

ORIGINAL ARTICLE

Annotating the Mycobacterium avium Genome: A Project in Bioinformatics

David D. Shersher[1†], Maksim Kirtsman[1†], Mikael Katz-Lavigne[1†], Marcel A. Behr, MD MSc[2], Makeda Semret, MD MSc[2†*]

ABSTRACT Bioinformatic tools facilitate efficient processing and formatting of experimental data and are becoming essential to research in the biological sciences. Whole genome sequencing projects, combined with DNA microarray technology, have allowed genomic comparisons between and within species of microorganisms. The genome of *Mycobacterium avium subsp. avium* (MAA) has been sequenced by The Institute for Genomic Research (TIGR), but a final and annotated version has not yet been made available. The goal of this project was to annotate the sequence of MAA as a foundation for microarray-based genomic comparisons. We used software to identify and predict open reading frames (ORFs) present in this organism. The ORFs were then compared to those catalogued in two large, online genetic databases for other microorganisms and matched to homologous sequences, allowing the determination of putative functions for each predicted gene. The genome of MAA was determined to contain 4480 genes, the majority of which are homologous to genes found in other Mycobacterial species.

Keywords: *Mycobacterium avium* complex, annotation, genomics, BLAST

INTRODUCTION

The combined availabilities of whole genome sequence information and new bioinformatic tools have fueled genomic research, allowing extensive comparisons between closely related species of microorganisms. Sequencing of the *Mycobacterium tuberculosis* genome in 1998 (1) laid the foundation for detailed study of the members of the *Mycobacterium tuberculosis* complex, which in turn has guided the establishment of evolutionary scenarios for this group of microorganisms as well as permitted inferences about their biology (2-4).

Since then, over 140 bacterial genomes, including several members of the M. tuberculosis complex, have been sequenced and genomic information has been made publicly available on the internet. In contrast, little genomic information is available for the *Mycobacterium avium* complex (MAC), a group of environmental organisms with tremendous pathogenic potential. The MAC is composed of two species, M. avium and M. intracellulare. The former is further subdivided into three subspecies, M. avium subsp. avium (MAA), M. avium subsp. paratuberculosis (MAP) and M. avium subsp. silvaticum (MAS). Defined as a complex based on their genetic similarities, these subspecies are phenotypically quite distinct. MAP is an obligate intracellular pathogen that is the cause of a chronic enteritis known as Johne's disease in cattle; it has also been suggested by some to be the cause of an inflammatory bowel disease in humans called Crohn's disease (5). MAA is predominantly an environmental organism which

*To whom correspondence should be addressed: Makeda Semret MD MSc FRCP(C), McGill University Health Centre. Research Institute RS1-105, 1650 Cedar Avenue , Montreal Quebec H3A 1G4 Canada, tel: (514) 934 1934 x 44621; fax: (514) 943 8261; e-mail:makeda.semret@mail.mcgill.ca .

1 Department of Microbiology and Immunology, Faculty of Medicine, McGill University, Montreal, Quebec, Canada.

2 Department of Infectious Diseases and Medical Microbiology, McGill University Health Centre

† These authors contributed equally to the manuscript.

causes a tuberculosis-like disease in birds, and which can also elicit a disseminated disease in immuno-compromised individuals. *M. avium subsp. silvaticum* also affects birds, although little is known about its true distribution. Moreover, members of the MAC present widely different laboratory characteristics. Their divergent patterns of disease manifestation and laboratory phenotypes are poorly understood but are likely to result from subtle variations within the organisms' respective genomes.

Genome sequencing projects for two of the members of the MAC (MAA and MAP) have been undertaken. Sequence information on MAA strain 104, a clinical isolate from an AIDS patient, has been released on the internet (TIGR, http://www.tigr.org), but a fully edited and annotated version of the genome is not yet available. The goal of this project was to annotate the complete genome sequence of MAA 104 in order to identify putative genes of this organism that will in turn serve as a template for DNA-microarray based genomic studies of the MAC. We describe herein the annotation of the 5.5 Mbp genome of MAA, which was accomplished through the use of various automation scripts.

