Efficient use of a small genome to generate antigenic diversity in tick-borne ehrlichial pathogens

Kelly A. Brayton*†, Donald P. Knowles*‡, Travis C. McGuire*, and Guy H. Palmer*

*Program in Vector Borne Diseases, Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, WA 99164-7040; and ‡Animal Disease Research Unit, Agricultural Research Service, U.S. Department of Agriculture, Ames, IA 50010

Communicated by Janice M. Miller, Agricultural Research Service, U.S. Department of Agriculture, Ames, IA, February 5, 2001 (received for review October 1, 2000)

Ehrlichiae are responsible for important tick-transmitted diseases, including anaplasmosis, the most prevalent tick-borne infection of livestock worldwide, and the emerging human diseases monocytic and granulocytic ehrlichiosis. Antigenic variation of major surface proteins is a key feature of these pathogens that allows persistence in the mammalian host, a requisite for subsequent tick transmission. In *Anaplasma marginale* pseudogenes for two antigenically variable gene families, *msp*2 and *msp*3, appear in concert. These pseudogenes can be recombined into the functional expression site to generate new antigenic variants. Coordinated control of the recombination of these genes would allow these two gene families to act synergistically to evade the host immune response.

Antigenic variation in major surface proteins of tick-borne bacterial pathogens is a primary mechanism for evasion of the host immune response and results in persistent infection. A number of different mechanisms have been reported. For instance, bacteria of the genus *Borrelia* generate antigenic diversity of the *vmp*/*vls* coat proteins through recombination from a tandem array of silent partial pseudogene cassettes into a telomeric expression site on a linear plasmid (1, 2). Ehrlichial genogroup I pathogens, *Ehrlichia chaffeensis*, *E. canis,* and *Cowdria ruminantium*, have recently been shown to contain tandemly repeated copies of the *omp*1/*map*1 gene family (3–6). These *omp*1/*map*1 tandem repeats are complete gene copies, and multiple copies may be transcriptionally active at a given time, resulting in polymorphic protein expression (3).

In contrast to a tandem array of genes or pseudogenes for antigenically variable proteins, *Anaplasma marginale*, a member of ehrlichial genogroup II, contains immunodominant major surface protein 2 (*msp*2) and *msp*3 families that include 10 or more variable genes widely dispersed throughout the 1.2-Mb genome (7–9). Recently we reported the identification of an operon of four ORFs, containing the *msp*2 gene at the 3′ terminus (10). The other ORFs of the operon occur as a single copy in the *A. marginale* genome, and the operon has been demonstrated to be a functional expression site for full-length *msp*2 transcripts. These *msp*2 transcripts have highly conserved 5′ and 3′ ends, interrupted by a central hypervariable region characterized by substitutions, insertions, and deletions. The hypervariable region encodes a diverse array of B cell epitopes that result in evasion of the host immune system. Similarly, MSP3 is also structurally and antigenically variable (11). Our data show that aside from the expression site in the operon, *msp*2 does not occur as a full-length gene, but rather as partial pseudogene cassettes, each containing a different hypervariable region and a portion of the 5′ and 3′ conserved regions. We have identified nine pseudogenes for *msp*2 and shown that these pseudogenes recombine into the operon expression site to generate a new hypervariable sequence. Partial pseudogene cassettes appear for the *msp*3 gene family as well, and the pseudogenes for the two gene families often appear close together. The two pseudogene families have the same 5′ sequence, indicating that they use the same mechanism to regulate recombination into the expression site.

Materials and Methods

Southern Analysis. Genomic DNA of the *A. marginale* South Idaho strain was digested with the use of *Kpn*I and separated on a 0.7% agarose gel. The blot was initially hybridized with a *msp*2 5′ end-specific probe corresponding to bp 2–335 and then stripped and rehybridized with an orf2-specific probe corresponding to bp 6–359. Generation and digoxigenin labeling of the probes, hybridization, and detection were as recommended by the manufacturer of the PCR labeling kit (Roche Molecular Biochemicals). High stringency wash conditions were as follows: two washes in 2× SSC, 0.1% SDS (wt/vol) at room temperature, one wash in the same buffer at 65°C, and a final wash in 0.2× SSC, 0.1% SDS (wt/vol) at 65°C (1× SSC = 0.15 M sodium chloride/0.015 M sodium citrate, pH 7). All washes lasted 15 min.

Bacterial Artificial Chromosome (BAC) Library Construction and Manipulation. Blood was collected from calf 836 during acute *A. marginale* (St. Maries strain) rickettsemia. Erythrocytes were isolated with the use of Histopaque (Sigma) and embedded in agarose blocks, and cells were lysed within the agarose blocks with the use of proteinase K and SDS (12). *A. marginale* genomic DNA was partially digested with *Hin*dIII, size selected on pulse field gels, ligated into the vector pBELOBAC11, and electroporated into *Escherichia coli* strain DH10B. A total of 1,536 BAC clones were arrayed into 384 well plates with an average insert size of 110 Kb. BAC 2439 was selected for sequence analysis after probing with a digoxigenin-labeled (Roche Molecular Biochemicals) *msp*2 probe (bp 375–965). Random shotgun libraries were constructed from partially digested BAC 2439 DNA. The randomly generated fragments were size selected, cloned into pCRScript (Stratagene), and electroporated into *E. coli* strain XL-1Blue. Insert DNA was sequenced with the use of BigDye terminator chemistry on an ABI 377XL-96 instrument (PE-Applied Biosystems). Data were assembled and analyzed by using SEQUENCHER (Gene Codes, Ann Arbor, MI) and PHRED & PHRAP software (Sanger Centre, Cambridge, UK). When required, gene walking or direct BAC sequencing was performed to ensure a minimum of 2× coverage, with an overall average of 3× coverage. The finished sequence contains 44,557 bp (accession number AF305077).

