INFECTION AND IMMUNITY, Feb. 2001, p. 865–868 Vol. 69, No. 2
0019-9567/01/$04.00+0 DOI: 10.1128/IAI.69.2.865–868.2001

Gene Discovery through Genomic Sequencing of *Brucella abortus*

DANIEL O. SÁNCHEZ,[1] RUBEN O. ZANDOMENI,[2] SILVIO CRAVERO,[3] RAMIRO E. VERDÚN,[1] ESTER PIERROU,[4] PAULA FACCIO,[2] GABRIELA DIAZ,[2] SILVIA LANZAVECCHIA,[2] FERNÁN AGÜERO,[1] ALBERTO C. C. FRASCH,[1] SIV G. E. ANDERSSON,[4] OSVALDO L. ROSSETTI,[3] OSCAR GRAU,[2,5] AND RODOLFO A. UGALDE[1]*

Instituto de Investigaciones Biotecnológicas-Instituto Tecnológico de Chascomús, CONICET/UNSAM, Universidad Nacional de General San Martín,[1] Laboratorio de Alta Complejidad (IMYZA),[2] and Instituto de Biotecnología, INTA,[3] Buenos Aires, and IBBM, Universidad de la Plata, La Plata,[5] Argentina, and Department of Molecular Evolution, Uppsala University, Uppsala, Sweden[4]

Received 29 June 2000/Returned for modification 29 August 2000/Accepted 6 November 2000

Brucella abortus **is the etiological agent of brucellosis, a disease that affects bovines and human. We generated DNA random sequences from the genome of *B. abortus* strain 2308 in order to characterize molecular targets that might be useful for developing immunological or chemotherapeutic strategies against this pathogen. The partial sequencing of 1,899 clones allowed the identification of 1,199 genomic sequence surveys (GSSs) with high homology (BLAST expect value $< 10^{-5}$) to sequences deposited in the GenBank databases. Among them, 925 represent putative novel genes for the *Brucella* genus. Out of 925 nonredundant GSSs, 470 were classified in 15 categories based on cellular function. Seven hundred GSSs showed no significant database matches and remain available for further studies in order to identify their function. A high number of GSSs with homology to *Agrobacterium tumefaciens* and *Rhizobium meliloti* proteins were observed, thus confirming their close phylogenetic relationship. Among them, several GSSs showed high similarity with genes related to nodule nitrogen fixation, synthesis of nod factors, nodulation protein symbiotic plasmid, and nodule bacteroid differentiation. We have also identified several *B. abortus* homologs of virulence and pathogenesis genes from other pathogens, including a homolog to both the Shda gene from *Salmonella enterica* serovar Typhimurium and the AidA-1 gene from *Escherichia coli*. Other GSSs displayed significant homologies to genes encoding components of the type III and type IV secretion machineries, suggesting that *Brucella* might also have an active type III secretion machinery.**

Bovine brucellosis is a disease that affects livestock with a high incidence in South America. The etiological agent is the pathogenic bacterium *Brucella abortus*, a gram-negative nonmotile coccobacilus that produces abortions, sterility, and orchitis. It also affects humans, producing undulant fever, migraine, nausea, arthritic pain, and partial or total motion incapability (20). Although *Brucella* affects humans, it does not spread among them.

Classically the characterization and identification of virulence genes was carried out by the generation of random mutants searching for avirulent phenotypes (7). Although useful, this approach has led to the identification of a limited number of genes. Some modifications, such as targeted mutagenesis, have been developed in order to improve the isolation of avirulent mutants (9). One of the drawbacks for the identification of virulence genes is that in most cases these genes are induced within the host and are not expressed under normal laboratory culture conditions (18).

Cloning, characterization, and identification of *B. abortus* virulence genes is important for understanding the molecular pathogenesis of this intracellular microorganism. It might also help for the identification of antigens useful for the development of diagnosis tests and new vaccines.

