**Supplementary Data for**

**Operator recognition by the ROK transcription factor family members, NagC and Mlc.**

Dominique Bréchemier-Baey[1], Lenin Domínguez-Ramírez[2] Jacques Oberto[3] and Jacqueline Plumbridge[1*]

**Table SI. Construction of mutant Mlc and NagC by 2-rounds-of-PCR oligonucleotide mutagenesis.**

| pXE plasmids | Template 1st PCR | Oligonucleotide | Template 2nd PCR |
|---|---|---|---|
| **Mlc constructs** | | | |
| $MN_{67-82}M$ ($MN_LM$) | pXEMlc$_{1-65}$Nag$_{67-406}$ | Nag82Mlc82 | pXENag82Mlc82 |
| $MN_{31-82}M$ ($MN_{HL}M$) | Mlc$_{1-30}$Nag$_{33-406}$ | Nag82Mlc82 | pXENag82Mlc82 |
| Mlc28 | pXEMlc wt | Mlc28 | pXEMlc wt |
| Mlc32 | pXEMlc wt | Mlc32 | pXEMlc wt |
| Mlc34 | pXEMlc wt | Mlc34 | pXEMlc wt |
| Mlc41 | pXEM$N_L$M | Nag97 | pXEM$N_L$M |
| Mlc43 | pXEM$N_L$M | Nag101 | pXEM$N_L$M |
| Mlc45 | pXEMlc$_{1-56}$Nag$_{59-406}$ | Mlc56Nag59 | pXEM$N_{31-82}$M |
| Mlc47 | pXEMlc$_{1-56}$Nag$_{59-406}$ | Nag102 | pXEM$N_{31-82}$M |
| Mlc48 | pXEM$N_{67-82}$M | Nag102 | pXEM$N_{31-82}$M |
| Mlc50 | pXEMlc wt | Mlc58Nag61 | pXEMlc45 |
| Mlc51 | pXEMlc wt | Mlc61Nag64 | pXEMlc45 |
| Mlc52 | pXEMlc wt | Mlc52 | pXEMlc wt |
| Mlc53c | pXEM$N_{67-82}$M | Mlc53 | pXEM$N_{67-82}$M |
| Mlc54 | pXEM$N_{67-82}$M | Mlc54 | pXEM$N_{67-82}$M |
| Mlc55 | pXEM$N_{67-82}$M | Mlc55 | pXEM$N_{67-82}$M |
| Mlc58 | pXEMlc45 | Nag97 | pXEMlc45 |
| Mlc59 | pXEMlc45 | Nag101 | pXEMlc45 |
| Mlc63 | pXEM$N_{67-82}$M | Mlc63 | pXEM$N_{67-82}$M |
| Mlc64 | pXEM$N_{67-82}$M | Mlc64 | pXEM$N_{67-82}$M |
| Mlc65 | pXEM$N_{67-82}$M | Mlc65 | pXEM$N_{67-82}$M |
| Mlc66 | pXEMlc wt | Mlc66 | pXEMlc wt |
| Mlc73 | pXEMlc45 | Mlc73 | pXEMlc45 |
| Mlc74 | pXEMlc45 | Mlc74 | pXEMlc45 |
| Mlc75 | pXEMlc45 | Mlc75 | pXEMlc45 |
| Mlc76 | pXEMlc45 | Mlc76 | pXEMlc45 |
| Mlc77 | pXEM$N_{31-82}$M | Nag97 | pXEM$N_{31-82}$M |
| Mlc78 | pXEM$N_{31-82}$M | Nag101 | pXEM$N_{31-82}$M |
| Mlc79 | pXEM$N_{31-82}$M | Nag102 | pXEM$N_{31-82}$M |
| Mlc80 | pXEMlc45 | Mlc80 | pXEMlc45 |
| Mlc81 | pXEMlc45 | Mlc81 | pXEMlc45 |
| Mlc82 | pXEMlc45 | Mlc82 | pXEMlc45 |
| Mlc83 | pXEMlc45 | Mlc83 | pXEMlc45 |
| Mlc84 | pXEMlc wt | Mlc84 | pXEMlc wt |
| Mlc85 | pXEMlc wt | Mlc85 | pXEMlc wt |
| Mlc92 | pXEMlc84 | Mlc91 | pXEMlc84 |
| Mlc93 | pXEMlc85 | Mlc91 | pXEMlc85 |
| Mlc110 | pXEMlc wt | Mlc110 | pXEMlc84 |
| Mlc111 | pXEMlc wt | Mlc111 | pXEMlc84 |
| Mlc112 | pXEMlc110 | Mlc54 | pXEMlc110 |
| | | | |

| **NagC constructs** | | | |
|---|---|---|---|
| $NM_{65-81}N$ ($NM_LN$) (1) | pXENagC wt | Nag66Mlc66 | $pXEMlc_{1-81}Nag_{83-406}$ |
| $NM_{33-81}N$ ($NM_{HL}N$) (1) | $pXENag_{1-32}Mlc_{31-406}$ | Mlc81Nag83 | $pXEMlc_{1-81}Nag_{83-406}$ |
| Nag44 | $pXEMlc_{1-58}Nag_{57-406}$ | Nag58Mlc57 | $pXENM_{33-81}N$ |
| Nag46 | $pXEMlc_{1-58}Nag_{57-406}$ | Mlc32 | $pXENM_{33-81}N$ |
| Nag49 | $pXENM_{65-81}N$ | Mlc32 | $pXENM_{65-81}N$ |
| Nag61 | $pXENM_{65-81}N$ | Mlc52 | $pXENM_{65-81}N$ |
| Nag62 | pXENag44 | Mlc52 | pXENag44 |
| Nag97 | pXENagC wt | Nag97 | pXENagC wt |
| Nag101 | pXENagC wt | Nag101 | pXENagC wt |
| Nag102 | pXENag97 | Nag102 | pXENagC wt |
| Nag103 | $pXENM_{33-81}N$ | Mlc34 | $pXENM_{33-81}N$ |
| Nag104 | $pXENM_{33-81}N$ | Mlc52 | $pXENM_{33-81}N$ |
| Nag105 | $pXENM_{33-81}N$ | Mlc32 | $pXENM_{33-81}N$ |
| Nag106 | $pXENM_{65-81}N$ | Mlc34 | $pXENM_{65-81}N$ |
| Nag107 | pXENag44 | Mlc34 | pXENag44 |
| | | | |

