Supplemental Methods

"Experimental and pan-cancer genome analyses reveal widespread contribution of acrylamide exposure to carcinogenesis in humans"

Zhivagui M. *et al*, 2019

TP53 genotyping of the Hupki MEF clones

The following primer pairs were used for Sanger sequencing of the human *TP53* knock-in construct in MEF cells, targeting exons 4 through 8. The sequences are listed in the 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.

Library preparation and whole-exome sequencing (WES)

Library preparation was carried out using the KAPA HyperPlus library preparation kit (KAPA Biosystems) according the manufacturer's instructions. Exome capture was performed using the SureSelect XT Mouse All Exon Kit (Agilent Technologies). Eighteen exome-captured libraries were sequenced in the paired-end 150 base-pair run mode using the Illumina HiSeq 4000 sequencer.

Processing of WES data

FASTQ files were analyzed for data amount and quality using FastQC (0.11.3) and were processed as follows: reads were trimmed with Trim Galore (v0.4.2,

https://github.com/FelixKrueger/TrimGalore), then mapped on the mouse mm10 genome (release GRCm38) using the BWA (v0.7.16) aligner (Li and Durbin 2010). Duplicate reads were flagged by Picard (v2.7.1) (http://broadinstitute.github.io/picard/), and the aligned reads further underwent base quality score recalibration and indel realignment with the corresponding tools from GATK v3.6. (McKenna et al. 2010) These components of the pipeline used are publicly available at https://github.com/IARCbioinfo/alignment-nf. The resulting alignment files had a mean depth-of-coverage of 135 and 175 for acrylamide and glycidamide samples, respectively.

Variant calling

Two somatic variant callers were employed with default parameters in order to detect single base substitutions (SBS) and indels (MuTect 1.1.6-4 (Cibulskis et al. 2013) and Strelka 1.015 (Saunders et al. 2012)) in exposed clones, using primary cells as normal samples. Each immortalized clone was compared to primary MEFs from three different embryos (conditions Prim_1, Prim_2, and Prim_3). The overlap of the variant calling outcome with respect to the different primary MEFs showed ~80% concordance (Supplemental Fig. S4) with MuTect exhibiting more stringent calling performance. Mutation data obtained from the MuTect variant caller were processed with the MutSpec suite (Ardin et al. 2016), downloadable from https://github.com/IARCbioinfo/mutspec) for annotation with ANNOVAR (Wang et al. 2010) and variant filtering to remove dbSNP142 contents, segmental duplicates, repeats and tandem repeat regions. To maximize the chance of robust variant calls and to exclude potential single nucleotide polymorphisms (SNP), we considered only variants unique to each sample. To estimate the extent of sequencing-related damage in the sample data, we determined the global imbalance value (GIV) score of each sample (Chen et al. 2017). No detectable contribution for any of the damage-associated mutation types was observed in our dataset. For more details, see the summary of sequencing metrics (Supplemental Table S1), the list of identified MuTect

SBS variants (Supplemental Table S2) and indels identified by Strelka (Supplemental Table S3).

Bioinformatics and statistical analyses of experimental signatures

The FactoMineR package (Lê 2008) in R (version 3.3.2.) (Team 2016) was used to perform the principal component analysis (PCA). We used the DNA damage estimator tool (Chen et al. 2017) to measure the Global Imbalance Value (GIV) score and to exclude sequencing-related DNA damage and artefacts that can confound the determination of treatment-specific variants. To perform the transcription strand bias (TSB) analyses, *p*-values were calculated using Pearson's $\chi 2$ test. As multiple comparisons were assessed, the *p*-value was adjusted by applying a false discovery rate (FDR). Statistical analyses were carried out using the *stats* package.in R version 3.3.2.

To estimate the number of signatures to extract, we applied factorization rank analysis (Brunet et al. 2004) within MutSpec suite (Ardin et al. 2016) which yielded the cophenetic correlation coefficient indicating that a minimum of 2 signatures is recommended although this resulted in insufficient separation of the background signal (signature 17 and a prominent G[C>G]G peak, artefacts of cell culture). Selecting three signatures separated well the GA-treated group from the ACR and control samples and the latter two groups are in effect further separated based on the higher or lower presence of background mutation patterns. A setting of 4 signatures led to major splitting of the exposure-related (the GA-specific C>A, T>A, T>C) as well as the background signals (C>G and T>G)(Supplemental Fig. S11).

