BMC Genomics

PROCEEDINGS Open Access

Accurate inference of isoforms from multiple sample RNA-Seq data

Masruba Tasnim[1], Shining Ma[2,1], Ei-Wen Yang[1], Tao Jiang[1,3*], Wei Li[4,1*]

From The Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015)
HsinChu, Taiwan. 21-23 January 2015

Abstract

Background: RNA-Seq based transcriptome assembly has become a fundamental technique for studying expressed mRNAs (*i.e.*, transcripts or isoforms) in a cell using high-throughput sequencing technologies, and is serving as a basis to analyze the structural and quantitative differences of expressed isoforms between samples. However, the current transcriptome assembly algorithms are not specifically designed to handle large amounts of errors that are inherent in real RNA-Seq datasets, especially those involving multiple samples, making downstream differential analysis applications difficult. On the other hand, multiple sample RNA-Seq datasets may provide more information than single sample datasets that can be utilized to improve the performance of transcriptome assembly and abundance estimation, but such information remains overlooked by the existing assembly tools.

Results: We formulate a computational framework of transcriptome assembly that is capable of handling noisy RNA-Seq reads and multiple sample RNA-Seq datasets efficiently. We show that finding an optimal solution under this framework is an NP-hard problem. Instead, we develop an efficient heuristic algorithm, called Iterative Shortest Path (ISP), based on linear programming (LP) and integer linear programming (ILP). Our preliminary experimental results on both simulated and real datasets and comparison with the existing assembly tools demonstrate that (i) the ISP algorithm is able to assemble transcriptomes with a greatly increased precision while keeping the same level of sensitivity, especially when many samples are involved, and (ii) its assembly results help improve downstream differential analysis. The source code of ISP is freely available at http://alumni.cs.ucr.edu/~liw/isp.html.

Introduction

Transcriptomic research has taken advantage of recent high-throughput sequencing methods, leading to a new experimental protocol, RNA-Seq [1]. A major application of RNA-Seq is transcriptome assembly and isoform (or transcript) abundance estimation, where full-length mRNA transcripts and their expression levels are inferred from RNA-Seq data. Transcriptome assemblies can help analyze both structural and quantitative differences of expressed isoforms between samples. Such (differential) analysis could, for example, lead to the detection of oncogenes that are associated with cancers [2] and splicing variants that are responsible for diseases [3].

If a reference genome is available, transcriptome assembly usually begins by mapping RNA-Seq reads to the reference genome. After that, different algorithms can be used to infer transcripts from mapped reads, including Cufflinks [4,5], IsoInfer [6], IsoLasso [7], SLIDE [8], CLIIQ [9], MITIE [10], *etc.* This *ab initio* approach is different from the *de novo* approach where reference genome is not used (such as AbySS [11], Trinity [12], *etc.*), and is able to take advantage of information provided by the reference genome. As a result, *ab initio* assemblers are able to recover transcripts with a better accuracy and yet demand less computational resource [13]. However, their results critically depend on the quality of the reference genome and mapping software, and they are not specifically designed to handle errors [13], which come

* Correspondence: jiang@cs.ucr.edu; wli@jimmy.harvard.edu
[1]Department of Computer Science and Engineering, University of California, Riverside, Riverside, CA, 92507, USA
[4]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA, 02215, USA
Full list of author information is available at the end of the article

from various sources including unwanted RNA fragments during the library preparation, mapping errors (due to sequencing errors and/or repeats), or "dark matters" from inter-genetic and intron regions [14].

In many RNA-Seq based studies, multiple sample RNA-Seq datasets are available. It is now common for an RNA-Seq project to sequence the whole transcriptomes of samples obtained from multiple replicates, tissues, populations, *etc*. For example, the ENCODE project [15] aims at creating functional element profiles of more than 100 human cell lines, and more than 200 RNA-Seq datasets from various tissues and experimental protocols are available for public use [16]. Other large research projects (including TCGA [17], modENCODE [18], *etc*.) are also producing many multiple sample RNA-Seq data. On one hand, RNA-Seq reads from multiple samples could potentially help assemble transcripts better than reads from only one sample, since the samples can be correlated. On the other hand, transcriptome assembly for multiple samples and subsequent differential analysis are more challenging because (i) multiple sample RNA-Seq data typically contains more noise and (ii) differential analysis is very sensitive to assembly and abundance estimation errors. Therefore, to analyze the structural and quantitative differences of isoforms from multiple samples, a highly accurate transcriptome assembly and abundance estimation tool is necessary.

A straightforward way to assemble transcriptomes for multiple samples is to "merge" all transcripts that are assembled from individual samples as a "universal" set of isoforms, which is then used for downstream applications including abundance estimation and differential analysis. An example of this approach is the "Cuffmerge" program in the Cufflinks software package [5]. However, as more samples are sequenced, errors from individual assemblies are likely to accumulate, which could seriously affect the isoform abundance estimation and result in unreliable (or even misleading) differential analysis results.

In this paper, we present a new framework for *ab initio* transcriptome assembly that is able to handle noisy RNA-Seq reads and multiple sample RNA-Seq datasets effectively. Instead of assembling transcripts separately for each sample and merging them together, our framework reconstructs transcripts directly from multiple samples. In fact, it takes advantage of the extra information contained in paired-end reads and in multiple sample RNA-Seq datasets (*e.g.*, correlation among the samples). We show that finding an optimal solution under this framework is NP-hard, and develop a heuristic algorithm, called ISP (for *Iterative Shortest Path*), to reconstruct isoforms efficiently under the framework. For a given gene, ISP solves either a linear programming (LP) or an integer linear programming (ILP) problem iteratively on a weighted graph derived from the input multiple sample RNA-Seq dataset. Our preliminary experimental results on both simulated and real datasets demonstrate that (i) ISP is able to assemble transcriptomes with high precision and sensitivity, especially when many samples are involved, and (ii) the assembly results of ISP help improve downstream differential analysis.

Methods

Multiple Sample Connectivity Graph (MSCG)

A set of RNA-Seq reads from F different samples are first mapped independently to the reference genome using a splice junction detection tool such as Tophat [19], SpliceMap [20], *etc*. The mapped reads are then clustered into genes, and the exon-intron boundary information for each may be derived from either its junction reads or existing annotations such as NCBI RefSeq [21] or UCSC known isoforms [22]. Based on this information, the sequence of a gene can be split into different *expressed segments* (or simply *segments*) [6], where a segment is a continuous region in the reference genome uninterrupted by any splicing events (or exon-intron boundaries).

Several transcriptome assemblers [7,9,23] use the *Connectivity Graph (CG)* to represent the splicing connections between segments or bases on single sample RNA-Seq data. Similarly, for multiple sample RNA-Seq data, we construct a *multiple sample connectivity graph (MSCG)* $G = (V, E)$ based on F sets of mapped RNA-Seq reads as follows. $V = \{s, t\} \cup \{v_i | 1 \leq i \leq M\}$ where $v_1, \ldots v_M$ represents the M segments contained in a gene. $(v_i, v_j) \in E$ if there is at least one read from the F samples joining both segments i and j. Also, $(s, v_i) \in E$, $(v_i, t) \in E$, $1 \leq i \leq M$ (see Figure 1(A)).

