Supplementary Figure Legends

Supplementary Fig. S1: Comparison of different normalization and single-nucleotide variant calling strategies. Variant calling with respect to primary cell normalization. Venn diagrams show the overlap of variants called in glycidamide (GA)-derived clones after normalization to three different batches of primary cells (Prim_1, Prim_2, and Prim_3).

Supplementary Fig. S2: Growth curves of Hupki MEFs. Primary cells were either left untreated (Spont) or were exposed to acrylamide (ACR±S9) or glycidamide (GA). X-axis represents days in culture. Y-axis represents the cumulative doubling populations. The dashed vertical line represents the threshold of p-value < 0.05. Arrow: compound exposure; S*: senescence; SBI: senescence bypass/immortalization.

Supplementary Fig. S3: Mutation spectra derived from exome sequencing data from immortalized Hupki MEF clones derived from exposure to (**A**) acrylamide (ACR) or (**B**) glycidamide (GA), or (**C**) by spontaneous immortalization (Spont). X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations in each context.

Supplementary Fig. S4: Illustration of the transcription strand bias derived from the analysis of exome sequencing data from immortalized Hupki MEF cell lines. GA: glycidamide-derived clones; ACR: acrylamide-derived clones; Spont: spontaneously immortalized clones. The six mutation types are represented by different colors. For each mutation type, the number of mutations occurring on the transcribed (T) and non-transcribed (N) strand, as well as the p-values for strand bias is shown on the y-axes. The dashed grey line in each graph indicates the p-values for strand bias for each mutation type. The horizontal, dashed black line represents a significance threshold of $p < 0.05$.

Supplementary Fig. S5: Distribution of mutations based on their allelic frequencies in the five glycidamide (GA)-derived clones (left). Mutations in individual cell lines were ranked and plotted based on decreasing allelic frequency. Percentage of mutations with allelic frequency between 25% and 75% is indicated. Percentages of the six mutation types, color-coded, among all mutations identified in GA clones (right). The overall mutation number for each sample is indicated in the centre of the pie chart.

Supplementary Fig. S6: Mutation type and mutation spectra analysis with respect to variant allele frequency (VAF). The analysis was carried out using exome sequencing data from immortalized Hupki MEF clones derived from exposure to glycidamide. Top left: Mutation counts were stratified into three VAF bins ([0-33% = low VAF]; [34-66% = medium VAF]; [67-100% = high VAF]). Top right: The relative contribution of the six mutation types to the overall number of mutations in each VAF bin is shown on the y-axis. Bottom panel: Mutation spectra (left) and strand bias (right) analysis for the different VAF bins. Mutation spectra analysis: X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations. The counts for each mutation type are indicated in parentheses. Strand bias analysis: For each mutation type, the number of mutations occurring on the transcribed and non-transcribed strand is shown on the y-axis. T: transcribed strand; N: non-transcribed strand.

Supplementary Fig. S7: The 'baiting' clean-up of background signature 17 and the quantification of its efficiency. COSMIC signature 17 (top track) marked by the arrows observed in GA mutation spectra as well as in GA-mutational signature before and after baiting (clean). The heat-map table on the right indicates the final proportionate reduction of signature 17-specific peaks after re-running the NMF with signature 17-rich ICGC ESAD data sets listed in the Supplementary Materials and Methods section.

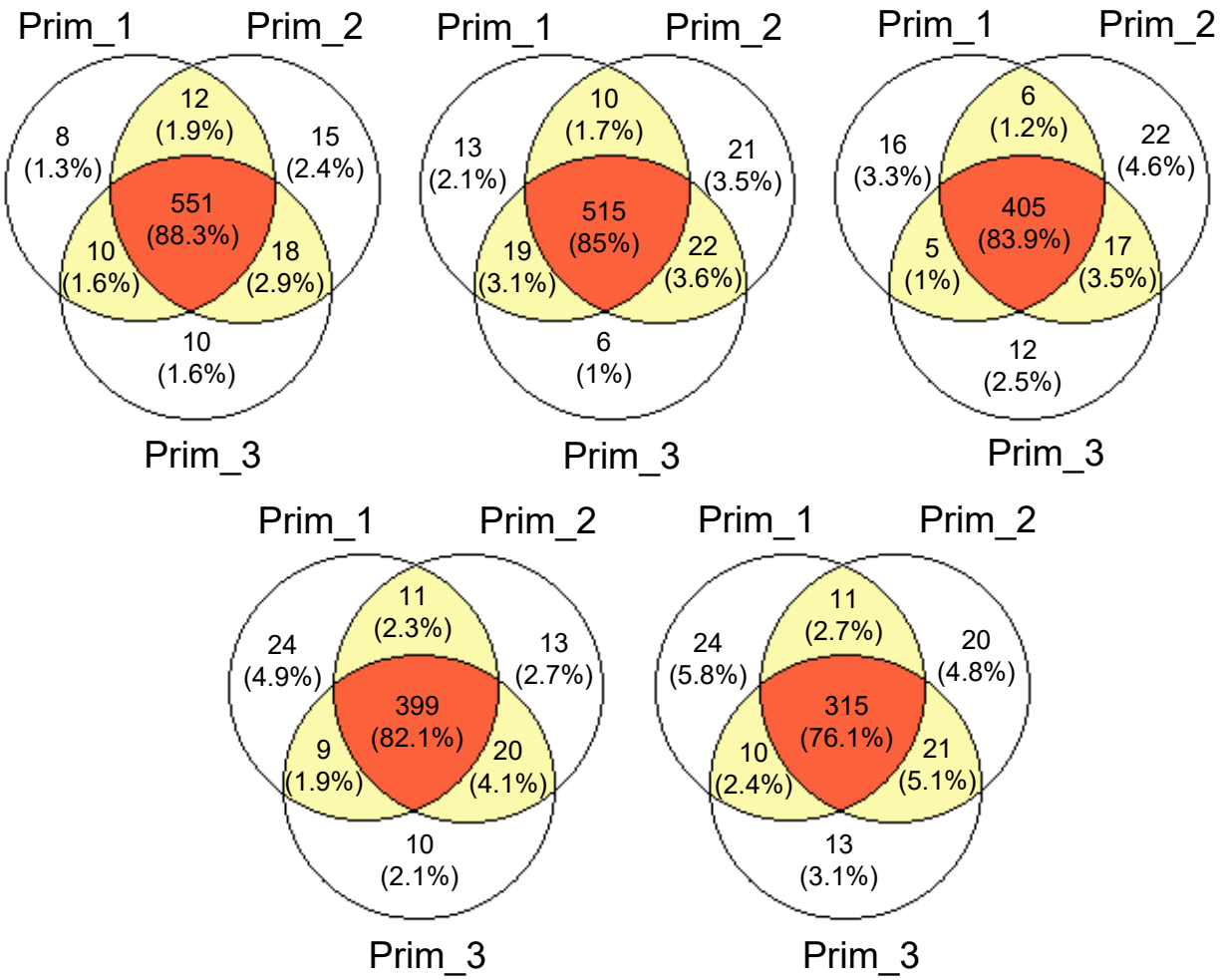
Supplementary Fig. S8: (**A**) The structures of N7-GA-Gua and N3-GA-Ade adducts analyzed by LC-MS/MS. (**B**) Representative multiple-reaction monitoring chromatograms (relative signal intensity vs time) for N7-GA-Gua and N3-GA-Ade adducts in DNA from ACR treatment in the presence of S9 fraction (ACR+S9) and GA-treated (GA) primary Hupki MEF. Internal standards (IS) were added in amounts of 1000 fmol for N7-GA-Gua and 200 fmol for N3-GA-Ade.

