Genetically encoded transcriptional plasticity underlies stress adaptation in Mycobacterium tuberculosis

Cheng Bei[1#], Junhao Zhu[2#], Peter H Culviner[2], Eric J. Rubin[2], Sarah M Fortune[2], Qian Gao[1,3*], Qingyun Liu[2*]

[1]Key Laboratory of Medical Molecular Virology (MOE/NHC/CAMS), School of Basic Medical Science, Shanghai Medical College, Shanghai Institute of Infectious Disease and Biosecurity, Fudan University, Shanghai, China.

[2]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA.

[3]National Clinical Research Center for Infectious Diseases, Shenzhen Third People's Hospital, Shenzhen, Guangdong Province, China

[#] Equal contributions.

[*] Correspondence: Qian Gao (qiangao@fudan.edu.cn), Qingyun Liu (qingyunliu@hsph.harvard.edu)

Abstract

Transcriptional regulation is a critical adaptive mechanism that allows bacteria to respond to changing environments, yet the concept of transcriptional plasticity (TP) remains largely unexplored. In this study, we investigate the genome-wide TP profiles of *Mycobacterium tuberculosis* (*Mtb*) genes by analyzing 894 RNA sequencing samples derived from 73 different environmental conditions. Our data reveal that *Mtb* genes exhibit significant TP variation that correlates with gene function and gene essentiality. We also found that critical genetic features, such as gene length, GC content, and operon size independently impose constraints on TP, beyond trans-regulation. By extending our analysis to include two other *Mycobacterium* species -- *M. smegmatis* and *M. abscessu*s -- we demonstrate a striking conservation of the TP landscape. This study provides a comprehensive understanding of the TP exhibited by mycobacteria genes, shedding light on this significant, yet understudied, genetic feature encoded in bacterial genomes.

Introduction

Cells must swiftly modulate the expression of their genes to cope with abrupt changes in external environment. Transcriptional plasticity (TP)[1] – the ability to alter the expression of a gene in response to different types of environmental stress – is pivotal to cellular adaptation and subject to natural selection[2-4]. In practice, TP can be estimated by quantifying the change in the level of expression across multiple conditions. For instance, Urchueguía et al. used a library of *E. coli* strains containing promoter-GFP (Green Fluorescence Protein) fusions to measure changes in fluorescence levels across different conditions, thereby quantifying expression plasticity[4]. Similarly, Lehner et al. used the normalized sum of squares of log2- expression changes to infer gene-level transcriptional plasticity from a *Saccharomyces. cerevisiae* microarray dataset[5]. These studies found that certain genetic traits, such as promoter architecture, nucleosome organization, and histone modification patterns may significantly influence eukaryotic gene transcriptional plasticity[6-10]. While the transcriptional machinery and the nucleoid organization of prokaryotic organisms fundamentally differ from those of eukaryotes[11,12], a recent investigation into *E. coli* promoter evolution showed that long-term natural selection favors the retention of high promoter TP despite the presence of segregating mutations[2]. The strong evolutionary constraint implies that, akin to eukaryotes, there may also be genetic traits in bacteria that determine TP, but the biological principles underlying TP in bacteria have not been adequately studied[4,13].

Exploring the genetic features contributing to TP in bacteria can enhance our understanding how of bacteria adapt to environmental pressures and guide the development of innovative strategies to combat bacterial pathogens. Tuberculosis (TB) remains the leading cause of death due to a single infectious agent[14]. Throughout the phases of infection, proliferation and transmission, *Mycobacterium* *tuberculosis* (*Mtb*), the causative agent of TB, faces a wide array of environmental challenges. Some of the stresses, such as hypoxia, are characteristic of the microenvironments where the bacilli reside within host, whereas others arise from host immune defenses such as toxic metal ions, nutrient restriction, acidic pH, and reactive oxygen or nitrogen species, etc. Over the past 75 years, *Mtb* has also faced constant pressure from antibiotics. To try understand how *Mtb* modulates its gene expression in response to different external challenges, studies have leveraged RNA sequencing (RNA-Seq) to query *Mtb*'s transcriptomic profiles across a broad panel of environmental conditions. These studies have revealed a complex transcriptional regulation network underlying the ability of *Mtb* to adapt to stresses. For example, over 50 transcriptional factors (such as *dosR* and *whiB3*) respond to hypoxia, allowing *Mtb*, an obligate aerobe, to survive in settings with oxygen depletion[15]. As these studies have been conducted under a multitude of experimental conditions, the resultant RNA-Seq datasets provide a comprehensive view of gene expression in *Mtb* that can be analyzed for insights into its transcriptional plasticity.

In this work, we systematically examine the TP profiles of *Mtb* genes by integrating publicly available RNA-Seq datasets. Our analysis uncovers significant variability in TP across genes and identifies overarching principles governing the amplitude of TP. We find a correlation between a gene's biological function and its TP and note that essential genes exhibit significantly lower levels of TP than non-essential genes. We further demonstrate that in addition to transcriptional regulators, genetic features such as operon architecture, gene length, and GC content (GC%) also appear to play substantial and distinct roles in shaping the TP of *Mtb* genes. In addition, by extending our study to *M. smegmatis* and *M. abscessus*, we show that the same principles appear to govern TP in other Mycobacteria. The findings in this study enrich our understanding of TP regulation and underscore the shared regulatory mechanisms governing gene expression dynamics.

Results

Quantifying the transcriptional plasticity of Mtb genes

To explore the transcriptome-wide pattern of gene expression in *Mtb*, we collected 894 previously published *Mtb* RNA-Seq datasets that were generated under a wide range of experimental conditions. All of the 894 datasets were obtained by studying the standard laboratory strain *Mtb* H37Rv, thus interrogating the physiological responses to various challenges in the same genetic background. These studies included antibiotic exposures, varied nutrient sources, host-mimicking conditions, and genetic manipulations such as gene knock-downs or deletions, as well as the corresponding untreated controls (Fig. 1a, see also *Methods* and Table S1). We reasoned that the wide diversity of these experimental conditions would provide a suitable resource for studying the transcriptional plasticity of *Mtb* genes (Fig. 1b).

We first employed standardized preprocessing criteria to facilitate analysis of the 894 RNAseq datasets (*Methods*). In brief, we excluded genes shorter than 150 bp, non-coding transcripts, and genes whose expression was not detected in most samples. We then normalized the expression data for the remaining 3,891 genes using the Trimmed Mean of M-values (TMM) method, a technique designed to account for varying sequencing depth and suppress batch effects (Table S2). For subsequent TP analysis, the expression values were indicated using log2-transformed Reads Per Kilobase Million (RPKM+1).

To estimate variations in gene expression, we initially calculated the range of expression values, or the *MinMax*, of the *Mtb* genes across the 894 samples. We noticed that the *MinMax* of *Mtb* genes varied from 2.8 to 18.1 (Fig. S1a), suggesting that the amplitude of the changes in the level of expression for certain *Mtb* genes could exceed the range of expression of other genes by a factor of more than 40,000.

We then examined gene expression at different percentiles of expression, including the most highly expressed 100th percentile (Max), the 75th (Q75), 50th (Median), 25th (Q25), and 1st (Min) percentiles - and observed significant differences in ranges of expression among *Mtb* genes (Fig. 1c). For instance, *hspX* - encoding a hypoxia-induced small heat shock protein[16]- displayed a markedly broader range of expression compared to *rpoB*, which encodes the β subunit of the RNA polymerase core enzyme. Conversely, the expression level of the lipoprotein peptidase gene, *lpqM*, remained relatively constant across all conditions (Fig. 1c).

We further characterized variations in expression with two additional metrics: the **I**nter-**Q**uantile-**R**ange (*IQR*) and the mean-adjusted **S**tandard **D**eviation (*adj-SD*) of the expression values (Fig. S1b, *Methods*). As expected, we found a high degree of correlation between *MinMax*, *IQR*, and *adj-SD* (Fig. S1c), indicating that these measures all represent variability of gene expression. To evaluate the robustness of these metrics, we performed a bootstrap analysis by comparing random subsamples with the complete dataset (*Methods*). This analysis indicated that while both *IQR* and *adj-SD* were more resilient to reductions in sample size than *MinMax* (Fig. 1d), *adj-SD* demonstrated a slight, but statistically significant advantage over *IQR* (Fig. 1d). Therefore, *adj-SD* was used to estimate TP in the subsequent analyses.

TP varies with gene function and gene essentiality

The calculated TPs for 3,891 *Mtb* genes displayed a predominantly normal distribution with a long tail representing genes with high TPs (Fig. 1e). Using a bootstrap approach similar to that described in Fig. 1d, we found that the 195 high-TP genes in the top 5% percentile demonstrated consistently high TP even when the sample size was reduced to just 10 genes (Fig. S1d, e). This pattern suggests that the skewed distribution wasn't caused by "outlier" values, but instead reflects a subset of genes with a wider range of expression levels. We then investigated the biological functions of the high-TP genes and found that the 195 high-TP genes (Fig. S2) were significantly enriched for genes involved in responding to stress, including hypoxia, host immune mechanisms, copper ions, etc., as per the DAVID database[17] (Fig. 2a). When we grouped *Mtb* genes based on previously established functional categories and compared their TP profiles[18,19], we found that genes involved in biomass production, cell wall biosynthesis, cellular metabolism, and respiration were primarily associated with the lowest TPs (Fig. 2b). This association is underscored by our observation that core genes conserved across mycobacteria species exhibited significantly lower TPs compared to the other genes in the genome (Fig. 2c).

