Application of RNA-seq transcriptomic analysis to reproductive physiology of the pig: Insights into differential trophoblast function within the late gestation porcine placenta

Anthony K. McNeel[1], Celine Chen[2], Steven Schroeder[3], Tad Sonstegard[3], Harry Dawson[2] and Jeffrey L. Vallet[1]

[1]Reproduction Research Unit, USDA, ARS U.S. Meat Animal Research Center, Clay Center, NE 68933-0166; [2] Diet, Genomics and Immunology Laboratory, [3]Bovine Functional Genomics Laboratory, USDA, ARS Beltsville Agricultural Research Center, Beltsville, MD 20705*

Next generation DNA sequencing is a high throughput method of sequencing DNA samples in parallel. During the last 10 years, this technology has expanded to include sequencing and quantification of an entire transcriptome. The advantage of this method of transcriptome analysis is that it allows the investigator to detect previously unknown genes and splice variants as well as detect potential DNA polymorphisms. Application of this technology, especially when used to perform cDNA sequencing, allows for comprehensive characterization of transcriptomes between cell types, tissues or under different physiological states. In this review, we summarize high throughput transcriptome analysis, the sequencing platforms currently available, some of the software needed to handle the data generated and how to develop a picture of what the data means from a physiologist's point of view. Lastly, we describe an example of this type of analysis applied to porcine placental trophoblast cell types.

Microarray analysis

A derivative of Southern blotting, one of the earliest reports of transcriptome analysis using a microarray was performed in *Arabidposis* on 45 genes (Schena *et al*. 1995). After almost 20

E-mail: jeff.vallet@ars.usda.gov

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

years, the supporting technology and bioinformatics associated with microarrays is relatively stable and easy to perform. In general, the RNAs of the transcriptome from a sample are labeled and hybridized to probes that have been previously arrayed on a supporting surface, and the amount of label hybridized to each probe is measured. The most significant drawback with microarrays is that only those genes and splice variants included on the array are detected. Previously unknown splice variants are also not easily detected. Lastly, it is usually necessary to validate differential expression of genes derived from microarray analysis, typically using real-time RT-PCR.

Serial Analysis of Gene Expression (SAGE) analysis

SAGE was developed in 1995 by Dr. Victor Velculescu and colleagues (Velculescu *et al*. 1995) called serial analysis of gene expression (SAGE. It is similar to next generation sequencing in that the process produces short sequence fragments (< 20 bp) that are then mapped back to a reference sequence/genome via computational analysis. Because of the procedure, the sequences produced in SAGE are usually from the 3′ end of a transcript. Depending on the completeness of the genome of interest, the short read lengths and bias toward the 3″ end of mRNA can limit the number of genes detectable using this technique.

High throughput (aka next generation) sequencing

Next generation sequencing (NGS) utilizes the sequencing by synthesis approach to DNA sequencing. Currently, there are four major platforms which utilize this approach: Roche (454), Life Technologies (Ion Torrent), Pacific BioSciences (SMRT), Illumina (Genome Analyzer, HiSeq, MiSeq). While each platform has specific advantages and disadvantages, each system sequences DNA strands by generating the complimentary strand. This approach is similar to the addition of dideoxynucleotides used in Sanger Sequencing (e.g ABI 3730) but instead of terminating the DNA synthesis reaction with the addition of a dideoxynucleotide, the nucleotides added during synthesis are either differentially labeled or are made sequentially available to the ongoing sequencing reaction for incorporation. The addition of nucleotides to the elongating strand is then detected, so that either the label or the sequence of added substrate allows one to determine which nucleotide was incorporated at each position.

The output of the ION Torrent system is currently limited to 1-2 Gb with 400 bp read length. Data quality is robust with 99.99% accuracy but the system has a difficultly sequencing through AT rich regions (Quail *et al*. 2012). The platform is currently most commonly used for sequencing of microbial genomes.

The Roche 454 system operates on a pyrosequencing approach where library fragments are attached to beads in a ratio of 1:1. Then, PCR amplification takes place in an emulsion to isolate beads and ensure purity of the DNA. Beads are then randomly allocated to nanopores and sequencing occurs as labeled nucleotides are flowed across the sequencing chamber and the resulting signal is generated via chemiluminesence. Maximal read length is 1000 bp with current chemistries with an average output of 0.7 Gb and consensus accuracy of 99.997%.

