Microbiology (2007), **153**, 1042–1058 DOI 10.1099/mic.0.2006/003657-0

Comparative analysis of the *Corynebacterium glutamicum* group and complete genome sequence of strain R

Hideaki Yukawa,[1,2] Crispinus A. Omumasaba,[1] Hiroshi Nonaka,[1] Péter Kós,[1]† Naoko Okai,[1] Nobuaki Suzuki,[1] Masako Suda,[1] Yota Tsuge,[1,2] Junko Watanabe,[1] Yoko Ikeda,[1] Alain A. Vertès[1] and Masayuki Inui[1]

[1]Microbiology Research Group, Research Institute of Innovative Technology for the Earth (RITE), Soraku, Kyoto 619-0292, Japan

[2]Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara 630-0101, Japan

Correspondence
Hideaki Yukawa
mmg-lab@rite.or.jp

Received 23 October 2006
Revised 19 December 2006
Accepted 2 January 2007

The complete genome sequence of *Corynebacterium glutamicum* strain R was determined to allow its comparative analysis with other corynebacteria. The biology of corynebacteria was explored by refining the definition of the subset of genes that constitutes the corynebacterial core as well as those characteristic of saprophytic and pathogenic ecological niches. In addition, the relative scarcity of corynebacterial sigma factors and the plasticity of their two-component system machinery reflect their relatively exacting nutritional requirements and reduced membrane-associated and secreted proteins. The conservation of key genes and pathways between corynebacteria, mycobacteria and *Nocardia* validates the use of *C. glutamicum* to study fundamental processes that are conserved in slow-growing mycobacteria, including pathogenesis-associated mechanisms. The discovery of 39 novel genes in *C. glutamicum* R that have not been previously reported in other corynebacteria supports the rationale for sequencing additional corynebacterial genomes to better define the corynebacterial pan-genome and identify previously undetected metabolic pathways in these organisms.

INTRODUCTION

Corynebacterium glutamicum is a fast-growing, aerobic, non-sporulating, non-motile, saprophytic, Gram-positive micro-organism that can be isolated from soil samples based on its properties to secrete large amounts of glutamic acid under suitable conditions (Kumagai, 2000; Vertès *et al.*, 2005). *Corynebacteria* belong to the order *Actinomycetales* of the eubacteria that is characterized by a high G+C content and that constitutes a different evolutionary line from that formed by low-G+C content micro-organisms such as bacilli or clostridia (Stackebrandt & Woese, 1981; Stackebrandt *et al.*, 1997). The genus *Corynebacterium* is closely related to *Mycobacterium* and *Nocardia*, among other genera, which form the *Corynebacterineae* suborder (Liebl, 2005). The genus includes both aerobes and facultative anaerobes, is phenotypically very diverse and forms a monophyletic group that exhibits considerable phylogenetic depth (Liebl, 2005). It consists of 59 validly described species, of which two taxon groups and 35 species are medically relevant (von Graevenitz & Bernard, 2001). Non-medical corynebacteria are widely disseminated in nature and have been isolated from a number of different environments other than soil, including dairy products, plant material, faeces and animal skin (Liebl, 2001). Except for phage-mediated transfer and a few conjugative plasmids, corynebacteria appear devoid of a natural competence system for exogenous DNA uptake (Vertès *et al.*, 2005). *C. glutamicum* has a long history of use for the industrial production of various primary metabolites, including amino acids and nucleotides (Demain, 2000; Hermann, 2003). Moreover, its potential as a commodity chemicals producer (Inui *et al.*, 2004) and for bioremediation applications (Shen *et al.*, 2005) is the focus of increasing research efforts.

The complete genome sequences of two variants of *C. glutamicum* ATCC 13032 have been published (Ikeda & Nakagawa, 2003; Kalinowski *et al.*, 2003), as have those of

†Present address: Institute of Plant Biology, Hungarian Academy of Sciences, Szeged, Hungary.

Abbreviations: CDS, coding sequence(s); COG, clusters of orthologous groups; ECF, extracytoplasmic function; PTS, phosphotransferase system; SSI, strain-specific island.

The DDBJ/EMBL/GenBank accession numbers for the complete sequence of the *C. glutamicum* R genome and its native episome PCgR1 are AP009044 and AP009045, respectively.

Tables of strains and plasmids, and of oligonucleotides and primers used in this study are available with the online version of this paper.

the closely related *Corynebacterium efficiens* YS-314 (Nishio *et al.*, 2003), and of the two human pathogens *Corynebacterium jeikeium* K411 (Tauch *et al.*, 2005) and *Corynebacterium diphtheriae* NCTC 13129 (Cerdeño-Tárraga *et al.*, 2003). We report here the genomic sequence of *C. glutamicum* R, a strain with industrial potential (Inui *et al.*, 2004) that was isolated from soil sampled in Japan.

The availability of these different genomic data allows the identification of the corynebacterial core genes and of those genes directly related to various ecological niches. In addition, comparative analysis of the sigma factors, secreted proteins, sugar metabolism and two-component systems present in *C. glutamicum* R enables further assessment of the industrial potential of this strain and its metabolic and regulatory specificities.

METHODS

Plasmids, bacterial strains and culture conditions. Strains and plasmids used in this study are shown in supplementary Table S1 (available with the online version of this paper). *C. glutamicum* R is a strain from our laboratory collection that was isolated in Japan from a meadow soil sample. *Escherichia coli* cells were grown at 37 °C in LB medium supplemented where necessary with 50 μg ml^{-1} of either ampicillin (Ap) or chloramphenicol (Cm) (Sambrook *et al.*, 1989). Unless otherwise stated, *C. glutamicum* and *C. efficiens* cells were grown in nutrient-rich A medium (Inui *et al.*, 2004) at 33 °C for 48 h. For sugar utilization experiments, a corynebacterial cell starter culture grown aerobically until late exponential phase in A-medium containing 40 g glucose l^{-1} was used to inoculate BT minimum medium (Inui *et al.*, 2004) containing 40 g l^{-1} of the sugar being tested, using glucose as a positive control for growth.

DNA techniques. Corynebacterial chromosomal DNA was isolated following standard methods (Sambrook *et al.*, 1989) modified by using 4 mg lysozyme ml^{-1} at 37 °C for 30 min. Cells were transformed as previously described (Vertès *et al.*, 1993b) using *E. coli* JM110 plasmid DNA. *E. coli* plasmid DNA isolation and strain transformation were performed following standard methods (Sambrook *et al.*, 1989). Restriction endonucleases, Klenow fragment and T4 DNA ligase were used as per the manufacturer's instructions (Takara). Restriction fragments were isolated from agarose gels with the GeneClean kit (Bio 101), according to the manufacturer's instructions. PCR was performed using Ex-*Taq* DNA polymerase (Takara). Prior to sequencing, exonuclease treatment was performed using ExoSAP-IT (USB), as per the manufacturer's instructions.

Library construction. Random fragments resulting from sonication of *C. glutamicum* R chromosomal DNA were separated on agarose gel into one 2–3 kb pool and another 8–9 kb pool. The fragments were blunted and ligated into *Sma*I-digested pUC119. The ligation mixture was used to transform *E. coli* JM109 and recombinants were selected on IPTG-supplemented plates. Gaps between contigs were closed using a Lambda FIX II/*Xho*I replacement phage library with a mean insert size of 20 kb, as per the manufacturer's instructions (Stratagene). Fragments (1–2 kb) at the end of the assembled contigs were amplified by PCR from the chromosomal DNA of *C. glutamicum* R, labelled using a Gene Image Random Prime Labelling Module (GE Healthcare Bio-Sciences Corp.) and used as probes to screen the phage library by plaque hybridization. Several genomic DNA fragments extracted from the positive phage clones were sequenced after subcloning into vector pUC18 (Sambrook *et al.*, 1989). For sequencing purposes, *E. coli* clones bearing *C. glutamicum* R chromosomal DNA fragments were grown overnight and the corresponding plasmids were isolated. Gaps in the assembled sequence were closed by PCR-mediated genome walking.

Genome sequencing. Using the whole-genome shotgun method, libraries of 2–3, 8–9 and 20 kb genomic inserts were sequenced from both ends using M13 universal forward and reverse primers (Sambrook *et al.*, 1989) and cycle-sequenced using the BigDye Terminator method in ABI 3700 CE and ABI 3730 DNA analysers (Applied Biosystems). The sequences were base-called and assembled using Phred, Phrap and Consed (Ewing *et al.*, 1998; Gordon *et al.*, 1998). The Pregap4 program of the Staden package (Bonfield *et al.*, 1995; Staden, 1996) was used for clipping vector sequences, as well as for quality clipping and contamination screening after base-calling by Phred (Ewing & Green, 1998; Ewing *et al.*, 1998). Gaps were closed by primer walking on gap-spanning plasmid clones and direct sequencing of PCR products. Repetitive sequences such as rDNA were confirmed by PCR. The error rate was lower than 2 bases per 10 kb as calculated using Consed (Gordon *et al.*, 1998).

Gene prediction and analysis. rRNAs were located by a BLASTN homology search against the 16S, 23S and 5S rRNA sequences of *C. glutamicum* ATCC 13032. tRNAs were predicted by tRNA scan SE (Lowe & Eddy, 1997). Protein coding sequences (CDS) were predicted using Glimmer3 (Delcher *et al.*, 1999a; Salzberg *et al.*, 1998) and GeneMarkS (Besemer *et al.*, 2001). Proteome prediction was performed by a BLASTP homology search using an e-value lower than $1 \times e^{-4}$ against GenBank release 152, GenPept release 152, UniProt release 4.6, the NCBI clusters of orthologous groups (COG) database (Tatusov *et al.*, 1997) and the Pfam family database (Accelrys GCG Wisconsin package version 11.0). Numerous annotations were checked manually after the auto-annotations were performed. The search for repeats at each extremity of each of the 10 major strain-specific islands (SSIs) present in the genome of *C. glutamicum* R (Suzuki *et al.*, 2005a) was performed using the EMBOSS package (Rice *et al.*, 2000).

Identification of orthologous CDS and genomic islands. All the CDS of *C. glutamicum* R, *C. glutamicum* ATCC 13032 (Ikeda & Nakagawa, 2003; Kalinowski *et al.*, 2003), *C. efficiens* YS-314 (Nishio *et al.*, 2003), *C. diphtheriae* NCTC 13129 (Cerdeño-Tárraga *et al.*, 2003) and *C. jeikeium* K411 (Tauch *et al.*, 2005) were compared to each other in a reciprocal manner using BLASTP with an e-value of $\times e^{-4}$ as a cut-off. Genes showing the highest similarity levels in a dual strain comparison were automatically parsed for every CDS present in both strains using an original Perl script. The CDS with a reciprocal best hit were defined as being orthologous CDS of the two strains.

Identification of genomic DNA islands in the *C. glutamicum* R and *C. glutamicum* ATCC 13032 genome was performed using MUMer2.1 (Delcher *et al.*, 1999b, 2002).

Transposon, deletion, and gene disruption and replacement mutagenesis. We used a combination of Tn*5* and Tn*31831* mutagenesis systems to assemble a library of 2300 different transposon mutants (Suzuki *et al.*, 2006; Vertès *et al.*, 2005), and the Cre-*loxP* system (Suzuki *et al.*, 2005b) to generate deletion mutants. Mutations were verified by PCR using the oligonucleotides and nucleotide primers shown in Supplementary Table S2 (available with the online version of this paper). Gene disruption and replacement mutagenesis were performed as described previously (Vertès *et al.*, 1993a) using primers indicated in Supplementary Table S2.

Gene identification numbers. Gene identification numbers are from the Virtual Institute of Microbial Stress and Survival (VIMSS) database (Alm *et al.*, 2005).

RESULTS AND DISCUSSION

General features of *C. glutamicum* R

The genetic basis of *C. glutamicum* R consists of a 49 120 bp native episome (PCgR1) and one circular chromosome of 3 314 179 bp encoding 2990 ORFs with a mean length of 957 bp (Table 1 and Fig. 1). The G+C content of the chromosome is 54.1 mol% overall, 55.2 mol% for the protein coding regions and 47.4 mol% for the non-coding regions. PCgR1 exhibits a G+C content of 53.9 mol% that is very similar to that of the chromosome, suggesting that its acquisition by strain R is not a recent event unless it was acquired by horizontal transfer from an organism with similar G+C content. It encodes 28 putative proteins, 647 bp long on average, and has a relatively low coding density (36.9 %). It does not appear to be similar to any episome previously identified in corynebacteria. The G+C content of the protein coding regions of PCgR1 is 55.4 mol% on average and that of the non-coding regions, 53.1 mol%.

