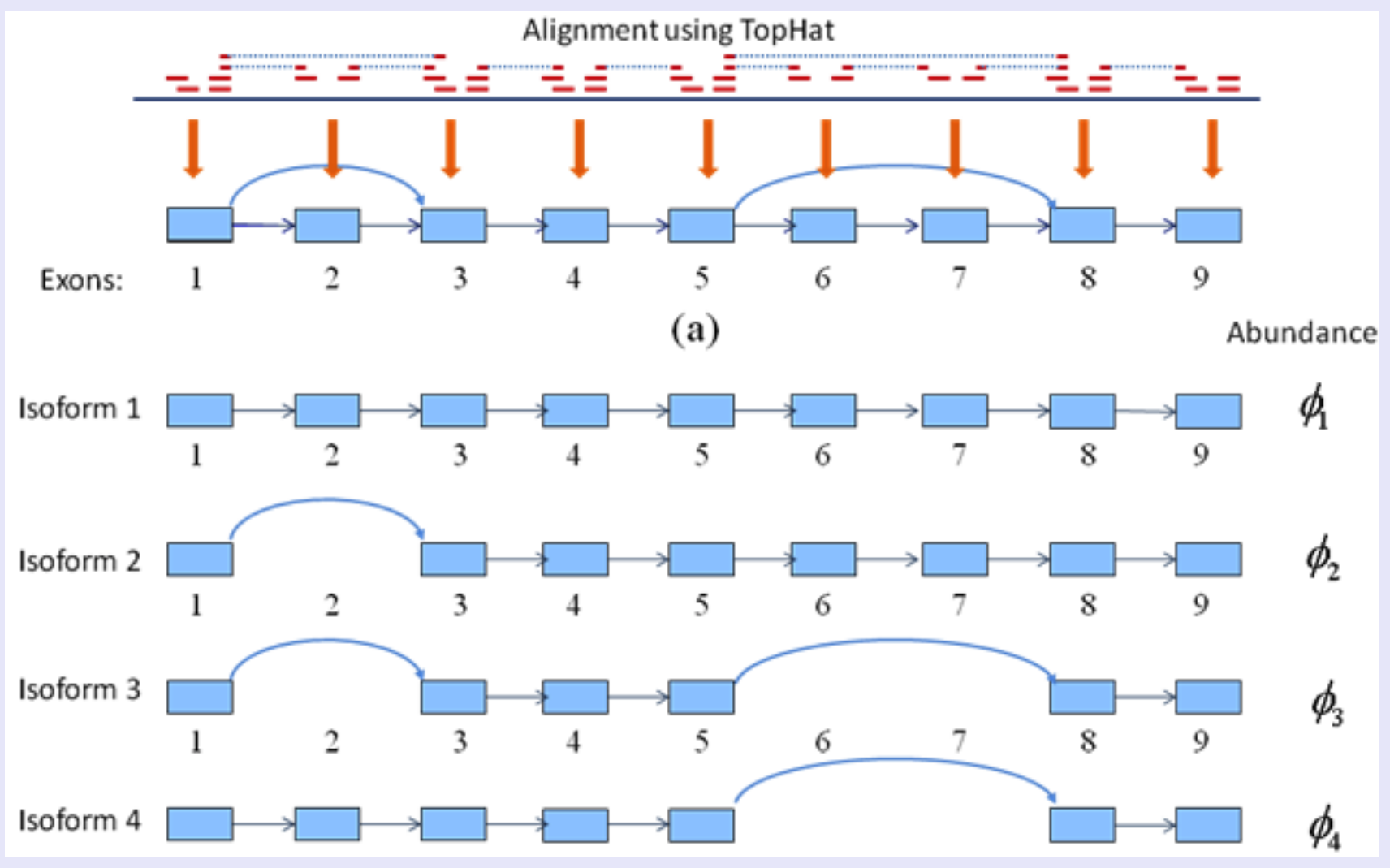
RNA-seq data



Costa et al., 2011

- RNA-seq measures abundance of each exon and exon-exon junction of a gene.

