Phyletic distribution and conservation of the bacterial transcription termination factor Rho

François D’Heygère,1,2 Makhlouf Rabhi1 and Marc Boudvillain1,3

1Centre de Biophysique Moléculaire, CNRS UPR4301, rue Charles Sadron, 45071 Orléans cedex 2, France
2École doctorale Santé, Sciences Biologiques et Chimie du Vivant (ED 549), Université d’Orléans, France
3ITP Sciences Biologiques et Chimie du Vivant, Université d’Orléans, France

Correspondence
Marc Boudvillain
marc.boudvillain@cnrs-orleans.fr

Received 1 March 2013
Accepted 15 May 2013

Transcription termination factor Rho is a ring-shaped, ATP-dependent molecular motor that targets hundreds of transcription units in Escherichia coli. Interest in Rho was renewed recently on the realization that this essential factor is involved in multiple interactions and cellular processes that protect the E. coli genome and regulate its expression on a global scale. Yet it is currently unknown if (and how) Rho-dependent mechanisms are conserved throughout the bacterial kingdom. Here, we mined public databases to assess the distribution, expression and structural conservation of Rho across bacterial phyla. We found that rho is present in more than 90% of sequenced bacterial genomes, although Cyanobacteria, Mollicutes and a fraction of Firmicutes are totally devoid of rho. Genomes lacking rho tend to be small and AT-rich and often belong to species with parasitic/symbiotic lifestyles (such as Mollicutes). By contrast, large GC-rich genomes, such as those of Actinobacteria, often contain rho duplicates and/or encode Rho proteins that bear insertion domains of unknown function(s). Notwithstanding, most Rho sequences across taxa contain canonical RNA-binding and ATP hydrolysis signature motifs, a feature suggestive of largely conserved mechanism(s) of action. Mutations that impair binding of bicyclomycin are present in ~5% of rho sequences, implying that species from diverse ecosystems have developed resistance against this natural antibiotic. Altogether, these findings assert that Rho function is widespread among bacteria and suggest that it plays a particularly relevant role in the expression of complex genomes and/or bacterial adaptation to changing environments.

INTRODUCTION

In bacteria, the Rho factor mediates one of the two major pathways that lead to the termination of transcription (Ciampi, 2006; Peters et al., 2011). Since eukaryotes do not contain Rho homologues, the protein makes an attractive target for the development of novel antibiotics. Early studies with the natural Rho inhibitor bicyclomycin (BCM) delineated the spectrum of organisms susceptible to Rho inhibition (Nishida et al., 1972). BCM is active against various Gram-negative bacteria, but inactive against all tested Gram-positive bacteria with the exception of Micrococcus (Mc.) luteus (Nowatzke et al., 1997b; Williams & Durham, 1988). This has been interpreted to suggest that Rho function is important in Gram-negative bacteria, but dispensable in most Gram-positive species (Rabhi et al., 2010b; Washburn et al., 2001). The rationale for this simple rule is, however, not firmly established. Notably, the phyletic distribution and structural conservation of Rho within living organisms were inferred more than a decade ago from very limited sets of DNA sequences (Opperman & Richardson, 1994; Washburn et al., 2001). Hence, the fractions of bacteria devoid of rho genes or containing Rho factors whose function may be unaffected by BCM are not known.

