Archives of Toxicology (2019) 93:2689–2701
https://doi.org/10.1007/s00204-019-02541-3

GENOTOXICITY AND CARCINOGENICITY

Genome-wide somatic mutation analysis via Hawk-Seq™ reveals mutation profiles associated with chemical mutagens

Shoji Matsumura[1] · Hirayuki Sato[2] · Yuki Otsubo[1] · Junichi Tasaki[1] · Naohiro Ikeda[1] · Osamu Morita[1]

Received: 5 June 2019 / Accepted: 14 August 2019 / Published online: 26 August 2019

Abstract

It is difficult to identify mutagen-induced genome-wide somatic mutations using next generation sequencing; hence, mutagenic features of each mutagen and their roles in cancer development require further elucidation. We described Hawk-Seq™, a highly accurate genome sequencing method and the optimal conditions, for using it to construct libraries that would enable the accurate (c.a. 1 error/10^7–10^8 bp) and efficient survey of genome-wide mutations. Genomic mutations in *gpt* delta mice or *Salmonella typhimurium* TA100 exposed to methylnitrosourea (MNU), ethylnitrosourea (ENU), diethylnitrosamine (DEN), benzo[a]pyrene (BP), and aristolochic acid (AA) were profiled using Hawk-Seq™ to analyse positions, substitution patterns, or frequencies. The resultant vast mutation data provided high-resolution mutational signatures, including for minor mutational fractions (e.g. G:C>A:T by AA), which enabled the clarification of the mutagenic features of all mutagens. The 96-type mutational signatures of MNU, AA, and BP indicate their partial similarity to signature 11, 22, and 4 or 29, respectively. Meanwhile, signatures attributable to ENU and DEN were highly similar to each other, but not to signature 11, suggesting that the mechanisms of these agents differed from those of typical alkylating agents. Thus, Hawk-Seq™ can clarify genome-wide chemical mutagenicity profiles at extraordinary resolutions, thereby providing insight into mutagen mechanisms and their roles in cancer development.

Keywords Mutational signatures · Next generation sequencing · Somatic mutation · Genotoxicity tests · Mutagenesis · Cancer genomics

Introduction

The analysis of genomic mutations using next generation sequencing (NGS) has enabled us to generate large-scale, genome-wide catalogues of somatic mutations in human cancer patients (Meyerson et al. 2010; Garraway and Lander 2013). These large-scale cancer mutation data, which reflect the features of mutation profiles for each type of cancer, have enabled us to analyse mutation spectra at extraordinary resolutions and the underlying processes for the development of each type of cancer (Stratton et al. 2009; Stratton 2011). For example, mutation catalogues were used to explore historical information about DNA damage and repair in genomes during cancer development (Nik-Zainal et al. 2012a, b). Signatures of mutational processes have been identified in patients with various types of human cancers using mathematical algorithms (Alexandrov et al. 2013a, b, 2016), and used not only to investigate the contribution of environmental mutagens to cancer development, but also to establish the link between mutagens and human cancer development.

To validate and utilize the knowledge derived from cancer genome analyses more effectively, data regarding mutations induced by cancer-related mutational processes, such as those associated with environmental mutagens, need to be obtained. The associated large-scale, genome-wide mutation data would provide clarity regarding the mutagenic features of mutagens at high resolutions, as observed during cancer research and would provide insight into their role in cancer

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00204-019-02541-3) contains supplementary material, which is available to authorized users.

✉ Shoji Matsumura
matsumura.shouji@kao.com

[1] R&D-Safety Science Research, Kao Corporation, 2606 Akabane, Ichikai-Machi, Haga-Gun, Tochigi 321-3497, Japan

[2] R&D-Analytical Science Research, Kao Corporation, 2606 Akabane, Ichikai-Machi, Haga-Gun, Tochigi 321-3497, Japan

development, while mutually referring to mutations in cancer genomes. Genome-wide mutational profiles were previously characterized via exposure of experimental models to mutagens through clonal expansion (Mimaki et al. 2016; Phillips 2018; Kucab et al. 2019). However, in these experimental models, some mutagens exhibited mutation patterns that were different from those observed in human cancers, which could be attributable to the differences in cell types, or effects of the cell selection process during clonal expansion. Therefore, to precisely elucidate the mutagenic profiles of mutagens in each cell type and thereby understand their role in human cancer development, it is important to develop a method for analysing the primary mutation profiles of mutagens using various experimental resources.

However, unlike mutation analysis in cancer genomes, it is difficult to detect mutagen-induced somatic mutations in very small fractions of cells using NGS. Because of the high error rate in NGS, which is primarily attributable to PCR errors that occur due to processes such as spontaneous DNA oxidation, NGS is not widely used for this purpose (Costello et al. 2013). Our knowledge of mutation profiles of environmental mutagens has mostly been obtained from results of traditional genotoxicity tests, such as the Ames test and transgenic mouse models (Mortelmans and Zeiger 2000; Lambert et al. 2005). Varied mutational data for multiple species, which have been accumulated using these models, form the basis of our knowledge about mutagenicity (Kirkland et al. 2014). However, the resolution of the currently available data is insufficient to thoroughly understand the difference between mutagenic profiles and their association with human cancer (Richardson et al. 1987; Watson et al. 1998; Ohta et al. 2000). Therefore, it is extremely important to develop a methodology for a large-scale, genome-wide survey of somatic mutations.

Several challenges are associated with the detection of rare mutations using NGS. Several promising approaches reportedly increased sequencing accuracy and enabled the utilization of data for both strands of double-stranded DNA (dsDNA) (Travers et al. 2010; Schmitt et al. 2012; Gregory et al. 2016). While PCR errors usually occur only in single strands of dsDNA, true mutations are fixed in both strands through DNA replication within the cell. These techniques involve the sequences of both original dsDNA strands and enable the accurate detection of true mutations, after they are distinguished from sequencing errors. The sequence accuracy of these techniques reportedly proved to be sufficient for the detection of mutations induced by chemical exposure (Matsuda et al. 2013; Chawanthayatham et al. 2017). However, most existing methods require either additional molecular barcodes during library preparation to discriminate between molecules or a specific sequence apparatus that has been used only in a few studies. Certain accurate sequencing techniques reportedly used endogenous sequences as an alternative to exogenous tags. However, none of these existing techniques were optimized to maximize sequence output or the covered genomic region; thus, the efficiency of genome-wide mutation profiling was not maximal (Kinde et al. 2011; Hoang et al. 2016). Therefore, a simple method with a substantial throughput is required to enhance our knowledge of mutation profiles rapidly.

Here, we developed a simple, highly accurate genome sequencing procedure for performing high-throughput mutation analysis across large genomic regions. We named our procedure 'Hawk-Seq™', after 'hypothesis alignment with weak overlap', in which we controlled the overlap rate of DNA fragments, by controlling the input DNA amount for PCR (IDAP) per unit genome length; this has been described in detail in another section. Additionally, it enables the accurate detection of single mutations from large genomic regions, such as that of the 'hawk-eye'. Hawk-Seq™ achieves an experimental throughput that is equivalent to that of standard Illumina libraries, because it requires only slight modifications in PCR steps and no additional external barcodes. Therefore, it can be easily applied to the evaluation of various samples in different laboratories; this is critically important for expanding our knowledge of mutation profiles via mutagen exposure. Here, we identified the IDAP to be a key parameter for maximizing the sequence efficiency and covered genomic regions, and determined the optimal experimental conditions for performing Hawk-Seq™. Furthermore, we prove that Hawk-Seq™ is applicable to the clarification of mutagen-induced, genome-wide somatic mutations in bacteria and mammals using *Salmonella typhimurium* TA100 and the transgenic mouse model (*gpt* delta mice). We assessed the ability of Hawk-Seq™ to characterize somatic mutations via mutagen exposure. This technology will further accelerate our understanding of mutagens as a superior alternative to genotoxicity tests.

Materials and methods

Mutagens

Methylnitrosourea (MNU; CASRN. 684-93-5) and Ethylnitrosourea (ENU; CASRN. 759-73-9) were purchased from Toronto Research Chemicals (Toronto, Canada). Aristolochic acid I (AA; CASRN. 313-67-7) was obtained from Sigma-Aldrich (MO, USA). Diethylnitrosamine (DEN; CASRN. 55-18-5) was purchased from the Tokyo Chemical Industry Co. Ltd. (Tokyo, Japan). Benzo[a]pyrene (BP; CASRN. 50-32-8) was obtained from FUJIFILM Wako Pure Chemical Corporation (Osaka, Japan).

