BMC Microbiology

RESEARCH ARTICLE Open Access

Correlation of drug resistance with single nucleotide variations through genome analysis and experimental validation in a multi-drug resistant clinical isolate of *M. tuberculosis*

Kausik Bhattacharyya[1,2], Vishal Nemaysh[3], Monika Joon[1], Ramendra Pratap[3], Mandira Varma-Basil[2], Mridula Bose[2] and Vani Brahmachari[1*]

Abstract

Background: Genome sequencing and genetic polymorphism analysis of clinical isolates of *M. tuberculosis* is carried out to gain further insight into molecular pathogenesis and host-pathogen interaction. Therefore the functional evaluation of the effect of single nucleotide variation (SNV) is essential. At the same time, the identification of invariant sequences unique to *M. tuberculosis* contributes to infection detection by sensitive methods. In the present study, genome analysis is accompanied by evaluation of the functional implication of the SNVs in a MDR clinical isolate VPCI591.

Result: By sequencing and comparative analysis of VPCI591 genome with 1553 global clinical isolates of *M. tuberculosis* (GMTV and tbVar databases), we identified 141 unique strain specific SNVs. A novel intergenic variation in VPCI591 in the putative promoter/regulatory region mapping between *embC* (*Rv3793*) and *embA* (*Rv3794*) genes was found to enhance the expression of *embAB*, which correlates with the high resistance of the VPCI591 to ethambutol. Similarly, the unique combination of three genic SNVs in RNA polymerase β gene (*rpoB*) in VPCI591 was evaluated for its effect on rifampicin resistance through molecular docking analysis.
The comparative genomics also showed that along with variations, there are genes that remain invariant. 173 such genes were identified in our analysis.

(Continued on next page)

* Correspondence: vani.brahmachari@gmail.com
[1]Dr. B. R. Ambedkar Center for Biomedical Research (ACBR), University of Delhi, 110007, New Delhi, India
Full list of author information is available at the end of the article

(Continued from previous page)
Conclusion: The genetic variation in *M. tuberculosis* clinical isolate VPCI591 is found in almost all functional classes of genes. We have shown that SNV in *rpoB* gene mapping outside the drug binding site along with two SNVs in the binding site can contribute to quantitative change in MIC for rifampicin. Our results show the collective effect of SNVs on the structure of the protein, impacting the interaction between the target protein and the drug molecule in *rpoB* as an example. The study shows that intergenic variations bring about quantitative changes in transcription in *embAB* and in turn can lead to drug resistance.

Keywords: *Mycobaterium tuberculosis*, Clinical isolate, Single nucleotide variations, rpoB, Drug resistance, Ethambutol, Rifampicin, Structural analysis, Expression analysis

Background

In spite of the worldwide efforts to combat mycobacterial diseases, it continues to be a great challenge to achieve this goal. In addition to the various strategies adopted by this pathogen to escape host immune system, *M. tuberculosis* has gainfully utilized genetic variability for its highly successful growth, pathogenesis, immunity and persistence [1]. The complete genome sequence of the strain of *M. tuberculosis* H37Rv and the recent surge in data on clinical isolates permits high-throughput whole-genome analysis, relationship and correlation with drug resistance [2–9]. This has revealed local, global and patient specific heterogeneity in *M. tuberculosis* strains [10]. The increased genetic diversity and variation in *M. tuberculosis* is implicated in strain specific immunogenicity and pathogenicity [11]. A large number and the most frequent occurrence of variation is seen in the genes for lipid metabolism and *PE*/*PPE* genes [12, 13].

The whole genome sequence of more than 11,000 clinical isolates of *M. tuberculosis* are reported in NCBI-SRA, however, the correlation between genetic alterations and the phenotype is carried out mainly in the drug resistance genes [14]. Thus, drug resistance in tuberculosis is a phenomenon more complex than previously assumed where the association of intergenic regions and the identification of new genes have resulted from whole-genome sequence based approach [15].

We analysed the genetic polymorphism in an Indian, multi-drug resistant (MDR) clinical isolate VPCI591, from a 50 year old male patient from Vallabhbhai Patel Chest Institute (VPCI), Delhi. It is TBd-1 positive and EAI clade spoligotype and the isolate is resistant to all the first-line tuberculosis drugs [16, 17]. The availability of this strain made it possible to validate selected variations and investigate their effect on the function of the gene. We carried out the comparative analysis of the variations detected in VPCI591 with the genome sequence of 1553 global clinical isolates, to identify the shared and unique Single Nucleotide Variations (SNV). This analysis also indicated the invariant regions/genes that could be of potential use as markers for *M. tuberculosis* complex.

Results

Sequencing of VPCI 591 genome

The sequencing depth of coverage was calculated to be 133X. The FastQC data analysis showed high quality reads after adapter removal. The quality score distribution of all the reads and GC nucleotide distribution of the reads are shown in (Additional file 1:Fig. 1a & b). The distribution pattern shows typical pattern of GC rich mycobacterial genomes. The mean quality score distribution is also high (Additional file 1: Fig. 1c).

To increase the confidence in the identification of SNVs, four different pipelines were used in data analysis, namely Maq (0.6.6), Bowtie 2 (2.0.4) for variant detection in massively parallel sequencing data, in combination with Samtools, GATK (3.6–0-g89b7209) and Picard Tools (1.119), Varscan and Mpileup of Samtools and inGAP as given under methods. The consensus of the four pipelines was considered to call the **S**ingle **N**ucleotide **V**ariations (SNV) in the genome of VPCI591. We compared VPCI591 with global datasets of variation totalling to 1553, obtained from GMTV (http://mtb.dobzhanskycenter.org) [18] and tbVar database (http://genome.igib.res.in/tbVar/) [19].

Analysis of single nucleotide variations. Distribution of genic variations

A total of 1917 SNVs were chosen as consensus variations with high confidence from the four pipelines as described under methods and were considered for further analysis (Fig. 1a). 1685 genic and 232 non genic/intergenic variations were obtained. 1059 NS, 596 SY, 18 SG and 12 SL variations were there distributed among 1685 SNVs (Fig. 1b).

We identified 1685 genic SNVs in VPCI591 distributed among 1245 genes (Fig. 2a). We have found 926 genes having single variation among which 737 are NS (Fig. 2a, b). The 1059 NS variations found in VPCI591 are distributed in 894 genes (Fig. 2b). It is known that the PE-PPE genes are highly variant among clinical isolates [20]. In VPCI591, among the PPE genes, *Rv3347c* has 8 SNVs out of which 5 are non-synonymous SNV and *Rv0355c*

