Complete genome of *Phenylobacterium zucineum* – a novel facultative intracellular bacterium isolated from human erythroleukemia cell line K562

Yingfeng Luo†1,2,3, Xiaoli Xu†1, Zonghui Ding†1, Zhen Liu1, Bing Zhang2,3, Zhiyu Yan1, Jie Sun1, Songnian Hu*2,3 and Xun Hu*1

Address: 1Cancer Institute (Key Laboratory for Cancer Intervention and Prevention, National Ministry of Education, PR China; Key Laboratory of Molecular Biology in Medical Sciences, Zhejiang Province, PR China), the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, PR China, 2James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou, PR China and 3Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, PR China

Email: Yingfeng Luo - luoyf@big.ac.cn; Xiaoli Xu - jessie_xuxiaoli@hotmail.com; Zonghui Ding - dingzonghui@yahoo.com.cn; Zhen Liu - liuzhen@cancer.ac.cn; Bing Zhang - zhangbing@big.ac.cn; Zhiyu Yan - yanzhiyu1982@yahoo.com.cn; Jie Sun - sunjie1030@yahoo.com.cn; Songnian Hu* - huasn@big.ac.cn; Xun Hu* - huxun@zju.edu.cn

* Corresponding authors †Equal contributors

Published: 13 August 2008

BMC Genomics 2008, 9:386 doi:10.1186/1471-2164-9-386

Received: 21 August 2007

Accepted: 13 August 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/386

© 2008 Luo et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

**Background:** *Phenylobacterium zucineum* is a recently identified facultative intracellular species isolated from the human leukemia cell line K562. Unlike the known intracellular pathogens, *P. zucineum* maintains a stable association with its host cell without affecting the growth and morphology of the latter.

**Results:** Here, we report the whole genome sequence of the type strain HLK1T. The genome consists of a circular chromosome (3,996,255 bp) and a circular plasmid (382,976 bp). It encodes 3,861 putative proteins, 42 tRNAs, and a 16S-23S-5S rRNA operon. Comparative genomic analysis revealed that it is phylogenetically closest to *Caulobacter crescentus*, a model species for cell cycle research. Notably, *P. zucineum* has a gene that is strikingly similar, both structurally and functionally, to the cell cycle master regulator CtrA of *C. crescentus*, and most of the genes directly regulated by CtrA in the latter have orthologs in the former.

**Conclusion:** This work presents the first complete bacterial genome in the genus *Phenylobacterium*. Comparative genomic analysis indicated that the CtrA regulon is well conserved between *C. crescentus* and *P. zucineum*.

**Background**

*Phenylobacterium zucineum* strain HLK1T is a facultative intracellular microbe recently identified by us [1]. It is a rod-shaped Gram-negative bacterium 0.3–0.5 × 0.5–2 μm in size. It belongs to the genus *Phenylobacterium* [2], which presently comprises 5 species, *P. lituiforme* (Fail3T) [3], *P. falsum* (AC49T) [4], *P. immobile* (ET) [2], *P. koreense* (Slu-01T) [5], and *P. zucineum* (HLK1T) [1]. They were isolated from subsurface aquifer, alkaline groundwater, soil, activated sludge from a wastewater treatment plant, and the human leukemia cell line K562, respectively. Except for *P. zucineum*, they are environmental bacteria, and there is no
evidence that these microbes are associated with eukaryotic cells. The HLK1T strain, therefore, represents the only species so far in the genus *Phenylobacterium* that can infect and survive in human cells. Since most, if not all, of the known microbes that can invade human cells are pathogenic, we proposed that HLK1T may have pathogenic relevance to humans [1]. Unlike the known intracellular pathogens that undergo a cycle involving invasion, overgrowth, and disruption of the host cells, and repeating the cycle by invading new cells, HLK1T is able to establish a stable parasitic association with its host, i.e., the strain does not overgrow intracellularly to kill the host, and the host cells carry them to their progeny. One cell line (SW480) infected with *P. zucineum* has been stably maintained for nearly three years in our lab (data not shown).

In this report, we present the complete genome sequence of *P. zucineum*.

**Results**

**Genome anatomy**

The genome is composed of a circular chromosome (3,996,255 bp) and a circular plasmid (382,976 bp) (Figure 1; Table 1). The G + C contents of chromosome and plasmid are 71.35% and 68.5%, respectively. There are 3,861 putative protein-coding genes (3,534 in the chromosome and 327 in the plasmid), of which 3,180 have significant matches in the non-redundant protein database. Of the matches, 585 are conserved hypothetical proteins and 2,595 are proteins with known or predicted functions. Forty-two tRNA genes and one 16S-23S-5S rRNA operon were identified in the chromosome.

There are 7 families of protein-coding repetitive sequences and a family of noncoding repeats in the genome (Table 2). Notably, identical copies of repeats 02–04 were found in both the chromosome and the plasmid, suggesting their potential involvement in homologous recombination.

On the basis of COG (Cluster of Orthologous Groups) classification, the chromosome is enriched in genes for basic metabolism, such as categories E (amino acid transport and metabolism) and I (lipid transport and metabolism), accounting for 8.29% and 6.09% of the total genes in the chromosome, respectively. On the other hand, the plasmid is enriched for genes in categories O (posttranslational modification, protein turnover, chaperones) and T

| Genomic Element | Chromosome | Plasmid |
|-----------------|------------|---------|
| Length (bp)     | 3,996,255  | 382,976 |
| GC content (%)  | 71.35      | 68.54   |
| Proteins        | 3,534      | 327     |
| Coding region of genome (%) | 88.85 | 81.94 |
| Proteins with known or predicted function | 2,394(67.73%) | 201(61.47%) |
| Conserved hypothetical proteins | 560(15.84%) | 25(7.65%) |
| Hypothetical proteins | 580(16.41%) | 101(30.88%) |
| rRNA operon     | 1          | 0       |
| tRNAs           | 42         | 0       |
| Proteins in each COG category | | |
| [J] Translation, ribosomal structure and biogenesis | 185 (5.24%) | 3 (1.21%) |
| [K] Transcription | 210 (5.94%) | 22 (8.91%) |
| [L] Replication, recombination and repair | 139 (3.93%) | 23 (9.31%) |
| [D] Cell cycle control, cell division, chromosome partitioning | 27 (0.76%) | 0 |
| [V] Defense mechanisms | 51 (1.44%) | 3 (1.21%) |
| [T] Signal transduction mechanisms | 166 (4.47%) | 24 (9.72%) |
| [M] Cell wall/membrane/envelope biogenesis | 195 (5.52%) | 15 (6.07%) |
| [N] Cell motility | 62 (1.75%) | 4 (1.62%) |
| [U] Intracellular trafficking, secretion, and vesicular transport | 96 (2.72%) | 13 (5.26%) |
| [O] Posttranslational modification, protein turnover, chaperones | 151 (4.27%) | 32 (12.96%) |
| [C] Energy production and conversion | 188 (5.32%) | 16 (6.48%) |
| [G] Carbohydrate transport and metabolism | 161 (4.56%) | 15 (6.07%) |
| [E] Amino acid transport and metabolism | 293 (8.29%) | 5 (2.02%) |
| [F] Nucleotide transport and metabolism | 58 (1.64%) | 3 (1.21%) |
| [H] Coenzyme transport and metabolism | 116 (3.28%) | 3 (1.21%) |
| [I] Lipid transport and metabolism | 215 (6.09%) | 12 (4.86%) |
| [P] Inorganic ion transport and metabolism | 232 (6.31%) | 24 (9.72%) |
| [Q] Secondary metabolites biosynthesis, transport and catabolism | 152 (4.3%) | 9 (3.64%) |
| [R] General function prediction only | 444 (12.57%) | 28 (11.34%) |
| [S] Function unknown | 307 (8.69%) | 20 (8.10%) |
Figure 1