Bacterial DNA sample preparation

The *Salmonella typhimurium* Ames tester strain, TA100, was supplied by the NITE Biological Resource Center (Tokyo, Japan). The TA100 strain was cultured for 4 h at 37 °C with Nutrient Broth No. 2 (Oxoid, UK) containing 24 µg/mL ampicillin (Sigma-Aldrich, MO, USA). The resulting bacterial cell suspensions were used for preparing DNA samples exposed to mutagens. Mutagen exposure was performed according to the pre-incubation procedure of the standard Ames test, which included slight modifications for DNA extraction (Mortelmans and Zeiger 2000; Matsumura et al. 2018). Briefly, 100 µL of bacterial cell suspensions was mixed with 500 µL of phosphate buffer, pH 7.4 (Nacalai Tesque, Kyoto, Japan), and 100 µL of DMSO or test substance solutions. During BP exposure, 500 µL of the rat liver S9-mix (Kikkoman Biochemifa Company, Tokyo, Japan) was used instead of phosphate buffer. The resultant mixture was agitated at 100 rpm and 37 °C, for 20 min. Then, to prepare DNA samples for sequencing, 50 µL of mixture was inoculated into 2 mL of nutrient broth and cultured at 37 °C for 14 h, to fix mutations. Subsequently, genomic DNA was isolated using a DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA), according to the manufacturer's instructions. The standard Ames test was also performed using minimum glucose medium plates (Tesmedia®AN; Oriental Yeast Co. Ltd., Tokyo, Japan). Plates were incubated at 37 °C for 48 h and the number of colonies generated was counted.

Mouse DNA sample preparation

Male C57BL/6JJmsSlc-Tg (*gpt* delta) mice (7–9 weeks old) supplied by Japan SLC, Inc. (Shizuoka, Japan) were acclimatized for 1 week before use. Then, 12.5 and 25 mg/kg of MNU, 75 and 150 mg/kg of ENU, and 40 and 80 mg/kg of DEN were dissolved into saline and administered intraperitoneally (i.p.) once daily, for 5 consecutive days. Meanwhile, 150 and 300 mg/kg of BP and 5 and 10 mg/kg of AA were dissolved into olive oil and administered orally (p.o.) once daily, for 5 consecutive days. Saline and olive oil were administered in the same manner as controls. Mice were euthanized by carbon dioxide inhalation, 7 days after final administration. Genomic DNA samples were extracted from the organs listed in Table 1 and Supplementary Table S3, using the RecoverEase DNA Isolation Kit (Agilent Technologies, CA, USA), according to the manufacturer's instructions. All animal experiment protocols were approved by the Animal Testing Committee at Kao Corporation.

Library construction and sequencing

TA100 and mice genomic DNA samples were sheared to 350 bp sized fragments using a sonicator (Covaris, MA, USA). The resultant DNA fragments were used for library construction, using the TruSeq nano DNA library preparation kit (TruSeq; Illumina, San Diego, USA), with a slight modification for Hawk-Seq™. Briefly, after fragmentation with a sonicator, DNA fragments were subjected to end repair, 3′ dA-tailing, and ligation to TruSeq indexed adaptors, according to the manufacturer's instructions. Then,

Table 1 The summary of the *gpt* assay indicating the number of colonies induced by exposure to mutagens ($n=4$ or 5)

Materials	Administration route (vehicle)	Organ	Dose (mg/kg/day)	No. of animals for gpt assay (for Hawk-Seq™)	Mutant frequency ($\times 10^{-6}$)		
					Mean	SD	Statistics
MNU	i.p. (saline)	Bone marrow	0	5 (4)	3.31	0.87	
			12.5	5 (4)	8.53	4.38	*(S)
			25	5 (4)	20.44	3.84	*(S)
ENU	i.p. (saline)	Bone marrow	0	5 (4)	2.36	1.26	
			75	5 (4)	37.80	28.24	*(S)
			150	5 (4)	126.33	38.64	*(S)
DEN	i.p. (saline)	Liver	0	5 (4)	3.85	3.19	
			40	5 (4)	26.54	25.90	*(S)
			80	4 (4)	31.48	2.40	*(S)
BP	p.o. (olive oil)	Bone marrow	0	5 (4)	1.27	0.73	
			150	5 (4)	15.22	8.36	*(S)
			300	5 (4)	48.66	22.23	*(S)
AA	p.o. (olive oil)	Kidney	0	5 (4)	4.64	2.40	
			5	5 (4)	6.70	1.54	
			10	4 (4)	13.45	4.70	*(D)

*(S): $p<0.05$ by Steel's test, *(D): $p<0.05$ by Dunnett's test

the DNA concentration of each sample was measured using Agilent 4200 TapeStation (Agilent technologies, CA, USA). Ligated products were diluted to an appropriate concentration with suspension buffer and subjected to PCR enrichment. To perform the experiment for determining the optimal IDAP, several dilutions of ligation products of TA100 DNA samples exposed to DMSO or ENU were prepared, with concentrations ranging from 800 to 0.2 amol/μL, and 25 μL of each sample was used for PCR enrichment. These were equivalent to 20,000, 2500, 1250, 625, 313, 156, 78, 39, 20, 10 and 5 amol of IDAP. After PCR enrichment, the resultant PCR products were sequenced on the HiSeq 2500 platform using v4 chemistry (Illumina, San Diego, USA). To determine the optimal conditions for maximizing the recovery of dsDNA consensus sequences (dsDCS), these samples were sequenced to obtain a yield of ~ 10 Gbp (50 M read pairs at 2 × 100 bp) per sample and the sequencing efficiency (SE, %) was calculated, by dividing the number of dsDCS read pairs by the number of read pairs used for each sample. To evaluate the effect of the original sequencing amount on SE, we also calculated the SE, using about 1/5th of the number of read pairs subsampled from original read pairs (i.e. 10 M read pairs per sample). To perform mutation analysis of TA100 and *gpt* delta mice samples, the ligated products were diluted to a concentration of ~ 3.1 amol/μL (i.e. 78 amol at 25 μL), subjected to enrichment by PCR, and sequenced to yield ~ 50 M of read pairs, using HiSeq 2500. For samples used to calculate the rate of 'overlap by accident (OBA)', two different TruSeq indexed adaptors were used per sample, to construct libraries. Index data were used as markers to calculate the rate of OBA. The index data were connected to each read pair upon the addition of the index sequence to the header region of the fastq file.

Data processing for Hawk-Seq™

Adapter sequences and low-quality bases were removed from the generated read pairs using Cutadapt (Martin 2011). Then, edited paired-end reads were mapped to reference genome sequences and their format was changed to the SAM format using Bowtie2 software (Langmead and Salzberg 2013). For TA100 and *gpt* delta mice genome samples, the *S. typhimurium* LT-2 genome (GCA000006945.2) and mouse genome (GRCm38) were used as reference genome sequences, respectively. SAM format processing was performed using SAMtools-1.2 (Li et al. 2009). Then, to prepare dsDCS sequences, read pairs that shared the same genomic locations were grouped into the SP-Gs (Fig. 1 and Supplementary Fig. S1). To utilize the sequence information from both strands of dsDNA fragments, SP-Gs that included at least one read pair from both R1R2-Gs and R2R1-Gs were identified and used to generate dsDCS (Fig. 1 and Supplementary Fig.S1). The output for the resulting dsDCS read pairs included new paired fastq files; they were mapped again to the reference genome sequence using Bowtie2 software. The resulting SAM files were processed using SAMtools and mutations were detected.

Mutation detection and statistical analyses

To analyse mutation frequency, the number of base substitutions for each type was separately enumerated. The mutation frequencies for each mutation type per 10^6 G:C or A:T base pairs were calculated, by dividing each mutation count by the total read base count mapped to the G:C or A:T base pair, respectively. Statistical analyses were performed based on the frequencies of each mutation type, per 10^6 bp using the Dunnett's multiple comparison test or Student's *t* test. In analyses using DNA samples of mice, to reduce background mutation call frequencies caused by single nucleotide polymorphisms (SNPs), the genomic positions listed in the ensemble variation list (version 92) were removed from the analysis (Chen et al. 2010). Additionally, 898 genomic positions in which mutations were frequently detected in control samples were removed from the analysis. For the analysis using TA100, known variant positions observed in our laboratory strain were removed from the analysis (Matsumura et al. 2017). To estimate the dependency of mutation frequencies on the sequence context, the bases immediately 5′ and 3′ of each mutation were analysed, and mutation frequencies were calculated within the context of each trinucleotide. To evaluate the similarity of each 96-trinucleotide format mutation pattern, the cosine similarities (CS) between signatures of mutagens or signatures listed in COSMIC were calculated (Alexandrov et al. 2013a). The detected mutations were annotated based on their genomic positions and base substitution types, using SnpEff (Cingolani et al. 2012).

