IJPSR (2017), Volume 8, Issue 12 (Review Article)

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 22 April, 2017; received in revised form, 17 July, 2017; accepted, 25 July, 2017; published 01 December, 2017

REVIEW ON COMPARATIVE GENOMICS FOR *MYCOBACTERIUM TUBERCULOSIS* STRAINS

C.R. Subhasree *[1], R. Sri Kamatchi Priya [1], M. Diwakar [1], S. Subramaniam [1] and S. Shyama [2]

Department of Biochemistry [1], Regenix Super Speciality Laboratories Private limited, No: 42, Loganathan Nagar, 2nd Street, Choolaimedu, Chennai - 600094, Tamil Nadu, India.
Department of Clinical Biochemistry [2], Apollo Hospitals, Chennai - 600006, Tamil Nadu, India.

Keywords:

Mycobacterium tuberculosis, Genomics, Extensive drug resistant strains, Multiple drug resistant strains

Correspondence to Author:
C. R. Subhasree

SFF,
Second Floor, New No.10,
Old No.18, Vallalar Street,
Padmanaba Nagar, Choolaimedu,
Chennai - 600094, Tamil Nadu,
India.

E-mail: c.r.subhasree@gmail.com

ABSTRACT: Tuberculosis caused by multidrug-resistant (MDR) and extensively drug-resistant (XDR) *Mycobacterium tuberculosis* (MTB) strains is a growing problem in many countries. The availability of the complete nucleotide sequences of several MTB genomes allows to use the comparative genomics as a tool to study the relationships of strains and differences in their evolutionary history including acquisition of drug-resistance. In our study, online resources for comparative genomics analysis between strains by using databases such as MGDD (Mycobacterial Genome Divergence Database), Biohealth base, GenomycDB, *Mycobacterium tuberculosis* database, and TBDB (Tuberculosis Database). Taken together, this study reveals that strain-specific variations in protein expression patterns have a meaningful impact on the biology of the pathogen.

INTRODUCTION: Tuberculosis is a worldwide pandemic. Every second someone in this planet is newly infected with *Mycobacterium tuberculosis*, the etiologic agent of tuberculosis (TB), of which 5-10% become sick infectious at some point time during their life. Overall, one-third of the human race is currently infected with the TB bacillus [1]. Globally, 9.2 million new cases and 1.7 million deaths from TB occurred in 2006 [2]. There were 9.27 million new TB cases identified in 2007 and a total of 1.77 million people died from TB in 2007 (including 456 000 people with HIV), equal to about 4800 deaths a day.

QUICK RESPONSE CODE

DOI: 10.13040/IJPSR.0975-8232.8(12).5022-42

Article can be accessed online on: www.ijpsr.com

DOI link: http://dx.doi.org/10.13040/IJPSR.0975-8232.8 (12).5022-42

80% of the new TB cases identified in 2007 were in just 22 countries. Per capita, the global TB incidence rate is falling, but the rate of decline is very slow - less than 1%. TB is a disease of poverty, affecting mostly young adults in their most productive years.

The vast majority of TB deaths are in the developing world, with more than half occurring in Asia. Among the 15 countries with the highest estimated TB incidence rates, 13 are in Africa, while half of all new cases are in six Asian countries (Bangladesh, China, India, Indonesia, Pakistan and the Philippines) [3].

Drugs and Drug Resistance: The history of tuberculosis (TB) changed dramatically after the introduction of antimycobacterial agents. Drug treatment is fundamental for controlling TB, promoting the cure of the patients and breaking the chain of transmission when the anti-tuberculosis drug regimen is completely and correctly followed.

Anti-tuberculosis drug treatment started in 1944, when streptomycin (SM) and paraaminosalicylic acid **(**PAS) were discovered. In 1950, the first trial was performed comparing the efficacy of SM and PAS both as monotherapy or combined. The study demonstrated that combined therapy was more effective and resulted in the first multidrug anti-tuberculosis treatment that consisted of a long course of both drugs. In 1952, a third drug, isoniazid (INH), was added to the previous combination, greatly improving the efficacy of treatment, but which still had to be administered for 18 - 24 months. In 1960, ethambutol (EMB) substituted PAS, and the treatment course was reduced to 18 months. In the 70's, with the introduction of rifampicin (RIF) into the combination, treatment was shortened to just nine months.

In 1980, pyrazinamide (PZA) was introduced into the antituberculosis treatment, acid treatment was reduced further to just six months. Soon after the introduction of the first anti-mycobacterial drugs, drug resistant bacilli started to emerge, but the launch of both combination therapy and new and more effective drugs seemed to be enough to control the disease. In fact, it was thought that TB could be eradicated by the end of 20th century. However, TB unexpectedly re-emerged in the 1980s, and in the following years there was a significant increase in the incidence of particularly dangerous forms of *viz*., multi-drug resistant strains and extensively drug-resistant strains. Since 1970, no new drug has been discovered for anti-tuberculosis treatment, which today seems insufficient to confront the disease [4].

Multi-drug resistant tuberculosis (MDR-TB) is defined as the disease caused by TB bacilli resistant to at least isoniazid and rifampicin, the two most powerful anti-TB drugs [5]. Rates of MDR-TB are high in some countries, especially in the former Soviet Union, and threaten TB control efforts. While drug-resistant TB is generally treatable, it requires extensive chemotherapy (for two years) with second-line anti-TB drugs, which are costlier than first-line drugs, and produce adverse drug reactions that are more severe, though manageable [6]. TB was curable till XDR-TB was reported. XDR-TB (Extensively Drug Resistant Tuberculosis) is defined as disease caused by a strain of *M. tuberculosis* that is resistant to isoniazid and rifampicin plus any of the fluoroquinolones and at least one of the three injectable second-line drugs (amikacin, kanamycin, capreomycin) [7]. Until 50 years ago, there were no medicines to cure TB. Now, strains that are resistant to a single drug have been documented in every country surveyed; what is more, strains of TB resistant to all major anti-TB drugs have emerged.

Drug-resistant TB is caused by inconsistent or partial treatment, when patients do not take all their medicines regularly for the required period because they start to feel better, because doctors and health workers prescribe the wrong treatment regimens, or because the drug supply is unreliable [8]. Quality-assured second-line anti-TB drugs are available at reduced prices for projects approved by the Green Light Committee [9]. The emergence of extensively drug-resistant (XDR) TB, particularly in settings where many TB patients are also infected with HIV, poses a serious threat to TB control, and confirms the urgent need to strengthen basic TB control and to apply the new WHO guidelines for the programmatic management of drug-resistant TB [10].

Causes of DR-TB: DR-TB is essentially a man-made phenomenon although causes of the drug resistant tuberculosis may be microbial, clinical and programmatic. From a microbiological perspective, resistance is caused by a genetic mutation that makes a drug ineffective against the mutant bacilli. From a clinical and programmatic perspective, it is an inadequate or poorly administered treatment regimen that allows a drug-resistant strain to become the dominant strain in a patient infected with TB [11].

Magnitude of the DR-TB Problem: The incidence of drug resistance has increased since the first drug treatment for TB was introduced in 1943. The emergence of MDR-TB following the widespread use of rifampicin beginning in the 1970s led to the use of second-line drugs. Improper use of these drugs has fuelled the generation and subsequent transmission of highly resistant strains of TB termed extensively DR-TB or XDR-TB [12]. Based on available information from the duration of the Global Project, the most recent data available from 116 countries and settings were weighted by

the population in areas surveyed, representing 2 509 545 TB cases, with the following results: global population weighted proportion of resistance among new cases: any resistance 17.0% (95% confidence limits (CLs), 13.6–20.4), isoniazid resistance 10.3% (95% CLs, 8.4 - 12.1) and MDR-TB 2.9% (95% CLs, 2.2 - 3.6). Global population weighted proportion of resistance among previously treated cases: any resistance 35.0% (95% CLs, 24.1 - 45.8), isoniazid resistance 27.7% (95% CLs, 18.7 - 36.7), MDR-TB 15.3% (95% CLs, 9.6 - 21.1). Global population weighted proportion of resistance among all TB cases: any resistance 20.0% (95% CLs, 16.1 - 23.9), isoniazid resistance 13.3% (95% CLs, 10.9 - 15.8) and MDR-TB 5.3% (95% CLs, 3.9 - 6.6). Based on drug resistance information from these 116 countries and settings reporting to this project, as well as nine other epidemiological factors, it is estimated that 489 139 (95% CLs, 455 093 - 614 215) cases emerged in 2006.

China and India carry approximately 50% of the global burden of MDR-TB and the Russian Federation a further 7% [13]. A survey was conducted to determine the extent of resistance to second-line drugs by the United States Centers for Disease Control and Prevention (CDC) and WHO in 2006. The survey found that of the isolates tested against second-line drugs in the 49 contributing countries, 20% were MDR-TB and 2% were XDR-TB [14]. Strains of XDR-TB have been reported in every region of the world, with as many as 19% of MDR-TB strains found to be XDR-TB, a proportion that has more than tripled in some areas since 2000 [15].

India is the worst affected country by tuberculosis in number among all the countries in the world and India ranks first in having 2.0 million out of 9.27 million identified new cases in 2007 and again India ranks at the top in harbouring the most number of MDR-TB cases (131 000 out of estimated 0.5 million cases) [16].

Co-infection with HIV: TB is an opportunistic disease that preys on weakened immune systems; if not diagnosed early it can progress rapidly in HIV-positive individuals. TB is the leading infectious killer of people living with HIV/AIDS. Up to 50% of people with HIV or AIDS develop TB. Worldwide, 14 million are co-infected with TB and HIV. In the South East Asia (SEA) Region approximately three million are co-infected. TB causes at least 11% of AIDS deaths. The lifetime risk of developing TB in an HIV-positive individual is 50% as compared to the 5 - 10% risk of someone who is HIV - negative. Most importantly, TB can successfully be treated even if someone is HIV - infected. Treatment of TB can prolong and improve the quality of life for HIV-positive people but cannot alone prevent people from dying of AIDS. In the South East Asia Region, five of the 11 countries are high- or moderate - TB / HIV burden countries: India, Indonesia, Myanmar, Nepal and Thailand [17]. TB is a leading cause of death among people who are HIV-positive [18].

Among 9.2 million new cases and 1.7 million deaths reported in 2006 globally 0.7 million cases and 0.2 million deaths were in HIV-positive people [19]. In Africa, HIV is the single most important factor contributing to the increase in incidence of TB since 1990 [20]. Of the 9.27 million incident TB cases in 2007, an estimated 1.37 million (15%) were HIV-positive; 79% of these HIV-positive cases were in the African Region and 11% were in the South-East Asia Region [21]. An estimated 1.3 million deaths occurred among HIV-negative incident cases of TB (20 per 100 000 population) in 2007. There were an additional 456000 deaths among incident TB cases who were HIV-positive; these deaths are classified as HIV deaths in the International Statistical Classification of Diseases (ICD-10).