The above findings suggested that those genes crucial for basic cellular activities exhibit more tightly regulated expression. To test this, we compared the TP distribution for genes previously annotated as essential with those annotated as not essential[20], and found that essential genes displayed significantly lower TPs than non-essential genes (Fig. 2d). We also noticed that those genes whose disruption by transposon insertion conferred a growth advantage exhibited significantly higher TPs than both essential and other non-essential genes (Fig. 2d). Recent studies have proposed gene vulnerability – the organism's susceptibility to perturbations in the transcription of the gene (e.g., by CRISPRi) – as a quantitative, orthogonal proxy to gene essentiality[21]. Consistent with the analysis by annotated essentiality, we found that genes identified as vulnerable also tended to exhibit lower *TP*s (Fig. 2e), and none of the highly vulnerable genes exhibited high TP (Fig. 2e). We hypothesized that high-TP genes may promote phenotypic diversification that confers a selective advantage in the ability of *Mtb* to survive in fluctuating environments, and therefore these genes might accumulate mutations more rapidly than the rest of the genome. To test this hypothesis, we utilized a recently established set of evolutionary metrics for *Mtb* genes, drawn from 10,209 *Mtb* genomes[22]. Consistent with our hypothesis, we found that high-TP genes exhibited higher base substitution rates than low-TP genes

(Fig. 2f). Overall, our analysis suggests that for those genes involved in essential cellular processes, stable levels of expression are advantageous to the bacteria. In contrast, for genes that provide a growth advantage in certain conditions, but are dispensable or even detrimental in others, a "plastic", inducible transcriptional program appears to be beneficial.

Genetic features underlying transcriptional plasticity

To identify the genetic factors influencing TPs of *Mtb* genes, we compiled a comprehensive list of 78 genetic features including sequence composition, transcriptional regulation and evolutionary parameters (Fig. 3a, Table S3, *Methods*). We then employed a decision-tree-based regression analysis to model the *Mtb* TP landscape with these 78 features (Fig. 3b). The regression model was trained on a randomly selected subset of 60% (2,335/3,891) of the total *Mtb* genes, and then used to predict the TPs of the remaining 40% (1,556/3,891) of *Mtb* genes. We iterated this process 100 times, with the derived models yielding an average R^2 value of 0.16 (Fig. S3a). For each model, the features were ranked by importance based on the contribution of each feature to the predictive power of the model. We then aggregated these feature ranks across all iterations to provide an average measure of each feature's contribution to TP prediction.

Our analysis highlighted four features – operon length, gene length, number of activating regulators, and GC percentage (GC%) – that consistently demonstrated high predictive importance across iterations (Fig. 3c). A support vector machine (SVM) model trained solely with these four features was able to predict a gene's TP (R^2=0.17) with an accuracy similar to that of a model trained with all 78 features (Fig. 3d). The feature contributing most to the predictive power of the SVM model was operon size, followed by gene length, number of activating regulators and GC% (Fig. S3b), consistent with the results of the decision-tree-based regression analysis (Fig. 3c). However, there was no correlation between the top features ($r<0.21$, Fig. S3c), indicating that these features play independent roles in shaping transcriptional plasticity.

The role of genetic features in affecting transcriptional plasticity

We then sought to understand how these feature influence TP. We first examined the role of gene length and found a negative correlation between gene length and TP, with longer genes tending to exhibit lower TPs than shorter genes (Fig. 4a). Contrary to gene length, however, the correlation between TP and GC content (GC%) was not monotonic. We found that genes with a GC% substantially different from the average for the *Mtb* genome (65.6%) generally had higher TPs (Fig. 4b, Fig. S4a). To confirm this observation, we binned *Mtb* genes according to their TPs and calculated the standard deviation (SD) for the median GC% of the genes in each bin. We observed an apparent linear correlation between the SD of the median GC% and the ranks of TP bins, such that the bins with higher TPs had larger SDs for GC%. This corroborated the hypothesis that the TP increases with greater GC% deviation from the genome average (Fig. 4c). Notably, both essential and non-essential genes whose GC% approximated the genome-average GC% exhibited lower TPs (Fig. S4b-c), implying that the association between GC% and TP was not confounded by gene essentiality.

Next, we evaluated the effect of operon size on TP. We found that genes located in polygenic operons, containing two or more genes, had significantly higher TPs than genes located in monogenic operons, consisting of only one gene (Fig. 4d). Furthermore, we also observed that the TPs of genes within the same operon were highly correlated (Fig. S4d). Despite the confounding TP differences between essential and non-essential genes, both exhibited higher TPs in polygenic operons (Fig. S4e-f). A recent study reported that *Mtb* undergoes frequent premature transcription termination[23], and we observed a decreased mean expression for downstream genes in an operon (Fig. S4g), but there was no similar trend for TP (Fig. S4h). Together, these analyses suggested that it is the size of the operon, rather than the position of the genes within the operon, that influences TP.

While gene length, GC%, and operon size are features related to the primary sequence of the gene, the number of activating regulators is a feature that pertains to process of transcriptional regulation. We found that the TP of a gene tended to be higher when its expression was modulated by a higher number of predicted transcriptional activators (Fig. 4e). We also observed a similar trend for transcriptional repressors, whereby genes with more predicted repressors tended to have higher TPs, although the TP dropped slightly in genes predicted to have only one repressor (Fig. 4f). Taken together, our analysis shows that not only the basic genetic composition of genes but also the complex network of transcriptional regulation can significantly influence the TP landscape of the *Mtb* genome (Fig. 4g).

Genetic features can explain TP variation in genes belonging to the same regulon

Because the different genes within a regulon are co-regulated, we speculated that they could also have similar TPs. We investigated 36 well-annotated gene regulons (*Methods*, Table S4) and found that the TP varied greatly between different regulons (Fig. 5a). For instance, the regulons Mce3R, KstR2, BkaR and FasR, which are thought to be involved in lipid metabolism, had the lowest TPs (Fig. 5a)[24-27]. By contrast, the hypoxia- and redox-sensing DosR regulon and metal related regulons such as Zur, RicR, M-box and IdeR, demonstrated high TPs (Fig. 5a). However, while the genes within the same regulon displayed similar expression patterns, coordinately regulated up or down, they differed significantly in the magnitude of their changes in expression, resulting in diverse TPs. For example, the TPs of the genes belonging to the DosR regulon varied substantially, with *dosT* exhibiting the lowest (0.74) and *hspX* exhibiting the greatest change in level of expression (3.98) (Fig. 5b-c), and similar TP variations were seen amongst the genes belonging to other regulons (Fig. 5a). Because the expression of genes within a regulon generally showed the same direction of change in response to stress, we speculated that the TP differences amongst the regulon's genes might derive from differences in the genetic features of the individual genes. Indeed, we found the two primary genetic features - gene length and GC% - could explain the TP variations of co-regulated genes in most regulons (Fig. S5-6). To show this, we selected five regulons that comprised of more than 20 genes each (*whiB1*, *whiB4*, *sigD*, *zur* and *Rv1828*/*sigH*) and demonstrated that shorter genes with a GC% deviating from the genomic average generally displayed higher TP than other co-regulated genes (Fig. 5d-e). These results highlight the ability of genetic features to affect the TP, independent of other transcriptional regulatory processes.

The transcriptional plasticity landscape is conserved across Mycobacterium species

The analyses above revealed that, in *Mtb*, a gene's TP is linked to its function, essentiality, and its evolutionary and genetic features, all of which are likely to be conserved in closely related homologous genes from other mycobacterial species. To demonstrate this, we curated published RNA-Seq datasets from 192 samples of *M. smegmatis* (*Msm*) and 106 samples of *M. abscessus* (*Mab*), and used *adj-SD* to estimate their genome-wide TP (Fig. S7a, Table S5). We found that all three species displayed long-tailed distributions at high TP values (Fig. 1e, Fig. S7b), and homologous genes in *Mtb*, *Msm*, and *Mab* showed similar amplitudes of TP (Fig. 6a-b, Fig. S7c). Moreover, as observed in *Mtb*, the essential/vulnerable genes in *Msm* exhibited lower TPs than non-essential or less-vulnerable genes (Fig. 6c-d). Also as seen in *Mtb*, the genes in *Msm* and *Mab* with higher TP values tended to be shorter in length and have GC% more deviated from the average (Fig. 6e-h). It is intriguing that the high-TP genes across all three species were enriched in iron-related functions (Fig. S7d, Fig. 3a). These findings suggest that despite the differences in natural lifestyles, the evolutionary principles underlying TP are likely conserved across mycobacterial species.

Discussion

In this work, we assessed the TP of *Mtb* genes by utilizing 894 RNA-Seq datasets that were previously collected when the bacteria were exposed to various environmental conditions. Our analyses revealed that TP varies significantly among *Mtb* genes in a manner that is associated their biological functions and subjected to natural selection. We identified primary genetic features that contribute to TP values, including gene length, GC%, operon size and transcriptional regulatory factors. Finally, we extended these findings to *Msm* and *Mab*, demonstrating that TP, and the factors that influence it, are likely to be biological features that are conserved across mycobacterial species.

