Corynebacterium pseudotuberculosis genome sequencing: Final Report

Summary

To provide an invaluable resource to assist in the development of diagnostics and vaccines against caseous lymphadenitis (CLA), the sequencing of the genome of a virulent, United Kingdom *Corynebacterium pseudotuberculosis* strain has been made possible through funding by MLC. Sequencing provided coverage of approximately 98% of the entire genome, contained within eight large contiguous segments. Subsequently, bioinformatic techniques were applied to the identification of gene coding regions within the genome, and where possible the proteins encoded by these putative coding regions were identified by comparison with other proteins resident in publicly accessible sequence databases. The overall structure of the genome was found to be similar to other sequenced corynebacterial genomes, especially that of the human pathogen *C. diphtheriae*. However, numerous differences between *C. pseudotuberculosis* and *C. diphtheriae* were identified which may be due to their adaptation to different host species. Preliminary insights suggest the evolution of these two pathogens from a common ancestor following the identification of numerous distinct bacteriophage insertions, each carrying with them genes encoding proteins involved in host colonisation and persistence. Significantly, numerous genes which appear to be unique to *C. pseudotuberculosis* have been identified, priming future work which will focus on these genes in order to determine their suitability for use as vaccine and/or diagnostic reagents.

Objectives

The *Corynebacterium pseudotuberculosis* genome sequencing project has been conducted as outlined in the original proposal to the MLC. The strain of *C. pseudotuberculosis* sequenced (designated strain 3/99-5) is a virulent isolate from a case of naturally-occurring CLA in a sheep in the Scottish Borders in 1999 (Connor *et al*. 2000. J. Clin. Microbiol. **38:**2633); this strain has been the focus of a significant body of research at Moredun, including efforts directed at vaccine and diagnostic test development (Fontaine *et al*. 2006. Vaccine. **14:**33). Its close relationship with other *C. pseudotuberculosis* isolates (including those from disease cases worldwide) has been well-defined in other research at MRI (Connor *et al*. 2000. J. Clin. Microbiol. **38:**2633; Connor *et al*. 2007. Vet. Res. **38:**613) and is therefore representative of the major epidemiological type.

Following a thorough analysis of the company profiles of the commercial sequence providers available at the time this project was initiated, the company chosen to conduct the genome sequencing was 454 Life Sciences™ (Branford, CT, USA), who offered a relatively low-cost, high-quality sequencing service. 454 are responsible for the development of a revolutionary technology, producing millions of raw bases of nucleotide sequence per hour using a GS FLX genome sequencing instrument. 454 has also developed system specific software enabling mapping or *de novo* assembly for whole genome shotgun sequencing of genomes up to 50 megabases. Many biologically meaningful and complex regions of genomes can be analyzed with this system without the time or cost constraints of current DNA sequencing methods. Through this technology 454 Life Sciences provides an enabling solution for ultra-high-throughput DNA sequencing. The quote obtained from 454 was for provision of assembled sequence derived from 30× coverage of a 2.5 mega base-pair (mb) genome (the higher the coverage, the better quality the resulting sequence data). In addition, paired-end sequencing runs were performed to reduce the number of gaps in the resulting genome sequence.

The bacterial genome relates to the total complement of DNA sequences carried in each individual cell, which comprises the chromosome, in addition to any other non-chromosomal, self-replicating elements, such as plasmids. *C. pseudotuberculosis* has not been reported to contain plasmid(s); therefore, in this instance "genomic" refers to the chromosome only. Bacterial chromosomes are circular elements of varying size, which in the case of *C. pseudotuberculosis* was estimated to be approximately 2.5 mb in size. Because of the large size and fragile nature of bacterial chromosomes, inevitable damage occurs during the purification process which results in shearing (breaking up) of the circular molecule; the extent of this damage depends greatly on the purification process employed. However, in order to allow the generation of high-quality sequence data, it was essential that 454 be provided with high-quality *C. pseudotuberculosis* genomic DNA. *C. pseudotuberculosis* is a Gram-positive organism belonging to the actinomycete family, which means that it has a characteristically tough cell-wall, to which is attached numerous mycolic acids (long-chain fatty-acid molecules). As a result, the cells are extremely difficult to break open to release genomic DNA, and the mycolic acids complicate downstream DNA purification. Unfortunately, relatively harsh treatment is required to lyse *C. pseudotuberculosis* cells, which inevitably results in damage to the DNA. Therefore, significant time and effort was afforded to the optimisation of DNA preparation methodology. Following discussion with staff at 454, it was ascertained that fragmented DNA in the size region below approx. 1 kilo base (kb) was especially detrimental to the sequencing process. Therefore, fractionation of DNA was conducted by electrophoresis through agarose gels, followed by excision of high molecular weight DNA and subsequent

