Supplemental Methods for

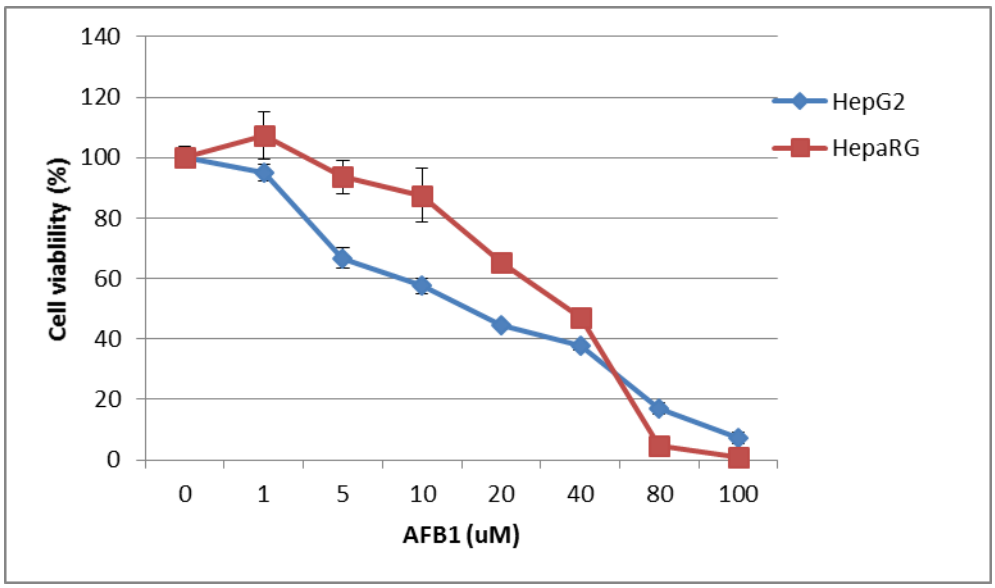
Genome-scale mutational signatures of aflatoxin in cells, mice and human tumors

Huang, Yu, et al., 2017

Human cell lines

We assessed AFB1 mutations in two human cell lines. HepaRG (Invitrogen) cells were derived from an HCC and can be differentiated to cells that resemble adult hepatocytes; we cultured and differentiated HepaRG cells as described previously (Gripon et al. 2002). HepG2 was derived from a hepatoblastoma (Ostlund et al. 1996); we cultured it in minimum essential medium supplemented with 10% fetal calf serum, nonessential amino acids, 100 units/ml penicillin, and 100 ug/ml streptomycin.

We determined the AFB1 (A 6636, Sigma) half maximal inhibitory concentration (IC50) in these cell lines as shown.



Inhibitory concentrations of AFB1 in HepaRG and HepG2 cells. AFB1 cytotoxicity was evaluated by an ATP assay (Cell Titer-Glo, Promega) on cells treated with AFB1 for 72 h. Cell viability corresponds to the ratio between AFB1- and DMSO-treated cells

To generate clones we treated HepaRG at 30 uM and HepG2 at 20 uM every 3 days for 45 days. We then separately sequenced three clones each of treated HepaRG and HepG2 cells.

Mouse models

Details of the mice, their treatment, and similarity of the mouse liver tumors to human HCCs are provided in (Teoh et al. 2015). Briefly, we studied tumors from 6 C57BL/6J mice, 3 of which bore an hepatitis B surface antigen (HBsAg) transgene (Chisari et al. 1985). The mice received 1 peritoneal injection of 19 nmol/g AFB1 at day 7 after birth and were sacrificed at 15 months. Animal experiments were approved by and performed in accordance with the guidelines of the SingHealth Animal Care and Use Committee.

Whole-genome sequencing of human cell lines and mouse tumors

For DNA extraction from frozen mouse tissue, we placed 15 to 25mg of frozen tissue in a 2ml microcentrifuge tube containing ATL buffer (220 µl, Qiagen) and Proteinase K (20 µl, Qiagen). Tissue was homogenized in the TissueLyser II (speed:20-30Hz, Qiagen) for 1-2 minutes then placed into thermomixer (56 ^{o}C, shaking speed: 900rpm, Eppendorf) for 3 hours. After homogenization, 220 µL of supernatant was transferred to another sample tube containing RNase A (4 µl, Qiagen) and incubated at room temperature for 2 minutes before DNA extraction using QIAsymphony (Qiagen). For DNA extraction from cell lines, cells were trypsinized, washed with PBS and centrifuged to form a pellet. We then added a mixture of buffer ATL and Proteinase K and extracted DNA using QIAsymphony (Qiagen). For both tissue and cell-line DNA, quality was assessed by an Agilent 2200 T before sequencing. Supplemental Table S13 provides details on the sequencing runs.

Whole-exome sequencing of human FFPE samples

Genomic DNA was extracted from macro-dissected paraffin sections, and 250 nanograms of each DNA sample were sheared by Covaris (Covaris, Inc.) to ~300 bp fragments, as previously described (Castells et al. 2015). We prepared libraries with the Kapa LTP Library Preparation Kit (Kapa Biosystems) according to the manufacturer's recommendations. Four libraries (250 ng each) were pooled per capture with the Nimblegen Roche SeqCap EZ Exome v3 reagent (Roche), and the exome-enriched mix was PCR-amplified in 10 cycles. The post-enrichment material was diluted in 420 mL water to a final concentration of 6 pmol/L and sequenced as described in Supplemental Table S12.