Gene vulnerability reflects the quantitative association between changes in bacterial fitness and the degree of CRISPR-i mediated inhibition of a gene's transcription[21]. Perturbing the expression of highly vulnerable genes can be deleterious, whereas the same level of expression inhibition of invulnerable genes can be tolerated[21]. Initially, we anticipated a linear-like relationship between TP and vulnerability, whereby more vulnerable genes would exhibit lower TPs. Although we observed a positive association between vulnerability and TP, this relationship could not be explained by a simple or log-linear model. Instead, we observed an intriguing pattern between vulnerability and TP that presented as a reversed "L"-shape, with the elbow point representing genes that were insensitive to transcription inhibition and invariant in expression. This observation could be due to several reasons. First, the effects on the bacteria caused by gene's transcriptional activation or transcriptional repression are not necessarily symmetrical. For instance, for some house-keeping genes, overexpression is better tolerated by the bacteria than repressed expression, whereas for protein toxins the outcomes would be the opposite[28]. Because TP considers both up and down-regulation of gene expression, it reflects gene-specific constraints on both transcriptional activation and repression, whereas studies of gene vulnerability and essentiality only consider transcriptional repression. Second, vulnerability is not a constant gene feature but rather is expected to vary depending on the specific environmental conditions. Therefore, we speculate that vulnerability estimated from different conditions could have a stronger correlation with TP. Finally, although essential genes showed significantly lower TP than non-essential genes, the TP variation in essential genes is overall quite close to that of non-essential genes. This suggests that bacteria may have the flexibility to alter the level of expression of essential genes as required for survival in changing environments (Fig. 2e).

It is noteworthy that genetic features play a more significant role than transcriptional regulation in determining TP, even though the mechanisms underlying this observation are not yet fully understood. For example, we found that shorter genes had higher TPs, a pattern that has been also observed in eukaryotes such as *Drosophila* and *Arabidopsis thaliana*[29,30]. The length of gene appears to be evolutionarily shaped to accommodate its functionality, with housekeeping genes tending to be longer while stress-responsive genes tend to be shorter[30-32]. We speculate that stress-responsive genes require efficient and diverse expression patterns to cope with fluctuating environments while conserving energy. A reduction in gene size may represent an adaptive strategy to achieve this efficiency, allowing for more efficient regulation of the expression of these genes in response to stress. However, further research is needed to test this hypothesis and fully understand the evolutionary relationship of gene size with stress response.

There was a significant association between gene expression patterns and GC content, indicating that GC content could be an important regulatory factor[33]. It was previously observed that AU-rich and GC-rich transcripts follow distinct decay pathways, with a linear relationship between higher GC content and greater RNA stability[34]. In our study, however, we found a "V" shaped relationship, whereby genes with low TP were clustered around a GC content of 65.6%, which is the average GC% of the *Mtb* genome. This finding contradicted our initial assumption that higher GC content would be associated with lower TP. The genes with extremely high-GC content (>75%) may result from recent horizontal gene transfer from other bacteria[35-37], and therefore one possible explanation is that the TP of these recently acquired genes has not yet been optimized to align with the local transcriptional network, resulting in noisy expression of these genes. Moreover, high GC content may have a detrimental effect on expression stability if it leads to the formation of secondary structures or interferes with the binding of regulatory factors. The clustering of low TP genes at the *Mtb* average 65.6% GC content, suggests that these genes have evolved to be both GC stable and expression stable, thereby representing an optimized state of gene regulation.

Though we successfully identified four significant contributing features, the models incorporating these features could not completely predict TP values, suggesting that there are likely other determinants that were not identified (Fig. 3d, S3a), such as the promoter. Recent work in *E. coli* showed that, for most genes, the range of protein abundance across different environmental conditions is constrained by the the TFs that regulate promoter activity[38]. Another study revealed that promoter characteristics, such as the length of the transcriptional initiation region and the presence of TATA-boxes, play important roles in determining the range of expression variation in eukaryotic genes[29]. Similarly, the positive correlation we found between TP and the number of transcriptional activators demonstrates the influence of promoter characteristics and trans regulatory mechanisms on TP in mycobacteria (Fig. 4e).

The inherent differences in the transcriptional plasticity (TP) of genes can be used to normalize expression differences in microbial transcriptional studies. Traditionally, different thresholds have been employed to identify meaningful changes in gene expression. The threshold for identifying genes that respond to particular conditions is often a 2-fold change in the level of expression or occasionally thresholds of 1.5-fold or 4-fold are used, but the genes exhibiting the largest transcriptional changes frequently receive the most attention. However, these thresholds are arbitrary because they don’t adjust for the inherent TP of each gene. As a result, high-TP genes are more likely to display changes in expression that surpass the threshold, while relatively large changes in the expression in low-TP genes may be overlooked because they don’t meet the arbitrary threshold. An alternative method would determine the degree of expression change that should be considered meaningful for each specific gene. To this end, we propose utilizing the expression changes corresponding to the 5th and 95th percentiles, based on the studies in our dataset, as a "soft-thresholding" benchmark for screening differentially expressed genes (Table S6). For instance, in the case of low-TP genes such as *lpqM* and *ribF*, the log2 fold-changes corresponding to the 95th quantile expression levels were 0.54 and 0.57, respectively, times the level of expression in the controls. An analysis using the arbitrary thresholds would miss changes in the expression of these genes that are equivalent to two standard deviations. Criteria based on the inherent TP for each gene could establish a more nuanced analysis for identifying differentially expressed genes. We believe that our integration of RNA-Seq data from 894 *Mtb* samples provides a comprehensive estimation of the transcriptional variations in *Mtb* genes across various conditions, and therefore the calculated TPs can serve as a reference for evaluating changes in expression. The TMM method employed in our analysis can be used to evaluate of the transcriptional signatures of genes of interest (Table S2). This will foster a deeper understanding of the differential gene expression landscape in *Mtb* and facilitate the exploration of gene-specific transcriptional patterns.

In summary, our study has characterized the landscape of TP in *Mtb* genes and established a framework for determining TP levels. This work thereby serves as a foundation for future investigation aimed at understanding the influences that determine a gene’s TP. Additionally, the proposed TP-based benchmark offers valuable guidance for the interpretation of differential expression changes in transcriptional studies. Moving forward, further research can build upon these findings to uncover the intricacies of TP and its impact on gene expression in *Mtb* and other microbial systems.

Methods

Collection and processing of RNA-Seq data

We used the keyword "tuberculosis" to search for publicly available RNA-Seq data of *Mtb* released on NCBI Sequence Read Archive (SRA) before January 1, 2022, and obtained a total of 1,084 datasets from 64 BioProjects with 47 associated research articles (Table S1). FASTQ files of all 1,084 samples were downloaded using Fastq-dump (version 2.8.0). Adaptor trimming and the removal of low quality sequencing reads were conducted using Trimmomatic (version 0.39)[39] with parameters of "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" for paired-end data and "ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" for single-end data. The filtered profiles were then mapped against the H37Rv reference genome (ASM19595v2) using Bowtie2 (version 2.2.9)[40], and duplicated reads were removed with SAMtools (version 1.9)[41].

To identify strand specificity of the RNA-Seq libraries, we measured the Pearson correlation coefficient of total read counts on two strands for each library using SAM files generated by Bowtie2. Libraries with a correlation coefficient lower than or equal to 0 would be considered as strand-specific while a coefficient higher than or equal to 0.6 would be considered as non-strand-specific. For libraries with coefficients between 0 and 0.6, we manually judged their strand specificities based on the description of the experimental design and strand specificities of other samples from the same experiment. Library read counts were then enumerated with htseq-count from the HTSeq framework (version 0.11.3)[42] using non-strand-specific or strand-specific parameters based on strand specificities identified above. Samples with small library size (< 1,000,000 reads) and from *Mtb* strains other than H37Rv were excluded. 962 samples from 58 BioProjects were eventually included for further analysis. RNA-Seq data of *Msm* (mc^{2}155) and *Mab* were collected and processed with the same pipeline used for *Mtb*. *Msm* data were mapped to the mc^{2}155 reference genome (ASM1334914v1) and *Mab* data were mapped to ASM402801v1. Included were 293 *Msm* samples from 36 BioProjects and 146 *Mab* samples from 9 BioProjects.

Quantification of transcriptional expression

Before library normalization, we removed small genes (<=150bp), non-coding transcripts (tRNA, rRNA, and annotated non-coding RNAs in the *Mtb* genome) as well as non-expressing genes (read counts in all samples were zero). Read counts from each BioProject were subsequently normalized to account for variations in library size using Trimmed Mean of M-values (TMM) factor[43], and the TMM normalized RPKMs were calculated using the edgeR package (version 3.30.3)[44]. Next, $\log_2$ (RPKM+1) were calculated and defined as transcriptional expression levels. The Shannon index (SI) was calculated for each gene using the diversity function from R package *vegan* (version 2.5-7). We then excluded samples from all three mycobacteria with a high proportion of zero-expressing genes (> 4% of total genes), and also excluded genes with low SI (SI < 6.5 in *Mtb*, < 4 in *Msm* and *Mab*) and genes that are not expressed in more than 1% of total samples. Downstream analyses thus included curated transcriptomic profiles of 894 samples and 3,891 genes from *Mtb*, 192 samples and 6,629 genes from *Msm*, 106 samples and 4,917 genes from *Mab* (Table S1).