METHODS

Identification of open reading frames

The 5.5 Mbp genomic sequence of MAA strain 104 was obtained from The Institute for Genomic Research (TIGR, http://www.tigr.org). Annotation was performed using Artemis™, a genomic visualization software package developed by The Sanger Institute (http://www.sanger.ac.uk/software). This tool allows the identification of open reading frames (ORFs) based on recognition of start and stop codons, and the in-silico translation of these nucleotide sequences into their respective amino acids. The minimum size of the ORFs identified was set at 100 bp. The Artemis™ version used (version 4) had limitations with respect to the size of the sequence that could be viewed at one time. For this reason, the 5.5 Mbp genome was split into 61 contiguous sequences (contigs) of 92,000 bp each. The contigs contained 500 bp overlaps at both ends in order to facilitate their later merging as well as to ensure that open reading frames spanning two adjacent contigs are appropriately identified.

The nucleotide sequences of each of the contigs was entered into Artemis™; once putative ORFs were assigned, the amino acid sequence of each ORF in the 61 contigs was exported into text files to permit alignment with sequences already deposited in public internet databases.

Alignment searches for predicted open reading frames

Alignment of each ORF sequence to existing databases was done using the Basic Local Alignment Search Tool (BLAST) (6). The BLAST algorithm allows comparisons of nucleotide or amino acid sequences for homology to previously annotated sequences or proteins that are deposited in public databases. For each predicted ORF within the MAA genome, Artemis™ provides an amino acid sequence that is then compared (using the blastp protocol) against amino acid sequences from other genomes. If those sequences had previously been assigned a probable gene function, by extension, we assigned the function of the most similar amino acid sequence to the corresponding MAA ORF.

The two major databases used were the *Mycobacterium tuberculosis* genome database (Release 5) from the Pasteur Institute (http://genolist.pasteur.fr/TubercuList/), and the Non-Redundant (NR) database from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/). Initially, the databases were accessed through an online BLASTing program called 'NetBlast'. This program takes local query sequences contained in text files and aligns them individually against the databases on their respective servers. The BLAST queries are queued based on the volume of searches occurring simultaneously, which causes delays proportional to the total number of queries. The solution to the query-delay problem was to use a local version of the BLASTing software called 'blastall' (ftp://ftp.ncbi.nih.gov/blast/) and to download and locally store the two databases.

Query batching

In order to query all of our MAA genome files in an easy, automated manner, text-based batch files were created. Both these files and the 'blastall' program are MSDOS™-based. The batch files were simple and contained the program execution command specifying the name and certain details of each contig file to be processed. Each BLAST result is reported with an Expectation (E) value, representing the number of hits with equal or better alignment scores than what one can expect to find by chance alone when searching a database of a certain size. Because the NR database is much larger than R5, we made the simplifying assumption that an E value of 0.01 in the former is more or less equivalent to an E value of 0.1 in the latter, and arbitrarily assigned these E values as our initial cutoffs.

Creating Microsoft Excel™ databases

In order to extract the useful information from the BLAST results and display it in a user-friendly format,

a Microsoft Windows™-based automation software was used. We custom-wrote a script to locate the desired information within the BLAST result file and transfer it to a Microsoft Excel™ spreadsheet. Furthermore, the E values obtained from homology searches against both databases were compared, and only the best hit (with the lowest E value) was imported into the Microsoft Excel™ spreadsheet. When Expectation values were similar, precedence was given to the TubercuList database as there is significant homology between M. avium and M. tuberculosis. If a putative ORF in MAA did not show a significant degree of homology to proteins in the M. tuberculosis database, the results from NR, if any, were retained. The script evolved numerous times during the process to improve on accuracy and efficiency. The fifth revision of the script had a proofreading function that could detect errors in the data transfers and correct them.

For each contig, the Microsoft Excel™ file generated was "cleaned" using another automation script that removed the empty ORFs with no hits in either of the two databases. For further processing, the Microsoft Excel™ files had to be reformatted into GFF ("Gene-Finding Format") files to permit future visualization in Artemis™.

Removal of overlapping ORFs in Artemis™

When the "cleaned" GFF files were imported into Artemis™, a graphical overview of the nucleotide and amino acid sequence with ORFs marked in the 6 possible reading frames was produced. The entire genome thus represented was visually scanned, specifically searching for overlapping ORFs. Once an overlap was detected, the longer ORF with the lower E value or the ORF with significant homology to a protein in R5, was retained while the other was manually removed. For example, if an ORF annotated as Equine Herpes Virus Protein with an E value of 0.01 overlapped another ORF whose best hit was with a M. tuberculosis gene that had an E value of 0.1, the *M. tuberculosis* homologue was retained.

