bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

# Large scale and significant expression from pseudogenes in *Sodalis glossinidius* - a facultative bacterial endosymbiont

Ian Goodhead[1,2], Frances Blow[1,3], Philip Brownridge[1], Margaret Hughes[1,4], John Kenny[1,4], Ritesh Krishna[1,5], Lynn McLean[1], Pisut Pongchaikul[1], Rob Beynon[1] and Alistair C. Darby[1,4]

1. Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool. L69 7ZB. UK.
2. School of Environment and Life Sciences, Peel Building, University of Salford. M5 4WT. UK.
3. Department of Entomology, Cornell University, Ithaca 14853, NY, USA.
4. Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool. L69 7ZB. UK.
5. IBM Research UK, STFC Daresbury Laboratory, Warrington, WA4 4AD, UK.

## Abstract

The majority of bacterial genomes have high coding efficiencies, but there are some genomes of intracellular bacteria that have low gene density. The genome of the endosymbiont *Sodalis glossinidius* contains almost 50% pseudogenes containing mutations that putatively silence them at the genomic level. We have applied multiple 'omic strategies, combining: Illumina and Pacific Biosciences Single-Molecule Real Time DNA-sequencing and annotation; stranded RNA-sequencing; and proteome analysis to better understand the transcriptional and translational landscape of *Sodalis* pseudogenes, and potential mechanisms for their control. Between 53% and 74% of the *Sodalis* transcriptome remains active in cell-free culture. Mean sense transcription from Coding Domain Sequences (CDS) is four-times greater than that from pseudogenes. Comparative genomic analysis of six Illumina-sequenced *Sodalis* isolates from different host *Glossina* species shows pseudogenes make up ~40% of the 2,729 genes in the core genome, suggesting that they are stable and/or *Sodalis* is a recent introduction across the *Glossina* genus as a facultative symbiont. These data further shed light on the importance of transcriptional and translational control in deciphering host-microbe interactions. The combination of genomics, transcriptomics and proteomics give a multidimensional perspective for studying prokaryotic genomes with a view to elucidating evolutionary adaptation to novel environmental niches.

## Importance

Bacterial genes are generally 1Kb in length, organized efficiently (i.e. with few gaps between genes or operons), and few open reading frames (ORFs) lack any predicted function. Intracellular bacteria have been removed from extracellular selection pressures acting on pathways of declining importance to fitness and thus, these bacteria tend to delete redundant genes in favour of smaller functional repertoires. In the genomes of endosymbionts with a recent evolutionary relationship with their host, however, this process of genome reduction is not complete; Genes and pathways may be at an intermediate stage, undergoing mutation linked to reduced selection and small population numbers being vertically transmitted from mother to offspring in their hosts, resulting in an increase in abundance of pseudogenes and reduced coding capacities. A greater knowledge of the genomic architecture of persistent pseudogenes, with respect to their DNA structure, mRNA transcription and even putative translation to protein products, will lead to a better understanding of the evolutionary trajectory of endosymbiont genomes, many of

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

which have important roles in arthropod ecology.

## Introduction

The genomes of intracellular parasites and endosymbiotic bacteria evolve under conditions that are fundamentally different to those of free-living organisms (1). In many arthropod systems, bacteria can provide nutrients that are otherwise scarce to their host (such as B-vitamins absent from blood meals, or essential amino acids absent from plant sap), in exchange for host provision of protection, nutrition and mechanisms for vertical or horizontal transmission (2, 3). Obligate intracellular symbionts are maintained by the host and have evolved strategies that ensure their vertical transmission to the next generation of hosts. Ultimately this intracellular lifestyle, small population size and strict vertical transmission can result in extremely reduced genomes (1, 3–6). The general theory and process of this extreme genome reduction has been well studied using genomic data for intracellular bacteria, including endosymbionts such as *Buchnera* in aphids (7, 8), and *Wigglesworthia* in tsetse flies (9). However, gene loss is not limited to obligate intracellular pathogen/symbionts with strict vertical transmission, it is also observed in free-living bacteria and facultative symbionts (10).

One of the most important mechanisms for gene loss is that of pseudogenisation, resulting from the accumulation of nonsense mutations in protein coding sequences (1). These mutations putatively silence the gene at the genomic level resulting in theoretically non-functional genes/proteins (11). Prokaryotic pseudogenes generally exist at levels approximately between 1% and 5% (12). Comparative genomic analysis between closely related strains suggests that pseudogenes are often associated with reduced selective pressure on redundant gene sets allowing mutation to accumulate and inactivate genes. This has been observed as *Salmonella* changes host range or utilizes a new environment (13). The low level of pseudogenes in most bacteria suggests that they are removed rapidly from the genomes due to strong selection for genome efficiency (11). There are however examples among the intracellular pathogens and endosymbionts of high levels of pseudogene presence, reducing coding capacity down towards 50% in *Sodalis glossinidius* (14) and *Mycobacterium leprae* (15). Likewise, pseudogenes can persist for long periods – the mean half-life of *Buchnera aphidocola* pseudogenes has been estimated to be 24 million years (16). Pseudogenes have been well studied in the context of comparative genomics to understand how gene loss has shaped bacterial genomes (17), but whether they continue to contribute to the genetic capabilities of the bacterium has seldom been assessed (18). It could, for instance, be suggested that if pseudogene-derived transcription retains some form of *cis/trans* regulatory function, then this could select for pseudogene retention in the genome (19). It is also clear that under some circumstances, specifically where polymerase infidelity corrects for a frameshift within homopolymeric tracts at the transcriptional level, pseudogenes can still produce functional proteins that contribute to the fitness of the bacterium (20).

In this study we aim to understand the importance of pseudogenes in bacterial genome evolution in a model of a degrading bacterial genome, that of *Sodalis glossinidius*. *Sodalis* is a facultative intracellular, secondary endosymbiont of the tsetse fly (Diptera: *Glossina*). The variable frequency of *Sodalis* in natural populations suggests that *Sodalis* is not an obligatory component of the tsetse microbiome (21), however, the occurrence of *Sodalis* in natural populations has been linked to an increased capacity of tsetse to vector African trypanosomes (22).

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Interestingly, *Sodalis* has a relatively large genome for a facultatively intracellular endosymbiont (~4Mbp) and two genome annotations suggest that pseudogene levels are between 29% (14) and 38% (23) of the total gene content. The *S. glossinidius* genome has amongst its coding repertoire, systems for flagella, transmembrane transport (14), quorum sensing (24), and, of note, type-III secretion systems, encoded by three *Sodalis* symbiosis regions (SSR 1-3), which are analagous to pathogenicity islands (e.g. Salmonella Pathogenicity Islands (SPIs (25)) and have been implicated in establishing or maintaining symbiosis (26). By combining the latest high-throughput sequencing and proteomics methods, we hope to shed light on potential post-transcriptional regulatory mechanisms that may be mitigating any potential deleterious effects. At the RNA level, riboswitches (27), or small RNAs (sRNAs) – short, 50-300bp transcripts mediated by imperfect base pairing interactions, have been shown to regulate genes in this manner (28). DNA methylation could also serve as a mechanism by which to control transcription and/or translation (29, 30).

*Sodalis glossinidius* represents an ideal system in which to test hypotheses surrounding pseudogene functionality and their evolution, as the organism maintains an unusually reduced coding capacity, yet remains amenable to cell culture allowing for sufficient DNA, RNA and peptides to be extracted for poly-omic analyses. First: Assuming genes with nonsense mutations are non-functional and therefore costly to the cell, pseudogenes should be evolving rapidly and be removed from the genome. Secondly: If pseudogene transcription or translation is deleterious, pseudogenes should be transcriptionally and translationally silent. Thirdly: Given hypothesis two, we can expect there to be genetic mechanisms to silence pseudogenes, and we will be able to identify genetic and transcriptional features that determine pseudogene status using a combination of genomic, expression and proteomic analysis. To that end, we tested these hypotheses by: 1. Establishing pseudogene content and evolution using pan-genome data; 2. Evaluating genome-wide methylation data and negative strand expression to elucidate potential expression control mechanisms; and 3. Correlating mRNA and protein expression levels to understand functional control of pseudogenes.

