Cell Host & Microbe, *Volume 4*

Supplemental Data

Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont

Eric C. Martens, Herbert C. Chiang, and Jeffrey I. Gordon

Supplemental Results and Discussion

In vivo specific PULs

A number of the PULs depicted in **Fig. 2** exhibited induction only during growth *in vivo* in the mouse gut, suggesting that the glycans to which these systems respond were not present in sufficient quantities in our PMG preparation, or were not present in porcine stomach compared to the distal intestine of the mouse. Of the 15 PUL operons that we categorized into Group 3 ('*in vivo*-specific'), two (*BT1619-20* and *BT0865-67*) show particularly robust induction *in vivo* (columns A and '*in vivo* max' in **Fig. 2**). The observation that the *BT1619-20* operon is controlled by an ECF-σ regulator and its modest responsiveness during the secondary phase of growth in PMG (**Fig. 2**), suggest that it too may be a mucin *O*-glycan PUL, although we do not have conclusive *in vitro* data to support this idea. The *BT0865-67* system (**Fig. S9**) is noteworthy because it is highly induced *in vivo* but is not associated with any conspicuous regulator. This system, and the similarly small *BT0317-19* system, is obviously subject to transcriptional activation, although the mechanism appears to be different than the ones observed for other PULs characterized in this study. Transcriptional cross-activation from other host glycan-responsive regulatory systems could account for the induction of these loci.

Beyond the three most highly induced Group 3 PULs (*BT1619-20, BT0865-67* and *BT2392-95*), the remaining systems show substantially lower *in vivo* induction. With the exception of *BT3854-62* and *BT1280-85,* which are responsive *in vitro* to α-mannan and *O*-

glycan fractions respectively, it is difficult to predict their *in vivo* target substrates. Note that the presence of the Sus system, which has a known specificity for starch-like glycans, among these Group 3 loci suggests that some residual dietary glycans may have been present in small quantities in these mice: *e.g*., 17d old suckling gnotobiotic animals were co-housed with adult females who were consuming a plant glycan-rich chow diet.

N*-glycan foraging *in vivo

Growth on α-mannan, a proxy for the α-mannosidic linkages contained in host *N*-glycans, stimulated expression of genes encompassed within two Group 2 PULs (*BT2615-33* and *BT3773-92*) and one Group 3 PUL (*BT3853-62*) (column D in **Fig. 2** and **Fig. S8**). The fact that all three of these systems are expressed *in vivo* in mice fed a diet lacking exogenous α-mannan indicates that these PULs are deployed in the mouse gut in response to *N*-glycans. However, these systems are only modestly induced compared to PULs showing the highest activity *in vivo*, suggesting that *N*-glycans are not the most abundant host glycans foraged in the mouse cecum (see column labeled '*in vivo* max' in **Fig. 2** and **Fig. S8**). In contrast to the two GAG-utilization loci described above, each of the two Group 2 PULs (expressed both *in vivo* and *in vitro* in MM-PMG) showed peak expression in the second phase of PMG growth (column B in **Fig. 2)**, indicating that *N*-glycans are not given similar metabolic priority by *B. thetaiotaomicron* as GAGs.

The genetic architecture of the three α-mannan responsive PULs is reminiscent of the GAG utilization loci described above (compare **Figs. 3** and **S8**). Two of these loci have associated HTCS regulators, suggesting that they operate via the same general regulatory mechanism observed for GAGs. The third locus lacks a conspicuous HTCS regulator and instead has a gene encoding a hypothetical protein (*BT3853*), which occupies a similar position just upstream of the *susC* homolog. Sequence analysis of *BT3853* revealed that it encodes a protein with an N-terminal signal peptide, a single trans-membrane segment and a predicted C-terminal

DNA-binding domain belonging to the OmpR/SARP family (**Fig. S10**). This sequence architecture yields an inner membrane-spanning protein with similar topology to the HTCS sensors (Sonnenburg et al., 2006) and to SusR, suggesting that BT3853 represents a third class of membrane-spanning sensor/regulators that activate PUL genes in response to glycans.

Like the GAG utilization PULs, all three α-mannan-responsive PULs contain a single pair of *susC/D* homologs, along with multiple glycoside hydrolases (**Table S5**) which harbor either periplasmic or membrane lipoprotein secretion motifs (**Fig. S8)**. In contrast to the GAG-specific enzymes associated with the GAG-responsive loci, the α-mannan-regulated PULs contain mostly family 38, 76 and 92 glycoside hydrolases - all of which include α-mannosidases. The *BT2615-33* locus also encodes two other glycoside hydrolases belonging to families 67 and 97, which to date include only α-glucoronidases and α-glucosidases, respectively. Although we cannot predict a role for the family 67 enzyme, it is possible that the family 97 enzyme has a role in cleaving the α-glucosidic linkages contained in unprocessed *N*-glycans (**Fig. S1**). This possibility is intriguing because it would mean that individual PULs have not only evolved to sense individual linkages (α-mannosides) contained in target glycans, but also encode other activities targeting linkages commonly encountered in the same 'glycan context' (α-glucosides).

A novel feature of ECF-σ linked PULs

A novel feature observed with the ECF-σ linked PULs is that each locus contains one or more flanking genes that show lower fold-changes in expression in response to host glycans. Closer evaluation of these genes (shaded by gray boxes in **Fig. 4**), which mostly encode glycolytic functions, revealed that the probable reason for their low fold-change values is that they show high basal transcriptional levels under non-inducing conditions (**Fig. S11**). This suggests that these genes are expressed from promoters with constitutive activity in the absence of glycans. Moreover, the added inducibility of these genes in response to glycans indicates that their expression may be augmented by ECF-σ induction from upstream promoters. The reason for

this different regulatory mode is unknown, but it raises the possibility that during PUL evolution new functions may be appended to existing systems via addition of genes that are not dependent on the sensor/regulator that regulates the core PUL. These appended functions may, over time, become less constitutive and more dependent on substrate-dependent regulation.

Several of the enzyme-coding genes described above as having elevated basal expression may play roles in *O*-glycan degradation. Combined among the five loci shown in **Fig. 4** are one family 2 and two family 20 glycoside hydrolases, which, based on their CAZy family membership, likely encode β-galactosidase and β-hexosaminidase functions. The products of two of these genes (*BT1621* and *BT4241*) have predicted secretion motifs without proximal cysteine residues, suggesting that they are secreted but are not lipidated. Thus, these glycolytic activities are likely located in the periplasm where they act upon imported oligosaccharides. The products of five other genes associated with the PULs shown in **Fig. 4**, do not have functions that are immediately attributable to *O*-glycan metabolism. These encode four family 92 and one family 89 glycoside hydrolases, which are only known to contain α-1,2-mannosidase and α-N-acetylglucosaminidase activities, respectively. Since the linkages targeted by these known activities do not commonly occur in *O*-glycans, these enzymes must either have activities that are not yet described for their families, or target linkages other than those present in *O*-glycans.

Supplemental Experimental Procedures

Growth of *B. thetaiotaomicron* for transcriptional profiling

Minimal medium (MM) cultures were grown either in 18mm-diameter test tubes (5 ml culture volume) or in a 1.2 L Bioflo 110 batch fermentor (800 ml culture volume; New Brunswick Scientific). Pure host glycans/glycan components were purchased from Sigma (chondroitin sulfate mixed isomers of A/C, hyaluronan, dermatan sulfate, heparin, *S. cerevisiae* α-mannan, galactose, *N*-acetylgalactosamine and *N*-acetylglucosamine), or from V-Labs (Covington, LA; mucin core 1 and *N*-acetyllactosamine). Tube cultures were made anaerobic

using the $NaHCO_3$/pyrogallol method (Holdeman et al., 1977). Batch fermentor cultures were pre-reduced by sparging under agitation (20% CO_2/80% CO_2; ~0.1L min^{-1} at 37°C; 100 rpm for 30 min) prior to inoculation with 2 ml of an overnight culture of *B. thetaiotaomicron* VPI-5482 grown in TYG (~2 x 10^9 cfu ml^{-1}).

Growth profiles on PMG *O*-glycans are summarized in **Fig. S4**. For transcriptional profiling on unfractionated PMG *O*-glycans, *B. thetaiotaomicron* was grown in MM containing *O*-glycans. Cultures grown on unfractionated glycans produced a biphasic growth curve consisting of two prominent logarithmic phases (**Fig. S4A-B**). Samples were collected from triplicate batch fermentor cultures within the first growth phase (OD_{600} = 0.31, 0.31 and 0.32) and second growth phase (OD_{600} = 0.67, 0.69, 0.70) (**Fig. S4B**). *B. thetaiotaomicron* produced monophasic logarithmic growth patterns on the neutral, 300mM NaCl and 1M NaCl *O*-glycan fractions (each at 0.5% w/v). Samples were collected from duplicate 5 ml tube cultures and were sampled during mid-log phase for each substrate (OD_{600} = 0.59, 0.63 for neutral; 0.62, 0.72 for 300mM NaCl; 0.60, 0.63 for 1M NaCl). When grown on 100mM NaCl *O*-glycans (0.5% w/v), *B. thetaiotomicron* produced a polyphasic growth pattern consisting of at least 5 phases suggesting numerous physiologic shifts between multiple nutrient sources. Samples were harvested for transcriptome analysis from duplicate cultures that represent a first prominent growth phase (OD_{600} = 0.3, 0.3). To represent the numerous and shorter secondary growth phases, we sampled six replicate cultures at three separate phases in this secondary growth profile (**Fig. S4F**). Following RNA extraction from these samples (see below), equivalent amounts of material (3.3 μg of RNA from each growth point) were subsequently pooled into two sample sets, each representing all three of the secondary growth phases sampled [RNA was pooled from individual cultures sampled at OD_{600} 0.47, 0.55, 0.69 for pool #1 ('late phase'); and at OD_{600} 0.47, 0.55, 0.71 for pool #2 ('early phase')].

