Published online 7 December 2012

Nucleic Acids Research, 2013, Vol. 41, No. 2 ***961–977***
doi:10.1093/nar/gks1260

High-resolution transcriptome and genome-wide dynamics of RNA polymerase and NusA in *Mycobacterium tuberculosis*

Swapna Uplekar[1,3], Jacques Rougemont[2,3], Stewart T. Cole[1,*] and Claudia Sala[1,*]

[1]Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, Station 19, CH-1015 Lausanne, Switzerland, [2]Bioinformatics and Biostatistics Core Facility, Ecole Polytechnique Fédérale de Lausanne, Station 15, CH-1015 Lausanne, Switzerland and [3]Swiss Institute of Bioinformatics, Bâtiment Génopode, Université de Lausanne, CH-1015 Lausanne, Switzerland

Received October 2, 2012; Revised October 31, 2012; Accepted November 2, 2012

ABSTRACT

To construct a regulatory map of the genome of the human pathogen, *Mycobacterium tuberculosis*, we applied two complementary high-resolution approaches: strand-specific RNA-seq, to survey the global transcriptome, and ChIP-seq, to monitor the genome-wide dynamics of RNA polymerase (RNAP) and the anti-terminator NusA. Although NusA does not bind directly to DNA, but rather to RNAP and/or to the nascent transcript, we demonstrate that NusA interacts with RNAP ubiquitously throughout the chromosome, and that its profile mirrors RNAP distribution in both the exponential and stationary phases of growth. Generally, promoter-proximal peaks for RNAP and NusA were observed, followed by a decrease in signal strength reflecting transcriptional polarity. Differential binding of RNAP and NusA in the two growth conditions correlated with transcriptional activity as reflected by RNA abundance. Indeed, a significant association between expression levels and the presence of NusA throughout the gene body was detected, confirming the peculiar transcription-promoting role of NusA. Integration of the data sets pinpointed transcriptional units, mapped promoters and uncovered new anti-sense and non-coding transcripts. Highly expressed transcriptional units were situated mainly on the leading strand, despite the relatively unbiased distribution of genes throughout the genome, thus helping the replicative and transcriptional complexes to align.

INTRODUCTION

To adapt and survive in a range of different environments, *Mycobacterium tuberculosis*, the etiologic agent of human tuberculosis (TB), has to fine-tune gene expression to support active growth, to survive periods of non-replicating persistence and to cope with the plethora of stresses encountered within the host (1–3).

In prokaryotes, regulation of gene expression mainly takes place at the transcriptional level, and this is particularly evident in *M. tuberculosis* that has 13 sigma factors, which direct RNA polymerase (RNAP) core enzyme to defined subsets of promoters, and nearly 200 potential transcriptional regulators (4). One particular class of regulators is represented by proteins affecting pausing, termination and anti-termination of transcription, the best characterized of them being NusA.

NusA is an essential RNAP- and RNA-binding modulator of gene expression, present in all bacteria whose genomes have been sequenced so far, and was named N-utilizing substance after the phage λ N-protein mediated anti-termination process (5). Its role in preventing termination at ribosomal RNA operons was first demonstrated in the paradigm organism *Escherichia coli*, where the ability of NusA to bind to nut-like sites (Box A, B and C) and to protect the nascent naked transcript from Rho-dependent termination has been proved (6). Likewise, more recent studies in *M. tuberculosis* confirmed that this mechanism is conserved among different bacterial species (7). Conversely, NusA has also been implicated in pausing and termination at intrinsic or Rho-independent terminator sequences (8,9). Cardinale *et al.* (10) later demonstrated that the protein plays a pivotal role in Rho-dependent silencing of foreign DNA in *E. coli*. This multi-functionality is reflected in the crystal structures of the *Thermotoga maritima* and *M. tuberculosis*

*To whom correspondence should be addressed. Tel: +41 21 693 0649; Fax: +41 21 693 1790; Email: claudia.sala@epfl.ch
Correspondence may also be addressed to Stewart T. Cole. Tel: +41 21 693 1851; Fax: +41 21 693 1790; Email: stewart.cole@epfl.ch

NusA proteins, which have a multidomain organization. The protein is in fact composed of an N-terminal RNAP-binding domain, connected through a flexible linker to the S1, KH1 and KH2 RNA-binding motifs (11,12).

In a comprehensive study aimed at understanding the dynamics of the transcriptional complex in *E. coli*, Mooney *et al.* (13) demonstrated the genome-wide association of RNAP with NusA, as RNA synthesis occurs and the enzyme moves away from the promoter. In addition, RNAP promoter-proximal peaks coincide with the distribution of NusA, therefore indicating that RNAP peaks reflect elongating rather than stalled transcriptional complexes. Recent work performed in *Bacillus subtilis* showed that, contrary to *E. coli* RNAP, the enzyme is distributed evenly from the promoter throughout the coding sequence (CDS) in this species and does not generate promoter-proximal signals (14).

Whole-genome research conducted thus far in *M. tuberculosis* has focused on the final product of transcriptional regulation, the RNA, and has analyzed the differential expression of genes in specific growth conditions by microarrays (15–17). More recently, deep sequencing was applied to transcriptomic profiling and revealed the existence of previously unknown small RNAs whose expression changes during the transition from exponential (Exp) to stationary (Stat) phase, and may undergo deregulation during infection (18). However, little is known of the molecular mechanisms taking place at the DNA level, and there is no genome-wide description of the RNAP and NusA interaction with the chromosome in different growth phases. In this work, we used two complementary, single-nucleotide resolution approaches, namely ChIP-seq and RNA-seq, to investigate the assembly, distribution and activity of the transcriptional complex throughout the *M. tuberculosis* genome.

MATERIALS AND METHODS

Bacterial strains and culture conditions

M. tuberculosis H37Rv was grown at 37°C in 7H9 broth (Difco) supplemented with 0.2% glycerol, 0.05% Tween 80 and 10% albumin-dextrose-catalase (ADC, Middlebrook) or on 7H10 plates supplemented with 0.5% glycerol and 10% oleic acid-albumin-dextrose-catalase (OADC, Middlebrook). Logarithmic-phase cultures (Exp phase) were harvested at an optical density (OD_{600}) of 0.4–0.6; Stat phase cultures were collected 4 weeks later.

Chemicals, antibodies and oligonucleotides used in this study

All chemicals were purchased from Sigma-Aldrich, unless otherwise stated. Monoclonal antibodies to the beta subunit (RpoB) of *E. coli* RNAP were purchased from Neoclone (clone 8RB13). Polyclonal anti-NusA antibodies were raised in rats at the Statens Serum Institut, Copenhagen, Denmark, according to the previously described procedure (19) using purified NusA, kindly provided by Dr. Kristine Arnvig and Dr. Ian Taylor. Sequences of the oligonucleotides used in this study will be provided on request.

Chromatin immunoprecipitation experiments

Chromatin immunoprecipitation experiments were performed as previously described (19) with the following modifications. Briefly, *M. tuberculosis* cultures, either in Exp or Stat phase, were cross-linked with 1% formaldehyde for 10 min at 37°C. Cross-linking was quenched by addition of glycine (125 mM). Cells were then washed twice with Tris-buffered saline (TBS, 20 mM Tris–HCl pH 7.5, 150 mM NaCl), resuspended in 4 ml immunoprecipitation (IP) buffer (50 mM Hepes-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, protease inhibitor cocktail from Roche) and sonicated to shear DNA using Bioruptor (Diagenode). Cell debris was removed by centrifugation and the supernatant used in IP experiments. Nucleo-protein extracts were incubated with 80 µl of either anti-RpoB or anti-NusA antibodies at 4°C over-night on a rotating wheel. Complexes were subsequently precipitated with Dynabeads (Dynal, anti-mouse or anti-rat, respectively) for 3 h at 4°C. Beads were washed twice with IP buffer, once with IP buffer plus 500 mM NaCl, once with buffer III (10 mM Tris–HCl pH 8, 250 mM LiCl, 1 mM EDTA, 0.5% Nonidet-P40, 0.5% sodium deoxycholate), once with Tris-EDTA buffer pH 7.5. Elution was performed in 50 mM Tris–HCl pH 7.5, 10 mM EDTA, 1% SDS for 40 min at 65°C. Samples were finally treated with RNAse A for 1 h at 37°C, and cross-links were reversed by incubation for 2 h at 50°C and for 8 h at 65°C in 0.5× elution buffer with 50 µg Proteinase K (Eurogentec). DNA was purified by phenol-chloroform extraction and quantified by Nanodrop and Qubit fluorometer, according to the manufacturer's recommendations (Invitrogen).

Library preparation for ChIP-seq analysis and Illumina high-throughput sequencing

DNA fragments obtained from the IP procedure and the Input controls were used for library construction and sequencing with the ChIP-Seq Sample Preparation Kit (Illumina), according to the protocol provided by the manufacturer. One lane per library was sequenced on the Illumina Genome Analyzer IIx using the SR Cluster Generation Kit v4 and SBS 36 Cycle Kit (for RNAP ChIP-seq and input samples) or TruSeq SR Cluster Generation Kit v3 and TruSeq SBS Kit v3 (for NusA ChIP-seq). Data were processed with the Illumina Pipeline Software v1.60 (for RNAP ChIP-seq), v1.70 (for input samples), v1.80 (for NusA ChIP-seq).

Genome annotation

All analyses in this study were carried out using the *M. tuberculosis* H37Rv annotation from the TubercuList database (http://tuberculist.epfl.ch/). There are 4019 protein CDS currently annotated in the genome, 73 genes encoding for stable RNAs, small RNAs and tRNAs.To quantify protein occupancy and transcription across the entire genome, 3080 intergenic regions (IGs)

(regions flanked by two non-overlapping CDS) were included, resulting in a total of 7172 features.

ChIP-seq data analysis

The single-ended sequence reads generated from ChIP-seq experiments were aligned to the *M. tuberculosis* H37Rv genome (NCBI accession NC_000962.2) using Bowtie (20) allowing up to 3 mismatches and up to 10 hits per read. Since the samples were sequenced using different protocols resulting in varied read lengths (38–50 nt) all the raw datasets were trimmed to 38 bases to enable unbiased comparison of experiments. Bowtie results were converted into SAM/BAM format using samtools (21). A custom perl script was then used to obtain the per-base coverage normalized to the total number of mapped reads for each dataset. The script also shifted (by 80 bp) and merged the read counts (RCs) for forward and reverse strands to generate wig files containing single ChIP-seq profiles that were visualized on the University of California Santa Cruz (UCSC) Genome Browser *Mycobacterium tuberculosis* H37Rv 06/20/1998 Assembly (22). To compute the RC for each feature, the number of reads mapping to all positions in the feature were summed up and normalized to feature length. The final RC for each feature was determined as the mean of the RCs of both replicates. To determine the level of ChIP-seq enrichment for each feature, an enrichment ratio (ER) was calculated by dividing the RC for the ChIP-seq sample by the RC for the Input (control) sample.

