Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

Novel heavy metal resistance gene clusters are present in the genome of *Cupriavidus neocaledonicus* STM 6070, a new species of *Mimosa pudica* microsymbiont isolated from heavy-metal-rich mining site soil

Agnieszka Klonowska
Universite de Montpellier
Lionel Moulin
Universite de Montpellier
Julie Ardley
Murdoch University
Florence Braun
Laboratoire des Symbioses Tropicales et Mediterraneennes
Margaret Gollagher
Curtin University
Jaco Daniel Zandberg
Murdoch University https://orcid.org/0000-0002-3780-1464
Dora Marinova
Curtin University
Marcel Huntemann
DOE Joint Genome Institute
T.B.K. Reddy
DOE Joint Genome Institute
Neha Varghese
DOE Joint Genome Institute
Tanja Woyke
DOE Joint Genome Institute
Natalia Ivanova
DOE Joint Genome Institute
Rekha Seshadri
DOE Joint Genome Institute
Nikos Kyrpides
DOE Joint Genome Institute
Wayne Reeve (W.Reeve@murdoch.edu.au)
Murdoch University https://orcid.org/0000-0002-3780-8965

Research article

Keywords: Rhizobia, Cupriavidus, nickel tolerance, HGT, Mimosa, rhizobial biogeography, heavy metal resistance, heavy metal efflux

Posted Date: January 20th, 2020

DOI: https://doi.org/10.21203/rs.2.15865/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at BMC Genomics on March 6th, 2020. See the published version at https://doi.org/10.1186/s12864-020-6623-z.

Abstract

Background Cupriavidus strain STM 6070 was isolated from nickel-rich soil collected near Koniambo massif, New Caledonia, using the invasive legume trap host Mimosa pudica. STM 6070 is a heavy metal-tolerant strain that is highly effective at fixing nitrogen with M. pudica. Here we have provided an updated taxonomy for STM 6070 and described salient features of the annotated genome, focusing on heavy metal resistance (HMR) loci and heavy metal efflux (HME) systems.

Results The 6,771,773 bp high-quality-draft genome consists of 107 scaffolds containing 6,118 protein-coding genes. ANI values show that STM 6070 is a new species of Cupriavidus. The STM 6070 symbiotic region was syntenic with that of the M. pudica-nodulating Cupriavidus taiwanensis LMG 19424T. In contrast to the nickel and zinc sensitivity of C. taiwanensis strains, STM 6070 grew at high Ni2+ and Zn2+ concentrations. The STM 6070 genome contains 55 genes, located in 12 clusters, that encode HMR structural proteins belonging to the RND, MFS, CHR, ARC3, CDF and P-ATPase protein superfamilies. These HMR molecular determinants are putatively involved in arsenic (ars), chromium (chr), cobalt-zinc-cadmium (czc), copper (cop, cup), nickel (nie and nre), and silver and/or copper (sil) resistance. Seven of these HMR clusters were common to symbiotic and non-symbiotic Cupriavidus species, while four clusters were specific to STM 6070, with three of these being associated with insertion sequences. Within the specific STM 6070 HMR clusters, three novel HME-RND systems (nieIC cep nieBA, czcC2B2A2, and hmxB zneAC zneR hmxS) were identified, which constitute new candidate genes for nickel and zinc resistance.

Conclusions STM 6070 belongs to a new Cupriavidus species, for which we have proposed the name Cupriavidus neocaledonicus sp. nov.. STM6070 harbours a pSym with a high degree of gene conservation to the pSyms of M. pudica-nodulating C. taiwanensis strains, probably as a result of recent horizontal transfer. The presence of specific HMR clusters, associated with transposase genes, suggests that the selection pressure of the New Caledonian ultramafic soils has driven the specific adaptation of STM 6070 to heavy-metal-rich soils via horizontal gene transfer.

Background

Rhizobia are nitrogen-fixing legume microsymbionts belonging to the alpha and beta subclass of Proteobacteria, and have been named for convenience alpha- and beta-rhizobia [1, 2]. Alpha-rhizobia are common symbionts of most legume species, whereas many of the beta-rhizobial strains have a particular affinity with *Mimosa* hosts [1, 3]. The competitiveness of beta-rhizobial *Paraburkholderia* or *Cupriavidus* strains for nodulation of *Mimosa* spp. varies as a function of the host species and/or ecotypes [4], and of soil characteristics such as nitrogen availability and pH [5, 6].

While *Paraburkholderia* symbionts are considered to be ancient partners of *Mimosa* spp. [7], the *Cupriavidus-Mimosa* symbiosis seems to have evolved more recently [6, 8]. Symbiotic *Cupriavidus* strains belonging mainly to the species *C. taiwanensis* have been isolated from nodules of the invasive species *Mimosa diplotricha* Sauvalle*, Mimosa pigra* L. and *Mimosa pudica* L., with the type strain *C. taiwanensis* LMG 19424^T being isolated from a nodule of *M. pudica* growing in Taiwan [6, 9-16]. Strains of *C. necator* and *Cupriavidus* sp. that nodulate the mimosoid legume *Parapiptadenia rigida* and native *Mimosa* spp. in Uruguay and in Texas, USA have also been described [17-19]. *Cupriavidus* strains have so far not been isolated from native species of *Mimosa* growing in Brazil [7] or in India [20], raising questions as to the origins and native hosts of rhizobial *Cupriavidus* species.

Within *Cupriavidus*, several species seem particularly adapted to metal-rich environments [21, 22]. The most well-known and studied strain is *C. metallidurans* CH34^T, which represents the model bacterium for metal resistance studies [21, 22]. Other *Cupriavidus* species, such as *C. necator* (formerly *C. eutrophus*) H16 [23, 24], are metabolically versatile organisms capable of growth in the absence of organic substrates and able to use H_2 and CO_2 as sole sources of energy and carbon [25]. The genome of *C. necator* H16 was shown to display high similarity to the genome of *C. taiwanensis* LMG 19424^T [8].

We were interested in questions concerning the origin and adaptation of *M. pudica* microsymbionts found in soils characterized by heavy metal contamination in New Caledonia [13]. *M. pudica*, which originates from the Americas [26], was introduced onto the island probably at the end of the 19th century. It has become a serious weed on many Pacific Islands, where it can form dense mats, resulting in land degradation, biodiversity loss and decreased agricultural yield and economic productivity [27, 28]. Conversely, the combination of *M. pudica* and associated *Cupriavidus* rhizobia has been advocated as a novel biosorption system for removing heavy-metal pollutants [29].

A study of rhizobia isolated from New Caledonian *M. pudica* trap hosts identified five different 16S RNA and REP-PCR *Cupriavidus* genotypes (I to V) that nodulated this host [13]. *Cupriavidus* strain STM 6070 is a representative strain of a group of 15 isolates belonging to genotype III. These isolates were obtained from plants grown in a soil characterized by high total nickel concentrations (1.56 g kg^{-1}) that was collected from an active nickel mine site at the bottom of the Koniambo Massif [13]. STM 6070 and the other genotype III isolates, initially ascribed to the *C. taiwanensis* species, are highly nickel-tolerant and appear to be well adapted to the ultramafic soils they were isolated from. Strain STM 6070 was selected as part of the DOE Joint Genome Institute 2010 Genomic Encyclopaedia for Bacteria and Archaea-Root Nodule Bacteria (GEBA-RNB) sequencing project [30, 31], to allow comparative genomic studies concerning the evolution of *Cupriavidus* symbionts and, in particular, their adaptation to metal-rich environments. In this study, whole-genome data of STM 6070 was compared with genomes of symbiotic *Cupriavidus* species [6, 8, 32, 33], non-symbiotic strains of *Cupriavidus* [25, 34-36], and two genomes of the closely related genus *Ralstonia* [37]. Here we show that the STM 6070 genome harbours a multitude of diverse **h**eavy **m**etal **r**esistance (HMR) loci, including putative *ars*, *czc*, *chr*, *cop* and *nre* operons. By comparing the STM 6070 HMR loci to those in other *Cupriavidus* genomes, we identified four gene clusters (clusters B, D, I and J) that are specific to STM 6070 and may be important genetic determinants that contribute to the adaptation of this strain to the heavy-metal-rich ultramafic Koniambo soil in New Caledonia.

Results And Discussion

General characteristics of *Cupriavidus* strain STM 6070

STM 6070 is a fast-growing, Gram-negative, motile, rod-shaped isolate that forms white-opaque, slightly domed and moderately mucoid colonies within 2-3 days when grown on solid media (Figure S1). Because STM 6070 was trapped from nickel-rich ultramafic soil, we compared its heavy metal tolerance with that of other symbiotic and non-symbiotic *Cupriavidus* strains. The growth of STM 6070 was compared to the growth of *C. metallidurans* CH34^T (a model organism for heavy metal resistance [21]) and its heavy metal-sensitive derivative AE104 (CH34^T devoid of the plasmids pMOL28 and pMOL30 that confer heavy-metal-resistance [38]) at various concentrations of Ni^{2+} (Figure S2). Of the tested strains, STM 6070 had the highest tolerance to Ni^{2+} and was the only strain capable of growth at 15 mM $NiSO_4$.

C. metallidurans CH34^T grew in the presence of 10 mM $NiSO_4$, while AE104 was unable to grow at 3 mM $NiSO_4$. Previous studies had established that other symbiotic *C. taiwanensis* strains LMG 19424^T from Taiwan [13] and *C. taiwanensis* STM 6018 from French Guiana [6] were also unable to grow at 3 mM $NiSO_4$ (data not shown).

In light of the observed Ni^{2+} tolerance of STM 6070, we examined the tolerance of the *Cupriavidus* symbionts to other metal ions. In the presence of Cu^{2+}, STM 6070, 6018 and LMG 19424^T were able to grow in media containing 1.0 mM Cu^{2+}, however, growth of STM 6070 was inhibited from 0.6 mM Cu^{2+} (Figure S3). In addition, STM 6070 was able to grow in media containing 15 mM Zn^{2+}, whereas STM6018 and LMG 19424^T were far more sensitive and could not grow at this concentration (data not shown). Since STM 6070 was highly tolerant to Ni^{2+} and Zn^{2+}, the genome of this strain was examined, in particular for putative HMR determinants.

STM 6070 Minimum Information for the Genome Sequence (MIGS) and genome properties

The classification, general features and genome sequencing project information for *Cupriavidus* strain STM 6070 are provided in Table S1, in accordance with the minimum information about a genome sequence (MIGS) recommendations [39] published by the Genomic Standards Consortium [40]. The genome sequence consisted of 6,771,773 nucleotides with 67.21% G+C content and 107 scaffolds (Table 1) and contained a total of 6,182 genes, of which 6,118 were protein encoding and 64 were RNA only encoding genes. The majority of protein encoding genes (81.69%) were assigned a putative function, whilst the remaining genes were annotated as hypothetical. The distribution of genes into COGs functional categories is presented in Table S2.

Phylogenetic placement of STM 6070 within the *Cupriavidus genus*

Previous studies have shown that STM 6070 is most closely related to *C. taiwanensis* LMG 19424^T [11] and *C. alkaliphilus* ASC-732^T [34], according to *recA* phylogenies [13]. This was confirmed by a phylogenetic analysis based on an intragenic fragment of the 16S rRNA gene (Figure S4). To determine the taxonomic placement of STM 6070 at the species level, the whole genome of STM 6070 was compared with sequenced genomes of five non-symbiotic and three symbiotic *Cupriavidus* species (Table S3) to establish the average nucleotide identity (ANI) (Table S4).

ANI [41-43] comparisons showed that the STM 6070 genome displayed the highest ANI values with the *C. taiwanensis* strains STM 6018 and LMG 19424^T, but the values were lower than the species affiliation cut-off scores (Table S4). This reveals that STM 6070 (and isolates of the same rep-PCR group isolated from New Caledonia soils [13]) represent a new *Cupriavidus* species, for which we propose the name *Cupriavidus neocaledonicus* sp. nov. (i.e. from New Caledonia). The ANI values also suggest that the UYPR2.512 and AMP6 strains represent new *Cupriavidus* species.

