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], and Osamu Morita[1]

[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

* To whom correspondence should be addressed. Tel: +81-285-68-7827; Fax: +81-285-68-7452; Email: matsumura.shouji@kao.com

SUPPLEMENTARY DATA

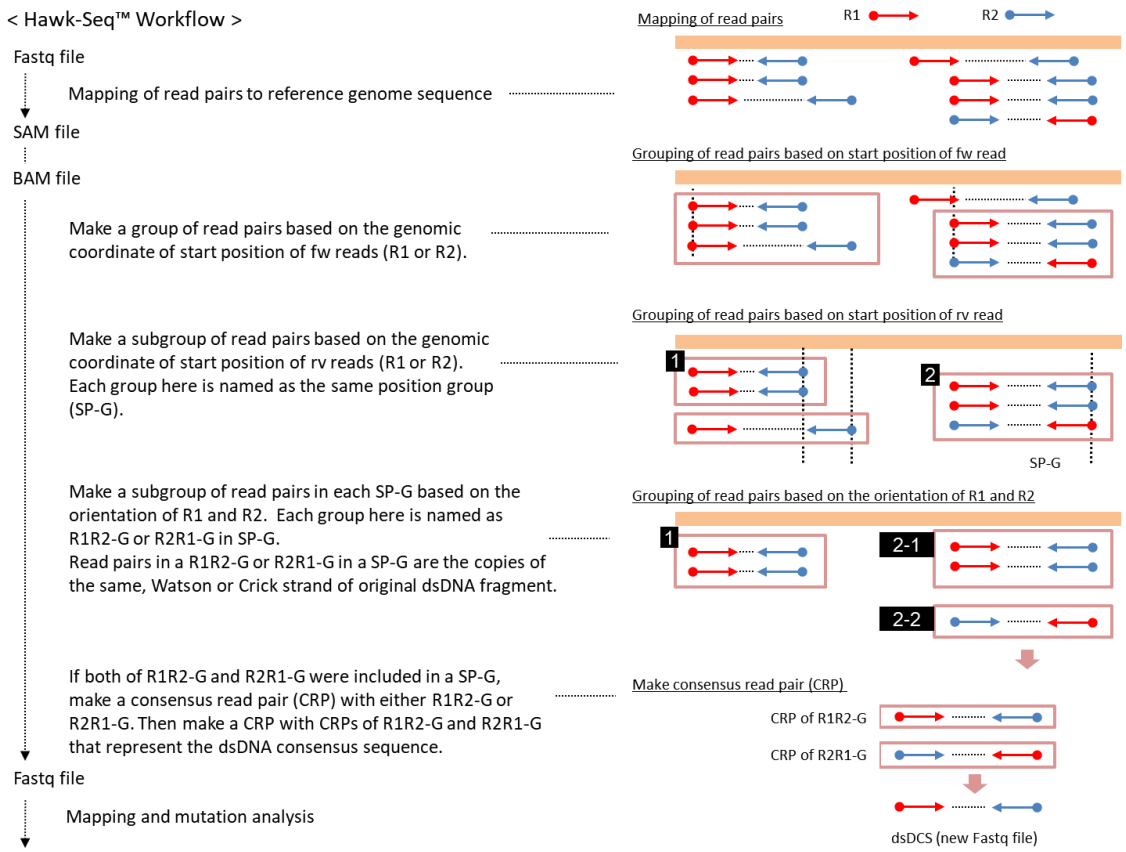


Fig. S1 An overview of Hawk-Seq™ workflow. The SP-Gs were made using information regarding the mapping positions of read pairs recorded in SAM files. Read pairs can be divided into 2 groups, i.e., R1R2-G, in which R1 is located on the left side of the reference genome sequence, as compared to the location of R2, and R2R1-G, in which R2 is conversely located on the left side. As their locations are determined by the strand that the read pairs derive, a SP-G includes both R1R2-G and R2R1-G, if they have read pairs derived from both double stranded DNAs. Therefore, each SP-G was further divided into R1R2-G and R1R2-G, and if both these groups are included in a SP-G, a dsDNA consensus sequence (dsDCS) was made. To make a dsDCS, the consensus read pair (CRP) within R1R2-G (R2R1-G) was prepared first; then, CRPs from both of these groups were compared.

a)