PCR Cloning of Pseudogenes and Operon-Linked *msp2* Fragments. Ticks fed on *A. marginale* (South Idaho strain)-infected calf 824 subsequently transmitted the infection to calf 828. Blood was taken during acute rickettsemia for both animals, and genomic DNA was isolated with a Puregene DNA extraction kit (Gentra Systems). Total RNA was isolated with the use of TRIzol

Abbreviations: BAC, bacterial artificial chromosome; msp, major surface protein.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF305077 and AF305503–AF305508).

†To whom reprint requests should be addressed. E-mail: kbrayton@vetmed.wsu.edu.

www.pnas.org/cgi/doi/10.1073/pnas.071056298

(GIBCO/BRL). RNA was treated with DNase I, followed by cDNA synthesis with random hexamers, with the use of the Thermoscript reverse transcription–PCR kit from GIBCO/BRL. The primers used to amplify the operon-linked hypervariable regions were orf 2 forward primer TCCTACCAAGCGTCTTTTCCCC and *msp*2 reverse primer: TTACCACCGATACCAGCACAA. PCR was performed with DNA, RNA (as negative control), or cDNA and fragments cloned with the pMOSBlue cloning system (Amersham Pharmacia). Inserts were sequenced in both directions with the Big Dye kit and an ABI PRISM automated sequencer (PE-Applied Biosystems). Sequences were compiled and analyzed with the VECTOR NTI (InforMax, North Bethesda, MD) and GCG (University of Wisconsin) software packages.

Results and Discussion

The *msp*2 gene is estimated to have 10 or more copies (7), whereas the operon containing the expressed copy of *msp*2 is reported to be a single copy (10). Importantly, as shown in Fig. 1*A*, the 5′ end of *msp*2 also occurs only as a single copy. Southern analysis of *Kpn*I-digested *A. marginale* DNA with a *msp*2 5′ end probe (bp 2–335) detects two fragments of 1.2 and 4 kb. These same two bands are detected when the blot is rehybridized with an operon-specific probe (ORF2, bp 6–359) and result from a *Kpn*I polymorphism in the hypervariable region of *msp*2 (Fig. 1*B*). Detection of both the 1.2- and 4-kb fragments reflects the oligoclonal nature of the infection *in vivo*. These data demonstrate that the operon is the only expression site for full-length *msp*2 transcripts. Therefore the remaining copies of *msp*2 throughout the genome were hypothesized to be truncated pseudogenes.

To identify potential pseudogenes, a BAC library from the St. Maries strain of *A. marginale* was screened for the presence of *msp*2. A 45-kb positive clone (BAC 2439) was sequenced, and two pseudogenes for *msp*2 and two pseudogenes for *msp*3 (Fig. 2) were found. A third *msp*2 pseudogene was previously cloned from the Florida strain in conjunction with *msp*3 (clone 3-11; ref. 8) but remained unrecognized as such until comparison with the sequences in BAC 2439. This third pseudogene is on the opposite strand of DNA, 321 bp downstream from the 3′ end of the *msp*3-11 coding sequence (Fig. 2). Although these three *msp*2 pseudogenes do not have uniform start and stop positions, their general structures are similar: they contain a portion of the 5′ conserved region, continue through the hypervariable region, and contain a portion of the 3′ conserved region. Thus, these pseudogenes have a structure capable of recombination into the operon expression site to generate new hypervariable regions of *msp*2.

The motif of a *msp*3 pseudogene on one strand and a *msp*2 pseudogene on the opposite strand was observed both in BAC 2439 and in the *msp*3-11 clone. The 321 bp separating *msp*2 from *msp*3 in each of these clones was identical, even though the respective *msp*2 and *msp*3 coding regions themselves were not identical between the two clones. Furthermore, ≈600 bp 5′ to *msp*2 pseudogenes 1 and 3 and *msp*3 pseudogenes 1 and 2 was highly conserved (93.8–99.7% identity). This 600-bp 5′ flanking region is indicated in Fig. 2 by ovals. *Msp*2 pseudogene 2 in BAC 2439 had a 5′ extension that was not seen in any of the other pseudogenes, such that it started at bp 236 relative to pCKR11.2. Southern analysis indicated that this 5′ extension occurred only once in the genome and was an aberration that was specific to the St. Maries strain of *A. marginale* (data not shown). This pseudogene is incapable of encoding a full-length *msp*2 gene, and it differs from the other *msp*2 pseudogenes in that it does not have the conserved 600-bp 5′ flanking region.

The two regions of sequence conservation, the 600-bp 5′ region and the 321-bp 3′ region, flanking the *msp*2 pseudogenes allowed sets of primers to be designed that could specifically amplify *msp*2 pseudogenes. Primer positions are shown in Fig. 2 relative to *msp*2 pseudogene 3. Five distinct DNA fragments were obtained by PCR from *A. marginale* (South Idaho, calf 824) genomic DNA, with four primer combinations (Fig. 3). These fragments were cloned, and 10 clones from each PCR reaction were sequenced. With this strategy, six additional pseudogenes were identified (Fig. 4). The ORFs for these pseudogenes start uniformly with the sequence PYQGYHSMLTALE, with identity to full-length MSP2 starting with the serine corresponding to amino acid 112 [numbering based on pCKR11.2, a full-length *msp*2 gene (7)]. The pseudogenes are 186–194 aa in length, with the variation in length due to differences in the hypervariable region. The four pseudogenes generated with primer R1 (Fig. 2) in the 321-bp intergenic region uniformly ended at amino acid 284 (Fig. 4). The remaining two pseudogenes were generated with primer R2 in the 3′ conserved region of *msp*2 (Figs. 2 and 4).

