Developing systems for high-throughput screening of infectious diseases using zebrafish

Veneman, W.J.

Chapter 5

Analysis of RNAseq datasets from a comparative infectious disease zebrafish model using GeneTiles bioinformatics

Wouter J. Veneman[1#], Jan de Sonneville[2#], Kees-Jan van der Kolk[2,3#], Anita Ordas[1], Zaid Al-Ars[3], Robert-Jan Raterink[4], Koen Egberts[4], Thomas Hankemeijer[4], Annemarie H. Meijer[1], Herman P. Spaink[1]

[1] Institute of Biology, Leiden University
[2] Life Science Methods BV
[3] Computer Engineering, Delft University
[4] Leiden Academic Centre for Drug Research, Leiden University

[#]Authors contributed equally

Parts of this chapter are published in Genomics 2014

Abstract

We present a RNA deep sequencing (RNAseq) analysis of a comparison of the transcriptome responses to infection of zebrafish larvae with *Staphylococcus epidermidis* and *Mycobacterium marinum* bacteria. We show how our developed GeneTiles software can improve RNAseq analysis approaches by more confidently identifying a large set of markers upon infection with these bacteria. For analysis of RNAseq data currently software programs such as Bowtie2 and Samtools are indispensable. However, these programs that are designed for a LINUX environment require some dedicated programming skills and have no options for visualization of the resulting mapped sequence reads. Especially with large data sets this makes the analysis time consuming and difficult for non-expert users. We have applied the GeneTiles software to the analysis of previously published and newly obtained RNAseq datasets of our zebrafish infection model and we have shown the applicability of this approach also to published RNAseq datasets of other organisms by comparing our data with a published mammalian infection study. In addition we have implemented the DEXSeq module in the GeneTiles software to identify genes, such as glucagon A, that are differentially spliced under infection conditions. In the analysis of our RNAseq data this has led to the possibility to improve the size of data sets that could be efficiently compared without using problem-dedicated programs, leading to a quick identification of marker sets. This approach will therefore also be highly useful for transcriptome analyses of other organisms for which well-characterized genomes are available. Based on the transcriptome analysis we have analysed the function of the leptin b gene that was identified as the most highly induced gene after infection. Using metabolomics analyses we have shown that infection with *M. marinum* results in a rapid wasting syndrome. Results of gene knock-down analysis suggest that the leptin b gene is essential for progression of the wasting syndrome after bacterial infection.

Introduction

In our previous research we have used zebrafish larval infection models to study the transcriptome response to infection by several pathogens [1-5]. In addition we have tested the response of zebrafish larvae to infection by the opportunistic bacterium *Staphylococcus epidermidis* as a model for biomaterial-associated infections that are often caused by this species in clinical practice [6-9]. These studies have led to a high-throughput model that resulted in a large set of RNAseq data sets highlighting a new bottleneck in our research: the fast and user-friendly analysis of large datasets that can be easily visualized for comparative purposes.

In the analysis of our former transcriptome data sets there was a need for specialized scripting languages to quickly find good marker genes for disease. We used an existing visualization program, Integrative Genomics Viewer (IGV) [10] that shows the data

solely along a line representing the genome, thereby requiring zooming in to view the aligned reads. IGV, and many other open source visualization programs such as MapView [11], Tablet [12], GenoViewer [13] and BamView [14] also require the user to scroll or manually search for other genes to bring these into focus and an overview of a selection of genes based on alignment results is not available. Finally, most of these data visualization programs do not allow for export of presented visual results, other than taking screenshots. In our previous analysis of RNAseq data we were reliant on manual counting of reads as guided by the Integrative GenomeViewer [1].

In this paper we used the transcriptome data set obtained from the zebrafish high-throughput screening system for *S. epidermidis* infection [1] as a case study to optimize and automate the data analysis pipeline. Using the resulting software package we also went further and added a larger RNAseq data set from *Mycobacterium marinum* infection data for comparisons of specificity of the transcriptome responses. We also integrated the DEXSeq algorithm that can be used to give an estimate of probability of the occurrence of differential splicing. This has led to the identification of genes that are differentially spliced after microbial infection in zebrafish larvae. Finally, we wanted to include a comparison with whole organism infection data in other vertebrate species. Unfortunately, there is still few RNAseq data for this available that can be mapped on ENSEMBL genome data and as a result we have only been able to compare our zebrafish infection data with RNAseq data from a bovine digital dermatitis (BDD) model as published by Scholey et al. 2013 [15]. However, the results are sufficient to show that our approach makes also such interspecies comparisons of RNAseq datasets very easy and can quickly lead to conclusions on conserved immune responses, even in comparisons between very different fish and mammalian infection models. One of the genes conserved between fish and mammals and highly induced by infection encodes the hormone leptin b and we selected this gene for functional analysis in this study.

Material and methods

Bacterial strains and growth conditions

The *S. epidermidis* strain O-47 containing a pWVW189 derived mCherry expression vector (De Boer L. unpublished) was grown as described in Veneman et al. (2013). The *M. marinum* strain E11 was grown as described in Carvalho et al. 2011 [16]. Two reaction vials with 1 ml of the culture were centrifuged at 14680 rpm for 1 min. The pellets were combined and washed three times with 1 ml phosphate-buffered saline (PBS). Suspensions were prepared based on the optical density at 600 nm and by plating and cfu determination. The inoculates were suspended in 2% polyvinylpyrrolidone$_{40}$ (PVP$_{40}$, CalBiochem) to 2.0×10^7 or 3.0×10^7 cfu/ml.

Morpholino knock-down experiments

For morpholino knock-down experiments, morpholino oligonucleotides (Gene Tools) were diluted to 0.5mM in nuclease free water. To block leptin b we used an ATG morpholino (5'- TTTTTTGCTTGTTAATATCATCCCT-3'). Embryos injected with leptin b morphlino were staged between the 16 and 128 cell stage and injected with 50 cfu of mCherry expressing *M. marinum* E11 bacteria suspended in 2% PVP_{40}. Embryos were screened at 3, 4 and 5 dpi for bacterial proliferation using the COPAS XL (Union BioMetrica) flow cytometer as described in Veneman et al. (2013).

Metabolic profiling

Samples for mass spectrometry were taken at 4 and 5 dpi. Eight embryos per group were collected into a 1.5 ml reaction vial. All egg water was removed and the embryos were washed 2 times with 200 µl Millipore water. Per sample 80 µl of stock C (table 1) was added and were snap frozen in liquid nitrogen for 30 minutes. Samples were

Table 1

	MeOH	I.S. Carnithine stock (table 2)	I.S. Amine stock (table 3)	Glucose-d7 5.5mM
Stock C	126.4 µl	16 µl	16 µl	1.6 µl

Table 2 (Carnithine stock)

		Step 1: stock 7.3.1.2		Step 2: Spike solution 7.3.1.3		
	Internal STD's (1mg/ml)	µl	µg/ml	µl	µg/ml	µM
1	Carnitine-d3 HCL			20	2.00	9.996
2	Deoxycarnitine-d9 HCL			10	2.00	10.487
3	Choline-d4 HCL			10	2.00	13.923
4	Betaine-d3 HCL			20	2.00	12.769
5	Acetyl-L-carnitine-d3 HCL			10	2.00	8.240
6	Butyryl-L-carni	10	12.5		0.025	0.092
7	Octanoyl-L-carnitine-d3 HCL	10	12.5		0.025	0.076
8	Octadecanoyl-L-carnitine-d3 HCL	10	12.5		0.025	0.054
	Stock solution 7.3.1.2			20		
	MeOH	770		9910		
	Total volume	800		10000		

Table 3 (Amine stock)

Internal Standard	Conc. (mg/ml)	Solvent
L-Ornithine D6	1	H_2O
Beta-Alanine D4	1	H_2O
L-2-aminobutyric acid d6	1	H_2O
2-(4-hydroxy-3methoxyphenyl)ethyl-1,1,2,2-D4-amine	1	H_2O
L-NT-Methyl-D3-L-Histidine	1	H_2O
L-3-(4-Hydroxy-3methoxy-D3-phenyl)alanine	1	H_2O
U-C13n15 cell free AA mix	0.875	H_2O
Ethanolamine-d7	1	H_2O
DL-Homocysteine-d4	1	H_2O
1.4-Butane-d8-diamine (Putrescine)	1	H_2O

Ratio: 1 ml labeled amino acid mix: 40 µl all other amine standards, in a total volume of 10 ml in MeOH.

sonicated for 2 minutes at 40 kHz continuously changing positions in order for local sweet spots in the sonicator. The freeze-sonicate cycle was repeated 3 times. The lysate was vortexed for 5 seconds after which they were centrifuged for 5 minutes at 4 ˚C at 16.1 krcf. Per sample 75 µl supernatant was collected and 8 µl supernatant was collected for quality control. The samples were evaporated using speedvac for 1 hour until they were completely dry. All samples were stored at -80 ˚C. Samples were defrosted in 30 minutes and per sample 40 µl borate buffer was added and vortexed. Per sample 10 µl AccQ-Tag (3 mg/ml) (Waters BV) reagent was added for derivatization. The samples were incubated at 55 ˚C for 10 minutes and were quickly spun down for 2 seconds after which they were diluted 4 times using H_2O.