Merging the contigs

The next step in the annotation was the merging of the GFF files based on the intentional overlaps left when the MAA genome was first split into contigs. This was done in Microsoft Excel™ by opening the manually edited GFF files and checking for identical ORFs found in the extremes of two adjacent contigs. The whole genome was thus reconstructed from the overlapping contigs.

Functional classification of the annotated genes

All putative gene products were subdivided into 10 functional categories of proteins based on the classification developed by the Pasteur Institute for the genomes of *Mycobacterium tuberculosis* and *Mycobacterium leprae*. The functional classification data was imported into Artemis™ by editing a GenBank (GBK) file of the annotation. A color-coding scheme was then used to separate the ORFs based on these different functional gene categories.

RESULTS

Annotation of the 5.5 Mbp chromosomal sequence of MAA strain 104 took several months. The initial 61 contigs were processed through the Artemis™ software package, and a total of 29,000 possible ORFs were identified. The contigs were subsequently batch-BLASTed, which generated two sets of files with the BLAST results: one set that showed homology results against the NCBI database and the other against the M. tuberculosis database. By applying the cut-offs for E values described above, we were able to decrease the number of ORFs to about 10,000.

After the final steps of removing overlapping ORFs and merging the 61 contigs, annotation was complete. A total of 4,480 putative ORFs were identified; 4,095 of these ORFs (91%) matched genes in the M. tuberculosis database (R5), while the rest matched to other organisms. Figure 1 shows an example of the

Functional Category*	ORFs in MAA 104 (% of total)
0 virulence, detoxification, adaptation	148 (3.3%)
1 lipid metabolism	436 (9.7%)
2 information pathways	222 (4.9%)
3 cell wall and cell processes	662 (14.8%)
5 insertion sequences and phages	155 (3.5%)
6 PE/PPE	53 (1.2%)
7 intermediary metabolism and respiration	1139 (25.4%)
8 unknown	93 (2.1%)
9 regulatory proteins	265 (5.9%)
10 conserved hypotheticals	862 (19.2%)
Unclassified (MTB orthologs)	60 (1.3%)
Unclassified (non MTB orthologs)	385 (8.6%)
Total # of genes annotated	4480

*functional categories are the same as those for the classification of *M. tuberculosis* (http://genolist.pasteur.fr/TubercuList/)

Table 1. Summary of the annotated M. avium subsp. avium genome, with functional classification of the predicted genes as per the classification for M. tuberculosis (http://genolist.pasteur.fr/TubercuList/). ORF = open reading frame, MAA = M. avium subsp. avium, MTB = M. tuberculosis. PE/PPE refers to a family of genes commonly found in the M. tuberculosis genome defined by the amino acid sequence motifs Pro-Glu (PE) and Pro-Pro-Glu (PPE). These motifs are located at the highly conserved N-terminal domains

