Deep sequencing of the zebrafish transcriptome response to mycobacterium infection

Hegedűs, Z.; Zakrzewska, A.; Agoston, V.C.; Ordas, A.; Rácz, P.; Mink, M.; ... ; Meijer, A.H.

Molecular Immunology 46 (2009) 2918–2930

Contents lists available at ScienceDirect

Molecular Immunology

journal homepage: www.elsevier.com/locate/molimm

Deep sequencing of the zebrafish transcriptome response to mycobacterium infection

Zoltán Hegedűs[b,c,1], Anna Zakrzewska[a,1], Vilmos C. Ágoston[b], Anita Ordas[a,d], Péter Rácz[a,b,d], Mátyás Mink[d], Herman P. Spaink[a], Annemarie H. Meijer[a,*]

[a] Institute of Biology, Leiden University, PO Box 9502, 2300 RA, Leiden, The Netherlands
[b] Zenon Bio Ltd., Maros u. 40, H-6721 Szeged, Hungary
[c] Bioinformatics Laboratory, Biological Research Center, Hungarian Academy of Sciences, Temesvári krt. 62., H-6726 Szeged, Hungary
[d] Szeged University, Department of Genetics, Középfasor 52, H-6726 Szeged, Hungary

A R T I C L E I N F O

Article history:
Received 1 June 2009
Accepted 1 July 2009
Available online 24 July 2009

Keywords:
Tuberculosis
Digital gene expression
Illumina sequencing
Microarray
Transcriptome profiling
Differential expression
Transcript isoforms

A B S T R A C T

Novel high-throughput deep sequencing technology has dramatically changed the way that the functional complexity of transcriptomes can be studied. Here we report on the first use of this technology to gain insight into the wide range of transcriptional responses that are associated with an infectious disease process. Using Solexa/Illumina's digital gene expression (DGE) system, a tag-based transcriptome sequencing method, we investigated mycobacterium-induced transcriptome changes in a model vertebrate species, the zebrafish. We obtained a sequencing depth of over 5 million tags per sample with strong correlation between replicates. Tag mapping indicated that healthy and infected adult zebrafish express over 70% of all genes represented in transcript databases. Comparison of the data with a previous multi-platform microarray analysis showed that both types of technologies identified regulation of similar functional groups of genes. However, the unbiased nature of DGE analysis provided insights that microarray analysis could not have achieved. In particular, we show that DGE data sets are instrumental for verification of predicted gene models and allowed us to detect mycobacterium-regulated switching between different transcript isoforms. Moreover, genomic mapping of infection-induced DGE tags revealed novel transcript forms for which any previous EST-based evidence of expression was lacking. In conclusion, our deep sequencing analysis revealed in depth the high degree of transcriptional complexity of the host response to mycobacterial infection and resulted in the discovery and validation of new gene products with induced expression in infected individuals.

1. Introduction

Cellular identity and function is determined by the transcriptome: the complete repertoire of expressed RNA transcripts. Development and disease processes in multicellular organisms are governed by complex variations in transcriptional activities. Deciphering the functional complexity of transcriptomes is extremely challenging, especially since recent studies indicate that much more of the genome is transcribed than previously thought and that the majority of genes is transcribed in a bidirectional manner (Carninci et al., 2005; Katayama et al., 2005; Yelin et al., 2003). Recent advances in the development of ultra high-throughput deep sequencing technologies are making a huge impact on genomic research. These next generation sequencing systems, such as the Solexa/Illumina Genome Analyzer and the ABI/SOLiD Gene Sequencer, can sequence in parallel millions of DNA molecules derived directly from mRNA, without the need to use bacterial clones (Cloonan and Grimmond, 2008; Morozova and Marra, 2008; Wang et al., 2009). The direct sequencing yields libraries of short sequences (25–50 nucleotides), referred to as RNA-Seq data or Digital Gene Expression (DGE) data. Sequencing-based methods generate absolute rather than relative gene expression measurements and avoid many of the inherent limitations of microarray analysis (Irizarry et al., 2005; Pedotti et al., 2008; t Hoen et al., 2008; Wilhelm and Landry, 2009), which has been the most commonly used technology for transcriptome profiling over the last decade.

The first results of transcriptome profiling using next generation sequencing technology have recently been published (Cloonan et al., 2008; Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Sultan et al., 2008; Wilhelm et al., 2008). These RNA-Seq studies were based on different procedures starting either with mRNA fragmentation or with cDNA synthesis that is followed by fragmentation. Dependent on the protocol used, information on

* Corresponding author. Tel.: +31 71 5274927.
E-mail address: a.h.meijer@biology.leidenuniv.nl (A.H. Meijer).
[1] These authors contributed equally to this work.

0161-5890/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.molimm.2009.07.002

transcript directionality, which is useful for annotation and detection of antisense transcription, was retained or not (Shendure, 2008). RNA-Seq studies of yeasts detected transcription of 92–99% of all known genes and demonstrated that most of the genome sequence is transcribed (75% in *Saccharomyces cerevisiae* and >90% in *Saccharomyces pombe*) (Nagalakshmi et al., 2008; Wilhelm et al., 2008). Evidence of transcript heterogeneity and of novel transcribed regions, including non-coding and antisense transcripts, was obtained. Furthermore, a genome-wide regulation of splicing was revealed (Wilhelm et al., 2008). Bioinformatical analysis of RNA-Seq data needs to deal with mapping of short reads to the genome, issues of multi-mapping, and mapping of splice-crossing reads. The analysis becomes even more challenging when dealing with the more complex genomes of vertebrate species. Nevertheless, the first applications of RNA-Seq methodology to mouse and human have now already provided rich information for new or revised gene models. By indicating the presence of additional promoters, alternative exons, alternative 3′ UTRs, non-coding transcripts and bidirectional transcripts, these studies have made it clear that genome annotation is far from complete (Cloonan et al., 2008; Morin et al., 2008; Mortazavi et al., 2008; Pan et al., 2008; Rosenkranz et al., 2008; Sultan et al., 2008; Wang et al., 2008).

An alternative to RNA-Seq is the use of tag-based transcriptome sequencing methods, such as serial analysis of gene expression (SAGE), which generates short signature sequences (tags) for the 3′-end regions of mRNA transcripts (Harbers and Carninci, 2005; Velculescu et al., 1995). SAGE-derived technologies include MPSS (massive parallel signature sequencing) and PMAGE (polony multiplex analysis of gene expression) that rely on amplification of 3′ tag sequences on microbeads (Brenner et al., 2000; Kim et al., 2007). Other useful tag-based sequencing technologies include the 5′-SAGE or CAGE (cap analysis of gene expression) methods that determine the 5′-ends of mRNAs by oligo-capping or cap-trapping (Harbers and Carninci, 2005; Hashimoto et al., 2009). The classical LongSAGE technique is based on the cleavage of cDNA with two restriction enzymes, an anchoring enzyme (commonly NlaIII, which cuts at CATG sites) and a tagging enzyme (MmeI) cutting 17 bp downstream of the anchoring enzyme's recognition site. Although SAGE, LongSAGE and similar SuperSAGE methods have greatly contributed to transcriptome analysis, their application has previously been limited by laborious ditag formation and concatemer cloning procedures and by the costs and throughput level of sequencing steps. However, SAGE-derived methods, which are particularly suitable for quantitative expression analysis, are back into view with next generation sequencing technology that entirely eliminates the need for tag cloning and provides a much greater sequencing depth.

Here, we used the second Illumina Genome Analyzer platform (GA II) to perform a SAGE-derived Digital Gene Expression (DGE) analysis of the zebrafish transcriptome response to mycobacterium infection. The zebrafish-mycobacterium model recapitulates hallmark features of human tuberculosis (Lesley and Ramakrishnan, 2008; Swaim et al., 2006). We have previously performed a multi-platform microarray study showing that mycobacterium-infected zebrafish express many homologs of human immune response genes and genes that have previously been implicated in the response to mycobacterial infection (Meijer et al., 2005). We also observed induction of genes with previously unknown relationship to the immune response, indicating that the use of the zebrafish-mycobacterium model can assist functional annotation of genes and provide new leads in the investigation of mycobacterial pathogenesis. The previous study was limited by contents of commercially available microarray designs, whereas the DGE analysis reported here provides wide unbiased coverage of the entire transcriptome. Comparison of DGE and microarray data revealed a substantial degree of overlap in differentially expressed transcripts as well as technology-dependent differences, indicating the value of using two complementary transcriptome analysis methods. Furthermore, by mapping our DGE tag data onto transcript databases and onto the genomic sequence we could show that many alternative transcript forms and novel transcripts are disease-specifically regulated, information that could not have been obtained by microarray analysis.

2. Materials and methods

2.1. Zebrafish husbandry and infection experiments

Zebrafish were handled in compliance with the local animal welfare regulations and maintained according to standard protocols (http://ZFIN.org). The infection experiment was approved by the animal welfare committee (DEC) of Leiden University. Adult male zebrafish were infected by intraperitoneal inoculation with approximately 1×10^3 *Mycobacterium marinum* bacteria as previously described (Meijer et al., 2005). Four of the RNA samples used for this study were identical to those from our previously published chronic infection study (control fishes c1, c2 and infected fishes i1, i2) (Meijer et al., 2005). Four additional RNA samples (control fishes c3, c4 and infected fishes i3, i4) were from an independent *M. marinum* E11 infection experiment performed under similar conditions. All four infected fish were sacrificed when they showed overt signs of fish tuberculosis, including lethargy, skin ulcers and extensive granuloma formation in organs, such as liver and kidney. Histological examination of fish from the same experiments confirmed that the pathology of infected fish corresponded to fish tuberculosis (Swaim et al., 2006; van der Sar et al., 2004) and that no characteristics of the disease were present in the control fish.

2.2. RNA isolation

Fish were snap frozen in liquid nitrogen, stored at −80 °C, and homogenized in liquid nitrogen using a mortar and pestle. Portions of 50–100 μg of powdered tissue were used for extraction of total RNA with 1 ml of TRIZOL® Reagent (Invitrogen) according to the manufacturer's instructions. The RNA samples were incubated for 20 min at 37 °C with 10 units of DNaseI (Roche Applied Science) to remove residual genomic DNA prior to clean up using RNeasy columns (Qiagen). The integrity of the RNA was confirmed by Lab-on-chip analysis using the 2100 Bioanalyzer (Agilent Technologies). The samples used had an average RIN value of 9.5 and a minimum RIN value of 8.9.

