**Jamie Alnasir[1] / Hugh P. Shanahan[1]**

# A Novel Method to Detect Bias in Short Read NGS Data

[1] Department of Computer Science, Royal Holloway, University of London, London TW20 0EX, England, UK, E-mail: Hugh.Shanahan@rhul.ac.uk

**Abstract:**
Detecting sources of bias in transcriptomic data is essential to determine signals of Biological significance. We outline a novel method to detect sequence specific bias in short read Next Generation Sequencing data. This is based on determining intra-exon correlations between specific motifs. This requires a mild assumption that short reads sampled from specific regions from the same exon will be correlated with each other. This has been implemented on Apache Spark and used to analyse two *D. melanogaster* eye-antennal disc data sets generated at the same laboratory. The wild type data set in drosophila indicates a variation due to motif GC content that is more significant than that found due to exon GC content. The software is available online and could be applied for cross-experiment transcriptome data analysis in eukaryotes.

**Keywords:** next-Generation sequencing, short reads, bias, transcriptomics, RNA-Seq

**DOI:** 10.1515/jib-2017-0025

**Received:** April 4, 2017; **Revised:** June 22, 2017; **Accepted:** August 10, 2017

## 1 Introduction

Next-Generation sequencing technologies have had a transformative effect in Genomics [1] and Transcriptomics (RNA-Seq) [2]. On the other hand, RNA-Seq remains susceptible to a variety of systematic biases. Some are comparatively generic, such as library preparation protocols [3]; others are specific to the sequencing platform [4], [5] (recently there has been attention drawn to cross-contamination in multiplexed samples [6] – the effect of this on RNA-Seq studies remains to be seen); others due to differential expression and splicing analysis [7], [8], [9]; finally there are biases due to assembly [10]. Disentangling these issues is difficult and it is clear that there is a need for approaches to detect bias that are not dependent on, for example, differential expression analysis or other types of analysis that are carried out at the end of the extensive pipeline of computational steps taken to derive the data set.

Sequence-specific bias in the RNA-Seq data has already been identified – namely GC-content and dinucleotide frequencies [11], [12] and motif content in hexamer primer regions [13]. In addition to this, transcript length is also a confounding factor and should be taken into account [11].

With this in mind, we propose a novel method to detect possible sources of sequence-specific bias in short read data. This is based on the observation that ideally the number of short reads around a region in an exon will be correlated with short read counts in another region on the same exon. This approach is based on a similar observation made for microarrays [14]. This approach requires assembly of the short reads to a reference genome but does not require any further analysis steps.

All of the software developed for this paper is available on request or can be downloaded directly from https://doi.org/10.5281/zenodo.801378.

## 2 Methods

### 2.1 Intra-Exon Motif Correlations

As discussed previously, a method that can quantify sequence specific biases in deep, transcript analysis is necessary. Here we will describe a novel analysis method, based on analysing sequence motif correlations, that employs the MapReduce formalism of Apache Spark [15] to quantify bias in next-generation sequencing (NGS)

**Hugh P. Shanahan** is the corresponding author.
©2017 Jamie Alnasir et al., published by De Gruyter.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

data at the exon level. This is necessary in order to provide the capacity to process the amounts of data typical in transcriptomic datasets [16].

In an ideal transcriptomic data set mapped reads would be of sufficient length to span entire exons, and would therefore be uniformly distributed across an exon (Figure 1, part A). However, RNA-Seq from NGS generates short reads, the length of which is dependent on the sequencing platform, and as a result mapped reads are typically not distributed uniformly across exons (Figure 1, part B) [17]. Furthermore the number of mapped reads is a function of the number of fragments sequenced and the feature length (i.e. length of the exon), for this reason a number of normalisation methods are used to quantify the number of mapped reads to a feature such as Reads per Kilobase Million (RPKM) [18], and Transcripts per million (TPM) [19]. A review of normalisation methods can be found in [20]. This method investigates the distribution of mapped reads in large datasets.