## Results

***Sodalis* Genomics:** To provide an updated, accurate reference for transcriptome mapping of the *S. glossinidius* isolate used in this study, we sequenced *de novo* a *Sodalis glossinidius* (from the host *Glossina morsitans morsitans*) isolate (SgGMMB4) using Pacific Biosciences sequencing. A single SMRTcell produced a total of 48,519 reads, with a mean length of 10,290 bp and mean read score of 0.85. The chromosome was assembled in to a single 4.1 Mbp contig and one copy of the three plasmid sequences pSG1 (90.747 bp); pSG2 (38,394 bp) and pSG3 (10,640 bp). Previously published annotations from the GMM4 isolate (14, 23, 31) were used, alongside a manually curated PROKKA-generated annotation of the PacBio sequenced isolate, to generate a new annotation of our *Sodalis* sequence (32). The overall mean genome GC content is 54.4 %, and the pseudogenes and CDS have a similar GC of ~55.5 %. The revised annotation presented contains 3,336 putative CDS and 2,286 putative pseudogenes. In addition, 43 putative riboswitch domains, belonging to seven families, were identified (Supplementary Table 2). A prodigal (v. 2.6.2) (33), gene prediction including a Ribosome Binding Site (RBS) identification scan suggested a prevalence of standard methionine-coding ATG start codons (78 %), which decreases in the case of pseudogenes (67 %). GTG start-codons are the next most common, found in 14 % of all genes, increasing to 21 % of the total number of pseudogenes);

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

followed by TTG (8 % overall; 12 % of pseudogenes). Sixty-one percent of all genes (48 % of pseudogenes) have an RBS predicted within 5-10 bp of the start codon, and 26 % of all genes (38 % pseudogenes) have no discernible RBS. A mummer-plot (34) of plasmid pSG1 reveals a 6,371 bp tandem repeat encompassing the inactive Type IV secretion system operon that is not present in the original *Sodalis* sequencing and annotation experiments, perhaps due to a collapsed repeat missed by first-generation sequencing assembly. Pseudogenes tend to be enriched for activities such as: transmembrane transport, including metal ions (GO:0099132; 23 pseudogenes / 10 CDS; Fisher Exact Test p=0.001); transposition (GO:0006313; 52 pseudogenes / 14 CDS; p<0.0001); receptors (GO:0004872; 27 pseudogenes / 13 CDS; p<0.0001) and glycerol metabolic processes (GO:0006071; 6 pseudogenes, 0 CDS; p<0.01).

***Sodalis* Transcriptomics:** To ascertain whether pseudogenes are being transcribed, or if their transcription is being regulated throughout growth, stranded RNA-sequencing was performed on three-replicates in three conditions across a bacterial growth experiment in cell-free media (Early Log Phase, ELP; Late Log Phase, LLP; Late Stationary Phase, LSP). Figure 1A shows boxplots of mean sense and antisense transcription (log (transcripts per million + 1) for each condition. From the density plots of sense transcription (Figure 1B) and of overall transcription (Figure 1C), it can be seen that there is a clear signal of no transcription from putatively inactive genes (logTPM+1=0). Other studies have used transcripts per million (TPM) of ≥1 (35) or ≥10 (36) as an indicator of activity, which are displayed on each of these figures. For this study, we have defined putatively active genes as having an arbitrary TPM value ≥ 1 in all three biological replicates in at least one condition. Genes with TPM≥10 in all three biological replicates in at least one condition are additionally described as being active. 53% of all combined genes and pseudogenes, 3,087 (2,237 CDS; 850 pseudogenes) exhibit active sense transcription (TPM≥10) in any given condition, with an additional 1,191 (547 CDS; 644 pseudogenes) putatively active (TPM ≥ 1; A total of 73.7% of all genes and pseudogenes). Additionally, 1,088 genes (703 CDS; 385 pseudogenes) showed active antisense transcription, with an additional 993 genes (629 CDS; 364 pseudogenes) exhibiting putatively active antisense transcription according to the same rules.

Across all conditions, mean sense CDS expression (TPM=245.33) is significantly greater than that of pseudogenes (TPM=60.96; Mann-Whitney U Test; W = 46797000, p-value < 2.2e-16). Mean antisense CDS expression (TPM=24.74) is also significantly greater than that for pseudogenes (TPM=13.64; W = 37651000, p-value < 2.2e-16). Actively expressed pseudogenes (TPM≥10) are enriched for organonitrogen compound metabolic processes (GO:1901564; 51 pseudogenes; 31 CDS; p<0.0001). Putatively expressed pseudogenes (TPM≥1 and TPM<10) tend to be involved in transposase activity (GO:0006313; 52 pseudogenes; 14 CDS; p<0.0001).

Differential expression analysis suggests that, of the actively transcribed genes, 938 CDS and 219 pseudogenes are being differentially expressed between either LLP or LSP when compared to ELP growth (FDR ≤ 0.05). In terms of antisense expression: 219 CDS and 106 pseudogenes showed differential antisense expression between timepoints. DE pseudogenes are involved with nucleic acid and organic cyclic compound binding activities (GO:0003677 and GO:0097159).

Two related examples of specific gene degradation putatively associated with positive or negative selection are the type-III secretion and motility systems. *S. glossinidius* has two broad regions that encode for flagella-related

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

proteins; one intact flagellum region (Flagellum Region 1), and one is undergoing significant degradation (Flagellum Region 2, Figure 2), and three T3SS regions (SSR1-3), all largely intact. Coding flagella genes were significantly more likely exhibit sense expression (chi-sq=31.06; p<0.0001) and more likely to be differentially expressed than pseudogenised flagella genes in Flagellum Region 2 (chi-sq=35.27; p<0.0001; Figure 2). The intact flagellum region is up-regulated in late log phase and down regulated in early log phase growth relative to average expression. SSR-1 is up-regulated in early log phase and down-regulated in late stationary phase. SSRs 2 and 3 are down-regulated in early growth and up-regulated in late log and late stationary phase, respectively (Figure 2).

Figure 3 displays volcano plots of log Fold Change against negative log FDR for Late Log Phase (Figure 3A), and for Late Stationary Phase (Figure 3B), versus Early Log Phase. Each shows that some pseudogenes are highly likely to be differentially expressed between conditions.

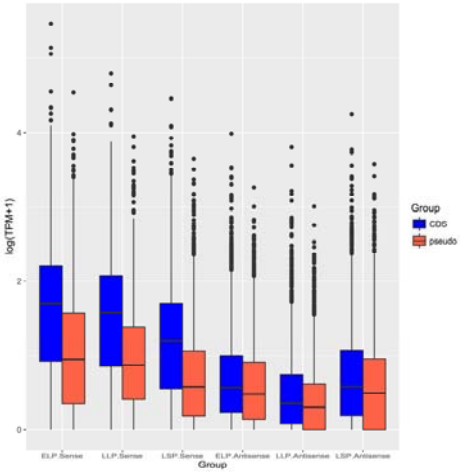



**Figure 1A** – Sense and antisense mean transcription of CDS (blue) and pseudogenes (red) in three cell-free cultures of *Sodalis glossinidius*: Early Log Phase (ELP); Late Log Phase (LLP) and Late Stationary Phase (LSP). Transcripts per million (TPM) derived from EdgeR counts per million have been transformed to log(TPM+1) to enable presentation of zero transcription.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

