ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

In-Silico Analysis And Identification Of Novel Drugs For Multidrug Resistant *Mycobacterium tuberculosis* (H37RV) Applying Whole Genome Sequence (WGS) Data Analyses.

Karthikeyen TRS [1]*, and Puniethaa Prabhu[2].

[1]Department of Biopharmaceutical technology, Anna university, Guindy, Chennai-638102, Tamil Nadu, India.
[2]Department of Biotechnology, K S Rangasamy college of technology, Tiruchengode, Tamil Nadu, India.

ABSTRACT

Globally, tuberculosis is proving to be a continuous threat despite the efforts taken by WHO and other health care organizations all over the world. The major hindrance towards eradication of tuberculosis is the development of resistant Mycobacterium tuberculosis against the anti-tuberculosis drugs and there have been no much new drug molecules able to overhaul the effect of mutated targets. The discovery of the new drugs targets of the drug resistant strains of Mycobacterium tuberculosis proves to be the best way for complete eradication of the tuberculosis. Whole Genome Sequence Analyses performed over the entire genome generated Next Generation Sequencing platforms involves the identification of variants against the reference genome. Current work includes the whole genome analysis performed over the datasets of Mycobacterium tuberculosis (H37RV) that were retrieved from WGS Repository. Annotation of the variants predicts the functional effects of the variants responsible for resistance. The metabolically important variants that present in the most of the datasets were screened. The compounds are screened with swissADME server and selected based upon the various parameter such as solubility, Lipophilicity and Drug likeliness property. The compounds are then docked with the protein produced by the resistant gene, which gives the inhibitory results. **1178 genes** in H37RV strains of Mycobacterium tuberculosis are responsible for the variants. **43562 potential SNPs** responsible for the variants. Finally, **12 genes** were found to be novel drug targets. The best binding has the highest binding affinity of 14 kcal/mol. Nearly **238 docking** scores are above -8 kcal/mol. The best ligands are selected based upon the effectiveness over all the proteins. The **Escobarine A and sesquiterpenoids** are the best drugs against multidrug resistant tuberculosis have average binding affinity **of -8.892765957 and -8.918297872 kcal/mol** respectively over all 13 gene targets.

Keywords: *Mycobacterium tuberculosis,* H37RV, WGS, data analyses.

https://doi.org/10.33887/rjpbcs/2019.10.6.23

Corresponding author

ISSN: 0975-8585

INTRODUCTION

According to the survey of AFP and agencies in the year 2015, MDR-TB estimates 480,000 new TB cases (83%) and 250,000 deaths (12-17%) in the year 2015.Multi-drug resistant tuberculosis (MDR-TB) is a type of tuberculosis (TB) infection caused by bacteria *Mycobacterium tuberculosis* that are resistant to treatment with at least two of the most specific first-line anti-mycobacterium and anti-tuberculosis drugs such as Isoniazid and Rifampicin. The decrease in the cost and improvement of base-calling accuracy of the Datasets of Next generation sequencing will untouched the genetic architectures of complex diseases and develop the large opportunities for genomic medicine. The analysis of the generated data cascades the sequential procedure initiating with quality checking of the raw data, alignment of reads to the reference genome, discovery of variants followed by their annotation and lastly visualizing the data[2]. Variant calling from the NGS analysis is important to detect the functional effect of those variants developed due to mutations for further analysis[7]. The prediction of effects of the variants involves different methods such as region based analysis, evaluation of the structural effects on mutations and sequence based analysis[3]. The cross checking process involves the reporting of the variants that are responsible for the specific causative, taking annotations into considerations such that they are having an effect, also the annotations to the known function of a specific phenotype[6]. The genetic polymorphisms from WGS provides the adequate differential power to determine the natural variation involves in the drug susceptibility determinants virulence factors, and factors that are able to modify the immune functions[4]. Whole genome sequencing helps in the prediction of SNPs and INDELS in genomic DNA that may be conferring mutations in relation with resistance to anti-tuberculosis drug. WGS of *Mycobacterium tuberculosis* brings the new SNPs in genes into the light that are previously linked with drug resistance[5]. Now-a-days enormous amount of anti-tuberculosis compounds have been reported worldwide from numerous sources such as plant origin, marine source, animal origin, etc. They may be ineffective due to the Drug resistant mutation in the strain. Computer-aided docking is an important tool for gaining understanding of the binding interactions between a ligand (small molecule) and its mutated target receptor. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using scoring functions. It is a time-saving, cost effective and reliable technique for identifying new potent lead compounds for the Mutated drug targets. Therefore it can completely evade the Drug resistant problem in the Tuberculosis.

