PLoS One. 2017 Sep 13;12(9):e0184119. doi: 10.1371/journal.pone.0184119. eCollection 2017.

Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome.

Hücker SM[1,2], Ardern Z[1,2], Goldberg T[3], Schafferhans A[3], Bernhofer M[3], Vestergaard G[4], Nelson CW[5], Schloter M[4], Rost B[3], Scherer S[1,2], Neuhaus K[1,6].

Author information

Abstract

In the past, short protein-coding genes were often disregarded by genome annotation pipelines. Transcriptome sequencing (RNAseq) signals outside of annotated genes have usually been interpreted to indicate either ncRNA or pervasive transcription. Therefore, in addition to the transcriptome, the translatome (RIBOseq) of the enteric pathogen Escherichia coli O157:H7 strain Sakai was determined at two optimal growth conditions and a severe stress condition combining low temperature and high osmotic pressure. All intergenic open reading frames potentially encoding a protein of ≥ 30 amino acids were investigated with regard to coverage by transcription and translation signals and their translatability expressed by the ribosomal coverage value. This led to discovery of 465 unique, putative novel genes not yet annotated in this E. coli strain, which are evenly distributed over both DNA strands of the genome. For 255 of the novel genes, annotated homologs in other bacteria were found, and a machine-learning algorithm, trained on small protein-coding E. coli genes, predicted that 89% of these translated open reading frames represent bona fide genes. The remaining 210 putative novel genes without annotated homologs were compared to the 255 novel genes with homologs and to 250 short annotated genes of this E. coli strain. All three groups turned out to be similar with respect to their translatability distribution, fractions of differentially regulated genes, secondary structure composition, and the distribution of evolutionary constraint, suggesting that both novel groups represent legitimate genes. However, the machine-learning algorithm only recognized a small fraction of the 210 genes without annotated homologs. It is possible that these genes represent a novel group of genes, which have unusual features dissimilar to the genes of the machine-learning algorithm training set.

PMID: 28902868 PMCID:PMC5597208 DOI: 10.1371/journal.pone.0184119