

- 19 Bishop, N. *et al.* (2002) Mammalian class E vps proteins recognize ubiquitin and act in the removal of endosomal protein-ubiquitin conjugates. *J. Cell Biol.* 157, 91–101
- 20 Pornillos, O. *et al.* (2002) Structure of the Tsg101 UEV domain in complex with the PTAP motif of the HIV-1 p6 protein. *Nat. Struct. Biol.* 9, 812–817
- 21 Pornillos, O. *et al.* (2002) Structure and functional interactions of the Tsg101 UEV domain. *EMBO J.* 21, 2397–2406
- 22 Zarrinpar, A. and Lim, W.A. (2000) Converging on proline: the mechanism of WW domain peptide recognition. *Nat. Struct. Biol.* 7, 611–613
- 23 Nguyen, J.T. *et al.* (1998) Exploiting the basis of proline recognition by SH3 and WW domains: design of N-substituted inhibitors. *Science* 282, 2088–2092
- 24 Huang, X. *et al.* (2000) Structure of a WW domain containing fragment of dystrophin in complex with beta-dystroglycan. *Nat. Struct. Biol.* 7, 634–638
- 25 Wittekind, M. *et al.* (1997) Solution structure of the Grb2 N-terminal SH3 domain complexed with a ten-residue peptide derived from SOS: direct refinement against NOEs, J-couplings and 1H and 13C chemical shifts. *J. Mol. Biol.* 267, 933–952

0966-842X/03/\$ - see front matter. Published by Elsevier Science Ltd.  
PII: S0966-842X(02)00013-6

## Genome Analysis

# Did the loss of sigma factors initiate pseudogene accumulation in *M. leprae*?

M. Madan Babu

MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK

**Pseudogenes are non-functional regions in the genome that have arisen as a consequence of accumulating mutations that either result in the premature termination of proteins during protein synthesis or the disruption of transcription. There have been various discussions of the origins of pseudogenes and the models for their formation, but there has been little input on how pseudogenes could have accumulated in an organism. In this brief communication, I propose a two-step model for the accretion of pseudogenes in the *Mycobacterium leprae* genome, triggered by the loss of different sets of sigma factors at different time points during the course of evolution.**

*Mycobacterium leprae*, the causative agent of leprosy, has extremely specialized requirements for growth because it has lost many genes as pseudogenes and hence can no longer survive under 'normal' conditions. On infection therefore, *M. leprae* must colonize highly specific niches to survive. When the complete genome sequence of *M. leprae* was first published, Cole *et al.* [1] revealed evidence of massive gene decay. They also demonstrated that this gene loss was not an optimization to reduce genome size, but was in fact indiscriminate as many pathways had been lost completely.