The PacBio SMRT system has the lowest output of all the systems but is capable of very long reads; up to + 20 kb. The PacBio is currently primarily used for sequencing prokaryotic genomes. The SMRT (Single Molecule Real Time) technology uses a high speed camera to capture sequence data as a DNA polymerase molecule incorporates a nucleotide in a single strand of DNA in real time. Due to the speed at which the polymerase incorporates nucleotides into the complimentary strand, the SMRT system has a higher incidence of incorrectly called bases. However, this error is random and individual DNA templates can be circularized and sequenced numerous times in a single run, so that errors can be eliminated by averaging multiple base calls at the same

position. The SMRT system performs particularly well when coupled with reads from the Illumina platform as the large amount of contiguous sequence available from the PacBio combined with the short error free reads from Illumina technology are very useful for dealing with repetitive elements in DNA.

The Illumina HiSeq platform provides the greatest amount of sequence data of the systems available, capable of generating 600 Gb of sequence in a single run through massively parallel sequencing. This is enough output to theoretically produce > 200x coverage of a single porcine genome. To accomplish this, DNA fragments in a sequencing library are fixed to a stationary media. This is accomplished by fragmenting the DNA followed by size selection and ligation of an adaptor sequence that includes a unique sequence identifier tag. The adaptor sequence facilitates the fixation of the DNA fragment to the flow cell while the unique identifier sequence tag allows for multiplexing of different samples (e.g., different tissue sources, animals, etc.) during sequencing. The DNA plus the adaptor is then amplified using PCR to facilitate bridge amplification of the DNA strands. During bridge amplification, the dsDNA is separated into single strands and non-labeled nucleotides are used to replicate a "cluster" of approximately 1000 identical fragments of DNA within 1 μm in diameter. Once cluster generation is complete, the non-labeled nucleotides are washed from the flow cell and a mix of primer, DNA polymerase and labeled nucleotides is applied to the flow cell to initiate the first detected cycle of cDNA synthesis. From this point forward the instrument performs repetitive cycles (up to 200) of nucleotide incorporation, removal of the non-incorporated nucleotides, identification of the incorporated nucleotide and application of fresh labeled nucleotides. The process is both time consuming (up to 11 days for 200 bp of sequence) and inefficient in its use of the labeled nucleotide (25%). However, balancing this is the production of very high quality data with more than 80% of the bases possessing a high quality score with an accuracy of 99.99%.

Samples throughput, depth of coverage and data files

The amount of sequence information obtainable combined with the ability to individually tag sequences from different libraries within a single sequencing run provides great flexibility. The distribution of libraries within a single run depends on the depth of sequencing desired for each library. This is dependent on the goals of the experiment and the tissues being examined. Transcript discovery projects, where the detection of novel transcripts is important, demand the greatest depth of sequencing for each library within the experiment. SNP discovery experiments require less depth of sequencing for individual libraries but the libraries are based on samples from a greater number of animals. Physiology experiments fall somewhere in between. A useful target for the evaluation of a transcriptome of a given sample is 20 million sequences, though the number continues to evolve as the technique is refined. As such, the sequencing files generated are typically very large (gigabytes) and the current Illumina software for exporting read data from the sequencer for individually tagged sequences truncates files at 4 million reads. Handling files this large necessitates access to software and hardware capable of handling large datasets.

Gene identification: Shareware vs proprietary software

Once sequences are in the correct format and ready for processing, they need to be matched to a set of reference sequences. For RNA-seq, the reference sequences are usually from a fully annotated genome, which is available for swine.

For RNA-seq, the gold standard of available freeware is a suite of related software packages called TopHat and CuffLinks (Trapnell *et al*. 2009, 2010, 2013). These two programs map the reads generated during sequencing to a reference sequence, and then make comparisons

between given treatment groups. Use of these programs requires some knowledge of the use of the UNIX operating system, which can be confusing to inexperienced users. However, proprietary software packages designed for the non-bioinformatician have been developed and are currently available.

