1732–1738 *Nucleic Acids Research, 2011, Vol. 39, No. 5* *Published online 4 November 2010*
doi:10.1093/nar/gkq1067

The pseudogenes of *Mycobacterium leprae* reveal the functional relevance of gene order within operons

Enrique M. Muro[1,*], Nancy Mah[1], Gabriel Moreno-Hagelsieb[2] and Miguel A. Andrade-Navarro[1]

[1]Computational Biology and Data Mining Group, Max Delbrück Center for Molecular Medicine, Robert-Rössle Strasse 10, 13125, Berlin, Germany and [2]Department of Biology, Wilfrid Laurier University. Waterloo, Ontario, Canada

Received March 26, 2010; Revised October 13, 2010; Accepted October 14, 2010

ABSTRACT

Almost 50 years following the discovery of the prokaryotic operon, the functional relevance of gene order within operons remains unclear. In this work, we take advantage of the eroded genome of *Mycobacterium leprae* to add evidence supporting the notion that functionally less important genes have a tendency to be located at the end of its operons. *M. leprae*'s genome includes 1133 pseudogenes and 1614 protein-coding genes and can be compared with the close genome of *M. tuberculosis*. Assuming *M. leprae*'s pseudogenes to represent dispensable genes, we have studied the position of these pseudogenes in the operons of *M. leprae* and of their orthologs in *M. tuberculosis*. We observed that both tend to be located in the 3′ (downstream) half of the operon (*P*-values of 0.03 and 0.18, respectively). Analysis of pseudogenes in all available prokaryotic genomes confirms this trend (*P*-value of 7.1×10^{-7}). In a complementary analysis, we found a significant tendency for essential genes to be located at the 5′ (upstream) half of the operon (*P*-value of 0.006). Our work provides an indication that, in prokarya, functionally less important genes have a tendency to be located at the end of operons, while more relevant genes tend to be located toward operon starts.

INTRODUCTION

Operons, consecutive genes in the same strand co-transcribed into a single messenger RNA, were first described in the pioneering work of Jacob *et al.* (1) in the early 1960s. The co-expression of genes in operons has well-known functional and regulatory implications (2). However, while in *Escherichia coli* there is some evidence of co-linearity of genes in lowly expressed operons associated to the temporal order of reactions in metabolic pathways (3), whether there is a more generic functional relevance to the specific gene order within operons is still unclear (4–6).

Here, we present an analysis of gene order within operons using the set of pseudogenes in the *Mycobacterium leprae* genome as markers of dispensable genes. Pseudogenes are a generic feature of many bacterial and archaeal genomes (7–9) and comparison of related species suggests that they can be acquired in a relatively rapid fashion (10).

The genome of *M. leprae* (Ml) currently constitutes the most extreme example known of an eroded genome with 1133 annotated pseudogenes and 1614 protein-coding genes (11). For comparison, the closely related genome of *M. tuberculosis* (Mt) has 3959 protein-coding genes and only six pseudogenes (11,12). A comparative analysis of Ml and Mt suggested that most of the pseudogenes in Ml have degenerated mostly after gene-by-gene inactivation (13,14).

Although the transcription of pseudogenes in Ml has been detected (15,16), the role of transcribed pseudogenes is not clear and, in any case, the majority of annotated pseudogenes are not translated into protein (16,17). Moreover, the presence of simple sequence repeats nearby Ml pseudogenes has been observed (18) and this can be taken as an indication that Mls pseudogenes are being actively removed as they disrupt RNA and DNA stability (19).

Under the assumption that the pseudogenes of Ml point to ancestral genes of relatively low functional importance, we set to study their distribution within the operon

*To whom correspondence should be addressed. Tel: (+49) 30 9406 4227; Fax: (+49) 30 9406 4240; Email: enrique.muro@mdc-berlin.de

structures of Ml. Our results indicate that dispensable genes have a tendency to be located toward the 3′-(downstream) end of the operon. This trend is confirmed by analysis of all the prokaryotic genomes annotated at NCBI with a completely sequenced chromosome. Correspondingly, we found that the genes selected from the 13 most comprehensive studies of gene essentiality have a tendency to be located toward the 5′ (upstream) half of operons. These results add evidence for the existence of gene order within prokarya operons and suggest that such order could be used to predict gene functional value.