The 456 000 deaths among HIV-positive incident TB cases equate to 33% of HIV-positive incident cases of TB and 23% of the estimated 2 million HIV deaths in 2007 [22]. The interface between TB and HIV is increased in countries like India where both TB and HIV infection are maximally prevalent in people of 15-49 years of age. Also, the socioeconomic factors of poverty, ignorance and stigma are common to both the diseases [23]. In 1993 TB was declared as global health emergency [24] recently too as a global emergency [25]. WHO and its international partners have formed the TB/HIV Working Group, which develops global policy on the control of HIV-related TB and advises on how those fighting against TB and HIV can work together to tackle this lethal combination.

The interim policy on collaborative TB / HIV activities describes steps to create mechanisms of collaboration between TB and HIV / AIDS programmes, to reduce the burden of TB among people and reduce the burden of HIV among TB patients [26]. In the research perspective, there is a need for finding out solutions for tuberculosis using the existing knowledge and technology. The developments in genomics and proteomics have significantly enhanced our knowledge on TB bacilli.

***Mycobacterium tuberculosis*:** *M. tuberculosis* is a highly contagious, airborne, slow-growing, Gram-positive, aerobic, rod-shaped, acid-fast bacillus. The cell wall has high lipid content and allows the bacteria to survive within macrophages. It also provides the organism with a resistant barrier to many common drugs [27]. Man is the primary host for *M. tuberculosis*. Infection is spread *via* airborne dissemination of aerosolized bacteria containing droplet nuclei of 1–5μm in diameter that carry *M. tuberculosis* from an individual with infectious TB disease to an uninfected individual. The infectious droplet nuclei are inhaled and lodge in the alveoli in the distal airways. *M tuberculosis* is then taken up by alveolar macrophages, initiating a cascade of events that result in either successful containment of the infection or progression to active disease (primary progressive TB). Risk of development of active disease varies according to time since infection, age, and host immunity, however, the life- time risk of disease for a newly infected young child has been estimated at 10% [28]. *M. tb*'s unique cell wall, which has a waxy coating primarily composed of mycolic acids, allows the bacillus to lie dormant for many years. Once active, TB attacks the respiratory system and other organs, destroying body tissue.

The disease is contagious, spreading through the air by coughing, sneezing, or even talking. Transmission of TB occurs primarily by the aerosol route but can also occur through the gastrointestinal tract. Coughing by people with active TB produces droplet nuclei containing infectious organisms, which can remain suspended in the air for several hours. Infection occurs if inhalation of these droplets results in the organism reaching the alveoli of the lungs. Only 10% of immunocompetent people infected with *M. tuberculosis* develop active disease in their lifetime - the other 90% do not become ill and cannot transmit the organism. However, in some groups such as infants or the immunodeficient (*e.g.* those with AIDS or malnutrition) the proportion WHO develop clinical TB is much higher. The characteristic features of the tubercle bacillus include its slow growth, dormancy, complex cell envelope, intracellular pathogenesis and genetic homogeneity. The cell envelope of *Mycobacterium tuberculosis*, a Gram-positive bacterium with a G* C-rich genome, contains an additional layer beyond the peptidoglycan that is exceptionally rich in unusual lipids, glycolipids and polysaccharides. Novel biosynthetic pathways generate cell wall components such as mycolic acids, mycocerosic acid, phenolthiocerol, lipoarabinomannan and arabinogalactan, and several of these may contribute to mycobacterial longevity, trigger inflammatory host reactions and act in pathogenesis [29].

The generation time of *M. tuberculosis*, in synthetic medium or infected animals, is typically ~24 hours. This contributes to the chronic nature of the disease, imposes lengthy treatment regimens and represents a formidable obstacle for researchers. The state of dormancy in which the bacillus remains quiescent within infected tissue may reflect metabolic shutdown resulting from the action of a cell-mediated immune response that can contain but not eradicate the infection. As immunity wanes, through ageing or immune suppression, the dormant bacteria reactivate, causing an outbreak of disease often many decades after the initial infection [30].

TABLE 1: THE SUMMARY OF COMPLETE GENOME OF *MYCOBACTERIUM TUBERCULOSIS* H37RV

Refseq	NC_000962
GenBank	AL123456
Length	4,411,532 nt
GC Content	65%
Percentage of Coding	90%
Topology	circular
Molecule	DNA
Genes	4048
Protein coding	3989
Structural RNAs	50
Pseudo genes	8

The integration of principles from different disciplines like genomics, proteomics and

bioinformatics will help us in enhancing our knowledge about the TB bacilli and enable us to develop new therapies against the TB. In 1998, Cole *et al*., published the complete genome sequence of *Mycobacterium tuberculosis* [31] that is freely available in the NCBI GenBank, the accession number is NC_000962. The information from this paper was incorporated into the public database Tuberculist [32].

The *M. tuberculosis* genome comprises of 4,411,529 base pairs (bp) with a G + C content of 65.6% and 3,924 open reading frames accounting for 91% of the potential coding capacity. The genome is rich in repetitive DNA, particularly insertion sequences, and in new multigene families and duplicated housekeeping genes. The G + C content is relatively constant throughout the genome indicating that horizontally transferred pathogenicity islands of atypical base composition are probably absent. Several regions showing higher than average G + C content are present; these correspond to sequences belonging to a large gene family that includes the polymorphic G +C-rich sequences (PGRSs). Fifty genes code for functional RNA molecules. By using various database comparisons, precise functions have been attributed to ~40% of the predicted proteins were attributed and found some information or similarity for another 44% have been found.

The remaining 16% resembled no known proteins and may account for specific mycobacterial functions. *M. tuberculosis* differs radically from other bacteria in that a very large portion of its coding capacity is devoted to the production of enzymes involved in lipogenesis and lipolysis, and to two new families of glycine-rich proteins with a repetitive structure that may represent a source of antigenic variation.

In 2002, a re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv was published [33]. This group manually reevaluated each of the coding sequences (CDS) previously annotated and presented the combined results of recent database searches and literature surveys and made new comparisons with the genome sequence of *Mycobacterium leprae* [34]. The re-annotation of *M. tuberculosis* H37Rv has incorporated many changes to the functional classifications of the predicted proteins. A comparison of the number of predicted proteins in each of the functional categories between 1998 and 2002 is shown in the table below. Among several improvements, an important change due to the re-annotation was that the number of unknown proteins decreased from 606 to 272.

From the point of complete genome for *M. tuberculosis* was completed the pace of research on tuberculosis is dramatically increased and improved our knowledge about tuberculosis. The complete genome has opened up a new door for exploring the bacteria in more intensive way which has given us access into the drug targets, molecular biology, genomics, gene expression by microarray and proteomics studies, regulated genes, structural bioinformatics, and many more. Due to the extensive approach of research on tuberculosis heavy amount of data about tuberculosis has been generated. This has opened up establishing several databases for tuberculosis. These databases are commonly found to be useful for acquiring data about the genes and proteins.

TABLE 2: COMPLETE GENOME ANNOTATION INFORMATION FROM COLE *et al*., 1998 AND JEAN-CHRISTOPHE CAMUS *et al*., 2002

Class	Function	Number of new genes	Gene no. (1998)	Gene no. (2002)
0	Virulence, detoxification, adaptation	1	91	99
1	Lipid metabolism	1	225	233
2	Information pathways	3	207	229
3	Cell-wall and cell processes	11	516	708
4	Stable RNAs	0	50	50
5	Insertion sequences and phages	6	137	149
6	PE and PPE proteins	2	167	170
7	Intermediary metabolism and respiration	6	877	894
8	Proteins of unknown function	14	606	272
9	Regulatory proteins	3	188	189
10	Conserved hypothetical proteins	35	910	1051

Complete genome sequences are important source for any organism to understand the basic principles necessary to make an organism. Such understanding would provide us to have clear insight into the pathogenesis of infectious diseases too. However comprehensive analysis of entire genomes is required to understand what the genome codes for and how the genes interact and carry out complex and coordinated cellular functions. Among many approaches, the comparative genomics would potentially help to identify drug targets, specific markers and to develop diagnostic methods.

To enable the comparative genomics, the data generated for any pathogens have to be incorporated and integrated for holistic analysis. Bioinformatics provides us such resource for many diseases and in particular for tuberculosis. However, various Bioinformatics resources are available in the net for researchers for different niche goals. These resources have to be understood in such a way that they can be properly utilized for research that can help us to solve research questions. Therefore, databases for comparative genomics of *Mycobacterium tuberculosis* are discussed here in this report.

Mycobacterial Genome Divergence Database (MGDD): MGDD (Mycobacterial Genome Divergence Database) is a repository of genetic differences among different strains and species of organisms belonging to *Mycobacterium tuberculosis* complex. The results from a specific region (based on boundary defined by nucleotide sequence) or a specific gene can be displayed based on user's choice. Presently, the database has precomputed analysis from three fully sequenced genomes of this complex. These are *Mycobacterium tuberculosis* H37Rv, *Mycobacterium tuberculosis* CDC1551 and *Mycobacterium bovis* AF2122/97. It has the scope for updating to include more strains as fully sequenced genomes become available. MGDD is a free web-based database that allows quick user-friendly search to find different types of genomic variations among *M. tuberculosis* complex. Different types of variations that can be searched are SNP, indels, tandem repeats and divergent regions. The searches can be designed to find specific variations either in a given gene or any given location of the query genome with respect to any other genome currently available in the database.

TABLE 3: DATA COMPOSITION OF MGDD

Type of Data	Total Number of Entries
Divergence	829
Insertions	7865
Repeat expansion	578
SNP	68768

Biological Prespective:

Display and Search Option:

Single Nucleotide Polymorphism: A single-nucleotide polymorphism is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome (or other shared sequence) differs between members of a species. Single-nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. After submission of the selected information a detailed query page appears that contains option for choosing one of the 20 different possible transitions in a user-defined menu-bar and the search can be restricted by specifying genomic coordinates or gene name. The output would show all the indicated SNP in the selected region along with annotation of genes that contain the SNP.

