Tuberculosis 93 (2013) 12–17

Contents lists available at SciVerse ScienceDirect

Tuberculosis

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

REVIEW

Database resources for the tuberculosis community

Jocelyne M. Lew [a,b,j], Chunhong Mao [c,j], Maulik Shukla [c], Andrew Warren [c], Rebecca Will [c], Dmitry Kuznetsov [d], Ioannis Xenarios [a,d,e], Brian D. Robertson [f], Stephen V. Gordon [g], Dirk Schnappinger [h], Stewart T. Cole [b,*,j], Bruno Sobral [c,i,**,j]

[a] Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
[b] Global Health Institute, École Polytechnique Fédérale de Lausanne, Station 19, 1015 Lausanne, Switzerland
[c] Virginia Bioinformatics Institute at Virginia Tech, Blacksburg, VA 24061, USA
[d] Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
[e] University of Lausanne, Center for Integrative Genomics, Lausanne, Switzerland
[f] MRC Centre for Molecular Bacteriology and Infection, Imperial College London, Exhibition Road, South Kensington, London SW7 2AZ, UK
[g] UCD Conway Institute of Biomolecular and Biomedical Research, Belfield, Dublin, Ireland
[h] Department of Microbiology and Immunology, Weill Cornell Medical College, New York, NY, USA
[i] Nestlé Institute of Health Sciences, Campus of École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland

ARTICLE INFO

Article history:
Received 31 October 2012
Accepted 27 November 2012

Keywords:
TubercuList
TB database
PATRIC
Genomics
Database

SUMMARY

Access to online repositories for genomic and associated "-omics" datasets is now an essential part of everyday research activity. It is important therefore that the Tuberculosis community is aware of the databases and tools available to them online, as well as for the database hosts to know what the needs of the research community are. One of the goals of the Tuberculosis Annotation Jamboree, held in Washington DC on March 7th–8th 2012, was therefore to provide an overview of the current status of three key Tuberculosis resources, TubercuList (tuberculist.epfl.ch), TB Database (www.tbdb.org), and Pathosystems Resource Integration Center (PATRIC, www.patricbrc.org). Here we summarize some key updates and upcoming features in TubercuList, and provide an overview of the PATRIC site and its online tools for pathogen RNA-Seq analysis.

1. Introduction

Bacterial genomes can now be sequenced in a matter of days for a few hundred dollars. Genomic, transcriptomic, and associated data-sets are becoming so large that the extent to which we provide user-friendly access will determine how much the Tuberculosis (TB) research community can learn from them. The TB community therefore needs to take stock of how we are placed to best exploit this data, and how we will deal with issues such as data analysis, curation and dissemination.

Web-accessible databases and analysis tools are an essential part of how we interpret and interact with genome data; as a community we need to be kept up to date with developments in these areas. This was one of the key aims of the TB Annotation Jamboree held in Washington on March 7th–8th 2012, where a session was devoted to databases and related issues. We focused our discussions on three of the key resources for TB genome data on the web, namely TubercuList (tuberculist.epfl.ch), TB Database (www.tbdb.org), and Pathosystems Resource Integration Center (PATRIC, www.patricbrc.org). The goal of this manuscript is to update and introduce the community to developments in TubercuList and PATRIC, as detailed below. Web-links to the databases and tools mentioned in this article can be found in supplemental Table 1.

2. TubercuList

2.1. Overview

TubercuList (tuberculist.epfl.ch) is a relational database for the genome sequence annotation of *Mycobacterium tuberculosis* H37Rv, the reference strain commonly used in the study of TB. This infectious disease continues to be a serious global health issue, killing 1.4 million people in 2010.[1] The database is a well-established resource, having been maintained since its inception in 1998.[2] It is a gene-centric database, and in its current form

* Corresponding author. Global Health Institute, École Polytechnique Fédérale de Lausanne, Station 19, 1015 Lausanne, Switzerland.
** Corresponding author. Virginia Bioinformatics Institute at Virginia Tech, Blacksburg, VA 24061, USA.
E-mail addresses: stewart.cole@epfl.ch (S.T. Cole), sobral@vbi.vt.edu (B. Sobral).
[j] Equal contributions.

1472-9792/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tube.2012.11.003

provides information on annotated *M. tuberculosis* H37Rv genes and proteins, including functional annotation, orthologous genes in closely related species, gene ontology terms, structural information, and cross-references to several external resources including the TB Drug Resistance Mutation Database, a comprehensive list of polymorphisms associated with drug resistance,[3] and The TDR Targets Database, designed to facilitate the prioritization of drug targets.[4]

One of the greatest strengths of the TubercuList database lies in the fact that it has been subject to continuous manual annotation since the first release of the genome sequence and annotation.[5,6] It is updated with experimental evidence from the scientific literature resulting in changes to gene boundaries, addition of new genes both protein- and RNA-encoding, improvements in functional annotation, and assignment or modification of gene names. This is enriched with data on the characterization of mutant strains, protein localization determined by proteomics studies, gene essentiality under different growth conditions, gene regulatory information, and operon structure. Citations are also provided for all such manually selected publications from which data has been extracted.