Mass spectrometry data processing

After Liquid Chromatography mass spectrometry (LC-MC) data acquisition, the data was processed in order to extract the proper target and internal standards (I.S.) peak area as follows. First, an Xcalibur (Thermo Fischer) Processing setup was generated in which all the targets and I.S. are identified (in terms of retention time and exact mass) from one of the QC samples and validated with the system suitability test (SST). Per batch the peak separation and retention time was checked and integrated into the processing setup. When every metabolite was correctly assigned and identified in the processing setup, the whole batch was analysed in the Xcalibur Quan browser. All peaks were integrated and manually checked in the Xcalibur Quan browser. When peaks were bad (i.e. bad peak shapes due to e.g. low concentration), the target was discarded from the study. From all the remaining targets the relative standard deviation (RSD) was calculated within all the QCs. Metabolites with RSDs above 20% were discarded from the study. On the remaining high-quality, normalized metabolite data set, univariate and multivariate analysis was performed. For the multivariate analysis, a principle component analysis (PCA) was used as a data exploratory tool [17]. For the univariate analysis, t-tests or ANOVAs were used.

Zebrafish husbandry

Zebrafish were handled in compliance with animal welfare regulations and maintained according to standard protocols (http://ZFIN.org). Embryos were grown at 28°C in egg water (60 µg/ml Instant ocean sea salt, Sera Marin). The egg water was refreshed every day.

Experimental outline

Infection experiments were performed with mixed egg clutches from wild type ABxTL strain zebrafish. Embryos were staged at 2 hours post fertilization by morphological criteria, and 20 cfu of mCherry expressing *S. epidermidis* O-47 or 30 cfu of mCherry expressing *M. marinum* E11 bacteria suspended in 2% PVP_{40} were injected into the yolk. Automated micro-injections were performed as described in Carvalho et al. [16].

At 5 days post fertilization, embryos (N ~100) were collected from the 2 hours post fertilization injected and non-injected group, snap frozen in liquid nitrogen, and stored at −80°C for RNA isolation.

RNA deep sequencing

RNA isolation was performed as described in Veneman et al. (2013). A total of 3 µg of RNA was used to make RNAseq libraries using the Illumina TruSeq RNA Sample Preparation Kit v2 (Illumina Inc., San Diego, USA). In the manufacturer's instructions two modifications were made. In the adapter ligation step 1 µl instead of 2.5 µl adaptor was used. In the library size selection step the library fragments were isolated with a double Ampure XP purification with a 0.7x beads to library ration. The resulting mRNAseq library was sequenced using an Illumina HiSeq2000 instrument according to the manufacturer's description with a read length of 2 × 50 nucleotides. Image analysis and base calling was done by the Illumina HCS version 1.15.1. The raw RNAseq data have been deposited in the NCBI GEO database with accession number GSE42846, GSE44351 and GSE57792.

Data analysis

GeneTiles was used for quantification and visualization of the RNAseq data. When using GeneTiles, the complete data processing pipeline, including the used parameters is available for download. This enables the user to perform the same analysis locally, or try small modifications (for bioinformaticians). Here we give a quick list of the used programs and their function within our current pipeline. A detailed explanation including the used parameters is available in Supplementary file 2.

In order of use in GeneTiles:

1. Bowtie2 [18] is used to align the reads in the fastq file to the genome (obtained from Ensembl). Bowtie2 generates SAM files that contain the reads together with the location on the genome. Upon multiple hits, the best quality hit is selected, or upon a tie of multiple best hits, the reads are randomly distributed (the manual of Bowtie2 is referred to for other default behaviour).

2. Samtools [19] is used to convert and compress the SAM files into a binary BAM file.

3. Samtools is furthermore used to sort the reads in the BAM files based on the aligned read location in the genome, resulting in a sorted BAM file.

4. The BAM files is indexed to be able to quickly find the aligned reads based on a location in the genome, i.e. to be able to quickly search the BAM file. The index is saved as a BAI file.

5. Using the available annotation from Ensembl we can search the BAM file for reads

within a gene. All reads that at least partially fall within the gene exon and intron regions are counted once. This is done with a python script which consists of the combination of HTseq and pysam [20]. The output of this script is a tab separated file (tsv) containing the read counts per gene.

6. We used DESeq, an R-script to perform statistical analysis. DESeq is used to normalise the reads using a DESeq scaling factor, computed as the median of the ratio, for each gene, of its read count over the geometric mean across samples. Then variance and average of the measurement compared to the control is expressed as a *P*-value, by calculating the dispersion per gene using DESeq. The size factors as well as the *P*-values are stored in 'tsv' files.

7. Using scripts, similarly as in step 5, also the input files for DEXSeq can be generated. DEXseq requires:

- A 'gtf' file containing the experiment design.
- For all samples a 'txt' file containing the counts obtained for the mapping data in the .sam files.
- Genome annotation (from Ensembl) in a 'gff' file.

Scripts to obtain these files are also available in the supplement.

8. Using DEXSeq, another R-script to perform more complex statistical analysis, we can look at the reads within exons, and compare the variance and average per exon between measurement and control groups of samples. DEXSeq uses binning, where exons are cut into bins, based on known exon boundaries. When a read overlaps multiple bins, it is counted in each bin. Per bin, based on the annotation, two comparisons can be made, a comparison between the same exon bins in different samples (groups), and a comparison between an exon-bin and its neighbour exon-bins within the same group of samples. Note that therefore DEXSeq requires at least two groups containing at least two samples. Based on both comparisons a likelihood test is performed resulting in a *P*-value. More details are available in Anders et al. 2012 [21]. The output of size factors and *P*-values are stored as 'tsv' files.

9. Using a script, all tsv files are combined into an excel file, available for download, e.g. per experiment, chromosome, per filtered results of most significant reads or highest ratio between measurement and control. In addition an index is built for fast visualization online (closed source).

Through the website of wikipathways [http://wikipathways.org/index.php/Download_Pathways] the SVG images were downloaded on the GeneTiles server. Using Javascript, on the client side, within the SVG images the gene-boxes are given a background colour

based on a user selection, e.g. *P*-value or ratio. For this, the genes and or proteins are matched to their Ensembl references on the selected genome. In addition also the human pathways are searched for find homologs of genes using Ensembl biomart. Using these homologs, predictions of homolog pathways can be accessed, this enables to search a larger set of pathways contained in the human section of wikipathways. It should be noted that using human pathways to find information about zebrafish biology should be treated with caution and can only lead to suggestions for further investigation.

Results and Discussion

Software design

RNAseq data, containing tens of millions of reads, is mostly processed using scripts. After processing, a selection of reads is analysed using RNAseq viewers. Directly browsing processed RNAseq data is difficult due to the large dynamic range of length scales of reads (50 bp), exons (~200 bp), introns (~3 kb), genes (~20 kb), and chromosomes (~65 Mb). Using a minimal size of one pixel per read, a computer screen allows only for ~50 kb to be visible. In addition, most RNAseq viewers show introns at the same scale as exons, which in most experiments means that 90 percent of the visible sequence data does not display aligned reads. We created an online viewer, GeneTiles (www.Genetiles.com), that does allow for browsing all the aligned reads, while eliminating almost completely the need for user intervention (such as zooming in). The genes in a chromosome are visible as tiles in a 2D array. The tile colour and intensity are a measure of the significance of the number of reads of experiment versus control, indicating changes in expression levels. When a tile is selected, the gene is loaded underneath, scaled to fit the width of the screen. In a schematic view all introns are shrunk to a fixed short length to visualize the aligned reads in a graph above the exons. To accomplish fast browsing, all reads are indexed on the server directly after data processing. This indexed data is also available for download to apply custom filtering in Excel or other programs. The export functions of the tiles and genes as scalable vector graphics makes it easy for the user to modify the final visualization for publication.