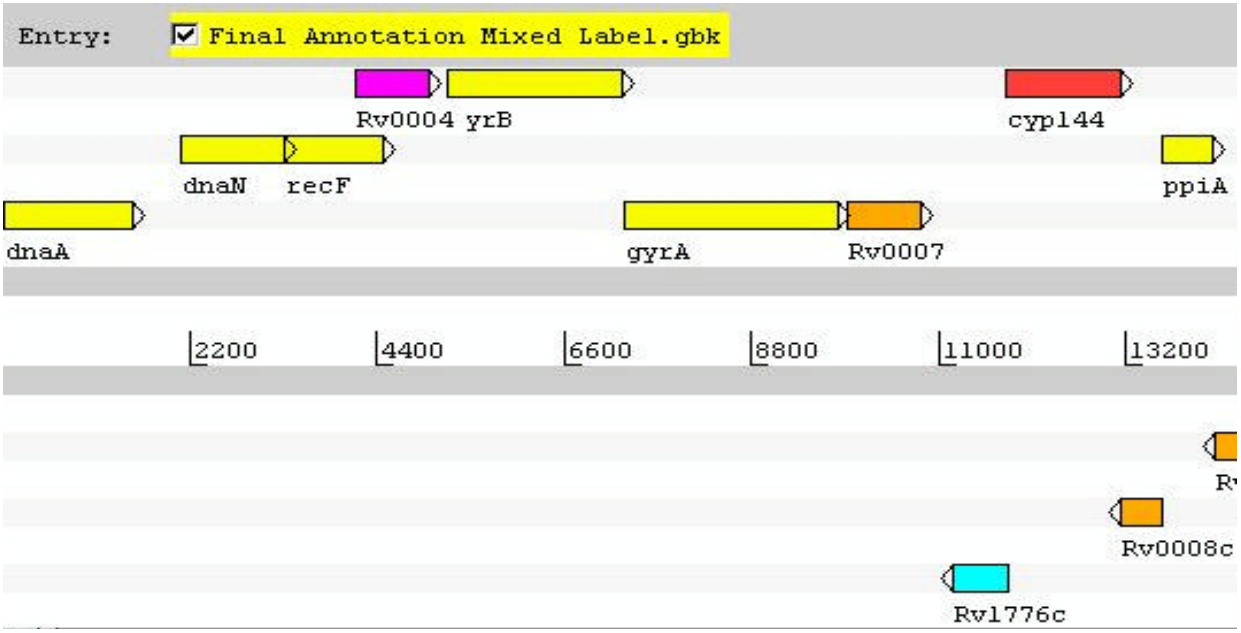

Figure 1. Screenshot of the final visual annotation as seen through Artemis. Colored ORFs represent genes in M. avium subsp. avium and are color-coded based on functional categories of genetic function. As there are 6 ways in which DNA sequence can encode amino acids (three frames forward and three frames in reverse), these colored arrows representing putative genes are found in any of these 6 reading frames. The direction of the colored arrow representing the ORF indicates the coding sense.

appearance of the ORFs in Artemis™. In this final annotation, there are few overlapping regions between the 4480 putative genes. Each of the ORFs is also color-coded based on its functional classification as defined by TubercuList (http://genolist.pasteur.fr/TubercuList /help/classif-search.html#Codes). Table 1 summarizes the number of ORFs per functional category and their percentage within the genome.

DISCUSSION

The *Mycobacterium avium subsp. avium* annotation project evolved from a need to cross-reference a previous annotation, which was performed in-house in the year 2000 but was based on a fragmented (non-sequential contigs) genomic sequence of MAA strain 104. This first annotation identified 5,574 genes that served as a template for the design of oligonucleotides used in the assembly of a DNA microarray. The array has since been used to perform extensive genomic comparisons between the members of the MAC and has been validated for transcriptome analysis for this species. However, the fragmented sequence released earlier most likely contained redundancy and errors. The release of an edited and circularized version of the MAA 104 sequence in June 2002 provided the opportunity to annotate the genome presenting the genes in their correct order, starting with dnaA in the origin of replication and ending with the gene rpmH.

The 5.5 Mbp genome of MAA 104 thus annotated contains 4,480 ORFs, for a gene density of 1,222 bp/gene. In contrast, the 4.4 Mbp genome of M. tuberculosis has 3,924 ORFs for a gene density of 1,114 bp/gene (1). Considering that there is a significant degree of genetic homology between Mycobacterial species, our annotation of the MAA genome may have slightly underestimated the total number of ORFs. Previous work has established that there is 97% homology between MAA and MAP over several large regions (7,8). In order to gauge the accuracy of our annotation, we compared one region of our annotated genome (the region surrounding the origin of replication, oriC) to the corresponding region in the MAP (8) and the M. tuberculosis genomes. We note over the oriC region that the number of genes and their order were similar among these three genomes (Figure 2), providing reassuring validation to our annotation.

