







Article

# Pangenome Analysis of *Mycobacterium tuberculosis* Reveals Core-Drug Targets and Screening of Promising Lead Compounds for Drug Discovery

Hamza Arshad Dar <sup>1</sup>, Tahreem Zaheer <sup>1</sup>, Nimat Ullah <sup>1</sup>, Syeda Marriam Bakhtiar <sup>2</sup>, Tianyu Zhang <sup>3</sup>, Muhammad Yasir <sup>4,5</sup>, Esam I. Azhar <sup>4,5,\*</sup> and Amjad Ali <sup>1,\*</sup>

<sup>1</sup> Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology, Islamabad 44000, Pakistan; darhamza000@gmail.com (H.A.D.); zaheertahreem@gmail.com (T.Z.); nimatscholar@gmail.com (N.U.)

<sup>2</sup> Department of Bioinformatics and Biosciences, Capital University of Science and Technology Islamabad expressway, Kahuta Road, Zone-V, Islamabad 44000, Pakistan; marriam@cust.edu.pk

<sup>3</sup> State Key Laboratory of Respiratory Disease, Guangzhou Institutes of Biomedicine and Health (GIBH), Chinese Academy of Sciences, Guangzhou 510530, China; zhang\_tianyu@gibh.ac.cn

<sup>4</sup> Special Infectious Agents Unit, King Fahd Medical Research Center, King Abdulaziz University, Jeddah 21589, Saudi Arabia; yamuhammad@kau.edu.sa

<sup>5</sup> Medical Laboratory Technology Department, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah 21589, Saudi Arabia

\* Correspondence: eazhar@kau.edu.sa (E.I.A.); amjad.ali@asab.nust.edu.pk (A.A.); Tel.: +92-3339191903 (A.A.)

Received: 16 October 2020; Accepted: 15 November 2020; Published: 17 November 2020



**Abstract:** Tuberculosis, caused by *Mycobacterium tuberculosis* (*M. tuberculosis*), is one of the leading causes of human deaths globally according to the WHO TB 2019 report. The continuous rise in multi- and extensive-drug resistance in *M. tuberculosis* broadens the challenges to control tuberculosis. The availability of a large number of completely sequenced genomes of *M. tuberculosis* has provided an opportunity to explore the pangenome of the species along with the pan-phylogeny and to identify potential novel drug targets leading to drug discovery. We attempt to calculate the pangenome of *M. tuberculosis* that comprises a total of 150 complete genomes and performed the phylo-genomic classification and analysis. Further, the conserved core genome (1251 proteins) is subjected to various sequential filters (non-human homology, essentiality, virulence, physicochemical parameters, and pathway analysis) resulted in identification of eight putative broad-spectrum drug targets. Upon molecular docking analyses of these targets with ligands available at the DrugBank database shortlisted a total of five promising ligands with projected inhibitory potential; namely, 2'-deoxy-thymidine-5'-diphospho-alpha-D-glucose, uridine diphosphate glucose, 2'-deoxy-thymidine-beta-L-rhamnose, thymidine-5'-triphosphate, and citicoline. We are confident that with further lead optimization and experimental validation, these lead compounds may provide a sound basis to develop safe and effective drugs against tuberculosis disease in humans.

**Keywords:** *Mycobacterium tuberculosis*; pangenome; drug targets; molecular docking; lead compounds; drug discovery; tuberculosis

## 1. Introduction

According to the 2019 global World Health Organization (WHO) report, tuberculosis (TB) is among the top ten causes of death and the leading cause of a single infectious agent (above HIV/AIDS) [1]. In 2018, this disease was responsible for 1.2 million deaths among HIV-negative people

and an additional 251,000 deaths among HIV-positive people. The causative agent of human TB is *Mycobacterium tuberculosis* (*M. tuberculosis*).

TB is majorly an airborne disease where *M. tuberculosis* establishes infection inside the human body by overcoming the immune responses directed by the host against the foreign pathogen. Once inside, the bacteria infiltrate the host macrophages and persist inside them for years thus causing a chronic infection that reflects the failure of host immunity [2]. The increased occurrence of multidrug resistant (MDR) and extensively drug-resistant (XDR) strains of *M. tuberculosis* has been attributed to spontaneous mutations in the bacterial genome, followed by the emergence of these mutant strains at the expense of wild type strains [3]. This, in turn, has led to the loss of effectiveness of standard anti-TB drugs isoniazid and rifampicin.

The availability of the first *M. tuberculosis* genome sequence in 1998 along with further developments in the genomics and associated disciplines has enabled us to understand the importance of many proteins encoded in its genome [4,5]. Especially, the concept of pangenome can now be explored to obtain the core genome, i.e., genes present in all the strains of the dataset [6]. These genes can be exploited to design broad-spectrum therapeutics against the pathogenic species.

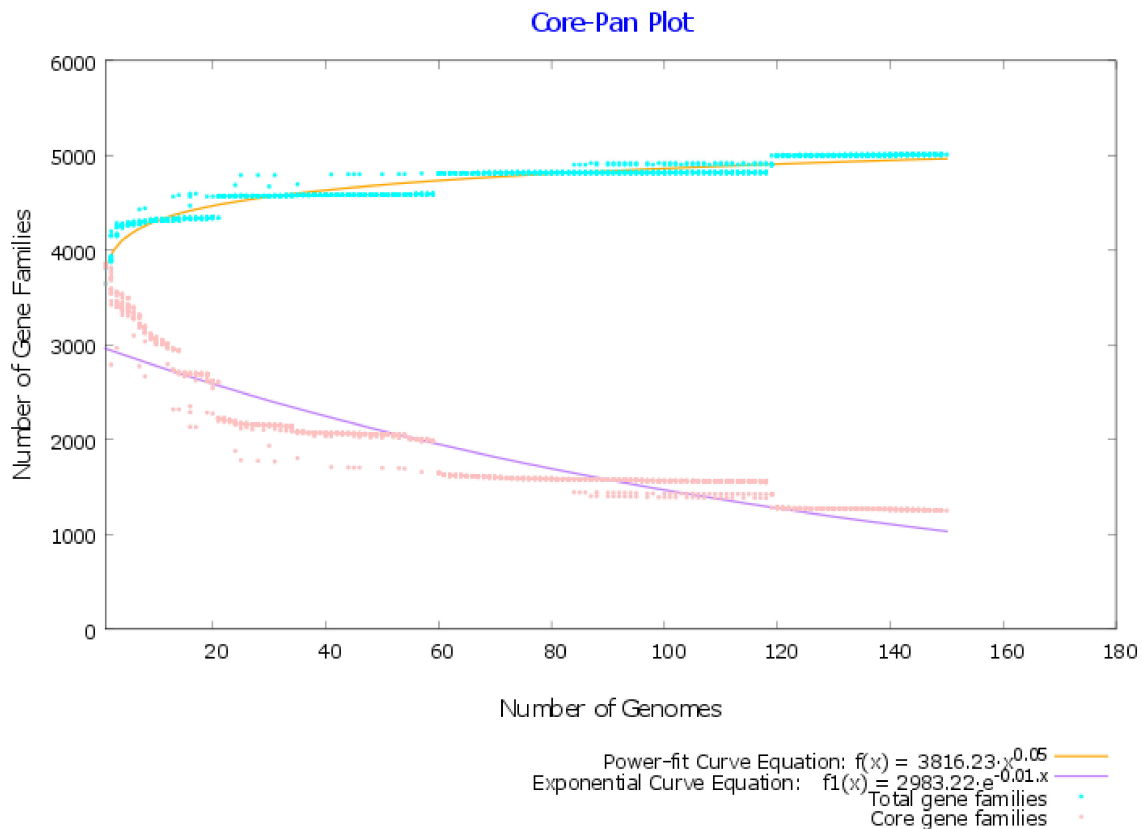
Various protocols in bioinformatics have helped scientists to process biological data of pathogens in order to prioritize drug targets, while the parallel developments in the cheminformatics help researchers access ligand databases such as the DrugBank and perform virtual screening for in silico-aided drug discovery [7,8]. Utilizing these useful tools, scientists have endeavored to identify lead compounds for tuberculosis [9,10]. Therefore, in this study, we used the integrative approach of pangenome analyses, subtractive genomics, and bioinformatics to prioritize potential drug targets in the *M. tuberculosis* genome. Further, we explored the structural association of these potential targets with FDA-approved drugs using molecular docking to shortlist promising lead molecules to aid drug development against TB.

