nature genetics

Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance

Hongtai Zhang[1,2,14], Dongfang Li[3,4,14], Lili Zhao[5,6,14], Joy Fleming[1,14], Nan Lin[7], Ting Wang[1], Zhangyi Liu[3], Chuanyou Li[8], Nicholas Galwey[1], Jiaoyu Deng[9], Ying Zhou[1], Yuanfang Zhu[3], Yunrong Gao[1], Tong Wang[3], Shihua Wang[7], Yufen Huang[3], Ming Wang[1], Qiu Zhong[10], Lin Zhou[10], Tao Chen[10], Jie Zhou[11], Ruifu Yang[3], Guofeng Zhu[12], Haiying Hang[1], Jia Zhang[1], Fabin Li[13], Kanglin Wan[5,6], Jun Wang[3], Xian-En Zhang[2,9] & Lijun Bi[1]

The worldwide emergence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) tuberculosis threatens to make this disease incurable[1,2]. Drug resistance mechanisms are only partially understood[3–5], and whether the current understanding of the genetic basis of drug resistance in *M. tuberculosis* is sufficiently comprehensive remains unclear. Here we sequenced and analyzed 161 isolates with a range of drug resistance profiles, discovering 72 new genes, 28 intergenic regions (IGRs), 11 nonsynonymous SNPs and 10 IGR SNPs with strong, consistent associations with drug resistance. On the basis of our examination of the dN/dS ratios of nonsynonymous to synonymous SNPs among the isolates[6–8], we suggest that the drug resistance–associated genes identified here likely contain essentially all the nonsynonymous SNPs that have arisen as a result of drug pressure in these isolates and should thus represent a near-complete set of drug resistance–associated genes for these isolates and antibiotics. Our work indicates that the genetic basis of drug resistance is more complex than previously anticipated and provides a strong foundation for elucidating unknown drug resistance mechanisms.

Although there has been an encouraging 1.3% decline in tuberculosis incidence worldwide each year since 2002 (ref. 9), drug-resistant tuberculosis is a serious and growing global challenge. Drug resistance is particularly acute in China, where 5.7% of new tuberculosis cases are MDR, and 8% of MDR cases are XDR[10]. Mutations in a limited number of coding genes and IGRs have been identified in drug-resistant strains[3–5], but the full spectrum of genetic causes of drug resistance is unclear. Although some recent *M. tuberculosis* genome-wide studies have addressed aspects of drug resistance[11–15], none have systematically searched for the genes, IGRs and SNPs most closely related to drug resistance. With this aim in mind, we performed a comprehensive genome-wide study of 161 isolates from China with a broad range of resistance profiles (44 drug-sensitive, 94 MDR and 23 XDR isolates; **Supplementary Fig. 1** and **Supplementary Table 1**).

After capturing genetic variation using Illumina DNA sequencing, we compared each genome with the *M. tuberculosis* H37Rv reference genome[16]. On average, 5.27 million sequence reads were obtained per genome at a depth of 112× and with coverage of 99.32% (**Supplementary Table 1**). Overall, 18,970 SNPs present in ≥95% of the isolates were identified, 17,824 of which were present in all 161 isolates and 22 other published genomes[17] and were used to construct a genome-wide phylogeny. Our phylogenetic tree (**Fig. 1**) agrees well with published trees[17,18] and restricts all but 2 of the 161 isolates to lineage 2 (mainly from East Asia), 95% of which belong to the Beijing family (**Supplementary Table 1**), or lineage 4 (mainly from Europe, North and South America and Africa)[19]. We identified 161 lineage 2–specific and 99 lineage 4–specific nonsynonymous SNPs distributed, respectively, across 151 and 99 genes (**Supplementary Fig. 2** and **Supplementary Table 2**) by comparison with the ancestral sequence. Ten genes contained both lineage 2–specific and lineage 4–specific SNPs (**Supplementary Table 2**) and thus could potentially be used as markers for the typing of lineage 2 and 4 isolates.

After discarding phylogenetically related and synonymous SNPs, we focused subsequent analyses on nonsynonymous SNPs and IGR

[1]Key Laboratory of Non-coding RNA, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. [2]State Key Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. [3]BGI-Shenzhen, Shenzhen, China. [4]Shenzhen Key Laboratory of Unknown Pathogen Identification, Shenzhen, China. [5]National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China. [6]State Key Laboratory for Infectious Disease Prevention and Control, Beijing, China. [7]College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China. [8]Beijing Chest Hospital, Capital Medical University, Beijing, China. [9]State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China. [10]Center for Tuberculosis Control of Guangdong Province, Guangzhou, China. [11]The Fourth People's Hospital, Foshan City, China. [12]Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China. [13]Heilongjiang Center for Tuberculosis Control and Prevention, Harbin, China. [14]These authors contributed equally to this work. Correspondence should be addressed to L.B. (blj@sun5.ibp.ac.cn), X.-E.Z. (x.zhang@wh.iov.cn), J.W. (wangj@genomics.org.cn) or K.W. (wankanglin@icdc.cn).

Received 7 February; accepted 30 July; published online 1 September 2013; doi:10.1038/ng.2735