Stress conditions of RNA-Seq samples

To investigate the diversity of selected samples, we generalized the conditions of 894 samples based on the description in each BioProject and the related research articles. We further divided these conditions into 6 groups to summarize the sample conditions (Fig. 1a, Table S1); group "Antibiotic" referred to samples treated with antibiotics and other antimicrobial compounds; group "Respiration" referred to hypoxia, reaeration, peroxide stress and nitric oxide stress; group "Genetic manipulation" referred to knockdown, knockout, complementation and over-expression of a gene; group “Nutrient” referred to alterations in carbon sources or other nutrient conditions; group “Infection” referred to samples isolated from *ex vivo* or *in vivo* infection models; group “Control” referred to the untreated control samples of each study. tSNE is archived using R package 'Rtsne' with following parameters: dims = 2, PCA = True, max_iter = 100, theta = 0.4, perplexity = 20, verbose = False.

Estimation of transcriptional plasticity (TP)

MinMax was calculated by subtracting the minimum $\log_2$ (RPKM+1) from the maximum $\log_2$ (RPKM+1) for each gene. *IQR* was calculated by subtracting the 25th percentile of $\log_2$ (RPKM+1) from the 75th percentile of $\log_2$ (RPKM+1) for each gene. Considering the underlying association between the variance and the mean of a gene’s expressions[29,45,46], the initial standard deviation (SD) measures were calibrated by an estimated global trend between the SD and the mean $\log_2$ (RPKM+1). This global trend was estimated using a local polynomial regression model (LOESS or Locally Estimated Scatterplot Smoothing) with a large sampling window with the R package *stats* (version 4.0.2; span = 0.7, degree = 1). A gene’s adjusted SD is defined as the sum of this gene’s corresponding SD residual of the LOESS fit and the global average of the LOESS fitted SD measures.

Evaluation of the robustness of expression variation metrics

To evaluate the robustness of the three expression variation metrics, MinMax, *IQR* and adjusted SD, we performed a bootstrapping analysis. Specifically, a subset of N (N=10, 20, 30, 50, 100, 200, 300, 500, or 800) samples were randomly drawn from dataset, and a Pearson’s correlation coefficient (*r*) was calculated for each metric (*MinMax*, *IQR*, or *SD*) by comparing the randomly sampled output and the corresponding metrics measured using the full dataset. This process was repeated for 100 times for each N and the means and the standard deviations of the coefficients (*r*) were depicted in Fig. 1d and Fig. S7a.

Enrichment analysis of high-TP genes

To identify high-TP genes, a density curve of adjusted SD was determined with a Gaussian kernel density function using the R package *stats* (version 4.0.2), and the high-TP subgroup consisted of genes whose TP measures were higher than the upper threshold defined by a probability cutoff of 0.05 based on the probabilistic density estimation of adjusted SD. Gene essentiality and vulnerability indices were referenced from a recently established work that leveraged genome-wide CRISPR interference (CRISPRi) and deep sequencing to render a comprehensive quantification the effect of differential transcriptional repression on cellular fitness for nearly all *Mtb* and *Msm* genes[21]. Enrichment analysis of high-TP genes was performed using the DAVID (https://david.ncifcrf.gov) online server, and enrichment results with FDR (false discovery rate) < 0.1 were considered significant.

Mycobacteria core genome

Homologous genes of mycobacteria including *Mtb*, *Msm* and *Mab* were identified by J. A. Judd et al[47]. Homologous genes existed in all three mycobacteria were identified as core genes (Fig 2c).

Collection of gene features

Gene length. To identify significant gene features that potentially contribute to TP, we first collected genome annotations of *Mtb* genes from NCBI Genome Database (ASM19595v2). Gene length was identified by the difference between start position and end position for each gene, and then divided by average length of all genes to calculate normalized length for each gene.

Codon usage. codon usage features, including codon adaptation index (CAI), codon bias index (CBI), frequency of optimal codons (Fop), effective number of codons (Nc), A/T/C/G/GC of silent 3rd codon position (A3s/T3s/C3s/G3s/GC3s), hydrophobicity (Gravy) and aromaticity (Aromo) of a protein were calculated based on gene sequences of *Mtb* H37Rv (ASM19595v2) by using CodonW (http://codonw.sourceforge.net/).

Base and amino acid composition. Based on the reference sequence of a gene, we further identified the percentage of each base type as well as percentages of GC content (GC%) and pyrimidine content (CT%) by calculating the number of each base in a gene divided by the gene length. Similarly, we calculated the percentage of each of the 20 amino acids found in the protein products of the 3,891 genes.

Start and stop codon. According to the reference genome sequence, we identified the first and the last three base of coding sequence (CDS) for each gene, referring to the start codon and the stop codon, respectively.

Direction of replication and transcription. To study the impact of conflict between replication and transcription on TP, we identified whether DNA replication and RNA transcription were in the same or opposite directions for each gene based on the strand and genome location relative to the *dif* site (2,232,640 bp) of the gene. The site of chromosomal segregation (*dif*) was identified by Cascioferro et al[48]. To be more specific, genes located on the positive strand and before the *dif* site (clockwise), or genes on the negative strand and after the *dif* site would have the same direction of replication and transcription, and *vice versa*.

Transcription factors. Considering the direct influences of transcription factors (TFs) on transcriptional expression, we collected the data of interactions between TFs and their targets from *MTB* Network Portal (http://networks.systemsbiology.net/*Mtb*). The data contained the interaction of 4,635 TF-target pairs with evidence of ChIP-seq experiments[49] and transcriptional profiling[50], including 136 TFs and 2,111 target genes. TF-target pairs were marked with 1 or -1, representing the TF was an activator or a repressor, respectively. We then counted the number of activators and repressors for each target gene based on the TF-target pairs. The number of target genes for each TF was also counted. In addition, interactions between TFs and their targets identified by ChIP-seq were also selected, including the number of targets located at intergenic and intragenic regions for each TF.

Selective pressure. Natural selective pressures (indicated as *dN/dS* ratio) on *Mtb* genes were estimated by GenomegaMap, a phylogeny-free statistical approach performed on 10,209 *Mtb* genomes to estimate substitution parameters[22], including the mean values and 95% CIs (Q2.5 and Q97.5) of *dN/dS* ratio, transition:transversion ratio, and substitution rate. The mean probability of an *dN/dS* ratio higher than 1 (Pr(*dN/dS* > 1)) and number of sites with Pr(*dN/dS*) > 1 for each gene were also included.

Transcription start sites. Features associated with a gene's transcription start site (TSS) included upstream TSS subtype (leadered or leaderless), total number of proximal TSS associated with this gene, maximum/minimum TSS coverage, and the corresponding base at the +1 position of each TSS. TSS annotations were adopted from a previous work by Shell and others[51].

Operon. Operons in *Mtb* were predicted by Roback et al[52]. We calculated the total number of genes of each operon as well as the position in the operon which was defined as the order of a gene in its operon. Operon length was defined as the sum of the lengths of all genes in the operon.

Regulon. Regulons of *Mtb* were identified by Yoo, R. et al.[53]. Regulons with less than three genes and annotated as "Unknown function", "KO", "Single gene" and "Uncharacterized" were removed in Fig 5b. To identify whether the TPs of the genes in a regulon were significantly higher or lower than the total TPs of the genes in genome, we performed Gene Set Enrichment Analysis (GSEA) with the R package clusterProfiler (version 3.16.1) to calculate normalized enrichment score (NES) and adjusted the *p* value for each regulon. NES represents the overall level of TP amplitude of a regulon, whereby higher positive NES values mean higher overall TP and lower negative NES values mean lower overall TP.

Other mycobacteria. Gene length and GC% of *Msm* and *Mab* were collected from mc^{2}155 and ATCC 19977 genome annotation files derived from Mycobrowser (https://mycobrowser.epfl.ch).

Machine learning model

To assess the importance of different gene features in determining the TP, we leveraged the recently advanced LightGBM, a decision-tree ensemble model, to perform a multiparametric regression analysis of the 3,891 genes and the corresponding 78 features[54]. This was achieved using the Python-compiled *lightgbm* package (version 3.3.2) with the following parameters: objective='regression', num_leaves=31, learning_rate=0.05, n_estimators=100, with the remaining parameters set to default. 3,891 genes were randomly divided into test and training sets in a ratio of 4:6 using “train_test_split” function from *sklearn*. Then, the LightGBM regression model was trained by training sets with the same parameters mentioned above. To evaluate the performance and robustness of the trained model, the genes were randomly split into test and training groups 100 times, and importance of each feature and performance (R^2) accuracy of the predicted TP with the TP in the test sets were calculated for each time, as shown in Fig. 3c and Fig. S3a, respectively.

LightGBM model predicted 4 robust features, which were operon size, gene length, activating regulator number and GC content, and we performed a support vector machines (SVM) model to assess the predictive power of these 4 features. This was archived using the R package '*e1071*' with the following parameters: types = 'eps-regression', kernel = 'radial', degree = 3, cost = 1, gamma = 0.25, coef0 = 0, epsilon = 0.1. Genes missing any feature value were removed so that a total of 2,016 genes were included in the analysis. Performance of this SVM model is shown in Fig. 3d. The Shapley additive explanations (SHAP) method was then applied to calculate the contribution of each feature to TP values predicted by SVM model[55]. We performed SHAP analysis using R package 'iBreakDown' (version 2.0.1), and the contribution value of each feature to the predicted TP of each gene was determined. As the contribution value can be positive or negative, representing the portion of the feature making the predicted TP value of a gene higher or lower than the average predicted TP value of all 2,016 genes, respectively, the absolute contribution value was taken (Fig. S3b).