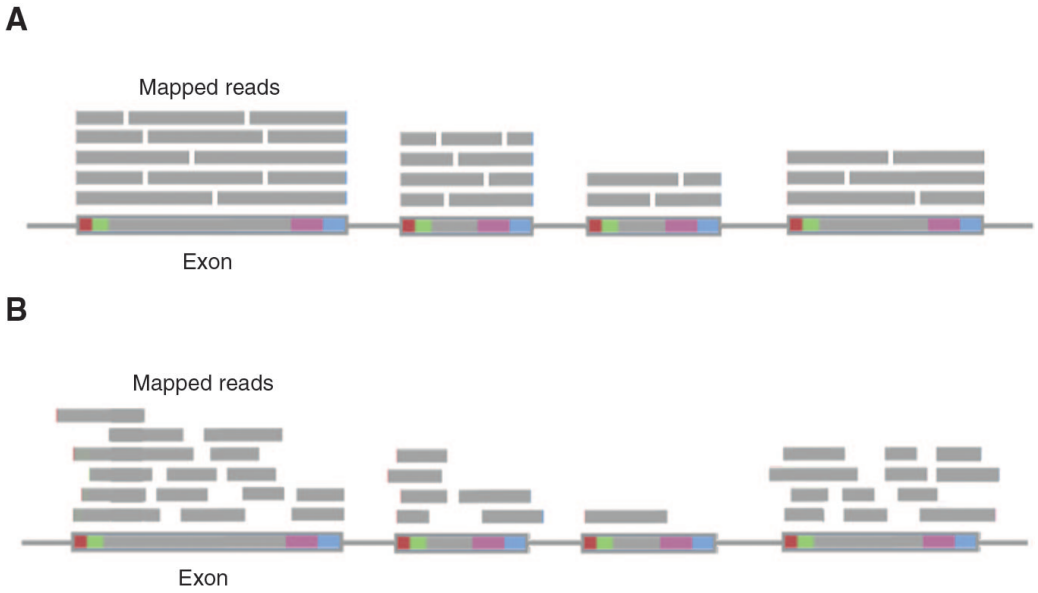



**Figure 1:** (A) Ideal distribution of RNA-Seq reads mapped to an exon if the reads were contiguous and of sufficient length. (B) Typical distribution of RNA-Seq reads mapped to an exon. Grey regions on the exon represent the CDS and other colours represent the 5′cap, 5′UTR, 3′UTR and Poly-Adenylated tail (red, green, purple, blue) correspondingly.

One can estimate the number of reads required for this calculation. Typically one would require as a minimum 10 reads per motif site on an individual exon to establish a reasonable signal. Assuming coverage of reads over an entire exon then one requires on average $10l/r$ reads per exon, where $l$ is the average length of an exon (approximately 200 bp, noted in Section 3) and $r$ the average length of a read (approximately 50 bp). As there are nearly $10^5$ exons in *D. melanogaster* [21] then a minimum of 4 million reads are required. The data sets discussed in this text have 12.9 million reads (wild-type) and 15.0 million reads (mutant) and hence satisfy this.

Quantifying sequence-specific deviations in the distribution of mapped reads across an exon is achieved by picking a short sequence motif – specifically motifs of length 4, which we refer to as *4-mers*, e.g. GGGG,GGGA, and so on. The choice of *4-mers* is motivated at the end of this section. These can occur at various positions within the sequence of the exon. Pairs of these motif occurrences are picked based on their distance apart from each other within the exon and the number of overlapping reads covering each motif position in the pair was counted (Figure 2). The distance between the motif pairs is termed the *motif-spacing*. Motif-spacings of 10, 50, 100 and 200 bp were chosen. In an ideal transcriptomic data set the counts for each motif in the pair would be identical as reads mapped to the exon under inspection would be uniformly distributed and hence over all such possible pairs in that transcriptome there would be perfect correlation. Noise and bias due to length of the transcript will significantly reduce the correlation but if there is no sequence specific effect these averaged correlations should not vary across different motifs.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd.