MATERIALS AND METHODS

A classical way to predict prokaryotic operons stems from the observation that consecutive co-directionally expressed genes belonging to the same operon generally have a small separation (<60 nt) or even a small overlap (up to 10 nt). Therefore, an intergenic distance between −10 and 60 nt has been used to consider that two adjacent co-directional genes are part of the same operon. This range correctly identifies 75–80% of known transcriptional units in bacterial species with a rate of false positives under 20% (20–22) and was followed here to predict operons.

Genomic data for Ml (*M. leprae* strain TN) were obtained from the Leproma World-Wide Web Server [http://genolist.pasteur.fr/Leproma/; Data Release R3, 20 July 2004; (11)]. The Mt data were based on the *M. tuberculosis H37Rv* genome sequence project and was obtained from genomic annotations from the Tuberculist World-Wide Server (http://genolist.pasteur.fr/TubercuList/; Data Release R11, 1 October 2008).

The binomial probability of a binomial experiment (b) can be computed according to the following equation: $b(x;\ n,\ P) = (n!\ /\ x!(n - x)!) * P^x * (1-P)^{(n-x)}$, where n is the number of trials, x the number of successes and P the probability of success on an individual trial. We used the cumulative binomial probability, that is the sum of the values of b for values of x equal or larger than the observed number of successes.

The genome annotations for the prokaryotic wide analysis were downloaded from NCBI [ftp://ftp.ncbi.nih.gov/genomes/Bacteria/; 1110 prokaryotic genomes; (23)]. We selected the 754 genomes that had a complete sequenced chromosome and annotated pseudogenes. We filtered further our selection by taking only one strain per organism (the one with more annotated pseudogenes), resulting in a total of 510 organisms. Lists of essential genes were obtained from the original publications collected at the database of essential genes [DEG 5.4; (24)].

RESULTS

The Ml genome contains 1614 genes and 1133 annotated pseudogenes. We computed the distances between contiguous co-directional elements in the genome of Ml, either genes or pseudogenes and compared their distributions (Figure 1). As previously shown (21), the distribution of the distances between the 750 pairs of contiguous co-directional genes has a maximum near zero distance (Figure 1), indicating that most contiguous genes possibly belong to the same operon. However, the distribution of distances between the 146 pairs of pseudogenes followed by a gene has its maximum after 60 nt. This suggests that many pseudogenes in Ml are located at the end of operons. In contrast, the distribution of distances when the pseudogene follows the gene (146 pairs) has its maximum before 60 nt. Surprisingly, the maximum of the distribution of distances between co-directional pseudogene pairs is also below 60 nt. Such pseudogene pairs may represent entire operons that have been inactivated (25).

The analysis of these distributions suggests that much of the operon structure that was present in an ancestor of Ml could have survived the subsequent widespread pseudogenization process. Therefore, we computed the operons of Ml using the distances between contiguous co-directional genomic elements (see 'Materials and Methods' section for details), these being either genes or pseudogenes, to allow for the analysis of the position of the Ml pseudogenes within those ancestral operons or the description of ancestral operons now entirely composed of pseudogenes. Applying this approach, we predicted 491 ancestral operons containing two or more elements of which 111 contain both genes and pseudogenes (Table 1 and Supplementary Table S1). For comparison, the same prediction ignoring the pseudogenes resulted in 314 operons.

The evidence indicated above suggested that we should find a number of pseudogenes at the end of the ancestral operons. In order to make a more general analysis, we defined the 5′ (upstream) and 3′ (downstream) halves of