2.3. Digital gene expression-tag profiling (DGE)

For DGE analysis RNA samples from the four control fish (c1, c2, c3, c4) were pooled, and RNA samples from the four infected fish (i1, i2, i3, i4) were pooled. Before pooling the individual RNA samples had been checked by microarray analysis for correlation between biological replicates. From each pool duplicate libraries for tag sequencing were prepared in order to assess technical reproducibility. Tag library preparation was performed using the DGE: Tag Profiling for NlaIII Sample Prep kit from Illumina according to the manufacturer's instructions. In brief, 1 μg of total RNA was used for mRNA capture using magnetic oligo(dT)beads. First- and second-strand cDNA was synthesized and bead-bound cDNA was subsequently digested with NlaIII. Fragments other than the 3′ cDNA fragments attached to oligo(dT) beads were washed away and a GEX NlaIII adapter was ligated to the free 5′-end of the digested bead-bound cDNA fragments. The GEX NlaIII adapter contains a restriction site for MmeI which cuts 17–18 bp downstream from the NlaIII site, thereby releasing 21–22 bp tags starting with the NlaIII recognition sequence, CATG. A second adapter (GEX adapter 2) was ligated at the site of MmeI cleavage, and the adapter-ligated cDNA tags were enriched using PCR-primers that anneal to the adaptor ends. The resulting 85 bp fragments were purified from a 6% acrylamide gel. Purity and yield were checked by Lab-on-chip analysis using the 2100 Bioanalyzer (Agilent Technologies). A total of 6 pmol of cDNA per tag library was used for cluster generation on individual lanes of Illumina's 1.4 mm channel flow cell, and sequencing by synthesis was performed using the Illumina Genome Analyzer II system (ServiceXS, Leiden, the Netherlands) according to the manufacturer's protocols. Image analysis, base calling, extraction of 17 bp tags and tag counting were performed using the Illumina pipeline. The raw data (tag sequences and counts) were deposited in the GEO database under submission number GSE14782.

2.4. Statistical evaluation of DGE libraries

The tag entities and count numbers of DGE libraries from control and infected fish were collected and summarized by custom perl and PHP scripts. Statistical comparison was performed using the Bayesian method described by Lash et al. (2000) with the software tool available from the SAGEmap resource. We accepted the change of a tag expression as significant if the chance of a false positive hit was less than 5% ($P < 0.05$). The correlation of the detected count numbers between parallel libraries was statistically assessed by calculation of Pearson correlation coefficient using the built-in function of Microsoft Excel.

2.5. Mapping of DGE tags

For mapping of DGE tags to different transcript databases or to the zebrafish genome we created virtual libraries containing all the possible 17 bases length sequences of these resources located next to an NlaIII restriction site. For transcript mapping we used the UniGene (Danio rerio build 105), Refseq (2007_07) and Ensembl (ZFISH7.49) transcript databases. For genomic mapping we used the native and the masked form of the zebrafish genome version Zv7, which were downloaded from the FTP server of the Ensembl database. For mapping against UniGene transcripts we used the file containing the sequence with the longest region of high-quality sequence data from each UniGene cluster. For monitoring the mapping events on both strands, both the sense and the complementary antisense sequences were included in the data collection. Information on the position of polyadenylation signals and the length of polyadenylation tails was also collected from the transcript databases. Virtual tag libraries and the DGE libraries were uploaded into an in house developed data warehouse using a relational database engine for mapping of DGE tags onto virtual libraries. Only perfect matches over the entire 21 bp length of the 17 bp tag plus the 4 bp NlaIII recognition site were allowed. During the mapping process the system tracked the so-called multiple mapping events where tags detected in the experiments could be assigned to more than one transcript or to more than one position in the genome.

2.6. Microarray analysis

Agilent microarray analysis was performed using a custom designed platform (GEO submission number GPL7735) with previously described conditions for labeling, hybridization, scanning and feature extraction (Stockhammer et al., 2009). Samples from control fish (c1–c4) were labeled with Cy3 dye and samples from infected fish (i1–i4) with Cy5 dye. After feature extraction, data were imported into Rosetta Resolver 7.1 (Rosetta Biosoftware, Seattle, Washington) and subjected to default ratio error modeling. Ratio results from four control fish vs. four infected fish (c1 vs. i1, c2 vs. i2, c3 vs. i3, c4 vs. i4) were combined using the default ratio experiment builder. Data were analyzed at the level of UniGene clusters (UniGene build #105). Microarray data were submitted to the GEO database under series GSE14782.

2.7. RT-PCR

Real-time quantitative RT-PCR (qPCR) was performed as previously described (Stockhammer et al., 2009). All reactions were done in duplicate. For normalization *ppial* (peptidylprolyl isomerase A like), which was unaffected by mycobacterium infection, was taken as reference. Sequences of forward and reverse primers are shown in Supplementary Table 1.

RT-PCR verification of the mycobacterium-induced novel transcript variant of the *zgc:112143* gene was performed using the SuperScript III One-Step RT-PCR System with Platinum Taq DNA Polymerase (Invitrogen). A primer overlapping with one of the significant tags located downstream of the known 3′ transcript end was used for the reverse transcription reaction (CCAGACACTCAAACAGACATG). The same primer was subsequently used in the PCR amplification step, in combination with a forward primer in exon 3 of the known transcript (ATGAGCGAGTTACCAACGGA). The resulting PCR product was sequenced using the sequencing service of ServiceXS (Leiden, The Netherlands) and submitted to GenBank under accession number FJ754358.

3. Results

3.1. Basic quantitative parameters and reproducibility of DGE library sequencing

In previous microarray studies we have show that mycobacterium infection of zebrafish results in a highly reproducible host response at the transcriptome level (Meijer et al., 2005; van der Sar et al., 2009). Here we have pooled biological replicates from these previous studies to make representative samples for deep sequencing analysis. We sequenced two DGE libraries (technical duplicates) of a pool of four control fish (cont-a, cont-b) and two DGE libraries of a pool of four mycobacterium-infected fish (inf-a, inf-b). The total count number of tags per library ranged from 5.3 to 8.4 million and the number of tag entities with unique nucleotide sequences

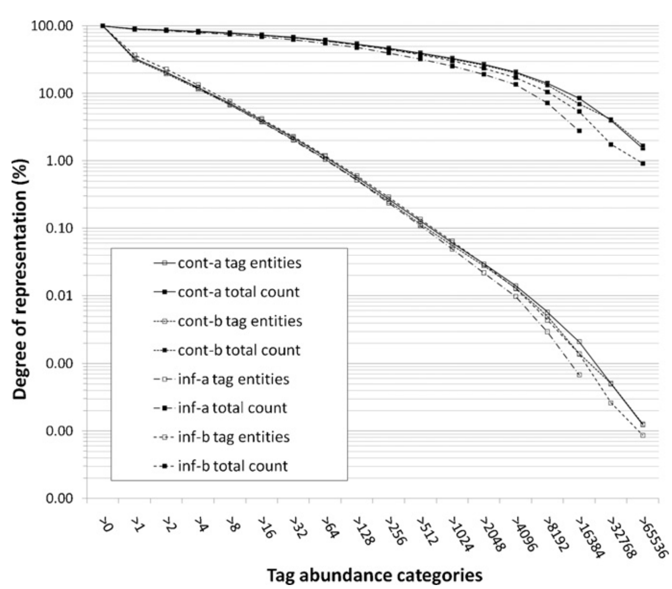


Fig. 1. Distribution of tag entities and total tag counts over different tag abundance categories. Categories of tag abundance were assigned by setting the lower limit of the count number that includes a tag as a category member. The percentages of total tag counts (filled squares) and number of different tag entities (open squares) per category are plotted on a logarithmic scale.

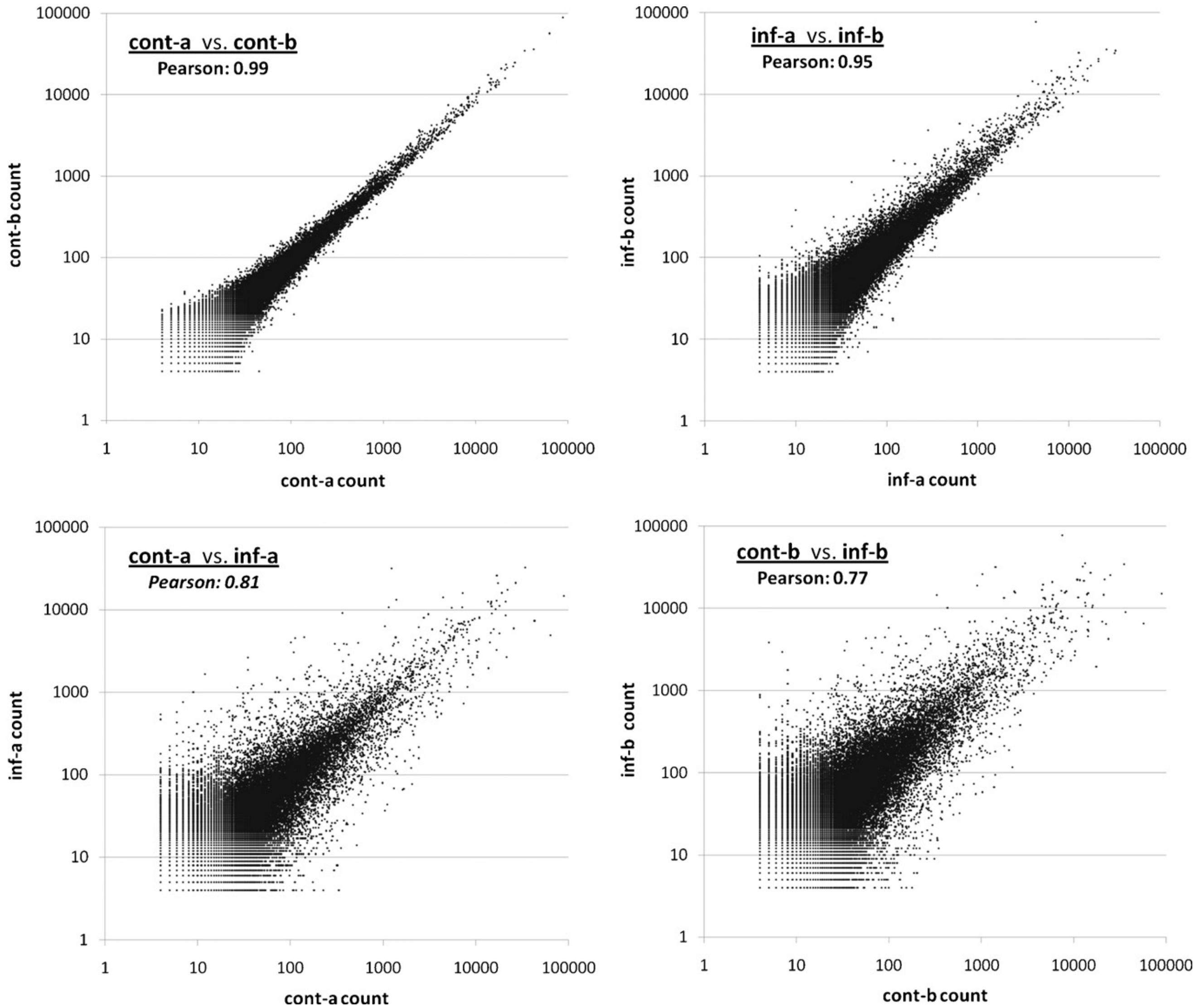


Fig. 2. Correlation analysis of DGE libraries. Correlation charts are shown of the tag entity counts of the four possible combinations of DGE libraries from control fish (cont-a and cont-b) and from mycobacterium-infected fish (inf-a and inf-b). Dots in the charts indicate individual tag entities. Pearson correlation coefficients are shown in the upper left corner of each plot.

