Australian Government

Rural Industries Research and Development Corporation

Determination of the genomic sequence of *Mycoplasma gallisepticum*

A report for the Rural Industries Research and Development Corporation

by Dr GF Browning and Dr PF Markham

December 2004

RIRDC Publication No 04/140
RIRDC Project No UM-45J

ISBN 1 74151 048 1
ISSN 1440-6845

Determination of the Genomic Sequence of Mycoplasma gallisepticum
Publication No. 04/140
Project No.UM-45J.

Researcher Contact Details
Dr Glenn Browning
School of Veterinary Science
The University of Melbourne
VIC 3010

Phone: 03 8344 7342
Fax: 03 8344 7374
Email: glenfb@unimelb.edu.au
In submitting this report, the researcher has agreed to RIRDC publishing this material in its edited form.

RIRDC Contact Details
Rural Industries Research and Development Corporation
Level 1, AMA House
42 Macquarie Street
BARTON ACT 2600
PO Box 4776
KINGSTON ACT 2604

Phone: 02 6272 4539
Fax: 02 6272 5877
Email: rirdc@rirdc.gov.au.
Website: http://www.rirdc.gov.au

Published in December 2004
Printed on environmentally friendly paper by Canprint

Foreword

Although disease due to *Mycoplasma gallisepticum* is effectively controlled by the use of strain ts-11 as a vaccine, this cannot be presumed to continue indefinitely, as past experience with Marek's disease vaccines has demonstrated. In addition, recent work in this research group's laboratory has demonstrated that *M. gallisepticum* vaccine strains may be used to deliver antigens to the respiratory tract of the bird. Thus ongoing research on the attenuation of *M. gallisepticum* is likely to yield benefits to the poultry industry in the longer term. One of the most efficient methods to achieve improvements in the current vaccine strains is to use information gained from examining other mycoplasmas and apply that to *M. gallisepticum*. The focus in studies on human mycoplasmas has shifted with the availability of the complete genomic sequence of three species. Investigators are now deleting specific genes from these mycoplasmas and determining whether these deletions attenuate the organism. If the equivalent genes in *M. gallisepticum* can be rapidly identified, then these could be deleted and an assessment made as to whether this attenuation resulted in an effective vaccine strain. Thus, the availability of a full genomic sequence for *M. gallisepticum* would facilitate application of the work on other mycoplasmas directly to *M. gallisepticum*.

Determination of a full bacterial genomic sequence, even the smaller ones such as those of mycoplasmas, is very expensive and generally beyond the scope of all but specialist genomic sequencing laboratories. In 1998 a collaborative arrangement was established between the group at The University of Melbourne and several groups in the USA. This collaboration enabled the genomic sequence to be determined at a reduced cost. The aim of this project was to collaborate with several international laboratories in the determination and analysis of the genomic sequence of *Mycoplasma gallisepticum.*

This project was funded from industry revenue which is matched by funds provided by the Australian Government.

This report, a new addition to RIRDC's diverse range of over 1,000 research publications, forms part of our Chicken Meat Program. The Chicken Meat Program aims to support increased sustainability and profitability in the chicken meat industry by focusing research and development on those areas that will enable the industry to become more efficient and globally competitive and that will assist in the development of good industry and product images. Funding was also provided by the RIRDC Egg Program, now the Australian Egg Corporation Limited.

Most of our publications are available for viewing, downloading or purchasing online through our website:

- downloads at www.rirdc.gov.au/reports/Index.htm
- purchases at www.rirdc.gov.au/eshop

Simon Hearn
Managing Director
Rural Industries Research and Development Corporation

Acknowledgements

This work was supported by USDA grant 58-1940-0-007 to Professor Steven J Geary as well as by RIRDC Project UM-45J.

The collaborative team consisted of Leka Papazisi, Timothy S. Gorton and Steven J. Geary of the Center of Excellence for Vaccine Research and the Department of Pathobiology and Veterinary Science, The University of Connecticut, Storrs, Connecticut; Gerald Kutish of Plum Island Animal Disease Center, U.S. Department of Agriculture, Greenport, New York, Di Kim Nguyen of the Department of Medicine, University of Washington, Seattle, Washington; Greg Mahairas and Steven Swartzell of Regulome Corporation, Seattle, Washington and Anup Madan of The Institute for Systems Biology, Seattle, Washington, USA as well as the authors of this report.