To test whether there were co-variants among the 4 features (Fig. 3c) found to affect TP, pairwise Spearman's correlation coefficients were calculated using the R package *stats* (Fig. S3c).

Statistical analysis

Pearson’s correlation coefficients and the corresponding *p* values (Fig. 3d, Fig. 6a-b, Fig. S1c, Fig. S4f, Fig. S7c) were calculated using the R package *stats*; Spearman's correlation coefficients (Fig. 2f, Fig. 4a, Fig. 4c, Fig. 5d-e, Fig. 6f, Fig. S4a, Fig. S5, Fig. S6) were calculated using the R package *stats*. The non-parametric Wilcoxon test was used to make un-paired comparisons and to render the *p* values depicted in Fig. 2c-d, Fig. 4d-f, Fig. 5a, Fig. 6c, Fig. S4d-e.

Data availability

No primary data has been generated in this study. RNA-Seq data sources are listed in Supplementary Table 1. The conditions of 894 samples are annotated in Supplementary Table 1. The integrated transcriptional profile containing 3,891 genes and 894 samples is available in Supplementary Table 2. Collected genetic features are listed in Supplementary Table 3. TP data of *Msm* and *Mab* are available in Supplementary Table 5. Benchmark of DEGs based on TP data of *Mtb* are shown in Supplementary

Table 6.

Code availability

Code for data analysis in this study is available from the following GitHub repository, https://github.com/ChengBEI-FDU/Transcriptional_Plasticity

Acknowledgement

This work was supported by the National Institutes of Health (grant P01 AI132130 to S.M.F.). This study was also financially supported by the National Natural Science Foundation of China (82272376 to QG) and Shanghai Municipal Science and Technology Major Project (ZD2021CY001to QG).

Competing interests

The authors declare no competing interests.

Reference

1 Silander, O. K. *et al.* A genome-wide analysis of promoter-mediated phenotypic noise in Escherichia coli. *PLoS Genet* **8**, e1002443, doi:10.1371/journal.pgen.1002443 (2012).

2 Vlková, M. & Silander, O. K. Gene regulation in Escherichia coli is commonly selected for both high plasticity and low noise. *Nat Ecol Evol* **6**, 1165-1179, doi:10.1038/s41559-022-01783-2 (2022).

3 Kenkel, C. D. & Matz, M. V. Gene expression plasticity as a mechanism of coral adaptation to a variable environment. *Nat Ecol Evol* **1**, 14, doi:10.1038/s41559-016-0014 (2016).

4 Urchueguía, A. *et al.* Genome-wide gene expression noise in Escherichia coli is condition-dependent and determined by propagation of noise through the regulatory network. *PLoS Biol* **19**, e3001491, doi:10.1371/journal.pbio.3001491 (2021).

5 Lehner, B. Conflict between noise and plasticity in yeast. *PLoS Genet* **6**, e1001185, doi:10.1371/journal.pgen.1001185 (2010).

6 Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genet* **2**, e222, doi:10.1371/journal.pgen.0020222 (2006).

7 Lehner, B. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol* **4**, 170, doi:10.1038/msb.2008.11 (2008).

8 Xiao, L., Zhao, Z., He, F. & Du, Z. Multivariable regulation of gene expression plasticity in metazoans. *Open Biol* **9**, 190150, doi:10.1098/rsob.190150 (2019).

9 Bajić, D. & Poyatos, J. F. Balancing noise and plasticity in eukaryotic gene expression. *BMC Genomics* **13**, 343, doi:10.1186/1471-2164-13-343 (2012).

10 Latorre, P. *et al.* Data-driven identification of inherent features of eukaryotic stress-responsive genes. *NAR Genom Bioinform* **4**, lqac018, doi:10.1093/nargab/lqac018 (2022).

11 Charoensawan, V., Wilson, D. & Teichmann, S. A. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res* **38**, 7364-7377, doi:10.1093/nar/gkq617 (2010).

12 Ishihama, A. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. *FEMS Microbiol Rev* **34**, 628-645, doi:10.1111/j.1574-6976.2010.00227.x (2010).

13 Roberfroid, S., Vanderleyden, J. & Steenackers, H. Gene expression variability in clonal populations: Causes and consequences. *Crit Rev Microbiol* **42**, 969-984, doi:10.3109/1040841x.2015.1122571 (2016).

14 World Health, O. *Global tuberculosis report 2019*. xi, 283 p. (World Health Organization, 2019).

15 Galagan, J. E. *et al.* The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature* **499**, 178-183, doi:10.1038/nature12337 (2013).

16 Yuan, Y. *et al.* The 16-kDa alpha-crystallin (Acr) protein of Mycobacterium tuberculosis is required for growth in macrophages. *Proc Natl Acad Sci U S A* **95**, 9578-9583, doi:10.1073/pnas.95.16.9578 (1998).

17 Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* **50**, W216-221, doi:10.1093/nar/gkac194 (2022).

18 Cole, S. T. *et al.* Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537-544, doi:10.1038/31159 (1998).

19 Kapopoulou, A., Lew, J. M. & Cole, S. T. The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinb)* **91**, 8-13, doi:10.1016/j.tube.2010.09.006 (2011).

20 DeJesus, M. A. *et al.* Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. *mBio* **8**, doi:10.1128/mBio.02133-16 (2017).

21 Bosch, B. *et al.* Genome-wide gene expression tuning reveals diverse vulnerabilities of M. tuberculosis. *Cell* **184**, 4579-4592.e4524, doi:10.1016/j.cell.2021.06.033 (2021).

22 Wilson, D. J. & Consortium, C. R. GenomegaMap: Within-Species Genome-Wide dN/dS Estimation from over 10,000 Genomes. *Mol Biol Evol* **37**, 2450-2460, doi:10.1093/molbev/msaa069 (2020).

23 Ju, X. *et al.* Incomplete transcripts dominate the Mycobacterium tuberculosis transcriptome. *bioRxiv*, doi:10.1101/2023.03.10.532058 (2023).

24 Santangelo, M. P. *et al.* Mce3R, a TetR-type transcriptional repressor, controls the expression of a regulon involved in lipid metabolism in Mycobacterium tuberculosis. *Microbiology (Reading)* **155**, 2245-2255, doi:10.1099/mic.0.027086-0 (2009).

25 Crowe, A. M. *et al.* Structural and functional characterization of a ketosteroid transcriptional regulator of Mycobacterium tuberculosis. *J Biol Chem* **290**, 872-882, doi:10.1074/jbc.M114.607481 (2015).

26 Balhana, R. J. *et al.* bkaR is a TetR-type repressor that controls an operon associated with branched-chain keto-acid metabolism in Mycobacteria. *FEMS Microbiol Lett* **345**, 132-140, doi:10.1111/1574-6968.12196 (2013).

27 Lara, J. *et al.* Mycobacterium tuberculosis FasR senses long fatty acyl-CoA through a tunnel and a hydrophobic transmission spine. *Nat Commun* **11**, 3703, doi:10.1038/s41467-020-17504-x (2020).

28 Nyström, T. Conditional senescence in bacteria: death of the immortals. *Mol Microbiol* **48**, 17-23, doi:10.1046/j.1365-2958.2003.03385.x (2003).

29 Sigalova, O. M., Shaeiri, A., Forneris, M., Furlong, E. E. & Zaugg, J. B. Predictive features of gene expression variation reveal mechanistic link with differential expression. *Mol Syst Biol* **16**, e9539, doi:10.15252/msb.20209539 (2020).

30 Cortijo, S., Aydin, Z., Ahnert, S. & Locke, J. C. Widespread inter-individual gene expression variability in Arabidopsis thaliana. *Mol Syst Biol* **15**, e8591, doi:10.15252/msb.20188591 (2019).

31 Aceituno, F. F., Moseyko, N., Rhee, S. Y. & Gutiérrez, R. A. The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in

Arabidopsis thaliana. *BMC Genomics* **9**, 438, doi:10.1186/1471-2164-9-438 (2008).

32 Lopes, I., Altab, G., Raina, P. & de Magalhães, J. P. Gene Size Matters: An Analysis of Gene Length in the Human Genome. *Front Genet* **12**, 559998, doi:10.3389/fgene.2021.559998 (2021).

33 Rao, Y. S., Chai, X. W., Wang, Z. F., Nie, Q. H. & Zhang, X. Q. Impact of GC content on gene expression pattern in chicken. *Genet Sel Evol* **45**, 9, doi:10.1186/1297-9686-45-9 (2013).

34 Courel, M. *et al.* GC content shapes mRNA storage and decay in human cells. *Elife* **8**, doi:10.7554/eLife.49708 (2019).

35 Teng, W., Liao, B., Chen, M. & Shu, W. Genomic Legacies of Ancient Adaptation Illuminate GC-Content Evolution in Bacteria. *Microbiol Spectr* **11**, e0214522, doi:10.1128/spectrum.02145-22 (2023).

36 Ford, C. B. *et al.* Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature genetics* **43**, 482-486, doi:10.1038/ng.811 (2011).

37 Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**, 32-42, doi:10.1038/nrg2899 (2011).