purification (Fig. 1). In this way, a suitable amount of DNA of the required quality was obtained for provision to 454.

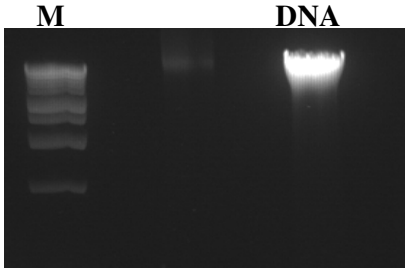

Figure 1. *C. pseudotuberculosis* genomic DNA electrophresed through a 2% (w/v) agarose gel and stained with ethidium bromide (M=DNA molecular weight marker).

Approximately one month following submission of genomic DNA, 454 completed the sequencing of the genome and made available the data. Through the use of paired-end sequencing it had been possible for 454 to assemble the generated sequence data into 8 large contiguous regions (contigs.), of approximately 2.3 mb in size, covering in excess of 98% of the predicted complete genome. Sequence information was downloaded from the 454 server, and subjected to bioinformatic analyses at MRI. Initially, it was necessary to attempt to assemble the 8 contigs. into the order that they would fall within an intact genome. To faciltate this, other published corynebacterial genomes were used as a scaffold against which the *C. pseudotuberculosis* contigs. were mapped. Genome sequences are available in the public database for *C. glutamicum*, *C. efficiens*, *C. diphtheriae* and towards the end of this project *C. jeikeium*. Both *C. glutamicum* and *C. efficiens* are highly-related, non-pathogenic organisms, while *C. jeikeium* is an opportunistic pathogen. *C. diptheriae* was the only obligate pathogen for which a genome existed, and previous studies have suggested that *C. diphtheriae* and *C. pseudotuberculosis* (and *C. ulcerans* for which no genome sequence exists) share a relatively high level of overall DNA homology. Therefore, it was decided that initial attempts to scaffold the *C. pseudotuberculosis* contigs. would be based upon alignment with the *C. diphtheriae* genome. Homology between the *C. diphtheriae* genome and the ends of each of the 8 *C. pseudotuberculosis* contigs. was used to map each contig. onto the *C. diphtheriae* genome, and in doing so the contigs. were arranged into the correct order. Subsequently, the contigs. were concatenated into a single, linear sequence by pasting together the sequence data. Each contig. was separated from its neighbours on either side by inclusion of 50×"N" to make it obvious where a gap existed in the sequence data (Fig. 2). Interestingly, one *C. pseudotuberculosis* contig. mapped to two separate locations in the *C. diphtheriae* genome. Both regions contain identical 16s rRNA genes (encoding a portion of the prokaryotic ribosome), and this duplicate region is conserved among other sequenced corynebacteria. Therefore, it is likely that the same is also true of *C. pseudotuberculosis*; however, because of the similarity in sequence between the two regions, during 454's assembly of the sequencing reads, sequences pertaining to both regions were unavoidably assembled into a single contig.

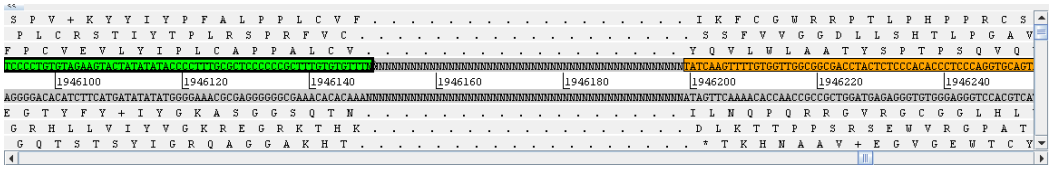

Figure 2. Example of 50×N spacer inserted between *C. pseudotuberculosis* contigs. to allow assembly into a linear, incomplete genome sequence.

