RNA sequencing: current and prospective uses in metabolic research

Petter Vikman, Joao Fadista and Nikolay Oskolkov

Diabetes and Endocrinology, Department of Clinical Sciences, Malmö University Hospital, CRC, Lund University, Building 60, Level 13, Entrance 72, S-205 02 Malmö, Skåne, Sweden

Correspondence should be addressed to P Vikman

Email
petter.vikman@med.lu.se

Abstract

Previous global RNA analysis was restricted to known transcripts in species with a defined transcriptome. Next generation sequencing has transformed transcriptomics by making it possible to analyse expressed genes with an exon level resolution from any tissue in any species without any a priori knowledge of which genes that are being expressed, splice patterns or their nucleotide sequence. In addition, RNA sequencing is a more sensitive technique compared with microarrays with a larger dynamic range, and it also allows for investigation of imprinting and allele-specific expression. This can be done for a cost that is able to compete with that of a microarray, making RNA sequencing a technique available to most researchers. Therefore RNA sequencing has recently become the state of the art with regards to large-scale RNA investigations and has to a large extent replaced microarrays. The only drawback is the large data amounts produced, which together with the complexity of the data can make a researcher spend far more time on analysis than performing the actual experiment.

Key Words

- ▶ gene expression
- ▶ gene transcription
- ▶ methodology
- ▶ microarray
- ▶ RNA
- ▶ sequencing

Journal of Molecular Endocrinology
(2014) 53, R93–R101

Introduction

The transcriptome is defined as all the RNAs in a cell at a given time and researchers have for a long time been interested in its composition. It has been reported that as many as 90% of eukaryotic genes can undergo alternative splicing (Pan *et al.* 2008) and that this is utilized to increase the complexity of the transcriptome (Modrek & Lee 2002, Pan *et al.* 2008, Wang *et al.* 2009). Further variation is in part done both in a tissue-specific manner (Le *et al.* 2004, Sugnet *et al.* 2006, Cooper *et al.* 2009) and by cellular signals (Ortis *et al.* 2010) and much of this is variation is not seen with the conventional techniques. In the early days, the amount of a given transcript was semi-quantitatively measured using hybridization techniques such as northern blotting analysis (Kevil *et al.* 1997) with

the subsequent development allowing for the use of macroarrays (Rast *et al.* 2000) and then microarrays (Schena *et al.* 1995) and finally a fully quantitative technique, real-time PCR also known as qPCR (Heid *et al.* 1996). All of these techniques were limited to the investigation of previously defined transcripts with a known sequence and the choice had to be made between a global or a quantitative analysis, producing little splice-specific information. As the work to define transcripts was both time consuming and cumbersome (Agulnik *et al.* 1998), the definition of new transcripts took time. These limitations meant that if you worked outside of the more common model organisms you would have to use the sequence from other species as a base for your work, and

when working with well-studied transcriptomes you would still have to choose the splice variant(s) of interest with the risk of other splice variants affecting your end result (Table 1).

RNA sequencing on the other hand allows for a hypothesis-free investigation of all the RNAs in a sample. It investigates the entire transcript in comparison with the other methods that focuses on a targeted part of the transcript and is able to discover novel exons, splice sites, and gene fusions (Mortazavi *et al.* 2008, Sultan *et al.* 2008, Asmann *et al.* 2009, Pickrell *et al.* 2010, Richard *et al.* 2010). It also gives the total expression level of a gene, as well as the ratio between different isoforms, splice variation (Fig. 1) and expression levels of exon resolution with a better correlation with the subsequent protein levels as compared with microarrays and a higher sensitivity (Degner *et al.* 2009, Wang *et al.* 2009).

Technical overview of RNA sequencing

There are currently three major suppliers of sequencing equipment: Illumina, Pacific Bioscience, and Life Technologies. All three technologies have their pros and cons and it is beyond the scope of this review to properly address them. An overview is presented in Table 2. Briefly Pacific Bioscience produces by far the longest reads, but carries a hefty price tag. Life Technologies have the lowest entry price, but have a higher running cost and a somewhat lower base quality. Illumina sports the highest throughput and lowest cost per base, although on their most expensive equipment with their less expensive equipment having a lower throughput and somewhat higher cost per base. For the last few years, Illumina's chemistry and technology has been by far the most commonly used sequencing equipment for RNA-sequencing studies.

