SOM MATERIALS AND METHODS

Sequencing and assembly of the *Kuenenia stuttgartiensis* genome

Operation of the laboratory bioreactor

The anaerobic ammonium oxidizing bioreactor consisted of an anoxic 10 l gas-lift reactor fed with synthetic wastewater (350 ml/day) and sparged with dinitrogen gas (10-20 ml/min) as previously described[1]. The inoculum originated from the nitrification stage of the Dokhaven wastewater treatment plant (Rotterdam, the Netherlands). *Kuenenia stuttgartiensis* made up less then one percent of the microbial community in the inoculum and during the one year operation of the bioreactor this percentage increased to about 73% as determined by quantitative FISH using specific probes[2] and the image analysis program DAIME[3]. At the time of sampling (one year after inoculation), the ammonium and nitrite concentrations in the influent were 250-300 mM and 100-200 ml/min of gas was recycled from the top to the bottom of the bioreactor for mixing.

DNA preparation

The same DNA extraction protocol was used for all genomic libraries. DNA extraction was performed in agarose plugs from an enrichment culture pellet. The plugs were incubated for 3 hours at 37°C in EDTA (100 mM), NaCl (50 mM), Tris-HCl (10 mM), Na Lauroyl sarcosine (0.5%) pH 8. This solution also contained lysozyme (1 mg/mL), mutanolysine (10 U/mL), lipase (1 mg/mL), peptidase (1 mg/mL) and β-glucuronidase (1 mg/mL). The agarose plugs were then incubated twice for 24 hours at 50°C in EDTA (0.5 M), Na Lauroyl sarcosine (0.5%) pH 8 containing protease (500 ng/mL) and proteinase K (2 mg/mL).

Genomic libraries

A BAC library comprising 8448 clones was constructed from partial HindIII digests using the vector pBeloBAC5 (Epicentre), as described previously[4]. In addition, a fosmid library containing 6432 clones was constructed into pEpiFOS5 (Epicentre), and a shotgun randomly sheared DNA plasmid library was constructed using pCDNA2 (a low copy cloning vector). Clones from all libraries were picked and bi-directionally sequenced by using standard protocols.

Assembly and assignment of contigs to *K. stuttgartiensis*

A preliminary global assembly of 192713 sequence reads was performed using Phrap. This produced a great number of small contigs, probably due to the presence of repeated sequences from *K. stuttgartiensis* and/or other microorganisms, as well as the presence of many other microorganisms in the original enrichment culture. All sequences harboring a motif of at least 10 nucleotides which were repeated more than 50 times in the whole set of reads were removed and a novel global assembly (127557 reads) was then performed; this resulted in 9787 contigs. Most of these were still very small, but 493 contigs could be organized into 68 supercontigs by BAC or fosmid bridging.

