JLINKS: A NOVEL ISOFORM ABUNDANCE ESTIMATION METHOD USING SPLICE JUNCTIONS

by

Jingyi Lu

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland

December, 2015

Abstract

Transcripts, or interchangeably referred to as isoforms, have been well known to be involved in many important biological pathways and disease mechanisms such as cancer and mental disorders. Understanding the roles of isoforms calls for precise quantification of isoform expression abundances from RNA-Seq reads. Yet state-of-the-art isoform quantification methods yield weak estimation accuracy, especially for datasets that undergo a wide range of isoform expression levels. Here we present a novel isoform quantification algorithm called Jlinks, designed to estimate isoform abundances using splice junctions. The key distinguishing feature of Jlinks is that it treats each isoform as a "link" of splice junctions and converts the abundance estimation problem into obtaining an optimal solution for a linear system. We demonstrate that Jlinks outperforms existing isoform quantification methods in both speed and accuracy.

Advisor: Yuan Gao, Ph.D.

Thesis Committee: Joel S. Bader, Ph.D.

Yuan Gao, Ph.D.

Hongkai Ji, Ph.D.

Acknowledgements

First and foremost, I would like to thank my advisor Dr. Yuan Gao for his guidance and support, without which this thesis would not be possible. His academic foresights and creative ideas have always been inspiring me. It's my great pleasure to complete this work under his supervision.

I would also like to thank Dr. Hun Ki Lim for his long-term help and advice on my thesis work, for being the person I could always turn to whenever faced with any problems and bottlenecks. Special thanks go to Christopher Hartl, for all the helpful discussions and suggestions that contribute to this work.

Last but not least, I would like to thank my thesis committee members Dr. Bader and Dr. Ji, for taking time off their busy schedules to review and edit my thesis. Thanks for all the help and support.

Contents

List of Tables

List of Figures

1. Introduction

In eukaryotic species, genes with multiple exons are known to undergo alternative splicing events that encode multiple spliced isoforms, also termed as transcripts, which encode distinct but related protein products [1]. Studies have shown that in humans, more than 90% of multi-exon genes are subject to alternative splicing [2] and 50% of disease-causing mutations affect splicing events [3,4]. Therefore, it is crucial to obtain precise estimates of isoform expression abundances as part of conducting differential expression analyses across samples and conditions.

Several technologies have been used to quantify isoform expression levels such as cloning cDNAs or expressed sequence tag (EST) libraries, followed by capillary sequencing [5-7]. Due to the high cost and limited resolution, these approaches could not provide a thorough characterization of the true complexity of alternative splicing and transcription [8]. Nowadays the massively parallel sequencing technologies from Illumina, Applied Biosystems and Roche 454 Life Sciences have revolutionized the study of transcriptomes [9]. High-throughput RNA sequencing (RNA-Seq) makes it possible to generate comprehensive pictures of transcriptomes, allowing isoform abundance estimations at unprecedented levels of resolution, accuracy and low cost.

Current RNA-Seq technologies generate RNA-Seq reads with lengths ranging from 25 nt to 300 nt. Limited read lengths result in a significant amount of *multireads*, *i.e.*, reads that map ambiguously to multiple isoforms or paralogs whose sequences are similar to each other. The key challenge in isoform abundance estimation is to accurately assign those *multireads* to isoforms.

Generally, isoform abundance estimation methods apply the Expectation-Maximization algorithm to maximize a likelihood function by adjusting isoform abundance parameters. Successful and popular methods of this class include IsoEM [10] and RSEM [11], where RNA-Seq reads are first assigned to isoforms, these assignments are then used to estimate isoform abundances, and these steps are iterated many times until final convergence. However, the first EM-based approaches were time consuming and did not scale well with the size of input datasets. To overcome this obstacle, eXpress [12] optimizes the EM procedure through a streaming algorithm, resulting in a linear run time and constant memory usage while still maintaining comparable quantification accuracy. Another method Sailfish [13] further accelerates the EM procedure through a lightweight algorithm, which uses counts of k-mers instead of alignments of reads to avoid mapping step, and uses k-mer equivalence classes to substantially reduce parametric complexity. Even for these improved approaches, processing large datasets remains to be a computational burden and fundamentally limits their scalability. Therefore, an isoform quantification method with better accuracy and faster run time, easily scalable for large datasets, is urgently needed in the field of transcriptome study.

In this paper we present Jlinks, a novel isoform quantification method that takes advantage of splice junction information to estimate isoform expression abundances. Through experiments on simulated RNA-Seq datasets under various sequencing depths and simulation patterns, we demonstrate that Jlinks consistently outperforms other isoform quantification methods in all scenarios, showing a significantly better global accuracy.

2. Background

2.1 RNA-Seq

The transcriptome is the complete set of transcripts in a cell for a specific physiological condition. Transcriptome study is essential for uncovering the functional elements of the genome and revealing their roles in development stages and disease pathways. Various approaches have been developed to characterize and quantify transcriptomes, including Sanger sequencing-based method [14] and hybridization-based microarray method [15].

The development of ultra high-throughout sequencing of RNA (RNA-Seq) allows transcriptome studies at a finer resolution and greater scale. Compared with the earlier approaches, RNA-Seq method excels in the following aspects: First, RNA-Seq provides digital quantitation rather than signals for gene expression profiling by mapping millions of short reads from transcriptome of interest to the reference genome. In addition, RNA-Seq has high sensitivity even for genes with little expression, providing a wide range of expression levels. Finally, without cloning step, RNA-Seq requires less amount of RNA sample compared to the other technologies. With these advantages, RNA-Seq has become the dominant method for transcriptome analyses in recent studies.

2.2 Alternative splicing

During the transcription process, most eukaryotic genes will be spliced into multiple isoforms sharing common parts of their sequences. By alternative splicing, different mature mRNAs are produced from a single precursor mRNA, resulting in multiple

protein products encoded by a single gene. This phenomenon happens to over 90% of multi-exon human genes, greatly increases the complexity of transcriptome studies.

There are various types of alternative splicing, among which five basic models are generally recognized: exon skipping, mutually exclusive exons, alternative donor site, alternative acceptor site and intron retention. Figure 2.1 shows the mechanism of these five classical types. Alternative splicing is believed to be involved in the regulations of various physiological functions. It has been known that cancer cells have higher levels of intron retention and lower levels of exon skipping, compared with normal cells [16]. A recent study of RNA-Seq and proteomics revealed striking differential expression of splice isoforms of key proteins in important cancer pathways [17].

2.3 Isoform abundance estimation

A typical process of transcriptome analysis consists of three steps: read alignment, isoform quantification and differential expression analysis. First, RNA-Seq reads are mapped to a reference transcriptome by unspliced aligners such as Bowtie [18] and BWA [19], or a reference genome by spliced aligners such as TopHat [20], MapSplice [21] and STAR [22]. Second, isoforms are either assembled from these alignments or provided by a known set, and their expression abundances are estimated. Lastly, the estimated isoform abundances are used to analyze differential expressions among samples, uncovering the roles of isoforms in biological pathways. Table 2.1 provides a list of currently available tools for each step.

There are two major tasks in isoform quantification from RNA-Seq data: isoform assembly and abundance estimation. The first task aims at assembling the complete set of

isoforms from RNA-Seq reads while the second task aims at quantifying the expression levels for a given set of isoforms. Assembling reads into full isoforms is challenging due to the limited information from single-end or paired-end short reads and the complicated isoform structures. Abundance estimation for a known set of isoforms is also challenging for that reads might be ambiguously mapped to multiple isoforms of a gene as well as multiple genes within a gene family. This ambiguity makes it difficult to estimate the expression abundances of isoforms, especially those with few unique regions.

Many methods have been developed to tackle the isoform assembly problem (*e.g.*, Trinity [23], Oases [24], Trans-ABySS [25]), or the abundance estimation problem (*e.g.*, RSEM, eXpress, Sailfish) or both (*e.g.*, IsoInfer [26], Scripture [27], Cufflinks [28]). In this paper we focus on the abundance estimation problem and propose a novel method Jlinks to estimate isoform expression abundances using spliced junctions.

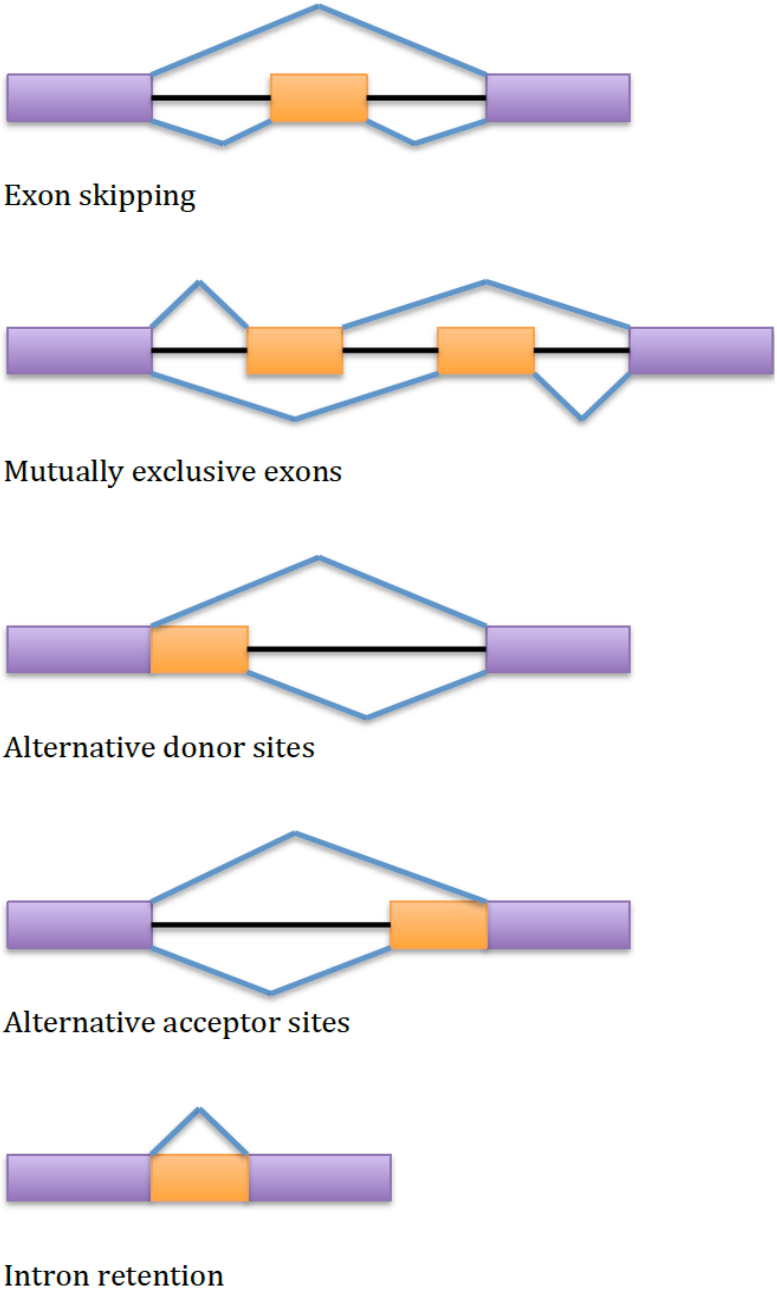

Figure 2.1: **Five basic models of alternative splicing.** Purple blocks represent constitutive exons, and orange blocks represent alternatively spliced exons.

Table 2.1: **Selected list of transcriptome analysis tools in each step.** Shaded part is the task we are focused on in this paper.

Step	Category	Tool	Usage
Read Alignment	Unspliced aligners	Bowtie, BWA	Align RNA-Seq reads to a reference transcriptome
	Spliced aligners	TopHat, MapSplice, STAR	Align RNA-Seq reads to a reference genome and identify splice junctions
Isoform Quantification	Isoform assembly	Trinity, Oases, Trans-ABySS	Assembly a set of isoforms from read alignments
	Abundance estimation	RSEM, eXpress, Sailfish	Estimate expression abundances for a given set of isoforms
	Both	IsoInfer, Scripture, Cufflinks	Simultaneously assembly isoforms and estimate their expression abundances
Differential Expression Analysis	Cuffdiff [29], DESeq [30], edgeR [31]		Identify differentially expressed isoforms across samples

2.4 Previous methods

The key challenge for isoform abundance estimation is to accurately assign *multireads* to isoforms. The early solution for this challenge was simply discarding any reads that mapped ambiguously, leaving only the "unique" reads to do the abundance estimation. A "rescue" method was then conducted to fractionally allocate *multireads* according to the estimated expression abundances [32]. However, this method did not make full use of the information from RNA-Seq reads and generated high variances and significant biases in the quantification.

Having realized that the "rescue" method is equivalent to a single iteration of the Expectation-Maximization algorithm, researchers extended this method to a full version of EM algorithm: At Expectation step, reads are probabilistically assigned to isoforms based on the current abundance estimates; at Maximization step, current estimates are updated to maximize the likelihood function given the current read assignments. These steps are iterated until reaching the threshold of convergence. With a concave likelihood function, the parameters will eventually converge to the maximum likelihood estimates. Most current methods for isoform abundance estimation are derived from this EM algorithm, with various forms of likelihood functions.

2.4.1 RSEM method

RSEM, as its name "RNA-Seq by Expectation Maximization" suggests, applies the EM algorithm to handle reads that map ambiguously. It computes the maximum likelihood values of the parameters θ, where θ_i represents the probability that a fragment

is derived from isoform i. Then the isoform fractions τ is computed from θ and effective lengths $\bar{l}$:

$$\tau_i = \frac{\theta_i / \bar{l}_i}{\sum_j \theta_j / \bar{l}_j}$$

The effective length $\bar{l}_i$ of isoform i is given by

$$\sum_{x \le l_i} \lambda_F(x)(l_i - x + 1)$$

where λ_F is the fragment length distribution.

RSEM computes a maximum likelihood estimate for θ using the EM algorithm. The iterations are terminated when all θ_i with value $\geq 10^{-7}$ have a relative change of less than 10^{-3}. The outputs of RSEM consist of the isoform fractions τ, as well as the expected number of fragments originating from each isoform, given the ML parameters.

2.4.2 eXpress method

eXpress uses a probabilistic graphical model for fragment assignment. Applying Bayes rule, the joint probability of obtaining a fragment f of length l sequenced from position p on target t is given by

$$P(L = l, T = t, P = p, F = f)$$

$$= P(L = l)P(T = t|L = l)P(P = p|T = t, L = l)P(F = f|P = p, T = t, L = l)$$

Use parameters to represent the conditional probabilities:

$$\lambda_l = P(L = l) \qquad \tau_{t|l} = P(T = t|L = l)$$

$$\pi_{p|t,l} = P(P = p|T = t, L = l) \qquad \phi_{f|p,t,l} = P(F = f|P = p, T = t, L = l)$$

The joint probability becomes

$$P(L = l, T = t, P = p, F = f) = \lambda_l \tau_{t|l} \pi_{p|t,l} \phi_{f|p,t,l}$$

The likelihood function for a set of sequenced fragments $\mathcal{F}$ originating from a set of target sequences $\mathcal{T}$ is given by

$$L(\lambda, \tau, \pi, \phi|\mathcal{F}) = \prod_{f\in\mathcal{F}} \sum_{l=1}^{M_L} \sum_{t\epsilon\mathcal{T}} \sum_{p=1}^{l(t)-l+1} \lambda_l \tau_{t|l} \pi_{p|t,l} \phi_{f|p,t,l}$$

Let τ_t denotes the relative abundance of target t, which satisfies $\tau_t = \sum_l \tau_{t|l}$, the likelihood function can be rewritten as

$$L(\lambda, \tau, \pi, \phi|\mathcal{F}) \propto \prod_{f\in\mathcal{F}} \sum_{l=1}^{M_L} \sum_{t\epsilon\mathcal{T}} \sum_{p=1}^{l(t)-l+1} \lambda_l \tau_t \frac{\omega_{p|t,l}}{\tilde{l}(t)} \phi_{f|p,t,l}$$

where M_L is the maximum length of fragment f and $\tilde{l}(t)$ is the effective length of target sequence t.