Circular representation of the *P. zucineum* strain HLK1^T^ chromosome and plasmid (smaller circle). Circles indicate (from the outside): (1) Physical map scaled in megabases from base 1, the start of the putative replication origin. (2) Coding sequences transcribed in the clockwise direction are color-coded according to COG functional category. (3) Coding sequences transcribed in the counterclockwise direction are color-coded according to COG functional category. (4) Proteins involved in establishment of intracellular niche are TonB-dependent receptors (orange) and pilus genes (sienna). (5) Functional elements responsible for environmental transition are extracytoplasmic function sigma factors (royal blue), transcriptional regulators (violet red), two-component signal transduction proteins (deep sky blue), heat shock molecular chaperons (spring green), type IV secretion systems (plum), chemotaxis systems (green yellow) and flagellum proteins (gray). (6) G + C percent content (10-kb window and 1-kb incremental shift for chromosome; 300 bp window and 150 bp for incremental shift for plasmid); values larger than average (71.35% in chromosome and 68.5% in plasmid) are in red and smaller in medium blue. (7) GC skew (10-kb window and 1-kb incremental shift for chromosome; 300 bp window and 150 bp for incremental shift for plasmid); values greater than zero are in gold and smaller in purple. (8) Repeat families, repeats 01-08 are in dark salmon, dark red, wheat, tomato, light green, salmon, dark blue and gold, respectively.
As to genes in the plasmid that cope with environmental stimuli, about half of the genes in category O are molecular chaperones (17/32), including 2 dnaJ-like molecular chaperones, 2 clusters of dnaK and its co-chaperonin grpE (PHZ_p0053-0054 and PHZ_p0121-122), a cluster of groEL and its co-chaperonin groES (PHZ_p0095-0096), and 9 heat shock proteins Hsp20. Of 23 genes in category T, there is one cluster (FixLJ, PHZ_p0187-0188), which is essential for the growth of *C. crescentus* under hypoxic conditions [6].

**General metabolism**

The enzyme sets of glycolysis and the Entner-Doudoroff pathway are complete in the genome. All genes comprising the pentose phosphate pathway except gluconate kinase were identified, consistent with our previous experimental result that the strain cannot utilize gluconate [1]. The genome lacks two enzymes (ldh, alpha ketoglutarate dehydrogenase and lgd, alpha ketoglutarate decarboxylase), making the oxidative and reductive branches of the tricarboxylic acid cycle operate separately. The genome has all the genes for the synthesis of fatty acids, 20 amino acids, and corresponding tRNAs. Although full sets of genes for the biosynthesis of purine and pyrimidine were identified, enzymes for the salvage pathways of purine (apt, adenine phosphoribosyltransferase; ade, adenine deaminase) and pyrimidine (cad, cytidine deaminase; ctdA, cytosine deaminase; tdh, thymidine kinase; denA, thymidine phosphorylase; upp, uracil phosphoribosyltransferase; udh, uridine kinase; and udp, uridine phosphorylase) were absent. The plasmid encodes some metabolic enzymes, such as those participating in glycolysis, the pentose phosphate pathway, and the citric acid cycle. However, it is worth noting that the plasmid has a gene (6-phosphogluconate dehydrogenase) that is the only copy in the genome (PHZ_p0183).

Like most other species in the genus *Phenylobacterium*, the strain is able to use L-phenylalanine as a sole carbon source under aerobic conditions [1]. A recent study revealed that phenylalanine can be completely degraded through the homogentisate pathway in *Pseudomonas putida* U [7]. *P. zucineum* may use the same strategy to utilize phenylalanine, because all the enzymes for the conversion of phenylalanine through intermediate homogentisate to the final products fumarate and acetoacetate are present in the chromosome (Table 3).

**Functional elements responding to environmental transition**

HLK1 is able to survive intracellularly and extracellularly. Consistently, the genome contains the fundamental elements to support the life cycle in different environments. The genome contains abundant two-component signal transduction proteins, transcriptional regulators, and heat shock response proteins, enabling the strain to respond to extra- and intra-cellular stimuli at transcriptional and post-translational levels. Among the total of 102 two-component signal transduction proteins (91 in the chromosome and 11 in the plasmid), there are 36 histidine kinases, 48 response regulators, and 18 hybrid proteins fused with histidine kinase and response regulator. Sixteen pairs of histidine kinase and response regulator (1 in the plasmid) are adjacent and may act as functional operons. These tightly linked modules make two-component signal transduction systems respond to environmental changes efficiently. The genome encodes 170 transcriptional regulators (16 in the plasmid) (Table 4). Notably, we annotated the proteins of 93 bacteria (see...
methods – comparative genomics) with the same annotation criteria used for *P. zucineum* and found that the fraction of two-component signal transduction proteins and transcriptional regulators was positively correlated with the capacity for environmental adaptation (Figure 2). The genome contains 17 extracytoplasmic function (ECF) sigma factors (3 in the plasmid) (Table 5). ECFs are suggested to play a role in environmental adaptation for *Pseudomonas putida* KT2440, whose genome contains 19 ECFs [8]. *P. zucineum* has 3 heat shock sigma factors (2 in the plasmid) and 33 heat shock molecular chaperons (17 in the plasmid) (Table 6), which can cope with a variety of stresses, including cellular energy depletion, extreme concentrations of heavy metals, and various toxic substances. [9].

The genes for cell motility include 3 chemotaxis operons, 7 MCP (methyl-accepting chemotaxis) genes, 15 other genes related to chemotaxis (Table 7), and 43 genes for the biogenesis of the flagellum (Table 8).

The genome contains sec-dependent, sec-independent, typical type II (Table 9) and IV secretion systems (Table 10), which are known to play important roles in adapting to diverse conditions [10,11].

To better understand the roles of proteins responsible for environmental transition, we computed the distributions of those proteins in 5 representative alphaproteobacteria with typical habitats (see methods – comparative genomics). Like other multiple bacteria and facultative bacteria, which can survive in multiple niches, *P. zucineum* encodes a higher fraction of ECFs, transcriptional regulators and two-component signal transduction proteins than obligate bacteria (Table 9). Notably, *P. zucineum* has the largest number of heat shock related proteins (Table 6), in comparison to the 5 representative alphaproteobacteria

### Table 3: Phenylalanine-degrading enzymes in the *P. zucineum* genome

| Gene | *P. zucineum* Locus | Length (bp) | Alignment coverage (%) | Score | Amino acid Identity (%) | Gene name                        |
|------|---------------------|-------------|------------------------|-------|------------------------|----------------------------------|
|      |                     |             |                        |       |                        | phenylalanine-4-hydroxylase       |
| phhA | PHZ_c1409           | 262         | 308                    | 83.59 | 71.75                  | carbinolamine dehydratase         |
| phhB | PHZ_c0077           | 118         | 97                     | 79.66 | 93.81                  | tyrosine aminotransferase         |
| tryB | PHZ_c1644           | 398         | 406                    | 60.05 | 57.39                  | 4-hydroxyphenylpyruvate dioxygenase |
| hpd  | PHZ_c2833           | 358         | 374                    | 98.32 | 93.58                  | homogenitase 1,2-dioxygenase      |
| hmgA | PHZ_c2831           | 433         | 377                    | 60.28 | 67.64                  | fumarylacetocetase hydrolase      |
| hmgB | PHZ_c0313           | 430         | 226                    | 9.77  | 18.14                  | maleylacetocetate isomerase       |
| hmgC | PHZ_c0314           | 210         | 212                    | 98.1  | 98.11                  |                                 |

