Tuberculosis 93 (2013) 47–59

Contents lists available at SciVerse ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

REVIEW

# Updating and curating metabolic pathways of TB

Richard A. Slayden [a,*], Mary Jackson [a], Jeremy Zucker [b], Melissa V. Ramirez [a], Clinton C. Dawson [a], Rebecca Crew [a], Nicole S. Sampson [c], Suzanne T. Thomas [c], Neema Jamshidi [d], Peter Sisk [b], Ron Caspi [e], Dean C. Crick [a], Michael R. McNeil [a], Martin S. Pavelka [f], Michael Niederweis [g], Axel Siroy [g], Valentina Dona [h], Johnjoe McFadden [i], Helena Boshoff [h], Jocelyne M. Lew [j]

[a] Colorado State University, Fort Collins, CO, USA
[b] Broad Institute of MIT and Harvard, Cambridge, MA, USA
[c] Stony Brook University, Stony Brook, NY, USA
[d] University of California, San Diego, USA
[e] SRI International, Menlo Park, CA, USA
[f] University of Rochester Medical Center, Rochester, NY, USA
[g] University of Alabama at Birmingham, AL, USA
[h] NIH/NIAID, Bethesda, USA
[i] University of Surrey, Guildford, UK
[j] Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, Station 19, CH-1015 Lausanne, Switzerland

ARTICLE INFO

*Article history:*
Received 16 November 2012
Accepted 25 November 2012

*Keywords:*
Metabolic pathways
Cell envelope
Outer membrane proteins
Cholesterol
Cell division
Regulation
Toxin antitoxin
Reannotation
Curate
Bioinformatics
*Mycobacterium tuberculosis*

SUMMARY

The sequencing of complete genomes has accelerated biomedical research by providing information about the overall coding capacity of bacterial chromosomes. The original TB annotation resulted in putative functional assignment of ~60% of the genes to specific metabolic functions, however, the other 40% of the encoded ORFs where annotated as conserved hypothetical proteins, hypothetical proteins or encoding proteins of unknown function. The TB research community is now at the beginning of the next phases of post-genomics; namely reannotation and functional characterization by targeted experimentation. Arguably, this is the most significant time for basic microbiology in recent history. To foster basic TB research, the Tuberculosis Community Annotation Project (TBCAP) jamboree exercise began the reannotation effort by providing additional information for previous annotations, and refining and substantiating the functional assignment of ORFs and genes within metabolic pathways. The overall goal of the TBCAP 2012 exercise was to gather and compile various data types and use this information with oversight from the scientific community to provide additional information to support the functional annotations of encoding genes. Another objective of this effort was to standardize the publicly accessible *Mycobacterium tuberculosis* reference sequence and its annotation. The greatest benefit of functional annotation information of genome sequence is that it fuels TB research for drug discovery, diagnostics, vaccine development and epidemiology.



## 1. Introduction

The success of *Mycobacterium tuberculosis* is due in large part to its unique metabolism, which provides the capability to survive in the host environment, resist treatment and resume growth to relapse disease. It is widely accepted that tuberculosis is a dynamic disease that results from a combination of phenotypically diverse populations of bacilli in a continually changing host environment. This phenotypic capacity is encoded within the bacterial genome and is derived from an orchestrated expression of genes and ORFs tailored to specific alternative growth conditions encountered in the host.

The sequencing of complete genomes has accelerated biomedical research by providing an understanding of the metabolic diversity encoded by an organism. TB research was ushered into the post-genomic era in 1998 with completion of the whole genome sequence and annotation.[1] Genomic information is now an integral part of the research workflow, and impacts all aspects of TB research including drug discovery, diagnostics, vaccine development and epidemiology. The greatest benefit of genome sequence is that it provides the foundational information about the extent of coding capacity. In combination with next generation sequencing

* Corresponding author. Tel.: +1 970 491 2902.
*E-mail address:* Richard.slayden@colostate.edu (R.A. Slayden).

1472-9792/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tube.2012.11.001

technologies, investigators are now able to sequence genomes, including clinical isolates of interest in a few weeks making accurate and comprehensive genome annotation more important than ever.

The original annotation of the *M. tuberculosis* H37Rv genome resulted in the identification of 3924 ORFs, with ~60% assigned to specific metabolic functions, and the remaining ORFs categorized as conserved hypothetical proteins (~25%) or hypothetical proteins or encoding proteins of unknown function (~16%).[1] In 2002, the *M. tuberculosis* genome was reannotated. Eighty-two new protein-coding sequences (CDS) were included and 22 of these had an annotated function. In addition, supporting information including more than 300 gene names and over 1000 publications were added to support the annotation. With this version of the annotation, it was possible to assign a function to 2058 proteins (~52% of the 3995 proteins predicted) and only 376 putative proteins shared no homology with known proteins.[2]