Developing a new drug through conventional methods takes at least ten years of comprehensive research and huge funding; Nevertheless, the inclusion of computer-aided analyses at the initial stages can decrease the associated time and costs [11,12]. In silico drug screening serves as a valuable method to shortlist/prioritize only the most relevant compounds that could be checked later through experimental studies, and hence remove biomolecules that do not meet the required specifications to optimize the overall research in drug discovery [13]. Instead of focusing our attention on one drug target, we intended to scrutinize all high-ranked drug targets obtained computationally. Subject to experimental validation, this research work will provide a sound basis for developing novel therapeutics against the most troublesome bacterial pathogen *M. tuberculosis*.

## 2. Results and Discussion

### 2.1. Pangenome and Pan-Phylogeny Analysis of *Mycobacterium tuberculosis* Genomes

A total of 150 complete genomes and associated proteomes of *M. tuberculosis* were downloaded from the NCBI. Their information such as the accession number, strain name, and genome statistics are provided in Supplementary Table S1. The metadata associated with these bacterial strains such as the isolation source and the country of isolation is provided in Supplementary Table S2. Pangenome analysis of 150 *M. tuberculosis* complete genomes revealed that there were 5009 gene families in total (pangenome), among them, 1251 proteins were common (core genome) in all the studied genomes. The ratio of the core and pangenome size was found to be 0.25, thus the core forms 25% of the pangenome. This signifies the low level of genetic diversity in the *M. tuberculosis* strains. This trend is also visible in the pan-core genome plot, which projects that the global gene repertoire of this species is difficult to alter considerably in the future (Figure 1).

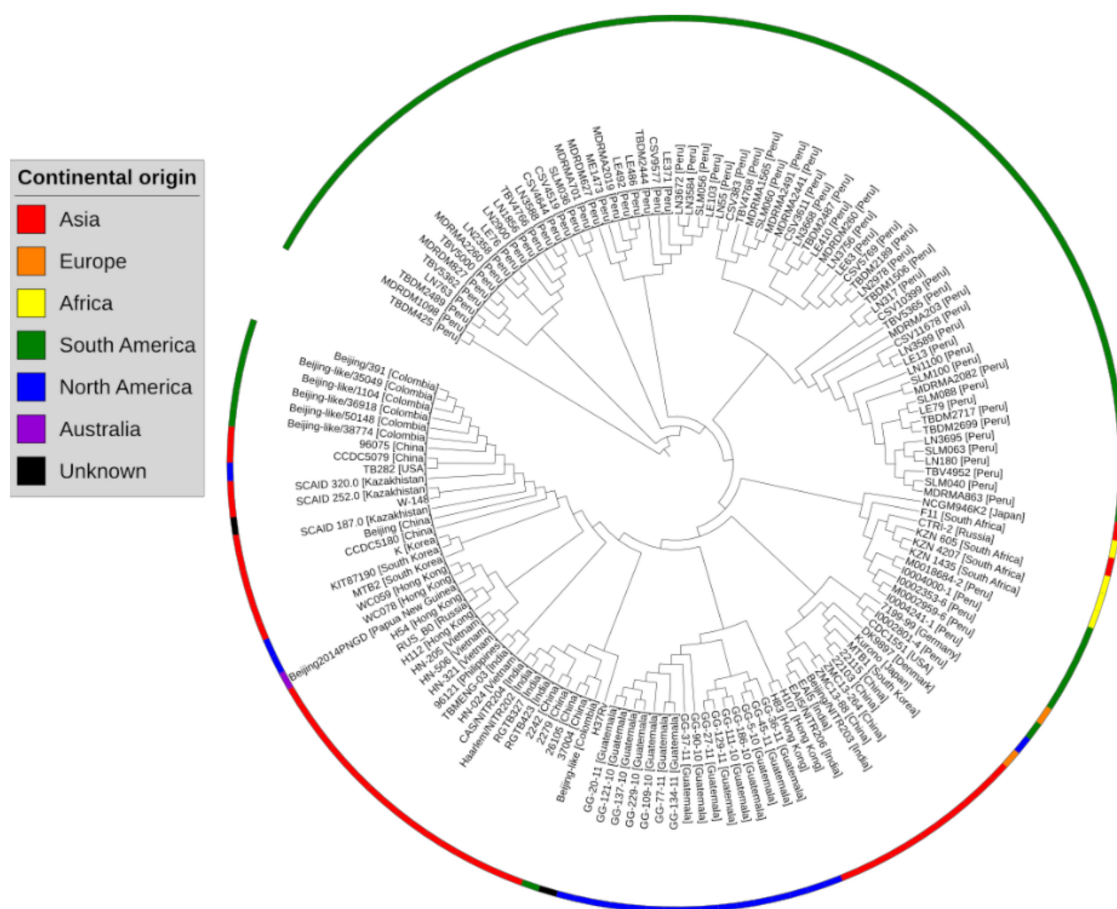


**Figure 1.** Pan-core plot of 150 *Mycobacterium tuberculosis* genomes. X-axis shows the number of genomes while the y-axis represents the number of gene families. With the addition of every genome, the pangenome size increased while the core genome size declined. The pangenome curve (shown in brown) has almost flattened, i.e., reached plateau. This suggests that the global gene repertoire of this species is unlikely to change significantly in the future and the pangenome is almost closed.

According to pangenome calculations, the b value of 0.0542 in the power-law regression model is indicative of a nearly close pangenome for *M. tuberculosis*. The pan-phylogeny based phylogenetic tree along with geographical source information is provided in Figure 2.

Each genome on average contained a total of 4102 protein-encoding genes. The core genome size (1251) thus accounted for 30.5% of the average genome size. Similarly, the minimum number of protein-encoding genes (3622) was present in strain RGTB423 while the maximum number of protein-encoding genes (4599) was present in strain 2279.