### Table 4: Transcriptional regulators in the *P. zucineum* genome

| Family name | Action type           | Chromosome | Plasmid | Proposed roles                           |
|-------------|-----------------------|------------|---------|------------------------------------------|
| AsnC family | Activator/repressor   | 8          | 0       | Amino acid biosynthesis                   |
| AraC family | Activator             | 10         | 1       | Carbon metabolism, stress response and pathogenesis |
| ArsR family | Repressor             | 8          | 0       | Metal resistance                          |
| BlaI family | Repressor             | 2          | 0       | Penicillin resistance                     |
| Cold shock family | Activator     | 6          | 0       | Low-temperature resistance                |
| Cro/CI family | Repressor         | 9          | 2       | Unknown\(^2\)                             |
| Crp/Fnr family | Activator/repressor | 7          | 2       | Global responses, catabolite repression and anaerobiosis |
| GntR family | Repressor             | 7          | 0       | General metabolism                        |
| Lact family | Repressor             | 4          | 0       | Carbon source utilization                 |
| LuxR family | Activator             | 5          | 1       | Quorum sensing, biosynthesis and metabolism, etc. |
| LysR family | Activator/repressor   | 15         | 1       | Carbon and nitrogen metabolism            |
| MarR family | Activator/repressor   | 6          | 0       | Multiple antibiotic resistance            |
| MerR family | Repressor             | 9          | 2       | Resistance and detoxification             |
| TetR family | Repressor             | 22         | 0       | Biosynthesis of antibiotics, efflux pumps, osmotic stress, etc. |
| XRE family | Repressor             | 2          | 2       | Unknown (initial function is lysisogen maintenance) |
| Other types\(^2\) | Repressor         | 34         | 5       | -                                        |
| Total      |                      | 154        | 16      | -                                        |

\(^1\)Initial function is related to controlling the expression of phage gene  
\(^2\)"Other types" include the transcriptional regulators with only one member in the *P. zucineum* genome or transcriptional regulators that could not be classified into any known family.
and 93 bacteria (data not shown). Among the plasmid-encoded heat shock related proteins are 2 RpoH (PHZ_p0049 and PHZ_p0288) and 2 DnaK-GrpE clusters (PHZ_p0053-0054 and PHZ_p0121-0122). Further phylogenetic analysis suggested that the plasmid-encoded DnaK-GrpE clusters may have undergone a genus-specific gene duplication event (Figure 3C &3D).

Adaptation to an intracellular life cycle

To survive intracellularly, *P. zucineum* must succeed in adhering to and subsequently invading the host cell [12], defending against a hostile intracellular environment [13-16], and capturing iron at very low concentration [17].

It is well known that the pilus takes part in adhering to and invading a host cell [12]. We identified one pili biosynthesis gene (*pilA*) and 2 operons for pili biosynthesis (Table 11).

The genes involved in defense against oxidative stress include superoxide dismutase (PHZ_c0927, PHZ_c1092), catalase (PHZ_c2899), peroxiredoxin (PHZ_c1548), hydroperoxide reductase (*ahpF*, alkyl hydroperoxide...

---

**Table 5: Extracytoplasmic function (ECF) sigma factors in the *P. zucineum* genome**

| Locus      | Location of proteins | Genomic element | Genomic element 5'-end | Genomic element 3'-end | COG category |
|------------|----------------------|-----------------|-------------------------|-------------------------|--------------|
| PHZ_p0151  | Plasmid              | 171,032         | 170,316                 | COG1595                 |
| PHZ_p0174  | Plasmid              | 208,703         | 208,053                 | COG1595                 |
| PHZ_p0192  | Plasmid              | 229,133         | 228,316                 | COG1595                 |
| PHZ_c0249  | Chromosome           | 249,840         | 250,553                 | COG1595                 |
| PHZ_c0301  | Chromosome           | 296,299         | 295,706                 | COG1595                 |
| PHZ_c1475  | Chromosome           | 1,676,920       | 1,677,492               | COG1595                 |
| PHZ_c1529  | Chromosome           | 1,730,783       | 1,731,403               | COG1595                 |
| PHZ_c1531  | Chromosome           | 1,732,219       | 1,732,800               | COG1595                 |
| PHZ_c1907  | Chromosome           | 2,134,971       | 2,135,507               | COG1595                 |
| PHZ_c2171  | Chromosome           | 2,447,581       | 2,448,396               | COG1595                 |
| PHZ_c2233  | Chromosome           | 2,526,836       | 2,527,369               | COG1595                 |
| PHZ_c2394  | Chromosome           | 2,724,759       | 2,725,307               | COG1595                 |
| PHZ_c2577  | Chromosome           | 2,965,250       | 2,964,390               | COG1595                 |
| PHZ_c2585  | Chromosome           | 2,970,368       | 2,969,811               | COG1595                 |
| PHZ_c2684  | Chromosome           | 3,077,272       | 3,076,727               | COG1595                 |
| PHZ_c0569  | Chromosome           | 605,441         | 604,233                 | COG4941                 |
| PHZ_c3154  | Chromosome           | 3,582,010       | 3,583,269               | COG4941                 |

1COG1595, DNA-directed RNA polymerase specialized sigma subunit, sigma24 homolog; 2COG4941, predicted RNA polymerase sigma factor containing a TPR repeat domain

---

**Table 6: Distribution of heat shock related proteins in *P. zucineum* and representative alphaproteobacteria with different living habitats**

| Content/Species | S. melliloti | B. suis | C. crescentus | P. zucineum | R. conorii | G. oxydans |
|-----------------|--------------|---------|---------------|-------------|------------|-----------|
| rpoH, heat shock sigma factor | 2 | 2 | 1 | 1 | 2 | 1 |
| dnaK, molecular chaperone (Hsp70) | 1 | 1 | 1 | 1 | 2 | 1 |
| grpE, molecular chaperone (co-chaperonin of Hsp70) | 1 | 1 | 1 | 1 | 2 | 1 |
| dnaK-like molecular chaperone | 1 | 1 | 1 | 1 | 0 | 1 |
| dnaJ, molecular chaperone | 1 | 1 | 1 | 1 | 0 | 1 |
| dnaJ-like molecular chaperone | 4 | 3 | 3 | 6 | 2 | 3 |
| groEL, molecular chaperone (hsp60) | 5 | 1 | 1 | 1 | 1 | 1 |
| groES, molecular chaperone (Hsp10, co-chaperonin of Hsp60) | 3 | 1 | 1 | 1 | 1 | 1 |
| molecular chaperone Hsp20 | 5 | 2 | 2 | 3 | 9 | 0 |
| molecular chaperone Hsp33 | 1 | 1 | 1 | 1 | 0 | 0 |

1rpoH may be responsible for the expression of some or all heat shock proteins
2The function of molecular chaperones is to protect unfolded proteins induced by stress factors through renaturation or degradation in cooperation with protease.
Comparative genomic analysis demonstrated that the species share only 57.8% (2,231/3,861) of orthologous proteins. Categories J (translation, ribosomal structure and biogenesis), F (nucleotide transport and metabolism), and L (replication, recombination and repair) are the top 3 conservative COG categories between the species, sharing 88.01%, 81.67%, and 80.65% of the orthologs, respectively.

