







RESEARCH ARTICLE

REVISED Correction of gene model annotations improves isoform abundance estimates: the example of ketohexokinase (*Khk*) [version 2; peer review: 3 approved]

Christophe D. Chabbert , Tanja Eberhart , Ilaria Guccini , Wilhelm Krek⁺, Werner J. Kovacs 

Institute of Molecular Health Sciences, ETH Zurich, Zurich, 8093, Switzerland

⁺ Deceased author

v2 First published: 19 Dec 2018, 7:1956 (<https://doi.org/10.12688/f1000research.17082.1>)
 Latest published: 03 Apr 2019, 7:1956 (<https://doi.org/10.12688/f1000research.17082.2>)

Abstract










Next generation sequencing protocols such as RNA-seq have made the genome-wide characterization of the transcriptome a crucial part of many research projects in biology. Analyses of the resulting data provide key information on gene expression and in certain cases on exon or isoform usage. The emergence of transcript quantification software such as Salmon has enabled researchers to efficiently estimate isoform and gene expressions across the genome while tremendously reducing the necessary computational power. Although overall gene expression estimations were shown to be accurate, isoform expression quantifications appear to be a more challenging task. Low expression levels and uneven or insufficient coverage were reported as potential explanations for inconsistent estimates. Here, through the example of the ketohexokinase (*Khk*) gene in mouse, we demonstrate that the use of an incorrect gene annotation can also result in erroneous isoform quantification results. Manual correction of the input *Khk* gene model provided a much more accurate estimation of relative *Khk* isoform expression when compared to quantitative PCR (qPCR measurements). In particular, removal of an unexpressed retained intron and a proper adjustment of the 5' and 3' untranslated regions both had a strong impact on the correction of erroneous estimates. Finally, we observed a better concordance in isoform quantification between datasets and sequencing strategies when relying on the newly generated *Khk* annotations. These results highlight the importance of accurate gene models and annotations for correct isoform quantification and reassert the need for orthogonal methods of estimation of isoform expression to confirm important findings.




Keywords

RNA-seq, quantification, gene expression, transcriptomics, *Khk*, Salmon, alternative splicing

Open Peer Review

Referee Status: 

	Invited Referees		
	1	2	3
REVISED			
version 2	report	report	report
published 03 Apr 2019			
version 1			
published 19 Dec 2018	report	report	report

- Patrick K. Kimes** , Dana-Farber Cancer Institute, USA
Harvard TH Chan School of Public Health, USA
- Luisa Mayumi Arake de Tacca**, University of California, Berkeley, USA
Stephen N Floor , University of California, San Francisco (UCSF), USA
- Magnus Rattray** , University of Manchester, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Christophe D. Chabbert (chrischabbert@gmail.com), Werner J. Kovacs (werner.kovacs@biol.ethz.ch)

Author roles: **Chabbert CD:** Conceptualization, Data Curation, Formal Analysis, Methodology, Project Administration, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Eberhart T:** Formal Analysis, Investigation, Methodology, Project Administration, Validation, Visualization, Writing – Review & Editing; **Guccini I:** Methodology, Project Administration, Validation, Writing – Review & Editing; **Krek W:** Funding Acquisition; **Kovacs WJ:** Methodology, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing

Competing interests: CDC is a full-time employee of Roche AG and a shareholder in AstraZeneca.

Grant information: This work was supported in part by a donation from Dr. Walter and Edith Fischli to W.K.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Chabbert CD *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Chabbert CD, Eberhart T, Guccini I *et al.* **Correction of gene model annotations improves isoform abundance estimates: the example of ketohexokinase (*Khk*)** [version 2; peer review: 3 approved] F1000Research 2019, 7:1956 (<https://doi.org/10.12688/f1000research.17082.2>)

First published: 19 Dec 2018, 7:1956 (<https://doi.org/10.12688/f1000research.17082.1>)

REVISED Amendments from Version 1

In this new version of the article, we show that adjusting *Khk* GENCODE UTR annotations prior to removing *Khk.R1* did not suffice to correct biases in *Khk.R1* expression estimates. We also verified that this modification alone was improving agreements between datasets and sequencing strategies but not to the levels observed when the *Khk.R1* transcript is removed. [Figure 3](#) and [Figure 4](#) have been updated to reflect these new results.

In addition, we used StringTie as an example of orthogonal quantification strategy to show that the observed biases were consistent across at least two different methods. These findings are reported in [Extended Data, Figure 10](#).

As modifying annotations can have a strong effect on gene-level abundance estimates, we assessed the impact of isoform annotation modifications on the overall *Khk* expression levels. We provide evidence that none of the changes brought to the *Khk* annotation had a major impact on gene-level abundance estimates (see new [Extended Data, Figure 7](#)).

We also discuss the potential impact of the choice of transcript reference databases in the discussion.

Finally, to improve the readability and coherence of the manuscript, we have clarified how expected *Khk* isoform expression levels were determined and have updated misleading phrasing in the result and introduction sections.

See referee reports

Introduction

Accurate measurement of mRNA expression levels is a crucial component in many modern biological studies. Common and standardized techniques such as reverse transcription real-time quantitative PCR (RT-qPCR) have remained limited in throughput, only allowing measurements for a handful of genes at a time. The emergence of Next Generation Sequencing (NGS) based protocols such as RNA-seq has overcome this limitation and enabled researchers to profile mRNA expression at the genome wide level¹⁻³. While such experiments are now routinely performed, the subsequent bioinformatics analysis and data interpretation still pose computational challenges. As sequencing reads are currently much shorter (usually 100bp – 150bp) than most isoforms, tailored approaches are necessary to study complex events such as splicing or isoform usage switch. In addition, low number of replicates per condition together with a high dynamic range in expression levels across the genome require appropriate statistical frameworks⁴⁻⁸.

One common approach to analyse RNA-seq datasets consists in identifying significant changes in expression levels between two or more experimental conditions using gene-level counts⁹. Such counts are usually obtained from the alignment of sequencing reads to a reference genome or transcriptome when available, and a subsequent counting step during which reads are assigned to annotated genes based on their mapping locations. In the absence of a large number of biological replicates, the following statistical analysis usually requires a reliable estimation of count dispersions for each gene^{4,6-8}. Statistical tools such as edgeR⁶, DESeq2⁴ or limma⁵ offer a panel of solutions to this problem

and have been shown to perform equally well in settings where few biological replicates are available¹⁰. Nevertheless, despite such progress in differential gene expression (DGE) analysis, investigating changes of splice variants and the methods for their quantification continue to be an active field of research. Being able to accurately measure such changes is all the more crucial since they are connected to various biological processes and pathologies and may be used as biomarkers and therapy targets¹¹.

Indeed, although early approaches have followed a framework similar to gene level studies and used exon level counts to implement differential exon usage analysis^{8,12-14}, results from such studies often remain difficult to interpret given the complexity of mammalian transcriptional units. In addition, the recent development of alignment-free transcript quantification methods has provided the possibility to efficiently and rapidly quantify each individual transcript¹⁵⁻¹⁹. Such approaches are indeed computationally much less demanding and faster than alignment-based methods¹⁵. Moreover, they have been shown to overcome the difficulty of handling multi-mapped reads, which can create biased results in count-based analysis²⁰. Although transcript quantification estimates may be used to improve gene-level inference in DGE²¹, testing for changes in isoform usage between conditions remains a challenging task with most approaches focusing on junction and exon read counts rather than transcript quantifications themselves. DESeq2-tximport²¹, sleuth²² and DRIMSeq²³ do make use of such quantifications, with sleuth incorporating estimates of inferential variances obtained during the quantification step. In contrast, DRIMSeq relies on a Dirichlet-multinomial model to estimate relative transcript usage and tests for differential transcript usage.

Regardless of the method used, several studies have now reported limitations and pitfalls associated with transcript abundance estimations^{21,24}. Systematic errors in estimation may stem from sample-specific GC content biases²⁵, which should be accounted for when comparing conditions. In addition, a systematic assessment of quantification performance on simulated datasets also revealed a weaker accuracy in transcript abundance estimates when compared to gene abundance estimates²¹. To explain such discrepancies, it has been suggested that certain transcript abundances cannot be reliably estimated from the data, in particular in cases where coverage is lacking in genomic regions allowing a distinction between transcripts²¹. To our knowledge, no systematic evaluation of the impact of the presence of low coverage on such key regions has been conducted and detailed reports of such examples in real datasets are still missing.

Additionally, quantification tools rely on an input reference transcriptome to compute quantifications with the exception of Cufflinks¹⁷, casper²⁶ and FlipFlop²⁷, which may be used to assemble *de-novo* transcripts. These tools are therefore limited to current gene models made available in databases such as Ensembl or RefSeq and it is unclear whether these models properly recapitulate the actual complexity of transcriptional units. An assessment of the impact of erroneous or incomplete annotations on transcript

quantifications is still missing. Additionally, meticulous examination of the concordance between transcript quantification and mRNA isoforms measured using gene-tailored experimental methods has not been undertaken.

In this study, we focus on the murine ketohexokinase (*Khk*) gene to better understand and evaluate the impact of genomic annotation on transcript quantifications. *Khk*, also known as fructokinase, is the first rate-limiting enzyme in the fructose metabolic pathway and catalyzes the conversion of fructose and ATP to fructose-1-phosphate (F1P) and ADP, respectively. Previous studies have shown that this gene predominantly expresses two usually exclusive isoforms, *KhkA* and *KhkC* that are generated via the specific excision of exon 3C and 3A, respectively²⁸ (Figure 1A). With a greater affinity for fructose²⁹, *KhkC* is thought to be responsible for the functional role of this gene in metabolism, in particular in liver where it is highly expressed³⁰. Epidemiological and animal studies implicate overconsumption of fructose in the development of non-alcoholic fatty liver disease. While the physiological substrate of *KhkA* is unknown, several studies have highlighted the importance of *Khk* isoforms choice in the development of clear cell renal cell carcinoma (ccRCC), hepatocellular carcinoma (HCC)³¹, and pathological cardiac hypertrophy³². Given the clear importance of *Khk* isoforms expression in several disease settings, it is therefore crucial to accurately quantify these variants in order to understand relevant biological mechanisms.

By re-processing publicly available RNA-seq data^{33,34}, we confirm that *Khk* isoforms are differentially expressed in various mouse tissues. Using DRIMSeq proportion estimations, we show that quantification of these isoforms as output by Salmon is biased by the presence of an annotated retained intron that is expressed at very low levels. We also highlight the importance of correct 3' and 5' UTR annotation to improve transcript quantification estimates and validate our computational findings by RT-qPCR. Finally, through the comparison of various datasets, we illustrate the importance of using correct annotations to avoid the emergence of discrepancies between library preparation protocols and datasets.

Results

Khk isoforms expression is tissue-specific

In order to assess tissue-specific expression patterns of *Khk* in mouse, we downloaded RNA-seq data generated from 14 different mouse tissues³³. The availability of 4 biological replicates (2 males and 2 females) per tissue enabled us to conduct a differential gene expression analysis using standard, gene-count based analytical workflows. As previously described in gene specific studies³⁵, we identified a strong tissue specific expression of *Khk* (adjusted p-value of 0 from DESeq2), with significantly higher expression levels in the liver, small intestine and kidney when compared to other tissues (Figure 1B Underlying data: Table 1). Interestingly, gender did not impact the overall *Khk* expression levels. These results were

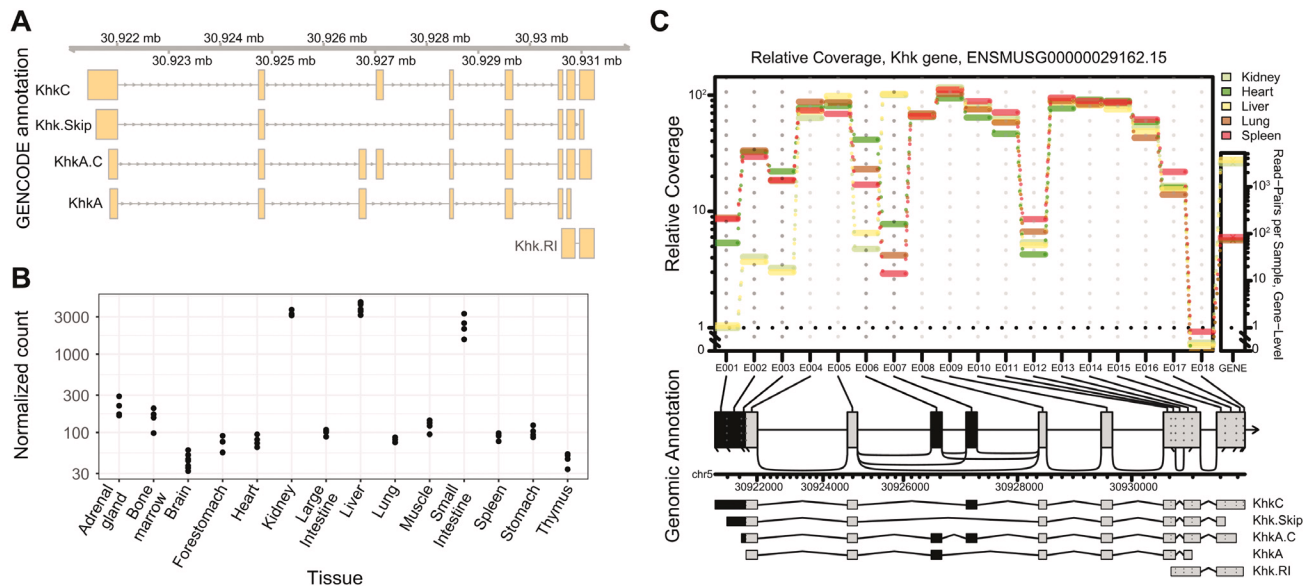


Figure 1. Murine *Khk* expression and splicing patterns are tissue dependent. **A** – *Khk* gene model provided by the Ensembl and GENCODE. Genomic coordinates are indicated on the top ribbon. Usual isoform names are indicated. Murine *Khk* is thought to express 4 main protein coding isoforms (*KhkA*, *KhkC* and ENSMUST00000201571.3 and ENSMUST00000031053.14 termed *Khk.Skip* and *KhkA.C* for this study) and one isoform with a retained intron (*Khk.RI*). **B** – Normalized *Khk* expression levels across mouse tissues (data from Li *et al.*, 2017³³) using gene count tables as input. The liver, kidney and small intestine are clearly expressing *Khk* mRNA at higher levels compared with other tissues. **C** – Relative exon coverage of the *Khk* gene. Normalized exon counts are indicated in the central panel, while normalized gene expression counts are plotted to the right. Exons coloured in black were called as differentially used across all considered tissues (adjusted p value < 0.001). *KhkC* and *KhkA* expression are mutually exclusive in each tissue, with *KhkC* strongly expressed in liver and kidney while *KhkA* is expressed in heart, lung, and spleen.