The level of conservation in the *M. tuberculosis* genomes is believed to be very high, with limited genetic variability between/across its various strains. Our genome-level analysis also indicates that the pangenome of this species is almost closed, with minimum variations. Multiple studies have corroborated this claim [14–16]. This is expected, as the lifestyle of this bacterium and its restricted niche makes it difficult to alter its gene pool [17]. Due to these unique conditions and their dependence on the host, it is suggested that there is less flexibility in its genome to accommodate more genes into the global gene repertoire. Hence, considering all this information now is the perfect opportunity to design and develop effective broad-spectrum therapeutics against this deadly human pathogen and to control the further spread of tuberculosis.



**Figure 2.** Pan-phylogeny tree of 150 complete genomes of *Mycobacterium tuberculosis*. Colored strips show the continental origin of strains. Overall, the strains originating from different geographical regions of the World clustered into different clades. This pan-phylogeny has been constructed based on accessory gene presence/absence data in different strains.

## 2.2. Subtractive Proteomics Revealed Putative *Mycobacterium tuberculosis* Drug Targets

Differential genome analysis was conducted on the core proteins of *M. tuberculosis* for the identification of therapeutic targets. Targets found to be human homologs could adversely affect the host metabolism, therefore, all those proteins were excluded in the first step that were identified as human homologs. The Rv numbers (gene identifiers or virulent strain of *Mycobacterium tuberculosis*) and the gene annotation were also conducted on these core proteins; the core genomes-associated data is provided in Supplementary Table S3. A total of 1185 proteins were identified as non-human homologous (Supplementary File S1) and were further screened based on their essentiality. Among them, 377 proteins were characterized as essential proteins and were considered crucial for pathogen survival (Supplementary File S2). If essential proteins are also functionally characterized as virulent, they are especially of vital significance to unveil novel therapeutic targets as these proteins help bacteria to modulate or degrade host defense mechanisms and may contribute to pathogenesis [18]. Therefore, the essential proteins were further screened to identify genes associated with pathogenicity. Among 377 essential proteins, VFDB and MvirDB identified a total of 93 virulence-related proteins involved in *M. tuberculosis* pathogenicity (Supplementary File S3). Since all the 93 proteins are non-human homologs and essential proteins associated with virulence, they represent an attractive dataset that could be explored for future vaccine production and drug design to tackle tuberculosis disease. However, for the purpose of this study, we further mined this large dataset to shortlist a few potential drug targets in order to facilitate the drug design and development against *M. tuberculosis*.

A total of 55 proteins were obtained after applying physicochemical checks, i.e., low molecular weights, a high value of the aliphatic index and negative GRAVY score (Supplementary File S4). Out of these, only eight proteins were identified by comparative pathway analysis to be involved in the unique bacterial metabolic pathway/s and were positively selected as *M. tuberculosis* drug targets to avoid targeting any human pathway upon drug therapy (Table 1).

**Table 1.** Metabolic pathway analysis of potential drug targets.

Protein Name	Rv Locus Number	KEGG Orthology	Metabolic Pathway(s)
dTDP-4-dehydrorhamnose reductase	Rv3266c	K00067	ko00521 Streptomycin biosynthesis ko00523 Polyketide sugar unit biosynthesis ko00541 O-Antigen nucleotide sugar biosynthesis ko01100 Metabolic pathways ko01110 Biosynthesis of secondary metabolites
glucose-1-phosphate thymidyltransferase	Rv0334	K00973	ko00521 Streptomycin biosynthesis ko00523 Polyketide sugar unit biosynthesis ko00525 Acarbose and validamycin biosynthesis ko00541 O-Antigen nucleotide sugar biosynthesis ko01100 Metabolic pathways ko01110 Biosynthesis of secondary metabolites
two-component system regulator trcR	Rv1033c	K07672	ko02020 Two-component system
two-component system regulator mtrA	Rv3246c	K07670	ko02020 Two-component system
two-component system regulator regX3	Rv0491	K07776	ko02020 Two-component system
two-component system regulator kdpE	Rv1027c	K07667	ko02020 Two-component system ko02024 Quorum sensing
two-component system regulator devR	Rv3133c	K07695	ko02020 Two-component system
dTDP-4-dehydrorhamnose 3,5-epimerase	Rv3465	K01790	ko00521 Streptomycin biosynthesis ko00523 Polyketide sugar unit biosynthesis ko00541 O-Antigen nucleotide sugar biosynthesis ko01100 Metabolic pathways ko01110 Biosynthesis of secondary metabolites

Finally, a total of seven prioritized proteins were found to have similarity with targets associated with FDA-approved drugs and thus are likely druggable. These seven proteins, and their corresponding ligands, are tabulated in Table 2.

**Table 2.** The Autodock vina score of all drug targets with their ligands from the DrugBank database. The binding energies are indicated by a bold black color in the case of top interactions.

Target	Rv Locus Number	Ligand	Binding Energy
Glucose-1-phosphate thymidyltransferase	Rv0334	2'-Deoxy-Thymidine-Beta-L-Rhamnose	-9.1
		2'-deoxy-Thymidine-5'-Diphospho-Alpha-D-Glucose	-10.1
		Alpha-D-Glucose-1-Phosphate	-5.8
		Citicoline	-8.4
		Citric acid	-5.7
		Thymidine-5'-Triphosphate	-8.9
		Thymidine	-7.1
		Thymidine monophosphate	-7.7
DNA-binding response regulator	Rv1027c	Uridine diphosphate glucose	-9.8
		Phosphoaspartate	-5.2
		Guanosine-5'-Monophosphate	-7.9
		AlphaBeta-Methyleneadenosine-5'-Triphosphate	-7.3
		Adenosine-5'-Rp-Alpha-Thio-Triphosphate	-6.8
dTDP-4-dehydrorhamnose 3,5-epimerase	Rv3465	2-Hydroxyestradiol	-7.9
		2'-deoxy-Thymidine-5'-Diphospho-Alpha-D-Glucose	-7.2
		3'-O-Acetylthymidine-5'-Diphosphate	-7.1
		D-tartaric acid	-4.6
		SS-(2-Hydroxyethyl)Thiocysteine	-4.6
		Thymidine_monophosphate	-6.8
Thymidine-5'-diphospho-beta-D-xylose	-6.7		

Table 2. Cont.

Target	Rv Locus Number	Ligand	Binding Energy
DNA-binding response regulator TrcR	Rv1033c	S-Methyl Phosphocysteine	-4.6
		Phosphoaspartate	-4.8
		Guanosine-5'-Monophosphate	-7.2
		Glycerine	-4
		AlphaBeta-Methyleneadenosine-5'-Triphosphate	-7.4
		Adenosine-5'-Rp-Alpha-Thio-Triphosphate	-7.6
		3-Aminosuccinimide	-4.5
DNA-binding response regulator RegX3	Rv0491	2-Hydroxyestradiol	-7.1
		3-Aminosuccinimide	-4.4
		Adenosine-5'-Rp-Alpha-Thio-Triphosphate	-6.7
		AlphaBeta-Methyleneadenosine-5'-Triphosphate	-6.6
		Glycerine	-3.8
		2-Hydroxyestradiol	-7.1
		3-Aminosuccinimide	-4.4

However, one prioritized target (two-component transcriptional regulatory protein DevR) did not show similarity to any drug target in the DrugBank database thus it is potentially a novel drug target worthy of experimental testing. Virtual screening may be performed with ligands from other ligand databases in order to find potent inhibitors of this protein. Literature studies were conducted on the seven proteins from Table 2 to understand their relevance as a drug target.

