BMC Microbiology

Database

MyBASE: a database for genome polymorphism and gene function studies of *Mycobacterium*

Xinxing Zhu†[1,2], Suhua Chang†[2], Kechi Fang[2], Sijia Cui[2], Jun Liu[3],
Zuowei Wu[4], Xuping Yu[5], George F Gao[4], Huanming Yang[6], Baoli Zhu*[4]

Background

Mycobacteria are notorious for its two species, *Mycobacterium tuberculosis* (*M. tb*) and *Mycobacterium leprae* (*M. leprae*), the causative agent of tuberculosis (TB) and leprosy, respectively. In addition to *M. tb* and *M. leprae*, a number of mycobacterial pathogens also cause human and animal diseases, including *Mycobacterium bovis* (*M. bovis*), the causative agent of classical bovine tuberculosis, and *Mycobacterium ulcerans* (*M. ulcerans*), which causes Buruli ulcers. The emergence of multi-drug resistant strains of pathogenic mycobacteria has made the development of better vaccines and new drugs and novel control strategies a top priority.

The genome of *M. tb* H37Rv was the first mycobacterial genome to be sequenced and was published in 1998 [1], which was followed by the genome of *M. leprae* in 2001 [2]. The complete sequencing of these genomes greatly contributed to the understanding of the unique physiology and pathogenesis of mycobacteria. With the development of DNA sequencing technologies in recent years, a total of 18 complete mycobacterial genomes have been available and deposited in public domains thus far. This progress offers an unprecedented opportunity to understand the virulence mechanisms of mycobacteria at the molecular level, which offers insight into the development of potential control strategies.

One of the most significant findings in mycobacterial research was from the genome-wide comparison between virulent (e.g. *M. tb* H37Rv or *M. bovis*) and avirulent strains (e.g. *M. bovis* BCG) [3]. This genomic comparison unveiled large sequence polymorphisms (LSPs), usually called regions of difference (RDs), which are believed to be the major source of genomic diversity [4,5] and probably contribute to the phenotypic differences [6]. Some of the LSPs/RDs have been shown be important for virulence and pathogenicity. For example, RD1, which is deleted in all BCG strains but is present in virulent strains of *M. tb* or *M. bovis*, has been shown to be essential for *M. tb* virulence [7-9]. The success of systematic genetic screening of mycobacterial mutants from different environments [10-13], coupled with focused investigation into individual virulence genes, has contributed to the functional genomic data of mycobacteria [14], which has provided useful information in understanding the physiology and pathogenesis of this unique bacterial genus.

Currently, several public resources for mycobacterial research are available, including the TB database [15], which is an integrated platform of genomic data with special interest in microarray analysis; GenoList http://genolist.pasteur.fr/, which focuses on the gene annotation of six mycobacterial strains [16]; MycoDB from xBASE [17,18], which provides search and visualization tools for genome comparison of mycobacteria; MycoperonDB [19], which is a database of predicted operons in 5 mycobacterial species; MGDD [20], a mycobacterial genome divergence database derived from an anchor-based comparison approach [21]; GenoMycDB [22], a database for pair-wise comparison of six mycobacterial genomes; and MtbRegList [23], which is dedicated to the analysis of transcriptional regulation of *M. tb*. Although each of these databases provides unique and useful information, none are focused on LSPs, essential genes, and the relationship between these and virulence. MyBASE was therefore developed to meet these needs. In addition to providing a platform for analyzing all published mycobacterial genomes, MyBASE features important information on genomic polymorphisms, virulence genes, and essential genes. As such, MyBASE will help researchers to easily explore and analyze data in a user-friendly and cross-referenced fashion, thereby facilitating functional genomic studies. This will inevitably enhance our understanding on the virulence mechanisms, genome structure, and molecular evolution of mycobacteria.

