2P-0059

（第34回日本分子生物学会年会
2011.12）

Analysis of accumulated mutation in the whole genome of *E. coli* mutator strains using a next generation sequencer

°Masami Yamada, Kenichi Masumura, Takehiko Nohmi
Division of Genetics and Mutagenesis, NIHS, Tokyo, JAPAN

次世代シークエンサーを用いた大腸菌ミューテーター株のゲノムに蓄積する突然変異の解析

°山田雅巳、増村健一、能美健彦（国立医薬品食品衛生研究所・変異遺伝部）

ABSTRACT

The whole genome sequencing data obtained from analyses with a next generation sequencer can cover comprehensive research of mutation frequency, distribution as well as accumulation in addition to mutation spectra. In this study, we focus on mutator strains of *Escherichia coli*, typical of which is a DNA repair deficient strain, due to their hyper mutation frequency, that is, 100 to 1,000 times higher than the wild type, and determined accumulated mutation on their genome. Strains used in this study were YG6156, *mutT*; YG2250, *mutM/mutY*; AB1157/pYG782, DinB-overproducing strain. Genomic DNA was prepared from over-night culture of theses strains as well as AB1157, which is their original strain. DNA sequencing was carried out with Genome Analyzer (GAIIx), illumine and the reference sequence, *E. coli* K12 subsrt. W3110 uid58567, was obtained from NCBI site. The specific mutation in mutator strains were extracted as 15 in *mutT*, 24 in *mutM/Y*, and 13 in DinB-overproducing strain. The mutation spectra reflected the feature of the deficiency of DNA repair system for each strain, as 40% of mutation (6/15) found in *mutT* strain was A to C, 75% of mutation (18/24) found in *mutM/Y* strain was G to T. On the other hand, DinB-overproducing strain seemed to be an exception since -1 frameshift at G's run was not typically observed as reported before. As for distribution of the mutation, more than half of the mutation observed in the case of *mutT* and *mutM/Y* were in ORF and most of them are base substitutions of missense mutation. This is possibly because cells might die with nonsense or frameshift mutations, which would disrupt the function of the product when they generated in ORF.

BACKGROUND

- **Using a next-generation sequencer, we analyzed whole genome of mutator strains* which lacks DNA repair system or overproduces an error-prone DNA polymerase, and determined how much is the frequency, which is typical spectra and in which locus the mutations are accumulated.**
- **The size of *E. coli* genome is about 5 Mb, and *E. coli* cells are roughly divided every 30 minutes, that is, 50 times a day and 1,500 times a month.**
- ***Mutator strains of *E. coli*: The mutator strains exhibiting high mutation frequency (MF) are called "mutator". Reported MF of the mutator strains used in this study examined using reverse mutation of *lacZ:* $1x10^{-3}$ in A to C in *mutT*, $2x10^{-5}$ in G to T in *mutM/Y*, $2x10^{-4}$ in -1 frameshift at GGGGGG in a DinB-overproducing strain.**

METHODS

Escherichia coli strains used in this study: AB1157 for wild type (WT), YG6156 for *mutT* (ΔT), YG2250 for *mutM*/*mutY* double mutant (ΔMY), AB1157/pYG782 for DinB-overproducing strain (+B).

Reference sequence: *E. coli* K12 subsrt. W3110 uid58567 obtained from NCBI

DNA sequencer: Genome Analyzer (GA II x), illumina

DNA preparation: Single colony isolation was carried out for each strain. One colony on the LB plate was selected randomly and inoculated into 10-mL LB broth. The genomic DNA was prepared for each overnight culture. Cell pellet was lysed by SDS and proteinase K for one hour at 37 ˚C, mixed thoroughly with 5M NaCl, added CTAB/NaCl solution, then incubated the solution for 10 min at 65˚C. Extraction with $CHCl_3$/isoamyl alcohol, and phenol/$CHCl_3$/isoamyl alcohol was carried out. The aqueous phase was mixed with 2-propanol to precipitate the genomic DNA in it. The genomic DNA was treated with RNase, then the RNase was removed by phenol extraction. The amount of the genomic DNA was measured by Nanodorop photometer, Qubit Fluorometer and agarose gel electrophoresis.

