Screening of xylose isomerase metabolic enzymes through metagenome approach for bioethanol production

February, 2015

Dini NURDIANI

Screening of xylose isomerase metabolic enzymes through metagenome approach for bioethanol production

February, 2015

Waseda University
Graduate School of Advanced Science and Engineering
Major in Life Science and Medical Bioscience,
Research on Biomolecular Engineering

Dini NURDIANI

Contents

Chapter 1

General Introduction

Metagenome-based screening prospects for xylose isomerase gene isolation and yeast metabolic engineering for xylose metabolism

Abstract

The conversion of Lignocellulosic raw materials, especially pentose sugar to bioethanol is being improved in recombinant *Saccharomyces cerevisiae* to obtain high ethanol productivity through metabolic engineering using microbial C5 metabolic enzymes. Introduction of microbial novel sequences that highly expressed C5 metabolic enzymes in recombinant *S. cerevisiae* strains becomes challenging in the attempts to produce efficient industrial bioethanol. Metagenomics has emerged as a powerful approach to mining novel biocatalysts through sequence- or activity-based screening. These approaches have uncovered a tremendous genetic and metabolic diversity of complex libraries deriving from metagenomes in the attempts to isolate novel catalytic enzymes and bioactive molecules from microorganisms.

1.1. Metagenome-based approach for screening metabolic enzymes

Nature is a rich source of both prokaryotic and eukaryotic microorganisms. Prokaryotic microorganisms are everywhere with the amount accounting for about 4-6x10^{30}. They have evolved and accumulated functional and physiological diversity, thus becoming part of the genetic diversity of the world's main reserve. However, approximately 95-99% of microorganisms cannot be cultured by standard laboratory techniques (Li et al., 2009). Uncultivated microorganisms can be accessed by employing culture-independent methods in order to figure out the genetic diversity, population structure and ecological performance of the bulk of organisms (Singh et al., 2008). The total DNA extracted from environmental samples referred as metagenome is assuring the source for uncovering novel biocatalyst (Kotik et al., 2009).

Tapping compounded microbial communities composing both cultivable and non-cultivable using metagenomic approach has appeared as a surrogate method to the conventional culture-dependent screening method. It facilitates expanded screening of microbial genomes in their natural environment.

1.1.1. Microbial and functional gene diversity

The Natural environment harbors a huge number of prokaryotic microbial diversity, and thus a great concern to analyze its functional diversity emerges. 16S ribosomal RNA gene sequence is used to measure the phylogenetic relationship and the biodiversity in prokaryotes and the analysis of those huge numbers of sequences which prove that nature harbors tremendous numbers of diverse microorganisms (Ferrer et al., 2009).

Part of the natural populations are uncultivated microbial communities which are considered as an unexplored repository of genetic and metabolic diversity.

Systematic characterization of these diversity will not only yield novel functionalities but also contribute in giving new knowledge of metabolic activities and mutual reliance behind microbial life in diverse habitats, and the roles of individual microorganisms in the ecosystem, and thereby yields new insights and clues that will aid to the cultivation of currently uncultured microbes (Ferrer et al., 2009). However, as they have immensely high diversity and uncultivated status, defining microbial diversity and its ecological role from environmental samples are very challenging. Next generation sequencing technologies provide large numbers of sequencing-based metagenomics that support new insights to the taxonomic diversity, metabolic potential, and ecological roles of microbial communities in various habitats (Zhou et al., 2013).

Two strategies that are generally used to screen and identify novel biocatalysts from metagenomic libraries are sequence-based and function-based screening.

1.1.2. Sequence-based screening

A sequence-based screening is relied on known conserved sequences. It can unveil target genes although it neglects the gene expression and protein folding in the host and discounts the full length of the gene's sequence (Li et al., 2009). The oligonucleotide primers need to be designed so they can reflect conserved amino acid sequence that matches with unknown target genes (Lorenz et al., 2002). Though the environment has a great molecular diversity, novel enzyme sequence can be retrieved by using this approach, as demonstrated for some enzyme classes such as polyketide synthases (Feng et al 2011), fumarase (Jiang et al., 2010), esterase (Park et al., 2011), etc. The advantage of the sequence-based screening approach is that it can access sequence space functionality validated by natural evolution, whereas it might remain

uncovered by expression-based screening due to undetected expression of particular gene function in the heterologous hosts. Partial gene sequences from metagenomic sources obtained by sequence-based screening may even be applied for gene shuffling experiments (Lorenz et al., 2002).

The disadvantage of sequence-based screening approach is its failure to discover basically distinct 'new' genes. Furthermore, there are challenges connected to this approach such as: (1) to measure a complex community needs a large sequencing effort. Recent sequencing technologies support the exploration of microbial communities in environmental samples resulting in greater output sequence reads, such as the next generation 454 pyrosequencing. (2) Metagenomic approach captures representative DNAs from diverse organisms with a different variety of sizes that may cause unassembled sequence reads. Sequence assembly is limited by the fragments length. Therefore, the output fragment should meet the required length to obtain full length of the gene target (3) bioinformatics tools are required for advanced data analysis which is not only based on sequence homology but also the prediction of protein structures, putative catalytic sites and activities (Li et al., 2009).

Sequence-based screening is usually performed by the following steps: (1) multiple sequence alignment of genes encoding a specific enzyme; (2) degenerate primers design based on conserved region; (3) isolation of environmental DNA; (4) amplification of the target DNA using degenerate primers with environmental DNA as the template; (5) sequence identification and specific primer design based on these sequences; (6) retrieval of full length genes by genome-walking PCR to determine unknown upstream and downstream sequences. Retrieving full length genes from metagenome DNA by using PCR-based strategy can rapidly expand the tools for obtaining future biocatalysts. This technique generates novel enzymes although the

enzymes retrieved are related in some point to the previous known enzymes (Kotik et al., 2009).

1.1.3. Function-based screening

Function-based screening does not require information of known sequence to detect full length genes or to identify the synthesis of new bioactive compounds and it is selective for full length genes and functional gene products as it employs a screening host that functionally expresses the targeted enzyme. Therefore, it has a prospect to detect an entirely new gene function (Daniel, 2004). However, this screening approach has a disadvantage, as it depends on the expression of the cloned genes in the heterologous host (such as *E. coli*) as well as the function that is encoded by gene cluster (Yun and Ryu, 2005). Generally, *E. coli* strains that are used as heterologous host have flexible prerequisites for their promoter recognition and translation initiation. However, many genes isolated from environmental samples have difficulty to express in their foreign host, possibly due to the reasons, such as dissimilarity of codon usage, transcription and/or translation initiation signals, protein-folding elements, post-translational modifications (such as glycosylation), or toxicity of the active enzyme (Li et al., 2009). Therefore, a high-throughput screening methods need to be established to obtain craved traits from a broad range libraries (Yun and Ryu, 2005).

In the function-based method, total metagenomic DNA is subjected to fragmentation into different sizes which then inserted into appropriate vectors distinguished based on its size on plasmids, cosmids, fosmids, or phages in order to create metagenome libraries. These libraries are subsequently introduced into an available host screening system. Construction of DNA libraries depends on the size of

the fragment inserted into the vector, such as 2-10 kb fragments, which are suitable for plasmid or lambda vectors. While a gene cluster with insert length 20-40 kb is constructed in cosmids and fosmids, and up to 100-200 kb in bacterial artificial chromosome vectors (Li et al., 2009).

Even though function-based screening has identified novel enzymes, such as lipase and esterase (Hong et al., 2007), glucosidase (Jiang et al 2010) or xylanase (Gong et al 2013) from metagenomes, there are some challanges connected to this method that result in no or low expression of genes that derive from metagenome in the screening host. Therefore, several attempts have been done to overcome these problems by establishing new vectors, host strains and transposable promoters, to achieve more efficient transcription of metagenomic DNAs or by establishing high-throughput methods (Singh et al., 2008).

1.2. Utilization of C5 metabolic enzymes for yeast metabolic engineering to produce bioethanol

Growing concern of bioethanol production from renewable sources such as plant biomass is supposed to have the potential to replace large part of fossil fuel that is being used today (Bertilsson et al., 2007) and it is considered as an environmentally friendly alternative energy (Bettiga et al., 2008). Efficient lignocellulose conversion into ethanol are dependent on the consumption of all fermentable sugars such as glucose, xylose and arabinose, which are available in the lignocellulosic residues at high rate and yields (Wisselink et al., 2009). Since xylose abundance comprises up to 25% of the total available pentose sugar in some hydrolysates (van Maris et al., 2007) and it is considered as the most abundant sugar after glucose (Grotkjær et al., 2005),

xylose conversion into ethanol becomes significantly important for the efficient bioethanol production from lignocellulosic materials.

Members of bacteria and fungi are known to have ability to utilize xylose and convert it into ethanol. However, yeasts have more advantages than the others for better productivity in producing ethanol from xylose (Chu and Lee, 2007). *Saccharomyces cerevisiae* (baker's yeast) is a well known hexose-fermenting yeast and presently becomes the organism of choice for industrial ethanol production (Wisselink et al., 2009; Chu and Lee, 2007). Some excellent qualifications of *S. cerevisiae,* which made it chosen as an industrial strain, are its high resistance to ethanol, its capacity of being cultivated under stringently anaerobic conditions and its insensitivity to contamination of bacteriophage or lactic acid bacteria and important characteristics differentiating it from prokaryotic organisms such as larger sizes, thicker cell wall and better growth at low pH (Jeffries, 2006; van Maris et al 2007). Although its wild type strains could efficiently ferment hexose sugar, they lack the ability to utilize pentose sugar D-xylose and L-arabinose although they have an entire xylose metabolic pathway (Wisselink et al., 2009; Chu and Lee, 2007). Improvement of xylose fermentation biotechnology in *S. cerevisiae* is attempted concerning the fact that *S. cerevisiae* considerably produces higher ethanol yields and productivity in glucose fermentation than xylose fermentation by naturally pentose-fermenting yeasts (Chu and Lee, 2007). Therefore, xylose fermentation in *S. cerevisiae* is possibly established by the introduction of a heterologous pathway from naturally-fermenting xylose organisms to convert xylose to ethanol. In order to obtain *S. cerevisiae, which* expresses a pentose utilization pathway, the genes encoding enzymes involved in pentose fermentation isolated from naturally pentose-utilizing bacteria and fungi have been introduced into *S. cerevisiae* (Hahn-Hägerdal et al., 2007).

1.2.1. Xylose utilization pathways

The initial xylose catabolic pathway in xylose-fermenting bacteria consists of one step i.e. xylose isomerase (XI) catalyzes direct isomerization of D-xylose to D-xylulose. D-xylulose is then phosphorylated by xylulokinase (XK) to xylulose-5-phosphate (Fig. 1.2.1.1), which then goes through the pentose phosphate pathway (PPP). Anaerobic fungus *Piromyces* sp. is later found to have the same initial pathway with bacteria (Harhangi et al., 2003). In naturally xylose-fermenting fungi, initial xylose catabolic pathway involves two enzymes to convert xylose to D-xylulose. First, D-xylose is reduced to xylitol by a NADPH-dependent xylose reductase (XR), and then oxidized it into D-xylulose by a NAD^+-dependent xylitol dehydrogenase (XDH) (Fig. 1.2.1.1) (Hahn-Hägerdal et al., 2007). The difference in cofactor specificity in xylose-fermenting fungi can produce redox imbalance. XR catalytic reaction requires NADPH, which is produced from D-xylose carbon through oxidative PPP. It results in a loss of some carbon as CO_2, which reduces the ethanol yield on D-xylose (van Maris et al., 2007).

Although wild-type *S. cerevisiae* strains are unable to metabolize xylose, they are capable of very slow growth on xylulose (Jeppson et al., 1996). *S. cerevisiae* also capable of xylose transport and D-xylulose can go through PPP via the native *S. cerevisiae* xylulokinase (Fig. 1.2.1.2) (Kuyper et al., 2003; Eliasson et al., 2000). Therefore, the introduction of xylose initial pathway converting xylose to xylulose may allow xylose conversion into ethanol in *S. cerevisiae.*

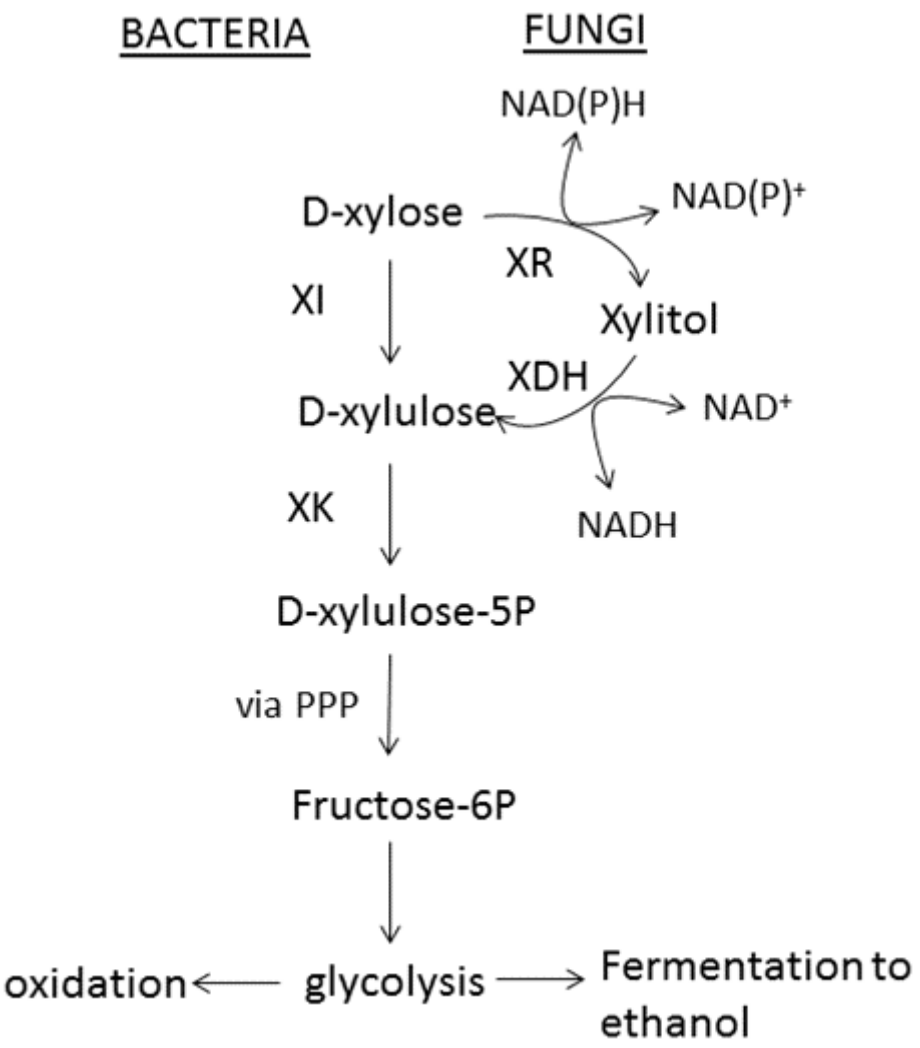


Fig. 1.2.1.1. Xylose metabolic pathway in bacteria and fungi (Hahn-Hägerdal et al., 2007).

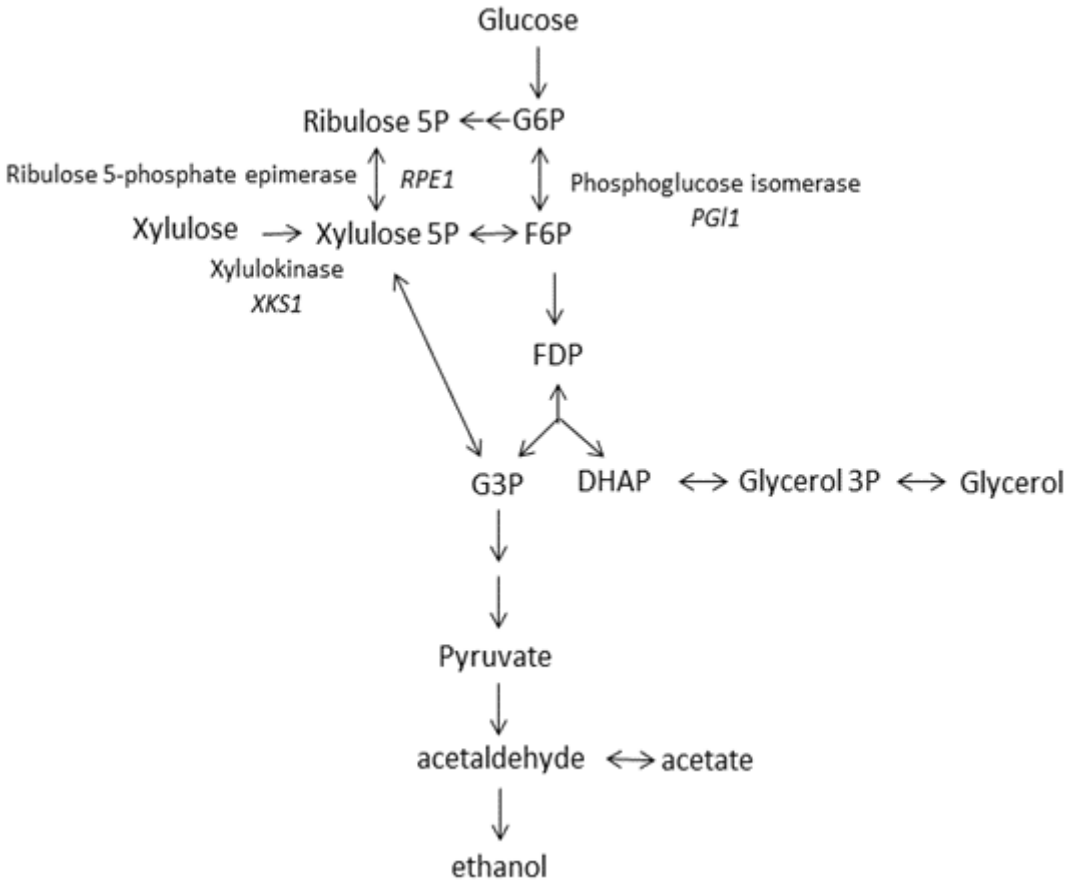


Figure 1.2.1.2. Glycolysis scheme and xylulose metabolism in yeast (Eliasson et al., 2000).

1.2.2. XR and XDH expression in *S. cerevisiae* recombinant strains

The genetic modification to generate an efficient xylose-fermenting *S. cerevisiae* has become one of the considerable challenges in yeast metabolic engineering (Kuyper et al., 2005). Most of the xylose-utilizing fungi that employ NADPH-dependent XR activity and NAD^+-dependent XDH produce considerable amounts of xylitol from xylose. The dissimilar cofactor between XR and XDH may inhibit ethanolic fermentation by yeast. The Heterologous expression of *XYL1* and *XYL2* genes encoding XR and XDH respectively, from *Pichia stipitis* in *S. cerevisiae*

with the endogenous *XKS1* gene encoding xylulokinase, results mainly in the transformation of xylitol in xylose consumption, with biomass and ethanol being minor products (Zaldivar et al., 2002). Grotkjær et al., (2005) reported that the deletion of *gdh1* (gene encoding NADPH-dependent glutamate dehydrogenase) and overexpression of *GDH2* (gene encoding NADH-dependent glutamate dehydrogenase) in recombinant *S. cerevisiae* strain CPB.CR4 resulted in a shift of xylose reductase activity where it preferred to using NADH to NADPH as a cofactor. This shift is favourable to fix redox imbalance and it may be related to the 25% increase of ethanol yield in strain CBP.CR4 compared to the reference strain TMB3001. Matsuhika and Sawayama (2008) examined the role of XR and XK activities affecting xylose in fermenting abilities of recombinant *S. cerevisiae* expressing XDH. They observed that strain I-PGK/AUR which has high activity of both XR and XDH and moderate XK activity has the most increased ethanol yield and lowered xylitol formation.

1.2.3. XI expression in *S. cerevisiae* recombinant strains

Unlike xylose metabolic pathway by XR and XDH, which have dissimilar cofactor preferences, it was thought that heterologous expression of an XI is promising since it can avoid cofactor imbalance. Therefore, several XI genes have been introduced into *S. cerevisiae*. However, after many years, no actively expressed XI enzyme has been reported (Hahn-Hägerdal et al., 2007). The first functionally bacterial XI in *S. cerevisiae* was reported from *Thermus thermophilus*. However it has low activity at 30^{o}C, the optimum temperature of *S. cerevisiae* growth. The breakthrough in the attempts to express XI gene in *S. cerevisiae* is the discovery of an XI from anaerobic fungus *Piromyces* sp. E2, which has an enzymatic function that

resembles a condition in bacteria (Harhangi et al., 2003). The cell extracts of *Piromyces* sp. E2 lack detectable activities of the two key enzymes (XR and XDH) for xylose metabolism in eukaryotic organisms. Surprisingly, XI and XK activity of *Piromyces* sp. E2 detects 0.08 U (mg protein)$^{-1}$ and 0.2 U (mg protein)$^{-1}$ respectively during growth on xylose. It is revealed that *Piromyces* sp. E2 shares bacterial XI in its xylose metabolic pathway (Kuyper et al., 2003). The selected clones sequence (pR3 and pAK44) from cDNA libraries of *Piromyces* sp. E2 show high similarities to XI and XK genes, respectively. It is predicted that XI pathway found in anaerobic fungus *Piromyces* sp. may be caused by horizontal gene transfer (Harhangi et al., 2003).

Several studies on metabolic engineering of *S. cerevisiae* have been reported by using *Piromyces*'s XI. XI is encoded by one gene named *xylA*. Kuyper et al. (2003) reported that the expression of *Piromyces xylA* in *S. cerevisiae* results in high heterologous isomerase enzyme activities in the cell extract observed (0.3-1.1 μmol (mg protein^{-1}) min^{-1} at 30^{o}C) (Table 1). *S. cerevisiae* cannot utilize D-xylose as the sole carbon source. However, it has genes that encode nonspecific NADH-dependent aldose reductase (*GRE3*) and a xylitol dehydrogenase (*XYL2*) (van Maris et al., 2007). Metabolic engineering has been done to an XI-expressing *S. cerevisiae* strain for rapid anaerobic xylose fermentation by deleting the *GRE3* gene to reduce xylitol production. This study shows that during the growth of engineered strain on xylose, xylulose formation is not present and it produces minor xylitol (Kuyper et al., 2004). The *S. cerevisiae* engineered strain with *Piromyces xylA* (RWB 217) is also studied for growth improvement in mix sugar utilization. This culture exhibits improved xylose consumption in glucose-xylose mixture (Kuyper et al., 2005).

Recently, several attempts to improve xylose and arabinose fermentation using bacterial XI in *S. cerevisiae* have been reported. Brat et al. (2009) have screened

nucleic acid database for sequences encoding putative XIs. They successfully cloned and expressed a highly active and new kind of XI from *Clostridium phytofermentans* in *S. cerevisiae* (Table 1). This new enzyme has low sequence similarities to the XI from *Piromyces* sp. Strain E2 and *Thermus thermophilus*. The performance of two isomerases within yeast cells were shown from the reaction velocities (Vmax). Conversion D-xylose to xylulose from cells containing codon optimized of the native clostridial XI gene catalyzed at a rate 0.0344 μmol.min^{-1}.mg protein^{-1} whereas in codon optimized *xylA* variant from *Piromyces* strain E2, a Vmax of 0.0538 μmol. min^{-1}.mg protein^{-1} was determined. This Vmax reveals higher XI activity in *Piromyces* strain E2 than the clostridial XI. Parachin and Gorwa-Grauslund (2011) reported two XI genes, *xym*1 and *xym*2, isolated from a soil metagenomic library by sequence and activity-based screening. These two genes have identity to the XI of *Sorangium cellulosum* (83%) and *Robiginitalea biformata* (67%).

Table 1 shows the comparison of the xylose fermentation performance between *Piromyces*'s XI, which is known as one of the best *xylA*-expresssing recombinant yeasts among *S. cerevisiae* recombinant strains (Table 1). Recent reported literature about *Orpinomyces xylA* gene which is constitutively expressed in *S. cerevisiae* strain ADAP8 at 35^{o}C in the presence of borate shows that the yeast yields the highest ethanol (0.48 g.g^{-1}) and XI activity (1.76 U.mg protein^{-1}) compared to other yeast strains (Table 1). These results indicate that genetically engineered *S. cerevisiae* has the prospect to produce ethanol from xylose and mixed sugar glucose-xylose through XI pathway.