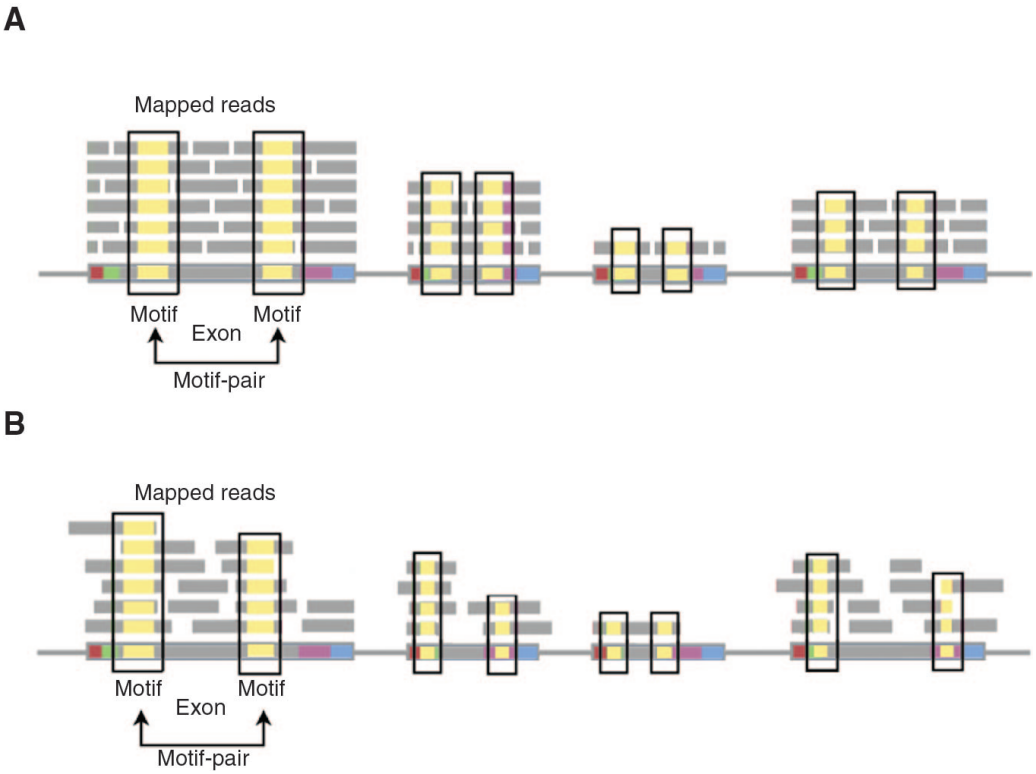



**Figure 2:** Quantification of read coverage using short pairs of sequence motifs (specifically *4-mers*) within the reads shown in yellow. The colours and designation are the same as in Figure 1. (A) Ideal distribution of RNA-Seq reads mapped to an exon if the reads were contiguous and of sufficient length – motif-pairs show perfect correlation. (B) Typical distribution of RNA-Seq reads mapped to an exon – motif-pairs show variable correlation.

In the implementation of this approach discussed in this paper reads which are mapped across exons are included in the analysis. It is key to note, however, that in this analysis the correlations being computed are between motifs that lie on the same exon and hence alternative splicing effects are not considered here.

In order to thoroughly examine the effect of sequence-specific motifs on uniformity of read distribution the Spearman's rank correlation coefficient for all *4-mer* motifs ranging from AAAA to GGGG (i.e. $4^4$ combinations) is computed. This choice of length of motif represents a trade-off. Specifically shorter lengths of motifs would increase the computational size of the problem significantly as the number of instances of each motif grow across the exon. On the other hand, larger lengths of motifs would significantly reduce the statistics and hence increase the noise accordingly.

## 2.2 Using Spark as a Computational Framework

The implementation of this analysis method comprises of two main phases (depicted in Figure 3). In the first phase, motif count and position information is distilled for all exons in the sam reads input tuple $S$. All reads in $S$ must first be partitioned by exon, and the occurrences of the motif in the read sequence $s_k$ and their positions counted. This is carried out using the MapReduce formalism on the Apache Spark platform. Details of the implementation of MapReduce, Spark and the Hadoop computational ecosystem can be found elsewhere [22], [23], [24].