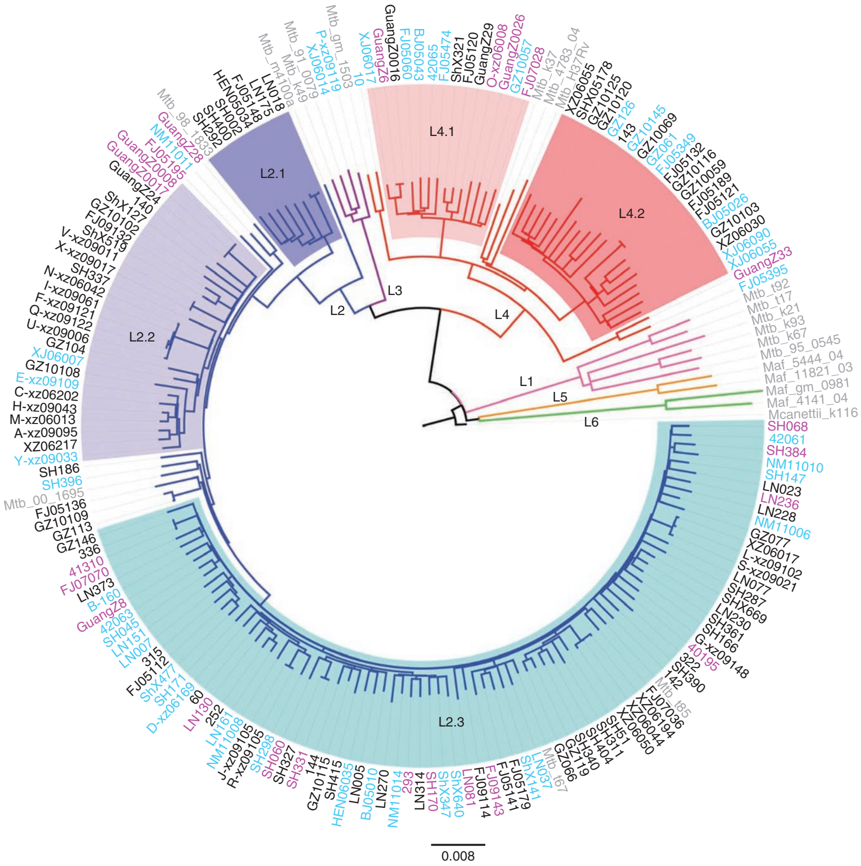

Figure 1 Phylogenetic analysis of *M. tuberculosis* isolates. Our phylogenetic tree of the genomes of the 161 *M. tuberculosis* isolates in this study along with 22 previously published genomes[17] restricts all but 2 isolates (p-xz09119 from Tibet and XJ06014 from Xinjiang, both assigned to lineage 3) to lineage 2 or lineage 4, further subdividing lineages 2 and 4 into 3 and 2 sublineages, respectively. Branch colors represent the lineages of *M. tuberculosis*: pink, lineage 1 (the Philippines and the Indian Ocean rim); blue, lineage 2 (East Asia); purple, lineage 3 (India, East Africa); red, lineage 4 (Europe, America); brown, lineage 5, *Mycobacterium africanum* (West Africa 1); green, lineage 6, *M. africanum* (West Africa 2). The blue and red shaded regions show the sublineages of lineages 2 and 4, respectively. Strains labeled in gray represent previously published genomes, those in blue represent drug-sensitive strains, those in black represent MDR strains, and those in pink represent XDR strains. The scale bar represents the number of SNP substitutions per SNP site.

SNPs (**Supplementary Fig. 3**). We anticipated that (i) genes or IGRs associated with drug resistance should be diversifying, i.e., they should have a higher density of nonsynonymous SNPs or IGR SNPs, respectively, than non-associated genes or IGRs, and (ii) that such genes or IGRs should be mutated more frequently in drug-resistant isolates than in drug-sensitive ones. We first examined differences between MDR and XDR isolates, identifying 15 genes (including the long noncoding gene for 16S rRNA, *rrs*) and 13 IGRs that met these criteria (**Fig. 2**, **Table 1**, Online Methods, **Supplementary Figs. 3** and **4**, and **Supplementary Tables 3–8**). Many well-known drug resistance genes (*gyrA*, *pncA*, *rpoB*, *rpoC*, *embB*, *ethA*, *katG*, *rpsL*, *thyA* and *rrs*)[20] and IGRs (*embC–embA*, *proA–ahpC*, *Rv1482c–fabG1* and *eis–Rv2417c*)[21–24] were selected, demonstrating the reliability of our approach. Whereas some genes were associated specifically with XDR isolates and presumably, therefore, with second-line drugs, almost half of the selected genes were associated with resistance in both MDR and XDR isolates. This finding is probably due to the overlapping drug resistance profiles of MDR and XDR isolates used here (**Supplementary Table 1**), arising in part from the somewhat artificial definitions of MDR (resistance to at least isoniazid and rifampicin) and XDR (also resistant to any fluoroquinolone and at least one second-line injectable drug)[25] and from clinical practice in the use of first- and second-line tuberculosis drugs. As a result, isolates resistant to, for example, several second-line drugs but not to fluoroquinolones are still classified as MDR. We repeated our analysis drug by drug, grouping all isolates resistant to a given drug and using all isolates sensitive to that drug as controls, thereby identifying 70 more genes (including *gyrB*, *embA* and *inhA*) and 19 IGRs (**Fig. 2**, **Table 2**, **Supplementary Figs. 3** and **4**, and **Supplementary Tables 3–8**) with strong, consistent associations with drug resistance. Many well-known drug resistance genes were associated with resistance to most drugs tested, suggesting that background drug resistance confounds this analysis. Development of resistance in *M. tuberculosis* is likely stepwise, occurring in parallel with the stepwise use of drugs (which act as a selective pressure) in response to increasingly resistant tuberculosis[26] and their potency, with the most likely sequence of resistance development for the drugs tested here (based on drug resistance

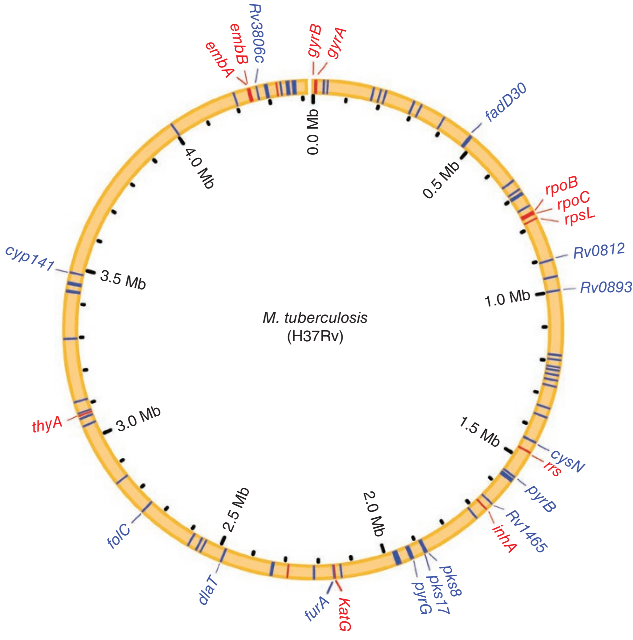

Figure 2 Genomic locations of the 85 drug resistance–associated genes. Genes previously reported to be associated with drug resistance are shown by red lines, whereas those associated with drug resistance for the first time in this study are shown by blue lines. The genes named are those considered essential for the *in vitro* growth of H37Rv[47].

Table 1 Genes and IGRs that have strong, consistent associations with MDR and XDR isolates

Gene	MDR	XDR
rrs[a]	15/9/16	14/3/16
gyrA[a]	–	9/3/9
Rv0147	–	3/3/3
Rv0265c	–	3/3/3
rpoB[a]	91/30/118	20/11/30
rpoC[a]	34/25/35	8/7/8
rpsL[a]	52/3/52	11/2/11
Rv1129c	11/10/11	–
katG[a]	83/21/87	20/5/20
pncA	32/30/33	9/9/9
Rv2752c	–	4/4/4
thyA[a]	–	4/3/4
pks15	–	3/3/3
embB[a]	65/19/67	18/9/21
ethA	–	7/9/9
Rv0744c–Rv0745	–	2/2/2
pip–Rv0841[b]	–	2/2/2
ctpE–Rv0909[b]	–	2/2/2
Rv0920c–Rv0921	–	2/2/2
Rv1482c–fabG1[b]	21/3/21	7/3/7
lipJ–cinA	–	2/2/2
eis–Rv2417c[a]	–	3/2/3
proA–ahpC[a]	13/7/14	3/3/3
thyX–hsdS.1[a]	10/2/10	3/1/3
Rv3185–Rv3186[a]	7/3/13	–
Rv3210c–rhlE	–	2/2/2
whiB2–fbiA	8/6/8	–
embC–embA[a]	14/5/14	5/3/5

Shown is the number of isolates in which the gene or IGR is mutated, the number of SNPs within the gene or IGR, and the total number of SNPs in the gene or IGR in all isolates carrying mutations in the gene or IGR.