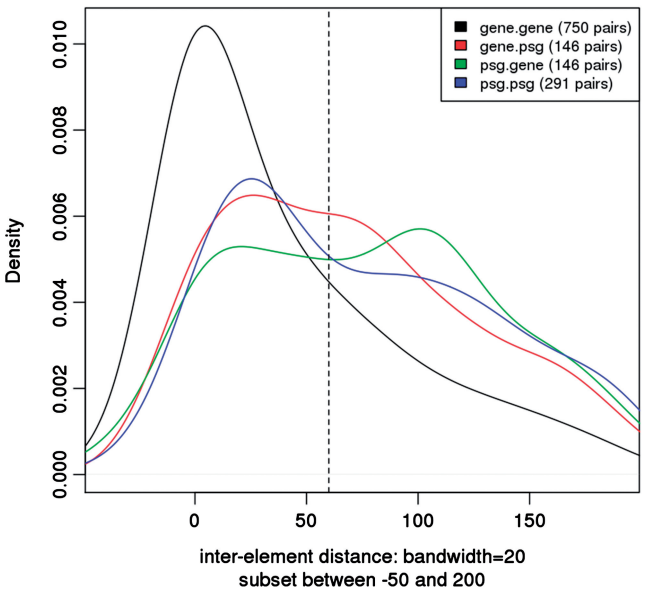

Figure 1. Distribution of distances between different genomic elements. All distributions present maxima under 60 nt (dashed vertical line), except for genes following pseudogenes. This suggests that pseudogenes tend to appear at the last position of operons. The plot lines have been smoothed using a Gaussian filter of bandwidth 20.

Table 1. Properties of predicted operons

	Mycobacterium leprae		*Mycobacterium tuberculosis*
No psg	278	No omp	559
Mixed	111	Mixed	234
Only psgs	102	Only omps	112
Total	491		905

psg, pseudogene; omp, Ortholog of Ml pseudogene

an operon and counted the number of genomic elements (genes or pseudogenes) present in each half (e.g. in an operon of five genes the first two genes are considered to be in the 5′ half and the last two in the 3′ half, with the middle one undefined). We observed the presence of 48 pseudogenes in the 5′ half and 69 in the 3′ half among the 111 predicted operons of *Ml* that contained both genes and pseudogenes. This suggests that the position of a pseudogene within an operon is biased toward the 3′ half. Using as null hypothesis that the number of pseudogenes observed in the 3′ half of operons follows a binomial distribution, we can compute the significance of our observation using a one-tailed binomial test (P-value = 0.03 for observing 69 or more pseudogenes in the 3′ half of an operon in a sample size of $48 + 69 = 117$ pseudogenes) (see 'Materials and Methods' section for details).

In order to test the effect of the operon prediction method on this result, we repeated the analysis using variable thresholds of intergenic distance in the prediction of the operons. The largest value for the fraction of pseudogenes in the 3′ half of the predicted operons was obtained when using intergenic distance values for operon prediction close to the classically used threshold of 60 nt (Figure 2, magenta line). We took this as an indication that our observation is relevant since it is maximally observed for optimal operon predictions (20–22).

Next, we tested the distribution of genes in operons of Ml according to an alternative measure of gene importance. We obtained a list of 482 Ml genes that were orthologs (according to the Leproma database; http://genolist.pasteur.fr/Leproma/) to one of 614 genes essential for Mt growth as obtained by transposon mutagenesis hybridization [essential genes; (26)] (Supplementary Table S2). Ml and Mt are phylogenetically closely related and therefore we expect that the Ml orthologs of Mt essential genes may represent essential genes too. Accordingly, most of the Mt essential genes have a corresponding Ml ortholog that survived the pseudogenization process (Figure 3, left), whereas only a small fraction of the Ml pseudogenes are orthologous to Mt essential genes (Figure 3, right).

If Ml pseudogenes are found in the 3′-regions of operons, we would expect to find the Ml genes orthologous to essential Mt genes in the 5′-region of operons. To study their relative position with a conservative test, we used the set of 314 operons predicted using only intergenic distance between genes (that is excluding pseudogenes from the analysis), since (as shown)

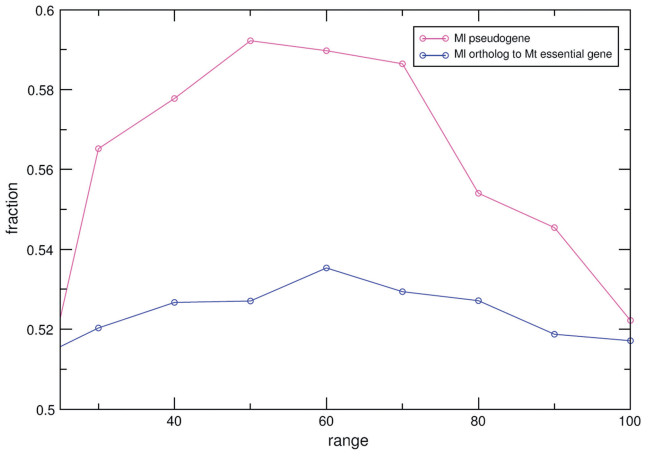

Figure 2. Fraction of genomic elements versus the maximum of the intergenic distance used to predict the operons. Magenta line, fraction of Ml pseudogenes in the 3′ (downstream) half of the operon; operons were defined considering genomic elements being either genes or pseudogenes. Blue line, fraction of Ml orthologs to Mt essential genes in the 5′ (upstream) half of the operon; operons were defined considering genes. Both fractions reach maximal values around 60 nt, which corresponds to the optimal distance for operon prediction.