These proprietary packages typically utilize a graphic user interface to facilitate utilization of the program (e.g. CLC Biosciences Genome Workbench). With this system the user directs the needed operations of the software using a mouse and keyboard. The specific algorithms used for mapping reads are proprietary, but one advantage of the analysis approach used by Genome Workbench is that it processes in parallel, reducing the amount of time required to map the reads.

Counting reads and statistical analysis

Quantifying reads mapped to a gene is a complex problem. Transcripts vary in size. During fragmentation of cDNA, if transcript numbers are equal, more fragments are generated from transcripts of greater length than those of shorter length, which must be taken into account when quantifying the abundance of the original transcripts. There are currently three different methods of quantifying RNA-seq data: TPM (Wagner *et al*. 2012), FPKM (**F**ragments **P**er **K**ilobase of exon model per **M**illion mapped reads) and RPKM (**R**eads **P**er **K**ilobase of exon model per **M**illion mapped reads; (Trapnell *et al*. 2010). All of the methods are based on mapping each sequence in a library within genes in a reference list (Figure 1). Each method takes into account the depth of sequencing of each library and the different sizes of transcripts. FPKM and RPKM are related methods and generate similar data. The RPKM method is the most common method of quantifying RNA-seq data as it treats each sequence fragment in a paired end read (see below) as one piece of the same unit. FPKM counts each fragment of a paired end read as an independent unit and counts them separately. Thus, with RPKM, a paired end read with two mappable fragments counts as 1, whereas in FPKM, the two fragments count as two. The advantage of FPKM is that when dealing with fragments of divergent quality, if only one fragment maps the value is one, compared to RPKM which would result in a count of 0. TPM makes similar adjustments for transcript size and the number of mapped reads, but also adjusts for the average size of transcripts in each library.

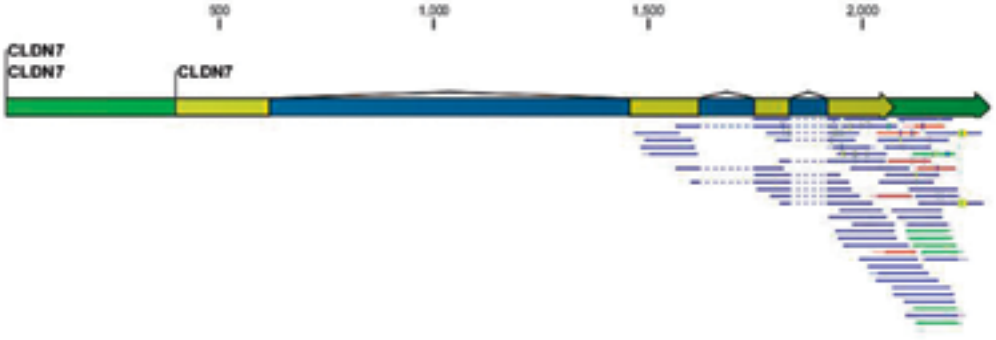

Fig. 1 Mapping results for Claudin-7 (CLDN7). The number of reads mapped to a reference sequence indicates the prevalence of a given sequence within a library. In this graphic, 5′ and 3′ UTR are annotated by the green areas of the gene, followed by the exons in yellow and introns in blue. Below the reference sequence, the color of the reads indicates if both fragments of the read in this paired end library mapped to the reference sequence (blue), if the forward read mapped (green) and if the reverse read mapped (red).

Once the abundance of the transcripts for each transcriptome is quantified, the abundance values can then be used for statistical analysis. Which analysis is conducted varies, but one of the most common methods is to use Cufflinks to determine if statistical differences exist between treatment groups. The statistical treatment used in Cufflinks utilizes a program named Cuffdiff to make comparisons using a Student's t-test. Alternatively, using an ANOVA with adaptive false discovery rate for multiple testing provides researchers with robust results that account for biological variation, replication, and the number of tests performed.