all confirmed using the gene level estimates based on Salmon quantifications (Extended Data Figure 1A and B). In particular, patterns of *Khk* expression across tissues were highly concordant between count-based and Salmon estimates. We also sought to evaluate changes in relative exon usage for this gene using JunctionSeq¹⁴, including both exon and junction counts in our analysis. Therefore, we selected five tissues (liver, spleen, lung, heart, kidney) with known variations in *Khk* isoform usage^{30–32,35} and different levels of overall expression. We clearly identified a preferential inclusion of exon 3C (Ensembl ID ENSMUSE00000186455) in liver and kidney, as previously described, while exon 3A (Ensembl ID ENSMUSE00001361691) is preferentially retained in heart, spleen and lung (p value < 0.001) (Figure 1C). This trend was also reflected in junctions spanning these exons (Underlying data: Table 2). Variations in 5' UTR were also observed, most likely resulting from alternative transcription start site usage in liver and kidney. In summary, using traditional count-based methods, we confirmed the tissue-specificity of *Khk* expression and identified exons preferentially retained in some tissues.

An annotated, unexpressed retained intron biases *Khk* isoform quantification

As count-based methods might not always reflect the full complexity of splicing patterns and isoform diversity, we sought to quantify relative proportions of *Khk* isoforms in each tissue to fully capture the complexity of its expression patterns. The GENCODE annotation together with the Salmon¹⁵ quantifier were used to obtain these estimations (Underlying data: Table 3). Interestingly, the quantification results showed that the annotated retained intron (*Khk.RI*) accounts for more than 15% of expressed isoforms in 8 tissues (Figure 2A). This trend was consistently observed in DRIMSeq²³ proportion estimates (Figure 2A; Underlying data: Table 3) and the raw TPM (transcripts per million) values output by Salmon (Extended data: Figure 2). Nevertheless, examination of the coverage tracks generated from the alignment of the reads to the reference genome showed hardly any detectable expression of the transcript (Figure 2B; Extended data: Figure 3). This observation was also supported by the differential exon usage analysis which clearly revealed very low expression levels for the only *Khk.RI*-specific exonic

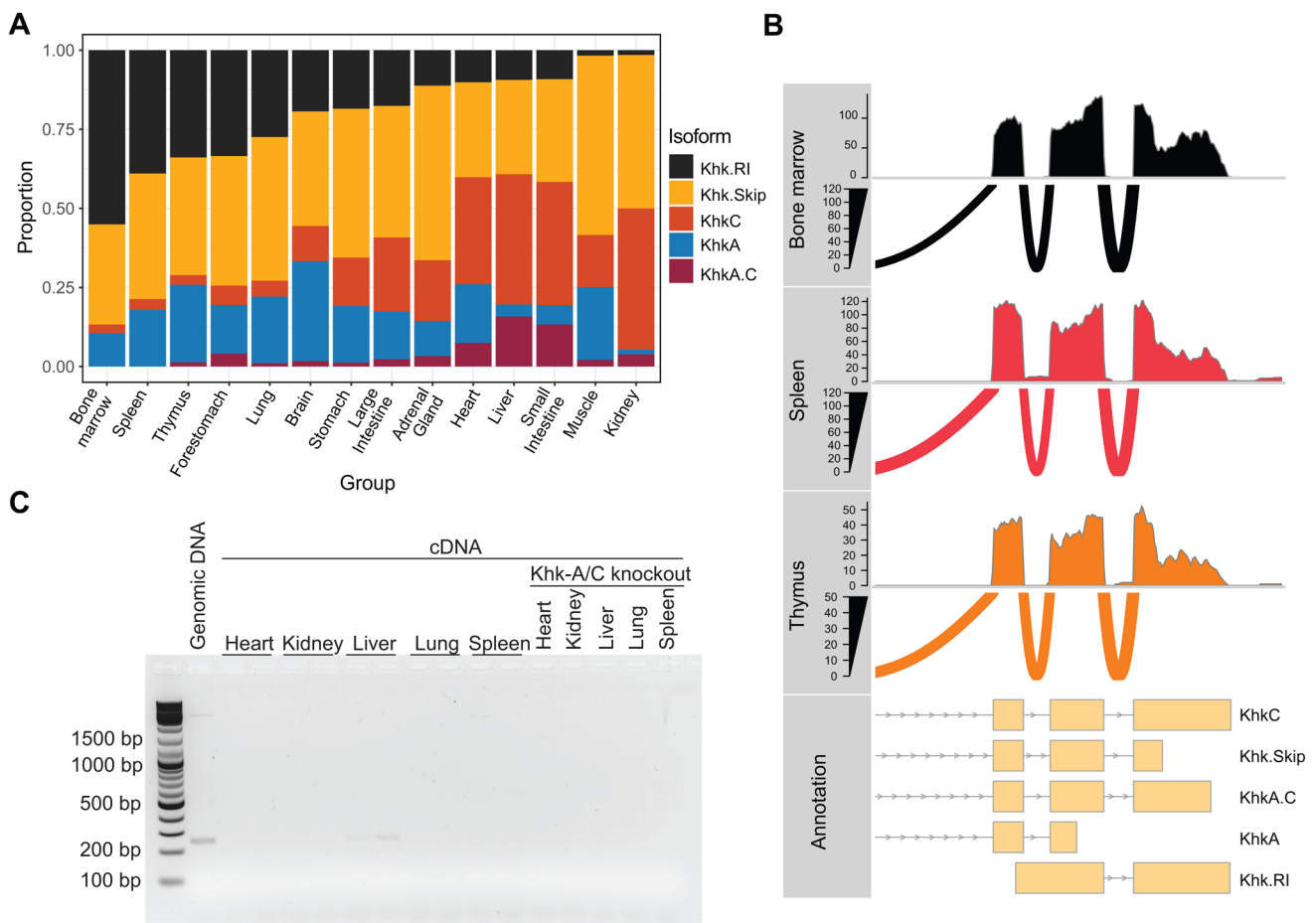


Figure 2. *Khk.RI* expression levels are overestimated using Salmon quantifications. **A** – Estimated relative *Khk* isoform expression across all tissues in mouse. Proportions were output by DRIMseq using Salmon quantification as an input. **B** – RNA-seq coverage tracks and sashimi plots highlighting the absence of *Khk.RI* expression in thymus, spleen and bone marrow samples. Data obtained from all biological replicates were merged prior to plotting. **C** – Products derived from semiquantitative RT-PCR analysis on cDNAs prepared from total RNA of different mouse organs using the primers specific for *Khk.RI*. Genomic DNA isolated from the liver was used as positive and RNA isolated from organs of *Khk-A/C* knockout mice as negative control. (n = 5 and products from two representative mice are shown).

region, E012 (Figure 1C; Extended data: Figure 4A). To experimentally confirm the absence of *Khk.RI* expression, we designed PCR primers amplifying fragments specific to this transcript (see Extended data: Figure 4B; Extended data: Table 1 for a list of primers used in the study) and used RT-qPCR and semiquantitative RT-PCR to measure its expression levels. *Khk.RI* could hardly be detected using this sensitive method (Extended data: Figure 4C and 4D) and comparison with the expression levels in a *Khk-A/C^{-/-}* mouse model showed that it is hardly expressed in heart, kidney, liver, lung and spleen, whereas a product could be amplified using genomic DNA as template (Figure 2C). This experimental validation further demonstrates that the contribution of *Khk.RI* to the overall *Khk* expression level was overestimated during the quantification step. Taken together, these results suggest that the presence of non-expressed transcripts in the “raw” gene annotation may result in erroneous detection by a transcript quantification software.

Manual update of the *Khk* transcript annotations improved quantification results

Since the current genomic annotation did not reflect *Khk* isoform expression and introduced biases in quantifications, we manually removed the *Khk.RI* transcript (Ensembl ID ENSMUST00000200978.1) prior to the quantification step. Despite this adjustment, inspection of the junction reads, coverage tracks, and normalised exon and junction counts derived from QoRTs (Figure 1C) revealed discrepancies between quantification estimates and results derived from alignment-based methods (Figure 3A and 3B; Extended Data: Figure 5). These quantification estimates also did not reflect the observations made by previous reports focusing on the characterisation of *Khk* expression patterns³⁵. An example of such discrepancy may be found examining the heart coverage tracks: while it is quite clear that the *KhkC*-specific exon 3C (Ensembl ID ENSMUSE00000186455) is hardly captured in comparison with exon 3A (Ensembl ID ENSMUSE00001361691), quantification estimates yielded a score of 39.9% and 16.7% for *KhkC* and *KhkA*, respectively (Underlying data: Table 3). Similarly, *KhkC* estimates were inflated in lung and spleen (14.9% and 18.6% to be compared with the absence of coverage of exon 3C), as were *Khk.Skip* estimates (32%) in liver (Underlying data: Table 3). Since such quantifications are relying on the reference transcripts provided during the indexing step¹⁵, we reasoned that incorrect transcript models might be the cause of the observed discrepancies and therefore compared coverage tracks, normalised exon counts and isoform models. We identified annotated differences in 3' end annotations between all isoforms which were not reflected on our coverage tracks (Figure 1C and Figure 3B). As *Khk* isoforms can be identified unambiguously based on the exclusion patterns of the exons 3A and 3C and regardless of differences in UTRs, we could investigate the impact of these UTR variations on transcript quantifications. We therefore manually updated *Khk* isoform annotations to provide an identical 3' end to all isoforms (Underlying data: File 1 and 2) and re-estimated isoform proportions (Figure 3A, second panel). Despite this adjustment, inspection of the proportion estimates still revealed erroneous estimations of isoform expression in particular in the case of liver where *KhkA* was detected in levels similar to *KhkC*.

Further examination of the results revealed that, while some differences in 5' end coverage in the dataset were concordant with the current gene annotation, they were not always reflected in the gene model (Figure 1C and Figure 3A, heart, lung and spleen 5' UTR coverage). Following a similar approach to the one described earlier for 3' UTRs, we finally manually modified *Khk* isoforms to provide an identical 5' and 3' end to all listed isoforms (Underlying data: File 3). Transcript quantification performed using this updated annotation yielded more concordant results when compared to coverage tracks and in light of previous reports, in particular for tissues such as liver³⁰, small intestine³⁶, heart³², spleen and lung (Figure 3A and 3B; Extended data: Figure 5). Interestingly, a substantial fraction of isoforms detected in kidney were still attributed to the *Khk.Skip* isoform while both junction count analysis and single gene studies reported a prevalence of *KhkC*³⁵. Additionally, we computed the relative *Khk* isoform usage using a new annotation with identical 5' and 3' end for all isoforms except *Khk.RI* which was retained as such in the gene model (Figure 3A). This modification was not sufficient to remove *Khk.RI* estimation biases, with the retained intron predicted to erroneously account for 20% of the overall gene expression in bone marrow or spleen. We therefore confirmed the impact and importance of both *Khk.RI* and UTRs annotations on *Khk* isoform expression estimates.

To confirm the biological relevance of our newly estimated proportions, we designed primer pairs to specifically target *Khk* isoforms (see Extended data: Figure 6A and 6B; Extended data: Table 1) and evaluated their relative expression using RT-qPCR. The expression of total *Khk* was ~60-fold higher in liver and kidney compared to heart, lung, and spleen (Figure 3C), confirming previously described findings³⁵. While we did not manage to achieve a reliable quantitative evaluation of *Khk.Skip* and *KhkA.C* expression alone, we accurately measured the expression of (*KhkC* + *KhkA.C*) and (*KhkA* + *KhkA.C*) in five mouse tissues and normalized it to the total *Khk* expression (Figure 3C and 3D). We thereby confirmed a strong prevalence of (*KhkA* + *KhkA.C*) expression in heart while (*KhkC* + *KhkA.C*) accounted for most of the expression measured in kidney and liver (Figure 3D). Therefore, we concluded that *KhkA.C* levels of expression were much lower than *KhkA* and *KhkC* in these three tissues and that the proportion estimates derived from our adjusted genomic annotation reflected the RT-qPCR measurements. The isoform measurements in lung and spleen highlighted low levels of *KhkA*, *KhkC* and *KhkA.C* compared to total *Khk* expression, strongly suggesting the prevalence of *Khk.Skip* expression, as observed on coverage tracks (Figure 3B) and in the proportion estimates derived from the updated genomic annotation (Figure 3A).