Contents

Executive Summary

Although mycoplasmosis has been controlled with the current generation of vaccines, future control programs may need to focus on improvements in these vaccines so that they are compatible with eradication or so that they are effective in different strains of bird. In the last few years a number of human pathogens have been fully sequenced, including four human mycoplasmas, and parallel information for avian mycoplasmas would enable rapid application of findings on human mycoplasma genes to improving avian mycoplasma vaccines. The University of Melbourne was invited to participate in a project to sequence the *Mycoplasma gallisepticum* genome and subsequently to collaboratively examine the basis of virulence, with the ultimate aim of developing defined third generation vaccines. Involvement in this collaborative exercise has ensured early access to this data, thus enhancing current research programs on mycoplasmosis in Australia. The project presented a unique opportunity to obtain this data at cost, using highly specialised facilities developed in the USA for sequencing large genomes. The strain chosen for sequencing was the standard virulent strain used for challenge trials in the USA.

The genome of the virulent R strain of *M. gallisepticum* was sheared into fragments, which were then randomly cloned and sequenced. The genome was found to be 996,422 base pairs. It was predicted to contain 742 genes encoding proteins and some function could be assigned to 469 of these. A further 150 predicted genes are similar to genes in other bacterial species and 123 are thus far unique to *Mycoplasma gallisepticum*. The *vlh*A gene family, which is thought to play a key role in chronicity of mycoplasmosis in birds, contains 43 genes distributed among five loci containing 8, 2, 9, 12 and 12 genes and comprised 10.4% (103 kb) of the total genome. A total of 80 genes are predicted to encode lipoproteins which are likely to be involved intimately in interactions with the host and a number of other putative virulence factors have been identified. These data lay a strong basis for ongoing studies on the basis of virulence in *M. gallisepticum*. The data are publicly available at http://cevr.uconn.edu and through Genbank.

Introduction

Although disease due to *Mycoplasma gallisepticum* is effectively controlled by the use of strain ts-11 as a vaccine, as experience with Marek's disease vaccines has demonstrated, this cannot be presumed to continue indefinitely. In addition, recent work in this research team's laboratory has demonstrated that *M. gallisepticum* vaccine strains may be used to deliver antigens to the respiratory tract of the bird. Ongoing research on attenuation of *M. gallisepticum* is likely to yield benefits to the poultry industry in the longer term. One of the most efficient methods to achieve improvements in the current vaccine strains is to use information gained from examining other mycoplasmas and apply that to *M. gallisepticum*. The focus in studies on human mycoplasmas has shifted with the availability of the complete genomic sequence of three species. Investigators are now deleting specific genes from these mycoplasmas and determining whether these deletions attenuate the organism. If the equivalent genes in *M. gallisepticum* can be identified, these can be deleted and an assessment made as to whether this attenuation resulted in an effective vaccine strain. Thus the availability of a full genomic sequence for *M. gallisepticum* will facilitate application of the work on other mycoplasmas directly to *M. gallisepticum*.

Determination of a full bacterial genomic sequence, even the smaller ones such as those of mycoplasmas, is generally beyond the resources of all but the most specialist laboratories and is very expensive. The University of Melbourne was invited to participate in a project to sequence the *Mycoplasma gallisepticum* genome and subsequently to collaboratively examine the basis of virulence, with the ultimate aim of developing defined third generation vaccines. Involvement in this exercise has ensured early access to this data, thus enhancing current research programs on mycoplasmosis in Australia. The project presented a unique opportunity to obtain this data at cost, using highly specialised facilities developed in the USA for sequencing large genomes.

Cost of *M. gallisepticum* infections to industry

M. gallisepticum continues to be one of the major bacterial pathogens of poultry worldwide. Vaccination with attenuated live vaccines has reduced its impact on bird health, but even where eradication has been the focus of control programs (for example in the broiler industry the USA) it remains an ongoing cost to poultry production. The experience with eradication campaigns suggests that it will continue to be a disease requiring ongoing control measures, and that the most reliable measure will probably be vaccination of broiler breeder and layer flocks. Although current vaccines appear to have controlled the problem, there is no guarantee they will continue to do so in the future. Furthermore, it is likely that further improvements can be made to the current live vaccine, not only to improve its performance, but also to use it to control other diseases by insertion of genes of their protective antigens into the genome of *M. gallisepticum*. Use of *M. gallisepticum* as a vector would reduce the relative cost of vaccination and may improve control of some other diseases.

