Accurate Diagnostics for *Bovine tuberculosis* Based on High-Throughput Sequencing

Alexander Churbanov[1]*, Brook Milligan[2]

[1] Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, Beijing, China, [2] Biology Department, New Mexico State University, Las Cruces, New Mexico, United States of America

Abstract

Background: Bovine tuberculosis (bTB) is an enduring contagious disease of cattle that has caused substantial losses to the global livestock industry. Despite large-scale eradication efforts, bTB continues to persist. Current bTB tests rely on the measurement of immune responses *in vivo* (skin tests), and *in vitro* (bovine interferon-γ release assay). Recent developments are characterized by interrogating the expression of an increasing number of genes that participate in the immune response. Currently used assays have the disadvantages of limited sensitivity and specificity, which may lead to incomplete eradication of bTB. Moreover, bTB that reemerges from wild disease reservoirs requires early and reliable diagnostics to prevent further spread. In this work, we use high-throughput sequencing of the peripheral blood mononuclear cells (PBMCs) transcriptome to identify an extensive panel of genes that participate in the immune response. We also investigate the possibility of developing a reliable bTB classification framework based on RNA-Seq reads.

Methodology/Principal Findings: Pooled PBMC mRNA samples from unaffected calves as well as from those with disease progression of 1 and 2 months were sequenced using the Illumina Genome Analyzer II. More than 90 million reads were splice-aligned against the reference genome, and deposited to the database for further expression analysis and visualization. Using this database, we identified 2,312 genes that were differentially expressed in response to bTB infection ($p<10^{-8}$). We achieved a bTB infected status classification accuracy of more than 99% with split-sample validation on newly designed and learned mixtures of expression profiles.

Conclusions/Significance: We demonstrated that bTB can be accurately diagnosed at the early stages of disease progression based on RNA-Seq high-throughput sequencing. The inclusion of multiple genes in the diagnostic panel, combined with the superior sensitivity and broader dynamic range of RNA-Seq, has the potential to improve the accuracy of bTB diagnostics. The computational pipeline used for the project is available from http://code.google.com/p/bovine-tb-prediction.

Citation: Churbanov A, Milligan B (2012) Accurate Diagnostics for *Bovine tuberculosis* Based on High-Throughput Sequencing. PLoS ONE 7(11): e50147. doi:10.1371/journal.pone.0050147

Editor: Matthew Baylis, University of Liverpool, United Kingdom

Received December 29, 2011; **Accepted** October 22, 2012; **Published** November 30, 2012

Funding: This study was supported by grant 2011Y1SA09 from the Chinese Academy of Sciences Fellowship for Young International Scientists, by grant 31150110466 from the National Natural Science Foundation of China (NSFC), and in part by NSF award DBI-0821806. No additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: alexander@big.ac.cn

Introduction

Bovine tuberculosis (bTB) is an insidious, progressive disease of livestock that has cost the United States livestock industry millions of dollars in losses prior to and since the establishment of a national eradication campaign in 1917 [1]. Despite this large-scale eradication effort, bTB is a reemerging infectious disease in the U.S. It is endemic in select areas of Michigan and recent outbreaks have occurred in Minnesota, California, and New Mexico.

Mycobacterium bovis, the causative agent of bovine tuberculosis, creates significant problems for agriculture at both the state and national levels. From a management and animal health perspective, it is essential that infected animals are reliably detected and removed to prevent the spread of the disease. Current diagnostic tests are primarily based on immune responses to crude protein extracts from *M. bovis* (PPDb) injected intradermally. Three days after injection of PPDb, excessive swelling at the injection site indicates that the animal may be infected with *M. bovis*.

The sensitivity (*Se*) and specificity (*Sp*) of the single intradermal test (SIT) depends on a cut-off value, and there is an inverse relationship between test *Se* and *Sp* values [2]. For example, the SIT test *Se* could be as high as 91.2%, with an *Sp* of only 75.5% [3], or the *Se* could be only 63.2%, with the *Sp* as high as 99.0% [4], depending on the cut-off. The *Sp* of the tuberculin skin test can be reduced by exposure to environmental non-tuberculous mycobacteria such as *M. avium* and *M. avium subsp. paratuberculosis* [5]. This reduction in *Sp* is due to immunological cross-reactivity between these species. To increase *Sp* while maintaining reasonable *Se*, animals that test positive to the caudal fold (CF) test are tested 60 days later using the Comparative Cervical Test (CCT). The CCT consists of injecting PPDb and a crude protein derivative from *M. avium* (PPDa) at adjacent sites on the neck. Three days later, the swelling at each injection site is compared. If the inflammation at the PPDb injection site is greater than that at the PPDa site, the animal is considered *M. bovis* infected. Conversely, if the swelling at the PPDa site is greater than that at the PPDb site, the animal is considered clinically negative.

Traditional skin testing requires at least 2 animal handling events, one for PPD injection, and another for the evaluation of the test. The need to hold animals for 72 h is a significant disadvantage of PPD testing [6]. The recognition of cytokines and their role in tuberculosis immunology has led to the development of an in vitro assay for bovine interferon-γ (IFN-γ) production [7,8]. The Bovigam^{TM} assay detects IFN-γ released in response to PPDb in a whole-blood culture assay [8,9]. Because the Bovigam^{TM} uses the same antigens as the skin test, it has *Se* similar to the SIT with a slightly lower corresponding *Sp* [2]. The reported *Se* of the use of IFN-γ release as a diagnostic tool was 91.4%, whereas the *Sp* was 86.7% [6]; no significant difference was seen between the reliability of the IFN-γ assay and that of the SIT [2,10,11]. Despite the national programs in Brazil, the limited sensitivity and specificity of current tests do not facilitate complete bTB eradication in many countries [2,12].