Finally, we evaluated whether the changes brought to the *Khk* transcript annotations affected the estimations of overall *Khk* expression levels. Gene level estimates obtained using quantifications based on 3 of the modified annotations were strongly correlated (Pearson, $r > 0.99$) with estimates derived from the raw GENCODE annotation (Extended data: Figure 7A). In addition, the high tissue specificity of *Khk* expression was observed in all cases, with identical expression

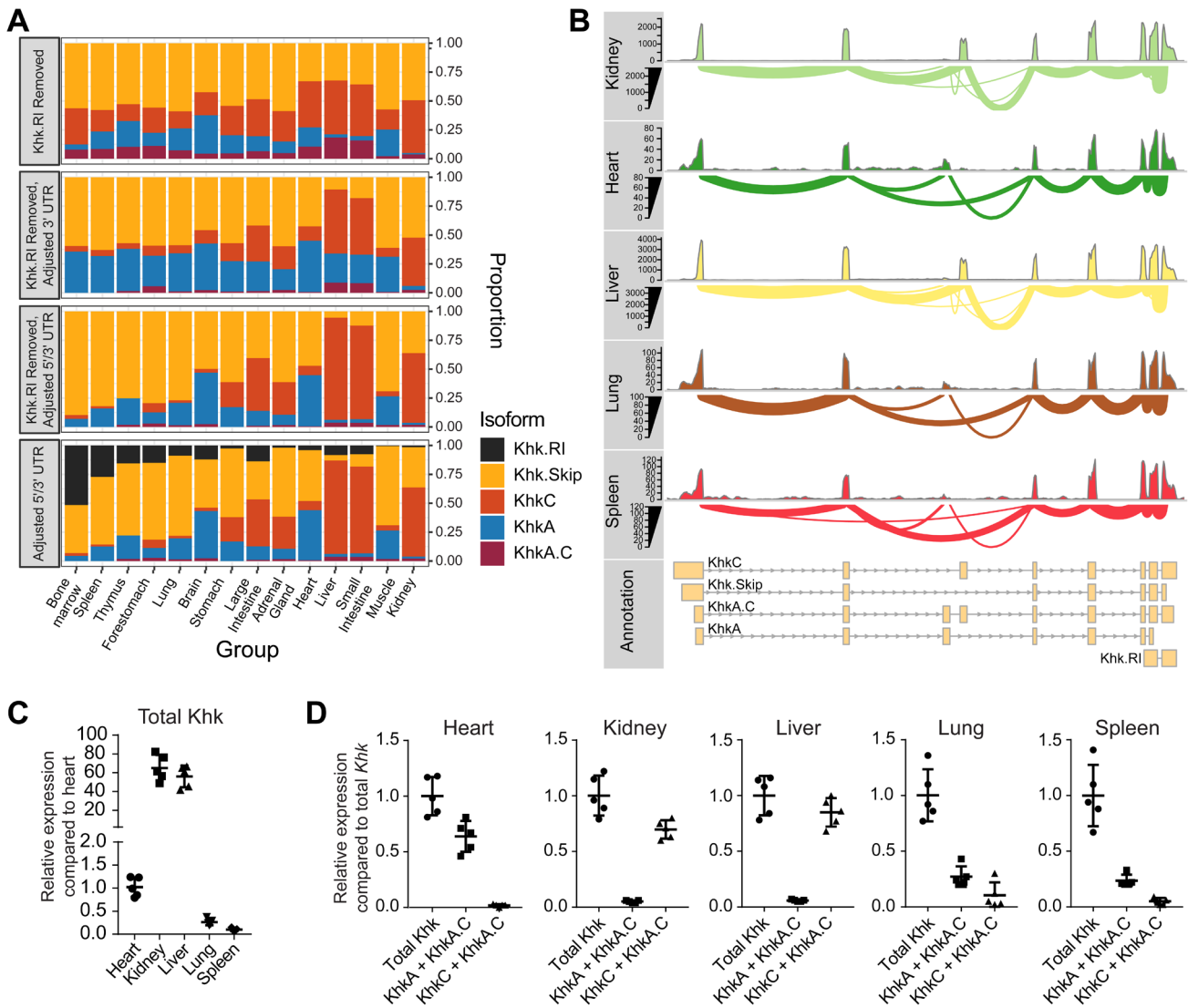


Figure 3. *Khk* isoform quantification estimates are improved after manual adjustment of the genomic annotation. **A** – Estimated relative *Khk* isoform expression across all tissues in mouse, using four different genomic annotations: a raw annotation after removal of the *Khk.RI* transcript, an annotation without *Khk.RI* and with adjusted 3' UTR, an annotation without *Khk.RI* and adjusted 3' and 5' UTR, and an annotation with adjusted 3' and 5' UTR for all transcripts except *Khk.RI*. The choice of annotation greatly impacts the quantification results. **B** – RNA-seq coverage track and sashimi plots illustrating the predominance of different isoforms in spleen, lung, liver, heart, and kidney. **C** – RT-qPCR analysis of total *Khk* expression in different mouse organs. Values are expressed as fold-change compared to the expression levels obtained for the heart, which was arbitrarily defined as 1. β -actin was used as the invariant reference gene. Data are mean \pm SD (n = 5). **D** – RT-qPCR analysis of total *Khk*, *KhkA* + *KhkA.C*, and *KhkC* + *KhkA.C* expression in different mouse organs. Values are expressed as fold-change compared to the expression levels obtained for total *Khk*, which was arbitrarily defined as 1. Data are mean \pm SD (n = 5).

patterns between annotations (Figure 1A, Extended data: Figure 7B).

Altogether, these findings demonstrate that UTRs and more generally 5' and 3' end annotation may greatly influence transcript quantification results. The resulting biases lead to the identification of isoforms that can hardly be detected in biological samples, therefore highlighting the importance of inspecting results for any given gene of interest.

Erroneous genomic annotation of *Khk* increases discrepancies between sequencing strategies and datasets

We next investigated whether such discrepancies could be observed when using different sequencing library strategies. To do so, we artificially created a single-end dataset by removing one read for each tissue sample. We also created a short-read dataset by trimming the remaining reads to only retain the first 50bp. In both cases, we aligned reads to the reference genome and independently quantified transcript expressions using

Salmon as previously described. Regardless of the sequencing strategy considered, we observed a similar overestimation of the *Khk.RI* fraction as well as discrepancies between proportion estimates and junction counts (Extended data: Figure 8A and 8B). Both differences were corrected using the aforementioned manually curated annotations. We then compared proportion estimates between datasets for each annotation. Quite strikingly, we noted a much better agreement in transcript estimates using our manually modified annotation (Figure 4A–D). While the use of the “raw” annotation only resulted in a 0.56 correlation (Pearson) between estimates from the paired-end and 50bp single-end libraries, the use of an updated annotation resulted in a 0.97 correlation between both platforms. This trend was also observed between paired-end and single-end and between single-end and 50bp single-end respectively (Extended data: Figure 9A–F). To further evaluate the impact of annotations on estimate concordance across datasets, we downloaded RNA-seq data from another study profiling mRNA expression across 13 tissues³⁴. We quantified isoform usage for each tissue and compared those estimates with the ones from the 50bp single-end dataset from Li *et al.* 2017³³ in order to avoid biases due to differences in read length. In total, 8 tissues could be compared between both studies. Quite strikingly, the agreement between both datasets was not high (Pearson correlation 0.72) when considering fractions derived using the original annotation (Figure 4E). While the removal of *Khk.RI* without an adjustment of the UTR did further hinder reproducibility between both datasets (Figure 4F), we observed a much stronger consistency of proportion estimates after UTR adjustment (Figure 4H, Pearson correlation 0.73). However, the highest consistency between datasets was reached when using the fully updated annotation (Figure 4G, Pearson correlation 0.91). These results therefore further underscore the importance of appropriate annotations during transcript isoform quantifications. The use of erroneous gene models may further increase discrepancies between sequencing libraries and datasets as exemplified in the case of the *Khk* gene.

Discussion

Gene and transcript quantifications are essential steps in many genomic studies where the characterization of gene expression patterns is of biological relevance. The recent development of quasi-mapping methods such as Salmon¹⁵ has drastically improved the computational speed of these quantifications steps while relying on reduced computational power. However, unlike more traditional and alignment-based methods, they strongly rely on the provided genomic annotation. Through the example of *Khk* gene expression in mouse, we describe the importance of using a properly curated annotation to avoid biases and erroneous isoform proportion estimates. We show that the inclusion of an annotated, yet not detected retained intron (*Khk.RI*) was sufficient to wrongly predict isoform usage in several tissues. We also found that differences in 5' and 3' end annotations may result in inaccurate transcript quantifications. Manual adjustment of such differences resulted in a better agreement between isoform proportion estimates, coverage tracks inspections, junction counts and qPCR results in at least 3 tissues. Both the removal of *Khk.RI* and UTR adjustments were necessary to reach this concordance between

profiling methods. Finally, comparison of these estimates across different datasets and sequencing library strategies revealed that the use of a corrected annotation strongly improves the reproducibility of estimations between each dataset.

The use of gene or exon level counts to assess differences in gene expression or exon usage between conditions has been described as a robust method by several independent studies⁸. Recent reports²¹ have also emphasized the reliability of gene-level quantification estimates and their biological relevance. It was therefore reassuring to observe a very good agreement between both methods when assessing the tissue specificity of *Khk* expression in this dataset. Identification of higher levels of expression in liver, small intestine and kidneys reflects previously described findings³⁵. We thereby provide further evidence of the reliability of gene quantification, albeit at the single gene level. Additionally, the variations observed in expression levels across tissues together with the previously reported alternative splicing events make *Khk* an ideal gene to study performance of bioinformatics tools.

We identified strong discrepancies between proportion estimates and coverage tracks when quantifying *Khk* isoforms while relying on the original GENCODE/Ensembl annotation. In particular, the annotated retained intron *Khk.RI* (ENSMUST00000200978.1) was identified as a predominantly expressed isoform in several tissues while hardly any read could be mapped to the genomic region specific to this isoform. RT-qPCR validations confirmed the extremely low levels of *Khk.RI* in 5 mouse tissues. Previous work relying mostly on simulated data showed that in the case of lowly expressed genes, transcript-level estimates lack accuracy²¹. However, through the instance of *Khk*, we provide a concrete example of a strong overestimation in transcript quantification and show that this is not limited to tissues with very low expression of the corresponding gene.

Differential exon usage analysis clearly revealed, as expected, a preferential usage of either exon 3A (Ensembl ID ENSMUSE00001361691) or 3C (Ensembl ID ENSMUSE00000186455) in various tissues³⁵. The discovery of a potential change in 5' transcription start site, while not previously described for *Khk*, further underscores the importance of alternative start and termination sites in transcript isoform diversification in mammals³⁷. Interestingly, inspection of the usage of other exons revealed that these new start sites are most likely specific to the retention of exon 3A or 3C, therefore suggesting that the determination of the isoform choice for *Khk* could be achieved solely by considering junction and exon counts in this region.

This idiosyncrasy was further confirmed by the various isoform proportion estimations performed using updated annotations of the *Khk* gene. Estimations reflected experimental measurements, junction counts and coverage tracks only when both 5' and 3' end annotations were harmonized across all annotated isoforms. Importantly, when using the naive GENCODE annotation, Salmon and DRIMseq failed to reliably quantify isoform proportions. It is also important to note that the