Workflow

Automated analysis of RNAseq data using GeneTiles does not need any programming steps anymore in a Linux environment by the user and performs directly a visualization of the differential expressed data, making it easier to interpret. To validate this new software package, RNAseq data from zebrafish bacterial infection experiments was obtained from Veneman et al. (2013) and used as the initial test model. All programs used by Veneman et al. (2013) are implemented and more visualization and export options are added in a server based environment (Figure 1, Supplementary file 1). Therefore the analysis pipeline of GeneTiles represents a combination of previously

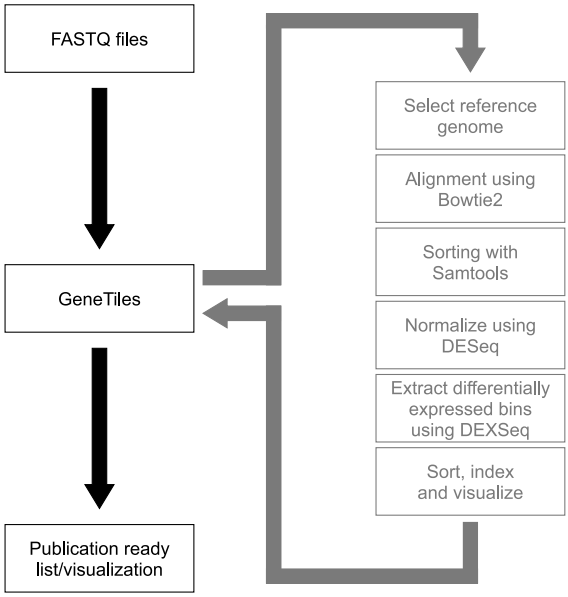


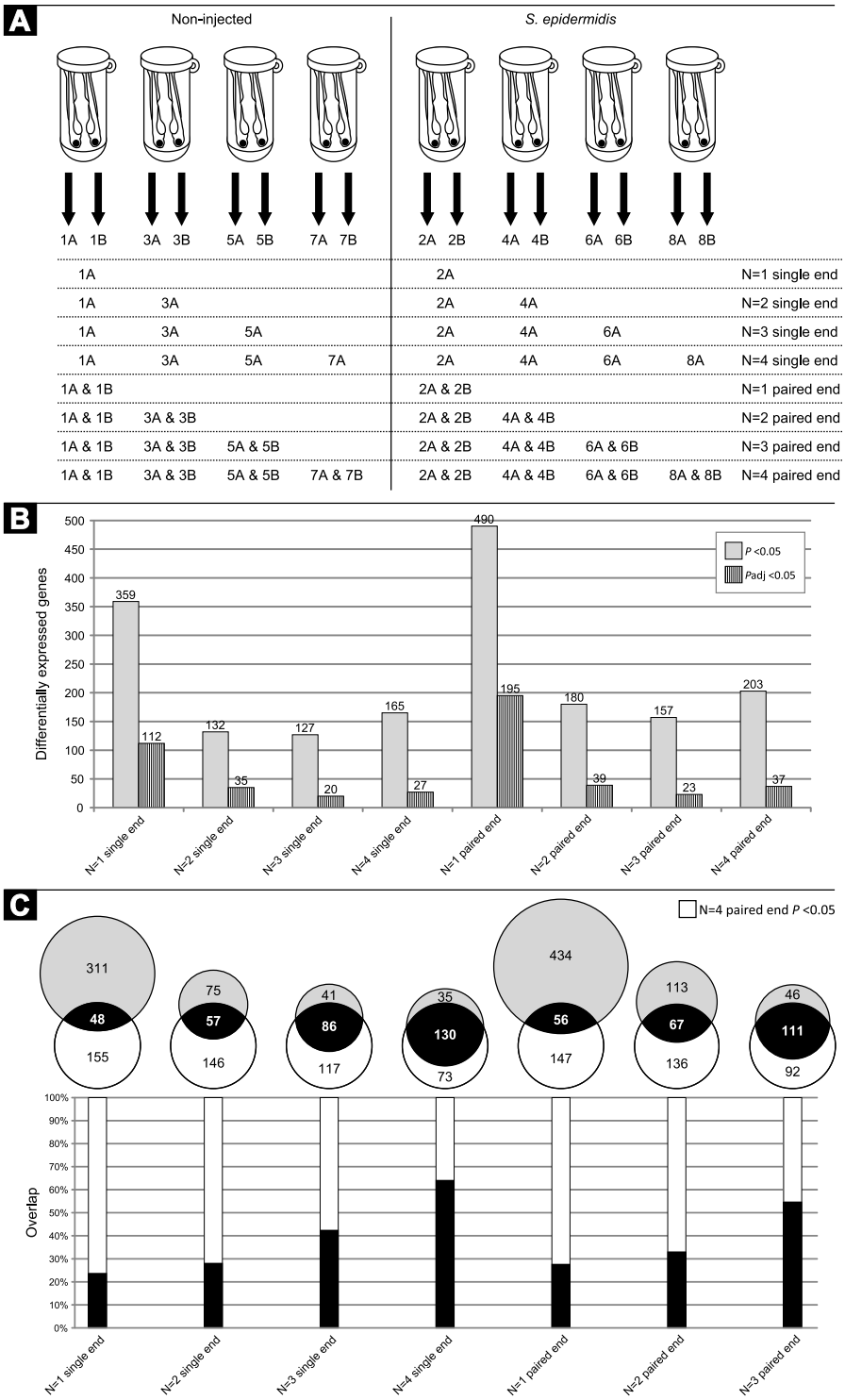
Figure 1: Pipeline of RNAseq data analysis. The diagram shows the workflow of the data analysis starting at the raw fastq files until the final visualization performed automatically by the GeneTiles server. The analysis pipeline is open source, and is available in the supplement.

described tools that have been previously shown to be useful for RNAseq analyses [22]. This makes it very manageable because it reduces the amount of high-end computers required in the research group for alignment and analysis, as all calculations are performed on the server. To start the analysis, the user can choose between various genomes that have been imported from ENSEMBL. Subsequently, the fastq files will be uploaded, followed by the option to analyse the data as single-end or paired-end. The files will be aligned automatically, after which the control or measurement treatment can be chosen. DESeq [23] will normalize the data, and subsequently DEXSeq [21] will extract differentially expressed bins that indicate differential splicing. A table containing the differentially expressed genes or visualizations containing tiles or individual genes can be exported at this point.

With respect to the use of Bowtie2 aligner, we want to point out that it will fail to map reads spanning exon-exon boundaries to the genome. This problem could be solved using a splice aware aligner based on Bowtie2, such as Tophat2 [24]. This option will be included in a future version of the GeneTiles package, However Tophat2 is more computationally intensive, and depends in the correct predictions of splice sites. Therefore the analysis without the splice aware aligner as used in this paper will remain present.

Different analysis methods

Considering that the RNA samples of the *S. epidermidis* infection experiments were paired-end sequenced we had the possibility to explore the added value of paired-end over single-end sequencing. We compared the outcome of the differential expression of these 2 methods as well as the difference in sample sizes as shown in Figure 2A. It can



(see legend on next page)

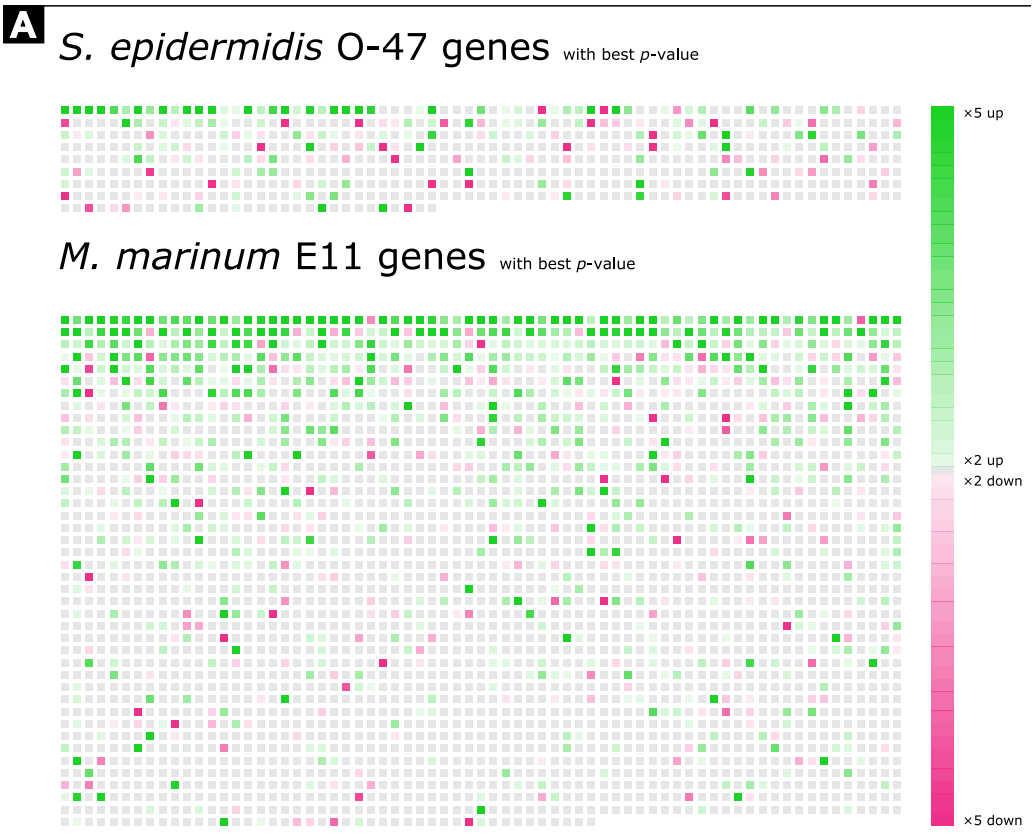
(see Figure on previous page)
Figure 2: Comparing single- and paired-end RNAseq analysis. (**A**) The different samples sizes (n=1-4) for analysis are visualized. (**B**) The total number of differentially expressed genes with a fold change larger than 2 or smaller than -2 and a *P*-value smaller than 0.05 (solid grey bars) or an adjusted *P*-value smaller than 0.05 (patterned dark bars). (**C**) The light grey circles of the Venn diagrams show the total number of differentially expressed genes for all the samples sizes with a fold change larger than 2 or smaller than -2 and a *P*-value smaller than 0.05 as shown in **B**, and the overlap of differentially expressed genes compared to the n=4 paired-end data set. The bar graphs show the overlap in percentage of these different sample sizes compared to the n=4 paired-end data set.