Synteny between genomes

To assess how the observed differences in genome size (6.48 – 7.86 Mb) affected the distribution of specific genes within the five symbiotic strains of *Cupriavidus*, we used progressive Mauve [44] to align the draft genomes of STM 6070, STM6018, UYPR2.512 and AMP6 to the finished genome of *C. taiwanensis* LMG19424^T (Figure S5). The alignments of the STM 6018 and STM 6070 genomes against that of *C. taiwanensis* LMG 19424^T showed a high similarity of collinear blocks within the two largest replicons (Figure S5A), the sequence of the LMG 19424^T chromosome 1 (CHR1) being more conserved than that of the chromosome 2 (CHR2 or chromid). We identified eight scaffolds specific to STM 6070 (A3AGDRAFT_scaffold_31.32_C, _43.44_C, _54.55_C, _39.40_C, _104.105_C, _101.102_C, _99.100_C, and _89.90_C) that could not be aligned to the LMG19424^T genome sequence, as well as two STM 6070 scaffolds (A3AGDRAFT_scaffold_84.85_C and _75.76_C) that were absent from LMG 19424^T but present in STM 6018. A putative genomic rearrangement was also detected within one scaffold of STM 6070 (A3ADRAFT_scaffold_0.1), in which one part of the scaffold mapped to chromosome CHR1 and another part mapped to the chromid CHR2 of LMG 19424^T (see shaded area on Figure S5A).

In contrast, the genome alignment of UYPR2.512 and AMP6 with LMG19424^T showed important differences in replicon conservation (Figure S5B). Earlier studies on comparative genomics of *Cupriavidus* species have suggested that the largest CHR1 replicon probably constitutes the ancestral one, while the smaller CHR2 replicon was acquired as a plasmid during the evolution of *Cupriavidus* and gradually evolved to a large-sized replicon following either gene transfer from CHR1 or horizontal gene transfer [35]. Large secondary replicons, or “chromids” [45], such as CHR2, have been detected in many bacterial species and carry plasmid-like partitioning systems [25, 35] and some essential genes, such as rRNA operons and tRNA genes (present in CHR2 of LMG19424 and the corresponding syntenic region of STM6070). This chromid also carries

many genes that are conserved within a genus, and genes conserved among strains within a species. This may well explain the greater degree of sequence divergence observed (Figure S5) in CHR2 as compared with CHR1 in the symbiotic *Cupriavidus* genomes.

Finally, we observed that whereas most of the LMG 19424^T pSym sequence was well conserved in the STM 6018 and STM 6070 genomes (Figure S5A), only a few LMG19424^T pSym genes (including the *nod*, *nif*, *fix* and *fdx* genes) were conserved across all five genomes. The *M. pudica* microsymbionts (LMG 19424^T, STM 6018 and STM 6070) had almost identical pSyms (conserved pSym synteny with *nod* genes characterized by 100% protein identity). In contrast, the *Parapiptadenia rigida* (UYPR2.512) and *Mimosa asperata* (AMP6) nodulating strains harboured divergent pSyms (low synteny, with *nod* genes characterized by 80-94 and 95-98.4% protein identity to those of LMG 19424^T, respectively). Based on phylogenetic analyses of symbiotic and housekeeping loci, our results support the hypothesis that symbiotic *Cupriavidus* populations have arisen via horizontal gene transfer [46].

Comparisons of *Cupriavidus neocaledonicus* STM 6070 with other sequenced genomes of symbiotic *Cupriavidus*

The comparison of gene orthologues of STM 6070 with those of the symbiotic *Cupriavidus* strains LMG 19424^T, STM 6018, UYPR2.512 and AMP6, performed using the “Gene Phyloprofile” tool in the Microscope MaGe platform [47] (Figure 1), showed that these strains have a large core set of 4673 genes, representing from 55.5 to 78.1% of the total number of genes in these organisms (70.2% for STM 6070). Each species harbours a set of unique genes, which range from 226 for LMG 19424^T to 1993 for UYPR2.512; larger genomes had a greater number of unique genes (Figure 1A). STM 6070 harbours 483 unique genes, which represent 7.2% of the total number of genes in the genome. The majority of these unique genes (376) encode hypothetical proteins. Only 22.2 % of the 483 STM 6070 unique genes could be ascribed to functional COG categories (Figure 1B). Within the functional COG category “Cellular processes and signaling”, the largest number of genes were found in Cell wall/membrane/envelop biogenesis, Signal Transduction, Defense mechanism and Intracellular trafficking, secretion, and vesicular transport. This may be related to processes required for plant host relationships and bacterial adaptation to the host environment. For example, within functional category M we detected several genes encoding glycosyl transferases, which are putatively involved in biosynthesis of exopolysaccharides and/or polysaccharides, products that have been shown to play a major role in rhizobial infection [48].

Unique STM 6070 genes within the signal transduction category included four genes encoding putative universal stress proteins (UspA family), additional response regulators and a sensor protein (RcsC), while the defense mechanism category includes genes encoding type I and III restriction modification systems, as well as genes encoding multidrug resistance efflux pumps, which could reflect adaptation to ultramafic soils. A high number of specific genes was assigned to “Information storage and processing”. For example, 38 genes encoded putative transcriptional regulators (COG category ‘transcription’) of various families (AraC, CopG, GntR, LacI, LysR, LuxR, MerR, NagC, TetR and XRE), suggesting a requirement for supplementary regulatory mechanisms of cellular and metabolic processes. Finally, a high number of specific genes was assigned to metabolic functions, represented mainly by amino acid, carbohydrate and inorganic ion transport and metabolism, energy production and conversion, lipid metabolism and secondary metabolites biosynthesis, transport and catabolism.

Metal resistance determinants in the STM 6070 genome

To understand the genetic basis of STM 6070 metal tolerance, we then searched for the presence of common and specific heavy metal resistance (HMR) markers within the genomes of STM 6070 and the other symbiotic *Cupriavidus* species, using the TransAAP tool on the TransportDB website (http://www.membranetransport.org/) [49] to find genes encoding predicted transporter proteins. Given that STM 6070 is nickel- and zinc-tolerant, we were particularly interested in identifying

HMR proteins within known transporter superfamilies (Transporter Classification Database: http://www.tcdb.org/) [50-52]. TransAAP analysis revealed a total of 834 putative transporters within STM 6070, of which 156 were classified within the MFS, CDF, RND, CHR, ACR3 and P-ATPase protein families (Table S5). Of the 156 TransportDB predicted transporters, 23 HME transporter genes were identified in the STM 6070 genome. Based on gene arrangements and homology with characterised HMR loci, a total of 55 structural HMR genes (TransportDB predicted HME genes plus associated genes) were located in 12 clusters (clusters A – L, Figure 2). The transporter superfamily genes were compared with those described for *C. metallidurans* CH34^T, *C. necator* H16, and the symbiotic species *C. taiwanensis* LMG19424^T [35], *Cupriavidus* sp. UYPR2.512 and *Cupriavidus* sp. AMP6 (Table 2, Table S6).

Major Facilitator Superfamily (MFS) proteins

The MFS is one of the two largest families of membrane transporters found in living organisms. Within the MFS permeases, 29 distinct families have been described, each transporting a single class of compounds [53]. Of the 106 STM 6070 TransAAP-identified genes encoding putative MFS proteins, two genes (*nreB* and *arsP*) were associated with HME functions. The *nreB* gene located in the *nreAB* operon (cluster I), and the *arsP* gene located in the *arsRIC1C2BC3H1P* operon (cluster K), encode putative nickel and arsenic efflux systems, respectively (Figure 2) [52].

Cation Diffusion Facilitator (CDF) proteins

The CDF proteins are single-subunit systems located in the cytoplasmic membrane that act as chemiosmotic ion-proton exchangers [51]. They include HMR proteins such as CzcD, which provides resistance to **c**obalt, **z**inc and **c**admium [52]. Four genes encoding CDF proteins were detected in the STM 6070 genome (Table S6), but only one, *czcD*, is located in an HME cluster (*czcDI2C3B3A3,* cluster K) (Figure 2). This locus encodes a CDF efflux protein with 67.2 % identity to CH34^T CzcD, which mediates the efflux of Co^{+2}, Zn^{+2}, and Cd^{+2} ions [54]. The second CDF gene (*dmeF*) encodes an efflux protein with highest identity (76.1 %) to the CH34^T DmeF protein, which has a role in cobalt homeostasis and resistance [54], while the other two CDF genes (*fieF1* and *fieF2*) encode efflux proteins with homology to CH34^T FieF (70.8 and 69.8 % identity, respectively). FieF has a role in ferrous iron detoxification but was also shown to mediate low level resistance to other divalent metal cations such as Zn^{2+} and Cd^{2+} [55, 56].

Resistance-Nodulation-Cell Division (RND)-HME systems

The RND-HME transporters are transmembrane proteins that form a tripartite protein complex consisting of the RND transmembrane transporter protein (component A), a membrane fusion protein (MFP) (component B), and an outer membrane factor (OMF) protein (component C). These components export toxic heavy metals from the cytoplasm, or the periplasm, to the outside of the cell and have been designated as CBA efflux systems, or CBA transporters [52], to differentiate them from ABC transporters. Within a CBA system, the RND transmembrane and MFP proteins [52, 57], mediate the active part of the transport process, determine the substrate specificity, and are involved in the assembly of the RND-HME protein complex.

The RND-HME transmembrane proteins contain a large periplasmic loop flanked by 12 transmembrane α-helices, TMH I to TMH XII [52]. They are classified into different groups according to the signature consensus sequence located in TMH IV, which is essential for proton/cation antiport and is used to predict the heavy metal substrate specificity [52, 58]. The five classes of efflux systems and their predicted heavy metal substrates include: HME1 (Co^{2+}, Zn^{2+}, Cd^{2+}), HME2 (Co^{2+}, Ni^{2+}), HME3a (divalent cations), HME3b (monovalent cations) HME4 (Cu^+ or Ag^+) and HME5 (Ni^{2+}) types [52, 59, 60].

Our phylogenetic analysis of the eight TransAAP predicted STM 6070 RND proteins, together with the analysis of the conserved motifs within the proteins, suggests that three of these proteins belong to the HME1 class, two belong to the

HME3a class and the remaining three proteins belong to the HME3b, HME4 and HME5 classes, respectively (Figure 3 and Table S6). The STM 6070 genome lacks genes encoding the HME2-type transmembrane proteins, such as the *C. metallidurans* CH34^T CnrA and NccA, which are involved in heavy metal resistance and have predicted substrate specificity for cobalt and nickel [52].

RND-HME1

STM 6070 contained 3 RND-HME1 encoded proteins (CzcA1-A3), characterized on the basis of homology to the canonical CzcA proteins of CH34^T. STM 6070 CzcA1 and CzcA3 grouped with CH34^T CzcA and CczA2, while STM 6070 CzcA2 formed an outgroup (Figure 3). The *C. necator czcA1* gene was within an operon located in cluster F (Figure 2) and annotated as *czcJ1I1C1B1A1*. In addition to the *czcCBA* genes, this cluster contained a *czcI1* homolog to a transcriptional regulator that has been shown to control the expression of *czcC1B1A1* [52, 61] and a *czcJ1* homolog, which was reported to be strongly induced by Cd^{2+}, Cu^{2+}, Ni^{2+}, and Zn^{2+} [35, 62]. This operon was located in a genomic region showing high synteny with corresponding regions in the other symbiotic *Cupriavidus* strains and in *C. necator* N-1, and the STM 6070 CzcA1 protein showed high identity with other *Cupriavidus* CzcA orthologues (Table S6). In *C. metallidurans* CH34^T, the corresponding *czc* cluster (*czcMNICBADRSEJ*, locus tags Rmet_5985-74) is located on the plasmid pMOL30 and contains additional genes that are not found in STM 6070 [35]. The second STM 6070 RND-HME1 efflux system (*czcC2B2A2*) formed part of a large group of HMR loci within cluster I (Figure 2). Immediately upstream of *czcC2B2A2* is a *nreB* gene, encoding a putative nickel resistance MFS protein. A similar arrangement has been observed for the CH34^T *nccCBA nreB* cluster found on plasmid pMOL30 [35]. Cluster I was delimited by transposases and no conserved syntenic arrangement with the six other *Cupriavidus* genomes was observed (Table 2). The third STM 6070 RND-HME1 efflux system (*czcD czcI2C3B3A*) was located within cluster K (Figure 2). The *czcI2* and *czcD* components encode putative regulator and CDF proteins, respectively. This cluster, which was also delimited by two Tn*3* transposases, had conserved synteny to corresponding regions in the genomes of *Cupriavidus* spp. LMG 19424^T and STM 6018, but not to AMP6 and UYPR2.512.