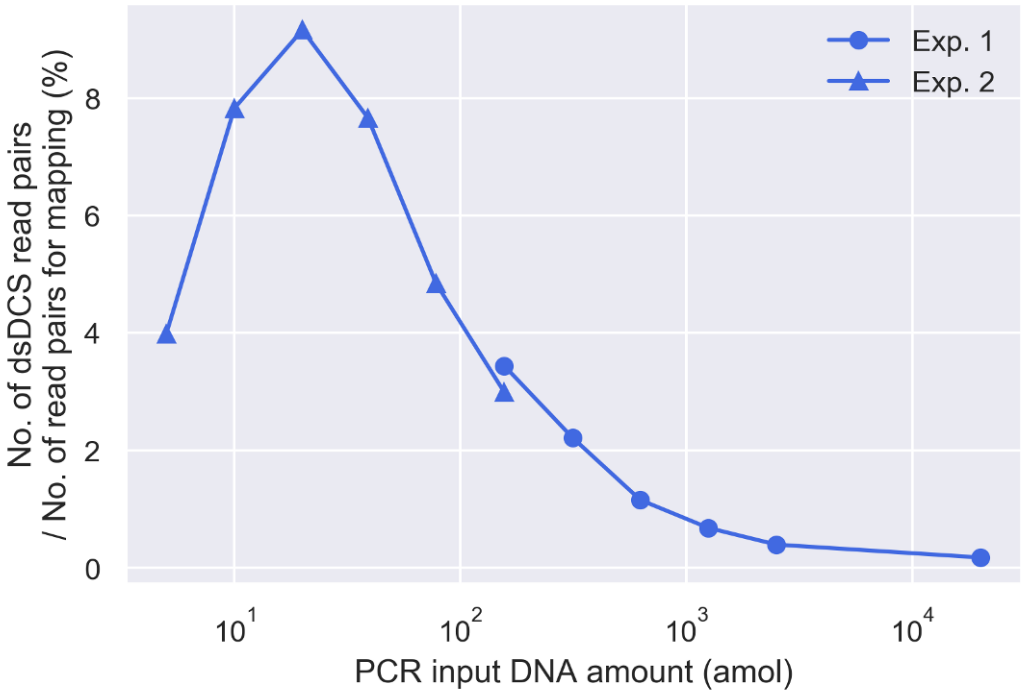


Fig. S2-a The optimization of Hawk-Seq™ parameters using 10 M of read pairs / sample subsampled from the original fastq file for each sample. As an indicator for SE, the mean rate of the number of generated dsDCS read pairs / 10 M read pairs was shown (mean of DMSO and ENU, i.e. n = 2). The optimal IDAP needed to generate peak sequence efficiency decreased proportionally with the amount of read pairs used, as compared to that used for the original experiment (Fig. 2a)

b)

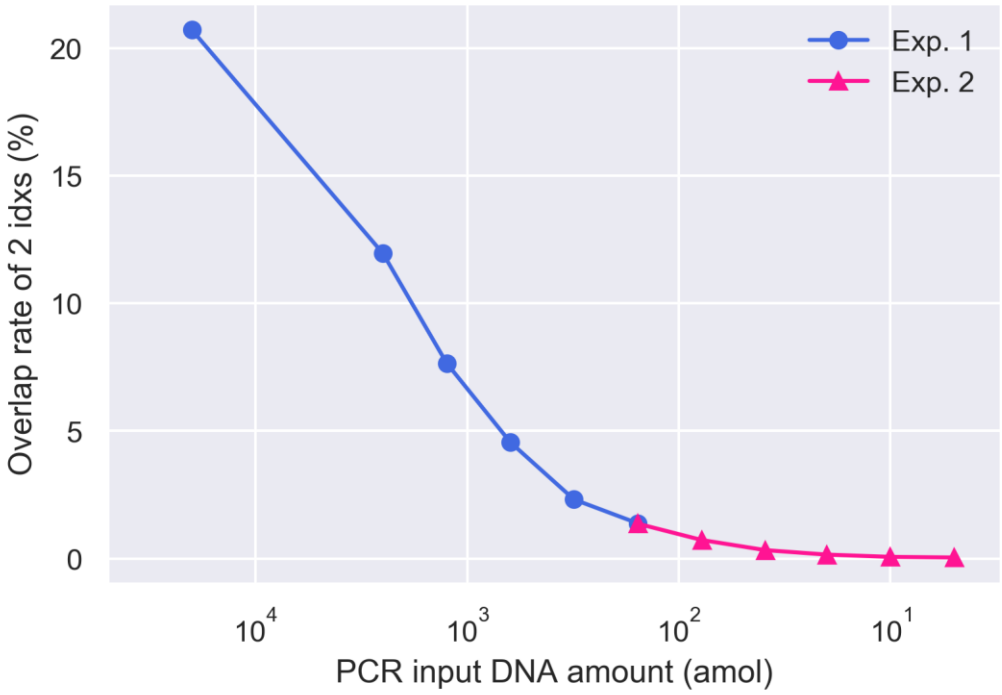
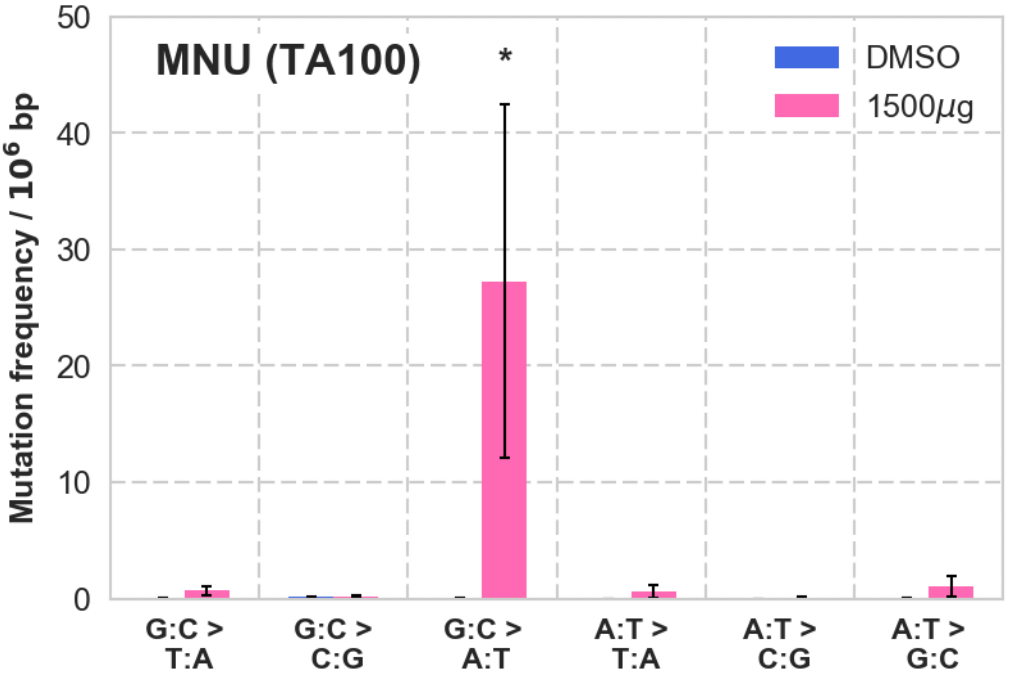
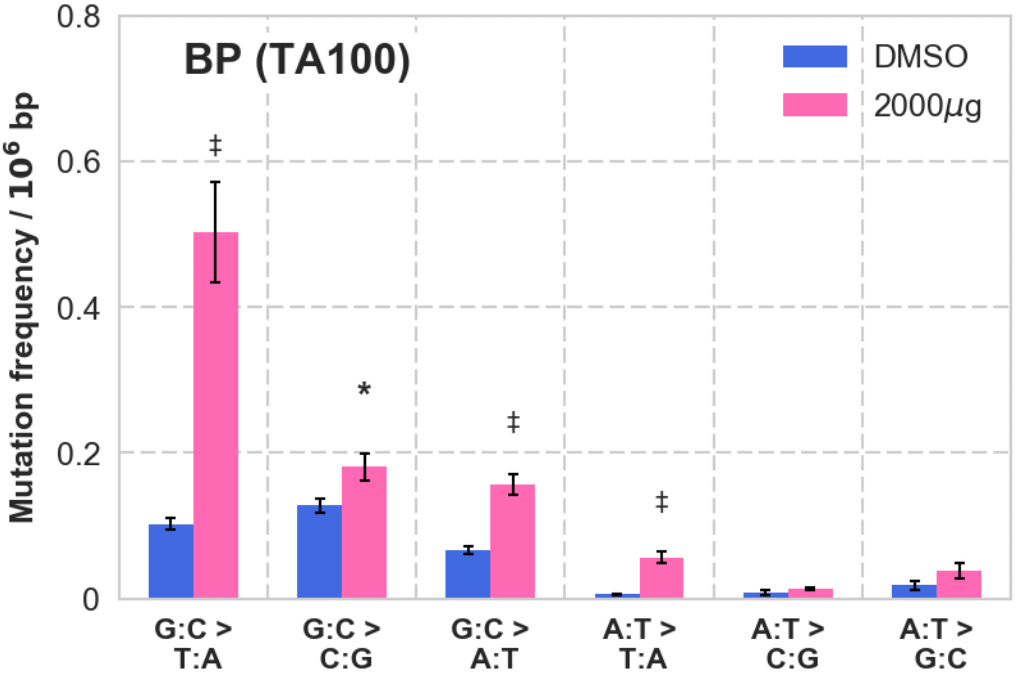