[a] Genes or IGRs that contain relatively high-frequency SNPs. [b] Newly identified IGRs associated with drug resistance in XDR isolates but not with resistance to the drugs tested in this study.

profiles, tuberculosis drug treatment guidelines[26] and differences in drug potency) being isoniazid followed by rifampicin, streptomycin, ethambutol, ofloxacin, ethionamide, kanamycin and capreomycin. When we applied a logistic regression model based on this sequence, all but 11 genes and 5 IGRs showed significant associations with drug resistance ($P \leq 0.05$), and many but not all associations that were likely due to confounding were resolved (**Supplementary Fig. 4** and **Supplementary Table 9**). Our results confirm the view that the emergence of XDR is due to the accumulation of nonsynonymous SNPs associated with resistance to second-line drugs rather than to mutations in one or a few genes conferring pan-antibiotic resistance. The designation of isolates as XDR, although helpful in the clinical setting, does not therefore have a clear-cut genetic basis.

We examined the distribution of nonsynonymous SNPs and IGR SNPs in the above genes and IGRs, respectively, in the genomes of 32 drug-resistant and 10 drug-sensitive isolates from Samara province in Russia[12] and in the 21 drug-sensitive isolates used in our phylogenetic analysis[17]. Although the well-known *proA–ahpC* drug resistance–associated IGR was not associated with drug resistance in this relatively small sample of Russian isolates, 23 genes and 3 IGRs newly associated with drug resistance in our analysis contained nonsynonymous SNPs and IGR SNPs (**Supplementary Table 10**), confirming their likely relevance to drug resistance. Two of the 686 SNPs present in the drug resistance–associated genes and IGRs from our analysis were found in the Russian drug-sensitive isolates, and 3 were found in the 21 drug-sensitive isolates used in our phylogenetic analysis[17] (**Supplementary Table 10**), suggesting the possible association of these SNPs with low levels of resistance.

Gene ontology (GO) analysis using DAVID (Database for Annotation, Visualization and Integrated Discovery)[27] showed that drug resistance–associated genes were enriched in GO classes 'response to chemical stimulus', 'pyrimidine metabolism' and 'DNA topoisomerase activity' ($P < 0.05$) (**Supplementary Table 11**). Analysis of interactions of the protein products of the 84 protein-encoding genes with the drugs tested in this study (STITCH V3.1; see URLs) identified the existence of a complex network (**Supplementary Fig. 5**), suggesting that there is still much to discover about drug resistance mechanisms; the involvement of many of these proteins in drug resistance is as yet undocumented.

Although many of the newly identified drug resistance–associated genes and IGRs are poorly characterized (**Supplementary Table 12**), some, such as representatives of the *fadD*, *pks* and *mmpL* families (*fadD14*, *fadD30*, *pks2*, *pks8*, *pks15*, *pks17* and *mmpL1*), are of notable interest. Mycobacterial cell wall biosynthetic pathways are targeted by current tuberculosis drugs[28]; for example, isoniazid and ethionamide inhibit mycolic acid synthesis[29,30], and ethambutol targets the synthesis of arabinogalactan components[31]. Accumulating evidence suggests that there is functional cross-talk between fadD (FAAL family) and pks enzymes in the generation of complex cell wall lipid components[28,32]. Several *fadD* and *pks* genes are located adjacent to each other and to *mmpL* genes, which encode a family of membrane proteins that function in the transport of lipids[33]; *mmpL* genes have recently received considerable attention since *mmpL3*, which encodes the only MmpL protein considered essential for *M. tuberculosis* growth[33], was identified as a target of the drug candidates BM212 and SQ109 (refs. 34–36). Here *mmpL3* (2,835 bp) had 2 independent nonsynonymous SNPs present in only 2 of the 161 isolates (**Supplementary Table 13**), and *mmpL1* (2,877 bp), previously thought not to be involved in drug resistance[33], had 11 independent nonsynonymous SNPs in 10 drug-resistant isolates. Further investigation of the functional relationships between these families of proteins will help to determine whether the drug resistance–associated *fadD*, *pks* and *mmpL* genes identified here have compensatory roles in drug resistance. Indeed, it is likely that some drug resistance–associated genes identified here do have compensatory roles. For example, 10 of 34 nonsynonymous SNPs identified here in the *rpoC* gene, and 3 of 9 SNPs in the *proA–ahpC* IGR have previously been reported as compensatory mutations[37–40] (**Supplementary Table 14**).

There are relatively few reports on the role of IGRs in drug resistance. Here, of the 28 new drug resistance–associated IGRs identified, IGRs of note included *thyA–Rv2765* and *thyX–hsdS.1* (*thyA* and *thyX* are thymidylate synthase genes important in DNA replication and repair[41]). Because IGR SNPs may affect the expression levels of flanking genes—for example, SNPs in the *eis* promoter that result in its overexpression are associated with kanamycin resistance[24]—we investigated whether the IGR SNPs identified here affected gene expression. The three SNPs in the *thyA–Rv2765* IGR and the two SNPs in the *thyX–hsdS.1* IGR resulted in approximately 5-fold and 18-fold increases, respectively, in the expression of *lacZ* in *Mycobacterium smegmatis* relative to unmutated IGR controls (**Fig. 3**). These results strongly suggest that the SNPs in these drug resistance–associated IGRs would lead to overexpression of the downstream genes in *M. tuberculosis*. Using Neural Network Promoter Prediction[42], we predicted that SNPs were located in promoter regions in 27 of the 32 drug resistance–associated IGRs (**Supplementary Table 15**). IGRs also encode small RNAs (sRNAs), which can regulate gene expression in response to environmental changes[43,44]. Previous reports[45,46]

Table 2 Drug resistance associations of the new genes and IGRs with strong, consistent gene–drug resistance associations in this study