ranged from 0.8 to 1.1 million (Supplementary Table 2). As shown in Fig. 1, the distribution of tag entities and total tag counts over different tag abundance categories showed very similar tendencies for all four DGE libraries. The lowest abundant tags that were still consistently detected in all four libraries occurred at a frequency of below 1 count per million. To further investigate the reproducibility of DGE library sequencing we performed correlation analyses of all four possible sample combinations (cont-a vs. cont-b, inf-a vs. inf-b, cont-a vs. inf-a, and cont-b vs. inf-b) (Fig. 2). Pearson correlations for the parallel libraries (cont-a vs. cont-b and inf-a vs. inf-b) were 0.99 and 0.95, indicating the high technical reproducibility of the DGE method. In contrast, Pearson correlations for control vs. infected library pairs were much lower (0.77 and 0.81), consistent with a large effect of mycobacterium infection on the host transcriptome as previously observed in microarray experiments (Meijer et al., 2005; van der Sar et al., 2009).

3.2. *Efficiency of tag mapping to transcript databases*

Mapping the tags to known transcripts is the most efficient way to reveal the molecular events behind DGE profiles. In our study the tag sequences of the four DGE libraries were mapped to the zebrafish transcript datasets of the UniGene, RefSeq and Ensembl databases and found to match with over 70% of all sequence entries in these databases (Table 1). The mapping efficiency increased significantly as the count number of tags increased. Specifically, a 50% mapping efficiency was observed for tags with around 10 copies and the mapping efficiency increased to about 80% for tags with an abundance of over 200 copies (Fig. 3). This means that the most abundant tags correspond to the most highly expressed transcripts, which in turn are most likely to be found in the existing zebrafish transcript collections. Tags mapping to a unique sequence position form the most important subset of the DGE libraries as they can unambiguously identify a transcript. Up to 54% of the sequence records in the different transcript databases could be unequivocally identified by unique tag mapping (Supplementary Table 3). Examples of the most abundantly expressed genes in all libraries that could be unequivocally identified by tag mapping comprised more than ten different ribosomal subunit genes, translation initiation and elongation factor genes (*eif5a* and *ef1a*), creatine kinase genes (muscles a and b) and several cytoskeletal protein genes, including those encoding alpha and beta actin, keratin and skeletal muscle myosin proteins.

Table 1
Representation of transcript databank entries in the DGE libraries.

Transcript databank	Total number of databank entries	cont-a + cont-b libraries[a]	inf-a + inf-b libraries[a]
UniGene	56,561	73%	77%
RefSeq	37,396	69%	73%
Ensembl	31,841	74%	86%

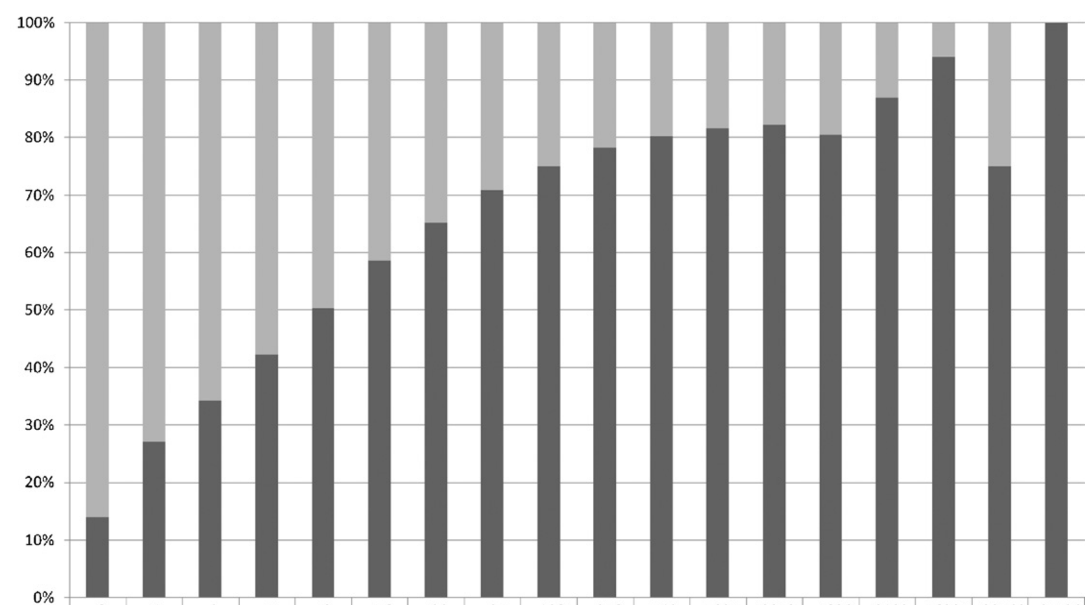
[a] Replicate DGE libraries (a and b) were merged *in silico* and indicated is the percentage of entries from the UniGene, RefSeq and Ensembl transcript databanks that were hit by mapping of DGE tags from the merged libraries.

3.3. Changes in tag profile induced by mycobacterium infection

In order to identify the tags showing a significant change in expression between the two control vs. infected library pairs we used the Bayesian approach described by Lash et al. that takes into account differences in library size (Lash et al., 2000). The cont-b vs. inf-b library pair showed the highest number of significant tag entities (Supplementary Fig. 1), which can be explained by the greater sequencing depth of inf-b library compared with the inf-a library (Supplementary Table 2). A total of 5049 significantly changed tag entities were detected in the intersection of the cont-a vs. inf-a and the cont-b vs. inf-b library pairs (Supplementary Fig. 1; Supplementary Table 4). Efficiencies of tag mapping to the UniGene, RefSeq and Ensembl transcript databases were very similar for the two parallel library pairs (Supplementary Table 5). In both cases the highest percentage of significant tags (ca. 40%) could be mapped to the UniGene database, which is the largest in size (Supplementary Table 5). The intersection of the two library pairs contained a total of 1051 different UniGene transcripts, constituting 2% of the total number of transcripts in this database (Supplementary Fig. 1). Mapping to a unique UniGene transcript was observed for 27% of the significant tags. Approximately 2-fold more significant tag entities mapped to the sense strand of the transcripts than to the antisense strand in any of the three investigated transcript databases (Supplementary Table 5). By comparison, the ratio of sense to antisense mapping of the total number of tags (significant and non-significant) was approximately 1:1 for all libraries. This suggests that in spite of the high number of antisense mapping events detected, the transcriptional regulation in the mycobacterium-induced immune response acts most strongly on the sense strand. The possibility that the host response to mycobacterium infection may also depend on the regulation of antisense expression is of great interest and will require further investigations. In the present study, we focused on the regulation of sense strand transcripts.

3.4. Comparison of DGE tag data with a reference set of mycobacterium-regulated genes

Previously we analyzed the zebrafish transcriptome response to mycobacterium infection using three different microarray platforms (Sigma-Compugen spotted oligonucleotide library, MWG spotted oligonucleotide library and Affymetrix). In that study we identified a set of genes whose differential expression was detected irrespective of the type of microarray used and that could therefore serve as a multi-platform reference for future transcriptome profiling studies in the mycobacterium—zebrafish model (Meijer et al., 2005). For the present study we took the exact same samples as used for DGE tag profiling and hybridized these to a fourth microarray platform (Agilent 44k). Next, a new reference set was defined consisting of 121 mycobacterium-regulated genes (55 up-regulated and 66 down-regulated) confirmed in two separate studies and by in total four microarray platforms (Supplementary Table 6). This reference set was used for comparison with our DGE tag data. As shown in Supplementary Table 6, 120 genes out of the 121 genes in the reference set were represented by transcript tags in our DGE libraries. The only gene not represented lacks the NlaIII restriction site that is required for detection by DGE. For 100 out of the 120



	>0	>1	>2	>4	>8	>16	>32	>64	>128	>256	>512	>1024	>2048	>4096	>8192	>16384	>32768	>65536
not mapped	692114	193944	108373	56991	28471	13433	6169	2695	1134	470	202	90	42	22	6	1	1	0
mapped	111752	72154	56398	41478	28809	18927	11511	6568	3399	1695	821	399	195	91	40	16	3	1

Fig. 3. Efficiency of tag mapping to the UniGene transcript database. Bars indicate the percentages of successful (dark grey) and non-successful (light grey) tag mapping events for the different categories of tag abundance as assigned in Fig. 1. The numbers of tag entities per category are indicated below the graph.