Fig. S2-b The mean overlap rate of 2 indexes was calculated by dividing the number of SPG-2idxs by the number of SP-Gs of 2 or more read pairs using the mapping results of 10 M read pairs subsampled from the original fastq file of libraries in Exp.1 and Exp.2 (mean of DMSO and ENU, i.e. n = 2). The mean rates of OBA were not changed with a change in the sequenced amount, but with the amount of IDAP

a)


MNU (TA100)
DMSO
1500μg
Mutation frequency / 10⁶ bp
50
40
30
20
10
0
*
G:C > T:A
G:C > C:G
G:C > A:T
A:T > T:A
A:T > C:G
A:T > G:C

b)


BP (TA100)
DMSO
2000μg
Mutation frequency / 10⁶ bp
0.8
0.6
0.4
0.2
0
‡
*
‡
‡
G:C > T:A
G:C > C:G
G:C > A:T
A:T > T:A
A:T > C:G
A:T > G:C

c)

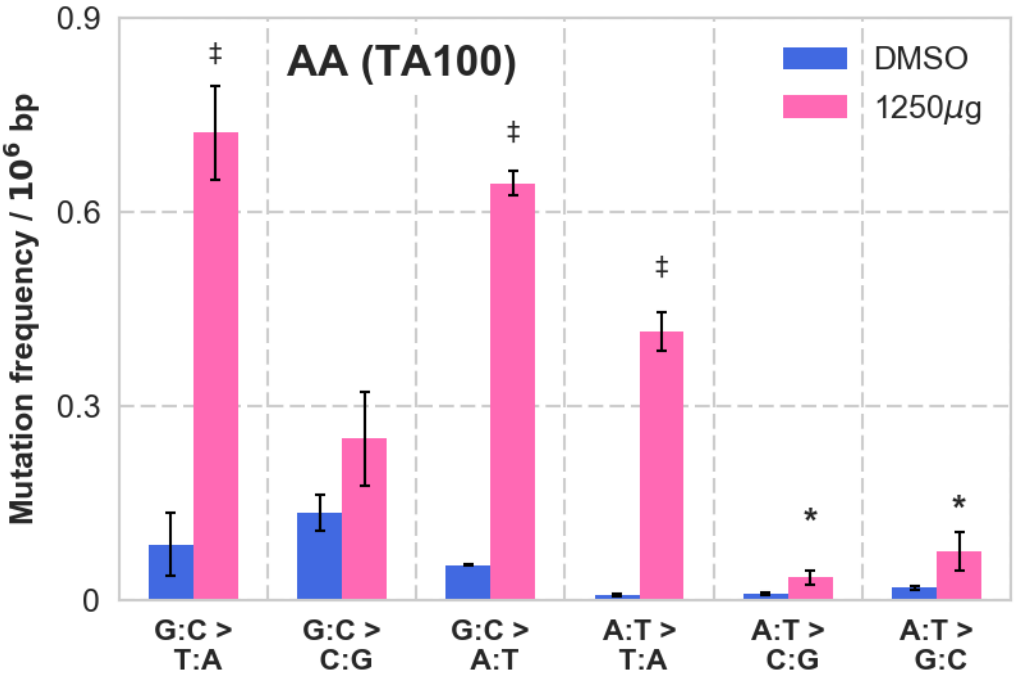
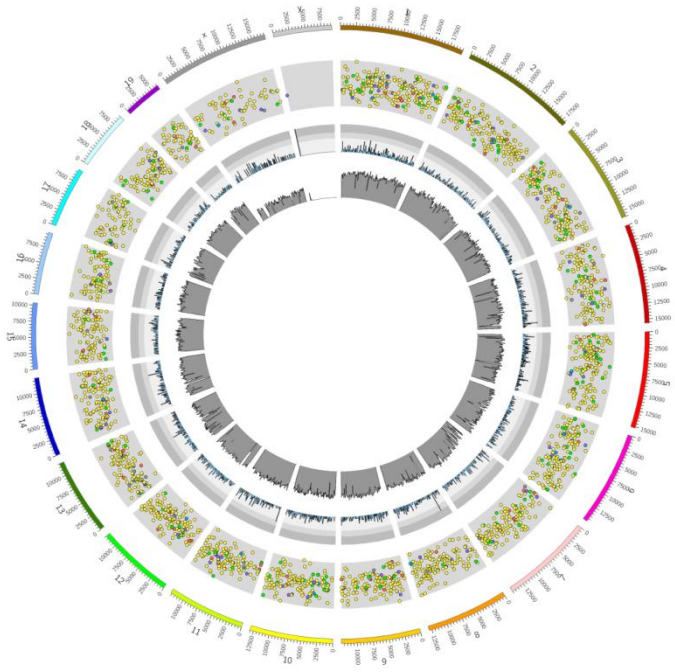
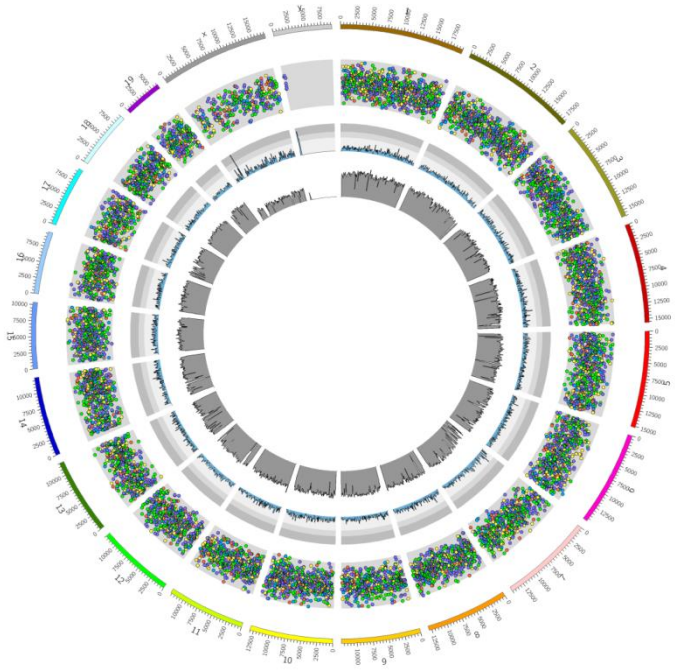


