BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–7

Statistics for K-mer Based Splicing Analysis

Ruofei Du, Hao Li, Hui Miao and Shangfu Peng

[1]Department of Computer Science, University of Maryland, College Park.

Team: Data Learner Miner Practitioner

CMSC 702 Project Report

ABSTRACT

It is well acknowledged that alternative splicing module plays a crucial role to identify the variations of the RNA transcriptomes. In high-throughput short-read RNA, splicing analysis is a challenging task due to the uncertainty and time complexity of reads alignments onto genome and transcriptome.

In this paper, we introduce k-mer based statistical method for splicing event analysis. The k-mer based representation avoids time-consuming reads alignment, and the significant differential k-mers between controlled group of samples are a good indicator of existence of certain types of splicing events. We explored statistical models including t-test, DESeq and likelihood ratio test to identify statistical significant differential k-mers. We also develop a faset k-mer mapping method instead of Bowtie for identifying whether a k-mer from reads data can be matched on genome or transcriptome.

1 INTRODUCTION

Alternative splicing is a regulated process during gene expression that results in a single gene coding for multiple proteins. Similar to the isoform abundance difference, the difference in alternative splicing events has been shown as an important method to tell the variations of the RNA transcriptomes, which may improve understanding of cell differentiation and classify disease types [3, 5, 6, 11]. Splicing event is often classified as different alternative splicing module (ASM). From [12], there are five traditional basic modes of alternative splicing events: Exon skipping or cassette exon, Mutually exclusive exons, Alternative donor site, Alternative acceptor site, and Intron retention. These five modes describe basic splicing mechanisms, but is inadequate to describe complex splicing events [4].

High-throughput short-read RNA sequencing technologies provide in-depth and high-speed sampling of the transcriptome, however, the outputs of the technique introduce uncertainty, and the time required for analysis has not kept up with the pace of data generation. Thus scalable method is greatly desired. Recently, k-mer and k-mer index have been shown as a very effective method to accelerate certain bio-statistics task, such as detecting isoform abundance [8, 9]. K-mer is a fixed sized (K) sequence of nucleic acid or amino acid bases [13]. K-mer index compresses the sequence by counting the co-occurrence and mapping each transcript to the set of its k-mers and mapping each k-mer to the sets of transcripts including it. In Sailfish [8], Patro, et al. use k-mer instead of the computationally intensive reads mapping method to develop a statistical method and find abundance of isoforms with an order magnitude performance gain.

In this project, our task is to investigate the scalable statistic methods to find the significant differential k-mer to infer alternative splicing events from RNA-Seq reads. Our K-mer based method does not require reads alignment but calculate K-mers counts for RNA-Seq reads. We explore several statistical methods based on k-mer counts to identify significantly differential counts including t-test, DESeq, and likelihood ratio test. Additionally, we develop fast method for k-mer matching, which is 2 orders of magnitude faster than Bowtie. To the best of our knowledge, we conducted the first study on k-mers based statistical model for splicing event analysis.

2 PROBLEM STATEMENT

Given RNA-Seq reads from a total of S treated and untreated samples, we can find all the unique k-mers (size K) appeared in the each samples and count their frequencies. Then we have a $K \times S$ matrix of k-mer counts. The problem is how to identify some unusual splicing events by inferring this frequency matrix.

A k-mer from sequencing data can be mapped to three groups: (1) in the genome (introns and exons); (2) in the transcriptome (exon-exon-junction); (3) only in the reads data. The k-mers mapped to genome are not useful for splicing event analysis. The k-mers in the second group are possibly generated from splicing events of "Exon skipping" and "Mutually exclusive exons". The k-mers that cannot be find in the genome or transcriptome may be generated from splicing events of "Alternative donor site", "Alternative acceptor site", "Intron retention" and the other anomalies.

Since an alternative splicing events could generate unique k-mers that cannot find in genome or transcriptome (generated by other splicing events), some k-mers in treated samples may have significant differential frequencies comparing to untreated samples. Thus the appearance of a certain k-mer with great count difference between the two groups could be a strong indicator for the existence of a certain splicing event in the treated samples.

Though the abnormal k-mers that do not appear in genome or transcriptome can be identified by a brute force matching, those k-mers could be generated by sequencing errors. To eliminate false discoveries, statical methods need to be developed for robust abnormal k-mer identification.

The workflow of our approach is as follows: first, build k-mer index for all the reads data and count their frequency; second, find the statistically significant differential k-mers set; third, map these significant differential k-mers on the three groups; last we can use the unmapped k-mers to identify splicing events.

