The Exploration of *Brucella* Transcriptome, from the ORFeome to RNAseq

5

Juan M. García-Lobo, María C. Rodríguez, Asunción Seoane, Félix J. Sangari and Ignacio López-Goñi

Abstract

In this chapter we will analyse the results available on the characterization of the *Brucella* transcriptome. After a summary of earlier work on transcription, two technical approaches will be mainly described, on one side the use microarrays, specially that derived from the *Brucella* ORFeome that allows hybridization with mRNA derived cDNA to determine the relative abundance of transcripts from each *B. melitensis* ORF. On the other, RNAseq, consisting in the massive sequencing of cDNA libraries derived from mRNA obtained from *B. abortus* grown in culture medium. Sequencing with the Illumina Genome-Analyser II platform<Please provide manufacturer name and address> produced 3 millions of 35-nt-long reads that annealed with single copy coding regions of the genome. This allowed a good coverage for every CDS and produced a new data set on the transcription of *Brucella*. We obtained a good correlation for the set of highly expressed genes from the microarrays and confirmed the observations obtained on the asymmetry between chromosome transcription. Preliminary conclusions on intracellular transcription have been drawn from real-time polymerase chain reaction (RT-PCR) on selected candidate genes and from microarray data sets obtained from virulence related conditions. The RNAseq derived data allowed more versatile data mining, giving some new details on transcription from pseudogenes or intergenic regions.

Introduction

During the last 50 years, to understand the mechanisms working in bacterial cells, we have applied a reductionist approach consisting in the analysis of isolated parts of the organism at different complexity levels ranging from the genetics to the three-dimensional determination of complex molecular structures. This approach has been really fruitful and undoubtedly it will be our principal source of information in the near future. Newer approaches to the study of biological systems want to analyse living organisms as a complex unit by looking at all the components of the organism as well as the interactions among them that result in an observable behaviour at the same time. This methodology is the hallmark of the field of systems biology. These procedures are being applied already into microbiology. Perhaps, the most appealing case has been the analysis of *Mycoplasma*, the free-living bacteria with the smallest genome (Ochman and Raghavan, 2009). After a lot of genomic data on *Mycoplasma* were available, a multinational team has analysed in depth the transcriptome of the bacteria using both, RNA sequencing and tiling microarrays (Güell *et al.*, 2009). They have also studied the proteome of the bacteria, including protein interactions (Kühner *et al.*, 2009) and finally they have reconstructed the metabolism by analysing any metabolic pathway integrating transcriptional, proteomic and metabolic flux data sets (Yus *et al.*, 2009). This piece of work signals the path for the study of more complex organisms using the same methodologies. Perhaps, the most important conclusion to take into account is that some classic paradigms of molecular microbiology are changing. This is mainly the case for bacterial transcription. The analysis of the transcriptome of *Mycoplasma* has revealed extensive transcription from both DNA strands of the same DNA, alternative transcripts from operons including only some of the operon genes, and existence of many short RNA's usually non-detected by automatic annotation pipelines.

As a desirable objective we should consider performing a similar study on *Brucella*. Taking into account genome size only, we expect a much more complex problem to solve. While this effort is organized, the progressive application of high throughput technologies to *Brucella* may build a set of data that could be later used to sketch the blueprint of our beloved bacteria.

Early transcriptional analysis

Application of molecular analysis to genetic information flow in *Brucella* started near 30 years ago. To our knowledge the first paper published on this field (that somehow inspired work of our group) was a description of mutagenesis produced by transposon Tn5 in *B. abortus* published in 1987 (Smith and Heffron, 1987). The first nucleotide sequences of *Brucella* rRNA were deposited in GenBank in 1989.

Focus on transcription started soon after, targeting first surface antigen genes with a clear interest in the improvement of brucellosis vaccines and diagnostic tests. For instance, the analysis of the 36-kDa outer membrane protein gene of *B. abortus*, included the detection of promoter activity by fusion to LacZ, identification of Shine-Dalgarno sequences, and rho-independent terminators (Ficht *et al.*, 1989).

Besides the transcriptional analysis of a few dozens of individual genes performed in the following years, a first experiment addressing global analysis of transcription was performed by R.C. Essemberg, who made a library of *Sau*3AI generated DNA fragments fused to a promoter-less luciferase reporter and studied transcription arising from the cloned fragments and the effect that different sugars produced on the transcription levels. In this way fragments containing promoters with activity dependent on erythritol, glucose, galactose and succinate were detected. This work was not published, but was communicated at the Chicago Brucellosis meeting and the sequence of the cloned fragments was deposited in GenBank (accession numbers AF075168, AF072121, AF072119, AF074323, AF073884, AF072580, AF072569, AF072120, AF072570, AF072571, AF072572, AF072573, AF072574, AF072575, AF072576, AF072577, AF072578, AF072579).

Methodological overview of massive transcriptomic analysis

The *Brucella* ORFeome

The ORFeome of *Brucella* is an ordered collection of plasmids devised to contain all the ORF's from the sequenced genome of *B. melitensis* 16M. The available annotation was refined to correct the start or the end points of the ORF's according to the best homology with known database products and proximity of Shine-Dalgarno sequences near ORF 5′ end. 908 ORF's limits were corrected out of 3198 ORF's examined. Amplification primers designed for each ORF containing a specific sequence from the ORF and a constant part containing *attB1* and *attB2* sequences at the 5′ and 3′ end of the ORF respectively. This strategy produced ds DNA segments containing the complete ORF flanked by constant sequences that served as substrate for site-specific recombination reactions for the process known as recombinational cloning. The 3198 amplification products were introduced into the Gateway compatible vector pDONR201 and transformed into *E. coli* DH5α. 3091 clones (96.7%) were successfully demonstrated to contain the appropriate ORF by sequencing (Dricot *et al.*, 2004). The genomic data available on *Brucella* have shown than more than 90% of the genes are common (over 99% identity) between different species indicating that the *Brucella* ORFeome can be used as a general tool for high throughput studies in the genus *Brucella*.