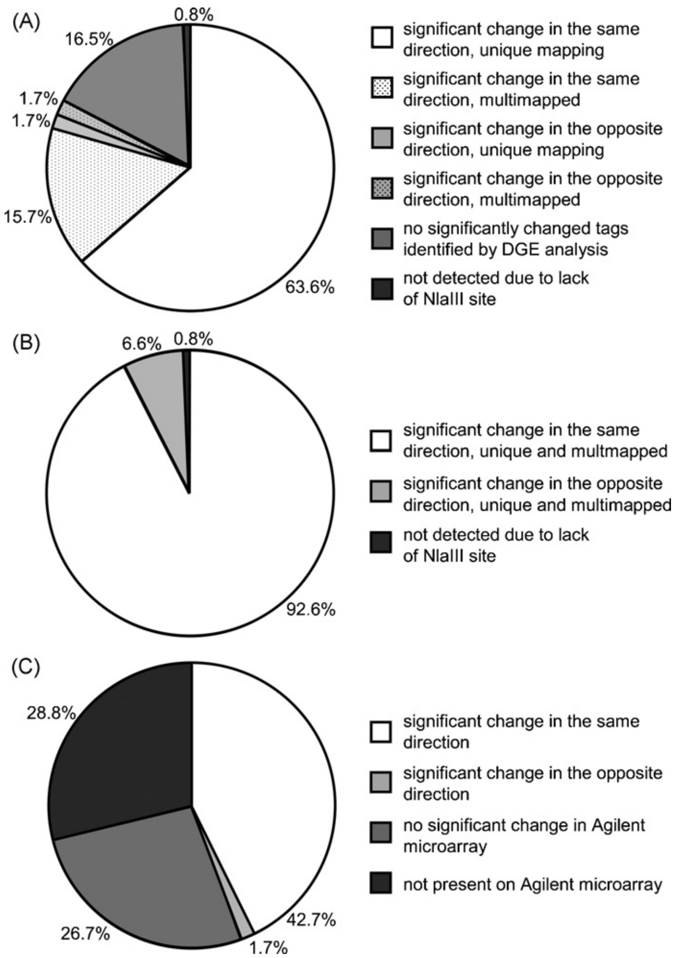


Fig. 4. Correlation between DGE and microarray analysis. (A) Comparison of DGE results with a reference set of mycobacterium-regulated genes based on four microarray platforms. DGE results were evaluated by *P*-value with the significance threshold at 0.05. (B) The same comparison as in (A), but based on the direction of change of all the tags (significant and non-significant) mapped on the same transcript. This approach gives a raw estimation of concordance even in those cases where DGE detected non-significant mappings. (C) Comparison of DGE and Agilent microarray results.

genes detected (83%) the DGE experiments identified significantly different tag counts between control and infected libraries, whereas the significance threshold was not met for 20 genes. 79 out of 100 significantly changing genes were unambiguously identified by the DGE tags. In 97% of these cases (77 out of 79) the direction of change positively correlated between DGE and microarray analysis (Fig. 4A; Supplementary Table 6). Identification of the remaining 21 genes was ambiguous because the tags mapped to multiple UniGene transcripts. However, in nearly all cases (20 out of 21) multiple mapping occurred to highly related transcripts, encoding a fragment or isoform of the same protein or a near-identical protein. Furthermore, the direction of change positively correlated between DGE and microarray data in 90% of the cases where multiple mapping occurred (19 out of 21) (Fig. 4A; Supplementary Table 6). Therefore, we conclude that in most cases even the multi-mapped tags, which are usually discarded in SAGE-based experimental systems, can also reveal useful information. Finally, when we extended the comparison of the direction of change between DGE and microarray analysis to include also the non-significant tags (mapping to 20 UniGenes in the reference set), we found a positive correlation in 93% of the cases (112 out of the 120 detected genes, Fig. 4B). In conclusion, comparison of our DGE tag data with the reference set of mycobacterium-regulated genes indicates a strong agreement between data obtained by different transcriptome profiling technologies.

3.5. Comparison of DGE tag data with Agilent microarray data and qPCR

The flexibility of the Agilent platform is a major advantage for microarray analyses of zebrafish since it allows continuous upgrading of custom designed probes. Agilent is also a preferred platform because previously used platforms (Meijer et al., 2005) are no longer available (MWG, Sigma-Compugen) or several years outdated (Affymetrix). In order to compare our DGE tag data to Agilent microarray data we used the subset of UniGene transcripts that were identified in both control-infected library pairs by unambiguous mapping of significant tags to the sense strand. A total of 815 different UniGene identifiers met these criteria but only 580 of these UniGenes were represented in the Agilent probe set. Agilent microarray analysis showed a significant change in the same direction as determined by DGE analysis for 348 of these genes (60%) (Fig. 4C; Supplementary Table 7). Only 14 genes showed a conflicting significant change in the opposite direction, and no significant change was detected for 218 genes (Fig. 4C; Supplementary Table 7). These results indicate a substantial overlap between DGE and Agilent microarray results in addition to technology-dependent differences.

In order to verify a subset of the DGE tag data by a third independent technology we used quantitative reverse transcriptase PCR (qPCR) analysis. The selection of genes tested included 6 genes that showed a conflict between DGE and Agilent microarray data (Fig. 5A), 3 genes that behaved similarly between the two technologies (Fig. 5B), and 5 genes whose differential expression was detected only by DGE analysis because they were not represented on the Agilent microarray (Fig. 5C). For almost all genes tested, with the exception of *jak1* only (Fig. 5A), qPCR analysis confirmed the direction of change detected by DGE analysis.

3.6. Functional annotation and human disease relationships

To achieve an unbiased functional annotation of the infection-responsive genes that were identified by DGE analysis we tested for significant enrichment of Gene Ontology (GO) groups (Supplementary Table 8). Among the up-regulated UniGene transcripts we observed specific enrichment of gene groups including immune response related transcription factors, proinflammatory cytokines and MHC class II proteins (with enriched GO-terms ‘immune system process’, ‘response to stimulus’, ‘extracellular region’, ‘catalytic activity’). Gene groups encoding proteolytic enzymes (such as matrix metalloproteinases, cathepsins and proteasome subunits) and lysosomal proton transporting ATPases were also enriched among the up-regulated transcripts. Statistical testing of the down-regulated UniGene transcripts revealed significant enrichment of genes groups encoding enzymes involved in carbohydrate, alcohol, steroid, amino acid and lipid metabolism (with enriched GO-terms ‘metabolic process’ and ‘catalytic activity’). Additionally down-regulated were genes encoding cytoskeletal proteins (among which skeletal muscle actin and myosin), tight junction proteins, solute carriers and fatty acid binding proteins (with enriched GO terms ‘structural molecule activity’, ‘transporter activity’, ‘binding’, ‘envelope’). The up- and down-regulated gene groups corresponded well with those identified in our previous microarray analysis of the zebrafish host response to mycobacterium infection (Meijer et al., 2005), indicating that DGE and microarray analysis identified similar functional groups of genes.

Gene ontology analysis allowed identification of only 41% of the significantly up- or down-regulated transcripts. In order to extend and improve the annotation of the DGE data set we retrieved the

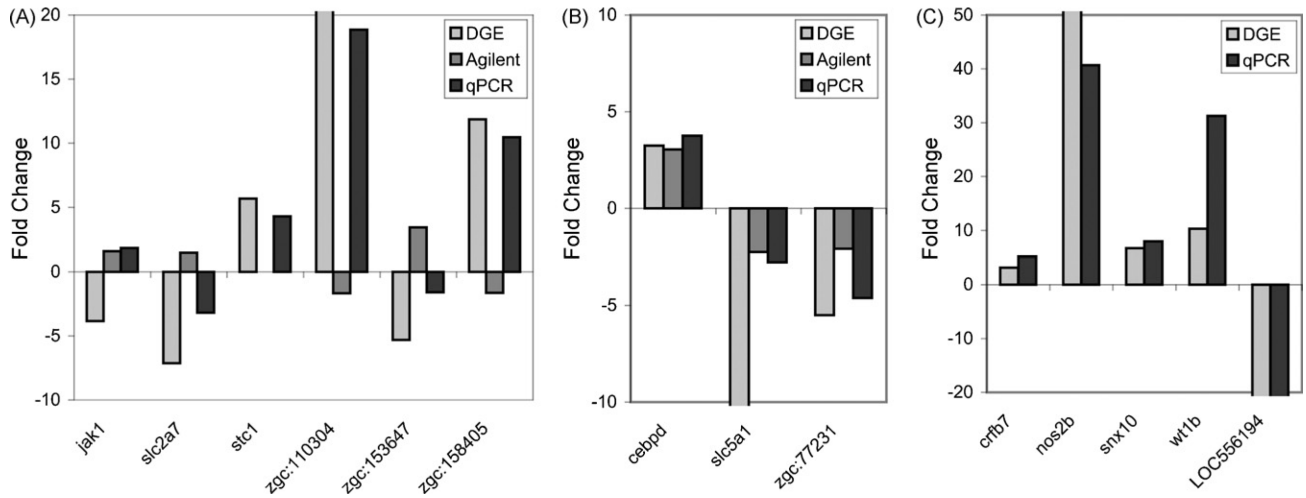


Fig. 5. qPCR validation of DGE tag data. (A) Genes showing a significant change by DGE analysis and for which Agilent microarray analysis showed a significant change in the opposite direction (*jak1*, *slc2a7*, *zgc:110304*, *zgc:153647*, *zgc:158405*) or no change (*stc1*). (B) Genes showing a significant change in the same direction by DGE and Agilent microarray analysis. (C) Genes showing a significant change by DGE analysis and not represented on the Agilent microarray platform. Up-regulation of gene expression by mycobacterium infection is indicated by a positive fold change and down-regulation by a negative fold change. Bars touching the chart border indicate infinite fold changes (zero tags in either the control or infected libraries). The following selection of genes was tested with their gene description, UniGene ID and Entrez Gene ID indicated between brackets: *jak1* (janus kinase 1, Dr.74470, 30280), *slc2a7* (slc2a7 solute carrier family 2 (facilitated glucose transporter), member 7, Dr.77040, 100006415), *stc1* (stanniocalcin 1, Dr.88421, 393511), *zgc:110304* (predicted tumor-associated calcium signal transducer 1 homolog, Dr.39071, 550255), *zgc:153647* (predicted cytochrome P450, family 2, subfamily J, polypeptide 2 homolog, Dr.79897, 768288), *zgc:158405* (predicted neutrophil cytosolic factor 2 homolog, Dr.66415, 562473), *cebpd* (CCAAT/enhancer binding protein (C/EBP), delta, Dr.1280, 140817), *slc5a1* (solute carrier family 5 (sodium/glucose cotransporter), member 1, Dr.87868, 93654), *crfb7* (cytokine receptor family member b7, Dr.91624, 777651), *nos2b* (nitric oxide synthase 2b, inducible, Dr.118320, 563654) *snx10* (sorting nexin 10a, Dr.13606, 403027), *wt1b* (wilms tumor 1b, Dr.91799, 568416), *LOC556194* (hypothetical LOC556194, Dr.77733, 556194), *zgc:77231* (predicted fast skeletal myosin alkali light chain 1 isoform 1f homolog, Dr.1448, 336165).

putative human orthologs of the zebrafish genes by using the g:Orth function of the g:Profiler web server (http://biit.cs.ut.ee/gprofiler) and by searching the NCBI HomoloGene database. In addition, we compiled information on protein similarities from the UniGene database. As a result, we were able to annotate 73% of the up-regulated and 83% of the down-regulated transcripts, whereas the remaining transcripts lacked any similarity to genes with known function (Fig. 6; Supplementary Table 9).