Having assembled the *C. pseudotuberculosis* genome, it was then necessary to predict protein-encoding sequences within the sequence. Automated software, known as "Glimmer" was used to predict open reading frames (ORFs), which are regions of DNA putatively corresponding to gene sequences (Fig. 3). The stringency of the ORF prediction was set so that extremely small ORFs (unlikely to encode real protein sequences) were excluded. Nevertheless, it is noted that ORF prediction in this way, although the generally accepted method, does not always identify the entire gene complement within a genome sequence. In this respect, alternative methods of genome annotation now exist which rely on the application of sensitive proteomic techniques; however, these approaches are expensive and were not possible within the limits of the funding provided.

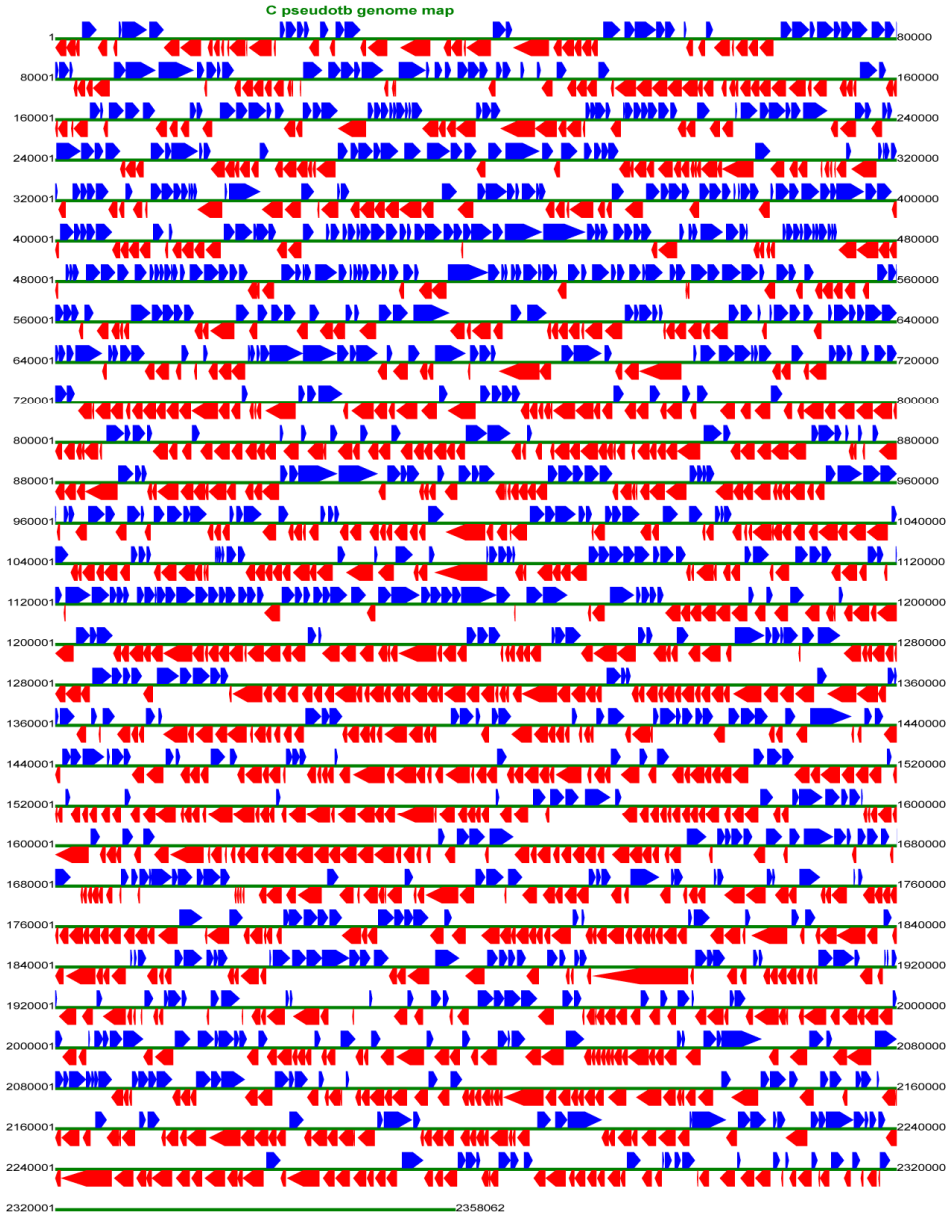

Figure 3. Schematic representation of the *C. pseudotuberculosis* genome (linear). Arrows denote predicted open reading frames (ORFs); blue arrows represent ORFs encoded by the leading DNA strand, while red arrows indicate ORFs encoded by the complementary DNA strand.

Having identified putative ORFs, the Basic Local Alignment Search Tool (BLAST) was employed to find regions of local similarity between sequences. The software program compares nucleotide or protein sequences to those sequence compiled in public databases (NCBI, SwissProt and TrEMBL), and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. Based on the BLAST results, a preliminary annotation of the *C. pseudotuberculosis* genome was carried out, and subsequently imported into the Artemis software package, a freely available genome viewer and annotation tool that allows visualization of sequence features and the results of analyses within the context of sequences and their six-frame translations.