RNA extraction

Forty milliliters of either Exp or Stat phase *M. tuberculosis* cultures were pelleted and cells were flash frozen in liquid nitrogen and stored at −80°C until use. Bacteria were re-suspended in 1 ml Trizol (Invitrogen) and added to a 2-ml screw-cap tube containing 0.5 ml zirconia beads (BioSpec Products). Cells were disrupted by bead-beating twice for 1 minute with a 2-minute interval on ice. The cell suspension was then transferred to a new tube, where chloroform-isoamylalcohol (24:1) extraction was performed. RNA was precipitated by adding 1/10 volume of sodium acetate (2 M, pH 5.2) and 0.7 volume of isopropanol, washed with 70% ethanol, air-dried and resuspended in DEPC-treated water. DNase treatment was carried out twice using RQ1 RNase-free DNase (Promega), following the manufacturer's recommendations, and the reactions were subsequently cleaned up by phenol-chloroform extraction and ethanol precipitation. RNA was stored at −80°C in DEPC-treated water. Amount and purity of RNA were determined spectrophotometrically, integrity of RNA was assessed on 1% agarose gel.

Library preparation for RNA-seq analysis and Illumina high-throughput sequencing

In all, 100 ng of total RNA from Exp and Stat phase were mixed with 5× Fragmentation buffer (Applied Biosystems), incubated for 4 min at 70°C and then transferred immediately on ice. RNA was purified using RNAClean XP beads (Beckman Coulter), according to the manufacturer's recommendations, and subsequently treated with Antarctic phosphatase (New England Biolabs). RNA was then re-phosphorylated at the 5′-end with polynucleotide kinase (New England Biolabs) and purified with Qiagen RNeasy MinElute columns. To ensure strand-specificity, v1.5 sRNA adapters (Illumina) were ligated at the 5′- and 3′-ends using RNA ligase. Reverse transcription was carried out using SuperScript III Reverse Transcriptase (Invitrogen) and SRA RT primer (Illumina). Twelve cycles of Polymerase Chain Reaction (PCR) amplification using Phusion DNA polymerase were then performed, and the library was finally purified with AMPure beads (Beckman Coulter) as per the manufacturer's instructions. A small aliquot (2.5 µl) was analyzed on Invitrogen Qubit and Agilent Bioanalyzer before sequencing on Illumina Genome Analyzer IIx using the TruSeq SR Cluster Generation Kit v3 and TruSeq SBS Kit v3. Data were processed with the Illumina Pipeline Software v1.82.

RNA-seq data analysis

The single-ended sequence reads generated from RNA-seq experiments were aligned to the *M. tuberculosis* H37Rv genome (NCBI accession NC_000962.2) using Bowtie (20), allowing one mismatch and no more than five hits per read. As the samples were sequenced using different protocols resulting in varied read lengths (40–50 nt) the raw data sets were trimmed to 38 bases to enable unbiased comparison of the experiments. Bowtie results were converted into SAM/BAM format using samtools (21). Preliminary analysis revealed a striking abundance of reads mapping to the ribosomal RNA operon (>95%) in all samples. Consequently, the data sets were normalized to the number of remainder reads, i.e. subtracting the number of ribosomal RNA operon reads from total number of mapped reads for each data set. The gene expression values were quantified in terms of reads per million (RPM), which can be defined as the total number of reads mapping to the feature divided by feature length (in bp) normalized to the number of remaining reads (in millions). The normalization factor was also used to generate wig files that were visualized on the UCSC Genome Browser (22). The mean RPM value was determined for the experimental replicates. The Magnitude-Amplitude (MA) plot was generated using a python script where the average expression in the Exp and Stat phase samples was plotted against the $\log_2$ fold-change between the two conditions. Analysis of differential expression was carried out using the DESeq package (23).

Transcriptional unit quantification

Enrichment of RNAP and NusA in transcriptional units (TUs) was quantified separately for the promoter and the body of the TU. The first feature was considered as the promoter, whereas the remaining features represented the body. The TU body RC and ER were computed as the mean of the RCs and ERs for all features in the TU. The level of transcription of the TU was calculated as the average RPM of all features (including promoter) in the TU.

Reverse transcription

Two micrograms of *M. tuberculosis* RNA were incubated with 50 ng random primers and 1 mM dNTPs at 65°C for 5 min. After cooling on ice, 40 U RNase inhibitor, 10 mM DTT, 1× reaction buffer, 5 mM $MgCl_2$ and 200 U SuperScript III reverse transcriptase (Invitrogen) were added in a final volume of 20 µl. A control reaction without reverse transcriptase was included as a control. Reactions were incubated at room temperature for 10 min, at 50°C for 1 h and at 55°C for 1 h. Reverse transcriptase was inactivated by incubation at 80°C for 2 min. RNase H treatment was carried out for 20 min at 37°C with 1 µl RNase H (Invitrogen). cDNA was stored at −20°C.

Quantitative PCR for ChIP-seq and RNA-seq data validation

All PCR primers were designed using Primer3 software (http://frodo.wi.mit.edu/primer3/). The 20 µl PCR reaction consisted of 1× Sybr Green PCR Master Mix (Applied Biosystems), 0.1 µM of each primer and 1 µl of cDNA or IP DNA from IP reactions. Reactions were carried out in duplicate in an Applied Biosystems 7900HT Sequence Detection System with the following protocol: denaturation at 95°C for 10 min, 40 cycles of denaturation at 95°C for 15 sec, annealing and extension at 60°C for 40 sec with data collection. Melting curves were constructed to ensure that only one amplification product was obtained.

Parallel reactions using different amounts of H37Rv chromosomal DNA were performed for each primer set to obtain the standard curve correlating the threshold cycle with the number of template molecules. The resulting equation was used to quantify the number of target molecules in the unknown samples.

In the case of quantitative PCR (qPCR) for RNA-seq data confirmation, normalization was obtained to the total amount of RNA used in the reverse transcription reaction, as previously described (24). Results were expressed as the log2 ratio of the number of molecules determined in the Exp phase versus the number of molecules in the Stat phase and correlated with the log2 ratio obtained with RPM values.

Regarding the qPCR for ChIP-seq data validation, the number of target molecules was normalized to the Input sample, after subtraction of the background represented by the mock-IP (no antibody control). Results were expressed as the log2 ER of the IP DNA versus Input and correlated with the log2 ER calculated as described earlier.

Statistical analysis

Statistical analyses were performed with the statistical language R and various Bioconductor packages (25) (http://www.bioconductor.org). Several plots were created using GraphPad Prism 5 software (www.graphpad.com). Custom Perl and Python scripts were developed by members of the EPFL Bioinformatics and Biostatistics core Facility (https://github.com/bbcf/bbcfutils). Correlations between replicates were based on Pearson's product moment correlation coefficient. The Wilcoxon signed-rank test (Mann–Whitney U test) was used to assess the differences between Exp and Stat phase values for the same set of features.

Data access

The ChIP-seq and RNA-seq data sets have been deposited in NCBI's Gene Expression Omnibus (26) under accession number GSE40862.

RESULTS

General strategy

The genome-wide dynamics of the transcriptional complex in *M. tuberculosis* was investigated by carrying out chromatin IP experiments followed by deep sequencing (ChIP-seq) using antibodies specific for a core component of RNAP, the beta subunit RpoB and for the transcriptional regulator NusA. To correlate the occupancy of the transcriptional complex with its activity, the global transcriptome was determined by using a strand-specific RNA-seq approach. ChIP-seq and RNA-seq data were obtained from Exp and Stat phase cultures. All data sets were mapped to the H37Rv genome sequence, and further analyses were based on the annotation from the TubercuList database (http://tuberculist.epfl.ch), which comprises 4019 protein CDS, 73 genes encoding ribosomal RNAs, small RNAs (sRNAs), tRNAs and other stable RNAs. In addition, the annotation used in this work includes 3080 IG, defined as regions flanked by two non-overlapping features, of which 2283 are ≥30 bases in length.

The genome-wide distribution of RNAP and NusA in *M. tuberculosis*

ChIP-seq experiments for RNAP and NusA, and sequencing of the Input DNA, were performed in duplicate, and the sequencing statistics are summarized in Supplementary Table S1. Supplementary Table S2 shows the correlation coefficients for the biological replicates, confirming the extremely high reproducibility of the Input and NusA data sets and, to a lesser extent, of the RNAP results.

Inspection of RNAP data, displayed on the UCSC genome browser, revealed that signals were present along the entire genome, although prominent accumulation of the enzyme at the putative promoter regions was detected in both of the conditions tested (promoter-proximal peaks). This is illustrated in Figure 1, where a global view of a 1 MB portion of the genome is shown (Figure 1A) together with three representative TUs (Figure 1B–D). The NusA-binding profile appeared to qualitatively match the RNAP distribution. Indeed, on closer inspection of the RNAP and NusA tracks, extensive overlap was observed in both growth phases, as evidenced from the correlation reported in Supplementary Figure S1, thus proving that, although NusA does not bind to DNA directly, it is part of the transcriptional complex and associates with RNAP.

Head-to-head comparison of the Exp and Stat data sets was possible by calculation of the ERs relative to the Input sample (ER) for RNAP and NusA for each

