# Supporting Information

## Ariffin et al. 10.1073/pnas.1417322111

### SI Materials and Methods

***TP53* Database and Anticipation Data.** The IARC *TP53* database compiles standardized information on all subjects with germ-line *TP53* mutation described in the public literature. This information includes *TP53* mutation identification, tumor topography and morphology, sex, age at diagnosis, number of generations analyzed in the pedigree, and position in the pedigree of each subject with documented cancer status. We used the R16 version of the database, updated in November 2012 (p53.iarc.fr/DownloadDataset.aspx), which contains data on 2,450 subjects with the germ-line *TP53* mutation. Data were filtered to select individuals who (*i*) are part of families matching LFS or LFL criteria or have a defined familial history (FH) of cancer; (*ii*) belong to pedigrees with one to four documented generations (families with five generations were excluded because of either low average age of individuals in the fifth generation or incomplete ascertainment of tumor diagnosis in the first generation); (*iii*) have their generational position in the pedigree clearly identified; and (*iv*) do not carry the "Brazilian founder" *TP53 R337H* mutation, a mutation of partial penetrance, which has been associated with extremely variable familial and nonfamilial tumor patterns. This search identified 1,771 individual records from 294 pedigrees. Families were subgrouped in four classes according to the number of generations documented (one to four generations). The mean age at first cancer onset (±SD) and proportion of diagnoses corresponding to childhood or adult forms of common LFS/LFL cancers in each generation were calculated separately for families with one, two, three, or four generations. To detect a possible shift in the severity of childhood cancer phenotype, the age-dependent accrual of cancer up to age 20 y was plotted for each generation according to pedigree structure. Data were analyzed using the standard *t* test (for age at first cancer onset) or $\chi^2$ test (for distribution of tumor types) using tools available at vassarstats.net/, without further adjustment for age, sex, or region of origin of selected patients and families.

**Patients and Family.** Whole-genome sequencing was performed on 13 members of a family of Malay origin matching strict LFS criteria (KA family). Briefly, the proband, an 8-y-old girl, first presented at the age of 8 mo with embryonal rhabdomyosarcoma of the trunk, which was treated by surgery and chemotherapy. Seven years later, she developed an intracranial lesion (left temporoparietal mass), which was histologically proven to be a recurrence of rhabdomyosarcoma, and died at age 9. Among her 5 siblings, a younger sister developed adrenal cortical carcinoma at 6 mo of age, which was successfully treated. Their mother was diagnosed with carcinoma of the right breast at age 26, and of left breast at age 28, for which she is currently in remission after surgery and chemotherapy. Her sister (aunt of the proband) was diagnosed with an osteosarcoma of the jaw at age 26 and is deceased. The maternal grandmother of the proband was deceased from breast cancer at age 38.

*TP53* mutation was detected using primers and protocols recommended by the International Agency for Research on Cancer (p53.iarc.fr/Download/TP53_DirectSequencing_IARC.pdf). A 6-bp insertion mutation at the second base of codon 334 (nucleotide 17579) in exon 10, causing in-frame insertion of residues repeating residues 334 (Gly) and 335 (Arg) was identified in the following family members: the proband, her affected sister, two younger siblings who have yet to manifest a malignancy, the affected aunt and her two unaffected children, and an unaffected uncle (brother of the mother). Two siblings of the proband and the father were confirmed noncarriers (Fig. 1). The mutation is predicted to be deleterious and to preclude p53 protein function by disrupting the oligomerization domain and preventing the formation of high-affinity DNA binding protein complexes.

DNA from five members of a second family (MM family) was analyzed by whole-exome sequencing (WES). This pedigree is shown in Fig. 2. The proband carried a *TP53* mutation at codon 245 (p.G245S) (mutation details) and developed rhabdomyosarcoma at age 2 y. A sister (unconfirmed mutation carrier) died of medulloblastoma at age 2.5 y. Two other siblings who are carriers of the p.G245S mutation (age 8 and 12 y) are yet to manifest a malignancy. The father (mutation carrier) has not developed any cancer by the age of 40. The mother is an ascertained noncarrier.

**Whole-Genome and -Exome Sequencing.** From the onset of the study (identification of the proband), the family was offered genetic counseling, and parents were asked to provide signed informed consent for themselves and for their children. For WGS of 13 subjects from this LFS kindred, a specific consent was provided by the father and the mother of the proband (who are acting as caretakers for the children of their deceased sister). For participants who are currently alive, DNA was extracted from peripheral white blood cells obtained by venipuncture using the Qiagen DNA Extraction kit according to the manufacturer's instructions (QIAamp DNA blood Maxi kit; Qiagen). For deceased patients, archival DNA (preserved at −80 °C) was used. WGS and WES were performed by the Beijing Genome Institute (BGI) (www.genomics.cn/en/index). WGS was performed using ABI SOLiD 4.0 with paired-end 50-base pair reads. Primary data analysis including image analysis, base calling, alignment, and variant calling including CNV were performed by BGI using Bioscope software. Exome sequencing of the MM family was performed by BGI using DNA fragments of 150–200 bp in length and enriched for exome with the Agilent SureSelect in Solution. Captured library was then sequenced on the HisEq. 2000 platform with sequences generated as 90-bp paired-end reads. Basic bioinformatics analysis was performed by BGI. Briefly, adapter and low-quality reads were filtered to generate clean data. Burrows–Wheeler Aligner (BWA) was used to align reads to the reference sequence. SOAPsnp was used to detect SNPs, and SAMtools and the Genome Analysis Toolkit (GATK) were used to detect SNVs and insertions/deletions (indels).