The model described here is similar to the RSEM model, which obtains the maximum value of the likelihood function by iterating and adjusting parameters τ_t. But eXpress optimizes RSEM's algorithm through an alternative optimization procedure: streaming EM algorithm. It approximates the batch EM without accessing the alignment of each fragment more than once, resulting in a significant reduction of time and memory.

2.4.3 Sailfish method

Another method Sailfish implements an alignment-free, accelerated EM algorithm for isoform abundance estimation. Unlike a typical alignment process, it creates a unique k-mer index $I_k(T)$ for the given isoform set T, and catalogs the k-mer counts for each read in the RNA-Seq read set $\mathcal{R}$. The isoform abundances are estimated using those k-mer counts instead of alignments of reads.

For each k-mer $s_i \in kmers(T) \bigcap kmers(\mathcal{R})$, let $C_{\mathcal{R}}(s_i)$ denotes the number of occurrences of s_i in $\mathcal{R}$. Define a k-mer equivalence class $[s_i]$ as the set of k-mers

occurring in the same set of isoforms with the same frequency, then the total amount of k-mers originating from equivalence class $[s_i]$ is

$$L(s_i) = \sum_{s_j \in [s_i]} C_{\mathcal{R}}(s_j)$$

Sailfish then applies the EM algorithm to estimate the relative abundances of isoforms. In the E-step, the fraction of k-mer equivalence class $[s_j]$'s total count allocated to isoform t_i is computed as

$$\alpha(j,i) = \frac{\mu_i' L(s_j)}{\sum_{t \supseteq [s_j]} \mu_t'}$$

where μ_i' is the current estimate of relative abundance of isoform t_i. In the M-step, the relative abundance of isoform t_i is updated as

$$\mu_i' = \frac{\mu_i}{\sum_{t_j \in T} \mu_j}$$

where μ_i is

$$\mu_i = \frac{\sum_{[s_j] \subseteq t_i} \alpha(j,i)}{l_i - k + 1}$$

By using k-mers to avoid the mapping step, and collapsing millions of k-mers into equivalence classes, the Sailfish algorithm reduces parametric complexity substantially, making the convergence of the EM algorithm much faster than other EM-based quantification methods.

3. Jlinks method

Since isoforms are generated from alternative splicing, each isoform can be regarded as a unique combination of splice junctions within this gene. Therefore we can infer isoform abundances from their junction coverage by solving a linear system. Based on this idea we developed a novel algorithm Jlinks to estimate isoform expression abundances using splice junctions.

3.1 Algorithm

Unlike the EM-based methods RSEM and eXpress, whose first step is to map RNA-Seq reads onto a known set of isoforms using unspliced aligners such as Bowtie, Jlinks requires spliced aligners such as TopHat, MapSplice and STAR. It takes advantage of the alignment files along with the junction files generated from those aligners, estimates the relative abundances of isoforms for each gene, and outputs the estimated fragment counts as well as FPKM (Fragments Per Kilobase of transcript per Million fragments mapped) values of each isoform. Figure 3.1 shows a brief workflow of Jlinks program.

Jlinks estimates the isoform abundances for each gene in a case-by-case manner. For a given gene, Jlinks first measures the amount of fragments originating from this gene by counting how many fragments fall into its genomic region. If this gene overlaps other genes, Jlinks merges all the overlapping genes into a single *supergene* and treats isoforms of each overlapping gene as isoforms of this *supergene*. Having obtained the fragment count F_{gene}, Jlinks performs isoform abundance estimation for this gene or *supergene* using splice junctions.

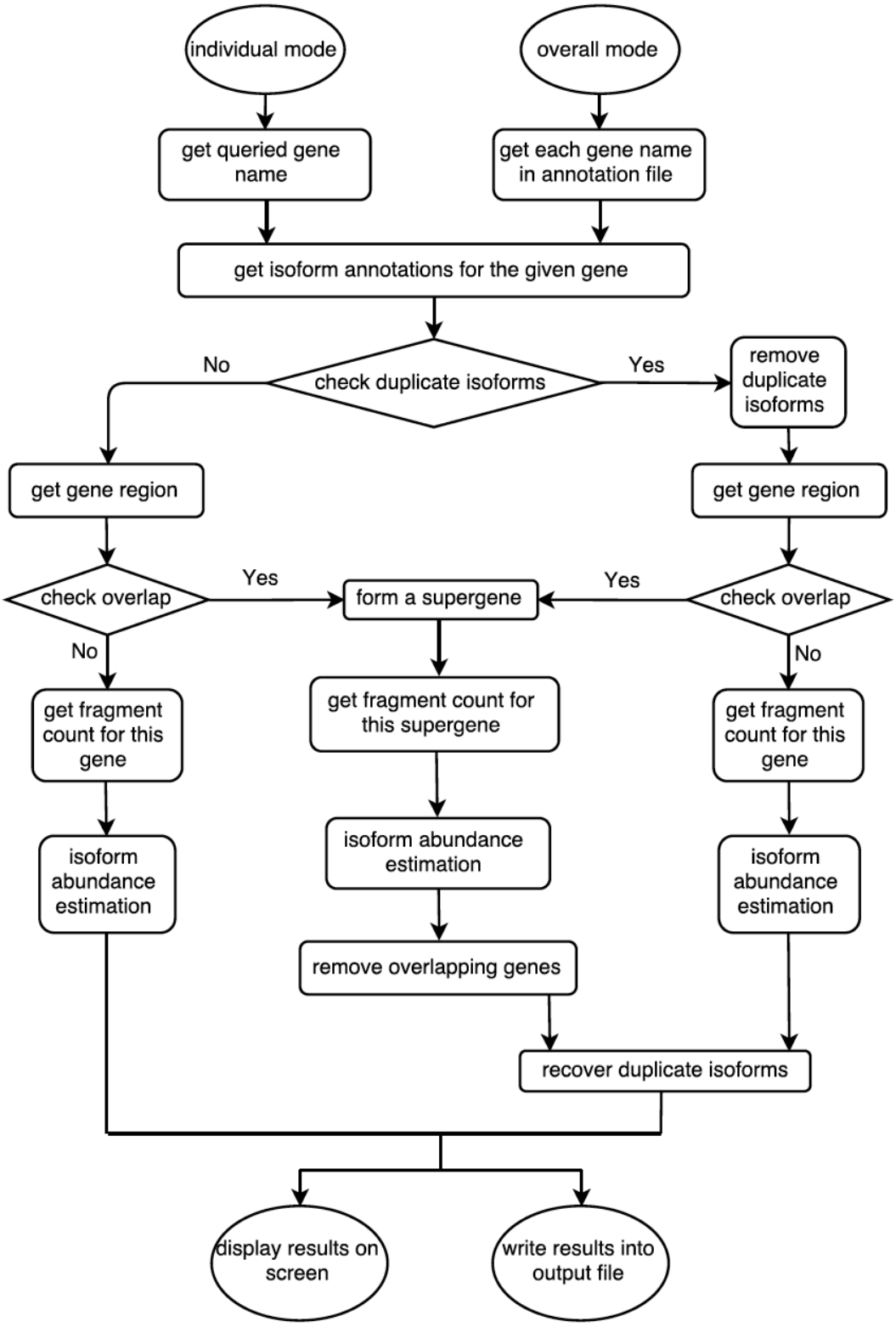

Figure 3.1: **Jlinks algorithm workflow**

Jlinks treats each isoform as a "link" of splice junctions. For a given gene, suppose it has M isoforms $T_1, T_2, \dots, T_M$ and N splice junctions within its genomic locus: $J_1, J_2, \dots, J_N$. Each isoform T_m can be represented as a unique set of those junctions:

$$T_m = \left(J_{m_1}, J_{m_2}, \dots, J_{m_{N_m}}\right) \quad m = 1, \dots, M$$

The fragment coverage for each isoform is simplified as a uniform distribution along the isoform sequence. Under this assumption, denoting the fragment coverage of splice junctions as $C = (c_1, \dots, c_N)^T$, we want to infer the fragment coverage of isoforms: $X = (x_1, \dots, x_M)^T$. The coverage of junction J_n is the sum of the coverage of all isoforms having this junction:

$$c_n = \sum_{k \in S_n} x_k \quad n = 1, \dots, N$$

where S_n is the subset of isoforms having splice junction J_n. The above equation set can be rewritten into a matrix form as described below.

Define an isoform-junction representation matrix $A = \{a_{ij}\}$ $i = 1, \dots, N; j = 1, \dots, M$, where a_{ij} = 1 if isoform T_j contains junction J_i; a_{ij} = 0 otherwise. Then we have

$$AX = C$$

This linear system can be categorized into four conditions according to the properties of matrix A:

Condition 1. If $rank(A) = m = n$, the problem has a unique solution

$$X = A^{-1}C$$

Condition 2. If $rank(A) = m < n$, the problem has a unique least-squares solution

$$X = (A^T A)^{-1} A^T C$$

Condition 3. If $rank(A) = n < m$, the problem has a unique minimum-norm least-squares solution

$$X = A^T(AA^T)^{-1}C$$

In other cases where A is a non-square singular matrix with neither inverse nor pseudo-inverse, we can still obtain a unique minimum-norm least-squares solution by applying the following theorem in Linear Algebra:

Rank Factorization Theorem: *Any $n \times m$ matrix A of rank r can be decomposed as $A = FG$, where F is a $n \times r$ full column rank matrix, G is a $r \times m$ full row rank matrix.*

Thus we have:

Condition 4. If $rank(A) < min(n, m)$, the problem has a unique minimum-norm least-squares solution

$$X = G^T(GG^T)^{-1}(F^TF)^{-1}F^TC$$

where $A = FG$ is a rank factorization of matrix A.

The mathematical proofs of Condition 1~4 and the *Rank Factorization Theorem* are provided in the **Appendix**. Therefore, for any n and m, we can always obtain an optimal solution for the fragment coverage of isoforms:

$$X^* = (x_1^*, \dots, x_M^*)$$

In addition to fragment coverage, isoform length is another factor influencing isoform expression abundance because longer sequences generate more fragments than shorter ones given the same coverage. Define the effective length of an isoform as

$$\bar{l}_m = \sum_{i=1}^{l_m} P(i)(l_m - i + 1) \quad m = 1, \ldots, M$$

where l_m is the length of isoform T_m, and P is the fragment length distribution. The fragment count for each isoform is given by

$$F_m = \frac{x_m^* \bar{l}_m}{\sum_{m=1}^{M} x_m^* \bar{l}_m} F_{gene} \quad m = 1, \ldots, M$$

Jlinks estimates the isoform abundances for each alternatively spliced gene. As for genes with single isoform and not overlapped by others, Jlinks skips the above estimation procedure and outputs the fragment count F_{gene} directly. This greatly reduces the computation time without any loss of quantification accuracy.

3.2 Implementation

3.2.1 Input files

Jlinks has three input files: a GTF format annotation file for known isoforms, a BED format junction file generated from any spliced aligner, and the BAM format alignment file from that aligner. The annotation files can be easily downloaded from databases such as Ensembl, Genbank and the UCSC Genome Browser Database. Jlinks can also use annotation files generated by *de novo* isoform assemblers as long as they are in GTF format. Since SAM and BAM files are convertible, spliced aligners which provide SAM format alignment files are also compatible. Moreover, if a spliced aligner does not provide junction files (which is rarely the case in spliced aligners), junction information can be extracted from the alignment SAM file to generate a BED format junction file following these steps:

```
samtools view -HS accepted_hits.sam > header.sam
samtools view -hS accepted_hits.sam | awk '($6 ~ /N/)' > spliced_hits.sam
cat header.sam spliced_hits.sam > spliced_hits_with_header.sam
samtools view -bS spliced_hits_with_header.sam > file.bam
bamToBed -bed12 -i file.bam > file.bed12
bed12ToBed6 -i file.bed12 > file.bed6
subtractBed -a file.bed12 -b file.bed6 -s | cut -f 1 -6 > pre.junctions.bed
```

The resulting file pre.junctions.bed is a BED format file containing the splice junction information for each spliced read. The preprocessing script junction_pileup.py can then pile up the junctions: python junction_pileup.py pre.junctions.bed junctions.bed

The output file junctions.bed is ready for use by Jlinks.

3.2.2 Jlinks modes

Jlinks can run in two modes.

1. **Individual mode**

This mode is designed for querying an individual gene without running Jlinks on the whole set of genes. In individual mode, Jlinks takes the name of the queried gene following the option -g, estimates the isoform abundances for this gene and displays the quantification results by standard output. If the queried gene is not in the annotation file, an error message will be printed to screen. Below is the command line:

python Jlinks.py -g GENE_NAME [options] annotationfile junctionfile alignmentfile

2. **Overall mode**

This mode is designed for a complete analysis of all genes contained in the annotation file. In overall mode, Jlinks writes the estimates of the whole gene set into the output file specified by the -o option. Below is the command line:

python Jlinks.py -o OUTPUT_FILE [options] annotationfile junctionfile alignmentfile

Jlinks deals with both single-end and paired-end alignments by specifying the value of the -s option: “yes” for single-end data and “no” for paired-end data. The alignments can be both with or without multi-hits by setting the value of the -m option: “yes” for alignments with multi-hits and “no” for alignments without multi-hits. Jlinks is designed to take advantage of multicore processors, and running the program with multiple threads is highly recommended. The number of threads used for running can be specified by an integer following the -p option.

3.2.3 Output file

When running in overall mode, Jlinks generates an output file containing the isoform abundance estimates for all genes contained in the annotation file. The output file is a single tab-delimited file consisting of five columns. The first column is the gene id of a given isoform, the second column is the transcript id and third column is the length of this isoform. The last two columns are the abundances estimated by Jlinks as represented by two measurements. The fourth column is an estimate of the number of fragments originating from each isoform, and these rounded counts can be used by downstream analysis tools such as DESeq and edgeR to conduct differential expression analyses. The fifth column contains the estimated isoform FPKM values.

4. Results

4.1 Test data

Since there is no ground truth for real transcriptome data, simulating RNA-Seq data has become a standard way to evaluate RNA-Seq analysis methods. We generated test datasets from the human transcriptome using a home-designed RNA-Seq simulator MadeSeq (see **RNA-Seq data simulation**). 100bp paired-end RNA-Seq data sets were simulated ranging from 20 million to 100 million reads, with both uniform and exponential simulation patterns. For the simulation we used the human reference genome (hg19) downloaded from UCSC Genome Browser Database, and annotation file (genes.gtf) downloaded from the RefSeq website containing a total of 35,066 isoforms from 19,088 genes. The isoform length distribution and the number of isoforms per gene are shown in Figure 4.1 and 4.2.

4.2 RNA-Seq data simulation

We used a home-designed java program MadeSeq to simulate RNA-Seq datasets with known ground truth. This program simulates paired-end RNA-Seq reads for an annotated transcriptome, and generates a SAM format answer file recording the origin of each read, *i.e.* which isoform it came from. In this way the true fragment count and FPKM value of each isoform are known, and this ground truth is then used to evaluate the accuracy of various isoform quantification methods.

MadeSeq has two simulation patterns: uniform and exponential. The uniform pattern samples reads uniformly and independently from isoforms in the transcriptome and

across all possible start sites. The exponential pattern samples the frequencies of isoforms from an exponential distribution.

