# 3 Pseudogenes in host restricted *Salmonella enterica*

## *3.1 Introduction*

Within subspecies I, hugely variable levels of host specificity and disease outcomes can be found among the 1,547 serovars. Based upon serotyping, there are even groups of serovars that share the same antigenic formula, but still display the variation in host range and severity of disease seen across the subspecies as a whole, e.g. 6,7:c:1,5 (Paratyphi C, and Typhisuis and Choleraesuis) and 1,4,(5),12:Hb:1,2 (Paratyphi B and Java) (Grimont and Weill 2007). Paratyphi C and Typhisuis are restricted to humans and swine respectively (S. Nair, personal communication) while Choleraesuis is strongly adapted to swine, although it is capable of causing a highly invasive disease in humans. Isolates of both Paratyphi C and Choleraesuis have been sequenced, revealing that the formation of pseudogenes is a common feature (Chiu et al. 2005; Liu et al. 2009). Serovars Paratyphi B and Java are distinguished on the basis of a single metabolic test, *d*-tartrate fermentation, yet Paratyphi B is believed to cause paratyphoid fever in humans, and Java mild gastroenteritis (Han et al. 2006).

Another lineage displaying differential host specificity includes Gallinarum, which is restricted to chickens. By multi locus enzyme electrophoresis (MLEE), Gallinarum forms a related cluster with host-generalist Enteritidis and bovine-adapted Dublin (Boyd et al. 1993). The genome sequence of Gallinarum has also been completed, again revealing a significant level of degradation, with over 300 pseudogenes (Thomson et al. 2008).

In recent years, as more genome sequences have become available for *S. enterica*, genome degradation has emerged as a common theme for both host-adapted and

restricted serovars. This phenomenon is also seen in other bacteria, including *Mycobacterium, Shigella* and *Yersinia* (Chain et al. 2004; Cole et al. 1998; Wei et al. 2003).

Prior to the advent of genome sequencing, *Salmonella* serovars were differentiated by the metabolic reactions they could or could not perform, and there were hints of an association between reduced metabolic capacity and a narrow host range (Winslow et al. 1919). Until recently however, there have not been the tools available to investigate whether the effect of extensive pseudogene formation is responsible for the lack of metabolic capability displayed by host restricted *Salmonella*.

By using the metabolic pathway databases described in Chapter 2, pseudogenes from multiple host-restricted serovars can now be identified and their effect upon metabolic capability examined. Further advances in technology now allow multiple metabolic phenotypes to be assayed at once, which aid in describing the full metabolic potential of serovars regardless of host range. Comparing the loss-of-function phenotypes found in host-restricted serovars with the fuller metabolic potential observed in host-generalists has begun to uncover the pseudogenes responsible.

## 3.2 Methods

### 3.2.1 Strains

The serovars and related genome sequences used in the pseudogene comparison and for Biolog phenotyping are given in Table 3-1. The Typhimurium and Typhi strains are described in greater detail in the following chapter.

**Table 3-1 Strains**

| Serovar | Strain (pseudogenes) | Strain (Biolog) | Host | Reference |
|---|---|---|---|---|
| Typhimurium | n/a | SL3261 | Multiple | This dissertation |
| Typhi | multiple* | WT174† | Human | (Holt et al. 2008)<br>(Langridge et al. 2009b) |
| Paratyphi A | ATCC 9150 | nd | Human | (McClelland et al. 2004) |
| | AKU 12601 | | Human | (Holt et al. 2009b) |
| Gallinarum | 287/91 | 287/91 | Chicken | (Thomson et al. 2008) |
| Typhisuis | 61-6 | 61-6 | Pig | This dissertation |

* Pseudogenes analysed from 19 Typhi strains in total, including CT18 (Parkhill et al. 2001a); † This is an attenuated strain of Typhi classified as hazard group 2; nd, not done as an attenuated hazard group 2 Paratyphi A was not available (normally hazard group 3).

### 3.2.2 Sequencing of a Typhisuis reference strain

The Typhisuis 61-6 genome sequencing and initial assembly was performed by Craig Corton.

Typhisuis 61-6 (kindly donated by S. Nair) was sequenced as a reference strain, using two second-generation sequencing platforms: 454 Roche GS FLX Titanium and the Illumina Genome Analyzer II. The sequence data produced by 454 Roche consisted of a paired end library with a 3 kb insert, generating 232,241 reads. The Illumina platform used a 200-300 bp standard paired end library and was run in one lane on a flow cell

generating 18,525,268 37 bp reads. The theoretical sequence depth coverage for the Roche 454 and Illumina platform was 10x and 146x respectively.

The Illumina sequences were assembled using Velvet, a *de novo* short read assembly program (Zerbino and Birney 2008). The parameters used for the assembly were optimised for the Typhisuis dataset and produced an assembly in 421 contigs with an N50 of 20,868 bp, representing approximately 97% of the entire genome (based upon the genome size of Enteritidis P125109). The contigs generated by the Velvet assembly were then combined with the 454 Roche sequences using Roche's GS *de novo* assembler, Newbler. The final combined assembly statistics were 1,041 contigs with an N50 of 36,665 bp. The combined assembly statistics were skewed due to slight contamination in the Roche 454 library with *Yersinia enterocolitica*. However, the majority of this contamination assembled in contigs less than 2 kb in length, and the final assembly contained 210 contigs > 2 kb.

The combined assembly was converted into a Gap4 database (Bonfield et al. 1995) to allow improvement by a round of *in silico* finishing. The 454scaffolds.fna file produced as part of the Newbler assembly output was used to guide gap closure based on the 454 read pair information. ABACAS, a script for ordering and orientating fasta sequences against references (Assefa et al. 2009) was used to align the fragmented Typhisuis assembly against the complete genomes of Enteritidis P125109 and Choleraesuis SC-B67 to help scaffold the contigs where no read pair information was available. This allowed a large number of small repeat regions to be correctly assembled (the cause of the majority of gaps in the sequence), aiding the reduction in contig numbers. After improvement there were 36 contigs > 2kb, containing approximately 99% of the total genome, again

based upon genome length of Enteritidis P125109. These final contigs were ordered and oriented against Choleraesuis SC-B67 using ABACAS. iCORN (Otto et al.) was then used to correct the assembled sequence using the Illumina sequences, primarily to check all homopolymer base discrepancies (errors inherent with the Roche 454 technology) and to highlight any potential problematic regions within the assembly. The corrections made by iCORN were checked and confirmed using the assembly in the Gap4 database.

#### *3.2.2.1 Pseudogene identification and validation*

The corrected sequence aligned against Choleraesuis SC-B67 was used to mark up the position of putative pseudogenes. By comparing the Choleraesuis and Typhisuis genomes using the Artemis Comparison Tool (ACT) (Carver et al. 2005), each Choleraesuis pseudogene was checked in Typhisuis for the same, or different, inactivating mutation. Every Choleraesuis coding sequence was subsequently checked for a possible pseudogene in Typhisuis.

Craig Corton then performed sequence checks of putative pseudogenes.

These were checked against both the 454 and Illumina sequence to determine whether coverage was sufficient to call them pseudogenes.

Maria Fookes performed assemblies of multiple Typhisuis genomes sequenced using Illumina.

Seven strains of Typhisuis were sequenced using the Illumina short read platform. Reads relating to each individual strain were assembled using Velvet (Zerbino and Birney 2008), and contigs were ordered relative to the reference strain 61-6, using ABACAS

(Assefa et al. 2009). All contigs not assembled against the reference were concatenated to the draft genome, which was then used to run a protein BLAST against the reference .

Ambiguous pseudogenes in Typhisuis 61-6 were checked against these seven other Illumina-sequenced Typhisuis genomes in order to determine their validity. A pseudogene was deemed valid if sequence from at least 4 of the other genomes was continuous across the relevant region and consistent with the 61-6 sequence.

### 3.2.3 Whole genome comparison

A full list of pseudogenes was obtained from the genome annotation of Gallinarum (GenBank accession number AM933173). The pseudogene complement of Paratyphi A ATCC9150 was taken from the genome annotation (GenBank CP000026) and extended to include others found by comparison with Paratyphi A AKU 120601 (GenBank FM200053). For Typhi, the pseudogene list was based upon those annotated in CT18 (GenBank AL513382) and again extended based upon comparison with multiple sequenced isolates (Holt et al. 2009b). Pseudogenes in Typhisuis 61-6 were determined as described above.

