平成２６年

博士（バイオサイエンス）

Establishment of methods for microbial genome analysis using next-generation sequencers and their applications for microbial genomics.

次世代シーケンサーを用いた微生物ゲノム解析手法の確立と その微生物ゲノミクスへの応用

志波　優

CONTENTS

General Introduction

General Introduction

In bacteriology, the genomic era began in 1995, when the first complete bacterial genome was sequenced using conventional Sanger sequencing method.[1] Due to limitations of Sanger sequencing with respect to throughput and costs,[2] genome sequencing projects had spent a lot of budgets and years of effort. In 2005, the advent of the first next-generation sequencing (NGS) instrument, Roche/454 GS20 became available.[3] Since then, NGS now delivers sequence data thousands of times more cheaply than Sanger sequencing. History of DNA sequencing is summarized in Table 1.

Classification of NGS

NGS platforms can be divided into two broad groups depending on its read length.[4] Short-read sequencing platforms typically generate shotgun sequences with 35-200 bp read length. Long-read sequencing platforms generally produce shotgun sequences with 500-3,000 bp read length. Within these broad categories, there is considerable variation in performance, including throughput and error rate (Table 2). In addition, factors affecting usability, such as cost and run time should be considered.

In 2013, the most widely used short-read sequencing platform is HiSeq 2000/2500 (Illumina).[5] This platform can produce 600 giga bases (Gb) with 2 x 100-bp read length in one run (run time, 11 days, Table 2). MiSeq (Illumina), the smaller and lower throughput version of the HiSeq 2000/2500 machine, is also widely used in microbiological laboratories. It produces only 1.5 Gb compared to HiSeq, but it greatly reduces run time to 27 hours. Now Illumina's platform also achieves the highest read accuracy in NGS systems. While it uses fluorescence imaging technologies

for the detection of base, Ion PGM/Ion Proton platform (Life Technologies) adopts semiconductor-based technologies.[6] Ion Proton can produce up to 10 Gb with up to 200-bp read length (Proton I chip) within 2 hours. However, this platform uses pyrosequencing technologies and is prone to make mistakes in homopolymers.[4]

The most widely used long-read sequencing platform is 454 GS FLX+ (Roche). This platform can produce 0.7 Gb with 700-800-bp read length in one run (23 hours, Table 2). This read length is almost comparable to conventional Sanger sequencing, thus 454 GS FLX+ is mainly used for *de novo* genome sequencing. But this platform provides lower throughput and needs higher per-base costs than other platforms. In addition, due to pyrosequencing technologies, it is prone to make mistakes in homopolymers. So the combination of 454 and Illumina platforms is common in *de novo* genome sequencing.

PacBio RS platform (Pacific Biosciences) is for so-called 'third-generation' sequencing. Compared to the above platforms, this platform does not need amplifications of sequencing molecule, which is called 'single-molecule' sequencing.[7] It can produce 3 Gb reads with 3-kb (maximum 23 kb) read length in one run (Table 2). This platform can produce very long read than any other platform, but the base accuracy of the first version was very low. However, as recent improvements have increased greatly its base accuracy,[8] this platform is very promising for *de novo* genome sequencing now. In this thesis, I adopted PacBio RS for *de novo* genome sequencing (See chapter III).

NGS applications for microbiology

There are four major NGS applications for microbiology: re-sequencing, *de novo* assembly, RNA-seq, and meta genomics (Fig. 1).[9] Re-sequencing is one of the major areas of

application for NGS. For re-sequencing, short-read sequencing platform is suitable, owing to its high throughput, high accuracy, and cost-effective features. Sequence reads that are mapped to a known reference sequence can be used to identify single-nucleotide polymorphisms (SNPs; also called single-nucleotide variations, SNVs), small insertions or deletions (collectively called InDels), and copy number variations (CNVs) or other structural variations (SVs). These genetic differences between a reference genome and a targeted genome help the better understanding of the genetic basis of phenotypic differences.

When a closely related reference genome is unavailable, we have to consider *de novo* assembly. For *de novo* assembly, long-read sequencing platform is suitable. It can produce a draft genome rapidly, typically with dozens to hundreds of contig sequences. However, because of the presence of repetitive sequences such as rRNA operons and insertion sequences (IS), no second-generation sequencing was able to generate accurate one-contig-per-replicon assemblies that might be equal to a finished genome. Finishing genome using NGS still remains a time-consuming task. Recently, PacBio RS platform successfully demonstrated the production of a complete genome sequence (See chapter III).

RNA-seq is sequencing of cDNA to quantify gene expression or/and identify novel transcripts. Compared to microarrays, RNA-seq offers high signal-noise ratio, nearly unlimited dynamic range, and single-base resolution, and is applicable to microorganisms without reference sequences. RNA-seq helps the better understanding of regulation of gene expression through an experiment with different culture conditions.

Meta genomics is defined as the direct genetic analysis of genomes contained with an environmental sample. While traditional microbiology relies upon cultivated clonal cultures, meta genomics can produce a profile of diversity in the environment. Meta genomics gives genetic

information on potentially novel biocatalysts or enzymes, genomic linkages between function and phylogeny for uncultured organisms, and evolutional profiles of function and structure of a community.[10]

Bioinformatical challenges of NGS for microbial genomics

Because NGS produces a huge amount of data and this field is new and progresses rapidly, bioinformatical challenges are increasing exponentially. The aim of this thesis is to establish novel bioinformatical methods for NGS and apply them to microbial genomics. Especially, I have dealt with applications of re-sequencing and *de novo* assembly.

Since I started NGS five years ago, a huge number of software for NGS has been released. However, even today, there is no integrated application program for every application of NGS. It is necessary to combine a variety of software tools for each purpose. In chapter I, to expedite re-sequencing analysis, I have developed the analysis pipeline, NODAI Sequence Annotation Pipeline (NSAP). I applied NSAP to re-sequencing of *Bacillus subtilis.* In chapter II, to annotate variations detected from re-sequencing analysis, I have developed the software, COVA (comparison of variants and functional annotation). I applied NSAP and COVA to re-sequencing of *Saccharomyces cerevisiae.* In chapter III, I have established the methods for determination of novel genome sequencing using NGS and genome annotation, and applied these methods to determine the complete genome sequencing of a lactic acid-producing bacterium.

References

1. Fleischmann, R. D., Adams, M. D., White, O., et al. 1995, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. Service, R. F. 2006, Gene sequencing. The race for the $1000 genome. *Science*, **311**, 1544–1546.
3. Margulies, M., Egholm, M., Altman, W. E., et al. 2005, Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
4. Loman, N. J., Constantinidou, C., Chan, J. Z. M., et al. 2012, High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat. Rev. Microbiol.*, **10**, 599–606.
5. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., et al. 2008, Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
6. Rothberg, J. M., Hinz, W., Rearick, T. M., et al. 2011, An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.
7. Eid, J., Fehr, A., Gray, J., et al. 2009, Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
8. Chin, C.-S., Alexander, D. H., Marks, P., et al. 2013, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, **10**, 563–569.
9. Nowrousian, M. 2010, Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot. Cell* , **9**, 1300–1310.
10. Thomas, T., Gilbert, J., and Meyer, F. 2012, Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.*, **2**, 3.

Table 1. History of DNA sequencing

Year	Events
1973	Publication of the DNA sequence of 24 bp
1977	Publication of Sanger sequencing method
1983	Development of polymerase chain reaction (PCR) method
1987	Invention of first automated sequencer
1995	First success in sequencing of the bacterial complete genome (*Haemophilus influenzae*)
1996	Invention of capillary sequencer
2005	Invention of 454 Life Sciences (Roche) Next Generation Sequencing system
2007	Invention of Solexa (Illumina) and ABI (Lifetech) Next Generation Sequencer
2009	Invention of Helicos single molecule sequencer
2011	Invention of Pacific Biosciences single molecule sequencer
2012	Demonstration of ultralong single molecule reads by Oxford Nanopore Technologies

Table 2. Performance of typical next-generation sequencers (NGS)

Sequencer	**ABI 3730xl**	**454 GS FLX+**	**Genome Analyzer IIx**	**HiSeq 2500**	**Pac Bio RS**
Maker	Life Technologies	Roche	Illumina	Illumina	Pacific Biosciences
Generation	1st	1st	2nd	2nd	3rd
Chemistry	Sanger	Pyro-sequencing	Sequence-by-synthesis	Sequence-by-synthesis	Single molecule real-time sequencing
Read length	400-900 bp	700-800 bp	2 x 100 bp	2 x 100 bp	3-15 kb
Data output per run	1.9~84 kb	0.7 Gb	60 Gb	600 Gb (HO mode)	0.1-0.7 Gb
Accuracy (%)	99.999[1]	99.9[1]	98[1]	98[1]	82.1-84.4[2]
Run time per run	20 minutes - 3 hours	23 hours	10 days	11 days (HO mode)	2-3 hours

Accuracy was referred from Liu, L. et al. (2012)[1] and Shin, S. C. et al. (2013)[2].

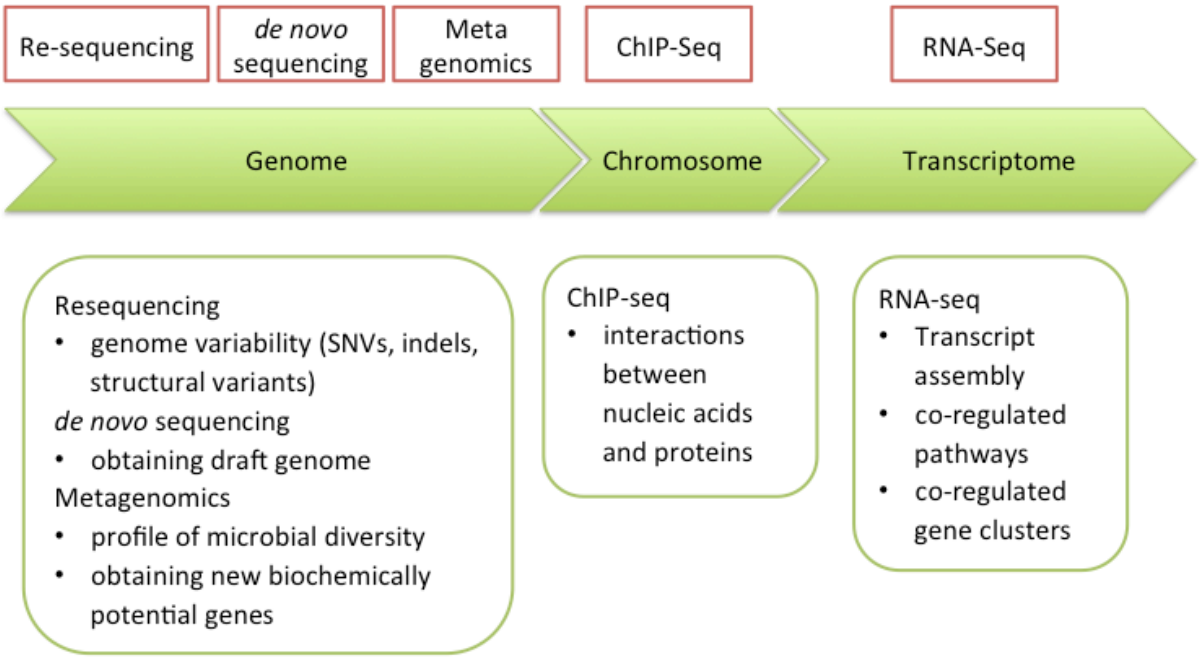

Figure 1. Application of NGS for microbial genomics.
This figure is modified from Nowrousian, M. (2010).[9]

CHAPTER I

Development of an analysis pipeline for NGS data and its application to re-sequencing of *Bacillus subtilis* laboratory strains

CHAPTER I

Development of an analysis pipeline for NGS data and its application to re-sequencing of *Bacillus subtilis* laboratory strains

Abstract

With the recent advent of next-generation sequencing technologies, one can achieve an accurate characterization of a mutant genome relevant to a parental reference strain. In order to identify mutations using NGS, it is necessary to combine a variety of software tools, such as for quality control of sequenced reads, alignment, detection of variations, and functional annotation. To facilitate re-sequencing analysis with multiple samples, I developed a suite of automated tools, NODAI Sequence Annotation Pipeline (NSAP), and applied it to re-sequencing of *Bacillus subtilis*. The strain of *B. subtilis* 168 used in laboratories in Japan was distributed in the 1990s when the sequencing consortium commenced operations. After 20 years of use of *B. subtilis* 168 in many laboratories, observations of variations in growth phenotypes have been reported. In this study, to uncover laboratory-specific variations of *B. subtilis* 168 strains in Japan, I re-sequenced 11 *B. subtilis* 168 strains from nine laboratories and analyzed how their genomes differed. It was found that the 168 strains from different laboratories differed by 1-7 variations. These variations might have been caused by differences in storage conditions in the laboratories or differences among colonies of the original stock. Based on my results, researchers ought to understand the genetic differences among wild-type (parental) strains in different laboratories and the reference strain by re-sequencing analysis, and ought to pay more attention to the management of laboratory strains. In addition, NSAP successfully demonstrated the dramatical reduction of time for bioinformatical analysis.

Introduction

With the recent advent of next-generation sequencing technologies, one can achieve an accurate characterization of a mutant genome relevant to a parental reference strain. Because of the relatively small genome size of bacteria, sample multiplexing (also called barcoding or indexing) makes it possible to sequence dozens of bacterial whole genomes in one sequencing run. This dramatically accelerates the speed of analysis in bacterial genetics.

Then researches are facing the challenge of bioinformatical analysis of enormous NGS data. In order to identify mutations using NGS, it is necessary to combine a variety of software tools, such as for quality control of sequenced reads, alignment, detection of variations, and functional annotation (Fig. 1). When one handles multiple samples, it takes a lot of time for running various software tools manually for each sample. So I developed a suite of automated tools, NODAI Sequence Annotation Pipeline (NSAP). NSAP integrates several software tools for alignment, *de novo* assembly, and detection of structural variants. Once researchers write a recipe file that describes a protocol of analysis, NSAP automatically executes those software tools. NSAP can dramatically save the time of bioinformatics analysis.

For re-sequencing application, an accurate reference sequence is necessary to identify the variations that cause a mutant phenotype. *Bacillus subtilis* is used as a model for Gram-positive bacteria for more than half a century. The genome of the *B. subtilis* 168 strain was sequenced by an international consortium in the 1990s, and was published in 1997.[1] Since sequencing the genome involved more than 30 laboratories, there was concern about sequencing errors. Srivatsan *et al.* identified 1,519 variations of the published sequences using a next-generation sequencer.[2] This published reference sequence was updated by re-sequencing using next-generation sequencing technologies in 2009.[3]

The strain of *B. subtilis* 168 used in laboratories in Japan was distributed from Dr. Naotake Ogasawara's laboratory (Nara Institute of Science and Technology) when the Japanese sequencing consortium commenced operations in the 1990s (Fig. 2).[4] I noticed that there were approximately 20 variations between the 168 strain in Dr. Hirofumi Yoshikawa's laboratory (Tokyo University of Agriculture) and the reference sequence[3] when I re-sequenced it using next-generation sequencing (data not shown). One possibility is that errors still exist in the updated reference. Another is that laboratory-specific variations exist. For example, it has been found that isolates of *E. coli* MG1655 from different laboratories can display considerable variations.[5] A similar observation was reported for cyanobacteria.[6] Similarly, after 20 years of use of *B. subtilis* 168 in many laboratories, observations of variations in the growth phenotype were reported (*e.g.*, variations in the growth rate on minimal media), and hence a new culture derived from the original collection of *B. subtilis* strain 168 was used in functional analysis and re-sequencing projects.[3]

In the post-genome era, study depends extensively on genomic information. Thus obtaining genome information on laboratory strains would aid in understanding the reproducibility of results between laboratories. In this study, to uncover laboratory-specific variations of *B. subtilis* 168 strains in Japan, I re-sequenced 11 *B. subtilis* 168 strains from nine laboratories including different source (BGSC-1A1) and analyzed how their genomes differed.

Materials and methods

Bacterial strains used in this study

B. subtilis 168 strains used in this study are described in Fig. 2. Preparation of sequence grade DNA was performed at respective laboratories.

Library preparation and sequencing with Illumina Genome AnalyzerII

Workflow from library preparation to sequencing is illustrated in Fig. 3. Sequenced libraries were prepared following the manufactures' protocols. Briefly, 3 μg of genomic DNA was fragmented to an average length of 200 bp by the Covaris S2 system (Covaris, Woburn, MA). The fragmented DNA was repaired, a single A nucleotide was ligated to the 3' end, Illumina Index PE adapters (Illumina, San Diego, CA) were ligated to the fragments, and the sample was size-selected aiming for a 300-bp product with E-Gel SizeSelect 2% (Invitrogen, Grand Island, NY). The size-selected product was amplified by PCR for 18 cycles with primers InPE1.0, InPE2.0, and Index primer containing a unique-index tag for each individual sample. The final product was validated by Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA). Pooled libraries were sequenced on Illumina Genome Analyzer II following the manufactures' protocols, generating 75- or 91-bp paired-end reads and 6-bp index tags. Sequence data were generated on two separate runs of an average size of 300 Mb per sample. Details of the output data are given in Table 1.

Implementation of NSAP

Architecture of NSAP is illustrated in Fig. 4. The NSAP analysis pipeline consists of three major steps: quality control of Illumina sequence files, primary analysis such as mapping analysis and *de novo* assembly, and secondary analysis such as statistics of mapping, detection of structure variation, and genome comparison of assembled sequences. NSAP is written in Ruby script language and runs on Linux systems. NSAP takes a recipe file that describes the analysis workflow in CSV format (Fig. 5). NSAP outputs an analysis report file that integrates results from various tools (Fig. 6).

Bioinformatics analysis

Sequence reads from each sample were analyzed using NSAP. In NSAP at first, quality control of sequence reads were performed. Then, sequence reads were mapped onto the *B. subtilis* 168 reference genome (accession no. NC_000964.3), with BWA software (ver. 0.5.1)[7] with default parameters. Possible variations (SNP/Indel) were listed using SAMtools software (ver. 0.1.9).[8] To identify correct variations, I applied the following filtering criteria to the candidates: (i) The coverage at the non-reference allele of at least 5; (ii) Indels must meet an SNP quality threshold of 50 and substitutions must meet an SNP quality threshold of 20 (SAMtools assigns SNP quality, which is the Phred-scaled probability that the consensus is identical to the reference); (iii) variations must meet a mapping quality of 30 (SAMtools assigns Mapping quality, which is the Phred-scaled probability that the read alignment is wrong); (iv) percentage of sequence reads showing the variant allele must exceed 55%. The filtered lists of variations were annotated by COVA (comparison of variants and functional annotation) (developed in chapter II). To discover structural variations, I computed the average read depth within 1-kb windows and used BreakDancer software (ver. 0.0.1r81)[9] with default parameters.

Confirmation of variations with Sanger sequencing

Genomic regions of 500-base around the SNPs and Indels detected by BWA and SAMtools were amplified by PCR and sequenced directly on a capillary sequencer by the Sanger method using the commercial sequence service of macrogen (Tokyo). The primers used are listed in Table 2.

Accession codes

The raw sequence reads used in this study are available at the DDBJ Sequence Read Archive (DRA) under accession no. DRA000969.

Results and Discussion

To facilitate re-sequencing analysis with multiple samples, I developed the NSAP pipeline and applied it for re-sequencing of *B. subtilis* laboratory strains. In order to reveal laboratory-specific variations among *B. subtilis* 168 strains in Japan, 11 *B. subtilis* 168 strains from nine laboratories including different source (BGSC-1A1) were re-sequenced and analyzed how their genomes differ using NSAP. The workflow of this study is illustrated in Fig. 7.

Comparison of variations among the 11 B. subtilis *168 strains used in Japan*

Two *B. subtilis* 168 phyletic lines are used in Japan. One is the strains labeled 168, the another is those labeled BGSC-1A1, an independent isolate of the 168 strain is deposited in the Bacillus Genetic Stock Center (BGSC).[10] The distribution of variations (SNPs, Indels and large deletions) detected in the strains is shown in Fig. 8.

The distribution of variations is clearly distinguished between the 168 and the BGSC strains. Compared with 168, the BGSC strains have additional variations. Despite the different sources of NAIST and BGSC, there were common variations among all strains, as well as unique variations in each strain. These common variations appeared not to be true variations, but sequencing errors of the reference sequence (see below). Compared with 168-C and its subculture of 168-C', an accumulation of variations was found in 168-C'. In 168-F, deletion of 48kb SKIN element was found. Details are discussed below.

The numbers of detected variations in seven *B. subtilis* 168 strains are summarized in Table 3. The 168-A strain was distributed from Europe to the Japanese sequencing consortium about 20 years ago, and 14 variations were found to be different from the updated reference sequence[3] (Table 4). All the variations of the 168-A strain were validated by Sanger sequencing. There are two possibilities to account for these discrepancies. One is that they reflect differences in the independent isolates (colonies) of the 168 strain. The second possibility is that errors remain in the updated reference sequence. Barbe *et al.* reported that the differences of isolates was 6 bases in the strain JH642, and 13 bases in case of the strain SMY.[2] Thus, a dozen variations would arise from independent isolates (colonies) of the same strain. 168-A has approximately twice as many indels as the SNPs. When the old reference sequence of 168[1] was updated by re-sequencing, 454 pyro-sequencing and Illumina sequencing were used.[3] Pyro-sequencing technology tends to miscall Indels in homopolymer runs. While erroneous Indels were tried being corrected by Illumina sequence reads, they may have remained. Since all variations of the 168-A strain were detected in all strains including different isolates of BGSC-1A1, it is likely that these 14 variations of 168-A were due to sequencing errors in the updated reference sequence.[3] However, most of them detected in this study fell into the intergenic region, and thus had little influence on the annotations (Table 4).

Genomic difference of B. subtilis *168 strains among laboratories in Japan*

To evaluate the divergence of 168 strains among the laboratories in Japan, further comparison of the variations in several strains was performed. While the distribution of variations in 168-C was identical to that in 168-A, some strains have additional variations. The 168-B to 168-F strains have one to seven variations as compared to 168-A (Fig. 9). Especially, 168-C' which is a subculture of 168-C strain in the same laboratory, accumulated six variations. It has been reported

that two isolates of the *B. subtilis* JH642 strain differed by as few as six base pairs.[2] The differences between 168-A and other the 168 strains are of the same magnitude. These variations might be caused by differences in storage conditions in the laboratories or differences among colonies of the original stock.

The laboratory-specific variations include two intergenic changes, five synonymous (silent) changes, three frame-shift changes, and eight missense mutations (Table 5). Frame-shift changes occurred in *phoA*, *opuD*, and *yqxJ,* whose functions are alkaline phosphatase, glycine betaine transporter, and unknown respectively. Missense mutations are found in *cspR*, *wprA*, *dppE*, *topA*, *yosH*, *ywfH*, *yxzC,* and *tetB*. *cspR* is an essential gene that encodes putative rRNA methylase.[11] *wprA* and *dppE* are localized in the cell membrane. They encode a cell wall-associated protease[12] and a dipeptide ABC transporter,[13] respectively. *topA* is perhaps important for DNA topology.[14] *yosH* and *ywfH* are of unknown function. *ywfH (bacG)* involves the biosynthesis of the antibiotic bacilysin.[15] *tetB* involves resistance to tetracycline.[16] The phenotypic differences resulting from these variations have not been examined, and it remains possible that there are phenotypic differences from 168-A. Moreover, deletion of a 48-kb SKIN region was found in the 168-F strain from the read-depth plot (Fig. 10 A, B). The shape of the read-depth plot also indicated a difference in growth conditions. In 168-A, an almost flat shape indicated that the cells were in the stationary phase. On the other hand, in 168-F, the ratio of sequence reads around the origin region was approximately 2-fold higher than of the *terC* region, indicating that the cells were in the log phase. The SKIN element interrupts a gene encoding a sporulation-specific sigma factor, SigK. When cells enter developmental phases, the SKIN element is excised from the genome only in the mother cell, allowing its flanking regions to join and form a complete gene, and thus turns on sporulation genes (Fig. 10 C).[17] However, some of the other isolates maintained in laboratory-F were found to retain the SKIN element (data

not shown). These results indicate a small amount of heterogeneity in the same laboratory.

Genomic divergence of B. subtilis *168 (BGSC-1A1) strains among laboratories in Japan*

The 168 strain distributed to the sequencing consortium was the one conserved by C. Anagnostopoulos.[3] On the other hand, BGSC-1A1 is an independent isolate of the same 168 strain that was deposited in BGSC in the mid 1970s by James Shapiro through P. Shaeffer and A.L. Sonenshein (Fig. 2).[18] With an old version of reference sequence,[1] Srivatsan *et al.* compared 168 between BGSC-1A1 strains, and reported 31 variations.[2] This accumulation of variations in BGSC-1A1 is compatible with the fact that BGSC-1A1 had a longer strain history than the 168 strain. With updated reference sequence of 168, 53-62 variations were found (Table 6). Excluding 14 variations considered sequencing errors in the reference sequence,[3] there were 33 common variations among the four strains of 1A1, corresponding closely to the early study (Table 7).[2] After excluding the common variations among the four strains of 1A1, each strain had 5-14 variations (Table 8).

Conclusion

In this study, I developed NSAP and applied it to re-sequence *B. subtilis* 168 strains including different isolates (BGSC-1A1), and the base variations among laboratories were identified. The differences in strain 168 among laboratories were few, some strains had unique variations and may have a hidden phenotype. These variations might have been caused by differences in storage conditions in the laboratories or differences among colonies of the original stock. Based on these results, it is necessary to understand the genetic differences between the wild-type (parental) strain at each laboratory and the reference strain by re-sequencing analysis, and to pay more attention in managing laboratory strains. Also NSAP successfully demonstrated the dramatical reduction of time

for bioinformatical analysis.

Acknowledgments

Gratitude is expressed to Associate Prof. Kei Asai (Saitama University), Prof. Mitsuhiro Itaya (Keio University), President Naotake Ogasawara (Nara Institute of Science and Technology), Prof. Mitsuo Ogura (Tokai University), Prof. Tsutomu Sato (Hosei University), Prof. Junichi Sekiguchi (Shinshu University), Emeritus Prof. Kunio Yamane (University of Tsukuba), and Prof. Kenichi Yoshida (Kobe University) for providing strains. Also I am grateful to Dr. Takashi Matsumoto for supporting experiments and interpretation of variations on *Bacillus subtilis*. This study was supported by the MEXT-Supported Program for the Strategic Research Foundation at Private Universities, 2008-2012 (S0801025).

References

1. Kunst, F., Ogasawara, N., Moszer, I., et al. 1997, The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249-256.
2. Barbe, V., Cruveiller, S., Kunst, F., et al. 2009, From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology*, **155**, 1758-1775.
3. Soupene, E., van Heeswijk, W. C., Plumbridge, J., et al. 2003, Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. *J. Bacteriol.*, **185**, 5611-5626.
4. Kanesaki, Y., Shiwa, Y., Tajima, N., et al. 2012, Identification of substrain-specific mutations by massively parallel whole-genome resequencing of *Synechocystis* sp. PCC 6803. *DNA Res.*, **19**, 67-79.
5. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.
6. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
7. Chen, K., Wallis, J. W., McLellan, M. D., et al. 2009, BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677-681.
8. Srivatsan, A., Han, Y., Peng, J., et al. 2008, High-Precision, Whole-Genome Sequencing of Laboratory Strains Facilitates Genetic Studies. *PLoS Genetics*, **4**, e1000139.

9. Kunkel, B., Losick, R. and Stragier, P. 1990, The *Bacillus subtilis* gene for the development transcription factor sigma K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes Dev.*, **4**, 525-535.

10. Zeigler, DR. 2000, "Bacillus Genetic Stock Center Catalog of Strains", 7th ed., Vol. 1, Bacillus Genetic Stock Center, Columbus, OH.

11. Kobayashi, K., Ehrlich, SD., Albertini, A., et al. 2003, Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 4678-4683.

12. Margot, P. and Karamata, D. 1996, The lytE gene of *Bacillus subtilis* 168 encodes a cell wall hydrolase. *Microbiology***, 142**, 3437-3444.

13. Slack, FJ., Serror, P., Joyce, E., et al. 1995, A gene required for nutritional repression of the *Bacillus subtilis* dipeptide permease operon. *Mol. Microbiol.*, **15**, 689-702.

14. Thomaides, HB., Davison, EJ., Burston, L., et al. 2007, Essential bacterial functions encoded by gene pairs. *J. Bacteriol.,* **189**, 591-602.

15. Inaoka, T., Takahashi, K., Ohnishi-Kameyama, M., et al. 2003, Guanine nucleotides, guanosine 5'-diphosphate 3'-dipahosphate and GTP co-operatively regulate the production of an antibiotic bacilysin in *Bacillus subtilis*. *J. Biol. Chem.*, **278**, 2169-2176.

16. Wang, W., Guffanti, AA., Wei, Y., et al. 2000, Two types of *Bacillus subtilis* tetA(L) deletion strains reveal the physiological importance of TetA(L) in K+ acquisition as well as in Na+, alkali, and tetracycline resistance. *J. Bacteriol.*, **182**, 2088-2095.

17. Kunkel, B., Losick, R., and Stragier, P. 1990, The *Bacillus subtilis* gene for the development transcription factor sigma K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes Dev.*, **4**, 525-535.

18. Zeigler, DR., Prágai, Z., Rodriguez, S., et al. 2008, The origins of 168, W23, and other *Bacillus subtilis* legacy strains. *J. Bacteriol.*, **190**, 6983-6995.