Among them, *trcR* has a crucial role in intracellular survival of *M. tuberculosis* and also in the regulation of Rv1057 expression, i.e., the  $\beta$ -propeller gene activated by the envelope stress [19]. In another study by Haydel et al.; they demonstrated that *trcR* and *trcS* two-component system genes are transcribed in broth-grown *M. tuberculosis* validated through reverse transcription PCR analysis. Moreover, through the SCOTS technique the expression of these genes are observed in liquid (broth) medium and after 18 h of *M. tuberculosis* growth in cultured human primary macrophages [20]. Nevertheless, the stimulatory signal for the TrcR-TrcS system is poorly understood and requires further research. Recently, another study, based on the drug target score, also ranked TrcR as the second-best therapeutic target in *M. tuberculosis* [21]. We believe that future studies are also needed to investigate this matter.

The dTDP-4-dehydrorhamnose reductase, another protein classified as a drug target by our study, is already known to be a high-confidence drug target [22]. This protein plays an important role in cell wall synthesis. Targeting the cell wall of *M. tuberculosis* through inhibitors is a good strategy, as has been recognized by earlier studies [23,24].

Similarly, *mtrA* (in Table 1) plays an important role in many crucial processes such as cell wall homeostasis and bacterial growth. It also has a role in cell division, DNA replication, and controls susceptibility of *M. tuberculosis* to the first line antimycobacterial drugs [25]. This signifies the necessity of *mtrA* for the growth of *M. tuberculosis*.

Another prioritized drug target RegX3 (in Table 1) has a role in aerobic respiration, virulence and phosphate absorption [26]. It is also required for optimal growth of *M. tuberculosis* in nutrient rich medium [27]. Moreover, RegX3 directly interacts with *exs-5*, another essential gene for the growth process. It also causes an increase in the production of membrane vesicles in *M. tuberculosis* [28].

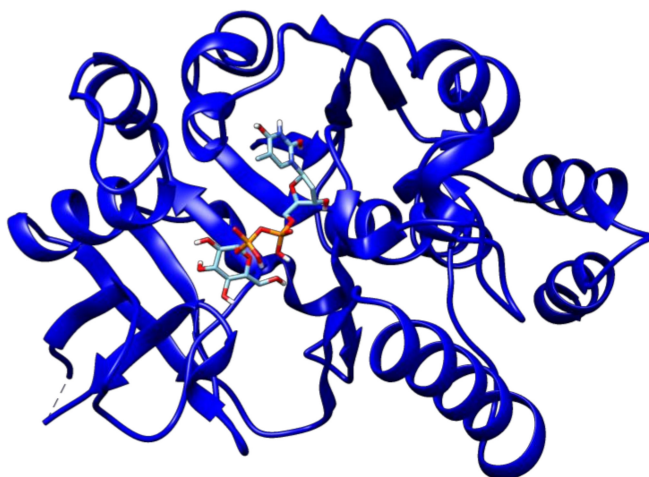
The *kdpE* participates in the two-component system pathway and is involved in transcriptional regulation of Kdp-ATPase [29]. *KdpE* also has a role in the regulation of key factors involved in stress tolerance and energy metabolism [30,31]. Moreover, it is also notorious for its role in virulence caused by *M. tuberculosis*, *Pseudomonas aeruginosa*, *Salmonella enterica*, and *Staphylococcus aureus* [29,32–35]. Considering the situation, *kdpE* can be considered as a broad-spectrum therapeutic target for multiple pathogens.

The present study also prioritized drug target glucose-1-phosphate thymidyltransferase. The protein is involved in three metabolic pathways namely, polyketide sugar unit biosynthesis, streptomycin biosynthesis, and the nucleotide sugars metabolism pathway [36–38]. All these pathways

are crucial for survival of *M. tuberculosis* and elaborate on the relevance of glucose-1-phosphate thymidyltransferase as a drug target. Additionally, the dTDP-4-dehydrorhamnose 3,5-epimerase protein obtained by our study is known to be a potential drug target in the rhamnose pathway [39].

### 2.3. Docking Analyses of Drug Targets Revealed Potential Lead Compounds for Drug Discovery

The Autodock vina docking scores of ligands with drug targets are elaborated in Table 2. Among all the drug–ligand interaction analyses, 2′ deoxy-thymidine-5′-diphospho-alpha-D-glucose was found to have the lowest binding energy (−10.1 kcal/mol) with glucose-1-phosphate thymidyltransferase (Figure 3).

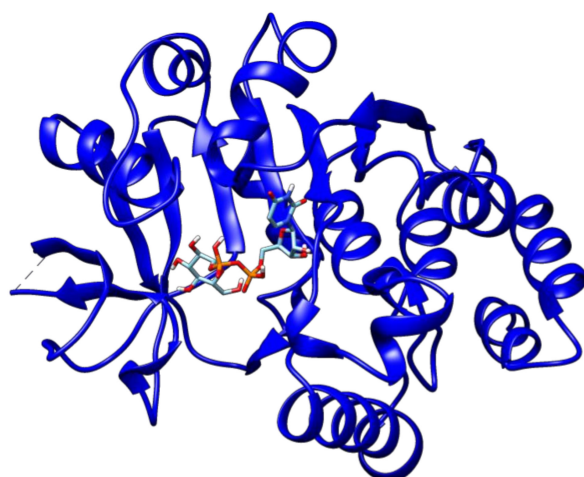


**Figure 3.** The docked complex of 2′ deoxy-thymidine-5′-diphospho-alpha-D-glucose and glucose-1-phosphate thymidyltransferase.

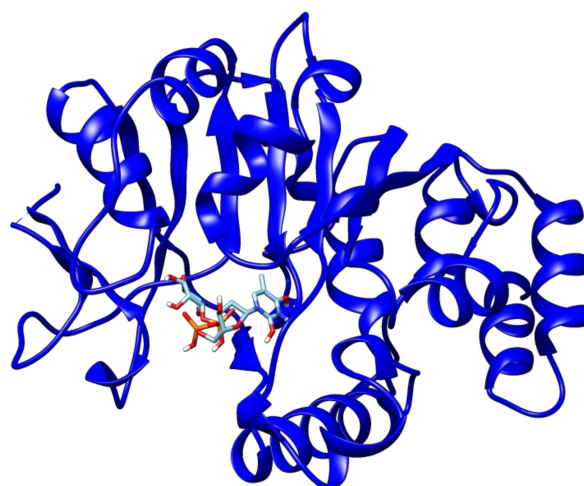
Interestingly, 2′ deoxy-thymidine-5′-diphospho-alpha-D-glucose was found to associate with another prioritized drug target protein of this study DNA response regulator, however, the binding energy was somewhat higher (−7.2 kcal/mol) in that case. Nevertheless, our docking analyses proposed the 2′ deoxy-thymidine-5′-diphospho-alpha-D-glucose compound as a high-ranked lead compound to develop a drug against tuberculosis.