38 Balakrishnan, R. *et al.* Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. *Science* **378**, eabk2066, doi:10.1126/science.abk2066 (2022).

39 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

40 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

41 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

42 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).

43 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25, doi:10.1186/gb-2010-11-3-r25 (2010).

44 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).

45 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

46 Eling, N., Richard, A. C., Richardson, S., Marioni, J. C. & Vallejos, C. A. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Syst* **7**, 284-294.e212, doi:10.1016/j.cels.2018.06.011 (2018).

47 Judd, J. A. *et al.* A Mycobacterial Systems Resource for the Research Community. *mBio* **12**, doi:10.1128/mBio.02401-20 (2021).

48 Cascioferro, A. *et al.* Xer site-specific recombination, an efficient tool to introduce unmarked deletions into mycobacteria. *Appl Environ Microbiol* **76**, 5312-5316, doi:10.1128/aem.00382-10 (2010).

49 Minch, K. J. *et al.* The DNA-binding network of Mycobacterium tuberculosis. *Nat Commun* **6**, 5829, doi:10.1038/ncomms6829 (2015).

50 Rustad, T. R. *et al.* Mapping and manipulating the Mycobacterium tuberculosis transcriptome using a transcription factor overexpression-derived regulatory network. *Genome Biol* **15**, 502,

doi:10.1186/preaccept-1701638048134699 (2014).

51 Shell, S. S. *et al.* Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet* **11**, e1005641, doi:10.1371/journal.pgen.1005641 (2015).

52 Roback, P. *et al.* A predicted operon map for Mycobacterium tuberculosis. *Nucleic Acids Res* **35**, 5085-5095, doi:10.1093/nar/gkm518 (2007).

53 Yoo, R. *et al.* Machine Learning of All Mycobacterium tuberculosis H37Rv RNA-seq Data Reveals a Structured Interplay between Metabolism, Stress Response, and Infection. *mSphere* **7**, e0003322, doi:10.1128/msphere.00033-22 (2022).

54 Ke, G. *et al.* in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 3149–3157 (Curran Associates Inc., Long Beach, California, USA, 2017).

55 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

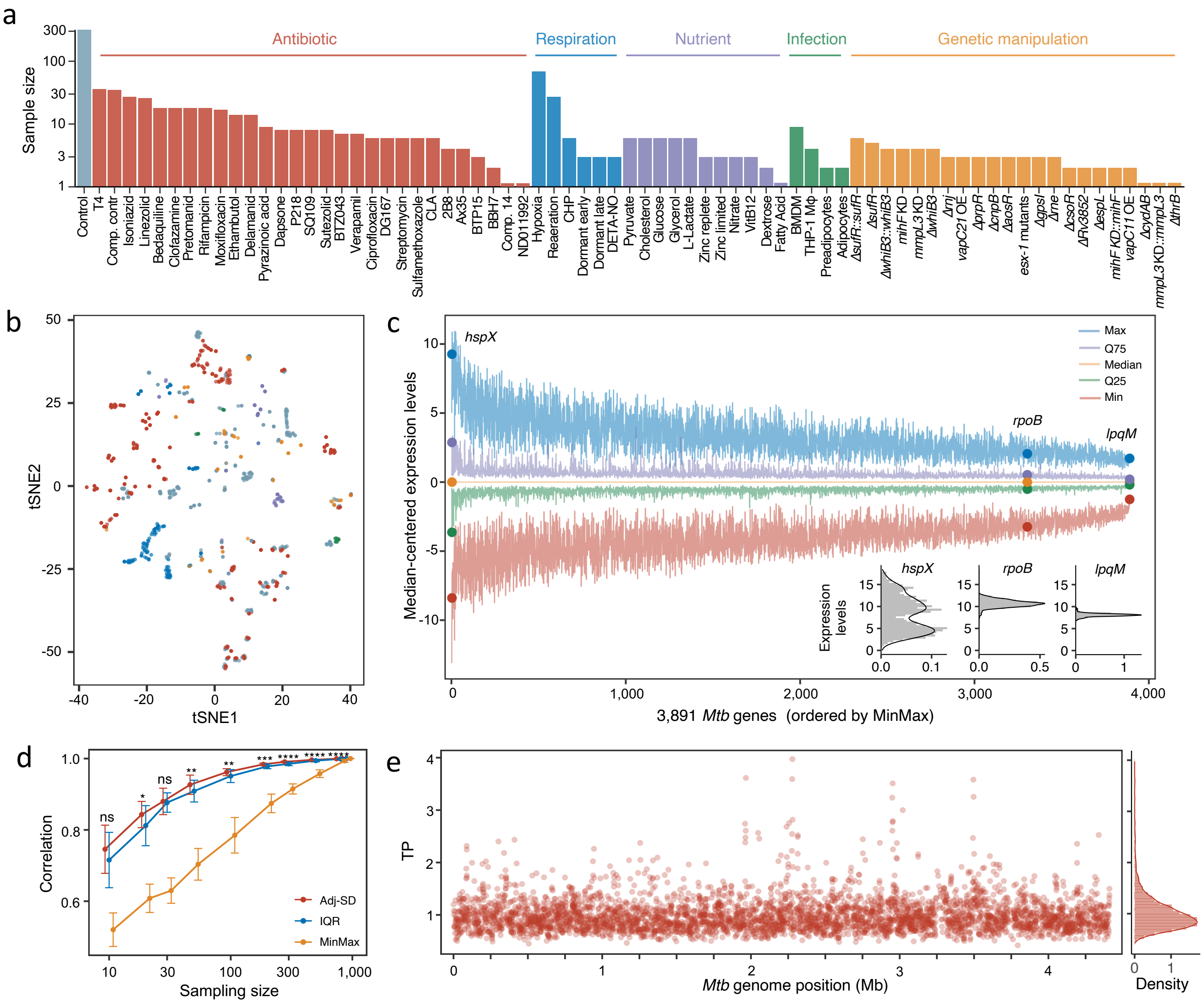

Fig. 1 Genome-wide estimation of *Mtb* transcriptional plasticity (TP). (a) A diagram illustrating the composition of the 894 samples from 73 different conditions. Detailed information about the samples can be found in Table S1. **(b)** Visualization of the 894 samples using t-distributed stochastic neighbor embedding (tSNE) grouped according to different experimental condition categories. **(c)** Primary expression statistics of *Mtb* genes across the 894 samples. Genes are horizontally ranked by the *MinMax* metric. The five line-plots represent the maximum (Max), 75 percentile (Q75), median, 25 percentile (Q25) and minimum (Min) expression levels which are centered by subtracting the median expression level of each gene. Expression statistics for three representative genes, *hspX*, *rpoB* and *lpqM*, are highlighted. **(d)** Comparing *adj-SD, IQR*, and *MinMax* metrics in describing TP of *Mtb* genes using a subsampling and bootstrap analysis (see *Materials and Methods*). Statistical significance between correlation coefficients of adj-SD and IQR was estimated by Wilcoxon tests. ns represents nonsignificant, * *p* value 0.01 ~ 0.05, ** *p* value 0.001 ~ 0.01, *** *p* value 0.0001 ~ 0.001, and **** *p* value < 0.0001. **(e)** Genome-wide TP profiles (*adj-SD*) of the 3,891 *Mtb* genes. The positively skewed genome-wide *TP* distribution is illustrated in the right panel.