Brucella microarrays

One of the uses of the *Brucella* ORFeome was the construction of a microarray containing all the *Brucella* ORF's printed on a glass slide. The ORFeome was used to isolate plasmid DNA from each clone and every insert was PCR amplified using a single pair of primers containing the *attB1* and *attB2* sequences flanking each ORF in the ORFeome. Each DNA was spotted by duplicate and appropriate positive and negative controls were spiked over the whole microarray. As a positive control we used a gene from *Brucella* that was considered to be constitutively expressed, namely the IF-1 gene (Eskra *et al.*, 2001; Hernández-Castro *et al.*, 2003). As negative controls we included 128 spots containing an *Arabidopsis thaliana* gene (*porB*, encoding NADPH-protochlorophyllide oxidoreductase), known not to hybridize with *Brucella* DNA, and another 128 spots containing spotting buffer but no DNA. Microarrays were validated by ethidium bromide staining showing that they were uniform and could be used to analyse transcription in *Brucella*. Total *B. abortus* 2308 RNA was prepared from 10 ml of cells growing to mid stationary phase in *Brucella* Broth (Pronadisa<Please provide manufacturer address>) with the RNeasy mini System (Quiagen<Please provide manufacturer address>). RNA was depleted of DNA by repeated treatment with DNAse and enriched in messenger RNA by subtraction of rRNA with the MicrobExpress kit (Ambion<Please provide manufacturer address>). This enriched mRNA was amplified with the Message AmpII-bacteria amplification kit (Ambion) and labelled with Cy3 fluorescent dye. 10 μg of this labelled aRNA was used to hybridize the microarrays (Viadas *et al.*, 2009).

This ORFeome derived microarray was used first to analyse global transcription in *B.* abortus 2308 grown in laboratory conditions and to evaluate what genes were being expressed or remained silent in these conditions. The same microarray was later used to analyse the function of the two component regulatory system *bvrR/S* (Viadas *et al.*, 2010). and the effect of erythritol on transcription in *Brucella*.

In addition to the ORFeome microarray, at least two more oligonucleotide based microarrays have been used to analyse the *Brucella* transcriptome. A microarray was built from 8768 70-mer unique synthetic oligonuleotides derived from *B. melitensis* ORFs and additional ORFs from *B. suis*. This microarray was hybridized with cDNA derived from total RNA and with genomic DNA as an internal control (Rossetti *et al.*, 2009). Another oligonucleotide based microarray constructed by NimbleGen Systems <Please provide manufacturer name and address>contained 20 24-mer oligonucleotides per CDS, and has also been used to investigate transcriptional regulation (Uzureau *et al.*, 2010, discussed below).

RNA seq: preparation of samples

The more recent approach to study the transcriptome of *Brucella* takes advantage of new generation sequencing systems. RNAseq is a method consisting in the massive sequencing of RNA from a cell, usually after conversion into the appropriate cDNA library. The Illumina platform is especially well suited to this aim since it is capable to produce 10^9 reads per run. Size of reads is not as crucial in RNAseq as is in genome sequencing, and short 35 nt reads are usually satisfactory. Preparation of RNA samples for *Brucella* RNAseq was essentially as described for microarrays with several modifications to improve DNA and rRNA removal as well as mRNA integrity. All samples were processed using the Qiagen RNeasy Protect bacteria mini kit with on-column RNase free DNase I digestion for effective removal of genomic DNA. Samples were assayed for RNA integrity using an Agilent 2100 Bioanalyzer and were quantified on a NanoDrop 1000 spectrophotometer <Please provide manufacturer name and address>. 5 μg of pure RNA was depleted of rRNA with the Ambion MicrobExpress kit. The kit is devised to use 10 μg RNA per reaction; using only half of the kit capacity we found a significant improvement in rRNA removal. This step results in a better proportion of non-rRNA reads in the final result. 100 ng of this enriched mRNA were used to construct the library for subsequent sequencing.

Libraries were generated using the Illumina® sample preparation protocols. Following purification, the mRNA was fragmented into small pieces using divalent cations at 94ºC during 5 min. Then, the cleaved RNA fragments were copied to first strand cDNA using SuperScript II reverse transcriptase (Invitrogen<Please provide manufacturer address>) and random primers. This was followed by second strand cDNA synthesis using DNA polymerase I and RNase H. These cDNA fragments went through an end repair process and then were ligated to the adapters. The products of the ligation reaction, of sizes around 250 bp, were further amplified by PCR using Phusion DNA polymerase (Finnzymes<Please provide manufacturer address>). These cDNA libraries were diluted to 6 pM. cDNA was hybridized to the flow-cell using a Single-Read Cluster Generation Kit v2 (Illumina Inc.). Further amplification of each DNA single molecule was performed by 35 cycles of isothermal amplification in the flow-cell on the Cluster Station. After amplification, one of the strands is removed and the genomic sequencing primer hybridized. The flow-cell was transferred to the Genome Analyser II×<Please provide manufacturer name and address> and sequencing was performed for 36 cycles using SBS Sequencing Kit v3 (Illumina Inc.)

RNAseq.: data analysis

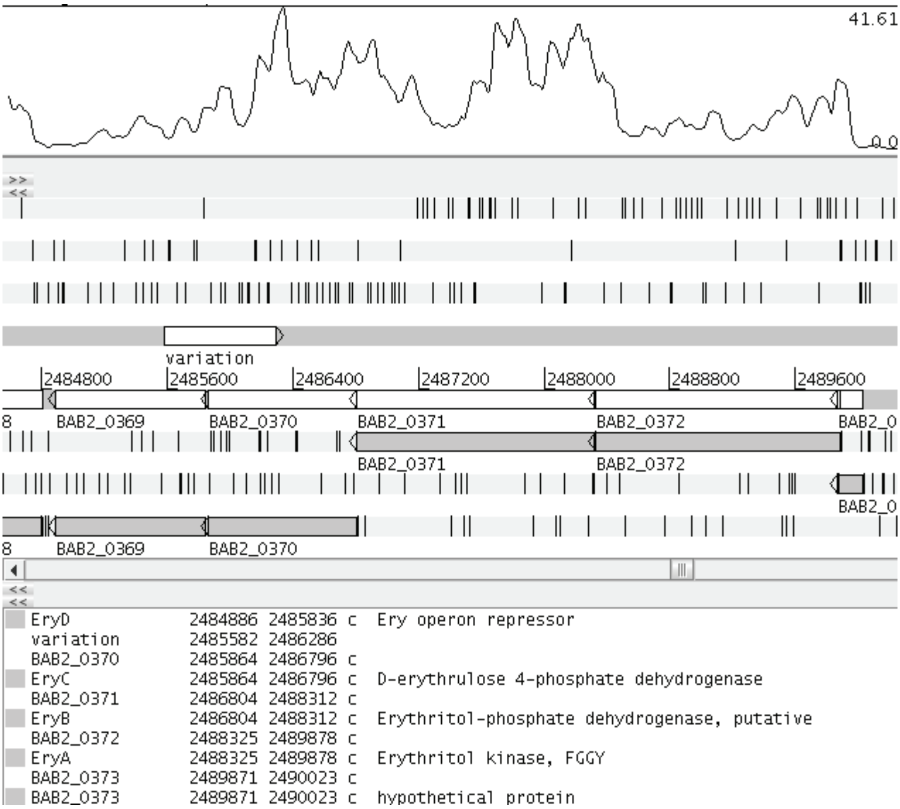
The raw data of the sequencing run consisted of 13,835,287 sequences 35 nt long. These sequences were aligned to the reference *B. abortus* 2308 genome by two separate methods. First we used the Illumina proprietary software ELAND (Illumina Corp). This procedure uses a filter file containing the rRNA coding sequences. Those reads aligning with sequences in the filter file are not reported in the output. A total 2,979,705 sequences were found to align with the coding region of the genome fulfilling at the same time the quality criteria of the software.