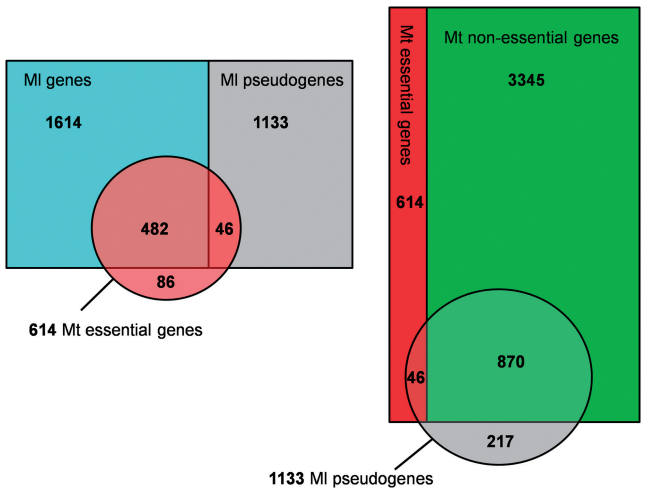

Figure 3. Orthology relationships between genes of *Mt* and genes and pseudogenes of *Ml*. Left: the comparison of 614 functionally important (essential) genes in Mt (26) to the genes and pseudogenes of Ml shows that only a small number of them have been turned into pseudogenes in Ml. Right: conversely, the majority of Ml pseudogenes are orthologs of Mt non-essential genes. These comparisons suggest that pseudogenes in Ml are good indicators of elimination of non-essential genes and that the present orthologs of Mt essential genes in Ml are good indicators of functionally relevant genes.

pseudogenes tend to be positioned in the 3′-region of the operon and this could bias the results. Using these 314 operons, we counted 159 essential genes in the 5′ region and 138 essential genes in the 3′ region, a tendency, however, much less significant than that found for pseudogenes (P-value = 0.12). Again, when testing variable thresholds of intergenic distance in the prediction of the operons, the fraction of essential genes in the 5′-region of

the predicted operons was maximally observed around the optimal 60 nt threshold of intergenic distance (Figure 2, blue line).

We asked whether similar biases could be observed in the related genome of Mt. We did a prediction of operons in Mt using again an intergenic distance between −10 and 60 nt (see 'Materials and Methods' section for details). This resulted in 905 operons of two or more genes (Table 1 and Supplementary Table S3). Then, we obtained the 916 Mt genes that were orthologs to one of 1133 Ml pseudogenes (Figure 3; Supplementary Table S4) and studied their distribution within the 234 predicted operons of Mt that included both orthologs and non-orthologs to Ml pseudogenes. We observed 256 versus 278 orthologs of Ml pseudogenes, in the 5′ and in the 3′ halves of the predicted operons, respectively, suggesting a tendency for these orthologs to be located within the 3′-region (*P*-value = 0.18). An analysis in Mt showed, again like in Ml, a tendency for essential genes to be positioned in the 5′ half of operons (219 versus 204 essential genes; *P*-value = 0.25).

The difference between the significance of the results in Ml and Mt when using Ml pseudogenes or their orthologs in Mt, respectively, may stem from the fact that the former is a more direct observation of the functional importance of genes in the context of the Ml genome.

In order to test the significance of our results in a broader analysis, we repeated the analysis of pseudogenes on all prokaryotic genomes with a complete sequenced chromosome and annotated with pseudogenes at NCBI (23). When considering one strain for each of 510 organisms (see 'Materials and Methods' section and Supplementary Table S5 and File F1), 5362 pseudogenes were located in the 5′ half of operons and 5874 in the 3′ half, showing again a tendency toward the 3′-end of the operon (*P*-value of 7.1×10^{-7}). (Considering all strains resulted in a similar *P*-value of 4.4×10^{-7}).