RND-HME3a

STM 6070 contained two putative RND-HME3a efflux systems located in clusters G and I. Cluster G contained an *hmv* operon, located in a region that was syntenic to corresponding regions in symbiotic *Cupriavidus* and *C. necator* N-1. Although the region was not syntenic in CH34^T, the encoded HmvCBA proteins all had high identity with STM 6070 HmvCBA proteins [35, 63], however, the role of the CH34^T proteins in heavy metal efflux has yet to be determined [35].

Cluster I contained a putative zinc efflux RND-HME3a system, annotated as *hmxB zneAC*, with the associated upstream genes *zneRhmxS* encoding a two-component sensor regulatory system. The encoded proteins had low identity (38-44%) to corresponding proteins in other *Cupriavidus* genomes, however, although the BAC gene arrangement is atypical to the characterised RND-HME CBA transporter gene arrangement, it is the same as that described in the characterised CH34^T HME3a zinc efflux system *zneSRBAC* [35, 57, 64] (Table S6). The STM 6070 ZneA protein contained highly conserved amino acids identified in the active and proximal heavy metal-binding sites of the characterised CH34^T ZneA protein [64] (Table S7). Based on conservation of the essential amino acid residues, these proteins would be divalent cation transporters, putatively involved in zinc efflux. Interestingly, the highest similarity to the STM 6070 HmxB ZneAC proteins (70, 86.5 and 69.5%, respectively) was to encoded proteins of the marine betaproteobacterium *Minibacterium massiliensis*, within an operon of similar architecture but of unknown function and substrate specificity [65].

RND-HME3b

An RND-HME3b *hmyFCBA* efflux system was identified in cluster A (Figure 2). This operon showed high identity to a corresponding CH34^T *hmyFCBA* cluster (locus tags Rmet_4119-4123), located on the chromid, and was also highly conserved in the four symbiotic *Cupriavidus* strains and *C. necator* N-1. The role of the Hmy efflux system in *Cupriavidus* is

currently unknown and this system is likely to be inactive in CH34^T since *hmyA* in this strain is insertionally inactivated by IS*1088* [66]. However, in the characterized *Escherichia coli* metal cation-transporting efflux system CusCFBA, *cusF* encodes a small auxiliary protein that is required for full resistance to copper and silver [67]. A similar role is predicted for the *Cupriavidus hmyF* even though there is low identity (< 30%) to *E. coli cusF*.

RND-HME4

An RND-HME4 *silDCBAF* efflux system was identified in cluster J and has been suggested to be important for monovalent cation efflux in CH34^T [68]. No syntenic regions were identified in the other *Cupriavidus* genomes. However, this operon is similar to the CH34^T pMOL30 *silDCBA* operon (Rmet_5030-5034), which encodes a putative silver efflux system, and to the CH34^T chromid *cusDCBAF* operon (Rmet_6133-6136), which encodes a putative copper efflux system [68]. Similar operons were also identified in the STM 6018, AMP6, N-1 and H16 genomes.

RND-HME5

An RND-HME5 *nielC cep nieBA* efflux system, identified in cluster B, was located 28 kb downstream of cluster A. This operon included a gene encoding a conserved exported protein (*cep*) situated between the *nieC* and *nieB* structural genes, disrupting the typical RND CBA operon arrangement. Among the *Cupriavidus* strains, a similar operon structure was found only in the AMP6 genome, with the structural proteins displaying high identity to the corresponding STM 6070 proteins. This operon structure was also found in the genome of *M. massiliensis* [65], with the encoded proteins having 41 to 79 % protein identity with those of STM6070. As there are no RND-HME5 efflux systems present in CH34^T, the protein encoded by STM 6070 *nieA* was compared with the characterized RND-HME5 proteins NrsA (involved in nickel resistance) and CopA (involved in copper resistance) of the cyanobacterium *Synechocystis sp*. PCC 6803 [69, 70]. The phylogenetic analysis (Figure 3) shows that although these proteins possess a common ancestor, they form two well separated clades, one comprising the HME5 proteins of STM 6070, AMP6 and *M. massiliensis*, and the second containing the NrsA and CopA of PCC 6803 together with RND-HME5 proteins from the cyanobacterium *Anabena sp*. PCC 7120 [71]. The betaproteobacterial and cyanobacterial RND-HME5 proteins share less than 41 % identity, resulting in totally different amino acids involved in putative proximal and distal metal-binding sites, as well as differences in the consensus sequence of the TMHIV α-helice (Table S7). Of particular interest was the finding that the three histidines, which are present in the proximal site of NieA and in the proteins of this clade, form part of conserved HAEGVH and HRLDH motifs, and match with putative nickel-binding motifs H-X4-H and H-X3-H that are predominant in Ni-binding proteins, as described for the Ni-binding proteins of *Streptococcus pneumoniae* [72]. Based on these findings, we suggest that this *nielC cep nieBA* operon encodes a new RND-HME system (class 6) putatively involved in nickel efflux and represents an interesting candidate for knockout mutation to determine if it is a major determinant of nickel tolerance in STM 6070.

Chromate Ion Transporter (CHR) proteins

The CHR proteins efflux chromate from the cytoplasm through an indirect active transport process [73]. Two STM 6070 genes (*chrA1* and *chrA2*) were identified as encoding putative CHR proteins. The STM 6070 ChrA1 and ChrA2 proteins showed higher identity to each other than to the CH34^T pMOL28 and chromid ChrA proteins. The *chrB1A1* operon (cluster B) was located up-stream of the putative RND-HME5 efflux system *nielC-cep-nieBA* (Figure 2). This *chr* operon was conserved in the genomes of the symbiotic *Cupriavidus* strains LMG19424^T and STM6018, forming part of a large synteny block. The second *chr* operon, annotated as *chrB2A2CF-cep-chrL (chrY)*, was located in cluster I, along with the RND-HME efflux systems *czcC2B2A2* and *hmxB-zneAC* and the *nreAB* operon (Figure 2). In addition to *chrB2A2*, this operon contained four other genes: *chrC*, encoding a putative superoxide dismutase that may reduce chromate and thereby decrease chromate toxicity [74]; *chrF*, encoding a putative transcriptional repressor [74]; *cep*, encoding a conserved exported protein containing a Concanavalin A-like lectins/glucanases domain; and finally, *chrL*, encoding a lipoprotein (protein family, LppY/LpqO [75]) with 71.1 % identity to CH34^T ChrL (also annotated as CH34^T ChrY). Corresponding gene clusters were identified in the

UYPR2.512 and CH34^T genomes. In CH34^T the corresponding *chrL* (*chrY*) gene (locus tag Rmet_6195) is induced by chromate [35]). Deletion of *chrL* in the Gram-positive *Arthrobacter* sp. strain FB24 resulted in a noticeable decrease in chromate resistance [75]. The STM 6070 *chr* operon lacks the *chrE*, *chrO*, *chrN*, *chrP* and *chrZ* orthologues found in the corresponding CH34^T *chr* operon. The different chromate resistance genes might affect tolerance to chromate, or to another metal-oxyanion [35]. The STM 6070 *chrB2* gene appears to be inactivated by an insertion that changes the reading frame after 214 amino acids, and shortens the protein to only 293 amino acids, instead of the full length 324 amino acid protein encoded by CH34^T *chrB*. Since ChrB seems to be important for chromate resistance in CH34^T [76], the tolerance of STM 6070 to chromate might be compromised. Indeed, in our experimental conditions STM 6070 only showed slight tolerance to Cr^{6+} (0.1 mM) [13].

Arsenical Resistance-3 (ACR3) proteins

The ACR3 family includes permeases involved in arsenate resistance. The two STM 6070 ACR-3 type *arsB1* and *arsB2* genes are located in two *ars* operons encoding putative arsenate detoxification systems. The first operon is located downstream of the *czc* operon in cluster K. Genes in this *ars* operon had high identity with genes of the CH34^T *arsMRIC2BC1HP* operon encoding an arsenite and arsenate detoxifying system [62, 77]. This *ars* cluster encoded a putative arsenite/arsenate transcriptional regulator/repressor (ArsR), a glyoxalase family of proteins (ArsI), three arsenate reductases (ArsC1, ArsC2, ArsC3), an arsenite efflux pump belonging to the ACR3 class of permeases (ArsB1), a NADPH-dependent FMN reductase (ArsH1), and a putative permease from the major facilitator family (MFS) (ArsP) [77]. The operon was highly conserved in the *Cupriavidus* symbionts LMG19424^T and STM6018 and formed a large syntenic region. The second *ars* operon *arsR2C4B2H2*, in cluster L, was present in all other *Cupriavidus* genomes except UYPR2.512, but was missing several genes (*arsI*, *arsC* and *arsP*) found in the cluster K operon.

P-type ATPase proteins

P-type ATPases directly utilise ATP to export metal ions from the cell cytoplasm. Of the ten STM 6070 genes assigned to the P-type ATPase protein family (Table S5), five encoded P-type ATPases putatively involved in HME (Figure 2 and Table 2). The *copF* P-type ATPase gene in cluster J was located upstream of the *silDCBAF* operon and could encode an essential copper efflux component, as shown for CH34^T [38]. However, the STM 6070 CopF appears to be truncated in its C-terminus and thus may not be functional. Two P-type ATPase-encoding genes were identified in cluster D and annotated as *silP* and *copP*. The encoded proteins had very low identity with proteins of the *Cupriavidus* genomes (Table S6), except for one P-type ATPase protein from AMP6 with 86% identity with the CopP protein. The proteins had higher identity with P-type ATPases encoded by *C. necator* H16, annotated as SilP (86%) and CopP (94.7%), and putatively involved in silver and copper ion transport, respectively [25]. Within cluster H, a P-type ATPase-encoding gene, annotated as *cupA*, was located next to a regulatory gene, *cupR* (Figure 2), in a conserved large syntenic block common to all compared *Cupriavidus* strains, with high identity between corresponding genes. The *cupA* and *cupR* genes are putatively involved in copper ion transport. Finally, *zntA* was located within cluster C in a group of genes annotated as *czcJ2-hns-czcLRS-ubiGI-zntA*. Genes in this cluster had high identity with loci in two gene clusters in CH34^T that have been annotated as *zntA czcICΔB* (locus tags Rmet_4594-4597) and *czcBA ubiG czcSRL* IS *hns mmmQ* (locus tags Rmet_4469-4461), respectively. These CH34^T clusters encode an RND system (*czcICBA*), the ZntA ATPase, a two-component regulatory system CzcRS and a 3-demethylubiquinone-9 3-methyltransferase (UbiG) [35, 63]. UbiG participates in the biosynthesis of ubiquinone and its activity could be related to the sensor kinase activity of the two-component system CzcRS [78, 79]. The *czcL*, *hns* and *mmmQ* genes encode an unknown protein, an H-NS like protein and a small stress responsive protein, respectively. Genes in the second CH34^T cluster may be inactivated by an insertion sequence located between *czcL* and *hns*. The synteny of the STM 6070 cluster C is perfectly conserved in the genomes of the four symbiotic *Cupriavidus* strains, suggesting that it is

functional, but it is devoid of the *czcCBA* RND system found in the corresponding CH34^T cluster. The role of the regulatory loci *czcLRS-ubiGI*, with regard to *zntA* expression, would thus be interesting to determine.

Other mechanisms of cation detoxification (not included in TransAAP)

The search for further heavy metal resistance determinants in STM 6070 that were orthologous to those described in CH34^T led to identification of a copper-resistance operon *copRSABCD* (cluster E). This had a similar structure to the CH34^T *cop* cluster (*copS2R2A2B2C2D2*) located on the chromid, which encodes a copper-resistance mechanism that is thought to sequester copper outside the cytoplasm [80, 81]. CopSR is a two-component sensor-regulator system and CopA is a putative multi-copper oxidase thought to oxidize Cu^{1+} to Cu^{2+}. CopA proteins contain several motif variants of MGGM/MAGM/MGAM/MSGM, possibly involved in binding numerous Cu^{1+} ions, as determined for *Pseudomonas syringae* CopA [82]. CopA is exported to the periplasm by the twin-arginine translocation pathway [81], where it may interact with an outer-membrane protein CopB, providing the minimum system required for low level copper resistance. CopD is a membrane protein involved in transfer of Cu^{1+} from the periplasm to the cytoplasm for CopA binding [80, 81]), and CopC is thought to regulate copper uptake by CopD. The STM 6070 CopA protein shows 75.8 % identity to both CH34^T CopA1 (pMOL30) and CopA2 (chromid) proteins. Interestingly, the alignment of corresponding proteins reveals the presence of a histidine-rich sequence (GHG GHS GDS GHS GDS $(GHS)_5$ GDS GHG AHA GHG) located in the middle of the methionine-rich CopA motif in the STM 6070 protein, which is absent from other CopA sequences deposited in the NCBI database. The *Escherichia coli* HRA-1 and 2, *Enterococcus hirae* CopB [83] and *Rhizobium leguminosarum* ActP [84] Cu-exporting P-type ATPase proteins also contain histidine-rich leaders, which we postulate bind to copper ions. The STM 6070 CopRSABCD putative copper sequestration system may provide a second line of defence against copper toxicity and is particularly well conserved in all of the symbiotic *Cupriavidus* isolates.