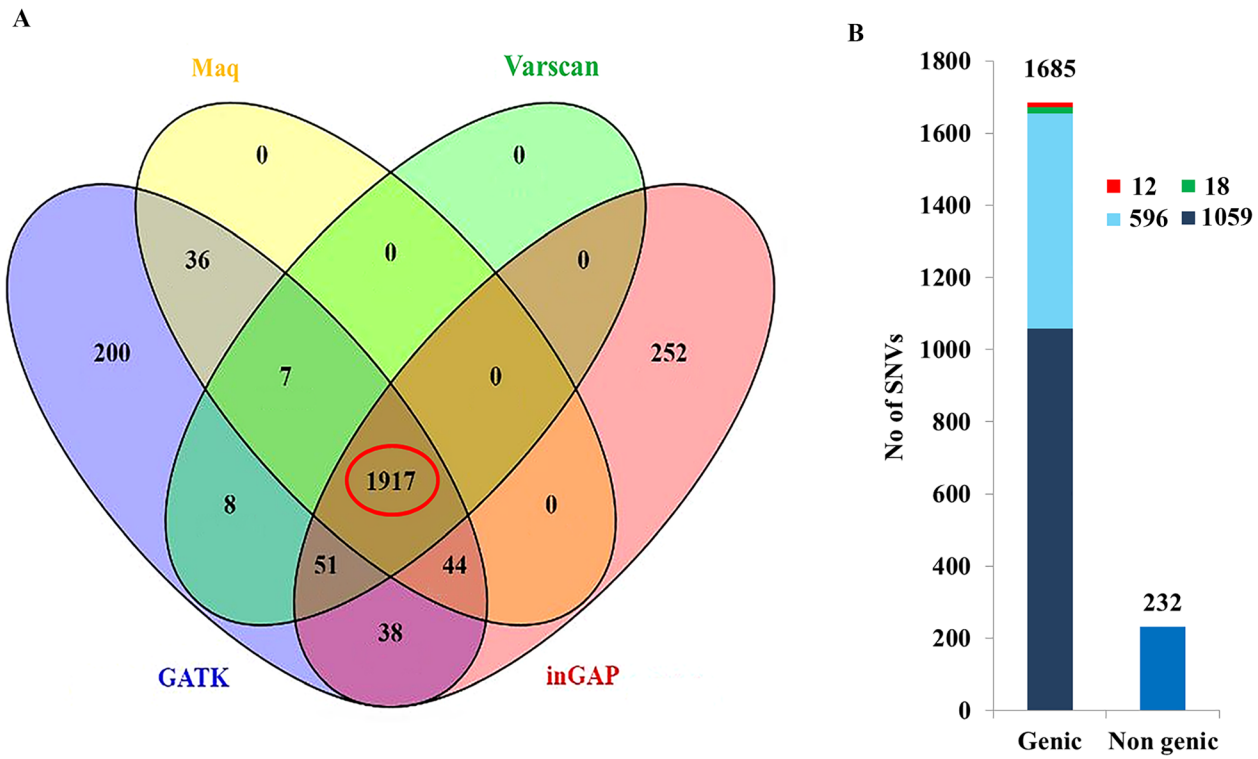


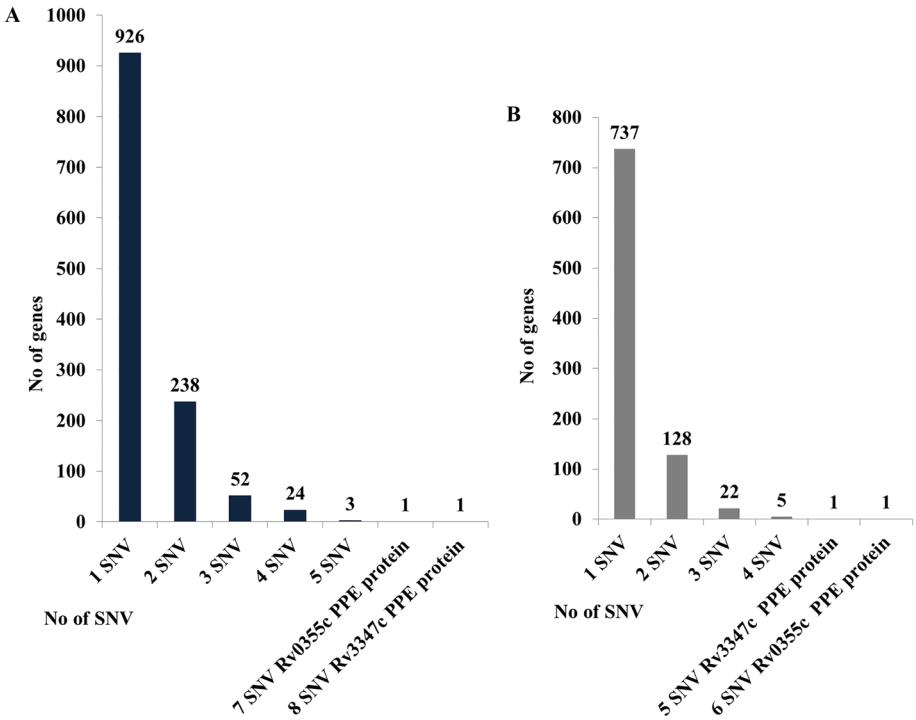
Fig. 1 Single Nucleotide variations in VPCI591. **a**. VENN diagram showing the number of SNVs called through the different software and the red circle indicates the consensus. **b**. Total number of variations and their distribution is represented. The Y axis shows the number of variations, the distribution of the 1917 SNVs along with their genic and non-genic distribution and subcategories are shown, stop gain (green), stop-loss (red), synonymous (cyan), non-synonymous (dark blue)

has 7 SNVs, 6 of which are NS (Fig. 2a, b). VPCI591 contains multiple non-synonymous variations which are also seen in 11 genes known for drug resistance and multiple variations are identified in *gyrA*, *embC*, *embB* and *rpoB*. Based on SIFT score (< 0.05), several variations were predicted to be deleterious; for example A384V in *gyrA*, I21T in *inhA*, N394D in *embC*, P913S in *embA*, D12A in *pncA* and all the 3 variations in *rpoB* (Fig. 2c).

The genetic variations were found in genes responsible for resistance to other drugs such as, fluoroquinolones, ethionamide, isoniazid, ethambutol, pyrazinamide and rifampicin (Fig. 2c). RNA polymerase β subunit (*rpoB)* harbours 3 variations known to cause resistance to rifampicin (Fig. 2c). We found similar variations in 1553 global clinical isolates of *M. tuberculosis*, that we analysed, however the co-occurrence of all the three is unique to VPCI591. In addition, based on SIFT score, we can predict deleterious effect of the SNVs in the following genes; *mmpL4* (*Rv0450c)* and *mmpL8* (*Rv3823c*) involved in membrane transport, *proC* (*Rv0500*, pyrroline-5-carboxylate reductase), *trcS* (sensor histidine kinase, *Rv1032c*), *cyp121* (*Rv2276*, cytochrome P450 and polyketide synthases *pks2* (*Rv3825c*) and *pks7* (*Rv1661*). VPCI591 has SNVs in the two component system genes, *mprA* (*Rv0981*) and *mprB* (*Rv0982*). We have investigated its effect on the expression of immune response genes and also on phago-lysosome fusion [21]. A complete list of NS variations with gene names, position of SNV, genomic location and amino acid (AA) change is given in (Additional file 2).

Functional classification of genes with non-synonymous SNV

The classification of genes containing SNV according to Camus et al., [4] and PANTHER [22] was carried out (Fig. 3a). Classification based on Camus et al., [4] identified 219 genes of intermediary metabolism, respiratory pathway genes, 200 genes involved in cell wall and cell processes, 49 PE and PPE proteins, 65 genes for lipid metabolism of Mycobacterium, 45 regulatory proteins and 11 insertion elements, and phage proteins (Fig. 3a). A large number of genes (211) are annotated as conserved hypothetical or uncharacterized proteins (Fig. 3a). The genes with NS variations (894) were classified according to their molecular function using PANTHER [22]. Among the mapped genes, 290 genes are associated with catalytic activity, 56 genes with binding function and 48 genes having transporter activity (Fig. 3b). On classifying the catalytic activity genes further, it was



C

Gene Name	Rv Id	Nucleotide Change	Amino Acid Change	Known Drug	Deleterious through SIFT
gyrB	Rv0005	G990C	M330I	FLQ	
gyrA	Rv0006	G61C	E21Q	FLQ	
gyrA	Rv0006	G284C	S95T	FLQ	
gyrA	Rv0006	C1151T	A384V	FLQ	Yes
gyrA	Rv0006	G2003A	G668D	FLQ	
ethA	Rv3854c	C1035A	N345K	Ethionamide	
inhA	Rv1484	T62C	I21T	Isoniazid	Yes
katG	Rv1908c	G1388T	R463L	Isoniazid	
embC	Rv3793	C809T	T270I	Ethambutol	
embC	Rv3793	A1180G	N394D	Ethambutol	Yes
embA	Rv3794	C2737T	P913S	Ethambutol	Yes
embB	Rv3795	A1133C	E378A	Ethambutol	
embB	Rv3795	A1512C	E504D*	Ethambutol	
embR	Rv1267c	G329A	C110Y	Ethambutol	
pncA	Rv2043c	A35C	D12A	Pyrazinamide	Yes
rpoB	Rv0667	A1304C	D441A	Rifampicin	Yes
rpoB	Rv0667	T1355C	L458P	Rifampicin	Yes
rpoB	Rv0667	C3318G	I1112M*	Rifampicin	Yes

Fig. 2 Distribution of Genic SNV in VPCI591. **a**. The overall distribution of genic SNVs. **b**. Distribution non-synonymous (NS) SNVs. **c**. Variation in genes known for drug resistance. FLQ: fluoroquinolones, * Represents variations unique to this clinical isolate

observed that 109 of these genes have transferase activity, 88 genes have hydrolase activity and 80 genes have oxido-reductase activity (Fig. 3c). Among the 109 genes for transferase activity, 40 genes are involved in transferring acyl group, 31 genes have methyltransferase activity, 12 are glycosyl group transferases and 10 have kinase activity (Fig. 3d). The genes in the following classes were under represented among the SNV containing genes: structural constituents of ribosome (GO:0003735), DNA-binding transcription factor activity (GO:0003700), transcription coregulator activity (GO:0003712) and guanyl-nucleotide exchange factor activity (GO:0005085).