Preparation and fractionation of porcine mucosal glycans

Porcine mucosal glycans (PMG) were purified from porcine gastric mucin (Type III, Sigma) using a combination of proteolysis and alkaline β-elimination to break down mucin glycopeptides and release *O*- and *N*-linked glycans. Material was suspended at 2.5% w/v in 100mM Tris (pH 7.4): the mixture was immediately autoclaved for 5 min to increase solubility and reduce potential contaminating glycoside hydrolase and polysaccharide lyase activity. The heated solution was subsequently cooled to 65ºC, Proteinase K (Invitrogen) was added to a final concentration of 0.1 mg ml^{-1} and the suspension was incubated at 65ºC for 16-20h. The proteolyzed solution was subsequently centrifuged at 21,000 x g for 30 min at 4°C to remove insoluble material, and NaOH and $NaBH_4$ were added to final concentrations of 0.1M and 1M, respectively. This solution was incubated at 55ºC for 16h to promote selective release of *O*-glycans (mucin *O*-glycans and GAGs) from mucin glycopeptides by alkaline β-elimination, although some *N*-glycan release was anticipated and observed. The pH of the solution was subsequently decreased to 7.0 with HCl: the neutralized mixture was centrifuged at 21,000 x g for 30 min at 4°C, and then filtered through a 0.22 μm filter (Millipore) to remove remaining insoluble material. The filtrate was subsequently dialyzed (1 kDa cutoff) against deionized distilled (dd)H_2O (>1:10^5 dilution) to remove salts and contaminating small molecules. The dialyzed glycans were lyophilized, dissolved in ddH_2O (at a final concentration of 20 mg ml^{-1}) and stored at -20ºC.

A portion of the porcine mucosal glycans was further fractionated using anion exchange chromatography. Glycans, purified as above from 20 g of porcine gastric mucin, were passed twice over a DEAE-Sepharose (Sigma) column (325 ml bed volume; equilibrated in 50mM Tris 7.4; flow rate of 1-1.5 ml min^{-1}). The flow through (neutral fraction) was collected and the column washed with 1L of 50mM Tris, pH 7.4. The charged glycans retained on the column were eluted with sequential NaCl washes (~800 ml each) of increasing concentration (100mM, 300mM

and 1M). The resulting fractions (neutral, 100mM NaCl, 300mM NaCl and 1M NaCl) were dialyzed against ddH_2O (1 kDa cutoff, $>1:10^5$ dilution), lyophilized and resuspended in ddH_2O at 20 mg ml^{-1}. The relative masses of material recovered after ion-exchange chromatography was 4:1:4:1 for the neutral, 100mM NaCl, 300mM and 1M fractions, respectively.

The monosaccharide compositions of unfractionated and fractionated PMG glycans were analyzed at the University of California San Diego Glycotechnology Core Resource by high pH anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD). Three individual assay conditions were required for analysis of neutral monosaccharides (fucose, N-acetylgalactosamine, N-acetylglucosamine, galactose, glucose and mannose), N-acetylneuraminic acid, and uronic acids (glucoronate, iduronate and galacturonate), respectively. To calculate individual monosaccharide proportions using the three different assays (**Table S1**), the molar concentrations of each of the sugars analyzed were normalized to the amount of material used in their respective HPAEC-PAD runs, and summed to give the total moles of sugar detected in all assays. GalA, a common component of plant glycans, was observed in the mucosal glycan samples and was likely generated through previously described desulfation and isomerization reactions involving IdoA-2-sulfate under the alkaline conditions employed (Conrad, 1998). All detected GalA was therefore counted as IdoA.

Genomic mapping and sequence analysis

A physical map of the *B. thetaiotaomicron* VPI 5482 (also known as ATCC 29148) genome was constructed based on the 4,779 ORFs originally reported for this species (Xu et al., 2003) using the program GenVision (DNAStar; Madison, WI). ORFs believed to be involved in sensing and directing glycan metabolism were labeled based on their previous annotations in the CAZy database (www.cazy.org), or by homology to the starch binding proteins SusC and SusD (Xu et al., 2007), or based on annotation as a sensor/regulator or other enzymatic functions (Xu et al., 2003).

Deduced protein sequences for BT3853, BT1770 and BT2204 were aligned using ClustalW in the MegaAlign component of the Lasergene software package (DNAStar). Signal sequences and transmembrane helices were predicted using SignalP (http://www.cbs.dtu.dk/services/SignalP/) and Toppred (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=toppred), respectively (Claros and von Heijne, 1994; Gardy et al., 2003).

Complementation and genomic tagging of *B. thetaiotaomicron* mutants

Genetic complementation of selected mutants was carried out using a set of novel constructs, p*NBU2-bla-tetQ*b and p*NBU2-bla-ermG*b (**Table S8**), that were based on two new *Bacteroides* suicide vectors created for this study (see below). Plasmids p*NBU2-bla-tetQ*b and p*NBU2-bla-ermG*b contain the integrase gene (*intN2*) and attachment site targeting sequence (*attN2*) from the non-replicating *Bacteroides* element *NBU2* (kindly provided by Abigail A. Salyers, University of Illinois, Urbana-Champaign) (Wang et al., 2000). This construct carries cloned DNA fragments into the *B. thetaiotomicron* genome, in single-copy, by integrating into one of two $tRNA^{ser}$ attachment sites without disrupting any genetic functions. Proper insertion of complementation constructs into $tRNA^{ser}$ loci was verified by PCR using primer pairs listed in **Table S9**.

Signature-tag constructs were created by annealing 5'-phosphorylated, complementary oligonucleotides designed to produce DNA duplexes with unique 24bp sequences (**Table S9**). Oligonucleotide pairs were combined in ddH_2O (100μM each, 200μl total volume), heated to 100°C, and allowed to cool slowly to room temperature. Sequences were designed such that each duplex also contained unique restriction sites at both ends (*Xba*I and *Sal*I, directly adjacent to the unique 24bp sequence), and these sites were used to ligate each tag into p*NBU2-bla-tetQ*b. Genomic DNA extracted from *B. thetaiotaomicron* strains harboring each of the three signature-tags used in this study were combined in test mixtures, which varied the concentration of each tag relative to the remaining two. qPCR assays of these test mixtures verified that each could be

specifically detected at a concentration as low as 10^{-4} relative to the remaining tags (data not shown). Proper insertion of signature tag *NBU2* constructs into $tRNA^{ser}$ loci was confirmed by PCR as described above. Detection of each signature-tag by qPCR was accomplished by using a variable primer that hybridized to each unique 24bp tag and a universal primer that hybridized to a position in the vector 210nt away. Thus, the amplicon size and sequence for each tag assay was nearly identical, varying only on the end bearing the signature-tag.

New *Bacteroides* spp. plasmid constructs

Several new plasmids were created during this work (**Table S8**): they include two new *ori*R6K suicide vectors for use in *Bacteroides* spp., pKNOCK-*bla-ermGb* and pKNOCK-*bla-tetQ,* that are derivatives of the Gram-negative suicide vector pKNOCK-Cm (Alexeyev, 1999). These plasmids were constructed by removing a *Mlu*I fragment carrying the *cat* (chloramphenicol resistance) gene from pKNOCK-Cm, and replacing it with a different *Mlu*I-ended fragment encoding either *bla-ermG* (β-lactamase gene for ampicillin selection in *E. coli* and *ermG* for erythromycin selection in *Bacteroides* spp.), or *bla-tetQ* (β-lactamase for ampicillin selection in *E. coli* and *tetQ* for tetracycline selection in *Bacteroides* spp.). *NBU2* based insertion vectors, p*NBU2-bla-ermG*b and p*NBU2-bla-tetQ*b, were each constructed by amplifying a fragment encoding the *intN2-att* fragment (with *Kpn*I and *Sal*I ends) from the plasmid pEP*intN2* (Wang et al., 2000). The complete nucleotide sequences of these plasmids are available at http://gordonlab.wustl.edu/plasmids.

Bacteroidetes phylogenetic analysis

A 16S rRNA phylogeny was constructed by collecting aligned sequences corresponding to 194 different Bacteroidetes genera from the GreenGenes database (www.greengenes.org; DeSantis et al., 2006). A single species from the Chlorobi, *Chlorobium phaebacteroides*, was chosen as an outgroup, and a representative neighbor-joining tree was generated using the program Mega (Kumar et al., 2008). Searches for species containing Sus-like PULs were performed using the *B. thetaiotaomicron* SusC amino acid sequences as a BLAST query against

publicly-searchable completed or active genomic databases for species belonging to 23 different Bacteroidetes genera. **Table S10** lists the genera searched along with the highest numbers of SusC homologs in each genus and the corresponding E-values of best BLAST-hits.

References

Alexeyev, M.F. (1999). The pKNOCK series of broad-host-range mobilizable suicide vectors for gene knockout and targeted DNA insertion into the chromosome of Gram-negative bacteria. BioTechniques *26*, 824-826, 828.

Cawley, T.N., and Ballou, C.E. (1972). Identification of two *Saccharomyces cerevisiae* cell wall mannan chemotypes. J. Bacteriol. *111*, 690-695.

Claros, M.G., and von Heijne, G. (1994). TopPred II: an improved software for membrane protein structure predictions. Comput. Appl. Biosci. *10*, 685-686.

Conrad, H.E. (1998). *Heparin-binding proteins* (London, Academic Press).

D'Elia, J.N., and Salyers, A.A. (1996a). Contribution of a neopullulanase, a pullulanase, and an alpha-glucosidase to growth of *Bacteroides thetaiotaomicron* on starch. J. Bacteriol. *178*, 7173-7179.

D'Elia, J.N., and Salyers, A.A. (1996b). Effect of regulatory protein levels on utilization of starch by *Bacteroides thetaiotaomicron*. J. Bacteriol. *178*, 7180-7186.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069-5072.

Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K*., et al.* (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nuc. Acids Res. *31*, 3613-3617.

Holdeman, L.V., Cato, E.D., and Moore, W.E.C. (1977). *Anaerobe Laboratory Manual* (Blacksburg, Va.: Virginia Polytechnic Institute and State University Anaerobe Laboratory).

Kumar, S., Nei, M., Dudley, J., and Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. *9*, 299-306.

Kuwahara, T., Yamashita, A., Hirakawa, H., Nakayama, H., Toh, H., Okada, N., Kuhara, S., Hattori, M., Hayashi, T., and Ohnishi, Y. (2004). Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. Proc. Natl. Acad. Sci. USA *101*, 14919-14924.

Moran, N.A., Tran, P., and Gerardo, N.M. (2005). Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. Appl. Environ. Microbiol. *71*, 8802-8810.

Shipman, J.A., Berleman, J.E., and Salyers, A.A. (2000). Characterization of four outer membrane proteins involved in binding starch to the cell surface of *Bacteroides thetaiotaomicron*. J. Bacteriol. *182*, 5365-5372.

Shipman, J.A., Cho, K.H., Siegel, H.A., and Salyers, A.A. (1999). Physiological characterization of SusG, an outer membrane protein essential for starch utilization by *Bacteroides thetaiotaomicron*. J. Bacteriol. *181*, 7206-7211.

Sonnenburg, E.D., Sonnenburg, J.L., Manchester, J.K., Hansen, E.E., Chiang, H.C., and Gordon, J.I. (2006). A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing *in vivo* to carbohydrate metabolism. Proc. Natl. Acad. Sci. USA *103*, 8834-8839.