Data availability

The *gpt* delta mice genome sequence data used in this study are available in the Sequence Read Archive of the DNA Data Bank of Japan, under Accession Number DRA008304.

Results

Overview of Hawk-Seq™ and definition of terms

Hawk-Seq™ utilizes the sequences of both strands of dsDNA in a manner similar to that for existing accurate sequencing techniques (Fig. 1). Here, to explain its optimization processes, we describe the algorithm briefly and define terms. In HiSeq, paired reads were obtained from both ends of the DNA fragment. Therefore, the read pairs originating

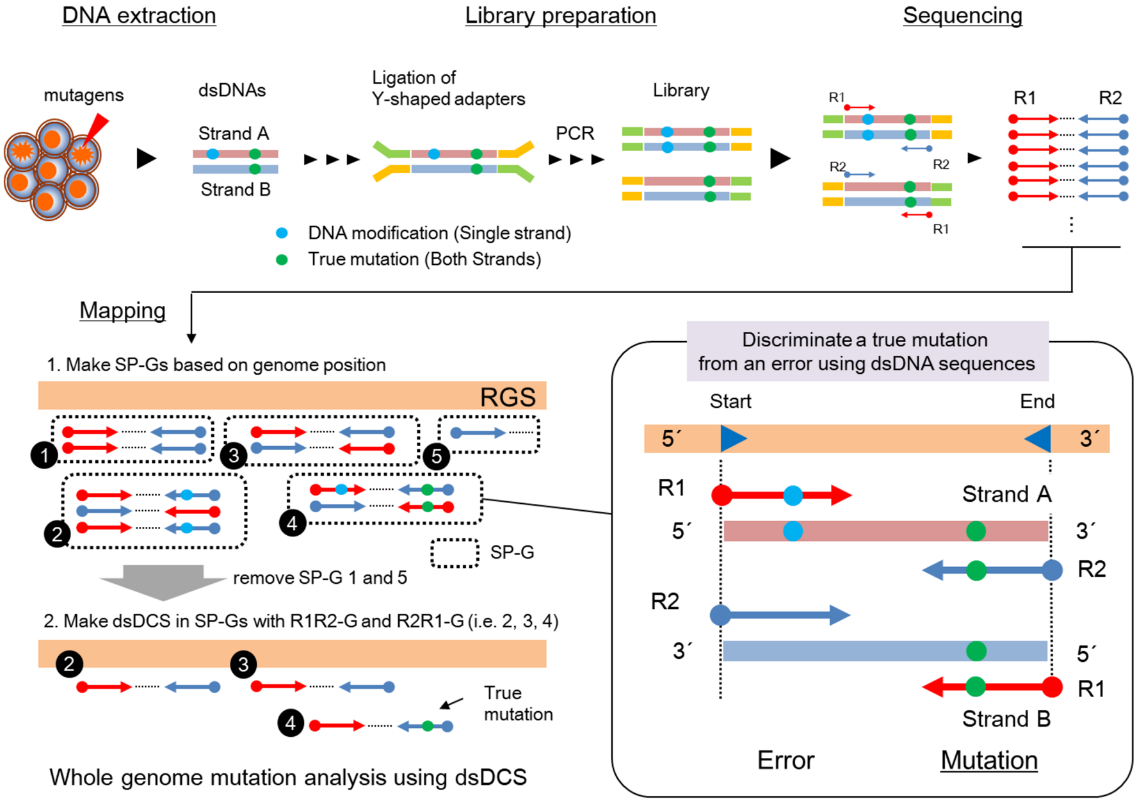
DNA extraction
Library preparation
Sequencing
mutagens
dsDNAs
Strand A
Strand B
Ligation of
Y-shaped adapters
PCR
Library
R1
R2
DNA modification (Single strand)
True mutation (Both Strands)
Mapping
1. Make SP-Gs based on genome position
RGS
1
2
3
4
5
SP-G
remove SP-G 1 and 5
2. Make dsDCS in SP-Gs with R1R2-G and R2R1-G (i.e. 2, 3, 4)
True
mutation
Whole genome mutation analysis using dsDCS
Discriminate a true mutation
from an error using dsDNA sequences
Start
End
5´
3´
R1
Strand A
R2
Strand B
Error
Mutation

Fig. 1 The Hawk-Seq™ concept. Genomic DNA samples extracted from cell populations (mice or bacteria) after mutagen exposure were used to perform Hawk-Seq™. First, libraries for Illumina sequences were constructed according to the standard library preparation methods and sequenced. The obtained read pairs were mapped to reference genome sequences. The mapped read pairs were grouped into SP-Gs (surrounded by dashed line above), based on the genomic coordinates of their R1 start (i.e. one end of the fragment) and R2 start (i.e. the other end of the fragment) positions. Then, read pairs in an SP-G were further classified into two groups, based on their R1 and R2 orientation (i.e. R1R2-G or R2R1-Gs), which depended on the strand they were derived from. Among the above 5 SP-Gs, SP-G 1 and 5 are excluded from the following analysis because they included either R1R2-G or R2R1-G alone. SP-Gs that included both R1R2-G and R2R1-Gs (i.e. SP-G 2, 3, and 4 above) are used for preparing dsDNA consensus sequences (dsDCS), followed by mutation analysis

from each dsDNA strand would be mapped to the same position in the reference genome sequence (RGS). These read pairs can be categorized as ‘same position groups (SP-Gs)’. Furthermore, because Illumina uses y-shaped asymmetric adapters, if two read pairs in an SP-G are derived from different strands of parent dsDNA, the direction of R1 and R2 would be opposite. Therefore, read pairs in an SP-G were further separated into two groups, based on the read direction (i.e. R1R2-G or R2R1-G), which represent the parent strand in the original dsDNA fragment. Thus, we developed an algorithm to identify dsDNA consensus sequences (dsDCSs) in which (1) read pairs were grouped into SP-Gs based on mapping positions, (2) read pairs in an SP-G were further classified as R1R2-G or R2R1-G, based on the read direction, and (3) if both R1R2-G and R2R1-G were included in an SP-G, the consensus read pairs (CRPs) were constructed separately within R1R2-G or R2R1-G; finally, (4) dsDCSs were prepared by comparing these CRPs (Fig. 1 and Supplementary Fig. S1). In Hawk-Seq™, an individual SP-G was made as far as the positions of either end of paired reads were different from each other by at least 1 bp. Additionally, we made dsDCSs from an SP-G as long as one read pair from each of R1R2-G and R2R1-G was included in that SP-G. Thus, the number of prepared dsDCSs was sufficient to cover large genomic regions, using only mapping results and no external barcode sequences.

Optimization of Hawk-Seq™ using *Salmonella* DNA samples

To optimize Hawk-Seq™, two conflicting parameters needed to be identified. First, the conditions for maximizing sequencing efficiency (SE) needed to be clarified to increase sequence output and the genomic regions covered. These would be influenced by the sequenced amount and molecular diversity of the library, which depends on the IDAP. Meanwhile, the possibility of incorrectly assigning read pairs from a different parent dsDNA fragment into an SP-G could inhibit SE maximization during the detection of each somatic mutation and needs to be minimized. This phenomenon, which we have named ‘overlap by accident (OBA)’, is especially critical for the performance of Hawk-Seq™, as it causes mutations to be overlooked as sequencing errors. This would occur if DNA fragments from multiple cells that were sheared at the same genomic locus were sequenced independently. It is speculated that the smaller the target genome size, the higher the possibility of OBA; besides, it would also be affected by IDAP. Thus, we optimized Hawk-Seq™ by manipulating the IDAP with regard to these two parameters, using genomic DNA samples of *S. typhimurium* TA100, whose genome is the smallest among resources used for evaluating mutagenicity.

