Essentiality, not expressiveness, drives gene-strand bias in bacteria

Eduardo P C Rocha[1,2] & Antoine Danchin[1,3]

Preferential positioning of bacterial genes in the leading strand was thought to result from selection to avoid high head-on collision rates between DNA and RNA polymerases. Here we show, however, that in *Bacillus subtilis* and *Escherichia coli*, essentiality (the transcript product), not expressiveness (the collision rate), selectively drives the biased gene distribution.

Replication and transcription occur simultaneously on the same DNA molecule. DNA polymerases (DNAPs) in *E. coli* proceed 10–20 times faster than RNA polymerases (RNAPs), and both head-on and co-oriented collisions often occur in replicating bacteria[1,2]. In co-oriented collisions (genes in the leading strand), DNAP tends to slow down until transcription is completed, and replication then resumes its normal pace. In head-on collisions (genes in the lagging strand), replication stalls, RNAP is displaced and transcription is aborted[2]. It has been proposed that highly expressed genes are preferentially positioned in the leading strand to allow faster DNA replication and lower transcript losses[1]. This is systematically observed for rDNA and ribosomal proteins in bacterial genomes[3]. But several lines of evidence are challenging this view[4]. First, there is no obvious correlation between bacterial growth rates and gene-strand bias. Second, gene-strand bias depends on the composition of DNAP, with bacteria having two dedicated DNAPs showing much stronger biases.

Here, we analyzed the *B. subtilis* gene distribution relative to gene essentiality and expression level. Essentiality was determined from extensive gene inactivation data[5]. Ribosomal proteins were used to build a codon adaptation index (CAI), which we applied to all genes in the genome[6]. We regarded the 10% of genes with highest CAI values as highly expressed (variation of this threshold resulted in no significant changes; **Supplementary Fig. 1** and **Supplementary Table 1** online). Notably, the frequency of essential genes in the leading strand of *B. subtilis* (94%) is higher than the frequency of highly expressed genes (78%). Classifying the genes into four categories according to expressiveness and essentiality, we discovered that essentiality is the primary determinant of gene-strand bias (**Fig. 1a** and **Table 1**). Furthermore, when essentiality was taken into account, gene-strand bias was independent of the gene expression level (**Table 1**). We tested whether this bias could be a simple consequence of related genes clustering in operons. Because an exhaustive list of *B. subtilis* operons was unavailable, we built one using information on gene orientation and rho-independent transcription terminators[7] (**Supplementary Table 2** online). Among such putative operons containing at least one essential gene, 92% are in the leading strand (compared with 65% for operons containing at least one non-essential highly expressed gene). This suggests an operon needs to contain only one essential gene to be preferentially positioned in the leading strand.

We then extended our analysis to the PEC database of *E. coli* essential genes, which classifies 60% of the genes according to essentiality using bibliographic information. Less accurate but still reliable, this information led to the same conclusion: the distribution of essential genes in the chromosome is highly biased, whereas the contribution of expression levels, when controlled for essentiality, is not significant (**Table 1** and **Fig. 1b**). The availability of gene expression data in *E. coli* substantiated the choice of CAI as a valid index of high expression. Of the 97 most highly expressed genes[8], the leading strand contains 62% of non-essential and 97% of essential genes. Furthermore, of the 10% of genes most expressed in rich medium[9], 64% of the 216 non-essential and 90% of the 85 essential genes are located in the leading strand. The latter data set corresponds to fast growth conditions. This confirms that essentiality, not expression level, is the basis of gene-strand bias.

If essentiality drives gene-strand bias and gene-strand bias is caused by collisions between polymerases, then the deleterious nature of collisions depends on the function of genes being transcribed, not on the rate of collisions. Some expression is still required to explain gene-strand bias through DNAP and RNAP collisions, but essential genes are, by definition, expressed in replicating bacteria. Given our results, the emphasis of the model must then shift from the rate of expression, which would distinguish highly expressed genes, to the effect of such collisions on the transcript, which would distinguish according to gene

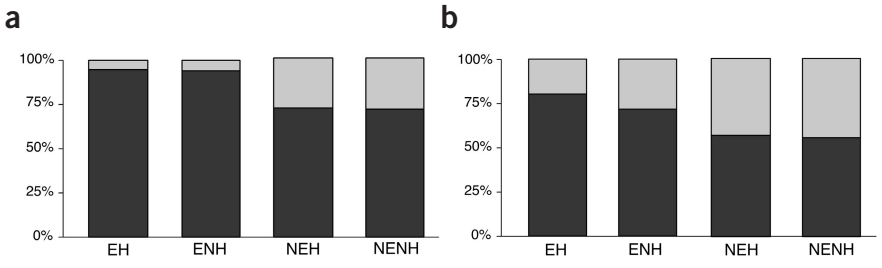

Figure 1 Distribution of genes between the leading (dark gray) and the lagging (light gray) strands of the genome of *B. subtilis* (**a**) and *E. coli* (**b**). H, highly expressed; NH, non-highly expressed; E, essential; NE, non-essential.