Varki, A., Cummings, R., Esko, J., Freeze, H., Hart, G., and Marth, J. (1999). *Essentials of Glycobiology* (Plainview, NY, Cold Spring Harbor Laboratory Press).

Wang, J., Shoemaker, N.B., Wang, G.R., and Salyers, A.A. (2000). Characterization of a *Bacteroides* mobilizable transposon, *NBU2*, which carries a functional lincomycin resistance gene. J. Bacteriol. *182*, 3559-3571.

Xu, J., Bjursell, M.K., Himrod, J., Deng, S., Carmichael, L.K., Chiang, H.C., Hooper, L.V., and Gordon, J.I. (2003). A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. Science (New York, NY *299*, 2074-2076.

Xu, J., Mahowald, M.A., Ley, R.E., Lozupone, C.A., Hamady, M., Martens, E.C., Henrissat, B., Coutinho, P.M., Minx, P., Latreille, P*., et al.* (2007). Evolution of Symbiotic Bacteria in the Distal Human Intestine. PLoS Biol. *5*, e156.

Supplemental Figure Legends

Figure S1. Sources and structures of host glycans. Representative host glycans are illustrated, together with their component monosaccharide and glycosidic linkages. Linkages are abbreviated with the assumption that they originate from the C1 hydroxyl group, except sialic acid, which originates from the C2 hydroxyl (*e.g.*, 'β3' between galactose residues indicates a β-1,3 linkage). Colored boxes indicate specific linkages that were tested for PUL induction in this study: core 1 disaccharide (orange), *N*-acetyllactosamine (LacNAc, green) and α-mannan (blue). (**A**) An illustration of host glycans. Glycosaminoglycans (GAGs) are components of epithelial cell surface proteoglycans and the extracellular matrix. Along with cell surface *N*-glycans, these molecules could be introduced into the lumen of the intestine through the continuous renewal of its epithelium: epithelial cells descended from multipotential stem cells positioned near or at the base of crypts of Lieberkühn, migrate up adjacent villi located in the small intestine, or to the surface epithelial cuff that surrounds the crypt orifice in the colon, and are then exfoliated. Members of the goblet epithelial cell lineage secrete copious amounts of proteins with abundant *O*-glycosylation and less abundant *N*-glycosylation into the lumen, producing a protective layer of mucus that covers the epithelium. (**B**) Representative chemical structures of common GAGs. Four of the GAG types exist as protein glycoconjugates (proteoglycans). Chondroitin sulfate, dermatan sulfate and heparin sulfate are *O*-linked via tetrasaccharides containing xylose, galactose and glucuronic acid, whereas keratan sulfate may be either *N*-linked, via a core structure similar to that of complex *N*-glycans (see panel C below), or *O*-linked, via a core structure similar to that for core 2 mucin *O*-glycans (see panel D below). In this view, *O*-linked keratan sulfate glycans are structurally very similar to *O*-glycans, differing only in the lack of fucosylation and presence of sulfation. (**C**) Representative structures of *N*-glycans. Unprocessed *N*-glycans, which exist in the endoplasmic reticulum (ER) prior to being processed and secreted

through the Golgi, consist of a core N-acetylglucosamine (GlcNAc) β1,4-linked disaccharide with an attached mannose cluster containing α1,2, α1,3 and α1,6 linkages (the same ones found in α-mannan from *S. cerevisiae* (Cawley and Ballou, 1972)). Mannose chains may contain α-linked glucose residues at their non-reducing ends: these residues are remnants of an initially synthesized dolichol-oligosaccharide precursor. After release from the ER, *N*-glycans are processed to remove terminal glucose and mannose residues. High mannose *N*-glycans contain unsubstituted terminal mannose residues, whereas complex *N*-glycans contain a smaller tri-mannose core with two or three oligosaccharide chains that may be extended by GlcNAc and galactose, and may terminate with sialic acid. Complex *N*-glycans may contain fucosylation of the core GlcNAc disaccharide. Hybrid *N*-glycans also exist that combine structural feature of high mannose and complex types. (**D**) Representative mucin *O*-glycan structures, each beginning with a serine- or threonine-linked *N*-acetylgalactosamine (GalNAc) residue that is subsequently elaborated with various 1,3 and/or 1,6 linkages to produce a series of 'core' oligosaccharide structures. The four most common core subtypes (1-4) are illustrated and are composed of galactose and/or GlcNAc linkages to GalNAc. Oligosaccharide chains, which consist of repeating LacNAc units (green) may extend from the core and can be heterogeneous in length. The branched core 2 and 4 structures may contain biantennary extensions that originate from each of GalNAc linked monosaccharides. Oligosaccharide chains are frequently fucosylated and may terminate in sialic acid residues. Glycan structures are adapted from (Varki et al., 1999).

Figure S2. Generic Polysaccharide Utilization Locus (PUL). A model based on the prototypic starch utilization system (Sus) initially characterized by Salyers and co-workers (D'Elia and Salyers, 1996a; Shipman et al., 2000; Shipman et al., 1999), plus observations of similar systems in *Bacteroides* spp. (**A**) Eight genes are illustrated and correspond to the genes contained in the *susRABCDEFG* PUL (labels below each gene). The definitive feature of each PUL is one or more

pairs of *susC/D* genes, which so far have always been observed immediately adjacent to one another in the same order. Gene content and order in PULs can be varied relative to the *Sus* cluster. However, glycolytic enzymes, sensor regulators and other hypothetical functions are typical PUL components. (**B**) A schematic of how PUL gene products participate in glycan catabolism. Extracellular outer membrane proteins (including SusD), with affinity for a specific glycan species, bind the target substrate to the bacterial cell surface where it is cleaved by an endo-acting glycolytic enzyme (SusG). Liberated oligosaccharides are imported into the periplasm through a SusC-like transporter and further digested by periplasmic enzymes prior to being transported into the cell through a sugar permease(s). PUL-associated sensor-regulators, many of which appear to have inner membrane spanning components that allow them to directly sense periplasmic sugars, detect glycan degradation products and activate expression of adjacent PUL genes. For example, SusR activates expression of *susABCDEFG* in response to maltose but not glucose, indicating it acts as a sensor of both monosaccharide content *and* glycosidic linkage (D'Elia and Salyers, 1996b).

Figure S3. Purification and fractionation of PMG glycans. Flow diagram illustrating steps used to purify host glycans from porcine mucosa. Materials used for bacterial growth and transcriptional profiling are labeled with the corresponding figure number in red.

Figure S4. Growth of *B. thetaiotaomicron* on PMG glycans. Growth profiles of *B. thetaiotaomicron* in minimal medium with purified host glycans added as the sole carbon sources. Cultures that generated polyphasic growth curves are indicated as bicolor curves with a horizontal dashed line separating the primary growth phase (red) from the later phase(s) (blue). Glycan substrates, substrate concentrations and culture method (tube-based or in batch fermentor) are noted in each panel. Arrows and numbers indicate the harvest points and corresponding optical density values (2-3 biological replicates per point) for transcriptional profiling experiments. With

the exceptions of (**B)** and (**F)**, each panel shows the results of three individual growth experiments for which optical density OD_{600} values were averaged at each time point. Error bars indicate standard deviations for each point. Panel B shows three batch fermentor cultures used for transcriptional profiling in **Fig. 1**. Individual data points were omitted to highlight the six points at which bacteria were harvested for transcriptional profiling during two growth phases (red and blue open circles). Note that OD_{600} values in panel B were measured using a 1.0cm path length, as opposed to a 1.8cm path length for all other tube cultures, and unfractionated PMG was at a concentration of 1% rather than 0.5% to increase RNA yield from cells harvested during the first phase. Thus, OD_{600} values presented in panels A and B do not correspond directly to one another, although both profiles remain biphasic. Panel F only shows a single representative replicate of growth in 100mM glycans, in order to highlight the small secondary growth phases.
Growth curves on all except the 100mM NaCl PMG fraction were monophasic (**panels C-E**), suggesting that the different glycan components that effected catabolite prioritization during growth on unfractionated PMG glycans had been separated by chromatography. Growth on the 100mM fraction, which contains all three classes of host glycans is characterized by an initial dominant growth phase followed by several shorter phases (**panel F**). Growth on a 1:1 mixture of the two most abundant glycan fractions ('neutral' and '300mM'), which contain mucin *O*-glycans and GAGs respectively, restores a diauxic growth pattern, suggesting that *B. thetaiotaomicron* prioritizes one of these glycan types (**panel G**). Growth on an equal mixture of the 'neutral', '300mM', and another purified GAG - chondroitin sulfate (**panel H**), produces a diauxic profile with a protracted primary phase, suggesting simultaneous use of GAGs (300mM glycans and chondroitin sulfate) prior to mucin *O*-glycans.

Figure S5. Genomic map of *B. thetaiotaomicron* VPI-5482 illustrating the locations of its 88 putative PULs. Individual ORFs are represented by rectangles that reflect the relative sizes of each coding region: ORFs below the horizontal axis code left to right (positive strand), those

below the axis code right to left (negative strand). Functions relating to glycan metabolism are color-coded (see accompanying symbol key for details). Each of the 88 putative PULs is delineated by a light blue box along with its deduced substrate specificity, where applicable. Capsular polysaccharide synthesis (*CPS*) loci are shown in orange for reference. Updated versions of this map will be available at http://gordonlab.wustl.edu/B_thetaVPI5482_glycobiome.

Figure S6. Identification of induced PUL genes. (**A**) Data analysis scheme for the 485 regulated genes that resulted in identification of host glycan-induced PULs. The rationale for each step is described in the main text. (**B**) Individual fold-change values for the 211 PUL genes that exhibited $\geq$10-fold induction in the host glycan growth conditions shown in **Fig. 1**.

Figure S7. Specificity of PUL responses to purified GAG species. (**A**) Induction of the *BT3328-34* locus and flanking genes, *BT3324* and *BT3348-50*, in response to PMG glycan fractions and purified GAGs. (**B**) Induction of the *BT4652-63* locus and *BT4675* in response to PMG glycan fractions and purified GAGs. Note the respective specificities of each PUL for chondroitin sulfate/hyaluronan and heparin. Values represent the mean $\pm$ range of two biological replicates performed for each substrate.

Figure S8. Architecture and induction of *N*-glycan PULs. (A) The genomic organization of three PULs, induced during α-mannan utilization, is illustrated along with predicted functional annotations of their component genes. Enzymatic functions are labeled according to their assigned CAZy glycoside hydrolase (GH) families. Other functions are color coded according to the legend. Gray areas connecting genes in the *BT2615-33* and *BT3773-92* loci indicate homologous regions. Fold-change induction values of the individual genes contained in three α-mannan/*N*-glycan sensing loci: *BT2615-33* (**B**); *BT3773-92* (**C**); and *BT3853-62* (**D**). Note the

stronger overall induction of genes in response to growth on pure α-mannan *in vitro* compared to *in vivo,* suggesting that mannose linkages in *N*-glycans are not a primary endogenous substrate foraged by *B. thetaiotaomicron* in the gut. Values represent the mean ± range of two biological replicates performed for each substrate.