* Corresponding author. Mailing address: Instituto de Investigaciones Biotecnológicas, Universidad Nacional de General San Martín, INTI (Ed. 24), Av. Gral Paz entre Constituyentes y Albarellos, 1650 San Martín, Provincia de Buenos Aires, Argentina. Phone: (54-11) 4580-7255. Fax: (54-11) 4752-9639. E-mail: rugalde@iib.unsam.edu.ar.

Today it is possible to obtain the complete genomic sequence of a microorganism and, by comparison with known sequences deposited in protein or nucleotide sequence databases, to assign functions to genes and identify regulatory sequences. As of October 2000, the complete sequences of the genomes of 28 bacterial species have been released (see http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html), and many others are in different stages of completion. With multiple genomes sequenced, it is possible to delineate highly conserved protein families (17). Such information may be critically important, for example, to assign virulence functions to those protein families that are conserved among different bacteria.

In this work we have begun a large-scale sequencing of random *B. abortus* genomic fragments in order to obtain the sequence of a representative number of genes and facilitate the identification of genes involved in virulence and other host interaction functions. We describe 1,899 *B. abortus* genomic sequence surveys (GSSs) which represent an interesting start that may be used in combination with targeted mutagenesis in functional studies and identification of virulence genes.

MATERIALS AND METHODS

Genomic libraries. DNA used for construction of the libraries was isolated by CsCl-ethidium bromide equilibrium centrifugation. Three different libraries from *B. abortus* strain 2308 were constructed, with two in the plasmid vector pBluescript SK(−) (Stratagene). The first library was made by partial digestion of total DNA with *Sau*3AI. Restriction fragments were size fractionated by agarose gel electrophoresis, and fragments between 0.8 and 2.2 kbp were recovered with Geneclean (Bio 101, Inc.) and cloned into the dephosphorylated *Bam*HI site of the vector. The second library was constructed with random DNA

fragments generated by using a nebulizer. After treatment with mung bean nuclease, phenol extraction, and ethanol precipitation, the DNA was blunt ended with T4 DNA polymerase and Klenow fragment. Fragments were size fractionated by agarose gel electrophoresis, and those in the range between 1.5 and 3 kbp were recovered and cloned into the dephosphorylated *Eco*RV site of the vector. The third library was constructed with DNA sheared by nebulization to an average size of 2 kb. The random fragments were cloned into a modified M13 vector using the double adaptor method (3).

Template preparation. Transformed bacteria were plated on LB agar containing ampicillin (100 μg/ml), X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (40 μg/ml), and IPTG (isopropyl-β-D-thiogalactopyranoside (100 μg/ml). White colonies were grown overnight at 37°C in 3 ml of 2× TY containing ampicillin (100 μg/ml) (15). Plasmid DNA templates for sequencing reaction were prepared from 1.5 ml of culture by an alkaline lysis method with minor modifications (15), followed by a polyethylene glycol 8000 precipitation. The amount of isolated DNA template was estimated on 1.0% agarose gel by comparison to serial dilutions of pBluescript II KS(+) (Stratagene). M13 phage DNA templates were prepared by using a glass fiber-filtration method (2).

DNA sequencing. Sequencing reactions at the San Martín University were performed in a Genius thermal cycler (Techne) using a Dye Terminator Cycle Sequencing Ready Reaction Kit with Ampli*Taq*r DNA polymerase (FS enzyme) (Applied Biosystems), following the protocols supplied by the manufacturer, and analyzed in an ABI prism 377 sequencer (Applied Biosystems). Single-pass sequencing was performed on each template using T7 primer. Sequencing reactions at the INTA Institute were carried out with primers 691 (GCCGCTCTA GACTAGTGGA) or 721 (GTCGACGGTATCGATAAGC) in a Perkin-Elmer 9600 thermocycler by using Dye Terminator Cycle Sequencing kits (Applied Biosystems). The fluorescent-labeled fragments were purified from the unincorporated terminators with Centri-Sep minicolumns (Princeton Separation, Adelphia, N.J.). The samples were resuspended in formamide and subjected to electrophoresis in an ABI 373 automatic sequencer. Sequencing reactions at the Department of Molecular Evolution, Uppsala University, Uppsala, Sweden, were carried out with the -21 M13 Forward primer on phage DNA templates and analyzed in an ABI prism 377 sequencer (Applied Biosystems).

