Enabling complete transcriptome sequencing

Mix² : A software tool for the accurate estimation of RNA concentration from RNA-Seq data

Mix² (rd. "mixquare") yields highly accurate concentration estimates for gene isoforms by adapting to the positional coverage bias in RNA-Seq data. This leads to higher accuracy in the detection of differential expression of genes and gene isoforms. Mix² enables repeatable concentration estimates across multiple library preparations and sequencing facilities and can be used as an explorative tool to investigate the positional biases present in RNA-Seq data. Mix² is highly efficient and runs significantly faster than current state-of-the-art RNA-Seq data analysis tools.

Introduction

Fragment bias in RNA-Seq poses a serious challenge to the accurate quantification of gene isoforms. Mix² makes no assumptions about coverage bias but fits for each gene isoform a mixture model to the data (**Fig. 1**). Mix² can therefore, for instance, accurately represent the 5' bias, as shown in **Fig. 1** (**a** and **b**), whereas Cufflinks is restricted to the uniform distribution (**Fig. 1c**).

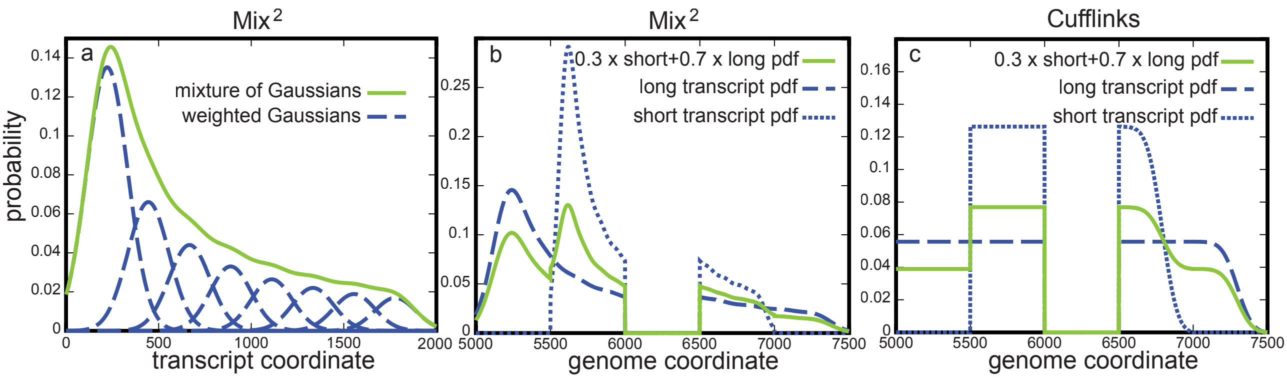

Figure 1 | Exemplary representation for positional fragment bias over a 2000 bps transcript modeled with a mixture of 8 normal distributions. (**a**) the green curve shows the combined probability density function over the whole transcript, while the blue curves show the individual mixture distributions. (**b**) and (**c**) panels display fragment distributions in a locus with two transcripts sharing one junction, as modeled by Mix² or Cufflinks. Long and short transcripts start at 5000 and 5500 bp from the beginning of the locus, and are 2000 and 1000 bp long, respectively. The junction spans the 6000 – 6499 bp region.

Experiments and results

Mix² was tested on the publicly available MicroArray Quality Control (MAQC) [1] and Association of Biomolecular Resource Facilities (ABRF) [2] datasets, containing RNA-Seq data from multiple sequencing facilities and library preparations which started with differently degraded RNA. For the Universal Human Reference (UHR) RNA and the Human Brain Reference (HBR) RNA samples in MAQC and ABRF, qPCR concentration measurements of around 1000 gene isoforms are available, which are regarded as the ground truth. In our experiments we compare the Mix² software to the widely used Cufflinks[3], PennSeq[4], RSEM[6] and eXpress[7] which has been shown to outperform a large number of methods for the estimation of isoform concentrations.