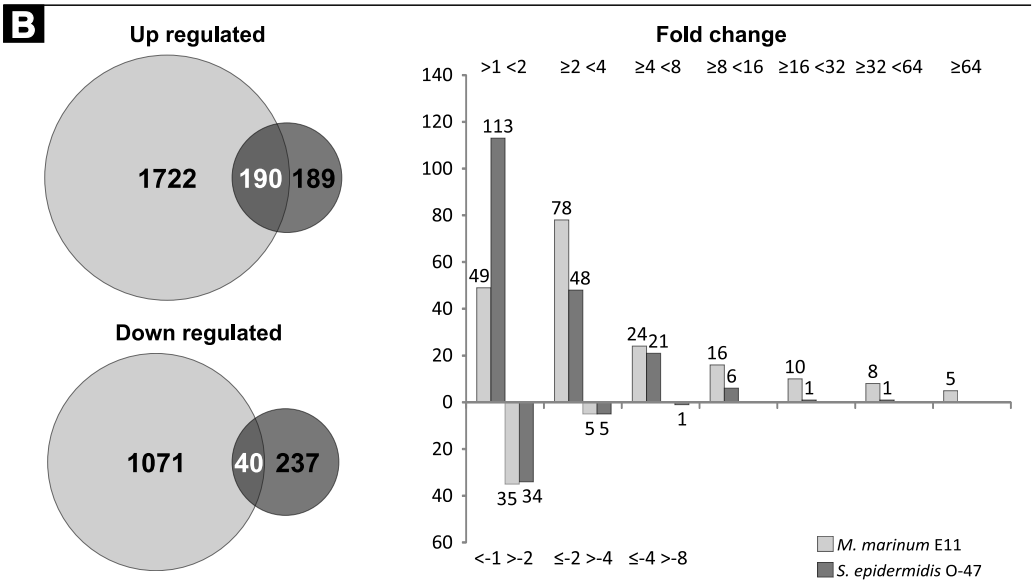
be noted that a the number of differentially expressed genes does not give an estimate of the reliability of the data, however, considering the high quality of RNAseq data it can be assumed that adding an extra biological sample provides more relevant information than analysing a smaller group of samples by paired-end sequencing. Our data support this assumption since we found only a slight increase of 23% in differentially expressed genes when performing paired-end analysis in the 4 *S. epidermidis* infected samples, using a *P*-value of 0.05 as cut-off filter.

Secondly we analysed the number of differentially expressed genes with all possible options regarding samples sizes (Figure 2B). We found a set of 359 differentially expressed genes when only analysing 1 sample with a *P*-value of 0.05 and compared this large set of genes to our reference set of 203 genes (resulting from the analysis of paired-end sequence data of all four samples) (Figure 2C. This comparison shows a rather small overlap of only 23.7% (single-end) and 27.6% (paired-end). As expected, this overlap increases and therefore the number of false-positives decreases when adding more samples (Figure 2C). In order to provide a statistically more stringent analysis of differentially expressed genes we have also included in the GeneTiles software a tool for minimizing false discovery based on the algorithm of Benjamini et al. 1995 [25], using the implementation of DESeq. The resulting adjusted values show far more stringent results (Figure 2B) but generally confirm the limited value of performing paired-end sequencing as compared to the added value of adding more biological controls.

Comparison *S. epidermidis* versus *M. marinum* infection in zebrafish embryos

We compared the different transcriptome host responses of zebrafish embryos upon *S. epidermidis* or *M. marinum* yolk injection. The previous comparison as shown by Veneman et al. (2013) was based on a single biological replicate of *M. marinum* infected zebrafish embryos. We used this single replicate and added 5 more independent biological replicas which led to a total of 6 replicas of *M. marinum*, 4 replicas of *S. epidermidis* infected zebrafish embryos and a total of 9 replicas of non-infected control samples. As found before [1], *S. epidermidis* infection elicits a much smaller transcriptional host response of immune related genes compared to *M. marinum* (Figure 3). However, the total number of differentially expressed genes was increased since several genes that were previously not found to be significantly regulated by *S. epidermidis* where they do show a response in this analysis. Another finding is the high induction of genes upon *M. marinum* infection compared to the *S. epidermidis*





(see legend on next page)

(see Figure on previous page)
Figure 3: An overview of differentially expressed genes. (**A**) The GeneTiles output shows a much larger set of genes with a *P*-value smaller than 0.05 with the *M. marinum* infected samples compared to the *S. epidermidis* infected samples. Each tile visualizes one gene, sorted on *P*-value. The colour and intensity are a function of the ratio between measurement and control samples. (**B**) Comparing the data from **3A** is shown in the Venn diagrams on the left and the overlaps in white digits are shown in a quantitative manner in the bar graph on the right.

infected samples (Figure 3B). An explanation could be that the *M. marinum* bacterium is a natural fish pathogen, and therefore is better recognized as intruder. *S. epidermidis* is in large quantities also pathogenic for fish, however, since it is not a natural pathogen it could very well be that it does not get recognized as well as *M. marinum*. For instance, we now observe a high induction of the leptin b gene (*lepb*) upon infection with *S. epidermidis* (Figure 4). In this case the difference with the previous study is caused by errors in the automated annotation of probes in the micro-array used in Veneman et al. (2013) which was used as bench mark for the RNAseq analysis. The high expression of *lepb* found in *M. marinum* infected samples is in line with the results earlier described by Wieland et al. 2005 [26], where they found a higher mycobacterial load in the lungs of leptin-deficient ob/ob mice.