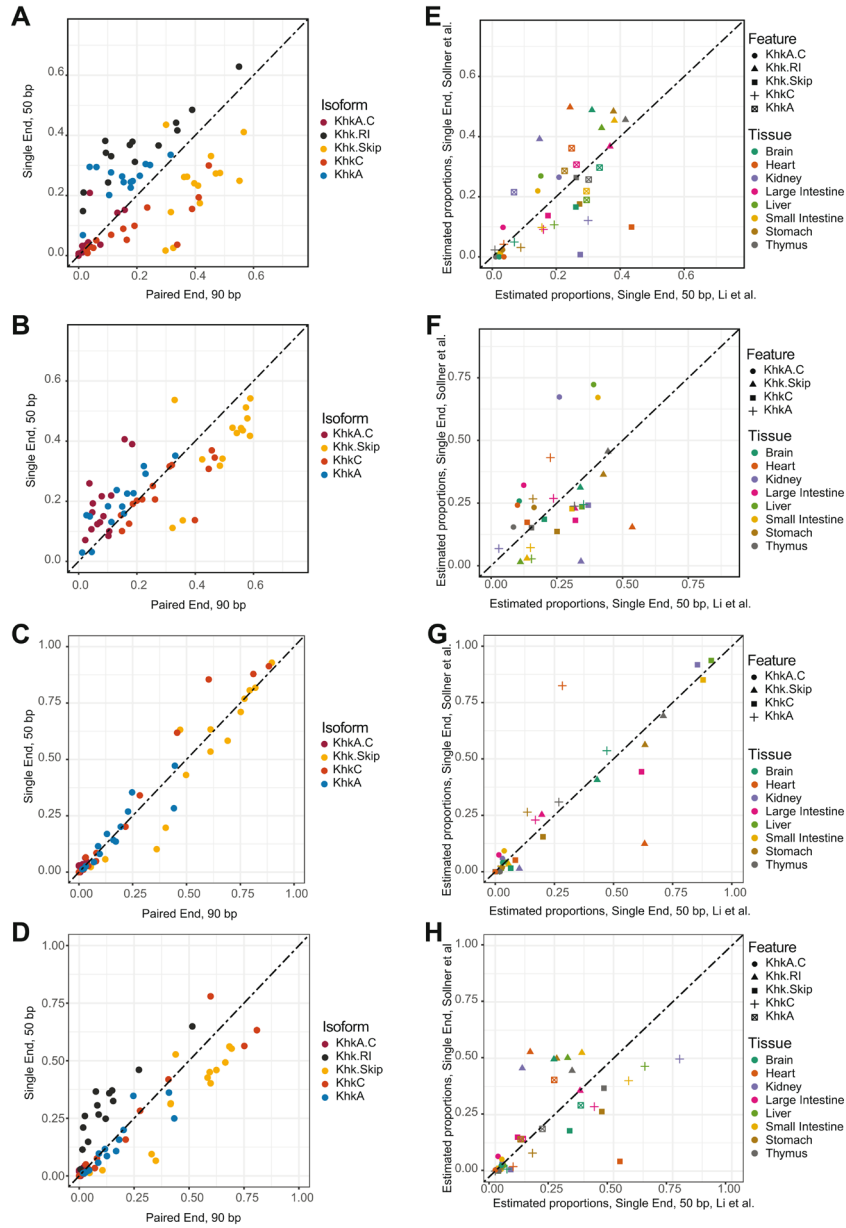


Figure 4. Manual adjustment of the *Khk* gene model improves the concordance between estimates across sequencing library strategies and datasets. **A** – Comparison of relative *Khk* isoforms expression estimates between the full length paired-end dataset from Li *et al.* 2017³³ and the short read, single-end dataset. The estimates were generated using the naive Ensembl annotation. **B** – Comparison of relative *Khk* isoforms expression estimates between the full length paired-end dataset from Li *et al.*, 2017³³ and the short read, single-end dataset. The estimates were generated using the modified Ensembl annotation where the *Khk.RI* transcript has been removed. **C** – Comparison of relative *Khk* isoforms expression estimates between the full length paired-end dataset from Li *et al.*, 2017³³ and the short read, single-end dataset. The estimates were generated using the modified Ensembl annotation where the *Khk.RI* transcript has been removed and the 5' and 3' UTR of other transcripts were all adjusted. **D** – Comparison of relative *Khk* isoforms expression estimates between the full length paired-end dataset from Li *et al.*, 2017³³ and the short read, single-end dataset. The estimates were generated using the modified Ensembl annotation where the 5' and 3' UTR of all transcripts except *Khk.RI* were all adjusted. **E** – Comparison of relative *Khk* isoforms expression estimates between the Li *et al.*, 2017³³ dataset and the Söllner *et al.*, 2017³⁴ dataset, using single-end, short (50bp) reads in each case. The estimates were generated using the naive Ensembl annotation. **F** – Comparison of relative *Khk* isoforms expression estimates between the Li *et al.*, 2017³³ dataset and the Söllner *et al.* 2017³⁴ dataset, using single-end, short (50bp) reads in each case. The estimates were generated using the modified Ensembl annotation where the *Khk.RI* transcript has been removed. **G** – Comparison of relative *Khk* isoforms expression estimates between the Li *et al.*, 2017³³ dataset and the Söllner *et al.* dataset, using single-end, short (50bp) reads in each case. The estimates were generated using the modified Ensembl annotation where the *Khk.RI* transcript has been removed and the 5' and 3' UTR of other transcripts were all adjusted. **H** – Comparison of relative *Khk* isoforms expression estimates between the Li *et al.*, 2017³³ dataset and the Söllner *et al.* dataset, using single-end, short (50bp) reads in each case. The estimates were generated using the modified Ensembl annotation where the 5' and 3' UTR of all transcripts except *Khk.RI* were all adjusted.

curated RefSeq *Khk* gene model differs from GENCODE as it is missing *Khk.RI*, *Khk.Skip* and the 3' and 5' UTRs of all curated transcripts are identical. While this would be a close configuration to the optimal annotation presented in this study, but the absence of *Khk.Skip* in the gene model would result in erroneous quantifications as well. Such misestimations are likely to be observed in other genes for which current annotations are either limited or inaccurately reflect experimental measurements. However, systematic harmonisation of all UTRs across annotated transcripts might not be a general approach, especially in cases when such differences are reflecting tissue-specific expression patterns³⁷.

During the preparation of this manuscript, a preprint from Sonesson *et al.* reported a similar observation and proposed the creation of a new index to flag such problematic genes³⁸. While the current manuscript strongly emphasizes the role of 3' UTRs in the emergence of estimation biases, we could pinpoint at least one example where 5' UTRs play a similar role in the issue. The use of the JCC (Junction Coverage Compatibility) score introduced by Sonesson *et al.* will be greatly useful to prevent misinterpretation of transcriptomics studies in the future but will tie quantifications to the results of computationally demanding alignment methods³⁸. Improvement of current genomic annotations might ultimately offer an alternative as they will allow for the sole use of fast quantification algorithms. This might partially be achieved using transcript catalogues obtained from large scale studies such as CHES³⁹ even though Sonesson *et al.*³⁸ reported very little to no improvement in their JCC scores using these new annotations.

Using an additional dataset from Söllner *et al.*, 2017³⁴ and in-silico single-end and short single-end datasets from Li, B *et al.* 2017³³, we showed that such updated annotations have the potential to reduce discrepancies between methods and experiments. While this is only exemplified at the level of the *Khk* gene, it is very likely that other instances will emerge as new metrics such as JCC will enable scientists to flag problematic genes. The main results of this study exclusively focus on the use of Salmon as a quantification software and DRIM-Seq to estimate relative proportions, in particular as recent reports have suggested that most quantification pipelines might perform similarly²⁴. To complement our main findings, we estimated transcripts abundance in the Li *et al.*, 2017 dataset using StringTie⁴⁰, as an example of a tool relying on the construction of a splicing graph to quantify isoforms. When using the GENCODE annotation as a guide for quantification, we found that StringTie overestimated the *Khk.RI* expression in a fashion similar to Salmon. This is in concordance with the report from Sonesson *et al.*³⁸. In addition, we used StringTie without any supporting annotation to enable transcript assembly from the alignments. Inspection of the resulting newly assembled transcripts showed that *Khk.RI* could not be detected (Extended Data, Figure 10). Nevertheless, *KhkA.C* was also not identified in the dataset while junction counts clearly indicate that it could be detected in a handful of tissues, albeit with low expression levels. We therefore suggest that the issue reported here in the case of Salmon might be commonly found across quantification softwares,

and it will be interesting to assess whether similar biases may arise with more tools. Results from Sonesson *et al.* strongly indicate that this is the case, at least for a group of human genes³⁸.

Finally, the results presented in our study will provide a valuable resource to the scientific community investigating the role of fructose metabolism and *Khk* in mammals. As each *Khk* isoform might harbor different functions, the complete mapping of their usage across tissues will help further pinpoint their role in different biological contexts. Our analysis was conducted on mouse tissues but further exploration of the results presented in Reyes *et al.* 2017³⁷ also showed that exons 3A and 3C of *KHK* are selectively included in human tissues, across individuals. Refining our understanding of these expression patterns in human will be critical in particular as the KHK protein product has already been identified as a promising target in the treatment of non-alcoholic steato-hepatitis (NASH)⁴¹. This study provides important background information to improve the results of the transcriptomic work that might therefore be necessary in the future.

Methods

Mice

C57BL/6J were obtained from The Jackson Laboratory, while *KhkA/C*^{-/-} mice, which are of C57BL/6 background and are lacking both ketohexokinase-A and ketohexokinase-C, were obtained from R. Johnson (University of Colorado) and used as negative control. All mice were housed in a pathogen-free facility at the ETH Phenomics Center (EPIC) under standard conditions (12 h light and 12 h dark cycle) with free access to food and water. 3 female and 2 male C57BL/6J mice and 2 male *KhkA/C*^{-/-} mice were euthanized with CO₂ at the age of 6 weeks and heart, kidney, liver, lung, and spleen were subsequently removed and shock-frozen in liquid nitrogen. The mice did not suffer during the euthanasia with CO₂; the mice were placed into a chamber that contained room air and then CO₂ was gradually introduced with no more than 6 psi to displace at least 20% of the chamber volume per minute. All protocols for animal use and experiments were reviewed and approved by the Veterinary Office of Zurich (Switzerland).

Evaluation of *Khk* isoform expression *in vivo*

Total RNA from C57BL/6J and *KhkA/C*^{-/-} mice was prepared from frozen tissues with RNeasy Mini Kit (QIAGEN, Hilden, Germany) and treated with DNase I to remove traces of DNA. First-strand complementary DNA (cDNA) was synthesized with random hexamer primers using the High-Capacity cDNA Reverse Transcription Kit (Cat. No. 4368813; Applied Biosystems). Quantitative reverse transcription PCR (RT-qPCR) was performed on a Roche LightCycler 480 in duplicates using 10 ng cDNA and the 2x KAPA SYBR FAST qPCR Master Mix LC480 (Sigma). Thermal cycling was carried out with a 5 min denaturation step at 95 °C, followed by 45 three-step cycles: 10 sec at 95 °C, 10 sec at 60 °C, and 10 sec at 72 °C. Finally, melt curve analysis was carried out to confirm the specific amplification of a target gene and absence of primer dimers. Relative mRNA amount was calculated using the comparative threshold

cycle (C_T) method. β -*actin* was used as the invariant reference gene. The PCR amplification efficiency of RT-qPCR primer sets was determined with serial dilutions of liver cDNAs and was similar for all primer sets. Semiquantitative RT-PCR was performed using Phusion High-Fidelity DNA Polymerase (New England Biolabs) and the following 3-step amplification protocol: 30 sec at 98 °C (denaturation), 30 cycles of 10 sec at 98 °C, 30 sec at 63 °C, and 40 sec at 72 °C, and a final elongation step for 5 min at 72 °C. PCR products were evaluated after gel-electrophoresis. Primer sequences are listed in [Extended data: Table 1](#).

Modification of *Khk* transcript annotation

A fasta file containing all manually modified *Khk* transcripts was created. All exonic sequences used to modify the gene model were downloaded from the Ensembl website. A fasta file containing nucleotide sequences of all transcripts from the GENCODE M14 annotation was loaded into R using the Biostings v 2.46.0 package. All original *Khk* transcripts were removed from the DNASTringSet object and the updated transcripts were then added. The resulting annotation was written to a fasta file to generate Salmon indexes (see following sections for more details).

RNA-seq data processing and alignment

Fastq files from Li *et al.*, 2017³³ and Söllner *et al.*, 2017³⁴ were downloaded from SRA using the sra-tools software v2.7.0⁴². As they only provided two replicates (instead of 4), we excluded the testis and ovary samples from Li *et al.*, 2017³³. Additionally, due to very low library complexities, we also excluded the pancreatic samples from Söllner *et al.*, 2017³⁴. Reads were aligned to the M14_GRCm38.p5 reference genome using STAR 2.4.2a⁴³ together with the GENCODE M14 annotation (STAR was run with default parameters). The search for novel junctions was allowed during the mapping step. Gene, exon and junction level read counts were generated using the QoRTs software v1.2.42⁴⁴ after excluding reads with multiple alignments (MAPK score less than 255). All workflows were orchestrated using Snakemake v 3.13.3⁴⁵.

Transcript quantifications

Salmon 0.9.1¹⁵ and StringTie 1.3.3b⁴⁰ were used for transcript quantifications. In the case of Salmon, indexes were built from each fasta files using the default quasi-mapping mode and a k-mer of length 31 as recommended in the software documentation. Transcript isoforms were quantified using the default VEBM algorithm. Library types were inferred by the software. Sequence and GC bias corrections were performed during each quantification (`--seqBias` and `--gcBias` options) and 100 bootstraps were run to estimate the variance of abundance estimates.

In order to quantify transcripts using StringTie, alignments to the reference genome were first re-generated as described in the “RNA-seq data processing and alignment” section with the additional STAR flag “`--outSAMstrandField intronMotif`”. Quantifications were then performed with and without the GENCODE annotation used as a guide (`-G` option in StringTie). All other parameters were set to default values. Newly assembled

transcripts were then merged using the transcript merge mode (`--merge`) with default parameters.

Differential gene expression

Gene count tables were loaded into R (v 3.4.1) as a DESeq2⁴ object to conduct differential gene expression analysis. For the purpose of differential gene expression analysis, we only retained features labelled as “gene” in the GENCODE annotation and of type *protein_coding*, *antisense*, *sense_intronic*, *3prime_overlapping_ncRNA*, *sense_overlapping* or *non_coding* in order to exclude transcription products requiring specific library preparations to be accurately measured. Genes with very low counts were excluded from downstream analysis: the threshold was set at 50 mapped reads across all samples, corresponding to less than one read per sample on average. Estimated size factors were used to correct for differences in library size. Following the standard DESeq2 workflow⁴⁶, changes in gene expression were modelled using a variable accounting for differences in tissue of origin for each sample. A Likelihood Ratio Test (“LRT” option in DESeq2) was performed to compare this model to a reduced model consisting of only an intercept. Results were extracted with the DESeq2 *results* function and multiple testing correction was performed using the Benjamini Hochberg procedure⁴⁷.