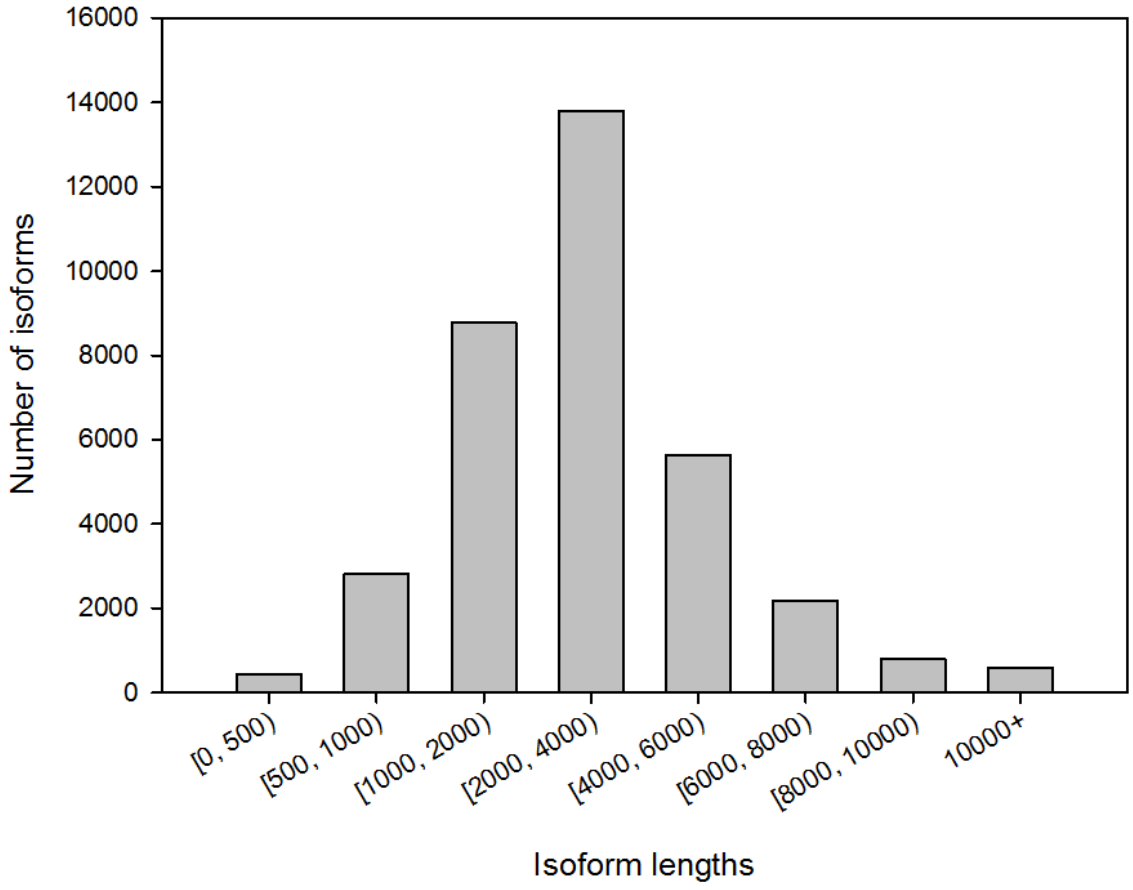

Figure 4.1: **Length distribution of isoforms in the RefSeq annotation file**

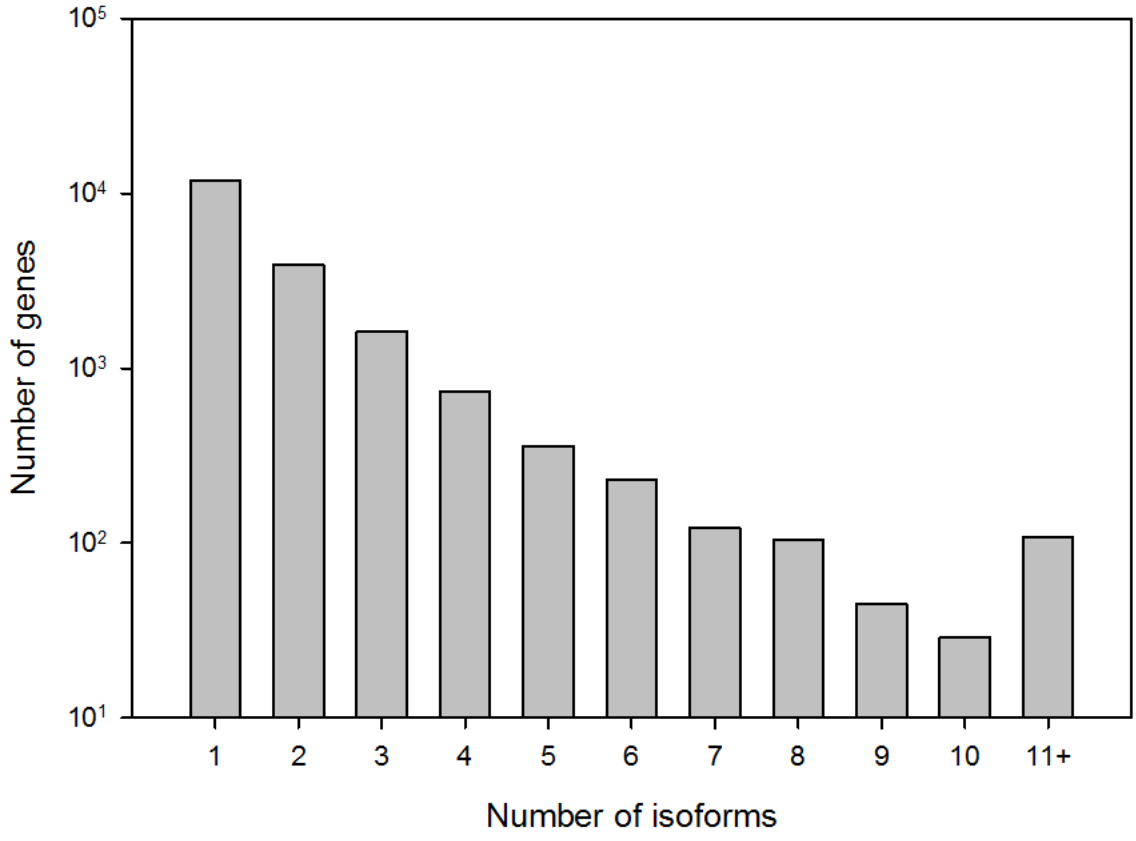

Figure 4.2: **Distribution of the number of isoforms per genes in the RefSeq annotation file**

4.3 Performance comparison for 100 million paired-end RNA-Seq simulation data

4.3.1 Uniform simulation pattern

Jlinks requires alignments to genome, while other EM-based methods such as RSEM and eXpress require alignments to the transcriptome. To generate the alignments suitable for use by Jlinks, we used three popular spliced aligners: (1) TopHat v2.0.14 with the option -G to supply the annotation file, option --no-novel-juncs to only align reads to these annotated isoforms, --microexon-search to find alignment incidents to micro-exons and --max-multihits=1 to align without multi-hits. With these parameters, 99.4% of the

read pairs mapped to the reference genome. (2) MapSplice v2.1.9 with --gene-gtf to supply the annotation file, --non-canonical to search for non-canonical in addition to canonical and semi-canonical junctions, and --filtering=1 to increase the sensitivity of splice junction detection. As a result, 99.93% of reads were successfully mapped. (3) STAR v2.4.1d with --sjdbGTFfile to supply annotation file and no other advanced options, resulting in a 99.6% alignment rate.

To generate the alignments used by eXpress and RSEM, we first extracted the isoform sequences from the annotation file and then used Bowtie2 to align reads to this set of target isoforms. We used Bowtie v2.2.5 with the option -a to report all mappings, -X 1000 to allow fragments up to length 1000 and -v 3 to allow up to three mismatches in each read. With these parameters, 98.81% of the simulated read pairs mapped to the target isoforms. Table 4.1 summarizes the upstream alignment tools for isoform quantification methods except for the alignment-free method Sailfish. Each alignment tool results in a mapping rate over 98%, demonstrating that the impact of alignment algorithms on the quantification results is very little. This allows us to tease out the contribution of various alignment tools to overall isoform abundance estimation accuracy.

Table 4.1: **Summary of upstream alignment tools**

Aligner Type	Alignment Tool	Version	Options	Alignment Rate
Spliced aligner (align to genome)	TopHat	2.0.14	-G --no-novel-juncs --microexon-search --max-multihits=1	99.4%
	MapSplice	2.1.9	--gene-gtf --non-canonical --filtering=1	99.93%
	STAR	2.4.1d	--sjdbGTFfile	99.6%
Unspliced aligner (align to transcriptome)	Bowtie2	2.2.5	-a -X 1000 -v 3	98.81%

We combined Jlinks with TopHat, MapSplice, STAR, and compared the quantification results of Jlinks with the results of RSEM and eXpress which use Bowtie2 as the upstream alignment tool, as well as the alignment-free quantification tool Sailfish. To evaluate the abundance estimation accuracy of each quantification tool, we compared the estimated isoform FPKM values with the true FPKM values. We used root mean square error as accuracy measurement, along with the Pearson correlation coefficient and Spearman correlation coefficient of the FPKM values across all isoforms. Table 4.2

summarizes the overall accuracy comparison of all these isoform quantification methods, Figure 4.3-4.5 display the individual comparisons of root mean square error, Pearson correlation coefficient and Spearman correlation coefficient.

Table 4.2: **Overall accuracy comparison of isoform quantification methods for uniform pattern simulation**

Upstream alignment tool	Isoform quantification tool	Root mean square error	Pearson correlation coefficient	Spearman correlation coefficient
TopHat	Jlinks	8.1116	0.8146	0.8009
MapSplice		8.2497	0.7980	0.7858
STAR		9.1940	0.7797	0.8000
Bowtie2	eXpress	11.7743	0.7099	0.6743
	RSEM	12.9709	0.6778	0.6349
Sailfish		15.8114	0.6146	0.5509

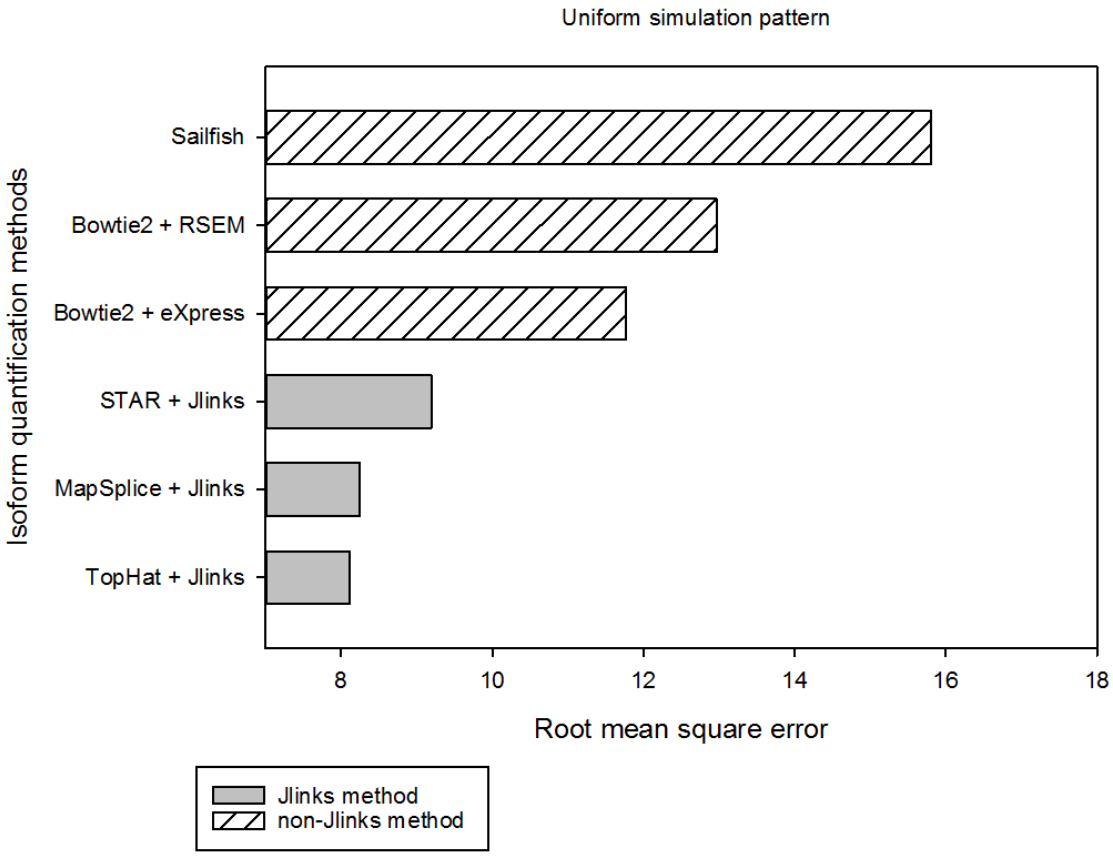

Figure 4.3: **Root mean square error comparison for uniform simulation pattern.** Root mean square errors of estimated isoform FPKM values compared with true FPKM values for isoform quantification methods. Grey bars indicate pipelines using Jlinks as the quantification tool, white striped bars indicate pipelines using other quantification tools.

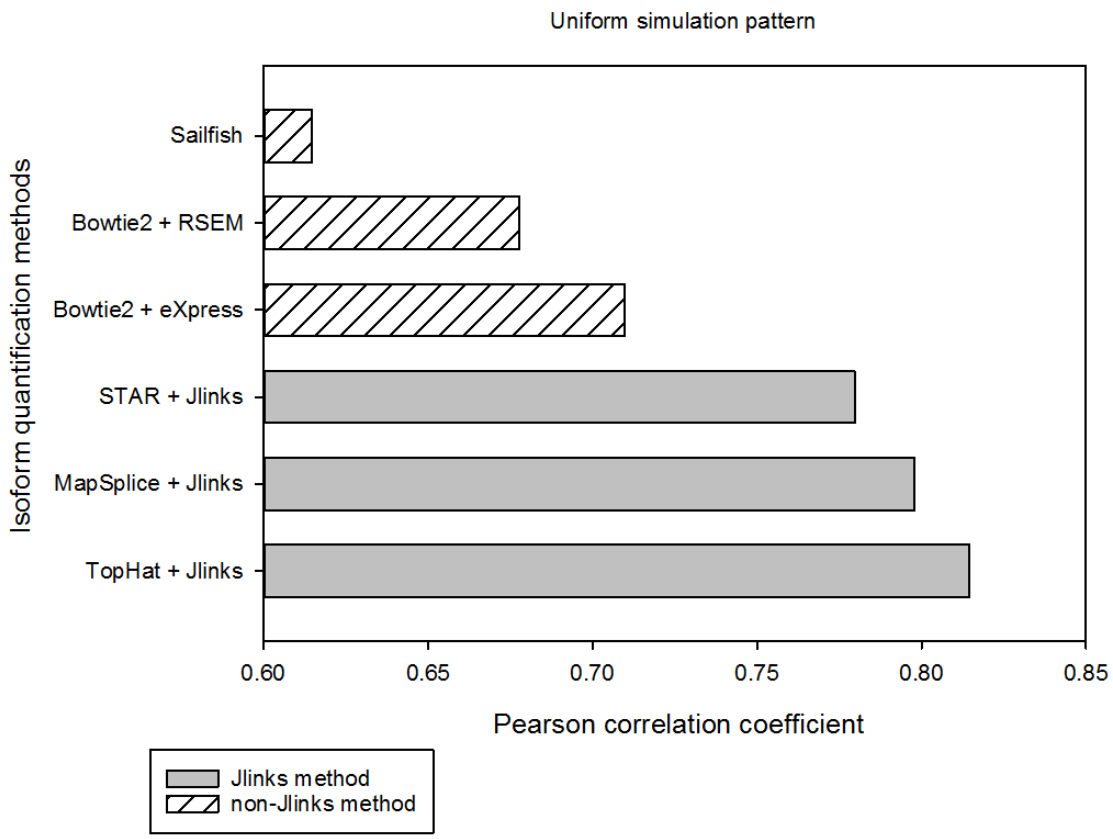

Figure 4.4: **Pearson correlation coefficient comparison for uniform simulation pattern.** Pearson correlation coefficients between estimated FPKM values and true FPKM values for these isoform quantification methods. Grey bars indicate pipelines using Jlinks as the quantification tool, white striped bars indicate pipelines using other quantification tools.