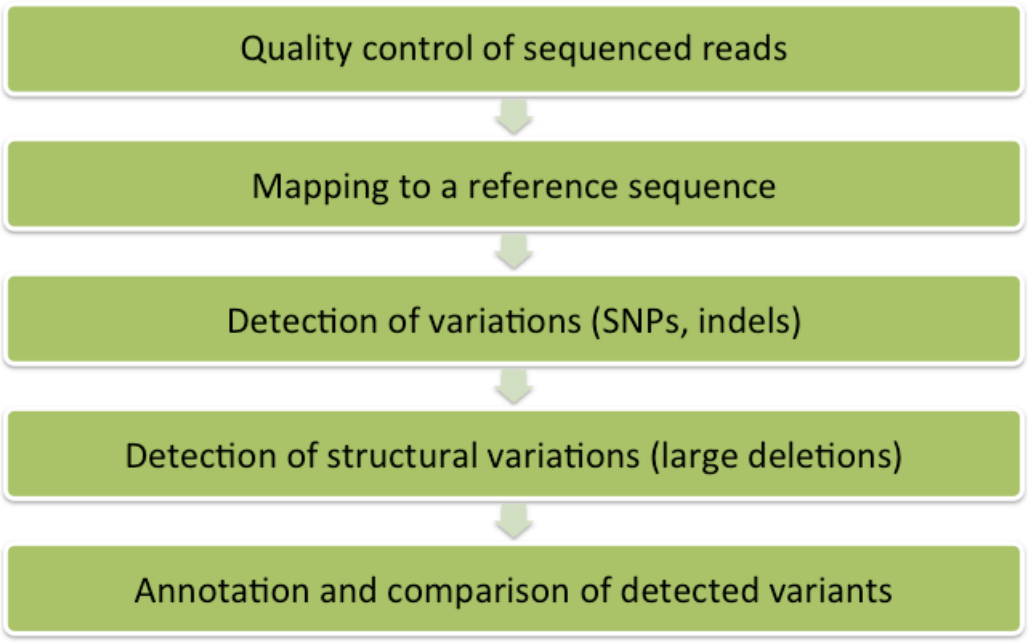

Figure 1. Workflow of bioinformatical analysis for detection of mutation

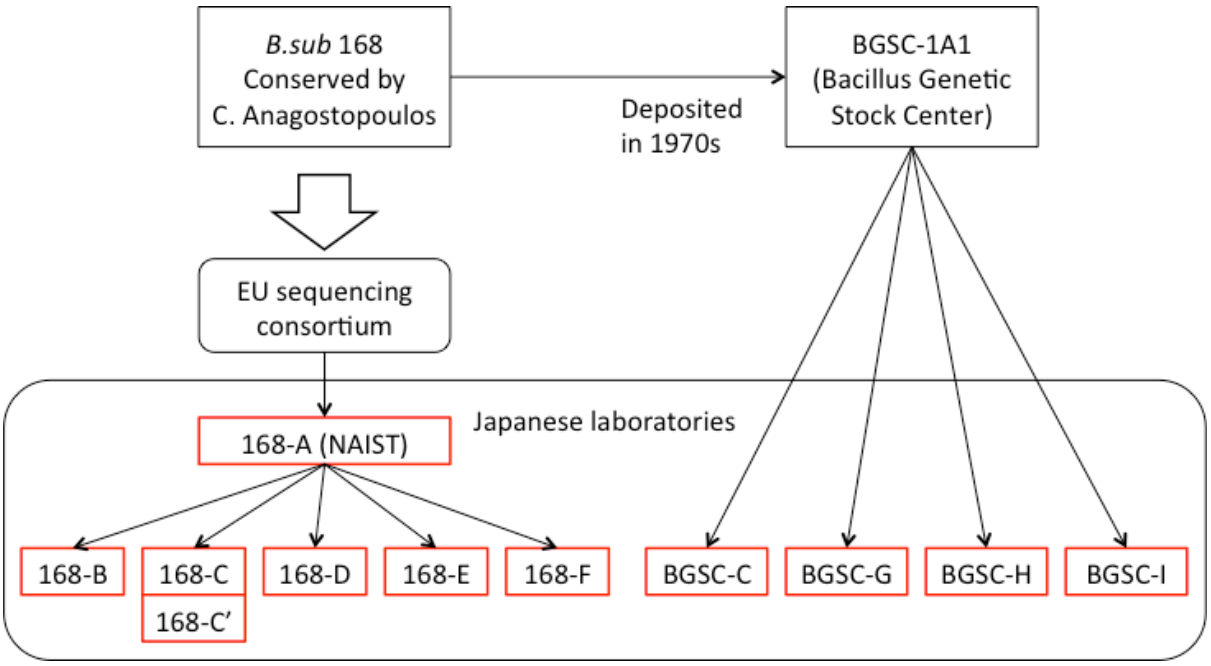

Figure 2. History of *Bacillus subtilis* 168 strain used in Japan

The *B. subtilis* 168 conserved by C. Anagostopoulos was distributed to the EU sequencing consortium in the late 1980s. This strain was distributed to the Japanese sequencing consortium via Ogasawara's laboratory at Nara Institute of Science and Technology (NAIST) in the early 1990s. Re-sequenced strains in this study are surrounded by red frames. These strains were provided by the following laboratories: 168-A, NAIST, Ogasawara's laboratory; 168-B, Kobe University, Yoshida's laboratory; 168-C and C', Saitama University, Asai's laboratory; 168-D, Hosei University, Sato's laboratory; 168-E, Tokai University, Ogura's laboratory; 168-F, Tokyo University of Agriculture, Yoshikawa's laboratory; BGSC-C, Saitama University, Asai's laboratory; BGSC-G, Shinshu University, Sekiguchi's laboratory; BGSC-H, University of Tsukuba, Yamane's laboratory; BGSC-I, Keio University, Itaya's laboratory. 168-C' is a subculture of 168-C.

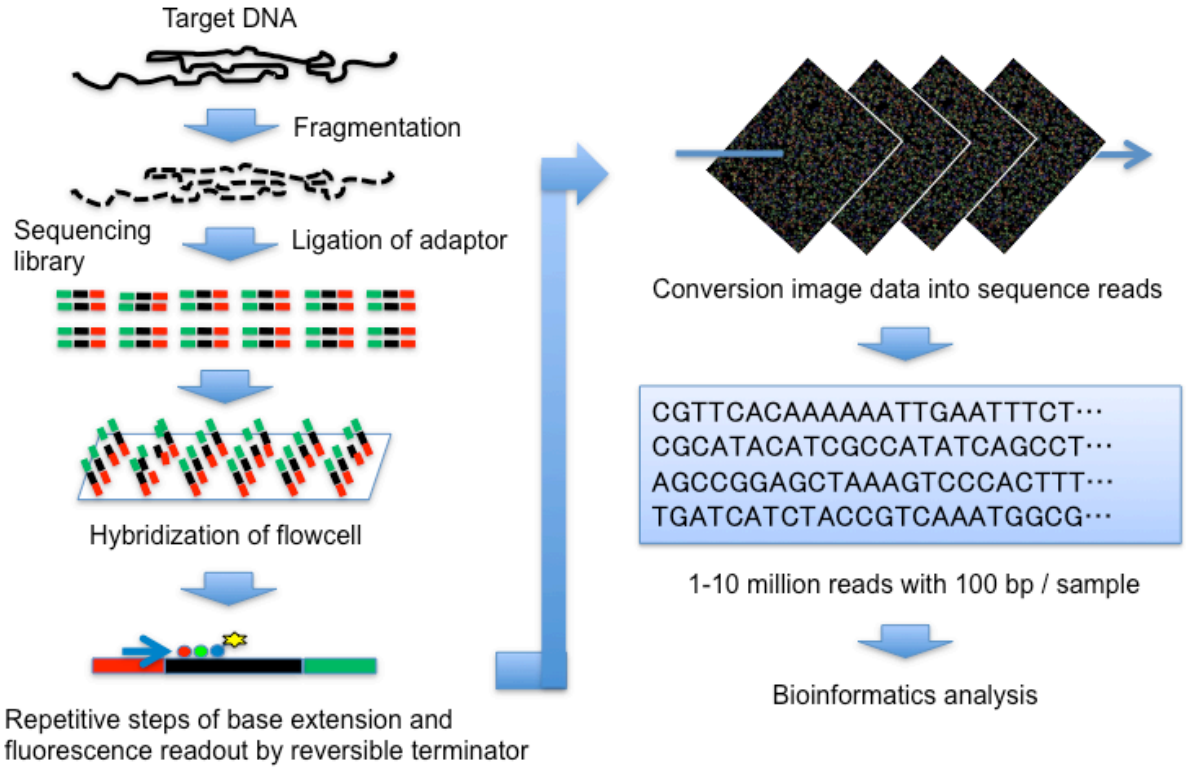

Figure 3. Workflow of re-sequencing analysis from sample preparations to sequencing on Illumina sequencer

Table 1. Summary of mapping analysis of 11 *B. subtilis* laboratory strains

Strain	Read length (bp)	Number of total reads	Total read bases (Mb)	% of mapped reads	% of properly paired	Unmapped region (bp)	% of covered genome	Genomic coverage
168-A	75	3,102,870	232	97.93	91.14	179	99.99	54
168-B	75	2,302,398	172	98.17	95.06	213	99.99	40
168-C	75	3,134,327	235	97.94	92.31	193	99.99	55
168-C'	75	3,942,439	295	98.31	95.22	173	99.99	69
168-D	75	4,146,924	311	97.36	90.85	182	99.99	72
168-E	75	4,071,253	305	97.46	91.25	160	99.99	71
168-F	75	3,985,900	298	97.48	92.27	45,304	98.85	70
BGSC-G	75	4,887,851	366	98.77	97.85	187	99.99	86
BGSC-H	91	3,392,281	308	99.04	98.62	225	99.99	73
BGSC-C	75	6,463,954	484	97.56	93.55	163	99.99	112
BGSC-I	75	4,100,055	307	98.07	92.56	178	99.99	72

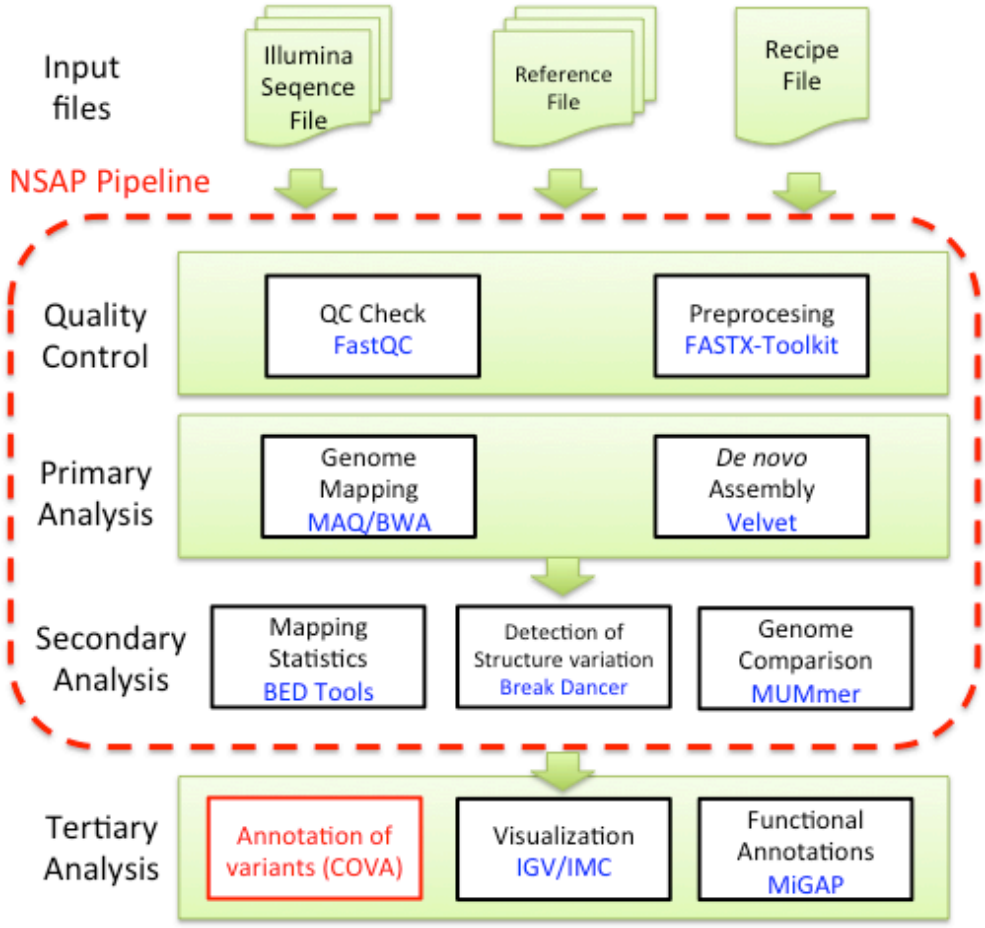

Figure 4. Architecture of NSAP

The NSAP analysis pipeline consists of three major steps: quality control of Illumina sequence files, primary analysis such as mapping analysis and *de novo* assembly, and secondary analysis such as statistics of mapping, detection of structure variation, and genome comparison of assembled sequences. NSAP pipeline (surrounded by the red broken line) includes various third-party software tools (names in blue text). NSAP takes three types of input files: Illumina sequence files, reference files used in mapping analysis, and a recipe file that describes the analysis workflow in CSV format. One software tool highlighted by red (COVA) was developed for an annotation of variants described in chapter II.

Project Dir					
/disk3/work/ChIP7942					
Ref name	Ref dir				
Syn7942	/disk3/share/ncbi_genomes/Synechococcus_elongatus_PCC_7942_uid58045				
Seq name	Read1	Read2	fastx_clipper	fastx_trimmer	OnlyRead1
Sample1	R1_filtered_t50.fastq	R2_filtered_t50.fastq	-v -l 30 -n -a GTATGCCGTCTTCTGCTTG	-v -f 1 -l 30	TRUE
Sample2	R1_filtered_t50.fastq	R2_filtered_t50.fastq	-v -l 30 -n -a GTATGCCGTCTTCTGCTTG	-v -f 1 -l 30	TRUE
Sample3	R1_filtered_t50.fastq	R2_filtered_t50.fastq	-v -l 30 -n -a GTATGCCGTCTTCTGCTTG	-v -f 1 -l 30	TRUE
Sample4	R1_filtered_t50.fastq	R2_filtered_t50.fastq	-v -l 30 -n -a GTATGCCGTCTTCTGCTTG	-v -f 1 -l 30	TRUE

Mapping	Key	Value
bwa	rmdup	FALSE
bwa	min_depth	5
bwa	num_haplotypes	1
bwa	unmap_assembly	FALSE

Figure 5. Example of a recipe file

A CSV-fomatted recipe file specifies what analysis should be done for each sample. The above example shows the recipe file for mapping analysis. ‘Project Dir’ specifies the root directory of an analysis project. ‘Ref name’ and ‘Ref dir’ specify a path to the directory including reference sequences. ‘Seq name’, ‘Read1’, and ‘Read2’ specify a name and file names of fastq files. ‘fastx_clipper’, ‘fastx_trimeer’, and ‘OnlyRead1’ columns specify parameters for preprocessing of sequence files. ‘Mapping’, ‘Key’, and ‘Value’ columns specify parameters for mapping analysis.

Infomation of input sequences		
Length of sequence(Mb):	1696	426

Result of bwa

	Sample1	Sample2
# reference sequences	1	1
Total reference sequences (bp)	4,097,429	4,097,429
Read length (bp)	100	100
# total reads	16,493,259	4,199,671
Total read bases (Mb)	1,649	419
# mapped reads	15,903,474	3,990,726
% mapped reads	96.4	95.0
% properly paired	95.0	92.1
Unmapped region (bp)	109,029	97,023
Covered genomic bases (bp) at a depth of at least 5x	3,982,145	3,988,211
% covered genomic bases (bp) at a depth of at least 5x	97.2	97.3
Genomic coverage at a depth of at least 1x	399	100

Result of BreakDancer

	Sample1	Sample2
Number of structural variants	12	0

Result of velvet

	unmap_Sample1	unmap_Sample2
Velvet hash value:	93	75
Total number of contigs:	57	120
n50:	4,094	1,563
length of longest contig:	22,939	13,362
Total bases in contigs:	110,599	109,779
Number of contigs > 1k:	23	27
Total bases in contigs > 1k:	96,569	70,885

Figure 6. Example of an analysis report

The uppermost section of this report describes a total length of sequenced base for each sample. The second section 'Result of maq' describes a statistics of mapping analysis using MAQ. The third section 'Result of BreakDancer' describes a number of detected structural variants using BreakDancer. The lowermost section 'Result of velvet' describes a statistics of *de novo* assembly using velvet.

Table 2. Primers used in this study

Position	Sequence
1317152F	ATATAGAAAACGGACAGCCCG
1317152R	ATTTAATGCTGACAATGAGAACG
2097080F	AGTATGAAGCACTGCTGGCAC
2097080R	AAGACGGAACTTATGGCGTG
2271424F	TCATATCTCCAACGTGCAATG
2271424R	ATGTCACCGGCAGAGTCC
2480646F	AGGACATTCCGATGAACCTG
2480646R	CTGAGCTTGACAACCTGCTTC
2581726F	AGAACCGCCAGCTTGTACC
2581726R	TACGACGCTTGTGGCTTATG
3770058F	ATTGCCATTAACTCCGCATC
3770058R	GGCTGGCGATATACAAGGTG
3935822F	TGGCATTCTGAGCATAGCAG
3935822R	ACAAGCTGGATTACGAAGCG
4155390F	GGTGTCTTAACCGCTTGACC
4155390R	TTGACGTCCTAGCAGGAATTG
557865F	CCTCGTCTGAGGAATCTTGG
557865R	CAATTGATTATATAAGCTGTAATCTCC

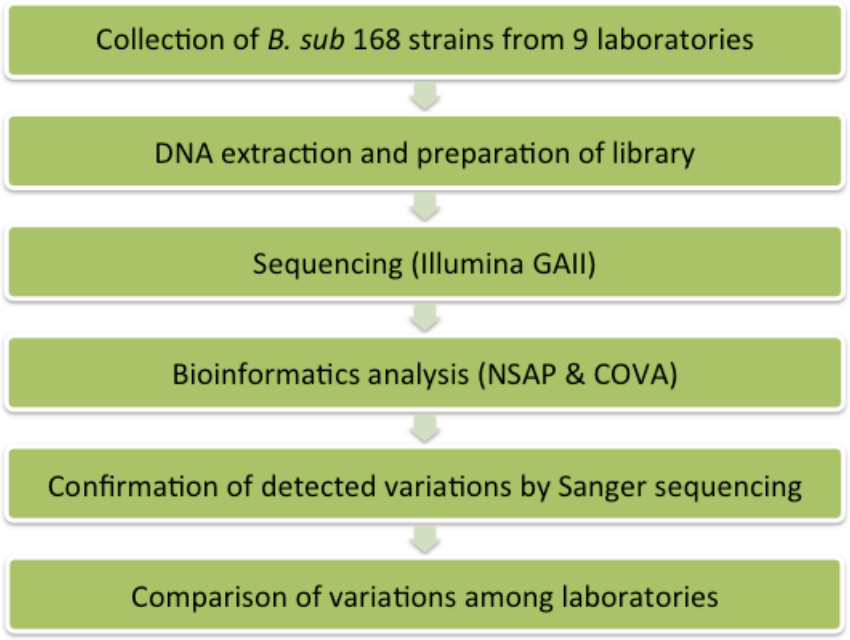

Figure 7. Workflow of this study

For the bioinformatical analysis, I used two software tools which were developed by myself, NSAP and COVA. COVA was designed for an annotation of variations described in chapter II.

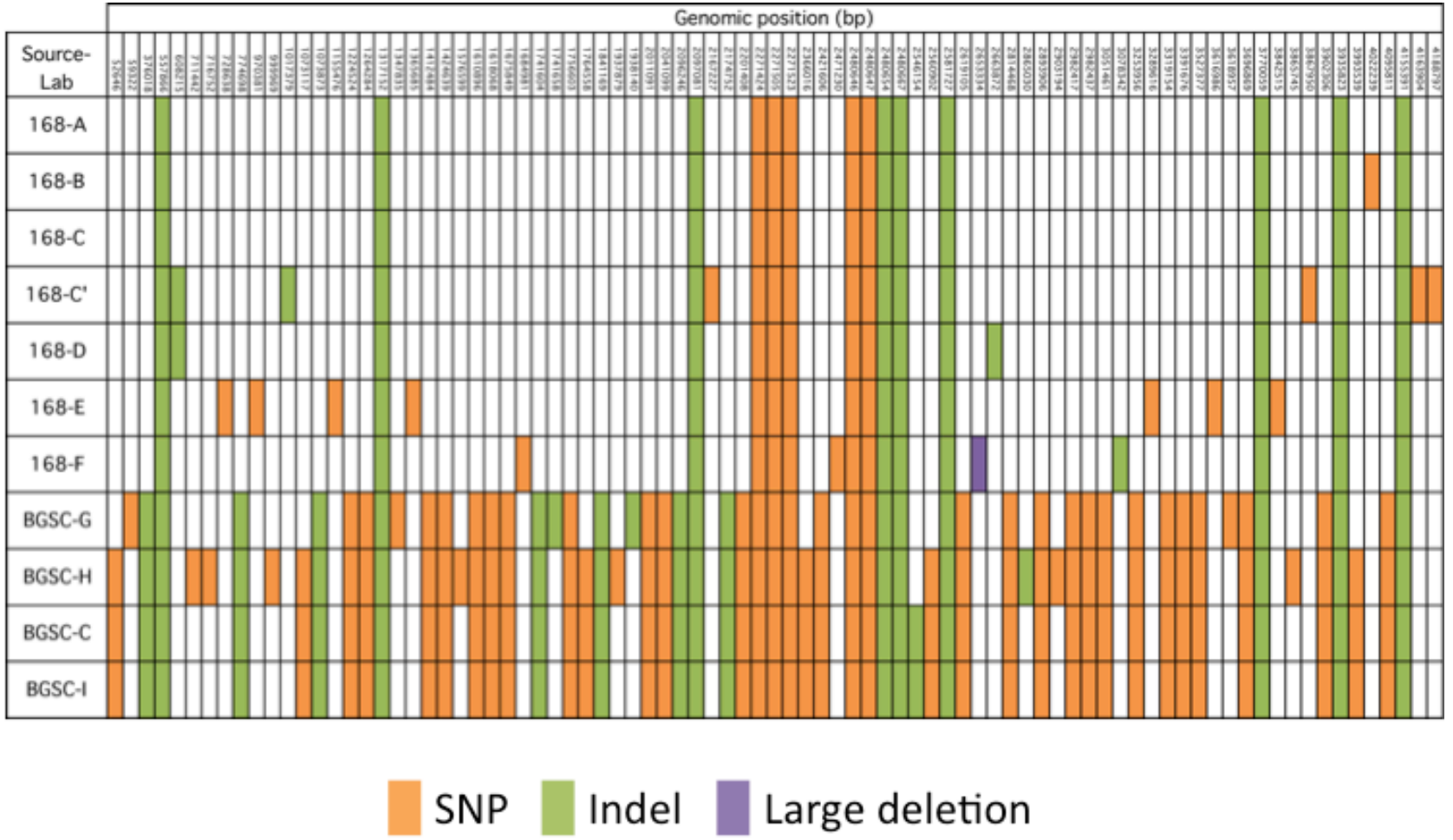

Figure 8. Comparison of variations among the 11 *B. subtilis* 168 strains used in Japan.

11 *B. subtilis* 168 strains from nine laboratories (A-H), including BGSC stock strains (BGSC-1A1), were sequenced. The genomic positions of identified variations against the *B. subtilis* 168 published genome[2] are shown.

Table 3. Numbers of variations in 7 *B. subtilis* 168 strains compared to the 168 reference sequence

Variation type	168-A	168-B	168-C	168-C'	168-D	168-E	168-F
SNP	5	6	5	9	5	12	8
Insertion	6	6	6	7	7	6	6
Deletion	3	3	3	4	4	3	4*
Total	14	15	14	20	16	21	18
Intergenic	11	11	11	12	12	12	12
Synonymous	2	2	2	3	2	5	18
Non-synonymous	1	2	1	5	2	4	3

All strains were compared to the 168 reference sequence.[3] Numbers in parentheses show differences between a given strain and 168-A.

* Include one large deletion.

Table 4. Base difference between 168-A strain and the published sequence

Position	Reference base	Variant base	Gene	Amino acid change	Product
557866	-	+T			
1317152	-	+CTT *			
2097081	-	+A			
2271424	T	C	*uvrX*	Silent	lesion bypass phage DNA polymerase
2271505	C	T	*uvrX*	Silent	lesion bypass phage DNA polymerase
2271523	A	C	*uvrX*	D45E	lesion bypass phage DNA polymerase
2480646	T	A			
2480647	A	T			
2480654	-	-T			
2480667	-	-T			
2581727	-	+T			
3770059	-	+A			
3935823	-	-T			
4155391	-	+A			

All strains were compared to the 168 reference sequence.[3] *BWA and SAMtools called this variation as +GT, but the sangar sequencing showed +CTT variation.

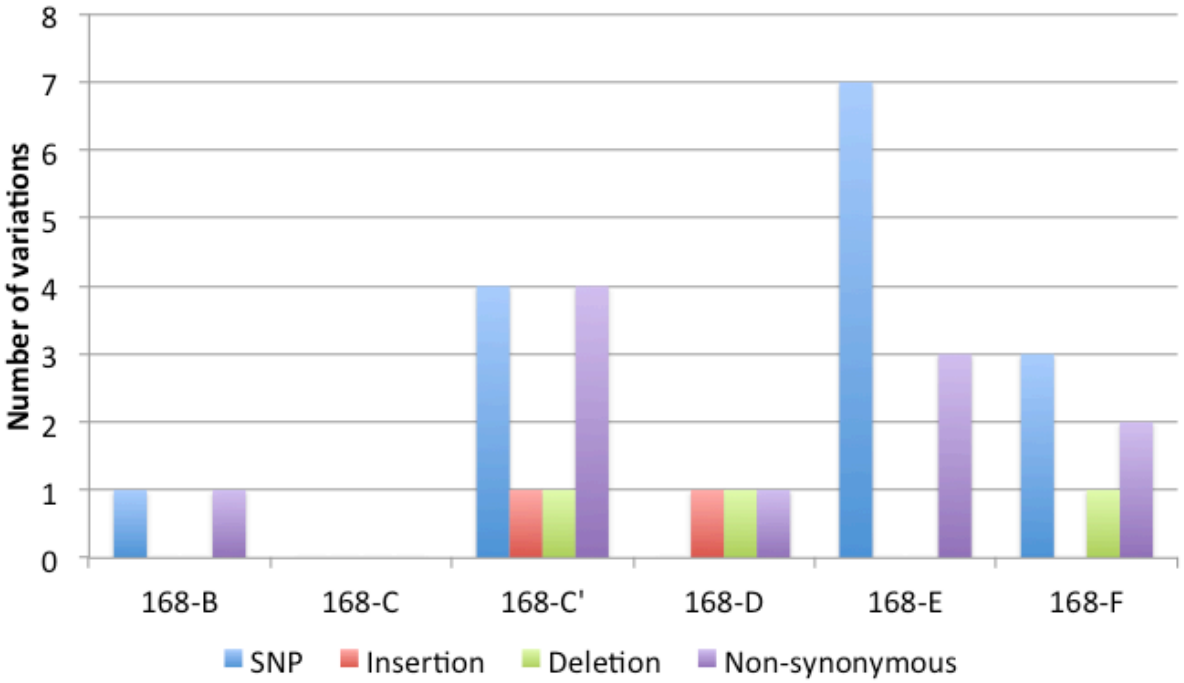

Figure 9. Numbers of laboratory specific variations in 6 *B. subtilis* 168 strains compared to the 168-A strain. All strains were compared to the 168 reference sequence.[3] Variations were categorized as snp, insertion, and deletion. When they changed amino acid, they were also counted as a non-synonymous variation.

Table 5. Base difference between 168-A strain and other 168 strains

Position	Reference base	Variant base	Gene	Amino acid change	Product	Strains
608215	-	+A				168-C', 168-D
728638	G	A				168-E
970381	A	G	*cspR*	T83A	putative rRNA methylase	168-E
1017379	-	-T	*phoA*	Freamshift	alkaline phosphatase A	168-C'
1155476	G	A	*wprA*	G563D	cell wall-associated protease	168-E
1365685	C	T	*dppE*	A512V	dipeptide ABC transporter (dipeptide-binding lipoprotein)	168-E
1684981	A	G	*topA*	T441A	DNA topoisomerase I	168-F
2167227	C	A	*yosH*	D95Y	conserved hypothetical protein; phage SPbeta	168-C'
2471230	C	A	*yqjQ*	Silent	putative metabolite dehydrogenase, NAD-binding	168-F
2653334	-	Large deletion				168-F
2663872	-	-T	*yqxJ*	Freamshift	hypothetical protein; skin element	168-D
3078342	-	-A	*opuD*	Freamshift	glycine betaine transporter	168-F
3289616	C	T	*dhbE*	Silent	2,3-dihydroxybenzoate-AMP ligase (enterobactin synthetase component E)	168-E
3616986	C	T	*yvkC*	Silent	putative phosphotransferase	168-E
3842515	A	G	*albG*	Silent	putative integral inner membrane protein involved in subtilosin production and immunity	168-E
3867950	C	T	*ywfH*	T153I	carrier protein reductase of bacilysin synthesis	168-C'
4022239	G	A	*yxzC*	T9I	putative nucleic acid binding protein	168-B
4163904	C	A	*yybT*	Silent	putative phosphodiesterase	168-C'
4188797	C	T	*tetB*	M87I	multifunctional tetracycline-metal/H+ antiporter and Na+(K+)/H+ antiporter	168-C'

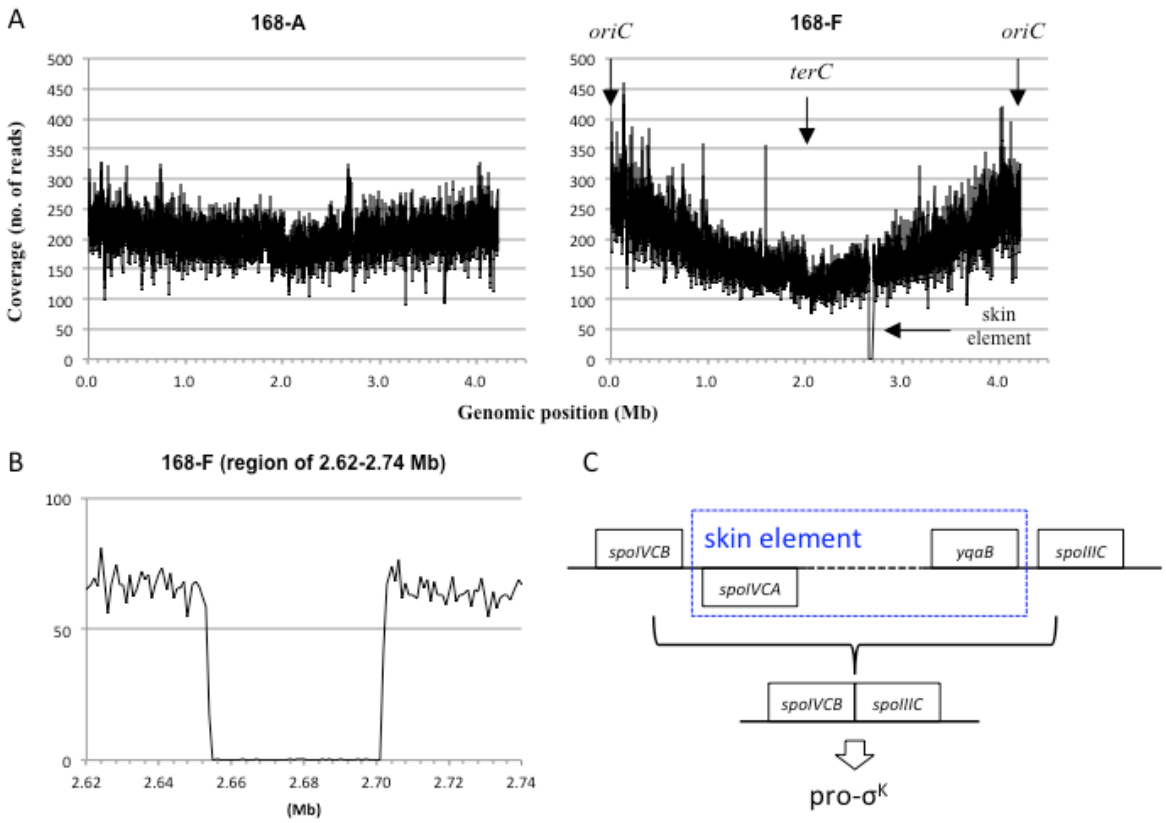

Figure 10. The number of sequencing reads at each genomic position for each strain.