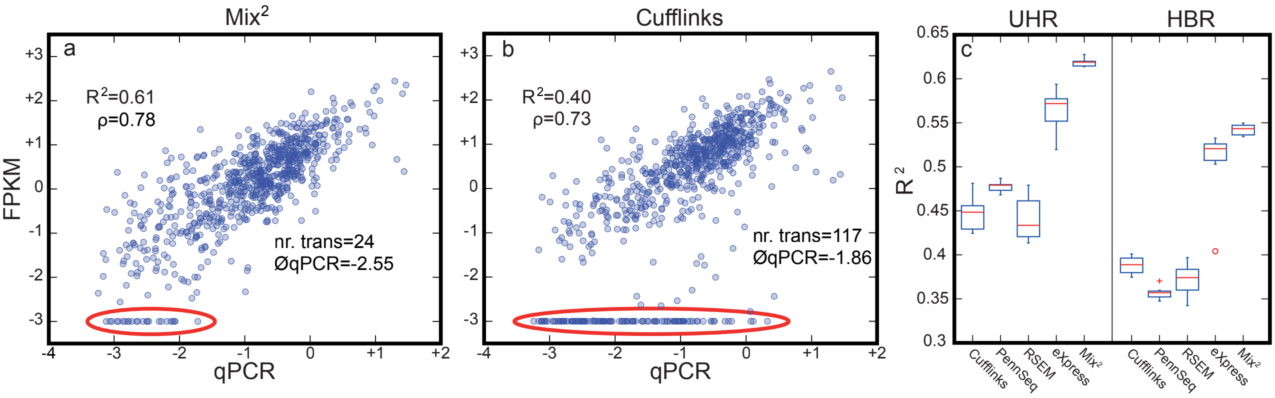

Figure 2 | Correlation between qPCR and FPKM values on MAQC data for Mix², Cufflinks, PennSeq, eXpress and RSEM. (a) and **(b)** Correlation for one lane of UHR RNA data. The ellipses at the bottom highlight transcripts which were not detected. **(c)** Average R² value over all 7 lanes of UHR and HBR RNA.

Accurate concentration estimates of gene isoforms

Our experiments show considerably better correlation between the qPCR and FPKM (expected Fragments Per Kilobase of transcript per Million fragments mapped [3]) values for the Mix^2 model than for Cufflinks with bias correction[3, 5], PennSeq[4], RSEM[6] and eXpress[7]. Cufflinks fails to detect 117 transcripts or 16 % of the complete test set in one lane of UHR RNA data (**Fig. 2b**). Similarly, the average qPCR value of -1.86, for undetected transcripts, is relatively high in comparison to the average for the complete test set, which is -1.02. Hence, Cufflinks fails to detect a large number of highly abundant transcripts. The number of undetected transcripts for Mix^2 is considerably smaller and equals 24, which is 3 % of the complete test set (**Fig. 2a**). In addition, the average qPCR value of -2.55 for Mix^2 is noticeably smaller than that for Cufflinks, hence Mix^2 fails to detect only transcripts of low abundance. Finally, the correlation between qPCR and FPKM values as measured by R^2 is considerably higher for Mix^2 than for for Cufflinks, PennSeq, eXpress and RSEM (**Fig. 2c**).

The higher accuracy of the concentration estimates of Mix^2 leads to better correlation between qPCR and FPKM fold-changes and consequently to higher accuracy in the detection of differential expression (**Fig. 3**). In order to determine the influence of the correlation of FPKM and qPCR fold changes on the detection of differential expression, a simple classification experiment was performed. Transcripts whose qPCR fold change was above 2 or below 0.5 were defined as differentially up- or down-regulated, respectively. The remaining transcripts were characterized as not differentially expressed. All transcripts were also classified according to their FPKM fold change: the transcript was treated as up- or down-regulated if its FPKM fold change was above a certain threshold or below the inverse of the threshold. Varying the threshold between 1.1 and 10^6 and recording the true and false positive rates for each value, with respect to the qPCR-based definitions, we obtained the Receiver Operating Characteristic (ROC) curve shown in **Fig. 3c**. Taking the average over all 49 combinations of lanes in UHR and HBR datasets, with Mix^2 we obtain a true positive rate of 0.71 at a false positive rate of 0.1 (indicated by a dotted line in the **Fig. 3c**), which is considerably higher than the 0.22 for Cufflinks, the 0.44 for PennSeq and the 0.46 for eXpress.