**Comparison of cell cycle genes between P. zucineum and C. crescentus**

Since *P. zucineum* is phylogenetically closest to *C. crescentus*, and since the latter is a model organism for studies of the prokaryotic cell cycle [19,20], we compared the genes regulating the cell cycle between these species.

The cell cycle of *C. crescentus* is controlled to a large extent by the master regulator CtrA, which controls the transcription of 95 genes involved in the cycle [19,20]. On the other hand, *ctrA* is regulated at the levels of transcription, phosphorylation, and proteolytic degradation by its target genes, e.g., DNA methyltransferase (*CcrM*) regulates the transcription of *ctrA*, histidine kinases (*CckA*, *PleC*, *DivJ*, *DivL*) regulate its activity, and ClpXP degrades it. These regulatory ‘loops’ enable CtrA to precisely control the progression of the cell cycle.

*P. zucineum* has most of the orthologs mentioned above (Table 13). Among the 95 CtrA-regulated genes in *C. crescentus*, 75 have orthologs in the *P. zucineum* genome (Additional file 1). The fraction of CtrA-regulated genes with orthologs in *P. zucineum* (76.9%, 73/95) is significantly greater than the mean level of the whole genome (57.8%, 2,231/3,861), indicating that the CtrA regulatory system is highly conserved. Genes participating in regulating central events of the cell cycle, such as CcrM (CC0378), Clp protease (CC1963) and 14 regulatory proteins, except for one response regulator (CC3286), are present in the *P. zucineum* genome. The genes without counterparts in *P. zucineum* are mostly for functionally unknown proteins.

Notably, the sequence of CtrA is strikingly similar between *P. zucineum* and *C. crescentus*, with 93.07% identity of amino acid sequence and 89.88% identity of nucleotide sequence. In addition, they share identical promoters (p1 and p2) [21] and the motif (GAnTC) recognized by DNA methyltransferase (*CcrM*) (Figure 6) [22], suggesting that they probably share a similar regulatory loop of CtrA.

Consistent with the results from in silico sequence analysis, the CtrA of *P. zucineum* can restore the growth of temperature-sensitive strain LC2195 (a CtrA mutant) of *C. crescentus* [23] at 37°C, indicating that the CtrA of *P. zucin-
neum can functionally compliment that of C. crescentus in our experimental conditions (data not shown).

Taken together, the comparative genomics of P. zucineum and C. crescentus suggests that the cell cycle of the former is likely to be regulated similarly to that of the latter.

**Presence of ESTs of the strain in human**

Since P. zucineum strain HLK1T can invade and persistently live in several human cell lines [1], we were curious about whether this microbe can infect humans. By blasting against the human EST database (dbEST release 041307 with 7,974,440 human ESTs) with the whole genome sequence of P. zucineum, we found 9 matched ESTs (Table 14), of which 3 were from a library constructed from tissue adjacent to a breast cancer, and 6 were from a library constructed from a cell line of lymphatic origin. The preliminary data suggest that P. zucineum may invade humans.

**Conclusion**

This work presents the first complete bacterial genome in the genus Phenyllobacterium. Genome analysis reveals the fundamental basis for this strain to invade and persistently survive in human cells. P. zucineum is phylogeneti-
Table 8: Flagella genes in the *P. zucineum* genome

| Locus      | 5'-end | 3'-end | Name                              | Gene symbol | Proposed role       |
|------------|--------|--------|-----------------------------------|-------------|---------------------|
| PHZ_c0080 | 75,413 | 76,462 | flagellin modification protein FlmA | flmA        | regulator           |
| PHZ_c0081 | 76,467 | 77,621 | flagellin modification protein FlmB | flmB        | regulator           |
| PHZ_c0745 | 816,772| 818,034| flagellar hook-length control protein FlkK | flkK        | flagellar structure |
| PHZ_c0787 | 868,051| 866,696| flagellar hook protein FlgE        | flgE        | flagellar structure |
| PHZ_c0788 | 868,820| 868,171| flagellar hook assembly protein FlgD | flgD        | flagellar structure |
| PHZ_c0789 | 870,604| 868,865| flagellar hook length determination protein flagE | flagE       | flagellar structure |
| PHZ_c0790 | 870,819| 872,918| flagellar hook-associated protein flaN | flaN        | flagellar structure |
| PHZ_c0791 | 872,933| 873,862| flagellin and related hook-associated proteins - flagellar structure | -           | flagellar structure |
| PHZ_c0853 | 945,008| 946,354| flagellum-specific ATP synthase Fil | flil        | protein export ATPase |
| PHZ_c0854 | 946,354| 946,758| fli protein                        | fli         | flagellar structure |
| PHZ_c0857 | 950,714| 948,621| flagellar biosynthesis protein FlmA | flmA        | export apparatus    |
| PHZ_c0859 | 952,470| 952,138| flagellar motor switch protein FlmN | flmN        | motor               |
| PHZ_c0860 | 953,126| 952,479| fliE protein                       | fliE        | regulator           |
| PHZ_c0861 | 954,151| 953,126| flagellar motor switch protein FlgG | flgG        | flagellar structure |
| PHZ_c0862 | 955,794| 954,151| flagellar M-ring protein FlfF       | flfF        | flagellar structure |
| PHZ_c0913 | 1,007,753| 1,006,992| flagellar L-ring protein FlgH       | flgH        | flagellar structure |
| PHZ_c0914 | 1,008,508| 1,007,753| distal basal-body ring component protein FlaD | flad        | flagellar structure |
| PHZ_c0915 | 1,009,300| 1,008,515| flagellar basal-body rod protein FlgG | flgG        | flagellar structure |
| PHZ_c0916 | 1,010,052| 1,009,318| flagellar basal-body rod protein FlgF | flgF        | flagellar structure |
| PHZ_c0917 | 1,017,085| 1,016,351| flagellar biosynthesis protein FliP | fliP        | export apparatus    |
| PHZ_c0918 | 1,017,272| 1,010,874| flagellar basal-body-associated protein FlIL | flIL        | flagellar structure |
| PHZ_c0919 | 1,017,942| 1,016,351| flagellar biosynthesis protein FliP | fliP        | flagellar structure |
| PHZ_c0920 | 1,017,924| 1,018,355| flagellar basal-body rod protein FlgC | flgC        | flagellar structure |
| PHZ_c0921 | 1,018,370| 1,018,678| flagellar hook-basal body complex protein FlIE | flIE        | flagellar structure |
| PHZ_c0922 | 1,021,796| 1,022,056| flagellar biosynthesis protein FilQ | flIQ        | export apparatus    |
| PHZ_c0923 | 1,022,079| 1,021,796| flagellar biosynthesis protein FilR | flIR        | export apparatus    |
| PHZ_c0924 | 1,022,837| 1,022,837| flagellar biosynthesis protein FilR | flIR        | export apparatus    |
| PHZ_c0925 | 1,023,913| 1,023,913| flagellar biosynthesis protein FilB | flIB        | export apparatus    |
| PHZ_c1380 | 1,563,281| 1,562,745| putative flagellin accessory protein FlaCE | flaCE       | flagellar structure |
| PHZ_c1381 | 1,565,145| 1,563,358| flagellin modification protein FlmG | flmG        | regulator           |
| PHZ_c1382 | 1,565,343| 1,565,765| flagellar repressor protein FblT    | flbT        | regulator           |
| PHZ_c1383 | 1,565,782| 1,566,093| flagellar biosynthesis regulator FlaF | flaf        | regulator           |
| PHZ_c1384 | 1,566,375| 1,567,202| flagellin FlfM                     | flfM        | flagellar structure |
| PHZ_c1385 | 1,567,469| 1,568,314| flagellin FlfM                     | flfM        | flagellar structure |
| PHZ_c1386 | 1,568,434| 1,568,724| flagellin FlfG                     | flfG        | flagellar structure |
| PHZ_c1387 | 1,568,887| 1,569,720| flagellin FlfL                     | flfL        | flagellar structure |
| PHZ_c1395 | 2,168,522| 2,169,634| flagellar P-ring protein FglI       | fglI        | flagellar structure |
| PHZ_c1397 | 2,169,942| 2,170,382| flagellar basal-body rod protein FlbY | flbY        | flagellar structure |
| PHZ_c2595 | 2,982,550| 2,983,593| flagellin modification protein FlmD | flmD        | regulator           |
| PHZ_c2597 | 2,984,874| 2,986,508| flagellin modification protein FlmG | flmG        | regulator           |
| PHZ_c2599 | 2,989,315| 2,989,974| flagellin modification protein FlmC | flmC        | regulator           |
| PHZ_c2600 | 2,990,549| 2,989,977| flagellin modification protein FlmH | flmH        | regulator           |