### The loss of sigma factors by *M. leprae*

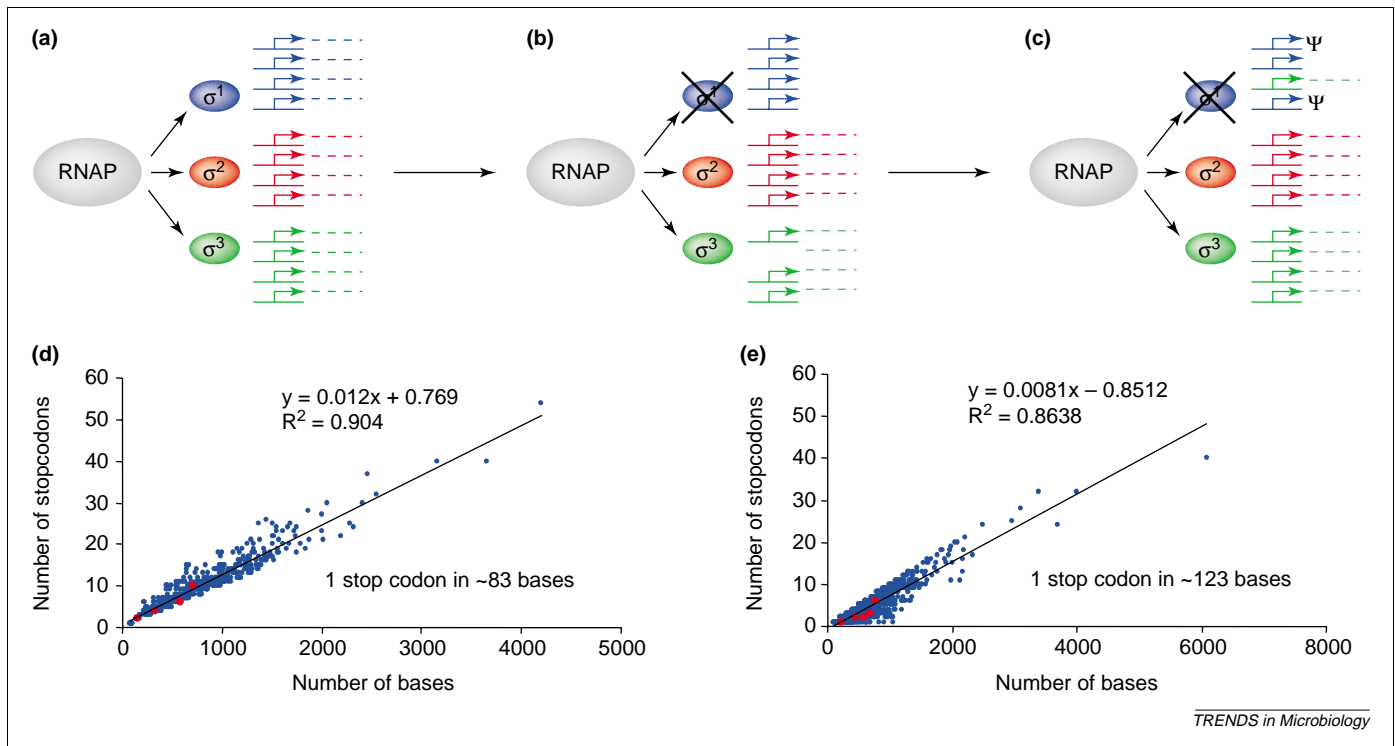
There are 1116 pseudogenes in the genome of *M. leprae*, whereas the genome of its close relative *Mycobacterium tuberculosis* contains only six [2]. The genes lost as pseudogenes in the *M. leprae* genome are quite diverse in function and include nine sigma factors, leaving only four functional sigma factors encoded in the genome [1]. Sigma factors are a class of proteins that bind and confer promoter specificity to the RNA polymerase core enzyme,

hence each sigma factor regulates different sets of genes [3]. Bacterial genomes usually encode one or two principal sigma factors, responsible for the expression of house-keeping genes, and also encode a variable number of alternative sigma factors that respond to specific environmental stimuli. One group of such alternative sigma factors is known as extra-cytoplasmic function (ECF) sigma factors [4]. Assuming there is an equal probability that any gene will be lost as a pseudogene, it appears that *M. leprae* has lost more sigma factors than would be expected by random chance ( $p$  value =  $2.8 \times 10^{-2}$ , obtained by simulating random gene loss of 1116 genes 100 000 times on a computer and calculating the number of instances when at least nine sigma factors were lost as pseudogenes).

### Genes have existed as pseudogenes for different amounts of time

It is a reasonable hypothesis to assume that the rate of stop codon accumulation should always be the same for a given length of gene. If genes have been pseudogenes (i.e. inactivated) for different periods of time, we would expect proteins of similar lengths to now have different numbers of stop codons. When I examined the accumulation of stop codons in the pseudogenes of *M. leprae*, two mutation rates were apparent (Fig. 1). This could be explained if one set of genes was inactivated at a given time first (and hence began accumulating mutations first) and then another set of genes was inactivated at a later point in time (and hence has been accumulating mutations for a shorter length of time than the first set). In Fig. 1, I propose that the formation of pseudogenes in the *M. leprae* genome has not been a continuous process over the course of time, but rather could have been initiated by at least two different events. The evidence for this is based on the best fit of the data to the two sets shown in Fig. 1. If the size of the genome was to be reduced then deletion events would

Corresponding author: M. Madan Babu (madanm@mrc-lmb.cam.ac.uk).