Gene	INH-RMP	STR	EMB	OFX	ETH	KAN	CPM	Gene	INH-RMP	STR	EMB	OFX	ETH	KAN	CPM
Rv0026	–	–	5/5/5	–	–	–	–	*yjcE*	–	–	–	4/5/5	–	–	–
Rv0039c	–	–	–	–	–	–	2/2/2	*Rv2314c*	–	–	–	–	–	2/3/3	–
Rv0147	–	–	–	–	–	3/3/3	–	*folC*[a]	–	9/7/10	–	–	7/6/8	3/2/3	3/2/3
Rv0165c	–	–	–	–	–	3/3/3	3/3/3	*Rv2530c*	–	–	–	–	–	–	2/2/2
lprO	–	–	6/6/6	6/6/6	–	–	–	*Rv2729c*	–	–	–	–	–	–	2/2/2
fadE5	–	–	–	–	–	3/4/4	–	*Rv2752c*	14/14/14	–	–	9/9/9	9/9/9	–	4/4/4
Rv0265c	–	–	–	–	3/3/3	3/3/3	2/2/2	*Rv2771c*	–	–	–	–	–	2/2/2	–
Rv0323c	–	–	–	–	–	2/2/2	2/2/2	*Rv2807*	–	–	–	–	–	–	2/2/2
Rv0401	–	–	–	–	–	–	2/2/2	*pks15*	–	–	7/7/7	–	–	3/3/3	–
mmpL1	–	–	–	–	8/9/9	–	4/4/4	*Rv3067*	–	–	–	–	–	2/2/2	2/2/2
fadD30	–	–	–	5/5/5	5/5/5	–	–	*Rv3071*	–	–	–	4/4/4	–	–	–
Rv0565c	–	–	–	4/4/4	5/5/5	3/3/3	3/3/3	*Rv3087*	–	–	–	–	–	3/3/3	–
yrbE2B	–	–	–	–	–	2/2/2	–	*Rv3090*	–	–	–	–	–	–	2/2/2
Rv0600c	–	–	–	–	–	2/2/2	–	*cyp141*	–	–	–	–	–	3/3/3	–
Rv0605	–	–	–	–	–	–	2/2/2	*fadE33*	–	–	–	–	3/3/3	–	–
Rv0608	–	–	–	–	–	–	2/2/2	*proZ*	–	–	5/5/5	–	–	–	–
lipG	–	–	–	4/4/4	–	–	–	*Rv3806c*[a]	–	–	–	6/5/7	–	–	–
Rv0812	–	–	–	–	–	2/2/2	2/2/2	*pks2*	–	–	–	10/12/12	–	–	–
fadA	–	–	–	4/4/4	–	2/2/2	–	*whiB6*[a]	–	–	–	–	6/5/6	–	–
Rv0893c	–	–	–	–	–	2/2/2	–	*Rv3877*	–	–	–	–	–	–	3/5/5
fadD14	–	–	6/6/6	–	–	–	–	*Rv3881c*	–	–	–	–	4/4/4	3/3/3	–
echA8	–	–	–	–	–	–	2/2/2	*Rv3889c*	–	–	–	–	–	–	3/3/3
celA2b	–	–	–	–	–	2/2/2	–	*Rv3894c*	–	–	–	8/8/8	–	–	–
Rv1096	–	–	–	–	–	2/2/2	2/2/2	*Rv0010c–Rv0011c*	–	–	–	–	4/4/4	–	–
Rv1112	–	–	–	–	–	–	2/2/2	*Rv0466–icl*	–	–	–	3/3/3	–	–	–
Rv1129c	13/12/13	–	8/8/8	8/7/8	7/7/7	–	–	*Rv0744c–Rv0745*	–	–	–	3/3/3	–	–	–
Rv1144	–	–	–	–	–	–	2/2/2	*PPE13–Rv0879c*	–	–	–	–	–	–	2/2/2
Rv1192	–	–	–	–	–	–	2/2/2	*Rv0920c–Rv0921*	–	–	–	–	–	–	2/2/2
Rv1218c	–	–	–	–	3/3/3	–	–	*Rv1042c–Rv1043c*	–	–	–	–	–	2/2/2	–
cysN	–	–	–	–	–	–	3/3/3	*greA–Rv1081c*	–	–	3/3/3	3/3/3	–	–	–
pyrB	–	–	–	–	–	–	2/2/2	*Rv1194c–PE13*	–	4/4/4	–	–	–	–	–
PPE20	–	–	–	–	–	3/3/3	–	*rfe–Rv1303*	–	–	3/3/3	–	–	–	–
Rv1393c	–	–	–	–	–	–	3/3/3	*Rv1347c–Rv1348*	–	–	3/3/3	–	–	–	–
Rv1465	–	–	–	–	–	2/2/2	–	*Rv1816–Rv1817*	–	–	–	4/4/4	–	2/2/2	–
Rv1520	–	–	–	–	–	–	2/2/2	*lipJ–cinA*	–	–	–	–	–	2/2/2	–
pks8	–	–	12/12/12	–	–	–	–	*blaC–sigC*	–	–	–	–	–	–	2/2/2
pks17	–	–	–	5/5/5	–	–	–	*cobS–Rv2209*	–	–	–	–	–	2/2/2	2/2/2
pyrG	–	–	6/6/6	–	–	–	–	*PE_PGRS39–lppQ*	–	–	–	–	3/3/3	–	–
cycA	–	–	–	5/5/5	–	–	–	*Rv2733c–Rv2734*	–	–	3/2/3	–	–	–	–
narX	–	–	–	–	–	3/3/3	3/3/3	*thyX–hsdS.1*[a]	13/2/13	12/2/12	9/2/9	–	–	–	–
Rv1741	–	–	–	–	–	2/2/2	–	*thyA–Rv2765*	–	–	–	–	–	–	2/2/2
pknF	–	6/8/8	–	–	–	–		*Rv3185–Rv3186*[a]	7/3/13	7/3/13	–	–	–	–	–
Rv1885c	–	–	–	–	–	–	2/2/2	*Rv3210c–rhlE*	–	–	–	–	–	2/2/2	–
furA	–	–	–	–	–	–	2/2/2	*whiB2–fbiA*	9/8/10	8/8/9	–	–	6/6/7	–	–
mce3D	–	–	–	5/5/5	4/4/4	3/3/3	–	*infA–Rv3463*	–	–	3/3/3	–	–	–	–
Rv2077c	–	–	–	–	3/3/3	–	–	*Rv3651–Rv3652*	–	–	–	3/3/3	–	–	–
lppJ	–	–	–	–	3/3/3	2/2/2	–	*Rv3765c–Rv3766*	–	–	–	–	–	–	2/2/2
dlaT	–	–	–	4/4/4	–	–	–	*Rv3796–fadE35*	–	–	–	–	–	–	2/2/2
Rv2274c	–	–	–	–	–	2/2/2	–	*whiB6–Rv3863*	–	5/5/5	–	–	3/3/3	–	–

Shown is the number of isolates in which the gene or IGR is mutated, the number of SNPs within the gene or IGR, and the total number of SNPs in the gene or IGR in all isolates carrying mutations in the gene or IGR. Drug resistance groups: INH-RMP, isoniazid-rifampicin; STR, streptomycin; EMB, ethambutol; OFX, ofloxacin; ETH, ethambutol; KAN, kanamycin; CPM, capreomycin.

[a]Genes or IGRs that contain relatively high-frequency SNPs.

indicate that 30 of the 32 drug resistance–associated IGRs express sRNAs, with 8 containing drug resistance–associated SNPs in sRNA-encoding regions (**Supplementary Table 16**). The role of IGRs in drug resistance therefore deserves greater attention.

Of the approximately 774 genes considered to be essential for *M. tuberculosis* growth[47], those with a high density of SNPs in drug-resistant isolates are likely to be useful drug targets or important players in drug resistance mechanisms. Consistent with this idea, 24 of the 84 drug resistance–associated genes from our analysis, including 10 of 12 well-known drug resistance–associated genes and *rrs*, are considered to be essential (**Fig. 2** and **Supplementary Table 7**). The role of the 14 newly identified drug resistance–associated essential genes in drug resistance mechanisms thus deserves further attention.