Estimated amount of DNA: 22.5 µg for WT, 3.6 µ g for DT, 27 µ g for DMY and 25 µ g for +B.

<<DNA sequencing service was provided by Hokkaido System Science Co., Ltd.>>

Libraty preparation for GAIIx: the procedure was based on "PCR Free Protocol" in the manual of the DNA Sample Prep Kit, Illumina . Fragmented genomic DNA was ligated with adaptors and ran agarose electrophoresis. The bands were cut out from agarose gel and DNA was purified from it.

Data mining procedure: gDNA pair-end method; four samples per lane; 75 base per read x2, ≥3 x 10^9 per sample as sequencing data

Extraction of mutation from sequencing data: Used software was IMC Genomics Edition, in silico biology, inc.

i. Base call - bases with the strongest signal among A, T, C and G are selected for each base in the cluster
ii. QV filtering - clusters with weak fluorescent intensity are removed after calculation
iii. Mapping on the reference sequence – placing of the part for each fragment whose sequence has been fixed among the reference sequence
iv. Extract mutation candidates - search of the sequences which are different from the reference sequence
v. Annotation information to mutation candidates – names of genes which has mutation site, with or without amino acid substitution…

RESULTS

Mutation spectra and distribution

	ΔT		ΔMY		+B	
No. of mutation	15		24		13	
in ORF (base change)	11(7)		19(17)		5(2)	
flanking region (base chage)	4(2)		5(3)		8(4)	
1	*copA*	A to C	*yahI*	A to C	*tnaB*	C to T
2	*ybjL*	A to C	*nei*	C to A	*tnaB*	A to C
3	*ydcC*	A to C	*ycfK*	C to G	*hepA*	+G
4	*atoS*	T to G	*nohA*	G to T	*avtA*	+G
5	*hyfD*	T to G	*ynhG*	C to A	*lacZ*	+TG
6	*yfgA*	T to G	*sufA*	G to T	*	C to T
7	*insI*	T to C	*yebU*	G to T	*	A to C
8	*abgT*	-AAAAAA	*cpsB*	G to T	*	C to A
9	*glnH*	+G	*yeiM*	G to T	*	G to T
10	*fcl*	+G	*hyfC*	G to T	*	+G
11	*ygeG*	+GGGG	*yhbV*	G to T	*	+G
12	*	G to A	*dtd*	G to T	*	+TC
13	*	T to G	*yihN*	C to A	*	-AATTAGAAGGTT
14	*	+C	*trkH*	C to A		
15	*	+TCTC	*yifB*	G to T		
16			*envZ*	C to A		
17			*yjjU*	G to T		
18			*yiaT*	+GG		
19			*rssB*	-AT		
20			*	C to A		
21			*	G to T		
22			*	C to A		
23			*	+TCCCCC		
24			*	-ATTT		

- ✓ **Comparing to the reference sequence, mutations found in the overnight culture of ΔT, ΔMY and +B were 15, 24, and 13, respectively.**
- ✓ **In the case of ΔT, 11 out of 15 mutations were found in ORF, six of which were A–to-C transversion, most of which were missense, i.e., with amino-acid substitution.**
- ✓ **In the case of ΔMY, 19 out of 24 mutations were found in ORF, 17 of which were G-to-T transversion, which was typically observed in this mutator, most of which were missense.**
- ✓ **In the case of +B, five out of 13 mutations were found in ORF, two of which were missense mutations. Particular spectrum was not observed in this strain.**
- ✓ **As for in/del, G/C for insertion and A/T for deletion seem to be targeted.**