Alignment and variant calling

For whole genome sequencing (WGS) we used BWA-MEM (v0.7.12) (Li and Durbin 2009) with the –R and –M options to align reads to hs37d5 (human) or mm10 (mouse), followed by sorting, PCR duplicate removal and merging using Sambamba (v0.5.8) (Tarasov et al. 2015). Strelka (v1.014) (Saunders et al. 2012) called somatic variants, with non-default parameters as follows: ssnvNoise=0.00000005, minTierMapq=15, ssnvQuality_LowerBound=25, sindelQuality_LowerBound=20 and isWriteRealignedBam=1; any variants with variant depth / total depth < 0.1 were removed. Within each cell line (HepaRG and HepG2) we used Strelka to call "somatic" variants in each clone, *c*, against the other two clones, and considered the intersection of the "somatic" variants as the mutations specific to *c*. For tumors from mice M1, M2, and M3, which were treated with AFB1 only, Strelka called variants from reads from the tumor and matched non-malignant tissue. For mice M4, M5, and M6, which carried the HBsAg transgene and were treated with AFB1, we used Strelka to call "somatic" variants in each tumor, *t*, against the other two tumors, and considered the intersection of the "somatic" variants as the somatic mutations in *t*.

For WES data, BWA-MEM (v0.7.9a) with default parameters aligned reads to hs37d5, followed by SAMtools (v 0.1.8 r613) (Li et al. 2009) to sort and reads and remove duplicates. GATK (v2.2-25-g2a68eab) was used for local realignment around indels and base quality recalibration (McKenna et al. 2010). Picard tools corrected mate pair information, and MuTect (v 1.1.4) (Cibulskis et al. 2013) with default parameters called variants on tumor and matched normal pairs. We restricted analysis to mutations in the capture target.

Additional filtering of mutations

For human samples, variants in dbSNPv132, The 1000 Genomes (The 1000 Genomes Project Consortium 2015), segmental duplications, microsatellites and homopolymers, and the GL

and decoy sequences were excluded. For mouse tumors, variants in dbSNPv142, regions of segmental duplication, or regions of tandem repeats were excluded. For human and mouse WGS data, we also excluded candidate somatic variants at sites where ≥ 2 normal samples each contained ≥ 2 reads with a variant base.

Principal components analysis

We used the R function "prcomp" for principal components analysis (PCA) (R Development Core Team 2017). Trinucleotide frequencies in the human genome, human exome and mouse genome are different. Therefore, to compare mutation spectra from these three sources, the spectra of human exomes and mouse genomes were normalized to human genome frequencies, as is a common practice in this field.

Analysis of variant allele frequencies in the context of tumor purity and aneuploidy

Tumor purity (proportion of malignant cells in the tumor), ploidy estimation, and copy number and depth ratios across each mouse tumor genome were calculated and visualized using Sequenza (Favero et al. 2015). In brief, we created Sequenza input files by applying bam2seqz to the the tumor and normal BAMs file along with information on the, reference genome. We generated binned input with seqz-binning with -w 500, and then it loaded into R using sequenza.extract() function from the R sequenza library. Tumor content, ploidy, and copy number and depth ratios were calculated using sequenza.fit(). We plotted results with sequenza.results().

References

Castells X, Karanovic S, Ardin M, Tomic K, Xylinas E, Durand G, Villar S, Forey N, Le Calvez-Kelm F, Voegele C et al. 2015. Low-Coverage Exome Sequencing Screen in Formalin-Fixed Paraffin-Embedded Tumors Reveals Evidence of Exposure to Carcinogenic Aristolochic Acid. *Cancer Epidemiology, Biomarkers & Prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **24**: 1873-1881.

Chisari F, Pinkert C, Milich D, Filippi P, McLachlan A, Palmiter R, Brinster R. 1985. A transgenic mouse model of the chronic hepatitis B surface antigen carrier state. *Science* **230**: 1157-1160.

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. *Nature Biotechnology* **31**: 213-219.

Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, Szallasi Z, Eklund AC. 2015. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* **26**: 64-70.

Gripon P, Rumin S, Urban S, Le Seyec J, Glaise D, Cannie I, Guyomard C, Lucas J, Trepo C, Guguen-Guillouzo C. 2002. Infection of a human hepatoma cell line by hepatitis B virus. *Proc Natl Acad Sci U S A* **99**: 15655-15660.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

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.

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.

Ostlund RE, Jr., Seemayer R, Gupta S, Kimmel R, Ostlund EL, Sherman WR. 1996. A stereospecific myo-inositol/D-chiro-inositol transporter in HepG2 liver cells. Identification with D-chiro-[3-3H]inositol. *J Biol Chem* **271**: 10073-10078.

R Development Core Team. 2017. *R: A Language and Environment for Statistical Computing*, http://www.r-project.org/. R Foundation for Statistical Computing, Vienna, Austria.

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.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**: 2032-2034.

Teoh WW, Xie M, Vijayaraghavan A, Yaligar J, Tong WM, Goh LK, Sabapathy K. 2015. Molecular characterization of hepatocarcinogenesis using mouse models. *Disease Models & Mechanisms*: **8**: 743-753.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.