The attempt to improve *S. cerevisiae* strains to utilize pentose sugars has been done in the engineered both laboratory (BWXY02.XA) and industrial (TMB3400) *S. cerevisiae* strains by introducing a fungal xylose and a bacterial arabinose pathway

(*Bacillus substilis AraA* and *E. coli AraB*) for examining their performace to co-ferment the pentose sugars D-xylose and L-arabinose. It results in strains capable of growing on both pentose sugars. In an arabinose-fermenting laboratory strain, introduction of a xylose pathway resulting almost entire arabinose is converted into arabitol due to the L-arabinose activity of the XR. Conversely, the industrial strain has shown lower arabitol yield and increased ethanol yield from xylose and arabinose (Karhumaa et al., 2006). Bettiga et al. (2008) introduced the bacterial arabinose isomerase (AI) pathway that consists of L-arabinose isomerase (*B. subtilis AraA*), L-ribulokinase (*E. coli AraB*) and L-ribulose-5-phosphate 4-epimerase (*E. coli AraD*) combined with two different xylose utilization pathways: the XR/XDH (*XYL1* and *XYL2 P. stipitis*) and XI (*Piromyces sp.*) pathway, respectively into *S. cerevisiae* to provide arabinose and xylose utilization. Higher uptake and ethanol yields of pentose sugar have been achieved by combining XR/XDH and bacterial AI pathway than the XI pathway and bacterial AI pathway.

Table 1. Comparison of xylose isomerase activity and ethanol yield in several recombinant strains of *S. cerevisiae*

Strain	Gene source	Growth condition	substrate	Ethanol Yield (g ethanol) $(g.sugar)^{-1}$	Xylose isomerase activity [U $(mg\ protein)^{-1}$]	Reference
RWB 202	*TPI1*/*XYLA Piromyces*	aerobic	glucose+xylose	0	1.10	Kuyper et al. 2003
		anaerobic	glucose+xylose	0.39	0.55	
BJ1991	GAL1/ *XYLA Piromyces*	anaerobic	xylose	Ns	0.025	Harhangi et al. 2003
BWY10Xyl	*codon-optimized XYLA C. phytofermentans*	anaerobic	xylose	Ns	0.0344	Brat et al. 2009
	XYLA Piromyces	anaerobic	xylose	Ns	0.0538	Brat et al. 2009
TMB 3076	XI *Piromyces* synthetic, *B. subtilis AraA*, mutated *E. coli AraB* and *E. coli AraD*	anaerobic	glucose+arabinose+xylose	0.18	ns	Bettiga et al. 2008
ADAP8	*XYLA Orpinomyces*/*XKS1*/*SUT1*, xylose adapted	anaerobic	complex media	0.48	1.72	Madhavan et al. 2009

ns = not stated

TPI1=constitutive promoter, *XYLA*= xylose isomerase gene, *GAL1*= galactose inducible promoter, *AraA*= L-arabinose isomerase gene, *AraB*= L-ribulokinase gene, *AraD*= L-ribulose-5-phosphate-epimerase gene, *XKS1 = S. cerevisiae* xylulokinase gene, *SUT1= P. stipitis* sugar transporter gene.

1.3. Objectives and significance of this study

In this work, both sequence- and function-based screening approaches were employed to obtain novel *xylA* genes from soil metagenome. Metagenome DNA from soils differed by plant vegetation was used as samples. The study consists of four chapters, in the first three chapters sequence-based screening approach was employed. First is the investigation of microbial and xylose isomerase genes diversity in soil metagenomes. In order to evaluate the complex community in the soil, 454 pyrosequencing-based technology was employed to obtain large sequence reads. The first chapter clarifies the diversity of *xylA* from soil metagenome samples and the partial *xylA* sequences obtained from the first study that can be used for subsequent study described in the second chapter. The second chapter is the identification of the partial *xylA* sequences as valuable information to retrieve unknown parts of upstream and downstream of the gene sequences by inverse PCR method in order to retrieve the full length *xylA* genes. The third chapter presents the expression of full length *xylA* genes retrieved from soil metagenomes in *S. cerevisiae* by cell surface display. On the other hand, the fourth chapter presents the attempts to establish a high-throughput screening of soil metagenome DNA, which was done by introducing partial metagenomic *xylA* into *Piromyces xylA* forming chimeras through homologous recombination in *S. cerevisiae*. The chimeras are then directly screened for their functionality in *S. cerevisiae* by assessing their ability to grow on xylose substrate. In this part, both sequence and activity-based screening are employed in order to obtain novel active *xylA* gene from soil metagenome in *S. cerevisiae*.

The ultimate goal of this study is to obtain the novel xylose isomerase genes from soil metagenome that is highly expressed in *S. cerevisiae*. As the heterologous expression of xylose isomerase genes in *S. cerevisiae* is one step to allow entire

pathway that improve the ability of *S. cerevisiae* to utilize xylose. This improvement is expected to enhance *S.cereviciae* productivity to produce ethanol from lignocellulosic biomass that is sufficient for industrial scale.

References

Bertilsson M, Andersson J, Liden G. 2007. Modelling simultaneous glucose and xylose uptake in *Saccharomyces cerevisiae* from kinetics and gene expression of sugar transporters. *Bioprocess Biosyst Eng*. 31: 369-377.

Bettiga M, Hahn-Hägerdal B, Gorwa-Grouslund MF. 2008. Comparing the xylose reductase/xylitol dehydrogenase and xylose isomerase pathway in arabinose and xylose fermenting *Saccharomyces cerevisiae* strains. *Biotechnol Biofuels*. 1:16(1-9).

Brat D, Boles E, Wiedemann B. 2009. Functional expression of a bacterial xylose isomerase in *Saccharomyces cerevisiae*. *Appl Environ Microbiol*. 75: 2304-2311.

Chu BCH, Lee H. 2007. Genetic improvement of *Saccharomyces cerevisiae* for xylose fermentation. *Biotechnol adv*. 25: 425-441.

Daniel R. 2004. The soil metagenome – a rich resource for the discovery of novel natural products. *Curr Opin Biotechnol.* 15(3):199-204.

Eliasson A, Boles E, Johansson B, Östenberg M, Thevelein JM, Spencer-Martins I, Juhnke H, Hahn-Hägerdal B. 2000. Xylulose fermentation by mutant and wild-type

strains of *Zygosaccharomyces* and *Saccharomyces cerevisiae*. *Appl Microbiol Biotechnol*. 53: 376-382.

Feng Z, Kallifidas D, Brady SF. 2011. Functional analysis of environmental DNA-derived type II polyketide synthases reveals structurallydiverse secondary metabolites. *Proc Natl Acad Sci U S A*. 108(31):12629-34.

Ferrer M, Beloqui A, Timmis KN, Golyshin PN. 2009. Metagenomics for Mining New Genetic Resources of Microbial Communities. *J Mol Microbiol Biotechnol.* 16(1-2):109-23.

Gong X, Gruniniger RJ, Forster RJ, Teather RM, McAllister TA. 2013. Biochemical analysis of a highly specific, pH stable xylanase gene identified from a bovine rumen-derived metagenomic library. *Appl Microbiol Biotechnol*. 97(6):2423-31.

Grotkjær T, Christakopoulos P, Nielsen J, Olsson L. 2005. Comparative metabolic network analysis of two xylose fermenting recombinant *Saccharomyces cerevisiae* strains. *Metab Eng*. 7: 437-444.

Hahn-Hägerdal B, Karhumaa K, Jeppsson M, Gorwa-Grauslund MF. 2007. Metabolic engineering for pentose utilization in *Saccharomyces cerevisiae*. *Adv Biochem Eng Biotechnol.* 108: 147-177.

Harhangi RH, Akhmanova AS, Emmens R, van der Drift C, de Laat WTAM, van Dijken JP, Jetten MSM, Pronk JT, Op den Camp HJM. 2003. Xylose metabolism in

the anaerobic fungus *Piromyces* sp. strain E2 follows the bacterial pathway. *Arch Microbiol*. 180: 134-141.

Hong KS, Lim HK, Chung EJ, Park EJ, Lee MH, Kim JC, Choi GJ, Cho KY, Lee SW. 2007. Selection and characterization of forest soil metagenome genes encoding lipolytic enzymes. *J Microbiol Biotechnol.* 17(10):1655-60.

Jeffries TW. 2006. Engineering yeasts for xylose metabolism. *Curr Opin Biotechnol*. 17: 320-326.

Jeppsson H, Yu S, Hahn-Hägerdal B. 1996. Xylulose and glucose fermentation by *Saccharomyces cerevisiae* in chemostat culture. *Appl Environ Microbiol*. 62(5): 1705-1709.

Jiang C, Li SX, Luo FF, Jin K, Wang Q, Hao ZY, Wu LL, Zhao GC, Ma GF, Shen PH, Tang XL, Wu B. 2010. Biochemical characterization of two novel b-glucosidase genes by metagenome expression cloning. *Bioresour Technol.* 102(3): 3272-8.

Karhumaa K, Wiedemann B, Hahn-Hägerdal B, Boles E, Gorwa-Grauslund MF. 2006. Co-utilization of L-arabinose and D-xylose by laboratory and industrial *Saccharomyces cerevisiae* strains. *Microb Cell Fact*. 5:18 (1-11).

Kotik M. 2009. Novel genes retrieved from environmental DNA by polymerase chain reaction: Current genome-walking techniques for future metagenome applications. *J Biotechnol*. 144: 75-82.

Kuyper M, Harhangi HR, Stave AK, Winkler AA, Jetten MS, de Laat WT, Den Ridder JJ, Op den Camp HJ, van Dijken JP, Pronk JT. 2003. High-level functional expression of fungal xylose isomerase: the key to efficient to ethanolic fermentation of xylose by *Saccharomyces cerevisiae*. *FEMS Yeast Res*. 4(1): 69-78.

Kuyper M, Hartog MMP, Toirkens MJ, Almering MJH, Winkler AA, van Dijken JP, Pronk JT. 2004. Metabolic engineering of xylose-isomerase expressing *Saccharomyces cerevisiae* strain for rapid anaerobic xylose fermentation. *FEMS Yeast Res*. 5: 399-409.

Kuyper M, Toirkens MJ, Diderich JA, Winkler AA, van Dijken JP, Pronk JT. 2005. Evolutionary engineering of mixed-sugar utilization by a xylose-fermenting *Saccharomyces cerevisiae* strain. *FEMS Yeast Res*. 5: 925-934.

Li LL, McCorkle SR, Monchy S, Taghavi S, van der Lelie D. 2009. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnol Biofuels*. 2:10 doi:10.1186/1754-6834-2-10.

Lorenz P, Liebeton K, Niehaus F, Eck J. 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Curr Opin Biotechnol.* 13(6):572-7.

Madhavan A, Tamalampudi S, Srivastava A, Fukuda H, Bisaria VS, Kondo A. 2009. Alcoholic fermentation of xylose and mixed sugars using recombinant *Saccharomyces*

cerevisiae engineered for xylose utilization. *Appl Microbiol Biotechnol*. 82: 1037-1047.

Matsuhika A, Sawayama S. 2008. Efficient bioethanol production from xylose by recombinant *Saccharomyces cerevisiae* requires high activity of xylose reductase and moderate xylulokinase activity. *J Biosci Bioeng*. 106(3): 306-309.

Parachin NS, Gorwa-Grauslund MF. 2011. Isolation of xylose isomerases by sequence- and function-based screening from a soil metagenome library. *Biotechnol Biofuels*, 4: 9.

Park SY, Shin HJ, Kim GJ. 2011. Screening and identification of a novel esterase EstPE from a metagenomic DNA library. *J Microbiol.* 49(1):7-14.

Singh J, Behal A, Singla N, Joshi A, Birbian N, Singh S, Bali V, Batra N. 2009. Metagenomics: Concept, methodology, ecological inference and recent advances. *Biotechnol J.* 4: 480-494.

van Maris AJA, Winkler AA, Kuyper M, de Laat WTAM, van Dijken JP, Pronk JT. 2007. Development of efficient xylose fermentation in *Saccharomyces cerevisiae*: xylose isomerase as a key component. *Adv Biochem.Engin/Biotechnol*. 108: 179-204.

Wisselink HW, Toirkens MJ, Wu Q, Pronk JT, van Maris AJA. 2009. Novel evolutionary engineering approach for accelerated utilization of glucose, xylose, and

arabinose, mixtures by engineered *Saccharomyces cerevisiae* strains. *Appl Environ Microbiol*. 75: 907-914.

Yun J, Ryu S. 2005. Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb Cell Fact.* 25;4(1):8.

Zaldivar J, Borges A, Johansson B, Smits HP, Villas-Bôas SG, Nielssen J, Olsson L. 2002. Fermentation performance and intracellular metabolite patterns in laboratory and industrial xylose-fermenting *Saccharomyces cerevisiae*. *Appl Microbiol Biotechnol*. 59: 436-442.

Zhou J, Jiang YH, Deng Y, Shi Z, Zhou BY, Xue K, Wu L, He Z, Yang Y. 2013. Random sampling process leads to overestimation of β-diversity of microbial communities. *MBio*. 4(3):e00324-13.

Chapter 2

Analysis of the bacterial xylose isomerase gene diversity with gene targeted metagenomics

Abstract

Bacterial xylose isomerases (XI) are promising resources for efficient biofuel production from xylose in lignocellulosic biomass. Here, xylose isomerase gene (*xylA*) diversity was investigated in three soil metagenomes differing in plant vegetation and geographical location, using an amplicon pyrosequencing approach and two newly-designed primer sets. A total of 158,555 reads from three metagenomic DNA replicates for each soil sample were classified into 1,127 phylotypes, detected in triplicate and defined by 90% amino acid identity. The phylotype coverage was estimated to be within the range of 84.0–92.7%. The *xylA* phylotypes obtained were phylogenetically distributed across the two known *xylA* groups. They shared 49–100% identities with their closest-related XI sequences in GenBank. Phylotypes demonstrating <90% identity with known XIs in the database accounted for 89% of the total *xylA* phylotypes. The differences among *xylA* members and compositions within each soil sample were significantly smaller than they were between different soils based on a UniFrac distance analysis, suggesting soil-specific *xylA* genotypes and taxonomic compositions. The differences among *xylA* members and their compositions in the soil were strongly correlated with 16S rRNA variation between soil samples, also assessed by amplicon pyrosequencing. This is the first report of *xylA* diversity in environmental samples assessed by amplicon pyrosequencing. Our data provide information regarding *xylA* diversity in nature, and can be a basis for the screening of novel *xylA* genotypes for practical applications.

2.1. Introduction

Xylose, a five-carbon sugar abundant in hardwood and agricultural residues, is a potential resource for biofuel production. The economic fermentation of xylose using xylose-fermenting microorganisms would greatly be helpful for utilization of lignocellulose biomass (Jeffries et al., 2007). *Saccharomyces cerevisiae* is a well known ethanol producer from glucose. However, there is no report that *S. cerevisiae* obtained from natural environments to date can efficiently ferment xylose. The use of genetically modified *S.cerevisiae* has been attempted to produce bioethanol from xylose (Karhumaa et al., 2005; Karhumaa et al., 2007; Kuyper et al., 2005; Brat et al., 2009).

S. cerevisiae lacks the ability to utilize xylose but it utilizes and ferments its isomer D-xylulose (Hsiao et al., 1982). On the other hand, a part of bacteria and other fungi possesses xylose isomerases (XIs) (van Maris et al., 2007), which directly convert xylose to D-xylulose. Therefore, the introduction of *xylA*, a gene encoding XI into *S. cerevisiae,* can theoretically result in recombinant yeast cells capable of fermenting xylose to ethanol (Harhangi et al., 2003; Brat et al., 2009). XIs have been classified into two groups, group I and II, based on their length, amino acid sequence similarity and divalent cation preference (Park and Batt, 2004). Group I XIs have 380 to 390 amino acid residues while group II XIs have 440 to 460 amino acids. They share only 20 to 30% amino acid identities with each other, but the active site residues are highly conserved among the two groups and no general difference in enzymatic properties has been known so far. Their taxonomic distribution are also different; group I XIs were mainly identified in the phyla *Actinobacteria* and *Deinococcus-Thermus* while group II XIs were mainly in the phyla *Proteobacteria* and *Firmicutes*.

Nevertheless, the xylose-fermenting activities have remained insufficient in heterologous hosts with known *xylA* (Hsiao et al., 1982; Harhangi et al., 2003; Park and Batt, 2004; van Maris et al., 2007; Brat et al., 2009; Parachin and Gorwa Grauslund, 2011). Further exploration of bacterial *xylA* is required for improving their XI activities. The information as to the genetic diversity of *xylA* in nature will be helpful for efficient screening of suitable genes in *S. cerevisiae.*

As diverse microbes inhabit soil environments that contain lignocellulosic biomass, soil metagenome could be a promising resource to explore potential bacterial *xylA* genes. The *xylA* gene diversity in soil may be affected by the following: (i) soil organic carbon and nutrient content, which vary dependent on soil (Bachar et al., 2010), and (ii) xylose sugar and carbon content, which differ among plant species on soil (Johansson, 1995; Ono et al., 2003), causing a different plant litter decomposition rate. Thus, it can be expected that soils of different plant vegetation can contain different genetic diversity of *xylA* genes.

High-throughput targeted metagenomics of amplified functional gene sequences based on next generation sequencing (NGS) has provided deep insights into their diversity, distribution and ecological roles as well as microbial assemblage with particular ecological function (Howard et al., 2011;Woodhouse et al., 2013; Zheng et al., 2013; Wang et al., 2013; Lema et al., 2013). Elucidating the x*ylA* diversity in soil by gene-targeted metagenomics will provide a basis for screening of novel *xylA,* useful for bioethanol production. Parachin and Gorwa-Grauslund (2011) reported the isolation of two novel *xylA* genes from a soil metagenomic library while they did not focus on the diversity and distribution of *xylA* genes in soil. In this study, we aimed to uncover sequence diversity of *xylA* genes in three different soil metagenomes by *xylA*-targeted metagenomics. Although two degenerate primer sets were reported to

amplify *xylA* gene sequences (Parachin and Gorwa-Grauslund, 2011), the primers were designed based on only 11 *xylA* gene sequences. Since our purpose is to obtain much more divergent *xylA* gene sequences, two sets of new degenerate primers were newly designed for exploration of *xylA* diversity over group I and II *xylA* groups. The microbial community structures in the soils were also analyzed by amplicon pyrosequencing, and the association between *xylA* and 16S rRNA gene was examined.

2.2. Materials and Methods

2.2.1. Soil sampling

Soil samples were collected from three of the following sites: two were located at Tsukuba mountain (Ibaraki prefecture, Japan) (site T1 and T2) while the other was located at Shirakaba lake (Nagano prefecture, Japan) (site S1). Soil sampling was conducted in May 2009 for S1 and in June 2009 for T1 and T2. After carefully removing the surface plant litter, soil samples were collected in 50 (width) x 50 (length) x 5 (top soil depth) cm dimensions. The soil samples were placed into individual sterile plastic bags and stored at 4^{o}C. Each soil sample was homogenized and the resultant homogenates were stored at 10 g aliquots at -20^{o}C for subsequent DNA extraction.

The three sampling sites were chosen due to their different plant vegetation, which provide larger soil microbial composition and abundance. The T1 sampling site is comprised of *Criptomeria japonica* (needle-leaf tree) vegetation, while the T2 and S1 sites are comprised plant vegetation of birch-leaf tree species, *Fagus crenata* and *Betula platyphylla*, respectively. Based on a reported analysis on plant litter

composition (Johansson, 1995), holocellulose fraction in needle- and birch-leaf litter, as a whole, was fairly constant while xylan composition in birch-leaf litter is 2- to 3-fold higher than that of needle-leaf litter.

2.2.2. DNA extraction

Metagenome DNA was extracted from the aliquot homogenates of each soil sample by using ISOIL (Nippongene, Tokyo, Japan) according to the manufacturer's instruction. Final DNA concentration was measured spectrophotometrically, which yielded DNA concentrations averaging between 4.6 – 5.2 μg/g soil. Each soil DNA was extracted in triplicate, and each metagenomic DNA (total nine samples) was used as the template DNA for PCR amplification of partial *xylA* sequences.

2.2.3. PCR amplification of *xylA* and 16S rRNA genes for pyrosequencing

To amplify *xylA* genes from soil metagenome, two degenerate primer sets named set A and B, respectively, were designed based on specified conserved regions of 112 known amino acid XI sequences, collected from the NCBI database. The primer set A consisted of xyl1 (5'-TGGGGNGGNCGNGARGGNA-3') and xyl2 (5'-RAAYTSRTCNGTRTCCCARCC-3') (Fig.2.2.3.1). The two oligonucleotides were designed manually against the conserved amino acid region WGGREGY and GWDTDEF, correspondingly. The primer set B consisted of xyl30F (5'-TGTGTTTTGGGGCGGNMKNGANGG-3') and xyl30.4R (5'-GTTATGGCCCGCCADNKKNKCRTG-3') (Fig. 2.2.3.1). The primers were designed using the CODEHOP program (Rose et al., 2003) against the conserved amino acid region VFWGGREG and HEQMAGHN, respectively.

Partial fragments of *xylA* genes were amplified from the respective soil metagenomic DNA with the following primer sets containing 454 pyrosequencing adaptors (underlined), common MID sequence tags for 454 pyrosequencing (as indicated in bold), and the degenerate primer sets as describe above. The primer set designated pyro-set A comprised of forward and reverse primers, (5'-CCATCTCATCCCTGCGTGCTCCGACTCAG-**[common MID]-**TGGGGNGGNCGNGARGGNTA-3'), and (5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGRAAYTSRTCNGTRTCCCARCC-3'), and the primers set designated pyro-set B comprised of forward and reverse primers (5'-CCATCTCATCCCTGCGTGCTCCGACTCAG**-[common MID]-**TGTGTTTTGGGGCGGNMKNGANGG-3') and (5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGGTTATGGCCCGCCADNKKNKCRTG-3'). PCR amplification with pyro-set A was performed in a 25 μL mixture (total volume) containing 10 ng soil DNA, 2.5 μL 10-fold reaction Ex Taq buffer, 0.2 mM dNTP, 4 μM of each primer, and 1.25 U Ex Taq^{TM} HS polymerase (TaKaRa-Bio Inc., Ohtsu, Japan) with the following PCR condition: 2 min of initial denaturation at 94^{o}C, 25 cycles of 30 s denaturation at 94^{o}C, 30 s annealing at 68^{o}C, and 30 s extension at 72^{o}C, followed by a 7 min final extension at 72^{o}C. As for amplification using pyro-set B, a second round of PCR was conducted with the PCR products amplified by primer set B as the template. The amplification using set B and pyro-set B was performed in a 25 μL mixture (total volume) containing 10 ng soil DNA, 4 μM of each primer, and 12.5μL Premix Ex Taq^{TM} HS polymerase (TaKaRa-Bio Inc., Ohtsu, Japan) with the following PCR condition: 2 min of initial denaturation at 94^{o}C, 25 cycles of 30 s denaturation at 94^{o}C, 30 s annealing at 58^{o}C, and 15 s extension at 72^{o}C, followed by a 7 min final extension at 72^{o}C. PCR products were then analyzed by agarose gel

electrophoresis and purified using a Gel Extraction Kit (Qiagen GmbH, Hilden, Germany) according to the manufacturer's instructions.

16S rRNA genes were amplified with a universal primer set (27F/338R) (Lane, 1991) accompanied with 454 pyrosequencing adaptors (underlined) and the common MID tag sequences for each metagenome sample (as indicated in bold). The primer sequences are as follows: pyro-27F(5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG-[**common MID**]-AGAGTTTGATCMTGGCTCAG-3') as the forward primer and pyro-338R (5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAGtgctgcctcccgtaggagt-3') as the reverse primer. For each sample, PCR amplification of 16S rRNA genes was performed in a 25 μL mixture (total volume) containing 10 ng soil metagenome DNA, 12.5 μl PrimeSTAR Max DNA polymerase (Takara) and 0.2 μM of forward and reverse primers with the following PCR condition: 5 min of initial denaturation at 98^{o}C, 25 cycles of 10 s denaturation at 98^{o}C, 15 s annealing at 49^{o}C, and 5 s extension at 72^{o}C, followed by a 5 min final extension at 72^{o}C. The amplified PCR products were then analyzed by agarose gel electrophoresis.