The size of the *C. glutamicum* R genome is similar to that of the other saprophytic corynebacteria that have been sequenced to date, and significantly larger than that of the genomes of the pathogenic organisms *C. jeikeium* and *C. diphtheriae* (Table 1). The lower number of genes observed in the clinical isolates sequenced to date perhaps reflects that fewer metabolic functions are needed for corynebacteria to occupy a clinical ecological niche rather than a soil-based environment, consistent with the observation that gene decay and gene reduction have played a central role in the evolution of the *Mycobacterium leprae* (Cole *et al.*, 2001) and *C. diphtheriae* (Nishio *et al.*, 2004) chromosomes. These various corynebacterial complete genome sequences confirm the observation that, of the organisms forming the

Table 1. General features of *C. glutamicum* R and related bacteria

Abbreviations: *Cg* R, *C. glutamicum* R; *Cg* K, Kitasato University *C. glutamicum* ATCC 13032 isolate; *Cg* B, Bielefeld University *C. glutamicum* ATCC 13032 isolate; *Ce*, *C. efficiens*; *Cj*, *C. jeikeium*; *Ms*, *Mycobacterium smegmatis*; *Mt*, *Mycobacterium tuberculosis*; *Nf*, *Nocardia farcinica*. The *M. smegmatis* sequence was obtained from The Institute of Genome Research (TIGR).

Feature	*Cg* R	*Cg* K	*Cg* B	*Ce*	*Cj*	*Cd*	*Ms*	*Mt*	*Nf*
Accession no.	AP009044	NC003450	BX927147	BA000035	CR931997	BX248353		AL123456	AP006618
Strain	R	ATCC 13032	ATCC 13032	YS-314	K411	NCTC 13129	MC2	H37Rv	IFM 10152
Reference	This work	Ikeda & Nakagawa (2003)	Kalinowski *et al.* (2003)	Nishio *et al.* (2003)	Tauch *et al.* (2005)	Cerdeño-Tárraga *et al.* (2003)	TIGR	Cole *et al.* (1998, 2001)	Ishikawa *et al.* (2004)
Total genome size (bp)	3 314 179	3 309 401	3 282 708	3 147 090	2 462 499	2 488 635	6 988 209	4 411 532	6 021 225
Genome G+C content (mol%)	54.10	53.80	53.80	63.40	61.40	53.50	67.40	65.60	70.80
No. of ORFs	2990	2993	3002	2950	2104	2320	6829	3959	5674
Mean ORF length (bp)	957	933	916	979	1030	964	1005	1012	960
Coding density (%)	86.30	86.80	88.30	90.40	89.20	89.60	99.20	90.80	90.4
tRNA	57	60	60	56	50	54	47	45	53
16S RNA	6	6	6	5	3	5	2	4	3
23S RNA	6	6	6	5	3	5	2	4	3
5S RNA	6	6	6	5	3	5	2	0	3
Episome 1									
Accession no.	AP009045			AP005225	AF401314				AP006619
Size (bp)	PCgR1: 49 120			pCE2: 23 743	pKW4: 14 323				pNF1: 184 026
G+C content (mol%)	53.9			54.4	53.8				67.2
Episome 2									
Accession no.				AP005226					AP006620
Size (bp)				pCE3: 48 672					pNF2: 87 093
G+C content (mol%)				56.4					68.4

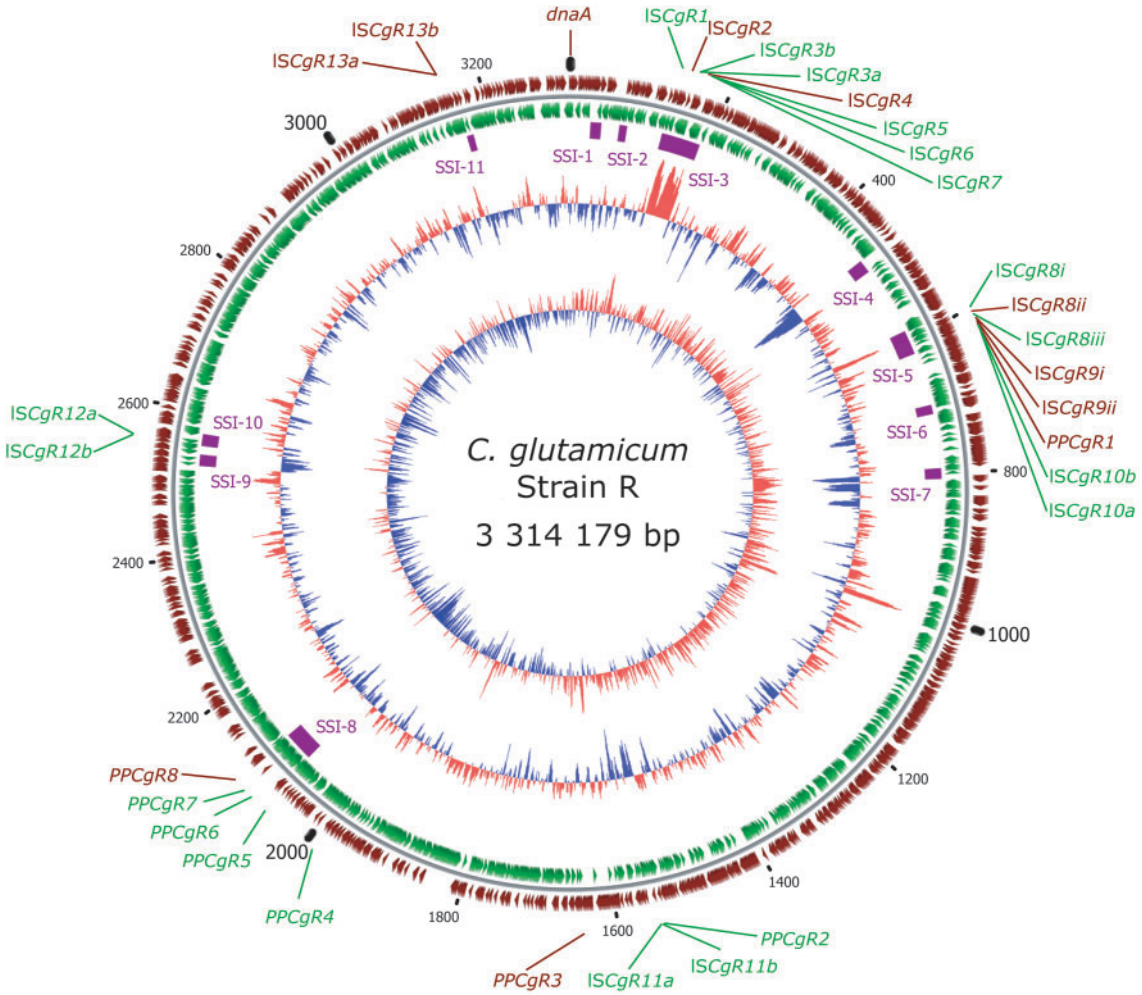

Fig. 1. Circular representation of the genome of *C. glutamicum* R. The first base of the initiation codon of the *dnaA* gene was set as the origin of the coordinates. Coordinates are given in kb. The two outermost circles represent the predicted CDS on the forward (maroon) strand on the outside and reverse (green) strand on the inside. The third circle from the outside shows the location of SSIs greater than 10 kb (SSI-1–SSI-11) along the genome. The localization of insertion sequences (IS*CgR1*–IS*CgR13*) and phage-derived proteins (*PPCgR1*–*PPCgR8*) is depicted. The colour of the labelled genes corresponds to the colour of the strand on which they are located. The fourth and fifth circles represent the G+C content and the GC skew (G–C)/(G+C), respectively, each plotted using a 3000 bp window with a 1000 bp window overlap. Red regions (pointing outwards) in these two circles are those of high G+C content or high GC skew, whereas blue regions (pointing inwards) are those of low G+C content or low GC skew. The map was created using original scripts and the CGView software (Stothard & Wishart, 2005).

Corynebacterineae phylogenetic cluster, corynebacteria have both the smallest genomes and the lowest coding densities (Table 1), although this difference is perhaps an artifact as gene finding in genomes of high G+C content tends to predict more or extended coding regions due to lower frequencies of stop codons.

The genome of strain R encodes 6 rRNA operons and 57 tRNA genes. The first base of the initiation codon of the *dnaA* gene was designated as the sequence coordinate origin. It is located near characteristic replication regions. The GC skew analysis (Grigoriev, 1998) presented in Fig. 1 supports the view that DNA replication in *C. glutamicum* is bidirectional, as observed in strain ATCC 13032 (Kalinowski *et al.*, 2003). It is noteworthy that the G+C content variation is low in this genome. Nevertheless, a comparison of the genome of *C. glutamicum* strain R with the published genome sequences of two isolates of *C. glutamicum* ATCC 13032 (Ikeda & Nakagawa, 2003; Kalinowski *et al.*, 2003) revealed the presence of hundreds of SSIs (Suzuki *et al.*, 2005b). In particular, strain R exhibits 11 SSIs larger than 10 kb (Fig. 1) that have a G+C content ranging from 45.7 to 60.7 mol% (Suzuki *et al.*, 2005a). Notably, none of these was observed to be flanked by obvious repeats. However, it is

noteworthy that strain R is devoid of the AT-rich 211 kb genomic island with clear boundaries (Zhang & Zhang, 2004) that is carried by strain ATCC 13032 and that contains genes typically associated with horizontal transfer, such as a restriction-modification system, transposases, recombination enzymes and phage-derived sequences (Kalinowski *et al.*, 2003). Likewise, the AT-rich 25 kb region harboured by strain ATCC 13032 containing the genes *cg0414–cg0440* is absent from the strain R genome. It is noteworthy that products of these genes are involved in cell wall formation, including cell surface polysaccharide (*wzz*, cg0414), lipopolysaccharide (various glycosyl transferases) or murein formation (*murA*, *cg0422*; *murB*, *cg0423*). These observations suggest that measurable differences could exist between the cell walls of the two strains that could perhaps form the basis of immunotyping procedures. Furthermore, the 14 kb G+C-rich region identified in the genome of strain ATCC 13032 (Kalinowski *et al.*, 2003) to contain *C. diphtheriae* sequences (95 % identity at the nucleotide level) is absent from the genome of strain R, which does not exhibit any sequence similar to the gene cluster *cg3276–cg3290*. This observation promotes the view that these genes have, on an evolutionary timescale, recently been acquired by *C. glutamicum* ATCC 13032 in a horizontal transfer event originating from *C. diphtheriae*. On the other hand, it is noteworthy that the strain R SSI-4, characterized by a low G+C content (45.7 mol%) (Suzuki *et al.*, 2005a), and thus probably resulting from a horizontal transfer, does not contain any obvious phage- or mobile-element-related sequence.

Strain R encodes 22 sequences homologous to mobile elements, including six incomplete insertion sequence signatures, and five insertion sequences that contain two ORFs which putatively form a full transposase protein via a frameshift (IS*CgR3a-b*, IS*CgR10a-b*, IS*CgR11a-b*, IS*CgR12a-b*, IS*CgR13a-b*) (Table 2). The presumably functional mobile elements found in *C. glutamicum* R originate from two different families (IS*3*, IS*CgR11*; IS*6*, IS*CgR1*, IS*CgR2*, IS*CgR9*), as defined by Mahillon & Chandler (1998). IS*CgR9* is an isoform of IS*1628* (Mahillon & Chandler, 1998). IS*CgR3* and IS*CgR13* constitute novel elements that share a relatively low level of identity at the amino acid level with a putative *Rhodococcus erythropolis* insertion sequence (Stecker *et al.*, 2003). Similar to what is observed in strain ATCC 13032, SSIs in strain R are rich in mobile elements, particularly SSI-3 (G+C content, 60.7 mol%) and SSI-5 (55.2 mol%) (Table 2). On the other hand, fewer genes of phage origin are present in the genome of *C. glutamicum* R as compared to *C. glutamicum* ATCC 13032 since only phage remnants are observed (Table 2). Nevertheless, SSI-8, 42.7 kb in size and with a G+C content of 52.1 mol% (Suzuki *et al.*, 2005a), could have originated from a previously unknown phage as it encodes numerous hypothetical proteins and several ORFs, the products of which share some homology to known phage proteins. This is in sharp contrast to what is observed in *C. glutamicum* ATCC 13032 isolates that harbour three to four different putative prophages (Kalinowski, 2005).