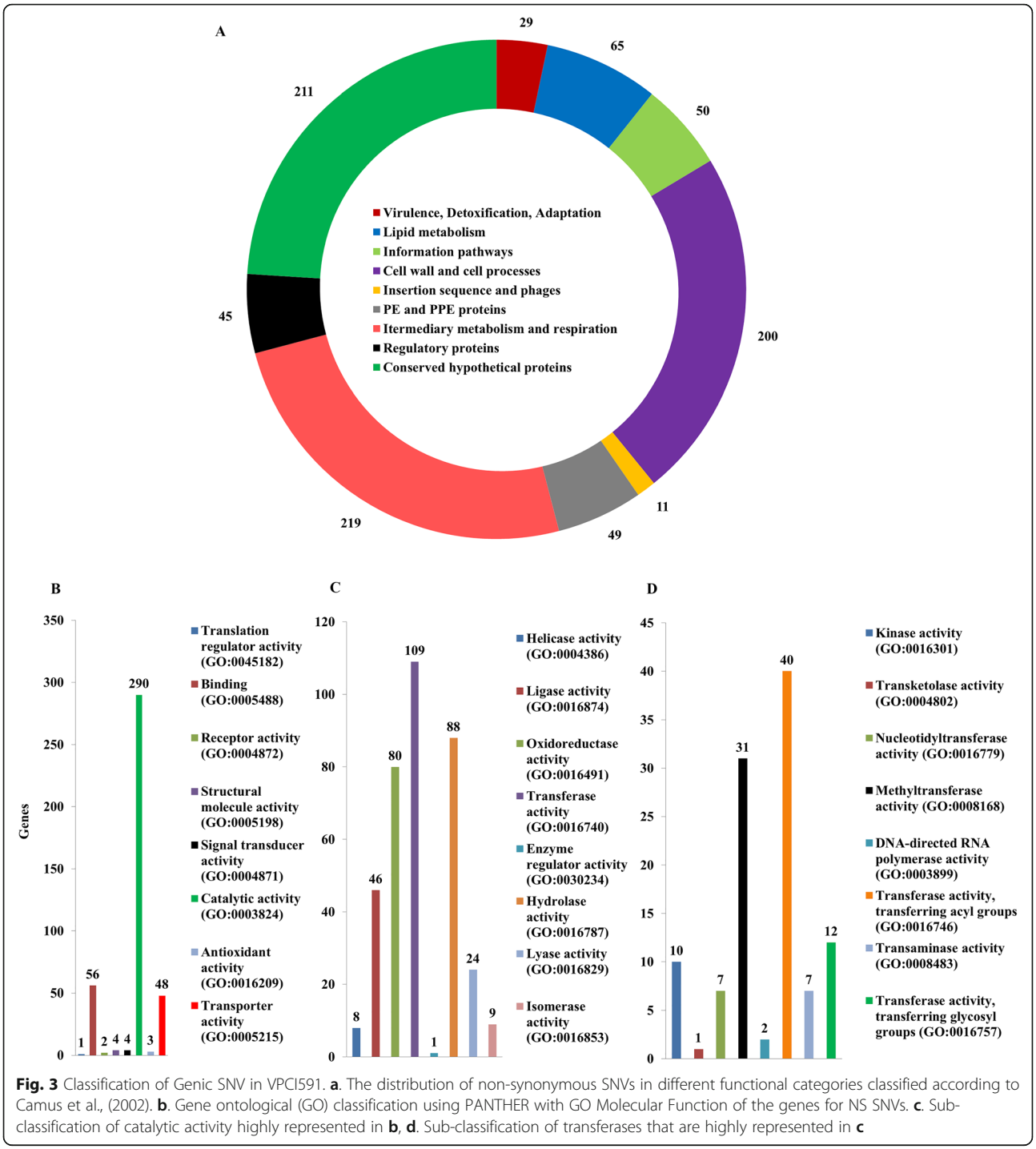


Fig. 3 Classification of Genic SNV in VPCI591. **a**. The distribution of non-synonymous SNVs in different functional categories classified according to Camus et al., (2002). **b**. Gene ontological (GO) classification using PANTHER with GO Molecular Function of the genes for NS SNVs. **c**. Sub-classification of catalytic activity highly represented in **b**, **d**. Sub-classification of transferases that are highly represented in **c**

SNV leading to loss/gain of stop codons

The SNVs that result in a premature stop codon (Stop-Gain, SG) or a loss of stop codon resulting in longer transcripts (Stop-Loss, SL) were identified. The SG variations were identified in 18 genes including ABC transporter *pstA1* (*Rv0930*), dihydroorotate dehydrogenase *pyrD* (*Rv2139*), fatty-acid-CoA synthase *fadD15* (*Rv2187*), NADPH: adrenodoxin oxidoreductase *fprA* (*Rv3106*) and NAD(P) H quinone reductase *lpdA* (*Rv3303c*) (Table 1). In these cases we identified the domains lost due to which the loss of function is predicted, which was correlated with the loss of a long peptide, including the loss of the functional domain. For example, in *Rv2187* (*fadD15*), 558 amino acids are lost which results in the loss of AMP synthetase/ligase domain. Since many of these genes are annotated as hypothetical

Table 1 Distribution of stop-gain variation in VPCI591

Genomic Position	Gene Id	Gene Name	Nucleotide Change	Amino acid changing to stop codon	No. AA lost	Function	Domains lost
162,226	Rv0134	ephF	G456A	W152X	148	Possible epoxide hydrolase EphF	Hydrolase domain
234,477	Rv0197	Rv0197	T2247G	Y749X	14	Possible oxidoreductase	
309,765	Rv0257	Rv0257	C67T	R23X	102	Hypothetical protein	
549,251	Rv0457c	Rv0457c	G357A	W119X	555	Probable peptidase	Oligopeptidase domain
704,997	Rv0610c	Rv0610c	C913T	Q305X	81	Hypothetical protein	
1,009,490	Rv0906	Rv0906	C547T	Q183X	190	Hypothetical protein	
1,037,911	Rv0930	pstA1	C913T	R305X	3	Probable phosphate-transport integral membrane ABC transporter PstA1	
2,399,782	Rv2139	pyrD	C1063T	Q355X	2	Probable dihydroorotate dehydrogenase	
2,448,288	Rv2187	fadD15	G129A	W43X	558	Actelyco asynthetase like	AMP synthetase ligase
2,882,317	Rv2563	Rv2563	C28T	Q10X	340	Probable glutamine-transport transmembrane ABC transporter	Cytoplasmic domain
2,929,354	Rv2601	speE	C967T	Q323X	202	Spermidine synthatase	
2,936,497	Rv2608	PPE42	C1452A	Y484X	97	PPE family protein PPE42	PPE C-terminal domain
3,097,349	Rv2788	sirR	C391T	Q131X	130	Probable transcriptional repressor SirR	
3,351,472	Rv2994	Rv2994	G204A	W68X	377	Probable integral membrane protein	
3,378,720	Rv3019c	esxR	G282A	W94X	3	Secreted ESAT-6 like protein EsxR	
3,442,240	Rv3079c	Rv3079c	G358T	E120X	156	Hypothetical protein	Luciferase like domain
3,689,523	Rv3303c	lpdA	C1416A	C472X	22	NAD(P) H quinone reductase LpdA	
4,351,039	Rv3872	PE35	G295T	E99X	1	PE family-related protein PE35	

proteins, the nature of functional deficiency could not be predicted. Among the 12 stop-loss variations, one of the SNVs results in loss of stop codon at 1467 A > C position in the fatty acid pathway gene polyketide synthase, *pks3* (*Rv1180*) which can now be read through into the adjacent gene, *pks4* (*Rv1181*) to form a longer transcript. This is annotated as *msl3*. Some of the other genes showing stop loss variations are *PPE33* (*Rv1809*), *PPE67* (*Rv3739c*) and epoxide hydrolase e*phF* (*Rv0134*). The predicted new function based on BLAST and InterProScan in VPCI591 due to SL variations are shown with the number of AA added (Table 2). *Rv0325* gains 155 amino acids to form class I SAM dependent methyl transferase. In *Rv1870*, 11 AA were added and it is converted in to an endonuclease.