A) Coverage was obtained by calculating the average number of sequence reads in a given 1-kb window. B) A magnification of deleted region in 168-F. C) This deleted region includes skin element. The SKIN element is excised from the genome in the developmental phases. Two separated genes (*spoIVCB* and *spoIIIC*) are fused and act as a *sigK* gene that turns on sporulation genes.

Table 6. Numbers of variations in four *B. subtilis* 168 (BGSC-1A1) strains compared to the 168 reference sequence

Variation type	BGSC-C	BGSC-I	BGSC-G	BGSC-H
SNP	36	36	33	43
Insertion	7	7	7	7
Deletion	10	10	11	10
Total	53	53	51	60
Intergenic	23	23	21	23
Synonymous	8	8	9	10
Non-synonymous	23	23	21	23

All strains were compared to the 168 reference sequence.[3]

Table 7. Base difference between BGSC-1A1 strain and the 168 published sequence

Position	Reference base	Variant base	Gene	Amino acid change	Product
376018	-	-G			
774698	-	-A	*yesY*	Freamshift	rhamnogalacturonan acetylesterase
1073873	-	-T			
1224524	T	G	*oppD*	V357G	oligopeptide ABC transporter (ATP-binding protein)
1264284	T	A	*yjcM*	K216N	hypothetical protein
1412484	T	G	*sigI*	L198R	putative RNA polymerase sigma factor SigI
1424639	T	G			
1610896	T	C	*sepF*	M11T	cell division machinery factor
1618068	C	T	*rluD*	P287S	pseudouridylate synthase
1675849	C	T	*trmD*	H227Y	tRNA (guanine-*N*(1)-)-methyltransferase
1741604	-	-G			
1756603	G	A	*ymfD*	A319T	putative multidrug resistance protein
1841169	-	+A	*pksN*	Freamshift	polyketide synthase of type I
2011091	G	A	*gltA*	A1181V	glutamate synthase (large subunit)
2041099	A	G	*yozT*	Silent	hypothetical protein
2096246	-	-A			
2174752	-	-CT			
2201408	A	G	*yoqA*	L23P	hypothetical protein; phage SPbeta
2421606	T	C	*rluB*	Silent	pseudouridine synthase
2619105	C	T	*yqeZ*	Silent	putative membrane bound hydrolase
2814468	C	T			
2814468	C	T			
2893906	G	A	*ilvC*	Silent	ketol-acid reductoisomerase
2982417	T	A			
2982437	T	C			
3051461	G	A	*ytpS*	P375S	putative DNA translocase stage III sporulation protein (modular protein)
3253956	T	C	*comP*	E628G	two-component sensor histidine kinase
3319154	C	T	*yutE*	A37T	hypothetical protein
3391676	A	G	*gerAA*	T299A	component of the GerA germination receptor
3527377	G	A	*epsC*	A276V	putative UDP-sugar epimerase
3696869	T	C	*pgdS*	Silent	gamma-*DL*-glutamyl hydrolase (PGA depolymerase)
3902306	C	A	*sacA*	L448F	sucrase-6-phosphate hydrolase
4095811	C	T	*yxbD*	V9I	putative acetyltransferase

All strains were compared to the 168 reference sequence.[3]

Table 8. Uncommon bases in four BGSC-1A1 strains

Position	Reference base	Variant base	Gene	Amino acid change	Product	Strains
52646	C	T				BGSC-C, BGSC-I, BGSC-H
59322	G	A	*ctc*	Silent	50S ribosomal protein L25/general stress protein Ctc	BGSC-G
711442	C	G				BGSC-H
716752	G	C				BGSC-H
999969	C	G				BGSC-H
1073117	C	A	*hpr*	V201L	transcriptional regulator Hpr	BGSC-C, BGSC-I, BGSC-H
1347835	C	T	*xlyA*	Silent	bacteriophage PBSX *N*-acetylmuramoyl-L-alanine amidase	BGSC-G
1576599	A	G	*ylbO*	Silent	putative spore coat protein regulator protein YlbO	BGSC-H
1741658	-	-T	*ylxY*	Freamshift	putative sugar deacetylase	BGSC-G
1764558	C	T				BGSC-C, BGSC-I, BGSC-H
1937879	G	C				BGSC-H
1938140	-	-T	*ynfC*	Freamshift	hypothetical protein	BGSC-G
2366016	T	C	*ypiB*	H87R	hypothetical protein	BGSC-C, BGSC-I, BGSC-H
2546154	-	-C	*gcvPB*	Freamshift	glycine dehydrogenase subunit 2	BGSC-C, BGSC-I
2560902	C	T	*yqxL*	E181K	putative CorA-type Mg(2+) transporter	BGSC-C, BGSC-I, BGSC-H
2865030	-	-AATTC	*comC*	Freamshift	membrane protease and transmethylase	BGSC-H
2903194	G	C				BGSC-H
3618957	G	A				BGSC-G
3865745	A	G	*pta*	Silent	phosphotransacetylase	BGSC-H
3993539	G	T	*yxjM*	Silent	two-component sensor histidine kinase [YxjL]	BGSC-C, BGSC-I, BGSC-H

All strains were compared to the 168 reference sequence.[3]

CHAPTER II

Development of a variant annotation software and its application to re-sequencing of the mutagenized yeast

CHAPTER II

Development of a variant annotation software and its application to re-sequencing of the mutagenized yeast

Abstract

A novel mutagenesis technique using error-prone DNA polymerase δ (*polδ*), based on the disparity mutagenesis model of evolution, has been successfully employed to generate novel microorganism strains with desired traits. However, little else is known about the spectra of mutagenic effects caused by disparity mutagenesis. In this study, the performance of the *polδMKII* mutator, which expresses the proofreading-deficient and low-fidelity *polδ*, was evaluated in *Saccharomyces cerevisiae* haploid strain. It was compared with that of the commonly used chemical mutagen ethyl methanesulfonate (EMS). This mutator strain possesses exogenous mutant *polδ* supplied from a plasmid, thereby leaving the genomic one intact. The mutation rate achieved by each mutagen was measured and high-throughput next generation sequencing with Illumina GAII was performed. To analyze the genome-wide mutation spectra produced by the 2 mutagenesis methods, I developed a variant annotation software, COVA (comparison of variants and functional annotation). The mutation frequency of the mutator was approximately 7 times higher than that of EMS. The strong G/C to A/T transition bias of EMS was confirmed, whereas it was found that the mutator mainly produces transversions, giving rise to more diverse amino acid substitution patterns. This study demonstrated that a proofreading-deficient and low-fidelity *polδMKII* mutator is a useful and efficient method for rapid strain improvement based on *in vivo* mutagenesis. Also COVA has successfully reduced the time for procedure of variant annotation, compared with manual annotation.

Introduction

Random mutagenesis is a powerful tool for generating enzymes, proteins, metabolic pathways, or even entire genomes with desired or improved properties.[1] Due to the technical simplicity and applicability to almost any organism, chemical or radiation mutagenesis is frequently used for the generation of genetic variability in a microorganism. However, these methods tend to be inefficient, because they can cause substantial cell damage when performed *in vivo*.[2]

A novel mutagenesis technique using error-prone DNA polymerase δ (*polδ*), based on the disparity mutagenesis model of evolution,[3] has been successfully employed to generate novel microorganism strains with desired traits.[4-11] In the disparity model, mutations occur preferentially on the lagging strand, due to the more complex, discontinuous DNA replication that takes place there (Fig. 1). Computer simulation shows that the disparity model accumulates more mutations than the parity model, in which mutations occur stochastically and evenly in both strands.[3] In addition, the disparity model produces greater diversity because some offspring will have mutant DNA while some offspring will have non-mutated, wild-type DNA.

Several studies have shown that the disparity mutagenesis method often achieved more satisfactory results (i.e., higher mutation rate and quick attainment of the desired phenotype) than conventional methods such as the chemical mutagen, ethyl methanesulfonate (EMS),[5,10] which is known to produce mainly G/C to A/T transitions.[12] However, little else is known about the spectra of mutagenic effects caused by disparity mutagenesis.

polδ is involved in the synthesis of the lagging strand of DNA.[13] Several mutants, including the proofreading-deficient *pol3-01* strain and several low fidelity mutants, have been shown to elevate the mutation rate.[14-18] To generate the strains with the greatest mutagenicity, Neo-Morgan Laboratory (Kanagawa, Japan) has developed the plasmid YCplac33/*polδMKII*,

expressing the *polδ* mutant allele with 2 mutations: one mutation to inactivate the proofreading activity (D321A and E323A)[15] and another mutation to decrease the fidelity of replication (L612M).[14,17,18]

With the recent advent of next-generation sequencing technologies, an accurate characterization of the mutant genome, relative to the parental reference strain, is now achievable. In fact, Flibotte et al. have analyzed the mutation spectra induced by various mutagens, such as EMS, ENU, and UV/TMP, in the whole genome of *Caenorhabditis elegans*.[12] Another group has also used these sequencing technologies to analyze the genetic variations between a parental and EMS-mutagenized strain of yeast.[19]

To evaluate the effect of detected mutations on genome, it needs a variant annotation such as a substitution of amino acid. Although there have been dozens of software for sequencing reads alignment and variant identification, only few software for functional annotation of variants for bacteria are available, and few tools can compare variants among multiple samples. Thus, I developed the software COVA (comparison of variants and functional annotation). COVA can annotate variations such as a single base substitution, an insertion, a deletion, and a structural variation. In addition, COVA also can compare variants among samples, which can help to pinpoint effective variant(s).

In this study, I applied COVA to evaluate the performance of the *polδMKII* mutator, which expresses the proofreading-deficient and low-fidelity *polδ*, in *Saccharomyces cerevisiae* haploid strain. It was compared with that of the commonly used chemical mutagen ethyl methanesulfonate (EMS). This mutator strain possesses exogenous mutant *polδ* supplied from a plasmid, thereby leaving the genomic one intact. The mutation rate of this mutator strain was measured and it was found that the mutation frequency of *polδMKII* was approximately 7 times higher than that of EMS. Also

high-throughput next generation sequencing with Illumina GAII was performed to analyze the genome-wide mutation spectra produced by the 2 different mutagenesis methods. It was uncovered that the mutator strain exhibited more pleiotropy and gave rise to more diverse amino acid substitution patterns. This study has demonstrated that a proofreading-deficient and low-fidelity *polδMKII* mutator is a useful and efficient method for rapid strain improvement based on *in vivo* mutagenesis. This mutator is also useful for studying the acceleration of evolution.

Materials and Methods

Construction of plasmid vectors

Plasmid YCplac33/*polδMKII* was constructed as follows: a 4.8 kb DNA fragment containing the *S. cerevisiae* BY2961 *pol3* gene, plus the UTR 1 kb upstream and 0.5 kb downstream, (*Matα ura3-52, his3-Δ300, trp1-Δ901, leu2-3, 112 lys2-801, ade2-2*) was inserted into the *Sal*I-*Eco*RI site of YCplac33, and 3 amino acid substitutions, D321A, E323A, and L612M, were introduced into the *pol3* gene using site-directed mutagenesis.[20] YCplac33 is low-copy number plasmid and is stably maintained in *S. cerevisiae*.[20]

Mutagenesis with mutator

YCplac33/*polδMKII* vector (and YCplac33 empty vector as non-mutator control) was introduced into *S. cerevisiae* BY2961 strain cells using the LiCl method, and the transformants (mutator strains) were selected on synthetic complete (SC)-agar plates without uracil. Five mutator strains were picked and independently cultivated in 1 ml SC medium at 30ºC for 24 h (about 30 generations) in order to introduce mutations into their chromosomes. To determine the mutation frequencies of the 5 mutator strains, aliquots were spread on SC-agar plates containing L-canavanine

sulfate salt (0.06 mg/ml) (Sigma, St. Louis, MO, USA) to identify *CAN1* mutants and incubated until resistant colonies were formed. The mutation frequencies were calculated as the number of drug resistant colonies divided by the number of colonies on SC-agar plate without drug. Forward mutation rates at *CAN1* were determined by fluctuation analysis using these 5 independent cultures.[21] In order to fix mutations, another aliquot of the mutator culture was spread on SC-agar plates containing 5-fluoroorotic acid monohydrate (Wako) to obtain de-mutatorized cells curing from YCplac33/*polδMKII* vector. The genomic DNA was prepared from the de-mutatorized cells using the procedure described in the following section.

Mutagenesis with EMS

S. cerevisiae BY2961 strain cells were suspended in 0.1 M phosphate-buffered saline (PBS) (pH 7.0) containing 1.5, 2.0, 2.5, or 3.0% ethyl methanesulfonate (EMS), and were incubated at 30ºC for 1 h to introduce chromosomal mutations. The cells were washed 3 times with 5% sodium thiosulfate, suspended in sterilized water, and spread on SC-agar plates containing L-canavanine sulfate salt (0.06 mg/ml) (Sigma) to identify *CAN1* mutants. The mutation frequencies were calculated as described above. Another aliquot of the EMS-treated cell suspension was spread on a YPD-agar plate to isolate single clones. The genomic DNA was prepared from 5 single clones derived from the cells treated with 1.5% EMS using the procedure described in the following section.

Library preparation and sequencing with Illumina Genome AnalyzerII

The genomic DNA from *S. cerevisiae* was extracted using the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA, USA). Each sequenced sample was prepared according to the Illumina protocols. Briefly, 3 μg of genomic DNA was fragmented to an average length of 200 bp by using

the Covaris S2 system (Covaris, Woburn, MA, USA). The fragmented DNA was repaired, a single 'A' nucleotide was ligated to the 3′ end, Illumina Index PE adapters (Illumina, San Diego, CA, USA) were ligated to the fragments, and the sample was size-selected for a 300 bp product using E-Gel SizeSelect 2% (Invitrogen, Grand Island, NY). The size-selected product was amplified by 18 cycles of PCR with the primers InPE1.0, InPE2.0, and the Index primer containing 6-nt barcodes (Illumina). The final product was validated using the Agilent Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA).

The 11 barcoded libraries (the parental strain BY2961, 5 colonies from the mutator strain, and 5 colonies from the EMS-treated strain) were used for cluster generation in several multiplexed flowcell lanes in the Illumina Genome Analyzer II system. Ninety-one cycles of multiplexed paired-end sequencing was performed, running phi X 174 genomic DNA as a control in a separate lane of the flow cell. After the sequencing reactions were complete, Illumina analysis pipeline (CASAVA 1.6.0) was used to carry out image analysis, base calling, and quality score calibration. Reads were sorted by barcode and exported in the FASTQ format.

Sequencing and Bioinformatics analysis

Sequence reads from each sample were analyzed using NSAP. In NSAP at first, quality control of sequence reads were performed. Once the raw sequence data were curated, the reads of each sample were aligned to the S288c reference genome (http://www.yeastgenome.org/) using the BWA software (Ver. 0.5.1) with default parameters.[22] To avoid false positives and mutations from repetitive regions, I removed repetitive reads from the alignment files. I then used the SAMtools software (Ver. 0.1.9)[23] to produce the lists of mutations. To identify mutations that were produced by mutagenesis, I applied the following filtering criteria to the lists of mutations: (a) The coverage at the

mismatch positions should be at least 10; (b) The variant is not present in the sequenced parental strain; (c) Indels meet a SNP quality threshold of 50 and substitutions meet a SNP quality threshold of 20 (SAMtools assigns SNP quality, which is the Phred-scaled probability that the consensus is identical to the reference); (d) Samples meet a mapping quality of 30 (SAMtools assigns Mapping quality, which is the Phred-scaled probability that the read alignment is wrong); (e) The percentage of reads showing the variant allele exceeds 90%. A variant must pass this filter to be considered a mutation. Alignments of all mutations were inspected by Integrative Genomics Viewer (IGV).[24]

Implementation of COVA and annotation of mutations

For annotation of mutations, I have developed the software, COVA (comparison of variants and functional annotation). The lists of mutations detected in this study were annotated using COVA. COVA is a Ruby-based tool for variant comparison and functional annotation, especially useful for bacterial mutation analysis. Workflow of COVA is illustrated in Fig. 2. COVA can annotate single nucleotide variants (SNVs), insertions/deletions (InDels) and other types of variants such as structural variations and coverage of genes. COVA is freely available at http://sourceforge.net/projects/cova/

The input file of variant list is obtained after step of variant calling. COVA can annotate the following variant file format: SAMtools-pileup, VCF, MAQ, BreakDancer, and GFF3 formatted coverage of gene generated by coverageBed. COVA can utilize annotation data sets conforming to Genbank Format which is easily downloadable from NCBI website.

COVA generates five types of output table files. The first file contains annotation for all variants (SNP/InDel), such as type and probability of variant, mutational spectrum, and its effect on the gene. The second file is a comparison table of variants (SNP/InDel) among multiple samples,

which helps to pinpoint causal variation(s) relating to phenotype (Fig. 3). The third file is a comparison table of structural variants among multiple samples, which helps to pinpoint causal variation(s) relating to phenotype. The forth file is a comparison table of gene coverage among multiple samples, which helps to find deletion of genes. The fifth file is a summary table of number of variants in each sample.

Accession codes

The raw reads used in this study are available on the DDBJ Sequence Read Archive (DRA) under accession number DRA000522.

Results

Comparison of mutation frequencies

In this study, the performance of the *polδMKII* mutator, which expresses the proofreading-deficient and low-fidelity *polδ*, in the *S. cerevisiae* haploid strain, was compared with that of the commonly used chemical mutagen EMS. The workflow of this study is illustrated in Fig. 4. To assess EMS efficiency, *S. cerevisiae* BY2961 cells were treated with different concentrations of EMS. The lethality and mutation frequencies of the canavanine resistant colonies are shown in Table 1. At an EMS concentration of 1.5%, the mutation frequency was approximately 18-fold higher than that in the control (untreated) strain. Above 2.0% EMS, the survival rate decreased with no increase in mutation frequency. Based on this result, cells treated with 1.5% EMS were used for whole-genome sequencing.

To assess the effectiveness of the mutator, the haploid BY2961 strain was transformed with a yeast expression plasmid, YCplac33/*polδMKII*, expressing the *polδ* mutant allele containing both the

mutation to inactivate the proofreading activity (D321A and E323A) and the mutation to decrease the fidelity of replication (L612M). The mutator strain harboring the YCplac33/*polδMKII* plasmid will be referred to from here on as "mutator." The mutation frequency by resistance to canavanine was determined. As summarized in Table 2, the mutation frequency of the mutator was approximately 132-fold higher than in the cells containing the empty vector. The forward mutation rate at the *CAN1* (arginine permease) locus was calculated to be 7.9×10^{-6} /cell division. These results show that the plasmid-generated mutated *polδ* protein effectively competes with the endogenous wild-type *polδ* protein that is produced from the chromosome, and the semi-dominant negative expression of mutated *polδ* was effective in introducing mutations. These results also demonstrate that the mutation frequency of the mutator was approximately 7 times higher than that of EMS.

Comparison of number of mutations introduced by mutagenesis

To analyze the genome-wide mutation spectra of the 2 different mutagenesis methods, I implemented a parallel sequencing approach with the Illumina Solexa technology (GAII instrument). I sequenced the parental haploid strain BY2961, each of the 5 clones from the mutator strains, and each of the 5 clones from the EMS-treated strains under non-selective conditions. Using NSAP developed in chapter I, sequencing reads were aligned to the S288c reference genome using the BWA software.[22] To avoid false positives due to mutations from repetitive regions, reads mapped to multiple locations were discarded and only uniquely mapped reads were used for subsequent analysis.

In the current study, the average genomic coverage ranged from 32× to 87× (Table 3). On average, 94.18% of the S288c reference genome was covered with at least 1 uniquely mapped read at each base. Subsequently, I analyzed the data for 2 kinds of mutational events: single nucleotide variants (SNVs), and small insertions and deletions (Indels). Illumina sequencing found 6,766

genetic differences between our parental strain BY2961 and the S288c. Mutations induced by these mutagens were identified by subtracting the parental mutations and analyzed using COVA. Sequence processing details can be found in the Materials and Methods.

Comparison of the mutational spectra introduced by mutagenesis

I compared the average number of mutations between mutator strains and EMS-mutagenized strains (Fig. 5). Mutator produced fewer SNVs than EMS (7.2 versus 55.8 per strain, respectively, $p < 0.05$). Mutator and EMS produced few deletions (1.6 versus 2.8 per strain, respectively), as well as few insertions (0.2 versus 0.6 per strain, respectively). An average of 1.14×10^7 nucleotide sites fulfilled our criteria of read depth (≥10), with an average base-substitutional mutation rate estimate of EMS: 4.87 (SE = 1.34) $\times 10^{-6}$ per site, Mutator: 2.09 (0.55) $\times 10^{-8}$ per site per cell division (about 30 generations). The rate calculated for the mutator is 100-fold higher than the previously reported spontaneous mutation rate, 3.3 (0.8) $\times 10^{-10}$, based on 454 analyses of 4 mutation-accumulation (MA)-lines.[25] The 2 mutagens generate mutations that are distributed similarly across the various gene features, although the mutator did produce more SNVs within exons than did EMS (Fig. 6).

The mutation spectra are shown in Fig. 7A. In the genome-wide profile, it was found that the mutator primarily induced transversions (72%), while EMS primarily induced transitions (97%), well in accord with the known mutagenic specificity of EMS.[12] Similarly, the mutator primarily induced transversions (69%) in the non-synonymous substitutions in exons (Fig. 7B), similar to what has been seen in *pol3-01* study using *URA3* reporter gene.[16] EMS treatment was also in agreement with the genome-wide spectra, induced transitions with a prevalence of 98%.

Comparison of amino acid substitution patterns

The mutation spectra of a given mutagenesis method influences the repertoire of changed amino acids at the protein level and it was able to evaluate the amino acid substitution patterns generated by our 2 protocols (Table 4). Initially, mutations were classified into those that preserved the corresponding amino acid, changed the amino acid, or generated a stop codon. A clear difference was seen between mutator and EMS. Of the total mutations, the mutator changed the amino acid in approximately 85%, whereas EMS changed the amino acid in approximately 61%. The mutator also generated more stop codons than EMS (7% versus 2%, respectively). While mutator generated more changes to the first or second nucleotide of the codon, EMS generated changes in all 3 positions in approximately equal proportions.

Amino acid changes were classified into conservative and nonconservative substitutions, where a conservative substitution changed the encoded amino acid to a similar amino acid according to the criteria of the BLOSUM62 matrix.[26] Of the amino acid changes, mutator produced more nonconservative substitutions than EMS (83% and 53%). For the comparison of random mutagenesis methods, Wong et al.[27] proposed a useful structure indicator that takes into account Gly and Pro substitutions as well as stop codons. In our study, the mutator produced an equivalent number of Gly/Pro and stop codon substitutions, whereas EMS generated only stop codon substitutions.

Discussion

In this study, the performance of a novel mutagenesis technique using error-prone proofreading-deficient and low-fidelity DNA polymerase δ was evaluated by determining the mutation rate of the strain harboring the enzyme. I also analyzed the spectra of mutations across the

entire *S. cerevisiae* genome using COVA developed in this study. Then I assessed the diversity of mutation types at the amino acid level.

Mutational rate of pol*δ*MKII *mutator*

Proofreading-deficient *polδ* mutants, such as *pol3-01* strain, and several low-fidelity *polδ* mutants, such as L612M, have been shown to present a mutator phenotype and to elevate the mutation rate.[14-18] A BY2961 strain expressing a *polδMKII* mutator was generated. This mutator has *polδ* mutant allele containing a combination of mutations to inactivate the proofreading activity (D321A and E323A) and to decrease the fidelity of replication (L612M). This mutant allele acts as a strong mutator, as evidenced by the high frequency of spontaneous mutations (131-fold over control, compared to 18-fold for EMS strains). Vencatesan et al. reported the forward *CAN1* mutation rates of *polδ* mutants as 1.5×10^{-6} in L612M, and 5.6×10^{-6} in *pol3-01*.[18] These mutant strains were constructed by integrating the *pol3-01* or *pol*3-L612M allele into the chromosomal *POL3* gene by targeted integration, thereby disrupting the endogenous *POL3* gene. In contrast, our mutator plasmid expressing the *polδ* mutant allele produced a mutation rate of 7.9×10^{-6}, which shows a high mutation rate as well as chromosomal integration. The use of the *polδMKII* mutator plasmid allows the continued expression of the endogenous wild-type *POL3* and provides for an efficient restoration of the wild-type mutation rate by curing the yeast strains of the mutator plasmid. Once the desired trait(s) has been selected, curing the cells from the mutator plasmid can stabilize the newly obtained phenotype.

Mutational spectra of polδMKII *mutator*

In general, all random mutagenesis methods developed to date are biased toward

transition mutations, although efforts have been made to overcome this.[28] While transition bias was observed in EMS, transversion bias with the mutator was actually observed (Fig. 7A). Because of this, the mutator yielded a broader spectrum of nucleotide changes across the entire genome. The mutator was also biased toward transversions in the non-synonymous substitutions (Fig. 7B). For EMS, the spectrum of mutation events observed is similar to what has been reported by others.[12]

At the protein level, the amino acid substitution pattern differed between the mutator and EMS (Table 4). Mutations generated by the mutator resulted in amino acid substitutions more often than did mutations generated by EMS (85% versus 61%, respectively). Most of the substitutions made by the mutator were nonconservative, whereas only half of the substitutions made by EMS were nonconservative. In addition, the mutator generated more structure-disturbing amino acid changes (Gly/Pro). The transversion bias of non-synonymous substitutions by the mutator generates more diverse amino acid substitution patterns than does the transition bias of EMS.

Gap between a higher apparent mutation frequency and fewer mutations

Although the average base-substitution mutation rate of EMS was approximately 100 times higher than that of the mutator, the mutation frequency of the mutator was approximately 7 times higher than that of EMS. This gap between a higher apparent mutation frequency and fewer mutations may be explained by the higher proportion of amino acid changes and the diversity of amino acid substitutions by the mutator. This suggests one plausible explanation for the effectiveness of the disparity mutagenesis.

The disparity mutagenesis technique has been successfully applied to not only eukaryotic microorganisms such as *S. cerevisiae*,[5,7-9] *S. pombe*,[9] and *Ashbya gossypii*,[10] but also to prokaryotic microorganisms such as *Escherichia coli*,[4] and *Bradyrhizobium japonicum*.[6] I believe that this novel

mutagenesis technique has the potential to be applied to a wide variety of microorganisms.

Conclusion

In this study, I developed COVA and applied it to re-sequence yeast mutagenized by novel mutagenesis technique. COVA has successfully reduced the time for procedure of variant annotation, compared with manual annotation. This study also demonstrated that a proofreading-deficient and low-fidelity *polδMKII* mutator is a useful and efficient method for rapid strain improvement based on *in vivo* mutagenesis. It has been suggested that organisms may accelerate evolution by decreasing the fidelity of the proofreading activity of *polδ* in nature.[29] Therefore, this mutator may also be useful for studying the acceleration of evolution.

Acknowledgments

I would like to express my gratitude to a member of Neo-Morgan Laboratory Inc., Dr. Sanae Fukushima-Tanaka, Dr. Ken Kasahara, and Dr. Takayuki Horiuchi. They provided mutagenized yeast strains and useful suggestions. Gratitude is also expressed to Dr. M. Furusawa, a founder of Neo-Morgan Laboratory Inc., for critical reading of the manuscript and for useful suggestions. This study was supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (S0801025).

References

1. Labrou, N. E. 2010, Random mutagenesis methods for in vitro directed enzyme evolution. *Curr. Protein Pept. Sc.*, **11**, 91-100.

2. Selifonova, O., Valle, F., and Schellenberger, V. 2001, Rapid evolution of novel traits in microorganisms. *Appl. Environ. Microbiol.*, **67**, 3645-3649.

3. Furusawa, M. and Doi, H. 1992, Promotion of evolution: disparity in the frequency of strand-specific misreading between the lagging and leading DNA strands enhances disproportionate accumulation of mutations. *J. Theor. Biol.*, **157**, 127-133.

4. Tanabe, K., Kondo, T., Onodera, Y., et al. 1999, A conspicuous adaptability to antibiotics in the *Escherichia coli* mutator strain, dnaQ49. *FEMS Microbiol. Lett.*, **176**, 191-196.

5. Shimoda, C., Itadani, A., Sugino, A. et al. 2006, Isolation of thermotolerant mutants by using proofreading-deficient DNA polymerase delta as an effective mutator in *Saccharomyces cerevisiae. Genes Genet. Syst.*, **81**, 391-397.

6. Itakura, M., Tabata, K., Eda, S., et al. 2008, Generation of *Bradyrhizobium japonicum* mutants with increased N2O reductase activity by selection after introduction of a mutated *dnaQ* gene. *Appl. Environ. Microbiol.*, **74**, 7258-7264.

7. Abe, H., Fujita, Y., Chiba, Y., et al. 2009, Upregulation of genes involved in gluconeogenesis and the glyoxylate cycle suppressed the drug sensitivity of an N-Glycan-deficient *Saccharomyces cerevisiae* mutant. *Biosci. Biotechnol. Biochem.*, **73**, 1398-1403.

8. Abe, H., Fujita, Y., Takaoka, Y., et al. 2009, Ethanol-tolerant *Saccharomyces cerevisiae* strains isolated under selective conditions by over-expression of a proofreading-deficient DNA polymerase delta. *J. Biosci. Bioeng.*, **108**, 199-204.

9. Abe, H., Takaoka, Y., Chiba, Y., et al. 2009, Development of valuable yeast strains using a novel mutagenesis technique for the effective production of therapeutic glycoproteins. *Glycobiology*, **19**, 428-436.

10. Park, E. Y., Ito, Y., Nariyama, M., et al. 2011, The improvement of riboflavin production in *Ashbya gossypii* via disparity mutagenesis and DNA microarray analysis. *Appl. Microbiol. Biotechnol.*, **91**, 1315-1326.

11. Matsuzawa, T., Fujita, Y., Tanaka, N., et al. 2011, New insights into galactose metabolism by *Schizosaccharomyces pombe*: isolation and characterization of a galactose-assimilating mutant. *J. Biosci. Bioeng.*, **111**, 158-166.