Location of HMR determinants

The detected STM 6070 HMR determinants in the 12 clusters (A to L, Figure 2, Table 2) were assigned to putative replicons of the STM 6070 genome, following alignment of contigs to the finished LMG19424^T genome. Two clusters (D and H) could be assigned to chromosome 1 (CHR1), one cluster (K) to the pSym, and nine clusters (A, B, C, E, F, G, I, J and L) to CHR2 (chromid, Figure S6). Therefore, STM 6070 appears to carry the great majority of its HMR clusters on CHR2. In contrast *C. metallidurans* CH34^T harbours 8 out of 24 HMR clusters on CHR2 (chromid) [35, 52, 63]. The genome synteny comparison revealed that six of the STM 6070 HMR clusters (A, C, E, F, G and H) are common to symbiotic and non-symbiotic *Cupriavidus* genomes. STM 6070 HME gene products from clusters A, C, E, F, G and H displayed highest identity (93 to 100 %) with corresponding proteins of *C. taiwanensis* isolates (LMG 19424^T and STM 6018, Table S6), reflecting the taxonomic relationship with *C. taiwanensis*.

Synteny analysis indicated that the specific STM 6070 HMR clusters B, D, I and J were absent from all other analysed *Cupriavidus* genomes, although some of the HMR genes within these clusters had orthologues (35 to 89 % of encoded protein identity) in the genomes of the other *Cupriavidus* strains. Cluster K was perfectly conserved within the LMG 19424^T and STM 6018 genomes (100 %) in a large syntenic region, whereas it was absent from the AMP6 and UYPR2.512 genomes. Only the separate *czc* and *ars* operons from cluster K were detected in the non-symbiotic *Cupriavidus* genomes, with encoded protein identities of 76 - 77 % and 83 - 88 %, respectively, to the STM 6070 *czc* and *ars* operon encoded proteins. This observation can be explained by the location of cluster K on the pSym, which, as proposed recently [46], seems to be largely shared between *M. pudica* microsymbionts of different genomic backgrounds. Indeed, we demonstrated by the progressive Mauve alignment (Figure S5) that the pSym seems to be conserved in the genomes of the *M. pudica*-nodulating LMG 19424^T, STM 6018 and STM 6070, in contrast to genomes of AMP6 and UYPR2.512, which nodulate different mimosoid legumes and harbour totally different symbiotic plasmids.

The analysis of genes adjacent to HMR clusters revealed that for clusters D and J contained a transposase-encoding gene at one end of the cluster and clusters I and K were flanked by transposase-encoding genes (Figure 2). Analysis of the GC% using a two-tailed Mann-Whitney U test revealed that cluster D and J did not contain a significantly different GC% (P-value >0.01) in comparison to the average GC% of the genome. In comparison, clusters I and K did contain a significantly different GC% (P-value <0.01) in comparison to the average GC% of the genome. This suggests acquisition of the clusters by horizontal gene transfer (HGT) for clusters I and K. Cluster I, located on the chromid, is the largest of these clusters (of approximately 25 kb), flanked by transposases of the Tn*3* and IS66 type, and carries four different HMR determinants, including *czcC2B2A2* and *hmxB zneAC*. Cluster K is flanked by two Tn*3* transposases, however, unlike Cluster I there is a high conservation of architecture and gene identity with the closely related *C. taiwanensis* strains (LMG 19424^T and STM 6018). This may indicate that Cluster I contains HME determinants that are important for survival in the New Caledonian ultramafic soils. In *C. metallidurans*, the acquisition of mobile genetic elements that contain metal resistance genes appears to be a strategy important for its adaptation to environments that contain elevated levels of heavy metals [62, 85].

In contrast, no transposases or insertion sequences could be found around cluster B, or more particularly, around the operon *nielC cep nieBA*). This operon, which is absent from LMG 19424^T and STM 6018 genomes, is located in a large highly conserved region, suggesting a gene loss from *C. taiwanensis* genomes. Interestingly, *nielC cep nieBA* (cluster B) and *hmxB zneAC* (cluster I), two unique RND-HME systems in terms of operon structure and protein sequences, showed significant structure and protein sequence similarity with two operons from the genome of *M. massiliensis* [65].

Conclusions

New Caledonian *Cupriavidus* microsymbionts isolated from *Mimosa pudica* nodules belong to one of five REP-PCR genotypes, which all harbour identical symbiotic *nodA* and *nifH* genes [13] but display different metal tolerance phenotypes. Fifteen strains belonging to the REP-PCR genotype III were found to be the most nickel-tolerant. The current study presents an analysis of the genome of strain STM 6070, a representative of the REP-PCR genotype III. STM 6070 was originally placed within *C. taiwanensis* on the basis of 16S and *recA* phylogenies [13], however, our analysis, combined with the genetic and phenotypic data described by Klonowska and colleagues [13], has revealed that STM 6070 represents a new species of *Cupriavidus*, for which we propose the name *Cupriavidus neocaledonicus* sp. nov.

The major aim of this study was to gain insights into the molecular basis of the tolerance of *C. neocaledonicus* to high levels of nickel and zinc. The genome of *C. neocaledonicus* STM 6070 contains a very large number of diverse putative HMR determinants belonging to the RND, MFS, CHR, ARC3, CDF and P-ATPase protein superfamilies (Table 2). These constitute putative efflux systems or ion pumps involved in arsenic (2 *ars* operons), chromium (2 *chr* operons), cobalt-zinc-cadmium (2 *czc* operons), copper and/or silver (*copA*, *copP*, and *silA* genes), and nickel (1 *nre* operon and 1 *nie* operon) tolerance. The HMR determinants are clustered in 12 loci (cluster A to cluster L) of which two clusters seem to be localised in CHR1, nine on CHR2 (chromid) and one on the pSym (Figure S6).

Among these clusters, six (A, C, E, F, G and H) are common to both symbiotic and non-symbiotic genomes, with the different levels of sequence similarity suggesting their presence in a bacterial ancestor and possible evolution under different evolutionary pressures. Conversely, cluster K, on the pSym, was present only in STM 6070 and the *C. taiwanensis* strains. The 100 % identity of cluster K encoded proteins among the STM 6070, LMG 19424^T and STM 6018 genomes could be explained by the “recent” transfer of pSym between the *M. pudica* microsymbionts, in accordance with the findings of Parker [46].

Four of the HMR clusters (B, D, I, and J) are specific to the STM 6070 genome and we propose that these clusters contain genes that are determinants for the adaptation of *C. neocaledonicus* to high concentrations of nickel and zinc in Koniambo soil in New Caledonia. Indeed, within clusters B and I, the identified *nie*, *czc2* and *zne* operons (encoding RND-HME5, -HME1 and -HME3a, efflux systems respectively) constitute good candidates for nickel and zinc tolerance molecular determinants.

Moreover, the finding that at least four HMR clusters (D, I, J and K) are directly associated with insertion elements suggests that mobile genetic elements play an important role in adaptation of the STM 6070 genome to the New Caledonian environment. Insertion elements have previously been found to play a role in enabling the host to adapt to new environmental challenges, and to contribute to the genetic adaptation of *C. metallidurans* to toxic zinc concentrations [85, 86]. Future work involving a targeted mutagenesis study should allow us to determine the precise role of the newly identified HMR operons in STM 6070 and will provide an understanding of the specific molecular determinants required for the evolution and adaptation of these bacterial symbionts to the heavy-metal-rich New Caledonian soils.

Methods

Bacterial strains and growth conditions

All strains used in this study can be found in Table S3. *C. neocaledonicus* STM 6070 was isolated using *M. pudica* as a trap-host, as previously described, from a soil characterized by high total nickel concentrations (1.56 g kg^{-1}) collected at the bottom of the Koniambo Massif, where active nickel mines are located [13]. Bacterial isolates were sub-cultured on yeast mannitol agar plates (YMA, Vincent, 1970) and incubated at 28°C for 48 h. For long-term maintenance, bacterial strains were grown in YM broth and preserved in 20 % glycerol at -80°C. For the comparison of metal tolerance, bacteria were grown in 30 mL liquid 284 Tris-culture medium [19] amended with $NiSO_4$ (0, 3, 5, 10 and 15 mM), $Cu(NO_3)_2$ (0, 0.3, 0.6 and 1.0 mM) and $ZnSO_4$ (0, 3, 5, 10 and 15 mM) at 28°C on a gyratory shaker set to 150 rpm. Bacterial growth was monitored by measuring the OD_{600nm} in a spectrophotometer.

Genomic DNA preparation

C. neocaledonicus STM 6070 was streaked onto TY solid medium [87] and grown at 28°C for three days to obtain well grown, well separated colonies, then a single colony was selected and used to inoculate 5 ml TY broth medium. The culture was grown for 48 h on a gyratory shaker (200 rpm) at 28°C. Subsequently, 1 ml was used to inoculate 60 ml TY broth medium that was incubated on a gyratory shaker (200 rpm) at 28°C until an OD_{600nm} of 0.6 was reached. DNA was isolated from 60 ml of cells using a CTAB bacterial genomic DNA isolation method [87]. Final concentration of the DNA was 0.5 mg ml^{-1}.

Genome sequencing and assembly

The genome of *C. taiwanensis* STM 6070 was sequenced at the Joint Genome Institute (JGI) using Illumina technology [88]. An Illumina standard shotgun library was constructed and sequenced using the Illumina HiSeq 2000 platform which generated 26,402,396 reads totaling 3,960.4 Mbp. All raw Illumina sequence data was passed through DUK, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts (Mingkun, L., Copeland, A. and Han, J., unpublished). The following steps were then performed for assembly: (1) filtered Illumina reads were assembled using Velvet [89] (version 1.1.04), (2) 1–3 kb simulated paired end reads were created from Velvet contigs using wgsim (https://github.com/lh3/wgsim), (3) Illumina reads were assembled with simulated read pairs using Allpaths–LG [90] (version r39750). Parameters for assembly steps were: 1) Velvet (–v –s 51 –e 71 –i 4 –t 1 –f "-shortPaired -fastq $FASTQ" –o "-ins_length 250 -min_contig_lgth 500") 10) 2) wgsim (-e 0 -1 100 -2 100 -r 0 -R 0 -X 0) 0) 3) Allpaths–LG (PrepareAllpathsInputs:PHRED64=1 PLOIDY=1 FRAGCOVERAGE=125 JUMPCOVERAGE=25 LONGJUMPCOV=50, RunAllpath-sLG: THREADS=8 RUN=stdshredpairs TARGETS=standard VAPIWARNONLY=True OVERWRITE=True). The final draft assembly contained 107 scaffolds. The total size of the genome is 6.8 Mb and the final assembly is based on 814 Mbp of Illumina data, which provides an average 120.3x coverage of the genome.

Genome annotation

For the general genome content description genes were identified using Prodigal [91] as part of the DOE-JGI annotation pipeline [92, 93]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. The tRNAScanSE tool [94] was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA [95]. Other non–coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL (http://infernal.janelia.org). Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG-ER) platform (http://img.jgi.doe.gov/er) [93]. The expert annotation of HMR genes was performed within the MaGe platform (https://www.genoscope.cns.fr/agc/microscope/mage) and therefore the gene numbers (CT6070v1_ XXXXXX-XX) are those from the MaGe platform. The corresponding locus tags of genes annotated in the MaGe and JGI platforms are indicated in Table S6.

Phylogenetic analyses

Gene fragments sequences were corrected with Chromas Pro v1.33 software (Technelysium) and aligned using either ClustalX [96] or MUSCLE as implemented in MEGA, version 6 [97]. Alignments were manually edited using GeneDoc software [98]. Phylogenetic analyses were performed in MEGA6 [97] using the Neighbor-Joining method [99]. Bootstrap analysis [100] with 1000 replicates was performed to assess the support of the clusters.