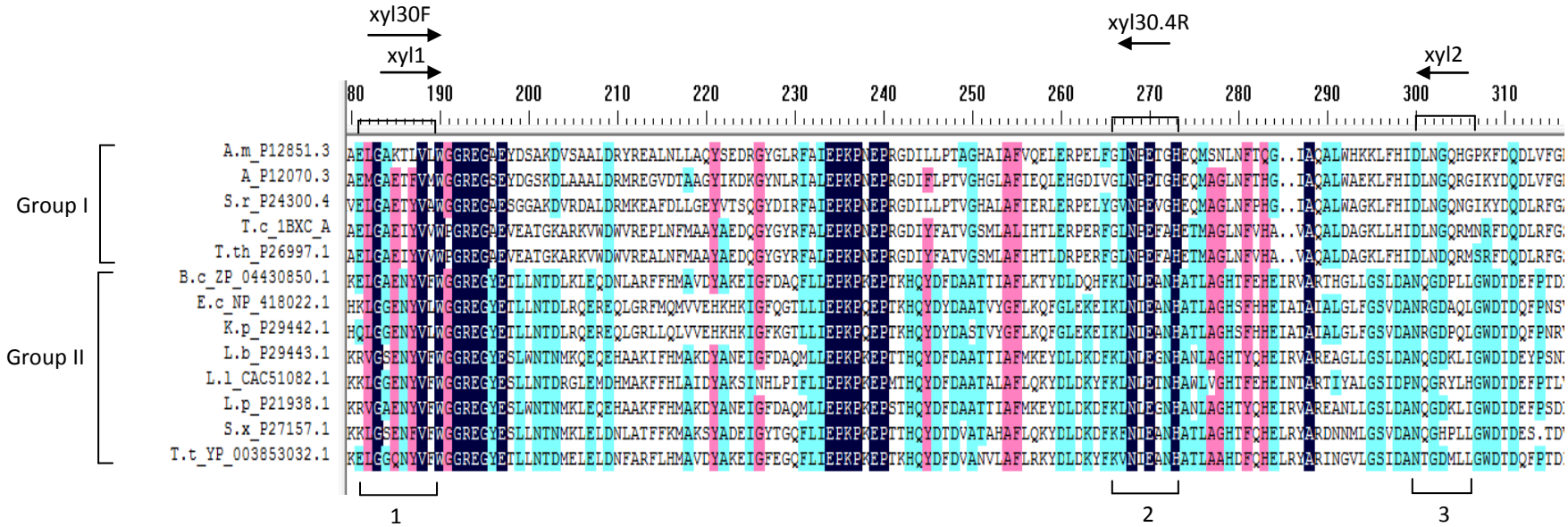


Fig. 2.2.3.1. The multiple sequence alignment of known *xylA* genes collected from the NCBI database. Species names are followed by accession numbers; A.m, *Actinoplanes missouriensis*; A, *Arthrobacter* sp.; S.r, *Streptomyces rubiginosus*; T.c, *Thermus caldophilus*; T.th, *Thermus thermophilus*; B.c, *Bacillus cereus*; E.c, *Escherichia coli*; K.p, *Klebsiella pneumoniae*; L.b, *Lactobacillus brevis*; L.l, *Lactococcus lactis*; L.p, *Lactobacillus pentosus*; S.x, *Staphylococcus xylosus*; T.t, *Thermoanaerobacterium thermosaccharolyticum*). The colors in the alignment indicate conservation among the sequences; dark blue, completely conserved; pink, ≥80% conserved; green, ≥50% conserved. The degenerate primer design for *xylA* gene amplification was based on the conserved regions. Primer set A was designed based on conserved regions 1 and 3 with a fragment length of approximately 380 bp, while primer set B was designed based on conserved regions 1 and 2 with a fragment length of approximately 290 bp. The arrows represent the primer regions.

2.2.4. Pyrosequencing

Amplicons of *xylA* and 16S rRNA genes were purified with AMPure XP magnetic purification beads (Beckman Coulter Inc., Brea, CA, USA). The quality of the purified products was examined on a high sensitivity DNA chip with Agilent Bioanalyzer 2100 system (Agilent technologies, Palo Alto, CA, USA). The products without primer dimers and other short DNA fragments (<300 bp) were quantified using Quant-iT^{TM}Picogreends DNA assay kit (Life Technologies Japan, Tokyo, Japan). Equal amplicon molecules from the nine metagenomic DNA samples (three soil metagenomic DNA were extracted each in triplicate) were mixed and used for the pyrosequencing procedure of 454 GS Junior (Roche Applied Science, Indianapolis, USA) based on the Lib-L protocol by the manufacturer. The sequence reads produced by GS Run processor were used as raw reads and subjected to further analyses.

2.2.5. Sequence data analysis

2.2.5.1. Filtering and annotation of 454 reads

The raw reads were firstly processed with 16S analysis pipeline (http://www.cb.k.u-tokyo.ac.jp/hattorilab/ja/protocol/16s_analysis; Kim et al., 2013). This pipeline discards the reads i) including ambiguous reads, ii) without tag sequence and iii) with average quality score below 25, and trims the sequence region corresponding to forward and reverse primers. The resultant reads in files with ".filtered.fna" filename extension were used for further analyses. In order to remove sequences likely not to originate from a *xylA* gene, blastx search was carried out against UniprotKB/Swiss-prot database and discarded the reads that did not show any similarity (e-value <1e^{-5}) to known XI amino acid sequences in the database. The reads passed the above steps were converted into amino acid using FrameBot (Wang

et al., 2013), a tool which can take frame shifts into consideration and translate nucleotide sequence precisely.

For taxonomic annotation of 16S rRNA gene sequences, RDP classifier was applied (Wang et al., 2007) in order to annotate taxonomic information. The taxonomic assignments were only accepted if the bootstrap values for 100 replicates were greater than 50. In contrast, BLASTX against the NCBI-nr database was used to estimate the phyla of origin for the *xylA* sequences. Any *xylA* sequence with a pairwise identity to its top blast hit sequence over 80% was assigned the original phyla. The sequences amplified with the primer set B were classified into group I and group II *xylA,* based on the phylogeny of representative sequences of each phyloptype (see the next 'phylotype clustering' section).

2.2.5.2. Phylotype clustering

The reads that passed the filtering described above were aligned with the mafft program (Katoh et al., 2002) and trimmed manually into the same length. The reads amplified by pyro-set A primers (*xylA*-set A), those by pyro-set B primers (*xylA*-set B) and 16S rRNA reads were adjusted to 110 aa, 70 aa and 270 bp, in that order. The reads shorter than the corresponding length were discarded. Thereafter, UPARSE (Edgar, 2013) was utilized to cluster *xylA* and the 16S rRNA gene sequences into phylotypes defined at 90% amino-acid and 97% nucleotide identity, respectively. Representative sequences of each *xylA* and 16S rRNA phylotype were chosen from the reads found at least three times and two times, respectively. Chao1 (Chao, 1984), ACE (Chanzdon et al., 1998) and Shannon index (Shannon et al., 1984) were determined from the defined phylotypes for evaluation of richness and diversity of *xylA* and 16S rRNA genes in soil metagenomes.

2.2.5.3. Phylogenetic analysis

Representative sequences of the 100 most abundant phylotypes were aligned together by Clustal omega program (Sievers et al., 2011). Subsequently, PHYLIP (Felsenstein, 2005) was used to calculate revolutionary distance with the alignment result based on Jones-Taylor-Thornton model. The phylogenetic tree containing XI amino acid sequences obtained in this study and those obtained in Swiss-prot database was constructed based on the neighbor-joining method and drawn by iTOL (Letunic and Bork, 2007).

2.2.5.4. UniFrac distance analysis

Based on the information of the phylotypes defined above, Fast UniFrac (Hamady et al, 2010) was used to calculate UniFrac distances and their principle coordinates. For unweighted UniFrac distance analysis, to avoid bias due to difference in sequencing depth, 5,000, 3,000 and 8,000 reads were randomly sampled and analyzed from *xylA*-set A, *xylA*-set B and 16S rRNA gene data sets, respectively. Phylogenetic trees for UniFrac analysis were constructed by neighbor-joining method using representative sequences of each phylotype.

2.2.5.5. Nucleotide sequence

The representative sequences of all triplicately-detected phylotypes of 16S rRNA gene, *xylA*-set A and *xylA*-set B were deposited in DNA Data Bank of Japan (DDBJ) under accession numbers AB992444 to AB994682, AB994683 to AB995706, and AB995707 to AB996596, respectively.

2.3. Results and Discussion

2.3.1. Sequence data acquisition and evaluation

To obtain diverse *xylA* gene sequences from our metagenomic DNA sources, two new degenerate primer sets were designed to target several conserved regions based on 112 amino acid sequences of known XI proteins (Fig.2.2.3.1). The primer set A was constructed based on the amino acid sequences WGGREGY and GWDTDEF of the region 1 and 3, respectively, which are highly conserved among group II XIs. On the other hand, the primer set B was constructed based on the amino acids VFWGGREG and HEQMAGHN of the region 1 and the region 2, respectively, primarily to detect group I *xylA*. The primer sets, set A and B were constructed to cover both groups I and II *xylA* genes (Fig. 2.2.3.1). Both of the two primer sets amplified the fragments of respective expected sizes.

Pyrosequencing of the amplicons was carried out in order to obtain sufficient XI sequence information from the soils. 256,234 raw 454 reads were attained across all samples, including 118,396 reads amplified by pyro-setA primers (*xylA*-set A) and 137,838 reads amplified by pyro-set B primers (*xylA*-set B) (Table 2.3.1.1). After subsequent quality control steps (see materials and methods section), the remaining 162,756 reads were translated into amino acid sequences by using FrameBot, a translation tool that detects and corrects frame-shift errors (Wang et al., 2013). It was used to recover the sequences with stop codons, possibly due to frame shifts derived from pyrosequencing errors (Huse et al., 2007; Zhang and Sun, 2011), increasing the number of sequences to be further analyzed by approximately 9.5%. This value was consistent with the previously-reported ratio (12.7%) of the reads with frame shift

errors by 454 pyrosequencing (Wang et al., 2013). Overall, 158,555 sequences passed all the filtering steps for the subsequent analyses (Table 2.3.1.1).

The resulting sequence datasets were subjected to phylotype classification. Since the clustering of 454 pyrosequencing reads in several widely-used clustering tools can produce more operational taxonomic units (OTUs) than the actual number of species (Quince et al., 2009; Kunin et al., 2010; Kim et al., 2013), UPARSE, the latest amplicon analysis pipeline was employed, resulting highly accurate and far fewer OTU sequences (Edgar, 2013). UPARSE contains a chimera checking step that does not need curated database or alignment of a target gene, rendering the pipeline suitable for phylotype binning of functional genes with no specific database such as *xylA*. This procedure resulted in 1,935 phylotypes containing 787 *xylA*-set A and 1,148 *xylA*-set B phylotypes across all samples (Table 2.3.1.2). Phylotypes commonly detected in triplicate metagenomic DNA were 1,136, consisting of 558 *xylA*-set A and 569 *xylA*-set B (Table 2.3.1.3). Rarefaction curves (Simberloff, 1978) derived from the obtained phylotypes were shown in Figure 2.3.1.1. The curves of *xylA*-set A and *xylA*-set B by UPARSE almost reached saturation (Fig. 2.3.1.1A and B), indicating that the number of reads sequenced was sufficient to determine x*ylA* diversity in the soil samples.

For 16S rRNA gene diversity analysis in each soil, 98,298 sequence reads were obtained across all soil samples with average read length being ca. 300 bp (Table 2.3.1.2). The 3,664 phylotypes of 16S rRNA amplicons were also obtained by using UPARSE (Table 2.3.1.2), among which common-in-triplicate phylotypes were 1526. Rarefaction curves based on all the triplicate sequence data almost reached saturation (Fig. 2.3.1.1C).

Table 2.3.1.1. Number of *xylA* sequences during data processing

Primer set	Soil	Sample no.	Number of 454 sequence reads					
			Raw	After QC	After translation			
					(-) Framebot	ratio	(+) Framebot	ratio
Set A	T1	1	9631	7707	7084	91.9	7501	97.3
		2	7574	5616	5116	91.1	5425	96.6
		3	15617	12849	11859	92.3	12551	97.7
	T2	1	13964	11554	10756	93.1	11256	97.4
		2	12064	9952	9234	92.8	9699	97.5
		3	15195	12500	11612	92.9	12218	97.7
	S1	1	18218	15039	13269	88.2	14046	93.4
		2	12362	10124	9252	91.4	9791	96.7
		3	13771	11117	10212	91.9	10851	97.6
Set B	T1	1	11699	9116	8567	95.1	8948	98.2
		2	14165	5919	5547	94.7	5820	98.3
		3	25718	3309	3060	94.0	3229	97.6
	T2	1	10634	8976	8543	95.6	8868	98.8
		2	16740	9825	9335	95.6	9692	98.7
		3	26987	6022	5647	94.8	5897	97.9
	S1	1	10357	7346	6983	95.6	7243	98.6
		2	12380	9673	1793	18.7	9493	98.1
		3	9158	6112	5640	93.0	6027	98.6
Total			256234	162756	143509	88.2	158555	97.4

2.3.2 Richness and diversity of xylose isomerase and 16S rRNA genes

Table 2.3.1.2 shows richness and diversity of phylotypes of *xylA* and 16S rRNA genes across all samples. The coverage values were estimated within the range of 84.0%-92.7% and 84.3-95.8% for *xylA* and 16S rRNA data set, respectively. Richness estimation based on both Chao1 and ACE indices for *xylA*-set A and *xylA*-set B suggested that the T2 soil sample, whose vegetation was birch leaf tree *Fagus crenata*, contained the highest number of *xylA* members. T1 and S1 soil samples, whose vegetation was needle leaf and birch leaf, respectively, contained less *xylA* genes. In case of 16S rRNA gene analysis, however, T1 and T2 soil samples showed similar richness (Table 2.3.1.2). This non-proportionality of the richness between *xylA* and 16S rRNA gene may imply that *xylA* is absent in some soil bacteria inhabiting these soils. This is often the case with soil bacteria; some genome sequences of soil bacteria such as *Pseudomonas putida* (Nelson et al., 2002) and *Sphingobium japonicum* (Nagata et al., 2011) do not contain any genes with a significant identity to *xylA*. Shannon indices were comparable among all soils for *xylA* genes (within the range of 4.30-5.03) and 16S rRNA genes (within the range of 5.97-6.35), suggesting that the diversity and evenness of *xylA* and 16 rRNA genes were similar among the three soil types.

Table 2.3.1.2. Richness and diversity of phylotypes of *xylA* and 16S rRNA gene defined with 10% amino-acid and 3% nucleotide identity, respectively.

Gene	Soil	Sample no.	Number of sequences[a]	Number of phylotypes	Richness		Coverage		Shannon index (H')
					Chao1	ACE	Chao1	ACE	
xylA-set A	T1	1	7501	425	472	482	90.1	88.1	4.69
		2	5425	389	455	452	85.5	86.1	4.76
		3	12551	485	525	528	92.5	91.8	4.86
		1+2+3.	25477	564	589	592	-	-	4.85
	T2	1	11256	510	567	569	89.9	89.6	4.96
		2	9699	483	539	534	89.5	90.4	4.97
		3	12218	521	575	570	90.5	91.4	4.93
		1+2+3	33173	626	650	653	-	-	5.00
	S1	1	14046	432	473	467	91.3	92.5	4.69
		2	9791	424	461	457	92.0	92.7	4.78
		3	10851	428	465	464	92.0	92.3	4.72
		1+2+3	34688	527	549	547	-	-	4.78
	Total of T1, T2 and S1		93338	787	-	-	-	-	-
xylA-set B	T1	1	8948	516	569	572	90.7	90.3	4.96
		2	5820	463	516	531	89.7	87.1	4.92
		3	3229	390	447	464	87.2	84.0	5.03
		1+2+3	17997	663	683	694	-	-	5.07
	T2	1	8868	571	632	644	90.4	88.7	4.82
		2	9692	569	620	627	91.8	90.8	4.70
		3	5897	523	591	606	88.5	86.3	4.86
		1+2+3	24457	792	810	821	-	-	4.87
	S1	1	7243	421	489	487	86.2	86.4	4.28
		2	9493	467	515	518	90.7	90.1	4.51
		3	6027	432	494	490	87.5	88.1	4.55
		1+2+3	22763	601	623	626	-	-	4.51
	Total of T1, T2 and S1		65217	1148	-	-	-	-	-
16S rDNA	T1	1	10112	1417	1676	1733	92.3	90.7	6.33
		2	11135	1471	1740	1795	87.7	85.8	6.35
		3	11771	1547	1795	1870	89.6	87.8	6.41
		1+2+3	33018	2196	2254	2304	-	-	6.53
	T2	1	8992	1269	1538	1592	87.1	86.9	6.11
		2	11467	1434	1647	1718	95.8	92.9	6.21
		3	10691	1408	1665	1743	84.4	84.3	6.21
		1+2+3	31150	2119	2222	2284	-	-	6.35
	S1	1	11756	1231	1433	1469	90.6	88.9	5.98
		2	10929	1216	1404	1453	90.1	88.9	5.98
		3	11445	1229	1453	1482	90.3	89.5	5.97
		1+2+3	34130	1804	1876	1911	-	-	6.11
	Total of T1, T2 and S1		98298	3664	-	-	-	-	-

[a],Sequences translated from the 454 reads that passed all quality control steps (see materials and methods)

Table 2.3.1.3. Distribution of identities between *xylA* sequences obtained and those in Genbank Database

Protein	Soil	Phylotypes common in triplicate	Pairwise identity to the closest xylose isomerase in nr database							
			< 70%		≥70 to <80%		≥80 to <90%		≥ 90%	
XylA-Set A	T1	309	4	(1.3)	96	(31.1)	166	(53.7)	43	(13.9)
	T2	385	6	(1.6)	145	(37.7)	176	(45.7)	58	(15.1)
	S1	330	7	(2.1)	134	(40.6)	143	(43.3)	46	(13.9)
	total	558	9	(1.6)	218	(39.1)	265	(47.5)	66	(11.8)
XylA-Set B	T1	261	36	(13.8)	99	(37.9)	94	(36.0)	32	(12.3)
	T2	344	31	(9.0)	147	(42.7)	121	(35.2)	45	(13.1)
	S1	285	24	(8.4)	126	(44.2)	109	(38.2)	26	(9.1)
	total	569	69	(12.1)	253	(44.5)	192	(33.7)	55	(9.7)

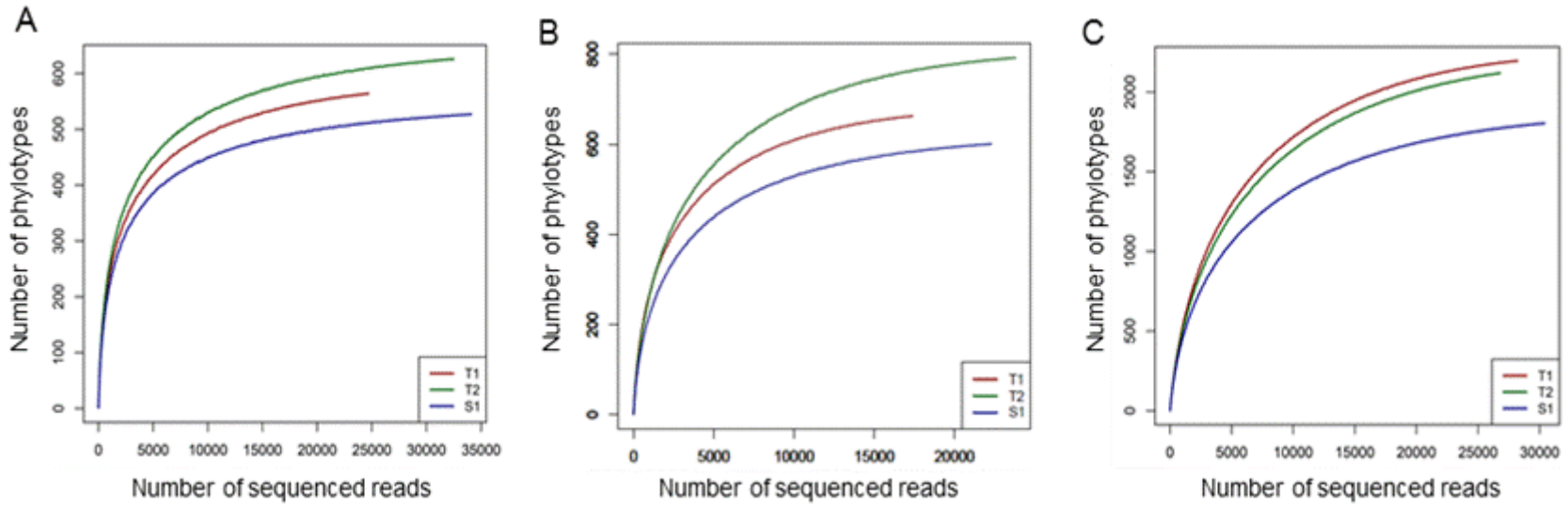


Fig. 2.3.1.1. Rarefaction curves of phylotypes derived from the pyrosequenced amplicons. The triplicate sequence data sets from each soil sample were combined and analyzed. Each panel represents the rarefaction curve of (A) *xylA*-set A, (B) *xylA*-set B and (C) 16S rRNA derived from the T1, T2 and S1 soil metagenomes. Red, T1; green, T2; blue, S1.

2.3.3. Phylogeny of xylose isomerase genes

The phylogeny of the attained *xylA* phylotypes and their similarity in amino-acid level with known XIs in the database was investigated. It seemed possible that the phylotypes produced by UPARSE still included artificial sequences which had formed during PCR amplification and/or 454 sequencing, as discussed by Edgar (2013). For appropriate evaluation of the data and future screening of *xylA* genes based on these sequence information, the artificial sequences should be carefully excluded. To extract certainly existing sequences in each soil, the focus was only on the 1,127 phylotypes commonly detected in triplicate metagenomic. The distribution of the phylotypes in each identity to known XI amino-acid sequences was summarized in Table 2.3.1.3. The *xylA* phylotypes shared 49-100% identities with each phylogenetically-closest XI in GenBank nr database. The phylotypes with ≥90% identity to known XIs (that is, within the same phylotype in this study) were only 66 out of 558 (11.8%) for *xylA*–set A and 55 out of 569 (9.7%) for *xylA*-set B. Furthermore, the phylotypes with <80% identity accounting for 40.7% for *xylA*-set A and 56.6% for *xylA*-set B. These results clearly indicate that our exploration of x*ylA* diversity in soils has uncovered a large number of *xylA* genes with sequential novelty.

Fig. 2.3.3.1 shows the phylogenetic tree of the obtained *xylA* phylotypes, combined with all XI amino-acid sequences in Uniprot/Swissprot database and their phylogenetically-closest XI in GenBank nr database described above *xylA*-set A constituted group II XI as expected, whereas *xylA*-set B showed identities to not only group I but also group II XIs. This broad amplification range of set B primers may contribute to the higher richness of *xylA*-set B than *xylA*-set A (Table 2.3.1.2). This characteristic of set B primers may be due to less conserved amino acid sequences of the region 2 among group I XIs, chosen for the set B reverse primer site (Fig. 2.2.3.1).

The relative representation of each taxonomic group based on the assignment of *xylA* genes is shown in Fig. 2.3.3.2. The group I and II XIs of *xylA*-set B were classified into each group and shown individually. Since the results were quite similar within the triplicate metagenomic samples, only the one of each representative sample (T1-1, T2-1, and S1-1) is shown. In this survey prior to the assignment of the taxonomic origins to the *xylA* sequences, known XI amino-acid sequences in the database showed up to 74% identity to a sequence in a different phylum (data not shown). To avoid misassignment, a threshold value was set at 80% amino-acid identity to known XIs derived from isolated strains. Using this threshold, the sequences of group II *xylA* including *xylA*-set A and *xylA*-set B were estimated to be derived from the phyla *Acidobacteria, Verrucomicrobia*, *Proteobacteria*, *Planctomycetes*, and *Bacteroidetes* while those of group I *xylA* were from *Acidobacteria, Actinobacteria, Chloroflexi and Armatimonadetes*. This result indicated that the *xylA* gene sequences were obtained from diverse taxonomic groups. The taxonomic members of *xylA* were almost consistent but their relative abundance was different among the three soils at phylum-level classification.The phyla observed were detected in the 16S rRNA gene-tageted analysis (Fig. 2.3.3.2), though *Verrucomicrobia* and *Planctomycetes* each represented less than 1% of sequences (these phyla are included in 'others' in Fig. 3). On the other hand, greater than 23% and 40% of group II and group I *xylA* genes, respectively, were unassigned ('unassigned' in Fig. 2.3.3.2). Those sequences additionally included XIs of the phyla *Firmicutes, Chloroflexi, Deinococcus-Thermus, Thermotogae* and unclassified phyla. For taxonomic identification of these *xylA* genes, further isolations of bacterial strains with a *xylA* gene, or the single cell-based techniques (Ishii et al., 2011; Wilson et al., 2014) will be helpful.