First, we calculated SE using libraries prepared with different IDAPs while sequencing a fixed amount per sample (c.a. 50 M read pairs) (Fig. 2a). In the first experiment (Exp. 1: 20,000–156 amol), the SE increased with a decrease in the IDAP and reached a maximum value of 8.5%, when the IDAP was 156 amol. In the second experiment (Exp. 2: 156–5 amol), the SE reached a maximum value of 7.1% when the IDAP was 78 amol, after which it decreased. Based on these results, after taking the experimental error into account, we concluded that the optimum IDAP value that maximized the SE was ~ 156–39 amol in about 50 M of sequencing read pairs, which yielded ~ 5–9% of dsDCSs, as compared to the number of read pairs used for mapping (Fig. 2a). These corresponded to 0.2–2 read pairs/IDAP (amol) or 1–2 read pairs/SP-G (Supplementary Table S1). To evaluate the effect of the original sequencing amount on SE, we also calculated the SE using ~ 1/5th of the number of read pairs subsampled from original read pairs (i.e. 10 M read pairs per sample). As a result, the most efficient IDAP was 20 amol and it decreased almost proportionally with the sequencing amount, as compared to that in original experiments (Fig. 2a, Supplementary Fig. S2a). The numbers of read pairs/IDAP or the number of read pairs/SP-G in optimal conditions were equivalent to those in the original experiments (Supplementary Fig. S2a). Therefore, these parameters can be used as indicators to determine the optimal experimental conditions for Hawk-Seq™.

Next, we calculated the rate of SP-Gs with read pairs of two indexes (SPG-2idxs), within SP-Gs with 2 or more read pairs in the Exp. 1 and Exp. 2 above (Fig. 2b), as they were indicative of OBA probability. As expected, the rate of SPG-2idxs decreased as IDAP decreased and became < 1%, when IDAP was ≤ 78 amol (i.e. approx. ≤ 16 amol/Mbp genome). These values were not significantly affected by

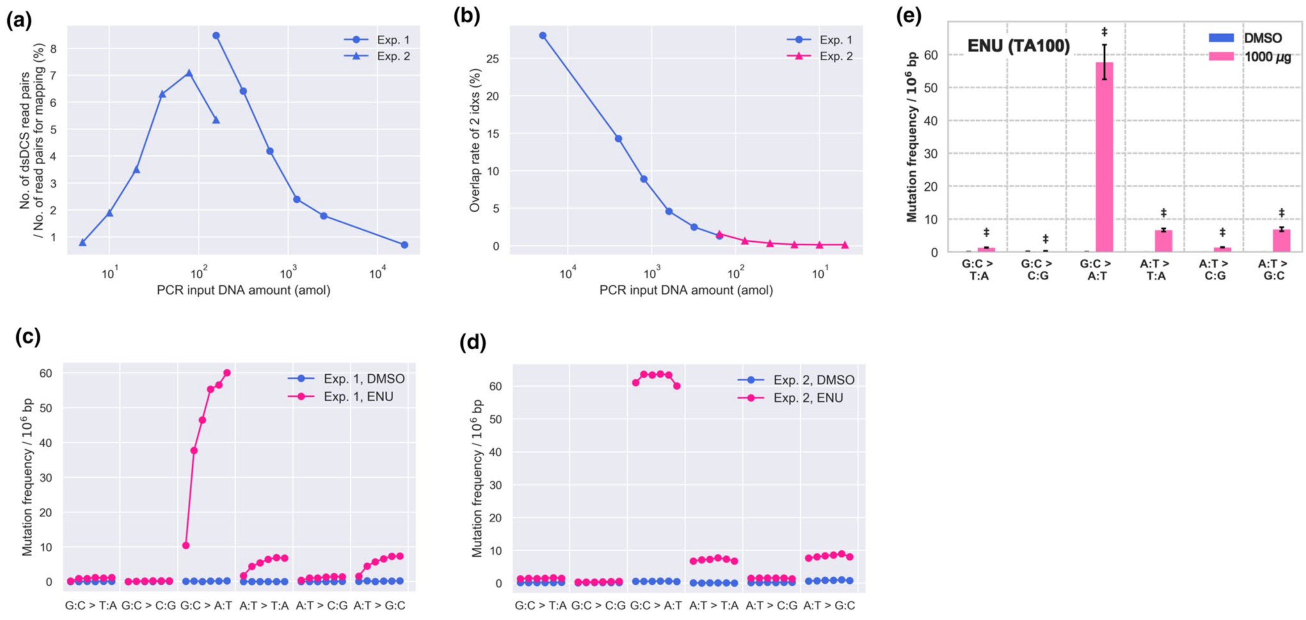

Fig. 2 Optimization of Hawk-Seq™ using DNA samples of TA100 exposed to DMSO or ENU. Libraries using 20,000, 2500, 1250, 625, 313, or 156 amol of IDAP (Exp. 1), and 156, 78, 39, 20, 10, or 5 amol of IDAP (Exp. 2) were subjected to Hawk-Seq™. **a** As an indicator for S.E., the mean rates of the number of generated dsDCS read pairs within the number of read pairs used for mapping to reference genome sequences were shown (mean of DMSO and ENU, i.e. $n=2$). **b** The mean overlap rate of 2 idxs was calculated by dividing the no. of SPG-2idxs by no. of SP-Gs of 2 or more read pairs using the mapping results for the libraries in Exp. 1 and Exp. 2 (mean of DMSO and ENU, i.e. $n=2$). **c** The mutation frequencies in six types of base substitutions were calculated using libraries in Exp. 1 of DMSO- and ENU-exposed samples. Each circle represents the value in each IDAP in Exp. 1 (i.e. 20,000, 2500, 1250, 625, 313, and 156 amol from the left). The frequency of each mutation type was negatively correlated with IDAP. **d** The mutation frequencies in six base substitution types were calculated with libraries in Exp. 2 (i.e. 156, 78, 39, 20, 10, and 5 amol from the left) of DMSO- and ENU-exposed samples. The frequency of each mutation type was almost unchanged with different IDAP values in Exp. 2. **e** Mutation frequencies in 6 types of base substitution in DNA samples of TA100 exposed to DMSO or ENU at 78 amol of IDAP ($n=3$). $^{‡}p<0.001$ by Student’s *t* test

the sequenced amount (Supplementary Fig. S2b) and were considered to be determined by IDAP per unit size of target genomic sequence. To evaluate the effect of OBA on mutation detection, we also analysed mutation frequencies in DNA samples from Exp. 1 and Exp. 2 that were exposed to DMSO or ENU (Fig. 2c, d). As the rate of OBA decreased, the mutation frequencies in the ENU-exposed samples were increased. Therefore, taking the optimal conditions for SE into account, we decided that the most optimal IDAP in ~ 5 Mbp of genomic regions and 50 M of sequence read pairs was 78 amol. As these conditions were applicable to larger genomic sequence analyses, we performed the following experiments using mice samples with this IDAP value. Figure 2e shows the results of analysis of mutations induced by ENU at 78 amol of IDAP. Hawk-Seq™ clearly detected increased mutation frequencies after bacteria were exposed to ENU, with a sensitivity that was comparable to that of the traditional Ames test (Supplementary Table S2). The mutations induced by MNU, BP, and AA in TA100 were also analysed and mutational spectra that reflected the mechanism of each mutagen were obtained (Supplementary Fig. S3).

Analyses of mutagen-induced mutations in *gpt* delta mice

Next, as a representative mammalian resource, we analysed genomic DNA samples of *gpt* delta mice exposed to the ENU, MNU, DEN, BP, and AA mutagens using Hawk-Seq™. DNAs were extracted from organs targeted by these mutagens [i.e. bone marrow (BM) for MNU, ENU, and BP; liver for DEN; kidney for AA (Table 1); and liver (BM for DEN) as the second organ (Supplementary Table S3)]. These representative mutagens were selected, because of their ability to efficiently induce base substitutions in mice, including *gpt* delta mice (Lambert et al. 2005). Additionally, their mutational patterns or signatures have been thoroughly studied in mutagenicity assays and human cancer samples (Richardson et al. 1987; Bigger et al. 2000; Hunter et al. 2006; Arlt et al. 2007; Poon et al. 2015). In the *gpt* assay, the number of colonies significantly increased after exposure to these mutagens (Table 1, Supplementary Table S3), which indicated a substantial level of induction of mutations. During the representative analyses of saline- or ENU-treated BM samples ($n = 12$), 69 ± 11 million (M) read pairs per sample were subjected to Hawk-Seq™, and 5.5 ± 1.0 M dsDCS read pairs were obtained; the mean SE was 7.97 ± 0.50% ($n = 12$). The mean percentage of the genomic region covered by at least 1 dsDCS was 25.8 ± 3.9%, and the mean depth within these mapped regions was 1.24 ± 0.05 ($n = 12$); this indicated that the large genomic regions were uniformly sequenced. The rate of SPG-2idxs within SP-Gs with 2 or more read pairs was analysed with multiple samples and confirmed to be sufficiently low (Supplementary Table S4), as observed during TA100 analysis. These data indicated that Hawk-Seq™ was successfully performed and mutations were analysed across large genomic regions. Within 8 *gpt* delta mice samples for each mutagen (i.e. 4 samples each for low and high doses, Table 1), 12,137, 63,320, 13,830, 6498, and 2478 base substitutions were detected for MNU, ENU, DEN, BP, and AA, respectively. The detected mutations in a sample for each mutagen were plotted onto the genome using a Circos plot (Supplementary Fig. S4) (Krzywinski et al. 2009). In mice, a majority of mutations occurred in intergenic or intronic regions, as expected (Table 2). Among mutations found in protein-coding regions, non-synonymous mutations were more frequently observed than synonymous mutations, for which the ratio was equivalent to that observed in cancer genomes (Greenman et al. 2007; Kandoth et al. 2013). In TA100, a majority of mutations were detected on coding genes (Supplementary Table S5). In TA100, the non-synonymous/synonymous mutation ratio exhibited values specific to each mutagen, with small errors between samples. This parameter might be indicative of the mutagenic features of mutagens.