12. Flibotte, S., Edgley, M. L., Chaudhry, I., et al. 2010, Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. *Genetics*, **185**, 431-441.

13. Fukui, T., Yamauchi, K., Muroya, T., et al. 2004, Distinct roles of DNA polymerases delta and epsilon at the replication fork in Xenopus egg extracts. *Genes Cells*, **9**, 179-191.

14. Li, L., Murphy, K. M., Kanevets, U., et al. 2005, Sensitivity to phosphonoacetic acid: a new phenotype to probe DNA polymerase delta in *Saccharomyces cerevisiae*. *Genetics*, **170**, 569-580.

15. Morrison, A., Johnson, A. L., Johnston, L. H., et al. 1993, Pathway correcting DNA replication errors in *Saccharomyces cerevisiae*. *EMBO J.*, **12**, 1467-1473.

16. Morrison, A. and Sugino, A. 1994, The 3'-->5' exonucleases of both DNA polymerases delta and epsilon participate in correcting errors of DNA replication in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.*, **242**, 289-296.

17. Nick McElhinny, S. A., Stith, C. M., Burgers, P. M., et al. 2007, Inefficient proofreading and biased error rates during inaccurate DNA synthesis by a mutant derivative of *Saccharomyces cerevisiae* DNA polymerase delta. *J. Biol. Chem.*, **282**, 2324-2332.

18. Venkatesan, R. N., Hsu, J. J., Lawrence, N. A., et al. 2006, Mutator phenotypes caused by substitution at a conserved motif A residue in eukaryotic DNA polymerase delta. *J. Biol. Chem.*, **281**, 4486-4494.

19. Timmermann, B., Jarolim, S., Russmayer, H., et al. 2010, A new dominant peroxiredoxin allele identified by whole-genome re-sequencing of random mutagenized yeast causes oxidant-resistance and premature aging. *Aging (Albany NY)*, **2**, 475-486.

20. Gietz, R. D. and Sugino, A. 1988, New yeast-*Escherichia coli* shuttle vectors constructed with in vitro mutagenized yeast genes lacking six-base pair restriction sites. *Gene*, **74**, 527-534.

21. Lea, D. E., and C. A. Coulson 1949, The distribution of the numbers of mutants in bacterial populations. *J. Genet.*, **49**, 264-285.

22. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.

23. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

24. Robinson, J. T., Thorvaldsdottir, H., Winckler, W., et al. 2011, Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24-26.

25. Lynch, M., Sung, W., Morris, K., et al. 2008, A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 9272-9277.

26. Henikoff, S. and Henikoff, J. G. 1992, Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10915-10919.

27. Wong, T. S., Roccatano, D., Zacharias, M., et al. 2006, A statistical analysis of random mutagenesis methods used for directed protein evolution. *J. Mol. Biol.*, **355**, 858-871.

28. Rasila, T. S., Pajunen, M. I. and Savilahti, H. 2009, Critical evaluation of random mutagenesis by error-prone polymerase chain reaction protocols, *Escherichia coli* mutator strain, and hydroxylamine treatment. *Anal. Biochem.*, **388**, 71-80.

29. Furusawa, M. 2011, Implications of double-stranded DNA structure for development, cancer and evolution. *Open Journal of Genetics*, **1**, 78-87.

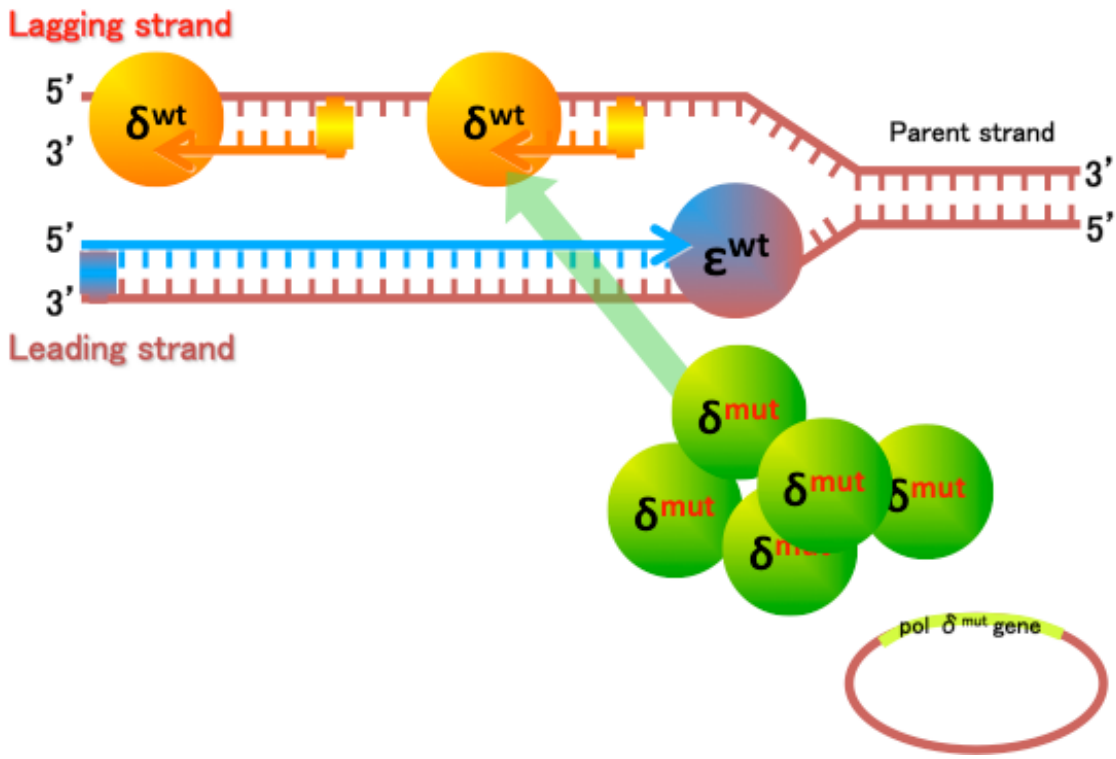

Figure 1. Scheme of DNA replication and the disparity mutagenesis.

In eukaryote, leading strand of DNA is continuously replicated by DNA polymerase ε (epsilon). Whereas, lagging strand of DNA is discontinuously replicated by DNA polymerase δ (delta). Due to its complex manner, the disparity model assumes that mutations occur preferentially on the lagging strand. In the disparity mutagenesis, the proofreading-deficient DNA polymerase δ is overexpressed from the plasmid vector, therefore the mutation rate is transiently increased.

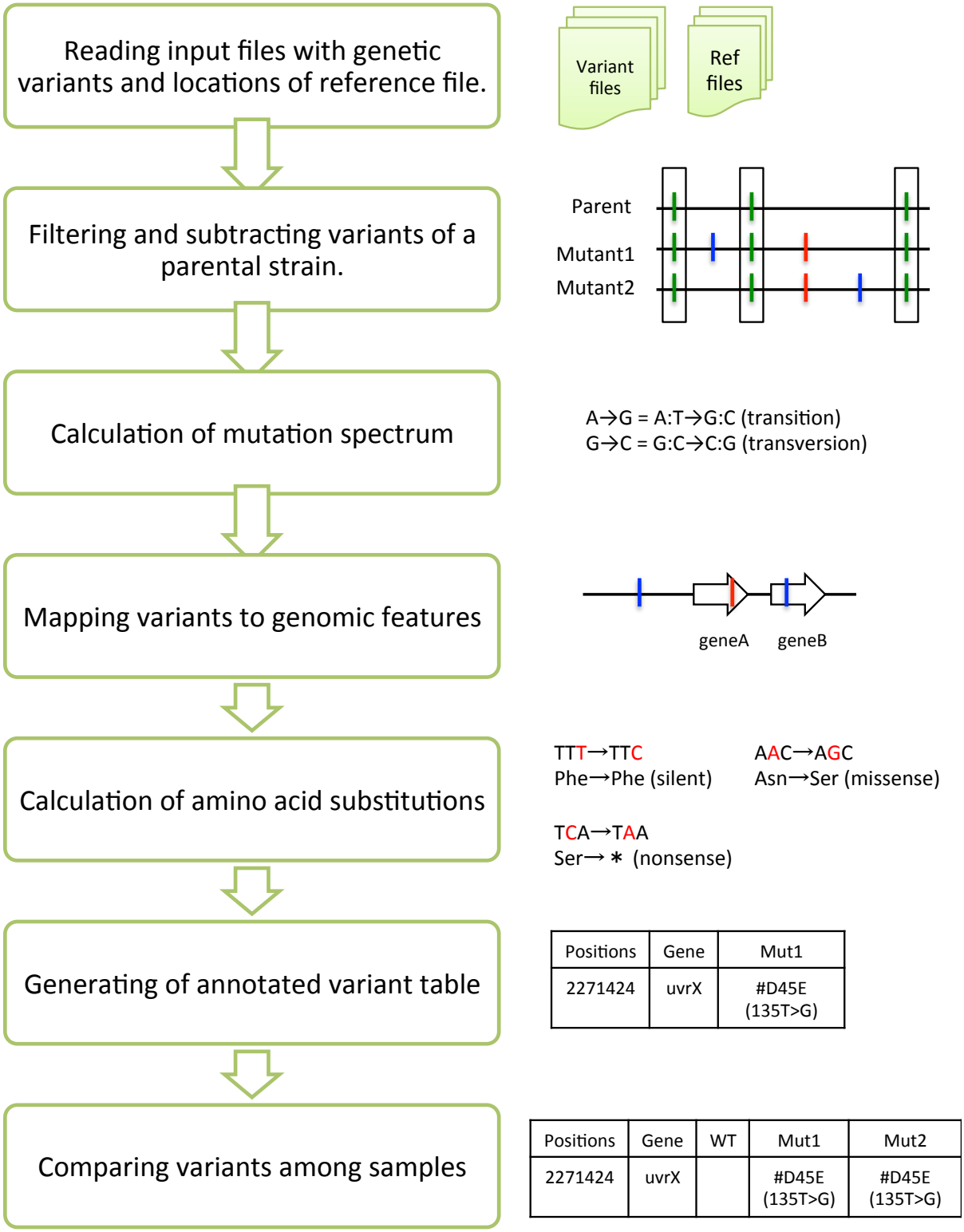

Figure 2. Workflow of annotation of variants using COVA.

The COVA consists of seven major steps to annotate variants.

Pos	PosType	VarType	Subtract	GeneName	Wildtype_V	Mutant1_V	Mutant2_V	Wildtype_R	Mutant1_R	Mutant2_R
165748	upstream	deletion	TRUE	trnI-Asn	19>-C	19>-C	19>-C	54 (0.83)	63 (0.92)	50 (0.8)
165825	downstream	insertion	TRUE	trnI-Asn	1>+C	1>+C	1>+C	66 (0.86)	92 (0.89)	63 (0.9)
557865	upstream	insertion	TRUE	ydzN	94>+T	94>+T	94>+T	45 (0.84)	65 (0.8)	65 (0.75)
728638	upstream	snp	FALSE	gatC		7G>A			63 (0.98)	
970381	CDS	snp	FALSE	cspR		#T83A (247A>G)			66 (0.97)	
1155476	CDS	snp	FALSE	wprA		#G563D (1688G>A)			71 (0.99)	
1317152	downstream	insertion	FALSE	yjpA		10>+GT	10>+GT		56 (0.71)	71 (0.75)
1317153	downstream	insertion	TRUE	yjpA	11>+T	11>+T	11>+T	52 (0.62)	55 (0.65)	71 (0.73)
1365685	CDS	snp	FALSE	dppE		#A512V (1535C>T)			60 (0.92)	
1684981	CDS	snp	FALSE	topA			#T441A (1321A>G)			57 (0.96)
2097080	downstream	insertion	TRUE	yocK	30>+A	30>+A	30>+A	35 (0.6)	51 (0.73)	47 (0.64)
2271424	CDS	snp	TRUE	uvrX	R78R (234A>G)	R78R (234A>G)	R78R (234A>G)	53 (1.0)	76 (1.0)	62 (1.0)
2271505	CDS	snp	TRUE	uvrX	V51V (153G>A)	V51V (153G>A)	V51V (153G>A)	43 (1.0)	61 (0.98)	61 (1.0)
2271523	CDS	snp	TRUE	uvrX	#D45E (135T>G)	#D45E (135T>G)	#D45E (135T>G)	45 (1.0)	54 (0.98)	56 (1.0)
2471230	CDS	snp	FALSE	yqjQ			P184P (552G>T)			54 (0.98)

Figure 3. Example of the output table for comparison of variants among samples

A comparison table of variants (SNP/InDel) among multiple samples helps to pinpoint causal variation(s) relating to phenotype. This file contains annotated variants from all samples, such as type and frequency of reads having variant and its effect on the gene. The variants observed in the parental strain are flagged, so one can subtract parental variations. This file is a comma-delimited file, so can be opened in Excel. In 'SampleName_V' field, annotated variant is reported as the following manner: for snp in CDS, the reference amino acid, the position of the corresponding codon, and the substituted amino acid, and in parentheses the position of the substituted base, the reference base, and the substituted base; * means a stop codon (not appeared in this example), # means a non-synonymous change, and % means a heterogeneous variant (also not shown in this example); for insertion/deletion in CDS (frame shift variant which did not appear in this example), the position of the corresponding codon and in parentheses the position of the corresponding base and the inserted (shown as +) or deleted (shown as -) base); for SNP in non-coding regions, the position of the substituted base, the reference base, and the substituted base; and for insertion/deletion in non-coding regions, the position of the corresponding base and the inserted (+) or deleted (-) base. 'SampleName_R' fields show number of reads covering a given variant and fraction of reads having the variant in parentheses.

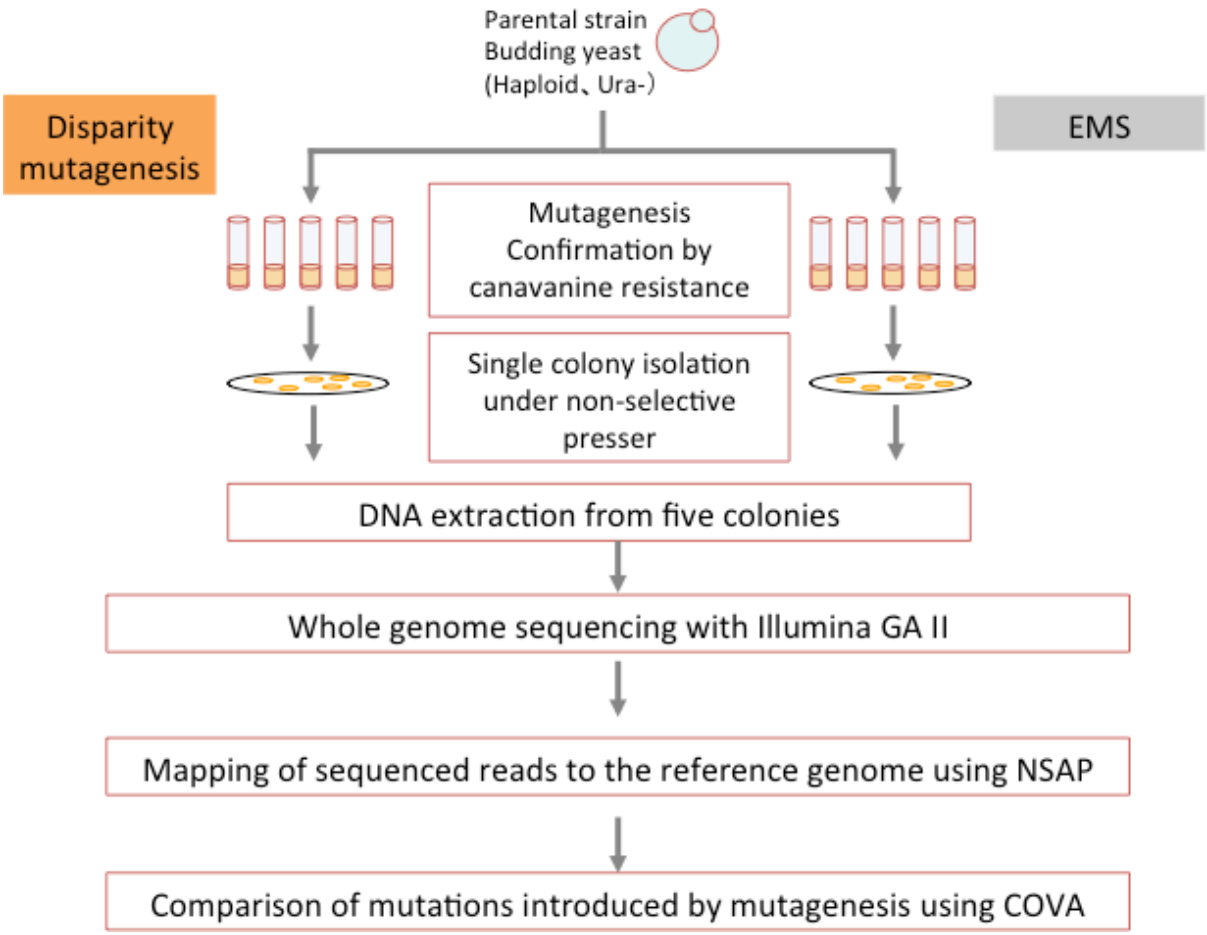

Figure 4. Workflow of this study.

Two mutagenesis techniques were evaluated in this study. The parental haploid yeast strain was mutagenized by each method. In disparity mutagenesis, a yeast expression plasmidYCplac33/*pol δ MKII* was transformed to the parental strain, and then five transformants were picked and independently cultivated in SC medium at 30ºC for 24 h (about 30 generations). In EMS mutagenesis, the parental strain was suspended in PBS containing EMS, and was incubated at 30ºC for 1 h. Aliquots were spread on plates containing canavanine, and mutation frequencies were calculated as the number of drug resistant colonies per the number of total colonies. Single colonies were isolated from each test tube and DNA was extracted from five colonies independently. DNA mutagenized by each method was sequenced with Illumina GA II. Sequenced reads were mapped to the reference genome, and detected mutations were compared.

Table 1. Relationship between mutation frequency and survival after EMS treatment

EMS concentration (%)	Mutation frequency of canavanine resistant ($\times 10^{-7}$)	Fold elevation*	Survival (%)
0.0	2	1	100
1.5	35	18	51
2.0	36	19	30
2.5	33	17	21
3.0	37	19	12

* Fold elevation is relative to untreated cells.

Table 2. Frequency of drug-resistant mutants in the mutator strains

Plasmids	Mutation frequency of canavanine resistant ($\times 10^{-7}$)	Fold elevation*
YCplac33	3.70	1
YCplac33/*polδMKII*	486.7 ± 145.0#	132

* Fold elevation is relative to the empty vector YCplac33.

Mean ± standard deviation of 3 SC plates.

Table 3. Sequencing and mapping statistics

Sample name	Number of mapped unique reads	% mapped reads	% genome covered by unique reads	Average coverage* by unique reads
BY2961	11,155,487	96.13	94.97	87.9 ×
EMS1	5,406,681	96.94	94.81	42.2 ×
EMS2	6,240,554	97.26	94.85	48.7 ×
EMS3	5,275,583	98.12	94.81	41.2 ×
EMS4	4,502,271	97.17	94.80	35.2 ×
EMS5	4,113,345	96.27	94.83	32.1 ×
Mutator1	9,612,541	93.93	94.90	75.8 ×
Mutator2	5,111,531	92.39	94.79	39.9 ×
Mutator3	5,649,822	96.11	94.95	44.1 ×
Mutator4	4,226,405	98.79	94.85	33.0 ×
Mutator5	9,855,938	97.36	95.10	77.6 ×

* Coverage is defined as the percentage of bases in the genome that have at least 1 uniquely mapped read at that position.

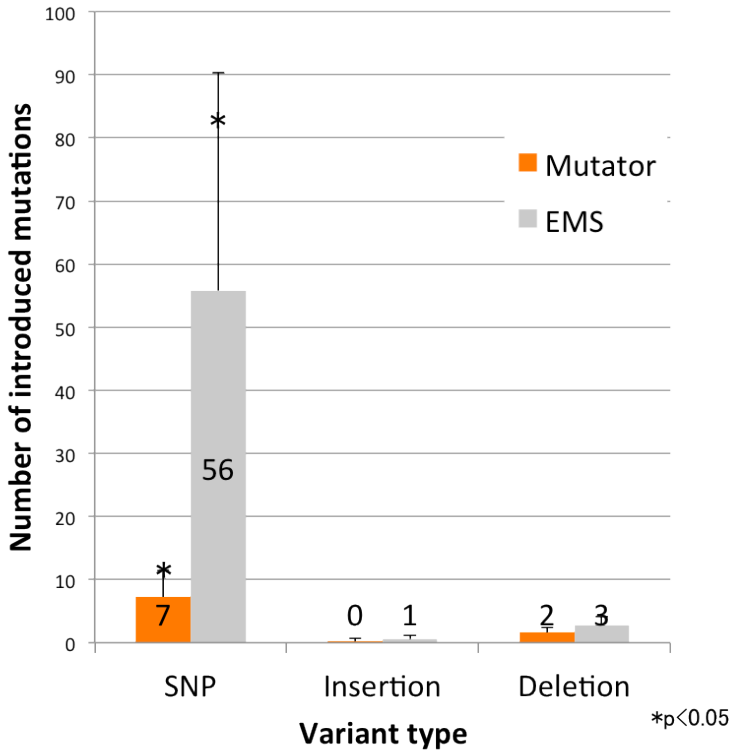

Figure 5. Average number of mutations introduced by mutagenesis.

By subtracting parental mutations from each mutagenized strain, I determined the number of mutations that were introduced by each mutagen. Bars represent mean ± standard error for 5 clones. *$p < 0.05$ versus mutator in a two-sample t-test.

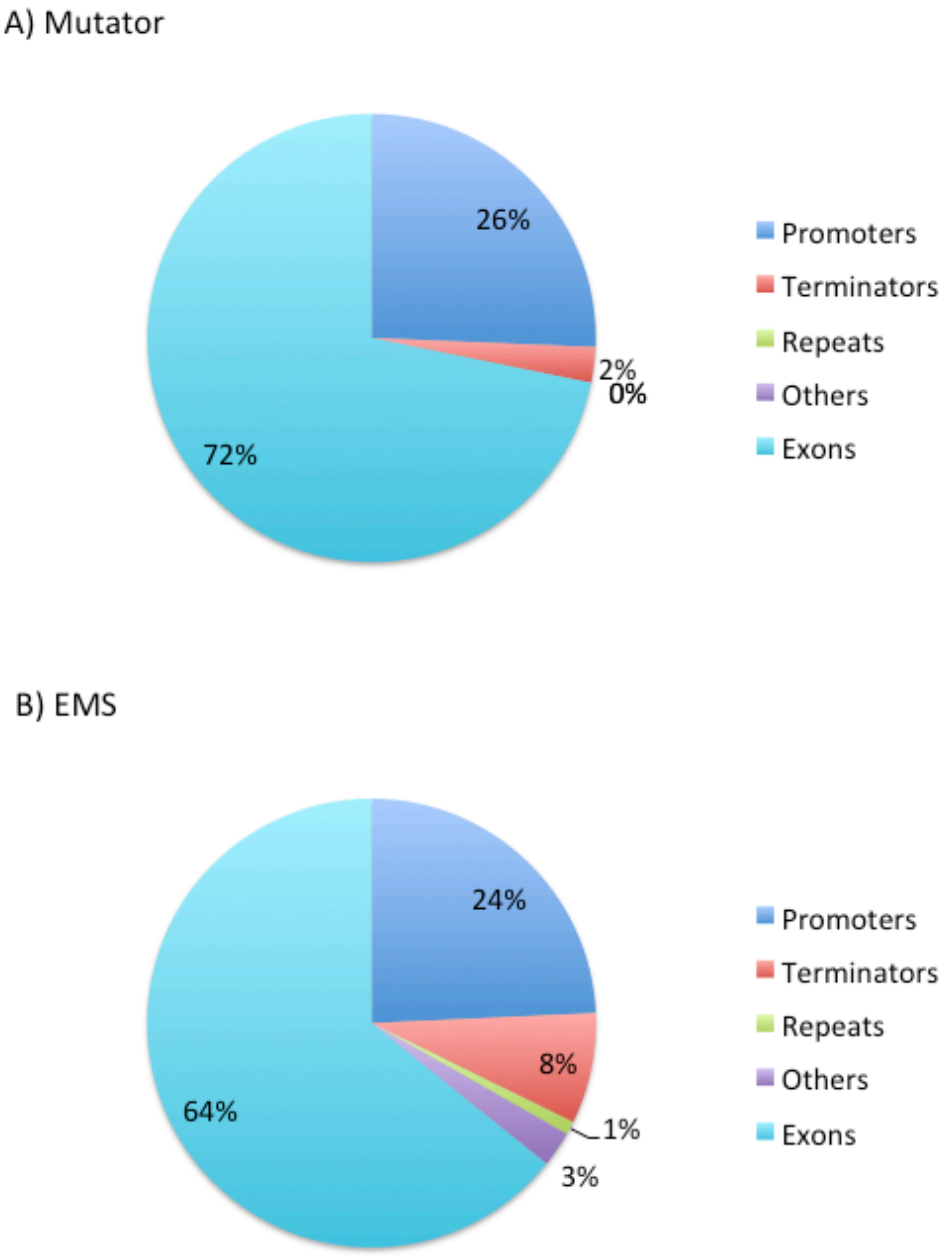

Figure 6. Distribution of SNVs across various gene features.

The mutator (A) and EMS (B) generated mutations that were distributed similarly across the various gene features. The data for individual strains were combined according to the mutagen used. Promoters indicate the region 1 kb upstream of each gene. Terminators indicate the region 200 bp downstream of each gene.

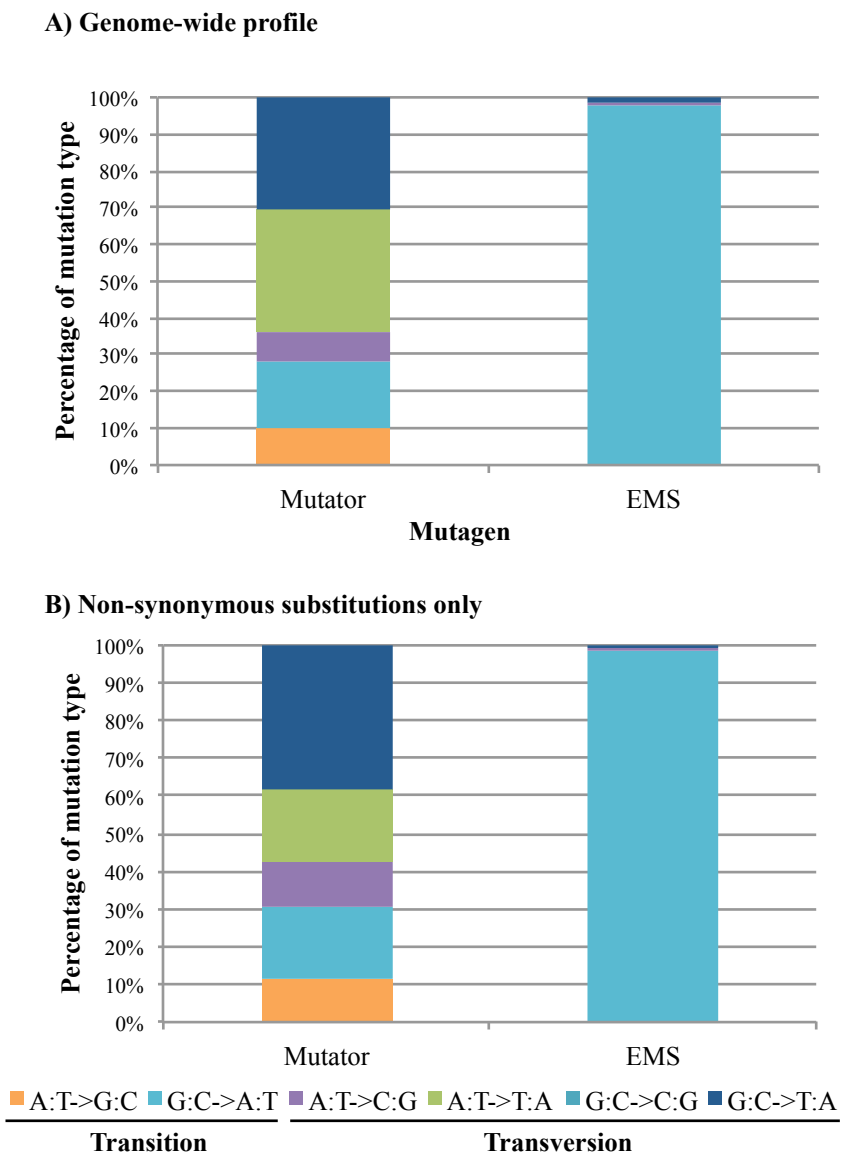

Figure 7. Proportion of transitions and transversions induced by *polδ* and EMS.

The mutations spectra show the frequency of transitions and transversions generated by the mutator and EMS. The data for the individual strains were combined according to the mutagen used. The color key identifying the type of mutation is provided in the inset. Complementary mutations, such as A→C and T→G, are pooled. (A) Genome-wide profile; (B) Non-synonymous substitutions only.

Table 4. Mutations at protein level

	Mutator		EMS	
	n	%	n	%
Total mutations	28	100	201	100
Preserved amino acids	2	7.1	74	36.8
Amino acid changes	24	85.7	123	61.2
Stop	2	7.1	4	2.0
Changes in codon letter	28	100	201	100
1st	11	39.3	64	31.8
2nd	13	46.4	65	32.3
3rd	4	14.3	72	35.8
Impact of amino acid change	24	100	123	100
Conservative[a]	4	16.7	57	46.3
Nonconservative	20	83.3	66	53.7
Stop and Gly/Pro codons	4	15.4	4	3.1
Stop	2	50.0	4	100.0
Gly/Pro	2	50.0	0	0.0

[a] Conservative and nonconservative amino acid substitutions were defined according to the BLOSUM62 matrix.[26]

CHAPTER III

Establishment of method for genome sequence determination using the 3rd generation sequencing and application to a novel lactic acid-producing bacterium

CHAPTER III

Establishment of method for genome sequence determination using the 3rd generation sequencing and application to a novel lactic acid-producing bacterium

Abstract

Draft genome sequences of microorganisms can be obtained rapidly and cost-effectively by using second-generation sequencing technologies. The recent advent of third-generation sequencing promises to offer a complete genome sequence. *Enterococcus mundtii* QU 25, a non-dairy bacterial strain of ovine faecal origin, can ferment both cellobiose and xylose to produce L-lactic acid. The use of this strain is highly desirable for economical L-lactate production from renewable biomass substrates. Genome sequence determination is necessary for the genetic improvement of this strain. In this study, the complete genome sequence of strain QU 25 is determined, primarily using Pacific Biosciences sequencing technology. The *E. mundtii* QU 25 genome comprises a 3,022,186-bp single circular chromosome (GC content, 38.6%) and five circular plasmids: pQY182, pQY082, pQY039, pQY024, and pQY003. In all, 2,900 protein-coding sequences, 63 tRNA genes, and 6 rRNA operons were predicted in QU 25 chromosome. Plasmid pQY024 harbours genes for mundticin production. It was found that strain QU 25 produces a bacteriocin, suggesting that mundticin-encoded genes on plasmid pQY024 were functional. For lactic acid fermentation, two gene clusters were identified—one involved in the initial metabolism of xylose and uptake of pentose and the second containing genes for the pentose phosphate pathway and uptake of related sugars. This is the first complete genome sequence of an *E. mundtii* strain. The data provide insights into lactate production in this bacterium and its evolution among enterococci.

