



HAL
open science

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

Elsa Bernard, Laurent Jacob, Julien Mairal, Eric Viara, Jean-Philippe Vert

► To cite this version:

Elsa Bernard, Laurent Jacob, Julien Mairal, Eric Viara, Jean-Philippe Vert. A convex formulation for joint RNA isoform detection and quantification from multiple RNA-seq samples. *BMC Bioinformatics*, 2015, 16 (1), pp.262. 10.1186/s12859-015-0695-9 . hal-01123141v2

HAL Id: hal-01123141

<https://minesparis-psl.hal.science/hal-01123141v2>

Submitted on 6 Sep 2015 (v2), last revised 15 Oct 2015 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Elsa Bernard^{1,2,3}, Laurent Jacob⁴, Julien Mairal⁵, Eric Viara⁶, Jean-Philippe Vert^{1,2,3}

September 6, 2015

Abstract

Detecting and quantifying isoforms from RNA-seq data is an important but challenging task. The problem is often ill-posed, particularly at low coverage. One promising direction is to exploit several samples simultaneously. We propose a new method for solving the isoform deconvolution problem jointly across several samples. We formulate a convex optimization problem that allows to share information between samples and that we solve efficiently. We demonstrate the benefits of combining several samples on simulated and real data, and show that our approach outperforms pooling strategies and methods based on integer programming. Our convex formulation to jointly detect and quantify isoforms from RNA-seq data of multiple related samples is a computationally efficient approach to leverage the hypotheses that some isoforms are likely to be present in several samples. The software and source code are available at <http://cbio.ensmp.fr/flipflop>.

1 Introduction

Most genes in eukaryote genomes are subject to alternative splicing Pan *et al.* (2008), meaning they can give rise to different mature mRNA molecules, called transcripts or isoforms, by including or excluding particular exons, retaining introns or using alternative donor or acceptor sites. Alternative splicing is a regulated process that not only greatly increases the repertoire of proteins that can be encoded by the genome Nilsen and Graveley (2010), but also appears to be tissue-specific Wang *et al.* (2008); Xu *et al.* (2002) and regulated in development Kalsotra and Cooper (2011), as well as implicated in diseases such as cancers Pal *et al.* (2012). Hence, detecting isoforms in different cell types or samples is an important step to understand the regulatory programs of the cells or to identify splicing variants responsible for diseases.

Next-generation sequencing (NGS) technologies can be used to identify and quantify these isoforms, using the RNA-seq protocol Mortazavi *et al.* (2008); Wang *et al.* (2009); Martin and Wang (2011). However, identification and quantification of isoforms from RNA-seq data, sometimes referred to as the *isoform deconvolution problem*, is often challenging because RNA-seq technologies usually only sequence short portions of mRNA molecules, called *reads*. A given read sequenced by RNA-seq can therefore originate from different transcripts that share a particular portion containing the read, and a deconvolution step is needed to assign the read to a particular isoform or at least estimate globally which isoforms are present and in which quantity based on all sequenced reads.

¹ MINES ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, Fontainebleau, France, ²Institut Curie, Paris, France, ³INSERM U900, Paris, France, ⁴LBBE, Lyon, France, ⁵LEAR Project-Team, INRIA Grenoble - Rhône Alpes, France, ⁶Sysra, Yerres, France

When a reference genome is available, the RNA-seq reads can be aligned on it using a dedicated splice mapper Trapnell *et al.* (2009); Li and Durbin (2009); Dobin *et al.* (2013), and the deconvolution problem for a given sample consists in estimating a small set of isoforms and their abundances that explain well the observed coverage of reads along the genome. One of the main difficulty lies in the fact that the number of candidate isoforms is very large, growing essentially exponentially with the number of exons. Approaches that try to perform *de novo* isoform reconstruction based on the read alignment include Cufflinks Trapnell *et al.* (2010), Scripture Guttman *et al.* (2010), IsoLasso Li *et al.* (2011b), NSMAP Xia *et al.* (2011), SLIDE Li *et al.* (2011a), iReckon Mezlini *et al.* (2013), Traph Tomescu *et al.* (2013), MiTie Behr *et al.* (2013), and FlipFlop Bernard *et al.* (2014). However, the problem is far from being solved and is still challenging, due in particular to identifiability issues (the fact that different combinations of isoforms can correctly explain the observed reads), particularly at low coverage, which limits the statistical power of the inference methods: as a result, the performance reported by the state-of-the-art is often disappointingly low.

One promising direction to improve isoform deconvolution is to exploit several samples at the same time, such as biological replicates or time course experiments. If some isoforms are shared by several samples, potentially with different abundances, the identifiability issue may vanish and the statistical power of the deconvolution methods may increase due to the availability of more data for estimation. For example, the state-of-the-art methods CLIIQ Lin *et al.* (2012) and MiTie Behr *et al.* (2013) perform joint isoform deconvolution across multiple samples, by formulating the problem as an NP-hard combinatorial problem solved by mixed integer programming. MiTie avoids an explicit enumeration of candidate isoforms using a pruning strategy, which can drastically speed up the computation in some cases but remains very slow in other cases. The Cufflinks/Cuffmerge Trapnell *et al.* (2010) method uses a more naive and straightforward approach, where transcripts are first predicted independently on each sample, before being merged (with some heuristics) in a unique set.

In this paper, we propose a new method for isoform deconvolution from multiple samples. When applied to a single sample, the method boils down to FlipFlop Bernard *et al.* (2014); thus, we simply refer to the new multi-sample extension of the technique as FlipFlop as well. It formulates the isoform deconvolution problem as a continuous convex relaxation of the combinatorial problem solved by CLIIQ and MiTie, using the group-lasso penalty Yuan and Lin (2006); Lounici *et al.* (2009) to impose shared sparsity of the models estimated on each sample. The group-lasso penalty allows to select a few isoforms among many candidates jointly across samples, while assigning sample-specific abundance values. By doing so, it shares information between samples but still considers each sample to be specific, without learning a unique model for all samples together as a merging strategy would do. Compared to CLIIQ or MiTie, FlipFlop addresses a convex optimization problem efficiently, and involves an automatic model selection procedure to balance the fit of the data against the number of detected isoforms. We show experimentally, on simulated and real data, that FlipFlop is more accurate than simple pooling strategies and than other existing methods for isoform deconvolution from multiple samples.

2 Method

The deconvolution problem for a single sample can be cast as a sparse regression problem of the observed reads against expressed isoforms, and solved by penalized regression techniques like the Lasso, where the ℓ_1 penalty controls the number of expressed isoforms. This approach is implemented by several of the referenced methods, including IsoLasso Li *et al.* (2011b) and FlipFlop Bernard *et al.* (2014). When several samples are available, we propose to generalize this

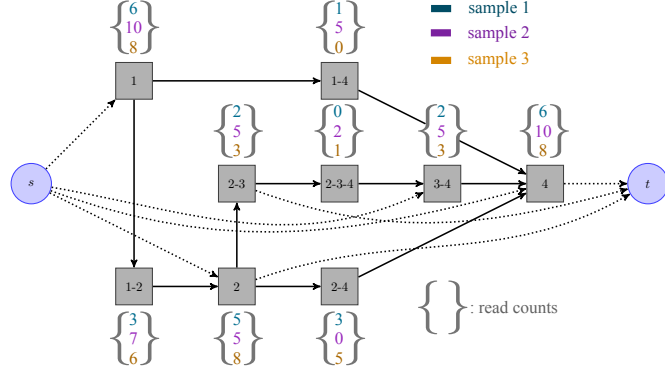


Figure 1: Multi-dimensional splicing graph with three samples. Each candidate isoform is a path from source node s to sink node t . Nodes denoted as grey squares correspond to ordered set of exons. Each read is assigned to a unique node, corresponding to the exact set of exons that it overlaps. Note that more than 2 exons can constitute a node, properly modeling reads spanning more than 2 exons. A vector of read counts (one component per sample) is then associated to each node of the graph. Note also that some components of a vector can be equal to zero.

approach by using a convex penalty that leads to small sets of isoforms jointly expressed across samples, as we explain below.