Genome analyses

The comparison of specific and common genes of symbiotic *Cupriavidus* species, presented in a Venn diagram (Figure 1), was performed using the “Gene Phyloprofile” tool in the Microscope MaGe platform (https://www.genoscope.cns.fr/agc/microscope/mage). The orthologous counterparts in the genomes were detected by applying parameters of a minimum of 30% for protein sequences identity over a minimum of 80% of the protein length (>30% protein MinLrap 0.8). The homologous genes were then removed from the resulting list. Transport systems were identified using the TransAAP tool [49] (TransportDB website (http://www.membranetransport.org/)) for prediction of efflux systems and transporter families.

Two methods were used for the comparison of average nucleotide identities (ANI): ANIg [42] and ANIb [43]. In order to perform the alignments using progressive Mauve software [51], the scaffolds of each draft genome (STM 6070, STM6018, UYPR2.512 and AMP6) were firstly reordered using Mauve software on the basis of the *C. taiwanensis* LMG19424^T concatenated genome. Then, reordered genomes were used to perform the alignment with progressiveMauve. Circular views by BlastAtlas were performed using the CGview server hosted at Stothard Research Group (http://stothard.afns.ualberta.ca/cgview_server/), for alignment of the STM 6070 sequence, aligned firstly on the LMG19424^T concatenated sequence (CHR1, pSym, CHR2).

The essential amino acid residues that form the proximal and distal heavy-metal-binding sites have been identified for the CH34^T zinc-specific RND HME3a transmembrane transporter ZneA [64]. Using the ZneA protein as a backbone, we aligned the eight STM 6070 RND transmembrane proteins with those used for phylogenetic analysis (Figure 3), to compare and identify the corresponding essential amino acid residues that form the putative proximal and distal heavy-metal-binding sites in these transporters (Table S7). In addition, we used the MaGe Microscope annotation platform [47] to analyse the syntenic arrangements of the eight HME-RND efflux systems present in STM 6070 and compare them with the HME-RND efflux systems found within six other *Cupriavidus* strains, as outlined below.

Abbreviations

½LA: Half strength Lupin Agar

ANI: Average Nucleotide Identity

GEBA-RNB: Genomic Encyclopedia for Bacteria and Archaea-Root Nodule Bacteria

HGT: Horizontal Gene Transfer

HME: Heavy Metal Efflux

HMR: Heavy Metal Resistance

IMG: Integrated Microbial Genomes

IS: Insertion Sequence

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231. We gratefully acknowledge the funding received from the French National Agency of Research (Project BETASYM ANR-09-JCJ-0046), Curtin University Sustainability Policy Institute, and the funding received from Murdoch University Small Research Grants Scheme.

Authors' contributions

AK, LM and FB supplied the strain and background information for this study, performed growth experiments and were responsible for phylogenetic analyses and genome alignments; WR supplied the DNA to the JGI; AK, WR, JA, MG and DM performed bioinformatics analyses and/or drafted the paper, MH, TBKR, NV, TW, NI, RS and NK were involved in sequencing the genome, and all authors were involved in editing the final paper. All authors read and approved the final manuscript.

Acknowledgements

We thank Gordon Thompson (Murdoch University) for the preparation of SEM and TEM photos. We thank Paul Jennsen for help with TransAAP analysis and Gilles Béna for helpful discussions. We are especially grateful to Max Mergeay for very helpful discussions.

References

1. Gyaneshwar P, Hirsch AM, Moulin L, Chen WM, Elliott GN, Bontemps C, Estrada-de Los Santos P, Gross E, Dos Reis FB, Sprent JI *et al*: **Legume-nodulating betaproteobacteria: diversity, host range, and future prospects**. *Mol Plant Microbe Interact* 2011, **24**(11):1276-1288.
2. Moulin L, Munive A, Dreyfus B, Boivin-Masson C: **Nodulation of legumes by members of the beta-subclass of Proteobacteria**. *Nature* 2001, **411**(6840):948-950.
3. Bournaud C, de Faria SM, Dos Santos JMF, Tisseyre P, Silva M, Chaintreuil C, Gross E, James EK, Prin Y, Moulin L: ***Burkholderia* species are the most common and preferred nodulating symbionts of the *Piptadenia* group (tribe Mimoseae)**. *PLoS One* 2013, **8**:e63478.
4. Melkonian R, Moulin L, Béna G, Tisseyre P, Chaintreuil C, Heulin K, Rezkallah N, Klonowska A, Gonzalez S, Simon M *et al*: **The geographical patterns of symbiont diversity in the invasive legume *Mimosa pudica* can be explained by the competitiveness of its symbionts and by the host genotype**. *Environ Microbiol* 2014, **16**:2099-2111.
5. Elliott GN, Chou J-H, Chen W-M, Bloemberg GV, Bontemps C, Martínez-Romero E, Velázquez E, Young JPW, Sprent JI, James EK: ***Burkholderia* spp. are the most competitive symbionts of *Mimosa*, particularly under N-limited conditions**. *Environ Microbiol* 2009, **11**(4):762-778.
6. Mishra RP, Tisseyre P, Melkonian R, Chaintreuil C, Miché L, Klonowska A, Gonzalez S, Bena G, Laguerre G, Moulin L: **Genetic diversity of *Mimosa pudica* rhizobial symbionts in soils of French Guiana: investigating the origin and diversity of *Burkholderia phymatum* and other beta-rhizobia**. *FEMS Microbiol Ecol* 2012, **79**(2):487-503.
7. Bontemps C, Elliott GN, Simon MF, Dos Reis Júnior FB, Gross E, Lawton RC, Neto NE, de Fátima Loureiro M, De Faria SM, Sprent JI *et al*: ***Burkholderia* species are ancient symbionts of legumes**. *Mol Ecol* 2010, **19**(1):44-52.
8. Amadou C, Pascal G, Mangenot S, Glew M, Bontemps C, Capela D, Carrère S, Cruveiller S, Dossat C, Lajus A *et al*: **Genome sequence of the beta-rhizobium *Cupriavidus taiwanensis* and comparative genomics of rhizobia**. *Genome Res* 2008, **18**(9):1472-1483.
9. Barrett CF, Parker MA: **Coexistence of *Burkholderia*, *Cupriavidus*, and *Rhizobium* sp. nodule bacteria on two *Mimosa* spp. in Costa Rica**. *Appl Environ Microbiol* 2006, **72**:1198-1206.
10. Chen WM, James EK, Chou JH, Sheu SY, Yang SZ, Sprent JI: **Beta-rhizobia from *Mimosa pigra*, a newly discovered invasive plant in Taiwan**. *New Phytol* 2005, **168**(3):661-675.
11. Chen WM, Laevens S, Lee TM, Coenye T, De Vos P, Mergeay M, Vandamme P: ***Ralstonia taiwanensis* sp. nov., isolated from root nodules of *Mimosa* species and sputum of a cystic fibrosis patient**. *Int J Syst Evol Microbiol* 2001, **51**(Pt 5):1729-1735.
12. Chen WM, Moulin L, Bontemps C, Vandamme P, Béna G, Boivin-Masson C: **Legume symbiotic nitrogen fixation by beta-proteobacteria is widespread in nature**. *J Bacteriol* 2003, **185**:7266-7672.

13. Klonowska A, Chaintreuil C, Tisseyre P, Miché L, Melkonian R, Ducousso M, Laguerre G, Brunel B, Moulin L: **Biodiversity of *Mimosa pudica* rhizobial symbionts (*Cupriavidus taiwanensis*, *Rhizobium mesoamericanum*) in New Caledonia and their adaptation to heavy metal-rich soils**. *FEMS Microbiol Ecol* 2012, **81**:618-635.
14. Liu X, Wei S, Wang F, James EK, Guo X, Zagar C, Xia LG, Dong X, Wang YP: ***Burkholderia* and *Cupriavidus* spp. are the preferred symbionts of *Mimosa* spp. in southern China**. *FEMS Microbiol Ecol* 2012, **80**:417-426.
15. Liu XY, Wu W, Wang ET, Zhang B, Macdermott J, Chen WX: **Phylogenetic relationships and diversity of beta-rhizobia associated with *Mimosa* species grown in Sishuangbanna, China**. *Int J Syst Evol Microbiol* 2011, **61**(Pt 2):334-342.
16. Verma SC, Chowdhury SP, Tripathi AK: **Phylogeny based on 16S rDNA and *nifH* sequences of *Ralstonia taiwanensis* strains isolated from nitrogen-fixing nodules of *Mimosa pudica*, in India**. *Can J Microbiol* 2004, **50**(5):313-322.
17. Taulé C, Zabaleta M, Mareque C, Platero R, Sanjurjo L, Sicardi M, Frioni L, Battistoni F, Fabiano E: **New Betaproteobacterial *Rhizobium* strains able to efficiently nodulate *Parapiptadenia rigida* (Benth.) Brenan**. *Appl Environ Microbiol* 2012, **78**(6):1692-1700.
18. Platero R, James EK, Rios C, Iriarte A, Sandes L, Zabaleta M, Battistoni F, Fabiano E: **Novel *Cupriavidus* strains isolated from root nodules of native Uruguayan *Mimosa* species**. *Appl Environ Microbiol* 2016, **82**(11):3150-3164.
19. Andam CP, Mondo SJ, Parker MA: **Monophyly of *nodA* and *nifH* genes across Texan and Costa Rican populations of *Cupriavidus* nodule symbionts**. *Appl Environ Microbiol* 2007, **73**(14):4686-4690.
20. Gehlot HS, Tak N, Kaushik M, Mitra S, Chen WM, Poweleit N, Panwar D, Poonar N, Parihar R, Tak A *et al*: **An invasive *Mimosa* in India does not adopt the symbionts of its native relatives**. *Ann Bot* 2013, **112**(1):179-196.
21. Mergeay M, Nies D, Schlegel HG, Gerits J, Charles P, Van Gijsegem F: ***Alcaligenes eutrophus* CH34 is a facultative chemolithotroph with plasmid-bound resistance to heavy metals**. *J Bacteriol* 1985, **162**(1):328-334.
22. Sato Y, Nishihara H, Yoshida M, Watanabe M, Rondal JD, Concepcion RN, Ohta H: ***Cupriavidus pinatubonensis* sp nov and *Cupriavidus laharis* sp nov., novel hydrogen-oxidizing, facultatively chemolithotrophic bacteria isolated from volcanic mudflow deposits from Mt. Pinatubo in the Philippines**. *Int J Syst Evol Microbiol* 2006, **56**:973-978.
23. Schlegel HG, Gottschalk G, Von Bartha R: **Formation and utilization of poly-beta-hydroxybutyric acid by Knallgas bacteria (*Hydrogenomonas*)**. *Nature* 1961, **191**:463-465.
24. Vandamme P, Coenye T: **Taxonomy of the genus *Cupriavidus*: a tale of lost and found**. *Int J Syst Evol Microbiol* 2004, **54**:2285-2289.
25. Pohlmann A, Fricke WF, Reinecke F, Kusian B, Liesegang H, Cramm R, Eitinger T, Ewering C, Potter M, Schwartz E *et al*: **Genome sequence of the bioplastic-producing "Knallgas" bacterium *Ralstonia eutropha* H16**. *Nature Biotechnol* 2006, **24**(10):1257-1262.
26. Simon MF, Grether R, de Queiroz LP, Särkinen TE, Dutra VF, Hughes CE: **The evolutionary history of *Mimosa* (Leguminosae): toward a phylogeny of the sensitive plants**. *Am J Bot* 2011, **98**:1201–1221.
27. **Pacific Island Ecosystems at Risk (PIER) *Mimosa pudica*** [http://www.hear.org/Pier/species/mimosa_pudica.htm]
28. **Report to the Government of Niue on invasive plant species of environmental concern** [http://www.hear.org/pier/pdf/niue_report.pdf]
29. Chen WM, Wu CH, James EK, Chang JS: **Metal biosorption capability of *Cupriavidus taiwanensis* and its effects on heavy metal removal by nodulated *Mimosa pudica***. *J Hazard Mater* 2008, **151**(2-3):364-371.
30. Reeve W, Ardley J, Tian R, Eshragi L, Yoon JW, Ngamwisetkun P, Seshadri R, Ivanova NN, Kyrpides NC: **A genomic encyclopedia of the root nodule bacteria: assessing genetic diversity through a systematic biogeographic survey**. *Stand Genom Sci* 2015, **10**:14.
31. Seshadri R, Reeve WG, Ardley JK, Tennessen K, Woyke T, Kyrpides NC, Ivanova NN: **Discovery of novel plant interaction determinants from the genomes of 163 Root Nodule Bacteria**. *Scientific Reports* 2015, **5**:16825.
32. De Meyer SE, Fabiano E, Tian R, Van Berkum P, Seshadri R, Reddy TBK, Markowitz V, Ivanova NN, Pati A, Woyke T *et al*: **High-quality permanent draft genome sequence of the *Parapiptadenia rigida* nodulating *Cupriavidus* sp strain**

UYPR2.512. *Stand Genom Sci* 2015, **10:80**.