Introduction

Optically pure lactic acid is necessary for the production of the bioplastic polylactic acid. The use of cellulosic biomass instead of food crops should lower the cost for commercial production of this green plastic. *Enterococcus mundtii* QU 25 is a non-dairy bacterial strain that was originally isolated from ovine feces.[1] Unlike most lactic acid bacteria (LAB), strain QU 25 can ferment both cellobiose and xylose to produce L-lactic acid.[1,2] This strain metabolizes a mixture of glucose and cellobiose simultaneously without apparent carbon catabolite repression[1], and it produces optically pure L-lactate (≥99.9%) with a yield of 1.41 mol/mol xylose consumed, without by-products such as acetic acid or ethanol.[1,2] Moreover, high productivity of L-lactic acid in an open repeated batch fermentation system under non-sterile conditions was demonstrated.[3] Therefore, the use of strain QU 25 is highly desirable for the economical production of L-lactate from renewable biomass substrates. Furthermore, determination of the genome sequence of this bacterium is necessary to generate optimized, recombinant strains for commercial use.

Draft genome sequences of microorganisms can be obtained rapidly and cost-effectively by using second-generation sequencing technologies. Owing to its relatively short read length of second-generation platforms (100–700 bp), obtaining a complete genome sequence requires additional costs and time-consuming finishing steps such as scaffolding and gap closing. A typical draft genome consists of dozens or hundreds of contigs/scaffolds (Fig. 1). However, repetitive DNA elements such as an rRNA operon, a phage region, and an insertion sequence, are usually determined as only partial sequences. This situation is not sufficient to examine the complete genome structure.

The recent advent of third-generation sequencing promises to offer a complete genome sequence. Third-generation single molecule real time (SMRT) sequencing technology developed by Pacific Biosciences (PacBio) can produce considerably long sequences (~23 kb). This technology

generates two types of sequences: CLR (continuous long reads) and CCS (circular consensus sequences) reads.[4] The read length of CLR can reach up to 23 kb; however the average base accuracy is only 82.1–84.4%.[5] On the other hand, CCS reads are consensus sequences obtained from multiple passes on a single sequence with relatively short lengths (~2 kb) and a low error rate.[6] Using the PBcR algorithm, read accuracy of CLR can be improved from 80% to over 99.9% by an error correction using CCS (Fig. 2)[5]. Complete genome sequencing using only PacBio sequence data was recently reported.[7] However, sequence and assembly methods of PacBio have not been well established as yet.

In this study, I aimed to establish methods for genome sequence determination using the 3rd generation sequencer PacBio RS and applied it to genome sequencing of *E. mundtii* QU 25. To date, there have been only two draft genome sequences available for *E. mundtii* (strains CRL1656[8] and ATCC 882). Here, the complete genome sequence of strain QU 25 was determined. Its chromosome sequence was sequenced using only PacBio sequence data. This is the first complete genome sequence of an *E. mundtii* strain. The data reveal useful insights on lactic acid fermentation in this bacterium, as well as phylogenetic relationships with other *Enterococcus* species.

Materials and methods

Media, growth conditions and extraction of genomic DNA for sequencing

E. mundtii QU 25 cells were grown to mid-log phase in GM17.[9] Cells from 100 ml-culture were harvested, suspended in 25 mL of a solution containing 2 g of polyethylene glycol 2000, 62.5 mg of egg-white lysozyme, and 5 mM Tris-hydrochloride (pH 8.0), and incubated for 60 min at 37°C. After centrifugation, the cells were resuspended in 12.5 mL of TES buffer (50 mM Tris-hydrochloride [pH 7.6], 20 mM sodium ethylenediaminetetraacetic acid, and 25% sucrose),

treated with 8 μg/mL of ribonuclease A for 60 min at 37°C, and then lysed with heat at 60°C for 2 h in the presence of 40 μg/mL of proteinase K and 1.7% sodium dodecyl sulphate. Total DNA was extracted gently from the lysate with PCI mixture (phenol, chloroform, and isoamyl alcohol in a ratio of 25:24:1, respectively) three times and precipitated with ethanol.

Short-read sequencing with Illumina

A 500-bp paired-end library was prepared following the manufactures' protocols. An 8-kb mate-paired library was prepared according to '454 GS FLX Titanium 20 kb and 8 kb Span Paired End Library Preparation Method Manual', with modification for an Illumina library preparation. The 500-bp paired-end library was sequenced using the Illumina Genome Analyzer IIx, generating 76-bp paired-end reads (48,724,736 reads, ~1234× coverage). The 8-kb mate-paired library was sequenced using the Illumina Genome Analyzer IIx, generating 100-bp paired-end reads (21,716,672 reads, ~723× coverage). The 500-bp and 8-kb reads were filtered and trimmed, then assembled using SOAPdenovo (http://soap.genomics.org.cn/). To evaluate accuracy of contig sequences determined by PacBio, the 500-bp paired-end reads were mapped to the contigs generated form PacBio RS, and detection of variants was carried out with threshold of frequency ≥90% using CLC Genomics workbench 6 (CLC bio, Aarhus, Denmark). The copy number of plasmid per chromosome was estimated based on the coverage ratio between the corresponding contig and the chromosome contig.

Long-read sequencing with Roche/454 sequencing

A 1.6-kb fragment library was sequenced using the GS-FLX+ system. A total of 302,056 reads (mean length, 455 bp; ~34× coverage) was generated and assembled using Newbler (ver. 2.6).

Ultra long-read sequencing with PacBio RS

Two types of SMRTbell DNA template libraries were created with 1-kb and 10-kb sheared genomic DNA, and prepared using the standard PacBio RS sample preparation methods with C2 chemistry specific to each insert size. The 10-kb library was sequenced on eight SMRT cells with a 1 × 120 min collection protocol, generating 189,953 post-filtered continuous long reads (CLR; mean length, 3,702 bp; maximum length 20,405 bp; ~234× coverage). The 1-kb library was sequenced on eight SMRT cells with a 2 × 55 min collection protocol, generating 258,068 post-filtered circular consensus sequences (CCS; mean length, 660 bp; ~57× coverage). After error correction, the resulting 6,806 PBcR of at least 7-kb length (mean length, 8,517 bp; maximum length, 16,733 bp; ~20× coverage) were assembled.

Processing of PacBio RS data and validation of assembly with optical mapping

The error correction of CLR reads was performed using the command pacBioToCA.[5] CCS (57× length coverage) reads were used for correction. After error correction, reads named PacBio-corrected Reads (PBcR) were selected for assembly. Assembly was performed using Celera Assembler (ver. 7.0).[10] To confirm the correctness of the assembly, DNA from QU 25 cells was digested using *Nco*I and a whole-genome optical map (OpGen, Inc., Gaithersburg, MD) was generated.[11]

Genome annotation and comparative genome

At first, the genome sequence was automatically annotated using the MiGAP, Microbial Genome Annotation Pipeline (www.migap.org) with g-MiGAP level. In the pipeline, open reading frames (ORFs) were identified using MetaGeneAnnotator,[12] and genes for tRNAs and rRNAs were

identified by tRNAscan-SE and RNAmmer, respectively.[13] Predicted ORFs were annotated using BLASTP searches against other *Enterococcus* genomes, and National Center of Biological Information (NCBI) nr[14] with an E-value of 1×10^{-10}. Additional annotation was performed using InterProScan[15] and KEGG pathway analysis.[16] Regions containing prophage were predicted by a phage search tool, PHAST (http://phast.wishartlab.com/)[17] and further manual inspection. The insertion sequence (IS) was detected by ISsaga.[18] CRISPR loci were detected by CRISPRfinder.[19] For comparative genomes of other *Enterococcus* spp., a draft genome sequence of *E. mundtii* ATCC 882 was obtained from the Broad Institute website (https://olive.broadinstitute.org/genomes/ente_mund_atcc882.1). In addition, complete genome sequences of five *Enterococcus* spp. with (*E. casseliflavus* EC20[20]: NC_021023; *E. faecalis* V583[21]: NC_004668, NC_004669, NC_004670, and NC_004671; *E. faecium* Aus0004[22]: NC_017022, NC_017023, NC_017024, and NC_017032; *E. faecium* DO[23]: NC_017960, NC_017961, NC_017962, and NC_017963; *E. hirae* ATCC 9790[24:] NC_015845 and NC_018081) were obtained from the NCBI website for genome comparisons. GenomeMatcher[25] and In Silico Molecular Cloning Genomics Edition (IMC-GE) software (In Silico Biology, Japan) were also used for intra- and inter-QU 25 genome comparisons. For genome analysis described above, I have developed several in-house scripts written in Ruby language.

Assay of vancomycin resistance

E. mundtii QU 25 was propagated in MRS medium (Oxoid, Basingstoke, UK) at 30°C for 12 h as a pre-culture. One hundred microliters of pre-culture was inoculated in 10 mL of MRS medium containing various concentrations of vancomycin (2–100 μg/mL, Sigma, St. Louis, MO). After incubation at 30°C for 24 h, the minimum concentration of vancomycin that resulted in the

absence of growth was considered to be the minimum inhibitory concentration (MIC).

Assay of bacteriocin activity

Lactobacillus sakei JCM 1157T and *E. faecalis* JCM 5803T were employed as indicator strains for bacteriocin activity. Together with *E. mundtii* QU 25, *E. mundtii* QU 2 (mundticin producer)[26] and JCM 8731T (non-bacteriocin producer) were tested as positive and negative controls, respectively. The three *E. mundtii* strains were also used as indicator strains for cross- and self-immunity. All strains were propagated in MRS medium at 30°C for 12–18 h before use. Strains QU 25, QU 2, and JCM 8731T were cultured in MRS medium at 30°C for 12 h for bacteriocin production. Bacteriocin activity assay was performed by the spot-on-lawn method, as described previously.[27] Briefly, 10 μL of each cell-free culture supernatant was spotted onto a double-layered agar plate containing 5 mL of Lactobacilli Agar AOAC (BD, Sparks, MD, USA) inoculated with an overnight culture of an indicator strain as an upper layer, and 10 mL of MRS medium supplemented with 1.2% agar as a bottom layer. After overnight incubation at appropriate temperatures for indicator strains, bacterial lawns were checked for inhibition zones.

Assay of catalase and hemolysin activity

Strain QU 25 was tested for catalase activity by two methods. First, cells cultured in MRS liquid medium at 30°C for 12 h were collected by centrifugation and examined for catalase activity by the addition of 3% hydrogen peroxide solution. Second, colonies formed on the sheep blood-containing agar plates (Eiken Chemical, Tokyo, Japan) after incubation at 30°C for 24 h were examined by the addition of 3% hydrogen peroxide solution. Generation of oxygen bubbles was considered to be indicative of catalase activity. Colonies formed on blood agar plates were also

examined for hemolysin activity. Discoloration or clearing of blood agar in the vicinity of the colonies was regarded as hemolysis.

Accession code

The complete genome sequence for *E. mundtii* QU 25 was deposited in GenBank/DDBJ/EMBL under accession numbers AP013036 to AP013041.

Results and Discussion

Genome sequencing, assembly, and annotation

To determine the full genome sequence of *E. mundtii* QU 25, three sequencing technologies (Illumina, Roche/454, and PacBio) were tried out for genome assembly. Workflow of assembly is illustrated in Fig. 3. Summary of assembly statistics is shown in Table 1. First, a genome assembly using Illumina was tried, generating 238 scaffolds. Second, Roche/454 GS-FLX+ sequencing generated 60 contigs. However it seemed that it takes a lot of time and effort to finish genome. Then an assemble using PacBio RS was tried, generating five contigs. The length of the largest contigs was 3.02 Mb, the other contigs seemed to be plasmid due to its length.

To evaluate accuracy of these contig sequences from PacBio RS, the 500-bp paired-end reads of Illumina were mapped to them. Only 113 base-pair differences and 17 insertions were detected, confirming highly accuracy of contig sequences generated from PacBio RS. Comparison of the *Nco*I-digest of the whole-genome optical map and in silico-generated physical maps of contigs showed that the largest contig (3 Mb) was mapped to the chromosome, confirming the correctness of the assembly (Fig. 4). Thus, it was concluded that the remaining four contigs were plasmids. The presence of the 2.5-kb plasmid, which had been detected as an extrachromosomal element by

agarose gel electrophoresis (data not shown), was not included in the five contigs generated by PacBio, due to its assembly threshold of 7-kb read length. Therefore, one contig produced using Roche/454 was adapted. As a result, the complete chromosome sequence of *E. mundtii* QU 25 was determined using only PacBio RS sequencing. After determination of genome sequence, I performed genome annotation using various software and databases as illustrated in Fig. 5.

Summary of genomic features

The genome of QU 25 comprises a single circular chromosome of 3,022,186 bp (GC content, 38.6%) and five circular plasmids: pQY182, pQY082, pQY039, pQY024, and pQY003, with lengths of 181,920 bp, 82,213 bp, 38,528 bp, 23,629 bp, and 2,584 bp, and GC contents of 36.2%, 35.8%, 33.8%, 35.3%, and 38.9%, respectively (Table 2, Fig. 6). Coordinates on the genome were designated as bp starting from the first nucleotide of the start codon of *dnaA*. It was confirmed that the likely location of the replication terminus by identifying a single *dif* sequence (ACTTTGTATAATATATATTATGTAAACT, position 1,449,088 to 1,449,115) that aligns with the shift in GC skew (Fig. 6). A total of 2,900 protein-coding sequences (CDS), 63 tRNA genes, and 6 rRNA operons were predicted in the QU 25 chromosome.

Comparative genomic with other Enterococcus species

From the phylogeny based on 16S rRNA sequences, *E. mundtii* was closely related to *E. hirae* and *E. faecium*, while *E. casseliflavus* and *E. faecalis* were more distantly related to *E. mundtii*.[28] To investigate the taxonomic position of *E. mundtii* QU 25 based on genome-wide comparisons, I first carried out ortholog analysis with the draft genome of *E. mundtii* ATCC 882 and five complete genomes of other *Enterococcus* spp., including two *E. faecium* strains (DO and

Aus0004), *E. hirae* ATCC 9790, *E. casseliflavus* EC20, and *E. faecalis* V583 (Table 3, Fig. 6). The results were consistent with the phylogeny based on 16S rRNA. While *E. mundtii* ATCC 882 showed the highest similarity to QU 25 as anticipated, two other species (*E. faecium* DO and Aus0004, and *E. hirae* ATCC 9790) showed moderate degrees of similarity, while *E. casseliflavus* EC20 and *E. faecalis* V583 were the least similar. DNA dot plot analysis showed the centre diagonal line between QU 25, the two *E. faecium* strains, and *E. hirae* (Fig. 7), indicating that not only each of the genes but also their genome structures (or gene orders) were related. There were gene regions unique to the QU 25 strain (Fig. 6). Except for prophages, the 9-kb region (positions 1,359,588–1,368,963) includes five hypothetical proteins and a cell-wall-anchored protein with a LPXTG motif.

Prophages and insertion sequence elements (ISEs)

Three chromosomal regions were identified as prophage loci. Their positions are indicated in Fig. 6, and are named phiEmqu1 (38.7 kb; positions 806,547–845,215), phiEmqu2 (47.9 kb; positions 2,327,297–2,375,151), and phiEmqu3 (40.8 kb; positions 2,556,843–2,597,594). Sixty, 70, and 54 phage-related genes were identified in these regions, respectively (Table 4). Prophages phiEmqu1 and phiEmqu3 contained several putative genes involved in DNA replication. However, no genes for DNA synthesis were found in the largest prophage phiEmqu2, suggesting that it is replication-defective. As shown in Fig. 6, the locations of *ori* (*dnaA*) and *ter* (*dif*) were not exactly opposite each other. The *dif* motif, which is strongly associated with replication termini,[29] was about 60 kb off the exact opposite position of *ori* (*dnaA*). Phage-mediated replicore imbalance has been observed in *E. faecium* Aus0004.[22] Thus, the replicore balance of QU 25 could be disrupted by the insertion of phiEmqu2 and/or phiEmqu3.

Clustered regularly interspaced short palindromic repeats (CRISPRs) are involved in a

recently discovered interference pathway that protects cells from bacteriophages and conjugative plasmids. Approximately 40% of sequenced bacterial genomes and ~90% of genomes from archaea contain at least one CRISPR locus.[30] No CRISPR loci were detected in the QU 25 genome. Many insertion sequence elements (ISEs) have been found in enterococci. The relatively closed species *E. mundtii* ATCC 882, *E. faecium* DO and Aus0004 strains, and *E. hirae* ATCC 9790, have 21, 180, 76, and 14 ISEs and transposase-related genes, respectively. At least 13 different ISEs were detected in the QU 25 genome, ranging in copy number from 1 to 5, representing 33 distinct copies and distributed around the chromosome and plasmids (Table 5). The most frequently observed ISE type was the ISL3 family.

Plasmids

Enterococcus spp. have been reported to possess a number of plasmids that often confer resistance to antimicrobials and particular heavy metals, and serve to enhance virulence and/or DNA repair mechanisms.[31–34] In strain QU 25, the plasmid copy number per chromosome was estimated by observing the distribution of read coverage of the Illumina sequence read, which indicated one copy of pQY182, pQY082, pQY039, and pQY024, and five copies of pQY003 (Table 2). BLASTN analyses showed that these five plasmids were similar to those of *E. mundtii, E. faecium, E. hirae*, or *E. faecalis* (Table 6).

Plasmid pQY182 harbours genes that encode a two-component regulatory system, a cellulose 1,4-beta-cellobiosidase, a toxin-antitoxin system, and several proteins with DNA repair functions. Plasmid pQY082 harbours duplicated regions of 8.3 kb, which include the IS1675 transposase, ubiD family decarboxylase, two cell surface proteins, and two proteins with unknown functions (EMQU_3088-3094 and EMQU_3155-3160 with 99% similarity). pQY082 also harbours

genes that encode several proteins with DNA repair functions and a toxin-antitoxin system. Toxin-antitoxin systems have been frequently reported in *E. faecium* strains,[35] and the QU 25 chromosome additionally has at least four such systems. Plasmid pQY039 contains several genes for a DNA damage-inducible protein and a gene encoding a toxin-antitoxin system. Plasmid pQY024 also harbours genes for a DNA damage-inducible protein, a toxin-antitoxin system, and mundticin KS genes (see discussion later). Plasmid pQY003 only harbours genes with unknown function, except for the replication initiation protein.

Vancomycin resistance

Because many *Enterococcus* isolates show vancomycin resistance which has been associated with hospital-acquired infections, the sensitivity of QU 25 to this antibiotic was tested for the safety of industrial use. The results showed that the vancomycin MIC for QU 25 was >2 μg/mL, indicating that this strain is vancomycin-sensitive. Several known genes involved in vancomycin resistance (*vanA, vanB, vanX, vanH, vanR,* and *vanS*) [36] were not detected in the QU 25 genome and also in plasmids.

Bacteriocin activity and self- and cross-immunity

Bacteriocins are ribosomally synthesized bacterial peptides or proteins that show antimicrobial activity, generally against species that are closely related to bacteriocin producers.[26] Mundticin is one of the bacteriocins produced by some *E. mundtii* strains. It is significant to clarify whether QU 25 produces mundticin or not for the resistivity to contamination in a large-scale culture. Three genes, *munA* (mundticin precursor), *munB* (ATP-binding cassette (ABC) transporter), and *munC* (mundticin KS immunity protein), are responsible for mundticin production in *E. mundtii*

NERI 7393.[37] The gene cluster containing these three genes was identified on plasmid pQY024, and *munA* (EMQU_3203; 100% nucleotide sequence identity), *munB* (EMQU_3204; 99.56% nucleotide sequence identity), and *munC* (EMQU_3205; 98.99% nucleotide sequence identity) showed high homology with corresponding genes in strain NERI 7393.

To examine whether the gene cluster for mundticin synthesis was functional, QU 25 was tested for bacteriocin production. QU 25 and *E. mundtii* QU 2 showed bacteriocin activity against three indicator strains (*Lactobacillus sakei* JCM 1157T, *E. faecalis* JCM 5803T, and *E. mundtii* JCM 8731T), none of which showed inhibitory activity (Table 7). QU 25 and QU 2 showed no activity against each other (Table 7), which indicated that these strains have self- and cross-immunity against their bacteriocins. Bacteriocin-producing strains are known to have immunity (tolerance) to their own bacteriocins and to bacteriocins with similar structures. Collectively, these results strongly suggest that QU 25 produces a bacteriocin with similar characteristics to mundticin produced by QU 2.

Hemolysin activity

Hemolysin is one of the putative enterococcal virulence factors,[22] so it is important to test the hemolysin activity for the safety of industrial use in a large-scale culture. Four putative hemolysin genes (hemolysin, hemolysin III, hemolysin A, and α-hemolysin) were identified (EMQU_0190 and EMQU_0948, EMQU_0449, EMQU_0841, and EMQU_1982, respectively). Hemolysin activity was tested and no changes to the blood agar in the vicinity of the colonies were observed (data not shown), suggesting that these putative hemolysin genes in QU 25 might be inactive or silent under the tested culture conditions.

Genes involved in lactic acid fermentation

QU 25 was previously reported to have two different pathways for xylose metabolism: the phosphoketolase (PK) pathway and the pentose phosphate (PP)/glycolytic pathway in low xylose concentrations.[2,38] When QU 25 was grown in high xylose concentrations, PK activity was not detected. However, higher transaldolase and transketolase activities were detected[2], indicating that strain QU 25 would utilize the PP/glycolytic pathway, not the PK pathway, as the main pathway for lactic acid fermentation.

Genes for xylose metabolism in the QU 25 chromosome were located in a 22-kb region (positions 2,904,895–2,926,710 bp) in two gene clusters: one involved in the initial metabolism of xylose and uptake of pentose, and the other involved in the PP pathway and uptake of related sugars (Fig. 8). The first gene cluster contained *xylR* (EMQU_2811; xylose repressor), *xylA* (EMQU_2810; xylose isomerase), *xynB* (EMQU_2809; xylan beta-1,4-xylosidase; additionally, there is another *xynB* gene (EMQU_2642) outside of this cluster), *xylB* (EMQU_2805; D-xylulose kinase), putative xylose transporter genes annotated as L-arabinose and D-Ribose ABC transporter (EMQU_2806-2808), and a hypothetical protein (EMQU_2804), the N-terminal and C-terminal regions of DNA mismatch repair protein (EMQU_2803 and EMQU_2802 respectively, which were thought as pseudogenes), ABC transporter ATP-binding protein (EMQU_2801), and its permease (EMQU_2800). Since pentose transporters have been shown to be promiscuous,[39] D-xylose would likely be transported by these gene products in QU 25.

The second gene cluster contains genes for the PP/glycolytic pathway, including a transketolase (EMQU_2812) and a transaldolase (EMQU_2814). Furthermore, this cluster contains an allulose-6-phosphate 3-epimerase gene (EMQU_2813), genes for a fructose-like phosphotransferase system (EMQU_2815-2819), and a putative transcriptional regulator

(EMQU_2820). Transketolase is a key enzyme in the PP/glycolytic pathway, and the QU 25 chromosome additionally harbours one transketolase gene (EMQU_1275). Since allulose-6-phosphate 3-epimerase catalyzes the conversion of D-allulose-6-phosphate to D-fructose-6-phosphate, this enzyme and the fructose transporters may supply various ketoses to the PP/glycolytic pathway. For genes involved in the PK pathway, phosphoketolase (EMQU_1837), acetate kinase (EMQU_2620), phosphotransacetylase (EMQU_2119), acetaldehyde dehydrogenase (EMQU_2205), and alcohol dehydrogenase (EMQU_1129, EMQU_1829, and EMQU_2109) were dispersed throughout the chromosome. Metabolic pathway and genes involved in lactic acid fermentation of strain QU 25 are illustrated in Fig. 9.

In order to get insights into the characteristics of strain QU 25 on lactic acid fermentation, the analysis of KEGG pathway was performed. The overview of KEGG pathway map of QU 25 is illustrated in Fig. 10. Then the comparison of gene number among related species using KO (KEGG Orthology) gene assignment was performed (Fig. 11). QU 25 genome possesses more genes than related species in categories of ABC transporters and phosphotransferase system (PTS), indicating that it might have high ability in sugar transport. Further closely examination of key enzymes related to lactic acid fermentation from xylose was performed (Table 8). Two clinical isolates of *E. faecium* (DO and Aus0004) and *E. hirae* ATCC 9790 lack genes necessary for the metabolism of xylose (transaldolase, phosphoketolase, xylulokinase, and xylose isomerase). Another clinical isolate, *E. faecalis* V583 lacks genes for transketolase, transaldolase, and phosphoketolase. Thus, it is most unlikely that these strains metabolize xylose. *E. casseliflavus* EC20 has genes for phosphoketolase, xylulokinase, and xylose isomerase, so that this strain can metabolize xylose by the PK pathway. It also has two complete genes for transketolase, but lacks transaldolase for the PP/glycolytic pathway. However, Kato et al. reported that *Lactococcus lactis* IO-1, which can utilize xylose, also lacks the

gene for transaldolase and is presumed to have an alternative PP/glycolytic pathway.[40] Therefore, EC20 might metabolize xylose also by the PP/glycolytic pathway. Two *E. mundtii* strains (QU 25 and ATCC 882) have two genes of full-length transketolase and all other genes necessary for the xylose metabolism by the two pathways. The results of our genomic analysis coincide with the description on their phenotype about xylose metabolism in Bergey's Manual (p. 599, Table 116).[28] From these results, remarkable genomic features related to lactic acid fermentation in QU 25 still remain unclear, and the analysis of transcriptional regulation of these genes could help its clarification.

QU 25 was able to metabolize a mixture of glucose and cellobiose simultaneously without apparent carbon catabolite repression (CCR).[1] In Gram-positive bacteria, the roles of catabolite control protein A (CcpA) and seryl-phosphorylated form of histidine-containing protein (P-Ser-HPr) in CCR have been well studied.[41] Genes encoding CcpA (EMQU_1943), HPr (EMQU_0954), and HPr kinase/phosphorylase (EMQU_1951) were also found in the QU 25 genome. The mechanism by which CCR is prevented in the presence of glucose is still unknown.

Strain QU 25 shows a predominant production of L-(+)-lactate when grown at high concentrations of cellobiose and xylose, whereas D-lactic acid was not detected in the culture broth.[1,2] However, two L-lactate dehydrogenase genes (L-LDH; EMQU_1380 and EMQU_2714) and one D-lactate dehydrogenase gene (D-LDH; EMQU_2453) were identified. Potentially, little or no D-LDH was expressed under these culture conditions. Lactate racemase, another gene involved in D-lactic acid formation,[42] was not found in the QU 25 genome.

Conclusions

This study successfully demonstrated that the determination of a complete genome sequence was achievable using only 3rd generation sequencing. Through this study, I established the

methods of assembly, annotation, and genome comparison of a novel bacterial genome. This study also has highlighted the phylogenetic relationship of *E. mundtii* QU 25 with related enterococcal species and characterized mobile genetic elements, including multiple prophages and ISEs, plasmids, and genes for metabolic pathways for lactic acid fermentation in this strain. In addition, the bacteriocin activity of QU 25 was demonstrated. The complete *E. mundtii* QU 25 genome sequence described here may be an important resource in the genetic engineering of recombinant strains for optimized production of lactic acid.

Acknowledgments

I would like to express my gratitude to a member of Kyushu University, Mr. Shohei Satomi, Assistant Prof. Takeshi Zendo, and Prof. Kenji Sonomoto for providing DNA of QU 25 and conducting assay of biological characteristics of QU 25. I would also like to thank to Assistant Prof. Yuu Hirose (Toyohashi University of Technology) for conducting 454 sequencing. I also thank Mr. Hiroaki Yanase for helping the constriction of Illumina mate-pair library. Also I am grateful to Dr. Yuu Kanesaki, Ms. Tomoko Araya-Kojima, and Prof. Mariko Shimizu-Kadota for helpful discussions. I also thank Pacific Biosciences, Inc. and Tomy Digital Biology Co., Ltd. for PacBio sequencing and assembly, and Hitachi Solutions, Ltd. for optical mapping. This study was supported by the MEXT-Supported Program for the Strategic Research Foundation at Private Universities, 2008-2012 (S0801025).

References

1. Abdel-Rahman, M. A., Tashiro, Y., Zendo, T., et al. 2011, Isolation and characterisation of lactic acid bacterium for effective fermentation of cellobiose into optically pure homo L-(+)-lactic acid. *Appl. Microbiol. Biotechnol.*, **89**, 1039–1049.

2. Abdel-Rahman, M. A., Tashiro, Y., Zendo, T., et al. 2011, Efficient homofermentative L-(+)-lactic acid production from xylose by a novel lactic acid bacterium, *Enterococcus mundtii* QU 25. *Appl. Environ. Microbiol.*, **77**, 1892–1895.

3. Abdel-Rahman, M. A., Tashiro, Y., Zendo, T., et al. 2013, Improved lactic acid productivity by an open repeated batch fermentation system using *Enterococcus mundtii* QU 25. *RSC Adv.*, **3**, 8437–8445.

4. Shin, S. C., Ahn, D. H., Kim, S. J., et al. 2013, Advantages of single-molecule real-time sequencing in high-GC content genomes. *PLoS One*, **8**, e68824.

5. Koren, S., Schatz, M. C., Walenz, B. P., et al. 2012, Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.*, **30**, 693–700.

6. Travers, K. J., Chin, C.-S., Rank, D. R., et al. 2010, A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, **38**, e159.

7. Chin, C.-S., Alexander, D. H., Marks, P., et al. 2013, Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, **10**, 563–569.

8. Magni, C., Espeche, C., Repizo, G. D., et al. 2012, Draft genome sequence of *Enterococcus mundtii* CRL1656. *J. Bacteriol.*, **194**, 550.

9. Machii, M., Watanabe, S., Zendo, T., et al. 2012, Chemically defined media and auxotrophy of the prolific l-lactic acid producer *Lactococcus lactis* IO-1. *J. Biosci. Bioeng.,* **115**, 481-484.

10. Miller, J. R., Delcher, A. L., Koren, S., et al. 2008, Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**, 2818–2824.