3 DATA DESCRIPTION

We conducted our experiments on the RNA-Seq data published by Brooks et al. [2]. The datasets are publicly available in the NCBI Gene Expression Omnibus under the accession GSE18508[1]. As shown in Table 1, we used "S2_DRSC_Untreated-3,4" as an untreated control group and "S2_DRSC_CG8144_RNAi-3,4" as an experimental group with the gene CG8144 of pasilla knocked down. Later we named the sample 1 and 2 as Treated1 and Treated2, sample 3 and 4 as Untreated1 and Untreated2. Each group contains two biological replicates. The gene CG8144 is known to bind to mRNA in the spliceosome, and is thought to be involved in the regulation of splicing.

In this project, our goal is to identify the patterns of the special splicing events caused by the gene CG8144 knocked down. A k-mer based method is to find the significant differential k-mers between the treated group and the untreated group.

Table 1. Read numbers statistics of the datasets

id	filer	type	number of reads
1	S2_DRSC_CG8144_RNAi-3	paired-end	$4.9 * 10^7 (\times 2)$
2	S2_DRSC_CG8144_RNAi-4	paired-end	$4.1 * 10^7 (\times 2)$
3	S2_DRSC_Untreated-3	paired-end	$3.8 * 10^7 (\times 2)$
4	S2_DRSC_Untreated-4	paired-end	$4.5 * 10^7 (\times 2)$

3.1 Data Pre-processing

3.1.1 K-mer counts matrix Given a set of RNA-Seq reads, we first calculate k-mer counts instead of aligning them to the annotated gene. Each reads set can be represented as a column that each cell is the number of times the corresponding k-mer appearing in the whole reads set. Then several RNA-Seq samples are represented as a matrix $C \in Z^{K \times S}$, where K is the k-mer sequences ordered by the value calculated by perfect hash function H_K, and S is the set of samples one of which is composed of RNA-Seq reads. We will choose a reasonable k (e.g. 25) to eliminate the possibility that one k-mer can be mapped to multiple places in the RNA while tolerating to errors.

In detail, we first preprocessed the RNA-Seq sample reads set Ψ using k-mer counts instead of mapping to the annotated gene or isoforms and at the same time construct a perfect hash H_K to map a k-mer string to an integer. The H_K we used is to map a gene string to a base 4 numbers, which A is 0, C is 1, G is 2, and T is 3. So the range of H_K is $[0, 4^{25}]$.

Considering the fact that smaller k-mers are not unique enough to be distinguished while larger k-mers cost too much disk and memory space to store, we choose a k-mer length $k = 25$ as the default parameter in our exploration. Given the k-mer length k, we computed a perfect hash function H_K on the set of $kmers(\Psi)$. In GSE18508 data, $|S| = 4$. The number of appeared k-mers is around 112.9 millions.

[1] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18508

3.1.2 Normalization Since the number of reads are different in the 4 reads samples, it requires a method to normalize the counts matrix. In this project, we tried two normalization methods. The first one is to calculate the k-mer ratio in its reads sample. The second one is the same normalization method from DESeq [1]. In the rest of the paper, we use the DESeq normalization method.

K-mer ratio. For each k-mer, we calculate its count ratio in its reads sample defined as follows:

$$r_{ij} = \frac{C_{ij}}{\sum_i C_{ij}}, j \in [1, 4] \quad (1)$$

DESeq. We first calculate the geometric mean for each row of C. Then we take the median of the ratios of observed counts as the size factors s_j of the corresponding samples j.

$$s_j = median_i(\frac{C_{ij}}{r_i}) \; for \; i \; where \; r_i \neq 0$$

$$r_i = (\prod_{j=1}^{j=4} C_{ij})^{1/4}$$

Here we give a sample distribution of $\frac{C_{ij}}{r_i}$ on Treated 1 in Figure 1. The distributions of other samples are similar to this one. Table 2 shows the medium s_j of the four samples, which are the size factors. The normalized matrix is calculated by the counts matrix that each column divides its size factor, e.g., $C_{ij}^N = \frac{C_{ij}}{s_j}$.