Figure S9. Mucin *O*-glycan PULs with non-ECF-σ regulators

(**A**) A mucin *O*-glycan responsive, HTCS-regulated PUL. The location of a HTCS regulator at the 3' end of the locus is in contrast to HTCS placement at the 5' ends of identified GAG utilization PULs. (**B**) Two small PULs that exhibit robust *in vivo* induction. Although each of these systems is clearly regulated by some mechanism, they are not physically linked to a conspicuous regulator. Note that the *BT0317-19* locus is just a pair of *susC/D* homologs: *BT0318/19* is a *susD* homolog for which the published sequence of strain VPI-5482 contains a frame shift, resulting in its annotation as two genes (Xu et al., 2003).

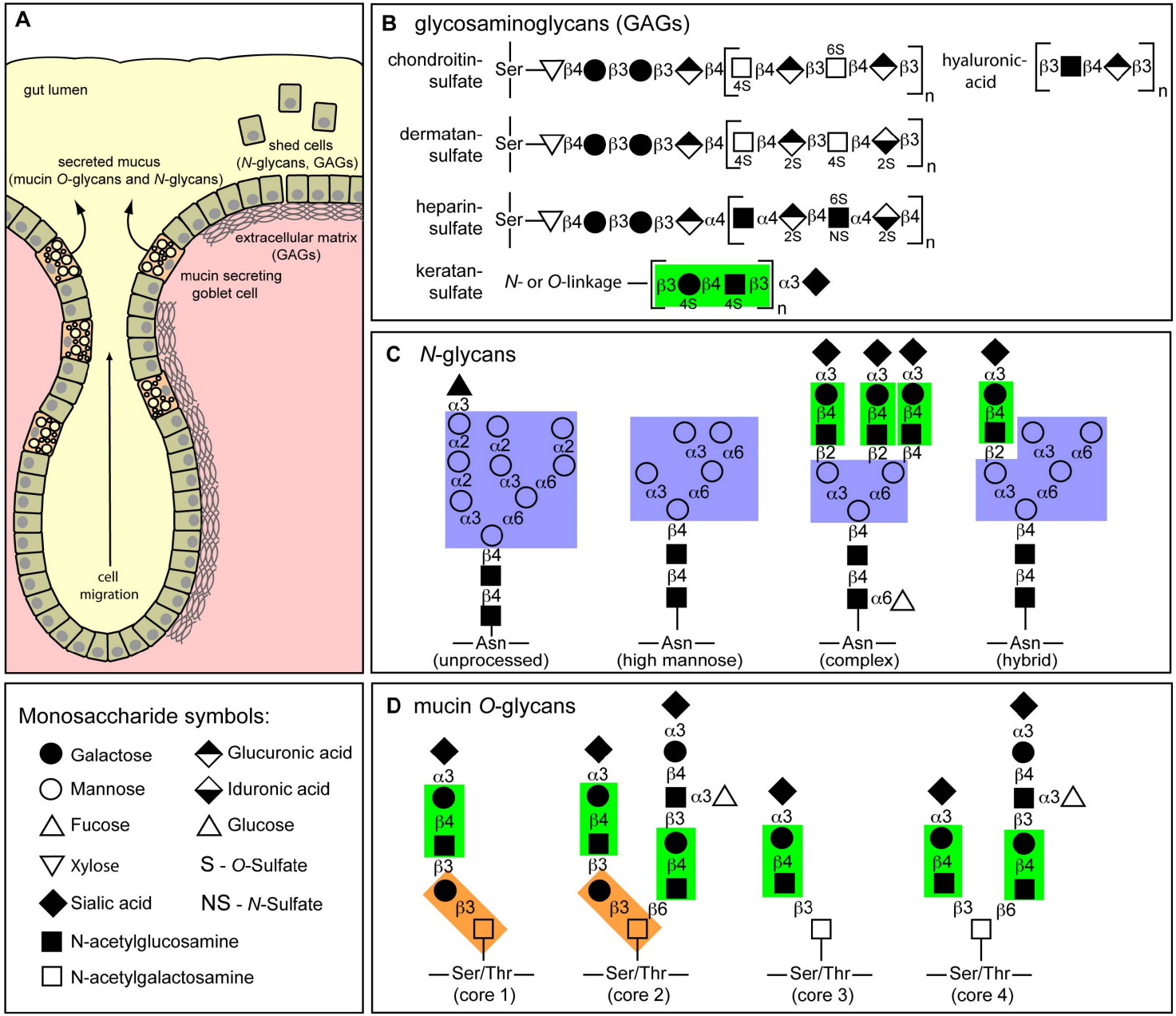
Figure S10. Alignment of putative SARP family glycan sensor/regulators. Alignment of BT3853, BT1770 and BT2204 illustrating homology over the entire protein sequence (amino acid identities are highlighted in yellow). Each protein has a putative cleavable signal peptide (blue boxes at N-terminus; predicted using SignalP) and a single internal transmembrane helix (green boxes; predicted using the program Toppred). The N-terminal domain (relative to the transmembrane helix) is predicted to be periplasmic and the C-terminal domain cytosolic. Residues in BT1770 that show homology to the *Thermotoga maritima* TM1119 SARP family transcriptional factor DNA binding domain (COG3629) are underlined in red. The presence of a hypothetical C-terminal DNA domain in these proteins is consistent with their predicted membrane topology, and the potential role of BT3853 as an α-mannan-sensing transcriptional regulator.

Figure S11. Normalized intensity values of genes from five *O*-glycan PULs. Normalized GeneChip intensity values are shown for the five ECF-σ-linked PULs illustrated in **Fig. 6A**: *BT1032-53* (panel **A**); *BT3983-94* (panel **B**); *BT4240-50* (panel **C**); *BT4355-59* (panel **D**); and *BT1617-22* (panel **E**). With the exception of panel B, ECF-σ gene intensities are not shown. PUL gene numbers are indicated under each histogram along with the deduced transcriptional units shown in **Fig. 4**. *In vitro* growth on MM-glucose (chemostat; yellow bars) is compared to three bacterial transcription profiles *in vivo* in adult NMRI mice fed a simple sugar diet (from **Fig. 6**): prototrophic 'wild-type' *B. thetaiotaomicron* (*tdk* parent strain, black bars); Δ5ECF-σ (red bars); and the complemented ECF-σ mutant (blue bars). Note the higher expression levels in MM-glucose of the genes highlighted by gray boxes (these correspond to **Fig. 4**), indicating higher constitutive expression in the absence of inducing glycans. Despite their higher basal expression, they are also induced when glycans are present (black bars), a response that is diminished in the ECF-σ mutant and restored upon complementation (blue bars). Note that the y-axis in each graph is discontinuous and increments are different on upper and lower axes. Values represent the mean ± standard deviation of three biological replicates performed for each mutant *in vivo*.

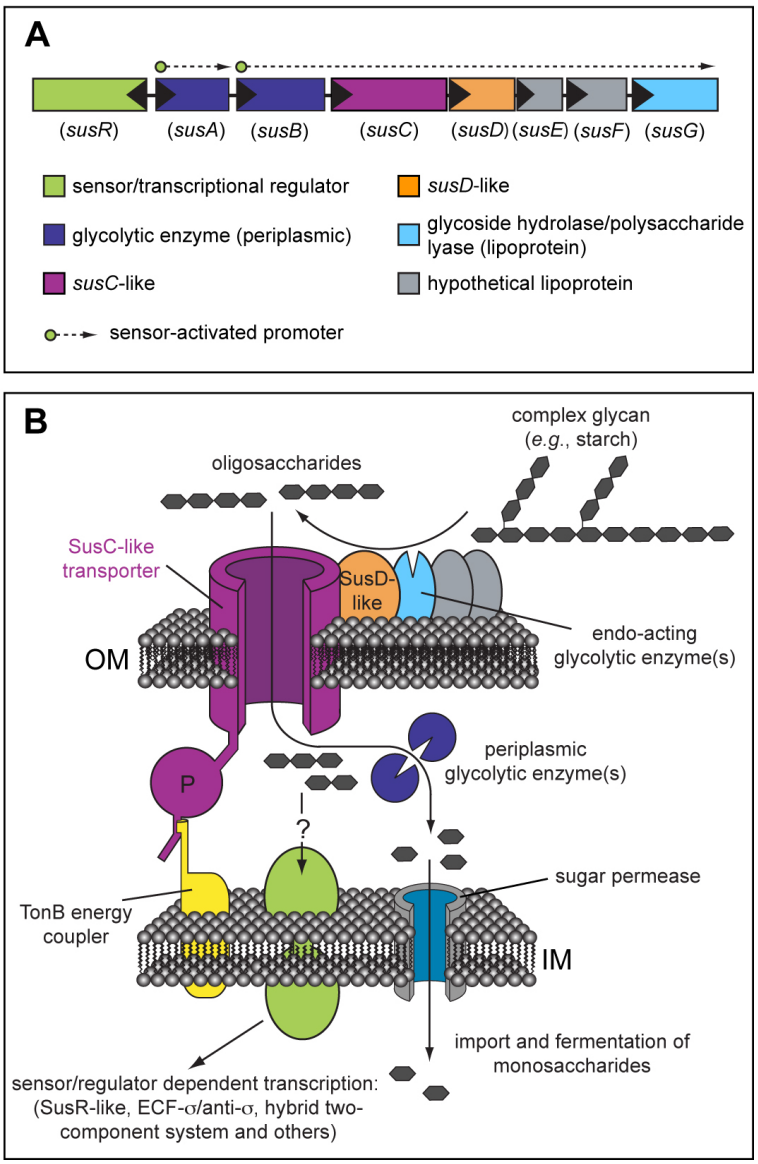
Figure S12. Inversion of the *BT4249* ORF fragment. (A) The original sequence assembly of the *B. thetaiotaomicron* VPI-5482 genome delineated two divergent ORFs, *BT4248* and *BT4249*, that appear to encode two fragments of a complete anti-σ factor gene (red ORFs). Genomic analysis of invertable elements in the *B. fragilis* and *B. thetaiotaomicron* genomes (Kuwahara et al., 2004) suggested that a fragment spanning *BT4249* (middle gene in the 'assembled orientation') is capable of inversion via two flanking copies of a 15bp sequence (5'-agttctaacagaact) containing a partial palindrome (underlined sequence). This inversion would yield a full-length anti-σ gene. Therefore, PCR primers were designed (1-3, **Table S7**) that

amplify fragments indicative of both orientations. (**B)** Results of PCR assays using primer pairs 1-2 and 1-3 on genomic DNA isolated from bacteria grown in TYG medium. Both genomic orientations are detectable, indicating that *BT4249* undergoes DNA inversion. The functional consequence of this event is unknown.