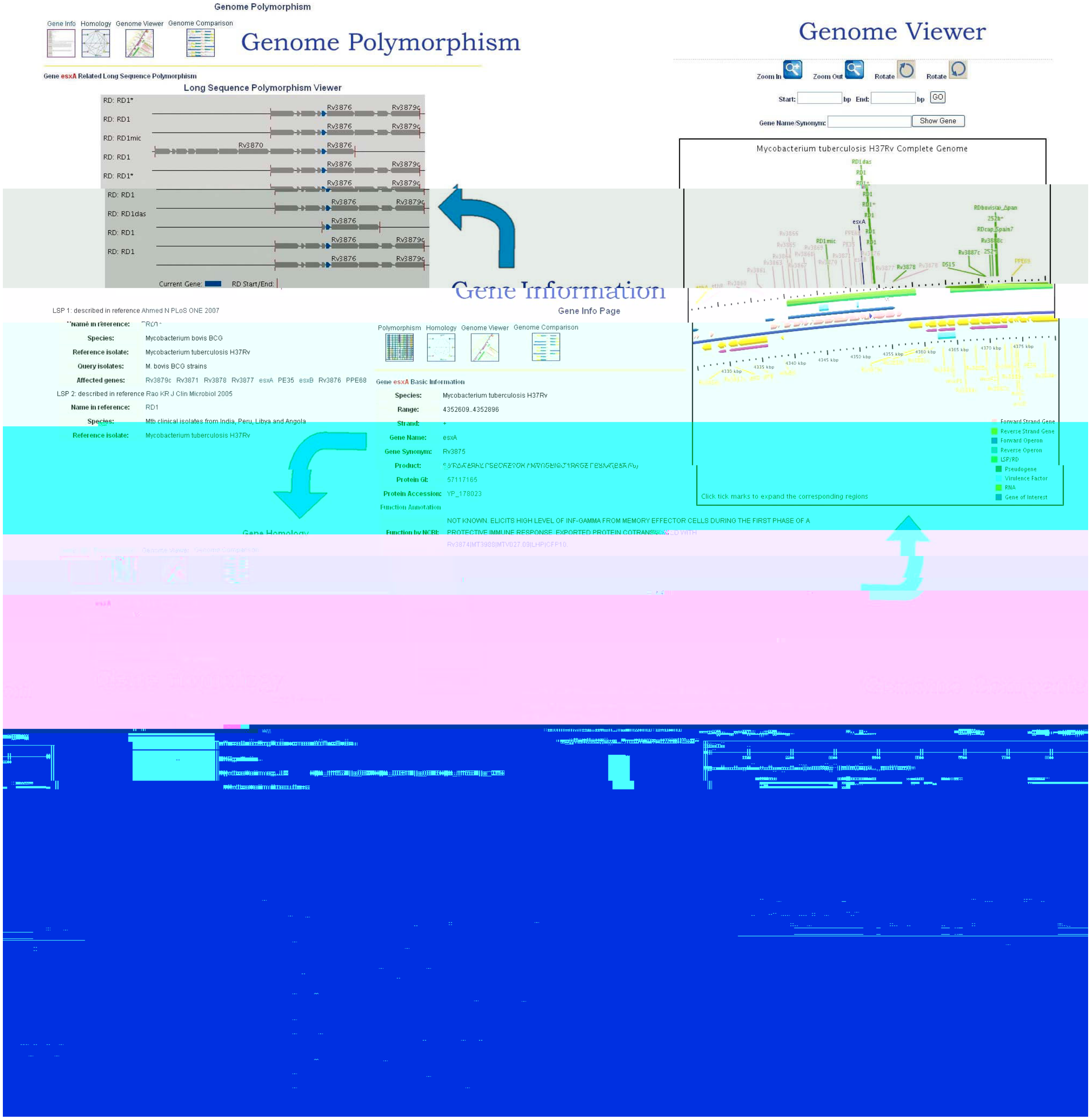
Construction and content

Data sources and curation

MyBASE contains data from both our own experiments and public resources. There are four main types of data: 1) genome sequences with curated annotations, 2) genome polymorphism data, particularly LSPs identified among different mycobacterial genomes, 3) functional gene annotations with a specific focus on virulence genes and essential genes, and 4) predicted operons.