Multi-dimensional splicing graph

The splicing graph for a gene in a single sample is a directed acyclic graph with a one-to-one mapping between the set of possible isoforms of the gene and the set of paths in the graph. The nodes of the graph typically correspond to exons, sub-exons Li *et al.* (2011b,a); Behr *et al.* (2013) or ordered sets of exons Montgomery *et al.* (2010); Bernard *et al.* (2014)—the definition we adopt here as it allows to properly model long reads spanning more than 2 exons Bernard *et al.* (2014). The directed edges correspond to links between possibly adjacent nodes.

When working with several samples, we choose to build the graph based on the read alignments of all samples pooled together. Since the exons used to build the graph are estimated from read clusters, this step already takes advantage of information from multiple samples, and leads to a more accurate graph. We associate a list of read counts, as many as samples, with each node of the graph. In other words, we extend the notion of splicing graph to the multiple-sample framework, using a shared graph structure with specific count values on each node. Our multi-dimensional splicing graph is illustrated in figure 1.

Notation

Throughout the paper, we call $G = (V, E)$ the multi-dimensional splicing graph where V is the set of vertices and E the set of edges. We denote by \mathcal{P} the set of all paths in G . By construction of the graph, each path $p \in \mathcal{P}$ corresponds to a unique candidate isoform. We denote by y_v^t the number of reads falling in each node $v \in V$ for each sample $t \in \{1, \dots, T\}$, where T is the number of samples. We denote by $\beta_p^t \in \mathbb{R}_+$ the abundance of isoform p for sample t . Finally, we define for every path p in \mathcal{P} the T -dimensional vector of abundances $\beta_p = [\beta_p^1, \beta_p^2, \dots, \beta_p^T]$, and denote by $\beta = [\beta_p]_{p \in \mathcal{P}}$ the matrix of all abundances values with $|\mathcal{P}|$ rows and T columns.

Joint sparse estimation

We propose to estimate β through the following penalized regression problem:

$$\min_{\beta} \mathcal{L}(\beta) + \lambda \sum_{p \in \mathcal{P}} \|\beta_p\|_2 \quad \text{such that } \beta_p \geq 0 \text{ for all } p \in \mathcal{P}, \quad (1)$$

where \mathcal{L} is a convex smooth loss function defined below, $\|\beta_p\|_2 = \sqrt{\sum_{t=1}^T (\beta_p^t)^2}$ is the Euclidean norm of the vector of abundances of isoform p across the samples, and λ is a non-negative regularization parameter that controls the trade-off between loss and sparsity. The $\ell_{1,2}$ -norm $\|\beta\|_{1,2} = \sum_{p \in \mathcal{P}} \|\beta_p\|_2$, sometimes called the group-lasso penalty Yuan and Lin (2006), induces a shared sparsity pattern across samples: solutions of (1) typically have entire rows equal to zero Yuan and Lin (2006), while the abundance values in the non-zero rows can be different among samples. This shared sparsity-inducing effect corresponds exactly to our assumption that only a limited number of isoforms are present across the samples (non-zero rows of β). It can be thought of as a convex relaxation of the number of isoforms present in at least one sample, which is used as criterion in the combinatorial formulations of CLIQ and MiTie.

We define the loss function \mathcal{L} as the sum of the T sample losses, thus assuming independence between samples as reads are sampled independently from each sample. The loss is derived from the Poisson negative likelihood (the Poisson model has been successfully used in several RNA-seq studies Jiang and Wong (2009); Salzman *et al.* (2011); Xia *et al.* (2011); Bernard *et al.* (2014)) so that the general loss is defined as

$$\mathcal{L}(\beta) = \sum_{t=1}^T \sum_{v \in V} [\delta_v^t - y_v^t \log \delta_v^t] \quad \text{with } \delta_v^t = \left(N^t l_v \sum_{p \in \mathcal{P}: p \ni v} \beta_p^t \right),$$

where N^t is the total number of mapped reads in sample t and l_v is the effective length of node v , as defined in Bernard *et al.* (2014). The sum $\sum \beta_p^t$ over all $p \in \mathcal{P}$ that contain node v represents the sum of expressions in sample t of all isoforms involving node v .

Candidate isoforms

Since $|\mathcal{P}|$ grows exponentially with the number of nodes in G , we need to avoid an exhaustive enumeration of all candidate isoforms $p \in \mathcal{P}$. FlipFlop efficiently solves problem (1) in the case where $T = 1$, *i.e.*, the ℓ_1 -regularized regression $\min_{\beta_p \in \mathbb{R}_+} \mathcal{L}(\beta) + \lambda \sum_{p \in \mathcal{P}} \beta_p$ using network flow techniques, without requiring an exhaustive path enumeration and leading to a polynomial-time algorithm in the number of nodes.

Unfortunately, this network flow formulation does not extend trivially to the multi-sample case. We therefore resort to a natural two-step heuristic: we first generate a large set of candidate isoforms by solving $T + 1$ one-dimensional problems—the T independent ones, plus the one corresponding to all samples pooled together—for different values of λ , and taking the union of all selected isoforms, and we then solve (1) restricted to this union of isoforms. This approach can potentially miss isoforms which would be selected by solving (1) over all paths $p \in \mathcal{P}$ and are not selected for any single sample or when pooling all reads to form a single sample, but allows to efficiently approximate (1). We observe that it leads to good results in various settings in practice, as shown in the experimental part.

Model selection

We solve (1) for a large range of values of the regularization parameter λ , obtaining solutions from very sparse to more dense (a sparse solution involves few non-zero abundance vectors β_p). Each solution, *i.e.*, each set of selected isoforms obtained with a particular λ value, is then re-fitted against individual samples—without regularization but keeping the non-negativity constraint—so that the estimated abundances do not suffer from shrinkage Tibshirani (1996). The solution with the largest BIC criterion Schwarz (1978), where the degree of freedom of a group-lasso solution is computed as explained in Yuan and Lin (2006), is finally selected. Note that although the same list of isoforms selected by the group-lasso is tested on each sample, the refitting step lets each sample pick the subset of isoforms it needs among the list, meaning that all samples do not necessarily share *all* isoforms at the end of the deconvolution.

3 Results

We show results on simulated human RNA-seq data with both increasing coverage and increasing number of samples, with different simulation settings, and on real RNA-seq data. In all cases, reads are mapped to the reference with TopHat2 Trapnell *et al.* (2009). We compare FlipFlop implementing the group-lasso approach (1) to the simpler strategy of pooling all samples together, running single-sample FlipFlop Bernard *et al.* (2014) on the merged data, and performing a fit for each individual sample data against the selected isoforms. We also assess the performance of MiTie Behr *et al.* (2013) and of the version 2.2.0 of the Cufflinks/Cuffmerge package Trapnell *et al.* (2010). Performances on isoform identification are summarized in terms of Fscore, the harmonic mean of precision and recall, as used in other RNA-seq studies Lin *et al.* (2012); Behr *et al.* (2013). Of note, in all the following experiments, we consider a *de novo* setting, without feeding any of the methods with prior transcript annotations (*i.e.*, MiTie and FlipFlop first reconstruct sub-exons and build the splicing graph, then perform isoform deconvolution).

Influence of coverage and sample number

The first set of simulations is performed based on the 1329 multi-exon transcripts on the positive strand of chromosome 11 from the RefSeq annotation Pruitt *et al.* (2005). Single-end 150bp reads are simulated with the RNASeqReadSimulator software (available at <http://alumni.cs.ucr.edu/~liw/rnaseqreadsimulator.html>). We vary the number of reads from 10 thousand to 10 million per sample (corresponding approximately to sequencing depth from 1X to 1000X) and the number of samples from 1 to 10. All methods are run with default parameters, except that we fix *region-filter* to 40 and *max-num-trans* to 10 in MiTie as we notice that choosing these two parameter values greatly increases its performances (see figure A.1 of the supplementary material for a comparison between MiTie with default parameters or not).

Figure 2 shows the Fscore in two different settings: the *Equal* setting corresponds to a case where all samples express the same set of transcripts at the same abundances (in other words each sample is a noisy realization of a unique abundance profile), while in the *Different* setting the abundance profiles of each sample are generated independently. Hence in that case the samples share the same set of expressed transcripts but have very different expression values (the maximum correlation between two abundance vectors is 0.088).