Since *M. marinum* infection of zebrafish is considered as an animal model of human tuberculosis, next we analyzed how the whole set of mycobacterium-regulated genes may associate with a known human pathological condition. A search using GeneALaCart, provided by GeneCards (www.genecards.org), linked our data set to 397 OMIM (Online Mendelian Inheritance in Man) disorders, of which 132 were additionally represented by UniProt disorders that are associated with monogenic human genetic diseases (Supplementary Table 9). This is a substantial proportion (16%) of the mycobacterium-regulated UniGene clusters identified by DGE analysis.

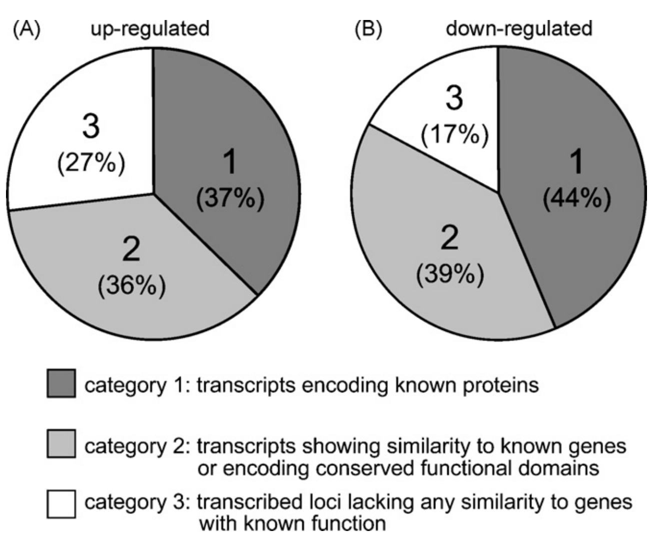


Fig. 6. Distribution of transcripts over different annotation categories. (A) Transcripts up-regulated by mycobacterium infection. (B) Transcripts down-regulated by mycobacterium infection.

3.7. Genomic mapping of DGE tag sequences

Disease-related genes may be underrepresented in the zebrafish transcript databases that are mostly based on sequencing of cDNA libraries from healthy fish. Therefore, additional mycobacterium-regulated transcripts can be revealed by genomic mapping of the DGE tag sequences (Supplementary tables 2 and 5). Genomic mapping of the significant tags had a success rate of 59% (of which 84% unique mapping events), which was approximately 1.5-fold more efficient than mapping to the UniGene transcript database (Supplementary Table 5). The fact that we did not allow mismatches and that the zebrafish genome is highly polymorphic may explain that still a substantial proportion of the significant tags could not be mapped. In addition, tags extending over intron boundaries would not be picked up in genomic mapping and gaps in the zebrafish genome sequence still exist. The chromosomal distribution of the complete collection of tag entities (significant and non-significant) was roughly proportional to the chromosome length (Fig. 7A). However, the chromosomal distribution of the significant tags was less flat (Fig. 7B) and specific regions could be identified within the individual chromosomes where a number of tags clustered together, suggestive of potential genomic hotspots linked with the physiological processes induced by mycobacterium infection. In Fig. 7C, we show an example of such a region on chromosome 3 where 10 mycobacterium-regulated genes were identified. These included *grb2* and *ras* family genes (*zgc:112091*, *rac2*, *arl6ip1*) linked with immune response and hematopoietic signaling (e.g. Diebold and Bokoch, 2001; Martin et al., 2005; Pettersson et al., 2000), several genes involved in inflammation and defense (encoding granulin a, ferritin-like, nucleobindin-1 and actinoporin proteins (Aroian and van der Goot, 2007; He and Bateman, 2003; Leclerc et al.,

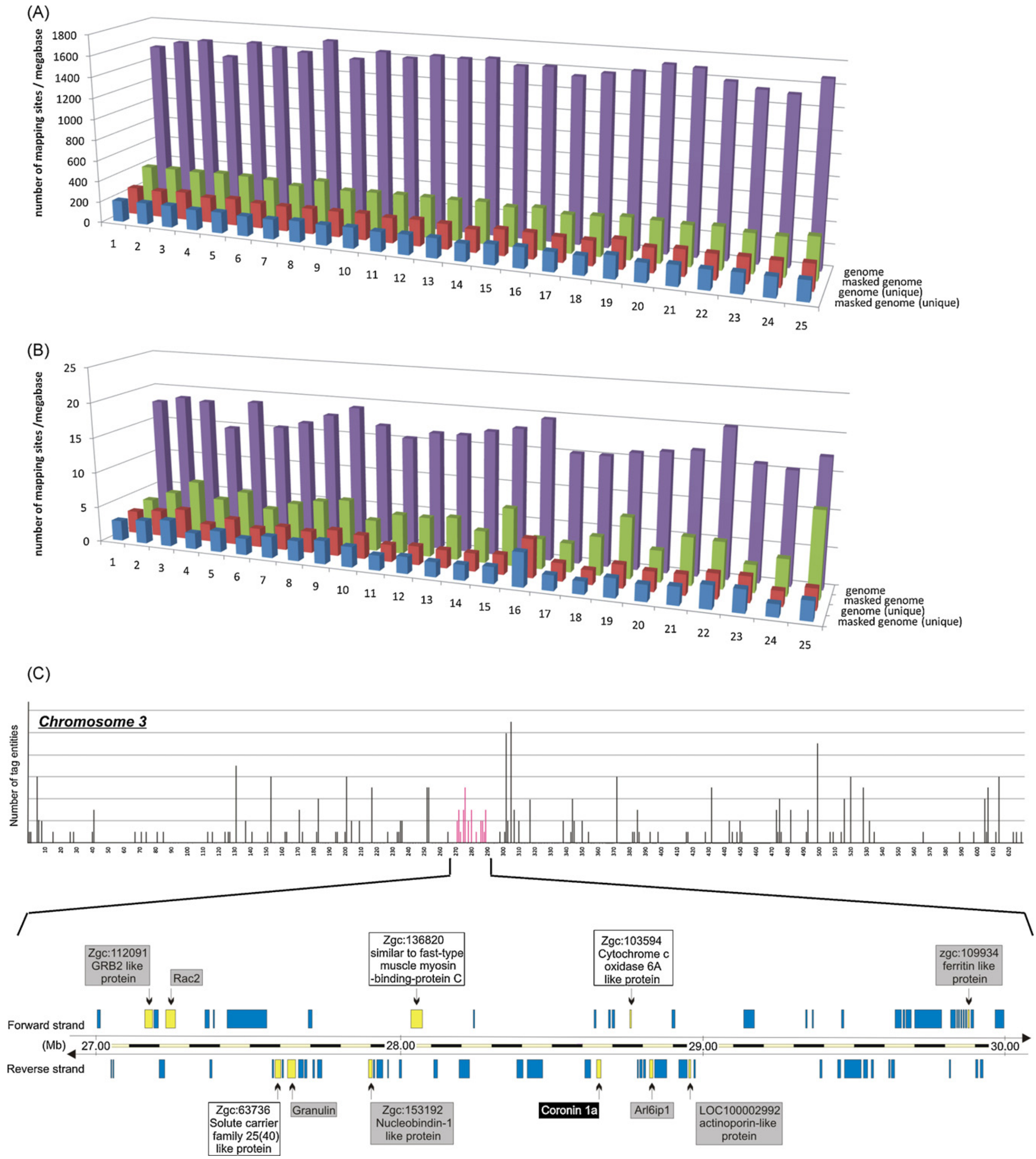


Fig. 7. Distribution of genomic DGE tag mapping sites over the different chromosomes. (A) Normalized chromosomal distribution of DGE tags (significant and non-significant) per megabase. Mapping results of one DGE library (cont-a) are shown as a representative example. The total number of mapping sites and the number of mapping sites with unique sequence were determined using the unmasked and masked versions of the Zv7 zebrafish genome sequence. Both the upper and lower strand of genomic DNA was equally populated by the tags of the DGE libraries (Supplementary Table 3). (B) Normalized chromosomal distribution of the DGE tags that were significantly changed by mycobacterium infection. (C) Distribution of the unique and significant tags mapping to chromosome 3. A region between 27.2 and 29.9 Mb where 29 significant tags map closely together, assigning an active chromosomal region during mycobacterium infection, is indicated in red with the genomic structure shown below. Genes that are hit on the sense strand by the significant tags are shown in yellow and other genes in blue. Annotations of encoded proteins on the two strands of the genomic sequence are indicated by rectangles above and below the ruler. The *coronin 1a* gene (*coro1a; coronin, actin binding protein, 1A*) that is hit by 2 up-regulated tags has a direct link with tuberculosis infection (Jayachandran et al., 2007; Kaul, 2008). The proteins encoded by other genes that were hit by up-regulated tags are indicated in grey boxes. All these proteins have previously been linked with responses to inflammation or infection (e.g. Aroian and van der Goot, 2007; Diebold and Bokoch, 2001; He and Bateman, 2003; Leclerc et al., 2008; Martin et al., 2005; Pettersson et al., 2000; Recalcati et al., 2008). Proteins encoded by genes hit by down-regulated tags are shown in white boxes (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.).

2008; Recalcati et al., 2008), and the zebrafish homolog of the human *CORO1A* gene, which encodes the actin-binding protein coronin that has long been implicated in mycobacterial pathogenesis (Jayachandran et al., 2007; Kaul, 2008).

3.8. Verification of predicted gene models based on DGE tag information

Since DGE is a SAGE-based transcript profiling method, theoretically all the observed tags in the DGE libraries should be mapped to the so called canonical site, a sequence region located following the last NlaIII cutting site on the sense strand at the 3′-end of the transcript. In our experiments only about 40% of the sense strand mappings fell to the canonical site. Furthermore, approximately 70% of the detected transcripts were represented by more than one tag. These observations are consistent with similar results in DGE analysis of the mouse transcriptome (t Hoen et al., 2008) and may be explained partly by incomplete NlaIII digestion during library preparation as well as by frequent use of alternative polyadenylation and/or alternative splicing sites (Pan et al., 2008; Wang et al., 2008). The existence of multiple tags per transcript can be conveniently utilized for the experimental verification of *ab initio* gene predictions from the Ensembl database. Many of these predicted transcripts gather more than one tag hit and many of them have count numbers over several hundred up to thousands. Detection of multiple tags with high count numbers specific for a predicted transcript indicates the reliability of the transcript sequence and the biological importance of the given gene. In a specific example (Ensembl: GENSCAN00000043834, encoding a homolog of the human olfactomedin 4 (*OLFM4*) gene induced by mycobacterium infection in our DGE study) we show how tag mapping over different Genscan exons can support predicted gene models (Fig. 8A).