**CNV Calling of WGS Data.** CNVs were inferred using depth-of-read coverage, and CNValidator (code.google.com/p/cnvalidator) was used to identify high-confidence CNVs using SNP information. CNVs were initially called using the Find Human CNV tool in Bioscope. Briefly, the algorithm divides the chromosome arms into windows of size 5 kb. The read-depth coverage was computed and corrected for GC content bias. Global normalization was used to compute log ratios using the median of average means of all chromosome arms as the expected value. A hidden Markov model was used to call the copy-number state of the windows, and contiguous windows of the same state were merged into CNV segments with *P* value significance. Only CNVs exceeding 10 kb in size were called because they must span at least two windows to be detected. A smaller *P* value indicates a greater probability of a predicted CNV to be true. Only CNVs with *P* value less than 0.05 were retained for further consideration. Chromosomes X and Y were excluded from the analysis.

In addition, the CNValidator software (code.google.com/p/cnvalidator) was used to filter out low-quality CNVs using SNP information, applying $P < 0.01$ as threshold. Briefly, rigorous statistical methods were used to assign *P* values to CNVs with null model consisting of regions of the genome after removing all purported CNVs as well as repeat regions including telomeric and pericentric regions. Deletions were filtered using homozygosity and, because the null model has very uniform distribution of heterozygous SNPs averaged over segments greater than 10 kb, the Poisson model was used to determine *P* value. Duplications were filtered by median of ratio of heterozygous SNP reads, and bootstrapping of the null model was used to determine *P* value.

**Detection of de Novo CNV from WGS.** For the children KA:III:1 to III:6, CNVs could be identified as inherited or de novo through pairwise comparison with the CNV results for their parents, II:1 and II:2 (Table S2). Any CNV from a child that overlaps with a CNV from either parent was considered inherited. Almost all CNVs that passed CNValidator ($P < 0.01$) were inherited. At most one de novo CNV was found in any child. This number is consistent with a low false-positive rate for CNV detection because de novo CNVs are rare per birth and inherited CNVs are detected independently in at least two members of the family. To further assess the significance of candidate de novo CNV detected, a lower *P* value was applied in filtering with CNValidator. No de novo CNVs passed CNValidator with $P < 0.0001$.

**Identification of SNV Segregating in LFS Family.** Single nucleotide variants (SNVs) were called using diBayes, the SNP calling tools of the Bioscope software. SNV calls were confirmed using the Genome Analysis Toolkit (GATK) (www.broadinstitute.org/gatk/). Briefly, duplicates were marked with Picard (picard.sourceforge.net/) and then analyzed with GATK to perform local realignment around indels and base quality score recalibration. SNVs were detected using the GATK Unified Genotyper. These variants were annotated and filtered using ANNOVAR (www.openbioinformatics.org/annovar/). The filters used include the following: annotate and identify exonic/splicing variants, remove synonymous variants, use conservation from 46-species alignment, remove variants in segmental duplication regions, and remove variants in 1000 Genome Projects and dbSNP. Nine missense mutations and one nonsense mutation were identified and confirmed by Sanger sequencing.

**Detection of de Novo SNV from WES.** MuTect (www.broadinstitute.org/cancer/cga/mutect) with default parameters was used to detect point mutations in offsprings not found in parents from whole-exome sequencing data. The parents' binary alignment/map (BAM) alignment files had duplicate reads removed, merged, and used as control. Because MuTect was originally developed for the identification of somatic point mutations in cancer genomes, we performed additional filtering of calls. The potential de novo mutations were ranked by the t_lod_fstar value, which represents the log of likelihood of a mutation being real to the likelihood of an event being a sequencing error. We inspected the top candidates for each offspring manually by looking at the alignment of reads at those positions in the genome across all family members with Integrative Genomics Viewer (IGV).

**CNV Detection from aCGH.** DNA from members of the KA family (II:1, II:2, III:1, III:4, and III:6) and the MM family (II:6, II:7, III:1, III:3, and III:4) were sent to Genotypic Technology for aCGH using the Agilent SurePrint G3 Human CGH Microarry Kit, 1 × 1 M. Each sample was hybridized along with Coriell Male or Female Reference DNA according to sex. Initial data analysis was done with Agilent Cytogenomics 2.7.70 software using Aberration Detection Algorithms (ADM-2 and Fuzzy Zero) with the minimum number of probes that should be present in an aberrant region taken to be 3, threshold score of 6.0 (default), and minimum average log ratio to be 0.25. Chromosomes X and Y were excluded from the analysis.