All complete genome sequences and original annotation files were downloaded from NCBI ftp://ftp.ncbi.nih.gov/genomes/Bacteria. Curations were made to clarify inconsistencies resulting from different annotations provided by the original sequence providers. For Clusters of Orthologous Groups (COGs) that were inconsistently designated [24], we refined the COGs using the algorithm previously described [25].

We have recently used the NimbleGen tiling microarray method for whole-genome comparison of 13 BCG strains with subsequent confirmation by DNA re-sequencing [26]. A total of 42 deletions were identified, four of which are novel [26]. These novel deletions are believed to have an impact on virulence or immunogenicity of the corresponding BCG strains [26]. All data and analytical results were incorporated into MyBASE. In addition to our self-generated data, other polymorphism datasets, particularly LSPs/RDs that were included in MyBASE were extracted from public literatures. After the first usage of microarray to study genome polymorphism in 1999 [3], a growing trend emerged to generate systematic genome polymorphism data [27-29]. We performed an extensive literature



LSP 1: described in reference Ahmed N PLoS ONE 2007

Name in reference:	RD1*
Species:	Mycobacterium bovis BCG
Reference isolate:	Mycobacterium tuberculosis H37Rv
Query isolates:	M. bovis BCG strains
Affected genes:	Rv3879c Rv3871 Rv3878 Rv3877 esxA PE35 esxB Rv3876 PPE68

LSP 2: described in reference Rao KR J Clin Microbiol 2005

Name in reference:	RD1
Species:	Mtb clinical isolates from India, Peru, Libya and Angola
Reference isolate:	Mycobacterium tuberculosis H37Rv

Gene esxA Basic Information

Species:	Mycobacterium tuberculosis H37Rv
Range:	4352609..4352896
Strand:	+
Gene Name:	esxA
Gene Synonym:	Rv3875
Product:	[illegible]
Protein GI:	57117165
Protein Accession:	YP_178023
Function Annotation	
Function by NCBI:	NOT KNOWN. ELICITS HIGH LEVEL OF INF-GAMMA FROM MEMORY EFFECTOR CELLS DURING THE FIRST PHASE OF A PROTECTIVE IMMUNE RESPONSE. EXPORTED PROTEIN COTRANS[illegible]ED WITH Rv3874|MT3988|MTV027.09|LHP|CFP10.

Figure 1

Schematic representation of the data repository and the interrelation of functional modules in MyBASE. After the gene of interest is found, users can check whether it is in a genomic polymorphic region, compare the selected genome with MCV, explore the details of its genome segment with Genome Viewer or view its homolog distributions.

review to extract information about each LSP/RD from original experiments. We found inconsistencies between the nomenclature of LSPs (RDs) used by different groups and so to avoid further confusion, we have kept the original nomenclature from each group. However, we have provided the reference information and a hyperlink to the PubMed entry for each LSP/RD dataset.

tional annotations by the research community will be added to MyBASE periodically to keep the database up-to-date. The functionality of the LSP search and viewer will be enriched and enhanced. In addition, new tools, such as software packages for phylogenomic study will be integrated. Finally, MyBASE also provides an opportunity for the mycobacterial research community to standardize nomenclature, data formats of gene, and polymorphism annotations.

Conclusion

MyBASE is a unique data warehouse and analysis platform for the mycobacterial research community, which features a collection and curation of a large amount of LSP and functional genomic data. By developing various tools, MyBASE can help researchers to easily explore and investigate genome deletions, virulence factors, essential genes, and operon structure of mycobacteria.

Availability and requirements

The database is freely available on http://mybase.psych.ac.cn.

Authors' contributions

XZ designed the database, collected, curated the data and wrote the manuscript. SC analyzed the data and developed the database. KF and SC developed the database and did the programming work. JL, ZW, and XY performed the microarray experiments and analyzed the data. GFG, and HY revised the manuscript. BZ and JW supervised the work, manage the team and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was sponsored by the National Natural Science Foundation of China (NSFC, Grant No. 30700441, 30221004) and Beijing Municipal Science and Technology Commission (Grant No: Z0005190043521).

References

1. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, *et al.*: **Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence.** *Nature* 1998, **393(6685):**537-544.