A few fundamental decisions on design need to be made before running the experiment. As can be seen in the summary in Table 2, all instruments produce a finite amount of data. Due to this one of the first decisions that needs to be made is how to focus the sequencing power

toward the transcripts of interest. The two most common approaches being the investigation of either rRNA-depleted or polyA-enriched samples, this is because otherwise ~99% of the sequencing power will go toward the rRNAs. The quality of the sample can affect this decision with poly(A) selection, working on high-quality samples and rRNA depletion being the recommended approach for lower quality RNA such as from formalin-fixed paraffin-embedded samples. The RNAs are then fragmented (optional step) and made into cDNA that is subsequently sequenced. This can either be done as a paired end (PE), meaning that the fragment is sequenced from both ends or single read (SR), meaning that it is only sequenced from one side. This will then produce either one stretch of sequence, one read, of a given length or two bi-directional reads per fragment sequenced. This is done on several million fragments simultaneously creating several million reads per sample. These reads are subsequently compared and aligned to the genome and/or the transcriptome of the species investigated and the expression levels of both known and novel transcripts are defined (Fig. 2).

If your aims are to investigate both known and unknown transcripts and splice variants, a longer read length together with PE sequencing is recommended for use. Longer reads will give more information with regards to the actual splice events, i.e., data from exon/exon a boundary that not only indicate the expression levels of the exon but also show which exons are linked. There is also the possibility to perform a so-called stranded sequencing, meaning that the data produced are defined as being from either of the two chromosomal strands. This gives more stringent information and can help in discerning between reads from genomic regions which are shared by genes on different strands. In the case of genes sharing exons and genes on the same strand, such as insulin (INS) and insulin-like growth factor 2 (INS-IGF2), the strand specific data do not help, instead longer reads and paired-end sequencing increases the possibility to clearly define which gene a read comes from.

Table 1 A summary of the most common techniques to investigate RNA. Previously there were a tradeoff between doing a global or a quantitative analysis, RNA sequencing can however be used to combine a global and quantitative analysis

	Northern blotting	Arrays, macro and micro	Real-time PCR	Sequencing
A priori knowledge	Yes	Yes	Yes	No
Quantitative	Semi	Semi	Yes	Yes
Number of RNAs detected	1–5	Up to 30–40.000	1–4 at a time	All
Dynamic range	10 ³ –10 ⁶	10 ¹ –10 ⁴	1–10 ⁹ copies	1–infinity

Figure 1

Different splice variations that has been observed with RNA sequencing. This includes but is not limited to differential exon usage, inclusion and exclusion, partial differential exon usage both on the 5' and 3' ends as well as intron inclusion and mutually exclusive exons. This kind of variation can be seen, thanks to the exon level resolution from RNA sequencing data and it is especially clear with data from exon–exon spanning reads.

If you only are interested in the gene expression then the read length is of little value except for the alignment precision. Here, however, the number of reads are very important because a higher number of reads will give a statistically more solid investigation, especially of transcripts with a lower expression. Hence, in this study, the focus need not be on the total amount of data produced, 100 million clusters worth of data from a single-end run with 76 bp long reads will be more valuable than 40 million 100 bp long paired-end reads despite the fact that the data amount produced is basically the same.