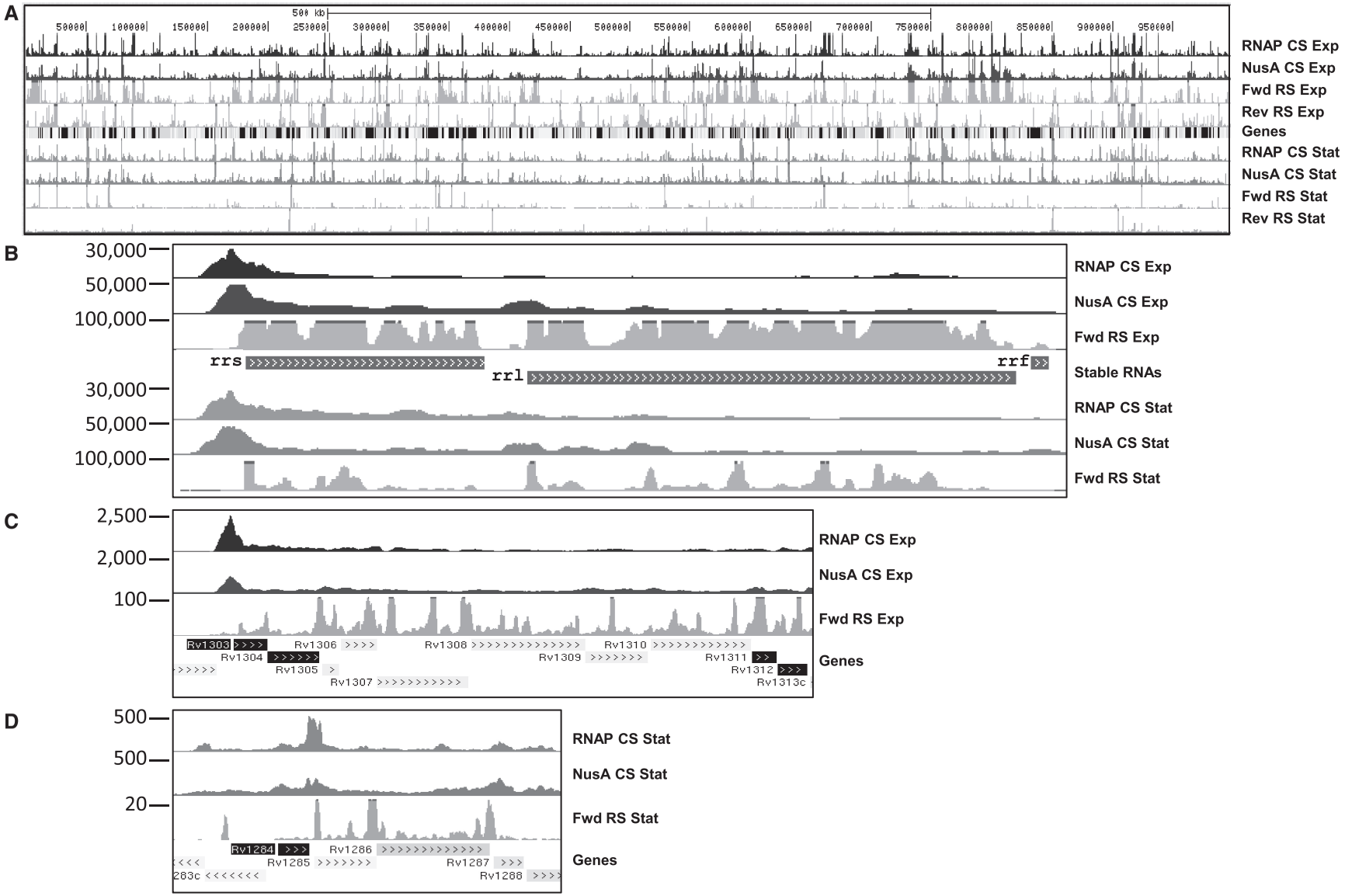


Figure 1. ChIP-seq profiles of RNAP and NusA and transcriptional levels measured by RNA-seq. (**A**) UCSC genome browser view of ChIP-seq (CS) profiles for RNAP and NusA and RNA-seq (RS) profiles on forward (Fwd) and reverse (Rev) strands across a 1 MB region of the *M. tuberculosis* H37Rv genome in Exp and Stat phases. The scale and the genome coordinates are reported at the top. (**B**) Profile of the ribosomal RNA operon in Exp and Stat phases. (**C**) Profile of the ATP synthase operon in Exp phase. (**D**) Profile of the *rv1285* to *rv1286* region in Stat phase. The scale for the number of reads is on the *y*-axis.

genomic feature. Results are reported in Supplementary Table S4. Approximately the same number of features was found to be enriched (ER $\geq$ 2) in Exp and Stat (1117 and 1062, respectively) and, in both phases, most of the signals were identified in IG regions (>60% versus <40% inside CDSs and RNA-encoding genes), where promoter sequences are most likely present. The strongest peaks for NusA in Exp phase were detected at the ribosomal RNA operon locus (Figure 1B), the ribosomal protein operon *rpsM-rpmJ*, in the IG region between *rv1534* and *rv1535* and preceding genes encoding sRNA and tRNAs. On the other hand, RNAP was mainly enriched in the IG region carrying the putative promoter of *rrn*, where the sRNA *mcr3* maps, at the *B11* locus and between *rv1733c* and *rv1734c*, where an anti-sense sRNA was identified as described later. Curiously, RNAP, but not NusA, was detected at some genes belonging to the DosR regulon. In both growth conditions, most of the regions showed enrichment for either RNAP and NusA together or RNAP only, whereas NusA alone was detected for a small fraction of signals (193 and 107 features in Exp and Stat, respectively). Independent support for the ChIP-seq findings was obtained by qPCR for a subset of CDSs and IGs, selected to cover a broad range of ER values (Supplementary Figure S2).

Detailed examination of the whole-genome profiles revealed unexpectedly high densities for RNAP and NusA at the end of convergent genes or inside CDSs, suggesting the existence of new features. For instance, these are the cases of the peak between *rv3661* and *rv3662c* or the peak inside *ino1* where, indeed, new sRNAs have been recently identified (18). In addition, the genomic regions encoding the putative excisionases of prophages phiRv1 and phiRv2 exhibited exceptionally high ER values for both RNAP and NusA, mainly in the Exp growth phase (21 and 7 for RNAP and NusA, respectively, for *rv1584c*; 17 and 6.7 for RNAP and NusA, respectively, for *rv2657c*). Deeper understanding of these peaks and of the entire ChIP-seq data set was possible on transcriptomic profiling as described later.

Deep RNA-seq analysis

To correlate the presence of RNAP and NusA, inferred from ChIP-seq, with transcriptional activity, global transcriptomic analysis was performed by deep sequencing, RNA-seq. More than 98% of the sequence tags could be mapped to the annotated CDSs in the sense orientation, whilst <2% was attributed to the anti-sense orientation, thereby confirming the strand-specific nature of the protocol. The correlation coefficient for biological replicates revealed high reproducibility ($r^2 > 0.95$), and further mapping statistics are reported in Supplementary Table S3. As expected, most of the reads aligned to *rrn* in both Exp and Stat phases, thus accounting for 82 and 95% of the total RPM in the two conditions, respectively. Overall, there was a clear predominance of stable RNAs such as sRNAs (1 and 0.6% in Exp and Stat), tRNAs (6.8 and 1.7%), the RNA component of RNAse P and the tmRNA *ssr* (together representing 1.9 and 1.7% of the total RPM in Exp and Stat). The remaining reads could be assigned to CDSs on both strands in almost equal proportion, and these provided sufficient coverage for quantification purposes. RPM values represent the unit chosen to allow comparison of the expression levels, and these are presented in Supplementary Table S5 for all of the features in the sense and anti-sense orientation, in the two growth conditions.

Codon usage and tRNAs

Measuring tRNA expression levels was informative in terms of understanding genome biology and generating confidence in the quantitation procedure. We counted the occurrence of each codon in every CDS and then multiplied this count by the value of gene expression obtained by RNA-seq (RPM) before summing these expression values across the whole genome to yield a per-codon 'expression level'. This codon usage was then compared with the expression of the corresponding tRNA. Figure 2A shows a clear correlation between these two metrics (Spearman correlation score 0.38, $P < 0.03$) in the Exp condition (with highly similar results in the Stat phase). Notably, AT-rich codons generally tend to be under used as opposed to their GC-rich alternatives with the cognate tRNAs being considerably more abundant.

Comparison of the Exp versus Stat phase transcriptomes

Grouping CDSs into functional categories according to TubercuList facilitated comparison of the total transcriptome in Exp and Stat phases. The box-and-whisker plots in Supplementary Figure S3 show that all categories are represented in both conditions, although the RPM dynamic range is higher in Exp phase, in particular for the lipid metabolism, information pathways and PE-PPE groups. On the contrary, features belonging to the virulence, detoxification and unknown subgroups are more enriched in Stat phase. Figure 2B illustrates the relative abundance of each category in terms of the proportion of genes with expression level (RPM) $\geq$1, the value chosen as cutoff. There is enrichment of genes involved in information pathways ($P < 10^{-4}$ in Exp and Stat, Fisher's exact test) and an abundance of stable RNAs ($P < 10^{-4}$ in Exp and $P < 10^{-21}$ Stat, Fisher's exact test) in both Exp and Stat phases. There is also a clear underrepresentation of genes belonging to intermediary metabolism and respiration ($P < 10^{-7}$) in the Stat phase.

Closer inspection of the differentially expressed genes revealed a Stat profile similar to the nutrient starvation condition previously described by Betts *et al.* (15), with down-regulation of the ATP synthase gene cluster, the *nad* genes, the housekeeping sigma factor *sigA*, the ribosomal protein operons, cell wall and cell membrane functions. By contrast, specific features induced in the Betts model were also found to be upregulated in Stat phase, such as *lat*, *rv0188*, *usfY*, *hsp* and the *cys* operon, and a steep increase in the expression level of the sigma factors *sigB* and *sigE* was observed (Figure 2C). Genes *rv1954c* and *rv2660c* had been reported as induced in starvation conditions (15), and our data confirmed this, but in the

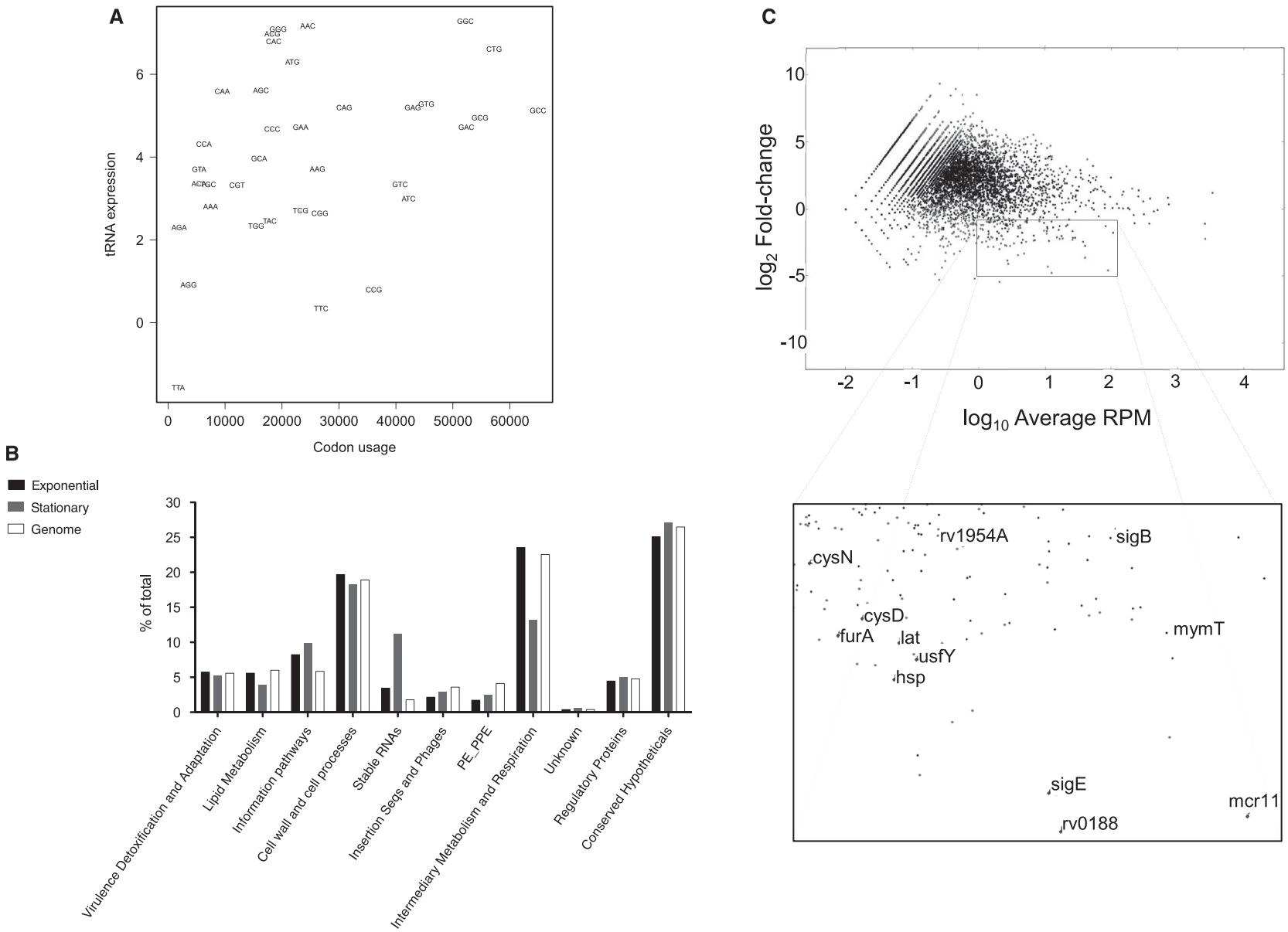


Figure 2. RNA-seq results. (**A**) Correlation between tRNA abundance and codon usage. (**B**) Distribution of genes transcribed in Exp and Stat phases across functional categories as defined in the TubercuList database (http://tuberculist.epfl.ch/) relative to the proportion observed in the *M. tuberculosis* H37Rv genome. (**C**) Magnitude-Amplitude-plot in which the mean ($\log_{10}$) expression value for each gene in Exp and Stat phase samples is plotted against the expression ratio ($\log_2$) between the two samples. Differentially expressed genes are colored in gray. The inset box indicates some of the genes induced in Stat phase. *rv1954A* indicates the anti-sense transcribed feature at the *rv1954* locus.