To similarly expand the scope of the analysis of essential genes presented above for Mt, we collected 13 sets of essential genes from the original studies on bacterial genomes compiled at the database of essential genes [DEG 5.4; (24)] (See Supplementary Table S6 and File S1). A total of 1202 essential genes were located in the 5′ half of operons and 1082 in the 3′ half, showing a tendency toward the 5′ half of the operons (*P*-value 0.006) that corroborates our initial results.

In the next paragraphs, we compare some operons of *Ml* and *Mt*, to illustrate how the properties we used to define gene functional importance apply in particular cases and to illustrate the limitations of our analysis.

Figure 4A illustrates four examples of predicted *Ml* operons that contain pseudogenes in the 3′ half of the operon. In the first example, a long operon containing mostly genes involved in histidine biosynthesis contains a pseudogene for the Mt homolog *impA*, the only gene in the Mt operon not defined as essential (Figure 4A, panel i). Curiously, *impA* is not involved in histidine biosynthesis and examination of the gene order in other bacterial species indicates that the insertion of *impA* in between *hisA* and *hisF* is unique to Corynebacterineae, which includes *Corynebacterium* sp., *Mycobacterium* sp., *Nocardia* sp. *and Rhodococcus* sp. The protein ImpA dephosphorylates inositol-1-phosphate to produce inositol, which is a precursor for intracellular redox reagents in actinomycetes and mycobacterial cell wall components (27,28). Mycobacteria can obtain inositol in two ways (29): (i) by importing it into the cell from the external environment; (ii) by synthesizing it from glucose-6-phosphate. Reliance upon external sources of inositol for survival is unlikely in Ml, since the putative inositol transporter from Mt in Ml has been converted into a pseudogene. Additionally, *de novo* synthesis of inositol is required for Mt pathogenicity (30) and may also be the case in Ml. Although there are no characterized *impA* mutants in Mt or Ml, transposon mutagenesis of *impA* in *M. smegmatis* showed that a mutation producing a defective *impA* affected cell wall permeability but not cell viability (31). The involvement of *impA* in cell wall synthesis, coupled with the possibility that Ml has lost the ability to import external inositol, suggests that Ml is dependent on *de novo* biosynthesis of inositol. In this case, the deficiency of *impA* in Ml could be compensated by another gene, such as ML0662, encoding a probable monophosphatase of the inositol-monophosphatase-like family whose homolog in Mt is essential for growth.

The next two examples represent cases in which changes in operon structure occur at the 3′ half of the Ml operon. In Mt, the operon containing *zwf2* is composed of six genes, of which five have homologs in Ml (Figure 4A, panel ii). This operon is involved in the pentose phosphate pathway, which produces 5-carbon sugars required for purine, pyrimidine and histidine metabolism. Coincidentally, *zwf2* is non-essential for growth in Mt and is a pseudogene in Ml. The introduction of a pseudogene also introduces a gap large enough to split the operon into two separate operons. The absence of both zwf isoforms in Ml (*zwf1* and *zwf2*) forces it to divert α-D-glucose-6P from glycolysis through alternate routes with more steps toward PRPP, the precursor for purines, pyrimidine and histidine metabolism. The function of the next operon example (Figure 4A, panel iii) is unknown. In Ml, the third element is a pseudogene and corresponds to a non-essential gene in Mt. The homologous operon in Mt coincidentally ends after the third position due to high overlap (−28) with the following gene.

Unlike the previous three operon examples, the last example (Figure 4A, panel iv) demonstrates a case where the Mt homolog of the Ml pseudogene was indeed essential for *Mt* growth. In contrast to Mt, Ml can bypass pyruvate and pyruvate carboxylase (pca) through an alternative pathway by directing phosphoenolpyruvate to phosphoenolpyruvate carboxylase (EC:4.1.1.31; not found in Mt) to produce oxaloacetate for entry into the TCA cycle (11).

Figure 4B illustrates examples of Ml operons that contain pseudogenes in the 5′ (upstream) half of the operon. The first is a compact operon (Figure 4B, panel i.) involved in amino acid transport. In Ml, the first element is a pseudogene, whereas all three genes in the operon are essential for growth in Mt. In the last two examples, the Ml operons are shorter than their homologous counterparts in Mt. In both cases, the introduction