The close positioning of *msp*2 and *msp*3 pseudogenes in a tail-to-tail arrangement was a recurring motif in the genome. In addition to *msp*2 pseudogenes 2 and 3 found in this arrangement,

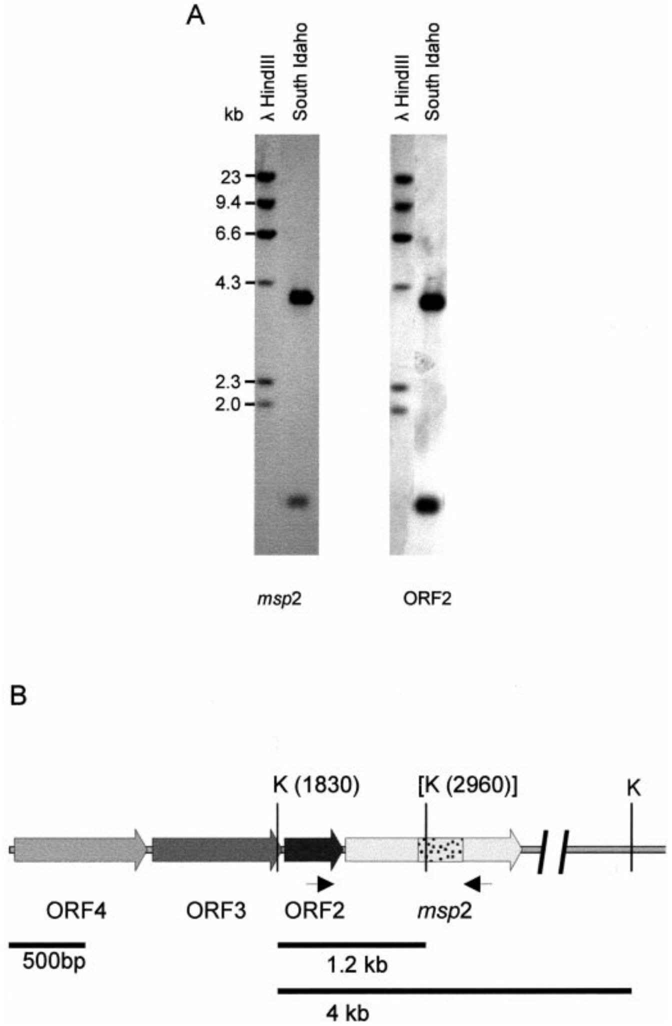


Fig. 1. (A) Presence of a single genomic expression site for *msp*2. (*Left*) A blot hybridized with a *msp*2 5′ end-specific probe corresponding to bp 2-335. (*Right*) The same blot hybridized with an orf2-specific probe corresponding to bp 6-359. Lambda *Hin*dIII markers are shown. (*B*) Schematic representation of the *msp*2 operon. The hypervariable region of *msp*2 is stippled. Positions of *Kpn*I sites are shown (K). The *Kpn*I site in the hypervariable region is polymorphic. Arrows indicate the positions of primers used to amplify DNA and cDNA clones.

