



**HAL**  
open science

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

Paul J Janssen, Rob van Houdt, Hugo Moors, Pieter Monsieurs, Nicolas Morin, Arlette Michaux, Mohammed A Benotmane, Natalie Leys, Tatiana Vallaey, Alla Lapidus, et al.

### ► To cite this version:

Paul J Janssen, Rob van Houdt, Hugo Moors, Pieter Monsieurs, Nicolas Morin, et al.. The Complete Genome Sequence of *Cupriavidus metallidurans* Strain CH34, a Master Survivalist in Harsh and Anthropogenic Environments. PLoS ONE, 2010, 5 (5), pp.e10433. 10.1371/journal.pone.0010433 . hal-04580603

**HAL Id: hal-04580603**

**<https://hal.science/hal-04580603>**

Submitted on 20 May 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Paul J. Janssen<sup>1\*</sup>, Rob Van Houdt<sup>1</sup>, Hugo Moors<sup>1</sup>, Pieter Monsieurs<sup>1</sup>, Nicolas Morin<sup>1</sup>, Arlette Michaux<sup>1</sup>, Mohammed A. Benotmane<sup>1</sup>, Natalie Leys<sup>1</sup>, Tatiana Vallaëys<sup>2</sup>, Alla Lapidus<sup>3</sup>, Sébastien Monchy<sup>4</sup>, Claudine Médigue<sup>5</sup>, Safiyh Taghavi<sup>6</sup>, Sean McCorkle<sup>6</sup>, John Dunn<sup>6</sup>, Daniël van der Lelie<sup>6</sup>, Max Mergeay<sup>1</sup>

**1** Molecular and Cellular Biology, Belgian Nuclear Research Center SCK-CEN, Mol, Belgium, **2** Laboratoire Ecosystèmes Lagunaires (ECOLAG), Centre National de la Recherche Scientifique - UMR 5119, Université Montpellier II, Montpellier, France, **3** Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, California, United States of America, **4** Laboratoire Microorganismes: Génome et Environnement (LMGE), Centre National de la Recherche Scientifique - UMR 6023, Université Blaise Pascal, Aubière, France, **5** Laboratoire d'Analyses Bioinformatiques pour la Génomique et le Métabolisme (LABGeM), Centre National de la Recherche Scientifique - UMR8030 and CEA/DSV/IG/Genoscope, Evry, France, **6** Biology 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 of, 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 19 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.

**Citation:** Janssen PJ, Van Houdt R, Moors H, Monsieurs P, Morin N, et al. (2010) The Complete Genome Sequence of *Cupriavidus metallidurans* Strain CH34, a Master Survivalist in Harsh and Anthropogenic Environments. PLoS ONE 5(5): e10433. doi:10.1371/journal.pone.0010433

**Editor:** Niyaz 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  $\beta$ -proteobacterium *Cupriavidus* (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 RP4::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 *Cupriavidus* 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  $Cu^+$ ,  $Cu^{2+}$ ,  $Ni^{2+}$ ,  $Zn^{2+}$ ,  $Co^{2+}$ ,  $Cd^{2+}$ ,  $CrO_4^{2-}$ ,  $Pb^{2+}$ ,  $Ag^+$ ,  $Au^+$ ,  $Au^{3+}$ ,  $HAsO_4^{2-}$ ,  $AsO_2^-$ ,  $Hg^{2+}$ ,  $Cs^+$ ,  $Bi^{3+}$ ,  $Tl^+$ ,  $SeO_3^{2-}$ ,  $SeO_4^{2-}$ , and  $Sr^{2+}$ , with at least 24 gene clusters distributed over four replicons, 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 a wide variety of systems for RND (for Resistance, Nodulation, cell Division) and CDF (for Cation Diffusion Facilitation). It also contains transporters specialized for certain metal ions, whereby efflux may be followed by metal sequestration or complexation (this study; [7,8,38–40]). Many of these metal responses overlap and are multilayered ([10] and unpublished results). In addition, we have noted changes in gene expression, cell morphology and physiology for *C. metallidurans* CH34 cells in response to spaceflight (i.e. cells grown at the International Space Station during 10-day visits) and oxidative stress [41,42], as well as simulated microgravity and ionizing radiation (unpublished). The organism has been isolated from very harsh environments such as ultra-clean spacecraft assembly rooms [43] and spent-nuclear-fuel pools [44]. Not

surprisingly, genome data indicate that CH34 has at its disposal intricate regulatory networks and many signal transduction systems.

Members of the genus *Cupriavidus* are metabolically diverse and occur in a wide range of habitats. In this study, we compared the CH34 genome with the genomes of three other species: The  $H_2$ -oxidising *C. eutrophus* H16 [45], the chloroaromatic pollutant degrading *C. pinatubonensis* JMP134 (unpublished), and the nitrogen-fixing plant symbiont *C. taiwanensis* [46]. While each of these four species carry plasmids that *de facto* determine their lifestyle, orthology and synteny analyses showed they share many 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 carboxylate-binding motif first noted as overrepresented in host-restricted *Bordetella* species [47]. The prevalence of these receptors in *Cupriavidus* 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, the many metal-resistance mechanisms, 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 replicons: 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 replicons. Nonetheless, in a separate subsection below, we give a brief update on the megaplasmids. Table 1 lists the general features of all four replicons. There are four sets of 5S, 16S, and 23S rRNA genes and 12 structural RNA genes that include one RNAaseP RNA (*mpB*; Rmet\_R0004), one tmRNA (*sszA*; Rmet\_R0007), one SRP RNA (*fff*; Rmet\_R0026), and nine riboswitches as predicted by Rfam (<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 COGnitor module [49]

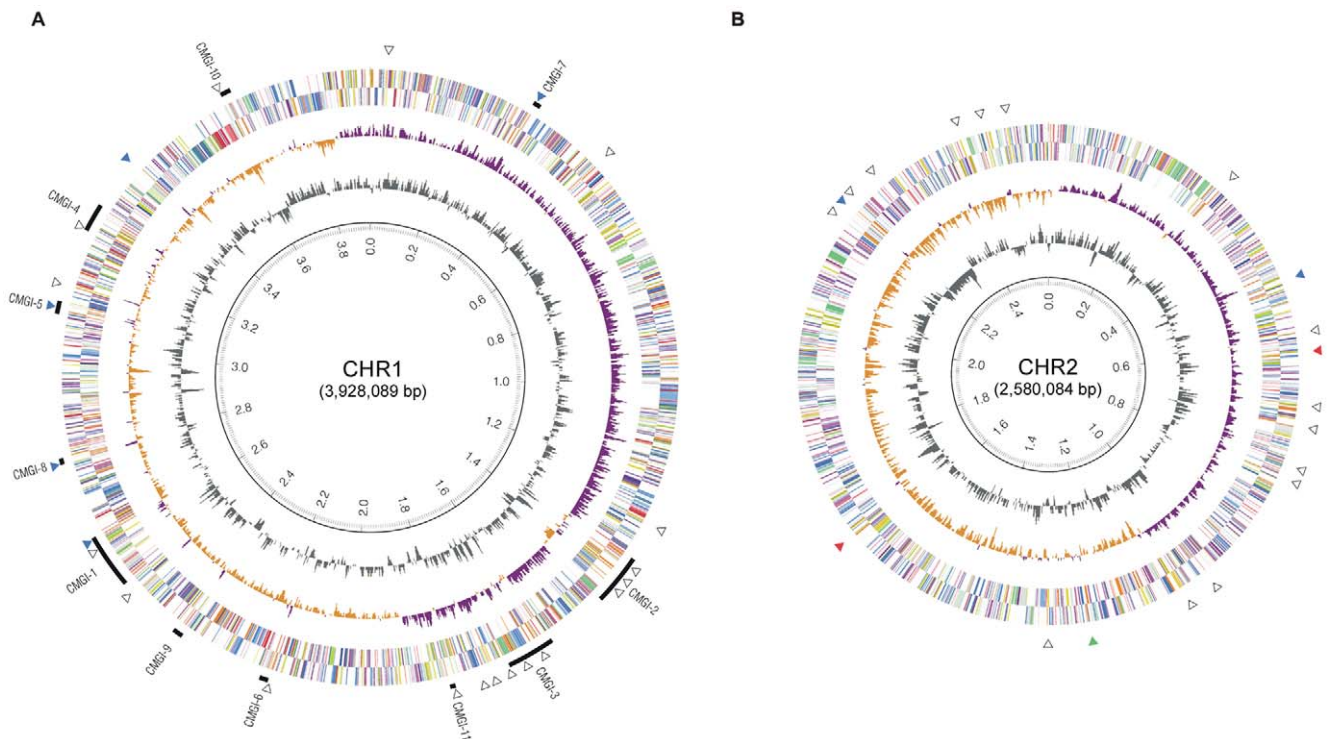
**Table 1.** General features of the *Cupriavidus metallidurans* CH34 genome.

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

doi:10.1371/journal.pone.0010433.t001

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

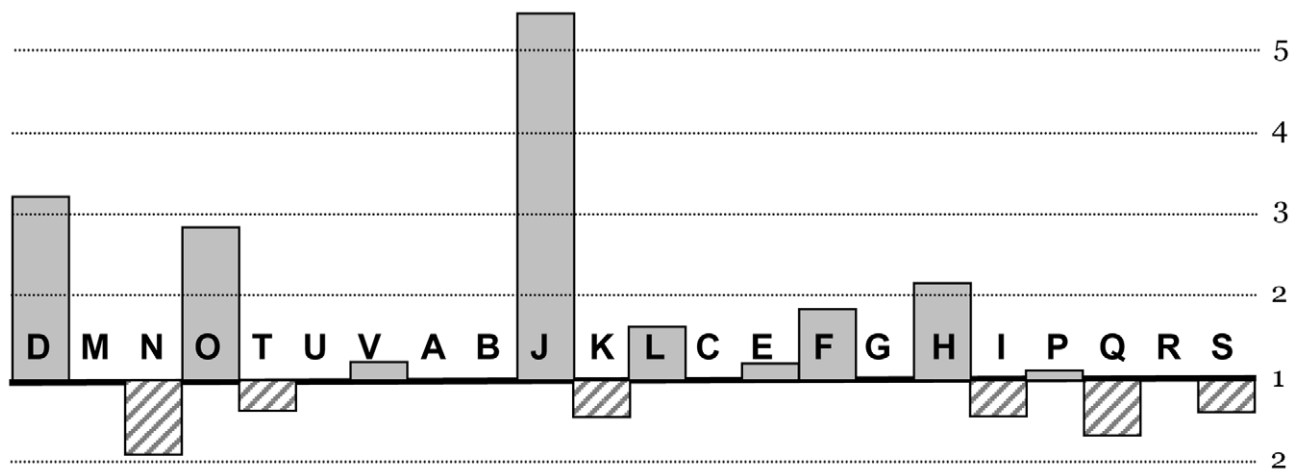


**Figure 1. Circular representation of the two main replicons of CH34.** Circles display from the inside outwards: (ring 1) scale in Mb, (ring 2) GC-content using a 2 kb window, (ring 3) GC-skew (G-C/G+C ratio) using a 2 kb window, (rings 4 and 5) predicted CDSs transcribed in a counterclockwise/clockwise 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

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\_5284) 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 *infA2* (Rmet\_5176), a second housekeeping sigma-70 factor encoded by *rpoD2* (Rmet\_4661) and six genes that code for alternative sigma factors *fliA* (Rmet\_3702), *sigJ* (Rmet\_3844), *rpoK* (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; *crtB*, Rmet\_4149), acetone utilisation and tolerance (*acxRABC*, 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\_3841), biofilm formation (*pelABCDEFG*), and exopolysaccharide synthesis (*epsAPB*). It carries most of the genes for motility and chemotaxis (*mot*, *flh*, *fli*, *flg*, *cpa*, *tad*, *pil*, *fim*, and *che* loci), peroxide



<b><u>Cellular processes</u></b>		<b><u>Metabolism</u></b>	
D	Cell cycle control, cell division, chromosome partitioning	C	Energy production/conversion
M	Cell wall/membrane/envelope biogenesis	E	Amino acid transport/metabolism
N	Cell motility	F	Nucleotide transport/metabolism
O	Posttranslational modification, protein turnover, chaperones	G	Carbohydrate transport/metabolism
T	Signal transduction mechanisms	H	Coenzyme transport/metabolism
U	Intracellular trafficking, secretion, and vesicular transport	I	Lipid transport/metabolism
V	Defense mechanisms	P	Inorganic ion transport/metabolism
<b><u>Information storage and processing</u></b>		Q	Secondary metabolism
A	RNA processing and modification	<b><u>Poorly characterized</u></b>	
B	Chromatin structure and dynamics	R	General function prediction only
J	Translation, ribosomal structure & biogenesis	S	Function unknown
K	Transcription	<div style="display: flex; justify-content: space-around;"> <div style="border: 1px solid black; width: 20px; height: 10px; background-color: #cccccc;"></div> <b>CHR1</b> <div style="border: 1px solid black; width: 20px; height: 10px; background: repeating-linear-gradient(45deg, transparent, transparent 2px, black 2px, black 4px);"></div> <b>CHR2</b> </div>	
L	Replication, recombination and repair		

**Figure 2. Functional distribution over CHR1 and CHR2 by COG classification.** The scale represents the normalised ratio of CDS numbers per replicon (i.e. the ratio of percentages for each class per replicon); a ratio of 1 means that genes classified in a particular COG are, numerwise, evenly distributed over both replicons, taking into account the different total count of genes per replicon. Detailed data are in Table S1. doi:10.1371/journal.pone.0010433.g002

removal and oxidative stress (*yhjA*, *ahpC*, *ahpD*, *katG*, *pcaCD*, *soxR*, *glrX*), DNA adenine methylation (*dam*), DNA repair (*ada*, *alkA*, *alkB1*, *alkB2*, *sbcCD*, *ogt2*, *ogt3*), adaptive mutagenesis (*imuA-imuB-dnaE2*, Rmet\_4128-29-30), and eight copies of the *uspA* 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 (*bugSA*) 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* layout (genes Rmet\_0001, Rmet\_0002, and Rmet\_0003, with *dnaA* starting at position 104) although the 400 bp upstream region of *dnaA* contains 1 perfect DnaA-binding box (TTATCCACA) and 3 putative DnaA boxes with one or two mismatches to this consensus (Figure S1A). Interestingly, the GC skew transition region located around coordinate 3,837,500 is preceded immediately by a short region of 213 nt comprising 3 perfect DnaA boxes. Four more putative DnaA boxes (tightly clustered in a 65 nt region and each with one mismatch to the consensus) occur even further upstream of *dnaA*, at coordinates 3,707,588–3,707,653 (Figure S1A). For CHR2 the switch in GC skew is situated around the *repA-parA-parB* genes (Rmet\_5817, Rmet\_5819, and Rmet\_5820) (point 2.43 of the map, Figure 1B)

indicating that CHR2 partitioning may be plasmid-like. The *repA* upstream region contains 3 putative DnaA boxes with only one mismatch to the consensus, at locations –392, –419, and –1080 with respect to the startcodon of *repA* (Figure S1B). This region also contains at least 12 repetitive 17 nt-long elements with a highly conserved motif (consensus CGCAGAA(A/T)(C/T)(A/G)GGTACG(C/T)) that may serve as RepA binding sites. In CHR1, the terminus region as predicted by a shift in GC skew is diametrically opposed to the origin of replication (Figure 1) leading to replichores of approximately equal length. However, the CHR2 replicon is clearly divided in two unequal replichores. The reason for this asymmetry is unclear but may be related to genome dynamics since the *C. metallidurans* CH34 genome harbours numerous Mobile Genetic Elements (MGEs) including many Insertion Sequence (IS) elements, transposons, and Genomic Islands (CMGIs). We further discuss details of the *C. metallidurans* CH34 mobilome in separate subsections.

**Phylogeny and genome comparative analyses**

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

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

Organism <sup>a,b</sup>	Main property	Family	Principle habitat	CHR1	CHR2	Replicon 3 <sup>c</sup>	Replicon 4	Year/Ref.
<i>C. metallidurans</i> * CH34	heavy metal resistance	<i>Burkholderiaceae</i>	heavy metal-rich sediments and soils	3.93	2.58	0.17 (pMOL28)	0.23 (pMOL30)	This report
<i>C. pinatubonensis</i> JMP134	degrader of xenobiotics	<i>Burkholderiaceae</i>	soil, aquatic; volcanic mudflow deposits	3.81	2.74	0.63 (MPL)	0.09 (pJP4)	unpublished
<i>C. eutrophus</i> H16	hydrogen oxidation, PHA <sup>1</sup> metabolism	<i>Burkholderiaceae</i>	soil, freshwater	4.05	2.91	0.45 (pHG1)	-	2006/[45]
<i>C. taiwanensis</i>	N <sub>2</sub> -fixing legume symbiont	<i>Burkholderiaceae</i>	plant root nodules	3.42	2.50	0.56 (pRALTA)	-	2008/[46]
<i>R. solanacearum</i> GMI1000	plant pathogen with wide host range	<i>Burkholderiaceae</i>	soil, rhizosphere	3.72	2.09	-	-	2002/[223]
<i>B. xenovorans</i> LB400	PCB <sup>2</sup> degrader	<i>Burkholderiaceae</i>	rhizosphere, polluted soils, landfills	4.90	3.36	1.47 (MPL)	-	2006/[224]
<i>D. acidovorans</i> * SPH-1	organotrophy, LAS <sup>3</sup> degrader	<i>Comamonadaceae</i>	soil, aquatic, sewage	6.77	-	-	-	unpublished
<i>B. petrii</i> * Se-1111R	degrader of aromatic compounds	<i>Alcaligenaceae</i>	river sediments, TCB <sup>4</sup> contaminated soils	5.29	-	-	-	2008/[148]
<i>Janthinobacterium</i> sp.	heavy metal resistance, small sized	<i>Oxalobacteraceae</i>	aquatic, 0.22 µm filtered water	4.11	-	-	-	2007/[225]
<i>H. arsenicoxydans</i>	arsenic oxidation and scavenging	<i>Oxalobacteraceae</i>	activated sludge, industrial waste water	3.42	-	-	-	2007/[226]
<i>P. necessarius</i> STIR1	obligate intracellular endosymbiont	<i>Burkholderiaceae</i>	freshwater ciliates	1.56	-	-	-	unpublished

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

doi:10.1371/journal.pone.0010433.t002

*taiwanensis*. 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 Ctai (for *C. taiwanensis*).