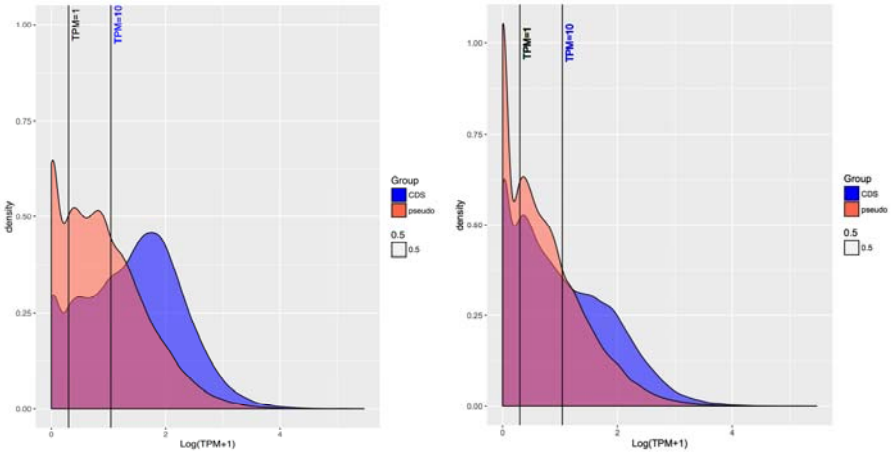
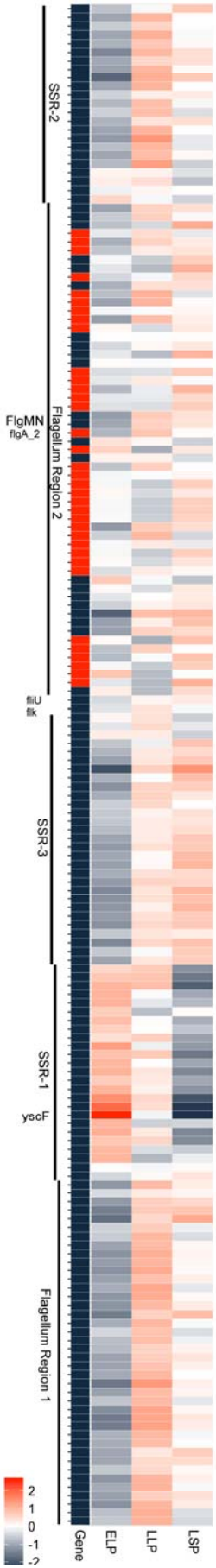



**Figure 1B (left) and 1C (right)** – Density plot of expression (Log (Transcripts per Million + 1)) showing transcription for all three cell-free culture conditions, grouped by CDS (blue) or pseudogene (red). Lines represents TPM = 1 and TPM = 10 that represents two different minimum thresholds to be considered as activity. Figure 1B is sense transcription only, and Figure 1C displays all transcription. CDS can be seen to show greater levels of higher expression levels than pseudogenes (red). Overlapping low-expression CDS and pseudogenes highlights the difficulty in identifying pseudogenes using transcription levels.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

**Figure 2:** *Sodalis glossinidius* flagellum and *Sodalis* Symbiosis Region expression, summarized by general 'region' (left bar). Two genes are not covered by general 'region' bars are fliU (top) and flk (bottom). CDS (blue) and pseudogenes (red) are displayed as the first coloured column. Early log phase (ELP), late log phase (LLP) and late stationary phase (LSP) expression is displayed as a heatmap of Log Fold Change relative to average expression, where red signifies up-regulation and blue represents down-regulation.

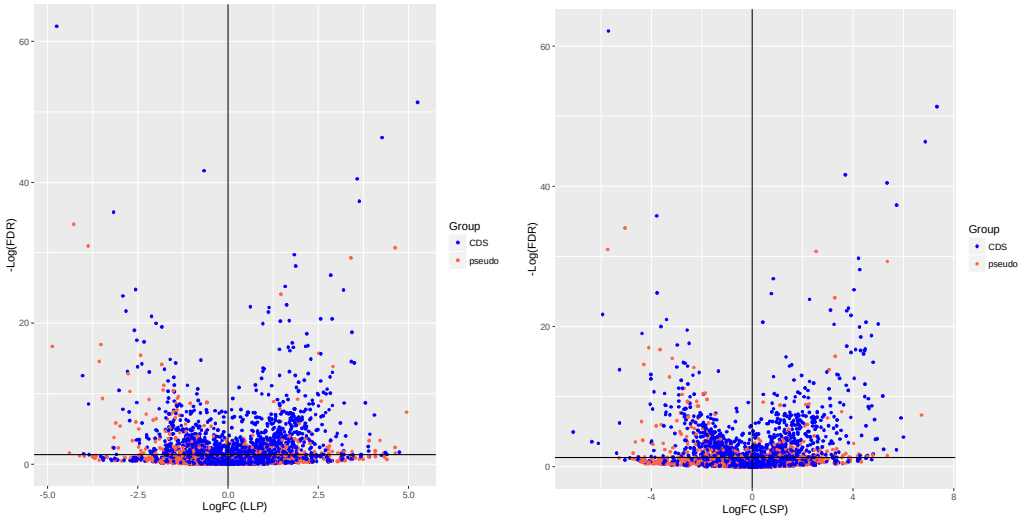



**Figure 3:** $-Log_2FC$ differential expression plotted against negative logFDR (False Discovery Rate), determined from EdgeR. Left pane (A) shows Late Log Phase and Right Pane (B) shows Late Stationary Phase transcription. Whilst CDS (blue) represent the majority of differentially expressed genes, some pseudogenes (red) are significantly differentially expressed in either condition.

**SgGMMB4 Proteome:** To support pseudogene annotation, and to assess whether transcription may relate to translation we performed proteomic analysis from pooled bacterial cells from all conditions in the cell-free growth experiment. The PROKKA gene models and the 6-frame translations comprised a total of 5,625 and 6,769 proteins respectively. With our MS/MS search on the combined search database, we identified 1,503 peptides attributable to the PROKKA annotation. In two instances, an alternative start codon in the annotation (GTG/Val) was manually changed to ATG to encode a methionine residue, in order to match. We identified 34 pseudogenes from previous annotations that corresponded to the presence of a protein product in our data (Supplementary 2 and 3). Seven pseudogenisation events that split ancestral orthologues into two or more ORFs were verified by observation of peptides from a protein product of one of the resultant ORFs. A further 27 peptides were detected, corresponding to (single-ORF) pseudogene predictions from the original annotation. These were re-classified as CDS and are detailed in Supplementary Data 3. It should be noted that these identified proteins are the *representative proteins* from each protein group as reported by ProteoAnnotator. ProteoAnnotator reports one representative protein from each protein ambiguity group (consisting of one or more proteins) formed due to sharing the same set or subset of peptide identifications. This strategy avoids double counting of proteins with no independent evidence. We further computed

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the Exponentially Modified Protein Abundance Index (emPAI) value as a semi-quantitative measure of protein abundance for the representative proteins using the mzidLibrary (Supplementary Figure 1) (37).

**SgGMMB4 Methylome and Codon Usage:** To assess whether methylation patterns differed between intact CDS and pseudogenes, we used the ability of PacBio SMRT sequencing to detect epigenetic modifications, including (for example), 6mA, 4mC or 5mC, by comparing the sequencing profiles (specifically comparing interpulse duration) between native DNA and PCR-amplified DNA (38). 24,869/29,832 (83.4%) of 5-GATC-3 motifs in the SgGMMB4 chromosome are predicted to be 6-Adenine methylated. No other epigenetic modifications or underlying motifs were detected. In CDS and pseudogenes: CDS display a significantly higher frequency of methylation (4.68 per gene) than pseudogenes (2.54 per pseudogene; Kruskal-Wallis chi-squared = 382.61, df = 2, p-value < 2.2e-16). Mean CDS length is larger than pseudogenes (714 bp vs 415 bp), and although CDS and pseudogenes do not significantly differ in their underlying mean GC-content (CDS=55.48%, pseudogenes=55.71%), there are therefore fewer 5-GATC-3 sites within pseudogenes (mean 3 per pseudogene) than in coding sequences (mean 5.6 per CDS; Kruskal-Wallis chi-squared = 428.82, df=3, p-value < 2.2e-16; Supplementary Data 6). There is an increased frequency of both GAT (increase of 2.7 codons per 1000) and ATC (increase of 2.2 codons per 1000) codons in CDS vs. pseudogenes (Supplementary Figure 2).