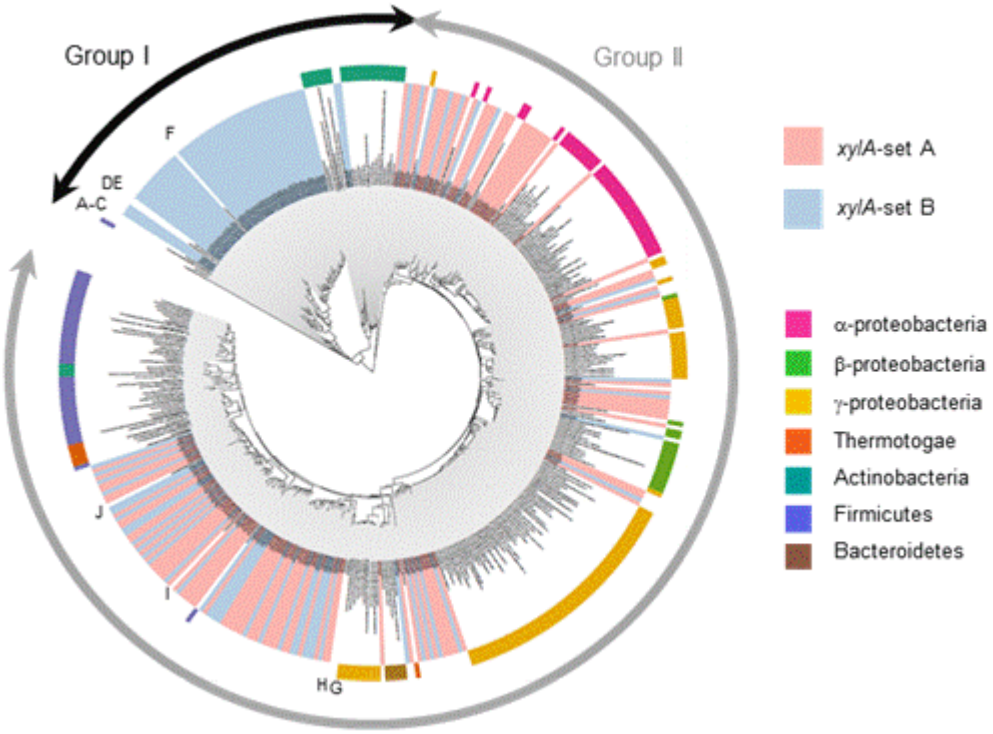


Fig. 2.3.3.1. Phylogeny of xylose isomerase (XI) phylotypes encoded by metagenomic XI gene amplicons. The phylotypes based on the primer pyro-set A (*xylA*-set A) and pyro-set B (*xylA*-set B) were highlighted in pink and blue, respectively. Black and gray arrows indicate the range of group I and group II XI, respectively. Reference XI sequences (not pink- or blue-highlighted) were derived from Uniprot/Swissprot database (with colors shown in legends) or from nr database (with alphabets). The taxonomic origins of the sequences from Uniprot/Swissprot were shown in class level (phylum *Proteobacteria*) or phylum level (the other phyla). The sequences with alphabets are the blast top hits of the XIs obtained in this study. A, *Thermus caldophilus* (phylum Deinococcus-Thermus); B, *Thermus thermophilus* (phylum Deinococcus-Thermus); C, *Herpetosiphon aurantiacus* (phylum Chloroflexi) ; D, *Roseiflexus castenholzii* (phylum Chloroflexi); E, *Thermobaculum terrenum* (unclassified); F, *Candidatus Koribacter versatiles* (phylum Acidobacteria); G, *Arabidopsis thaliana* (phylum Streptophyta); H, *Hordeum vulgare* (phylum Streptophyta); I, *Rhodopirellula baltica* (Phylum Planctomycetes); J, uncultured marine bacterium HF10 49E08 (unclassified).

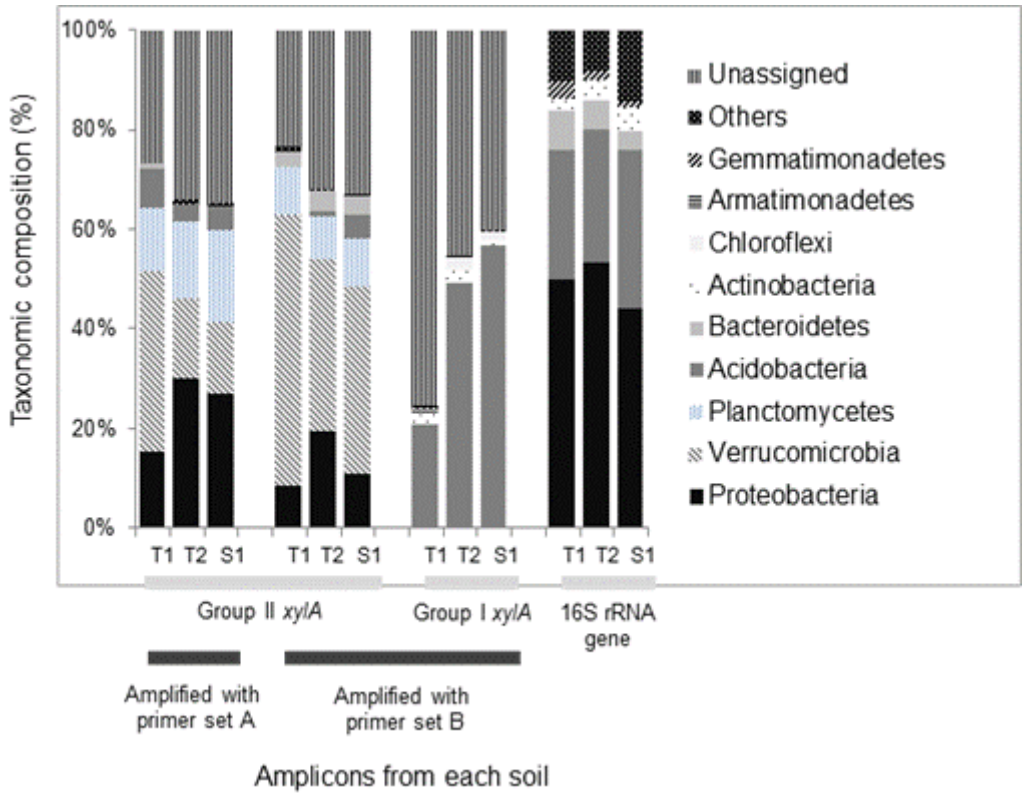


Fig. 2.3.3.2. Taxonomic composition (phylum level) of the amplicons of group II *xylA* amplified with the primer set A, group II *xylA* with the primer set B, group I *xylA* with the primer set B, and 16S rRNA gene.'Others' include the phyla whose abundance were less than 1% of total sequence reads. For *xylA* amplicons, the estimated taxonomic origin was assigned when pairwise identity to the BLASTX top hit were ≥80%. For 16S rRNA gene amplicons, the estimated taxonomic origin was assigned when the bootstrap values in the RDP classifier were more than 50%.

2.3.4. Distribution and similarity of xylose isomerase repertories in three different soil metagenomes

The distribution of *xylA* phylotypes among the three different soils was examined in Venn diagram by using common-in-triplicate phylotypes (Fig. 2.3.4.1). There were 'general phylotypes',which were shared in all soils and also 'specific phylotypes', which were detected specifically in each soil (Fig. 2.3.4.1). This study focused on the common phylotypes between any two soils and found less common phylotypes between S1-T1 soils compared to T1-T2 and T2-S1 soils; the common *xylA*-set A phylotypes between T1-T2, T2-S1, and S1-T1 were 47.0, 48.0 and 33.4% in the total phylotypes of each two soils, respectively, and those of *xylA*-set B were 30.0, 33.3 and 22.7%, respectively. Similarly, those of 16S rRNA between T1-T2, T2-S1, and S1-T1 were 50.4, 52.0, and 43.2%, respectively. These results suggest that the difference of both soil bacterial communities and *xylA* gene members were larger between S1 and T1 soils compared to the other two soil pairs.

The similarity of *xylA* gene members and compositions in each metagenomic DNA was analyzed using the principal coordinate analysis (PCoA) based on unweighted and weighted UniFrac Distance (UD) (Fig. 2.3.4.2). Unweighted UD reflects the difference of the member while weighted UD reflects the difference of both the member and the composition. In both of the analyses, triplicate *xylA*-set A and *xylA*-set B sequences from each soil clearly clustered together, strongly suggesting that each soil has each-specific x*ylA* repertories and compositions. These results were consistent with two other results: the existence of the 'specific *xylA* phylotypes' in each soil (Fig. 2.3.4.1A and B), and the phylotype diversity of *xylA* (Table 2.3.1.2), where the richness and diversity of *xylA*-set A and *xylA*-set B were similar among triplicate metagenomic DNAs. The PCoA of 16S rRNA gene

composition also resulted in the clusters consisting of each soil bacterial flora (Fig. 2.3.4.2C).

The association between the repertories and compositions of *xylA* genes and bacterial community was examined by plotting the UDs between each sample (Fig. 2.3.4.3). Overall, the UDs of *xylA*-set A and *xylA*-set B between each soil strongly correlated to those of soil bacterial communities estimated with 16S rRNA gene (Pearson correlation, 0.86-0.98). Thus, it seemed reasonable to presume that the *xylA* member and composition in soil primarily depends on the member and composition of bacteria inhabiting there. Notably, the UDs of inter soil samples could be classified into two groups; the UDs of S1-T1 were larger than those of T1-T2 and T2-S1 (Fig. 2.3.4.3). These results clearly coincided with the differences of common phylotype ratio between each pair of soil described above. The sampling sites of T1 and T2 soils were in the same mountain (Mt. Tsukuba, Japan). The plant vegetations in T2 and S1 were both birch-leaf trees although they were different in plant species. However, S1 and T1 were different in both. The soil type or the plant vegetation can affect the soil bacterial community (Hansel et al., 2008; Will et al., 2010). Bao et al. (2012) reported that soil bacterial communities were strongly affected by geographical sites, which in turn the geographical sites are strongly associated with soil characteristics. The differences of UDs of *xylA* and 16S rRNA genes might reflect the differences of plant vegetation, geographical location and soil characteristics among soil samples.

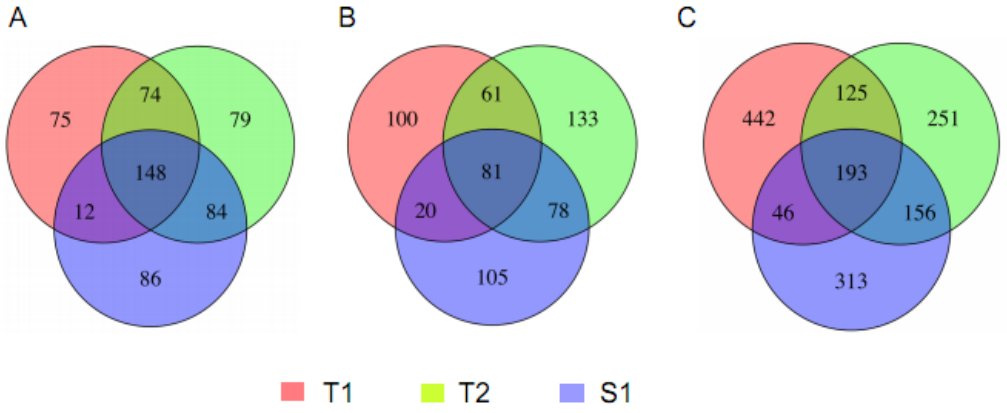


Fig. 2.3.4.1. Distribution of phylotypes of (A) *xylA*-set A, (B) *xylA*-setB and (C) 16S rRNA gene in three different soils T1, T2 and S1. The phylotypes detected commonly in triplicate experiments were used for analysis.

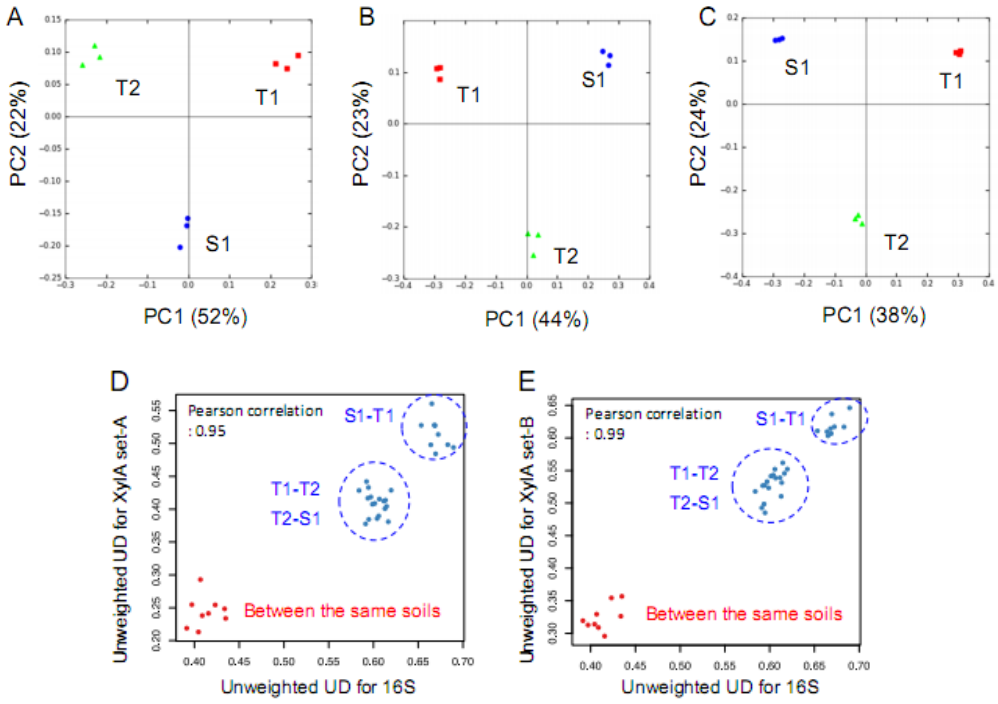


Fig. 2.3.4.2. UniFrac distance (UD) analysis. (A), (B) and (C), Principal coordinate analyses (PCoA) of the member of *xylA*-set A, *xylA*-set B, and16S rRNA genes, respectively, based on unweighted UD. (D) and (E), correlation of unweighted UDs between16S and *xylA*-set A , and 16S and *xylA*-set B, respectively. In (D) and (E),the UDs between replicate soil samples were shown in red and those between different soil samples were in blue.

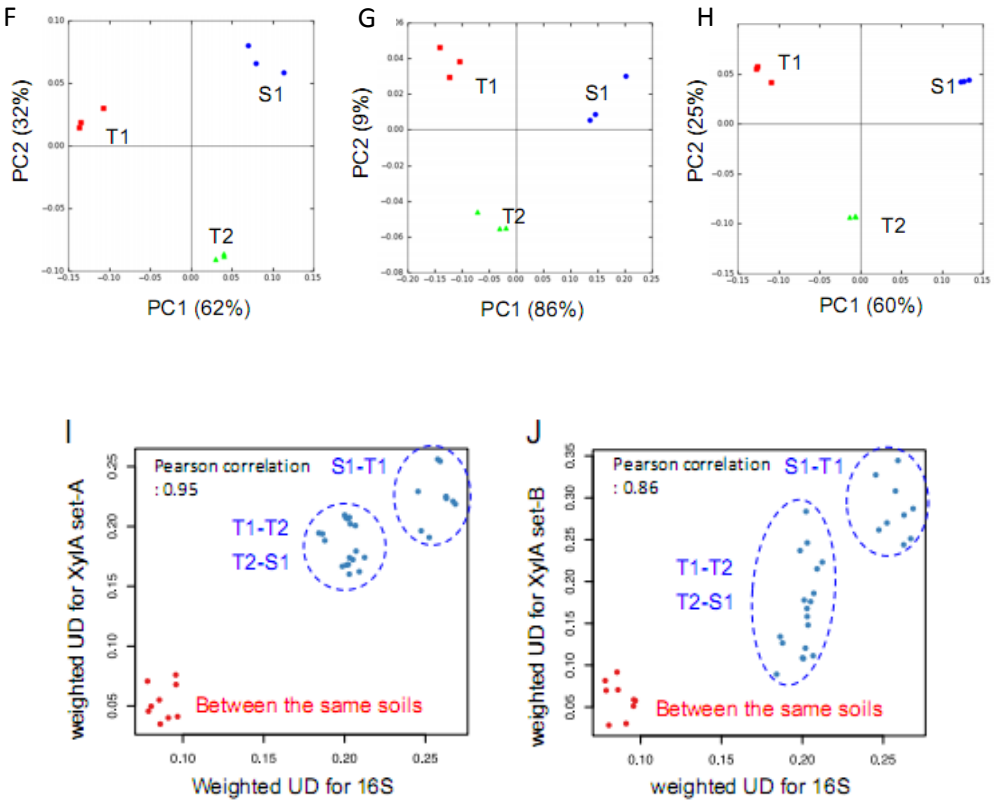


Fig. 2.3.4.3. (F), (G), and (H), PCoA of the member and composition of *xylA*-set A, *xylA*-set B, and16S rRNA gene, respectively, based on weighted UD. (I) and (J), correlation of weighted UDs between16S and *xylA*-set A, and 16S and *xylA*-set B, respectively. In (I) and (J), the UDs between replicate soil samples were shown in red and those between different soil samples were shown in blue.

2.4. Concluding Remarks

In this study, the diversity, composition and distribution of XI genes in three different soils were explored by *xylA*-targeted metagenomics. The construction of two novel *xylA*-sepecific primer sets resulted in the discovery of *xylA* sequences over two known Groups I and II XIs from diverse bacterial phyla. The data demonstrated, by far, a higher diversity of *xylA* sequences than had ever been known in the environment. Each soil had each member and composition of XI genes and 16S rRNA, which might be associated with plant vegetation, geographical sites and characteristics of soil. The overall data in this study greatly extended our knowledge of *xylA* diversity in soil and will be a basis for the exploration of novel *xylA* sequences.

References

Bachar A, Al-Ashhab A, Soares MI, Sklarz MY, Angel R, Ungar ED, Gillor O. 2010. Soil microbial abundance and diversity along precipitation gradient. *Microb Ecol*. 60: 453-461.

Bao Z, Ikunaga Y, Matsushita Y, Morimoto S, Takada-Hoshino Y, Okada H, Oba H, Takemoto S, Niwa S, Ohigashi K, Suzuki C, Nagaoka K, Takenaka M, Urashima Y, Sekiguchi H, Kushida A, Toyota K, Saito M, Tsushima S. 2012. Combined analyses of bacterial, fungal and nematode communities in andosolic agricultural soils in Japan. *Microbes Environ*. 27: 72-79.

Brat D, Boles E, Wiedemann B. 2009. Functional expression of bacterial xylose isomerase in *Saccharomyces* cerevisiae. *Appl Environ Microbiol*. 75: 2304-2311.

Chao A. 1984. Non-parametric estimation of the number of classes in a population. Scandinavian Journal of Statistics. 11: 265-270.

Chanzdon R, Colwell R, Denslow J, Guariguata M. 1998. *Statistical Methods for Estimating Species Richness of Woody Regeneration in Primary and Secondary Rain Forests of Northeastern Costa Rica.,* p 289-309. *In* Dallmeier F. and Comiskey J. (eds.). *Monitoring and Modeling: Conceptual Background and Old World Case Studies Forest Biodiversity Research,* Parthenon Publishing, New York & UK.

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods*. 10: 996-998.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5, 164-166.

Hamady M, Lozupone C, Knight R. 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and Phylochip data. *ISME J*. 4: 17-27.

Hansel CM, Fendorf S, Jardine PM, Francis CA. 2008. Changes in bacterial and archaeal community community structure and functional diversity along geochemically variable soil profile. *Appl Environ Microbiol*. 74: 1620-1633.

Harhangi HR, Akhmanova AS, Emmens R, van der Drift C, de Laat WT, van Dijken JP, Jetten MS, Pronk JT, Op den Camp HJ. 2003. Xylose metabolism in the anaerobic fungus *Piromyces* sp. Strain E2 follows the bacterial pathway. *Arch Microbiol*. 180: 134-141.

Howard EC, Sun S, Reisch CR, del Valle DA, Bürgmann H, Kiene RP, Moran MA. 2011. Changes in dimethylsulfoniopropionate demethylase gene assemblages in response to an induce phytoplankton bloom. *Appl Environ Microbiol*. 77: 524-531.

Hsiao HY, Chiang LC, Chen LF, Tsao GT. 1982. Effect of borate on isomerization and fermentation of high xylulose solution and acid hydrolysate of hemicellulose. *Enzyme Microb Tech.* 4: 25-31.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*. 8:R143. doi: 10.1186/gb-2007-8-7-r143.

Ishii S, Ohno H, Tsuboi M, Otsuka S, Senoo K. 2011. Identification and isolation of active N_2O reducers in rice paddy soil. *ISME J*. 5: 1936-1945.

Jeffries TW, Grigoriev IV, Grimwood J, Laplaza JM, Aerts A, Salamov A, Schmutz J, Lindquist E, Dehal P, Shapiro H, Jin YS, Passoth V, Richardson PM. 2007. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nature Biotechnol*. 25: 319-326.

Johansson MB. 1995. The chemical composition of needle and leaf litter from Scots pine, Norway spruce, and white birch in Scandinavian forest. *Forestry*. 68: 49-62.

Karhumaa K, Hahn-Hägerdal B, Gorwa-Grauslund MF. 2005. Investigation of limiting metabolic steps in the utilization of xylose by recombinant *Saccharomyces cerevisiae* using metabolic engineering. *Yeast*. 22: 359-368.

Karhumaa K, Garcia Sanchez R, Hahn-Hägerdal B, Gorwa-Grauslund MF. 2007. Comparison of the xylose reductase-xylitol dehydrogenase and the xylose isomerase pathway for xylose fermentation by recombinant *Saccharomyces cerevisiae*. *Microb Cell Fact*. 6:5.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequences alignment based on fast Fourier transform. *Nucleic Acids Res*. 30: 3059-3066.

Kielak A, Pijl AS, Van Veen JA, Kowalchuk GA. 2009. Phylogenetic diversity of *Acidobacteria* in a former agricultural soil. *ISME J*. 3: 378-382.

Kim SW, Suda W, Kim S, Oshima K, Fukuda S, Ohno H, Morita H, Hattori M. 2013. Robustness of gut microbiota of healthy adults in response to probiotic intervention revealed by high-throughput pyrosequencing. *DNA Res*. 20: 241-253.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol*. 12, 118-123.

Kuyper M, Hartog MM, Toirkens MJ, Almering MJ, Winkler AA, van Dijken JP, Pronk JT. 2005. Metabolic engineering of a xylose-isomerase-expressing Saccharomyces cerevisiae strain for rapid anaerobic xylose fermentation. FEMS Yeast Res. 5: 399-409.

Lane DJ. 1991. 16S/23S rRNA sequencing, p. 115-175. *In* Stackebrandt, E., and Goodfellow, M. (ed), Nucleic acid techniques in bacterial systematics. John Wiley & Sons Ltd., West Sussex, United Kindom.

Lema KA, Willis BL, Bourne DG. 2013. Amplicon pyrosequencing reveals spatial and temporal consistency in diazotroph assemblages of the *Acroporamillepora. Environ Microbiol*.doi: 10.1111/1462-2920.12366.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. 23: 127-128.

Nagata Y, Natsui S, Endo R, Ohtsubo Y, Ichikawa N, Ankai A, Oguchi A, Fukui S, Fujita N, Tsuda M. 2011. Genomic organization and genomic structural rearrangements of *Sphingobium japonicum* UT26, an archetypal γ-hexachlorocyclohexane-degrading bacterium. *Enzyme Microb Technol*. 49: 499-508.

Nelson KE, Weinel C, Paulsen IT, Dodson RJ, Hilbert H, Martins dos Santos VA, Fouts DE, Gill SR, Pop M, Holmes M, Brinkac L, Beanan M, DeBoy RT, Daugherty S, Kolonay J, Madupu R, Nelson W, White O, Peterson J, Khouri H, Hance I, Chris Lee P, Holtzapple E, Scanlan D, Tran K, Moazzez A, Utterback T, Rizzo M, Lee K, Kosack D, Moestl D, Wedler H, Lauber J, Stjepandic D, Hoheisel J, Straetz M, Heim S, Kiewitz C, Eisen JA, Timmis KN, Düsterhöft A, Tümmler B, Fraser CM. 2002. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol*. 4: 799-808.

Ono K, Hiraide M, Amari M. 2003. Determination of lignin, hollocellulose, and organic solvent extractives in fresh leaf, litterfall, and organic material on forest floor using near-infrared reflectance spectroscopy. *J Forest Res*. 8: 191-198.

Parachin NS, Gorwa-Grauslund MF. 2011. Isolation of xylose isomerases by sequence- and function-based screening from a soil metagenome library. *Biotechnol Biofuels*. 4: 9.

Park JH, Batt CA. 2004. Restoration of a defective *Lactococcuslactis* xylose isomerase. *Appl Environ Microbiol*. 70: 4318-4325.

Quince C, Lanzen A, Curtis TP,Davenport RJ, Hall N, Head IM, Read LF, Sloan WT.2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods*. 6: 639–641.

Rose TM, Henikoff JG, and Henikoff S. 2003. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res*. 31: 3763-3766.

Shannon CE, Weaver W. 1984. The Mathematical Theory of Communication, University of Illinois Press, Urbana, IL.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 7: 539.

Simberloff D. 1978. Use of rarefaction and related methods in Biological Data, p.150-165. in Dickson, K. L., Cairns, J., Jr., and Livingstone, R. J. (ed.), Water Pollution

Assessment Quantitative and Statistical Analyses, American Society for Testing and Materials, Philadelphia.

van Maris AJ, Winkler AA, Kuyper M, de Laat WT, van Dijken JP, Pronk JT. 2007. Development of efficient xylose fermentation in *Saccharomyces cerevisiae*: xylose isomerase as a key component. *Adv Biochem Eng Biotechnol*. 108, 179–204.

Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*. 73: 5261-5267.

