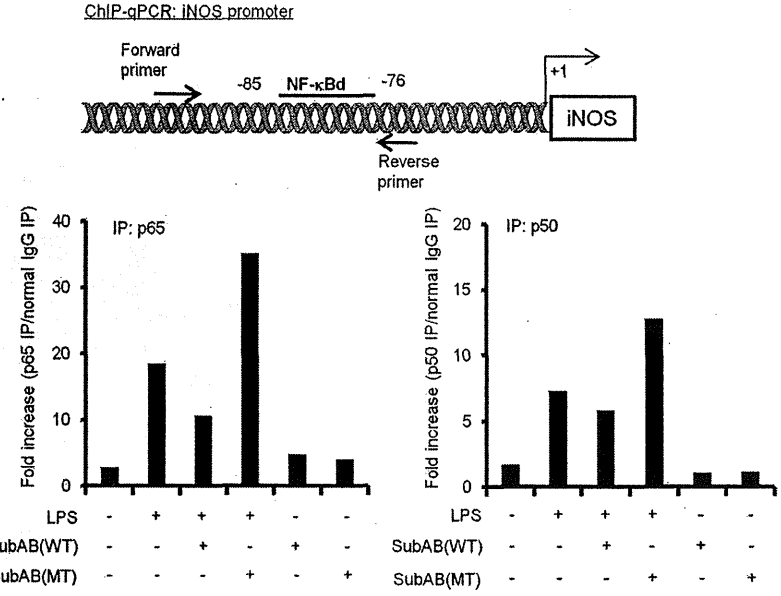



FIG 5 SubAB suppresses LPS-induced NF-κB p65/p50 heterodimer binding to the iNOS promoter. RAW 264.7 cells ($2 \times 10^6$ cells/dish) were treated with or without LPS (10 μg/ml) in the presence or absence of SubAB (WT; 0.5 μg/ml) or mSubAB (MT; 0.5 μg/ml) for 4 h. Cells were fixed with formaldehyde, lysed in SDS buffer, and then sonicated to fragment the chromatin. Immunoprecipitation was performed using anti-p65 (p65 IP) or anti-p50 (p50 IP) antibodies or normal rabbit IgG. Normal IgG IP refers to immunoprecipitation of chromatin with a nonimmune IgG. After purification of immunoprecipitated DNA, targeted promoter regions of iNOS were amplified by qPCR. Results are expressed as the percent input for each ChIP fraction, and fold changes were calculated to normal IgG IP. Data are representative of three independent experiments.

2 h or 16 h. During the initial period of infection for 2 h, similar number of *E. coli* were taken up by RAW 264.7 cells in the presence or absence of PBS control, L-NAME, SubAB, or mSubAB ($1.4 \times 10^6$, $1.6 \times 10^6$, $1.6 \times 10^6$, or $1.6 \times 10^6$ CFU/well, respectively). As shown in Fig. 7A, treatment of L-NAME and SubAB, but not mSubAB, increased intracellular survival of *E. coli* in RAW 264.7 cells at 16 h. *E. coli*-induced NO production by RAW 264.7 cells was reduced by L-NAME and SubAB but not mSubAB. Next, we examined the effect of expression of SubAB or mSubAB on *E. coli* BL21(DE3) survival within RAW 264.7 cells. During the initial 2 h of infection with BL21/pET-SubAB or BL21/pET-mSubAB, the same number of the *E. coli* CFU ($0.7 \times 10^6$ CFU/well) were taken up by RAW 264.7 cells. As shown in Fig. 7B, expression of SubAB by *E. coli*, but not that of mSubAB, increased intracellular survival in RAW 264.7 cells (top) and inhibited *E. coli*-induced iNOS expression and NO production (bottom) at 16 h. These results suggest that NO generated by macrophages results in killing of the bacteria and SubAB contributes to *E. coli* survival in macrophages.

## DISCUSSION

The reactive free radical, NO, is a major inflammatory mediator which is involved in host defense against bacterial infection and is produced following activation of phagocytes, such as neutrophils and macrophages (2, 28). In this study, LPS-induced NO production by RAW 264.7 cells was strongly inhibited by SubAB but not by mSubAB (Fig. 1). Because iNOS expression is responsible for LPS-induced NO production in macrophages (25, 29, 56, 57), we first investigated the effect of SubAB on LPS-induced iNOS protein and mRNA expression. In agreement with a previous report (25), iNOS mRNA and protein expressions were significantly induced within 4 h after treatment with LPS. SubAB, but not mSubAB, inhibited LPS-induced iNOS expression by an ER stress-mediated mechanism (Fig. 2B to D). These results indicate that SubAB inhibited LPS-induced NO production by RAW 264.7 cells through ER stress-mediated inhibition of iNOS expression. Lowenstein et al. (29) and Xie et al. (57) identified the murine

(10 μg/ml) for 30 min. Cells were fixed for immunofluorescence staining with anti-NF-κB p65 (red) or anti-NF-κB p50 (green) antibodies as described in Materials and Methods. A merged picture shows colocalization in RAW 264.7 cells. The nuclei were stained with DAPI. Solid arrows show colocalization of NF-κB p65 and p50. Bars represent 20 μm. Experiments were repeated two times with similar results. (C) RAW 264.7 cells were transiently transfected with HIV-NF-κB luciferase reporter construct, as well as the reference plasmid phRL-TK. Cells were treated with LPS (10 μg/ml) in the presence or absence of SubAB (WT; 0.5 μg/ml) or mSubAB (MT; 0.5 μg/ml) for 4, 8, or 24 h. Relative changes in luciferase expression were measured as described in Materials and Methods. Luciferase activity was normalized to *Renilla* luciferase activity. Results are expressed as the fold induction of luciferase activities in PBS-treated cells. Data are means ± SD of values from three independent experiments, with $n = 3$ per experiment. Statistical significance: *, $P < 0.01$; **, $P < 0.05$; N.S., not significant. (D) RAW 264.7 cells ($5 \times 10^4$ cells/well) were incubated with LPS (10 μg/ml) in the presence or absence of SubAB (WT; 0.5 μg/ml) or mSubAB (MT; 0.5 μg/ml) for the indicated times. Cell lysates were analyzed by immunoblotting using anti-phospho-IκB α, anti-α-tubulin, or anti-BiP/Grp78 antibodies (top). Data are representative of three independent experiments. IκB α phosphorylation was normalized to α-tubulin using densitometry. Relative densities of phospho-IκB α (phospho-IκB/tubulin) were compared with densities obtained at 0 min for each of three treated cells (bottom). Data are means ± SD of values from triplicate experiments. **, $P < 0.05$.

Downloaded from http://iai.asm.org/ on October 10, 2012 by CHIBA UNIVERSITY

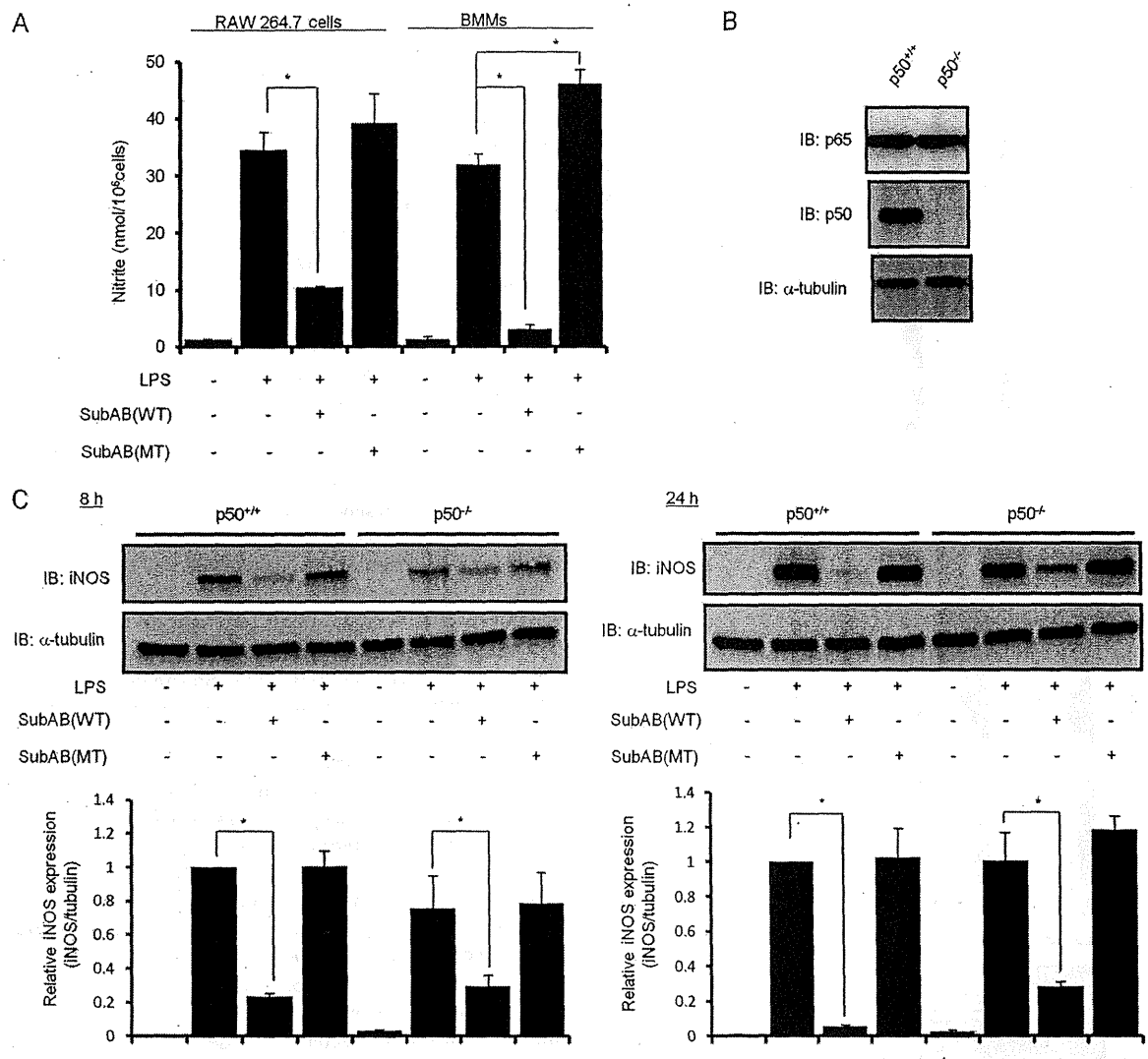



FIG 6 NF-κB p50 is involved in SubAB-mediated inhibition of LPS-induced iNOS expression. (A) RAW 264.7 cells ($5 \times 10^4$ cells/well) in a 48-well dish were grown in DMEM-1% FBS for 24 h. Mouse BMMs ($5 \times 10^4$ cells/well) in a 48-well dish were grown in DMEM-1% FBS containing 20 ng/ml M-CSF for 24 h. Cells were treated with LPS (10 μg/ml) in the presence or absence of SubAB (WT; 0.5 μg/ml) or mSubAB (MT; 0.5 μg/ml) for 24 h. The accumulation of nitrite in culture supernatants was quantified by Griess assay as described in the legend to Fig. 1. Data are means ± SD of values from triplicate experiments. *, $P < 0.01$. (B) Wild-type (NF-κB1$^{+/+}$) or NF-κB1 knockout (NF-κB1$^{-/-}$) BMMs ($5 \times 10^4$ cells/well) in a 48-well dish were grown in DMEM-1% FBS containing 20 ng/ml M-CSF for 24 h. Cell lysates were analyzed by immunoblotting with anti-p65, anti-p50, or anti-α-tubulin antibodies. (C) Wild-type (NF-κB1$^{+/+}$) or NF-κB1 knockout (NF-κB1$^{-/-}$) BMMs ($5 \times 10^4$ cells/well) in a 48-well dish were grown in DMEM-1% FBS containing 20 ng/ml M-CSF for 24 h. Cells were treated with LPS (10 μg/ml) in the presence or absence of SubAB (WT; 0.5 μg/ml) or mSubAB (MT; 0.5 μg/ml) for 8 and 24 h. Cell lysates were analyzed by immunoblotting with anti-iNOS or anti-α-tubulin antibodies. iNOS expression was normalized to α-tubulin using densitometry (iNOS/tubulin) and is depicted in bar graphs. Data are means ± SD of values from triplicate experiments. *, $P < 0.01$.

iNOS promoter sequence and showed that *E. coli*-derived LPS induced iNOS transcriptional activation. The published sequences of iNOS promoter in murine cells exhibit homologies to binding sites for numerous transcription factors (i.e., NF-κB, STAT-1, AP-1, C/EBP, CREB, IRF-1, Oct-1), which were known to be involved in the LPS- and cytokine-mediated induction of transcription (28, 29, 57). Our results from reporter assays using mouse iNOS promoter-luciferase reporter plasmid or iNOS promoter-deletion plasmids showed that the −93 bp region from the transcription initiation site of iNOS promoter, which contains an NF-κB-binding site, is responsible for SubAB-mediated inhibition of iNOS (Fig. 3). These findings suggest that SubAB controls LPS-induced NF-κB nuclear translocation and binding to iNOS promoter.