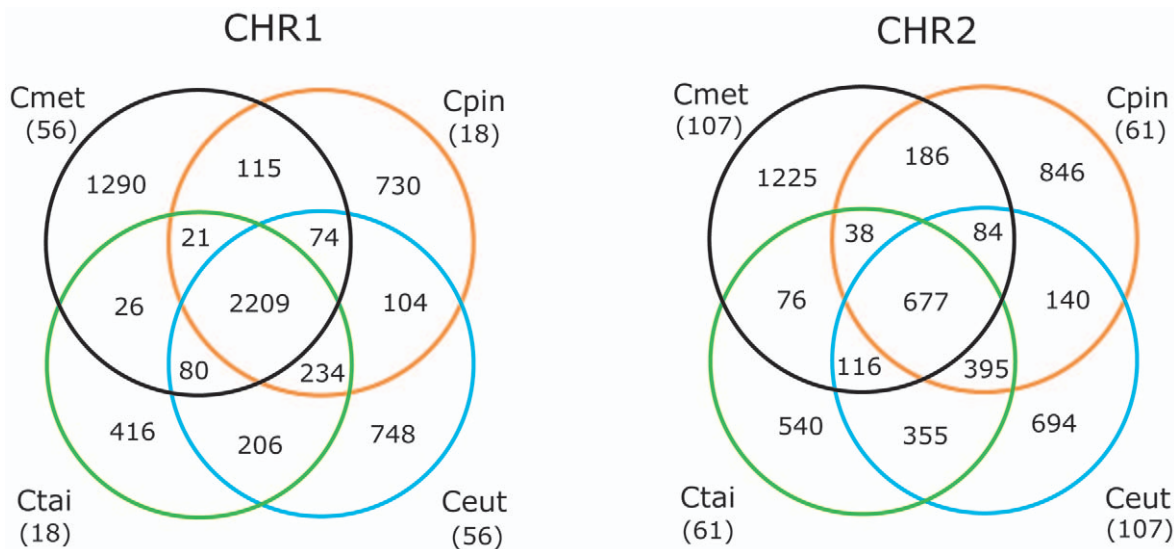
The genome organisation of these four species is very similar. All four carry two large replicons, that we further nominated as CHR1 and CHR2, with each replicon being of comparable size to their counterparts. The main difference in terms of genome organisation stems from the presence of plasmids (Table 2). Whereas Cmet harbours two megaplasmids of 171- and 233-kb, Ceut and Ctai 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 *Herminiimonas arsenicoxydans* (all having a single replicon). We further included *Polynucleobacter 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 Ctai [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 genetransfer 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* (Ctai) 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 Ctai 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*, *Janthinobacterium*, 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 (Ctai), 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 parantheses 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*; Ctai, *C. taiwanensis*. doi:10.1371/journal.pone.0010433.g003

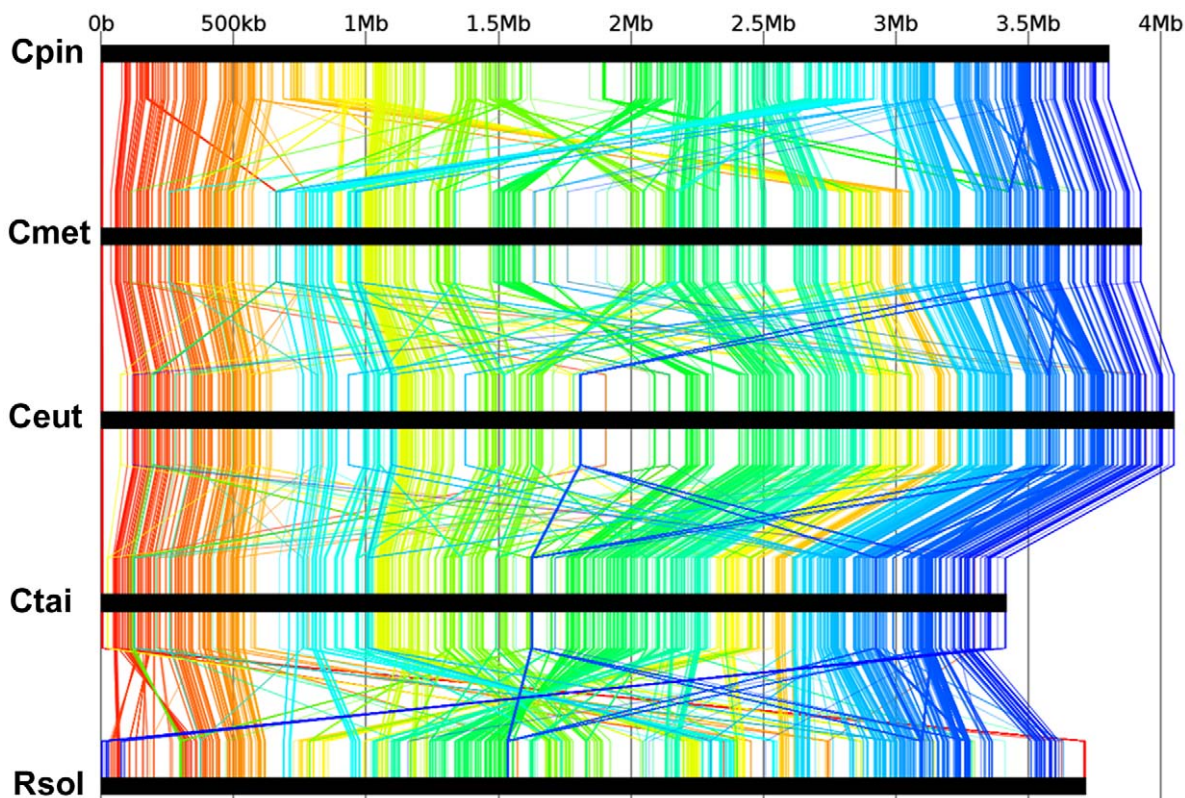
three plasmids in cluster C have the same backbone, i.e., contain the essential genes for plasmid replication and maintenance ([7] and see further). Interestingly, the genome of *P. necessarius* STIR1, an obligate intracellular symbiont of the freshwater ciliate *Euplotes aediculatus*, not only clusters 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 *Cmet* 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  $S \leq 5$  e.g. considering five or less genes for each synton 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



**Figure 4. Comparative DNA analysis of *Cupriavidus* chromosomes.** Nucleotide based comparison of the CHR1 replicons of the four *Cupriavidus* species (abbreviations as in Figure 3) and *R. solanacearum* GMI1000 (denoted as *Rsol*) using the anchor-alignment software Murasaki (<http://murasaki.dna.bio.keio.ac.jp/>) [218]. Scale in Mb is shown on top. doi:10.1371/journal.pone.0010433.g004



transduction (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 dehydratases, acyl-CoA dehydrogenases, and acyl-CoA synthetases). These findings are in line with those of Gevers *et al.* [57] 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 paralog sizes of 7% for intracellular organisms, such as *Rickettsia conorii*, 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 paralogome. 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.* [58]) 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 butyryl-CoA:acetate-CoA transferase, an enzyme that may be involved in acetate detoxification or instrumental in delivering butyrate to produce polyhydroxybutyrate (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 replicons (Table 1). Using Van Houdt *et al.*'s criteria [59], 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 replicon 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 campinensis* 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 between (industrial) environmental reservoirs 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 Tn4371 family, first described in strain *C. oxalaticus* 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 toluene corresponding to CH34's ability to grow with toluene, 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 CO<sub>2</sub> (*cbb* 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 CO<sub>2</sub> and H<sub>2</sub> are confined to CMGI-2 and CMGI-3 in CHR1: this differs in the closely related H<sub>2</sub>-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 CO<sub>2</sub> fixation) in both strains clearly have different origins since the CH34 *cbb* genes are phylogenetically much more closely related to the *cbb* homologues of photosynthetic- 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. acidovorans* SPH-1 genome [69–71] except that the latter does not contain Tn6048 but carries a complete Tn4371-like structure flanked by Tn6051 [59].

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 CMGI-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 IS1087 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 transposon: 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 *gspA* because its product shows strong similarities to putative general secretion pathway ATPases. We also note that although the core region (*tnpA-tnpR*) 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 Tn3-related transposons i.e., alone, as in Tn1013 from the IncP-1 $\alpha$  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 ISRme3 (an IS3 family member), with 10 copies, and IS1088 (an IS30 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 *hyp* and *hox* 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 ISRme5 copies on CMGI-2 are located amidst multiple transposase genes and fragments, and probably were the driving force toward integrating the *hyp/hox* cluster on CMGI-2. In addition, both copies of ISRme7 flank the genomic island CMGI-11 while two copies of ISRme3 are associated with the genomic islands in pMOL30, further

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

Element	Total nr.	CHR1	CHR2	pMOL28	pMOL30
Tn4378	1			1	
Tn4380	1				1
Tn6048	3	1	2		
Tn6049	12	7	3	1	1
Tn6050	2		2		

doi:10.1371/journal.pone.0010433.t003

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

Element	Family (sub-)	Total nr.	CHR1	CHR2	pMOL28	pMOL30
IS1071	Tn3	3	3			
IS1086	IS30	3	1	1	1	
IS1087	IS3 (IS2)	2	2			
IS1088	IS30	9	3	6		
IS1090	IS256	4	4			
ISRme1	IS5 (IS427)	4	2	2		
ISRme3	IS3 (IS3)	10	3	5		2
ISRme4	IS21	2	2			
ISRme5	IS481	4	3	1		
ISRme6	IS5 (IS427)	1	1			
ISRme7	IS6	2	2			
ISRme8	IS4	2	1	1		
ISRme9	IS21	1			1	
ISRme10	IS30	1				1
ISRme11	IS3	2		2		
ISRme12	IS3 (IS150)	1	1			
ISRme15	IS3 (IS51)	2		1		1
ISRme17	IS3 (IS150)	1	1			
ISRme18	Tn3	1			1	
ISRme19	IS66	1				1
ISRme20	IS21	1	1			

doi:10.1371/journal.pone.0010433.t004

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  $\alpha$ - and  $\beta$ -proteobacteria and termed them *recombinase in trio* or *rit* elements. One trio (*ritA<sub>2</sub>B<sub>2</sub>C<sub>2</sub>*) (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 (*ritA<sub>1</sub>B<sub>1</sub>C<sub>1</sub>*) (Rmet\_1271 to Rmet\_1273) also lies on CMGI-2 and is part of a complex accessory module; it inactivates a *tnpA* gene of the IS66 family. The *C. metallidurans ritA<sub>1</sub>B<sub>1</sub>C<sub>1</sub>* element exhibits high similarity to the *rit* elements on plasmid pHG1 of *C. eutrophus*, showing very high aa sequence similarities (>92% identity) for *ritA<sub>1</sub>* and *ritB<sub>1</sub>* to either PHG129 and PHG141, or either PHG130 and PHG140, respectively. Further, there are high aa sequence similarities (40–54% identity) between *ritA<sub>1</sub>B<sub>1</sub>C<sub>1</sub>* and the trio PHG134/135/136. Immediately adjacent to this maze of pHG1 recombinase genes we discerned an IS66 type *tnpA* 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

elements we list here is the *bimAB* module, named after the *bimA* recombinase 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\_5987/88 (linked to a *czc* locus), respectively [7]. The latter module (21 genes) is fully conserved in *Ralstonia picketti* 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 arrangements, insertions, or deletions [6,22,77,78]. One well studied case is the integration of a 44 kb DNA fragment in the pMOL28 plasmid and the identification of IS1086 participating in this event [77,79]. The resulting plasmid pMOL50 has a much higher conjugation frequency than either pMOL28 or pMOL30 and is able to very efficiently (retro)mobilise very large DNA fragments from the CH34 chromosome (of up to 1 Mb), a property that was exploited to construct the genetic map of the CH34 main chromosome, CHR1 [22].

### A brief update on megaplasmids in strain CH34

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] derived 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 (398 kb) of *B. vietnamiensis* 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 [5,75,83,86] 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 sited immediately downstream of the chromate-regulated operon *chrIA<sub>1</sub>B<sub>1</sub>CEF<sub>1</sub>* (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  $\alpha$ -helical segments), *chrY* (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 *Methylobacterium* sp. 4–46, *B. vietnamiensis* G4, and *B. pseudomallei*. However, the *chr* locus in *Methylobacterium* sp. 4–46 lacks *chrE*, *chrN*, and *chrP* orthologues, the four *chr* loci of *B. vietnamiensis* G4 all lack *chrP* and *chrZ* orthologues, and the majority of *chr* loci across 21 *B. pseudomallei* strains lack *chrC*, *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-oxyanion.

We also identified hypothetical genes on CMGI-28b that encode membrane proteins rich in YD (tyrosyl-aspartyl) motifs and named them *rhsA* (Rmet\_6263), *rhsB* (Rmet\_6261), and *rhsC* (Rmet\_6258) because such YD-repeat proteins apparently are linked to RHS (Recombination Hot Spot) regions [87,88]. Noteworthy, these genes are centered around the *tnpA* and *tnpB* genes (respectively Rmet\_6260 and Rmet\_6259) of the ISRme9 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 *czc* loci on CMGI-30a is speckled by very small hypothetical genes, transposases (*tnp*), site-specific recombinases (*int*), 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 *sil* 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.

Type <sup>a</sup>	Replicon	Gene cluster <sup>b</sup>	Rmet_ number	Metal(s)	GI/transposon <sup>c</sup>	Mechanism <sup>d</sup>	Fc <sup>e</sup>	Refs.
partially plasmid-redundant	CHR1	<i>cdfX pbrR<sub>2</sub> cadA pbrC<sub>2</sub>, pbrR<sub>3</sub></i> (§)	2299, 2302–2304, 3456	Pb <sup>2+</sup>	CMGI-1 (P)	ExP, pCDF	+	[59]
	pMOL30	<i>pbrUb UaTR_ pbrABCD</i> (#)	6180, 5944–5949		CMGI-30a	ExP, seq	+	[85]
	CHR1	<i>merRTPA' A''</i>	2312–2315	Hg <sup>2+</sup>	CMGI-1 (P)	HgRed	?	[59]
	pMOL28	<i>merRTPADE urf-2</i>	6344–6346, 6183–6186		CMGI-28a/Tn4378	HgRed	+	[4]
	pMOL30	<i>merRTPADE urf-2</i>	6171–6177		CMGI-30a/Tn4380	HgRed	+	[4]
	pMOL30	<i>merRTAP</i>	5990–5992		CMGI-30a	HgTra	?	[10]
	CHR2	<i>chrBAF</i>	3866–3864	CrO <sub>4</sub> <sup>2-</sup>	-	ExChr	?	[5]
	pMOL28	<i>chrIBACEF_ONPYZ</i>	6204–6194		CMGI-28a	ExChr, FS, MFS	+	[12]
	CHR2	<i>cusDCBAF</i>	5030–5034	Ag <sup>+</sup> , Cu <sup>+</sup> , Cu <sup>2+</sup>	-	ExHME4	+	[8]
	pMOL30	<i>silDCBA cusAF</i>	6133–6136, 5953		CMGI-30b	ExHME4	(+)	[10]
	CHR2	<i>copSRABCD</i>	5673–5668	Cu <sup>2+</sup>	-	ExCop, seq	+	[8]
	pMOL30	<i>copVTKMNSRABCDIJGFOLQHEW</i>	6105–6119, 6382, 6120–6124		CMGI-30b	ExCop, ExP, seq	+	[75]
	CHR2	<i>zntA czcICABA ubiG czc SRL  hns mmmQ</i> (†)	4594–4597, 4469–4461	Cd <sup>2+</sup> , Zn <sup>2+</sup> , Co <sup>2+</sup>	-	ExHME1, ExP	(+)	[59]
	pMOL30	<i>czcMNICBADREJ ompP czcP</i>	5985–5974, 5970		CMGI-30a	ExHME1, ExP	+	[10]
	CHR2	<i>nimBA AC</i> (*)	5682–5677	Ni <sup>2+</sup> , Co <sup>2+</sup>	-	ExHME3b	-	[8]
	pMOL28	<i>cnrYXHCBAT</i>	6205–6211		CMGI-28a	ExHME2, CDF	+	[80,81]
pMOL30	<i>nccCB'B'A nreB mmmQ</i>	6148–6143		CMGI-30b	ExHME2, MFS	(+)	[10]	
not plasmid-redundant	CHR1	<i>arsPHC<sub>1</sub>BC<sub>2</sub>IRM</i>	0327–0334	HAsO <sub>4</sub> <sup>2-</sup> , AsO <sub>2</sub> <sup>-</sup>	CMGI-7	AsRed (ExAs)	+	[94]
	CHR1	<i>cupric</i>	3523–3525	Ag <sup>+</sup> , Cu <sup>+</sup> , Cu <sup>2+</sup>	-	ExP	+	[75]
	CHR1	<i>agrCBARS</i>	1748–1752		-	ExHAE	?	[8]
	CHR1	<i>hmzRS yodB hmzBAA</i>	3016–3011	divalent metals	CMGI-4	ExHME3b	-	[59]
	CHR2	<i>hmvCBAA</i>	3836–3838		-	ExHME3a	-	[8]
	CHR2	<i>zniABC<sub>SR</sub> znePRSCAB</i>	5319–5323, 5325–5330		-	ExHME3a, MSF	?	[58]
	CHR2	<i>hmyCB hmyA</i> (‡)	4120–4123		-	ExHME3b	+	[8]

<sup>a</sup>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.