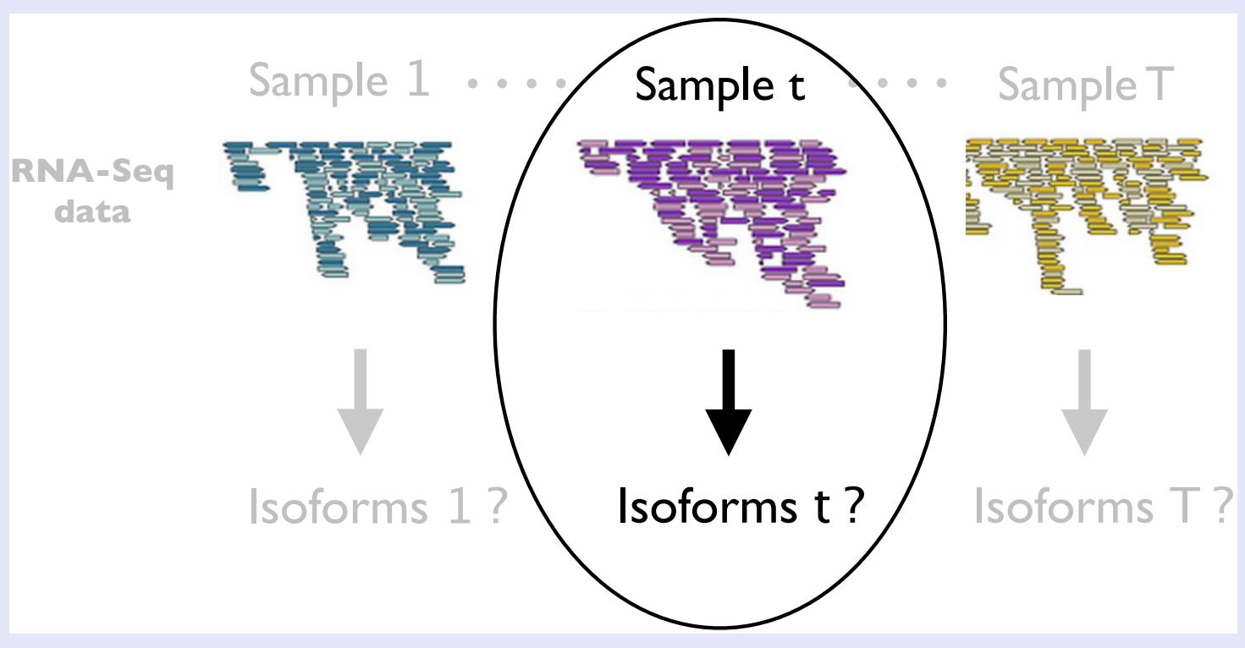
Isoform Deconvolution



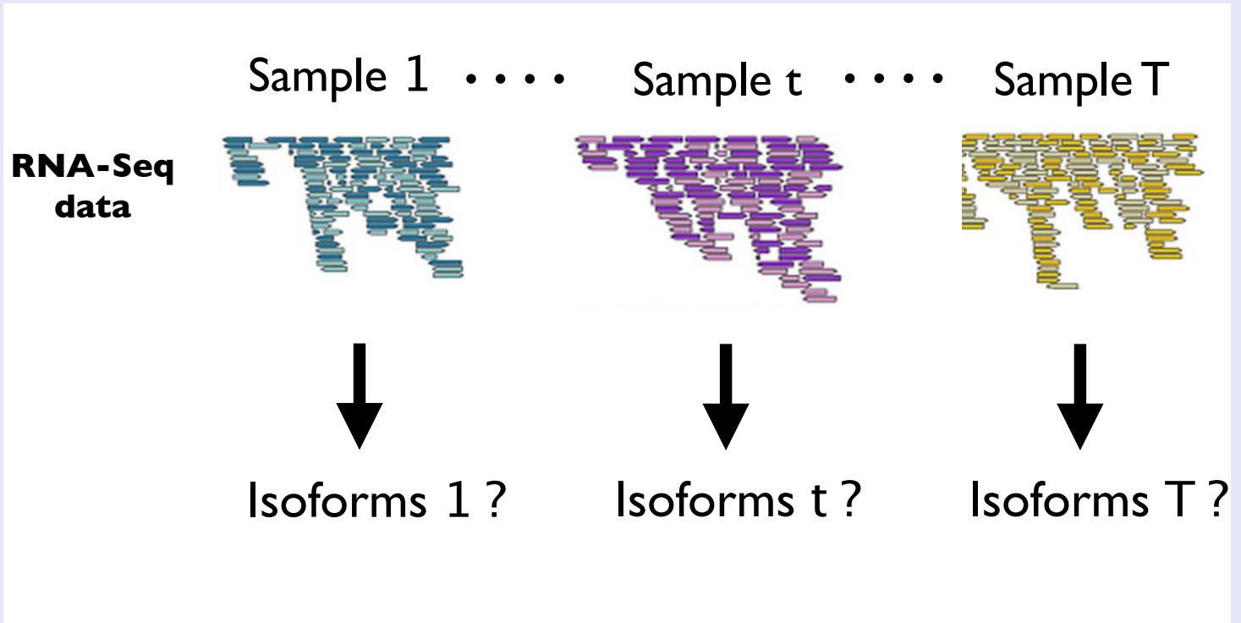
Xia et al., 2011

- Isoforms are paths in a directed acyclic graph (splicing graph).