Reliability of transcript measurement

The repeatability of transcript measurements using the Illumina system is high (Spearman correlation value of replicate runs equals 0.96 (Marioni *et al*. 2008). The generation of reliable sequence data of 100 bp or greater makes sequence matching to reference sequences very reliable and has the added benefit of detection of previously unknown splice variants. Unmatched sequences provide an opportunity to detect mRNA that are not in the reference gene set if desired, which can be important depending on the quality of the reference gene set available. These advantages are even greater with the ability to generate paired-end sequence reads. Paired-end reads are generated when sequence data is collected from both ends of each fragment, which is available for selection during an Illumina sequencing run. Knowing that the paired- end sequence information conforms to this arrangement increases the probability of matching to a reference gene set. It can also improve identification of splice variants, when matching to the reference gene indicates that the distance between the paired- ends exceeds the predicted 100 bp intervening sequence. Finally, the paired-end sequence can be used for *de novo* assemblies of cDNA, which can then be used to improve the correspondence between the transcriptome and the genome, as well as the detection of SNP within cDNA.

Pathway analysis

The result of a transcriptome analysis is a list of differences in gene expression along with p values. This would normally allow one to make interpretations of the results, but in the case of a comparison between transcriptomes, the list of genes that are statistically different can be in the thousands and does not take into consideration the biological roles of genes that were not different. Understanding what the genes in a given cell/tissue contribute to biological function is an important consideration when drawing conclusions from RNA-seq data and is usually accomplished through the use of pathway analysis.

Similar to software for mapping of reads to a reference sequence, there are shareware and proprietary software options for understanding the biological functions found in a transcriptome analysis. These software packages assemble known functions and/or associations from published reports, and provide the researcher with the ability to interpret their transcriptomic results on the basis of those known functions. This type of analysis can allow the formulation of testable hypotheses for future research. Two of the most common software packages available are: Database Annotation, Visualization and Integrated Discovery (DAVID; Dennis *et al*. 2003), an open source software package, and Ingenuity Pathway Analysis (IPA, www.igenuity.com) a commercially available software. Both are available online. While these software packages are incredibly beneficial to the biologist, the results generated in their analysis can be very biased by the version of the database/program used (Henderson-Maclennan *et al*. 2010). The number of genes analyzed by these programs can also affect the results of the analysis.

Usually, it is recommended that investigators filter their results according to abundance of the transcript and/or changes between treatment groups, so that results of pathway analyses are more meaningful. Taking such an approach makes it easier for researchers to focus on the main functions found within the gene list and what functions are undergoing the most change. Lastly, these programs make it possible to identify upstream regulators that may be responsible for the changes observed.

A transcriptome analysis: Pig trophoblast cells

Trophoblast epithelial cells in the folded bilayer of the porcine placenta differ in morphology, with tall columnar (TC) trophoblasts located at the peaks and short cuboidal (SC) trophoblasts lining the sides and bottoms of the folds (Friess *et al*. 1980). We used transcriptomic analysis to determine 1) how these two trophoblast cell types contribute to placental function and 2) how they differ from each other. To accomplish this, we performed laser capture microdissection followed by high throughput sequencing of cDNA to characterize the transcriptome of each cell type.

Total RNA was extracted from the two cell types of interest from 4 different animals on day 85 of gestation. RNA was reverse transcribed and resulting cDNA was amplified. Sequencing libraries were then prepared for each animal/cell type combination, each library was tagged, and libraries were sequenced on an Illumina Hi-Seq 2000. Reads were then mapped to the porcine genome (version 9.2). Transcript abundance was expressed using the RPKM method. Contrasts of the RPKM values between treatment groups were made using ANOVA with an adaptive false discovery rate set at 0.05 in JMP genomics v 5.0. Functional analyses were generated through the use of Ingenuity Pathway Analysis (IPA; www.ingenuity.com).

Main (20 RPKM threshold) functions of the average trophoblast

Mapping of sequencing reads to the *Sus scrofa* genome (version 9.2) resulted in the identification of 7398 unique genes/sequences with an RPKM value $\geq$ 1.8. The RPKM values for SC and TC trophoblasts were averaged to provide values for the average trophoblast, and the list of genes above each threshold were used for IPA analysis. The top five Main non-disease/non-disorder biological functions for trophoblasts were Cell death and survival, Protein synthesis, Cellular growth and proliferation, Post-transcriptional modification and Protein folding.