Fig. S3 Analysis of the mutation frequencies of TA100 DNA samples exposed to a) MNU (1500 μg / tube), b) BP (2000 μg / tube), and c) AA (1250 μg / tube). The mean ± S.D. for mutation frequencies per 10^6 of G:C or A:T base pairs of 3 independent biological experiments are presented. Asterisks and daggers indicate p-values in the Student's *t* test (*p < 0.05, †p < 0.01, and ‡p < 0.001)

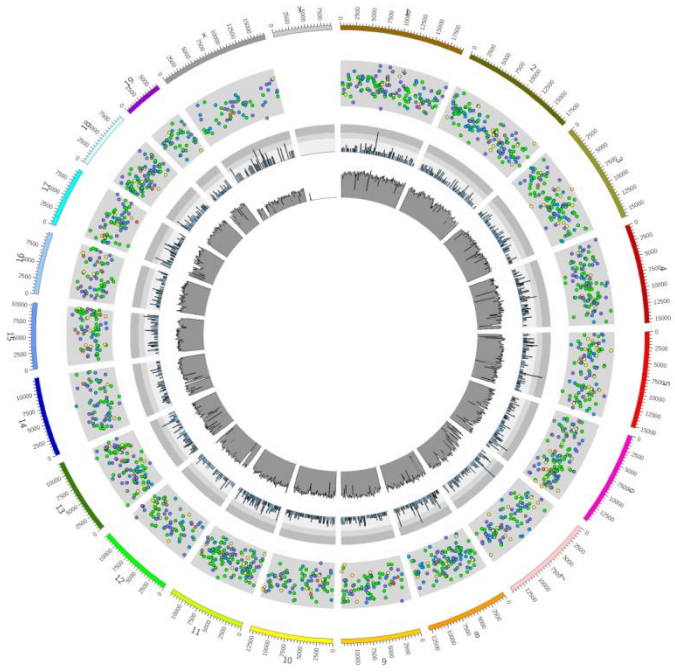
a)



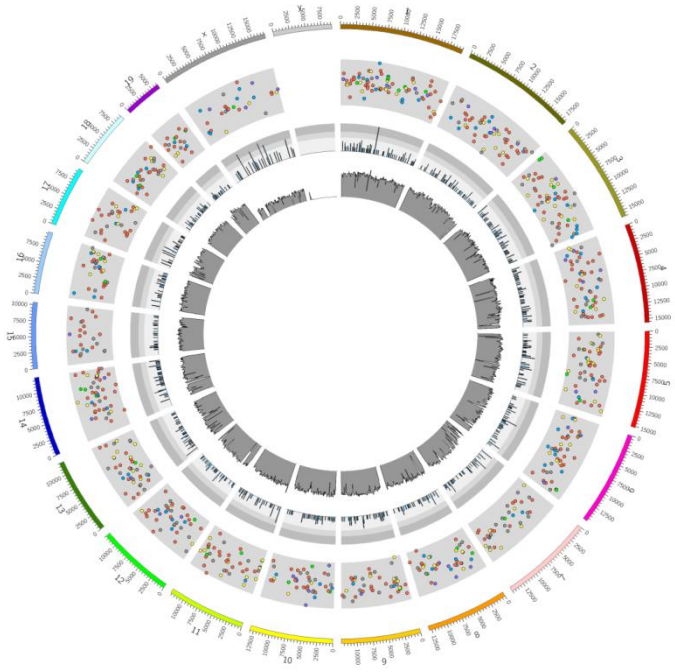
b)



c)



d)



e)