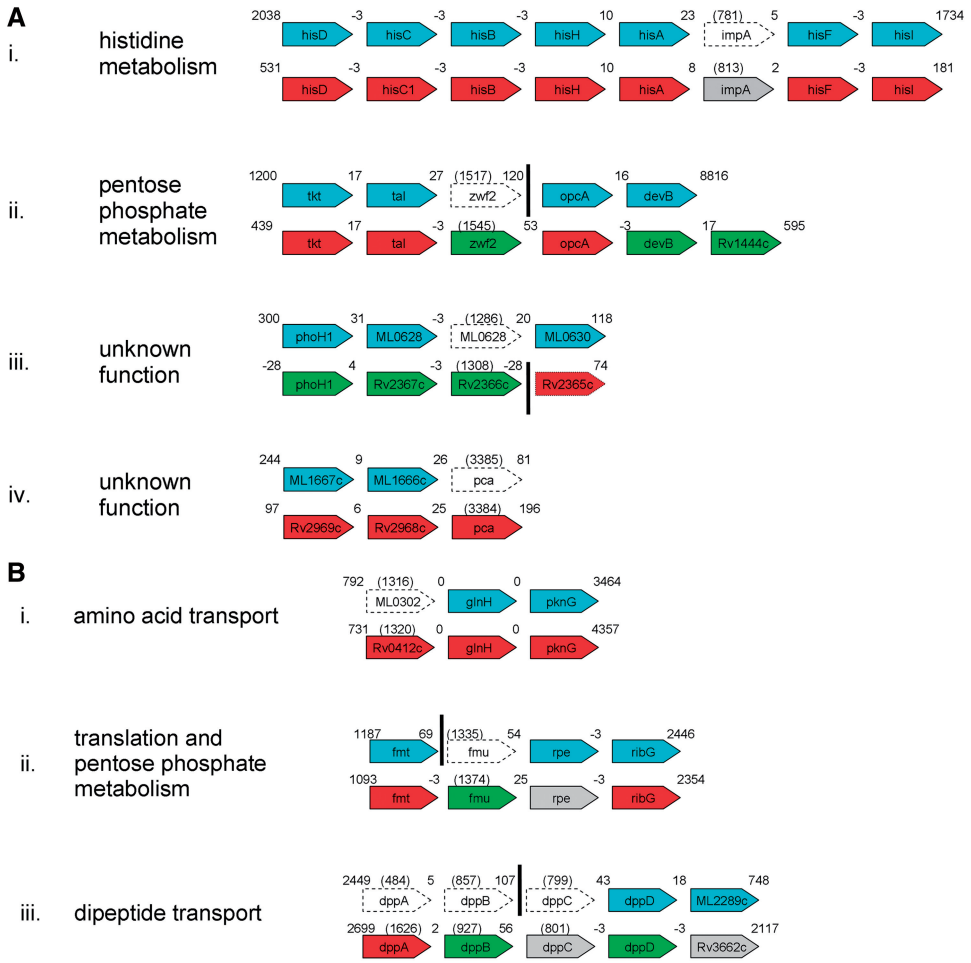

Figure 4. Examples of *Ml* operons containing pseudogenes and the homologous operons in *Mt*. Operons in Ml were defined using an intergenic distance between −10 and 60 nt between either genes or pseudogenes. Top row of each pair represents the Ml genome, with gene elements (cyan) and pseudogenes (white with dashed border). Bottom row represents the homologous Mt operon, colored according to the results of transposon mutagenesis on growth by Sassetti *et al.* (26). Genes were essential for growth (red), non-essential for growth (green) or not determined (gray). A thick vertical bar indicates an operon split. Numbers at the top of each operon string indicate the distance (nt) to the next gene or pseudogene. Numbers in parentheses show the gene/pseudogene size (nt). A general description of the function of the genes in the operon in written on the left side. (A) Four cases of Ml operons with pseudogenes in the 3′ half. (B) Three cases of Ml operons with pseudogenes in the 5′ half.

of a pseudogene in the Ml operon increases the distance between the neighbouring elements, leading to a break in the operon prediction relative to Mt.