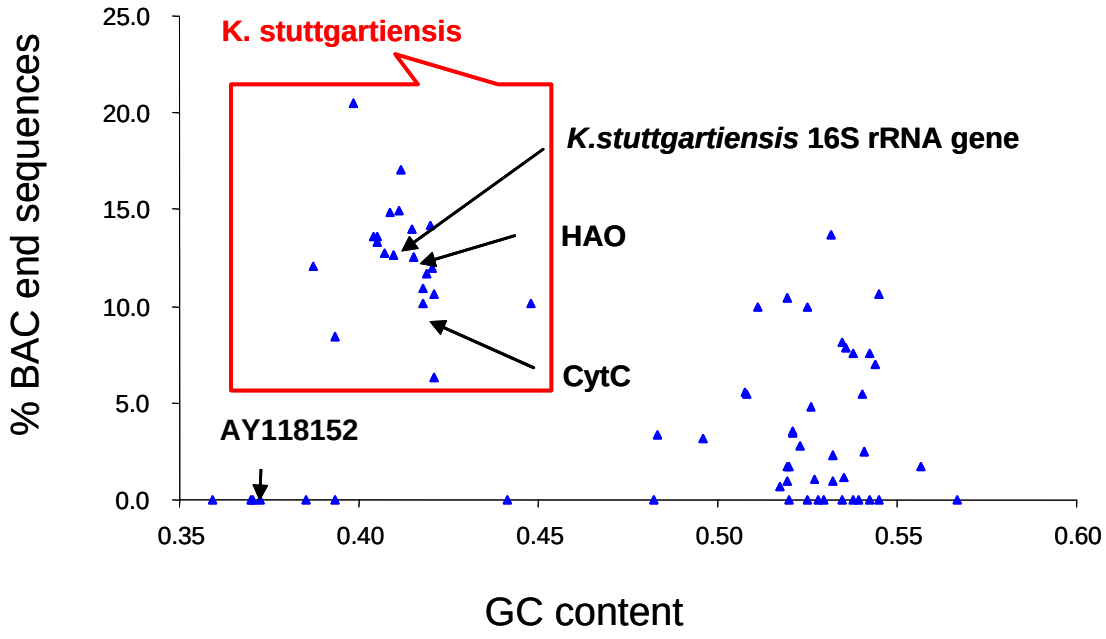

Figure: Distribution of the percentage of the BAC end sequences versus the GC%.

To classify the supercontigs into probable *K. stuttgartiensis* and non-*K.stuttgartiensis* entities, several strategies were utilised. Firstly, a contig containing the *K. stuttgartiensis* 16S rRNA gene (85265 bp) was found to have a GC content of 41%, guiding later selection of supercontigs based on GC content. Furthermore, it was found that some contigs contained BAC end sequences while others did not. Two additional genes indicative of anammox organisms (hydroxylamine oxidoreductase, HAO and cytochrome C, CytC) were then found in two contigs (28823 and 4567 bp, respectively). These contigs were composed more than 10% by BAC end sequences while a contig (40892 bp) containing a 16S rRNA gene (AY118152) affiliated to an uncultured *Chlorobi* did not contain any BAC end sequences,

suggesting that the % of BAC ends is a suitable parameter for binning supercontigs. Thus, 20 *K. stuttgartiensis* candidate supercontigs were chosen for subsequent assembly work based on the GC content and % of BAC end sequences, (see Figure above). In addition, another contig apparently representing a plasmid fulfilled both criteria (29731 bp) but was not further considered.

After careful analysis of the 20 supercontigs, together with gap filling by sequencing of relevant PCR fragments and BAC or fosmid clones, 5 contigs (totalling 4,218,325 nt) were finally obtained (**SOM Fig. 2**). We tried to order and /or bridge these contigs by different methods (e.g. PCR, XL PCR between all contigs using genomic DNA) and by looking for all possible BAC extensions. However, despite all these efforts we were unfortunately unable to bridge the remaining gaps, although the possibility exists that one or more of the contigs represent linear chromosomes. The mean read coverage over the 5 contigs is 22 x, more than is usually achieved in a "traditional" genome project. Moreover, 1.3 Mb (32% of the assembled genome) has been covered with individual BAC/fosmid sequences, further supporting the robustness of the assembled sequence of *K. stuttgartiensis*. Interestingly, the representation of *Kuenenia* genome fragments in the BAC vectors roughly matched the abundance of this organism in the enrichment (about 70% of the BACs contained *Kuenenia* genome fragments), the representation was much lower in the high-copy plasmids used for shotgun-sequencing (about 30%), clearly indicating a substantial cloning bias.

Accession numbers. The 5 contigs were submitted to EMBL and have the following numbers: CT030148, CT573071, CT573072, CT573073, CT573074.

Assembly confirmation and *Brocadia* homologues. The robustness of the assembly was further supported by the phylogenetic analysis of conserved genes. Of all phylogenetic marker genes analysed (**SOM Table 6**) only those proteins or rRNA genes (n = 33) were considered as confirmation points for which *K. stuttgartiensis* clusters with at least one of the planctomycetes *Rhodopirellula baltica* or *Gemmata obscuriglobus* in neighbour-joining (NJ) or maximum parsimony trees. In a complementary approach, 88 shotgun sequences were obtained from a 99.5% pure cell preparation of the related anammox bacterium *Brocadia anammoxidans*[5]. These were blasted (Blastn as well as Blastp on ORFs longer than 150 bp) against 180 complete bacterial genomes (including *R. baltica*) plus that of *K. stuttgartiensis.* If the best Blastn hit was with a *K. stuttgartiensis* genome sequence fragment and the E-value was < 0.01, the hit was listed. Additionally all predicted ORFs as inferred from the *B. anammoxidans* shotgun sequences were searched against all bacterial protein sequences deposited at GenBank plus all predicted proteins from the *K. stuttgartiensis* genome. Again, if

the best Blastp hit was with *K. stuttgartiensis* and the E-value was < 0.01, the hit was listed. Finally, both lists were merged and all *B. anammoxidans* sequences were removed which were only supported by either a genome hit with an E-value > 1E-05 or a protein hit with an E-value > 1E-05. After this procedure, 73 of the *B. anammoxidans* sequences remained on the list, strongly suggesting that the almost complete *K. stuttgartiensis* genome is represented in the 5 contigs.