MICROBIOLOGY

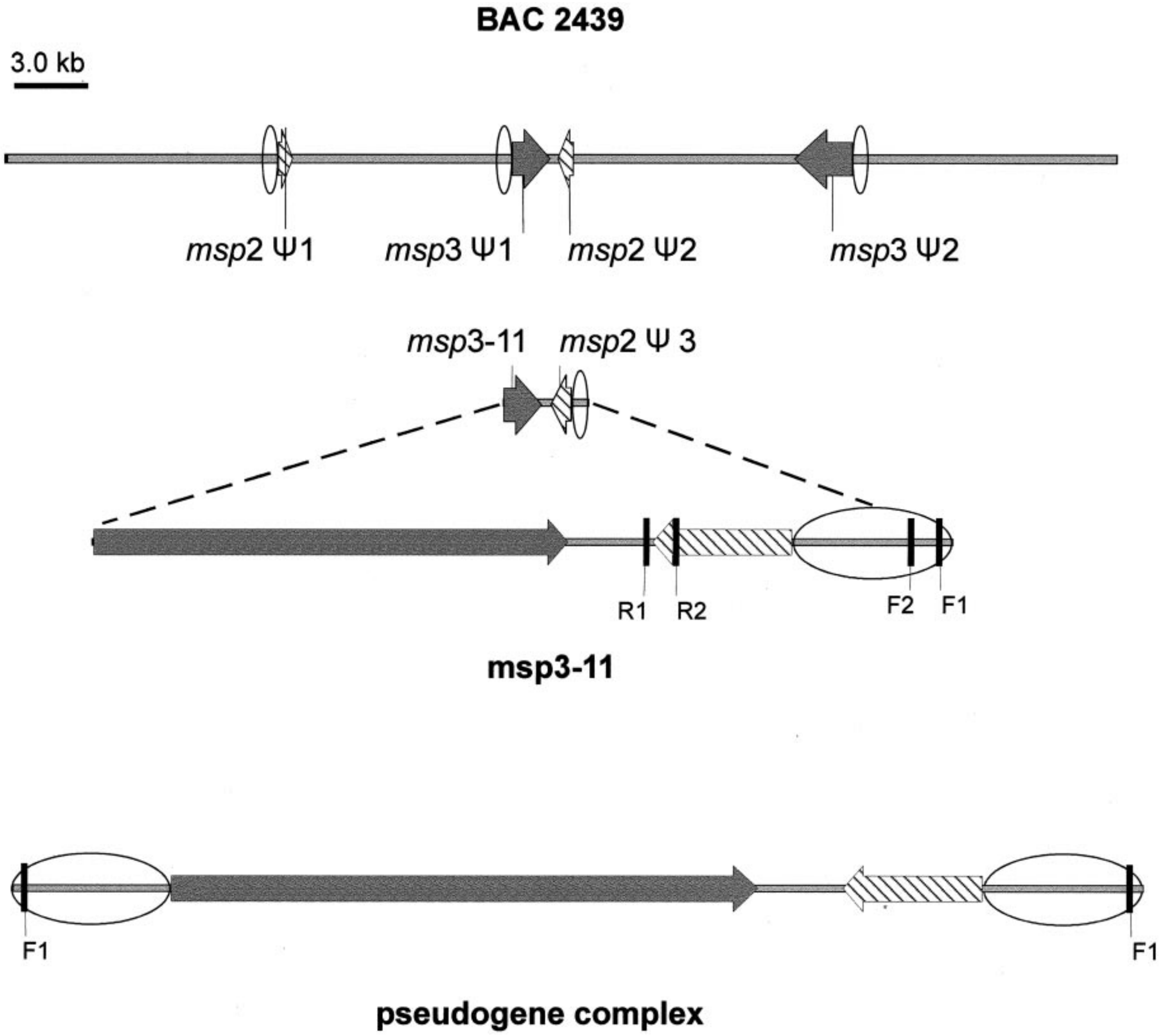


Fig. 2. Schematic representation of genomic pseudogenes. BAC 2439 is 45 kb in length. *Msp*3-11 is 3263 bp in length. It is shown to the same scale as BAC 2439 for comparison and is enlarged to show the positions of primers F1, F2, R1, and R2 used for amplification and cloning of pseudogenes. The sequence of the primers is as follows: F1: GCACCAAAGAATATAGCTGTAAATAC; F2: CCCAGCTTCTGCACAAAC; R1: TGTTGCCCGCCATCC; R2: CTCTAGCACCTTCAGCATC. The complete pseudogene complex is illustrated. *Msp*2 pseudogenes are striped and *msp*3 pseudogenes are dark gray. The orientation of pseudogenes is indicated by the direction of the arrow. Ovals indicate the presence of the conserved 600-bp sequence.

four of the six *msp*2 pseudogenes cloned in this study were amplified with the use of a primer specific for the intergenic region between these two genes, indicating that they, too, were arranged in the same manner. Furthermore, amplification of *A. marginale* DNA with the use of a single primer, F1, specific for the 5′ flanking region generates a product of ≈4.3 kb, the expected size of the pseudogene complex, corresponding to a 5′ flanking region, a *msp*3 pseudogene, the 321-bp intergenic region, with a *msp*2 pseudogene, and a second 5′ flanking region on the opposite strand (Fig. 2).

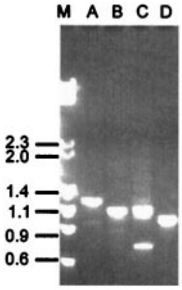
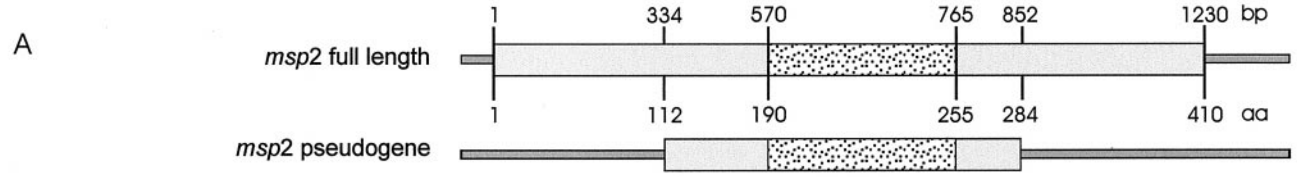


Fig. 3. Pseudogene amplicons generated with primers F1, F2, R1, and R2. Amplicons were generated with the use of *A. marginale* South Idaho DNA from calf 824. The amplicon in Lane A was generated with primers F1 and R1, Lane B used primers F2 and R1, Lane C used primers F1 and R2, and Lane D used primers F2 and R2. Lane M contains the molecular weight standards λ*Hin*dIII and $\phi\chi$*Hae*III.

The pseudogenes are not functional transcription units, as there are no promoter consensus sequences within 350 bp of the first methionine of each pseudogene. Furthermore, polyclonal antibodies that detect a full-length MSP2 (42 kDa) in Western blots do not detect any proteins in the region of 18 kDa, the expected size of the product if the pseudogenes were expressed, indicating that the pseudogenes are not functional coding regions (data not shown). Consequently pseudogene expression would require recombination into the expression site.