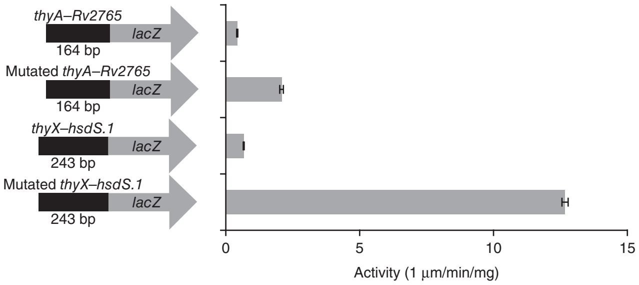

Figure 3 IGR SNPs alter the expression level of downstream genes. The effects of SNPs in the *thyA–Rv2765* and *thyX–hsdS.1* IGRs on the expression of *lacZ* in *M. smegmatis* were examined by performing β-D-galactosidase activity assays on IGR-*lacZ* constructs. The IGR sequences upstream of *thyA* and *thyX* were inserted into the pSD5B mycobacterial shuttle vector that contains a promoterless *lacZ* reporter gene immediately downstream of the cloning site. IGR-*lacZ* constructs are shown on the left. The *thyA–Rv2765* and *thyX–hsdS.1 lacZ* constructs, with or without the SNPs identified in this study, were electroporated into *M. smegmatis*, and β-D-galactosidase activity was assayed. Data presented are mean values from three independent experiments. Error bars, 95% confidence intervals.

SNPs present in a high proportion of isolates, such as those in codons 516, 526 and 531 of *rpoB* associated with rifampicin resistance[37,48], are likely functionally important[37]. To identify new SNPs that might be important in drug resistance mechanisms, we used similar reasoning to that outlined above to identify drug resistance–associated genes and IGRs, finding 33 relatively high-frequency nonsynonymous SNPs and 16 IGR SNPs closely associated with drug resistance, 40 of which (**Supplementary Table 17**), including 11 new nonsynonymous SNPs and 10 IGR SNPs, were located in 11 of the 85 drug resistance–associated genes and 6 of the 32 IGRs, respectively, further validating their strong association with drug resistance. Ten of these 11 genes (*gyrA*, *rpoB*, *rpoC*, *rpsL*, *katG*, *folC*, *thyA*, *embB*, *Rv3806c* and *rrs*) are considered essential for the *in vitro* growth of *M. tuberculosis*[47]. Of these genes, *folC* and *Rv3806c* are largely uninvestigated and thus deserve attention.

In addition, we examined the distribution of lineage-specific SNPs among drug resistance–associated genes and IGRs; 21 drug resistance–associated genes and 1 IGR contained lineage-specific SNPs, providing preliminary evidence that these SNPs have contributed to phylogenetic differentiation in *M. tuberculosis* (**Supplementary Table 18**).

Our findings generally validate previous reports and suggest that traditional investigations of drug resistance mechanisms have indeed identified many of the key players in drug resistance. However, a different approach is necessary to complete the picture and reflect the layers of complexity encountered in clinical situations. Because the *M. tuberculosis* genome carries relatively few mutations[17,49] (in the current study, lineage 2 isolates had 1,302.48 ± 24.25 SNPs per genome, and lineage 4 isolates had 769.78 ± 34.74 SNPs per genome) and the overall difference in numbers of SNPs between drug-sensitive and drug-resistant isolates in these lineages is not significant (**Supplementary Fig. 6**), genome-wide sequencing is the most efficient method to comprehensively detect drug resistance–related genetic variation. Here we have analyzed a large set of isolates, taking a statistical approach designed to identify the genes, IGRs and nonsynonymous SNPs across the genome with the strongest associations with drug resistance. Our results uncover new relationships between drug resistance and previously unassociated genes and suggest that some drug resistance–associated genes and IGRs may be involved in resistance to more than one drug (**Table 2**). The precise nature of these associations, however, requires experimental validation.

Although the *M. tuberculosis* genome is under purifying selection[50], comparing dN/dS ratios (the ratio of nonsynonymous to synonymous SNPs, a measure of individual protein evolution as well as of the impact of selection on the genome)[6–8] in drug-resistant and drug-sensitive isolates (**Fig. 4** and **Supplementary Table 19**) indicates that, whereas drug pressure has exerted a small but positive selective effect on the genome (drug-sensitive isolates, 0.432; drug-resistant isolates, 0.463; $P = 6.47 \times 10^{-7}$), it has had a relatively large positive selective effect on the 84 drug resistance–associated genes (drug-sensitive isolates, 0.489; drug-resistant isolates, 1.031; $P = 3.26 \times 10^{-13}$), particularly on the 24 essential drug resistance–associated genes (drug-sensitive isolates, 0.270; drug-resistant isolates, 1.427; $P = 5.95 \times 10^{-11}$), 10 of which, as noted above, are well-known drug resistance genes. When these 84 genes were excluded from calculations, the dN/dS values for drug-resistant isolates were not significantly different from those for the whole genomes of drug-sensitive isolates (84 drug resistance–associated genes excluded, 0.441; whole genome, drug-sensitive isolates, 0.432; $P = 0.055$), suggesting that these 84 genes contain essentially all the nonsynonymous SNPs that have arisen due to drug pressure and should thus represent a near-complete set of drug resistance–associated genes for these isolates and antibiotics.

In identifying a set of new drug resistance–associated genes, IGRs and SNPs, this study has demonstrated that the genetic basis of drug resistance is more complex than previously anticipated and has laid a strong foundation for more comprehensive investigations of drug resistance mechanisms in both *M. tuberculosis* and other bacterial pathogens treated with the same antibiotic regimens.

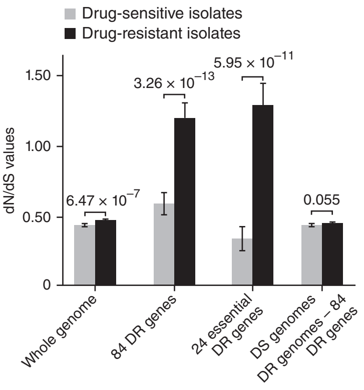

Figure 4 The 84 drug resistance–associated protein-encoding genes identified in this study contain essentially all the nonsynonymous SNPs that have arisen as a result of drug pressure in this set of isolates and antibiotics. Comparisons of dN/dS values were made between the whole genomes of drug-sensitive isolates versus those of drug-resistant isolates; the 84 drug resistance–associated genes in drug-sensitive isolates versus drug-resistant isolates; the 24 essential drug resistance–associated genes in drug-sensitive isolates versus drug-resistant isolates; and the whole genomes of drug-sensitive isolates versus the whole genomes of drug-resistant isolates when the 84 drug resistance–associated genes are excluded. The dN/dS value for the genomes of drug-resistant isolates is not significantly different from that of drug-sensitive isolates when the 84 drug resistance–associated genes are excluded. The dN/dS ratio is expected to be near unity in the absence of selection, <1 under purifying selection and >1 under positive selection. DR, drug resistant; DS, drug sensitive. Data presented are mean values. Error bars, 95% confidence intervals. *P* values (calculated using two-tailed Mann-Whitney *U* tests) for the appropriate comparisons are shown above the bars.