On the other hand, the reads were aligned using the open software package MAQ<Please provide manufacturer name and address> (Li *et al.*, 2008). This program allows a better user control on the aligning process. 8,762,781 (63.3% of the total reads) were found to align with the reference genome with two or less mismatches. 5,731,599 (65% of the aligned reads) corresponded to rRNA sequences. The remaining 3,031,182 reads were considered to align to 'single copy' DNA. The summary data of the RNAseq experiment is shown in Table 5.1.

To count how many times each base of the reference genome was contained in a single read we used the Maq pileup utility. From the output of this program we parsed a file containing the complete reads per base of the genome. This file contains a single column of numbers indicating how many times this base was read and a row per base of the genome (3,278,307 rows). Such a file may be used by the visualizing program Artemis <Please provide manufacturer name and address> (Rutherford *et al.*, 2000) as a user plot to indicate graphically the transcription level along the complete genome. Fig. 5.1 shows a screenshot of the Artemis program containing the erythritol operon region of the *B. abortus* 2308 genome (chromosome II). The results of the RNAseq experiment are shown in the top of the figure as user loaded plot containing the reads per base of the complete genome. The plot shows the data for 100 bp sliding window centred in each base. The region shown contains the four genes of the *ery* operon (*eryABCD*: BAB2_0372, BAB2_0371, BAB2_0370, BAB2_0369) encoded in the bottom chain of the DNA. The location of the *ery* deletion in strain S19 is also shown. The transcription plot for the region ranges from very low values (0 counts upstream the *eryA* gene) to a maximum of 41.6 counts, indicating how many reads containing that base have been sequenced. The characteristic saw-shaped profile of the plot is usually seen in RNAseq experiments (Güell *et al.*, 2009). The observed

Table 5.1 Summary of the RNAseq sequencing data

Total reads	1,383,5287
Aligned reads (maq, *n*=35, mis=2)	8,762,781 (63.3%)
rRNA	5,731,599 (65%)
Unique	3,031,182
Aligned ELAN	2,979,705



```
EryD            2484886 2485836 c  Ery operon repressor
variation       2485582 2486286
BAB2_0370       2485864 2486796 c
EryC            2485864 2486796 c  D-erythrulose 4-phosphate dehydrogenase
BAB2_0371       2486804 2488312 c
EryB            2486804 2488312 c  Erythritol-phosphate dehydrogenase, putative
BAB2_0372       2488325 2489878 c
EryA            2488325 2489878 c  Erythritol kinase, FGGY
BAB2_0373       2489871 2490023 c
BAB2_0373       2489871 2490023 c  hypothetical protein
```

Figure 5.1 Artemis screenshot showing the Chr II region containing the *ery* operon of *B. abortus* 2308. In the top part of the figure appears the corresponding RNA transcription plot read from the reads per base file. Base numbering corresponds to the two fused chromosomes (2,121,359 bp is the length of *B. abortus* 2308 Chr I)

transcription level seems to be higher for *eryB* and *eryC* than for *eryA* and *eryD*, which is unexpected from the paradigm of operon behaviour. In the case shown in the screen, there is a sharp increase in transcription upstream the *eryA* gene in a position neatly coincident with the experimentally determined transcriptional start point for the *ery* operon (Sangari *et al.*, 2000). From these data, the existence of the gene BAB2_0373, annotated in this genome as a hypothetical protein could be questioned. Fig. 5.1 summarizes the informative content of the data obtained. We may visualize the relative level of transcription of any gene. In some cases, such as the one presented in Fig. 5.1, the absence of background allows detection of the approximate site of the translation initiation point, which may be useful to identify and characterize *Brucella* promoters. Nevertheless specific libraries are needed in order to selectively keep the 5′ ends of every transcript, as has been applied for the global detection of transcriptional start sites of the gastric pathogen *Helicobacter pylori* (Sharma *et al.*, 2010). Other applications of the RNAseq method as new gene discovery, or identification of sRNAs from antisense regions, may obtain some information from this approach, but they will need to ascertain the actual DNA chain that is being transcribed in each position which in turn will require a different method of library preparation keeping directionality of mRNA (Vivancos *et al.*, 2010), which is lost in the process of library construction followed in this experiment.

The reads per base obtained as described were integrated using the GenBank annotation of *B. abortus* 2308 selecting both the CDSs and genes annotated as pseudogenes which totalized

3350 annotations (3034 CDSs and 316 pseudogenes). The result of this operation provided for each annotation (either gene or pseudogene) an expression value that represents the sum of the reads for all the bases of the CDS (or pseudogene), and accordingly gives a relative value to the extent of transcription of the element. This annotation could be improved specially in the part concerning to pseudogenes; however, we decided to use it in order to allow more consistent comparison between the RNAseq and microarray data.

Brucella transcriptional snapshots

Identification of gene sets in exponentially growing B. abortus with high and low level expression