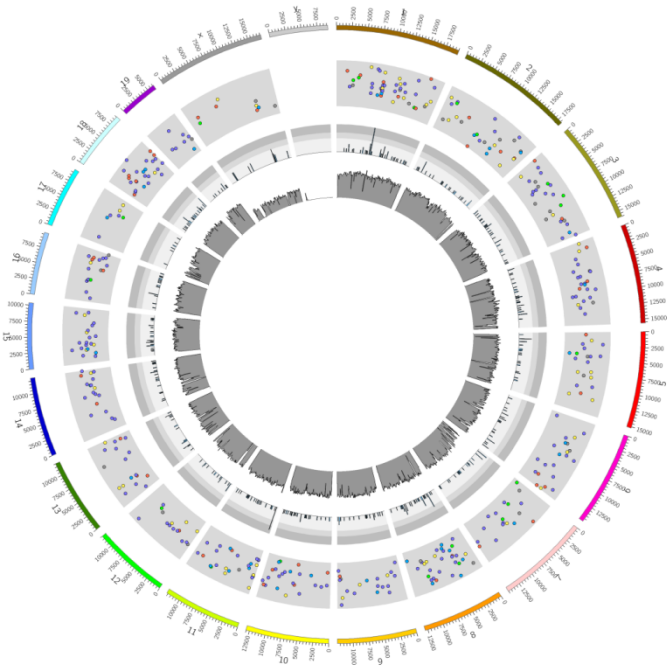
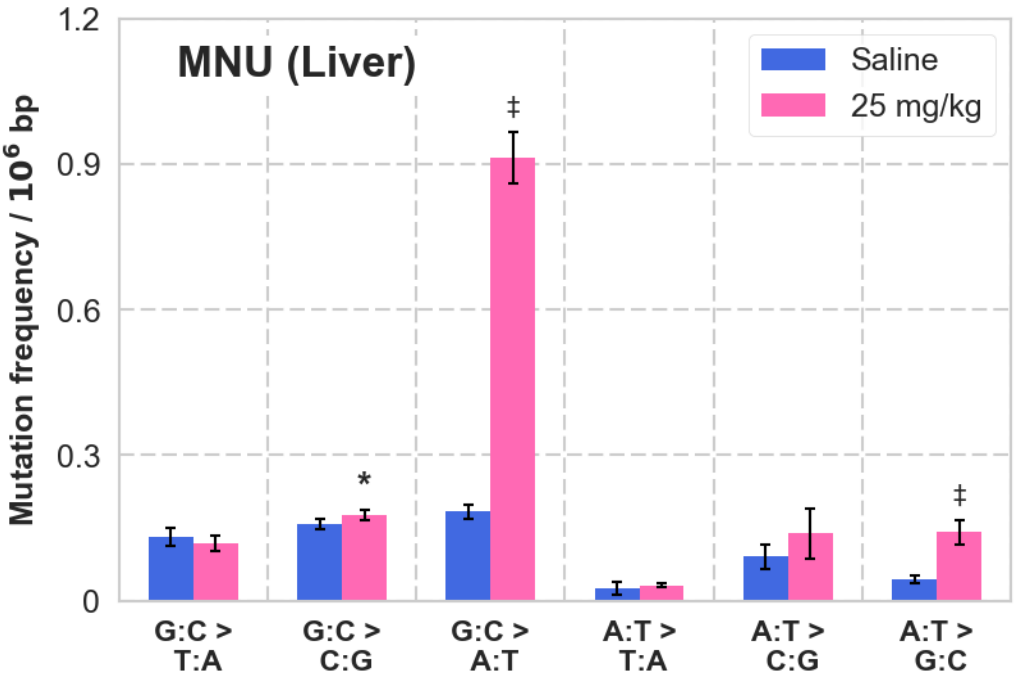
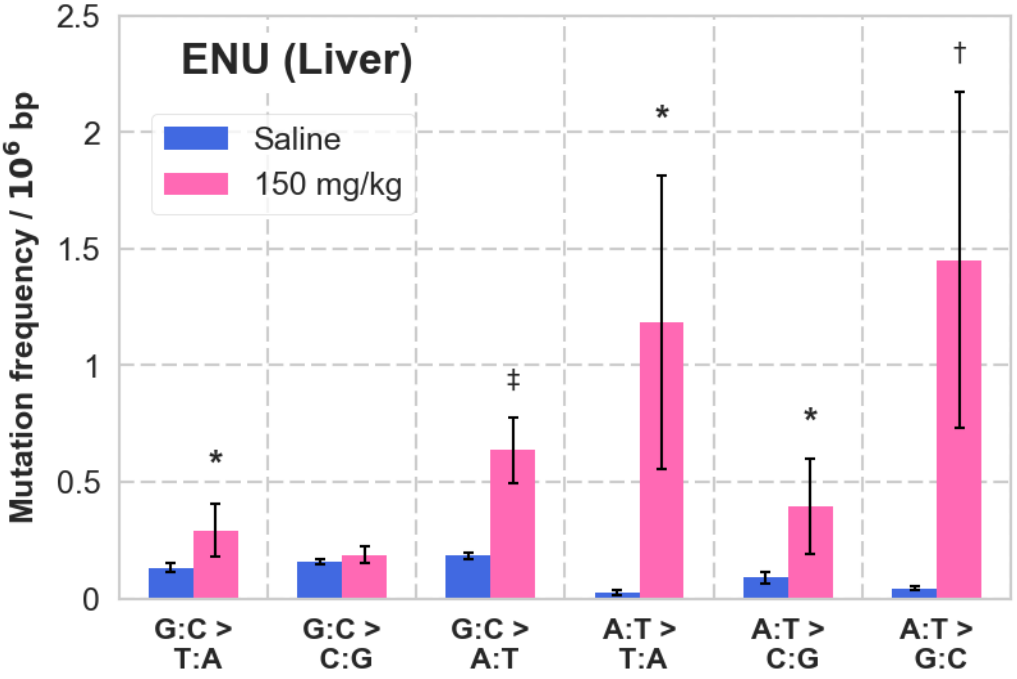