Using real-time PCR, it has been reported that the expression of IFN-γ, tumor necrosis factor alpha (TNF-α), inducible nitric oxide synthase (iNOS), and interleukin (IL)-4 by peripheral blood mononuclear cells (PBMCs) increased in response to infection, whereas that of IL-10 decreased. PPDb-stimulated PBMCs from animals in the high-pathology (with lesions in the lungs and associated lymph nodes) group expressed more IFN-γ, TNF-α, iNOS, and IL-4 mRNA than did those from animals in the low-pathology (only had lesions in the head lymph nodes) group at early time points. PBMC expression of the IL-10 gene decreased faster among animals in the high-pathology group, whereas the expression patterns of T-helper (TH) 1 and TH2 cytokines were different among the animals in the high- and low-pathology groups [13]. The maximal difference in expression occurred within the first month after experimental infection. However, over the next 2 months, the IFN-γ responses between the 2 groups reached similar levels. These data suggest that the outcome of disease may be established early after infection. Similar responses were detected in *M. bovis* infected white-tailed deer [14].

Measuring changes in cell products other than IFN-γ after *in vitro* stimulation can yield useful diagnostic assays. For example, an IL-2 receptor A (IL2RA) enzyme-linked immunosorbent assay (ELISA) exhibited a reported sensitivity of 94% and specificity of 98% [15]. ELISA-based and Griess reaction assays were used to determine that changes in TNF-α and iNOS expression in PBMCs exposed to PPDa or PPDb antigen could serve as additional diagnostic indices complementing IFN-γ measurements [16]. The advent of high-throughput functional genomics has facilitated studies based on targeted immunospecific bovine cDNA microarrays to discover changes in the expression levels of hundreds of genes, many of which are cytokines [17–19].

Diagnostics specific for *M. bovis* that can reliably detect early infection are critical for the eradication program. In this study, we explore the possibility of using next-generation sequencing from PBMC mRNA for the purpose of diagnosing bTB. By quantifying the host immunological response to infection by comparing the transcriptome of known infected and uninfected individuals, we can enhance our ability to detect *M. bovis* in agriculturally important species and, in the future, in potential wildlife reservoirs.

Whole transcriptome sequencing technology (RNA-Seq) based on second-generation sequencing platforms, such as the Illumina Genome Analyzer II, have revolutionized the field of transcriptomics [20]. Quantitative PCR (qPCR) has confirmed the accuracy of RNA-Seq in quantifying gene expression levels [21]. RNA-Seq analysis of spike-in RNA controls of known concentrations also confirmed the high fidelity of the novel technique [22]. Compared to microarray platforms, RNA-Seq delivers higher sensitivity, accuracy, and a broader dynamic range in a hypothesis-neutral way that can help elucidate and annotate novel transcripts [20,23,24]. The results of RNA-Seq are highly reproducible, for both technical and biological replicates [21,25].

Table 1. Number of reads.

Sample name	Number of reads	Number of reads mapped
		against reference genome
TCT1	10,421,654	7,616,528 (73.08%)
TCT2	29,316,410	18,324,561 (62.51%)
TCT3	13,881,201	10,143,019 (73.07%)
TCT4	14,512,900	10,645,842 (73.35%)
TCT5	14,606,602	10,627,915 (72.76%)
TCT6	15,189,216	11,071,421 (72.89%)

Number of reads from different pooled transcriptome samples and the total number of reads mapped against the Btau 4.0 reference genome.
doi:10.1371/journal.pone.0050147.t001

In this case-control study, RNA-Seq reads from PBMCs have been splice-aligned against the Btau 4.0 reference genome. We converted the alignment results and Btau 4.0 genome annotation to the general feature format GFF3 and uploaded the results to a MySQL database connected to the Generic Model (GMOD) Generic Genome Browser (GBrowse). Gene expression levels were measured by counting the number of reads mapped against the NCBI annotated gene loci. Based on the fact that RNA-Seq provides highly sensitive measures of absolute and relative gene expression levels, we constructed a probabilistic model for bTB diagnosis. We demonstrated that reliable classification of infected animals could be achieved using only 7,500 reads for each sample.