Supplementary Fig. S9: T:A>A:T enriched mutational signatures used for cosine similarity analysis (see Fig. 3D). The individual signatures were originally derived from human cancer sequencing data or experimental models (animal bioassays, cell lines) of carcinogen exposure. X-axis represents the trinucleotide sequence context. Y-axis represents the frequency distribution of the mutations. The predominant trinucleotide context for T:A>A:T mutations is indicated by an arrow in the signature landscape. AA: aristolochic acid; DMBA: 7,12-dimethylbenz[a]anthracene.

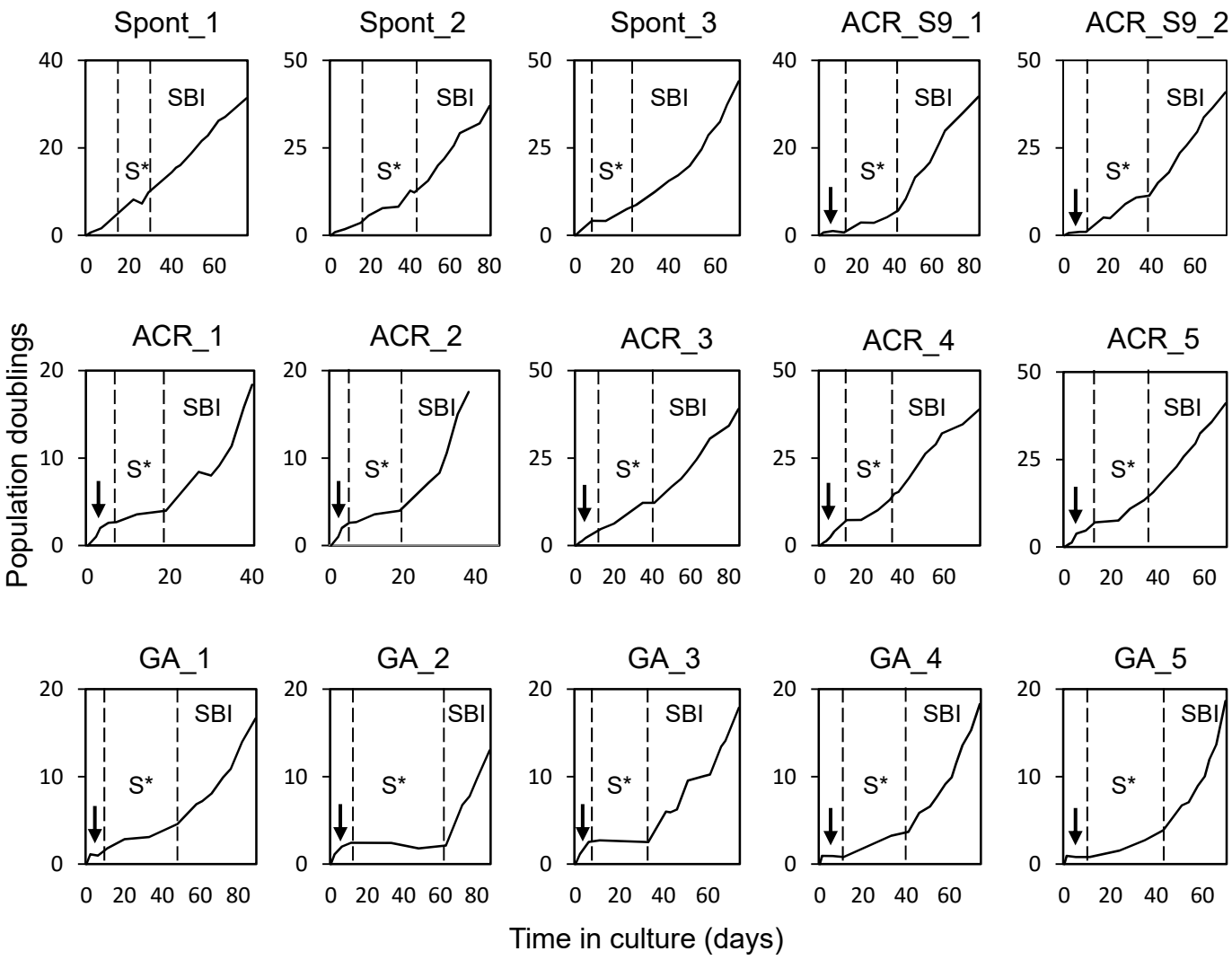
Supplementary Fig. S10: (**A**) Scatter plots show the measure of correlation of the GA-signature versus B[a]P-signature (used to reconstruct COSMIC signature 4) in PCAWG lung adenocarcinomas (ADCA), lung squamous cell carcinomas (SCC) and hepatocellular

carcinomas (HCC). (**B**) Bar-plots representing the proportion of the assignment of the experimental GA_Exp and B[a]P_Exp signatures in lung adenocarcinomas, lung squamous cell carcinomas and hepatocellular carcinomas from the PCAWG data set. The asterisk denotes liver HCC samples harboring GA-signature only (no B[a]P-signature detected), indicating possible dietary or occupational exposure. Full list of these samples is accessible from Suppl. Table S5.