DISCUSSION

To demonstrate position-dependent functional importance of genes within operons, we took advantage of the extensive presence of pseudogenes in Ml and used them as markers of dispensable gene function. We found that pseudogenes are biased toward being located in the 3′ half of Ml operons. Furthermore, we showed that essential genes in Mt and their orthologs in Ml have a tendency to be positioned in the 5′ half of the operon. Together, these observations suggest that in Ml and Mt there is selective pressure to locate or rearrange genes so that functionally important genes are situated upstream of less important genes in operons. This bias is maximally observed when using the optimal distance for operon prediction (Figure 2) and is consistently observed across many prokaryotic organisms in both pseudogenes and essential genes. We illustrated with some examples the limitations of the data sets and orthology relationships that we used to define gene importance. Most essential genes in Mt are conserved in Ml and most pseudogenes in Ml correspond to non-essential Mt genes (Figure 3), like in the case of non-essential Mt *impA*, which is a pseudogene in Ml (Figure 4A, panel i). But there are exceptions: pca, essential for Mt, is a pseudogene in Ml (Figure 4A, panel iv). As a tendency, Ml pseudogenes were observed toward the 3′-end of operons. But indeed we observed operons where the first element was a pseudogene (Figure 4B). Only, when pooling the comparisons of the complete genomes of Ml and Mt was it possible to detect a positional tendency of pseudogenes, essential genes and their orthologs. The imperfections of the properties we used to measure gene functional importance were counterbalanced by the possibility of studying them over a large enough set of predicted operons.

Due to the polycistronic nature of the bacterial operon, the physical position of the genes in the operon determines

their order of transcription. Although the existence of alternative promoters internal to operons has been described (32–34), in most cases one would expect that a gene at a given position would be transcribed after and only if the upstream genes in the operon have been already transcribed. A cessation of the sequential transcription of the genes in the operon might happen due to stochastic events, but it might also be intended as a mechanism of control where a particular condition interrupts the transcription of a number of genes. This would explain our observation that functional relevance of Mt/Ml genes has a tendency to decrease in operons from beginning to end: the genes conditionally expressed at the end of the operon tend to be less essential than the genes expressed at the beginning of the operon. Consistent with this, a recent work on *Mycoplasma pneumoniae* shows that almost half of 139 polycistronic identified operons have decaying expression toward the 3′-end of the operon in a staircase-like manner, suggesting a reduced functionality at the end of its operons (35).

We have provided a strategy that can be applied to any complete genome with alternative measurements of gene functional importance. This avenue of work should lead to further insights on gene order within operons, which would guide functional predictions and experimental target prioritization.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Francisco Alonso-Sanchez (CSIC, Madrid, Spain) for helpful discussions and the anonymous referees for their comments.

FUNDING

Helmholtz Alliance in Systems Biology for the Max-Delbrück Center Systems Biology Network; MedSys initiative of the Bundesministerium für Bildung und Forschung (Germany). Funding for open access charge: Helmholtz Alliance in Systems Biology.

Conflict of interest statement. None declared.