In all cases and for all methods, the higher the coverage or the number of samples, the higher the Fscore. In the *Equal* case, the group-lasso and merging strategies give almost identical results, which shows the good behavior of the group-lasso, as pooling samples in that case corresponds

to learning the shared abundance profile. In the *Equal* case again, for all methods the different Fscore curves obtained with increasing number of samples converge to different plateaux. None of these levels reaches a Fscore of 100, but the group-lasso level is the highest (together with the merging strategy). In the *Different* case, the group-lasso shows equal or higher Fscore than the merging strategy, with a great improvement when the coverage or the number of samples increases. The group-lasso also outperforms the Cufflinks/Cuffmerge method for all numbers of samples when the coverage is larger than 80. When using more than 5 samples the group-lasso shows greater Fscore as soon as the coverage is bigger than 15 (see table B.1 of the supplementary material for statistical significance). Finally, the group-lasso outperforms MiTie for all number of samples and all coverages. Of note, the group-lasso performances are better in the *Different* setting than in the *Equal* setting, showing that our multi-sample method can efficiently deal with diversity among samples.

We also investigate the influence of the read length on the performance of the compared methods in the *Different* setting. Figure 3 shows the obtained Fscore when using either 2 or 5 samples with a fixed 100×10^4 coverage, while read length varies from 75bp to 300bp. Because we properly model long reads in our splicing graph the group-lasso performance greatly increases with the read length, proportionally much more than other state-of-the-art methods. When using 5 samples and long 300bp reads, the group-lasso reaches a very high Fscore of 90 (compared to 84 for the second best Cufflinks/Cuffmerge method), showing that our method is very well adapted to RNA-Seq design with long reads and several biological replicates.

Note finally that our method generalizes to paired-end reads. We show in figure C.1 of the supplementary material a comparison of the tested methods on simulations in the *Different* setting using both paired or single-end reads at comparable coverage.

Influence of hyper-parameters with realistic simulations

The second set of simulations is performed using a different and more realistic simulator, the Flux Simulator Griebel *et al.* (2012), in order to check that our approach performs well regardless the choice of the simulator. Coverage and single-end read length are respectively fixed to 10^5 reads and 150bp, and we run experiments for one up to five samples. We study the influence of hyper-parameters on the performances of the compared methods, and show that our approach leads to better results with optimized parameters as well. Hyper-parameters are first tuned on a training set of 600 transcripts from the positive strand of chromosome 11, which is subsequently left aside from the evaluation procedure after tuning. We start by jointly optimizing a set of pre-processing hyperparameters. We then keep the combination that leads to the best training Fscore, and we jointly optimize a set of prediction hyperparameters. More specifically, we optimize 7 values of 3 different pre-processing or prediction parameters (hence 7^3 different combinations in both cases), except that for MiTie we add 2 values of one pre-processing parameter and 3 values of a fourth prediction parameter (hence optimizing over 9×7^2 and 3×7^3 parameters). A more detailed description of the optimized parameters is given in tables D.1 and D.2 of the supplementary.

Fscore is shown on figure 4 for 600 other test transcripts, for both default and tuned settings (except that again we set *region-filter* to 40 and *max-num-trans* to 10 in MiTie instead of using all default parameters as it greatly improves its performances, see figure A.2 for a comparison of several versions of MiTie). For all methods and for both default and tuned settings, performances increase with the number of samples. Except for Cufflinks/Cuffmerge for the last three sample numbers, all methods improve their results after tuning of their hyper-parameters. When using default parameter values, the group-lasso shows the largest Fscore for the first three sample numbers, while Cufflinks/Cuffmerge is slightly better for the very last sample number. When using tuned

parameter values, the group-lasso approach outperforms all other methods for the first three sample numbers, and is slightly better or equal to the default version of Cufflinks/Cuffmerge for the last two sample numbers.

Experiments with real data

We use five samples from time course experiments on *D. melanogaster* embryonic development. Each sample corresponds to a 2-hour period, from 0 to 10 hours (0-2h, 2-4h, . . . , 8-10h). Data is available from the modENCODE Celniker *et al.* (2009) website. For each given period we pooled all 75bp single-end technical replicate reads available, ending up with approximately 25 to 45 million mapped reads per sample. A description of the samples is given in table C.1. Data from the same source were also used in the MiTie paper Behr *et al.* (2013).

Because the exact true sets of expressed transcripts is not known, we validated predictions based on public transcript annotations. We built a comprehensive reference using three different databases available on the UCSC genome browser Karolchik *et al.* (2004), namely the RefSeq Pruitt *et al.* (2005), Ensembl Cunningham *et al.* (2015) and FlyBase Marygold *et al.* (2013) annotations. More specifically, we took the union of the multi-exon transcripts described in the three databases, while considering transcripts with the same internal exon/intron structure but with different length of the first or the last exon as duplicates. Reads were mapped to the reference transcriptome in order to restrict predictions to known genomic regions, and we perform independent analysis on the forward and reverse strands. All methods are run with default parameters.

Figure 5 shows the Fscore per sample when FlipFlop, MiTie, and Cufflinks are run independently on each sample or when multi-sample strategies are used. Results on the forward and reverse strands are extremely similar. All methods give better results than their independent versions, and the performances of the multi-sample approaches increase with the number of used samples. Again, the group-lasso strategy of FlipFlop seems more powerful than the pooling strategy, and gives better Fscore than MiTie and Cufflinks/Cuffmerge in that context.

Considering running time, each method was run on a 48 CPU machine at 2.2GHz with 256GB of RAM using 6 threads (all tools support multi-threading). When using only a single sample and 6 threads, Cufflinks, FlipFlop and MiTie respectively completed in ~ 4.2 min, ~ 9.5 min and ~ 26.6 min. while when using 5 samples and 6 threads, Cufflinks/Cuffmerge, FlipFlop with group-lasso and MiTie took ~ 0.45 h, ~ 1 h and ~ 25 h (see figure G.1 of the supplementary).

Illustrative examples

We describe an example as a proof of concept that multi-sample FlipFlop with the group-lasso approach (1) can be much more powerful in some cases than its independent FlipFlop version, and than the merging strategy of Cufflinks/Cuffmerge. Figure 6 shows transcriptome assemblies of gene CG15717 on the first three modENCODE samples presented in the previous section, denoted as 0-2h, 2-4h and 4-6h on the figure. For each sample, we display the read coverage along the gene, the junctions between exons, and the single-sample FlipFlop and Cufflinks predictions. At the bottom of the figure, we show the 6 RefSeq records as well as the multi-sample predictions obtained with FlipFlop or with Cuffmerge. A predicted transcript is considered as valid if all its exon/intron boundaries match a RefSeq record (✓ and ✗ denote validity or not). The estimated abundances in FPKM are given on the right-hand side of each predicted transcript. Of note, the group-lasso predictions come with estimated abundances (one specific value per sample), whereas Cufflinks/Cuffmerge only reports the structure of the transcripts.

For single-sample predictions, FlipFlop and Cufflinks report the same number of transcripts

for each sample (respectively 2, 2 and 3 predictions for samples 0-2h, 2-4h and 4-6h), with the same number of valid transcripts, except for the first sample where FlipFlop makes 2 good guesses against 1 for Cufflinks. This difference might be due to the fact that FlipFlop not only tries to explain the read alignment as Cufflinks does, but also the coverage discrepancies along the gene.

For multi-sample predictions, FlipFlop gives much more reliable results, with 4 validated transcripts (among 4 predictions), while Cufflinks/Cuffmerge makes only 1 good guess out of 2 predictions. FlipFlop uses evidences from all samples together to find transcripts with for instance missing junction reads in one of the sample (such as the one with 30, 7 and 20 FPKM) or lowly expressed transcripts (such as the one with 0, 0.5 and 2 FPKM). Cufflinks/Cuffmerge explains all read junctions but does not seek to explain the multi-sample coverage, which seems important in that example.

Importantly, one can note that the results of multi-sample group-lasso FlipFlop are different from the union of all single-sample FlipFlop predictions (the union coincides here to the results of FlipFlop on the merged sample—data not shown). This illustrates the fact that designing a dedicated multi-sample procedure can lead to more statistical power than merging individual results obtained on each sample independently. We display an additional example in figure H.1 of the supplementary.