Questions



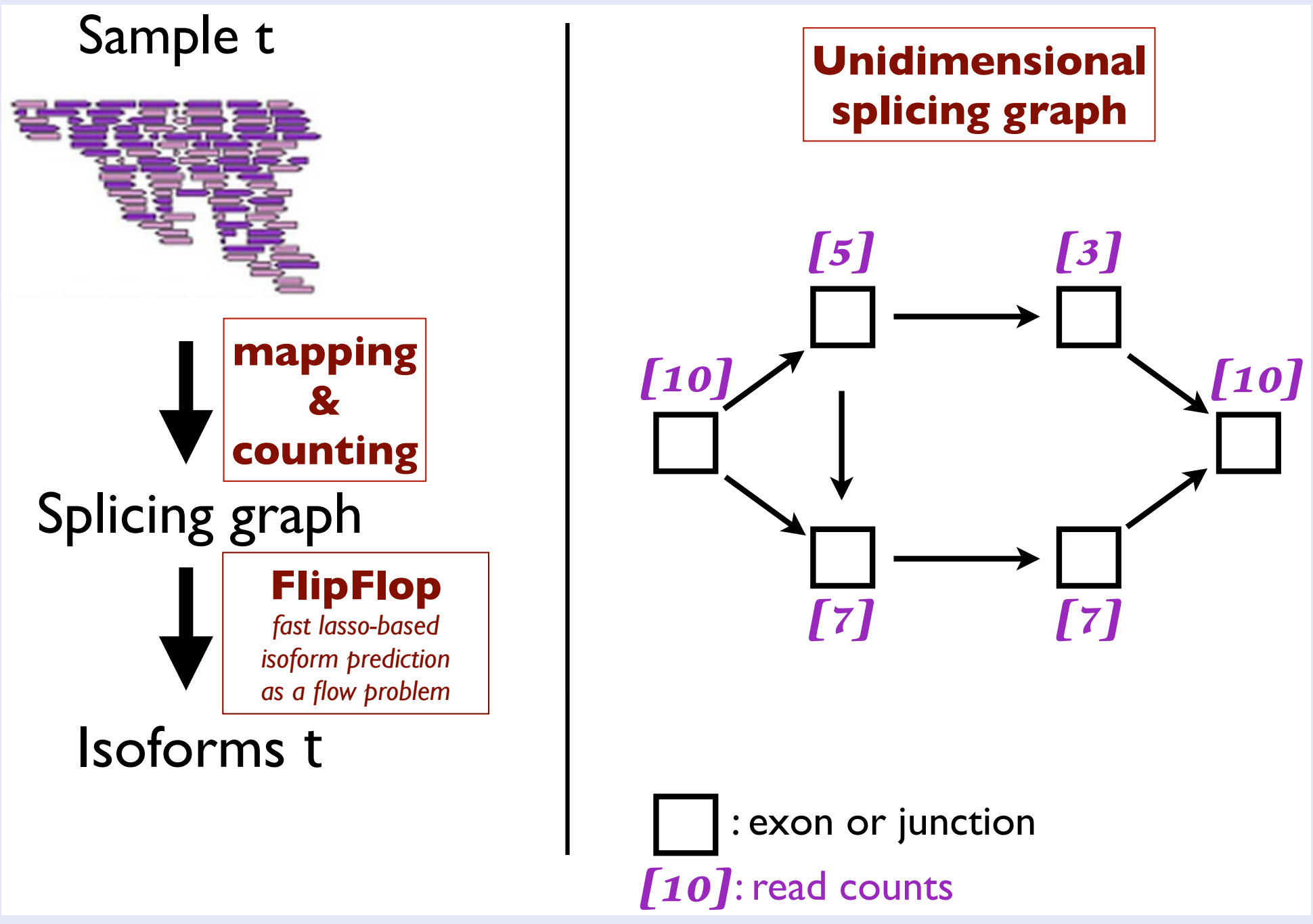
One sample: can we perform fast and accurate de novo isoform reconstruction for one given sample?