Typically, the Rho factor is a ring-shaped homo-hexameric enzyme with RNA-dependent ATP hydrolysis activity (reviewed by Banerjee et al. 2006; Boudvillain et al., 2010; Ciampi, 2006; Peters et al., 2011). This ATPase activity fuels the translocation of Rho along nascent transcripts and the ensuing displacement of roadblocks formed by RNA:DNA duplexes (e.g. transcriptional R-loops) or by RNA polymerases halted at Rho-dependent termination sites (Fig. 1a). Rho function is tightly controlled in vivo and
usually requires uncoupling of translation from transcription (or lack of translation), which exposes nascent RNA rut (Rho utilization) sites to binding by Rho. Activation of Rho function is nonetheless frequent in \textit{Escherichia coli} where Rho mediates many transcription termination events (current estimates, 20–50 %) and regulates hundreds of transcription units encoding proteins as well as non-coding RNAs (transfer RNAs, small RNAs and antisense RNAs) (Cardinale et al., 2008; Peters et al., 2009, 2012). Accordingly, diverse biological roles were proposed for Rho in \textit{E. coli} and related bacteria (reviewed by Boudvillain et al., 2013). They include control of faithful gene expression \textit{e.g.} by transcriptional polarity (Ciampi, 2006; Peters et al., 2011), regulation of specific genes/operons in response to environmental cues \textit{notably through ribo-switch- and small RNA-dependent mechanisms (Bossi et al., 2012; Hollands et al., 2012)}], prevention of spurious transcription (Cardinale et al., 2008; Peters et al., 2009, 2012) and of R-loop formation (Krishna Leela et al., 2013), and roles in maintenance of genome integrity (Dutta et al., 2011; Washburn & Gottesman, 2011). These functions rely more-or-less directly on Rho cooperating with various endogenous factors. Notably, Rho binds to transcription factors NusA and NusG (Li et al., 1993; Schmidt & Chamberlin, 1984), antitermination factor YaeO (Pichoff et al., 1998) and riboregulation factor Hfq (Rabhi et al., 2011b), and co-purifies with a host of other factors (Butland et al., 2005; Hu et al., 2009) that vary with growth conditions to include, for instance, RNA degradosome components PNPase and RNase E (Tuckerman et al., 2011). At present, the extent to which these various roles and interactions are conserved throughout the bacterial kingdom remains largely uncertain. This is due to past research that focused primarily on Rho from \textit{E. coli} and on very few other bacteria. For instance, the possibility that some Rho homologues operate by distinctive mechanisms is supported by only one recent \textit{in vitro} study devoted to the Rho factor from \textit{Mycobacterium tuberculosis} (Kalarickal et al., 2010). The authors show that the transcription termination activity of \textit{Mb. tuberculosis} Rho can be totally uncoupled from ATP hydrolysis, an unexpected feature that may be linked tentatively to the presence of a large
N-terminal domain (NTD) insertion within the Rho polypeptide (Kalarickal et al., 2010). Since rho inactivation (Griffin et al., 2011), but not BCM treatment (Nishida et al., 1972), inhibits mycobacterial growth, the data also argue that sensitivity to the ATPase inhibitor BCM (Skordalakes et al., 2005) is not always a reliable gauge of the functional significance of Rho.

To better appraise the importance of Rho across the bacterial kingdom, we explored public databases for clues supporting the presence and functional role of rho in the thousands of bacterial genomes that have now been sequenced. We also assessed the conservation of important residues/motifs revealed by previous studies of Rho, including (1) the cold shock domain (CSD)-like RNA-binding domain (RBD) that is located within Rho’s NTD and ensures recruitment of transcript rut regions on the crown-like primary binding site (PBS) of the Rho hexamer (Fig. 1b, c, in cyan), (2) the positively charged N-terminal helix bundle (NHB) domain located on the periphery of the hexamer (Fig. 1b, c, in yellow) that may facilitate RNA recruitment to the PBS (Canals et al., 2010), (3) ATP hydrolysis signature motifs [Walker A (P-loop) and B motifs] located within Rho’s C-terminal domain (CTD) (Fig. 1b, c, in red), (4) CTD residues involved in coordinating ATP hydrolysis or structural transitions within the hexamer [e.g. ‘catalytic’ Glu, Arg ‘valve’ and Arg ‘finger’ residues (Fig. 1b, c, in purple)] and (5) CTD residues forming the secondary binding site (SBS) (Q and R-loops, in green and blue in Fig. 1b, c, respectively) and required for RNA translocation within the ring channel (Rabhi et al., 2010a). The arrangement of these components in a crystal structure of E. coli Rho (ecRho), which shows the hexamer in a closed, asymmetrical configuration (Fig. 1c), suggests a mechanism for RNA translocation, whereby each nucleotide is SBS-bound by a single Rho subunit and driven through the central channel in response to sequential chemical changes in the ATP-binding pockets (Thomsen & Berger, 2009). However, this monotonous, ATPase-driven ‘escort’ mechanism does not account for the ~7 nt periodic stepping behaviour of ecRho (Boudvillain et al., 2010; Schwartz et al., 2009) nor explain how Rho from Mb. tuberculosis triggers efficient transcription termination in spite of very poor ATPase activity (Kalarickal et al., 2010). Thus, we have also looked for features that could reflect alternative Rho mechanisms, notably the presence of atypical domain insertions and/or aa substitutions. We show that the Rho factor is generally well conserved and expressed across the bacterial kingdom, but highlight intriguing features in a number of cases that will require further exploration and, eventually, experimental assessment. We observe that rho is absent from a limited number of species (~8% of our compiled dataset), which include all Cyanobacteria and Mollicutes as well as a significant fraction of Firmicutes. Species from various phyla were found to have duplicated and/or atypical rho genes, including ones bearing deletions/insertions of unknown function(s) or mutations that are likely to alter enzymic activity or to confer resistance to BCM. Interestingly, genomes containing rho sequence(s) tend to be larger and GC-richer than rho-less genomes, suggesting a link between Rho function and the complexity of genome structure (or expression) and/or lifestyle of the bacterial species. Overall, our work significantly extends early phylogenetic analyses (Opperman & Richardson, 1994; Washburn et al., 2001) that were performed with limited sets of bacterial sequences and before detailed structural information on Rho became available, thereby providing a comprehensive framework for future analyses of this central regulatory factor.