Analysis of mutagen-induced mutation spectra in *gpt* delta mice

As mutations were widely distributed throughout the entire genome, it was assumed that Hawk-Seq™ provides less biased mutation spectra. Based on the obtained mutation data, we determined the mutation spectra in the six base substitution subtypes induced by each mutagen (Fig. 3a–e). The major mutation patterns induced by mutagens, such as those of A:T>T:A, A:T>G:C, and G:C>A:T in ENU-exposed samples, and G:C>A:T in MNU, G:C>T:A in BP,

Table 2 Summary of annotations of mutations observed in samples of *gpt* delta mice exposed to mutagens (eight samples per mutagen)

	MNU	ENU	DEN	BP	AA
Organ	Bone marrow	Bone marrow	Liver	Bone marrow	Kidney
Total	12,137	63,320	13,830	6498	2478
Outside of the gene	7727	40,731	8449	4235	1562
CDS	211	735	216	117	47
Synonymous	70	189	60	25	12
Non-synonymous	141	546	155	92	35

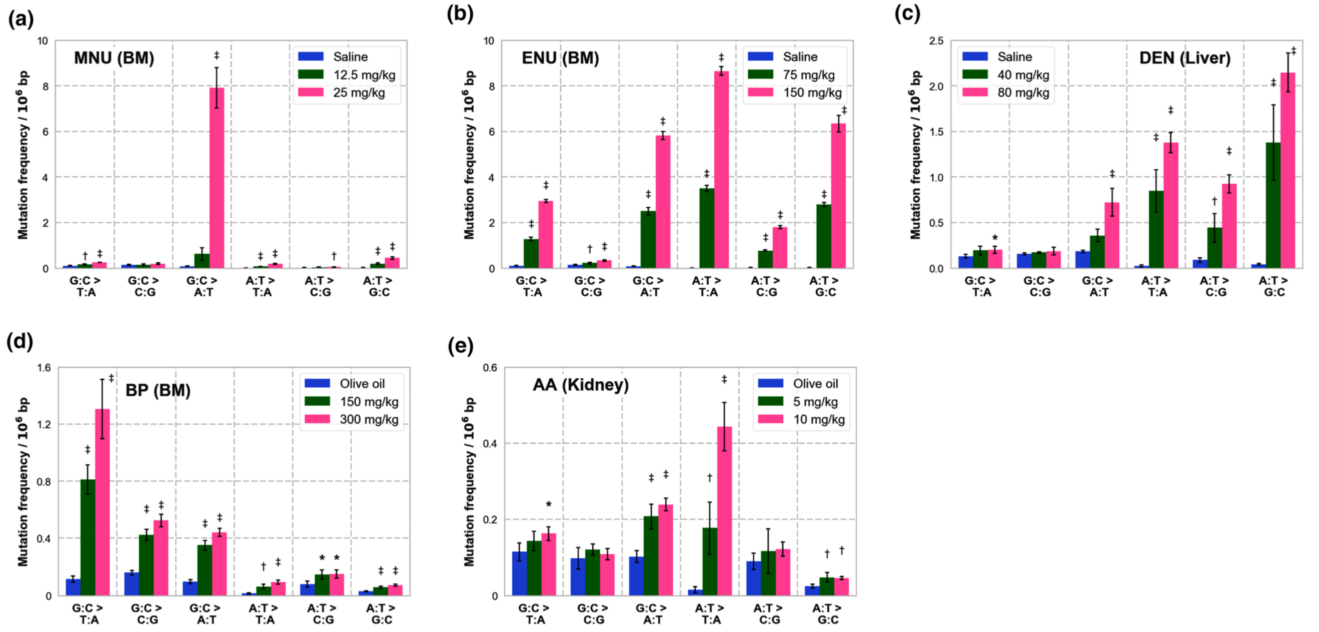

Fig. 3 Analysis of the frequencies of mutations induced by exposure to **a** MNU (BM, 12.5, 25 mg/kg/day), **b** ENU (BM, 75, 150 mg/kg/day), **c** DEN (liver, 40, 80 mg/kg/day), **d** BP (BM, 150, 300 mg/kg/day), and **e** AA (kidney, 5, 10 mg/kg/day) in *gpt* delta mice. The mutation frequencies per 10^6 of G:C or A:T base pairs are presented ($n=4$). Asterisks and daggers indicate p values in Dunnett's multiple comparison test ($^*p<0.05$, $^†p<0.01$, and $^‡p<0.001$)

and A:T>T:A in AA-exposed samples (Richardson et al. 1987; Bigger et al. 2000; Arlt et al. 2007), were consistent with those reported previously. We also identified other previously unreported minor mutation patterns that had been induced, including G:C>T:A in ENU, G:C>A:T in BP, and G:C>A:T in AA-exposed samples. Furthermore, regarding ENU and AA, there were clear differences in the mutation spectra of *gpt* delta mice and TA100 (Supplementary Fig. S3). In both these mutagens, mutations on G:C base pairs were more frequently observed in TA100 than in *gpt* delta mice. These results indicated that Hawk-Seq™ could clarify the mutation spectra of mutagens at sufficiently high resolutions and enable us to speculate about the mechanisms of mutagens, including in mammals. Thus, Hawk-Seq™ provides useful data for the analysis of the mutagenic reactions of each mutagen in multiple species, which were difficult to obtain via traditional genotoxicity assays.

Analysis of trinucleotide mutational signatures in *gpt* delta mice

Mutation signatures usually occurring in 96 types of trinucleotide formats (6 types of base substitution × 4 types of 5′ bases × 4 types of 3′ bases) have been extracted from a variety of human cancers; currently, 30 signatures are registered in COSMIC (Alexandrov et al. 2013a). We obtained mutation signatures associated with ENU, MNU, DEN, BP, and AA exposure in mice, by examining the 5′ and 3′ bases of mutations in the immediate vicinity (Fig. 4a, b, each mutation type was represented by the pyrimidine bases C and T). Furthermore, we analysed the similarity of these signatures to those in COSMIC (Fig. 4c). Although MNU, ENU, and DEN are categorized as alkylating agents, the signature for MNU was not similar to that for ENU and DEN [cosine similarity (CS) was 0.41 and 0.24, respectively]. Only the signature associated with MNU exhibited a significantly high CS (0.91) with signature 11 (alkylating agent) and had peaks representing G:C>A:T mutations on NpCpY (Y: pyrimidine base) consensus sequences, which are patterns known to be typical to alkylating agents (Alexandrov et al. 2013a; Matsumura et al. 2018). Meanwhile, the signatures for ENU and DEN were not similar to signature 11 (CSs are 0.28 and 0.18, respectively). The signatures for ENU and DEN exhibited a comparatively high CS with regard to each other (0.69). Particularly, it was observed in the liver that the CS for the signatures for ENU and DEN was 0.93 (Supplementary Figs. S5, S6). These results suggested that ENU and DEN have common mutagenic mechanisms that probably originated from ethyl cations. Meanwhile, the mutational signature for ENU observed in TA100 exhibited different patterns, which indicated a high CS (0.92) to signature 11 (supplementary Fig. S7). These results suggested that ENU might cause mutagenicity via a typical mode of action of an alkylating agent in *Salmonella*, but not in mice. Signatures of BP exhibited CS values that were comparatively high, as compared to COSMIC signatures 4, 24, and 29 (0.55,