The pXE1 derived plasmids carrying the Mlc and NagC mutants and chimeras listed in column 1 (Constructs) were made by the 2-rounds-of-PCR method described previously (1,2) using the templates and mutagenic oligos indicated. The first round of PCR used XE1 and the mutagenic oligonucleotide with the 1st template plasmid. After purification, this PCR fragment was used as a megaprimer with RBP22 on the 2$^{nd}$ template plasmid. The sequences of the oligonucleotides are given in Table S2. For final sequences of the HTH and linker regions of the constructs see Figures 2, 3, 4 and Supporting data S2, S3, S4. All constructs were verified by sequencing the plasmids with RBP22 and Lac22. Alternative names of plasmids are from reference (1). $pXEMN_{67-82}M$ was previously called $pXEMN_LM$ , $pXEMN_{31-82}M$ was $pXEMN_{HL}M$, $pXENM_{33-81}N$ was $pXENM_{HL}N$ and $pXENM_{33-81}N$ was $pXENM_{HL}N$.

**Table S2: Oligonucleotides used**

| Oligonucleotide | Sequence |
|---|---|
| Nag66Mlc66 | GACGGCCACGGTTCCCCGCTTCTTTATCAACTTCTTTGATCAGCCCGCGT |
| Mlc65Nag67 | CGGCCCCCGGTGGAGGCCTGCTGGATTTCCAGCTCTTGCACCAGGTGTGC |
| Mlc81Nag83 | CGCCGATTGCGTGGAAATTGCGGGTTTCAACCACCAGCCCCACCGCCG |
| Nag82Mlc82 | GAGAAAGATAGTGCCAGGCTTCAGTTTCGGTGACGATGGAGATAGCGCGG |
| Mlc56Nag59 | ATCAACTTCTTTGATCAGCCCGCGTTCGAGCATCTCACGGACAATTTTAG |
| Nag58Mlc57 | TTCCAGCTCTTGCACCAGGTGTGCTTCGATAAGCTGACGCGTAATTTTGG |
| Mlc61Nag64 | GGTGGAGGCCTGCTGATCAACTTCTTGCACCAGGTGTGCTTCGAG |
| Mlc28 | CCAGGCTTCAGTTTCAACCACCAGGGAGATAGCGCGGCGGCCCCCGTTCC CCGCTTCTTTGATTTCCAG |
| Mlc32 | AACCACCAGCCCCACCGCGCGACGGCCCCCGTTCCCCGCTTCTTTGAT |
| Mlc34 | CACCAGCCCCACCGCGCGACGGCCACGGTTCCC |
| Mlc52 | CACCGCCGGACGGCCACCGTTCCCCGCTTC |
| Mlc53 | CCCGGTGGAGGCCTGCTGTTCCAGCTCTTGCACCAG |
| Mlc54 | GATTTCCAGCTCTTTCACCAGGTGTGCTTCGAG |
| Mlc55 | CCGGTGGAGGCCTGCTGTTCCAGCTCTTTCACCAGGTGTGCTTCGAGC |
| Mlc63 | CCCGGTGGAGGCCTGCTGATCCACCTCTTTCACCAGGTGTGCTTCGAG |
| Mlc64 | CCCGGTGGAGGCCTGCTGTTCCACCTCTTTCACCAGGTGTGCTTCGAG |
| Mlc65 | CCCGGTGGAGGCCTGCTGATCCAGCTCTTTCACCAGGTGTGCTTCGAG |
| Mlc66 | GTTCCCCGCTTCTTTATCAACTTCTTTGATCAGGTGTGCTTCGAGCAT |
| Mlc73 | GCCCCCGGTGGAGGCTTCTTTATCAACTTCTTTGATCAGCCC |
| Mlc74 | GATAGCGCGGCGGCCCCCGTTCCCGGCCTGCTGATCAACTTC |
| Mlc75 | TTCAGTTTCGGTGACGATCCCCACAGCGCGGCGGCCCCCGGT |
| Mlc76 | GTGCCAGGCTTCAGTTTCAACCACCAGGGAGATAGCGCGGCGGCC |
| Mlc80 | GCGGCGGCCCCCGGTCCCGGCCTGCTGATCAACTTC |
| Mlc81 | AGCGCGGCGGCCCCCGTTGGAGGCCTGCTGATCAAC |
| Mlc82 | GGCTTCAGTTTCGGTGACCAGGGAGATAGCGCGGCGGCC |
| Mlc83 | GTGCCAGGCTTCAGTTTCAACGACGATGGAGATAGCGCG |
| Mlc84 | CACCGCCGGACGGCCCCCGGTGGACGCTTCTTTTTCCAGCTCTTTCACCA GGTGTGCTTC |
| Mlc85 | CACCGCCGGACGGCCCCCGGTGGACGCTTCTTTTTCCA |
| Mlc91 | TTCAGTTTAACCACGATCCCCACCGCCGGACG |
| Nag97* | GATAGCGCGGCGGCCCCGGGTGGAGGCCTGCTG |
| Nag101 | GACGATGGAGATAGCGGGGCGGCCCCCGGTGGA |
| Nag102 | GGTGACGATGGAGATAGCCGGGCGGCCACGGGTGGAGGCCTGCTGATC |
| Mlc110 | CACCGCCGGACGGCCCCCGGTGGACGCTTCTTTGATTTCCAG |
| Mlc111 | ACCGCCGGACGGCCCCCGGTGGACGCTTCTTTTTCCAGCTCTTGCACCAG |
| Nag115 | GCCACGGGTGGAGGCTTCTTTATCAACTTCTTTGATCAGCCC |
| XE1 | GGTTGGCTCCAATTATTTGTATATTC |
| RBP22 | CCGAAAAGTGCCACCTGACGTC |
| | |

Oligonucleotides XE1 and RBP22 hybridise to plasmid sequences around the insertion site sequences in pXE1 (Fig. S1). All other oligonucleotides were used in the first round PCR of mutagenesis as described in Table S1.