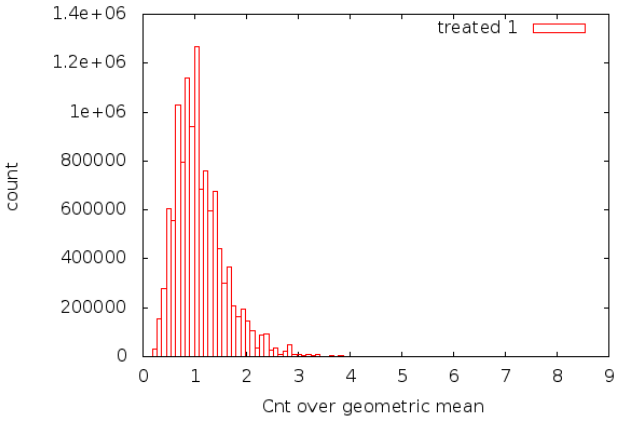


Fig. 1. The distribution of counts over geometric means on Treated 1

Table 2. The size factors of the four samples

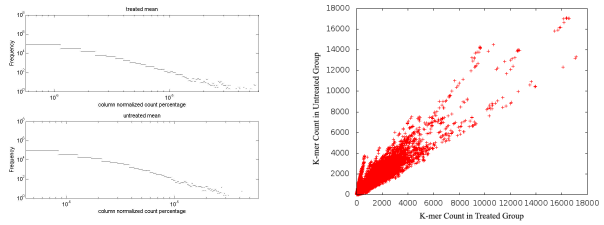
Treated1	Treated2	Untreated1	Untreated2
1.000	1.092	0.931	1.003

3.2 Data Exploration

3.2.1 K-mer Distribution In order to build a sound statistical model, we look into the distribution of k-mer counts, and the

correlation between the k-mer counts of the same k-mer in untreated and treated group.

In Fig. 3(a), we show the distribution of normalized k-mer count. The x axis is the k-mer counts in log scale, and the y axis is the frequency in log scale. As we can see, it is close to linear shape and the heavy long tail show a powerlaw distribution. We can use Pareto type I distribution, $Pareto(1, k)$, where k is the skewness shape parameter, to describe the data. Interestingly, the skewness of both groups are very similar. We also investigate the rank based distribution, and it follows a powerlaw distribution as well, in other words, the k-mer count follows Zipf distribution.



(a) Normalized k-mer count distribution for treated and untreated samples.

(b) K-mer correlation between untreated and treated group among the three samples

Fig. 3. Data Exploration Result

Regarding the flat head of the log-log scaled distribution, we listed the k-mers in Table 3.

3.2.2 K-mer Covariance Intuitively, large amount of k-mer index ratio should be the same in treated and untreated group. Thus the counts of the same k-mer in two groups are not independent. We verify our proposition by plotting the covariance in Fig. 3(b), as we can see the shape is radical shape. Each dot in the plot is a 2-dimensional vector. Also the Pareto distribution of the k-mer counts can be seen from this figure as well, as the density of the dots are very skewed along the diagonal.

From the k-mer distribution analysis and the covariance analysis, we investigate to use a bivariate Pareto Prior to describe the k-mer matrix.

In Figure 2(a) and Figure 2(b), we explored the dependence of variance and squared variance coefficient versus mean.

The inference in DESeq relies on an estimation of the relationship between the k-mer counts' variance and their mean, or equivalently the data's dispersion and their mean. The dispersion could be understood as the square of the coefficient of biological variation. Specifically, the dispersion is defined as:

$$\alpha = \frac{v - s\mu}{s^2\mu^2},$$

where s, v, μ correspond to the size factor, variance and mean for each k-mer.

From the scatter plots of dispersion versus mean of the k-mer counts table in Figure 2(c), we observed that when mean is smaller than $K = 100$, the dispersion decreases as mean rises; as the mean is large enough, the dispersion remains the same.

4 STATISTIC METHOD

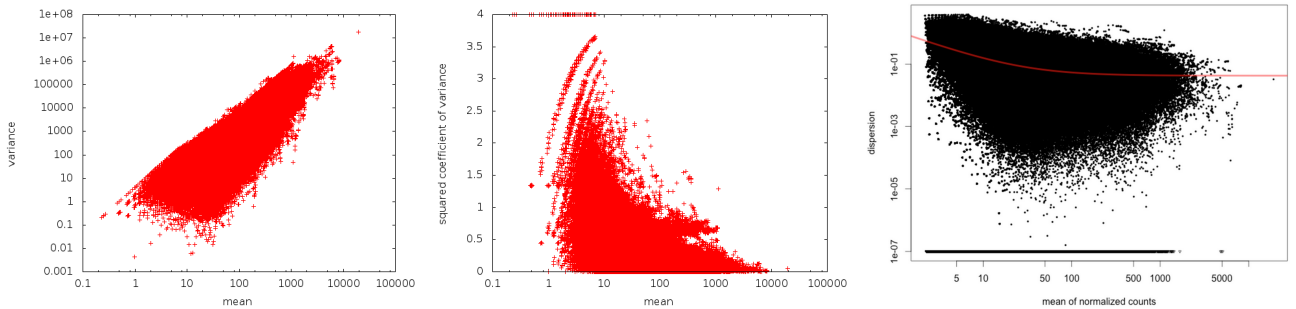
Different splicing events result in the different in k-mer counts of the two groups. Essentially, we want to identify some k-mers that are statistically significant different from other k-mers, which is a superset of the different k-mer counts caused by splicing events.