The second high-ranked ligand in our study is uridine diphosphate glucose. This compound showed a binding energy of −9.8 kcal/mol with glucose-1-phosphate thymidyltransferase (Figure 4), and thus may also represent a good target for lead optimization. Meanwhile, 2′-deoxy-thymidine-beta-L-rhamnose ligand exhibited docked energy of −9.1 kcal/mol with glucose-1-phosphate thymidyltransferase (Figure 5).

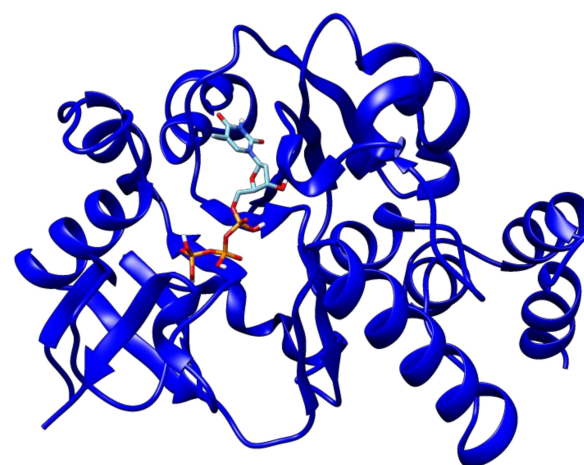
This ligand also interacted with another prioritized protein of our study dTDP-4-dehydrorhamnose reductase with the binding energy of −8.3 kcal/mol. Another high-ranked ligand thymidine-5′-triphosphate ligand was found to have −8.9 kcal/mol binding energy with glucose-1-phosphate thymidyltransferase (Figure 6) while citicoline, on the other hand, showed −8.4 kcal/mol binding energy with this same drug target (Figure 7).



**Figure 4.** The docked complex of uridine diphosphate glucose and glucose-1-phosphate thymidyltransferase.

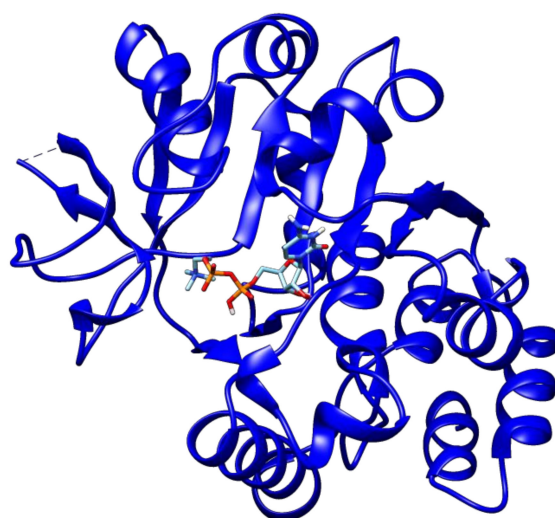


**Figure 5.** The docked complex of 2'-deoxy-thymidine-beta-L-rhamnose and glucose-1-phosphate thymidyltransferase.



**Figure 6.** The docked complex of thymidine-5'-triphosphate and glucose-1-phosphate thymidyltransferase.





**Figure 7.** The docked complex of citicoline and glucose-1-phosphate thymidyltransferase.

Thus, overall, this study shortlisted five top-ranked ligands with good binding affinity to drug targets and thus may guide drug development studies to counter *M. tuberculosis*. These ligands are 2'-deoxy-thymidine-5'-diphospho-alpha-D-glucose, uridine diphosphate glucose, 2'-deoxy-thymidine-beta-L-rhamnose, thymidine-5'-triphosphate, and citicoline, which have good binding affinity with metabolic pathway proteins especially glucose-1-phosphate thymidyltransferase. The protein is involved in three metabolic pathways namely, polyketide sugar unit biosynthesis, streptomycin biosynthesis, and nucleotide sugars metabolism pathway [36–38]. All these pathways are crucial for survival of *M. tuberculosis* and shortlisted repurposed drugs have higher affinity with the enzyme, i.e., glucose-1-phosphate thymidyltransferase. Moreover, these drugs have a tendency to bind with dTDP-4-dehydrothymidine 3,5-epimerase that is a potential drug target in rhamnose pathway [39]. These findings may provide a basis for the development of effective therapeutics against tuberculosis and/or replace currently available drugs to treat this deadly pathogen, however, experimental validation is needed to confirm these results.

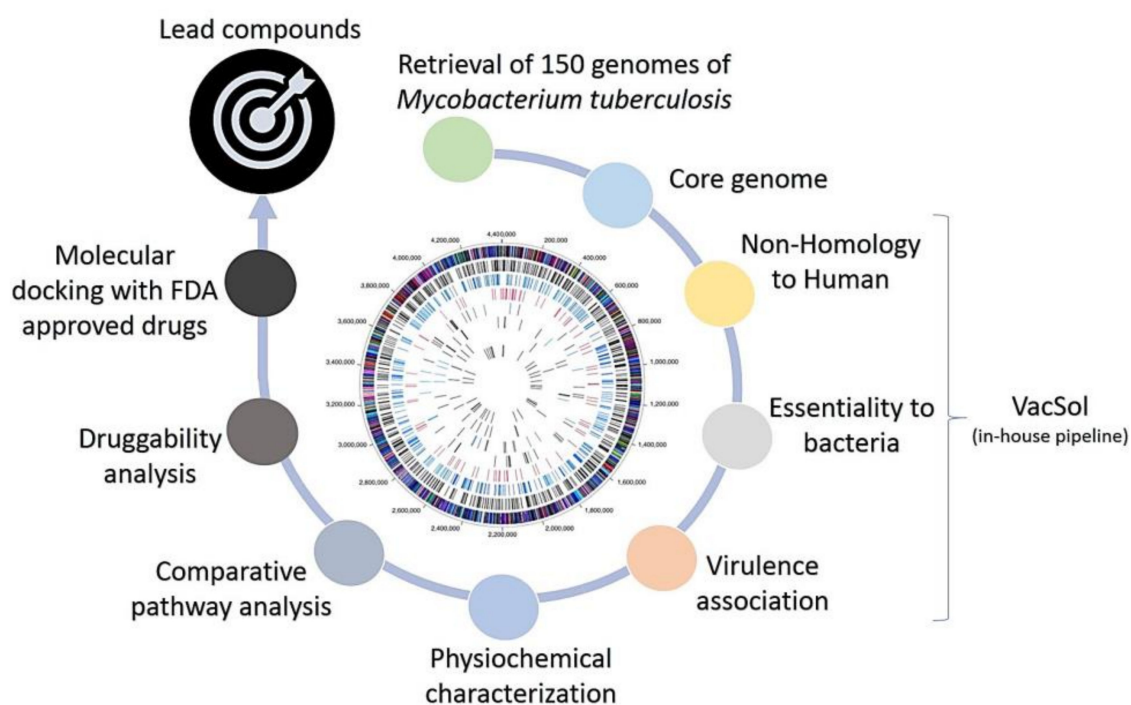
### 3. Materials and Methods

The methodology adopted in this study was visualized in Figure 8. The different steps were elaborated below.