<sup>b</sup>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 Tn-insertions are indicated by a pipe symbol (|); truncated genes are indicated with a delta ( $\Delta$ ) symbol; (§), *pbrR<sub>3</sub>* acts on *zntA* (Rmet\_4594); (#), insertion of Tn6049 in *pbrU*; (†), *czcB* truncated by Tn6050-mediated rearrangement (inversion) and insertion of IS1088 between *czcL* and *hns*; (\*), insertion of ISRme3 in *nimA*; (‡), insertion of IS1088 between *hmyB* and *hmyA*.

<sup>c</sup>CMGI, *Cupriavidus metallidurans* CH34 Genomic Island; (P), PGI-2 family.

<sup>d</sup>Mechanisms of HM detoxification: ExP, efflux (P1-ATPase); pCDF, putative CDF-like; seq, periplasmic sequestration; HgRed, reduction of Hg<sup>2+</sup> into Hg<sup>0</sup> which then volatilizes out of the cell; HgTra, tentative transport of Hg<sup>2+</sup>; ExChr, 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 Cu<sup>2+</sup>; AsRed (ExAs), reduction of HAsO<sub>4</sub><sup>2-</sup> into AsO<sub>2</sub><sup>-</sup> and efflux of AsO<sub>2</sub><sup>-</sup>.

<sup>e</sup>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 cations, including Cu<sup>+</sup>, Ag<sup>+</sup>, Cu<sup>2+</sup>, Co<sup>2+</sup>, Cd<sup>2+</sup>, Ni<sup>2+</sup>, Pb<sup>2+</sup> and Zn<sup>2+</sup> [8,10,59,75,80,83,85,90,92,93]. Alternative mechanisms with specificity to other cations or oxyanions include the chromate CrO<sub>4</sub><sup>2-</sup> efflux system [5], the reduction of arsenate HAsO<sub>4</sub><sup>2-</sup> by ArsC reductases followed by the extrusion of the formed arsenite AsO<sub>2</sub><sup>-</sup> by the ArsB permease [94], the CopABCD/PcoABCD mediated periplasmic detoxification for Cu<sup>2+</sup> [95,96], and the reduction of Hg<sup>2+</sup> 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* YiiP zinc transporter protein [103,104].

It is worth mentioning that RND systems for HME (Heavy Metal Efflux) were first discovered in *C. metallidurans* CH34 with

the archetypes CzcCBA, conferring resistance to Co, Zn and Cd [9,105], 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 antiporter 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 antiporter efflux systems include the NccCBA complex (Ni, Co, Cd) of *C. metallidurans* 31A (formerly known as *Alcaligenes/Achromobacter xylosoxydans*) [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 *czc* cluster of pMOL30 contains besides the RND-HME module *czcCBA* also *czcD* (Rmet\_5979; a CDF protein) and *czcP* (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. *cdjX pbrR cadA pbrC* in CMGI-1, *hmz* in CMGI-4, and *ars* in CMGI-7), and for the large genomic islands on the megaplasmids pMOL28 and pMOL30 encoding intertwined 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-cadA/pbrA-pbrC*, and *czcICBA-RS* that have counterparts on either pMOL28 (*chr*) or pMOL30 (*cop*, *pbr*, *czc*); furthermore 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 *merA* gene of CMGI-1 and the *nccB* gene of CMGI-30b) or by partial deletion (e.g. *hmzA* 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 those novel genes are strongly induced by heavy metals (not

shown), most notably *czcJ* (Rmet\_5975) and *mmrQ* (Rmet\_6143; formerly *orf-167*), and to a lesser extent *copQ* (Rmet6121). We noted previously that CzcJ, MmrQ, and CopQ have a signal sequence (well defined by SignalP embedded in MaGe) and have a common domain motif, RDXXTDG, repeated 3 to 4 times [10]. A closer look learns that these domains are connected with spacers of defined length and sequence. Three other genes that are highly induced in the presence of metals, *mmmQ* (Rmet\_4461), *mmtQ* (Rmet\_4187), and *mmsQ* (Rmet\_4908), encode proteins with apparent signal sequences and similar domains and spacers. Note that *mmsQ* has an almost identical paralog *mmsQJ* (Rmet\_5594) thus both genes are covered by the same probe in transcriptomic analysis. The *mmmQ* gene is linked to the *czc* locus on CHR2 (Table 5) displaying excellent synteny with metal efflux loci in the three other *Cupriavidus* genomes (in spite of an IS1088 insertion between the *mmmQ* and the *czc* genes). In contrast, the other genes *mmtQ*, *mmsQ* and *mmsQJ* appear to be unlinked to any HMR locus. The metal induction of these seven genes (*czcJ*, *mmrQ*, *copQ*, *mmmQ*, *mmtQ*, and *mmsQ/J*) follows a varied and overlapping response, with very high induction by some metals and no induction by other metals. For example, *czcJ* and *mmrQ* are strongly induced by Cd<sup>2+</sup>, Cu<sup>2+</sup>, Ni<sup>2+</sup>, and Zn<sup>2+</sup>, moderately induced by Pb<sup>2+</sup>, and not induced by Hg<sup>2+</sup>, while the *mmtQ* gene is highly induced by Cd<sup>2+</sup>, Hg<sup>2+</sup>, and Pb<sup>2+</sup>, only moderately by Zn<sup>2+</sup> and not by Cu<sup>2+</sup> or Ni<sup>2+</sup> (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 *fieF* (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 Co<sup>2+</sup>, indicating that *dmeF* is essential in cobalt homeostasis [112]. Correspondingly, *fieF* mutants of strain CH34 showed decreased resistance to Ni<sup>2+</sup>, Co<sup>2+</sup>, Cd<sup>2+</sup> and Zn<sup>2+</sup> [112] and *atmA* mutants of strain CH34 exhibit lower resistance to Ni<sup>2+</sup> and Co<sup>2+</sup> [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: 1.A: alpha-Type channels; 2.A: Porters (uniporters, symporters, antiporters); 3.A: P-P-bond-hydrolysis-driven transporters; 4.A: Phosphotransfer-driven group translocators; and 9.A: 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  $\beta$ -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 (TC2.A), closely followed by the primary transporters (TC3.A). The largest families herein are the Major Facilitator Superfamily (MSF) (TC2.A.1) and the ATP-binding Casette (ABC) Superfamily (TC3.A.1) 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 (TC2.A.80) 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 (1.A.35), CDF (2.A.4), ZIP (2.A.5), RND (2.A.6), CHR (2.A.51), NiCoT (2.A.52), P-type ATPase (3.A.3), MerT (9.A.2), and TerC (9.A.30) (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-Hydroxycarboxylate Transporter (2-HCT) Family (TC2.A.24), and the Iron/Lead Transporter (ILT) Family (TC9.A.10). 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 *czc* 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<sup>+</sup> or Na<sup>+</sup>):Proton Antiporter-3 (CPA3) Family (TC2.A.63) 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 *mnhABCDEFG* 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. pectii* 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 (twin-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 peptidases 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 2.A.9 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 2.A.64 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 *gspGHIJKLMNDEF* (Rmet\_3384-94) is situated on CHR1 and is well conserved in the four *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 trans-membrane 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 *flh* 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\_3689-99, Rmet\_5258, Rmet\_5261, Rmet\_5263-64, Rmet\_5299, and Rmet\_5301-03) together with all additional *fli*, *flh*, *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\_1546 through \_1559). These clusters are located on CHR1 as part of Tn4371-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 sets of genes required for chemolithotrophic growth and degrading aromatic compounds (addressed elsewhere in this discussion). A third homologous cluster of ten genes, mostly denoted as *trb* or *tra* genes (Rmet\_6299 through \_6309), is present on plasmid pMOL28 preceding the *uvrD-parAB* genes. These *trb/tra* genes probably contribute to the maintenance and conjugal transfer of the pMOL28 plasmid and are part of the plasmid backbone common to a range of megaplasmids, including pHG1 of *C. eutrophus* strain H16 and pRALTA of *C. taiwanensis* [7,10].

## Sigma factors and transcriptional regulators

An important strategy of bacteria is to use alternative sigma factors to cope with environmental changes; *C. metallidurans* CH34, with eighteen sigma factors, appears to be particularly poised to combat stress. We generally adopted the earlier proposed description [76,114,123] for annotating these eighteen sigma factors, except for the housekeeping sigma-70 factor. In the initial annotation [123], gene Rmet\_4661 located on CHR2 was named as *rpoD* while a second RpoD-like protein (encoded by Rmet\_2606 situated on CHR1) was named as RpoG [114]. However, this latter annotation was cumbersome. In updating the genome sequence, it became evident that this ambiguity reflected a sequencing error by which only the carboxyterminal part of RpoG displays conserved sigma factor domains (regions 1.1 and 1.2), while the erroneous aminoterminal part does not reveal any features typical of sigma factors. In this updated version, the Rmet\_2606 protein now displays all proper characteristics of a sigma-70 housekeeping factor, and has a high homology with the *E. coli* RpoD (56% aa identity, minLrap 0.99). Consequently we propose to rename RpoG as RpoD1 and its gene Rmet\_2606 as *rpoD1*. In addition, since the Rmet\_4661 protein displays a good similarity with the *E. coli* RpoD (34% aa identity, minLrap 0.99) and clearly contains all necessary domains to be a functional

housekeeping sigma factor, we propose renaming RpoD as RpoD2 with its gene Rmet\_4661 as *rpoD2* (i.e. replacing *rpoD*).

The presence of multiple sigma-70 housekeeping factors is not unique to *C. metallidurans*: *C. eutrophus* H16 and *C. taiwanensis* both contain two sigma-70 factors (albeit both located on the main replicon) while *C. pinatubonensis* JMP134 contains three copies of this sigma factor (one on each chromosome and one on the megaplasmid) (Table S4). In our study, only *B. xenovorans* 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*, *Janthinobacterium*, *Polynucleobacter*, *B. petrii*, *D. acidovorans*, and *E. coli*).

Pairwise Smith-Waterman alignments of RpoD protein sequences for all analysed genomes (represented in Figure S5 as a heatmap) showed that *C. metallidurans* RpoD1 (Rmet\_2606) forms a tight group with orthologous RpoD proteins from the three other *Cupriavidus* species (Reut\_A0888, RALTA\_A2226 and H16\_A2725) 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 orthologs in the *Cupriavidus* family (Figure S5); indeed, it had more significant BlastP hits with the RpoD proteins of *Bordetella* and *Burkholderia* 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:FecI group (type 1) to indicate a similar function in iron assimilation as FecI in *E. coli*. Interestingly, the other three *Cupriavidus* species (*Cpin*, *Ceut*, and *Ctai*) 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 (<http://www.transcriptionfactor.org/>) [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 life-style [127,128]. Nonetheless, our analysis shows that only *Bordetella pertussis* strain Tohama I, six *Burkholderia* species, and *Streptomyces coelicor* 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)

(Table S5). Members of this family typically repress transcription from their own promoter, but activate transcription from a divergent promoter. The products of these activated genes perform a wide range of functions [130] while their activation is mediated by co-inducers, such as extracellular signaling molecules, metabolites, or ions. Other important one-component systems in *C. metallidurans* were GntR (n = 43), TetR (n = 35), AraC (n = 25), MarR (n = 16), and IclR (n = 15) (Table S5), all of which contribute to the transcriptional control of stress responses, carbon metabolism, catabolic pathways, multidrug efflux pumps, biosynthesis of antibiotics, response to osmotic stress and toxic chemicals, differentiation processes, and virulence and pathogenicity.

In terms of one-component metalloregulatory proteins in CH34, three families MerR (n = 10), ArsR (n = 6), and Fur (n = 2) deserve our attention. Although these regulators are evolutionary related in their DNA binding properties, their mode of action is different [131]. While MerR-type regulators nearly exclusively function as transcriptional activators, ArsR-type regulators generally act as repressors that upon metal binding dissociate from the DNA giving way to the RNA polymerase (coined as transcriptional derepression). Fur proteins although being typical repressors can also activate gene expression, either directly, or indirectly via the Fur-dependent repression of a small regulatory antisense RNA. Importantly, there is a wide range in metal selectivity within these three families. Metal sensing in MerR regulators is governed by the effector binding domain in their C-terminal region. Structural diversity in this domain allows MerR family proteins to sense not only various metal ions, e.g., Zn<sup>2+</sup> by ZntR, Au<sup>+</sup> by GolS, Cd<sup>2+</sup> by CadR, Pb<sup>2+</sup> by PbrR, or Hg<sup>2+</sup> by MerR, but also oxidative stress, e.g., by SoxR via an [2Fe-2S]<sup>2+</sup> cluster (reviewed in [131]). Preferential sensing by MerR regulators of one metal ion over another, or monovalent ions over divalent ions, is defined by the positioning of Cys and His but also other conserved residues, mainly in the DNA binding loop. The stringency of this preference probably varies among MerR regulators as some respond to more than one metal. In strain CH34 there are four MerR proteins and three PbrR proteins encoded by the *mer* and *pbr* loci (regulators are underlined in Table 5), and one CupR (Rmet\_3523) which responds to Cu<sup>+</sup> and Au<sup>+</sup> [35]. The remaining MerR-type regulators in CH34, including SoxR (Rmet\_4538), are not physically linked to any structural HMR gene.

The metal sensing in ArsR regulators in contrast is not defined by a single metal binding pocket but is distributed over distinct locations of the repressor involving a number of alpha helices containing key Cys or His residues [131]. Members of this family sense, next to As(III) and Sb(III), a wide variety of divalent metal ions including Zn<sup>2+</sup>, Ni<sup>2+</sup>, Co<sup>2+</sup>, Pb<sup>2+</sup>, and Cd<sup>2+</sup>. Some ArsR family members display sensitivity to multiple metal ions and are able to respond to either monovalent or multivalent ions, probably as a result of evolutionary adaptations in their protein structure particularly in the  $\alpha$ 3N and  $\alpha$ 5 helical regions. The CH34 genome enlists only one *bona fide* ArsR regulator (Rmet\_0333 of the *ars* locus on CHR1; Table 5). This repressor has been shown to dissociate from its operator DNA in a metal and valence qualifying manner [94], with three groups of heavy metals displaying decreasing degrees of ArsR dissociation: As(III)/Bi(III)/Co<sup>2+</sup>/Cu<sup>2+</sup>/Ni<sup>2+</sup>; Cd<sup>2+</sup>; and Pb<sup>2+</sup>/Zn<sup>2+</sup>. No ArsR dissociation was observed in the presence of As(V). The other five ArsR family proteins of CH34 appear to be unlinked to any HMR gene cluster except Rmet\_5676 which is located adjacent to the defective *nim* locus on CHR2 (Table 5). The two Fur proteins in CH34 are discussed in the next section on siderophores. Besides the MerR, ArsR, and Fur families, no other structural families of one-

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 *czcRS*, and one pair each of *agrRS*, *hmzRS*, *zniRS*, and *zneRS* (underlined in Table 5). From MaGe data and the DBD and MiST2 databases we discovered that the second replicon of *C. metallidurans* (CHR2) carries proportionally almost twice the number of regulatory genes as compared to the main replicon 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 staphyloferrin B [134] because it resembled the siderophores typically produced by staphylococci. It since was recognized that other organisms produce staphyloferrin B, such as *R. solanacearum* [135]. Notably, *C. metallidurans* staphyloferrin B can interact with Cd<sup>2+</sup> 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 staphyloferrin B biosynthesis (Rmet\_1109-1118) and two genes encoding ferric/ferrichrome siderophore TonB-dependent receptors BfrC and FcuA (Rmet\_1104 and Rmet\_1108, respectively) are located on CHR1 within a 52 kb region that lacks synteny with Cpin, Ceut, and Cai. Although a tRNA-Ser gene precedes 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 (*ssd*, 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 cyclodeaminase (*ocd*, Rmet\_1116), needed to biosynthesize alcaligin [137], a cysteine synthase (*cysK*, 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 *fur* 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 [138]. In both *C. metallidurans* and *R. solanacearum*, a gene (Rmet\_2977) encoding the global virulence regulator PhcA is located next to the *fur* 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* mutant 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\_2277/8/9 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 desferrioxamine B, D2 and E, desferriferrocinn, desferriferrochrysin, desferriferrochrome A, and ornibactin [133]. In addition, its genome encodes proteins involved in iron storage (*bfr*, Rmet\_0195, Rmet\_0248) and soluble ferrous Fe(II) transport (*feoB*, 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://genomics.ornl.gov/mist/>) that holds the predicted signal transduction proteins for nearly 750 bacterial and archaeal 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. arsenicoxydans* 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 synteny with *C. metallidurans* CHR1 (see section on Phylogeny), 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 replicon 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 replicon 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, IclR) 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 freeliving 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* *bugT* gene identified by Antoine *et al.* [142], was reported as the first member of a large family of periplasmic solute-binding receptors present in great numbers (up to 181 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>). Nonetheless, 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 genomes had a comparable count: *Delftia acidovorans* SPH-1 (159), *Variovorax paradoxus* S110 (137), *Verminophrobacter eiseniae* EF01-2 (131), and *Polaromonas* sp. JS666 (97). The only other genomes with a large count (>40) were *Diaphorobacter* sp. TPSY (45), *Acidovorax avenae* ssp. *citrulli* (55), *Acidovorax* JS42 (46), and *Polaromonas 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. petrii*, 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 suppurative 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