Multi-samples: can we improve isoform reconstruction by using all samples simultaneously?

One sample: FlipFlop

Strategy for 1 sample



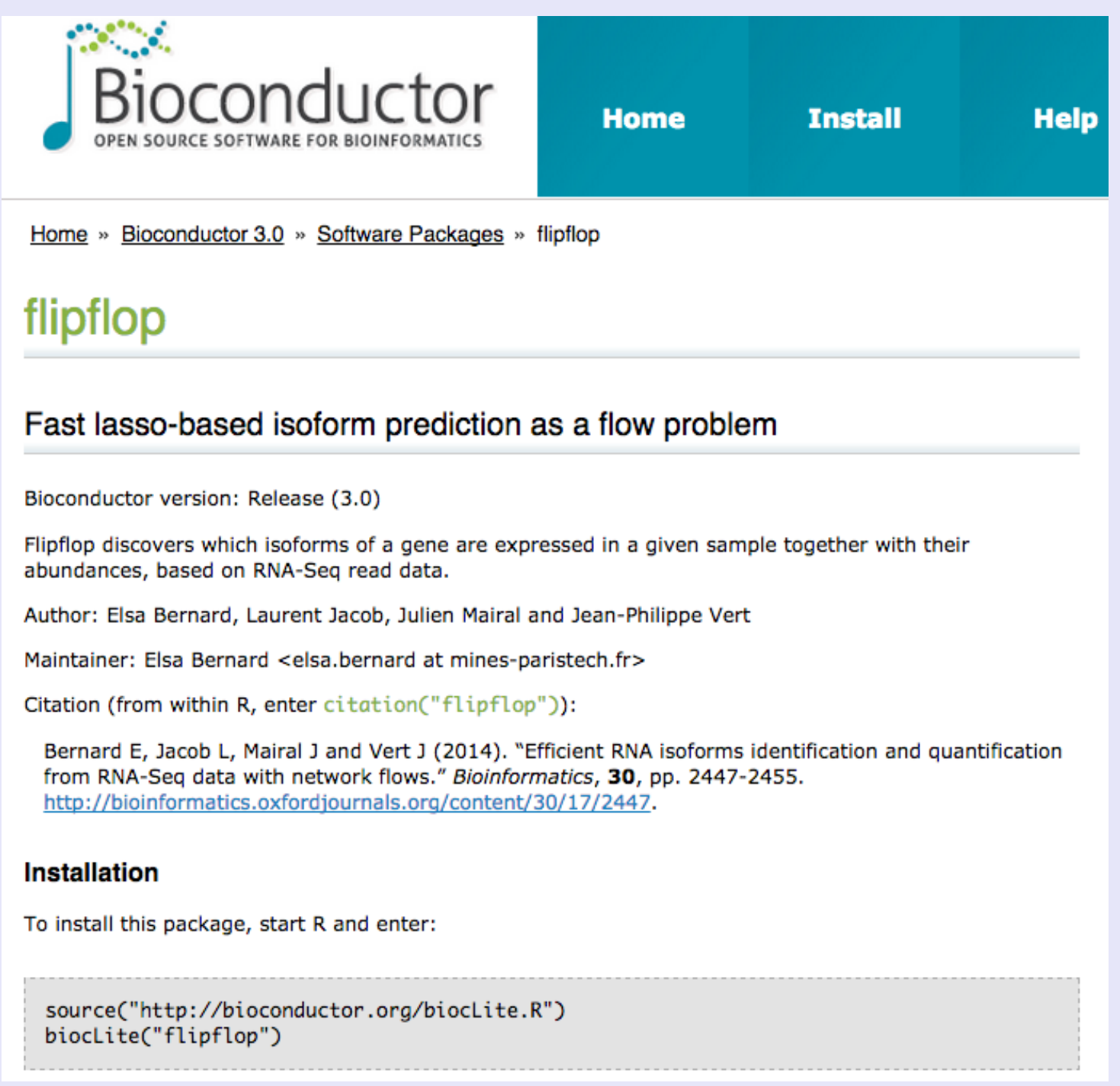
FlipFlop

`http://cbio.ensmp.fr/flipflop/`

Main features

- Solve the isoform deconvolution problem in polynomial time with the number of nodes of the splicing graph
  1. candidate isoforms = all paths in the splicing graph
  2. find a sparse set of paths that explains the observed read counts
  3. network flow formulation with efficient algorithm
- R package

FlipFlop software



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.