Results

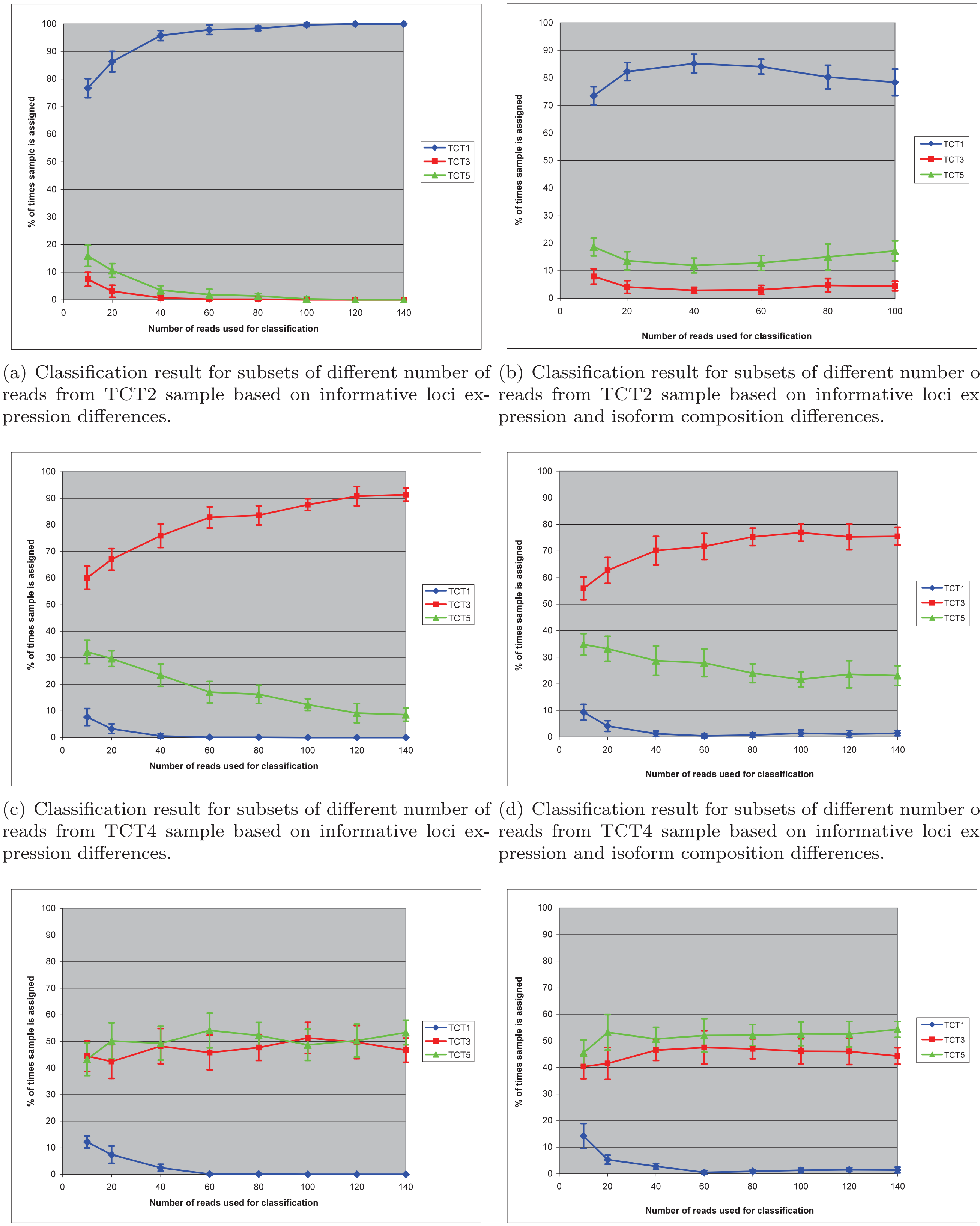
PBMC transcriptome sequencing result

The following numbers of reads were obtained for each pooled transcriptome sample with the Illumina Genome Analyzer II and mapped against the Btau 4.0 reference genome as shown in Table 1. We applied the Fisher exact test to get the list of loci with significant expression changes in response to bTB as presented in Supporting Information S4. We also applied the Fisher exact test to obtain the list of annotated exons with significant coverage changes in response to bTB as presented in Supporting Information S5, which might indicate alternative exon inclusion levels.

Table 2. Number of mapped reads.

Sample name	Number of reads mapped	Total number	Fraction
	to informative loci	of reads	
TCT1	78,434	7,616,528	1.03%
TCT3	137,912	10,143,019	1.36%
TCT5	142,741	10,627,915	1.34%

Number of reads mapped against informative loci.
doi:10.1371/journal.pone.0050147.t002



(a) Classification result for subsets of different number of reads from TCT2 sample based on informative loci expression differences.

(b) Classification result for subsets of different number of reads from TCT2 sample based on informative loci expression and isoform composition differences.

(c) Classification result for subsets of different number of reads from TCT4 sample based on informative loci expression differences.

(d) Classification result for subsets of different number of reads from TCT4 sample based on informative loci expression and isoform composition differences.

(e) Classification result for subsets of different number of reads from TCT6 sample based on informative loci expression differences.

(f) Classification result for subsets of different number of reads from TCT6 sample based on informative loci expression and isoform composition differences.

Figure 1. Classification performance. Classification performance for subsets of RNA-Seq reads from samples corresponding to various bTB post infection periods against control (TCT1), one month progression (TCT3) and two months progression (TCT5) trained profiles.
doi:10.1371/journal.pone.0050147.g001

The number of reads mapped against informative loci listed in Supporting Information S3 are shown in Table 2. As mentioned in the section *ssec:classification results*, our experiments demonstrated that reliable classification could be achieved with 100 or more reads that map against informative loci. Table 2 shows that the fraction of RNA-Seq reads mapped against the informative loci is approximately 1.35%, which translates to 7,500 RNA-Seq reads per sample that are necessary for reliable classification.

Classification results

We used pools TCT1, TCT3, and TCT5 to train probabilistic profiles as described in the subsection *Samples classification*. Samples TCT2, TCT4, and TCT6 were used to estimate classification performance, where we used reads mapping against the informative loci mentioned in Supporting Information S3. From the reads that are known to map to informative loci, we randomly sampled groups of sizes $10,20,40,\ldots,140$ and conducted maximum a posteriori (MAP) classification according to the formulas mentioned in the subsection *Samples classification*. In each category, we formed 10 groups, each containing 100 read sets of sizes in the range $10,\ldots,140$ and reported the means and standard deviations of the classification accuracy in these groups. The results of these classifications are shown in Figures 1(a), 1(c) and 1(e).

Figure 1 compares the performance of 2 classification methods. One method aligns the reads against the profiles to calculate forward probability as discussed in the section *Samples classification*. The performance of the alignment-based method is represented in Figures 1(b), 1(d), and 1(f). Another method, based on a much simpler technique, assigns a constant logarithm of probability to all the reads that map against informative loci, listed in Supporting Information S3, according to the genomic short-read nucleotide alignment program (gsNap). This type of classification uses only gene expression mixture proportions for sample classification, as shown in equation (1). This simpler technique results in improved performance, as can be seen in Figures 1(a), 1(c), and 1(e).

Significant gene expression changes

A heat map of statistically significant gene changes ($p<10^{-30}$), along with their product names, is provided in Supporting Information S2. A histogram of the expression changes for these loci with error bars is provided in Figure 2. The locations and heat map of exons with statistically significant inclusion discrepancies ($p<10^{-8}$) relative to the expression changes of the containing gene locus are provided in Supporting Information S5. Examples of the expression changes of the cytokines IFN-γ and IL-17 in response to bTB, as displayed in GBrowse, are shown in Figure 3.

Materials and Methods

RNA-Seq protocol

Seven Holstein calves were obtained from a TB-free herd and housed at the National Animal Disease Center in a biosafety level 3 facility. All animals were housed and cared for in accordance with institutional policies, and procedures were approved by the Institutional Animal Care and Use Committee. The calves received *M. bovis* strain 95-1315 by aerosol at 6 months of age, as described previously [16].