The annotation process led to the identification of beneficial informatic tools. For instance, the version of the Artemis™ software used in this project (version 4) had inherent limitations in the size of the sequence that could be viewed at once. This was resolved by artificially splitting the genome into 61 files. The most recent release of Artemis™ (version 5) is able to display the entire 5.5Mbp genome without division into multiple contigs. As another example, some of the repetitive steps involved in annotation can be expedited by custom-written automation scripts, and/or with the use of Visual Basic™ macros for performing some of

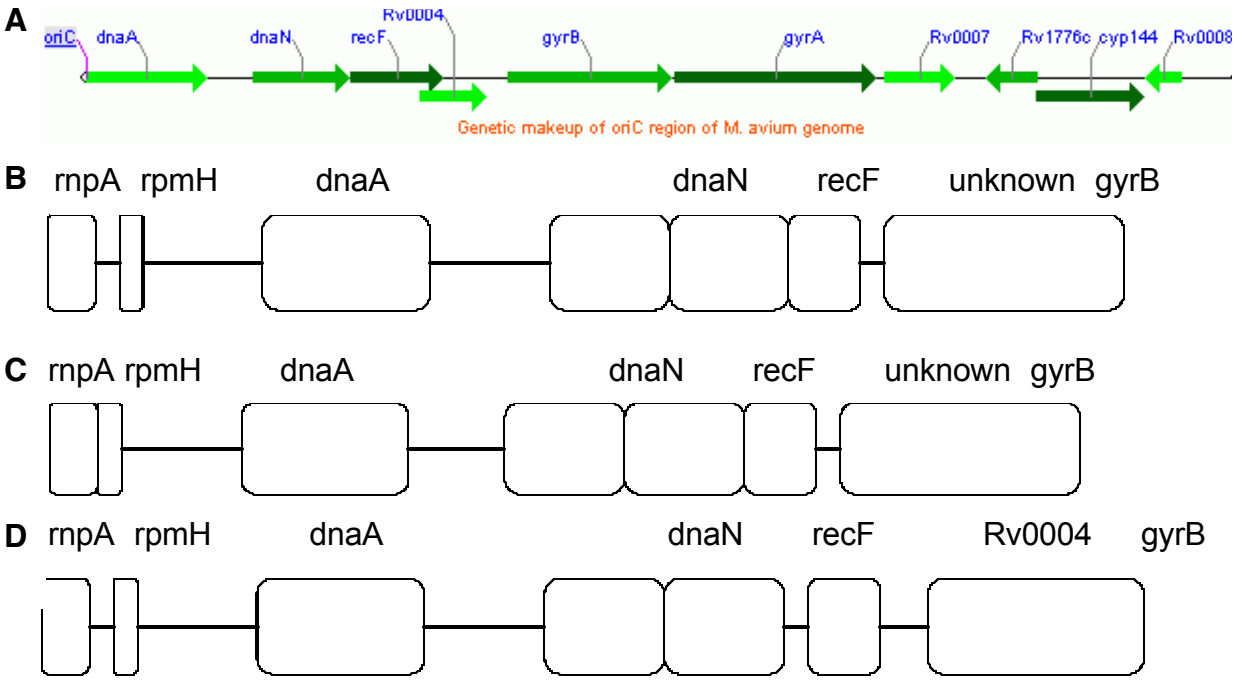

Figure 2. Diagram of the origin of replication (oriC) region of our annotation of the *M. avium subsp. avium* genome (a), compared with a published annotation of the same region for *M. avium subsp. avium* (b), *M. avium subsp. paratuberculosis* (c) and *M. tuberculosis* (d) (8) .

these automated steps within Microsoft Word™ or Microsoft Excel. As a final example, BlastParser™ (http://cbi.swmed.edu/computation/blastparser/) proved to be a tool of great value in the transfer of properly formatted text-based alignment result files. This software is able to convert BLAST output files into tab-delimited text that can be opened within Microsoft Excel™ , thus simplifying BLAST result comparison between the databases. Table 2 lists some of the resources that we consider to be useful for future genomic projects.

Sequencing inaccuracies were encountered during the annotation process and highlighted an important limitation of bioinformatic work in general; the downstream analysis is only as good as the precision of raw data. The sequencing errors we were able to detect were noticed through careful comparison of the sequences of MAA and the M. tuberculosis genomes. For instance, in some cases consecutive ORFs overlapped in MAA while their M. tuberculosis counterparts were shorter and contiguous. By

Table 2. List of useful resources for annotation of bacterial genomes.