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

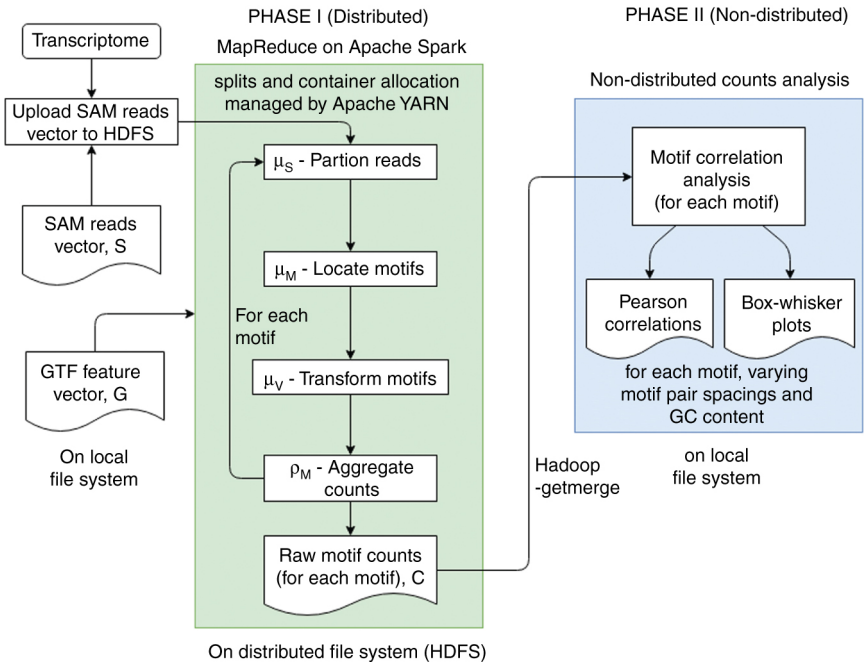



**Figure 3:** Overview of method for quantifying sequence-specific deviations in read distribution. Phase I, the distributed phase, comprises 3 map steps and a reduce step on Apache Spark, with intermediate data being stored on HDFS. Phase II, the non-distributed phase, counts analysis phase utilises raw motif count and position data generated by phase I, which has been stored on the local file system.

The key point to note is that computations revolve around simple data structures called tuples. These are ordered lists; the simplest of which is a key-value pair. Defining $k$ as the key and $v$ as the value then the tuple in this case is written in the form $\langle k,v\rangle$. We note that typically $k$ is a string which acts as an identifier. For example, one key we used was a string constructed from the start and end position, recorded in absolute genomic positions with respect to the chromosomes, e.g. the key "00024790580002481026" is constructed from positions "2479058" and "2481026". Sets of $N$ tuples of this type are described as $\langle k, v\rangle_{i=1}^{N}$. The value $v$ can be a more complex data type. In addition tuples can be composed of keys with a specified list of values.

The Map step, in general, implements a function which converts a tuple into a set of other tuples, i.e.

$$\langle k, v\rangle \rightarrow \langle I_i, w_i\rangle_{i=1}^{N}, \tag{1}$$

i.e. a tuple which is a simple key-value pair is mapped to a set of key-value pair tuples.

The Reduce step converts a tuple composed of values with the same key into a single key-value tuple, i.e.

$$\langle k, v_1, \ldots, v_N\rangle \rightarrow \langle I, W\rangle. \tag{2}$$

With these decompositions, the analysis of tuples can be massively parallelised. In this case, this analysis requires 3 map steps and a reduce step.

The MapReduce steps are daisy-chained such that the output of one MapReduce step is the input of the next step until the final reduce step. The order and function of these steps is outlined in Table 1.

**Table 1:** MapReduce steps employed in the distributed phase I of our analyses method.

| Name of step | Designation | Purpose |
|---|---|---|
| 1. GTF-SAM map | $\mu_S$ | Partitions reads in $S$ by exons in $G$ |
| 2. MOTIF map | $\mu_M$ | Returns a set of positions in which the motif occurs in the read |
| 3. VECTOR map | $\mu_V$ | Maps each occurrence of the motif with a value of 1 |
| 4. MOTIF reduce | $\varrho_M$ | Aggregates counts for the motif at given positions in the exon |

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

Phase I utilises as input widely-used transcriptomic data file formats: SAM aligned reads and GTF genome annotation (which we represent at tuple sets $S$ and $G$ respectively) and yields distilled motif counts data $C_{n,m}$ comprising of exon, position of motif and count for each exon and each motif. Data in $C_{n,m}$ is then processed by a non-distributed phase II to compute correlations which can quantify sequence-specific deviations in the distribution of mapped reads across exons.

The first map step is $\mu_S$ map which partitions reads by exon. This utilises the annotation data (stored in the tuple $G$) and returns a key for the exon the read is mapped to (mapping occurs prior to our analysis by the read-alignment software). After application, the output of the $\mu_S$ map step are the partitioned reads $E$. $E$ is a list of key-value pair tuples comprising of the exon-key (which we discuss later in this section) and the raw read data as the value, and is defined below:

$$E = \langle e, S(e) \rangle_{i=1}^{N} \tag{3}$$

The second map step $\mu_M$ takes a read and returns an exon key $e$ and vector $P$ of the positions in which the motif occurs for that exon. The third map step $\mu_V$ and final reduce step $\wr_M$ transform and aggregate the data by exon $e_{n,m}$, position $q_{n,m}$ and count $z_{n,m}$. This processing yields tuple $C_{n,m}$, which is the final output from the distributed phase, where $q$ is the motif position on the exon $e$ and $z$ is the count of motifs overlapping position $q$. The indices $n$, $m$ are the indices of the count tuple and *4-mer* motif on a given exon $e$ respectively. $C_{n,m}$ is defined below:

$$C_{n,m} = \langle e_{n,m}, q_{n,m}, z_{n,m} \rangle \tag{4}$$

Counts in $C_{n,m}$ that are in the lowest quartile (typically 2) are discarded to reduce noise. Each set of counts is of the overlapping motif $m$ at positions separated at a spacing of $d$ bp on the exon with a tolerance $t$ of $\pm 2$ bp for 10 and 50 spaced motifs, and $\pm 4$ bp for 100 and 200 bp spaced motifs. The resulting motif-pair counts tuple $W_{n,m}$ takes the form below:

$$W_{n,m} = \langle e, q_{n,m}, v_{n,m}, q_{o,m}, w_{n,m} \rangle \quad \text{where } q_{o,m} = q_{n,m} + d \pm t. \tag{5}$$

where $v_{n,m}$, $w_{n,m}$ are the counts in the motif-pair for motif $m$ on each exon $e$ at a fixed separation of $d$ bp.

With this tuple, correlations for each motif are computed over all exons. In order to probe the effect of the GC content of the exon, the GC content of each exon is estimated by computing the average GC content of all the reads that have overlap with each exon (i.e. using $E$). $C_{n,m}$ is partitioned into bins of varying GC-content.

#### 2.2.1 Computing Correlations

The tuple $W_{n,m}$ carries all the pairs of counts $v_{n,m}$ and $w_{n,m}$ of short reads in an exon that have overlap with a given motif $m$ at two positions on the exon that are a specific spacing apart, as outlined in Figure 2. Hence one can compute the corresponding paired ranks of the counts for each motif $m$ and compute the Spearman's rank correlation coefficient.

## 3 Results

In order to test this approach we used RNA-Seq data generated from the eye-antennal imaginal discs of *D. melanogaster*. This data set is composed of wild type and the homozygous glass mutant gl[60j] [25] using Illumina HiSeq 2000's. This data is deposited at GEO with the ID GSE39781. The data was also assembled previously using TopHat V2 [26] using default settings. In this analysis one replicate from each case was used.

The data quoted here is for a spacing of 200 bp. In Figure 4 a histogram of exon lengths for *D. melanogaster* is plotted (data derived from version 6.15) of the Exon Fasta entries of Flybase [21]. Over 50% of the exons have a length that is greater than 200 bp.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