**Methods**

**Bacterial growth and genomic library construction**

*P. zucineum* strain HLK1 was grown in LB (Luria-Bertani) broth at 37°C and then harvested for the preparation of genomic DNA[1]. Genomic DNA was prepared using a bacterial genomic DNA purification kit (V-Gene Biotech, Hangzhou, China) according to the manufacturer's instructions. Sheared DNA samples were fractionated to construct three different genomic libraries, containing average insert sizes of 2.0–2.5 kb, 2.5–3.0 kb and 3.5–4.0 kb. The resulting pUC18-derived library plasmids were extracted using the alkaline lysis method and subjected to direct DNA sequencing with automated capillary DNA sequencers (ABI3730 or MegaBACE1000).

**Sequencing and finishing**

The genome of *P. zucineum* was sequenced by means of the whole genome shotgun method with the phred/phrap/consed software packages [24-27]. Sequencing and
Table 9: Distributions of proteins involved in environmental adaptation in *P. zucineum* and representative alphaproteobacteria with different living habitats

| Species         | S. meliloti | B. suis | C. crescentus | P. zucineum | R. conorii | G. oxydans |
|-----------------|-------------|---------|---------------|-------------|------------|------------|
| Genome size (Mb)| 6.69        | 3.32    | 4.02          | 4.38        | 1.27       | 2.92       |
| GC content (%)  | 62.2        | 57.3    | 67.2          | 71.1        | 32.4       | 60.8       |
| Habitat         | Multiple1    | Faculative1 | Aquatic1   | Faculative2 | Obligate1 | Multiple1 |
| ECF, extracytoplasmic function sigma factor (/Mb) | 11 (1.6) | 2 (0.6) | 15 (3.7) | 17 (3.9) | 0 (0) | 2 (0.7) |
| Transcriptional regulator (/Mb) | 433 (64.7) | 149 (44.9) | 183 (45.5) | 170 (38.8) | 11 (8.7) | 89 (30.1) |
| Two-component signal transduction protein (/Mb) | 113 (16.3) | 44 (13.3) | 111 (27.6) | 102 (23.3) | 7 (5.5) | 41 (14.1) |
| Molecular chaperone | 23 | 12 | 14 | 33 | 8 | 14 |
| Flagellar protein | 41 | 37 | 42 | 43 | 10 | 40 |
| Chemotaxis protein | 42 | 4 | 48 | 41 | 0 | 11 |
| Pilus protein | 13 | 4 | 9 | 16 | 2 | 4 |
| Sec-dependent secretion system | 11 | 11 | 11 | 11 | 11 | 12 |
| Sec-independent secretion system | 4 | 4 | 4 | 4 | 3 | 4 |
| Type II secretory protein | 2 | 0 | 8 | 13 | 0 | 3 |
| Type IV secretory protein | 9 | 8 | 9 | 31 | 15 | 1 |

1The habitats of *S. meliloti*, *B. suis*, and *R. conorii* were indicated in a recent publication [42].
2According to our recent publication [1], *P. zucineum* was classified as "facultative". 3Given that *G. oxydans* is often isolated from sugary niches (such as flowers and fruits) and associated soil (such as garden soil and baker’s soil) [43], we classified *G. oxydans* as "multiple".

Table 10: Type IV secretion systems in the *P. zucineum* genome

| Locus     | Location of protein | Name                        |
|-----------|---------------------|-----------------------------|
|           | Genomic element     | 5'-end | 3'-end | type IV secretion protein, VirB1 |
| PHZ_p0007 | Plasmid             | 6,786  | 7,445  | type IV secretion protein, VirB2 |
| PHZ_p0008 | Plasmid             | 7,483  | 7,800  | type IV secretion protein, VirB3 |
| PHZ_p0009 | Plasmid             | 7,816  | 8,148  | type IV secretion protein, VirB4 |
| PHZ_p0010 | Plasmid             | 8,144  | 10,546 | type IV secretion protein, VirB5 |
| PHZ_p0011 | Plasmid             | 10,546 | 11,298 | type IV secretion protein, VirB6 |
| PHZ_p0012 | Plasmid             | 11,553 | 12,488 | type IV secretion protein, VirB7 |
| PHZ_p0013 | Plasmid             | 12,816 | 13,493 | type IV secretion protein, VirB8 |
| PHZ_p0014 | Plasmid             | 13,493 | 14,320 | type IV secretion protein, VirB9 |
| PHZ_p0015 | Plasmid             | 14,320 | 15,543 | type IV secretion protein, VirB10 |
| PHZ_p0016 | Plasmid             | 15,543 | 16,538 | type IV secretion protein, VirB11 |
| PHZ_c1506 | Chromosome          | 1,709,481 | 1,709,999 | type IV secretion protein, TraF |
| PHZ_c1508 | Chromosome          | 1,711,058 | 1,712,773 | type IV secretion protein, VirD2 |
| PHZ_c1509 | Chromosome          | 1,712,790 | 1,714,763 | type IV secretion protein, VirD4 |
| PHZ_c1512 | Chromosome          | 1,716,262 | 1,717,242 | conjugal transfer protein, TrbB |
| PHZ_c1513 | Chromosome          | 1,717,242 | 1,717,559 | conjugal transfer protein, TrbC |
| PHZ_c1514 | Chromosome          | 1,717,562 | 1,717,828 | conjugal transfer protein, TrbD |
| PHZ_c1515 | Chromosome          | 1,717,836 | 1,720,283 | conjugal transfer protein, TrbE |
| PHZ_c1516 | Chromosome          | 1,720,283 | 1,721,014 | conjugal transfer protein, TrbL |
| PHZ_c1517 | Chromosome          | 1,721,238 | 1,722,398 | conjugal transfer protein, TrbF |
| PHZ_c1518 | Chromosome          | 1,722,401 | 1,723,084 | conjugal transfer protein, TrbG |
| PHZ_c1519 | Chromosome          | 1,723,087 | 1,724,064 | conjugal transfer protein, TrbH |
| PHZ_c1520 | Chromosome          | 1,724,070 | 1,725,212 | conjugal transfer protein, TrbI |
| PHZ_c2348 | Chromosome          | 2,660,517 | 2,660,813 | type IV secretion protein, VirB2 |
| PHZ_c2349 | Chromosome          | 2,660,809 | 2,661,444 | type IV secretion protein, VirB3 |
| PHZ_c2350 | Chromosome          | 2,661,119 | 2,663,497 | type IV secretion protein, VirB4 |
| PHZ_c2352 | Chromosome          | 2,664,374 | 2,665,309 | type IV secretion protein, VirB6 |
| PHZ_c2353 | Chromosome          | 2,665,482 | 2,666,159 | type IV secretion protein, VirB8 |
| PHZ_c2354 | Chromosome          | 2,666,159 | 2,667,004 | type IV secretion protein, VirB9 |
| PHZ_c2355 | Chromosome          | 2,667,004 | 2,668,041 | type IV secretion protein, VirB10 |
| PHZ_c2356 | Chromosome          | 2,668,046 | 2,669,035 | type IV secretion protein, VirB11 |
| PHZ_c2357 | Chromosome          | 2,669,091 | 2,670,872 | type IV secretion protein, VirD4 |
subsequent gene identification was carried out as described in our earlier publications [28-30]. Briefly, during the shotgun sequence phase, clones were picked randomly from three shotgun libraries and then sequenced from both ends. 44,667 successful sequence reads (>100 bp at Phred value Q13), accounting for 5.47× sequence coverage of the genome, were assembled into 563 sequence contigs representing 60 scaffolds connected by end-pairing information.