Resource	Description and Website
Artemis™	Genomic visualization software; The Sanger Institute: http://www.sanger.ac.uk/Software/Artemis/
Genamics Expression™	DNA and protein sequence analysis software; Genamics Inc.: http://genamics.com/expression/index.htm
National Center for Biotechnology Information (NCBI) Website	Online BLAST tools, including: http://www.ncbi.nlm.nih.gov/BLAST/ http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
BlastParser™	A program that parses/reformats BLAST query results; Informatics Group at the Center for Biomedical Inventions: http://cbi.swmed.edu/computation/blastparser/
Microsoft Office™	Word Processor and Database software; Microsoft Corporation, www.microsoft.com
TubercuList Website	Genome database: *Mycobacterium tuberculosis* H37Rv; Pasteur Inst.: http://genolist.pasteur.fr/TubercuList
The Institute for Genomic Research(TIGR) website	The Institute for Genome Research, currently determining the Mycobacterium avium subsp. avium strain 104 sequence, http://www.tigr.org/
Mycobacterium avium comparative genomics	Comparative Genomics of the MAC at the Behr Lab, http://www.molepi.mcgill.ca/mac.htm

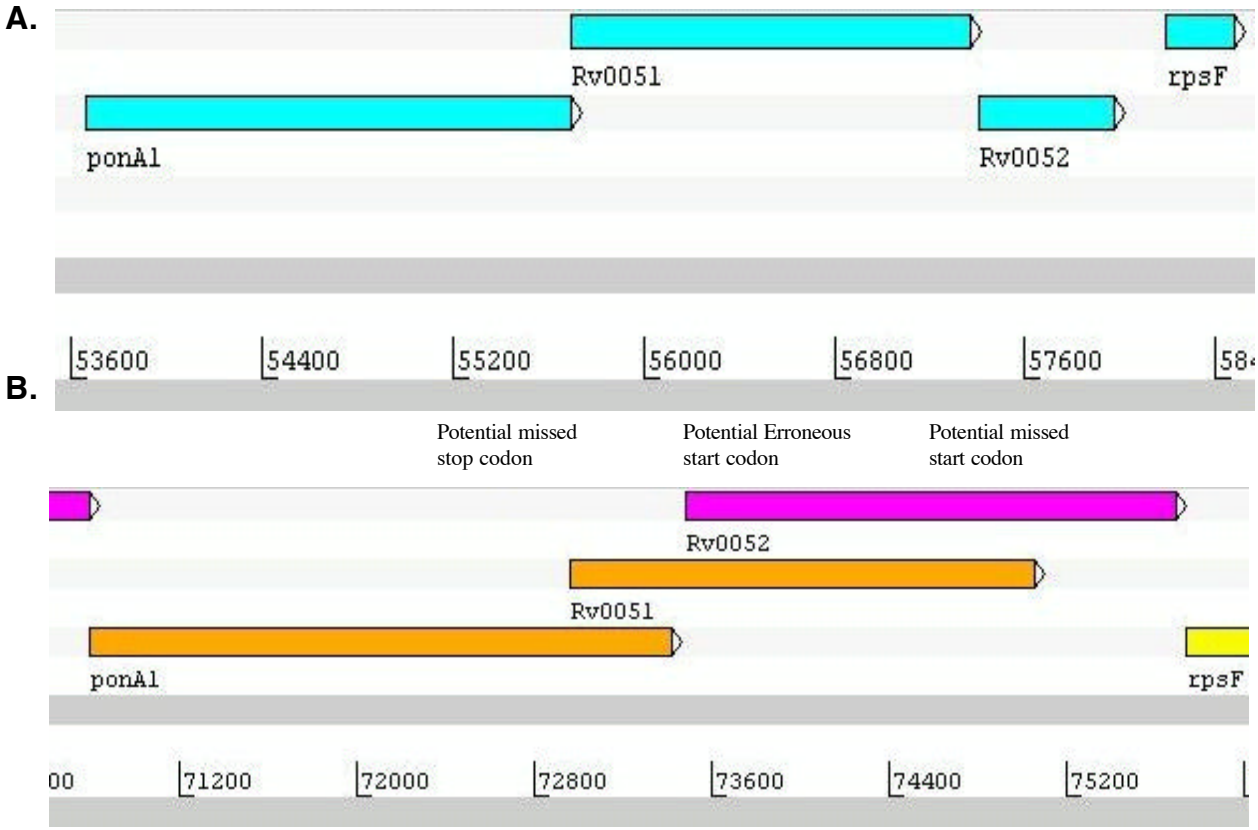

Figure 3. A possible sequencing error when comparing *M. tuberculosis* (a) and the complementary ORFs in *M. avium* subsp. avium (b). The ORFs in MAA overlap and are longer than the corresponding *M. tuberculosis* genes. These overlaps and gene length discrepancies are most likely due to sequencing inaccuracies, where stop and start codons are missed.