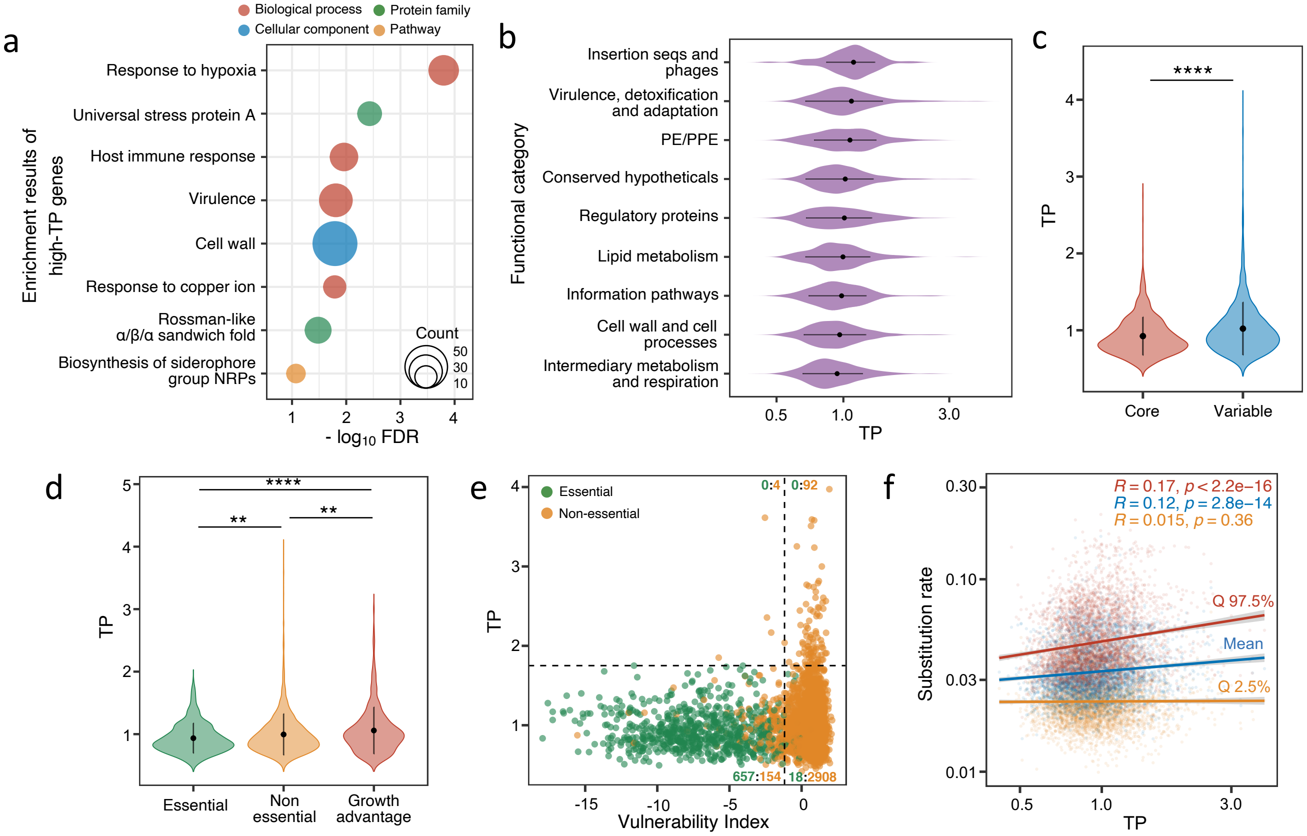

Fig. 2 ***TP*** **is associated with gene function and gene essentiality. (a)** Functional enrichment analysis of the 195 high-*TP* genes. Circle size corresponds to the number of genes in each category **(b)** Violin plots showing the TP profiles of genes in different functional categories. Error bars denote mean ± SD of TPs. The X-axis is presented on a log scale. **(c)** Genes of mycobacterial core-genome exhibit lower TPs than other genes of the variable genome. Error bars represent mean ± SD of TPs. Statistical significance was assessed by the Wilcoxon test, **** p value < 0.0001. **(d)** TP comparison between essential genes, non-essential genes and genes whose disruption confer growth advantage under axenic culture conditions. Statistical significance was assessed by the Wilcoxon test, error bars represent mean ± SD of TPs, ** p value 0.0001 ~ 0.01, **** p value < 0.0001. **(e)** *Mtb* Genes vulnerable to transcriptional perturbation exhibit low TPs. The horizontal black dashed line represents the maximum TP value of essential genes, and the vertical line shows the 5th vulnerability index of non-essential genes. The counts of essential and non-essential genes in each quadrant are displayed in green and yellow, respectively. **(f)** TP positively correlates with genes' substitution rate, as simulated by *genomegaMap* (Wilson, 2020). Mean value and 95% credibility intervals of substitution rates are presented in colored points. Colored Lines depict the linear fit between TP and substitution rate. *R* and *p* represent Spearman's correlation coefficient and the associated *p* values, respectively.

a

Category	N	Features
Sequence	50	Gene length; Strand and replication orientation; % of amino acid; % of base; Codon usage (CAI, CBI, Fop, Nc, Gravy, Aromo, start and stop codon)
Transcription	17	Operon (operon size and operonic location); Relative orientation between replication and transcription; Transcriptional regulation (activator/repressor number, and activated/repressed targets number of regulators); Transcription start sites (coverage, number, distance to downstream genes, and base type at +1 position)
Evolution	11	Ratio of non-synonymous to synonymous substitutions (dN/dS, ω); Ratio of transition to transversion (Ts/Tv, κ); Substitution rate (θ)

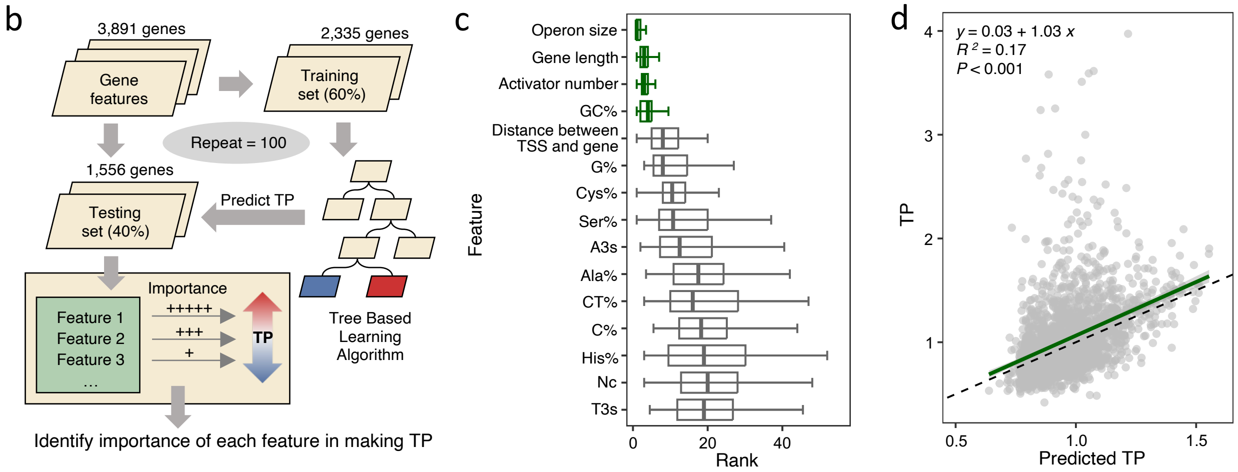

Fig. 3 Identification of genetic features underlying TP. (a) A table summary of the 78 candidate genetic features. N denotes the number of features in each category. **(b)** Schematic diagram illustrating our machine-learning workflow. **(c)** The top 15 genetic features ranked by their average feature importance in predicting TP. Lower ranks signify higher feature importance for TP prediction, whereas a tight rank distribution indicates higher consistency in predictions across randomized sample splits and modeling iterations. The four genetic features consistently rank low across random repeats are highlighted in green. Error bars represent the median ± 1.5*IQR of feature importance ranks across experiments. **(d)** An SVM model constructed using only the top 4 features effectively predicts TP. The green line represents the linear fit between SVM-modeled and observed TPs.

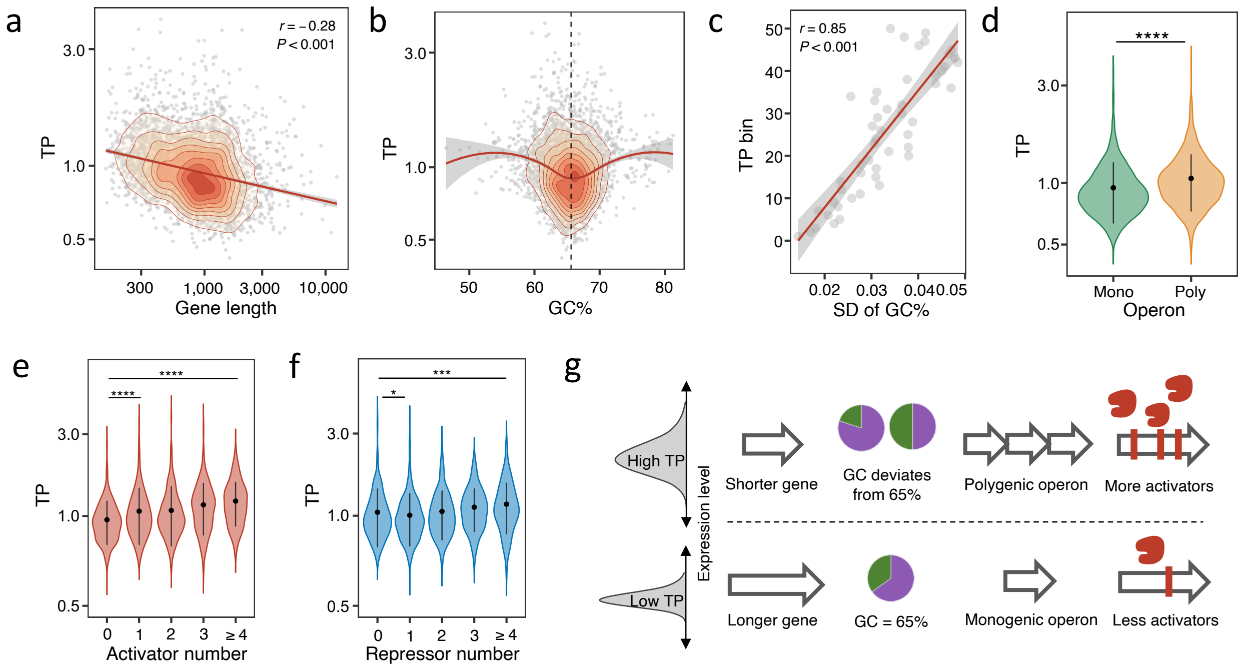

Fig. 4 Impact of key genetic features on TP. (a) A negative correlation exists between gene length and TP, illustrated by the 2D density contour plot of genes by TP and gene length. The red line depicts the linear fit. (**b**) Deviation in GC% from the genome-wide average GC% (65.6%, black dashed line) is positively linked with TP, depicted by the LOESS trendline and the 2D density contours. This trend is signified by the strong positive association between average TP and standard deviation (SD) of GC% of genes belonging to the 50 TP quantiles, as illustrated in **(c)**. **(d)** Genes in polygenic operons exhibit significantly higher TPs than those in monogenic operons. Wilcoxon tests, **** indicates *p* value < 0.0001. **(e-f)** TP increases as genes are regulated by more regulators. Boxplots demonstrate a monotonic relationship between TP and the number of activators. (e). Genes targeted by only one repressor display the lowest TPs. Error bars represent mean ± SD of TPs. Statistical significance was assessed by Wilcoxon tests, * *p* value 0.0001 ~ 0.05, **** *p* value < 0.0001. **(g)** A schematic illustrating the relationships between the four genetic features and TP.