11. Nagarajan, N., Read, T. D., and Pop, M. 2008, Scaffolding and validation of bacterial genome assemblies using optical restriction maps. *Bioinformatics*, **24**, 1229–1235.

12. Noguchi, H., Taniguchi, T., and Itoh, T. 2008, MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.*, **15**, 387–396.

13. Sugawara, H., Ohyama, A., Mori, H., et al. 2009, Microbial genome annotation pipeline (MiGAP) for diverse users. *The 20th International Conference on Genome Informatics (GIW2009) Poster and Software Demonstrations (Yokohama), S001-1-2*.

14. Wheeler, D. L., Barrett, T., Benson, D. A., et al. 2007, Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **35**, D5–D12.

15. Zdobnov, E. M. and Apweiler, R. 2001, InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.

16. Moriya, Y., Itoh, M., Okuda, S., et al. 2007, KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–185.

17. Zhou, Y., Liang, Y., Lynch, K. H., et al. 2011, PHAST: a fast phage search tool. *Nucleic Acids Res.*, **39**, W347–W352.

18. Varani, A. M., Siguier, P., Gourbeyre, E., et al. 2011, ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biol.*, **12**, R30.

19. Grissa, I., Vergnaud, G., and Pourcel, C. 2007, CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.

20. Palmer, K. L., Carniol, K., Manson, J. M., et al. 2010, High-quality draft genome sequences of 28 *Enterococcus* sp. isolates. *J. Bacteriol.*, **192**, 2469–2470.

21. Paulsen, I. T., Banerjei, L., Myers, G. S. A., et al. 2003, Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science*, **299**, 2071–2074.

22. Lam, M. M. C., Seemann, T., Bulach, D. M., et al. 2012, Comparative analysis of the first complete *Enterococcus faecium* genome. *J. Bacteriol.*, **194**, 2334–2341.

23. Qin, X., Galloway-Peña, J. R., Sillanpaa, J., et al. 2012, Complete genome sequence of *Enterococcus faecium* strain TX16 and comparative genomic analysis of *Enterococcus faecium* genomes. *BMC Microbiol.*, **12**, 135.

24. Gaechter, T., Wunderlin, C., Schmidheini, T., et al. 2012, Genome sequence of *Enterococcus hirae* (*Streptococcus faecalis*) ATCC 9790, a model organism for the study of ion transport, bioenergetics, and copper homeostasis. *J. Bacteriol.*, **194**, 5126–5127.

25. Ohtsubo, Y., Ikeda-Ohtsubo, W., Nagata, Y., et al. 2008, GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics*, **9**, 376.

26. Zendo, T., Eungruttanagorn, N., Fujioka, S., et al. 2005, Identification and production of a bacteriocin from *Enterococcus mundtii* QU 2 isolated from soybean. *J. Appl. Microbiol.*, **99**, 1181–1190.

27. Ennahar, S., Asou, Y., Zendo, T., et al. 2001, Biochemical and genetic evidence for production of enterocins A and B by *Enterococcus faecium* WHE 81. *Int. J. Food Microbiol.*, **70**, 291–301.

28. Whitman, W. B., Goodfellow, M., Kämpfer, P., et al. 2012, *Bergey's manual of systematic bacteriology*. Springer Publishing Company: New York.

29. Hendrickson, H. and Lawrence, J. G. 2007, Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. *Mol. Microbiol.*, **64**, 42–56.

30. Marraffini, L. A. and Sontheimer, E. J. 2010, CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat. Rev. Genet.*, **11**, 181–190.

31. Arias, C. A., Panesso, D., Singh, K. V., et al. 2009, Cotransfer of antibiotic resistance genes and a hylEfm-containing virulence plasmid in *Enterococcus faecium*. *Antimicrob. Agents Chemother.*, **53**, 4240–4246.

32. Garcia-Migura, L., Hasman, H., and Jensen, L. B. 2009, Presence of pRI1: a small cryptic mobilizable plasmid isolated from *Enterococcus faecium* of human and animal origin. *Curr. Microbiol.*, **58**, 95–100.

33. Garcia-Migura, L., Liebana, E., and Jensen, L. B. 2007, Transposon characterization of vancomycin-resistant *Enterococcus faecium* (VREF) and dissemination of resistance associated with transferable plasmids. *J. Antimicrob. Chemother.*, **60**, 263–268.

34. Hasman, H., Kempf, I.., Chidaine, B., et al. 2006, Copper resistance in *Enterococcus faecium*, mediated by the tcrB gene, is selected by supplementation of pig feed with copper sulfate. *Appl. Environ. Microbiol.*, **72**, 5784–5789.

35. Moritz, E. M. and Hergenrother, P. J. 2007, Toxin-antitoxin systems are ubiquitous and plasmid-encoded in vancomycin-resistant enterococci. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 311–316.

36. Pootoolal, J., Neu, J., and Wright, G. D. 2002, Glycopeptide antibiotic resistance. *Annu. Rev. Pharmacol. Toxicol.*, **42**, 381–408.

37. Kawamoto, S., Shima, J., Sato, R., et al. 2002, Biochemical and genetic characterization of mundticin KS, an antilisterial peptide produced by *Enterococcus mundtii* NFRI 7393. *Appl. Environ. Microbiol.*, **68**, 3830–3840.

38. Abdel-Rahman, M. A., Tashiro, Y., and Sonomoto, K. 2013, Recent advances in lactic acid production by microbial fermentation processes. *Biotechnol. Adv.*, **31**, 877–902.

39. Song, S. and Park, C. 1998, Utilization of D-ribose through D-xylose transporter. *FEMS Microbiol. Lett.*, **163**, 255–261.

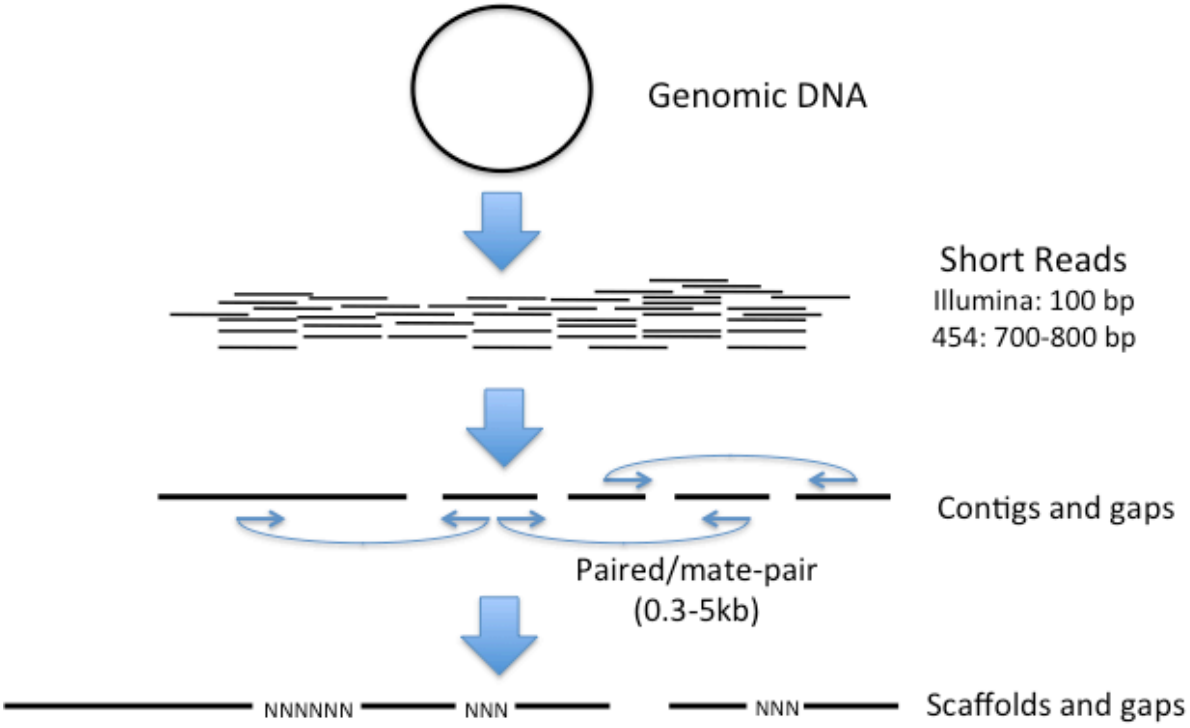

Figure 1. Scheme of determination of draft sequence using 2nd generation sequencing

Genomic DNA was fragmented, generating random and overlapping DNA fragments. These fragments can range from 300bp to 800bp in length. Next, whole genome shotgun sequence is performed. Illumina platform generates short reads with 100 bp lengths. 454 platform generates short reads with 700-800 bp length. In assembly step, short read sequences are assembled, generating dozens to hundreds of contigs. Paired-end/mate-pair reads contain sequences of both ends of fragments. They are useful for joining contigs. Joining contigs using paried/mate end reads are called as scaffolding. Scaffolds usually contain gap sequences representing as 'N'.

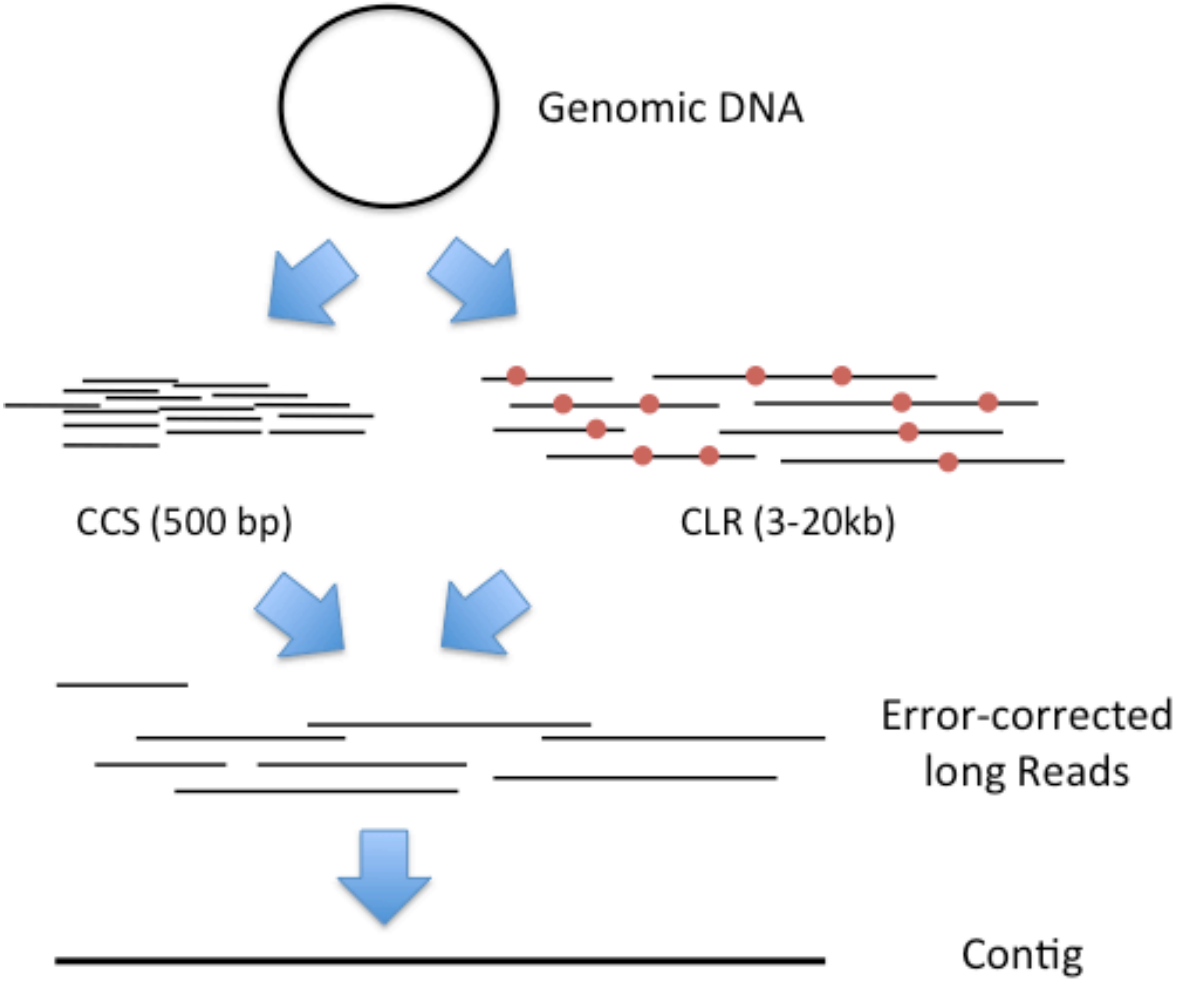

Figure 2. Scheme of determination of a genome sequence using the 3rd generation sequencer PacBio RS

Genomic DNA was fragmented, generating random and overlapping DNA fragments. For circular consensus sequences (CCS), DNA is fragmented to less than 1kb. Mean length of CCS is 500 bp with a low error rate. For continuous long reads (CLR), DNA is fragmented to approximately 10kb. The length of CLR ranges from 3-20kb, however CLR contain many sequencing errors (represented as red circles). Error correction with mapping CCS to CLR can generate long reads with high read accuracy. These long reads are assembled by consensus overlap assembly, generating long contigs without gap.

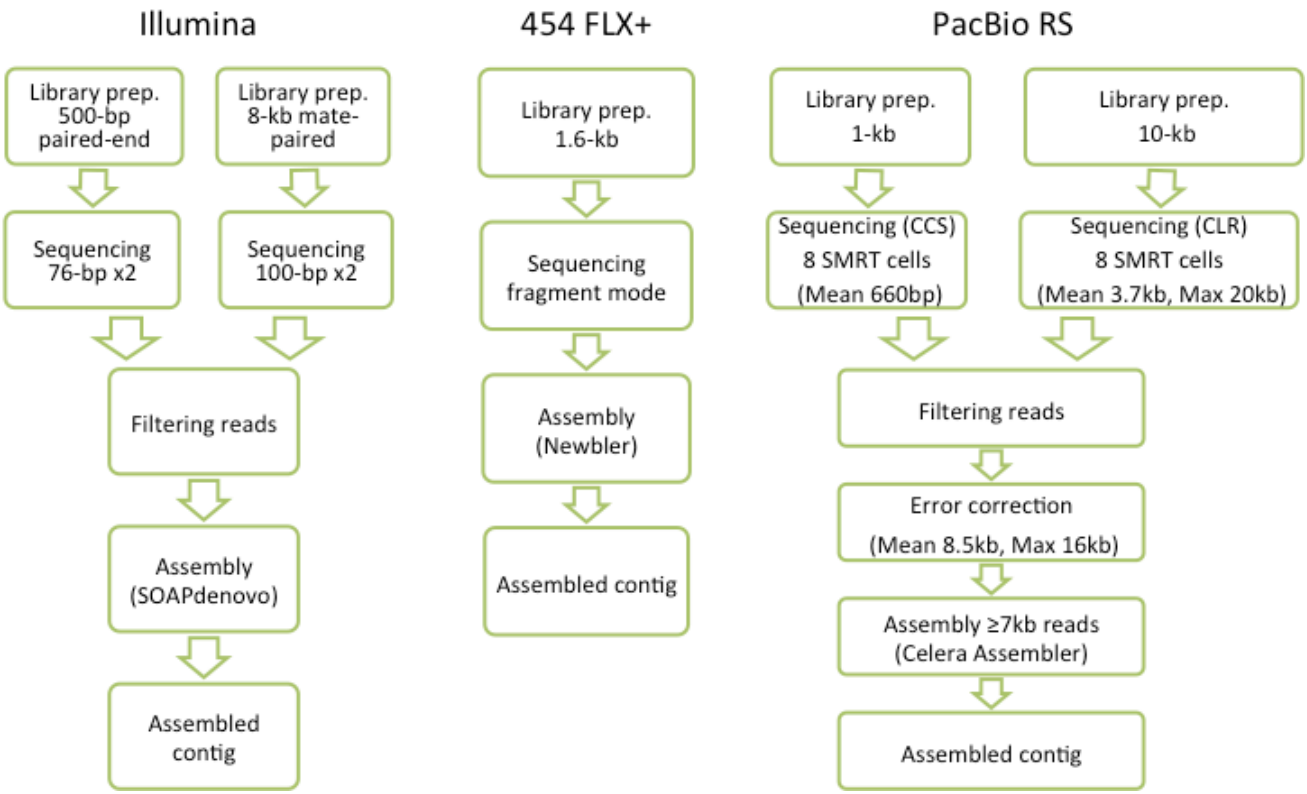

Figure 3. Workflow of genome assembly of *E. mundtii* QU 25 using three sequencing technologies

For Illumina sequencing, two types of library were prepared, a 500-bp insert size paired-end library and an 8-kb insert size mate-paired library. After sequencing, filtered reads were assembled using SOAPdenovo. For 454 FLX+ sequencing, a 1.6-kb size fragment library was prepared, and assembled using Newbler. For PacBio RS sequencing, two type of library were prepared, a 1-kb library and a 10-kb library. After sequencing using two methods (CCS and CLR), filtering reads were used for error correction of CLR, generating corrected long reads. Long reads with over 7 kb were assembled using Celera Assembler.

Table 1. Assembly statistics summary for the three different sequencing technologies

Technology	Illumina Genome Analyzer II	Roche/454 GS-FLX+	PacBio RS
Sequenced library	400X PE 500bp + 30X Mate 8kb	40x Fragment	234x CLR 10kb 57x CCS 1kb
Assembler	SOAPdenovo	Newbler	SMRT Pipe(AHA)
Read Length	100 bp	Avg 455bp	Avg 3.7kb Max 20 kb
Total Scaffolds	238	60	5
Total Contigs	310	60	5
N50 Contigs (bp)	81,003	133,269	3,022,186
Max Contig (bp)	172,673	133,269	3,022,186
Total num of N's in scaffolds	85,141	0	0
Total bp	3,274,156	3,262,871	3,348,476

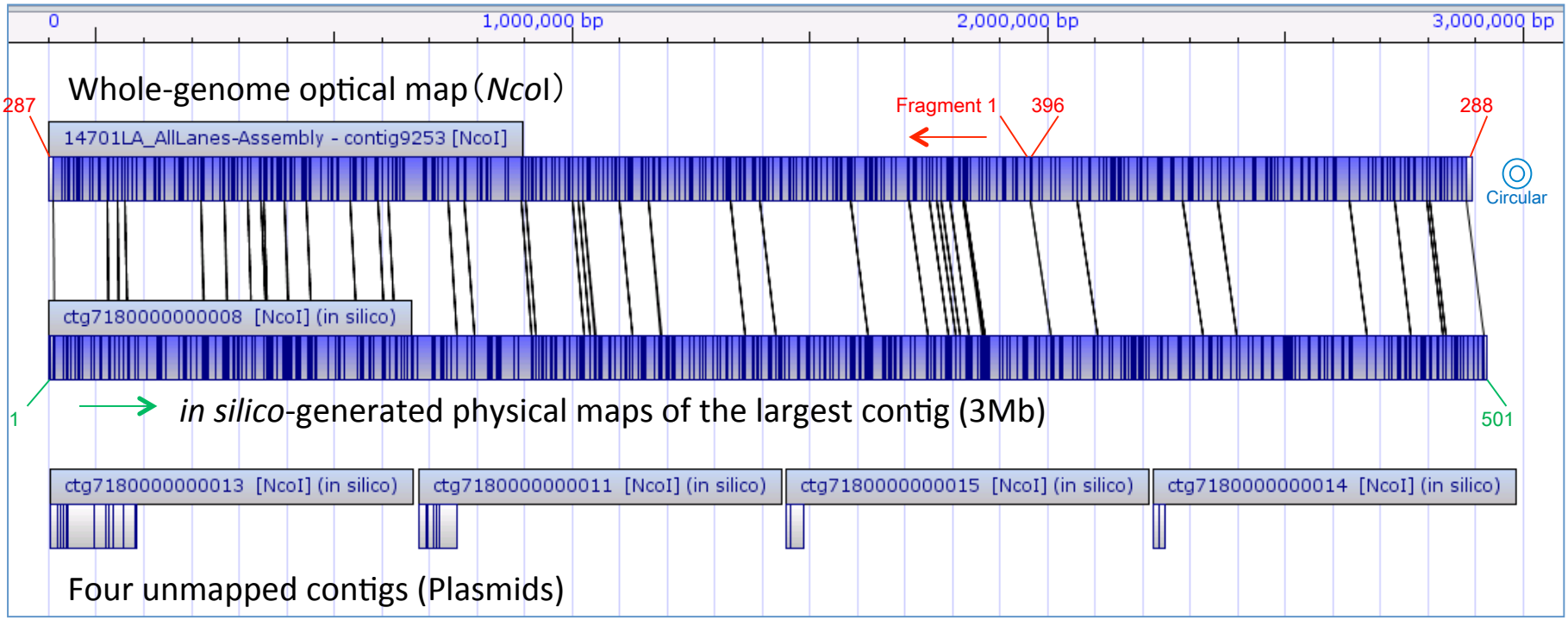

Figure 4. Comparison of the *Nco*I-digest of the whole-genome optical map and *in silico*-generated physical maps of contigs

Upper rectangle represents *Nco*I-digest of the whole-genome (chromosome) optical map. Vertical black line indicates *Nco*I-digested position. Middle rectangle represents *in silico*-generated physical maps of the largest contig (3Mb). Lines between maps indicate the position of identical sequences on the two maps, and can be used to visually identify misassembles and inversions. The whole-genome optical map also confirms circular configuration of the chromosome. Bottom small rectangles are non-aligned contigs to the chromosome, indicating they are plasmids.

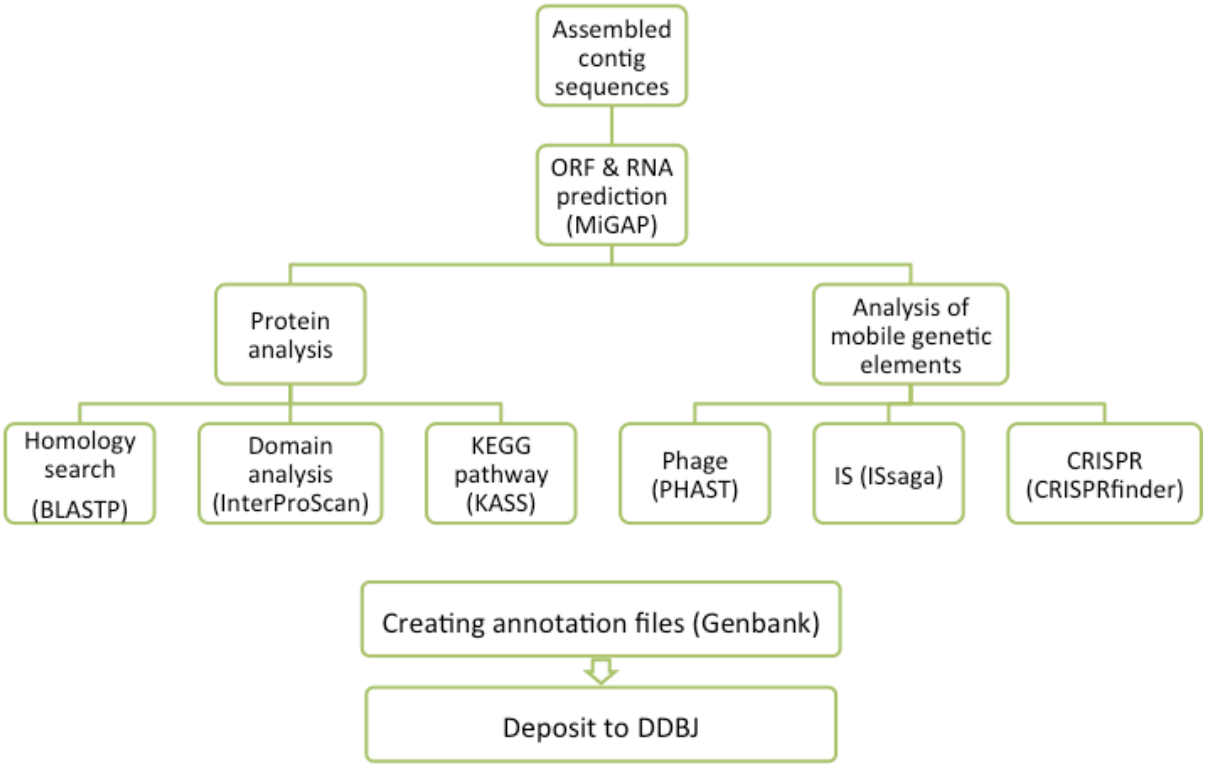

Figure 5. Workflow of genome annotation

Prediction of ORF and RNA from assembled contig sequences was performed using MiGAP (Microbial Genome Annotation Pipeline). After prediction of ORF, they were further analyzed using homology search, domain analysis, and KEGG pathway. For analysis of mobile genetic elements, phage, insertion sequence (IS), and CRISPR were predicted using third-party web services. Programs or web services used for each analysis are noted in brackets. After each analysis, annotations were integrated as Genbank format, and they were deposited to DDBJ (DNA Data Bank of Japan).

Table 2. General features of the *Enterococcus mundtii* QU 25 genome

Features	Chromosome	pQY182	pQY082	pQY039	pQY024	pQY003
Size (bp)	3,022,186	181,920	82,213	38,528	23,629	2,584
G+C content (%)	38.6	36.2	35.8	33.8	35.3	38.9
No. of rRNA operons	6	0	0	0	0	0
No. of tRNA genes	63	0	0	0	0	0
CDS (protein coding)	2,900	178	82	36	21	4
Avg. of CDS length (bp)	884	813	818	850	903	436
Estimated copy number of replicon	1	1	1	1	1	5

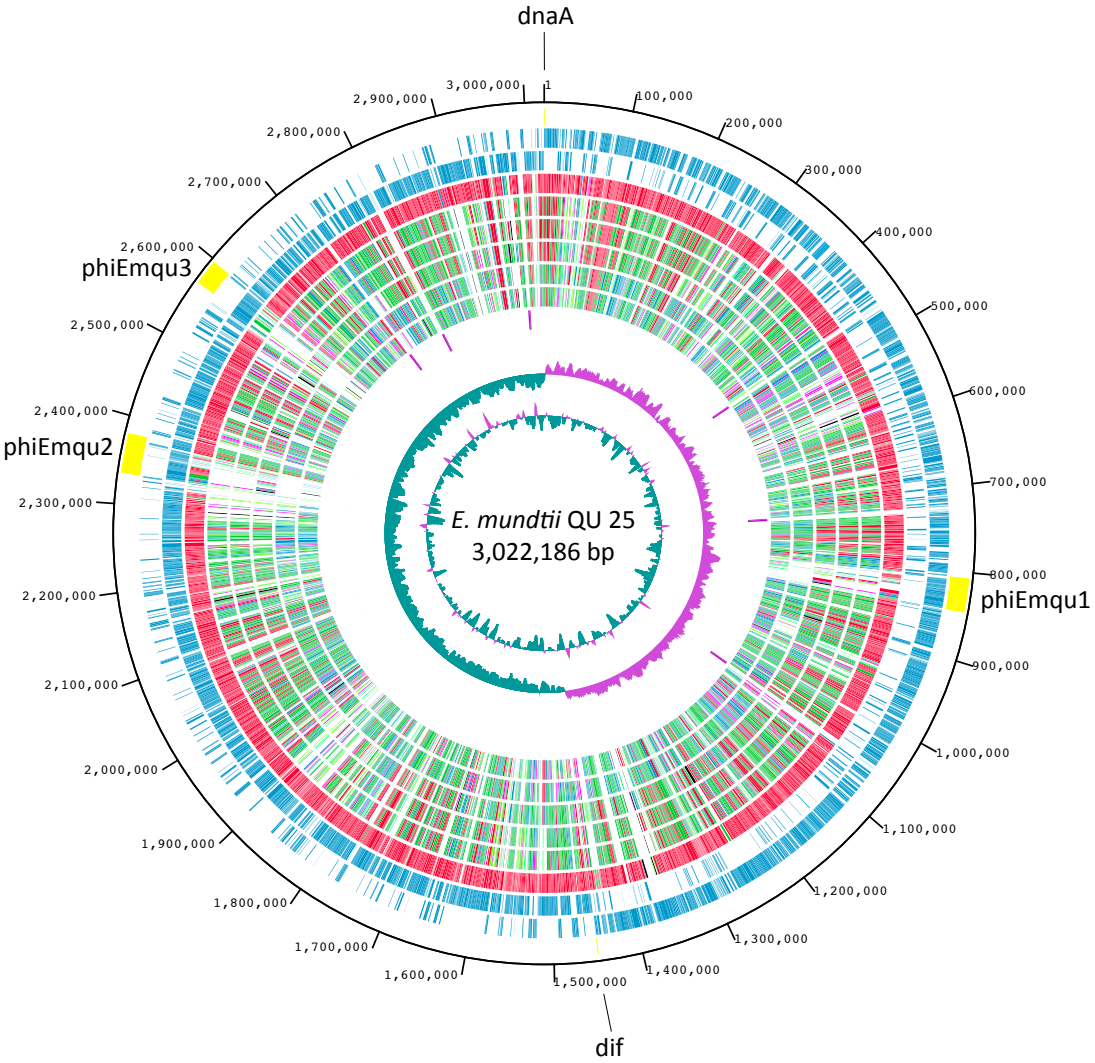

Figure 6. Circular map of *Enterococcus mundtii* QU 25 complete genome

Genome map of the QU 25 strain. In the outermost circle, three prophages of phiEmqu1, phiEmqu2 and phiEmqu3, replication origin (*dnaA*), and terminus (*dif*) are shown. In the second circle, the ORFs transcribed in a clockwise manner are shown as bars. The third circle shows ORFS transcribed in a counter-clockwise manner. The fourth to ninth circles depict the results of ortholog analyses (BLASTP E-value $\leq 1 \times 10^{-10}$) with *E. mundtii* ATCC 882, *E. faecium* DO, *E. faecium* Aus0004, *E. hirae* ATCC 9790, *E. casseliflavus* EC20, and *E. faecalis* V583, respectively. The extent of homology relative to QU 25 is depicted using a

heat map of arbitrarily chosen bins. The colour scheme and percentage identity for orthologs are as follows: red, orthologs with >90% identity; green, 70–90% identity; blue, 50–70% identity; black, <50% identity. The tenth circle shows the positions of rRNA operons. The last two (innermost) circles represent G+C content (purple > 39.5% average; green < 39.5% average; range from 32 to 47%) and G+C skew, both calculated for a 10-kb window with 1-kb stepping.