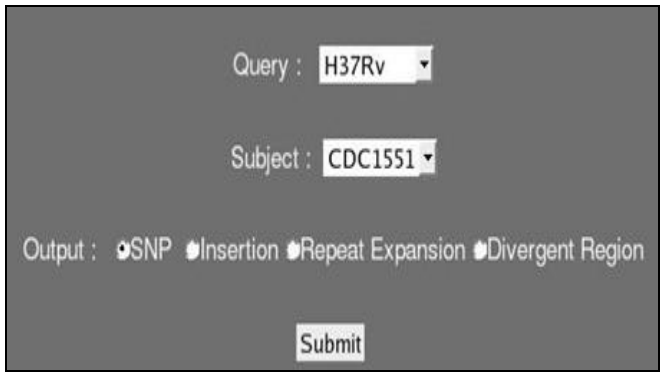


FIG. 1: THIS SHOWS QUERY WITH *MYCOBACTERIUM TUBERCULOSIS* H37RV AND CDC1551

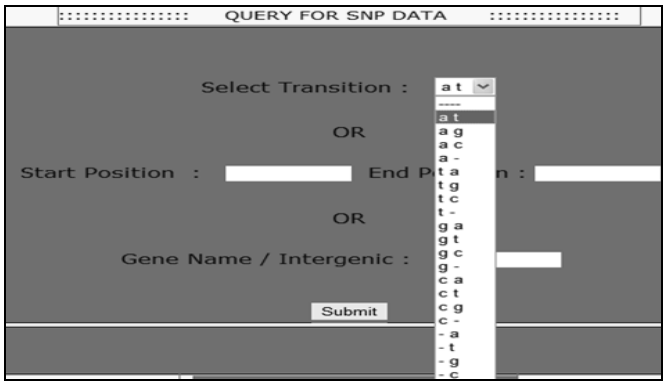


FIG. 2: DIFFERENT POSSIBILITIES OF TRANSITION

Transition Query Result for :: t g

SNP Position	Transition	Gene	Function
338101 t g	MRA_0288	PE-PGRS family protein	
338104 t g	MRA_0288	PE-PGRS family protein	
468948 t g	MRA_0395	PPE family protein	
1092859 t g	MRA_0984	PE-PGRS family protein	
1092890 t g	MRA_0984	PE-PGRS family protein	
1192695 t g	MRA_1078	PE-PGRS family protein	
1619491 t g	intergenic	------	

FIG. 3: THIS RESULT SHOWS THE SNP POSITION AS INTERGENIC REGION, TRANSITION REGION, GENE NAME AND WHICH FAMILY IT BELONGS TO

Insertion: Insertions can be either due to insertion of a stretch of nucleotides or due to increase in number of copies of repetitive elements. The information of such repetitive elements is present in the database. The database can be search in the following ways. By specifying a range of sizes, defining the start and end position of the region of interest of the query genome or entering either the accession number of gene.

TABLE 4: INSERTION PRESENT IN *M. TUBERCULOSIS* CDC1551 COMPARED TO *M. TUBERCULOSIS* H37RV

Position	Size	Gene	Function
79486	8	MT0077	Hypothetical protein
125821	1	MT0116	Cation- transporting ATPase, E1-E2 family
131166	1	intergenic	-
191560	1	intergenic	-
234609	1	intergenic	-
293743	1	intergenic	-
453339	1	intergenic	-
467587	1	MT0400	PPE family protein
467599	1	MT0400	PPE family protein
584324	1	MT0512	Hypothetical protein
635038	1	MT0566	Hypothetical protein

Repeat Expansion: Trinucleotide repeat expansion, also known as triplet repeat expansion, is the DNA mutation responsible for causing any type of disorder categorized as a trinucleotide repeat disorder. These are labeled in dynamical genetics as dynamic mutations. Triplet expansion is caused by slippage during DNA replication. Due to the repetitive nature of the DNA sequence in these regions, loop out structures may form during DNA replication while maintaining complementary base paring between the parent strand and daughter strand being synthesized. If the loop out structure is formed from sequence on the daughter strand it will result in an increase in the number of repeats. However if the loop out structure is formed on the parent strand a decrease in the number of repeats occurs.

The database can be searched in three different ways .By specifying a range of sizes, defining the start and end position of the region of interest of the query genome or entering either the accession number of gene.

Divergence: Divergent Regions are the domains that have diverged in terms of nucleotide sequences. These can be found in three different search ways. By specifying a range of sizes, defining the start and end position of the region of interest of the query genome or entering either the accession number of gene.

Under Construction:

Cilinical Isolates and the KZN Tb Strains: Three strains of *M. tuberculosis* isolated from patients in Kwa Zulu - Natal, South Africa have been sequenced using both Solexa and Sanger sequencing technology. The raw sequence data for the unassembled Solexa sequence runs are now available from the Broad Institute web site and NCBI. These three strains represent a range of important drug resistance phenotypes spanning fully drug-sensitive (DS) to multiply drug resistant (MDR), to extensively drug resistant (XDR). Recently, a high mortality rate for patients infected with XDR TB was reported in the KwaZulu-Natal region (Gandhi *et al*., 2006, Lancet (368): 1575-1580) (list these reference). These three strains are reported to be incorporated in the MGDD in the future for analysis.

Informatics Prespective:

Application Layer: MGDD is implemented by using three - tier architecture. The web based application is created by using Apache web server which is connected to the database using MYSQL through an application layer written in Perl-CGI.

Significance: A detailed analysis of these genomic sequences can help us to decipher and establish genotype to phenotype relationship. Characterization of sequence alterations in closely related organisms can help us to understand genome evolution at the molecular level in short time span, for example emergence of new endemic strains in a few decades. Thus the MGDD provides a catalog of all the sequence differences *Mycobacterium tuberculosis* complex that will be useful for the better understanding of tuberculosis.

BioHealth Base: The Biodefense and Public Health Database (BioHealth Base) Bioinformatics Resource Center (BRC) provides a comprehensive genomic and proteomic data repository for five pathogenic organism groups for the purpose of public health. It provides an analysis platform and appropriate tools to facilitate genomic and proteomic study of these pathogens. The goal of the BioHealth Base is to provide a resource to the scientific research community to facilitate the development of vaccines, diagnostics and therapeutics for these pathogens. It provides tools for comparative genomics studies on tuberculosis strains. By connecting the genomics studies it really targeting the unique genes that present in the virulent type. It majorly concentrates on the identification of drug targets.

Genomes in Biohealth Base:

TABLE 5: BIOHEALTHBASE PROVIDES FOR FIVE IMPORTANT MICROORGANISMS ARE GIVEN

Organism	Kingdom	Strain
Francisella tulresis	Bacteria	2
Mycobacterium tuberculosis	Bacteria	2
Mycobacterium avium	Bacteria	1
Mycobacterium leprae	Bacteria	1
Mycobacterium bovis	Bacteria	1
Encephalitozoon cuniculi	Fungi	1
Influenza A virus	Virus	7211
Influenza B virus	Virus	1127
Influenza C virus	Virus	152

Display Options: There are several options are given in the database for search through gene search, locus identifier, public database identifier, gene ontology, domain, motif, protein localization, protein structure, ortholog, epitope, and non-protein feature. The screenshot for the forms of gene search and locus search given in the **Fig. 4** and **5** respectively. This database provides information on various aspects like the

- operon prediction WUSTL
- B cell epitopes and T epitopes curated by IEDB
- Protein domains
- Motifs formed by interproscan algorithm and MHC bound surface epitopes,
- coexpression based operon prediction
- PSIPRED scores

Top 5 hits from swissprot protein database and coding sequences.

The operon predicton tool, user can easily access about the operon, gene details, genome location, annotation of that operon where the coding sequences that start, end, length as well as protein sequences. Operon identification through locus tag, evidence code and data sources of that operon. An important feature of this WUSTL tool as gene coexpresses with operon as gives the co expression coefficient and biological process involved and rank list of the expression genes.