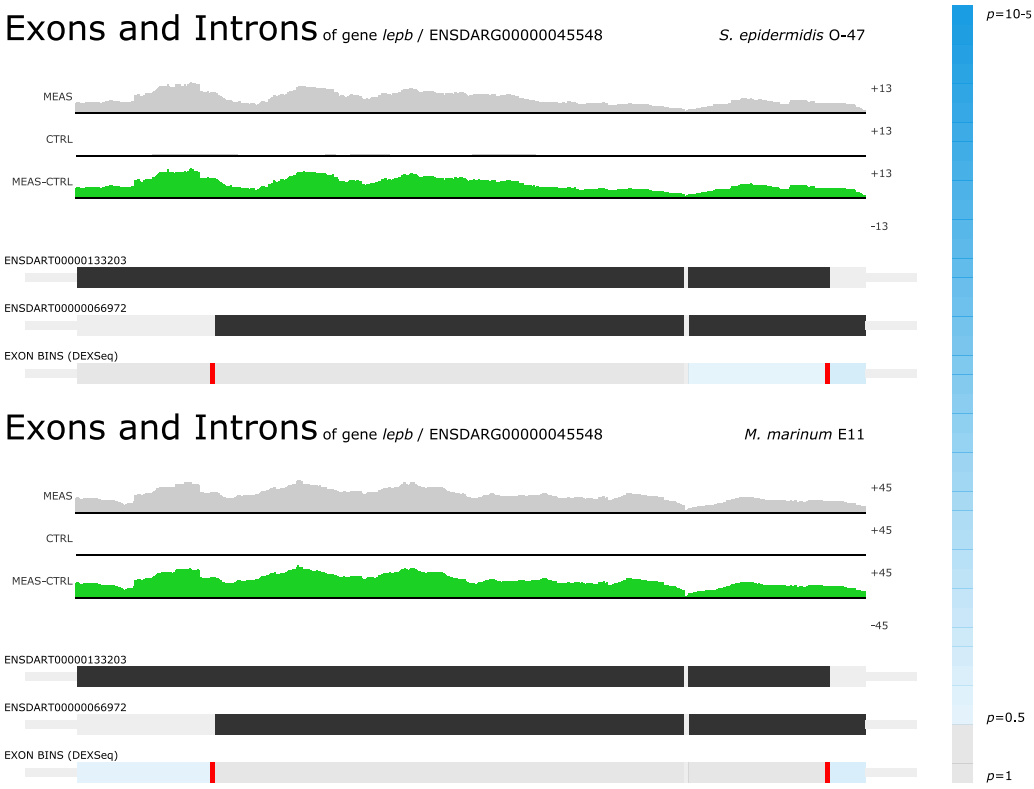
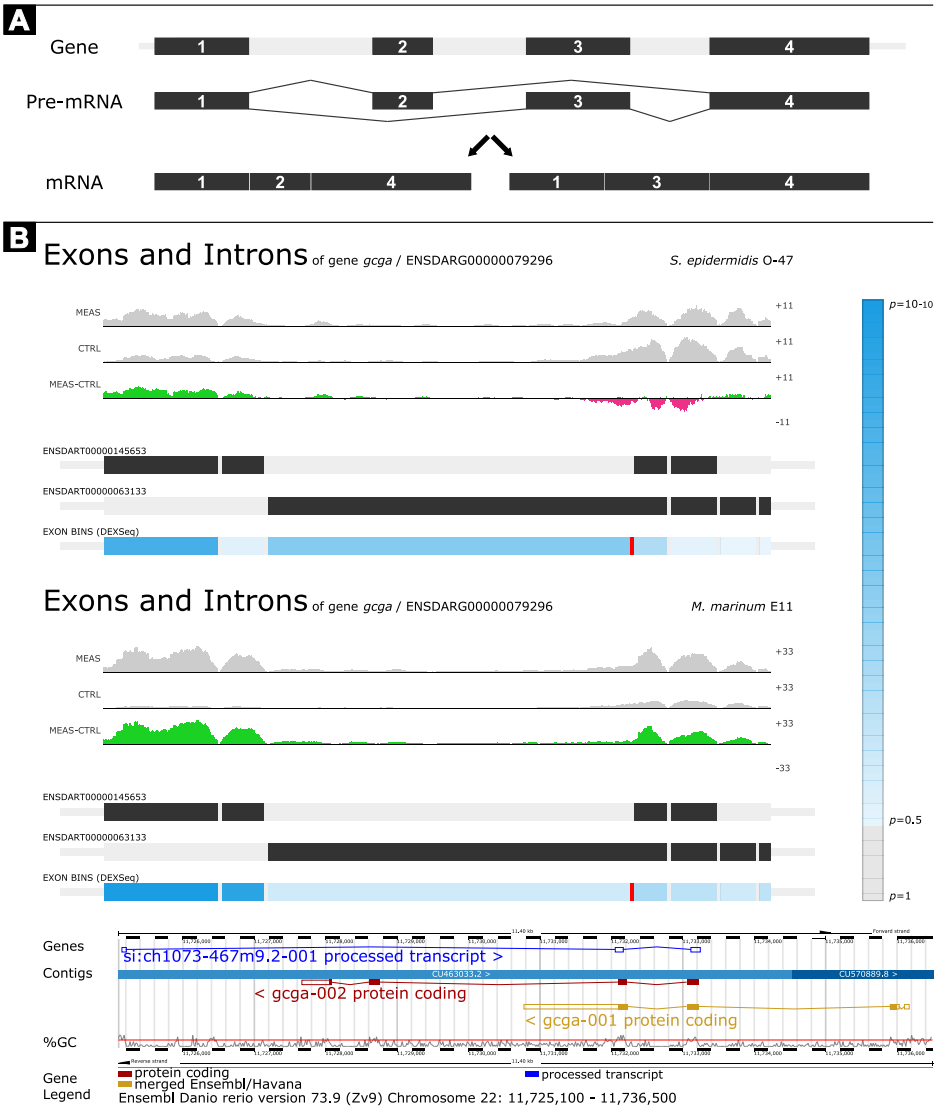


Figure 4: *Lepb* as highest induced gene. For both *S. epidermidis* O-47 (fc: 36, *P*-value: 4.99×10^{-6}) and *M. marinum* E11 (fc: 148, *P*-value: 1.54×10^{-10}) infected samples *lepb* was found to be the highest induced gene.

Differential splicing

Another feature in the GeneTiles software is the integration of the DEXSeq analysis tool [21], which allows searching for genes that are differentially spliced. The analysis strategy of differential splicing is schematically shown in Figure 5A, with an example that shows that 1 of 4 exons is spliced out from a pre-mRNA to form the mature mRNA. With both the *S. epidermidis* and *M. marinum* infection data set we found glucagon a (*gcga*) as top candidate to be differentially spliced with large enrichment of two 5' exons (Figure 5B), which are indicated by the dark blue bars underneath the representing exons. Supporting this finding, the pro-glucagon gene has been described before as being differentially spliced into multiple peptides in teleost fish [27].



(see legend on next page)

(See Figure on previous page)
Figure 5: Finding differentially spliced genes. (**A**) Schematic view of the principal of differential splicing, where two different mRNAs are formed from one gene. (**B**) For both the *S. epidermidis* (fc: 1.17, *P*-value $4.62x10^{-1}$) and *M. marinum* (fc: 4.70, *P*-value $2.06x10^{-10}$) infected samples glucagon a (*gcga*) was found to be differentially spliced as shown by the dark blue bar under the left exon. The screenshot of the GeneTiles visualization demonstrates a schematic view, where introns are compressed to allow for more space for visualization of reads on exons. The gradient-blue bar on the right indicates the *P*-value predicting differential splicing. The bottom image is a gene representation from ENSEMBL.

We also demonstrate that GeneTiles can quickly point out false negative results based on DEXSeq as a result of ambiguous ENSEMBL annotations. For instance in our analysis of infection markers granulin antisense (*grnas*) appeared as a candidate for differential splicing. However, the actual differential expression found of grnas occurs from a fusion of two genes, granulin 1 (*grn1*) and granulin 2 (*grn2*), which are located at this same position as shown in Figure 6.

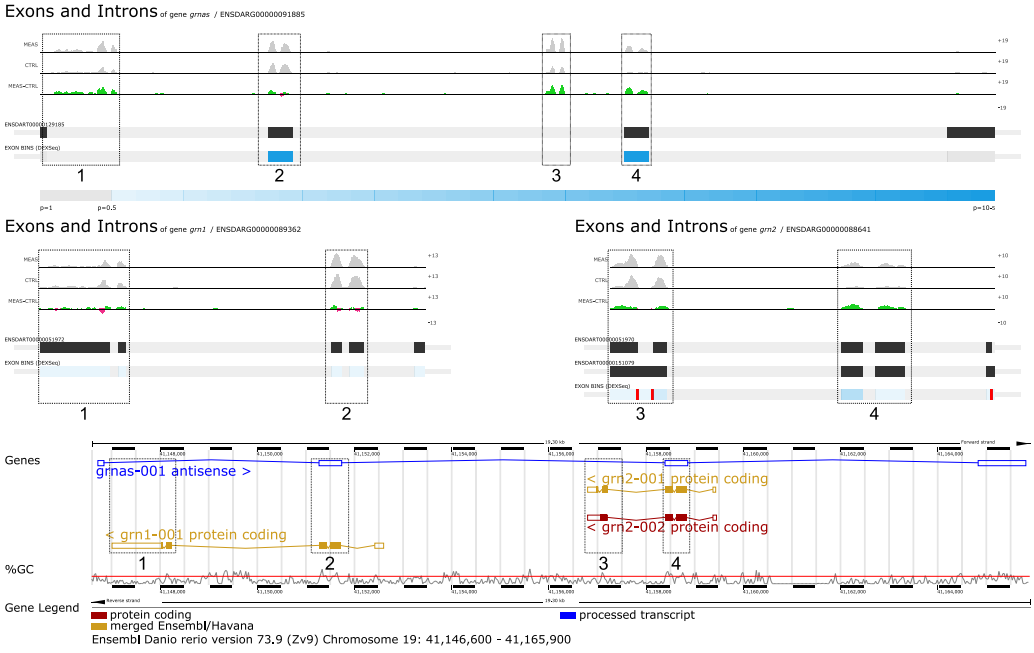
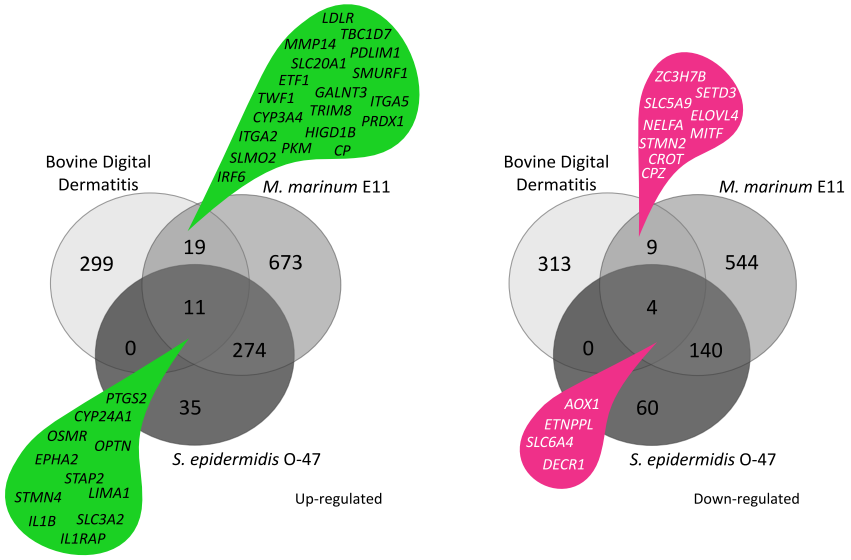


Figure 6: False discovery. The differential splicing indicated by the dark blue bars at exon 2 and 4 from grnas proved to be incorrect. The differential expression found indicated by the 4 boxes at *grnas* (fc: 1.87, *P*-value $3.05x10^{-3}$) derived from *grn1* (fc: 1.57, *P*-value $1.80x10^{-2}$) and *grn2* (fc: 3.33, *P*-value $2.23x10^{-3}$). The screenshot of the GeneTiles visualization demonstrates a non-schematic view, where introns are not compressed showing the actual length of the introns and exons. The gradient-blue bar indicates the *P*-value predicting differential splicing. The bottom image is a gene representation from ENSEMBL.