NF-κB is known to be a primary regulator of iNOS expression by an LPS-dependent pathway (29, 56, 57). Several reports showed that ER stress activates NF-κB (6, 23) but can also inhibit inflammatory responses in association with suppression of the NF-κB pathway (8, 19–21, 49). A recent study showed that SubAB pretreatment of macrophages inhibited LPS-induced NF-κB activation and MCP-1 and TNF-α production through inhibition of IκB-α degradation (17). Consistent with these studies, our results demonstrated that LPS-induced NF-κB nuclear translocation and activation were blocked by SubAB (Fig. 4A to C). Although these

Downloaded from http://iai.asm.org/ on October 10, 2012 by CHIBA UNIVERSITY

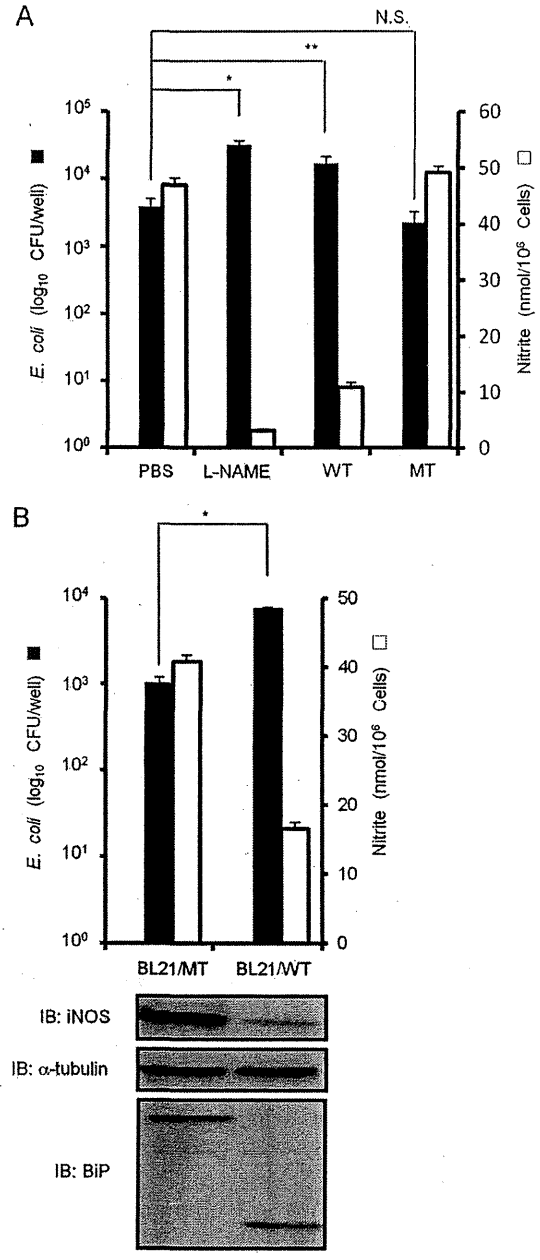



FIG 7 SubAB contributes to *E. coli* survival in RAW 264.7 cells. (A) RAW 264.7 cells ($1 \times 10^6$ cells/well) in a 24-well dish were grown in antibiotic-free DMEM-10% FBS for 24 h. Cells were infected with *E. coli* BL21/pET23b in the presence or absence of L-NAME (10 mM), SubAB (WT; 0.5 μg/ml), or mSubAB (MT; 0.5 μg/ml) as described in Materials and Methods. Intracellular bacteria were quantified 16 h later. Data are means ± SD of values from triplicate experiments. *, $P < 0.01$; N.S., not significant. (B) RAW 264.7 cells ($1 \times 10^6$ cells/well) in a 24-well dish were infected with *E. coli* BL21/pET-SubAB (BL21/WT) or BL21/pET-mSubAB (BL21/MT) as described above. After 16 h, intracellular bacteria were quantified by CFU (top). Cell lysates were analyzed by immunoblotting using anti-iNOS, anti-α-tubulin, or anti-BiP/Grp78 antibodies (bottom). Data are means ± SD of values from triplicate experiments. *, $P < 0.01$.

data raise the possibility that SubAB inhibits IκB degradation by effects on proteasomal function or IKK-mediated IκB phosphorylation, we now present evidence that SubAB treatment of RAW 264.7 cells suppressed LPS-induced IκB-α phosphorylation, a trigger for its degradation (Fig. 4D).

In addition, we observed that mSubAB treated in RAW 264.7 cells slightly enhanced LPS-induced NO production (Fig. 1A), iNOS mRNA expression (Fig. 2D), and iNOS transcription (Fig. 3). These results suggest that mSubAB binding to the target cells induced activation of an NF-κB signaling pathway. Our previous study demonstrated that SubAB binds to functional receptors, i.e., L1CAM, α2β1 integrin, and Met (60). Interestingly, these membrane proteins are known to modulate NF-κB activity (4, 14, 48). Thus, SubAB binding to these receptors could stimulate NF-κB signaling, followed by increased LPS-induced iNOS expression. However, the stimulatory effect of toxin interaction with its receptors is diminished by SubAB-induced ER stress after BiP cleavage, suggesting that NF-κB activation by toxin binding does not affect iNOS expression in the presence of wild-type SubAB.

NF-κB subunits can form various dimers, but the classical, best-characterized complex is composed of p50 and p65 (1, 50, 53). Mouse iNOS promoter contains two NF-κB-binding sites, NF-κBu site (−971 to −962) and NF-κBd site (−85 to −76), which play an important role in iNOS expression. NF-κB subunits (p65, p50, p52, and c-Rel) bind to the NF-κBd site in response to LPS stimulation. In particular, p50 may have a more pivotal role in NF-κB transactivation activity at the NF-κBd site since p50 binding was demonstrated by electrophoretic mobility shift assay (EMSA) (7, 56). NF-κB p50 displayed higher-affinity DNA binding than p65 (12), suggesting that p50 is an essential DNA-binding subunit of the p65/p50 heterodimer. In NF-κB1$^{-/-}$ BMMs, SubAB-mediated iNOS inhibition was partially abolished (Fig. 6C), indicating that the p65/p50 heterodimer is, at least, involved in SubAB-mediated iNOS inhibition. However, it is still possible that SubAB may have an inhibitory effect on other NF-κB dimers composed of p65, c-Rel, and p52 or other transcription factors which induce iNOS expression by LPS.

A number of pathogenic bacteria have developed a variety of mechanisms to evade host defense mechanisms and enhance their virulence (44). Previous studies showed that Stx or other bacterial factors of STEC, such as EspA, B, and D, downregulate NF-κB activation (15, 18). In addition, a recent study showed that an undefined effector of O157 inhibited iNOS expression (52). Therefore, to exclude the effect of the iNOS inhibitors, we used nonpathogenic *E. coli* strain BL21(DE3) for macrophage infection. Inhibition of NO production by treatment with NOS inhibitor L-NAME resulted in increased *E. coli* survival in RAW 264.7 cells (Fig. 7A), indicating that NO acts as a intracellular antibacterial molecule. Further, treatment with purified SubAB or expression of SubAB by *E. coli* also increased the intracellular survival of *E. coli* in RAW 264.7 cells (Fig. 7A and B). NO is a key component of the host immune response and is encountered by pathogenic bacteria during their lives outside and within the host (11). The effects of NO and ROS can be synergistic and efficiently kill microorganisms by causing double-stranded DNA breakage, bacterial $Fe^{2+}$ release, and depletion of the antioxidant glutathione (31, 40, 55). The reaction of $O_2^-$ and NO leads to the generation of peroxynitrite ($ONOO^-$), which can decompose to species with an extremely potent oxidizing potential, capable of killing *E. coli* (40). Thus, our findings suggested that NO represents an im-

Downloaded from http://iai.asm.org/ on October 10, 2012 by CHIBA UNIVERSITY

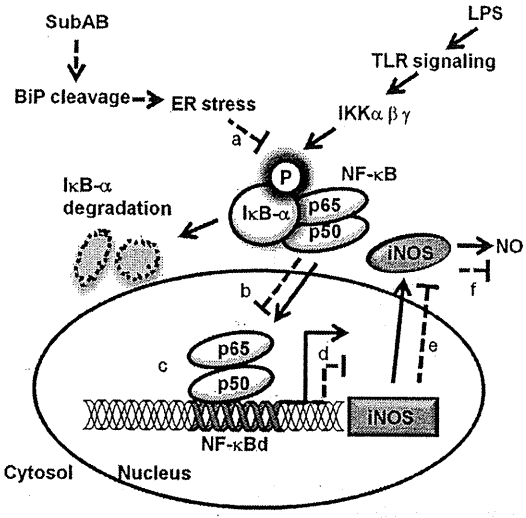



FIG 8 Proposed model of the inhibitory effect of SubAB on NO production in macrophages. SubAB-mediated inhibition of LPS-induced iNOS can be summarized as follows. The solid lines represent mechanisms of LPS-induced NO production. LPS induces IκB-α phosphorylation through a TLR-signaling pathway, followed by IκB-α degradation and NF-κB p65/p50 heterodimer nuclear translocation. Binding of NF-κB p65/p50 heterodimer to the NF-κBd site in the iNOS promoter is responsible for iNOS expression and NO production. The dashed lines show inhibitory effects of SubAB on LPS-induced NO production. First, SubAB cleaves BiP/Grp78, leading to ER stress, which suppresses IκB-α phosphorylation (a). Suppression of IκB-α phosphorylation prevents IκB-α degradation and NF-κB p65/p50 heterodimer translocation to nuclei (b), followed by reduction of NF-κB p65/p50 heterodimer binding to the NF-κBd site in iNOS promoter (c), resulting in inhibition of iNOS mRNA expression (d). Inhibition of LPS-induced iNOS mRNA expression by SubAB suppresses iNOS protein expression (e). Finally, SubAB-mediated inhibition of LPS-induced iNOS expression leads to inhibition of NO production by macrophages (f), which permits survival of intracellular *E. coli*.

portant antimicrobial agent of special significance to *E. coli*, and reduction of NO production by SubAB results in increased STEC viability and infectivity. Moreover, a recent study reported that NO inhibits Stx synthesis in STEC (47, 51). Thus, increasing NO production in infected patients might represent an alternative strategy to limit the development of HUS. This finding suggests that reduction of NO production during the host immune response may result in Stx synthesis by STEC and maintenance of STEC virulence. Taken together, SubAB by inhibiting NO production may enhance STEC pathogenicity.

LPS is a major component of the outer membrane of Gram-negative bacteria, such as STEC. LPS interacts with TLR4 on macrophages during STEC infection, leading to increased phagocytosis (5), production of cytokines, and/or ROS/RNS generation (25). Interestingly, posttreatment with SubAB after LPS stimulation for 3 h also significantly inhibited NO production (Fig. 1C). These findings further support our proposal that SubAB-producing STEC can inhibit NO production effectively during infection.

In conclusion, we provide molecular mechanisms for the inhibitory effect of SubAB on LPS-induced NO production in mouse macrophages as shown in Fig. 8. This inhibition occurs through a process involving SubAB-mediated reduction of p65/p50 heterodimer binding to the NF-κB binding site on iNOS promoter. As NO is an essential factor for antimicrobial host defense in *E. coli* infections, we propose that SubAB might be involved in the upregulation of STEC virulence. Future studies involving macrophage infection with isogenic SubAB-deficient STEC or an animal infection model will help to determine the apparent role of SubAB in STEC survival.

## ACKNOWLEDGMENTS

This work was supported by grants in aid for scientific research from the Ministry of Education, Science, and Culture of Japan and Improvement of Research Environment for Young Researchers from Japan Science and Technology Agency. Joel Moss was supported by the Intramural Research Program, National Institutes of Health, National Heart, Lung, and Blood Institute.

We thank T. Hirayama (Institute of Tropical Medicine, Nagasaki University) and N. Morinaga (Chiba University) for helpful discussions and suggestions. We acknowledge the expert technical assistance of D. Nakagomi, C. Noritake, and A. Kiuchi (Chiba University).

## REFERENCES

1. Baeuerle PA, Baltimore D. 1996. NF-kappa B: ten years after. Cell 87:13–20.
2. Bogdan C. 2001. Nitric oxide and the immune response. Nat. Immunol. 2:907–916.
3. Brunelli L, Crow JP, Beckman JS. 1995. The comparative toxicity of nitric oxide and peroxynitrite to *Escherichia coli*. Arch. Biochem. Biophys. 316:327–334.
4. Chung CH, Lin KT, Chang CH, Peng HC, Huang TF. 2009. The integrin alpha2beta1 agonist, aggretin, promotes proliferation and migration of VSMC through NF-kB translocation and PDGF production. Br. J. Pharmacol. 156:846–856.
5. Cooper PH, Mayer P, Baggiolini M. 1984. Stimulation of phagocytosis in bone marrow-derived mouse macrophages by bacterial lipopolysaccharide: correlation with biochemical and functional parameters. J. Immunol. 133:913–922.
6. Deng J, et al. 2004. Translational repression mediates activation of nuclear factor kappa B by phosphorylated translation initiation factor 2. Mol. Cell. Biol. 24:10161–10168.
7. Deng WG, Tang ST, Tseng HP, Wu KK. 2006. Melatonin suppresses macrophage cyclooxygenase-2 and inducible nitric oxide synthase expression by inhibiting p52 acetylation and binding. Blood 108:518–524.
8. Du S, et al. 2009. Suppression of NF-kappaB by cyclosporin A and tacrolimus (FK506) via induction of the C/EBP family: implication for unfolded protein response. J. Immunol. 182:7201–7211.
9. Ducut Sigala JL, et al. 2004. Activation of transcription factor NF-kappaB requires ELKS, an IkappaB kinase regulatory subunit. Science 304:1963–1967.
10. Fagerlund R, Kinnunen L, Kohler M, Julkunen I, Melen K. 2005. NF-κB is transported into the nucleus by importin α3 and importin α4. J. Biol. Chem. 280:15942–15951.
11. Fang FC. 2004. Antimicrobial reactive oxygen and nitrogen species: concepts and controversies. Nat. Rev. Microbiol. 2:820–832.
12. Fujita T, Nolan GP, Ghosh S, Baltimore D. 1992. Independent modes of transcriptional activation by the p50 and p65 subunits of NF-kappa B. Genes Dev. 6:775–787.
13. Furukawa T, et al. 2011. Fatal hemorrhage induced by subtilase cytotoxin from Shiga-toxigenic *Escherichia coli*. Microb. Pathog. 50:159–167.
14. Gavert N, Ben-Shmuel A, Lemmon V, Brabletz T, Ben-Ze'ev A. 2010. Nuclear factor-kappaB signaling and ezrin are essential for L1-mediated metastasis of colon cancer cells. J. Cell Sci. 123:2135–2143.
15. Gobert AP, et al. 2007. Shiga toxin produced by enterohemorrhagic *Escherichia coli* inhibits PI3K/NF-kappaB signaling pathway in globotriaosylceramide-3-negative human intestinal epithelial cells. J. Immunol. 178:8168–8174.
16. Guilbert LJ, Stanley ER. 1986. The interaction of 125I-colony-stimulating factor-1 with bone marrow-derived macrophages. J. Biol. Chem. 261:4024–4032.
17. Harama D, et al. 2009. A subcytotoxic dose of subtilase cytotoxin prevents lipopolysaccharide-induced inflammatory responses, depending on its capacity to induce the unfolded protein response. J. Immunol. 183:1368–1374.
18. Hauf N, Chakraborty T. 2003. Suppression of NF-kappa B activation and