Pseudogene orthologues between all the genomes were established either from the genome annotation or from reciprocal nucleotide BLAST searches. Genes were deemed to be orthologues if they were the best reciprocal BLAST hit for each other. Frameshifted pseudogenes were typically un-matched, so orthologues were determined manually by checking for conservation and synteny between genomes using ACT.

### 3.2.4 Functional classification of pseudogenes

The 'lost' functions of pseudogenes in all four serovars were assessed by comparison to intact orthologues annotated in Typhimurium SL1344 (in-house annotation) and Enteritidis P125109 (GenBank AM933172). Functional categories were taken from standard Sanger annotation.

### 3.2.5 Metabolic pathway analysis

Each serovar was matched to either StyCyc or StmCyc depending upon which contained the highest number of orthologous pseudogenes. The list of appropriate pseudogene orthologues was overlaid on the relevant metabolic map in order to visualise where pseudogenes were interrupting pathways and transport reactions. All pathways containing pseudogenes were recorded for each serovar. Where possible from the literature, regulators for pathways were identified and checked against pseudogene lists.

### 3.2.6 Biolog phenotyping

The Biolog Phenotype MicroArray (PM) assays (Technopath, Ballina, Ireland) for some of the strains were carried out by Theresa Feltwell, as indicated in Table 3-2.

Each strain was assayed in triplicate, with PM plates 1 and 2A, which contain 192 individual carbon sources. All strains were assayed at the body temperature of their natural host: 37 $^{o}$C for mammalian hosts and 42 $^{o}$C for avian (Gallinarum).

**Table 3-2 Metabolic phenotyping**

| Serovar | Strain details | Additives | Temperature |
|---|---|---|---|
| Typhi | WT174 | Aro mix | 37 °C |
| Typhimurium | SL3261 | none | 37 °C |
| Gallinarum* | 287/91 | Nicotinic acid and thiamine | 42 °C |

*This strain and the third replicate of Typhi and Typhimurium were phenotyped by Theresa Feltwell; aro mix is comprised of 40μg/mL each of L-phenylalanine and L-tryptophan and 10μg/mL each of *p*-aminobenzoic acid and 2,3-dihydroxybenzoic acid, final concentration.

### *3.2.6.1 Preparation of strains to be phenotyped*

Luria-Bertani (LB) agar plates were inoculated with frozen stocks (stored at -80°C) of each strain and incubated overnight at 37 °C. New LB agar plates were inoculated from the overnight growth and again incubated overnight at 37 °C. Colonies from these plates were used to inoculate the PM plates.

### *3.2.6.2 Preparation of PM Inoculating Fluids*

125 mL 1.2x IF-0 (Technopath) was added to 25 mL distilled $H_20$, of which 16 mL was aliquoted into a sterile, capped test tube. A separate 'IF-0 + dye' mixture was prepared by adding 1.8 mL of dye mix (Technopath) and 23.2 mL distilled $H_20$ to 125 mL 1.2x IF-0. Where required, 'aro mix' was added at 1 in 100, and nicotinic acid and thiamine at 0.5 μM, final concentration (see Table 3-2 for serovar-specific supplements). 50 mL aliquots were pipetted into sterile glass bottles for plates PM-1 and PM-2A .

#### *3.2.6.3 Preparation of cell suspension*

Colonies were taken from LB agar plates using a sterile swab and transferred into the sterile capped tube containing 16 mL IF-0. Cells were added to this suspension until a 42%T (transmittance) had been achieved according to the Biolog Turbidimeter (Technopath). 10 mL of the 42%T cell suspension was added to the vial containing 50 mL 'IF-0 + dye' and gently mixed to a final cell density of 85%T.

#### *3.2.6.4 Inoculation of plates PM-1 and PM-2A*

22 mL of the 85%T cell suspension was transferred to a sterile reservoir and used to inoculate PM-1 and PM-2A at 100 µL / well. The PM plates were incubated for 48 hours at 37 $^{o}$C or 42 $^{o}$C, and readings were taken every 15 minutes. The readings measured the strength of the dye colour in each well, which increased as the substrate in that well was utilised.

### 3.2.7 Analysis of Biolog phenotypes

#### *3.2.7.1 Phenotype calling*

Raw data from all 3 replicates of the two PM plates for each strain were loaded into the OmniLog File Management software (version 12.0, Technopath). The value determining the area under the curve (AUC) was taken as a representative parameter for the 48 hour timepoint and was exported into comma-separated files. The AUC values were averaged across replicates and plotted on a frequency distribution to determine an arbitrary cutoff

below which the strain was deemed unable to utilise the relevant substrate. This value was placed at 100, which correlated well with the kinetic graphs produced by the OmniLog software. Thus, substrates where the AUC value exceeded 100 at 48 hours were deemed positive metabolic phenotypes.

#### *3.2.7.2 Differential phenotypes*

As above, the raw data from all 3 replicates were loaded into the OmniLog File Management Software (Technopath). In this instance, the AUC values were exported for every timepoint, which covered measurements taken every 15 minutes from 0 to 48 hours. A pair-wise analysis was used to compare both Gallinarum and Typhi with Typhimurium.

Lars Barquist performed this analysis in R, using `limma`.

The data were read into R using the `PMarray` package (unpublished, L. Barquist). Each well was modelled with a spline-fit curve using `grofit` (Kahm et al.). `limma` (Smyth 2004) was then used to assign Benjamani-Hochberg corrected *P*-values to the log fold differences in AUC values between strains. A cutoff of $P < 0.001$ was used, corresponding to a false discovery rate of 0.1%.

## 3.3 Results

A summary of the total number of pseudogenes determined for each genome or set of genomes analysed is given in Table 3-3. The genomes of Typhi, Paratyphi A and Typhisuis displayed similar levels of genome degradation of ~200 pseudogenes each, with Gallinarum showing the greatest loss of gene function with over 300 pseudogenes.

**Table 3-3 Pseudogenes across host restricted *Salmonella***

| Serovar | Strain | Pseudogenes | Genome size |
|---|---|---|---|
| Typhi | multiple | 211 | 4.8 Mbp |
| Paratyphi A | multiple | 187 | 4.6 Mbp |
| Gallinarum | 287/91 | 306 | 4.65 Mbp |
| Typhisuis | 61-6 | 190 | 4.65 Mbp |

An analysis of the functions lost from each of these serovars revealed that the nature of the genome degradation seen was broadly similar across them all (Figure 3-1). Approximately one third of pseudogenes are found in membrane/surface structures, and alongside loss of function in regulation and pathogenicity/adaptation/chaperones account for at least half of the genome degradation. Pseudogenes that represent conserved hypothetical proteins and proteins of unknown function make up the next largest proportion, indicating that there may be functions relating to the top 3 categories that have yet to be elucidated.

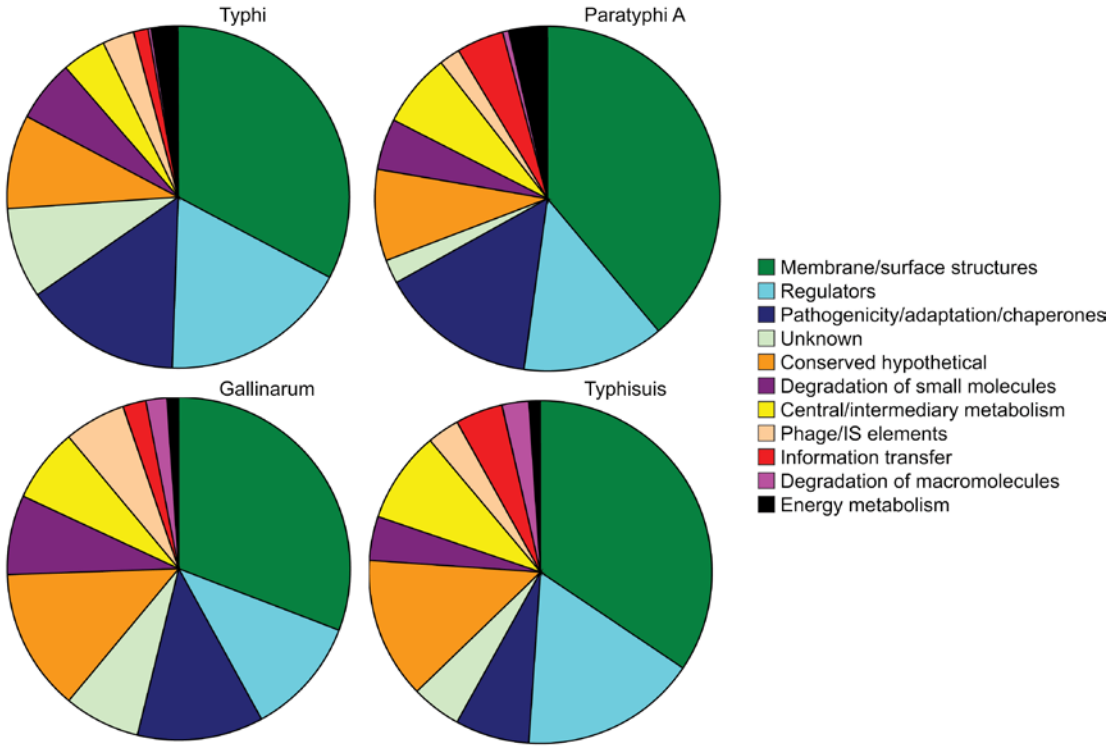


**Figure 3-1 Functional classification of pseudogenes**

Pseudogene function determined based upon comparison with intact orthologues in Typhimurium and Enteritidis. Categories and colours based upon genome annotation.