METHODS

Genomic variant discovery is much more complicated due to multiple sources of error: amplification biases during the sample libraries preparation in the wet lab, machine errors, software errors, and mapping artifacts when the reads are aligned. A good variant-calling workflow must involve data-preparation methods that correct the various error modes. Variant discovery can then be performed on the appropriately processed data, using a robust calling algorithm that leverages meta-information. At this stage, it is preferable to prioritize sensitivity over specificity. Once a highly sensitive call set has been generated, appropriate filters can be applied to achieve the desired balance between sensitivity and specificity. BWA and GATK are publicly available software packages that can be used to construct a variant-calling workflow following those principles. Specifically, BWA is a robust read-mapping algorithm that can be used to map sequence data to a reference genome. Its output can be used with minimal processing by utility software packages as input to the GATK tools. Identification of the variants present throughout the sample can provide a “eye-opener” Targets for the development of the novel drugs from the various plant resources. The drug targets network analysis will provide the functional correlation with the resistance. The Energy minimization and simulation should be done before the docking studies and the characteristics plots are generated. The draggability pockets are predicted using Dogsitescorer. The optimized ligands should be allowed to pharmacokinetic modules in order to reveal the ADME/T properties. The Docking is performed in Autodock Vina 1.1.2 and the best inhibitory results are collected based upon the docking confirmation.

RESULTS

Collection and quality control of Datasets

Totally fifteen metadata of the sequence files are found in the archives were downloaded in fastq file which contains both read and the Quality information of the reads. The reference genome *Mycobacterium tuberculosis* H37Rv for the mapping of raw reads was downloaded from NCBI GenBank as fasta file. A good read

ISSN: 0975-8585

should have a Phred quality score of above 30, per base content should be equal. Overrepresented sequences in the retrieved reads are not allowable and verified datasets are then taken into analysis.

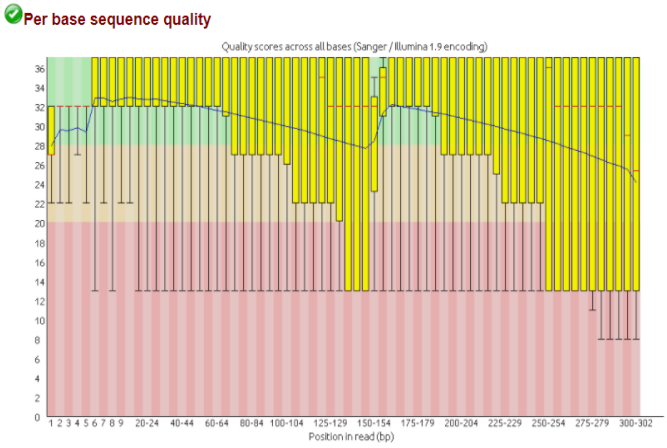


Figure 1: per base Sequence quality of the dataset

Mapping and removing the Duplicates

Mapping was done with Burrows Wheeler Alignment-MEM performs local alignment. The aligned SAM file which contains the details of alignment in tab delimited format and then converted to bam file to reduce the memory requirement, using Picard tools. Duplicate read should be flagged and removed by running MarkDuplicates on a sorted aligned bam file. The duplicate reads were removed if REMOVE_DUPLICATES was set true. A metrics file as shown in figure 2 contains the details of the reads that were detected and processed during the process is produced. It was found that from the metrics file that no duplicate reads were detected from all the datasets used.

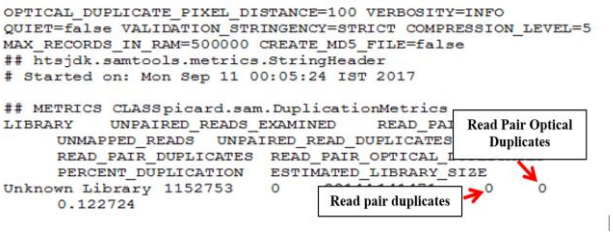


Figure 2 Metrics file produced after duplicates marking

Local realignment

Local realignment was done to remove the errors produced by the mapping algorithms to improve the downstream analyses that were to be done after this such as base recalibration and variant calling. As a first step for realignment regions where realignment to be done was determined using **RelignerTargetCreator** (GATK) which produced a file with the extension .intervals and it contains details of the regions where realignment need to be performed. These intervals was used by IndelRealigner GATK tool which realigns according to regions provided in the file and created a new realigned bam file and also indexed the new bam file. The new realigned bam file was viewed in IGV as shown in figure 3 and was confirmed that realignment had been done successfully.