## Beta-proteobacteria

### *Burkholderiales*

#### *Alcaligenaceae*

##### *Bordetella*

- B. bronchiseptica* RB50 (183)**
- B. parapertussis* 12822 (140)**
- B. pertussis* Tohama I (79)**
- B. petrii* DSM 12804 (107)**
- B. avium* 197N (26)

#### *Burkholderiaceae*

##### *Burkholderia*

- B. xenovorans* LB400 (13)

##### *Ralstonia*

- R. solanacearum* GMI1000 (2)

##### *Polynucleobacter*

- P. necessarius* subsp. *asymbioticus* QLW (16)
- P. necessarius* subsp. *necessarius* STIR1 (8)

##### *Cupriavidus*

- C. metallidurans* CH34 (123)**
- C. eutropha* H16 (155)**
- C. pinatubonensis* JMP134 (131)**
- C. taiwanensis* (103)**

#### *Comamonadaceae*

##### *Acidovorax*

- A. citrulli* AAC00-1 (55)**
- Acidovorax* sp. JS42 (46)**

##### *Delftia*

- D. acidovorans* (159)**

##### *Diaphorobacter*

- Diaphorobacter* sp. TPSY (45)**

##### *Polaromonas*

- Polaromonas* sp. JS666 (97)**
- P. naphthalenivorans* CJ2 (41)**

##### *Variovorax*

- V. paradoxus* S110 (137)**

##### *Verminophrobacter*

- V. eiseniae* EF01-2 (131)**

#### *Oxalobacteraceae*

##### *Herminiimonas*

- Herminiimonas arsenicoxydans* (5)

##### *Janthinobacterium*

- Janthinobacterium* sp. Marseille (10)

**Figure 5. Distribution of the Bug domain.** Overrepresentation of ‘Bug’ domains in the family *Burkholderiales* according to MiST data (<http://mistdb.com/>), with numbers in parentheses. Organisms encoding more than 40 Bug domain-containing proteins are in bold. Only relevant taxa and species are shown (see text). Detailed data are in Table S8.  
doi:10.1371/journal.pone.0010433.g005

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 aminoacids (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 carboxylates naturally accumulating in the rhizosphere.

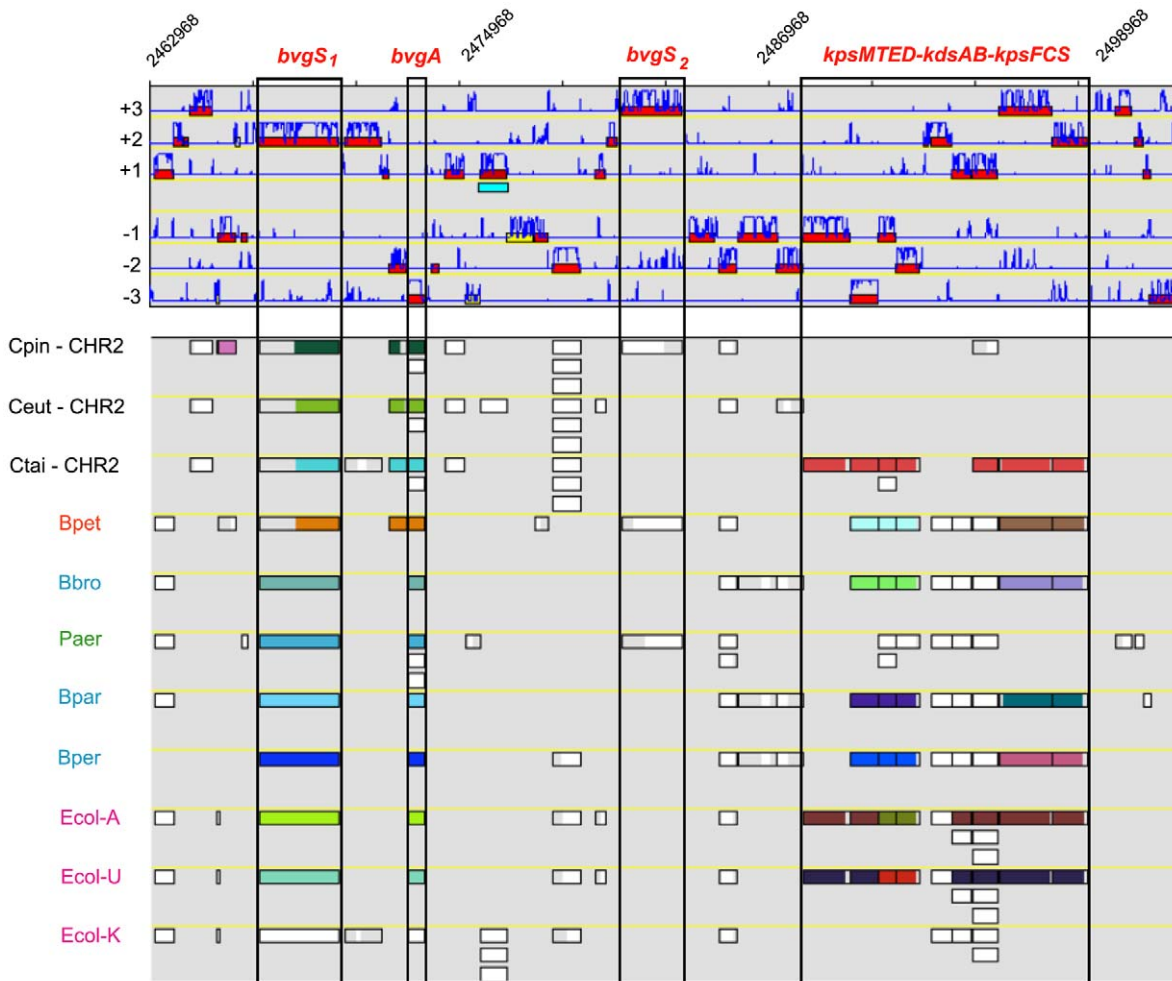
Another unusual feature of signal transduction in the *C. metallidurans* CH34 genome is the presence of an intact phosphorelay two-component system entailing a multidomain histidine-kinase sensor protein BvgS<sub>1</sub> (Rmet\_5710) and a response-transcriptional activator BvgA (Rmet\_5714). This system 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 37 EvgA-activated genes, including stationary-

phase acid resistance genes and multidrug efflux pump genes [152], 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* (58% aa identity), and displays a good similarity to EvgA and RocA, and to the BvgA regulators of pathogenic *Bordetella* species (39–45% aa identity) (all identities with minLrap >0.95). Likewise, the length and sequence of the *C. metallidurans* BvgS<sub>1</sub> 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 BvgS<sub>2</sub> (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 *bvgS*<sub>1</sub> (Rmet\_5710) to *bvgS*<sub>2</sub> (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 cassettes 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 syntenic maps are given for 11 organisms. Abbreviations: Cpin, *C. pinatubonensis* JMP134; Ceut, *C. eutrophus* H16; Ctai, *C. taiwanensis*; Bpet, *B. petrii* DSM 12804; Bbro, *Bordetella bronchiseptica* RB50; Paer, *Pseudomonas aeruginosa* PAO1; Bpar, *B. parapertussis* 12822; Bper, *B. pertussis* Tohama I; Ecol-A, avian pathogenic *Escherichia coli* (APEC01); Ecol-U, uripathogenic *E. coli* (UT189); 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, KpsC, 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, *kdsA* and *kdsB* have duplicate copies on CHR1, respectively, as Rmet\_1053 and Rmet\_0533. The genomes of *C. eutrophus* H16 and *C. pinatubonensis* JMP134 do not carry a *kps*-SSR-*kps* cassette. In fact, of all the species so far discussed in this paper, only pathogenic *Bordetella* species and *B. petrii* contain a similar cassette *kpsMT-kpsE/wcbD*-SSR-*kpsCS/wcbAO* 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 *kpsMTED*-SSR-*kpsCS* module. This module is encoded in both *S. meliloti* strains Rm41 and 2101 by two separate *rkp* 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-chained signaling molecules. In this context, the presence of a full set of *pelABCDEFGHI* 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. eutrophus* 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 (Molk2, Ipo1609, 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 (PelC and PelF), and a sucrose 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 (*ISRme8*), Rmet\_4153 (*ISRme1*), and Rmet\_4152 (*ISRme5*) 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*, to a lesser extent to the other *Cupriavidus* and *Ralstonia* species, and very intermittently 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\_5835 and Rmet\_5843, encoding proteins that respectively display a fair-to-good aa identity with the *P. aeruginosa* PslA (32%) and PslB (43%) proteins. Psl proteins in *P. aeruginosa* are implicated in the synthesis of a mannose-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 *epsAPB* (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* cassette 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 swarm and swim. Most of the *flh*, *fli*, and *flg* genes necessary for synthesizing flagella are grouped within four gene clusters: A *flh* locus Rmet\_3698-3703, two *fli* loci Rmet\_5252-66 and Rmet\_5298-5303, and the *flg* locus represented by Rmet\_3731-44. In addition, the CH34 genome encodes *cpa*, *tad*, and *pil* loci essential for synthesizing type IV pili (Tfp), allowing cells to move across solid surfaces via a mode of twitching. These include the *pilVWX* 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-tadABCD* 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 *fimAC-htrE* genes Rmet\_0574/5/6, FimA-type adhesin genes Rmet\_1667, Rmet\_4250, and Rmet\_4961, and a *fim* locus Rmet\_4250-64. Our genome survey shows that more than 150 genes may be involved in generating or controlling flagella, pili, and fimbriae. 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 *phc* genes of this system are conserved in CH34,

the other three *Cupriavidus* species, and *R. pickettii*. 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 effect 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 ( $C_4$  to  $C_{11}$ ) [9]. CH34 can derive glucose-6-phosphate from gluconeogenesis (see below) or via the phosphorylation of cytoplasmic glucose by glucokinase (Rmet\_5799). It then is further converted to 6-phosphogluconate (6GP) in a two-step reaction involving the *pgl* and *zef* genes (encoded by Rmet\_5800 and Rmet\_5801, respectively). The 6GP is metabolized through the alternative glycolytic Entner-Doudoroff (ED) pathway involving a 6-phosphogluconate dehydratase (EDD; Rmet\_5802) and a 2-keto-3-deoxy-6-phosphogluconate aldolase (EDA; Rmet\_4768), thus generating two pyruvate molecules that finally are processed to acetyl-CoA. In many prokaryotes, this acetyl-CoA is fully oxidized to  $CO_2$  via citrate, D-isocitrate, 2-ketoglutarate, succinate, fumarate, malate, and oxaloacetate in the tri carboxic-acid (TCA) cycle, that is, the Krebs cycle. However, *C. metallidurans* apparently also can use the glyoxylate bypass in which D-isocitrate is cleaved by isocitrate lyase (AceA) (Rmet\_1385, Rmet\_4737) into succinate and glyoxylate. As in full TCA cycles, succinate is converted to fumarate by succinate dehydrogenase (succinate:quinone reductase, SQR), a multimeric enzyme encoded by four genes *sdhCDAB* (Rmet\_2483/4/5/6) and further processed to malate by fumarase FumA (Rmet\_2273, Rmet\_5752). In parallel, glyoxylate condenses with acetyl-CoA yielding malate via malate synthase (AceB) (Rmet\_1390). Malate finally is changed into oxaloacetate via the malate dehydrogenase enzyme MDH (Rmet\_2489, Rmet\_5095), completing a full cycle in which acetate is again transformed to citrate by citrate synthase CitA (Rmet\_2481, Rmet\_4144, Rmet\_4268, Rmet\_4946, and Rmet\_5380) and citrate to D-isocitrate via aconitase AcnA (Rmet\_1585, Rmet\_2492, Rmet\_4240, and Rmet\_5296). We note that D-isocitrate, as for full TCA cycles, is transfigured into 2-ketoglutarate by isocitrate dehydrogenase ICDH (Rmet\_2895, Rmet\_3729). The carbon skeleton of 2-ketoglutarate then is used, depending on cellular ammonia levels and energy status, by either one of the two enzymes glutamate dehydrogenase GDH (Rmet\_1181) or gluta-

mate synthase (GOGAT) (GltAB encoded by Rmet\_3262/3) to generate glutamate that in turn feeds the nitrogen metabolic pathways. We offer three important observations here. First, although the *C. metallidurans* CH34 genome encodes the E1, E2, and E3 components of an  $\alpha$ -ketoglutarate dehydrogenase (KGDH) complex (Rmet\_2048 through Rmet\_2050), their role in the TCA cycle for converting 2-ketoglutarate to succinate via succinyl-CoA is undecided. Possibly, an alternative route is followed wherein the gene products of Rmet\_2049 and Rmet\_2050 actually form a ketoglutarate decarboxylase (KGD) complex that changes 2-ketoglutarate into the intermediate succinate semialdehyde followed by further conversion to succinate by a succinate-semialdehyde dehydrogenase (SSDH) encoded by the *gabD* genes (Rmet\_1898, Rmet\_3961, and Rmet\_4942). Interestingly, these SSDH enzymes may have a role in 4-hydroxybutyric acid catabolism, as was reported for *C. eutrophus* H16 [165]. Second, neither succinate nor oxaloacetate are produced via a reductive TCA cycle as the essential enzymes are missing (see further). And third, most TCA enzymes in *C. metallidurans* CH34 appear to have multiple isoenzymes whose genes are scattered throughout both replicons CHR1 and CHR2 (i.e., we discerned at least five citrate synthase genes, four aconitase genes, three genes for SSDH, and two genes each for ICDH, MDH, fumarase, and isocitrate lyase).

None of the sequenced *Cupriavidus* strains contains a gene for 6-phosphogluconate dehydrogenase (6-PGDH), basically ruling out the oxidative branch of the pentose-phosphate pathway. Nonetheless, all essential genes for gluconeogenesis, in which glucose is generated from non-sugar carbon substrates (such as pyruvate, lactate, oxaloacetate, and glycerol) and glucogenic amino-acids (primarily alanine and glutamine) are present in *C. metallidurans*. They include genes to transform malate via oxaloacetate (Rmet\_2489, Rmet\_2750) 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-bisphosphatase (FBP; Rmet\_0875, Rmet\_1507, Rmet\_1511) and finally into  $\beta$ -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 keto adipate via catechol and protocatechuate, respectively, as it has all the genes for both ortho-cleavage pathways (e.g., the *benABCDM* and *catABCD* genes Rmet\_4875 through Rmet\_4885, *pcaA* gene Rmet\_4270, *pcaBCDHGKQ* genes Rmet\_4011 through Rmet\_4016, and *pobAR* genes Rmet\_4017/8). The keto adipate thus formed then is processed to acetyl-CoA by the products of the *pcaTJF* genes (Rmet\_3667/8/9). In strain CH34, benzoate might be degraded anaerobically via the central metabolite benzoyl-CoA (instead of catechol; see Harwood *et al.* [166]) in an alternate pathway requiring the products of Rmet\_1220 through Rmet\_1225. Therein, the benzoyl-CoA reductase BoxAB (Rmet\_1220/1) will open the benzene ring, followed by  $\beta$ -oxidation of the resulting product (oxidation occurs at the  $\beta$ -carbon C3) leading to 3-keto adipyl-CoA, and finally to succinyl-

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,  $\beta$ -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 Harwood [167] implicated the *pimFABCDE* operon for the  $\beta$ -oxidative degradation of C<sub>7</sub>-C<sub>14</sub> dicarboxylic- and fatty-acids in *Rhodospseudomonas 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 IclR-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, FADH<sub>2</sub>, 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;  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ , 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 discerned 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 (SQR; see previous section on the TCA cycle), a *bc<sub>1</sub>*-type cytochrome complex (encoded by Rmet\_3228/9/30), a *cbb<sub>3</sub>*-type cytochrome c oxidase (encoded by *cco* genes Rmet\_2041/2/3/4), two cytochrome *bd*-type quinol oxidases CydAB (encoded by the genepairs Rmet\_4955/6 and Rmet\_5232/3), a *aa<sub>3</sub>*-type oxidase encoded by the *cta*-gene cluster (Rmet\_0261 through Rmet\_0271), a *bb<sub>3</sub>*-type cytochrome oxidase encoded by the *cox*-gene cluster (Rmet\_3955/6/7/8/9), and three quinol oxidases encoded by the clusters *cyo1* (Rmet\_0948/49/50/51), *cyo2* (Rmet\_5608/09/10/11), and *cyo3* (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 *nasAB* genes Rmet\_4820/1), a membrane-bound quinol-oxidizing enzyme (NAR; encoded by the *narGHJI* genes Rmet\_2074 through Rmet\_2077, accompanied by two ABC-type transporter-genes Rmet\_2072/3), and a periplasmic enzyme (NAP; encoded by the *napEDABC* genes Rmet\_4061 through Rmet\_4065). 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\_4817), 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 *nirS* gene (Rmet\_3172) encoding a NO-forming nitrite-reductase (NNR), the *norB* gene (Rmet\_3167) encoding a nitric-oxide reductase (NOR), and a *nosZ* 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 *nirENJ* (Rmet\_3168/69/70) and *nirCFGH* genes (Rmet\_2079/80/81, RALME2203), (ii) the *norR* gene (Rmet\_3166) encoding a  $\sigma^{54}$  dependent transcriptional regulator, and, (iii) the NOS accessory genes *nosLIFDR* (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 (RNRase): NrdAB (Rmet\_3087/8), NrdJ (Rmet\_2130), and NrdDG (Rmet\_5137, Rmet\_5139), along with NrdR (Rmet\_2681) as a transcriptional repressor of these genes. According to the RNRase database (<http://mrdb.molbio.su.se/>) less than 8% of all bacteria have all three classes of RNRase: *C. eutrophus* H16 and *C. metallidurans* CH34 are among the bacteria with these genes, but *C. pinatubonensis* JMP134 and *C. taiwanensis* lack the *nrdDG* genes. Our identification of *nrdDG* genes in strain CH34 is important because the NrdDG enzyme, which operates only under anaerobic conditions, is essential for complete denitrification in *C. eutrophus* H16 [168]. As the latter's *nrdDG* genes lie on the pHG1 plasmid, a pHG1-free mutant strain grows very poorly on nitrate and does not achieve full denitrification. This defect in the mutant is countered by adding cobalamin (VitB<sub>12</sub>) to the medium [169], most likely because the coenzyme B<sub>12</sub>-dependent ClassII RNRase, 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-relay 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 phyloquinone (VitK).

