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 1, Lionel Moulin 1, Julie Kaye Ardley 2, Florence Braun 3, Margaret Mary Gollagher 4, Jaco Daniel Zandberg 2, Dora Vasileva Marinova 4, Marcel Huntemann 5, T. B. K. Reddy 5, Neha Jacob Varghese 5, Tanja Woyke 5, Natalia Ivanova 5, Rekha Seshadr 5, Nikos Kyrpides 5 and Wayne Gerald Reeve 2*

**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 6118 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 19424 T. 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, 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.

(Continued on next page)
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 noduleation 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\textsuperscript{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\textsuperscript{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\textsubscript{2} and CO\textsubscript{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\textsuperscript{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 nineteenth 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\textsuperscript{-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 heavy metal resistance (HMR) loci, including putative ars, ccc, 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* CH34T (a model organism for heavy metal resistance [21]) and its heavy metal-sensitive derivative AE104 (CH34T devoid of the plasmids pMOL28 and pMOL30 that confer heavy-metal-resistance [38]) at various concentrations of Ni2+ (Figure S2). Of the tested strains, STM 6070 had the highest tolerance to Ni2+ and was the only strain capable of growth at 15 mM NiSO4.

*C. metallidurans* CH34T grew in the presence of 10 mM NiSO4, while AE104 was unable to grow at 3 mM NiSO4. Previous studies had established that other symbiotic *C. taiwanensis* strains LMG 19424T from Taiwan [13] and *C. taiwanensis* STM 6018 from French Guiana [6] were also unable to grow at 3 mM NiSO4 (data not shown).

In light of the observed Ni2+ tolerance of STM 6070, we examined the tolerance of the *Cupriavidus* symbionts to other metal ions. In the presence of Cu2+, STM 6070, 6018 and LMG 19424T were able to grow in media containing 1.0 mM Cu2+, however, growth of STM 6070 was inhibited from 0.6 mM Cu2+ (Figure S3). In addition, STM 6070 was able to grow in media containing 15 mM Zn2+, whereas STM 6018 and LMG 19424T were far more sensitive and could not grow at this concentration (data not shown). Since STM 6070 was highly tolerant to Ni2+ and Zn2+, 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 6182 genes, of which 6118 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 19424T [11] and *C. alkaliphilus* ASC-732T [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 19424T, 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, STM 6018, UYPR2.512 and AMP6 to the

| 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) | 4,551,463 | 67.21     |
| Number of scaffolds | 107       |            |
| Total gene         | 6182      | 100.00     |
| RNA genes          | 64        | 1.04       |
| rRNA operons*      | 1         | 0.02       |
| Protein-coding genes | 6118   | 98.96     |
| Genes with function prediction | 5050 | 81.69 |
| Genes assigned to COGs | 4500   | 72.79     |
| Genes assigned Pfam domains | 5305 | 85.81     |
| Genes with signal peptides | 677   | 10.95     |
| Genes with transmembrane helices | 1402 | 22.68     |
| CRISPR repeats     | 1         |            |

*1 copy of 16S rRNA and 4 copies of 5S rRNA
finished genome of \textit{C. taiwanensis} LMG 19424$^\text{T}$ (Fig. 1). The alignments of the STM 6018 and STM 6070 genomes against that of \textit{C. taiwanensis} LMG 19424$^\text{T}$ showed a high similarity of collinear blocks within the two largest replicons (Fig. 1a), the sequence of the LMG 19424$^\text{T}$ 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 LMG 19424$^\text{T}$ genome sequence, as well as two STM 6070 scaffolds (A3AGDRAFT\_scaffold\_84.85.C and \_75.76.C) that were absent from LMG 19424$^\text{T}$ 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$^\text{T}$ (see shaded area on Fig. 1a).