Variation in intergenic region between divergently transcribed genes

There are several intergenic regions that are flanked by divergently transcribed genes. The presence of promoters or regulatory regions within such intergenic noncoding regions are known in *M. tuberculosis* [23]. Thus intergenic SNVs can affect the transcription/regulation of the downstream gene. We have mapped the intergenic variations occurring upstream of genes and also those mapping between divergently transcribed genes (Fig. 4a; I and II) and computed the relative distance between the SNV position and the gene(s). We have identified a total of 232 intergenic variations, 32 of them are positioned between genes transcribing in opposite directions and these have the potential to affect both the genes. These were binned into categories depending upon the distance between the downstream genes (Fig. 4b). There are 77 variations that map within 50 bp from transcription start site, 39 SNVs between 50 and 100 bp region and 35 SNVs between 100 and 150 bp regions indicating that majority of the intergenic variations are close to the transcription start site. Such variations in putative promoter or regulatory variations can affect both the flanking genes transcribed in opposite directions. Some of the genes whose transcription could be affected by the SNV in VPCI591 were identified (Fig. 4c). The complete list of the intergenic variations is given in (Additional file 3 and Additional file 4) respectively. We found a variation within the 85 bp intergenic region between *embC* and *embAB*, 4 bp upstream of

Table 2 Distribution of stop-loss variation in VPCI591

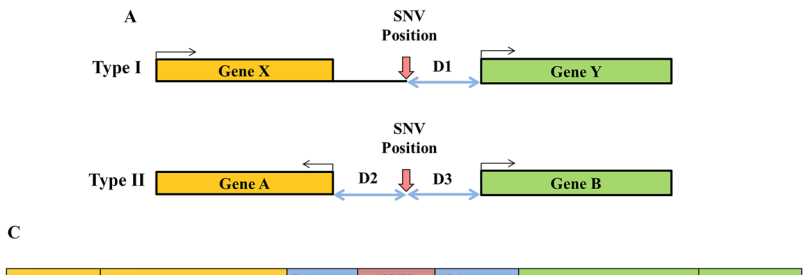
Genomic Position	Gene Id	Gene Name	Nucleotide Change	Stop codon changing to Amino Acid	Function	AA added	Predicted function[a]
392,261	Rv0325	none	T223C	X75Q	Hypothetical protein	155	Class I SAM-dependent methyltransferase
715,266	Rv0621	none	A1065G	X355W	Possible membrane protein	31	
932,280	Rv0836c	none	A653G	X218W	Hypothetical protein	23	
1,315,191	Rv1180	pks3	A1467C	X489Y	Probable polyketide beta-ketoacyl synthase Pks3	1597	Msl3 (mass like protein3)
1,694,547	Rv1504c	none	T598G	X200E	Hypothetical protein	182	
2,020,563	Rv1783	none	A1307T	X436L	ESX-5 type VII secretion system protein	15	
2,052,687	Rv1809	PPE33	G1406C	X469S	PPE protein	1	
2,120,796	Rv1870c	none	A635T	X212L	Hypothetical protein	11	Endonuclease
3,007,238	Rv2690c	none	T1972C	X658R	Integral membrane alanine and valine and leucine rich protein	6	
3,788,365	Rv3373	echA18	T640G	X214G	Probable enoyl-CoA hydratase EchA18	83	
4,190,285	Rv3739c	PPE67	A233G	X78W	PPE67	18	
4,383,655	Rv3898c	none	T331C	X111Q	Hypothetical protein	215	

[a]Predicted function of the longer protein resulting from loss of stop codon due to SNV

transcription start site of *embA*. The evaluation of the functional significance of this variation is discussed later.

Unique variations in VPCI591 and invariant genes in clinical isolates of *M. tuberculosis*

We found several invariant genes among the 1553 clinical isolates and also VPCI591. This includes genes coding for secretory proteins (*Rv3875*/ESAT-6), EsxM), ribosomal proteins (RpmJ, RpmH, RpsJ, RpsH), toxin-antitoxin pathway proteins (VapB3, VapB26, VapC7, VapB10), several transposase for IS elements (*Rv2278*, *Rv2354*, *Rv2355*) and many putative PhiRv2 prophage proteins (Rv2654c, Rv2656c, Rv2657c) known in *M. tuberculosis.* Interestingly there are several PE, PPE and PGRS genes identified like *Rv2099c* (PE21), *Rv1195* (PE13), *Rv2431c* (PE25), *Rv3622c* (PE32), *Rv2098c*



C

Gene_name	Function	Distance (bp)	SNV Pos	Distance (bp)	Function	Gene_name
Rv0193c	Hypothetical	105	226676	201	Multidrug efflux pump	Rv0194
Rv0439c	Probable dehydrogenase	40	528354	253	Chaperonin 2 groEL-2 hsp65	Rv0440
Rv0653c	Transcriptional regulator tetr	39	749968	31	Probable dioxygenase	Rv0654
Rv0853c	Pyruvate decarboxylase	24	951142	40	Hypothetical	Rv0854
Rv0876c	Hypothetical	106	976043	31	Hypothetical	Rv0877
Rv0966c	Hypothetical	86	1077921	53	Cu sensitive operon repressor	Rv0967
Rv1776c	Transcriptional regulatory	59	2010614	41	Cytochrome P450	Rv1777
Rv1872c	Lactate dehydrogenase	18	2123169	4	Hypothetical	Rv1873

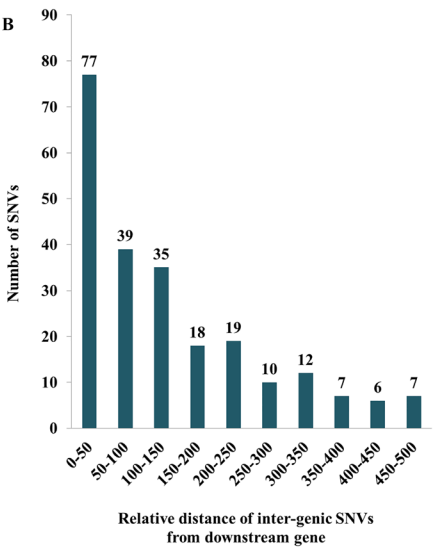


Fig. 4 Intergenic SNV in VPCI591. **a**. Schematic representation of intergenic variations leading to type I and type II configuration with reference to the direction of transcription of the adjacent genes. SNV position is indicated with vertical red arrows. The two-way arrows (D1, D2, D3), show the distance between the SNV and the downstream genes, bent arrows indicate direction of transcription. **b**. Distribution of non-genic SNVs relative to the distance from the downstream gene (Type I and II) **c**. The list flanking genes and their relative distance (D2 and D3) from the intergenic SNVs in Type II configuration

(PGRS36) proteins that do not vary among the genomes we compared (1553). The complete list of the invariant genes is shown in (Additional file 5).

The comparative analysis with global clinical isolates, led to the identification of 141 unique variations in VPCI591, among which 125 are genic and 16 are intergenic. Most of the genic variations were non-synonymous and stop-gain SNVs. Ontology classification revealed that majority of them fall under genes with catalytic function (75%) and transportation (17%). The complete list of the unique genes with amino acid alteration along with the SNV position in the genome is shown in (Additional file 6). A number of intergenic variations lie within the known promoter/regulatory regions. Intergenic variation in the cis-regulatory sequences of mce1 operon between *Rv0166* (fadD5) and *Rv0167* (yrbE1A) were identified [23]. SNV was identified within IS 1560 elements and mutation hotspot for anti-tuberculosis drug ethambutol, between *embC* and *embAB* gene. The complete list of unique intergenic variations identified in VPCI591 along with their genomic position and the adjacent genes is shown in (Additional file 7).

The effect of SNV in RpoB subunit and its possible implication

As mentioned earlier, 3 variations co-occurring in RpoB (*Rv0667*) at D441A, L458P and I1112M which is unique to VPCI591. We examined the effect of these on the structure of RpoB and its interaction with rifampicin. To understand the binding of rifampicin at the active site of RpoB, molecular docking analysis was performed using the crystal structure of RpoB (PDB Id: 5UHB). The locations of the 3 variations are mapped in the crystal structure at D441, L458 and I1112 (Additional file 8: Fig. A). The wild type protein showed His451, Phe439, Gln438, Arg454, Arg465, Ser456, Asn493 interacting with the drug rifampicin at the active site (Fig. 5a, c). Each variation was examined for its effect separately as well as in combination with each other; D441A showed loss of all the interactions with drug except Gly438, Phe439 and Arg465 (Additional file 8: Fig. B & C) and, while L458P also showed loss of all interactions with drug except Gln438, His441, Phe439 and Arg465 (Additional file 8: Fig. D and E). The variation I1112M showed loss of interactions with drug except Gln438, Phe439 and Arg465 (Additional file 8: Fig. F & G). The combination of all the three variations, D441A, L458P and I1112M as present in the clinical isolate VPCI591, showed loss of all interactions with the drug retaining the interaction with only with Arg454 (Fig. 5b & d). The binding energy and the respective Ki values for the 3 variations were calculated for individual variations. The wild type shows (ΔG) − 9.27 kcal/mol while the variants D441A, L458P and I1112M have − 6.67 kcal/mol, − 7.11 kcal/mol and − 7.18 kcal/mol respectively. The combination of the three, D441A + L458P + I1112M has ΔG equal to − 5.82 kcal/mol showing loss of binding of rifampicin at the active site. This could result in drug resistance. The Ki values calculated from the wild type is 160 nM whereas RpoB from VPCI591 shows 53.96 μM. The complete representation of the individual variations as well as the combined, the number of hydrogen bonding and the relative distances in Å is shown in (Table 3).