**Comparative Genomics to other *Sodalis* isolates:** Comparing SNP rates and pseudogene carriage between multiple genomes of *Sodalis* isolated from different tsetse hosts could reveal whether pseudogenes are being deleted at different rates, or if there is relaxed selective pressure acting on pseudogenes. ROARY pan genome analysis of the six Illumina sequenced isolates compared to our PacBio-sequenced annotation assigned 3,183 CDS and 2,301 pseudogenes to either the core genome (all seven genomes), soft core (2 to 6 genomes) or cloud (one genome). ROARY suggests that there are 2,729 core CDS, 358 soft core CDS, and 184 cloud CDS. 1,796 pseudogenes are assigned to the core genome, which represents ~40% of the overall core genome, despite the phylogenetic distance between hosts. There are an additional 280 soft core and 137 cloud pseudogenes (Figure 4B). Low SNP rates and stable pseudogene carriage, indicated by high numbers of pseudogenes contributing to the *Sodalis* core genome, confirms the suggestion of a recent association of *Sodalis* with the tsetse hosts, and that *Sodalis* developed an association prior to tsetse diversification. This may also imply that either: pseudogenes may not be under relaxed selective pressure, due to the low rate of SNP accumulation; or that there has not been sufficient time since the association for SNP to accumulate. Core single nucleotide polymorphisms (SNP) are not targeted towards pseudogenes: Core single-nucleotide polymorphism using SNIPPY suggests there are 474 core SNP loci in pseudogenes, 814 in intact CDS and 540 intergenic core SNP. Enriched functions of core pseudogenes are transporters (GO:0005215; 144 pseudogenes) and oxidoreductase activity (GO:0016491; 92 pseudogenes). Unique pseudogenes to SgGMMB4 are enriched for transposase (GO:0004803; insertion sequences and phage activity).

Functional corrections of pseudogenes could be important for ecological contribution or evolutionary trajectory of pseudogenes. To test for RNA-based (functional) corrections of genomic SNPs at the transcriptional level, a SNIPPY analysis of the RNA-seq data was performed. A single base insertion in a C(8) homopolymeric tract corrects a frameshift in a putative proline/betaine transporter (Base position 869745, SGGMMB4_01173-01174/SG0498).

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Neither of the proP2/3 ORFs have been annotated as pseudogenes in this analysis, although no peptide has been detected for either ORF. A second insertion was found to correct for a frameshift in a putative transposase (C(9)>C(10); SGGMMB4_02124/SG0946). Additionally, but not affecting pseudogenes, a single base insertion (G>GG) was found upstream of the annotation for the CDS SlyA_2, which is a transcriptional regulator (SGGMMB4_03264/SG1443). The SNP putatively either alters the region immediately upstream of the 5'-end or extends and alters the first 22 amino acids if re-annotating using an alternative start codon. A BLASTP search of the sequence suggests the current annotation for SlyA_2 is very similar to *SlyA*/*MarR* orthologues in other endosymbionts (Query coverage ≥ 98%; e ≤ 4e75). Each of these SNPs was only identified in the RNA-seq data and neither in the Illumina nor PacBio DNA sequencing data.

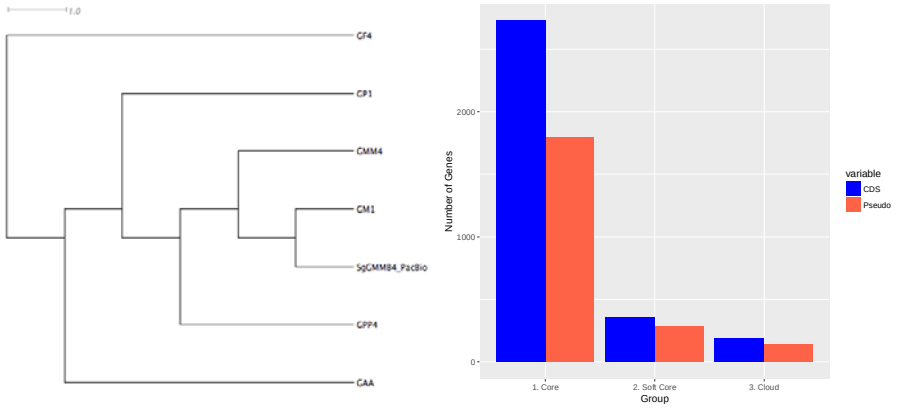



**Figure 4: Pan-genome analysis of** CDS (blue) and pseudogenes (red) from genomes of six *S. glossinidius* isolated derived from four different tsetse species. 2A: GUBBINS-derived whole genome alignment tree showing relatedness between sequenced strains; 2B: Plot representing the number of CDS/Pseudogenes in the ROARY-derived Pan-Genome. Core = 7 genomes; Soft-Core = 2-6 Genomes. Cloud = 1 genome. GAA = *Glossina austeni*; GF = *Glossina fuscipes*; GM(M) = *Glossina morsitans (morsitans)*; GP(P) = *Glossina palpalis (palpalis)*.

## Discussion

It seems logical to consider pseudogenes as potentially maintaining a function until their association with transcriptional processes has been silenced. This is particularly pertinent in the case of Secondary (S) symbionts with high proportions of pseudogenes, like *Sodalis*, which are presumed to be evolving towards an obligate association with their host. In secondary symbiosis, the current opinion is that degeneration can largely be attributed to small, vertically transmitted, populations with little diversity reducing the ability of the organism to purge deleterious mutations (i.e. leading to the generation and persistence of pseudogenes) (1, 39, 40).

**Pseudogenes can harbour residual sense and antisense transcription:** We have shown that bacterial pseudogenes can be both actively transcribed and dynamically regulated during growth. Expressed genes hypothesized to be

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

under negative selection (i.e. annotations that were significantly more likely to be present in pseudogenes than CDS) were likely to have functions related to transmembrane transport (particularly Major Facilitator Superfamily (MFS) transporters), glycerol metabolism and transposition. MFS pseudogenes permeate genome annotations: A search for "MFS pseudogene" in Ensembl bacteria shows 1,984 genes at the time of writing, across many genera including *Escherichia*, *Fransicella* and *Pseudomonas* (41). Genes linked to substrate transport (such as MFS transporters) and metabolism are commonly found to be degraded in insect symbiont genomes, linked to their restricted diets (42). Because positive selection is intrinsically linked to substrate bioavailability – as environmental conditions change (such as a lack of nutrients or the provision of a processed metabolite directly to the symbiont) – relaxed selection leads to the accumulation of deleterious mutations and pseudogenisation of redundant genes. Nevertheless, residual expression from these may be indicative of a recent switch in environment, where mutation at the genomic level could outpace overarching transcriptional control, in so much that transcription may be positively selected for due to links to other, maintained, mechanisms.

We have highlighted the related flagella and Type-III secretion / Symbiosis Regions as examples of regions undergoing clear patterns of pseudogenisation (i.e. pseudogenes are clustered with few exceptions, e.g. FlgMN; Figure 2). It is clear that coding genes associated with motility and T3SS apparatus were more likely to be differentially expressed than pseudogenes, and thus the canonical definition of pseudogenes may hold true in these cases. The FlgMN operon remains intact, despite residing within the degrading flagellum region 2, and is differentially expressed in line with Flagellum Region 1 (Supplementary Data 5). Additionally, the only flagellum/T3SS-related pseudogene exhibiting differential expression was flgA_2, which is the only other gene in the FlgMN operon, suggesting that pseudogene differential expression could be a factor of co-regulation of intact genes, or as a result of polycistronic regulation across operons. One of the most differentially expressed pseudogenes, without a hypothetical annotation, is trehalase, an enzyme involved in the breakdown of trehalose – a sugar commonly found in insect haemolymph (43). Amongst a number of hypotheses as to the differential expression of trehelase in cell free culture (which lacks trehalose as a constituent) are implications of residual global control of metabolic processes beyond single sugars. Given the importance of this sugar in insect systems, it would be interesting to further test this pseudogene for residual function.