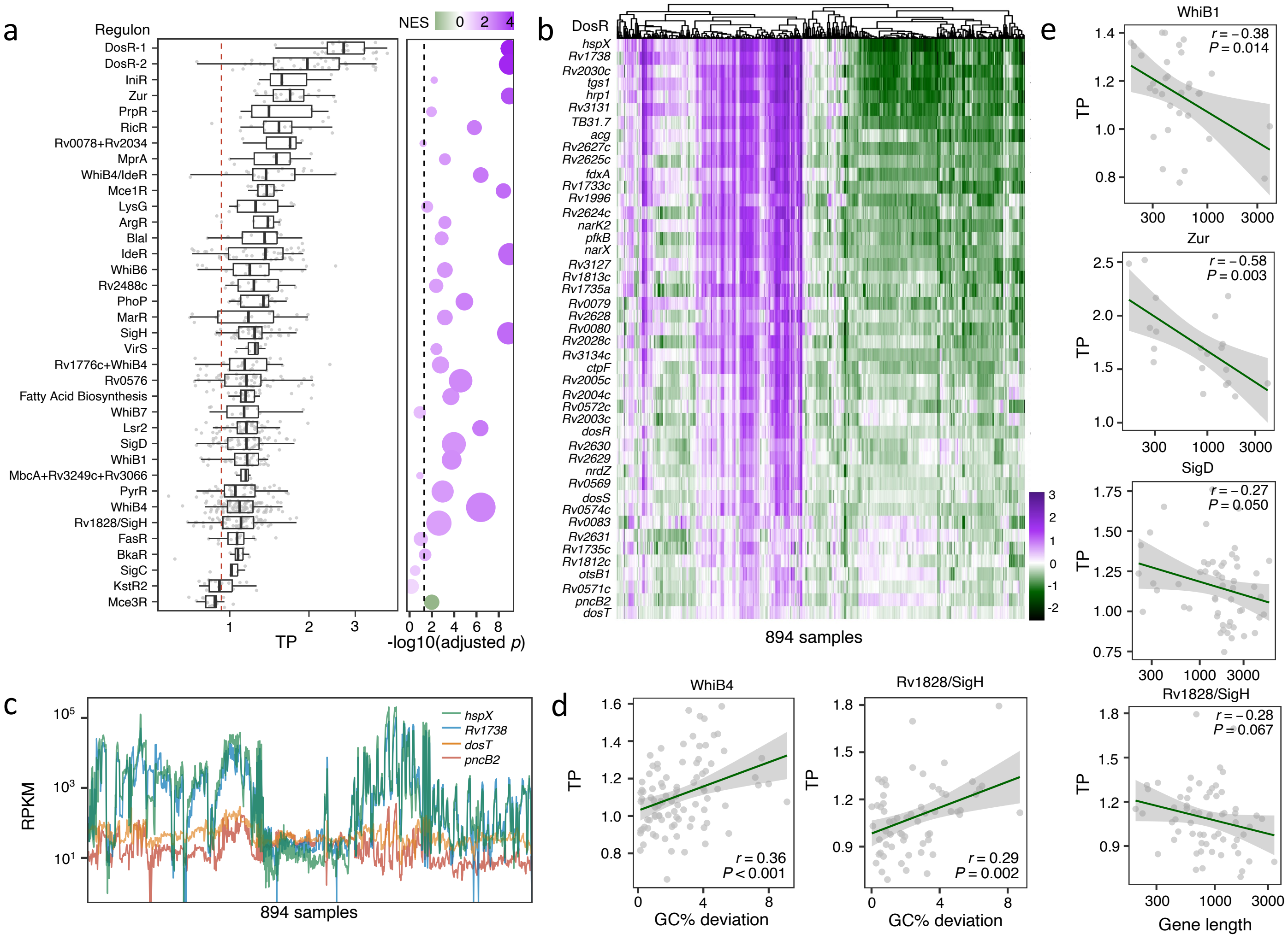

Fig. 5 The impact of primary sequence features on TP is partially independent of transcription regulation. (a) *Mtb* regulons display varying degrees of transcriptional plasticity. Error bars denote median ± 1.5*IQR of TPs, and the red dashed line represents the median TP of all 3,891 genes. The bubble plot to the right summarizes the statistical significance (adjusted *p*-value) and normalized enrichment score (NES) of each regulon by single-sample Gene Set Enrichment Analysis (ssGSEA). A higher NES indicates that the operon is enriched for genes with higher TPs. Bubble size corresponds to the number of genes in each regulon. (**b**) Expression profiles of DosR regulon genes ranked by TP. The color gradient represents the Z-score normalized log-RPKM. (**c**) Variations in TP within the DosR regulon, exemplified by comparing expression profiles of two high-TP genes (*hspX* and *Rv1738*) with two low-TP genes (*dosT* and *pncB2*). **(d)** Deviation in GC% from the genome average partially explains TP variations of genes of the same regulon. Linear fits and Spearman's correlation coefficients are shown for two representative regulons, WhiB4 and Rv1828/SigH. **(e)** TPs of co-regulated genes negatively correlate with their gene lengths. Spearman's correction coefficient and the corresponding *p* values are provided. The associations between primary genetic features and TP for genes in additional regulons are illustrated in Fig. S5.

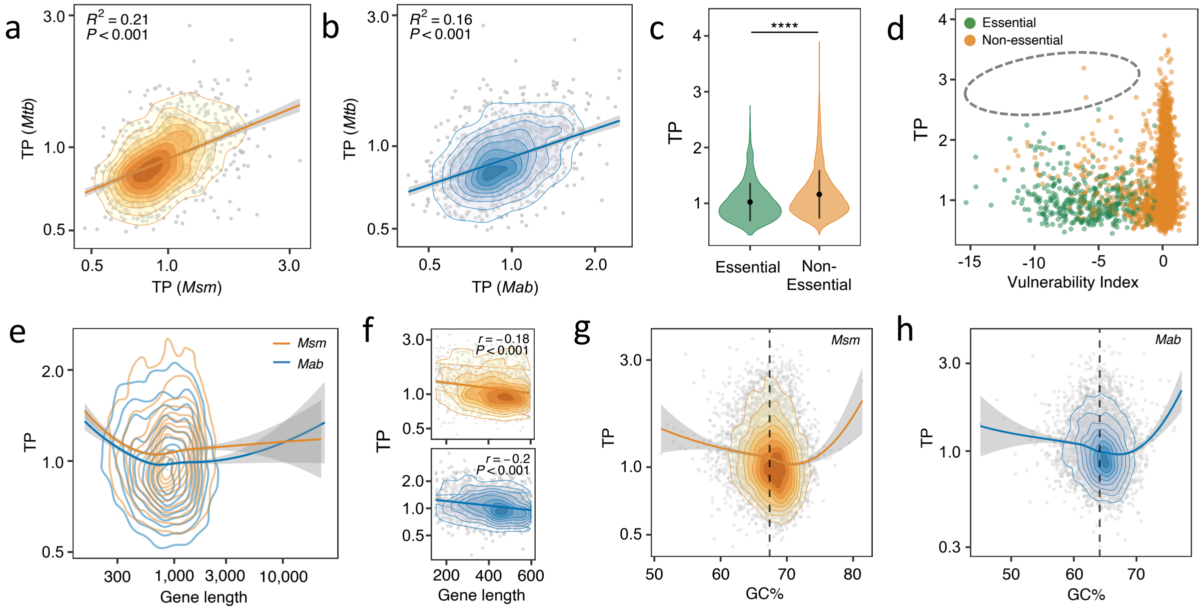

Fig. 6 TP and its underlying genetic determinants are conserved in other *Mycobacterium* species. (a-b). The TP profiles of *M. smegmatis* (*Msm*) and *M. abscessus* (*Mab*) genes resemble those of the *Mtb* homologs. The 2D density contour plots illustrate the distribution of gene orthologs according to their TPs in corresponding *Mycobacterium* species. Red lines denote the linear fits. © Non-essential *Msm* genes have higher TPs than their essential *Msm* counterparts. Error bars represent mean ± SD of TPs. Statistical significance was measured by Wilcoxon tests, **** p value < 0.0001. **(d)** *Msm* genes vulnerable to transcriptional perturbation exhibit low TPs. The grey circle highlights the lack of genes with both high TP and high vulnerabilit©**(e)** Gene length is negatively associated with TP in *Msm* (orange) and *Mab* (blue). The 2D density contour plots illustrate the distribution of genes based on TP and gene length. **(f).** A linear correlation is observed between TPs and gene lengths for genes shorter than 600 bp. **(g-h)** Genes with GC% close to the genome-wide average (67.4% in *Msm* and 64.1% in *Mab,* annotated by black dashed lines) display lower TP in both *Msm* (g) and *Mab* (h). The 2D density contour plots depict the distribution of genes by their TPs and GC%.