[1]Unité GGB, URA 2171, Institut Pasteur, 28 rue Dr. Roux, 75015 Paris, France. [2]Atelier de BioInformatique, Université Pierre et Marie Curie, 12 rue Cuvier, 75005 Paris, France. [3]HKU-Pasteur Research Centre, Dexter HC Man Building, 8, Sassoon Road, Pokfulam, Hong Kong. Correspondence should be addressed to E.P.C.R. (erocha@pasteur.fr).

Table 1 Number of genes and statistical tests

	Number of genes		Percentage of genes lead				Statistical tests		
Genome	E	NE	EH	ENH	NEH	NENH	E > NE	NEH > NENH	ENH > NEH
B. subtilis	277	3,813	96%	93%	71%	72%	$P < 0.001$	NS	$P < 0.001$
E. coli	206	2,204	80%	71%	57%	56%	$P < 0.001$	NS	$P < 0.01$

Genes were classed into two dichotomous classes: H, highly expressed, and NH, non-highly expressed; E, essential, and NE, non-essential. As an example, NEH > NENH tests the hypothesis that among non-essential genes (NE), the highly expressed genes (H) are more biased than the others (NH). Tests are regarded as non-significant (NS) for $P > 0.1$.

function. This cannot be explained solely by the differential availability of the corresponding essential proteins or RNAs. Even if all head-on clashes between polymerases resulted in transcription abortion, this would still result in a small reduction of transcript availability for protein synthesis. For example, in fast-replicating *E. coli*, 70 RNAPs simultaneously transcribe each of the seven rRNA operons, so there is potential for 490 collisions. This represents only ~2% of the total number of rRNAs per cell[10]. If a large number of head-on collisions lead to aborted transcripts, a substantial proportion of these will translate into truncated proteins. This can happen either by saturation of tmRNAs, which rescue stalled ribosomes, or because ribosome drop-off prevents tmRNAs from acting. Truncated peptides are usually non-functional, but if they belong to multi-subunit complexes, they often produce dominant-negative phenotypes[11]. This is deleterious because inactive complexes of essential proteins are a waste of resources and disrupt essential functions required for the cell's organization. In the case of genes encoding components of the ribosome, DNAP or RNAP, this may even lead to error catastrophe.

These results identify an unexpected role for essentiality in the organization of the bacterial chromosome. Rearrangements resulting in strand switch of essential genes may explain the existence of chromosomal inversions that lead to lethality without apparently disrupting important genes[12,13]. How genomes lacking gene-strand bias deal with these problems is not known. The phage T4, which lacks well defined leading strands, has evolved mechanisms to solve head-on collisions[14]. In *Saccharomyces cerevisiae*, head-on collisions are also a problem[15], but eukaryotes have different DNAPs and multiple facultative origins of replication, complicating the *in silico* analysis of this problem.

URLs. The PEC database of *E. coli* essential genes can be found at http://www.shigen.nig.ac.jp/ecoli/pec/. The list of *B. subtilis* essential genes can be found at http://bacillus.genome.ad.jp.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank A. Blanchard, F. Dionísio and F. Taddei for comments and criticisms on the manuscript and the scientists who disrupted and studied all the genes of the *B. subtilis* genome, in particular those who made the enterprise happen: S. D. Ehrlich, F. Kunst, N. Ogasawara and H. Yoshikawa.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 14 April; accepted 9 June 2003
Published online 6 July 2003; doi:10.1038/ng1209

1. Brewer, B. *Cell* **53**, 679–686 (1988).
2. French, S. *Science* **258**, 1362–1365 (1992).
3. McLean, M.J., Wolfe, K.H. & Devine, K.M. *J. Mol. Evol.* **47**, 691–696 (1998).
4. Rocha, E.P.C. *Trends Microbiol.* **10**, 393–396 (2002).
5. Kobayashi, K. *et al. Proc. Natl. Acad. Sci. USA* **100**, 4678–4683 (2003).
6. Sharp, P.M. & Li, W.H. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
7. Kunst, F. *et al. Nature* **390**, 249–256 (1997).
8. Wei, Y. *et al. J. Bacteriol.* **183**, 545–556 (2001).
9. Tao, H., Bausch, C., Richmond, C., Blattner, F.R. & Conway, T. *J. Bacteriol.* **181**, 6425–6440 (1999).
10. Bremer, H. & Dennis, P.P. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* (eds. Neidhardt, F. *et al.*) 1553–1569 (ASM Press, Washington DC, 1996).
11. Pakula, A.A. & Sauer, R.T. *Annu. Rev. Genet.* **23**, 289–310 (1989).
12. Louarn, J.M., Bouche, J.P., Legendre, F., Louarn, J. & Patte, J. *Mol. Gen. Genet.* **201**, 467–476 (1985).
13. Segall, A., Mahan, M.J. & Roth, J.R. *Science* **241**, 1314–1318 (1988).
14. Liu, B. & Alberts, B.M. *Science* **267**, 1131–1137 (1995).
15. Deshpande, A.M. & Newlon, C.S. *Science* **272**, 1030–1033 (1996).