In contrast, the genome alignment of UYPR2.512 and AMP6 with LMG 19424$^\text{T}$ showed important differences in replicon conservation (Fig. 1b). Earlier studies on comparative genomics of \textit{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 \textit{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” [46], 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 LMG 19424$^\text{T}$ 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 (Fig. 1) in CHR2 as compared with CHR1 in the symbiotic Cupriavidus genomes.

Finally, we observed that whereas most of the LMG 19424 pSym sequence was well conserved in the STM 6018 and STM 6070 genomes (Fig. 1a), only a few LMG 19424 pSym genes (including the nod, nif, fix and fdx genes) were conserved across all five genomes. The M. pudica microsymbionts (LMG 19424, STM 6018 and STM 6070) had almost identical pSym (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 pSym (low synteny, with nod genes characterized by 80–94 and 95–98.4% protein identity to those of LMG 19424, 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 [47].

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, STM 6018, UYPR2.512 and AMP6, performed using the “Gene Phyloprofile” tool in the Microscope MaGe platform [48] (Fig. 2), 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 to 1993 for UYPR2.512; larger genomes had a greater number of unique genes (Fig. 2). 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 (Fig. 1b). Within the functional COG category “Cellular processes and signaling”, the largest number of genes were found in Cell wall/membrane/envelope 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 [49].

Unique STM 6070 genes within the signal transduction category included four genes encoding putative universal stress proteins (UsP 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 (ArcA, CopG, GntR, LaCl, 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/) [50] 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/) [45, 51, 52]. TransAAP analysis revealed a total of 834 putative transporters within STM 6070, of which 156 were classified within the MFS, CDF, RND, or CBA transporters [45], to differentiate them from the cytoplasm, or the periplasm, to the outside of the cell, and have been designated as CBA efflux systems, or CBA transports [45], to differentiate them from ABC transporters. Within a CBA system, the RND transmembrane and MFP proteins [45, 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 proteins, which 156 were classified within the MFS, CDF, RND, or CBA transporters [45]. 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].

**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 nreP) were associated with HME functions. The nreB gene located in the nreAB operon (cluster I), and the nreP gene located in the arsenic efflux systems, respectively (Fig. 3) [45].

**Cation diffusion facilitator (CDF) proteins**
The CDF proteins are single-subunit systems located in the cytoplasmic membrane that act as chemiosmotic ion-proton exchangers [52]. They include HMR proteins such as CzcD, which provides resistance to cobalt, zinc and cadmium [45]. Four genes encoding CDF proteins were detected in the STM 6070 genome (Table S6), but only one, czcD, is located in an HME cluster (czcDl2C3B3A3, cluster K) (Fig. 3). 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 transports [45], to differentiate them from ABC transporters. Within a CBA system, the RND transmembrane and MFP proteins [45, 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 proteins, which 156 were classified within the MFS, CDF, RND, or CBA transporters [45]. 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].
α-helices, TMH I to TMH XII [45]. 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 [45, 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 [45, 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 (Fig. 3 and Table S6). The STM 6070 genome lacks genes encoding the HME2-type transmembrane proteins, such as the *C. metallidurans* CH34T CnrA and NccA, which are involved in heavy metal resistance and have predicted substrate specificity for cobalt and nickel [45].