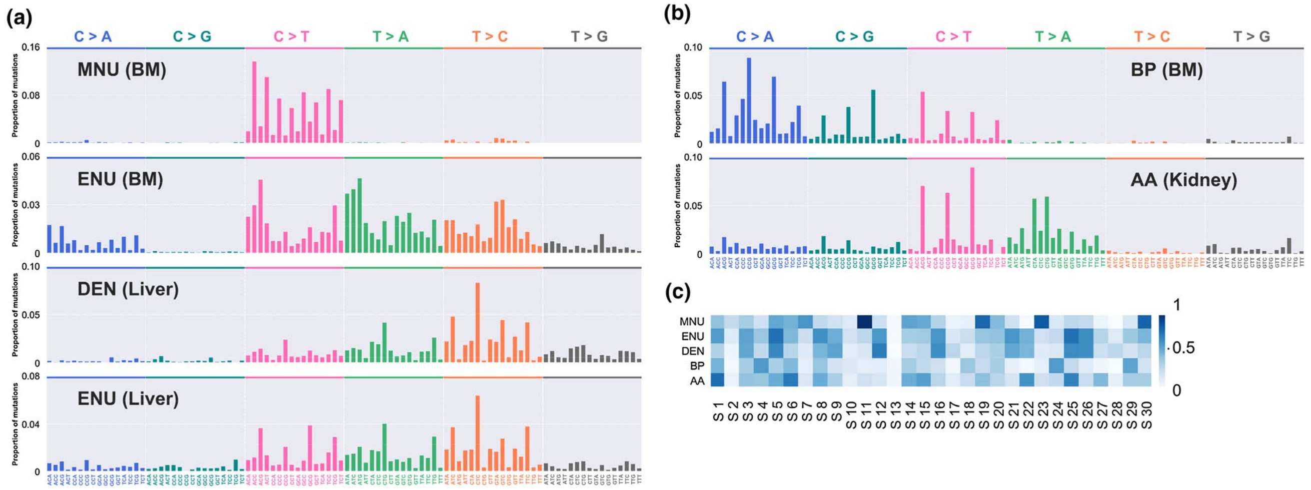

Fig. 4 The pattern of mutational signatures after exposure to mutagens in the 96-trinucleotide format in *gpt* delta mice. **a** The mutation patterns obtained from the analysis of samples exposed to MNU (BM, 25 mg/kg), ENU (BM and liver, 150 mg/kg), and DEN (liver, 80 mg/kg). MNU exhibited a similar pattern to signature 11 in COSMIC, which has peaks in the context of the trinucleotide consensus, NpCpY (where Y indicates a pyrimidine base). **b** The mutation patterns obtained from the analysis of samples exposed to BP (BM, 300 mg/kg) and AA (kidney, 10 mg/kg). **c** The cosine similarities between obtained mutation signatures of mutagens and COSMIC signatures were calculated and shown as a heatmap

0.54, and 0.51, respectively). Among these, signatures 4 and 29 were known to be caused by tobacco carcinogens. High CS values were observed for the signatures for AA and signatures 1, 6, 22, and 25 (0.70, 0.66, 0.58, and 0.63, respectively). The mutation patterns of A:T>T:A in AA signatures especially indicated that they were highly similar to the mutation pattern for signature 22 (CS: 0.94), caused by AA exposure. Interestingly, the G:C mutations observed in samples exposed to BP or AA were apparently more frequently observed in the CpG context. This might be attributable to the previously reported tendency of mutagens to bind to methylated CpG during in vitro and in vivo analysis in mice (Chen et al. 1998; Weisenberger and Romano 1999; Beal et al. 2015; O'Brien et al. 2016).

Discussion

We developed the Hawk-Seq™, a novel, highly accurate genome sequencing method, which enabled the high-throughput characterization of somatic mutations across large genomic regions. We proved that this method enabled the accurate and efficient genome-scale characterization of mutagen-induced somatic mutations. Additionally, because no additional exogenous barcodes are necessary, the flow of the library preparation process is as simple as that of standard Illumina library preparation. These characteristics are critical while evaluating a substantial number of mutagens during high-throughput analysis. Therefore, Hawk-Seq™ is extremely important for the accumulation of data of various environmental mutagens, thereby enabling us to achieve a systemic understanding of mutagenic profiles of mutagens.

The possibility of OBA during Hawk-Seq™ could be disadvantageous, as it might adversely affect the detection of true mutations. However, as indicated in our results, this possibility could be lowered sufficiently, by controlling IDAP per unit genome size. In our experiments, the rate could be lowered to ~ ≤ 1%, which was sufficiently lower than the standard deviation in mutation frequencies, even for the small bacterial genome. Because this possibility would decrease as the size of the target genome increases, Hawk-Seq™ can be applied for examining the overall mutation landscape across large genomic regions without any discernible influence of OBA, using various biological resources. Hawk-Seq™ is the simplest sequencing technique that can be used for this purpose. However, during the analysis of mutations in relatively small genomic regions, OBA might substantially influence the detection of mutations. In our experiments using the TA100 genome, this possibility increased in proportion to IDAP, and the detected mutation frequency in the ENU-exposed TA100 sample decreased accordingly (Fig. 2). Therefore, it is advisable to estimate the OBA rate using multiple indexed adapters per sample, while analysing relatively small genomic regions. Otherwise, it is recommended to use other sequencing techniques that utilize molecular barcodes (Schmitt et al. 2012; Gregory et al. 2016). Here, this possibility was sufficiently low (approx. ≤ 1%) at a concentration of 78 amol during bacterial genome analysis (approx. 5 Mbp genome). Considering the SE, this IDAP is sufficient for producing ~ 1 Gbp of dsDCS per sample (i.e. approx. × 200 coverage). Therefore, Hawk-Seq™ can be used for at least × 200 genomic coverage and analysis, without being influenced by OBA. As the genome of mammals is much larger than that of bacteria, the effect of OBA would be even lower, as compared to that of bacteria.

In analyses performed using *gpt* delta mice and TA100 samples, Hawk-Seq™ sensitively provided catalogues of genome-wide somatic mutations induced by 5 mutagens. In the Ames assay, the n-fold increase in mean colony number was the smallest in AA-treated samples, and the maximum *n*-fold increase was observed to be 7.26 in the group treated with 313 µg/tube of mutagens. The coefficient of variation (CV) for this group was 0.43. During Hawk-Seq™ analysis, the frequencies of A:T>T:A mutations increased by 57.7 fold in AA-treated samples (Supplementary Fig. S3), and the CV for this mutation pattern was 0.73. Meanwhile, in *gpt* assays, the minimum *n*-fold increase in mean colony number was 2.90 and was observed in kidneys of mice treated with 10 mg/kg of AA. The CV for this group was 0.35. However, in mutation spectra obtained using Hawk-Seq™, the n-fold increase in the mean A:T>T:A mutation frequency for this group was 29.5 and the associated CV was 0.14. Thus, Hawk-Seq™ would enable us to conduct a more sensitive and stable evaluation of mutations than the existing genotoxicity assays. Furthermore, we exhibited that in the TG animal model, Hawk-Seq™ could detect mutation induction in three organs, i.e. BM, liver, and kidney. As these organs are susceptible to the toxicity associated with mutagens and widely evaluated in the toxicological field (Lambert et al. 2005), mutation analysis by Hawk-Seq™ is considered as a valuable replacement for existing TG animal model assays. Thus, Hawk-Seq™ would enable the replacement of mutagenicity assay models using bacteria or mammalian cells and would dramatically improve the current system for evaluating the mutagenicity of chemical substances.