With advances in next-generation sequencing technologies and decreasing costs, the number of genome projects is increasing at a remarkable rate. According to the Genomes OnLine Database, there were 11472 genome sequencing projects as of September 2011, with 2907 complete.[7] Although these numbers are impressive, for the majority of newly sequenced genomes, the annotation will not go beyond computer-generated predictions.[8,9] Moreover, vast amounts of empirical data are constantly being produced at the bench, particularly from high-throughput and genome-wide studies, and it is critical to extract key findings and apply them to the genome annotation so that it is readily accessible in a useful form for the entire research community. It is through this challenging task of manual annotation from the literature, collecting and organizing data from disparate sources, that we strive to extend the value of TubercuList for TB researchers as a current and reliable resource.

Through a partnership between the École Polytechnique Fédérale de Lausanne (EPFL) and the SIB Swiss Institute of Bioinformatics (SIB), updating of the TubercuList database is now carried out by the SIB, which is part of the UniProt consortium with the European Bioinformatics Institute (EBI) and the Protein Information Resource (PIR), and produces the Swiss-Prot section of the UniProt Knowledgebase (UniProtKB/Swiss-Prot). UniProtKB/Swiss-Prot is an expertly curated database for protein information and *M. tuberculosis* is one of several model organisms on which the database focuses (www.uniprot.org).[10] Annotation carried out by curators for UniProtKB/Swiss-Prot and for TubercuList is exchanged thereby maximizing the results of manual annotation efforts made by both groups.

As improvements, modifications, and the incorporation of new data to the *M. tuberculosis* H37Rv genome annotation are continually being made in TubercuList, we describe in the following sections some of the recent changes that have been made and the updates that will appear in the next release of the database (R26).

2.2. Updates to TubercuList annotation, release 25

The TubercuList database is updated approximately every four months with information from the literature as well as with new or updated cross-references to external databases. The number of genes annotated in the current release of the database, version R25 completed in April 2012, has not changed significantly, now at 4095, and the coordinates of only four coding sequences (CDS) have been altered. New genes added since version R20 from June 2010[6] include four CDS, one of which is a replacement for *rv0061*, now annotated in the opposite orientation as *rv0061c* as indicated by RNA-Seq data[11,12] (Uplekar et al., in preparation). Continuing with the trend reported previously,[6] new CDS added to the annotation are typically small, being less than or close to 100 amino acids. Also added are two non-coding RNAs, regulatory molecules that are a topic of increasing interest.[13,14] There are now a total of 23 such small RNA genes annotated in TubercuList.

Advances in mass spectrometry-based proteomic methods are providing the ability to identify wider ranges of proteins, reliably and accurately.[15–17] In TubercuList, 2828 proteins are annotated as having been identified in a proteomics study, 1114 more than in the R20 version of the database. This validates 70% of protein-coding features annotated in the genome, although a recent study, whose results await addition to the database, reports a higher ~80% coverage of the predicted genes.[18] Of the 2828 proteins, 23% are categorized as *8-Unknown* or *10-Conserved hypothetical proteins*, verifying that these predicted CDS are actual proteins produced by the bacterium.

The current distribution of all *M. tuberculosis* H37Rv genes across eleven functional categories is shown in Table 1. The function of one quarter (1048) of annotated CDS remains unknown, although this number is steadily being reduced as more proteins are characterized. Changes have been made to the functional category of 55 CDS (See Table 1) and approximately half of these changes move CDS from *10-Conserved hypothetical proteins* to involvement in *1-Lipid metabolism*. In addition to this functional annotation, 85 gene names have been added or modified, and more than half of these concern toxin-antitoxin genes, mainly for *vapBC* gene pairs.

Structural biology plays an important role in understanding the mechanisms of protein function as well as in predicting functions for unknown proteins, and can also make a significant contribution to drug development.[19] TubercuList now links to 1019 structures in the Protein Data Bank (www.pdb.org), representing 365 unique proteins. This is a significant achievement in the field of TB research. However, as the protein structures of most of the CDS annotated in the genome remain unknown, protein structure prediction methods are a necessary tool to be used where experimental structure determination has not yet succeeded[20] (see Mao et al., this issue).

In this period, we have also added information on gene regulation involving eight regulatory proteins[21–29]; 544 genes now have annotation on regulation. This includes data on predicted and confirmed regulons, identification of DNA-binding motifs as well as demonstration of DNA-binding by the regulatory proteins, and changes in expression levels in the absence or overexpression of the regulatory protein. Experimental evidence of operon structure has also been added for 38 genes.

Table 1
Distribution of genes across functional categories (TubercuList version R25).

Functional category		Gene number	Change from R20
0	Virulence, detoxification, adaptation	238	10
1	Lipid metabolism	272	25
2	Information pathways	242	1
3	Cell wall and cell processes	773	–
4	Stable RNAs	73	2
5	Insertion sequences and phages	147	–
6	PE/PPE	168	–
7	Intermediary metabolism and respiration	936	13
8	Unknown	16	–
9	Regulatory proteins	198	3
10	Conserved hypothetical proteins	1032	−49