* hybridises with bottom strand

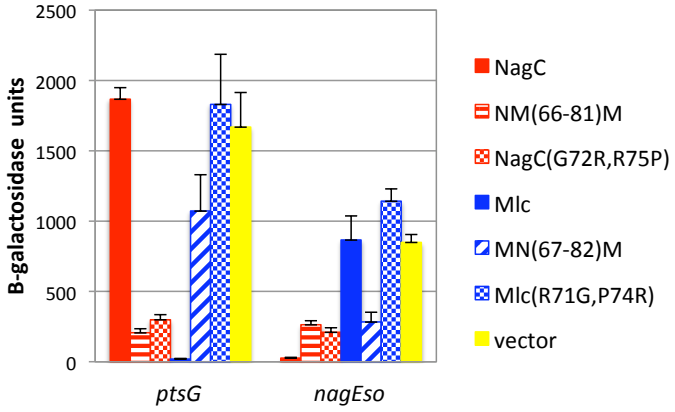
**Table S3. The 35 proteins used to extract ROK proteins from genomes**

| | GI | Definition from Genebank | Alternative nomenclature |
|---|---|---|---|
| Thermotoga maritima MSB8 | 15642807 | XylR family transcriptional regulator | TM0032 BglR[1] |
| Thermotoga maritima MSB8 | 15642885 | XylR family transcriptional regulator | TM0110 XylR [1] |
| Thermotoga maritima MSB8 | 15643159 | XylR family transcriptional regulator | TM0393 TreR [1] |
| Thermotoga maritima MSB8 | 15643177 | XylR family transcriptional regulator | TM0411 IolR [1] |
| Thermotoga maritima MSB8 | 15643571 | XylR family transcriptional regulator | TM0808 ChiR [1] |
| Thermotoga maritima MSB8 | 15643980 | XylR family transcriptional regulator | TM1224 ManR [1] |
| Thermotoga maritima MSB8 | 15644590 | ROK family protein, partial | TM1847 GluR [1] |
| Escherichia coli str. K-12 substr. MG1655 | 16128652 | NagC | |
| Escherichia coli str. K-12 substr. MG1655 | 16129552 | Mlc | |
| Escherichia coli str. K-12 substr. MG1655 | 90111457 | YphH | |
| Streptomyces coelicolor A3(2) | 21218825 | transcriptional repressor protein | SCO02754 |
| Streptomyces coelicolor A3(2) | 21219259 | transcriptional repressor | SCO0734 |
| Streptomyces coelicolor A3(2) | 21219316 | transcriptional regulator | SCO0794 |
| Streptomyces coelicolor A3(2) | 21219555 | regulatory protein | SCO1039 |
| Streptomyces coelicolor A3(2) | 21219575 | transcriptional repressor | SCO1060 |
| Streptomyces coelicolor A3(2) | 21219769 | transcriptional regulator | SCO1261 |
| Streptomyces coelicolor A3(2) | 21219949 | transcriptional regulator | SCO1447 |
| Streptomyces coelicolor A3(2) | 21221114 | transcriptional regulator | SCO2657 CsnR [2] |
| Streptomyces coelicolor A3(2) | 21221296 | transcriptional regulator | SCO2846 |
| Streptomyces coelicolor A3(2) | 21224343 | transcriptional repressor protein | SCO6008 NgcR [3] |
| Streptomyces coelicolor A3(2) | 21224900 | transcriptional regulator | SCO6600 |
| Streptomyces coelicolor A3(2) | 21225754 | transcriptional regulator | SCO7486 |
| Streptomyces coelicolor A3(2) | 21225808 | CRP family transcriptional regulator | SCO7543 |
| Streptomyces coelicolor A3(2) | 32141306 | ROK family protein, partial | SCO6566 |
| Mycobacterium smegmatis str. MC2 155 | 118467726 | NagC regulator | MSMEG_0316 |
| Mycobacterium smegmatis str. MC2 155 | 118469208 | xylose repressor | MSMEG_3313 |
| Mycobacterium smegmatis str. MC2 155 | 118469375 | ROK family protein | MSMEG_0932 |
| Mycobacterium smegmatis str. MC2 155 | 118473203 | xylose repressor, ROK-family protein transcriptional regulator | MSMEG_6022 |
| Bifidobacterium longum DJO10A | 189439069 | NagC family transcriptional regulator | BLD_0206 |
| Bifidobacterium longum DJO10A | 189439071 | NagC family transcriptional regulator | BLD_0208 |
| Bifidobacterium longum DJO10A | 189440346 | NagC family transcriptional regulator | BLD_1485 |
| Bifidobacterium longum DJO10A | 189440565 | NagC family transcriptional regulator | BLD_1703 |
| Bifidobacterium longum DJO10A | 189440720 | NagC family transcriptional regulator | BLD_1858 |
| Bifidobacterium longum DJO10A | 189440749 | NagC family transcriptional regulator | BLD_1887 |
| Bacillus subtilis subsp. subtilis str. 168 | 255767416 | XylR | |

[1] Putative identification from (3)
[2] Identified in *S. lividans* (4)
[3] Identified in *S. olivaceoviridis* (5)

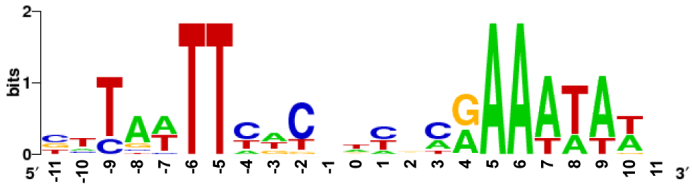
A wide range of ROK protein sequences was used to analyse each of a set of single genus genomes for ROK homologues. The initial screening set comprised Mlc, NagC and the third ROK family repressor, YphH, of unknown function in *E. coli* and the XylR protein of *B. subtilis*. Using just these proteins as screening set pulled out only five of the additional clusters shown in Figures 5 and S2 and failed to pull out some other known ROK repressor sequences such as the 7 ROK repressor in *Thermotoga maritima* identified by Bioinformatic analysis (3). ROK proteins are also known in *Streptomyces* sequences and preliminary screening detected 12 ROK repressor proteins in *S. coelicolor*. *Bifidobacterium longum,* which also has diverse sugar utilisation operons, was found to have 6 ROK repressor homologues and *Mycobacterium smegmatis* 4 ROK repressors. We included all the ROK proteins identified above to constitute the seeding set of 35 proteins.