3.9. Infection-induced transcript isoform switching

Next we have tried to identify transcripts where the infection triggered unbalanced alterations in tag expression pattern that may be caused by the selective induction or repression of particular transcript isoforms generated by alternative splicing, alternative polyadenylation or alternative transcription initiation. To this extent we selected transcripts with at least two mapped tags, and we investigated how their tag expression patterns changed

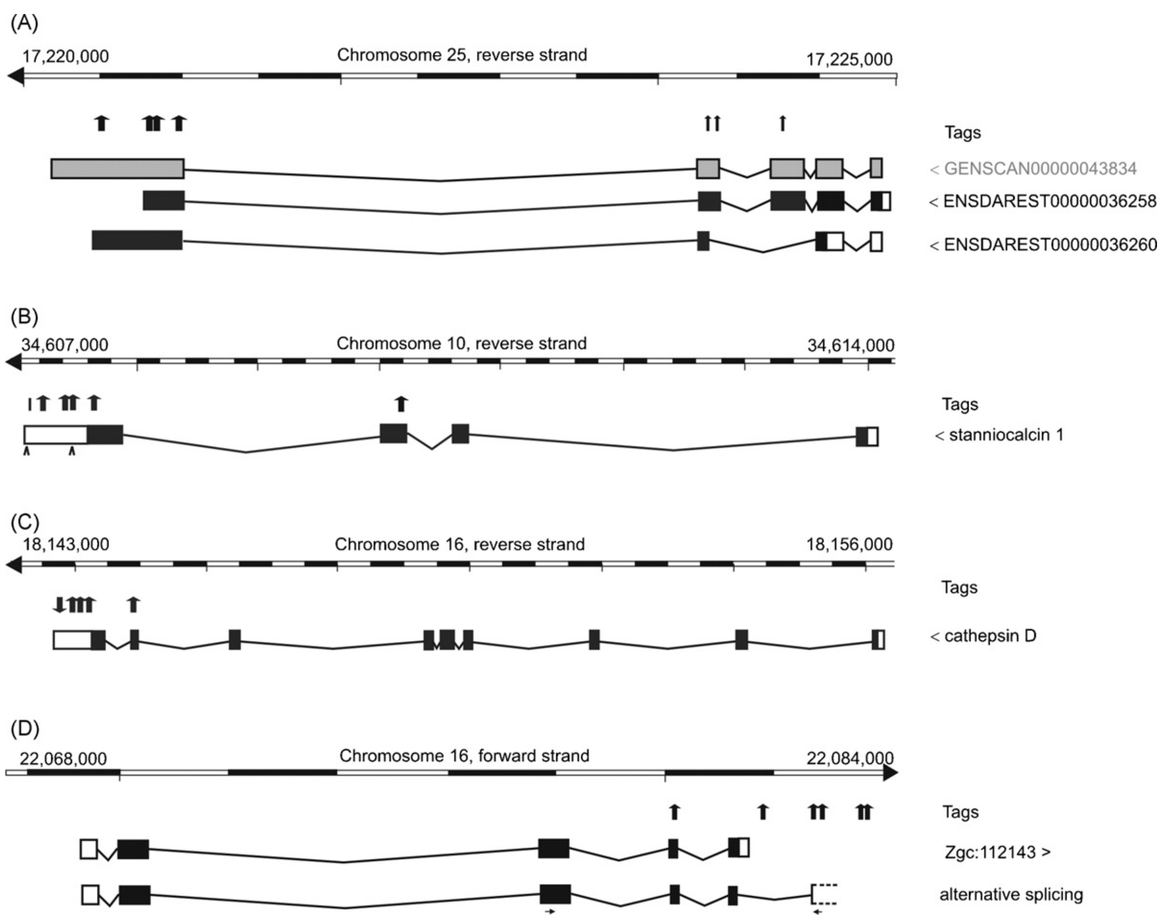


Fig. 8. Genome data mining using DGE tag information. Arrows pointing up indicate the mapping positions of tags with significantly increased expression under mycobacterium-infection conditions. Arrows pointing down indicate tags significantly down-regulated by mycobacterium infection. Lines indicate non-significant tags. Bold arrows correspond to tag counts above 100 in the DGE libraries of mycobacterium-infected fish. (A) Verification of an *ab initio* gene prediction from Ensembl. Three exons of the predicted GENSCAN00000043834 gene, a homolog of the human olfactomedin 4 gene, are supported by DGE tag information. The tag distribution further supports the existence of at least two alternative splice forms, represented by ENSDARESTT00000036260 and ENSDARESTT00000036258. (B) Infection-dependent isoform switching. Arrowheads below the gene indicate the positions of the last two polyadenylation signals in the *stc1* transcript. The unchanged expression of the most 3′ tag vs. the significant induction of more upstream tags suggest alternative polyadenylation induced by mycobacterium infection leading to a transcript with a shorter 3′ UTR. (C) Infection-dependent up- and down-regulation of two transcript isoforms of the *ctsd* gene. (D) Infection-dependent alternative splicing leading to a novel gene product. A cluster of 5 significant tags was detected downstream of the known *zgc:112143* gene, encoding the zebrafish homolog of the tumor-necrosis factor alpha-induced protein 9. Horizontal arrows indicate the position of RT-PCR primers used to verify the expression of an alternative transcript form that comprises the tag cluster region. From sequence analysis of the amplified product we could derive alternative splicing of the 5th exon to a new exon 6, encoding a different C-terminal domain of the protein. The 3′-end of exon 6 was not determined (open rectangle).

within the transcripts upon mycobacterium infection. Among the mycobacterium-regulated gene set we identified over 30 cases where part of the detected tag repertoire changed to a much greater extent than other tags, or where different tags for the same transcript were changed in the opposing direction. An example of a case where only part of the tag repertoire was changed is *stanniocalcin 1* (*stc1*), a gene for a calcium-regulating glycoprotein hormone. Up-regulation of this gene upon mycobacterium infection was confirmed by qPCR analysis (Fig. 5A). Five tags at the 3′-end of the transcript showed a significant change with 4–10 times increase in count number in the two control vs. infected library pairs (Fig. 8B). However, the most 3′ tag of this transcript was unchanged in both library pairs. Since the last four tags are located in the 3′ UTR of this transcript, this observation most probably reflects a selectively induced alternative polyadenylation event, where the infection results in the expression of a transcript variant having a shorter 3′ UTR region. This hypothesis is also supported by the finding that the 3′ UTR contains two polyadenylation signals, one at the very end of the sequence after the last weakly changing tag and another in a more upstream position. Alternative polyadenylation might be a mechanism for regulation of *stc1* mRNA stability, which was previously shown to be affected by external stimuli (Ellis and Wagner, 1995). An example of a case where tags for the same transcript changed in opposite directions is *cathepsin D* (*ctsd*). The tag expression pattern for this gene indicates selective down-regulation of a longer transcript form with concomitant up-regulation of a transcript with a shorter 3′ UTR (Fig. 8C).

3.10. *Identification of novel transcript forms induced by mycobacterium infection*

Finally, we attempted to find novel mycobacterium-regulated transcript forms (splice variants) without any previous indications from different prediction methods. As above, we took advantage of the observation that the majority of known transcripts were represented by more than one tag. We anticipated that the same would likely be true for currently unknown transcript forms and therefore searched for tag clusters with significant tags that could not be mapped to any of the three transcript databases used in this work, but having genomic localizations close to each other. Next, we examined the neighboring genomic environment of these tag clusters, and inspected the gene annotations in that region for the presence of potential immune relevant functions. Out of 98 investigated clusters 29 proved to be located close to a gene that might have a function in immune defense (Supplementary Table 10). As an example, we experimentally confirmed the expression of a novel splice form of the *zgc:112143* gene, which is homologous to the human tumor necrosis factor alpha-induced protein 9 gene (*TNFAIP9*, also known as *STEAP4*), encoding a six transmembrane putative channel or transporter protein (Moldes et al., 2001). To this extent we performed RT-PCR on RNA samples from control and infected fish, taking a forward primer in one of the known exons of the gene (exon 3) and a reverse primer overlapping with one of the mycobacterium-induced tags located downstream of the known 3′ transcript end in a region without any predicted or EST-based indications of transcription. An RT-PCR product could be amplified from infected but not from control RNA samples, in agreement with the much higher tag counts in the DGE libraries from infected fish (total count numbers: cont-a/cont-b: 3/2; inf-a/inf-b: 86/369). Sequence analysis of the resulting product revealed alternative splicing in the known exon 5 resulting in a novel transcript form comprising a 6th exon that is the source of the significant DGE tags that we identified in our study (Fig. 8D). Alternative splicing changes the C-terminal domain of the TNFAIP9 homolog, downstream of the six predicted transmembrane domains of this protein.

4. Discussion

Major advances in transcriptomics have become possible as a result of novel technology developments in deep sequencing (Cloonan and Grimmond, 2008; Morozova and Marra, 2008; Wang et al., 2009; Wilhelm and Landry, 2009). Here we report on the first deep sequencing study of the vertebrate host response to infectious disease. We chose the mycobacterium-zebrafish infection model as a case study because of its relevance to human tuberculosis, a disease that is characterized by an intricate interaction between host and pathogen (Lesley and Ramakrishnan, 2008; Pieters, 2008). Furthermore, multi-platform microarray data sets of infection in this model were available for validation purposes (Meijer et al., 2005). Transcriptome alterations underlying a complex process like infectious disease are characterized not only by massive gene induction and repression responses, but also by subtle changes in transcript levels and presumably in the expression of alternative transcript forms (Vos et al., 2007). Our results demonstrate that deep sequencing is a major improvement over microarray analysis for detection of transcriptional changes over such a wide range. Moreover, the unbiased nature of deep sequencing data gives a fundamental advantage for gene discovery and genome annotation.

In this study we used Solexa/Illumina's Digital Gene Expression (DGE) system, which is essentially a SAGE-based tag profiling approach. We could reach a sequencing depth of 5–8 million tags per library and confirmed the reproducibility of replicate DGE library constructions by correlation analysis. Using a Bayesian method for statistical evaluation of our DGE tag data from control and mycobacterium-infected animals, we found over 5000 sequence tags to be differentially expressed. Count numbers of the significant tags spanned a dynamic range of three orders of magnitude. An example of one of the most abundant tag sequences represented skeletal muscle actin alpha 1, which was detected by over 40,000 copies in the controls and down-regulated to around 10,000 copies upon mycobacterium infection. As an example of the sensitivity of the method, a tag sequence specific for the transcript encoding peptidoglycan recognition protein 1 was undetectable in the DGE libraries from control fish and was significantly up-regulated to around 20 copies in the DGE libraries from infected fish. The sensitivity of DGE profiling is further demonstrated by the fact that our analysis of a single developmental stage of the zebrafish (adult stage) could detect over 70% of all transcripts in the UniGene, RefSeq and Ensembl databases. The significantly changed tag sequences mapped to approximately 2% of the transcripts in these databases.