For simplicity, s is assigned number 0, t is assigned number M +1, and the vertices $v_1, \ldots v_M$ are all assigned numbers 1 through M. Thus, a vertex in V or an edge in E can be represented as a tuple (i, j) with $0 \leq i, j \leq M + 1$. For example, $(0, i) = (s, v_i)$ and $(i, M + 1) = (v_i, t)$ for $1 \leq i \leq M$. Similarly, $(i, j) \in E$ if $(v_i, v_j) \in E$ for $i \neq j$, and $(i, j) \in V$ if $i = j$ $(1 \leq i, j \leq M)$.

Isoform discovery on MSCG

In MSCG, every path P from s to t represents a possible isoform of the gene. We further assign weights to each vertex (or edge) in MSCG to represent the probability that the corresponding segment (or junction) is included in an isoform; the higher the weights, the lower probability that they are included. As a result, a shortest path (minimum weight path) represents the most possible isoform of the gene. Furthermore, we make use of paired-end read information (if any) and expression correlation between segments among the samples that could help the reconstruction of isoforms.

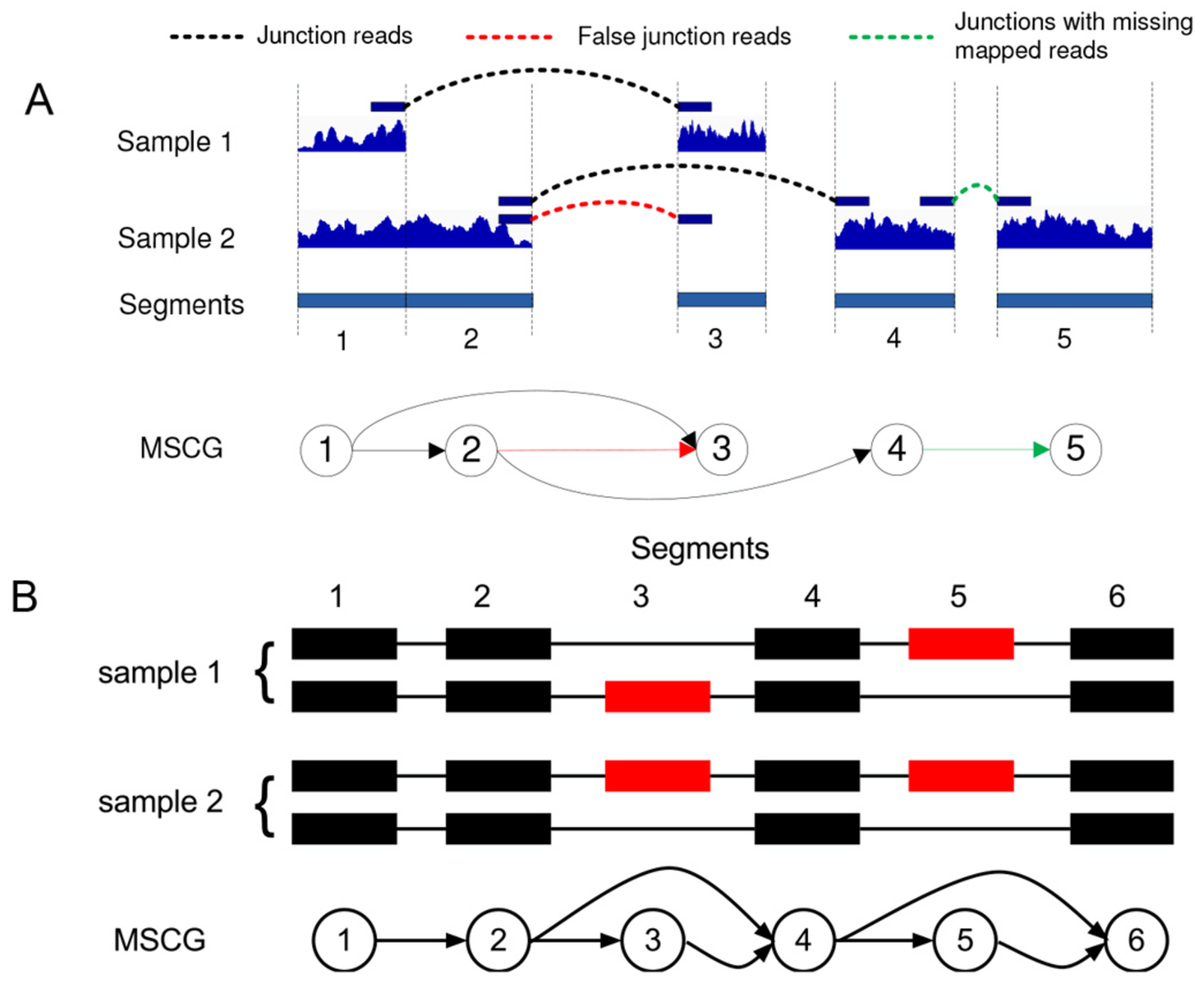

Figure 1 Illustration of MSCG. (A) An example of MSCG from two samples. For simplicity, the vertices *s* and *t* as well as the corresponding edges connecting *s* or *t* are not shown. (B) An example of two samples with coexpressed segments and exclusively expressed segments (see Section). Segments 3 and 5 (marked in read) are exclusively expressed in two isoforms of sample 1, but are coexpressed in two isoforms of sample 2. Both samples generate the same MSCG.

Our approaches to assigning weights and utilizing information from paired-end reads and segment expression correlations are described in detail below.

Weight assignment

A weight $w_{i,j}$ is assigned for each edge $(i, j) \in E$ (if $i \neq j$) and each vertex $v_i \in V$ (if $i = j$) to reflect the likelihood that the corresponding segment (or junction) is "problematic": a higher weight is assigned if the segment (or junction) is more likely the product of some incorrectly mapped reads. Notice that $w_{i,j}$ may be either positive (considered as "cost") or negative (considered as "reward"). For simplicity, the weight of a path P is represented as), $\sum_{(i,j)\in P} w_{i,j}$, which is the sum of all weights of the vertices and edges included on P.

Assume that the same number of reads are generated from F samples. For every vertex $v_i \in V \backslash \{s, t\}$, we assign $w_{i,i} = -\log(d_i + 1)$, where d_i is the average read density of segment i in the F samples:

$$d_i = \sum_{k=1}^{F} C_i^k/(l_i - L + 1) \tag{1}$$

Here, C_i^k is the number of reads mapped to segment i from the kth RNA-Seq sample, l_i is the length of the segment i and L is the read length. Since we will look for a shortest path, paths going through segments with high densities are preferred.

Because noisy junctions may result in incorrect assembly results, a higher positive cost is assigned for junction edges that are more likely to be problematic. For every edge $(i, j) \in E$, where $1 \le i, j \le M$ and $i \ne j$, we set

$$w_{i,j} = -\log P(i,j) = -\log(\frac{d_{i,j}}{\sum_h d_{i,h}} \frac{d_{i,j}}{\sum_h d_{h,j}}) \tag{2}$$

where $1 \le h \le M$, $P(i, j)$ represents the probability that the junction between segments i and j is included in an isoform, and $d_{i,j}$ is the average read density of edge $(v_i, v_j) \in G$:

$$d_{i,j} = \sum_{k=1}^{F} C_{i,j}^{k}/(L-1) \tag{3}$$

where $C_{i,j}^{k}$ is the number of reads mapped to the corresponding junction from the kth RNA-Seq sample.