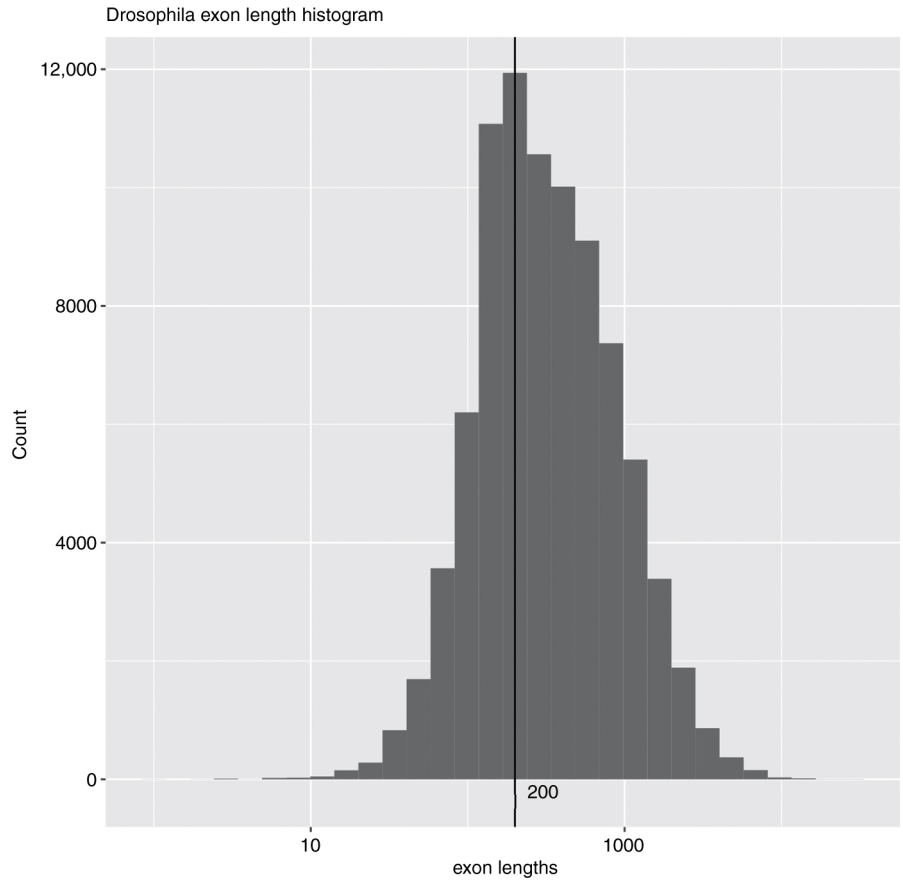



**Figure 4:** Histogram of exon lengths. The horizontal scale is logarithmic.

In the supplementary materials, equivalent data in Section 3.1 and Section 3.2 is shown for the other separations and the picture is largely the same.

## 3.1 Motif Correlations

In Table 2 the motifs with highest and lowest 10 Spearman's rank correlations are listed for the wild type and mutant data.

**Table 2:** Lowest and highest correlations for motifs in wild-type and Glass mutant data with a separation of 200 bp using Spearman's rank correlation.

| | Wild-type | | Glass mutant | |
|---|---|---|---|---|
| | **Motif** | **Correlation** | **Motif** | **Correlation** |
| 11*Lowest | | | | |
| | GGGG | 0.047 | GGGT | 0.009 |
| | CCCC | 0.076 | AGGG | 0.016 |
| | CCCG | 0.077 | ACCC | 0.030 |
| | GCCC | 0.083 | GCCC | 0.040 |
| | ACCC | 0.088 | CCCT | 0.042 |
| | CCCT | 0.090 | GGGC | 0.048 |
| | CGGG | 0.108 | CGGG | 0.054 |
| | GGGC | 0.109 | CCCA | 0.054 |
| | GGGA | 0.121 | TGGG | 0.057 |
| | CGCG | 0.125 | TCCC | 0.063 |
| 11*Highest | | | | |
| | CGTG | 0.616 | CTTG | 0.636 |
| | CAAG | 0.611 | CACT | 0.618 |
| | CTTA | 0.607 | CAGT | 0.609 |
| | CTTG | 0.606 | ACTG | 0.605 |
| | TACG | 0.606 | CTCG | 0.588 |

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

| | | | |
|---|---|---|---|
| CACT | 0.602 | AGTT | 0.584 |
| GTAC | 0.600 | CAAG | 0.582 |
| ACGT | 0.599 | CGAT | 0.581 |
| AGTC | 0.597 | TCGT | 0.573 |
| CTAC | 0.595 | ATTG | 0.566 |

A number of observations can be drawn from this data. In the first instance there are a very wide range of correlations in both data sets. The outliers in the wild-type and mutant data with low correlations tend have to higher GC-content with repeats. These can still indicate statistically significant relationships but the overall trend of small correlations for this class of motifs is noticeable. The wild-type and mutant high outliers on the other hand have similar correlations and a lower GC content with little or no repeats (there are no repeats longer that 2 and those are composed of A or T). This picture is borne out at the other spacings examined which are listed (along with their statistical significance) in the Supplementary Information.

### 3.2 GC Content of Motifs Versus GC Content of Exons

As noted previously, there is a clear effect on estimated expression levels due to GC content and dinucleotide frequencies [11], [12]. As outlined in Section 2.2 the raw count data was sub-divided as a function of the GC-content of the motif and exon and correlations recomputed. In Figure 5 boxplots are drawn of the resulting data. In (A) of Figure 5 it can be seen that there is a notable variation of the correlation as a function of the GC content of the motif in the wild-type data set. On the other hand, no variation can be seen in (B) with the variation of the exon GC content indicating that there is an effect is due to the motif rather than the overall GC content of the exon. In (C) and (D) we see no variation for the mutant-type data set. We note also that the correlations for the mutant data in (C) and (D) are significantly smaller than those in (A) and (B).