**Figure S1A. Effect of mutations in the linker motif GGRR and RGRP of NagC and Mlc on repression of *ptsG* and *nagEso*.**



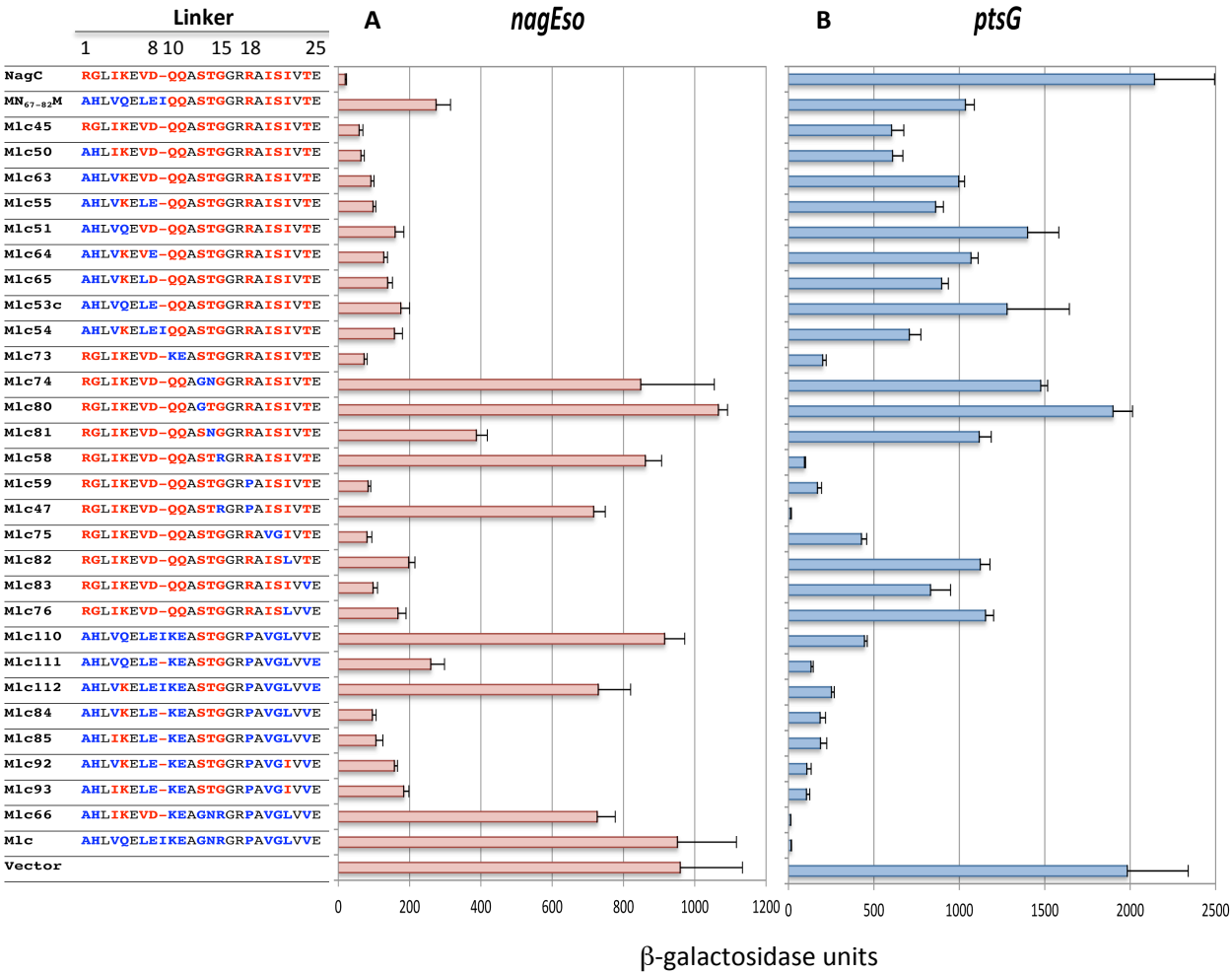

Repression by NagC and its derivatives (shown in red or red shadings) and Mlc and its derivatives (shown in blue or blue shading) compared to the control vector (yellow) was measured on a ptsG-lacZ fusion (a Mlc-target) and *nagEso-lacZ* (a NagC target). Data are taken from (1,2). NM(66-81)N was previously called $NM_LN$ and MN(67-82) called $MN_LM$. NagC(G72R,R75P) is called Nag102 and Mlc(R71G,P74R) is called Mlc32 in subsequent figures. The mutations in positions NagC(G72R,R75P) (Nag102) and Mlc(R71G,P74R)(Mlc32) correspond to linker positions 15 and 18 in the numbering of Fig. 1A. These changes convert the sequence of GGRR found in the linker of NagC to RGRP found in Mlc and vice versa.