ISSN: 0975-8585

Figure 3 Comparison of bam file before and after Realignment

Variant calling and Hard filtering of the variants

Variants were called using Unified genotyper a sophisticated tool that identify variants much better than previous variant callers developed by GATK team. The GATK version which was used in this work detected variants and provided them in Variant Call Format (VCF) version 4.2. The file begins with the meta information that will be provided in the main body need to be there, and other lines regarding such as file format, assembly field format, alternative allele field format, contig field format etc., starting with ##. After the meta information we have tab delimited details on the variants which provides information in the following order name of the chromosome, position of the reference genome, identifier ID if present or simply, reference base in the respective position, alternate base present in the query at the same position, Phred scaled quality for the alternate base call, status of the filter and information regarding the parameters. The index file for the .vcf file is created that helps in viewing the variants in IGV as shown in figure 4. The red arrow indicates the position of the mutation in the genome frame and the Information about the mutation is visualized in the pop up screen upon clicking of the variants.

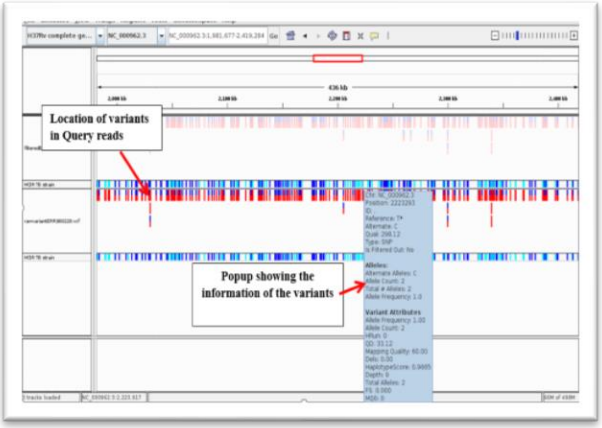


Figure 4 Variant file of the reads in IGV

Annotation of variants and visualization

The filtered individual variants was annotated by loading them in the online database GMTV (Genome-wide Mycobacterium tuberculosis Variation)

database. Once the annotation was completed the results can be viewed in the web page itself or can be downloaded as a table in text format as shown in the figure 1.5 which contains various information such as variation type, region of variation, NCBI's gene ID, name of the gene, locus tag, start and end position, KEGG ID, Metabolic pathway's KEGG name, function of the protein.

ISSN: 0975-8585

	Type	SNP/Indel disposition	Gene ID (NCBI)	Gene name	Locus tag	Start	E
2	"SNP"	"CDS"	"887089"	"recF"	"Rv0003"	"3280"	"4438"	"+"	"3446"	"GTC"
3	"SNP"	"CDS"	"887089"	"recF"	"Rv0003"	"3280"	"4438"	"+"	"4013"	"ACC"
4	"SNP"	"CDS"	"887081"	"gyrB"	"Rv0005"	"5240"	"7268"	"+"	"6361"	"AAA"
5	"SNP"	"CDS"	"887081"	"gyrB"	"Rv0005"	"5240"	"7268"	"+"	"6362"	"CTG"
6	"SNP"	"CDS"	"887081"	"gyrB"	"Rv0005"	"5240"	"7268"	"+"	"6364"	"TTC"
7	"SNP"	"CDS"	"887105"	"gyrA"	"Rv0006"	"7302"	"9819"	"+"	"7362"	"CAG"
8	"SNP"	"CDS"	"887105"	"gyrA"	"Rv0006"	"7302"	"9819"	"+"	"7582"	"GCC"
9	"SNP"	"CDS"	"887105"	"gyrA"	"Rv0006"	"7302"	"9819"	"+"	"7585"	"ACC"
10	"SNP"	"CDS"	"887105"	"gyrA"	"Rv0006"	"7302"	"9819"	"+"	"9304"	"GAC"
11	"SNP"	"CDS"	"887085"	"Rv0008c"	"Rv0008c"	"11874"	"12312"	"-"	"11879"	"
12	"SNP"	"CDS"	"887082"	"Rv0010c"	"Rv0010c"	"13133"	"13559"	"-"	"13368"	"
13	"SNP"	"CDS"	"887083"	"Rv0012"	"Rv0012"	"14089"	"14878"	"+"	"14785"	"