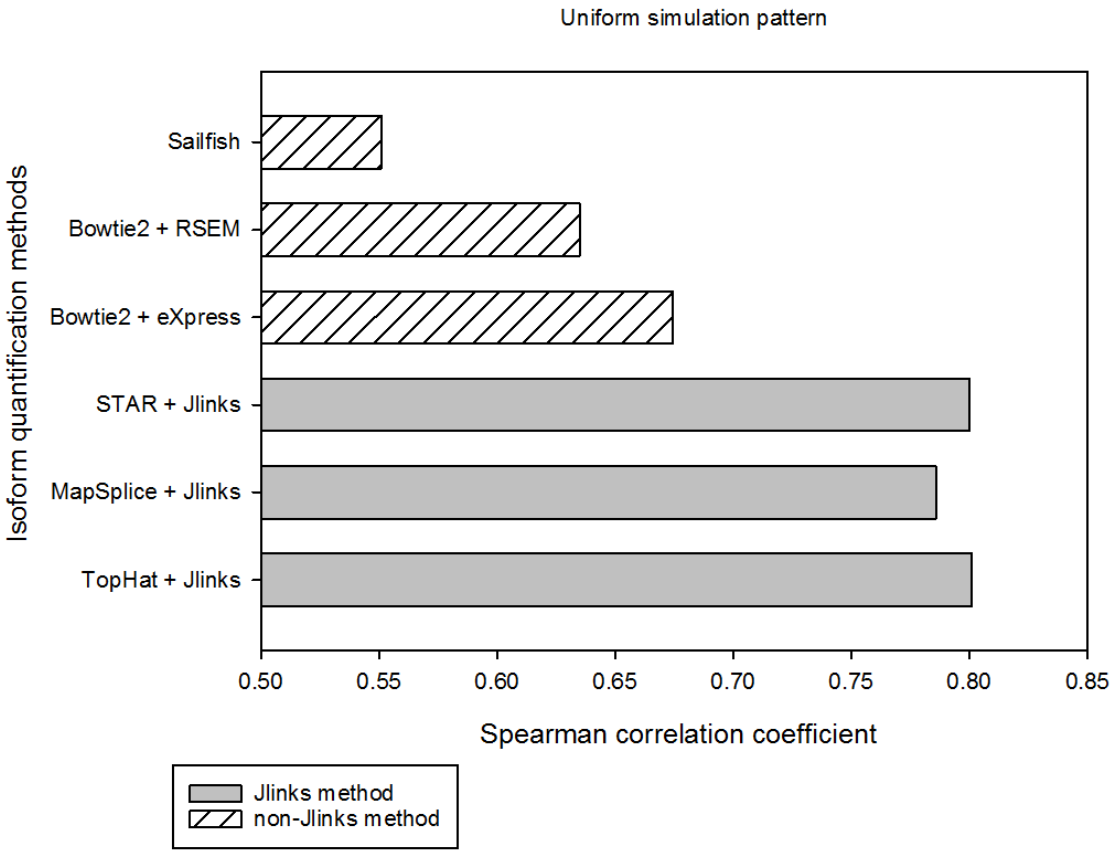

Figure 4.5: **Spearman correlation coefficient comparison for uniform simulation pattern.** Spearman correlation coefficients between estimated FPKM values and true FPKM values for these isoform quantification methods. Grey bars indicate pipelines using Jlinks as the quantification tool, white striped bars indicate pipelines using other quantification tools.

To further examine the performances of these tools in quantifying genes with multiple isoforms, we looked deeper into the results by extracting genes with more than one isoform and dividing them into five categories according to the number of isoforms they have, namely two-isoform genes, three-isoform genes, four-isoform genes, five-isoform genes and six-plus-isoform genes. For each category, we compared the accuracy of these quantification tools on that set of genes. Figure 4.6, 4.7, 4.8 display detailed comparisons for the root mean square error, Pearson correlation coefficient and Spearman correlation coefficient of these isoform quantification methods on all categories. As the number of isoforms for each gene increases, the accuracy improvement of Jlinks over other methods also increases, indicating that Jlink is superior to other methods for quantifying genes with multiple isoforms.

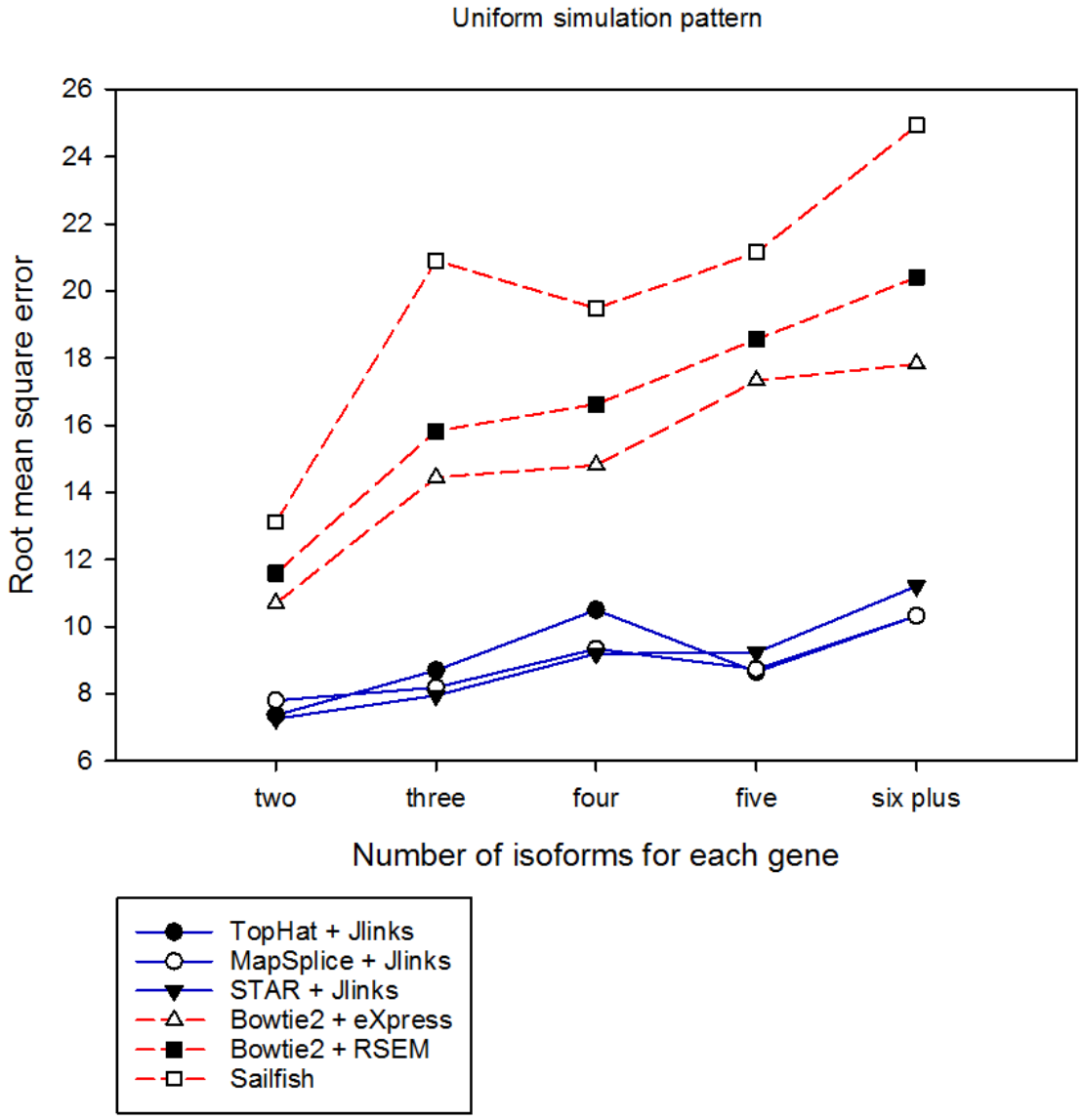

Figure 4.6: **Root mean square error comparison for uniform simulation pattern on various gene categories.** Root mean square errors of estimated isoform FPKM values compared with true FPKM values of these isoform quantification methods on various gene categories. Blue solid lines indicate pipelines using Jlinks as the quantification tool, red dash lines indicate pipelines using other quantification tools.

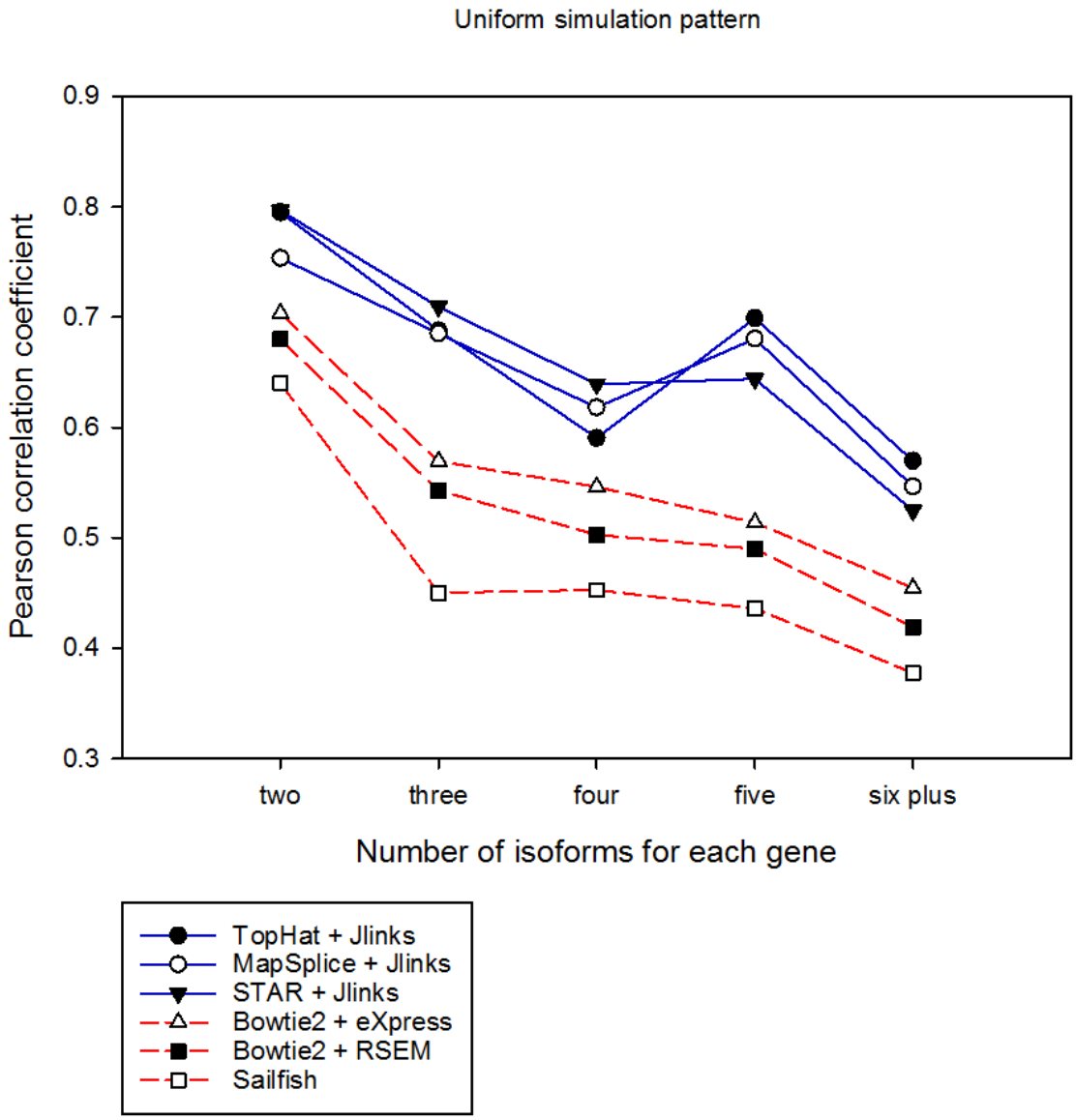

Figure 4.7: **Pearson correlation coefficient comparison for uniform simulation pattern on various gene categories.** Pearson correlation coefficients between estimated FPKM values and true FPKM values of these isoform quantification methods on various gene categories. Blue solid lines indicate pipelines using Jlinks as the quantification tool, red dash lines indicate pipelines using other quantification tools.

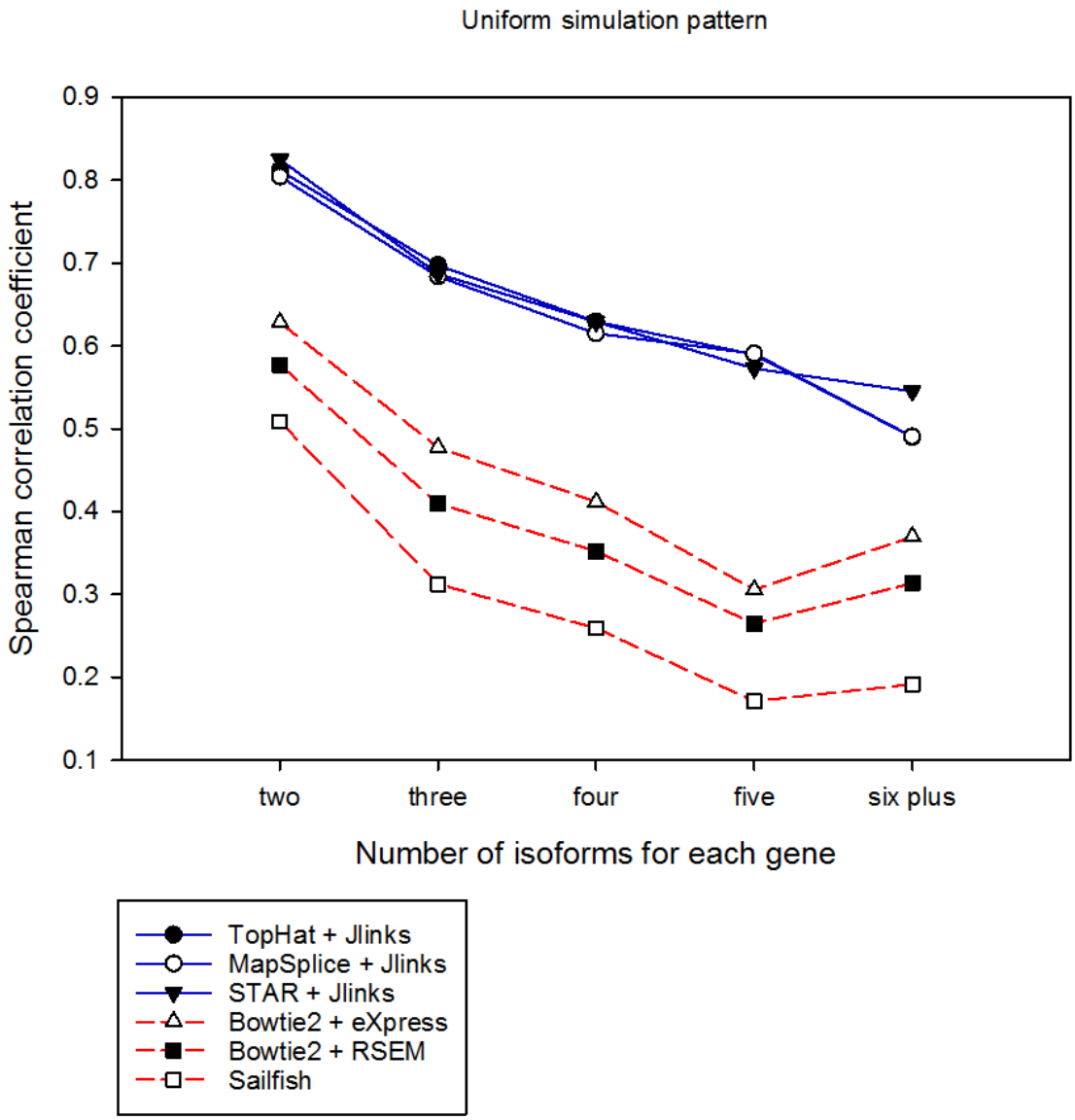

Figure 4.8: **Spearman correlation coefficient comparison for uniform simulation pattern on various gene categories.** Spearman correlation coefficients between estimated isoform FPKM values and true FPKM values of these isoform quantification methods on various gene categories. Blue solid lines indicate pipelines using Jlinks as the quantification tool, red dash lines indicate pipelines using other quantification tools.