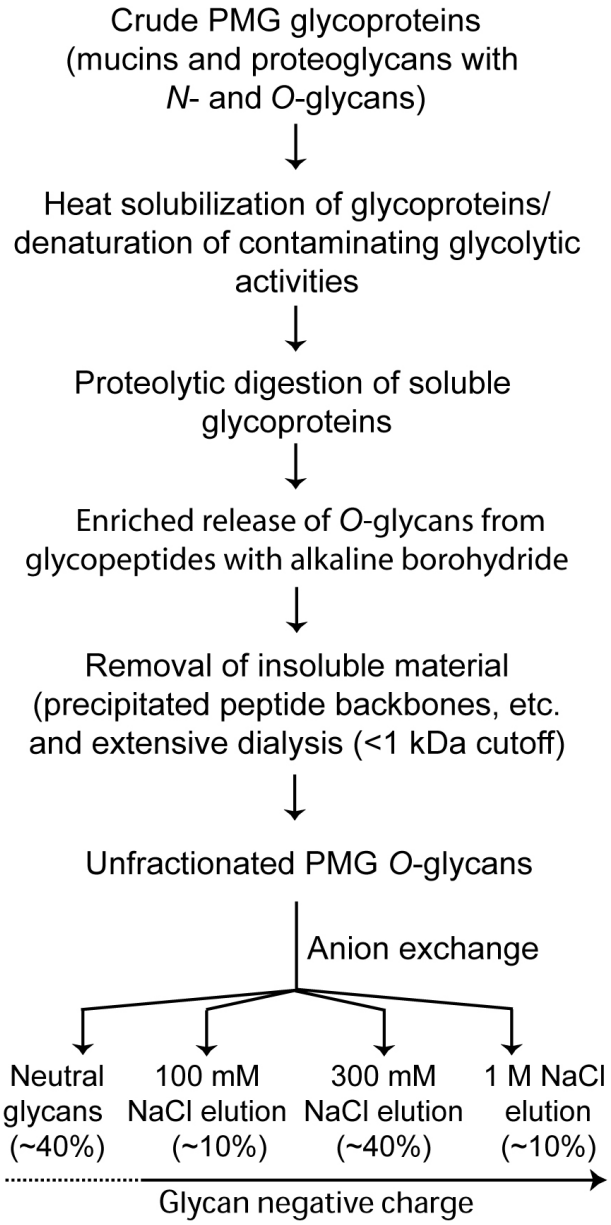
Figure S13. Diversity of Bacteroidetes species containing Sus-like PULs. A phylogenic tree of gut and environmental Bacteroidetes spp. Individual taxa correspond to representative Bacteroidetes species from 194 different genera, plus *Chlorobium phaebacteroides*, which is used as an outgroup. Taxa are color-coded according to their inclusion in three major Bacteroidetes classes: Flavobacteria (blue), Sphingobacteria (green) and Bacteroidetes (red). All genera commonly associated with mammalian alimentary tracts (including mouth, rumen and distal gut) or the insect gut are contained in the *class* Bacteroidetes branch (red). Genera for which searches of completed or active genome sequencing projects yielded detectable SusC homologs are noted with a black asterisk. The red 'X' highlights *Candidatus sulcia muelleri*. No susC homologs are discernable in its finished genome sequence. However, *Candidatus sulcia muelleri* is an endosymbiont of Hemipteran insects: its genome is only 0.25kb, indicating that it has undergone massive genome reduction (Moran et al., 2005). See **Table S10** for the names of genera that contain SusC homologs.


A
gut lumen
shed cells
(N-glycans, GAGs)
secreted mucus
(mucin O-glycans and N-glycans)
extracellular matrix
(GAGs)
mucin secreting
goblet cell
cell
migration
B glycosaminoglycans (GAGs)
chondroitin-sulfate
dermatan-sulfate
heparin-sulfate
keratan-sulfate
N- or O-linkage
hyaluronic-acid
Ser
4S
2S
6S
NS
C N-glycans
—Asn—
(unprocessed)
(high mannose)
(complex)
(hybrid)
D mucin O-glycans
—Ser/Thr—
(core 1)
(core 2)
(core 3)
(core 4)
Monosaccharide symbols:
Galactose
Mannose
Fucose
Xylose
Sialic acid
N-acetylglucosamine
N-acetylgalactosamine
Glucuronic acid
Iduronic acid
Glucose
S - O-Sulfate
NS - N-Sulfate