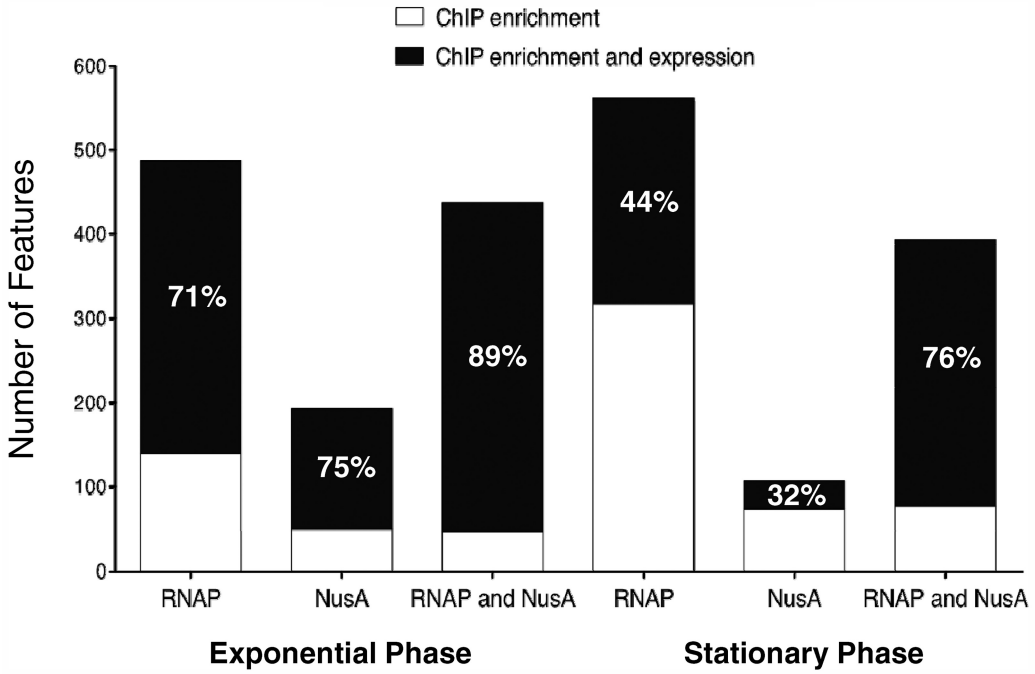


Figure 3. Overlap between enriched (ER $\geq$ 1) and transcribed (RPM $\geq$ 1) features in Exp and Stat phases. Stacked bar charts show the total number of features enriched with RNAP, NusA and both (RNAP and NusA) in the two phases. Black stacks represent the proportion of enriched features that are also transcribed.

anti-sense orientation. In addition, the high resolution provided by deep sequencing allowed the quantification of small transcripts, namely *mcr11,* reported as inducible in Stat phase (18), the recently discovered gene *mymT* (27) and the sRNA encoded in the IG between *rv3661* and *rv3662c* whose expression increases significantly in Stat phase (18).

To validate the quantification method, 12 features covering a broad range of RPM values were selected for quantitative reverse transcription PCR (qRT-PCR) analysis. Results are reported in Supplementary Figure S4 and confirm the good correlation ($r^2 > 0.91$) between the Exp/Stat ratio calculated from RPM values and the corresponding figures obtained from qRT-PCR, thus justifying the use of RPM values for absolute comparisons between the different growth phases.

Correlation between the RNAP and NusA profiles and transcription

The availability of the ChIP-seq and RNA-seq data sets from two different growth conditions allowed correlation studies to understand the relationship between the presence of the transcriptional complex and RNA production. For this purpose, features with RNAP and NusA ER $\geq$ 2 and with RPM $\geq$ 1 were selected. From the histogram reported in Figure 3, it is evident that most of the features enriched in RNAP and NusA in Exp phase were also associated with transcription (89%). A smaller percentage of features with ER $\geq$ 2 for RNAP but no detectable NusA were transcribed (71%), suggesting that, indeed, the presence of NusA in the transcriptional complex favors transcription. The values changed markedly in the Stat phase, where a reduced number of features enriched in both RNAP and NusA were found to be transcribed (76%). Even more relevant, less than half of the RNAP-only-containing signals were associated with RPM values >1. Overall, the vast majority of the RNAP- and/or NusA-enriched features in Exp phase were also expressed, whereas approximately half of those enriched in Stat were associated with detectable transcripts, indicating that RNAP and NusA interact with the genome, but are not involved in active transcriptional activity during the Stat condition. Regarding the NusA-only-containing peaks (often intragenic), most of them correlated with a transcript at least in Exp phase, thereby leaving a small number (53) with unusual NusA binding.

Most of the surprising signals previously reported for ChIP-seq could be explained on direct comparison with the RNA-seq results. In the next section, we will describe two subsets of these peculiar features: those corresponding to anti-sense and to intergenic transcription. Importantly, this information allowed refinement of the existing *M. tuberculosis* genome annotation (4).

Unusual ChIP-seq peaks reflect anti-sense transcription in *M. tuberculosis*

As mentioned earlier, <2% of the RNA-seq reads mapped to the anti-sense orientation as compared with the existing annotation. Quantification uncovered 134 anti-sense transcribed features in Exp phase and 41 in Stat phase with 30 transcripts being expressed in both conditions (Supplementary Table S6). The majority of them could be explained as RNAs that are much longer than the corresponding annotated CDS, thereby generating anti-sense transcripts to the following, reverse-oriented, feature.

Comparison of the RPM values of the transcripts of the correct length and of the 3′ extension confirmed that the latter should be considered as spillover of the flanking RNAs, thus suggesting that weak transcriptional terminators might exist in *M. tuberculosis* (for instance, the terminators downstream of *rv2711*, *rv3878*, *rv2204c* and *rv0188*). In this section, we will not provide details of these, but rather focus on the anti-sense RNAs that are transcribed independently. On clustering them into functional categories, two groups appeared prominent, the 'Insertion sequences and phages' and the 'Unknown' categories, where a >2- or 3-fold increase was noticed compared with their abundance in the genome, respectively.

The top-scoring feature in Exp phase is the anti-sense transcript inside the gene *ino1*. Interestingly, this region contains one of the features that was unexpectedly highly enriched in RNAP and NusA ChIP-seq studies (Supplementary Figure S5A).

A closer look into the fraction of RNAP- and NusA-enriched features with no corresponding sense transcript revealed that some of these were associated with an anti-sense RNA. This is exemplified by *rv0842*, where RNAP and NusA densities at the beginning of the gene do not correlate with expression of the CDS but rather with transcription of an anti-sense RNA that extends in the upstream region (Supplementary Figure S5B). Furthermore, an anti-sense transcript was mapped inside the phiRv1 excisionase-encoding gene *rv1584c*. Its expression level increased in Exp and was reduced to the background level in Stat phase. On the contrary, no obvious explanation could be obtained for the RNAP peak inside the phiRv2 gene *rv2657c*. RNAP and NusA may just be 'sitting' there without transcribing (at least in the conditions tested), or the transcript could be undetectable because of stability issues. Finally, curious RNAP and NusA peaks were detected at the end of *rv0061*. After looking at the transcriptomic profile, these signals could be associated with anti-sense transcription of the whole CDS (Supplementary Figure S5C), thereby allowing re-annotation of this open reading frame as *rv0061c,* encoding a protein similar to *M. marinum* MMAR_3839, with 76% identity in 112 amino acid overlap.

Concerning the Stat condition, the top-scoring anti-sense RNA was noted inside *rv3684*. Accordingly, RNAP and NusA peaks are high in this position, thus cross-validating the finding (Supplementary Figure S6A). The genomic region encompassing the end of *rv1734c* was found to be highly transcribed in the anti-sense orientation in Stat phase (Supplementary Figure S6B). Correspondingly, the NusA density profile increases in this condition.

Intergenic ChIP-seq densities as hallmarks of intergenic transcripts

Within the 2283 IG regions of at least 30 bases in length, 1153 showed an RPM ≥1. We could identify most of these as part of TUs as described later, and IGs constituting the 5′- or 3′-UTR of flanking genes (15%) (Supplementary Table S6).

Examples of 5′-UTRs longer than 100 bp (listed in Supplementary Table S6) are represented by *rv3219* (Supplementary Figure S7A) and by the *rv3648c* mRNA (Supplementary Figure S7B). The latter feature, expressed at high levels in the Exp growth phase (28), showed high ERs for RNAP and NusA and suggested the existence of attenuation mechanisms in controlling gene expression.

Other examples include IG regions transcribed as independent features. This is the case of the IG at the end of the convergent genes *rv3661* and *rv3662c*: an sRNA (18), highly induced in Stat phase, is encoded here, and this is corroborated by enrichment of both RNAP and NusA in the ChIP-seq data.

The last type of IG directs anti-sense transcripts with respect to both of the flanking genes and therefore representing independent features such as sRNAs. The top-scoring instance is the IG between *rv1144* and *rv1145*, where an sRNA was identified, and high ER values for both RNAP and NusA were calculated (Supplementary Figure S7C).

Integration of ChIP-seq and RNA-seq data sets

To provide more quantitative correlation patterns between the dynamics of RNAP and NusA and transcription levels, a subset of well-known highly expressed TUs was selected. Specifically, 24 ribosomal protein operons, comprising a total of 75 CDSs and 57 IGs, represented the test set in the systems biology approach described hereafter. The level of transcription and enrichment of RNAP and NusA were quantified for each operon, with the 'operon promoter' being defined as the first feature of the TU (usually an IG), and the 'body' representing the remaining features. RNA-seq data suggested that the majority of these TUs were expressed in both Exp (21 of 24) and Stat (18 of 24) growth conditions as long, continuous transcripts, but the RPM values were significantly higher in the Exp compared with the Stat phase ($P < 10^{-6}$, Wilcoxon signed-rank test, Figure 4A). Promoter regions of all transcribed TUs showed a median ER greater than two for RNAP and NusA (the latter in Exp phase only). These values were considerably different in the body of the TUs: the median ER of RNAP and NusA was reduced, confirming the presence of the promoter-proximal peaks mentioned earlier rather than homogeneous distribution of the transcriptional complex throughout the operon. Interestingly, a significant higher level of NusA in Exp phase compared with Stat was detected in the body ($P < 10^{-4}$), whereas the RNAP ER in the operon body did not vary between the two conditions. Altogether, these results suggest that increased NusA occupancy in the operon body is associated with higher levels of transcription in the Exp compared with the Stat growth phase. In addition, the criteria that define a TU (i.e. promoter at the first feature and uninterrupted RNA) and the patterns identified with this subgroup of operons were applied to and examined in a larger data set, as described below.

Genome-wide TU identification

An algorithm (Figure 5A) was designed to discover TUs using the genome-wide ChIP-seq and RNA-seq profiles.