All primary readings were edited to remove vector sequences from the 5′ ends and unreliable data from the 3′ ends using the program Factura (Perkin-Elmer). Sequences longer than 100 nucleotides were further analyzed.

Sequence analysis. Local homology searches were performed in a PC computer running Linux, using the BLAST suite of programs (1). BLAST searches against the national Center for Biotechnology Information (NCBI) nonredundant protein database were performed remotely using the Netblast client. The BLAST programs and the Netblast client are distributed by the NCBI (ftp://ncbi.nlm.nih.gov).

GSSs were classified according to BLASTx analysis. Those GSSs having a positive match (best BLAST hit with an expect value [*E*] of $<10^{-5}$) against a non-*Brucella* protein in the nonredundant database were considered putative genes, whereas GSSs without a significant match (best BLAST hit with an *E* of $\geq 10^{-5}$) or without hits were classified as having no database matches. To detect putative coding sequences we used the testcode algorithm developed by Fickett (6), which measures the positional randomness of a sequence and is independent of the reading frame. Fickett's test was implemented as a Perl program and used to calculate a testcode value for each sequence.

To perform the clustered orthologous group (COG) analysis, GSSs were first compared locally using BLASTx against a database containing the protein products of 21 complete bacterial genomes. The query sequence, subject sequence, score, and expect and positional information derived from this analysis were used as input for dignitor, a program that uses this information to establish relationships between the query sequence and the orthologous sequences grouped in COGs. Dignitor, the COG database, and the bacterial protein database are distributed by the NCBI (ftp://ncbi.nlm.nih.gov/pub/COG and ftp://ncbi.nlm.nih.gov/pub/tatusov).

Nucleotide sequence accession numbers. Sequence data have been deposited in the dbGSS division of GenBank under the following accession numbers: AQ752928 to AQ752940, AZ048471 to AZ49844, and AZ302564 to AZ303170.

RESULTS AND DISCUSSION

Three libraries were constructed as described in Materials and Methods, with inserts with an average size of 2 kbp. A total of 1,899 clones were successfully sequenced. After deleting vector sequences and unreliable data, an average length of 421 bases per clone was obtained and used for database searches.

TABLE 1. Database match categories of GSSs sequenced in *B. abortus*

Database match	No. of GSSs (%)
Any	1,199 (63.1)
Brucella spp.	74 (3.9)
Other α-2 proteobacteria	267 (14.0)
Other organisms	858 (45.2)
None[a]	700 (36.9)
Total	1,899 (100)

[a] GSSs without significant matches ($E \geq 10^{-5}$) to database sequences or without hits.

About 805,000 bp of genomic sequences were generated. In order to identify overlapping sequences, all sequences were subjected to contig assembly by using the Phred/Phrap/Consed system (courtesy of B. Ewing, P. Green, and D. Gordon, University of Washington, Seattle). This analysis generated 362 contigs and 978 isolated singlets (sequences having no nonvector match to any other read), representing 633,500 bp of unique genomic sequence. This represents ~20% of the estimated 3.2-Mb *B. abortus* genome.

Sequence similarities identified by the BLASTx program were considered statistically significant, with an *E* of $<10^{-5}$. Among the 1,899 GSSs obtained, 1,199 matched sequences deposited in the GenBank databases, and 700 either did not have a significant match or had no match at all (Table 1). To detect putative coding sequences within the latter group we used the testcode algorithm. About 37% of the sequences were found to be potentially coding; hence, they might represent novel or *B. abortus*-specific protein coding genes. About 3.9% of the GSSs matched *Brucella* sp. sequences, while 14.0 and 45.2% matched sequences from other members of the α-2 subgroup of the division *Proteobacteria* and from other organisms, respectively.