To achieve this, we investigate sevearl methods, namely, t-test, DESeq test and likelihood ratio test in our project. We present the findings in this section.

4.1 t-test

In order to identify statistically significant differential k-mers in the hundres of millions appeared k-mers, one intuitive method is to apply t-test to compare the two means of k-mer counts between treated and untreated groups, by assuming the counts within one group are normally distributed and different k-mers are independent. Though the two assumptions are strong, t-test is simple to conduct and could be serve as the baseline of our analytical task.

For each row, we use $\mathbb{H}_0$: $\mu_t = \mu_u$ as our hypothesis, in other words we want to test the rows that the treated and untreated group k-mer counts mean are not the same. For the t-test, we vary $\alpha \in \{0.1, 0.05, 0.01\}$ to see how does it performs. In Fig 4, we first show the statistically significant different k-mers plotted together



(a) Dependence of the variance on the mean for the k-mer count table.

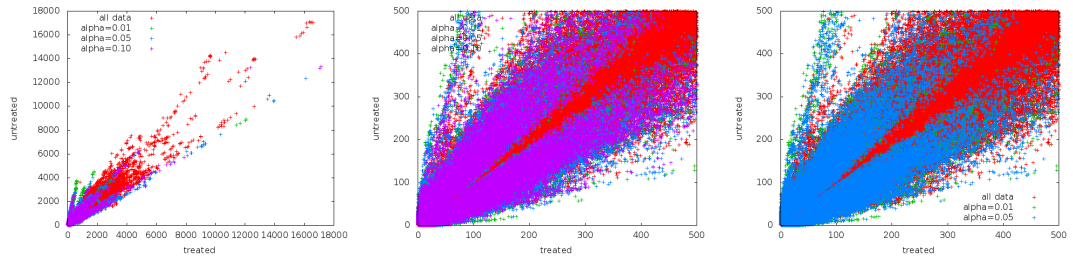
(b) Dependence of the squared variance coefficient on the mean for the k-mer count table.

(c) Empirical (black dots) and fitted (red lines) dispersion values plotted against the mean of the normalized read counts

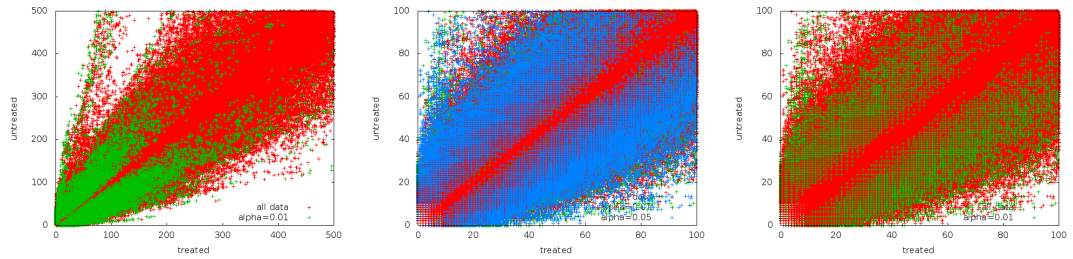
Fig. 2. Data Exploration Result

Table 3. K-mer Count Percentage between Treated and Untreated Groups

Treated		Untreated	
K-mer	$R^{(+)}$ ($\times 10^{-4}$)	K-mer	$R^{(-)}$ ($\times 10^{-4}$)
ACCATTCATTCCAGCCTTCAATTAA	0.569	TTCGTACTAAAATATCACAATTTTT	0.424
AACCATTCATTCCAGCCTTCAATTA	0.567	CGTACTAAAATATCACAATTTTTTA	0.423
CAACCATTCATTCCAGCCTTCAATT	0.566	GTACTAAAATATCACAATTTTTTAA	0.421
CCATTCATTCCAGCCTTCAATTAAA	0.566	TACTAAAATATCACAATTTTTTAAA	0.421
ATTCATTCCAGCCTTCAATTAAAAG	0.561	TCGTACTAAAATATCACAATTTTTT	0.421
CATTCATTCCAGCCTTCAATTAAAA	0.557	TTTCGTACTAAAATATCACAATTTT	0.412
TTCATTCCAGCCTTCAATTAAAAGA	0.557	CTAAAATATCACAATTTTTTAAAGA	0.401
TCGTACTAAAATATCACAATTTTTT	0.553	TAAAATATCACAATTTTTTAAAGAT	0.397
CGTACTAAAATATCACAATTTTTTA	0.552	CTTTCGTACTAAAATATCACAATTT	0.394
GTACTAAAATATCACAATTTTTTAA	0.551	ACTAAAATATCACAATTTTTTAAAG	0.388