**METHODS**

*Database searches.* We searched the Xbase, Uniprot, GenBank and NCBI protein databases for rho coding sequences (CDSs) or ORFs and for Rho proteins using a combination of keyword queries, as well as BLAST searches with selected target sequences. Sequences that displayed homology with Rho NTD (PSSM id: 208474) only, such as hypothetical protein MAE_06000 (414 aa) from Microcystis aeruginosa, were not considered significant hits. We thoroughly checked all available genomes of species lacking rho to ensure absence of the gene itself or lack of any significant remnant. Control searches were conducted with PSI-BLAST (threshold E-value=0.05) and distinct Rho baits, including ones from bacteria having phyletic proximity with the rho-less species (proximity based on SILVA’s tree of life; http://www.arb-silva.de). Rho baits were b3783 (E. coli), Rv1297 (Mb. tuberculosi), BSU57080 (Bacillus subtilis), HMPreF00871232 (Enterococcus casseliflavus), sce7932 (Sorangium cellulosum) and Hprae_1978 (Halanaerobium praevalens). Consistent with long-known homology (Opperman & Richardson, 1994), usually the best PSI-BLAST hits for rho-less species were ATP synthase subunits with 40-60% coverage (corresponding to Rho’s CTD) and 20-30% identity scores.

When several completed genomes existed for a given species, we arbitrarily selected the strain with the largest genome or the most representative strain (e.g. E. coli K-12 MG1655) to build our datasets. Whenever possible, we also favoured fully assembled genomes over draft or incomplete ones. To perform all subsequent analyses on uniform datasets, searches for rho− and rho-less species were stopped at the end of January 2012. Locus tags for the annotated rho genes were used to retrieve Rho sequences from the Uniprot database and to build FASTA files. Theoretical isolectric points and hydrophobicity GRAVY (grand average of hydropathicity) indexes were calculated with PROTPARAM on the ExPASy portal (http://www.expasy.org).

**Sequence alignments and phylogeny trees.** Sequence alignments and phylogeny trees were constructed with SEAVIEW 4.3.3 (Gouy et al., 2010). Sequences were aligned with the MUSCLE algorithm implemented in SEAVIEW using default parameters. Unrooted phylogeny trees were built with the BioNJ distance method (bootstrap=100; gap sites ignored). Iterative cycles of sequence alignment and phylogenetic ordering were performed until convergence of both was obtained. Detection and alignment of NHB and NTD insertion domain residues over the full dataset were more satisfactory with this approach than by direct alignment with alternative algorithms such as CLUSTAL-Ω (not shown). Phylogeny was then evaluated with maximum-likelihood (PHYLML) and/or parsimony (PHYLIP) algorithms, as implemented in SEAVIEW. Bootstrap support was estimated based on 100 replicates. Conversion of alignments into image files was performed with JALVIEW (Waterhouse et al., 2009) using the FASTA files generated with SEAVIEW. Histograms for sequence conservation (based on physicochemical properties), quality (unlikelyhood to observe a
non-conservative substitution in the alignment column) and consensus (percentage of consensus nucleotide in the column) were generated with the default parameters implemented in Jalview. Conservation and quality histograms are shaded by default from dark brown (low values) to yellow (high values). Shannon entropy analysis was performed on the Protein Variability Server (Garcia-Boronat et al., 2008).

RESULTS AND DISCUSSION

Rho homologues are found in many bacterial phyla

Database mining provided evidence for a rho ORF (or CDS) in 1250 distinct bacterial species, including 706 species whose genome had been fully sequenced and assembled at the time of analysis. This search essentially covered the ~3600 Rho entries that were present in InterPro (accession no. IPR004665; http://www.ebi.ac.uk/interpro). About 99.9% of these entries derive from rho sequences found in bacterial genomes, often from several strains of the same species (not shown). For instance, we retrieved more than 250 entries for ecRho corresponding to various E. coli strains or genome annotation projects. To reduce representation biases, analysis was thus conducted with a curated dataset that contained Rho sequence(s) from the most advanced annotation project for only one representative strain per bacterial species (Table S1, available in Microbiology Online).