Taking into account that the main aim of this project was gene discovery, we grouped 1,125 GSSs with significant homology to non-*Brucella* sequences according to matches to the same database entry. After deleting redundant GSSs, 925 out of 1,125 remained as nonredundant GSSs, thus representing 925 different genes that have not been previously described in the *Brucella* genus (see http://www.iib.unsam.edu.ar/genomelab/brucella/gss.html).

To get an insight into the functional diversity of our random sequences we compared our GSSs to the sequences present in the COG database developed by Tatusov et al. (17) and derived the functional classification that is associated with each COG in the COG database. Our results show that 470 nonredundant GSS could be related to 316 COGs and could be classified into 15 broad functional categories. Some GSSs were classified into more than one category and thus were included in the "mixed function" group, the remaining sequences were grouped under the "no related COG" category (see abovementioned website). The distribution of putative genes with assigned COG category is shown in Fig 1. The largest number (18%) was related to the mixed-function group. Other categories include sequences related to general functions (15%); amino acid transport and metabolism (14%); translation, ribosomal structure, and biogenesis (11%); and energy production and conversion (8%).

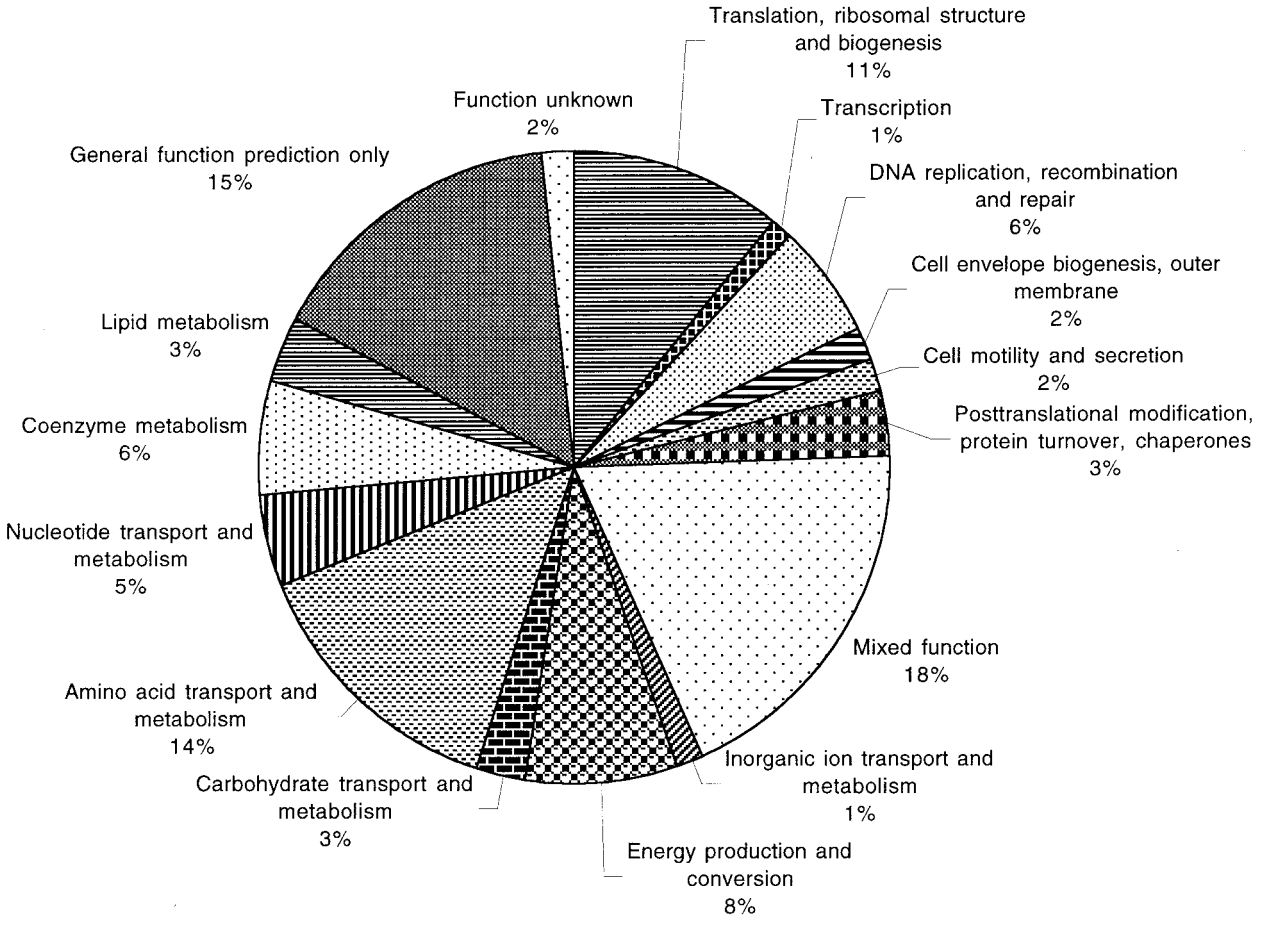

FIG. 1. Functional classification of *B. abortus* GSSs, showing the distribution of predicted genes according to their putative biological functions. A total of 470 nonredundant GSSs out of 925 with an E value of $<10^{-5}$ were classified into 15 categories. GSSs that were classified into more than one category in the COG database were included in the mixed-function group (see website mentioned in Results and Discussion).

A detailed analysis of the putative genes identified is not within the scope of this work and will certainly be done by interested researchers in the field. It is, however, worthwhile mentioning the finding of an important number of interesting matches to sequences from other bacteria such as those of the genera *Rickettsia, Agrobacterium*, and *Rhizobium*, which, like *Brucella*, belong to the α-2 subgroup of proteobacteria. In our survey, 220 out of 925 putative new genes identified showed similarities to bacterial sequences belonging to this group of bacteria (see above-mentioned website). This evidence reinforces the close phylogenetic relationship among these bacteria, a fact already pointed out by using 16S RNA sequence analysis (13).