Downloaded from http://iai.asm.org/ on October 10, 2012 by CHIBA UNIVERSITY

proinflammatory cytokine expression by Shiga toxin-producing *Escherichia coli*. J. Immunol. 170:2074–2082.
19. Hayakawa K, et al. 2009. Acquisition of anergy to proinflammatory cytokines in nonimmune cells through endoplasmic reticulum stress response: a mechanism for subsidence of inflammation. J. Immunol. 182: 1182–1191.
20. Hayakawa K, et al. 2008. Blunted activation of NF-kappaB and NF-kappaB-dependent gene expression by geranylgeranylacetone: involvement of unfolded protein response. Biochem. Biophys. Res. Commun. 365:47–53.
21. Hayakawa K, et al. 2010. ER stress depresses NF-kappaB activation in mesangial cells through preferential induction of C/EBP beta. J. Am. Soc. Nephrol. 21:73–81.
22. Hiramatsu Y, et al. 2010. c-Maf activates the promoter and enhancer of the IL-21 gene, and TGF-beta inhibits c-Maf-induced IL-21 production in CD4+ T cells. J. Leukoc. Biol. 87:703–712.
23. Hu P, Han Z, Couvillon AD, Kaufman RJ, Exton JH. 2006. Autocrine tumor necrosis factor alpha links endoplasmic reticulum stress to the membrane death receptor pathway through IRE1alpha-mediated NF-kappaB activation and down-regulation of TRAF2 expression. Mol. Cell. Biol. 26:3071–3084.
24. Huang T, et al. 2009. Downregulation of gap junction expression and function by endoplasmic reticulum stress. J. Cell Biochem. 107:973–983.
25. Jacobs AT, Ignarro LJ. 2001. Lipopolysaccharide-induced expression of interferon-beta mediates the timing of inducible nitric-oxide synthase induction in RAW 264.7 macrophages. J. Biol. Chem. 276:47950–47957.
26. Karin M, Ben-Neriah Y. 2000. Phosphorylation meets ubiquitination: the control of NF-[kappa]B activity. Annu. Rev. Immunol. 18:621–663.
27. Karmali MA. 1989. Infection by verocytotoxin-producing *Escherichia coli*. Clin. Microbiol. Rev. 2:15–38.
28. Kleinert H, Pautz A, Linker K, Schwarz PM. 2004. Regulation of the expression of inducible nitric oxide synthase. Eur. J. Pharmacol. 500:255–266.
29. Lowenstein CJ, et al. 1993. Macrophage nitric oxide synthase gene: two upstream regions mediate induction by interferon gamma and lipopolysaccharide. Proc. Natl. Acad. Sci. U. S. A. 90:9730–9734.
30. Lowenstein CJ, Snyder SH. 1992. Nitric oxide, a novel biologic messenger. Cell 70:705–707.
31. Marcinkiewicz J. 1997. Nitric oxide and antimicrobial activity of reactive oxygen intermediates. Immunopharmacology 37:35–41.
32. Matsuura G, et al. 2009. Novel subtilase cytotoxin produced by Shiga-toxigenic *Escherichia coli* induces apoptosis in Vero cells via mitochondrial membrane damage. Infect. Immun. 77:2919–2924.
33. Morinaga N, Yahiro K, Matsuura G, Moss J, Noda M. 2008. Subtilase cytotoxin, produced by Shiga-toxigenic *Escherichia coli*, transiently inhibits protein synthesis of Vero cells via degradation of BiP and induces cell cycle arrest at $G_1$ by downregulation of cyclin D1. Cell Microbiol. 10:921–929.
34. Morinaga N, et al. 2007. Two distinct cytotoxic activities of subtilase cytotoxin produced by Shiga-toxigenic *Escherichia coli*. Infect. Immun. 75:488–496.
35. Morita H, et al. 2004. Antimicrobial action against verotoxigenic *Escherichia coli* O157:H7 of nitric oxide derived from sodium nitrite. Biosci. Biotechnol. Biochem. 68:1027–1034.
36. Morrison DC, Jacobs DM. 1976. Binding of polymyxin B to the lipid A portion of bacterial lipopolysaccharides. Immunochemistry 13:813–818.
37. Nakajima S, et al. 2011. Selective abrogation of BiP/GRP78 blunts activation of NF-kappaB through the ATF6 branch of the UPR: involvement of C/EBPbeta and mTOR-dependent dephosphorylation of Akt. Mol. Cell. Biol. 31:1710–1718.
38. Nakamura Y, et al. 2006. Potentiation by high potassium of lipopolysaccharide-induced nitric oxide production from cultured astrocytes. Neurochem. Int. 48:43–49.
39. Okamoto T, et al. 2010. A new paradigm for antimicrobial host defense mediated by a nitrated cyclic nucleotide. J. Clin. Biochem. Nutr. 46:14–19.
40. Pacelli R, et al. 1995. Nitric oxide potentiates hydrogen peroxide-induced killing of *Escherichia coli*. J. Exp. Med. 182:1469–1479.
41. Paton AW, et al. 2006. AB5 subtilase cytotoxin inactivates the endoplasmic reticulum chaperone BiP. Nature 443:548–552.
42. Paton AW, Srimanote P, Talbot UM, Wang H, Paton JC. 2004. A new family of potent AB(5) cytotoxins produced by Shiga toxigenic *Escherichia coli*. J. Exp. Med. 200:35–46.
43. Paton JC, Paton AW. 1998. Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. Clin. Microbiol. Rev. 11:450–479.
44. Poirier K, et al. 2008. *Escherichia coli* O157:H7 survives within human macrophages: global gene expression profile and involvement of the Shiga toxins. Infect. Immun. 76:4814–4822.
45. Schreiber E, Matthias P, Muller MM, Schaffner W. 1989. Rapid detection of octamer binding proteins with 'mini-extracts,' prepared from a small number of cells. Nucleic Acids Res. 17:6419.
46. Sha WC, Liou HC, Tuomanen EI, Baltimore D. 1995. Targeted disruption of the p50 subunit of NF-kappa B leads to multifocal defects in immune responses. Cell 80:321–330.
47. Shimizu T, Tsutsuki H, Matsumoto A, Nakaya H, Noda M. 2012. The nitric oxide reductase of enterohaemorrhagic *Escherichia coli* plays an important role for the survival within macrophages. Mol. Microbiol. 85:492–512.
48. Tacchini L, De Ponti C, Matteucci E, Follis R, Desiderio MA. 2004. Hepatocyte growth factor-activated NF-kappaB regulates HIF-1 activity and ODC expression, implicated in survival, differently in different carcinoma cell lines. Carcinogenesis 25:2089–2100.
49. Takano Y, et al. 2007. Suppression of cytokine response by GATA inhibitor K-7174 via unfolded protein response. Biochem. Biophys. Res. Commun. 360:470–475.
50. Thanos D, Maniatis T. 1995. Identification of the *rel* family members required for virus induction of the human beta interferon gene. Mol. Cell. Biol. 15:152–164.
51. Vareille M, de Sablet T, Hindre T, Martin C, Gobert AP. 2007. Nitric oxide inhibits Shiga-toxin synthesis by enterohemorrhagic *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. 104:10199–10204.
52. Vareille M, et al. 2008. Heme oxygenase-1 is a critical regulator of nitric oxide production in enterohemorrhagic *Escherichia coli*-infected human enterocytes. J. Immunol. 180:5720–5726.
53. Verma IM, Stevenson JK, Schwarz EM, Van Antwerp D, Miyamoto S. 1995. Rel/NF-kappa B/I kappa B family: intimate tales of association and dissociation. Genes Dev. 9:2723–2735.
54. Wolfson JJ, et al. 2008. Subtilase cytotoxin activates PERK, IRE1 and ATF6 endoplasmic reticulum stress-signalling pathways. Cell Microbiol. 10:1775–1786.
55. Woodmansee AN, Imlay JA. 2003. A mechanism by which nitric oxide accelerates the rate of oxidative DNA damage in *Escherichia coli*. Mol. Microbiol. 49:11–22.
56. Xie QW, Kashiwabara Y, Nathan C. 1994. Role of transcription factor NF-kappa B/Rel in induction of nitric oxide synthase. J. Biol. Chem. 269: 4705–4708.
57. Xie QW, Whisnant R, Nathan C. 1993. Promoter of the mouse gene encoding calcium-independent nitric oxide synthase confers inducibility by interferon gamma and bacterial lipopolysaccharide. J. Exp. Med. 177: 1779–1784.
58. Yahiro K, Morinaga N, Moss J, Noda M. 2010. Subtilase cytotoxin induces apoptosis in HeLa cells by mitochondrial permeabilization via activation of Bax/Bak, independent of C/EBP-homologue protein (CHOP), Ire1alpha or JNK signaling. Microb. Pathog. 49:153–163.
59. Yahiro K, et al. 2006. Identification and characterization of receptors for vacuolating activity of subtilase cytotoxin. Mol. Microbiol. 62:480–490.
60. Yahiro K, et al. 2011. Identification of subtilase cytotoxin (SubAB) receptors whose signaling, in association with SubAB-induced BiP cleavage, is responsible for apoptosis in HeLa cells. Infect. Immun. 79:617–627.
61. Yamazaki H, et al. 2009. Activation of the Akt-NF-kappaB pathway by subtilase cytotoxin through the ATF6 branch of the unfolded protein response. J. Immunol. 183:1480–1487.
62. Yuhas Y, Kaminsky E, Mor M, Ashkenazi S. 1996. Induction of nitric oxide production in mouse macrophages by Shiga toxin. J. Med. Microbiol. 45:97–102.

Downloaded from http://iai.asm.org/ on October 10, 2012 by CHIBA UNIVERSITY

Takami et al. BMC Genomics 2012, 13:699
http://www.biomedcentral.com/1471-2164/13/699

BMC Genomics

METHODOLOGY ARTICLE Open Access

# Evaluation method for the potential functionome harbored in the genome and metagenome

Hideto Takami[1*], Takeaki Taniguchi[2], Yuki Moriya[3], Tomomi Kuwahara[4], Minoru Kanehisa[3] and Susumu Goto[3]

## Abstract

**Background:** One of the main goals of genomic analysis is to elucidate the comprehensive functions (functionome) in individual organisms or a whole community in various environments. However, a standard evaluation method for discerning the functional potentials harbored within the genome or metagenome has not yet been established. We have developed a new evaluation method for the potential functionome, based on the completion ratio of Kyoto Encyclopedia of Genes and Genomes (KEGG) functional modules.

**Results:** Distribution of the completion ratio of the KEGG functional modules in 768 prokaryotic species varied greatly with the kind of module, and all modules primarily fell into 4 patterns (universal, restricted, diversified and non-prokaryotic modules), indicating the universal and unique nature of each module, and also the versatility of the KEGG Orthology (KO) identifiers mapped to each one. The module completion ratio in 8 phenotypically different bacilli revealed that some modules were shared only in phenotypically similar species. Metagenomes of human gut microbiomes from 13 healthy individuals previously determined by the Sanger method were analyzed based on the module completion ratio. Results led to new discoveries in the nutritional preferences of gut microbes, believed to be one of the mutualistic representations of gut microbiomes to avoid nutritional competition with the host.

**Conclusions:** The method developed in this study could characterize the functionome harbored in genomes and metagenomes. As this method also provided taxonomical information from KEGG modules as well as the gene hosts constructing the modules, interpretation of completion profiles was simplified and we could identify the complementarity between biochemical functions in human hosts and the nutritional preferences in human gut microbiomes. Thus, our method has the potential to be a powerful tool for comparative functional analysis in genomics and metagenomics, able to target unknown environments containing various uncultivable microbes within unidentified phyla.

## Background

One of the main goals of genomic and metagenomic analyses is to extract the comprehensive functions (functionome) harbored in an individual organism or a whole community in various environments. However, evaluating the potential functionome is still difficult when compared with the functional annotation of individual genes or proteins; i.e. based on a similarity search against a reference database such as the NCBI-NR database of non-redundant protein sequences [1], usually employing a variant of the BLAST program [2], or on the protein domain search against a protein family database such as PFAM [3]. This is mainly because a standard methodology for extracting functional category information, such as individual metabolism, energy generation and transportation systems, has not yet been fully established. Traditionally, clusters of orthologous groups (COGs) have been used for functional classification of proteins, particularly in microbial genome sequencing projects. The COGs database provides 17 functional categories for orthologous groups in order to facilitate functional studies and serves as a platform for functional annotation of newly sequenced genomes and studies on genome evolution [4]. Although the COG functional categories are often used within Standards in Genomic Sciences (http://standardsingenomics.org/index.php/sigen) as a standard analysis, through combination with the Integrated Microbial Genomes (IMG) system [5], no large functional

* Correspondence: takamih@jamstec.go.jp
[1]Microbial Genome Research Group, Japan Agency for Marine-Earth Science & Technology (JAMSTEC), 2-15 Natsushima, Yokosuka 237-0061, Japan
Full list of author information is available at the end of the article

BioMed Central

© 2012 Takami et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

differences are usually observed in such broad categories; even between phenotypically different organisms (http://img.jgi.doe.gov/cgi-bin/w/main.cgi?section) and also whole microbial communities in different environments [6-8]. Thus, it is difficult to differentiate the functional potentials between different genomes and metagenomes by analysis based on COG classification.

Recently, more detailed and comprehensive functional categories facilitated in KEGG [9] and SEED [10] have been used for comparative genomics and as metagenomics tools to highlight functional features represented by KAAS (KEGG Automatic Annotation Server) [11], MG-RAST (Rapid Annotations using Subsystems Technology server for metagenomic project) [12] and MEGAN [13,14] (Figure 1). They all employ a similarity-based method for functional annotations, but utilize different databases for protein sequences, default threshold values and orthology IDs for mapping annotated sequences to functional categories depending on their desired outputs, namely pathways in KEGG or subsystems in SEED. Notably, KAAS has been applied to protein coding sequences from several metagenomic samples, and their annotated KEGG pathways and other classifications are already available (http://www.genome.jp/kegg/catalog/org_list3.html).