Comparing different host infection models

To date, we have only found 1 publication in PubMed describing RNAseq analysis of the host after a bacterial infection *in vivo* other than in zebrafish [15]. The data of Scholey et al. 2013 [15], describing bovine digital dermatitis (BDD, an infectious foot disease) was used to compare the host response in the cow and zebrafish. The raw data with accession number GSE41732 was obtained from the GEO database and analysed with the GeneTiles software package. Comparing the differential expression data from Scholey et al. 2013 [15], we could not validate all transcripts, since 23% of the transcripts are retired and are not available anymore, due to the updated version of the *Bos taurus* 4.0 ENSEMBL annotation to the COW UMD3.1 ENSEMBL annotation. All other transcripts could be validated.



	Cluster #	Enrichment score cluster	Ontology	GO #	Name	# Genes	*P*-value
Upregulated	1	3.2	Biological Process	GO:0007584	Response to nutrient	5	1.00E-04
	1	3.2	Biological Process	GO:0031667	Response to nutrient levels	5	3.70E-04
	1	3.2	Biological Process	GO:0009991	Response to extracellular stimulus	5	5.60E-04
	2	2.82	Biological Process	GO:0050896	Response to stimulus	13	1.00E-02
	4	2.5	Biological Process	GO:0006979	Response to oxidative stress	4	2.90E-03
	6	2.41	Biological Process	GO:0009719	Response to endogenous stimulus	5	5.20E-03
	6	2.41	Biological Process	GO:0010033	Response to organic substance	6	7.80E-03
	14	1.7	Biological Process	GO:0050727	Regulation of inflammatory response	3	7.90E-03
	19	1.62	Biological Process	GO:0080134	Regulation of response to stress	5	1.30E-03
	21	1.59	Biological Process	GO:0001666	Response to hypoxia	3	2.30E-02
	29	1.29	Biological Process	GO:0048522	Positive regulation of cellular process	8	3.70E-02
	34	1.15	Biological Process	GO:0009611	Response to wounding	4	6.60E-02
	55	0.7	Biological Process	GO:0050794	Regulation of cellular process	18	3.80E-02
Downregulated	1	2.13	Biological Process	GO:0019752	Carboxylic acid metabolic process	4	6.30E-03
	1	2.13	Biological Process	GO:0043436	Oxoacid metabolic process	4	7.20E-03
	2	1.94	Molecular Function	GO:0048037	Cofactor binding	3	1.10E-02
	3	1.56	Biological Process	GO:0006631	Fatty acid metabolic process	3	7.20E-03
	3	1.56	Biological Process	GO:0044255	Cellular lipid metabolic process	3	5.60E-02
	4	0.59	Cellular Component	GO:0005622	Intracellular	11	2.30E-01

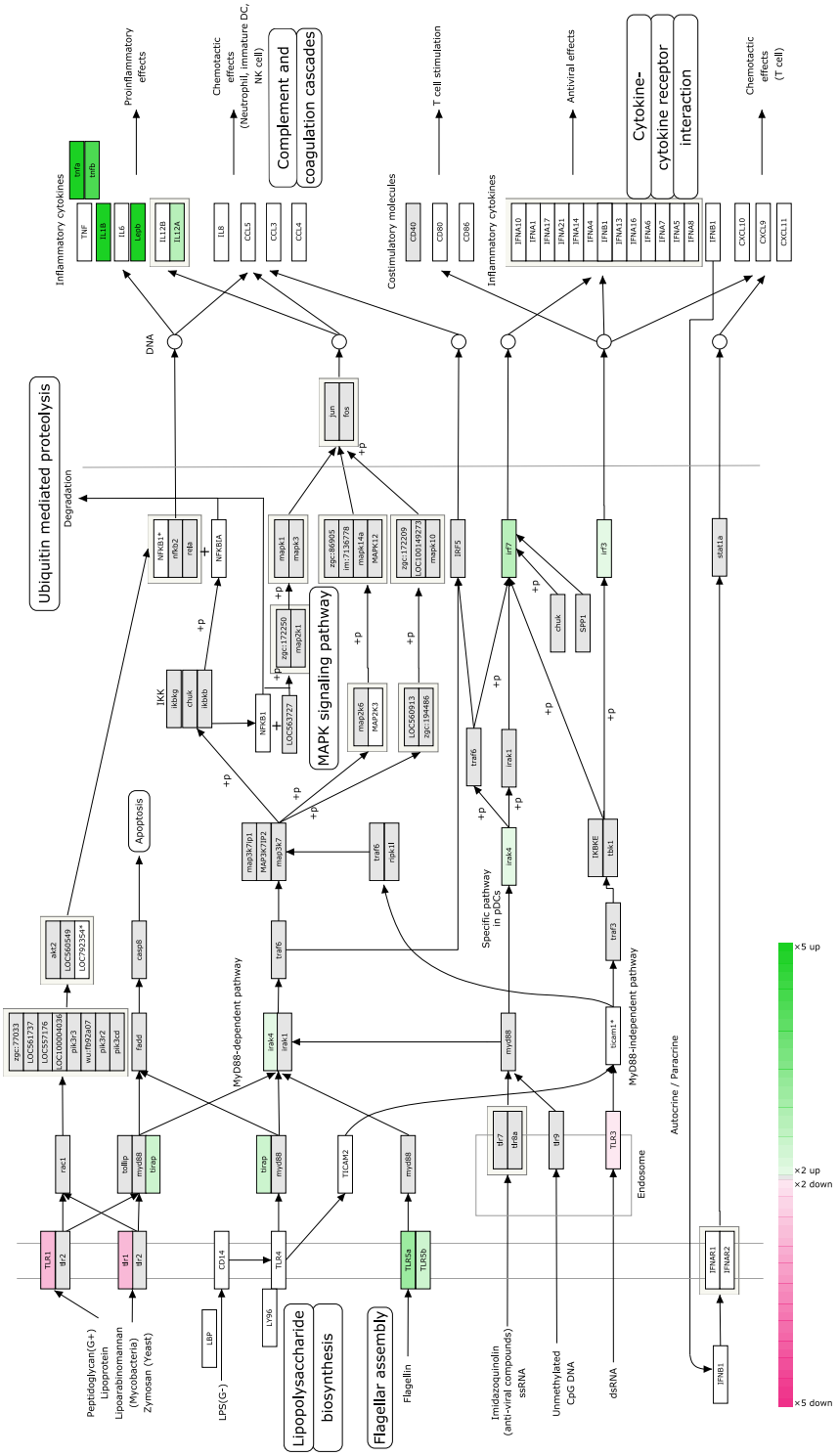
(see legend on next page)

(See Figure on previous page)
Figure 7: Overlap with other host pathogen RNAseq experiments. The Venn diagrams show the number of human orthologs of differentially expressed (fc: >2 or <-2 and *P*-value <0.05) genes from the Bovine digital dermatitis, *M. marinum* E11 and *S. epidermidis* O-47 infection data. The gene ontology analysis [28, 29] is based on the overlapping groups indicated by the green and magenta drops.

Both zebrafish and bovine gene identifiers were linked to the human orthologs using the ENSEMBL database. The results show that 61% of the zebrafish genes and 95% of the bovine genes could be translated to human orthologs; a comparison of the differentially expressed gene sets in both disease models (*P*-value < 0.05) is shown in the Venn diagrams in Figure 7. Gene ontology (GO) [28, 29] analysis on the overlapping set of differentially expressed genes between the cow and zebrafish showed that the differentially expressed genes are categorized in multiple response processes (Figure 7). The group of up-regulated overlapping genes between BDD, *M. marinum* and *S. epidermidis* infection includes the following genes; prostaglandin-endoperoxide synthase (*PTGS2*), cytochrome P450 family 24 subfamily A polypeptide 1 (*CYP24A1*), oncostatin M receptor (*OSMR*), Optineurin (*OPTN*), EPH receptor A2 (*EPHA2*), signal transducing adaptor family member 2 (*STAP2*), stathmin-like 4 (*STMN4*), LIM domain and actin binding 1 (*LIMA1*), interleukin 1, beta (*IL1b*), solute carrier family 3 (amino acid transporter heavy chain), member 2 (*SLC3A2*) and interleukin 1 receptor accessory protein (*IL1RAP*). As expected, most of these genes are related to the immune system such as *OPTN* which can activate Fas-ligand pathways to induce apoptosis or anti-inflammatory responses [30], and *IL1B* that is a well-known cytokine produced by activated macrophages, which then can indirectly activate *PTGS2* that also is significantly expressed [31]. The *IL1RAP* is essential for signal transduction of *IL1* in order to induce proinflammatory proteins upon infection [32].