Species containing rho belong to nearly all bacterial phyla with the noticeable exceptions of Tenericutes and Cyanobacteria (see below), which confirms previous observations (Opperman & Richardson, 1994; Washburn et al., 2001). Analysis of published data also indicates that Rho proteins are expressed in representative species across phyla (Table S2). The putative (or experimentally detected) proteins encoded by the compiled rho ORFs have an average length of 480 aa, although the most frequent Rho length is closer to 420 aa (Fig. 2a). However, length distributions may be biased by disproportionate representations of phyla among the sequenced genomes. Indeed, the typical Rho length appears to vary significantly with the phylum and is noticeably larger for Actinobacteria and Bacteroidetes (Fig. 2b) due to the presence of NTD insertions and/or extensions (see below). Variations in Rho length are also larger within these phyla as well as among Firmicutes, Spirochaetes and members of the Planctomycetes–Verrucomicrobia–Chlamydia superphylum (Fig. 2b). We did not detect any correlation between Rho length and specific features of the corresponding genome such as its size, number of genes or GC content (data not shown).

The longest Rho sequence (865 aa) in our dataset is from the marine extremophile Thermarobacter marianensis (Firmicutes). As observed for Rho proteins from other organisms, this is due to a long NTD region (~490 aa) upstream from the CSD-like RBD (Fig. 1b) that bears little resemblance to other known sequences (no significant BLAST hit; data not shown). Although very short Rho sequences (<300 aa) are present in the dataset, they usually correspond to partial reads from uncompleted genome assemblies (WGG5 genomes in Table S1). In fact, the shortest Rho sequences deduced from fully assembled genomes are those of two γ-Proteobacteria, Colwellia psychrerythraea (314 aa) and Marinomonas sp. MWYL1 (318 aa), and both lack important RNA-binding motifs (RNP1 and RNP2; Fig. 1b) in their NTD. The two γ-Proteobacteria, however, contain another, longer rho ORF (Table S1) that encodes all key motifs supporting Rho activity, including canonical RNA-binding motifs (see below). Although it is unknown if Rho factors are expressed in these two bacteria (Table S2), these observations may indicate that one of the rho duplicates has been inactivated, possibly to prevent functional redundancy. Inactivation can also result from frame-shift mutations that disrupt rho, as is observed for Dehalococcoides ethenogenes (Chloroflexi) and B. pumilus (Firmicutes). In these cases, however, there is no evidence for a second rho ORF (Table S1), suggesting that Rho function is dispensable in these two micro-organisms.

We were surprised to detect rho in a few eukaryotic genomes. In two instances (Drosophila ananassae and Callosobruchus chinensis), rho presence appears to be due to horizontal gene transfers (HGTs) from Wolbachia endosymbionts (Dunning Hotopp et al., 2007; Nikoh et al., 2008). In other instances, however, we could not find a better explanation for the presence of rho than bacterial contamination of the genomic DNA used for sequencing.

The rho gene is absent from Cyanobacteria, Mollicutes and a fraction of Firmicutes

We did not find rho in any of the fully assembled genomes of Cyanobacteria (29 species) and Tenericutes (27 Mollicutes) that were present in GenBank at the time of our analysis (Table S3). We also did not find rho in the fully assembled genomes of 43 other bacterial species (Table S3), most of which are Firmicutes from Clostridia, Bacilli and Negativicutes classes. If one also considers uncompleted genome assemblies, it appears that a minority of Clostridia species but all negativicutes and most bacilli of the order Lactobacillales (e.g. Streptococcus pneumoniae, a species unaffected by BCM [Nishida et al., 1972]) do not have rho (Tables S1 and S3; data not shown). By contrast, all bacilli of the order Bacillales (e.g. B. subtilis or Staphylococcus aureus) contain a rho CDS (Table S1). In recently constructed trees of Firmicutes (Wolf et al., 2004; Zhao et al., 2005; see also SILVA project at http://www.arb-silva.de), rho-less species are polyphyletic, suggesting that the rho gene was repeatedly lost in several, independent events during evolution.

Finally, we did not find rho in the fully assembled genomes of four Proteobacteria (Table S3). Three are insect endosymbionts that have very small genomes due to
reductive evolution processes (McCutcheon et al., 2009). Massive gene loss and pseudogenization typically occur during adaptation to intracellular life (Bordenstein & Reznikoff, 2005), and it is possible that \(\rho\) has been lost in the process. Reductive evolution may also explain the absence of \(\rho\) in Mollicutes (Table S3), a class of bacteria with small genomes and parasitic behaviour (e.g. Mycoplasma). The obligate symbiotic/parasitic lifestyles of these organisms may have helped render Rho-dependent regulation superfluous. However, this determinant is not necessarily sufficient or compelling. For instance, obligate intracellular Rickettsia and Chlamydia contain CDS encoding canonical Rho proteins (Table S1). Furthermore, Rho is expressed in at least some Chlamydia species (Table S2).