DISCUSSION

- **Whole genome sequencing is used in the field of medicine and many reports have been published for comparison between healthy and cancer cells, mutation frequency and the identified locus. On the other hand, a systems biology approach monitors various changes on genes, products etc. of bacteria upon long-term cultivation. Here, we showed the accumulated mutations and their distribution on the genome in the pilot experiment. It would probably be the first time to use mutator strains, each of which has biased mutation spectrum as well as high mutation frequency, for determination of accumulated mutation on their genome. This novel strategy used in this study would make it possible that the information about distribution of mutation under the high mutation frequency.**

CONCLUSION

- **Considering the remarkable progress in the performance of the sequencer, it would be better to determine whole genome with the next-generation sequencer than to analyze mutation frequency and spectra with reporter genes derived from transgenic animals. Because the more information would be available since the former can even cover the distribution of mutations, not only frequency nor spectra.**

Disclosure Information: First speaker, Masami Yamada has no financial relationships to disclose.

CONTACT

山 田 雅 巳, Dr. Masami YAMADA, myamada@nihs.go.jp, 03-3700-9873
〒158-8501東京都世田谷区上用賀1-18-1, 国立医薬品食品衛生研究所 変異遺伝部 第二室

3P-0749

（第35回日本分子生物学会年会 2012.12）

Analysis of mutations accumulated for three months in the genome of *E. coli mutM/mutY* double mutants using a next-generation sequencer

Masami Yamada, Kenichi Masumura, Makiko Takamune, and Takehiko Nohmi/ *Division of Genetics and Mutagenesis, NIHS, Tokyo, JAPAN*

山田 雅巳・増村 健一・高宗 万希子・能美 健彦（国立衛研）

ABSTRACT

Whole-genome sequence data from next-generation sequencers can provide comprehensive information about mutation frequency, distribution and accumulation even with a few samples. Using Hiseq2000 Illumina, we determined mutations generated in the genome of the mutator strain, YG2250, whose mutation rate reaches about 100 to 1,000 times higher than the parental wild-type strain due to the lack of *mutM* and *mutY* genes. Sequential inoculation of the culture into LB medium, i.e., 500 times dilution every 24 hour, was continued for three months. Kanamycin was added into the tube every one week to avoid contamination. Sample genomic DNA was prepared from the culture at 1, 4, 8 and 12 weeks after the first inoculation. Removing the mutation supposed that the original genome of YG2250 has possessed and counting the common mutation among the four samples as one mutation, 88 mutations were totally observed within all four samples, and 47 out of 88 (53%) were base substitutions. Most of them (41/47) happened at GC pairs and transversions which tend to be increased among the generated mutations for three months account for 70% of the substitutions. Eighty percent of the base substitutions were found in ORFs and the numbers of missense and nonsense mutations were almost equal. Besides the substitutions, 41 out of 88 were insertions (34) and deletions (7). The insertions were mainly one-base insertions whereas all the deletions were larger than one base. Rifampicin assay carried out every one week exhibited that the numbers of observed mutations did not seem to be related to the mutation frequency (MF) in the assay. The MF tends to be deeply decreased after gradual increase.

METHODS

Escherichia coli strains used in this study: YG2250 for *mutM/mutY* double mutant and its parental strain is AB1157.

Sample clones preparation: Single colony isolation was carried out on LB plate for YG2250. Four colonies on the LB plate were selected randomly and inoculated into 10-mL LB broth for separately. Incubation overnight and inoculation of 500-times-diluted overnight cultures in LB medium were repeated every day for three months, 12 weeks. Kanamycin was added into the medium every one week to avoid contamination. Four parallel cultures were made. Genomic DNA was prepared from the four cultures at week 1, 4, 8 and 12 in DNA preparation method indicated below.

DNA preparation: The genomic DNA was prepared for overnight culture. Cell pellet was lysed by SDS and proteinase K for one hour at 37 °C, mixed thoroughly with 5M NaCl, added CTAB/NaCl solution, then incubated the solution for 10 min at 65°C. Extraction with $CHCl_3$/isoamyl alcohol, and phenol/$CHCl_3$/isoamyl alcohol was carried out. The aqueous phase was mixed with 2-propanol to precipitate the genomic DNA in it. The genomic DNA was treated with RNase, then the RNase was removed by phenol extraction. The amount of the genomic DNA was measured by Nanodorop photometer, Qubit Fluorometer and agarose gel electrophoresis. The obtained DNA was 197.8 μg for 1W, 311.0 μg for 4W, 101.2 μg for 8W, 151.8 μg for 12W.