Effect of SNV in intergenic region upstream of *embA* (*Rv3794*) gene

Several NS variations were identified with the associated genes for ethambutol resistance in *embC* (C809T, A1180G), *embA* (P913S), *embB* (E378A, E504D) (Fig. 6a). We found NS variation at *embR* (C110Y) which is not a part of the operon. In addition to the genic variations, an intergenic SNV at position 4,243,299 C > T at -4 bp position upstream to *embA* gene is identified in VPCI591. The SNV was confirmed by PCR followed by Sanger sequencing. The expression level of *embAB* was compared with that of the reference strain H37Rv by quantitative PCR, using expression of *sigA* for normalization. No SNV is identified in *sigA* in VPCI591. There is 3.6 fold increases in the expression of *embA* gene in VPCI591 as compared to H37Rv in log phase and 1.25 fold increase in expression in the stationary phase of growth, indicating a gain-of-function due to the intergenic variation (Fig. 6b).

Discussion

The clinical isolate VPCI591, is a multidrug resistant strain that has been used in various analysis [16, 23]. The sensitivity of VPCI591 for first line anti-tuberculosis drugs is tested and it shows high MIC; INH-(MIC > 300 mg/L), RIF-(MIC > 125 mg/L), STR-(MIC 40 mg/L), EMB-(MIC 15 mg/L [17];. We have earlier carried out a directed polymorphism analysis of the *mce* 1 operon of VPCI591 and identified a gain-of-function SNV at 196800 (G > C) in the intergenic region between *Rv0166-Rv0167* [23]. This strain has also been analysed for the effect of an SNV in MprA (*Rv0981*) in host-pathogen interaction [21]. In the light of these background studies, the availability of the genome sequence provides the complete genetic profile that can be analysed for its correlation with various phenotypes. This led us to predict the molecular correlates for the resistance phenotype of VPCI591 for two of the first line drugs, rifampicin and ethambutol. Second line drug-susceptibility testing had not been put up for the isolate. However, WGS results showed variations at *gyrA* (E21Q, S95T, A384V and G668D) and *gyrB* (M330I) for fluoroquinolones.

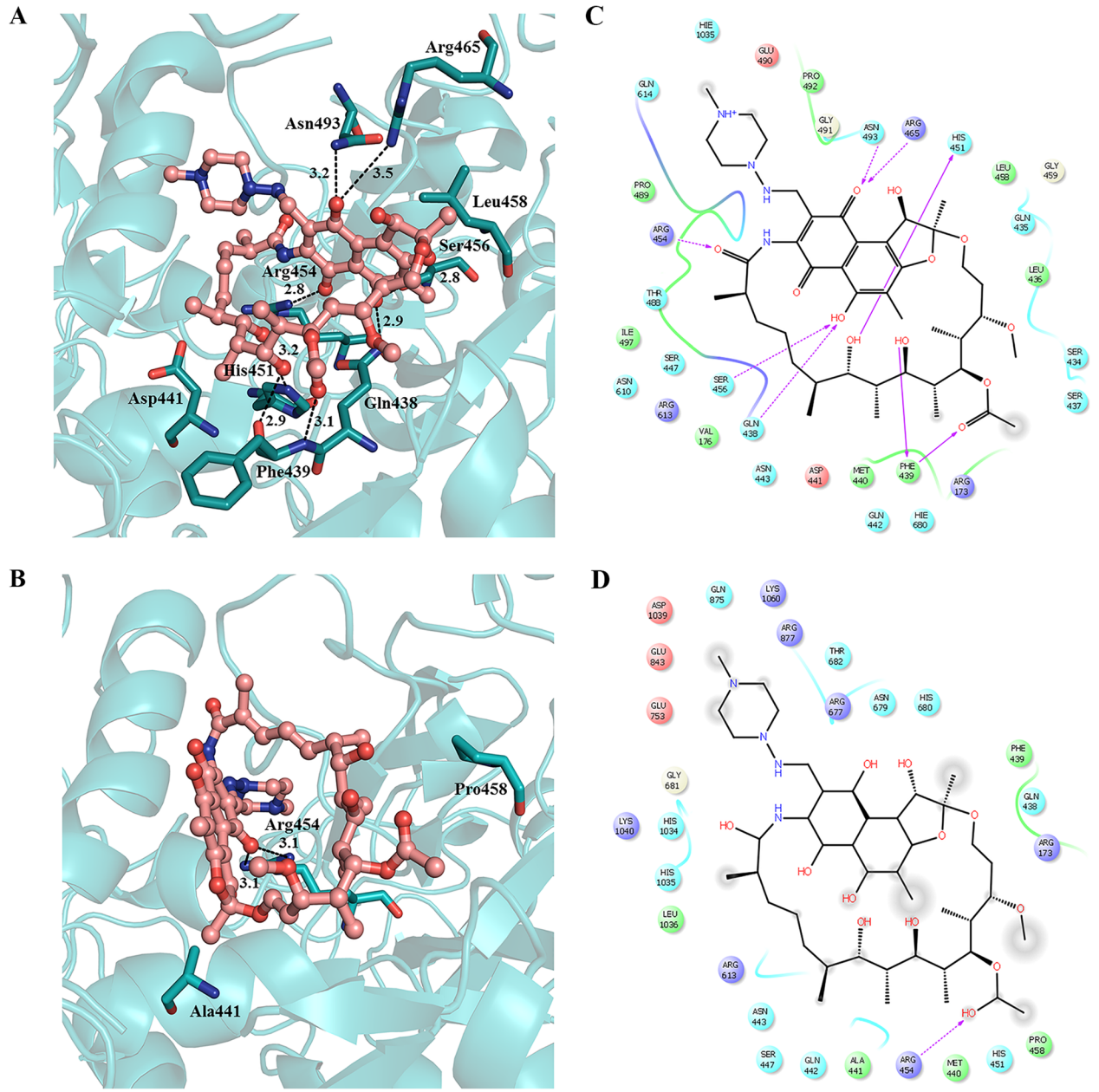


Fig. 5 Docking analysis of SNV in RpoB. **a** & **c** are for (RpoB-Wild type-H37Rv) and **b** & **d** are for (RpoB-clinical isolate-VPCI591). **a** & **b** represent 3D view, amino acid residues are shown with sticks and rifampicin (red) is shown as ball and stick model, hydrogen bonds as dotted lines (black). **c** & **d** represent the corresponding 2D sketch of molecular interaction of rifampicin with adjacent amino acid residues indicated by purple arrow

Among the high confidence SNVs identified, the antigenic *PE-PPE* proteins show high variation. They have multiple variations in a given gene in VPCI591 as in the 1553 global clinical isolates of *M. tuberculosis*, that we analysed from the repositories tbVar [19] and GMTV [18]. However there are also certain invariant *PE, PPE* family of genes. The number of SNVs per gene is highest in the PPE genes *Rv0355c* and *Rv3347c*; out of the 7 SNVs in *Rv0355c,* 6 are non-synonymous, while 5 out of 8 SNVs in *Rv3347c* are non-synonymous. The members such as *PPE 35*, *PPE55*, *PPE8*, *PPE54*, *PPE34*, *PPE24* genes are known to be highly polymorphic among clinical isolates [20]. SNVs leading to stop-loss (SL) in *PPE33*, *PPE67* and stop-gain (SG) in *PPE42*, *PPE35,* in addition to SG in fatty acid ligase gene, *fadD15* are also identified. However the phenotypic consequence of these variations is not reflected in the ability of the pathogen to infect and survive in the human host or in in-vitro growth.

The classification of the SNV containing genes based on their biological function indicates that genes involved

Table 3 Molecular docking analysis of RpoB crystal structure of H37Rv PDB-5UHB with Rifampicin by Autodock 4.2

Protein Model	Binding Energy, ΔG (Kcal mol^{-1})	$K_{i\ observed}$ (nM/μM)	Hydrogen Bonding	
			Residues	Distance (Å)
H37Rv	-9.27	160.26 (nM)	Gln438	2.9
			Phe439	2.9, 3.1
			His451	3.2
			Arg454	2.8
			Ser456	2.8
			Arg465	3.5
			Asn493	3.2
D441A[a]	−6.76	11.06 (μM)	Gln438	3.2
			Phe439	2.7, 2.7
			Arg465	3.4
L458P[a]	−7.11	6.16 (μM)	Gln438	2.8
			Phe439	2.7, 2.8
			His441	3.2
			Arg465	3.3, 3.5
I1112M[a]	−7.18	5.50 (μM)	Gln438	2.8
			Phe439	2.6, 2.8, 3.3
			Arg465	3.5
D441A + L458P + I1112M[b]	−5.82	53.96 (μM)	Arg454	3.1, 3.1

Binding parameters computed based on molecular docking analysis of the crystal structure of RpoB H37Rv (PDB-5UHB) with Rifampicin by Autodock 4.2. The binding parameters were calculated for RpoB of H37Rv and VPCI591; [a] with individual SNV and [b] the three SNVs co-occurring in VPCI591

in transferase activity were the major class represented among the non-synonymous SNV containing genes in VPCI591 as also in others as reflected in tbVar [19] and GMTV [18] databases. It can be speculated that transferases including those involved in the post-translational modifications can tolerate variations and thus show plasticity, due their role in conferring functional diversity and having various classes of proteins as substrates.