The outputs of these systems include functional distributions of each sample by hierarchical classification using KEGG and/or SEED and comparisons between several samples when necessary. However, it is still difficult to evaluate the functional potentials via the current classification systems (such as pathway map-based analysis) because the functional information from different organisms such as microbes, plants, and animals has been mixed up. On the other hand, KEGG MODULE, a database that collects pathway modules and other functional units, presents a promising tool for functional classification [16]. Pathway modules in KEGG MODULE are smaller pieces of subpathways, manually defined as consecutive reaction steps, operon or other regulatory units, and phylogenetic units obtained by genome comparisons (Figure 2A). This database also contains molecular complex modules, comprising multiple molecules such as the subunits of transporters and receptors, functional sets, and signature modules (Figure 2B). As of December 2011, 179 pathway modules have been defined for energy, carbohydrate, lipid, nucleotide, and amino acid metabolism, including genetic and environmental information processing pathways. In total, 434 KEGG modules (179

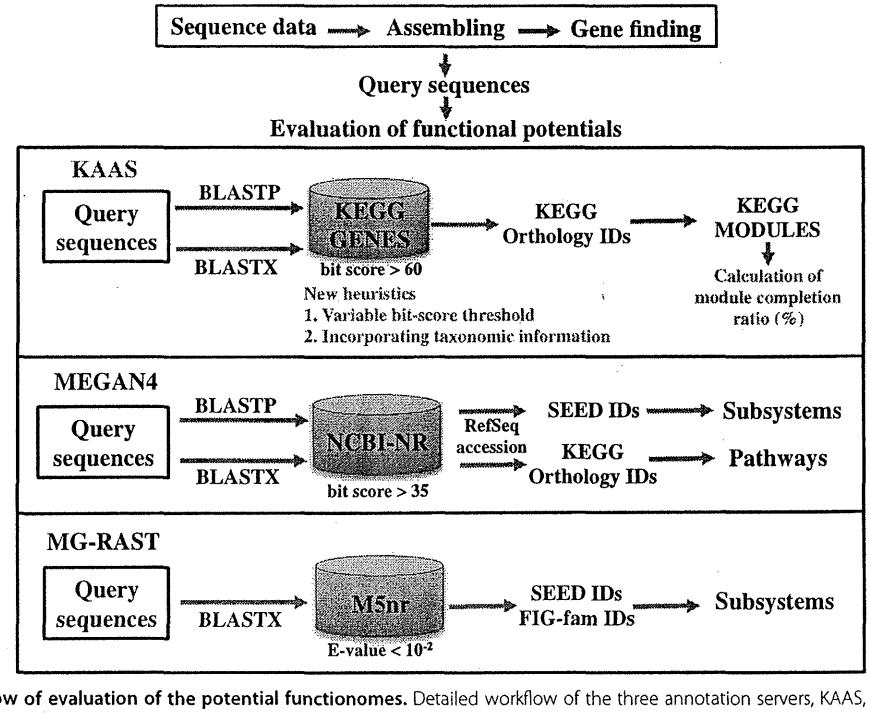
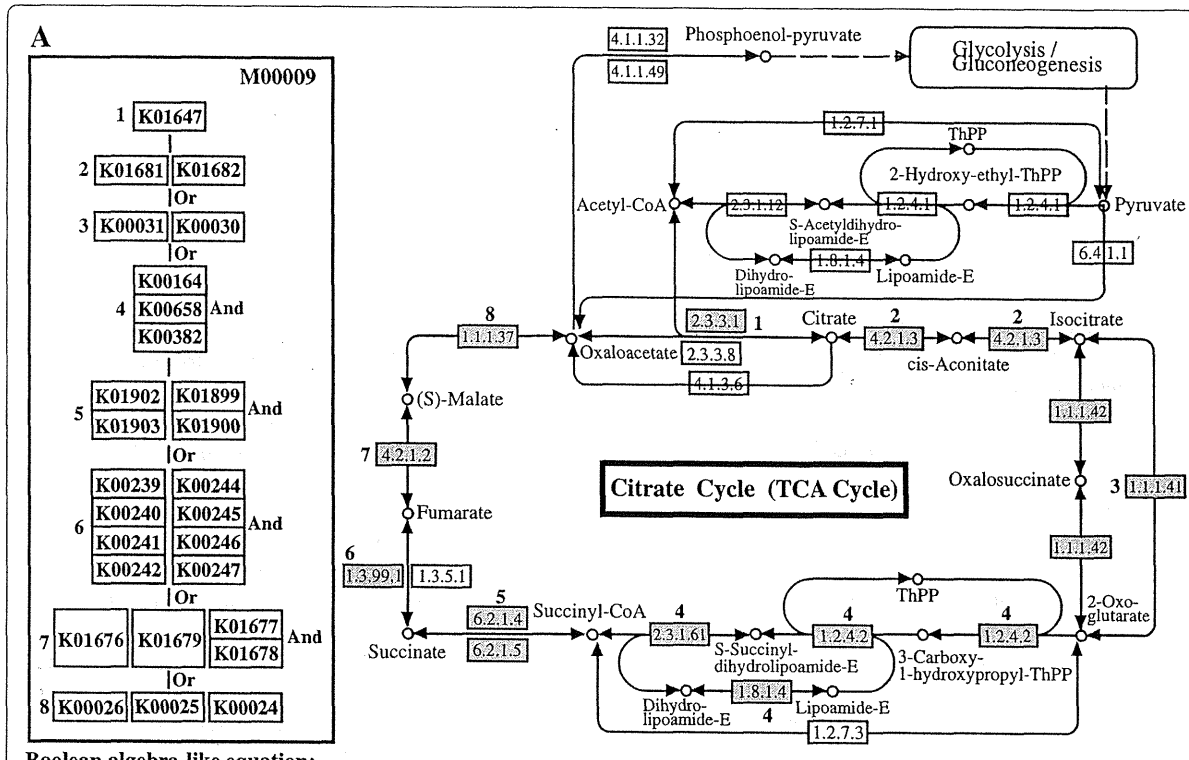
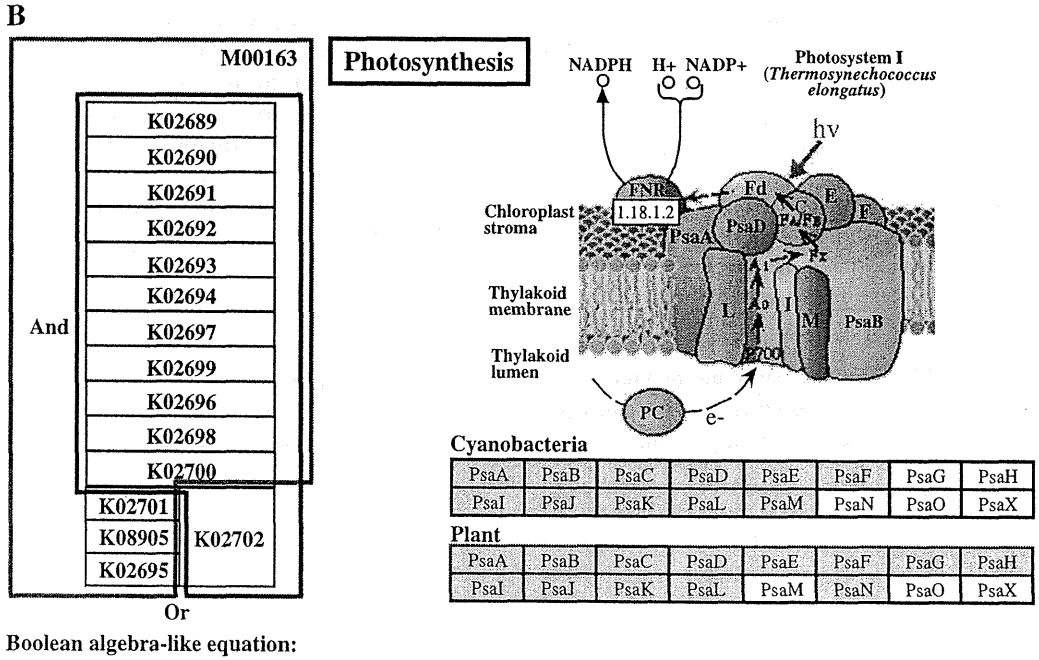



**Figure 1 Workflow of evaluation of the potential functionomes.** Detailed workflow of the three annotation servers, KAAS, MEGAN4, and MG-RAST using query sequences after gene finding process of sequenced data; KAAS and MEGAN4 use BLASTP and BLASTX for amino acid and nucleotide query sequences respectively and the MG-RAST uses only BLASTX. All use different databases, i.e. KEGG GENES for KAAS, NCBI-NR for MEGAN4, and M5nr [15] for MG-RAST (M5nr includes the SEED as a subset.), and different default threshold values for the BLAST hits. Each server converts the hit entries to the corresponding orthology IDs for functional annotation and pathway/module/subsystem mapping. Red colored texts of KAAS indicate its improvements in the current study (see Assignment of the query sequences to KO identifiers in the Methods section).




**Boolean algebra-like equation:**

**K01647 (K01681,K01682) (K00031,K00030) (K00164+K00658+K00382) (K01902+K01903,K01899+K01900) (K00239+K00240+K00241+K00242,K00244+K00245+K00246+K00247) (K01676,K01679,K01677+K01678) (K00026,K00025,K00024)**




**Boolean algebra-like equation:**

**(K02689+K02690+K02691+K02692+K02693+K02694)+(K02697+K02699+K02696+K02698+K02700)+((K02701+K08905+K02695), K02702)**

**Figure 2** (See legend on next page.)

(See figure on previous page.)
**Figure 2 KEGG functional modules. A**: A pathway module. The module M00009 comprising 8 reactions is defined for the citrate cycle (TCA cycle) core module and represented as a Boolean algebra-like equation of KO identifiers or K numbers for computational applications. The relationship between this module and the corresponding KEGG pathway map is also shown by indicating corresponding K number sets in the module and EC numbers in the pathway map using the same index. In each K number set, vertically connected K numbers indicate a complex and therefore represent "And" or "+" in the Boolean algebra-like equation, whereas horizontally located K numbers indicate alternatives and represent "Or" or "," in the equation. **B**: A structural complex module. The structural complex module M00163 comprising 12 (cyanobacteria) or 14 (plant) components is defined for the type I photosystem. The Boolean algebra-like equation and the corresponding KEGG pathway map are also shown. The KEGG pathway map shows the *Thermosynechococcus elongatus* (cyanobacteria) photosystem. Green and purple boxes indicate plant and cyanobacteria components, respectively.

pathways, 248 structure complexes, 4 functional sets, and 3 signatures) can be accessed through the website (http://www.genome.jp/kegg-bin/get_ htext?ko00002.keg).

This background motivated us to develop a new evaluation method using the KEGG MODULE database to differentiate the comprehensive and detailed functional potentials between different genomes and metagenomes. In this study, we first calculated the completion ratio of each KEGG module in reference species whose genomic sequences have been completely determined. Then we characterized the functional potentials between phenotypically different bacilli and human gut microbiomes from 13 healthy individuals. Finally, we validated the effect of database dependency on the accuracy of KO assignment.

## Results and discussion

### Distribution patterns of the module completion ratio in 768 prokaryotic species

KEGG modules are modular functional units derived from the KEGG pathways, and are categorized into pathway modules, structural complexes, functional sets and genotypic signatures. Each KEGG module is designed for automatic functional annotation by a Boolean algebra-like equation of KEGG Orthology IDs (see Methods for more details). However, it remains uncatalogued as to which species possess common modules or if certain modules demonstrate universality or rareness between specific species, phyla etc. Specific information regarding the phylogenetic profiles of each module holder would be especially useful for annotating metagenomes.

Thus, we first examined distribution patterns of the completion ratios of the KEGG modules in the 768 prokaryotic species whose genomic sequences have been completed (Additional file 1: Table S1). Although distribution of the module completion ratios in the 768 species varied greatly depending on the kind of module (Additional file 2: Figure S1-S3 and Additional file 1: Table S2), we found that it could be categorized into 4 patterns (universal, restricted, diversified and non-prokaryotic) regardless of the module type (pathway, structural complex, signature, or functional set), when considering 70% of all species to represent a majority measurement for the patters (Table 1 and Figure 3).

Pattern A defined as "universal" comprised modules completed by more than 70% of the 768 species (Figure 3A-1), and more than 70% of the 768 species possessed a module completion ratio of >80% (Figure 3A-2). Of 205 pathway modules containing submodules, modules grouped into pattern A account for only 11.3% (Table 1) and mainly belong to the categories of central carbohydrate metabolism and cofactor and vitamin biosynthesis. Although there are many species, more than 70% of the 768 prokaryotes possessed a module completion ratio of 80%, species with 100% completion ratio is very limited in the pattern A-2. M00019_1, shown as a representative of pattern A-2 (Figure 3), is a pathway

**Table 1 Classification of the KEGG modules based on the module completion ratio of 768 prokaryotes**

| Completion pattern | Definition of module type | Subtype | Pathways [203] | | Structural complexes [263] | | Functional sets [4] | | Signatures [3] | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | No. of modules (%) | | No. of modules (%) | | No. of modules (%) | | No. of modules (%) | |
| | | | Total | Rare | Total | Rare | Total | Rare | Total | Rare |
| A | Universal | A-1 | 15 (7.4) | 0 (0) | 9 (3.4) | 0 (0) | 1 (25) | 0 (0) | 0 (0) | 0 (0) |
| | | A-2 | 8 (3.9) | 1 (1.9) | 0 (0) | 0 (0) | 1 (25) | 0 (0) | 0 (0) | 0 (0) |
| B | Restricted | - | 22 (10.8) | 17 (31.5) | 119 (45.2) | 95 (89.6) | 0 (0) | 0 (0) | 3 (100) | 3 (100) |
| C | Diversified | - | 79 (38.9) | 36 (66.7) | 54 (20.5) | 11 (10.4) | 1 (25) | 1 (100) | 0 (0) | 0 (0) |
| D | Non-prokaryotic | - | 79 (38.9) | - | 81 (30.8) | - | 1 (25) | - | 0 (0) | - |

[ ] shows total number of the KEGG modules containing branched modules. "Rare" indicates the modules completed by less than 10% of 768 prokaryotic species. Universal: the modules completed by more than 70% of 768 prokaryotic species, Restricted: the modules completed by less than 30% of 768 prokaryotic species. Diversified: the modules that varies in the module completion ratio among 768 prokaryotic species, Non-prokaryotic: the modules not to be completed by any prokaryotic species.