4 Discussion

We proposed a multi-sample extension of FlipFlop, which implements a new convex optimization formulation for RNA isoform identification and quantification jointly across several samples. Experiments on simulated and real data show that an appropriate method for joint estimation is more powerful than a naive pooling of reads across samples. We also obtained promising results compared to MiTie, which tries to solve a combinatorial formulation of the problem.

Accurately estimating isoforms in multiple samples is an important preliminary step to differential expression studies at the level of isoforms Anders *et al.* (2012); Trapnell *et al.* (2013). Indeed, isoform deconvolution from single samples suffers from high false positive and false negatives rates, making the comparison between different samples even more difficult if isoforms are estimated from each sample independently. Although the FlipFlop formulation of joint isoform deconvolution across samples provides a useful solution to define a list of isoforms expressed (or not) in each sample, variants of FlipFlop specifically dedicated to the problem of finding differentially expressed isoforms may also be possible by changing the objective function optimized in (1).

Finally, as future multi-sample applications such as jointly analyzing large cohorts of cancer samples or many cells in single-cell RNA-seq are likely to involve hundreds or thousands of samples, more efficient implementations involving in particular distributed optimization may be needed.

Fundings: This work was supported by the European Research Council [SMAC-ERC-280032 to J-P.V., E.B.]; the European Commission [HEALTH-F5-2012-305626 to J-P.V., E.B.]; and the French National Research Agency [ANR-09-BLAN-0051-04, ANR-11-BINF-0001 to J-P.V., E.B., ANR-14-CE23-0003-01 to J.M, L.J.].

References

- Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from rna-seq data. *Genome Res*, **22**, 2008–2017.
- Behr, J., Kahles, A., Zhong, Y., Sreedharan, V. T., Drewe, P., and Ratsch, G. (2013). Mitie: Simultaneous rna-seq based transcript identification and quantification in multiple samples. *Bioinformatics*, **29**(20), 2529–2538.

- Bernard, E., Jacob, L., Mairal, J., and Vert, J.-P. (2014). Efficient rna isoform identification and quantification from rna-seq data with network flows. *Bioinformatics*, **30**(17), 2447–2455.
- Celniker, E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Kerpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., *et al.* (2009). Unlocking the secrets of the genome. *Nature*, **459**(7249), 927–930.
- Cunningham, F., Amode, M. R., Barrell, D., *et al.* (2015). Ensembl 2015. *Nucleic Acids Res*, **43**(D1), D662–D669.
- Dobin, A., Carrie, A., Schlesinger, F., Drenkow, J., Zaleski, C., Sonali, J., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1), 15–21.
- Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigo, R., and Sammeth, M. (2012). Modelling and simulating generic rna-seq experiments with the flux simulator. *Nucleic Acids Res*, **40**(20), 10073–10083.
- Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., *et al.* (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotech*, **28**(5), 503–510.
- Jiang, H. and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**(8), 1026–1032.
- Kalsotra, A. and Cooper, T. A. (2011). Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet*, **12**(10), 715–729.
- Karolchik, D., Hinrichs, A. S., Furey, T. S., *et al.* (2004). The ucsc table browser data retrieval tool. *Nucleic Acids Res*, **32**(suppl), D493–D496.
- Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics*, **25**(14), 1754–1760.
- Li, J. J., Jiang, C.-R., J., B. B., Huang, H., and Bickel, P. J. (2011a). Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *P Natl Acad Sci USA*, **108**(50), 19867–19872.
- Li, W., Feng, J., and Jiang, T. (2011b). IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J Comput Biol*, **18**(11), 1693–1707.
- Lin, Y.-Y., Dao, P., Hach, F., Bakhshi, M., Mo, F., Lapuk, A., Collins, C., and Sahinalp, S. C. (2012). Cliq: Accurate comparative detection and quantification of expressed isoforms in a population. In B. J. Raphael and J. Tang, editors, *WABI*, volume 7534 of *Lecture Notes in Computer Science*, pages 178–189.
- Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. In *Proceedings of the 22nd Conference on Information Theory*, pages 73–82.
- Martin, J. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet*, **12**(10), 671–682.
- Marygold, S. J., Leyland, P. C., Seal, R. L., *et al.* (2013). Flybase: improvements to the bibliography. *Nucleic Acids Res*, **41**(D1), D751–D757.
- Mezlini, A. M., M., S. E. J., Fiume, M., Buske, O., Savich, G., Shah, S., Aparicion, S., Chiang, D., Goldenberg, A., and Brudno, M. (2013). iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res*, **23**(3), 519–529.
- Montgomery, S. B. *et al.* (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**(7289), 773–777.
- Mortazavi, A., Williams, B. A., McCue, K., and Schaeffer, L. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**(7), 621–628.
- Nilsen, T. W. and Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, **463**(7280), 457–463.
- Pal, S., Gupta, R., and Davuluri, R. V. (2012). Alternative transcription and alternative splicing in cancer. *Pharmacol Ther*, **136**(3), 283–294.
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **40**(12), 1413–1415.
- Pruitt, K., Tatusova, T., and Maglott, D. R. (2005). Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**(suppl), D501–D504.
- Salzman, J., Jiang, H., and Wong, W. H. (2011). Statistical modeling of RNA-Seq data. *Stat Sci*, **26**(1), 62–83.
- Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann Stat*, **6**(2), 461–464.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B*, **58**(1), 267–288.

- Tomescu, A., Kuosmanen, A., Rizzi, R., and Makinen, V. (2013). A novel min-cost flow method for estimating transcript expression with rna-seq. *BMC Bioinformatics*, **14** (Suppl 5), S15.
- Trapnell, C., Patcher, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9), 1105–1111.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A. M., Kwan, G., van Baren, M. J., L., S. S., Wold, B., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**(5), 511–515.
- Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Patcher, L. (2013). Differential analysis of gene regulation at transcript resolution with rna-seq. *Nat Biotechnol*, **31**(1), 46–53.
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., and Mayr, C. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(721), 470–476.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**(1), 57–63.
- Xia, Z., J., W., Chang, C.-C., and Zhou, X. (2011). NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics*, **12**, 162.
- Xu, Q., Modrek, K., and Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*, **30**(17), 3754–3766.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser.*, **68**(1), 49–67.

Appendices

- A** Some influence of MiTie parameters on human simulations
- B** Statistical significance on human simulations
- C** Paired-end simulations
- D** Parameter optimization on human simulations
- E** Description of real RNA-seq data
- F** Abundance comparison on real RNA-seq data
- G** Running time on real RNA-seq data
- H** More illustrative examples