For every edge $(0, i) \in E$, we "inactivate" it by setting $w_{0,i}$ to infinity if v_i can be reached from another vertex v_j in the MSCG G; otherwise, we set $w_{0,i} = 0$. Similarly for edge $(i, M + 1) \in E$, $w_{i,M+1}$ is assigned infinity if there is an edge from v_i to another vertex v_j in the MSCG G, or 0 otherwise.

Incorporating paired-end read information

Paired-end RNA-Seq reads provide more information than single-end reads in transcriptome assembly, since both ends of a paired-end read come from the same RNA (or cDNA) fragment. To incorporate paired-end read information into our framework, we try to find a path in MSCG to simultaneously minimize the cost of the path and maximize the number of paired-end reads that are *compatible* with the isoform represented by the path. Reads that are compatible with an isoform are the reads that are possibly generated from the isoform. If a read is compatible with an isoform, the splicing patterns implied by the read and the isoform must be identical. More precisely, a single-end read b containing k segments can be represented as a vector $b = (b_1, b_2,..., b_k)$, where $1 \le b_1 < ... < b_k \le M$ are the segments included in b. An isoform I (or a path P) that b is compatible with must include all the segments $b_1,..., b_k$, and must not include any other segment between b_1 and b_k. A paired-end read $p = (b, b')$ is compatible with I if and only if both b and b' are compatible with I.

For each paired-end read $p = (b, b')$, where $b = (b_1,... b_k)$ and $b' = (b'_1, \cdots b'_k)$, we define the set of "inclusion segments" IS_p and "exclusion segments" ES_p as follows:

$$IS_p = b \cup b' \tag{4}$$

$$ES_p = \{i : b_1 < i < b_k \text{ or } b'_1 < i < b'_k, i \notin IS_p\} \tag{5}$$

Intuitively, IS_p (and ES_p) represents the set of segments that an isoform I must (and must not) include, based on the information of p. For example, if $p = ((b_1, b_3), (b_5, b_7))$, then $IS_p = \{b_1, b_3, b_5, b_7\}$ (as they are included in p), and $ES_p = \{b_2, b_6\}$ (as they are spliced out in p). For each paired-end read p, we define a binary variable $q_{p\sim P} \in \{0, 1\}$ to indicate whether p is compatible with a path P implied in the solution. Given a set of paired-end reads R, maximizing the number of compatible reads with P is equivalent to maximizing), $\sum_{p \in R} q_{p\sim P}$. For each gene, paired-end reads that are mapped to only one segment (*i.e.*, $|IS_p| = 1$) are excluded from R, since these reads do not provide any useful information in the assembly.

Resolving ambiguities using Jensen-Shannon metric

In a complicated gene model, an MSCG may give rise to several sets of isoforms due to the existence of segments that introduce ambiguities (named as "uncertain" segments). For example, the MSCG in Figure 1(B) has two edges at each end of segment 4 due to the two uncertain segments, segments 3 and 5. Different combinations of these pairs of edges would lead to two possible sets of isoforms. Paired-end reads can be used to resolve such ambiguity (as in [23]), but it only works if there are paired-end reads mapped to uncertain segments. In [5], isoforms are decomposed such that the expression levels of the segments in one isoform are similar, but this strategy does not consider positional biases [24] and is applied only to a single sample.

In this work, we use *Jensen-Shannon metric* (or JS metric) to resolve the ambiguity of uncertain segments. JS metric measures the similarity of the expression patterns between samples and was used to analyze differential alternative splicing events [5]. It is defined as the square root of the *Jensen-Shannon divergence* [25]:

$$JS(i,j) = \left(H(\frac{p_i + p_j}{2}) - \frac{H(p_i) + H(p_j)}{2}\right)^{1/2} \tag{6}$$

where $H(x)$ stands for the entropy of the probability distribution x and p_i is the distribution of segment i among the samples. The latter is calculated based on the read density of segment i (defined in Equation (1)) over all F samples.

If the JS metric of the expression levels of two uncertain segments is low (which means both segments are positively correlated), then both segments are likely to be included on the same isoform (termed "coexpressed segments", see Figure 1(B) for an example). Otherwise if the JS metric is high, they are likely to appear in different isoforms (termed "exclusively expressed segments"). To determine whether two uncertain segments are coexpressed or exclusively expressed, we randomly permute the expression of each segment in a gene 1000 times, and calculate the "background" JS metric distribution P_{bg}. For a given false-discovery rate (FDR) β% (controlled by the user; the default is 5%), segments i and j are considered coexpressed (or exclusively expressed) if

$JS(i, j)$ is located in the lowest (or highest) β% of P_{bg}, respectively. For coexpressed segments i and j, we add some "pseudo" paired-end reads p_c spanning segments i and j (*i.e.*, $IS_{pc} = \{i, j\}$) to the read set R. These reads will encourage our algorithm (*i.e.*, ISP) to prefer paths that include both segments i and j. Similarly for exclusively expressed segments i and j, paired-end reads p_e with $IS_{pe} = \{i\}$ and $ES_{pe} = \{j\}$ are added to R.

The objective function and complexity of the problem

Using the notations defined in previous sections, given an MSCG G and a single-end and/or paired-end read set R, our objective function is to find a path P from s to t in the MSCG to maximize

$$\sum_{(i,j)\in P} -w_{i,j} + \sum_{p\in R'} \alpha_p q_{p\sim P} \tag{7}$$

where the set R' includes all "pseudo" reads and excludes all reads $p \in R$ with $|IS_p| = 1$. Here, $\alpha_p > 0$ is a user-defined parameter and should be smaller for organisms with simple splicing patterns (like fruit fly or warm) and relatively larger for organisms with more complicated splicing patterns (like human or mouse). For the convenience of presentation, we will refer to this problem as a (constrained) *shortest path* problem on G, because when $\alpha = 0$, the problem reduces to finding the minimum weight path (or shortest path) from s to t.

Unfortunately, it is hard to find a path to maximize Equation (7), since we can show that the corresponding decision problem is NP-complete even when $w_{i,j} = 0$ and $\alpha_p = 1$.

Theorem: The following decision problem is NP-complete:

Input: An MSCG $G = (V, E)$ and a set of mapped paired-end reads R; an integer k.

Question: Is there a path P in G such that), $\sum_{p\in R} q_{p\sim P} \geq k$?