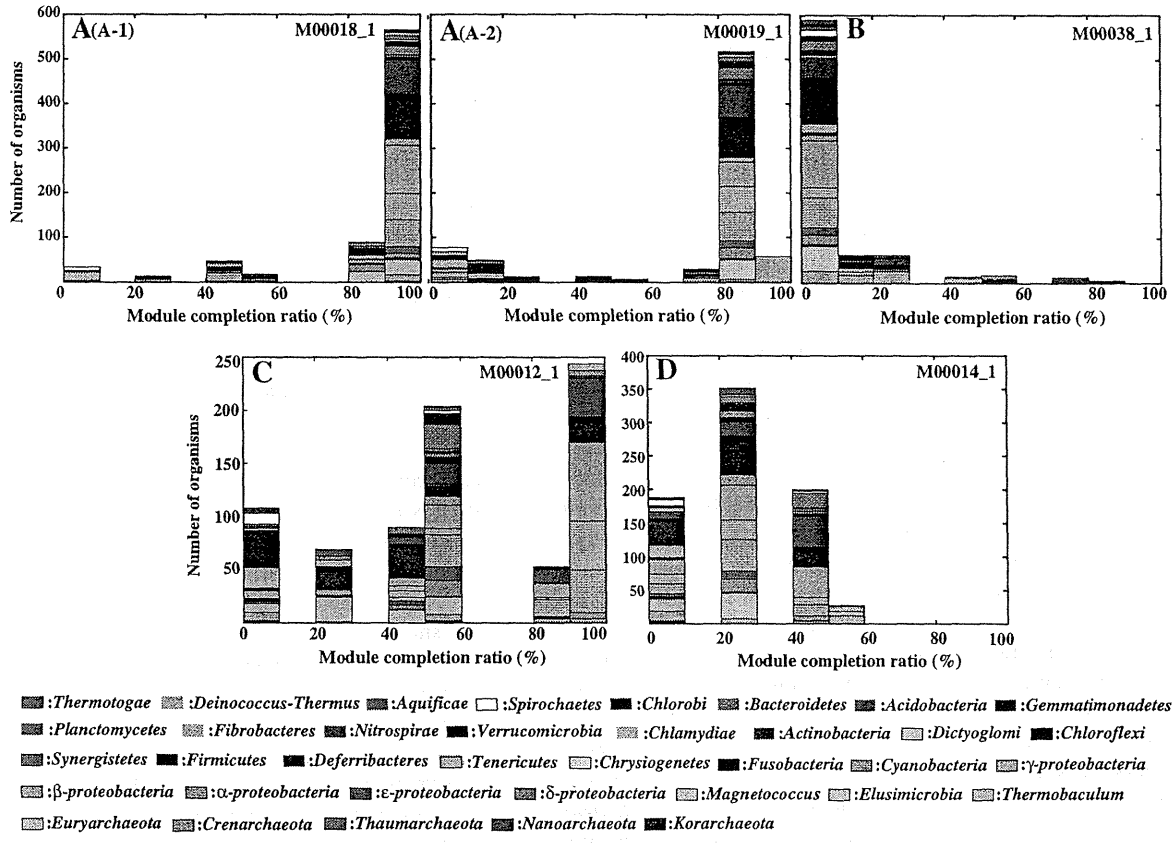



**Figure 3 Typical completion patterns to the KEGG modules by 768 prokaryotic species. A**: universal modules. (A-1) The modules completed by more than 70% of 768 prokaryotic species. M00018_1, which is threonine biosynthesis (aspartate homoserine threonine) is one of examples of the pattern A-1. (A-2) The modules for which more than 70% of 768 prokaryotic species show a module completion ratio of >80%. M00019_1, which is leucine biosynthesis (pyruvate 2-oxoisovalerate leucine) is one of examples of the pattern A-2. **B**: Restricted modules completed by less than 30% of 768 prokaryotic species. M00038_1, which is tryptophan metabolism (tryptophan kynurenine 2-aminomuconate) is one of examples of the pattern B. **C**: Diversified modules. These are the modules that vary in the module completion ratio among 768 prokaryotic species. M00012_1, which is glyoxylate cycle is one of examples of the pattern C. **D**: Non-prokaryotic modules completed by no prokaryotic species. M00014_1, which is glucuronate pathway (uronate pathway) is one of examples of the pattern D. Breakdown of taxonomic variations that complete each KEGG module is summarized in Table 1 and shown in Supplementary Table S2 in detail.

module for leucine biosynthesis comprising 7 reaction steps. The 1st reaction in this module, from pyruvate and thiamine diphosphate to 2-(α-hydroxyethyl) thiamine diphosphate plus $CO_2$ is catalyzed by acetolactate synthase (EC 2.2.1.6) comprising 3 (K01652, K01653, and K11258) subunits. However, since most of the species, except for 58 species within *Gammaproteobacteria*, do not have the genes assigned to K11258 of the acetolactate synthase II small subunit, the module completion ratio in the remaining 518 species becomes 85.7%. Thus, this small subunit may not necessarily be crucial for the pyruvate and thiamine reaction to occur in these species. Pattern B defined as "restricted" comprised modules completed by less than 30% of the species (Figure 3B), and accounted for 10.8% of all the pathway modules, and 17 modules were rare modules completed by less than 10% of the 768 species (Table 1). Pattern C defined as "diversified" accounted for 38.9% of all the pathway modules, and comprised modules ranging widely in completion ratios. M00012_1 (the glyoxylate cycle comprising 5 reactions) is one of representatives of pattern C (Figure 3C). As shown in Figure S4 (Additional file 2). 1 or several KO identifiers were assigned to each reaction in this module; however, KO identifiers, except for K01637 and K01638 assigned to the 3rd and 4th reactions, were also assigned to other pathway modules such as the TCA (Krebs) cycle (M00009_1), 1st carbon oxidation (M00010_1), 2nd carbon oxidation (M00011_1), reductive TCA cycle (M00173_1) and C4-dicarboxylate cycle (nicotinamide adenine dinucleotide (NAD)$^+$-malic enzyme type) (M00171_1). Some KO identifiers assigned to many of the modules, categorized into pattern C, were also

assigned to several other independent modules. Thus, when the module completion ratio is low, the relationship between the module completion ratio of the targeted module and others to which the same KO identifiers are assigned should be considered. Pattern D, which accounted for 38.9% of all pathway modules, comprised nonprokaryotic modules that are not completed by prokaryotic species (Figure 3D).

Of the 263 structural complex modules containing submodules redefined from modules with various complex patterns, 119 modules were categorized into pattern B (45.6%) and 95 were rare modules (Table 1). Pattern C accounted for only 20.5% in the structural modules compared with 38.9% in the pathway modules. Thus, it was hypothesized that most of the structural complex modules, except for pattern D, are shared only in limited prokaryotic species.

Non-prokaryotic modules account for 38.9% of pathway and 30.8% of structural complex modules respectively, and other modules were classified into various taxonomic patterns such as prokaryotic, Bacteria-specific and Archaea-specific based on the module completion profiles as shown in Table 2. These 4 patterns indicate the universal and unique nature of each module and also the versatility of the KO identifiers mapped to each module. Thus, the 4 criteria and taxonomic classification for each module should be helpful for interpretation of results based on module completion profile. A breakdown of all the modules grouped into the 4 patterns is summarized in Table S3-S5 (Additional file 1).

### Comparative functionome analysis of bacilli based on the KEGG modules

*Bacillus* and its related species in genera such as *Oceanobacillus* and *Geobacillus* reclassified from genus *Bacillus* (*Bacillus*-related species) are known to thrive in a wide range of environmental conditions: pH 2–12, temperatures between 5–78°C, salinity from 0 to 30% NaCl, and pressures from 0.1 Mpa (atmospheric pressure) to at least 30 MPa (pressure at a depth of 3000 m) [17]. The genome structure of these species within family *Bacillaceae* is comparatively similar, and the core structure comprising more than 1,400 orthologous groups is well conserved among *Bacillaceae* [18]. Therefore, moderately related bacillar genomes from 8 species with different phenotypic properties were selected to test our evaluation method for potential functionome using KEGG modules, in order to differentiate the functional potentials harbored in their genomes.

The gene products from 8 bacillar genomes were assigned to KO identifiers constructing each module in 111 pathway and 84 structural complex modules as shown in Figure S5 (Additional file 2). There was significant difference in the module completion ratio by 8 bacilli in terms of at least 19 pathway and 35 complex modules (Figure 4A and B). In particular, the completion ratio in *Oceanobacillus iheyensis*, a mesophilic, extremely halotolerant alkaliphile [19], was very low in 4 modules for thiamine biosynthesis (M00127_1), NAD biosynthesis (M00115_1), phosphatidylethanolamine biosynthesis (M00092_1) and biotin biosynthesis (M00123_1). These 4 modules were completed by all bacilli except for *O. iheyensis* although they are categorized into one of the diversified modules (pattern C). In addition, all bacilli almost completed the module for C5 isoprenoid biosynthesis (M00096_1), categorized into one of the universal modules (pattern A-2) in spite of very low module completion ratio by *O. iheyensis*. Conversely, 2 modules belonging to pattern C for tryptophan biosynthesis (M00023_1) and ketone body biosynthesis (M00088_1) were completed by only *O. iheyensis*, although other species partially completed them. Through these results it was evident that *O. iheyensis* differs from other bacilli in its metabolic potentials.

Some of the completed structural complex modules were found to be shared in bacilli with the same phenotypic properties, or to be independently species specific (Figure 4B). For example, the *Firmicutes*-specific modules for the teichoic acid transport system (M00251_1) were shared only among 3 mesophilic neutrophiles (*Bacillus subtilis* [20], *Bacillus amyloliquefaciens* [21], and *Bacillus licheniformis* [22]), although this module is widely shared in other genera such as *Staphylococcus*, *Clostridium*, and *Listeria* within phylum *Firmicutes*. Similarly, *Bacillus*-specific uncharacterized ATP-binding cassette (ABC) transport system (M00315_1) was also found to be shared among 3 mesophilic neutrophiles. On the other hand, 2 other modules, the iron (III) transport system (M00190_1) and phosphonate transport system (M00223_1) which are shared in many prokaryotic species within various phyla and belonged to pattern C, were shared only among 3 mesophilic alkaliphiles (*Bacillus halodurans* [23], *Bacillus pseudofirmus* [24], and *O. iheyensis*). Although it has been previously reported that the orthologous genes for the phosphonate transport system were shared between *O. iheyensis* and *B. halodurans* [19], it could be easily visualized using our new evaluation method that this system was also shared in other mesophilic and alkaliphilic *B. pseudofirmus*, whose genome sequence has been completed recently. In addition, another putative phosphonate transport system (M00224_1) and the N-acetylglucosamine transport system (M00205_1) categorized into one of the restricted modules (pattern B) were found to be conserved only in *O. iheyensis*. Although how the differentiated functional modules confer phenotypic properties directly or indirectly is still unclear, a series of the above results should be helpful in better understanding of the physiological properties.

**Table 2 Breakdown of taxonomic patterns of the KEGG modules**

| **Pathway [203]** | | **Structural complex [263]** | |
|---|---|---|---|
| Major taxonomic pattern | Number (%) | Major taxonomic pattern | Number (%) |
| Non-prokaryote | 79 (38.9) | Non-prokaryote | 81 (30.8) |
| Prokaryote | 52 (25.6) | Bacteria-specific | 45 (17.1) |
| Bacteria-specific | 25 (12.3) | Prokaryote | 42 (16) |
| *Gammaproteobacteria*-specific | 8 (3.9) | *Proteobacteria*-specific | 24 (9.1) |
| *Euryarchaeota*-specific | 6 (3) | Archaea-specific | 10 (3.8) |
| *Cyanobacteria*-specific | 4 (2) | *Cyanobacteria*-specific | 10 (3.8) |
| *Proteobacteria*-specific | 4 (2) | *Firmicutes*-specific | 10 (3.8) |
| *Alphaproteobacteria*-specific | 3 (1.5) | *Gammaproteobacteria*-specific | 8 (3) |
| *Proteobacteria/Firmicutes/Actinobacteria* | 3 (1.5) | *Proteobacteria/Firmicutes* | 4 (1.5) |
| Archaea-specific | 2 (1) | *Actinobacteria*-specific | 3 (1.1) |
| *Chloroflexi*-specific | 2 (1) | *Alphaproteobacteria*-specific | 3 (1.1) |
| *Crenarchaeota*-specific | 2 (1) | *Proteobacteria/Actinobacteria* | 3 (1.1) |
| *Firmicutes*-specific | 2 (1) | *Gammaproteobacteria/Firmicutes/Fusobacteria* | 2 (0.8) |
| *Actinobacteria*-specific | 1 (0.5) | *Actinobacteria/Verrucomicrobia/Nitrospirae* | 1 (0.4) |
| *Betaproteobacteria*-specific | 1 (0.5) | *Betaproteobacteria*-specific | 1 (0.4) |
| *Betaproteobacteria/Actinobascteria/ Cyanobacteria* | 1 (0.5) | *Euryarchaeota*-specific | 1 (0.4) |
| *Betaproteobacteria/Chloroflexi* | 1 (0.5) | *Firmicutes/Actinobacteria* | 1 (0.4) |
| *Cyanobacteria/Euryarchaeota* | 1 (0.5) | *Firmicutes/Fusobacteria* | 1 (0.4) |
| *Cyanobacteria/Chlorobi* | 1 (0.5) | *Frmicutes*-specific | 1 (0.4) |
| *Gammaproteobacteria/ Firmicutes/Cyanobacteria* | 1 (0.5) | *Gammaproteobacteria/Firmicutes* | 1 (0.4) |
| *Proteobacteria/Actinobacteria* | 1 (0.5) | *Proteobacteria/Actinobacteria/Deonococcus-Thermus* | 1 (0.4) |
| *Proteobacteria/Bacteroidetes* | 1 (0.5) | *Proteobacteria/Actinobacteria/Verrcomicrobia* | 1 (0.4) |
| *Proteobacteria/Fusobacteria/ Gemmatimonadetes* | 1 (0.5) | *Proteobacteria/Chloroflexi/Deonococcus-Thermus* | 1 (0.4) |
| *Proteobacteria/Firmicutes* | 1 (0.5) | *Proteobacteria/Chlamydiae/Chlorobi* | 1 (0.4) |
| **Functional set** [4] | | *Proteobacteria/Chlamydiae* | 1 (0.4) |
| Major taxonomic pattern | Number (%) | *Proteobacteria/Chrysiogenetes/Spirochaetes* | 1 (0.4) |
| Prokaryote | 3 (75) | *Proteobacteria/Cyanobacteria* | 1 (0.4) |
| Non-prokaryote | 1 (25) | *Proteobacteria/Cyanobacteria/Chlorobi* | 1 (0.4) |
| **Signature** [3] | | *Proteobacteria/Firmicutes/Actinobacteria* | 1 (0.4) |
| Major taxonomic pattern | Number (%) | *Proteobacteria/Magnetococcus* | 1 (0.4) |
| *Proteobacteria*-specific | 1 (33.3) | *Proteobacteria/Magnetococcus/Aquificae* | 1 (0.4) |
| *Betaproteobacteria*-specific | 1 (33.3) | | |
| *Gammaproteobacteria*-specific | 1 (33.3) | | |

[ ] shows total number of the KEGG modules containing branched modules.

### Comparative functionome analysis of humans and human gut microbiomes

The completion ratio of each KEGG module was compared between humans and human gut microbiomes to illustrate their metabolic linkage. The metagenomic data of gut microbiomes from 13 healthy Japanese individuals, previously reported on, was used [6]. Detailed information for all metagenomic samples are summarized in Table S7 (Additional file 1). The gene products from metagenomes of the microbiomes from the 13 individuals were assigned to KO identifiers constructing each module in 158 pathway and 150 structural complex modules as shown in Figure S6 (Additional file 2). Similarly, the gene products from the human genome were completely or partially assigned to KO identifiers in 144 pathway and 84 structural complex modules. There was a significant difference in the module completion ratios of 13 individuals in terms of at least 35 pathway modules (Figure 5A).