A total of 150 complete genome sequences of *Mycobacterium tuberculosis* were retrieved from the NCBI GenBank database and pangenome analysis was conducted to identify the genes shared by all the strains of our dataset (i.e., the core genome). The core genome was then subjected to various sequential filters of subtractive genomics (non-homology to humans, bacterial essentiality and virulence, physiochemical checks, comparative pathway analysis, and druggability analysis) to select potential drug targets. Finally, these proteins were subjected to molecular docking with ligands from the DrugBank database to identify potential lead compounds that could be useful to develop a drug against tuberculosis.

#### 3.1. Collection of Genomic Data

The complete genome sequences of 150 *M. tuberculosis* strains and their associated proteomes were retrieved from the GenBank database (<https://www.ncbi.nlm.nih.gov/genbank/>) available at the National Center for Biotechnology Information (NCBI). These strains had been earlier isolated from all the geographical regions of the World.



**Figure 8.** The methodology adopted in the study to find lead compounds (drugs) against *Mycobacterium tuberculosis*.

### 3.2. Pangenome Analysis of *Mycobacterium tuberculosis* Strains

Pangenome analysis was conducted on the genomes using the Bacterial Pan Genome Analysis (BPGA) tool to obtain highly conserved proteins of *M. tuberculosis* [40]. For this, we used the default threshold of 50% identity to generate orthologous protein clusters by the USEARCH algorithm [41]. BPGA calculates pan and core genome/proteome size by randomly considering 20 permutations and sequentially stating median values after the addition of each genome. Graphically, the core and pan-genome curves are generated by comparing the total number of common and unique gene families against the total number of genomes, respectively. The output also shows the number of new genes with the addition of every genome. The pan phylogeny was generated using the pan-matrix data and the associated pangenome tree was constructed using the neighbor-joining method with a default combination value of 20 iterations. Protein sequences associated with the core genome were retrieved and subjected to further screening steps to identify potential therapeutic targets.

### 3.3. Identification of Non-Host Homologous, Essential, and Virulence-Associated Proteins

To aid the preliminary identification of novel therapeutic targets in the core genome/proteome, we used an in-house pipeline VacSol for non-homology, essentiality, and virulence screening steps [42]. To avoid harmful responses in humans due to drug therapy, the drug targets must be non-homologous to human proteins, so for the identification of non-host homologous targets, the NCBI BLASTp tool (E-value  $1 \times 10^{-3}$ ) was used to compare the core genome with the human genome [43]. From this non-host homologous conserved proteome, essential genes were identified using BLASTp against the Database of Essential Genes (DEG) at an E-value  $<0.0001$  and bit score  $>100$  [44]. The DEG comprises experimentally validated data collected from archaea, bacteria, and eukaryotes, including currently reported essential genomic elements such as genes that are required for cellular life. We subjected the selected proteins to BLASTp search against the Virulence Factor Database (VFDB) and the Microbial Virulence database (MvirDB) to identify potential virulence-associated factors [45,46]. The VFDB is a comprehensive online database that contains information about virulence factors of bacterial pathogens. Virulence factors are known gene products that enable a microorganism to establish itself within a host

and thus enhance its disease-inducing potential [47]. MvirDB is another integrated online database comprising publicly available, organized sequences related to known virulence factors, toxins, and antibiotic resistance genes [46].

#### 3.4. Identification of Putative *Mycobacterium tuberculosis* Drug Targets

To identify potential drug targets, the non-human homolog, essential, and virulence-associated core proteins of *M. tuberculosis* were further filtered based on physicochemical parameters such as molecular weight, pI (isoelectric point), grand average of hydropathicity (GRAVY) value, and aliphatic index using the ProtParam tool [48]. Low molecular weight proteins are considered good drug targets, as they are accessible to drugs [49]. A higher value of the aliphatic index of proteins indicates thermostability whereas the negative GRAVY (grand average of hydropathicity) value indicates the hydrophilic nature of putative drug targets [50]. After physicochemical filters, comparative pathway analysis of filtered proteins was performed using the KEGG Automatic Annotation Server (KAAS) version 2.1 to identify the proteins associated with pathogen-specific pathways to enable specific targeting of the pathogen [51].

Finally, druggability assessment of the shortlisted proteins was performed using BLASTp against the DrugBank database at default settings [52]. The druggability of potential targets is also a crucial screening step that evaluates the potential of prioritized targets to be modulated by a drug or drug-like entity [50]. The potential targets should bind to a drug or drug-like compound with high affinity. However, this was not considered an inclusion criterion as novel drug targets may not display homology to targets associated with FDA approved drugs and show specific affinity to other drugs from natural products [53].

#### 3.5. Molecular Docking of Putative Drug Targets with Drugs

The sequences of drug targets were checked for the availability of crystal structures, if any, from the RCSB Protein DataBank [54]. Structural file PDB was downloaded, and all the attached ligands were removed from the crystal structure. The drug targets having no associated crystal structure available in the database were modeled using the i-TASSER server [55,56]. Structure with the highest C-Score was selected in each case as these models are of good quality and stability. Using the PyRx-incorporated Autodock vina tool, the structures of drug targets were subjected to molecular docking with the FDA-approved drugs identified earlier in the druggability analysis [57,58]. Blind docking was conducted by adjusting the grid-box in order to cover the whole protein space, and the default value of exhaustiveness (8) was selected. Both ligands and their targets were prepared in the PDBQT format as per requirements and molecular docking analysis was performed. Docked complexes with the lowest binding energies were comparatively analyzed to identify those ligands that associated strongly with drug targets. These ligands were thus shortlisted as lead compounds for developing a potent drug against tuberculosis.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2079-6382/9/11/819/s1>, File S1: The core proteins of 150 *Mycobacterium tuberculosis* genomes found to be non-homologous to the human genome given in the FASTA format. File S2: The essential and non-human homolog core proteins identified in genomes under study given in the FASTA format. File S3: Projected virulence-associated proteins from the non-homologous essential core proteins given in the FASTA format. File S4: Protein targets found to have appropriate physicochemical parameters provided. Table S1: Information about 150 complete genome sequences of *Mycobacterium tuberculosis* such as strain name, biosample, bioproject, and genome statistics are provided. Table S2: The metadata associated with these bacterial strains such as isolation source and the country of isolation is provided. Table S3: The annotation of core gene targets and their Rv numbers.