Suppl. Fig. S1

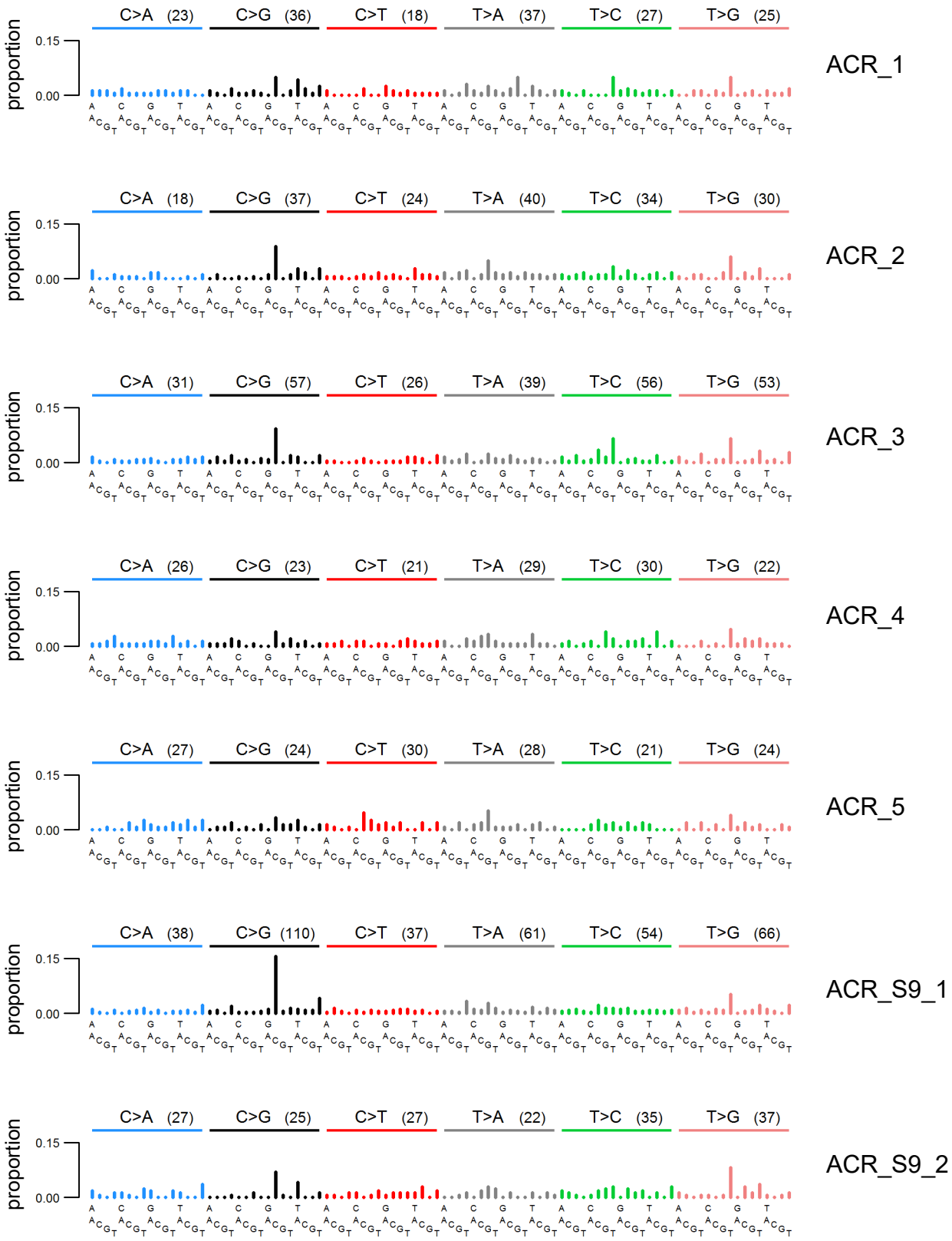

Prim_1
Prim_2
Prim_3
8 (1.3%)
12 (1.9%)
15 (2.4%)
551 (88.3%)
10 (1.6%)
18 (2.9%)
10 (1.6%)
13 (2.1%)
10 (1.7%)
21 (3.5%)
515 (85%)
19 (3.1%)
22 (3.6%)
6 (1%)
16 (3.3%)
6 (1.2%)
22 (4.6%)
405 (83.9%)
5 (1%)
17 (3.5%)
12 (2.5%)
24 (4.9%)
11 (2.3%)
13 (2.7%)
399 (82.1%)
9 (1.9%)
20 (4.1%)
10 (2.1%)
24 (5.8%)
11 (2.7%)
20 (4.8%)
315 (76.1%)
10 (2.4%)
21 (5.1%)
13 (3.1%)

Suppl. Fig. S2


Spont_1
Spont_2
Spont_3
ACR_S9_1
ACR_S9_2
ACR_1
ACR_2
ACR_3
ACR_4
ACR_5
GA_1
GA_2
GA_3
GA_4
GA_5
SBI
S*
Population doublings
Time in culture (days)