Figure 5 Annotation results downloaded from GMTV

Identification of drug targets from variants

From the list of annotated variants non synonymous variants that were present in the datasets were obtained by filtering out the synonymous SNPs. The genes which had undergone variation was compared with list of variants in genes that were responsible for resistance against the first and second line anti-tuberculosis drugs that was retrieved from TB Drug Resistant Mutation (TBDReaM) database (https://tbdreamdb.ki.se/). Variants that were involved in three or more metabolic pathways was chosen from all the datasets and were prioritized based on their repetition in each dataset based on the metabolic pathway and then comparing with genes present in Database of Essential Genes (DEG) is tabulated in the table 1.1.

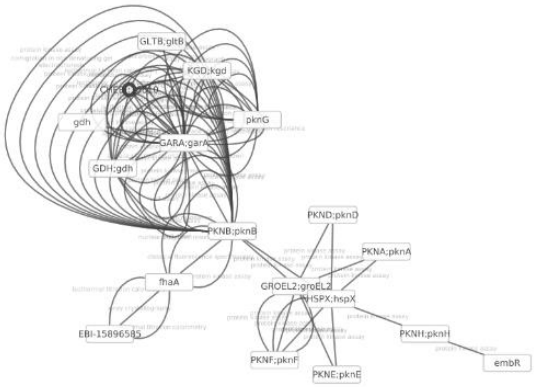
Table 1 Gene targets and its pathway

GENE	PATHWAYS INVOLVED	GENE	PATHWAYS INVOLVED
accD6	Lipid biosynthesis	pnCA	Antibiotic resistance
EmbC	Cell wall and cell processes	KatG	Biosynthesis of secondary metabolites
EmbR	Regulatory functions	KasA	Lipid biosynthesis
Fbpc	phospholipid synthesis	inhA	Lipid biosynthesis
ThyA	Polyamine synthesis	gyrA	DNA replication and repair

NETWORK ANALYSIS OF THE GENE TARGET

Cytoscape's software Core provides basic functionality to layout and query the network; to visually integrate the network with expression profiles, phenotypes, and other molecular states; and to link the network to databases of functional annotations. The KEGG database are installed as a plugin within the cytoscape. The Query genes are allowed to form the network with other genes based upon the pathway and expression profiles. The interaction between embR and its cascade is shown in the figure 6.

Figure 6 Network interaction between the embR gene and its co-expression



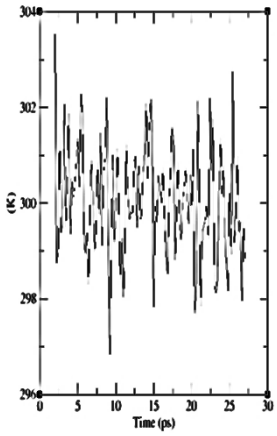
Energy minimization

The target protein was solvated and then energy minimized. The protein was minimized in water at 310K to obtain global energy minima because these are necessary to obtain a better model of entire receptor

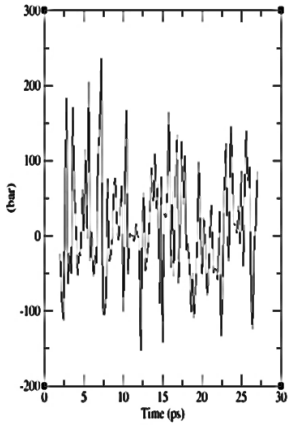
ISSN: 0975-8585

which can mimic the human physiological conditions. The various plots generated during the GROMACS are shown in the figure. The pressure, temperature and potential values of the various protein can\ be determined by performing GROMACS. The average value and the total drift from the initial position can be visualized using the terminal commands and the Plots generated upon the drift vs time is shown in the figure 7.