Chromosome structure

As previously observed (Kalinowski *et al.*, 2003; Nakamura *et al.*, 2003), corynebacterial genomes, including that of strain R, show a very high degree of synteny with a striking lack of detectable inversions (not shown). The *C. jeikeium* genome, however, exhibits 10 apparent breakpoints of synteny with *C. glutamicum* ATCC 13032 (Tauch *et al.*, 2005). Likewise, the genomes of the relatively closely related mycobacteria show that extensive rearrangements have occurred in these species throughout the course of evolution. This phenomenon has been ascribed to the lack of a complete RecBCD recombination repair system in corynebacteria (Nakamura *et al.*, 2003), despite the presence of *recA* and *recB* genes. Nevertheless, as demonstrated by the rearrangements that have occurred in the genome of *C. jeikeium*, other recombination mechanisms may be at play that allow moderate reorganizations in the chromosome architecture.

The *C. glutamicum* R and ATCC 13032 genomes contain similar numbers of putative operons in concordance with the numbers of rho-independent transcription terminators containing a stem–loop and poly-U tail. Using the TIGR Comprehensive Microbial Resource (Ermolaeva *et al.*, 2001), we calculated that *C. glutamicum* ATCC 13032 contains 455 operons and 485 rho-independent terminators. Notably, *cgR1278*, encoding the transcription termination factor rho, is probably an essential gene in *C. glutamicum* R (Suzuki *et al.*, 2006).

Corynebacterial core genes

The corynebacterial backbone, defined by the set of orthologous genes present in all corynebacteria sequenced to date [*C. glutamicum* ATCC 13032 (Ikeda & Nakagawa, 2003; Kalinowski *et al.*, 2003), *C. glutamicum* R (this work), *C. efficiens* YS-314 (Nishio *et al.*, 2003), *C. jeikeium* K411 (Tauch *et al.*, 2005) and *C. diphtheriae* NCTC 13129 (Cerdeño-Tárraga *et al.*, 2003)], comprises 835 genes assigned to COG categories (Table 3), with the total number of shared genes corresponding to approximately a third of the genes present in the saprophytic corynebacteria, in agreement with the value reported by other authors (Tauch *et al.*, 2005) (1089 genes). However, the subset of genes that are essential is significantly lower than the core corynebacterial genome, as demonstrated by the successful transposon mutagenesis of at least 75 % of the ORFs of strain R using transposons Tn*5* and Tn*31831* (Suzuki *et al.*, 2006; Vertès *et al.*, 2005). It has been demonstrated that the genes that are indispensable for growth of *Bacillus subtilis* cells in rich medium under standard laboratory conditions belong to the following known categories: DNA and RNA metabolism, protein synthesis, cell envelope, cell shape and division, glycolysis, respiratory pathways, nucleotides and cofactors (Kobayashi *et al.*, 2003). Based on the genomic sequences of *C. glutamicum* R (this work) and of other corynebacteria (Cerdeño-Tárraga *et al.*, 2003; Ikeda & Nakagawa, 2003; Kalinowski *et al.*, 2003; Nishio *et al.*, 2003;

Table 2. Insertion sequences and phage-derived sequences in the *C. glutamicum* R genome

The sizes of ORFs are given in bp. The direction of transcription is given relative to the *dnaA* gene (+, clockwise, −, anticlockwise). Identity levels are calculated based on putative amino acid sequences. Homologous genes: IS*1673* is from the *C. glutamicum* plasmid pCG4 (Tauch *et al.*, 2003), IS*1870* is from the *C. glutamicum* pTET3 plasmid (Tauch *et al.*, 2002), PBD2.162 and PBD2.163 are from the *Rhodococcus erythropolis* linear plasmid pBD2 (Stecker *et al.*, 2003), IS*30* is from *E. coli* K-12 (Blattner *et al.*, 1997), IS*6110* is from *Mycobacterium avium* (Li *et al.*, 2005), IS*1628* is from the *C. glutamicum* plasmid pCG4 (Tauch *et al.*, 2003), IS*1206* [isoform IS*Cg14* (Kalinowski *et al.*, 2003)] is from the *C. glutamicum* chromosome (Bonamy *et al.*, 1994). Putative *tnp*, Putative transposase gene; IR, inverted repeats; NF, not found. SSI-3 has a G+C content of 60.7 mol%; SSI-5, 55.2 mol%; SSI-8, 52.1 mol%; and SSI-10, 53.0 mol% (Suzuki *et al.*, 2005a).

Name	ORF no.	Size	G+C (mol)%	Direction	Function	IRs	Promoter	Identity (%)	SSI
IS*CgR1*	*cgR0118*	702	55.2	−	Probably functional IS	Probable	Probable	IS*1673*: 88	3
IS*CgR2*	*cgR0128*	480	53.2	+	Probably functional IS	Probable	Probable	IS*1673*: 74	3
	cgR0136	243	47.9	+	IS signature	NF	Probable	IS*1870*: 77	3
IS*CgR3b*	*cgR0141*	642	56.5	−	Frameshift fusion with CgR0142?	NF	Probable	PBD2.163: 52	3
IS*CgR3a*	*cgR0142*	369	53.5	−	Frameshift fusion with CgR0141?	NF	NF	PBD2.162: 44	3
IS*CgR4*	*cgR0149*	1092	59.1	+	Probably functional IS	Probable	Probable	Putative *tnp*	3
	cgR0150	231	55.6	−	IS signature	NF	NF	Putative *tnp*	3
	cgR0151	321	53.5	−	IS signature	NF	NF	Putative *tnp*	3
IS*CgR5*	*cgR0152*	648	60.2	−	Probably functional IS	NF	NF	Putative *tnp*	3
IS*CgR6*	*cgR0153*	1152	58.6	−	Probably functional IS	NF	NF	IS*30*: 47	3
	cgR0154	198	54.2	−	IS signature	NF	NF	IS*6110*: 44	3
IS*CgR7*	*cgR0171*	717	55.0	−	Probably functional IS	Probable	Probable	Putative *tnp*	3
	cgR0534	153	45.1	+	IS signature	Probable	Probable	Putative *tnp*	5
IS*CgR8*	*cgR0537*	453	52.8	−	Probably functional IS	Probable	Probable	IS1673: 88	5
IS*CgR8*	*cgR0544*	453	52.8	+	Probably functional IS	Probable	Probable	IS*1673*: 76	5
IS*CgR8*	*cgR0548*	453	52.8	−	Probably functional IS	Probable	Probable	IS*1673*: 76	5
IS*CgR9*	*cgR0552*	711	55.4	+	Probably functional IS	Probable	Probable	IS*1628*: 100	5
IS*CgR9*	*cgR0557*	711	55.9	+	Probably functional IS	Probable	Probable	IS*1628*: 100	5
IS*CgR10b*	*cgR0562*	453	53.3	−	Frameshift fusion with CgR0563?	Probable	Probable	IS*1628*: 92	5
IS*CgR10a*	*cgR0563*	138	49.9	−	Frameshift fusion with CgR0562?	Probable	Probable	IS*1628*: 90	5
	cgR0565	162	48.7	+	IS signature	NF	NF	IS*1628*: 92	5
IS*CgR11b*	*cgR1410*	897	54.5	−	Frameshift fusion with CgR1411?	Probable	NF	IS*1206*: 83	
IS*CgR11a*	*cgR1411*	309	54.1	−	Frameshift fusion with CgR1410?	Probable	Probable	IS*1206*: 75	
IS*CgR12b*	*cgR2323*	930	47.8	−	Frameshift fusion with CgR2324?	Probable	NF	Putative *tnp*	10
IS*CgR12a*	*cgR2324*	177	43.6	−	Frameshift fusion with CgR2323?	Probable	Probable	Putative *tnp*	10
IS*CgR13a*	*cgR2839*	435	58.0	+	Frameshift fusion with CgR2840?	NF	NF	PBD2.162: 50	
IS*CgR13b*	*cgR2840*	414	58.0	+	Frameshift fusion with CgR2839?	NF	NF	PBD2.16: 53	
PPCgR1	*cgR0559*	609	64.8	+	Putative invertase				5
PPCgR2	*cgR1409*	759	45.7	−	Probable phage type integrase				
PPCgR3	*cgR1480*	915	53.8	+	Putative integrase recombinase				
PPCgR4	*cgR1791*	831	51.6	−	Similar to phage shock protein A				
PPCgR5	*cgR1856*	948	50.3	−	Putative site specific recombinase				
PPCgR6	*cgR1882*	4791	53.6	−	Putative phage tail protein				8
PPCgR7	*cgR1895*	1650	49.6	−	Putative phage associated protein				8
PPCgR8	*cgR1924*	1182	53.9	+	Putative integrase recombinase				

Tauch *et al.*, 2005), these fundamental cellular processes are also essentially conserved in corynebacterial genomes.

Sigma factors

Ecological niches in soil are characterized by the availability of numerous growth substrates. The pan-genome of saprophytic organisms contains numerous enzymic activities necessary for the breakdown of a large variety of complex molecules. While saprophytic fungi play a predominant role in the recycling of organic matter, numerous enzymes are secreted by bacterial saprophytes to hydrolyse polysaccharides, proteins and lipids. Likewise, saprophytic organisms need to deploy a variety of adaptive responses to cope with various environmental stresses that can range, for a soil bacterium, from nutrient limitation, external osmolality fluctuations and oxygen deprivation to temperature shock. These responses are typically regulated

Table 3. Classification of *C. glutamicum* R CDS by COG

ORFs were analysed by BLASTP for each of the genomes of the corynebacteria sequenced to date against the NCBI COG database. Results were parsed for hits with an e-value cutoff of $1 \times e^{-25}$. Proteins containing multiple functional domains were forced into only one COG category by elimination of duplicates. In accordance with their COG identities, the resultant hits were grouped into the following categories: *Cg* R, total hits of *C. glutamicum* R; Core, hits with orthologues in all the corynebacteria sequenced to date; Sap, hits with orthologues only in the saprophytic corynebacteria sequenced to date; *Cglut*, hits with orthologues only in both *C. glutamicum* R and the Kitasato University *C. glutamicum* ATCC 13032 isolate; R spec, non-core, non-*C. glutamicum* and non-saprophytic hits of strain R; *Cg* K, *C. glutamicum* ATCC 13032 Kitasato (Ikeda & Nakagawa, 2003); K spec, non-core, non-*C. glutamicum* and non-saprophytic hits of ATCC 13032 Kitasato strain; *Ce*, total hits of *C. efficiens* YS-314; *Ce* spec, non-core and non-saprophytic hits of *C. efficiens*; Path, non-core genes present in both of the pathogenic corynebacteria sequenced to date; *Cd*, total hits of *C. diphtheriae* NTCT 13129; *Cd* spec, non-core and non-pathogen-specific hits of *C. diphtheriae*; *Cj*, total hits of *C. jeikeium* K411; *Cj* spec, non-core and non-pathogen-specific hits of *C. jeikeium*. Genes that could not be assigned to a COG category are not included in the Table.