Differential exon and junction usage

Exon and junction count tables from QoRTs were loaded in R (v 3.4.1) using the JunctionSeq (v 1.8.0) package¹⁴. Changes in exon usage were modelled using the tissue of origin as a main variable. Size factors and dispersions were estimated using default parameters (options “byGenes” and “advanced” respectively, see the JunctionSeq vignette for more details). Dispersion function fits, test for differential usage and estimation of effect sizes were also run using default parameters. Final test results were extracted using the *writeCompleteResults* function. Feature-level p values were adjusted using the Benjamini Hochberg procedure⁴⁷. Only features with an adjusted p-value below 0.05 were retained.

Estimation of relative isoform proportions

Transcript isoform quantifications from Salmon were loaded into R (v 3.5.1) using the tximport package²¹. Relative isoform proportions and differential transcript usage were then modelled using a Dirichlet-multinomial model as described in Nowicka *et al.*, 2016²³ and implemented in the DRIMSeq package (v 1.6.0). Model precisions were estimated by the *dmPrecision* function run with default parameters, except for the search grid ranges which were set to -15 and 15. For each sample, feature proportions were then computed using the *dmFit* function (default parameters).

Data visualization

Genomic annotations and coverage tracks were plotted using the Gviz package (v 1.22.3)⁴⁸. Data from all available biological replicates were pooled together prior to plotting each track. Results from differential exon and junction usage were visualized using JunctionSeq. Other plots were generated with the ggplot2 package.

Data availability

Underlying data

Fastq files from Li *et al.*, 2017³³ are available from: <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA375882>

Fastq files from Söllner *et al.*, 2017³⁴ are available from: <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6081/>

Scripts used to generate the analysis presented in this paper are freely and publicly available on Github: https://github.com/chbtchris/Khk_quantifications (Archived scripts: <http://doi.org/10.5281/zenodo.2583233>⁴⁹).

All underlying data are available at: <https://doi.org/10.17605/OSF.IO/NMKFA>⁵⁰. Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication). Files available are as follows:

- **Table 1:** Normalized *Khk* counts for each sample (and therefore tissue) considered.
- **Table 2:** Normalized counts for each exonic regions and tissue group output by JunctionSeq.
- **Table 3:** Relative *Khk* isoform expression across all tissues using the different genomic annotations.
- **Table 4:** Relative *Khk* isoform expression across all tissues using the GENCODE annotation and StringTie as a quantification software.
- **File 1:** Fasta file containing the sequences of *Khk* transcript isoforms downloaded from GENCODE.
- **File 2:** Fasta file containing the sequences of *Khk* transcript isoforms after removal of the *Khk.RI* isoform and adjustment of the 3' UTR for all remaining transcripts.
- **File 3:** Fasta file containing the sequences of *Khk* transcript isoforms after removal of the *Khk.RI* isoform and adjustment of the 5' and 3' UTR for all remaining transcripts.
- **File 4:** Fasta file containing the sequences of *Khk* transcript isoforms after adjustment of the 5' and 3' UTR for all remaining transcripts except *Khk.RI*.

Extended data

All extended data are available at: <https://doi.org/10.17605/OSF.IO/NMKFA>⁵⁰. Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication). Files available are as follows:

- **Table 1:** List of qPCR primers used in this study
- **Figure 1:** A - Normalized *Khk* expression levels across tissues in mouse (data from Li *et al.* 2017³³) using transcript abundance estimates from Salmon as input. Similar to observations made using raw counts, the liver, kidney and small intestine are clearly expressing *Khk* mRNA at higher levels compared to other tissues. B – Comparison of gene expression estimates using count tables and transcript abundance estimates.

- **Figure 2:** Transcript Per Million (TPM) values for each annotated isoform of the *Khk* gene as returned by Salmon. The naive Ensembl annotation was used to estimate the abundances.
- **Figure 3:** RNA-seq coverage tracks and sashimi plots highlighting the absence of *Khk.RI* expression in all the tissues investigated in this study. Data obtained from all biological replicates were merged prior to plotting.
- **Figure 4:** A – Normalized fractions of reads mapped to all exonic region overlapping the *Khk.RI* transcript. E12, which is the only *Khk.RI* specific region, clearly show a reduced coverage when compared to other regions. Exonic regions were determined using the JunctionSeq package and their associated genomic coordinates are available in **Underlying Data, Table 2**. B – Schematic representing the location of *Khk.RI* specific primer targets. C – Amplification plots of RT-qPCR analysis of *Khk.RI* expression in different mouse organs (n = 5 mice). Note that the Ct or threshold cycle value at which the fluorescence generated within a reaction crosses the threshold, a numerical value assigned for each run reflecting a statistically significant point above the calculated baseline, is very high (> 40) and can be considered as noise. D – Melt curves from RT-qPCR analysis of *Khk.RI* expression in different mouse organs.
- **Figure 5:** RNA-seq coverage track and sashimi plots illustrating the predominance of different isoforms in each tissue inspected in this study. Data obtained from all biological replicates were merged prior to plotting.
- **Figure 6:** A – Schematic representing the location of *Khk.A*, *KhkC*, *KhkA.C* and total *Khk* specific primer targets. B – Amplification plots of RT-qPCR analysis of total *Khk*, *KhkC*, and *KhkA* expression in different mouse organs (n = 5 mice). C – Melt curves from RT-qPCR analysis of total *Khk*, *KhkC*, and *KhkA* expression in different mouse organs.
- **Figure 7:** A - Normalized *Khk* expression levels across mouse tissues (data from Li *et al.*, 2017³³) using the transcript abundance estimates from Salmon using four annotations used in the study: a naive annotation after removal of the *Khk.RI* transcript, an annotation without *Khk.RI* and with adjusted 3' UTR, an annotation without *Khk.RI* and adjusted 3' and 5' UTR, and an annotation with adjusted 3' and 5' UTR for all transcripts except *Khk.RI*. B – Comparison of the normalized *Khk* expression levels obtained using four different annotations used in the study: a naive annotation after removal of the *Khk.RI* transcript, an annotation without *Khk.RI* and with adjusted 3' UTR, an annotation without *Khk.RI* and adjusted 3' and 5' UTR, and an annotation with adjusted 3' and 5' UTR for all transcripts except *Khk.RI*. All reported coefficients are Pearson correlations.
- **Figure 8:** A - Estimated relative *Khk* isoform expression across all tissues in mouse, using the single-end dataset created from Li *et al.*, 2017³³. Three different genomic annotations were considered: a naive annotation,

a naive annotation after removal of the *Khk.RI* transcript, and an annotation without *Khk.RI* and with adjusted 3' and 5' UTR. The choice of annotation greatly impacts the quantification results. B - Estimated relative *Khk* isoform expression across all tissues in mouse, using the 50bp single-end dataset created from Li *et al.*, 2017³³. Three different genomic annotations were considered: a naive annotation, a naive annotation after removal of the *Khk.RI* transcript, and an annotation without *Khk.RI* and with adjusted 3' and 5' UTR. The choice of annotation greatly impacts the quantification results.

- **Figure 9:** A – Comparison of relative *Khk* isoforms expression estimates between the full length paired-end dataset from Li *et al.*, 2017³³ and the full length, single-end dataset. The estimates were generated using the naive Ensembl annotation. B – Comparison of relative *Khk* isoforms expression estimates between the full length paired-end dataset from Li *et al.*, 2017³³ and the full length, single-end dataset. The estimates were generated using the modified Ensembl annotation where the *Khk.RI* transcript has been removed. C – Comparison of relative *Khk* isoforms expression estimates between the full length paired-end dataset from Li *et al.*, 2017³³ and the full length, single-end dataset. The estimates were generated using the modified Ensembl annotation where the *Khk.RI* transcript has been removed and the 5' and 3' UTR of other transcripts were all adjusted. D – Comparison of relative *Khk* isoforms expression estimates between the full length single-end dataset generated from Li *et al.*, 2017³³ and the short read, single-end dataset. The estimates were generated using the naive Ensembl annotation. E – Comparison of relative *Khk* isoforms expression estimates between the full length single-end dataset generated from Li *et al.*, 2017³³ and the short read, single-end dataset. The estimates were generated using the modified Ensembl annotation where the *Khk.RI* transcript has been removed. F – Comparison

of relative *Khk* isoforms expression estimates between the full length single-end dataset generated from Li *et al.*, 2017³³ and the short read, single-end dataset. The estimates were generated using the modified Ensembl annotations where the *Khk.RI* transcript has been removed and the 5' and 3' UTR of other transcripts were adjusted.

- **Figure 10:** A - Estimated relative *Khk* isoform expression across all tissues in mouse using StringTie as a quantification software and the naive GENCODE annotation. B – Comparison of the *Khk* gene model provided by GENCODE and the newly assembled transcripts from StringTie. Transcripts assembled by StringTie are colored in red (transcript id 25257.1, 25257.2, 25257.3). No transcript could be associated to *Khk.RI* or *Khk.Skip* using the StringTie approach.

Grant information

This work was supported in part by a donation from Dr. Walter and Edith Fischli to W.K.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to thank the whole Krek Laboratory for useful discussions and comments on this manuscript. We also would like to thank Malgorzata Nowicka for her help in getting the DRIMSeq package to run. Finally, we would like to thank Alejandro Reyes for helpful discussions.

Author information

Christophe D. Chabbert is currently at Roche Innovation Center Zurich, Wagistrasse 10, 8952 Schlieren, Switzerland.

References

1. Cloonan N, Forrest AR, Kelle G, *et al.*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods.* 2008; 5(7): 613–619. [PubMed Abstract](#) | [Publisher Full Text](#)
2. Miyoshi T, Kanoh J, Saito M, *et al.*: **Fission yeast Pot1-Tpp1 protects telomeres and regulates telomere length.** *Science.* 2008; 320(5881): 1341–1344. [PubMed Abstract](#) | [Publisher Full Text](#)
3. Goodwin S, McPherson JD, McCombie WR: **Coming of age: ten years of next-generation sequencing technologies.** *Nat Rev Genet.* Nature Publishing Group. 2016; 17(6): 333–351. [PubMed Abstract](#) | [Publisher Full Text](#)
4. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; 15(12): 550. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Ritchie ME, Phipson B, Wu D, *et al.*: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res.* 2015; 43(7): e47. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. McCarthy DJ, Chen Y, Smyth GK: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.** *Nucleic Acids Res.* 2012; 40(10): 4288–4297. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; 26(1): 139–140. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from RNA-seq data.** *Genome Res.* 2012; 22(10): 2008–2017. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Huber W, Carey VJ, Gentleman R, *et al.*: **Orchestrating high-throughput genomic analysis with Bioconductor.** *Nat Methods.* Nature Publishing Group. 2015; 12(2): 115–121. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Schurch NJ, Schofield P, Gierliński M, *et al.*: **How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?** *RNA.* 2016; 22(6): 839–851. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Climente-González H, Porta-Pardo E, Godzik A, *et al.*: **The Functional Impact of Alternative Splicing in Cancer.** *Cell Rep.* 2017; 20(9): 2215–2226. [PubMed Abstract](#) | [Publisher Full Text](#)
12. Reyes A, Anders S, Weatheritt RJ, *et al.*: **Drift and conservation of differential exon usage across tissues in primate species.** *Proc Natl Acad Sci U S A.* 2013; 110(38): 15377–15382. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Law CW, Chen Y, Shi W, *et al.*: **voom: Precision weights unlock linear model**