(a) On all k-mers, the whole dataset, $\alpha = \{0.1, 0.05, 0.01\}$ (b) On k-mer count less than 500, $\alpha = \{0.1, 0.05, 0.01\}$ (c) On k-mer count less than 500, $\alpha = \{0.05, 0.01\}$



(d) On k-mer count less than 500, $\alpha = 0.01$ (e) On k-mer count less than 100, $\alpha = \{0.05, 0.01\}$ (f) On k-mer count less than 100, $\alpha = 0.01$

Fig. 4. T-test result, Varying $\alpha = \{0.1, 0.05, 0.01\}$, Zoom in to $\{500, 100\}$

with original data in Fig 4(a). The red color is original k-mer count pairs, purple color is the different k-mers when $\alpha = 0.1$, blue color shows the ones when $\alpha = 0.05$, and finally green color plotted the k-mers found when $\alpha = 0.01$.

When we increase α, differential k-mers number increases. We show this different in the Fig 4(b), Fig 4(c) and Fig 4(d). We also show the case when we restrict the k-mer count to be less than 100, in order to see what kind of k-mer rows did t-test found in relatively rare k-mers in Fig 4(e) and Fig 4(f).

As we can see, t-test is a good baseline, since it can found almost all the outliers (the arms not along the $t = u$ diagonal). On the other hand, the precision cannot be high, as we can see in Fig 4(f), many inner points are reported as differential ones. Main reason is the two mean comparision t-test uses the same threshold for different k-mer counts. As we can see in dispertion figure in the data exploration, this assumption is not true. Other drawbacks of this approach, is the data itself may not be normally distributed and not independent.

4.2 DESeq test

DESeq developed a general framework to identify the significant differential reads counts in different samples. Inspired by DESeq, we assume that the read counts of k-mer i in sample j could be modeled by a negative binomial (NB) distribution with two parameters p and r.

$$K_{ij} \sim Pr(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k$$

which is equivalent to be parametrized in terms of its mean μ and variance σ^2:

$$p = \frac{p}{\sigma^2}; r = \frac{\mu^2}{\sigma^2 - \mu}$$

Furthermore, DESeq introduced a variance mean model to reduce the over-dispersion problem:

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_{i,\rho(j)}$$

where s_j is the size factor and $v_{i,\rho(j)}$ is the per k-mer raw variance parameter. DESeq proposes to use local regression on a gamma family linear model to estimate the parameters.

Then DESeq propose a test based on fitted empirical distribution to find the statistically significant differential k-mers.

Due the high time complexity of the fitting algorithm in DESeq, the entire dataset takes more than 5 hours to be processed. As a compromise, we extract the unique k-mers, i.e. representatives of k-mers with the same counts in all the samples, and then plot the log2 fold changes against the mean normalized counts in Fig. 5 to evaluate the negative binomial model. Each scatter point corresponds to a k-mer while the red ones indicate the differentially expressed k-mers at 10% false discovery rate.

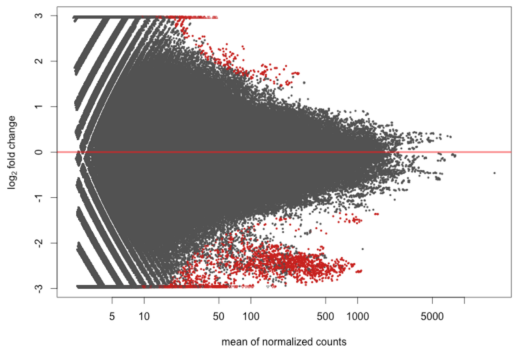


Fig. 5. Testing for differential expression between treated and untreated groups: Scatter plot of log2 ratio (fold change) versus mean.

We also plot our t-test result (with $\alpha = 0.01$) on the same chart as shown in Fig. 6. Comparing with DESeq test result shown in Fig. 5, we found that t-test precision is not satisfactory.