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

| Locus | Proposed gene | IMG Locus | HM transporter | Putative substrate | Orthologous genes |
|-------|---------------|-----------|----------------|-------------------|-------------------|
|       | annotation    | tag       |                |                   |                   |
| A     | hmyFCBA       | A3AGDRATF_00876–00879 | HME3b         | monovalent cations | STM 19424T   |
| B     | *nieC cep nieBA* | 00906–00910 | HME5 | nd                | STM 19424T   |
|       | *chaB1A9*     | 00912–00911 | CHR | CrO4^2−           | STM 19424T   |
| C     | *czcJ2 hns czcLRS ubIG* | 00978 (zntA) | P-ATPase | Zn2^+           | STM 19424T   |
| D     | *sip pep*     | 00999     | P-ATPase | Ag               | STM 19424T   |
| E     | *copSRABCD*   | 01005     | nd           | Cu2^+            | STM 19424T   |
| F     | *czcJ11C1B1A1* | 03877–03882 | HME1 | Co2^+, Zn2^+, Cd2^+ | STM 19424T |
| G     | *hmxB*        | 04395–04397 | HME3a | divalent cations | STM 19424T   |
| H     | *cupA cupR*   | 04504     | P-ATPase | Cu2^+            | STM 19424T   |
| I     | *czcC2B2A2*   | 04975–04973 | HME1 | nd                | STM 19424T   |
|       | *nieAB*       | 04978–04977 | MFS | Ni2^+             | STM 19424T   |
| J     | *Δcbr2 2A2CF exp* | 04985–04980 | CHR | CrO4^2−           | STM 19424T   |
| K     | *hmxB zneAC zneR* | 04993–04989 | HME3a | nd                | STM 19424T   |
| L     | *ΔcopF*       | 05405     | P-ATPase | Cu2^+            | STM 19424T   |
|       | *silDCRAF*    | 05406–05409 | HME4 | Cu2^+, Ag         | STM 19424T   |
|       | *czcD2C3B3A3* | 05655–05651 | CDF | HME1 | Co2^+, Zn2^+, Cd2^+ | STM 19424T |
|       | *arsR1K1C2B1C3H1P* | 05663–05656 | ARC3 | nd                | STM 19424T   |
|       | *arsR2C4B2K2* | 05853–05856 | ARC3 | nd                | STM 19424T   |
|       | *dmeF*        | 04892     | CDF +        | +MFS             | STM 19424T   |
|       | *fieF1*       | 03199     | CDF +        | +MFS             | STM 19424T   |
|       | *fieF2*       | 06164     | CDF +        | +MFS             | STM 19424T   |

Footnote: STM 6070, *Cupriavidus* sp. STM 6070; LMG 19424T, *C. taiwanensis* LMG 19424T; STM 6018, *C. taiwanensis* STM 6018; UYPR2.512, *Cupriavidus* sp. UYPR2.512; AMP6, *Cupriavidus* sp. AMP6; CH34T, *C. metallidurans* CH34T; N-1, *C. necator* N-1 T

* a, HM loci names from Fig. 2 and Table S6
  b, IMG Locus tag (JGI);
  c, classification of the annotated HM proteins from Table S5