Two comparable data sets were obtained in our laboratories on the global transcription of *B. abortus* 2308 growing at mid-exponential phase on *Brucella* Broth medium, one using the *Brucella* microarray (Viadas *et al.*, 2009) and the other by RNAseq. We used these data to analyse with them the genes more expressed and those with the poorest expression in the mentioned conditions. In both sets of data, the expression indexes were normalized to gene size and Llog$_2$ transformed to compact the data. Then, the Gene Expression Index (GEI) was sorted and the top and bottom 10% of each list were selected as the more and less expressed genes in *B. abortus* 2308 growing in exponential conditions. The dynamic range of the quantitative data obtained by each method was as different as could be expected by the different technologies. Microarray results were obtained from the fluorescence intensity that needs to be normalized and background corrected before analysis. In our microarray experiment the log$_2$ of the fluorescence intensity range from the more to the less expressed genes was 14.23–2.73. The cut points for the high and low expression categories were 9.19 and 6.50. In the RNAseq experiment we count how many reads were contained in the interval defining a gene, then to avoid selecting for large genes, counts were normalized by gene size and expressed as counts per kb. Twenty-six genes obtained 0 counts in the RNAseq experiment showing that they are indeed not expressed. On the other hand we have a null background due to contaminant genomic DNA. Perhaps this is one of greatest differences that favour RNAseq versus microarrays for estimation of low expression genes (see below). The highest expressed gene (after filtration of rRNAs) in the RNAseq experiment was the porin BAB1_0660. This was also the most expressed gene in microarray experiment. BAB1_0660 showed nearly 5 million counts, this means that the 1128 nucleotides that the gene contains were read 4,955,906 times or that 141,600 reads contained BAB1_0660 sequences. The GEI calculated for BAB1_0660 was 4397.43 (counts/size in bp). The last gene in the high expression set in RNAseq showed a GEI of 73.5. These values were similar to those observed in the microarrays (12.2–6.2 log$_2$ transformed). The low expression gene set deduced from the RNAseq experiment ranged in GEI value between and 1.18 and 0.001 (log$_2$ transformed from 1.18 to –10). There is a large difference between the lower values detected in the two methods. In the microarray experiment the lower log$_2$ observed was 2.73. Detection of lower values, as those obtained in the RNAseq method are not reliable owing to the high background. The range of values observed spanned 22 logs for RNAseq but only 12 in the microarrays.

We then analysed the coincidence among the gene sets obtained by both methodologies. Results were rather consistent in the high expression data set, 185 genes (near 60%) were common in the two data sets, while only 59 genes (less than 20%) were common in the low-expression sets. The Venn diagram in Fig. 5.2 was produced with BioVenn <Please provide manufacturer name and address> (Hulsen *et al.*, 2008) accessed at its web server (http://www.cmbi.ru.nl/cdd/biovenn/index.php). The level of coincidence was as expected, taken into account the inherent divergence found in global assays and the differences in methodology. As discussed above, probably the low expression data set obtained from RNAseq should reflect best the reality than the microarray-derived data.

A more functional way to compare the results was to analyse the distribution of the selected genes in the functional COG categories (Tatusov *et al.*, 2003). The COG category of the selected

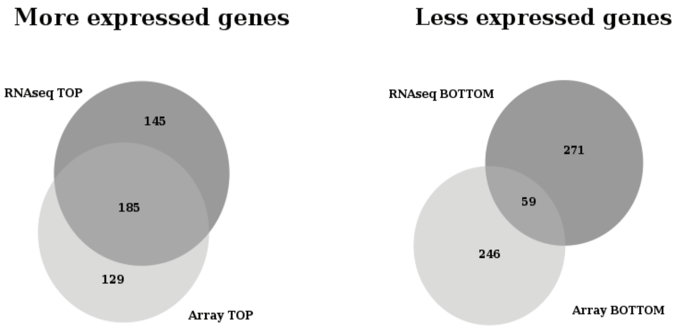


Figure 5.2 Coincidence degree among the high and low expression gene sets determined by the microarray and the RNAseq technologies.

genes was taken from the precomputed refseq ptt files corresponding to the *B. abortus* 2308 genome obtained from Gene Bank. The distribution of the more and less expressed genes in COG categories is shown in the Table 5.2. The table contains the COG distribution for the whole genome as a reference. This allowed to detect differences between the two distributions. Looking only to RNAseq results, in the group of seq results, in the group of more expressed genes the more overrepresented COG categories were 'translation', 'post-translational modification, protein turnover, chaperones' and 'energy production and conversion'. As expected, the categories overrepresented in the high expression gene-set were almost absent in the low expression data set. The low expression group of genes was enriched in the 'cell motility' category owing to the very low expression of many of the genes related to biosynthesis of the flagellum. The possibility of the presence of a flagellum in *Brucella* was deduced from the existence of flagellar genes in the *Brucella* genome (Halling, 1998). Later it was shown *Brucella* may produce a flagellar structure involved in persistence in mice and produced only at the beginning of the exponential phase (Fretin *et al.*, 2005). Another category overrepresented in this set was 'replication, recombination and repair'. This result seems to be a result of the presence in the set of a large number of transposases, and site-specific recombinases, rather to the lack of expression of DNA polymerase or general recombination genes. It was also noteworthy that more than half of the genes in this set could not be assigned to any of the COG categories. This proportion is much more lower (28.8%) in the whole *Brucella* genome. This enrichment could be explained by the existence in the *B. abortus* 2308 annotation of some genes, many of them small, predicted by automatic annotation systems but not corresponding to real genes. Most of these will not be assigned to any COG. A summary of the analysis of distribution among COG categories was obtained by plotting in an histogram the frequency of the genes in each category relative to frequency in *B. abortus* 2308 genome, piling the high and low expression data sets in the same column. This plot contained in Fig. 5.3, shows the extreme enrichment of the J category in the high expression and N and L categories in the low expression data set.

Another way to look at the functional classification of the above data sets was by means of the Database for Annotation, Visualization and Integrated Discovery (DAVID v6.7<Please provide manufacturer address>) accessible via web at http://david.abcc.ncifcrf.gov/ (Huang *et al.*, 2007). The gene functional classification tool of DAVID computes a list of genes to detect whether some descriptive terms taken from the more used databases are enriched in the list with respect to the reference genome. This enrichment is calculated with a variation of the Exact Fisher *P*-value. DAVID also clusters the list in groups of genes associated with some biological function allowing a functional analysis of the gene list. Using this tool we found that the more represented functional groups among the more expressed genes in exponentially growing *Brucella* were those

Table 5.2 COG category distribution of the genes in the low and high expression data sets

Code	Description	Genome	Top_arr	Top_seq	Bot_arr	Bot_seq
A	RNA processing and modification	0	0	0	0	0
B	Chromatin structure and dynamics	0	0	0	0	0
C	Energy production and conversion	155	37	45	17	12
D	Cell cycle control, mitosis and meiosis	27	3	5	5	0
E	Amino acid transport and metabolism	294	12	11	38	26
F	Nucleotide transport and metabolism	58	3	3	3	0
G	Carbohydrate transport and metabolism	131	6	13	22	12
H	Coenzyme transport and metabolism	105	10	6	16	5
I	Lipid transport and metabolism	100	7	7	15	5
J	Translation	131	61	63	5	0
K	Transcription	172	15	12	22	6
L	Replication, recombination and repair	114	4	1	20	22
M	Cell wall/membrane biogenesis	143	21	19	18	8
N	Cell motility	24	0	0	9	16
O	Post-translational modification, protein turnover, chaperones	103	14	24	6	4
P	Inorganic ion transport and metabolism	136	10	12	28	13
Q	Secondary metabolites biosynthesis, transport and catabolism	61	4	2	12	4
R	General function prediction only	283	19	13	33	14
S	Function unknown	164	25	24	17	10
T	Signal transduction mechanisms	74	3	6	13	2
U	Intracellular trafficking and secretion	46	7	8	10	9
V	Defence mechanisms	27	0	0	5	2
W	Extracellular structures	1	0	0	0	0
Z	Cytoskeleton	0	0	0	0	0
–	Not in COGs	950	79	69	29	179