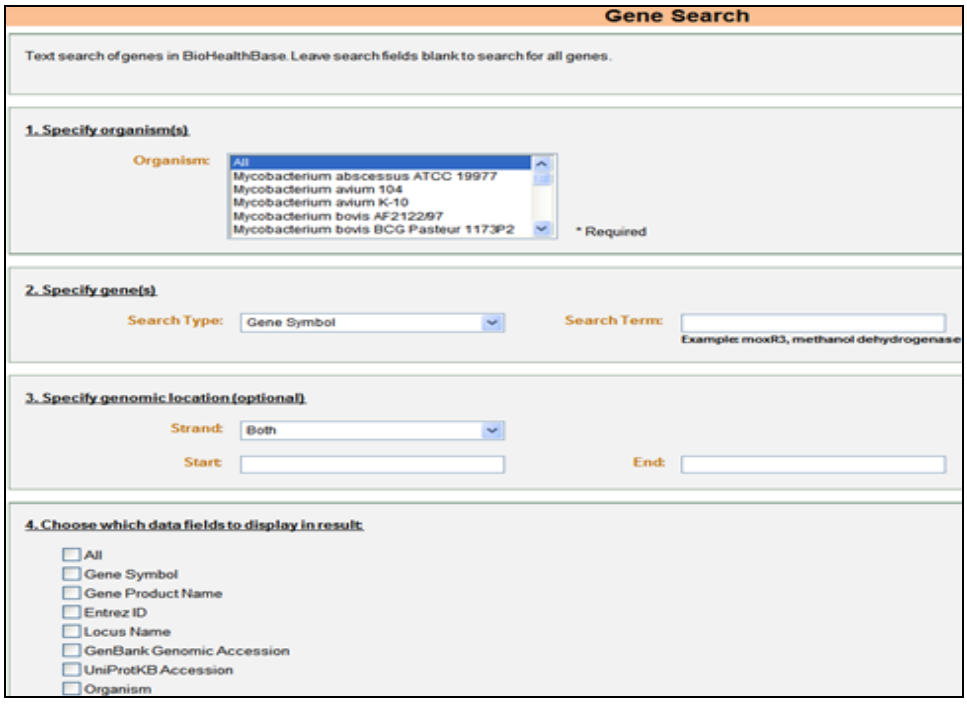


FIG. 4: FORM FOR GENE SEARCH

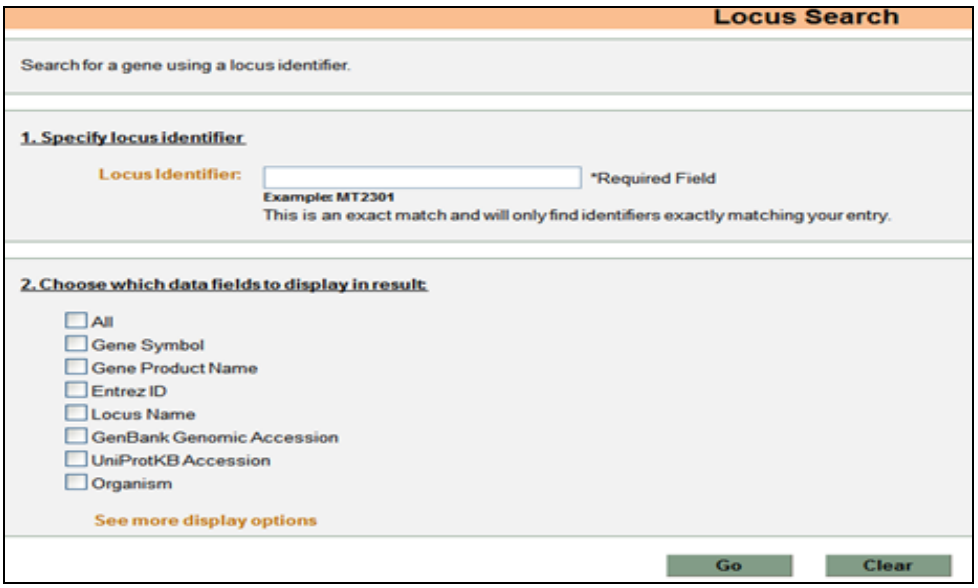


FIG. 5: FORM FOR LOCUS SEARCH

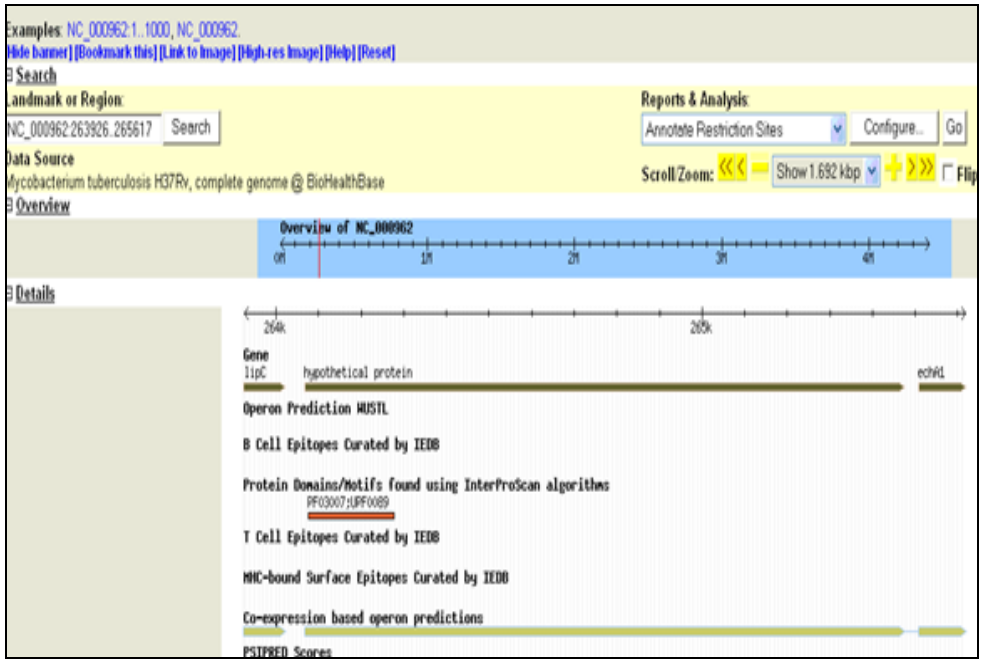


FIG. 6: THIS SHOWS THE GENOME BROWSER TOOL

The protein annotation gives as information of that gene, isoeletric point, protein localization whether as extra celluar, periplasmic, transmembrane through the program as subloc, signalp, PSORTb, DAS, and WOLPSORT. In protein structure search, it specifically gives about the display options about the molecule change the modelling, spinning, highlighting the ligand molecules, residues by chain type, or residue type. In the epitope search, as curated epitopes derived from the immune epitope database and analysis resource and predicted epitopes are derived based on the sequence similarity to the curated IEDB epitope sequence within an ortholog group.

TABLE 6: ALGORITHMS AND ITS DATA TYPES

Algorithm	Data type
NCBI BLASTP	Sequences similarities
PSIPRED	Predicted protein secondary structure
Psortb, signalp, subloc, DAS, wolf-psort	Protein subcellular localization predictions
Ptools, wustl	Operon predictions
Biocyc pathway tools	Metabolic pathways
Glimmer	ORF predictions
Interproscan	Domains/motifs
OrthoMCL	Ortholg predictions
Isoeletric point and molecular weight	Isoeletric point
Lipoprotein	Predictions

Tools:

Blast Search: Retrieve an alignment search or by new search. If user entered through retrieve alignment search it asked for ticket number where we added to the workbench option or collect the details through new search option.

Whole Genome Comparison: Whole genome alignment programs are used for comparative genome analysis to learn about the large-scale structural similarities and differences between genomes and to aid in the detection of complex genomic changes such as rearrangements, duplications and gene insertions or deletions. Users can begin by choosing a reference genome and comparison genomes provided in the drop-down lists, then select the program of choice (MUMmer, NUCmer or PROmer) and hit the go button. The program generates a dot-plot graph of the alignment. This can be saved as a .pdf or be added to the workbench for further analysis.

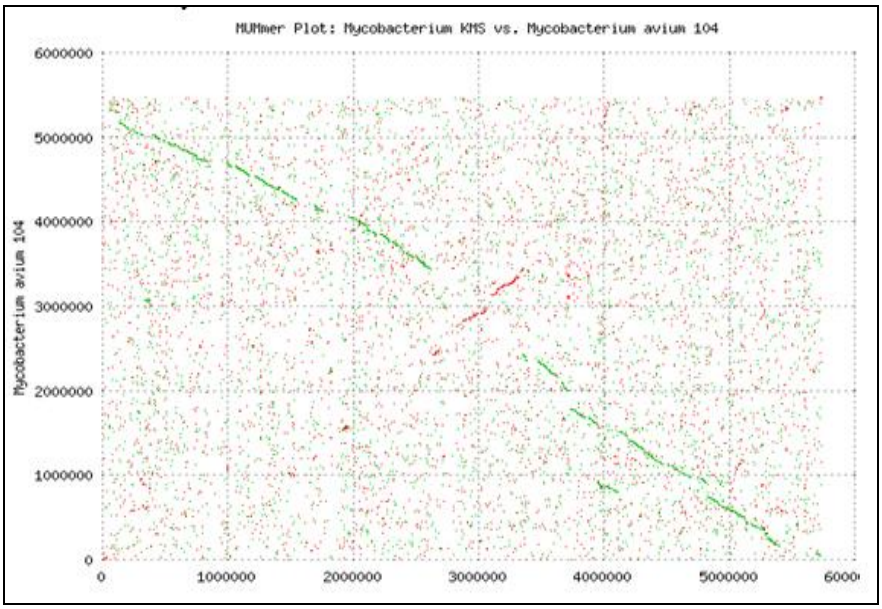


FIG. 7: THIS SHOWS THE MUMMER PLOT BETWEEN TWO DIFFERENT SPECIES

In dot-plot, the reference genome is along the X-axis, while the query genome is on the Y-axis. Wherever the two sequences match, a colored dot is plotted. The forward matches are displayed in red, while the reverse matches are displayed in green Mummer can also align incomplete genomes; it can easily handle the 100s or 1000s of contigs from a shotgun sequencing project, and will align them to another set of contigs or a genome using the NUCmer program included with the system. If the species are too divergent for a DNA sequence alignment to detect similarity, then the PROmer program can generate alignments based upon the six-frame translations of both input sequences.

Highlights of Mummer Tool:

- Open source
- Improved efficiency
- Ability to find non-unique, repetitive matches as well as unique matches
- New graphical output modules

Other Tools:

- **Insignia:** A web service for the identification of DNA signatures suitable for real-time pathogen detection assays.
- **AMOS:** A genome assembly toolkit including AMOScmp, a comparative genome assembler built with MUMmer.
- **Synteny Miner:** A visualization tool for interrogation of multiple whole genome alignments.
- **Tandemizer:** A visualization tool for the analysis of tandem array blocks across multiple genomes.

Comparative Genomics Viewer: Synteny refers to the occurrence of two or more genes on the same chromosome (or same region of genome) in different organisms. The Synteny analysis tool

converts homology data into visually interpretable gene alignments, allowing users to study relationship between strains, complex genomic changes such as rearrangements, genome duplication, gene insertions and deletions and study of evolution of genomes. Clicking on any gene graph in the current Synteny View, one can choose to:

- Center-align all its orthologs with color coding
- Display the operon the gene belongs to and all operons of its orthologs.
- View all details about the gene or protein.
- View the alignment of all orthologous proteins in an interactive viewer.

Metabolic Pathway Tools: The initial datasets of pathways at BHB were downloaded from BioCyc, which were generated at SRI by comparing the annotations to a reference database (MetaCyc). A class hierarchy allows users to retrieve information according to categories of interest. Its subclasses divide pathways into groups based on their biological functions, and based on the classes of metabolites that they produce and/or consume. In pathway tools query page, it opens a form provides several different mechanism for querying pathway, genome browser, it shows chromosome segment with every gene present in that, start, end codons and edit the tracks of segments. This query results shows all pathways of *e.g. Mycobacterium tuberculosis* H37Rv totally it has 234 pathways.

Other Data Uniquely Linked to This Database:

***Mycobacterial tuberculosis* H37RV Essential Genes:** Essential Genes are genes that are vital to sustain life and cellular growth in an organism. The reason for genes to be 'essential for growth and existence' depends on various factors such as the type of organism, the environmental conditions related to its way of life and the genetic factors. The study of essential genes provides an insight to their metabolic function in the organism and the various pathways in which they could be involved. Different methodologies are used to determine the essentiality of genes, such as generating a library of mutants to recognize strains with growth defect etc. Genes in the table below are collected from Database of Essential Genes. These genes are found to be essential for the growth and survival of *Mycobacterium tuberculosis* based on the high-density mutagenesis work carried out by Sassetti *et al*.,

Drug Information: Drug information specifically for tuberculosis provided with generic name, chemical formulae, molecular weight, and structural therapeutic indication.

Database Cross Reference:

- INTERPRO
- PFAM
- PIR
- TBDB
- TIGR
- TIGRFAMS
- TUBERCULIST
- METCYC
- BIOCYC
- GENBANK
- UNIPROT

GenoMycDB: GenoMycDB (http://www.dbbm.fiocruz.br/GenoMycDB) is a relational database built for large scale comparative analyses of completely sequenced mycobacterial genomes, based on their predicted protein content. Six *Mycobacterial* species are included in this database. They are:

- *Mycobacterium tuberculosis* H37Rv
- *Mycobacterium tuberculosis* CDC1551
- *M. bovis* AF2122/97
- *M. avium* subsp. paratuberculosis K10
- *M. leprae* TN and
- *M. smegmatis* MC2 155.

The structure of the DB is composed of the results obtained after pair-wise sequence alignments among all the predicted proteins coded by these genomes. For each of the gene, the following information can be obtained from this database:

Searches can be restricted according to the predicted subcellular localization of the protein, the DNA strand of the corresponding gene and/or the description of the protein. Massive data search and/or retrieval are made available. GenoMycDB provides an on-line resource for the functional classification of mycobacterial proteins as well as for the analysis of genome structure, organization, and evolution.