33. De Meyer SE, Parker M, Van Berkum P, Tian R, Seshadri R, Reddy TBK, Markowitz V, Ivanova N, Pati A, Woyke T *et al*: **High-quality permanent draft genome sequence of the *Mimosa asperata* nodulating *Cupriavidus* sp strain AMP6**. *Stand Genom Sci* 2015, **10:80**.
34. Estrada-de los Santos P, Martínez-Aguilar L, López-Lara IM, Caballero-Mellado J: ***Cupriavidus alkaliphilus* sp nov., a new species associated with agricultural plants that grow in alkaline soils**. *Syst Appl Microbiol* 2012, **35**(5):310-314.
35. Janssen PJ, Van Houdt R, Moors H, Monsieurs P, Morin N, Michaux A, Benotmane MA, Leys N, Vallaeys T, Lapidus A *et al*: **The complete genome sequence of *Cupriavidus metallidurans* strain CH34, a master survivalist in harsh and anthropogenic environments**. *PLoS One* 2010, **5**(5).
36. Poehlein A, Kusian B, Friedrich B, Daniel R, Bowien B: **Complete genome sequence of the type strain *Cupriavidus necator* N-1**. *J Bacteriol* 2011, **193**(18):5017-5017.
37. Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, Billault A, Brottier P, Camus JC, Cattolico L *et al*: **Genome sequence of the plant pathogen *Ralstonia solanacearum***. *Nature* 2002, **415**(6871):497-502.
38. Monchy S, Benotmane MA, Janssen P, Vallaeys T, Taghavi S, van der Lelie D, Mergeay M: **Plasmids pMOL28 and pMOL30 of *Cupriavidus metallidurans* are specialized in the maximal viable response to heavy metals**. *J Bacteriol* 2007, **189**(20):7417-7425.
39. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen M, Angiuoli SV *et al*: **Towards a richer description of our complete collection of genomes and metagenomes "Minimum Information about a Genome Sequence " (MIGS) specification**. *Nature Biotechnol* 2008, **26**:541-547.
40. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I *et al*: **The Genomic Standards Consortium**. *PLoS Biol* 2011, **9**(6):e1001088.
41. Richter M, Rosselló-Móra R: **Shifting the genomic gold standard for the prokaryotic species definition**. *Proc Natl Acad Sci U S A* 2009, **106**:19126-19131.
42. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A: **Microbial species delineation using whole genome sequences**. *Nucleic Acids Res* 2015, **43**(14):6761-6771.
43. Goris J, Konstantinidis KT, Klappenbach J, Coenye T, Vandamme P, Tiedje JM: **DNA-DNA hybridization values and their relationship to whole-genome sequence similarities**. *Int J Syst Evol Microbiol* 2007, **57**:81-91.
44. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement**. *PLoS One* 2010, **5**:e11147.
45. Harrison PW, Lower RPJ, Kim NKD, Young JPW: **Introducing the bacterial "chromid": not a chromosome, not a plasmid.** *Trends Microbiol* 2010, **18**:141-148.
46. Parker MA: **A single sym plasmid type predominates across diverse chromosomal lineages of *Cupriavidus* nodule symbionts**. *Syst Appl Microbiol* 2015, **38**(6):417-423.
47. Vallenet D, Belda E, Calteau A, Cruveiller S, Engelen S, Lajus A, Le Fevre F, Longin C, Mornico D, Roche D *et al*: **MicroScope-an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data**. *Nucleic Acids Res* 2013, **41**(D1):E636-E647.
48. Jones KM, Kobayashi H, Davies BW, Taga ME, Walker GC: **How rhizobial symbionts invade plants: the *Sinorhizobium-Medicago* model**. *Nature Rev Microbiol* 2007, **5**(8):619-633.
49. Ren QH, Chen KX, Paulsen IT: **TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels**. *Nucleic Acids Res* 2007, **35**:D274-D279.
50. Saier MH, Tran CV, Barabote RD: **TCDB: the Transporter Classification Database for membrane transport protein analyses and information**. *Nucleic Acids Res* 2006, **34**:D181-D186.
51. Saier MH, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G: **The Transporter Classification Database (TCDB): recent advances**. *Nucleic Acids Res* 2016, **44**(D1):D372-D379.

52. Nies DH: **Efflux-mediated heavy metal resistance in prokaryotes**. *FEMS Microbiol Rev* 2003, **27**:313-339.
53. Saier MH: **Genome archeology leading to the characterization and classification of transport proteins**. *Curr Opin Microbiol* 1999, **2**(5):555-561.
54. Munkelt D, Grass G, Nies DH: **The chromosomally encoded cation diffusion facilitator proteins DmeF and FieF from *Wautersia metallidurans* CH34 are transporters of broad metal specificity**. *J Bacteriol* 2004, **186**(23):8036-8043.
55. Grass G, Otto M, Fricke B, Haney CJ, Rensing C, Nies DH, Munkelt D: **FieF (YiiP) from *Escherichia coli* mediates decreased cellular accumulation of iron and relieves iron stress**. *Arch Microbiol* 2005, **183**(1):9-18.
56. Wei YN, Fu D: **Selective metal binding to a membrane-embedded aspartate in the *Escherichia coli* metal transporter YiiP (FieF)**. *J Biol Chem* 2005, **280**(40):33716-33724.
57. De Angelis F, Lee JK, O'Connell JD, Miercke LJW, Verschueren KH, Srinivasan V, Bauvois C, Govaerts C, Robbins RA, Ruysschaert JM *et al*: **Metal-induced conformational changes in ZneB suggest an active role of membrane fusion proteins in efflux resistance systems**. *Proc Natl Acad Sci U S A* 2010, **107**(24):11038-11043.
58. Goldberg M, Pribyl T, Juhnke S, Nies DH: **Energetics and topology of CzcA, a cation/proton antiporter of the resistance-nodulation-cell division protein family**. *J Biol Chem* 1999, **274**(37):26065-26070.
59. Nies DH: **Heavy metal-resistant bacteria as extremophiles: molecular physiology and biotechnological use of *Ralstonia* sp CH34**. *Extremophiles* 2000, **4**(2):77-82.
60. Tseng TT, Gratwick KS, Kollman J, Park D, Nies DH, Goffeau A, Saier MH, Jr.: **The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins**. *J Mol Microbiol Biotechnol* 1999, **1**(1):107-125.
61. Diels L, Dong Q, van der Lelie D, Baeyens W, Mergeay M: **The *czc* operon of *Alcaligenes eutrophus* CH34: from resistance mechanism to the removal of heavy metals**. *J Ind Microbiol* 1995, **14**(2):142-153.
62. Monsieurs P, Moors H, Van Houdt R, Janssen PJ, Janssen A, Coninx I, Mergeay M, Leys N: **Heavy metal resistance in *Cupriavidus metallidurans* CH34 is governed by an intricate transcriptional network**. *Biometals* 2011, **24**(6):1133-1151.
63. Van Houdt R, Monchy S, Leys N, Mergeay M: **New mobile genetic elements in *Cupriavidus metallidurans* CH34, their possible roles and occurrence in other bacteria**. *Anton Leeuw Int J G* 2009, **96**(2):205-226.
64. Pak JE, Ekende EN, Kifle EG, O'Connell JD, De Angelis F, Tessema MB, Derfoufi KM, Robles-Colmenares Y, Robbins RA, Goormaghtigh E *et al*: **Structures of intermediate transport states of ZneA, a Zn(II)/proton antiporter**. *Proc Natl Acad Sci U S A* 2013, **110**(46):18484-18489.
65. Audic S, Robert C, Campagna B, Parinello H, Claverie JM, Raoult D, Drancourt M: **Genome analysis of *Minibacterium massiliensis* highlights the convergent evolution of water-living bacteria**. *PloS Genet* 2007, **3**(8):1454-1463.
66. Van Houdt R, Mergeay M: **Genomic context of metal response genes in *Cupriavidus metallidurans* with a focus on strain CH34**. In: *Metal Response in Cupriavidus metallidurans: Volume I: From Habitats to Genes and Proteins.* Edited by Mergeay M, Van Houdt R. Cham: Springer International Publishing; 2015: 21-44.
67. Franke S, Grass G, Rensing C, Nies DH: **Molecular analysis of the copper-transporting efflux system CusCFBA of *Escherichia coli***. *J Bacteriol* 2003, **185**(13):3804-3812.
68. Mergeay M, Monchy S, Vallaeys T, Auquier V, Benotmane A, Bertin P, Taghavi S, Dunn J, van der Lelie D, Wattiez R: ***Ralstonia metallidurans,* a bacterium specifically adapted to toxic metals: towards a catalogue of metal-responsive genes**. *FEMS Microbiol Rev* 2003, **27**(2-3):385-410.
69. García-Domínguez M, López-Maury L, Florencio FJ, Reyes JC: **A gene cluster involved in metal homeostasis in the cyanobacterium *Synechocystis* sp. strain PCC 6803**. *J Bacteriol* 2000, **182**(6):1507-1514.
70. López-Maury L, García-Domínguez M, Florencio FJ, Reyes JC: **A two-component signal transduction system involved in nickel sensing in the cyanobacterium *Synechocystis* sp PCC 6803**. *Mol Microbiol* 2002, **43**(1):247-256.
71. Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T *et al*: **Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120**. *DNA*

Res 2001, **8**(5):205-213; 227-253.

72. Sun X, Yu G, Xu Q, Li N, Xiao C, Yin X, Cao K, Han J, He Q-Y: **Putative cobalt- and nickel-binding proteins and motifs in *Streptococcus pneumoniae***. *Metallomics* 2013, **5**(7):928-935.
73. Diaz-Perez C, Cervantes C, Campos-Garcia J, Julian-Sanchez A, Riveros-Rosas H: **Phylogenetic analysis of the chromate ion transporter (CHR) superfamily**. *Febs J* 2007, **274**(23):6215-6227.
74. Juhnke S, Peitzsch N, Hubener N, Grosse C, Nies DH: **New genes involved in chromate resistance in *Ralstonia metallidurans* strain CH34**. *Arch Microbiol* 2002, **179**(1):15-25.
75. Henne KL, Nakatsu CH, Thompson DK, Konopka AE: **High-level chromate resistance in *Arthrobacter* sp strain FB24 requires previously uncharacterized accessory genes**. *BMC Microbiol* 2009, **9**:199.
76. Nies A, Nies DH, Silver S: **Nucleotide sequence and expression of a plasmid-encoded chromate resistance determinant from *Alcaligenes eutrophus***. *J Biol Chem* 1990, **265**(10):5648-5653.
77. Zhang YB, Monchy S, Greenberg B, Mergeay M, Gang O, Taghavi S, van der Lelie D: **ArsR arsenic-resistance regulatory protein from *Cupriavidus metallidurans* CH34**. *Anton Leeuw Int J G* 2009, **96**(2):161-170.
78. Alvarez AF, Rodriguez C, Georgellis D: **Ubiquinone and menaquinone electron carriers represent the yin and yang in the redox regulation of the ArcB sensor kinase**. *J Bacteriol* 2013, **195**(13):3054-3061.
79. Sharma P, Stagge S, Bekker M, Bettenbrock K, Hellingwerf KJ: **Kinase activity of ArcB from *Escherichia coli* is subject to regulation by both ubiquinone and demethylmenaquinone**. *PLoS One* 2013, **8**(10).
80. Monchy S, Benotmane MA, Wattiez R, van Aelst S, Auquier V, Borremans B, Mergeay M, Taghavi S, van der Lelie D, Vallaeys T: **Transcriptomic and proteomic analyses of the pMOL30-encoded copper resistance in *Cupriavidus metallidurans* strain CH34**. *Microbiol* 2006, **152**:1765-1776.
81. Rensing C, Grass G: ***Escherichia coli* mechanisms of copper homeostasis in a changing environment**. *FEMS Microbiol Rev* 2003, **27**(2-3):197-213.
82. Cha JS, Cooksey DA: **Copper resistance in *Pseudomonas syringae* mediated by periplasmic and outer-membrane proteins**. *Proc Natl Acad Sci U S A* 1991, **88**(20):8915-8919.
83. Trenor C, Lin W, Andrews NC: **Novel bacterial P-type ATPases with histidine-rich heavy-metal-associated sequences**. *Biochem Biophys Res Commun* 1994, **205**(3):1644-1650.
84. Reeve WG, Tiwari RP, Kale NB, Dilworth MJ, Glenn AR: **ActP controls copper homeostasis in *Rhizobium leguminosarum* bv. viciae and *Sinorhizobium meliloti* preventing low pH-induced copper toxicity**. *Mol Microbiol* 2002, **43**(4):981-991.
85. Vandecraen J, Monsieurs P, Mergeay M, Leys N, Aertsen A, Van Houdt R: **Zinc-induced transposition of Insertion Sequence elements contributes to increased adaptability of *Cupriavidus metallidurans***. *Front Microbiol* 2016, **7**(359).
86. Vandecraen J, Chandler M, Aertsen A, Van Houdt R: **The impact of insertion sequences on bacterial genome plasticity and adaptability**. *Crit Rev Microbiol* 2017, **43**(6):709-730.
87. Howieson JG, Dilworth MJ: **Working with rhizobia**. Australia: Australian Centre for International Agricultural Research (ACIAR); 2016.
88. Bennett S: **Solexa Ltd**. *Pharmacogenomics* 2004, **5**(4):433-438.
89. Zerbino DR: **Using the Velvet *de novo* assembler for short-read sequencing technologies**. *Current Protocols in Bioinformatics* 2010, **Chapter 11**:Unit 11 15.
90. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S *et al*: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data**. *Proc Natl Acad Sci U S A* 2011, **108**(4):1513-1518.
91. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: Prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinformatics* 2010, **11**:119.