Blood was collected prior to infection, and at 1 month and 2 months post-infection. PBMCs were isolated after stimulation with PPD for 16 hours and the RNA isolated as previously described [14]. The quality of the RNA was tested using an Agilent 2100 Bioanalyzer using the Agilent RNA 6000 Nano Kit according to the manufacturer's instructions. All RNA samples had a RNA integrity number (RIN) value greater than 7.0. Samples (3.3 μg of RNA) from each animal were randomly assigned to 1 of 2 pools at the 0-, 1-, and 2-month timepoints of disease progression. Two RNA pools were generated for each time point, each containing randomly assigned RNA samples from 3 animals. RNA pools from uninfected animals were designated as TCT1 and TCT2, those from the 1-month progression animals were designated as TCT3 and TCT4, and those from the 2-month progression animals were designated as TCT5 and TCT6. Pooled samples were sent to the Iowa State DNA Facility for library preparation and sequencing (75 base run) on the Illumina Genome Analyzer II (one pooled sample per channel).

Processing the mapped samples

The resulting cDNA reads were splice-aligned against the reference genome Btau 4.0, listed in Supporting Information S6, using the gsNap [26,27] program. The gsNap tool has been cited [28] as one of the most accurate programs for RNA-Seq reads alignment in a splicing-aware fashion. The alignment results were parsed and deposited into a custom-designed MySQL database and then converted to GFF3 format. The Genbank files were parsed using a BioJava [29]-based parser. All the GFF3 files were uploaded to a MySQL database connected to GBrowse. We used a 2×2 Fisher test to compare the number of reads that map against a gene locus, as annotated in the Btau 4.0 reference genome, to the number of reads mapped against the chromosome containing the locus minus the number of reads that map to the locus in a case-control experiment. We also estimated patterns of differential inclusion of exons by comparing the number of reads that map against an exon, as annotated in Btau 4.0 reference genome, to the number of reads that map against the containing locus minus the number of reads mapped against the exon in case-control experiments. In our experiments, we reported statistically significant differences in gene expression and exon inclusion patterns at significance levels of 0.01 or less in the following tests: $TCT1 \Leftrightarrow TCT3$, $TCT1 \Leftrightarrow TCT4$, $TCT1 \Leftrightarrow TCT5$, and $TCT1 \Leftrightarrow TCT6$. The probability that all 4 tests are significant is 10^{-8}.

We show expression changes for genes with significance levels 1×10^{-30} using the heat map built using Bioconductor http://www.bioconductor.org/. The heat map dendrogram shown in Supporting Information S2, was used to identify genes in the top 3 clusters as the most informative classification loci, as listed in Supporting Information S3. These clusters group the most closely related expression profiles having the shortest dendrogram branches. Attempts to use loci in outgroups of these clusters results in suboptimal classification performance in our experiments.

Samples classification

In this work, we introduce a hierarchical mixture model for the classification of transcriptomes based on individual RNA-Seq reads. In our model, differential isoform expression patterns can be modeled with various probabilities of exonic isoforms, as shown in Figure 4. The mixture of profiles for different conditions, as seen in Figure 4, form a hierarchical model based on which we generate a MAP classification based on equation (1).