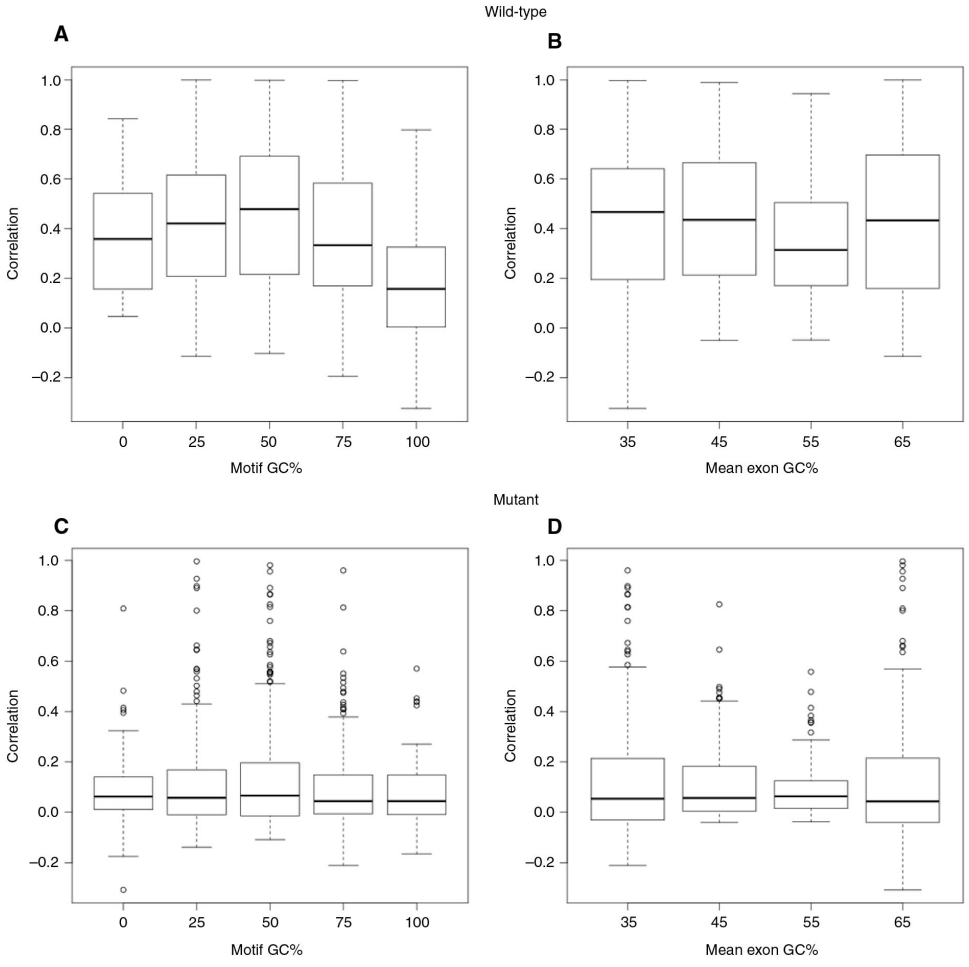



**Figure 5:** Comparison of correlations as a function of Motif and Exon GC content. The data exhibits a wider spread of correlations as each data point is based on data that has been binned as a function of Motif and Exon GC content.

A similar pattern occurs for the other spacings used in this analysis which are listed in the Supplementary Information.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

## 4 Conclusions

In this paper we have proposed a novel method to probe sequence-specific biases in short read RNA-Seq data. The approach is based on the assumption that short reads from one region on an exon will be correlated with short reads from another region of the same exon. The short reads must be assembled with a reference genome but requires no further pre-processing. The assembly is used to identify which short reads overlap with the position of motifs (in this case of length 4) within all exons. In this respect, this could be obviated further by using motif identification approaches outlined in [27].

We have presented an initial analysis of two data sets drawn from *D. melanogaster* [25]. In this case we have shown that both data sets exhibit a bias that is dependent on GC content, namely in the motifs that exhibit the lowest correlations for both the wild-type and mutant data. The effect appears not to be specific to particular sequences – as we can see in Table 2 the motif GGGG and CCCC have the lowest correlations in the wild-type data but do not appear in the ten lowest correlations for the mutant data. Outliers with low correlations may still represent a statistically significant relationship between the counts of overlapping reads with a specific motif on an exon with read counts of the same motif on another site on the same exon a specific distance apart. On the other hand, when we distinguish the effect of GC content of the exon and motif there is evidence that the wild-type data set exhibits bias that is specific to the GC content of the motif rather than the overall GC content of the exon and the there is no noticeable effect for the mutant data set and more specifically, the correlations are significantly smaller for the mutant data set. These effects are largely independent of changing the separation between occurrences of motifs.