URLs. STITCH V3.1, http://stitch.embl.de/; TreeBeST, http://treesoft.sourceforge.net/treebest.shtml; PAML, http://abacus.gene.ucl.ac.uk/software/paml.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing reads have been submitted to the NCBI Sequence Read Archive (SRA) under accession SRA065095.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank S.M. Wang for helpful discussions and input and the anonymous reviewers for their constructive comments. We thank D. Chatterji (Indian Institute of Science) for the pSD5B vector. L.B. was supported by the National Natural Science Foundation of China (grant 31170132), the National Basic Research Program of China (grants 2009CB825402 and 2012CB518703), the Chinese Academy of Sciences (grant KSZD-EW-Z-006) and the Key Project Specialized for Infectious Diseases of the Chinese Ministry of Health (grants 2012ZX10003002 and 2013ZX10003006).

AUTHOR CONTRIBUTIONS

H.Z. conducted the experiments. D.L., J.F., H.Z., N.L., Y.H., Z.L., Y. Zhu, S.W., M.W., Y. Zhou, J.D., J.W., R.Y., G.Z., N.G., H.H., L. Zhao, J. Zhou and J. Zhang contributed to data analysis. K.W., L. Zhao, C.L., Q.Z., L. Zhou, T.C. and F.L. collected the isolates. Y.G., H.Z., L. Zhao, N.L. and T.C. cultured the isolates and prepared DNA samples. Tong Wang performed genome sequencing and assembly. H.Z. and Ting Wang performed the IGR promoter experiments. J.F. and H.Z. wrote and revised the manuscript. X.-E.Z. directed part of the work, and L.B. conceived, designed and provided overall direction for the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Gandhi, N.R. *et al.* Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet* **375**, 1830–1843 (2010).
2. Zumla, A. *et al.* Drug-resistant tuberculosis—current dilemmas, unanswered questions, challenges, and priority needs. *J. Infect. Dis.* **205** (suppl. 2), S228–S240 (2012).
3. Goldberg, D.E., Siliciano, R.F. & Jacobs, W.R. Jr. Outwitting evolution: fighting drug-resistant TB, malaria, and HIV. *Cell* **148**, 1271–1283 (2012).
4. Laurenzo, D. & Mousa, S.A. Mechanisms of drug resistance in *Mycobacterium tuberculosis* and current status of rapid molecular diagnostic testing. *Acta Trop.* **119**, 5–10 (2011).
5. Zhang, Y. & Yew, W.W. Mechanisms of drug resistance in *Mycobacterium tuberculosis*. *Int. J. Tuberc. Lung Dis.* **13**, 1320–1330 (2009).
6. Elena, S.F. & Lenski, R.E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**, 457–469 (2003).
7. Barrick, J.E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247 (2009).
8. Woods, R., Schneider, D., Winkworth, C.L., Riley, M.A. & Lenski, R.E. Tests of parallel molecular evolution in a long-term experiment with *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **103**, 9107–9112 (2006).
9. World Health Organization. *Global Tuberculosis Control 2011* (World Health Organization, Geneva, 2011).
10. Zhao, Y. *et al.* National survey of drug-resistant tuberculosis in China. *N. Engl. J. Med.* **366**, 2161–2170 (2012).
11. Ford, C.B. *et al. Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* **45**, 784–790 (2013).
12. Casali, N. *et al.* Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res.* **22**, 735–745 (2012).
13. Walker, T.M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
14. Ford, C.B. *et al.* Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat. Genet.* **43**, 482–486 (2011).
15. Ioerger, T.R. *et al.* Genome analysis of multi- and extensively-drug-resistant tuberculosis from KwaZulu-Natal, South Africa. *PLoS ONE* **4**, e7778 (2009).
16. Cole, S.T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544 (1998).
17. Comas, I. *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503 (2010).
18. Hershberg, R. *et al.* High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* **6**, e311 (2008).
19. Gagneux, S. & Small, P.M. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *Lancet Infect. Dis.* **7**, 328–337 (2007).
20. Müller, B., Borrell, S., Rose, G. & Gagneux, S. The heterogeneous evolution of multidrug-resistant *Mycobacterium tuberculosis*. *Trends Genet.* **29**, 160–169 (2013).
21. Ramaswamy, S. & Musser, J.M. Molecular genetic basis of antimicrobial agent resistance in *Mycobacterium tuberculosis*: 1998 update. *Tuber. Lung Dis.* **79**, 3–29 (1998).
22. Sandgren, A. *et al.* Tuberculosis drug resistance mutation database. *PLoS Med.* **6**, e2 (2009).
23. Sekiguchi, J. *et al.* Detection of multidrug resistance in *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **45**, 179–192 (2007).
24. Zaunbrecher, M.A., Sikes, R.D. Jr., Metchock, B., Shinnick, T.M. & Posey, J.E. Overexpression of the chromosomally encoded aminoglycoside acetyltransferase *eis* confers kanamycin resistance in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **106**, 20004–20009 (2009).
25. World Health Organization. *Guidelines for Surveillance of Drug Resistance in Tuberculosis* (World Health Organization, Geneva, 2009).
26. World Health Organization. *Treatment of Tuberculosis: Guidelines for National Programmes* 4th edn. (World Health Organization, Geneva, 2009).
27. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
28. Mohanty, D., Sankaranarayanan, R. & Gokhale, R.S. Fatty acyl-AMP ligases and polyketide synthases are unique enzymes of lipid biosynthetic machinery in *Mycobacterium tuberculosis*. *Tuberculosis (Edinb.)* **91**, 448–455 (2011).
29. Schroeder, E.K., de Souza, N., Santos, D.S., Blanchard, J.S. & Basso, L.A. Drugs that inhibit mycolic acid biosynthesis in *Mycobacterium tuberculosis*. *Curr. Pharm. Biotechnol.* **3**, 197–225 (2002).
30. Heath, R.J., White, S.W. & Rock, C.O. Lipid biosynthesis as a target for antibacterial agents. *Prog. Lipid Res.* **40**, 467–497 (2001).
31. Birch, H.L. *et al.* Biosynthesis of mycobacterial arabinogalactan: identification of a novel α(1→3) arabinofuranosyltransferase. *Mol. Microbiol.* **69**, 1191–1206 (2008).
32. Gavalda, S. *et al.* The Pks13/FadD32 crosstalk for the biosynthesis of mycolic acids in *Mycobacterium tuberculosis*. *J. Biol. Chem.* **284**, 19255–19264 (2009).
33. Domenech, P., Reed, M.B. & Barry, C.E. III. Contribution of the *Mycobacterium tuberculosis* MmpL protein family to virulence and drug resistance. *Infect. Immun.* **73**, 3492–3501 (2005).
34. La Rosa, V. *et al.* MmpL3 is the cellular target of the antitubercular pyrrole derivative BM212. *Antimicrob. Agents Chemother.* **56**, 324–331 (2012).
35. Tahlan, K. *et al.* SQ109 targets MmpL3, a membrane transporter of trehalose monomycolate involved in mycolic acid donation to the cell wall core of *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **56**, 1797–1809 (2012).
36. Deidda, D. *et al.* Bactericidal activities of the pyrrole derivative BM212 against multidrug-resistant and intramacrophagic *Mycobacterium tuberculosis* strains. *Antimicrob. Agents Chemother.* **42**, 3035–3037 (1998).
37. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44**, 106–110 (2012).
38. Ramaswamy, S.V. *et al.* Single nucleotide polymorphisms in genes associated with isoniazid resistance in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* **47**, 1241–1250 (2003).
39. Rindi, L. *et al.* Mutations responsible for *Mycobacterium tuberculosis* isoniazid resistance in Italy. *Int. J. Tuberc. Lung Dis.* **9**, 94–97 (2005).
40. Gagneux, S. *et al.* Impact of bacterial genetics on the transmission of isoniazid-resistant *Mycobacterium tuberculosis*. *PLoS Pathog.* **2**, e61 (2006).
41. Fivian-Hughes, A.S., Houghton, J. & Davis, E.O. *Mycobacterium tuberculosis* thymidylate synthase gene *thyX* is essential and potentially bifunctional, while *thyA* deletion confers resistance to *p*-aminosalicylic acid. *Microbiology* **158**, 308–318 (2012).
42. Reese, M.G. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.* **26**, 51–56 (2001).
43. Nellen, W. & Hammann, C. *Small RNAs: Analysis and Regulatory Functions* (Springer, Heidelberg, Germany, 2005).
44. Song, T. & Wai, S.N. A novel sRNA that modulates virulence and environmental fitness of *Vibrio cholerae*. *RNA Biol.* **6**, 254–258 (2009).
45. Arnvig, K.B. *et al.* Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathog.* **7**, e1002342 (2011).
46. Miotto, P. *et al.* Genome-wide discovery of small RNAs in *Mycobacterium tuberculosis*. *PLoS ONE* **7**, e51950 (2012).
47. Griffin, J.E. *et al.* High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog.* **7**, e1002251 (2011).
48. Telenti, A. *et al.* Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis*. *Lancet* **341**, 647–650 (1993).
49. Koenig, R. Few mutations divide some drug-resistant TB strains. *Science* **318**, 901–902 (2007).
50. Fleischmann, R.D. *et al.* Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**, 5479–5490 (2002).