Code: One letter code of COG categories. Genome: indicates the number of genes in each COG category in the complete *B. abortus* 2308 genome. Top_arr, Top_seq, Bot_arr, Bot_seq: indicate the number of genes in each COG category in the 10% more expressed genes from the microarray, 10% more expressed genes from RNAseq, 10% more less <Should this be more OR less?> expressed genes from the microarray and 10% less expressed genes from RNAseq, respectively.

involved in ribosomal biogenesis (55 and 36 genes with this function were present in the sequencing and microarray gene sets). Another cluster of genes in this set were related with respiratory electron transport mainly implicated in reduction of NADH. Enriched terms were 'electron transport', 'iron–sulphur', and 'NADH-quinone oxyreductase'. Oxidative phosphorylation, ATP synthase and tricarboxylic acid cycle related terms were also enriched. The more significant category identified by DAVID in the low expression gene sets were again the flagellum related genes, also detected by COG distribution comparison. Another functional group of genes identified by the DAVID engine was a large and heterogeneous group of genes encoding nucleotide binding proteins many of them transport related. This group included many ATP-binding proteins from small molecule transport systems but also components of the type IV secretion system. A group of eight genes encoding transposases and related proteins also was found significant enriched only in the RNAseq low expression gene set.

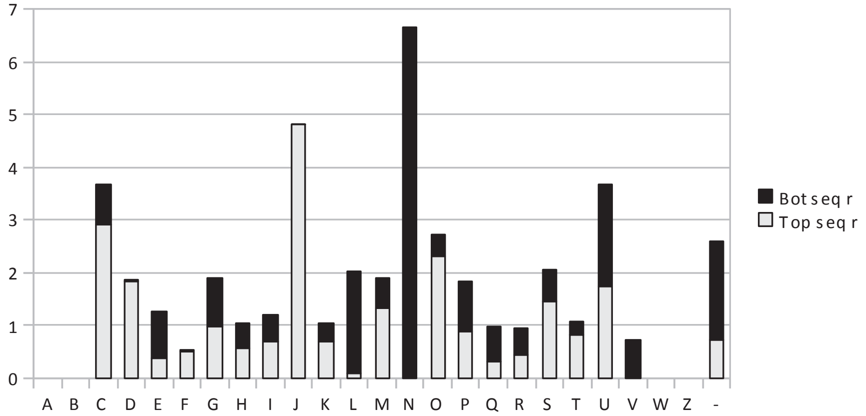


Figure 5.3 Distribution in COG categories of the more and less expressed genes of *B. abortus* 2308. The histogram shows the per cent on each class (black or light grey bars) of genes in each COG category as indicated. Similar height of the colours in each column indicates similar distribution in the two classes. Colour asymmetries indicate different distribution. This is extreme in some classes (D, F and J categories are poorly represented in the low expression set and L and N are almost absent in the high expression gene set). One letter code to COG categories are described in Table 5.2.

Transcription during cell invasion

While no direct analysis of global intracellular transcription has been performed, there have been several lateral approaches that are worth a comment in this section.

From the observation that *B. melitensis* 16M at late-log phase of growth were more invasive to epithelial cells that were bacteria at mid-log or stationary growth phases, it has been assumed that genes preferentially expressed at this phase could be implicated at least in the initial phases of the cell invasion process (Rossetti *et al.*, 2009). Then, analysis of transcription at late-log phase of growth may be considered a surrogate of the invasion process. Comparative microarray analysis revealed a greater number of genes up-regulated in late-log phase cultures than in stationary phase cultures: of the 454 genes significantly altered in *B. melitensis* during late-log phase (the most invasive), 414 (91%) were up- and 40 (9%) were down-regulated compared with the bacteria in stationary phase (the less invasive). According to the conditions compared, the majority of gene expression changes were associated with growth and metabolism. Among the up-regulated genes were those associated with DNA replication, transcription and translation (57 genes), nucleotide, amino acid, lipid and carbohydrate metabolism (65 genes), energy production and conversion (24 genes), membrane transport (56 genes) and cell envelope, biogenesis and outer membrane (26 genes), while the 40 down-regulated genes were more heterogeneous in nature, demonstrating no predominant functional category. Several genes whose products are known to be associated with *Brucella* virulence were differentially expressed between the most and the less invasive cultures. These included: SP41 gene, a surface protein that enables *B. suis* to attach and penetrate non-phagocytic cells (Castaneda-Roldán *et al.*, 2006); three genes components of the *virB* locus (*virB1, virB3* and *virB10*), that plays a critical role in *Brucella* virulence and intracellular multiplication (see also Chapter 11); five genes (*fliC, fliF, fliN, flhA* and *flgD*) which encode parts of the flagellar apparatus; several transcriptional regulators; and cell envelope and outer membrane biogenesis genes (LPS and peptidoglycan biosynthesis and outer membrane proteins and lipoproteins). Interestingly 22% of the differentially expressed genes identified in this study had an uncharacterized function, and may contain unknown essential information to completely understand the virulence factors utilized for *Brucella* to invade and infect the host.

Other transcriptional problem targeted with microarray experiments was the effect of the transcriptional regulator VjbR. Mutants missing this regulator are attenuated in all models of infection and suggest that quorum sensing plays an essential role in the regulation of virulence, specially of those genes required for adaptation to growth in the host. Nimblegen oligonucleotide microarrays were used to compare transcriptional profiles of wild-type and *vjbR* mutants (Uzureau *et al.*, 2010). This analysis showed near a 10% of *B. melitensis* genes differentially expressed and henceforth probably involved in the process of host invasion. These differentially expressed genes, corroborated by other experimental methods, were involved in central metabolic pathways, virulence and stress responses which has been interpreted as proof of the central role of *vjbR* (and quorum sensing) in the adaptation of *Brucella* to the oxidative, pH, and nutritional stresses encountered within the host.