Thiamin (VitB1) biosynthesis is likely governed in *C. metallidurans* CH34 by the *thiCOSGED* 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 *dxs* (Rmet\_2615), *icsS* (Rmet\_1025), and *thiL* (Rmet\_3047) 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 pMOL28 plasmid); we did not recognize the *thiK* gene.

The biosynthesis of riboflavin (VitB2) in CH34 might be linked to the genes *ribDC* (Rmet\_2688/9), *rib(BA)E* (Rmet\_2693/4), *ribA1* (Rmet\_0194), *ribF* (Rmet\_2884), *ribA2* (Rmet\_4572), and *ribB* (Rmet\_4758). 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  $\beta$ -alanine and L-valine. In *C. metallidurans* CH34, the synthesis of VitB5 could be connected to genes *ilvE* (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), *pdxA* (present in two copies Rmet\_0438, Rmet\_2590), a gene encoding an FMN-binding protein (Rmet\_0820), *pdxF* (Rmet\_2415), *dxs* (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 pimeloyl-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 and 3-hydroxypimeloyl-CoA, while the direct use of pimelate necessitates a fifth gene, *pauA* (pimelic acid-utilizing A; Rmet\_2598). The *pauA* gene codes for a 6-carboxyhexanoate-CoA ligase (often referred to as a 'pimeloyl-CoA synthetase'), essentially 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\_1820), *folP* (Rmet\_2187), *folC* (Rmet\_2463), *folA* (Rmet\_2569), *purX* (Rmet\_2878), *folK* (Rmet\_2915), *tpgG* (Rmet\_3179), and a putative *folE* (Rmet\_3990). We failed to identify a hydrolase gene for converting 7,8-dihydroneopterin 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 sirohydrochlorin cobaltochelate gene *cbiX* (Rmet\_2810) 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 *cbi*, *cob* and *btu* genes Rmet\_2777 through Rmet\_2789) with a cobalamin riboswitch we identified at the far end (Rmet\_R0047). This entire operon, which according to the MaGe interface displays good synteny with all *Cupriavidus* and *Ralstonia* genomes, contains all the essential genes for producing adenosylcobalamin from cobyrinic acid (including those controlling corrinoid transport, viz. *btuB*, *btuC*, *btuD*, and *btuF*). Only the *bluB* gene (Rmet\_2213), coding for a cobalt reductase, is at a different location on CHR1. Whilst the early and late stages of the biosynthesis generally are covered by these listed genes, other genes involved in the intermediate corrin biosynthesis in *C. metallidurans* 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 (rTCA) 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 *cbf* 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 *cbfL* (Rmet\_1501) and *cbfS* (Rmet\_1500) of RubisCo (ribulose 1-5 biphosfaat carboxylase oxygenase) as well as of the accessory genes *cbfQ* (Rmet\_1499) and *cbfO* (Rmet\_1498) needed to assemble RubisCo.

Several regulators control the CBB cycle. First, there are three copies of *cbfR*, a LysR-type transcriptional activator of *cbf* gene expression. Two copies juxtapose the central *cbf* gene clusters, *cbfR1* (Rmet\_1502) and *cbfR2* (Rmet\_1508), while a third copy, *cbfR3* (Rmet\_4365), is located on the second replicon CHR2 (unlike other *cbf* genes that all are 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 *cbf*-genes appear to be transcriptionally linked to the *cbf* operon: *cbfY* (Rmet\_1517), *cbfZ1* (Rmet\_1514), and *cbfZ2* (Rmet\_3177). Under certain conditions, autotrophic bacteria carrying these exceptional *cbf* 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 *cbfX* 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 (CoxS), a flavoprotein (CoxM), and a molybdoprotein (CoxL). 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 *coxSLM* (Rmet\_0365 – Rmet\_0363) and the accessory genes *coxDEGI* (Rmet\_0362 – Rmet\_0359). The order of the CODH genes in this operonic structure and the presence of an AYRGAGR motif in CoxL (encoded by Rmet\_0364) strongly suggest that the *C. metallidurans*' CODH is of form II. Although the ability of *C. metallidurans* CH34 to undertake substantial oxidation of carbon monoxide still needs to be tested, CODH-II-encoding *cox* genes were found in *Bradyrhizobium japonicum* USDA 110, an organism relying solely on CODH-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 lyase, 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 CODH-II 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 oxygenase reductase genes (*sor*) (Figure S7A) [181]. The first set of genes comprises the *soxBWXA<sub>2</sub>YFE* 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: SoxYZ, a sulfur carrier protein, SoxXA, a c-type cytochrome complex, and SoxB, a sulfate thiol hydrolase. In this system, the oxidation of thiosulfate ( $S_2O_3^{2-}$ ) 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 ( $SO_3^-$ ) group by SoxB (Figure S7B). The SoxXA complex then covalently binds other pathway intermediates to the SoxY cysteine thus forming extended chains of sulfane ( $S^-$ ) groups preceding the terminal sulfone group ( $SO_3^-$ ) [182]. These sulfane 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 oxydise 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 S7C). 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 ( $SO_3^-$ ) through the action of the *sorA* and *sorB* genes (present on CHR2 as Rmet\_4891 and Rmet\_4892) (Figure S7D) that are orthologues of the *C. eutrophus* H16 *sorA* and *sorB* genes encoding, respectively, a periplasmic sulfite dehydrogenase and a membrane-bound c-type cytochrome [71].

These routes of sulfur oxidation (SoxB-XA-YZ, SoxEF, and SorAB) 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 SsuABC (Rmet\_1370/2/3) and TauABC (Rmet\_2857/8/9), an alkanesulfate monooxygenase SsuD (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 SsuE (Rmet\_5075) that with AsfA/AsfB may form an electron chain for aromatic desulfonation. We did not identify other genes for enzymatic desulfonation 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 S8A). The first cluster, governing MBH, harbors the *h<sub>2</sub>* hydrogenase *h<sub>2</sub>* operon encompassing six genes *hypE<sub>1</sub>D<sub>1</sub>C<sub>1</sub>F<sub>1</sub>B<sub>1</sub>A<sub>1</sub>* (Rmet\_1281 – Rmet\_1286), followed by a set of genes required for *h<sub>2</sub>* hydrogen oxidation, *hoxVTRQOLM<sub>2</sub>ZGK* (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  $\Delta hoxV$  (Rmet\_1534), nine intact *hox* genes *hoxFUYHWIXA* (Rmet\_1522 through Rmet\_1527, Rmet\_1533, and Rmet\_1541-42), and six genes *hypA<sub>2</sub>B<sub>2</sub>F<sub>2</sub>C<sub>2</sub>D<sub>2</sub>E<sub>2</sub>* (Rmet\_1535 through Rmet\_1540) all on genomic island CMGI-3. The latter group of *hyp* genes (index 2) appears to have arisen from a reverse duplication of the *hyp* operon of the first *hyp* cluster (index 1) on CMGI-2. We note that the *hyp* and *hox* genes for SH are located, together with the *cbf* 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 *hyfBCEFGI* (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 hydrogenase 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 4Fe-4S 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 HoxHUFIK, with HoxY holding the  $e^-$  transferring 4Fe-4S 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  $H_2$  instantaneously as a reducing power and hence as an alternative source of energy when nutrients are scarce indicates that chemolithoautotrophic 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  $H_2$  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,  $H_2$  affinity, and  $O_2$  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].

## Organoautotrophy 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 acetoacetate 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  $\beta$ -oxidation of fatty acids and dicarboxylates, or PHB metabolism) and has two CoATases 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 ( $\alpha_2\beta_2\gamma_2$ ) 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 acetone 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* Py2 [192]. Both in *X. autotrophicus* and *C. metallidurans* the AcxR protein contains a SELFGXXXGAFTGA motif that is typical for  $\sigma^{54}$  subunit binding regulators [193], and there is a clearly identifiable  $\sigma^{54}$  promoter sequence TGGCATGNNNNTTGC in the *acxRA* 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 *acxABC* 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  $\sigma^{54}$  dependent transcriptional activator, but for a LuxR/FixJ 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 isopropanol dehydrogenase (iPDH) is Rmet\_5645 because its product is highly similar in length and sequence to the multifunctional ADH of *C. glutamicum* H16 (H16\_A0757) [194]. In addition, the CH34 genome contains six NPD encoding genes (Rmet\_0222, Rmet\_2083, Rmet\_2605, 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.

Benzene, toluene (methylbenzene), ethylbenzene, and xylene (dimethylbenzene) jointly are known as BTEX compounds, with phenol as a major chemical intermediate. In *C. metallidurans* CH34 aerobic degradation of BTEX involves ring hydroxylation by one of three different multicomponent monooxygenases 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 monooxygenases (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], Cafaro *et al.* [196] suggested that multiple BMMs in the same system may be beneficiary in terms of BMM specificity and efficiency of substrate utilization. We identified two Group 1 (phenol hydroxylases; PHs) and one Group 2 (toluene-benzene monooxygenases; 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 monooxygenase, encoded by six genes Rmet\_1311 to Rmet\_1316. All genes are present for the *meta*-cleavage of catechol (by an 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 (*patR*) was identified as a  $\sigma^{54}$ -dependent Fis-type transcriptional regulator. This entire region is in high synteny with the aromatic degradation gene clusters in the related genomes of *C. pinatubonensis* JMP134 (Reut\_B5671 through Reut\_B5693) 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\_1788 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 *Wautersia numadzuensis* strain TE26 (as the *phlRABCDEF-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 GBI 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. glutamicum* H16 which encompasses only three enzymes: A PHA synthase (or polymerase) encoded by *phaC*; an acetoacetyl-CoA reductase encoded by *phaB*; and, a  $\beta$ -ketothiolase encoded by *phaA*. The synthesis of PHB has been studied most extensively in *C. glutamicum* H16 because it can generate PHB solely using simple carbon substances, such as gluconate, via the Entner-Doudoroff (ED) pathway [45]. Additional genes playing a role in PHB synthesis are *phaR* encoding a transcriptional regulator, and *phaP* 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. glutamicum* 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 *phaP*, three copies of *phaB*, 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* (*bktB*, *atoB*) encoding additional acetoacetyl-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. glutamicum* H16 identified by York *et al.* [201]. In addition, the CHR1 replicon carries a gene for a PHB-oligomer hydrolase (designated by *phaY* 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 *phaC2* gene (Rmet\_1671) is predicted by the MaGe system to encode a defective  $\gamma$ -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*, *mutY*, *tag*, *ung*, *fpg*, *nth*, *xthA*) (all Rmet numbers are given in Table S10) and the complete pathway for nucleotide-excision repair (*mfd*, *uvrA*, *uvrB*, *uvrC*, *uvrD*). 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  $\beta$ - and  $\alpha$ -proteobacteria and firmicutes [207]. The CH34 genome carries the genes encoding the Holliday-junction protein complex RuvAB, the resolvase RuvC, a dsDNA exonuclease SbcD, and the mismatch-repair proteins MutS and MutL. These genes *ruvAB*, *ruvC*, *sbcD*, 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 (*relA* and *spoT*), the main regulatory genes of the SOS response (*lexA* and *recA*), and the oxidative-stress-response regulators, OxyR and SoxR (all Rmet numbers in Table S10). *C. metallidurans* CH34, however, lacks a SoxS homolog and does not carry any typical SoxR 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 SoxR-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.belspo.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 (fosmids) 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 Paracel Genome Assembler (PGA; <http://www.paracel.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 genetic experimental 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 [48]. 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.cns.fr/agg/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 scrutinised 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 normalised 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 *gplots* 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  $[1 - \text{number of orthologs}/\text{size of the smallest genome}]$  [217]. For the clustering of different RpoD proteins (Figure S5) we used the distance measure  $[100 - \text{the percentage of protein similarity}]$ .

### 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  $10^{-10}$  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 TTATCCACA 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. Found at: doi:10.1371/journal.pone.0010433.s001 (0.05 MB PDF)

**Figure S2** Genome dynamics in the second chromosome of *Cupriavidus* and *Ralstonia*. 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: Cmet, *C. metallidurans* CH34; Cpin, *C. pinatubonensis* JMP134; Ceut, *C. eutrophus*; Ctai, *C. taiwanensis*. Found at: doi:10.1371/journal.pone.0010433.s002 (1.86 MB TIF)

**Figure S3** Similarity matrix between CH34 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. pinatubonensis* JMP134; Ceut, *C. eutrophus* H16; Ctai, *C. taiwanensis*; Ecol, *Escherichia coli* K-12; Bxen, *Burkholderia xenovorans* LB400; Bpet, *Bordetella petrii*; Jmar, *Janthinobacterium* sp. Marseille; Hars, *Hermiimonas arsenicoxydans*; Daci, *Delftia acidovorans* SPH-1; Rsol, *Ralstonia solanacearum* GMI1000; and Pnec, *Polynucleobacter necessarius* STIR1. Replicons are denoted by their commonly known name (e.g., pRALTA) or subindexed (e.g., Cmet\_CHR1).

Found at: doi:10.1371/journal.pone.0010433.s003 (2.72 MB TIF)

**Figure S4** Protein-based synteny between the *C. metallidurans* CH34 chromosomes and the genome of *P. necessarius* STIR1, an obligate symbiont of freshwater protozoans. (A) Synteny displays were generated by the LinePlot tool of the MaGe system using a synten 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.

Found at: doi:10.1371/journal.pone.0010433.s004 (0.18 MB PDF)

**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. pinatubonensis* JMP134; RALTA, *C. taiwanensis*; H16, *C. eutrophus* H16; Bpet, *B. petrii*; Daci, *D. acidovorans* SPH-1; Pnec, *P. necessarius* STIR1; mma, *Janthinobacterium* sp. Marseille; HEAR, *H. arsenicoxydans*; RSc, *R. solanacearum* GMI1000; b3067 denotes the RpoD protein sequence of *E. coli* str. K-12 substr. MG1655.

Found at: doi:10.1371/journal.pone.0010433.s005 (0.73 MB PDF)

**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).

Found at: doi:10.1371/journal.pone.0010433.s006 (0.07 MB PDF)

**Figure S7** Genes and pathways for sulfur oxidation (sulfur autotrophy) in *C. metallidurans* CH34. (A) Genetic loci encompass-

ing sulfur oxidizing genes (*sox*) and sulfur oxygenase reductase genes (*sor*). Note that strain CH34 lacks the *soxCD* 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). (B) Proposed pathway for the oxidation of thiosulfate ( $S_2O_3^{2-}$ ) to sulfate ( $SO_4^-$ ) by the combined action of the sulfur carrier protein SoxYZ, a cytochrome c-type complex SoxXA, and the sulfate thiol hydrolase SoxB. (C) Proposed pathway for the oxidation of sulfide ( $HS^-$ ) by the SoxEF c-type flavocytochrome complex in which SoxF acts as a sulfide dehydrogenase. (D) Proposed pathway for the oxidation of sulfite ( $SO_3^-$ ) to sulfate ( $SO_4^-$ ) through the action of a periplasmic sulfite dehydrogenase, SorA, and a membrane-bound c-type cytochrome, SorB.