Martens *et al.*, Figure S2



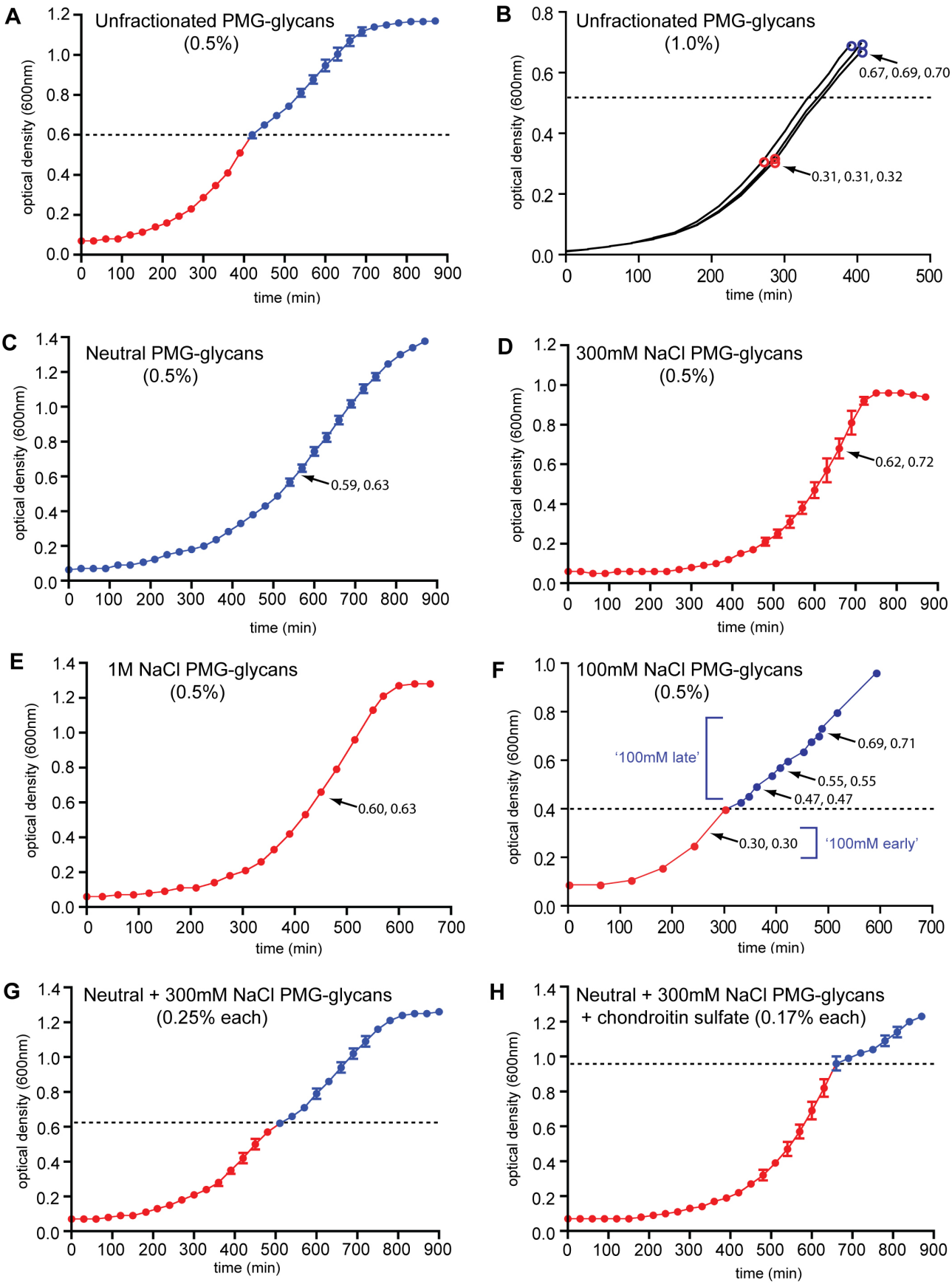
Martens *et al.*, Figure S3



Figs. 1-2

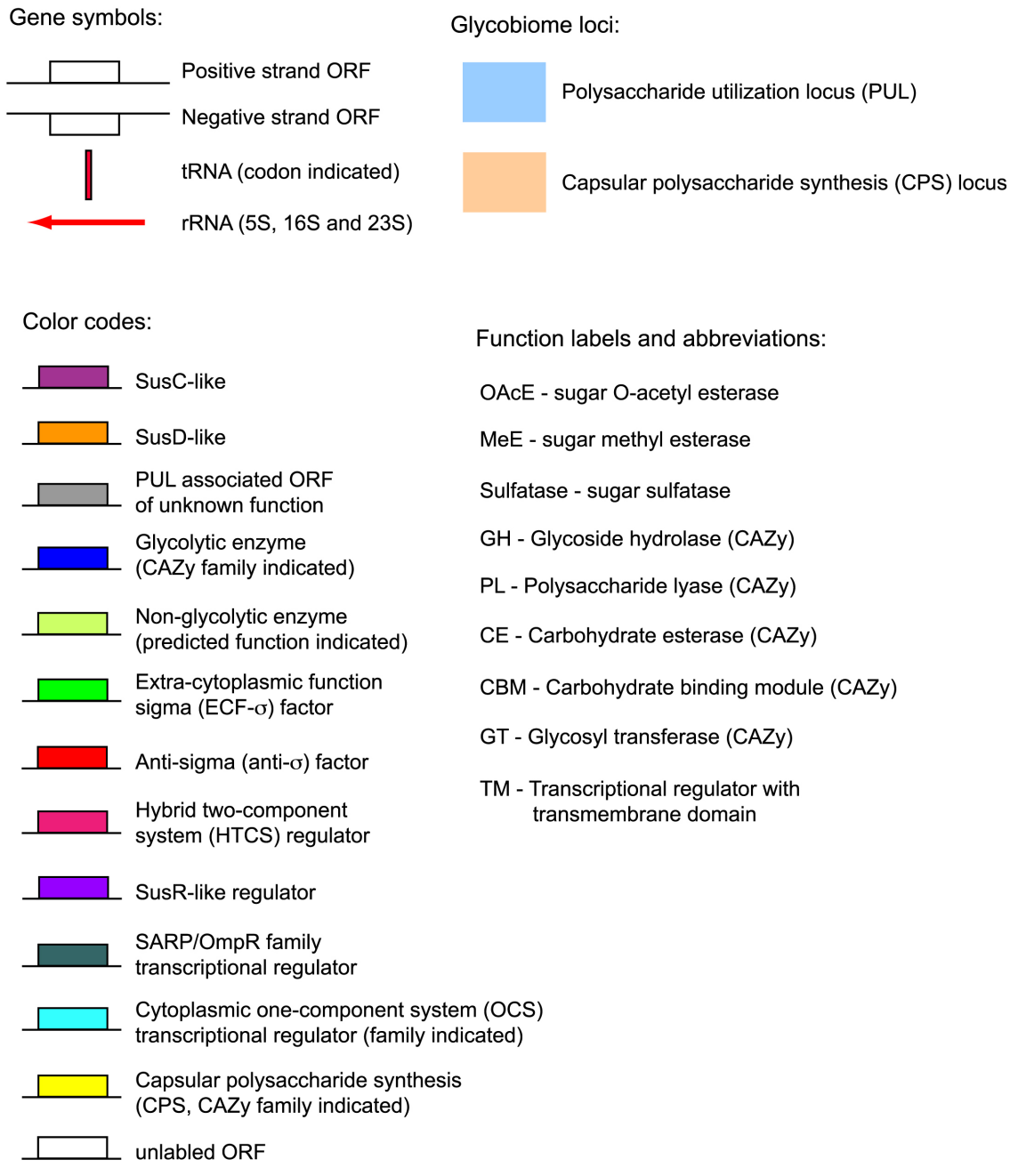
Fig. 2

Martens *et al.*, Figure S4



Note: Fig. S5 is provided as a separate PDF due it size and is also available at this URL:
http://gordonlab.wustl.edu/Glycobiome.html

Martens *et al.*, Figure S5 symbol key



Martens *et al.*, Figure S6

A

Total regulated genes
(411 up, 74 down)

↓

PUL associated genes
(211 up, 0 down)

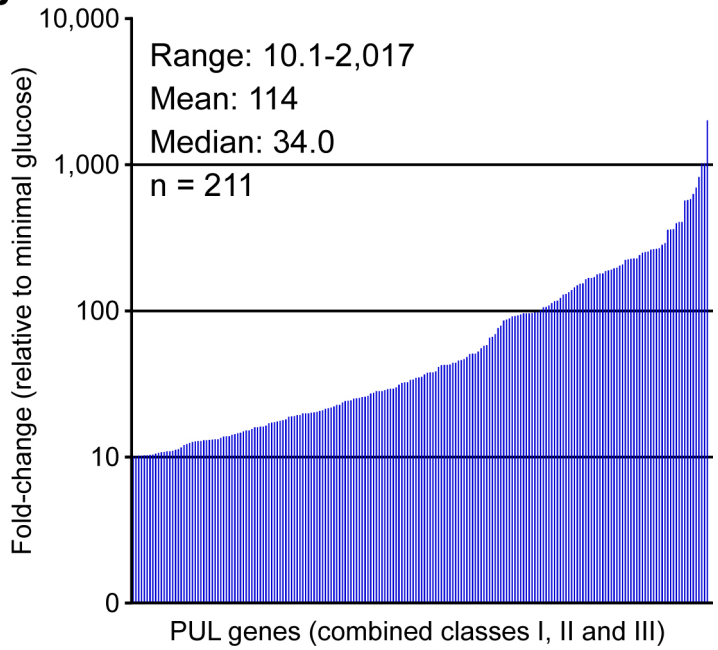
↓

Repopulated PULs
(424 genes)

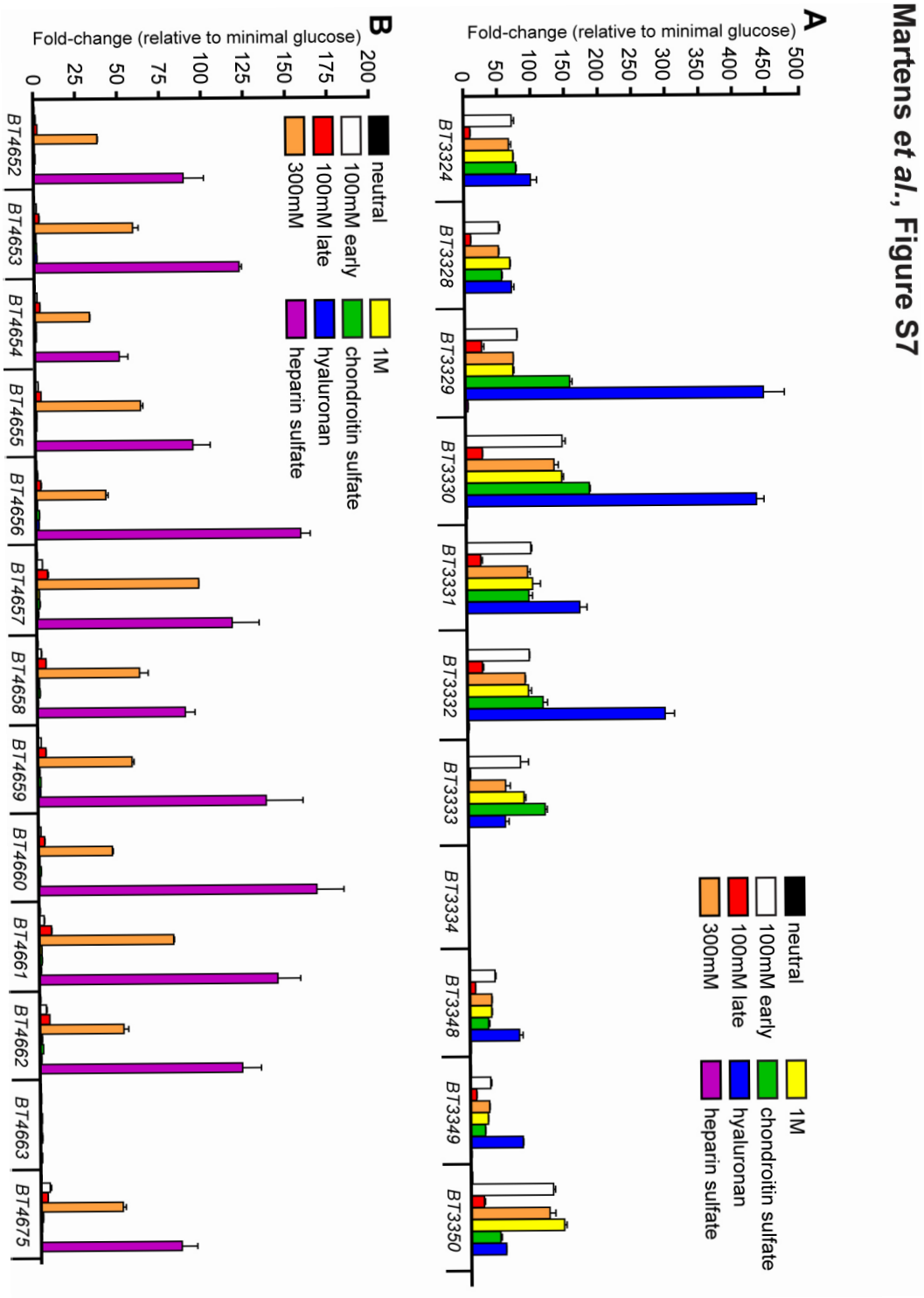
↓

PULs with operons >10-fold
(264 genes, 39 PULs)