**Fig. 1.** (a) Represents the 'normal' situation, where the RNA polymerase core enzyme (RNAP) associates with different alternative sigma factors (depicted as blue, red and green) to regulate the expression of different set of genes under different environmental conditions or stresses. The dashed lines represent transcribed mRNA. (b) During the course of evolution, when a sigma factor (here  $\sigma^1$ ) is mutated, the organism is unable to express the set of genes that this sigma factor controls. However, it can survive until this environmental condition is experienced. (c) This results in pressure to choose a selective environment, hence forcing the organism to adopt a specialised niche. At this point in time, the organism cannot survive when it encounters the condition (blue). As this set of genes will never be expressed, they are equivalent to any non-coding region in the genome and hence there is no selective pressure for the organism to maintain these genes without accumulating mutations. This can result in a situation where these genes start to mutate, leading to an accumulation of pseudogenes in the genome (represented as  $\psi$ ). Thus, the loss of sigma factors as an early event will lead to accumulation of mutations and pseudogenes in a genome. If a particular protein is essential, selective pressure will allow mutations in the upstream region to incorporate a different recognition site and hence expression by a different sigma factor (shown by the green dashed line). (d) and (e) show two different mutation rates in pseudogenes of similar length, suggesting that a set of sigma factors that was lost first initiated accumulation of mutations (d), followed by the loss of a second set of sigma factors (e), leading to accumulation of mutations in another set of genes. Division of the dataset into two groups produced the best linear fits. The red dots in (d) and (e) represent the sigma factors that are pseudogenes in *Mycobacterium leprae*. For a comprehensive list of pseudogenes, their predicted function, protein length and the number of stop codons, see the supplementary website at [http://www.mrc-lmb.cam.ac.uk/genomes/madann/mlep\\_pseudo/](http://www.mrc-lmb.cam.ac.uk/genomes/madann/mlep_pseudo/).

have removed the unwanted genes [5]. The fact that this has not occurred again suggests that pseudogenes were formed as independent events without selective pressure on the organism to reduce genome size.

### Loss of sigma factors leads to the accumulation of pseudogenes

The fact that the *M. leprae* genome has lost many sigma factors as pseudogenes suggests a possible pathway for pseudogene accumulation, depicted in Fig. 1. The loss of a set of sigma factors is the first event and triggers pseudogene formation. The formation of pseudogenes shuts down the expression of a set of genes that would normally be expressed when the organism experiences a particular stress or different environmental condition. As these genes will now not be expressed when this stress or condition is experienced, the environmental conditions and stresses that the pathogen is able to survive is restricted, thus forcing it to occupy a specific niche suitable for survival. This can be linked to the observation that *M. leprae* colonizes cooler regions of the body such as the skin and ears because of its inability to trigger the synthesis of SigB (a principal sigma factor,  $\sigma^{70}$  class) as SigH, one of the ECF sigma factors involved in the synthesis of SigB, is encoded by a pseudogene. By contrast,

in *M. tuberculosis*, SigH triggers the production of SigB and hence *M. tuberculosis* can grow at higher temperatures than *M. leprae* [6].

Once a suitable niche has been found, the genes controlled by alternative sigma factors are no longer under selective pressure. These genes are now equivalent to any non-coding region in the genome and hence start to accumulate mutations. If a gene was absolutely essential, then mutations in the upstream recognition site would revert its ability to be expressed by allowing a different sigma factor to recognize the upstream region, as shown in Fig. 1.

### Alternative sigma factors and their regulated genes in *M. tuberculosis* are pseudogenes in *M. leprae*

In a very recent study, using quantitative RT-PCR and microarray technology, it was shown that the alternative sigma factor SigH, which responds to heat and oxidative stress and is a pseudogene in *M. leprae*, positively regulates at least 48 genes in *M. tuberculosis* and some of these genes encode proteins involved in cysteine and molybdopterin biosynthesis [7]. Strikingly, in *M. leprae*, these SigH-regulated genes have also been lost as pseudogenes. Similarly, it has been shown by another research group that the alternative sigma factor SigJ, which is also a pseudogene in *M. leprae*, positively

regulates at least 82 genes in *M. tuberculosis* and some of these genes encode proteins involved in nitrogen metabolism during late stationary phase [8]. Interestingly, these SigJ-regulated genes have also been lost as pseudogenes in *M. leprae*. Both SigJ and SigH have higher rates of stop codon accumulation compared with the genes they regulate, as shown in Fig. 2. in the supplementary material online. These observations fit the proposed model very well. As the experimental methods used in this research detect only highly expressed genes, one cannot immediately obtain estimates for the number of genes expressed in low quantities. This is an important factor that must be considered to obtain a correct estimate of the number of genes regulated by these alternative sigma factors. Assuming that each set of alternative sigma factors regulates approximately the same number of genes, one could arrive at a conservative estimate that the alternative sigma factors could regulate ~480–820 genes in all.