Proof : The theorem can be proven by a straightforward reduction from the wellknown CLIQUE problem. The reduction is presented in Additional file 1. ■

An efficient heuristic algorithm to identify expressed isoforms in multiple samples

The ILP and LP approaches for finding an optimal path

In this section, we present two different approaches to find a path on the MSCG G maximizing Equation (7). First, a binary variable $f(i, j) \in \{0, 1\}$ is introduced to indicate whether each vertex (or edge) in G is included in a path P. The following ILP problem is formulated to find a path maximizing Equation (7):

$$max \quad \sum_{0\leq i,j\leq M+1} -w_{i,j}f(i,j) + \sum_{p\in R'} \alpha_p q_p \tag{8}$$

$$\text{s.t} \sum_{i=1}^{M} f(0,i) = 1,\ 1 \leq i \leq M \tag{9}$$

$$\sum_{0\leq k\leq M} f(i,M) = f(i,i) = \sum_{0\leq k\leq M} f(k,i),\ 1 \leq i \leq M \tag{10}$$

$$q_p \leq f(i,i),\ i \in IS_p \tag{11}$$

$$q_p \leq 1 - f(i,i),\ i \in ES_p \tag{12}$$

$$f(i,j), q_p \in \{0,1\},\ 0 \leq i,j \leq M+1 \tag{13}$$

Equations (9)-(10) are constraints ensuring that the final solution represents a path (and thus an isoform) from s to t, while Equation (11)-(12) guarantee that $q_p = 1$ if and only if the path P is compatible with paired-end read p. Solving the above ILP problem may be time-consuming since the number of variables may be large for some genes. Instead, we could relax the binary constraints in Equation (13) as follows, which turns the problem into an LP problem:

$$0 \leq q_p, f(i,j) \leq 1 \tag{14}$$

Ideally, the solution to the above LP problem is integral (*i.e.*, q_p, $f(i, j) \in \{0, 1\}$), which would represent a path from s to t. However, in some cases (for about 0.1% of the genes in our simulated and real data experiments), the LP problem may not lead to an integral solution. For these genes, we can solve the corresponding ILP problem instead. We use GNU Linear Programming Kit (GLPK, [26]) to solve both the ILP and LP problems.

The Iterative Shortest Path algorithm

A gene may have multiple isoforms expressed in the samples, but only one isoform is extracted by solving the above LP/ILP problem. To recover more expressed isoforms of the gene, we apply the "weight-decay" strategy [27] to modify the weights in the graph G and iterate the algorithm several times. In each iteration, the weights are adjusted to encourage the algorithm to look for an isoform different from all previously found isoforms. The details of this ISP algorithm are described in Additional file 1.

Results

Simulation results

We simulated RNA-Seq reads and evaluated the performance of different algorithms following the method described in [7,28]. Briefly, we used UCSC known human (and mouse) transcripts [22] to simulate single-end and paired-end reads and evaluate the *sensitivity* and *precision* of different assemblers on noisy RNA-Seq data and multiple samples. Following the definition in [6] and [7], two transcripts are matched if their exon coordinates are identical except the start of the first exon and the end of the last exon. If K of M predicted transcripts match K of N known transcripts, then the sensitivity and precision are defined as K/N and K/M , respectively. We added two different types of noisy reads in the simulation to capture noise in real RNA-Seq data: *noisy junction reads* and *noisy intron reads*. Noisy junction reads are generated by randomly shifting the splicing positions of some normal junction reads by 1 to 3 bases. These reads are added since in reality, splicing regulators may shift the splice site a few bases to the proximal or distal intron boundaries [29,30]. Noisy intron reads are reads coming randomly from the intron regions of a transcript. They are added since it has been observed that a fair amount of reads come from intronic regions in practice, possibly due to intron retention, non-coding RNAs or other unknown mechanisms [14].

We compared the performance of ISP with two existing assembly algorithms for multiple samples, Cufflinks/Cuffmerge [4,5] and MITIE [10]. Cufflinks and Cuffmerge are algorithms incorporated in the Cufflinks software package. For multiple RNA-Seq samples, Cufflinks first constructs a set of isoforms from multiple samples, followed by Cuffmerge merging assembly results from each individual sample. MITIE, on the other hand, constructs isoform structures by solving a mixed integer programming problem defined on multiple samples. CLIIQ [9] is another recent tool for assembling isoforms from multiple sample RNA-Seq data based on integer programming. However, we have had great difficulty in getting CLIIQ to run on our servers (even with the help of the authors of CLIIQ). Hence, we will make a comparison with CLIIQ indirectly and present the comparison results in Additional file 1.

The effect of noisy RNA-Seq reads on single sample data

We added different amounts of noisy reads of both types to a single sample RNA-Seq dataset, and the sensitivity and precision of ISP and Cufflinks are presented in Figure 2. Here, a total of 80 million single-end or paired-end reads are used, and "error rate" shows the percentage of the randomly shifted junction reads and noisy intron reads added to the dataset. When more errors are added, both programs keep the same level of sensitivity (about 10%), but the precision of both programs gradually drops. Compared with Cufflinks, ISP is less affected by the errors, showing that ISP is able to handle read errors better on single sample RNA-Seq data.

It is worth noting that when the simulated RNA-Seq data is error-free, mapping tools may still result in incorrectly mapped reads and thus the input to Cufflinks/ISP could still be noisy. Also, the low sensitivity of both programs is due to the fact that many of the transcripts are assigned very low expression levels (or they are not expressed at all) based on the log-normal model [31]. These transcripts with few (or no) mapped reads decrease the value of sensitivity.

Assembly for multiple sample RNA-Seq data

To compare the performance of these algorithms on multiple sample RNA-Seq data, we generated six RNA-Seq datasets with different numbers of samples and evaluate the sensitivity and precision of the programs. For each dataset, the expression level of an isoform is independently assigned and 10% noisy reads are added as errors. To reconstruct all isoforms from multiple samples, a straightforward algorithm is to merge the RNA-Seq reads from all samples together and apply a transcriptome assembly tool (such as Cufflinks) designed for single sample RNA-Seq data. As a comparison to ISP and Cuffmerge, we also tested Cufflinks and ISP on pooled data where RNA-Seq reads from all samples are merged together.

Figure 3(A-B) shows both sensitivity and precision of the four programs on different numbers of samples. When only one sample is considered, the sensitivity of all programs is the same. As more samples are added, more transcripts are correctly predicted, and both ISP and Cuffmerge achieve similar improvements of sensitivity on six samples. As for the precision, ISP has a clear advantage, maintaining 40% to 60% higher values than Cuffmerge, and 60% to 80% higher values than Cufflinks. The increasing trend of sensitivity and precision for both ISP and Cuffmerge shows that both programs are able to take advantage of the existence of multiple samples and improve their sensitivity and precision simultaneously. Instead, the precision of Cufflinks and ISP on the pooled data (denoted as Cufflinks and ISPpool in the figure) drops slightly while their sensitivity falls behind ISP and Cuffmerge. This is because as reads from more samples are merged, the detectable splicing patterns become more complicated. Although more isoforms can be discovered (thus improving the sensitivity), many incorrect isoforms are also predicted (thus hurting the precision) because of the increased difficulty in dealing with complex splicing patterns. Therefore, the straightforward approach for dealing with multiple samples is not a good way to treat multiple sample RNA-Seq data.

