The Complete Genome Sequence of *Cupriavidus metallidurans* Strain CH34, a Master Survivalist in Harsh and Anthropogenic Environments

Paul J. Janssen¹, Rob Van Houdt¹, Hugo Moors¹, Pieter Monsieurs¹, Nicolas Morin¹, Arlette Michaux¹, Mohammed A. Benotmane¹, Natalie Leys¹, Tatiana Vallaeys², Alla Lapidus³, Sébastien Monchy⁴, Claudine Médigue⁵, Safiy Taghavi⁶, Sean McCorkle⁶, John Dunn⁶, Daniël van der Lelie⁶, Max Mergeay¹

¹Molecular and Cellular Biology, Belgian Nuclear Research Center SCK/CEN, Mol, Belgium, 2Laboratoire Ecosystèmes Lagunaires (ECOLAG), Centre National de la Recherche Scientifique - UMR 5119, Université Montpellier II, Montpellier, France, 3Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, California, United States of America, 4Laboratoire Microorganismes: Génome et Environnement (LMGE), Centre National de la Recherche Scientifique - UMR 6023, Université Blaise Pascal, Aubière, France, 5Laboratoire d’Analyses Bioinformatiques pour la Génomique et le Métabolisme (LABGeM), Centre National de la Recherche Scientifique - UMR8030 and CEA/DSV/IG/Genoscope, Evry, France, 6Biological Department, Brookhaven National Laboratory, Upton, New York, United States of America

**Abstract**

Many bacteria in the environment have adapted to the presence of toxic heavy metals. Over the last 30 years, this heavy metal tolerance was the subject of extensive research. The bacterium *Cupriavidus metallidurans* strain CH34, originally isolated by us in 1976 from a metal processing factory, is considered a major model organism in this field because it withstands milli-molar range concentrations of over 20 different heavy metal ions. This tolerance is mostly achieved by rapid ion efflux but also by metal-complexation and -reduction. We present here the full genome sequence of strain CH34 and the manual annotation of all its genes. The genome of *C. metallidurans* CH34 is composed of two large circular chromosomes CHR1 and CHR2, respectively, 3,928,089 bp and 2,580,084 bp, and two megaplasmids pMOL28 and pMOL30 of, respectively, 171,459 bp and 233,720 bp in size. At least 25 loci for heavy-metal resistance (HMR) are distributed over the four replicons. Approximately 67% of the 6,717 coding sequences (CDSs) present in the CH34 genome could be assigned a putative function, and 9.1% (611 genes) appear to be unique to this strain. One out of five proteins is associated with either transport or transcription while the relay of environmental stimuli is governed by more than 600 signal transduction systems. The CH34 genome is most similar to the genomes of other *Cupriavidus* strains by correspondence between the respective CHR1 replicons but also displays similarity to the genomes of more distantly related species as a result of gene transfer and through the presence of large genomic islands. The presence of at least 57 IS elements and 9 transposons and the ability to take in and express foreign genes indicates a very dynamic and complex genome shaped by evolutionary forces. The genome data show that *C. metallidurans* CH34 is particularly well equipped to live in extreme conditions and anthropogenic environments that are rich in metals.


Editor: Nyiaz Ahmed, University of Hyderabad, India

Received February 1, 2010; Accepted March 29, 2010; Published May 5, 2010

Copyright: © 2010 Janssen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The work was partially supported by internal funds of SCK/CEN, by the European Space Agency through the MESSAGE contracts (PRODEX agreements No. 90037 and 90094), and by Laboratory Directed Research and Development funds at the Brookhaven and Lawrence Berkeley National Laboratories. JD, SMcC, ST and DvdL acknowledge support by the United States Department of Energy, Office of Science, Biological and Environmental Research (BER), project number KP1102010 under contract DE-AC02-98CH10886, and by Laboratory Directed Research and Development funds (LDRD09-005) at the Brookhaven National Laboratory under contract with the U.S. Department of Energy. SM and NM acknowledge the support via a SCK/CEN PhD grant (AWM-programme). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pjanssen@sckcen.be

Introduction

The abundant metal-containing minerals in our environment have shaped the world we live in. While metals such as calcium or sodium are essential micronutrients sustaining life, other (‘heavy’) metals, such as cobalt, copper, nickel, and zinc are vital cofactors for metallo-proteins and enzymes. At excessive concentrations, these metals may be toxic. Nonessential heavy metals, such as mercury, lead, and cadmium are considered toxic at any concentration: hence they are termed ‘toxic metals’.

Perhaps the largest natural environmental sources of heavy metals are volcanic emissions, forest fires, deep-sea vents, and geysers [1]. Because of geobiochemical transformations and natural routes of accumulation, metal-rich sites are fairly common across the planet. Such sites became colonised by long-lasting microbial consortia that overcame metal toxicity by adapting to metal stress and developing the underlying cellular mechanisms to overcome metal toxicity (such as complexation, efflux, or reduction of metal ions). However, consequent upon industrialisation, the concentrations of heavy and toxic metal in soils further increased worldwide over the past few hundred years [2]. A major contribution to this acceleration is the large-scale burning of fossil fuels that releases a wide range of toxic metals into the environment. Anthropogenic sources of metal contamination also include mining and smelting sites, metal-manufacturing plants, painting- and coating-industries, and tanneries. Other human
practices such as applying heavy metals in herbicides and pesticides or using sewage sludge as fertiliser recently raised concern as they have led to the presence of toxic metals in crops, and subsequently, into the foodchain. Typical hot spots of metal contamination are urbanised areas with large industrial activities surrounded by agricultural areas.

The facultative chemolithoautotrophic β-proteobacterium Capnidiavirus (formerly Wautersia, Ralstonia, Alcaligenes) metallidurans was isolated in 1976 as a cadmium-resistant and hydrogen-oxidising pseudomonad from a decantation tank at a metal processing factory in Belgium [3]. This strain, later called CH34, was found to be highly resistant to Zn\(^{2+}\), Cd\(^{2+}\), and Co\(^{2+}\); the extrachromosomal genetic determinants conferring this resistance were transferable to related bacteria. Subsequent studies identified multiple loci for metal resistance, many of which were located on two large plasmids pMOL28 and pMOL30 [4–14].

Many environmental microbes evolved new or better strategies to survive metal-stressed conditions via the exchange of DNA i.e. by gene transfer. Novick and Roth [15] were among the earliest to report on plasmid-mediated heavy metal resistance when they discovered separate genetic loci on the plasmids of Staphylococcus aureus encoding resistance to arsenate, arsenite, antimony, lead, cadmium, zinc, bismuth, and mercury [15]. Since then, the transfer of genes involved in heavy metal resistance (HMR) was documented for many different bacterial species [6,16–18]. For C. metallidurans strain CH34, some important observations related to horizontal gene transfer and the acquisition of HMR can be made. First, upon its discovery, strain CH34 was described as a good acceptor of foreign genes in homologous and heterospecific matings that involved the broad host range IncP plasmid RPl4::miniMu (aka pULB113) containing a very efficient transposon (prophage miniMu3A). This transposon promotes the generation of R-prime plasmids via an “in vivo” cloning process [19,20] later used to map the largest chromosome of C. metallidurans CH34 [21,22]. Second, strain CH34 is a good recipient of large plasmids that are involved in the HMR or hydrogenotrophy of other Capnidiavirus strains [23]. And third, strain CH34 served as recipient and donor in soil microcosms to either assess (recombinant) gene dissemination promoted by broad host range plasmids [24–26] or to directly capture new broad host range plasmids [27–30] and plasmids carrying genes for degradation of xenobiotics [31].

Today, following several taxonomic amendments, strain CH34 is known as the type strain of the species C. metallidurans [32–34]. This strain tolerates high concentrations of heavy metal ions, including but not necessarily limited to Cu\(^{2+}\), Cu\(^{2+}\), Ni\(^{2+}\), Zn\(^{2+}\), Co\(^{2+}\), Cd\(^{2+}\), CrO\(_4^{2-}\), Pb\(^{2+}\), Ag\(^{+}\), Au\(^{+}\), Au\(^{3+}\), HAsO\(_4^{2-}\), AsO\(_4^{2-}\), Hg\(^{2+}\), Cs\(^{+}\), Bi\(^{3+}\), Ti\(^{4+}\), SeO\(_4^{2-}\), SeO\(_2^{2-}\), and Sr\(^{2+}\), with at least 24 gene clusters distributed over four replications, i.e., two main chromosomes and two megaplasmids (8,10,14,35–37 and unpublished results). C. metallidurans CH34 accomplishes metal detoxification by the concerted action of efflux systems including at least eight P-type ATPases and the many metal-resistance mechanisms, their mobile genetic elements, the many metal-resistance mechanisms, their genetic characteristics defined by their bipartite chromosomal structure (denoted as CHR1 and CHR2 throughout this study).

The CH34 genome was also compared with the genomes of more distantly related bacteria belonging to the family Burkholderiales. The species included in this analysis were selected on the basis of high overall synteny between their genomes and the CH34 genome, exemplifying their phylogenetic relationship. One striking feature across the analyzed genomes is the overrepresentation of periplasmic receptors containing the Bug domain, a carbohydrate-binding motif first noted as overrepresented in host-restricted Bordetella species [47]. The prevalence of these receptors in Capnidiavirus and closely related free-living bacteria may point to a common adaptation strategy to polluted environments rich in metal-carboxylates or carboxylated compounds (i.e., sewage, activated sludge, effluents, and industrial wastewaters), and in rhizospheres affected by anthropogenic pollution or in which carboxylates exuded from plant roots have accumulated.

Here we describe the complete sequencing of C. metallidurans strain CH34, and discuss its gene functions derived from manual curation. New insights into its genome structure, its anthology of mobile genetic elements, its regulatory networks, vitamin biosynthesis, and siderophore binding and translocation are revealed. Respiration and autotrophy, the oxidation of hydrogen and sulfur, the synthesis of PHB polymeric granules, its metabolic capacities, capsular-protein production, its signal sensory and transduction systems, and its secretion and transport machinery were also considered.

Results and Discussion

Genome Structure and General Features

The total genome of C. metallidurans CH34, of 6,913,352 bp, comprises four circular replications: Chromosome 1 (CHR1) (3,928,089 bp); chromosome 2 (CHR2) (2,580,084 bp); and two megaplasmids pMOL28 (171,459 bp), and pMOL30 (233,720 bp). Since the two megaplasmids are detailed by Monchy et al. [10] and Mergeay et al. [7] this report mainly covers the sequencing and expert annotation of the two large replications. Nonetheless, in a separate subsection below, we give a brief update on the megaplasmids. Table 1 lists the general features of all four replications. There are four sets of 5S, 16S, and 23S rRNA genes and 12 structural RNA genes that include one RNaseP RNA (rplP, Rmet_R0004), one tmRNA (srrf, Rmet_R0007), one SRP RNA (rscf, Rmet_R0026), and nine riboswitches as predicted by Rfam [http://rfam.sanger.ac.uk/](http://rfam.sanger.ac.uk/) (Rmet_R0001, Rmet_R0012, Rmet_R0043, Rmet_R0047, Rmet_R0067, Rmet_R0068, Rmet_R0083, Rmet_R0084, and Rmet_R0085). Sixty-two transfer RNA (tRNA) genes with specificities for all 20 amino acids were identified, most located on CHR1. Five tRNA genes are duplicated on CHR2 (two genes once, and three genes twice) [Table 1]. Genome analysis using the MaGe system [48] predicts 6,717 coding sequences (CDS). Of those, 4,518 CDS were assigned a function using one of five evidence classes, while 1,274 and 611 CDS remained, respectively, ‘conserved hypothetical’ or ‘hypothetical’. Using the automated COGnit module [49]...
imbedded in the MaGe system, 5,133 CDS were assigned to one or more COG functional classes (COGs are Clusters of Orthologous Groups of proteins generated by comparing the protein sequences of complete genomes - each cluster contains proteins or groups of paralogs from at least three lineages). (Figure 1; Table 1). The G+C content of the replicons all are very similar to each other, with an

![Figure 1.](https://example.com/figure1.png)

**Figure 1. Circular representation of the two main replicons of CH34.** Circles display from the inside outwards: (ring 1) scale in Mb, (ring 2) GC-content using a 2 kb window, (ring 3) GC-skew (G-C/G+C ratio) using a 2 kb window, (rings 4 and 5) predicted CDSs transcribed in a counterclockwise/clockwise direction; genomic islands are indicated in black solid bars and given a numbering; black open triangles represent the IS elements; colored triangles depict transposons: red, Tn6049; blue, Tn6049; and green, Tn6050.

doi:10.1371/journal.pone.0010433.g001

<table>
<thead>
<tr>
<th>Feature</th>
<th>Chromosome 1 (CHR1)</th>
<th>Chromosome 2 (CHR2)</th>
<th>Plasmid 1 (pMOL28)</th>
<th>Plasmid 2 (pMOL30)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size (bp)</td>
<td>3,928,089</td>
<td>2,580,084</td>
<td>171,459</td>
<td>233,720</td>
</tr>
<tr>
<td>G + C ratio (mol %)</td>
<td>63.82</td>
<td>63.60</td>
<td>60.50</td>
<td>60.13</td>
</tr>
<tr>
<td>rRNA</td>
<td>54</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rRNA operons</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IS elements</td>
<td>30</td>
<td>19</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Transposable elements</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Genomic Islands</td>
<td>11</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Total number of CDSs</td>
<td>3,766</td>
<td>2,493</td>
<td>175</td>
<td>283</td>
</tr>
<tr>
<td>Overlapping CDSs</td>
<td>104</td>
<td>92</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>CDSs with assigned function</td>
<td>2,784</td>
<td>1,468</td>
<td>100</td>
<td>166</td>
</tr>
<tr>
<td>Conserved hypothetical</td>
<td>690</td>
<td>478</td>
<td>54</td>
<td>52</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>250</td>
<td>292</td>
<td>15</td>
<td>54</td>
</tr>
<tr>
<td>CDSs assigned to COGs</td>
<td>3,045</td>
<td>1,888</td>
<td>77</td>
<td>123</td>
</tr>
<tr>
<td>Genes involved in transport</td>
<td>433</td>
<td>288</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>Sigma factors</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bug genes</td>
<td>37</td>
<td>87</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0010433.t001

Table 1. General features of the *Cupriavidus metallidurans* CH34 genome.
average of 63.7% for the large chromosomes, and of 60.3% for the two plasmids (Table 1).

The distribution over the replicons of the genome’s functional content also was analyzed using the MaGe system. As expected, the main replicon CHR1 carries most of the essential housekeeping genes, including those needed for DNA replication, DNA repair, cell division, transcription and translation, along with sets of genes required for protein processing, folding, and post-translational modification (Figure 2; Table S1). However, a bias towards particular COG classes was also seen for the smaller replicon (classes I, K, N, Q, and T). These functional classes generally point to adaptation and survival in terms of energy storage and responsive cellular behavior. In addition, the CHR2 replicon contains several genes paralogous to CHR1 genes. Most notably are dbpA (Rmet_5294) encoding a second ATP-dependent RNA helicase, a second excinuclease gene uvrA2 (Rmet_4549), a third copy of an exodeoxyribonuclease gene xthA3 (Rmet_4910), a second glutamate dehydrogenase gene gdhA2 (Rmet_5114), a second elongation factor G encoded by fusA2 (Rmet_5930), a second translation initiation factor IF-1 encoded by usfA2 (Rmet_5176), a second housekeeping sigma-70 factor encoded by rpoD2 (Rmet_4661) and six genes that code for alternative sigma factors rpoD2 (Rmet_3702), sigJ (Rmet_3844), rpoH (Rmet_4001), rpoJ (Rmet_4499), rpoQ (Rmet_4686), and rpoM (Rmet_5400) - see also dedicated subsection on transcription factors. Furthermore, the CHR2 replicon harbours many genes and gene cassettes involved in specialised metabolic- and biosynthetic-activities and adaptive responses, including those involved in carotenoid biosynthesis (carX, Rmet_5644; carB, Rmet_4114), acetone utilisation and tolerance (acsaABC, Rmet_4104-07), polyhydroxyalkanoic acid (PHA) synthesis and conversion (phaZ, phaZ3, phaB3, phaC3), sulfonate/taurine transport and utilisation (asfRAB, tauABC, tauD, tauE), sulfite oxidation (sorAB), squalene/hopene synthesis (sqhc, Rmet_4148), tannin degradation (tanA, Rmet_3041), biofilm formation (pelABCDEFG), and exopolysaccharide synthesis (epsAPB). It carries most of the genes for motility and chemotaxis (mot, flh, fli, flag, spa, tad, pil, fim, and che loci), peroxide

![Figure 2](image-url)
removal and oxidative stress (\( \gamma \)-phA, abpC, abpD, katG, pcaCD, ssoR, gltX); DNA adenine methylation (dam); DNA repair (ads, alkC, alkB1, alkB2, shoCD, ogt2, ogt3); adaptive mutagenesis (imuA-imuB-dnaE2, Rmet_4128-29-30), and eight copies of the \( \omega p \)-\( dp \) gene (two of which are located on transposons) encoding the universal stress protein [most genes and loci without Rmet numbers are also discussed in separate sections]. The CHR2 replicon also carries an intact master two-component phosphorelay system (\( bgbU \)) that, in Bordetella species, controls hundreds of genes involved in an extensive variety of responses and physiological functions, including virulence, biofilm formation, chemotaxis, and motility [50].

Both chromosomes display clear GC skew transitions that likely correspond with their replication origin and terminus [51] (Figure 1). For CHR1 the GC skew transition region (at point 3.84 on the map, Figure 1A) is about 90 kb away from the typical dnaA-dnaN-gyrB consensus) occur even further upstream of a GC skew transition region located around coordinate 3,837,500 is 4 mismatches to this consensus (Figure S1A). Interestingly, the GC skew transition region (reclassified as C. pinatubonensis) with its closest relative C. pauculus [33,52]. The complete genome sequences of three other Capriavirus species have been published (Table 2): The pollutant degrading R. eutrophus strain JMP134 (reclassified as C. pinatubonensis by Sato et al. [52]): the lithiotrophic “Knallgass” bacterium C. neator strain H16 (with the proposed name C. eutrophus H16, Vandamme & Coenye, pers. com.); and the beta-rhizobial nitrogen-fixing plant symbiont C. eutrophus.

Phylogeny and genome comparative analyses
Phylogenetic analyses based on 16S rRNA sequences placed Capriavirus metallosalivorans CH34 (formerly Rajstonia) at one point transferred to Wautersia [34] firmly into the genus Capriavirus, with its closest relative C. pauculus [33,52]. The complete genome sequences of three other Capriavirus species have been published (Table 2): The pollutant degrading R. eutrophus strain JMP134 (reclassified as C. pinatubonensis by Sato et al. [52]): the lithiotrophic “Knallgass” bacterium C. neator strain H16 (with the proposed name C. eutrophus H16, Vandamme & Coenye, pers. com.); and the beta-rhizobial nitrogen-fixing plant symbiont C. eutrophus.

Table 2. Genome organisation of representative species (discussed in this contribution) within the order of Burkholderiales.

<table>
<thead>
<tr>
<th>Organism**</th>
<th>Main property</th>
<th>Family</th>
<th>Principle habitat</th>
<th>CHR1</th>
<th>CHR2</th>
<th>Replicon 3 C°</th>
<th>Replicon 4</th>
<th>Year/Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. metalloidurans CH34</td>
<td>heavy metal resistance</td>
<td>Burkholderiaceae</td>
<td>heavy metal-rich sediments and soils</td>
<td>3.93</td>
<td>2.58</td>
<td>0.17 (pMOL28)</td>
<td>0.23 (pMOL30)</td>
<td>This report</td>
</tr>
<tr>
<td>C. pinatubonensis JMP134</td>
<td>degrader of xenobiotics</td>
<td>Burkholderiaceae</td>
<td>soil, aquatic, volcanic mudflow deposits</td>
<td>3.81</td>
<td>2.74</td>
<td>0.63 (MPL)</td>
<td>0.09 (pP4)</td>
<td>unpublished</td>
</tr>
<tr>
<td>C. eutrophus H16</td>
<td>hydrogen oxidation, PHA¹ metabolism</td>
<td>Burkholderiaceae</td>
<td>soil, freshwater</td>
<td>4.05</td>
<td>2.91</td>
<td>0.45 (phG1)</td>
<td>-</td>
<td>2006/45</td>
</tr>
<tr>
<td>C. taiwanensis</td>
<td>N₂-fixing legume symbiont</td>
<td>Burkholderiaceae</td>
<td>plant root nodules</td>
<td>3.42</td>
<td>2.50</td>
<td>0.56 (prALTA)</td>
<td>-</td>
<td>2008/46</td>
</tr>
<tr>
<td>R. solanacearum GM1000</td>
<td>plant pathogen with wide host range</td>
<td>Burkholderiaceae</td>
<td>soil, rhizosphere</td>
<td>3.72</td>
<td>2.09</td>
<td>-</td>
<td>-</td>
<td>2002/223</td>
</tr>
<tr>
<td>B. xenovorans LB400</td>
<td>PCB² degrader</td>
<td>Burkholderiaceae</td>
<td>rhizosphere, polluted soils, landfills</td>
<td>4.90</td>
<td>3.36</td>
<td>1.47 (MPL)</td>
<td>-</td>
<td>2006/224</td>
</tr>
<tr>
<td>D. acidovorans Sph-1</td>
<td>organotrophy, LAS³ degrader</td>
<td>Comamonadaceae</td>
<td>soil, aquatic, sewage</td>
<td>6.77</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>unpublished</td>
</tr>
<tr>
<td>B. petrii Se-1111R</td>
<td>degrader of aromatic compounds</td>
<td>Alcaligenaceae</td>
<td>river sediments, TCB⁴ contaminated soils</td>
<td>5.29</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2008/148</td>
</tr>
<tr>
<td>Jahnthinobacterium sp.</td>
<td>heavy metal resistance, small sized</td>
<td>Oxalobacteraceae</td>
<td>aquatic, 0.22 ( \mu )m filtered water</td>
<td>4.11</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2007/225</td>
</tr>
<tr>
<td>H. arsenicayadyans</td>
<td>arsenic oxidation and scavenging</td>
<td>Oxalobacteraceae</td>
<td>activated sludge, industrial waste water</td>
<td>3.42</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2007/226</td>
</tr>
<tr>
<td>P. necessarius STIR1</td>
<td>obligate intracellular endosymbiont</td>
<td>Burkholderiaceae</td>
<td>freshwater ciliates</td>
<td>1.56</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>unpublished</td>
</tr>
</tbody>
</table>

*species that are rarely recovered from clinical samples are indicated with an asterisk (*);
1 Only the propensity for which the species and/or particular strain is best known is given;
2 MPL, megaplasmid.
3 PHA, poly(3-hydroxybutyric acid) (cfr. ‘bioplastic’);
4 PCB, polychlorinated biphenyl (cfr. ‘herbicide’);
5 LAS, linear alkylbenzene sulfonate (cfr. ‘laundry surfactant’);
6 TCB, trichlorobenzene.

http://doi:10.1371/journal.pone.0010433.t002
taivanensis. Throughout this paper we will often refer to these genomes as Cmet (for C. metallidurans), Cpin (for C. pinatubonensis JMP134), Ceut (for C. eutrophus H16) and Ctau (for C. taiwanensis).

The genome organisation of these four species is very similar. All four carry two large replicons, that we further nominated as CHR1 and CHR2, with each replicon being of comparable size to their counterparts. The main difference in terms of genome organisation stems from the presence of plasmids (Table 2). Whereas Cmet harbours two megaplasmids of 171- and 233-kb, Ctau and Ceut carry a single megaplasmid, respectively, of 452 kb and 557 kb. The Cpin genome includes a small plasmid of 88 kb (pJP4) along with a megaplasmid of 635 kb.

We determined orthology between all those replicons by analysing the reciprocal best BlastP hits, and then performed cluster analysis using calculated distance scores (see Methods). Our analysis also encompassed representative genomes that, according to the MaGe system, display a well conserved gene order (synteny) to C. metallidurans (measuring synteny to the CHR1 replicon). These were Ralstonia solanacearum GMI1000 (two large replicons), Burkholderia xenovorans LB400 (two large replicons and a plasmid), and Bordetella petrii DSM12804, Delftia acidovorans SPH-1, Janthinobacterium sp. Marseille, and Hermanninomonas arsenicoxydans (all having a single replicon). We further included Polyvalensbacillus necessarius STIR1 (a single replicon with surprisingly good synteny to Cmet CHR1) to represent, next to the genera Ralstonia, Cupriavidus, and Burkholderia, the fourth genus in the family of Burkholderiaceae [53]. Table 2 gives general information on all those species. Lastly, we employed the genome of Escherichia coli K-12 in this analysis as a reference and outlier.

Considering only the four Cupriavidus species, some interesting observations were made. First, the core number of genes common to all four species clearly is larger for the first replicon with 2,209 genes for CHR1 compared to 677 genes for CHR2 (Figure 3): calculated against the average number of genes present on each replicon, this amounts to 62% and 28%, respectively. This observation, together with a strong bias for essential genes, as we noted for Cmet, and also seen for Ceut and Ctau [45,46], indicates that CHR1 constitutes the ancestral replicon. The smaller replicon CHR2 probably was acquired as a plasmid during the evolution of Cupriavidus and gradually evolved into a second chromosome through genetic transfer from either CHR1 or external sources [54]. This idea may explain why CHR2 carries features typical for plasmid-type partitioning, as well as its unique core cellular functions. Here, the delegation of essential genes from the primary replicon to a second replicon would ensure the maintenance of all genes on that replicon, a strategy seemingly common in multipartite prokaryotic genomes [55]. Second, C. taiwanensis (Ctau) and C. eutrophus H16 (Ceut) have more genes in common (Figure 3) i.e., are more closely related than is the case for the other Cupriavidus species, confirming the findings of Amadou et al. [46]. And third, both large replicons of Cmet harbour more species-specific genes than do the large replicons of the other Cupriavidus genomes (Figure 3).

Additional patterns emerge from the similarity matrix (Figure S3) holding protein-based orthology data for these genomes. As expected, the CHR1 replicons of the four Cupriavidus species group tightly together. This is also the case for the Cupriavidus CHR2 replicons. For both replicons, the highest orthology is between Ctau and Ceut, and the ranking of the species according to orthology in this cluster remains the same for both replicons, with Cmet displaying the least orthology. There is a clear division among all analysed replicons into two large clusters A and B (Figure S3): Cluster A contains the CHR1 group and the main replicons of B. xenovorans and R. solanacearum, along with the chromosomes of D. acidovorans, B. petrii, H. arsenicoxydans, J. variabilis, and P. necessarius (with the E. coli chromosome inside this cluster but clearly an outlier). Cluster B consists of the CHR2 group together with the C. pinatubonensis, B. xenovorans and R. solanacearum megaplasmids and the B. xenovorans second chromosome. The grouping of the remaining Cupriavidus replicons (all plasmids) from Blast analyses is less clear, although pMOL28 (Cmet), pRALTA (Ctau), and pHG1 (Ceut) appear to be somewhat more related (cluster C) to each other than is the case for pMOL30 (Cmet) and pJP4 (Cpin). Probably, this reflects the fact that the

---

**Figure 3. Shared and unique proteins of Cupriavidus chromosomes.** 4-way comparison Venn diagram illustrating the intersection and differences between chromosomal replicons of completed Cupriavidus genomes. Intersections show the number of shared proteins between two or more organisms based on reciprocal best BLAST hits (see Methods). Numbers in parentheses depict the missing overlap sectors due to circular drawing and represent proteins shared between opposite genomes but absent in the other two genomes. Abbreviations: Cmet, C. metallidurans CH34; Cpin, C. pinatubonensis JMP134; Ceut, C. eutrophus; Ctau, C. taiwanensis.

doi:10.1371/journal.pone.0010433.g003
three plasmids in cluster C have the same backbone, i.e., contain the essential genes for plasmid replication and maintenance ([7] and see further). Interestingly, the genome of *P. necessarius* STIR1, an obligate intracellular symbiont of the freshwater ciliate *Euplotes aediculatus*, not only clusters well with the CHR1 group but does so with nearly two thirds (63% on average) of its proteins having an ortholog in any of the other members of this group (in contrast to only 40% of its proteins having an ortholog in the outlier *E. coli*). This high orthology may denote genomic reduction in *P. necessarius* STIR1, in which mainly essential genes were preserved, and is reflected in a significant degree of gene synteny between the *P. necessarius* STIR1 genome and CHR1 of *C. metallidurans* but much less with CHR2 (Figure S4A). This high synteny with CHR1 was fully supported by MUMmer3 analysis on protein level using the PROmer module (Figure S4B).

Full length alignment of the main replicons of the four *Cupriavidus* genomes and the closely related *R. solanacearum* genome clearly suggests a common ancestral architecture for CHR1 (Figure 4); overall homology between their respective CHR2 replicons is much lower (Figure S2) probably owing to different evolutionary histories. Such multiple alignments also reveal a large number of DNA rearrangements in the CH34 genome compared to the other analysed genomes (and vice versa), including some large inversions, and multiple sequence duplications causing significant paralogy in both CH34 replicons (discussed further). Pairwise synteny analysis using the LinePlot tool embedded in the MaGe interface fully supported these results (data not shown) while limited synteny was observed between CHR1 of *C. metallidurans* and the single replicons of the other organisms listed in Table 2. The intragenomic synteny between the CHR1 and CHR2 replicons of *C. metallidurans* was not very apparent. However, choosing a small synton size in the LinePlot analysis tool (with S=5 e.g. considering five or less genes for each synteny group), highlighted a myriad of sequence matches (data not shown). This type of local synteny may be partly explained by evolutionary DNA exchange between the two replicons or by parallel insertion of homologous sequences of incoming DNA.

**Paralogy and protein families**

Gene duplications (paralogs) are seen as evolutionary events contributing to the innovative drive of organisms in order to adapt to changing environments. We measured paralogy within the *C. metallidurans* genome at the protein level using the GeneRAGE algorithm that facilitates clustering of large protein sequence sets into families according to their sequence similarity [56]. To allow GeneRage to analyse genome paralogy, we undertook a basic similarity search by an all-to-all BLAST analysis of all *C. metallidurans* CH34 proteins against each other. Smith-Waterman dynamic programming (see Methods) automatically removed ambiguities in the resulting similarity matrix, assuring the method's high accuracy. Multi-domain proteins detected in the genome were disregarded. Half of the remaining proteins (3,311 or 52%) had one or more paralog, while the other half (3,008 or 48%) remained singletons having no match with any of the other proteins in the genome. Remarkably, 1,358 proteins (21%) grouped together into 43 clusters of 10 or more members, while 4 clusters held more than 110 members. These large paralogous families represented proteins involved in two-component signal...
transcription (n = 138) (predominantly containing GGDEF, EAL, PAS, and HisKA domains), extra-cytoplasmic solute receptors of the tripartite tricarboxylate transporter (TTT) family (n = 123), LysR-type transcriptional regulators (n = 121), and ABC-transporter related proteins (n = 110). Other large families (size > 30) comprised proteins involved in membrane structure and transport (membrane fusion proteins, porins, and major facilitator superfamily transporters), or enzymes with general metabolic functions (short-chain dehydrogenases, enoyl-CoA reductases, L-carnitine dehydrogenases, acyl-CoA dehydrogenases, and acyl-CoA synthetases). These findings are in line with those of Gevers et al. (37) who studied paralogy in 106 bacterial genomes, reporting a significant paralog enrichment in functional classes involved in cellular defense mechanisms, metabolism, and transcriptional responses. These authors broadly correlated the number of paralogs in the genomes with the organism’s lifestyle, with paralogue sizes of 7% for intracellular organisms, such as Rickettsia conori, and up to 41% for free-living highly versatile soil bacteria, such as Streptomyces coelicolor. Accordingly, it is not surprising that the survivalist C. metallidurans possesses a large, functionally biased paralogue. In exploring the smaller clusters holding one- to five-paralogs (656 clusters with a total of 1,724 proteins), it is difficult to discern a functional bias or propose a plausible explanation for paralogy. Many of them contain conserved hypothetical proteins, or proteins similar in function but differing in substrate or cellular roles. Others contain cellular- or transcriptional-regulators specifically directed to heavy-metal resistance (see also Nies et al. (38)) or encompass functions related to mobile genetic elements. A quick survey points to several interesting findings. For instance, apparently, there are no less than six copies of 2-nitropropane dioxygenase that catalyzes the enzymatic reaction 2-nitropropane + oxygen ↔ acetone + nitrite. This could be an important finding in view of the known acetone tolerance of C. metallidurans CH34 (see further in this discussion), or in terms of converting environmental nitroalkanes for bioremediation purposes. Likewise, there are three copies of butyl-CoA:Acetate-CoA transferase, an enzyme that may be involved in acetate detoxification or instrumental in delivering butyrate to produce polyhydroxybutyrates (PHBs), biopolymers that are stored as an energy reserve in the cell, and which have industrial applications as bioplastics (see further, section on polyester synthesis PHA/PHB). Undoubtedly, it will be exciting to follow these leads but a systematic functional breakdown of these clusters is beyond the scope of this report.

Mobile genetic elements

A striking feature of the C. metallidurans CH34 genome is the presence of at least 16 laterally acquired genomic islands and the abundance of transposase- and recombinase-related gene sequences across the four replications (Table 1). Using Van Houdt et al.’s criteria (39), 11 GIs were identified on CHR1 but none on CHR2. Defining GIs on CHR2 is not straightforward since its structure is very different from that of CHR1 and it appears to consist of a complicated patchwork of gene clusters and subclusters generated by multiple integration events with the subsequent retention (and occasional loss) of acquired genes. Further, the CHR2 replication contains fewer tRNA genes (i.e. only eight compared to 54 in CHR1) and only very few xerC/D-like or int-like recombinase genes. Previously identified GIs of the plasmids pMOL28 and pMOL30, respectively referred to as CMGI-28a, -28b, -28c, and CMGI-30a and -30b, respectively, all carry heavy-metal response genes; discussed in detail elsewhere (7,10) and in the next subsection on megaplasmids.

The largest island CMGI-1 is 109 kb and is almost identical (>99.95% DNA sequence similarity) to the PAGI-2 island in Pseudomonas aeruginosa clone C, and to the highly conserved PAGI-2C-type islands from many other isolates of this species. It also corresponds to several C. metallidurans and Cupriavidus campestris strains isolated from environments polluted with sewage and heavy metals (60,61). The observation that clinical strains of nosocomial infections and environmental strains isolated from industrial sites have such a large segment of genetic material in common is worrying since it suggests a recent breach of the natural and artificial barriers of bioremediation and clinical environments. Clearly, P. aeruginosa has expanded its genomic repertoire to survive in a wide range of different environments, including the human host, in which lateral gene-acquisition might have been enhanced by co-inhabitation and by genetic relatedness (62). In this respect there is a certain resemblance between the evolutionary forces shaping the genomes of P. aeruginosa and C. metallidurans, as the latter also acquired and retained many new genetic elements, and consequently new traits, enabling it to survive in many different, often harsh, environments.

Moreover, two C. metallidurans strains identified by 16S rDNA sequence analysis were recently isolated from the respiratory tract of cystic fibrosis patients (63) warranting concern about the disease-causing potential of C. metallidurans.

Three of the C. metallidurans GIs (CMGI-2, CMGI-3, and CMGI-4) belong to the Tr371 family, first described in strain C. acetovorans A5 (formerly R. eutropha A5) (64–66). This family is defined by a quadripartite structure consisting of an integrase module, a region with plasmid/phage/GI maintenance genes, the accessory genes, and the conjugative genes (59) and was recently revisited to describe genomic islands of other environmental bacteria (67).

The 101 kb large CMGI-2 contains at least 25 genes involved in degrading tolulene corresponding to CH34’s ability to grow with tolulene, benzene, or xylene as their sole carbon source (see section on organoautotrophy and degradation of aromatics). Apparently, additional accessory genes are inserted near this catabolic cluster, including genes involved in chemolithotrophy (e.g. hyp and hox genes – see section on hydrogenotrophy).

The 97 kb CMGI-3 island also shelters accessory genes in chemolithotrophy organized as two gene cassettes flanked by IS1071: one contains the genes for the fixation of CO2 (hyp genes; described in the section on autotrophic metabolism) and the other encompassing additional genes for hydrogenotrophy (hyp and hox genes). All the genes allowing C. metallidurans to grow aerobically at the expense of CO2 and H2 are confined to CMGI-2 and CMGI-3 in CHR1: this differs in the closely related H2-oxidising chemolithotrophic bacterium C. eutrophus H16 wherein the hox and hyp genes are situated on a 452 kb megaplasmid, pHG1 (45,68). It is unclear whether strains H16 and CH34 inherited these homologous sets of hyp and hox genes from a common ancestor through vertical descent, or whether they were passed on by lateral gene transfer. In contrast, the cbb loci (for CO2 fixation) in both strains clearly have different origins since the CH34 cbb genes are phylogenetically much more closely related to the cbb homologues of photosynthetically- and nitrifying-organisms (45).

The 56 kb CMGI-4 island has a mosaic structure and lacks the conjugative module. Instead, its extremity carries genes involved in the biosynthesis of methionine, and the degradation of phosphonates; the transposon Tn6048 also occurs there. The DNA sequence of the CMGI-4 island is highly similar to a region in the D. desulfuricans SPH-1 genome (69–71) except that the latter does not contain Tn6048 but carries a complete Tn4371-like structure flanked by Tn6051 (39).

The contribution of the remaining GIs (between 11- and 24-kb in size) towards the particular lifestyle of C. metallidurans is unclear.
Seemingly, CMGI-5 is composed of plasmid remnants, while CMGI-7 might be involved in resistance to arsenic. The CMG-11 island harbours a putative fimbrial operon of the chaperone-usher-dependent assembly pathway. The islands CMGI-6 and CMGI-8 are cryptic islands holding copies of IS1007 and Tn6049, while CMGI-9 and CMGI-10 consist almost entirely out of hypothetical genes.

The *C. metallidurans* CH34 genome accommodates 19 transposons of five different types [59] (Table 3). Two 'mercury transposons' Tn4378 and Tn4380 (both of the Tn501 family) are located, respectively, on pMOL28 and pMOL30, and were previously reported [10]. We identified three novel types of transposons' Tn6048 (10.4 kbp), containing 8 genes of unknown function but highly induced by zinc and lead, Tn6049 (3.5 kbp), often associated with genomic islands, and Tn6050 (6.8 kbp) found twice in CHR2 in opposite orientations and separated by 143 kb. Remarkably, all 12 copies of Tn6049 carry the same ATPase-encoding gene, putatively named *gfpM* because its product shows strong similarities to putative general secretion pathway ATPases. We also note that although the core region (*tnpA-tnpB*) of Tn6050 is very similar to that of the mercury transposons Tn4378 and Tn4380, its accessory gene cassette is uninvolved in mercury resistance and encodes proteins that are only rarely associated with transposons, viz., a sulfate permease, a universal stress protein (*UspA*), and a DksA-like DnaK suppressor protein. We reason that there is a functional linkage between these genes, as this module, in full or part, also occurs on other Tn5-related transposons i.e., alone, as in Tn1013 from the IncP-1 plasmid pBS228 [72], or in combination with antibiotic resistance determinants, as in Tn6001, Tn1403, and Tn1404* transposons detected in clinical- and environmental-strains of *Pseudomonas* [73,74]. Closer examination of the chromosomal regions flanking the Tn6050 transposons in strain CH34 suggests the possibility of chromosomal inversion through homologous recombination, thereby effecting several loci involved in metal resistance [75,76].

Significantly, the *C. metallidurans* CH34 genome bears no fewer than 57 insertion elements evenly distributed over the four replicons (Table 4). The most abundant are ISRme^3^ (an IS3 family member), with 10 copies, and IS1008 (an IS39 family member), with nine copies. Conceivably, many of those 57 IS elements have played an important role in modulating the architecture of the CH34 genome and in the continuous acquisition of new traits. For instance, the incorporation of *hyl* and *hex* genes into CMGI-3 likely results from active IS1071 elements, while in contrast, the ability of strain CH34 to grow autotrophically may be lost owing to IS1071-mediated excision. Similarly, the two ISRme^3^ copies on CMGI-2 are located amidst multiple transposase genes and fragments, and probably were the driving force toward integrating the *hyl/hex* cluster on CMGI-2. In addition, both copies of ISRme^3^ flank the genomic island CMGI-11 while two copies of ISRme^3^ are associated with the genomic islands in pMOL30, further strengthening our belief that IS elements principally shaped the CH34 genome.

We also detected in the CH34 genome a series of putative mobile elements that previously were unnoticed (detailed by Van Houdt et al. [59]). The first group of such elements contain tyrosine-based site-specific recombinase genes consecutively placed in series of three. We observed such trios in a wide variety of *α*- and *β*-proteobacteria and termed them *recombinase in trio or rit* elements. One trio (*ritA1-B1-C1*) (Rmet_1239 to Rmet_1241) belongs to the integrase module of CMGI-2 and apparently inactivates a putative DNA repair gene, *radC*. This trio is highly similar to the *B. petrii* rit element (Bpet1498 to Bpet1500) of genomic island GI3, having corresponding aa (amino acid) sequence identities of 77% or higher. Intriguingly, the *B. petrii* rit element inactivates a *radC* gene (Bpet1497), and all genes in both *rit* elements display ATGA overlaps strongly suggesting translational coupling. A second trio (*ritA1-B1-C1*) (Rmet_1271 to Rmet_1273) also lies on CMGI-2 and is part of a complex accessory module; it inactivates a *tpaA* gene of the IS66 family. The *C. metallidurans* *ritA1-B1-C1* element exhibits high similarity to the *rit* elements on plasmid pHG1 of *C. eutrophus*, showing very high aa sequence similarities (>92% identity) for *ritA1* and *ritB1* to either PHG129 and PHG141, or either PHG130 and PHG140, respectively. Further, there are high aa sequence similarities (40–54% identity) between *ritA1-B1-C1* and the trio PHG134/135/136. Immediately adjacent to this maz of pHG1 recombinase genes we discerned an IS66 type *tpaA* gene (PHG142) but we are uncertain whether it is interrupted or not. We also noticed four-gene elements, comprising three recombinases close to each other and invariably accompanied by a fourth gene encoding a transcriptional regulator of the TetR family [59]. The last group of mobile

---

**Table 3. Distribution of transposons in *C. metallidurans* CH34.**

<table>
<thead>
<tr>
<th>Element</th>
<th>Total nr.</th>
<th>CHR1</th>
<th>CHR2</th>
<th>pMOL28</th>
<th>pMOL30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tn4378</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tn4380</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Tn6048</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tn6049</td>
<td>12</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tn6050</td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

---

**Table 4. Distribution of IS elements in *C. metallidurans* CH34.**

<table>
<thead>
<tr>
<th>Element</th>
<th>Total nr.CHR1</th>
<th>CHR2</th>
<th>pMOL28</th>
<th>pMOL30</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS1071</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS1086</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IS1087</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>IS1088</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>IS1090</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme1</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>ISRme2</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>ISRme3</td>
<td>17</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>ISRme4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme5</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>ISRme6</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ISRme7</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme8</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ISRme9</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme10</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme11</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme12</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme13</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme14</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme15</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme16</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme17</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme18</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme19</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ISRme20</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

doi:10.1371/journal.pone.0010433.t003

doig:10.1371/journal.pone.0010433.t004
elements we list here is the bimAB module, named after the bimA recombaine and bimB putative repressor genes first noted on the pMOL28 and pMOL30 plasmids as genepairs Rmet_6192/93 (linked to a chr locus) and Rmet_5807/88 (linked to a cze locus), respectively [7]. The latter module (21 genes) is fully conserved in Ralstonia pickettii 12J, indicating a horizontal transfer event. The bimAB element has also been associated to HMR loci in other bacteria [59].

An interesting feature of C. metallidurans strain CH34, and characteristic for related metallotolerant C. metallidurans strains, is the increased activity of transposable elements when cells are grown at elevated temperature i.e. 37°C instead of 30°C. This phenomenon is referred to as temperature induced mutagenesis in which survivor cells appear to have undergone DNA arrangement phenomena. The increased activity of transposable elements when cells are grown at elevated temperature i.e. 37°C has also been associated with HMR loci in other bacteria [59].

Both megaplasmids pMOL28 and pMOL30 carry key genes for resistance to heavy metals [9]. Plasmid pMOL28 encodes resistance to nickel and cobalt [80,81], chromate [5], mercury [4,10], and thallium [82,83]. Plasmid pMOL30 specifies resistance to zinc, cadmium, and cobalt [13,84], copper [75], silver [8], and lead [37,85]. The corresponding gene clusters on pMOL28 and pMOL30, and the replication region of pMOL28 were identified subsequently (reviewed by [7,8,10]). Monchy et al. [10] identified detailed genetic maps of both megaplasmids from their full DNA sequence. This distinguished the genes involved in plasmid maintenance, replication, and partition, together constituting the ‘backbone’ of the plasmid, next to specialized ‘accessory’ genes that group together on genomic islands i.e., CMGI-28a, -28b, -28c, and CMGI-30a and -30b. In pMOL28 and pMOL30, these GIs contain all the genes involved in metal resistance [7,59]. Notably, the pMOL28 replicon belongs to a family of megaplasmids that share a similar backbone and include the megaplasmids pHG1 (452 kb) of C. eutrophus H16 and pRALTA (557 kb) of C. taiwanensis. Owing to the distinctly specialized functions on pMOL28 (heavy metal resistance), pHG1 (hydrogenotrophy and chemolithotrophy), and pRALTA (nitrogen fixation and legume symbiosis), these megaplasmids might well be regarded as the decisive factor in fixing their hosts to their ecological niche. A direct relationship of the megaplasmid pMOL30 to other megaplasmids is less certain. Nonetheless, through synteny analysis and advanced annotation using the MaGe platform we noted that the backbone of pMOL30 shares, over a region of approximately 60 kb, many characteristics with the pBVIE01 megaplasmid (396 kb) of B. vietnamensis G4 [7].

Despite detailed annotation of megaplasmids pMOL28 and pMOL30 [10], continuous manual curation of the CH34 genome via updated databases, and earlier functional information from proteomic- and transcriptomic-approaches [5,7,8,38,86] many enigmatic genes remain in both plasmids (Table 1).

Recent transcriptomic analyses revealed that five such genes, located on CMGI-28a, are induced by submillimolar concentrations of chromate (no data given). All five genes are situated immediately downstream of the chromate-regulated operon chrLA,5BCEF (Rmet_6199 through Rmet_6204): we called them chrO (Rmet_6198), chrN (Rmet_6197), chrP (Rmet_6196, encoding an MFS family permease with ten transmembrane a-helical segments), chrF (Rmet_6195, encoding a putative membrane protein), and chrZ (Rmet_6194) (Table 5). Synteny analysis using MaGe shows a fair to good resemblance between the pMOL28 chr locus of C. metallidurans and the chr loci of Methyllobacterium sp. 4–46, B. vietnamensis G4, and B. pseudomallei. However, the chr locus in Methyllobacterium sp. 4–46 lacks chrE, chrN, and chrP orthologues, the four chr loci of B. vietnamensis G4 all lack chrP and chrZ orthologues, and the majority of chr loci across 21 B. pseudomallei strains lack chrN, chrE, chrO, and chrN orthologues. This puzzling variation in chr locus structure may point to differences among these bacteria in terms of selectivity to chromate or another metal-oxoanion.

We also identified hypothetical genes on CMGI-29b that encode membrane proteins rich in YD (tyrosyl-aspartyl) motifs and named them tbr (Rmet_6263), tbs (Rmet_6261), and tbrB (Rmet_6258) because such YD-repeat proteins apparently are linked to RHS (Recombination Hot Spot) regions [87,88]. Noteworthily, these genes are centered around the tbrA and tbrB genes (respectively Rmet_6260 and Rmet_6259) of the ISRic9 mobile element.

The pMOL28 backbone region covered by Rmet_6288 through Rmet_6297 provisionally was named the pil locus because of some resemblance in its sequence and gene order with other pil loci [7,10]. However, although this region and the flanking ones display good synteny with the corresponding backbone regions on the C. taiwanensis pRALTA and C. eutrophus pHG1 megaplasmids, the individual sequence similarity of genes Rmet_6288 to Rmet_6294 to known pil genes is very low. Either these genes are not involved in conjugation (but rather in secretion), or pMOL28 has evolved a sui generis conjugative system. For now, we will retain the pil epitaph. We note that the Rmet_6291 gene, formerly annotated as pilQ [10], now is annotated as pilB because its product shows no aa sequence similarity to the PilQ usher protein encoded in the pilMNOPQ locus on CHR1 (Rmet_3268 to Rmet_3272); rather, it probably acts as an ATPase.

A fair number of plasmid-born genes are seemingly involved in either gene exchange or membrane processing. For instance, the entire region between the pbr and cze loci on CMGI-30a is speckled by very small hypothetical genes, transposases (tnp), site-specific recombinases (inf), or remnants thereof, some of which are upregulated by heavy metals [10]. In addition, both genomic islands on pMOL30 contain a gtrMAB locus encoding proteins that are required for translocating oligosaccharide precursors across the cytoplasmic membrane and the subsequent glucosylation of lipopolysaccharide (LPS). Several genes of both gtr loci are upregulated by heavy metals [10] while the gtr loci are centrally placed between the pbr and mer clusters on CMGI-30a (as genes Rmet_5966/67/68) and between the cop and sin clusters on CMGI-30b (as genes Rmet_6128/29/30). A possible reason for recruiting gtr loci and embedding them in metal resistance clusters may be that modified (i.e., shortened) LPS chains could positively affect heavy-metal extrusion. Alternatively, GtrMAB proteins may act on other components outside of the inner membrane thereby limiting heavy-metal damage inflicted on membranes.

### Heavy metal resistance (HMR) genes in CH34

Metal detoxification in C. metallidurans CH34 mainly occurs via ion efflux although other systems are present (Table 5; Table S2). A variety of export systems that enlist members of three protein families assure ion efflux: The RND (for Resistance, Nodulation, cell Division) family of permeases [74], the CDF (for Cation Diffusion Factor) family of heavy metal transporters [89,90], and...
### Table 5. Metal detoxification in *C. metallidurans* CH34 occurs via a variety of export systems encoded by 24 HMR gene clusters on 4 replicons.

<table>
<thead>
<tr>
<th>Type</th>
<th>Replicon</th>
<th>Gene cluster</th>
<th>Rmet. number</th>
<th>Metal(s)</th>
<th>Gl/transposon</th>
<th>Mechanism</th>
<th>Fc</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>partially plasmid-redundant</td>
<td>CHR1</td>
<td>cdIX pbr12 cda pbrC2 pbrH2 (†)</td>
<td>2299, 2302–2304, 3456</td>
<td>Pb2+</td>
<td>CMGI-1 (P)</td>
<td>ExP, pCDF</td>
<td>+</td>
<td>[59]</td>
</tr>
<tr>
<td></td>
<td>pMOL30</td>
<td>pbrUB(UatR) pbrABCD (‡)</td>
<td>6180, 5944–5949</td>
<td></td>
<td>CMGI-30a</td>
<td>ExP, seq</td>
<td>+</td>
<td>[85]</td>
</tr>
<tr>
<td></td>
<td>CHR1</td>
<td>merRTPA A+</td>
<td>2312–2315</td>
<td>Hg2+</td>
<td>CMGI-1 (P)</td>
<td>HgRed</td>
<td>?</td>
<td>[59]</td>
</tr>
<tr>
<td></td>
<td>pMOL30</td>
<td>merRTPADE urf-2</td>
<td>6171–6177</td>
<td></td>
<td>CMGI-30a/Tn4380</td>
<td>HgRed</td>
<td>+</td>
<td>[4]</td>
</tr>
<tr>
<td></td>
<td>pMOL30</td>
<td>merRT1P</td>
<td>5990–5992</td>
<td></td>
<td>CMGI-30a</td>
<td>HgTra</td>
<td>?</td>
<td>[10]</td>
</tr>
<tr>
<td></td>
<td>CHR2</td>
<td>chrBAC</td>
<td>3866–3864</td>
<td>CrO42-</td>
<td>-</td>
<td>ExChr</td>
<td>?</td>
<td>[5]</td>
</tr>
<tr>
<td></td>
<td>pMOL28</td>
<td>chrBACDEFONPYZ</td>
<td>6204–6194</td>
<td></td>
<td>CMGI-28a</td>
<td>ExChr, FS, MFS</td>
<td>+</td>
<td>[12]</td>
</tr>
<tr>
<td></td>
<td>CHR2</td>
<td>cuxDCBAF</td>
<td>5030–5034</td>
<td>Ag+, Cu2+,Zn2+</td>
<td>-</td>
<td>ExHME4</td>
<td>+</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>pMOL30</td>
<td>silDCBA cuxFl</td>
<td>6133–6136, 5953</td>
<td></td>
<td>CMGI-30b</td>
<td>ExHME4 (+)</td>
<td>[10]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CHR2</td>
<td>copSRABCDO</td>
<td>5763–5668</td>
<td>Cu2+</td>
<td>-</td>
<td>ExCop, seq</td>
<td>+</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>pMOL30</td>
<td>copVIFMNRSABCDOJF-OLQHIEW</td>
<td>6105–6119, 6382, 6120–6124</td>
<td></td>
<td>CMGI-30b</td>
<td>ExCop, ExP, seq</td>
<td>+</td>
<td>[75]</td>
</tr>
<tr>
<td>not plasmid-redundant</td>
<td>CHR1</td>
<td>arsPHICBCBbM</td>
<td>3027–0334</td>
<td>HAsO42-, AsO33-</td>
<td>CMGI-7</td>
<td>AsRed (ExAs)</td>
<td>+</td>
<td>[94]</td>
</tr>
<tr>
<td></td>
<td>CHR1</td>
<td>cupric</td>
<td>3523–3525</td>
<td>Ag+, Cu2+,Zn2+</td>
<td>-</td>
<td>ExP</td>
<td>+</td>
<td>[75]</td>
</tr>
<tr>
<td></td>
<td>CHR1</td>
<td>agrCBARS</td>
<td>1748–1752</td>
<td></td>
<td>-</td>
<td>ExHAE</td>
<td>?</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>CHR1</td>
<td>hmrRS yodB hmx2B1A</td>
<td>3016–3011</td>
<td></td>
<td>divalent metals CMGI-4</td>
<td>ExHME1</td>
<td>+</td>
<td>[10]</td>
</tr>
<tr>
<td></td>
<td>CHR2</td>
<td>hmxCB1A</td>
<td>3836–3838</td>
<td></td>
<td>-</td>
<td>ExHME3a</td>
<td>-</td>
<td>[8]</td>
</tr>
<tr>
<td></td>
<td>CHR2</td>
<td>zniABCzr znePRICAB</td>
<td>5319–5323, 5325–5330</td>
<td></td>
<td>-</td>
<td>ExHME3a, MFS</td>
<td>?</td>
<td>[58]</td>
</tr>
<tr>
<td></td>
<td>CHR2</td>
<td>hmyCBhmyA (‡)</td>
<td>4120–4123</td>
<td></td>
<td>-</td>
<td>ExHME3b</td>
<td>+</td>
<td>[8]</td>
</tr>
</tbody>
</table>

*Chromosomal HMR resistance loci with genes that have a counterpart with a homologous function on a plasmid are called partially redundant, other loci are not redundant.*

*Underlined genes are regulators; bold genes are new in respect to an earlier report [8] (see also Table S2); gene parts created by a frameshift are indicated by ‘and’; IS- or Tr-insertions are indicated by a pipe symbol (|); truncated genes are indicated with a delta (∆); insertion of Tn6049 in pbrH2 (†); cuxB truncated by Tn6055-mediated rearrangement (inversion) and insertion of IS1088 between cuxL and hns (‡); insertion of IS8088 between hmy6 and hmyA.*

*CMGI, *Cupravidus metallidurans* CH34 Genomic Island; (P), PAGI-2 family.*

*Mechanisms of HM detoxification: ExP, efflux (P1-ATPase); pCDF, putative CDF-like; seq, periplasmic sequestration; HgRed, reduction of Hg2+ into Hg0 which then volatilizes out of the cell; HgTra, tentative transport of Hg2+; ExCh, transport of chromate; FS, Fe-SOD (chrC); MFS, major facilitator superfamily permease (i.e., chrP); CDF, cation diffusion facilitator; ExHME, efflux by RND-HME type 1 to 4 (see text); ExHAE, efflux by HAE-RND (unknown substrate); ExCop, extrusion of Cu2+; AsRed (ExAs), reduction of HAsO42- into AsO33- and efflux of AsO33-.*

*Functionality: +, active; (+), partly active; ?, unknown; -, inactive.*

doi:10.1371/journal.pone.0010433.t005

The P-type ATPase family of ion pumps [91]. These export systems may act on a range of systems, including Cu2+, Ag+, Cu2+, Co2+, Cd2+, Ni2+, Pb2+ and Zn2+ [8,10,59,75,80,83,85,90,92,93]. Alternative mechanisms with specificity to other cations or oxyanions include the chromate CrO42- efflux system [5], the reduction of arsenate HAsO42- by ArsC reductases followed by the extrusion of the formed arsenite AsO33- by the ArsB peromase [94], the CopABC/ProABC efflux system [95,96]. The reduction of Hg2+ into metallic mercury [4,59,97,98] (Table 5; Table S2). However, the function and mode of action of many proteins encoded by putative heavy-metal resistance genes remain undetermined. Structural analysis of proteins at the molecular level offered some mechanistic information (i.e., via protein crystallization, nuclear magnetic resonance analysis, and cryo electron microscopy) for the *C. metallidurans* proteins CopK (Rmet_6108), CopH (Rmet_6122), and CzcE (Rmet_5976) expressed by pMOL30, and CnrX (Rmet_6206) expressed by pMOL28 [99–102]. In other cases, such data were gained from aa sequence based comparison of closely related proteins, such as for the CzcD protein (Rmet_5979) that greatly resembles the recently studied *E. coli* YIp zinc transporter protein [103,104].

It is worth mentioning that RND sytems for HME (Heavy Metal Efflux) were first discovered in the *C. metallidurans* CH34 with
the archetypes CzcCBA, conferring resistance to Co, Zn and Cd [9,103], and CnrCBA, conferring resistance to Co and Ni [80,106], with their genes, respectively, located on megaplasmids pMOL30 and pMOL28. Such RND-HME complexes comprise three components spanning the complete cell wall and mediating cation/proton antipporter efflux via a chemiosmotic gradient from the cytoplasm towards the periplasm and into the exterior of the cell. The main component of these complexes (i.e. CzcA, CnrA) is associated with the inner membrane and typically contains 12 transmembrane domains. An outer-membrane efflux protein (i.e. CzcC, CnrC) functions as an auxiliary element to export metals and is characterized by a channel formed by 12 beta-sheets extending further into the periplasm by alpha-helices. The third component (i.e. CzcB, CnrB) is a membrane fusion protein (MFP) bridging the two membrane-bound export proteins. This configuration effectively avoids the release of metal ions in the periplasmic space, and ensures their full disposal by the cell.

Other chemiosmotic antipporter efflux systems include the NccCBA complex (Ni, Co, Cd) of C. metallidurans 31A (formerly known as Alcaligenes/Lachnobacter xyllosoxidans) [107], the CzcCBA complex (Cd and Zn) of Pseudomonas aeruginosa [108], and the CusCBA complex (Cu and Ag) of E. coli [109]. Remarkably, C. metallidurans CH34 possesses 12 such transenvelope RND-HME export devices (Table 5).

In contrast, the cation diffusion facilitator (CDF) family transporters, which act as chemiosmotic ion-proton exchangers, and the P-type ATPases, that use ATP-energy to pump metal ions out of the cytoplasm, are single-subunit systems located in the cytoplasmic membrane. Sometimes, all three types of metal-ion-efflux systems, e.g. RND-HME, P-ATPase, and CDF, occur in the same locus. For instance the cze cluster of pMOL30 contains besides the RND-HME module czeCBA also czeD (Rmet_5979; a CDF protein) and czeF (Rmet_5970; a cation efflux P1-ATPase). We and others [8,37,75,76,110,111] suggested that this combination entails much higher resistance to heavy metals and better detoxification of cellular compartments, especially the cytoplasm (thiol pool). Together, there are at least 30 heavy-metal-efflux systems in C. metallidurans CH34 encoded by 155 genes in 25 loci (Table 5; Table S2). To our knowledge no other soil bacterium has acquired so many ion efflux systems specialized in removing heavy metals. We think that genomic islands likely played an important role in this procurement. This is most notable for CHR1 (viz. copSRABCD) and CHR2 (viz. cop, pbr, czc) encoding the cop and czc systems of pMOL28 and pMOL30 encoding interwined nested HMR loci.

In addition to this interplay of HMR determinants there is a redundancy of HMR genes across the four replicons of C. metallidurans CH34 (Table 5). For instance the two chromosomes of CH34 contain a number of HMR loci, i.e., chrBAF, copSRABCD, pbrR-cadaA/pbrA-pbrC, and czeICBRA-RS that have counterparts on either pMOL28 (chr) or pMOL30 (cop, pbr, czeF); further their products display a high sequence similarity (i.e., for the products of the copSRABCD locus this ranges between 50 and 87% identity over their entire length). Such duplicates may reflect multiple acquisition of the same HMR locus (possibly derived from different lineages), or may represent internal rearrangements generated by DNA mobility. However, not all of these loci (or genes within loci) are functional, as some HMR loci appear to be affected by either a frameshift mutation (e.g. the merQ gene of CMGI-1 and the nccB gene of CMGI-30b) or by partial deletion (e.g. hmmA of CMGI-4 and merP of CMGI-30a ). A number of HMR loci could be extended or updated via MaGe-based annotation (Tables 5 and S2, gene names in bold face). Some of these novel genes are strongly induced by heavy metals (not shown), most notably czcF (Rmet_5975) and mmrQ (Rmet_6143; formerly ref-167), and to a lesser extent copQ (Rmet6121). We noted previously that CzcF, MmrQ and CopQ have a signal sequence (well defined by SignalP embedded in MaGe) and have a common domain motif, RIXXTDG, repeated 3 to 4 times [10]. A closer look reveals that these domains are connected with spacers of defined length and sequence. Three other genes that are highly induced in the presence of metals, mmrQ (Rmet_4461), mmrR (Rmet_4187), and mnrQ (Rmet_4906), encode proteins with apparent signal sequences and similar domains and spacers. Note that mmrQ has an almost identical paralog mmrQ' (Rmet_5594) thus both genes are covered by the same probe in transcriptional analysis. The mmrQ gene is linked to the cze locus on CHR2 (Table 5) displaying excellent synteny with metal efflux loci in the three other Cupriavidus genomes (in spite of an IS1008 insertion between the mmrQ and the cze genes). In contrast, the other genes mnrQ, mnrR and mnrQ' appear to be unlinked to any HMR locus. The metal induction of these seven genes (czcF, mmrQ, copQ, mmrR, mmrQ, mnrQ and mnrQ'/Q) follows a varied and overlapping response, with very high induction by some metals and no induction by other metals. For example, czeF and mnrQ are strongly induced by Cd2+, Cu2+, Ni2+ and Zn2+; moderately induced by Pb2+ and not induced by Hg2+; while the mmrQ gene is highly induced by Cd2+, Hg2+ and Pb2+, only moderately by Zn2+ and not by Cu2+ or Ni2+ (data not shown). Much to our surprise, according to BLAST analyses against GenBank, these genes (with copQ as prototype) and their orthologs are restricted to the genera Cupriavidus and Ralstonia. We are currently extending this analysis using pattern matching software to search for CopQ-like protein sequence motifs across multiple genomes. We are also performing additional transcriptomic studies on these genes induced by a wider range of metals and other types of stress, and are investigating their gene upstream regions for regulatory DNA sequences.

Through genome annotation, we also found additional candidate HMR genes that warrant further investigation. They are genes dmeF (Rmet_0198 on CHR1) and jsF (Rmet_3406 on CHR2) encoding CDF-type transporters, and the atmA gene (Rmet_0391 on CHR1) coding for an ABC-type transporter. According to MaGe data, these genes have orthologues in the three other Cupriavidus species and are well-conserved in other bacteria. Interestingly, disrupting dmeF in CH34 causes the Cnr and Czc systems of pMOL28 and pMOL30, respectively, to become ineffective against Co2+, indicating that dmeF is essential in cobalt homeostasis [112]. Correspondingly, jsF mutants of strain CH34 showed decreased resistance to Ni2+, Co2+, Cd2+ and Zn2+ [112] and atmA mutants of strain CH34 exhibit lower resistance to Ni2+ and Co2+ [113].

Clearly, C. metallidurans CH34 has a history of colonizing biotopes with moderate to high-levels of bioavailable metals and, via the uptake and reshuffling of HMR loci, adapted very well to such extreme conditions. This adaptation required the dedicated management of incoming HMR systems to avoid functional redundancy (i.e., under stress when resources are scarce), the optimization of plasmid-born genes to counter elevated levels of metal toxicity, and the economical interplay between plasmid- and chromosomal-genes, supported by a large diversity of sigma subunits of RNA polymerase [114]. We expect that whole-genome transcriptomic studies using a wide range of metal cations and -oxyanions will reveal additional HMR genes and highlight new interplays between chromosomal and plasmid borne genes. Such analyses also will refine our knowledge about the intricate regulation patterns apparently governing the response to high concentrations of heavy metals.
Small-molecule transport and protein secretion systems

The four *Cupriavidus* genomes were submitted to the TransportDB [115,116] website (http://www.membranetransport.org/) to identify efflux systems and transporter families by the TransAAP (Transporter Automatic Annotation Pipeline) tool. The TransportDB relational database holds the complete membrane transport complements of 288 bacteria classified into protein families according to the Transport Classification (TC) system [117,118]. The recently updated TC system organizes nearly 5,000 representative transporters in more than 500 families [118] and is freely accessible at http://www.tcdb.org/. From the mode of transport and the mechanisms of energy coupling, TCDB discerns four major classes of transporters: Channels; secondary carriers; primary transporters; and group translocators. Each class is divided in families and subfamilies according to their function, phylogeny, and substrate specificity. In our analysis with TransAAP we consider only five trans-membrane families: 1A: alpha-Type channels; 2A: Porters (uniporters, symporters, antiporters); 3A: P-P-bond-hydrolysis-driven transporters; 4A: Phosphotransfer-driven group translocators; and 9A: Recognized transporters of unknown biochemical mechanism. A more extensive analysis using all 9 TC classes was reported albeit on the unfinished genome sequence [119]. The numbers of transport proteins predicted by our TransAAP analysis were about the same in all four *Cupriavidus* species (Table S3), amounting to 12 to 14% of their total number of proteins. This appears to be typical for β-proteobacteria [120]. In addition the distribution of transporter classes by species is very similar (Table S3). The largest group of transporters (ca. 400 per species) is presented by the porters (TC2A), closely followed by the primary transporters (TC3A). The largest families herein are the Major Facilitator Superfamily (MSF) (TC2A1) and the ATP-binding Cassette (ABC) Superfamily (TC3A1) that are involved in the uptake of various nutrients, such as amino acids, sugars, nucleosides and various organic and inorganic anions. They also remove antibiotics, metabolites, and toxic byproducts. A third large group is the Tricarboxylate Transporter (TTT) Family of solute receptors (TC2A80) that carry a typical carboxylate-binding motif (the “Bug domain”), and probably is involved in diverse metabolic and transport activities (discussed further under ‘Signal Transduction’). These three large families constitute up to 70% of all predicted transporter proteins in the four *Cupriavidus* strains. Of immediate interest to our research on *C. metallidurans* are the systems involved in transporting transition metals: MIT (1A35), CDF (2A4), ZIP (2A5), RND (2A6), CHR (2A51), NiCoT (2A52), P-type ATPase (3A3), MerT (9A2), and TerC (9A30) (with TC class in parentheses). These systems, their genes, and their regulation were discussed in detail elsewhere [8,10,75,76]. Evidently, the cellular household of *C. metallidurans* CH34 is tuned towards adapting to a wide range of heavy-metal species (see our brief update on HMR genes earlier in this discussion) or to surviving in biotopes that are either occasionally or constitutively rich in heavy metals. Unsurprisingly therefore, *C. metallidurans* and its related strains most often are found in heavily polluted anthropogenic- or industrial-environments [6,8].

Only a few transporter systems are unique to *C. metallidurans* CH34 compared to the other three *Cupriavidus* species (Table S3): they are the 2-Hydroxyxycarboxylate Transporter (2-HCT) Family (TC2A24), and the Iron/Lead Transporter (ILT) Family (TC9A10). The former is represented in *C. metallidurans* by gene Rmet_4998 (situated on CHR2 and listed in MaGe as encoding a citrate carrier) while the latter is represented by two members Rmet_5945 and Rmet_6106 (both located on plasmid pMOL30 as part of the cex and pbr operons, respectively).

Nine transport systems not present in *C. metallidurans* were found in at least one of the other *Cupriavidus* species (indicated in gray in Table S3) but apparently this observation lacks physiological relevance. However, no member of the Monovalent Cation (K+ or Na+)-Proton Antiporter-3 (CPA3) Family (TC2A63) was detected in *C. metallidurans*, while the other three *Cupriavidus* species have seven to nine predicted transport proteins belonging to this family. A closer examination reveals that these predicted proteins are subunits of a large multicomponent antiporter with their encoding genes organized as a single operon (e.g., denoted as the mnaABCDDEF operons in *C. eutrophus* H16 and *C. taiwanensis*).

Such antiporters were linked in *Bacillus* species with cellular responses to short-term changes in salinity or pH [121]. Interestingly, MaGe synteny maps show that *D. acidovorans* and *B. petrii* have exactly the same set of genes in one single locus and so does *Sinorhizobium meliloti* and *Bacillus halodurans* but not *Ralstonia pickettii* nor any of the three *R. solanacearum* species listed in the MaGe databases.

*C. metallidurans* CH34 possesses fully intact Sec-dependent and Sec-independent (twins-arginine translocation “TAT”) protein export pathways for secreting proteins across the inner/periplasmic membrane. The Sec-dependent pathway consists of a protein translocation channel made by the contact of two heterotrimERIC complexes SecYEG (encoded by Rmet_3297, Rmet_3340, and Rmet_0926) and SecDFYajC (encoded by Rmet_2945-47). The SecA and SecB proteins (encoded by Rmet_3118 and Rmet_0254, respectively) cooperatively bind with the signal sequence of secretory proteins that, after translocation, is cleaved off by signal peptides SPaseI (Rmet_2420) or SPaseII (of which there are three candidate enzymes in *C. metallidurans* i.e. encoded by Rmet_2304, Rmet_2886, and Rmet_5948). Here, an important accessory component is the membrane insertase YidC (Rmet_3613) that may act independently from the Sec translocase and is involved in folding the inner membrane proteins and the biogenesis of respiratory-chain complexes (listed under TC class 2A9 in Table S3). Genome analysis of *C. metallidurans* also identified Rmet_3094 encoding a signal recognition particle (SRP54) that acts with FtsY (Rmet_0281) as an alternative to the SecA-SecB pair for targeting protein precursors. The TAT secretion pathway is specialized to translocate proteins that already have undergone folding in the cytoplasm [122]. It consists of the three subunits Tata, TatB and TatC (encoded by Rmet_3235-27; listed under TC class 2A64 in Table S3) that form distinct multimeric Tata or TatABC complexes acting together at the point of translocation.

Extracellular secretion of folded proteins across the outer membrane depends on the Type II Secretion System (T2SS), also often referred to as the “general secretion pathway (gsp)”. In proteobacteria this apparatus comprises 12 to 16 different gene products forming a multiprotein complex spanning the periplasmic compartment; it shares many features with the type IV pilus biogenesis system, including the ability to assemble a pilus-like structure. There are two clusters of gsp genes in *C. metallidurans* CH34. The largest cluster of 11 genes gspGHJKNLMNDEF (Rmet_3384-94) is situated on CHR1 and is well conserved in the *Cupriavidus* species, while CHR2 (Rmet_4704 through Rmet_4713) supports another smaller cluster consisting of gspD, gspEFG, two additional copies of gspG, and four conserved hypothetical genes (three of which predictably code for transmembrane proteins). Lateral gene transfer may have allowed the acquisition of this genomic region, which does not display any synteny with the genomes of the other three *Cupriavidus* species. Interestingly, this cluster of genes is flanked by two IS elements in *R. solanacearum* GMI1000. We detected around 30 genes required
for type-IV pilus biogenesis and assembly in C. metallidurans, some of which occurred in multiple copies across the genome (see also section on motility).

There is no direct evidence for a functional pathogenicity-related Type III Secretion System (T3SS) in C. metallidurans. However, all the fli and fla genes encoding the structural proteins for a basal flagella body, i.e., those that display homology to T3SS genes through common ancestry, are present (Rmet_3089-99, Rmet_5258, Rmet_5261, Rmet_5263-64, Rmet_5299, and Rmet_5301-03) together with all additional fli, fla, mot, and flg genes required for assembling full flagella (see section on motility).

The Type IV Secretion System (T4SS) is a multifunctional family of translocation pathways that mediate the transfer of macromolecules (proteins, DNA or DNA-protein complexes) across the bacterial cell envelope either to the outside medium or to other cells (prokaryotic or eukaryotic). These systems, widespread in nature, are evolutionary related to bacterial conjugation systems and largely contribute to the rapid dissemination among bacterial populations of a range of catabolic- and metabolic-genes and of antibiotic resistance genes and other virulence traits. Moreover, because often they are sited on plasmids or other mobile genetic elements, many T4SS are laterally transferred between species. We discerned two T4SS gene clusters in the C. metallidurans genome, each with nine genes homologous to the vir genes of the Ti-plasmid of Agrobacterium tumefaciens (Rmet_1336 through _1350 and Rmet_1346 through _1359). These clusters are located on CHR1 as part of TrnA371-like mobile genetic elements (MGEs) and may be involved in GI maintenance or conjugational transfer [66]. We denoted these MGEs as genomic islands CMGI-2 and CMGI-3; they carry large maintenance or conjugational transfer [66]. We denoted these like mobile genetic elements (MGEs) and may be involved in GI related Type III Secretion System (T3SS) in C. metallidurans. However, all the section on motility).

The presence of multiple sigma-70 housekeeping factors is not unique to C. metallidurans; E. coli H16 and C. taiwanensis both contain two sigma-70 factors (albeit both located on the main replicon) while C. finitumunonae JMP134 contains three copies of this sigma factor (one on each chromosome and one on the megaplasmid) (Table S4). In our study, only B. selenwurans contained four sigma-70 genes, all on its second chromosome. Only one copy was detected in the other genomes we analyzed (of R. solanacearum, H. arsenicoxydans, J. magdonicus, and E. coli).

Pairwise Smith-Waterman alignments of RpoD protein sequences for all analysed genomes (represented in Figure S3 as a heatmap) showed that C. metallidurans RpoD1 (Rmet_2606) forms a tight group with orthologous RpoD proteins from the three other Cupriavidus species (Rmet_A0888, RALTA_A2226 and H16/A2275) whose genes all sit on CHR1. In contrast, in this same analysis, C. metallidurans RpoD2 (Rmet_4661) did not group with any of the other RpoD orthology in the Cupriavidus family (Figure S5); indeed, it had more significant BlastP hits with the RpoD proteins of Bordetella and Bordetella species. Therefore we believe that the C. metallidurans RpoD2 protein originates from another lineage. The co-existence of multiple sigma-70 factors may be useful for regulatory intricacy. Assuming that these factors would recognize the same set of promoters, slight nuances in factor binding and/or DNA interaction support finely tuned expression patterns such as when cells enter a different growth mode, or at the onset of a particular stress condition. It is unclear how this complexity in transcriptional regulation is attained. It would be interesting to undertake mutational analysis on either one of the C. metallidurans rpoD genes (Rmet_2606, Rmet_4661) to check their functionality, to follow cell viability of mutants, and subsequently to measure whole-genome gene expression under a variety of conditions.

Grosse et al. [114] described at least ten sigma factors in C. metallidurans CH34 that belong to the extracytoplasmic function (ECF) family, most of which were implicated in HMR. These authors also showed that one particular sigma factor, RpoI (Rmet_1120) is essential for producing siderophores in CH34 (see following section) and that this sigma factor and two other related ones, RpoJ (Rmet_4001) and RpoK (Rmet_4499), were induced by iron depletion. Consequently, the researchers classified these three factors as belonging to the ECF:FeI group (type 1) to indicate a similar function in iron assimilation as FeCr in E. coli. Interestingly, the other three Cupriavidus species (Cpin, Ceut, and Csu) do not possess RpoJ or RpoK factors while only C. taiwanensis has an RpoI ortholog (RALTA_B1239) albeit with weak similarity (37% aa identity, minLrap 0.98) (Table S4).

The complex transcriptional behavior of C. metallidurans not only is manifested by the many alternative sigma factors, but also its high count of transcriptional regulators. Using MaGe-based data for CH34 at least 543 transcription factors were discerned, corresponding to 9% of all genes in the genome, while the average percentage transcription factors is only 2.4% for the 673 bacteria in the DBD database [124,125]. This large difference partially may be explained by the facts that the relative number of transcriptional regulators increases with genome size [126], and generally depends on the organism’s lifestyle [127,128]. Nonetheless, our analysis shows that only Bordetella pertussis strain Tohama I, six Burkholderia species, and Streptomyces coelicolor display a higher density of transcriptional regulators. Based on Pfam predictions [129], roughly a quarter of all regulators in C. metallidurans belong to the LysR family (n = 126).
component metal responsive regulators are currently recognized in the CH34 genome. Although also LysR, TetR, and MarR families reportedly contain members that are directly involved in metal homeostasis or metal-regulated responses [131], only very few well-documented cases are known so far. Nonetheless, these three families are represented in CH34 by a total of a few hundred proteins so it may be worthwhile to look also within this pool for further candidate metal response regulators.

Among two-component systems the OmpR system constituted the largest group (n = 33) while the LuxR family (best known for its quorum sensing regulators) also was well represented (n = 41) (Table S5). According to the MiST2 database [132], there are 56 two-component pairs of adjacent genes in the CH34 genome. Several of those have been implicated in HMR, notably two pairs of copRS and czfRS, and one pair each of agrRS, hmzRS, zincRS, and znuRS (underlined in Table 5). From MaGe data and the DBD and MiST2 databases we discovered that the second repliction of C. metallidurans (CHR2) carries proportionally almost twice the number of regulatory genes as compared to the main repliction CHR1. This enrichment in regulatory genes agrees well with the earlier notion that CHR2 carries many functions relevant to the survival and responsive behavior of C. metallidurans CH34 (Figure 2).

Clearly, the wide diversity of sigma factors and transcriptional regulators in C. metallidurans CH34 points to intricate regulatory mechanisms for activating genes in response to environmental stimuli. For instance, microarray analyses of genome-wide gene expression studies on strain CH34 pointed to layered multiple responses when cells are challenged with heavy metals [10]. Similar complex expression patterns were shown for CH34 cells grown under other harsh conditions, including simulated microgravity, growth on basalt, and spaceflight (data not shown). To unravel these regulatory mechanisms and better understand the molecular basis of C. metallidurans’ ability to survive prolonged periods of stress, we now are focussing on reconstructing the regulatory networks by systematically identifying regulatory motifs in the upstream regions of stress-induced and co-regulated genes.

Siderophores

At neutral pH and oxic conditions, bacteria cannot readily utilize low-solubility Fe(III) minerals. Accordingly, they synthesise and secrete soluble iron-complexing compounds, called siderophores. C. metallidurans CH34 grown under iron-limiting conditions produces a unique phenolate-type siderophore containing neither hydroxamate nor catecholate groups [133]. This siderophore, originally referred to as “alcaligen E”, eventually was classified as a polycarboxylate staphyloferin B [134] because it resembled the siderophores typically produced by staphylococci. It since was recognized that other organisms produce staphyloferin B, such as R. solanacearum [135]. Notably, C. metallidurans staphyloferin B can interact with Cd to thereby decreasing the metal’s bioavailability and hence lowering its cellular toxicity in the bacterium. Also, siderophore-producing C. metallidurans tested in bioaugmentation-assisted phytoextraction increased, by a factor of 5, the accumulation of Cr(VI) in plant shoots [136]. Future studies on siderophore production in C. metallidurans CH34 might thus open up exciting lines of research in heavy metal tolerance and bioremediation.

Genes for staphyloferin B biosynthesis (Rmet_1109-1118) and two genes encoding ferric/ferriichrome siderophore TonB-dependent receptors BfrC and FcuA (Rmet_1104 and Rmet_1108, respectively) are located on CHRI within a 52 kb region that lacks synteny with Cpin, Ceut, and Cai. Although a tRNA-Ser gene preceeds this region, we could not identify any tyrosine-based site-
specific recombinases, flanking IS elements, or direct repeats; hence, we did not classify it as a genomic island (see our discussion on Mobile Genetic Elements). It includes genes encoding a siderophore synthase decarboxylase (sdi, Rmet_1110), an aldolase (Rmet_1111), staphyloferrin B biosynthesis proteins (Rmet_1112, Rmet_1113, and Rmet_1115), a permease of the Major Facilitator Superfamily (MSF) (Rmet_1114), an ornithine cyclod diamase (ocd, Rmet_1116), needed to biosynthesize alcaligin [137], a cysteine synthase (gydF, Rmet_1117), and the staphyloferrin B receptor (aleB, Rmet_1118). This core set of genes (Rmet_1110 to Rmet_1117) displays high synteny with the staphyloferrin B biosynthesis regions of Ralstonia solanacearum GMI1000 and Staphylococcus aureus MW2 [135] and of all other S. aureus strains present in the MaGe database. In addition, R. solanacearum GMI1000 and C. metallidurans CH34 share the aleB gene (Rmet_1118) and a gene for the sigma-19 transcription factor FecI (RpoI; Rmet_1120). The insertion of a mini-transposon Tn5 into aleB decreased iron uptake three to four times in iron-starved cells compared to the wild type strain CH34 [133]. The C. metallidurans CH34 genome lacks the entA, entB, entC, entD, entF, and entS genes required for enterobactin biosynthesis but seems to contain an entE gene (Rmet_3777) on CHR2.

The regulation of staphyloferrin B biosynthesis is controlled by two systems: A fecRI regulon located immediately upstream of aleB, and the far gene distant from aleB. The regulator FecR (Rmet_1119) is a sensor for the presence of iron dicitrate in the periplasm and regulates its transport by inactivating FecI (Rmet_1120). The Fur protein (for ferric uptake regulation) tightly represses staphyloferrin B biosynthesis under iron-replete conditions [130]. In both C. metallidurans and R. solanacearum, a gene (Rmet_2979) encoding the global virulence regulator PhcA is located next to the far gene (Rmet_2976). In R. solanacearum, PhcA negatively regulates staphyloferrin B biosynthesis [135]. Furthermore, the phcA gene of C. metallidurans CH34 can complement a phcA− mutation of R. solanacearum [139]. Interestingly, two Fur-like regulators, Rmet_0128 and Rmet_5746, respectively, were found on CHR1 and CHR2.

Besides the aleB locus, the genome of C. metallidurans CH34 contains additional genes that are probably involved in binding siderophores. These are Rmet_0837 and Rmet_4617 encoding hydroxamate-type ferrisiderophore receptors, Rmet_4496 and Rmet_4497 encoding ferric siderophore receptors, Rmet_5806 and Rmet_5807 encoding FecA-type ferric citrate outer membrane receptors, a small locus Rmet_2277/9 encoding a complete TonB-ExbB-ExbD type energy-transducing complex. Supplementary genes encode a TonB-like protein (Rmet_3055) and ExbB/ExbD homologues (Rmet_0536/7). The presence of multiple siderophore binding and translocation systems is consistent with the observation that C. metallidurans CH34 employs various ferric ion chelators for iron uptake, including desferririoxamine B, D2 and E, desferriferricocin, desferriferrihydrin, desferriferrichrome A, and ornibactin [133]. In addition, its genome encodes proteins involved in iron storage (bfr, Rmet_0195, Rmet_0246) and soluble ferrous Fe(II) transport (fesB, Rmet_5890 and Rmet_5890). Table S6 offers an overview of CH34 genes implemented in iron uptake and -metabolism.

Signal transduction

Like many other environmental bacteria, C. metallidurans CH34 deploys early warning systems and response mechanisms to cope with many different processes such as growth and development, metabolic regulation, and sensing, and to react swiftly to various stresses. In bacteria, the linkage of input signals to output responses entails a cascade of protein modifications, basically of two modes: One-component and two-component systems (denominated as 1CST and 2CST, respectively). In 1CST systems, the input-and output-domains are fused in a single protein, whereas in 2CST systems, the sensory- and receiver-domains lie on two different proteins (reviewed by Galperin and Gomelsky [140]). The latter has the advantage that the sensory component can be membrane-associated to receive extracellular stimuli while the receiver domain can be part of the actual regulator, often acting at the level of transcription. Until recently, 2CST systems were thought to be the primary mode of signal transduction in prokaryotes, but it became evident that 1CST systems carry out much of the signal transduction [128]. The input and output domains of these systems display a remarkable diversity, are widely distributed across species, and are mainly involved in DNA binding and the intracellular detection of environmental stimuli [128]. To survey signal transduction proteins in C. metallidurans and related organisms we used the Microbial Transduction System (MiST) database [141] (http://genomicsornl.gov/mist/) that holds the predicted signal transduction proteins for nearly 750 bacterial and archaela organisms, using Pfam and SMART reference domains (added note: an extended database MiST2 was very recently established at http://mistdb.com/ [132]). These data demonstrated that all four Cupriavidus species contained approximately the same proportional number of proteins putatively involved in signal transduction (approx. 9–10% of their proteome) (Table S7). The same proportion was evident in D. acidovorans, Janthinobacterium, and B. xenovorans, while B. petrii, R. solanacearum, and H. arsenic oxidans displayed slightly lower values (8.3–8.2- and 7.0%, respectively). In contrast, and in line with its deterministic lifestyle, the intracellular symbiont P. necessarius STIR1, whose single replicon displays good synergy with C. metallidurans CHR1 (see section on Phyllogeny), deploys only 1.6% of its protein complement for signal transduction. For all the organisms we compared, there were two- to three-fold more predicted proteins involved in 1CST systems than in 2CST systems (Table S7) (the only exception being B. petrii that has a slightly higher preference for 1CST systems). Also, this comparison demonstrated that the second replication of all organisms with a multipartite genome (i.e. the four Cupriavidus species, R. solanacearum, and B. xenovorans) displayed an enrichment on their second replicon for proteins involved in signal transduction (data not shown), once more underpinning the importance of this replication for adaptation and survival.

According to MiST data, protein domains commonly used for signal transduction in C. metallidurans CH34 were Bug (123), LysR_substrate (122), Response_reg (98), HATPase_c (72), HisKA (55), HAMP (45), TetR_N (36), GGDEF (30), PAS (30), FCD (24), EAL (23), MCPsignal (16), Sigma54_activat (16), and GAF (10) (domain names are in Pfam notation with numbers in parentheses). Most of these domains are integrated as signal receivers into transcriptional regulators (e.g. LysR_substrate, TetR_N, FCD) whether in 2CST or 1CST format, while others have more specific functions suggesting a role in chemotaxis (e.g. Response_reg, MCPsignal), His-Asp phosphorelay (HisKA, HAMP, HATPase_c), cyclic di-GMP turnover (GGDEF, EAL, GAF), or solute binding (e.g. Bug). Undoubtedly, many output domains constitute DNA binding motifs for the transcriptional regulation of genes. These domains typically have helix-turn-helix (e.g. HTH_1, IcIR) or winged helix (e.g. Trans_reg_C, MarR) structures (Table S7, not all data shown).

In general, the distribution of signal transduction domains is the same across the four Cupriavidus species (Table S7) and is not very different from their distributions in the other free-living species. However, we noted an important deviation for the Bug domain that occurs at high frequency (>100) for all four Cupriavidus species
but is far less prevalent in either *R. solanacearum* or *B. xenovorans* (respectively, 2 and 13 Bug domains). The Bug domain, named after the *Bordetella pertussis* bug*T* gene identified by Antoine et al. [42], was reported as the first member of a large family of periplasmic solute-binding receptors present in in great numbers (up to 101 per genome) in *Bordetella* species, with much lower numbers (1 to 11) in some other bacterial genomes [47]. The only exception in their study was *C. metallidurans* for which 102 Bug homologs were predicted. Typically, Bug proteins are highly conserved in size (median size 325 residues), contain an identifiable signal-sequence, and share well-conserved carboxylate-binding motifs that form a characteristic Venus-fly-trap fold [143,144]. This latter characteristic and the fact that the family member TctC, a periplasmic citrate-binding protein of *Salmonella typhimurium*, is part of a novel transport-system TctABC, led researchers to classify many members of this family to the tripartite tricarboxylate transporter (TTT) family (listed as TC 2.A.80 in the Transport Classification Database; http://www.tcdb.org). Nevertheless, although Bug proteins clearly form a distinct family of solute receptors, their sequences have diverged widely and their genes often are not linked to transport-related operons. Antoine and co-workers [47,145] posited that these bug genes may be used by other transport systems, or involved in other metabolic activities or other pathways activated under stress. In this context it is interesting to note that many of the bug genes in CH34 seem linked or are in the close vicinity of transporter and regulatory genes. We also find it intriguing that *C. metallidurans*, and the other Cupriavidus species, harbour so many bug genes: Out of 200 genomes analysed by Antoine et al. [47], only *Bordetella* species and *C. metallidurans* were listed with a prodigious number of bug genes. We might argue that *Cupriavidus* species are free-living organisms wherein the use of solute receptors for specialised strategies of adaptation and survival evolved, but then, we question why so few other free-living organisms failed to develop this approach? We surveyed for the Bug domain in the 1,216 complete and draft, bacterial and archaeal genomes currently in the MiST2 database [132] and found 2,377 proteins with Bug domains. Outside the genera *Bordetella* and *Cupriavidus* only four genera had a comparable count: *Delftia acidovorans* SPI-1 (159), *Variovorax paradoxus* S110 (137), *Vernimicrobium eiseniae* EF1-2 (131), and *Polamaromas* sp. JS666 (97). The only other genomes with a large count (>40) were *Diaphorobacter sp.* TPSY (45), *Acidovorax avenae* sp. *citrulli* (55), *Acidovorax*, JS42 (46), and *Polamaromas naphthalenivorans* CJ2 (41). To our surprise, all organisms that showed an overrepresentation of bug genes belonged to the same beta-proteobacterial order of Burkholderiales (Figure 5; Table S8). We propose to call this group of prokaryotes the “Bug Receptor Bacteria” or BuRBs. Such overrepresentation may be a recent innovation in bacteria, or may be more useful to those with a particular lifestyle and habitat. Whether the propensity to utilise large numbers of Bug-domain containing receptors arose within the genus *Bordetella* and was radiated out to closely related genera, or whether it has originated elsewhere within the order of Burkholderiales is an open question. It is useful here to mention that *B. pertis*, which carries 107 bug genes, is not host-restricted like the other *Bordetella* species. It is an environmental species typically obtained from river sediment, grass root, and polluted soil [146,147] and has the capacity to degrade aromatic compounds and detoxify heavy metals [148]. However, *B. petrii* strains subsequently were isolated from patients with mandibular osteomyelitis and chronic supplicative mastoiditis [149,150]; thus, this species appears to be an opportunistic pathogen after all. It also contains many mobile genetic elements [148], some strikingly resembling those found in CH34 [59], that may have played an intermediate role in the evolution of BuRBs and the spread of the Bug domain. We speculate that BuRBs evolved many Bug-domain proteins as periplasmic receptors to expand their cellular capacity to bind and transport, via the Venus fly trap fold, a wide range of environmental carboxylated compounds. They might include amino acids (glutamate, aspartate), other forms of C-skeletons useful for growth (oxalate, malate, succinate, lactate, citrate), humic-acid derivatives from soil, metal-complex carboxylated acids in polluted environments, or plant-root-exuded carbohydrates naturally accumulating in the rhizosphere.

Another unusual feature of signal transduction in the *C. metallidurans* CH34 genome is the presence of an intact phosphor- clay two-component system entailing a multidomain histidine- kinase sensor protein BvgS$_1$ (Rmet_5710) and a response-transcriptional activator BvgA (Rmet_5714). This system BvgSA controls, via the transduction of environmental signals, the expression of numerous virulence- and colonization-factors in pathogenic *Bordetella* species [50,151] and resembles, respectively, the EvgSA and RocSA systems in *E. coli* and *P. aeruginosa*. *E. coli* contains at least 57 EvgA-activated genes, including stationary-
phase acid resistance genes and multidrug efflux pump genes [132], while *P. aeruginosa* RocA activates fimbrial adhesin genes for forming biofilms [153]. The BvgA transcriptional activator is well conserved in the four *Cupriavidus* species (63–68% aa identity) and *B. petrii* (38% aa identity), and displays a good similarity to EvgA and RocA, and to the BvgA regulators of pathogenic *Bordetella* species (39–43% aa identity) (all identities with minLrap >0.95). Likewise, the length and sequence of the *C. metallidurans* BvgS1 protein is conserved to EvgS and RocS, as well as to the BvgS sensory proteins of pathogenic *Bordetella* species (29–33% aa identity, minLrap 1.0). By contrast, the BvgS equivalents of *B. petrii* and the three other *Cupriavidus* strains are devoid of the large periplasmic domain spanning nearly 400 residues (Figure 6). Thus, only the BvgSA system of *C. metallidurans* is fully conserved in respect to the *E. coli* EvgSA, *P. aeruginosa* RocSA, and BvgSA systems of pathogenic *Bordetella*. Adding to the complexity of the BvgS structures in *C. metallidurans* and *B. petrii*, both have a second Bvg histidine-kinase sensor. It is called BvgS2 (encoded by Rmet_5724 and Bpet4472) and lacks the periplasmic domain. Also, in both organisms the phosphotransfer domain HPt, which is in pathogenic *Bordetella* part of the BvgS sensor, is an independent protein (Rmet_5713 and Bpet4470, respectively) (Figure 6). Certainly, the region spanning bvgS1 (Rmet_5710) to bvgS2 (Rmet_5724) has undergone multiple insertions and DNA rearrangements, proof of which is a nearby IS1088 copy (Rmet_5718) inserted into a defective serine protease gene (the IS element is indicated with a blue rectangle in the +1 reading frame of Figure 6).

Capsular proteins

Immediately downstream of the bvg gene region (Rmet_5710-24; see above) there is a stretch of genes kpsMTED-kdsAB-kpsFCS putatively involved in synthesizing and transporting capsular polysaccharides (genes Rmet_5729 through Rmet_5737) (Figure 6). The kdsA and kdsB genes (Rmet_5733-34) located between these kps loci encode enzymes for the biosynthesis of 2-keto-3-deoxyoctulosonic acid (KDO), a common component of (lipo) polysaccharides. A similar gene cassette is present in the *C. taiwanensis* genome as a cluster of nine genes (RALTA_B1091 through RALTA_B1100). Such gene casettes appear to be

Figure 6. The *bvgSA* and *kps* loci of *C. metallidurans* CH34. Genomic region on the CHR2 replicon of *C. metallidurans* CH34 covering the *bvgSA* locus encoding a master virulence regulon (Rmet_5710, _5714, _5724) and the kps locus for synthesis and transport of capsular polysaccharides (Rmet_5729 to _5737). Relevant synteny maps are given for 11 organisms. Abbreviations: Cpin, *C. pinatubonensis* JMP134; Ceut, *C. eutrophus* H16; Cta, *C. taiwanensis*; Bpet, *B. petrii* DSM 12804; Bbro, *Bordetella bronchiseptica* RB50; Paer, *Pseudomonas aeruginosa* PA01; Bpar, *B. parapertussis* 12822; Bper, *B. pertussis* Tohama I; Ecol-A, avian pathogenic *Escherichia coli* (APEC01); Ecol-U, uripathogenic *E. coli* (UTI89); Ecol-K, *E. coli* K12. A nearby IS1088 element (Rmet_5718) is indicated with a blue rectangle in the +1 reading frame (see text).

doi:10.1371/journal.pone.0010433.g006
organised in three regions in analogy with the group II/III capsular polysaccharide gene clusters of K-type Escherichia coli [154,155] in which a serotype-specific region (SSR) is flanked by conserved kps genes encoding a transport machinery for bringing polysaccharides to the cell surface. This entails a heterodimeric ABC transporter KpsMT, a multimeric KpsD outer membrane porter, and a periplasmic connector KpsE [156,157]. The function of the other kps proteins KpsF, KpsG, and KpsS is less clear but they were implicated in lipid metabolism, translocation and surface expression of capsule polymers, and determining chain length. We point out that the C. metallidurans kpsF gene (Rmet_5735), encoding a D-arabinose 5-phosphate isomerase, has a duplicate copy on CHR1 as gene Rmet_0307. Likewise, kdcA and kdcB have duplicate copies on CHR1, respectively, as Rmet_1053 and Rmet_0533. The genomes of C. autotrophicus H16 and C. pinatubonensis JMP134 do not carry a kps-SSR-kps cassette. In fact, of all the species discussed in this paper, only pathogenic Bordetella species and B. petrii contain a similar cassette kpsMT-kpsE/wechD-SSR-kpsCS/wechAO although they all seem to lack the kpsD gene encoding the outer membrane porter (in these species it may have been replaced by an unknown alternative protein). Interestingly, the plant symbiont Sinorhizobium meliloti produces a K-antigen that is exported by a typical kpsMT-SSR locus module. This module is encoded in both S. meliloti strains RM1021 and 2101 by two separate klp loci of which one is sited on the chromosome and the other one on the symbiotic megaplasmid [158,159] (not shown).

Although there is no direct evidence for an K-antigen polysaccharide (KPS) in C. metallidurans CH34 or C. taiwanensis, such a Kps ‘secretin’ complex may be used to secrete other types of exopolysaccharides, lipo-oligosaccharides, or short-signal-containing molecules. In this context, the presence of a full set of pelABCDDEF genes (Rmet_4155 through Rmet_4561) in the C. metallidurans CH34 captured our interest. This pel operon first was described in P. aeruginosa PA14 as producing a glucose-rich biofilm matrix [160]; it later was identified in P. aeruginosa PAK and R. solanacearum GMI1000 [161]. Comparative analysis of related genomes via MaGe synteny maps (data not shown) failed to detect a pel locus in C. pinatubonensis JMP134; however, both C. autotrophicus H16 and C. taiwanensis carry a full set of seven pel genes. Other genomes listed in the MaGe database with a complete set are those of three other R. solanacearum strains (Mokk2, Ip01609, and UW551), R. picketti 12D, B. xenovorans LB400, Pseudomonas fluorescens Pf-5, and D. acidovorans. The pel operon encodes sugar-processing enzymes, such as oligogalacturonide lyase (PelA), two glycosyltransferases (PelG and PelF), and a succrose synthase (PelE), along with proteins involved in cellular signalling or transmembrane transport (i.e., PelD, a cyclic-di-GMP signal receptor and PelG, a putative ‘flippase’). Implication of the pel locus in biofilm formation was apparent by a variety of pel mutants of P. aeruginosa strains and R. solanacearum [160–162]. Immediately upstream of the C. metallidurans CH34 pel locus we identified three transposase (tnpA) genes belonging to different families of insertion elements namely Rmet_4154 (ISRme7), Rmet_4153 (ISRme1), and Rmet_4152 (ISRme2) indicating a hotspot for insertion.

In searching for additional genes that might be involved in capsular protein synthesis and transport, we found two more regions of interest. The first is a large cluster of 20 genes (Rmet_2711 through Rmet_2731 situated on CHR1) clearly involved in lipopolysaccharide biosynthesis although the exact role of most proteins remains enigmatic. Based on the patched synteny with other genomes, seemingly this region, which is highly conserved compared to C. taiwanensis, is of a lesser extent to the other Cupriavidus and Ralstonia species, and very intermitantly with the genomes of any other of the species in the MaGe database, displays an amalgam of genes encoding secretion proteins and sugar-processing enzymes, possibly of different origins, that may or may not act in concert. A second cluster of nearly 30 genes (Rmet_5832 through Rmet_5860 situated in CHR2) exhibits a similar mosaic structure with many genes being hypothetical or conserved hypothetical while others generally seem to be involved in transport, signal transduction, or sugar processing activities. Some genes are worth mentioning, for instance Rmet_3833 and Rmet_5843, encoding proteins that respectively display a fair-to-good aa identity with the P. aeruginosa PdA (32%) and PeB (43%) proteins. PeB proteins in P. aeruginosa are implicated in the synthesis of a mannos-rich component of the biofilm matrix [162]. Just downstream of these genes, on the opposite strand of CHR2, lies a small cluster of three genes pdsABP (Rmet_5851-5850-5849) encoding proteins involved in exporting exopolysaccharides; this locus is, on the same strand, immediately preceded by four genes encoding glycosyltransferases (Rmet_5853 through Rmet_5856), and two more genes encoding sugar-processing enzymes mannose epimerase (Rmet_5859) and mannose dehydratase (Rmet_5860). We believe that these two large regions (in addition to the better defined pel locus) deserve special attention in regard to the kps-SSR-kps case or alternative transport systems encoded by the C. metallidurans CH34 genome, particularly in view of the organism’s association with biofilms [14,39], and the role of exopolysaccharides in bacterial complexation of metals [163].

**Motility, chemotaxis, biofilms, and Q-sensing**

C. metallidurans displays flagella uniformly distributed over the cell surface [34] allowing it to swim and swarm. Most of the flh, fli, and fli genes necessary for synthesising flagella are grouped within four gene clusters: A flh locus Rmet_3698-3703, two flh loci Rmet_5252-66 and Rmet_5298-5303, and the fli locus represented by Rmet_3731-44. In addition, the CH34 genome encodes cpa, tad, and pel loci essential for synthesising type IV pilis (Tfp), allowing cells to move across solid surfaces via a mode of twitching. These include the pilIIVX genes (Rmet_0189-91), various cpa and tad genes within the region Rmet_0639 through Rmet_0697, pilTU genes (Rmet_2935/6), the pilBCD genes (Rmet_3108-10), the pilMNOPQ genes (Rmet_3268-72), and the cpaBCD-tadABC genes (Rmet_3653-59). The CH34 genome also contains many genes required for the biogenesis of fimbriae important for cell adhesion, colonization, and biofilm formation, including the finAChhE genes Rmet_0574/5/6, FinA-type adhesin genes Rmet_1667, Rmet_4250, and Rmet_4961, and a fin locus Rmet_4250-64. Our genome survey shows that more than 150 genes may be involved in generating or controlling flagella, pili, and fimbria. Most occur within the above loci, or are situated within a large cluster of 18 chemotaxis (che), motility (mot) and flagellar (flh, fli) genes (Rmet_3678 through Rmet_3695); others are part of smaller sets of genes encoding sensory- or regulatory-proteins.

One major mechanism of regulating motility in strain CH34 is through the action of a Quorum Sensing (QS) system encompassing the phcBSRQ operon (Rmet_2949-52) and the phcA gene (Rmet_2977) [139]. This QS or cell-to-cell communication system was first discovered in R. solanacearum and is neither related to the classical LuxIR system utilizing acyl-homoserine lactone (AHL) autoinducers nor the LuxS system with autoinducer-2. In R. solanacearum, PhcB directs the synthesis of the diffusible signal molecule 3-OH-palmitic acid methyl ester (3-OH-PAME). Typically, with higher cell densities also 3-OH-PAME levels increase to an extent that it saturates the membrane-bound sensor PhcS, effectively lifting the PhcR-mediated repression of the PhcA regulator. All five phe genes of this system are conserved in CH34,
the other three Cupriavidus species, and R. jetkii. However, while 3-OH PAME could be extracted and purified from R. solanacearum, this was not the case for CH34, indicating that 3-OH PAME-like activity in CH34 may result from a structural variant of this molecule [139]. In the plant-pathogenic bacterium R. solanacearum activation of the virulence regulator PhcA is followed by an enhanced production of exoenzymes and extracellular polysaccharides leading to severely decreased motility, presumably in preparation for host colonization [164]. This is not the case in strain CH34 where PhcA activation actually stimulates motility [139]. Likewise, the inactivation of phcA in R. solanacearum increased siderophore production tenfold [135] but not in C. metallidurans CH34 where no affect on siderophore production was seen [139]. Possibly, the Phc regulatory system may connect to different sets of genes in C. metallidurans compared to R. solanacearum; this is not unduly surprising bearing in mind the fundamental differences between their ecology and physiology. Interestingly, CH34 does possess an AHL acylase (Rmet_4680), which enables CH34 to modulate AHL levels and thereby AHL-dependent regulatory systems of other members in a multispecies microbial community.

**Heterotrophic carbon metabolism**

C. metallidurans CH34 cannot grow on glucose, fructose, or galactose [9] because it lacks adequate glucose uptake systems and does not possess a functional 6-phosphofructokinase. Consequently, energy is not produced through the Embden-Meyerhof-Parnas pathway, but by assimilating acidic sugar-derivatives such as D-gluconate, acetate, pyruvate, lactate, L-malate, benzoate, 4-hydroxybenzoate, and various dicarboxylic acids (C4 to C11) [9]. CH34 can derive glucose-6-phosphate from gluconeogenesis (see below) or via the phosphorylation of cytoplasmic glucose by CH34 cannot grow on glucose, fructose, or galactose [9] because it lacks adequate glucose uptake systems and does not possess a functional 6-phosphofructokinase. Consequently, energy is not produced through the Embden-Meyerhof-Parnas pathway, but by assimilating acidic sugar-derivatives such as D-gluconate, acetate, pyruvate, lactate, L-malate, benzoate, 4-hydroxybenzoate, and various dicarboxylic acids (C4 to C11) [9]. CH34 can derive glucose-6-phosphate from gluconeogenesis (see below) or via the phosphorylation of cytoplasmic glucose by glucose 6-phosphate isomerase (GPI; Rmet_0875, Rmet_1507, Rmet_1511) and finally into fructose-6-phosphate by fructose-bisphosphatase (FBP; Rmet_0420), phosphoglycerate kinase (PGK; Rmet_0501, Rmet_1176), phosphoglycerate mutase (PGM; Rmet_0251, Rmet_1518), glyceraldehyde-3-phosphate dehydrogenase (GAP; Rmet_0196, Rmet_1492, Rmet_1518) or via pyruvate (Rmet_0878, Rmet_1453) to phosphoenolpyruvate that in turn is processed to fructose-1,6-biphosphate through a cascade of enzymatic actions by, for example, phosphopyruvate hydratase (ENO; Rmet_1055, Rmet_1176), phosphoglycerate mutase (PGM; Rmet_0251, Rmet_0420), phosphoglycerate kinase (PGK; Rmet_0501, Rmet_1516), glyceraldehyde-3-phosphate dehydrogenase (GAP; Rmet_1096, Rmet_1515, Rmet_2979), and fructose-1,6-biphosphate aldolase (FBA; Rmet_0503, Rmet_1492, Rmet_1518). The resulting fructose-1,6-biphosphate is further converted to fructose-6-phosphate by fructose-biphosphatase (FBP; Rmet_0875, Rmet_1507, Rmet_1511) and finally into β-D-glucose-6-phosphate by glucose 6-phosphate isomerase (GPI; Rmet_1879). Without exception, CHR1 encompasses all the genes required for gluconeogenesis, some of which also are involved in the reductive pentose-phosphate pathway (Calvin-Benson-Bassham cycle).

Under aerobic conditions, C. metallidurans CH34 potentially can degrade benzoate and 4-hydroxybenzoate to ketoacidate via catechol and protocatechuate, respectively, as it has all the genes for both ortho-cleavage pathways (e.g., the benABCDM and catABCDM genes Rmet_4875 through Rmet_4885, pcaA gene Rmet_4270, pcaBDCDHGRQ genes Rmet_4011 through Rmet_4016, and pohAR genes Rmet_4017/8). The ketoacidate thus formed then is processed to acetyl-CoA by the products of the pcaDF genes (Rmet_3667/8/9). In strain CH34, benzoate might be degraded anaerobically via the central metabolite benzyol-CoA (instead of catechol; see Harwood et al. [166]) in an alternate pathway requiring the products of Rmet_1220 through Rmet_1225. Therein, the benzyol-CoA reductase BoxAB (Rmet_1220/1) will open the benzene ring, followed by β-oxidation of the resulting product (oxidation occurs at the β-carbon C3) leading to 3-ketoacyl-CoA, and finally to succinyl-

---

**Note:** The text above is a natural representation of the document's content as if it were being read naturally, without hallucinations. It may require context from other parts of the document to understand the full implications of the scientific statements and findings discussed. The full scientific context and implications are not explicitly provided here.
CoA and acetyl-CoA. Five more genes of this operon (Rmet_1226 through Rmet_1230) probably encode an ABC-like transport system of benzoate.

Similarly, β-oxidation may play an important role in the formation of acetyl-CoA from dicarboxylated compounds because reportedly, _C. metallidurans_ CH34 grows on a range of dicarboxylated compounds [9]. However, without experimental evidence, it is difficult to predict the genes and enzymes that would be involved in these conversions in CH34. An interesting report by Harrison and Woodward [167] implicated the _pinFAB/CDE_ operon for the β-oxidative degradation of C4-C14 dicarboxylic- and fatty-acids in _Rhodopseudomonas palustris_. This operon encodes a set of acyl-CoA ligase, enoyl-CoA hydratase-, acyl-CoA dehydrogenase-, and acyl-CoA transferase enzymes, and is accompanied by a IeIR-type regulatory gene, and several genes for ABC-like transport. Our search through the MaGe annotated data of CH34 yielded more than a hundred of these enzymes. Though most of them might degrade fatty acids, we believe that in CH34 many others are recruited to specifically degrade dicarboxylates.

### Aerobic- and anaerobic-respiration

The citric acid cycle is followed by oxidative phosphorylation in which NADH, FADH2, and GTP are reduced by cytochromes and other terminal oxidases in the cell's inner membrane, while protons are expelled into the periplasm. The resulting proton gradient across the membrane then drives ATP synthesis by the conversion of ADP (ATPase; α, β, γ, δ, ε, a, b, and c units encoded by Rmet_3493 through Rmet_3500). A key locus in this aerobic energy metabolism is a 14-gene cluster located on CHR1 (Rmet_0927 through Rmet_0940) encoding a large membrane-bound complex of NADH dehydrogenase (NADH:quinone oxidoreductase) that is highly conserved within the four sequenced _Cupriavidus_ strains, and in other closely related species. According to BioCyc pathway analysis, alternative NADH dehydrogenases may exist in _C. metallidurans_. These are encoded by Rmet_2623 as a single-subunit FAD-binding NDHase, by Rmet_0177 as a NAD-dependent epimerase/dehydratase, and by Rmet_4310 as a FAD-dependent pyridine nucleotide-disulphide oxidoreductase. We also discovered the Rmet_4300 gene that apparently codes for a NuoF homologue (NADH ubiquinone oxidoreductase, chain F). Other membrane-associated protein complexes that may be important to the electron transport chain in strain CH34 are succinate:quinone reductase (SQOR; see previous section on the TCA cycle), a _bb3-type_ cytochrome complex (encoded by Rmet_3228/9/30), a _cbb3-type_ cytochrome _c_ oxidase (encoded by _cco_ genes Rmet_2041/2/3/4), two cytochrome _bc_ type quinol oxidases CydAB (encoded by the gene pairs Rmet_4955/6 and Rmet_5322/3), a _aa3-type_ oxidase encoded by the _cto_-gene cluster (Rmet_0261 through Rmet_0271), a _bb3-type_ cytochrome oxidase encoded by the _oxy_-gene cluster (Rmet_5955/6/7/8/9), and three quinol oxidases encoded by the clusters _cyt_ _A_ (Rmet_0948/9/50/51), _cyt_ _B_ (Rmet_5608/9/10/11), and _cyt_ _C_ (Rmet_5790/1/2/3).

Under anoxic conditions, _C. metallidurans_ CH34 may resort to alternate modes of energy generation by using electron acceptors other than oxygen. Thus, the bacterium reduces nitrate [32], which is reflected by three loci encoding distinct types of nitrate reductases: a cytoplasmic assimilatory enzyme (NAS; encoded by the _nasaAB_ genes Rmet_4820/1), a membrane-bound quinol-oxidizing enzyme (NAR; encoded by the _nartGH_ genes Rmet_2074 through Rmet_2071), accompanied by two ABC-type transporter-genes Rmet_2072/3, and a periplasmic enzyme (NAP; encoded by the _napFEDABC_ genes Rmet_4061 through Rmet_4063). Probably, the NAS-, NAP-, and NAR-systems have different physiological roles depending on metabolic conditions.

Consequently, their pathways may be interconnected to foster rapid adaptation to changing nitrogen and/or oxygen conditions, a big asset for survival in nature.

Although strain CH34 reportedly does not reduce nitrite [32], it contains two genes Rmet_4018/9 encoding the large- and small-subunits of a putative assimilatory nitrite-reductase (ANR). These genes, preceded by a nitrate transporter gene (Rmet_4017), lie immediately upstream of the NAS genes (Rmet_4820/1). Theoretically, full denitrification by the successive reduction of nitrate to molecular nitrogen seems possible in CH34 because of three key genes. They are the _nis_ gene (Rmet_3172) encoding a NO-forming nitrite-reductase (NNR), the _nor_ gene (Rmet_3167) encoding a nitric-oxide reductase (NOR), and a _nos_ gene (Rmet_4917) coding for a nitrous-oxide reductase (NOS). These three are centered among additional genes that might be required for functional denitrification: (i) The _nisEN_ genes (Rmet_3168/9/70) and _nirCFGH_ genes (Rmet_2079/30/81, RALME2203), (ii) the _nor_ gene (Rmet_5166) encoding a _s_ _4-dependent_ transcriptional regulator, and, (iii) the _nos_ accessory genes _nosLTFDDR_ (Rmet_4912/3/4/5/6) and _nosCX_ (Rmet_4918/9). A regulatory two-component system NarX-NarL that may be important for nitrate- or nitrite-controlled transcription is encoded by Rmet_2086/7.

In view of the ability of strain CH34 to metabolize anaerobically and its potential for full denitrification, we note that it has all the genes for the three known classes of ribonucleotide reductase (RNase): _NrdAB_ (Rmet_3087/8), _NrdJ_ (Rmet_2130), and _NrdDG_ (Rmet_5137, Rmet_5139), along with _NrdR_ (Rmet_2601) as a transcriptional repressor of these genes. According to the RNase database [http://rndrb.molbio.su.se/] less than 8% of all bacteria have all three classes of RNase _C. eutrophus_ H16 and _C. metallidurans_ CH34 are among the bacteria with these genes, but _C. pinutubonensis_ JMP134 and _C. taiwanensis_ lack the _nrdDG_ genes. Our identification of _nrdDG_ genes in strain CH34 is important because the _NrdDG_ enzyme, which operates under anaerobic conditions, is essential for complete denitrification in _C. eutrophus_ H16 [168]. As the latter's _nrdDG_ genes lie on the _phg_ _H_ plasmid, a _phg_ _H_ -free mutant strain grows very poorly on nitrate and does not achieve full denitrification. This defect in the mutant is countered by adding cobalamin (VitB12) to the medium [169], most likely because the coenzyme _B12-dependent_ ClassII RNase, _NrdJ_, assumes the role of _NrdDG_, while at the same time _C. eutrophus_ H16 readily takes up and assimilates cobalamin. Likewise, _C. metallidurans_ CH34 encodes most, if not all genes for the uptake and assimilation of cobalamin (see next section). Together, these facts strongly indicate that anaerobic metabolism, with nitrate as the terminal electron acceptor, runs through very similar pathways in strains H16 and CH34.

### Vitamin biosynthesis

_C. metallidurans_ CH34 has multiple genes for assimilating a variety of B family vitamins, or their _de novo_ production, including thiamin (VitB1), riboflavin (VitB2), pantothenic acid (VitB5), pyridoxal 5'-phosphate (VitB6), biotin (VitB7), folic acid (VitB9), and cobalamin (VitB12). These vitamins, important cofactors for synthesizing fatty acids, proteins, carbohydrates, and branched-chain amino acids, aid in electron-rely systems and energy production. In our genomic survey for vitamin-related genes we did not find any that are relevant to the biosynthetic pathways of retinol (VitA), niacin (VitB3), ascorbic acid (VitC), calciferol (VitD), tocopherol (VitE), or phylloquinone (VitK).

Thiamin (VitB1) biosynthesis is likely governed in _C. metallidurans_ CH34 by the _thiCOGESD_ locus (Rmet_0162/6, Rmet_0169). This set of genes is needed to generate the thiazole- and pyrimidine-
moieties of the vitamin as well as to control the condensation reaction of both moieties to form thiamin phosphate. Additional genes involved in thiamin biosynthesis are *dks* (*Rmet_2615*), *icsS* (*Rmet_1025*), and *bluC* (*Rmet_3047*). *bluC* encoding an oxidase and catalytic enzymes. Two more kinases, ThiF and ThiK, usually are involved in thiamin synthesis but we identified only three putative *thiF* genes (*Rmet_0249* and *Rmet_2639*, both located on CHR1, and *RALMEp10091*, present on the pMOL26 plasmid); we did not recognize the *thik* gene.

The biosynthesis of foliobolin (VitB2) in CH34 might be linked to the genes *ribDC* (*Rmet_2688/9*), *rib(BA)E* (*Rmet_2693/4*), *ribAI* (*Rmet_0194*), *ribF* (*Rmet_2834*), *ribA2* (*Rmet_4572*), and *ribB* (*Rmet_4575*). No other *rib* genes were detected.

Besides its involvement in energy metabolism, pantothenic acid (VitB5) is regarded as a crucial component of coenzyme A biosynthesis, starting from β-alanine and L-valine. In *C. metallidurans* CH34, the synthesis of VitB5 could be connected to genes *ileE* (*Rmet_0493*), *panB* (*Rmet_2774*), *panC* (*Rmet_2917*), and two copies of *panE* (*Rmet_4536*, *Rmet_5770*).

The biosynthesis of pyridoxal 5'-phosphate (VitB6) in this strain proceeds via two complementary pathways. The biosynthetic pathway involves the genes *pdxC* (serC; *Rmet_0715*), *pdxC* (present in two copies *Rmet_0438*, *Rmet_2590*), a gene encoding an FMN-binding protein (*Rmet_0820*), *pdxJ* (*Rmet_2415*), *dxx* (*Rmet_2615*; also involved in VitB1 synthesis, see above), *pdxH* (*Rmet_2642*), and a putative pyridoxine oxidase gene (*Rmet_4050*). We did not detect the genes involved in generating 2-oxo-3-hydroxy-4-phosphobutanoate, an early step in VitB6 synthesis and the actual substrate of the SerC catalytic action. The second pathway, the salvage pathway, basically converts pyridoxine, pyridoxamine, or pyridoxal, into pyridoxal-5'-phosphate, the active form of VitB6. This involves only two genes *pdxH* (*Rmet_2642*) and *pdxK* (*Rmet_4114*). We note that the former also is part of the de novo biosynthetic pathway (above).

The *bioAFDB* locus (*Rmet_0114/5/6/7*) in CH34 controls the production of biotin (VitB7). There are slight variations in the pathway, depending on whether pimelate or pimelyl-CoA is the starting substrate; nonetheless, the gene products of all four genes are involved in both. Pimeloyl-CoA may be derived from the conversion of aromatic compounds via the intermediates benzoyl-CoA ligase (often referred to as a ‘pimeloyl-CoA synthetase’), eventually synthesising pimeloyl-CoA from pimelate and coenzyme A.

The variant pathways and intermediate reactions leading to the formation of folate (VitB9) from chorismate or GTP are complex and behind the scope of our genome study. We identified several genes that might be involved in folate biosynthesis in *C. metallidurans* CH34; they are scattered throughout the genome and linkage between individual ones is blurred. These genes are *folB* (*Rmet_0183*), *folD* (*Rmet_1192*), *purT* (*Rmet_1829*), *folP* (*Rmet_2187*), *folC* (*Rmet_2463*), *folA* (*Rmet_2569*), *pum* (*Rmet_2687*), *folE* (*Rmet_2915*), *topG* (*Rmet_3179*), and a putative *folE* (*Rmet_3990*). We failed to identify a hydroxase gene for converting 7,8-dihydropterin triphosphate into 7,8-dihydro-D-neopterin.

There have been extensive studies on cobalamin, a vitamin of a particular interest in public health and medical sciences, and produced only by microorganisms [170]; indeed, its biosynthesis is one of the best known vitamin pathways to date [171]. Two pathways for corrin ring formation were described in bacteria, differing from each other in the way cobalt is incorporated [172].

In *C. metallidurans* CH34, the presence of a sirohdrochlorin cobaltotchelatase gene *cbhX* (*Rmet_2010*) suggests that this organism synthesises cobalamin anaerobically, thereby including an early insertion step of cobalt. Genes involved in the late stage of the biosynthesis are grouped in a large operon of 13 genes (entailing *cbh*, *cob* and *btu* genes *Rmet_2777* through *Rmet_2797*) with a cobalamin riboswitch we identified at the far end (*Rmet_R0047*). This entire operon, which according to the MaGe interface displays good synerhy with all *Cupriavidus* and *Ralstonia* genomes, contains all the essential genes for producing adenosylcobalamin from cobyric acid (including those controlling corrinoid transport, viz. *btuA*, *btuB*, *btuC*, *btuD* and *btuF*). Only the *btuB* gene (*Rmet_2213*), coding for a cobalt reductase, is at a different location on CHR1. Whilst the earlier and late stages of the biosynthesis generally are covered by these listed genes, other genes involved in the intermediate corrin biosynthesis in *C. metallidurans* CH34 await elucidation.

### Autotrophic metabolism

In the absence of heterotrophic carbon sources, *C. metallidurans* CH34 can switch to an autotrophic lifestyle by fixing and reducing carbon dioxide for synthesizing organic molecules. Generally, autotrophic prokaryotes have four different biochemical routes to accomplish this (reviewed by Atomi [173]): (i) The reductive pentose-phosphate pathway; (ii) the reductive acetyl-CoA pathway; (iii) the 3-hydroxypropionate cycle; and, (iv) the reductive tricarboxylic-acid (TCA) cycle.

The *C. metallidurans* CH34 genome encodes all enzymes necessary for the 13 biochemical reactions of the reductive pentose phosphate pathway, also known as the Calvin-Benson-Bassham (CBB) cycle. Figure S6 is a schematic view of the *cbb* loci in CH34, using Tabita et al.’s gene notations [174]. We note especially that this region contains some IS elements (marked in red by their *tnpA* transposase genes in Figure S6), including two exterior IS1071 elements, *Rmet_1491* and *Rmet_1544*. This may explain the earlier observed loss of autotrophic growth in CH34 under mutagenic stress [22] since the IS1071-mediated excision of this region would remove, amongst other genes, the two structural genes *cbIL* (*Rmet_1501*) and *cbBH* (*Rmet_1500*) of RubisCo (ribulose-1,5-bisphosphate carboxylase oxygenase) as well as of the accessory genes *cbBQ* (*Rmet_1499*) and *cbBO* (*Rmet_1498*) needed to assemble RubisCo.

Several regulators control the CBB cycle. First, there are three copies of *cbBR*, a LysR-type transcriptional activator of *cbb* gene expression. Two copies juxtapose the central *cbb* gene clusters, *cbBR1* (*Rmet_1502*) and *cbBR2* (*Rmet_1508*), while a third copy, *cbBR3* (*Rmet_4365*), is located on the second replicon CHR2 (unlike other *cbb* genes that are all sited on CHR1). Furthermore, the *C. metallidurans* CH34 genome harbours a set of *regA* (*Rmet_0135*) and *regB* (*Rmet_0136*) genes probably regulating carbon fixation via the CBB-cycle, as in other organisms [173,175]. Remarkably, three non-essential *cbb*-genes appear to be transcriptionally linked to the *cbb* operon: *cbbY* (*Rmet_1517*), *cbbZ1* (*Rmet_1514*), and *cbbZ2* (*Rmet_3177*). Under certain conditions, autotrophic bacteria carrying these exceptional *cbb* genes display a phosphoglycolate phosphatase (PGP) enzyme activity [176] that may help cells to cope with toxic levels of phosphoglycolate accumulated by the oxygenase reaction of RubisCo whilst the carbon skeleton of phosphoglycolate is salvaged as glycolate for further use in glycine synthesis. We did not detect a *cbbX* gene in the *C. metallidurans* CH34 genome.

Besides the CBB cycle, *C. metallidurans* CH34 apparently has functional genes to fix carbon dioxide via the reductive acetyl...
coenzyme A pathway (also termed the Wood-Ljungdahl pathway). One key biochemical reaction in this pathway is the transition between carbon dioxide and carbon monoxide mediated by a metalloenzyme, CO dehydrogenase (CODH). This enzyme is a dimeric heterotrimer consisting of a small, medium, and large subunit representing respectively, an iron-sulfur protein (CosX), a flavoprotein (CosM), and a molybdoprotein (CosL). Two forms of aerobic CO-DH are known: Form I was characterized explicitly for its ability to oxidize carbon monoxide, but the function of form II CODH, distinct from form I but phylogenetically closely related, remains unclear [177]. Strain CH34 carries the genes cosSLM (Rmet_0365 – Rmet_0363) and the accessory genes cosDEGI (Rmet_0362 – Rmet_0359). The order of the CODHI genes in this operonic structure and the presence of an AYRGAGR motif in CosL (encoded by Rmet_0364) strongly suggest that the C. metallidurans' CODHI is of form II. Although the ability of C. metallidurans CH34 to undertake substantial oxidation of carbon monoxide still needs to be tested, CODHI-II-encoding cos genes were found in Bradyrhizobium japonicum USDA 110, an organism relying solely on CODD-II for CO oxidation and hence, growth on carbon monoxide as a sole carbon- and energy-source [178].

Based on our genomic survey and supported by KEGG analysis, C. metallidurans CH34 lacks the cycle-specific malonyl-CoA reductase (EC:4.1.1.9), an essential enzyme for the 3-hydroxypropionate cycle [179,180], and hence, cannot employ this pathway to assimilate carbon dioxide. Likewise, the C. metallidurans CH34 genome does not code for the key enzymes of a rTCA cycle, i.e., ATP-citrate lyase and 2-oxoglutarate synthase nor does it encode enzymes for alternative formation of oxaloacetate, for example, via the combined action of citryl-CoA synthetase and citryl-CoA hydrolase, thus effectively ruling out the rTCA-cycle as a route for fixing carbon dioxide.

From this genomic information it appears that C. metallidurans CH34 relies primarily on a RuBiScO Type I enzyme for fixing carbon dioxide as part of a complete CBB-cycle. Although an intact CODDII-based system is present, it is yet unclear whether strain CH34 can use CO as a sole energy and carbon source.

Sulfur oxidation (sulfur autotrophy)

C. metallidurans CH34 has the genetic capacity to oxidize inorganic sulfur compounds by the action of sulfur oxidizing genes (sox) and sulfur oxygene oxidase reductase genes (soi) [Figure 57A] [181]. The first set of genes comprises the soxBWXAZYFE gene cluster situated on CHR1 as Rmet_3417 through Rmet_3427. Basically, sulfur respiration requires only five sox genes whose products form three key periplasmic protein complexes: SoxY, a sulfur carrier protein, SoxXA, a c-type cytochrome complex, and SoxB, a sulfate thiol hydrolase. In this system, the oxidation of thiosulfate (S\(^{2-}\)) to sulfate entails the oxidative linkage, mediated by SoxXA, of thiosulfate to the conserved C-terminal cysteine of SoxY, followed by hydrolysis of the terminal sulfone (SO\(_3^2\)) group by SoxB [Figure 57B]. The SoxXA complex then covalently binds other pathway intermediates to the SoxY cysteine thus forming extended chains of sulfate (S\(^{-}\)) groups preceding the terminal sulfone group (SO\(_3^2\)) [182]. These sulfate chains are broken by the iterated actions of the SoxB hydrolase and a periplasmic sulfur dehydrogenase, SoxCD. Although C. metallidurans CH34 lacks the soxCD genes, this would not impair its ability to oxidise thiosulfate to sulfate in the first step of the cycle. However, fewer electrons would be liberated compared to the electron generation of complete sox systems.

In C. metallidurans CH34, the likely route for sulfide (HS\(^{-}\)) oxidation would be via the soxE and soxF gene products, so forming a typical c-type flavocytochrome complex in which SoxF functions as a sulfide dehydrogenase [Figure 57C]. Lastly, the C. metallidurans sox gene cluster includes two genes of unknown function coding for a conserved hypothetical protein (Rmet_3421), putatively exported, and an unknown c-type cytochrome precursor (Rmet_3424).

C. metallidurans probably oxidizes sulfites (SO\(_3^2\)) through the action of the soxD and soxS genes [present on CHR2 as Rmet_4891 and Rmet_4892] (Figure 57D) that are orthologues of the C. eutrophus H16 soxD and soxS genes encoding, respectively, a periplasmic sulfite dehydrogenase and a membrane-bound c-type cytochrome [71].

These routes of sulfur oxidation (SoxB-XA-YZ, SoxF, and SoxAB) in combination with an autotrophic lifestyle would enable C. metallidurans to use reduced sulfur compounds as electron donors, while employing carbon dioxide as an electron acceptor, and hence it could function as a true chemolithotrophic sulfur-oxidizing bacterium. Such a sulfur-based growth is sustainable only when sufficient amounts of thiosulfate ions are present. A biological source for them is from the fermentation of taurine (2-aminoethanesulfonate), together with the potential fermentation of other organosulfonates in anoxic environments [183]. Interestingly, C. metallidurans thrives not only on thiosulfate but also on a wide range of aromatic sulfonates, alkanesulfonates, and sulfate esters [184]. Therefore, it is not surprising that the C. metallidurans CH34 genome encodes alkanesulfonate- and taurine-specific ABC transporter systems, such as SauABC (Rmet_1370/2/3) and TauABC (Rmet_2857/8/9), an alkanesulfonate monoxygenase SauD (Rmet_1371), at least six copies of taurine dioxygenase TauD (encoded by Rmet genes _4136, _4409, _4633, _5119, _5292, and _5823), a reductase-ferredoxin AsfA/AsfB (Rmet_4334/5), and a putative NAD(P)H-dependent FMN reductase SauE (Rmet_5075) that with AsfA/AsfB may form an electron chain for aromatic desulphonation. We did not identify other genes for enzymatic desulphonation from the C. metallidurans CH34 genome data.

Hydrogen oxidation (hydrogenotrophy)

C. metallidurans CH34 can gain its energy from oxidising hydrogen gas (H\(_2\)) into electrons (e\(^{-}\)) and protons (H\(^+\)). The underlying enzymatic engine has two parts: A periplasmic membrane-bound hydrogenase (MBH), and a cytosolic soluble hydrogenase (SH). This allows the bacterium to switch between an H\(_2\)-driven energy supply and a classic lithotrophic growth mode. In contrast to the H\(_2\)-oxidising C. eutrophus H16, the expression of MBH and SH in C. metallidurans CH34 is not controlled by an H\(_2\)-sensor in the form of a third hydrogenase. Therefore, the nearby concentration of hydrogen gas does not regulate the lithotrophic growth of CH34. Nonetheless, the lack of a specific H\(_2\)-sensor-hydrogenase does not entail the constitutive expression of the energy-related hydrogenases, such as in the absence of H\(_2\). Rather, both the hydrogenases and the RubisCo enzyme (see autotrophic metabolism) are regulated by the carbon source afforded by the growth substrate [9].

All the genes needed to form and activate MBH and SH are grouped in two large gene clusters situated on CHR1 [Figure 58A]. The first cluster, governing MBH, harbors the hydrogenase kleitoprotic operon encompassing six genes hyp\(_E\)D\(_C\)G\(_F\)B\(_A_1\) (Rmet_1281 – Rmet_1286), followed by a set of genes required for hydrogen oxidation, box\(_V\)Q\(_L\)M\(_Z\)G\(_K\) (Rmet_1287 through Rmet_1293, Rmet_1295, and Rmet_1297-98). This entire cluster is located on genomic island CMGI-2. A second cluster, governing...
the SH, contains one partial hox gene denoted ΔhoxV' (Rmet_1534), nine intact hox genes hoxFUWHX11 (Rmet_1522 through Rmet_1527, Rmet_1533, and Rmet_1541-42), and six genes hyfbA/B/C/D/E (Rmet_1535 through Rmet_1540) all on genomic island CMGI-3. The latter group of hyf genes (index 2) appears to have arisen from a reverse duplication of the hyf operon of the first hyf cluster (index 1) on CMGI-2. We note that the hyf and hox genes for SH are located, together with the chb genes, on a region flanked by IS elements (Figures S6 and S8A) and that the IS-mediated excision of this region would lead to the loss of RubisCo and the entire SH system, thereby impairing not only autotrophic growth but causing the loss of hydrogenotrophic growth as well. Indeed, stress-induced mutants of CH34 lacking this genomic region are unable to utilize hydrogen gas as a source of energy [6].

Besides these hydrogenase gene clusters on CHR1, there are six hydrogenase ‘four’ (hyf) genes grouped together on CHR2 as hyfBCDEFG (Rmet_4666 through Rmet_4671). This resembles an incomplete operon that, in E. coli K12, exists as a 10-gene cluster encoding a multimeric hydrogenase complex. Presumably, this complex interacts with formate dehydrogenase FdhF to form a hydrogenlyase system cleaving formate to dihydrogen and carbon dioxide [185]. The same six hyf genes also occur on the first replicon CHR1 of C. eutrophus H16 as genes H16_A2196 through H16_A2201 [45,169]. However, because of the lack of at least four subunit-encoding genes, and the structural differences in the CH34 and H16 Hyf proteins compared to the E. coli K12 Hyf units, it is very unlikely that the hyf genes in either strain are involved in any aspect of lithotrophic growth.

The C. metallidurans MBH is a typical [Ni-Fe]-hydrogenase [186] composed of three subunits (Figure S8B); HoxG, holding the active hydrogen-splitting Ni-Fe centre; HoxK, an electron-transferring subunit containing two Fe4S8 clusters; and HoxZ, a membrane-bound di-heme cytochrome b. The HoxK signal sequence contains the SRRSFMK sequence recognized by the Twin arginine translocation (Tat) system. Consequently, MBH is exported from the cytoplasm to the periplasmic space, to where its enzymatic activity is limited. The C. metallidurans SH is classified as a cytoplasmic bidirectional or hydrogen-evolving [Ni-Fe]-hydrogenase [186,187] consisting of five subunits HoxHUFK, with HoxY holding the c’ transferring Fe4S8 cluster. While the reducing powers of MBH reside solely under the form of reduced ubiquinone (ubiquinol), the SH can directly reduce NAD+ to NADH (Figure S8B). Because SH acts as a bidirectional enzyme complex, it reverses its activity with increasing NADH to NAD+ ratios, effectively acting as a “redox valve”. Accordingly, SH is an important factor in protecting the electron transfer chain from becoming hyper-reductive. The fact that C. metallidurans CH34 can utilize H2 instantaneously as a reducing power and hence as an alternative source of energy when nutrients are scarce indicates that chemoholotrophic growth is an integral part of the organism’s lifestyle.

We point out that, although C. eutrophus H16 is known as the model ‘knallgas’ bacterium, the MBH of C. metallidurans CH34 appears to be far more efficient than its counterpart in C. eutrophus H16 since it can oxidise H2 gas at concentrations as low as 3% (v/v) and this under atmospheric conditions (i.e., at 20.9% v/v oxygen) [188]. The remarkable features of the C. metallidurans MBH (high catalytic activity, H2 affinity, and O2 tolerance) might be valuable in the emerging field of microbial fuel-cell technology [189,190]. Further, the oxygen tolerance of MBH may account for the ability of C. metallidurans CH34 to grow autotrophically, even in high-oxygen atmospheres [9].

Organooautrophy and degradation of aromatics

C. metallidurans CH34 can utilize various recalcitrant organics as sole source of carbon and energy, including volatile compounds such as 1-propanol, 2-propanol, 1-butanol, 2-butanol, and acetone, as well as the monoaromatic hydrocarbons benzene, toluene, o-xylene, and phenol. The CH34 genome encodes at least 18 alcohol dehydrogenases (ADHs) and at least 10 aldehyde dehydrogenases (ALDHs). This multitude of fermentation enzymes may seem unusual for an organism that typically follows respiratory metabolism. However, as discussed in earlier separate sections, under limiting oxygen and carbon conditions CH34 will degrade short-chain fatty acids, dicarboxylated compounds, or energy-storage products, such as polyhydroxybutyrate (PHB). This may engender elevated levels of acetyl-CoA while simultaneously, fewer terminal electron acceptors are available, thus giving rise to partially oxidized toxic products including alcohols and aldehydes. Hence, the primary role of the multiple ADHs and ALDHs in CH34 would be cellular detoxification. Further accumulation of acetyl-CoA may initiate the condensation of two acetyl-CoA molecules to acetoacetyl-CoA by the action of an acetyl-CoA acetyltransferase (AAT; EC 2.3.1.9). This compound then may be converted by an acetyl-CoA:Acetoacetate CoA transferase (CoAT; EC 2.8.3.8) to acetoacetate, which may have various metabolic fates. However, some bacteria change acetoacetate further into acetone via acetooacetate decarboxylase (ADC; EC 4.1.1.4). The CH34 genome carries the genes for ten AATs very similar in length and sequence (though some may have alternate functions in the cell, for example, the β-oxidation of fatty acids and dicarboxylates, or PHB metabolism) and has two CoAT genes encoded by two independent atoAD loci (Rmet_1153/4 and Rmet_5059/60). However, we detected no ADC in the CH34 sequence data (by BlastP analysis against the ADC sequence of Clostridium acetobutylicum ATCC 824, a model organism for bacterial solventogenesis). Hence, it is doubtful that CH34 produces acetone directly from acetoacetate through decarboxylation.

We were very surprised to learn that the CH34 genome encodes an acetone carboxylase (ACX; EC 6.4.1.6) (encoded by Rmet_4105/6/7), an (αβ2γ2) multimeric enzyme involved in acetone metabolism and identified in only very few bacteria [191,192]. The ACX catalytic reaction basically is the reverse of the ADC reaction in that it carboxylates acetate to generate acetoacetate that is metabolized to acetoacetyl-CoA and further to acetyl-CoA for oxidation in the TCA cycle. The acxABC genes in C. metallidurans CH34 encoding the three subunits of ACX seemingly are regulated by the AcxR transcriptional activator, whose gene (Rmet_4104) immediately precedes acxABC and lays in the opposite orientation, closely resembling the acxRABC operon of Xanthobacter autotrophicus P82 [192]. Both in X. autotrophicus and C. metallidurans the AcxR protein contains a SELFGIXXXGAFGTA motif that is typical for Ζ4 subunit binding regulators [193], and there is a clearly identifiable Ζ4 promoter sequence TGGCATGNNNTTGC in the acxR1 intergenic region. Importantly, proteomic studies demonstrated that the growth of strain CH34 on acetone or 2-propanol as the sole carbon source induces ACX expression (C. Rosier, pers. comm.). A survey of CH34-related genomes using the MaGe interface reveals that the C. pinatubonensis JMP134 megaplasmid displays an acxRABC operon almost identical to the CH34 acx locus in organization and sequence (Ceut_C6170/72/73/74; with aa sequence identities of 89-93% between the respective ACX subunits and 71% identity between the AcxR proteins). None of the other organism listed in Table 2 appears to harbor an acxRABC locus. According to current MaGe data, only Azorhizobium caulinodans ORS 571 carries an
acxRABC locus closely resembling the CH34 acx locus (with 62-71% sequence identity between the respective ACX subunits and 41% sequence identity between the AcxR proteins) albeit that its acxR gene is oriented in the same direction as the acxRABC genes. We also detected acx loci in Mesorhizobium loti MAFF 303099 and Sinorhizobium loti 1021 with 37-40% sequence identity of their ACX subunits to the respective ACX subunits of CH34. However, the M. loti and S. loti loci do not code for an AcxR-type σ54-dependent transcriptional activator, but for a LuxR/Fis type transcriptional regulator.

Bacteria can produce acetone through the dehydrogenation of 2-propanol (iso-propanol) or via the oxidation of 2-nitropropane by a 2-nitropropane dioxygenase (NPD; EC 1.3.11.32). Among the 18 ADH genes we detected in CH34, the most likely candidate for an iso-propanol dehydrogenase (2PDH) is Rmet_5645 because its product is highly similar in length and sequence to the multifunctional ADH of C. eutrophus H16 (H16_A0757) [194]. In addition, the CH34 genome contains six NPD encoding genes (Rmet_0222, Rmet_2083, Rmet_2603, Rmet_3674, Rmet_5434, and Rmet_5534). The production of acetone by CH34 would explain the presence of an ACX because this way cells can utilize it for growth and keep (toxic) acetone levels low.

Benzoate, toluene (methylbenzene), ethylbenzene, and xylene (dimethylbenzene) jointly are known as BTX compounds, with phenol as a major chemical intermediate. In C. metallidurans CH34 aerobic degradation of BTX involves ring hydroxylation by one of three different multicomponent mono-oxygenases and proceeds via typical meta-cleavage pathways initiated via catechol decyclation by a 2,3-dioxygenase (in contrast to 1,2-dioxygenase ortho-cleavage pathways that cleave the catechol ring between the two hydroxyl groups). Bacterial multicomponent mono-oxygenases (BMMs), transcribed from single operons encoding four to six subunits, are classified in six distinct groups [195]. Although the genomes of aromatic-degrading strains only rarely carry more than one BMM [195], Cafafo et al. [196] suggested that multiple BMMs in the same system may be beneficial in terms of BMM specificity and efficiency of substrate utilization. We identified two Group 1 (phenol hydroxylases; PHs) and one Group 2 (toluene-benzene mono-oxygenases; TOMOs) BMMs in two different loci of the CH34 genome. The first locus (Rmet_1305 to Rmet_1331) is part of the CMGI-2 genomic island and contains the genes for two BMMs: a phenol hydroxylase, encoded by six genes Rmet_1326 to Rmet_1331, and a benzene-toluene mono-oxygenase, encoded by six genes Rmet_1311 to Rmet_1316. All genes are present for the meta-cleavage of catechol (by a 2,3-dioxygenase encoded by Rmet_1324), and its subsequent enzymatic processing (by the products of genes Rmet_1317 to Rmet_1322) to the final product acetyl-CoA. The Rmet_1305 gene [phaR] was identified as a σ54-dependent Fis-type transcriptional regulator. This entire region is in high synteny with the aromatic degradation gene clusters in the related genomes of C. pasteurianus JMP134 (Reut_B3671 through Reut_B3693) and Ralstonia pickettii 12D (Rpic_5178 through Rpic_5201). The second locus (Rmet_1781 to Rmet_1788) encodes a second phenol hydroxylase (PH) also consisting of six subunits (Rmet_1782/3/4/5/6/7), with Rmet_1787 encoding the Fis-type regulator PoxR. However, gene Rmet_1781 encodes a 1,2-dioxygenase with high similarity (64% aa sequence identity) to the 1,2-dioxygenase for ortho-cleavage of benzoate (encoded by Rmet_4881; see ‘Heterotrophic carbon metabolism’ in this discussion). This eight-gene cassette, including a 1,2-dioxygenase gene, also occurs in strain JMP134 (as Reut_A1698 to Reut_A1705) and in Wintersia namadzuensis strain TE26 (as the phiRIBCDEF-catA locus; [197]). The very high sequence similarities between the eight corresponding gene products of the CH34, JMP134, and TE26 loci (82–99% sequence identity) point to their acquisition via gene transfer as one cassette.

The combination of a phenol hydroxylase gene cluster with a 1,2-dioxygenase gene seems to be very rare and confined to the above-mentioned strains. As far as we are aware, P. putida strain GB1 is the only other organism with a similar locus (as evidenced by MaGe data) although the 1,2-dioxygenase gene is not located downstream of the hydroxylase genes, but on the other side of the regulatory gene. Also, the sequence similarity between the corresponding hydroxylase subunits is much lower (48–68% sequence identity). Notomista et al. [195] postulated that the modularity of multiple BMMs in the same system may foster the invention of new mixed pathways, or even hybrid BMMs with novel substrate specificities. Here, it might be valuable to explore systematically the aromatic-substrate specificity of C. metallidurans CH34, as was done recently for C. pinatubonensis JMP134 [198]. Only then can we correlate aromatics degradation with pathway predictions and determine the true catabolic potential of CH34.

Polyester synthesis PHA/PHB

Bacteria produce polyhydroxybutyrate (PHB), a polyester polymer belonging to the family of polyhydroxyalkanoates (PHAs), as a carbon- and energy-reserve to be metabolized after exhausting all other carbon resources (reviewed by Lenz and Marchessault [199]). The bacterial synthesis of PHAs, such as PHB, has gathered renewed interest because their physical properties can be remarkably similar to those of the plastic polypropylene, and because, unlike polypropylene, it is rapidly biodegradable (hence the name “bioplastic” or “green plastic”). Although there are various bacterial pathways to create PHAs [200], the key model one presently used is that from C. eutrophus H16 which encompasses only three enzymes: A PHA synthase (or polymerase) encoded by phaC; an acetooacetyl-CoA reductase encoded by phaB; and, a β-ketothiolase encoded by phaA. The synthesis of PHB has been studied most extensively in C. eutrophus H16 because it can generate PHB solely using simple carbon substances, such as glucose, via the Entner-Doudoroff (ED) pathway [45]. Additional genes playing a role in PHB synthesis are phaR encoding a transcriptional regulator, and phaB encoding a phasin acting on the surface of PHA granules. Phasins inhibit the coalescence of the granules to regulate their number, and may take part in mobilizing the storage polymer. Lastly, several phaZ paralogs encode the intracellular depolymerases needed for PHB homeostasis and degradation.

Like its relative C. eutrophus H16, C. metallidurans CH34 synthesizes PHB and probably accumulates it in the form of intracellular storage granules. Interestingly, PHB synthesis in CH34 is initially (i.e., after 24 hours of growth) enhanced under simulated microgravity [41]. The CH34 genome has three copies of phaC, one copy each of phaA and phaB, three copies of phaZ, and one regulatory gene phaR (for Rmet numbers see Table S9). These PHB synthesis genes mainly are organized into two cassettes: phaC1-phaA-phaB1-phaR (Rmet_1356-1359) located on the CHR1, and phaC3-phaB3 (Rmet_5122-5123) located on CHR2. All the other genes are scattered throughout the genome (Table S9). We also found isologs of phaA (bitB, aolB) encoding additional acetooacetyl-CoA thiolases but these were annotated as having a different substrate specificity than PhaA (Table S9). For intracellular PHB depolymerization, we identified three phaZ genes on the C. metallidurans CH34 genome: One on CHR1 and two on CHR2. These phaZ genes correspond with the phaZ1, phaZ2 and phaZ3 genes of C. eutrophus H16 identified by York et al. [201]. In addition, the CHR1 replicon carries a gene for a PHB-oligomer hydroxylase (designated by phaF or Rmet_1960) that is an ortholog of...
the earlier identified 3-hydroxybutyrate-oligomer hydrolase (3HBH) of *C. eutrophus* H16 [202-204]. The 3HBH enzyme in *C. eutrophus* H16 degrades the trimeric products formed by intracellular PHB depolymerases [205]; based on the very high aa similarity (63%), PhaY seems to have the same function in *C. metallidurans*. Lastly, a gene fragment (Rmet_1670) located adjacent to the phoC2 gene (Rmet_1671) is predicted by the MaGe system to encode a defective γ-aminobutyrate transporter, indicating that these genes might be remnants of an obsolete PHB biosynthesis operon.

DNA repair genes and stress

The robustness of *C. metallidurans* CH34 under a wide variety of harsh environmental conditions does not rest just on its efficient transport and detoxification of harmful substances, ions or molecules, but also greatly depends on safeguarding its genetic material from irreversible damage. Generally, genome integrity is maintained by a number of DNA-repair mechanisms, including excision repair, direct repair, and recombinational repair. The CH34 genome contains DNA-repair genes for base-excision repair (alkA, alkT, tag, ung, fpg, nth, xthA) and DNA repair in strain CH34 is mediated by photoreactivation (phr), oxidative demethylation (alkB1, alkB2), and alkyltransferases activity (ada, ogt1, ogt2, ogt3). Its genome does not seem to carry a RecBCD or a RecFOR pathway to repair single-strand gaps and double-strand DNA breaks by homologous recombination. The lack of these genes was previously postulated to allow the transformation in CH34 of large linear ds DNA fragments [206]. Nonetheless, strain CH34 has a functionally analogous system in the form of AddAB that is prevalent among β- and α-proteobacteria and firmicutes [207]. The CH34 genome carries the genes encoding the Holliday-junction protein complex RuvAB, the resolvase RuvC, a dsDNA exonuclease ShbD, and the mismatch-repair proteins MutS and MutL. These genes *ruvAB*, *ruvC*, *shbD*, and *mutSL*, often co-occur within genomes [207]. Interestingly, CH34 does not contain a *mutH* gene encoding a dGATC endonuclease but a dam homologous gene is present (Rmet_5483), located on CHR2. Dam affects the transposition frequency of several transposons and IS elements (Tn10, Tn5, Tn903, IS3, and possibly others), and modulates gene expression by the adenine methylation of DNA regulatory sites [208]. The presence of Dam in *C. metallidurans* CH34 might well be significant in view of the complex transcriptional networks and the large number of mobile genetic elements intrinsic to its flexible, fluid genome.

In addition, CH34 harbors the main genes of the RpoH regulon (heat shock) and stringent response (*recA* and *spoT*), the major regulatory genes of the SOS response (*lexA* and *recA*), and the oxidative-stress-response regulators, OxyR and SoxR (all Rmet numbers are given in Table S10) and the complete pathway for nucleotide-excision repair (*uwrA*, *uwrB*, *uwrC*, *uwrD*). Direct DNA repair in strain CH34 is mediated by photoreactivation (phr), oxidative demethylation (alkB1, alkB2), and alkyltransferases activity (ada, ogt1, ogt2, ogt3). Its genome does not seem to carry a RecBCD nor a RecFOR pathway to repair single-strand gaps and double-strand DNA breaks by homologous recombination. The lack of these genes was previously postulated to allow the transformation in CH34 of large linear ds DNA fragments [206]. Nonetheless, strain CH34 has a functionally analogous system in the form of AddAB that is prevalent among β- and α-proteobacteria and firmicutes [207]. The CH34 genome carries the genes encoding the Holliday-junction protein complex RuvAB, the resolvase RuvC, a dsDNA exonuclease ShbD, and the mismatch-repair proteins MutS and MutL. These genes *ruvAB*, *ruvC*, *shbD*, and *mutSL*, often co-occur within genomes [207]. Interestingly, CH34 does not contain a *mutH* gene encoding a dGATC endonuclease but a dam homologous gene is present (Rmet_5483), located on CHR2. Dam affects the transposition frequency of several transposons and IS elements (Tn10, Tn5, Tn903, IS3, and possibly others), and modulates gene expression by the adenine methylation of DNA regulatory sites [208]. The presence of Dam in *C. metallidurans* CH34 might well be significant in view of the complex transcriptional networks and the large number of mobile genetic elements intrinsic to its flexible, fluid genome.

In addition, CH34 harbors the main genes of the RpoH regulon (heat shock) and stringent response (*recA* and *spoT*), the major regulatory genes of the SOS response (*lexA* and *recA*), and the oxidative-stress-response regulators, OxyR and SoxR (all Rmet numbers are given in Table S10). *C. metallidurans* CH34, however, lacks a SoxS homolog and does not carry any typical SoxS binding site, as indicated by a genome-scale pattern search using the SoxR binding site consensus sequence [209] and the Regulatory Sequence Analysis Tools webservice (RSAT) [210] (results not shown). Therefore, the involvement of SoxS-mediated regulation of the oxidative-stress response is unresolved. Finally, CH34 harbors 15 proteins belonging to the universal stress protein (UspA) superfamily (Table S11). Generally, UspA proteins are induced in response to a variety of starvation and stress conditions and have partially overlapping but distinct biological functions [211]. Eleven genes encode small proteins harboring one Usp domain, two encode intermediate proteins with one Usp domain (*uspA13*, *uspA14*), and two encode large proteins consisting of two Usp domains (*uspA6*, *uspA11*). The *uspA11* gene is present in 2 copies and located on transposon *Tn6050*. All but five (uspA3, *uspA4*, *uspA13*, *uspA14*, *uspA15*) contain the G-2X-G-9X-[S/T] motif that is shared by many members of the UspA family [212].

Conclusion

We have embarked on an exciting journey through the genome of a versatile soil bacterium, *C. metallidurans* strain CH34, best known for its tolerance to a wide range of heavy metals and representative for a species that is mainly found in areas linked to mining sites and metallurgical or chemical industries. Genome data show that this genome underwent numerous genetic adaptations in its evolutionary history by continuously gathering novel genes and pathways, and by rearranging and combining them in new clusters useful for its survival. This continuous acquisition of genes, and clusters of genes, has led to the co-occurrence of many complementary systems either involved in heavy-metal resistance or other types of stress, requiring intricate fine-tuning of cellular responses on many different levels. Accordingly, its complex genome boasts a vast network of regulatory proteins, including 18 sigma factors and hundreds of transport, signal sensory, and transduction systems. Although we only have just begun to understand this complexity, genome data presented here will now allow us to systematically link whole genome expression studies with high quality annotation data that are publicly available via the MaGe system. In addition, the comparison of the *C. metallidurans* CH34 genome with the genomes of three other *Cupriavidus* strains, each from a species with a particular lifestyle and habitat, will offer many clues on their physiology, metabolic capacities, and distinct interactions with their environment.

Materials and Methods

*C. metallidurans* strain CH34 isolation and propagation

Strain CH34 of *C. metallidurans* was isolated previously as described [3]. This strain is the type strain of the species (currently holding at least 17 strains [32]), and was deposited into the BCCM/LMG public strain collection (http://bccm-belgium.be/) as strain LMG1195. The bacterium was routinely grown at 28–30°C in rich Luria-Bertani medium or in Tris-salt mineral medium with gluconate as sole carbon source (MM284). The latter medium especially was used for its growth in the presence of heavy metals [9].

Sequencing, assembly, and gap closure

The complete genome sequence of *C. metallidurans* strain CH34 was determined using the whole-genome shotgun method at the Joint Genome Institute (JGI), USA, using a combination of three randomly sheared libraries that contain inserts in the 3kb and 8kb (plasmids), and 40 kb (fossils) size range. All general aspects of library construction and sequencing performed at the JGI are described at http://www.jgi.doe.gov/. These libraries were sequenced to a total depth of 9.9X, and all reads assessed for quality. Before assembling the reads, all were trimmed by removing vector-specific sequences. Libraries were assembled using the Parachute Genome Assembler (PGA; http://www.parachute.com/). Possible faulty assemblies, resulting from large repeat elements such as IS elements, were corrected by verifying the correspondence between the left- and right-direct repeat (i.e., of the IS element) via PCR amplification and sequence analysis of the predicted product. Gaps between contigs were closed by using the...
Consed editing software [213], custom primer walks, and/or PCR amplification. The final assembly consisted of 184,165 Sanger reads (i.e., using conventional dye-terminator chemistry). Of those, 3,752 were finishing reads used to correct errors in assembling the draft. The estimated error rate of the completed genome sequence was 1 in 2.5 Mb, or less. The final assembly contained two chromosomes (3,928,089 bp and 2,580,084 bp in size) and two plasmids (233,720 bp and 171,459 bp in size). We verified the consistency of the chromosome whole assembly by comparing the genome sequence map with a gene content map obtained by conjugation [22]. The sequence and annotation data of the four replicons of 
C. metallidurans CH34 were deposited in the NCBI database [http://www.ncbi.nlm.nih.gov/Genbank/] under the GenBank accession numbers NC_007971 (for pMOL30), NC_007972 (for pMOL28), NC_007973 (for CHR1), and NC_007974 (for CHR2).

Structural and functional genome annotation
Prediction of ORFs and other genome features (RNA encoding genes, riboswitches, ribosome binding sites, signal sequences, etc.) as well as functional predictions were done by the MaGe platform [40]. This system allowed the manual curation of all genes and their functions. MaGe is also interlinked to metabolic pathway tools such as KEGG and BioCyc, displays synteny maps for easy genome comparison, and has an exploration interface to query the underlying SQL databases. All C. metallidurans CH34 project data are freely available at https://www.genoscope.cnrs.fr/aage/mage/.

Circular plots
Circular maps for the two main replicons CHR1 and CHR2 (Figure 1) were drawn using the Circos software [214]. We calculated the GC content and GC-skew data employing the gc2viz.pl and gcskew2viz.pl scripts from the GenomeViz package [215].

Identification of ori and ter regions
Regions representing the putative origin of replication (ori) and termination (ter) were predicted by Ori-Finder software [216] that combines gene finding, detection of base composition asymmetry (e.g., GC-skew), and analysis of DnaA-box distribution. These regions were scrutinized for additional features and repetitive elements (Figure S1).

Distribution of functional classes
Functional classes were classified in COGs (Clusters of Orthologous Groups) by the COGnitor software imbedded into the MaGe system. All COG tables were retrieved from MaGe via the Export pages and the derived CDS numbers were used to calculate the normalized distribution of functional content over the two main replicons (Table S1, Figure 2).

Clustering
Hierarchical clustering was performed via the heatmap.2 function of the ggplot2 library in R (version 2.8.0) using single linkage clustering. For clustering of the different replicons (Figure 3; Figure S3), we normalized the obtained numbers of orthologs shared between two replicons according to the actual size of the replicon, i.e., the distance scores used in hierarchical clustering were calculated as [1 – number of orthologs/size of the smallest genome] [217]. For the clustering of different RpD proteins (Figure S5) we used the distance measure [100 – the percentage of protein similarity].

Genome comparative analysis
Whole genome alignments of the four Cupriavidus and the R. solanacearum genomes (Figure 4; Figure S2) were performed using the anchor-alignment software Murasaki [http://murasaki.dna.bio.keio.ac.jp/] [218]. Genome comparisons between C. metallidurans CH34 and P. necessarius STIR were done using either the LinePlot tool of MaGe [48] or the PROmer module of the MUMmer v3.0 package [219].

Clustering of protein families (paralogy)
To classify protein sequence sets into families according to sequence similarity we used the GeneRAGE algorithm [56]. This algorithm constructs a binary matrix holding all similarity relationships from an all-to-all protein sequence comparison performed with BLAST v2.0 [220] and deploys the CAST algorithm [221] to filter sequences for low-complexity regions. The matrix then is processed for symmetry and transitivity relationships using successive rounds of the Smith-Waterman dynamic programming algorithm [222]. In this way, false relationships within the matrix are detected, as well as multi-domain protein families are detected. We have used a cut-off E-value threshold of $1 \times 10^{-10}$ for all BLAST comparisons, and the default Z-score values for symmetrification and multi-domain detection (10 and 3, respectively).

Identification of orthologs
Protein sequence similarity was determined using BLAST (v2.0) software [220]. Proteins between different organisms were defined as orthologs when (i) they displayed a minimal percentage sequence identity of 35%; (ii) they are each other’s reciprocal best BLAST hit; and, (iii) the sequence identity was measured over at least 80% of the smallest protein.

Supporting Information

Figure S1 Origins of replication for the main replicons (chromosomes) CHR1 and CHR2 of C. metallidurans CH34 (panels A and B, respectively). (A) DnaA-binding boxes with consensus sequence TATCCACA are given as blocked arrows. Dark shaded boxes have a perfect match with the consensus sequence, while lighter shaded boxes have one or two mismatches (marked in red in their respective sequences). (B) Putative RepA-binding sites (see text) are indicated as ovals. Blue shaded ovals indicate putative RepA-binding sites on the opposite strand.

Figure S2 Genome dynamics in the second chromosome of Cupriavidus andRalstonia. Nucleotide based comparison of the CHR2 replicons of four Cupriavidus species and R. solanacearum GMI1000 (denoted as Rsol) using the anchor-alignment software Murasaki [http://murasaki.dna.bio.keio.ac.jp/]. Scale in Mb is shown on top. Abbreviations: Cneet, C. metallidurans CH34; Cpin, C. pinatubonensis JMP134; Cet, C. eutrophus; Ctdt, C. taiwanensis. Found at: doi:10.1371/journal.pone.0010433.s001 (0.05 MB PDF)

Figure S3 Similarity matrix between CHR34 replicons and replicons of related organisms using protein based orthology data. Protein contents of all replicons for four Cupriavidus species and related organisms were compared by scoring the reciprocal best BlastP hits (values in lower part of the matrix) and subsequent cluster analysis of normalised values (upper part of the matrix) (see Methods). Most analysed replicons broadly fall into two large clusters A and B, with main chromosomes grouping tightly together, while three plasmids group together in cluster C owing to
a backbone of essential genes (see also text). General information on all those species is given in Table 2. The genome of Escherichia coli K-12 served as a reference and outlier. Abbreviations: MPL, megaplasmid; Cmet, C. metallidurans CH34; Cpin, C. putrefaciens MPM134; Cesat, C. eutrophus H16; CtaJ, C. taiwanensis; Ecol, Escherichia coli K-12; Bxen, Burkholderia xenovorans LB400; Bpet, Bordetella petrii; Jmar, Janthinobacterium sp. Marseille; Hars, Hermoniaea arsenoxydans; Daci, Delftia acidovorans SPH-1; Rsol, Ralstonia solanacearum GMI1000; and Pnec, Polynucleobacter necessarius STIR1. Replicons are denoted by their commonly known name (e.g., pRALT) or subindexed (e.g., Cmet_CHR1).

**Figure S4** Protein-based synteny between the C. metallidurans CH34 chromosomes and the genome of P. necessarius STIR1, an obligate symbiote of freshwater protozoans. (A) Synteny displays were generated by the LinePlot tool of the MaGe system using a synton size (S) of 3 genes. Both species belong to the family of Burkholderiaceae and a plausible reason for the high synteny between CH34-CHR1 and the STIR1 genome would be the preservation of essential genes during genomic reduction in the STIR1 symbiont. Genes indicated on the positive (+) and negative (-) strand are rRNA (blue), tRNA (green), and transposons (pink). Strand conservations and strand inversions between de compared replicons are colored green and red, respectively. (B) Dot plot analysis of MUMs (maximal unique matching subsequences) between the main chromosome of C. metallidurans CH34 and the P. necessarius STIR1 genome. MUMs between both forward strands are shown in red and MUMs between the forward and reverse strands are shown in blue. The analysis was performed at protein level (using the six reading frames) deploying the PROmer module of MUMmer software with default parameters.

**Figure S5** Sigma-70 sequence similarity heatmap based on pairwise Smith-Waterman alignments of RpoD sequences. While the RpoD1 protein of CH34 (Rmet_2606 sited on CHR1) groups tightly with RpoD orthologs from the other three Cupriavidus species, the second sigma-70 factor RpoD2 (Rmet_4661 sited on CHR2) does not group with any of the RpoD orthologues in Cupriavidus but displays more significant BlastP hits with the RpoD proteins of Bordetella and Burkholderia species. All gene notations are from GenBank. Abbreviations: Rmet, C. metallidurans CH34; Bxe, B. xenovorans LB400; Reut, C. putrefaciens MPM134; RALT, C. taiwanensis; H16, C. eutrophus H16; Bpet, B. petrii; Daci, D. acidovorans SPH-1; Pnec, P. necessarius STIR1; Hars, J. arsenoxydans; Daci, D. acidovorans SPH-1; Rsol, R. solanacearum GMI1000; b3067 denotes the RpoD protein sequence of E. coli str. K-12 substr. MG1655.

**Figure S6** Schematic view of the cbb loci in C. metallidurans CH34. Loci involved in the Calvin-Benson-Bassham cycle (reductive pentose phosphate pathway) are dispersed over the CH34 genome although most are within the Rmet_1491 through Rmet_1544 region. Note that IS-mediated excision of this region (or parts thereof) would lead to the impairment of autotrophic growth in CH34, a phenomenon that can be observed when strain CH34 is under mutagenic stress (see text). Note that the SH enzyme effectively acts as a ‘redox valve’. (B) Proposed pathway for the oxidation of sulfite to sulfite (SO4^2-) through the action of a periplasmic sulfite dehydrogenase, SorA, and a membrane-bound c-type cytochrome, SorB. Genes and pathways for sulfur oxidation (sulfur autotrophy) in C. metallidurans CH34. (A) Genetic loci encompassing sulfur oxidizing genes (sox) and sulfur oxygenase reductase genes (soy). Note that strain CH34 lacks the soxGD genes for a periplasmic sulfur dehydrogenase which is not essential for the oxidation of thiosulfate although fewer electrons are generated (see text and next panel).
proteins are omitted from the table.

Table S5 Transcription factors - meta analysis. Transcription factors across 965 organisms, including 673 bacteria (source: http://www.transcriptionfactor.org/). Note that this table consist of multiple sheets: Sheet 1, DBDOverview - general information and data from DBD; Sheet 2, DBD organism - breakdown of transcriptional regulators for each of the organisms listed in Table 2 and E. coli K-12; Sheet 3, MageReplicon - distribution of transcriptional regulators across the four replicons of CH34, using MaGe data.

Table S6 Functional distribution and chromosomal location of C. metallidurans CH34 genes implicated in iron uptake and -metabolism.

Table S7 Signal transduction systems. Domain profiles of C. metallidurans CH34 and related organisms listed in Table 2 and E. coli K-12 (source: http://mistdb.com/). Only domains that occur 10 times or more in CH34 are considered. Extreme underrepresentations of domains (according to actual numbers versus genome size) are shaded in orange. Rows 25–27 hold the distributions of predicted one-component and two-component signal transduction systems (denoted as 1CST and 2CST, respectively) as well as the total number of predicted systems. The latter was used to calculate the proportional number (in percentages) of proteins putatively involved in signal transduction.

Table S8 Prevalence of Bug domains according to the MistDB database. Distribution of 2,377 proteins with Bug domains across 1,216 complete and draft, bacterial and archaeal genomes (source: http://mistdb.com). Organisms without Bug domain-containing proteins are omitted from the table.

References


resistant to Cupiavis metabolidurans strain CH34. Microbiology 152: 1763–1776.