Wang Q, Quensen JF 3rd, Fish JA, Lee TK, Sun Y, Tiedje JM, Cole JR. 2013. Ecological patterns of nifH genes in four terrestrial climatic zones explored with targeted metagenomics using FrameBot, a new informatics tool. *Mbio*. 4, e00592-13.

Will C, Thürmer A, Wollherr A, Nacke H, Herold N, Schrumpf M, Gutknecht J, Wubet T, Buscot F, Daniel R. 2010. Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. *Appl Environ Microbiol*. 76: 6751-6759.

Wilson MC, Mori T, Rückert C, Uria AR, Helf MJ, Takada K, Gernert C, Steffens UA, Heycke N, Schmitt S, Rinke C, Helfrich EJ, Brachmann AO, Gurgui C, Wakimoto T, Kracht M, Crüsemann M, Hentschel U, Abe I, Matsunaga S, Kalinowski J, Takeyama H, Piel J. 2014. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature*. 506: 58-62.

Woodhouse JN, Fan L, Brown MV, Thomas T, Neilan BA. 2013. Deep sequencing of non-ribosomal peptide synthetases and polyketide synthases from the microbiomes of Australian marine sponges. *ISME J*. 7: 1842-1851.

Zhang Y, Sun Y. 2011. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. *BMC Bioinformatics.* 12: 198. doi:10.1186/1471-2105-12-198.

Zheng H, Bodington D, Zhang C, Miyanaga K, Tanji Y, Hongoh Y, Xing XH. 2013. Comprehensive phylogenetic diversity of [Fe-Fe]-hydrogenase genes in termite gut microbiota. *Microbes Environ*. 28, 491-494.

Chapter 3

Isolation of full length *xylA* genes from soil metagenomes and their functional expression in *S. cerevisiae*

Abstract

Metagenomics has appeared as one useful approach to obtain novel useful genes from microorganisms. Amplifying flanking sequence of internal gene sequences are required to determine full length of the genes and their functional activity. This study attempted to retrieve full length *xylA* genes from soil metagenome. Inverse PCR primer sets were designed from selected partial *xylA* sequences, which were chosen based on the phylogenetic clades of partial *xylA* sequences amplified from soil metagenomes. Pre-amplified inverse PCR (PAI-PCR) method was applied to amplify the flanking sequence of target *xylA* sequences from soil metagenome. Six target *xylA* sequences were successfully amplified by using PAI-PCR method. Four putative full length *xylA* genes, which have identity 73, 81, 62, and 67% to the XI gene of *Candidatus Koribacter, bacterium Ellin, Opitutus terrae,* and *Mesorhizobium* sp., respectively, were successfully isolated and cloned. These full length *xylA* genes were subsequently screened for their functional expression in *S. cerevisiae* by cell surface display. Four putative full length *xylA* genes designated as BE, CK, OT, and MS were constructed in yeast expression vector for cell surface display, pULD1. Three putative full length *xylA* genes, CK, OT, and MS were successfully displayed on *S. cerevisiae* cell surface.

3.1. Introduction

Soil is a rich source of microbial population and it is recognized as the source of the highest diversity of microbes compared to other environments (Roesch et al., 2007). Each gram of soil comprises billions of microorganisms with thousands of identified distinct species (Delmont et al., 2010). Due to the difficulty of cultivating microbes from their natural environment, it was surveyed that less than 1% of soil bacteria can be cultivated under standard cultivation techniques (Kellenberger, 2001). Metagenome approaches involving the extraction of total DNA from soil may provide access to novel genetic sources of uncultivated bacteria from soil.

Two strategies that are usually used in metagenome screening are activity-based and sequence-based screening. Both activity- and sequence-based screening have been applied successfully to discover catalytic enzymes from metagenome such as chitinase (Cottrell et al., 1999), lipase (Henne et al 2000), xylanase (Yamada et al., 2008; Liu et al., 2010), esterase (Park et al., 2011), fumarase (Jiang et al., 2010), etc. In this study, the sequence-based screening was chosen to obtain full-length xylose isomerase gene sequences from soil metagenomes. Dissimilar to the activity-based screening, this sequence-based screening does not require the heterologous expression of the genes encoded within metagenome clones in the particular host screening system. Generally, it is based on the consensus oligonucleotide primers obtained from alignment of known gene sequences (Yun and Ryu, 2005). Here, homology sequence-based screening was used to recover the entire genes from three soil metagenomes by employing PCR-based strategy. There are some difficulties to isolate full length genes from metagenomes by using inverse polymerase chain reaction (IPCR) technique and most probably because the copy number of target DNA sequences had been quite low.

Pre-amplified inverse PCR method (PAI-PCR) is an inverse PCR method which was developed to amplify desired nucleotide sequences from environmental DNA. This method involves enrichment of the target sequences by rolling circle amplification (RCA). A primer containing locked nucleic acids (LNAs) that anneal to the specific site of the target was used to increase inverse PCR sensitivity higher than usual inverse PCR, as reported earlier by Yamada et al. (2009). In this study, the amplification of *xylA* flanking sequences was conducted by PAI-PCR method from soil metagenomes in order to obtain full length of the genes and four putative full length *xylA* genes were successfully isolated and cloned from soil metagenome samples.

The heterologous expression of bacterial XI genes which catalyze direct convertion of xylose into xylulose in *S. cerevisiae* was assumed to be more prospective than XR/XDH pathway, as it can avoid different cofactor preferences. However, for many years, several attempts having been done to introduce bacterial *xylA* in *S. cerevisiae* have not resulted detectable XI enzyme expression (Hahn Hagerdal et al., 2007). The first functionally expressed XI in *S. cerevisiae* originated from *Thermus thermophilus* (Walfridson et al., 1996). However, the activity of this enzyme in *S. cerevisiae* was lower than that of the XI gene isolated from fungus *Piromyces* sp. As observed in the highly active XI from *Piromyces*, the expression of this enzyme alone in *S. cerevisiae* only allowed very slow growth on xylose (Kuyper et al., 2003), which may also correspond to the failure of early trials for heterologous XI expression, where it was only assayed as growth on xylose (Hahn Hagerdal, 2007). Similarly, the putative full length *xylA* genes from soil metagenome were screened for their intracellular heterologous expression in *S. cerevisiae,* however *S. cerevisiae* recombinant strains failed to grow on xylose. This occurred possibly because the

activity of the enzyme was too low to be detected as growth on xylose. In order to confirm that the putative full length *xylA* genes are expressing the XI protein, some attempts to observe the *xylA* expression were done through the cell surface engineering.

The putative full length *xylA* genes isolated from soil metagenomes besides known active and inactive *xylA* in *S. cerevisiae* were introduced into a yeast expression vector that enables the heterologous proteins to be displayed on the *S. cerevisiae* cells surface. In this method, the displayed protein can be directly observed for its activity on the cell state without purification process (Kuroda et al., 2009). A successful attempt to display XI protein on the *S. cerevisiae* cell surface has been reported for the expression of XI protein of the bacterium *Clostridium cellulovorans* (Ota et al., 2013). In this study, the full length *xylA* gene retrieved from soil metagenome was fused with 3'half of α-agglutinin gene as the anchoring motif for the XI protein to be displayed on *S. cerevisiae* cell surface.

3.2. Materials and Methods

3.2.1. Soil samples and DNA extraction

Three soil samples named T1, T2, and S1 were collected from three sites, which have different plant vegetation. The T1 soil contained biomass of *Criptomeria Japonica* (needle-leaf tree), T2 and S1 soils contained biomass of *Fagus crenata* and *Betula platyphylla*, both of which are birch-leaves. Each 10 g of soil homogenates was used for the metagenome DNA extraction using ISOIL DNA extraction Kit (Nippongene, Tokyo Japan).

3.2.2. Oligonucleotide primers

Two degenerate primer sets (set A: xyl1/xyl2, set B: xyl30F/xyl30.4R; Table 3.2.2.1) were designed based on specified conserved regions of 112 known amino acid XI sequences collected from the NCBI database. Primer A (xyl1/xyl2) was designed manually against the conserved amino acid region WGGREGY and GWDTDEF, respectively. While Primer set B (xyl30F/xyl30.4R) was designed by using the CODEHOP program (Rose et al., 2003) against the conserved amino acid region VFWGGREG and HEQMAGHN, respectively.

Locked nucleic acid-containing RCA primer (xylAmeta; Table 3.2.2.1) was designed manually based on highly conserved region of known amino acid *xylA* sequences alignment. Inverse PCR primers (AS1.21Fi/AS1.21Ri, BS1.62Fi/BS1.62Ri, AT2.90Fi/AT2.90Ri, Ot-T1Fi/Ot-T1Ri, Bj-T2Fi/Bj-T2Ri, AT1.55Fi/AT1.55Ri; Table 3.2.2.1) were designed by using Primo inverse 3.4 (Chang Bioscience). These primers facilitated the specific amplification of the unknown part of the targeted *xylA*

fragments. Those primers were designed based on identified partial *xylA* fragment, amplified by degenerate primers. Oligonucleotides primers were supplied by Sigma-Aldrich Japan (Tokyo, Japan) and Eurofins MWG Operon (Tokyo, Japan) and LNA oligonucleotides by Gene Design Inc. (Ibaraki, Japan).

Table 3.2.2.1. Primers used for retrieving full length *xylA* genes from soil metagenomes

Primer	Sequence (5'-3')[a]
xyl1	TGGGGNGGNCGNGARGGNA
xyl2	RAAYTSRTCNGTRTCCCARCC
xyl30F	TGTGTTTTGGGGCGGNMKNGANGG
xyl30.4R	GTTATGGCCCGCCADNKKNKCRTG
xylAmeta	G**A**RC**C**N**A**ARCC
AS1.21Fi	CCGGGCACACCATGCATCATGAATGTG
AS1.21Ri	TCGACAGCCATATGCAGGAATTTTGCGAGG
BS1.62Fi	CACGTGCAGACATCTACATGGCCACTACGG
BS1.62Ri	TTCATGCCATACCCGCGTTCGATGTTG
AT2.90Fi	CCATTCGTTCGAGCATGAGATCGCGCTG
AT2.90Ri	AGGATCACGCCCTTGTAGCCGATCTTG
AT1.55Fi	AGGTAGCTGTCGATGCAGGCATGCTG
AT1.55Ri	AAATCCCTGTTTGCGTGCATAGTCCCTG
Ot-T1Fi	CTCAACGTCGAAGCCAACCATGCCAA
Ot-T1Ri	GTTCCTTCGGCTTGGGCTCGATCAGC
Bj-T2Fi	CTCAACATCGAGCAGAACCACGCCATC
Bj-T2Ri	GCCCCTTGAACCCAATCTTGTGCTTGTG

[a]LNA residues are shown in boldface.

3.2.3. Amplification of internal *xylA* gene sequences from soil metagenomes and cloning

Metagenome DNA extracted from soil samples were used as the template for amplification of internal *xylA* gene sequences by using two degenerate primer sets. Based on amino acid, similarity, bacterial XIs are classified into two groups, which differ by approximately an additional 50 amino acid residues at the N terminus of group II XIs (Park and Batt, 2004). To amplify the internal fragment of group II XIs, PCR amplification with the primer set A was performed in a 25 µL mixture (total volume) containing 10-50 ng soil DNA, 2.5 µL 10-fold reaction Ex Taq buffer, 0.2 mM dNTP, 4 µM of each primer, and 1.25 U TaKaRa Ex Taq^{TM} HS polymerase (TaKaRa-Bio Inc., Ohtsu, Japan) with the following PCR condition: 2 min of initial denaturation at 94^{o}C, 25 cycles of 30 s denaturation at 94^{o}C, 30 s annealing at 68^{o}C, and 30 s extension at 72^{o}C, followed by a 7 min final extension at 72^{o}C. Whereas for amplifying the internal fragment of group I XIs, a touchdown PCR was performed by using primer set B. Each 25 µl PCR mixture contained 10-50 ng soil DNA, 2.5 µl 10-fold AccuPrime^{TM} PCR Buffer II, 2 µM of each primer, and 0.5 µl AccuPrime^{TM} DNA Polymerase (Invitrogen, Carlsbad, CA, USA). Touchdown PCR consisted of one cycle of pre-denaturation at 94^{o}C for 2 min, followed by the first 10 cycles with 5 cycles each of denaturation at 94^{o}C for 30 s, varrying annealing for 30 s (from 58^{o}C decreased to 57^{o}C), and extension at 68^{o}C . A further 15 cycles constituted the following: 94^{o}C for 30 s, 56^{o}C for 30 s, and 68^{o}C for 30 s. PCR was terminated after a final cycle 68^{o}C for 4 min. PCR products were then analyzed by agarose gel electrophoresis and purified using a Gel extraction Kit (Qiagen GmbH, Hilden, Germany). The purified PCR ligated to the pGEM-T Easy Vector System (Promega, Madison, USA) transformed into ECOS^{TM}-competent *E. coli* JM109 cells

(Nippongene). Transformants were selected on LB agar plate supplemented with 50 µg/ml ampicillin. Positive colonies harboring inserts were picked up and grown in 100 µl of LB medium supplemented with ampicillin in the each well of 96-well plate (Qiagen, Hilden, Germany) at 37°C for 8 h. DNA sequencing of 596 clones was performed by TaKara, Japan.

3.2.4. Clone library sequence data analysis

Clone library sequences of internal *xylA* gene sequences amplified by both primer sets A and B were analyzed by NCBI BLASTx to remove non-hit sequences against *xylA* genes uploaded onto the database and low quality sequence harboring stop codons or unknown amino acid residues. The identified internal *xylA* gene sequences were aligned and trimmed equally to the amino acid sequences from primer A (WGGREGY) to the reverse primer of primer B (HEQMAGHN) to construct a phylogenetic tree. The internal *xylA* gene sequences representing each phylogenetic clade were chosen to retrieve the full length of the genes and IPCR primer sets were designed based on those selected sequences.

3.2.5. Amplification of full length *xylA* genes and cloning

The flanking sequences of *xylA* genes were amplified by pre-amplified inverse polymerase chain reaction (PAI-PCR), which was previously described by Yamada et al. (2008) as follows. Approximately 3 µg of DNA isolated from the soil metagenome was digested with 10 U of a *BamH*I, *Bgl*II, *EcoR*I, *Nco*I, *Xba*I or *Xho*I restriction enzyme (TaKaRa-Bio) at 37°C overnight, then self-circularized using a Mighty Mix DNA ligation kit (TaKaRa-Bio) at 16°C overnight in 20 µl of the ligation buffer. RCA first mixture was prepared by mixing 1 µl of the self-circularzed DNA solution

and the 3.2 µM LNA oligonucleotide (xylAmeta) (Table 3.2.2.1) with Phi29 DNA polymerase buffer (New England Biolabs, Beverly, MA) to a final volume of 10 µl. The first mixture was pre-treated by heat at 95°C for 3 min. The denatured template in the first mixture immediately cooled on ice. Subsequently, second mixture of 10 µl of phi29 DNA polymerase reaction buffer containing 0.2 mM each of dNTP, 100 µg/ml BSA and 5 U of Phi29 DNA polymerase (New England Biolabs) was added to the first RCA mixture, and the RCA reaction was incubated at 30°C overnight. The RCA product of 1 µl was then mixed with 24 µl of GC Buffer I containing 0.2 mM dNTP, 0.4 mM each of IPCR primer pairs (AS1.21Fi/AS1.21Ri, BS1.62Fi/BS1.62Ri, AT2.90Fi/AT2.90Ri, Ot-T1Fi/Ot-T1Ri, Bj-T2Fi/Bj-T2Ri, AT1.55Fi/AT1.55Ri; Table 3.2.2.1) and 2.5 U of LA Taq DNA polymerase (TaKaRa-Bio). The inverse PCR was set as follows: denaturation at 94°C for 1 min, the next 30 cycles of denaturation at 94°C for 30 s continued by annealing and elongation at 68°C for 5 min, and the final extension at 68°C for 7 min. PCR amplicon was visualized by agarose gel electrophoresis and purified using a Gel Extraction Kit (Qiagen GmbH, Hilden, Germany). Purified IPCR products were cloned and positive colonies harboring inserts were analyzed by sequencing. Sequencing analysis of longer inserts was done by primer walking. Identification of full open reading frame (ORF) was performed by using the ORF finder (NCBI). The analysis of putative ORFs was done by BLASTP searches of NCBI database. The full length *xylA* genes were confirmed by amplifying the genes from soil metagenome using newly designed specific primer sets based on putative ORFs. The amplicons were then cloned and sequenced.

3.2.6. Phylogenetic analysis

Known XIs were collected from NCBI database and aligned with the XI sequences isolated from soil metagenome. MEGA software version 4.1 (Tamura et al., 2007) was employed to build the phylogenetic tree using the Neighbor-Joining method and bootstrap resampling analysis for 1,000 replicates.

3.2.7. Strains, plasmid and media for cell surface display

Constucted *xylA* genes from soil metagenome in pULD1 plasmid were cloned in *Escherichia coli* JM109. Mutated *S. cerevisiae* BY4741 (*Δsed* or Δ*kex*) strain (Kuroda et al., 2009) was used for displaying XI protein on its cell surface. *E. coli* JM109 harboring recombinant vector was grown on Luria-Bertani medium (1% tryptone, 0.5% yeast extract, 1% sodium chloride) supplemented with 100 μg/mL ampicillin. Yeast transformants were aerobically cultivated at 30^{o}C in synthetic dextrose (SD) medium [Yeast nitrogen base w/o amino acid 6.7 g/L (Difco Laboratories), Glucose 20 g/L, 10 × DO supplement (-URA, -trp) 0.72 g/L, Tryptophan 0.2 g/L (Clontech Laboratories Inc.)].

3.2.8. Construction of plasmid for xylose isomerase display

Six plasmids were constructed for the expression of four putative full length *xylA* genes isolated from soil metagenome and two known full length *xylA* genes from database. All primers used for plasmid constuction were listed in table 3.2.8.1. PCR reaction was carried out using PrimeSTAR® MAX DNA polymerase (TaKaRa, Japan). Three plasmids pULD1-Be, pULD1-Pi and pULD1-Ec for the cell surface display of *Bacterium Ellin*-like, *Piromyces* sp., and *E. coli xylA* genes were constructed as follow. The 1035, 1314, and 1323 bp of *Bgl*II-*Xho*I fragment encoding *xylA* genes of

B. Ellin-like, *Piromyces* sp., and *E. coli*, were respectively prepared by PCR performed with primers BeF/BeR, PiF/PiR, and EcF/EcR. Each PCR product was digested with *Bgl*II and *Xho*I and introduced into *Bgl*II-*Xho*I site of the cell surface expression plasmid pULD1 (Kuroda et al., 2009). pULD1 plasmid harbor glucoamylase gene from *Rhizopus oryzae* encoding secretion signal sequence and the 3'-half region of the α-agglutinin gene.

While two plasmids pULD1-Ck and pULD1-Ms for the cell surface display of *Candidatus Koribacter*- and *Mesorhizobium* sp.-like *xylA* genes were constructed as follow. The 1230 and 1311 bp of *Bgl*II-*Nhe*I fragment encoding *xylA* genes of *C. Koribacter*-like and *Mesorizobium* sp.-like, were respectively prepared by PCR performed with primers CkF/CkR and MsF/MsR. Each PCR product was digested with *Bgl*II and *NheI* and introduced into *Bgl*II-*NheI* site of the cell surface expression plasmid pULD1.

One other plasmid pULD1-Ot for the cell surface display of *Opitutus terrae*-like *xylA* was constructed as follow. The 1332 bp of *NotI*-*Nhe*I fragment encoding *xylA* of *O. terrae*-like was prepared by PCR performed with primers OtF/OtR. PCR product was digested with *NotI* and *NheI* and introduced into *NotI*-*NheI* site of the cell surface expression plasmid pULD1.

Table 3.2.8.1. Strains, plasmid and primers used for full length *xylA* genes expression in *S. cerevisiae* by cell surface display

Strain, plasmid, or primer	Relevant features or sequence[a]	Reference or source
S. cerevisiae		
BY4741 Δ*kex*	*MATa, his3, leu2, ura3, met15, kex2::kanMX4*	Kuroda et
BY4741 Δ*sed*	*MATa, his3, leu2, ura3, met15, sed1::kanMX4*	al., 2009
BY4741-1.1	BY4741 Δ*kex* (pULD1-Be)	This study
BY4741-2.1	BY4741 Δ*sed* (pULD1-Ck)	This study
BY4741-1.2	BY4741 Δ*kex* (pULD1-Ot)	This study
BY4741-1.3	BY4741 Δ*kex* (pULD1-Ms)	This study
BY4741-1.4	BY4741 Δ*kex* (pULD1-Ec)	This study
BY4741-2.2	BY4741 Δ*sed* (pULD1-Pi)	This study
E. coli		
JM109	F'[*tra*D36, *pro*AB, *lac*lq, *lacZ*ΔM15], Δ (*lac-pro*AB), *hsd*R17 ($rk^{-}mk^{+}$), *rec*A1, *end*A1, *rel*A1, *sup*E44, *thi*-1, *gyr*A96, $e14^{-}$($mcrA^{-}$)	Nippon gene
Plasmid		
pULD1	Cassette vector for cell surface display of proteins with FLAG-tag, *URA3*, and *leu2-d*	Kuroda et al 2009
Primers		
BeF	5'-AAAAAAAGATCTATGGCCGAACATCTGCGCTT-3' (*Bgl*II)	This study
BeR	5'-AAAACTCGAGTTCGGGCAGGCGGCCAT-3' (*Xho*I)	This study
CkF	5'-ACTGCCAGATCTATGAAGTCGATTCTAAAATTGATG-3' (*Bgl*II)	This study
CkR	5'-AAAAGCTAGCCCGTACACCTAGCAGAACGTC-3' (*Nhe*I)	This study
OtF	5'-AAAAGCGGCCGCATGCCGTACGTACTTTCGGAACATA-3' (*Not*I)	This study
OtR	5'-AAAAAGCTAGCCCATTTCCGGCTAAACAGGTATTG-3' (*Nhe*I)	This study
MsF	5'-ACTGCCAGATCTATGAGCAAGCCCTTTTTC-3' (*Bgl*II)	This study
MsR	5'-AAAAGCTAGCCAGATAGCGGTTAAGCAGGT-3' (*Nhe*I)	This study
EcF	5'-ACTGCCAGATCTATGCAAGCCTATTTTGACCA-3' (*Bgl*II)	This study
EcR	5'-ACCGCTCGAGTTTGTCGAACAGATAATGGT-3' (*Xho*I)	This study
PiF	5'-AAAAAAGATCTATGGCTAAGGAATACTTCCC-3' (*Bgl*II)	This study
PiR	5'-AAAAACTCGAGTTGGTACATAGCAACAATTG-3' (*Xho*I)	This study

[a]Underlines indicate the restriction sites for the enzymes shown in parentheses

3.2.9. Yeast transformation

Yeast transformation was carried out by using Frozen-EZ Yeast Transformation II Kit^{TM} (Zymo Research, California, USA) according to the manufacturer's instructions. The plasmids pULD1-Be, pULD1-Ck, pULD1-Ot, pULD1-Ms, pULD1-Ec, pULD1-Pi were used to transform *S. cerevisiae* BY4741 Δ*kex* or Δ*sed* . The resulting yeast strains are summarized in Table 3.2.8.1.

3.2.10. Immunofluorescence labeling of cells

Overnight culture of yeast cells were washed with phosphate-buffered saline (PBS) buffer (pH 7.2). The immunofluorecence labeling of the yeast recombinant strains were performed as previously described by Kuroda et al. (2009).

3.2.11. Immunofluoresence microscopy and flow cytometry analysis

Immunofluorescence labeled yeast cells were observed with fluorescence microscope (BX51, Olympus, Tokyo, Japan). The detection of green fluorescence of Alexa fluor 488 was done through NIBA filter. Images were captured by DP73 camera system (Olympus) regulated by Cell Sens Standard Program (Olympus).

Total fluorescing cells of immunofluorecence labeled yeast cells were analyzed using a flow cytometer BD FACSAria II cell sorting system (BD Bioscience). The fluorescence intensity of the cells was detected with an excitation at 488 nm blue laser. Total fluorescing cells were estimated from the number of fluorescing cells against the total cells observed.

3.3. Results and Discussion

3.3.1. Amplification of internal *xylA* gene sequences from soil metagenome and cloning

Partial *xylA* sequences of both group I and II XIs were obtained by PCR amplification using degenerate primer sets. The amplification with both degenerate primer sets A and B yielded fragment sizes as expected, approximately 384 and 291 bp, respectively (data not shown) and was further confirmed by clone library sequence analysis. After sequencing 596 clone libraries, a BLASTX search was performed against the presently GenBank database of NCBI. All sequences which have identity to the reported *xylA* gene sequences were used to build a phylogenetic tree against representative known *xylA* gene sequences which belong to group I and II XIs. Based on representative phylogenetic clades, 11 sequences with identity ranging from 61-92% to the known XI gene sequences were selected to retrieve full length of the *xylA* genes from soil metagenomes by using PAI-PCR method (Table 3.3.1.1). Based on these selected sequences, IPCR primer sets were designed for the amplification of the flanking sequences.