Table 2 This is a summary of the sequencing options from three different companies on the market. The longest read length is by far that from Pacific Bioscience; however, their equipment is expensive and the through put is somewhat limited and expensive. Life Technologies have relatively inexpensive equipment; their major drawback is their throughput and running cost per sample. Illumina have a range of different equipment ranging in price and output. Here a tradeoff needs to be made between initial investment, through put, read length, and cost per base

	Illumina	Pacific Bioscience	Life Technologies
Run type	Single end or pair end	Single end	Single end or pair end
Read length	Up to 2*350	Up to 30–40 kb, 5–8 kb on average	Up to 400
Data per run	Up to 1 Tb	Up to 400 Mb	Up to 1 Gb
Reads per run	Up to 4 billion	Up to 50.000	Up to 80 million
Cost per 10 ⁶ bases	0.05–0.15 USD	0.33–1 USD	One USD
Throughput	High	Moderate	Moderate
Run time	1–10 days	30–120 min	2–8 h

Statistical analysis

Data analysis can broadly be divided into four main steps: alignment, count summarization, normalization, and identification of differentially expressed genes. The alignment process is carried out by comparing the data produced with the genome and/or the transcriptome to see where there is a high degree of sequence similarity. The output after alignment will be genomic positions where the sequence reads match the genome. If a read comes from more than one exon, i.e. exon–exon spanning, the genomic region where they map is specified and a junction where there is a presumed intronic region. Paired-end data will increase the mapping quality because the information from both reads is taken into account increasing the alignment specificity. There is the option to perform the analysis and create transcripts using only the sequencing data by aligning the reads together without the help of a scaffold. However, this is very time-consuming and tend to be artifact-prone, especially with short-read data such as that produced with Illumina and Life Technologies equipment. Unless the organism investigated completely lacks genomic and transcriptomic data, something that is becoming increasingly rare, the best option is to use previously produced data as a scaffold for the alignment. The two main aligners used are Tophat and STAR, and in our experience they produce similar results. STAR is much faster but requires more RAM to run (Engstrom *et al.* 2013). The recommended pipeline from BROAD (<http://gatkforums.broadinstitute.org/>) to enable variant calling in the data uses STAR as an aligner; however, the output from Tophat works for the said pipeline. This process has been compared with taking all papers in a news stand, shredding them and then trying to piece them back together (Korf 2013).

Figure 2

An overview of RNA sequencing as performed using Illumina Technology here depicted as a paired-end sequencing set-up. The two most common RNA sequencing approaches are to either sequence mRNA selected with an poly(A)n tail selection step or a negative selection step where the rRNAs are removed (A). rRNA depletion is common when working with RNA samples where the RNA quality is lower such as for RNA from formalin-fixed paraffin-embedded samples. The RNAs are then fragmented and turned to

cDNA and adaptors ligated to the ends (B, C, and D). Following this, the fragments are sequenced from both end creating two reads from the same fragment (E and F). The reads are then aligned to the genome and/or the transcriptome and the transcripts expression levels are measured (G and H). It is possible to extract information regarding the expression levels of individual exons as well as how the different exons are linked together from exon-exon junction spanning reads.

Following the alignment, the gene expression is calculated. The most straight forward way to do this is to count the number of reads that is aligned to a given gene. This is done by comparing the reads genomic coordinates with those of genes genomic, and the sum of the number of reads that map to a certain gene is summarized defining its expression level. Here, longer reads and PE data helps. The longer reads can be mapped more accurately and they contain more exon-exon junction information due to their length. The PE data help because the mapping information from both reads is used together to define the optimal alignment. By dividing the number of counts with the gene length and the total number of reads sequenced from the sample, it is possible to get a value that is comparable between genes as well as between samples and runs. This is

usually defined either as fragments per kilobase of exon per million fragments mapped (FPKM) or read per kilobase of exon per million fragments mapped (RPKM). So for a SR experiment, the FPKM value would be analogous to the RPKM while for a PE experiment the RPKM value would be two times the FPKM value. This is not always true because during the data production the reads can be produced without a pair and sometimes only one read from a pair maps. However, for most applications this simplification works. This calculation can also be done on an exon level. If the aim is to identify genes that differ between conditions the normalization for gene length is usually omitted as it introduces problems on its own (Cai *et al.* 2012).