**Figure S1B. Sequence logo derived from all 16 NagC sites**



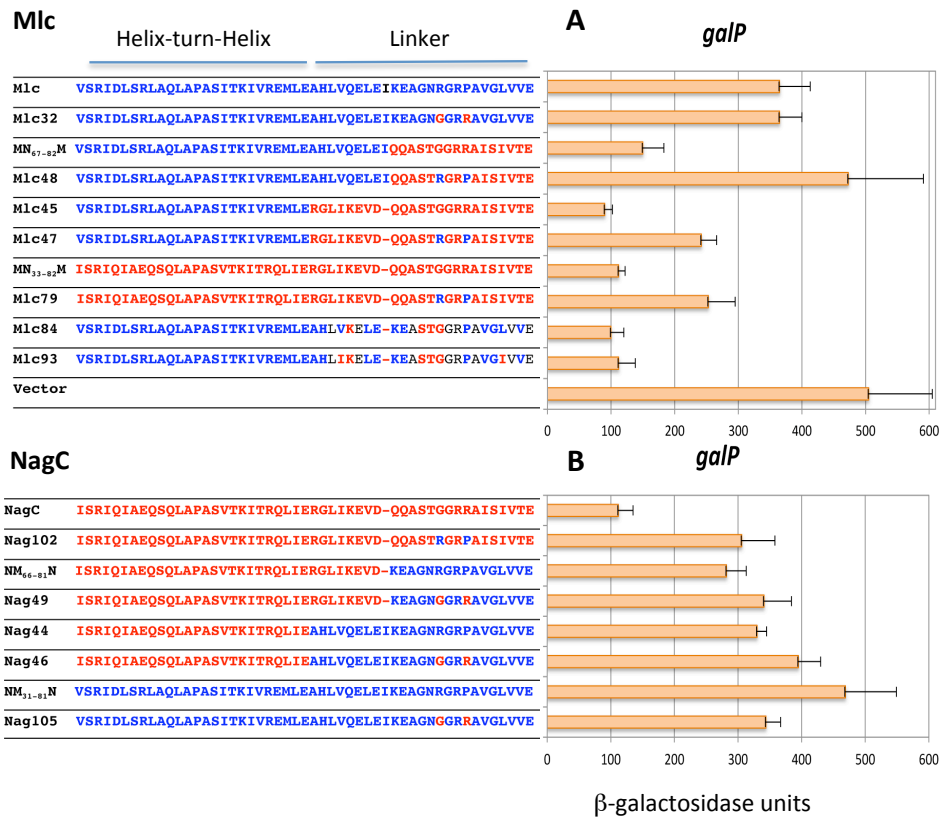

Note that NagC usually regulates genes by binding to two sites, of which one site has higher affinity. The higher affinity sites generally have a C or G at +/-11. whereas the lower affinity sites often do not. When all known NagC sites are included in the logo, there is no consensus for the bases at positions +/-11. The logo for higher affinity sites is shown in Figure1B. Sequences of operator sites are given in reference (1).

**Figure S2. Identification of amino acid replacements required to convert Mlc into a NagC-type repressor.**



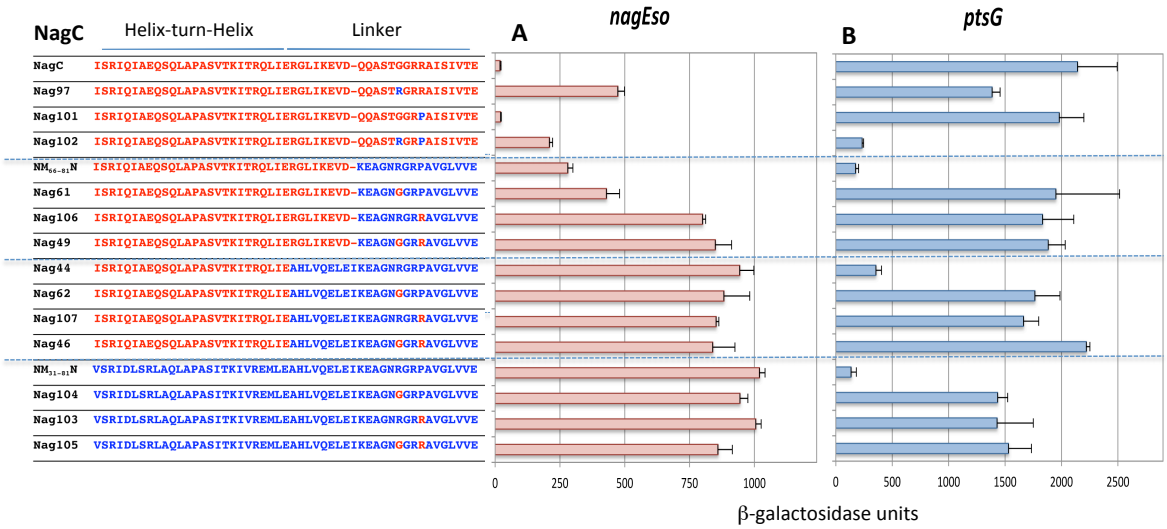

The sequences of the linker in NagC, Mlc and in the Mlc derived proteins are shown. Linker positions are numbered as in Figure 1A. Amino acids specific to NagC are shown in red, those specific to Mlc in blue and those identical in the two linkers in black. All the proteins are expressed from the single copy pXE plasmid. The ability of these Mlc derivatives to repress *nagE*so (A) and *ptsG-lacZ* (B) fusions is shown, compared to the control plasmid vector without insert. Activities are the mean of at least two (and generally more) independent cultures with standard deviation.

**Figure S3. Effect of mutations in the linker sequences of Mlc and NagC on the expression of *galP-lacZ* expression.**




The sequences of the HTH and linker regions of Mlc and NagC are shown and that of the hybrid proteins derived from them. Blocks of amino acids derived from NagC are shown in red, while those from Mlc are shown in blue. The rest (body) of the protein is derived from Mlc (A) or NagC (B). The effect of Mlc and its derivatives (A) or NagC and its derivatives (C, D) on expression of *galP-lacZ* is shown. Activities are the mean of at least two (and generally more) independent cultures with standard deviation. Note that the pattern of repression of *galP* is identical to that of *nagE*so (Figures 2, 4, S2 and S4)

**Figure S4. Effect of exchanging G and R at position 15 and R and P at position 18 of the linkers in NagC-derived constructs on *nagE*so-*lacZ* and *ptsG-lacZ* expression.**




Sequences of the HTH and linker region inserts in various NagC derivatives are given. Blocks of amino acids derived from NagC are in red, while those from Mlc are in blue. The ability of NagC and its derivatives to repress *nagEso-lacZ* (A) and *ptsG-lacZ* (B) is shown. Activities are the mean of at least two (and generally more) independent cultures with standard deviation.

To note:

1) Only NagC derivatives with both R15 and P18 repress *ptsG*.

2) The NagC derivatives with the Mlc linker do not repress *nagE*so or repress badly because they lack the secondary determinants (K5,Δ9,S13,T14).