Extended input NMF to obtain refined experimental GA and B[a]P signatures

The following ICGC esophageal carcinoma patient data (Secrier et al. 2016) were used for the extended input NMF in order to clean further the residual COSMIC signature 17 from the GA mutational signature obtained from the 5 GA-treated MEF clones: 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). The TCGA data used for the GA-signature clean-up from signature-17 were as follows: 15 TCGA ESCA positive for signature 17: TCGA-2H-A9GI, TCGA-2H-A9GK, TCGA-2H-A9GL, TCGA-2H-A9GR, TCGA-IC-A6RE, TCGA-IG-A4QS, TCGA-L5-A4OJ, TCGA-L5-A4OT, TCGA-L5-A4OU, TCGA-L5-A4OW, TCGA-L5-A88Y, TCGA-L5-A8NE, TCGA-L5-A8NJ, TCGA-L5-A8NR, TCGA-RE-A7BO; 15 TCGA ESCA negative for signature 17: TCGA-2H-A9GJ, TCGA-JY-A93C, TCGA-JY-A93E, TCGA-L5-A4OE, TCGA-L5-A4OH, TCGA-L5-A4OI, TCGA-L5-A4OO, TCGA-L5-A88V, TCGA-L5-A891, TCGA-L5-A8NM, TCGA-L5-A8NN, TCGA-L5-A8NS, TCGA-L5-A8NV, TCGA-V5-A7RE, , TCGA-V5-AASW; TCGA STAD positive for signature 17: TCGA-BR-4357, TCGA-BR-6458, TCGA-BR-8297, TCGA-BR-8485, TCGA-BR-8589, TCGA-CD-5801, TCGA-CD-8535, TCGA-CD-A48A, TCGA-CG-4469, TCGA-D7-A4Z0, TCGA-HF-7134, TCGA-HU-A4GH, TCGA-VQ-A8P3, TCGA-VQ-A91S, TCGA-VQ-A91V ; TCGA STAD negative for signature 17: TCGA-BR-7707, TCGA-BR-7851, TCGA-BR-8361, TCGA-BR-8591, TCGA-BR-A4QL, TCGA-CG-5723, TCGA-CG-5728, TCGA-D7-A6EY, TCGA-HF-A5NB, TCGA-HU-8602, TCGA-HU-A4G8, TCGA-HU-A4GQ, TCGA-HU-A4GU, TCGA-MX-A5UJ, TCGA-VQ-A8PP.

The experimental B[a]P mutational signature was extracted from the mutation spectra obtained from the HMEC post-stasis clones using the NMF algorithm with an extended input including SBS sets from the TCGA lung cancer collection. The TCGA samples used were as follows: 15 Lung.AdenoCA positive (>50% presence) for tobacco-smoking SBS4: TCGA-05-4382, TCGA-05-4410, TCGA-44-7670, TCGA-49-AARE, TCGA-55-7907, TCGA-55-7994, TCGA-55-A490TCGA-64-5781, TCGA-69-7979, TCGA-78-7155, TCGA-78-8662, TCGA-86-8073, TCGA-95-7039, TCGA-L9-A7SV,TCGA-NJ-A4YQ; 15 Lung.AdenoCA negative for SBS4: TCGA-17-Z026, TCGA-49-6743, TCGA-49-AARN, TCGA-50-5946, TCGA-53-7626, TCGA-55-6979,

TCGA-55-8092, TCGA-55-8506, TCGA-55-A4DF, TCGA-69-7978, TCGA-78-7536, TCGA-86-A4JF, TCGA-93-8067, TCGA-95-7567, TCGA-NJ-A4YP; 15 Lung.SCC positive (>50%) for SBS4: TCGA-33-4566, TCGA-33-4583, TCGA-34-2604, TCGA-37-3792, TCGA-37-4130, TCGA-37-5819, TCGA-39-5011, TCGA-39-5034, TCGA-43-2576, TCGA-43-2581, TCGA-43-5668, TCGA-46-3769, TCGA-60-2714, TCGA-90-7769, TCGA-O2-A52Q; and Lung.SCC negative for SBS4: TCGA-18-3416, TCGA-18-3421, TCGA-21-1080, TCGA-22-5473, TCGA-37-3789, TCGA-39-5031, TCGA-63-A5MM, TCGA-66-2773, TCGA-66-2785, TCGA-77-7141, TCGA-85-6561, TCGA-85-A4CL, TCGA-94-8490, TCGA-98-8021, TCGA-NC-A5HG).

REFERENCES

Ardin M, Cahais V, Castells X, Bouaoun L, Byrnes G, Herceg Z, Zavadil J, Olivier M. 2016. MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17**: 170.

Brunet J-P, Tamayo P, Golub TR, Mesirov JP. 2004. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* **101**: 4164-4169.

Chen L, Liu P, Evans TC, Ettwiller LM. 2017. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* **355**: 752-756.

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**: 213-219.

Lê S, Josse, J. & Husson, F. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **25**: 1-18.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. 2012. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**: 1811-1817.

Secrier M, Li X, de Silva N, Eldridge MD, Contino G, Bornschein J, MacRae S, Grehan N, O'Donovan M, Miremadi A, et al. 2016. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature Genetics* **48**: 1131-1141.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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