The most complete 16S rRNA gene sequence-based enumerations available in human gut microbiomes

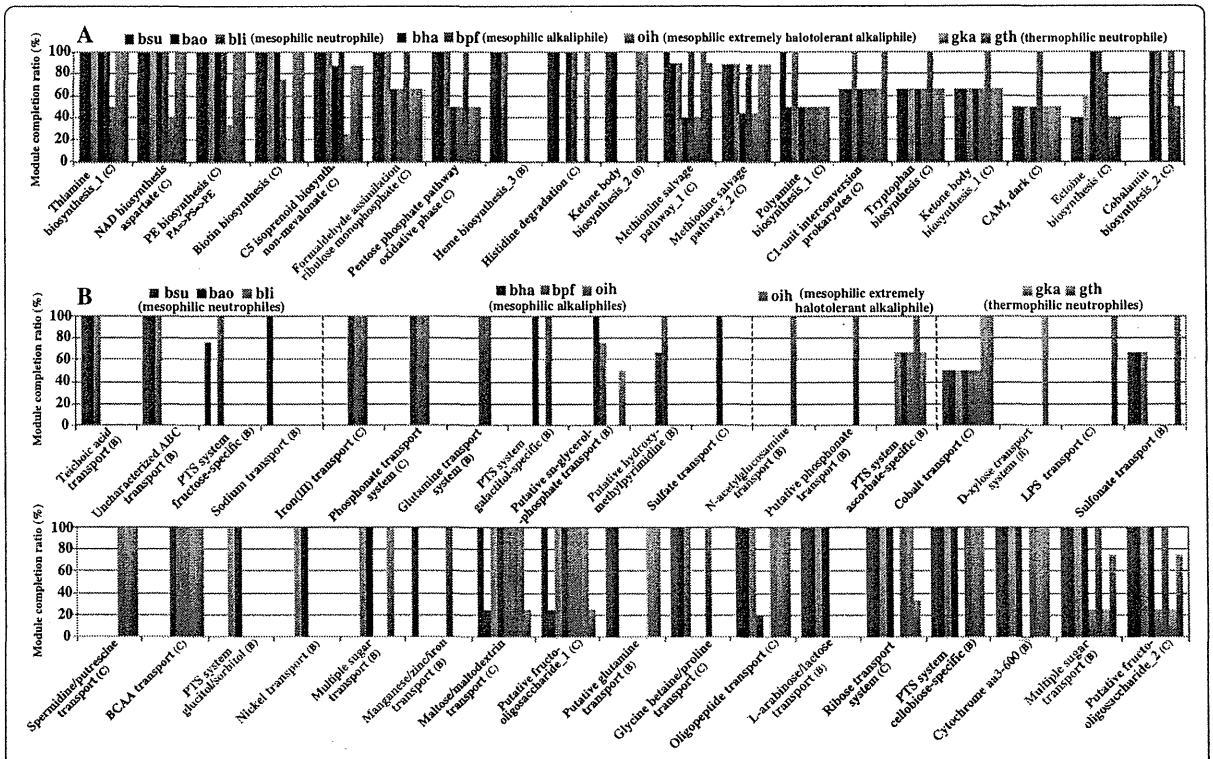



**Figure 4 Comparison of module completion patterns in 8 phenotypically different *Bacillus*-related species. A**: Pathway modules showing remarkable differences appeared among the 8 species. **B**: Structural complex modules showing remarkable differences appeared among the 8 species. Upper histogram indicates common or specific modules in the species possessing each phenotype (from left to right; mesophilic neutrophile; mesophilic alkaliphile; mesophilic, extremely halotolerant alkaliphile; and thermophilic neutrophile). Green letters show rare modules completed by less than 10% of 768 prokaryotic species described in Figure 3. Alphabet in parentheses shows the patterns of completion profile based on the module completion ratio as shown in Table 1 and Figure 3. **A**: Universal module, **B**: Restricted module, **C**: Diversified module, **D**: Non-prokryotic module. bsu, *B. subtilis*; bao, *B. amyloliquefaciens*; bli, *B. licheniformis*; bha, *B. halodurans*; bpf, *B. pseudofirmus*; oih, *O. iheyensis*; gka, *G. kaustophilus*; and gth, *G. thermoglucosidasius*.

indicate that more than 90% of phylotypes belong to just two of the 70 known divisions of Bacteria, the *Bacteroidetes* and the *Firmicutes*, with the remaining phylotypes distributed among eight other phyla [25]. Pairwise comparison of the completion ratio of the KEGG module clearly demonstrated the well-recognized functional complementation of the gut microbiome to the human host, which includes essential amino acid and vitamin biosynthesis (Additional file 2: Figure S6). The contributors completing the modules for vitamin production are *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Gammaproteobacteria*. Completion patterns of the KEGG module for these amino acids and vitamins mainly fall into patterns C (diversified module) and D (non-prokaryotic module) except for riboflavin biosynthesis (M00125_1~_3) belonging to one of the universal modules (A-1), indicating that these modules are involved in the nutritional supply for the gut microbiome as well as for the host (Figure 5B). Inter-individual variation was also evident in the completion ratio of the module for vitamins. For example, the module (M00124_1) belonging to pattern C for pyridoxal (vitamin B6) biosynthesis was mainly attributable to *Bacteroidetes* in adults and *Gammaproteobacteria* in infants; however, its completion ratio in 2 male infants (In-B and In-E) was extremely low (33.33%) (Figure 5A). Inter-individual variations in completion ratios were also observed in modules (M00133_1 and M00134_1) for polyamine biosynthesis, for example, putrescine, spermidine, and spermine. Similarly, the completion ratio of the KEGG modules (M00135_1 and M00136_1) for γ-aminobutyric acid (GABA) varied among individuals, and *Gammaproteobacteria* mainly contributed to GABA production. Because these polyamines and GABA are essential biological substances that act as cell growth promoters and inhibitory neurotransmitters respectively, in humans, these variations may be linked to susceptibilities to certain diseases. Indeed, a recent report on metabolic changes in gut microbiomes after bariatric surgery for obese patients demonstrated their potential for polyamine production in the gut; elevated protein putrefaction because of the bypassed food passage promoted putrescine and GABA production from gut microbiota [26].

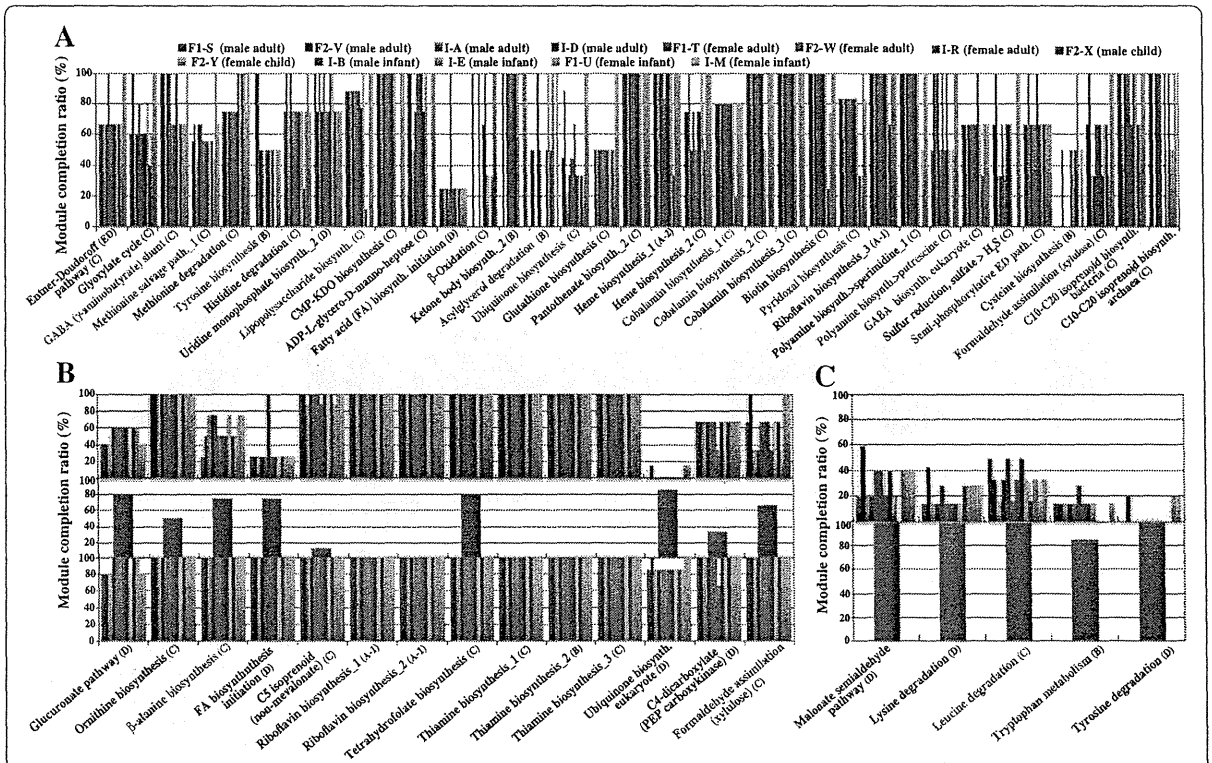



**Figure 5 Comparison of module completion patterns in humans and human gut microbiomes from 13 healthy individuals. A**: Typical pathway modules showing remarkable differences in the module completion ratio appeared among human gut microbiomes from 13 healthy individuals. **B**: Typical pathway modules possessing complementary relationships between humans and human gut microbiomes in the module completion ratio. **C**: Typical pathway modules for which the completion ratio in the human gut microbiome is very low in contrast to that in humans. Green letters show rare modules completed by less than 10% of 768 prokaryotic species described in Figure 3. Detailed information of the 13 individuals has been previously described [6]. Alphabet in parentheses shows the patterns of completion profile based on the module completion ratio as shown in Table 1 and Figure 3. **A**: Universal module, **B**: Restricted module, **C**: Diversified module, **D**: Non-prokryotic module.

Interestingly, gut microbiomes showed preference for amino acid catabolism. As shown in Figure 5C, the gut microbiome did not seem to utilize exogenous lysine (M00032_1), leucine (M00036_1), and aromatic amino acids such as tryptophan (M00038_1) and tyrosine (M00044_1). To our knowledge, this is a novel finding on the nutritional preference of gut microbes. This may be one of the mutualistic representations of gut microbiomes to avoid nutritional competition with the host because these aromatic amino acids are precursors of various biological substances such as catecholamines, melatonin, serotonin, thyroid hormones, and NAD. To assess the taxonomic composition of gut microbiomes, the module for glycolysis (M00002_1) was analyzed (Figure 6). Each gene product mapped to this module was taxonomically assigned, and distribution at the phylum level was calculated. Consistent with a previous report [6], adult and child gut microbiomes are constituted by 2 major phyla, *Firmicutes* and *Bacteroidetes*. Analysis of the module for glycolysis clearly differentiated the gut microbial composition between adults and infants as well as among infants. In particular, it was highlighted that *Actinobacteria* was a major phylum in breast-fed infants, whereas *Gammaproteobacteria* was predominant in a bottle-fed infant. It was also evident that the microbial composition in an infant with mixed feeding (breast and bottle) showed a pattern intermediate between those in breast- and bottle-fed infants. Thus, the new evaluation method based on the KEGG modules is expected not only to highlight the metabolic linkage between host and commensal microbes but also to identify microbiome-based biomarkers for particular diseases.

## Conclusions

We developed a new evaluation method for potential functionomes based on the KEGG modules. Modules with branching or different component structures in a complex were redefined depending on the number of branching or types of component structures. The module completion ratio was calculated by counting the number of genes assigned to KO identifiers constructing each module defined by a Boolean algebra-like equation.

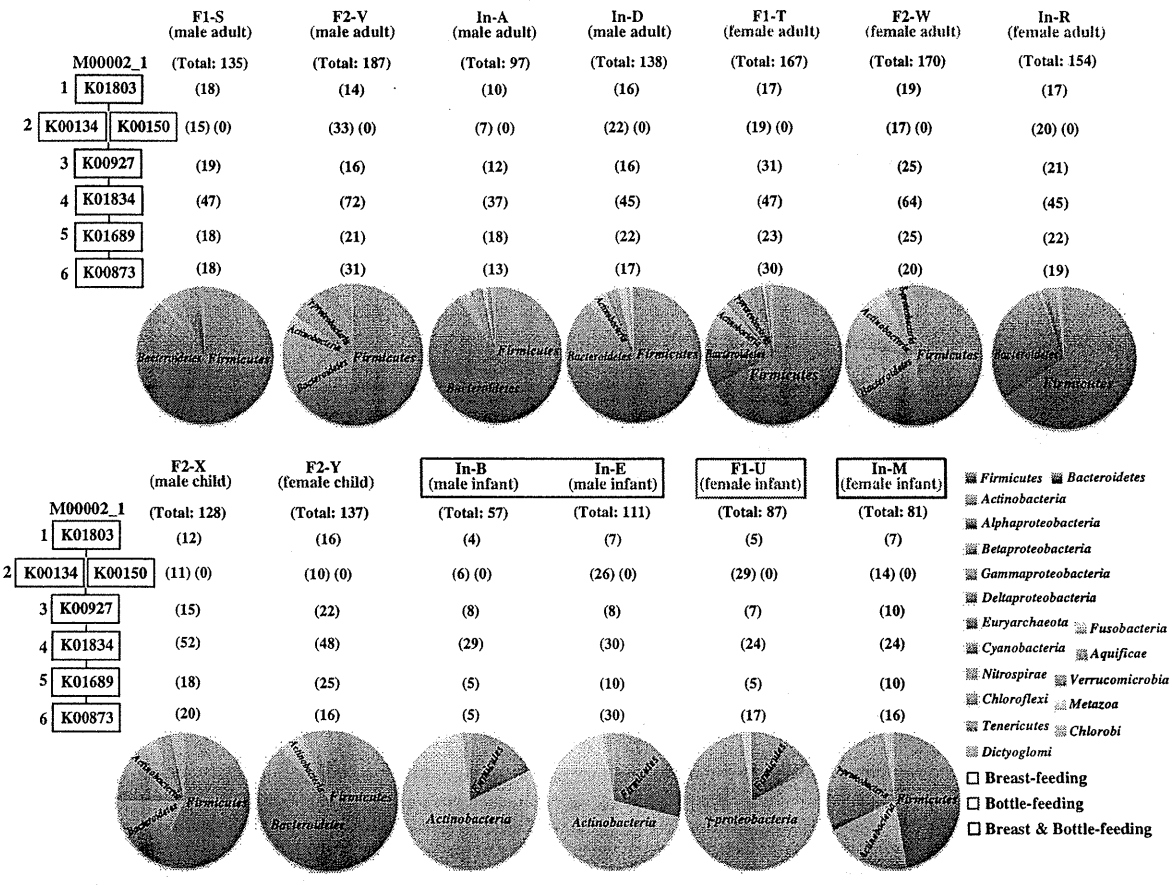



**Figure 6 Taxonomic variation in genes assigned to KOs associated with glycolysis in human gut microbiomes from 13 healthy individuals.** The pathway module M00002_1 comprising 6 steps shows glycolysis (core module involving 3-carbon compounds), and K number in each box indicates KO assigned for every individual reaction. Because K00134 and K00150 have the relationship of "Or" as explained in Figure 2, if the gene from human gut microbiomes is assigned to either K00134 or K00150, the reaction at the 2nd step can be executed. Pie charts show taxonomic breakdown of the genes assigned to KOs in all 6 steps. Numbers in parentheses indicate the number of the genes assigned to KO in each reaction step.