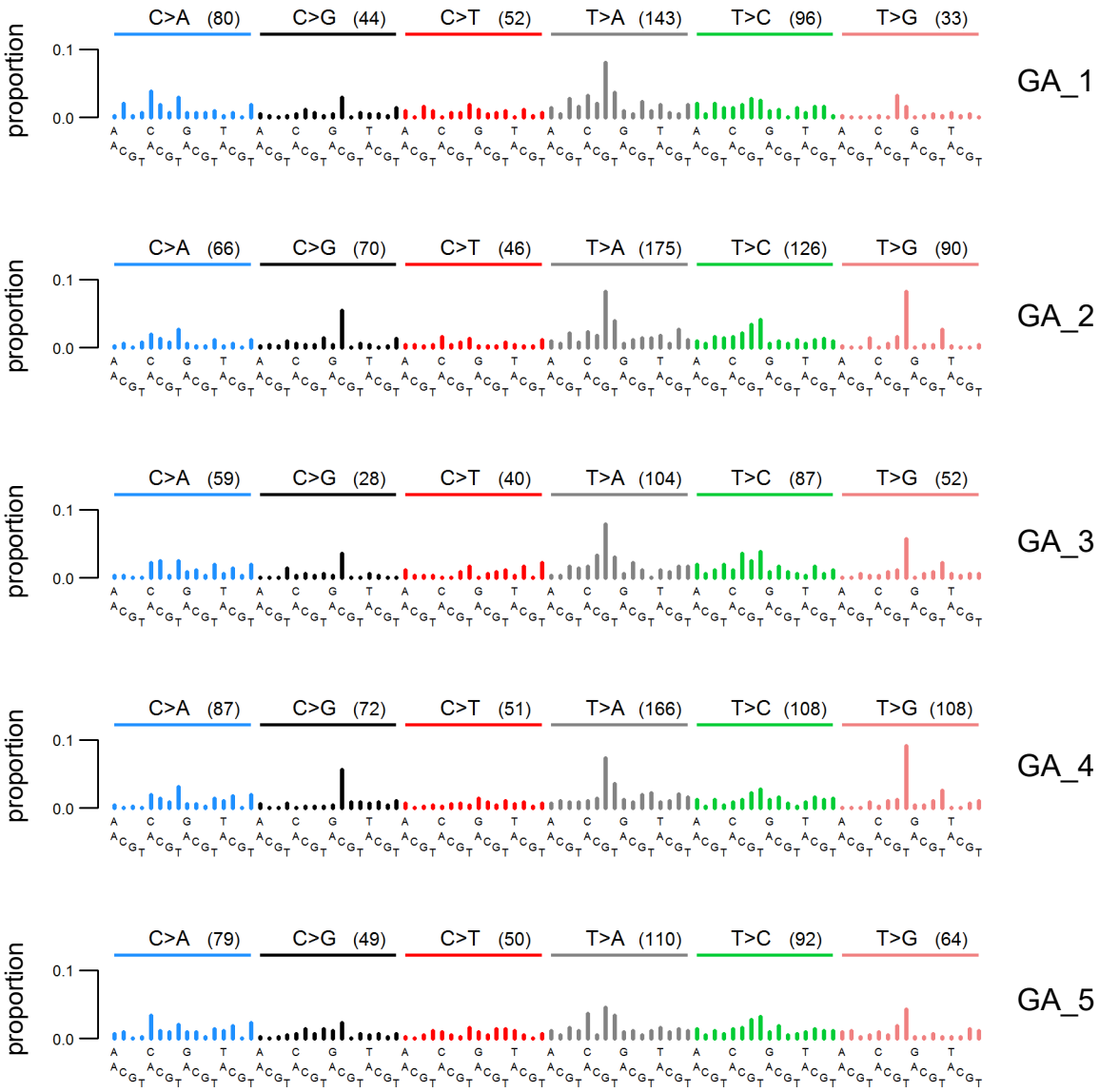
Suppl. Fig. S3

A



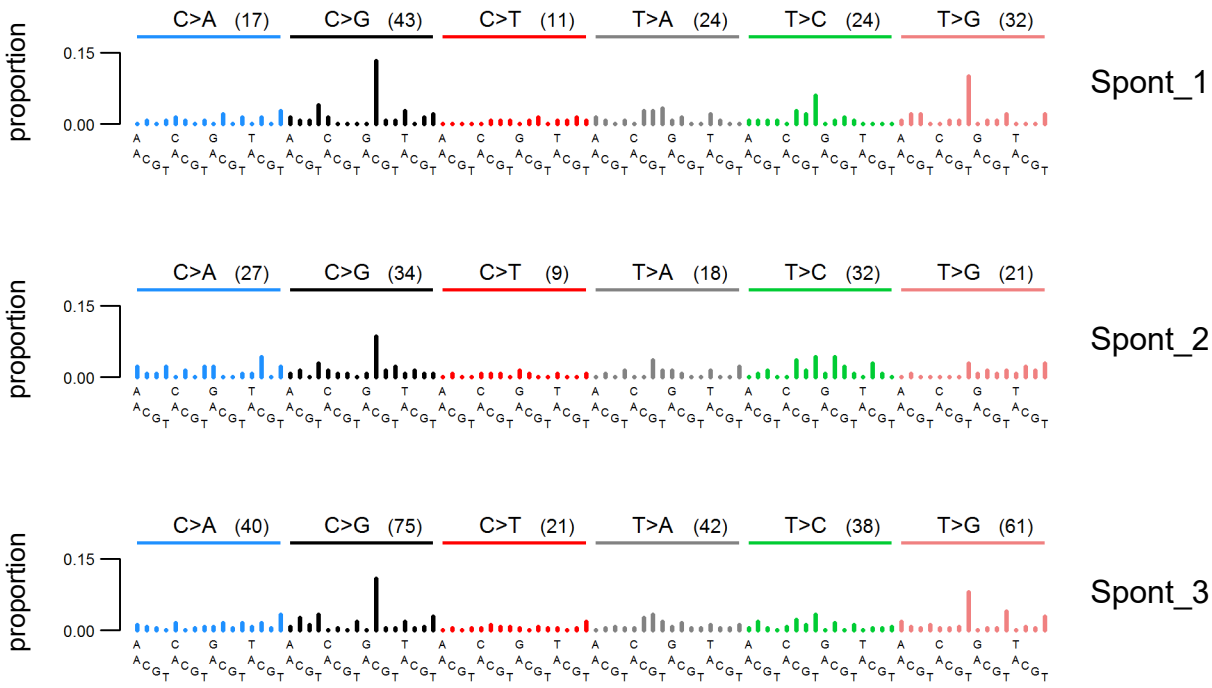
Suppl. Fig. S3

B

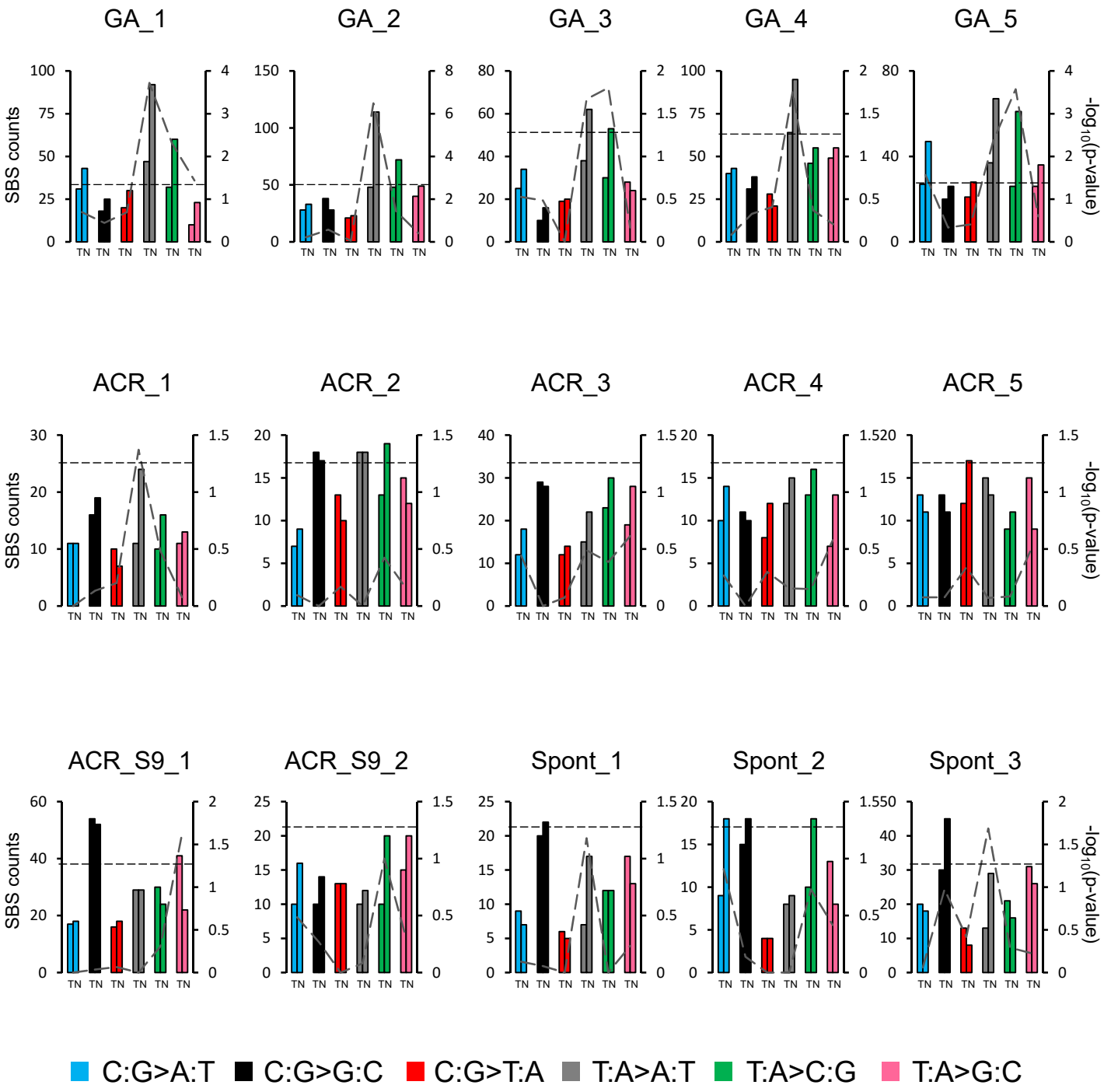


Suppl. Fig. S3

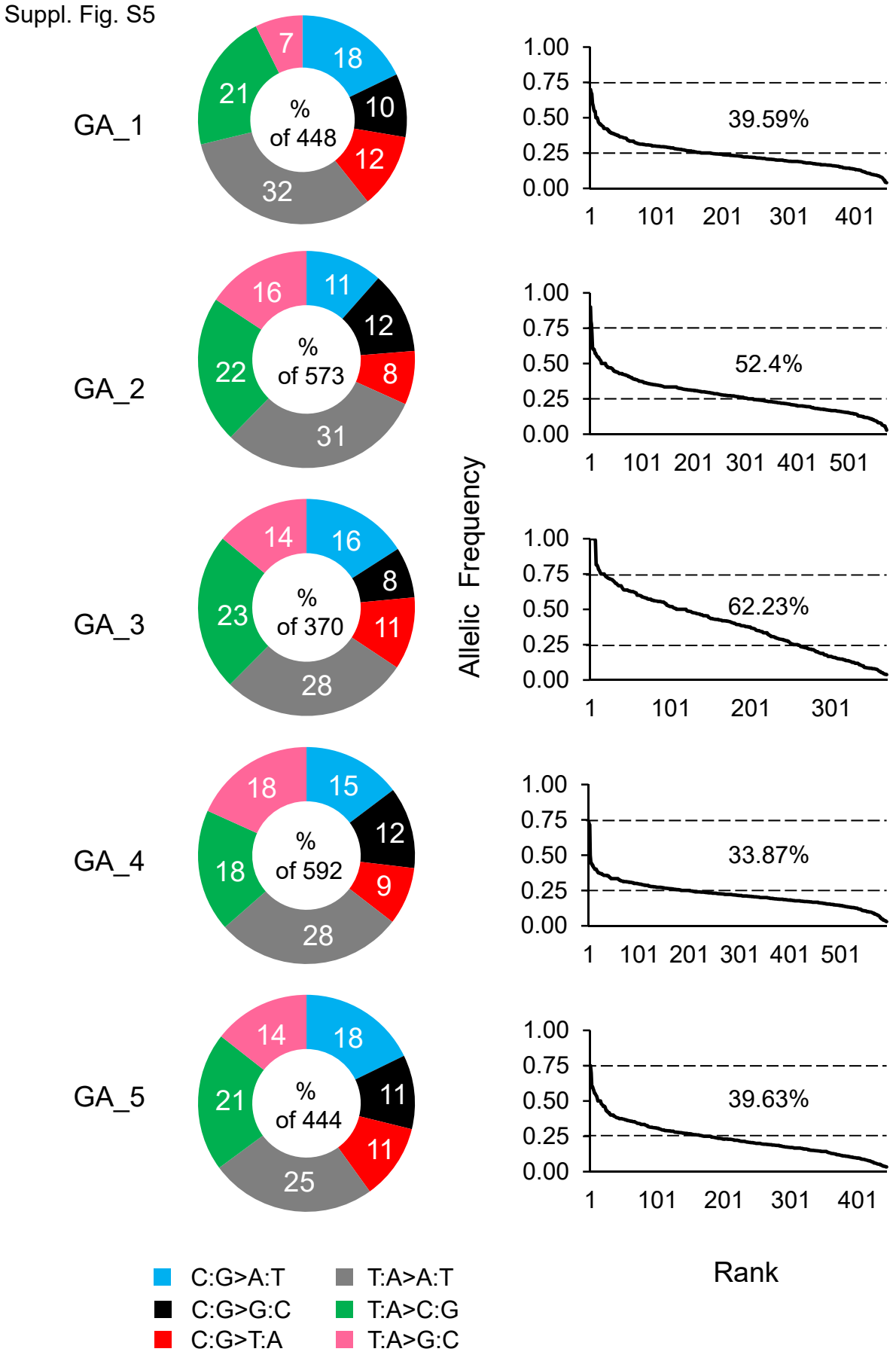
C