Since only one other Illumina-based DGE study has presently been reported (t Hoen et al., 2008), it was important to validate our DGE results by independent transcriptome profiling technologies. Our DGE data showed strong correlation (over 90%) with a reference set of mycobacterium-regulated genes that had previously been confirmed in a multi-platform microarray study. Comparison with microarray data obtained from a single microarray platform (Agilent) also showed a substantial level of correlation (60%). Furthermore, gene ontology analysis showed significant enrichment of similar functional groups of genes in DGE and Agilent microarray data sets. In most of the investigated cases where DGE and microarray analysis gave conflicting results, qPCR analysis supported the DGE results. Conflicts between DGE and microarray results may occur for technical reasons, but may also result from the fact that microarray probes will often detect a mixture of different transcript isoforms, whereas DGE analysis can discriminate between specific transcript isoforms (t Hoen et al., 2008). Both methods have their intrinsic limitations, as discussed extensively by t Hoen et al. (2008). In short, important drawbacks of microarray analysis concern limited sensitivity, cross-hybridization problems and inadequate probe design. On the other hand, DGE will fail to identify some transcripts that lack a unique tag sequence or cutting site for the DGE anchoring enzyme (NlaIII in this study). Therefore, microarray analysis and DGE can be considered complementary to each other. DGE compares favorably to microarray analysis in standardization between laboratories (Irizarry et al., 2005; Marioni et al., 2008; Pedotti et al., 2008; t Hoen et al., 2008). The most important advantage of DGE analysis is that the method is not limited by predefined array content. In our study, we detected differential expression of many transcripts that were not represented on the available microarray platforms. By qPCR analysis we confirmed the differential expression of a subset of these transcripts, with tag abundances in the range of 20–400 copies.

Classical SAGE studies (Velculescu et al., 1995), based on conventional sequencing of cloned tags, have focused on analysis of the canonical tags, which are the tags resulting from the most 3′ located NlaIII restriction site on the sense strand of the transcript. The drawback of mapping only the canonical tags is that any information on alternative polyadenylation or alternative splicing is lost. In addition to biological mechanisms, partial digestion may also contribute to non-canonical mappings. As long as partial digestion is carefully controlled for by simultaneous preparation of parallel libraries with the same batch of reagents it does not present a problem. Rather, as a result of the enormous sequencing depth achieved in DGE analysis, the more 5′ located non-canonical tags can provide a useful source of additional information. Furthermore, valuable insights can be obtained from genomic mapping of those tags that fail to hit the transcript databases. Here we have shown that the information obtained from genomic mapping of the entire set of DGE tags can be used (i) to investigate genomic clustering of co-regulated transcriptome responses, (ii) to verify the expression of *ab initio* predicted genes, (iii) to detect switching between alternative transcript forms induced by mycobacterial infection, and (iv) to detect and map novel mycobacterium-regulated transcript forms for which any previous EST-based evidence of expression was lacking. Altogether, we found a 50% higher efficiency for genomic mapping of the significant tag entities than for mapping to the UniGene transcript database. This shows the limitations of the use of cDNA sequence data for disease studies, a problem that is solved by deep sequencing strategies.

Contrary to tag-based DGE analysis, full RNA sequencing procedures (RNA-Seq), in which mRNA or cDNA is fragmented mechanically, result in overlapping short fragments that cover the entire transcriptome. Clearly this approach is even more powerful for unraveling transcriptome complexity. However, tag sampling methods, which produce a less complex data mass, are currently more suitable and affordable for comparative expression studies of larger numbers of samples. For the mouse it has been estimated that 40 million RNA-Seq reads are required to achieve coverage of most transcripts and to allow accurate analysis of expression differences between samples (Mortazavi et al., 2008). In contrast, approximately 2 million transcript-specific DGE tags from mouse were shown to be sufficient to reliably detect low abundant genes (t Hoen et al., 2008). Whereas the quantification of RNA-Seq results may be complicated by unequal coverage of transcripts by sequence reads, the quantitative comparison of DGE tag libraries can build directly on the vast knowledge of statistical evaluation of conventional SAGE experiments (Lash et al., 2000; t Hoen et al., 2008; Zhu et al., 2008).

Unlike some of the RNA-seq protocols, SAGE-derived DGE analysis provides information on transcript directionality. In our study approximately 40% of the mapping events were detected on the antisense strand of UniGene, RefSeq or Ensembl database transcripts. This is comparable to data reported for mouse DGE analysis, where evidence for bidirectional transcription was found for 51% of all detectable UniGene clusters (t Hoen et al., 2008). Similar as in their study, we found no correlation between the abundance of sense and antisense tags for the same transcript. To what extent antisense transcripts might result from transcriptional noise is currently unknown. However, there is accumulating evidence for the widespread occurrence of antisense transcription and its biological relevance (Beiter et al., 2009; Carninci et al., 2005; Katayama et al., 2005). For example, the proper dosage of expression of the human and murine hematopoietic transcription factor PU.1, critical for suppression of leukemia, was shown to rely on antisense RNA modulators (Ebralidze et al., 2008). We found that mycobacterium infection significantly induced both sense and antisense DGE tags of the zebrafish homolog of the *PU.1* gene (*spi1*), suggesting that antisense regulation of the expression of this gene may be an ancient evolutionary mechanism.

The mycobacterium-infected zebrafish analyzed in this study were at the late stage of chronic tuberculosis. As previously observed in microarray analysis, our DGE study showed massive changes in the expression levels of known immune response genes and of genes that have been implicated in human tuberculosis (Meijer et al., 2005). Many of the mycobacterium-regulated genes that we could functionally annotate were homologous to human genes that have been linked with genetic diseases. These included inflammatory and hematologic disorders, for example, anemia, which is commonly associated with tuberculosis (Lee et al., 2006). We also identified many interesting genes that had not been linked to tuberculosis in previous studies. For example, we demonstrated infection-dependent up-regulation of an *ab initio* predicted homolog of olfactomedin 4 (*OLFM4*), also known as granulocyte colony stimulating factor stimulated clone-1 (*hGC-1*). This gene encodes a glycoprotein of unknown function, suggested to be involved in cell adhesion, cancer progression and myeloid development (Chin et al., 2008; Liu et al., 2008). We also observed induction of a specific transcript variant of the zebrafish homolog of *stanniocalcin 1* (*stc1*), a gene suggested to be involved in macrophage chemotaxis and transendothelial migration of inflammatory cells (Chakraborty et al., 2007; Kanellis et al., 2004). Furthermore, by genomic tag mapping we demonstrated induction of a novel alternative splice form of a gene homologous to the human tumor necrosis factor alpha-induced protein 9 gene (TNFAIP9, also known as STEAP4). This gene encodes a six transmembrane putative channel or transporter protein implicated in inflammatory responses and cancer progression (Korkmaz et al., 2005; Wellen et al., 2007). The above mentioned genes are merely given as examples to illustrate how our deep sequencing-based systems approach can quickly lead from mapping of total transcriptome responses to detailed functional annotation. More importantly, the entire description of all mycobacterial regulated genes provides an unbiased basis for subsequent candidate gene approaches that was not possible with prior microarray-based transcriptome profiling. In future studies we will use deep sequencing technology to compare the present data set of the late stage of tuberculosis with earlier stages of disease progression, including granuloma formation in the zebrafish embryo model that is more amenable to experimental manipulation (Lesley and Ramakrishnan, 2008).

5. Conclusions

We have demonstrated here that deep sequencing analysis, using Solexa/Illumina's digital gene expression (DGE) system, provides a robust, sensitive and unbiased alternative to microarray analysis with major advantages for detection of the broad range of transcriptional responses that occur during the process of an infectious disease, as exemplified in this study using the zebrafish-mycobacterium model. First, by mapping of DGE sequence tags to transcript databases we showed that mycobacterium infection induced quantitative changes in the expression levels of many genes, including those previously implicated in human tuberculosis. These data showed over 90% concordance with a reference set of mycobacterium-regulated genes that had previously been confirmed in a multi-platform microarray study. Second, by genomic mapping of the DGE sequence tags we could reveal transcriptional responses that microarray analysis would have failed to detect, such as the switching between alternative transcript isoforms, the expression of novel splice products not present in the current transcript and EST databases, and a high level of antisense transcription. Our DGE study substantiates recent RNA sequencing results in other model species indicating a much larger extent of genome transcription than previously thought. Furthermore, it demonstrates the advantages of a deep sequencing approach for gene discovery and genome annotation and provides new leads for functional studies of candidate genes involved in host–pathogen interaction.

Acknowledgements

We thank Peter-Bram 't Hoen (Leiden University Medical Center, the Netherlands) for helpful discussions, Astrid van der Sar and Wilbert Bitter (Vrije Universiteit Medical Center, Amsterdam) for the *M. marinum* E11 strain, and Susan Wijting and Ulrike Nehrdich for help with infection experiments. This work was supported by the European Commission 6th framework project ZF-TOOLS (LSHG-CT-2006-037220).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.molimm.2009.07.002.

References

Aroian, R., van der Goot, F.G., 2007. Pore-forming toxins and cellular non-immune defenses (CNIDs). Curr. Opin. Microbiol. 10, 57–61.

Beiter, T., Reich, E., Williams, R.W., Simon, P., 2009. Antisense transcription: a critical look in both directions. Cell. Mol. Life Sci. 66, 94–112.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J., Corcoran, K., 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. 18, 630–634.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al., 2005. The transcriptional landscape of the mammalian genome. Science 309, 1559–1563.

Chakraborty, A., Brooks, H., Zhang, P., Smith, W., McReynolds, M.R., Hoying, J.B., Bick, R., Truong, L., Poindexter, B., Lan, H., Elbjeirami, W., Sheikh-Hamad, D., 2007. Stanniocalcin-1 regulates endothelial gene expression and modulates transendothelial migration of leukocytes. Am. J. Physiol. Renal Physiol. 292, F895–F904.

Chin, K.L., Aerbajinai, W., Zhu, J., Drew, L., Chen, L., Liu, W., Rodgers, G.P., 2008. The regulation of OLFM4 expression in myeloid precursor cells relies on NF-kappaB transcription factor. Br. J. Haematol. 143, 421–432.

Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., Robertson, A.J., Perkins, A.C., Bruce, S.J., Lee, C.C., Ranade, S.S., Peckham, H.E., Manning, J.M., McKernan, K.J., Grimmond, S.M., 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods 5, 613–619.

Cloonan, N., Grimmond, S.M., 2008. Transcriptome content and dynamics at single-nucleotide resolution. Genome Biol. 9, 234.

Diebold, B.A., Bokoch, G.M., 2001. Molecular basis for Rac2 regulation of phagocyte NADPH oxidase. Nat. Immunol. 2, 211–215.

Ebralidze, A.K., Guibal, F.C., Steidl, U., Zhang, P., Lee, S., Bartholdy, B., Jorda, M.A., Petkova, V., Rosenbauer, F., Huang, G., Dayaram, T., Klupp, J., O'Brien, K.B., Will, B., Hoogenkamp, M., Borden, K.L., Bonifer, C., Tenen, D.G., 2008. PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. Genes Dev. 22, 2085–2092.

Ellis, T.J., Wagner, G.F., 1995. Post-transcriptional regulation of the stanniocalcin gene by calcium. J. Biol. Chem. 270, 1960–1965.

Harbers, M., Carninci, P., 2005. Tag-based approaches for transcriptome research and genome annotation. Nat. Methods 2, 495–502.

Hashimoto, S., Qu, W., Ahsan, B., Ogoshi, K., Sasaki, A., Nakatani, Y., Lee, Y., Ogawa, M., Ametani, A., Suzuki, Y., Sugano, S., Lee, C.C., Nutter, R.C., Morishita, S., Matsushima, K., 2009. High-resolution analysis of the 5′-end transcriptome using a next generation DNA sequencer. PLoS ONE 4, e4108.

He, Z., Bateman, A., 2003. Progranulin (granulin-epithelin precursor, PC-cell-derived growth factor, acrogranin) mediates tissue repair and tumorigenesis. J. Mol. Med. 81, 600–612.

Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S.C., Hoffman, E., Jedlicka, A.E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S.Q., Yu, W., 2005. Multiple-laboratory comparison of microarray platforms. Nat. Methods 2, 345–350.

Jayachandran, R., Sundaramurthy, V., Combaluzier, B., Mueller, P., Korf, H., Huygen, K., Miyazaki, T., Albrecht, I., Massner, J., Pieters, J., 2007. Survival of mycobacteria in macrophages is mediated by coronin 1-dependent activation of calcineurin. Cell 130, 37–50.

Kanellis, J., Bick, R., Garcia, G., Truong, L., Tsao, C.C., Etemadmoghadam, D., Poindexter, B., Feng, L., Johnson, R.J., Sheikh-Hamad, D., 2004. Stanniocalcin-1, an inhibitor of macrophage chemotaxis and chemokinesis. Am. J. Physiol. Renal Physiol. 286, F356–F362.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., Suzuki, H., Carninci, P., Hayashizaki, Y., Wells, C., Frith, M., Ravasi, T., Pang, K.C., Hallinan, J., Mattick, J., Hume, D.A., Lipovich, L., Batalov, S., Engstrom, P.G., Mizuno, Y., Faghihi, M.A., Sandelin, A., Chalk, A.M., Mottagui-Tabar, S., Liang, Z., Lenhard, B., Wahlestedt, C., 2005. Antisense transcription in the mammalian transcriptome. Science 309, 1564–1566.

Kaul, D., 2008. Coronin-1A epigenomics governs mycobacterial persistence in tuberculosis. FEMS Microbiol. Lett. 278, 10–14.

Kim, J.B., Porreca, G.J., Song, L., Greenway, S.C., Gorham, J.M., Church, G.M., Seidman, C.E., Seidman, J.G., 2007. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science 316, 1481–1484.

Korkmaz, C.G., Korkmaz, K.S., Kurys, P., Elbi, C., Wang, L., Klokk, T.I., Hammarstrom, C., Troen, G., Svindland, A., Hager, G.L., Saatcioglu, F., 2005. Molecular cloning and characterization of STAMP2, an androgen-regulated six transmembrane protein that is overexpressed in prostate cancer. Oncogene 24, 4934–4945.

Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., Altschul, S.F., 2000. SAGEmap: a public gene expression resource. Genome Res. 10, 1051–1060.

Leclerc, P., Biarc, J., St-Onge, M., Gilbert, C., Dussault, A.A., Laflamme, C., Pouliot, M., 2008. Nucleobindin co-localizes and associates with cyclooxygenase (COX)-2 in human neutrophils. PloS ONE 3, e2229.

Lee, S.W., Kang, Y.A., Yoon, Y.S., Um, S.W., Lee, S.M., Yoo, C.G., Kim, Y.W., Han, S.K., Shim, Y.S., Yim, J.J., 2006. The prevalence and evolution of anemia associated with tuberculosis. J. Korean Med. Sci. 21, 1028–1032.

Lesley, R., Ramakrishnan, L., 2008. Insights into early mycobacterial pathogenesis from the zebrafish. Curr. Opin. Microbiol. 11, 277–283.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., Ecker, J.R., 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133, 523–536.

Liu, W., Liu, Y., Zhu, J., Wright, E., Ding, I., Rodgers, G.P., 2008. Reduced hGC-1 protein expression is associated with malignant progression of colon carcinoma. Clin. Cancer Res. 14, 1041–1049.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18, 1509–1517.

Martin, M., Del Valle, J.M., Saborit, I., Engel, P., 2005. Identification of Grb2 as a novel binding partner of the signaling lymphocytic activation molecule-associated protein binding receptor CD229. J. Immunol. 174, 5977–5986.

Meijer, A.H., Verbeek, F.J., Salas-Vidal, E., Corredor-Adamez, M., Bussman, J., van der Sar, A.M., Otto, G.W., Geisler, R., Spaink, H.P., 2005. Transcriptome profiling of adult zebrafish at the late stage of chronic tuberculosis due to *Mycobacterium marinum* infection. Mol. Immunol. 42, 1185–1203.

Moldes, M., Lasnier, F., Gauthereau, X., Klein, C., Pairault, J., Feve, B., Chambaut-Guerin, A.M., 2001. Tumor necrosis factor-alpha-induced adipose-related protein (TIARP), a cell-surface protein that is highly induced by tumor necrosis factor-alpha and adipose conversion. J. Biol. Chem. 276, 33938–33946.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., Marra, M., 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. BioTechniques 45, 81–94.

Morozova, O., Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255–264.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344–1349.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40, 1413–1415.

Pedotti, P., t Hoen, P.A., Vreugdenhil, E., Schenk, G.J., Vossen, R.H., Ariyurek, Y., de Hollander, M., Kuiper, R., van Ommen, G.J., den Dunnen, J.T., Boer, J.M., de Menezes, R.X., 2008. Can subtle changes in gene expression be consistently detected with different microarray platforms? BMC. Genomics 9, 124.

Pettersson, M., Bessonova, M., Gu, H.F., Groop, L.C., Jonsson, J.I., 2000. Characterization, chromosomal localization, and expression during hematopoietic differentiation of the gene encoding Arl6ip, ADP-ribosylation-like factor-6 interacting protein (ARL6). Genomics 68, 351–354.

Pieters, J., 2008. *Mycobacterium tuberculosis* and the macrophage: maintaining a balance. Cell Host Microbe 3, 399–407.

Recalcati, S., Invernizzi, P., Arosio, P., Cairo, G., 2008. New functions for an iron storage protein: the role of ferritin in immunity and autoimmunity. J. Autoimmun. 30, 84–89.
Rosenkranz, R., Borodina, T., Lehrach, H., Himmelbauer, H., 2008. Characterizing the mouse ES cell transcriptome with Illumina sequencing. Genomics 92, 187–194.
Shendure, J., 2008. The beginning of the end for microarrays? Nat. Methods 5, 585–587.
Stockhammer, O.W., Zakrzewska, A., Hegedus, Z., Spaink, H.P., Meijer, A.H., 2009. Transcriptome profiling and functional analyses of the zebrafish embryonic innate immune response to Salmonella infection. J. Immunol. 182, 5641–5653.
Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., Yaspo, M.L., 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321, 956–960.
Swaim, L.E., Connolly, L.E., Volkman, H.E., Humbert, O., Born, D.E., Ramakrishnan, L., 2006. *Mycobacterium marinum* infection of adult zebrafish causes caseating granulomatous tuberculosis and is moderated by adaptive immunity. Infect. Immun. 74, 6108–6117.
t Hoen, P.A., Ariyurek, Y., Thygesen, H.H., Vreugdenhil, E., Vossen, R.H., de Menezes, R.X., Boer, J.M., van Ommen, G.J., den Dunnen, J.T., 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. 36, e141.
van der Sar, A.M., Abdallah, A.M., Sparrius, M., Reinders, E., Vandenbroucke-Grauls, C.M., Bitter, W., 2004. *Mycobacterium marinum* strains can be divided into two distinct types based on genetic diversity and virulence. Infect. Immun. 72, 6306–6312.
van der Sar, A.M., Spaink, H.P., Zakrzewska, A., Bitter, W., Meijer, A.H., 2009. Specificity of the zebrafish host transcriptome response to acute and chronic mycobacterial infection and the role of innate and adaptive immune components. Mol. Immunol. 46, 2317–2332.
Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., 1995. Serial analysis of gene expression. Science 270, 484–487.
Vos, J.B., Datson, N.A., Rabe, K.F., Hiemstra, P.S., 2007. Exploring host–pathogen interactions at the epithelial surface: application of transcriptomics in lung biology. Am. J. Physiol. Lung Cell Mol. Physiol. 292, L367–L377.
Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B., 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470–476.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
Wellen, K.E., Fucho, R., Gregor, M.F., Furuhashi, M., Morgan, C., Lindstad, T., Vaillancourt, E., Gorgun, C.Z., Saatcioglu, F., Hotamisligil, G.S., 2007. Coordinated regulation of nutrient and inflammatory responses by STAMP2 is essential for metabolic homeostasis. Cell 129, 537–548.
Wilhelm, B.T., Landry, J.R., 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48, 249–257.
Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., Bahler, J., 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453, 1239–1243.
Yelin, R., Dahary, D., Sorek, R., Levanon, E.Y., Goldstein, O., Shoshan, A., Diber, A., Biton, S., Tamir, Y., Khosravi, R., Nemzer, S., Pinner, E., Walach, S., Bernstein, J., Savitsky, K., Rotman, G., 2003. Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. 21, 379–386.
Zhu, J., He, F., Wang, J., Yu, J., 2008. Modeling transcriptome based on transcript-sampling data. PLoS ONE 3, e1659.