4.3.2 Exponential simulation pattern

To test Jlinks performance in a more complex and realistic scenario, we simulated 100 million paired-end reads with an exponential simulation pattern: that is, the frequencies of isoforms being sampled follow an exponential distribution (Figure 4.9). Table 4.3 gives a summary of the performance of upstream alignment tools. Again, each alignment tool results in a mapping rate over 98%, demonstrating that the comparison of isoform quantification methods should not be greatly affected by the choices of alignment tools. Table 4.4 summarizes the overall accuracy comparison of the isoform quantification methods, and Figures 4.10, 4.11, 4.12 display the individual comparisons of root mean square error, Pearson correlation coefficient and Spearman correlation coefficient. We also examined the performances of these methods on various gene categories defined as in previous section, the comparison results are shown in Figures 4.13, 4.14, 4.15. The improvement in accuracy is even more evident when looking at genes with more isoforms, demonstrating Jlinks's strong advantage in quantifying genes with multiple isoforms.

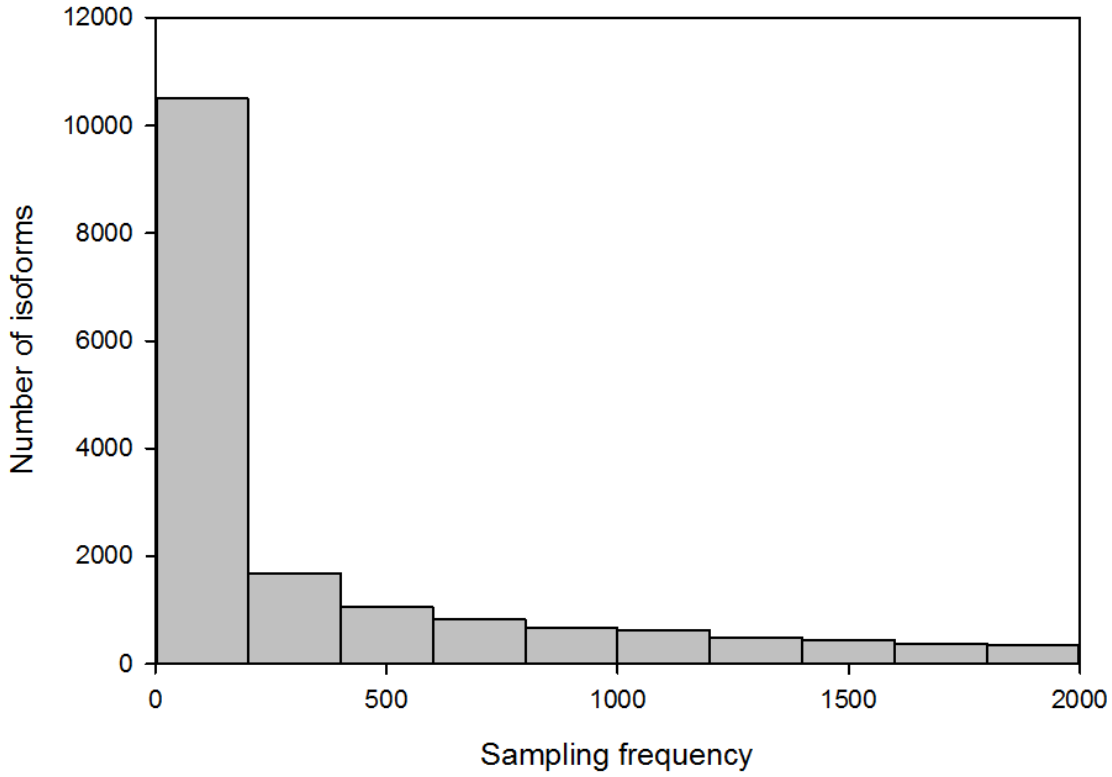

Figure 4.9: **Sampling frequency distribution in exponential simulation pattern.** The x-axis displays the frequency of isoforms being sampled during this simulation; the y-axis displays the number of isoforms having that sampling frequency.

Table 4.3: **Summary of upstream alignment tools for the exponential simulation pattern**

Aligner Type	Alignment Tool	Version	Options	Alignment Rate
Spliced aligner (align to genome)	TopHat	2.0.14	-G --no-novel-juncs --microexon-search --max-multihits=1	99.4%
	MapSplice	2.1.9	--gene-gtf --non-canonical --filtering=1	99.97%
	STAR	2.4.1d	--sjdbGTFfile	99.1%
Unspliced aligner (align to transcriptome)	Bowtie2	2.2.5	-a -X 1000 -v 3	98.61%

Table 4.4: **Overall accuracy comparison of isoform quantification methods for the exponential simulation pattern.**

Upstream alignment tool	Isoform quantification tool	Root mean square error	Pearson correlation coefficient	Spearman correlation coefficient
TopHat	Jlinks	14.1709	0.8689	0.7929
MapSplice		14.5775	0.8598	0.7902
STAR		15.4989	0.8045	0.7949
Bowtie2	eXpress	17.8476	0.8236	0.8035
	RSEM	19.3270	0.8008	0.7195
Sailfish		23.0211	0.7435	0.7076

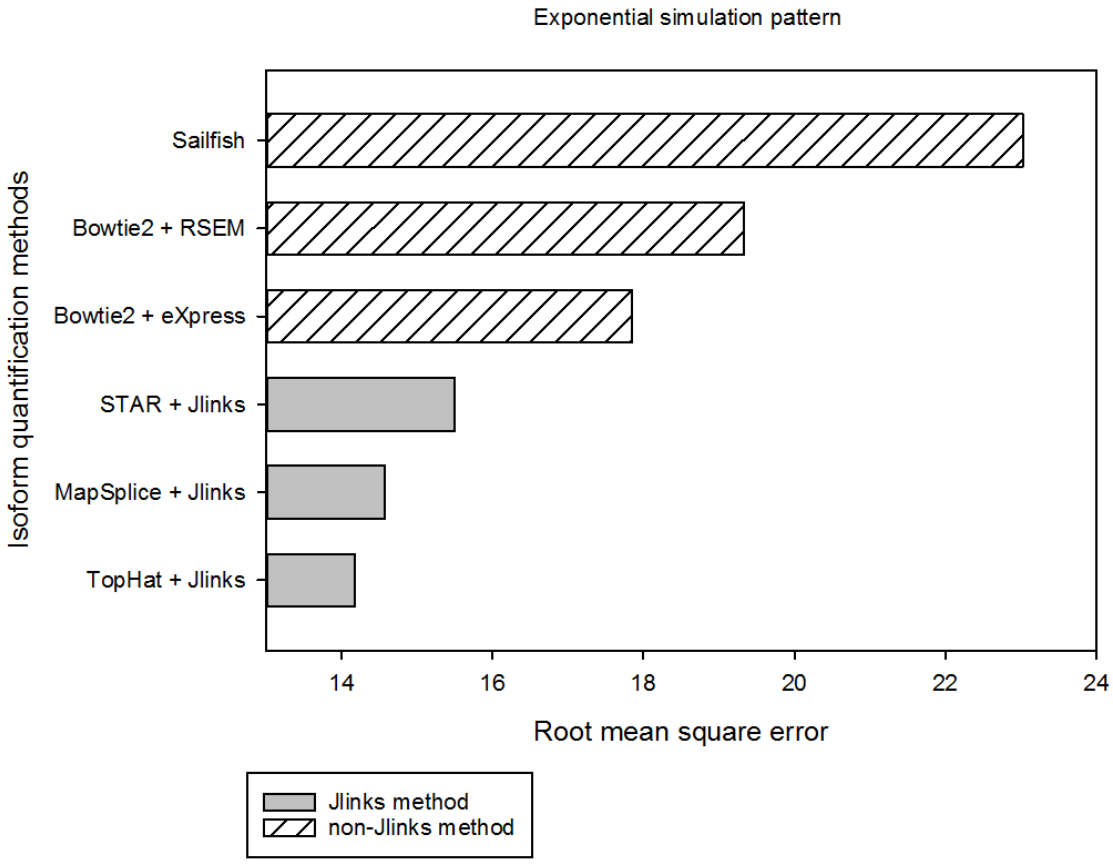

Figure 4.10: **Root mean square error comparison for exponential simulation pattern.** Root mean square errors of estimated isoform FPKM values compared with true FPKM values for these isoform quantification methods. Grey bars indicate pipelines using Jlinks as the quantification tool, white striped bars indicate pipelines using other quantification tools.

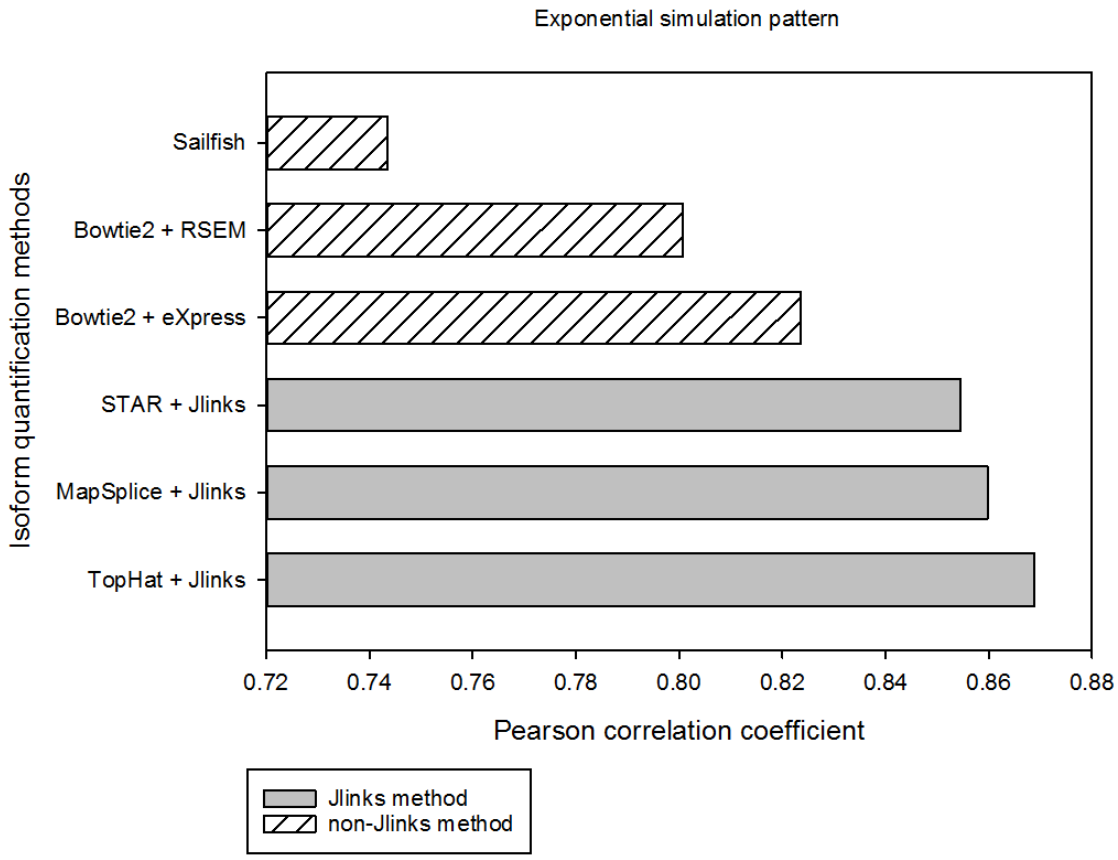

Figure 4.11: **Pearson correlation coefficient comparison for exponential simulation pattern.** Pearson correlation coefficients between estimated isoform FPKM values and true FPKM values for these isoform quantification methods. Grey bars indicate pipelines using Jlinks as the quantification tool, white striped bars indicate pipelines using other quantification tools.

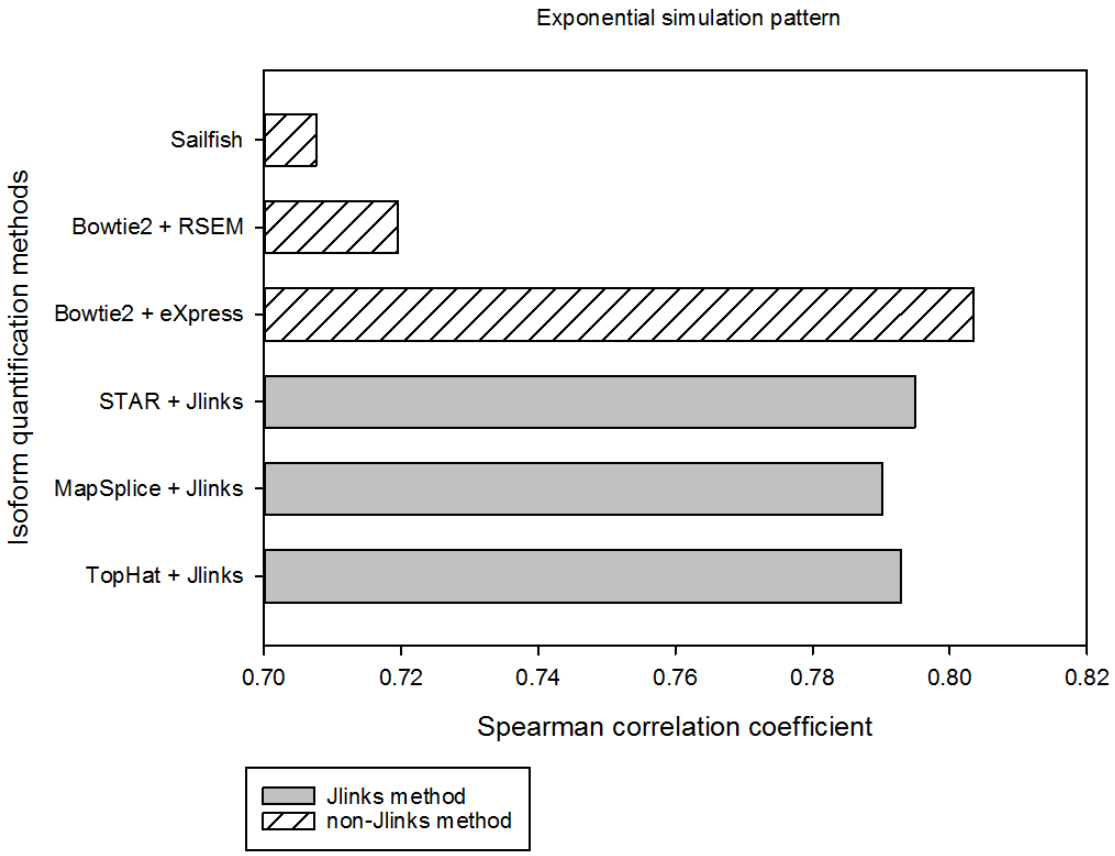

Figure 4.12: **Spearman correlation coefficient comparison for exponential simulation pattern.** Spearman correlation coefficients between estimated isoform FPKM values and true FPKM values for these isoform quantification methods. Grey bars indicate pipelines using Jlinks as the quantification tool, white striped bars indicate pipelines using other quantification tools.