We observed that \(\rho\)-less species tend to have smaller, GC-poorer genomes than species containing \(\rho\) ORFs \((P<10^{-5};\) Fig. 3). This trend remains significant even when Mollicutes are excluded from the analysis (Fig. 3). The amount of GC is probably not an independent factor because it is itself positively correlated with genome size [for reasons that remain speculative (see Bentley & Parkhill, 2004; Hildebrand et al., 2010)]. Our observations thus suggest that \(\rho\) function is linked to some form of complexity in genome organization and/or expression, even though some exceptions are notable. For instance, we could not find traces of \(\rho\) in the large, GC-rich genome of Haliangium ochraceum (Table S3), a marine microorganism (\(\delta\)-Proteobacteria) with complex metabolic and morphogenesis features, including social motility and fruiting body behaviours (Ivanova et al., 2010). The limited genomic information available for other members of the Nannocystineae suborder suggests that \(\rho\) may be totally absent from this branch of Myxococcales (data not shown). This would be different from other myxobacteria, such as Sorangium or Myxococcus species, which bear at least one and often two \(\rho\) ORFs within their large genomes (see Table S1). Considering that Rho is expressed in Sorangium cellulosum (Table S2), it will be interesting to determine if that is the case in other myxobacteria and why Rho is absent from members of the Nannocystineae suborder.

We did not detect other unusual features in genome architecture, such as specific biases in strand composition, or in orientation of CDSs, which would be characteristic of \(\rho\)-less genomes (data not shown). We also did not detect obvious correlations between the absence of Rho and that of major transcription factors or Rho-interacting proteins. We found hfq genes and even a hypothetical \(\rho\) (\(\gamma\)eO) gene in \(\rho\)-less Firmicutes while we could not find hfq and/or \(\rho\) genes in many Rho-containing species (data not shown). Moreover, we found transcription factors NusA and NusG in nearly all \(\rho\)-less species, probably because these factors have other essential functions in transcription.
elongation and intrinsic termination mechanisms (Mooney et al., 1998; Yang & Lewis, 2010). Intrinsic transcription termination mechanisms could actually compensate for the absence of Rho. In agreement with this proposal, computational searches uncovered high frequencies of putative intrinsic terminator motifs in \( \rho \)-less genomes of Firmicutes (de Hoon et al., 2005), Mollicutes (Kingsford et al., 2007) and Cyanobacteria (Vijayan et al., 2011). Notable exceptions are genomes of Mycoplasma genitalium, My. pneumoniae and Synechocystis PCC 6803 in which putative intrinsic terminators were only scarcely detected (de Hoon et al., 2005; Kingsford et al., 2007; Washio et al., 1998). However, it remains currently uncertain if these exceptions are due to limitations of the computational search methods or if they reflect predominance of alternative, unidentified termination mechanisms (de Hoon et al., 2005; Kingsford et al., 2007).

**Some bacterial genomes contain two rho ORFs**

During our analysis, we observed that a number of species from diverse phyla contain two distinct rho ORFs (Table S1), a feature suggestive of HGT or chromosomal duplication. To better profile rho duplications, we have built a phylogenetic tree of all Rho duplicates found in the fully assembled bacterial genomes available at the time of analysis. As shown in Fig. 4, phylogenetic relationships between Rho duplicates vary from close relatedness (e.g. in \( \alpha \)-Proteobacteria) to deeply rooted divergence (e.g. in \( \gamma \)- and \( \delta \)-Proteobacteria). Rho duplicates are monophyletic for \( \alpha \)-Proteobacteria, Chloroflexi or Actinobacteria species (Fig. 4), suggesting that they originate from chromosomal duplication events. In the case of Ehrlichia and Anaplasma spp., synteny also supports chromosomal duplication/inversion events (Fig. S1), which tend to be frequent for these species (Dunning Hotopp et al., 2006).

In contrast, \( \gamma \)- and \( \delta \)-Proteobacteria, Planctomycetes or Bacteroidetes that contain two Rho homologues do not appear to have duplicated a pre-existing rho gene. Such organisms contain one Rho duplicate that clusters with Rho sequences from \( \alpha \)-Proteobacteria, Chloroflexi or Actinobacteria species (Fig. 4, group A), and a second that clusters with Rho sequences from Verrucomicrobia (group B). The great phylodivergence between duplicates from groups A and B (Fig. 4) and lack of obvious compositional bias or synteny (data not shown) make it difficult to assert the origin of these rho duplications without a more detailed analysis, which is beyond the scope of the present study.