Found at: doi:10.1371/journal.pone.0010433.s007 (0.05 MB XLS)

**Figure S8** Genetic loci in *C. metallidurans* CH34 for the oxidation of hydrogen. (A) *hox* and *hyp* genes needed to form and activate the two hydrogenases of CH34. The first cluster of genes governs the membrane-bound hydrogenase (MBH) and is sited on CHR1 as part of the CMGI-2 genomic island. The second gene cluster encodes the cytoplasmic hydrogenase (SH) and is sited on CHR2 as part of the genomic island CMGI-3. (B) Multimeric composition of the MBH and SH metalloproteins and their cellular location in *C. metallidurans* CH34. While the MBH enzyme is translocated to the periplasmic space and transfers electrons in one direction, the SH remains inside the cytoplasm and is able to transfer electrons in both directions, depending on NADH to  $NAD^+$  ratios in the cell. As such, the SH enzyme effectively acts as a 'redox valve'.

Found at: doi:10.1371/journal.pone.0010433.s008 (0.06 MB XLS)

**Table S1** Functional classification. Normalised distribution of functional content over the two large replicons of *C. metallidurans* CH34, based on COG assignments. These data were used for colorcoding in Figure 1 and the drawing of Figure 2.

Found at: doi:10.1371/journal.pone.0010433.s009 (0.08 MB PDF)

**Table S2** Novel HMR genes in *Cupriavidus metallidurans* CH34. Correspondance between HMR loci reported here with HMR loci reported in an earlier overview [8]. Novel genes are given in bold letter type.

Found at: doi:10.1371/journal.pone.0010433.s010 (0.02 MB XLS)

**Table S3** Efflux and transport proteins in *C. metallidurans* CH34. Transport proteins predicted by TransAAP analysis for the four *Cupriavidus* species by submitting the genomes of these species to the TransportDB website (<http://www.membranetransport.org/>). Only five classes were considered (indicated with blue bars). Transporters that are not present in *C. metallidurans* CH34 but are present in one or more of the other *Cupriavidus* strains are shaded in gray. Reversely, transporters that are present in CH34 but are not in either of the three other strains are shaded in green.

Found at: doi:10.1371/journal.pone.0010433.s011 (0.03 MB XLS)

**Table S4** Sigma factors of *C. metallidurans* CH34. The 17 sigma factors of *C. metallidurans* CH34 and their orthologs in related organisms (as listed in Table 2) and *E. coli* K-12. Note that some organisms, including strain CH34, possess multiple sigma-70 factors (RpoD proteins). All gene notations are from GenBank (and are as in Figure S5).

Found at: doi:10.1371/journal.pone.0010433.s012 (0.02 MB XLS)

**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.

Found at: doi:10.1371/journal.pone.0010433.s013 (0.62 MB XLS)

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

Found at: doi:10.1371/journal.pone.0010433.s014 (0.12 MB DOC)

**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.

Found at: doi:10.1371/journal.pone.0010433.s015 (0.02 MB XLS)

**Table S8** Prevalence of Bug domains according to the MiST2 database. Distribution of 2,377 proteins with Bug domains across 1,216 complete and draft, bacterial and archaeal genomes (source: [http://mistdb.com](http://mistdb.com/)). Organisms without Bug domain-containing proteins are omitted from the table.

## References

- Fergusson JE (1990) The Heavy Elements: Chemistry, Environmental Impact, and Health Effects. Oxford, UK: Pergamon Press.
- Adriano DC (2001) Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals. New-YorkUS: Springer-Verlag.
- Mergeay M, Houba C, Gerits J (1978) Extrachromosomal inheritance controlling resistance to cadmium, cobalt, copper and zinc ions: evidence from curing in a *Pseudomonas* [proceedings]. Arch Int Physiol Biochim 86: 440–442.
- Diels L, Faelen M, Mergeay M, Nies D (1985) Mercury transposons from plasmids governing multiple resistance to heavy metals in *Alcaligenes eutrophus* CH34. Arch Intern Physiol Bioch 93: 27–28.
- Juhnke S, Peitzsch N, Hubener N, Grosse C, Nies DH (2002) New genes involved in chromate resistance in *Ralstonia metallidurans* strain CH34. Arch Microbiol 179: 15–25.
- Mergeay M (2000) Bacteria adapted to industrial biotopes: the metal resistant *Ralstonia*. In: Hengge-Aronis R, Storz G, eds. Bacterial Stress Responses. Washington, DC: ASM Press. pp 403–414.
- Mergeay M, Monchy S, Janssen P, Van Houdt R, Leys N (2009) Megaplastids in *Cupriavidus* Genus and Metal Resistance. In: Schwartz E, ed. Microbial Megaplastids. Berlin: Springer Berlin/Heidelberg. pp 209–238.
- Mergeay M, Monchy S, Vallaeys T, Auquier V, Benotmane A, et al. (2003) *Ralstonia metallidurans*, a bacterium specifically adapted to toxic metals: towards a catalogue of metal-responsive genes. FEMS Microbiol Rev 27: 385–410.
- Mergeay M, Nies D, Schlegel HG, Gerits J, Charles P, et al. (1985) *Alcaligenes eutrophus* CH34 is a facultative chemolithotroph with plasmid-bound resistance to heavy metals. J Bacteriol 162: 328–334.
- Monchy S, Benotmane MA, Janssen P, Vallaeys T, Taghavi S, et al. (2007) Plasmids pMOL28 and pMOL30 of *Cupriavidus metallidurans* are specialized in the maximal viable response to heavy metals. J Bacteriol 189: 7417–7425.
- Nies A, Nies DH, Silver S (1989) Cloning and expression of plasmid genes encoding resistances to chromate and cobalt in *Alcaligenes eutrophus*. J Bacteriol 171: 5065–5070.
- Nies A, Nies DH, Silver S (1990) Nucleotide sequence and expression of a plasmid-encoded chromate resistance determinant from *Alcaligenes eutrophus*. J Biol Chem 265: 5648–5653.
- Nies DH, Silver S (1989) Plasmid-determined inducible efflux is responsible for resistance to cadmium, zinc, and cobalt in *Alcaligenes eutrophus*. J Bacteriol 171: 896–900.
- Reith F, Rogers SL, McPhail DC, Webb D (2006) Biomineralization of gold: biofilms on bacteriiform gold. Science 313: 233–236.
- Novick RP, Roth C (1968) Plasmid-linked resistance to inorganic salts in *Staphylococcus aureus*. J Bacteriol 95: 1335–1342.
- Nies DH (1999) Microbial heavy-metal resistance. Appl Microbiol Biotechnol 51: 730–750.
- Silver S, Misra TK (1988) Plasmid-mediated heavy metal resistances. Annu Rev Microbiol 42: 717–743.
- Silver S, Phung LT (1996) Bacterial heavy metal resistance: new surprises. Annu Rev Microbiol 50: 753–789.
- Lejeune P, Mergeay M, Van Gijsegem F, Faelen M, Gerits J, et al. (1983) Chromosome transfer and R-prime plasmid formation mediated by plasmid pULB113 (RP4::mini-Mu) in *Alcaligenes eutrophus* CH34 and *Pseudomonas fluorescens* 6.2. J Bacteriol 155: 1015–1026.
- Van Gijsegem F, Toussaint A (1982) Chromosome transfer and R-prime formation by an RP4::mini-Mu derivative in *Escherichia coli*, *Salmonella typhimurium*, *Klebsiella pneumoniae*, and *Proteus mirabilis*. Plasmid 7: 30–44.
- Mergeay M, Lejeune P, Sadouk A, Gerits J, Fabry L (1987) Shuttle transfer (or retrotransfer) of chromosomal markers mediated by plasmid pULB113. Mol Gen Genet 209: 61–70.

Found at: doi:10.1371/journal.pone.0010433.s016 (0.07 MB XLS)

**Table S9** Functional distribution and chromosomal location of *C. metallidurans* CH34 genes involved in polyester (PHA/PHB) biosynthesis and degradation.

Found at: doi:10.1371/journal.pone.0010433.s017 (0.06 MB DOC)

**Table S10** *C. metallidurans* CH34 genes for DNA repair and stress-related regulators.

Found at: doi:10.1371/journal.pone.0010433.s018 (0.08 MB DOC)

**Table S11** *C. metallidurans* CH34 genes encoding universal stress proteins (UspA superfamily) that are induced by a variety of starvation- and stress-conditions.

Found at: doi:10.1371/journal.pone.0010433.s019 (0.04 MB DOC)

## Acknowledgments

We thank A. D. Woodhead for excellent scientific editing of the manuscript and D. Vallenet and Z. Rouy of Génoscope (Centre National de Séquençage, Evry, France) for implementing additional features in MaGe and their essential advice in genome annotation. We are also grateful to I. T. Paulsen and Q. Ren for advice and data analyses using TransAAP, and to L. Ulrich for providing SQL-derived data from the Microbial Signal Transduction database (MistDB and MiST2).

## Author Contributions

Wrote the paper: PJJ RVH HM PM NM NL AL SM ST DvdL MM. Coordinated the genome annotation: PJJ. Contributed to genome annotation: RVH HM PM NM SM MM. Performed genome data analysis: PJJ. Analysed genome information: PJJ RVH HM PM NM SM CM ST DvdL MM. Organised, reviewed, and edited the manuscript: PJJ. Performed bioinformatics: PJJ PM CM. Performed associated miscellaneous bioinformatics: SMcC. Performed genetic experiments and the sequencing and finishing of the genome: AM MAB TV AL SM ST JD. Maintained and updated the Cupriavidu2Scope project with her team: CM. Developed and applied methods to identify and correct misassemblies in the original shotgun assembly: SMcC.

22. Sadouk A, Mergeay M (1993) Chromosome mapping in *Alcaligenes eutrophus* CH34. *Mol Gen Genet* 240: 181–187.
23. Diels L, Dong Q, van der Lelie D, Baeyens W, Mergeay M (1995) The *czc* operon of *Alcaligenes eutrophus* CH34: from resistance mechanism to the removal of heavy metals. *J Ind Microbiol* 14: 142–153.
24. De Rore H, Top E, Houwen F, Mergeay M, Verstraete W (1994) Evolution of heavy metal resistant conjugants in soil environments with a concomitant selective pressure. *FEMS Microbiol Ecol* 14: 163–174.
25. Top E, De Rore H, Collard J-M, Gellens V, Slobodkina G, et al. (1995) Retromobilization of heavy metal resistance genes in unpolluted and heavy metal polluted soil. *FEMS Microbiol Ecol* 18: 191–203.
26. Top E, Mergeay M, Springael D, Verstraete W (1990) Gene escape model: transfer of heavy metal resistance genes from *Escherichia coli* to *Alcaligenes eutrophus* on agar plates and in soil samples. *Appl Environ Microbiol* 56: 2471–2479.
27. Tauch A, Schneiker S, Selbitschka W, Puhler A, van Overbeek LS, et al. (2002) The complete nucleotide sequence and environmental distribution of the cryptic, conjugative, broad-host-range plasmid pIPO2 isolated from bacteria of the wheat rhizosphere. *Microbiology* 148: 1637–1653.
28. Top E, De Smet I, Verstraete W, Dijkmans R, Mergeay M (1994) Exogenous Isolation of Mobilizing Plasmids from Polluted Soils and Sludges. *Appl Environ Microbiol* 60: 831–839.
29. Van der Auwera GA, Krol JE, Suzuki H, Foster B, Van Houdt R, et al. (2009) Plasmids captured in *C. metallidurans* CH34: defining the PromA family of broad-host-range plasmids. *Antonie Van Leeuwenhoek* 96: 193–204.
30. van Elsas JD, Gardener BB, Wolters AC, Smit E (1998) Isolation, characterization, and transfer of cryptic gene-mobilizing plasmids in the wheat rhizosphere. *Appl Environ Microbiol* 64: 880–889.
31. Diels L, Springael D, van der Lelie N, Top E, Mergeay M (1993) Use of DNA probes and plasmid capture in a search for new interesting environmental genes. *Sci Total Environ* 139–140: 471–478.
32. Goris J, De Vos P, Coenye T, Hoste B, Janssens D, et al. (2001) Classification of metal-resistant bacteria from industrial biotopes as *Ralstonia campinensis* sp. nov., *Ralstonia metallidurans* sp. nov. and *Ralstonia basiliensis* Steinle et al. 1998 emend. *Int J Syst Evol Microbiol* 51: 1773–1782.
33. Vandamme P, Coenye T (2004) Taxonomy of the genus *Cupriavidus*: a tale of lost and found. *Int J Syst Evol Microbiol* 54: 2285–2289.
34. Vanechoutte M, Kampfer P, De Baere T, Falsen E, Verschraegen G (2004) *Wautersia* gen. nov., a novel genus accommodating the phylogenetic lineage including *Ralstonia eutropha* and related species, and proposal of *Ralstonia* [*Pseudomonas*] *syzygii* (Roberts et al. 1990) comb. nov. *Int J Syst Evol Microbiol* 54: 317–327.
35. Jian X, Wasinger EC, Lockard JV, Chen LX, He C (2009) Highly Sensitive and Selective Gold(I) Recognition by a Metalloregulator in *Ralstonia metallidurans*. *J Am Chem Soc* 131: 10869–10871.
36. Roux M, Sarret G, Pignot-Paintrand I, Fontecave M, Coves J (2001) Mobilization of selenite by *Ralstonia metallidurans* CH34. *Appl Environ Microbiol* 67: 769–773.
37. Taghavi S, Lesaulnier C, Monchy S, Wattiez R, Mergeay M, et al. (2009) Lead(II) resistance in *Cupriavidus metallidurans* CH34: interplay between plasmid and chromosomally-located functions. *Antonie Van Leeuwenhoek* 96: 171–182.
38. Nies DH (2003) Efflux-mediated heavy metal resistance in prokaryotes. *FEMS Microbiol Rev* 27: 313–339.
39. Diels L, Van Roy S, Taghavi S, Van Houdt R (2009) From industrial sites to environmental applications with *Cupriavidus metallidurans*. *Antonie Van Leeuwenhoek* 96: 247–258.
40. Hynninen A, Touze T, Pitkanen L, Mengin-Lecreulx D, Virta M (2009) An efflux transporter PbrA and a phosphatase PbrB cooperate in a lead-resistance mechanism in bacteria. *Mol Microbiol* 74: 384–394.
41. De Gelder J, Vandenabeele P, De Boever P, Mergeay M, Moens L, et al. (2009) Raman Spectroscopic Analysis of *Cupriavidus metallidurans* LMG 1195 (CH34) Cultured in Low-shear Microgravity Conditions. *Microgravity Science and Technology* 21: 217–223.
42. Leys N, Baatout S, Rosier C, Dams A, s'Heeren C, et al. (2009) The response of *Cupriavidus metallidurans* CH34 to spaceflight in the international space station. *Antonie Van Leeuwenhoek* 96: 227–245.
43. Moissl C, Osman S, La Duc MT, Dekas A, Brodie E, et al. (2007) Molecular bacterial community analysis of clean rooms where spacecraft are assembled. *FEMS Microbiol Ecol* 61: 509–521.
44. Saró MI, Garcia AM, Moreno DA (2005) Biofilm formation in spent nuclear fuel pools and bioremediation of radioactive water. *Int Microbiol* 8: 223–230.
45. Pohlmann A, Fricke WF, Reinecke F, Kusian B, Liesegang H, et al. (2006) Genome sequence of the bioplastic-producing “Knallgas” bacterium *Ralstonia eutropha* H16. *Nat Biotechnol* 24: 1257–1262.
46. Amadou C, Pascal G, Mangenot S, Glew M, Bontemps C, et al. (2008) Genome sequence of the beta-rhizobium *Cupriavidus taiwanensis* and comparative genomics of rhizobia. *Genome Res* 18: 1472–1483.
47. Antoine R, Jacob-Dubuisson F, Drobeq H, Willery E, Lesjean S, et al. (2003) Overrepresentation of a gene family encoding extracytoplasmic solute receptors in *Bordetella*. *J Bacteriol* 185: 1470–1474.
48. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, et al. (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* 34: 53–65.
49. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 29: 22–28.
50. Nicholson TL (2007) Construction and validation of a first-generation *Bordetella bronchiseptica* long-oligonucleotide microarray by transcriptional profiling the Bvg regulon. *BMC Genomics* 8: 220.
51. Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* 26: 2286–2290.
52. Sato Y, Nishihara H, Yoshida M, Watanabe M, Rondal JD, et al. (2006) *Cupriavidus pinatubonensis* sp. nov. and *Cupriavidus lahariae* sp. nov., novel hydrogen-oxidizing, facultatively chemolithotrophic bacteria isolated from volcanic mudflow deposits from Mt. Pinatubo in the Philippines. *Int J Syst Evol Microbiol* 56: 973–978.
53. Springer N, Amann R, Ludwig W, Schleifer KH, Schmidt H (1996) *Polymycolobacter necessarius*, an obligate bacterial endosymbiont of the hypotrichous ciliate *Euplates aediculatus*, is a member of the beta-subclass of Proteobacteria. *FEMS Microbiol Lett* 135: 333–336.
54. Fricke WF, Kusian B, Bowien B (2009) The genome organization of *Ralstonia eutropha* strain H16 and related species of the *Burkholderiaceae*. *J Mol Microbiol Biotechnol* 16: 124–135.
55. Egan ES, Fogel MA, Waldor MK (2005) Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes. *Mol Microbiol* 56: 1129–1138.
56. Enright AJ, Ouzounis CA (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 16: 451–457.
57. Gevers D, Vandepoele K, Simillon C, Van de Peer Y (2004) Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* 12: 148–154.
58. Nies DH, Rehbein G, Hoffmann T, Baumann C, Grosse C (2006) Paralogs of genes encoding metal resistance proteins in *Cupriavidus metallidurans* strain CH34. *J Mol Microbiol Biotechnol* 11: 82–93.
59. Van Houdt R, Monchy S, Leys N, Mergeay M (2009) New mobile genetic elements in *Cupriavidus metallidurans* CH34, their possible roles and occurrence in other bacteria. *Antonie Van Leeuwenhoek* 96: 205–226.
60. Klockgether J, Wurdemann D, Reva O, Wiehlmann L, Tummeler B (2006) Diversity of the abundant pKLC102/PAGI-2 family of genomic islands in *Pseudomonas aeruginosa*. *J Bacteriol* 189: 2443–2459.
61. Larbig KD, Christmann A, Johann A, Klockgether J, Hartsch T, et al. (2002) Gene islands integrated into tRNA(Gly) genes confer genome diversity on a *Pseudomonas aeruginosa* clone. *J Bacteriol* 184: 6665–6680.
62. Mathee K, Narasimhan G, Valdes C, Qiu X, Mawish JM, et al. (2008) Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proc Natl Acad Sci U S A* 105: 3100–3105.
63. Coenye T, Spilker T, Reik R, Vandamme P, Lipuma JJ (2005) Use of PCR analyses to define the distribution of *Ralstonia* species recovered from patients with cystic fibrosis. *J Clin Microbiol* 43: 3463–3466.
64. Merlin C, Springael D, Mergeay M, Toussaint A (1997) Organisation of the bph gene cluster of transposon Tn4371, encoding enzymes for the degradation of biphenyl and 4-chlorobiphenyl compounds. *Mol Gen Genet* 253: 499–506.
65. Toussaint A, Merlin C, Monchy S, Benotmane MA, Leplac R, et al. (2003) The biphenyl- and 4-chlorobiphenyl-catabolic transposon Tn4371, a member of a new family of genomic islands related to IncP and Ti plasmids. *Appl Environ Microbiol* 69: 4837–4845.
66. Merlin C, Springael D, Toussaint A (1999) Tn4371: A modular structure encoding a phage-like integrase, a *Pseudomonas*-like catabolic pathway, and RP4/Ti-like transfer functions. *Plasmid* 41: 40–54.
67. Ryan MP, Pembroke JT, Adley CC (2009) Novel Tn4371-ICE like element in *Ralstonia pickettii* and genome mining for comparative elements. *BMC Microbiol* 9: 242.
68. Schwartz E, Henne A, Cramm R, Eitingier T, Friedrich H, et al. (2003) Complete nucleotide sequence of pHG1: a *Ralstonia eutropha* H16 megaplasmid encoding key enzymes of H(2)-based lithoautotrophy and anaerobiosis. *J Mol Biol* 332: 369–383.
69. Rosch V, Denger K, Schleich D, Smits TH, Cook AM (2008) Different bacterial strategies to degrade taurocholate. *Arch Microbiol* 190: 11–18.
70. Schleich D, Knepper TP, Fischer K, Cook AM (2004) Mineralization of individual congeners of linear alkylbenzenesulfonate by defined pairs of heterotrophic bacteria. *Appl Environ Microbiol* 70: 4053–4063.
71. Denger K, Weinitzschke S, Smits TH, Schleich D, Cook AM (2008) Bacterial sulfite dehydrogenases in organotrophic metabolism: separation and identification in *Cupriavidus necator* H16 and in *Deffia acidovorans* SPH-1. *Microbiology* 154: 256–263.
72. Haines AS, Jones K, Batt SM, Kosheleva IA, Thomas CM (2007) Sequence of plasmid pBS228 and reconstruction of the IncP-1 $\alpha$  phylogeny. *Plasmid* 58: 76–83.
73. Stokes HW, Elbourne LD, Hall RM (2007) Tn1403, a multiple-antibiotic resistance transposon made up of three distinct transposons. *Antimicrob Agents Chemother* 51: 1827–1829.
74. Tseng SP, Hsueh PR, Tsai JC, Teng IJ (2007) Tn6001, a transposon-like element containing the blaVIM-3-harboring integron In450. *Antimicrob Agents Chemother* 51: 4187–4190.
75. Monchy S, Benotmane MA, Wattiez R, van Aelst S, Auquier V, et al. (2006) Transcriptomic and proteomic analyses of the pMOL30-encoded copper