### 3.3.1 Individual pseudogenes

Across the four genomes of host restricted *Salmonella*, there are only 2 individual pseudogenes that are shared between them. Both are intact in the host generalist Typhimurium. The first is *sopA*, encoding a secreted effector protein known to be translocated into eukaryotic cells and to play a role in bovine enteritis in serovar Dublin (Wood et al. 2000). The second, *mglA*, encodes one of the inner membrane components of a galactose transport system and acts as an ATPase specifically stimulated by galactose (Richarme et al. 1993). The inactivating mutations for these genes are different in each serovar (with one exception), indicating a convergent loss of function. The exception is

*sopA* in Paratyphi A and Typhi which has the same inactivating mutations since the gene falls into one of the low divergence regions indicating recombination between the serovars (Didelot et al. 2007).

### 3.3.2 Relating metabolic phenotype to 'pseudo' genotype

A total of 192 carbon sources were tested for their ability to support the growth of Typhi, Typhimurium and Gallinarum (Appendix 8.2.1 and 8.2.2). The range of substrate utilisation is shown in Figure 3-2, indicating that Typhi and Gallinarum are capable of utilising less than two thirds of the number of substrates that support the growth of Typhimurium.

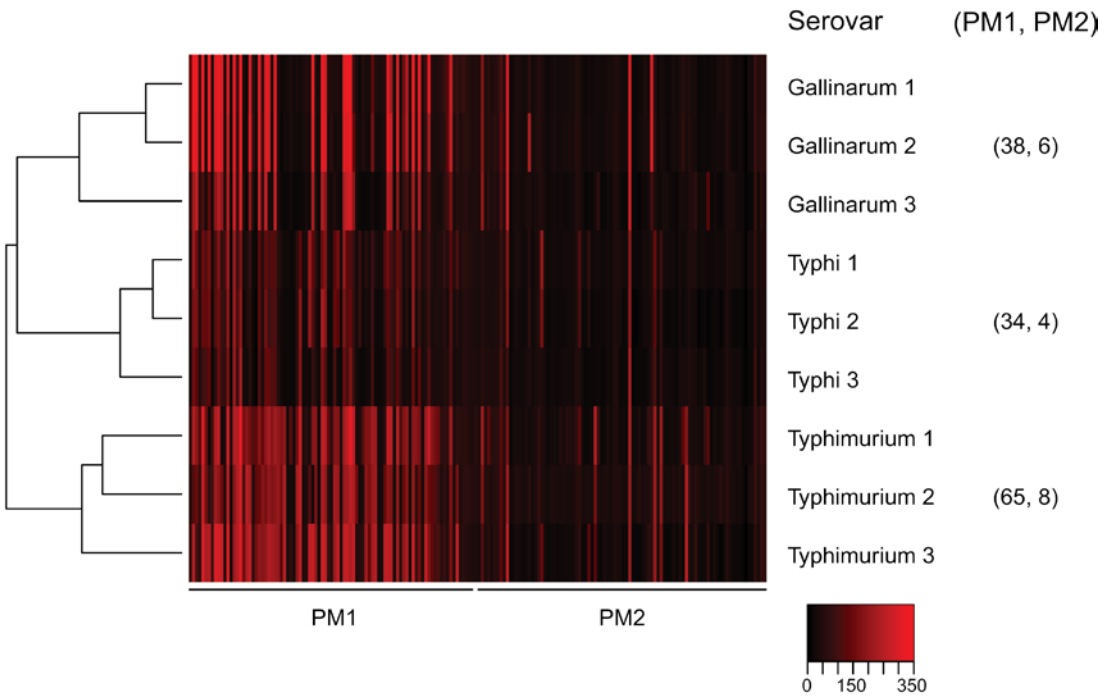


**Figure 3-2 Heatmap of carbon source utilisation**

Drawn using the R heatmap2 package (http://www.r-project.org). Values are a measure of the area under curve (AUC) for dye reduction of each substrate. Higher values (red) indicate capacity to utilise substrate. (PM1, PM2) numbers represent the number of positive metabolic phenotypes per Phenotype MicroArray.

#### *3.3.2.1 Negative metabolic phenotypes*

Both Typhi and Gallinarum yielded negative phenotypes for utilisation of L-ornithine (AUC < 100), a result consistent with the use of this substrate in the API20E test that distinguishes serovars of *Salmonella enterica*. According to StyCyc, L-ornithine is degraded into putrescine by either *speC* or *speF*. The former is constitutive and is inactivated in both Typhi and Gallinarum. The latter is inducible but appears not to compensate for the lack of *speC*, since it remains intact in Gallinarum. Alternatively, Gallinarum may harbour an additional inactivating mutation in the induction mechanism.

#### *3.3.2.2 Differential metabolic phenotypes*

When compared with Typhimurium, differential phenotypes were observed for 29 substrates in Typhi and 30 in Gallinarum ($P < 0.001$). In Typhi, all 29 displayed a negative log fold change (Log FC) but 2 of the phenotypes in Gallinarum showed a slight increase in utilisation relative to Typhimurium (Table 3-4). These were α-D-glucose and D-fructose (Log FCs 0.98 and 1.08 respectively), perhaps indicating the preference of Gallinarum for these carbon sources.

As mentioned in the previous chapter, Typhi contains a pseudogene in *rhaD* that affects its ability to ferment rhamnose. As a sole carbon source, rhamnose would provide energy by its degradation into dihydroxyacetone phosphate which is further utilised in central metabolism. This phenotype was borne out in the Biolog data for Typhi (Log FC -3.4), and for Gallinarum (Log FC -2.2). Gallinarum retains all the genes required for rhamnose

transport and degradation, but the likely cause of the loss of function is a pseudogene in *rhaS*, which encodes the transcriptional activator of the rhamnose operon.

Typhi displayed no growth on L-arabinose (LogFC -1.38, $P < 0.0005$), which would typically sustain growth as a single carbon source by its 3-step conversion to D-xylulose-5-phosphate, an intermediate of the pentose phosphate pathway. All the genes involved in this conversion are intact in Typhi. However, there are two uptake systems for L-arabinose in *E. coli*, a low affinity transporter *araE*, and a high affinity transport complex encoded by *araFGH*. While *araE* is intact in Typhi, only a truncated form of *araH* remains of the high-affinity transporter. Thus, it appears *araE* alone is not sufficient to support growth.