TABLE 7: DISPLAY OPTION AND ITS DESCRIPTION

Display Option	Description
QSpecies	Query species
QName	Query name
QDesc	Query description
QLen	Query length
QGQBank	GenBank identifying number of the query sequence
QSProt	SwissProt/TrEMBL identifying number of the query sequence
QKEGG	KEGG identifying number of the query sequence
QPDB	PDB identifying number of the query sequence
QPSbLocal	PSORTb subcellular prediction of the query sequence
QPSbScore	PSORTb subcellular prediction score of the query sequence
QMycName	GenoMycDB derivative name of the query sequence
QGene	Name of the query sequence gene
QGSynonym	Synonym of the query sequence gene
QGStart	Start position of the query sequence gene in the genome
QGEnd	End position of the query sequence gene in the genome
QGStrand	DNA strand where the query sequence gene is located
QGProduct	Description of the query protein sequence
QGCOG	Protein query sequence assigned COG(s)
HSpecies	Hit species
HName	Hit name
HDesc	Hit description
HLen	Hit length
HGBank	GenBank identifying number of the hit sequence
HSProt	SwissProt/TrEMBL identifying number of the hit sequence
HKEGG	KEGG identifying number of the hit sequence
HPDB	PDB identifying number of the hit sequence
HPSbLocal	PSORTb subcellular prediction of the hit sequence
HPSbScore	PSORTb subcellular prediction score of the hit sequence
HMycName	GenoMycDB derivative name of the hit sequence
HGene	Name of the hit sequence gene
HGSynonym	Synonym of the hit sequence gene
HGStart	Start position of the hit sequence gene in the genome
HGEnd	End position of the hit sequence gene in the genome
HGStrand	DNA strand where the hit sequence gene is located
HGProduct	Description of the hit protein sequence
HGCOG	Protein hit sequence assigned COG(s)
HIdent(%)	Overall fraction of identical positions across all HSPs (aligned regions only)
HPos(%)	Overall fraction of conserved positions across all HSPs (aligned regions only)
HAlnQuery(%)	Fraction of the query sequence which has been aligned across all HSPs (not including intervals between non-overlapping HSPs)
HAlnHit(%)	Fraction of the hit sequence which has been aligned across all HSPs (not including intervals between non-overlapping HSPs)
Score	Raw score
Bits	Bit score
Evalue	Expect value for the HSP (e-value)
Ident	Number of identical residues
Ident(%)	Fraction of identical positions for a given HSP
Pos	Number of conserved residues
Pos(%)	Fraction of conserved positions for a given HSP
QGaps	Number of gaps in the query alignment
HGaps	Number of gaps in the hit alignment
HSPLen	Length of HSP (full length of the alignment)
QOverlap	Length of query participating in alignment minus gaps
SSHOverlap	Length of hit participating in alignment minus gaps
AlnQuery(%)	Fraction of the query sequence which has been aligned within a given HSP
AlnHit(%)	Fraction of the hit sequence which has been aligned within a given HSP
QStart	Query start position from the alignment
QEnd	Query end position from the alignment
HStart	Hit start position from the alignment
HEnd	Hit end position from the alignment

FIG. 8: THE FORM FOR SEARCHING THE GENOMYCDB BROWSER

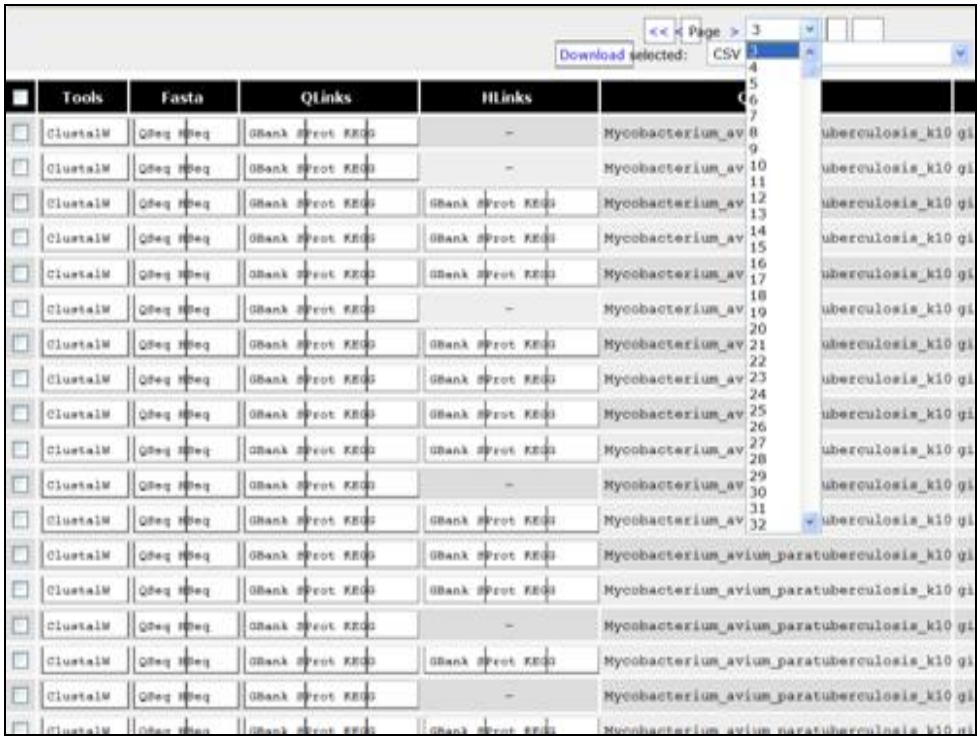


FIG. 9: THIS SHOWS THE OUTPUT OF THE COMPARATIVE GENOMICS BETWEEN *M. TUBERCULOSIS* H37RV AND *M. LEPRAE* TN IS GIVEN HERE. KATG WAS USED AS A QUERY TERM. ALL THE DISPLAY OPTIONS WERE SELECTED IN THE FORM FOR THIS QUERY

The GenoMycDB offers a flexible, scalable, functional, cross-referenced, and user-friendly system for the comparative genomic analyses of representatives of the genus *Mycobacterium*. Furthermore, the same structure and database interface can easily be applied to other groups of genomes, extending the potential of our system.

Currently, the developers of the GenoMycDB are using this database to study the nucleotide evolutionary rates among protein-coding regions of mycobacteria, to analyze point mutations and polymorphisms among selected protein-coding regions of *M. tuberculosis* complex species, and to investigate the factors shaping codon usage in

mycobacteria. In addition, the database is also being used by them to annotate the genome of BCG Moreau, a vaccine strain derived from *M.bovis* used to prevent tuberculosis in the Brazilian population. Therefore, GenoMycDB provides a valuable tool for the comparative analyses of mycobacterial genomes, making it possible to identify evolutionary, structural, and functional relationships between proteins in such genomes.

***Mycobacterium tuberculosis* Database:** The Microbial Sequencing Center at the Broad Institute is sequencing 8 strains of *Mycobacterium tuberculosis* towards the ends of better understanding, treating, and ultimately eradicating this deadly human pathogen. The genome sequences of several strains of *M. tuberculosis* and several related Mycobacterium species have provided valuable insights into the biology of this organism (1 - 4). This project is focused on understanding the molecular basis of fundamental clinical phenotypes such as TB transmission and multidrug-resistance. The sequencing and comparative analysis of patient isolates with the most important clinical phenotypes and disease epidemiology would help us to understand the varied degree of spread, drug resistance, and clinical severity of tuberculosis. The major aim of the project is to sequence and analyze the drug resistant stains of *M. tuberculosis*.

Three strains of *M. tuberculosis* isolated from patients in KwaZulu-Natal, South Africa have been sequenced using both Solexa and Sanger sequencing technology. The raw sequence data for the unassembled Solexa sequence runs are now available from the Broad Institute web site and NCBI. Draft genome assemblies for these 3 strains can be downloaded from this site.

These three strains were selected because they represent a range of important drug resistance phenotypes spanning fully drug-sensitive (DS) to multiply drug resistant (MDR), to extensively drug resistant (XDR). XDR TB is extremely difficult to treat because it is resistant to the first line anti-TB drugs INH and rifampin, and it is resistant to at least one fluoroquinolone plus at least one second line injectable drug (capreomycin, kanamycin, or amikacin). Recently, a high mortality rate for patients infected with XDR TB was reported in the KwaZulu-Natal region. The XDR (KZN 605), MDR (KZN 1435), and DS (KZN 4207) strains were selected for sequencing from among other strains in the KZN region. The three strains represent three levels of drug resistance:

- KZN 4207: drug sensitive
- KZN 1435: multiple drug resistant
- KZN 605: extensively drug resistant

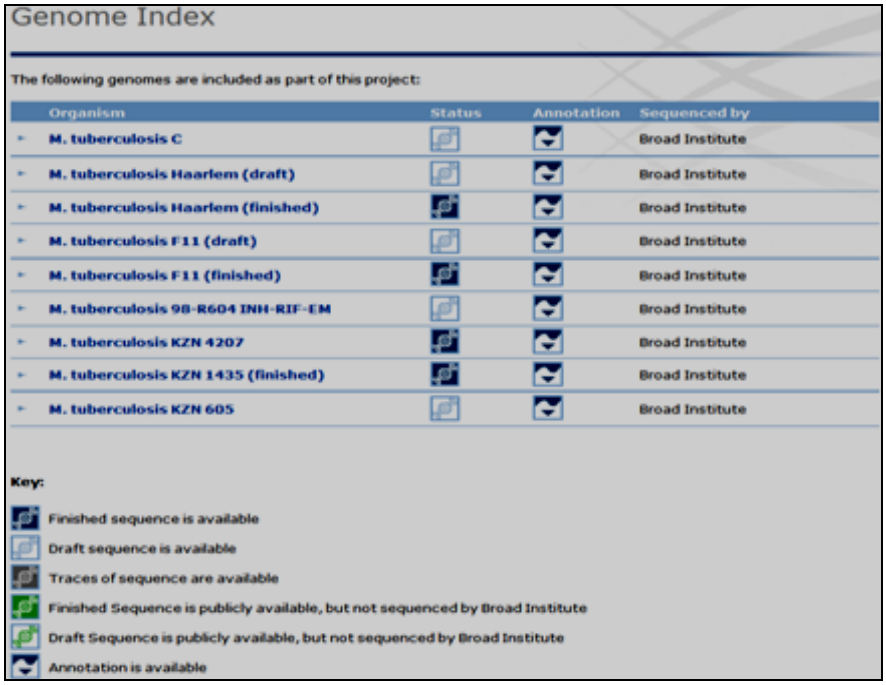


FIG. 10: LIST OF GENOMES OF *M. TUBERCULOSIS* STRAINS AVAILABLE IN THE BROAD INSTITUTE DATABASE

This database contains several search options and tools for comparative analysis of M. tuberculosis. Most of the features available in this database are incorporated in the TBDB by collaboration between Stanford School of Medicine and Broad Institute. Therefore, the tools and features of this database are discussed along with the TBDB (TB DATABASE).