MITIE [10] is a recently published algorithm that assembles transcripts from multiple samples. Since MITIE uses mixed linear programming to infer isoforms, it requires very long execution time and large memory space to process human RNA-Seq datasets. For practical considerations, we compared MITIE with ISP and Cuffmerge on RNA-Seq samples that were simulated from 500 randomly selected genes on human chromosome 1, including 1206 annotated transcripts (or 2.41 transcripts/gene). Figure 3(C-D) shows the sensitivity

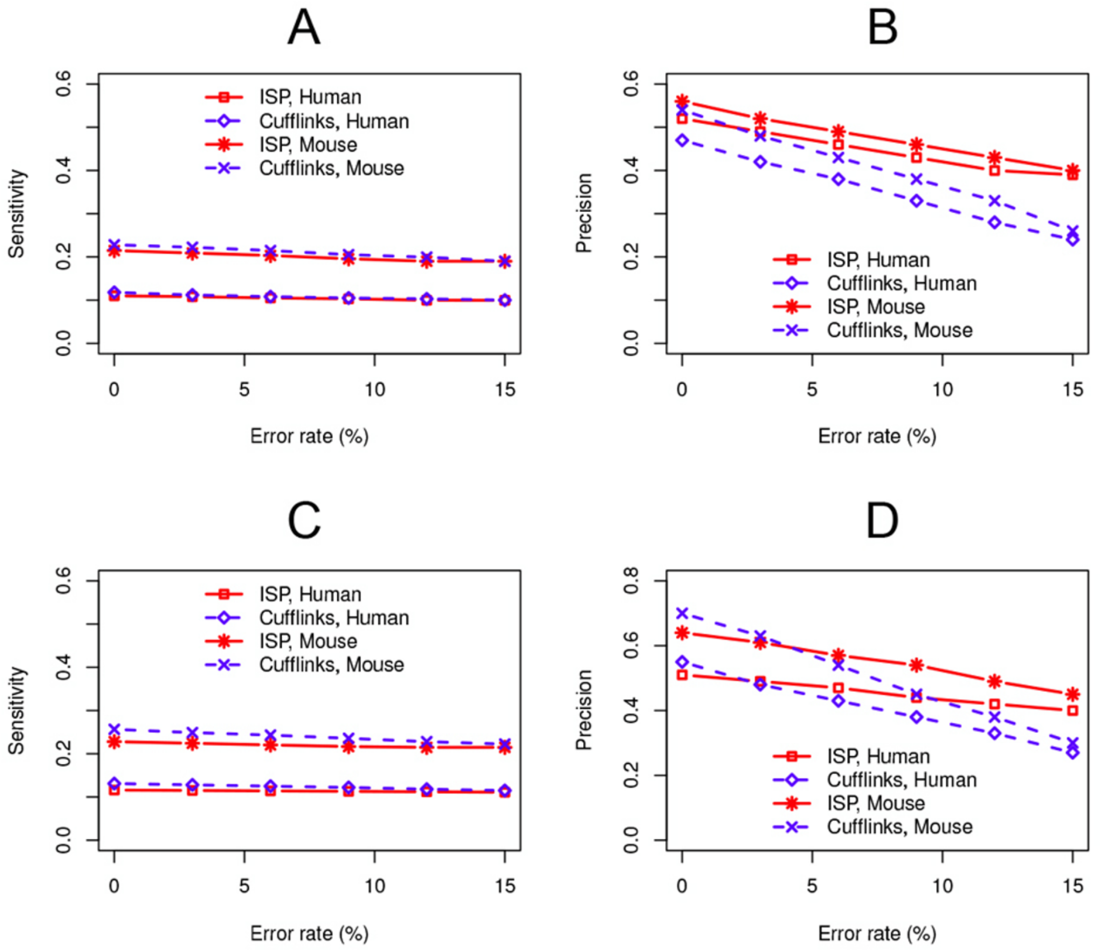

Figure 2 The performance of ISP and Cufflinks on different error rates. (A-B) The sensitivity and precision of ISP and Cufflinks on a single RNA-Seq sample with various error rates using single-end reads. Errors come from either noisy junction reads or noisy intron reads. (C-D) are the corresponding sensitivity and precision on paired-end reads.

and precision of different programs on different numbers of samples. Both ISP and Cuffmerge achieve similar sensitivity and precision in these tests with smaller samples, compared with Figure 3(A-B) using the whole transcriptome for the simulation. MITIE has much lower levels of sensitivity and precision (below 12%), and both values do not increase as more samples are used.

Transcriptome assembly and differential analysis

In a typical differential analysis, we are interested in finding and ranking genes (or isoforms) that are differentially expressed between two samples (or two groups of samples). Since isoforms assembled from individual samples may be different, it is necessary to construct a "universal" set of isoforms from all samples, from which the expression level estimation and statistical analysis can be performed. For example, Cufflinks includes a set of programs for differential analysis, and all of them are based on merging isoforms from individual assemblies (using Cuffmerge).

We are interested in the effect of multiple sample transcriptome assembly on differential analysis. We simulated two RNA-Seq datasets and generate a set of isoforms for both samples by running (i) ISP and (ii) Cufflinks followed by Cuffmerge. To avoid using different expression level estimation methods preferred by both methods, we used Cuffdiff 2 [5], the differential expression analysis tool in Cufflinks package, for expression level estimation and differential analysis (including fold change calculation and statistical evaluation) after running ISP and Cuffmerge.

We selected different numbers of isoforms that show the greatest changes of expression levels. Figure 4 shows the percentage of isoforms that match UCSC human known genes, and the percentage of the matched ones that have correct fold change estimations (defined as estimated fold changes within the [-2,+2] range of their corresponding true fold changes). We showed the trends