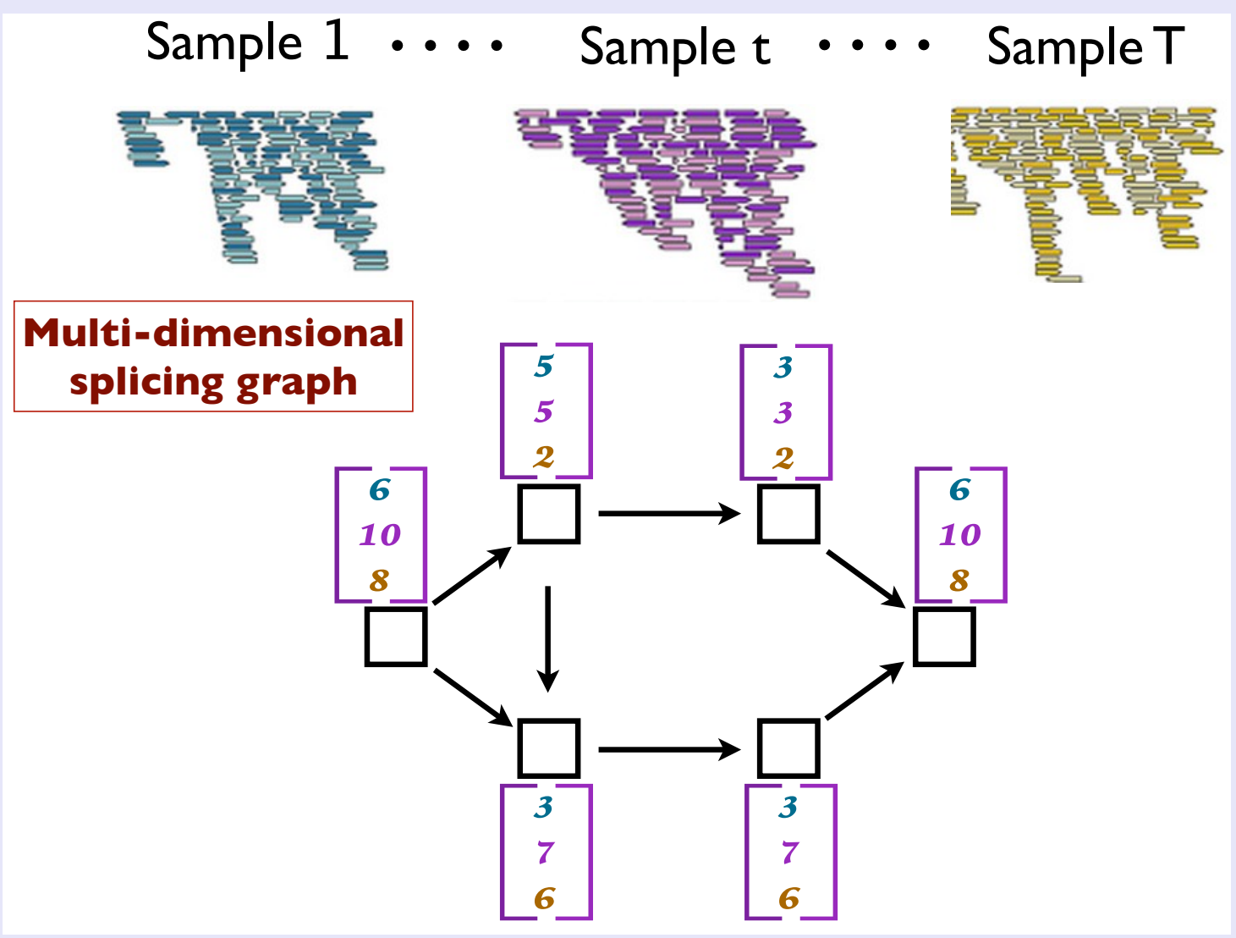
Installation

To install this package, start R and enter:

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

Multi-samples: Group-Lasso

Multi-sample case



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

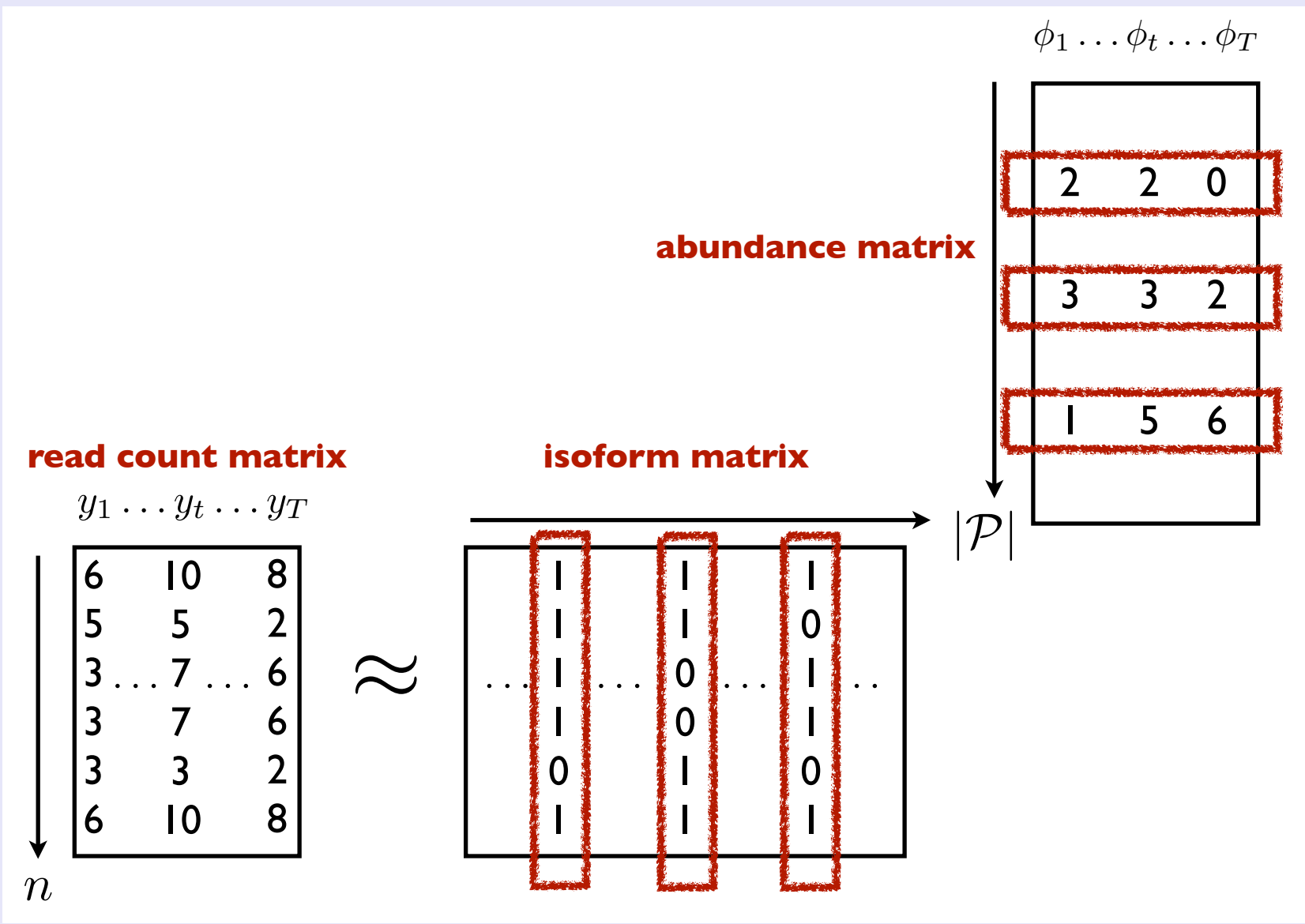
Notations

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

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

  $\phi_1 \dots \phi_t \dots \phi_T$

Idea



$$\begin{pmatrix} 6 & 10 & 8 \\ 5 & 5 & 2 \\ 3 & 7 & 6 \\ 3 & 7 & 6 \\ 3 & 3 & 2 \\ 6 & 10 & 8 \end{pmatrix} \approx \begin{pmatrix} \cdots & 1 & \cdots & 1 & \cdots & 1 & \cdots \\ & 1 & & 1 & & 0 & \\ & 1 & & 0 & & 1 & \\ & 1 & & 0 & & 1 & \\ & 0 & & 1 & & 0 & \\ & 1 & & 1 & & 1 & \end{pmatrix} \begin{pmatrix} 2 & 2 & 0 \\ 3 & 3 & 2 \\ 1 & 5 & 6 \end{pmatrix}$$