To address the question of whether these pseudogenes recombine into the operon expression site of *msp*2, we examined the expression site at various time points during infection. *A. marginale* DNA was isolated from two calves infected with the South Idaho strain and amplified with the use of primers



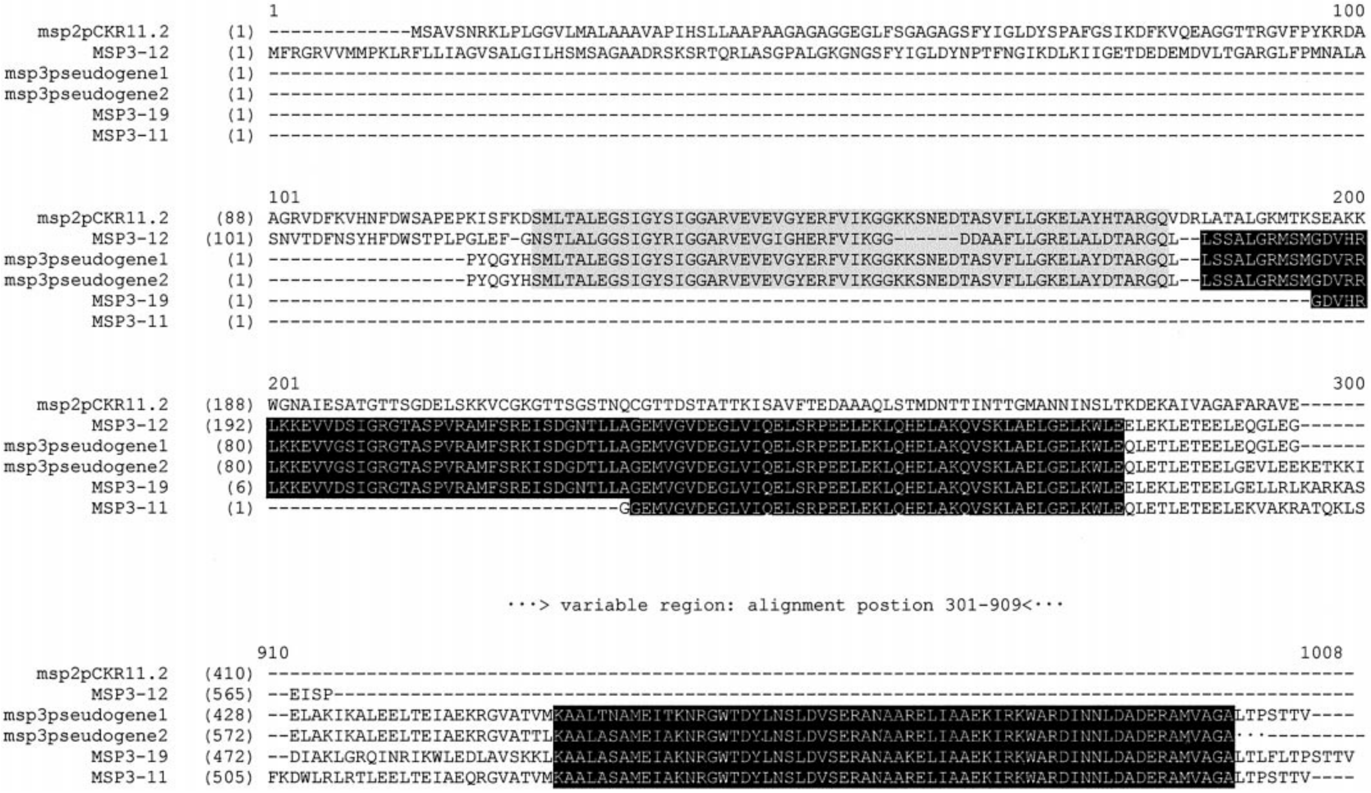
B

```
                          1                                                                                        90
   msp2pseudogene1    (1) ------------------------------------------------------------------------------------------
   msp2pseudogene2    (1) --------------------------------------------------------------------------------FPYKRDAAGR
   msp2pseudogene3    (1) ------------------------------------------------------------------------------------------
     msp2pseudoA-2    (1) ------------------------------------------------------------------------------------------
     msp2pseudoA-3    (1) ------------------------------------------------------------------------------------------
    msp2pseudoC1-1    (1) ------------------------------------------------------------------------------------------
     msp2pseudoD-2    (1) ------------------------------------------------------------------------------------------
    msp2pseudoB-10    (1) ------------------------------------------------------------------------------------------
    msp2pseudoA-10    (1) ------------------------------------------------------------------------------------------
      msp2pCKR11.2    (1) MSAVSNRKLPLGGVLMALAAAVAPIHSLLAAPAAGAGAGGEGLFSGAGAGSFYIGLDYSPAFGSIKDFKVQEAGGTTRGVFPYKRDAAGR

                          91                                                                                      180
   msp2pseudogene1    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
   msp2pseudogene2   (11) VDFKVHNFDWSAPEPKISFKDSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
   msp2pseudogene3    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
     msp2pseudoA-2    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVGIGHERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
     msp2pseudoA-3    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
    msp2pseudoC1-1    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
     msp2pseudoD-2    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
    msp2pseudoB-10    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYHTARGQVDRLATALGKM
    msp2pseudoA-10    (1) ---------------PYQGYHSMLTALEGSIGYSIGGARVEVGIGHERFVIKGGKKSNEDTASVFLLGKELAYDTARGQVDRLATALGKM
      msp2pCKR11.2   (91) VDFKVHNFDWSAPEPKISFKDSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYHTARGQVDRLATALGKM

                          181                                                                                     270
   msp2pseudogene1   (76) TKSEAKKWGNAIESATGTTSGDELSKKVCGKGTTSG--NQCGVNATSGSTN-NG-KLSTVFNTDGAEAISSMDTTASGTSNTISLQGMAG
   msp2pseudogene2  (101) TKSEAKKWGNAIESATGTTSGDELSKKVCGKGEGSNGTKKCGTTDSTATT------KISEVFTEGTDTLLSVEG----NKDTINLQGMAN
   msp2pseudogene3   (76) TKGEAKKWGKAVE---GTTNGEKVSQNVCGKGEGSNGTKKCGTTDSTATT------KISEVFTEGTDTLLSVEG----NKDTINLQGMAN
     msp2pseudoA-2   (76) TKSEAKKWGTTIENTTG---SDELSKKVCGKGDSSK----CGVNANSGSTT--ATTKISAVFTEGTEAISSMET-ASGASGTISLQGMAT
     msp2pseudoA-3   (76) TKSEAKKWGNAIESVTGTTNGQTVSQKVCGKGEGSNGTKKCGTNDGTTATQ----HKISEVFTEGTDTATLLSA----AGDTINTTGMAG
    msp2pseudoC1-1   (76) TKSEAKKWGNAIESATGTTNGDELSKKVCGKGTTSGTTNQCGKNTADGTTT-----KISAVFTEGTEAISSMET-ASGASGTISLQGMAG
     msp2pseudoD-2   (76) TKSEAKKWGNAIESATGTTNGDELSKKVCGKGEGSNGTKKCGTNDGTTATQ----HKISEVFTEGTDTATLLSA----AGDTINTTGMAG
    msp2pseudoB-10   (76) TKSEAKKWGNAIESATGTTNGEKVSQKVCGNGTGSSG-TQCGKNSGDTNGSSTTQHKISAVFTEGTEAISSMET-ASGASGTISLQGMAT
    msp2pseudoA-10   (76) TKSEAKKWGTTIENTTG---SHELSKKVCGKGDSSK----CGVNANSGSTT--ATTKISAVFTEGTEAISSMET-DSGASGTISLQGMAT
      msp2pCKR11.2  (181) TKSEAKKWGNAIESATGTTSGDELSKKVCGKGTTSGSTNQCGTTDSTATT------KISAVFTEDAAAQLSTMD-----NTTINTTGMAN

                          271                                                                                     360
   msp2pseudogene1  (162) NINSLFLLTQRHGCSHTFLSFALPYTHIINRAWVDALVEEG-------------------------------------------------
   msp2pseudogene2  (181) NINNLSKEDKAVVAGAFARAVEGAESCGELG-----------------------------------------------------------
   msp2pseudogene3  (153) NINNLSKEDKAVVAGAFARAVEGAESCGELG-----------------------------------------------------------
     msp2pseudoA-2  (156) NINGLSKEDKAVVAGAFARAVEGAESCGELG-----------------------------------------------------------
     msp2pseudoA-3  (158) NINGLSKEEKAIVAGAFARAVEGAESCGELG-----------------------------------------------------------
    msp2pseudoC1-1  (160) N-----------------------------------------------------------------------------------------
     msp2pseudoD-2  (158) N-----------------------------------------------------------------------------------------
    msp2pseudoB-10  (164) NINGLSKEDKAVVAGAFARAVEGAESCGELG-----------------------------------------------------------
    msp2pseudoA-10  (156) NINGLSKEDKAVVAGAFARAVKGAESCGELG-----------------------------------------------------------
      msp2pCKR11.2  (260) NINSLTKDEKAIVAGAFARAVEGAEVIEVRAIGSTSVMLNACYDLLTDGIGVVPYACAGIGGNFVSVVDGHINPKFAYRVKAGLS·····
```