92. Markowitz VM, Chen IA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang J, Woyke T, Huntemann M *et al*: **IMG 4 version of the integrated microbial genomes comparative analysis system**. *Nucleic Acids Res* 2014, **42**(D1):D560-D567.
93. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC: **IMG ER: a system for microbial genome annotation expert review and curation**. *Bioinformatics* 2009, **25**(17):2271-2278.
94. Lowe TM, Eddy SR: **tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic Acids Res* 1997, **25**(5):955-964.
95. Pruesse E, Quast C, Knittel K, Fuchs BdM, Ludwig W, Peplies J, Glöckner FO: **SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB**. *Nucleic Acids Res* 2007, **35**(21):7188-7196.
96. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. *Nucleic Acids Res* 1997, **25**(24):4876-4882.
97. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular evolutionary genetics analysis version 6.0**. *Mol Biol Evol* 2013, **30**(12):2725-2729.
98. Nicholas KB, Nicholas HB, Deerfield DW: **GeneDoc: analysis and visualization of genetic variation**. *EMBnetNews* 1997, **4**:14.
99. Saitou N, Nei M: **Reconstructing phylogenic trees.** *Mol Biol Evol* 1987:406-425.
100. Felsenstein J: **Confidence limits on phylogenies: An approach using the bootstrap**. *Evolution* 1985, **39**(4):783-791.

Tables

Table 1. Genome Statistics for *Cupriavidus* strain STM 6070.

Attribute	Value	% of Total
Genome size (bp)	6,771,773	100.00
DNA coding region (bp)	5,928,188	87.54
DNA G+C content (bp)	4551463	67.21
Number of scaffolds	107	
Total gene	6,182	100.00
RNA genes	64	1.04
rRNA operons*	1	0.02
Protein-coding genes	6,118	98.96
Genes with function prediction	5,050	81.69
Genes assigned to COGs	4,500	72.79
Genes assigned Pfam domains	5,305	85.81
Genes with signal peptides	677	10.95
Genes with transmembrane helices	1,402	22.68
CRISPR repeats	1	

Table 2. HME determinants in STM 6070 genome and their comparison with those detected in other *Cupriavidus* species.

Locus [a]	Proposed gene annotation	IMG Locus tag [b]	HM transporter [c]	Putative substrate	Orthologous genes						
		A3AGDRAFT_			LMG19424^T	STM6018	UYPR2.512	AMP6	CH34^T	N-1	H16
A	*hmyFCBA*	00876-00879	HME3b	monovalent cations	+	+	+	+	+	+	+
B	*nieIC cep nieBA*	00906-00910	HME5	nd	+	+	+	+	-	+	+
	chrB1A1	00912-00911	CHR	CrO_4^{2-}	+	+	+	-	+	-	-
C	*czcJ2 hns czcLRS ubiG zntA*	00978 (zntA)	P-ATPase	Zn^{2+}	+	+	+	+	+	+	+
D	*silP*	00999	P-ATPase	Ag	-	-	-	+	-	-	+
	copP	01005	P-ATPase	Cu^{2+}	-	-	+	+	+	-	+
E	*copSRABCD*	02911-02906	-	nd	+	+	+	+	+	+	+
F	*czcJ1I1C1B1A1*	03877-03882	HME1	Co^{2+}, Zn^{2+}, Cd^{2+}	+	+	+	+	+	+	+
G	*hmvCBA*	04395-04397	HME3a	divalent cations	+	+	+	+	+	+	+
H	*cupA cupR*	04504	P-ATPase	Cu^{2+}	+	+	+	+	+	+	+
I	*czcC2B2A2*	04975-04973	HME1	nd	+	+	+	+	+	+	+
	nreAB	04978-04977	MFS	Ni^{2+}	-	-	+	+	+	-	-
	ΔchrB 2A2CF exp chrL	04985-04980	CHR	CrO_4^{2-}	+	+	+	+	+	+	+
	hmxB zneAC zneR hmxS	04993-04989	HME3a	nd	+	+	+	+	+	+	+
J	*DcopF*	05405	P-ATPase	Cu^{2+}	+	+	+	+	+	+	+
	silDCBAF	05406-05409	HME4	Cu^{2+}, Ag	+	+	+	+	+	+	+
K	*czcDI2C3B3A3*	05655-05651	CDF +HME1	Co^{2+}, Zn^{2+}, Cd^{2+}	+	+	+	+	+	+	+
	arsR1IC1C2B1C3H1P	05663-05656	ARC3	nd	+	+	-	+	+	+	+
			+MFS		+	+	-	+	+	+	+
L	*arsR2C4B2H2*	05853-05856	ARC3	nd	+	+	-	+	+	+	+
Not in HMR clusters	*dmeF*	04892	CDF +HME1	nd	+	+	-	+	+	+	+
	fieF1	03199	CDF +HME1	nd	+	+	-	+	+	+	+
	fieF2	06164	CDF +HME1	nd	+	+	-	+	+	+	+

Footnote: STM 6070, *Cupriavidus* sp. STM 6070; LMG19424T, *C. taiwanensis* LMG19424T; STM6018, *C. taiwanensis* STM6018; UYPR2.512, *Cupriavidus* sp. UYPR2.512; AMP6, *Cupriavidus* sp. AMP6; CH34T, *C. metallidurans* CH34T; N-1, *C. necator* N-1T.

[a], HM loci names from Figure 2 and Table S6;

[b], IMG Locus tag (JGI);

[c], classification of the annotated HM proteins from Table S5

Table 3. Percentage of average nucleotide identities (ANI) for ANIb and ANIg (in brackets), and percentage of conserved DNA (in bold) among *Cupriavidus* and *Ralstonia* genomes.

Target	LMG 19424^T	STM 6018	STM 6070	UYPR2.512	AMP6	NBRC 13593^T	N1^T	JMP134	ASC-732^T	CH34^T
Query										
C. taiwanensis LMG 19424^T	---	**91.89**	**80.49**						**77.02**	
C. taiwanensis STM 6018	98.72 (99.03)	---	**84.55**						**57.96**	
Cupriavidus neocaledonicus. STM 6070	93.33 (94.6)	93.4 (94.7)	---						**76.48**	
Cupriavidus sp. UYPR2.512	87.22 (89.75)	87.23 (89.67)	87.29 (89.62)	---			**63.87**			
Cupriavidus sp. AMP6	85.66 (88.40)	85.71 (88.43)	85.65 (88.35)	85.74 (88.11)	---	**73.38**				
C. oxalaticus NBRC13593^T	85.84 (88.72)	85.81 (88.75)	85.93 (88.84)	86.23 (88.69)	92.69 (94.44)	---				
C. necator N1^T	86.7 (89.68)	86.62 (89.60)	86.46 (89.48)	93.43 (95.27)	85.3 (88.28)	86.36 (88.80)	---			
C. pinatubonensis JMP134	81.36 (83.91)	81.25 (83.84)	81.42 (83.97)	81.76 (84.05)	81.48 (83.92)	81.7 (84.12)	81.58 (83.91)	---		
C. alkaliphillus ASC-732^T	92.54 (93.90)	92.59 (93.89)	93.32 (94.62)	88.18 (90.03)	86.11 (88.42)	86.0 (88.88)	88.28 (90.07)	81.31 (83.90)	---	
C. metallidurans CH34^T	78.88 (81.60)	78.67 (81.38)	78.75 (81.39)	78.71 (81.29)	78.67 (81.34)	78.82 (81.68)	78.78 (81.29)	78.45 (80.87)	78.88 (81.48)	---

ANIb values were calculated with JSpecies (based on whole genome BLAST alignment) [42]. ANIg values (in brackets) were calculated using the ANI tool in IMG [43]. The % of conserved DNA values are shown for ANIb values higher than 90% (in bold). Values in red font reveal strains that belong to the same species. Genomes were downloaded from Genbank accessions when already published. Species compared: *Cupriavidus taiwanensis* strains LMG 19424^T and STM 6018; *Cupriavidus neocaledonicus* STM 6070; *Cupriavidus* sp. UYPR2.512; *Cupriavidus* sp. AMP6; *C. alkaliphilus* strain ASC-732^T; *C. necator* N1^T; *C. oxalaticus* NBRC 13593^T; *C. pinatubonensis* JMP134; *C. metallidurans* CH34^T.

Table 4. Comparison of TransAAP identified transporter genes in the genomes of *Cupriavidus neocaledonicus* STM 6070 and other *Cupriavidus* species.

Transporter classification	Number[a]	Number of genes detected in genomes of *Cupriavidus* species[b]						
		Cmet CH34[T]	*Cnec* H16	*Cnec* N-1[T]	*Cpin* JMP134	*Ctai* LMG 19424[T]	*Cneo* STM 6070	*Cneo* STM 6070 HMR genes[c]
MFS Superfamily	2.A.1	89	108	142	125	87	106	**2**
CDF Family	2.A.4	5	3	2	4	4	4	**4**
RND Superfamily	2.A.6	26	22	28	22	18	31	**8**
CHR Family	2.A.51	5	2	4	4	2	3	**2**
NiCoT Family	2.A.52	1	2	0	0	0	0	**0**
ACR3 Family	2.A.59	1	1	1	1	2	2	**2**
P-ATPase Superfamily	3.A.3	13	11	8	10	6	10	**5**
Total		140	149	185	166	119	156	**23**

[a]Subclass of transporters as defined in [51, 52], also see in the text. [b]*Cmet, C. metallidurans*; *Cnec, C. necator*; *Cneo, C. neocaledonicus; Cpin, C. pinatubonensis*; and *Ctai, C. taiwanensis.* Data includes information from Janssen *et al.* [35]. [c]Genes identified in this study as encoding proteins putatively involved in metal tolerance.