NCBI COG		*Cg* R	Core	Sap	*Cglut*	R spec	*Cg* K	K spec	*Ce*	*Ce* spec	*Cd*	Path	*Cd* spec	*Cj*	*Cj* spec
Information storage and processing															
J	Translation, ribosomal structure and biogenesis	139	117	13	6	3	140	4	135	5	131	3	11	129	9
A	RNA processing and modification	1	1	0	0	0	1	0	1	0	1	0	0	1	0
K	Transcription	145	42	58	35	10	143	8	107	7	69	1	26	55	12
L	Replication, recombination and repair	96	53	24	9	10	113	27	102	25	79	9	17	77	15
Cellular processes and signalling															
D	Cell cycle control, cell division, chromosome partitioning	17	13	2	2	0	18	1	16	1	17	1	3	22	8
V	Defence mechanisms	42	12	13	15	2	43	3	39	14	35	4	19	43	14
T	Signal transduction mechanisms	37	18	13	5	1	40	4	34	3	22	1	3	25	6
M	Cell wall/membrane/envelope biogenesis	84	49	16	8	11	89	16	81	16	71	3	20	65	13
U	Intracellular trafficking, secretion and vesicular transport	16	11	3	1	1	16	1	15	1	15	1	3	12	0
O	Post-translational modification, protein turnover, chaperones	67	35	21	9	2	71	6	61	5	56	5	16	45	5
Metabolism															
C	Energy production and conversion	117	42	44	20	11	113	7	99	13	78	3	33	68	23
G	Carbohydrate transport and metabolism	136	43	48	27	18	124	6	107	16	83	4	36	64	17
E	Amino acid transport and metabolism	161	64	67	19	11	160	10	145	14	123	9	50	112	39
F	Nucleotide transport and metabolism	60	35	20	4	1	62	3	58	3	55	4	16	51	12
H	Coenzyme transport and metabolism	95	45	40	6	3	95	4	95	10	92	8	39	66	13
I	Lipid transport and metabolism	59	22	18	7	12	58	11	63	23	42	5	15	56	29
P	Inorganic ion transport and metabolism	159	23	75	43	18	147	6	109	11	70	16	31	81	42
Q	Secondary metabolites biosynthesis, transport and catabolism	28	3	16	5	4	26	2	25	6	10	2	5	14	9
Poorly characterized															
R	General function prediction only	223	80	100	38	5	238	20	200	20	133	6	47	123	37
S	Function unknown	165	48	70	33	14	176	25	139	21	104	5	51	76	23
Other		251	79	101	48	23	253	25	214	34	150	10	61	141	52
Total		**2098**	**835**	**762**	**340**	**160**	**2126**	**189**	**1845**	**248**	**1436**	**100**	**502**	**1326**	**378**

via specialized sigma factors of the extracytoplasmic function (ECF) subfamily (Missiakas & Raina, 1998). The saprophytic organisms *Mycobacterium smegmatis* (Waagmeester *et al.*, 2005) and *Streptomyces avermitilis* (Ikeda *et al.*, 2003) exhibit 26 and 60 putative sigma factors, respectively. In the case of *S. avermitilis*, which belongs to a different suborder of the *Actinomycetales* from *M. smegmatis* and *C. glutamicum*, 47 of these belong to the ECF subfamily. On the

Table 4. Putative sigma factors of corynebacteria

Putative sigma factors were identified by BLASTP searches against the GenBank and VIMSS databases. Gene numbers for *Cg* K, *Ce*, *Cd* and *Cj* refer to the VIMSS database (Alm *et al.*, 2005). *Cg* R, *C. glutamicum* strain R; *Cg* K, *C. glutamicum* ATCC 13032 Kitasato (Ikeda & Nakagawa, 2003); *Ce*, *C. efficiens* YS-314 (Nishio *et al.*, 2003); *Cd*, *C. diphtheriae* NCTC 13129 (Cerdeño-Tárraga *et al.*, 2003); *Cj*, *C. jeikeium* K411 (Tauch *et al.*, 2005). Percentage identity scores (Id) were measured against the corresponding *C. glutamicum* R genes. The designation of the sigma factors is consistent with that used for *M. tuberculosis* (Waagmeester *et al.*, 2005). The genes listed as *pvdS1* and *pvdS2* are conserved hypothetical proteins that show a limited homology to sigma factors. In *C. glutamicum* R, the gene encoding σ^D is interrupted by a spacer region. Sizes of genes refer to the number of base pairs.

Gene	*Cg* R	Size	*Cg* K	Size	Id	*Ce*	Size	Id	*Cd*	Size	Id	*Cj*	Size	Id
sigC	*cgR0331*	193	*374306*	193	99	*299112*	202	90	*519963*	188	76	*844944*	205	68
sigD (front)	*cgR0718*	61	*374631*	188	90	*299494*	190	65	*520259*	188	59	*844664*	189	38
sigD (end)	*cgR0719*	114	*374631*	188	100	*299494*	190	88	*520259*	188	74	*844664*	189	56
sigH	*cgR0876*	207	*374789*	206	100	*299671*	250	88	*520390*	207	77	*844558*	209	73
pvdS1	*cgR1018*	321	*374936*	320	98	*299868*	351	76	*520561*	301	85	*844437*	326	78
sigE	*cgR1204*	214	*375131*	213	100	*300066*	216	86	*520673*	239	83	*844311*	196	74
sigA	*cgR1740*	495	*375892*	498	99	*300693*	506	82	*521096*	502	76	*844005*	544	68
sigB	*cgR1749*	329	*375900*	331	99	*300700*	333	93	*521103*	329	90	*844012*	343	85
pvdS2	*cgR2616*	307	*376676*	306	98	*301447*	315	94	*521713*	298	89	*843139*	296	84
sigM	*cgR2979*	202	*377039*	201	100	*301821*	251	66	*522059*	200	40	*845047*	205	27
sigK									*519777*	194		*842888*	198	
sigL									*520559*	179				
sigW												*843374*	191	
rshA									*520391*	83		*844557*	87	
Anti-sigma antagonist						*300093*	109							

other hand, *C. glutamicum* R harbours fewer of these components of the RNA polymerase complex. As shown in Table 4, in addition to σ^A, the sigma factor that directs the transcription of most genes in growing cells, and σ^B, which plays a crucial role in the maintenance of the stationary phase in *M. smegmatis* (Mukherjee & Chatterji, 2005), corynebacteria possess five alternative sigma factors to regulate genetic expression in response to extracellular changes. *C. glutamicum* R and *C. glutamicum* ATCC 13032 exhibit the same number of these ECF sigma factors to regulate extracytoplasmic functions, with the exception of σ^D which is disrupted by a spacer region in *C. glutamicum* R. In *M. tuberculosis* H37Rv, σ^D has been shown to control the expression of ribosome-associated gene products in the stationary phase and to be required for full virulence (Calamita *et al.*, 2005). Similarly, σ^H has been observed to regulate major components of oxidative and heat stress responses (Raman *et al.*, 2001), and σ^L, an orthologue of which is present in *C. diphtheriae* NCTC 13129 but not in the other corynebacteria sequenced to date, has been observed to regulate polyketide synthases and secreted or membrane proteins (Hahn *et al.*, 2005). In *C. jeikeium* K411 and *C. diphtheriae* NCTC 13129, *sigH* is located upstream of *rshA*, a putative anti-σ^H factor, with which it forms a putative operon, as observed in *Nocardia farcinica* IFM 10152 (Ishikawa *et al.*, 2004). All corynebacteria sequenced to date contain a *sigC* and a *sigE* gene. Interestingly, σ^C has been shown to be required for lethality of *M. tuberculosis* in mice (Sun *et al.*, 2004), and σ^E is induced upon treatment of *M. tuberculosis* cells with hydrogen peroxide or upon macrophage infection (Jensen-Cain & Quinn, 2001). In contrast to what is observed in mycobacteria, where the saprophytic *M. smegmatis* harbours twice as many sigma factors (26 sigma factors) as its pathogenic relative *M. tuberculosis* (13) (Waagmeester *et al.*, 2005), no sigma factor specific to saprophytic corynebacteria could be identified by homology searches, despite the presence of pathogenic-specific sigma factors in *C. diphtheriae* NCTC 13129 (σ^K, σ^L) and *C. jeikeium* K411 (σ^K, σ^L, σ^W). This latter observation particularly reinforces the view that σ^K is involved in bacterial virulence, as a σ^K orthologue is present in *M. tuberculosis* H37Rv but absent from *M. smegmatis* MC2 (Waagmeester *et al.*, 2005). Consistent with these predicted functions, we could isolate *C. glutamicum* R cells mutated by transposon insertion in any of the genes *sigH*, *sigE*, *sigB*, *sigM*, or *pvdS1* and *pvdS2* (these latter genes encode two conserved hypothetical proteins that show limited homology with sigma factors, though they are unlikely to encode sigma factors), but not in *sigC* or *sigA*. Notably, *sigH* and *sigE* are also dispensable in *C. glutamicum* ATCC 13032 (Engels *et al.*, 2004).

Secreted proteins

The relative paucity of sigma factor-mediated adaptive mechanisms in corynebacteria reflects the exacting nutrient requirements typically exhibited by these organisms and their relative limitation in extracellular enzymes to digest

complex molecules present in their environments. Corynebacteria excrete a limited number of proteins, as demonstrated by the identification of only approximately 40 protein spots in the 4.0 to 5.0 pI range during proteome analysis experiments of supernatants of late-exponential growth phase cultures (Hermann *et al.*, 2001), and by the identification of 49 cell-surface protein spots and 89 extracellular protein spots in the 3 to 7 pI range from proteome fractions of *C. efficiens* YS-314 (Hansmeier *et al.*, 2006). In particular, it has long been known that *C. glutamicum* cell extracts and supernatants do not have broad-spectrum proteolytic activity, as they show only limited extracellular protease activity on skim milk plates. Likewise, these bacteria exhibit limited extracellular lipolytic and nuclease activity, and no cellulase or amylase activity (Yukawa *et al*, 2007). The large number of transporters found in these bacteria to ensure the uptake of amino acids and peptides (Winnen *et al.*, 2005) can perhaps be ascribed to the limited capability of these organisms to breakdown complex molecules. Nevertheless, mining of the *C. glutamicum* R genome reveals the presence of gene *cgR1176* which encodes a putative secreted protease which is 29.6 % identical to the *B. subtilis epr* gene, the product of which is an extracellular serine protease. Based on *in silico* homology searches, orthologues of *cgR1176* appear to be common in members of the *Actinobacteria* group. Similarly, *cgR1002*, a probable pullulanase gene, which is part of a putative three-gene operon in *C. glutamicum* R, can also be observed in the genome of *C. glutamicum* ATCC 13032 (*VIMSS374909*) and in *C. efficiens* YS-314 as part of a putative four-gene operon, but not in *C. diphtheriae* NCTC 13129 nor *C. jeikeium* K411. Interestingly, this operon also encodes two hypothetical membrane proteins which are conserved in both saprophytic and pathogenic corynebacteria (*cgR1000*, *cgR1001*). In *C. efficiens*, this operon is predicted (VIMSS database) to be controlled by CE2422, a transcription regulator of the GntR family of proteins. Homologues of CE2422 are present in both *C. glutamicum* R (*cgR2434*) and ATCC 13032 (*VIMSS376496*), and in various *Streptomyces* species, but not in the pathogenic corynebacteria sequenced to date nor in mycobacteria.

Sugar metabolism

Corynebacteria are able to utilize only a limited array of different sugars. For example, wild-type *C. glutamicum* R cells are able to utilize fructose, glucose, glucuronic acid, glucosamine, maltose, mannose, α-methylglucoside, ribose, sucrose, trehalose, arbutin and salicin as sole carbon sources, but not arabinose, galactose, lactose, cellobiose, mannitol, rhamnose, xylose or xylitol (this work). *C. glutamicum* transports several sugars, including sucrose, fructose and glucose, by the phosphotransferase system (PTS) (Moon *et al.*, 2005; Saier, 2002). Sugar transport plays an important role in the observed substrate range limitation of *C. glutamicum*, as exemplified by the isolation of a spontaneous PTS mutant of *C. glutamicum* R that is able to extend the catabolic capabilities of this organism to the degradation of cellobiose (Kotrba *et al.*, 2003). However, this limitation in carbon substrate spectrum is also due to the lack of specific catabolic genes, such as the lack of xylose isomerase which prevents the catabolism of the pentose xylose, despite the presence of ATP-binding cassette proteins putatively involved in the transport of this sugar (specifically that encoded by *cgR1331*, which is part of a five-gene operon containing a *lacI*-type transcriptional regulator, and which shows, respectively, 48 and 44 % homology to the *xylF* genes of *Geobacillus kaustophilus* HTA426 and *E. coli* K-12 which encode the periplasmic xylose-binding subunit of a high-affinity xylose ABC-transporter; moreover, synteny comparisons suggest the presence also of the *xylG* (*cgR1329*) and *xylH* genes (*cgR1330*) encoding, respectively, the ATP-binding and membrane components of a xylose ABC transporter). In addition, several sugar/proton symports could also be involved in the uptake of xylose by wild-type *C. glutamicum* cells, since *cgR0261*, *cgR2943*, *cgR2864*, *cgR2290* and *cgR2267* all show homology levels greater than 40 % with *E. coli xylE*, a gene that encodes a D-xylose/proton symporter from the major facilitator superfamily of transporters.