SNVs mapping in the intergenic region between two genes that are transcribed in opposite directions, can impact the transcription of both the genes, if they map within the putative promoters or regulatory regions. There are several instances of such intergenic SNV in VPCI591, including SNV at genomic position 2,010,614, between *Rv1776c* (transcriptional regulator) and *Rv1777* (cytochrome P450), *Rv2333c* (drug efflux protein) and *Rv2334* (cysteine synthase).

In the present analysis, we have validated such an effect on *embAB* gene. We note that, VPCI591 harbours several SNVs which might be associated with ethambutol resistance; in *embC*/*Rv3793* (T270I) and (N394D), *embA*/*Rv3794* (P913S) and two variations in *embB*/*Rv3795* (E378A) and (E504D). There is also another non-synonymous change in *embR*/*Rv1267c* (C110Y). In addition, VPCI591 has a variation in *pknH*/*Rv1266c* (R607Q). *PknH*-mediated increase in the transcription of *embAB* genes significantly alters resistance to ethambutol [24]. Though among the various mechanisms of ethambutol resistance in *M. tuberculosis*, the over expression of *embAB* gene due to the presence of intergenic mutation, is one of the most well-understood mechanisms [25]. In case of VPCI591, the SNVs in several genes conferring resistance to ethambutol, and a threefold increase in the transcription of *embAB* collectively contribute to the high MIC of 15 μg/ml of VPCI591 [17] relative to MIC of 0.5–2 μg/ml of H37Rv [17, 26].

Similarly, the high resistance of VPCI591 to rifampicin is attributed to the co-occurrence of three genic non-synonymous SNVs which is unique to VPCI591 and weakens the interaction of RNA polymerase β subunit with rifampicin, as reflected in the increase in binding affinity. However, the effect of SNVs on the affinity of RpoB for rifampicin may not be the only reason for this level of resistance. It is known that in addition to *RpoB* mutation, the active efflux pumps like *Rv1258c*, *Rv1410c*, and *Rv0783*, the major facilitator superfamily (MFS) proteins, contribute to rifampicin resistance [27]. VPCI591 bears no variation in these genes except SNV in *Rv2936*/*drrA* (H309D), implicated in rifampicin efflux [27]. VPCI591 has NS variations in *rpoC* (A172V) and no variation in *rpoA*. The presence of these off-target variations are also known to contribute to the evolution and survival of drug-resistant *M. tuberculosis* [28, 29].

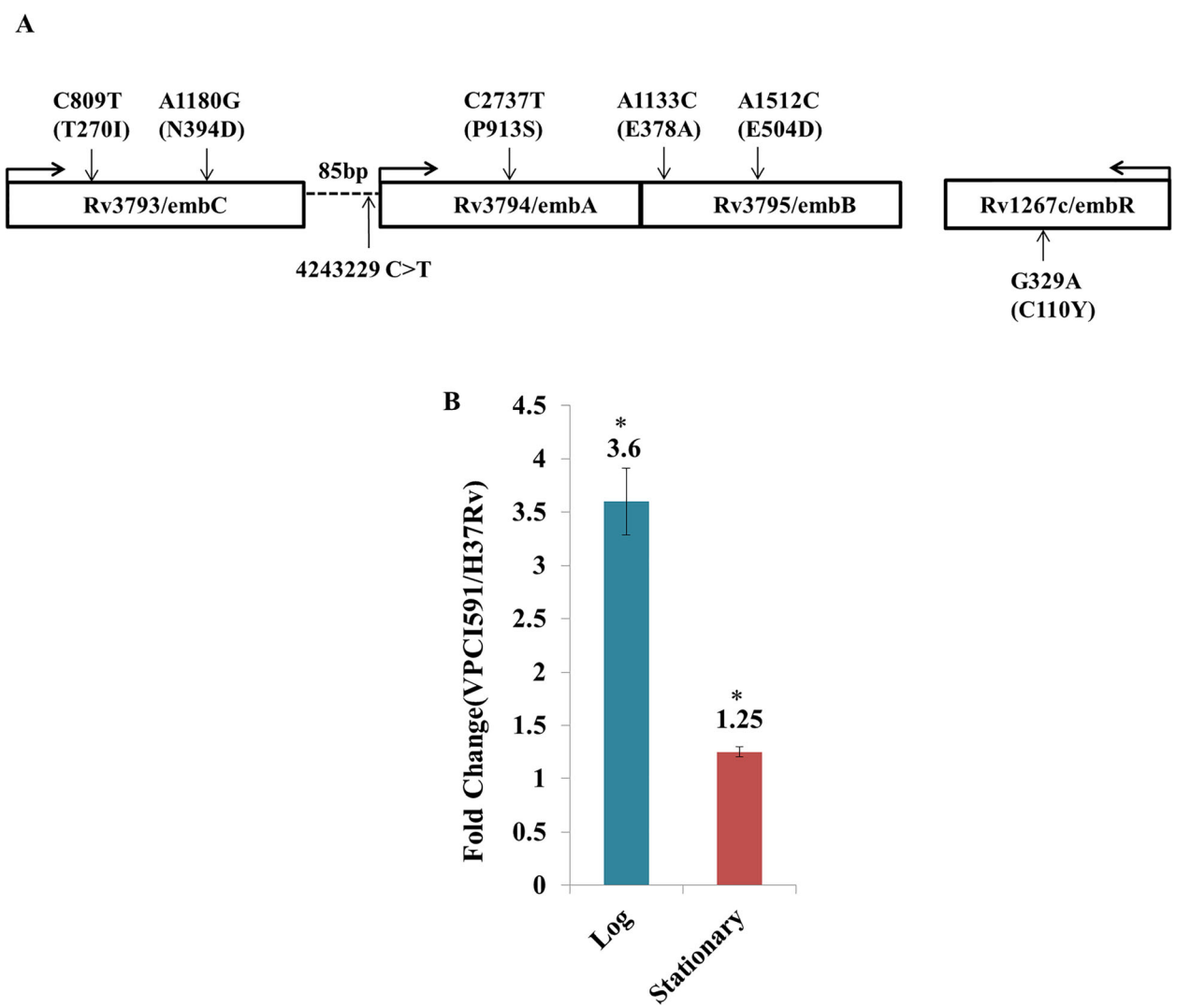


Fig. 6 Effect of intergenic SNV on the expression of *embAB*. **a**. Diagrammatic representation of *embCAB* operon of VPCI591 with the SNVs (vertical arrows). The direction of transcription is indicated by horizontal arrows. The intergenic region between *embC* and *embAB* (85 bp) is represented by dotted lines. **b**. Relative expression of *embAB* gene in VPCI591 with respect to H37Rv at log phase and stationary phase, represented as mean ± SEM (*$p < 0.05$). The results from two biological replicates are shown

The identification of genes that do not show any SNV in a large number of global clinical isolates (invariant regions), would be potential diagnostic markers for identification of *M. tuberculosis* complex [30–32]. We found several invariant genes in 1553 global clinical isolate including VPCI591. It has been shown that the invariant region in 190 bp region of *Rv1458c* and partial regions of *Rv0440* can be used as diagnostic markers for the identification of *M. tuberculosis* complex (MTBC) [30]. The major group of invariant genes in our study includes, toxin antitoxin pathway genes like *Rv0661c*, *Rv2760c*, *Rv0664*, and *Rv0581*, and transposase of IS elements (Insertion elements) like *IS6110*. Several ribosomal subunit genes and the putative mycobacteriophage proteins like probable PhiRv1 phage protein are also devoid of variations in the clinical isolates. Several secretory pathway proteins such as *Rv3875* (ESAT-6), TatA, SecG and sugar transport proteins like SugB, proteins of electron transport chain *nuoE*, *nuoC*, *nuoK* are also found to be invariant. Thus such invariant regions present in clinical isolates can be used as diagnostic markers.

Conclusion

The genetic variation in *M. tuberculosis* clinical isolates VPCI591 is common among almost all functional classes of genes. We demonstrate that the variations bring about quantitative changes in transcription and can lead to altered structure of the protein and interaction with the drug molecule. Likewise intergenic distance for non-genic SNVs plays a very important role if they lie within

the putative promoter or regulatory region. However, the clinical isolates survive and are pathogenic implying that there is no drastic negative effect that is obvious. On the other hand, the invariant genes can be explored for their potential as drug targets.