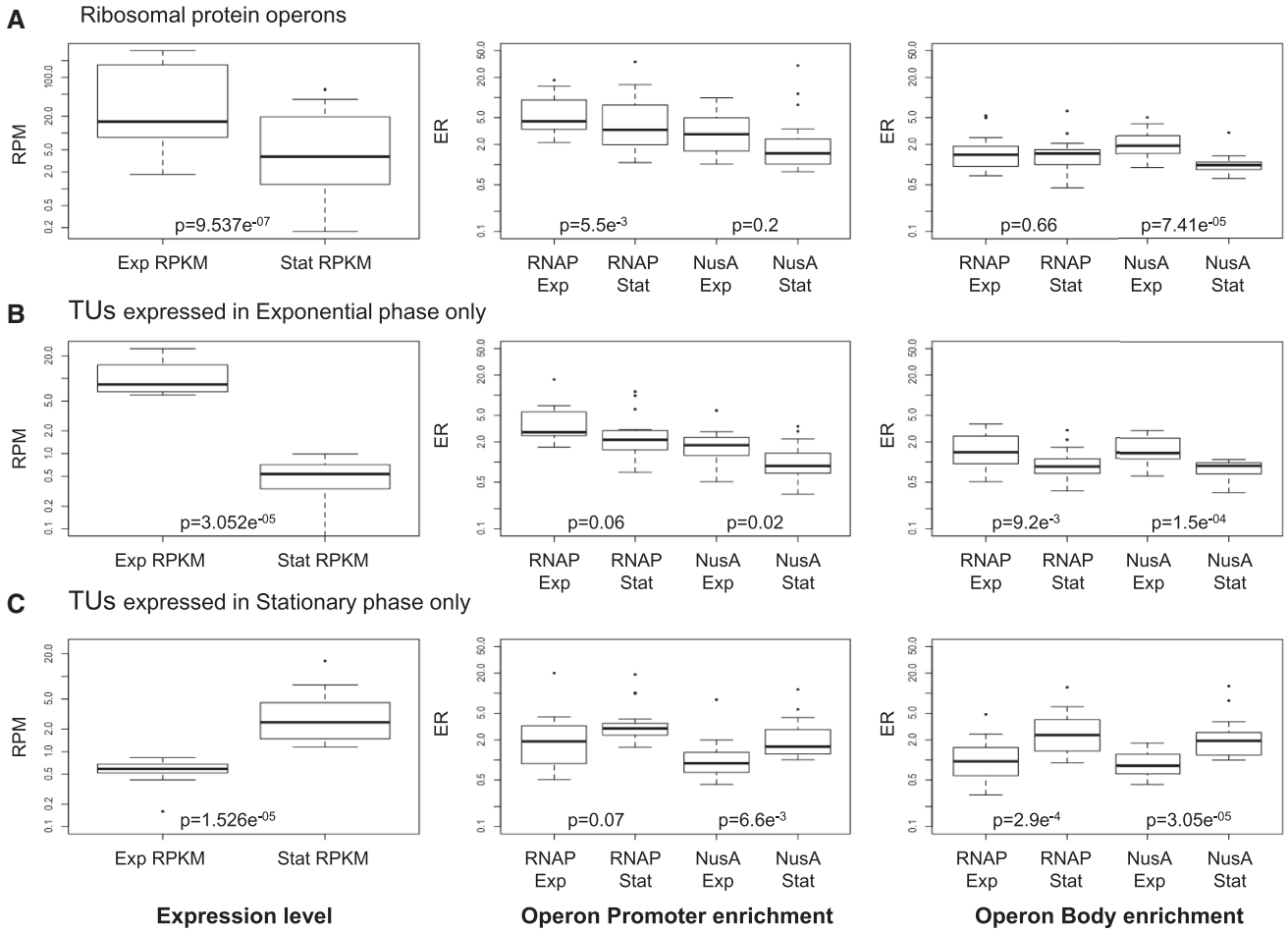
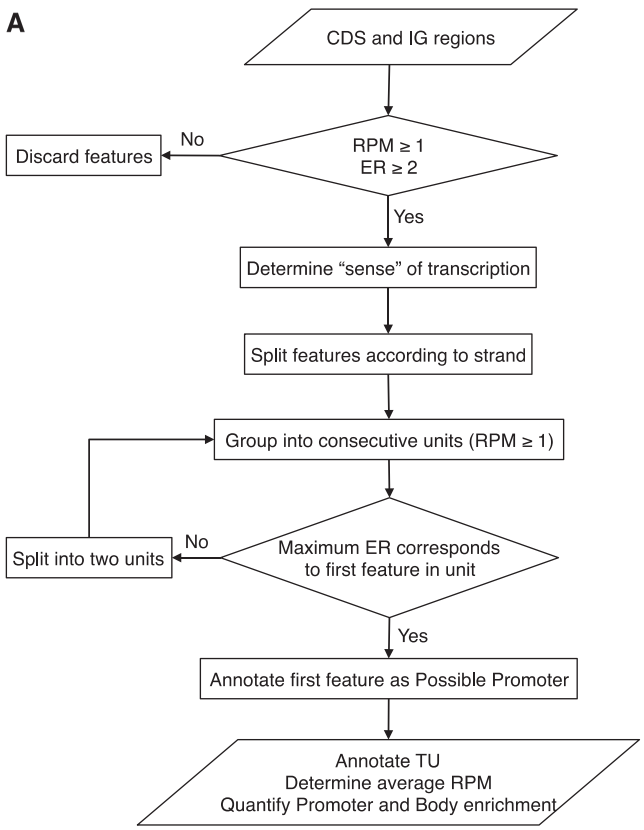
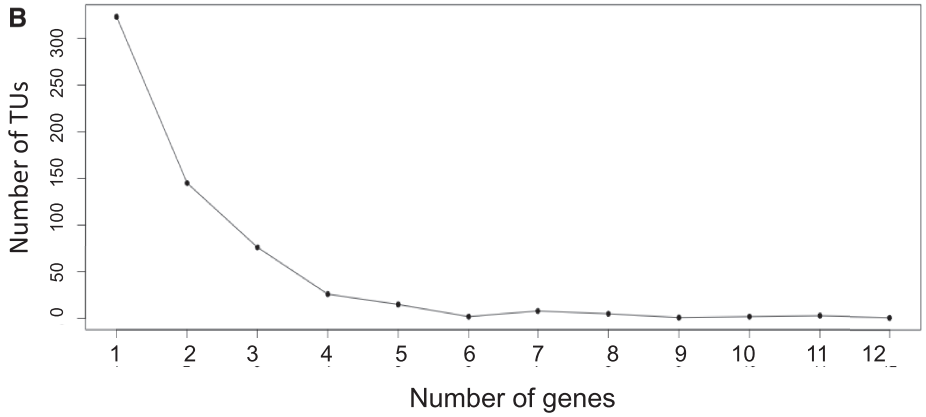


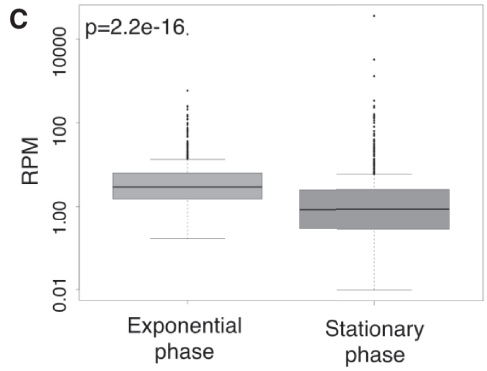
Figure 4. Box plots showing the distribution of RNAP and NusA in the promoter and body of TUs along with the transcription levels (RPM). (**A**) Ribosomal protein operons. (**B**) TUs expressed in Exp phase only and (**C**) TUs expressed in Stat phase only. *P*-values for comparison of different groups are based on the Wilcoxon signed-rank test.

Genomic features, including CDS and IGs, that displayed NusA or RNAP ER ≥ 2 and RPM ≥ 1 in at least one of the two growth phases were used as input. Features with RPM ≥ 1 were split into two groups based on the strand that was associated with higher RPM values. The forward and reverse sets of features were then sorted according to the genomic order and grouped into TUs. A TU was defined as a set of two or more consecutive features that are transcribed in the same direction and controlled by a promoter located at the first feature. Following the identification of transcription boundaries for all operons by means of RNA-seq data, promoters were mapped based on the RNAP and NusA ER. If the highest ER value was observed at the first feature, it was considered as a possible promoter. If the feature showing maximum ER did not correspond to the first one, the TU was split into two separate units at that feature, and promoter identification was repeated. This ensured that all TUs displayed a uniform organization characterized by enrichment of RNAP/NusA at the promoter and an uninterrupted transcript from the start to the end of the unit. Of the 817 continuous transcripts identified from RNA-seq, 606 were associated with significant promoter enrichment (ER ≥ 2) and therefore defined as 'high-quality TUs'. Of these, 301 were mapped on the forward strand and 305 on the reverse strand (Supplementary Table S7). Basic features of these TUs are reported in Figure 5. In all, 323 of these were single-gene units (composed of two features, one of them serving as a promoter), and the remainder comprised between 2 and 12 genes. The largest TUs were represented by the ATP synthase operon, the ESX-3 locus (*eccA3* to *eccE3*), and two ribosomal protein operons (from *rpsJ* to *rpsQ* and from *rplN* to *rplO*). A total of 268 operons were transcribed in both Exp and Stat phases, whereas 272 were transcribed only in the Exp phase and 17 only in the Stat phase, for instance the latter included *lipX-PPE17*, *rv2557-rv2558*, *usfY*, *lat*, *cysD, rv2660c*. Importantly, all of the previously described ribosomal protein operons were correctly identified with the developed workflow, underscoring the validity of the algorithm.

To appraise differences in RNAP and NusA distribution between Exp and Stat phase, the same parameters used for the characterization of the ribosomal protein operons were employed. Specifically, the RNAP and NusA ER values in the promoter and body of two subsets, namely the top 17 TUs transcribed in Exp phase only (Figure 4B) and the 17 TUs that were transcribed in







D

Growth Phase	RNAP TR	NusA TR	RNAP:NusA promoter	RNAP:NusA body
Exponential	0.51	0.75	1.78	1.11
Stationary	0.50	0.76	1.73	1.16

Figure 5. Integration of ChIP-seq and RNA-seq data to map TUs. (**A**) Workflow for TU identification. (**B**) Distribution of TU composition. (**C**) TU expression levels in Exp and Stat conditions. (**D**) Median TR for RNAP and NusA in Exp and Stat phases and RNAP:NusA ratio at the TU promoters and bodies.

Stat phase only (Figure 4C), were compared. The RPM and ER figures for those transcribed in Exp phase mirrored the observations made in the case of ribosomal protein operons (Figure 4A and B). Once again, the level of NusA in the body was significantly higher in Exp versus Stat phase ($P < 10^{-4}$, Wilcoxon signed-rank test). However, in contrast to what was observed for ribosomal protein operons, the RNAP ER in the TU body was also significantly increased in Exp phase ($P < 10^{-2}$), consistent with the lack of expression in Stat phase. Regarding the TUs transcribed in Stat phase only, complementary patterns to those in Figure 4B were observed (Figure 4C). Indeed, the levels of RNAP and NusA in the operon body were higher in the Stat phase compared with the Exp phase ($P < 10^{-3}$ and 10^{-4}, respectively) reflecting the difference in transcription between the two conditions. In conclusion, it is noteworthy that Exp phase-only and Stat phase-only transcribed TUs displayed RNAP ER at the promoters in both growth conditions, irrespective of the transcriptional levels, supporting the notion of a stationary enzyme in the non-transcribed phase.