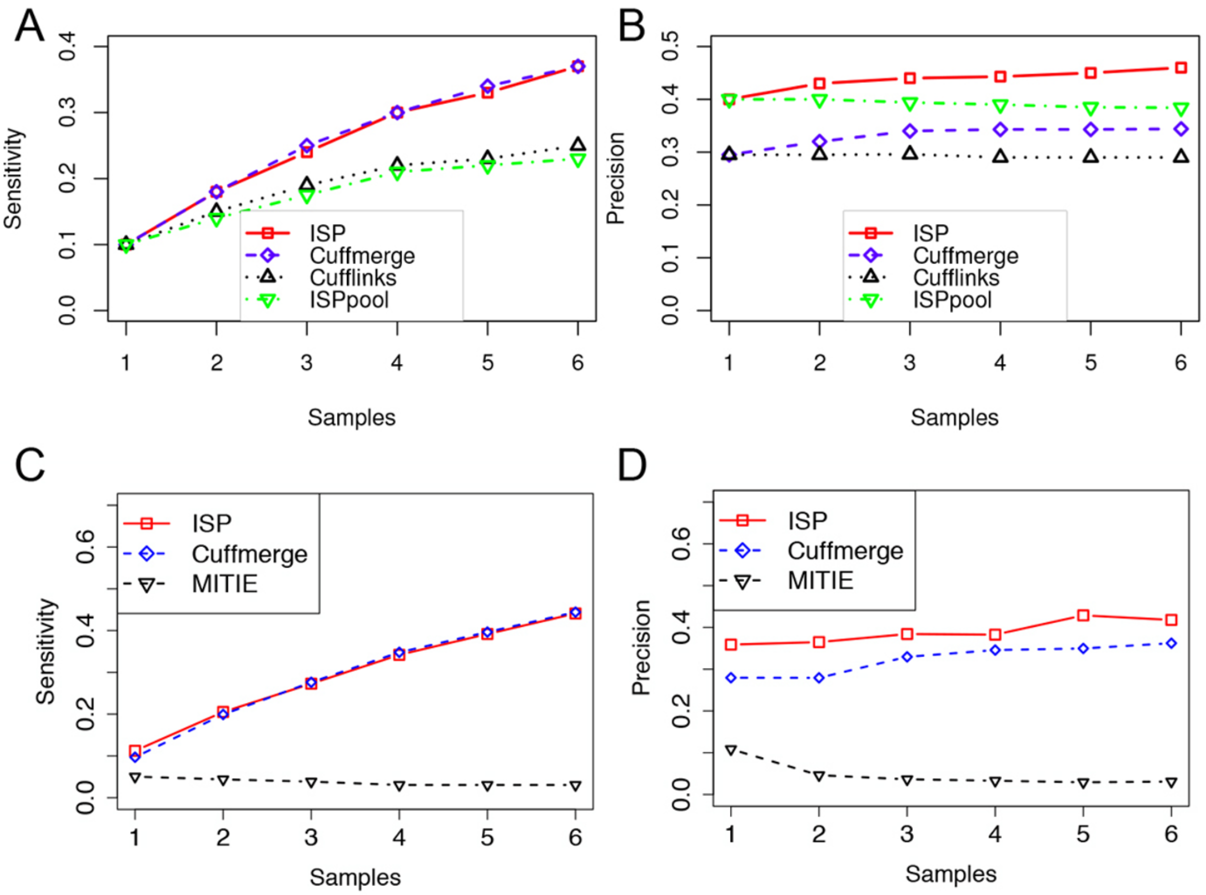

Figure 3 The performance of different algorithms on multiple samples. (A-B) The sensitivity and precision of ISP and Cuffmerge on multiple samples, and on the pooled data (denoted as ISPpool and Cufflinks). The pooled data were generated by merging reads from all samples. (C-D) The sensitivity and precision of ISP, Cuffmerge and MITIE on 500 randomly selected genes from human chromosome 1.

as we decrease the number of selected isoforms, since we usually prefer finding fewer isoforms with higher expression changes between samples.

ISP is able to find a larger number of matched (*i.e.*, true) isoforms than Cuffmerge when more than 10 isoforms are selected. This is consistent with the previous experiments showing a higher precision of ISP than Cuffmerge. Furthermore, ISP outputs more isoforms with correct fold change estimations. With different numbers of top ranked isoforms selected, 74%-90% have their fold changes correctly identified, which is higher than Cuffmerge (58%-82%). Because we use the same algorithm (Cuffdiff 2) for expression level estimation and differential analysis, we suspect that the low precision of Cuffmerge assembly led to the low accuracy in expression level estimation, hence reducing its performance in fold change estimation.

Real RNA-Seq data results

To compare the performance of the algorithms on real RNA-Seq data, we used the public RNA-Seq datasets of 7 cancer cell lines downloaded from the ENCODE project [32]. These cell lines (GM12878, H1-hESC, K562, HeLa-S3, HepG2, HUVEC, NHEK; NCBI GEO accession code: GSE23316) include normal and cancer cells of different tissues, and are the major cell models extensively used in biological and biomedical research [16].

Transcriptome assembly results

It is difficult to measure exactly which isoform is expressed in real RNA-Seq data since the current experimental techniques limit the ability to detect full-length transcripts efficiently. Instead, we treat all UCSC known transcripts as “canonical” isoforms and calculate both sensitivity and precision with respect to these isoforms, the same as in the simulation experiments.

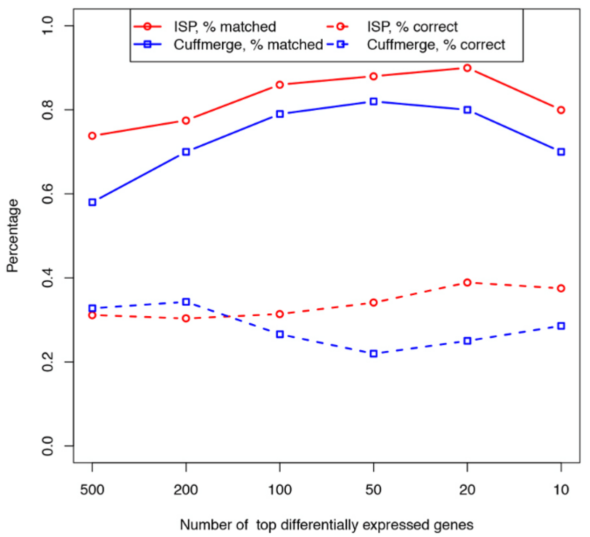

Figure 4 The effect of ISP and Cuffmerge on differential expression analysis. Here, different numbers of the most differentially expressed isoforms are considered, and the percentage of "matched" and "correct" isoforms are calculated. "% matched" is the ratio of these isoforms matching UCSC known isoforms, and "% correct" is the ratio of the matched isoforms that have correct fold change estimations.

Figure 5(A) shows the numbers of predicted isoforms, together with the numbers of matched UCSC known transcripts by using different numbers of samples. For a single sample, the number of isoforms predicted by Cuffmerge is over 60, 000, which is approximately twice as much as ISP. As more RNA-Seq samples are added, more transcripts are merged by Cuffmerge, and this number reaches 150, 000 (over 100% growth) when all seven samples are included. In contrast, ISP shows a moderate increase, with only 40% more predicted isoforms for seven samples compared to using only one sample. However, the numbers of matched UCSC known transcripts remain roughly the same for both programs, with ISP achieving over 90% of the number attained by Cuffmerge. This illustrates that ISP is able to keep a high precision while sacrificing sensitivity a little when the number of samples increases. An example transcriptome assembly results by ISP and Cuffmerge on some ENCODE data can be found in Section 4 of Additional file 1.

Cuffmerge predictions include a large number of single-exon transcripts that do not match any UCSC known transcripts. To study the effect of multiple samples on the inference of multi-exon isoforms, we exclude these single-exon transcripts and calculate the precision for isoforms grouped by their numbers of exons (see Figure 5(B-E)). ISP shows a higher precision than Cuffmerge on all multi-exon isoforms, and when all seven samples are used, its precision is almost doubled compared to Cuffmerge (see Figure 5(B)) for isoforms containing 5-10 exons. For isoforms with more than 10 exons, the difference between the two algorithms becomes smaller, but ISP still maintains a 70% higher precision than Cuffmerge (Figure 5(D-E)). Isoforms with more exons are difficult to assemble since more errors may occur around splice junctions (see Section 5 of Additional file 1). As a result, the high precision of ISP may be attributed to its ability to handle noise effectively and to use information from multiple samples.