- analysis tools for RNA-seq read counts.** *Genome Biol.* 2014; **15**(2): R29.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Hartley SW, Mullikin JC: **Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq.** *Nucleic Acids Res.* 2016; **44**(15): e127.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 15. Patro R, Duggal G, Love MI, *et al.*: **Salmon provides fast and bias-aware quantification of transcript expression.** *Nat Methods.* Nature Publishing Group. 2017; **14**(4): 417–419.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 16. Patro R, Mount SM, Kingsford C: **Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.** *Nat Biotechnol.* 2014; **32**(5): 462–464.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 17. Trapnell C, Williams BA, Pertea G, *et al.*: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol.* 2010; **28**(5): 511–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 18. Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
[PubMed Abstract](#) | [Publisher Full Text](#)
 19. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics.* 2011; **12**: 323.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 20. Robert C, Watson M: **Errors in RNA-Seq quantification affect genes of relevance to human disease.** *Genome Biol.* 2015; **16**(1): 177.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 21. Soneson C, Love MI, Robinson MD: **Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; peer review: 2 approved].** *F1000Res.* 2015; **4**: 1521.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 22. Pimentel H, Bray NL, Puente S, *et al.*: **Differential analysis of RNA-seq incorporating quantification uncertainty.** *Nat Methods.* 2017; **14**(7): 687–690.
[PubMed Abstract](#) | [Publisher Full Text](#)
 23. Nowicka M, Robinson MD: **DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics [version 2; peer review: 2 approved].** *F1000Res.* 2016; **5**: 1356.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 24. Teng M, Love MI, Davis CA, *et al.*: **A benchmark for RNA-seq quantification pipelines.** *Genome Biol.* 2016; **17**: 74.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 25. Love MI, Hogenesch JB, Irizarry RA: **Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation.** *Nat Biotechnol.* 2016; **34**(12): 1287–1291.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 26. Rossell D, Stephan-Otto Attolini C, Kroiss M, *et al.*: **QUANTIFYING ALTERNATIVE SPLICING FROM PAIRED-END RNA-SEQUENCING DATA.** *Ann Appl Stat.* 2014; **8**(1): 309–330.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 27. Bernard E, Jacob L, Mairal J, *et al.*: **Efficient RNA isoform identification and quantification from RNA-Seq data with network flows.** *Bioinformatics.* 2014; **30**(17): 2447–55.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 28. Hayward BE, Bonthon DT: **Structure and alternative splicing of the ketohexokinase gene.** *Eur J Biochem.* 1998; **257**(1): 85–91.
[PubMed Abstract](#) | [Publisher Full Text](#)
 29. Asipu A, Hayward BE, O'Reilly J, *et al.*: **Properties of normal and mutant recombinant human ketohexokinases and implications for the pathogenesis of essential fructosuria.** *Diabetes.* 2003; **52**(9): 2426–32.
[PubMed Abstract](#) | [Publisher Full Text](#)
 30. Ishimoto T, Lanaspas MA, Le MT, *et al.*: **Opposing effects of fructokinase C and A isoforms on fructose-induced metabolic syndrome in mice.** *Proc Natl Acad Sci U S A.* 2012; **109**(11): 4320–25.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 31. Li X, Qian X, Peng LX, *et al.*: **A splicing switch from ketohexokinase-C to ketohexokinase-A drives hepatocellular carcinoma formation.** *Nat Cell Biol.* 2016; **18**(5): 561–71.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 32. Mirtschink P, Krishnan J, Grimm F, *et al.*: **HIF-driven SF3B1 induces KHK-C to enforce fructolysis and heart disease.** *Nature.* 2015; **522**(7557): 444–449.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 33. Li B, Qing T, Zhu J, *et al.*: **A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq.** *Sci Rep.* 2017; **7**(1): 4200.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 34. Söllner JF, Leparo G, Hildebrandt T, *et al.*: **An RNA-Seq atlas of gene expression in mouse and rat normal tissues.** *Sci Data.* Nature Publishing Group. 2017; **4**: 170185.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 35. Diggle CP, Shires M, Leitch D, *et al.*: **Ketohexokinase: expression and localization of the principal fructose-metabolizing enzyme.** *J Histochem Cytochem.* 2009; **57**(8): 763–74.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 36. Jang C, Hui S, Lu W, *et al.*: **The Small Intestine Converts Dietary Fructose into Glucose and Organic Acids.** *Cell Metab.* 2018; **27**(2): 351–361.e3.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 37. Reyes A, Huber W: **Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues.** *Nucleic Acids Res.* 2018; **46**(2): 582–592.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 38. Soneson C, Love MI, Patro R, *et al.*: **A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs.** *Life Sci Alliance.* 2019; **2**(1): pii: e201800175.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 39. Pertea M, Shumate A, Pertea G, *et al.*: **CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise.** *Genome Biol.* 2018; **19**(1): 208.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 40. Pertea M, Pertea GM, Antonescu CM, *et al.*: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol.* 2015; **33**(3): 290–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 41. Hansen HH, Feigh M, Veidal SS, *et al.*: **Mouse models of nonalcoholic steatohepatitis in preclinical drug development.** *Drug Discov Today.* 2017; **22**(11): 1707–1718.
[PubMed Abstract](#) | [Publisher Full Text](#)
 42. Leinonen R, Sugawara H, Shumway M, *et al.*: **The sequence read archive.** *Nucleic Acids Res.* 2011; **39**(Database issue): D19–D21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 43. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 44. Hartley SW, Mullikin JC: **QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments.** *BMC Bioinformatics.* 2015; **16**(1): 224.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 45. Köster J, Rahmann S: **Snakemake—a scalable bioinformatics workflow engine.** *Bioinformatics.* 2012; **28**(19): 2520–2522.
[PubMed Abstract](#) | [Publisher Full Text](#)
 46. Love MI, Anders S, Kim V, *et al.*: **RNA-Seq workflow: gene-level exploratory analysis and differential expression [version 1; peer review: 2 approved].** *F1000Res.* 2015; **4**: 1070.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 47. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Royal Statistical Society.* 2018; **1**–13.
 48. Hahne F, Ivaneck R: **Visualizing Genomic Data Using Gviz and Bioconductor.** *Methods Mol Biol. Statistical Genomics.* (Springer New York). 2016; **1418**: 335–351.
[PubMed Abstract](#) | [Publisher Full Text](#)
 49. chbtchris: **chbtchris/Khk_quantifications: Second release (Version v1.1).** Zenodo. 2019.
<http://www.doi.org/10.5281/zenodo.2583233>
 50. Chabbert C: **Correction of Gene Model Annotations Improves Isoform Abundance Estimates: The Example of Ketohexokinase (KHK).** *OSF.* 2018.
<http://www.doi.org/10.17605/OSF.IO/NMKFA>

Open Peer Review

Current Referee Status:   

Version 2

Referee Report 26 April 2019

<https://doi.org/10.5256/f1000research.20430.r46714>

 **Luisa Mayumi Arake de Tacca**¹, **Stephen N Floor** ²

¹ Comparative Biochemistry Program, University of California, Berkeley, Berkeley, CA, USA

² Department of Cell and Tissue Biology (CTB), University of California, San Francisco (UCSF), San Francisco, CA, USA

The authors have addressed review comments - thanks!

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Gene expression, computational biology, RNA biology.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 12 April 2019

<https://doi.org/10.5256/f1000research.20430.r46712>

 **Magnus Rattray** 

Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

The authors have addressed all my comments in my original review and I'm happy with this version.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian inference, probabilistic models, genomic data analysis

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 11 April 2019

<https://doi.org/10.5256/f1000research.20430.r46713>



Patrick K. Kimes  1,2

¹ Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

² Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA, USA

I appreciate the additional work done by the authors to address my comments from the earlier version of the manuscript. The new analyses and clarifications made to the text have fully addressed my previous concerns.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Statistical Genomics, Statistical Learning

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 18 January 2019

<https://doi.org/10.5256/f1000research.18677.r42140>



Magnus Rattray 

Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

Chabbert et al. provide an interesting study of the effect of gene model annotation on isoform abundance estimation from RNA-Seq data. They take a gene with 5 annotated alternative transcripts and look at expression over several tissues. They explore how well a transcript-based estimation method (Salmon) and differential transcript usage method (DRIMSeq) does when considering additional evidence from visualisation of read coverage and experimental validation of alternative transcripts by qPCR. They demonstrate that the inferred transcript abundance using the standard model is inconsistent with read coverage (e.g. an inferred highly expressed transcript has very low read density or junction read evidence for its only unique exon) and they show that better results are obtained when the model is modified by hand in a number of ways. They also show that through using their handcrafted gene model, consistency is improved when the dataset is changed to single-ended and shorter reads, which I found to be an interesting experiment. They also look at concordance across different RNA-Seq datasets measuring the same tissues and find the modification of UTRs in the gene model to improve concordance.

Overall I found this to be an interesting and useful contribution. Recent work (cited by the authors) has looked at metrics to identify discordance of the type observed here but nevertheless this careful study of one gene model is a useful contribution, especially give the experimental validation using qPCR. I think the data will be of great interest to methodology developers grappling with this problem although I guess long-read datasets may provide evidence for more genes in future studies than this kind of qPCR approach.

Major comments:

1. As far as I can understand, in the end two kinds of change were made to the annotated gene model at the same time. A transcript Khk.RI was removed (this was tried first) and then both the UTRs were modified to be the same for all isoforms (first 3' and then 5') which means that new

isoforms are added to the gene model and some were removed. The removal of a transcript strongly affecting results is arguably more problematic for transcript quantification methods because we would often expect the set of isoforms to contain unexpressed isoforms in a specific dataset. The other issue (an incomplete gene model) is already known to be problematic ¹. However, I was not sure whether the UTR issue could potentially make the Khk.RI isoform issue worse. Therefore, I think it would be helpful to check whether Khk.RI inclusion remains problematic after making the UTR changes to the other isoforms. If adding an unexpressed isoform to the gene model can cause such an issue then this is a problem even when the gene model is correct. Also, there is a statement later on that “the removal of Khk.RI did further hinder reproducibility between both datasets” which seems to contradict the idea that removing this isoform is a good idea. I think this sentence is also a bit unclear and could do with some more explanation.

2. The modifications to the gene model in the end involves creating a new set of isoforms by combining existing exons and UTRs in a new configuration. Algorithms exist to do this in a data-driven way, e.g. the FlipFlop algorithm is designed to learn a set of isoforms from data by searching over all exon combinations using sparse inference on a splicing graph. Another example is StringTie and there may be later ones I'm not aware of. I think it would be useful to attempt to apply an algorithm of this type to this dataset to see how it would perform as this may provide a computational way to do something similar to what has been done manually in this example. Alternatively, these methods may fail and that would also be of interest.

Minor comments:

1. Page 3 - “the recent development of alignment-free...has provided the possibility to quantify each individual transcript”. Methods that predate Salmon were already available for transcript quantification e.g. RSEM, eXpress, bitseq etc. I don't think we should confuse the speed-up introduced in Salmon and Kallisto with the transcript-level inference model introduced by RSEM and I think RSEM could have been used in this paper with similar conclusions.
2. It would be good to know whether gene expression estimates based on the different gene models are performing differently, i.e. estimating gene expression by summing up isoform expression rather than counting. This has been discussed in the literature ² and this issue could be explored in this dataset.

Typos:

Page 7 - “on estimate concordance” should be estimated.

References

1. Sonesson C, Matthes KL, Nowicka M, Law CW, Robinson MD: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol.* 2016; **17**: 12 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Sonesson C, Love MI, Robinson MD: Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* 2015; **4**: 1521 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian inference, probabilistic models, genomic data analysis

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Mar 2019

Christophe Chabbert, ETH, Switzerland

We thank all the reviewers for their insightful comments and suggestions which have helped improve the quality of the manuscript. The main changes are summarized in the "Amendments from Version 1" section describing the new version of the manuscript. We will address the specific comments in this section.

Major Point 1

We have subjected the gene model annotation to two rounds of modifications to assess the impact

of *Khk.RI* and UTRs on relative *Khk* isoform expression estimates. To assess the importance of the order in which these modifications happen, we have generated an additional annotation where *Khk.RI* is retained and all other *Khk* annotated transcripts have identical 5' and 3' UTRs. This annotation was used to estimate *Khk* isoform proportions in the Li *et al* (using the paired end and 50bp single-end data) and the Sollner *et al* datasets. We observed that implementing this new UTR model was not sufficient to remove the *Khk.RI* estimation bias in the paired-end data (updated Figure 3A). This bias was still particularly important in the bone marrow and spleen datasets where *Khk.RI* was estimated to account for more than 20% of the overall gene expression.

Comparison of the estimates between the paired-end dataset and the 50bp dataset derived from Li *et al* showed that this modification was however more beneficial than the removal of *Khk.RI* (Pearson correlation of 0.86 as opposed to 0.79 when *Khk.RI* is removed). Nevertheless, the best agreement between both libraries was still obtained after removal of *Khk.RI* and extension of the UTRs (Pearson correlation of 0.97).

Finally, comparison of the estimates of the Li *et al* and Sollner *et al* datasets revealed that the extension of the UTRs (Pearson correlation of 0.73) was more beneficial than *Khk.RI* removal (Pearson correlation of 0.46) but provided hardly any improvement over the naïve GENCODE annotations (Pearson correlation of 0.73). The complete set of modifications including *Khk.RI* removal and homogenization of the UTR was required to reach higher concordance between both datasets (Pearson correlation 0.91).

Overall, these observations suggest that both modifications are indeed needed to improve *Khk* transcript quantification results in the datasets considered in this study.

We have updated main Figure 3, main Figure 4 and the underlying data table 3 to report these findings. The result and discussion section of the article have also been modified accordingly:

“Additionally, we computed the relative Khk isoform usage using a new annotation with identical 5' and 3' end for all isoforms except Khk.RI which was retained as such in the gene model (Figure 3A). This modification was not sufficient to remove Khk.RI estimation biases, with the retained intron predicted to erroneously account for 20% of the overall gene expression in bone marrow or spleen. We therefore confirmed the impact and importance of both Khk.RI and UTRs annotations on isoform expression estimates for that gene.”

“Both the removal of Khk.RI and UTR adjustments were necessary to reach this concordance between profiling methods.”

Finally, we have modified the mentioned statement to emphasise the importance of performing both modifications on the concordance between estimates:

“While the removal of Khk.RI without an adjustment of the UTR did further hinder reproducibility between both datasets, we observed a much stronger consistency of proportion estimates when using the manually updated annotation”

Major Point 2

Following this recommendation, we used StringTie as a complementary method to estimate transcript abundance across all tissues, with and without the reference annotation from GENCODE. Estimations output using the reference annotation still showed that the *Khk.RI* isoform still accounts for at least 15% of expressed isoform in 9 tissues. The results are presented in the Extended data, Figure 10A and the quantification results stored in the underlying data table 4. We also obtained StringTie quantifications without using a supporting genomic annotation, thereby

allowing for the assembly of transcript by relying only on the current dataset. After merging the discovered transcript, we examined the resulting annotation for the *Khk* gene and compared it with the current GENCODE gene model (Extended data, Figure 10B). Quite interestingly, no corresponding *Khk.RI* isoform could be detected but *Khk.A*, *Khk.C* and *Khk.Skip* were all identified, albeit with UTR regions differing from the GENCODE annotation. No *KhkA.C* isoform could be detected while junction counts and coverage tracks indicate that it is expressed in low levels in certain tissues (liver or small intestine for example).