- resistance in *Cupriavidus metallidurans* strain CH34. *Microbiology* 152: 1765–1776.
76. von Rozycki T, Nies DH (2009) *Cupriavidus metallidurans*: evolution of a metal-resistant bacterium. *Antonie Van Leeuwenhoek* 96: 115–139.
  77. Dong Q, Sadouk A, van der Lelie D, Taghavi S, Ferhat A, et al. (1992) Cloning and sequencing of IS1086, an *Alcaligenes eutrophus* insertion element related to IS30 and IS4351. *J Bacteriol* 174: 8133–8138.
  78. Mergeay M, Sadouk A, Diels L, Faelen M, Gerits J, et al. (1987) High level spontaneous mutagenesis revealed by survival at non-optimal temperature in *Alcaligenes eutrophus* CH34. *Arch Internat Physiol Bioch* 95: 35–36.
  79. Taghavi S, Mergeay M, van der Lelie D (1997) Genetic and physical maps of the *Alcaligenes eutrophus* CH34 megaplasmid pMOL28 and its derivative pMOL50 obtained after temperature-induced mutagenesis and mortality. *Plasmid* 37: 22–34.
  80. Liesegang H, Lemke K, Siddiqui RA, Schlegel HG (1993) Characterization of the inducible nickel and cobalt resistance determinant *cnr* from pMOL28 of *Alcaligenes eutrophus* CH34. *J Bacteriol* 175: 767–778.
  81. Tibazarwa C, Wuertz S, Mergeay M, Wymys L, van Der Lelie D (2000) Regulation of the *cnr* cobalt and nickel resistance determinant of *Ralstonia eutropha* (*Alcaligenes eutrophus*) CH34. *J Bacteriol* 182: 1399–1409.
  82. Collard JM, Corbisier P, Diels L, Dong Q, Jeanthon C, et al. (1994) Plasmids for heavy metal resistance in *Alcaligenes eutrophus* CH34: mechanisms and applications. *FEMS Microbiol Rev* 14: 405–414.
  83. Monchy S, Vallaeys T, Bossus A, Mergeay M (2006) Metal transport ATPase genes from *Cupriavidus metallidurans* CH34: a transcriptomic approach. *International Journal of Environmental Analytical Chemistry* 86: 677–692.
  84. Grosse C, Grass G, Anton A, Franke S, Santos AN, et al. (1999) Transcriptional organization of the *cze* heavy-metal homeostasis determinant from *Alcaligenes eutrophus*. *J Bacteriol* 181: 2385–2393.
  85. Borremans B, Hobman JL, Provoost A, Brown NL, van Der Lelie D (2001) Cloning and functional analysis of the *pbr* lead resistance determinant of *Ralstonia metallidurans* CH34. *J Bacteriol* 183: 5651–5658.
  86. Noel-Georis I, Vallaeys T, Chauvaux R, Monchy S, Falmagne P, et al. (2004) Global analysis of the *Ralstonia metallidurans* proteome: prelude for the large-scale study of heavy metal response. *Proteomics* 4: 151–179.
  87. Feulner G, Gray JA, Kirschman JA, Lehner AF, Sadosky AB, et al. (1990) Structure of the *rhsA* locus from *Escherichia coli* K-12 and comparison of *rhsA* with other members of the *rhs* multigene family. *J Bacteriol* 172: 446–456.
  88. Hill CW, Sandt CH, Vlazny DA (1994) *Rhs* elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. *Mol Microbiol* 12: 865–871.
  89. Montanini B, Blaudez D, Jeandroz S, Sanders D, Chalot M (2007) Phylogenetic and functional analysis of the Cation Diffusion Facilitator (CDF) family: improved signature and prediction of substrate specificity. *BMC Genomics* 8: 107.
  90. Nies DH, Silver S (1995) Ion efflux systems involved in bacterial metal resistances. *J Ind Microbiol* 14: 186–199.
  91. Axelsen KB, Palmgren MG (1998) Evolution of substrate specificities in the P-type ATPase superfamily. *J Mol Evol* 46: 84–101.
  92. Dong Q, Mergeay M (1994) *Czc/cnr* efflux: a three-component chemiosmotic antiport pathway with a 12-transmembrane-helix protein. *Mol Microbiol* 14: 185–187.
  93. Saier MH, Jr., Tam R, Reizer A, Reizer J (1994) Two novel families of bacterial membrane proteins concerned with nodulation, cell division and transport. *Mol Microbiol* 11: 841–847.
  94. Zhang YB, Monchy S, Greenberg B, Mergeay M, Gang O, et al. (2009) ArsR arsenic-resistance regulatory protein from *Cupriavidus metallidurans* CH34. *Antonie Van Leeuwenhoek* 96: 161–170.
  95. Brown NL, Barrett SR, Camakaris J, Lec BT, Rouch DA (1995) Molecular genetics and transport analysis of the copper-resistance determinant (*pco*) from *Escherichia coli* plasmid pRJ1004. *Mol Microbiol* 17: 1153–1166.
  96. Cha JS, Cooksey DA (1991) Copper resistance in *Pseudomonas syringae* mediated by periplasmic and outer membrane proteins. *Proc Natl Acad Sci U S A* 88: 8915–8919.
  97. Liebert CA, Hall RM, Summers AO (1999) Transposon Tn21, flagship of the floating genome. *Microbiol Mol Biol Rev* 63: 507–522.
  98. Summers AO, Silver S (1972) Mercury resistance in a plasmid-bearing strain of *Escherichia coli*. *J Bacteriol* 112: 1228–1236.
  99. Bersch B, Favier A, Schanda P, van Aelst S, Vallaeys T, et al. (2008) Molecular structure and metal-binding properties of the periplasmic CopK protein expressed in *Cupriavidus metallidurans* CH34 during copper challenge. *J Mol Biol* 380: 386–403.
  100. Pompidor G, Girard E, Maillard A, Ramella-Pairin S, Bersch B, et al. (2009) Biostructural analysis of the metal-sensor domain of CnrX from *Cupriavidus metallidurans* CH34. *Antonie Van Leeuwenhoek* 96: 141–148.
  101. Sendra V, Cannella D, Bersch B, Fieschi F, Menage S, et al. (2006) CopH from *Cupriavidus metallidurans* CH34. A novel periplasmic copper-binding protein. *Biochemistry* 45: 5557–5566.
  102. Zoropogui A, Gambarelli S, Coves J (2008) CzcE from *Cupriavidus metallidurans* CH34 is a copper-binding protein. *Biochem Biophys Res Commun* 365: 735–739.
  103. Lu M, Chai J, Fu D (2009) Structural basis for autoregulation of the zinc transporter YiiP. *Nat Struct Mol Biol* 16: 1063–1067.
  104. Lu M, Fu D (2007) Structure of the zinc transporter YiiP. *Science* 317: 1746–1748.
  105. Nies D, Mergeay M, Friedrich B, Schlegel HG (1987) Cloning of plasmid genes encoding resistance to cadmium, zinc, and cobalt in *Alcaligenes eutrophus* CH34. *J Bacteriol* 169: 4865–4868.
  106. Taghavi S, Provoost A, Mergeay M, van der Lelie D (1996) Identification of a partition and replication region in the *Alcaligenes eutrophus* megaplasmid pMOL28. *Mol Gen Genet* 250: 169–179.
  107. Schmidt T, Schlegel HG (1994) Combined nickel-cobalt-cadmium resistance encoded by the *ncc* locus of *Alcaligenes xyloxydans* 31A. *J Bacteriol* 176: 7045–7054.
  108. Hassan MT, van der Lelie D, Springael D, Romling U, Ahmed N, et al. (1999) Identification of a gene cluster, *czt*, involved in cadmium and zinc resistance in *Pseudomonas aeruginosa*. *Gene* 238: 417–425.
  109. Franke S, Grass G, Rensing C, Nies DH (2003) Molecular analysis of the copper-transporting efflux system CusCFBA of *Escherichia coli*. *J Bacteriol* 185: 3804–3812.
  110. Leedjarv A, Ivask A, Virta M (2008) Interplay of different transporters in the mediation of divalent heavy metal resistance in *Pseudomonas putida* KT2440. *J Bacteriol* 190: 2680–2689.
  111. Legatzki A, Grass G, Anton A, Rensing C, Nies DH (2003) Interplay of the Czc system and two P-type ATPases in conferring metal resistance to *Ralstonia metallidurans*. *J Bacteriol* 185: 4354–4361.
  112. Munkelt D, Grass G, Nies DH (2004) The chromosomally encoded cation diffusion facilitator proteins DmeF and FieF from *Wautersia metallidurans* CH34 are transporters of broad metal specificity. *J Bacteriol* 186: 8036–8043.
  113. Mikolay A, Nies DH (2009) The ABC-transporter AtmA is involved in nickel and cobalt resistance of *Cupriavidus metallidurans* strain CH34. *Antonie Van Leeuwenhoek* 96: 183–191.
  114. Grosse C, Friedrich S, Nies DH (2007) Contribution of extracytoplasmic function sigma factors to transition metal homeostasis in *Cupriavidus metallidurans* strain CH34. *J Mol Microbiol Biotechnol* 12: 227–240.
  115. Ren Q, Chen K, Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* 35: D274–279.
  116. Ren Q, Kang KH, Paulsen IT (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res* 32: D284–288.
  117. Saier MH, Jr., Tran CV, Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* 34: D181–186.
  118. Saier MH, Jr., Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res* 37: D274–278.
  119. von Rozycki T, Nies DH, Saier MH (2005) Genomic Analyses of Transport Proteins in *Ralstonia metallidurans*. *Comp Funct Genomics* 6: 17–56.
  120. Ren Q, Paulsen IT (2005) Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLoS Comput Biol* 1: e27.
  121. Kitada M, Kosono S, Kudo T (2000) The Na<sup>+</sup>/H<sup>+</sup> antiporter of alkaliphilic *Bacillus* sp. *Extremophiles* 4: 253–258.
  122. Palmer T, Berks BC (2003) Moving folded proteins across the bacterial cell membrane. *Microbiology* 149: 547–556.
  123. Nies DH (2004) Incidence and function of sigma factors in *Ralstonia metallidurans* and other bacteria. *Arch Microbiol* 181: 255–268.
  124. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD - taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res* 36: D88–92.
  125. Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res* 34: D74–81.
  126. van Nimwegen E (2003) Scaling laws in the functional content of genomes. *Trends Genet* 19: 479–484.
  127. Cases I, de Lorenzo V, Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol* 11: 248–253.
  128. Ulrich LE, Koonin EV, Zhulin IB (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol* 13: 52–56.
  129. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36: D281–288.
  130. Schell MA (1993) Molecular biology of the LysR family of transcriptional regulators. *Annu Rev Microbiol* 47: 597–626.
  131. Ma Z, Jacobsen FE, Giedroc DP (2009) Coordination chemistry of bacterial metal transport and sensing. *Chem Rev* 109: 4644–4681.
  132. Ulrich LE, Zhulin IB (2009) The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res*.
  133. Gillis A, Khan MA, Cornelis P, Meyer JM, Mergeay M, et al. (1996) Siderophore-mediated iron uptake in *Alcaligenes eutrophus* CH34 and identification of *aleB* encoding the ferric iron-alcaligin E receptor. *J Bacteriol* 178: 5499–5507.
  134. Munzinger M, Taraz K, Budzikiewicz H (1999) Staphyloferrin B, a citrate siderophore of *Ralstonia eutropha*. *Zeitschrift Fur Naturforschung C-a Journal of Biosciences* 54: 867–875.
  135. Bhatt G, Denny TP (2004) *Ralstonia solanacearum* iron scavenging by the siderophore staphyloferrin B is controlled by PhcA, the global virulence regulator. *J Bacteriol* 186: 7896–7904.