Pathway analysis

The visualization is not only limited to the coloured tiles as described above, but can also be used for functional analysis using WikiPathways [33]. With 96 zebrafish and 267 human pathways at the moment implemented in the software package this allows the user a fast overview of differential expression in biological networks. An example is given in Figure 8, where the Toll-like receptor signalling pathway is showing the differential expression data of *M. marinum* infected embryos using the pathway we submitted to WikiPathways that has been accepted in the curated collection.



(see legend on next page)

(See Figure on previous page)
Figure 8: Toll-like receptor pathway showing *M. marium* expression data. RNAseq expression data is shown of *M. marinum* E11 infected zebrafish at 5 dpi. The green and magenta boxes show differential expression respectively up- and down-regulation (fc >2 or <-2) with a *P*-value smaller than 0.05. The genes with an asterisk are discontinued in current databases, the white boxes could not be identified and the grey boxes did not meet the expression criteria.

Lepb as highest induced gene

As mentioned earlier, we found *lepb* as highest induced gene upon infection with *S. epidermidis* and *M. marinum*. Although leptin has been studied over decades for its function in controlling fat balance via brain signalling, its function in innate immune responses has only recently been discovered [34, 35]. For instance, it was shown to be an important factor in mucosal immunity; however, little is known about the underlying mechanisms. For that reason, we designed a morpholino to knock down the function of *lepb* in embryos. At a concentration of 0.5mM we did not found any observable phenotype in development. Combining the *lepb* knock down with the injection of *M. marinum* E11 suggested that loss of *lepb* leads to an increase of the bacterial burden (Figure 9), indicating that *lepb* could have a role during infection. Since leptin has been described to be involved in metabolic processes and the wasting syndrome [36], we performed liquid chromatography mass spectrometry (LC-MS)on 5 day old *lepb* morphants versus wild type control embryos, each with and without *M. marinum* infection. We used a reference set of wasting syndrome associated metabolites (R. Marín-Juez, unpublished results) and found 7 metabolites not influenced by the *M. marinum* infection in the *lepb* knockdown embryos (Figure 10). This indicates that there could be a cross regulation between the nutritional status and immune response against *M. marinum*. Based on these results it could very well be that weight loss due to tuberculosis in humans could be linked to the high levels of *lepb*, as a result of the infection as illustrated in Figure 11. This might be the result of the role of leptin in suppressing appetite or a hitherto undiscovered other function of leptin in metabolic control. However, there are some

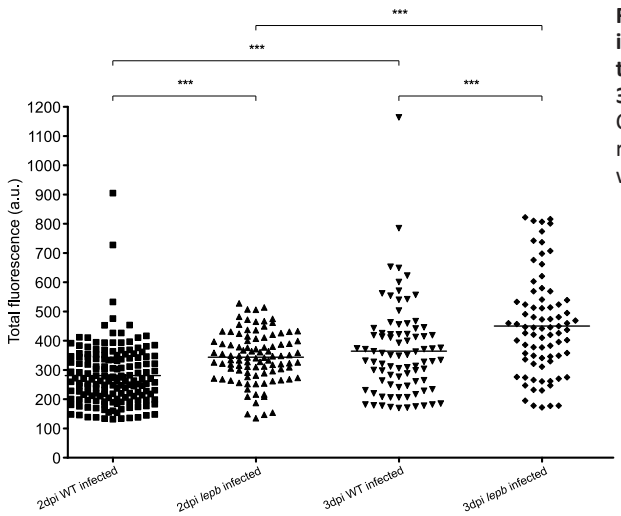


Figure 9: The *lepb* morphants show an increased bacterial burden compared to the wild type infected embryos at 2 and 3 dpi. Analysis was performed using the COPAS XL flow cytometer. The asterisk represent the *P*-value obtained by a one way anova; ****P*<0.001.

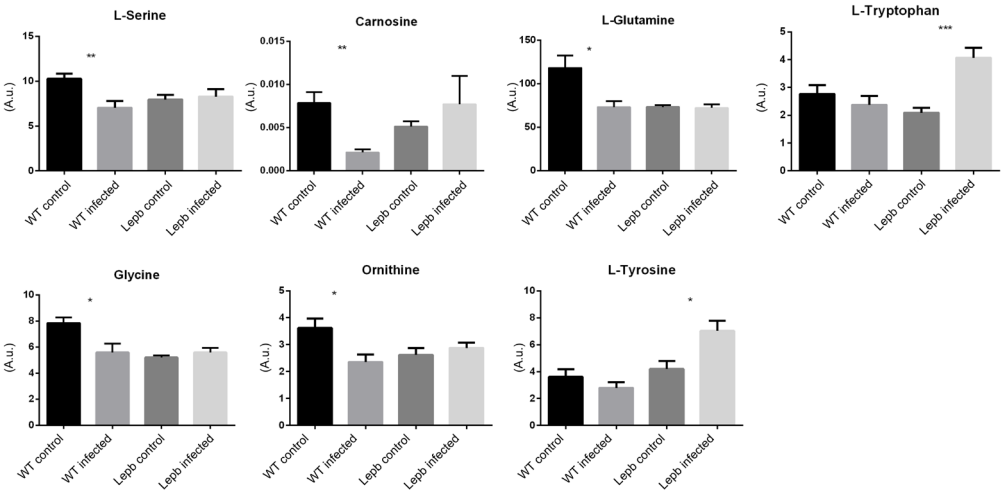


Figure 10: Bar graphs showing 7 wasting metabolites differentially changes in *lepb* morphants with or without *M. marinum* infection. The asterisk represent *P*-values obtained by a student T-test; **P*<0.05, ***P*<0.01, ****P*<0.001 and *****P*<0.0001. n=8

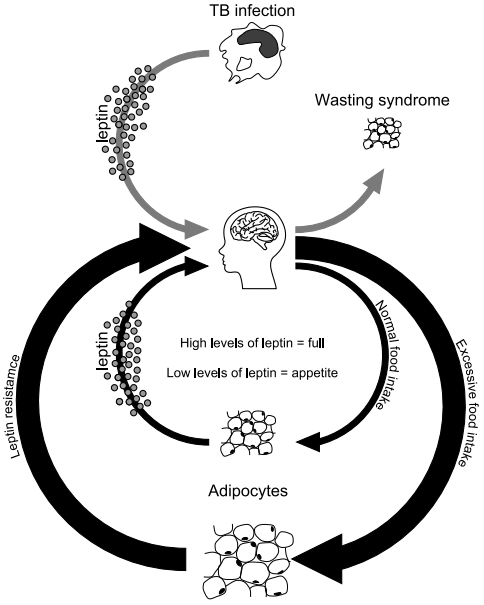


Figure 11: Model of the potential link between leptin production and TB progression. Under normal circumstances leptin regulates appetite and prevents overnutrition. Due to overnutrition leptin resistance can occur, which then can lead to uncontrolled appetite. However, leptin levels are also increased upon TB infection, which in turn can lead to wasting.

papers describing that leptin is not the missing link between the immune response against TB and weight loss [37-39]. For example Schwenk et al. (2003) used in their study heparinized blood samples from humans with pulmonary TB and did not find a correlation of high leptin levels and progression of TB. However, plasma concentrations may not always reflect the biological activity of compounds such as leptin and therefore do not represent the whole systemic situation. This means that our approach of using a whole model organism could still indicate new networks between the function of leptin in the context of a TB infection that are also relevant to the human situation.

Conclusions

The described toolbox for RNAseq data analysis offers two different levels of support in an integrative setting. First, the software combines several programs needed for open source RNAseq analysis such as Bowtie2, Samtools, the 'R' statistical package, DESeq, DEXSeq, HTSeq and pysam. These programs are placed in a pipeline (script) that runs these programs in the required order, with correct in- and output settings. Secondly, the processed data is visualized in a user friendly way and made available for export with a choice of quantitative settings.

The advantages and ease of use of this combined toolbox is demonstrated by analysis of previously published RNAseq datasets from zebrafish and cow infectious disease models, as well as new RNAseq data of a zebrafish mycobacterial infection experiments. This resulted in a highly confident innate marker set for systemic innate immune response to infection by pathogenic and non-pathogenic bacterial species in zebrafish. Furthermore, the data is viewable in a pathway view using the pathways stored at WikiPathways. In this way it was also possible to quickly determine the effect of the number of replicates and the evaluation of potential false positive results, as is the case for the analysis of differential splicing using the DEXSeq algorithm. Comparing our experiences with our previous analyses [1] and the re-analyses performed here we can estimate that we have saved several months of working time while obtaining far superior output files that could be rapidly compared to new RNAseq data sets also from other organisms. The data analysed in this study is available at the GeneTiles website for further analysis and as demonstration material. This makes it possible to rapidly evaluate new immune markers in the datasets described in this paper but also can be used to identify new markers based on other search criteria such as the discovery of the *lepb* induction after infection, which is currently under further investigation.