As noted above, the absence of important RNA recognition motifs in group B’s Rho factors from *C. psychrerythraea* and *Marinomonas* sp. *MWYL1* (Fig. S2) should severely compromise their enzymic activity. Significant deletions within Rho’s NTD are also apparent for additional
Fig. 4. Phylogram of the putative Rho proteins encoded by genomes containing rho duplicates. Numbers at nodes are bootstrap support values obtained upon maximum-likelihood analysis. Groups A and B contain Rho duplicates from γ- and δ-Proteobacteria, Planctomycetes and Bacteroidetes. For each species, one Rho duplicate belongs to group A, whereas the other falls within the phylo-distinct group B.

members of group B as well as for half of the Rho duplicates from Actinobacteria (Fig. 5, orange bars; data not shown). This may indicate that functional redundancy is often suppressed by enzymic inactivation of one of the Rho duplicates. Alternatively, the altered Rho duplicate may have evolved towards a new function or mechanism that does not require the deleted domains, yet provides a selective advantage (Näsivall et al., 2012). It is also possible
Large insertion domain region

Actinobacteria
Spirochaetes
Firmicutes
-Bacteroidetes
-Proteobacteria
-Proteobacteria
-Proteobacteria
-Proteobacteria
-Proteobacteria

Conservation: NHB CSD-like RBD Wa Wb
Quality:
Consensus:

N-terminal domain (NTD) (64–577 aa)
C-terminal domain (CTD) (217–345 aa)
that the Rho duplicate is not expressed at all. To conclude decisively, it will be necessary to determine the respective expression patterns of the Rho duplicates. This may be particularly relevant for species containing Rho duplicates that also bear aa substitutions conferring resistance to BCM (see below), an obvious selective advantage.

**Sequence conservation in Rho homologues**

To assess the conservation of important Rho motifs, we have aligned full-length Rho sequences listed in Table S1. Sequence alignments reveal a number of highly conserved regions along the Rho polypeptide that are well outlined by conservation histograms and by red/blue hydrophobic/phyllic ‘vertical lines’ in Fig. 5 (due to the ‘zoomed out’ alignment of colour-boxed residues required to show all the 1259 Rho sequences). These regions include NHB and CSD-like RBD segments within Rho’s NTD, the region that connects the NTD and CTD, as well as most of Rho’s CTD (Fig. 5; see also below). Overall, Rho’s CTD appears more conserved than that of NTD, which confirms previous observations (Opperman & Richardson, 1994). Of particular notice is the fact that most Rho factors from Actinobacteria and Bacteroidetes, a significant fraction of Rho factors from Firmicutes and a few additional Rho factors from other phyla contain large insertion domains between the NHB and CSD-like RBD regions (Fig. 5). Altogether, these NTD insertion domains are found in ~35% of our compiled Rho sequences and, contrary to what was envisioned from limited datasets (Nowatzke et al., 1997a), are not restricted to high G + C Gram-positive bacteria. Their sizes and compositions are not conserved (Fig. 5), although they often contain R/K-rich (mostly in Actinobacteria) or Q/N-rich (mostly in Bacteroidetes) stretches (data not shown). The NTD insertion domains are predominantly hydrophilic (Fig. 5 and data not shown; whole domain hydrophilic characters are also supported by negative GRAVY index values) but some are significantly basic (e.g. theoretical pl>9 for the whole domain), while others are acidic (theoretical pl<6). These insertion domains may promote specific Rho behaviours entailed by particular ecosystems, cellular metabolisms, genome structures and so on. For instance, the insertion domain of *Mc. luteus* Rho facilitates promoter-proximal transcription termination with structured RNA transcripts, a feature that seems consistent with the high GC content of the *Mc. luteus* genome (Nowatzke et al., 1997a). Similar observations have been made for the Rho factor from *M. tuberculosis* whose efficient termination activity in spite of poor ‘motor’ features may also be due to the presence of the insertion domain (Kalarickal et al., 2010).