Often the aim of an RNA experiment is to investigate whether the expression levels are different between tissues

Figure 3

The use of a sashimi plot can show differential exon usage (A). Here it shows the exclusion/inclusion of an exon in the blue and the green samples, an event that is not seen in the red and the brown samples. The numbers indicate the number of reads that spans from the 4th to the 6th exon in the

figure (A). In (B) it shows complete or partial usage of an exon, an event present in all four of the samples. This here is a partial usage of the second exon in the figure and it indicates that there are two 3' splice sites.

and/or treatment groups. Following the development of RNA sequencing, softwares have been developed to aid in this analysis. The group sizes and expression levels of the genes of interest will be pivotal in deciding which software works for a given experiment and there are reviews of the most commonly used softwares (Soneson & Delorenzi 2013). These softwares work for the analysis of both entire genes and transcripts or exons.

An alternative way to analyse the data is to create transcripts from each gene found expressed by combining the junction and exon level data. The most common way of doing this is through Cufflinks (Trapnell *et al.* 2012). This software package contains tools to create the transcripts and to subsequently compare the transcript

levels between groups. The creation of complete transcripts is hard; however, as there are linked events such as mutually exclusive exons which Cufflinks can have issues handling.

There are several software packages that can help in the visualization of RNA sequence data, IGV with the Sashimi plot add on being one. In Fig. 3, both the exon levels and the number of reads that link different exons together are visualized with both an exon skipping event and a 3' splice site being evident. This kind of data is where the strength of junction reads, reads spanning exon–exon junctions, comes in. These unequivocally show exon skipping and exon–exon linkage, something that is only indirectly seen in the levels of an exon.

Figure 4

RNA sequencing data can be used to call SNPs as much in the same way as with genomic sequencing data. It is however mainly possible to call exonic SNPs due to the nature of the source of the data. This data can be used to investigate whether there is an allelic imbalance in the gene expression, from imprinting or from an other source. In (A) there is clear evidence of

a C/G SNP with an even distribution, with 57 C vs 60 G carrying reads, between the alleles. In (B) however there is an uneven distribution, with 20 C vs 10 T carrying reads, between the two alleles. This uneven distribution indicates an allelic imbalance with one allele expressed with twice the level.

Imprinting and allele-specific expression

Imprinting and allele-specific expression are events when an allele is predominately expressed due to preference of, for example, DNA methylation or when a genomic variant acts in cis and affect a genes expression. Several articles have been published on this subject (Nag *et al.* 2013, Savova *et al.* 2013). The option to use RNA sequencing data to call SNPs (Fig. 4) in genes expressed makes it possible to investigate whether the expected 50/50 ratio is

significantly changed indicating imprinting. To enable the calling of SNPs, the authors behind GATK, arguably the most widely used software for variant calling, released guidelines for RNA sequencing data. However, calling variants from RNA-seq has to be executed carefully because a number of caveats exist. These guide lines helps in adjusting the genomic-based softwares to RNA sequence data before SNP calling. For imprinting and allele-specific expression studies genomic, especially exonic, variant data help. This is because if there is an

allele that is completely imprinted and not expressed at all it will not be possible to use RNA sequencing data alone to separate a homozygote variant/variant call from a mono allelic or completely imprinted gene. The same is of course true for a gene where only the ref/ref is seen.

Applications

As has been stated, RNA sequencing allows for a global analysis of all transcripts in a sample. Eizirik *et al.* (2012) has used this to measure the response of islets of Langerhans to cytokines, something they previously done with microarrays (Eizirik *et al.* 2008, Ortis *et al.* 2010, Gurzov *et al.* 2012) to model islet response during the development of type 1 diabetes (T1D). The aim of the investigation is to fill the gap between the strong genetic components of T1D and the environmental factors that contribute to the disease. In their initial investigation, they were able to show that out of the 8000 genes seen expressed in the islets 20% were regulated in response to cytokines. RNA sequencing detected 29 776 transcripts and here 35% of these were regulated. They also saw that > 60% of the T1D candidate genes were expressed in the islets of Langerhans and that several of them were regulated by the cytokines. This is a clear example whereby the microarray data, albeit informative, still misses information that can be seen with RNA sequencing.