BLASTing only a part of these ORFs, we often found that high homology with the corresponding M. tuberculosis ORF was restricted to only part of the ORF, while the overlapping area would then show homology to an adjacent M. tuberculosis ORF. The observation that different ORFs were not distinguished in MAA suggested to us that sequencing errors at stop codons had obscured the predicted end of the ORF. Figure 3 shows an example of where a sequencing error occurred. The degree to which such errors impact on other aspects of the annotation unfortunately can only be determined when a "clean", final version of the MAA genome is released.

The annotation of the MAA 104 genome was intended for internal use, in order to facilitate genomic research of the MAC. In ongoing work, microarray-based genomic comparisons between members of the MAC have revealed large sequence polymorphisms, which are being used to construct a tentative phylogeny of the complex and to explore the biology of these organisms. Furthermore, the process and automations that were developed during the annotation project can be transported to annotating other sequenced genomes.

ACKNOWLEDGEMENTS

Funding for this project was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC). MK and MKL obtained research bursaries from the Fonds de la Recherche en Santé du Québec (FRSQ) and NSERC respectively. MB is a New Investigator of the Canadian Institutes of Health Research. MS is a recipient of the CIDS/CIHR/Bayer Healthcare fellowship award and is currently funded by the Fonds de la Recherche en Sante du Quebec (FRSQ). Preliminary sequence data was obtained from The Institute for Genomic Research website at http://www.tigr.org. We acknowledge Serge Mostowy for his original annotation of the MAA genome that served as an example for our efforts and provided the template for the DNA microarray design. We thank Melinda Pryor-Stinear of the Pasteur Institute for supplying us with Release 5 of the M. tuberculosis sequence and annotation well before its public release date, and Laura Carranza for insightful comments and discussions.

REFERENCES

1. Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, B. G. Barrell, and . 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537-544.
2. Behr, M. A., M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane, and P. M. Small. 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science 284:1520-1523.
3. Mostowy, S., D. Cousins, J. Brinkman, A. Aranaz, and M. A. Behr. 2002. Genomic deletions suggest a phylogeny for the *Mycobacterium tuberculosis* complex. J.Infect.Dis. 186:74-80.
4. Kato-Maeda, M., J. T. Rhee, T. R. Gingeras, H. Salamon, J. Drenkow, N. Smittipat, and P. M. Small. 2001. Comparing genomes within the species *Mycobacterium tuberculosis*. Genome Res. 11:547-554.
5. El Zaatari, F. A., M. S. Osato, and D. Y. Graham. 2001. Etiology of Crohn's disease: the role of *Mycobacterium avium* paratuberculosis. Trends Mol.Med. 7:247-252.
6. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J.Mol.Biol. 215:403-410.
7. Bannantine, J. P., E. Baechler, Q. Zhang, L. Li, and V. Kapur. 2002. Genome scale comparison of *Mycobacterium avium* subsp. paratuberculosis with *Mycobacterium avium subsp. avium* reveals potential diagnostic sequences. J.Clin.Microbiol. 40:1303-1310.
8. Bannantine, J. P., Q. Zhang, L. L. Li, and V. Kapur. 2003. Genomic homogeneity between *Mycobacterium* a*vium subsp. avium* and *Mycobacterium avium subsp. paratuberculosis* belies their divergent growth rates. BMC.Microbiol. 3:10.

David D. Shersher, Maksim Kirtsman, and **Mikael Katz-Lavigne** plan to undertake studies in medicine by the fall of 2004. They hope their undergraduate training in Microbiology/Immunology and their exposure to research will serve them well in a career combining clinical medicine and active research. **Dr. Makeda Semret** is a postdoctoral fellow and **Dr. Marcel A. Behr** is an Associate Professor at McGill University. They perform genomic studies of Mycobacteria with the aim to improve diagnosis and prevention of mycobacterial diseases.