Intracellular transcription of Brucella genes

The description of the bacterial genes being expressed in the intracellular compartment would be extremely informative to understand the mechanisms the *Brucellae* use to thrive inside cells. A lot of effort has been dedicated to the characterization of *Brucella* genes required for *in vivo* infection, either in animals or in cultured cells. In some way a correlation may be established between requirement of the genes for *in vivo* growth and gene expression. In other words, the set of genes required for intracellular growth must be equivalent to the genes expressed in intracellular conditions. Kohler *et al.* used expression of Green Fluorescent Protein (GFP) as a means to identify *Brucella* promoters induced specifically in macrophages (Kohler *et al.*, 1999). The gene *nikA* was found to be induced in *B. suis* growing into J774 macrophages, which prompted an extensive analysis of the *nik* operon involved in nickel uptake and urease activity (Jubier-Maurin *et al.*, 2001). The analysis of intracellular expression with the use of gene-fusions with GFP to measure intracellular fluorescence either by microscopy or flow cytometry was also applied to study the behaviour of the *virB* promoter of *B. suis* growing inside J774 macrophages (Boschiroli *et al.*, 2022).

Later, a collection of 10,272 mini-Tn5 transposon mutants of a constitutively fluorescent *B. suis* were screened by fluorescence microscopy for lack of intracellular multiplication in human macrophages, and 131 attenuated mutants affected in 59 different genes were detected. This identified genes involved in global adaptation to intracellular environment, amino acid and nucleotide synthesis, sugar and nitrogen metabolism, redox reactions, regulation of transcription, disulphide bond formation, and LPS biosynthesis. This analysis allowed to define the environment that the bacteria encountered in the macrophage and to better understand the nature of the phagosome: it is poor in nutrients, as shown by the requirement of functional genes of various biosynthesis pathways; it most probably has a neutral pH; it is also characterized by low oxygen tension and, alternatively, nitrate ions are available for anaerobic respiration.

The limitation in the amount and quality of *Brucella* RNA recovered from infected cells has prevented by now the global analysis of intracellular transcription either by microarray or RNAseq experiments. For this reason direct analysis of intracellular transcription has been performed only on selected genes by RT-PCR reactions using as template RNA obtained from intracellular bacteria. This analysis was done on a set of 32 genes selected because of their known relation with virulence (Viadas *et al.*, 2010). Previously, a whole-genome microarray analysis determined the genes regulated by the two-component system BvrR/BvrS, related to *Brucella* virulence (see also Chapter 10). Most of the genes candidate to be regulated by BvrR/BvrS identified in the microarray experiments can be involved with the changes needed for intracellular survival of *Brucella*. In order to investigate if these genes were expressed intracellularly, bacterial RNA was obtained from *B. abortus* wild type recovered from BHK21 infected cells and from the same strain (*B. abortus* 2308) grown in laboratory conditions (Viadas *et al.*, 2010) and RT-PCR performed with specific primers for the selected genes. The results showed significant differences in the expression of at least fifteen genes: (i) those highly expressed intracellularly included genes related to virulence (*virB8*), carbon and nitrogen metabolism (*malF*, *pckA, fumB, norC*), fatty acid and LPS biosynthesis and a transcriptional regulator (*vjbR*), and (ii) genes with lower levels of intracellular expression were related to cell envelope proteins (*omp25d*, lipoproteins) and denitrification (*nosZ, nirK* and glutaminase). Interestingly, *virB8* was the gene with the highest level of expression intracellularly, which is in concordance with previous reports (Boschiroli *et al.*, 2002; Sieira *et al.*, 2004). In relation to nitrogen metabolism, it is significant that where the nitrite reductase gene (*nirK*) was poorly expressed intracellularly, the nitric oxide reductase (*norC*) was highly expressed. This regulation of the denitrification route could be use by *Brucella* to survive using nitrogen oxides as terminal electron receptors and limiting the production of reactive nitrogen intermediates by the host.

Highly transcribed intergenic regions

One of the advantages of the RNAseq approach is that data are obtained for all the positions of the genome (counts per base) in an annotation independent way. This raw result may be integrated for any set of subsequences the analyst could imagine by only defining start and endpoints of the features of interest. Taking advantage of this versatility, we have also analysed transcription from the intergenic regions of the *Brucella* 2308 genome. We added to the CDS and pseudogene annotation the data for tRNA and mRNA. To avoid short intercistronic regions of operons, intergenic spaces smaller than 20 bp were eliminated. The total counts per base were integrated for these intergenic spaces as was made for CDSs, and a relative 'intergene' expression index was also calculated. GEI for the more expressed CDS varied from 73 to 4400. Surprisingly 156 intergenic spaces showed GEI's higher than 73, reaching a maximal value of 1800, very close to the more expressed genes. This finding indicates that in addition to annotated CDS and RNAs there are many loci in the genome that are actively transcribed and presumably play functions in the biology of the *Brucella*. We have analysed the top members of the transcribed intergenic regions and detected that some of them corresponded to well known structural or functionally relevant non-translated RNAs. While most of the highly expressed intergenic regions are between highly expressed genes (i.e. ribosomal proteins), some of them encoded for elements related to mRNA processing or stability such as RNAse P (Scott *et al.*, 2009) and tmRNA (Liang and Deutscher, 2010).

Transcription of pseudogenes

Twenty-four of the genes selected from the RNAseq data were annotated as pseudogenes in the *B. abortus* 2308 genome, which was considered a rather unexpected finding. By comparison with other *Brucella* genomes we can reduce the list of highly expressed pseudogenes to 16 (often, truncated parts of a gene are annotated as different pseudogenes especially in *B. abortus* 2308). This seems contradictory since high transcription of these genes, which should be not able to translate into functional proteins, will be contrary to biological economy. The high levels of transcription observed for these genes strongly suggest that they could be active genes and their products may perform functions unreported in metabolic reconstructions. High pseudogene expression may also indicate that these are very recently produced pseudogenes that did not turned down transcription yet by accumulation of mutations in their promoter or control regions. It is also possible that these pseudogenes may contain sequencing errors and they are indeed active genes. On the other hand eighty two pseudogenes were found in the low expression class. This represents a significant enrichment of pseudogenes in the low expression set. This subset of pseudogenes showed a level of transcription consistent with their identification as inactivated, non-working genes, which also turned out transcription to avoid the energy burden of expressing partial or altered gene products.

A similar result showing pseudogene translation has been recently reported in a compilation of proteomic analysis, that described five peptide spots in a *B. abortus* 2308 proteome analysis corresponding to genes annotated as pseudogenes (Lamontagne *et al.*, 2010).

The absence of pseudogenes in the microarray data set was due to the method used to convert *B. melitensis* annotation to *B. abortus* 2308. This conversion was done by BlastP and Tblastn comparisons against *B. abortus* 2308 protein database, in which pseudogenes are not included.