Fig. 4. (*A*) Schematic representation of full-length *msp*2 compared with the *msp*2 pseudogenes. Numbering of full-length *msp*2 is indicated in base pairs and amino acids. The hypervariable region is indicated by stippling. (B) The deduced amino acid sequences of the *msp*2 pseudogenes obtained in this study are shown in comparison to the full-length *msp*2 sequence. The 5′ and 3′ conserved regions are on a gray background and flank the hypervariable region. The full-length pCKR11.2 continues for 134 aa (not shown). Pseudogenes 1 and 2 are from an *A. marginale* St. Maries strain, pseudogene 3 is from an *A. marginale* Florida strain (8), and the remaining pseudogenes are from an *A. marginale* South Idaho strain and are lettered A-D as denoted in Fig. 3.

spanning the hypervariable region of *msp*2 and linking these products to the operon, thus establishing that they were derived from the operon expression site (primer positions indicated in Fig. 1). PCR-derived amplicons were cloned, and genomic clones were sequenced. Pseudogene A-3 (AF305503) had a corresponding operon-linked genomic clone (828-45 g), demonstrating that pseudogenes, which reside elsewhere in the South Idaho strain genome, are recombined into the expression site during infection with this strain. To verify that the recombined pseudogene A-3 gave rise to functional transcripts, the experiment was repeated with RNA as the template, and the fragments from the resulting reverse transcription–PCR were cloned. Sequence analysis identified two clones (828-1c and 824-52c) that corresponded to A-3 and 828-45 g. These results demonstrate that the pseudogenes serve as a source of new hypervariable regions for the operon expression site of *msp*2.