Taken together, these results suggest that, when relying on current annotations, StringTie is subject to similar biases as Salmon (in accordance with the observations made in Soneson et al) when it comes to the erroneous detection of the *Khk.RI* unexpressed transcript. When relying on transcript assembly, StringTie does provide different UTRs from the GENCODE annotation but fails to detect the *KhkAxC* isoform. As it is beyond the scope of this study to deliver a computational solution to what was achieved manually and at the single gene level, we have reported our observations but will not evaluate the feasibility of the scalability of an improved StringTie approach to tackle this problem.

We have modified the discussion section of the article as follows:

“To complement our main findings, we estimated transcripts abundance in the Li et al dataset using StringTie, as an example of tool relying on the construction of a splicing graph to quantify isoforms. When using the GENCODE annotation as a guide for quantification, we found that StringTie overestimated the Khk.RI expression in a fashion similar to Salmon. This is in concordance with the report from Soneson et al³⁹. In addition, we used StringTie without any supporting annotation to enable transcript assembly from the alignments. Inspection of the resulting newly assembled transcripts showed that Khk.RI could not be detected (Extended Data, Figure 10). Nevertheless, KhkA.C was also not identified in the dataset while junction counts clearly indicate that it could be detected in a handful of tissues, albeit with low expression levels. We therefore suggest that the issue reported here in the case of Salmon might be commonly found across quantification software, and it will be interesting to assess whether similar biases may arise with more tools. Results from Soneson et al. strongly indicate that this is the case, at least for a group of human genes³⁹.”

The method section has also been updated to document the parameters and settings used for the StringTie quantifications:

“In order to quantify transcript using stringTie, alignments to the reference genome were first re-generated as described in the “RNA-seq data processing and alignment” section with the additional STAR flag “--outSAMstrandField intronMotif”. Quantifications were then performed with and without the GENCODE annotation used as a guide (-G option in stringTie). All other parameters were set to default values. Newly assembled transcripts were then merged using the transcript merge mode (--merge) with default parameters.”

Minor Point 1

This is true and we have updated this section of the introduction to make sure this distinction is made clear. The section is now written as follows:

“In addition, the recent development of alignment-free transcript quantification methods has provided the possibility to efficiently and rapidly quantify each individual transcript^{15–19}. Such approaches are indeed computationally much less demanding and faster than alignment-based methods¹⁵. Moreover, they have been shown to overcome the difficulty of handling multi-mapped

reads, which can create biased results in count-based analysis ²¹”

Minor Point 2

This is a good suggestion and following this recommendation and the major point 1 raised by Patrick Kimes, we have used *tximport* to obtain gene level estimates based on Salmon transcript quantifications from 3 of the modified annotations and the original GENCODE annotation. The results are now presented in the Extended Data, Figure 7. We found the gene level estimates to be consistent across all annotations and tissues and the patterns of expression across tissues to be conserved. In particular, *Khk* was found to be mostly expressed in liver, kidney and small intestine and that its expression was strongly tissue specific (adjusted p-value of 0 using the LRM from DESeq2 as described in the first version of the article). We have consequently added the following paragraph in the result section:

“Finally, we evaluated whether the changes brought to the Khk transcript annotations affected the estimations of overall Khk expression levels. Gene level estimates obtained using quantifications based on the 3 modified annotations were strongly correlated (Pearson, $r > 0.99$) with estimates derived from the naïve GENCODE annotation (Extended data: Figure 7A). In addition, the high tissue specificity of Khk expression was observed in all cases, with expression patterns identical between annotations (Figure 1A, Extended data: Figure 7B).”

Competing Interests: CDC is a full time employee of Roche

Referee Report 16 January 2019

<https://doi.org/10.5256/f1000research.18677.r42139>



Stephen N Floor ¹, **Luisa Mayumi Arake de Tacca**²

¹ Department of Cell and Tissue Biology (CTB), University of California, San Francisco (UCSF), San Francisco, CA, USA

² Comparative Biochemistry Program, University of California, Berkeley, Berkeley, CA, USA

Chabbert and colleagues did intricate work to measure how erroneous gene annotations without careful curation can lead to errors in gene expression analysis. This is a recurring problem in analysis of gene expression that has recently started to be addressed. In this work, they used a mouse gene (ketoheokinase) as a model to analyze publicly available data employing current alignment-free approaches to transcript quantification. The authors bring a real problem to the table and presents a good alternative as to how to proceed before analysis. Thanks for the deep study on how isoform annotations can affect gene expression quantification.

Comments –

- In Figure 2, is it possible to show the same coverage tracks and assayed tissues (2B vs 2C)? What does the R1 isoform look like in the spleen? Do the authors have any theories about why the software is misusing the retained intron information during quantification?

- Why might correcting 3' UTRs fix the problem? Would manually curating the 3' UTR for each isoform make a difference? Same question for the 5' UTR. A lot of times, the difference in isoforms is in transcription start sites, which have been observed to relate to downstream transcript processing events including splicing and poly-A site selection.
- Do the authors have any theories about how to generalize this approach for studies about many genes?
- How do the annotations for Khk change between databases? Would a unified annotation set using data from beyond Ensembl improve isoform quantification?

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: gene expression, computational biology, RNA biology

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 18 Mar 2019

Christophe Chabbert, ETH, Switzerland

We thank all the reviewers for their insightful comments and suggestions which have helped improve the quality of the manuscript. The main changes are summarized in the "Amendments from Version 1" section describing the new version of the manuscript. We will address the specific comments in this section.

Point 1:

The coverage tracks for all available tissues are shown either in Figure 2 or in the Extended Data, Figure 3 (accessible via this [link](#) - please make sure to expand the explorer tree on the left to make all files visible).

In this study we focused on identifying an example of gene model configuration that might be problematic. Although identifying which aspects of the quantification algorithms are causing this issue is definitely of interest, this question is beyond the scope of this study. Nevertheless, we invite the reviewers to consult the paper from Sonesson *et al* as it provides additional insights into this aspect and explores other parameters, such as the choice of quantification method. We are hoping that the bioinformatics community will be able to identify the root cause of these biases.

Point 2:

This study focuses on providing an example of quantification bias and on highlighting some aspects of a gene model that have a strong impact on this bias. Identifying the algorithmic cause of the problem would be very interesting but is unfortunately beyond the scope of the work reported here. In the example of *Khk*, we have shown in this paper that the use of homogenous 3' UTRs and 5' UTRs across transcripts has helped overcome the quantification issue. Additional work will be required to determine whether that would be the case for all problematic genes and but this is also not part of the current study.

As pointed out, some reports have indeed shown that changes in 5' or 3' end are an important source of isoform diversity, which might complicate the interpretation of quantification results. As we mentioned in the abstract and in the discussion, it will be important to always thoroughly confirm such findings derived from transcript quantification studies. In addition, it is important to note that standard RNA-seq protocols are usually not suited to identify such changes in 5' or 3' end usage. Indeed, it is recommended to use protocols such as CAGE or 3'T-fill capture that can unambiguously identify transcription start or end. Therefore, such cases should be handled with an adequate experimental setting rather than standard RNA-seq strategies.

Point 3:

As previously commented in Points 1 and 2, our study did not aim at providing a generalized method that identifies problematic genes such as *Khk* or to provide a universal solution to the issue. The study from Sonesson *et al* introduces the JCC index, which should be very helpful to identify problematic genes. In addition, in order to reduce the risk of making erroneous conclusions from large data analyses, we recommend confirming important findings using orthogonal methods

(as stated in the abstract) such as count based approaches or well-targeted low throughput experiments such as qPCR for example.

Point 4:

In mouse, *Khk* annotations differ between Ensembl and RefSeq (please see the RefSeq annotation [here](#), curated transcript NM_). When focusing on the curated transcripts, the most striking differences with the GENCODE annotation are the lack of *Khk.RI*, *Khk.Skip* and identical 5' and 3' UTRs. This would therefore be an equivalent situation to the one presented in this paper but without the *Khk.Skip* transcript available for quantification. This would therefore result in erroneous predictions as well. To comment on this point, we have added the following statement in the discussion section of the paper:

"It is also important to note that the curated RefSeq Khk gene model differs from the GENCODE as it is missing Khk.RI, Khk.Skip and the 3' and 5' UTRs of all curated transcripts are identical. While this would be a close configuration to the optimal annotation presented in this study, the absence of Khk.Skip in the gene model would result in erroneous quantifications as well."

While a consolidated annotation based on transcript model derived from large datasets might be an appealing idea, recent results from Sonesson *et al* showed that in practice, this did not improve transcript estimate accuracy for problematic genes. We updated the discussion section of the paper in order to elaborate on this matter:

*"Improvement of current genomic annotations might ultimately offer an alternative as they will allow for the sole use of fast quantification algorithms. This might partially be achieved using transcript catalogues obtained from large scale studies such as CHESS even though Sonesson *et al* reported very little to no improvement in their JCC scores using these new annotations."*

Competing Interests: CDC is a full time employee of Roche

Referee Report 08 January 2019

<https://doi.org/10.5256/f1000research.18677.r42141>



Patrick K. Kimes  1,2

¹ Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA

² Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA, USA

Using the example of murine ketohexokinase (*Khk*), the authors present an analysis of the limitations of a current alignment-free approach (Salmon) to isoform-level quantification using RNA-seq data. In particular, using publicly available data sets, primarily from the Mouse BodyMap project ¹, the authors show that transcript quantification using Salmon is highly sensitive to the specified reference transcriptome. These results are corroborated using qRT-PCR measurements of relative isoform abundance across several mouse tissues included in the study.

The authors apply two approaches to modify the commonly used GENCODE annotations for *Khk* to obtain improved estimates of isoform abundance using Salmon. First, the authors consider removing an annotated transcript, *Khk.RI*, with low support in unique exonic regions. Second, the authors modify the

original annotations by trimming all 5' and 3' differences between isoforms. The authors show that with these two procedures, the resulting isoform estimates are more consistent across library protocols and across experiments - comparing against data from a separate study of mouse tissues². The advantage of removing weakly supported isoforms is in line with previous work which showed the benefits of prefiltering isoforms for false discovery rate control in studies of differential transcript usage³.

The conclusion is the need for "manual curation" of transcript annotations to improve quantification by current alignment-free approaches, and no general approaches are provided for improving all annotations. This is fine, as providing general guidelines (aside from careful review of isoform annotations) does not appear to be the purpose of this article.

Overall, the paper provides an interesting and detailed analysis of a single gene that complements the growing literature on challenges in isoform-level quantification with RNA-seq. However, a few issues should be addressed before the paper and the analysis can be considered complete. These issues and a few other minor points are described below.

Major Issues

1. It appears that all gene-level analyses were carried out using gene count tables obtained using the QoRTs software, presumably based on the original ("naive") GENCODE transcript annotations. However, as one of the references cited by the authors points out⁴, isoform-level quantification provides improved estimates of gene expression over simple count-based approaches. I am curious to see how the Salmon-based estimates (using tximport) of gene-level expression 1) compare with the QoRTs count tables, and 2) change with modifications to the transcript annotations. While the modified annotations appear to improve isoform-level estimates, it is also important to verify that they do not reduce the accuracy of gene-level estimates. This would further confirm the "reliability of gene expression", as described by the authors.
2. In the analysis of Salmon-based isoform quantification results, the authors claim that "inspection of the junction reads and coverage tracks revealed discrepancies between quantification estimates and expected results for several tissues." How were the expected results determined? From the text, this appears to be based on visual inspection of the coverage plots. However, per-base coverage can be tricky to visually interpret due to biases in RNA-seq data, e.g. sequence specific bias and fragment-level GC bias⁵. Ideally, the definition of "expected results" should be made more concrete, e.g. using a metric such as the expected bias-corrected junction coverage as in Sonesson et al., (2018)⁶ or quantification of exon coverage as in Figure 1C. Regardless, the wording should be updated to more clearly state what constitutes "expected results" and how they were determined.
3. While the authors claim that "differences in 5' and 3' end annotations between all isoforms ... were not reflected on our coverage tracks," this does not appear to be true. In fact, there appears to be a clear difference in TSS coverage between tissues (Figure 3B). Notably, differences in 5' coverage appear to correspond to differences in coverage of the Khk.Skip splice junction (in agreement with GENCODE annotations). Additionally, clear TSS coverage differences were also observed and noted in the exon-level count analysis (Figure 1C), further supporting the presence and importance of TSS differences. The problem appears to be a mis-annotation of the 5' start site for the KhkC isoform (see Figure 1C where liver and kidney, two tissues with high KhkC preference according to RT-qPCR analysis, show significantly lower coverage in alternate start sites E01-E03). In light of these observations, the decision to completely trim all 5' and 3' differences seems rather extreme, and referring to the trimmed annotations as a "correction" of the gene models may not be accurate. Especially since such a trimming would result in the complete loss of the differential 5' behavior discovered earlier in the manuscript. This is particularly important as others have shown that start

and termination site differences are the primary sources of isoform differences across human tissues ⁷. More justification and careful discussion of the limitations of this general trimming procedure are needed. Rather than a "correction", this appears to be a particularly aggressive modification that works in this setting because all isoforms differ by more than just the trimmed 5' and 3' regions (and not because "differences were not reflected on our coverage tracks", as stated in the text).