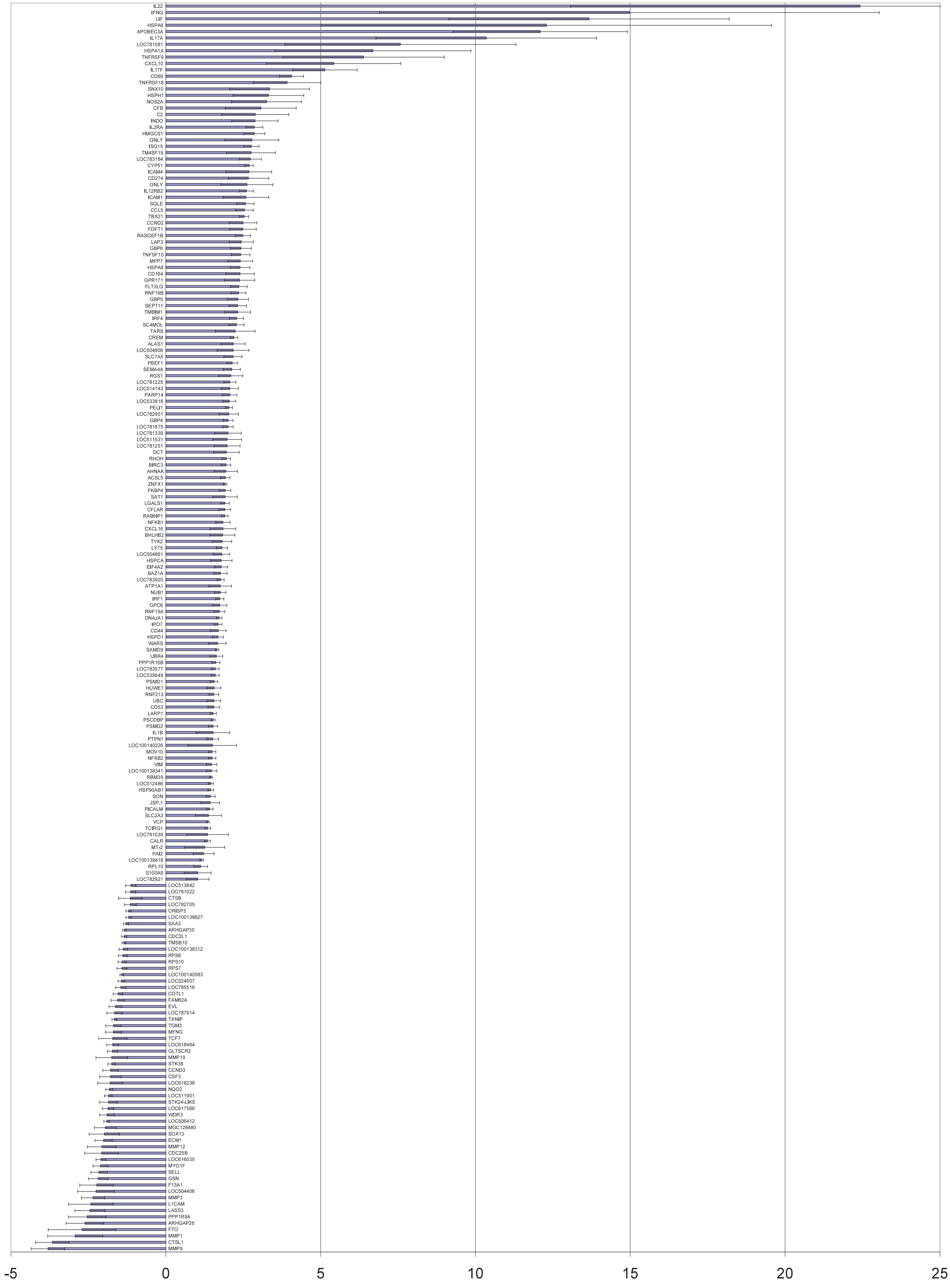


Figure 2. Gene expression changes.
doi:10.1371/journal.pone.0050147.g002

The hidden Markov model (HMM) is a widely accepted stochastic modeling tool [30] used in various domains, such as speech recognition [31] and bioinformatics [32]. HMM is a stochastic finite state machine where each transition between hidden states culminates in the emission of a symbol. The HMM can be represented as a directed graph with $\mathcal{N}$ states where each state can emit either a discrete character or a continuous value drawn from a probability density function (PDF). In order to describe the HMM, we need the following parameters:

- Set of states, we label individual states as $S=\{S_1,S_2,\ldots,S_N\}$, and denote the state visited at time t as $q(t)$,
- Set of PDFs $B=b_j(o)$ from where emission is drawn $b_j(o_t)=P(o_t|q_t=S_j)$, $1\le j\le N$. where o_t is observation at time moment t from the sequence of observations $\mathcal{O}=o_1,o_2,\ldots,o_T$.
- The state-transmission probability matrix $A=a(i,j)$, where $a(i,j)=P(q(t+1)=j|q(t)=i)$,
- The initial state distribution vector $\Pi=\{\pi_1,\ldots,\pi_N\}$.

The set of parameters $\lambda=(\Pi,A,B)$ completely specifies the HMM.

Here we adopt the notation from [33]. We need to calculate the expected probability of being at a certain state at a certain moment in time using a forward-backward procedure.

Forward procedure. By definition $\alpha_t(i)=P(o_1,o_2,\ldots,o_t,x_t=i|\lambda)$ is calculated the following way

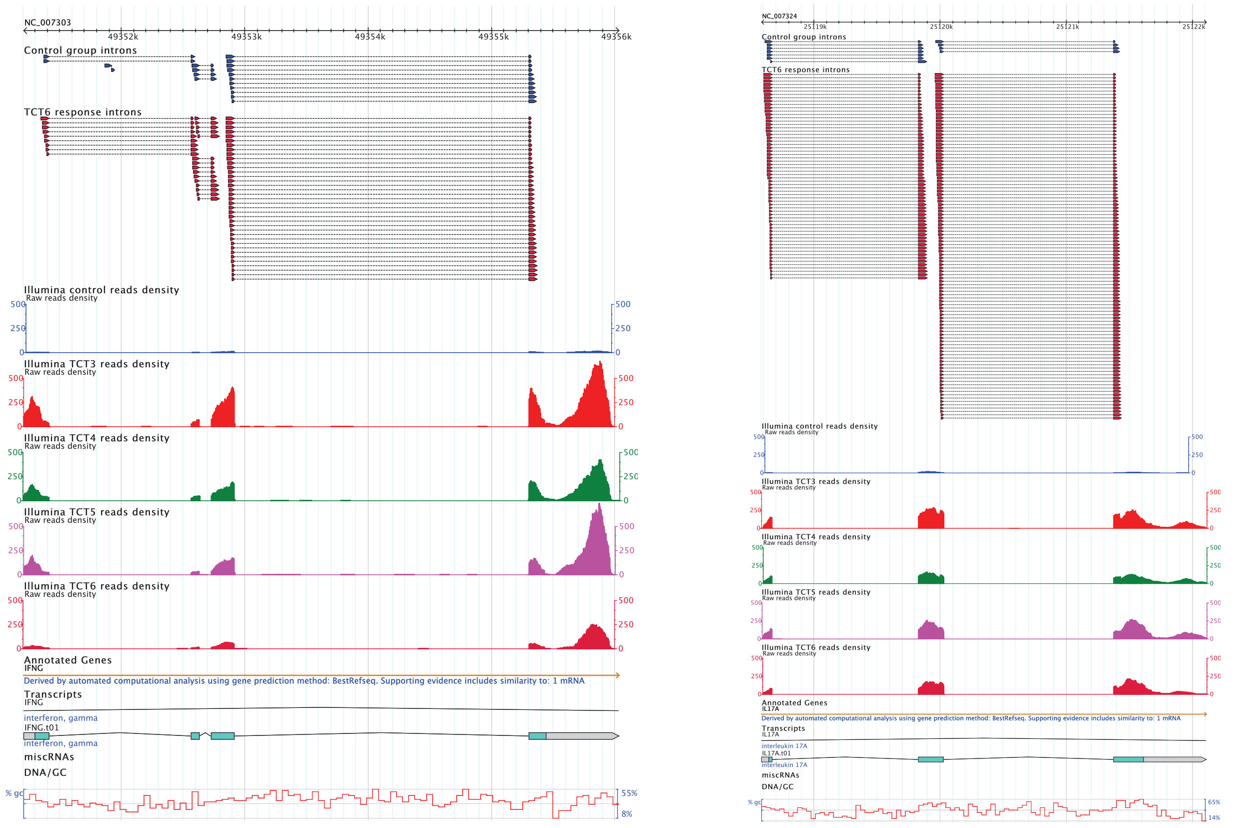
1. Initially $\alpha_1(i)=\pi_i\mathcal{N}(o_1|\Theta_i)$ *for* $1\le i\le N$,
2. $\alpha_t(j)=\left[\sum_{i=1}^{N}\alpha_{t-1}(i)a_{ij}\right]\mathcal{N}(o_t|\Theta_j)$ *for* t = 2,3,…,T and $1\le j\le N$,
3. Finally $P(\mathcal{O}|\lambda)=\sum_{i=1}^{N}\alpha_T(i)$ is the sequence *likelihood* according to the model.