The finishing phase involved iterative cycles of laboratory work and computational analysis. To reduce the numbers of scaffolds, reads were added into initial contig assembly by using failed universal primers as primers and by using plasmid clones that extended outwards from the scaffolds as sequence reaction templates. To resolve the low-quality regions, resequencing of the involved reads in low quality regions with universal primers and primer walking the plasmid clones were the first choice, otherwise, rese-

### Table 11: Pilus proteins in the *P. zucineum* genome

| Locus   | 5'-end | 3'-end | Name            | Gene symbol |
|---------|--------|--------|-----------------|-------------|
| PHZ_c0356 | 362,116 | 362,289 | pilus subunit protein PilA | pilA         |
| PHZ_c2992 | 3,412,800 | 3,413,318 | Flp pilus assembly protein TadG | tadG         |
| PHZ_c2995 | 3,415,220 | 3,415,468 | Flp pilus assembly protein, pilin Flp | -           |
| PHZ_c2996 | 3,415,532 | 3,416,023 | Flp pilus assembly protein, protease CpaA | cpaA         |
| PHZ_c2997 | 3,416,039 | 3,416,899 | pilus assembly protein CpaB | cpaB         |
| PHZ_c2998 | 3,416,899 | 3,418,350 | pilus assembly protein CpaC | cpaC         |
| PHZ_c2999 | 3,418,355 | 3,419,587 | pilus assembly protein CpaE | cpaE         |
| PHZ_c3000 | 3,419,594 | 3,420,991 | pilus assembly protein CpaF | cpaF         |
| PHZ_c3001 | 3,420,103 | 3,421,944 | Flp pilus assembly protein TadB | tadB         |
| PHZ_c3002 | 3,421,944 | 3,422,903 | Flp pilus assembly protein TadC | tadC         |
| PHZ_c3027 | 3,451,637 | 3,452,566 | Flp pilus assembly protein CpaB | cpaB         |
| PHZ_c3028 | 3,452,580 | 3,453,893 | Flp pilus assembly protein, secretin CpaC | cpaC         |
| PHZ_c3029 | 3,453,893 | 3,455,056 | Flp pilus assembly protein, ATPase CpaE | cpaE         |
| PHZ_c3030 | 3,455,059 | 3,456,489 | Flp pilus assembly protein ATPase CpaF | -           |
| PHZ_c3031 | 3,456,489 | 3,457,445 | Flp pilus assembly protein CpaF | cpaF         |
| PHZ_c3032 | 3,457,492 | 3,458,391 | Flp pilus assembly protein TadC | tadC         |

Figure 3

**Neighbor-joining trees of 5 representative alphaproteobacteria and *P. zucineum*, inferred from (A) 16S rRNA genes, (B) RpoH proteins, (C) DnaK proteins and (D) GrpE proteins.** The node labels are bootstrap values (100 replicates). The plasmid-encoded DnaK and GrpE of *P. zucineum* may have undergone a genus-specific gene duplication event (C &
sequencing with alternate temperature conditions resolved the remaining low-quality regions. New sequence reads obtained from the above laboratory work were assembled into existing contigs, which yielded new contigs and new scaffolds connected by end-pairing information. Then, consed interface helped us to do nest round of laboratory work based on new arisen contig assembly. After about four iterative cycles of the above "finish" procedures to close gaps and to resolve the low-quality regions, the PCR product obtained by using total genomic DNA as template was sequenced from both ends to close the last physical gap. In addition, the overall sequence quality of the genome was further improved by using the following criteria: (1) two independent high-quality reads as minimal coverage, and (2) Phred quality value = Q40 for each given base. Collectively, 3,542 successful reads were incorporated into initial assembles during the finishing phase. The final assembly was composed of two circular "contigs", of which a smaller one with a protein cluster (including repA, repB, parA and parB) related to plasmid replication was assigned as the plasmid, and the larger one was the chromosome.

**Annotation**

tRNA genes were predicted with tRNAscan-SE [31]. Repetitive sequences were detected by REPutter [32,33], coupled with intensive manual alignment. We identified and annotated the protein profiles of chromosome and plasmid with the same workstream. For the chromosome, the first set of potential CDSs in the chromosome was established with Glimmer 2.0 trained with a set of ORFs longer than 500 bp from its genomic sequence at default settings [34]. The resulting 5,029 predicted CDSs were BLAST searched against the NCBI non-redundant protein database to determine their homology [35]. 1,174 annotated proteins without the word "hypothetical" or "unknown" in their function description, and without frameshifts or in-frame stop codons, were selected as the second training set. The resulting second set of 4,018 predicted CDSs (assigned as "predicted CDSs") were searched against the NCBI non-redundant protein database. Predicted CDSs that accorded with the following BLAST search criteria were considered "true proteins": (1) 80% of the query sequence was aligned and (2) E-value ≤ 1e-10. Then, the ORFs extracted from the chromosome region among "true proteins" were searched against the NCBI non-redundant protein database. The ORFs satisfying the same criteria as true proteins were considered "true ORFs". Overlapping proteins were manually inspected and resolved, according to the principle we described previously [30]. The final version of the protein profile comprised three parts: true proteins, true ORFs, and predicted CDSs located in the rest of the genome. The translational start codon of each protein was identified by the widely used RBS script [36] and then refined by comparison with homologous proteins [30].

To further investigate the function of each protein, we used InterProScan to search against the InterPro protein family database [37]. The up-to-date KEGG pathway database was used for pathway analysis [38]. All proteins were searched against the COG database which included 66 completed genomes [39,40]. The final annotation was manually inspected by comprehensively integrating the results from searching against the databases of nr, COG, KEGG, and InterPro.