4. In modifying the GENCODE annotations, the authors use two steps. First, the removal of unexpressed transcript annotations, and second, the trimming of 5' and 3' differences. Figure 4D-F shows that simply removing the Khk.RI isoform from the annotations actually reduces the agreement between the two data sets, while additionally trimming 5' and 3' ends appears to greatly improve agreement. Have the authors considered how estimates change by applying the 5' and 3' trimming procedure without removing the Khk.RI transcript from the annotations? From the current analysis, it is unclear how much removing Khk.RI is improving the final result (with trimming), because these changes are not additive.
5. While details are included for the Khk A/C-/- negative control mice, details on the non-control mouse tissue samples used for RT-PCR and qRT-PCR analyses are missing. How were these samples obtained and matched with the publicly available Mouse BodyMap RNA-seq data? This should be described in the Methods section.

Minor Issues

1. The authors claim that alignment-free transcript quantification methods have "provided the possibility to quantify each individual transcript", and that "this task is often impossible to complete using traditional count approaches." However, several methods for transcript quantification predate alignment-free methods, including some referenced later in the same section (Cufflinks, casper, FlipFlop), among others (RSEM, eXpress). It is also unclear what is meant by "often impossible." The wording in this section should be clarified.
2. It is unclear what "one common approach" is referencing at the beginning of the second paragraph describing differential gene expression (DGE) analysis. This is particularly jarring as the previous paragraph ends with noting the challenges of studying "complex events such as splicing or isoform usage switch". DGE analysis does not address these questions.
3. The authors claim that "DESeq2-tximport and sleuth [incorporate] estimates of inferential variances obtained during the quantification". While this is true of sleuth, I do not believe this is true of DESeq2-tximport. If it is, an appropriate reference should be added (none of the currently cited references describe this feature). Additionally, the reference linked to DESeq2-tximport (reference 22 in the article) ⁸ describes benchmarking several DGE methods to scRNA-seq data, and is a primary source for neither DESeq2 nor tximport. The more relevant reference would seem to be (reference 20 in the article) ⁴ which describes the tximport software package.
4. It is unclear what is being referred to by "these configurations", when claiming "no systematic evaluation of the impact of these configurations has been conducted."
5. The figure referenced in ".. hardly any detectable expression of the transcript (Figure 2B ...)", should be "Figure 2C". It would also be helpful if the exon ID from the figure (E012) was also included in the text for easier reference.
6. Description and axes of Figure 3C,D should be corrected to reflect the fact that KhkA and KhkC estimates are actually (KhkA + KhkA.C) and (KhkC + KhkA.C) estimates, as described in the text.
7. In the Results section, "junctionSeq" should be stylized "JunctionSeq" for consistency.
8. Throughout, "Genecode" is probably meant to be "GENCODE" (or "Gencode").

References

1. Li B, Qing T, Zhu J, Wen Z, Yu Y, Fukumura R, Zheng Y, Gondo Y, Shi L: A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. *Scientific Reports*. 2017; **7** (1). [Publisher Full Text](#)
2. Söllner J, Leparc G, Hildebrandt T, Klein H, Thomas L, Stupka E, Simon E: An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Scientific Data*. 2017; **4**. [Publisher Full Text](#)
3. Sonesson C, Matthes KL, Nowicka M, Law CW, Robinson MD: Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biol*. 2016; **17**: 12 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Sonesson C, Love MI, Robinson MD: Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*. 2015; **4**: 1521 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Love MI, Hogenesch JB, Irizarry RA: Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat Biotechnol*. 2016; **34** (12): 1287-1291 [PubMed Abstract](#) | [Publisher Full Text](#)
6. Sonesson C, Love M, Patro R, Hussain S, Malhotra D, Robinson M: A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. *bioRxiv*. 2018. [Publisher Full Text](#)
7. Reyes A, Huber W: Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res*. 2018; **46** (2): 582-592 [PubMed Abstract](#) | [Publisher Full Text](#)
8. Sonesson C, Robinson MD: Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods*. 2018; **15** (4): 255-261 [PubMed Abstract](#) | [Publisher Full Text](#)
9. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011; **12**: 323 [PubMed Abstract](#) | [Publisher Full Text](#)
10. Roberts A, Pachter L: Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat Methods*. 2013; **10** (1): 71-3 [PubMed Abstract](#) | [Publisher Full Text](#)
11. Glaus P, Honkela A, Rattray M: Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*. 2012; **28** (13): 1721-8 [PubMed Abstract](#) | [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Statistical Genomics, Statistical Learning

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 Mar 2019

Christophe Chabbert, ETH, Switzerland

We thank all the reviewers for their insightful comments and suggestions which have helped improve the quality of the manuscript. The main changes are summarized in the “Amendments from Version 1” section describing the new version of the manuscript. We will address the specific comments in this section.

Major Point 1

Thank you for making this suggestion. In order to address this point, we have used tximport to compute *Khk* expression levels from salmon estimates generated using the naïve and all modified annotations (including the new one added to this version, see major comment 4). A comparison between the naïve annotation estimates and QoRTs counts is provided in the extended data, Figure 1B. A comparison between estimates from all tested annotations is now presented in the extended data, Figure 7 A and B. Examination of the tissue expression patterns of *Khk* (Panel A) and the correlation between gene level estimates clearly show a very good agreement between all annotations, thereby suggesting that the modifications have little impact on these estimates. We also verified that the very high tissue specificity of *Khk* in every case (adjusted p-value of 0 with DESeq2, following the similar analytical framework as described in the first version of the manuscript).

In order to clarify this point, we have added the following statement in the result section:

“These results were all confirmed using the gene level estimates based on Salmon quantifications (Extended Data Figure 1A and B). In particular, patterns of Khk expression across tissues were highly concordant between count-based and Salmon estimates.”

“Finally, we evaluated whether the changes brought to the Khk transcript annotations affected the estimations of overall Khk expression levels. Gene level estimates obtained using quantifications based on the 3 modified annotations were strongly correlated (Pearson, $r > 0.99$) with estimates derived from the naïve GENCODE annotation (Extended data: Figure 7A). In addition, the high tissue specificity of Khk expression was observed in all cases, with identical expression patterns between annotations (Figure 1A, Extended data: Figure 7B).”

Major Point 2

The expected results were determined using an inspection of the coverage tracks, the normalized exon coverage results obtained QoRTs and DEXSeq and previous reports from single gene studies. We have modified the wording in this section:

“Despite this adjustment, inspection of the junction reads, coverage tracks, and normalised exon and junction counts derived from QoRTs (Figure 1C) revealed discrepancies between quantification estimates and results derived from alignment-based methods (Figure 3A and 3B). These quantification estimates also did not reflect the observations made by previous reports focusing on the characterisation of Khk expression patterns³⁶. An example of such discrepancy may be found examining the heart coverage tracks”

Major Point 3

Thank you for pointing to this lack of clarity in the manuscript. It is indeed not the objective of this manuscript to suggest that a trimming or a standard homogenisation of the UTRs of all annotated transcripts of a gene will improve quantification results. Such an update of the annotation is possible because all *Khk* isoforms can be detected unambiguously just by considering the exclusion or inclusion levels of exon 3A and 3C (as partially stated in the discussion section). We have now made this clearer in the result section as well and explicitly mentioned this idiosyncrasy of the *Khk* gene model. We have also expanded the discussion section to elaborate on the limitations of this method for cases where UTRs are the main source of variability between isoforms of the same gene.

Finally, we would like to emphasize that the term correction is never used in this result section. In the title and abstract of this work, it refers primarily to the removal of the *Khk.R1* which is indeed a correction as this transcript expression levels remain below the threshold of detectability in the highly sensitive RT-qPCR assays.

The newly modified result section now reads as follows:

“Since such quantifications are relying on the reference transcripts provided during the indexing step¹⁵, we reasoned that incorrect transcript models might be the cause of the observed discrepancies and therefore compared coverage tracks, normalized exon counts and isoform models. We identified annotated differences in 3' end annotations between all isoforms which were not reflected on our coverage tracks (Figure 1C and Figure 3B). As Khk isoforms can be identified unambiguously based on the exclusion patterns of the exons 3A and 3C and regardless of differences in UTRs, we could investigate the impact of these UTR variations on transcript quantifications. We therefore manually updated Khk isoform annotations to provide an identical 3' end to all isoforms (Underlying data: File 1 and 2) and re-estimated isoform proportions (Figure 3A, middle panel). Despite this adjustment, inspection of the proportion estimates still revealed

erroneous estimations of isoform expression in particular in the case of liver where KhkA was detected in levels similar to KhkC. Further examination of the results revealed that, while some differences in 5' end coverage in the dataset were concordant with the current gene annotation, they were not always reflected in the gene model (Figure 1C and Figure 3A, Heart, Lung and Spleen 5' UTR coverage). Following a similar approach to the one described earlier for 3' UTRs, we finally manually modified Khk isoforms to provide an identical 5' and 3' end to all listed isoforms (Underlying data: File 3)."

The updated discussion section reads as follows:

"However, systematic harmonization of all UTRs across annotated transcripts might not be a general approach, especially in cases when such differences are reflecting tissue specific expression patterns"

Major Point 4

To address this point, we have generated an additional annotation where *Khk.RI* is retained while all other *Khk* annotated transcripts have identical 5' and 3' UTRs. This annotation was used to estimate *Khk* isoform proportions in the *Li et al* (using the paired end and 50bp single-end data) and the Söllner *et al* datasets. We observed that implementing this new UTR model was not sufficient to remove the *Khk.RI* estimation bias in the paired-end data (updated Figure 3A). This bias was still particularly important in the bone marrow and spleen datasets where *Khk.RI* was estimated to account for more than 20% of the overall gene expression.

Comparison of the estimates between the paired-end dataset and the 50bp dataset derived from *Li et al* showed that this modification was however more beneficial than the removal of *Khk.RI* (Pearson correlation of 0.86 as opposed to 0.79 when *Khk.RI* is removed). Nevertheless, the best agreement between both libraries was still obtained after removal of *Khk.RI* and extension of the UTRs (Pearson correlation of 0.97).

Finally, comparison of the estimates of the *Li et al* and Söllner *et al* datasets revealed that extension of the UTRs (Pearson correlation of 0.73) was more beneficial than *Khk.RI* removal (Pearson correlation of 0.46) but provided hardly any improvement over the naïve GENCODE annotations (Pearson correlation of 0.73). The complete set of modifications including *Khk.RI* removal and homogenization of UTR was required to reach higher concordance between both datasets (Pearson correlation 0.91).

Overall, these observations suggest that both modifications are indeed needed to improve *Khk* transcript quantification results in the datasets considered in this study.

We have updated main Figure 3, main Figure 4 and the underlying data table 3 to report these findings. The result and discussion section of the article have also been modified accordingly:

"Additionally, we computed the relative Khk isoform usage using a new annotation with identical 5' and 3' end for all isoforms except Khk.RI which was retained as such in the gene model (Figure 3A). This modification was not sufficient to remove Khk.RI estimation biases, with the retained intron predicted to erroneously account for 20% of the overall gene expression in bone marrow or spleen. We therefore confirmed the impact and importance of both Khk.RI and UTRs annotations on isoform expression estimates for that gene."

"Both the removal of Khk.RI and UTR adjustments were necessary to reach this concordance between profiling methods."

Major Point 5

11 mice used in this study, C57BL/6J control and *KhkA/C*^{-/-} mice, were treated in a similar fashion and tissue were harvested on the same day. Additionally, RNA extraction, cDNA synthesis and gene expression analysis via RT-qPCR and semiquantitative RT-PCR were performed within the same experimental run. Differences in mRNA expression for *Khk* and *Khk* isoforms were evaluated via normalization to the internal reference gene *β-actin* for each sample. The wording in the methods section of the article has been modified accordingly:

“C57BL/6J were obtained from The Jackson Laboratory while KhkA/C^{-/-} mice, which are of C57BL/6 background and are lacking both ketohexokinase-A and ketohexokinase-C, were obtained from R. Johnson (University of Colorado) and used as negative control. All mice were housed in a pathogen-free facility at the ETH Phenomics Center (EPIC) under standard conditions (12 h light and 12 h dark cycle) with free access to food and water.”

“Total RNA from C57BL/6J and KhkA/C^{-/-} mice was prepared from frozen tissues with RNeasy Mini Kit (QIAGEN, Hilden, Germany) and treated with DNase I to remove traces of DNA.”

Minor Point 1

This sentence is indeed confusing and misleading and has therefore been removed from the introduction.

Minor Point 2

We have modified the start of the second paragraph to make the statement clearer:

“One common approach to analyse RNA-seq datasets consists in identifying significant changes in expression levels between two or more experimental conditions using gene-level counts”

Minor point 3

Indeed, DESeq2-tximport cannot incorporate estimates of inferential variances. We have therefore updated the statement and corrected the reference, which should be the publication describing the tximport package.

Minor point 4

We have corrected the sentence to hopefully clarify its meaning:

“To our knowledge, no systematic evaluation of the impact of the presence of low coverage on such key regions has been conducted and detailed reports of such examples in real datasets are still missing.”

Minor point 5

We have updated the figure reference in the main text and referenced the exon E012.

Minor point 6, 7, 8

Thank you for taking the time to highlight these discrepancies. We have corrected the wording in the Figure and the changed the “JunctionSeq” and “GENCODE” formatting in the main text.

Competing Interests: CDC is a full-time employee of Roche

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research