**Table 3-4 Differential Biolog phenotypes**

| Ty LogFC | Ty *P*-value | Substrate | Gal LogFC | Gal *P*-value |
|---|---|---|---|---|
| -1.58 | 0.00017 | acetic acid | | |
| | | alpha-D-glucose | 0.98 | 0.00065 |
| | | alpha-hydroxy butyric acid | -2.14 | 1.91E-05 |
| -2.29 | 8.22E-05 | **alpha-keto-butyric acid** | -2.71 | 2.87E-05 |
| | | alpha-methyl-D-galactoside | -3.25 | 6.08E-06 |
| -2.63 | 7.24E-06 | **bromo succinic acid** | -2.52 | 1.05E-05 |
| -2.99 | 6.03E-06 | **D,L-malic acid** | -3.33 | 6.08E-06 |
| -2.47 | 6.03E-06 | **D-aspartic acid** | -2.62 | 6.16E-06 |
| -1.53 | 0.000429 | dextrin | | |
| | | D-fructose | 1.08 | 0.00073 |
| -2.06 | 1.47E-05 | **D-galatonic acid-gamma-lactone** | -1.52 | 0.00012 |
| -3.19 | 4.05E-06 | **D-glucosaminic acid** | -2.86 | 6.34E-06 |
| | | D-lactic acid methyl ester | -1.77 | 4.09E-05 |
| | | D-melibiose | -2.14 | 4.04E-05 |
| -1.57 | 0.000777 | D-ribose | | |
| -2.81 | 4.63E-06 | **D-saccharic acid** | -2.46 | 9.93E-06 |
| | | D-sorbitol | -2.24 | 9.93E-06 |
| -2.90 | 7.24E-06 | **D-tagatose** | -2.48 | 1.95E-05 |
| -2.81 | 4.05E-06 | **D-tartric acid** | -2.51 | 6.08E-06 |
| -2.67 | 5.99E-05 | **fumaric acid** | -2.62 | 6.70E-05 |
| -1.54 | 8.22E-05 | glycyl-L-aspartic acid | | |
| -2.32 | 4.16E-05 | L-alanine | | |
| -1.38 | 0.000343 | L-arabinose | | |
| -1.88 | 1.14E-05 | **L-asparagine** | -1.89 | 1.08E-05 |
| -2.94 | 4.63E-06 | **L-aspartic acid** | -2.42 | 1.08E-05 |
| -2.36 | 8.87E-06 | **L-fucose** | -1.88 | 4.09E-05 |
| -1.30 | 8.22E-05 | **L-glutamic acid** | -1.06 | 0.00038 |
| -1.63 | 6.11E-05 | **L-glutamine** | -1.66 | 5.61E-05 |
| -1.98 | 4.13E-05 | **L-malic acid** | -1.25 | 0.00096 |
| -3.41 | 1.44E-05 | **L-rhamnose** | -2.19 | 0.00034 |
| | | melibionic acid | -2.62 | 9.93E-06 |
| -1.77 | 5.90E-05 | **m-hydroxy phenyl acetic acid** | -1.64 | 8.87E-05 |

Log fold change (FC) and adjusted *P*-values were calculated with respect to Typhimurium. Substrates (in alphabetical order) in bold are differential phenotypes shared by Typhi (Ty) and Gallinarum (Gal).

D-tagatose was poorly utilised by both Typhi and Gallinarum with respect to Typhimurium (LogFCs -2.9 and -2.5 respectively). Tagatose, a naturally occurring isomer of fructose, is transported into the cell by a phosphotransferase (PTS) system encoded by *tagTH*. Tagatose-1-phosphate is phosphorylated by *tagK* to form tagatose-1,6-bisphosphate which is degraded by the *gatYZ*-encoded aldolase in a reaction that also forms the final step in galactitol degradation (Figure 3-3) (Mayer and Boos 2005; Shakeri-Garakani et al. 2004). The final products in these pathways are D-glyceraldehyde-3-phosphate and dihydroxyacetone phosphate, both ubiquitous metabolites that feed into a variety of metabolic pathways to provide sufficient energy for growth. In Typhi, *gatZ* is inactivated by a premature stop codon, and this also explains why Typhi produced a negative phenotype for galactitol (named dulcitol in the Biolog system). Gallinarum also displayed a negative phenotype for galactitol, but both PTS systems for tagatose and galactitol are intact, as are all the genes depicted in Figure 3-3. The mechanism behind these phenotypes in Gallinarum therefore remains unclear.

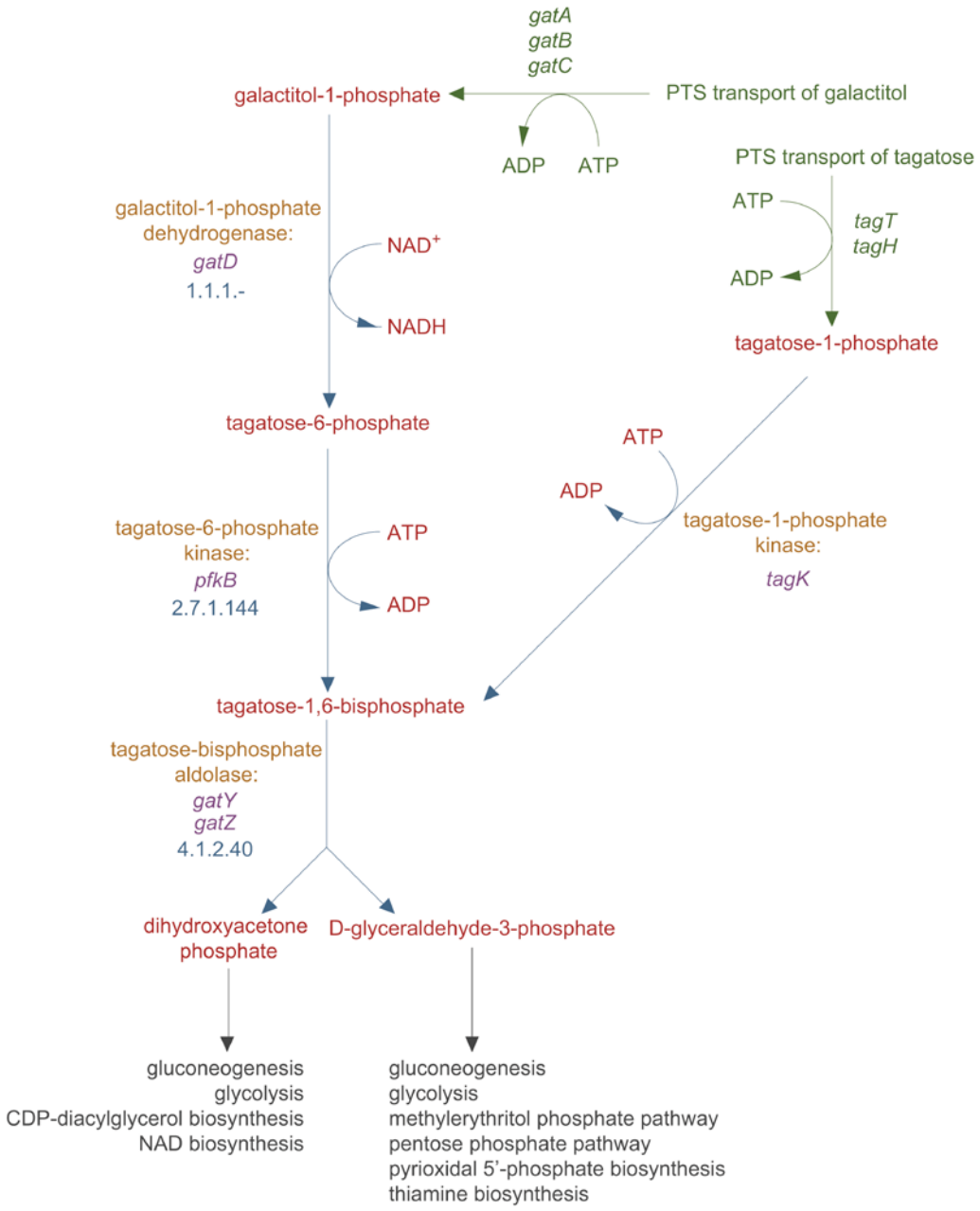


**Figure 3-3 Galactitol (dulcitol) and tagatose degradation**

Metabolic pathway diagram from StyCyc 7.0 demonstrating that the degradation of galactitol and tagatose share the final enzyme, which in Typhi is inactive. Blue arrows indicate enzymatic reactions, substrates shown in red, enzymes in gold, genes in purple and enzyme commission numbers in blue. Green arrows indicate transport reactions, substrates and genes also shown in green. Grey arrows and names represent pathways into which dihydroxyacetone phosphate and D-glyceraldehyde-3-phosphate subsequently feed into.

Gallinarum alone displayed a differential phenotype for sorbitol (glucitol), with a Log FC of -2.2. Sorbitol is a sugar alcohol transported into the cell via a phosphotransferase (PTS) permease encoded by *srlAEB* and degraded into D-fructose-6-phosphate by *srlD*. Again, all of these genes are intact, but given that this reaction may be used to differentiate Gallinarum from its close relative Pullorum, this suggests there is some other interaction affecting its ability to utilise this substrate.