ONLINE METHODS

Bacterial isolates. Several hundred isolates from patients with active disease were obtained from 12 provinces and municipalities across China, namely Beijing, Fujian, Guangdong, Guizhou, Heilongjiang, Henan, Shaanxi, Shanghai, Liaoning, Inner Mongolia, Xinjiang and Tibet, via the Chinese Center for Disease Control and Prevention. After genetic identification of isolates as strains of *M. tuberculosis* and testing for drug sensitivity, 161 isolates were selected for further study on the basis of their drug resistance profiles, including 44 drug-sensitive isolates, 94 MDR isolates and 23 XDR isolates (**Supplementary Fig. 1** and **Supplementary Table 1**). All isolates were cultured on Lowenstein-Jensen slants at 37 °C for 4–6 weeks.

Drug susceptibility testing. Drug susceptibility was tested on solid medium using the standard proportion method recommended by the World Health Organization[51]. Drug concentrations used for testing were 4.0 μg/ml for streptomycin (STR), 0.2 μg/ml for isoniazid (INH), 40.0 μg/ml for rifampicin (RMP), 2.0 μg/ml for ethambutol (EMB), 40.0 μg/ml for capreomycin (CPM), 30.0 μg/ml for kanamycin (KAN), 2.0 μg/ml for ofloxacin (OFX) and 40.0 μg/ml for ethionamide (ETH). Isolates were classified into the following drug resistance groups during data analysis on the basis of consistent drug resistance profiles obtained from three replicate tests (**Supplementary Table 1**): INH-RMP resistant ($n = 117$), EMB resistant ($n = 67$), STR resistant ($n = 83$), CPM resistant ($n = 22$), KAN resistant ($n = 21$), OFX resistant ($n = 54$) and ETH resistant ($n = 36$), drug sensitive ($n = 44$), MDR ($n = 94$) and XDR ($n = 23$).

Spoligotyping. To investigate the population structure of the 161 isolates, we performed spoligotyping by hybridizing amplified DNA from the direct repeat region by reverse line blotting to a Hybond membrane to which 43 oligonucleotides derived from the spacer sequences of *M. tuberculosis* H37Rv and *Mycobacterium bovis* BCG P3 were covalently bound, as previously described by Kamerbeek *et al.*[52]. Results were compared with the SpolDB4 spoligotyping database[53].

Genome sequencing and SNP identification. DNA libraries were constructed with genomic DNA extracted using a CTAB method[54] with kits provided by Illumina according to the manufacturer's instructions. PCR-free libraries with an insert size of 350 bp were prepared for each isolate. Methods for DNA manipulation, including the formation of single-molecule arrays, cluster growth and paired-end sequencing, were performed on an Illumina HiSeq 2000 sequencer according to standard protocols. The Illumina base-calling pipeline (version HCS1.4/RTA1.12) was used to process raw fluorescent images and call sequences. Raw reads of low quality from paired-end sequencing (those with consecutive bases covered by fewer than five reads) were discarded.

Paired-end reads from each strain (and those from 21 published genomes[17]) were mapped to the H37Rv reference genome (GenBank accession NC_000962) using SOAP2 (ref. 55). Base coverage of each position of H37Rv was assessed using an in-house C/C++ program. Bases with a quality score of <20 were filtered out. To validate the resulting non-redundant candidate SNPs in H37Rv and the alleles of the other 182 genomes, the numbers of the most abundant (n1) and the second most abundant (n2) nucleotides at each SNP in each strain (counted according to the number of reads in each strain supporting the presence of the nucleotide) were examined. High-quality SNPs satisfied the following criteria: (i) the most abundant base was different from that in the H37Rv genome, (ii) $n1 + n2 \geq 10$ and (iii) $n1/n2 \geq 5$. SNPs called in repetitive regions of the H37Rv genome, defined as exact repetitive sequences of ≥25 bp in length, identified using either BLAST, RepeatMasker or Trf[56,57] were excluded. If at least 95% of the strains had a non-redundant SNP in a certain position, it was included in the SNP set.

To assess the accuracy of our sequencing, we amplified the *rpoB* gene from a random selection of 80 isolates and sequenced the PCR products on an ABI 3730 sequencer (Applied Biosystems). There was 100% consensus between Illumina HiSeq 2000 and ABI 3730 sequencing results.