There has also been studies on how islets of Langerhans respond to saturated fatty acids, something that is known to cause B-cell failure and lead to type 2 diabetes (T2D; Wilding 2007). In a recent study, islets of Langerhans have been incubated with palmitate (Cnop *et al.* 2014). Here they showed a regulation of 1325 genes and alternative splicing of 3525 genes, which would have been missed using a microarray approach.

Hence an RNA sequencing experiment compared with a 'normal' gene expression experiment can give information regarding many other aspects, among those indicating different splice variants expressed. As there are data showing that different splice variants have different functions (Small & Pickering 2009, Kakhkhiaie *et al.* 2014, Michel *et al.* 2014), it is important to remember that a genes function is so much more than just gene expression levels. Therefore in the future it is likely that RNA sequencing can be used much in the same way as HDL/LDL measurements are being used today, but instead of looking at the lipid levels in the blood the biological response in the form of specific splice variants can be used.

Main problems and future of RNA sequencing

All techniques have their Achilles heel and the major problem in RNA sequencings is with sampling. When performing RNA sequencing experiments, more abundant transcripts are sampled more than lowly expressed transcripts and as much as 70% of the gene expression can come from very few genes (Shin *et al.* 2014). While it still is relatively easy to investigate gene expression levels of transcripts that are highly as well as lowly expressed at the same time more complex analysis such as allelic imbalance and splice variation can be harder. This is because only a fraction of the data from a given gene is going to be from the SNP investigated or the exon/exon boundaries of interest, increasing the risk of not enough data for a statistically stable investigation. Therefore, when investigating specific splice variants from a gene, the total amount of data needed is higher. However, it might not be practically or economically feasible to increase the read depth enough, especially when the gene(s) of interest is very lowly expressed.

There are ways to alleviate the problem. In part, this is built into the sample preparations themselves because the poly(A) selection or rRNA depletion focuses the sequencing power vs the type of RNA of interest. If this is not sufficient, there is a possibility to either remove more transcripts or to select for the transcripts of interest. When investigating the expression levels in blood, it is common to remove globin RNAs because these transcripts can make up to 70% of the total amount of transcripts in the samples (Shin *et al.* 2014). If you have a specific set of transcripts that you are interested in, then it is possible to perform a pull down of the said transcripts using probes (Mercer *et al.* 2011), thereby ensuring that the sequencing is predominantly performed on transcripts of interest. Of these two methods, globin depletion still allows for a set up without any prior assumptions, while the active selection is somewhat dependent on the targets being known. A targeted pull down followed by sequencing, in the future if possible, could even replace qPCR as the main quantitative method to measure even only a few transcripts when the transcriptome in a model system is clearly defined. By using a smart probe placement one could target all the transcripts of interest and ensure an in-depth analysis of all splice variants, with potentially different functions (Small & Pickering 2009, Kakhkhiaie *et al.* 2014, Michel *et al.* 2014), simultaneously. However, for it to be feasible to replace qPCR for one or a few genes, the cost of targeting needs to drop drastically.

One other drawback or limitation is the read length. The fact that the read length is shorter than the transcripts investigated means that there is the need for alignments

and transcript reassembly, two processes that comes with inherent issues. The sequencing technology is under constant development and longer reads lengths are possible all the time. Notably, the read length from Pacific Bioscience have been reported to produce reads of up to 30–40 kb in length with an average read length per run close to 10 kb. This would largely solve the problem that the data are uncoupled, i.e., that the exon usage in the beginning and the end of a transcript is ‘only’ statistically coupled. By sequencing the entire transcript in one long read, the exact sequence of a transcript would be defined helping in the analysis of allele-specific expression and allowing for a definite link between exon usage, clearly showing when there are mutually exclusive exons in transcripts (Stephan *et al.* 2007). There are however other limitations to the Pacific Bioscience set-up and one major factor limiting in its use is the cost. Oxford nanopor is one company that promises long read lengths without the costly initial investigation. They are currently performing a worldwide test to show their capabilities and this will tell whether their technology works.