Transcriptional differences between the two chromosomes

The structure of the *Brucella* genome in two separate replicons has been interpreted as a solution to organize the more used genes in the large (or principal) chromosome while the small chromosome contains genes useful only for particular situations. An analysis of COG category distribution among the two chromosomes confirms this view. Genes related to replication and recombination and translation, post-translational modification, protein turnover, chaperones, etc. are overrepresented in Chr I, while Chr II contains more genes related to transport. According to this view a higher level of transcription would be expected from genes in Chr I than from genes in Chr II in fast growth favourable conditions as those used in our experiments. The asymmetric distribution of highly expressed genes in Chr I had been already detected in the microarray analysis of transcription. The RNA seq data confirmed this observation as demonstrates the plot shown in Fig. 5.4. The plot of density frequencies of genes by expression level shows a global difference of near one log (two times) between expression of the genes in both chromosomes. In the RNAseq gene lists, 273 genes of the high expression class were in Chr I while only 146 of the low expressed genes were from Chr I.

Some differences in gene expression have been attributed to gene dosage effects. Genes located near the replication origin may have copy numbers higher than those close to the terminus, specially at fast growth rates (Cooper *et al.*, 2010), as is the case considered in our data sets. To analyse this possibility, we decided to study the correlation between the level of expression and the distance to the origin for the two chromosomes for each gene. To locate the origin of replication of *Brucella* chromosomes we used the DoriC program <Please provide manufacturer name and address>, which is a systematic method comprising: (i) the Z-curve analysis for nucleotide distribution asymmetry; (ii) DnaA box distribution; (iii) the presence of genes adjacent to candidate *ori*Cs in other species; and (iv) phylogenetic relationships (http://tubic.tju.edu.cn/doric/). The *B. abortus* 2308 Chr I origin was located at 2007638..2008089. This *oriC* region was defined as a 452 bp containing three putative DnaA boxes (TGTGGAAAA) and is adjacent to *haem,* a gene found close to the replication origin in other α-Proteobacteria (Brassinga *et al.*, 2001). For Chr II the origin of replication was predicted

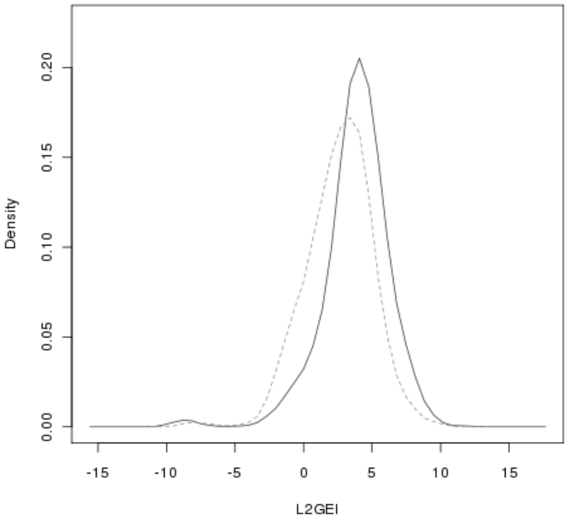


Figure 5.4 The figure shows the plot of the density of frequencies of genes showing a given transcription level expressed as the Log2 of the gene expression index (GEI) from either the Chr I (continuous line) or the Chr II (dashed line).

as a 368 nt sequence (1156735..154) containing two putative DNA boxes and close to a *repC* gene. Considering that replication of the two chromosomes was bidirectional from the origin we computed the distance of each gene to the origin but failed to detect any effect of the distance to the origin on the observed levels of transcription.

Future perspectives and conclusions

The transcriptome of *B. abortus* 2308 has been analysed in a comprehensive manner using two powerful technologies, a DNA microarray derived from the *Brucella* ORFeome and RNAseq. The data sets were used to identify the more and less expressed genes in bacteria grown in laboratory conditions. The two methods used performed well to identify the more expressed genes. The low expression sets were more divergent but still allowed the identification some gene families in this class, particularly those related to flagellar production in *Brucella*.

RNAseq data allow more possibilities to obtain information on the transcriptional landscape of the bacteria than the ORFeome derived microarray. Some interesting results on pseudogene expression, transcription from non-annotated regions, identification of clusters of genes and location of promoters may be obtained from the results. However, a different methodology, conserving directionality of mRNA in the sequencing libraries will be needed for a full definition of the transcriptome. Another objective that has to be pursued is the analysis of transcription in intracellular conditions. This has done already in other intracellular bacteria as *Yersinia pestis* (Fukuto *et al.*, 2010) and new methods available (Azhikina *et al.*, 2010) to improve purification of RNA from intracellular bacteria combined to the high power of RNAseq, should allow soon the detailed analysis of the transcriptome of intracellular *Brucella*.

Several proteomic analysis of *Brucella* have been performed using different species and experimental conditions. The results from many of them have been compiled in a recent publication (Lamontagne *et al.*, 2010) and are also discussed in Chapter 6. The combination of proteomic data with those of the transcriptome should reveal in the near future details of *Brucella* biology that we should transform in new tools for the practical control of brucellosis on earth.

References

Azhikina, T., Skvortsov, T., Radaeva, T., Mardanov, A., Ravin, N., Apt, A., and Sverdlov, E. (2010). A new technique for obtaining whole pathogen transcriptomes from infected host tissues. BioTechniques *48*, 139–144.

Boschiroli, M.L., Ouahrani-Bettache, S., Foulongne, V., Michaux-Charachon, S., Bourg, G., Allardet-Servent, A., Cazevieille, C., Liautard, J.P., Ramuz, M., and O'Callaghan, D. (2002). The *Brucella suis virB* operon is induced intracellularly in macrophages. Proc. Natl. Acad. Sci. U.S.A. *99*, 1544–1549.

Brassinga, A.K.C., Siam, R., and Marczynski, G.T. (2001). Conserved gene cluster at replication origins of the α-proteobacteria *Caulobacter crescentus* and *Rickettsia prowazekii*. J. Bacteriol. *183*, 1824–1829.

Castañeda-Roldán, E.I., Ouahrani-Bettache, S., Saldaña, Z., Avelino, F., Rendón, M.A., Dornand, J., and Girón, J.A. (2006). Characterization of SP41, a surface protein of *Brucella* associated with adherence and invasion of host epithelial cells. Cell. Microbiol. *8*, 1877–1887.

Cooper, V.S., Vohr, S. H., Wrocklage, S.C., and Hatcher, P.J. (2010). Why genes evolve faster on secondary chromosomes in bacteria. PLoS Comput. Biol. *6*, e1000732.