Phylogenetic analysis. We further filtered the unique high-quality SNPs to obtain a set of high-quality SNPs present in all strains that were used to construct a phylogenetic tree of our 161 isolates and 22 published genomes[17] on the basis of maximum likelihood using TreeBeST. A strain of *Mycobacterium canettii* was included as an outlier[18]. We reconstructed ancestral alleles according to the phylogenetic tree using PAML.

Filtering of phylogenetically related SNPs likely to influence the accuracy of selecting drug resistance–associated genes and IGRs. We further filtered our SNP data set to remove (i) phylogenetically related SNPs, defined as those that were present only within one clade, and (ii) SNPs present in two or more different subclades, one of which contained at least one drug-sensitive isolate (**Supplementary Fig. 3a**). The remaining SNPs were analyzed to identify drug resistance–associated genes, IGRs and SNPs.

Identification of lineage-specific SNPs. Lineage 2 and 4 isolates and their sublineage isolates, identified according to the phylogenetic tree, were compared separately with the ancestral sequence to identify linage 2–, lineage 4– and sublineage-specific SNPs. Lineage-specific SNPs were defined as those present only in all lineage 2 or all lineage 4 isolates or in their sublineage isolates.

Identification of drug resistance–associated genes, IGRs and SNPs. Genes and IGRs related to drug resistance were identified as outlined in **Supplementary Figure 3**. Two independent methods were used.

(1) Although SNPs generally occur randomly in any given sequence (gene, IGR or genome) or isolate, in line with the Poisson distribution, drug-related selection pressures will likely give rise to sequences and strains with more SNPs than predicted by the Poisson distribution. We identified such genes and IGRs by examining the distribution of nonsynonymous SNPs and IGR SNPs for each gene and IGR, respectively, among the 117 drug-resistant isolates (**Supplementary Tables 3** and **4** contain genes and IGRs whose density of nonsynonymous SNPs and IGR SNPs was greater than anticipated by the Poisson distribution ($P < 0.05$)).
(2) Identification of genes and IGRs mutated more frequently in isolates resistant to a given drug than in all other isolates. Nine drug resistance groupings were used to separate the influence of SNPs related to MDR and XDR (all drug-resistant isolates were rifampicin and isoniazid resistant) from those associated with resistance to other drugs. The proportion of drug-resistant isolates (F_R) in which a given gene or IGR was mutated minus the proportion of drug-sensitive isolates (F_S) in which the same gene or IGR was mutated (i.e., $F_R - F_S$) was calculated, and the mean value and standard deviation of this data set was used to obtain the quantile P value ($P < 0.01$) of the corresponding normal distribution of each gene or IGR (**Supplementary Tables 5** and **6**).

Genes and IGRs selected by both methods were considered to be strongly associated with drug resistance.

Similar reasoning was used to identify the nonsynonymous SNPs and IGR SNPs most closely related to drug resistance; nonsynonymous SNPs and IGR SNPs whose distributions among the 117 drug-resistant isolates did not fit the Poisson distribution ($P < 0.05$) were compared with the nonsynonymous SNPs and IGR SNPs, respectively, represented in a higher proportion of drug-resistant than drug-sensitive isolates ($F_R - F_S > 0.00$; quantile $P < 0.01$) in at least one of the nine drug profile groups. The resulting nonsynonymous SNPs and IGR SNPs were designated as relatively high-frequency nonsynonymous SNPs and IGR SNPs, and their distributions among the genes and IGRs strongly associated with drug resistance were examined.

Logistic regression. To estimate the strength of association between each drug resistance–associated gene or IGR and resistance to each drug, we performed logistic regression to adjust for preexisting drug resistances. The sequence of resistance accumulation was estimated to be INH-RMP, STR, EMB, OFX, ETH, KAN and CPM (resistance to RMP was completely confounded with resistance to INH). We fitted a regression model of the form 'resistance of interest ~resistances earlier in sequence + genetic variant (drug resistance–associated gene/IGR)'. For example, the model for estimating the association of resistance to ETH with *Rv0005* was ETH ~STR + INH + EMB + OFX + *Rv0005*. Models were fitted by logistic regression because drug resistance is a binary response variable, assuming a binomial distribution of the response. The null hypothesis

doi:10.1038/ng.2735

of no association (H_0) was tested using the *F* statistic by comparing the full model (above) with a reduced model from which the term for the drug resistance–associated gene or IGR was omitted. The *P* value obtained from this significance test is an approximation, which may be poor if only a small proportion of isolates carry the genetic variant or show resistance.

Construction of mutated IGRs. IGR sequences upstream of *thyX* (*Rv2754c*) and *thyA* (*Rv2764c*) were amplified by PCR, and sequences containing point mutations corresponding to IGR SNPs identified in this study were synthesized by Biomed (Beijing, China). Amplified products were digested with XbaI and SphI and then ligated into pSD5B, a mycobacterial shuttle vector with a promoterless *lacZ* reporter gene immediately downstream of the cloning site (kindly provided by D. Chatterji). Vector constructs were electroporated into *Escherichia coli* and selected on medium containing 30 μg/ml kanamycin. Inserts were validated by sequencing, and selected constructs were then electroporated into *M. smegmatis* and selected for on medium containing 30 μg/ml kanamycin.

β-D-galactosidase assays. Cultures (5 ml) of *M. smegmatis* containing vector constructs were grown to an OD_{600} of 2.0, collected by centrifugation and resuspended in 0.5 ml of 100 mM sodium phosphate buffer (pH 7.5). Bacteria were ruptured by ultrasonication, and supernatants collected after centrifugation were incubated in 0.1 M sodium phosphate (pH 7.5) containing 0.9 mg/ml *o*-nitrophenyl-β-D-galactopyranoside (ONPG), 1 mM $MgCl_2$ and 45 mM β-mercaptoethanol for 15–30 min at 37 °C. Reactions were stopped by the addition of 1 M Na_2CO_3. Optical density was read at 420 nm, and β-D-galactosidase activities were calculated.

Calculation of dN/dS values. We inferred the ancestral and mutated alleles of each SNP in the whole population on the basis of the phylogenetic tree using PAML. dN/dS values were calculated using the KaKs_calculator[58].

51. World Health Organization. *Policy Guidance on TB Drug Susceptibility Testing (DST) of Second-Line Drugs* (World Health Organization, Geneva, 2008).
52. Kamerbeek, J. *et al.* Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* **35**, 907–914 (1997).
53. Demay, C. *et al.* SITVITWEB—a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology. *Infect. Genet. Evol.* **12**, 755–766 (2012).
54. van Soolingen, D., Hermans, P.W., de Haas, P.E., Soll, D.R. & van Embden, J.D. Occurrence and stability of insertion sequences in *Mycobacterium tuberculosis* complex strains: evaluation of an insertion sequence–dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J. Clin. Microbiol.* **29**, 2578–2586 (1991).
55. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
56. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
57. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* Chapter 4, Unit 4.10 (2009).
58. Zhang, Z. *et al.* KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**, 259–263 (2006).