TB Database: An Integrated Platform for Tuberculosis Research: The Tuberculosis Database (TBDB) is an integrated database providing access to TB genomic data and resources, relevant to the discovery and development of TB drugs, vaccines and biomarkers. The current release of TBDB houses genome sequence data and annotations for 28 different *Mycobacterium tuberculosis* strains and related bacteria. TBDB stores pre- and post publication gene expression data from *M. tuberculosis* and its close relatives. TBDB currently hosts data for nearly 1500 public tuberculosis microarrays and 260 arrays for Streptomyces. In addition, TBDB provides access to a suite of comparative genomics and microarray analysis software. By bringing together *M. tuberculosis* genome annotation and gene expression data, comparative genome annotation with a suite of analysis tools.

Overview of Tbdb: It has led to the use of genome wide expression profiling and comparative genomics methods to better understand *M. tuberculosis* pathology, latency, emerging drug resistance and evolution. However, despite the wide-spread use of functional and comparative genomics to study *M. tuberculosis*, there has been no single repository for these large scale datasets, complete with high quality experimental annotation, and connected to upto date gene annotation and comparative genomic information.

This database brings together powerful genomics tools to advance *M. tuberculosis* research in ways that will contribute to the identification of new drug targets, vaccine antigens, and diagnostics and host biomarkers. This annotated genome sequence data and microarray and RT–PCR expression data from *in vitro* experiments and TB-infected tissues. TBDB houses genome sequence data for several *M. tuberculosis* strains as well as data for numerous related species. These data and annotations include publicly available sequences from a number of sequencing centres and groups, including sequences being produced by the Broad Institute's Microbial Sequencing Centre.

Summary of TBDB Data Content (as of September 2008):

TABLE 8: TBDB DATA STATISTICS

Number of genomes	28
Number of all microarrays	~5500
Number of public microarray	~1800
Number of publications	27
Number of experiment sets	160

TABLE 9: LIST OF ANNOTATED GENOMES IN THIS DATABASE THAT IS USED FOR STUDIES OF COMPARATIVE GENOMICS

Organism	Size (mb)	Genes
M. tuberculosis H37Rv	4.41	3999
M. tuberculosis CDC1551	4.4	4189
M. tb f11	4.42	3959
M. tb C	4.38	3851
M. tb haarlem	4.4	3866
M. bovis AF2122/97	4.35	3920
M. bovis BCG	4.37	3952
M. leprae TN	3.27	1605
M. avium 104	5.48	5120
M. avium k10	4.83	4350
M. smegmatis MC2 155	6.99	6716
M. marinum	6.64	5423
M. ulcerans Agy99	5.63	4160
M. vanbaalenii PYR-1	6.49	5979
M. sp. KMS	6.26	5975
M. sp. MCS	5.71	5391
Rhodococcus sp. RHA1	9.7	9145
Nocardia farcinica IFM 10152	6.02	5683
Corynebacterium glutamicum ATCC 13032	3.28	3057
C. diphtheriae NCTC 13129	2.49	2272
C. efficiens YS-314	3.15	2950
C. jeikeium K411	2.48	2120
Streptomyces avermitilis MA-4680	9.12	9.12
S. coelicolor A3(2)	8.67	7825
Propionibacterium acnes KPA171202	2.56	2297
Acidothermus cellulolyticus 11B	2.44	2157
Bifidobacterium longum NCC2705	2.26	1727

Biological Prespective: The database currently houses genome sequence data for M. *tuberculosis* strain H37Rv (a standard prototype strain long used for experimental and animal infection studies), as well as other *M. tuberculosis* strains and bacteria from related taxa, focusing on members of the Actinomycetes family of high G+C content, Gram-positive organisms of which M. *tuberculosis* is a

member. These genomes sequences have been annotated with a variety of genomic features including genes, operons, sequence similarity to GenBank sequences using BLAST, transfer RNAs using tRNA Scan, protein domains and families using PFAM and non coding RNAs based on RFAM.

TB Genomes Tools:

- **BLAST Search:** Find similarities to other sequences.
- **Feature Search:** Search and view annotated features on the sequence.
- **Browse Regions:** Retrieve DNA, find clones, and graphically view sequence regions.
- **Gene Index:** Find genes by a variety of methods.
- **Genome Statistics:** View basic statistics about genome size, gene density, *etc*.

TABLE 10: TB GENOMES FACILITIES

Gene finding method	Information on how this putative gene set was created
Search for genes	Search for genes Search database of all TB Genomes genes by name, locus, position, or PFAM domain
BLAST against protein set	Search predicted proteins using blastp
Download FASTA files and supplementary material	FASTA files for genome sequence, genes, translated proteins are available
Genes by PFAM domain	Genes categorized by HMMER hit to a PFAM protein domain
Genes for functional RNA by RFAM family	Features categorized by INFERNAL hit to an RFAM family
Operon	Searchable list of operon predictions
Genes indexed by functional annotation	KEGG: Kyoto Encyclopedia of Genes and Genomes COG: COG Functional Categories PWY: BioCyc Pathways EC: Enzyme Committee Nomenclature GO: Gene Ontology
Essential genes by TraSH	TraSH List Genes found to be essential in transposon site hybridization

Display Options:

- Genome visualization tool-Argo applet
- Synteny map
- Dot plot
- Operon browser
- Circular genome viewer
- Genome map

Comparative Analysis:

Genome Visualization Tool:

ARGO: The Argo Genome Browser (an interactive applet) and the Feature Map (a lighter weight version of the Argo Genome Browser) provide linear views of genome sequences along with all associated annotated features.

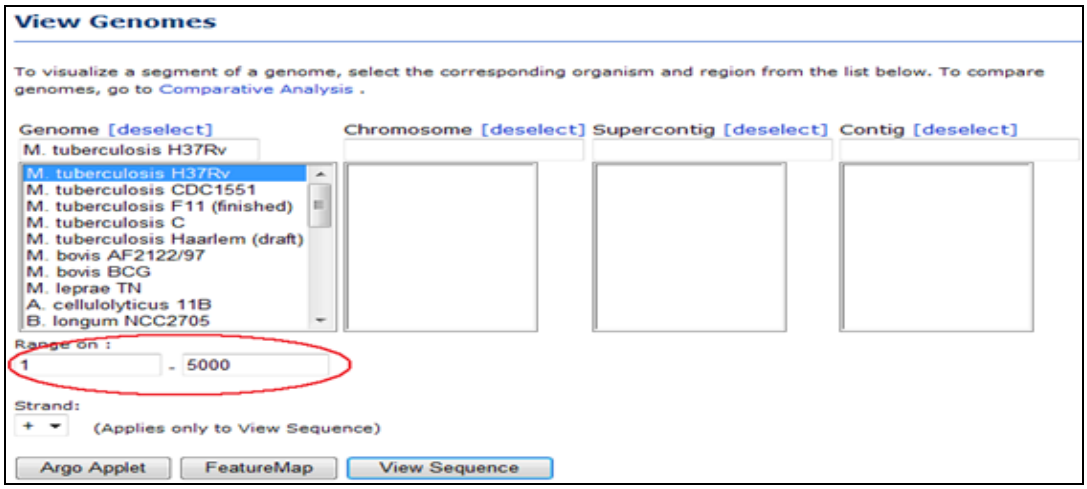


FIG. 11: THIS SHOWS SELECTION OF THE GENOME AND VISUALIZATION OF THE GENOME AND THE CORRESPONDING REGION

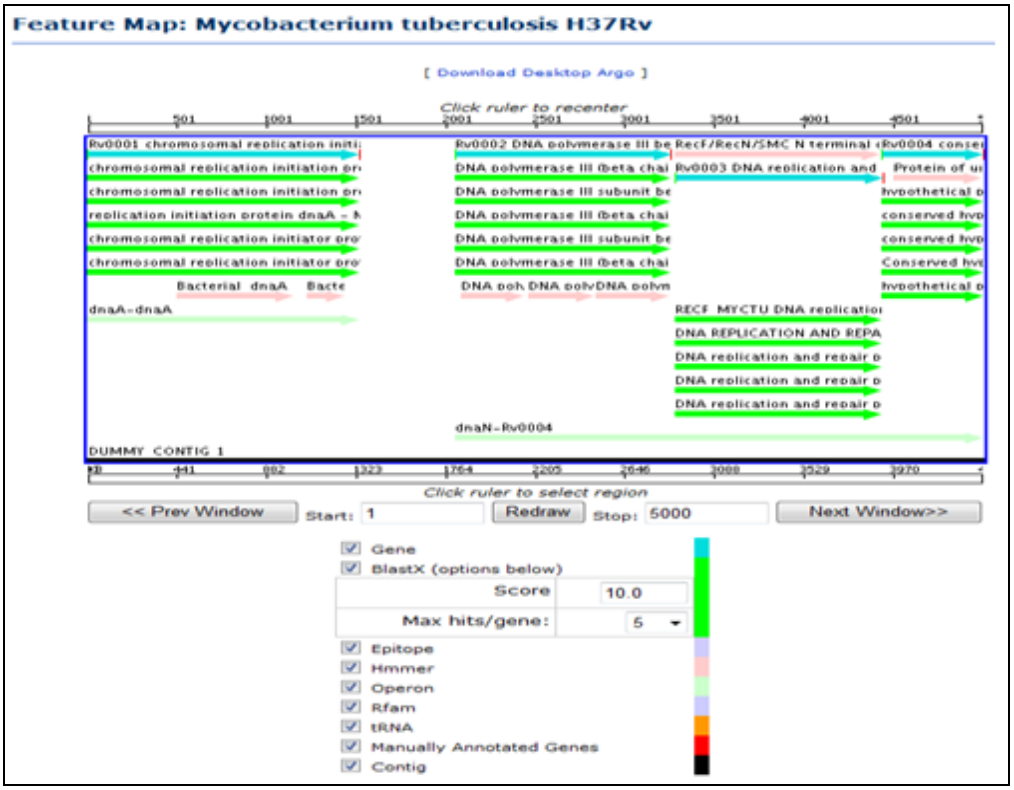


FIG. 12: THIS SHOWS THE ARGO BROWSER TOOL WITH THE SELECTED REGION

Argo in particular provides a dynamic interface to visualizing genome data that allows users to zoom from whole chromosomes to individual nucleotides, navigate within sequences, and select individual features to retrieve additional information. The Argo Genome Browser is the Broad Institute's production tool for visualizing and manually annotating whole genomes. It display of sequence and annotation tracks. Interactive zoom from megabase to nucleotide resolution. If user can edit of individual features, supporting manual annotation, the user can compare the chromosome and view the genome by selecting the region and by using the feature map.