Fig. S4 The representative Circos plots for mutations observed in *gpt* delta mice exposed to a) MNU (BM, 25 mg / kg), b) ENU (BM, 150 mg / kg), c) DEN (Liver, 80 mg / kg), d) BP (BM, 300 mg / kg), and e) AA (Kidney, 10 mg/ kg). From the outside, the scatter plot for each mutation, histogram of mutation frequency normalized by read coverage per 1 M bp, and histogram of average depth per 1 M bp are displayed. The color of each mutation plot indicates the types of base substitutions; red: G:C > T:A, grey: G:C > C:G, yellow: G:C > A:T, purple: A:T > T:A, blue: A:T > C:G, and green: A:T > G:C

a)


MNU (Liver)
Saline
25 mg/kg
Mutation frequency / 10⁶ bp
1.2
0.9
0.6
0.3
0
*
‡
‡
G:C > T:A
G:C > C:G
G:C > A:T
A:T > T:A
A:T > C:G
A:T > G:C

b)


ENU (Liver)
Saline
150 mg/kg
Mutation frequency / 10⁶ bp
2.5
2
1.5
1
0.5
0
*
‡
*
*
†
G:C > T:A
G:C > C:G
G:C > A:T
A:T > T:A
A:T > C:G
A:T > G:C

c)

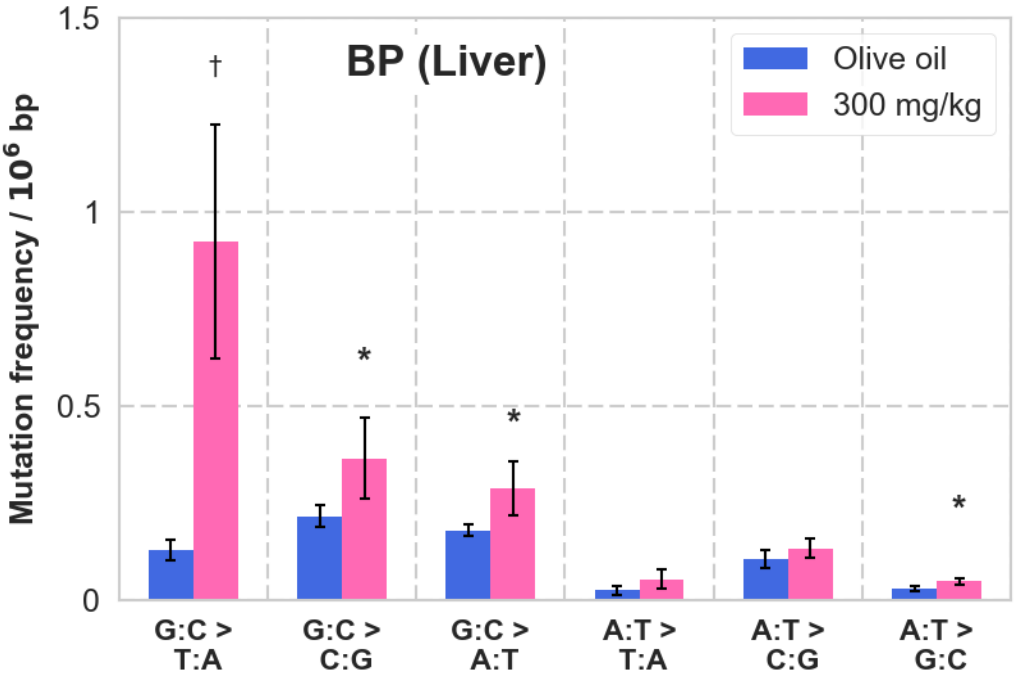


Fig. S5 Analysis of the mutation frequencies of liver DNA samples of *gpt* delta mice exposed to a) MNU (25 mg/kg/day), b) ENU (150 mg/kg/day) and c) BP (300 mg/kg/day). The means ± S.D. mutation frequencies per 10^6 of G:C or A:T base pairs are presented (n = 4). Asterisks and daggers indicate p-values in Student's *t* test (*p < 0.05, †p < 0.01, and ‡p < 0.001)