Additionally, the mutation catalogues obtained by Hawk-Seq™ enable the construction of refined mutation spectra and trinucleotide signatures, including the minor fractions of mutation patterns induced by each mutagen, which would help us to understand the mutagenic mechanisms of mutagens and their association with cancer in humans. For example, during the analysis of *gpt* delta mice exposed to AA, which was thought to mainly cause A:T>T:A mutations (Arlt et al. 2007), our results showed that G:C>A:T mutation frequencies were also increased. This probably reflects the mechanism of action of AA, which is known to form adducts on G bases (Arlt et al. 2002). The pattern of G:C>A:T mutations in 96-trinucleotide signatures also indicated that they were similar to signatures observed in AA-related cancers (Poon et al. 2015). However, the increase in the number of G:C>A:T mutations has generally been attributed to ageing. Our results suggested that mutations observed in these cancers could possibly be attributed to AA exposure. The signatures of MNU and BP exhibited similarities to signatures 11 and 4 (or 29), respectively. These results are thought to be reasonable, considering the mechanisms of these mutagens. However, in previous studies that analysed mutational signatures, the patterns associated with the signatures for MNU were not similar to patterns for signature 11 (Phillips 2018; Kucab et al. 2019). Although this could be attributed to several reasons, such as differences in cell type, it is difficult to achieve clarity using previous approaches. These results also indicated the usefulness of Hawk-Seq™ for precisely clarifying the mutational signatures in various cell types and directly linking mutagens with cancer in humans. The relatively small CS values between the pattern of BP and those of tobacco carcinogens were probably attributable to concurrent exposure to other mutagens in tobacco, such as aldehydes (Hecht 2008; Weng et al. 2018). This suggests that BP exposure is only one of the causes of mutagen-induced lung cancer. Therefore, obtaining mutational signatures with various mutagens using Hawk-Seq™ would help us to determine the signatures associated with human cancers and achieve a systemic understanding of the roles of mutagens in cancer development.

In BP- or AA-treated animals, mutations within G:C base pairs were frequently observed in the CpG context, but were not clearly observed in MNU, ENU, and DEN-treated samples, suggesting that this phenomenon is specific only to certain mutagens. Mutagens such as BP and aflatoxin B1 (AFB1) could reportedly preferentially form adducts with G bases in the methylated CpG context (Chen et al. 1998; Weisenberger and Romano 1999; Beal et al. 2015; O'Brien et al. 2016). Therefore, it can be hypothesized that these kinds of mutagens could possibly influence epigenetic regulation, by introducing mutations into the methylated genomic region. BP and AFB1 reportedly induced epigenetic alterations in cultured human liver cells (Tryndyak et al. 2018). BP also reportedly affected CpG methylation levels in zebrafish and adversely affected normal cellular development (Fang et al. 2013). This feature might influence the carcinogenicity or developmental toxicity of these materials. Therefore, an analysis to determine whether these agents targeted methylated CpG in vivo and thereby disturbed epigenetic regulation needs to be performed in the future.

Although these trinucleotide-based mutational signature analyses would increase the level of knowledge about mutagens and cancer, some questions about signatures remain unanswered. For example, the factors that determine characteristic trinucleotide patterns in each mutagen have remained largely unknown. In our analysis, unlike MNU, signatures attributable to ENU did not exhibit a pattern similar to that for signature 11 in mice, while their pattern was highly similar to that for signature 11 in *Salmonella*. One possible cause for this difference is the involvement of DNA repair enzymes. In bacteria, ethyl adducts on G and T bases were repaired efficiently by alkylguanine DNA transferase (AGT). However, the mammalian AGT enzyme is known to be relatively inefficient for the repair of O4-alkylT or O2-alkylT, as compared to the repair of these enzymes in bacteria; other repair systems were reportedly involved (Singer 1986; Jenkins et al. 2005; Nieminuszczy and Grzesiuk 2007). Additionally, the DNA sequence specificity of repair enzymes has been reported (Li et al. 2017). Therefore, the differences in relevant repair systems might cause differences in ENU-induced trinucleotide patterns. Therefore, the analysis of mutation patterns of various mutagens associated with relevant repair systems using experimental models is important. Thus, as Hawk-seq™ enables high-throughput analysis to be performed using multiple biological resources, the data obtained in these analyses would promote our understanding of mutational signatures.

Because Hawk-Seq™ provides a simple and high-throughput platform for accurate genome sequencing, it would be also useful for analysing clinical sequences. Yamashita et al. demonstrated that the number of somatic mutations in pre-cancerous tissues could be used as a marker for risk quantification in some types of cancer (Yamashita et al. 2018). Because the throughput ability of Hawk-Seq™ is superior to that of other accurate genome sequencing techniques, it would be especially useful in clinical applications for cancer risk assessment. The risk of emergence of neoplastic cells because of the accumulation of somatic mutation data for various pre-cancerous and cancerous tissues would become quantifiable.

This study is the first to use Hawk-Seq™ to derive genome-wide somatic mutation profiles of multiple mutagens in mice and identify some of the mutagenic features of chemical mutagens. The accumulation of these large-scale mutation data would provide clarity about the mutagenic features of mutational processes at extraordinary resolutions and their roles in human cancer development.

Acknowledgements We thank Dr. Sawako Kasamoto from the BioSafety Research Center Inc. for providing technical support, during our experiments involving *gpt* delta mice. We also thank Dr. Masayuki Yamane and Dr. Hiroshi Honda for their helpful advice regarding genotoxicity tests. We also thank Eurofins Genomics K.K. for participating in helpful discussions and providing technical support during NGS.

Compliance with ethical standards

Conflict of interest All authors are employees of Kao Corporation, the company that has applied for a patent for this method.

Ethical approval All animal experiment protocols were approved by the Animal Testing Committee at Kao Corporation, as indicated in "Materials and methods".

References

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al (2013a) Signatures of mutational processes in human cancer. Nature 500:415–421. https://doi.org/10.1038/nature12477

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR (2013b) Deciphering signatures of mutational processes operative in human cancer. Cell Rep 3:246–259. https://doi.org/10.1016/j.celrep.2012.12.008

Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata Tatsuhiro, Campbell PJ, Vineis P, Phillips DH, Stratton MR (2016) Mutational signatures associated with tobacco smoking in human cancer. Science 354:618–622. https://doi.org/10.1126/science.aag0299

Arlt VM, Stiborová M, Schmeiser HH (2002) Aristolochic acid as a probable human cancer hazard in herbal remedies: a review. Mutagenesis 17:265–277. https://doi.org/10.1093/mutage/17.4.265

Arlt VM, Stiborová M, vom Brocke J, Simões ML, Lord GM, Nortier JL, Hollstein M, Phillips DH, Schmeiser HH (2007) Aristolochic acid mutagenesis: molecular clues to the aetiology of Balkan endemic nephropathy-associated urothelial cancer. Carcinogenesis 28:2253–2261. https://doi.org/10.1093/carcin/bgm082

Beal MA, Gagné R, Williams A, Marchetti F, Yauk CL (2015) Characterizing Benzo[a]pyrene-induced lacZ mutation spectrum in transgenic mice using next-generation sequencing. BMC Genom 16:1–13. https://doi.org/10.1186/s12864-015-2004-4

Bigger CA, Pontén I, Page JE, Dipple A (2000) Mutational spectra for polycyclic aromatic hydrocarbons in the supF target gene. Mutat Res Fundam Mol Mech Mutagen 450:75–93. https://doi.org/10.1016/S0027-5107(00)00017-8

Chawanthayatham S, Valentine CC, Fedeles BI, Fox EJ, Loeb LA, Levine SS, Slocum SL, Wogan GN, Croy RG, Essigmann JM (2017) Mutational spectra of aflatoxin B 1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. Proc Natl Acad Sci USA 114:E3101–E3109. https://doi.org/10.1073/pnas.1700759114

Chen JX, Zheng Y, West M, Tang MS (1998) Carcinogens preferentially bind at methylated CpG in the p53 mutational hot spots. Cancer Res 58:2070–2075. https://doi.org/10.1089/lap.2012.0132

Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, Spudich GM, Brent S, Kulesha E, Marin-Garcia P, Smedley D, Birney E, Flicek P (2010) Ensembl variation resources. BMC Genom 11:293. https://doi.org/10.1186/1471-2164-11-293

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80–92. https://doi.org/10.4161/fly.19695

Costello M, Pugh TJ, Fennell TJ, Stewart S, Lichtenstein L, Meldrim JC, Fostel JL, Friedrich DC, Perrin D, Dionne D, Kim S, Gabriel SB, Lander ES, Fisher S, Getz G (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 41:e67. https://doi.org/10.1093/nar/gks1443

Fang X, Thornton C, Scheffler BE, Willett KL (2013) Benzo[a]pyrene decreases global and gene specific DNA methylation during zebrafish development. Environ Toxicol Pharmacol 36:40–50. https://doi.org/10.1016/j.etap.2013.02.014

Garraway LA, Lander ES (2013) Lessons from the cancer genome. Cell 153:17–37. https://doi.org/10.1016/j.cell.2013.03.002

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C et al (2007) Patterns of somatic mutation in human cancer genomes. Nature 446:153–158. https://doi.org/10.1038/nature05610