Notably, all corynebacteria sequenced to date, including *C. glutamicum* R (*cgR1200*) and ATCC 13032 (*VIMSS376610*), as well as *C. diphtheriae* NCTC 13129 (*VIMSS520668*) and *C. jeikeium* K411 (*VIMSS844316*), exhibit a putative β-fructofuranosidase gene encoding a protein that is 49 % similar (34 % identical) to the *fruA* gene product of *Bacillus megaterium* ATCC 14581 (Chiou *et al.*, 2002). The capacity of corynebacteria to synthesize fructans, and particularly pathogenic corynebacteria, would thus be interesting to verify since fructans can trigger inflammatory reactions (Shilo & Wolman, 1958) and thus contribute to disease progression. However, only *C. glutamicum* R (*cgR2548*) and *C. glutamicum* ATCC 13032 (*VIMSS376610*) exhibit a putative *sacA* gene encoding a second β-fructofuranosidase. The *C. glutamicum sacA* gene product shares 48 % homology with the *B. subtilis* 168 SacA (Glaser *et al.*, 1993) and is only 26 % identical to the product of the putative *C. glutamicum fruA* gene. In both *C. glutamicum* R and ATCC 13032, *sacA* is part of a putative four-gene operon that includes a phosphotransferase system component (PTS enzyme IIC) (respectively, *cgR2547* and *VIMSS376609*). Both *fruA* and *sacA* are dispensable, as demonstrated by the disruption and replacement of these genes in *C. glutamicum* R (this work).

Furthermore, consistent with the notion that corynebacteria use glycogen as their major polyglucan reserve, glycogen metabolism genes are highly conserved among these bacteria. In particular, corynebacterial genomes exhibit sequences homologous to a glycosyl transferase (*cgR1201*) linked in a putative operon to *glgC*, the gene encoding ADP-glucose pyrophosphorylase (strain R, *cgR1202*; strain ATCC 13032, *VIMSS375129*). Also observed are a putative two-gene operon, *glgB*, encoding glycosyl transferase (*cgR1302-cgR1303*), and *glgX*, encoding a glycogen debranching

Table 5. Putative two-component systems of corynebacteria

Putative two-component system genes were identified by BLASTP searches against the GenBank and VIMSS databases. Gene numbers for *Cg* K, *Ce*, *Cd*, *Cj*, *Mtb* and *Nf* refer to the VIMSS database, except CE3P006 and CE3P005 which refer to the GenBank database and gene numbers for *Ms* which refer to the TIGR-CMR database. *Cg* R, *C. glutamicum* strain R; *Cg* K, *C. glutamicum* ATCC 13032 Kitasato (Ikeda & Nakagawa, 2003); *Ce*, *C. efficiens* YS-314 (Nishio *et al.*, 2003); *Cd*, *C. diphtheriae* NCTC 13129 (Cerdeño-Tárraga *et al.*, 2003); *Cj*, *C. jeikeium* K411 (Tauch *et al.*, 2005); *Ms*, *M. smegmatis* MC2 (TIGR-CMR data); *Mtb*, *M. tuberculosis* H37Rv (Cole *et al.*, 1998); *Nf*, *N. farcinica* IFM 10152 (Ishikawa *et al.*, 2004). Percentage identity scores (Id) were measured against the corresponding *C. glutamicum* R genes, or, when no orthologue is present in *C. glutamicum* R, to a reference gene indicated by the annotation REF. Only genes with identity levels greater than or equal to 35 % are indicated, except for genes for which a possible significant link could be identified by identity, homology or synteny, for example the *citA/citB* genes of *N. farcinica*. The G+C content (mol%) is given only for the *C. glutamicum* R genes. Sizes of gene products refer to the number of amino acids. Fnct, Predicted physiological function; S, sensor kinase; R, transcriptional regulator. Known orthologous genes with the highest identity scores are given in bold type. By homology with previously described two-component systems, these genes are as follows: *yycF/yycG* (CgR0122 is 63 % homologous to the *yycF* gene product of *Bacillus subtilis*; Fukuchi *et al.*, 2000); *ykoG/ykoH* (CgR0360 is 50 % homologous to the *ykoH* gene product of *B. subtilis*; Fabret *et al.*, 1999); *senX3/regX3* (CgR0476 is 70 % identical to the *regX3* gene product of *M. tuberculosis* H37Rv; Fontan *et al.*, 2004); *cutS/cutR* (CgR0540 and CgR0541 are, respectively, 50 and 70 % homologous to the *cutS* and *cutR* gene products of *Streptomyces lividans*; Hutchings *et al.*, 2004); *mtrA/mtrB* (CgR0863 is 68 % identical to the *mtrA* gene product of *M. tuberculosis* H37Rv; Fontan *et al.*, 2004; Zahrt & Deretic, 2000); *mprB/mprA* (CgR0988 is 64 % identical to the *mprA* gene product of *M. tuberculosis* H37Rv; Fontan *et al.*, 2004; Zahrt & Deretic, 2000); *yocF/yocG* (CgR1050 is 62 % homologous to the *B. subtilis yocG* gene product; Fabret *et al.*, 1999); *yvqE/yvqC* (CgR1838 and CgR1839 are, respectively, 62 and 49 % homologous to the *B. subtilis yvqC* and *yvqE* gene products; similarly, CgR2844 and CgR2845 are 63 and 51 % homologous to the *B. subtilis yvqC* and *yvqE* gene products); *fixL/fixJ* (CgR2292 and CgR2299 are, respectively, 50 and 61 % homologous to the *fixL* and *fixJ* gene products of *Azorhizobium caulinodans* (Fischer, 1994); *phoR/phoP* (CgR2511 is 67 % identical to the *phoP* gene product of *M. tuberculosis* H37Rv; Fontan *et al.*, 2004); *baeS/baeR* (CgR2567 and CgR2566 are 60 and 53 %, respectively, homologous to the *E. coli baeS* and *baeR* gene products; Nishino *et al.*, 2005). The orphan putative regulator protein CgR0730 is 51 % homologous to the NarL protein of *E. coli* (Goh *et al.*, 2005). The products of the *C. diphtheriae* genes *VIMSS519770* and *VIMSS519769* are 48 and 57 % homologous to the *yxjM* and *yxjL* putative gene products of *B. subtilis* 168, respectively. The *C. glutamicum* ATCC 13032 gene *VIMSS376724* is 52 % homologous to the *yfiK* putative gene product of *B. subtilis* 168. The *M. smegmatis* MC2 genes MSMEG 0149 and MSMEG 0150 are adjacent but transcribed in opposite orientations. PAS, PAS sensing domain-containing protein, as defined by Taylor & Zhulin (1999).

Fnct	Gene	*Cg* R	G+C	Size	*Cg* K	Size	Id	*Ce*	Size	Id	*Cd*	Size	Id	*Cj*	Size	Id	*Ms*	Id	*Mtb*	Id	*Nf*	Id
S	***citA***	*cgR0088*	54	564	*374123*	551	94	*301794*	556	61											*770361*	28
R	***citB***	*cgR0089*	55	218	*374124*	218	93	*301793*	218	68											*770362*	31
R	***oxyS***													*844678*	304	31	0150	REF	31888	61		
S	***kdpD***													*844195*	898	REF	5352	44	32801	39	*767267*	41
R	***kdpE***													*844196*	230	REF	5376	42	32800	52	*767268*	52
S	***yxjM***							CE3P006	425	31	*519770*	409	REF								*776093*	37
R	***yxjL***							CE3P005	219	46	*519769*	219	REF								*769975*	40
R	***yycF***	*cgR0122*	67	240	*376919*	240	91	*299167*	240	68	*519745*	240	89	*844363*	240	90						
S	***yycG***	*cgR0123*	65	399	*376918*	399	83	*299168*	398	57	*519744*	375	78	*844362*	375	79						
S	***PAS***	*cgR0262*	48	145	*374236*	145	99	*299046*	166	71							0149	43				
R	***ykoG***	*cgR0359*	54	222	*374324*	222	95	*299122*	224	76							0235	41	32373	43		
S	***ykoH***	*cgR0360*	54	489	*374325*	489	94	*299123*	491	69												
S	***senX3***	*cgR0475*	57	413	*374447*	386	96	*299313*	413	82	*520071*	384	65	*844857*	432	54	0928	51	32261	49	*770922*	55
R	***regX3***	*cgR0476*	56	232	*374448*	232	93	*299314*	232	87	*520072*	226	76	*844856*	230	70	0929	72	32262	70	*770921*	78
S	***cutS***	*cgR0540*	67	344																		
R	***cutR***	*cgR0541*	66	224													5468	36			*766375*	51
R	***cstA***							*300245*	221	REF	*520143*	233	45	*844592*	227	47					*768832*	54
S	***cstS***							*300244*	395	REF	*520142*	408	36	*844593*	405	33					*768834*	36

Table 5. cont.

Fnct	Gene	*Cg* R	G+C	Size	*Cg* K	Size	Id	*Ce*	Size	Id	*Cd*	Size	Id	*Cj*	Size	Id	*Ms*	Id	*Mtb*	Id	*Nf*	Id
S	***narQ***				*374641*	352	REF	*299058*	401	76	*520280*	403	45								*766586*	44
R	***narL***	*cgR0730*	59	230	*374642*	230	79	*299512*	230	70	*520281*	231	60								*766587*	55
R	***mtrA***	*cgR0863*	57	226	*374777*	226	93	*299658*	226	87	*520375*	232	76	*844571*	225	67	1877	70	35020	68	*770338*	69
S	***mtrB***	*cgR0864*	57	499	*374778*	499	90	*299659*	489	73	*520376*	480	64	*844570*	578	51	1878	50	35019	48	*770337*	50
R	***mprA***	*cgR0988*	51	255	*374895*	232	90	*299837*	233	72	*520533*	230	72	*844465*	231	66	5468	65	32753	64	*770695*	65
S	***mprB***	*cgR0989*	50	479	*374896*	455	94	*299838*	488	56	*520534*	496	44	*844464*	517	39	5467	41	32754	40	*770694*	42
S	***yvfT***	*cgR1049*	56	419	*374967*	419	90	*299910*	419	51				*843797*	375	28						
R	***yvfU***	*cgR1050*	57	204	*374968*	203	92	*299911*	207	69				*843796*	200	37						
R	***yocG***							*300169*	220	REF											*769590*	46
S	***yocF***							*300170*	752	REF											*769589*	32
R	***yvqC***	*cgR1838*	57	210	*375990*	210	91	*301671*	225	53	*521954*	212	50	*844079*	216	55					*769968*	35
S	***yvqE***	*cgR1839*	54	383	*375991*	377	81	*301672*	429	37	*521955*	459	37	*844080*	364	45	1491	35				
S	***fixL***	*cgR2292*	48	432																		
R	***fixJ***	*cgR2299*	47	205																	*768547*	49
S	***phoR***	*cgR2510*	56	485	*376573*	485	89	*301382*	475	66	*521634*	519	49	*843214*	536	47	5836	39	32529	37	*766222*	41
R	***phoP***	*cgR2511*	54	240	*376574*	235	98	*301383*	240	91	*521635*	236	80	*843213*	233	73	5837	68	32528	67	*766223*	63
R	***baeR***	*cgR2566*	54	241	*376628*	241	92	*299167*	240	55	*519745*	240	57	*844363*	240	56					*770921*	39
S	***baeS***	*cgR2567*	58	372	*376629*	372	91	*299168*	398	47	*519744*	375	51	*844362*	375	50						
R	***yvqC***	*cgR2844*	61	212	*376890*	212	88	*301671*	225	87	*521954*	212	78	*844079*	216	43	6198	36			*768832*	41
S	***yvqE***	*cgR2845*	54	444	*376891*	444	94	*301672*	429	76	*521955*	459	51	*844080*	364	38						
S	***chrS***										*522014*	417	REF									
R	***chrA***										*522015*	199	REF									
R	***yfiK***				*376724*	206	REF	*301231*	267	38												
S	***yfiJ***				*376723*	380	REF	*301230*	380	22												

enzyme (*cgR1991*) that forms a putative operon with a regulatory protein of the TetR family (*cgR1990*). Likewise, all corynebacterial trehalose metabolic genes identified previously (Tzvetkov *et al.*, 2003) (*treS*, *cgR2175*; *treY*, *cgR2002*; *treZ*, *cgR2009*; *otsA*, *cgR2531*; *otsB*, *cgR2533*) are present in *C. glutamicum* R. Therefore, it is likely that in this bacterium all three trehalose metabolic pathways found in bacteria are enabled [TreS pathway from maltose, TreY/TreZ pathway from α(1–4)glucose polymers, OtsA/OtsB pathway from glucose 6-phosphate and UDP-glucose]. This observation reinforces the view that this non-reducing sugar plays a major physiological role in *Actinobacteria* as energy storage and as an environmental protectant against various stresses such as low water activity (desiccation, dehydration), external osmolality fluctuations, heat, cold and oxidation. It is also noteworthy that the genes *otsA* and *otsB* are part of a putative five-gene operon which is adjacent to a transcriptional regulator of the *lacI* family (*cgR2534*) located downstream and in the opposite orientation.