Dricot, A., Rual, J., Lamesch, P., Bertin, N., Dupuy, D., Hao, T., Lambert, C., Hallez, R., Delroisse, J., Vandenhaute, J., *et al.* (2004). Generation of the *Brucella melitensis* ORFeome version 1.1. Genome Res. *14*, 2201–2206.

Eskra, L., Canavessi, A., Carey, M., and Splitter, G. (2001). *Brucella abortus* genes identified following constitutive growth and macrophage infection. Infect. Immun *69*, 7736–7742.

Ficht, T.A., Bearden, S.W., Sowa, B.A., and Adams, L.G. (1989). DNA sequence and expression of the 36-kilodalton outer membrane protein gene of *Brucella abortus*. Infect. Immun. *57*, 3281–3291.

Fretin, D., Fauconnier, A., Köhler, S., Halling, S., Léonard, S., Nijskens, C., Ferooz, J., Lestrate, P., Delrue, R., Danese, I., *et al.* (2005). The sheathed flagellum of *Brucella melitensis* is involved in persistence in a murine model of infection. Cell. Microbiol. *7*, 687–698.

Fukuto, H.S., Svetlanov, A., Palmer, L.E., Karzai, A.W., and Bliska, J.B. (2010). Global gene expression profiling of *Yersinia pestis* replicating inside macrophages reveals the roles of a putative stress-induced operon in regulating type III secretion and intracellular cell division. Infect. Immun. *78*, 3700–3715.

Güell, M., van Noort, V., Yus, E., Chen, W., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., *et al.* (2009). Transcriptome complexity in a genome-reduced bacterium. Science *326*, 1268–1271.

Halling, S.M. (1998). On the presence and organization of open reading frames of the nonmotile pathogen *Brucella abortus* similar to class II, III, and IV flagellar genes and to LcrD virulence superfamily. Microb. Comp. Genomics 3, 21–29.

Hernández-Castro, R., Rodríguez, M.C., Seoane, A., and García Lobo, J.M. (2003). The aquaporin gene *aqpX* of *Brucella abortus* is induced in hyperosmotic conditions. Microbiology (Reading, Engl.) *149*, 3185–3192.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H.C., *et al.* (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. *35*, W169–W175.

Hulsen, T., de Vlieg, J., and Alkema, W. (2008). BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics *9*, 488–488.

Jubier-Maurin, V., Rodrigue, A., Ouahrani-Bettache, S., Layssac, M., Mandrand-Berthelot, M.A., Kohler, S., and Liautard, J.P. (2001). Identification of the *nik* gene cluster of *Brucella suis*: regulation and contribution to urease activity. *183*, 426–434.

Kohler, S., Ouahrani-Bettache, S., Layssac, M., Teyssier J., and Liautard, J.P.(1999). Constitutive and inducible expression of green fluorescent protein in *Brucella suis*. *67*, 6695–6697.

Kohler, S., Foulongne, V., Ouahrani-Bettache, S., Bourg, G., Teyssier, J., Ramuz, M., and Liautard, J.P. (2002). The analysis of the intramacrophagic virulome of *Brucella suis* deciphers the environment encountered by the pathogen inside the macrophage host cell. Proc. Natl. Acad. Sci. U.S.A. *99*, 15711–15716.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., *et al.* (2009). Proteome organization in a genome-reduced bacterium. Science *326*, 1235–1240.

Lamontagne, J., Béland, M., Forest, A., Côté-Martin, A., Nassif, N., Tomaki, F., Moriyón, I., Moreno, E., and Paramithiotis, E. (2010). Proteomics-based confirmation of protein expression and correction of annotation errors in the *Brucella abortus* genome. BMC Genomics *11*, 300.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.

Liang, W., and Deutscher, M.P. (2010). A novel mechanism for ribonuclease regulation: transfer-messenger RNA (tmRNA) and its associated protein SmpB regulate the stability of RNase R. J. Biol. Chem. *285*, 29054–29058.

Ochman, H., and Raghavan, R. (2009). Systems biology. Excavating the functional landscape of bacterial cells. Science *326*, 1200–1201.

Rossetti, C.A., Galindo, C.L., Lawhon, S.D., Garner, H.R., and Adams, L.G. (2009). *Brucella melitensis* global gene expression study provides novel information on growth phase-specific gene regulation with potential insights for understanding *Brucella*:host initial interactions. BMC Microbiol. *9*, 81.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics *16*, 944–945.

Sangari, F.J., Agüero, J., and García-Lobo, J.M. (2000). The genes for erythritol catabolism are organized as an inducible operon in *Brucella abortus*. Microbiology*146,* 487–495.

Scott, W.G., Martick, M., and Chi, Y. (2009). Structure and function of regulatory RNA elements: ribozymes that regulate gene expression. Biochim. Biophys. Acta *1789*, 634–641.

Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., *et al.* (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature *464*, 250–255.

Smith, L.D., and Heffron, F. (1987). Transposon Tn5 mutagenesis of *Brucella abortus*. Infect. Immun *55*, 2774–2776.

Sieira, R., Comerci, D.J., Pietrasanta, L.I., and Ugalde, R.A. (2004). Integration host factor is involved in transcriptional regulation of the *Brucella abortus virB* operon. Mol. Microbiol. *54*, 808–822.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., *et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Uzureau, S., Lemaire, J., Delaive, E., Dieu, M., Gaigneaux, A., Raes, M., De Bolle, X., and Letesson, J. (2010). Global analysis of quorum sensing targets in the intracellular pathogen *Brucella melitensis* 16 M. J. Proteome Res. *9*, 3200–3217.

Viadas, C., Rodríguez, M.C., García-Lobo, J.M., Sangari, F.J., and López-Goñi, I. (2009). Construction and evaluation of an ORFeome-based *Brucella* whole-genome DNA microarray. Microb. Pathog. *47*, 189–195.

Viadas, C., Rodríguez, M.C., Sangari, F.J., Gorvel, J., García-Lobo, J.M., and López-Goñi, I. (2010). Transcriptome analysis of the *Brucella abortus* BvrR/BvrS two-component regulatory system. PLoS ONE *5*, e10216.

Vivancos, A.P., Güell, M., Dohm, J.C., Serrano, L., and Himmelbauer, H. (2010). Strand-specific deep sequencing of the transcriptome. Genome Res. *20*, 989–999.

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W., Wodke, J. A. H., Güell, M., Martínez, S., Bourgeois, R., *et al.* (2009). Impact of genome reduction on bacterial metabolism and its regulation. Science *326*, 1263–1268.