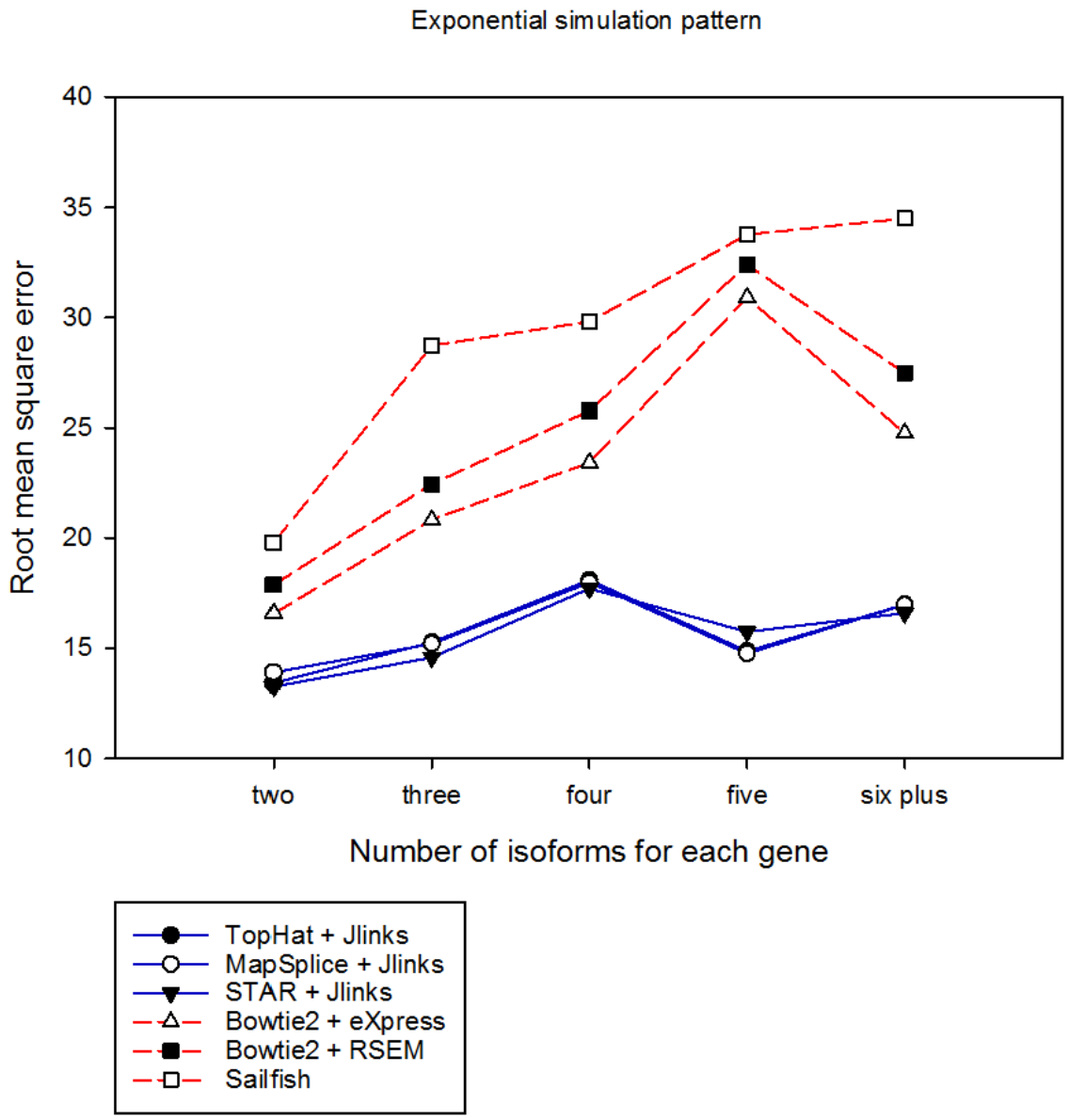

Figure 4.13: **Root mean square error comparison for exponential simulation pattern on various gene categories.** Root mean square errors of estimated isoform FPKM values compared with true FPKM values for these isoform quantification methods on various gene categories. Blue solid lines indicate pipelines using Jlinks as the quantification tool, red dash lines indicate pipelines using other quantification tools.

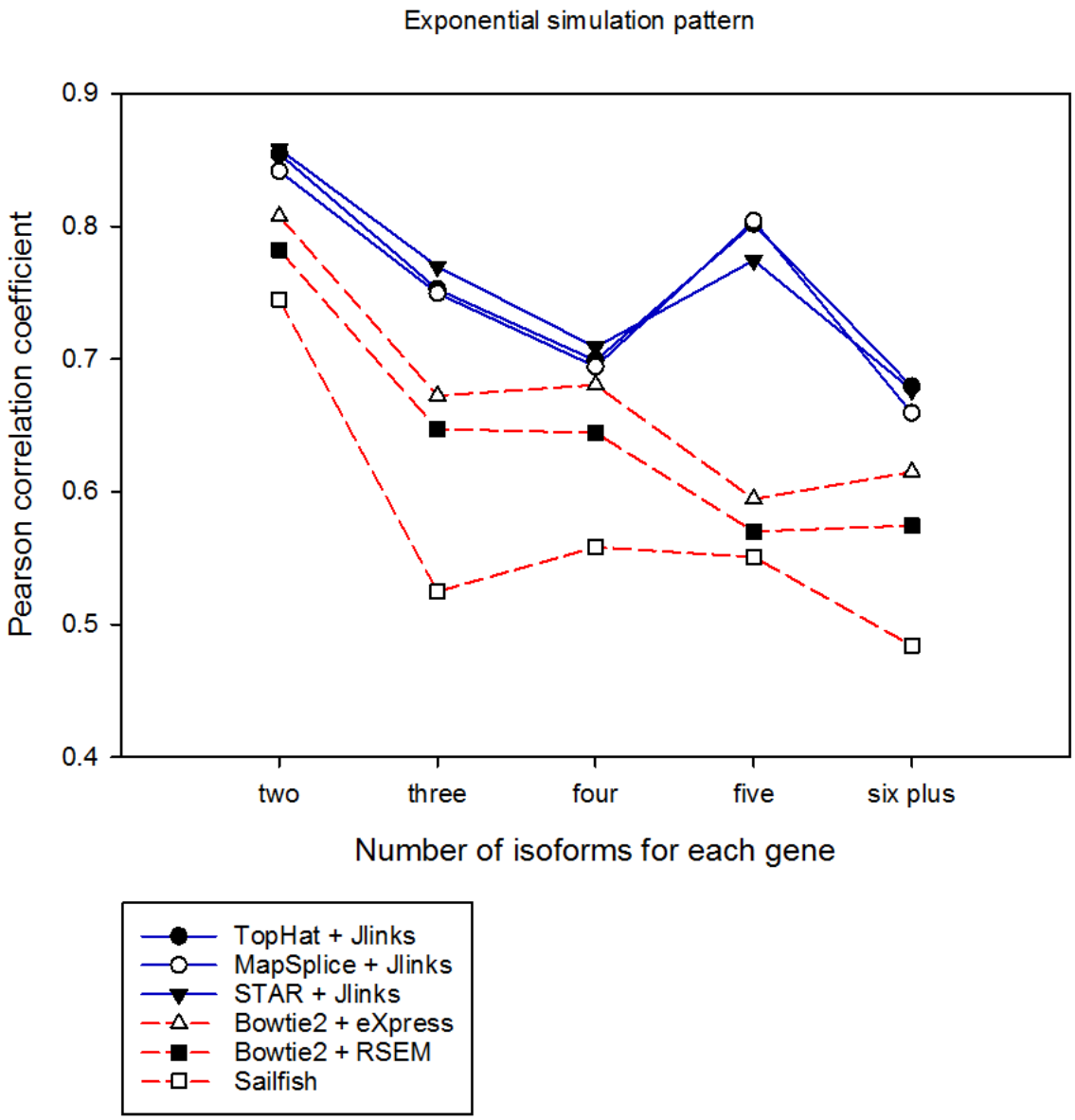

Figure 4.14: **Pearson correlation coefficient comparison for exponential simulation pattern on various gene categories.** Pearson correlation coefficients between estimated isoform FPKM values and true FPKM values for these isoform quantification methods on various gene categories. Blue solid lines indicate pipelines using Jlinks as the quantification tool, red dash lines indicate pipelines using other quantification tools.

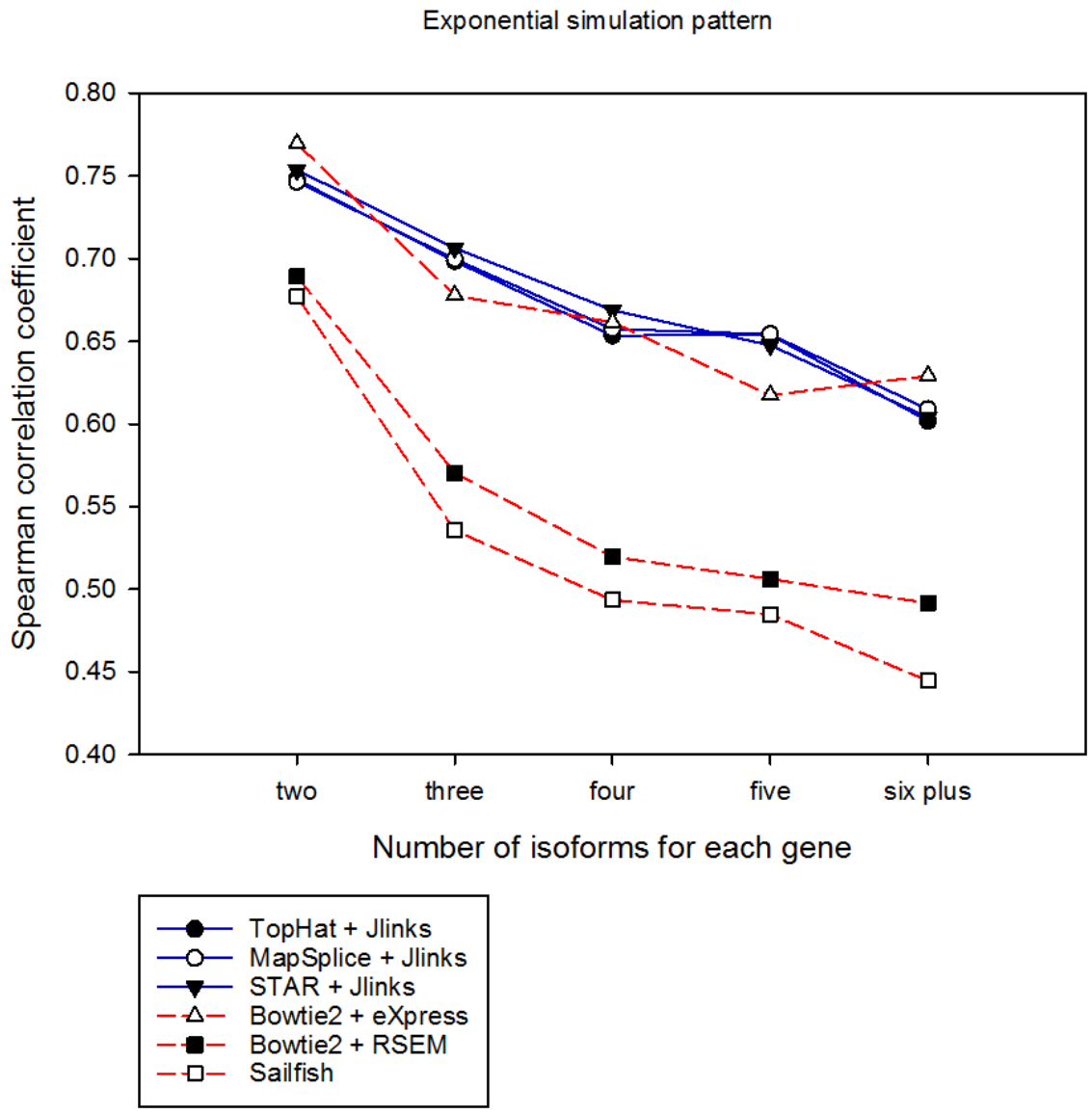

Figure 4.15: **Spearman correlation coefficient comparison for exponential simulation pattern on various gene categories.** Spearman correlation coefficients between estimated isoform FPKM values and true FPKM values for these isoform quantification methods on various gene categories. Blue solid lines indicate pipelines using Jlinks as the quantification tool, red dash lines indicate pipelines using other quantification tools.

4.3.3 Running time comparison

In addition to comparing the accuracies of quantification methods, we also measured their running times, as listed in Table 4.5. All quantification tools were run with 8 threads except for eXpress which does not have an option to set the number of threads. Sailfish is the fastest quantification tool with running time less than half an hour. eXpress and Jlinks have similar running times and RSEM is the most time-consuming tool in the comparison. It should be noted that the running times of eXpress and Jlinks are not completely comparable, as eXpress could not set the number of threads, whereas Jlinks could be faster if running with more threads. For instance, when running Jlinks with 16 threads on these two datasets, both running times were within an hour, less than those of eXpress.

Table 4.5: **Running time comparison of isoform quantification methods**

Upstream alignment tool	Isoform quantification tool	Quantification tool running time		Number of threads
		Uniform	Exponential	
TopHat	Jlinks	1h 48min 19s	1h 27min 36s	8
MapSplice		1h 48min 32s	1h 29min 07s	8
STAR		1h 46min 39s	1h 26min 29s	8
Bowtie2	eXpress	1h 42min 52s	1h 31min 22s	N/A*
	RSEM	2h 34min 06s	3h 08min 18s	8
Sailfish		0h 29min 54s	0h 16min 27s	8

*eXpress does not have an option to set the number of threads. The peak CPU usage for running eXpress is 280%

4.4 Performance comparison for paired-end RNA-Seq simulation datasets under various sequencing depths

To better evaluate the performances of Jlinks and other quantification methods, we generated simulation datasets of four other sequencing depths (20 million, 40 million, 60 million and 80 million paired-end reads) with both uniform and exponential simulation patterns. Figures 4.16-4.21 show the comparisons of these isoform quantification methods under various sequencing depths, in terms of root mean square error, Pearson correlation coefficient and Spearman correlation coefficient. eXpress provides the best Spearman correlation coefficient for datasets with the exponential simulation pattern, while in all the other cases Jlinks methods consistently outperform non-Jlinks methods in all the assessments.