We also compared ISP, Cuffmerge and MITIE on RNA-Seq reads from 500 randomly selected genes from human chromosome 1, similar to the comparisons in the simulation tests. The test was performed on a linux cluster node with a 4-core 2.50GHz CPU and 16G memory. Figure 5(F-H) show the sensitivity, precision and running time of the three algorithms on different numbers of samples. In contrast to the simulation tests, MITIE has a much better performance, achieving similar levels of precision with ISP and Cuffmerge when more than two samples are used. However, the sensitivity of MITIE is still lower than those of ISP and Cuffmerge, and it takes much longer time to run, especially when more samples are used (Figure 5(H)). For example, MITIE takes 163.7X longer than ISP (and 92.4X longer than Cuffmerge) to run when 7 samples are used, making it difficult to use on datasets with a large number of samples, especially when the large transcriptomes such as human or mouse are being studied.

Differential analysis

The expression profiles of some ENCODE cell lines have been measured by both RNA-Seq and Affymetrix Human Exon 1.0 ST Array (NCBI GEO accession code: GSE19090). To validate the differential analysis results from RNA-Seq reads, transcripts are assembled from two cell lines (GM12878 and K562) that have corresponding microarray data, and their expression level changes are compared with the microarray measurements. An Affymetrix Human Exon Array uses "probesets" (*i.e.*, sets of probes) to measure the expression levels of exons. To calculate the expression levels of transcripts, we only keep probesets whose measured exons correspond to only one RefSeq transcript (called "unique" probesets). For those RefSeq transcripts that include at least one such unique probeset, their expression levels are calculated by averaging the measurements of all unique probesets. As in the simulation experiments, we used Cuffdiff 2 for expression level estimation and differential analysis.

ISP and Cuffmerge are able to identify 4468 and 4627 transcripts that have corresponding expression level estimations in the microarray data, respectively. Comparing the fold change calculations with the microarray data, ISP assembly results reach a higher PCC (Pearson

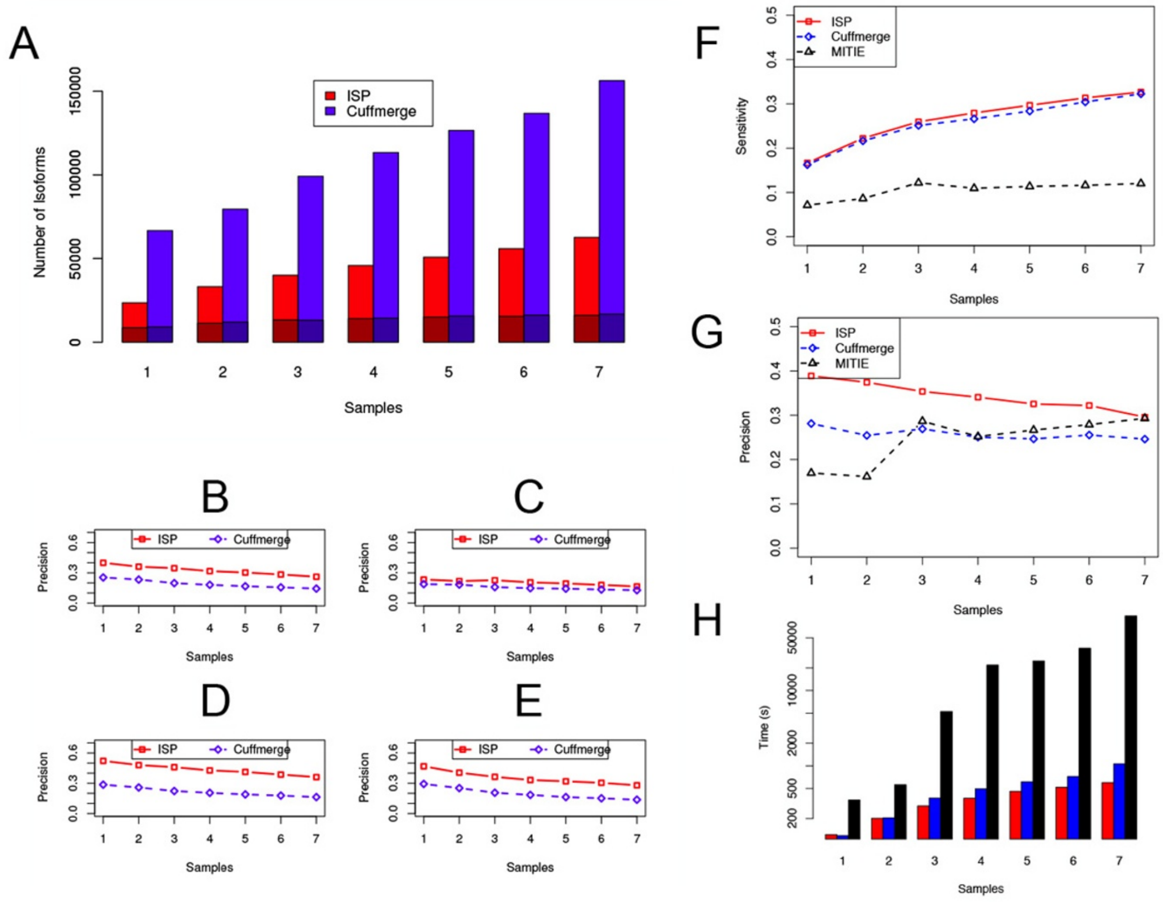

Figure 5 The performance of different algorithms on ENCODE RNA-Seq datasets. (A) The number of predicted isoforms by ISP and Cuffmerge using multiple samples. The shaded region at the bottom of the bar shows the number of predicted isoforms that match UCSC human known transcripts. (B-E) The precision of both algorithms on multi-exon isoforms using all samples (B), and the precision of both algorithms on isoforms grouped by the number of exons (C-E). The plots in C, D and E show the corresponding precision for isoforms with 2-4, 5-10 and over 10 exons, respectively. (F-H) The sensitivity, precision and running time of ISP, Cuffmerge and MITIE on 500 randomly selected genes in chromosome 1.

Correlation Coefficient) value than Cuffmerge (0.68 vs 0.58). To further compare the differential analysis results, we select transcripts that show the largest fold changes between samples (similar to the simulation experiments). For the corresponding microarray measurements of these transcripts, we used Student's t-test to check the statistical significance that these transcripts are differentially expressed between both samples. Table 1 shows the PCC values of fold change calculations between RNA-Seq and microarray measurements, and the numbers of top ranked differentially expressed transcripts confirmed by microarray data. The fold change calculations based on the assembly results of ISP are more accurate since they achieve higher PCC values than Cufflinks, and a higher number of predictions are confirmed by microarray measurements. This shows that by using the isoforms inferred by ISP, we are able to obtain a more accurate differential analysis than Cuffmerge.

Conclusion

With the advance of next generation sequencing technologies, it is now possible to reconstruct full-length transcripts, estimate their expression levels, and compare the structural and quantitative differences between samples. Transcriptome assembly may benefit from the existence of multiple sample RNA-Seq data, but may also be confused by inherent RNA-Seq errors, which in turn affects downstream differential analysis. In this paper, we have designed an algorithm (ISP) to reconstruct transcriptomes for multiple sample RNA-Seq data that is able to handle errors effectively by using an

Table 1 The correlation to microarray fold change calculations, and the numbers of differentially expressed isoforms confirmed by microarray measurements among the top ranked isoforms

Top isoforms		10	50	100	200
PCC	ISP	0.95	0.86	0.87	0.86
	Cuffmerge	0.89	0.82	0.82	0.81
confirmed ($p < 0.05$)	ISP	9	45	94	170
	Cuffmerge	8	43	82	148
confirmed ($p < 0.001$)	ISP	8	35	77	135
	Cuffmerge	7	32	64	108

iterative linear programming (or integer linear programming) approach. Both simulated and real experimental results demonstrate that, obtaining a set of accurately assembled transcripts is crucial for downstream differential analysis. A large number of false positives decrease the accuracy of estimating the expression fold changes of isoforms between samples, and ISP is able to achieve a better differential analysis performance by accurately assembling transcripts from multiple samples directly.