REFERENCES

1. Jacob,F., Perrin,D., Sanchez,C. and Monod,J. (1960) [Operon: a group of genes with the expression coordinated by an operator.]. *C. R. Hebd. Seances Acad. Sci.*, **250**, 1727–1729.
2. Rocha,E.P. (2008) The organization of the bacterial genome. *Annu. Rev. Genet.*, **42**, 211–233.
3. Kovacs,K., Hurst,L.D. and Papp,B. (2009) Stochasticity in protein levels drives colinearity of gene order in metabolic operons of Escherichia coli. *PLoS Biol.*, **7**, e1000115.
4. Huynen,M., Snel,B., Lathe,W. 3rd and Bork,P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.
5. Tamames,J. (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol.*, **2**, RESEARCH0020.
6. Poyatos,J.F. and Hurst,L.D. (2007) The determinants of gene order conservation in yeasts. *Genome Biol.*, **8**, R233.
7. Liu,Y., Harrison,P.M., Kunin,V. and Gerstein,M. (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.*, **5**, R64.
8. Lerat,E. and Ochman,H. (2004) Psi-Phi: exploring the outer limits of bacterial pseudogenes. *Genome Res.*, **14**, 2273–2278.
9. van Passel,M.W., Smillie,C.S. and Ochman,H. (2007) Gene decay in archaea. *Archaea*, **2**, 137–143.
10. Lerat,E. and Ochman,H. (2005) Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res.*, **33**, 3125–3132.
11. Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honore,N., Garnier,T., Churcher,C., Harris,D. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature*, **409**, 1007–1011.
12. 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.
13. Dagan,T., Blekhman,R. and Graur,D. (2006) The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. *Mol. Biol. Evol.*, **23**, 310–316.
14. Gomez-Valero,L., Rocha,E.P., Latorre,A. and Silva,F.J. (2007) Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction. *Genome Res.*, **17**, 1178–1185.
15. Akama,T., Suzuki,K., Tanigawa,K., Kawashima,A., Wu,H., Nakata,N., Osana,Y., Sakakibara,Y. and Ishii,N. (2009) Whole-genome tiling array analysis of Mycobacterium leprae RNA reveals high expression of pseudogenes and noncoding regions. *J. Bacteriol.*, **191**, 3321–3327.
16. Williams,D.L., Slayden,R.A., Amin,A., Martinez,A.N., Pittman,T.L., Mira,A., Mitra,A., Nagaraja,V., Morrison,N.E., Moraes,M. *et al.* (2009) Implications of high level pseudogene transcription in Mycobacterium leprae. *BMC Genomics*, **10**, 397.
17. de Souza,G.A., Softeland,T., Koehler,C.J., Thiede,B. and Wiker,H.G. (2009) Validating divergent ORF annotation of the Mycobacterium leprae genome through a full translation data set and peptide identification by tandem mass spectrometry. *Proteomics*, **9**, 3233–3243.
18. Guo,X. and Mrazek,J. (2008) Long simple sequence repeats in host-adapted pathogens localize near genes encoding antigens, housekeeping genes, and pseudogenes. *J. Mol. Evol.*, **67**, 497–509.
19. van Passel,M.W. and Ochman,H. (2007) Selection on the genic location of disruptive elements. *Trends Genet.*, **23**, 601–604.
20. Janga,S.C., Lamboy,W.F., Huerta,A.M. and Moreno-Hagelsieb,G. (2006) The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic Acids Res.*, **34**, 3980–3987.
21. Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18(Suppl. 1)**, S329–S336.
22. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
23. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
24. Zhang,R. and Lin,Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.*, **37**, D455–D458.
25. Madan Babu,M. (2003) Did the loss of sigma factors initiate pseudogene accumulation in M. leprae? *Trends Microbiol.*, **11**, 59–61.
26. Sassetti,C.M., Boyd,D.H. and Rubin,E.J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.*, **48**, 77–84.

27. Newton,G.L., Ta,P., Bzymek,K.P. and Fahey,R.C. (2006) Biochemistry of the initial steps of mycothiol biosynthesis. *J. Biol. Chem.*, **281**, 33910–33920.
28. Rawat,M. and Av-Gay,Y. (2007) Mycothiol-dependent proteins in actinomycetes. *FEMS Microbiol. Rev.*, **31**, 278–292.
29. Reynolds,T.B. (2009) Strategies for acquiring the phospholipid metabolite inositol in pathogenic bacteria, fungi and protozoa: making it and taking it. *Microbiology*, **155**, 1386–1396.
30. Movahedzadeh,F., Smith,D.A., Norman,R.A., Dinadayala,P., Murray-Rust,J., Russell,D.G., Kendall,S.L., Rison,S.C., McAlister,M.S., Bancroft,G.J. *et al.* (2004) The Mycobacterium tuberculosis ino1 gene is essential for growth and virulence. *Mol. Microbiol.*, **51**, 1003–1014.
31. Parish,T., Liu,J., Nikaido,H. and Stoker,N.G. (1997) A Mycobacterium smegmatis mutant with a defective inositol monophosphate phosphatase gene homolog has altered cell envelope permeability. *J. Bacteriol.*, **179**, 7827–7833.
32. Bauerle,R.H. and Margolin,P. (1967) Evidence for two sites for initiation of gene expression in the tryptophan operon of Salmonella typhimurium. *J. Mol. Biol.*, **26**, 423–436.
33. Morse,D.E. and Yanofsky,C. (1968) The internal low-efficiency promoter of the tryptophan operon of *Escherichia coli*. *J. Mol. Biol.*, **38**, 447–451.
34. Koide,T., Reiss,D.J., Bare,J.C., Pang,W.L., Facciotti,M.T., Schmid,A.K., Pan,M., Marzolf,B., Van,P.T., Lo,F.Y. *et al.* (2009) Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol.*, **5**, 285.
35. Guell,M., van Noort,V., Yus,E., Chen,W.H., Leigh-Bell,J., Michalodimitrakis,K., Yamada,T., Arumugam,M., Doerks,T., Kuhner,S. *et al.* (2009) Transcriptome complexity in a genome-reduced bacterium. *Science*, **326**, 1268–1271.