At this time, it has not been possible to exploit the *C. pseudotuberculosis* genome to its full capacity. However, preliminary analyses have revealed that the annotated proteins fall into three classes, namely those which are identical to other sequenced corynebacteria, those which manifest some divergence to the equivalent proteins of other sequenced corynebacteria, and those which are not found in other sequenced corynebacteria. This observation is entirely expected, and relates to the conservation of a "core" genome (encoding predominantly biosynthetic proteins involved in metabolism, cell genesis, cell division, *etc.*) with variation evident among other proteins, reflecting increasing evolutionary distance between corynebacteria, and adaptation to survival within different environments (*e.g.* the small ruminant host in the case of *C. pseudotuberculosis*). To investigate the overall similarity between *C. pseudotuberculosis* and *C. diphtheriae*, annotated genome files were compared using the Artemis Comparison Tool (ACT) a DNA sequence comparison viewer, based on Artemis. Interestingly, it was evident that both organisms were very similar to each other with respect to the order of genes within the genome (Fig. 4). However, it was apparent that horizontal gene transfer (*i.e.* the integration of foreign DNA sequences, such as from bacteriophages) into either bacterial chromosome was responsible for the gain/loss of certain characteristics which has contributed to the divergence of the two bacterial species and the evolution of their ability to cause different types of infection in different hosts (*e.g.* Fig. 5). *C. diphtheriae* especially has been subject to numerous bacteriophage insertions, bringing with them the genes encoding novel cell-surface proteins involved in host colonisation, and novel iron uptake mechanisms.

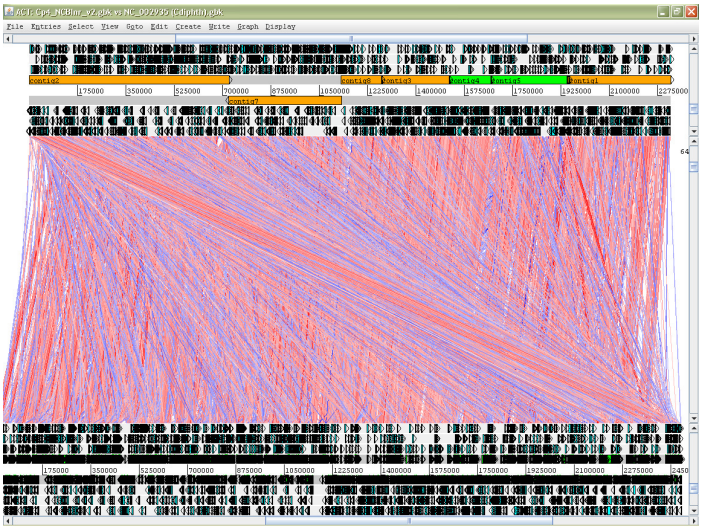

Figure 4. Alignment between *C. pseudotuberculosis* (top) and *C. diphtheriae* (bottom) genomes in ACT. Homology between the two genomes is indicated by the presence of lines between homologous ORFs. The absence of any gaps in the lines of homology indicates that both genomes are highly similar overall.