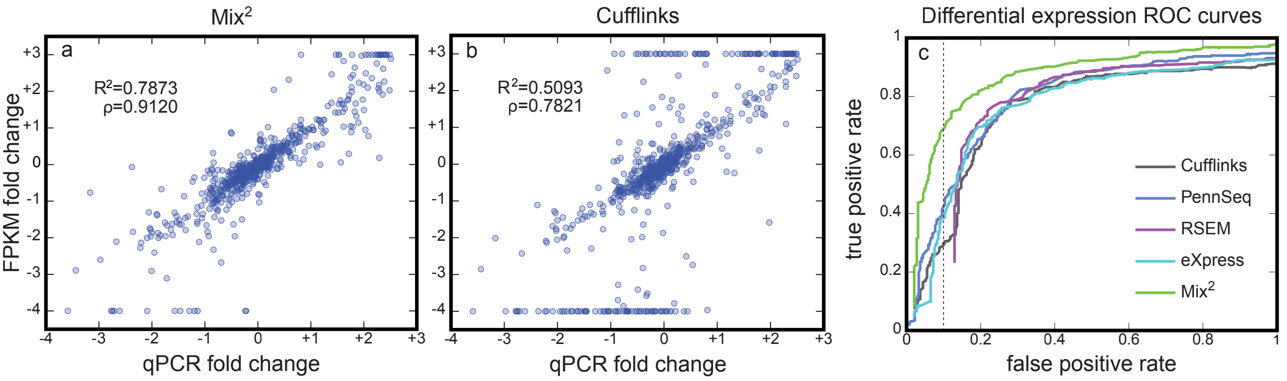

Figure 3 | Correlation between qPCR and FPKM fold changes between UHR and HBR RNA for Mix^2 vs Cufflinks, and the ROC curve for a classification experiment based on FPKM values of UHR and HBR RNA lanes. Since the FPKM and qPCR fold changes should be identical, the range of FPKM fold changes was restricted to the range of qPCR values, as shown in (**a**) and (**b**), and thus to a range between 10^{-4} and 10^3. (**b**) Cufflinks produces a large number of transcripts whose FPKM fold change lies considerably above or below the majority, as can be seen by the long straight clusters at FPKM fold changes of 10^{-4} and 10^3. The Mix^2 model, on the other hand, greatly improves the correlation between qPCR and FPKM fold changes for the UHR and HBR RNA samples, and as shown in the classification experiment (**c**) leads to a substantially higher accuracy in the detection of differential expression. The dotted line in (**c**) indicates a false positive rate of 0.1.

Repeatable concentration estimates for variable conditions

We performed repeatability experiments using data from the ABRF NGS study [4]. Our test set included RNA-Seq data generated with Illumina HiSeq 2000/2500 from the UHR and HBR RNA samples at 6 different sequencing facilities. Libraries were produced using either poly(A) enrichment or rRNA depletion protocols, where the input RNA was either intact or fragmented by one of three different degradation methods. We ran Mix^2 and Cufflinks on this test set and correlated the resulting FPKM values for identical samples and different conditions of RNA degradation, library preparation, and sequencing facility. These experiments revealed a 28.23 % average increase in R^2 value and a 4.08 % average increase of Spearman correlation for Mix^2 over Cufflinks for all 46 tested comparisons.

We also correlated the FPKM fold change between UHR and HBR RNA, whose RNA-Seq data were generated under different conditions, to the qPCR fold change (**Fig. 4**). In our experiments we distinguished between similar and variable conditions, where a condition was labeled similar if the facilities, which sequenced the UHR and HBR RNA samples, were either identical or generated RNA-Seq data of similar quality with similar protocols. The input RNA for both UHR and HBR had to be intact or degraded in an identical manner. Mix^2 produced a substantial increase in both R^2 value and Spearman correlation, both under similar and variable conditions (**Fig. 4**). The correlation for Mix^2, in terms of R^2 value, is higher under variable conditions than the correlation for Cufflinks under similar conditions. Thus, the FPKM values generated by Mix^2 under variable conditions are more comparable than the FPKM values generated by Cufflinks under similar conditions (**Fig. 4**). Overall, in the fold change experiments Mix^2 achieved a substantial increase in Spearman correlation and R^2 value over Cufflinks of on average 12.24 % and 48.36 %, respectively. The ROC curves (**Fig. 4a** and **d**) show the effect of the higher correlation between qPCR and FPKM fold changes for Mix^2 on the accuracy of the detection of differential expression. This clearly highlights the superiority of Mix^2 over Cufflinks at all true and false positive rates.