Availability and requirements

The analysis is open source and available for download, as well as offered as supplementary material. All visualization images are also available for download on the demo page of http://www.genetiles.com. The analysis pipeline, including the source and a complete script to run the same analysis locally is available for download. It is offered together with the open demo for everyone, and it is possible to apply small changes locally to change the analysis or to change the input files.

Acknowledgements

We thank Leonie de Boer and Bas Zaat (Academic Medical Center) for providing us with the *S. epidermidis* O-47 strain containing a pWVW189 derived mCherry expression vector, Davy de Witt, Ulrike Nehrdich, and Laura van Hulst for fish caretaking, other

colleagues from Leiden University for helpful discussions, and Hans Jansen and Ron Dirks (ZF-screens BV) for RNAseq service. This research forms part of the Project P5.03 IBIZA of the research program of the BioMedical Materials institute, co-funded by the Dutch Ministry of Economic Affairs, and of the Smart Mix Program (NWOA_6QY9BM) of The Netherlands Ministry of Economic Affairs and of The Netherlands Ministry of Education, Culture and Science. Additional support was obtained from the EU project ZF-Health (FP7-Health-2009-242048). The authors further acknowledge financial support from the Leiden University Fund (LUF) for robotics and from Cyttron, in the Besluit Subsidies Investeringen Kennisinfrastructuur program, which in turn is financially supported by the Netherlands Organization for Scientific Research for imaging facilities.

Supplementary material

Supplementary material can be found on:

http://link.springer.com/article/10.1007%2Fs00251-014-0820-3

References

1. Veneman WJ, Stockhammer OW, de Boer L, Zaat SA, Meijer AH, Spaink HP: A zebrafish high throughput screening system used for Staphylococcus epidermidis infection marker discovery. BMC Genomics 2013, 14(1):255.

2. van der Vaart M, van Soest JJ, Spaink HP, Meijer AH: Functional analysis of a zebrafish myd88 mutant identifies key transcriptional components of the innate immune system. Dis Model Mech 2013, 6(3):841-854.

3. van Soest JJ, Stockhammer OW, Ordas A, Bloemberg GV, Spaink HP, Meijer AH: Comparison of static immersion and intravenous injection systems for exposure of zebrafish embryos to the natural pathogen Edwardsiella tarda. BMC Immunol 2011, 12:58.

4. Ordas A, Hegedus Z, Henkel CV, Stockhammer OW, Butler D, Jansen HJ, Racz P, Mink M, Spaink HP, Meijer AH: Deep sequencing of the innate immune transcriptomic response of zebrafish embryos to Salmonella infection. Fish Shellfish Immunol 2011, 31(5):716-724.

5. Stockhammer OW, Rauwerda H, Wittink FR, Breit TM, Meijer AH, Spaink HP: Transcriptome analysis of Traf6 function in the innate immune response of zebrafish embryos. Mol Immunol 2010, 48(1-3):179-190.

6. Broekhuizen CA, Schultz MJ, van der Wal AC, Boszhard L, de Boer L, Vandenbroucke-Grauls CM, Zaat SA: Tissue around catheters is a niche for bacteria associated with medical device infection. Crit Care Med 2008, 36(8):2395-2402.

7. Boelens JJ, Dankert J, Murk JL, Weening JJ, van der Poll T, Dingemans KP, Koole L, Laman JD, Zaat SA: Biomaterial-associated persistence of Staphylococcus epidermidis in pericatheter macrophages. J Infect Dis 2000, 181(4):1337-1349.

8. Busscher HJ, van der Mei HC, Subbiahdoss G, Jutte PC, van den Dungen JJ, Zaat SA, Schultz MJ, Grainger DW: Biomaterial-associated infection: locating the finish line in the race for the surface. Sci Transl Med 2012, 4(153):153rv110.

9. Zaat S, Broekhuizen C, Riool M: Host tissue as a niche for biomaterial-associated infection. Future Microbiol 2010, 5(8):1149-1151.

10. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. Nat Biotechnol 2011, 29(1):24-26.

11. Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S: MapView: visualization of short reads alignment on a desktop computer. Bioinformatics 2009, 25(12):1554-1555.

12. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: Tablet--next generation sequence assembly visualization. Bioinformatics 2010, 26(3):401-402.

13. Laczik M, Tukacs E, Uzonyi B, Domokos B, Doma Z, Kiss M, Horvath A, Batta Z, Maros-Szabo Z, Torok Z: Geno viewer, a SAM/BAM viewer tool. Bioinformation 2012, 8(2):107-109.

14. Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA: BamView: visualizing and interpretation of next-generation sequencing read alignments. Brief Bioinform 2013, 14(2):203-212.

15. Scholey RA, Evans NJ, Blowey RW, Massey JP, Murray RD, Smith RF, Ollier WE, Carter SD: Identifying host pathogenic pathways in bovine digital dermatitis by RNA-Seq analysis. Vet J 2013, 197(3):699-706.

16. Carvalho R, de Sonneville J, Stockhammer OW, Savage ND, Veneman WJ, Ottenhoff TH, Dirks RP, Meijer AH, Spaink HP: A high-throughput screen for tuberculosis progression. PLoS One 2011, 6(2):e16779.

17. Xia J, Psychogios N, Young N, Wishart DS: MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 2009, 37(Web Server issue):W652-660.

18. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods 2012, 9(4):357-359.

19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079.

20. Anders S, Pyl PT, Huber W: HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics 2014.

21. Anders S, Reyes A, Huber W: Detecting differential usage of exons from RNA-seq data. Genome Res 2012, 22(10):2008-2017.

22. Hatem A, Bozdag D, Toland AE, Catalyurek UV: Benchmarking short sequence mapping tools. BMC Bioinformatics 2013, 14:184.

23. Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biol 2010, 11(10):R106.

24. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013, 14(4):R36.

25. Benjamini Y, Hochberg Y: Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met 1995, 57(1):289-300.

26. Wieland CW, Florquin S, Chan ED, Leemans JC, Weijer S, Verbon A, Fantuzzi G, van der Poll T: Pulmonary Mycobacterium tuberculosis infection in leptin-deficient ob/ob mice. Int Immunol 2005, 17(11):1399-1408.

27. Holland LZ, Short S: Alternative splicing in development and function of chordate endocrine systems: a focus on Pax genes. Integr Comp Biol 2010, 50(1):22-34.

28. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009, 4(1):44-57.

29. Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009, 37(1):1-13.

30. Wild P, Farhan H, McEwan DG, Wagner S, Rogov VV, Brady NR, Richter B, Korac J, Waidmann O, Choudhary C et al: Phosphorylation of the autophagy receptor optineurin restricts Salmonella growth. Science 2011, 333(6039):228-233.

31. Lappas M: NOD1 and NOD2 regulate proinflammatory and prolabor mediators in human fetal membranes and myometrium via nuclear factor-kappa B. Biol Reprod 2013, 89(1):14.

32. Subramaniam S, Stansberg C, Cunningham C: The interleukin 1 receptor family. Dev

Comp Immunol 2004, 28(5):415-428.

33. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: WikiPathways: building research communities on biological pathways. Nucleic Acids Res 2012, 40(Database issue):D1301-1307.

34. Paz-Filho G, Mastronardi C, Franco CB, Wang KB, Wong ML, Licinio J: Leptin: molecular mechanisms, systemic pro-inflammatory effects, and clinical implications. Arq Bras Endocrinol Metabol 2012, 56(9):597-607.

35. Matarese G, Moschos S, Mantzoros CS: Leptin in immunology. J Immunol 2005, 174(6):3137-3142.

36. Gaetke LM, Oz HS, de Villiers WJS, Varilek GW, Frederich RC: The leptin defense against wasting is abolished in the IL-2-deficient mouse model of inflammatory bowel disease. J Nutr 2002, 132(5):893-896.

37. Schwenk A, Hodgson L, Rayner CF, Griffin GE, Macallan DC: Leptin and energy metabolism in pulmonary tuberculosis. Am J Clin Nutr 2003, 77(2):392-398.

38. van Crevel R, Karyadi E, Netea MG, Verhoef H, Nelwan RH, West CE, van der Meer JW: Decreased plasma leptin concentrations in tuberculosis patients are associated with wasting and inflammation. J Clin Endocrinol Metab 2002, 87(2):758-763.

39. Santucci N, Diaz A, Bianchi E, Spinelli S, D'Attilio L, Bongiovanni B, Didoli G, Brandan N, Nannini L, Bay ML et al: Leptin does not enhance cell-mediated immune responses following mycobacterial antigen stimulation. Int J Tuberc Lung Dis 2014, 18(8):981-987.

40. Inferring deferential exon usage in RNA-Seq data with the DEXSeq package [http://www.bioconductor.org/packages/2.13/bioc/vignettes/DEXSeq/inst/doc/DEXSeq.pdf]