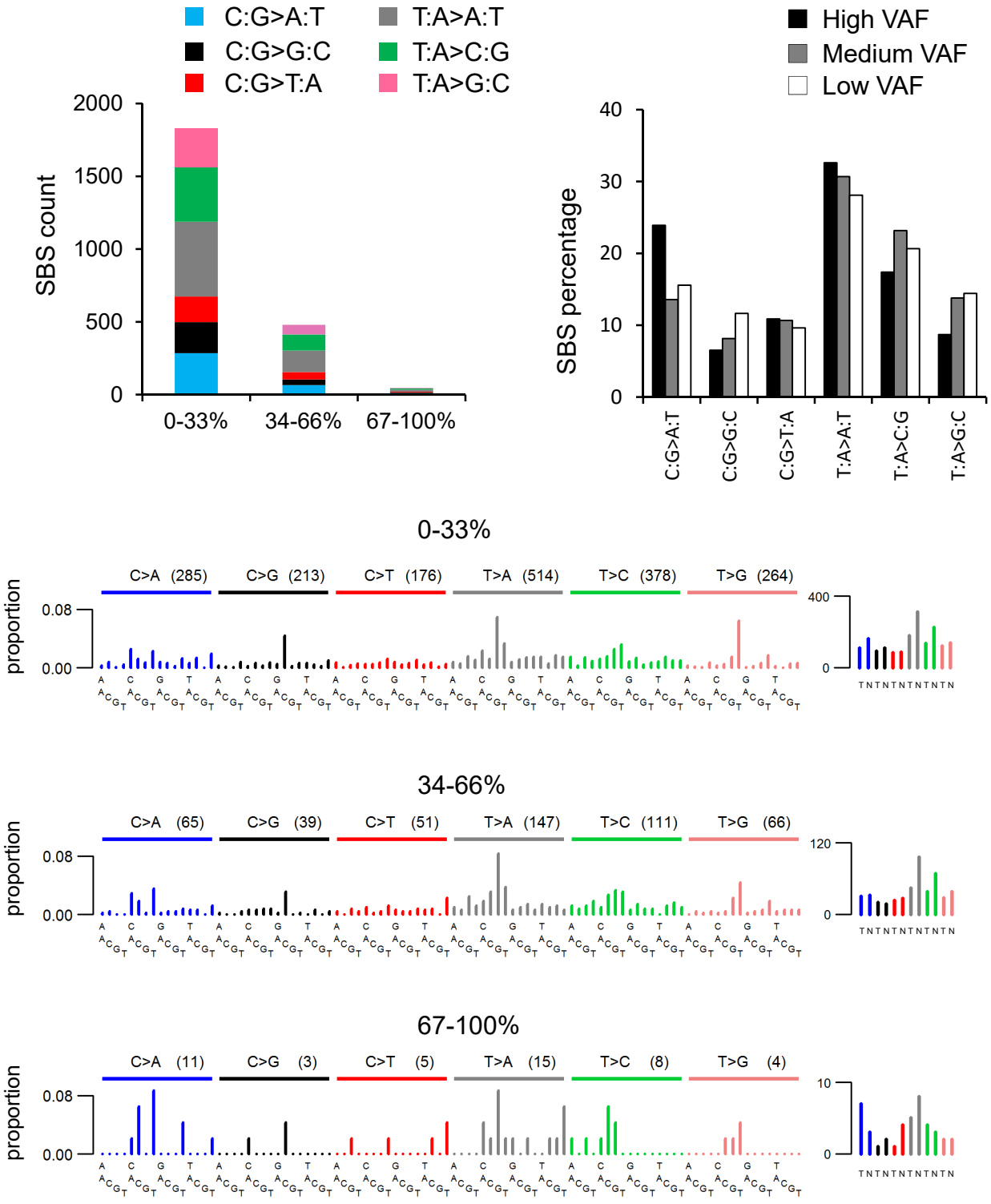
Suppl. Fig. S4


GA_1
GA_2
GA_3
GA_4
GA_5
ACR_1
ACR_2
ACR_3
ACR_4
ACR_5
ACR_S9_1
ACR_S9_2
Spont_1
Spont_2
Spont_3
SBS counts
-log10(p-value)
TN
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C