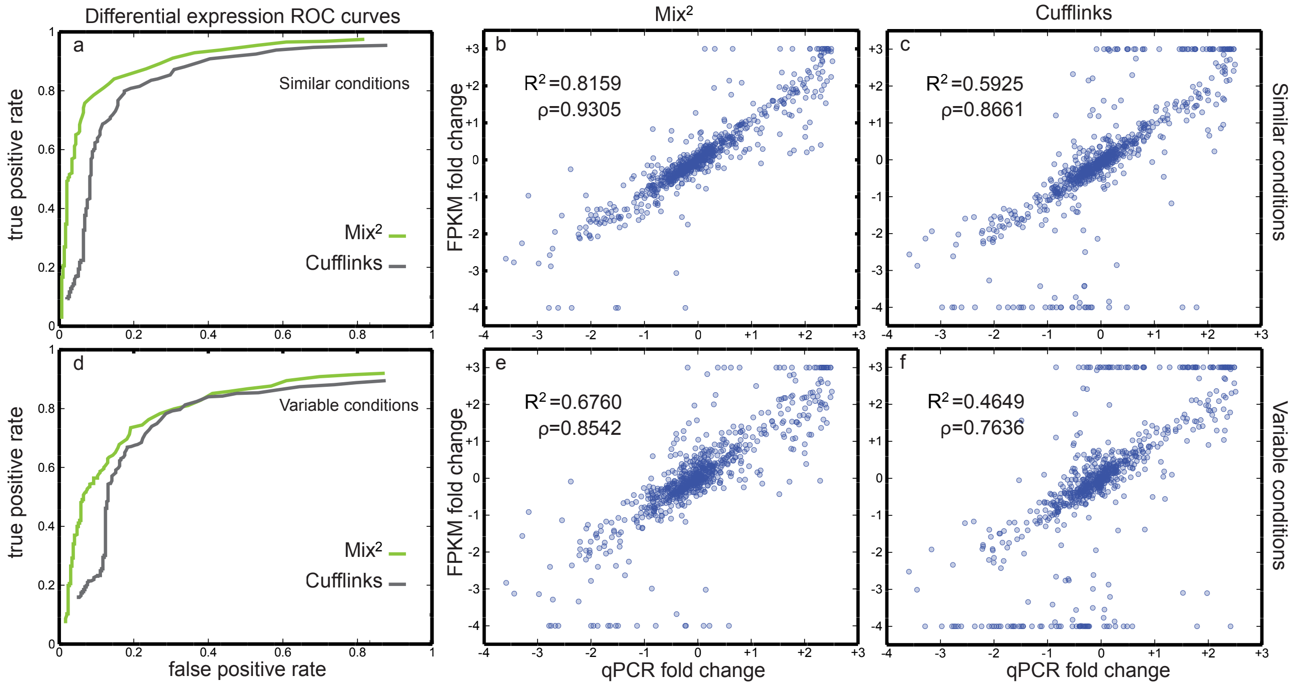

Figure 4 | FPKM vs qPCR fold change correlation between UHR and HBR RNA, where RNA-Seq data for UHR and HBR were generated under similar or variable conditions. (**a**)-(**c**) show differential expression ROC curves and correlations for similar conditions (LB1 and LA1 as described in [2]): the samples were prepared using the same protocol and sequenced at the same facility. (**d**)-(**f**) display differential expression ROC curves and correlations for variable conditions (NBR1 and MAS1, see [2]), where the samples were prepared using different fragmentation methods and sequenced at different facilities. Mix² ((**b**) and (**e**)) yields higher correlations than Cufflinks in both types of experiments. Cufflinks ((**c**) and (**f**)) produces a large number of transcripts whose FPKM fold change lies considerably above or below the majority, as can be noticed by the long straight clusters at FPKM fold changes of 10^{-4} and 10^{3}.