Comparative Genome Visualization Tool: An additional number of tools are also provided specifically for comparative analyses between genome sequences, including the Synteny Map, Dot Plot, Operon Browser and Gene Family Search.

Circular Genome Viewer: It provides a circular plot of genome sequences along with a plot with, GC content and GC skew. Potential pathogenicity elements in *Mycobacterial* genomes can be identified by analyzing gene gain or loss through comparative genomics.

Syteny Map: The Synteny Map uses precompiled genome alignments to graphically display regions of genomic similarity between a single reference genome and one or more other genomes-in effect providing the results of *in silico* genome hybridization between sets of genomes. Using this tool, the user can select regions of interest and then click a region to zoom in and view genes, genome sequence, and features.

Dot Plot: The Dot Plot displays a navigable map of computed synteny between genomes in the form of dot-plot lines. When comparing multiple genomes, the colour of the plotted synteny indicates which genome is aligned to the reference at that position. A dot plot is a visually appealing, intuitive, but qualitative tool for comparison of sequences. A dot is placed only when the residues labelling the element are complementary to each other. A well-filtered dot plot of the comparison of portions of two bacterial genomes. The broken diagonal line indicates the large degree of similarity between the two. This plot shows in x-axis *Mycobacterium tuberculosis* H37Rv and in y axis *Mycobacterium bovis* BCG .In this plot user can select the segment of the genome and zoom it by navigating and selection by view by feature map also. The diagonal line shows the alignment shows

complementary to each other. Below the plot there is an search option if user can searching for purine metabolism genes it shows triangle shape and by drag the in that position of that gene and locus id also.

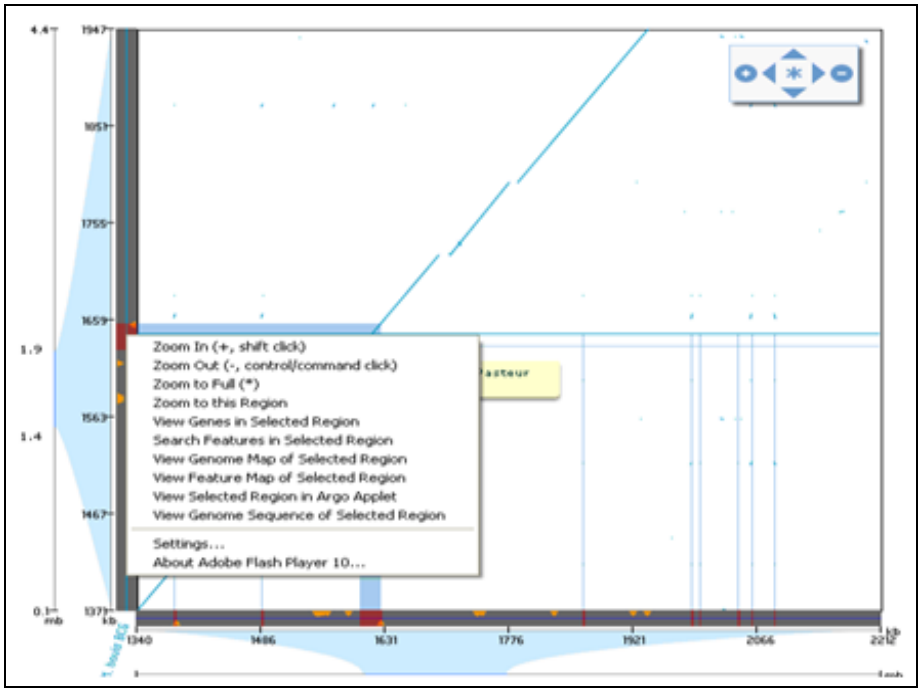


FIG. 13: THIS DOT - PLOT SHOWS THE SYNTENY BETWEEN TWO DIFFERENT GENOMES

Operon Browser: The Operon Browser is a tool that simultaneously displays the expression correlation between genes in a genomic region of the *M. tuberculosis* H37Rv strain while showing syntenic gene order of orthologs in related species. A heat map derived from expression correlation data is provided along with an alignment of syntenic areas. Mousing over the genes provides additional information such as locus ID, gene symbol and description. Color coding of genes indicate orthologous relationships across different species.

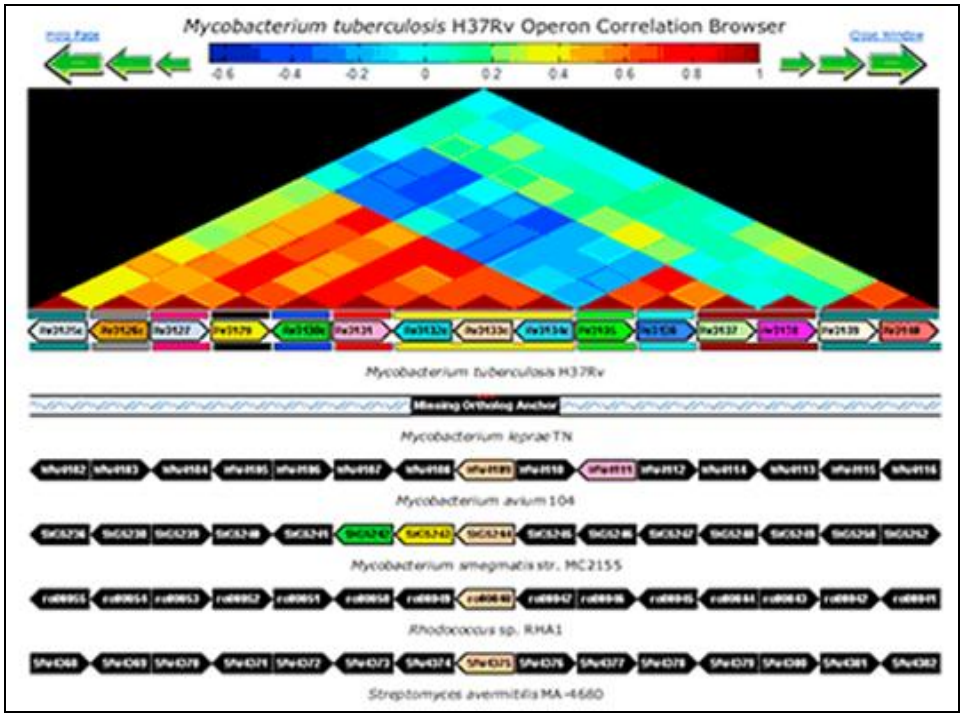
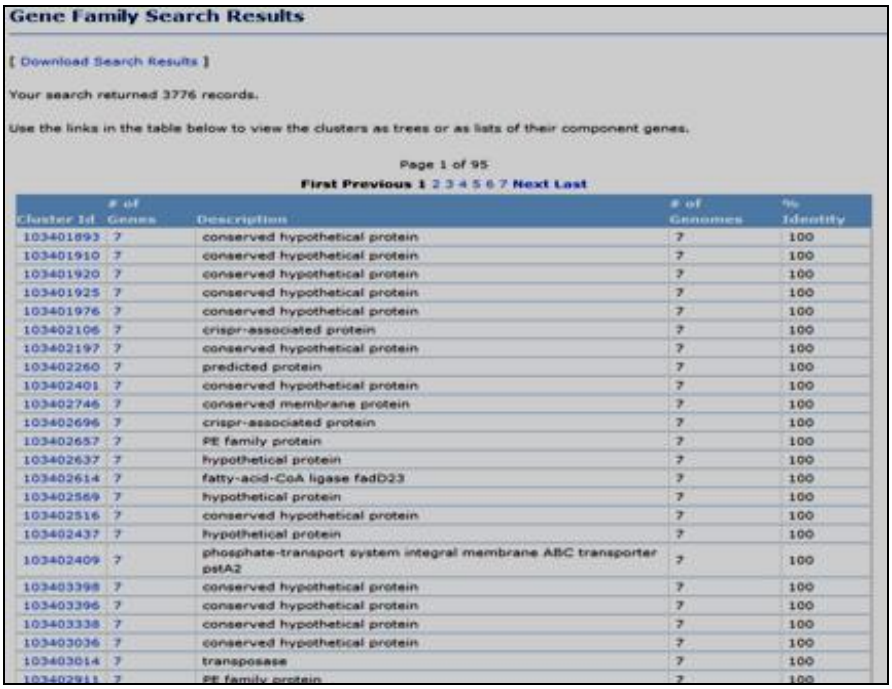


FIG. 14: THIS SHOWS THE OPERON CORRELATION BROWSER TOOL

Search Gene Families: The Gene Family Search displays phylogenetic trees and sequence alignments of predicted orthologous gene families within the genome sequences in TBDB.



Gene Family Search Results

[Download Search Results]

Your search returned 3776 records.

Use the links in the table below to view the clusters as trees or as lists of their component genes.

Page 1 of 95

First Previous 1 2 3 4 5 6 7 Next Last

Cluster Id	# of Genes	Description	# of Genomes	% Identity
103401893	7	conserved hypothetical protein	7	100
103401910	7	conserved hypothetical protein	7	100
103401920	7	conserved hypothetical protein	7	100
103401925	7	conserved hypothetical protein	7	100
103401976	7	conserved hypothetical protein	7	100
103402106	7	crispr-associated protein	7	100
103402197	7	conserved hypothetical protein	7	100
103402260	7	predicted protein	7	100
103402401	7	conserved hypothetical protein	7	100
103402746	7	conserved membrane protein	7	100
103402696	7	crispr-associated protein	7	100
103402657	7	PE family protein	7	100
103402637	7	hypothetical protein	7	100
103402614	7	fatty-acid-CoA ligase fadD23	7	100
103402569	7	hypothetical protein	7	100
103402516	7	conserved hypothetical protein	7	100
103402437	7	hypothetical protein	7	100
103402409	7	phosphate-transport system integral membrane ABC transporter pstA2	7	100
103403398	7	conserved hypothetical protein	7	100
103403396	7	conserved hypothetical protein	7	100
103403338	7	conserved hypothetical protein	7	100
103403036	7	conserved hypothetical protein	7	100
103403014	7	transposase	7	100
103402911	7	PE family protein	7	100

FIG. 15: THIS SHOWS THE LISTS OF COMPONENT GENES

Application Layer: Apache web server and HTML: which stands for Hypertext Markup Language. It provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists *etc.*, as well as for links, quotes, and other items.

TB Resources: The following resources of tuberculosis is connected from the TBDB

- Tuberculist
- Biocyc
- Biohealthbase

External Data Links:

dbEST: dbESTlib, dbEST.
LocusLink: locusLink, LL_Mult, LL_Set
RhDB: RhDB, rhLink.
SwissProt: swissProt, spLink, spDesc, spFeature.
UniGene: ugLIB, ugCluster, ugSTS, ugProtsim, ugTxMap.
Uniprot: uniprotid, clusterid, uniprot, locuslink.