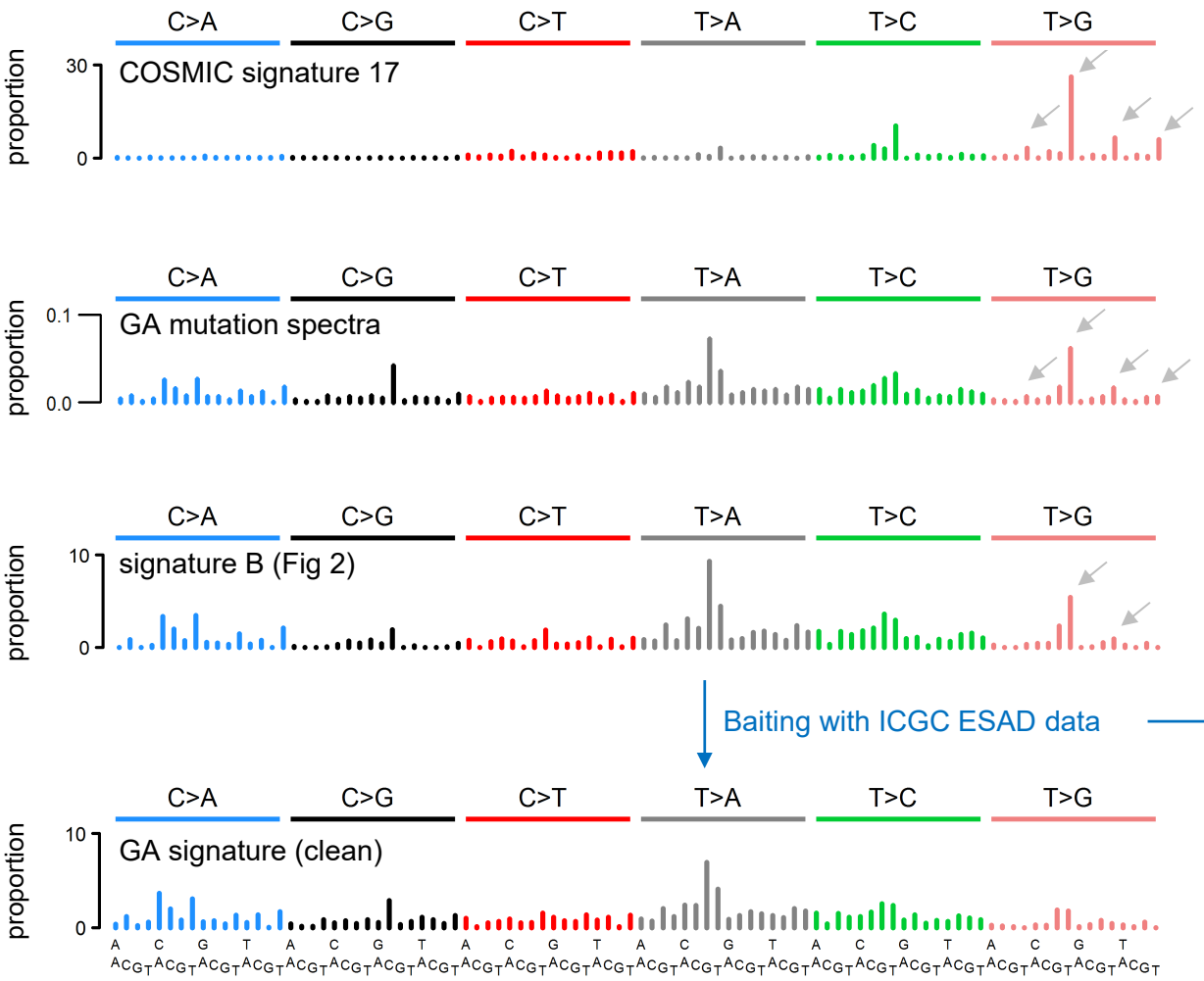
Suppl. Fig. S5


GA_1
GA_2
GA_3
GA_4
GA_5
% of 448
% of 573
% of 370
% of 592
% of 444
18
10
12
32
21
7
11
12
8
31
22
16
16
8
11
28
23
14
15
12
9
28
18
18
18
11
11
25
21
14
39.59%
52.4%
62.23%
33.87%
39.63%
1.00
0.75
0.50
0.25
0.00
1
101
201
301
401
501
Allelic Frequency
Rank
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C