One deficiency shared between Typhi and Gallinarum is lack of growth on L-glutamine (Log FCs both -1.6). In the Typhimurium strain tested, the utilisation of L-glutamine was relatively slow, with a lag of ~15 hours before reaching the AUC threshold of 100. This phenotype is supported by evidence from *E. coli*, where L-glutamine can support growth as a single carbon source, but growth is slow (McFall and Newman 1996). One multimeric complex has been demonstrated to transport L-glutamine, encoded by *glnQPH*. Gallinarum has a pseudogene in *glnH*, the periplasmic binding protein of the transport complex. Since *E. coli* mutants with enhanced transport are known to grow well, such a defect in this transporter is a highly probable cause of this phenotype. However, *glnQPH* appears intact in Typhi, raising the possibility that it is not a pseudogene causing this phenotype, but that transcription of the intact required genes is not at a high enough level to support growth.

#### *3.3.2.3 Potential for improving annotation*

Another shared phenotype was observed for mucic acid, also known as D-galactarate. Degradation of this substrate is by the products of four *gar* genes and *glxK*, and results in the production of pyruvate and 2-phosphoglycerate. A transporter for mucic acid has been

proven to be present in *E. coli*, but no gene has yet been shown to encode it (Hubbard et al. 1998). Given the shared phenotype, there is a possibility that this transporter is encoded in *Salmonella* by one of the pseudogenes in common between Typhi and Gallinarum. One candidate is STY2501/SG2301 which is a putative transmembrane protein that has a Pfam hit to the Major Facilitator superfamily (Finn et al. 2007). A site-specific mutation/knockout of this gene in Typhimurium could answer this hypothesis, and indicates the potential of high throughput metabolic phenotyping for improving genome annotation.

#### *3.3.2.4 Other metabolic phenotypes*

One metabolic reaction not part of the Biolog system but commonly used in *Salmonella* differentiation is the focus of a test for tetrathionate reduction, and the production of hydrogen sulphide ($H_2S$). Both Gallinarum and Paratyphi A test negative for $H_2S$ production, with Typhi only weakly positive. The $H_2S$ phenotypes can all be explained by pseudogenes present in the tetrathionate reduction pathway (Figure 3-4). Both Paratyphi A and Gallinarum have pseudogenes in tetrathionate reductase, encoded by *ttrABC*, located within SPI-2. In Typhi however, an inactivating mutation occurs in *ttrS*, the sensory element of a two-component regulator with *ttrR* that positively regulates the activity of *ttrABC* (Hensel et al. 1999). It is the activity of the tetrathionate reductase that is usually measured, but further degradation of the pathway has occurred in both Paratyphi A and Gallinarum.

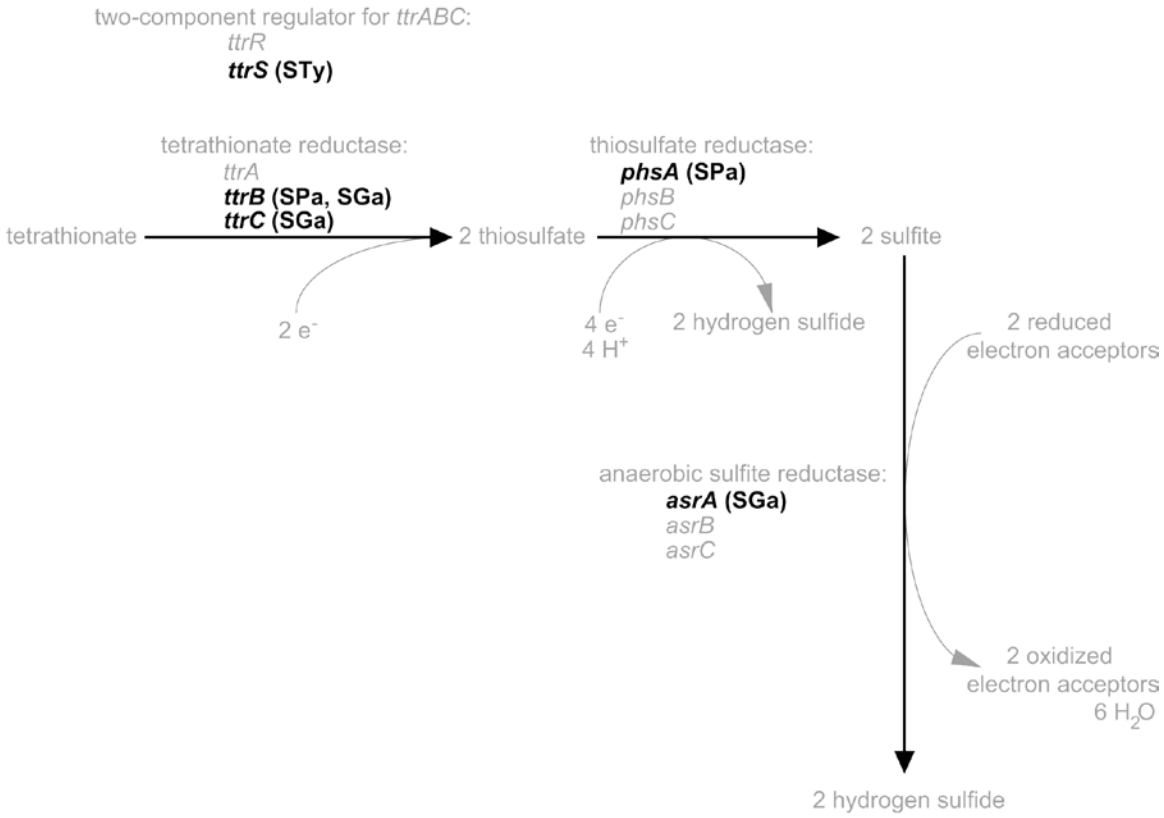


**Figure 3-4 Metabolic pathway of tetrathionate reduction**

Reaction lines and gene names in bold represent pseudogenes in host-restricted *Salmonella*. STy, Typhi; SGa, Gallinarum; Spa, Paratyphi A.

While Paratyphi A could not be tested with the Biolog system (as a hazard group 3 organism), the main reaction used to distinguish this serovar is a negative phenotype for L-lysine. From StyCyc, the genes involved in lysine degradation are *ldcC* and *cadA*, both of which are intact in Paratyphi A. The primary transporter for lysine is encoded by *cadB*, which is also intact, but the most likely cause of this phenotype is the pseudogene present in *cadC*, the transcriptional activator of *cadAB*.

### 3.3.3 Shared interruption of metabolic pathways

While few individual pseudogenes are shared between all serovars, similar losses of metabolic capability are found and can be explained by different pseudogenes interrupting the same metabolic pathway.

#### *3.3.3.1 Pathways inactivated in all host restricted serovars*

The anaerobic biosynthesis of vitamin B12 has been noted previously as inactive in both Typhi and Paratyphi A via mutations in different *cbi* genes. Both Gallinarum and Typhisuis also have pseudogenes in this pathway, shown in Figure 3-5.

Typhi has three and Gallinarum contains two pseudogenes that halt the addition of the eight methyl groups required to form the corrin ring in cobyrinate, while it is the enzyme required for the subsequent amidation of cobyrinate to cobyrinate *a,c*-diamide that is inactivated in Paratyphi A. Cobalt ions are required during biosynthesis and are provided from an extracellular location via a transport complex of CbiMNOQ. Typhisuis, Gallinarum and Typhi have different pseudogenes in this complex and given the requirement for these cobalt ions, it is likely that the loss of transport alone inactivates the pathway. It is therefore possible that the other pseudogenes in Gallinarum and Typhi were inactivated after this loss of function, when any selective pressure to maintain the remainder of the pathway would also have been lost.

The latter stages of vitamin B12 biosynthesis involve the *cob* genes, in which a further pseudogene occurs in Gallinarum (*cobD*).