Two-component systems

Sensing of environmental conditions to ensure variability and adaptability to environmental changes is enabled by two-component systems linked to signal transduction cascades that function via phosphorelays (Taylor & Zhulin, 1999). Basic two-component systems comprise a membrane-associated sensor kinase that detects external stimuli and a transcriptional regulator that acts upon the cellular machinery to bring about the necessary adaptive changes (Fontan *et al.*, 2004). The genes encoding these two proteins are typically organized in two-gene operons. A similar organization is observed in the genome of *C. glutamicum* R where 27 ORFs are present that share a high level of homology with two-component system genes (Table 5). One orphan regulator (*cgR0730*), which is dispensable as demonstrated by the absence of phenotypic change upon its deletion (this work), and 13 putative two-component systems organized in two-gene operons encoding a sensor kinase and a regulator, have been identified. Despite being absent from the genome of *C. glutamicum* R, *narQ*, the putative cognate kinase gene of *cgR0730*, is present in *C. glutamicum* ATCC 13032, *C. efficiens* YS-314 and *C. diphtheriae* NCTC 13129. On the other hand, orthologues of *senX3/regX3*, *mtrA/mtrB*, *mprB/mprA* and *phoR/phoP* are present in all the corynebacteria and mycobacteria sequenced to date (Table 5).

The stimuli sensed by most of these two-component systems remain unknown, despite, for example, the fact that the role of *citA/citB* in the uptake and metabolism of citrate dicarboxylates has been firmly established (Gerharz *et al.*, 2003). In addition, in *M. tuberculosis*, genes under the positive control of PhoP include genes encoding proteins involved in lipid metabolism, cell wall synthesis, membrane transport and oxidative stress response (Fontan *et al.*, 2004). The role of PhoP in adaptation to phosphate-limited conditions has been confirmed in *C. glutamicum* ATCC 13032 (Kočan *et al.*, 2006). *M. tuberculosis phoP* mutants have an altered rounded shape and show altered levels of lipoarabinomannan derivatives compared to the wild-type (Fontan *et al.*, 2004), as well as having an impaired ability to synthesize methyl-branched fatty-acid-containing acyltrehaloses (Gonzalo Asensio *et al.*, 2006). Disruption of the putative *phoP* and *phoR* genes of *C. glutamicum* R demonstrated that these genes are not essential under standard laboratory conditions and that their deletion does not result in any obvious phenotype. Likewise, it has been shown that the *mprA/mprB* gene pair is linked to the sensing of cell-wall- or outer-membrane-related stress (He & Zahrt, 2005). It is also noteworthy that *C. glutamicum* R and *C. glutamicum* ATCC 13032 display, downstream of the putative *mprA-mprB* genes, a putative *pepD* gene (respectively, *cgR0990* and *VIMSS374897*), the product of which is a trypsin-like serine protease which is thought to aid in the degradation of misfolded proteins that are generated in response to various stresses. PepD has been demonstrated in *M. tuberculosis* to be secreted into the culture medium (Skeiky *et al.*, 1999). A similar genetic organization, where the genes *mprA-mprB-pepD* are part of an operon, has been observed in pathogenic mycobacteria (He & Zahrt, 2005). The conservation of this gene cluster in both saprophytic and pathogenic organisms suggests that it constitutes an important adaptive mechanism of the bacteria of the *Corynebacterineae* group. However, none of these genes appears to be essential under laboratory conditions, as demonstrated by the disruption of *mprA* and *mprB* in *M. tuberculosis* (Zahrt & Deretic, 2001). Similarly, we achieved the disruption of *cgR0988* (*mprA*), *cgR0989* (*mprB*) and *cgR0990* (*pepD*) in *C. glutamicum* R by transposon mutagenesis (this work).

Of these two-component systems, only *mtrA/mtrB* has been shown to be essential to the growth of *M. tuberculosis* as demonstrated by unsuccessful attempts to disrupt the *mtrA* gene, despite the fact that *mtrB* appears to be non-essential (Fontan *et al.*, 2004; Zahrt & Deretic, 2000). The two-component system *mtrA/mtrB* is conserved in all actinobacteria (Hoskisson & Hutchings, 2006) and its regulator is the only response regulator that is induced in *M. tuberculosis* during macrophage infection, but not in broth culture (Zahrt & Deretic, 2000). The two-component system MtrA/MtrB has been shown to strongly influence cellular morphology, antibiotic susceptibility and genetic expression of osmoprotection as demonstrated by the phenotype exhibited by a *C. glutamicum* ATCC 13032 double mutant (Möker *et al.*, 2004). In *C. glutamicum* R we succeeded in disrupting either *cgR0863* (putative *mtrA*) or *cgR0864* (putative *mtrB*) by inserting a transposon in the central regions of these genes. Neither of the resulting single-gene disruptants showed any significantly altered cellular morphology or altered growth pattern, perhaps indicative of cross-talk between the MtrA and MtrB proteins and one or more other two-component system proteins. The gene immediately downstream of (and overlapping) *mtrB* is conserved in all actinobacteria, where it encodes the actinobacteria signature protein LpqB (Gao *et al.*, 2006).

In the corynebacteria and mycobacteria sequenced to date, *mtrA* and *mtrB* are part of a putative operon that also contains *sahH* (*cgR0861* and *VIMSS374775* in *C. glutamicum* R and ATCC 13032, respectively), encoding *S*-adenosylhomocysteine hydrolase, and *tmpk* (*cgR0862* and *VIMSS374776*, respectively), encoding thymidylate kinase. On the other hand, the *cgR0122-cgR0123* two-component system, a putative orthologue of the *yycF/yycG* system of *B. subtilis* that modulates the expression of the *ftsAZ* operon (Fukuchi *et al.*, 2000), can be deleted from the *C. glutamicum* R genome by excision of SSI-3 without any significant phenotypic change (Suzuki *et al.*, 2005b). Brocker & Bott (cited in Kočan *et al.*, 2006) suggested that this two-component system is involved in the genetic regulation of copper metabolism. Interestingly, although this two-component system is borne by an SSI in *C. glutamicum* R, it shows a high level of identity with orthologues found in all the other corynebacteria sequenced to date. On the other hand, its relatively high G+C content (greater than 60 mol%) in all corynebacteria sequenced to date would suggest that its acquisition is a recent event in evolutionary terms. Last, the response regulator encoded by gene *cgR2566*, the product of which shares a high level of homology with the BaeR protein of *E. coli*, is also conserved in all corynebacteria sequenced to date (Table 5). We did not succeed in disrupting or deleting gene *cgR2566* despite being able to mutate *cgR2567* by transposon insertion, its cognate gene encoding a sensor kinase.

Interestingly, two corynebacterial two-component systems appear to be specific to saprophytic corynebacteria. Their highest levels of identity are shared with the putative *B. subtilis* systems *ykoG/ykoH* (*cgR0359/cgR0360*) and *yvfT/yvfU* (*cgR1049/cgR1050*) (Fabret *et al.*, 1999). Similar to what has been observed in *B. subtilis*, these genes appear to be non-essential in corynebacteria as demonstrated by the isolation of several transposon mutants of *C. glutamicum* R where these genes had been inactivated (this work). However, saprophytic corynebacteria appear to be devoid of *chrS/chrA* or *cstS/cstA* orthologues. These latter two-component systems, present in both *C. diphtheriae* and *C. jeikeium*, have been linked to haem and haemoglobin sensing (Schmitt, 1999). Nevertheless, ORFs *cgR2844* and *cgR1838* show low identity but relatively high homology to *chrA* (67 and 63 %, respectively) and *cstA* (53 %) from pathogenic corynebacteria. Likewise, corynebacteria appear to be devoid of *devS/devR* orthologues, which have been linked in *M. tuberculosis* and *M. bovis* with hypoxic dormancy (Boon & Dick, 2002; Fontan *et al.*, 2004; Saini *et al.*, 2004), albeit that *cgR2844* and *cgR1838* are 52 and 57 % homologous, respectively, to *M. bovis devR*. Genes *cgR2844* and *cgR2845* are dispensable in *C. glutamicum* R, as demonstrated by their simultaneous deletion via Cre/*loxP*-mediated rearrangements (this work).

Moreover, the genome of *C. glutamicum* R reveals the presence of two unique two-component systems, exhibiting high similarity to *cutS/cutR* of *Streptomyces coelicolor* involved in negative regulation of actinorhodin synthesis (Hutchings *et al.*, 2004), and to *fixL/fixJ* of *Azorhizobium caulinodans* involved in nitrogen fixation in response to low-oxygen conditions (Fischer, 1994). These genes are present on two different SSIs (SSI-5, a region rich in transposase genes, and SSI-9, respectively) (Suzuki *et al.*, 2005a) and are characterized by G+C contents that differ significantly from the typical G+C content of corynebacterial genomes (Table 5). These observations corroborate the view that these sequences have been acquired by *C. glutamicum* R through horizontal transfer events that occurred relatively recently in evolutionary terms. As previously reported, these sequences are dispensable as demonstrated by the absence of a specific phenotype in the corresponding *C. glutamicum* R deletion mutants (Suzuki *et al.*, 2005a). Similarly, *C. glutamicum* ATCC 13032 exhibits one unique two-component system (*VIMSS376724* and *VIMSS376723*) that shows weak identity to two *C. efficiens* genes. It is also worth noting that *C. diphtheriae* also possesses a unique two-component system on its chromosome (*VIMSS519770* and *VIMSS519769*), albeit that two similar genes are borne by the *C. efficiens* plasmid pCE3 (Table 5). Likewise, *C. jeikeium* K411 encodes two two-component system genes, *kdpD* and *kdpE*, putatively involved in turgor pressure sensing and regulation (Hutchings *et al.*, 2004), that are absent from the other corynebacteria sequenced to date, but present in mycobacteria and *Streptomyces* species (Hutchings *et al.*, 2004).

These different observations promote the view that, except for a few sensor–regulator couples that are highly conserved in *Corynebacterineae*, the sensing and regulation machinery of corynebacteria is highly plastic. This raises the question whether horizontal transfers of these important adaptation genes, rather than gene decay, has played a major role during the course of evolution of these bacteria. Such a view is consistent with the observed presence in a few corynebacterial strains of two-component systems for which orthologues can only be found in mycobacteria, *Streptomyces* species or *N. farcinica*, but not in the other corynebacteria sequenced to date. Furthermore, a few of these observed systems either are borne by an episome or are part of an SSI. On the other hand, the overall conservation of these systems between corynebacteria and mycobacteria further validates corynebacteria as important model organisms to contribute to the understanding of the biology of slow-growing pathogenic mycobacteria.

Global differences between saprophytic and pathogenic corynebacteria

Based on our calculations and model, the presence in the corynebacteria sequenced to date of up to 762 genes identified in COG appears limited to the saprophytes (Table 3). In contrast, pathogenic corynebacteria display fewer candidate pathogen-specific genes (Table 3). For example, we calculated that the saprophytic corynebacteria sequenced to date exhibit approximately 30 % more

transporters per number of ORFs in their genomes than the pathogenic organisms *C. diphtheriae* and *C. jeikeium*. This difference constitutes an indication of the relatively greater metabolic versatility of the former organisms. This observation also suggests a higher degree of transporter substrate specificity and a higher number of secondary carriers, since overall the substrate specificities of transport systems of an organism are correlated to its ecological niche and the diversity and relative concentrations of nutrients it encounters (Paulsen *et al.*, 2000).