Analysis of repeat sequences. Repeats were analyzed using the programs REPuter[6] and Nosferatu[7]. A total number of 382 regions consisting of repeats with lengths ranging from the lower threshold of 180 nucleotides up to 3791 nucleotides could be found. 213 repetitive regions (containing 220,280 nucleotides) cover mostly intergenic regions, whereas 169 regions (207,882 nt) are dominated by coding sequences.

Genome annotation. Prediction of coding sequences was performed with the programs dps/orpheus[8], AMIGene[9], glimmer/rbsfinder[10] and genemarks/genemark.hmm[11]. The predictions of the three programs were clustered and overlapping genes were removed according to their length and homology to proteins in public databases. If multiple predictions for the same gene did not agree on the gene start, the decision was either based on alignments with proteins in public databases or the most upstream gene start prediction was kept. The PEDANT software system[12] was used for genome sequence analysis and annotation. Translated CDSs longer than 180 nucleotides were searched using BLAST[13] and FASTA[14] against a non-redundant protein sequence database (cut-off e-value <=1e-04) and further analyzed by searching against various motif and domain libraries (see http://pedant.gsf.de/about.html for a detailed list of methods and databanks used). The comprehensive data collected automatically for each CDS were subsequently used as a basis for careful manual annotation. CDSs were assigned to functional categories according to the functional role catalogue FunCat[15]. Proteins with amino acid sequence homology to characterized proteins (either by knock-out/complementation experiments, protein expression, or by known 3D structure) were annotated as strongly similar to known proteins (> 40% amino acid sequence identity to characterized protein) or as similar to known proteins (> 20% amino acid sequence identity to characterized protein). CDSs with sequence homology to not yet functionally characterized proteins were classified as conserved hypothetical proteins (> 30% amino acid sequence identity to first blast hit) or hypothetical proteins (> 20% amino acid sequence identity to first blast hit). CDSs showing no significant homology to proteins in

public databases were annotated as unknown protein. tRNA genes were identified with the tRNAscan-SE program[16] and rRNA genes were located by homology. Genome synteny against other complete genomes was computed with LAGAN[17].

Phylogenetic analysis. Phylogenetic analysis of *K. stuttgartiensis* was based on concatenated datasets of amino acid and ribosomal RNA sequences. Of the bacterial genomes which were publicly available in mid-2004, those representing different species under the 97% 16S rRNA similarity definition were selected. This approach yielded 98 genomes, to which were added *K. stuttgartiensis* and *Gemmata obscuriglobus*, a planctomycete currently being sequenced at The Institute for Genomic Research (TIGR). For selected analyses, ribosomal protein and ribosomal RNA sequences were also obtained from five archaeal genomes. Amino acid sequences were obtained directly from genome annotations (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html) or via BLASTP searches. In total, sequences were extracted for 44 ribosomal proteins, 3 DNA-directed RNA polymerase subunits and 3 further proteins commonly used for bacterial phylogeny (**SOM Table 6**). Individual proteins were aligned in ClustalW[18] using the default settings (GAPOPEN=-12; GAPEXT=-2) and subsequently concatenated. Species for which sequences were missing or markedly truncated were excluded from the relevant single-gene analyses and also from the concatenated alignments. In no case did this result in removal of the only representative(s) of a given phylum. Distance, maximum parsimony and maximum likelihood methods were applied for each concatenated alignment using the programs FITCH, PROTPARS and MOLPHY, respectively, implemented in the ARB package[19]. Positional conservation filters of 30% were applied. Phylogenetic consensus trees were constructed using the approach outlined by Ludwig *et al.*[20], and distance- and parsimony-based bootstrapping (100 resamplings) was performed using PHYLIP 3.61[21]. Ribosomal RNA sequences (5S, 16S and 23S) were aligned in ARB using FastAligner, and then the alignment was refined manually. Aligned sequences were concatenated, and phylogenetic analysis performed in ARB using neighbour-joining, maximum parsimony and maximum likelihood algorithms. Positional conservation filters of 50% were used, while bootstrapping and consensus tree construction were as described for amino acids.