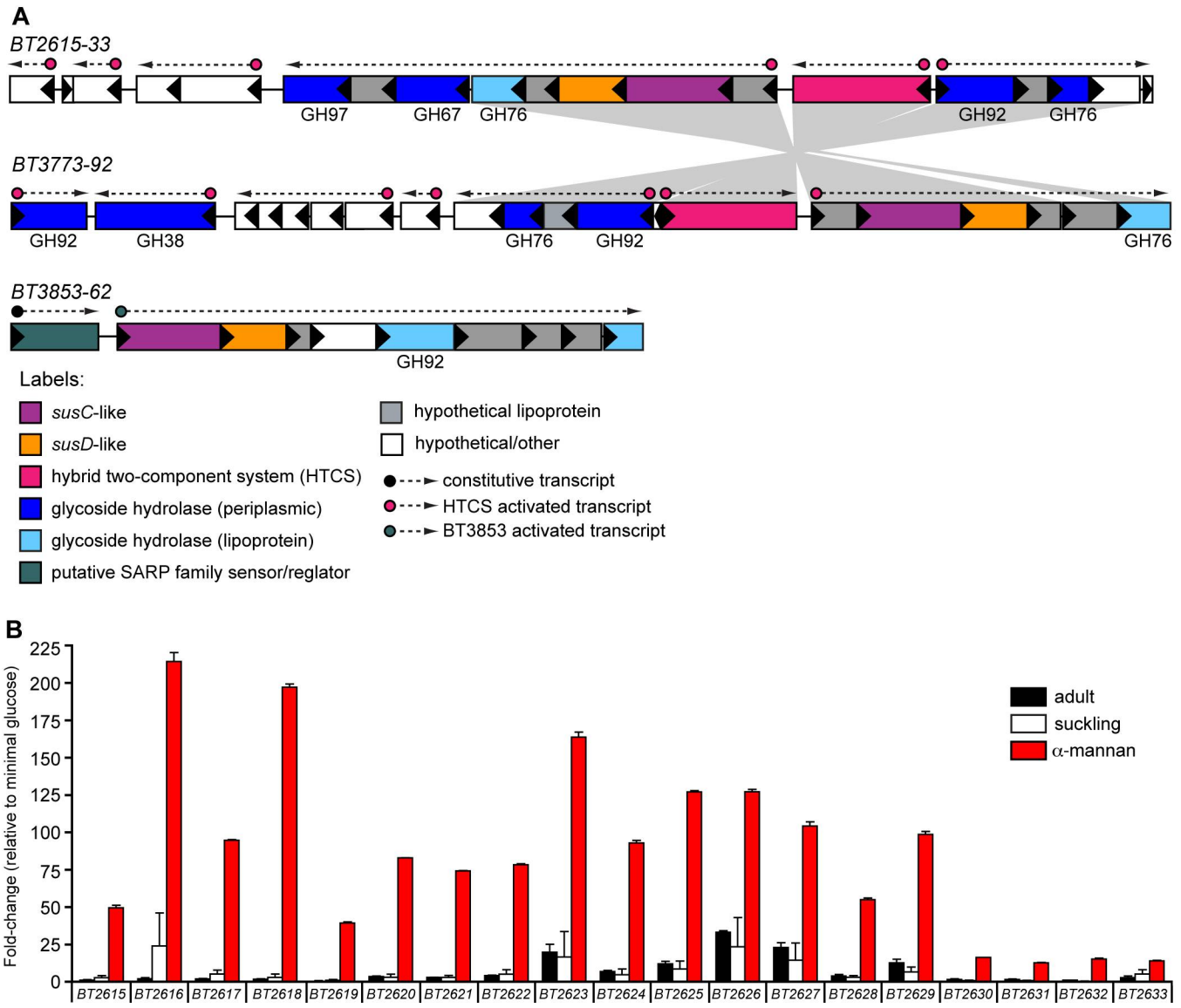
B



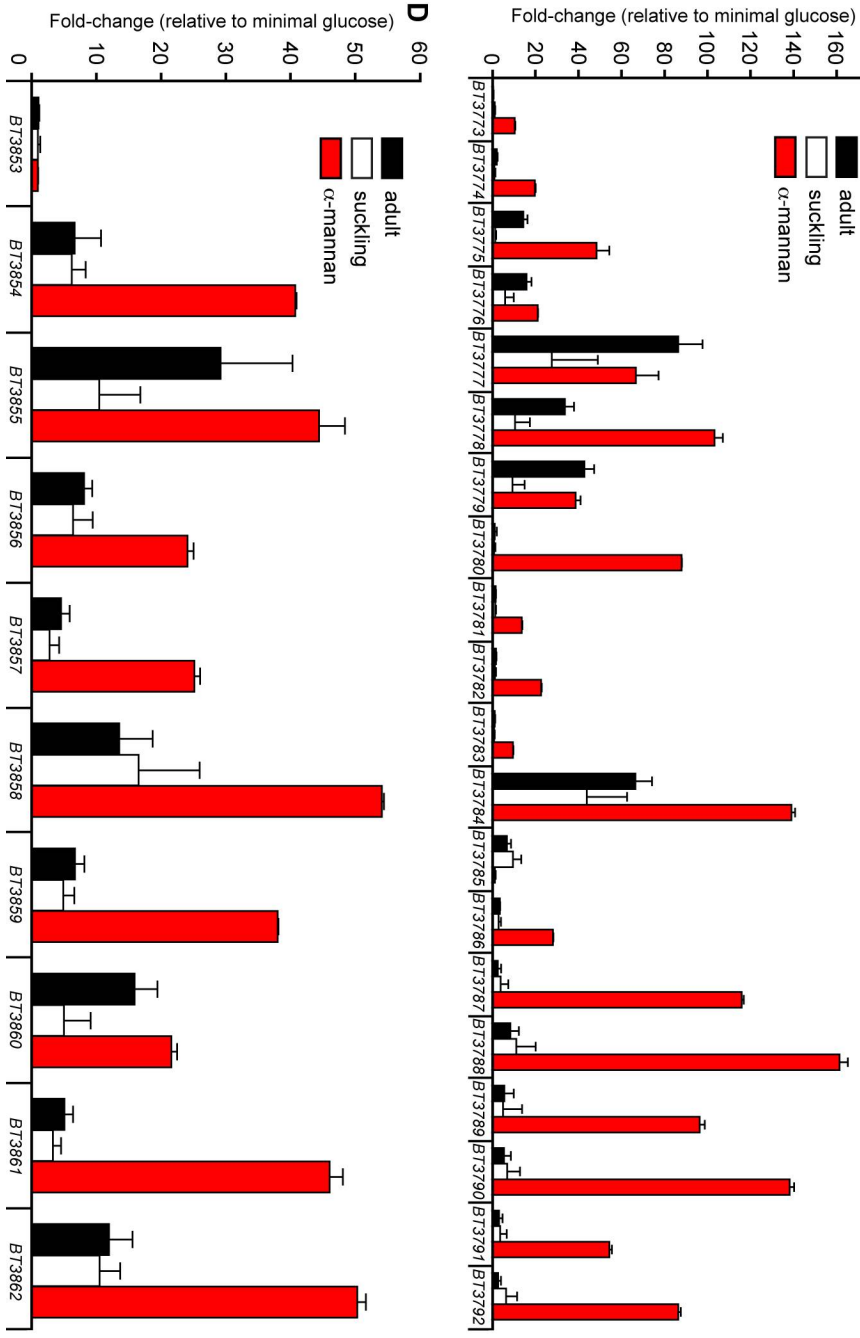
Martens *et al.*, Figure S7



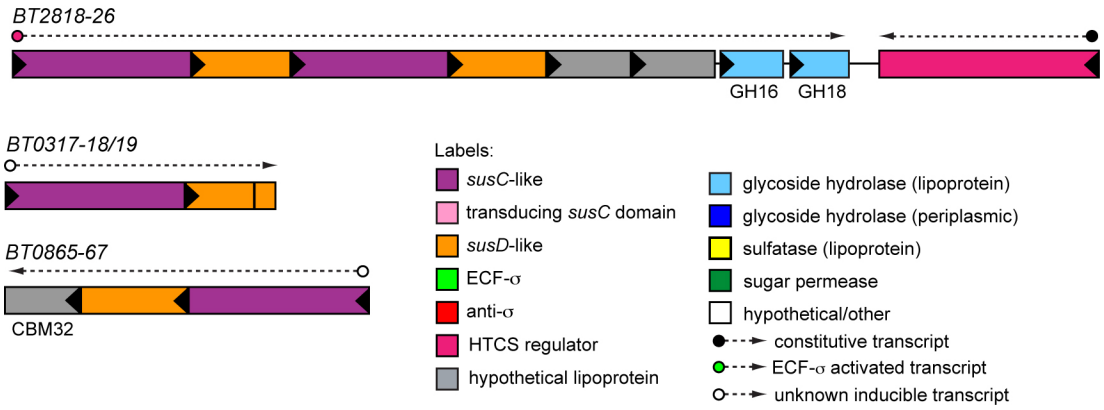
Martens *et al.*, Figure S8



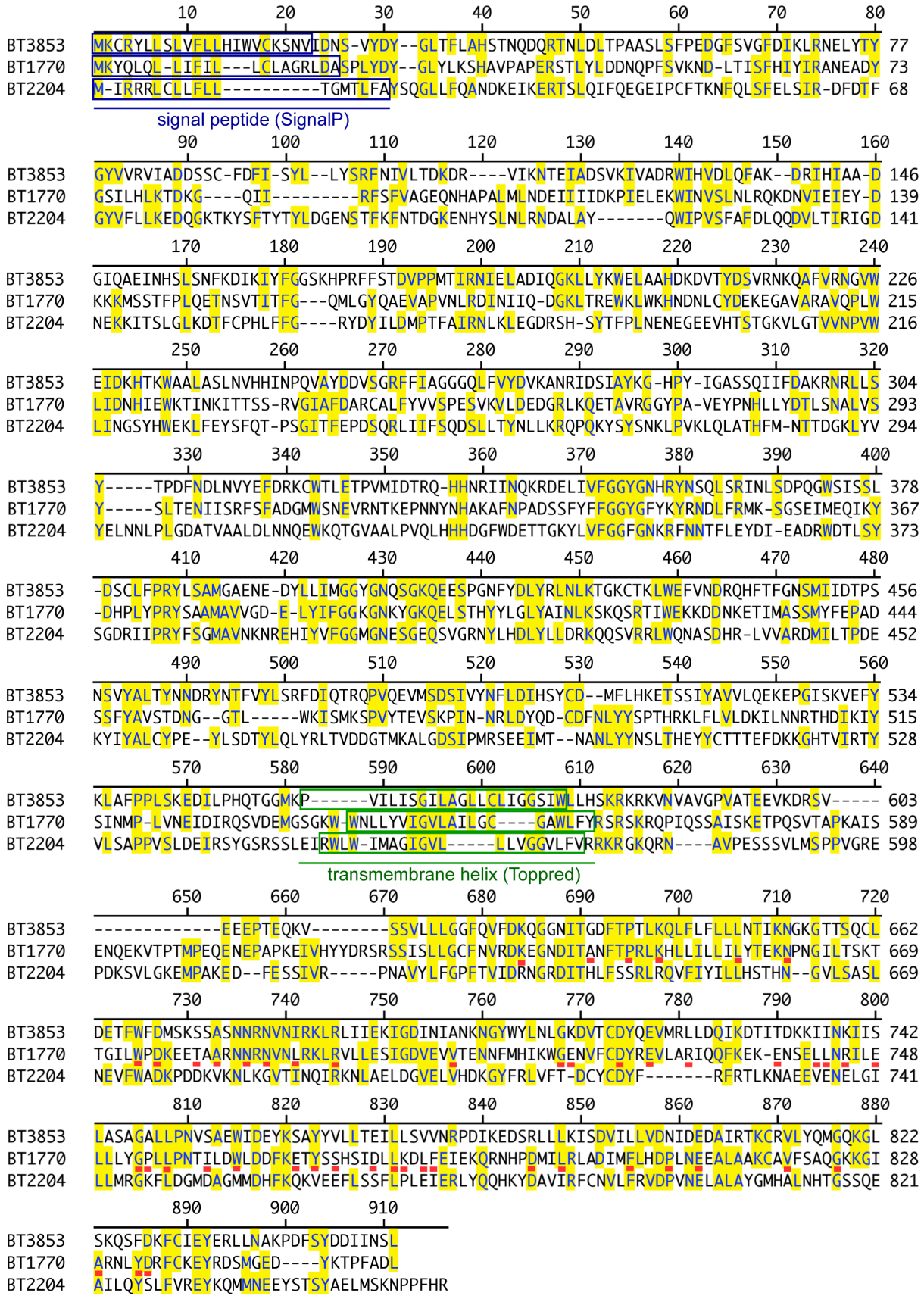
Martens *et al.*, Figure S8 (continued)