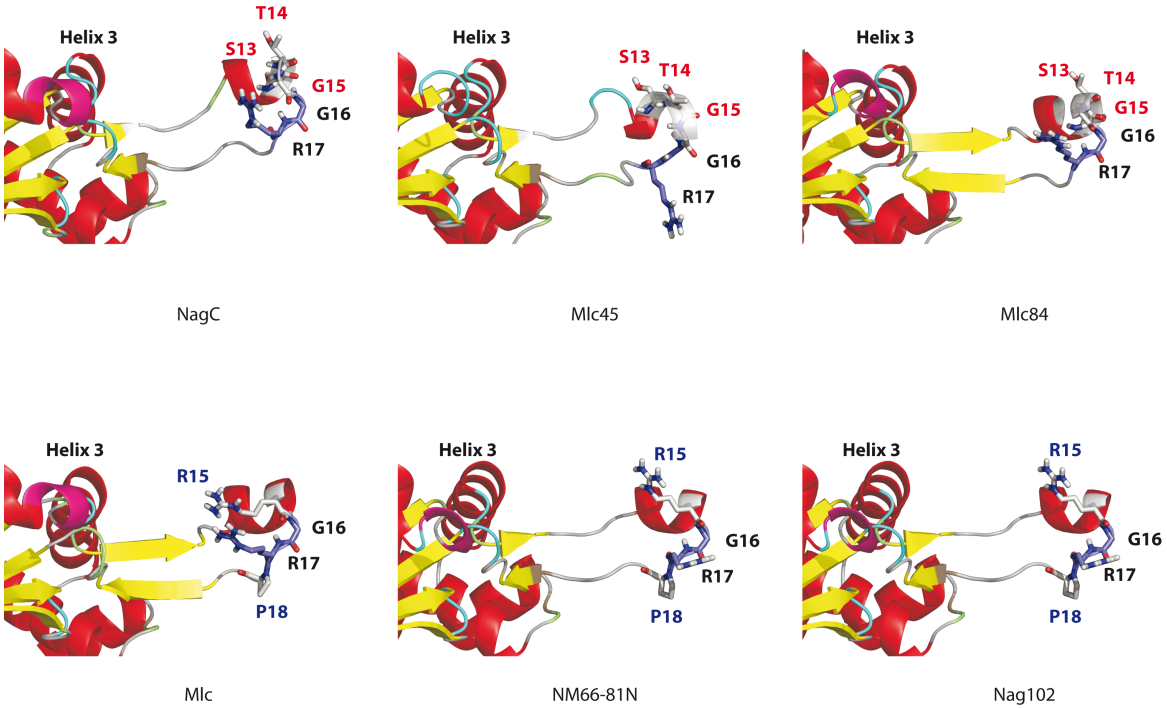
3) Within wild-type NagC loss of G15 does not lead to complete loss of repression, presumably because the secondary determinants and the HTH domain partially compensate.

4) Rather surprisingly, Nag102 (with two changes to the Mlc determinants R15, P18) represses *nagE*so better than Nag97 (with just R15). This confirms that R18 is not implicated in NagC-type recognition and could reflect subtle differences in the conformation of the amino acids at the apex of the linker so that RGRP in both Nag102 and $NM_{66-81}N$ is preferable to RGRR in Nag97 and Nag61.

5) Also unexpectedly, $NM_{66-81}N$ with the Mlc-specific R15,P18 represses *nagE*so better than Nag61, 106, and 49 where the R15 and/or P18 are changed to the NagC-specific G and R. This suggests that other factors could be influencing the conformation of the linker.

6) It can also be noted that all the constructs that have G13,N14,G15 in the linker (Nag61,49,62,46,104,105 (this figure), Mlc74 (Figure S2), Mlc32,52 (Figure 4) are essentially incapable of repressing either *ptsG* or *nagE*so. Possibly the G13,N14,G15 sequence produces an unacceptable configuration in the linker.

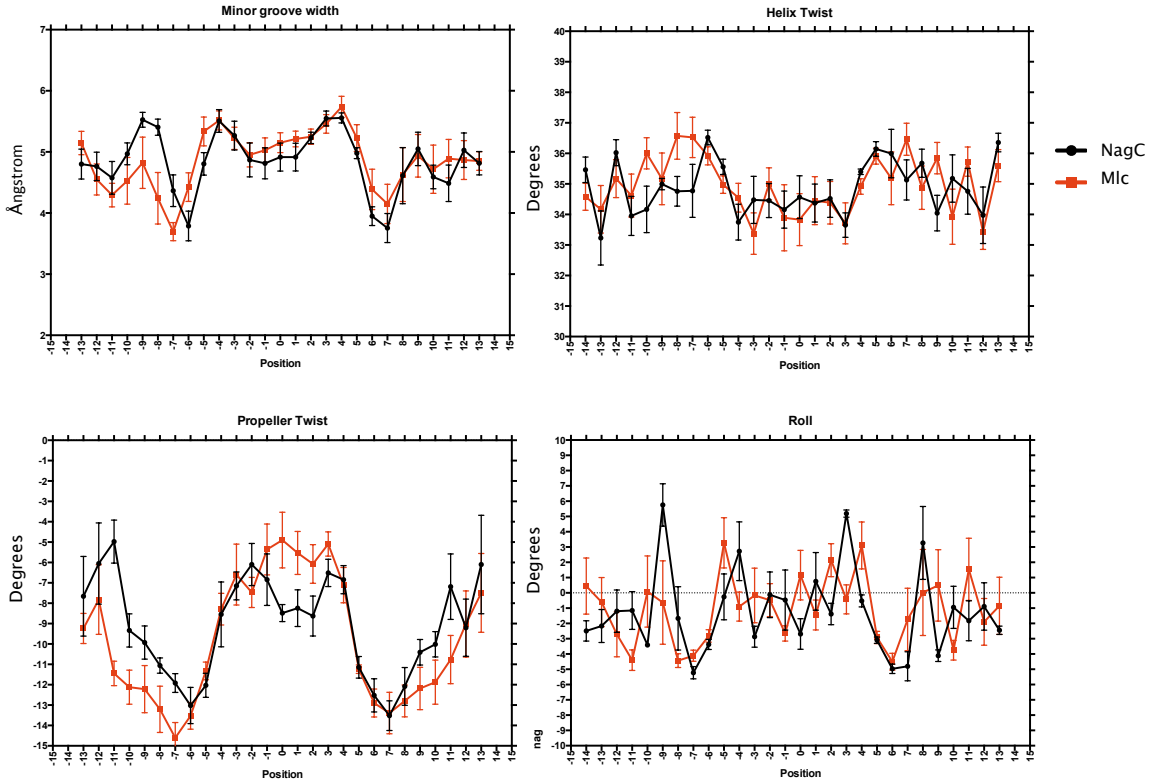
**Figure S5. Structures of the modelled linker motif in NagC, Mlc and derivatives with changed specificity.**




VC2007 structure (PDBID 1Z05) was taken from the RCSB database (6) and its stereochemical quality was checked with Molprobity (7). This structure was used as a seed for the reconstruction of the missing region of Mlc (PDBID 1Z6R). NagC and Mlc's derivatives were constructed using Rosetta via its web interface (8) and energy minimized with AMBER (9) before use. Visualization was done using PyMol. Secondary structure assignment was performed with DSSP (10) and represented with the DSSP and Stride Pymol plugin (http://www.biotec.tu-dresden.de/~hongboz/dssp_pymol/dssp_pymol.html).