136. Braud A, Jezequel K, Bazot S, Lebeau T (2009) Enhanced phytoextraction of an agricultural Cr- and Pb-contaminated soil by bioaugmentation with siderophore-producing bacteria. *Chemosphere* 74: 280–286.
137. Brickman TJ, Armstrong SK (1996) The ornithine decarboxylase gene *odc* is required for alcaligin siderophore biosynthesis in *Bordetella* spp.: Putrescine is a precursor of alcaligin. *J Bacteriol* 178: 54–60.
138. Gilis A, Corbisier P, Baeyens W, Taghavi S, Mergeay M, et al. (1998) Effect of the siderophore alcaligin E on the bioavailability of Cd to *Alcaligenes eutrophus* CH34. *J Ind Microbiol Biotechnol* 20: 61–68.
139. Garg RP, Yindeeoungyeon W, Gilis A, Denny TP, Van Der Lelie D, et al. (2000) Evidence that *Ralstonia eutropha* (*Alcaligenes eutrophus*) contains a functional homologue of the *Ralstonia solanacearum* Phc cell density sensing system. *Mol Microbiol* 38: 359–367.
140. Galperin MY, Gomelsky M (2005) Bacterial signal transduction modules: from genomics to biology. *ASM News* 71: 326–333.
141. Ulrich LE, Zhulin IB (2007) MiST: a microbial signal transduction database. *Nucleic Acids Res* 35: D386–390.
142. Antoine R, Raze D, Loch C (2000) Genomics of *Bordetella pertussis* toxins. *Int J Med Microbiol* 290: 301–305.
143. Herrou J, Bompard C, Antoine R, Leroy A, Rucktooa P, et al. (2007) Structure-based mechanism of ligand binding for periplasmic solute-binding protein of the Bug family. *J Mol Biol* 373: 954–964.
144. Huvent I, Belrhali H, Antoine R, Bompard C, Loch C, et al. (2006) Crystal structure of *Bordetella pertussis* BugD solute receptor unveils the basis of ligand binding in a new family of periplasmic binding proteins. *J Mol Biol* 356: 1014–1026.
145. Antoine R, Huvent I, Chemlal K, Deray I, Raze D, et al. (2005) The periplasmic binding protein of a tripartite tricarboxylate transporter is involved in signal transduction. *J Mol Biol* 351: 799–809.
146. von Wintzingerode F, Schattke A, Siddiqui RA, Rosick U, Gobel UB, et al. (2001) *Bordetella petrii* sp. nov., isolated from an anaerobic bioreactor, and emended description of the genus *Bordetella*. *Int J Syst Evol Microbiol* 51: 1257–1265.
147. Wang F, Grundmann S, Schmid M, Dorfler U, Rohrer S, et al. (2007) Isolation and characterization of 1,2,4-trichlorobenzene mineralizing *Bordetella* sp. and its bioremediation potential in soil. *Chemosphere* 67: 896–902.
148. Gross R, Guzman CA, Sebahia M, dos Santos VA, Pieper DH, et al. (2008) The missing link: *Bordetella petrii* is endowed with both the metabolic versatility of environmental bacteria and virulence traits of pathogenic *Bordetellae*. *BMC Genomics* 9: 449.
149. Fry NK, Duncan J, Malnick H, Warner M, Smith AJ, et al. (2005) *Bordetella petrii* clinical isolate. *Emerg Infect Dis* 11: 1131–1133.
150. Stark D, Riley LA, Harkness J, Marriott D (2007) *Bordetella petrii* from a clinical sample in Australia: isolation and molecular identification. *J Med Microbiol* 56: 435–437.
151. Cummings CA, Bootsma HJ, Relman DA, Miller JF (2006) Species- and strain-specific control of a complex, flexible regulon by *Bordetella BvgAS*. *J Bacteriol* 188: 1775–1785.
152. Masuda N, Church GM (2002) *Escherichia coli* gene expression responsive to levels of the response regulator EvgA. *J Bacteriol* 184: 6225–6234.
153. Kulasekara HD, Ventre I, Kulasekara BR, Lazdunski A, Filloux A, et al. (2005) A novel two-component system controls the expression of *Pseudomonas aeruginosa* fimbrial cup genes. *Mol Microbiol* 55: 368–380.
154. Clarke BR, Pearce R, Roberts IS (1999) Genetic organization of the *Escherichia coli* K10 capsule gene cluster: identification and characterization of two conserved regions in group III capsule gene clusters encoding polysaccharide transport functions. *J Bacteriol* 181: 2279–2285.
155. Russo TA, Wenderoth S, Carlino UB, Merrick JM, Lesse AJ (1998) Identification, genomic organization, and analysis of the group III capsular polysaccharide genes *kpsD*, *kpsM*, *kpsT*, and *kpsE* from an extraintestinal isolate of *Escherichia coli* (CP9, O4/K54/H5). *J Bacteriol* 180: 338–349.
156. Arrecubieta C, Hammarton TC, Barrett B, Chareonsudjai S, Hodson N, et al. (2001) The transport of group 2 capsular polysaccharides across the periplasmic space in *Escherichia coli*. Roles for the KpsE and KpsD proteins. *J Biol Chem* 276: 4245–4250.
157. Vimr ER, Stenbergen SM (2009) Early molecular-recognition events in the synthesis and export of group 2 capsular polysaccharides. *Microbiology* 155: 9–15.
158. Capela D, Barloy-Hubler F, Gouzy J, Bothe G, Ampe F, et al. (2001) Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proc Natl Acad Sci U S A* 98: 9877–9882.
159. Kiss E, Kereszt A, Barta F, Stephens S, Reuhs BL, et al. (2001) The *Ahp-3* gene region of *Sinorhizobium meliloti* Rm41 contains strain-specific genes that determine K antigen structure. *Mol Plant Microbe Interact* 14: 1395–1403.
160. Friedman L, Kolter R (2004a) Genes involved in matrix formation in *Pseudomonas aeruginosa* PA14 biofilms. *Mol Microbiol* 51: 675–690.
161. Vasseur P, Vallet-Gely I, Soscia C, Genin S, Filloux A (2005) The *pel* genes of the *Pseudomonas aeruginosa* PAK strain are involved at early and late stages of biofilm formation. *Microbiology* 151: 985–997.
162. Friedman L, Kolter R (2004b) Two genetic loci produce distinct carbohydrate-rich structural components of the *Pseudomonas aeruginosa* biofilm matrix. *J Bacteriol* 186: 4457–4465.
163. Guiné V, Spadini L, Sarret G, Muris M, Delolme C, et al. (2006) Zinc sorption to three gram-negative bacteria: combined titration, modeling, and EXAFS study. *Environ Sci Technol* 40: 1806–1813.
164. Clough SJ, Flavier AB, Schell MA, Denny TP (1997) Differential Expression of Virulence Genes and Motility in *Ralstonia (Pseudomonas) solanacearum* during Exponential Growth. *Appl Environ Microbiol* 63: 844–850.
165. Lutke-Eversloh T, Steinbuechel A (1999) Biochemical and molecular characterization of a succinate semialdehyde dehydrogenase involved in the catabolism of 4-hydroxybutyric acid in *Ralstonia eutropha*. *FEMS Microbiol Lett* 181: 63–71.
166. Harwood CS, Burchhardt G, Herrmann H, Fuchs G (1998) Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway. *FEMS Microbiology Reviews* 22: 439–458.
167. Harrison FH, Harwood CS (2005) The *pimFABCDE* operon from *Rhodospseudomonas palustris* mediates dicarboxylic acid degradation and participates in anaerobic benzoate degradation. *Microbiology* 151: 727–736.
168. Siedow A, Cramm R, Siddiqui RA, Friedrich B (1999) A megaplasmid-borne anaerobic ribonucleotide reductase in *Alcaligenes eutrophus* H16. *J Bacteriol* 181: 4919–4928.
169. Cramm R (2009) Genomic view of energy metabolism in *Ralstonia eutropha* H16. *J Mol Microbiol Biotechnol* 16: 38–52.
170. Roth JR, Lawrence JG, Bobik TA (1996) Cobalamin (coenzyme B12): synthesis and biological significance. *Annu Rev Microbiol* 50: 137–181.
171. Warren MJ, Raux E, Schubert HL, Escalante-Semerena JC (2002) The biosynthesis of adenosylcobalamin (vitamin B12). *Nat Prod Rep* 19: 390–412.
172. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS (2003) Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J Biol Chem* 278: 41148–41159.
173. Atomi H (2002) Microbial enzymes involved in carbon dioxide fixation. *J Biosci Bioeng* 94: 497–505.
174. Tabita FR, Gibson JL, Bowien B, Dijkhuizen L, Meijer WG (1992) Uniform designation for genes of the Calvin-Benson-Bassham reductive pentose phosphate pathway of bacteria. *FEMS Microbiol Lett* 78: 107–110.
175. Elsen S, Swem LR, Swem DL, Bauer CE (2004) RegB/RegA, a highly conserved redox-responding global two-component regulatory system. *Microbiol Mol Biol Rev* 68: 263–279.
176. Gibson JL, Tabita FR (1997) Analysis of the *cbbXYZ* operon in *Rhodobacter sphaeroides*. *J Bacteriol* 179: 663–669.
177. King GM, Weber CF (2007) Distribution, diversity and ecology of aerobic CO<sub>2</sub>-oxidizing bacteria. *Nat Rev Microbiol* 5: 107–118.
178. Lorite MJ, Tachil J, Sanjuan J, Meyer O, Bedmar EJ (2000) Carbon monoxide dehydrogenase activity in *Bradyrhizobium japonicum*. *Appl Environ Microbiol* 66: 1871–1876.
179. Alber B, Olinger M, Rieder A, Kockelkorn D, Jobst B, et al. (2006) Malonyl-coenzyme A reductase in the modified 3-hydroxypropionate cycle for autotrophic carbon fixation in archaeal *Metallosphaera* and *Sulfolobus* spp. *J Bacteriol* 188: 8551–8559.
180. Hügl M, Menendez C, Schagger H, Fuchs G (2002) Malonyl-coenzyme A reductase from *Chloroflexus aurantiacus*, a key enzyme of the 3-hydroxypropionate cycle for autotrophic CO<sub>2</sub> fixation. *J Bacteriol* 184: 2404–2410.
181. Friedrich CG, Bardischewsky F, Rother D, Quentmeier A, Fischer J (2005) Prokaryotic sulfur oxidation. *Curr Opin Microbiol* 8: 253–259.
182. Sauvé V, Bruno S, Berks BC, Hemmings AM (2007) The SoxYZ complex carries sulfur cycle intermediates on a peptide swinging arm. *J Biol Chem* 282: 23194–23204.
183. Denger K, Laue H, Cook AM (1997) Thiosulfate as a metabolic product: the bacterial fermentation of taurine. *Arch Microbiol* 168: 297–301.
184. Schmalenberger A, Kertesz MA (2007) Desulfurization of aromatic sulfonates by rhizosphere bacteria: high diversity of the *asfA* gene. *Environ Microbiol* 9: 535–545.
185. Andrews SC, Berks BC, McClay J, Ambler A, Quail MA, et al. (1997) A 12-cistron *Escherichia coli* operon (*hyf*) encoding a putative proton-translocating formate hydrogenlyase system. *Microbiology* 143(Pt 11): 3633–3647.
186. Vignais PM, Colbeau A (2004) Molecular biology of microbial hydrogenases. *Curr Issues Mol Biol* 6: 159–188.
187. Böck A, King PW, Blokesch M, Posewitz MC (2006) Maturation of hydrogenases. *Adv Microb Physiol* 51: 1–71.
188. Vincent KA, Cracknell JA, Clark JR, Ludwig M, Lenz O, et al. (2006) Electricity from low-level H<sub>2</sub> in still air - an ultimate test for an oxygen tolerant hydrogenase. *Chem Commun (Camb)*. pp 5033–5035.
189. Lovley DR (2006) Bug juice: harvesting electricity with microorganisms. *Nat Rev Microbiol* 4: 497–508.
190. Wrighton KC, Coates JD (2009) Microbial Fuel Cells: Plug-in and Power-on Microbiology. *Microbe* 4: 281–287.
191. Boyd JM, Ellsworth H, Ensign SA (2004) Bacterial acetone carboxylase is a manganese-dependent metalloenzyme. *J Biol Chem* 279: 46644–46651.
192. Sluis MK, Larsen RA, Krum JG, Anderson R, Metcalf WW, et al. (2002) Biochemical, molecular, and genetic analyses of the acetone carboxylases from *Xanthobacter autotrophicus* strain Py2 and *Rhodobacter capsulatus* strain B10. *J Bacteriol* 184: 2969–2977.
193. Studholme DJ, Dixon R (2003) Domain architectures of sigma54-dependent transcriptional activators. *J Bacteriol* 185: 1757–1767.

194. Jendrossek D, Kruger N, Steinbuechel A (1990) Characterization of alcohol dehydrogenase genes of derepressible wild-type *Alcaligenes eutrophus* H16 and constitutive mutants. *J Bacteriol* 172: 4844–4851.
195. Notomista E, Lahm A, Di Donato A, Tramontano A (2003) Evolution of bacterial and archaeal multicomponent monooxygenases. *J Mol Evol* 56: 435–445.
196. Cafaro V, Izzo V, Scognamiglio R, Notomista E, Capasso P, et al. (2004) Phenol hydroxylase and toluene/o-xylene monooxygenase from *Pseudomonas stutzeri* OX1: interplay between two enzymes. *Appl Environ Microbiol* 70: 2211–2219.
197. Kageyama C, Ohta T, Hiraoka K, Suzuki M, Okamoto T, et al. (2005) Chlorinated aliphatic hydrocarbon-induced degradation of trichloroethylene in *Wautersia numadzuensis* sp. nov. *Arch Microbiol* 183: 56–65.
198. Pérez-Pantoja D, De la Iglesia R, Pieper DH, Gonzalez B (2008) Metabolic reconstruction of aromatic compounds degradation from the genome of the amazing pollutant-degrading bacterium *Cupriavidus necator* JMP134. *FEMS Microbiol Rev* 32: 736–794.
199. Lenz RW, Marchessault RH (2005) Bacterial polyesters: biosynthesis, biodegradable plastics and biotechnology. *Biomacromolecules* 6: 1–8.
200. Khanna S, Srivastava AK (2005) Recent Advances in microbial polyhydroxyalkanoates. *Process Biochemistry* 40: 607–619.
201. York GM, Lupberger J, Tian J, Lawrence AG, Stubbe J, et al. (2003) *Ralstonia eutropha* H16 encodes two and possibly three intracellular poly[D(-)-3-hydroxybutyrate] depolymerase genes. *J Bacteriol* 185: 3788–3794.
202. Kobayashi T, Shiraki M, Abe T, Sugiyama A, Saito T (2003) Purification and properties of an intracellular 3-hydroxybutyrate-oligomer hydrolase (PhaZ2) in *Ralstonia eutropha* H16 and its identification as a novel intracellular poly(3-hydroxybutyrate) depolymerase. *J Bacteriol* 185: 3485–3490.
203. Kobayashi T, Uchino K, Abe T, Yamazaki Y, Saito T (2005) Novel intracellular 3-hydroxybutyrate-oligomer hydrolase in *Wautersia eutropha* H16. *J Bacteriol* 187: 5129–5135.
204. Saegusa H, Shiraki M, Saito T (2002) Cloning of an intracellular D(-)-3-hydroxybutyrate-oligomer hydrolase gene from *Ralstonia eutropha* H16 and identification of the active site serine residue by site-directed mutagenesis. *J Biosci Bioeng* 94: 106–112.
205. Sugiyama A, Kobayashi T, Shiraki M, Saito T (2004) Roles of poly(3-hydroxybutyrate) depolymerase and 3HB-oligomer hydrolase in bacterial PHB metabolism. *Curr Microbiol* 48: 424–427.
206. Taghavi S, van der Lelie D, Mergeay M (1994) Electroporation of *Alcaligenes eutrophus* with (mega) plasmids and genomic DNA fragments. *Appl Environ Microbiol* 60: 3585–3591.
207. Rocha EP, Cornet E, Michel B (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* 1: e15.
208. Low DA, Weyand NJ, Mahan MJ (2001) Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect Immun* 69: 7197–7204.
209. Palma M, Zurita J, Ferreras JA, Worgall S, Larone DH, et al. (2005) *Pseudomonas aeruginosa* SoxR does not conform to the archetypal paradigm for SoxR-dependent regulation of the bacterial oxidative stress adaptive response. *Infect Immun* 73: 2958–2966.
210. Defrance M, Janky R, Sand O, van Helden J (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat Protoc* 3: 1589–1603.
211. Siegle DA (2005) Universal stress proteins in *Escherichia coli*. *J Bacteriol* 187: 6253–6254.
212. Sousa MC, McKay DB (2001) Structure of the universal stress protein of *Haemophilus influenzae*. *Structure* 9: 1135–1141.
213. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8: 195–202.
214. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19: 1639–1645.
215. Ghai R, Hain T, Chakraborty T (2004) GenomeViz: visualizing microbial genomes. *BMC Bioinformatics* 5: 198.
216. Gao F, Zhang CT (2008) Ori-Finder: a web-based system for finding *oriCs* in unannotated bacterial genomes. *BMC Bioinformatics* 9: 79.
217. Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. *Nat Genet* 21: 108–110.
218. Sakakibara Y, Osana Y, Popendorf K (2007) Development of a large-scale comparative genome system and its application to the analysis of mycobacteria genomes. *Nihon Hansenbyo Gakkai Zasshi* 76: 251–256.
219. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5: R12.
220. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.
221. Promponas VJ, Enright AJ, Tsoka S, Kreil DP, Leroy C, et al. (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Complexity analysis of sequence tracts. Bioinformatics* 16: 915–922.
222. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147: 195–197.
223. Salanoubat M, Genin S, Artiguenave F, Gouzy J, Mangenot S, et al. (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum*. *Nature* 415: 497–502.
224. Chain PS, Denef VJ, Konstantinidis KT, Vergez LM, Agullo L, et al. (2006) *Burkholderia xenovorans* LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. *Proc Natl Acad Sci U S A* 103: 15280–15287.
225. Audic S, Robert C, Campagna B, Parinello H, Claverie JM, et al. (2007) Genome analysis of *Mimibacterium massiliensis* highlights the convergent evolution of water-living bacteria. *PLoS Genet* 3: e138.
226. Muller D, Médigue C, Koechler S, Barbe V, Barakat M, et al. (2007) A tale of two oxidation states: bacterial colonization of arsenic-rich environments. *PLoS Genet* 3: e53.