Transition from transcription initiation to elongation is rate limiting

The rate of transition from transcription initiation to elongation was measured as the ratio of the RNAP and NusA ER in the body relative to the corresponding promoter ER. This value defines the traveling ratio (TR), and the medians for all of the TUs are reported in Figure 5D. The TR for RNAP was ~0.5, reflecting the presence of promoter-proximal peaks and implying that the transition from initiation to elongation is a rate-limiting step at most transcribed regions. The TR for NusA was slightly higher (0.75), suggesting that NusA-containing transcriptional complexes move more efficiently from the promoter throughout the TU. Further support for this concept came from the calculated RNAP:NusA ratio at promoters and in gene bodies. Indeed, this was in favor of RNAP at promoter regions and close to one in bodies, suggesting that a proportion of RNAP molecules, probably not associated with NusA, contacts the promoter but either does not progress or leaves abortively.

Global RNAP and NusA profiles throughout the TUs

Global profiles for RNAP and NusA were obtained by averaging the RCs for all of the identified TUs. Two highly expressed and enriched loci, namely *rrn* and the sRNA *mcr11*, were excluded from the analysis, so as to avoid a biased output. The TUs were split into two groups based on their genomic orientation, and each one was divided into 100 equal sized bins. The average number of reads for RNAP and NusA was calculated within each bin, and the mean for all operons was used to generate a single operon profile. The average enrichment in 50 bp regions flanking each TU was also included. Figure 6A presents the averaged forward and reverse profiles for RNAP and NusA in both phases of growth. Prominent enrichment was evident close to the start of the TUs on both strands, as pointed out earlier, and the average level of RNAP was higher than that of NusA in agreement with the previously calculated ratios. The decrease in the level of RNAP and NusA along the operon body reflects the polarity of transcription. As half of the identified operons are transcribed in both Exp and Stat phases, the profiles for RNAP and NusA in the two phases almost coincide.

Finally, detailed promoter profiles were generated for transcribed TUs as follows: the average number of reads per nucleotide was calculated for the RNAP and NusA data sets in an interval spanning 200 bp centered at the maximum of the RNAP peak for each promoter (i.e. the first feature of the TU). The forward and reverse strand results were merged to generate a single Exp and Stat phase profile for RNAP and NusA (Figure 6B). The averaged plots not only confirmed enrichment of NusA at promoters but also its co-localization with RNAP in both Exp and Stat phase. As a control, a subset of features enriched with RNAP (in both Exp and Stat phases) but not associated with transcription was chosen and the analysis repeated. Unlike the profile observed for transcribed TUs, the NusA enrichment around the RNAP peak was much lower for non-transcribed features and did not display the same shape nor co-localize with the RNAP peak (Figure 6C), thereby confirming NusA as a transcription promoting factor in *M. tuberculosis* in both of the experimental conditions tested.

DISCUSSION

Our aim was to map the genome of the major human pathogen, *M. tuberculosis*, by systematically studying the activity of the transcriptional complex and its interaction with DNA. Using high-resolution ChIP-seq analyses, we demonstrated the ubiquitous association of RNAP and NusA with the genome and the significant overlap of the two binding profiles, thus confirming that NusA is part of the RNAP complex. ChIP-seq was an extremely powerful tool to monitor NusA, a protein that does not interact directly with DNA but rather with RNAP and/or with the nascent transcript.

Quantitative reproducibility between biological replicates was extremely high in the case of NusA in both of the conditions tested, whereas a reduced correlation was noted for RNAP. This could be the consequence of a lower affinity of the monoclonal antibody against RpoB as compared with the polyclonal NusA anti-serum or could result from intrinsic higher variability of the RNAP dynamics on the *M. tuberculosis* genome caused by non-specific or random interactions of the enzyme with DNA, as previously reported in *E. coli* (29,30). Interactions between the transcriptional complex and DNA might be stronger when NusA is present, thus explaining the increased reproducibility of the NusA profiles. Indeed, the density of RNAP-only-containing peaks was found to be less reproducible than that of peaks associated with NusA, supporting our hypothesis.

In addition to ChIP-seq, RNA-seq was exploited to obtain a global view of transcriptional regulation in *M. tuberculosis* and to explain unusual peaks observed

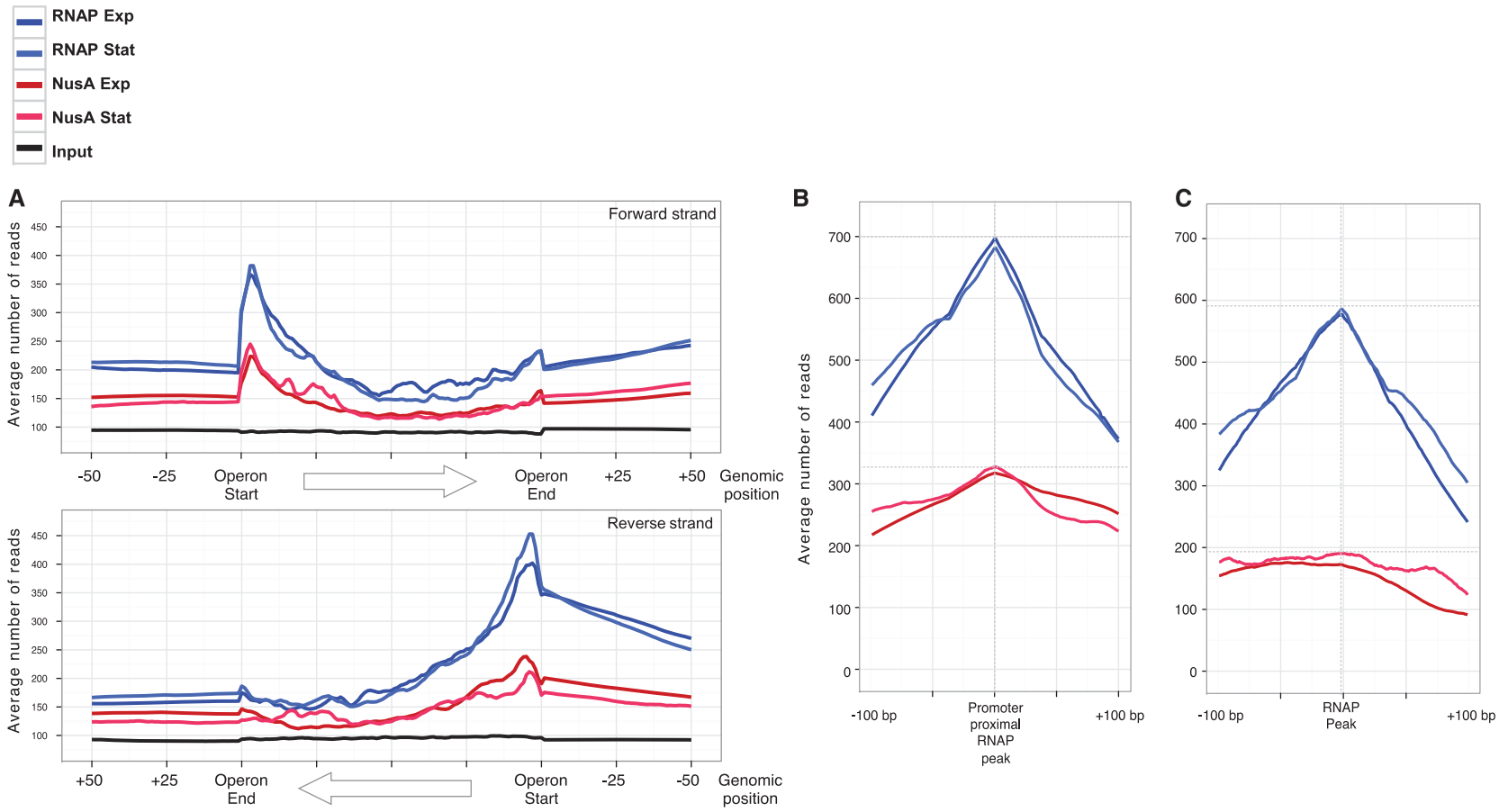


Figure 6. Global RNAP and NusA profiles of the TUs. (**A**) Averaged profiles for RNAP and NusA across forward and reverse TUs. The *Y*-axis indicates the average number of reads, the *X*-axis the genomic position with an arbitrary length for the TU. (**B**) Averaged RNAP and NusA profiles at promoter regions for RNAP-enriched transcribed features. (**C**) Averaged RNAP and NusA profiles at promoter regions for RNAP-enriched non-transcribed features. The legend is provided on top.

in the ChIP-seq profiles. Quantification of RNA-seq results by RPM values with normalization to the total number of reads allowed absolute comparison of the conditions tested, in a way that resembles microarray-based transcriptomic approaches, where total fluorescence was used as normalizing factor (24). The Stat profile was similar to that of a starved culture (15), and hallmarks of this condition were readily identified. In this model, a 4-week old *M. tuberculosis* culture may mimic the starvation condition, as all of the nutrients have been consumed. Comparison of the RNA-seq data generated here with those published in a previous study (18) (full comparison in Supplementary Table S8) revealed an overall correlation of 0.40 for the CDS in Exp phase (Supplementary Figure S8), whereas no correlation was observed in the Stat growth phase (data not shown). Discrepancies can be explained by the different culture conditions used, the significantly different age of the Stat phase cultures and distinct RNA-seq protocols.

Importantly, the deep-sequencing transcriptomic experiments reported here exhibited striking strand specificity, with a minor portion of the sequence tags (<2%) mapping to the CDSs in anti-sense orientation. We therefore considered this anti-sense transcription as genuine and exploited the data to improve the *M. tuberculosis* genome annotation (4). As an illustration, the CDS *rv0061* was renamed *rv0061c* following the results of this work. Several anti-sense transcripts mapped by RNA-seq found additional confirmation by ChIP-seq profiling. BLAST analysis performed on the top-scoring anti-sense-transcribed features revealed that some of them are specific to the MTB complex (e.g. *rv1374c*, *rv1734c*), whereas others, like *ino1*, showed >85% conservation at the nucleotide level in other mycobacterial species such as *M. marinum, M. smegmatis* and *M. leprae*, suggesting that the anti-sense RNA encoded there might be conserved. Curiously, two features (*rv2660c* and *rv1954c*) that were described as induced on starvation (15) were also found to be upregulated in Stat phase but in the anti-sense orientation. The strand-specificity of RNA-seq uncovered this discrepancy, whereas previously used PCR-based microarrays did not allow correct strand identification (15). Consistent with our observation, a recent publication reported a possible CDS on the forward strand (*rv1954A*) at the *rv1954c* locus (31). On the contrary, no alternative open reading frame can be predicted at the *rv2660c* locus, suggesting that the anti-sense RNA may play regulatory roles. Ironically, Rv2660c has been proposed as part of a multi-subunit TB vaccine (32), even though its gene is transcribed in the opposite direction.

Intergenic transcription and corresponding RNAP/NusA ER added further value. A number of long 5′-UTRs has been identified, such as those of *whiB1* (*rv3219*) and of *cspA* (*rv3648c*), suggesting the existence of *cis*-regulatory elements affecting transcription elongation or attenuation mechanisms, including those mediated by riboswitches. To date, some of these have been predicted or validated in *M. tuberculosis* (33–35). The presence of NusA at these genomic positions confirms the role of this protein in the transcription attenuation process, in agreement with earlier studies (36,37). The single-nucleotide resolution conferred by RNA-seq allowed the identification of intergenic sRNAs, and independent confirmation by ChIP-seq was obtained. Most of the previously described small transcripts (18,38–40) were easily recognized in our study, and ~90% overlap was observed. Some of the discrepancies encountered could be related to different culture conditions or sequencing methodologies. Prediction of the conservation and of the functional role carried out by these transcripts remains elusive, given the reduced sequence identity occurring in IGs among the various mycobacterial species.