Rooting of ribosomal protein trees with members of another domain is often not attempted due to alignment difficulties and the risk of biases due to long-branch attraction[22]. Here, to test whether the current data set suggests deep-branching of the *Planctomycetes* within the *Bacteria* tree, we included ribosomal protein sequences from five archaea according to the

alignments of Vishwanath et al[23]. From this concatenated alignment conserved positions were selected (for details see legend of **SOM Figure 5**) and phylogenetic analysis were performed by using the programs FITCH, PROTPARS and MOLPHY implemented in the ARB package[19] as well as PHYML[24] for inferring the Maximum Likelihood-based phylogeny with the following evolutionary models: Dayhoff[25], JTT[26], MtRev[27], and WAG[28]. For each PHYML inference topology optimization was applied on a BIONJ[29] starting tree, the proportion of invariable sites and the gamma distribution parameter were estimated, and 16 substitution rate categories were applied. These calculations were performed on Quad AMD Opteron 2,4 GHz computing nodes, equipped with 8 GB shared main memory.

References to the SOM Materials and Methods

1. Sliekers, A. O., Third, K., Abma, W., Kuenen, J. G. & Jetten, M. S. M. CANON and Anammox in a gas-lift reactor. *FEMS Microbiol. Lett.* **218,** 339-344 (2003).
2. Schmid, M., Twachtmann, U., Klein, M., Strous, M., Juretschko, S., Jetten, M., Metzger, J.W., Schleifer, K.H. & Wagner, M. Molecular evidence for genus level diversity of bacteria capable of catalyzing anaerobic ammonium oxidation. *Syst. Appl. Microbiol*. **23**, 93-106 (2000).
3. Daims, H., Lücker, S. & Wagner, M. DAIME, a novel image analysis program for microbial ecology and biofilm research. *Environ. Microbiol.* **8**, 200-213 (2006).
4. Le Paslier, M. C., Pierce, R. J., Merlin, F., Hirai, H. *et al.* Construction and characterization of a *Schistosoma mansoni* bacterial artificial chromosome library. *Genomics* **65,** 87-94 (2000).
5. Strous, M., Fuerst, J. A., Kramer, E., Logemann, S. *et al.* Missing lithotroph identified as new planctomycete. *Nature* **400**, 446-449 (1999).
6. Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., Giegerich, R. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **22,** 4633-4642 (2001).
7. Achaz, G., Coissac, E., Viari, A. & Netter, P. Analysis of intrachromosomal duplications in yeast *Saccharomyces cerevisiae*: a possible model for their origin. *Mol. Biol. Evol.* **17,** 1268-1275 (2000).
8. Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* **26,** 2941-2947 (1998).
9. Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., Medigue, C. AMIGene: Annotation of

microbial genes. *Nucleic Acids Res.* **31**, 3723-3726 (2003).

10. Delcher, A. L., Harmon, D., Kasif, S., White, O. & S.L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27,** 4636-4641 (1999).
11. Besemer J., Lomsadze A. & Borodovsky M. GeneMarkS: a self-training method for predicition of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29,** 2607-2618 (2001).
12. Frishman, D., Albermann, K., Hani, J., Heumann, K. et al. Functional and structural genomics using PEDANT. *Bioinformatics* **17,** 44-57 (2001).
13. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. Et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389-2402 (1997).
14. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444-2448 (1988).
15. Ruepp, A., Zollner, A., Maier, D., Albermannm K, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* **32,** 5539-5545 (2004).
16. Lowe, T. M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955-964 (1997).
17. Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13,** 721-731 (2003).
18. Chenna, R., Sugawara, H., Koike, T., Lopez, *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31,** 3497-3500 (2003).
19. Ludwig, W., Strunk, O., Westram, R., Richter, et al. ARB: a software environment for sequence data. *Nucleic Acids Res.* **32,** 1363-1371 (2004).
20. Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N. *et al*. Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 554-568 (1998).
21. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle (2004).
22. Brochier, C., Forterre, P. & Gribaldo, S. 2005. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol. Biol.* **5**, 36 (2005).
23. Vishwanath, P., Favaretto, P., Hartman, H., Mohr, S.C. & Smith, T.F. Ribosomal protein-sequence block structure suggests complex prokaryotic evolution with implications for the

origin of eukaryotes. *Mol. Phylogenet. Evol.* **33**, 615-625 (2004).

24. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).
25. Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. A model of evolutionary change in proteins. In: Dayhoff, M. O. (ed.) Atlas of Protein Sequence Structur, Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington DC, pp. 345-352 (1978).
26. Jones, D.T., Taylor, W. R. & Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *CABIOS* **8**, 275-282 (1992).
27. Adachi, J. & Hasegawa, M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42**, 459-468 (1996).
28. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691-699 (2001).
29. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**, 685-695 (1997).

Legends to the figures of the "Supplementary online material"

SOM Fig. 1.

Maximum-likelihood tree displaying the16S rRNA gene diversity in the anammox enrichment culture from which the *K. stuttgartiensis* genome was reconstructed. In total, 90 partial 16S rRNA gene sequences (representing 29 operational taxonomic units, OTUs) were obtained from BAC, fosmid or shotgun clones during the course of the project. Partial 16S rRNA gene sequences retrieved from the enrichment were added to the maximum likelihood tree without changing the overall tree topology. The partial sequences were assigned to OTUs by analyzing the tree topology and by applying a 97% similarity threshold on the respective phylogenetically closest related, nearly full-length 16S rRNA gene sequence available in public databases. Lineages without shading represent those sequences of uncertain affiliation; numbers inside wedges indicate the number of sequences present within the wedge. Scale bar represents 10% estimated sequence divergence.

SOM Fig. 2.

Representation of the *K. stuttgartiensis* chromosome including annotated coding sequences and RNA genes. Very short features and gaps were enlarged to enhance visibility. Local nucleotide composition measures (outer rings) are indicated with pseudo-color assignments.

The correct assignment of every contig was independently confirmed by phylogenetic analysis of established phylogenetic markers and by comparison with shotgun sequences from *Brocadia anammoxidans*.

SOM Fig. 3.

Functional redundancy in catabolism and respiration in *K. stuttgartiensis* compared to other bacteria. Neu, *Nitrosomonas europaea*; Ppu, *Pseudomonas putida*; Rba, *Rhodopirellula baltica*; Kst, *Kuenenia stuttgartiensis*; Gsu, *Geobacter sulfurreducens*; Mma, *Methanosarcina mazei*; Cac, *Clostridium acetobutyricum*. * Complex I refers to NADH/formate:quinon oxidoreductase.

SOM Fig. 4.

Phylogenetic consensus tree based on concatenated 5S-16S-23S rRNA sequences, showing the phylogenetic positioning of *K. stuttgartiensis* within the *Bacteria*. Tree is rooted with archaeal sequences; note that in no cases were the *Planctomycetes* the deepest-branching bacterial phylum. Values at nodes represent distance- and maximum parsimony-based

bootstrap support, respectively. Scale bar represents 10% estimated sequence divergence.

SOM Fig. 5.

Distance (FITCH with the Dayhoff-PAM model), maximum parsimony (MP) and maximum likelihood (PHYML with the Dayhoff-PAM model) phylogenetic trees based on 26 concatenated ribosomal protein sequences (see **SOM Materials and Methods**), showing the phylogenetic positioning of *K. stuttgartiensis* within the *Bacteria*. Tree is rooted with archaeal sequences; note that in no cases were the *Planctomycetes* the deepest-branching bacterial phylum. Consistent results regarding the positioning of the *Planctomycetes* were also obtained if other models (JTT, MtRev and WAG) were used with PHYML (data not shown). 3178 alignment positions were included in the phylogenetic analyses. Distance- and parsimony-based bootstrap support is indicated as follows: open circles indicate >75% support; filled circles indicate >90% support. Scale bar represents 10% estimated sequence divergence for the FITCH tree. For the ML tree, branch lengths were computed according to a model of protein evolution based on the Dayhoff matrix, applying a discrete gamma model (gamma shape parameter: 1.18) with 16 rate categories and a proportion of invariant sites of 0.019. The deep branching of the *Mollicutes* observed in some analyses could be due to long branch attraction; thus we make no assertions about which is the deepest branching bacterial phylum, but do note that there is no indications at this stage of *Planctomycetes* filling this role. Future sequencing of genomes from other closely related organisms (such as *Verrucomicrobium spinosum* currently being sequenced at TIGR) should help to elucidate the precise phylogenetic positioning of the *Planctomycetes*.