3.3.2. Amplification of full length *xylA* genes and cloning

In order to isolate bacterial *xylA* genes from soil metagenome, sequence-based screening was used by applying PCR strategy based on the conserve amino acid sequences of known *xylA* genes. However, due to the difficulty in isolating full length genes from metagenome by standard IPCR, pre-amplification of DNA target was applied using RCA with LNA-containing oligonucleotide as the primer to enrich target sequences and to increase the sensitivity of IPCR, as reported previously by

Yamada et al. (2008). Degenerate LNA-containing primer with the LNAs evenly distributed throughout its sequence was used to increase the sensitivity of RCA primer against a very low copy number of the target genes. Figure 3.3.2.1 shows the PAI-PCR results of target *xylA* genes, which mostly obtained one specific band except for the amplification of *xylA* target of *O. terrae*-like, which resulted two bands. Unlike Yamada et al 2008, who used site specific LNA-containing primer, the degeneracy of LNA-containing primer used in this study may have influence to the non-specific amplification of PAI-PCR result.

Table 3.3.1.1. Internal *xylA* sequences selected for retrieving full length of the genes

Clone library	Homology[a]		
	Species	Phylum	Identity (%)
BS1.62	*Candidatus koribacter*	*Acidobacteria*	73
AS1.21	*Chthoniobacter flavus*	*Verrucomicrobia*	78
AT2.90	*Mesorhizobium oportunistum*	*Proteobacteria*	84
AT1.55	*Chitinophaga pinensis*	*Bacteroidetes*	90
BT2.2	*Pedobacter heparinus*	*Bacteroidetes*	90
AS1.25	*Gemmata obscuriglobus*	*Planctomycetes*	81
AT2.93	*bacterium Ellin*	*Verrucomicrobia*	83
BT1.11	*Herpetosiphon aurantiacus*	*Chloroflexi*	61
BT1.52	*Candidatus solibacter*	*Acidobacteria*	77
Ot-T1	*Opitutus terrae*	*Verrucomicrobia*	74
Bj-T2	*Bradyrhizobium japonicum*	*Proteobacteria*	92

[a]Homology search was conducted by BLASTX NCBI

In this study, six target sequences were successfully amplified by PAI-PCR and four putative full length *xylA* genes retrieved from soil metagenome. They were designated as BE, CK, OT and MS. Full length *xylA* genes retrieved identification was summarized in table. 3.3.2.1. All putative XI sequences exhibited identities from 20-51% to the XI of *Piromyces* sp. Phylogenetic tree analysis (Figure 3.3.2.2) shows that three putative full length *xylA* genes, namely BE, OT, and MS are classified in group II XIs and one other putative *xylA* full length gene, CK was classified in group I XIs.

Interestingly, BE XI, which is classified in group II XIs, has a shorter amino acid sequence length (344 aa) compared to the average length of group II members, which range from 440-460 amino acids (Park & Batt 2004; Epting et al., 2005) and also shorter than the average length of group I XIs which range from 380-390 amino acids (Park & Batt 2004; Epting et al., 2005). Conversely, CK XI, which is classified in group I XIs, has a longer amino acid sequence length (409 aa) than the average length of group I XIs, but is still below the average length of group II XIs. These findings can be revealed that *xylA* genes isolated from metagenome possibly have more variety of sizes than that from known ones due to the diversity and complexity of soil metagenome samples. As described in the previous chapter, these metagenome samples contained high bacterial and *xylA* gene diversity.

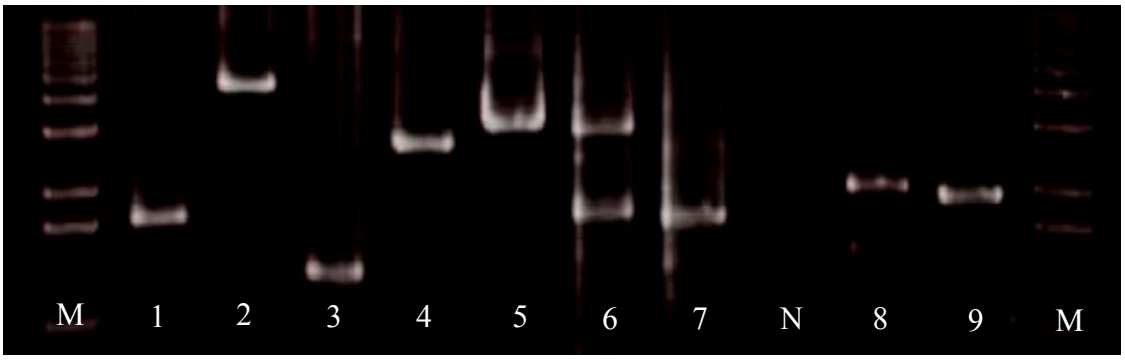


Fig. 3.3.2.1. Putative *xylA* fragments amplified by employing pre-amplified Inverse PCR method (PAI-PCR). The gel electrophoresis of PCR products corresponding to each targeted *xylA* is shown by numerical order, while restriction enzymes used and the size of PCR products (kb) are indicated in parenthesis, respectively. Lane1, *Chthoniobacter flavus* (*Bgl*II, 1.8); Lane 2-4, *Candidatus Koribacter* (*Bam*HI, 5; *Nco*I, 1.3; *Eco*RI, 3); Lane 5 *Mesorhizobium oportunistum* (*Xba*I, 3); Lane 6 and 7, *Opitutus terrae* (*Bam*HI , 1.8; *Xho*I, 1.8) ; Lane 8, *Bradyrhizobium japonicum* (*Bam*HI, 2); *Chitinophaga pinensis* (B*gl*II, 2); N, non-template control; M, marker 1 kb plus DNA Ladder.

Table 3.3.2.1 Sequence analysis of full length *xylA* genes retrieved from soil metagenome

Putative *xylA* genes	Size[a]		Homology[b]			
	Length (bp)	aa	Organisms	Acc. No.	aa	Identity (%)
CK	1230	409	*Candidatus Koribacter versatilis Ellin345*	YP_589980.1	385	73
BE	1035	344	bacterium Ellin514	ZP_03627970.1	437	81
OT	1332	443	*Opitutus terrae* PB90-1	YP_001818641.1	444	62
MS	1311	436	*Mesorhizobium* sp. BNC1	YP_675361.1	436	67

[a] Estimated by ORF Finder analysis of NCBI

[b]Homolog protein showing the highest similarity to putative *xylA* identified by BLASTP of NCBI

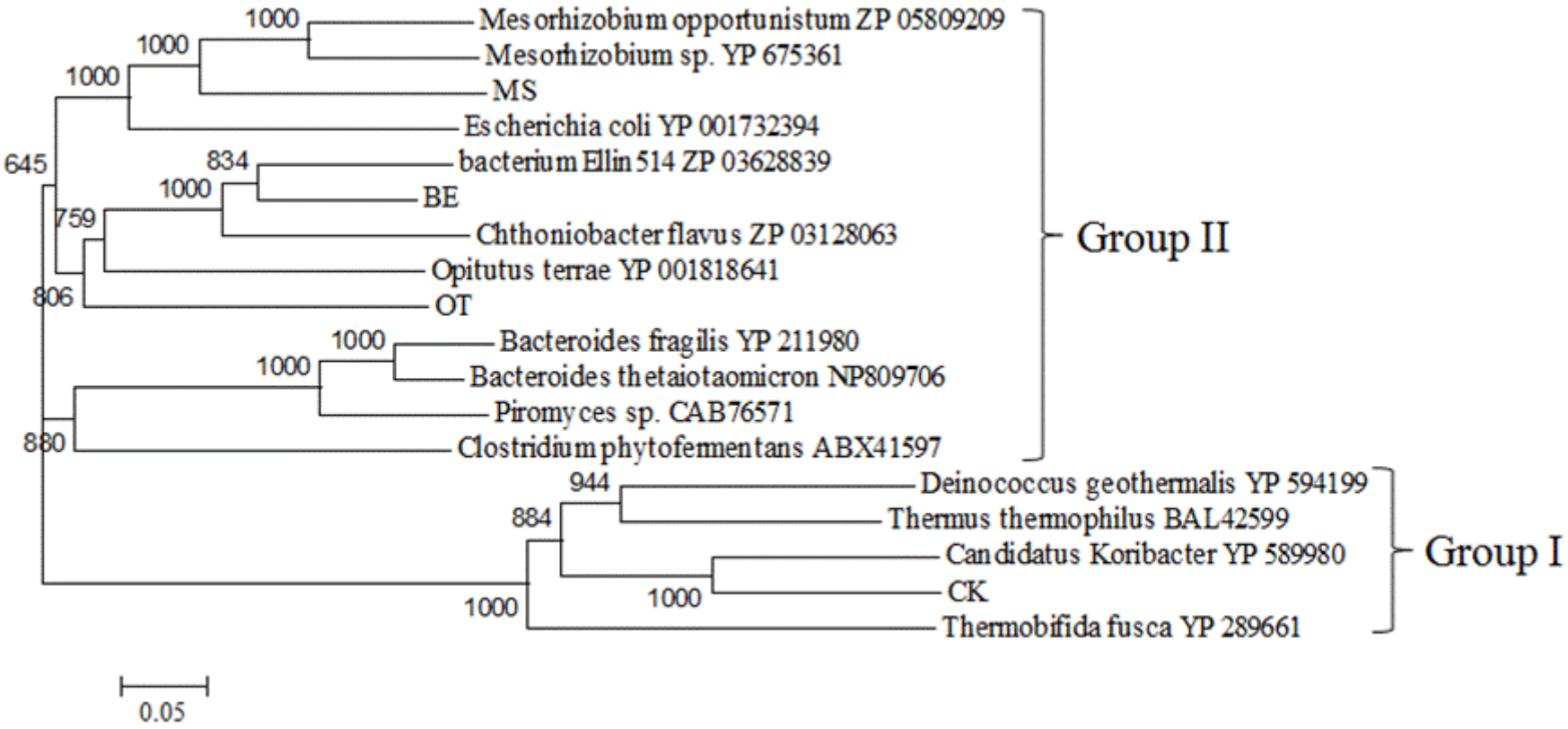


Fig. 3.3.2.2. Phylogenetic analysis of putative full length *xylA* genes from soil metagenome against known XI protein from database. Phylogenetic tree analysis based on translated amino acid sequences of full length *xylA* genes. The bar represents 5% divergence. Numbers at the nodes represent the bootstrap values (1000 resamplings). Reference sequences are given in species name followed by accession number.

3.3.3. Construction of plasmids for cell surface display of full length *xylA* genes

The constructed recombinant plasmids pULD1-Be, pULD1-Ck, pULD1-Ot, pULD1-Ms, pULD1-Ec and pULD1-Pi were successfully established for cell surface display of *xylA* genes. The genes to be displayed on the cell surface were fused with α-agglutinin and FLAG tag to confirm that they were successfully displayed. The constructed recombinant plasmids which contained the fusion gene consist of the glyceraldehydes-3-phosphate dehydrogenase promoter, glucoamylase signal sequence, full-length putative *xylA* genes from soil metagenome or known *xylA* genes from database belong to *E. coli* or *Piromyces* sp., FLAG-tag, linker, and 3′-half of α-agglutinin as the cell-wall anchoring domain. The BY4741 wild-type strain without plasmid introduced was used as negative control.

3.3.4. Cell surface display of full length *xylA* genes

XI protein that was successfully displayed on the *S. cerevisiae* cell surface will be able to be observed by fluoresence microscope. A FLAG tag was constructed to fuse with the XI protein or the C terminal domain of XI was labeled by immunofluorescence. Anti-FLAG antibody was used as the primary antibody whereas Alexa Flour 488-conjugated goat anti-mouse IgG antibody was used as the secondary antibody. Cells harboring pULD1-Ck, pULD1-Ot, pULD1-Ms, pULD1-Ec or pULD1-Pi for display of XIs fused with FLAG-tag showed green fluorescence on their surface (Figure 3.3.4.1.D-H). In contrast, cells harboring pULD1-Be and negative control showed no fluorescence (figure 3.3.4.1.A-C). The results of immunofluorescence labeling indicated that three metagenomic full length *xylA* genes were displayed correctly on the yeast cell surface.

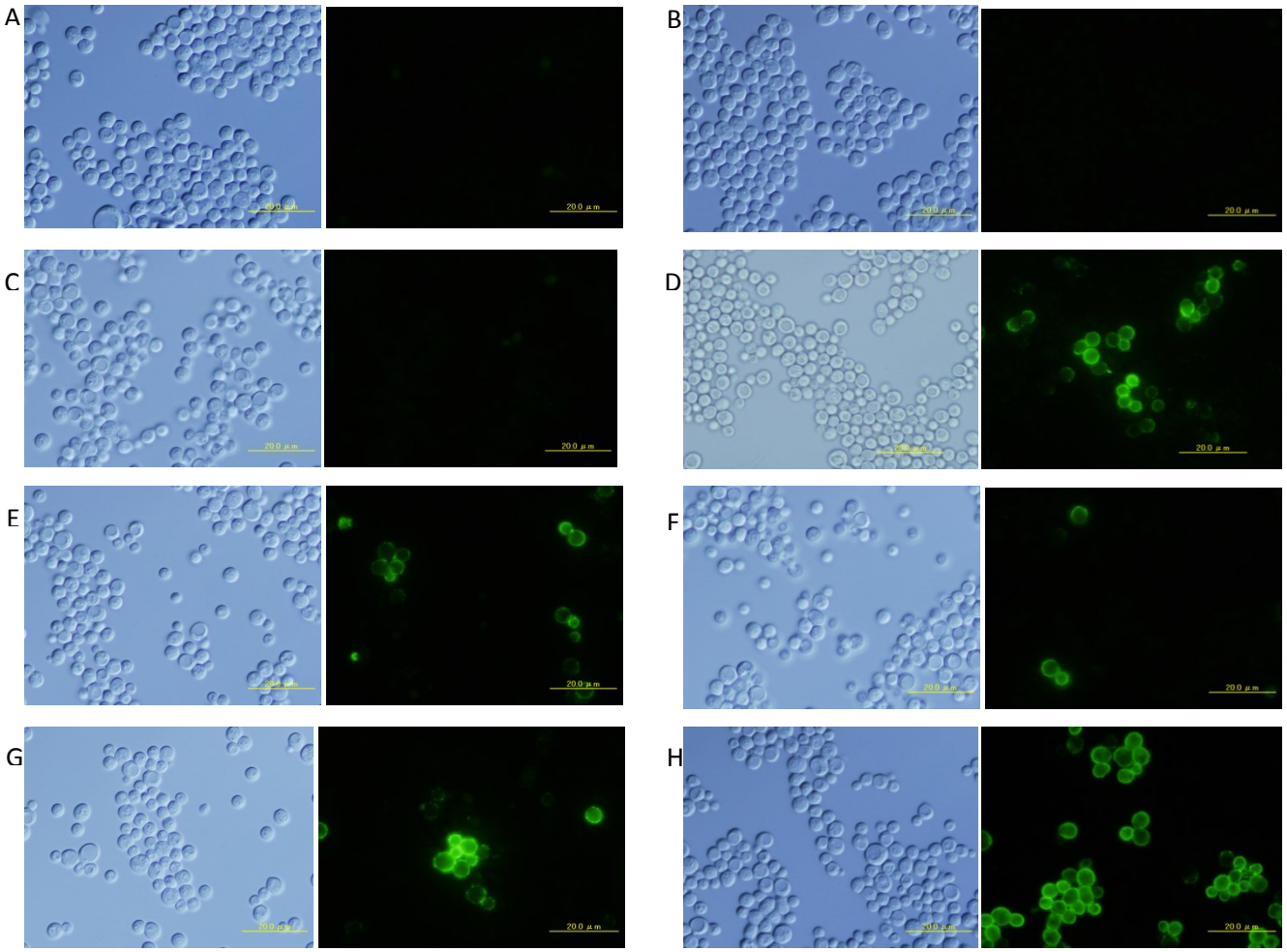


Fig. 3.3.4.1. Fluorescence microscopy observation of immunofluorescence labeled *S. cerevisiae* recombinant strains harboring full length *xylA* genes. Cells were labeled with anti-FLAG antibody and Alexa Fluor 488 anti-mouse IgG. Phase-contrast micrograph (left column), anti-FLAG antibody and Alexa Fluor 488 anti-mouse IgG (right column). A: BY4741 Δ*kex*, B: BY4741 Δ*sed*, C: BY4741-1.1, D: BY4741-2.1, E: BY4741-1.2, F: BY4741-1.3, G: BY4741-1.4, H: BY4741-2.2.

Furthermore, to evaluate the amount of cells expressing XIs on their surface, the flow cytometry analysis was conducted after immunolabeling. The cells harboring pULD1-Ck or pULD1-Pi more efficiently displayed the XI protein on their surface than the cells harboring pULD1-Ot, pULD1-Ms or pULD1-Ec. These are revealed by the microscopy observation result (Figure 3.3.4.1), which corresponds to the total amount of fluorescing cells (%) that was measured by flow cytometry analysis (Figure 3.3.4.2).

First, this study attempted to express the putative full length *xylA* genes retrieved from soil metagenome in *S.cerevisiae* cells intracelluarly, however the recombinant strains harboring putative full length *xylA* genes from metagenome were unable to grow on xylose media. Conversely, the recombinant strain harboring *Piromyces xylA* as a positive control was grown on xylose media (data not shown). Then this study attempted to confirm the heterologous expression of full length *xylA* genes by cell surface engineering. Cell surface engineering could facilitate the expression of the functional protein by displaying it on the cell surface of yeast. The constructed cells can be used for whole-cell biocatalyst (Kuroda and Ueda, 2011). Based on fluorescence microscopy analysis, three putative XI proteins from soil metagenome (CK, OT and MS) and two known XIs from *E. coli* and *Piromyces* sp. were successfully displayed on *S. cerevisiae* cells as shown in figure 3.3.4.1 (D-H) where the recombinant *S. cerevisiae* cells have the green fluorescence on their surface, which was not found on the surface of the cells harboring BE *xylA* and negative control (Fig 3.3.4.1 A-C). This result may reveal that *S. cerevisiae* recombinant strain harboring BE *xylA* was not expressing BE XI on its surface or the expression was too low to be observed by microscopy analysis.

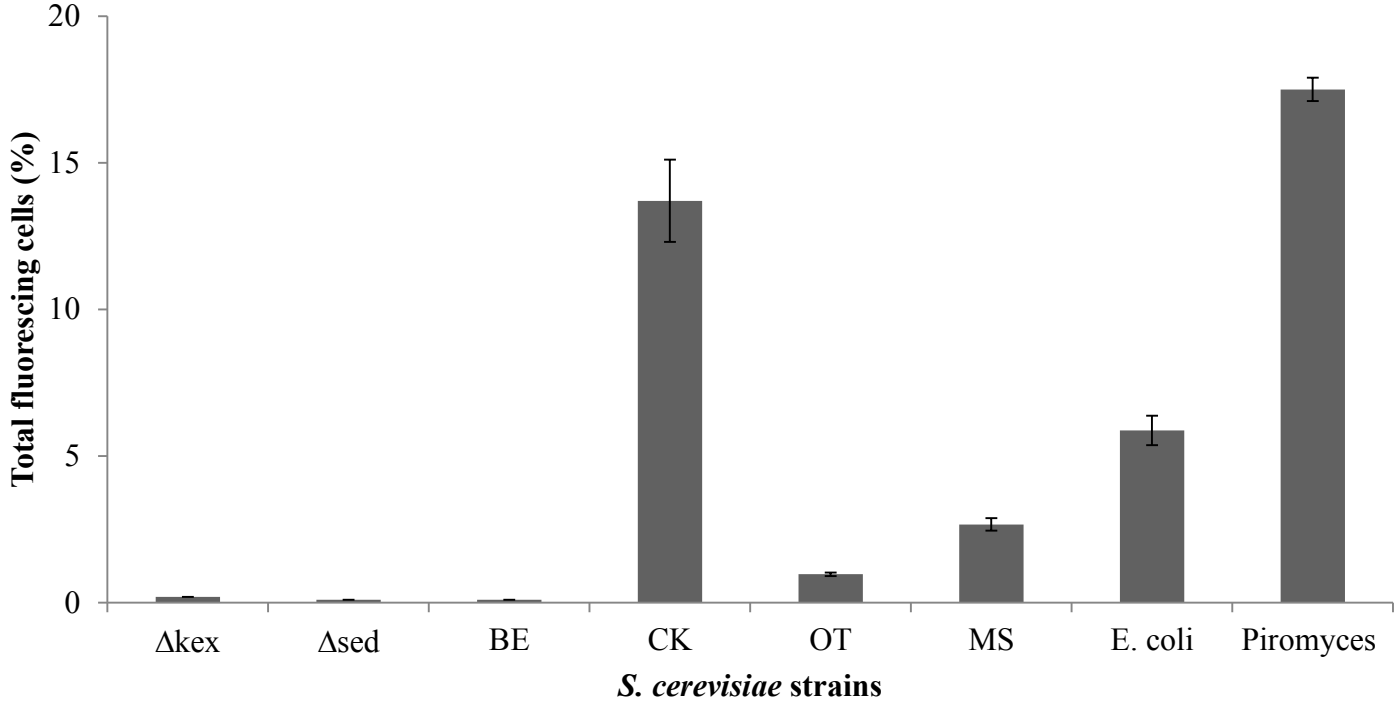


Fig. 3.3.4.2. Flow cytometry analysis of immunofluorescence labeled yeast recombinant strains. Δkex = BY4741Δkex; Δsed= BY4741Δsed; BE, CK, OT, MS, *E.coli*, and *Piromyces* are representing *S. cerevisiae* recombinant strains harboring particular full length *xylA* genes displayed on the cell surface.

A quantitative measurement of total fluorescing cells of *S. cerevisiae* recombinant strains was done by the flow cytometry analysis. Two different hosts, BY4741 Δ*kex* and BY4741 Δ*sed*, were used to introduced the constructed vector harboring full length *xylA* genes. Kex2 protease has an activity to cleave substrate having Arg-Arg, Pro-Arg, Ala-Arg, and Thr-Arg dipeptide (Brenner and Fuller, 1991). The host with with Kex2 disruption was used in order to avoid the cleavage of the XI protein, which contains the peptide that is recognized by Kex2 protease, such as XI from BE, OT, MS and *E. coli*. While for XI of CK and *Piromyces,* which are not recognized by Kex2 protease, were introduced into BY4741 Δ*sed* strain. Sed1p is a major stuctural protein in the stationary phase induced by stress and starvation and it is one of GPI-anchored proteins which could compete with the α-agglutinin fused protein for the cell surface expression (Kuroda et al., 2009). The host with Sed1 disrupted was used to increase the expression of XI fused with α-agglutinin on *S. cerevisiae* cell surface. Figure 3.3.4.2 shows total fluorescing cells among *S. cerevisiae* strains. Total fluorescing cells varied among *S. cerevisiae* recombinant strains.

This study also attempted to confirm the xylose isomerase enzyme activity catalyzing xylose conversion into xylulose by performing the whole cell biocatalyst. However, xylose isomerase protein activities were not detected from all *S. cerevisiae* recombinant strains expressing metagenomic *xylA* genes on their surface (data not shown). Conversely, Ota et al (2013) successfully confirmed that their *S. cerevisiae* recombinant strain displaying *C. cellulovorans* XI protein on its surface maintained the XI protein activity for catalyzing xylose to xylulose. *C. cellulovorans* XI required a specific cofactor for its activity. It prefered Co^{2+} as the cofactor instead of Mg^{2+} for its activity in the conversion of xylose to xylulose. This may reveal that these putative

full length *xylA* genes were expressed, however, they have no functional enzymes activity in *S. cerevisiae* possibly due to the unknown suitable cofactor for each XIs activity of metagenomic *xylA* or other reasons as previously assumed for the difficulty of bacterial XIs expression in *S. cerevisiae,* such as improper protein folding, posttranslational modifications, and intermolecular and intramolecular disulfide bridge formations (Amore et al., 1989; Walfridson et al., 1996). Similarly to these study results, Sarthy et al. (1987) reported that they found the heterologous expression of *E. coli xylA* gene in *S. cerevisiae*, and the major portion of the heterologously made *E. coli* xylose isomerase protein in *S. cerevisiae* was inactive. They subsequently examined the possible reasons related to the non-functional *E. coli* XI enzyme in *S. cerevisiae*. They found that it was not related to the posttranslational modifications or intermolecular and intramolecular disulfide bridge formations. However, they excluded the possibility that the absence of an essential cofactor or metal ion in *S. cerevisiae* may also influence to the XI enzyme activity.

References

Amore R, Wilhelm M, Hollenberg CP. 1989. The fermentation of xylose - an analysis of the expression of *Bacillus* and *Actinoplanes* xylose isomerase genes in yeast. *Appl Microb Biotechnol*. 30: 351-357.