Expected benefits of research to industry

The benefits of this research to the industry will not be immediate. The project has derived data that will serve as the basis for future work, which will be directly focussed on development of better vaccines and diagnostic reagents. The benefit to the industry has been primarily to make use of a unique opportunity to prepare for potential changes in control of mycoplasmosis and to provide a foundation for the future improvement of current control measures.

Environmental benefits

The major environmental effect of improved control of *M. gallisepticum* would be a reduction in antimicrobial drug usage, which could possibly lead to reduced development of antimicrobial resistance.

Dissemination and commercialisation of results

The data from this project will be made publicly available over the internet once the primary paper describing it has been published and thus will be able to be used by all people working on mycoplasma pathogenesis. There has been no attempt to patent the sequence as this is of doubtful value given the lack of information about which genes are crucial for development of vaccines at this stage. Commercialisation may flow from future projects aimed at specific genes identified from the data generated in this project.

Review and interactions

When this project commenced the complete genomic sequences of *Mycoplasma genitalium, M. pneumoniae* and *Ureaplasma urealyticum*, all human mycoplasmas, had been determined over a few years (Fraser *et al*., 1995; Glass *et al*., 2000; *Himmelreich et al*., 1996). In addition, a number of other human bacterial pathogens had also been completely sequenced, including *Mycobacterium tuberculosis, Helicobacter pylori, Haemophilus influenzae, Treponema, pallidum, Escherichia coli, Borrelia burgdorferi, Chlamydia trachomatis* and *Rickettsia prowazekii.* A number of other pathogenic bacteria are currently being sequenced, including *Actinobacillus actinomycetemcomitans, Bartonella henselae, Bordetella pertussis, Campylobacter jejuni, Chlamydia pneumoniae, Clostridium difficile, Enterococcus faecalis, Francisella tularensis, Legionella pneumophila, Listeria monocytogenes, Mycobacterium avium, Mycobacterium leprae, Mycoplasma mycoides ss. mycoides SC, Neisseria gonorrhoeae, Neisseria meningitidis, Porphyromonas gingivalis, Pseudomonas aeruginosa, Pseudomonas putida, Salmonella typhimurium, Staphylococcus aureus, Streptococcus pneumoniae, Streptococcus pyogenes, Treponema denticola, Vibrio cholerae and Yersinia pestis.* The genomes of some environmental bacteria had also been sequenced. However, at the time this project was initiated not a single species of specifically veterinary significance (as opposed to zoonotic pathogens) had been sequenced. Even today relatively few important bacterial pathogens of domestic animals have been sequenced. Thus, while the value of complete genomic sequences to attempts to control bacterial disease can be gauged by the resources devoted to the major medical pathogens, it is clear that the expense of fully funding such projects continues to influence the attention paid to veterinary pathogens.

The determination of the complete genomic sequence of *Mycoplasma pneumoniae* has enabled examination of the role of specific genes in the pathogenesis of disease caused by this human pathogen (Hahn *et al*., 1998; Hedreyda and Krause, 1995; Hedreyda *et al*., 1993; Krause *et al*., 1997; Reddy *et al*., 1996). The focus thus far has been on the genes involved in adherence to the respiratory epithelium. This information is likely to be directly applicable to *M. gallisepticum* as the limited information available indicates that homologous genes are present in *M. gallisepticum*. In fact *M. gallisepticum* is one of the most closely related species to *M. pneumoniae*. These genetic studies are based on identification of genes by disruption using one of two methods. The first is the use of random mutagenesis with transposons, particularly Tn4001 or Tn916 (Cao *et al*., 1994; Hedreyda *et al*., 1993; Tigges and Minion, 1994). The second is directed mutagenesis in a specific gene by homologous recombination. Both techniques have been established in the laboratory of the research team undertaking this project and have been applied to *M. gallisepticum* (Markham *et al*., 2003) and are being used to investigate significant genes now the complete genomic sequence has been determined. Furthermore, this research group's well-established model for reproduction of mycoplasmosis by aerosol exposure will enable rapid testing of mutants for the degree of attenuation achieved.