**Author Contributions:** Conceptualization, H.A.D. and T.Z. (Tahreem Zaheer); Data curation, N.U. and S.M.B.; Formal analysis, T.Z. (Tahreem Zaheer) and S.M.B.; Investigation, T.Z. (Tianyu Zhang) and M.Y.; Methodology, H.A.D.; Project administration, A.A. and E.I.A.; Software, H.A.D. and N.U.; Supervision, A.A.; Validation, H.A.D, T.Z. (Tahreem Zaheer); N.U. and S.M.B.; Visualization, M.Y.; Writing—original draft, H.A.D. and T.Z. (Tahreem Zaheer); Writing—review and editing, E.I.A. and T.Z. (Tianyu Zhang) All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, grant number FP-1-42.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. World Health Organization. *Global Tuberculosis Report 2019*; World Health Organization: Geneva, Switzerland, 2019.
2. Gengenbacher, M.; Kaufmann, S.H.E. *Mycobacterium tuberculosis*: Success through dormancy. *FEMS Microbiol. Rev.* **2012**, *36*, 514–532. [[CrossRef](#)]
3. Gandhi, N.R.; Nunn, P.; Dheda, K.; Schaaf, H.S.; Zignol, M.; Van Soolingen, D. Multidrug-resistant and extensively drug-resistant tuberculosis: A threat to global control of tuberculosis. *Lancet* **2010**, *375*, 1830–1843. [[CrossRef](#)]
4. Brindha, S.; Vincent, S.; Velmurugan, D.; Ananthakrishnan, D.; Sundaramurthi, J.C.; Gnanadoss, J.J. Bioinformatics approach to prioritize known drugs towards repurposing for tuberculosis. *Med. Hypotheses* **2017**, *103*, 39–45. [[CrossRef](#)]
5. Cole, S.; Brosch, R.; Parkhill, J.; Garnier, T.; Churcher, C.; Harris, D. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **1998**, *393*, 537. [[CrossRef](#)]
6. Vernikos, G.; Medini, D.; Riley, D.R.; Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **2015**, *23*, 148–154. [[CrossRef](#)]
7. Sundaramurthi, J.C.; Brindha, S.; Reddy, T.B.K.; Hanna, L.E. Informatics resources for tuberculosis—Towards drug discovery. *Tuberculosis* **2012**, *92*, 133–138. [[CrossRef](#)]
8. Ekins, S.; Freundlich, J.S.; Choi, I.; Sarker, M.; Talcott, C. Computational databases, pathway and cheminformatics tools for tuberculosis drug discovery. *Trends Microbiol.* **2011**, *19*, 65–74. [[CrossRef](#)]
9. Karunakar, P.; Girija, C.R.; Krishnamurthy, V.; Krishna, V.; Shivakumar, K.V. In Silico antitubercular activity analysis of benzofuran and naphthofuran derivatives. *Tuberc. Res. Treat.* **2014**, *2014*. [[CrossRef](#)]
10. Perryman, A.L.; Yu, W.; Wang, X.; Ekins, S.; Forli, S.; Li, S.-G. A virtual screen discovers novel, fragment-sized inhibitors of *Mycobacterium tuberculosis* InhA. *J. Chem. Inf. Model* **2015**, *55*, 645–659. [[CrossRef](#)]
11. Timo, G.O.; Reis, R.; de Melo, A.F.; Costa, T.V.L.; de Magalhães, P.O.; Homem-de-Mello, M. Predictive power of in Silico approach to evaluate chemicals against *M. tuberculosis*: A systematic review. *Pharmaceuticals* **2019**, *12*, 135. [[CrossRef](#)]
12. Langer, T.; Wolber, G. Pharmacophore definition and 3D searches. *Drug Discov. Today Technol.* **2004**, *1*, 203–207. [[CrossRef](#)] [[PubMed](#)]
13. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E.W. Computational methods in drug discovery. *Pharmacol Rev.* **2014**, *66*, 334–395. [[CrossRef](#)] [[PubMed](#)]
14. Comas, I.; Chakravartti, J.; Small, P.M.; Galagan, J.; Niemann, S.; Kremer, K. Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **2010**, *42*, 498. [[CrossRef](#)] [[PubMed](#)]
15. Kavvas, E.S.; Catoi, E.; Mih, N.; Yurkovich, J.T.; Seif, Y.; Dillon, N. Machine learning and structural analysis of *Mycobacterium tuberculosis* pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **2018**, *9*, 4306. [[CrossRef](#)]
16. Cubillos-Ruiz, A.; Morales, J.; Zambrano, M.M. Analysis of the genetic variation in *Mycobacterium tuberculosis* strains by multiple genome alignments. *BMC Res. Notes* **2008**, *1*, 110. [[CrossRef](#)]
17. Medini, D.; Donati, C.; Tettelin, H.; Maignani, V.; Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **2005**, *15*, 589–594. [[CrossRef](#)]
18. Naz, A.; Awan, F.M.; Obaid, A.; Muhammad, S.A.; Paracha, R.Z.; Ahmad, J. Identification of putative vaccine candidates against *Helicobacter pylori* exploiting exoproteome and secretome: A reverse vaccinology based approach. *Infect. Genet. Evol.* **2015**, *32*, 280–291. [[CrossRef](#)]
19. Pang, X.; Cao, G.; Neuenschwander, P.F.; Haydel, S.E.; Hou, G.; Howard, S.T. The  $\beta$ -propeller gene Rv1057 of *Mycobacterium tuberculosis* has a complex promoter directly regulated by both the MprAB and TrcRS two-component systems. *Tuberculosis* **2011**. [[CrossRef](#)]
20. Haydel, S.E.; Benjamin, W.H.; Dunlap, N.E.; Clark-Curtiss, J.E. Expression, autoregulation, and DNA binding properties of the *Mycobacterium tuberculosis* TrcR response regulator. *J. Bacteriol.* **2002**. [[CrossRef](#)]