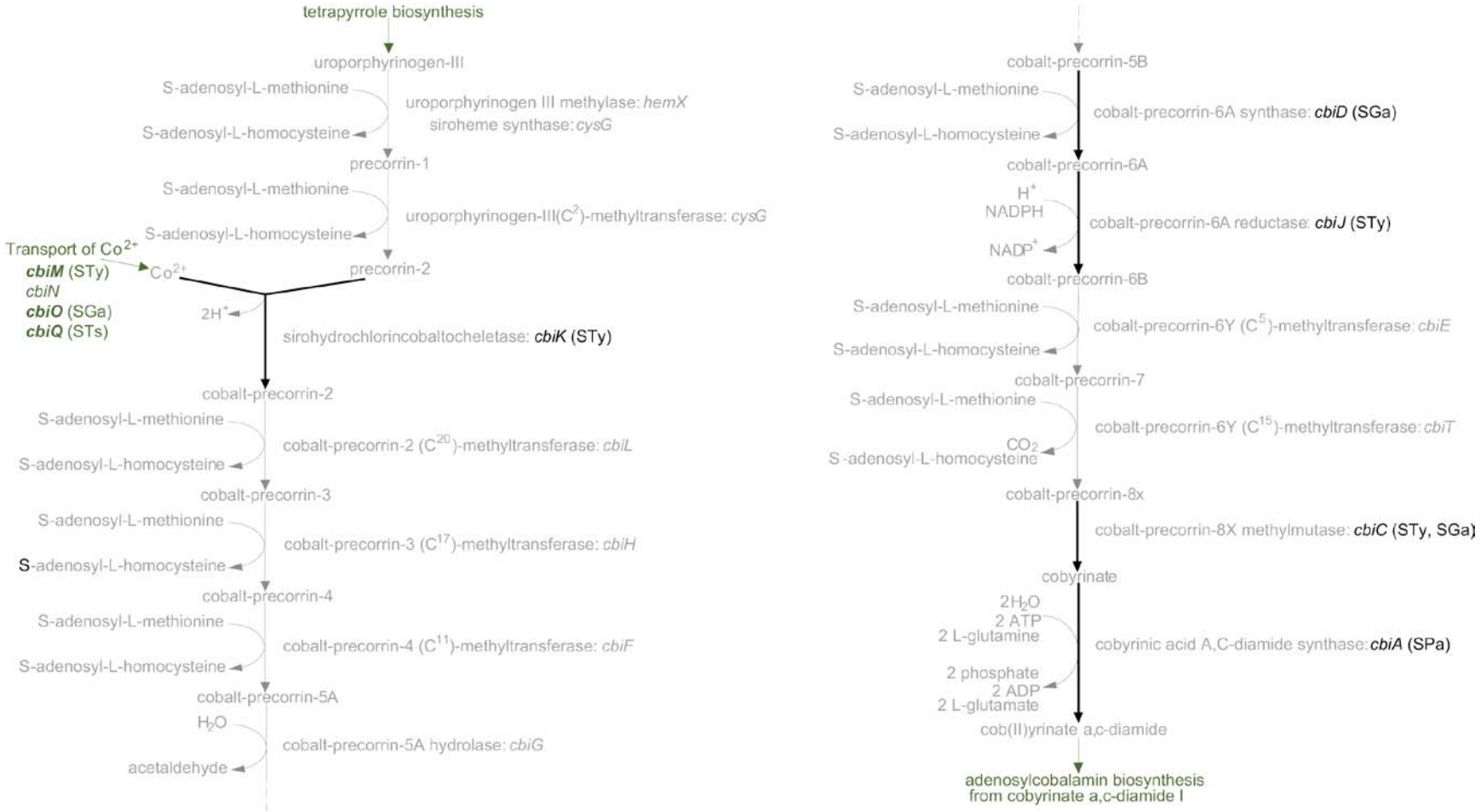


**Figure 3-5 Early stages of vitamin (coenzyme) B12 biosynthesis**

Reaction lines and gene names in bold represent pseudogenes in host-restricted *Salmonella*. STy, Typhi; SGa, Gallinarum; STs, Typhisuis; Spa, Paratyphi A. Green lines and names represent links to/from other metabolic pathways and transporters.

A related process is the degradation of L-1,2-propanediol (PDL), which is dependent upon the presence of vitamin B12. The metabolic pathway is shown in Figure 3-6 and each of the genomes analysed contains pseudogenes that affect PDL degradation.

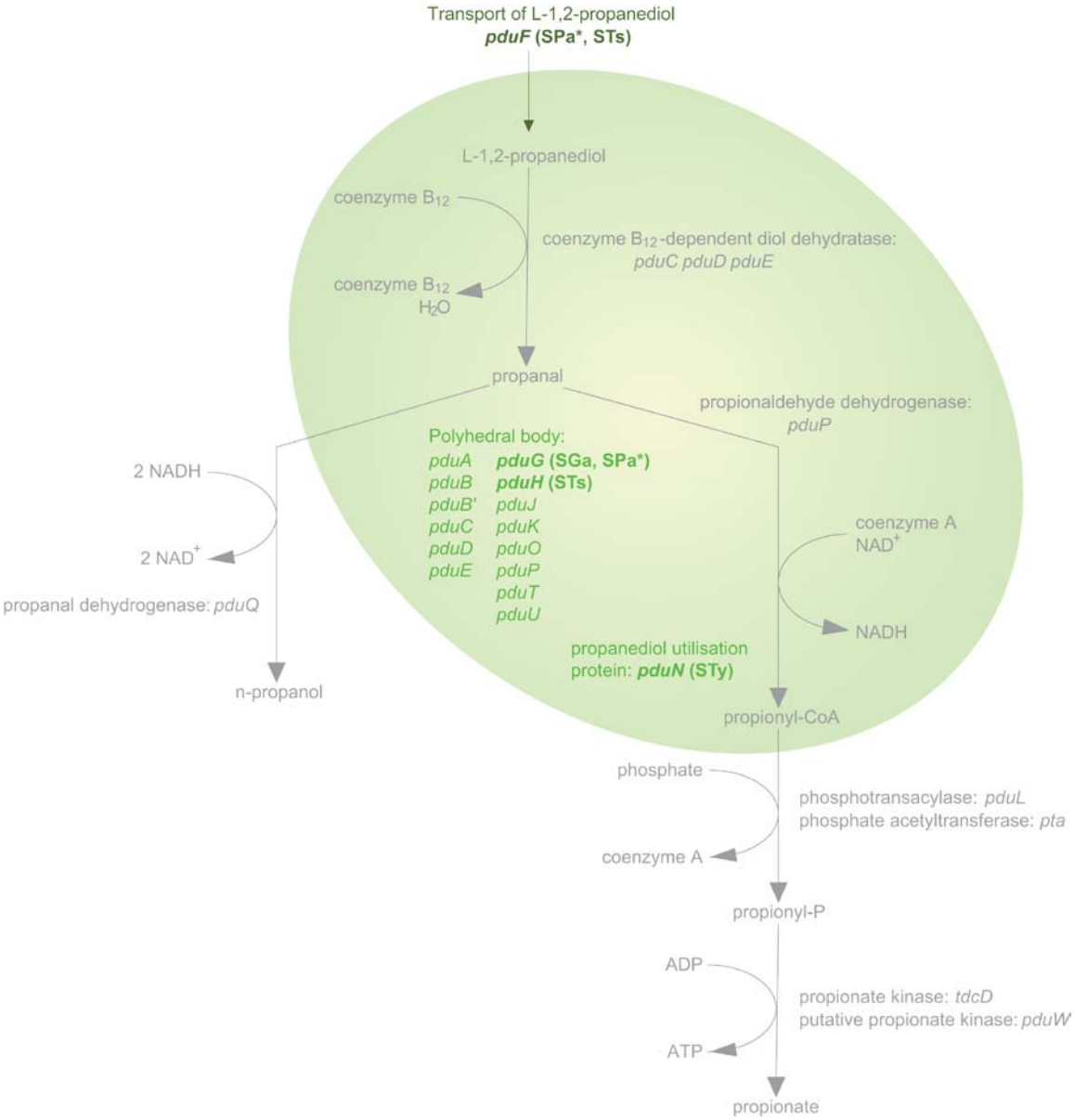


**Figure 3-6 Metabolic pathway for 1,2-propanediol degradation**

Gene names in bold represent pseudogenes in host-restricted *Salmonella*. STy, Typhi; SGa, Gallinarum; STs, Typhisuis; Spa, Paratyphi A. *, in the two sequenced Paratyphi A strains, one harbours a pseudogene in *pduF*, the other in *pduG*. Reactions inside the green ellipse occur inside the polyhedral body; gene names in bright green are those involved in its structure and formation.