Group-sparse regression

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

$$\mathcal{L}(\phi) = \sum_{t=1}^{T} \text{loss}(y_t, \phi_t)$$

- ideally we want to solve the NP-hard L_0 problem

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

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

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

Results

Simulations

- Equal: $\forall t \in \{1, \dots, T\}, \phi_t = \phi_o + \epsilon$
- Different: $\forall t \in \{1, \dots, T\}, \text{supp}\phi_t = \text{supp}\phi_o$

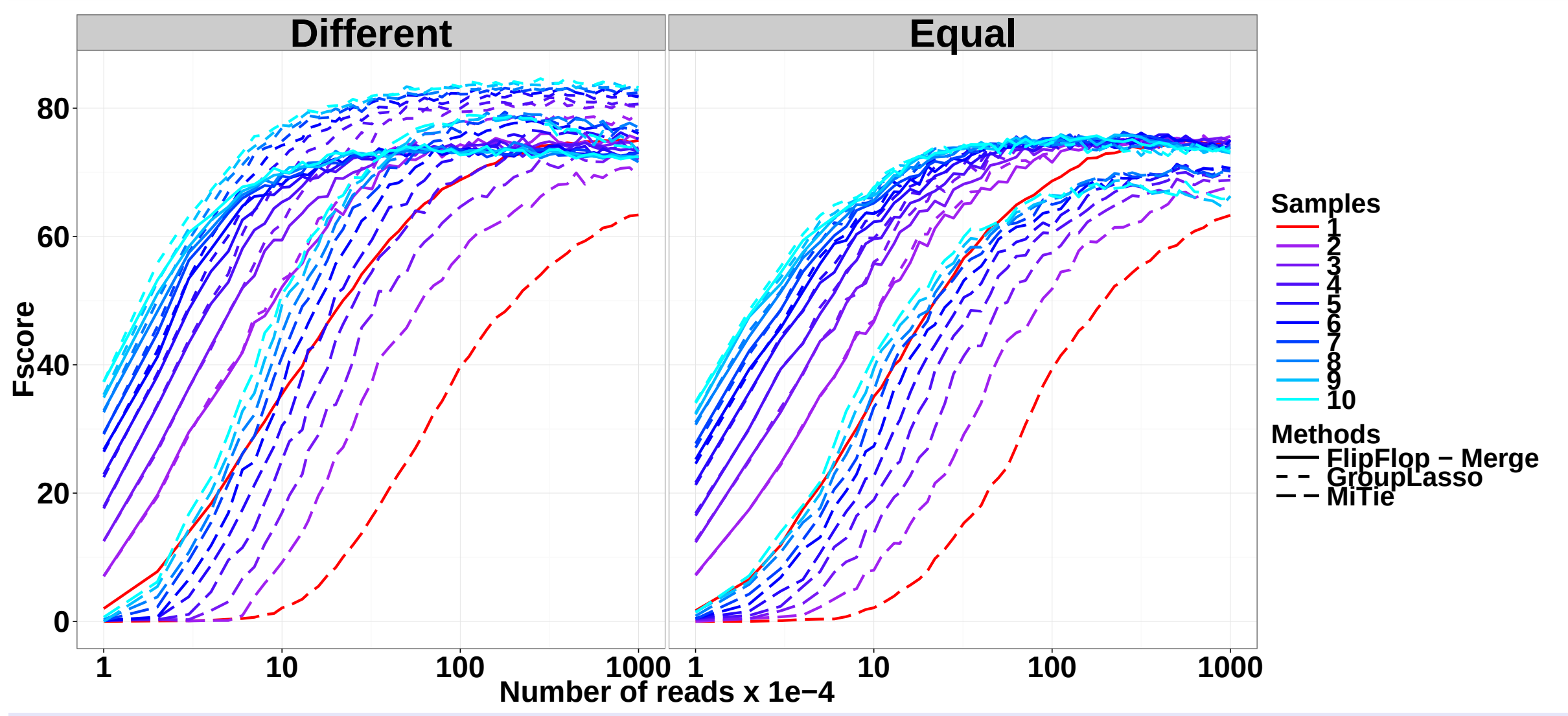


Figure : Fscore on human simulations with increasing coverage and number of samples