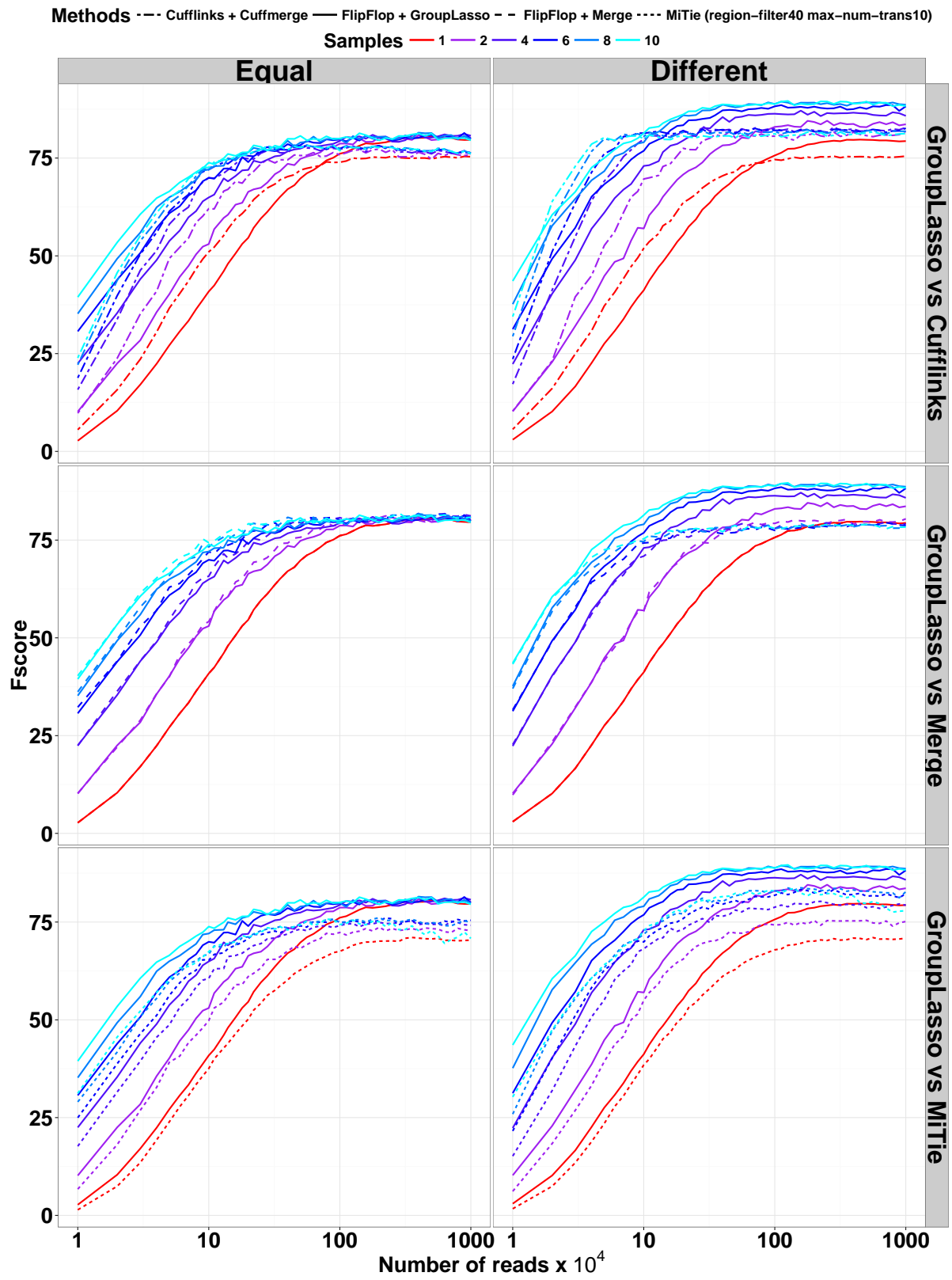


Figure 2: Human simulations with increasing coverage and number of samples.

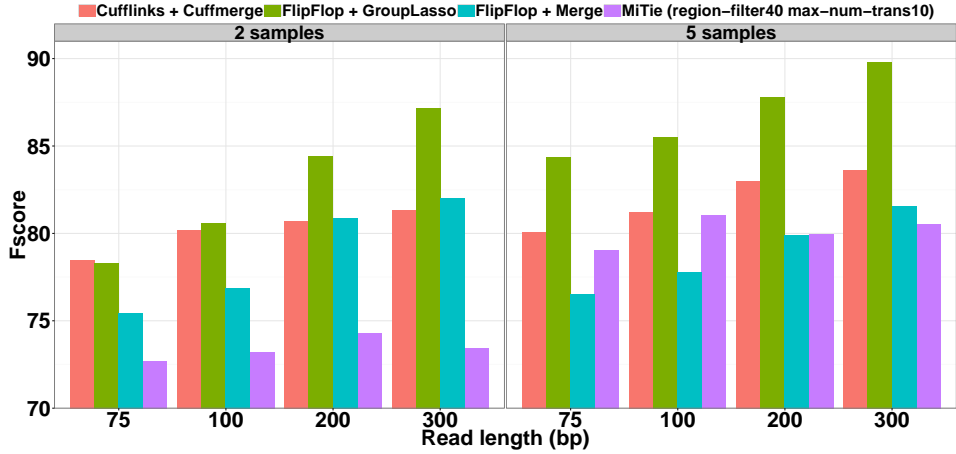


Figure 3: Human simulations with various read lengths.

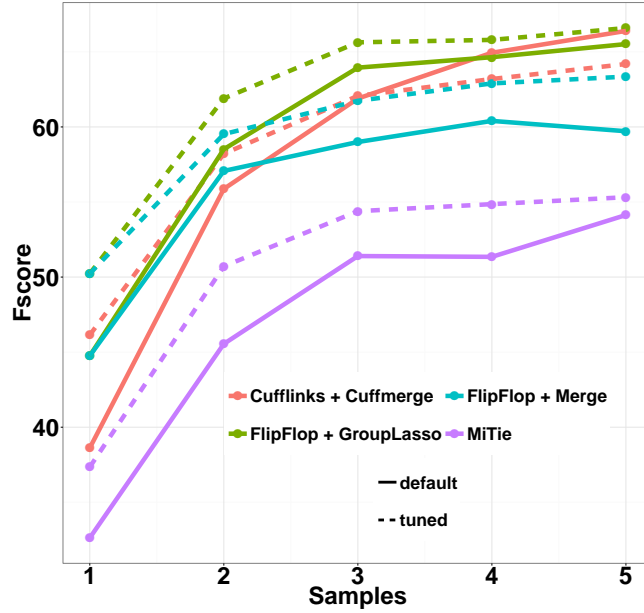


Figure 4: Fscore results on the Flux Simulator simulations.

	1-3 samples	4-6 samples	7-10 samples
1-10 coverage ($\times 10^4$ reads)	1	1	1
10-50 coverage ($\times 10^4$ reads)	1	0.035	$< 10^{-16}$
50-100 coverage ($\times 10^4$ reads)	0.040	$< 10^{-16}$	$< 10^{-16}$
100-1000 coverage ($\times 10^4$ reads)	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$

Table B.1: Statistical testing to assess performances in the *Different* human simulation setting, for different ranges of coverage and number of samples. Numbers correspond to the BH adjusted p-values when testing the null hypothesis that the Fscore obtained with FlipFlop+GroupLasso are lower than the Fscore obtained with Cufflinks+Cuffmerge (one-sided paired t-test). Note that when testing FlipFlop+GroupLasso against MiTie, all adjusted p-values are extremely small.

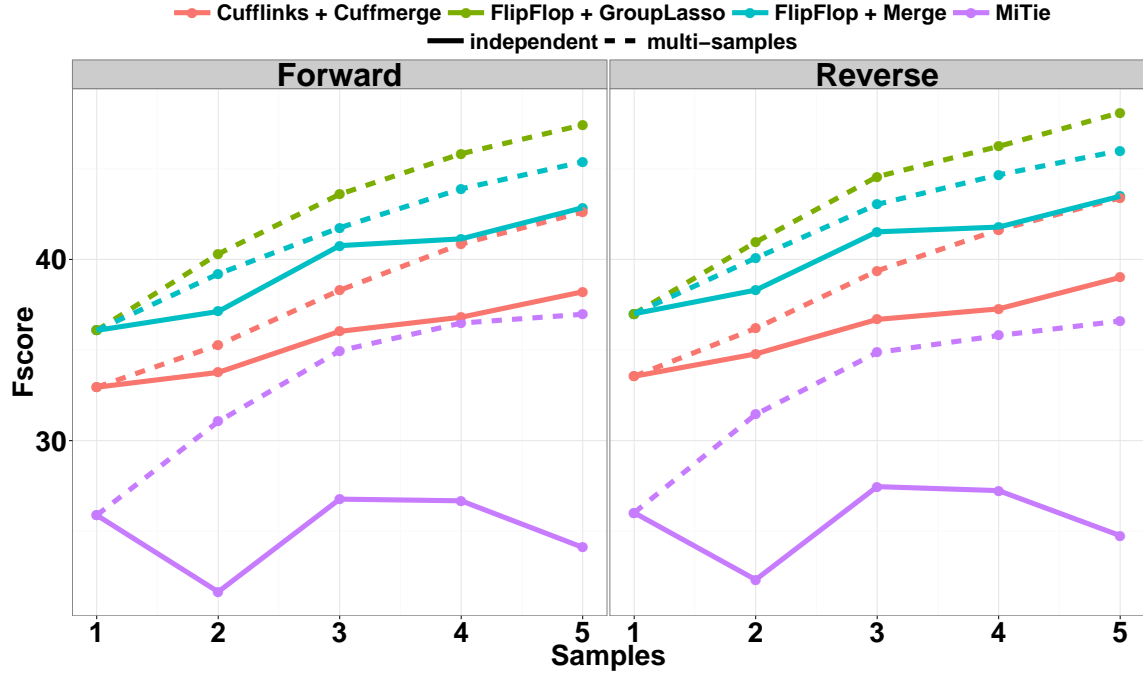


Figure 5: Fscore results on the modENCODE data.