Figures

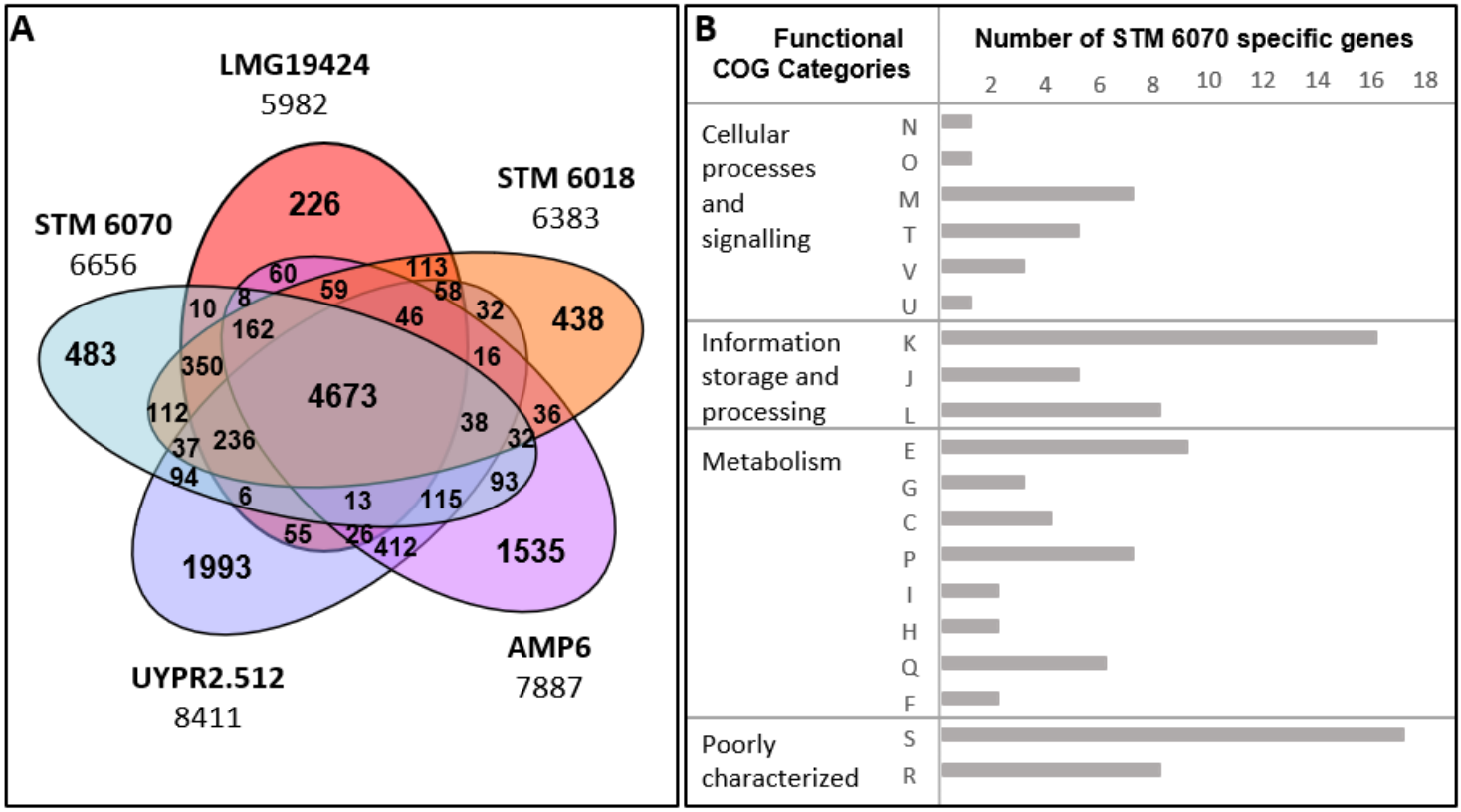

Figure 1

Gene content analysis of the STM 6070 genome. A: Venn diagram of gene number counts of symbiotic Cupriavidus strains; B: functional COG categories of STM 6070 specific genes (107 assigned genes out of 483). STM 6070, Cupriavidus neocaledonicus STM 6070; STM 6018, C. taiwanensis STM 6018; LMG 19424T, C. taiwanensis LMG 19424T; AMP6, Cupriavidus sp. AMP6; UYPR2.512, Cupriavidus sp. UYPR2.512. Numbers under the strain names describe the total number of genes for each corresponding genome. The analysis was performed using the “Gene Phyloprofile” tool in the Microscope MaGe platform [47], https://www.genoscope.cns.fr/agc/microscope/mage). The orthologous counterparts in the genomes were detected by applying a minimum of 30% for protein sequences identity over a minimum of 80% of the protein length (>30% protein MinLrap 0.8).

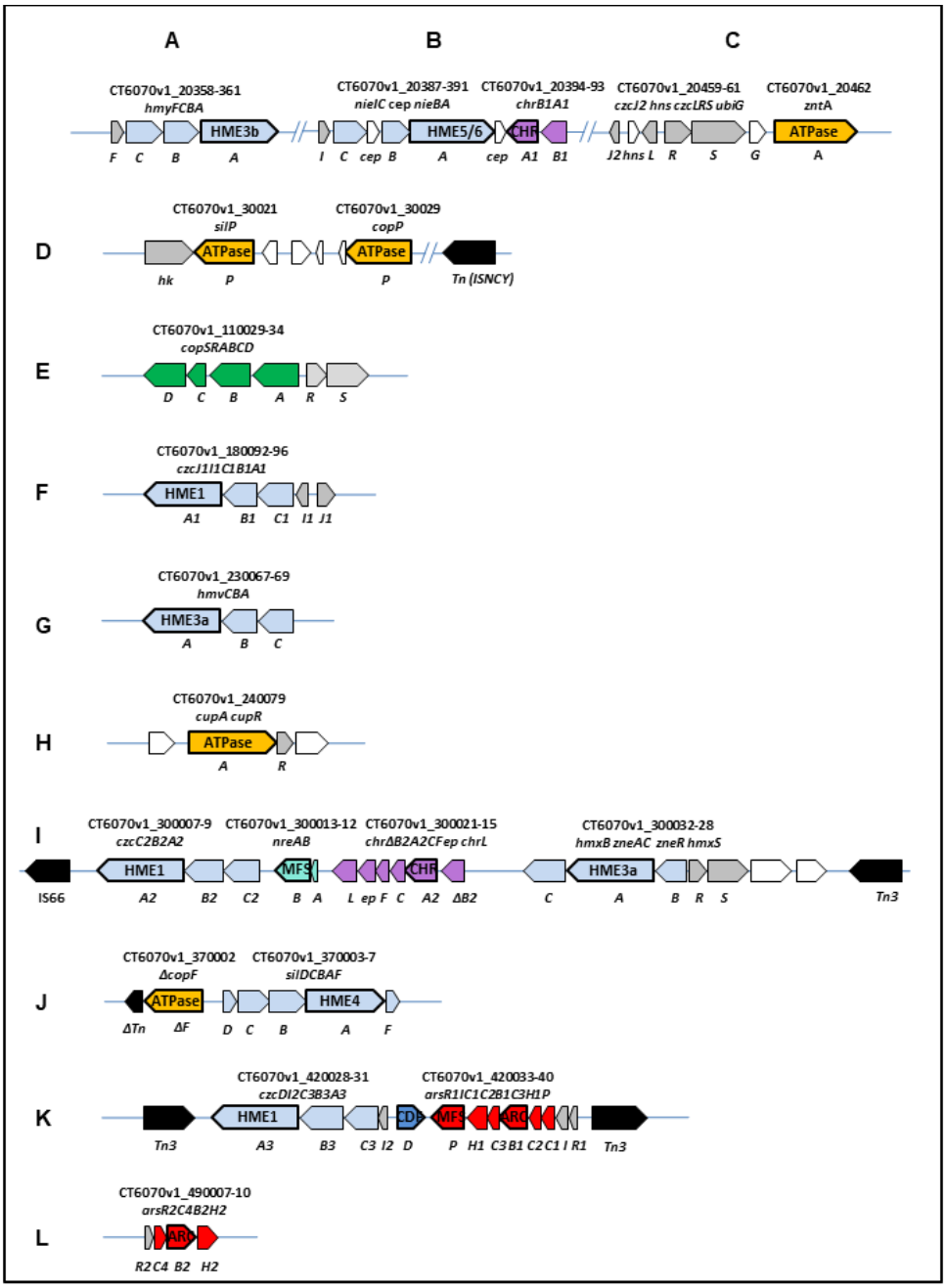

Figure 2

Cupriavidus neocaledonicus STM 6070 HMR gene clusters containing annotated putative genes encoding proteins involved in heavy metal efflux (HME). A to L: HMR loci (see also Table S6). Colour coding: light blue, HME-RND systems composed of canonical CBA genes [52]; dark blue, czcD encoding a CDF type protein; turquoise, nre genes; dark and light grey, putative corresponding regulatory genes; green, cop genes; purple, chr genes; red, ars genes; yellow, P-ATPase encoding genes; white,

genes encoding putative proteins of unknown function; black, transposases; cep: conserved exported protein; ep, exported protein; hk, histidine kinase. Truncated genes are identified with a delta (Δ) symbol. Thick lines identify genes encoding the transmembrane proteins. Gene coordinates for STM 6070 (CT6070v1_XXXXXX-XX) correspond to the annotation in the MaGe Microscope platform (https://www.genoscope.cns.fr/agc/microscope/mage/viewer.php) (see Table S6 for the corresponding IMG locus tags).

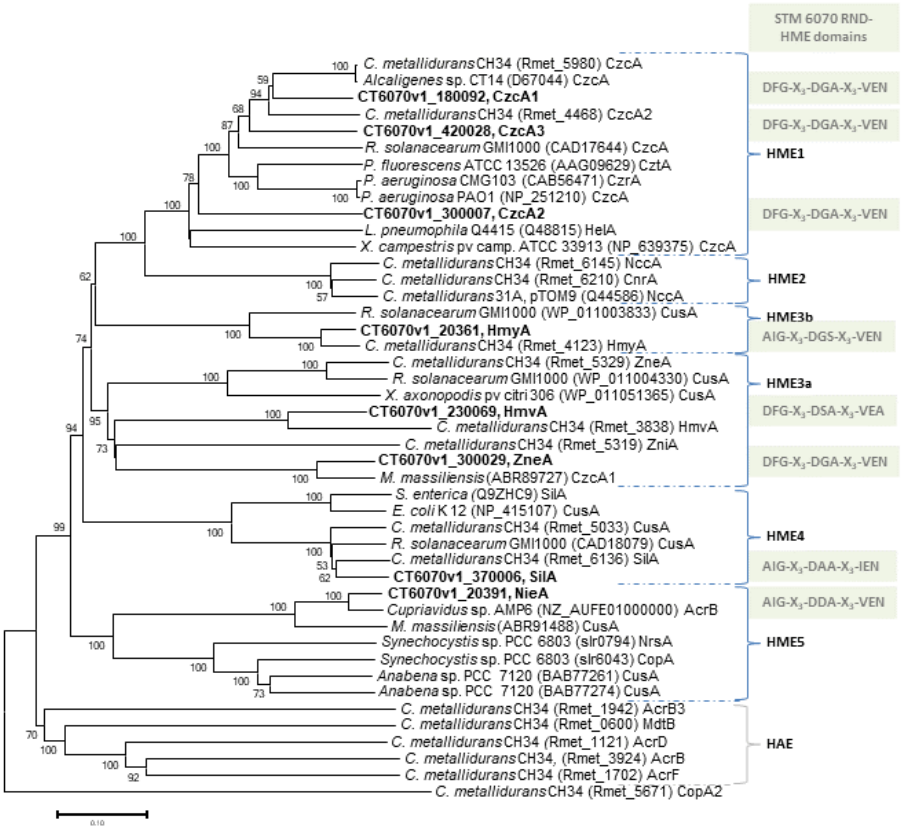

Figure 3

Phylogenetic tree of the RND transmembrane proteins (component A of the CBA transport system) from Cupriavidus neocaledonicus STM 6070 (in bold) and from sequenced genomes of reference Cupriavidus (C.), Escherichia (E.), Legionella (L.), Microbacterium (M.), Pseudomonas (P.), Ralstonia (R.) and Xanthomonas (X.) and other named reference strains. The HME class of the protein is designated according to the current classification scheme [52, 59, 60]. HME1 to 5 represent five classes of HME systems, HAE represents here an RND protein group involved in in export of hydrophobic and amphiphilic compounds. The evolutionary history was inferred by the Neighbor-Joining method with a bootstrap consensus tree inferred from 500 replicates. The evolutionary distances were computed using the Poisson correction method and are presented as

the number of amino acid substitutions per site. The rate variation among sites was modelled with a gamma distribution (shape parameter = 1). Evolutionary analyses were conducted in MEGA6. STM 6070 HME locus tags are displayed in Figure 2 and Table S6. GenBank accession numbers or locus tags are given in brackets. For C. metallidurans CH34T only the gene numbers of the annotated genome (NC_007973) are given.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2STM6070.docx
- TableS7STM6070.xlsx
- TableS5STM6070.docx
- FigureS3STM6070.pptx
- FigureS1STM6070.pptx
- FigureS6STM6070.pptx
- FigureS5STM6070.pptx
- TableS1STM6070.docx
- FigureS2STM6070.pptx
- TableS3STM6070.docx
- FigureS4STM6070.pptx
- TableS4STM6070.docx
- TableS6STM6070.xlsx