Obtaining the complete chromosomal sequence of *B. abortus* will enable the identification of most of the potential virulence genes by comparison with other pathogens. Nevertheless, in the short run, a reasonable accomplishment might be the identification of a large proportion of its gene content by approaches like random genomic sequencing. In this study we have identified 925 genes, most of them representing novel genes for the *Brucella* spp. Among them, we have identified several *B. abortus* homologs of virulence and pathogenesis genes from other pathogens, including a homolog to both the Shda gene from *Salmonella enterica* serovar Typhimurium and the AidA-1 gene from *Escherichia coli* (GSSBru133). The AidA-1 gene encodes a cell adhesion molecule in enteropathogenic strains of *E. coli* which facilitates cell colonization. The same GSS is also homologous to the virulence gene *virG* of *Shigella* spp. which encodes a product involved in motility of the bacteria into the cell cytoplasm and spreading to the neighboring cells (11). No adhesion molecules have been described so far in *Brucella*. Thus, it will be interesting to obtain null mutants for the *Brucella* homolog of AidA-1 and study its possible role in cell adhesion and invasion. On the other hand, the *Brucella* homolog of *virG* might be involved in actin polymerization during phagocytosis.

Recently, it has been shown that some proteins implicated in the export of flagellar components are similar to components of the type III secretion machinery (8). Furthermore, it has been reported that the flagellar apparatus itself can secrete virulence factors (21). In our dataset we found several GSSs with high similarity to flagellar basal body genes like *flip* (GSSBru021), *fliM* (GSSBru838), and *flgG* (GSSBru1354) and to *fliI* (GSSBru1959), an ATPase associated to the flagellum biosynthesis (5). These findings suggest that, although *Brucella* is a nonmotile bacterium, it might have an active type III secretion machinery and/or a flagellar type secretion system not related to the biogenesis of the flagellum.

