Interrupted coding sequences in Mycobacterium smegmatis: authentic mutations or sequencing errors?

Submitted by Caroline Deshayes on Thu, 03/30/2017 - 10:38

BACKGROUND: In silico analysis has shown that all bacterial genomes contain a low percentage of ORFs with undetected frameshifts and in-frame stop codons. These interrupted coding sequences (ICDSs) may really be present in the organism or may result from misannotation based on sequencing errors. The reality or otherwise of these sequences has major implications for all subsequent functional characterization steps, including module prediction, comparative genomics and high-throughput proteomic projects.

RESULTS: We show here, using Mycobacterium smegmatis as a model species, that a significant proportion of these ICDSs result from sequencing errors. We used a resequencing procedure and mass spectrometry analysis to determine the nature of a number of ICDSs in this organism. We found that 28 of the 73 ICDSs investigated correspond to sequencing errors.

CONCLUSION: The correction of these errors results in modification of the predicted amino acid sequences of the corresponding proteins and changes in annotation. We suggest that each bacterial ICDS should be investigated individually, to determine its true status and to ensure that the genome sequence is appropriate for comparative genomics analyses.