TABLE 11: A LIST OF ONLINE RESOURCES FOR COMPARATIVE GENOMICS OF *MYCOBACTERIUM TUBERCULOSIS*

Database	Purpose	URL
MGDD	MGDD (Mycobacterial Genome Divergence Database) is a repository of genetic differences among different strains and species of organisms belonging to *Mycobacterium tuberculosis* complex.	http://mirna.jnu.ac.in/mgdd/
BioHealthBase	The BioHealth Base web-interface provides a tool for researchers to query, retrieve, visualize and download data, and to apply bioinformatics tools such as BLASTs and multiple sequence alignments to aid in their research for tuberculosis.	http://www.biohealthbase.org/GSearch/home.do?decorator=Mycobacterium
GenoMycDB	GenoMycDB is a relational database built for large-scale comparative analyses of completely sequenced mycobacterial genomes, based on their predicted protein content.	http://xbase.bham.ac.uk/mycodb/
Mycobacterium tuberculosis Database (Broad Institute)	This database contains 9 strains of *Mycobacterium tuberculosis* genome sequence including drug resistant strains.	http://broadinstitute.org/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html

TBDB	The current release of TBDB houses genome sequence data and annotations for 28 different *Mycobacterium tuberculosis* strains and related bacteria and TBDB stores pre- and post-publication gene-expression data from *M. tuberculosis* and its close relatives.	http://www.tbdb.org/

The TBDB maintains microarray and RT-PCR data as well as genomic data for *Mycobacterium tuberculosis* organism along with several other relevant host organisms and similar model organisms. Genome sequence, microarray or RT-PCR data from published experiments may be used for further research from this resource.

CONCULSION: The comparative genomics helps to compare and analyze genetic material from different species. Such comparison would through light on our understanding of the evolution. Further, it also helps us make functional interpretation of the genes of various genomes, to find or group protein families, to identify unique set of genes involved in a specific process that would reveal the mechanism of such process, to identify genes involved in the pathogenesis of infectious organisms and to identify markers for developing diagnostic methods. The significance of the comparative genomics is very felt with the help of the recent development in the bioinformatics. The bio-informatics have paved the way for development of database discussed in this report for comparative genomics and they would be useful for the further development in the field of tuberculosis.

ACKNOWLEDGMENTS: We thank the Regenix super speciality laboratories at Chennai for support in managing the facility and Dr. Subramaniam. S, Lab Director, for comments that greatly improved the manuscript.

CONFLICT OF INTEREST: The authors have no conflicts of interest.

REFERENCES:

1. Global Tuberculosis report 2017- World Health Organisation (2017, October). Tuberculosis Fact Sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
2. Global Tuberculosis report 2008- World Health Organisation (2008). Summary on Tuberculosis.
3. World health Organisation (2015). Fact files on Tuberculosis. Retrieved From http://www.who.int/features/factfiles/tb_facts/en/index.html
4. Palamino *et al.*: Tuberculosis 2007 book-From basic sciences to patient care. Flying Publisher 2007.
5. Frank G.J Cobelens *et al.*: Scaling Up Programmatic Management of Drug-Resistant Tuberculosis: A Prioritized Research Agenda. PLoS Med. 2008; 5(7): e150.
6. Global Tuberculosis report 2015- World Health Organisation (2015, October). Tuberculosis Fact Sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
7. Frank G.J Cobelens *et al.:* Scaling Up Programmatic Management of Drug-Resistant Tuberculosis: A Prioritized Research Agenda. PLoS Med. 2008; 5(7): e150.
8. Global Tuberculosis report 2015- World Health Organisation (2015, October). Tuberculosis Fact Sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
9. Global Tuberculosis report 2015- World Health Organisation (2015, October). Tuberculosis Fact Sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
10. Global Tuberculosis report 2015- World Health Organisation (2015, October). Tuberculosis Fact Sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
11. WHO: Guidelines for the programmatic management of drug-resistant tuberculosis Emergency update 2008.
12. WHO: Guidelines for the programmatic management of drug-resistant tuberculosis Emergency update 2008.
13. WHO: Guidelines for the programmatic management of drug-resistant tuberculosis Emergency update 2008.
14. The Global MDR-TB and XDR-TB Response Plan 2007–2008. Geneva, World Health Organization, 2007 WHO/HTM/TB/2007.387.
15. Emergence of Mycobacterium tuberculosis with extensive resistance to second-line drugs – worldwide, 2000–2004. Morbidity and Mortality Weekly Report, 2006, 55(11): 301–305.
16. Global tuberculosis control - epidemiology, strategy, financing WHO Report 2009 WHO/HTM/TB/2009.411.
17. Standards for TB care in India. WHO Report 2010. Retrieved from http://www.searo.who.int/india/mediacentre/events/2014/stci_book.pdf?ua=1
18. Global Tuberculosis report 2015- World Health Organisation (2015, October). Tuberculosis Fact Sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
19. Global Tuberculosis report 2008- World Health Organisation (2008). Summary on Tuberculosis. http://www.who.int/tb/publications/global_report/2008/summary/en/index.html
20. Global Tuberculosis report 2015- World Health Organisation (2015, October). Tuberculosis Fact Sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs104/en/
21. Global tuberculosis control - epidemiology, strategy, financing WHO Report 2009 WHO/HTM/TB/2009.411.
22. Global tuberculosis control - epidemiology, strategy, financing WHO Report 2009 WHO/HTM/TB/2009.411.
23. TB-HIV Co-Infection in India. Preetish S. Vaidyanathan & Sanjay Singh; NTI Butletin 2003, 39 / 3&4, 11-18.

24. TB Alert India. WHO: Antimicrobial resistance a global health emergency. Retrieved from http://www.tbalertindia.org.
25. WHO 2003. Annual risk of tuberculous infection in north India. Bulletin of the World Health Organization 2003; 81(8), 551-628. Retrieved from http://www.who.int/bulletin/volumes/81/8/en/
26. Katzung *et al*.: Basic and clinical Pharamacology 13E, (2001), McGraw-Hill Education - Europe. ISBN13 97800 71825054.
27. Lisa *et al*.: Latent Tuberculosis Infection in Children, 2017. Data retrieved fro http://www.uptodate.com/contents/latent-tuberculosis-infection-in-children
28. Cole *et al*.: Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 1998; 393(6695):537-44.
29. Chan J and Kaufmann: S.H.E. in Tuberculosis: Pathogenesis, Protection, and Control (ed. Bloom, B. R.) 1994; 271–284.
30. Cole ST, Brosch R, Parkhill J *et al*.: Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998; 393: 537-544.
31. http://genolist.pasteur.fr/TubercuList/
32. Jean-Christophe Camus, Melinda J. Pryor, Claudine Medigue and Stewart T. Cole: Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, 2002; 148: 2967–2973.
33. Cole ST, Eiglmeier K, Parkhill J *et al*.: Massive gene decay in the leprosy bacillus. Nature 2001; 409; 1007-1011.
34. Vishnoi A, Srivastava A, Roy R and Bhattacharya A MGDD: *Mycobacterium tuberculosis* Genome Divergence Database BMC Genomics. 2008; 9: 373.
35. http://www.biohealthbase.org/
36. Marcos Catanho, Daniel Mascarenhas, Wim Degrave and Antonio Basílio de Miranda: GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. Genet. Mol. Res. 2006; 5 (1): 115-126.
37. http://www.broadinstitute.org/annotation/genome/mycobacterium_tuberculosis_spp/MultiHome.html
38. Reddy TBK, Riley R, Wymore F Montgomery P, DeCaprio D, *et al*.: TB database: an integrated platform for tuberculosis research Nucleic Acids Research, 2008; 1–10.
39. Sundaramurthi JC, Ramanandan P, Brindha S, Subhasree, CR, *et al*: DDTRP – Database of drug targets for resistant pathogens. Bioinformation; 2011; 7; 98-101.
40. Sundaramurthi JC, Brindha S, Reddy TB, Hanna LE: Informatics resources for tuberculosis – towards drug discovery. Tuberculosis 2012; 92(2): 133-138.
41. Unissa AN, Hassan S, Selvakumar N: Elucidating isoniazid resistance in *Mycobacterium tuberculosis* using molecular docking approach. Int. J Pharma Bio Sci 2012; 3(1): 314-26.
42. Raghavan S, Alagarasu K, Selvaraj P: Immunogenetics of HIV and HIV associated tuberculosis. Tuberculosis 2012; 92(1): 18-30.
43. Bourai N, Jacobs WR Jr, Narayanan S: Deletion and overexpression studies on DacB2, a putative low molecular mass penicillin binding protein from *Mycobacterium tuberculosis* H 37 Rv. Microb Pathog 2012; 52(2):109-116.
44. Sundaramurthi JC, Brindha S, Shobitha SR, Swathi A, Ramanandan P, Hanna LE: *In silico* identification of potential antigenic proteins and promiscuous CTL epitopes in *M. ycobacterium* tuberculosis. Infect Genet Evol 2012; 12(6): 1312-8.
45. Narayanan S and Deshpande U: Whole-Genome sequences of four clinical isolates of Mycobacterium tuberculosis from Tamil Nadu, South India. Genome Announc. 2013; 1:1.
46. Fei Liu, Yongfei Hu and Qi Wang: Comparative genomic analysis of *Mycobacterium tuberculosis* clinical isolates. BMC Genomics 2014; 15: 469.
47. Subhasree CR and Subramaniam S: Virtual screening of small molecules against deoxycytidine triphosphate deaminase of *Mycobacterium tuberculosis*. World Journal of Pharmaceutical Research. 4(8): 2826-2861.
48. Hamilton CD, Swaminathan S, Christopher DJ, Ellner J, Gupta A, Sterling TR, Rolla V, Srinivasan S, Karyana M, Siddiqui S, Stoszek SK and Kim P: RePORT International: Advancing tuberculosis biomarker research through global collaboration. Clin Infect Dis. 2015; 61(S3):S155-S159.
49. Gagan Deep Jhingan, Sangeeta Kumari, Shilpa V: Jamwal, Haroon Kalam, *et al*., Comparative Proteomic Analyses of Avirulent, Virulent, and Clinical Strains of Mycobacterium tuberculosis Identify Strain-specific Patterns. The Journal of Biological Chemistry. 2016; 291, 14257-14273.

How to cite this article:
Subhasree CR, Priya RSK, Diwakar M, Subramaniam S and Shyama S: Review on comparative genomics for *Mycobacterium tuberculosis* strains. Int J Pharm Sci Res 2017; 8(12): 5022-42.doi: 10.13040/IJPSR.0975-8232.8(12).5022-42.