Table 3. The number of orthologous genes between *E. mundtii* QU 25 and each *Enterococci* species

	E. casseliflavus EC20	*E. faecalis* V583	*E. faecium* Aus0004	*E. faecium* DO	*E. hirae* ATCC 9790	*E. mundtii* ATCC 882
No. of orthologous genes*	1,815	1,714	1,918	1,927	1,892	2,595
Average identity (%)	66.7	66.2	78.8	78.8	78.0	98.3

*Numbers of orthologous genes between *E. mundtii* QU 25 and each *Enterococci* species were calculated by the best-hit analysis using BLASTP program with a threshold E-value of 1×10^{-10}.

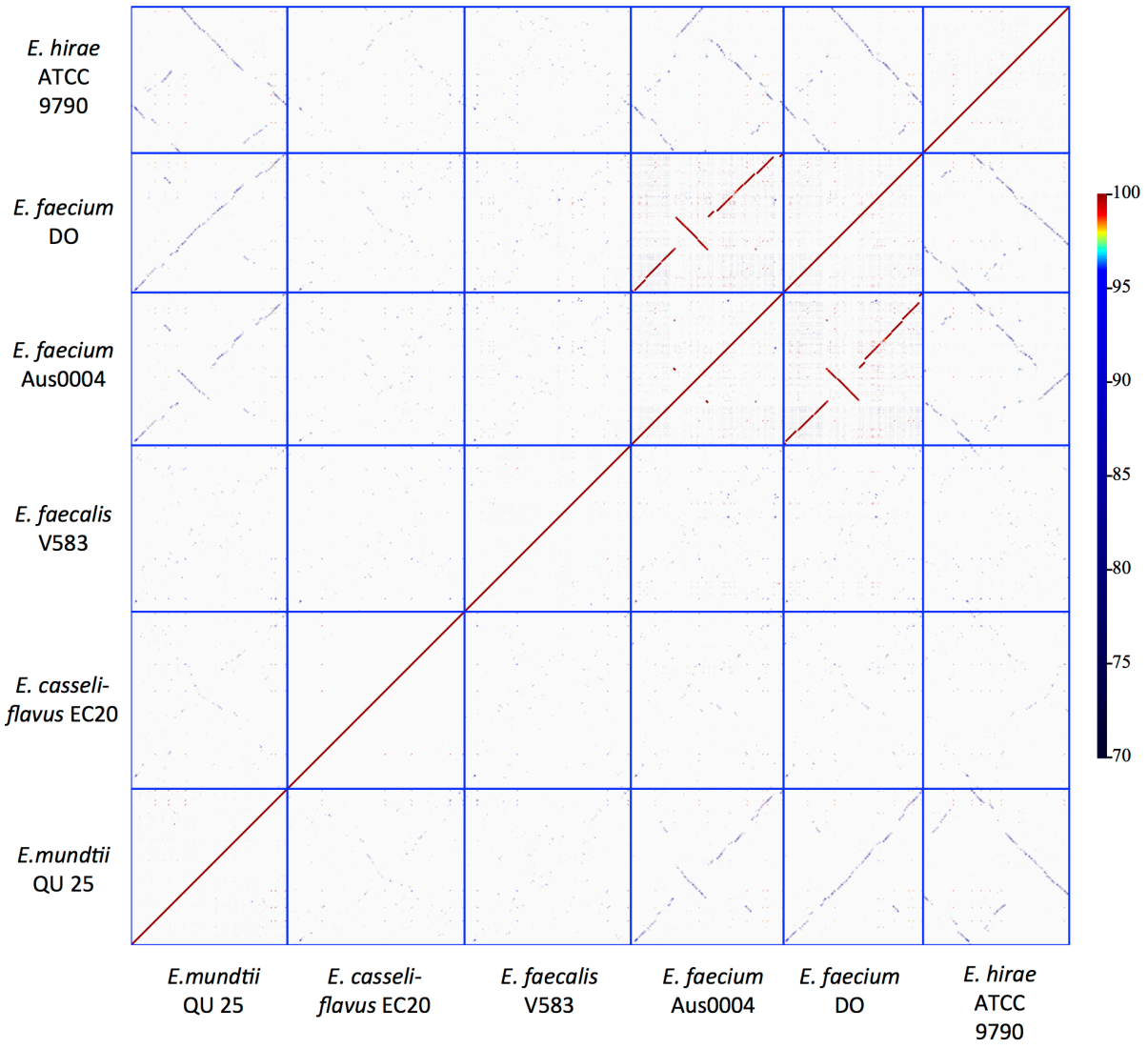

Figure 7. Dot plots comparison of the six *Enterococcus* species

Dot plots were generated using BLASTN and GenomeMatcher software. Color scale indicates the percentage of sequence homology.

Table 4. Prophage loci and genes on *E. mundtii* QU 25 genome

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
phiEmqu1	806547	845215			
	806547	806564	+	terminal direct repeat: CTCCTGTGACGTAAAAAA	
EMQU_0754	806636	807763	-	integrase	InterPro: IPR002104: Integrase, catalytic; IPR011010: DNA breaking-rejoining enzyme, catalytic core; IPR013762: Integrase-like, catalytic core; IPR023109: Integrase/recombinase, N-terminal; PHAST: PHAGE_Lactoc_bIL309
EMQU_0755	807878	808522	-	hypothetical protein	
EMQU_0756	808613	809452	-	putative S24-like peptidase	InterPro: IPR001387: Helix-turn-helix type 3; IPR010982: Lambda repressor-like, DNA-binding; IPR011056: Peptidase S24/S26A/S26B/S26C, beta-ribbon domain; IPR015927: Peptidase S24/S26A/S26B/S26C; IPR019759: Peptidase S24/S26A/S26B; PHAST: PHAGE_Geobac_E2
EMQU_0757	809625	809858	+	Cro-like protein phage associated	InterPro: IPR001387: Helix-turn-helix type 3; IPR010982: Lambda repressor-like, DNA-binding; PHAST: PHAGE_Lactob_AQ113
EMQU_0758	809919	810668	+	gp34	KO: K07741; InterPro: IPR013557: AntA/AntB antirepressor; PHAST: PHAGE_Brocho_BL3
EMQU_0759	810682	810996	+	hypothetical protein SPTP3101_gp10	InterPro: IPR008489: Bacteriophage bIL285, Orf7; PHAST: PHAGE_Staphy_1

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_0760	811009	811176	+	hypothetical protein	
EMQU_0761	811274	811609	+	hypothetical protein	
EMQU_0762	811606	811830	+	hypothetical protein	
EMQU_0763	811842	812195	+	hypothetical protein EFJG_00046	
EMQU_0764	812188	812379	+	hypothetical protein	
EMQU_0765	812382	813404	+	RecT protein	KO: K07455; InterPro: IPR018330: DNA single-strand annealing protein RecT-like; PHAST: PHAGE_Lactob_LF1
EMQU_0766	813367	814182	+	hypothetical protein	InterPro: IPR016974: Uncharacterised phage-associated protein; IPR024432: Putative exodeoxyribonuclease 8, PDDEXK-like domain
EMQU_0767	814200	814991	+	DNA replication protein	PHAST: PHAGE_Lactob_Lrm1
EMQU_0768	814991	815839	+	putative DnaC protein	KO: K02315; InterPro: IPR002611: IstB-like ATP-binding protein; IPR003593: ATPase, AAA+ type, core; PHAST: PHAGE_Bacill_WBeta
EMQU_0769	815836	816078	+	hypothetical protein	
EMQU_0770	816075	816263	+	hypothetical protein	
EMQU_0771	816221	816667	+	Orf19	InterPro: IPR009414: Bacteriophage 92, Orf34; PHAST: PHAGE_Lactoc_bIL286
EMQU_0772	816688	817122	-	hypothetical protein	
EMQU_0773	817174	817746	+	gp58	PHAST: PHAGE_Lister_B054
EMQU_0774	817746	817934	+	hypothetical protein	

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_0775	817945	818139	+	hypothetical protein	
EMQU_0776	818136	818267	+	hypothetical protein	
EMQU_0777	818264	818659	+	YopX superfamily protein	InterPro: IPR010024: Conserved hypothetical protein CHP1671; IPR019096: YopX protein; IPR023385: YopX-like domain, beta barrel type; PHAST: PHAGE_Entero_phiFL2A
EMQU_0778	818656	819714	+	putative DNA methylase	KO: K00558: DNA (cytosine-5-)-methyltransferase [EC:2.1.1.37]; InterPro: IPR001525: C-5 cytosine methyltransferase; IPR018117: DNA methylase, C-5 cytosine-specific, active site; PHAST: PHAGE_Strept_5093
EMQU_0779	819728	819943	+	hypothetical protein	
EMQU_0780	819936	820511	+	phage protein, putative	InterPro: IPR012865: Protein of unknown function DUF1642
EMQU_0781	820508	820729	+	hypothetical protein	
EMQU_0782	820726	821082	+	hypothetical protein	
EMQU_0783	821072	821269	+	hypothetical protein	
EMQU_0784	821269	821463	+	hypothetical protein	
EMQU_0785	821814	822290	+	conserved phage protein	InterPro: IPR013249: RNA polymerase sigma factor 70, region 4 type 2; IPR013324: RNA polymerase sigma factor, region 3/4; PHAST: PHAGE_Bacill_WBeta
EMQU_0786	822498	823568	+	hypothetical protein SGHV062	PHAST: PHAGE_Glossi_virus
EMQU_0787	823671	823898	+	hypothetical protein	
EMQU_0788	823895	824194	+	hypothetical protein	

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_0789	824181	824522	+	HNH endonuclease domain protein	InterPro: IPR002711: HNH endonuclease; IPR003615: HNH nuclease; PHAST: PHAGE_Lactob_LF1
EMQU_0790	824680	825144	+	phage terminase small subunit	InterPro: IPR006448: Streptococcus phage 7201, Orf21; PHAST: PHAGE_Lactob_LF1
EMQU_0791	825183	826925	+	bacteriophage terminase large subunit	InterPro: IPR005021: Bacteriophage bIL285, Orf41, terminase; PHAST: PHAGE_Lactob_LF1
EMQU_0792	826940	827110	+	hypothetical protein	
EMQU_0793	827137	828399	+	bacteriophage portal protein	InterPro: IPR006427: Bacteriophage 16-3, portal protein; IPR006944: Bacteriophage/Gene transfer agent portal protein; PHAST: PHAGE_Lactob_LF1
EMQU_0794	828335	828955	+	phage head maturation protease	KO: K06904; InterPro: IPR006433: Prohead protease, HK97 family; PHAST: PHAGE_Lactob_LF1
EMQU_0795	829013	830227	+	phage capsid protein	InterPro: IPR006444: Bacteriophage 16-3, major capsid protein; IPR024455: Caudovirus, capsid; PHAST: PHAGE_Lactob_LF1
EMQU_0796	830245	830538	+	hypothetical protein	
EMQU_0797	830525	830878	+	DNA packaging protein	InterPro: IPR006450: Bacteriophage HK022, Gp6; IPR021146: Bacteriophage QLRG family, putative DNA packaging; PHAST: PHAGE_Lactob_LF1
EMQU_0798	830865	831200	+	head-tail joining protein	InterPro: IPR008767: Bacteriophage SPP1, head-tail adaptor; PHAST: PHAGE_Lactob_LF1
EMQU_0799	831197	831562	+	head-tail joining protein	InterPro: IPR010064: Bacteriophage HK97, Gp10; PHAST:

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
					PHAGE_Lactob_LF1
EMQU_0800	831559	831975	+	head-tail joining protein	PHAST: PHAGE_Lactob_LF1
EMQU_0801	832016	832594	+	major tail protein	InterPro: IPR006490: Bacteriophage bIL285, Orf50, major tail protein; IPR006724: Bacteriophage bIL286, Orf50, major tail; PHAST: PHAGE_Lactob_LF1
EMQU_0802	832665	832991	+	hypothetical protein SpyM3_0934	PHAST: PHAGE_Strept_2
EMQU_0803	833009	833173	+	hypothetical protein	
EMQU_0804	833185	837780	+	phage tail tape measure protein	InterPro: IPR010090: Bacteriophage bIL285, Orf52, tail tape measure protein; PHAST: PHAGE_Lactob_LF1
EMQU_0805	837777	838598	+	phage tail family protein	InterPro: IPR008841: Siphovirus tail component; PHAST: PHAGE_Staphy_SMSAP5
EMQU_0806	838607	839635	+	phage-associated protein/endopeptidase	InterPro: IPR010572: Bacteriophage 53, Orf003; PHAST: PHAGE_Lactob_LF1
EMQU_0807	839635	840681	+	gp19	InterPro: IPR011050: Pectin lyase fold/virulence factor; IPR012334: Pectin lyase fold; PHAST: PHAGE_Lister_A500
EMQU_0808	840682	842868	+	hypothetical protein	InterPro: IPR018913: Domain of unknown function DUF2479
EMQU_0809	842881	843228	+	hypothetical protein	
EMQU_0810	843221	843349	+	hypothetical protein	InterPro: IPR010022: Protein of unknown function XkdX
EMQU_0811	843360	843593	+	hypothetical protein EfmU0317_0103	

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_0812	843595	843852	+	holin	InterPro: IPR009708: Bacteriophage A118, holin; PHAST: PHAGE_Lister_A118
EMQU_0813	843921	844814	+	putative N-acetylmuramoyl-L-alanine amidase	InterPro: IPR002502: N-acetylmuramoyl-L-alanine amidase domain; PHAST: PHAGE_Pseudo_phi15
	845198	845215	+	terminal direct repeat: CTCCTGTGACGTAAAAAA	
phiEmqu2	2327297	2375151			
	2327297	2327315	+	terminal direct repeat: GTGGCAAATTTGTGGCAAA	
EMQU_2239	2328181	2329473	+	D-serine dehydratase	KO: K01753: D-serine dehydratase [EC:4.3.1.18]; InterPro: IPR000634: Serine/threonine dehydratase, pyridoxal-phosphate-binding site; IPR001926: Pyridoxal phosphate-dependent enzyme, beta subunit; IPR011780: D-serine ammonia-lyase
EMQU_2240	2329599	2329931	-	hypothetical protein	
EMQU_2241	2331190	2331336	-	hypothetical protein	
EMQU_2242	2333101	2333745	+	chitin binding protein	KO: K03933; InterPro: IPR004302: Chitin-binding, domain 3; IPR014756: Immunoglobulin E-set
EMQU_2243	2333945	2334787	+	TIR protein	InterPro: IPR000157: Toll/interleukin-1 receptor homology (TIR) domain
EMQU_2244	2334776	2335012	-	hypothetical protein EFJG_02379	

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_2245	2335068	2335961	-	putative N-acetylmuramoyl-L-alanine amidase	InterPro: IPR002502: N-acetylmuramoyl-L-alanine amidase domain; PHAST: PHAGE_Pseudo_phi15
EMQU_2246	2336030	2336287	-	holin	InterPro: IPR009708: Bacteriophage A118, holin; PHAST: PHAGE_Lister_A118
EMQU_2247	2336289	2336522	-	hypothetical protein EfmU0317_0103	
EMQU_2248	2336533	2336661	-	hypothetical protein	InterPro: IPR010022: Protein of unknown function XkdX
EMQU_2249	2336654	2337001	-	hypothetical protein	
EMQU_2250	2337014	2339200	-	hypothetical protein	InterPro: IPR018913: Domain of unknown function DUF2479
EMQU_2251	2339201	2340247	-	gp19	InterPro: IPR011050: Pectin lyase fold/virulence factor; IPR012334: Pectin lyase fold; PHAST: PHAGE_Lister_A500
EMQU_2252	2340247	2341275	-	gp18	InterPro: IPR010572: Bacteriophage 53, Orf003; PHAST: PHAGE_Lister_A500
EMQU_2253	2341284	2342114	-	tail protein	InterPro: IPR008841: Siphovirus tail component; PHAST: PHAGE_Staphy_StB20
EMQU_2254	2342107	2347362	-	tail protein	InterPro: IPR013491: Caudovirus, tape measure, N-terminal; PHAST: PHAGE_Temper_1
EMQU_2255	2347355	2347924	-	hypothetical protein BCBBV1cgp48	InterPro: IPR009660: Bacteriophage A500, Gp15; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2256	2347933	2348394	-	hypothetical protein	PHAST: PHAGE_Bacill_BCJA1c

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
				BCBBV1cgp47	
EMQU_2257	2348422	2348583	-	hypothetical protein	
EMQU_2258	2348538	2349014	-	tail shaft protein	PHAST: PHAGE_Bacill_BCJA1c
EMQU_2259	2349025	2349402	-	hypothetical protein BCBBV1cgp45	InterPro: IPR024411: Minor capsid protein, bacteriophage; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2260	2349402	2349764	-	hypothetical protein BCBBV1cgp44	InterPro: IPR021080: Minor capsid protein; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2261	2349764	2350099	-	hypothetical protein BCBBV1cgp43	InterPro: IPR019612: Minor capsid protein, putative; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2262	2350096	2350551	-	hypothetical protein BCBBV1cgp42	PHAST: PHAGE_Bacill_BCJA1c
EMQU_2263	2350578	2350742	-	hypothetical protein	
EMQU_2264	2350739	2351626	-	major capsid protein gp34	PHAST: PHAGE_Lactob_H
EMQU_2265	2351640	2352230	-	scaffold protein	InterPro: IPR009636: Bacteriophage mv4, Gp20; PHAST: PHAGE_Entero_phiFL4A
EMQU_2266	2352464	2353612	-	hypothetical protein BCBBV1cgp37	InterPro: IPR009319: Bacteriophage A118, Gp4, minor capsid; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2267	2353617	2354954	-	portal	InterPro: IPR006432: Portal protein, putative, A118-type; IPR021145: Portal protein; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2268	2355003	2355155	-	putative minor capsid protein 1	PHAST: PHAGE_Strept_MM1
EMQU_2269	2355170	2356456	-	terminase large subunit	InterPro: IPR004921: Terminase, large subunit; IPR006437: Bacteriophage terminase, large subunit; PHAST:

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
					PHAGE_Bacill_PBC1
EMQU_2270	2356440	2357318	-	terminase small subunit	InterPro: IPR018925: Enterococcus phage phiFL1A, terminase small subunit; PHAST: PHAGE_Entero_phiFL1A
EMQU_2271	2357375	2357854	-	hypothetical protein	
EMQU_2272	2358038	2358280	-	hypothetical protein	
EMQU_2273	2358504	2359004	-	putative N-acetylmuramoyl-L-alanine amidase	InterPro: IPR002502: N-acetylmuramoyl-L-alanine amidase domain; PHAST: PHAGE_Entero_phiEF24C
EMQU_2274	2359754	2359975	+	hypothetical protein	
EMQU_2275	2360037	2360219	-	hypothetical protein	
EMQU_2276	2360750	2361166	-	transcriptional regulator ArpU family	InterPro: IPR006524: Transcription activator, ArpU family; PHAST: PHAGE_Entero_phiFL1A
EMQU_2277	2361247	2361684	-	hypothetical protein HMPREF9495_01321	
EMQU_2278	2362324	2362446	-	hypothetical protein	
EMQU_2279	2362449	2362838	-	hypothetical protein	
EMQU_2280	2362835	2363125	-	hypothetical protein	
EMQU_2281	2363115	2363663	-	Gp35 protein	InterPro: IPR012865: Protein of unknown function DUF1642; PHAST: PHAGE_Lister_2389
EMQU_2282	2363660	2363860	-	hypothetical protein	

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_2283	2363857	2364324	-	DNA cytosine methyltransferase	InterPro: IPR013216: Methyltransferase type 11; PHAST: PHAGE_Entero_phiFL1A
EMQU_2284	2364340	2364537	-	hypothetical protein	
EMQU_2285	2364964	2365341	-	ORF029	InterPro: IPR019096: YopX protein; IPR023385: YopX-like domain, beta barrel type; PHAST: PHAGE_Staphy_2638A
EMQU_2286	2365338	2365469	-	hypothetical protein	
EMQU_2287	2365481	2365702	-	Hypothetical protein EfmE4452_2680	
EMQU_2288	2365709	2366116	-	hypothetical protein	
EMQU_2289	2366119	2366616	-	hypothetical protein	
EMQU_2290	2366603	2366923	-	hypothetical protein	
EMQU_2291	2366917	2367081	-	hypothetical protein	
EMQU_2292	2367071	2368033	-	gp46	InterPro: IPR006343: Replication protein, DnaD/DnaB domain; IPR010056: Bacteriophage A500, Gp45, replisome organiser, N-terminal; PHAST: PHAGE_Brocho_BL3
EMQU_2293	2368052	2368738	-	conserved hypothetical protein	InterPro: IPR010373: Protein of unknown function DUF968; PHAST: PHAGE_Entero_phiEf11
EMQU_2294	2368689	2369477	-	Orf14	InterPro: IPR009425: Single-strand annealing protein SAK3; PHAST: PHAGE_Lactoc_bIL285
EMQU_2295	2369470	2369736	-	hypothetical protein	
EMQU_2296	2369840	2370004	-	hypothetical protein	

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_2297	2370001	2370126	-	hypothetical protein	
EMQU_2298	2370265	2370453	-	hypothetical protein	
EMQU_2299	2370642	2370911	+	hypothetical protein	
EMQU_2300	2370908	2371306	-	hypothetical protein	
EMQU_2301	2371318	2372073	-	putative anti-repressor protein	KO: K07741; InterPro: IPR003497: BRO N-terminal domain; IPR005039: Bacteriophage P1, Ant1, C-terminal; PHAST: PHAGE_Staphy_phiMR25
EMQU_2302	2372091	2372399	-	hypothetical protein HMPREF0348_2593	InterPro: IPR008489: Bacteriophage bIL285, Orf7
EMQU_2303	2372403	2372606	-	hypothetical protein	
EMQU_2304	2372912	2373250	+	CI phage repressor protein	InterPro: IPR001387: Helix-turn-helix type 3; IPR010982: Lambda repressor-like, DNA-binding; PHAST: PHAGE_Entero_phiFL1A
EMQU_2305	2373260	2373664	+	hypothetical protein	InterPro: IPR010359: Protein of unknown function DUF955
EMQU_2306	2373718	2373900	+	hypothetical protein	
EMQU_2307	2373944	2374450	+	integrase	InterPro: IPR023109: Integrase/recombinase, N-terminal; PHAST: PHAGE_Strept_YMC_2011
EMQU_2308	2374447	2375115	+	putative integrase	KO: K14059; InterPro: IPR002104: Integrase, catalytic; IPR011010: DNA breaking-rejoining enzyme, catalytic core; IPR013762: Integrase-like, catalytic core; PHAST: PHAGE_Clostr_phiC2
	2375133	2375151	+	terminal direct repeat: GTGGCAAATTTGTGGCAAA	
phiEmqu3	2556843	2597594			

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
	2556843	2556862	+	terminal direct repeat: TAAAATTATTTAACTGTTAC	
EMQU_2478	2557383	2557706	-	hypothetical protein	
EMQU_2479	2558030	2558395	+	hypothetical protein	
EMQU_2480	2558984	2559577	-	hypothetical protein FN1087	
EMQU_2481	2559567	2560919	-	Transporter	KO: K06926; InterPro: IPR007406: Prokaryotic chromosome segregation/condensation protein MukB, N-terminal
EMQU_2482	2561039	2561932	-	N-acetylmuramoyl-L-alanine amidase	InterPro: IPR002502: N-acetylmuramoyl-L-alanine amidase domain
EMQU_2483	2562001	2562258	-	conserved hypothetical protein	InterPro: IPR009708: Bacteriophage A118, holin
EMQU_2484	2562260	2562493	-	hypothetical protein EfmU0317_0103	
EMQU_2485	2562526	2564145	-	hypothetical protein	InterPro: IPR018913: Domain of unknown function DUF2479
EMQU_2486	2564206	2566005	-	hypothetical protein	
EMQU_2487	2566009	2568336	-	hypothetical protein	InterPro: IPR007119: Phage minor structural protein N-terminal domain; IPR010572: Bacteriophage 53, Orf003
EMQU_2488	2568333	2569088	-	phage putative tail component protein	InterPro: IPR008841: Siphovirus tail component
EMQU_2489	2569081	2574003	-	putative minor tail protein	InterPro: IPR013491: Caudovirus, tape measure, N-terminal; PHAST: PHAGE_Strept_MM1
EMQU_2490	2573990	2574571	-	hypothetical protein MM1p45	InterPro: IPR009660: Bacteriophage A500, Gp15; PHAST: PHAGE_Strept_MM1

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_2491	2574576	2574956	-	hypothetical protein MM1p44	PHAST: PHAGE_Strept_MM1
EMQU_2492	2575011	2575529	-	putative major tail shaft protein	PHAST: PHAGE_Strept_MM1
EMQU_2493	2575530	2575928	-	putative minor capsid protein 4	InterPro: IPR024411: Minor capsid protein, bacteriophage; PHAST: PHAGE_Strept_MM1
EMQU_2494	2575928	2576293	-	putative minor capsid protein 3	InterPro: IPR021080: Minor capsid protein; PHAST: PHAGE_Strept_MM1
EMQU_2495	2576293	2576628	-	hypothetical protein BCBBV1cgp43	InterPro: IPR019612: Minor capsid protein, putative; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2496	2576625	2577080	-	hypothetical protein BCBBV1cgp42	PHAST: PHAGE_Bacill_BCJA1c
EMQU_2497	2577111	2578220	-	major capsid protein	InterPro: IPR024455: Caudovirus, capsid; PHAST: PHAGE_Lister_A118
EMQU_2498	2578233	2578799	-	minor capsid protein	InterPro: IPR009636: Bacteriophage mv4, Gp20; PHAST: PHAGE_Lactob_phig1e
EMQU_2499	2579104	2580252	-	minor capsid protein	InterPro: IPR009319: Bacteriophage A118, Gp4, minor capsid; PHAST: PHAGE_Lactob_phig1e
EMQU_2500	2580257	2581822	-	portal	InterPro: IPR006432: Portal protein, putative, A118-type; IPR021145: Portal protein; PHAST: PHAGE_Bacill_BCJA1c
EMQU_2501	2581834	2583150	-	putative large terminase subunit	KO: K06909; InterPro: IPR006437: Bacteriophage terminase, large subunit; IPR006701: Caudovirales, terminase large subunit; PHAST: PHAGE_Strept_MM1

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_2502	2583147	2583857	-	gp1	PHAST: PHAGE_Brocho_BL3
EMQU_2503	2583898	2584176	-	hypothetical protein	
EMQU_2504	2584205	2584387	-	hypothetical protein	
EMQU_2505	2584534	2584743	-	hypothetical protein	
EMQU_2506	2584917	2585336	-	phage autolysin transcriptional regulator ArpU family	InterPro: IPR006524: Transcription activator, ArpU family; PHAST: PHAGE_Entero_phiEf11
EMQU_2507	2585418	2585705	-	hypothetical protein	
EMQU_2508	2585702	2586004	-	hypothetical protein	
EMQU_2509	2586001	2586222	-	hypothetical protein	
EMQU_2510	2586219	2586905	-	phage protein, putative	InterPro: IPR012865: Protein of unknown function DUF1642
EMQU_2511	2586919	2588208	-	putative DNA methylase	KO: K00558: DNA (cytosine-5-)-methyltransferase [EC:2.1.1.37]; InterPro: IPR001525: C-5 cytosine methyltransferase; IPR018117: DNA methylase, C-5 cytosine-specific, active site; PHAST: PHAGE_Strept_MM1
EMQU_2512	2588211	2588558	-	YopX superfamily protein	InterPro: IPR019096: YopX protein; IPR023385: YopX-like domain, beta barrel type; PHAST: PHAGE_Entero_phiFL2A
EMQU_2513	2588555	2588857	-	hypothetical protein	
EMQU_2514	2588850	2589014	-	hypothetical protein	
EMQU_2515	2589004	2589297	-	hypothetical protein	
EMQU_2516	2589310	2590032	-	putative DNA replication protein	InterPro: IPR011991: Winged helix-turn-helix transcription repressor DNA-binding; PHAST: PHAGE_Lactob_c5

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_2517	2590036	2590722	-	conserved hypothetical protein	InterPro: IPR010373: Protein of unknown function DUF968; PHAST: PHAGE_Entero_phiEf11
EMQU_2518	2590755	2591639	-	putative replication protein	InterPro: IPR009785: Lactobacillus prophage Lj928, Orf309; PHAST: PHAGE_Entero_phiEf11
EMQU_2519	2591640	2592332	-	putative single-strand DNA binding protein	InterPro: IPR007499: ERF; PHAST: PHAGE_Entero_phiEf11
EMQU_2520	2592478	2592642	-	hypothetical protein	
EMQU_2521	2592639	2592764	-	hypothetical protein	
EMQU_2522	2592911	2593153	-	hypothetical protein	
EMQU_2523	2593193	2593420	-	hypothetical protein	
EMQU_2524	2593432	2594157	-	anti-repressor	InterPro: IPR018873: KilA-N, DNA-binding domain; IPR018878: Bacteriophage bIL285, Orf6, C-terminal; PHAST: PHAGE_Lactoc_bIL285
EMQU_2525	2594224	2594508	-	hypothetical protein	
EMQU_2526	2594508	2594702	-	toxin-antitoxin system, antitoxin component, Xre family	InterPro: IPR010982: Lambda repressor-like, DNA-binding
EMQU_2527	2594806	2595009	+	hypothetical protein	
EMQU_2528	2595006	2595212	-	hypothetical protein	
EMQU_2529	2595516	2595863	+	repressor protein	InterPro: IPR001387: Helix-turn-helix type 3; IPR010982: Lambda repressor-like, DNA-binding; PHAST: PHAGE_Strept_MM1

Locus Tag	Start	End	Strand	Predicted Gene Product	Note
EMQU_2530	2595872	2596303	+	hypothetical protein MM1p03	InterPro: IPR010359: Protein of unknown function DUF955; PHAST: PHAGE_Strept_MM1
EMQU_2531	2596361	2597500	+	integrase	InterPro: IPR002104: Integrase, catalytic; IPR011010: DNA breaking-rejoining enzyme, catalytic core; IPR013762: Integrase-like, catalytic core; IPR023109: Integrase/recombinase, N-terminal; PHAST: PROPHAGE_Oceano_HTE831
	2597575	2597594	+	terminal direct repeat: TAAAATTATTTAACTGTTAC	