Both Paratyphi A and Typhisuis contain mutations in *pduF*, which encodes the transporter protein for PDL. The two initial steps of the pathway are believed to occur inside the proteinaceous PDL degradation polyhedral body, which protects the cell from aldehyde toxicity (Havemann and Bobik 2003). Paratyphi A, Gallinarum and Typhisuis pseudogenes occur in *pduG* and *pduH*, which together perform the reactivation of diol dehydratase in the polyhedral body. In addition, PduG may also be involved in the adenosylation of vitamin B12 (Bobik et al. 1999). In Typhi, *pduN* is inactivated, which encodes a close relative of the CcmL-CchB family of proteins required for the proper assembly and function of carboxysomes (Bobik et al. 1999). Hence, *pduN* likely functions to aid the formation of the polyhedral bodies.

#### *3.3.3.2 Pseudogenes inactivating pathways in human-restricted serovars*

Typhi and Paratyphi A have in common two inactivated pathways that show no pseudogenes in the other host-restricted serovars. The first is the degradation of 4-hydroxyphenyl-acetate (4-HPA), which was identified in *Klebsiella* as being degraded into the TCA cycle intermediates pyruvate and succinate, thus providing energy as a carbon source (Martin et al. 1991)(Figure 3-7). 4-HPA is a metabolite produced during tyrosine degradation in some bacterial species, although this process has not been demonstrated to occur in *Salmonella* or *E. coli*.

The two pseudogenes in Typhi disrupt the pathway at the beginning and the one in Paratyphi A in the middle. Interestingly, while all the genes in this pathway, and in the associated transporter (encoded by *hpaX*) are intact in Gallinarum, it displayed a

differential phenotype (LogFC -1.64) equivalent to Typhi (Log FC -1.77, suggesting that 4-HPA utilisation is also inactivated in this serovar.

Aromatic compounds like tyrosine are abundant in soil and water, so the likelihood that these will be utilised as sole carbon sources is quite high, but only for bacteria that have a substantial life cycle in the environment (e.g. *E. coli*) (Diaz et al. 2001). It has been postulated that similar compounds can also be found in the animal gut, as the product of various degradative reactions by the intestinal flora, but it is possible that the systemic infection caused by host-restricted *Salmonella* means that the organism does not remain long enough in the gut to utilise any 4-HPA that may be available and therefore no longer retains the use of this pathway.

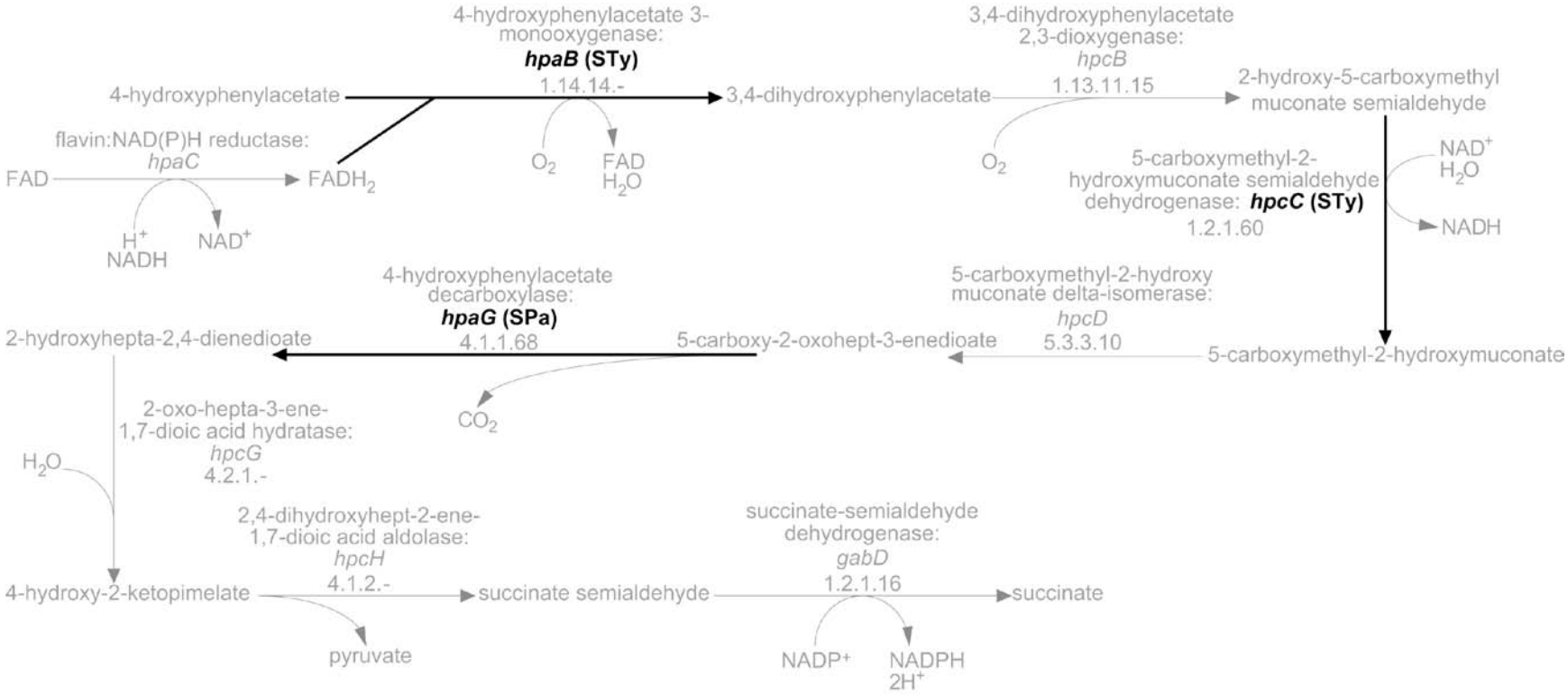


**Figure 3-7 Metabolism of 4-hydroxyphenylacetate**

Gene names and reactions in bold represent pseudogenes in Typhi (STy) and Paratyphi A (SPa).

One pathway that is not tested by the Biolog system and was therefore only analysed by pseudogene comparison is colanic acid biosynthesis, encoded by the *wca* gene cluster. Typhi contains pseudogenes in *wcaA*, *wcaD*, and *wcaK* while Paratyphi A harbours inactivating mutations in *wcaJ* and *wcaK*. The particular effect of this loss of function may be understood from the work that has been done on this pathway in *E. coli*, where colanic acid is made primarily at low temperatures and therefore believed to be important for survival in the environment rather than in the animal host (Ryu and Beuchat 2004; Whitfield and Keenleyside 1995). Host-restricted serovars have little theoretical need for a capsule that promotes long-term survival outside the host.

### 3.3.4 Loss of function in transport across the membrane

By overlaying pseudogenes onto StyCyc and StmCyc, it was possible to determine which broad functional areas (as defined by Pathway Tools) contained the highest levels of genome degradation (Table 3-5). Displaying the data in this manner indicated that the dominance of pseudogenes in membrane/surface structures (shown earlier in Figure 3-1), could be largely explained by those encoding transporter proteins as in each serovar, between 10 and 20% of pseudogenes occur in transporters. A large number of these occur in sugar transport systems, perhaps reflecting the reduced availability of many carbon sources in the restricted host environment.

**Table 3-5 Metabolic functions of pseudogenes**

| | Typhi | | Paratyphi A | | Gallinarum | | Typhisuis | |
|---|---|---|---|---|---|---|---|---|
| **Function** | Reactions with a pseudogene (possible redundancy[a]) | % of total | Reactions with a pseudogene (possible redundancy[a]) | % of total | Reactions with a pseudogene (possible redundancy[a]) | % of total | Reactions with a pseudogene (possible redundancy[a]) | % of total |
| **Biosynthesis** | 4(1) | 1% | 23(12) | 12% | 14(8) | 4.5% | 10(7) | 6% |
| **Energy** | 1(1) | 1% | 1(1) | 0.5% | 5(5) | 1.5% | 4(4) | 2% |
| **Degradation** | 5(2) | 5% | 12(3) | 6% | 13(5) | 4% | 11(2) | 6.5% |
| **Transport** | 18(-) | 15% | 19(-) | 10% | 34(-) | 11% | 33(-) | 20% |
| **Standalone** | 11(7) | | 10(4) | | 20(6) | | 10(3) | |
| **Regulator[b]** | 37(-) | | 23(-) | | 34(-) | | 31(-) | |
| **Total pseudogenes** | 211 | | 187 | | 306 | | 167 | |
| **Metabolic[c]** | 35 | | 39 | | 51 | | 43 | |

[a], pseudogenes for reactions which retain an intact enzyme were recorded as possibly redundant. Enzyme redundancy was determined from StyCyc and StmCyc. [b], pseudogenes in regulators were identified from genome annotation. [c], represents the actual number of pseudogenes present in the relevant metabolic map from StyCyc and StmCyc. This number may vary from the number of reactions with a pseudogene as one enzyme may catalyse multiple reactions. Typhi and Paratyphi A were mapped to StyCyc, the remainder to StmCyc.