Brenner C, Fuller RS. 1991. Structural and enzymatic characterization of a purified prohormone-processing enzyme: secreted, soluble Kex2 protease. *Proc Natl Acad Svi*. 89, 922-926.

Cottrell MT, Moore JA, Kirchman DL. 1999. Chitinases from uncultured marine microorganisms. *Appl Environ Microbiol*. 65: 2553-2557.

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM. 2010. Accessing microbial diversity for soil metagenomic studies. *Appl Environ Microbiol.* 77: 1315-1324.

Epting KL, Vieille C, Zeikus JG, Kelly RM. 2005. Influence of divalent cations on the structural thermostability and thermal inactivation kinetics of class II xylose isomerases. *FEBS J*. 272: 1454-1464.

Hahn-Hägerdal B, Karhumaa K, Jeppsson M. 2007. Metabolic engineering pentose utilization in *Saccharomyces cerevisiae*. *Adv Biochem Engin/Biotechnol*. 108: 147-177.

Henne A, Schmitz RA, Bömeke M, Gottschalk G, Daniel R. 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl Environ Microbiol*. 66: 3113-3116.

Jiang C, Wu LL, Zhao GC, Shen PH, Jin K, Hao ZY, Li SX, Ma GF, Luo FF, Hu GQ, Kang WL, Qin XM, Bi YL, Tang XL, Wu B. 2010. Identification and characterization of a novel fumarase gene by metagenome expression cloning from marine microorganisms. *Microb Cell Fact*. 9:91.

Kellenberger E. 2001. Exploring the unknown. The silent Revolution of microbiology. *EMBO Rep*. 2: 5-7.

Kuroda K, Matsui K, Higuchi S, Kotaka A, Sahara H, Hata Y, Ueda M. 2009. Enhancement of display efficiency in yeast display system by vector engineering and gene disruption. *Appl Microbiol Biotechnol*. 82, 713-719.

Kuroda K, and Ueda M. 2011. Cell surface engineering of yeast for applications in white biotechnology. *Biotechnol. Lett*. 33, 1-9.

Kuyper M, Harhangi HR, Stave AK, Winkler AA, Jetten MS, de Laat WT, Den Ridder JJ, Op den Camp HJ, van Dijken JP, Pronk JT. 2003. High-level functional expression of fungal xylose isomerase: the key to efficient to ethanolic fermentation of xylose by *Saccharomyces cerevisiae*. *FEMS Yeast research*. 4(1): 69-78.

Liu N, Yan X, Zhang M, Xie L, Wang Q, Huang Y, Zhou X, Wang S, Zhou Z. 2011. Microbiome of fungus-growing termites: a new reservoir for lignocellulose genes. *Appl Environ Microbiol*. 77: 48-56.

Ota M, Sakuragi H, Morisaka H, Kuroda K, Miyake H, Tamaru Y, Ueda M. 2013. Display of *Clostridium cellulovorans* xylose isomerase on the cell surface of *Saccharomyces cerevisiae* and its direct application to xylose fermentation. *Biotechnol Prog*. 29(2):346-51

Park JH, Batt CA. 2004. Restoration of a defective *Lactococcus lactis* xylose isomerase. *Appl Environ Microbiol*. 70: 4318-4325 .

Park SY, Shin HJ, Kim GJ. 2011. Screening and identification of a novel esterase EstPE from a metagenomic DNA library. *J Microbiol.* 49: 7-14.

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*. 1: 283-290.

Rose TM, Henikoff JG, Henikoff S. 2003. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res*. 31: 3763-3766.

Sarthy AV, McConaughy BL, Lobo Z, Sundstrom JA, Furlong CE, Hall BD. 1987. Expression of the *Escherichia coli* Xylose Isomerase Gene in *Saccharomyces cerevisiae*. *Appl Environ Microbiol*. 53(9):1996-2000.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA 4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*. 24: 1596-1599.

Walfridsson M, Bao X, Anderlund M, Lilius G, Bülow L, Hahn-Hägerdal B. 1996. Ethanolic fermentation of xylose with *Saccharomyces cerevisiae* harboring the *Thermus thermophilus xylA* gene, which expresses an active xylose (glucose) isomerase. *Appl Environ Microbiol*. 62(12):4648-51.

Yamada K, Terahara T, Kurata S, Yokomaku T, Tsuneda S, Harayama S. 2008. Retrieval of entire genes from environmental DNA by inverse PCR with pre-amplification of target genes using primers containing locked nucleic acids. *Environ Microbiol*. 10: 978-987.

Yun J, Ryu S. 2005. Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb Cell Fact*. 4:8.

Chapter 4

Chimeric *xylA* gene construction for a high-throughput screening of xylose isomerase genes from soil metagenome

Abstract

Conversion of xylose as the second most abundant sugar is significantly important for bioethanol production from lignocellulosic biomass. A baker yeast, *Saccharomyces cerevisiae* naturally cannot produce ethanol from xylose. The heterologous expression of bacterial xylose isomerase (XI) encoded by *xylA* in *S. cerevisiae* is one approach that is recently used for efficient bioethanol production. It allows direct isomerization of xylose into xylulose which can be utilized as a substrate by *S. cerevisiae, which* converts it into ethanol. However, the expression of bacterial *xylA* gene in *S. cerevisiae* remains problematic as it has low or no expression in *S. cerevisiae* recombinant strains. *Piromyces* XI is the first known highly active XI in *S. cerevisiae*. Therefore, in this study a screening system of bacterial *xylA* genes from soil metagenome was established by constructing chimeric genes. Partial *xylA* genes from soil metagenome were introduced into the deleted region of *Piromyces* XI in the PRS436GA-PiXI-opt vector which allowed homologous recombination in *S. cerevisiae*. Three *S. cerevisiae* recombinant strains which were able to grow in selective xylose medium (SX) carried chimeric *xylA* genes. Amino acid sequences of partial *xylA* genes from soil metagenome inserted into PRS436GA-PiXI-opt vector were identified by BLASTP. The inserted partial *xylA* genes have identity 95, 91 and 87% to the XI of *Niastella koreensis*, *Pedobacter heparinus* and Uncultured bacterium, respectively.

4.1. Introduction

Lignocellulosic biomass is considered as a renewable source that can be utilized as the substrate for bioethanol production. Complete conversion of all available sugar such as glucose and xylose in the lignocellulosic hydrolysate is required to obtain high ethanol yields. Xylose is one of abundant hemicelluloses composed of 25-35% of total lignocellulose (Gray et al., 2006). *Saccharomyces cerevisiae* remains the selected organism for bioethanol production. Although its wild type strain rapidly ferments hexose to ethanol in high rate and yields, they cannot utilize pentose sugar D-xylose (Wisselink et al. 2009; Chu and Lee 2007). Therefore, xylose fermentation by *S. cerevisiae* received considerable attention to efficiently produce bioethanol from lignocellulosic biomass.

The apparent way to improve *S. cerevisiae's* ability to metabolize xylose is to introduce initial xylose metabolic pathway, which converts xylose to xylulose (Hahn Hagerdal et al., 2007. The xylose isomerase (XI) gene pathway was found in naturally xylose-utilizing bacteria and some fungi (Harhangi et al., 2003; Madhavan et al., 2009). Even though some metabolic engineered *S. cerevisiae* strains have been reported to successfully express XI genes with relatively high ethanol yields from xylose (Kuyper et al., 2003; Karhumaa et al., 2007; Brat et al., 2009; Parachin and Gorwa-Grauslund, 2011), it remains insufficient for *S. cerevisiae* recombined with known XI genes to efficiently produce bioethanol. Thus, the exploration of bacterial XI genes is required in order to obtain diverse *xylA* genes for efficient screening in *S. cerevisiae*.

Soil is a compounded microbial environment (Will et al., 2010) and considered to harbor the most diverse population of bacteria of any environment on earth (Roesch et al., 2007). One gram of soil is reported to contain up to 10 billion

microorganisms and thousands of different species (Delmont et al., 2010). However, only 1% of the soil bacteria can be cultivated under standard cultivation techniques (Kellenberger, 2001). Metagenomic approaches involving the extraction of DNA from soil may provide access to novel genetic sources of uncultivated bacteria from soil. Two strategies that are generally used in metagenome screening are activity-based and sequence-based screenings. Both activity- and sequence-based screenings have individual advantages and disadvantages, and they have been applied successfully to discover biocatalysts from metagenomes such as chitinase (Cottrell et al., 1999), lipase (Henne et al., 2000), xylanase (Yamada et al., 2008; Liu et al., 2010), esterase (Park et al., 2011), fumarase (Jiang et al., 2010), and many more. Differing from the activity-based screening, which requires heterologous expressions of the genes encoded within metagenome clones and high-throughput function assays for clone identification, the sequence-based screening is not dependent on the expression of cloned genes in the heterologous host. Generally, it is based on the conserved DNA sequences of target genes (Yun and Ryu, 2005). It can disclose target genes, regardless of gene expression and protein folding in the host, and irrespective of the completeness of the gene's sequence (Li et al., 2009).

The heterologous expressions of bacterial XI genes in *S. cerevisiae* proved to be challenging as, for many years, no actively expressed enzyme was reported (Hahn Hagerdal et al., 2007). The first functionally expressed XI in *S. cerevisiae* originated from *Thermus thermophilus* (Walfridson et al., 1996). However, the activity of this enzyme in *S. cerevisiae* was lower than that of the fungus *Piromyces* sp. As observed in the highly active XI from *Piromyces*, the expression of this enzyme alone in *S. cerevisiae* only allowed very slow growth on xylose (Kuyper et al., 2003), which may also correspond to the failure of early trials for heterologous XI expression, where it

was only assayed as growth on xylose (Hahn Hagerdal et al., 2007). In this study, to overcome such a challenge, a high-throughput functional-based screening technique was established by using chimeric XI as an initial attempt to identify potentially novel, highly active and functional XI encoding genes, *xyl*A within *S. cerevisiae* from soil metagenome. This study showed that the established screening technique proved to be promising upon the identification of several partial bacterial XI homologs when expressed enzymes had activity comparable to the *Pyromices* XI in *S. cerevisiae*.

4.2. Materials and Methods

4.2.1. Strains, plasmids and media

S. cerevisiae W600 [ade2 :: ADE2ADH3 :: TAL1-TKL1-HIS3 gre3 :: RPE1-RKI1-LEU2 HIS3 :: XKS1-hph (ura3, trp1, hygromycin B resistant)] and the plasmid pRS436 carrying the *xylA* derived from *Piromyces* (PRS436GA-PiXI-opt) for homologous recombination were obtained from Toyota Central R & D laboratory. pGEMT-Easy vector system (Promega) was used as the cloning vector of partial *xylA* fragments amplified by degenerate primers. Yeast transformants were aerobically cultivated in synthetic dextrose (SD) medium [Yeast nitrogen base w/o amino acid 6.7 g/L, Glucose 20 g/L, 10 × DO supplement (-URA, -trp) 0.72 g/L, Tryptophan 0.2 g/L] and SX selection medium [yeast nitrogen base w/o Amino Acid 6.7 g/L, xylose 20 g/L, 10 × DO supplement (-URA, -trp) 0.72 g/L, Tryptophan 0.2 g/L] . *Escherichia coli* DH10B strain (Invitrogen) [F$^-$*mcr*AΔ(*mrr-hsd*RMS-*mcr*BC)φ80*lac*ZΔM15Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara*, *leu*)7697 *gal*U *gal*K λ-*rps*L *nup*G /pMON14272 / pMON7124] was used as a host for the cloning of manipulated DNA vector. *E. coli* JM 109 (Nippongene) {F'[*traD*36, *proAB*, *lacI*^{q}, *lacZ*⊿M15], ⊿ (*lac-proAB*),*hsdR*17 (r_k－m_k+), *recA*1, *endA*1, *relA*1, *supE*44, *thi*-1,*gyrA*96, *e*14－(*mcrA*－)} was used as a host for cloning partial *xylA* fragments amplified from soil metagenome. *E. coli* transformants were grown in Luria–Bertani medium (1% tryptone, 0.5% yeast extract, 1% sodium chloride) containing 100 μg/mL ampicillin.

4.2.2. Soil metagenome DNA preparation

Two soil samples were collected from Tsukuba mountain (Japan) below cedar (*Criptomeria japonica*) and beech (*Fagus crenata*) trees, named T1 and T2, respectively. Metagenomes were extracted from two soil samples using the Soil DNA Extraction Kit, ISOIL (Nippongene, Tokyo, Japan) according to the manufacturer's instruction. Metagenome DNA samples from this two sites were subsequently used as template for PCR amplification of partial *xylA* genes.

4.2.3. Construction of *xylA* cassette for vector insertion by homologous recombination

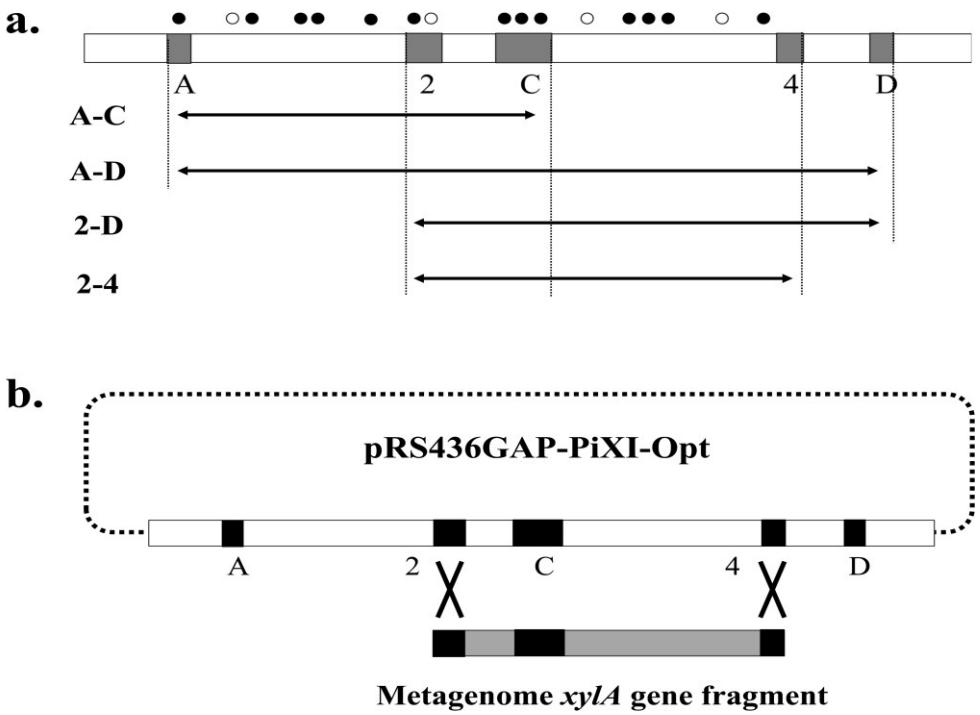
This study was aimed to establish an efficient screening system by excluding the recombination region of known active *xylA* gene in *S. cerevisiae,* derived from *Piromyces* carried by the pRS436 vector. By recombining partial unknown *xylA* genes from the environment, it was expected that active chimeric *xylA* in *S. cerevisiae* containing partial *xylA* from the environment, which may also be active in the full length form, would be obtained so that their activity could be further characterized in *S. cerevisiae* either in the chimeric or parent enzyme form.

xylA fragment subjected to the recombination was divided into four types based on its conserve regions (Figure 4.2.3.1a). Each type of partial *xylA* fragment amplified by degenerate primers was flanked by 20-25 nucleotide homolog to 5' and 3' of the linearized vector. Degenerate primers were designed for amplification of each *xylA* fragment type (Figure 4.2.3.1c). To amplify partial *xylA* genes from soil metagenome, four degenerate primer sets were designed based on specified conserved regions of 44 known amino acid XI sequences collected from the NCBI database, which resulted in four types of partial *xylA* cassettes for recombination. Figure

4.2.3.1a shows the four conserve regions that were chosen to cover all active sites and mediating sub unit interaction sites which will be subjected to the replacement by partial *xylA* genes from soil metagenome. All types of partial *xylA* fragments were amplified by PCR, analyzed by agarose gel electrophoresis and extracted from gel using electrophoresis in a dialysis membrane and purified by ethanol precipitation. In order to confirm sufficient diversity of partial *xylA* cassettes obtained by amplification using degenerate primer sets, purified PCR products were ligated into pGEMT-Easy vector (Promega) and transformed into *E. coli* JM109. White colony transformants were picked up and the sequences of cloned DNAs were analyzed.

4.2.4. Construction of pRS436 vectors for *xylA* cassette replacement

Plasmid pRS436 carrying *xylA* derived from *Piromyces* (pRS436GA-PiXI-opt) was used as the vector. Inverse PCR was carried out using PrimeSTAR® MAX DNA polymerase (TaKaRa, Japan) in order to obtain a linearized vector which proceeded the deletion of particular targeted region of PiXI. Figure 4.2.3.1a shows subjected regions for deletion and insertion of partial *xylA* in the vector. The deleted region in the vector corresponded to the *xylA* target region to be inserted into the vector. Figure 4.2.3.1b describes the replacement strategy of particular region of PiXI within pRS436GAP-PiXI-Opt by partial metagenome *xylA* fragment through homologous recombination in *S. cerevisiae*. Highly homolog sites from both vector and metagenome *xylA* fragment facilitated homologous recombination.



c.

Site 2:	**5' - TGGGGNGGNMGNGARGGNTA**
Site 4:	**5' - RAAYTSRTCNGTRTCCCANCC**
Site A:	**5' - TAYTGGCAYACNTTY**
Site C:	**5' - NGGYTCYTGNGGYTTNGGYTCDAT**
Site D:	**5' - YTTNGCRTCRAARTTNARNCCNCC**

Fig. 4.2.3.1. The high throughput screening technique for identification of *xylA* genes from soil metagenome in *S. cerevisiae*. a. Schematics of the regions employed for functional screening of *xylA* genes. Black and white dots represent the active sites and sub unit mediating interaction sites, respectively, which are conserved within the *xylA* genes. b. Schematics of the construct and homologous recombination using metagenome *xylA* fragments. Only region 2-4 is shown. c. Degenerate primers used in the amplification of the conserved regions within the *xylA* genes.

*Bam*HI and *Eco*RI restriction site and 5' phosphate-end were introduced in the primers for inverse PCR of the vector. The vector amplified by inverse PCR, which has phosphate at its both ends, was self circularized using Mighty Mix DNA ligation kit (TaKaRa-Bio) at 16°C, overnight. Circularized vector was then cloned into *E. coli* DH10B strain (Invitrogen) by heat shock method. *E. coli* transformants were selected on LB agar supplemented with ampicillin. Each colony grown on selection media carrying the amplified vector and the vector sequences were analyzed by sequencing (FasMacCo., Ltd.sequencing service) to confirm correct vector sequences. The correct vector sequences were further used for the subsequent recombination step. Linearized vector for recombination was then obtained by the digestion of circularized inverse PCR product using *Bam*HI and *Eco*RI restriction enzymes.

4.2.5. Screening of partial *xylA* genes from soil metagenome mediated by homologous recombination in *S. cerevisiae*

Both linearized pRS436GA-PiXI-opt vector and partial *xylA* cassette were transformed into *S. cerevisiae* W600 by using Frozen EZ Yeast Transformation II Kit (zymo research) according to the manufacturer's protocol. Yeast transformants were recovered in SD-URA medium for two days and selected in SX-URA medium for up to three weeks. Transformants were also grown in SD media to confirm that chimeric genes were successfully established by homologous recombination. Fifty yeast colony transformants were selected for analyzing the chimeric genes by sequencing the inserted region. On the other hand, yeast transformants which could grow in SX medium were expected to harbour plasmid pRS436 that has been circularized through homologous recombination with partial *xylA* cassette forming an active chimeric *xylA*.

Chimeric *xylA* from a single colony transformant was sequenced and analyzed. All inserted partial *xylA* sequences were then identified by BLASTP.

4.2.6. D-xylose consumption analysis and ethanol production via fermentation

Yeast clones harboring *Piromyces xylA* integrated with metagenomic partial *xylA* genes were tested for xylose degradation and subsequently tested for ethanol production via fermentation. The positive clones were inoculated at OD600 = 1 cell concentration, in 1 mL of SX medium supplemented with adenine and Complete Supplement Mixture (CSM) w/o URA in 96 deep-well plates. Fermentation of the clones was conducted at 30^{o}C in an MBR-024 Bioshaker (Taitec, Japan) at maximum speed for 72 h. Sampling was conducted every 24 h and the concentration of xylose and ethanol was conducted accordingly.

4.3. Results and Discussion

4.3.1. Construction of *xylA* cassette replacement of pRS436GA-PiXIopt vector by partial metagenome *xylA* genes through homologous recombination

As reported previously in Umemoto et al. 2011, XIs consist of thirteen active site residues (W49, F60, H101, D104, F145, W188, E232, K234, E237, N266, E268, H271, D296) and four residues mediating subunit interactions (D58, R191, L254, A275) which are conserved. In this study, we designed four sets of degenerate primers based on five conserved regions named A, 2, C, 4, and D (Figure 4.2.3.1a) in order to amplify partial metagenomic *xylA* fragments which were going to replace the same region in PiXI harbored by pRS436 vector by homologous recombination (Figure 4.2.3.1b). Four primer pairs consisting of A-C, A-D, 2-D, and 2-4 were used, covering all *xylA* active sites and subunit mediating interaction sites. Four *xylA* fragments were successfully amplified using newly designed degenerate primers consisting of A-C, A-D, 2-D, 2-4 with the PCR product fragment lengths 573, 882, 510, and 462 bp, respectively.

To confirm the homologous recombination of partial metagenome *xylA* into PiXI, 50 colonies grown on SD plate were checked by sequencing. 45 colonies positively carrying metagenome *xylA* from different species were inserted into PiXI. These partial metagenome *xylA* gene sequences, which were identified by BLASTX, have identity 75-95% to the *xylA* gene sequences in the database (data not shown). This analysis of yeast recombinant strains revealed that diverse chimeric *xylA* genes have successfully been established by homologous recombination in *S. cerevisiae* W600 strain. These chimeric genes were then further subjected for the functional

screening in the recombinant strains on a selective media with xylose as sole carbon source.

4.3.2. Diversity analysis of partial *xylA* genes amplified by degenerate primers from soil metagenome

Four sets of degenerate primers were designed based on known reported amino acid *xylA* gene sequences from database. These primers were used to amplify partial *xylA* genes from soil metagenome subjected for homologous recombination. Partial *xylA* amplified by 2-4 primer pair was evaluated. The PCR amplicons were cloned into pGEMT-Easy vector, 93 white colonies were picked for sequence analysis. Sequence analysis by BLASTX showed that 93 clones have the identity to the *xylA* genes of 86 different species (data not shown). These 86 *xylA* genes species were classified into four dominant phyla i.e *Acidobacteria*, *Proteobacteria*, *Planctomycetes*, and *Verrucomicrobia* (Figure 4.3.2.1a). It was revealed that the degenerate primers successfully amplified diverse *xylA* genes from soil metagenome and were sufficient for screening.

The subsequent diversity analysis was also performed for the yeast transformants harbouring chimeric genes obtained by homologous recombination. 45 out of 50 clones positively harboured chimeric genes analyzed by sequencing. BLASTX analysis revealed that the partial *xylA* phylotypes identified from 45 clones belong to five phyla i.e *Acidobacteria*, *Proteobacteria*, *Bacteroidetes*, *Planctomycetes*, and *Verrucomicrobia.* Partial metagenome *xylA* genes inserted into pRS436GAP-PiXI-Opt were dominated by *Proteobacteria* and *Acidobacteria* members shared 35 and 31%, respectively (Figure 4.3.2.1b).

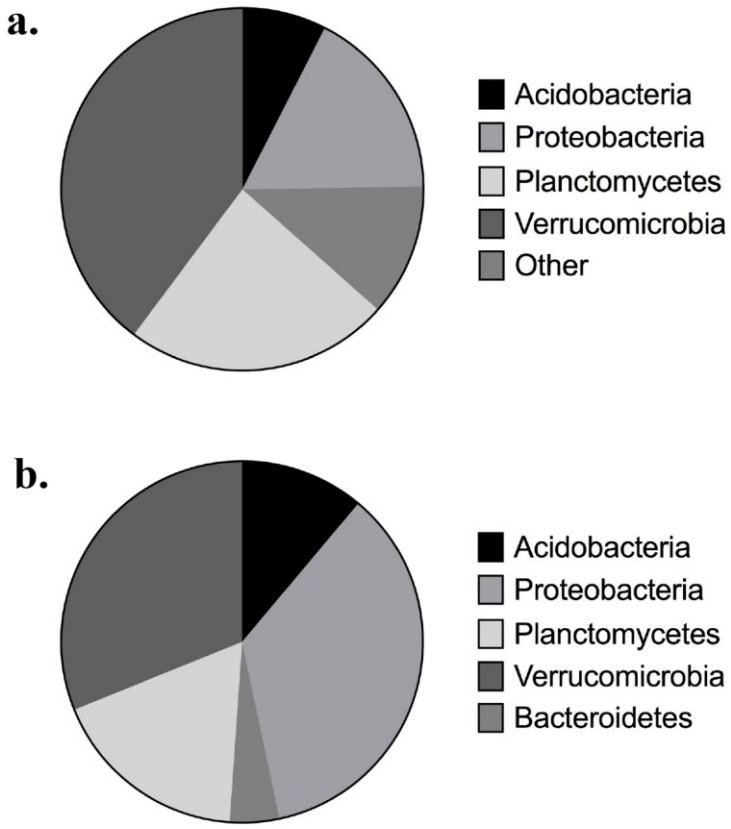


Fig. 4.3.2.1. Diversity analysis of bacterial partial *xylA* genes attained from soil metagenome. a. Diversity of partial *xylA* genes amplified using the 2-4 degenerate primers. b. Diversity of partial *xylA* genes identified after homologous recombination in *S. cerevisiae*.

