iScience, Volume *25*

Supplemental information

Spotlight on alternative frame coding: Two long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection

Michaela Kreitmeier, Zachary Ardern, Miriam Abele, Christina Ludwig, Siegfried Scherer, and Klaus Neuhaus

Data S1

Alternative start codons of olg1, related to Figure 2

Five alternative start codons (CTG_{291295}, TTG_{291370}, CTG_{291436}, ATG_{291508} & GTG_{291535}) are located within the upstream region of *olg1*, providing the possibility of an N-terminal extension of the coding region of 21 to 261 nt. This hypothesis is supported by the results of RiboSeq (figure 2A, second track) indicating translational signals upstream of the selected start codon. In addition, proven transcription upstream of the predicted start site (figure S3, lane 3 to 6) potentially facilitates the opportunity of a prolongated or alternative version of *olg1*. However, the following aspects argue against an N-terminal extension: Firstly, ATG is the most frequent start codon in *Pseudomonas aeruginosa* (West and Iglewski, 1988) whereas the usage of alternative start codons is rather rare in prokaryotes (Bachvarov et al., 2008). Despite the selected start codon (ATG_{291556}), only one additional ATG is located further upstream at position 291508. For this ATG, a SD and a putative promoter were predicted, but in suboptimal spacing to the start codon and with a lower probability compared to the selected ATG_{291556}. With the exception of GTG_{291535}, all further start codons lacked either a SD sequence or a putative promoter. However, we just tested for the presence of σ^{70} promoters and, therefore, transcription initiation driven by one of the other 23 known promoters (Potvin et al., 2008) might be conceivable. A second and probably the most descriptive aspect indicating that the coding region of *olg1* starts at the selected ATG_{291556} is provided by prediction programs. When deleting all start codons within *olg1,* Prodigal (Hyatt et al., 2010) predicted ATG^{291556} to be the correct start position. Furthermore, DeepRibo (Clauwaert et al., 2019) also predicted translation of *olg1* starting from ATG_{291556} with the highest likelihood. Thirdly, no mass spec peptides were detected when searching against an N-terminal extended Olg1 sequence in the MS data obtained by the DDA proteomics experiment. Absence of peptides, though, does not necessarily correlate with protein presence because of various reasons, including the absence of tryptic cleavage sites (Landry et al., 2015; Slavoff et al., 2013), the efficiency of protein digest and extraction (Baldwin, 2004) or challenging peptide properties like high hydrophobicity (Bagag et al., 2013). Due to contradictory results, further experiments, e.g. frameshift mutagenesis (Smollett et al., 2009) or modified RiboSeq with translation inhibitors like tetracycline (Nakahigashi et al., 2016) or retapamulin (Meydan et al., 2019) are necessary to unveil the correct translation start site of *olg1*.

Bachvarov, B., Kirilov, K., and Ivanov, I. (2008). Codon usage in prokaryotes. Biotechnology & Biotechnological Equipment *22*, 669-682.

Bagag, A., Jault, J.-M., Sidahmed-Adrar, N., Réfrégiers, M., Giuliani, A., and Le Naour, F. (2013). Characterization of hydrophobic peptides in the presence of detergent by photoionization mass spectrometry. PLoS ONE *8*, e79033.

Baldwin, M.A. (2004). Protein identification by mass spectrometry: issues to be considered. Molecular & Cellular Proteomics *3*, 1-9.

Landry, C.R., Zhong, X., Nielly-Thibault, L., and Roucou, X. (2015). Found in translation: functions and evolution of a recently discovered alternative proteome. Current opinion in structural biology *32*, 74-80.

Nakahigashi, K., Takai, Y., Kimura, M., Abe, N., Nakayashiki, T., Shiwa, Y., Yoshikawa, H., Wanner, B.L., Ishihama, Y., and Mori, H. (2016). Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. DNA Research *23*, 193-201.

Potvin, E., Sanschagrin, F., and Levesque, R.C. (2008). Sigma factors in *Pseudomonas aeruginosa*. FEMS microbiology reviews *32*, 38-55.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nature chemical biology *9*, 59-64.

Smollett, K.L., Fivian-Hughes, A.S., Smith, J.E., Chang, A., Rao, T., and Davis, E.O. (2009). Experimental determination of translational start sites resolves uncertainties in genomic open reading frame predictions–application to Mycobacterium tuberculosis. Microbiology *155*, 186.

West, S.E., and Iglewski, B.H. (1988). Codon usage in *Pseudomonas aeruginosa*. Nucleic acids research *16*, 9323-9335.

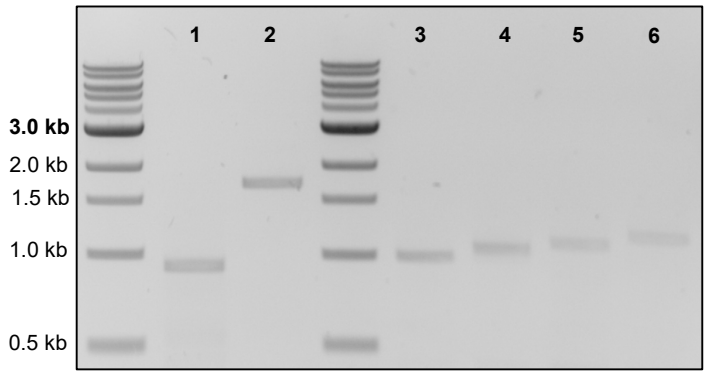

Figure S1. Transcriptional signals of *olg1* and *olg2*, related to Figure 2.
RT-PCR using primers binding at the beginning and at the end of the coding regions of *olg1* (lane 1, target length=917 nt, primer 8 & 9) and *olg2* (lane 2, target length=1696 nt, primer 10 & 11) confirmed transcription throughout the entire ORF length. Primer binding 45 nt (lane 3, target length=995 nt, Primer 15), 110 nt (lane 4, target length=1060 nt, Primer 16), 161 nt (lane 5, target length=1111 nt, Primer 17) and 240 nt (lane 6, target length=1190 nt, Primer 18) upstream of the start codon of *olg1* verified a minimum transcript length of 1190 nt.

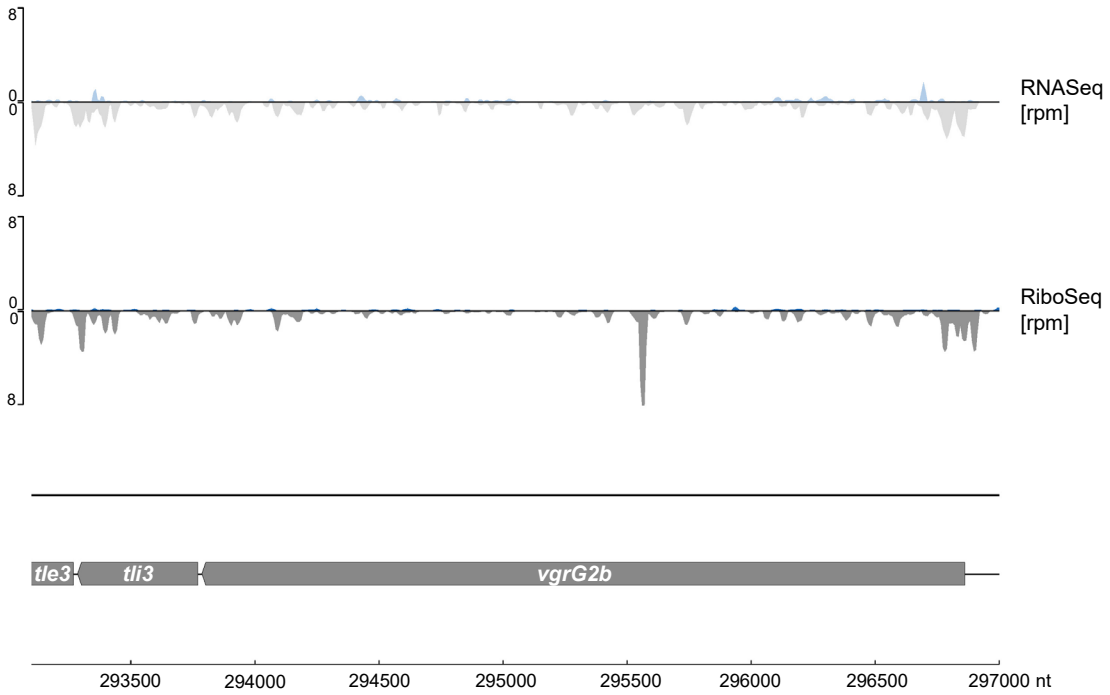

Figure S2. RNASeq and RiboSeq signals upstream of the *olg1-tle3* locus, related to Figure 2.
Shown are the mean normalized rpm values of all transcriptome (first track) and translatome reads (second track) of this study (n=2) for the annotated genes *tli3* and *vgrG2b* (both grey) located upstream of gene *tle3* (grey). Possible background reads antisense to the listed genes are highlighted in blue.

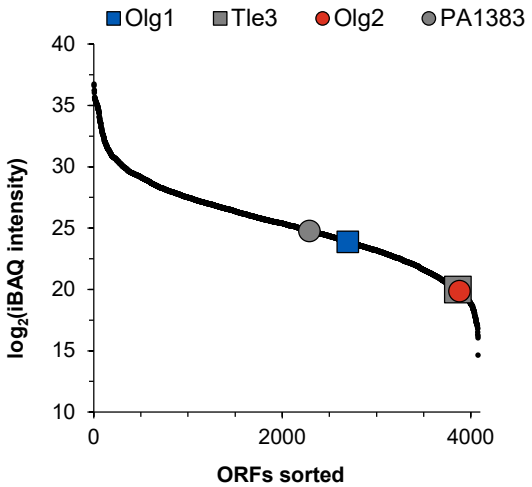

Figure S3. Intensities of all proteins measured by mass spectrometry in descending order, related to Figure 3.
Mass spectrometric intensities (iBAQ[102] values) of all proteins detected in the sample taken at OD_{600nm}=1. The proteins encoded by the OLGs are represented by coloured symbols, those of their mother genes by the respective grey-shaded symbols. Black dots represent the intensities of all quantifiable proteins encoded by annotated genes (n=4,073).

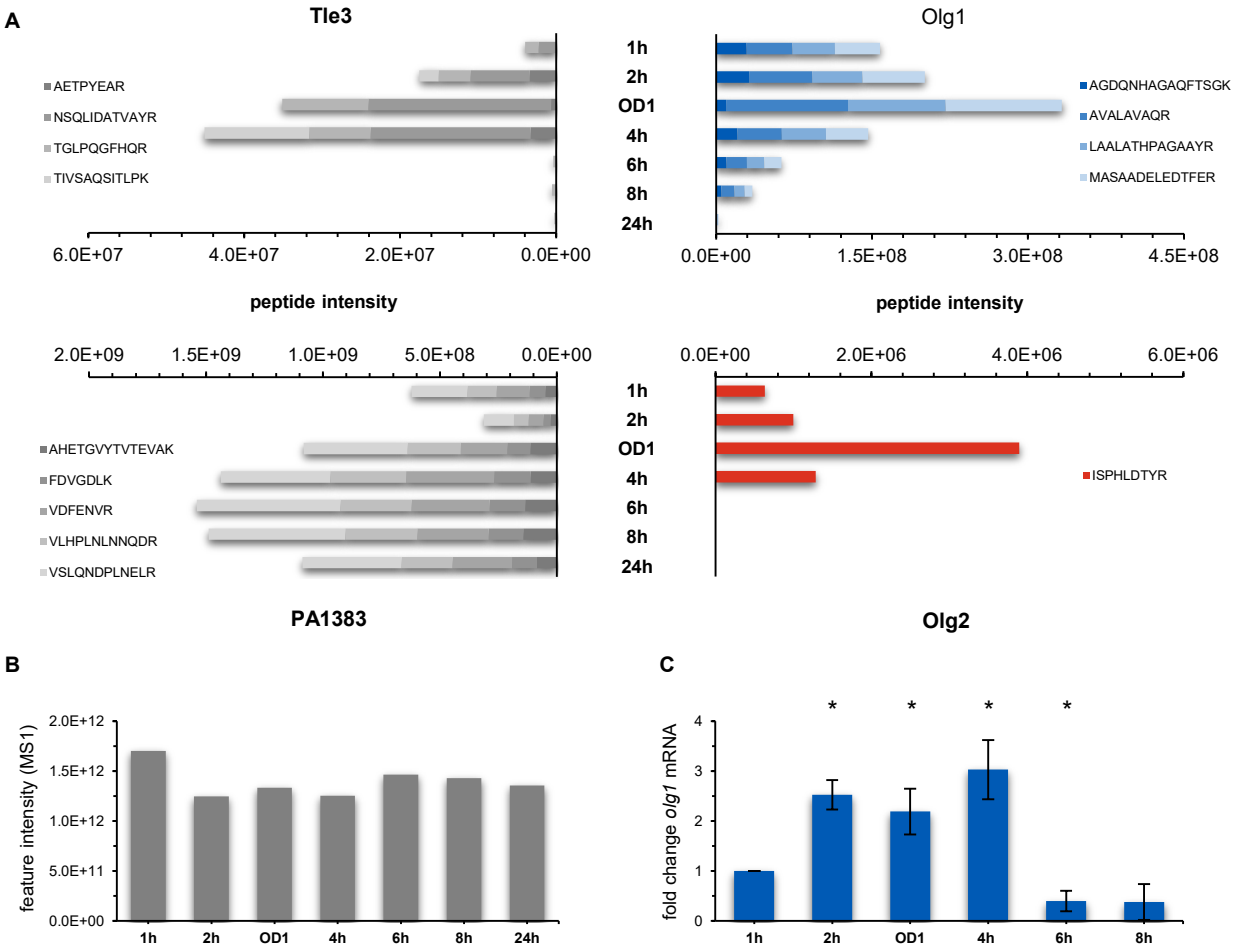

Figure S4. Temporal control of Olg1 and Olg2 expression, related to Figure 3.
(A) Shown are the intensities of all proteins at diverse time points during growth (1h, 2h, 4h, 6h, 8h and 24h as well as at OD_{600nm}=1) measured by targeted proteomics. **(B)** Loading control of the targeted proteomic experiment. Shown are the summed intensities of all measured peptides per sample. **(C)** mRNA levels measured for *olg1* via quantitative PCR. Ct values were normalized to the expression of the housekeeping gene *gyrA*. The values of the sample taken at 1 h served as a reference for fold change calculation. Shown are the mean values of three biological replicates. Comparison between the 1h sample and one of the other samples was performed by using a two-tailed Welch two sample t-test (*p ≤0.05).

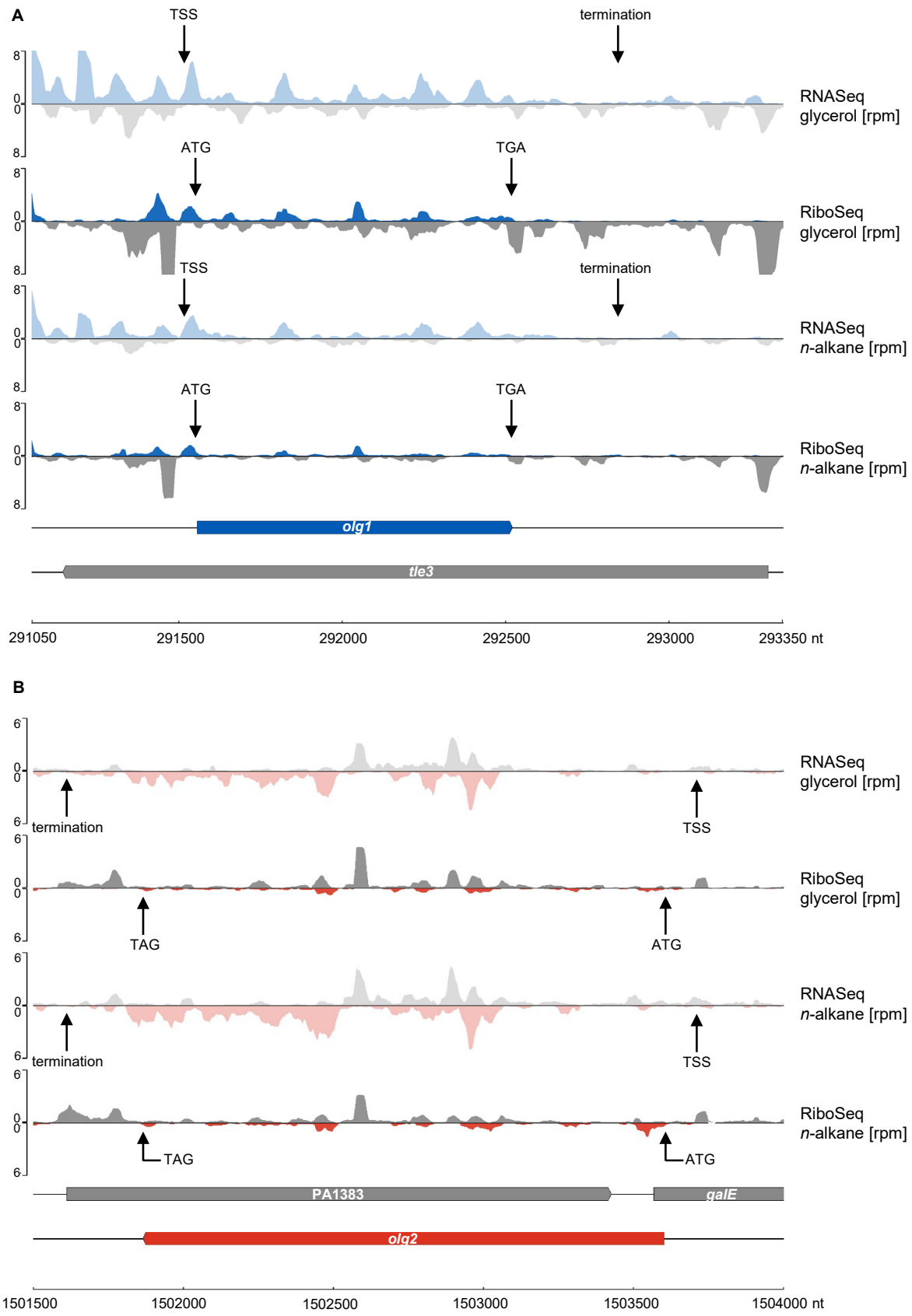

Figure S5. RNASeq and RiboSeq signals at the *olg1-tle3* (A) and *olg2*-PA1383 locus (B), related to Figure 3. Data obtained obtained after reanalysis of the data published by Grady et al. (2017). Shown are the mean normalized rpm values of all transcriptome (first & third track) and translatome reads (second & fourth track) of the datasets "M9+glycerol" and "M9+*n*-alkane" (n=3, each) published by Grady et al.[69] for *olg1* (blue), *olg2* (red) and their mother genes *tle3* and PA1383 (both grey). Transcription start (TSS) and stop sites (termination) as well as the positions of start and stop codons are indicated by arrows.

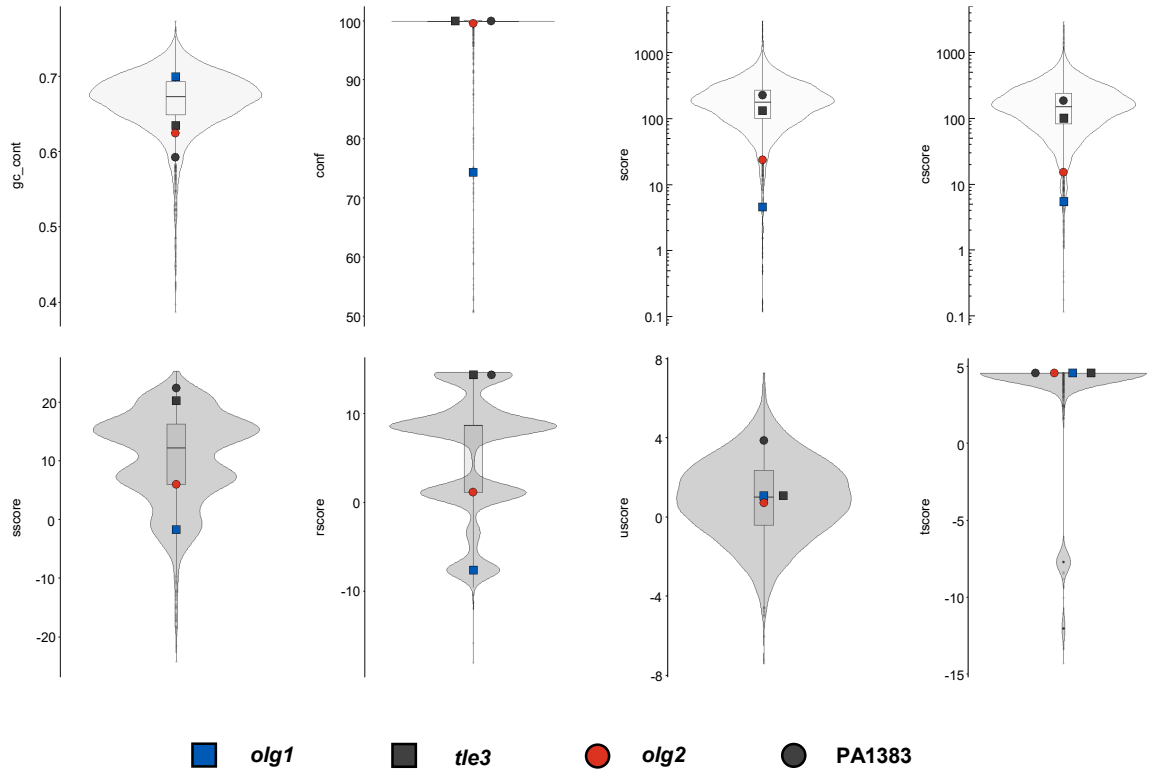

Figure S6. Predicted scores of *olg1* and *olg2* in relation to all protein coding genes obtained by Prodigal (Hyatt et al.[62]), related to Figure 4. Shown are violin plots displaying the GC content of the gene sequence (gc_cont), the confidence score (conf) indicating the likelihood of the gene to be real, the overall score (score), the hexamer coding proportion score (cscore), the translation initiation site score (sscore), the ribosome binding site score (rscore), the score for the region adjacent to the start codon (uscore) and the start codon sequence score (tscore). Included boxplots indicate 25%, 50% and 75% quartile values for all predicted, protein-coding genes (n=5,681). Values of the overlapping ORFs are represented by coloured symbols; their mother genes by the respective grey-shaded symbol.

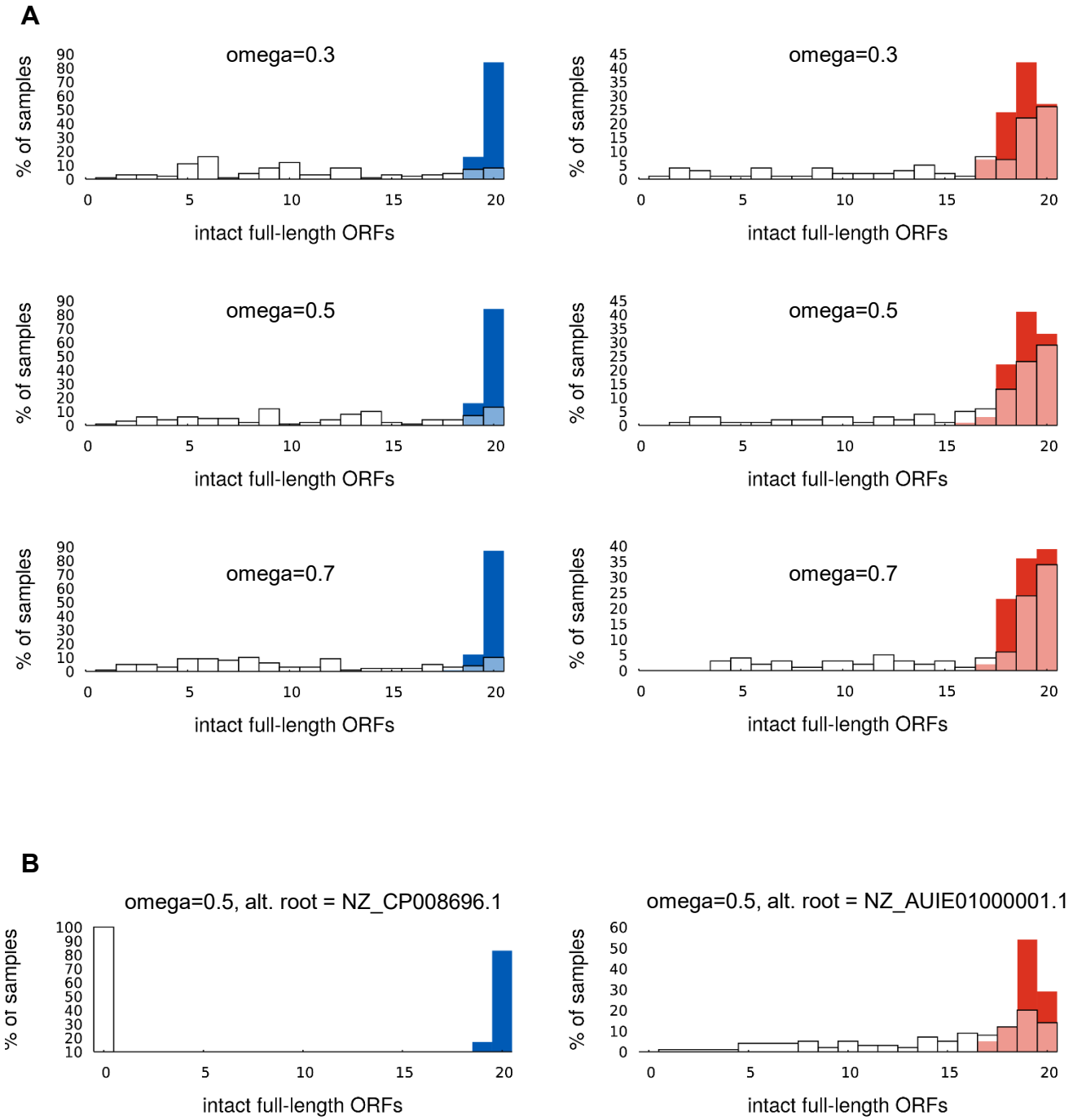

Figure S7. Additional evolutionary simulation analyses, related to Figure 5.
(A) The effect on different values of omega, which is approximately the dN/dS ratio, on the number of observed intact (no premature stop codon) ORFs in evolutionary simulations of the mother genes *tle3* (blue) and PA1383 (red). Simulations were conducted in Pyvolve. This shows only very limited effect of this parameter in these simulations. The plots for omega=0.5 are the same as in Figure 4C. **(B)** Examples of the effect of choosing a different (closer) genome as root for the tree along which evolution is simulated. It is apparent that the choice of root has a much greater effect than the variation in omega above. In these cases evidence of purifying selection against stop codons in the natural sequences is retained.