Shared across all serovars are pseudogenes in iron uptake systems (Table 3-6). *Salmonella* contain multiple complexes for obtaining iron, which is required for growth (Earhart 1996). Inactivation of *fhuE* was the most common, occurring in three of the four serovars. FhuE is a receptor protein for both ferric-coprogen and ferric-rhodotorulic acid (Hantke 1983), which are both produced by various fungal species, hence it is unlikely that a host-restricted serovar comes across these potential iron sources. Inactivation of the *fhu* complex would therefore not be associated with a biological cost.

**Table 3-6 Pseudogenes in iron uptake systems**

| Serovar | Iron uptake system pseudogene(s) |
|---|---|
| Typhi | *fepE*, *fhuA*, *fhuE* |
| Paratyphi A | *fhuA*, *fhuE* |
| Gallinarum | *iroD* |
| Typhisuis | *iroD*, *fhuE* |

Both Gallinarum and Typhisuis contain an inactivated *iroD*, one of two hydrolases present in the siderophore salmochelin uptake system. IroD acts to cleave salmochelin into several substrates, allowing the bound iron to be reduced and removed (Zhu et al. 2005). Typhi also contains a pseudogene in *fepE*, but this does not appear to affect the activity of the *fepBCDG* complex for uptake of another siderophore (enterobactin) that was found to be essential under laboratory conditions (see Chapter 4 (Langridge et al. 2009b)).

## 3.4 Discussion

Comparison of pseudogenes across host-restricted *Salmonella* has been concentrated in the past upon individual pseudogenes commonly shared. While these are likely to be important, expanding this comparison to look at commonly inactivated metabolic pathways provides a wider context in which to interpret the presumed loss of function. Examining these pathways in relation to the role they play when intact in host-generalists also gives an insight into how the loss of function may affect these serovars.

In the Biolog experiment, each serovar was grown at the temperature associated with their natural host; 37 $^{o}$C for mammalian Typhimurium and Typhi, and 42 $^{o}$C for avian Gallinarum. Previous work in our laboratory (unpublished) had shown that another host-associated serovar tested at varying temperatures was the most metabolically active at the temperature associated with the host. The temperatures chosen for the serovars in this study were therefore expected to be those the serovars were best adapted to, and yield the most informative results. This approach was validated by the number of metabolic phenotypes that were in accordance with those currently in use for the identification of Enterobacteriaceae, and *Salmonella* serovars in particular.

Other phenotypes were also identified, with Gallinarum showing a negative reaction for growth on L-glutamine, which could be traced to a pseudogene in the periplasmic binding protein of the transport system. While Typhimurium may not be the optimum serovar to use as a control for pseudogene formation, it did provide a basis for determining relatively straightforward phenotypes. In this study, it was shown that Typhimurium was capable of utilising over twenty substrates that neither Typhi nor Gallinarum could.

A number of the differential phenotypes displayed by Typhi and Gallinarum, with respect to Typhimurium, could be linked to 'pseudo' genotypes. Unsurprisingly, given the large proportion of pseudogenes that occur in genes encoding transporter proteins, mutations that affect the cells ability to take up substrates were often the cause of growth deficiency. However, the rhamnose example in Gallinarum showed the importance of being able to identify which genes are involved in regulation of metabolic pathways or transport reactions. Here, a pseudogene in the transcriptional activator of the rhamnose operon was the most likely cause of Gallinarum's inability to utilise rhamnose. Inactivation of transport and regulatory genes also circumvent issues of pathway intermediates accumulating inside the cell, to possible harmful effects.

The Biolog Phenotype MicroArrays represent a high-throughput system for identifying metabolic phenotypes, and analysed in conjunction with the metabolic pathway databases of StyCyc and StmCyc allowed the metabolic effect of pseudogenes to be better understood.

However, there were some phenotypes that could not be directly explained with a relevant pseudogene. In order to gain further understanding of the functional relevance of these losses of function, and perhaps uncover the genotypic cause, these results should be examined in the context of the relevant evolutionary lineage. For example, the effect of pseudogenes in Gallinarum would ideally be assessed in the context of Enteritidis, the closest non-adapted relative. Obtaining Biolog data for Enteritidis would be the first step in such an analysis and is currently under way.

Not all metabolic functions can be or are tested by the Phenotype MicroArrays. Therefore, comparing pseudogenes from host-restricted serovars across metabolic

pathways complements and extends the high throughput approach. The most striking pathway inactivation seen in all the host restricted serovars analysed in this study, is that involving vitamin B12 biosynthesis and 1,2-propanediol (PDL) degradation.

Originally, *Salmonella* was documented to grow on PDL as a sole carbon source aerobically but not under anaerobic conditions. However, its degradation requires adenosylcobalamin (vitamin B12) in the first enzymatic step, and this is only produced anaerobically. This paradox was solved when it was discovered that using tetrathionate as an alternative electron acceptor allows anaerobic growth on PDL, making vitamin B12 biosynthesis essential for this process, and revealing that *Salmonella* therefore require 40-50 genes (including the *cbi*, *cob*, and *pdu* genes) for PDL degradation (Bobik et al. 1999; Price-Carter et al. 2001). Both the *pdu* and *cob* genes are induced by PDL, implying that PDL utilisation is the primary reason for maintaining vitamin B12 biosynthesis. Other evidence to support this comes from the evolutionary history of this metabolic interaction. It has been proposed that the ancestor of most enteric bacteria could synthesise vitamin B12 and hence degrade PDL, but that these abilities were lost in the lineage leading to *E. coli* and *Salmonella* (Roth et al. 1996). *Salmonella* however, subsequently acquired the *pdu* and *cob* genes as a single chromosomal fragment and regained the ability to synthesis vitamin B12 and degrade PDL (Roth et al. 1996). Hence, if the ability to degrade PDL provides the selective pressure for maintaining such a large number of genes, then this pressure is apparently no longer exerted upon the genomes of host-restricted serovars. Unravelling the reasons behind this may shed more light on the host-pathogen interaction.

A pathway inactivated only in the human-restricted serovars was colanic acid biosynthesis, carried out by proteins encoded in the *wca* cluster. In Typhimurium, a *wcaE* mutant was shown to be attenuated for intestinal colonisation in calves and chicks, although the authors suggested such a mutant may also be more susceptible to acid stress (Morgan et al. 2004). Biofilm formation has also been connected with colanic acid production, where evidence has been presented indicating the *wca* genes are required for biofilm formation on HEp-2 cells, mammalian tissue culture cells and in the chicken intestinal epithelium (Ledeboer and Jones 2005). It is possible then that the inability to produce colanic acid is of twofold consequence: one, Typhi and Paratyphi A have a reduced need for a capsule providing long-term protection in the environment since they are host-restricted, and two, a reduced ability to colonise the intestine would correlate with evidence that these bacteria act as 'stealth' pathogens that cross the intestinal epithelium without causing inflammation and diarrhoea.

A pathway level analysis of the effects of pseudogenes is an important step towards understanding these effects upon the interconnected network of bacterial metabolism. This study has shown the importance of transport and of regulation in determining how a metabolic phenotype may be caused by a 'pseudo' genotype. A systems-level description of *Salmonella* metabolism would be the optimum scale at which to examine the metabolic activity of host restricted and host generalist serovars. But before this can be put in place, further curation is required to associate regulators and transporters with metabolic pathways, and experimental validation of any new pathways predicted from growth upon particular substrates in the Biolog system.

## 3.5 Conclusions

This is the first *Salmonella*-specific pathway analysis of the effect pseudogenes have upon metabolism. By putting pseudogenes from host-restricted serovars into context, it has been shown that there are more shared characteristics between host-restricted *Salmonella* than found by comparing individual genes. Use of a high throughput system like Biolog is advantageous for large-scale metabolic phenotyping, and there are possibilities that in the future, substrate analysis could be directly linked to the relevant pathway. The importance of both regulation and transport is key and curation efforts should focus upon improving StyCyc and StmCyc in that regard. This will help to untangle the mechanisms behind the negative metabolic phenotypes that cannot be explained directly by pseudogenes related to the substrate.