DNA sequencing: The sequencing service was provided by Hokkaido System Science Co., Ltd., using the following DNA sequencers (Illumina): HighSeq2000 for Experiment II. *E. coli* K12 substr. W3110 uid58567 obtained from NCBI was used as reference sequence for the experiments. Library preparation for the DNA sequencers was based on "PCR Free Protocol" in the manual of the DNA Sample Prep Kit, Illumina . Fragmented genomic DNA was ligated with adaptors and ran by agarose electrophoresis. The bands were cut out from agarose gel and DNA was purified from it. gDNA pair-end method was used for *Data mining procedure. F*our samples per lane; 75 base per read x2, ≥3 x 10^9 per sample as sequencing data. Using IMC Genomics Edition, in silico biology, inc., *mutations were extracted from sequencing data*.

CONTACT

Masami Yamada, Ph.D., myamada@nihs.go.jp
Division of Genetics and Mutagenesis, National Institute of Health Sciences, 1-18-1, Kamiyoga, Setagaya-ku, Tokyo 158-8501 JAPAN

BACKGROUND

- Using a next-generation DNA sequencer, we analyzed whole genome of mutator strain* which lacks DNA repair system and determined how much is the frequency, which is typical spectra and in which locus the mutations are accumulated.
- The size of *E. coli* genome is about 5 Mb, and *E. coli* cells are roughly divided every 30 minutes, that is, 50 generations a day and 1,500 generations a month.
- *Mutator strains of *E. coli*: The mutator strains exhibiting high mutation frequency (MF) are called "mutator". Reported MF of the mutator strain, *mutM/Y*, examined using reverse mutation of *lacZ* was $2x10^{-5}$ in *G* to T.

RESULTS

For the identification of mutations, we ignored the ones that were supposed to be present in the original genome of the strain.

☆The common mutations observed in the four cultures were counted as one mutation.

- ✓ Total 88 independent mutations were identified in the four cultures.
- ✓ Of 88 mutations, 47 (53%) were base substitutions.
- ✓ Most of them (41/47) were identified at GC pairs and this transversions accounted for 70% of the substitutions.
- ✓ Eighty percent of the base substitutions were identified in ORFs and the numbers of missense and nonsense mutations were almost equal.
- ✓ Besides the substitutions, 41 out of 88 were insertions (34) and deletions (7).
- ✓ The insertions were mainly one-base insertions whereas all the deletions were larger than one base.

MUTATION SPECTRA OBSERVED AT EACH SAMPLING POINT

	1W	4W	8W	12W
Base substitution				
GC to AT	3	4	5	3
GC to CG	1	0	0	1
GC to TA	8	18	21	25
AT to CG	1	2	1	1
AT to GC	1	0	0	2
AT to TA	1	1	1	0
insertion				
1 bp	16	16	15	16
+A	2	0	0	2
+C	1	4	4	4
+G	12	12	10	9
+T	1	0	2	1
≧2 bp	1	0	2	2
deletion				
1 bp	0	0	0	0
≧2 bp	2	0	5	0
total	**34**	**41**	**51**	**50**

DISCUSSION

- Whole genome sequencing is used in the field of medicine and many reports have been published for comparison between healthy and cancer cells, mutation frequency and the identified locus. On the other hand, a systems biology approach monitors various changes on genes, products etc. of bacteria upon long-term cultivation.
- Here, we showed the accumulated mutations and their distribution on the genome in the pilot experiment using mutator strains, each of which has biased mutation spectrum as well as high mutation frequency.
- This novel strategy for determination of accumulated mutation on their genome would make it possible that the information about distribution of mutation under the high mutation frequency.