Using this new method, we found significant difference in module completion ratio by 8 bacilli in terms of at least 19 pathway and 35 complex modules, although how the differentiated functional modules confer phenotypic properties directly or indirectly is unclear thus far. Because the coverage of KEGG modules over whole metabolic and signaling networks is continuously increasing, differences in module completion ratio will provide some important clues to the understanding of phenotypic properties. Furthermore, variations in the functional potential of human gut microbiomes from 13 healthy individuals could be characterized by the pathway and structural complex module units, and the complementarity between biochemical functions in human hosts and nutritional preferences in human gut microbiomes identified. In addition, taxonomic variation of the contributors to each module clarified by this method will prove informative when considering ecological dynamics.

Functional annotations to metagenomic sequences remain difficult because metagenomic data targeting various environments still contains incomplete genes from various unidentified species, absent in a reference database. In this study, we used the KAAS system for functional annotation to the human metagenomes and also attempted to estimate database dependency on the accuracy of the KO assignment using the *E. coli* draft genome. As a result, the KAAS system could correctly assign to KO groups with an accuracy rate of approximately 80%, even if the gene hosts were not classified into known phyla within the reference database (see the Methods section). Thus, our method will work well for comparative functional analysis in metagenomics, able to target unknown environments containing various uncultivable microbes within unidentified phyla, although further verification studies on database dependency for metagenomics should be performed.

## Methods

### KEGG MODULE

KEGG MODULE [16] is a collection of pathway modules and other functional units designed for automatic functional annotation or pathway enrichment analysis. Pathway modules such as the TCA cycle core module (Figure 2A) are tighter functional units than KEGG pathway maps and are defined as consecutive reaction steps, operon or other regulatory units, and phylogenetic units obtained by genome comparisons. Other functional units include (1) structural complexes representing sets of protein subunits for molecular machineries such as photosystems (Figure 2B), (2) functional sets representing other types of essential sets such as aminoacyl-tRNA synthases, and (3) signature modules representing markers of phenotypes such as enterohemorrhagic *E. coli* pathogenicity signature for Shiga toxin. The latest KEGG MODULE is available from the KEGG FTP site (http://www.kegg.jp/kegg/download). Each module is defined by the combination of KO identifiers so that it can be used for annotation and interpretation purposes in individual genomes or metagenomes. Notations of the Boolean algebra-like equation (Figure 2 and Additional file 1: Table S6) for this definition include space-delimited items for pathway elements, comma-separated items in parentheses for alternatives, a plus sign to define a complex, and a minus sign for an optional item. Some modules have branching points in their reaction cascades (Additional file 2: Figure S7), leading to different products or alternative reaction pathways. These modules are divided into several parts depending on the branching patterns and are redefined as submodules for accurate calculation of the completion ratio. In the example shown in Figure S7 (Additional file 2), there are 2 pathways for the module for heme biosynthesis (M00121_1): 1 for protoheme (C00032) and the other for siroheme (C00748). Consequently, 3 submodules were defined: M00121_1 for the original, M00121_2 for protoheme production, and M00121_3 for siroheme production. By redefinition of the modules based on the above mentioned policy the number of pathway and structural complex modules increased from 179 to 205 and from 248 to 263, respectively. The module completion ratio (see the section below) was calculated for each submodule in this study to examine fine-grained functional categories. A breakdown of the functional categories of all KEGG modules containing newly reidentified modules is summarized in Table 3.

#### Calculation of the module completion ratio based on a Boolean algebra-like equation

The completion ratio of all KEGG functional modules in each organism was calculated based on a Boolean algebra-like equation (Additional file 1: Table S6). For this analysis, 1 genome was selected from each of the 768 available prokaryotic species shown in Table S1 (Additional file 1) and a reference genome set was constructed to cover all completely sequenced prokaryotes, excluding draft genomes as of December 2011.

As one of examples, M00009_1 is a core pathway module for the TCA cycle comprising 8 reactions as shown in Figure 2A. In each KO number set, vertically connected KO identifiers indicate a complex and therefore represent "And" or "+" in the Boolean algebra-like equation, whereas horizontally located K numbers indicate alternatives and represent "Or" or "," in the equation. When genes are assigned to all KO identifiers in each reaction according to the Boolean algebra-like equation, the module completion ratio becomes 100%. If genes are not assigned to KO identifiers in 2 reactions, the module completion ratio is calculated as 75% (6/8 x 100 = 75). On the other hand, M00163_1 and M00163_2 comprising 14 components in plants and 12 in cyanobacteria represent a complex module for photosystem I. If genes assigned to KO identifiers in 2 of those components are missing in plants, the module completion ratio is calculated as 85.7% and 83.3% in the case of cyanobacteria (Figure 2B). A stand-alone calculation system of module completion ratio for Linux and Mac OS X is available from the download site of ExtremoBase (http://www.jamstec.go.jp/gbrowser/cgi-bin/top.cgi).

#### Assignment of the query sequences to KO identifiers

Efficient and accurate computational methods are required for the functional annotation of rapidly growing sequence data from complete genomes and metagenomes. KAAS is an efficient tool for assigning KO identifiers to genes from complete genomes based on a BLAST search of the KEGG GENES database combined with a bidirectional best-hit method [11]. Because of the efficiency, KAAS is used to assign KO identifiers to protein sequences from metagenome projects (http://www.genome.jp/kegg/catalog/org_list3.html) and to users' own data from other genome and metagenome projects.

We applied a slightly modified version of the KAAS system that has improved the accuracy of KO assignments by (i) using a variable bit-score threshold instead of a fixed one (60 in the original KAAS system) to avoid missed annotations when there are sufficient high-scoring hits for KO assignment, and (ii) considering taxonomic information of each KO when more than 1 candidate KO is obtained (Figure 1). This modification resulted in improved positive predictive value (#true positives / #all positives) by 2–5% in the KO reassignment tests for 30 selected species (Additional file 2: Figure S8). The latest stand-alone KAAS system for Linux and Mac OS X is available from the web site of KAAS HELP (http://www.genome.jp/tools/kaas/help.

**Table 3 Breakdown of small functional categories of the KEGG modules**

| **Pathway modules [203]** | | **Structural complex modules [263]** | |
|---|---|---|---|
| Small functional category | Number (%) | Small functional category | Number (%) |
| Cofactor & vitamin biosynthesis | 30 (14.6) | Saccharide and polyol transport system | 29 (11.0) |
| Carbon fixation | 14 (6.8) | ATP synthesis | 27 (10.3) |
| Central carbohydrate metabolism | 14 (6.8) | Phosphotransferase system (PTS) | 24 (9.1) |
| Lipid metabolism | 14 (6.8) | Mineral and organic ion transport system | 23 (8.7) |
| Glycan metabolism | 13 (6.3) | Phosphate and amino acid transport system | 19 (7.2) |
| Aromatic amino acid metabolism | 11 (5.4) | ABC-2 type and other transport systems | 16 (6.1) |
| Methane metabolism | 11 (5.4) | Bacterial secretion system | 14 (5.3) |
| Fatty acid metabolism | 10 (4.9) | RNA processing | 13 (4.9) |
| Sterol biosynthesis | 10 (4.9) | Ubiquitin | 13 (4.9) |
| Cystein & methionine metabolism | 7 (3.4) | Metallic cation, iron-siderophore and vitamin B12 transport system | 12 (4.6) |
| Glycosaminoglycan metabolism | 7 (3.4) | Protein processing | 9 (3.4) |
| Other carbohydrate metabolism | 7 (3.4) | Spliceosome | 9 (3.4) |
| Polyamine biosynthesis | 6 (2.9) | Repair system | 8 (3.0) |
| Telpenoid backbone biosynthesis | 6 (2.9) | DNA polymerase | 7 (2.7) |
| Lysine metabolism | 5 (2.4) | Photosynthesis | 7 (2.7) |
| Pyrimidine metabolism | 5 (2.4) | RNA polymerase | 7 (2.7) |
| Akaloid & other secondardy metabolite | 4 (1.9) | Peptide and nickel transport system | 6 (2.3) |
| LPS metabolism | 4 (1.9) | Replication system | 6 (2.3) |
| Other terpenoid biosynthesis | 4 (1.9) | Carbohydrate metabolism | 5 (1.9) |
| Arginine & proline metabolism | 3 (1.5) | Proteasome | 5 (1.9) |
| BCAA metabolism | 3 (1.5) | Ribosome | 3 (1.1) |
| Other amino acid metabolism | 3 (1.5) | Glycan metabolism | 1 (0.4) |
| Phenylpropanoid & flavonoid biosyntesis | 3 (1.5) | **Functional set modules** [4] | |
| Purine metabolism | 3 (1.5) | Small functional category | Number (%) |
| Histidine metabolism | 2 (1.0) | Aminoacyl-tRNA | 2 (50) |
| Metabolic capacity | 2 (1.0) | Nucleotide sugar | 2 (50) |
| Serin & threonine metabolism | 2 (1.0) | **Signature modules** [3] | |
| Nitrogen fixation | 1 (0.5) | Small functional category | Number (%) |
| Sulfur metabolism | 1 (0.5) | Genotypic signature | 3 (100) |

[ ] shows total number of the KEGG modules containing branched modules.

html). We used this new KAAS to estimate database dependency on accuracy of the KO assignment. We selected *Escherichia coli* as a representative of prokaryotic species and constructed 4 different types of datasets: without *E. coli* and closely related species (1,239 species), without all species within family *Enterobacteriales* (1,200 species), without all species within class *Gammaproteobacteria* (1,040 species), and without all species within phylum *Proteobacteria* (755 species).

In addition, we created artificially fragmented protein sequences from *E. coli* to confirm the accuracy of KO assignment to the truncated proteins because the gene products predicted from the end of assembled contigs and singletons were often truncated in the sequences produced by 2$^{nd}$ generation DNA sequencer. The draft genome of *E. coli* isolated from infants in Trondheim, Norway (accession: ERX127960), which appears to contain several sequencing errors, was used for this analysis. The short read sequences of *E. coli* produced by a 454 GS FLX Titanium sequencer (92.4 Mb in total) were assembled by Newbler ver. 2.6 and the contigs longer than 500 bp were used in this analysis. By using MetaGeneAnnotator [27], 4,410 complete and 178 partial CDSs were identified from the contigs. The amino acid sequences of complete CDSs were randomly fragmented to 50, 60, 80, 100, 120, 150, and 200 residues in length,

and each fragment was subjected to verification of database dependency based on the accuracy of KO identifier assignment (Additional file 2: Figure S9).

In general, because most microbes thriving in natural environments are uncultivable, many genes in environmental metagenomes do not show significant similarity to those from known species in the public genome database. Especially when microbial genomes belonging to the same phylum as the query microbe are missing in the genome database, the accuracy rate of KO assignment to proteins phylogenetically distant from known phyla is expected to be low. In fact, when all species within phylum *Proteobacteria* were not included in the data set, the accuracy rate of KO assignment to full proteins of *E. coli* decreased to 80%, but the accuracy rate of approximately 70% was maintained even in the proteins fragmented to about 100 residues (Additional file 2: Figure S9). Considering these results, even if the genes from unidentified phyla of the so-called Candidate division are included in the metagenomes, the KAAS system can presumably assign KO identifiers to genes longer than 300 bp (100 amino acids) with an accuracy rate of approximately 70%.

### Application of the evaluation method for functionome to genomic analysis

The completed genomic sequences of 8 moderately related species within *Bacillaceae* were selected and obtained from the KEGG FTP site (ftp://ftp.bioinformatics.jp/). A breakdown of the selected species is as follows: *B. subtilis* [20], *B. amyloliquefaciens* [21], and *B. licheniformis* [22] (mesophilic neutrophile); *B. halodurans* [23] and *B. pseudofirmus* [24] (mesophilic alkaliphile); *O. iheyensis* [19] (mesophilic, extremely halotolerant alkaliphile); and *Geobacillus kaustophilus* [28] and *G. thermoglucosidasius* (thermophilic neutrophile). Amino acid sequences were used for assignment of KO identifiers to the gene products from each species by the KAAS system (Figure 1) and the KO assigned ones were mapped to the KEGG modules. The completion ratio of each module in the 8 species was calculated based on the Boolean algebra-like equation, and the ratios were compared with each other to differentiate the potential functionome among the 8 species.

### Application of the evaluation method for functionome to metagenomic analysis

The metagenomic Sanger sequences of human gut microbiomes were selected from 13 healthy individuals [6] to apply our evaluation method for potential functionome using the KEGG modules to comparative metagenomic analysis focusing on differentiation of functional potentials between individuals. Amino acid sequences of the gene products predicted by MetaGeneAnnotaor [27] from the assembled contigs and also singlets in metagenomic sequences were used in this study and obtained from KEGG Metagenomes (http://www.genome.jp/kegg/catalog/org_list3.html). Sample ID and number of coding sequences (CDSs) used in this study are as follows: Male adult: F1-S (54,151), F2-V (65,156), and In-A (35,260); Female adult: F1-T (65,156), F2-W (57,213), and In-R (63,356); Female child: F2-X (57,446); F2-Y (64,942); Male infant: In-B (20,063) and In-E (37,652); and Female infant: F1-U (35,260) and In-M (34,330). Complete CDSs account for 40-55% of all CDSs in each sample and their average length was about 180 to 217 a.a. Average length of partial CDSs in each sample was 146 to 175 a.a. and the CDSs longer than 100 a.a. account for more than 80% of all CDSs in all 13 samples. The detailed information for metagenomic samples of human gut microbiomes are summarized in Table S7 (Additional file 1). KO identifiers were assigned to 35-55% of the complete and partial CDSs identified in the metagenomic sequences from the 13 individuals using the KAAS system and 31-39% of KO assigned CDSs were then mapped to the KEGG functional modules (Additional file 1: Table S7). The completion ratio of each module by human gut microbiomes from 13 healthy individuals was calculated based on the Boolean algebra-like equation and compared with each other to differentiate the potential functionome between the 13 individuals. Human genome sequence was also obtained from the KEGG FTP site and KO identifiers were assigned to 10,508 CDSs. The KO assigned 1,321 CDSs were mapped to the KEGG functional modules similar to human gut microbiomes.