It is important to note the consequences of this. These are two RNA-Seq data sets that have been generated in the same lab on the same species. Hence it is reasonable to assume that the protocols to prepare the sample and performing the sequencing are the same. Furthermore the data sets are gathered from the same type of tissue. There will be differences in the transcriptome because of the genetic perturbation; however we expect only a fraction of changes in expression and splicing. In the same respect, multi-mapped reads, as outlined in [28], [29], may represent a source of bias if reads map to many locations but are only recorded in the above count data in one location. However, as we expect only a fraction of the transcriptome to be perturbed the difference between them, the changes in correlations should remain overall relatively small. What is observed are changes in correlations that represent a much more significant change in the distribution of the short reads between these two data sets. Hence there is a source of variance that is unaccounted for between these data sets.

Beyond this present study, the method described in this paper could potentially be used as a tool to more specifically probe the biases in RNA-Seq data (apart from the differential expression and splicing analysis) individually with relevant data sets.

### Acknowledgements

The authors wish to thank the referees for many helpful comments – in particular suggesting the use of the Spearman's rank correlation coefficient. J.A. is supported through direct funding from the department of Computer Science at Royal Holloway, University of London.

**Conflict of interest statement:** Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

## References

[1] Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet 2010;11:31–46.
[2] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63.
[3] Alnasir J, Shanahan HP. Investigation into the annotation of protocol sequencing steps in the sequence read archive. GigaScience 2015;4:23.
[4] Whiteford N, Skelly T, Curtis C, Ritchie ME, Lohr A, Zaranek AW, et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics 2009;25:2194–9.
[5] Ledergerber C, Dessimoz C. Base-calling for next-generation sequencing platforms. Briefings in Bioinformatics. 2011;12:489–497. DOI: 10.1093/bib/bbq077.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

[6] Ledergerber C, Dessimoz C. Base-calling for next-generation sequencing platforms. Brief Bioinform 2011;12:489–97.
[7] Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. bioRxiv. 2017. DOI: 10.1101/125724 .
[8] Hooper JE. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. Hum Genomics 2014;8:3.
[9] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13.
[10] Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinform 2017;18:38.
[11] Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010;95:315–27.
[12] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-Seq data. BMC Bioinform 2011;12:480.
[13] Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-sequencing data. BMC Bioinform 2011;12:290.
[14] Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 2010;38:e131.
[15] Langdon WB, Upton GJ, Harrison AP. Probes containing runs of guanines provide insights into the biophysics and bioinformatics of Affymetrix GeneChips. Brief Bioinform 2009;10:259–77.
[16] Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. HotCloud 2010;10:95.
[17] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big data: astronomical or genomical? PLoS Biol 2015;13:1–11.
[18] Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 2011;12:R22.
[19] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 2008;5:621–8.
[20] Wagner GP, Kin K, Lynch VJ. Measurement of mrna abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 2012;131:281–5.
[21] Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC Bioinform 2015;16:347.
[22] Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM. FlyBase: genomes by the dozen. Nucleic Acids Research. 2007;35:D486–D491. DOI: 10.1093/nar/gkl827.
[23] Fish B, Kun J, Lelkes AD, Reyzin L, Turán G. On the computational complexity of mapreduce. In: International symposium on distributed computing, vol. 9363. Berlin: Springer, 2015;1–15.
[24] Zaharia M, Chowdhury M, Das T, Dave A, Ma J, McCauley M, et al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on networked systems design and implementation. USENIX Association, 2012:2–2.
[25] Naval-Sanchez M, Potier D, Haagen L, Sanchez M, Munck S, Van de Sande B, et al. Comparative motif discovery combined with comparative transcriptomics yields accurate targetome and enhancer predictions. Genome Res 2013;23:74–88.
[26] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009;25:1105–11.
[27] Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol 2014;32:462–4.
[28] Ji Y, Xu Y, Zhang Q, Tsui K-W, Yuan Y, Norris Jr, C, et al. BM-Map: Bayesian mapping of multireads for next-generation sequencing data. Biometrics 2011;67:1215–24.
[29] Feng W, Sang P, Lian D, Dong Y, Song F, Li M, et al. ResSeq: enhancing short-read sequencing alignment by rescuing error-containing reads. IEEE/ACM Trans Comput Biol Bioinform 2015;12:795–8.

**Supplemental Material:** The online version of this article offers supplementary material (https://doi.org/10.1515/jib-2017-0025).

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd