**Supplementary Material for**

# *De novo*, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus* nematodes

Neel Prabh[1,2], Christian Rödelsperger[1,*]

[1] Department of Integrative Evolutionary Biology,
Max-Planck-Institute for Developmental Biology,
Max-Planck-Ring 9, 72076 Tübingen, Germany

[2] Department of Evolutionary Genetics,
Max-Planck-Institute for Evolutionary Biology,
August Thienemann Str. 2, 24306 Plön, Germany

[*] Author for Correspondence: christian.roedelsperger@tuebingen.mpg.de

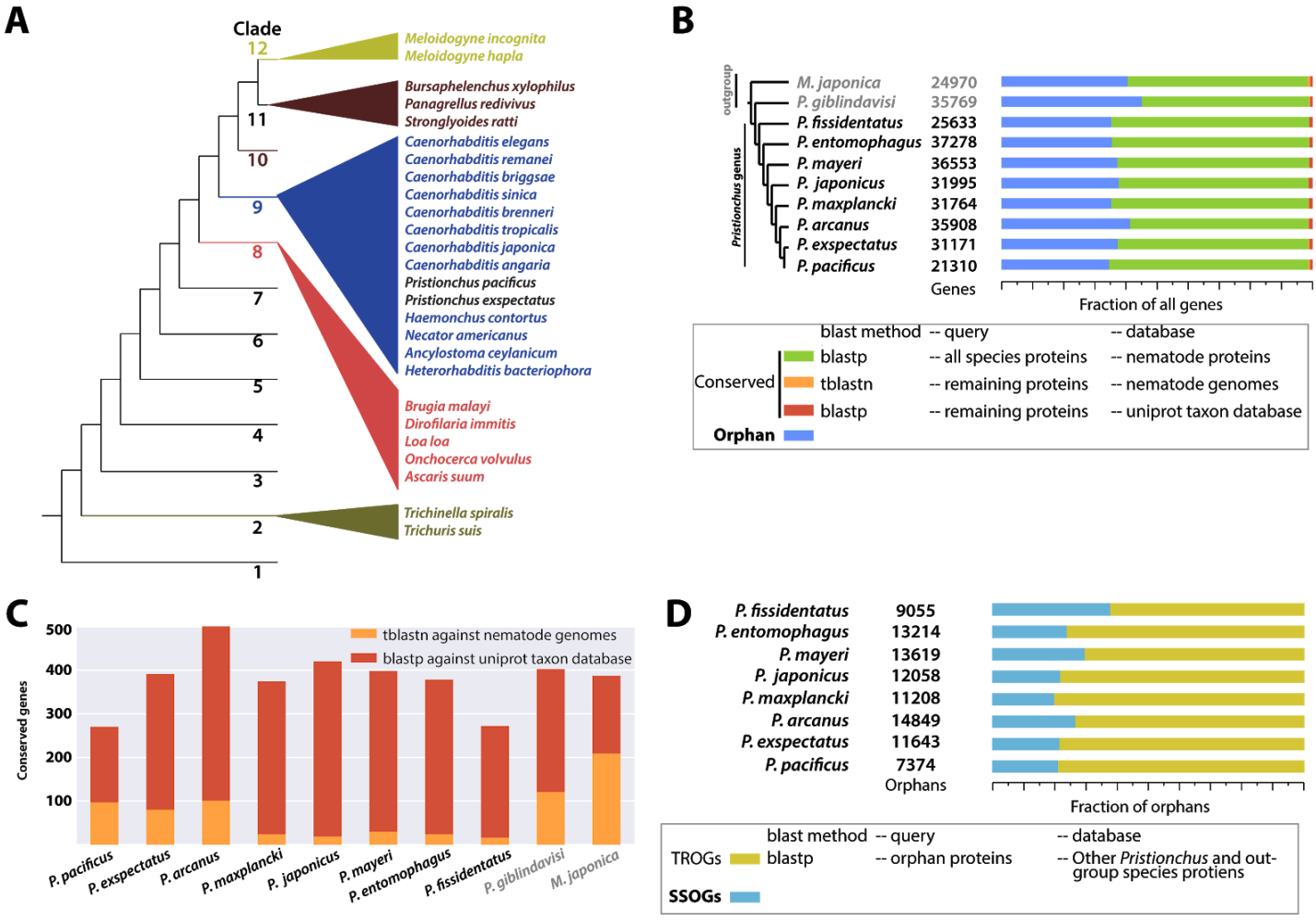


**Figure S1** *Pristionchus* orphan gene identification. (A) Cartoon representing the distribution of nematode species with assembled genomes on Wormbase till 2017. The two *Pristionchus* species are labeled in black. (B) The total number of protein-coding genes for the eight *Pristionchus* species and the two non-*pristionchus* diplogastrid species is shown, followed by the fraction of orphan and conserved genes as horizontally stacked bars. The box shows the different blast methods and databases used to identify the conserved genes in panel a and b. Nematode proteins do not include proteins from the diplogastrid family nematodes (C) Number of conserved genes identified using the additional filtering steps. (D) TROGs and SSOGs as a fraction of orphan genes in each *Pristionchus* species. The box shows the different blast methods and databases used.

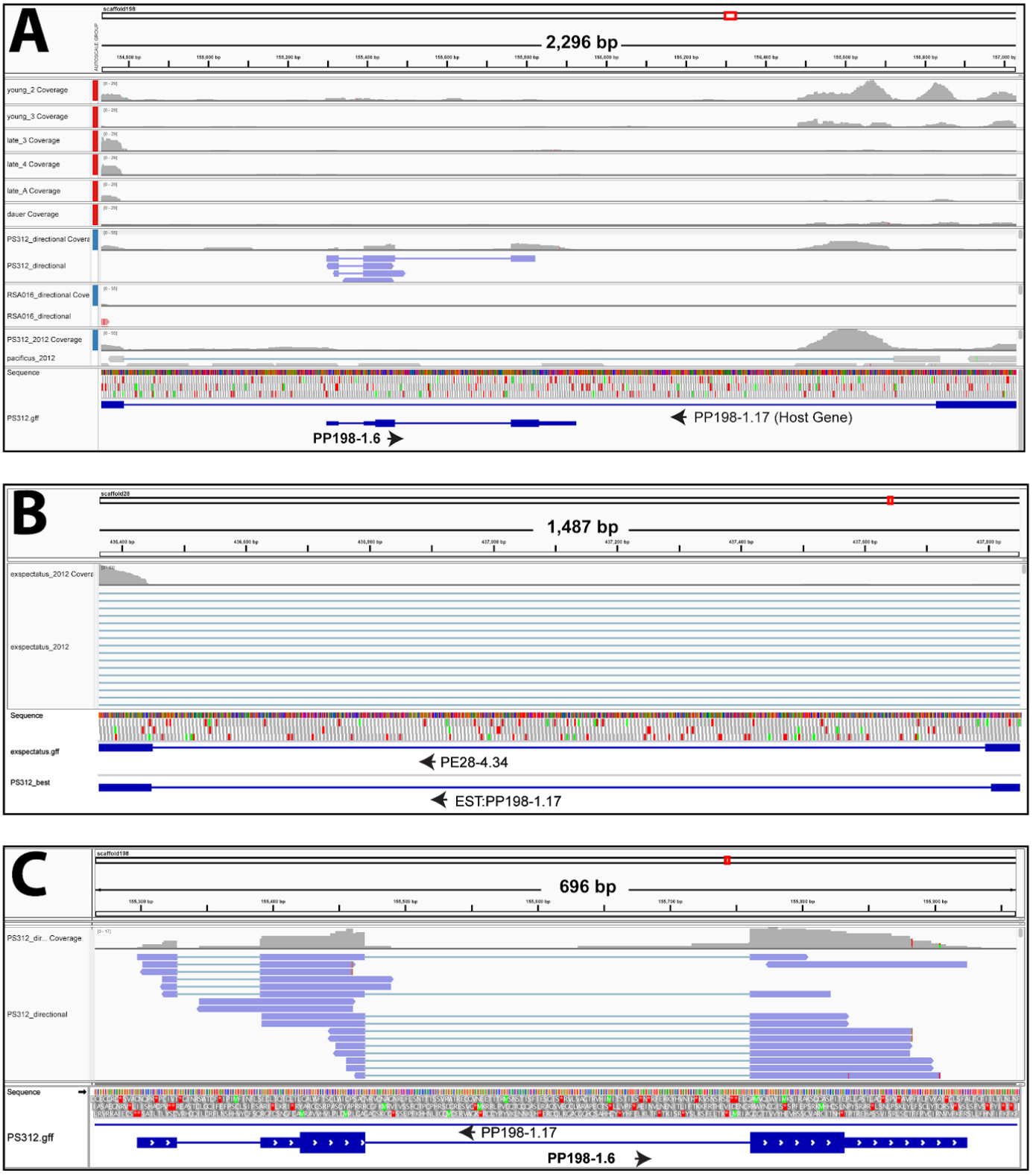


**Figure S2 Novel gene formation by duplication and insertion of exonic sequences into an intron.** (A) This IGV screenshot shows a 2.3kb region on scaffold198 of the *P. pacificus genome*. Different tracks denote gene annotations, coverage profiles and alignments of various RNA-seq samples. The *P. pacificus* candidate SSOG PP198-1.6 is located within the intron of another gene (PP198-1.17, host gene). The same intron also contains a second transcriptionally active region (around position 156,600 bp) which presumably represents a short isoform of the host gene. (B) This screenshot shows the orthologous intron in *P. exspectatus* that was identified by exonerate alignment of the host gene. The genomic span is roughly 800 bp less compared with the *P. pacificus* region suggesting one or multiple insertions of a novel sequences in the *P. pacificus* lineage which gave rise to the candidate SSOG. (C) The genomic span carrying our candidate SSOG is roughly equal to the difference in the intron size between *P. pacificus* and *P. exspectatus*. Alignment of strand-specific raw reads shows that many spliced reads cover the two coding exons in the correct orientation.

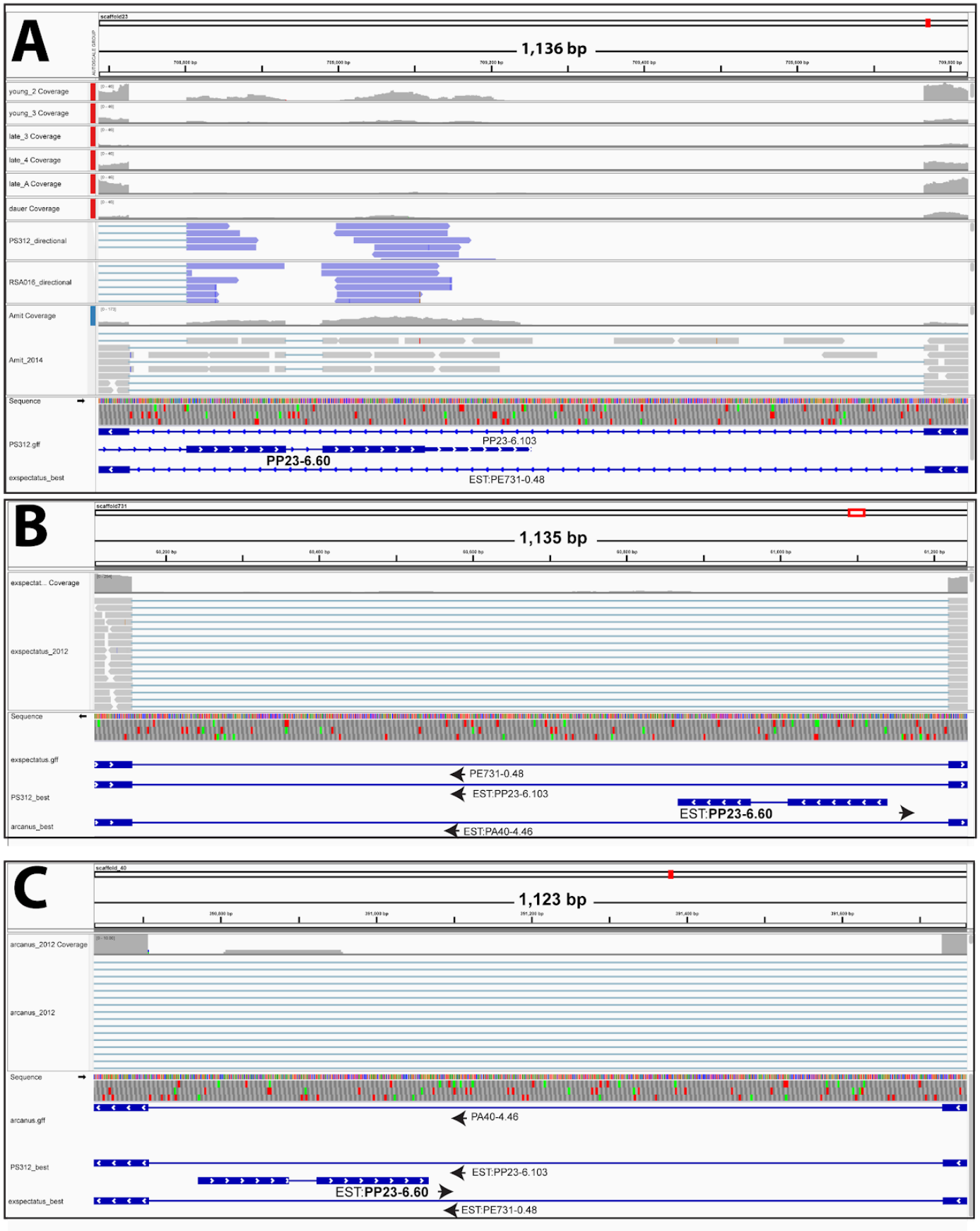


**Figure S3 Intronic *de novo* gene.** (A) This IGV screenshot shows a 1.1kb region on scaffold23 of the *P. pacificus genome* harboring the candidate *de novo* SSOG, PP23-6.60. The candidate SSOG is within the intron of another gene (PP23-6.103, host gene). Strand-specific RNA-seq reads confirm that the gene is predicted in the correct orientation. Raw reads spanning the two coding exons are not found. The ends of spliced reads exceeding the left boundary of the displayed region align to the next intron and form the 5'UTR of the candidate SSOG. (B, C**)** The length of corresponding introns from *P. exspectatus* (B) and *P. arcanus* (C) genomes are comparable with the *P. pacificus* intron. The spliced alignment of our candidate genes onto the genome of sister species allows extraction of corresponding ORFs from these species. Except for a single unspliced read in *P. arcanus*, no transcriptional evidence is found in the two sister species.

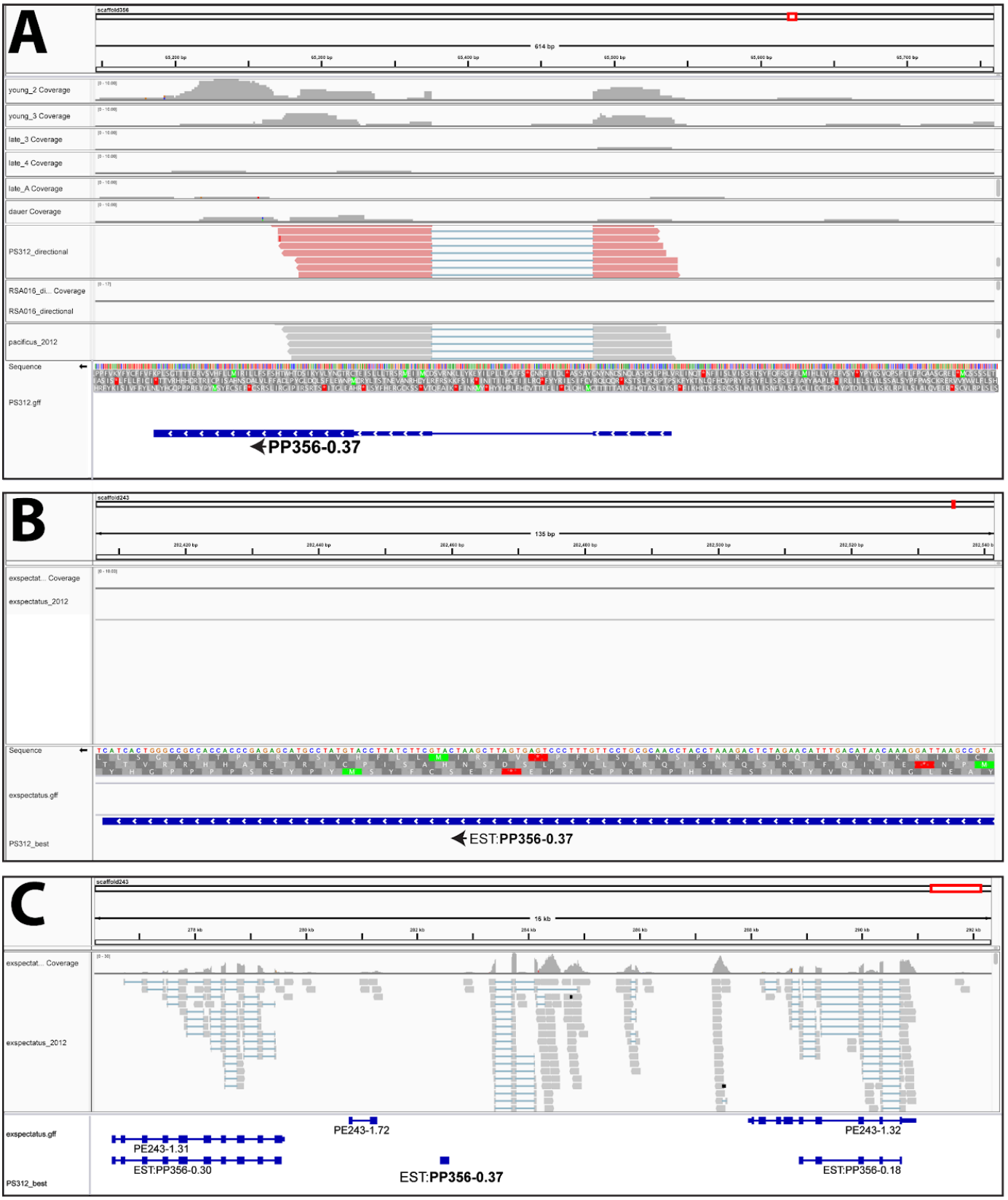


**Figure S4 Intergenic *de novo* gene.** (A) This IGV screenshot shows a 614bp region on scaffold356 of the *P. pacificus* genome harboring the candidate SSOG PP356-0.37. (B) Spliced alignment of our SSOG on the *P. exspectatus* genome shows no ORF exists in the sister species and raw RNA-seq reads do not align at this locus. (C) The neighboring *P. exspectatus* genes are syntenic with other *P. pacificus* genes mapped to the *P. exspectatus* genome and our candidate has emerged within this syntenic block.

**Table S1 Gene origin mechanisms.**

| Gene Id | Divergence | *De novo* | Chimeric | Gene Split | ORF switch | Overprinting | Blast failure | Artifact | Inconclusive |
|---|---|---|---|---|---|---|---|---|---|
| PP142-0.63 | X | | | | | | X | | |
| PP81-0.14 | X | | | | | | X | | |
| PP49-3.6 | X | | | | | X | | | |
| PP130-3.55 | X | | | | | X | | | |
| PP245-0.71 | | | | | | X | | | |
| PP355-0.56 | X | | | | | X | | | |
| PP293-1.0 | | | | | | X | | | |
| PP241-2.35 | | | | | | X | | | |
| PP60-1.28 | | | | | | X | | | |
| PP390-0.42 | X | X | | X | X | | | | |
| PP153-1.8 | X | X | | X | | | | | |
| PP10-2.1 | X | X | | X | | | | | |
| PP9-8.49 | | | | | X | | | | |
| PP48-2.0 | | | | | X | | | | |
| PP198-1.6 | | | | | X | | | | |
| PP378-0.29 | X | | X | | | | | | |
| PP60-1.24 | X | | | | | | | | |
| PP402-0.43 | X | | | | | | | | |
| PP23-6.60 | | X | | | | | | | |
| PP356-0.37 | | X | | | | | | | |
| PP317-0.10 | X | X | | | | | | | |
| PP121-1.22 | | | | | | | | X | |
| PP102-2.17 | | | | | | | | X | |
| PP251-0.100 | | | | | | | | | X |
| PP272-0.50 | | | | | | | | | X |
| PP51-7.48 | | | | | | | | | X |
| PP91-4.32 | | | | | | | | | X |
| PP6-7.26 | | | | | | | | | X |
| PP127-3.37 | | | | | | | | | X |
| **Total** | **12** | **6** | **1** | **3** | **4** | **7** | **2** | **2** | **6** |

The table indicates (marked as X) the proposed mechanisms or reasons behind classification of the 29 high-confidence candidates as SSOGs. 'Divergence' indicates that homology can only be established after synteny analysis. '*De novo*' indicates that at least parts of a gene are of *de novo* origin. 'Overprinting' defined genes that can have more than one overlapping ORFs and in the absence of evidence of translation, we cannot establish which ORFs are real.

**Table S2 Gene identifiers.**

| Abbreviated Gene ID | pristionchus.org<br>Maker Annotation (Prabh et al. 2018) | WormBase / ParaSite<br>(WS269 / WBPS13) |
|---|---|---|
| PP49-3.55 | PS312-ag_msk-S49-3.55-mRNA-1 | PPA35007 |
| PP356-0.37 | PS312-man-S356-0.37-mRNA-1 | PPA35168 |
| PP356-0.18 | PS312-sn_msk-S356-0.18-mRNA-1 | - |
| PP356-0.30 | PS312-mkr-S356-0.30-mRNA-1 | PPA19305 |
| PP142-0.63 | PS312-mkr-S142-0.63-mRNA-1 | PPA40727 |
| PP198-1.6 | PS312-mkr-S198-1.6-mRNA-1 | PPA35042 |
| PP198-1.17 | PS312-mkr-S198-1.17-mRNA-1 | PPA08818 |
| PP378-0.29 | PS312-mkr-S378-0.29-mRNA-1 | - |
| PP23-6.60 | PS312-mkr-S23-6.60-mRNA-1 | PPA38739 |
| PP23-6.103 | PS312-mkr-S23-6.103-mRNA-1 | PPA17250 |
| PP49-3.6 | PS312-mkr-S49-3.6-mRNA-1 | - |
| PP390-0.42 | PS312-mkr-S390-0.42-mRNA-1 | PPA35851 |
| PP81-0.14 | PS312-mkr-S81-0.14-mRNA-1 | - |
| PP241-2.35 | PS312-mkr-S241-2.35-mRNA-1 | - |
| PP51-7.48- | PS312-mkr-S51-7.48-mRNA-1 | - |
| PP245-0.71 | PS312-mkr-S245-0.71-mRNA-1 | - |
| PP9-8.49 | PS312-mkr-S9-8.49-mRNA-1 | PPA17285 |
| PP130-3.55 | PS312-mkr-S130-3.55-mRNA-1 | - |
| PP272-0.50 | PS312-mkr-S272-0.50-mRNA-1 | PPA40848 |
| PP251-0.100 | PS312-mkr-S251-0.100-mRNA-1 | PPA44314 |
| PP60-1.28 | PS312-mkr-S60-1.28-mRNA-1 | - |
| PP91-4.32 | PS312-mkr-S91-4.32-mRNA-1 | PPA43420 |
| PP102-2.17 | PS312-mkr-S102-2.17-mRNA-1 | - |
| PP60-1.24 | PS312-mkr-S60-1.24-mRNA-1 | PPA34695 |
| PP293-1.0 | PS312-mkr-S293-1.0-mRNA-1 | PPA41196 |
| PP121-1.22 | PS312-mkr-S121-1.22-mRNA-1 | PPA35325 |
| PP127-3.37 | PS312-mkr-S127-3.37-mRNA-1 | - |
| PP317-0.10 | PS312-mkr-S317-0.10-mRNA-1 | - |
| PP48-2.0 | PS312-mkr-S48-2.0-mRNA-1 | PPA34605 |
| PP6-7.26 | PS312-mkr-S6-7.26-mRNA-1 | PPA14495 |
| PP10-2.1 | PS312-mkr-S10-2.1-mRNA-1 | - |
| PP153-1.8 | PS312-mkr-S153-1.8-mRNA-1 | - |
| PP402-0.43 | PS312-mkr-S402-0.43-mRNA-1 | - |

| | |
|---|---|
| PE440-0.48 | exspectatus-mkr-S_440-0.48-mRNA-1 |
| PE68-1.70 | exspectatus-mkr-S_68-1.70-mRNA-1 |
| PE158-0.48 | exspectatus-mkr-S_158-0.48-mRNA-1 |
| PE1052-0.1 | exspectatus-mkr-S_1052-0.1-mRNA-1 |
| PE296-0.70 | exspectatus-mkr-S_296-0.70-mRNA-1 |
| PE242-0.104 | exspectatus-mkr-S_242-0.104-mRNA-1 |
| PE28-4.34 | exspectatus-ag_msk-S_28-4.34-mRNA-1 |
| PE731-0.48 | exspectatus-mkr-S_731-0.48-mRNA-1 |
| PE243-1.31 | exspectatus-mkr-S_243-1.31-mRNA-1 |
| PE243-1.72 | exspectatus-sn_msk-S_243-1.72-mRNA-1 |
| PE243-1.32 | exspectatus-mkr-S_243-1.32-mRNA-1 |
| PA40-4.46 | arcanus-ag_msk-S_40-4.46-mRNA-1 |
| PA7-2.29 | arcanus-ag_msk-S_7-2.29-mRNA-1 |
| PA73.-2.42 | arcanus-mkr-S_73-2.42-mRNA-1 |
| PA61-4.37 | arcanus-mkr-S_61-4.37-mRNA-1 |

Gene abbreviations as used throughout the manuscript are shown with their full gene identifiers from pristionchus.org and their corresponding gene models on WormBase (WS269) and WormBase Parasite (WBPS 13). As we have used a different version of the *P. pacificus* genome (Prabh et al. 2018) than what is currently available at WormBase, some *P. pacificus* genes do not have a correspondence on WormBase WS269.