---

**RND-HME1**

STM 6070 contained 3 RND-HME1 encoded proteins (CzcA1-A3), characterized on the basis of homology to the canonical CzcA proteins of CH34T. STM 6070 CzcA1 and CzcA3 grouped with CH34T CzcA and CczA2, while STM 6070 CzcA2 formed an outgroup (Fig. 4). The *C. necator* czcA1 gene was within an operon located in cluster F (Fig. 3) and annotated as *czcI11C1B1A1*. In addition to the *czcCBA* genes, this cluster contained a *czcI* homolog to a transcriptional regulator that has been shown to control the expression of *czcI1B1A1* [45, 61] and a *czcI* 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 \textit{Cupriavidus} strains and in \textit{C. necator} N-1, and the STM 6070 CzcA1 protein showed high identity with other \textit{Cupriavidus} CzcA orthologues (Table S6). In \textit{C. metallidurans} CH34\(^T\), the corresponding \textit{czc} cluster (\textit{czc}MNICBADRSEE, locus tags Rnet_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 (\textit{czc}CB2A2) formed part of a large group of HMR loci within cluster I (Fig. 3). Immediately upstream of \textit{czc}CB2A2 is a \textit{nreB} gene, encoding a putative nickel resistance MFS protein. A similar arrangement has been observed for the CH34\(^T\) \textit{ncc}B\textit{A} \textit{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 czc2C3B3A) was located within cluster K (Fig. 3). The czc2 and czcD components encode putative regulator and CDF proteins, respectively. This cluster, which was also delimited by two Tn3 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 (Fig. 3). 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 ISJ088 [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 silDCBAF operon (Rmet_5030–5034), which encodes a putative silver efflux system, and to the CH34 T chromosome *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 nieIC 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 (Fig. 4) 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-helix (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 *nieC 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 CH34T *pMOL28* and chromid *ChrA* proteins. The *chrB1A1* operon (cluster B) was located upstream of the putative RND-HME5 efflux system *nieIC-cep-nieBA* (Fig. 3). This *chr* operon was conserved in the genomes of the symbiotic *Cupriavidus* strains LMG 19424T and STM 6018, 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 (Fig. 3). 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 Concanaavalin A-like lectin/glucanases domain; and finally, *chrL*, encoding a lipoprotein (protein family, LppY/LpqO [75]) with 71.1% identity to CH34T *ChrL* (also annotated as CH34T *ChrY*). Corresponding gene clusters were identified in the UYPR2.512 and CH34T genomes. In CH34T 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 CH34T *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 CH34T *chrB*. Since ChrB seems to be important for chromate resistance in CH34T [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 CH34T *arsMRJC2BC1HP* 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 LMG 19424T and STM 6018 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 clade 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 (Fig. 2 and Table 2). The *copF* P-type ATPase gene in cluster J was located upstream of the *silDCAFB* operon and could encode an essential copper efflux component, as shown for CH34T [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 SiIP (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* (Fig. 3), in a conserved large synteny 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\textsuperscript{T} that have been annotated as \textit{zntA czcICAB} (locus tags Rmet\textsubscript{4594}–4597) and \textit{czcBA ubiG czcSRL IS hns mmmQ} (locus tags Rmet\textsubscript{4469}–4461), respectively. These CH34\textsuperscript{T} clusters encode an RND system (\textit{czcICBA}), the ZntA ATPase, a two-component regulatory system CzcRS and a 3-demethlyubiquinone-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 \textit{czcL}, \textit{hns} and \textit{mmmQ} genes encode an unknown protein, an H-NS like protein and a small stress responsive protein, respectively. Genes in the second CH34\textsuperscript{T} cluster may be inactivated by an insertion sequence located between \textit{czcL} and \textit{hns}. The synteny of the STM 6070 cluster C is perfectly conserved in the genomes of the four symbiotic \textit{Cupriavidus} strains, suggesting that it is functional, but it is devoid of the \textit{czcICBA} RND system found in the corresponding CH34\textsuperscript{T} cluster. The role of the regulatory loci \textit{czcLRS-ubiGI}, with regard to \textit{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\textsuperscript{T} led to identification of a copper-resistance operon \textit{copRSABCD} (cluster E). This had a similar structure to the CH34\textsuperscript{T} \textit{cop} cluster (\textit{copS2-R2A2B2C2D2}) 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\textsuperscript{1+} to Cu\textsuperscript{2+}. CopA proteins contain several motif variants of MGGM/MAGM/MGAM/MSGM, possibly involved in binding numerous Cu\textsuperscript{1+} ions, as determined for \textit{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\textsuperscript{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\textsuperscript{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 (GHS)), 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 \textit{Escherichia coli} HRA-1 and 2, \textit{Enterococcus hirae} CopB [83] and \textit{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 \textit{Cupriavidus} isolates.

Location of HMR determinants

The detected STM 6070 HMR determinants in the 12 clusters (A to L, Fig. 3, Table 2) were assigned to putative replicons of the STM 6070 genome, following alignment of contigs to the finished LMG 19424\textsuperscript{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 S5). Therefore, STM 6070 appears to carry the great majority of its HMR clusters on CHR2. In contrast \textit{C. metallidurans} CH34\textsuperscript{T} harbours 8 out of 24 HMR clusters on CHR2 (chromid) [35, 45, 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 \textit{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 \textit{C. taiwanensis} isolates (LMG 19424\textsuperscript{T} and STM 6018, Table S6), reflecting the taxonomic relationship with \textit{C. taiwanensis}.

Synteny analysis indicated that the specific STM 6070 HMR clusters B, D, I and J were absent from all other analysed \textit{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 \textit{Cupriavidus} strains. Cluster K was perfectly conserved within the LMG 19424\textsuperscript{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 \textit{czc} and \textit{ars} operons from cluster K were detected in the non-symbiotic \textit{Cupriavidus} genomes, with encoded protein identities of 76–77% and 83–88%, respectively, to the STM 6070 \textit{czc} and \textit{ars} operon encoded proteins. This observation can be explained by the location of cluster K on the pSym, which, as proposed recently [47], seems to be largely shared between \textit{M. pudica} microsymbionts of different genomic backgrounds. Indeed, we demonstrated by the progressive Mauve alignment (Fig. 1) that the pSym seems to be conserved in the genomes of the \textit{M. pudica}-nodulating LMG 19424\textsuperscript{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 (Fig. 3). 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 Tn3 and IS66 type, and carries four different HMR determinants, including czcC2B2A2 and hmxB zneAC. Cluster K is flanked by two Tn3 transposases, however, unlike Cluster I there is a high conservation of architecture and gene identity with the closely related C. taiwanensis strains (LMG 19424T 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 nieIC cep nieBA). This operon, which is absent from LMG 19424T and STM 6018 genomes, is located in a large highly conserved region, suggesting a gene loss from C. taiwanensis genomes. Interestingly, nieIC 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].

### Conclusion

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 cze 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 S5).

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 19424T 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 [47].

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, cze2 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 NiSO₄ (0, 3, 5, 10 and 15 mM), Cu(NO₃)₂ (0, 0.3, 0.6 and 1.0 mM) and ZnSO₄ (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 OD₆₀₀nm in a spectrophotometer.

Genomic DNA preparation

C. neocaledonicus STM 6070 was streaked onto TY solid medium [87] and grown at 28°C for 3 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 OD₆₀₀nm 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 3960.4 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 0–1100–2100 -r 0 -R 0 -X 0) 0) 3) Allpaths–LG (PrepareAllpathsInput:PHRED64 = 1 PLOIDY = 1 FRAGCOVERAGE = 125 JUMPCOVERAGE = 25 LONGJUMPCOV = 50, RunAllpath-s-LG: 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 (Fig. 2), 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 MinL.rap 0.8). The homologous genes were then removed from the resulting list. Transport systems were identified using the TransAAP tool [50] (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): ANIlg [42] and ANIB [43]. In order to perform the alignments using progressive Mauve software [51], the scaffolds of each draft genome (STM 6070, STM 6018, UYPR2.512 and AMP6) were firstly reordered using Mauve software on the basis of the C. taiwanensis LMG 19424T 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 LMG 19424T 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 CH34T 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 (Fig. 4), 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 syntentic arrangements of the eight HME-RND efflux systems present in STM 6070 and compare with the HME-RND efflux systems found within six other Cupriavidus strains, as outlined below.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6623-2.

Additional file 1: Figure S1. Images of Cupriavidus neocaledonicus STM 6070 using scanning (Left) and transmission (Centre) electron microscopy and the appearance of colony morphology on solid media (Right). Images depicted here were imaged by the authors.

Additional file 2: Figure S2. Bacterial growth in 284 Tris-medium, in absence (-) and in presence of NISO4 (● 5 mM, • 10 mM, ◆ 15 mM), STM 6070, studied isolate; CH34, Cupriavidus metallidurans CH34; AE104, plasmid cured derivative of C. metallidurans CH34.

Additional file 3: Figure S3. Tolerance to copper of symbiotic Cupriavidus strains STM 6070, 6018 and LMG 19424T, performed in 284 Tris-culture medium (Mergeay et al, 1985) with 0, 0.3, 0.6 and 1.0 mM of Cu (NO3)2 concentrations.

Additional file 4: Figure S4. Phylogenetic tree showing the relationship of Cupriavidus sp. STM 6070 (shown in bold print) to other members of the order Burkholderiales based on aligned sequences of the 16S rRNA gene (1290 bp internal region). All sites were informative and there were no gap-containing sites. Phylogenetic analyses were performed using MEGA, version 6 [97]. The tree was built using the Neighbor-Joining method [99]. Bootstrap analysis [100] with 1000 replicates was performed to assess the support of the clusters. Type strains are indicated with a superscript T, strains with available genomes are indicated with *, and symbiotic Cupriavidus strains are indicated with blue asterisks. Brackets after the strain name contain a DNA database accession number. For LMG 19424, STM 6018, STM 6070, AMP6, JMP134, H16, N-1 and CH34CH34T the 16S rRNA sequences were recovered from sequenced genomes. Published genomes are indicated with an asterisk.

Additional file 5: Figure S5. Circular representation of symbiotic Cupriavidus genomes (by BlastAtlas using the CGView) aligned to the STM 6070 genome. The STM 6070 contigs were first aligned to the three replicons of LMG 19424T Chr1/pSym/Chromid. Circles, from inside out, show GC skew (purple), GC content (black) and genomes of (1, dark grey) STM 6070; (2, green) LMG 19424T; (3, green) STM 6018; (4, orange) AMP6 (5, purple) UYPR2.512 and (6, red) UYMMa02A. The HME clusters A to L are marked with corresponding letters. Triangles with black borders represent clusters unique to STM 6070 and triangles without borders represent general HMR clusters.

Additional file 6: Table S1. General attributes and Minimum Information for the Genome Sequence (MIGS) of Cupriavidus strain STM 6070.

Additional file 7: Table S2. Number of protein coding genes of STM 6070 associated with the general COG functional categories.

Additional file 8: Table S3. Cupriavidus strains compared in this study (as bacterial isolates and/or sequenced genomes).

Additional file 9: Table S4. Percentage of average nucleotide identities (ANI) for ANIlg and ANIB (in brackets), and percentage of conserved DNA (in bold) among Cupriavidus and Ralstonia genomes.

Additional file 10: Table S5. Comparison of TransAAP identified transporter genes in the genomes of Cupriavidus neocaledonicus STM 6070 and other Cupriavidus species.

Additional file 11: Table S6. HME determinants in STM 6070 genome and their comparison with those detected in Cupriavidus species.

Additional file 12: Figure S7. A comparison of proximal and distal sites in Cupriavidus neocaledonicus STM 6070 RND-HME proteins compared to related proteins in other bacterial strains.

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

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

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, JZ, 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.

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-JCI-0046, Curtin University Sustainability Policy Institute, and the funding received from Murdoch University Small Research Grants Scheme.

Availability of data and materials
The datasets generated and/or analysed during the current study are available in the NCBI repository under bioproject ID 165313, [https://www.ncbi.nlm.nih.gov/bioproject?term=165313] and the accession number 2513237165. All protein sequences are available in the databases above and can also be accessed through Uniprot [https://www.uniprot.org/].

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
1IRD, Cirad, Univ. Montpellier, Interactions Plantes Microorganismes Environnement (IPME), 34394 Montpellier, France. 2College of Science, Curtin University, Bentley, Australia. 3DOE Joint Genome Institute, Walnut Creek, USA.

Received: 2 October 2019 Accepted: 25 February 2020

Published online: 06 March 2020

References
1. Gyampah P, Hirsch AM, Moulin L, Chen WM, Elliott GN, Bontemps C, Estrada–Los Santos P, Gross E, Dos Reis FR, Sprent JI, et al. Legume-nodulating betaproteobacteria: diversity, host range, and future prospects. Mol Plant-Microbe Interact. 2011;24(11):1276–88.

2. Moulou S, Munve A, Dreyfus B, Boivin-Masson C. Nodulation of legumes by members of the beta-subclass of Proteobacteria. Nature. 2001;408(6804):948–50.

3. Bournaud C, de Faria SM, Dos Santos JMF, Tisseyre P, Silva M, Chaintreuil C, Curtin University Sustainability Policy Institute, Curtin University, Bentley, Australia. 3DOE Joint Genome Institute, Walnut Creek, USA.

4. Klonowska A, Chaintreuil C, Tisseyre P, Miché L, Melkonian R, Ducuzzo M, Laguerre G, Brunel B, Moulou L. Biodiversity of Mimosa pudica rhizobial symbionts (Cupriavidus taiwanensis, Rhizobium mesoamericanum) in New Caledonia and their adaptation to heavy metal-rich soils. FEMS Microbiol. 2012;88(6):1692–700.

5. 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. 2012;80:417–26.

6. Mishra RP, Tisseyre P, Melkonian R, Chaintreuil C, Miché L, Klonowska A, Gonzalez S, Simon M, et al. The geographical patterns of nodulating betaproteobacteria: diversity, host range, and future prospects. Environ Microbiol. 2012;14:234–42.

7. Verma SC, Chowdhury SP, Tripathi AK. Phylogeny based on 16S rDNA and nifH sequences of Raistonia taiwanensis strains isolated from nitrogen-fixing nodules of Mimosa pudica, in India. Can J Microbiol. 2004;50(5):313–22.

8. Taulé C, Zabaleta M, Mareque L, Platero R, Sanjurjo L, Scardà M, Frioni L, Battistoni F, Fabiano E. New Betaproteobacterial Burkholderia strains able to efficiently nodulate Parapiptadenia rigida (Benth.) Brenan. Appl Environ Microbiol. 2012;78(6):1962–700.

9. Andam CP, Mundo 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):4686–90.

10. Gehlot HS, Tak N, Kaushik M, Mitra S, Chen WM, Poweleit N, Panwar D, Poornar 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):179–96.

11. Chen WM, James EK, Chou JH, Sheu SY, Yang SZ, Sprent JI. Beta-rhizobia and other beta-rhizobia. FEMS Microbiol. Ecol. 2014;88:225–37.

12. Verma SC, Chowdhury SP, Tripathi AK. Phylogeny based on 16S rDNA and nifH sequences of Raistonia taiwanensis strains isolated from nitrogen-fixing nodules of Mimosa pudica, in India. Can J Microbiol. 2004;50(5):313–22.

13. Verma SC, Chowdhury SP, Tripathi AK. Phylogeny based on 16S rDNA and nifH sequences of Raistonia taiwanensis strains isolated from nitrogen-fixing nodules of Mimosa pudica, in India. Can J Microbiol. 2004;50(5):313–22.

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;80:417–26.
30. Reeve W, Ardyte J, Tian R, Eshagi 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, Ardyte JK, Tennenhouse K, Woyte K, Kyrpides NC, Ivanova NN. Discovery of novel plant interaction determinants from the genomes of 163 root nodule Bacteria. Sci Rep. 2015;5:16825.

32. De Meyer SE, Fabiano E, Tian R, Van Berkum P, Seshadri R, Reddy TBK, Markowitz V, Ivanova NN, Patti A, Woyte T, et al. High-quality permanent draft genome sequence of the pathogenic fungus nodulating Cupriavidus sp strain UYPF2512. 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, Patti A, Woyte T, et al. High-quality permanent draft genome sequence of the Mimosaaspera nodulating Cupriavidus sp strain AMP6. Stand Genom Sci. 2015;10:80.

34. Estrada-de los Santos P, Martinez-Aguilar L, López-Lara IM, Caballero-Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I, et al. The genomic encyclopedia of the root nodule bacteria: assessing genetic diversity through a systematic biogeographic survey. Stand Genom Sci. 2012;3(3):310–4.

35. Janssens PJ, Van Houdt R, Moors H, Monsieurs P, Morin N, Michaux A, Benomtane MA, Leys N, Vallaeyts 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):e10433.

36. Poehlein A, Kusian B, Friedrich B, Daniel R, Bowien B. Complete genome sequence of the type strain Cupriavidus necator H16. J Bacteriol. 2011;193(18):5017.

37. Silanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, Arlat M, Billault A, Markowitz V, Ivanova N, Patti A, Woyte T, et al. High-quality permanent draft genome sequence of the Mimosaaspera nodulating Cupriavidus sp strain AMP6. Stand Genom Sci. 2015;10:80.

38. Monchy S, Benotmane MA, Leys N, Vallaeys T, Lapidus A, et al. The complete genome sequence of the plant pathogen Ralstonia solanacearum. Nature. 2002;415(6871):497–502.

39. Monchy S, Benomtane MA, Janssens P, Vallaeyts T, Taghavi S, van der Lelie D, Markowitz V, Ivanova N, Patti A, Woyte T, et al. High-quality permanent draft genome sequence of the plant pathogen Ralstonia solanacearum. Nature. 2002;415(6871):497–502.

40. Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity GM, Gilardi J, Glöckner FO, Hirschman L, et al. The genomic encyclopedia of the root nodule bacteria: assessing genetic diversity through a systematic biogeographic survey. Stand Genom Sci. 2012;3(3):310–4.

41. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106:19126–31.

42. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Markowitz V, Ivanova N, Patti A, Woyte T, et al. High-quality permanent draft genome sequence of the Mimosaaspera nodulating Cupriavidus sp strain AMP6. Stand Genom Sci. 2015;10:80.

43. Goris J, Konstantinidis KT, Klappenbach J, Coenye T, Vandamme P, Tiedje JM. DNA-DNA hybridization values and their relationship to whole-genome sequence similarity. Microbiol. 2007;5(8):619–26.

44. Longin C, Mornico D, Roche D, et al. MicroScope—an integrated microbial genome resource for the curation and comparative analysis of genomic and metagenomic sequences. Nucleic Acids Res. 2016;44(D1):D372–47.

45. Grass G, Otto M, Fricke B, Haney CJ, Ren Qing H, Chen KX, Paulsen IT. TransportDB: a comprehensive database for membrane transport protein families. Nucleic Acids Res. 2006;34:D181–6.

46. Saier MH, Tran CV, Barabote RD. TransportDB: the transporter classification database for membrane transport protein families and analyses. Nucleic Acids Res. 2006;34:D181–6.

47. 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(41):D372–9.

48. Saier MH. Genome archeology leading to the characterization and classification of transport proteins. Curr Opin Microbiol. 1999;2(5):555–61.

49. Munkelt D, Grass G, Nies DH. The chromosomally encoded cation diffusion facilitator proteins DmeF and FieF from Wautheria metallidurans CH34 are transporters of broad metal specificity. J Bacteriol. 2004;186(23):8036–43.

50. Grass G, Otto M, Fricke B, Haney CJ, Ren Qing H, Paulsen IT. TransportDB: the transporter classification database (TCDB): recent advances. Nucleic Acids Res. 2016;44(41):D372–47.

51. Saier MH. TransportDB: a comprehensive database for membrane transport protein families and analyses. Nucleic Acids Res. 2006;34:D181–6.

52. 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(41):D372–9.

53. Saier MH. Genome archeology leading to the characterization and classification of transport proteins. Curr Opin Microbiol. 1999;2(5):555–61.

54. Munkelt D, Grass G, Nies DH. The chromosomally encoded cation diffusion facilitator proteins DmeF and FieF from Wautheria metallidurans CH34 are transporters of broad metal specificity. J Bacteriol. 2004;186(23):8036–43.

55. Grass G, Otto M, Fricke B, Haney CJ, Ren Qing H, Paulsen IT. TransportDB: the transporter classification database (TCDB): recent advances. Nucleic Acids Res. 2016;44(41):D372–47.

56. Saier MH. Genome archeology leading to the characterization and classification of transport proteins. Curr Opin Microbiol. 1999;2(5):555–61.