Summary

The development of NGS sequencing has been a mile stone in the investigation of RNAs. For the first time, we are able to investigate the expression in a given sample without priori knowledge in a quantitative manner. Based on this, we can detect what happened during the development of disease, after treatment, or during development without filtering the investigation on previously known transcripts for the first time. Considering the complexity of the transcriptome with tissue-specific splicing creating functionally different transcripts from the same a tools such as RNA sequencing is sorely needed.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of this review.

Funding

We would like to acknowledge The Swedish Research Councils generous contributions (grant number 2010-3490 and 2007-4037).

Acknowledgements

To the authors acknowledge The Swedish Research Councils generous contributions.

References

- Agulnik SI, Papaioannou VE & Silver LM 1998 Cloning, mapping, and expression analysis of TBX15, a new member of the T-box gene family. *Genomics* **51** 68–75. (doi:10.1006/geno.1998.5278)
- Asmann YW, Klee EW, Thompson EA, Perez EA, Middha S, Oberg AL, Therneau TM, Smith DI, Poland GA, Wieben ED *et al.* 2009 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics* **10** 531. (doi:10.1186/1471-2164-10-531)
- Cai G, Li H, Lu Y, Huang X, Lee J, Muller P, Ji Y & Liang S 2012 Accuracy of RNA-Seq and its dependence on sequencing depth. *BMC Bioinformatics* **13** (Suppl 13) S5. (doi:10.1186/1471-2105-13-S13-S5)
- Cnop M, Abdulkarim B, Bottu G, Cunha DA, Igoillo-Esteve M, Masini M, Turatsinze JV, Griebel T, Villate O, Santin I *et al.* 2014 RNA-sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. *Diabetes* **63** 1978–1993. (doi:10.2337/db13-1383)
- Cooper TA, Wan L & Dreyfuss G 2009 RNA and disease. *Cell* **136** 777–793. (doi:10.1016/j.cell.2009.02.011)
- Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y & Pritchard JK 2009 Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25** 3207–3212. (doi:10.1093/bioinformatics/btp579)
- Eizirik DL, Moore F, Flamez D & Ortis F 2008 Use of a systems biology approach to understand pancreatic β -cell death in type 1 diabetes. *Biochemical Society Transactions* **36** 321–327. (doi:10.1042/BST0360321)
- Eizirik DL, Sammeth M, Bouckennooghe T, Bottu G, Sisino G, Igoillo-Esteve M, Ortis F, Santin I, Colli ML, Barthson J *et al.* 2012 The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. *PLoS Genetics* **8** e1002552. (doi:10.1371/journal.pgen.1002552)
- Engstrom PG, Steijger T, Sipos B, Grant GR, Kahles A, Alioto T, Behr J, Bertone P, Bohnert R, Campagna D *et al.* 2013 Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* **10** 1185–1191. (doi:10.1038/nmeth.2722)
- Gurzov EN, Barthson J, Marfour I, Ortis F, Naamane N, Igoillo-Esteve M, Gysemans C, Mathieu C, Kitajima S, Marchetti P *et al.* 2012 Pancreatic β -cells activate a JunB/ATF3-dependent survival pathway during inflammation. *Oncogene* **31** 1723–1732. (doi:10.1038/onc.2011.353)
- Heid CA, Stevens J, Livak KJ & Williams PM 1996 Real time quantitative PCR. *Genome Research* **6** 986–994. (doi:10.1101/gr.6.10.986)
- Kakhkhaie KR, Moaven O, Abbaszadegan MR, Montazer M & Gholamin M 2014 Specific MUC1 splice variants are correlated with tumor progression in esophageal cancer. *World Journal of Surgery* **38** 2052–2057. (doi:10.1007/s00268-014-2523-1)
- Kevil CG, Walsh L, Laroux FS, Kalogeris T, Grisham MB & Alexander JS 1997 An improved, rapid Northern protocol. *Biochemical and Biophysical Research Communications* **238** 277–279. (doi:10.1006/bbrc.1997.7284)
- Korf I 2013 Genomics: the state of the art in RNA-seq analysis. *Nature Methods* **10** 1165–1166. (doi:10.1038/nmeth.2735)
- Le K, Mitsouras K, Roy M, Wang Q, Xu Q, Nelson SF & Lee C 2004 Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Research* **32** e180. (doi:10.1093/nar/gnh173)
- Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS & Rinn JL 2011 Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nature Biotechnology* **30** 99–104. (doi:10.1038/nbt.2024)
- Michel LY, Verkaart S, Koopman WJ, Willems PH, Hoenderop JG & Bindels RJ 2014 Function and regulation of the Na^+ - Ca^{2+} exchanger NCX3 splice variants in brain and skeletal muscle. *Journal of Biological Chemistry* **289** 11293–11303. (doi:10.1074/jbc.M113.529388)
- Modrek B & Lee C 2002 A genomic view of alternative splicing. *Nature Genetics* **30** 13–19. (doi:10.1038/ng0102-13)

- Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5** 621–628. (doi:10.1038/nmeth.1226)
- Nag A, Savova V, Fung HL, Miron A, Yuan GC, Zhang K & Gimelbrant AA 2013 Chromatin signature of widespread monoallelic expression. *eLife* **2** e01256. (doi:10.7554/eLife.01256)
- Ortis F, Naamane N, Flamez D, Ladriere L, Moore F, Cunha DA, Colli ML, Thykjaer T, Thorsen K, Orntoft TF *et al.* 2010 Cytokines interleukin-1 β and tumor necrosis factor- α regulate different transcriptional and alternative splicing networks in primary β -cells. *Diabetes* **59** 358–374. (doi:10.2337/db09-1159)
- Pan Q, Shai O, Lee LJ, Frey BJ & Blencowe BJ 2008 Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40** 1413–1415. (doi:10.1038/ng.259)
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y & Pritchard JK 2010 Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464** 768–772. (doi:10.1038/nature08872)
- Rast JP, Amore G, Calestani C, Livi CB, Ransick A & Davidson EH 2000 Recovery of developmentally defined gene sets from high-density cDNA microarrays. *Developmental Biology* **228** 270–286. (doi:10.1006/dbio.2000.9941)
- Richard H, Schulz MH, Sultan M, Nurnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M *et al.* 2010 Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research* **38** e112. (doi:10.1093/nar/gkq041)
- Savova V, Vigneau S & Gimelbrant AA 2013 Autosomal monoallelic expression: genetics of epigenetic diversity? *Current Opinion in Genetics & Development* **23** 642–648. (doi:10.1016/j.gde.2013.09.001)
- Schena M, Shalon D, Davis RW & Brown PO 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270** 467–470. (doi:10.1126/science.270.5235.467)
- Shin H, Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, Wilson-McManus JE, Ng RT, McManus BM & Tebbutt SJ 2014 Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PLoS ONE* **9** e91041. (doi:10.1371/journal.pone.0091041)
- Small TW & Pickering JG 2009 Nuclear degradation of Wilms tumor 1-associating protein and survivin splice variant switching underlie IGF-1-mediated survival. *Journal of Biological Chemistry* **284** 24684–24695. (doi:10.1074/jbc.M109.034629)
- Soneson C & Delorenzi M 2013 A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14** 91. (doi:10.1186/1471-2105-14-91)
- Stephan M, Moller F, Wiehe T & Kleffe J 2007 Self-alignments to detect mutually exclusive exon usage. *In Silico Biology* **7** 613–621.
- Sugnet CW, Srinivasan K, Clark TA, O'Brien G, Cline MS, Wang H, Williams A, Kulp D, Blume JE, Haussler D *et al.* 2006 Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Computational Biology* **2** e4. (doi:10.1371/journal.pcbi.0020004)
- Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D *et al.* 2008 A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321** 956–960. (doi:10.1126/science.1160342)
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL & Pachter L 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7** 562–578. (doi:10.1038/nprot.2012.016)
- Wang Z, Gerstein M & Snyder M 2009 RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics* **10** 57–63. (doi:10.1038/nrg2484)
- Wilding JP 2007 The importance of free fatty acids in the development of type 2 diabetes. *Diabetic Medicine* **24** 934–945. (doi:10.1111/j.1464-5491.2007.02186.x)

Received in final form 6 July 2014

Accepted 24 July 2014