Methods

Bacterial strains and culture conditions

M. tuberculosis clinical isolate VPCI591 and *M. tuberculosis* H37Rv were grown at 37 °C in Middlebrook 7H9 broth (Becton Dickinson) supplemented with 10% OADC (Oleic acid, Bovine albumin fraction V, dextrose, catalase) from Himedia (FD 329),with 0.2% glycerol (SRL) and 0.05% tween 80 (Sigma).

Genome sequencing and pipeline for analysis of genome variation

Genomic DNA was isolated using standard protocol [33]. The genome sequencing was done using Illumina GAIIx analyser at CSIR-Institute of genomics and Integrative biology (IGIB). TruSeq2_SE adapters were used for sequencing. Sequence read file for the clinical isolate VPCI591 have been deposited in SRA format, NCBI: SRX5802345, Bioproject accession number PRJNA540936, BioSample accession numbers: SAMN11568242.

FastQC (0.11.5) was used for deciphering the overall quality statistics of the data [34]. Trimmomatic (0.36) [35] was used to remove adapter contamination shown by FastQC. *M. tuberculosis* H37Rv (NC_000962) sequence was used as the reference strain to identify the genetic variations.

Tools used for data analysis

To avoid errors, four different pipelines were used in data analysis: (a) Maq (Mapping and Assembly with Quality) (0.6.6) [36] (b) Bowtie 2 (2.0.4) was used for variant detection in massively parallel sequencing data [37] in combination with Samtools [38], GATK (3.6–0-g89b7209) and Picard Tools (1.119) (http://broadinstitute.github.io/picard) [39, 40], (c) Varscan 2 [41] and Mpileup of Samtools [38] (d) inGAP, an integrated genome analysis pipeline [42]. The consensus of the four software pipeline was considered to call the **S**ingle **N**ucleotide **V**ariations (SNV) in the genome of VPCI591. The variations were mapped using ANNOVAR software [43].

Classification of the genome sequence data

The whole genome SNVs were classified into genic and non-genic/intergenic on the basis of their location on the genome. The genic variations were further classified into (a) Synonymous (SY), (b) Non-synonymous (NS), (c) Stop-gain (SG), and (d) Stop-loss (SL) categories depending on their effect on the protein sequence. In case of SG and SL variations we predicted the number of amino acids lost or gained due to premature termination or extension respectively. In SL variations, the extended protein resulting from the addition of amino acids was analysed for its predicted function by BLASTP [44] and InterProScan [45]. The ontology for genes harbouring NS variations were carried out using PANTHER [22] and further classified according to the major pathway [4]. The non-synonymous SNVs in drug resistance genes were retrieved by literature mining as well as the analysis of TBDReaMDB, a comprehensive drug resistance database [14]. The SNVs were screened for their predicted effect through SIFT [46]. For non-genic/intergenic variations, the relative distance between the SNV and the downstream gene was determined. From global comparative analysis with tbVar [19] and GMTV [18] database, invariant genes and variations unique to VPCI591 were identified.

Structural analysis of RpoB (*Rv0667*)

To assess the effect of the NS-SNV in RpoB detected in VPCI591, structural analysis was carried out. The atomic coordinates of RpoB complex with rifampicin (PDB Id: 5UHB; Resolution: 4.29-Å) [47] was obtained from Protein Data Bank (http://www.rcsb.org/pdb). The effect of variation on the structure and the binding of rifampicin was investigated by molecular docking to examine the protein-ligand (rifampicin) interaction using AutoDock 4.2 [48]. The structure of the protein with SNVs was generated using Swiss-PdbViewer [49]. All the hetero atoms were removed, and both the protein and ligand files were prepared and saved in PDBQT format and used as initial input for AutoDock following the standard protocols. The Lamarckian genetic algorithm (LGA) method was applied for docking and to deal with the protein–antagonist interactions [50]. The polar hydrogen atoms were added geometrically and Gasteiger charge to all the atoms of the protein was assigned using AutoDocTools (ADT). The 3D affinity grid fields with grid map of $50 \times 50 \times 50$ points spaced equally at 0.375 Å and centre of the grid box was $164.1 \times 163.34 \times 20.42$ using auxiliary program AutoGrid to evaluate the binding energies between the ligand and receptor. The standard protocol of ADT utility was used to generate both the grid parameter file (gpf) and docking parameter file (dpf). The resultant docked models of wild-type and mutant complexes with rifampicin were analyzed using Schrödinger and PyMOL to visualize molecular interactions.

Expression analysis

The intergenic variation mapping between *embC* and *embAB* genes inVPCI591 was validated by Sanger sequencing using the primers, forward (CCTAGGAACG GTGACTCG) and reverse (AGACGACGGCTGCTAG GC). For expression analysis total RNA was extracted

from the clinical isolate VPCI591 and H37Rv using the RNeasy mini kit (Qiagen) and was treated with DNase using TURBO™ DNase kit (Invitrogen) and cDNA was prepared using First strand cDNA synthesis kit (Fermentas K1612). Quantitative PCR was performed with *sigA*/ *Rv2703* gene as control and fold change was measured by ΔΔCt method using FastStart universal master mix (Roche). The following primers were used: embA (F- GTAATGAGCGATCTCACCGG/ R- CGGTGATCTG GGTGATGTTG); sigA (F- AACGCACCGCCACCAA GTC/ R- TGGTGCTGGTCGTAGTGTCCTG).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12866-020-01912-6.

Additional file 1. FastQC analysis. A. Assessment of quality of the sequence reads. The data shows quartile deviation with mean deviation (blue curve). Sequence quality is shown on the X axis and base position on Y axis. Green- high quality reads, yellow-average and red-low quality reads. B. Mean distribution of GC content, estimated from VPCI591 sequence data (red) compared to the theoretical distribution of the GC rich genome (blue). C. Mean sequence quality score, ranging from 32 to 34 indicating high quality reads.

Additional file 2. Complete list of Non-Synonymous SNV in VPCI591.

Additional file 3. Intergenic Variations in VPCI591 (Type I).

Additional file 4. Intergenic Variations identified in VPCI591 (Type II).

Additional file 5. Invariant genes in global clinical isolates (including VPCI591).

Additional file 6. Unique genic variations in VPCI591.

Additional file 7. Inter-genic variations unique to VPCI591.

Additional file 8. Docking analysis of SNV in RpoB. A. The crystal structure of RpoB (5UHB) is represented harbouring all the 3 variations, as present in VPCI591 in green. B, D and F represent the 3D view for D441A, L452P & I1112M respectively. Amino acid residues are shown with sticks and rifampicin (red) is shown with ball and stick model. Hydrogen bonds are shown as broken line (black). C, E and G represent the corresponding 2D sketch. Rifampicin is represented in black interacting to amino acids (colour circles) through purple arrows.

Abbreviations

SY: Synonymous; NS: Non-synonymous; SG: Stop-gain; SL: Stop-loss; AA: Amino acids; SNV: Single nucleotide variation

Acknowledgements

The authors acknowledge Dr. Vinod Scaria at CSIR- Institute of Genomics and Integrative Biology (IGIB) for his help with the genome sequencing.

Authors' contributions

M.B., V.B., K.B., M.V. conceived and designed the experiments. K.B., V.N., and M.J., performed the experiments. K.B., V.N. VB analysed the data and wrote the manuscript. R. P provided inputs through discussion. All authors discussed, reviewed and approved the final manuscript.

Funding

The authors thank OSDD (Open Source Drug Discovery) and CSIR for funding. VB acknowledges UGC Special Assistance Program (UGC-SAP-II) and DU-DST PURSE Grant. We acknowledge DBT Bioinformatics Information facility at ACBR.

Availability of data and materials

NCBI: SRX5802345, Bioproject accession number PRJNA540936, BioSample accession numbers: SAMN11568242 Additional files 1-8

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have no competing financial interests and are solely responsible for the experimental designs and data analysis.

Author details

[1]Dr. B. R. Ambedkar Center for Biomedical Research (ACBR), University of Delhi, 110007, New Delhi, India. [2]Department of Microbiology, Vallabhbhai Patel Chest Institute, University of Delhi, New Delhi, India. [3]Department of Chemistry, University of Delhi, New Delhi, India.