Overall, Actinobacteria contain the most atypical Rho factors with multiple insertions that alter the whole NTD (upstream from, between and within the NHB and CSD-like RBD) and that extend the CTD (Fig. 5). By contrast, Rho sequences from Proteobacteria form a rather homogeneous set of sequences, as evidenced by both length distribution (Fig. 2b) and alignment of sequences (Fig. 5). Homogeneity is particularly high when considering the γ-Proteobacteria alone (~235 species). Although a few Rho homologues contain NTD insertions, most γ-Proteobacteria sequences start right upstream from the NHB domain, as best exemplified by ecRho (Fig. S2 and data not shown). Given the importance of Rho in many γ-Proteobacteria, we have performed Shannon entropy analysis (Garcia-Boronat et al., 2008) on the γ-Proteobacteria subset of Rho sequences (cleaned of NTD insertions) to better assess evolutionary sequence constraints imposed by Rho function. The γ-Proteobacteria-derived functional ‘core’ that was thereby analysed (Fig. 6, top) contains all the regions/motifs that are significantly conserved across phyla (Fig. 5). To facilitate interpretation of the entropy data, we have calculated a smoothed entropy curve whereby core motifs/regions of highest conservation are easily identified by local minima or ‘valleys’ (Fig. 6, red curve). For instance, the deepest entropy valley identifies the β7/ω8 junction region which contains the ‘catalytic’ Glu and Arg ‘valve’ residues (Fig. 6; Glu211 and Arg212 in ecRho) involved in ATP→ADP transitions (Thomsen & Berger, 2009) as the most conserved region of γ-Proteobacteria Rho proteins. This region is also extremely well conserved across phyla (see sequence logo in Fig. 6) with motif ERPEEVT (catalytic Glu and Arg valve residues in bold) found in all but two of the Rho sequences examined. Interestingly, the two divergent Rho sequences also contain non-conservative substitutions in other key motifs (Table S4 and data not shown) that suggest that Rho has been inactivated or relies on a different mechanism in the corresponding species (actinobacteria *Conexibacter woesei* and *Trophyerma whipplei*).
Three shallower entropy valleys identify the best conserved regions of Rho’s NTD within the CSD-like RBD (Fig. 6). These regions include the RNP1 and RNP2 motifs as well as all the PBS residues that contact RNA/DNA ligands in crystal structures of ecrh (Fig. 6, blue stars). These motifs and PBS residues are also relatively well conserved across phyla (Fig. 6; see also Table S4), although they are missing in a subset of Rho duplicates (Fig. 5, orange bars; data not shown). By contrast, the NHB subdomain is much less conserved among γ-Proteobacteria Rho factors or across phyla (Fig. 6 and Fig. S3). Hence, we have found only weak phylogenetic evidence to support the mechanism of electrostatic RNA recruitment by a positively charged NHB subdomain that has been inferred from the crystal structure of Thermotoga maritima Rho (Canals et al., 2010).

Another region of high conservation in γ-Proteobacteria connects the NTD and CTD and contains a few residues that are also well conserved across phyla (Fig. 6). To the best of our knowledge, these residues have never been subjected to systematic mutagenesis, although a K→Q substitution at position 130 (consensus position 132 in Fig. 6) spontaneously emerged in a genetic screen of impaired Rho function in Salmonella (Bossi et al., 2012). Although it has been suggested that the linker region could be important for allosteric communication between the PBS and SBS (see Rabhi et al., 2011a, and references within) or for interaction with NusG (Bossi et al., 2012), these conjectures await experimental testing.

As is observed for the complete set of Rho sequences (Fig. 5), the most conserved region in γ-Proteobacteria is the central part of Rho’s CTD (~210–350 region in Fig. 6). In addition to the β7/α8 junction region discussed above, this segment of Rho’s CTD includes the Walker B motif, Q- and R-loops, and the α12 helix, which have all been implicated in mechanochemical transduction within ecrh (Thomsen & Berger, 2009). Key residues within these motifs, notably those contacting RNA in the central channel of ecrh, are reasonably well conserved across phyla (Fig. 6, bottom), although a few Rho sequences...
contain non-conservative substitutions (Table S4) that may be hallmarks of inactivation or alternative mecanochemistry (Adelman et al., 2006). Side-chains forming the BCM-binding pocket in eC-Rho (Skordalakes et al., 2005) are also highly conserved across phyla (Fig. 6, pink triangles) with only ~5% of rho-containing species bearing a substitution at a critical position (Fig. S4). Interestingly, a majority of these substitutions are homologous to the single-point S266A and G337S substitutions that confer resistance to E. coli due to a weakened interaction between eC-Rho and BCM (Magyar et al., 1999; Skordalakes et al., 2005). This suggests that the corresponding bacterial species (Fig. S4) will also be poorly sensitive to BCM. Of particular notice are the cases of Catenulispora and Streptomyces species that contain Rho duplicates (Fig. S4 and Table S1). One of the duplicates almost systematically bears a substitution altering BCM binding (homologous to G337S; Fig. S4) as well as other changes in the CSD-like RBD (not shown). Whether these mutated Rho duplicates are expressed, functional and capable to confer resistance to BCM has not been determined. If true, this would contribute to explaining why S. lividans, which contains such a mutated Rho duplicate (Fig. S4), is poorly sensitive to BCM (Ingham et al., 1996). Similarly, it will be worth checking if BCM-producing Streptomyces spp. also contain a mutated Rho duplicate and/or if they use alternative transcription termination (Kalarickal et al., 2010) or efflux (Bentley et al., 1993) mechanisms to outwit their own metabolite.