Main (20 RPKM) differences between short cuboidal and tall columnar trophoblasts

Restricting the gene lists to only those genes with an RPKM 20 and p-value 0.05 reduced the number of genes included in the analysis to 445 total genes (372, 343 genes in SC and TC trophoblasts respectively). Results from pathway analysis indicated pathways responsible for Cellular growth and proliferation, Cell death and survival, Cell morphology, and Hematological system development and function.

Analysis of downstream effects of these differences in gene expression indicated that SC trophoblasts were associated with increased Engulfment of cells, Engulfment of myeloid cells, Response of myeloid cells, Adhesion of lymphocytes and Engulfment of antigen presenting cells (Z-score: 3.101, 2.688, 2.648, 2.642 and 2.608 respectively). The pathway in SC trophoblasts with the highest p-value and increased function was Invasion of cells ($P = 5.54x10^{-9}$, Z-score: 2.278). Short Cubodial trophoblasts were predicted to be more mobile than TC trophoblasts

as the top five functional annotations for Cellular movement are all increased (Z-score range: 2.426- 2.010). These results support our earlier observations that heparanase expression is primarily found in the SC trophoblasts (Miles *et al*., 2009), and suggest that these cells are the primary contributors to placental fold development based on their capacity for increased cellular movement, specifically those genes known to facilitate cellular invasion such as heparanase (Koliopanos *et al*. 2001, Goldshmidt *et al*. 2003, Gingis-Velitski *et al*. 2004, 2007), prostaglandin-endoperoxide synthase-2 (COX-2) (Neufang *et al*. 2001, Debruyne *et al*. 2002, Yiu and Toker 2006, Banu *et al*. 2008, Slaby *et al*. 2009, Ding *et al*. 2010, Mitchell *et al*. 2010) a rate-limiting enzyme in the biosynthesis of prostaglandin E(2 and CD44 (Bourguignon *et al*. 2004, Kim *et al*. 2004, Celetti *et al*. 2005).

Increased expression of genes in TC trophoblasts was consistent with an increase in the Quantity of cells, Formation of cells, Protein metabolism, the Quantity of reactive oxygen species and Apoptosis of epithelial cells (Z score: 2.694, 2.476, 2.270, 2.207, 2.128 respectively). An increase in pathways involved in Protein metabolism was supported by increased mRNA expression of *UBE2B*, *TPP2*, and *ITCH*, which are known to increase protein degradation (Pati *et al*. 1999, Heissmeyer *et al*. 2004, Bhutani *et al*. 2007)few in vivo targets of the mammalian Cdc34 and Rad6 ubiquitin-conjugating enzymes are known. A yeast-based genetic assay to identify proteins that interact with human Cdc34 resulted in three cDNAs encoding bZIP DNA binding motifs. Two of these interactants are repressors of cyclic AMP (cAMP. The prediction of increased activity in reactive oxygen species scavenging pathways was supported by increased mRNA expression of *PRDX1* and *PRDX6* (Kang *et al*. 1998, Kim *et al*. 2000, Manevich *et al*. 2002, Wang *et al*. 2003, Egler *et al*. 2005), which are known to reduce H_2O_2 concentrations.

Conclusions

High throughput sequencing is a valid and versatile method for performing a comprehensive transcriptomic analysis. Important details for consideration are the experimental design, the depth of sequencing required to provide useful data, and the availability and quality of reference sequences for matching. Despite some limitations imposed by the quality of the swine genome reference sequence used for the analysis, transcriptomic analysis using high throughput sequencing has provided intriguing clues to functions of pig placental trophoblast cells. While further research is clearly needed to explore the functions of these two placental trophoblast cell types, these results provide results from which testable hypotheses can be generated.

Bibliography

Banu SK, Lee J, Speights VO Jr, Starzinski-Powitz A & Arosh JA 2008 Cyclooxygenase-2 regulates survival, migration, and invasion of human endometriotic cells through multiple mechanisms. *Endocrinology* **149** 1180–1189.