Additional material

Additional file 1: Supplementary Materials. This file includes the description of the ISP algorithm, the NP-completeness proof of the ISP problem, the indirect comparison between ISP and CLIIQ, an example of transcriptome assembly results on ENCODE samples, and the comparison of detecting alternative splicing events between ISP and Cuffmerge.

Competing interests

The authors declare that they have no competing interests.

Declarations

The research was partially supported by National Science Foundation grant DBI-1262107.

This article has been published as part of *BMC Genomics* Volume 16 Supplement 2, 2015: Selected articles from the Thirteenth Asia Pacific Bioinformatics Conference (APBC 2015): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/16/S2

Authors' details

[1]Department of Computer Science and Engineering, University of California, Riverside, Riverside, CA, 92507, USA. [2]MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST / Department of Automation, Tsinghua University, Beijing, 100084, China. [3]MOE Key Lab of Bioinformatics and Bioinformatics Division, TNLIST / Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China. [4]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA, 02215, USA.

Published: 21 January 2015

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by rna-seq.** *Nature Methods* 2008, **5(7)**:621-628.
2. Hon GC, Hawkins RD, Caballero OL, Lo C, Lister R, Pelizzola M, Valsesia A, Ye Z, Kuan S, Edsall LE, Camargo AA, Stevenson BJ, Ecker JR, Bafna V, Strausberg RL, Simpson AJ, Ren B: **Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.** *Genome Research* 2011.
3. Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z, Chen F, Wu X, Zhang G, Fang X, Kang Y, Anderson GA, Harris TW, Campbell BE, Vlaminck J, Wang T, Cantacessi C, Schwarz EM, Ranganathan S, Geldhof P, Nejsum P, Sternberg PW, Yang H, Wang J, Wang J, Gasser RB: **Ascaris suum draft genome.** *Nature* 2011, **479(7374)**:529-533.
4. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nature Biotechnology* 2010, **28(5)**:511-515.
5. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with rna-seq.** *Nature biotechnology* 2013, **31(1)**:46-53.
6. Feng J, Li W, Jiang T: **Inference of isoforms from short sequence reads.** In *Research in Computational Molecular Biology. Lecture Notes in Computer Science. Volume 6044.* Springer;Berger, B 2010:138-157.
7. Li W, Feng J, Jiang T: **IsoLasso: A LASSO Regression Approach to RNA-Seq Based Transcriptome Assembly.** In *Research in Computational Molecular Biology. Lecture Notes in Computer Science. Volume 6577.* Springer, Berlin, Heidelberg;Bafna, V., Sahinalp, S 2011:168-188, Chap. 18.
8. Li JJ, Jiang C-R, Brown JB, Huang H, Bickel PJ: **Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation.** *Proceedings of the National Academy of Sciences* 2011, **108(50)**:19867-19872.
9. Lin Y-Y, Dao P, Hach F, Bakhshi M, Mo F, Lapuk A, Collins C, Sahinalp SC: **CLIIQ: Accurate Comparative Detection and Quantification of Expressed Isoforms in a Population Algorithms in Bioinformatics.** In *Lecture Notes in Computer Science. Volume 7534.* Springer, Berlin, Heidelberg; 2012:178-189, Chap. 14.
10. Behr J, Kahles A, Zhong Y, Sreedharan VT, Drewe P, Rätsch G: **Mitie: Simultaneous rna-seq-based transcript identification and quantification in multiple samples.** *Bioinformatics* 2013, **29(20)**:2529-2538.
11. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome research* 2009, **19(6)**:1117-1123.
12. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nature Biotechnology* 2011, **29(7)**:644-652.
13. Martin JA, Wang Z: **Next-generation transcriptome assembly.** *Nature Reviews Genetics* 2011, **12(10)**:671-682.
14. Ponting CP, Belgard TG: **Transcribed dark matter: meaning or myth?** *Human Molecular Genetics* 2010, **19(R2)**:162-168.
15. The ENCODE Project Consortium: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306(5696)**:636-640.
16. The ENCODE Project Consortium: **A User's Guide to the Encyclopedia of DNA Elements (ENCODE).** *PLoS Biol* 2011, **9(4)**:1001046.
17. The Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474(7353)**:609-615.
18. The modENCODE Consortium: **Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE.** *Science* 2010, **330(6012)**:1787-1797.
19. Trapnell C, Pachter L, Salzberg SL: **Tophat: discovering splice junctions with rna-seq.** *Bioinformatics* 2009, **25(9)**:1105-1111.
20. Au KF, Jiang H, Lin L, Xing Y, Wong WH: **Detection of splice junctions from paired-end RNA-seq data by SpliceMap.** *Nucleic Acids Research* 2010, **38(14)**:4570-4578.
21. Pruitt KD, Tatusova T, Brown GR, Maglott DR: **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.** *Nucleic Acids Research* 2012, **40(D1)**:130-135.
22. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The ucsc known genes.** *Bioinformatics* 2006, **22(9)**:1036-1046.
23. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas.** *Nature Biotechnology* 2010, **28(5)**:503-510.

24. Wu Z, Wang X, Zhang X: **Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq.** *Bioinformatics* 2011, **27(4)**:502-508.
25. Fuglede B, Topsoe F: **Jensen-Shannon Divergence and Hilbert Space Embedding.** 2004, 31.
26. GNU Linear Programming Kit (GLPK): 2008 [http://www.gnu.org/software/glpk/].
27. Werbos PJ: **Backpropagation: past and future.** *Neural Networks, 1988., IEEE International Conference On* IEEE; 1988, 343-3531.
28. Li W, Jiang T: **Transcriptome assembly and isoform expression level estimation from biased rna-seq reads.** *Bioinformatics* 2012, **28(22)**:2914-2921.
29. Wang Z, Xiao X, Van Nostrand E, Burge CB: **General and specific functions of exonic splicing silencers in splicing control.** *Molecular cell* 2006, **23(1)**:61-70.
30. Matlin AJ, Clark F, Smith CW: **Understanding alternative splicing: towards a cellular code.** *Nature reviews. Molecular cell biology* 2005, **6(5)**:386-398.
31. Bengtsson M, Ståahlberg A, Rorsman P, Kubista M: **Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels.** *Genome Research* 2005, **15(10)**:1388-1392.
32. The ENCODE Project Consortium: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447(7146)**:799-816.

doi:10.1186/1471-2164-16-S2-S15
Cite this article as: Tasnim *et al.*: **Accurate inference of isoforms from multiple sample RNA-Seq data.** *BMC Genomics* 2015 **16**(Suppl 2):S15.