Methods	Pre-processing parameters (with default values)	Optimal values for each number of samples				
		1	2	3	4	5
MiTie	region-filter (1000)	50	50	50	50	10
	seg-filter (0.05)	0.01	0.01	0.01	0.01	0.01
	tss-tts-pval (10^{-4})	6×10^{-5}	6×10^{-5}	2×10^{-5}	6×10^{-5}	6×10^{-5}
Cufflinks	min-frags-per-transfrag (10)	29	17	17	17	29
	max-multiread-fraction (0.75)	0.15	0.15	0.15	0.15	0.15
	overlap-radius (50)	146	85	85	85	146
FlipFlop + Merge	minReadNum (40)	23	40	23	8	14
	minJuncCount (1)	1	1	1	1	1
	minCvgCut (0.05)	0.02	0.03	0.01	0.01	0.01
FlipFlop + GroupLasso	minReadNum (40)	23	40	23	8	14
	minJuncCount (1)	1	1	1	1	1
	minCvgCut (0.05)	0.02	0.01	0.01	0.01	0.01

Table D.1: Details on the optimized pre-processing parameters.

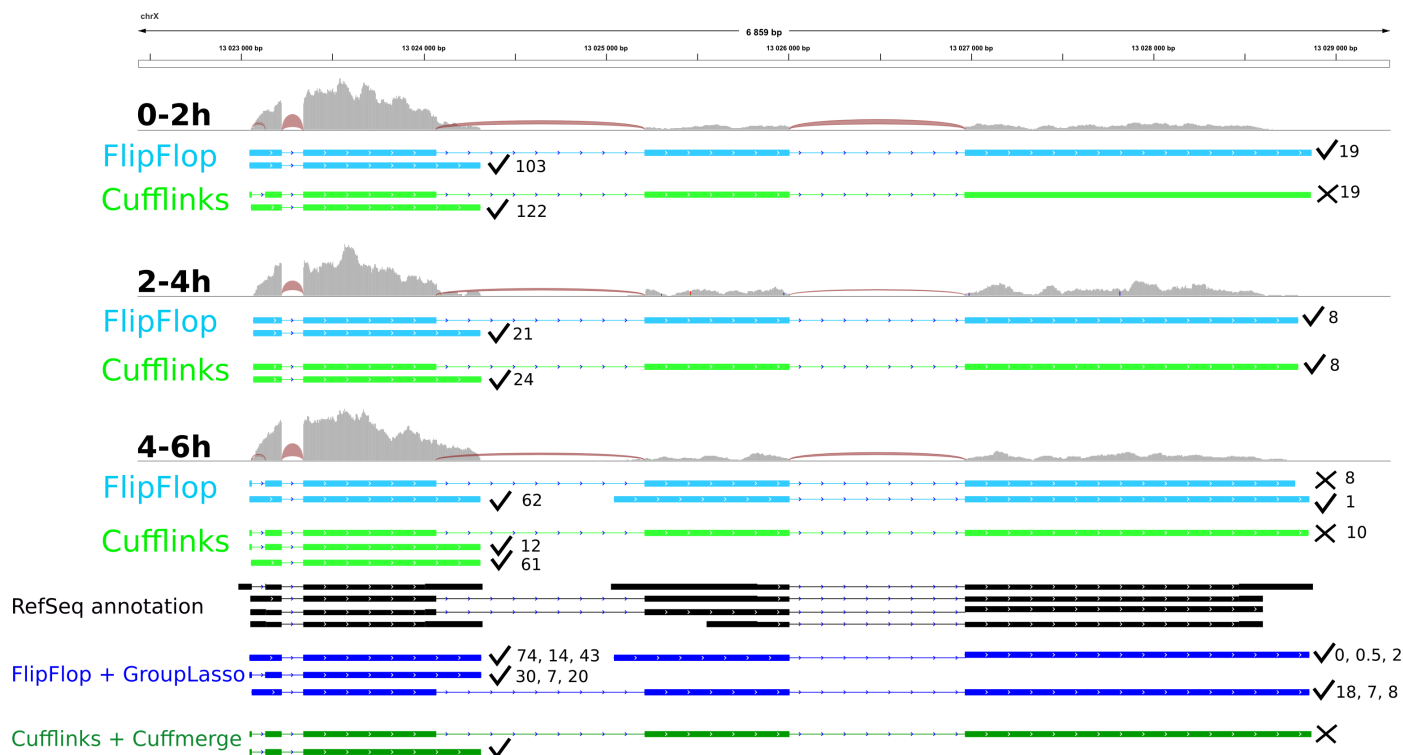


Figure 6: Transcriptome predictions of gene CG15717 from 3 samples of the modENCODE data. Samples name are 0-2h, 2-4h and 4-6h. Each sample track contains the read coverage (light grey) and junction reads (red) as well as FlipFlop predictions (light blue) and Cufflinks predictions (light green). The bottom of the figure displays the RefSeq records (black) and the multi-sample predictions of the group-lasso (dark blue) and of Cufflinks/Cuffmerge (dark green).

Methods	Prediction parameters (with default values)	Optimal values for each number of samples				
		1	2	3	4	5
MiTie	max-num-trans (2)	5	5	10	10	10
	C-exon (10)	29	50	17	50	29
	C-intron (100)	100	20	58	292	171
	C-num-trans (100)	20	20	20	20	34
Cufflinks	min-isoform-fraction (0.10)	0.02	0.03	0.02	0.02	0.02
	pre-mrna-fraction (0.15)	0.08	0.08	0.03	0.03	0.03
	junc-alpha (10^{-3})	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
FlipFlop + Merge	BICcst (50)	10	50	50	85	50
	cutoff (1)	0	1	1	3	3
	delta (10^{-7})	10^{-11}	10^{-11}	10^{-10}	10^{-10}	10^{-11}
FlipFlop + GroupLasso	BICcst (50)	10	29	29	50	50
	cutoff (1)	0	0	0	0	1
	delta (10^{-7})	10^{-11}	10^{-9}	10^{-10}	10^{-10}	10^{-10}

Table D.2: Details on the optimized prediction parameters.