Other GSSs displayed significant homologies to genes en-

coding components of the type IV secretion machinery; among these, GSSBru0998 exhibited sequence similarity to *traL* of *E. coli*, and GSSBru162 and GSSBru1401 showed high similarity to *virB9* and *virB10* from *Bordetella pertussis* and *Rhizobium etli*, respectively. Tra genes are homologous to and colinear with genes found in the *virB* operon of *Agrobacterium tumefaciens*, which transfer the Ti plasmid from the bacteria to the nucleus of the plant cell (4), whereas in *B. pertussis* this system exports the pertussis toxin into the host cell (19). We have recently knocked out the gene homolog to *virB10* in a virulent strain of *B. abortus* and showed that this gene is essential for intracellular survival and virulence (16). In addition, it has been recently shown that a region of the *Brucella suis* genome which is highly homologous to the *A. tumefaciens virB* operon is also required for intracellular multiplication (14).

Other GSSs displayed a significant homology with several *Rhizobium meliloti* genes (see above-mentioned website) related to nodule nitrogen fixation (GSSBru219 and -1400), synthesis of nod factors (GSSBru355), nodulation protein symbiotic plasmid (GSSBru223), and nodule bacteroid differentiation (GSSBru550). Among them, GSSBru550 was highly similar to the *bacA* gene, which encodes a putative cytoplasmic membrane transport protein that is essential for the symbiosis between *Rhizobium meliloti* and alfalfa (10). Recently, it has been shown that a mutation of the *bacA* homolog in *B. abortus* decreased intracellular survival (12). Thus, it will be interesting to know the functions of these proteins in *Brucella*. Also, it will be interesting to know the function of the *R. meliloti* homolog genes in *Brucella*, a mammalian intracellular pathogen.

Interestingly, we present here evidence showing a number of GSSs that are highly similar to both *Agrobacterium* virulence and *Rhizobium* nodulation genes. Why genes that seem to be very specific for the interaction of these soil bacteria with their hosts are conserved in *Brucella* is a question that remains to be answered but suggests that the ancestors of these organisms might have shared a common environment.

Particularly in bacteria, where gene density is high, it is increasingly apparent that random sequencing of genomic DNA is an efficient way to identify protein homologs. To date (October 2000), there are 333 protein sequences reported in the GenBank database for the *Brucella* genus; 198 of them correspond specifically to *B. abortus*. (htpp://www.ncbi.nlm.nih.gov/Taxonomy/). In the present study, random sequences totalling only 19.8% of the genome resulted in the identification of 925 putative new *Brucella* genes, which represents a large improvement over the known genes reported to date. These data, already available to researchers interested in the field, might provide new insights into the biology of *Brucella*.

ACKNOWLEDGMENTS

Daniel O. Sánchez and Ruben Zandomeni contributed equally to this work.

We thank J. J. Cazzulo and Diego Comerci for critical reading of the manuscript. We also thank Diego Comerci for providing DNA from *B. abortus* for the construction of the M13 library. We are indebted to Nancy Lopez, Fernanda Peri, Diego Rey Serantes, and Rodrigo Pavón for their valuable help in DNA purification and sequencing and to Martin Sarachu for computer assistance.