**Pseudogenes are difficult to define:** By using stranded RNA sequencing, and comparing transcription between *Sodalis* CDS and pseudogenes, we have shown that intact CDS show significantly greater mean levels of transcription than pseudogenes, but there remains a proportion of CDS with little or no expression – any of which could be non-functional and mis-annotated. It should be noted that this experiment relied on cell-free culture; CDS may therefore not be expressed due to their functional redundancy in cell-free culture, and follow-up experiments may be required in further media types, or in insect-cell co-culture, to fully ascertain the *Sodalis* transcriptional repertoire. Similarly, however, there remain pseudogenes with residual activity, going against the classical definition of a pseudogene, and it is clear, therefore, that problems remain with the identification and annotation of pseudogenes. We and others have identified novel genes, including genes potentially important in regulating flagellum and/or Type III secretion machinery (*HilA*) or in Quorum Sensing (*SlyA*/*MarR*) (44). Simply defining pseudogenes using any individual genomic assay is difficult: ORFs may be shortened by frameshift mutations, yet may retain functional domains and

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

appropriate transcriptional architecture. Coding sequences are generally characterized following a set of canonical rules of gene structure: the presence of an open reading frame (ORF), a promoter and ribosomal binding site (RBS); a methionine (or, occasionally, alternative) start-codon and a stop-codon. Similarly, pseudogenes were predicted wherever such rules break down: in the case of *S. glossinidius*, pseudogenes were predicted where <50 % of the functional homologue remained intact. Although studies at the single cell level in *E. coli* (45), and in some conditions at the population level in *Clostridium* (46), suggest that levels of mRNA and protein can remain uncorrelated and be regulated independently of one another, our data suggest that a balance may exist between mRNA transcript and protein abundance, as a semi-quantitative measure of peptide abundance correlates with sense expression. It is likely that each tier of control (i.e. at the DNA, transcription and translation levels), that each may act on another – for instance sRNA may impact mRNA levels, or protein interactions may regulate transcription. There remains a range of bacterial transcriptional processes still to fully characterize, including 5'-UTRs, alternative promoters or alternative transcriptional start and stop sites, and further experiments using techniques such as terminal endonuclease linked RNA-seq, which has been employed in similar experiments in *Salmonella enterica* serovar Typhimurium (47), would shed further light on the transcriptional landscape of this bacterium. It is important to stress that, in this study, we have defined pseudogenes specifically where peptides have *not* been detected; Given the difficulty for proteomics analysis to detect membrane proteins (48), further work may well reveal the presence of hitherto undiscovered peptides and thus that some of these differentially expressed pseudogenes may be mis-annotated and further explain the presence of residual transcription.

**Transcriptional- and post-transcriptional pseudogene control mechanisms remain to be ascertained:** Given its dual role in mismatch repair and the regulation of gene expression, Dam-mediated methylation of 5-GATC-3 motifs in bacterial genomes represents a potentially important factor to investigate. While Pacific Biosciences sequencing allowed for the examination of methylation status by comparing modified to unmodified DNA, the potential role methylation might play in pseudogene control remains difficult to ascertain: Pseudogenes displayed a significantly decreased rate of 6mA methylation, when compared to CDS, probably due to the tendency for pseudogenes to have fewer 5-GATC-3 methylation motifs at the genetic level (because pseudogenes are smaller than CDS). Dam-mediated methylation is predicted to post-transcriptionally regulate gene expression by altering the affinity of proteins for DNA, such as at the origin of replication (oriC) (49). In *S. enterica* serovar Typhimurium, Adenine methylation has been implicated in regulating quorum sensing derived virulence factors and as such Dam inhibitors or Dam-silenced pathogens have been studied for their antimicrobial or vaccine potential, respectively (50). Adenine methylation has also been implicated in protecting symbionts from heat-stress (51).

**Pseudogene abundance is stable between *Sodalis* genomes:** Given that we expect *Sodalis* to be routinely undergoing population bottlenecks through vertical transmission of their host, we could expect genetic drift to be acting on genes under little selective pressure, increasing SNP and/or pseudogenisation rates, or even driving their deletion. As accessory genomes diverge prior to SNP arising in the core genome (52), examining the *Sodalis* pan-genome derived from *S. glossinidius* species from multiple tsetse hosts, enabled us to examine pseudogene stability.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The high number of pseudogenes in the *Sodalis* core genome suggests pseudogenes are stable across *Sodalis* strains infecting different tsetse species – in line with the suggestion that *Sodalis* shares an evolutionarily recent association with its tsetse host due to a lack of concordant phylogeny, in contrast to that of *Wigglesworthia*, its primary symbiont (53).. Lawrence *et al* (2001) suggested that intracellular lifestyle promotes protection from bacteriophages and insertion sequences, reducing recombination and promoting pseudogene persistence (54). *S. glossinidius* GMBB4 has 44 chromosomal prophage elements, and an active circular phage (pSG3). Furthermore, sexual transmission of *Sodalis* from father to mother prior to vertical transmission has been reported, effectively increasing diversity and rates of recombination (55). Assessing *Sodalis* genetic variability across wild populations of tsetse, rather than lab-reared tsetse will be essential in understanding pseudogene persistence. Similarly, further research into expression levels of both coding and pseudogene orthologues may reveal further insight into transcriptional co-regulation – and gene redundancy linked to reductive evolution (56). Kuo and Ochman have previously suggested that *Salmonella* pseudogenes may lack sufficient negative pressure for deletion (11). In models of cyanobacterial genomes, increased resource levels and decreased mortality have been suggested to select for slower reproduction and streamlined genomes (57). Experimental evolution experiments in *Methylobacterium* have shown that accessory gene deletion confers a direct fitness benefit under selective environments, rather than the associated benefit of the reduced fitness costs of maintaining a shorter genome in its own right (58), and similar experiments would be interesting to perform in this system.

Defining which associations constitute significant function for which positive pressure ensures their persistence is difficult: associations with promoters, transcription factors or *cis*/*trans*-acting transcriptional regulators could all select for pseudogene retention in the genome, and reduce the likelihood of full deletion. An increasing number of bacterial small RNAs have been identified through transcriptomic analyses, including in *Streptococcus* (59) and *Borrelia* (60). Pseudogene-derived antisense RNA may be involved in the complex interactions between genome, transcriptome and proteome: sRNAs have been implicated in gene regulation of multiple target genes through processes such as translational inhibition and activation, or transcript stability (61). *Hfq*, a crucial chaperone involved in bacterial sRNA processing, is maintained in *Sodalis* (SGGMMB4_00878). Therefore, further studies into the roles of *Sodalis* sRNAs – including pseudogene-derived sRNAs and the role of *Hfq* or other chaperones – will be critical for full understanding of the complexity of gene regulation in this degrading bacterial genome. This is in line with previous work using tiling arrays in *Mycoplasma pneumonia*, wherein frequent antisense and non-coding transcripts were identified in a degrading bacterial genome (62). Pseudogene-derived transcripts (such as antisense small RNAs derived from pseudogenes) could act as regulators for orthologues elsewhere in the genome (63) and such an association may reduce the selective pressure towards their deletion. As this study relied on RNA-shearing based library preparation, it would be interesting to follow up this study with full-length third-generation RNA sequencing (i.e. using Pacific Biosciences cDNA (Isoseq) or direct RNA sequencing using Oxford Nanopore technologies) to elucidate the sequences of full-length mRNA, including primary, polycistronic transcripts, which would further enhance our knowledge as to how pseudogenes continue to contribute to overall transcription and its control despite ongoing genomic degradation.

## Conclusions

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The primary goal of this study was to establish whether bacterial pseudogenes remain active despite genomic degradation, using *Sodalis* as a model given the number of putatively inactivated, functionless genes persisting in its genome. We sought to combine DNA sequencing, stranded RNA-sequencing and proteomic analysis to fully describe the *Sodalis* transcriptional and translational landscape with a view to better understand the evolution and functional control of bacterial pseudogenes and the process of endosymbiont genome degradation.

We have revealed that whilst transitioning from a free-living to symbiotic status, *Sodalis* pseudogenes are often transcribed, but at a significantly lower level than intact CDS. Some pseudogenes even remain under active transcriptional control, exhibiting differential expression throughout growth, however proteomic analysis suggests they ultimately do not contribute to the protein content of the cell. The lack of some expression from intact CDS and pseudogenes underpins the difficulty in pseudogene identification – especially in cell-free culture where the correct conditions for their expression may be lacking. That a combination of sense and antisense transcription of pseudogenes persists implies a role of pseudogene transcription in control mechanisms: e.g. cis/*trans* small RNA transcriptional control, and could even be playing a role in wide-reaching mechanisms such as host-symbiont interaction, or symbiont-symbiont interaction. The persistence of pseudogenes in the *Sodalis* pangenome implies that the maintenance of function of degraded genes may outweigh any deleterious effects, or that there exists a mechanism by which such deleterious effects are mitigated. Given the proximity of *Sodalis* to medically important parasites and other bacteria within the tsetse host, further study on these mechanisms is of interest for identifying novel therapeutic interventions.

## Materials and Methods

**DNA Sequencing:** For PacBio sequencing: *Sodalis glossinidius* strain GMMB4 (SgGMMB4) was isolated from *Glossina morsitans morsitans* (Westwood) from the Langford derived long-term colony maintained at the University of Edinburgh in 2005. Six further *S. glossinidius* isolates were cultured from lab-based tsetse for Illumina sequencing: GP1 and GPP4 were isolated from *Glossina palpalis palpalis*; GAA from *G. austeni*; GF4 from *G. fuscipes*; GM1 and GMM4 from *G. morsitans* as previously described (64).