The TB research community is now at the beginning of the next phases of post-genomics; namely reannotation and functional characterization by targeted experimentation. Arguably, this is the most significant time for basic microbiology in recent history. The objective of the Tuberculosis Community Annotation Project (TBCAP) jamboree, and the development of a TBCAP community is not to overwrite the original annotation and re-assign all the ORFs encoded in the *M. tuberculosis* genome, but rather provide additional information for previous annotations, and refine and substantiate the functional assignment of ORFs and genes within metabolic pathways, and to assign ORFs to gaps in metabolic pathways. Over the course of the last year, two NIAID-funded Centers, the Genomic Sequencing Center for Infectious Diseases at the Broad Institute and the Bacterial Bioinformatics Resource Center PATRIC, joined forces with the Gates-funded Tuberculosis Database (TBDB), and members of the TB research community to plan and support a collaborative TBCAP to achieve a full reannotation of the *M. tuberculosis* genome based on currently available information. The overall goal of TBCAP is to gather and compile various data types and use this information with oversight from the scientific community to provide additional information to support the functional annotations of encoding genes. Another objective of this effort is to standardize the publicly accessible *M. tuberculosis* reference sequence and its annotation. This report describes the results of TBCAP 2012 Jamboree of the group of curators collectively known as the metabolism working group (TBCAP MWG), and key metabolic areas of contemporary importance.

## 2. Methods

*Information used by TBCAP MWG for assignment of gene function*. Selected mycobacterial investigators participated in the metabolism working group during TBCAP jamboree 2012. TBCAP participants performed manual annotations in their respective areas of expertise. Information substantiating each annotation consisted of up to five parts: (1) an ontology term that described the function of the gene, (2) an evidence code that described the kind of experiment performed to determine the function, (3) a citation to a publication that described the experiment in detail, (4) a free text comment, and (5) the name of the curator who submitted the annotation. Curators uploaded annotations and information to the Basecamp content management system. The discussion boards of Basecamp are used to communicate additional information and comments. All the annotation information received was parsed and formatted using Python. The resulting formatted annotation information was uploaded to the Tuberculosis Database (TBDB website at http://tbdb.org) and automatically integrated with previous annotations. These annotations then served as the foundation and the functional annotations were projected onto a metabolic chart generated by Pathway-tools.[100] The annotations can be searched, viewed and exported from TBDB and PATRIC (Figure 1).

*The TBCAP is a "collaborative investigator community" of experts to share current functional annotation in there respective areas of interest.* Since the TBCAP 2012 meeting, the TBCAP MWG has transitioned to a supervised dispersed community annotation project[3] The leaders of each of the working groups serve as points of contact to coordinate future functional annotation efforts.

## 3. Results and discussion

The impact of the TB genome sequence and annotation ultimately depends on functional characterization. Indeed, being able to substantiate the functional annotation of genes is the next hurdle that the TB scientific community faces. To address this critical need, the TBCAP was created to bring together TB investigators to contribute expertise to define the function of encoded gene products. TBCAP promises to power the functional assignment of genes and importantly, make the information rapidly accessible to all investigators. Whether this transition from information being generated by bioinformatics analysis alone to involving the greater research community is being driven from the diminishing return of functional knowledge from homology-based comparative genomics is not the focus. Rather, the importance of experimental information and its rapid dissemination, and how this information is maintained is the goal. Ultimately, the conception behind the TBCAP is continuing community investment.

Developments in analytical techniques and in genetic methodologies allowing for the expression and disruption of genes in *M. tuberculosis* combined with the definition of the genome of this bacterium in 1998 have stimulated a rapid evolution of knowledge resulting in a relatively thorough understanding, not only of the basic metabolic tendencies and capabilities, and the structure of the mycobacterial cell envelope, but also its biosynthesis and underlying genetics. The availability of the genome sequence of *M. tuberculosis*, however, inevitably has resulted in the annotation of a multitude of ORFs for which biochemical or genetic evidence substantiating function is still lacking.

While the exact experimental evidence required for a functional annotation is debatable, it is generally agreed that a combination of experimental information about essentiality, phenotype, enzymatic or structural role, protein–protein interaction, protein localization and conditions of expression and production is needed. Toward the functional annotation of the encoded gene products of *M. tuberculosis*, TBCAP MWG assembled all the available information including ontology, EC number, PubMed citations for each gene (Figure 2). This provided a framework for assignment of a gene to a categorization based on extent of information and grouped genes into one of 3-categories (Categories I–III) indicating functional annotation or another category not represented by annotated genes (Category IV).

More than 30,000 annotations were contributed from 83 annotators using 13 kinds of experimental evidence from 1104 publications. Together, 5246 unique ontology terms, 1171 different gene symbols, 1230 specific gene names and 2714 free text comments have been identified for 2183 genes in *M. tuberculosis* (Table 1), derived by numerous sources. Functional gaps in pathways of contemporary importance such as cell envelope macromolecular synthesis, cholesterol metabolism, septum formation and cell division regulation and toxin–antitoxin loci were then targeted for subsequent focused functional curation.

*Category I. Metabolic functions assigned to genes*. Assignment of a gene to this category is based on information of standard ontology including EC number, MetaCyc or KEGG. Evidence for assignment includes at least one PubMed citation, along with an evidence code from a standard ontology that specifies the type of experiment