A major strength of our work lies in the combination of ChIP-seq with RNA-seq technologies, and therefore the cross-validation of the respective data sets. Approximately the same number of features was found to be enriched in RNAP and/or NusA in the two conditions tested, but there was good correlation with transcription in Exp phase only. On the contrary, roughly half of the RNAP/NusA signals were associated with RNA in Stat phase. Overall, although on the one hand, RNA stability effects cannot be ruled out, on the other, these data suggest a model where RNAP and NusA bind throughout the *M. tuberculosis* genome, but their activity might be impeded by the lack of additional transcription factors and/or by the lack of nutrients, especially in Stat phase. It has been reported that ~23% of the RNAP signals in growing *E. coli* were not associated with RNA, thus leading to the definition of 'poised' RNAP (41). This is similar to the percentage (21%) of RNAP- and/or NusA-enriched features that do not correspond to a transcript in Exp phase in our study. Indeed, RNAP sitting on promoter sequences has to undergo several steps of abortive initiation and requires the presence of activators, such as GreA (42), and/or small molecules (43) or a particular DNA topology (44) to proceed fruitfully. A recent study identified transcription start site associated RNAs in *Mycoplasma pneumoniae* that overlap RNAP pausing sites (45). These small RNA molecules may exist in *M. tuberculosis* as well, thus explaining some of the unusual RNAP and/or NusA peaks that were not associated with a detectable transcript. As the results of RNA-seq and RNAP ChIP-seq are not necessarily the same in *M. tuberculosis*, care should be taken with replacing transcriptomic profiling by detection of RNAP-binding sites. Indeed, RNA studies (and nowadays powerful deep transcriptomic profiling) represent the gold standard for gene expression analysis.

The NusA-only-containing signals deserve some discussion. The majority of these peaks were associated with RNA either in the same or in the neighboring feature, whereas only 53 signals could not be correlated to transcription. Technical reasons could explain why RNAP was not detected above the background, for instance IP might be difficult because the epitope is masked by other transcription factors or by intricate chromatin structures in those regions. Moreover, interaction of NusA with other proteins cannot be excluded *a priori*, as well as a role of NusA in looping the chromosome. In the latter case,

ChIP-seq would detect distal sites that are contacted by the same NusA-containing complexes.

Integration of ChIP-seq and RNA-seq data in a systems biology context provided the most relevant output to this work. Based on the criteria defined with the ribosomal protein operons (i.e. presence of an RNAP/NusA peak at the first feature and a continuous transcript throughout the TU), we developed an algorithm that identified 606 high-quality TUs widely distributed along the *M. tuberculosis* genome. More TUs may be present on the chromosome but will have escaped identification, as our algorithm was based on data generated *in vitro* in specific growth conditions. This is the first study of its kind to have performed genome-wide TU identification relying on empirical data rather than on predictions as done previously (46). We could confirm those TUs for which experimental evidence in the literature exists, e.g. the *furA-katG* operon (47), the *yrbE1A-mce1F* locus (48), the *pstB-pstA2* unit (49) and the *rv3134c-devS* operon (50). Our approach is complementary to the transcriptional start site (TSS) mapping developed in *Helicobacter pylori* by means of a differential RNA-seq method (51). Although the latter procedure allows precise identification of the 5′-end of primary transcripts, the combination of RNAP/NusA ChIP-seq and RNA-seq provides a more dynamic view of the interactions of the transcriptional complex with the genome and circumvents potential technical issues related to RNA stability.

By studying RNAP and NusA binding across TUs and their correlation with transcription, we established some common patterns in both growth phases that shed light on the biology of the transcriptional complex in *M. tuberculosis*. Firstly, promoter-proximal peaks for RNAP (and, to a lesser extent, for NusA) were recognized in the global TU profile. These signals may correspond to RNAP trapped at promoters before promoter clearance (i.e. abortive initiation), represent pause sites or result from premature transcription termination or attenuation mechanisms. Similar findings were previously described in *E. coli* (13), where the simultaneous occurrence of RNAP and NusA peaks allowed the classification of those promoter-proximal signals as elongating (EC) rather than stalled complexes. Given the strong analogy with our results, we infer that this is the case for

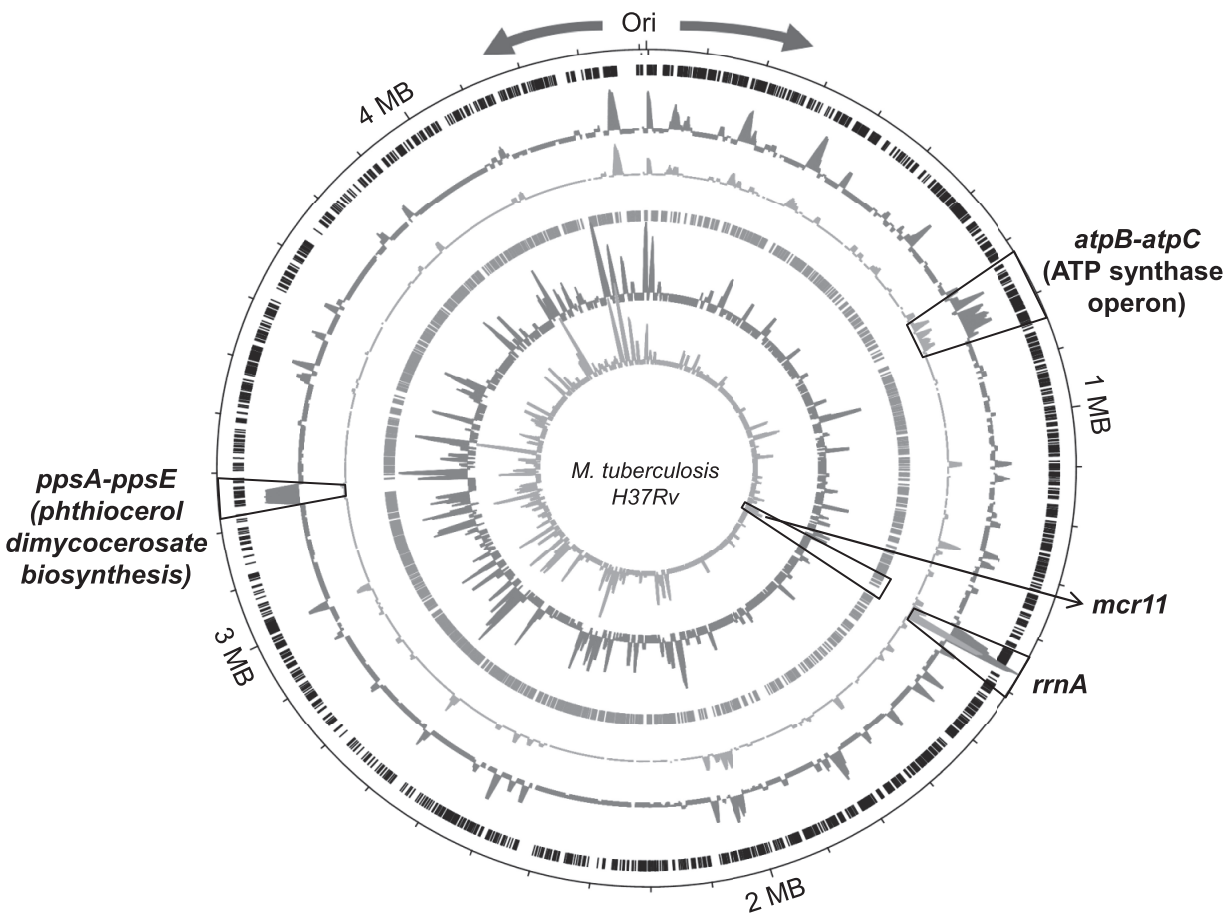


Figure 7. Circular genome view of the transcription profiles for TUs in Exp and Stat phases. *Ori* indicates the origin of replication. The outermost circle represents CDSs on the forward strand, followed by the Exp phase expression profile for TUs on the forward strand in dark gray, the Stat phase expression profile for TUs on the forward strand in light gray. The fourth circle represents CDSs on the reverse strand, followed by the Exp phase expression profile for TUs on the reverse strand in dark gray and the Stat phase expression profile for TUs on the reverse strand in light gray. Genomic coordinates are marked at every 1 MB. Arrows indicate the direction of the replication forks. Some of the highly transcribed loci are boxed and annotated.

M. tuberculosis as well, thus revealing Actinobacteria to be more similar to Proteobacteria than to the Firmicute *B. subtilis* (14) in the mechanistic basis of gene expression. Secondly, significant abundance of NusA in the body of the transcribed TUs was observed, underlining its role in promoting transcription by anti-termination processes. In addition, the TR for RNAP was found to be <1, suggesting that the transition from initiation to elongation is a rate-limiting step in *M. tuberculosis* as well, similarly to *E. coli* (42,52). Slightly different from the RNAP TR, the TR for NusA was 0.75, suggesting that NusA-containing complexes move through the TUs faster. Indeed, the ratio between RNAP and NusA ER at promoters was in favor of RNAP, as witnessed by the global promoter profile, indicating that a proportion of the RNAP molecules does not associate with NusA and probably does not move along the genes. Overall, the importance of NusA in promoter clearance was also highlighted, as features lacking the NusA peak in the corresponding RNAP enrichment were not expressed. Finally, a striking polarity in the distribution of RNAP and NusA was noticed in the TU profiles.

The whole-genome distribution of the TUs (Figure 7) revealed a strong bias in the expression levels with respect to the direction of the replication forks. The majority of the highly transcribed TUs were localized on the leading strand thus resulting in an image where the *M. tuberculosis* chromosome is split into two non-symmetrical halves with the replication terminus located tentatively at ~2.1 MB. Interestingly, while the orientation of the CDSs in the H37Rv genome is only slightly biased [59% with the same polarity as DNA replication (4)], their expression levels are more affected, as confirmed by applying a Kolmogorov–Smirnov test to the cumulative RPM along each strand resulting in a $P < 10^{-10}$ (Supplementary Figure S9). This is different from what was observed in other bacteria, such as *E. coli* (53) and *B. subtilis* (54), where the orientation bias is more pronounced. Our genome-wide transcriptional study shows that in a slow-growing pathogenic bacterium the DNA and RNA polymerases proceed in the same direction through most of the highly expressed features thus avoiding potential collisions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–8 and Supplementary Figures 1–9.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Kristine Arnvig and Dr. Ian Taylor (NIMR, London, UK) for providing purified NusA, Dr. Ida Rosenkrands (Statens Serum Institut, Copenhagen, Denmark) for generating anti-NusA antibodies, Dr. Keith Harshman (Lausanne Genomic Technologies Facility, University of Lausanne, Switzerland) for advice on RNA-seq experiments. S.T.C. and C.S. designed the study; C.S. performed the experiments; S.U., J.R., S.T.C., C.S. analyzed data; S.U., S.T.C. and C.S. wrote the article.

FUNDING

The European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement [260872]; SystemsX.ch and the Swiss National Science Foundation [31003A-125061]. Funding for open access charge: European Community Seventh Framework Programme FP7/2007-2013 under grant agreement 260872.

Conflict of interest statement. None declared.