Temperature plot



Pressure plot



Root mean square difference (RMSD Plot)

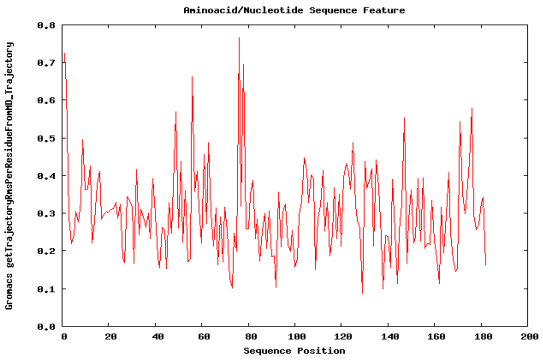


Figure 7 Plots generated during the energy minimization

DOCKING IN AUTODOCK VINA

The best binding affinity (> -10 kcal/mol) of the docking are tabulated after the docking was completed in the table 2. The interaction between the compounds and protein are visualized using autodock vina. The binding residues with the compounds present in the binding pockets are noted and the results are correlated with the pockets binding results calculated from the DOGSITESCORER online server. The best interaction between the PHE176 residue of the 4FQS coded by the mutated gene pncA and the novel drugs escobarine A & Sesquiterpenoids are shown in the figure 8

ISSN: 0975-8585

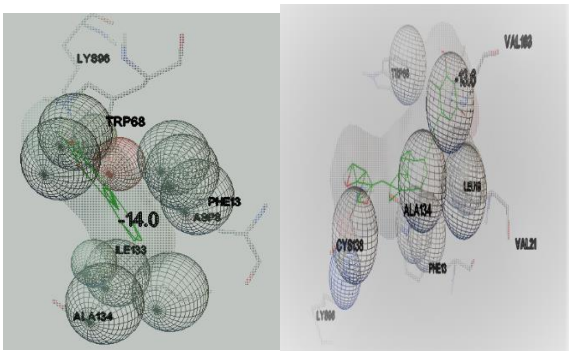


Figure 8 Docked conformation of the best docked ligands with its binding residues

DISCUSSION

Growing tuberculosis epidemic poses a much risk globally and the need to identify new innovations that could help to overcome the disease is on the increase. KEGG database information provided with annotation results was used to screen the important variant genes. **1178 genes** in H37RV strains of Mycobacterium tuberculosis are responsible for the variants. **43562 potential SNPs** responsible for the variants. Finally, **13 genes** were found to be novel drug targets. The best binding has the highest binding affinity of 14 kcal/mol. Nearly **238 docking** scores are above -8 kcal/mol. The best ligands are selected based upon the effectiveness over all the proteins. The **Escobarine A and sesquiterpenoids** are the best drugs against multidrug resistant tuberculosis have average binding affinity **of -8.892765957 and -8.918297872 kcal/mol** respectively over all 13 gene targets. These drugs can be best against the multidrug resistant strains of Mycobacterium tuberculosis H37RV.

REFERENCES

[1] Chung-Delgado, K., S. Guillen-Bravo, A. Revilla-Montag (2015) "Mortality among MDR-TB cases: comparison with drug-susceptible tuberculosis and associated factors." PLoS One 10(3): e0119332

[2] Asho Ali, Zahra Hasan, Ruth McNerney, Kim Mallard (2015) “Whole Genome Sequencing Based Characterization of Extensively Drug-Resistant Mycobacterium tuberculosis Isolates from Pakistan”, *PLoS ONE*, Vol.10, No.2, pp. 1-17.

[3] Cole S.T., Brosch R., Parkhill J (1998) “Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence”, *Nature*, Vol. 393, pp. 537–440

[4] David Stucki and Sebastien Gagneux (2013) “Single nucleotide polymorphisms in Mycobacterium tuberculosis and the need for a curated database”, *Tuberculosis*, Vol. 93, No. 1, pp. 30-39.

[5] Ekaterina N Chernyaeva, Marina V. Shulgina, Mikhail S. Rotkevich (2014) “Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology” *BMC Genomics*, Vol. 15, No. 1, pp. 68-75.

[6] Elizabeth A. Worthey (2013) “Analysis and Annotation of Whole-Genome or Whole-Exome Sequencing–Derived Variants for Clinical Diagnosis”, *Current Protocols in Human Genetics*, Vol. 79, pp. 9.24.1-9.24.24.

[7] Francesc Coll, Mark Preston, Jose Afonso Guerra-Assuncao (2014) “PolyTB: A genomic variation map for Mycobacterium tuberculosis”, *Tuberculosis*, Vol. 94, No. 3, pp. 346-354.