**Phylogenetic tree construction**

16S rRNA genes were retrieved from 63 alphaproteobacteria, *P. zucineum* and *Escherichia coli* O157:H7 EDL933. A neighbor-joining tree with bootstrapping was built using...
Figure 5
Neighbor-joining tree of the alphaproteobacteria, inferred from 16S rRNA genes. The node labels are bootstrap values (100 replicates). C. crescentus is phylogenetically the closest to P. zucineum.
MEGA [41]. The gammaproteobacterium E. coli was used as the outgroup to root the tree. To illustrate the evolutionary history of heat shock related proteins (RpoH, DnaK and GrpE), neighbor-joining trees based on the 16S rRNA genes and the above three proteins of 5 representative alphaproteobacteria (Sinorhizobium meliloti 1021, Brucella suis 1330, C. crescentus CB15, Rickettsia conorii str. Malish 7, Gluconobacter oxydans 621H), P. zucineum and E. coli O157:H7 EDL933 were constructed.

Comparative genomics

Sequence data for comparative analyses were obtained from the NCBI database ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria/. The database has 520 completely sequenced bacterial genomes (sequences downloaded on 2007/06/05). All P. zucineum ORFs were searched against the ORFs from all other bacterial genomes with BLASTP. The number of P. zucineum ORFs matched to each genome with significance (E value = 1e-10) was calculated.

Ortholog identification

All proteins encoded by one genome were BLASTP searched against a database of proteins encoded by another genome [35], and vice versa. The threshold used in these comparisons was 1e-10. Orthology was identified if two proteins were each other’s best BLASTP hit (best reciprocal match).

Data accessibility

The sequences reported in this paper have been deposited in the GenBank database. The accession numbers for chromosome and plasmid are CP000747 and CP000748, respectively.
Table 14: Human ESTs matching the genome sequences of P. zucineum

| Query GI | Sample origin       | Query GI | Sample origin       | Query Length | Query Position | Chromosome Position | Score | E Value | Similarity (%) |
|----------|---------------------|----------|---------------------|--------------|----------------|---------------------|-------|----------|----------------|
| 14251638 | Breast tissue¹      | 226      | 41                  | 175          | 1,276,914       | 1,277,048           | 204   | 2.00E-53 | 94.07          |
| 8261474  | Breast tissue       | 116      | 1                   | 108          | 1,277,042       | 1,276,937           | 167   | 2.00E-42 | 96.31          |
| 4251634  | Breast tissue       | 142      | 19                  | 134          | 1,277,054       | 1,276,937           | 204   | 1.00E-53 | 97.46          |
| 33194938 | Lymphatic cell line² | 441      | 8                   | 441          | 1,029,575       | 1,029,142           | 749   | 0        | 96.77          |
| 33194696 | Lymphatic cell line | 652      | 8                   | 652          | 1,029,575       | 1,028,931           | 1,166 | 0        | 97.67          |
| 33193754 | Lymphatic cell line | 654      | 8                   | 654          | 1,029,575       | 1,028,929           | 1,191 | 0        | 98.15          |
| 7117824  | Lymphatic cell line | 405      | 7                   | 405          | 1,558,831       | 1,558,433           | 735   | 0        | 98.25          |
| 33194587 | Lymphatic cell line | 638      | 7                   | 638          | 2,864,470       | 2,863,838           | 1,191 | 0        | 98.89          |
| 7114909  | Lymphatic cell line | 347      | 6                   | 347          | 3,498,624       | 3,498,283           | 654   | 0        | 99.12          |

¹All of three sequences come from the library BN0075 containing 182 ESTs; the original dataset was produced by a modification of the EST sequencing strategy ORESTES (open reading frame expressed sequences tags)[44,45].

²All six sequences come from the library NIH_MGC_51 containing 2,381 ESTs; the original dataset was produced and released by the "Mammalian Gene Collection" project [46].

Abbreviations
EST: Expressed Sequence Tag; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Authors' contributions
XH and SH designed the project; YL, XX, ZD, ZL, ZY and JS performed the research; SH and BZ contributed new reagents/analytical tools; YL, XX, and ZD analyzed the data; and XH, YL, and SH wrote the paper. All authors read and approved the final manuscript.

Additional material

Additional file 1
Supplemental Table 1 Comparison of genes directly regulated by CtrA between P. zucineum and C. crescentus.
Click here for file [http://www.biomedcentral.com/content-supplementary/1471-2164-9-386-S1.xls]

Acknowledgements
This work was supported in part by the Cheung Kong Scholars Programme (National Ministry of Education, China, and the Li Ka Shing Foundation, Hong Kong) to XH, a Natural Science Foundation of China grant (30672382) to XH, and a Zhejiang Natural Science Foundation, China, grant (R204204) to XH. We thank Dr. Lucy Shapiro (Department of Developmental Biology, Stanford University) for the gifts of the C. crescentus temperature sensitive strain LC2195 and the plasmid pSAL14. We are grateful to Dr. Iain Bruce (Department of Physiology, Zhejiang University School of Medicine) for English editing.

References
1. Zhang K, Han W, Zhang R, Xu X, Pan Q, Hu X. Phenylobacterium zucineum sp. nov., a facultative intracellular bacterium isolated from a human erythroleukemia cell line K562. Syst Appl Microbiol 2007, 30(3):207-212.
2. Lingens F, Blecher R, Blecher H, Blobel F, Eberspacher J, Frohner C, Gorisch H, Gorisch H, Layh G. Phenylobacterium immobile gen. nov., sp. nov., a gram-negative bacterium that degrades the herbicide chloridazon. Int J Syst Bacteriol 1985, 35:26-39.
3. Kanse S, Patel BK. Phenylobacterium itiniforme sp. nov., a moderately thermophilic bacterium from a subsurface aquifer, and emended description of the genus Phenylobacterium. Int J Syst Evol Microbiol 2004, 54(Pt 6):2141-2146.
4. Tiago I, Mendes V, Pires C, Morais P, Verssimo A: Phenylbutanobacterium falso sp. nov., an Alphaproteobacterium isolated from a nonsaline alkaline groundwater, and emended description of the genus Phenyllobacterium. Syst Appl Microbiol 2005, 28(4):295-302.
5. Adam Z, Im YW, Ten LN, Lee ST: Phenylobacterium koreense sp. nov., isolated from South Korea. Int J Syst Evol Microbiol 2005, 55(Pt 5):2001-2005.
6. Crosson S, McGrath PT, Stephens C, McAdams HH, Shapiro L: Conservation of modular design of an oxygen sensory/signaling network with species-specific output. Proc Natl Acad Sci U S A 2005, 102(22):8018-8023.
7. Arias-Barrau E, Olivera ER, Luengo JM, Fernandez C, Galan B, Garcia JL, Diaz E, Minambres B: The homogenisate pathway: a central catabolic pathway involved in the degradation of L-phenylalanine, L-tyrosine, and 3-hydroxyphenylacetate in Pseudomonas putida. J Biol Chem 2004, 186(15):5062-5077.
8. Martinez-Bueno MA, Toes R, Rey M, Ramos JL: Detection of multiple extracytoplasmic function (ECF) sigma factors in the genome of Pseudomonas putida KT2440 and their counterparts in Pseudomonas aeruginosa PA01. Environ Microbiol 2003, 52(3):621-630.
9. Missiakas D, Raina S: The extracytoplasmic function sigma factors: role and regulation. Mol Microbiol 1998, 28(6):1059-1066.
10. Pallen MJ, Chaudhuri RR, Henderson IR: Genomic analysis of secretion systems. Curr Opin Microbiol 2003, 6(5):519-527.
11. Wickner W, Schekman R: Protein translocation across biological membranes. Science 2005, 310(5753):1452-1456.
12. Pizarro-Cerda J, Cossart P: Bacterial adhesion and entry into host cells. Cell 2006, 124(4):715-727.
13. Roop RM 2nd, Bailleu BH, Valderramas MV, Cardelli JA: Adaptation of the Brucellae to their intracellular niche. Mol Microbiol 2004, 52(3):621-630.
14. Miller RA, Britigan BE: Role of oxidants in microbial pathophysiology. Clin Microbiol Rev 1997, 10(1):1-18.
15. Master SS, Springer B, Sander P, Boettger EC, Deretic V, Timmins GS: Oxidative stress response genes in Mycobacterium tuberculosis: role of ahpC in resistance to peroxynitrite and stage-specific survival in macrophages. Microbiology 2002, 148(Pt 1):3139-3144.
16. Nathan C, Shihol MU: Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. Proc Natl Acad Sci U S A 2000, 97(16):8841-8848.
17. Rataratde C, Dover LG: Iron metabolism in pathogenic bacteria. Annu Rev Microbiol 2000, 54:881-941.