Let us consider a set of K sequences $\mathcal{O}=\{\mathcal{O}_1,\mathcal{O}_2,\ldots,\mathcal{O}_L\}$ to find the likelihood of the mixture shown in Figure 4.

We calculate the likelihoods matrix as follows:

$$\left.\underbrace{\begin{array}{cccc} p(\mathcal{O}_1|\lambda_1) & \ldots & p(\mathcal{O}_1|\lambda_L) & \\ p(\mathcal{O}_2|\lambda_1) & \ldots & p(\mathcal{O}_2|\lambda_L) & \\ p(\mathcal{O}_3|\lambda_1) & \ldots & p(\mathcal{O}_3|\lambda_L) & \%hbrace \\ \ldots & \ldots & \ldots & \\ p(\mathcal{O}_K|\lambda_1) & \ldots & p(\mathcal{O}_K|\lambda_L) & \end{array}}_{L\text{ mixture components}}\right\} K\text{ sequences}$$

Let us define mixture parameters as $\Theta=(A,\Lambda)$ where $A=\{\alpha_1,\alpha_2,\ldots,\alpha_L\}$ and $\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_L\}$



(a) Interferon gamma average normalized expression level increased 22.42 times in response to bTB.

(b) Interleukin 17A average normalized expression level increased 10.35 times in response to bTB.

Figure 3. Example of IFN-γ and IL17A unnormalized gene expression changes in response to bTB. On these GBrowse views we show coverage for mapped RNA-Seq reads along with Illumina short reads spanning across introns, i.e. cDNA reads that that partially map to two different exons thus anchoring the exonic boundaries. Here the control reads are the reads from TCT1 pool.

doi:10.1371/journal.pone.0050147.g003

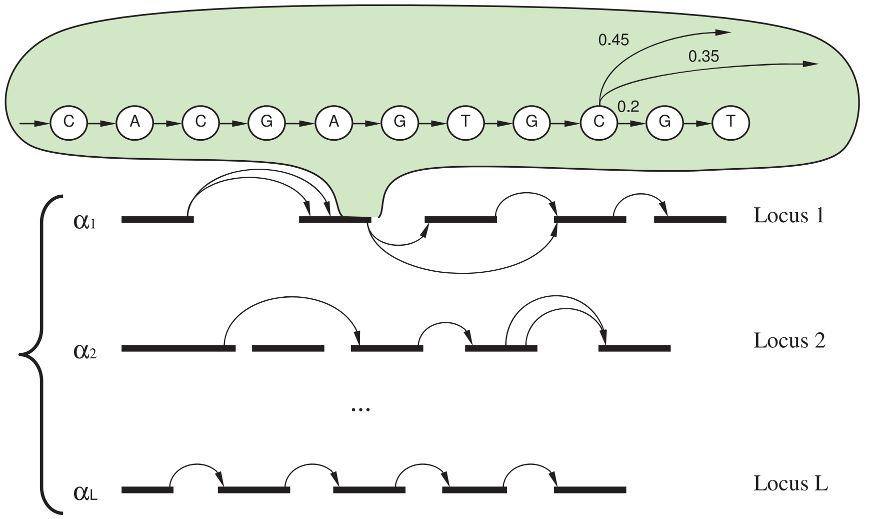


Figure 4. Symbolic representation of classification profiles mixture. Here the mixture components $\alpha_1, \alpha_2, \ldots, \alpha_L$ indicate the fraction of all hits mapping against loci used to build a profile. Transition frequencies across introns match the Illumina coverage density at each particular splice site.
doi:10.1371/journal.pone.0050147.g004

Mixture likelihood is then

$$p(\mathcal{O}|\Theta) = \prod_{k=1}^{K} \sum_{l=1}^{L} \alpha_l \, p(\mathcal{O}_k|\lambda_l).$$

We use Bayes rule to find the posterior probability (responsibility) of a mixture component as shown in Figure 4 with parameters Θ_m and emission sequences $\mathcal{O}$ where $\mathbf{Q} = \{\Theta_1, \Theta_2, \ldots, \Theta_M\}$

$$p(\Theta_m|\mathcal{O}, \mathbf{Q}) = \frac{\beta_m p(\mathcal{O}|\Theta_m)}{\sum_{m=1}^{M} \beta_m p(\mathcal{O}|\Theta_m)}. \quad (1)$$

Discussion

Evidence of immunomodulating response

In this study, we demonstrated that RNA-Seq can be used for the early diagnosis of bTB. We identified an extensive panel of genes undergoing differential expression changes in response to infection, as shown in Figure 2. Many of the genes that show increased expression are cytokines, the immune response modulators. We observed significant expression changes in cytokines, including interleukins (IL-22, IL17A, IL17F, IL1A, IL1B) and interferons (IFNG).

According to a study by Meade *et.al.* [18] based on an immune microarray, 378 genes were differentially expressed at the level $p = 0.05$ in bTB-infected and control animals. A significant proportion of genes (65%) were expressed at lower levels, among these genes are immune response modulators such as TLR2, TLR4, IFNG, IL-2, IL-4, and the bovine major histocompatibility complex proteins BoLA and BoLA-DRA. Suppression of the key genes modulating the immune response was suggested as one of the mechanisms by which bTB survives host immune defenses. A significant increase in the expression of IFN-γ, IL-22, CXCL9, CXCL10, GZMA, and IL17A has been reported based on high-density microarray gene expression profiling of the murine immune response against *M. bovis* infection [19], suggesting the use of the elevated expression of these genes as an additional biomarker for bTB diagnosis ante-mortem.