This work was supported by grants from the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina; the Ministerio de Cultura y Educación, Argentina; the Agencia Nacional de Promoción Científica y Tecnológica, Argentina (PICT97-01767), the Swedish Foundation for Strategic Research (to S.G.E.A.); and the World Bank/UNDP/WHO Special Program for Research and Training in Tropical Diseases (TDR). D.O.S., R.Z., A.C.C.F., and R.A.U. are members of the Research Career of the CONICET; O.G. is member of the Research Career of the Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC), Buenos Aires, Argentina. R.Z., S.C., O.L.R., and O.G. are members of the INTA.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Andersson, B., J. Lu, K. E. Edwards, D. M. Muzny, and R. A. Gibbs.** 1996. Method for 96-well M13 DNA template preparations for large-scale sequencing. BioTechniques **20:**1022–1027.
3. **Andersson, B., M. A. Wentland, J. Y. Ricafrente, W. Liu, and R. A. Gibbs.** 1996. A "double adaptor" method for improved shotgun library construction. Anal. Biochem. **236:**107–113.
4. **Christie, P. J.** 1997. *Agrobacterium tumefaciens* T-complex transport apparatus: a paradigm for a new family of multifunctional transporters in eubacteria. J. Bacteriol. **179:**3085–3094.
5. **Fan, F., and R. M. Macnab.** 1996. Enzymatic characterization of FliI. An ATPase involved in flagellar assembly in *Salmonella typhimurium*. J. Biol. Chem. **271:**31981–31988
6. **Fickett, J. W.** 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. **10:**5303–5318.
7. **Finlay, B. B., and S. Falkow.** 1997. Common themes in microbial pathogenicity revisited. Microbiol. Mol. Biol. Rev. **61:**136–169.
8. **Galan, J. E., and A. Collmer.** 1999. Type III secretion machines: bacterial devices for protein delivery into host cells. Science **284:**1322–1328.
9. **Hensel, M., J. E. Shea, C. Gleeson, M. D. Jones, E. Dalton, and D. W. Holden.** 1995. Simultaneous identification of bacterial virulence genes by negative selection. Science **269:**400–403.
10. **Ichige, A., and G. C. Walker.** 1997. Genetic analysis of the *Rhizobium meliloti bacA* gene: functional interchangeability with the *Escherichia coli sbmA* gene and phenotypes of mutants. J. Bacteriol. **179:**209–216.
11. **Kotloff, K. L., F. R. Noriega, T. Samandari, M. B. Sztein, G. A. Losonsky, J. P. Nataro, W. D. Picking, E. M. Barry, and M. M. Levine.** 2000. *Shigella flexneri* 2a strain CVD 1207, with specific deletions in *virG*, *sen*, *set*, and *guaBA*, is highly attenuated in humans. Infect. Immun. **68:**1034–1039.
12. **LeVier, K., R. W. Phillips, V. K. Grippe, R. M. Roop, 2nd, and G. C. Walker.** 2000. Similar requirements of a plant symbiont and a mammalian pathogen for prolonged intracellular survival. Science **287:**2492–2493.
13. **Moreno, E., E. Stackebrandt, M. Dorsch, J. Wolters, M. Busch, and H. Mayer.** 1990. *Brucella abortus* 16S rRNA and lipid A reveal a phylogenetic relationship with members of the alpha-2 subdivision of the class *Proteobacteria*. J. Bacteriol. **172:**3569–3576.
14. **O'Callaghan, D., C. Cazevieille, A. Allardet-Servent, M. L. Boschiroli, G. Bourg, V. Foulongne, P. Frutos, Y. Kulakov, and M. Ramuz.** 1999. A homologue of the *Agrobacterium tumefaciens* VirB and *Bordetella pertussis* Ptl type IV secretion systems is essential for intracellular survival of *Brucella suis*. Mol. Microbiol. **33:**1210–1220.
15. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
16. **Sieira, R., D. J. Comerci, D. O. Sánchez, and R. A. Ugalde.** 2000. A homologue of an operon required for DNA transfer in *Agrobacterium* is required in *Brucella abortus* for virulence and intracellular multiplication. J. Bacteriol. **182:**4849–4855.
17. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631–7.
18. **Ugalde, R. A.** 1999. Intracellular lifestyle of *Brucella spp*. Common genes with other animal pathogens, plant pathogens, and endosymbionts. Microbes Infect. **1:**1211–1219.
19. **Weiss, A. A., F. D. Johnson, and D. L. Burns.** 1993. Molecular characterization of an operon required for pertussis toxin secretion. Proc. Natl. Acad. Sci. USA **90:**2970–2974.
20. **Young, E. J.** 1983. Human brucellosis. Rev. Infect. Dis. **5:**821–842.
21. **Young, G. M., D. H. Schmiel, and V. L. Miller.** 1999. A new pathway for the secretion of virulence factors by bacteria: the flagellar export apparatus functions as a protein-secretion system. Proc. Natl. Acad. Sci. USA **96:**6456–6461.

Editor: V. J. DiRita