Bacteria were recovered from -80 ºC storage by incubation at 25 ºC on columbia agar plates supplemented with 10 % defibrinated horse blood (TCS Biosciences) in microaerophilic conditions (~5-12% $CO_2$ CampyGen, Oxoid, UK). An individual colony was picked and grown to late stationary phase in cell-free culture medium at 25 ºC in Schneiders Insect Medium (Sigma, UK) supplemented with 10 % Fetal Calf Serum (Life Technologies, UK). High molecular weight whole genomic DNA (gDNA) was extracted from the subsequent bacterial pellet using the Zymo Research Universal gDNA extraction kit (SgGMMB4) or the Qiagen DNeasy kit (other isolates) according to the manufacturers instructions.

SgGMMB4 gDNA was sequenced on the Pacific Biosciences RS-II instrument (PacBio) at the Centre for Genomic Research at the University of Liverpool on a single SMRTcell using P6-C4 chemistry with no prior size selection.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Reads were assembled and contigs polished using HGAP.3 resulting in a polished assembly consisting of a single chromosomal contig and nine further contigs. Comparison of the sequence to the available reference by MUMMER (65) and ACT (66) suggested that two contigs were a result of chimeras derived from pSG2. A further five contigs were repetitive phage-derived sequences. The chromosome was subsequently manually edited to begin at the start of the dnaA gene. The putative protein-coding, ncRNA, and tRNA gene sequences were annotated using PROKKA (v. 1.10) (67). Pseudogenes in this study were initially conservatively annotated where the PROKKA-defined ORFs overlapped with the Belda-annotated pseudogenes or predicted to be pseudogenes by PROKKA (67). In the latter case, PROKKA predicted pseudogenes based on identical annotations in sequential open reading frames (except for hypothetical protein annotations). Sequences matching possible riboswitch domains were predicted using the Denison Riboswitch Detector online webserver (Supplementary Data 4) (68). Additionally, the two available annotations for SgGMM4 (14, 23) were transferred to the PacBio SgGMMB4 scaffold using the RATT software package (32), for comparison and pseudogene prediction. An additional *S. glossinidius* sample (isolated from *Glossina palpalis*) was sequenced and assembled in the same manner using two SMRTcells using P6-C4 chemistry.

Sequencing libraries for the six further isolates of *Sodalis glossinidius* from multiple tsetse species (GAA; GF4; GM1; GMM4; GP1; GPP4) were prepared using a TruSeq library preparation kit (according to the manufacturer's instructions) and sequenced on a single lane of an Illumina HiSeq (High-output run; 2x100bp paired end reads) at the Centre for Genomic Research. The Illumina HiSeq data from the six further *S. glossinidius* isolates were initially processed using CASAVA 1.8 to produce FASTQ files. FASTQ data files were trimmed for the presence of Illumina adapter sequences using Cutadapt (v1.2.1), using the –O 3 option (69). The reads were further trimmed using Sickle (v1.200) (https://github.com/najoshi/sickle) with a minimum window quality score of 20. Data were assembled *de novo* with SPADES using default parameters and annotated using PROKKA as previously described. PROKKA-derived GFF annotations were processed through the ROARY pan-genome package to ascertain core and accessory genome coverage (70). Reads were mapped to the SgGMMB4 PacBio reference, and core SNP phylogenies derived using the SNIPPY package (https://github.com/tseemann/snippy).

**Methylome sequencing:** In addition to gDNA sequencing and assembly, the PacBio RSII instrument can detect epigenetic modification either in silico or by comparing native DNA to a PCR control. To that end, a Whole Genome Amplified (WGA) control was generated from SgGMMB4 as follows: 1µg gDNA was split into three equal reactions and Whole Genome Amplified using the Qiagen Repli-g Turbo kit according to the manufacturers instructions. These were then pooled and cleaned using a 2:1 ratio of homemade SPRI bead cleanup system analogous to Ampure XP beads (71). The WGA control was sequenced using one SMRTcell in the same way as described previously. Comparison to the native DNA was performed using the Motifs and Modifications module within the SMRT® Analysis Server with a mapping quality cutoff set at QV70 and the subsequent modifications and motifs file filtered for those with a quality ≥ Q50 ($P<0.0001$).

**RNA sequencing:** Individual 10 mL cell-free liquid cultures, as described above, were set up in quintuplets for seventeen time-points at six-hourly intervals. At each time point, the contents of the culture flasks were transferred to 15 mL Falcon tubes, an Optical Density (600nm) measurement taken, and then centrifuged for 10 min at 10 ºC. The

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

bacterial pellet was immediately resuspended in 1 mL Trizol reagent (Life Technologies) and total RNA was extracted using Zymo Research DirectZol columns. RNA cleanups were performed using a 2:1 ratio of SPRI beads as described previously. Three timepoints, representing: Early Log (12 hours); Late Log (72 hours) and Late Stationary phase (108 hours) according to the OD measurements were selected and DNase I treated using a Life Technologies DNAfree Turbo kit (according to the manufacturers instructions; (data not shown)).

Ribosomal RNA (rRNA) was depleted using a Ribo-Zero™ bacterial (low-input) rRNA Removal Kit (Epicentre), and individually barcoded, strand-specific Illumina cDNA libraries were prepared using a NEBNext® Ultra™ RNA Library Prep Kit for Illumina. Sequence data was generated using one Illumina MiSeq run with v2 chemistry generating 250 bp paired-end reads. All RNA and cDNA cleanups were performed using SPRI beads as described previously. All raw Illumina sequence Fastq files were trimmed for the presence of adapter sequences using Cutadapt version 1.2 using option -O 3 (69) and quality-trimmed using Sickle version 1.200 (72) with a minimum window quality score of 20. Any reads shorter than 10 bp after trimming were removed. Quality scores for all sequences were assessed using FASTQC v0.9.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). RNA-seq reads were aligned to the PacBio RSII-generated SgGMMB4 scaffold using Bowtie2 v. 2.1.0 (73). The resulting SAM files were converted to BAM and sorted using samtools v.0.1.18-r580 (74)s. For transcript-based annotation, reads were counted against the PROKKA/PacBio annotation using HTSeq version 0.5.3p9 using the --stranded option, to count in both the sense and antisense directions, and using the intersection-nonempty mode (75). EdgeR analysis was implemented using the DEGUST web package (http://degust.erc.monash.edu), which outputs counts per million, logFC and differential expression statistics (Supplementary Data 1). EdgeR counts per million were transformed to TPM according to Wagner *et al* (2012) (76). Further statistical analysis and figure plotting were implemented using R version 3.32. Operon structure was predicted from the RNA-seq data using Rockhopper with the default settings (77). The combined RNAseq sequencing reads were fed through the SNIPPY pipeline (as previously described) to identify potentially correcting SNPs. Gene enrichment analyses were performed in BLAST2GO 5, using Interpro GO IDs (Supplementary Data 1).

**Proteomics:** Liquid cultures were grown as previously described for early- mid-log and late-stationary growth phase *S. glossinidius* GMMB4. PBS-washed pellets were suspended in 250 µl of 25 mM ammonium bicarbonate and sonicated, using a Sonics Vibra Cell (Sonics and Materials Inc., Newton, U.S.A.) and 630-0422 probe (250 µl to 10 mL) for a total of 120 joules. The sample was then analysed for protein content, 50 µg was added to 0.05% RapiGest™ (Waters, Manchester) in 25 mM ammonium bicarbonate and shaken at 550 rpm for 10 min at 80°C. The sample was then reduced (addition of 10 µL of 60 mM DTT and incubation at 60 °C for 10 minutes) and alkylated (addition of 10 µl of 180 mM iodoacetamide and incubation at room temperature for 30 minutes in the dark). Trypsin (Promega U.K. Ltd., Southampton, proteomics grade) was reconstituted in 50 mM acetic acid to a concentration of 0.2 µg/µl and 10 µL added to the sample followed by overnight incubation at 37 °C. The digestion was terminated and RapiGest™ removed by acidification (1 µl of TFA and incubation at 37 °C for 45 min) and centrifugation (15,000 x g for 15 min). To check for complete digestion each sample was analysed pre- and post-acidification by SDS-PAGE.