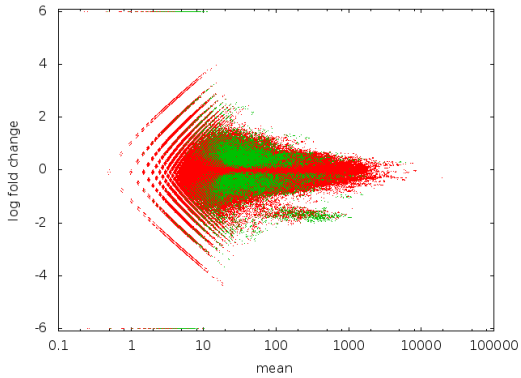


Fig. 6. t-test ($\alpha = 0.01$ result plotted on the log2 ratio (fold change) versus mean figure

4.3 Likelihood ratio test

Though DESeq method seems promising, it is computationally expensive in the model fitting stage, also the empirical distribution based test is non-linear. The model fitting in DESeq paper uses a local regression on gamma family linear models. It cannot finish on the entire k-mer count table in our dataset. Our DESeq result reported in the previous section is produced after unique the k-mer pairs, i.e. when two rows have the same normalized numbers.

In our project, we also investigate the likelihood ratio test approach. The approach requires us to build a joint distribution model to describe the data, and use the test to report differential k-mers. As an overview, the likelihood ratio test compares two competing models, a null model with joint probability $p_{null}(x, y)$ and an alternative model $p_{alter}(x, y)$. The null model is relatively a simpler model and has less degree of freedom f_n while the alternative is a richer model with more freedom f_a. The two models are separately fitted with data to estimate their parameters.

$$D = -2ln(\frac{p_{null}(x, y)}{p_{alter}(x, y)}) \quad (2)$$

Then the log likelihood ratio (Eq 2) is applied for a given k-mer row. One can use $\chi^2_{f_a - f_n}$ test to produce statistically significant differential k-mers.

Using this approach, ideally one can propose rich model that can be fitted fast to work on the hundreds of millions of k-mers rows, in order to make this approach highly scalable.

4.3.1 The Null Model The null model can be built using a Pareto distribution to describe our data. Pareto type I distribution has analytical MLE solutions, whose pdf is

$$f(x) = \frac{\alpha x_{min}^{\alpha}}{x^{\alpha+1}}, x \geq x_{min}$$

where x_{min} is the minimum x, and the only parameter is α. The MLE for α has analytical solutions, and can be solved in $O(n)$ time.

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log(\frac{x_i}{x_{min}})}$$

Also notice that many k-mer counts having the same values, one can group by the same count k-mers in the denominator. Thus the total number of operations can be reduced significantly. In our dataset, though there are around 0.1 billion distinct k-mers, the unique x_i is only 70k.

On the other hand, type II pareto distribution (Lomax distribution) with the following pdf definition does not have an analytical solution when estimating α and θ in MLE, thus it is not ideal to be used for the likelihood ratio test when we want to derive a scalable test.

$$f(x) = \frac{\theta\alpha}{(1 + \theta x)^{\alpha+1}}$$

4.3.2 The Alternative Model The alternative model should be available to describe the data more precisely using more parameters. To build this alternative model, we tried two approaches in our project.

Joint bivariate Pareto Model: First, we used a bivariate pareto model published in recent statistics journal [10]. The model is based

on the following survival function for $x > 0, y > 0$:

$$S_{X,Y}(x,y) = P(X > x, Y > y) = (1 + \alpha_1 x + \alpha_2 y + \alpha_0 xy)^{-\theta}$$

and the related pdf function is defined as:

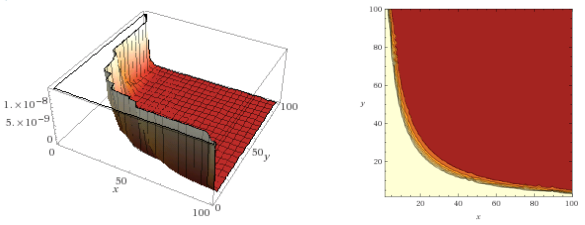
$$f_{X,Y}(x,y) = \frac{\theta(\theta(\alpha_1 + \alpha_0 y)(\alpha_2 + \alpha_0 x) + \alpha_1\alpha_2 - \alpha_0)}{(1 + \alpha_1 x + \alpha_2 y + \alpha_0 xy)^{\theta+2}}$$

The MLE is discussed thoroughly in their paper [10], which requires to solve two nonlinear optimization problems, one is two dimension problem for (α_1, α_2), the other is a single dimension problem for (α_0). We used the grouping technique and wrote our own random seed hill climbing optimization algorithm in C++ and the parameter fitting took about 10 minutes on the whole dataset. We show the fitted parameters in Table 4.

Table 4. Bivariate Model Parameter Fitting Result

α_0	α_1	α_2	θ
0.01441	0.0425	0.0441	6.67

Plug in the parameter, we plot the pdf function of the bivariate model in Fig 7. Unfortunately, as we see from the figure, we found that the bivariate model [10] is lack of shape parameter to model the positive correlation appeared in our data (Fig 3(b)).