MSP3, like MSP2, is a variable immunodominant surface protein belonging to a multigene family. BAC 2439 sequence data suggest that antigenic diversity of MSP3 is generated by the same mechanism as MSP2. The previously identified *msp*3-12 clone (8) appears to encode an authentic 5′ end for the gene and is truncated toward the 3′ end, whereas *msp*3-11 and 3-19 (8) start at aa positions 224 and 187, respectively, relative to *msp*3-12 (Fig. 5). When these *msp*3 sequences are spliced *in silico* they do not encode a fused domain of the appropriate size (86 kDa). The two *msp*3 genes found in this study (*msp*3-1 and *msp*3-2) start at



```
                       1                                                                                                  100
   msp2pCKR11.2    (1) -------------MSAVSNRKLPLGGVLMALAAAVAPIHSLLAAPAAGAGAGGEGLFSGAGAGSFYIGLDYSPAFGSIKDFKVQEAGGTTRGVFPYKRDA
        MSP3-12    (1) MFRGRVVMMPKLRFLLIAGVSALGILHSMSAGAADRSKSRTQRLASGPALGKGNGSFYIGLDYNPTFNGIKDLKIIGETDEDEMDVLTGARGLFPMNALA
msp3pseudogene1    (1) ----------------------------------------------------------------------------------------------------
msp3pseudogene2    (1) ----------------------------------------------------------------------------------------------------
        MSP3-19    (1) ----------------------------------------------------------------------------------------------------
        MSP3-11    (1) ----------------------------------------------------------------------------------------------------

                       101                                                                                                200
   msp2pCKR11.2   (88) AGRVDFKVHNFDWSAPEPKISFKDSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYHTARGQVDRLATALGKMTKSEAKK
        MSP3-12  (101) SNVTDFNSYHFDWSTPLPGLEF-GNSTLALGGSIGYRIGGARVEVGIGHERFVIKGG------DDAAFLLGRELALDTARGQL--LSSALGRMSMGDVHR
msp3pseudogene1    (1) ------------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQL--LSSALGRMSMGDVRR
msp3pseudogene2    (1) ------------------PYQGYHSMLTALEGSIGYSIGGARVEVEVGYERFVIKGGKKSNEDTASVFLLGKELAYDTARGQL--LSSALGRMSMGDVRR
        MSP3-19    (1) -----------------------------------------------------------------------------------------------GDVHR
        MSP3-11    (1) ----------------------------------------------------------------------------------------------------

                       201                                                                                                300
   msp2pCKR11.2  (188) WGNAIESATGTTSGDELSKKVCGKGTTSGSTNQCGTTDSTATTKISAVFTEDAAAQLSTMDNTTINTTGMANNINSLTKDEKAIVAGAFARAVE------
        MSP3-12  (192) LKKEVVDSIGRGTASPVRAMFSREISDGNTLLAGEMVGVDEGLVIQELSRPEELEKLQHELAKQVSKLAELGELKWLEELEKLETEELEQGLEG------
msp3pseudogene1   (80) LKKEVVGSIGRGTASPVRAMFSRKISDGDTLLAGEMVGVDEGLVIQELSRPEELEKLQHELAKQVSKLAELGELKWLEQLETLETEELEQGLEG------
msp3pseudogene2   (80) LKKEVVGSIGRGTASPVRAMFSRKISDGDTLLAGEMVGVDEGLVIQELSRPEELEKLQHELAKQVSKLAELGELKWLEQLETLETEELGEVLEEKETKKI
        MSP3-19    (6) LKKEVVDSIGRGTASPVRAMFSREISDGNTLLAGEMVGVDEGLVIQELSRPEELEKLQHELAKQVSKLAELGELKWLEELEKLETEELGELLRLKARKAS
        MSP3-11    (1) --------------------------------GGEMVGVDEGLVIQELSRPEELEKLQHELAKQVSKLAELGELKWLEQLETLETEELEKVAKRATQKLS


                                 ···> variable region: alignment postion 301-909<···

                      910                                                                                               1008
   msp2pCKR11.2  (410) ---------------------------------------------------------------------------------------------------
        MSP3-12  (565) --EISP---------------------------------------------------------------------------------------------
msp3pseudogene1  (428) --ELAKIKALEELTEIAEKRGVATVMKAALTNAMEITKNRGWTDYLNSLDVSERANAARELIAAEKIRKWARDINNLDADERAMVAGALTPSTTV----
msp3pseudogene2  (572) --ELAKIKALEELTEIAEKRGVATTLKAALASAMEIAKNRGWTDYLNSLDVSERANAARELIAAEKIRKWARDINNLDADERAMVAGA···---------
        MSP3-19  (472) --DIAKLGRQINRIKWLEDLAVSKKLKAALASAMEIAKNRGWTDYLNSLDVSERANAAKELIAAEKIRKWARDINNLDADERAMVAGALTLFLTPSTTV
        MSP3-11  (505) FKDWLRLRTLEELTEIAEQRGVATVMKAALASAMEIAKNRGWTDYLNSLDVSERANAARELIAAEKIRKWARDINNLDADERAMVAGALTPSTTV----
```

Fig. 5. Alignment of the deduced amino acid sequences of known *msp*3 (pseudo)genes in comparison with *msp*2. The gray region highlights a region of sequence similarity between *msp*2 and *msp*3. The black background indicates regions of sequence conservation in *msp*3 flanking the hypervariable region. The hypervariable region of *msp*3 from alignment position 301–909 is not shown.

position 130 relative to *msp*3-12. These two coding sequences start with the sequence PYQGYHSMLTALE and continue for 51 aa with 100% identity to the beginning of the *msp*2 pseudogenes (Fig. 5). Although these four putative pseudogenes (*msp*3-1, -2, -11, and -19) are quite divergent, they have two regions of high sequence similarity: a stretch of 93 aa of 100% sequence identity near the amino terminus, following the region of identity to *msp*2, and a stretch of 61 aa near the carboxy terminus (95–100% identity). Like *msp*2, these pseudogenes encode a portion of a 5′ conserved region, a divergent region with changes in sequence and length, and a 3′ conserved region. An additional similarity is the 600 bp of the 5′ flanking region that is highly conserved for *msp*2 pseudogenes 1 and 3 and for *msp*3 pseudogenes 1 and 2. The same 600-bp region 5′ to the *msp*2 and *msp*3 pseudogenes is a substantially sized repeat that likely has a role in the recombination of both *msp*2 and *msp*3 pseudogenes. As the 5′ regions of the *msp*3 and *msp*2 pseudogenes are the same, specificity for each target expression site must be ensured by the respective 3′ recombinatorial site in the coding region of each gene.