REFERENCES

1. Russell,D.G. (2011) *Mycobacterium tuberculosis* and the intimate discourse of a chronic infection. *Immunol. Rev.*, **240**, 252–268.
2. Russell,D.G., VanderVen,B.C., Lee,W., Abramovitch,R.B., Kim,M.J., Homolka,S., Niemann,S. and Rohde,K.H. (2010) *Mycobacterium tuberculosis* wears what it eats. *Cell Host Microbe*, **8**, 68–76.
3. Stokes,R.W. and Waddell,S.J. (2009) Adjusting to a new home: *Mycobacterium tuberculosis* gene expression in response to an intracellular lifestyle. *Future Microbiol.*, **4**, 1317–1335.
4. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. 3rd *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.
5. Greenblatt,J. and Li,J. (1981) The nusA gene protein of *Escherichia coli*. Its identification and a demonstration that it interacts with the gene N transcription anti-termination protein of bacteriophage lambda. *J. Mol. Biol.*, **147**, 11–23.
6. Vogel,U. and Jensen,K.F. (1997) NusA is required for ribosomal antitermination and for modulation of the transcription elongation rate of both antiterminated RNA and mRNA. *J. Biol. Chem.*, **272**, 12265–12271.
7. Arnvig,K.B., Pennell,S., Gopal,B. and Colston,M.J. (2004) A high-affinity interaction between NusA and the rrn nut site in *Mycobacterium tuberculosis*. *Proc. Natl Acad. Sci. USA*, **101**, 8325–8330.
8. Gusarov,I. and Nudler,E. (2001) Control of intrinsic transcription termination by N and NusA: the basic mechanisms. *Cell*, **107**, 437–449.
9. Schmidt,M.C. and Chamberlin,M.J. (1987) nusA protein of *Escherichia coli* is an efficient transcription termination factor for certain terminator sites. *J. Mol. Biol.*, **195**, 809–818.
10. Cardinale,C.J., Washburn,R.S., Tadigotla,V.R., Brown,L.M., Gottesman,M.E. and Nudler,E. (2008) Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. *coli*. *Science*, **320**, 935–938.
11. Gopal,B., Haire,L.F., Gamblin,S.J., Dodson,E.J., Lane,A.N., Papavinasasundaram,K.G., Colston,M.J. and Dodson,G. (2001) Crystal structure of the transcription elongation/anti-termination factor NusA from *Mycobacterium tuberculosis* at 1.7 A resolution. *J. Mol. Biol.*, **314**, 1087–1095.
12. Worbs,M., Bourenkov,G.P., Bartunik,H.D., Huber,R. and Wahl,M.C. (2001) An extended RNA binding surface through arrayed S1 and KH domains in transcription factor NusA. *Mol. Cell*, **7**, 1177–1189.
13. Mooney,R.A., Davis,S.E., Peters,J.M., Rowland,J.L., Ansari,A.Z. and Landick,R. (2009) Regulator trafficking on bacterial transcription units in vivo. *Mol. Cell*, **33**, 97–108.
14. Ishikawa,S., Oshima,T., Kurokawa,K., Kusuya,Y. and Ogasawara,N. (2010) RNA polymerase trafficking in Bacillus subtilis cells. *J. Bacteriol.*, **192**, 5778–5787.
15. Betts,J.C., Lukey,P.T., Robb,L.C., McAdam,R.A. and Duncan,K. (2002) Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol. Microbiol.*, **43**, 717–731.
16. Dahl,J.L., Kraus,C.N., Boshoff,H.I., Doan,B., Foley,K., Avarbock,D., Kaplan,G., Mizrahi,V., Rubin,H. and Barry,C.E. III (2003) The role of RelMtb-mediated adaptation to stationary phase in long-term persistence of *Mycobacterium*

tuberculosis in mice. *Proc. Natl Acad. Sci. USA*, **100**, 10026–10031.
17. Voskuil,M.I., Visconti,K.C. and Schoolnik,G.K. (2004) *Mycobacterium tuberculosis* gene expression during adaptation to stationary phase and low-oxygen dormancy. *Tuberculosis (Edinb)*, **84**, 218–227.
18. Arnvig,K.B., Comas,I., Thomson,N.R., Houghton,J., Boshoff,H.I., Croucher,N.J., Rose,G., Perkins,T.T., Parkhill,J., Dougan,G. *et al.* (2011) Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathog.*, **7**, e1002342.
19. Sala,C., Haouz,A., Saul,F.A., Miras,I., Rosenkrands,I., Alzari,P.M. and Cole,S.T. (2009) Genome-wide regulon and crystal structure of BlaI (Rv1846c) from *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **71**, 1102–1116.
20. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
21. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
22. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
23. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
24. Manganelli,R., Dubnau,E., Tyagi,S., Kramer,F.R. and Smith,I. (1999) Differential expression of 10 sigma factor genes in *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **31**, 715–724.
25. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
26. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
27. Gold,B., Deng,H., Bryk,R., Vargas,D., Eliezer,D., Roberts,J., Jiang,X. and Nathan,C. (2008) Identification of a copper-binding metallothionein in pathogenic mycobacteria. *Nat. Chem. Biol.*, **4**, 609–616.
28. Hu,Y., Butcher,P.D., Mangan,J.A., Rajandream,M.A. and Coates,A.R. (1999) Regulation of hmp gene transcription in *Mycobacterium tuberculosis*: effects of oxygen limitation and nitrosative and oxidative stress. *J. Bacteriol.*, **181**, 3486–3493.
29. deHaseth,P.L., Lohman,T.M., Burgess,R.R. and Record,M.T. Jr (1978) Nonspecific interactions of *Escherichia coli* RNA polymerase with native and denatured DNA: differences in the binding behavior of core and holoenzyme. *Biochemistry*, **17**, 1612–1622.
30. Grigorova,I.L., Phleger,N.J., Mutalik,V.K. and Gross,C.A. (2006) Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. *Proc. Natl Acad. Sci. USA*, **103**, 5332–5337.
31. Smollett,K.L., Fivian-Hughes,A.S., Smith,J.E., Chang,A., Rao,T. and Davis,E.O. (2009) Experimental determination of translational start sites resolves uncertainties in genomic open reading frame predictions - application to *Mycobacterium tuberculosis*. *Microbiology*, **155**, 186–197.
32. Aagaard,C., Hoang,T., Dietrich,J., Cardona,P.J., Izzo,A., Dolganov,G., Schoolnik,G.K., Cassidy,J.P., Billeskov,R. and Andersen,P. (2011) A multistage tuberculosis vaccine that confers efficient protection before and after exposure. *Nat. Med.*, **17**, 189–194.
33. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.
34. Vitreschak,A.G., Mironov,A.A., Lyubetsky,V.A. and Gelfand,M.S. (2008) Comparative genomic analysis of T-box regulatory systems in bacteria. *RNA*, **14**, 717–735.
35. Warner,D.F., Savvi,S., Mizrahi,V. and Dawes,S.S. (2007) A riboswitch regulates expression of the coenzyme B12-independent methionine synthase in *Mycobacterium tuberculosis*: implications for differential methionine synthase function in strains H37Rv and CDC1551. *J. Bacteriol.*, **189**, 3655–3659.
36. Sha,Y., Lindahl,L. and Zengel,J.M. (1995) Role of NusA in L4-mediated attenuation control of the S10 r-protein operon of *Escherichia coli*. *J. Mol. Biol.*, **245**, 474–485.
37. Yakhnin,A.V. and Babitzke,P. (2002) NusA-stimulated RNA polymerase pausing and termination participates in the *Bacillus subtilis* trp operon attenuation mechanism invitro. *Proc. Natl Acad. Sci. USA*, **99**, 11067–11072.
38. Arnvig,K.B. and Young,D.B. (2009) Identification of small RNAs in *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **73**, 397–408.
39. DiChiara,J.M., Contreras-Martinez,L.M., Livny,J., Smith,D., McDonough,K.A. and Belfort,M. (2010) Multiple small RNAs identified in *Mycobacterium bovis* BCG are also expressed in *Mycobacterium tuberculosis* and *Mycobacterium smegmatis*. *Nucleic Acids Res.*, **38**, 4067–4078.
40. Pellin,D., Miotto,P., Ambrosi,A., Cirillo,D.M. and Di Serio,C. (2012) A genome-wide identification analysis of small regulatory RNAs in *Mycobacterium tuberculosis* by RNA-Seq and conservation analysis. *PLoS One*, **7**, e32723.
41. Reppas,N.B., Wade,J.T., Church,G.M. and Struhl,K. (2006) The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. *Mol. Cell*, **24**, 747–757.
42. Stepanova,E., Lee,J., Ozerova,M., Semenova,E., Datsenko,K., Wanner,B.L., Severinov,K. and Borukhov,S. (2007) Analysis of promoter targets for *Escherichia coli* transcription elongation factor GreA *in vivo* and *in vitro*. *J. Bacteriol.*, **189**, 8772–8785.
43. Lee,S.J. and Gralla,J.D. (2004) Osmo-regulation of bacterial transcription via poised RNA polymerase. *Mol. Cell*, **14**, 153–162.
44. Shin,M., Song,M., Rhee,J.H., Hong,Y., Kim,Y.J., Seok,Y.J., Ha,K.S., Jung,S.H. and Choy,H.E. (2005) DNA looping-mediated repression by histone-like protein H-NS: specific requirement of Esigma70 as a cofactor for looping. *Genes Dev.*, **19**, 2388–2398.
45. Yus,E., Guell,M., Vivancos,A.P., Chen,W.H., Lluch-Senar,M., Delgado,J., Gavin,A.C., Bork,P. and Serrano,L. (2012) Transcription start site associated RNAs in bacteria. *Mol. Syst. Biol.*, **8**, 585.
46. Roback,P., Beard,J., Baumann,D., Gille,C., Henry,K., Krohn,S., Wiste,H., Voskuil,M.I., Rainville,C. and Rutherford,R. (2007) A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic Acids Res.*, **35**, 5085–5095.
47. Pym,A.S., Domenech,P., Honore,N., Song,J., Deretic,V. and Cole,S.T. (2001) Regulation of catalase-peroxidase (KatG) expression, isoniazid sensitivity and virulence by furA of *Mycobacterium tuberculosis*. *Mol. Microbiol.*, **40**, 879–889.
48. Casali,N., White,A.M. and Riley,L.W. (2006) Regulation of the *Mycobacterium tuberculosis* mce1 operon. *J. Bacteriol.*, **188**, 441–449.
49. Torres,A., Juarez,M.D., Cervantes,R. and Espitia,C. (2001) Molecular analysis of *Mycobacterium tuberculosis* phosphate specific transport system in *Mycobacterium smegmatis*. Characterization of recombinant 38 kDa (PstS-1). *Microb. Pathog.*, **30**, 289–297.
50. Bagchi,G., Chauhan,S., Sharma,D. and Tyagi,J.S. (2005) Transcription and autoregulation of the Rv3134c-devR-devS operon of *Mycobacterium tuberculosis*. *Microbiology*, **151**, 4045–4053.
51. Sharma,C.M., Hoffmann,S., Darfeuille,F., Reignier,J., Findeiss,S., Sittka,A., Chabas,S., Reiche,K., Hackermuller,J., Reinhardt,R. *et al.* (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, **464**, 250–255.
52. Wade,J.T. and Struhl,K. (2008) The transition from transcriptional initiation to elongation. *Curr. Opin. Genet. Dev.*, **18**, 130–136.
53. Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
54. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.