Real Data

- Time course development of D.melanogaster

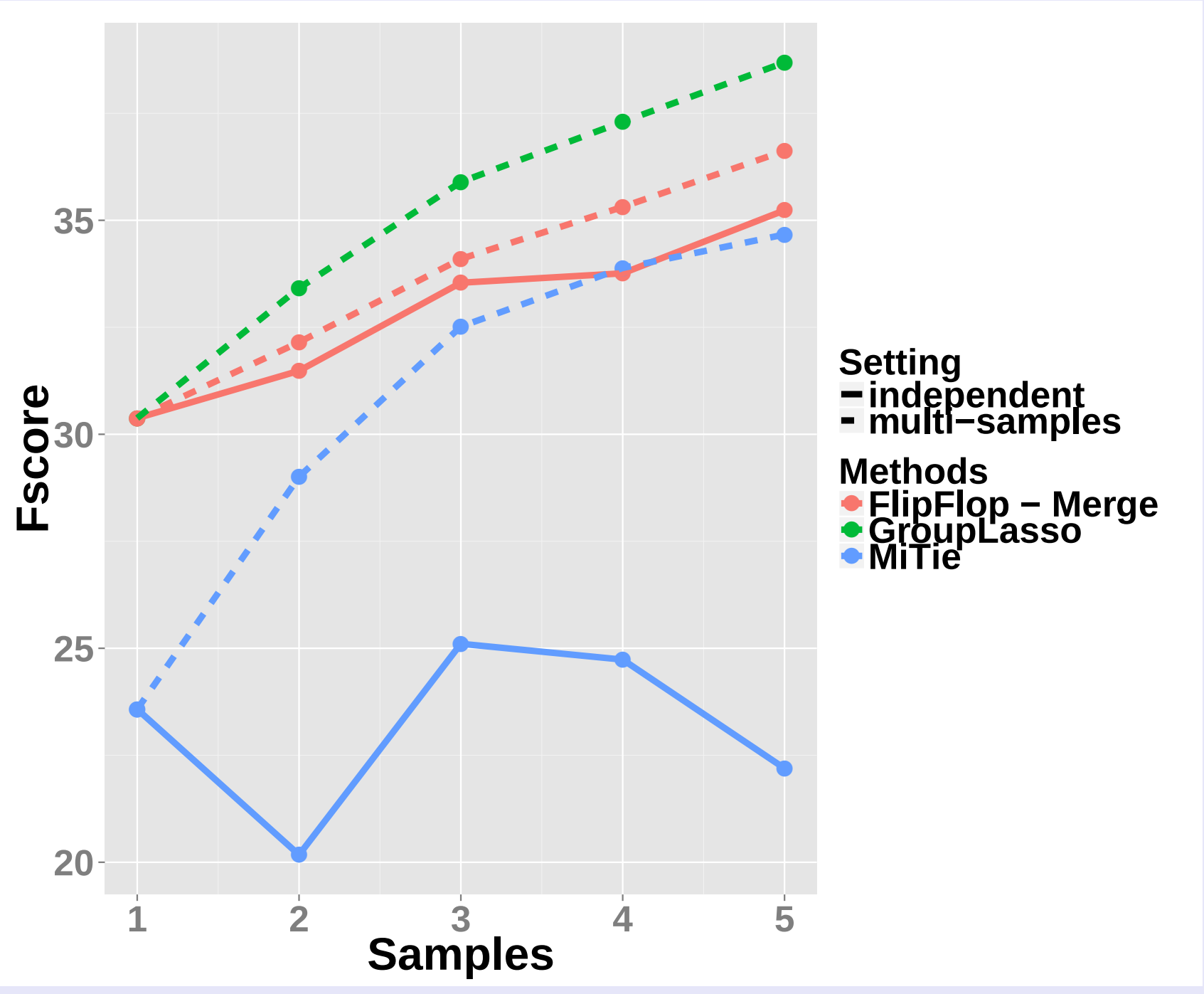


Figure : Fscore on modENCODE data with increasing number of samples

Summary

- New convex optimization formulation for RNA isoform identification and quantification jointly across several samples
- Joint estimation is more powerful than pooling reads across samples
- Competitive with state-of-the-art methods that try to solve a combinatorial formulation of the problem

References

[1] E. Bernard el al. Efficient RNA Isoform Identification and Quantification from RNA-Seq Data with Network Flows. *Bioinformatics*, 2014.

[2] SParse Modelling Software SPAMS `http://lear.inrialpes.fr/people/mairal/software.php`

[3] J. Behr el al. MITIE: Simultaneous RNA-Seq-based transcript identification and quantification in multiple samples. *Bioinformatics*, 2013.

[4] J.Huang et al. A Selective Review of Group Selection in High-Dimensional Models. *Stat Science*, 2012.