21. Chakraborty, A.K.; Sarkar, I.; Sen, A. Herbal medicine meets bioinformatics for remedy of tuberculosis by *Mycobacterium tuberculosis* RGTB423. *Int. J. Data Min. Bioinform.* **2019**. [[CrossRef](#)]
22. Raman, K.; Yeturu, K.; Chandra, N. targetTB: A target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. *BMC Syst. Biol.* **2008**, *2*, 109. [[CrossRef](#)]
23. Ma, Y.; Stern, R.J.; Scherman, M.S.; Vissa, V.D.; Yan, W.; Jones, V.C. Drug targeting *Mycobacterium tuberculosis* cell wall synthesis: Genetics of dTDP-rhamnose synthetic enzymes and development of a microtiter plate-based screen for inhibitors of conversion of dTDP-glucose to dTDP-rhamnose. *Antimicrob. Agents Chemother.* **2001**, *45*, 1407–1416. [[CrossRef](#)]
24. Barry, C.E., III. New horizons in the treatment of tuberculosis. *Biochem. Pharmacol.* **1997**, *54*, 1165–1172. [[CrossRef](#)]
25. Gorla, P.; Plocinska, R.; Sarva, K.; Satsangi, A.T.; Pandeeti, E.; Donnelly, R. MtrA response regulator controls cell division and cell wall metabolism and affects susceptibility of mycobacteria to the first line antituberculosis drugs. *Front. Microbiol.* **2018**. [[CrossRef](#)]
26. Li, X.; Lv, X.; Lin, Y.; Zhen, J.; Ruan, C.; Duan, W. Role of two-component regulatory systems in intracellular survival of *Mycobacterium tuberculosis*. *J. Cell Biochem.* **2019**, *120*, 12197–12207. [[CrossRef](#)]
27. Rifat, D.; Belchis, D.A.; Karakousis, P.C. SenX3-independent contribution of regX3 to *Mycobacterium tuberculosis* virulence. *BMC Microbiol.* **2014**. [[CrossRef](#)]
28. White, D.W.; Elliott, S.R.; Odean, E.; Bemis, L.T.; Tischler, A.D. *Mycobacterium tuberculosis* Pst/SenX3-RegX3 regulates membrane vesicle production independently of ESX-5 activity. *MBio* **2018**, *9*. [[CrossRef](#)]
29. Freeman, Z.N.; Dorus, S.; Waterfield, N.R. The KdpD/KdpE two-component system: Integrating K<sup>+</sup> homeostasis and virulence. *PLoS Pathog.* **2013**, *9*, e1003201. [[CrossRef](#)]
30. Njoroge, J.W.; Gruber, C.; Sperandio, V. The interacting Cra and KdpE regulators are involved in the expression of multiple virulence factors in enterohemorrhagic *Escherichia coli*. *J. Bacteriol.* **2013**, *195*, 2499–2508. [[CrossRef](#)]
31. Parker, C.T.; Russell, R.; Njoroge, J.W.; Jimenez, A.G.; Taussig, R.; Sperandio, V. Genetic and mechanistic analyses of the periplasmic domain of the enterohemorrhagic *Escherichia coli* QseC histidine sensor kinase. *J. Bacteriol.* **2017**, *199*. [[CrossRef](#)]
32. Parish, T.; Smith, D.A.; Kendall, S.; Casali, N.; Bancroft, G.J.; Stoker, N.G. Deletion of two-component regulatory systems increases the virulence of *Mycobacterium tuberculosis*. *Infect. Immun.* **2003**, *71*, 1134–1140. [[CrossRef](#)]
33. Alegado, R.A.; Chin, C.-Y.; Monack, D.M.; Tan, M.-W. The two-component sensor kinase KdpD is required for *Salmonella typhimurium* colonization of *Caenorhabditis elegans* and survival in macrophages. *Cell Microbiol.* **2011**, *13*, 1618–1637. [[CrossRef](#)]
34. Xue, T.; You, Y.; Hong, D.; Sun, H.; Sun, B. The *Staphylococcus aureus* KdpDE two-component system couples extracellular K<sup>+</sup> sensing and Agr signaling to infection programming. *Infect. Immun.* **2011**, *79*, 2154–2167. [[CrossRef](#)]
35. Feinbaum, R.L.; Urbach, J.M.; Liberati, N.T.; Djonovic, S.; Adonizio, A.; Carvunis, A.-R. Genome-wide identification of *Pseudomonas aeruginosa* virulence-related genes using a *Caenorhabditis elegans* infection model. *PLoS Pathog.* **2012**, *8*, e1002813. [[CrossRef](#)]
36. Kornfeld, S.; Glaser, L. The enzymic synthesis of thymidine-linked sugars. I. Thymidine diphosphate glucose. *J. Biol. Chem.* **1961**, *56*, 184–185.
37. Brown, H.A.; Thoden, J.B.; Tipton, P.A.; Holden, H.M. The structure of glucose-1-phosphate thymidyltransferase from *Mycobacterium tuberculosis* reveals the location of an essential magnesium ion in the RmlA-type enzymes. *Protein Sci.* **2018**. [[CrossRef](#)] [[PubMed](#)]
38. Qu, H.; Xin, Y.; Dong, X.; Ma, Y. An rmlA gene encoding D-glucose-1-phosphate thymidyltransferase is essential for mycobacterial growth. *FEMS Microbiol. Lett.* **2007**. [[CrossRef](#)]
39. Kantardjieff, K.A.; Kim, C.-Y.; Naranjo, C.; Waldo, G.S.; Lakin, T.; Segelke, B.W. *Mycobacterium tuberculosis* RmlC epimerase (Rv3465): A promising drug-target structure in the rhamnose pathway. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2004**, *60*, 895–902. [[CrossRef](#)] [[PubMed](#)]
40. Chaudhari, N.M.; Gupta, V.K.; Dutta, C. BPGA- an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* **2016**, *6*, 24373. [[CrossRef](#)]

41. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [[CrossRef](#)]
42. Rizwan, M.; Naz, A.; Ahmad, J.; Naz, K.; Obaid, A.; Parveen, T. VacSol: A high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology. *BMC Bioinform.* **2017**, *18*, 106. [[CrossRef](#)] [[PubMed](#)]
43. Nazir, Z.; Afridi, S.G.; Shah, M.; Shams, S.; Khan, A. Reverse vaccinology and subtractive genomics-based putative vaccine targets identification for *Burkholderia pseudomallei* Bp1651. *Microb. Pathog.* **2018**, *125*, 219–229.
44. Luo, H.; Lin, Y.; Gao, F.; Zhang, C.-T.; Zhang, R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res.* **2014**, *42*, D574–D580. [[CrossRef](#)] [[PubMed](#)]
45. Chen, L.; Yang, J.; Yu, J.; Yao, Z.; Sun, L.; Shen, Y. VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res.* **2005**, *33*, D325–D328. [[CrossRef](#)] [[PubMed](#)]
46. Zhou, C.E.; Smith, J.; Lam, M.; Zemla, A.; Dyer, M.D.; Slezak, T. MvirDB—A microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* **2007**, *35*, D391–D394. [[CrossRef](#)] [[PubMed](#)]
47. Peterson, J.W. Bacterial pathogenesis. In *Medical Microbiology*, 4th ed.; University of Texas Medical Branch at Galveston: Galveston, TX, USA, 1996.
48. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. Protein identification and analysis tools on the ExpASY server. In *The Proteomics Protocols Handbook*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 571–607.
49. Chawley, P.; Samal, H.B.; Prava, J.; Suar, M.; Mahapatra, R.K. Comparative genomics study for identification of drug and vaccine targets in *Vibrio cholerae*: MurA ligase as a case study. *Genomics* **2014**, *103*, 83–93. [[CrossRef](#)]
50. Azam, S.S.; Shamim, A. An insight into the exploration of druggable genome of *Streptococcus gordonii* for the identification of novel therapeutic candidates. *Genomics* **2014**, *104*, 203–214. [[CrossRef](#)] [[PubMed](#)]
51. Moriya, Y.; Itoh, M.; Okuda, S.; Yoshizawa, A.C.; Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **2007**, *35*, W182–W185. [[CrossRef](#)]
52. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2017**, *46*, D1074–D1082. [[CrossRef](#)]
53. Mondal, S.I.; Ferdous, S.; Jewel, N.A.; Akter, A.; Mahmud, Z.; Islam, M.M. Identification of potential drug targets by subtractive genome analysis of *Escherichia coli* O157: H7: An in silico approach. *Adv. Appl. Bioinforma. Chem. AABC* **2015**, *8*, 49. [[CrossRef](#)]
54. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]
55. Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725. [[CrossRef](#)] [[PubMed](#)]
56. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinform.* **2008**, *9*, 40. [[CrossRef](#)] [[PubMed](#)]
57. Dallakyan, S.; Olson, A.J. Small-molecule library screening by docking with PyRx. In *Chemical Biology*; Humana Press: New York, NY, USA, 2015; pp. 243–250.
58. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [[CrossRef](#)]

**Publisher’s Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).