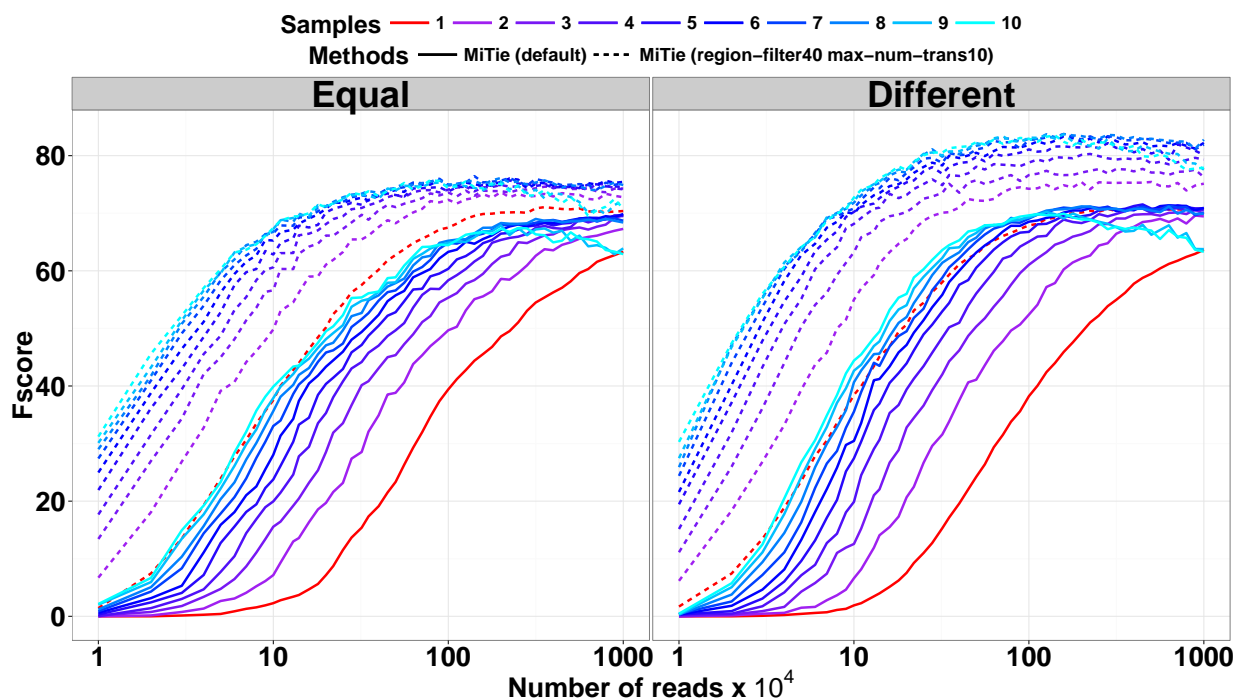


Figure A.1: MiTie results on a first set of human simulations when using default parameters or setting *region-filter* to 40 and *max-num-trans* to 10.

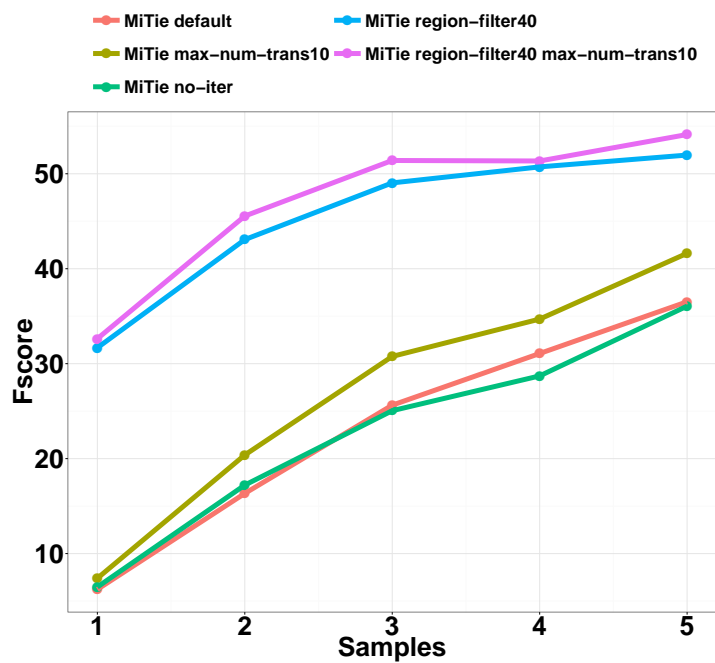


Figure A.2: MiTie results on a second set of human simulations when varying some parameters.

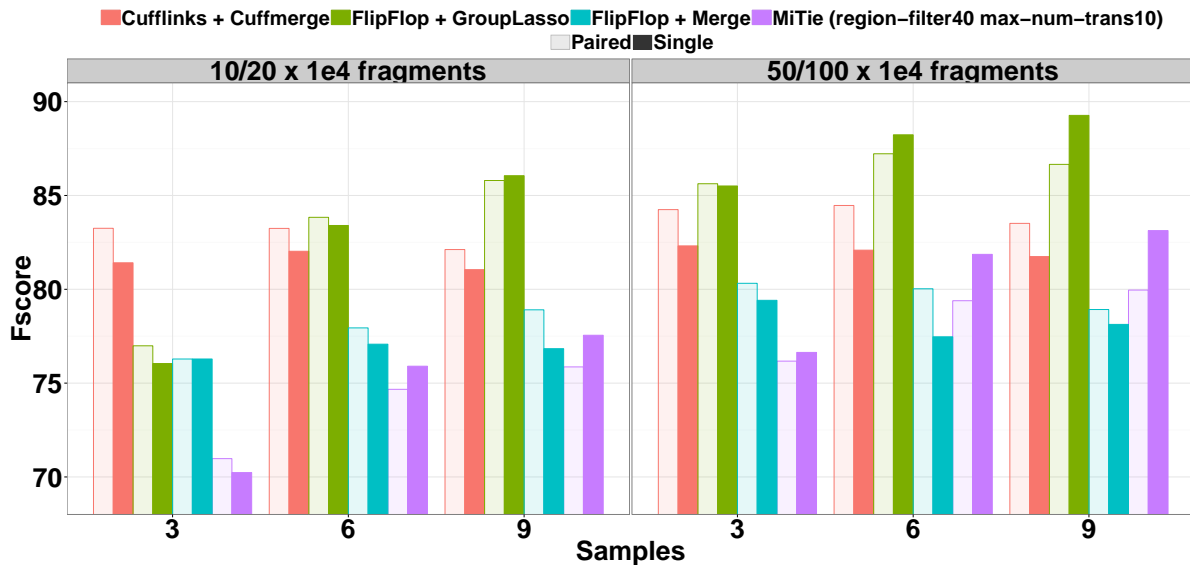


Figure C.1: Simulation using both paired or single-end reads at comparable coverage. The legend 10/20 or 50/100 represents $10^4 \times$ the number of sequenced fragments in the paired-end setting versus the single-end setting (the number of sequenced reads is then equal in the two settings, while the number of sequenced fragments is twice higher in the single-end setting). The read length is fixed to 150bp and the mean fragment size to 350bp in the paired-end setting.

Overall, our FlipFlop+GroupLasso method clearly achieves the best performances in both the paired and single-end settings in the high 50/100 coverage case, and also in the low 10/20 coverage case when using 9 samples.

For Cufflinks+Cuffmerge, the paired-end setting is systematically a bit better than the single-end one, while for both FlipFlop+GroupLasso and MiTie the two settings are either comparable or better in the single-end case.

Sample descriptions	SRA accession names	Total number of reads mapped on the reference transcriptome
0-2h embryos	SRR023659 SRR023755 SRR023671 SRR023663 SRR023747	25 388 344
2-4h embryos	SRR023722 SRR023745 SRR023705 SRR023660	24 541 397
4-6h embryos	SRR023746 SRR023836 SRR023696 SRR023669 SRR035220	46 722 946
6-8h embryos	SRR023691 SRR023732 SRR023654 SRR023668 SRR024217	32 231 644
8-10h embryos	SRR023754 SRR023657 SRR023749 SRR023701 SRR023759 SRR024219 SRR023750	29 544 727

Table E.1: Description of the *D.melanogaster* RNA-seq data from the modENCODE project. Data can be found at the following adress: <http://intermine.modencode.org/query/experiment.do?experiment=Developmental+Time+Course+Transcriptional+Profiling+of+D.+melanogaster+Using+Illumina+poly\%28A\%29\%2B+RNA-Seq>

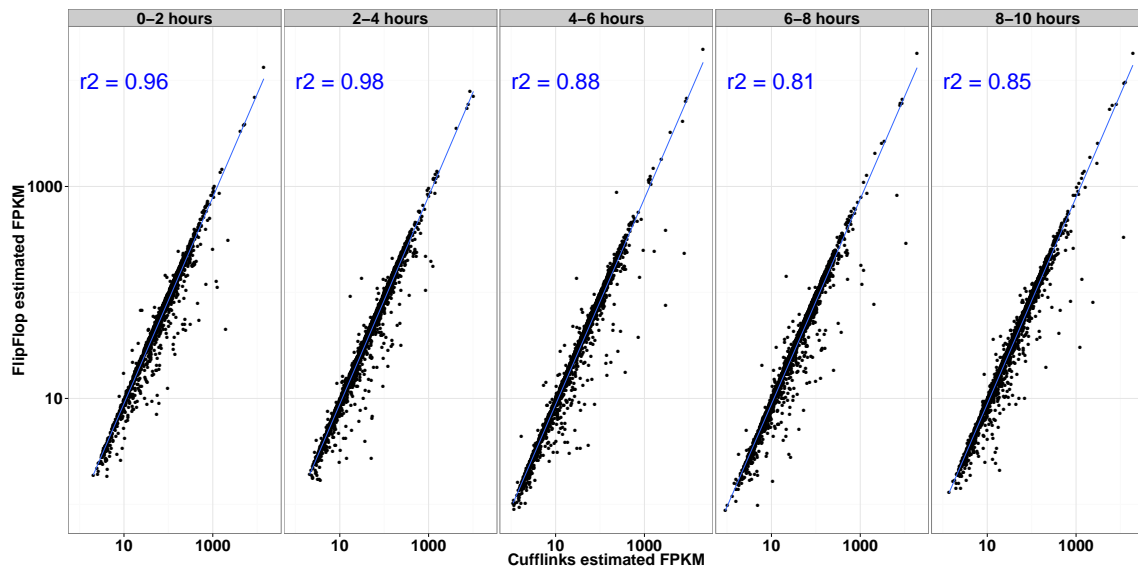


Figure F.1: Abundance estimation on the five samples of *D.melanogaster* RNA-seq data (forward strand). Scatter-plot are between the one-sample FlipFlop method and Cufflinks for genes where both method found the same set of expressed transcripts.