4.3.3. Functional analysis of the partially inserted XI attained from metagenome samples after homologous recombination

To efficiently screen bacterial *xylA* from soil metagenome, combination of sequence- and direct function-based screening approach in *S. cerevisiae* was used in this study. Sequence-based screening approach was conducted to obtained partial *xylA* genes from soil metagenome which targeted particular sites of the genes. Meanwhile, the functional screening of chimeric *xylA* genes constructed from particular region of PiXI and partial *xylA* from soil metagenome was conducted in synthetic xylose media. Yeast recombinant strain that has the ability to grow on synthethic xylose media may harbour active chimeric *xylA*. In this study, the yeast transformed by the combination of pRS436GA-PiXIopt vector and *xylA* fragment of 2-4 and A-C regions were able to grow on media with xylose as the sole carbon source. Chimeric *xylA* gene construct was confirmed by sequencing from three recombinant strains named W600-S1, W600-S2, and W600-S3, which were able to grow on xylose (Fig. 4.3.3.1). Amino acid sequence analysis of *xylA* insert was done by BLASTP. Sequence identification showed the identity of inserts in the recombinant strains W600-S1, W600-S2, and W600-S3 were 95, 91, and 87% to the XI of *Niastella koriensis* (*Nk*), *Pedobacter heparinus* (*Ph*), and Uncultured bacterium (Ub) (Table 4.3.3.1). From four different metagenomic *xylA* fragments that were introduced into PiXI based on its active sites, three chimeric *xylA* genes were obtained, two derived from 2-4 regions and one from A-C regions, which rescued *S. cerevisiae* recombinant strains in the selective xylose media. Even though 2-4 and A-C regions consist of *xylA* active sites and mediating subunit interaction residues which are important for the enzyme activity, the replacement of these regions by partial metagenome *xylA* genes resulted in active chimeras. Conversely, the replacement of PiXI by two other *xylA* fragments i.e A-D,

and 2-D fragments resulted in inactive chimeras in *S. cerevisiae* recombinant strains for xylose consumption. It may reveal that the 2-4 and A-C regions have the prospect for screening bacterial *xylA* genes from metagenome compared to other regions and the Nk-like and Ph-like *xylA* genes screened from soil metagenome are promising active XI protein in *S. cerevisiae* as only the replacement by these *xylA* fragments resulted in active chimeras in *S. cerevisiae* for xylose consumption.

Growth evaluation of *S. cerevisiae* recombinant strains carrying chimeric *xylA* genes were conducted in SX media in order to verify active chimeric XI enzymes compared to the Piromyces XI (PiXI) expressed in pRS436GA vector. Colonies of recombinant strain W600-S2 and W600-S3, which harbour Ph-like and Ub-like XI insert has slower growth compared to the recombinant strain W600-Pi (Fig. 4.3.3.1b-c). Conversely, colonies of recombinant strains W600-S1 which harbour Nk-like *xylA* have similar growth compared to the recombinant strain W600-Pi, which harbours *Piromyces* XI (Fig. 4.3.3.1a).

Table 4.3.3.1. Inserted metagenomic *xylA* sequence identification as part of chimeric *xylA* genes expressed in *S. cerevisiae*

Recombinant strain	Soil metagenome source (amplified region)	Homology (BLASTP)			
		Query length (aa)	Organisms	Accession No.	Identity (%)
W600-S1	T1 (2-4)	80	*Niastella korensis* XI	YP005006479.1	95
W600-S2	T2 (2-4)	125	*Pedobacter heparinus* XI	YP_003091879.1	91
W600-S3	T2 (A-C)	52	Uncultured bacterium XI	AEG75766.1	87

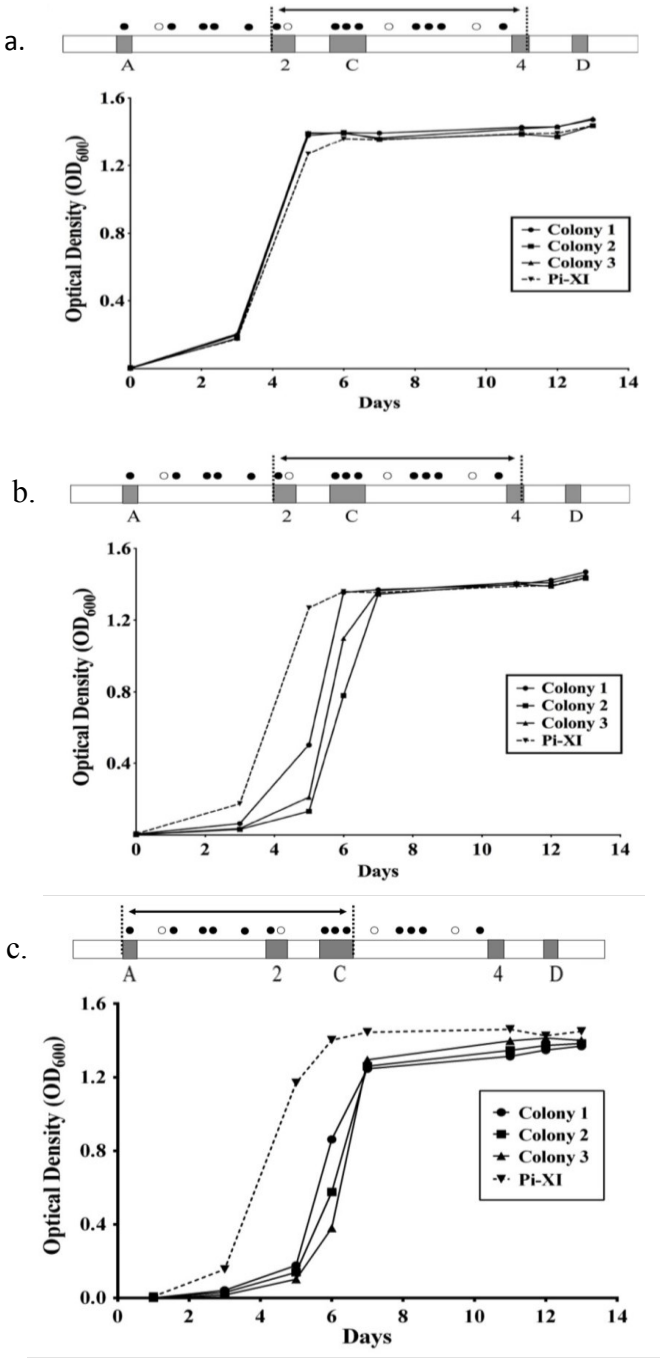


Fig. 4.3.3.1. Functional analysis of three bacterial partial *xylA* genes attained, screened from soil metagenome in *S. cerevisiae*. Reproducibility of the isolated *xylA* was determined by the random selection of 3 single colonies. The *Piromyces* sp. *xylA* was used as a control. a. Clone harbouring the partial *xylA* gene homolog of *Niastella koreensis*. b. Clone harbouring the partial *xylA* gene homolog of *Pedobacter heparinus*. c. Clone harbouring the partial *xylA* gene homolog of uncultured bacterium.

4.3.4. Fermentation characteristics of *S. cerevisiae* recombinant strains expressing the chimeric XI attained from soil metagenome

Xylose consumption and ethanol production were analyzed in two recombinant strains harbouring chimeric XI W600-S1 and W600-S2 (Fig.4.3.4.1). Anaerobic fermentation was performed in synthetic medium starting with approximately 18 g/L of D-xylose. Residual ca 7-11 g/L D-xylose was still left in the medium after 60 days of cultivation. *S. cerevisiae* recombinant strains W600-S1 (T1_2-4) have similar xylose consumption and ethanol production rate to the W600-Pi (PiXI). Whereas W600-S2 (T2_2-4) have lower xylose consumption and ethanol production rate than the W600-Pi (PiXI). These results corresponded to the growth assay of these strains described in figure 4.3.3.1. The heterologous expression of the chimeric gene harbouring partial metagenomic *xylA* homolog to *N. koreensis* has the potential to have a competitive ethanol productivity to the *Piromyces* XI.

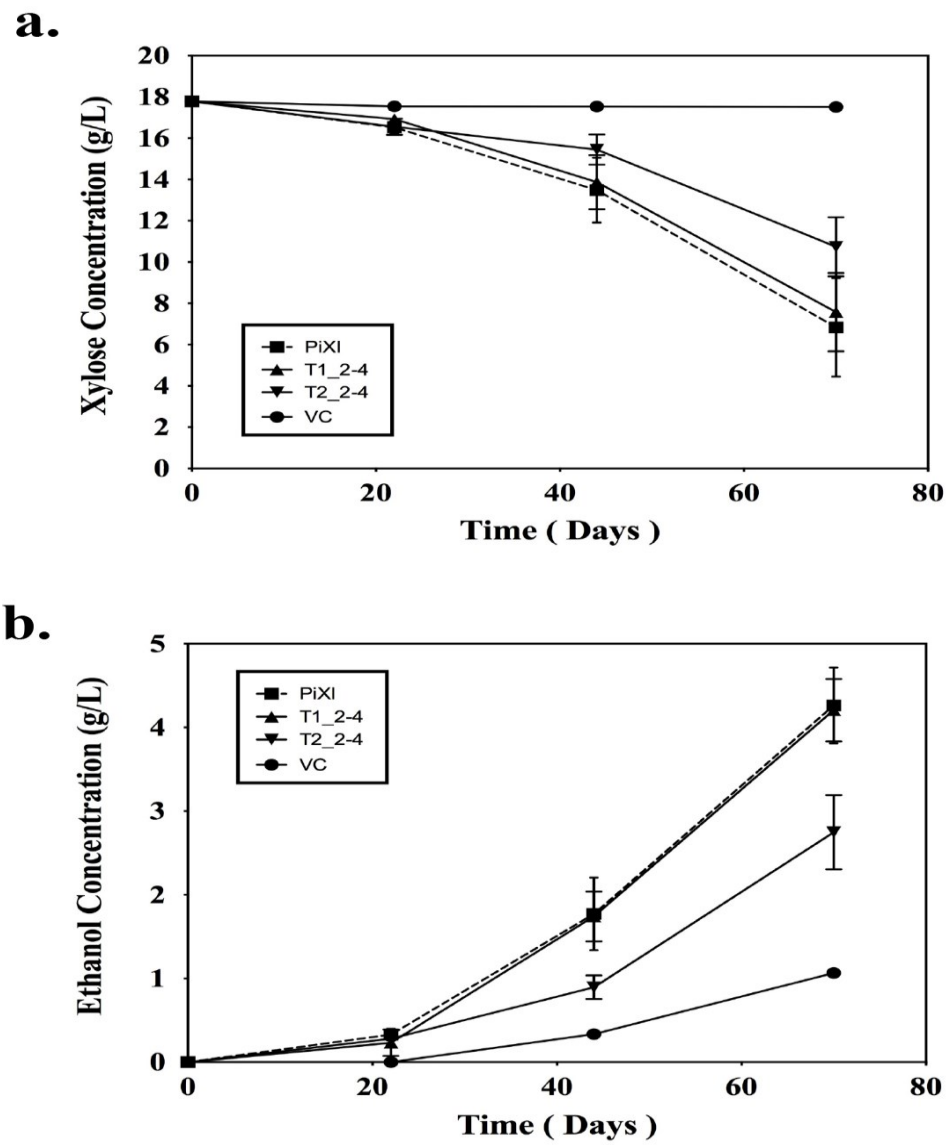


Fig. 4.3.4.1. Fermentation of *S. cerevisiae* clones harbouring the chimeric XI sequences attained from soil metagenome. a. Measurement of substrate depletion during fermentation. b. Measurement of ethanol production. Pi: *Piromyces* sp., XI: Xylose Isomerase.

References

Brat D, Boles E, Wiedemann B. 2009. Functional expression of a bacterial xylose isomerase in *Saccharomyces cerevisiae*. *Appl Environ Microbiol.* 75: 2304-2311.

Chu BC, Lee H. 2007. Genetic improvement of *Saccharomyces cerevisiae* for xylose fermentation. *Biotechnol Adv.* 25: 425-441.

Cottrell MT, Moore JA, Kirchman DL. 1999. Chitinases from uncultured marine microorganisms. *Appl Environ Microbiol*. 65(6):2553-2557.

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, and Vogel TM. 2010. Accessing microbial diversity for soil metagenomic studies. *Appl Environ Microbiol.* 77: 1315-1324.

Gray KA, Zhao L, Emptage M. 2006. Bioethanol. *Curr Opin Chem Biol.* 10(2):141-146.

Hahn-Hägerdal B, Karhumaa K, and Jeppsson M. 2007. Metabolic engineering pentose utilization in *Saccharomyces cerevisiae*. *Adv Biochem Engin/Biotechnol*. 108: 147-177.

Harhangi RH, Akhmanova AS, Emmens R, van der Drift C, de Laat WTAM, van Dijken JP, Jetten MSM, Pronk JT, Op den Camp HJM. 2003. Xylose metabolism in the anaerobic fungus *Piromyces* sp. Strain E2 follows the bacterial pathway. *Arch Microbiol.* 180: 134-141.

Henne A, Schmitz RA, Bömeke M, Gottschalk G, Daniel R. 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl Environ Microbiol.* 66(7): 3113-3116.

Jiang C, Wu LL, Zhao GC, Shen PH, Jin K, Hao ZY, Li SX, Ma GF, Luo FF, Hu GQ, Kang WL, Qin XM, Bi YL, Tang XL, Wu B. 2010. Identification and characterization of a novel fumarase gene by metagenome expression cloning from marine microorganisms. *Microb Cell Fact*. 9:91.

Karhumaa K, Garcia-Sanchez R, Hahn-Hägerdal B, Gorwa-Grauslund MF. 2007. Comparison of the xylose reductase-xylitol dehydrogenase and the xylose isomerase pathways for xylose fermentation by recombinant Saccharomyces cerevisiae. *Microb Cell Fact* 6:5.

Kellenberger E. 2001.Exploring the unknown. The silent Revolution of microbiology. *EMBO Rep*. 2: 5-7.

Kuyper M, Harhangi HR, Stave AK, Winkler AA, Jetten MS, de Laat WT, Den Ridder JJ, Op den Camp HJ, van Dijken JP, Pronk JT. 2003. High-level functional expression of fungal xylose isomerase: the key to efficient to ethanolic fermentation of xylose by *Saccharomyces cerevisiae*. *FEMS Yeast research*. 4(1): 69-78.

Li LL, McCorkle SR, Monchy S, Taghavi S, and van der Lelie D. 2009. Bioprospecting metagenomics: glycosyl hydrolases for converting biomass. *Biotechnol Biofuels*. 2:10.

Liu N, Yan X, Zhang M, Xie L, Wang Q, Huang Y, Zhou X, Wang S, Zhou Z. 2011. Microbiome of fungus-growing termites: a new reservoir for lignocellulose genes. *Appl Environ Microbiol*. 77(1): 48-56.

Madhavan A, Tamalampudi S, Srivastava A, Fukuda H, Bisaria VS, Kondo A. 2009. Alcoholic fermentation of xylose and mixed sugars using recombinant *Saccharomyces cerevisiae* engineered for xylose utilization. *Appl Microbiol Biotechnol* 82: 1037-1047.

Parachin NS, Gorwa-Grauslund MF. 2011. Isolation of xylose isomerases by sequence- and function-based screening from a soil metagenome library. *Biotechnol Biofuels* 4: 9.

Park SY, Shin HJ, Kim GJ. 2011. Screening and identification of a novel esterase EstPE from a metagenomic DNA library. *J Microbiol.* 49 (1): 7-14.

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*. 1: 283-290.

Walfridsson M, Bao X, Anderlund M, Lilius G, Bülow L, Hahn-HägerdalB. 1996. Ethanolic fermentation of xylose with *Saccharomyces cerevisiae* harboring the *Thermus thermophilus xylA* gene, which expresses an active xylose (glucose) isomerase. *Appl Environ Microbiol.* 62(12): 4648-4651.

Will C, Thürmer A, Wollherr A, Nacke H, Herold N, Schrumpf M, Gutknecht J, Wubet T, Buscot F, and Daniel R. 2010. Horizon-specific bacterial community composition of German grassland soils, as revealed by pyrosequencing-based analysis of 16S rRNA genes. *Appl Environ Microbiol*. 76: 6751-6759.

Wisselink HW, Toirkens MJ, Wu Q, Pronk JT, van Maris AJA. 2009. Novel evolutionary engineering approach for accelerated utilization of glucose, xylose, and arabinose, mixtures by engineered *Saccharomyces cerevisiae* strains. *Appl Environ Microbiol*. 75: 907-914.

Yamada K, Terahara T, Kurata S, Yokomaku T, Tsuneda S, Harayama S. 2008. Retrieval of entire genes from environmental DNA by Inverse PCR with pre-amplification of target genes using primers containing lock nucleic acids. *Environ Microbiol*. 10(4): 978-987.

Yun J, and RyuS. 2005. Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb Cell Fact.* 4:8.

Chapter 5

Conclusions

5. Conclusions

The heterologous expression of bacterial XI genes in *S. cerevisiae* is currently facing two major problems. The first known bacterial XI genes for screening in *S. cerevisiae* are limited and the second *S. cerevisiae* recombinant strains expressing bacterial XIs still need to increase their ethanol productivity. This research has conducted the diversity analysis of xylose isomerase genes from soil metagenome and developed the technique to screen xylose isomerase genes from soil metagenome.

Soil as compounded environment was proven to have high genetic diversity. In this study, the pyrosequencing-based study was conducted to measure xylose isomerase gene diversity as the higher the diversity, the larger the sequence data needs to be analyzed. The diversity analysis of xylose isomerase genes from soil metagenomes described in chapter 2. The total sequence reads of the pyrosequencing efforts conducted in this study resulted 158,555 sequence reads of *xylA* gene sequences analyzed from three soil samples, which were classified into 1,127 distinct phylotypes. Richness and diversity indicated by Shannon index of XI genes ranging from 4.3-5.03 revealed that the soil metagenome samples used in this study have high diversity. The high diversity of *xylA* gene in the soil metagenome analyzed in this study shows that these metagenome samples are promising for screening novel xylose isomerase genes. Therefore, in chapter 3, the clone library sequences of *xylA* gene were used to retrieve full length of the genes. Flanking sequences of six partial metagenome *xylA* sequences were successfully obtained and four sequences identified as full length genes through ORF finder analysis. These full length *xylA* genes were subjected to the heterologous expression analysis in *S. cerevisiae* cells. Functional analysis by intracellular expression in *S. cerevisiae* showed no detectable XI protein

expression by assaying the recombinant strains based on their growth on xylose media. Thus, the full length *xylA* genes expression were analyzed by using cell surface engineering where the protein expressed by *S. cerevisiae* will be diplayed on their cell surface. Three putative full length *xylA* genes designated as CK, OT and MS were successfully displayed on the *S. cerevisiae*'s cell surface, which confirmed that XI protein can be detected on the *S. cerevisiae* recombinant strains. However, measuring the activity of XI protein in catalyzing the conversion of xylose to xylulose by the whole cell biocatalyst remains problematic. It happens possibly because the activity of the XI protein was still too low to be detected or because the heterologous xylose isomerase proteins from soil metagenome *xylA* genes in *S. cerevisiae* have no functional activity.

On the other hand, chapter 4 describes the attempts to obtain highly active *xylA* genes from soil metagenome by shuffling partial *xylA* gene of *Piromyces* XI, which is known as active XI in *S. cerevisiae*. A high-throughput screening system was established by employing a particular region of *Piromyces* XI, which is replaced by partial metagenome *xylA* forming chimeric genes. In this system, sequence- and function-based screening were simultaneously applied. Constructed chimeric genes directly screened for their activity as growth on xylose media. Three chimeric genes harboring partial metagenome *xylA* homolog to *Niastella koreensis*, *Pedobacter heparinus, and* Uncultured bacterium were screened as active chimera as the *S. cerevisiae* recombinant strains harboring these three chimeras showed the ability to grow on xylose. However, subsequent xylose substrate depletion and ethanol yields analysis of *S. cerevisiae* recombinant strains harboring chimeric *xylA* genes only observed in two recombinant strains harboring partial metagenome *xylA* genes

homolog to *Niastella koreensis* and *Pedobacter heparinus* while they were being grown on xylose.

In summary, to overcome two major problems mentioned above, some improvement in the results has been obtained for xylose isomerase genes expression in *S. cerevisiae*. In this study, diverse xylose isomerase genes identified from soil metagenomes were obtained and based on these partial sequences, full length *xylA* genes were retrieved from soil metagenome. On the other hand, by using the newly developed screening system, functional chimeric XIs in *S. cerevisiae* were obtained.

In the future work, codon usage optimization will be conducted for both full length and chimeric XI genes obtained in this study in order to further improve the XI genes activity in *S. cerevisiae.*

Aknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Haruko Takeyama, for allowing me to join her research group. Without her guidance and persistent help, this dissertation would not have been possible. Her invaluable help of supervision, persistent support, constructive comments and suggestions throughout this work have contributed to the success of this research.

I would also like to thank the committee members, Professor Toshio Ohshima, Professor Masamitsu Sato, and Professor Tsuyoshi Tanaka for their evaluation, suggestions, and comments to this dissertation.

I would like to express many thanks to Dr. Michihiro Ito for conducting the pyrosequencing analysis of *xylA* and 16S rRNA genes, bioinformatic analysis, and his significant contribution to writing the paper "analysis of bacterial xylose isomerase gene diversity with gene-targeted metagenomics". Many thanks also go to Toru Maruyama for the bioinformatic analysis for this study.

I would like to address my deep thanks and appreaciation to Yukari Tani and Takeshi Sakamoto for the screening of chimeric *xylA* genes in *S. cerevisiae* and growth evaluation of the recombinant strains.

I would like to address my deep thanks to Yuma Hamamoto for the analysis of *xylA* genes activity by whole cell biocatalyst.

I would like to deeply aknowledge Dr. Takeshi Terahara, Dr. Tetsushi Mori, and Dr. Yoshiko Okamura for their willingness to share their knowledge and expertise in molecular biology works, scientific knowledge discussions, suggestions, and their invaluable help during my PhD study.

Many thanks also go to Dr. Masahito Hosokawa for his invaluable help to the completion of this dissertation.

Furthermore, I would like to express my appreciation and many thanks to all the staff and members in Professor Takeyama's Laboratory for their invaluable help and support in laboratory work, data analysis, suggestions, paperwork and presentations.

I would like to thank Global COE Program of Ministry of Education, Culture, Sports, Science and Technology (MEXT) Center for Practical Chemical Wisdom and to the Ministry of Research and Technology of Indonesia for their financial support towards my PhD program.

Finally, I would also like to express special appreciation and thanks to my mom, brothers and sisters and to all of my family members and friends for their love, prayers and support in my struggling moments throughout my PhD course.

List of Publications and Conference Presentations

Publications

D. Nurdiani, M. Ito, T. Maruyama, T. Terahara, T. Mori, S. Ugawa, H. Takeyama. Analysis of the bacterial xylose isomerase gene diversity with gene targeted metagenomics. J. Biosci. Bioeng. (in printing).

D. Nurdiani, T. Mori, Y. Tani, T. Sakamoto, C. Imamura, K. Tokuhiro, H. Takeyama. Chimeric xylose isomerase gene construction for a high-throughput screening of xylose isomerase genes from soil metagenome. (in preparation).

Conference Presentations

D. Nurdiani, Y. Okamura, T. Terahara, N. Takehiro, H. Takeyama. Efficient screening of xylose isomerase genes from soil metagenome for bioethanol production. International Union of Microbiological Societies 2011 Congress, September 6-16, (2011), Sapporo, Japan.

D. Nurdiani, Y. Hamamoto, T. Mori, K. Kuroda, M. Ueda, H. Takeyama. Efficient screening of xylose isomerase genes from soil metagenome for bioethanol production. German-Japanese 2nd Joint Symposium for Diamond Researchers on Sustainable Life Science Innovation and Biomedical Research, February, (2012), Tokyo, Japan.

D. Nurdiani, Y. Okamura, T. Terahara, N. Takehiro, H. Takeyama. Molecular diversity of xylose isomerase for bioethanol production. 第 4 回バイオ関連化学シンポジウム, September, (2010), Osaka, Japan.