Similar to this study [19], we observed that the expression of key immune response players such as IFNG, IL-22, IL17A, IL17F, NOS2A, TNF, and IL1A increased, as seen in Figures 3 and 2, and in the Supporting Information S4. The majority of the genes (56%) mentioned in Supporting Information S4 show increased expression, as represented in Figure 2. In this study, we confirm the previously observed [13,16] important roles of IFN-γ (IFNG), TNF-α (TNF), and iNOS (NOS2A) in the bTB immune response. As seen in Figure 2, IL2RA gene expression is significantly higher; this confirms diagnostic utility of this gene as described earlier [15]. Expression of IL2RB is also higher, as seen in Supporting Information S4.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway involved in the bovine TB immune response was

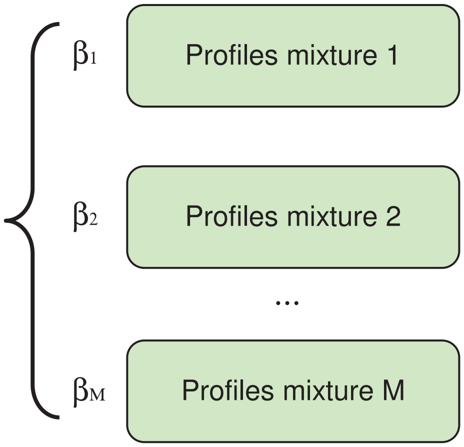


Figure 5. Transcriptome model. Here $\beta_1, \beta_2, \ldots, \beta_M$ are the prior probabilities of cattle being infected based on previous experience. Each profiles mixture has structure as shown in Figure 4.
doi:10.1371/journal.pone.0050147.g005

identified as listed in Supporting Information S1; the majority of immune modulators participating in this pathway were shown in this study to change expression. This study clearly reveals that IL17A responds to bTB infection synergistically with interleukin-1, TNF, iNOS, IFN-γ, IL2RA and other immune modulators, as shown in Supporting Information S1. We used this coordinated expression as a disease signature for early bTB diagnostics based on RNA-Seq technology. RNA-Seq has been previously reported to be a very sensitive and accurate method of evaluating gene expression with a literally unlimited dynamic range. The inherent stochasticity of RNA-Seq reads is convenient for interrogating the expression of multiple loci.

Further studies are needed to determine if naturally infected animals can also be easily classified using RNA-Seq at the early stages of infection, because experimentally infected animals usually have extremely high immune responses to *M. bovis* antigens. In this study we did not investigate if RNA-Seq is more efficient in distinguishing immune responses from *M. avium* vs. those from *M. bovis*. Further investigation is needed to tell if RNA-Seq has any advantage in *Se* and *Sp* compared to the SIT or the in vitro γ-interferon assay. One of the advantages of pooling RNA samples from different animals, as in this study, is the ability to estimate the generalized immune response to bTB infection. However, this approach does not allow estimating differences in the immune reaction that might exist between individual calves, including assessments of the impact of shared sires on the classification results.

Advantages and limitations of the RNA-Seq classification framework

Figures 1(a),1(c), and 1(e) indicate that for reliable classification, we would need 100 or more reads mapping to informative loci, as seen in Supporting Information S3. The performance of the classification based on the forward algorithm, as discussed in the section *Samples classification*, is represented in Figures 1(b), 1(d), and 1(f). The alignment-based classification is less accurate compared to the simple classification based on gene expression proportions shown in Figures 1(a), 1(c), and 1(e). This is explained by the fact that transitions corresponding to exonic isoform frequencies, as seen in Figure 4, have fluctuations associated with a limited number of reads interrupted by introns. These fluctuations generate some random noise in the loci classification profiles. We detected some gene isoform changes associated with the bTB response in the form of differential inclusion of the exons mentioned in Supporting Information S5. Although we did not use these predicted isoform changes in our bTB classification experiments, they might be useful for future developments in improved bTB diagnostics.

As expected, we had limited resolution between the 1-month and 2-month disease progression animals, as shown in Figures 1(c) and 1(e). In the case of the classification of the first month response to bTB, we have been able to achieve reliable classification, with an accuracy of more than 90%, as seen in Figure 1(c). The accuracy of classification of the 2-month progression samples was approximately 50%, mostly non-discriminatory between the first and second month profiles. However, in all rounds of classification, we had reliable resolution between the infected and uninfected animals. It is possible that extending the panel of the informative loci used for classification can further improve performance.

The proposed bTB classification framework based on RNA-Seq reads from the PBMC transcriptome has several advantages. It can easily handle the uneven coverage biases of modern sequencers such as Illumina and SOLID [34], and different lengths of underlying loci. The framework can also easily accommodate sequencing errors and known SNPs. The mixture framework is flexible and can model different numbers of genes. The classification process does not require isoform reconstruction, as in some cases it is impossible to infer isoforms from short cDNA reads [35]; it rather presents different isoforms as independent assortments of alternative exonic isoforms. The transcriptome model, as shown in Figure 5, can incorporate prior believes for optimal performance that can be adjusted based on the number of cattle found infected with bTB in a certain area.

Supporting Information

Supporting Information S1 Genes participating in immune response.
(PDF)

Supporting Information S2 Genes and the product names.
(PDF)

Supporting Information S3 Genes used for classification.
(PDF)

Supporting Information S4 Genes changing expression.
(PDF)

Supporting Information S5 Exons changing inclusion.
(PDF)

Supporting Information S6 Reference sequences used.
(PDF)

Acknowledgments

We would like to thank Tyler Thacker and Ray W. Waters for providing us with many important suggestions that significantly improved quality of the manuscript.

Author Contributions

Conceived and designed the experiments: AC BM. Performed the experiments: AC. Analyzed the data: AC BM. Contributed reagents/materials/analysis tools: BM. Wrote the paper: AC. Designed the software used in analysis: AC.