Likewise, among the genes potentially involved in host interactions, microbiofilm formation, DNA transfer and bacteriophage attachment, *C. diphtheriae* NCTC 13129 harbours 11 genes putatively encoding fimbrial proteins or fimbria-associated proteins, whereas *C. glutamicum* R only exhibits four such genes which are also present in *C. efficiens* (*cgR2789*, *VIMSS301626*; *cgR2790*, *VIMSS301627*; *cgR2791*, *VIMSS301628*; *cgR2793*, *VIMSS301630*, respectively). Notably, only one cluster of pili genes is found in saprophytic bacteria, whereas the genome of *C. diphtheriae* NCTC 13129 exhibits three such clusters (Gaspar & Ton-That, 2006). Similarly, *C. diphtheriae* NCTC 13129 and *C. jeikeium* K411 harbour, in comparison with *C. glutamicum* and *C. efficiens*, a relatively larger number of secreted or surface-anchored proteins, in relation to the ecological niche occupied by these species.

Notably, genes related to amino acid transport and metabolism are fewer in number in pathogenic corynebacteria than in the saprophytic ones (Table 3). In addition to horizontal DNA transfer that has occurred in the saprophytic corynebacteria, this difference has been ascribed to gene decay that has occurred during the course of the evolution of the former organisms (Nishio *et al.*, 2004). Interestingly, this specialization is not the rule in *Corynebacterineae*, since for example the chromosome sequence of *N. farcinica* reveals that this organism includes many genes for virulence, drug resistance and secondary metabolism. Interestingly, analyses of paralogous protein families suggest that gene duplications have resulted in *N. farcinica* being able to survive not only in soil environments, but also in animal tissues (Ishikawa *et al.*, 2004).

Functional differences between *C. glutamicum* R and ATCC 13032

A detailed comparison of the genomic sequence of *C. glutamicum* R with that of ATCC 13032 reveals that only 60 and 189 genes, respectively, are strain-specific. Relatively, both of these strains encode the same number of genes per COG category (Table 3), with a few genes unique to either *C. glutamicum* strain R or ATCC 13032 when compared to the corynebacteria sequenced to date (Table 3). With the exception of sequences from mobile elements, the most striking differences in the number of genes are observed in amino acid transport and metabolism, and secondary metabolite transport and metabolism.

Genes unique to *C. glutamicum* R

A total of 9 % (263) of the predicted genes of *C. glutamicum* ATCC 13032 remain hypothetical or specific to *C. glutamicum* when compared to *C. efficiens* YS-314 and *C. diphtheriae* NCTC 13129 (Kalinowski *et al.*, 2003). It has been suggested that sequencing more than one or two genomes per species is necessary to access bacterial pan-genomes (Tettelin *et al.*, 2005). For example, the sequencing of eight strains of *Streptococcus agalactiae* was found to be sufficient to define the core genome of these organisms with 95 % confidence, whereas each new *S. agalactiae* genome sequence would reveal an extrapolated 33 new genes, with a 6×10^{-4} probability that this number falls to zero. The complete genome sequence of *C. glutamicum* R reveals 39 genes that to the best of our knowledge have not been previously identified in corynebacteria. In addition to the previously described β-glucoside phosphotransferase (*cgR0436*) that confers cellobiose utilization properties upon occurrence of a single amino acid substitution (Kotrba *et al.*, 2003), and several mobile-element-derived ORFs (IS*CgR3a*, IS*CgR5*, IS*CgR12b*, IS*CgR13a*, IS*CgR13b*), *C. glutamicum* R notably encodes five novel conserved hypothetical proteins (*cgR0052*, *cgR0067*, *cgR1134*, *cgR2375*, *cgR2798*), four membrane proteins putatively involved in transport mechanisms (*cgR0768*, *cgR2326*, *cgR2800*, *cgR2956*), as well as four regulatory proteins (*cgR0139*, *cgR0414*, *cgR1106*, *cgR2822*) in addition to the two two-component systems *cgR2292/cgR2299* and *cgR0540/cgR0541* discussed previously. Interestingly, *C. glutamicum* R encodes several genes that reflect its ecological niche in soil, as exemplified by a putative tyramine oxidase gene (*cgR0016*) that could be involved in the degradation of phenethylamines, compounds present in natural environments, or by a putative L-asparaginase (*cgR2808*). As a result, to access the diversity of the metabolism of corynebacteria, a significant number of additional genome sequences would need to be determined.

ACKNOWLEDGEMENTS

This research was supported by New Energy and Industrial Technology Development Organization (NEDO), Japan.

REFERENCES

Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L. & Arkin, A. P. (2005). The MicrobesOnline Web site for comparative genomics. *Genome Res* **15**, 1015–1022.

Besemer, J., Lomsadze, A. & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**, 2607–2618.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K. & other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474.

Bonamy, C., Labarre, J., Reyes, O. & Leblon, G. (1994). Identification of IS*1206*, a *Corynebacterium glutamicum* IS*3*-related

insertion sequence and phylogenetic analysis. *Mol Microbiol* **14**, 571–581.

Bonfield, J. K., Smith, K. & Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res* **23**, 4992–4999.

Boon, C. & Dick, T. (2002). *Mycobacterium bovis* BCG response regulator essential for hypoxic dormancy. *J Bacteriol* **184**, 6760–6767.

Calamita, H., Ko, C., Tyagi, S., Yoshimatsu, T., Morrison, N. E. & Bishai, W. R. (2005). The *Mycobacterium tuberculosis* SigD sigma factor controls the expression of ribosome-associated gene products in stationary phase and is required for full virulence. *Cell Microbiol* **7**, 233–244.

Cerdeño-Tárraga, A. M., Efstratiou, A., Dover, L. G., Holden, M. T., Pallen, M., Bentley, S. D., Besra, G. S., Churcher, C., James, K. D. & other authors (2003). The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC 13129. *Nucleic Acids Res* **31**, 6516–6523.

Chiou, C. Y., Wang, H. H. & Shaw, G. C. (2002). Identification and characterization of the non-PTS *fru* locus of *Bacillus megaterium* ATCC 14581. *Mol Genet Genomics* **268**, 240–248.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S. & other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honore, N., Garnier, T., Churcher, C. & other authors (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011.

Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999a). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**, 4636–4641.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O. & Salzberg, S. L. (1999b). Alignment of whole genomes. *Nucleic Acids Res* **27**, 2369–2376.

Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478–2483.

Demain, A. L. (2000). Small bugs, big business: the economic power of the microbe. *Biotechnol Adv* **18**, 499–514.

Engels, S., Schweitzer, J. E., Ludwig, C., Bott, M. & Schäffer, S. (2004). *clpC* and *clpP1P2* gene expression in *Corynebacterium glutamicum* is controlled by a regulatory network involving the transcriptional regulators ClgR and HspR as well as the ECF sigma factor sigmaH. *Mol Microbiol* **52**, 285–302.

Ermolaeva, M. D., White, O. & Salzberg, S. L. (2001). Prediction of operons in microbial genomes. *Nucleic Acids Res* **29**, 1216–1221.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* **8**, 186–194.

Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* **8**, 175–185.

Fabret, C., Feher, V. A. & Hoch, J. A. (1999). Two-component signal transduction in *Bacillus subtilis*: how one organism sees its world. *J Bacteriol* **181**, 1975–1983.

Fischer, H. M. (1994). Genetic regulation of nitrogen fixation in rhizobia. *Microbiol Rev* **58**, 352–386.

Fontan, P. A., Walters, S. & Smith, I. (2004). Cellular signaling pathways and transcriptional regulation in *Mycobacterium tuberculosis*: stress control and virulence. *Current Sci* **86**, 122–134.

Fukuchi, K., Kasahara, Y., Asai, K., Kobayashi, K., Moriya, S. & Ogasawara, N. (2000). The essential two-component regulatory system encoded by *yycF* and *yycG* modulates expression of the *ftsAZ* operon in *Bacillus subtilis*. *Microbiology* **146**, 1573–1583.

Gao, B., Paramanathan, R. & Gupta, R. S. (2006). Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. *Antonie Van Leeuwenhoek* **90**, 69–91.

Gaspar, A. H. & Ton-That, H. (2006). Assembly of distinct pilus structures on the surface of *Corynebacterium diphtheriae*. *J Bacteriol* **188**, 1526–1533.

Gerharz, T., Reinelt, S., Kaspar, S., Scapozza, L. & Bott, M. (2003). Identification of basic amino acid residues important for citrate binding by the periplasmic receptor domain of the sensor kinase CitA. *Biochemistry* **42**, 5917–5924.

Glaser, P., Kunst, F., Arnaud, M., Coudart, M. P., Gonzales, W., Hullo, M. F., Ionescu, M., Lubochinsky, B., Marcelino, L. & other authors (1993). *Bacillus subtilis* genome project: cloning and sequencing of the 97-kb region from 325 degrees to 333 degrees. *Mol Microbiol* **10**, 371–384.

Goh, E. B., Bledsoe, P. J., Chen, L. L., Gyaneshwar, P., Stewart, V. & Igo, M. M. (2005). Hierarchical control of anaerobic gene expression in *Escherichia coli* K-12: the nitrate-responsive NarX-NarL regulatory system represses synthesis of the fumarate-responsive DcuS-DcuR regulatory system. *J Bacteriol* **187**, 4890–4899.

Gonzalo Asensio, J., Maia, C., Ferrer, N. L., Barilone, N., Laval, F., Soto, C. Y., Winter, N., Daffe, M., Gicquel, B. & other authors (2006). The virulence-associated two-component PhoP-PhoR system controls the biosynthesis of polyketide-derived lipids in *Mycobacterium tuberculosis*. *J Biol Chem* **281**, 1313–1316.

Gordon, D., Abajian, C. & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195–202.

Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* **26**, 2286–2290.

Hahn, M. Y., Raman, S., Anaya, M. & Husson, R. N. (2005). The *Mycobacterium tuberculosis* extracytoplasmic-function sigma factor SigL regulates polyketide synthases and secreted or membrane proteins and is required for virulence. *J Bacteriol* **187**, 7062–7071.

Hansmeier, N., Chao, T. C., Pühler, A., Tauch, A. & Kalinowski, J. (2006). The cytosolic, cell surface and extracellular proteomes of the biotechnologically important soil bacterium *Corynebacterium efficiens* YS-314 in comparison to those of *Corynebacterium glutamicum* ATCC 13032. *Proteomics* **6**, 233–250.

He, H. & Zahrt, T. C. (2005). Identification and characterization of a regulatory sequence recognized by *Mycobacterium tuberculosis* persistence regulator MprA. *J Bacteriol* **187**, 202–212.

Hermann, T. (2003). Industrial production of amino acids by coryneform bacteria. *J Biotechnol* **104**, 155–172.

Hermann, T., Pfefferle, W., Baumann, C., Busker, E., Schaffer, S., Bott, M., Sahm, H., Dusch, N., Kalinowski, J. & other authors (2001). Proteome analysis of *Corynebacterium glutamicum*. *Electrophoresis* **22**, 1712–1723.

Hoskisson, P. A. & Hutchings, M. I. (2006). MtrAB-LpqB: a conserved three-component system in actinobacteria? *Trends Microbiol* **14**, 444–449.

Hutchings, M. I., Hoskisson, P. A., Chandra, G. & Buttner, M. J. (2004). Sensing and responding to diverse extracellular signals? Analysis of the sensor kinases and response regulators of *Streptomyces coelicolor* A3(2). *Microbiology* **150**, 2795–2806.

Ikeda, M. & Nakagawa, S. (2003). The *Corynebacterium glutamicum* genome: features and impacts on biotechnological processes. *Appl Microbiol Biotechnol* **62**, 99–109.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. & Omura, S. (2003). Complete

genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol* **21**, 526–531.

Inui, M., Murakami, S., Okino, S., Kawaguchi, H., Vertès, A. A. & Yukawa, H. (2004). Metabolic analysis of *Corynebacterium glutamicum* during lactate and succinate productions under oxygen deprivation conditions. *J Mol Microbiol Biotechnol* **7**, 182–196.

Ishikawa, J., Yamashita, A., Mikami, Y., Hoshino, Y., Kurita, H., Hotta, K., Shiba, T. & Hattori, M. (2004). The complete genomic sequence of *Nocardia farcinica* IFM 10152. *Proc Natl Acad Sci U S A* **101**, 14925–14930.