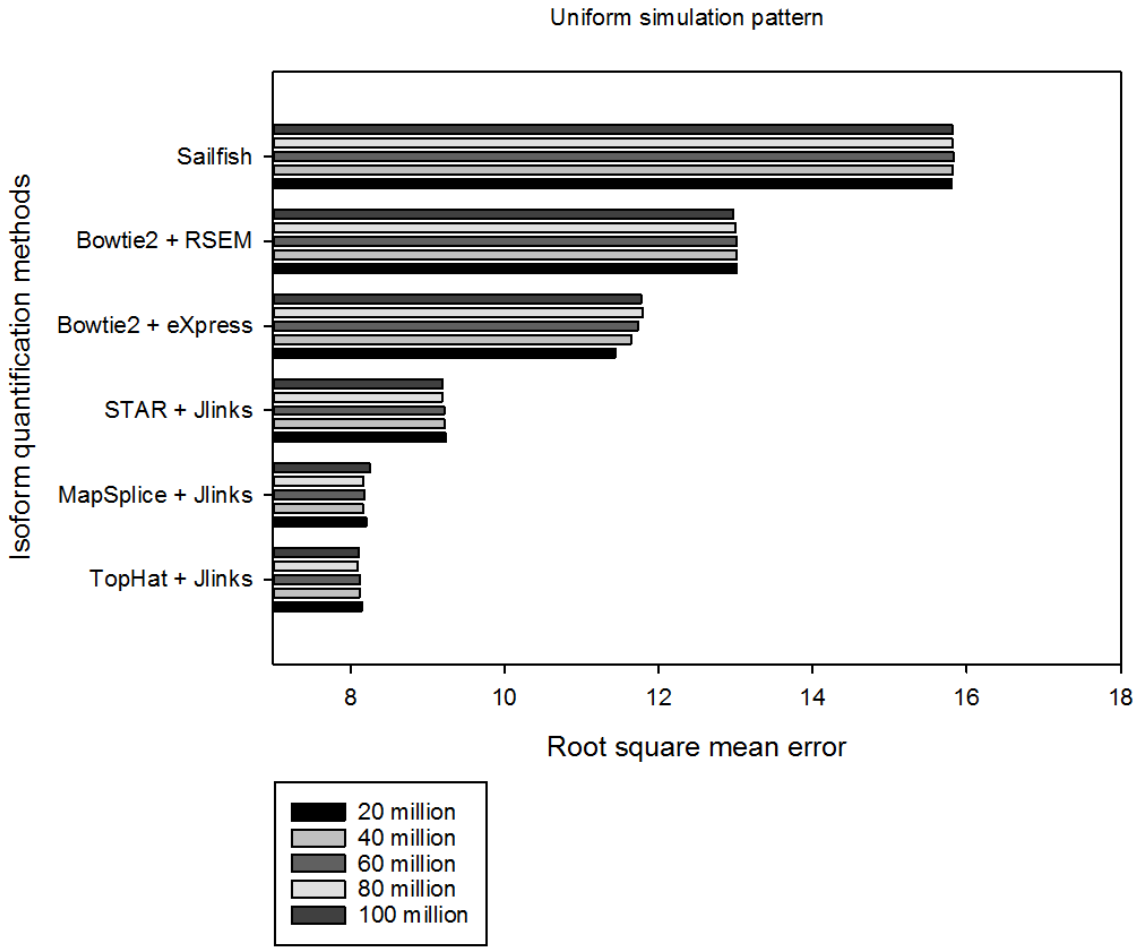

Figure 4.16: **Root mean square error comparison for uniform simulation pattern under various sequencing depths.** Root mean square errors of estimated isoform FPKM values compared with true FPKM values for these isoform quantification methods under various sequencing depths. All datasets were simulated with the uniform pattern.

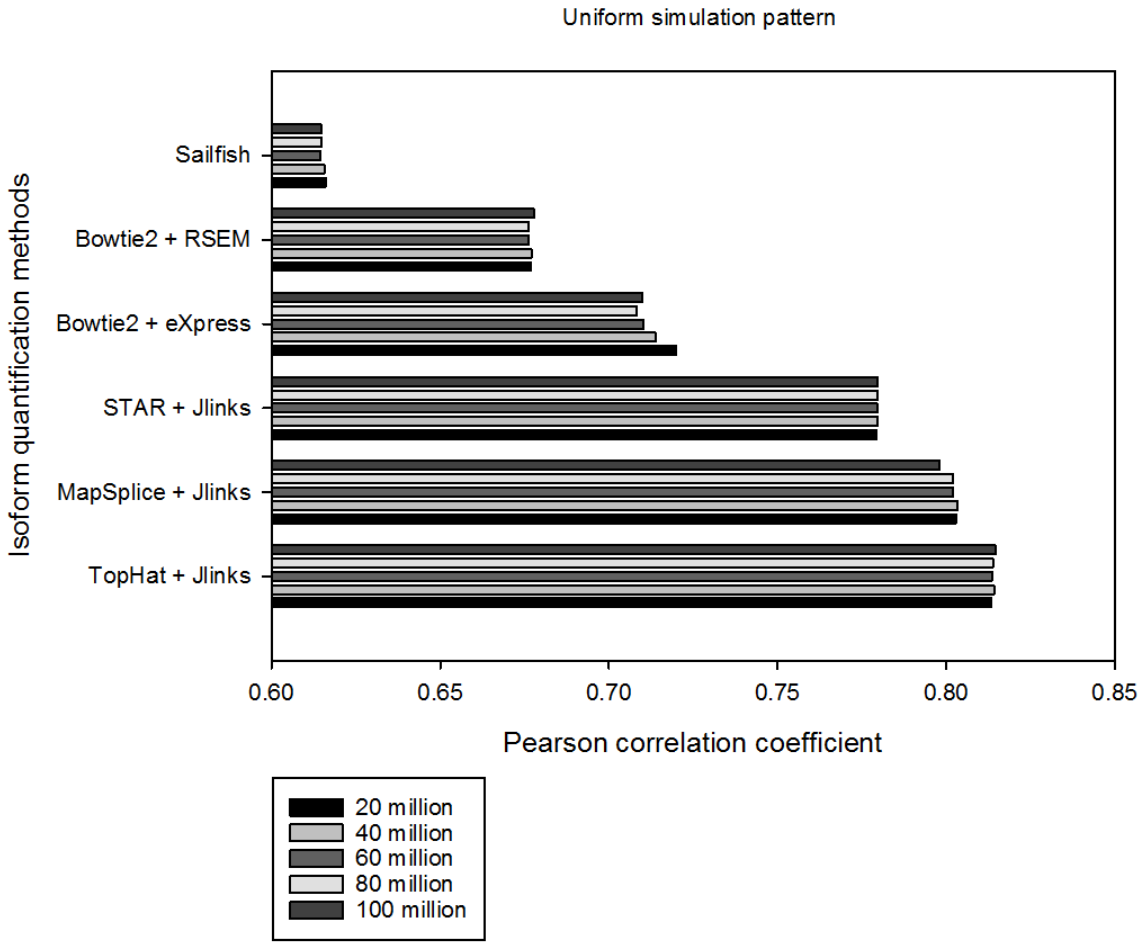

Figure 4.17: **Pearson correlation coefficient comparison for uniform simulation pattern under various sequencing depths.** Pearson correlation coefficients between estimated isoform FPKM values and true FPKM values for these isoform quantification methods under various sequencing depths. All datasets were simulated with the uniform pattern.

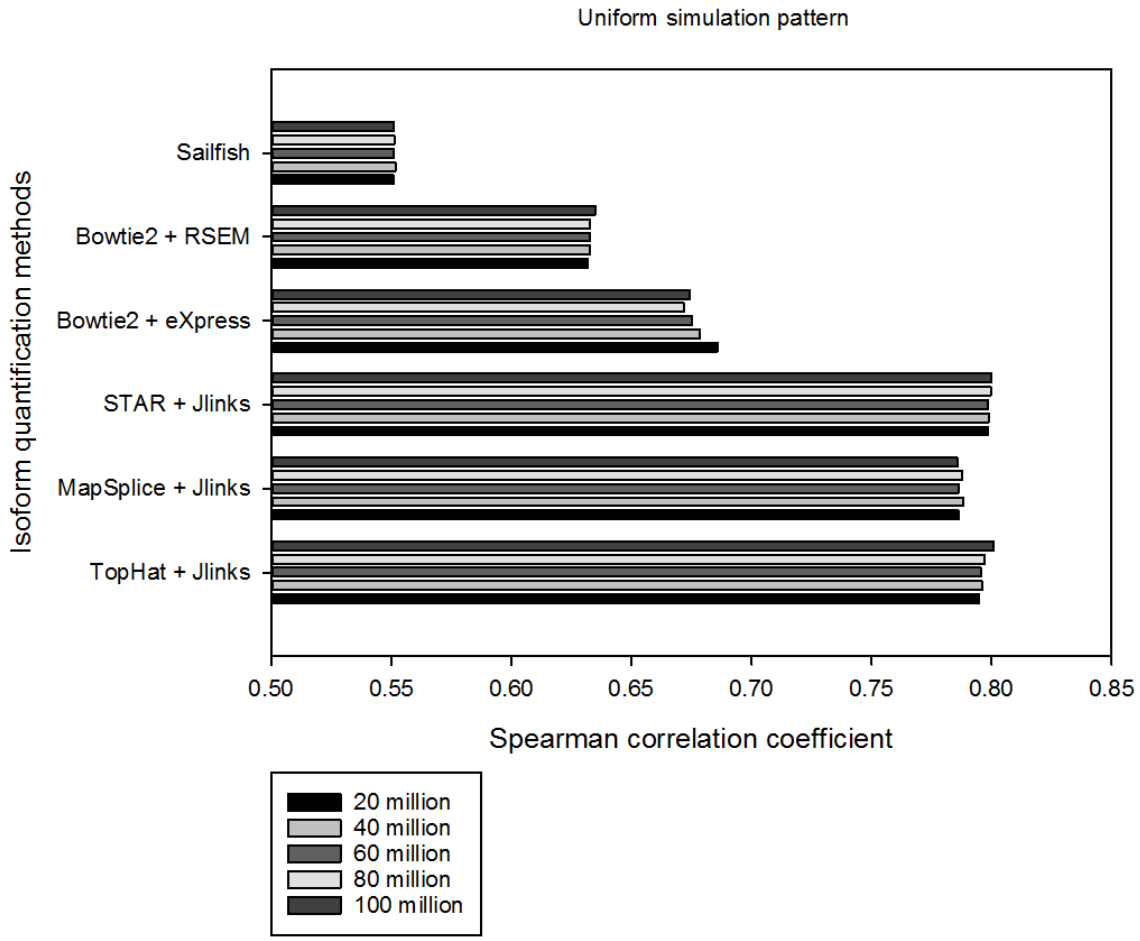

Figure 4.18: **Spearman correlation coefficient comparison for uniform simulation pattern under various sequencing depths.** Spearman correlation coefficients between estimated isoform FPKM values and true FPKM values for these isoform quantification methods under various sequencing depths. All datasets were simulated with the uniform pattern.

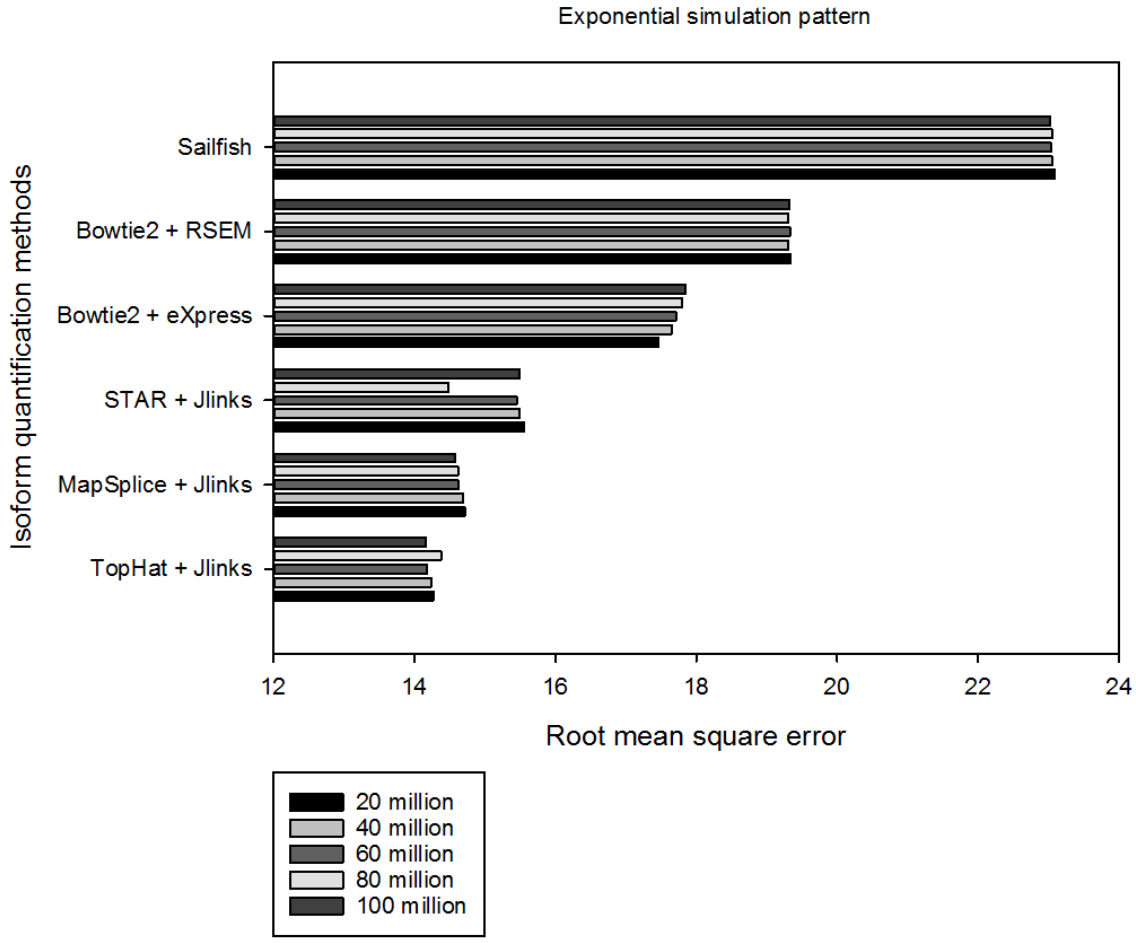

Figure 4.19: **Root mean square error comparison for exponential simulation pattern under various sequencing depths.** Root mean square errors of estimated FPKM values compared with true FPKM values for these isoform quantification methods under various sequencing depths. All datasets were simulated with the exponential pattern.