Exploring the positional biases in RNA-Seq with Mix²

The transcript-specific fragment distributions estimated by Mix² can be clustered to reveal biases present in RNA-Seq data. **Fig. 5** shows some of the dominant bias types in the MAQC data detected by Mix² for a subset of transcripts in one lane of UHR RNA. The classes of distributions with 5'-bias (7.93 %) and 3'-bias (17.72 %) depicted in **Fig. 5a** and **c** are just two examples of the classes representing these biases. In total, 20.16 % of transcripts can be assigned to classes with 5'-bias, while 26.34 % can be assigned to classes with 3'-bias. On the other hand, only 26.92 % of transcripts exhibit uniform distributions, as can be seen in **Fig. 5b**. The dominance of non-uniform distributions is not immediately evident, when looking at the accumulated distribution of unclustered transcripts (**Fig. 5d**). In fact, the accumulated distribution gives the false impression that no relevant biases are present. The source of the biases in **Fig. 5** is obscure, however the 3'-bias can be addressed to the type of library preparation, as it is typical for libraries made with cDNA fragmentation [2]. The 5'-bias can arise from RNA degradation.

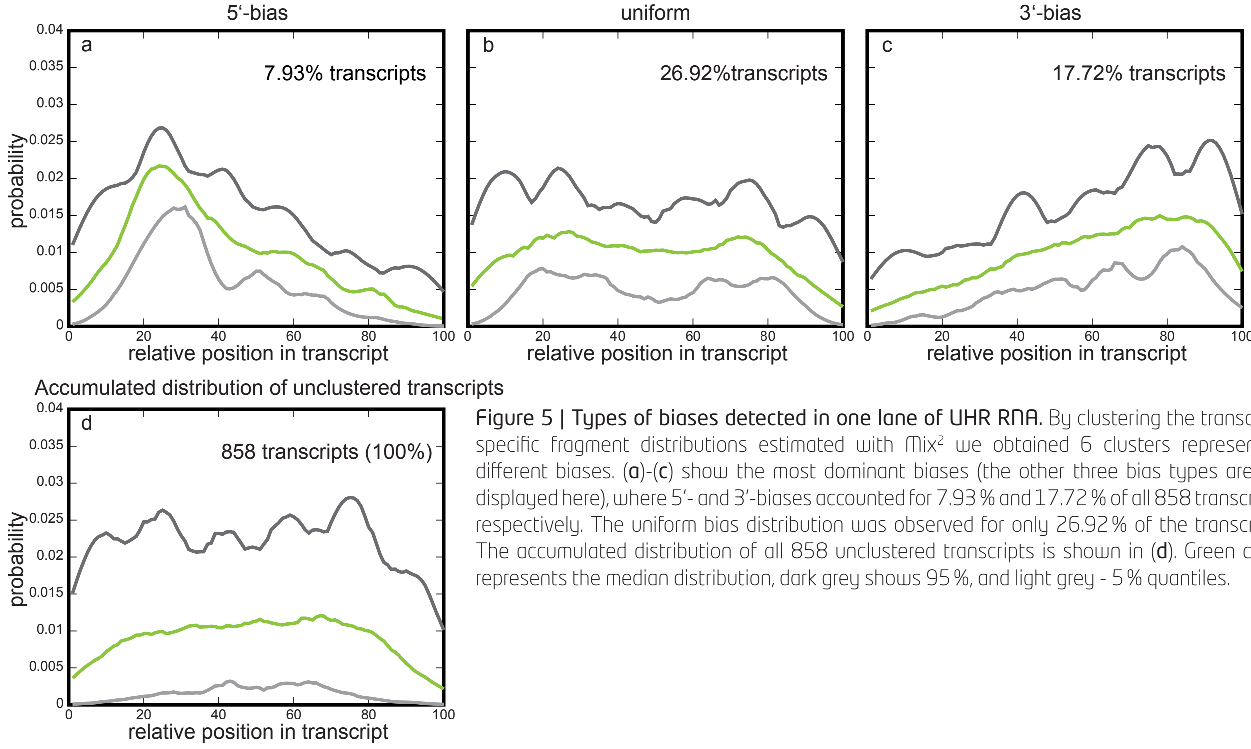

Figure 5 | Types of biases detected in one lane of UHR RNA. By clustering the transcript-specific fragment distributions estimated with Mix² we obtained 6 clusters representing different biases. (**a**)-(**c**) show the most dominant biases (the other three bias types are not displayed here), where 5'- and 3'-biases accounted for 7.93 % and 17.72 % of all 858 transcripts, respectively. The uniform bias distribution was observed for only 26.92 % of the transcripts. The accumulated distribution of all 858 unclustered transcripts is shown in (**d**). Green curve represents the median distribution, dark grey shows 95 %, and light grey - 5 % quantiles.

Implementation and run-time performance

Lexogen has implemented the Mix^2 software as a 64-bit Linux command line tool, which has been tested on a number of Linux distributions. For an up-to-date list of supported distributions please refer to the User Guide of the Mix^2 software. The Mix^2 software can process paired-end and single-end data and has been tested on bam files generated from Illumina and SOLiD RNA-Seq reads.