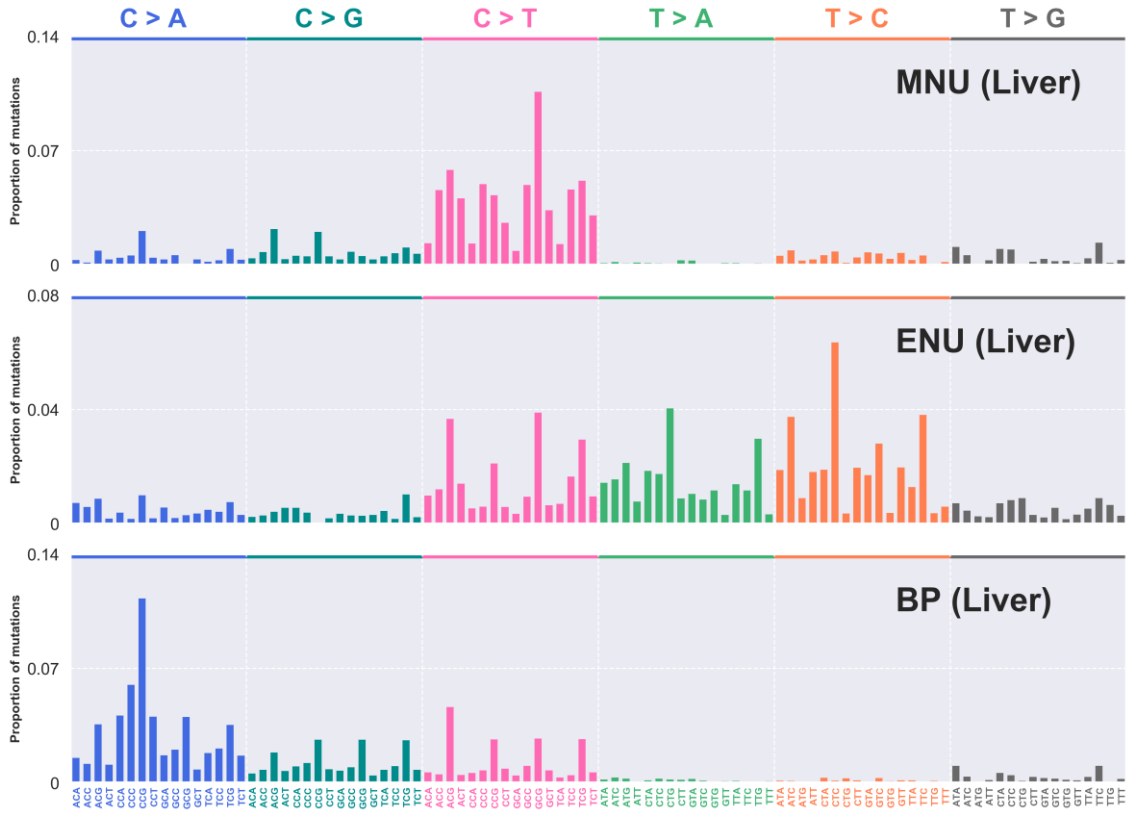


Fig. S6 The 96- trinucleotide mutational signatures obtained from the analysis of liver DNA samples of *gpt* delta mice exposed to MNU (25 mg/kg/day), ENU (150 mg/kg/day), and BP (300 mg/kg/day). The pattern induced by ENU exposure was highly similar to that induced by DEN in the liver (CS: 0.93, Fig. 4-a)

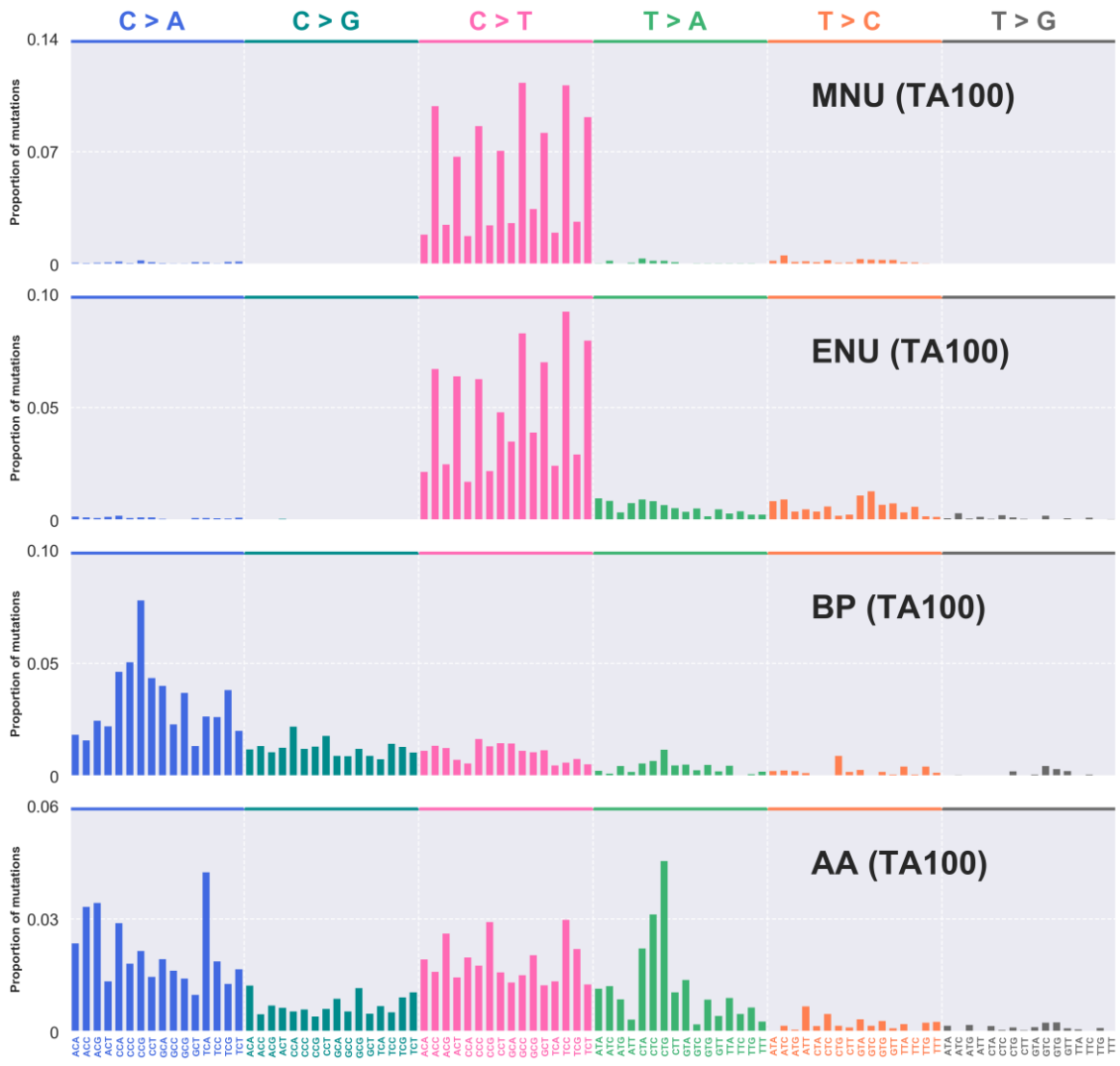


Fig. S7 The 96-trinucleotide mutational signatures obtained from the analysis of TA100 DNA samples exposed to MNU (1500 μg / tube), ENU (1000 μg / tube), BP (2000 μg / tube), and AA (1250 μg / tube). The pattern induced by ENU exposure was highly similar to that induced by signature 11 in human cancer patients (CS: 0.92)

SUPPLEMENTARY TABLES

a)