(a) Pdf 3D Plot (b) Pdf Contour Plot

Fig. 7. Bivariate Model Fitting Result

Thus, in the last phase of our project, we moved on to build our own model to fit better the data set we have.

Conditional Model In general, a joint bivariate model $f(x, y)$ can be modeled separately on $f(x)$ and the conditional probability $f(y|x)$. The $f(x)$ in our dataset has powerlaw distribution as shown in Fig 3(a) (DESeq proposes to use Negative Binomial Distribution to model the count probability $f(x)$). On the other hand, modeling the conditional probability $f(y|x)$ is more challenging, as we can see from Fig 3(b) the mean and variance of $\{y\}$ of each x_i is tricky to model cleanly.

We propose the following prerequisites for this model:

1. In order to use the likelihood ratio test, the richer model must not have too many parameters, otherwise the χ^2 test with too many degree of freedom helps little to justify the differential k-mers.
2. On the other hand, for a scalable model, the parameter estimation should be fast to deal with the huge amount of k-mers.

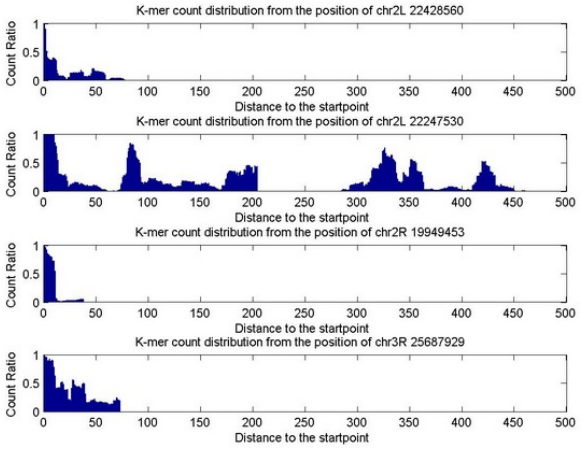


Fig. 8. The count correlation of consequent K-mers at specific genome position.

One thought we had is to fix mean μ_{y_i} as a function of x_i, and variance $\sigma^2_{y_i}$ as a function of μ_{y_i}, motivated by DESeq paper. However, the $\sigma^2 = \mu + \alpha\mu^2$ is the assumption in DESeq, it seems not fit well with our data, where our dataset is under dispersion (Fig 2(c)) and the relationship of mean and variance shown in Fig 2(a) is hardly to say it is quadratic. At the time of writing, we didn't come up with a concrete model that is simple enough and fast enough to be fitted against our data. We leave it as our future work.

K-mer correlation Beyond independent models, the k-mer counts may not be independent. In order to build better model, we verify this independence assumption in this project as well. We explored the consequent k-mers counts to see how their counts are related.

We first randomly choose some k-mers that have at least 2000 counts and calculate the count ratio of consequent 500 k-mers counts. As shown in Figure 8, the count of consequent $k-1$ k-mers is strongly related with the beginning k-mer. On the other hand, the correlation relationship is complicated. As the k-mer may near the intron region, the consequent k-mers may not appear.

How to model correlation among k-mers is an open question, and we leave this as our future work.

5 K-MER MAPPING

Beside identifying significant differential k-mers by statistical model, we also like to know how many k-mers that can be perfectly matched to genome and transcriptome. There are a total of 112,919,483 unique K-mers from reads data and 16,461,925 k-mers from genome and 14,084,301 k-mers from transcriptome. To map the k-mers of sequencing data to genome or transcriptome, Bowtie[7] can build index for genome or transcriptome and then locate the k-mers in the reads data. However, Bowtie is not efficient for perfect k-mer matching with fixed length. Instead, we build k-mer index on the reads data and then match k-mers of genome and transcriptome on it. The size of k-mers on genome and transcriptome much less than the k-mers from reads data, thus the matching is more efficient.

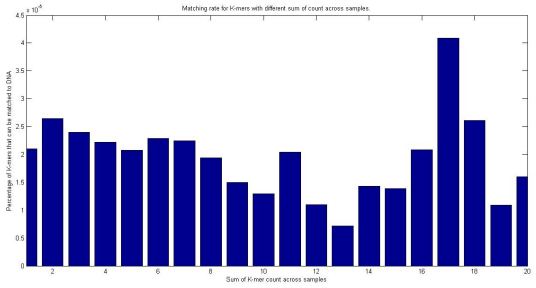


Fig. 9. The K-mer mapping rate for k-mers with small row count sums.

5.1 T-test Mapping Result

Given the K-mers selected by the T-test, we mapped them to the genome and transcriptome and the result is shown in Table 5. It can be seen that a larger α results in more "significant differential" k-mers and a larger matching rate.