Table 5. Mobile elements in the *E. mundtii* QU 25 genome

Locus Tag	Start	End	Predicted Gene/Family
Chromosome			
EMQU_0167	167,691	168,935	ISCbt3 (IS607 family) transposase
EMQU_0168	168,913	169,287	ISCbt3 (IS607 family) resolvase
EMQU_0274	281,667	282,857	transposase like protein
EMQU_0329	343,456	344,751	ISEfa11 (ISL3 family) transposase
EMQU_0547	585,939	586,103	transposase like protein
EMQU_0548	586,320	587,471	ISEfa4 (IS200/IS605 family) transposase
EMQU_0622	659,557	660,564	ISEfa4 (IS200/IS605 family) transposase
EMQU_0623	660,872	661,372	ISEfa5 (ISL3 family) transposase
EMQU_0625	668,887	669,339	ISAac3 (IS200/IS605 family) transposase
EMQU_0867	894,133	895,428	ISEfa11 (ISL3 family) transposase
EMQU_1329	1,405,266	1,405,475	ISH7A (ISNCY family) transposase
EMQU_1442	1,515,200	1,515,487	transposase like protein
EMQU_1566	1,626,597	1,628,087	transposase like protein
EMQU_1579	1,636,717	1,638,012	ISEfa11 (ISL3 family) transposase
EMQU_1657	1,732,315	1,732,443	transposase like protein
EMQU_1804	1,874,140	1,874,724	ISBce13 (IS3 family) integrase
EMQU_1805	1,874,788	1,874,961	transposase like protein
EMQU_1806	1,875,012	1,875,194	ISBce13 (IS3 family) transposase
EMQU_1936	2,024,246	2,025,682	ISEnfa2 (IS1182 family) transposase
EMQU_2069	2,158,940	2,159,881	ISEnfa2 (IS1182 family) transposase
EMQU_2070	2,159,881	2,160,375	ISEnfa2 (IS1182 family) transposase
EMQU_2237	2,324,316	2,325,725	transposase like protein
EMQU_2469	2,548,026	2,549,321	ISEfa11 (ISL3 family) transposase
EMQU_2619	2,690,281	2,690,475	ISEfa12 (IS1182 family) transposase
EMQU_2784	2,886,612	2,888,132	transposase like protein
EMQU_2859	2,978,145	2,979,440	ISEfa11 (ISL3 family) transposase
pQY182			
EMQU_3010	118,891	119,145	ISBce13 (IS3 family) transposase
EMQU_3011	119,196	120,017	ISBce13 (IS3 family) integrase
EMQU_3055	154,453	155,754	IS1476 (ISL3 family) transposase
EMQU_3064	165,064	166,293	ISEnfa110 (IS110 family) transposase
pQY082			
EMQU_3088	5,781	7,100	IS1675 (IS4 family) transposase
EMQU_3094	13,439	14,758	IS1675 (IS4 family) transposase
EMQU_3155	74,049	75,368	IS1675 (IS4 family) transposase

Table 6. BLAST homology search against nr database for five plasmids

Description	Query cover	Identity	E value
pQY182			
Enterococcus mundtii plasmid pCRL10, partial sequence	5%	84%	0.0
Enterococcus faecium DO plasmid 1, complete sequence	2%	89%	0.0
Enterococcus faecium plasmid pM7M2, complete sequence	2%	89%	0.0
pQY082			
Enterococcus faecium Aus0004 plasmid AUS0004_p1, complete sequence	48%	89%	0.0
Enterococcus faecium DO plasmid 3, complete sequence	46%	86%	0.0
pQY039			
Enterococcus hirae ATCC 9790 plasmid pTG9790, complete sequence	23%	86%	0.0
pQY024			
Enterococcus faecalis strain E99 plasmid pBEE99, complete sequence	3%	89%	0.0
pQY003			
Enterococcus faecium strain 9631160-1 (AHA15) plasmid pRI1, complete sequence	45%	83%	0.0
Enterococcus faecium Aus0004 plasmid AUS0004_p3, complete sequence	42%	77%	0.0

Table 7. Bacteriocin activity and immunity of *Enterococcus mundtii* strains

	Bacteriocin producer [a]		
Indicator strain	*E. mundtii* QU 25	*E. mundtii* QU 2	*E. mundtii* JCM 8731^T
E. mundtii QU 25	–	–	–
E. mundtii QU 2	–	–	–
E. mundtii JCM 8731^T	+	+	–
E. faecalis JCM 5803^T	+	+	–
L. sakei JCM 1157^T	+	+	–

[a] +, growth inhibition of indicator strains; –, no inhibition of indicator strains

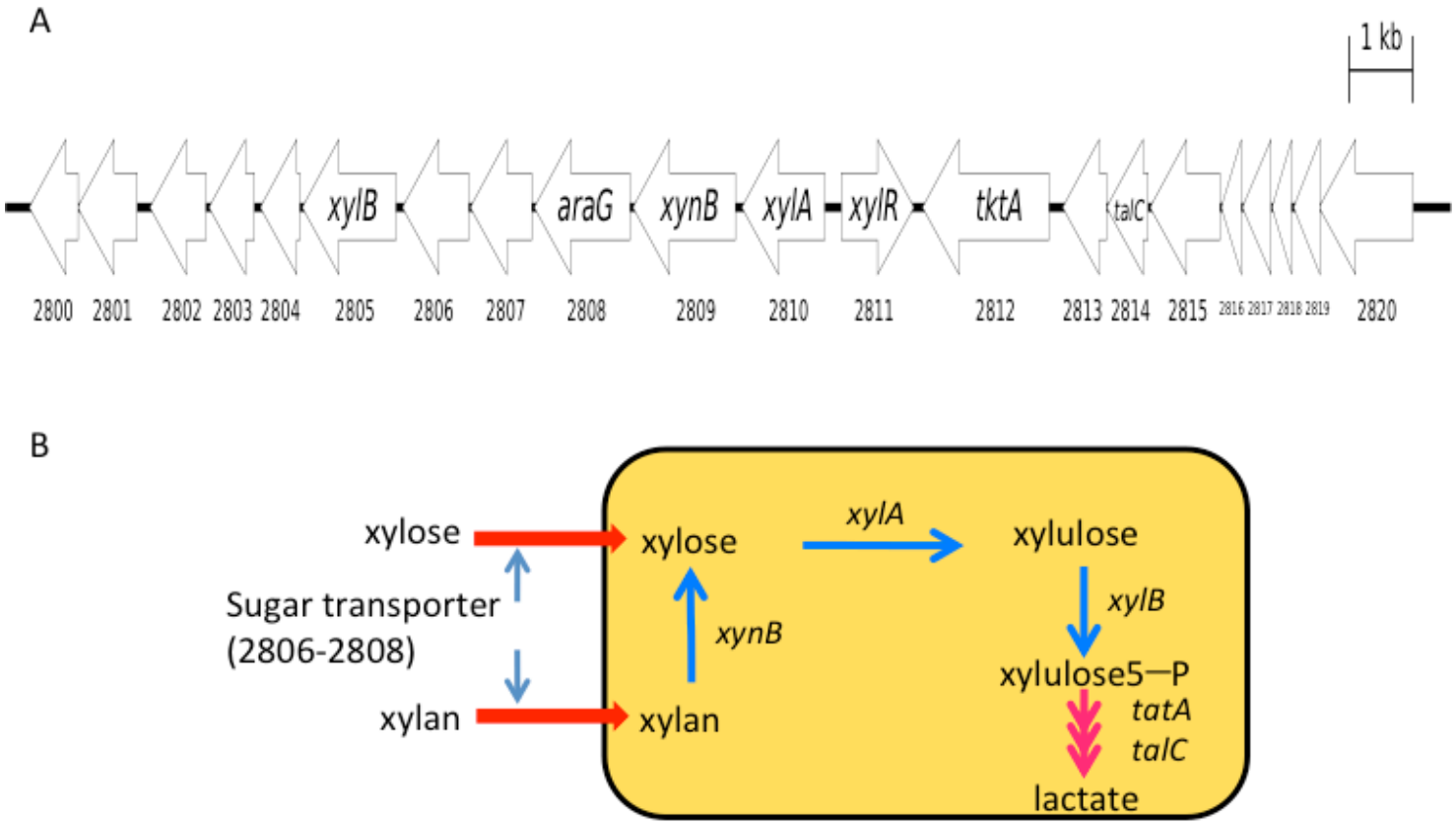

Figure 8. Gene clusters for early metabolism of xylose and the pentose phosphate pathway in *E. mundtii* QU 25

(A) Organization of the gene cluster. Numbers below arrows indicate feature code (EMQU_XXXX). (B) Metabolic pathway for lactic acid production from xylose and genes presented in the gene clusters.

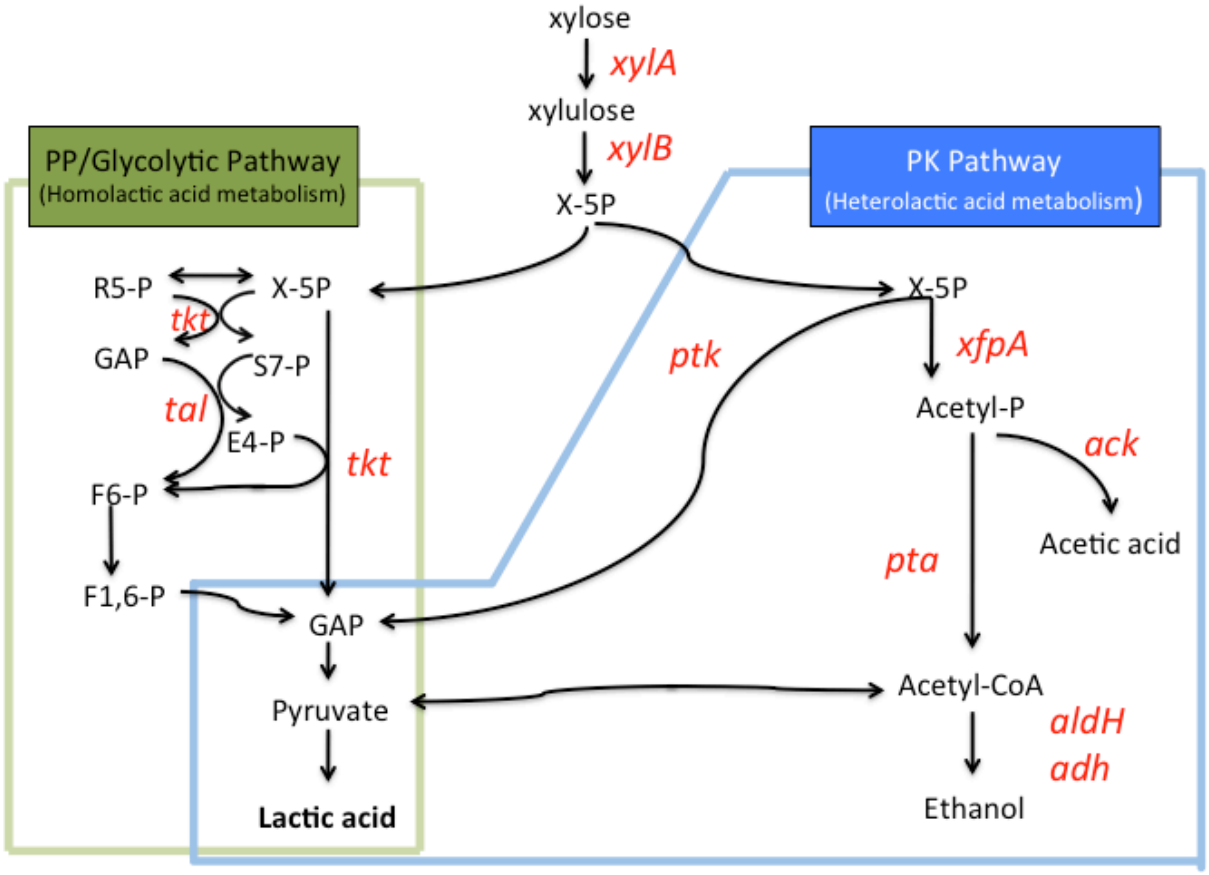

Figure 9. Metabolic pathways for lactic acid production from xylose and genes encoded by QU 25 genome

xylA, xylose isomerase (EMQU_2810); *xylB*, D-xylulose kinase (EMQU_2805); *tkt* transketolase (EMQU_1275, EMQU_2812); *tal*, transaldolase (EMQU_2814); *xfpA,* phosphoketolase (EMQU_1837); *ack*, acetate kinase (EMQU_2620); *pta*, phosphotransacetylase (EMQU_2119); *aldH*, acetaldehyde dehydrogenase (EMQU_2205); *adh*, alcohol dehydrogenase (EMQU_1129, EMQU_1829, and EMQU_2109)

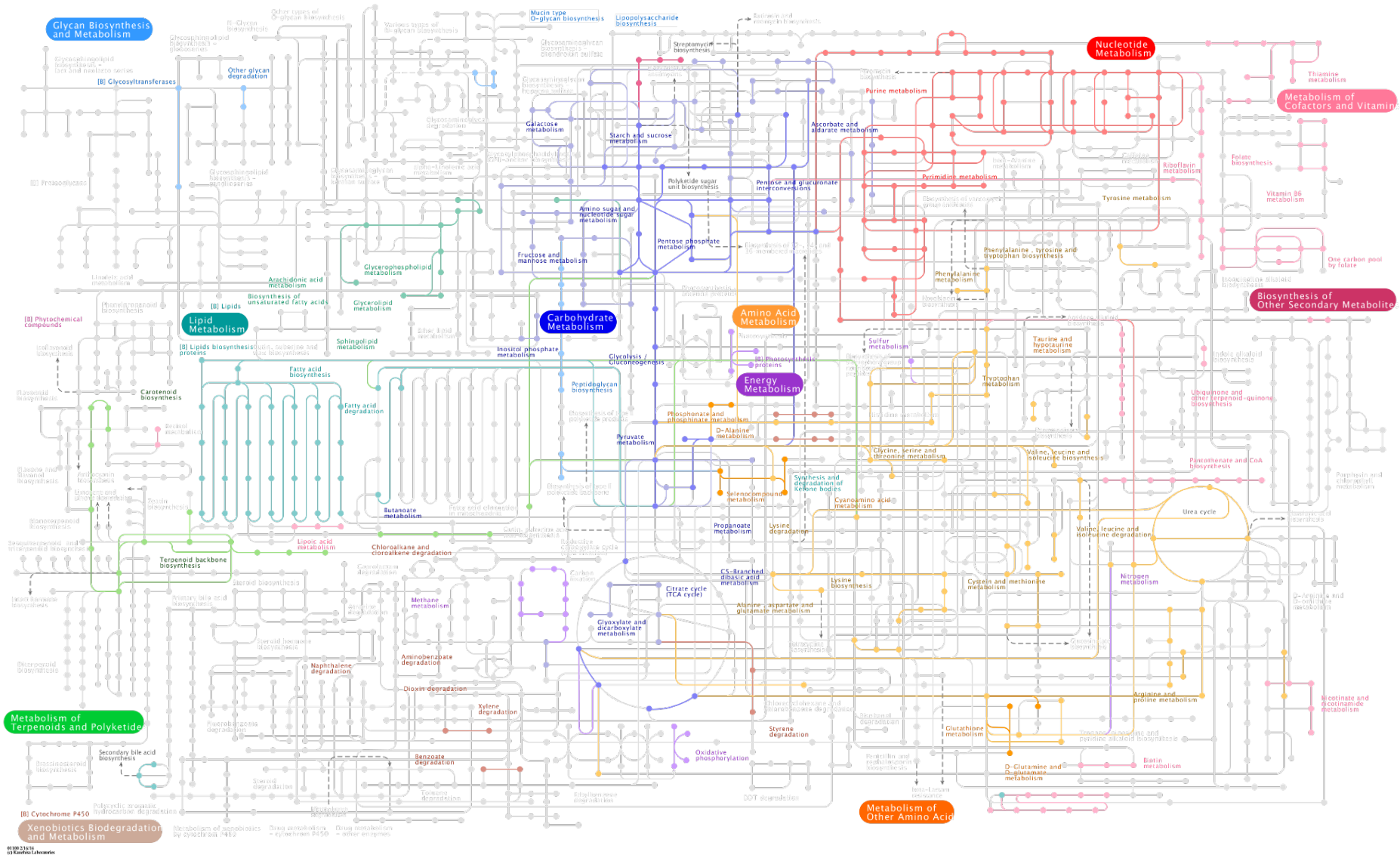

Figure 10. Overview of KEGG pathway map of QU 25. This map was generated by KEGG Atlas (http://www.genome.jp/kegg/atlas.html).

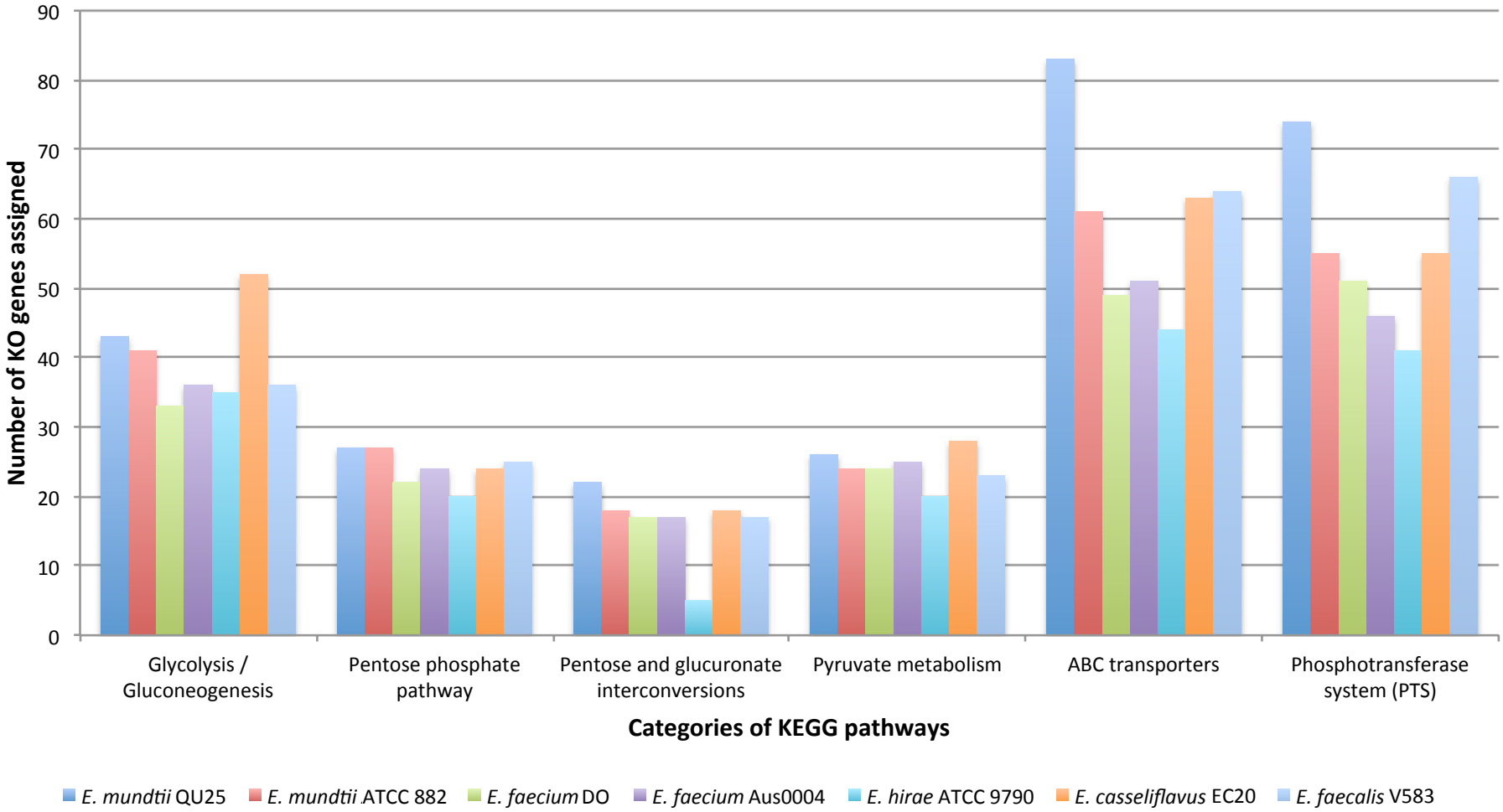

Figure 11. Number of genes related to lactic acid fermentation among allied enterococcal species using KO (KEGG Orthology) gene assignment

Table 8. Number of lactate production-related KO (KEGG Orthology) genes in each *Enterococci* species.

KEGG Pathway	KO	Description	*E. mundtii* QU 25	*E. mundtii* ATCC 882	*E. faecium* DO	*E. faecium* Aus0004	*E. hirae* ATCC 9790	*E. casseliflavus* EC20	*E. faecalis* V583
00030 Pentose phosphate pathway	K00615	transketolase [EC:2.2.1.1]	4[a]	3[b]	1	1	1	3[b]	0
00030 Pentose phosphate pathway	K00616	transaldolase [EC:2.2.1.2]	1	2	0	0	0	0	0
00030 Pentose phosphate pathway	K01621	phosphoketolase [EC:4.1.2.9]	1	1	0	0	0	1	0
00040 Pentose and glucuronate interconversions	K00854	xylulokinase [EC:2.7.1.17]	1	1	0	0	0	1	1
00040 Pentose and glucuronate interconversions	K01805	xylose isomerase [EC:5.3.1.5]	1	1	0	0	0	1	1
00500 Starch and sucrose metabolism	K01198	xylan beta-1,4-xylosidase [EC:3.2.1.37]	2	2	1	1	0	1	0
00620 Pyruvate metabolism	K00925	acetate kinase [EC:2.7.2.1]	1	1	1	1	1	1	1
00620 Pyruvate metabolism	K00625	phosphate acetyltransferase [EC:2.3.1.8]	1	1	1	1	1	1	1
00620 Pyruvate metabolism	K00016	L-lactate dehydrogenase [EC:1.1.1.27]	2	2	2	2	2	1	2
00620 Pyruvate metabolism	K03778	D-lactate dehydrogenase [EC:1.1.1.28]	1	1	1	2	0	0	1

KO (KEGG Orthology) assignments were done using KAAS (KEGG Automatic Annotation Server).[14]

[a] Two genes are considered to be pseudo-genes due to their short length.

[b] One gene is considered to be a pseudo-gene due to its short length.

General Discussion

General Discussion

In this thesis, I address the establishment of analysis methods for microbial genomics using NGS and their applications to resequencing and *de novo* assembly of microbial genomes.

Development of the analysis pipeline and its application to Bacillus subtilis

Since NGS produces a huge amount of data, researchers face difficulties in the post-sequencing in the bioinformatical analysis. This field is new and progresses rapidly, and numerous software programs have been actively developed. However, even today, there is no integrated software for every application of NGS at least to my knowledge. To exert maximum power of NGS, it is necessary to combine a variety of software tools and integrate their analysis results. These situations in bioinformatical analysis are laborious and time-consuming, especially in dealing with a large number of samples.

In order to make bioinformatical analysis more efficient, I developed the analysis pipeline for NGS data, NSAP. It is a command-line-based pipeline running in Linux environment, which is written in the Ruby script language. With a recipe file that describes a protocol of analysis, NSAP executes automatically a suite of software for quality control of reads, alignment, *de novo* assembly, detection of structural variants, etc.

One of the popular pipelines for NGS data used in Japan is DDBJ read annotation pipeline.[1] This pipeline was developed by DNA Data Bank of Japan (DDBJ). It is a web-based pipeline and uses the supercomputer in DDBJ. A web-based pipeline is user-friendly for users with limited skill in informatics. However, a web-based pipeline is

difficult to customize for own purpose and not suitable for handling a large number of samples, while NSAP is customizable due to Ruby script language. In addition, users can specify many samples using a csv formatted recipe file which is editable by Excel. NSAP markedly shortens time of bioinformatical analysis. NSAP is the fundamental software used in all studies described in this thesis.

In chapter I, I applied NSAP to re-sequence *Bacillus subtilis* strain 168. This strain is widely used in Japan as a model microorganism of Gram-positive bacteria. It was distributed to various laboratories in Japan in the 1990s when the sequencing consortium of *B. subtilis* 168 commenced operations. After 20 years from its distribution, variations in growth phenotypes have been reported from several laboratories. To uncover laboratory-specific variations of *B. subtilis* 168 strain in Japan, I re-sequenced these laboratory strains including different isolates (BGSC-1A1), and identified the base variations among them using NSAP and COVA (one of the software which I developed and described in chapter II).

Although the differences in each laboratory stock of strain 168 were few, some had unique variations which might include a still hidden phenotype. These variations might have been caused by the differences in storage conditions in the laboratories or the differences among colonies of the original stock. These results revealed the necessity to understand the genetic differences between the wild-type (parental) strain in each laboratory and the reference strain by re-sequencing analysis, and to pay more attention in managing laboratory strain stocks. This study also demonstrated the dramatically accelerating progress of re-sequencing application in bacterial genetics and the usefulness of NSAP.

Development of the variant annotation software and its application to yeast

To understand genotype-phenotype relationship, an evaluation of mutational effect such as a substitution of amino acid is needed. In addition, it is effective to sequence genomes of multiple strains showing the same phenotype and to pinpoint common variant(s) among them. Although there have been dozens of software for sequencing reads alignment and variant identification, only few software for functional annotation of variants for bacteria are available. One of the popular software for this purpose is ANNOVAR.[2] This tool is suitable for higher organisms such as human. Moreover, few tools can compare variants among multiple samples.

Thus, in chapter II, I developed the software COVA, which is a Ruby-based tool for variant comparison and functional annotation, especially useful for bacterial mutation analysis. COVA is used in re-sequencing studies in this thesis (Chapter I and II). COVA can annotate SNVs, InDels, and other types of variants such as SVs and coverage of genes. In addition, COVA can compare variants among multiple samples, which helps to pinpoint causal variation(s) relating to phenotype. COVA can utilize annotation data sets conformed to Genbank Format which is easily downloadable from NCBI website. Before development of COVA, I had done this process manually using online genome databases, Excel software, and sequence analysis software. COVA automates such laborious and time-consuming process and greatly shortens time of bioinformatical analysis. COVA is freely available at http://sourceforge.net/projects/cova/

In chapter II, I applied COVA to re-sequence *Saccharomyces cerevisiae.* Re-sequencing study of *S. cerevisiae* has done as the evaluation of a novel mutagenesis technique using error-prone DNA polymerase δ (Polδ). This technique is based on the

disparity mutagenesis model of evolution, and has been successfully employed to generate novel microorganism strains with desired traits. However, little is known about the spectra of mutations caused by disparity mutagenesis. Using NSAP and COVA, I evaluated the introduced mutations caused by disparity mutagenesis and showed that they have a broader spectrum of nucleotide changes compared with that of the commonly used chemical mutagen ethyl methanesulfonate (EMS). I demonstrated that a proofreading-deficient and low-fidelity *polδMKII* mutator is a useful and efficient method for the rapid strain improvement based on *in vivo* mutagenesis. I also demonstrated the possibility of application of NGS to the molecular breeding of microorganisms.

Establishment of methods for de novo assembly and annotation and their applications to the determination of the complete genome of enterococci

NGS makes it easy to determine a draft genome of bacteria. A typical draft genome comprises several tens or hundreds of contigs/scaffolds. However, draft genomes contain only partial sequences including repetitive sequences such as rRNA operons, phage regions, and insertion sequences. Thus they are insufficient to analyze the entire genome structure.. In chapter III, I aimed to determine a complete genome sequence with only NGS. I targeted the bacterium *Enterococcus mundtii* QU 25, which can ferment both cellobiose and xylose to produce L-lactic acid efficiently.

I challenged the combination of three sequencing platform, Illumina GAII, 454 GS FLX+, and PacBio RS. With only the third-generation sequencer PacBio RS, I successfully obtained the complete genome sequence of *E. mundtii* QU 25 without conventional Sanger sequencing. I also evaluated an accuracy of sequencing and assembly of PacBio using a

whole-genome mapping technology and second-generation sequencing. It demonstrated that the chromosomal sequence generated by PacBio RS achieved high accuracy. Through the genome annotation and comparative genome analysis, I tried to elucidate the mechanism of the efficient L-lactate production from xylose of the QU 25 strain. This is the first reported complete genome sequence in *E. mundtii*. So the obtained data in this study provide insights into lactate production in this bacterium and its evolution among enterococci.

Conclusion

In this thesis, I established the analysis methods for microbial genomics using NGS. In chapter I, I developed the automated analysis pipeline NSAP for efficient bioinformatical analysis with a large number of samples, and I applied it to re-sequence *B. subtilis*. NSAP successfully shortened the analysis time. NSAP also enabled repetitive analysis for optimization of parameters, contributing the improvement of accuracy of analysis. In chapter II, I developed the variant annotation software COVA for evaluation and comparison of detected variations, and I applied it to re-sequence *S. cerevisiae*. COVA enabled rapid analysis of the evaluation of the effects of variations, and narrowed down causative mutation(s) of a given phenotype. In chapter III, using a third-generation sequencer, I established the methods for determination of a novel complete genome sequence, genomic annotation, and genomic comparison. Then I applied them to determine the novel genome sequence of *E. mundtii* QU 25. With the recent progress of the third-generation sequencer, I demonstrated the determination of a complete genome sequence without any finishing procedure. Now NGS is becoming a new and powerful research method for all microbiologists and certainly changing their working way. I believe the analysis methods

established in this thesis will contribute to not only microbiology but also industrial applications such as molecular breeding.

References

1. Nagasaki, H., Mochizuki, T., Kodama, Y., et al. 2013, DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res.*, **20**, 383–390.

2. Wang, K., Li, M., and Hakonarson, H. 2010, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

ABBREVIATIONS

EMS: ethyl methanesulfonate

InDels: insertions and deletions

SC: synthetic complete

SNP: single nucleotide polymorphism

SNV: single nucleotide variant

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Prof. Hirofumi Yoshikwa. I would like to thank him for encouraging my research using NGS and for allowing me to grow as a research scientist. I would also like to thank Assistant Prof. Satoru Watanabe and Prof. Taku Chibazakura for encouraging my research. I am also grateful to Prof. Mariko Shimizu-Kadota for encouraging my research on lactic acid bacteria. Finally, I would like to thank Prof. Shunsuke Yajima and Prof. Mitsuhiro Itaya for taking time out from their busy schedule to critical reading of this thesis and useful suggestions.

Very special thanks go out to Ms. Tomoko Araya-Kojima for critical reading of this thesis and for useful suggestions. I would like to thank to members of NODAI Genome Research Center (NGRC), Dr. Ryouka Kawahara-Miki, Dr. Kaoru Tsuda, Dr. Mariko Hatta, Dr. Yuko Arai-Kichise, Dr. Takashi Matsumoto, Dr. Yu Kanesaki, and Visiting Prof. Kyo Wakasa for encouraging my research. I would also like to thank to Mr. Taichiro Ishige, Mr. Satoshi Sano, Ms. Hiroko Abe, and Prof. Tomohiro Kono for supporting NGRC. Also I am grateful to my current laboratory member of Iwate Medical University School of Medicine, Assistant Prof. Hideki Ohmomo, Assistant Prof. Ryohei Furukawa, Associate Prof. Tsuyoshi Hachiya, and Prof. Atsushi Shimizu for encouraging me in writing this thesis.

I would also like to thank my family for their devoted support through my entire life and in particular, I must acknowledge my father, Dr. Mitsuharu Shiwa. He encouraged me when I decided to get PhD. In conclusion, I recognize that this research would not have been possible without the financial assistance of the MEXT-Supported program for the Strategic Research Foundation at Private Universities, 2008-2012.