Bhutani N, Venkatraman P & Goldberg AL 2007 Puromycin-sensitive aminopeptidase is the major peptidase responsible for digesting polyglutamine sequences released by proteasomes during protein degradation. *The EMBO journal* **26** 1385–1396.

Bourguignon LYW, Singleton PA, Diedrich F, Stern R & Gilad E 2004 CD44 interaction with Na+-H+ exchanger (NHE1) creates acidic microenvironments leading to hyaluronidase-2 and cathepsin B activation and breast tumor cell invasion. *The Journal of biological chemistry* **279** 26991–27007.

Card C, Anderson EJ, Zamberlan S, Krieger K-BE, Kaproth M & Sartini BL 2013 Cryopreserved Bovine Spermatozoal Transcript Profile as Revealed by High-Throughput Ribonucleic Acid Sequencing. *Biology of reproduction*.

Celetti A, Testa D, Staibano S, Merolla F, Guarino V, Castellone MD, Iovine R, Mansueto G, Somma P, De Rosa G, Galli V, Melillo RM & Santoro M 2005 Overexpression of the cytokine osteopontin identifies aggressive laryngeal squamous cell carcinomas

and enhances carcinoma cell proliferation and invasiveness. *Clinical cancer research: an official journal of the American Association for Cancer Research* **11** 8019–8027.

Debruyne PR, Bruyneel EA, Karaguni I-M, Li X, Flatau G, Müller O, Zimber A, Gespach C & Mareel MM 2002 Bile acids stimulate invasion and haptotaxis in human colorectal cancer cells through activation of multiple oncogenic signaling pathways. *Oncogene* **21** 6740–6750.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC & Lempicki RA 2003 DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology* **4** P3.

Ding Q, Bai Y-F, Wang Y-Q & An R-H 2010 TGF-beta1 reverses inhibition of COX-2 with NS398 and increases invasion in prostate cancer cells. *The American journal of the medical sciences* **339** 425–432.

Egler RA, Fernandes E, Rothermund K, Sereika S, De Souza-Pinto N, Jaruga P, Dizdaroglu M & Prochownik EV 2005 Regulation of reactive oxygen species, DNA damage, and c-Myc function by peroxiredoxin 1. *Oncogene* **24** 8038–8050.

Friess AE, Sinowatz F, Skolek-Winnisch R & Träautner W 1980 The placenta of the pig. I. Finestructural changes of the placental barrier during pregnancy. *Anatomy and embryology* **158** 179–191.

Gingis-Velitski S, Ishai-Michaeli R, Vlodavsky I & Ilan N 2007 Anti-heparanase monoclonal antibody enhances heparanase enzymatic activity and facilitates wound healing. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **21** 3986–3993.

Gingis-Velitski S, Zetser A, Flugelman MY, Vlodavsky I & Ilan N 2004 Heparanase induces endothelial cell migration via protein kinase B/Akt activation. *The Journal of biological chemistry* **279** 23536–23541.

Gingras A-C, Raught B & Sonenberg N 1999 eIF4 INITIATION FACTORS: Effectors of mRNA Recruitment to Ribosomes and Regulators of Translation. *Annual Review of Biochemistry* **68** 913–963.

Goldshmidt O, Zcharia E, Cohen M, Aingorn H, Cohen I, Nadav L, Katz B-Z, Geiger B & Vlodavsky I 2003 Heparanase mediates cell adhesion independent of its enzymatic activity. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **17** 1015–1025.

Harris TE, Chi A, Shabanowitz J, Hunt DF, Rhoads RE & Lawrence JC Jr 2006 mTOR-dependent stimulation of the association of eIF4G and eIF3 by insulin. *The EMBO journal* **25** 1659–1668.

Heissmeyer V, Macián F, Im S-H, Varma R, Feske S, Venuprasad K, Gu H, Liu Y-C, Dustin ML & Rao A 2004 Calcineurin imposes T cell unresponsiveness through targeted proteolysis of signaling proteins. *Nature immunology* **5** 255–265.

Henderson-Maclennan NK, Papp JC, Talbot CC Jr, McCabe ERB & Presson AP 2010 Pathway analysis software: annotation errors and solutions. *Molecular genetics and metabolism* **101** 134–140.