It is noteworthy that all of the Rho sequences from Rickettsia and fusobacteria species that are present in our dataset contain a substitution that confers a BCM-binding defect (Fig. S4 and Table S1). This suggests that all of these species, most of which are pathogens, will be poorly sensitive to BCM. It is also significant that substitutions conferring BCM-binding defects are found in both Gram-negative and Gram-positive bacteria (Fig. S4) from very diverse ecosystems (e.g. soils, mucus membranes, host cells). This is intriguing given that the only known natural sources for BCM are soil Streptomyces spp. (Miyamura et al., 1973; Miyoshi et al., 1972). In any case, the widespread occurrence of aa substitutions conferring resistance to BCM lends additional support to the view that Rho is important in a large and diverse set of bacterial species.

CONCLUSIONS

More than 90% of the bacterial genomes that were sequenced at the time of our analysis contain a canonical rho gene (Tables S1 and S3). The prevalence of rho across the bacterial kingdom suggests that its origin is very ancient and that Rho-dependent functions are not easily substituted or discontinued in bacterial cells.

Rho is expressed in representative species across phyla (Table S2), yet is dispensable in some of them under laboratory conditions (Quirk et al., 1993; Washburn et al., 2001). Even spurious transcription elicited by rho inactivation is not necessarily catastrophic, as observed for B. subtilis (Nicolas et al., 2012), although condition-dependent or long-term effects may not be readily detected. For instance, genome instability may arise from uncontrolled encounters between transcription and replication machineries, since Rho inactivation should favour their co-directional collisions and the resultant formation of DNA double-strand breaks (Dutta et al., 2011). Traffic control may also explain why Rho is sometimes more important or abundant under conditions such as cold (Mykytczuk et al., 2011; Piette et al., 2010; Quirk et al., 1993) that favour formation (and thus require clearance) of roadblocks such as arrested RNA polymerases and R-loops (Erie, 2002; Massé & Drolet, 1999). Termination mechanisms that protect genomes from transcription-associated instability seem universal (Mischo et al., 2011; Skourtì-Stathaki et al., 2011) and may thus well constitute the least common denominator of Rho function across the bacterial kingdom.

Small genomes tend to belong to species restricted to a stable niche, often with a parasitic/symbiotic lifestyle, whereas bacteria with large genomes tend to occupy highly complex and variable environments (Bentley & Parkhill, 2004). Since species without rho tend to have smaller genomes than species that contain rho (Fig. 3), requirement for rho may be linked, at least in part, to the complexity of the bacterium ecosystem and to the need for mechanisms that ensure protection/adaptation to environmental changes. Selective pressure for rho may notably arise from ‘immunity-like’ mechanisms whereby Rho contributes to repressing harmful transcription of HGT sequences (Cardinale et al., 2008; Mitra & Nagaraja, 2012; Peters et al., 2009). By limiting exposure to foreign DNA (Bordenstein & Reznikoff, 2005), the restrictive lifestyles of obligate intracellular bacteria may thus reduce the selective pressure for Rho function(s).

Although alterations of functionally important Rho motifs are sometimes observed, key residues involved in Rho’s oligomerization, ATP hydrolysis, RNA binding and RNA translocation are highly conserved (Fig. 6). The most noticeable disparity among Rho sequences is the presence of a large NTD insertion domain in about a third of them (Fig. 5). This insertion domain has been proposed to facilitate Rho binding to GC-rich transcripts in Actinobacteria (Nowatzke et al., 1997a). However, its length and composition are not conserved and it is also found in Rho factors from GC-poor Firmicutes and Bacteroidetes (Fig. 5). Thus, the role(s) of the insertion domains remain(s) unclear and will have to be investigated for a large and representative set of species. Notably, it will be important to determine if insertion domains can elicit alternative termination mechanisms such as the one proposed for Mb. tuberculosis Rho (Kalarickal et al., 2010). Together with the present study, this information should expand our understanding of Rho functions and mechanisms across the bacterial kingdom and may also
facilitate the rational design of high-throughput screening assays in the search for new antibiotics targeting Rho.

ACKNOWLEDGEMENTS

We gratefully acknowledge Drs Elena Bidnenko, Soraya Ait-Bara, Lionello Bossi and Nara Figueroa-Bossi for insightful discussions and critical reading of the manuscript. This work was supported by a joint PhD scholarship from the University of Orleans and Region Centre to F.D’H. and by CNRS core funding to M.B.

REFERENCES


Edited by: R. Lan