Results V1

# Results

Known biofilm genes were compared with all four *Acidihalobacter* assemblies (refs). Consequently, a comprehensive literature search was performed to identify known candidate genes (Supplementary: table with gene IDs). We found 1039 candidate genes among 8 families, none of which belonged to Ectothiorhodospiraceae.

Orthofinder (get version from Philipp) was used to arrange the candidate and *Acidihalobacter* proteomes. Subsequently, the 1039 candidate genes were compared with the four *Acidihalobacter* genomes. 779 orthogroups were identified corresponding to the 1039 candidate genes, of which, 528 candidate genes (50.8%) are present in all *Acidihalobacter* genomes (Figure 1). Conversely, 337 (32.4%) candidate genes are not present in any of the *Acidihalobacter* genomes.

Diagram, venn diagram

Description automatically generated

Figure : Venn diagram showing core, dispensable, and unique candidate genes in Acidihalobacter assemblies

## Orthogroups with more than three candidate genes

First to be investigated were orthogroups with the largest number of candidate genes. Eleven orthogroups were identified that contained between four and nine candidate genes each (Supplementary: table with orthologues).

The largest orthogroup contained nine overlapping candidate genes representing sigma-54 dependent transcriptional regulator (FlhA, MifR, and AlgB) from *Acidithiobacillus thiooxidans, Burkholderia pseudomallei, Leptospirillum ferriphilum, Leptospirillum ferrooxidans, Escherichia coli,* and *Pseudomonas aeruginosa*. This orthogroup is present in all *Acidihalobacter* species and demonstrates substantial expansion with eleven copies in *A. aeolianus* and *A. prosperus*, ten copies in *A. yilgarnensis*, and nine copes in *A. ferrooxydans*.

The following two orthogroups contained eight and five overlapping candidate genes respectively representing glycosyltransferases (WapR, MigA, and PslI) from *B. pseudomallei, L. ferriphilum, L. ferrooxidans,* and *P. aeruginosa*. The first orthogroup is absent in all *Acidihalobacter* species except for *A. ferrooxydans* which contains two copies. The second orthogroup is present in all *Acidihalobacter* species except for *A. ferrooxydans*, with a single copy in *A. aeolianus* and *a. yilgarnensis*, and two copies in *A. prosperus*.

An orthogroup containing five overlaps representing PilZ type IV pilus assembly all from *A. thiooxidans* were absent from all *Acidihalobacter* species. Similarly, two orthogroups containing four overlaps each representing a set of porin proteins (OprE) all from *P. aeruginosa* and a TonB-dependent receptor from *S. oneidensis* were also absent from all *Acidihalobacter* species.

Two orthogroups contained four overlapping candidate genes each, representing flagellar hook protein (FlgE) from *E. coli* and *Shewanella oneidensis* and flagellar motor switch (FliM) from *B. pseudomallei, E. coli, L. ferrooxidans,* and *Stenotrophomonas maltophilia*. Both orthogroups are present in all *Acidihalobacter* species with a single copy in each.

The remaining three orthogroups contained four overlapping candidate genes each. The first represents lipopolysaccharide glucosyltransferase I from *E. coli* and *L. ferriphilum* and is present in all *Acidihalobacter* species and demonstrates substantial expansion with six copies present in *A. aeolianus, A. ferrooxydans,* and *A. yilgarnensis*, and three copies in *A. prosperus*. The second represents UTP-glucose-1-phosphate uridylyltransferase (GalU) from *E. coli, L. ferriphilum,* and *S. oneidensis* and is present in all *Acidihalobacter* species with a single copy in each. The remaining orthogroup represents glucan biosynthesis protein D and G from *P. aeruginosa* and *S. oneidensis* and is present in all *Acidihalobacter* species with two copies in *A. prosperus* and *A. yilgarnensis*, and a single copy in *A. aeolianus* and *A. ferrooxydans*.

## Orthogroups with substantial expansion

Further investigated were orthogroups showing substantial expansion in the *Acidihalobacter* genomes that were not already found with four or more candidate genes. Thirty-four orthogroups were identified with at least three copies in one of the four *Acidihalobacter* species (Supplementary: table with orthologues). All but three of the thirty-four orthogroups were present in all *Acidihalobacter* species. Two orthogroups representing a hybrid sensor histidine kinase/response regulator and a c-type cytochrome were not in *A. ferrooxydans.* The other orthogroup representing a ribonucleoside-diphosphate reductase alpha subunit was not in *A. aeolianus*.

There were three orthogroups demonstrating considerably substantial expansion of more than six candidate gene copies across all or most of the *Acidihalobacter* species. The most expanded orthogroup represented a methyl-accepting chemotaxis protein from *P. aeruginosa, S. maltophilia,* and *S. oneidensis*. *A. yilgarnensis* demonstrates the most expansion with fourteen copies, *A. aeolianus* and *A. prosperus* with twelve copies, while *A. ferrooxydans* has only two copies.

The second most expanded orthogroup represented a LysR family transcriptional regulator from *B. pseudomallei*. This orthogroup demonstrated heavy expansion in all *Acidihalobacter* species, with ten copies in *A. aeolianus,* eight copies in both *A. prosperus* and *A. yilgarnensis*, and six copies in *A. ferrooxydans*.

The third most expanded orthogroup represented PilT, a type IV pilus twitching motility protein from *A. ferrooxidans*. This orthogroup demonstrated heavy expansion in *A. ferrooxydans, A. prosperus,* and *A. yilgarnensis* with seven copies each, and moderate expansion in *A. aeolianus* with five copies.

The remaining orthogroups demonstrating expansion consisted predominantly of various metabolic and regulatory proteins. Most interesting of these is an orthogroup representing ExbD, a biopolymer transporter from *A. ferrooxidans*. This orthogroup demonstrated moderate expansion in all *Acidihalobacter* species, with four copies in *A. yilgarnensis*, three copies in *A. aeolianus* and *A. prosperus*, and two copies in *A. ferrooxydans*.

Do something with the GO/proteinfer data?

Do we also want to investigate the MAGs?

Any cluster comparisons?

Anything else?