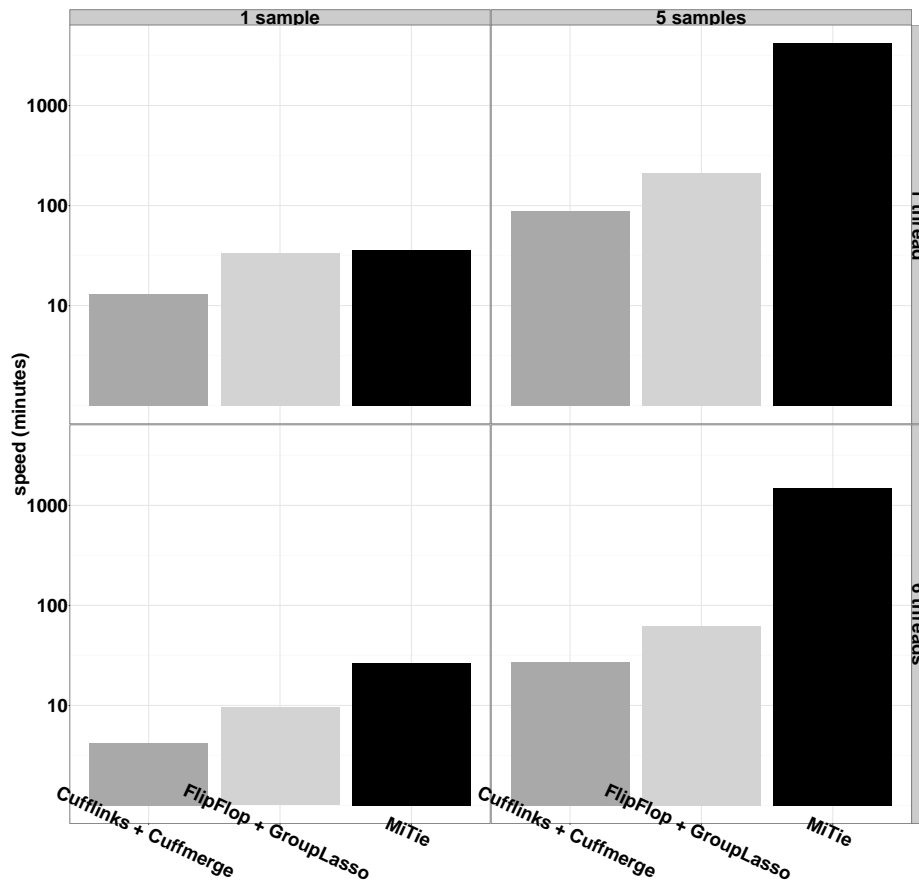


Figure G.1: Running time on the *D.melanogaster* RNA-seq data (forward strand). Each method was run on a 48 CPU machine at 2.2GHz with 256GB of RAM, on either 1 or 6 threads (all tools support multi-threading). MiTie is more than 20 times slower than FlipFlop+GroupLasso when using 5 samples.

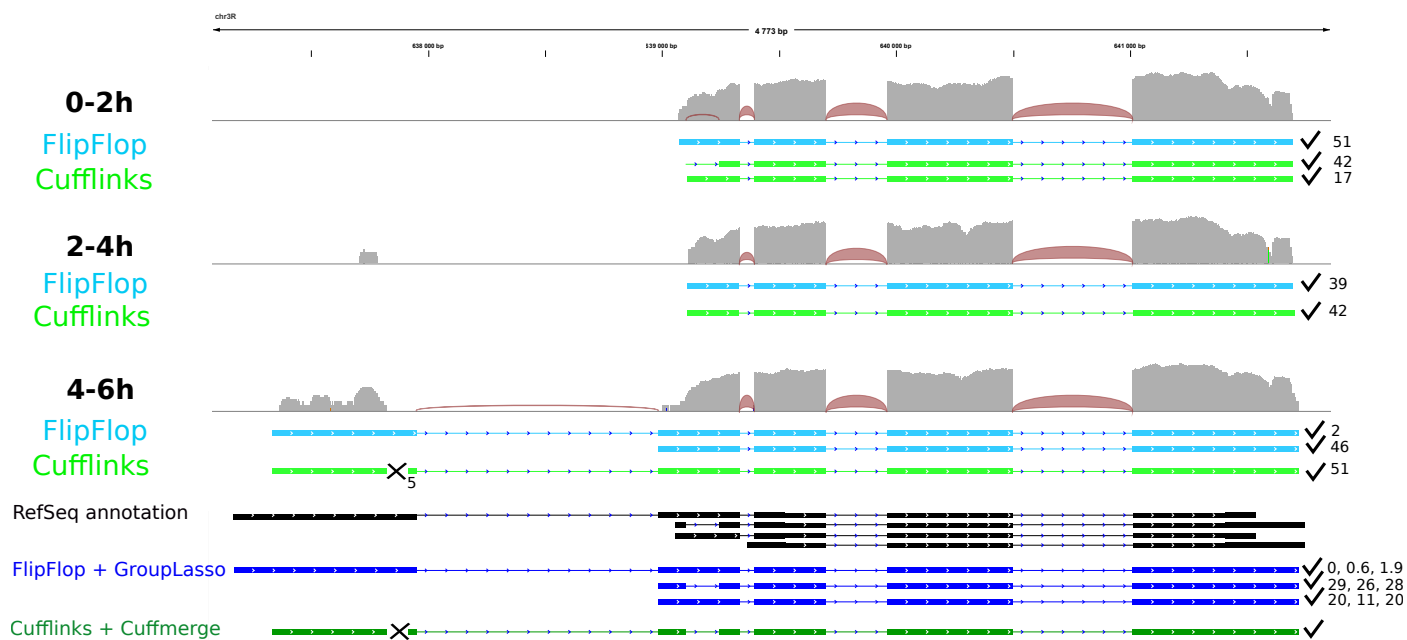


Figure H.1: Transcriptome predictions of gene CG1129 from 3 samples of the modENCODE data. Samples name are 0-2h, 2-4h and 4-6h. Each sample track contains the read coverage (light grey) and junction reads (red) as well as FlipFlop predictions (light blue) and Cufflinks predictions (light green). Here coverage is log-scale. The bottom of the figure displays the RefSeq records (black) and the multi-sample predictions of the group-lasso (dark blue) and of Cufflinks/Cuffmerge (dark green). Symbols ✓ and ✗ indicate if a predicted transcript matches a RefSeq record of not. Estimated abundances in FPKM are given on the right hand side of each transcript.

Figure H.1 illustrates that our group-lasso approach can be more powerful than individual predictions and than the merging strategy of Cuffmerge. Indeed, when using evidences from several samples (both junctions and coverage discrepancies) our approach finds a lowly expressed transcript (that was found in only 1 sample with individual predictions), and two well expressed transcripts, including one that was not previously found with individual predictions. On the other hand, Cufflinks/Cuffmerge is very conservative and only predicts a long transcript that does not explain the variations of coverage from the left to the right part of the gene.