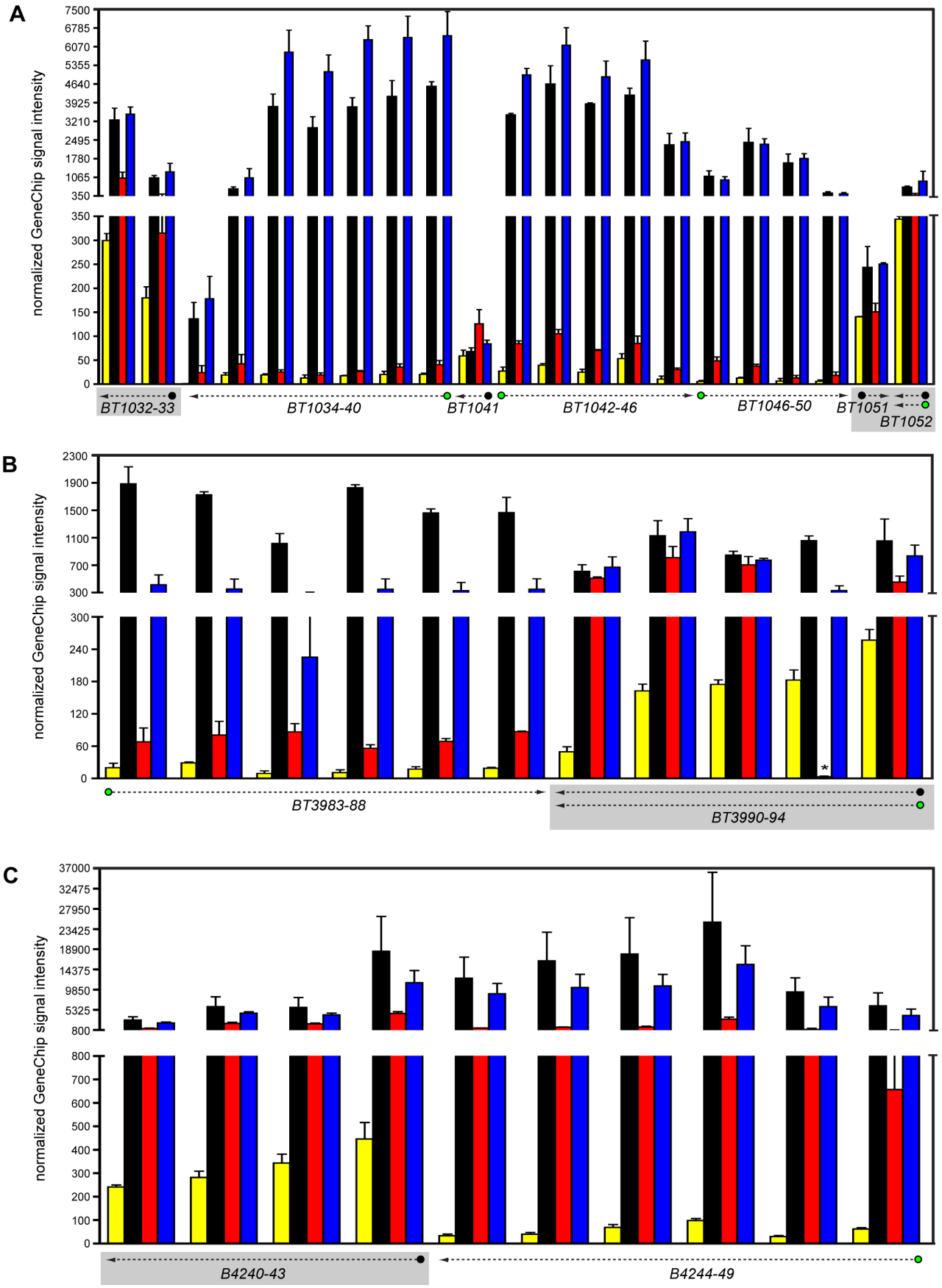
Martens *et al.*, Figure S9



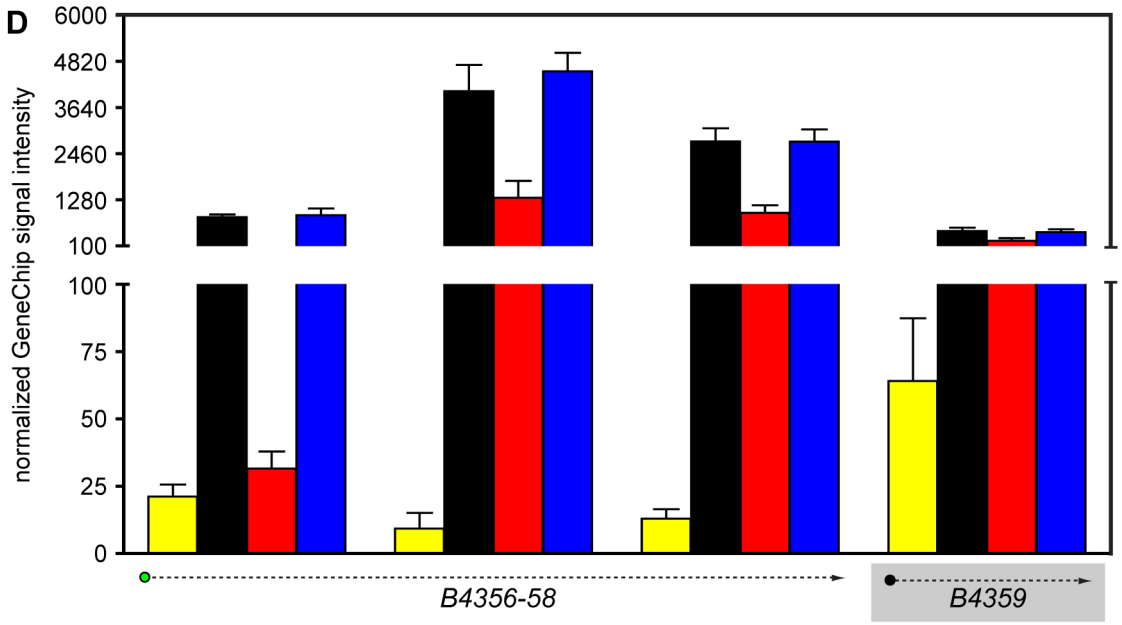
Martens *et al.*, Figure S10

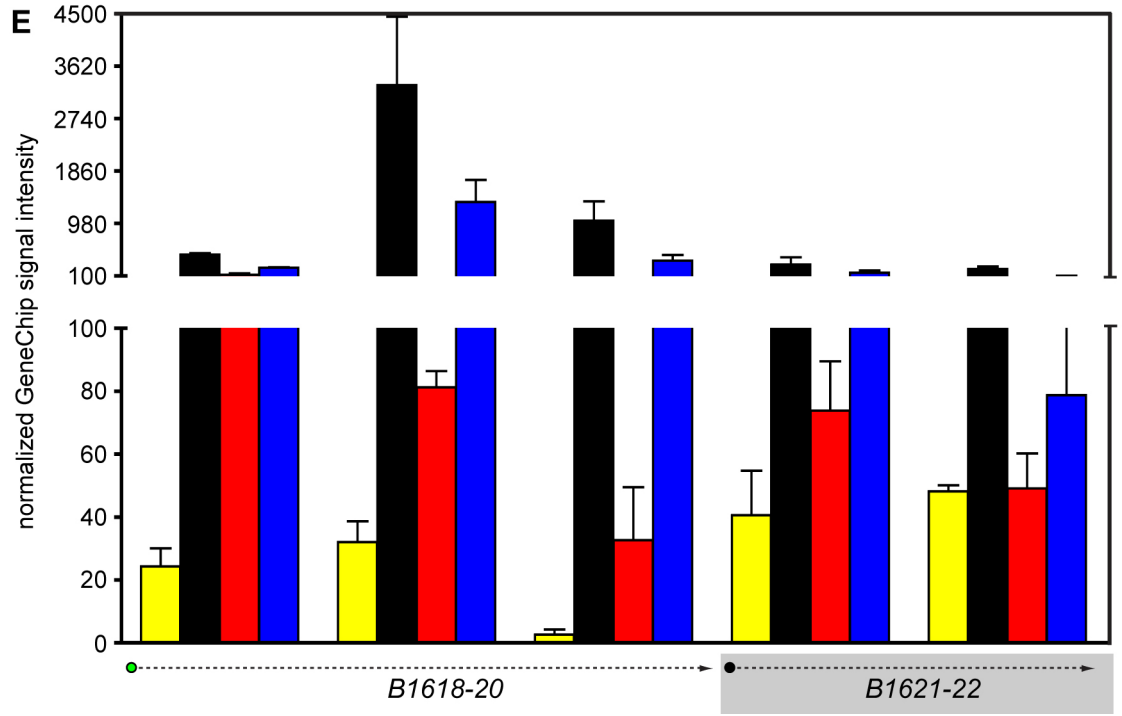


Martens *et al.*, Figure S11



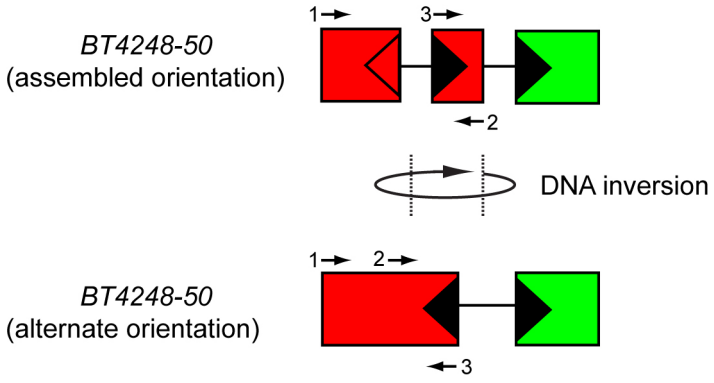
Martens *et al.*, Figure S11 (continued)



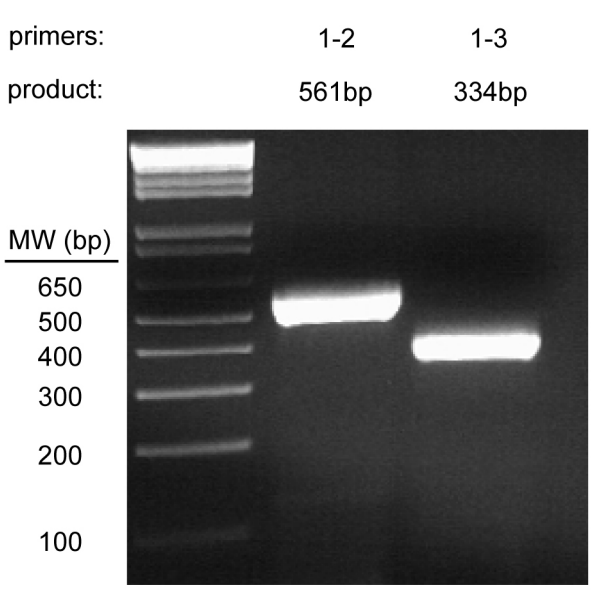


Martens *et al.*, Figure S12

A



B



Martens *et al.*, Figure S13