The structures of the DNA binding domains, as predicted by Rosetta, show the linker sequence in a long finger-like projection in all proteins. Alpha helices of the HTH domain and the short alpha helix at the apex of the linker are shown in red. The beta sheets at the base of the linker projection are shown as yellow arrows. Random coils are in white, turns in light green and beta bends in blue. Models for three proteins capable of NagC-type repression, NagC, Mlc45 and Mlc84 (top) and three capable of Mlc-like repression, Mlc, $NM_{66\text{-}81}N$, Nag102 (bottom) are shown. The amino acids identified as required for NagC-type repression (K5, S13, T14, G15) and Mlc-like repression (R15, P18) are shown in stick form. The G16, R17 motif present in both NagC and Mlc and identified as conserved in all ROK family repressors is also indicated. The short alpha helix is found at a similar position in all the models. The amino acids required for Mlc-like repression seem to project down and towards the viewer (bottom row), while those required for NagC-type repression (top row) are at the C-terminal extremity of the small alpha helix and seem to project upward and away from the viewer. Arg15 in the Mlc-like structures on the other hand appears to be directed towards helix 3 of the HTH. We stress that these models are essentially based on that of VC2007 (1Z05), where the linker was visible in the crystal structure. Although the presence of Pro18 should mean that the apex of the Mlc-like linkers are stiffer than in the NagC-like, this whole region in all proteins is likely to be flexible in solution and their structures can be expected to be significantly modified by contact with the DNA.

**Figure S6. DNAshape predictions of Mlc and NagC operators.**




DNA structural features of Mlc and NagC operator sites were predicted using the DNAshape program (11). The 6 known Mlc operators (*ptsG1, ptsG2, ptsH, man1, mlc, malT*) and 6 of the high-affinity NagC sites (*nagEso, nagB, man2, fim2, glm2, galP*) were analysed (see (1) for operator sequences) including 4 bp flanking sequences on either side of the 23 bp consensus. Operator sequences are numbered about the centre of symmetry of the quasi-palindrome as in Fig. 1B. The 6 sites were processed separately and the average calculated with standard deviation. NagC operators are shown in black and Mlc operators in red. The minor groove is predicted to be at a minimun at positions +/-5 to 8 for both Mlc and NagC sites but to be wider and not significantly different at positions +/-11.

**Figure S7. Phylogeny and evolution of ROK proteins and their linker motif**

The 35 seed proteins (Table S3) were used to collect homologues from a representative set of complete genomes, which were then were ranked in a phylogenetic tree as described in Material and Methods. The scale bar corresponds to 0.01 amino acid substitutions per site. Throughout we have only considered the repressor class of ROK proteins with an N-terminal HTH domain and about 400 amino acids in length. The kinase class (with about 300 amino acids) are also well represented in these genomes but were not used for the extraction or considered in the final tree. Very few kinases were extracted using the repressors as bait.

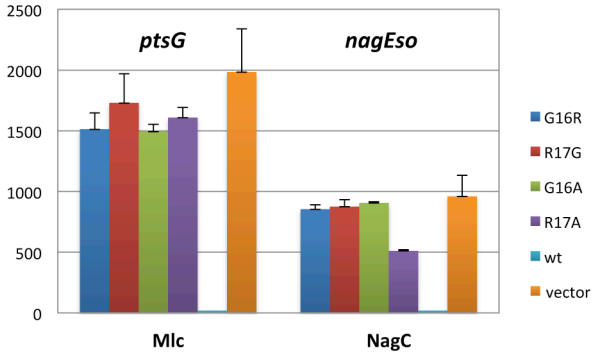
Clusters were chosen by eye with the aim of including a reasonable number of proteins per group to enable ClustalW alignment and logo generation. Proteins within each cluster were aligned with ClustalW and sequences corresponding to the central part of the linker (12 amino acids) and the HTH (24 amino acids Fig. 1A) of NagC. Proteins are noted by species name and GI number. The 35 seeding set of ROK proteins used as bait for protein extraction are marked by red dots. The Mlc and NagC protein clusters form two adjacent and closely evolved genomes presumably reflecting their occurrence in the multiple Enterobacteriales in the genome sequence databases. The highly conserved logos for both the HTH and linker demonstrate the close relationship of proteins within these groups. The YphH cluster is very small and the protein is found in just 4 species not close relatives of *E. coli*. For example, YphH is missing in *Salmonella* and *Shigella*. The XylR protein, which has been studied in *Bacillus* species (12-16) is grouped with proteins from a wide range of Firmicutes. The logos for both HTH and linker are less conserved.

The other seeding proteins are distributed in the different clusters, although clusters C2 and C3 are without a seeding protein. On the other hand cluster C1 includes 6 of the 7 *Thermotoga maritima* seed proteins, confirming that they evolved from lineage-specific gene duplication (3). The conservation of a GR in nearly all of these linker sequences was striking, with more or less of the surrounding amino acids significantly represented. Several clusters (C2 C8, C9) had the GGR of NagC; others had the GRP of Mlc (C9, C12, C13 C14). Conservation of amino acids within the HTH is more restricted, with the hydrophobic amino acids characteristic of HTH motif predominating (17-19), particularly R in position 3, L or I in position 6 and A in position 7 within the second helix (numbering as in the logos); L in position 12 of the turn and I/L/V in position 15 in the third (recognition helix). The very strongly conserved L at position 24 in all clusters is outside the standard HTH motif and could be characteristic of the ROK family.

The solution we present in here is not a unique solution to the data but is offered as a representative solution for the range of proteins studied. Bootstrapping methods to refine the tree are not practical for such a large data set. Similar trees with fewer clusters were generated if, for example, the *M. smegmatis* proteins were removed from the tree. Including more distantly related ROK proteins in the seeding set might enlarge the tree. The selection of 35 seeding proteins is, in some ways arbitrary, but consisted of genomes from diverse bacteria with known wide carbohydrate utilisation potential and several representative ROK proteins.