Table 5. K-mers identified by T-test with different parameters and matching result on genome and transcriptome.

	$\alpha = 0.01$	$\alpha = 0.05$
# k-mers	2,889,408	3,971,516
% matched to genome	25.57%	28.55%
% matched to transcriptome	12.98%	14.63%

5.2 All K-mers Mapping Result

Below are some statistics about the k-mers mapping:

1. 14.58% of all unique K-mers from reads data can be mapped to genome.
2. 12.69% of all unique K-mers from reads data can be mapped to transcriptome.
3. 297,522 (0.26%) K-mers have zero count variance across samples, and 80.57% of them are all ones.
4. 873,078 (0.77%) K-mers have different count patterns across samples.

We can see that approximately 86% of unique k-mers from reads data can not be mapped to anywhere. To explain this low rate matching, we analyzed the unmatched k-mers in the reads data. We find that 70% of all unique K-mers appeared only once across all samples and only 0.002% of these K-mers can be matched to the genome. We calculate the proportion of k-mers that can be matched for k-mers with the total counts less than 10. The proposition is shown in Figure 9. We can see that the probability of matching is quite low for k-mers with small counts.

Those unmatched k-mers could be generated from alternative splicing events and but a k-mer with too small is not reliable and could probably be generated by sequencing erors. Fr example, if the sequencing error rate is 3%, the probability of generating an error k-mer is $1 - 0.97^{25} = 0.533$, which may explain why a large proportion of k-mers cannot be matched.

6 CONCLUSIONS AND FUTURE WORK

In this project, we conduct the first study on k-mer based statistical method for alternative splicing event analysis. We explore several statistical methods based on k-mer counts to identify significantly differential K-mers including T-test, DESeq, and a joint distribution model. We also develop a fast method for k-mer mapping, which is 2 orders of magnitude faster than Bowtie.

For future work, we plan to map k-mers with 1 or 2 error tolerance onto genome or transcriptome data in order to verify whether the unmapped k-mers are resulted from splicing events or sequencing errors. Besides, taking the strong correlation of consecutive k-mers into account, we could design a more accurate model to identify splicing events. Additionally, more experiments need to be conducted to verify the robustness of the statistical models.

ACKNOWLEDGEMENT

This project is advised by Prof. Hctor Corrada Bravo and Prof. Steve Mount. Our team is sincerely grateful for their instructions and help.

REFERENCES

[1]Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[2]Angela N Brooks, Li Yang, Michael O Duff, Kasper D Hansen, Jung W Park, Sandrine Dudoit, Steven E Brenner, and Brenton R Graveley. Conservation of an rna regulatory map between drosophila and mammals. *Genome Research*, 21(2):193–202, 2011.

[3]Malachi Griffith, Obi L Griffith, Jill Mwenifumbo, Rodrigo Goya, A Sorana Morrissy, Ryan D Morin, Richard Corbett, Michelle J Tang, Ying-Chen Hou, Trevor J Pugh, et al. Alternative expression analysis by rna sequencing. *Nature methods*, 7(10):843–847, 2010.

[4]Roderic Guigó Serra, Michael Sammeth, and Sylvain Foissac. A general definition and nomenclature for alternative splicing events. *PLoS Computational Biology 2008; 4 (8): e1000147*, 2008.

[5]Yin Hu, Yan Huang, Ying Du, Christian F Orellana, Darshan Singh, Amy R Johnson, Anaïs Monroy, Pei-Fen Kuan, Scott M Hammond, Liza Makowski, et al. Diffsplice: the genome-wide detection of differential splicing events with rna-seq. *Nucleic acids research*, 41(2):e39–e39, 2013.

[6]Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and eesign of rna sequencing experiments for identifying isoform regulation. *Nature methods*, 7(12):1009–1015, 2010.

[7]Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[8]Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish: Alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *arXiv preprint arXiv:1308.3700*, 2013.

[9]Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature Biotechnology*, 2014.

[10]P.G. Sankaran and Debasis Kundu. A bivariate pareto model. *Statistics*, 48(2):241–255, 2014.

[11]Shihao Shen, Juw Won Park, Jian Huang, Kimberly A Dittmar, Zhi-xiang Lu, Qing Zhou, Russ P Carstens, and Yi Xing. Mats: A bayesian framework for flexible detection of differential alternative splicing from rna-seq data. *Nucleic acids research*, 40(8):e61–e61, 2012.

[12]Wikipedia. Alternative splicing — wikipedia, the free encyclopedia, 2014. [Online; accessed 25-April-2014].

[13]Wikipedia. K-mer — wikipedia, the free encyclopedia, 2014. [Online; accessed 25-April-2014].