For LC-MS/MS analysis, a 2 µl (1 µg) injection was analysed using an Ultimate 3000 RSLC™ nano system (Thermo Scientific, Hemel Hempstead) coupled to a QExactiveHF™ mass spectrometer (Thermo Scientific). The sample was

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

loaded onto the trapping column (Thermo Scientific, PepMap100, C18, 300 µm X 5 mm), using partial loop injection, for seven minutes at a flow rate of 4 µl/min with 0.1 % (v/v) FA. The sample was resolved on the analytical column (Easy-Spray C18 75 µm x 500 mm 2 µm column) using a gradient of 97 % A (0.1 % formic acid) 3 % B (99.9 % ACN 0.1 % formic acid) to 70 % A 30 % B over 120 minutes at a flow rate of 300 nl $min^{-1}$. The data-dependent program used for data acquisition consisted of a 60,000 resolution full-scan MS scan (AGC set to 3e6 ions with a maximum fill time of 100 ms) the 18 most abundant peaks were selected for MS/MS using a 30,000 resolution scan (AGC set to 1e5 ions with a maximum fill time of 45 ms) with an ion selection window of 1.2 m/z and a normalised collision energy of 28. To avoid repeated selection of peptides for MSMS the program used a 30 second dynamic exclusion window.

The protein identification for the MS/MS dataset was performed using an open-source software tool – ProteoAnnotator (78). ProteoAnnotator provides an automated pipeline for various interconnected computational steps required for inferring statistically robust identifications. The tool produces a variety of output files compliant with the data standards developed by Proteomics Standard Initiative (79). Mass spectra in form of a MGF (Mascot Generic format) file were provided as input to the tool, along with the search criteria and protein database as described in the sections below. The search parameters for the MS/MS dataset were fixed modification of carbamidomethylation of cysteine and variable modification of oxidation of methionine. A single missed trypsin cleavage was allowed. The product tolerance was set as ±0.5 Da and the precursor tolerance was set as 10 ppm. The protein search database comprised of the gene model predicted by PROKKA, as previously described, plus a six-frame translation of the SgGMMB4 genome. Six-frame translated sequences with length less than eight were excluded from the search database. Decoy sequences were added to the database with a true:decoy ratio of 1:1 to create a final protein database for performing the MS/MS search. For the post-processing of results, we applied a threshold of 5 % for both peptide level and protein group level FDRs as described in Ghali *et al* (2004).

**Nucleotide sequence database identifiers**

The Pacific Biosciences assembly and annotation have been submitted to the European Nucleotide Archive under accessions LN854557-LN854560. Illumina sequence reads for the six additional *Sodalis* isolates are available under Project accession PRJEB9474 (accessions ERR2036891-ERR2036896). RNAseq data are available under project PRJEB20150.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (80) partner repository with the dataset identifier PXD007068.

Acknowledgements

This research was funded by a BBSRC project grant BBJ017698/1 to ACD. The Pacific Biosciences RS-II instrument purchase was funded by the BBSRC under grant BB/L014777/1. IG is supported by a Wellcome Trust Seed Award in Science (Ref: 200690/Z/16/Z). Pisut Pongchaikul was supported by a Maihidol studentship. All sequencing was carried out at the Centre for Genomic Research at the University of Liverpool, UK. Thanks also to Dr. Chandan Pal, and the

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hinton and Kröger research groups for their support and useful discussions, and to anonymous reviewers whose comments helped us to significantly improve the manuscript.

References

1. McCutcheon JP, Moran N a. 2011. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26.

2. Douglas AE. 2016. How multi-partner endosymbioses function. Nat Rev Microbiol 14:731–743.

3. Wernegreen JJ. 2002. Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet 3:850–861.

4. Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A. 2006. A small microbial genome: the end of a long symbiotic relationship? Science (80- ) 314:312–313.

5. Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci 93:2873–2878.

6. Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596.

7. Tamas I, Klasson L, Canbäck B, Näslund AK, Eriksson A-S, Wernegreen JJ, Sandström JP, Moran NA, Andersson SGE. 2002. 50 Million Years of Genomic Stasis in Endosymbiotic Bacteria. Science 296:2376–2379.

8. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature 407:81–6.

9. Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S. 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. Nat Genet 32:402–7.

10. Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol 12:841–850.

11. Kuo C-H, Ochman H. 2010. The extinction dynamics of bacterial pseudogenes. PLoS Genet 6:e1001050.

12. Liu Y, Harrison PM, Kunin V, Gerstein M. 2004. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol 5:R64.

13. Foley SL, Johnson TJ, Ricke SC, Nayak R, Danzeisen J. 2013. Salmonella pathogenicity and host adaptation in chicken-associated serovars. Microbiol Mol Biol Rev 77:582–607.

14. Toh H, Weiss BL, Perkin SAH, Yamashita A, Oshima K, Hattori M, Aksoy S. 2006. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

15. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honoré N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG. 2001. Massive gene decay in the leprosy bacillus. Nature 409:1007–11.

16. Gómez-Valero L, Latorre A, Silva FJ. 2004. The evolutionary fate of nonfunctional DNA in the bacterial endosymbiont Buchnera aphidicola. Mol Biol Evol 21:2172–81.

17. Goodhead I, Darby AC. 2015. Taking the pseudo out of pseudogenes. Curr Opin Microbiol 23:102–109.

18. Ochman H, Davalos LM. 2006. The nature and dynamics of bacterial genomes. Science 311:1730–3.

19. Nuccio SP, Bäumler AJ. 2014. Comparative analysis of Salmonella genomes identifies a metabolic network for escalating growth in the inflamed gut. MBio 5.

20. Tamas I, Wernegreen JJ, Nystedt B, Kauppinen SN, Darby AC, Gomez-Valero L, Lundin D, Poole AM, Andersson SGE. 2008. Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. Proc Natl Acad Sci U S A 105:14934–9.

21. Dennis JW, Durkin SM, Horsley Downie JE, Hamill LC, Anderson NE, MacLeod ET. 2014. Sodalis glossinidius prevalence and trypanosome presence in tsetse from Luambe National Park, Zambia. Parasit Vectors 7:378.

22. Maudlin I. 1982. Inheritance of susceptibility to _Trypanosoma congolense_ infection in _Glossina morsitans_. Ann Trop Med Parasitol 76:225–227.

23. Belda E, Silva FJ, Peretó J, Moya A. 2012. Metabolic networks of Sodalis glossinidius: a systems biology approach to reductive evolution. PLoS One 7:e30652.

24. Pontes MH, Babst M, Lochhead R, Oakeson K, Smith K, Dale C. 2008. Quorum sensing primes the oxidative stress response in the insect endosymbiont, Sodalis glossinidius. PLoS One 3.

25. Hansen-Wester I, Hensel M. 2001. Salmonella pathogenicity islands encoding type III secretion systems. Microbes Infect 3:549–559.

26. Dale C, Young SA, Haydon DT, Welburn SC. 2001. The insect endosymbiont Sodalis glossinidius utilizes a type III secretion system for cell invasion. Proc Natl Acad Sci 98:1883–1888.

27. Breaker RR. 2012. Riboswitches and the RNA world. Cold Spring Harb Perspect Biol 4.

28. Storz G, Vogel J, Wassarman KM. 2011. Regulation by Small RNAs in Bacteria: Expanding Frontiers. Mol Cell.

29. Sánchez-Romero MA, Cota I, Casadesús J. 2015. DNA methylation in bacteria: From the methyl group to the methylome. Curr Opin Microbiol.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

30. Casadesús J, Low D. 2006. Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol Rev 70:830–56.

31. Darby AC, Lagnel J, Matthew CZ, Bourtzis K, Maudlin I, Welburn SC. 2005. Extrachromosomal DNA of the Symbiont Sodalis glossinidius Extrachromosomal DNA of the Symbiont Sodalis glossinidius † 187.

32. Otto TD, Dillon GP, Degrave WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res 39.

33. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

34. Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics Chapter 10:Unit 10.3.

35. Priam P, Krasteva V, Rousseau P, D'Angelo G, Gaboury L, Sauvageau G, Lessard JA. 2017. SMARCD2 subunit of SWI/SNF chromatin-remodeling complexes mediates granulopoiesis through a CEBP[epsiv] dependent mechanism. Nat Genet advance on.