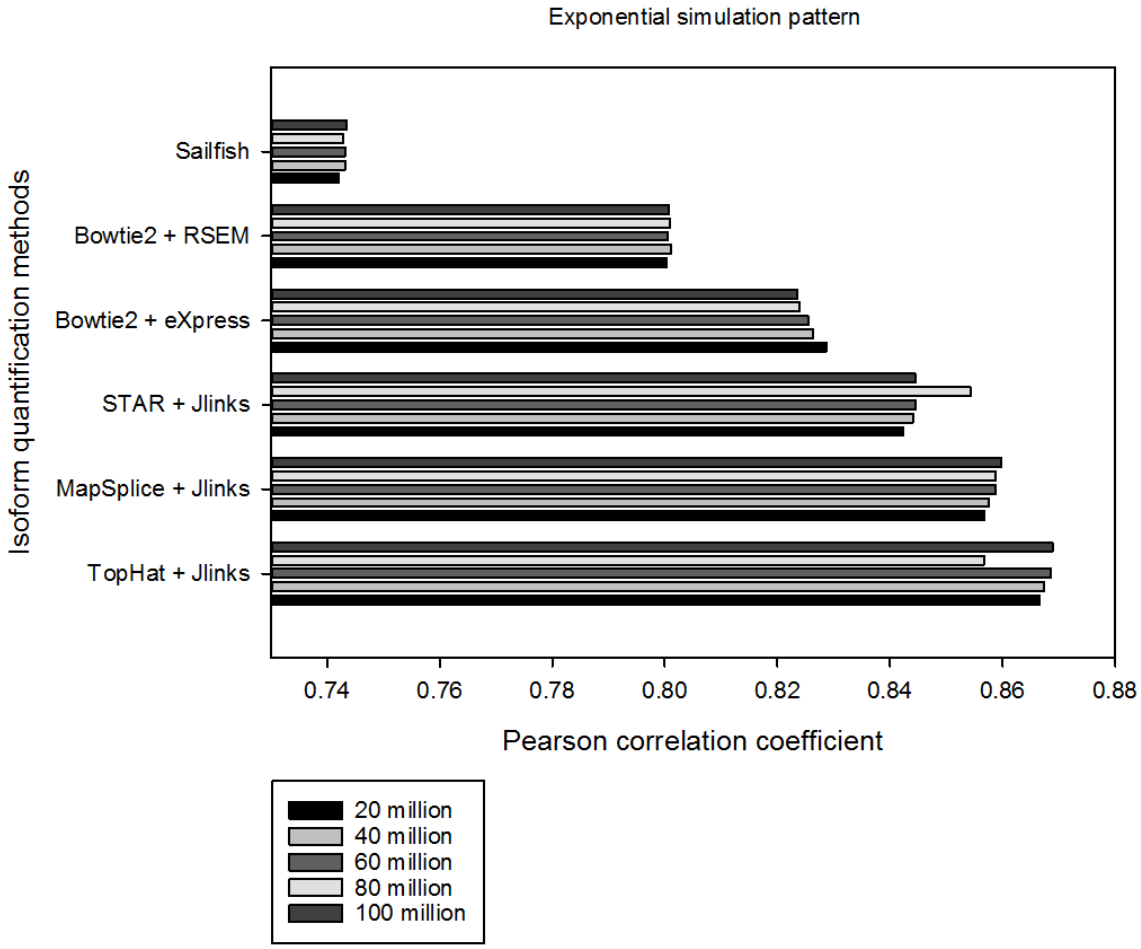

Figure 4.20: **Pearson correlation coefficient comparison for exponential simulation pattern under various sequencing depths.** Pearson correlation coefficients between estimated FPKM values and true FPKM values for these isoform quantification methods under various sequencing depths. All datasets were simulated with the exponential pattern.

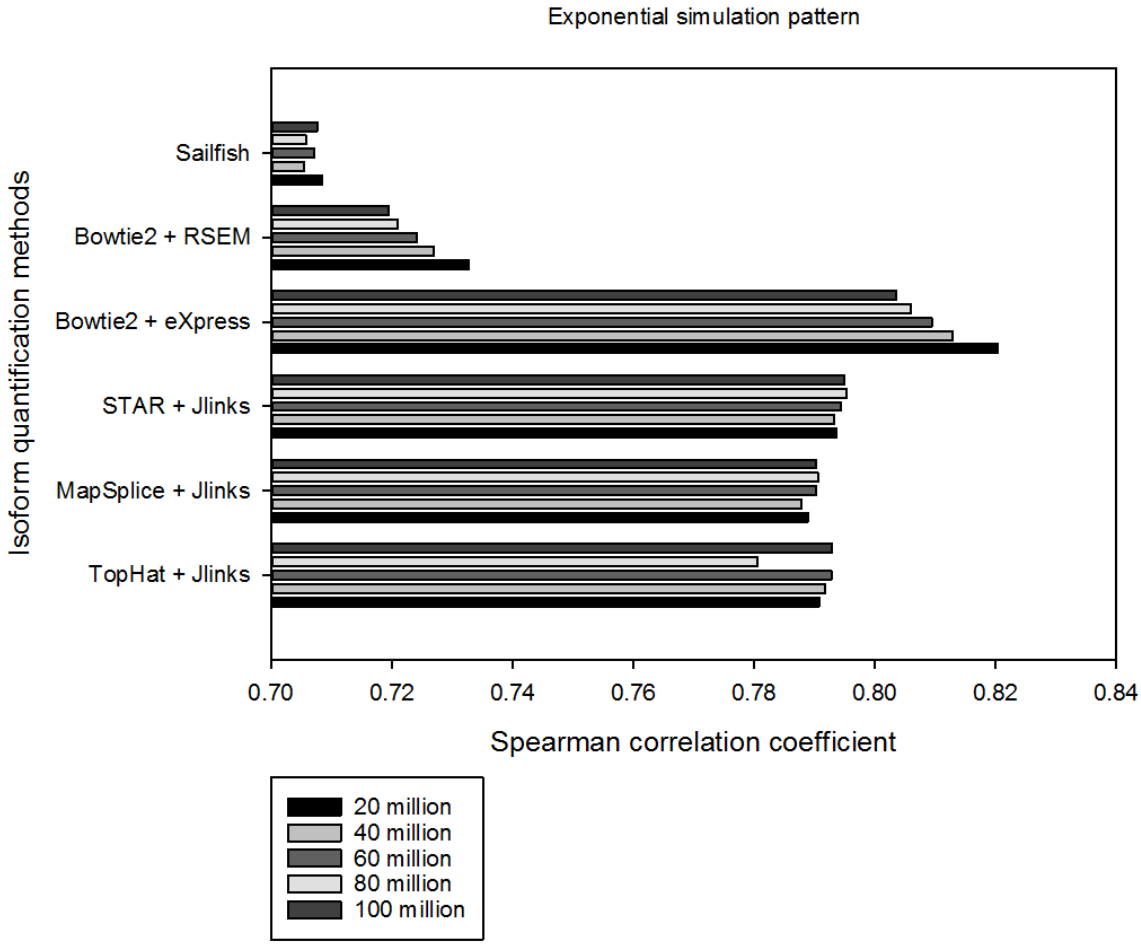

Figure 4.21: **Spearman correlation coefficient comparison for exponential simulation pattern under various sequencing depths.** Spearman correlation coefficients between estimated FPKM values and true FPKM values for these isoform quantification methods under various sequencing depths. All datasets were simulated with the exponential pattern.

5. Conclusion

In this paper we introduced a novel algorithm of isoform-level abundance estimation for a known set of isoforms. Our algorithm, referred to as Jlinks, treats each isoform as a unique "link" of splice junctions and converts the abundance estimation problem into obtaining an optimal solution for a linear system. Experiments on synthetic RNA-Seq datasets generated with both uniform and exponential simulation patterns under various sequencing depths demonstrate that Jlinks has superior performances relative to existing state-of-the-art isoform quantification methods.

Appendix

Proofs of the theorems applied in Jlinks algorithm:

For linear system $AX = C$, where A is a $n \times m$ matrix, X is an unknown vector with dimension $m \times 1$, C is a known vector with dimension $n \times 1$:

Condition 1. If $rank(A) = m = n$, the problem has a unique solution

$$X = A^{-1}C$$

Proof.

When $rank(A) = m = n$, A is a full rank square matrix, thus has inverse matrix A^{-1} satisfying $A^{-1}A = I$. $AX = C \Longrightarrow A^{-1}AX = A^{-1}C \Longrightarrow X = A^{-1}C$

Condition 2. If $rank(A) = m < n$, the problem has a unique least-squares solution

$$X = (A^T A)^{-1}A^T C$$

Proof.

In this case there are more constraints than unknowns, the system is over-determined and has no exact solution. We can obtain a least-squares solution that minimizes the error. We want to find X that minimizes

$$\|C - AX\|^2$$

or

$$(C - AX)^T(C - AX)$$

or

$$C^TC - C^TAX - X^TA^TC + X^TA^TAX$$

Differentiating w.r.t. X and setting the result equal to zero yields

$$-(C^TA)^T - (A^TC) + 2A^TAX = 0$$

so

$$X = (A^TA)^{-1}A^TC$$

where $(A^TA)^{-1}A^T$ is called left pseudo-inverse of A.

Condition 3. If $rank(A) = n < m$, the problem has a unique minimum-norm least-squares solution

$$X = A^T(AA^T)^{-1}C$$

Proof.

In this case there are fewer constraints than unknowns, the system is under-determined and has infinite amount of solutions. We can pick one with the minimum norm. That is, we will minimize $\|X\|^2$ subject to the constraint $AX = C$ using Lagrange multiplier method, which becomes

$$\|X\|^2 + \lambda^T(C - AX)$$

Differentiating w.r.t. X and setting the result equal to zero yields

$$2X - A^T\lambda = 0$$

so

$$2AX - AA^T\lambda = 0$$

and using $C = AX$ gives us

$$2C = AA^T\lambda$$

so

$$\lambda = 2(AA^T)^{-1}C$$

and hence

$$X = A^T(AA^T)^{-1}C$$

where $A^T(AA^T)^{-1}$ is called right pseudo-inverse of A.

Rank Factorization Theorem: Any $n \times m$ matrix A of rank r can be decomposed as $A = FG$, where F is a $n \times r$ full column rank matrix, G is a $r \times m$ full row rank matrix.

Proof.

Since $rank(A) = r$, A has r linearly independent column vectors $a_{i_1}, a_{i_2}, \dots, a_{i_r}$. Denote $F = (a_{i_1}, a_{i_2}, \dots, a_{i_r})$, so F is a $n \times r$ full column rank matrix.

Each column of A is a linear combination of column vectors of F. That is, there exists a $r \times m$ matrix G satisfying $A = FG$.

Now we have

$$r = rank(A) = rank(FG) \leq rank(G) \leq r$$

so

$$rank(G) = \square$$

So G is a $r \times m$ full row rank matrix.

Condition 4. If $rank(A) < min(n, m)$, the problem has a unique minimum-norm least-squares solution

$$X = G^T(GG^T)^{-1}(F^TF)^{-1}F^TC$$

where $A = FG$ is a rank factorization of matrix A.

Proof.

Suppose $rank(A) = r$, according to *Rank Factorization Theorem*, F is a $n \times r$ full column rank matrix and G is a $r \times m$ full row rank matrix. The problem becomes

$$FGX = \square$$

where $rank(F) = r < n$. Applying the solution in Condition 2, we have

$$GX = (F^T F)^{-1} F^T C$$

where $rank(G) = r < m$. Applying the solution in Condition 3, we have

$$X = G^T (GG^T)^{-1} (F^T F)^{-1} F^T C$$

Bibliography

[1] Black, Douglas L. (2003) “Mechanisms of alternative pre-messenger RNA splicing”. Annual Reviews of Biochemistry 72 (1): 291-336.

[2] Pan, Q; Shai O; Lee LJ; Frey BJ; Blencowe BJ. (2008) "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing". *Nature Genetics* 40 (12): 1413–1415.

[3] Lopez-Bigas N. *et al.* (2005) “Are splicing mutations the most frequent cause of hereditary disease?”. *FEBS Lett.* 579 (9): 1900-3.

[4] Wang ET. *et al.* (2008) “Alternative isoform regulation in human tissue transcriptomes”. *Nature* 456 (7221): 470-6.

[5] Marra, M. *et al.* (1999) “An encyclopedia of mouse genes”. *Nat Genet* 21: 191–194.

[6] Carninci, P. *et al.* (2003) “Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia”. *Genome Res* 13: 1273–1289.

[7] de Souza, S. J. *et al.* (2000) “Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags”. *Proc Natl Acad Sci USA* 97: 12690–12693.

[8] Guttman, M. *et al.* (2009) “Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals”. *Nature* 458: 223–227.

[9] Wang Z; Gerstein M; Snyder M. (2009) “RNA-Seq: a revolutionary tool for transcriptomics”. *Nature Reviews Genetics* 10 (1): 57-63.

[10] Nicolae M. *et al.* (2011) “Estimation of alternative splicing isoform frequencies from RNA-Seq data”. *Algorithm Mol Biol* 6 (1): 9.

[11] Bo Li; Colin N Dewey. (2011) “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. *BMC Bioinformatics* 12: 323.

[12] Adam Roberts; Lior Pachter. (2013) “Streaming fragment assignment for real-time analysis of sequencing experiments”. *Nature Methods* 10: 71-73.

[13] Rob Patro. *et al.* (2014) “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms”. *Nature Biotech* 32: 462-4.

[14] Sanger F; Coulson AR. (1975) “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. *J Mol Biol* 94 (3): 441-8.

[15] Yamada K. *et al.* (2003) “Empirical analysis of transcriptional activity in the Arabidopsis genome”. *Science* 302 (5646): 842-6.

[16] Kim E; Goren A; Ast G. (2008) "Insights into the connection between cancer and alternative splicing". *Trends Genet* 24 (1): 7–10.

[17] Omenn, GS; Guan, Y; Menon, R. (2014) "A New Class of Protein Cancer Biomarker Candidates: Differentially-Expressed Splice Variants of ERBB2 (HER2/neu) and ERBB1 (EGFR) in Breast Cancer Cell Lines". *Journal of proteomics* 107C: 103–112.

[18] Ben Langmead. *et al.* (2009) "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". *Genome Biology* 10: R25.

[19] Li H; Durbin R. (2009) "Fast and accurate short read alignment with Burrows-Wheeler transform". *Bioinformatics* 25 (14): 1754-60.

[20] Cole Trapnell; Lior Pachter; Steven L. Salzberg. (2009) "TopHat: discovering splice junctions with RNA-Seq". *Bioinformatics* 25 (9): 1105-1111.

[21] Wang K. *et al.* (2010) "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery". *Nucleic Acids Res* 38 (18): e178.

[22] Alexander Dobin. *et al.* (2012) "STAR: ultrafast universal RNA-seq aligner". *Bioinformatics* 10: 1093.

[23] Brian J. Haas. *et al.* (2013) "*De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity". *Nat Protoc* 10 (8): 1038.

[24] Marcel H. Schulz. *et al.* (2012) "*Oases:* robust *de novo* RNA-seq assembly across the dynamic range of expression levels". *Bioinformatics* 28 (8): 1086-1092.

[25] Gordon Robertson. *et al.* (2010) "*De novo* assembly and analysis of RNA-seq data". *Nature Methods* 7: 909-912.

[26] Feng J. *et al.* (2011) "Inference of isoforms from short sequence reads". *J Comput Biol* 18 (3): 305-21.

[27] Guttman M. *et al.* (2010) "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs". *Nat Biotechnol* 28 (5): 503-10.

[28] Trapnell C. *et al.* (2010) "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". *Nat Biotechnol* 28 (5): 511-5.

[29] Trapnell C. *et al.* (2013) "Differential analysis of gene regulation at transcript resolution with RNA-seq". *Nat Biotechnol* 31: 46-53.

[30] Simon Anders; Wolfgang Huber. (2010) "Differential expression analysis for sequence count data". *Genome Biology* 11: R106.

[31] Mark D. Robinson. *et al.* (2010) "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* 26 (1): 139-140.

[32] Ali Mortazavi. *et al.* (2008) "Mapping and quantifying mammalian transcriptomes by RNA-Seq". *Nature Methods* 5: 621-628.

Vita

Jingyi Lu received her B.S. degree in Mathematics and Physics from Tsinghua University in 2012. During her college period she won the Excellent Freshman Award in 2008, the National Scholarship in 2009, the Samsung Scholarship in 2010 and the Microsoft Young Fellowship in 2011. She enrolled in the Biomedical Engineering master program at Johns Hopkins University in 2013, where she worked as a Teaching and Research Assistant. Her research area is computational biology and bioinformatics, including RNA-Seq, data mining, algorithm design and software development.