Kang SW, Chae HZ, Seo MS, Kim K, Baines IC & Rhee SG 1998 Mammalian peroxiredoxin isoforms can reduce hydrogen peroxide generated in response to growth factors and tumor necrosis factor-alpha. *The Journal of biological chemistry* **273** 6297–6302.

Kim H, Lee TH, Park ES, Suh JM, Park SJ, Chung HK, Kwon OY, Kim YK, Ro HK & Shong M 2000 Role of peroxiredoxins in regulating intracellular hydrogen peroxide and hydrogen peroxide-induced apoptosis in thyroid cells. *The Journal of biological chemistry* **275** 18266–18270.

Kim H-R, Wheeler MA, Wilson CM, Iida J, Eng D, Simpson MA, McCarthy JB & Bullard KM 2004 Hyaluronan facilitates invasion of colon carcinoma cells in vitro via interaction with CD44. *Cancer research* **64** 4569–4576.

Koliopanos A, Friess H, Kleeff J, Shi X, Liao Q, Pecker I, Vlodavsky I, Zimmermann A & Büchler MW 2001 Heparanase expression in primary and metastatic pancreatic cancer. *Cancer research* **61** 4655–4659.

Leiser R & Kaufmann P 1994 Placental structure: in a comparative aspect. *Experimental and clinical endocrinology* **102** 122–134.

Manevich Y, Sweitzer T, Pak JH, Feinstein SI, Muzykantov V & Fisher AB 2002 1-Cys peroxiredoxin overexpression protects cells against phospholipid peroxidation-mediated membrane damage. *Proceedings of the National Academy of Sciences of the United States of America* **99** 11599–11604.

Marioni JC, Mason CE, Mane SM, Stephens M & Gilad Y 2008 RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18** 1509–1517.

Mitchell K, Svenson KB, Longmate WM, Gkirtzimanaki K, Sadej R, Wang X, Zhao J, Eliopoulos AG, Berditchevski F & Dipersio CM 2010 Suppression of integrin alpha3beta1 in breast cancer cells reduces cyclooxygenase-2 gene expression and inhibits tumorigenesis, invasion, and cross-talk to endothelial cells. *Cancer research* **70** 6359–6367.

Neufang G, Furstenberger G, Heidt M, Marks F & Müller-Decker K 2001 Abnormal differentiation of epidermis in transgenic mice constitutively expressing cyclooxygenase-2 in skin. *Proceedings of the National Academy of Sciences of the United States of America* **98** 7629–7634.

Pati D, Meistrich ML & Plon SE 1999 Human Cdc34 and Rad6B ubiquitin-conjugating enzymes target repressors of cyclic AMP-induced transcription for proteolysis. *Molecular and cellular biology* **19** 5001–5013.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP & Gu Y 2012 A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13** 341.

Schena M, Shalon D, Davis RW & Brown PO 1995 Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science* **270** 467–470.

Slaby O, Svoboda M, Michalek J & Vyzula R 2009 MicroRNAs in colorectal cancer: translation of molecular biology into clinical application. *Molecular cancer* **8** 102.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL & Pachter L 2013 Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology* **31** 46–53.

Trapnell C, Pachter L & Salzberg SL 2009 TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25** 1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ & Pachter L 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28** 511–515.

Vallet JL & Freking BA 2007 Differences in placental structure during gestation associated with large and small pig fetuses. *Journal of Animal Science* **85** 3267–3275.

Velculescu VE, Zhang L, Vogelstein B & Kinzler KW 1995 Serial analysis of gene expression. *Science (New York, N.Y.)* **270** 484–487.

Wagner GP, Kin K & Lynch VJ 2012 Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften* **131** 281–285.

Wang X, Phelan SA, Forsman-Semb K, Taylor EF, Petros C, Brown A, Lerner CP & Paigen B 2003 Mice with targeted mutation of peroxiredoxin 6 develop normally but are susceptible to oxidative stress. *The Journal of biological chemistry* **278** 25179–25190.

Yiu GK & Toker A 2006 NFAT induces breast cancer cell invasion by promoting the induction of cyclooxygenase-2. *The Journal of biological chemistry* **281** 12210–12217.