The tree should be viewed by zooming in on different regions of the pdf page (P. 14 of this Supporting data file)

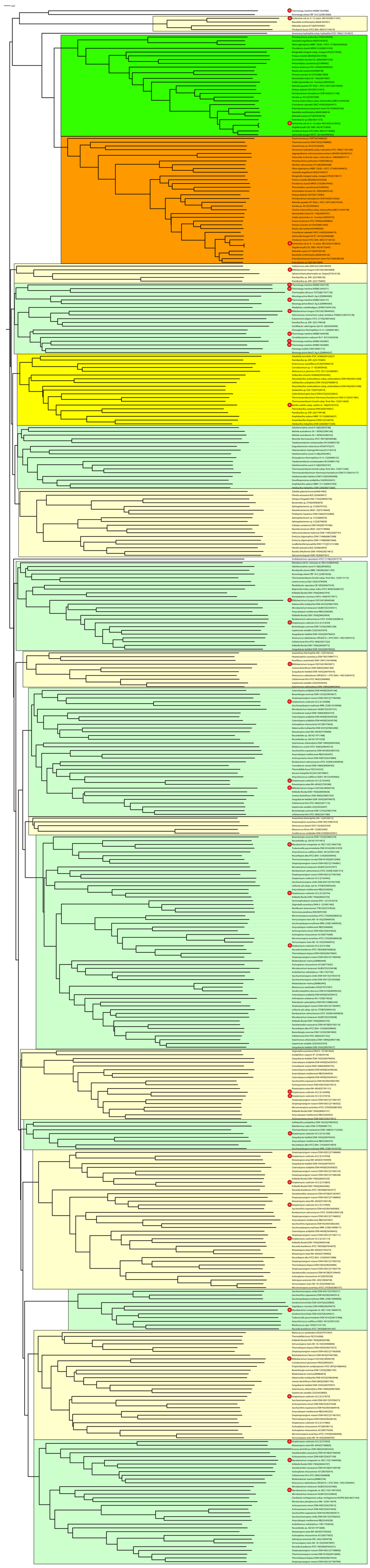
**Figure S8. Effect of mutations in linker positions G16, R17 of Mlc and NagC on repression of *ptsG* and *nagE*so.**




Mlc and NagC carrying the mutations indicated in the conserved GR motif in the centre of ROK family repressors (linker positions G16,R17) were tested for repression of *ptsG-lacZ* and *nagE*so-l*acZ* fusions and their activity compared to the the wild-type Mlc and NagC proteins and the vector without insert. Activities are the mean with standard deviation of two independent cultures. All mutations in G16 or R17 of the linker produced complete or almost complete derepression of their target fusion, showing the importance of the linker for DNA binding per se and not just target recognition. R17A in NagC exhibits about 50% residual repression.

**References for Supplementary Data**


1. Bréchemier-Baey, D., Domínguez-Ramírez, L. and Plumbridge, J. (2012) The linker sequence, joining the DNA-binding domain of the homologous transcription factors, Mlc and NagC, to the rest of the protein, determines the specificity of their DNA target recognition in *Escherichia coli*. *Mol. Microbiol.*, **85**, 1007-1019.
2. Pennetier, C., Domínguez-Ramírez, L. and Plumbridge, J. (2008) Different regions of Mlc and NagC, homologous transcriptional repressors controlling expression of the glucose and *N*-acetylglucosamine phosphotransferase systems in *Escherichia coli*, are required for inducer signal recognition. *Mol. Microbiol.*, **67**, 364-377.
3. Kazanov, M.D., Li, X., Gelfand, M.S., Osterman, A.L. and Rodionov, D.A. (2013) Functional diversification of ROK-family transcriptional regulators of sugar catabolism in the Thermotogae phylum. *Nucleic Acids Res.*, **41**, 790-803.
4. Dubeau, M.P., Poulin-Laprade, D., Ghinet, M.G. and Brzezinski, R. (2011) Properties of CsnR, the transcriptional repressor of the chitosanase gene, *csnA*, of *Streptomyces lividans*. *J. Bacteriol.*, **193**, 2441-2450.
5. Xiao, X., Wang, F., Saito, A., Majka, J., Schlosser, A. and Schrempf, H. (2002) The novel *Streptomyces olivaceoviridis* ABC transporter Ngc mediates uptake of N-acetylglucosamine and N,N'-diacetylchitobiose. *Mol. Genet. Genomics : MGG*, **267**, 429-439.
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235-242.
7. Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallograph. D, Biol. Crystallograph.*, **66**, 12-21.
8. Song, Y., DiMaio, F., Wang, R.Y., Kim, D., Miles, C., Brunette, T., Thompson, J. and Baker, D. (2013) High-resolution comparative modeling with RosettaCM. *Structure*, **21**, 1735-1742.
9. Yang, L., Tan, C.H., Hsieh, M.J., Wang, J., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P.A. and Luo, R. (2006) New-generation amber united-atom force field. *J. Phys. Chem. B*, **110**, 13166-13176.
10. Frishman, D. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566-579.
11. Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R. and Rohs, R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56-62.
12. Dahl, M.K., Degenkolb, J. and Hillen, W. (1994) Transcription of the *xyl* operon is controlled in *Bacillus subtilis* by tandem overlapping operators spaced by four base-pairs. *J. Mol. Biol.*, **243**, 413-424.
13. Scheler, A. and Hillen, W. (1994) Regulation of xylose utilization in *Bacillus licheniformis*: Xyl repressor-*xyl*-operator interactions studied by DNA modification protection and interferance. *Mol. Microbiol.*, **13**, 505-512.
14. Sizemore, C., Wieland, B., Götz, F. and Hillen, W. (1992) Regulation of *staphylococcus xylosus* xylose utilization genes at the molecular level. *J. Bacteriol.*, **174**, 3042-3048.
15. Lokman, C., van Santen, P., Verdoes, J.C., Krüse, J., Leer, R.J., Posno, M. and Pouwels, P.H. (1991) Organisation and characterisation of the three genes involved in D-xylose catabolism in *Lactobacillus pentosus. Mol. Gen. Genet.*, **230**, 161-169.
16. Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2001) Transcriptional regulation of pentose utilisation systems in the *Bacillus/Clostridium* group of bacteria. *FEMS Microbiol. Lett.*, **205**, 305-314.
17. Dodd, I.B. and Egan, J.B. (1990) Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res.*, **18**, 5019-5026.
18. Zhang, R.G., Joachimiak, A., Lawson, C.L., Schevitz, R.W., Otwinowski, Z. and Sigler, P.B. (1987) The crystal structure of *trp* aporepressor at 1.8 A shows how binding tryptophan enhances DNA affinity. *Nature*, **327**, 591-597.
19. Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M. and Pabo, C.O. (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature*, **298**, 447-451.



**YphH**

**Mlc**

**NagC**

**C0**

**C1**

**XylR**

**C2**

**C3**

**C4**

**C5**

**C6**

**C7**

**C8**

**C9**

**C10**

**C11**

**C12**

**C13**

**C14**