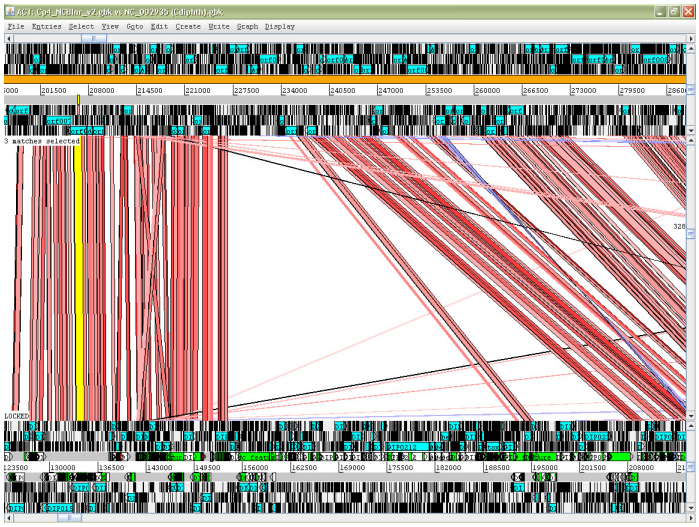

Figure 5. Zoomed-in view of a region in Fig. 4. The presence of a large gap in the lines of homology indicates ORFs that are present in *C. diphtheriae* (bottom) but not in *C. pseudotuberculosis* (top). In this instance the unique region corresponds to a phage-encoded island, having been acquired by horizontal gene transfer through infection and integration of a lysogenic bacteriophage.

The next logical stage in the progression of the *C. pseudotuberculosis* genome project will be to fill in missing sequences, to allow the closure of the gaps between each of the 8 contigs. This objective may be achieved in two ways, either through the use of PCR (where specific oligonucleotide primers are designed to anneal at the ends of adjoining contigs) or by taking advantage of new high-throughput sequencing technology (*e.g.* Solexa sequencing). Significantly, based upon homology with the *C. diphtheriae* genome, it is not anticipated that the gaps are particularly large, and can therefore be amplified by conventional PCR. However, to ensure that the predicted 16S rRNA-encoding regions are amplified and sequenced in their entirety, it may be necessary to employ long-range PCR.

Currently, it has been possible to integrate the availability of the genome with work being conducted by a PhD student at Moredun. The student's work has been focusing on the regulation of gene expression in *C. pseudotuberculosis*, and involves the application of proteomic techniques for the identification of genes differentially expressed under different environmental conditions. With the availability of the genome sequence, it has been possible to evaluate the proteomic data at the genome level, and in several instances conserved motifs have been identified upstream of differentially-expressed genes, which likely serve as binding sites for regulatory protein(s).

Unfortunately, the currently funded PhD student is due to complete their thesis work before the end of 2008 and hence there is no capacity for them to exploit the genome sequence to its full extent. Therefore, the availability of the genome forms the basis of a PhD proposal submitted to MLC in response to a studentship call earlier this year. Further funding would allow commitment of staff to a full investigation of the genome. In addition, further work would allow cataloguing of those proteins which are novel in *C. pseudotuberculosis* (compared to other bacteria including corynebacteria), and it is our belief that these proteins will be particularly suited towards the development of vaccines and diagnostic reagents; however, it will first be necessary to determine the conservation of these novel proteins among other *C. pseudotuberculosis* isolates. It is our intention to prepare a manuscript detailing the genome sequence following completion of gap closure and further annotation. Subsequently, the availability of the sequence will also contribute to further publications. There may also be IP potential through protection of novel *C. pseudotuberculosis* antigens.