Suppl. Fig. S6

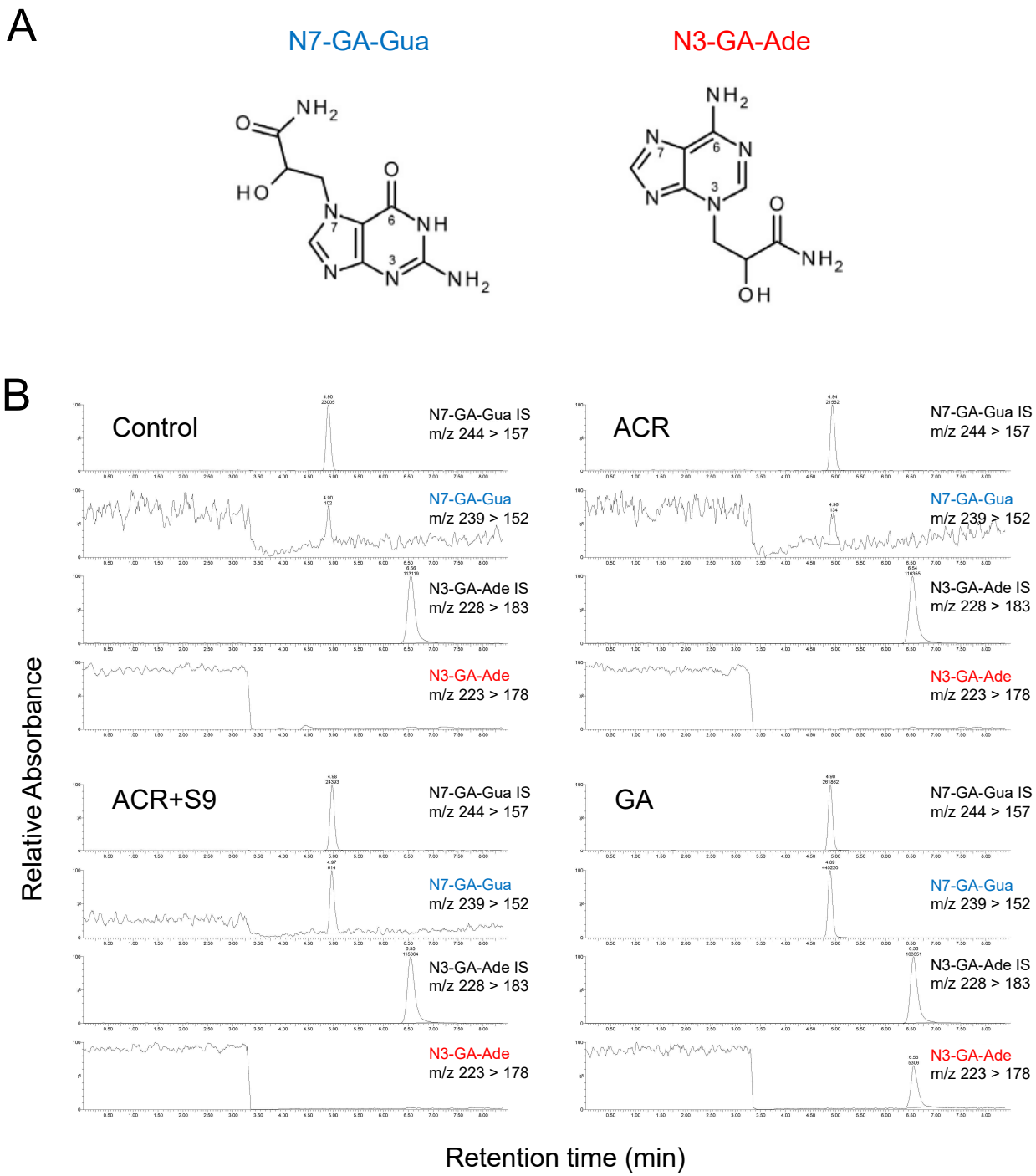

C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C
SBS count
2000
1500
1000
500
0
0-33%
34-66%
67-100%
High VAF
Medium VAF
Low VAF
SBS percentage
40
30
20
10
0
C:G>A:T
C:G>G:C
C:G>T:A
T:A>A:T
T:A>C:G
T:A>G:C
0-33%
C>A (285)
C>G (213)
C>T (176)
T>A (514)
T>C (378)
T>G (264)
proportion
0.08
0.00
400
0
TNTNTNTNTNTN
34-66%
C>A (65)
C>G (39)
C>T (51)
T>A (147)
T>C (111)
T>G (66)
proportion
0.08
0.00
120
0
TNTNTNTNTNTN
67-100%
C>A (11)
C>G (3)
C>T (5)
T>A (15)
T>C (8)
T>G (4)
proportion
0.08
0.00
10
0
TNTNTNTNTNTN

Suppl. Fig. S7

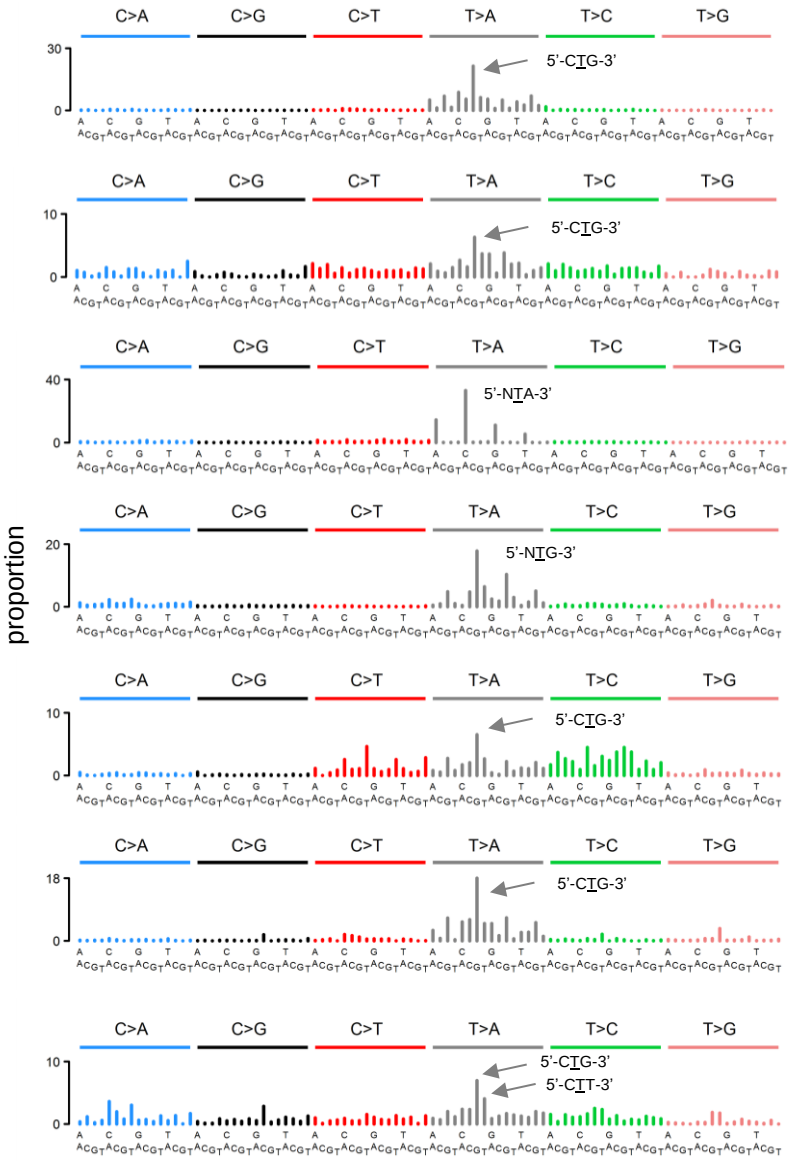


Sig. 17 peaks	Decrease (%)
C[T>A]C	0.0
C[T>A]G	25.7
C[T>A]T	8.0
C[T>C]C	22.4
C[T>C]G	29.7
C[T>C]T	20.0
A[T>G]T	100.0
C[T>G]C	42.9
C[T>G]G	19.1
C[T>G]T	68.2
G[T>G]T	52.7
T[T>G]T	0.0