36. Kröger C, Dillon SC, Cameron ADS, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JCD. 2012. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci U S A 109:E1277-86.

37. Ghali F, Krishna R, Lukasse P, Martinez-Bartolome S, Reisinger F, Hermjakob H, Vizcaino JA, Jones AR. 2013. A toolkit for the mzIdentML standard: the ProteoIDViewer, the mzidLibrary and the mzidValidator. Mol Cell Proteomics 3026–3035.

38. Davis BM, Chao MC, Waldor MK. 2013. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. Curr Opin Microbiol.

39. Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol 2:1–12.

40. Lo WS, Huang YY, Kuo CH. 2016. Winding paths to simplicity: Genome evolution in facultative insect symbionts. FEMS Microbiol Rev.

41. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. 2017. Ensembl 2018. Nucleic Acids Res.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

42. McCutcheon JP, McDonald BR, Moran NA. 2009. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci U S A 106:15394–9.

43. Heddi A, Vallier A, Anselme C, Xin H, Rahbe Y, Wäckers F. 2005. Molecular and cellular profiles of insect bacteriocytes: Mutualism and harm at the initial evolutionary step of symbiogenesis. Cell Microbiol 7:293–305.

44. Clayton AL, Enomoto S, Su Y, Dale C. 2017. The regulation of antimicrobial peptide resistance in the transition to insect symbiosis. Mol Microbiol 103:958–972.

45. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, Emili A, Xie XS. 2010. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Sci (New York, NY) 329:533–538.

46. Venkataramanan KP, Min L, Hou S, Jones SW, Ralston MT, Lee KH, Papoutsakis ET. 2015. Complex and extensive post-transcriptional regulation revealed by integrative proteomic and transcriptomic analysis of metabolite stress response in Clostridium acetobutylicum. Biotechnol Biofuels 8:81.

47. Kröger C, Dillon SC, Cameron ADS, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JCD. 2012. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. Proc Natl Acad Sci U S A 109:E1277-86.

48. Rabilloud T. 2009. Membrane proteins and proteomics: Love is possible, but so difficult. Electrophoresis.

49. Marinus MG, Casadesus J. 2009. Roles of DNA adenine methylation in host-pathogen interactions: Mismatch repair, transcriptional regulation, and more, p. 488–503. *In* FEMS Microbiology Reviews.

50. Heithoff DM, Sinsheimer RL, Low D a, Mahan MJ. 1999. An essential role for DNA adenine methylation in bacterial virulence. Science 284:967–970.

51. Reingold V, Luria N, Robichon A, Dombrovsky A. 2014. Adenine methylation may contribute to endosymbiont selection in a clonal aphid population. BMC Genomics 15:999.

52. Andreani NA, Hesse E, Vos M. 2017. Prokaryote genome fluidity is dependent on effective population size. ISME J.

53. Chen X, Li S, Aksoy S. 1999. Concordant evolution of a symbiont with its host insect species: Molecular phylogeny of genus Glossina and its bacteriome-associated endosymbiont, Wigglesworthia glossinidia. J Mol Evol 48:49–58.

54. Lawrence JG, Hendrix RW, Casjens S. 2001. Where are the pseudogenes in bacterial genomes? Trends Microbiol 9:535–540.

55. De Vooght L, Caljon G, Van Hees J, Van Den Abbeele J. 2015. Paternal transmission of a secondary symbiont during mating in the viviparous tsetse fly. Mol Biol Evol 32:1977–1980.

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

56. Andersson JO, Andersson SG. 1999. Insights into the evolutionary process of genome degradation. Curr Opin Genet Dev 9:664–671.

57. Bentkowski P, van Oosterhout C, Ashby B, Mock T. 2017. The effect of extrinsic mortality on genome size evolution in prokaryotes. ISME J 11:1011–1018.

58. Lee MC, Marx CJ. 2012. Repeated, selection-driven genome reduction of accessory genes in experimental populations. PLoS Genet 8.

59. Le Rhun A, Beer YY, Reimegård J, Chylinski K, Charpentier E. 2016. RNA sequencing uncovers antisense RNAs and novel small RNAs in *Streptococcus pyogenes*. RNA Biol 13:177–195.

60. Arnold WK, Savage CR, Brissette CA, Seshu J, Livny J, Stevenson B. 2016. RNA-Seq of Borrelia burgdorferi in Multiple Phases of Growth Reveals Insights into the Dynamics of Gene Expression, Transcriptome Architecture, and Noncoding RNAs. PLoS One 11:e0164165.

61. Hansen AK, Degnan PH. 2014. Widespread expression of conserved small RNAs in small symbiont genomes. ISME J 8:2490–502.

62. Güell M, Yus E, Lluch-Senar M, Serrano L. 2011. Bacterial transcriptomics: what is beyond the RNA horizome? Nat Rev Microbiol 9:658–69.

63. Guo X, Zheng D. 2011. Regulatory roles of novel small RNAs from pseudogenes, p. 193–208. *In* Non Coding RNAs in Plants. Springer.

64. Matthew CZ, Darby AC, Young S a, Hume LH, Welburn SC. 2005. The rapid isolation and growth dynamics of the tsetse symbiont Sodalis glossinidius. FEMS Microbiol Lett 248:69–74.

65. Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics Chapter 10:Unit 10.3.

66. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT: The Artemis comparison tool. Bioinformatics 21:3422–3423.

67. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics btu153-.

68. Havill JT, Bhatiya C, Johnson SM, Sheets JD, Thompson JS. 2014. A new approach for detecting riboswitches in DNA sequences. Bioinformatics 30:3012–3019.

69. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

70. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693.

71. Rohland N, Reich D. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

capture. Genome Res 22:939–946.

72. Joshi N, Fass J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle 2011.

73. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

74. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

75. Anders S, Pyl PT, Huber W. 2015. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169.

76. Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 131:281–285.

77. Tjaden B. 2015. De novo assembly of bacterial transcriptomes from RNA-seq data. Genome Biol 16:1.

78. Ghali F, Krishna R, Perkins S, Collins A, Xia D, Wastling J, Jones AR. 2014. ProteoAnnotator - Open source proteogenomics annotation software supporting PSI standards. Proteomics 14:2731–2741.

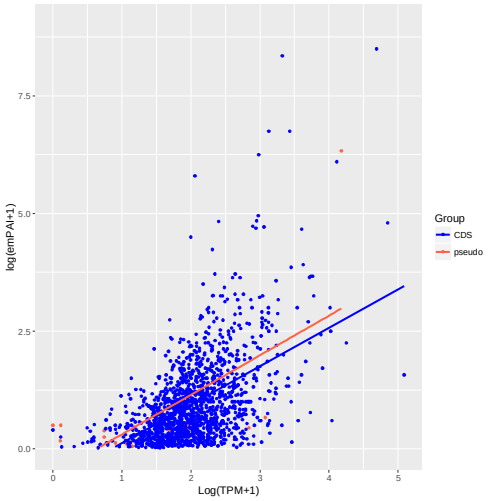
79. Medina-Aunon JA, Krishna R, Ghali F, Albar JP, Jones AJ. 2013. A guide for integration of proteomic data standards into laboratory workflows. Proteomics.

80. Vizcaíno J, Deutsch EEW, Wang R, Vizcaino JA, Deutsch EEW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz P-A, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus H-J, Albar JP, Martinez-Bartolome S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotech 32:223–226.

**Supplementary Data:**

**Supplementary Data 1: Differential Expression analysis for *Sodalis glossinidius* SgGMMB4 (SuppData1.xlsx)**

bioRxiv preprint doi: https://doi.org/10.1101/124388; this version posted April 17, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.




**Supplementary Figure 1:** Transcription vs. semi-quantitative protein abundance (for all detected peptides). Line represents linear regression (implemented using ggplot2 in R).

**Supplementary Data 2-4 (SuppData2_3_4.docx)**

Supplementary Data 2: Pseudogenised genes with residual active translation

Supplementary Data 3: Table of Single ORF Pseudogenes re-annotated due to detected peptide products

Supplementary Data 4: Riboswitch predictions

**Supplementary Data 5: Flagellum / Type III Secretion Data (SuppData5.xlsx)**

**Supplementary Data 6: Summary of Methylation and Codon Usage analysis (SuppData6.xlsx)**