Received: 5 March 2020 Accepted: 19 July 2020
Published online: 25 July 2020

References

1. Cubillos-Ruiz A, Morales J, Zambrano MM. Analysis of the genetic variation in Mycobacterium tuberculosis strains by multiple genome alignments. BMC Res Notes. 2008;1(1):110.
2. Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S, Eiglmeier K, Gas S, Barry Iii C. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998;393(6685):537.
3. Behr M, Wilson M, Gill W, Salamon H, Schoolnik G, Rane S, Small P. Comparative genomics of BCG vaccines by whole-genome DNA microarray. Science. 1999;284(5419):1520–3.
4. Camus J-C, Pryor MJ, Médigue C, Cole ST. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. Microbiology. 2002;148(10):2967–73.
5. Advani J, Verma R, Chatterjee O, Pachouri PK, Upadhyay P, Singh R, Yadav J, Naaz F, Ravikumar R, Buggi S, et al. Whole genome sequencing of Mycobacterium tuberculosis clinical isolates from India reveals genetic heterogeneity and region-specific variations that might affect drug susceptibility. Front Microbiol. 2019;10:309.
6. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZ, Depledge DP, Nikolayevskyy V, Broda A, Stone MJ, et al. Rapid whole-genome sequencing of Mycobacterium tuberculosis isolates directly from clinical samples. J Clin Microbiol. 2015;53(7):2230–7.
7. Kidenya BR, Mshana SE, Fitzgerald DW, Ocheretina O. Genotypic drug resistance using whole-genome sequencing of Mycobacterium tuberculosis clinical isolates from North-Western Tanzania. Tuberculosis (Edinb). 2018;109:97–101.
8. Takiff HE, Feo O. Clinical value of whole-genome sequencing of Mycobacterium tuberculosis. Lancet Infect Dis. 2015;15(9):1077–90.
9. Roa MB, Tablizo FA, Morado EKD, Cunanan LF, Uy IDC, Ng KCS, Manalastas-Cantos KG, Reyes JM, Ganchua SKC, Ang CF, et al. Whole-genome sequencing and single nucleotide polymorphisms in multidrug-resistant clinical isolates of Mycobacterium tuberculosis from the Philippines. J Glob Antimicrob Resist. 2018;15:239–45.
10. Ford C, Yusim K, Ioerger T, Feng S, Chase M, Greene M, Korber B, Fortune S. Mycobacterium tuberculosis–heterogeneity revealed through whole genome sequencing. Tuberculosis (Edinb). 2012;92(3):194–201.
11. Gagneux S, Small PM. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. Lancet Infect Dis. 2007;7(5):328–37.
12. Gao Q, Kripke KE, Saldanha AJ, Yan W, Holmes S, Small PM. Gene expression diversity among Mycobacterium tuberculosis clinical isolates. Microbiology. 2005;151(1):5–14.
13. Rehren G, Walters S, Fontan P, Smith I, Zárraga AM. Differential gene expression between Mycobacterium bovis and Mycobacterium tuberculosis. Tuberculosis (Edinb). 2007;87(4):347–59.
14. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med. 2009;6(2):e1000002.
15. Zhang H, Li D, Zhao L, Fleming J, Lin N, Wang T, Liu Z, Li C, Galwey N, Deng J. Genome sequencing of 161 Mycobacterium tuberculosis isolates

from China identifies genes and intergenic regions associated with drug resistance. Nat Genet. 2013;45(10):1255.
16. Sharma M, Bose M, Sharma L, Diwakar A, Kumar S, Gaur SN, Banavalikar JN. Intracellular survival of Mycobacterium tuberculosis in macrophages is modulated by phenotype of the pathogen and immune status of the host. Int J Mycobacteriol. 2012;1(2):65–74.
17. Tandon R, Ponnan P, Aggarwal N, Pathak R, Baghel AS, Gupta G, Arya A, Nath M, Parmar VS, Raj HG. Characterization of 7-amino-4-methylcoumarin as an effective antitubercular agent: structure–activity relationships. J Antimicrob Chemother. 2011;66(11):2543–55.
18. Chernyaeva EN, Shulgina MV, Rotkevich MS, Dobrynin PV, Simonov SA, Shitikov EA, Ischenko DS, Karpova IY, Kostryukova ES, Ilina EN, et al. Genome-wide Mycobacterium tuberculosis variation (GMTV) database: a new tool for integrating sequence variations and epidemiology. BMC Genomics. 2014;15:308.
19. Joshi KR, Dhiman H, Scaria V. tbvar: a comprehensive genome variation resource for Mycobacterium tuberculosis. Database. 2014;2014:bat083.
20. McEvoy CR, Cloete R, Müller B, Schürch AC, Van Helden PD, Gagneux S, Warren RM, van Pittius NCG. Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. PLoS One. 2012;7(4):e30593.
21. Bhattacharyya K, Bandopadhyay U, Singh A, Prakash A, Nemaysh V, Jain S, Varma-Basil M, Lynn AM, Bose M, Luthra PM. Modulation of macrophage defense responses by Mycobacterial persistence protein MprA (Rv0981) in human THP-1 cells: effect of single amino acid variation on host-pathogen interactions. bioRxiv. 2020.04.27.063602.
22. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013; 8(8):1551.
23. Joon M, Bhatia S, Pasricha R, Bose M, Brahmachari V. Functional analysis of an intergenic non-coding sequence within mce1 operon of M. tuberculosis. BMC Microbiol. 2010;10(1):128.
24. Sharma K, Gupta M, Pathak M, Gupta N, Koul A, Sarangi S, Baweja R, Singh Y. Transcriptional control of the mycobacterial embCAB operon by PknH through a regulatory protein, EmbR, in vivo. J Bacteriol. 2006;188(8):2936–44.
25. Cui Z, Li Y, Cheng S, Yang H, Lu J, Hu Z, Ge B. Mutations in the embC-embA region contribute to M. tuberculosis resistance to ethambutol. Antimicrob Agents Chemother. 2014;58(11):6837–43.
26. Rastogi N, Labrousse V, Goh KS. In vitro activities of fourteen antimicrobial agents against drug susceptible and resistant clinical isolates of Mycobacterium tuberculosis and comparative intracellular activities against the virulent H37Rv strain in human macrophages. Curr Microbiol. 1996;33(3): 167–75.
27. Pang Y, Lu J, Wang Y, Song Y, Wang S, Zhao Y. Study of the rifampin monoresistance mechanism in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2013;57(2):893–900.
28. Comas I, Borrell S, Roetzer A, Rose G, Malla B, Kato-Maeda M, Galagan J, Niemann S, Gagneux S. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. Nat Genet. 2012;44(1):106.
29. Q-j L, W-w J, Q-q Y, Xu F, Li J-Q, Sun L, Xiao J, Li Y-J, Mokrousov I, Huang H-R. Compensatory mutations of rifampin resistance are associated with transmission of multidrug-resistant Mycobacterium tuberculosis Beijing genotype strains in China. Antimicrob Agents Chemother. 2016;60(5):2807–12.
30. Shrivastava K, Garima K, Narang A, Bhattacharyya K, Vishnoi E, Singh RK, Chaudhry A, Prasad R, Bose M, Varma-Basil M. Rv1458c: a new diagnostic marker for identification of Mycobacterium tuberculosis complex in a novel duplex PCR assay. J Med Microbiol. 2017;66(3):371–6.
31. Varma-Basil M, Garima K, Pathak R, Dwivedi SKD, Narang A, Bhatnagar A, Bose M. Development of a novel PCR restriction analysis of the hsp65 gene as a rapid method to screen for the Mycobacterium tuberculosis complex and nontuberculous mycobacteria in high-burden countries. J Clin Microbiol. 2013;51(4):1165–70.
32. Singh A, Kashyap VK. Specific and rapid detection of Mycobacterium tuberculosis complex in clinical samples by polymerase chain reaction. Interdiscip Perspect Infect Dis. 2012;2012:654694.
33. Bose M, Chander A, Das R. A rapid and gentle method for the isolation of genomic DNA from mycobacteria. Nucleic Acids Res. 1993;21(10):2529.
34. Andrews S. FastQC: a quality control tool for high throughput sequence data; 2010.
35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
36. Li H, Ruan J, Durbin R. Maq: mapping and assembly with qualities. Version. 2008;06:3.
37. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25(14):1754–60.
38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
40. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1):11.10. 11–11.10. 33.
41. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76.
42. Qi J, Zhao F, Buboltz A, Schuster SC. inGAP: an integrated next-generation genome analysis pipeline. Bioinformatics. 2009;26(1):127–9.
43. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164.
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
45. Zdobnov EM, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17(9):847–8.
46. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
47. Lin W, Mandal S, Degen D, Liu Y, Ebright YW, Li S, Feng Y, Zhang Y, Mandal S, Jiang Y. Structural basis of *Mycobacterium tuberculosis* transcription and transcription inhibition. Mol Cell. 2017;66(2):169–79 e168.
48. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785–91.
49. Johansson MU, Zoete V, Michielin O, Guex N. Defining and searching for structural motifs using DeepView/Swiss-PdbViewer. BMC Bioinformatics. 2012;13(1):173.
50. Fuhrmann J, Rurainski A, Lenhof HP, Neumann D. A new Lamarckian genetic algorithm for flexible ligand-receptor docking. J Comput Chem. 2010;31(9):1911–8.

Publisher's Note