Table 1 gives a comparison of the usage statistics of Cufflinks running with and without bias correction to Mix^2 on the 7 lanes of UHR and HBR RNA in the MAQC data set. Mix^2 is faster than Cufflinks without bias correction by an average factor of 4.85. In comparison to Cufflinks with bias correction, Mix^2 is even faster by an average factor of 77.42 for UHR. Similar numbers can be observed for HBR, where Mix^2 is faster by a factor of 6.40 and 107.20, respectively, than Cufflinks without and with bias correction. Running Cufflinks with bias correction produces more accurate results and was the mode in which Cufflinks was used for the experiments in the previous sections. These results therefore show that Mix^2 does not only produce substantially more accurate concentration estimates than Cufflinks but is also significantly faster.

Table 1 | **Average memory usage and run-time statistics on the 7 lanes of UHR and HBR RNA in the MAQC data set.** Min stands for run-time in minutes, GB for memory usage in gigabytes. xRT and xMEM are the factors by which run-time and memory usage increases, respectively, in comparison to Mix^2.

	Mix^2		Cufflinks w/o bias correction				Cufflinks with bias correction			
lane	Min	GB	Min	xRT	GB	xMEM	Min	xRT	GB	xMEM
Avg (UHR)	7	1.26	34	4.9	0.99	0.79	542	77.4	1.32	1.05
Avg (HBR)	5	1.02	32	6.4	0.90	0.88	536	107.2	1.22	1.20

Conclusions

The Mix^2 software developed by Lexogen enables:

- Considerably better correlation between known and estimated isoform concentration than current state-of-the-art RNA-Seq data analysis methods.
- Improved transcript concentration fold-change estimates which yield more accurate detection of differential expression.
- Repeatable concentration estimates across different sequencing facilities, library preparations, and types of RNA degradation.
- Detection and classification of bias types in RNA-Seq data.
- Extremely fast run-times and small memory footprint.

References

1. MAQC Consortium (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 24, 1151-1161.
2. Li, S. et al (2014). Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat. Biotechnol. 32, 915-925.
3. Trapnell, C. et al (2010). Transcript assembly and quantification by RNA-Seq reveals unnatotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511-515.
4. Hu, Y. et al (2014). PennSeq: accurate isoform-specific gene expression quantification in RNA-Seq by modeling non-uniform read distribution. Nucleic Acids Res. 42:e20.
5. Roberts, A. et al (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 12, R22.
6. Bo Li and Colin Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12(1):323, 2011.
7. Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Meth, 10(1):71–73, January 2013.

More information and download options,
you can find at:
https://www.lexogen.com/mix-analysis-software/

Lexogen GmbH · Campus Vienna Biocenter 5 · 1030 Vienna · Austria

Find more about Mix^2 at www.lexogen.com.
Contact us at info@lexogen.com or +43 1 345 1212-41.

MIX^2

Accurate Analysis of RNA-Seq Data