18. Siddibingham TV, Laub MT, Paulsen IT, Nelson KE, Eisen JA, Heidelberg JF, Allevy MR, Ohta N, Maddock JR, Potocka I, Nelson WC, Newton A, Stephens C, Phadke ND, Ely B, DeBoy RT, Dodson RJ, Durkin AS, Gwinn ML, Haf DF, Kolonay JF, Smit J, Craven MB, Khouri H, Shetty J, Berry K, Utterback T, Tran K, Wolf A, vanBermameth J, Ermakoff DM, White O, Salzberg SL, Venter JC, Shapiro L, Fraser C: Complete genome sequence of Caulobacter crescentus. Proc Natl Acad Sci U S A 2001, 98(7):4136-4141.

18. Kurtz S, Laub MT: Cell-cycle progression and the generation of asymmetry in Caulobacter crescentus. Nat Rev Microbiol 2004, 2(4):325-337.

19. Reisenauer A, Kahng LS, McCollum S, Shapiro L: Parte M, Kasthan N, Alon U: REPuter: the manifold applications of repeat analysis from comparative genomics of bacterial cell-cycle regulator. Proc Natl Acad Sci U S A 2001, 98(3):195-202.

20. Quon KC, Marczynski GT, Shapiro L: REPuter: fast computation of maximal repeat analysis. Bioinformatics 2003, 19(3):44-41.

21. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1998, 27(22):4636-4641.

22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389-3402.

23. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawasaki S, Katayama T, Araki M, Hiraoka M: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2003, 31(3):387-390.

24. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. Science 1997, 278(5338):631-637.

25. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhodiev SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Wallner V, Wolf YI, Yin JJ, Natale DA: The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003, 4:41.

26. Parter M, Kashant N, Alon U: Environmental variability and modularity of bacterial metabolic networks. BMC Biol 2007/09/25 edition. 2007, 7:169.

27. Gupta A, Singh VK, Qazi GN, Kumar A: Glucanobacter oxydans: its biotechnological applications. J Mol Microbiol Biotechnol 2001/05/22 edition. 2001, 3(3):445-456.

28. Camargo AA, Samaia HP, Diaz-Neto E, Simao DF, Silva WA Jr., Silva RA, Sousa JF, Stecconi D, Tsubouchi Y, Valente V, Soares M, Nunes DN, Correa RG, Salberg H, Carvalho AF, Reis LF, Brentani RR, Simpson AJ, dos Souza S: The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proc Natl Acad Sci U S A 2001, 98(21):12103-12108.

29. Dias Neto E, Correa RG, Verjovski-Almeida S, Briones MR, Nagai MA, da Silva WR, Jorge MA, Bordin S, Costa FF, Goldmann GH, Carvalho AF, Matsukuma A, Baia GS, Simpson DH, Brunstein A, de Oliveira PS, Bucher P, Jongeneel CV, O'Hare MJ, Soares F, Brentani RR, Reis LF, de Souza SJ: Shotgun sequencing of the human transcriptome with a customized OntoExpressed sequence tag database. Proc Natl Acad Sci U S A 2000, 97(7):3491-3496.

30. Gerhard DS, Wagner L, Feingold EA, Shenmen CH, Grouse LH, Schulker G, Klein SL, Old S, Rasoloy R, Good P, Guyer M, Peck AM, Derge JG, Lipman D, Collins FS, Jang W, Sherry S, Feolo M, Misquita L, Lee E, Romistiropovsky K, Greenhut SF, Schaefer CF, Becket D, Brandt TI, Haasler D, Kent J, Kiekkas M, Furey T, Trent M, Prange C, Scherier K, Shapiro N, Blat NK, Hopkins RF, Hsie K, Driscoll T, Soboros MB, Casavant TL, Schettle TE, Brownstein MI, Usdin TB, Toshiyuki S, Carnacci P, Piao Y, Dudaekel DB, Ko MS, Kawakami K, Suzuki Y, Sugano S, Gruber CE, Smith MR, Simmons B, Moore T, Izumi S, Hashimoto E, Dixon S, Munden S, Gonzalez P, Gao X, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389-3402.

31. Yu J, Hu S, Wang J, Wu T, Li S, Li S, Bao B, Diao L, Zhou W, Yang J, Zhan S, Zhao X, Cao M, Lu J, Sun T, Tang J, Chen Y, Huang F, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Gao W, Hu D, Huang X, Li W, Li J, Liu Z, Li L, Liu Q, Qi L, Liu J, Li T, Wang X, Lu H, Wu T, Zhi H, Ni P, Han J, An R, Xin F, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhan J, He X, Zhan J, Xu J, Zhan K, Zhan X, Zhe J, Zeng J, Zeng T, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian X, Chick A, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Liu W, Chen W, Xiang W, Zhang T, Hu J, Wang J, Liu S, Yang J, Zhang X, Xiong Y, Li J, Mao L, Zhou C, Zhu J, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zou L, Yuan Y, Yang J: A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 2002, 296(5573):79-92.

32. Chen C, Tang J, Dong W, Wang C, Feng Y, Wang J, Zheng F, Pan X, Liu D, Li M, Song Y, Zhou X, Xu H, Feng T, Guo Z, Ju A, Ge J, Dong J, Yang Y, Sun W, Jiang Y, Wang J, Yan H, Yang W, Gao GF, Wang R, Yang J, Yu J: A genome-wide analysis of KEGG orthology in human genomic sequence. BMC Genomics 2008, 9:386 http://www.biomedcentral.com/1471-2164/9/386.
Waterman R, Johnson SL, Ruan Y, Wei CL, Mathavan S, Gunaratne PH, Wu J, Garcia AM, Hulyk SW, Fuh E, Yuan Y, Sneed A, Kowis C, Hodgson A, Muzny DM, McPherson J, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodriguez S, Sanchez A, Whiting M, Madari A, Young AC, Wetherby KD, Granite SJ, Kwong PN, Brinkley CP, Pearson RL, Bouffard GG, Blakesly RW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Griffith M, Griffith OL, Krzywinski MI, Liao N, Morin R, Palmquist D, Petrescu AS, Skliska U, Smalius DE, Stott JM, Schnirch A, Schein JE, Jones SJ, Holt RA, Baross A, Marra MA, Clifton S, Makowski KA, Bosak S, Malek J: The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 2004, 14(10B):2121-2127.