Jensen-Cain, D. M. & Quinn, F. D. (2001). Differential expression of *sigE* by *Mycobacterium tuberculosis* during intracellular growth. *Microb Pathog* **30**, 271–278.

Kalinowski, J. (2005). The genomes of amino acid-producing corynebacteria. In *Handbook of Corynebacterium glutamicum*, pp. 37–56. Edited by L. Eggeling & M. Bott. Boca Raton, FL: CRC Press.

Kalinowski, J., Bathe, B., Bartels, D., Bischoff, N., Bott, M., Burkovski, A., Dusch, N., Eggeling, L., Eikmanns, B. J. & other authors (2003). The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. *J Biotechnol* **104**, 5–25.

Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S. & other authors (2003). Essential *Bacillus subtilis* genes. *Proc Natl Acad Sci U S A* **100**, 4678–4683.

Kočan, M., Schaffer, S., Ishige, T., Sorger-Herrmann, U., Wendisch, V. F. & Bott, M. (2006). Two-component systems of *Corynebacterium glutamicum*: deletion analysis and involvement of the PhoS-PhoR system in the phosphate starvation response. *J Bacteriol* **188**, 724–732.

Kotrba, P., Inui, M. & Yukawa, H. (2003). A single V317A or V317M substitution in Enzyme II of a newly identified β-glucoside phosphotransferase and utilization system of *Corynebacterium glutamicum* R extends its specificity towards cellobiose. *Microbiology* **149**, 1569–1580.

Kumagai, H. (2000). Microbial production of amino acids in Japan. *Adv Biochem Eng Biotechnol* **69**, 71–85.

Li, L., Bannantine, J. P., Zhang, Q., Amonsin, A., May, B. J., Alt, D., Banerji, N., Kanjilal, S. & Kapur, V. (2005). The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*. *Proc Natl Acad Sci U S A* **102**, 12344–12349.

Liebl, W. (2001). *Corynebacterium* – nonmedical. In *The Prokaryotes: A Handbook on the Biology of Bacteria, Vol. 3: Archaea and Bacteria: Firmicutes, Actinomycetes*, pp. 796–818. Edited by M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer & E. Stackebrandt. New York: Springer.

Liebl, W. (2005). *Corynebacterium* taxonomy. In *Handbook of Corynebacterium glutamicum*, pp. 9–34. Edited by L. Eggeling & M. Bott. Boca Raton, FL: CRC Press.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964.

Mahillon, J. & Chandler, M. (1998). Insertion sequences. *Microbiol Mol Biol Rev* **62**, 725–774.

Missiakas, D. & Raina, S. (1998). The extracytoplasmic function sigma factors: role and regulation. *Mol Microbiol* **28**, 1059–1066.

Möker, N., Brocker, M., Schaffer, S., Krämer, R., Morbach, S. & Bott, M. (2004). Deletion of the genes encoding the MtrA-MtrB two-component system of *Corynebacterium glutamicum* has a strong influence on cell morphology, antibiotics susceptibility and expression of genes involved in osmoprotection. *Mol Microbiol* **54**, 420–438.

Moon, M. W., Kim, H. J., Oh, T. K., Shin, C. S., Lee, J. S., Kim, S. J. & Lee, J. K. (2005). Analyses of enzyme II gene mutants for sugar transport and heterologous expression of fructokinase gene in *Corynebacterium glutamicum* ATCC 13032. *FEMS Microbiol Lett* **244**, 259–266.

Mukherjee, R. & Chatterji, D. (2005). Evaluation of the role of sigma B in *Mycobacterium smegmatis*. *Biochem Biophys Res Commun* **338**, 964–972.

Nakamura, Y., Nishio, Y., Ikeo, K. & Gojobori, T. (2003). The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. *Gene* **317**, 149–155.

Nishino, K., Honda, T. & Yamaguchi, A. (2005). Genome-wide analyses of *Escherichia coli* gene expression responsive to the BaeSR two-component regulatory system. *J Bacteriol* **187**, 1763–1772.

Nishio, Y., Nakamura, Y., Kawarabayasi, Y., Usuda, Y., Kimura, E., Sugimoto, S., Matsui, K., Yamagishi, A., Kikuchi, H. & other authors (2003). Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Res* **13**, 1572–1579.

Nishio, Y., Nakamura, Y., Usuda, Y., Sugimoto, S., Matsui, K., Kawarabayasi, Y., Kikuchi, H., Gojobori, T. & Ikeo, K. (2004). Evolutionary process of amino acid biosynthesis in *Corynebacterium* at the whole genome level. *Mol Biol Evol* **21**, 1683–1691.

Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H., Jr (2000). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J Mol Biol* **301**, 75–100.

Raman, S., Song, T., Puyang, X., Bardarov, S., Jacobs, W. R., Jr & Husson, R. N. (2001). The alternative sigma factor SigH regulates major components of oxidative and heat stress responses in *Mycobacterium tuberculosis*. *J Bacteriol* **183**, 6119–6125.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277.

Saier, M. H., Jr (2002). *The Bacterial Phosphotransferase System*. Norwich: Horizon.

Saini, D. K., Malhotra, V., Dey, D., Pant, N., Das, T. K. & Tyagi, J. S. (2004). DevR–DevS is a bona fide two-component system of *Mycobacterium tuberculosis* that is hypoxia-responsive in the absence of the DNA-binding domain of DevR. *Microbiology* **150**, 865–875.

Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**, 544–548.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Biology: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.

Schmitt, M. P. (1999). Identification of a two-component signal transduction system from *Corynebacterium diphtheriae* that activates gene expression in response to the presence of heme and hemoglobin. *J Bacteriol* **181**, 5330–5340.

Shen, X. H., Jiang, C. Y., Huang, Y., Liu, Z. P. & Liu, S. J. (2005). Functional identification of novel genes involved in the glutathione-independent gentisate pathway in *Corynebacterium glutamicum*. *Appl Environ Microbiol* **71**, 3442–3452.

Shilo, M. & Wolman, B. (1958). Activities of bacterial levans and of lipopolysaccharides in the processes of inflammation and infection. *Br J Exp Pathol* **39**, 652–660.

Skeiky, Y. A., Lodes, M. J., Guderian, J. A., Mohamath, R., Bement, T., Alderson, M. R. & Reed, S. G. (1999). Cloning, expression, and immunological evaluation of two putative secreted serine protease antigens of *Mycobacterium tuberculosis*. *Infect Immun* **67**, 3998–4007.

Stackebrandt, E. & Woese, C. R. (1981). The evolution of prokaryotes. In *Molecular and Cellular Aspects of Microbial Evolution*, pp. 1–31. Edited by M. J. Carlile, J. F. Collins & B. E. B. Moseley. Cambridge: Cambridge University Press.

Stackebrandt, E., Rainey, F. A. & Ward-Rainey, N. L. (1997). Proposal for a new hierarchical classification system, *Actinobacteria* classis nov. *Int J Syst Bacteriol* **47**, 479–491.

Staden, R. (1996). The Staden sequence analysis package. *Mol Biotechnol* **5**, 233–241.

Stecker, C., Johann, A., Herzberg, C., Averhoff, B. & Gottschalk, G. (2003). Complete nucleotide sequence and genetic organization of the 210-kilobase linear plasmid of *Rhodococcus erythropolis* BD2. *J Bacteriol* **185**, 5269–5274.

Stothard, P. & Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. *Bioinformatics* **21**, 537–539.

Sun, R., Converse, P. J., Ko, C., Tyagi, S., Morrison, N. E. & Bishai, W. R. (2004). *Mycobacterium tuberculosis* ECF sigma factor *sigC* is required for lethality in mice and for the conditional expression of a defined gene set. *Mol Microbiol* **52**, 25–38.

Suzuki, N., Okayama, S., Nonaka, H., Tsuge, Y., Inui, M. & Yukawa, H. (2005a). Large-scale engineering of the *Corynebacterium glutamicum* genome. *Appl Environ Microbiol* **71**, 3369–3372.

Suzuki, N., Tsuge, Y., Inui, M. & Yukawa, H. (2005b). Cre/*loxP*-mediated deletion system for large genome rearrangements in *Corynebacterium glutamicum*. *Appl Microbiol Biotechnol* **67**, 225–233.

Suzuki, N., Okai, N., Nonaka, H., Tsuge, Y., Inui, M. & Yukawa, H. (2006). High throughput transposon mutagenesis of *Corynebacterium glutamicum* and construction of a single-gene disruptant mutant library. *Appl Environ Microbiol* **72**, 3750–3755.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631–637.

Tauch, A., Gotker, S., Puhler, A., Kalinowski, J. & Thierbach, G. (2002). The 27.8-kb R-plasmid pTET3 from *Corynebacterium glutamicum* encodes the aminoglycoside adenyltransferase gene cassette *aadA9* and the regulated tetracycline efflux system Tet 33 flanked by active copies of the widespread insertion sequence IS*6100*. *Plasmid* **48**, 117–129.

Tauch, A., Puhler, A., Kalinowski, J. & Thierbach, G. (2003). Plasmids in *Corynebacterium glutamicum* and their molecular classification by comparative genomics. *J Biotechnol* **104**, 27–40.

Tauch, A., Kaiser, O., Hain, T., Goesmann, A., Weisshaar, B., Albersmeier, A., Bekel, T., Bischoff, N., Brune, I. & other authors (2005). Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. *J Bacteriol* **187**, 4671–4682.

Taylor, B. L. & Zhulin, I. B. (1999). PAS domains: internal sensors of oxygen, redox potential, and light. *Microbiol Mol Biol Rev* **63**, 479–506.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L. & other authors (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc Natl Acad Sci U S A* **102**, 13950–13955.

Tzvetkov, M., Klopprogge, C., Zelder, O. & Liebl, W. (2003). Genetic dissection of trehalose biosynthesis in *Corynebacterium glutamicum*: inactivation of trehalose production leads to impaired growth and an altered cell wall lipid composition. *Microbiology* **149**, 1659–1673.

Vertès, A. A., Hatakeyama, K., Inui, M., Kobayashi, M., Kurusu, Y. & Yukawa, H. (1993a). Replacement recombination in coryneform bacteria: high efficiency integration requirement for non-methylated plasmid DNA. *Biosci Biotechnol Biochem* **57**, 2036–2038.

Vertès, A. A., Inui, M., Kobayashi, M., Kurusu, Y. & Yukawa, H. (1993b). Presence of *mrr*- and *mcr*-like restriction systems in coryneform bacteria. *Res Microbiol* **144**, 181–185.

Vertès, A. A., Inui, M. & Yukawa, H. (2005). Manipulating corynebacteria from individual genes to chromosome. *Appl Environ Microbiol* **71**, 7633–7642.

von Graevenitz, A. & Bernard, K. (2001). The genus *Corynebacterium* – medical. In *The Prokaryotes: A Handbook on the Biology of Bacteria, Vol. 3: Archaea and Bacteria: Firmicutes, Actinomycetes*, pp. 819–842. Edited by M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer & E. Stackebrandt. New York: Springer.

Waagmeester, A., Thompson, J. & Reyrat, J. M. (2005). Identifying sigma factors in *Mycobacterium smegmatis* by comparative genomic analysis. *Trends Microbiol* **13**, 505–509.

Winnen, B., Felce, J. & Saier, M. H., Jr (2005). Genomic analyses of transporter proteins in *Corynebacterium glutamicum* and *Corynebacterium efficiens*. In *Handbook of Corynebacterium glutamicum*, pp. 149–186. Edited by L. Eggeling & M. Bott. Boca Raton, FL: CRC Press.

Yukawa, H., Inui, M. & Vertès, A. A. (2007). Genomes and genome-level engineering of amino-acid producing bacteria. In *Amino Acid Biosynthesis - Pathways, Regulation and Metabolic Engineering*. Edited by V. F. Wendisch. Heidleberg: Springer (in press).

Zahrt, T. C. & Deretic, V. (2000). An essential two-component signal transduction system in *Mycobacterium tuberculosis*. *J Bacteriol* **182**, 3832–3838.

Zahrt, T. C. & Deretic, V. (2001). *Mycobacterium tuberculosis* signal transduction system required for persistent infections. *Proc Natl Acad Sci U S A* **98**, 12706–12711.

Zhang, R. & Zhang, C. T. (2004). A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics* **20**, 612–622.

Edited by: C. W. Chen