	IDAP (amol)	SE (%)	No. of read pairs / SP-G	No. of read pairs / IDAP (amol)
Exp.1	20000	0.71	1.03	0.002
	2500	1.79	1.07	0.023
	1250	2.40	1.09	0.034
	625	4.19	1.18	0.072
	313	6.42	1.32	0.140
	156	8.48	1.56	0.272
Exp. 2	156	5.35	1.51	0.374
	78	7.10	2.33	0.904
	39	6.31	3.64	1.629
	20	3.51	7.48	3.291
	10	1.90	16.44	6.769
	5	0.80	34	13.83

b)

	IDAP (amol)	SE (%)	No. of read pairs / SP-G	No. of read pairs / IDAP (amol)
Exp. 1	20000	0.18	1.01	0
	2500	0.40	1.02	0
	1250	0.68	1.03	0.01
	625	1.16	1.04	0.02
	313	2.22	1.09	0.03
	156	3.44	1.14	0.06
Exp. 2	156	3.00	1.14	0.06
	78	4.84	1.25	0.13
	39	7.66	1.55	0.26
	20	9.16	2.21	0.5
	10	7.83	4.12	1
	5	3.99	8.59	2

Supplementary Table S1

A comparison of the number of read pairs per SP-G (Parameter 1), number of read pairs per IDAP (Parameter 2), and SE are shown. (a) The results of original experiments using all sequence output data. (b) The results of experiments using 10 M of read pairs subsampled from the original sequence output data. Although the IDAP to be used for producing the highest SE changed according to the sequence amount used, the parameters 1 and 2 did not change at optimal conditions. At optimal conditions (i.e. 156 – 39 amol in Table-a or 39 – 10 amol in Table-b), the parameter 1 was approximately 1.5 – 4 and the parameter 2 was 0.2 – 2.

Materials	No. of colonies Mean	S.D.	
MNU (μg / plate)			
0	114.7	5.9	
750	3572.3	299.7	
1000	2936.0	29.3	
1250	1860.7	234.6	G
1500	979.0	160.5	G
ENU (μg / plate)			
0	119.7	8.1	
62.5	207.3	14.2	
125	381.3	54.2	
250	1389.0	82.2	
500	3950.7	178.6	
1000	2534.0	793.2	
BP (μg / plate)			
0	108.0	3.0	
250	722.0	52.9	
500	792.0	63.7	
1000	1042.0	188.9	
1500	1406.3	173.8	P
AA (μg / plate)			
0	105.7	1.5	
78	374.7	14.3	
156	602.7	16.5	
313	766.7	32.6	
625	118.0	139.5	G

Supplementary Table S2

The summary of Ames test results using TA100 is shown. TA100 was exposed to MNU, ENU, BP, and AA, and the No. of revertant colonies were counted. The mean ± S.D. for 3 independent biological experiments are presented. G: growth inhibition. P: precipitate of test materials.

Supplementary Table S3. The summary of *gpt* assay in additional organs (n = 4 or 5).

Materials	Administration route (vehicle)	Organ	Dose (mg/kg/day)	No. of animals for *gpt* assay (for Hawk-Seq™)	Mutant frequency (× 10^{-6})		
					Mean	S.D.	Statistics
MNU	i.p. (Saline)	Liver	0	5 (4)	3.85	3.19	
			12.5	5 (4)	4.5	1.21	
			25	5 (4)	9.6	4.03	*(D)
ENU	i.p. (Saline)	Liver	0	5 (4)	3.85	3.19	
			75	5 (4)	6.61	2.73	
			150	5 (4)	23.48	18.58	*(S)
DEN	i.p. (Saline)	Bone Marrow	0	5	2.36	1.26	
			40	5	2.24	0.94	
			80	4	3.08	1.66	
BP	p.o. (Olive oil)	Liver	0	5 (4)	2.43	0.81	
			150	5 (4)	9.14	2.35	*(S)
			300	5 (4)	19.91	9.79	*(S)
AA	p.o. (Olive oil)	Liver	0	5	3.52	1.67	
			5	5	4.06	4.66	
			10	4	3.59	1.18	

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

Supplementary Table S4. The overlap rate of 2 indexes (4 samples).

	No. of SPG-2idxs / No. of SP-Gs with 2 or more read pairs (%)			
	Saline		ENU 150 mg/kg	
Sample No.	1	2	1	2
Chr 1	0.023	0.027	0.022	0.018
2	0.026	0.025	0.022	0.020
3	0.025	0.023	0.025	0.017
4	0.026	0.023	0.022	0.016
5	0.022	0.026	0.024	0.019
6	0.023	0.027	0.021	0.018
7	0.023	0.019	0.020	0.017
8	0.023	0.025	0.021	0.020
9	0.026	0.023	0.021	0.019
10	0.022	0.024	0.017	0.020
11	0.023	0.026	0.026	0.022
12	0.024	0.024	0.024	0.018
13	0.028	0.025	0.024	0.023
14	0.022	0.028	0.027	0.021
15	0.025	0.026	0.020	0.018
16	0.021	0.026	0.027	0.020
17	0.030	0.027	0.023	0.020
18	0.024	0.028	0.020	0.021
19	0.026	0.026	0.022	0.020
X	0.026	0.021	0.020	0.019
Y	0.034	0	0.038	0
MT	0.127	0	0	0
Mean *	0.024	0.025	0.022	0.019

*: Mean overlap rate in all chromosomes except for Y and MT

Supplementary Table S5. The summary of annotations of mutations detected in TA100 exposed to mutagens (3 samples per mutagen).

	MNU			ENU			BP			AA		
Sample No.	1	2	3	1	2	3	1	2	3	1	2	3
Total	16631	27749	9446	35672	27021	32779	469	406	459	965	947	1128
CDS	14435	24099	8279	31215	23534	28634	406	341	400	820	823	988
Non-synonymous[1]	9646	16097	5574	21010	15965	19293	270	221	284	598	611	744
Synonymous[2]	4789	8002	2705	10205	7570	9341	136	120	116	222	212	244
Ratio (1 / 2)	2.01	2.01	2.06	2.06	2.11	2.07	1.99	1.84	2.45	2.69	2.88	3.05