Alternatively, we also employed assembled contig sequences of short reads produced (>500 bp) by an Illumina GA to compare their functionome with those by the Sanger sequences. Short reads from the metagenomics of human intestinal tract (MetaHIT) project [29] were subjected to CDS identification, KO assignment and KEGG module mapping (of the KO assignment CDSs). We obtained two sections of the MetaHit sequence data (MH_0001: Danish female and MH_0005: Danish male) from the MetaHIT database (http://www.bork.embl.de/~arumugam/Qin_et_al_2010/) and approximately 28,000 and 27,000 CDSs containing partial sequences were identified by MetaGeneAnnotator [27], respectively. KO identifiers were assigned to 40.5-41.5% to the CDSs from the 2 Danish individuals and 33-34% of KO assigned CDSs were mapped to the KEGG modules, in a similar manner as the Sanger sequences (Additional file 1: Table S7). Ultimately, it was found that there were no discernible differences in KO ID assignment and mapping ratios of KO assigned CDSs to the KEGG modules when comparing between Sanger and Illumina reads using contigs longer than 500 bp for analysis.

## Additional files

**Additional file 1: Tables S1–S7. Table S1.** List of 768 prokaryotic species used in this study. The additional data are available with the online version of this paper. **Table S2.** Taxonomic patterns of the prokaryotes which complete the KEGG modules (205 pathways, 263 structural complexes, 4 functional sets, and 3 signatures). Functional annotation of each module is listed in Table S3-S5. Figures S1-S3 were drawn based on this table. **Table S3.** Characterization of the 205 KEGG pathway modules containing submodules based on the module completion patterns in 768 prokaryotic species. **Table S4.** Characterization of the 263 KEGG structural complex modules containing submodules based on the module completion patterns in 768 prokaryotic species. **Table S5.** Characterization of the 7 KEGG modules (4 functional sets and 3 signatures) based on the module completion patterns in 768 prokaryotic species. Table S6. Notations of Boolean algebra-like equations for all KEGG modules containing redefined ones. Table S7. Summary of metagenomic sequences of human gut microbiome.

**Additional file 2: Figures S1–S9. Figure S1.** Distribution patterns of the completion ratio of the KEGG pathway modules in 768 prokaryotic species. The completion ratio of 205 pathway modules containing submodules were evaluated in this study. **Figure S2.** Distribution patterns of the completion ratio of the KEGG structural complex modules in 768 prokaryotic species. The module completion ratio of 263 structural complex modules containing submodules was evaluated in this study. **Figure S3.** Distribution patterns of the completion ratio of the KEGG functional set and signature modules in 768 prokaryotic species. The module completion ratio of 7 functional set and signature modules was evaluated in this study. **Figure S4.** Distribution of KO identifiers mapped to the module for glyoxylate cycle (M00012) in other pathway modules. KO identifiers, except for K01637 and K01638 colored light green, are also shared in several other modules. **Figure S5.** Module completion patterns in 8 phenotypically different *Bacillus*-related species. (A) Pathway module. (B) Structural complex module. bsu, *Bacillus subtilis* ; bao, *Bacillus amyloliquefaciens* ; bli, *Bacillus licheniformis* ; bha, *Bacillus halodurans* ; bpf, *Bacillus pseudofirmus* ; oih, *Oceanobacillus iheyensis* ; gka, *Geobacillus kaustophilus* ; and gth, *Geobacillus thermoglucosidasius*. Green characters show rare modules, which are completed by less than 10% of 768 prokaryotic species. **Figure S6.** Module completion patterns in human and human gut microbiomes. (A)-1–3, Pathway module. (B)-1–3, Structural complex module. Upper histogram shows the module completion pattern in gut microbiomes from 13 healthy individuals [18]. Middle histogram shows module completion patterns in humans. Lower histogram shows module completion patterns in human gut microbiomes plus humans. Green characters show rare modules, which are completed by less than 10% of 768 prokaryotic species. **Figure S7.** Definition of submodules for the KEGG module with branching. The heme biosynthesis pathway (glutamate => protoheme => siroheme) module (M00212) has branching at the intermediate compound uroporphyrinogen III (C01051), where this module was divided into 2 parts. Submodules are defined as M00121_1 (original), M00121_2 (left-side branching), and M00121_3 (right-side branching). Ovals with C numbers, rectangles with R numbers, and K numbers represent metabolites, enzymatic reactions, and KO, respectively. KO is used for mapping functional annotation of genes to the modules. Black K numbers indicate KO common to all 3 newly redefined submodules (M00121_1, M00121_2, and M00121_3), and blue and red K numbers correspond to reactions specific to M00121_2 and M00121_3, respectively. **Figure S8.** Positive predictive values (ppv) of the KO reassignment tests by KAAS. We performed KO reassignment tests for 30 species (7 eukaryotes, 20 bacteria, 3 archaea) by original (old) and improved (new) KAAS and found that new KAAS showed 2-5% improvements compared with the old KAAS. Three letter codes in X axis indicate species. abbreviations as follows: hsa: *Homo sapiens*, dre: *Danio rerio*, dme: *Drosophila melanogaster*, cel: *Caenorhabditis elegans*, ath: *Arabidopsis thaliana*, sce: *Saccharomyces cerevisiae*, cho: *Cryptosporidium hominis*, eco: *Escherichia coli*, nme: *Neisseria meningitidis*, hpy:*Helicobacter pylori*, rpr: *Ricketsia prowazekii*, bsu:*Bacillus subtilis*, sau: *Staphylococcus aureus*, lmo: *Listeria monocytogenes*, lla: *Lactococcus lactis*, lpl: *Lactobacillus plantarum*, cau: *Chloroflexus aurantiacus*, mge: *Mycoplasma genitalium*, mtu: *Mycobacterium tuberculosis*, blo: *Bifidobacterium longum*, ctr: *Chlamydia trachomatis*, pcu:*Protochlamydia amoebophila*, bbu:*Borrelia burgdorferi*, syn: *Synechocystis* sp., bth: *Bacteroides thetaiotaomicron*, dra: *Deinococcus radiodurans*, aae:*Aquifex aeolicus*, mja: *Methanocaldococcus jannaschii*, ape:*Aeropyrum pernix*, neq:*Nanoarchaeum equitans*.Blue bar: old KAAS, Red bar: new KAAS. **Figure S9.** Effect of database dependency on accuracy of the KO assignment. *Escherichia coli* isolated from Norwegian infant (Draft genome sequenced by 454 GS FLX Titanium). Blue diamonds show the results using the data set without proteins from the genera *Escherichia, Salmonella, Shigella,* and *Yersinia* (1,239 species). Similarly, red squares, green triangles, and purple dots show the results without proteins from the order *Enterobacteriales* (1,200 species), class *Gammaproteobacteria* (1,040 species), and phylum *Proteobacteria* (755 species), respectively. KO identifiers specific to the genera *Escherichia, Salmonella, Shigella,* and *Yersinia* (16 KO identifiers), order *Enterobacteriales* (90), class *Gammaproteobacteria* (203), or phylum *Proteobacteria* (370) were removed in advance from the protein data set. Here, the accuracy is defined by the sensitivity TP/(TP+FN), where TP and FN are the numbers of true positives and false negatives, respectively. We also used truncated proteins to confirm effect of amino acid (a.a.) sequence lengths on the accuracy of KO assignments. The 4,410 proteins from *E. coli* isolate were randomly fragmented into 50, 60, 80, 100, 120, 150, and 200 a.a. in length, and each length of a.a. sequences was used for verification of the accuracy of KO assignment.

### Abbreviations

KEGG: Kyoto Encyclopedia of Genes and Genomes; KO: KEGG Orthology; KAAS: KEGG Automatic Annotation Server; COGs: Clusters of orthologous groups; GABA: γ-aminobutyric acid; NAD: Nicotinamide adenine dinucleotide; ABC: ATP-binding cassette.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contribution

HT conceived the study and performed data analysis throughout the study. TT redefined the KEGG modules and analyzed the data. SG, YM, and MK confirmed the KEGG modules. TK performed data analysis of the human microbiome. HT, TK, and SG wrote the manuscript. All authors have read and approved the final manuscript.

### Acknowledgements

We wish to thank Prof. T. Itoh of Tokyo Institute of Technology for his help in analyzing the metagenomic samples of the microbiome from the 13 healthy individuals and W. Arai of JAMSTEC for his analytical assistance. Thanks are also due to Drs. M. Kawai & Y. Takagi of JAMSTEC for their useful suggestions. Finally, we thank Dr. G. M. Kirwan of Kyoto University for her critical reading of the manuscript and useful suggestions. This work was supported in part by a collaborative research program of Institute for Chemical Research, Kyoto University to HT and SG (grant #2012-25) and by the National Bioscience Database Center of the Japan Science and Technology Agency to SG and YM.

### Author details

[1]Microbial Genome Research Group, Japan Agency for Marine-Earth Science & Technology (JAMSTEC), 2-15 Natsushima, Yokosuka 237-0061, Japan. [2]Advanced Science & Innovation Group, Mitsubishi Research Institute Inc. (MRI), 2-10-3, Nagata-cho, Chiyoda-ku, Tokyo 100-8141, Japan. [3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [4]Department of Microbiology, Faculty of Medicine, Kagawa University, 1750-1 Miki, Kagawa 761-0793, Japan.

Received: 2 June 2012 Accepted: 6 December 2012
Published: 12 December 2012

### References

1. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW: GenBank. *Nucleic Acids Res* 2012, 40:D48–D53.
2. Altshul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403–410.

3. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40**:D290–D301.
4. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29**:22–28.
5. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC: **IMG: the integrated microbial genomes database and comparative analysis system.** *Nucleic Acids Res* 2012, **28**:27–30.
6. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, Taylor TD, Noguchi H, Mori H, Ogura Y, Ehrlich DS, Itoh K, Takagi T, Sakaki Y, Hayashi T, Hattori M: **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiome.** *DNA Res* 2007, **14**:169–181.
7. Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH: **Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment.** *Proc Natl Acad Sci USA* 2008, **105**:10583–10588.
8. Biddle JF, White JR, Teske AP, House CH: **Metagenomics of the subsurface Brazos-Trinity Basin (IOPD site 1320): comparison with other sediment and pyrosequenced metagenomics.** *ISME J* 2011, **5**:1038–1047.
9. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28**:27–30.
10. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, *et al*: **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** *Nucleic Acids Res* 2005, **33**:5691–5702.
11. Moriya Y, Itoh M, Okuda S, Yoshizawa A, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Res* 2007, **35**:W182–W185.
12. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA: **The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes.** *BMC Bioinformatics* 2008, **9**:386.
13. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17**:377–386.
14. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4.** *Genome Res* 2011, **21**:1552–1560.
15. Willke A, Harrison T, Wilkening J, Field D, Glass EM, Kyrpides N, Mavrommatis K, Meyer F: **The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools.** *BMC Bioinformatics* 2012, **13**:141.
16. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y: **KEGG for linking genomes to life and environment.** *Nucleic Acids Res* 2008, **36**:D480–D484.
17. Takami H: **Genomic diversity of extremophilic Gram-positive endospore-forming *Bacillus*-related species.** In *Trends in Genome Research*. Edited by Williams CR. New York: NOVA Publisher; 2006:25–85.
18. Uchiyama I: **Multiple genome alignment for identifying the core structure among moderately related microbial genomes.** *BMC Genomics* 2008, **9**:515.
19. Takami H, Takaki Y, Uchiyama I: **Genome sequence of *Oceanobacillus iheyensis* isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments.** *Nucleic Acids Res* 2002, **30**:3927–3935.
20. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Bertero MG, Bassieres P, Bolotin A, Borchert S, Borriss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Brouillet S, Bruschi CV, Caldwell B, Capuano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Cummings NJ, Daniel RA, Denizot F, Devine KM, Düsterhöft A, Ehrlich SD, Emmerson PT, *et al*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, **390**:249–256.
21. Chen XH, Koumoutsi A, Scholz R, Eisenreich A, Schneider K, Heinemeyer I, Morgenstern B, Voss B, Hess WR, Reva O, Junge H, Voigt B, Jungblut PR, Vater J, Süssmuth R, Liesengang H, Strittmatter A, Gottschalk G, Borriss R: **Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42.** *Nat Biotechnol* 2007, **25**:1007–1014.
22. Rey MW, Ramaiya P, Nelson BA, Brody-Karpin SD, Zaretsky EJ, Tang M, de Leon AL, Xiang H, Gusti V, Clausen IG, Olsen PB, Rasmussen MD, Andersen JT, Jørgensen PL, Larsen TS, Sorokin A, Bolotin A, Lapidus A, Galleron N, Ehrlich SD, Berka RM: **Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species.** *Genome Biol* 2004, **5**:r77.
23. Takami H, Nakasone K, Takaki Y, Maeno G, Sasaki R, Masui N, Fuji F, Hirama C, Nakamura Y, Ogasawara N, Kuhara S, Horikoshi K: **Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*.** *Nucleic Acids Res* 2000, **28**:4317–4331.
24. Janto B, Ahmed A, Ito M, Liu J, Hicks DB, Pagni S, Fackelmayer OJ, Smith TA, Earl J, Elbourne LDH, Hassan K, Paulsen IT, Kolstø AB, Tourasse NJ, Ehrlich GD, Boissy R, Ivey DM, Li G, Xue Y, Ma Y, Hu FZ, Krulwich TA: **Genome of alkaliphilic *Bacillus pseudofirmus* OF4 reveals adaptations that support the ability to grow in an external pH range from 7.5 to 11.4.** *Environ Microbiol* 2011, **13**:3289–3309.
25. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the human intestinal microbial flora.** *Science* 2005, **308**:1635–1638.
26. Li JV, Ashrafian H, Bueter M, Kinross J, Sands C, le Roux CW, Bloom SR, Darzi A, Athanasiou T, Marchesi JR, Nicholson JK, Holmes E: **Metabolic surgery profoundly influences gut microbial-host metabolic cross-talk.** *Gut* 2011, **60**:1214–1223.
27. Noguchi H, Taniguchi T, Itoh T: **MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes.** *DNA Res* 2008, **15**:387–396.
28. Takami H, Takaki Y, Chee GJ, Nishi S, Shimamura S, Suzuki H, Matsui S, Uchiyama I: **Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*.** *Nucleic Acids Res* 2004, **32**:6292–6303.
29. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Paslier DL, Linneberg A, Nielsen HB, Pelletier E, Renault P, *et al*: **A human gut microbial gene catalogue established by metagenomic sequencing.** *Nature* 2010, **464**:59–65.

doi:10.1186/1471-2164-13-699
**Cite this article as:** Takami *et al.*: Evaluation method for the potential functionome harbored in the genome and metagenome. *BMC Genomics* 2012 13:699.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central