Recombination of surface proteins to generate antigenic diversity occurs in several other bacterial pathogens. Tick-borne pathogens of the genus *Borrelia* undergo recombination of the *vmp*/*vls* genes (1, 2). Two distinct features of this system are the telomeric placement of the expression site for each of these genes on a linear plasmid and the tandem array of the pseudogene reservoir. In contrast, the *msp*2/*msp*3 system of antigenic diversity differs from *Borrelia* in that *A. marginale* contains no extrachromosomal plasmid, and the pseudogenes are distributed throughout the genome. Perhaps more similar to the antigenic variation of *A. marginale* are the non-tick-transmitted pathogens *Neisseria gonorrhoeae* (13) and *Mycoplasma* (14, 15). These bacteria have pseudogenes distributed throughout the genome, with more than one type of pseudogene for a given full-length gene.

The agent of human granulocytic ehrlichiosis is the closest relative to *A. marginale* in ehrlichial genogroup II. Human granulocytic ehrlichiosis has an ortholog of *msp*2 called p44 or human granulocytic ehrlichiosis *msp*2 (BLASTP value e^{-100}), which is also encoded by a multigene family (16). Several expressed hypervariable regions of p44 were detected by reverse transcription–PCR, cloning, and sequencing (16). Cloning of the genomic counterpart for these expressed hypervariable regions resulted in two genomic clones (p44-15 and 18) that were truncated at both the 5′ and 3′ ends and did not have consensus promoters. Although these two sequences did not start or end in the same positions relative to each other or to a full-length p44, the pattern of pseudogenes distributed throughout the genome appears to be similar to *A. marginale msp*2. Despite the fact that human granulocytic ehrlichiosis infects host species and cell types different from those infected by *A. marginale*, immune evasion is likely achieved through the same method of recombination from widely distributed pseudogenes into a functional expression site.

There are two unique features to the *A. marginale* antigenic variation system described here: the concerted appearance of pseudogenes for two different gene families and the 600 bp of the highly conserved 5′ flanking region. Repeats of this length are highly unlikely to exist in bacteria (17), and that this repeat should be so highly conserved when not encoding a functional gene product indicates that the composition is important. The juxtaposition of this repeat next to the pseudogenes for two gene families that undergo an extremely high rate of recombination is surely important. This arrangement of two pseudogenes in close proximity and the potential for these two gene families to use the same recombinatorial mechanism is a system that allows a greater potential for antigenic variation from a small genome.

We thank Lowell Kappmeyer for excellent technical assistance, Ralph Horn for animal management, and Wendy C. Brown and Fred R. Rurangirwa for helpful discussion. BAC library construction and BAC sequencing were provided by Amplicon Express (Pullman, WA). This work was supported by National Institutes of Health Grants RO1 AI44005 and RO1 AI45580 and U.S. Department of Agriculture Grants USDA-ARS-CRIS 5348-32000-012-00D and USDA-SCA 58-5348-8-044.

1. Zhang, J.-R., Hardham, J. M., Barbour, A. G. & Norris, S. J. (1997) *Cell* **89,** 275–285.
2. Barbour, A. G. (1993) *Trends Microbiol.* **1,** 236–239.
3. Yu, X.-J., McBride, J. W., Zhang, X.-F. & Walker, D. H. (2000) *Gene* **248,** 59–68.
4. Reddy, G. R. & Streck, C. P. (1999) *Mol. Cell. Biol. Res. Commun.* **1,** 167–175.
5. Ohashi, N., Unver, A., Zhi, N. & Rikihisa, Y. (1998) *J. Clin. Microbiol.* **36,** 2671–2680.
6. Sulsona, C. R., Mahan, S. M. & Barbet, A. F. (1999) *Biochem. Biophys. Res. Commun.* **257,** 300–305.
7. Palmer, G. H., Eid, G., Barbet, A. F., McGuire, T. C. & McElwain, T. F. (1994) *Infect. Immun.* **62,** 3808–3816.
8. Alleman, A. R., Palmer, G. H., McGuire, T. C., McElwain, T. F., Perryman, L. E. & Barbet, A. F. (1997) *Infect. Immun.* **65,** 156–163.
9. Alleman, A. R., Kamper, S. M., Viseshakul, N. & Barbet, A. F. (1993) *J. Gen. Microbiol.* **139,** 2439–2444.
10. Barbet, A. F., Lundgren, A. Yi, J., Rurangirwa, F. R. & Palmer, G. H. (2000) *Infect. Immun.* **68,** 6133–6138.
11. Alleman, A. R. & Barbet, A. F. (1996) *J. Clin. Microbiol.* **34,** 270–276.
12. Chapter 2. (1997) in *Genome Analysis: A Laboratory Manual,* eds. Birren, B., Green, E. D., Klapholz, S., Myers, R. M. & Roskams, J. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 136–138.
13. Haas, R. & Meyer, T. F. (1986) *Cell* **44,** 107–115.
14. Kenri, T., Taniguchi, R., Sasaki, Y., Okazaki, N., Narita, M., Izumikawa, K., Umetsu, M. & Sasaki, T. (1999) *Infect. Immun.* **67,** 4557–4562.
15. Noormohammadi, A. H., Markham, P. F., Kanci, A., Whithear, K. G. & Browning, G. F. (2000) *Mol. Microbiol.* **35,** 911–923.
16. Zhi, N., Ohashi, N. & Rikihisa, Y. (1999) *J. Biol. Chem.* **274,** 17828–17836.
17. Rocha, E. P. C., Danchin, A. & Viari, A. (1999) *Res. Microbiol.* **150,** 725–733.

MICROBIOLOGY