The project has been an international collaboration between The University of Melbourne, University of Connecticut and University of Washington. Two of the laboratories collaborating in the project are the major laboratories investigating the molecular basis of pathogenesis of avian mycoplasma infections and will further collaborate in analysis of genes crucial in pathogenesis in future projects.

The University of Melbourne research team and its collaborators in Connecticut have made significant advances in understanding some aspects of the pathogenesis of *M. gallisepticum* in recent years, particularly in the area of identifying the major adhesins and in determining the mechanisms responsible for high frequency antigenic variation of these adhesins (Glew et al., 1995;Glew et al., 1998; Glew et al., 2000; Gorton and Gleary, 1997; Markham *et al*., 1992; Markham *et al*., 1993; Markham *et al*., 1994; Markham *et al*., 1998; Markham *et al*., 1999). While these studies have enabled some insight to be gained into some processes involved in pathogenesis, they have been focussed on only two genes.

While one approach to identification of virulence genes is to generate mutants by transposon mutagenesis and the availability of the genomic sequence enables rapid identification of the gene disrupted by the transposon, a more directed approach has been adopted in work recently established in the University of Melbourne laboratory (Markham *et al*., 2003). This entails interruption of specific genes by homologous recombination with a cloned version of the gene into which the tetracycline resistance gene, *TetM*, has been inserted; the resistance to tetracycline enabling selective growth of the rare recombinants that contain the disrupted target gene. The availability of complete genomic sequences for *M. gallisepticum* will enable identification of all the homologues of putative virulence genes in *M. gallisepticum* and enable the specific disruption of these genes so that their function can be studied.

Objectives

The objectives of this project were to:

- determine the complete genomic sequence of *Mycoplasma gallisepticum;*
- facilitate identification of genes which are likely to play a role in virulence; and
- lay a foundation for subsequent studies to improve the performance of mycoplasma vaccines and to improve diagnosis of mycoplasmosis.

Methodology

The R strain of *M. gallisepticum* was chosen as the strain to be sequenced because it has been widely used as a reference virulent strain in the USA and in some other countries. The genome was randomly sheared into fragments averaging 3 to 5 kilobases in length. These fragments were then cloned in the plasmid vector pBlueScript SKII (+) and sequenced.

Plasmid preparations were processed automatically using a robotic workstation and sequenced at both ends. Sequence data were then assembled and analysed. Additional reactions were performed to close gaps in the assembly by designing primers near the gaps and using these to sequence plasmid clones that spanned the gaps. Where no clone was available to span the gap, PCR products were generated using custom primers on genomic DNA and the subsequent PCR products were sequenced and added to the assembly. Sequencing coverage, of both strands, was 11x (or 22 reads per kilobase), done to an error rate less than 1 per 10 kb.

Coding sequences were defined as those that would encode an open reading frame of at least 33 amino acids and did not overlap more than 30 nucleotides with neighbouring coding sequences. The amino acid sequences encoded by these putative coding sequences were compared to sequences in the NCBI

non-redundant protein databases, the conserved orthologous groups database, and proteins from the mycoplasma genomes sequenced to date. A coding sequence was assigned a designation based on homology when its translated product produced a BLAST score of at least 100 in the COG database and 150 in the non-redundant database. Scores of 300 or more were considered definitive.

The sequence data have been submitted to the NCBI database under accession number AE015450. Future updates will be available through the Center of Excellence for Vaccine Research website (http://cevr.uconn.edu/).

Detailed Results

The genome is composed of 996,422 bp with an overall GC content of 31%. It was predicted to contain 742 genes encoding proteins and some function could be assigned to 469 of these. A further 150 predicted genes are similar to genes in other bacterial species and 123 are thus far unique to *Mycoplasma gallisepticum*.

Perhaps the best described example of a multi-gene family in mycoplasmas is that encoding the VlhA or pMGA lipoproteins (Baseggio *et al*., 1996; Liu *et al*., 1998; *Markham et al*., 1993). It has been well established that *M. gallisepticum* generally expresses a single member of the family at any one time (Glew *et al*., 1995) and that the specific gene expressed can be influenced by growth in the presence of cognate antibody (Markham *et al*., 1998). The probable role of this family of genes in generating antigenic variation has been demonstrated in infected chickens, with both phase variation demonstrated during the acute stages of disease, and antigenic switching during the chronic stages (Glew *et al*., 2000). These findings have led to the suggestion that the principal function of this family of genes is to generate antigenic diversity and hence facilitate immune evasion during chronic infections. These genes formed the largest paralogous gene family in the *M. gallisepticum* genome.