Gregory MT, Bertout JA, Ericson NG, Taylor SD, Mukherjee R, Robins HS, Drescher CW, Bielas JH (2016) Targeted single molecule mutation detection with massively parallel sequencing. Nucleic Acids Res 44:e22. https://doi.org/10.1093/nar/gkv915

Hecht SS (2008) Progress and challenges in selected areas of tobacco carcinogenesis. Chem Res Toxicol 21:160–171. https://doi.org/10.1021/tx7002068

Hoang ML, Kinde I, Tomasetti C, McMahon KW, Rosenquist TA, Grollman AP, Kinzler KW, Vogelstein B, Papadopoulos N (2016) Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. Proc Natl Acad Sci 113:9846–9851. https://doi.org/10.1073/pnas.1607794113

Hunter C, Smith R, Cahill DP, Stephens P, Stevens C, Teague J, Greenman C, Edkins S, Bignell G, Davies H, O'Meara S, Parker A, Avis T, Barthorpe S et al (2006) A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. Cancer Res 66:3987–3991. https://doi.org/10.1158/0008-5472.CAN-06-0127

Jenkins GJS, Doak SH, Johnson GE, Quick E, Waters EM, Parry JM (2005) Do dose response thresholds exist for genotoxic alkylating agents? Mutagenesis 20:389–398. https://doi.org/10.1093/mutage/gei054

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L (2013) Mutational landscape and significance across 12 major cancer types. Nature 502:333–339. https://doi.org/10.1038/nature12634

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA 108:9530–9535. https://doi.org/10.1073/pnas.1105422108

Kirkland D, Zeiger E, Madia F, Gooderham N, Kasper P, Lynch A, Morita T, Ouedraogo G, Morte JMP, Pfuhler S, Rogiers V, Schulz M, Thybaud V, Benthem J, Vanparys P, Worth A, Corvi R (2014) Can in vitro mammalian cell genotoxicity test results be used to complement positive results in the Ames test and help predict carcinogenic or in vivo genotoxic activity? I. Reports of individual databases presented at an EURL ECVAM Workshop. Mutat Res Genet Toxicol Environ Mutagen 775–776:55–68. https://doi.org/10.1016/j.mrgentox.2014.10.005

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645. https://doi.org/10.1101/gr.092759.109

Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP, Arlt VM, Phillips DH, Nik-Zainal S (2019) A compendium of mutational signatures of environmental agents. Cell 177:1–16. https://doi.org/10.1016/j.cell.2019.03.001

Lambert IB, Singer TM, Boucher SE, Douglas GR (2005) Detailed review of transgenic rodent mutation assays. Mutat Res 590:1–280. https://doi.org/10.1016/j.mrrev.2005.04.002

Langmead B, Salzberg S (2013) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li W, Hu J, Adebali O, Adar S, Yang Y, Chiou YY, Sancar A (2017) Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. Proc Natl Acad Sci USA 114:6752–6757. https://doi.org/10.1073/pnas.1706021114

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12. https://doi.org/10.14806/ej.17.1.200

Matsuda T, Takamune M, Matsuda Y, Yamada M (2013) A pilot study for the mutation assay using a high-throughput DNA sequencer. Genes Environ 35:53–56. https://doi.org/10.3123/jemsge.35.53

Matsumura S, Ito Y, Morita O, Honda H (2017) Genome resequencing analysis of *Salmonella typhimurium* LT-2 strains TA98 and TA100 for the establishment of a next-generation sequencing-based mutagenicity assay. J Appl Toxicol 37:1125–1128. https://doi.org/10.1002/jat.3463

Matsumura S, Fujita Y, Yamane M, Morita O, Honda H (2018) A genome-wide mutation analysis method enabling high-throughput identification of chemical mutagen signatures. Sci Rep 8:9583. https://doi.org/10.1038/s41598-018-27755-w

Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696. https://doi.org/10.1038/nrg2841

Mimaki S, Totsuka Y, Suzuki Y, Nakai C, Goto M, Kojima M, Arakawa H, Takemura S, Tanaka S, Marubashi S et al (2016) Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes. Carcinogenesis 37:817–826. https://doi.org/10.1093/carcin/bgw066

Mortelmans K, Zeiger E (2000) The Ames Salmonella/microsome mutagenicity assay. Mutat Res Fundam Mol Mech Mutagen 455:29–60. https://doi.org/10.1016/S0027-5107(00)00064-6

Nieminuszczy J, Grzesiuk E (2007) Bacterial DNA repair genes and their eukaryotic homologues: 3. AlkB dioxygenase and Ada methyltransferase in the direct repair of alkylated DNA. Acta Biochim Pol 54:459–468

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA et al (2012a) Mutational processes molding the genomes of 21 breast cancers. Cell 149:979–993. https://doi.org/10.1016/j.cell.2012.04.024

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A et al (2012b) The life history of 21 breast cancers. Cell 149:994–1007. https://doi.org/10.1016/j.cell.2012.04.023

O'Brien JM, Beal MA, Yauk CL, Marchetti F (2016) Next generation sequencing of benzo(a)pyrene-induced lacZ mutants identifies a germ cell-specific mutation spectrum. Sci Rep 6:36743. https://doi.org/10.1038/srep36743

Ohta T, Watanabe-Akanuma M, Yamagata H (2000) A comparison of mutation spectra detected by the *Escherichia coli* Lac + reversion assay and the *Salmonella typhimurium* His + reversion assay. Mutagenesis 15:317–323. https://doi.org/10.1093/mutage/15.4.317

Phillips DH (2018) Mutational spectra and mutational signatures: insights into cancer aetiology and mechanisms of DNA damage and repair. DNA Repair (Amst) 71:6–11. https://doi.org/10.1016/j.dnarep.2018.08.003

Poon SL, Huang MN, Choo Y, McPherson JR, Yu W, Heng HL, Gan A, Myint SS, Siew EY, Ler LD et al (2015) Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Med 7:38. https://doi.org/10.1186/s13073-015-0161-3

Richardson KK, Richardson FC, Crosby RM, Swenberg JA, Skopek TR (1987) DNA base changes and alkylation following in vivo exposure of *Escherichia coli* to *N*-methyl-*N*-nitrosourea or *N*-ethyl-*N*-nitrosourea. Proc Natl Acad Sci USA 84:344–348. https://doi.org/10.1073/pnas.84.2.344

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA (2012) Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci USA 109:14508–14513. https://doi.org/10.1073/pnas.1208715109

Singer B (1986) *O*-Alkyl pyrimidines in mutagenesis and carcinogenesis: occurrence and significance. Cancer Res 46:4879–4885

Stratton MR (2011) Exploring the genomes of cancer cells: progress and promise. Science 331:1553–1558. https://doi.org/10.1126/science.1204040

Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458:719–724. https://doi.org/10.1038/nature07943

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res 38:e159. https://doi.org/10.1093/nar/gkq543

Tryndyak V, Kindrat I, Dreval K, Churchwell MI, Beland FA, Pogribny IP (2018) Effect of aflatoxin B1, benzo[a]pyrene, and methapyrilene on transcriptomic and epigenetic alterations in human liver HepaRG cells. Food Chem Toxicol 121:214–223. https://doi.org/10.1016/j.fct.2018.08.034

Watson DE, Cunningham ML, Tindall KR (1998) Spontaneous and ENU-induced mutation spectra at the cII locus in Big Blue Rat2 embryonic fibroblasts. Mutagenesis 13:487–497. https://doi.org/10.1093/mutage/13.5.487

Weisenberger DJ, Romano LJ (1999) Cytosine methylation in a CpG sequence leads to enhanced reactivity with benzo[a]pyrene diol epoxide that correlates with a conformational change. J Biol Chem 274:23948–23955. https://doi.org/10.1074/jbc.274.34.23948

Weng MW, Lee HW, Park SH, Hu Y, Wang HT, Chen LC, Rom WN, Huang WC, Lepor H, Wu XR, Yang CS, Tang MS (2018) Aldehydes are the predominant forces inducing DNA damage and inhibiting DNA repair in tobacco smoke carcinogenesis. Proc Natl Acad Sci USA 115:E6152–E6161. https://doi.org/10.1073/pnas.1804869115

Yamashita S, Kishino T, Takahashi T, Shimazu T, Charvat H, Kakugawa Y, Nakajima T, Lee YC, Iida N, Maeda M, Hattori N et al (2018) Genetic and epigenetic alterations in normal tissues have differential impacts on cancer risk among tissues. Proc Natl Acad Sci USA 115:1328–1333. https://doi.org/10.1073/pnas.1717340115