References

1. Essey MA, Koller MA (1994) Status of bovine tuberculosis in North America. Veterinary Micro-biology 40: 15–22.
2. de la Rua-Domenech R, Goodchild A, Vordermeier H, Hewinson R, Christiansen K, et al. (2006) Ante mortem diagnosis of tuberculosis in cattle: a review of the tuberculin tests, gamma-interferon assay and other ancillary diagnostic techniques. Res Vet Sci 81: 190–210.
3. Francis J, Seiler R, Wilkie W, O'Boyle D, Lumsden M, et al. (1978) The sensitivity and specificity of various tuberculin tests using bovine PPD and other tuberculins. Veterinary Record 103: 420–435.
4. Wood P, Corner L, Rothel J, Baldock C, Jones S, et al. (1991) Field comparison of the interferon-gamma assay and intradermal tuberculin test for the diagnosis of bovine tuberculosis. Australian Veterinary Journal 68: 286–290.
5. Palmer MV, Waters WR (2006) Advances in bovine tuberculosis diagnosis and pathogenesis: what policy makers need to know. Vet Microbiol 112: 181–90.
6. Marassia C, Medeirosa L, Lilenbaum W (2010) The use of a Gamma-Interferon assay to confirm a diagnosis of bovine tuberculosis in Brazil. Acta Tropica 113: 199–201.
7. Wood P, Corner L, Plackett P (1990) Development of a simple, rapid in vitro cellular assay for bovine tuberculosis based on the production of γ interferon. Research in Veterinary Science 49: 46–49.

8. Rothel J, Jones S, Corner L, Cox J, Wood P (1990) A sandwich enzyme-immunoassay for bovine interferon-gamma and its use for the detection of tuberculosis in cattle. Australian Veterinary Journal 67: 134–137.
9. Wood P, Jones S (2001) BOVIGAM^{TM}: an in vitro cellular diagnostic test for bovine tuberculosis. Tuberculosis 81: 147–155.
10. Ameni G, Miörner H, Roger F, Tibbo M (2000) Comparison between comparative tuberculin and gamma-interferon tests for the diagnosis of Bovine tuberculosis in ethiopia. Tropical Animal Health and Production 32: 267–276.
11. Antognolia M, Remmengaa M, Bengtsona S, Clarka H, Orloskib K, et al. (2011) Analysis of the diagnostic accuracy of the gamma interferon assay for detection of bovine tuberculosis in U.S. herds. Preventive Veterinary Medicine 101: 35–41.
12. Medeiros L, Marassi C, Figueiredo E, Lilenbaum W (2010) Potential application of new diagnostic methods for controlling *bovine* tuberculosis in Brazil. Brazilian Journal of Microbiology.
13. Thacker T, Palmer M, Waters W (2007) Associations between cytokine gene expression and pathology in *Mycobacterium bovis* infected cattle. Veterinary Immunology and Immunopathology 119: 204–213.
14. Thacker TC, Palmer MV, Waters WR (2009) T-cell mRNA expression in response to *Mycobacterium bovis* BCG vaccination and *Mycobacterium bovis* infection of white-tailed deer. U.S. Patent 8. Clinical and vaccine immunology 16: 1139–1145.
15. O'Nuallain E, Davis W, Costello E, Pollock J, Monaghan M (1997) Detection of *Mycobacterium bovis* infection in cattle using an immunoassay for bovine soluble interleukin-2 receptor-a (sIL-2R-a) produced by peripheral blood T-Lymphocytes following incubation with tuberculin PPD. Veterinary Immunology and Immunopathology 56: 65–76.
16. Waters W, Palmer M, Whipple D, Carlson M, Nonnecke B (2003) Diagnostic implications of antigen-induced Gamma Interferon, Nitric Oxide, and Tumor Necrosis Factor Alpha production by peripheral blood mononuclear cells from *Mycobacterium bovis*-infected cattle. Clinical and Diagnostic Laboratory Immunology 10: 960–966.
17. Meade K, Gormley E, O'Farrelly C, Park S, Costello E, et al. (2008) Antigen stimulation of peripheral blood mononuclear cells from *Mycobacterium bovis* infected cattle yields evidence for a novel gene expression program. BMC Genomics 9.
18. Meade K, Gormley E, Doyle M, Fitzsimons T, O'Farrelly C, et al. (2007) Innate gene repression associated with *Mycobacterium bovis* infection in cattle: toward a gene signature of disease. BMC Genomics 8.
19. Aranday-Cortes E, Hogarth P, Kaveh D, Whelan A, Villarreal-Ramos B, et al. (2012) Transcriptional profiling of disease-induced host responses in bovine tuberculosis and the identification of potential diagnostic biomarkers. PlOS ONE 7: e30626.
20. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev Gen 10: 57–63.
21. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.
22. Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.
23. Bradford J, Hey Y, Yates T, Li Y, Pepper S, et al. (2010) A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. BMC Genomics 11: 282.
24. Pan Q, Shai O, Lee L, Frey B, Blencow B (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 40: 1413–1415.
25. Cloonan N, Forrest A, Kolle G, Gardiner B, Faulkner G, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613–619.
26. Wu T, Watanabe C (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21: 1859–1875.
27. Wu T, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873–881.
28. Grant G, Farkas M, Pizarro A, Lahens N, Schug J, et al. (2011) Comparative analysis of rna-seq alignment algorithms and the rna-seq unīed mapper (rum). Bioinformatics 27: 2518–2528.
29. Holland R, Down T, Pocock M, Prlić A, Huen D, et al. (2008) BioJava: an open-source framework for bioinformatics. Bioinformatics 24: 2096–2097.
30. Bilmes J (2002) What HMMs can do. Technical report, University of Washington, Seattle.
31. Rabiner L, Juang BH (1993) Fundamentals of speech recognition. Printice Hall.
32. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge University press.
33. Rabiner L (1989) A tutorial on hidden markov models and selected applications in speach recognition. Proceedings of IEEE 77: 257–286.
34. Ozsolak F, Milos P (2011) RNA sequencing: advances, challenges and opportunities. Nature reviews genetics 12: 87–98.
35. Lacroix V, Sammeth M, Guigo R, Bergeron A (2008) Exact transcriptome reconstruction from short sequence reads. In: WABI '08 Proceedings of the 8th international workshop on Algorithms in Bioinformatics. Springer-Verlag Berlin, Heidelberg.