### Conclusion

In conclusion, I propose that the loss of a set of sigma factors could have been the triggering step for the accumulation of pseudogenes in the *M. leprae* genome. A further set of sigma factor inactivation events could have occurred at a later point in time, shutting down the expression of another set of genes and forcing the pathogen to adopt a more specialized environmental niche for survival. This set of genes would now start to accumulate mutations. If this scenario occurred, at this point in time the latter subset would have accumulated fewer mutations than the former set for proteins of similar length, suggesting that pseudogene accumulation could have been triggered by at

least two independent events by the loss of sets of sigma factors.

### Acknowledgements

I would like to thank Drs Teichmann, Sankaran and Rogozin for reading the manuscript and the anonymous referees for their very useful comments. I am grateful to the Medical Research Council, Cambridge Commonwealth Trust and Trinity College, Cambridge for financial support.

### References

- 1 Cole, S.T. *et al.* (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011
- 2 Cole, S.T. *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544
- 3 Lewin, B. (1998) *Genes VI*, Oxford University Press
- 4 Missiakas, D. and Raina, S. (1998) The extracytoplasmic function sigma factors: role and regulation. *Mol. Microbiol.* 28, 1066–1069
- 5 Petrov, D.A. *et al.* (2000) Evidence for DNA loss as a determinant of genome size. *Science* 287, 1060–1062
- 6 Eiglmeier, K. *et al.* (2001) The decaying genome of *Mycobacterium leprae*. *Lepr. Rev.* 72, 387–398
- 7 Manganelli, R. *et al.* (2002) Role of the extracytoplasmic-function sigma factor, SigH in *Mycobacterium tuberculosis* global gene expression. *Mol. Microbiol.* 45, 365–374
- 8 Hu, Y. and Coates, R.M. (2001) Increased levels of sigJ mRNA in late stationary phase cultures of *Mycobacterium tuberculosis* detected by DNA array hybridization. *FEMS. Microbiol. Lett.* 202, 59–65

### Supplementary material

Supplementary material and the dataset used for this analysis is available at [http://www.mrc-lmb.cam.ac.uk/genomes/madanm/mlep\\_pseudo/](http://www.mrc-lmb.cam.ac.uk/genomes/madanm/mlep_pseudo/)

0966-842X/03/\$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.  
PII: S0966-842X(02)00031-8

## Genome-specific higher-order background models to improve motif detection

Kathleen Marchal<sup>1</sup>, Gert Thijs<sup>1</sup>, Sigrid De Keersmaecker<sup>2</sup>, Pieter Monsieurs<sup>1</sup>, Bart De Moor<sup>1</sup> and Jos Vanderleyden<sup>2</sup>

<sup>1</sup>ESAT SISTA-SCD, K.U.Leuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium

<sup>2</sup>Centre of Microbial and Plant Genetics, K.U.Leuven, Kasteelpark Arenberg 20, 3001 Leuven-Heverlee, Belgium

Motif detection based on Gibbs sampling is a common procedure used to retrieve regulatory motifs *in silico*. Using a species-specific background model was previously shown to increase the robustness of the algorithm. Here, we demonstrate that selecting a non-species-adapted background model can have an adverse effect on the results of motif detection. The large differences in the average nucleotide composition of prokaryotic sequences exacerbate the problem of

exchanging background models. Therefore, we have developed complex background models for all prokaryotic species with available genome sequences.

DNA motifs are short patterns of DNA. In the promoter regions of genes, motifs constitute the recognition site of transcriptional regulators, and thus reflect the underlying transcriptional networks active at the cellular level. Elucidating such regulatory elements will help to unravel these networks and gain insights into global cellular regulation.

Corresponding author: Kathleen Marchal (Kathleen.Marchal@esat.kuleuven.ac.be).