Suppl. Fig. S8

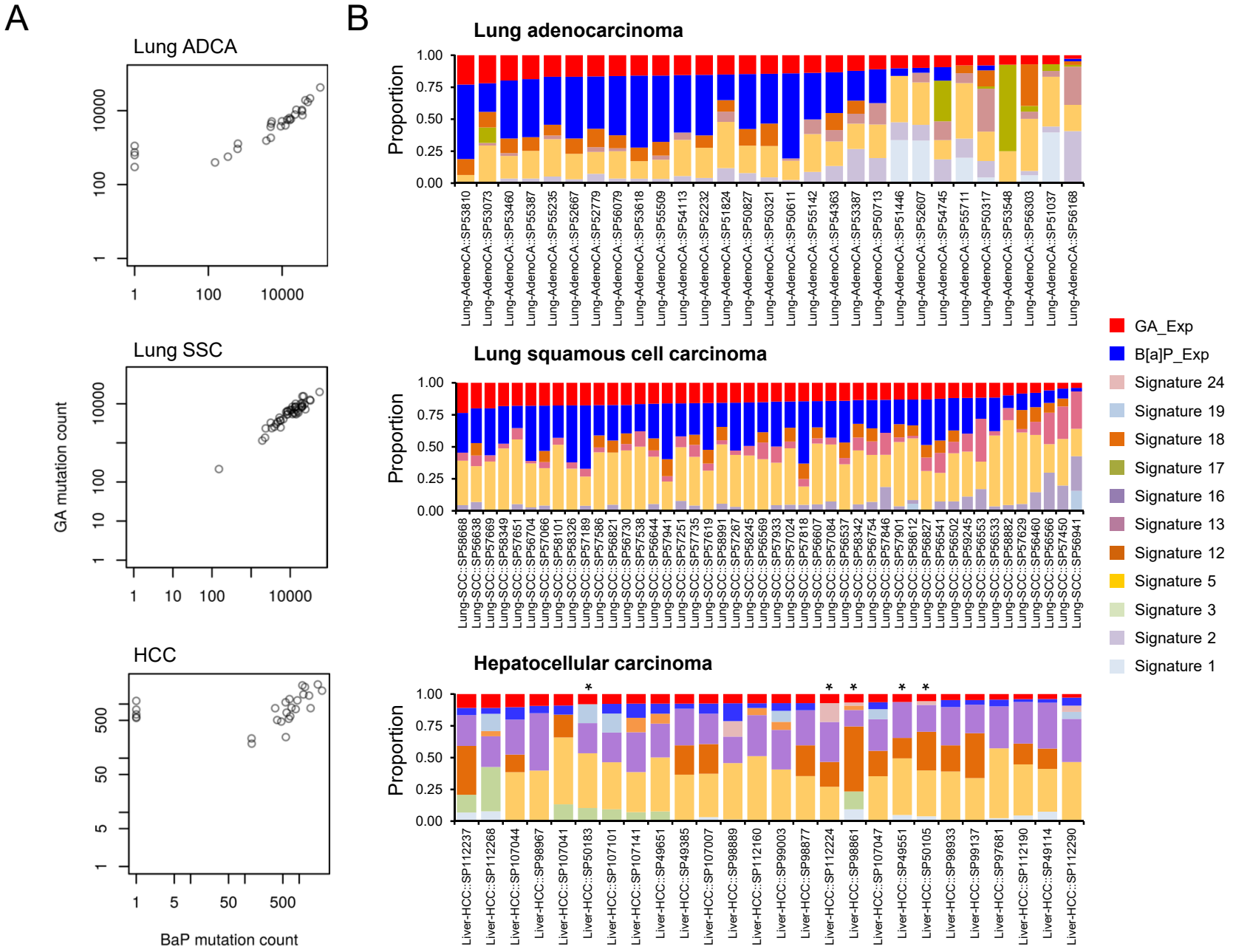

A
N7-GA-Gua
N3-GA-Ade
B
Control
ACR
ACR+S9
GA
N7-GA-Gua IS
m/z 244 > 157
N7-GA-Gua
m/z 239 > 152
N3-GA-Ade IS
m/z 228 > 183
N3-GA-Ade
m/z 223 > 178
Relative Absorbance
Retention time (min)

Suppl. Fig. S9



Exposure	Signature	Sample origin
AA [ref. 2]	sig 22	Urothelial carcinoma
Unknown [ref. 2]	sig 25	Hodgkin's lymphomas
Unknown [ref. 2]	sig 27	Kidney clear cell carcinomas
DMBA [ref. 48,49]	-	Skin cancer (mouse)
Urethane [ref. 50]	-	Lung cancer (mouse)
AA [ref. 37]	-	Hupki MEF
Glycidamide	-	Hupki MEF

Supplementary Figure S10



Supplementary Materials and Methods

DNA adduct analysis

The DNA was isolated from the cells using standard digestion with proteinase K, followed by phenol-chloroform extraction and ethanol precipitation. The DNA was subsequently treated with RNase A and T1, extracted with phenol-chloroform, and reprecipitated with ethanol. N7 GA-Gua and N3 GA-Ade were released by neutral thermal hydrolysis for 15 minutes, using Eppendorf Thermomixer R (Eppendorf North America) set to 99 °C. The samples were filtered through Amicon 3K molecular weight cutoff filters (Merck Millipore) to separate the adducts from the intact DNA.

TP53 genotyping

The following are the *TP53* primers used for amplicon sequencing of mutations accumulated in human TP53 of the Hupki MEFs. The sequences are presented in 5' to 3' orientation: Exon 4: fwd – TGCTCTTTTCACCCATCTAC, rev – ATACGGCCAGGCATTGAAGT; Exons 5-6: fwd – TGTTCACTTGTGCCCTGACT, rev – TTAACCCCTCCTCCCAGAGA; Exon 7: fwd – CTTGCCACAGGTCTCCCC, rev – CACTTGCCACCCTGCACA; Exon 8: fwd – TCCTTACTGCCTCTTGCTTCTCTT; rev – CCAAGGGTGCAGTTATGCCT. Sequences and their alterations were analyzed using the CodonCode Aligner software.

Processing of WES data

Prior to variant calling, recalibrated .bam files were interrogated for imbalanced base mismatch distribution between Read 1 and Read 2 sequences. We used the DNA damage estimator tool (as per (1); (https://github.com/Ettwiller/Damage-estimator)) to measure the Global Imbalance Value (GIV) score and to exclude sequencing-related DNA damage and artefacts due to oxidative damage that can confound the determination of treatment-specific variants. The MutSpec suite included tools for annotation of the vcf files with Annovar and variant filtering to remove dbSNP142 contents, segmental duplicates, repeats, and tandem repeat regions. Finally, to maximize the chance of robust variant calls and to exclude potential unfiltered single nucleotide polymorphisms (SNP), we considered only variants unique to each sample.

Bioinformatics and statistical analyses

The following are the International Cancer Genome Consortium (ICGC) esophageal carcinoma patient data (2,3) that were used in the step of cleaning the experimental signature from the COSMIC signature 17 signal: ESAD-UK-SP119768.hg19; ESAD-UK-SP191660.hg19; ESAD-UK-SP111113.hg19; ESAD-UK-SP111173.hg19; ESAD-UK-SP192267.hg19; ESAD-UK-SP111026.hg19; ESAD-UK-SP192494.hg19; ESAD-UK-SP111019.hg19; ESAD-UK-SP111058.hg19.

References

1. Chen, L.*, et al.* (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752-756.
2. Secrier, M.*, et al.* (2016) Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature Genetics*, **48**, 1131-1141.
3. Cancer Genome Atlas Research Network (2017) Integrated genomic characterization of oesophageal carcinoma. *Nature*, **541**, 169-175.