Based on the analysis of the genomic sequence determined in this project we now know that in strain R this family contains 43 genes, comprising a total of 103 kb or 10.4% of the genome. The 43 *vlh*A genes are distributed among five loci containing 8, 2, 9, 12 and 12 genes. Of the 43 genes, 38 possess the signature *vlh*A gene features, which include a GAA repeat motif 5' of a GTG start codon and conserved regions flanking the start codon (Markham *et al*., 1994). The sequence identity among these 38 genes ranges between 41% and 99%. Five of the genes have sequence similarity to *vlh*A but lack the *vlh*A signature motifs. The genes within each of the five loci are in the same transcriptional orientation, with only one exception, which also lacks the *vlh*A signature motifs and is adjacent to a putative transposase, suggesting that rearrangements by a transposon may account for the differences observed in this region. Five of the genes have been interrupted by mutations that introduce shifts in the reading frames.

Transcription of *vlh*A genes has been shown to correlate with the occurrence of twelve repeats within the GAA motif (Glew *et al*., 1998; Liu *et al*., 2000). The size of the GAA motifs ranges between 2 and 27 repeats, with a median of 9 repeats. Of the 38 *vlh*A genes possessing GAA motifs, only one possesses the 12 repeats reportedly required for transcription. Interestingly, of all the *vlh*A genes sequenced from different *M. gallisepticum* strains, four of the five that are preceded by 12 GAA repeats are predicted to encode proteins that share very high levels of amino acid sequence identity (95-98%). That different strains express near identical genes from the large repertoire of *vlh*A genes when cultured *in vitro* suggests that the product of these very closely related genes may have a specific function that is required during growth *in vitro*.

Other significant virulence associated genes identified included additional homologues of the GapA and CrmA cytadherence molecules. Approximately 10% of the genes encoded putative lipoproteins,

which are generally exposed on the surface of mycoplasmas, and 20% encoded putative membrane proteins.

Discussion of Results

Mycoplasma gallisepticum strain R has a genome of 996 kb genome with 742 putative protein encoding genes. Approximately one-third of the genes cannot be assigned a function and 17% of the genes appear to be unique to *M. gallisepticum*.

The vlhA family, totaling 43 genes, comprises the largest paralogous set of genes in *M. gallisepticum*. A large percentage of the *M. gallisepticum* genome is devoted to membrane-associated molecules, which are likely to be involved intimately with the host during infection.

Implications

The completion of the sequence has facilitated a number of projects directed at identifying the most important virulence genes in *M. gallisepticum*, paving the way for developing improved vaccines in the future.

Dissemination and Adoption

The sequence data have been submitted to the NCBI database under accession number AE015450. Future updates to the *M. gallisepticum* genome will be available through the Center of Excellence for Vaccine Research website (http://cevr.uconn.edu/).

Publication

The complete genome sequence of the avian pathogen *Mycoplasma gallisepticum* strain R_{low}. Leka Papazisi, Timothy S. Gorton, Gerald Kutish, Philip F. Markham, Glenn F. Browning, Di Kim Nguyen, Steven Swartzell, Anup Madan, Greg Mahairas, and Steven J. Geary. *Microbiology*, in press.

Recommendations

Future projects identifying the role of the numerous putative virulence genes are needed to capitalise on the data generated in this project. Some of these studies have already commenced and are expected to identify some of the genes involved in infection.

Intellectual Property

The data from this project will be made publicly available. Intellectual property may be developed in future studies that determine the role of specific genes in virulence.

Communications Strategy

The completed genome sequence has been submitted to the NCBI database under accession number AE015450. This is a publicly available sequence database, accessible without charge. The sequence is also available at the University of Connecticut Website at http://cevr.uconn.edu/. The availability of the data through these routes has already been disseminated to the avian mycoplasma research community through the email list of the International Research Program in Collaborative Mycoplasmology.

Bibliography

EN.REFLIST