**CNV Accrual in WT and *Trp53* Knockout Mice.** *Trp53* WT mice that were 3 mo of age were mated to either *trp53* WT mice (at 3 mo, 9 mo, and 12 mo of age), or *trp53* heterozygous mice (at 3 mo, 9 mo, and 12 mo of age), or *trp53* null mice (at 3 mo of age) (Dataset S3). Whole embryo of offspring from the crosses were taken at 17.5 d for analysis. DNA was extracted from both parents (liver) and all offspring and sent for genotyping by The Jackson Laboratory using the Mouse Diversity Genotyping Array to identify any potential de novo CNV in the offsprings. The fathers and mothers were either 129Sv$^{SL}$ or C57BL/6 but always different strains so offsprings were mixed of 129Sv$^{SL}$ and C57BL/6 strains. We identified the set of SNPs that were homozygous and different in the parents so the offspring should be heterozygous. From this set of SNPs, we computed the B allele frequency (BAF) and the log R ratio (LRR) using PennCNV (www.openbioinformatics.org/penncnv/). For the offspring, the BAF should be 0.5 and the LRR should be 0. Any significant deviation from these values may indicate a CNV. A de novo germ-line deletion would have BAF close to 0 or 1 and an LRR that is significantly negative. A de novo germ-line duplication would have BAF close to 0.33 or 0.67 and an LRR that is significantly positive. We took a sliding window of 31 SNPs and used the median value to reduce the noise for the BAF and LRR. To calculate the threshold for significance in BAF and LRR, we used the bootstrap approach. We randomized the location of SNPs across the genome and used the sliding windows to get the maximum and minimum value of BAF and LRR. We repeated this process 100 times and used the distribution of extremal BAF and LRR values to estimate threshold values with *P* value of 0.05. The BAF and LRR are shown for all mice with the threshold values as horizon lines (Dataset S3). Almost all of the CNVs detected in the offsprings were inherited. The evidence for inheritance was found in the parent contributing the CNV by looking at the LRR value at the CNV location. The second evidence was that multiple pups in the same litter or different litters would have the same CNV because the parents in these studies are often related by few generations. One embryo, batch 1/sample 16 from C57Bl6 *trp53*$^{+/+}$ female 9 mo and 129SvSL *trp53*$^{+/+}$ male 3 mo, seemed to have a de novo duplication in chromosome 16 (Dataset S3). Two siblings (batch 1/samples 3 and 4 from C57Bl6 *trp53*$^{+/+}$ female 3 mo and 129SvSL *trp53*$^{+/+}$ male 3 mo) seemed to have an identical de novo deletion on chromosome 11 not found in any other pups. Similarly, two siblings (batch 3/samples 10 and 11 from 129SvSL *trp53*$^{-/-}$ female 3 mo and C57Bl6 *trp53*$^{+/+}$ male 3 mo) had an identical de novo duplication on chromosome 17 not found in any other pups. These CNVs might be a result of clonal mosaicism of gametes from the parents. One pup (batch1/sample13 from C57Bl6 *trp53*$^{+/+}$ female 9 mo and 129 Sv SL *trp53*$^{+/+}$ male 3 mo) had a germ-line trisomy of chromosome 18 (Dataset S3). Two adults, batch 2/sample 50 (C57Bl6 *trp53*$^{-/-}$ male 3 mo) and batch 3/sample 34 (129SvSL *trp53*$^{-/-}$ male 3 mo), seemed to have somatic mosaicism with a large number of aneuploidy in the liver (Dataset S3).

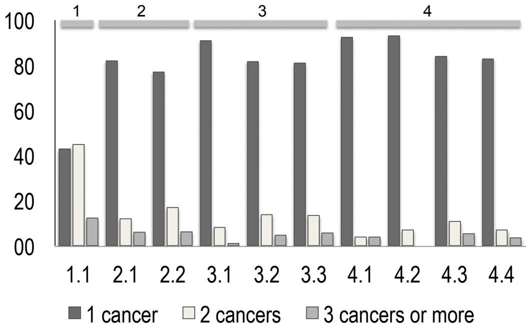

**Fig. S1.** Families are grouped in categories according to the number of documented generations (one, two, three, or four generations, gray bars at the top of the graph). In each group, generations are numbered, 1 being the oldest documented generation. Histograms show the proportion of patients with one, two, and three or more cancers diagnosed at any age. Only families with a single generation affected show a significantly higher number of subjects with more than one diagnosis (distribution in families with one generation versus the sum of most recent generations of families with two, three, or four generations ($P <$ 0.001; $\chi^2$ analysis). The proportion of patients with more than one diagnosis is comparable in all families with two or more documented generations, with no tendency toward increased multiple diagnoses with successive generations (comparison between generation 1 and generation 2 in families with two generations, $P = 0.6004$, Fisher's exact test; comparison between generation 1 and generation 3 in families with three generations, $P = 0.0581$, Fisher's exact test).

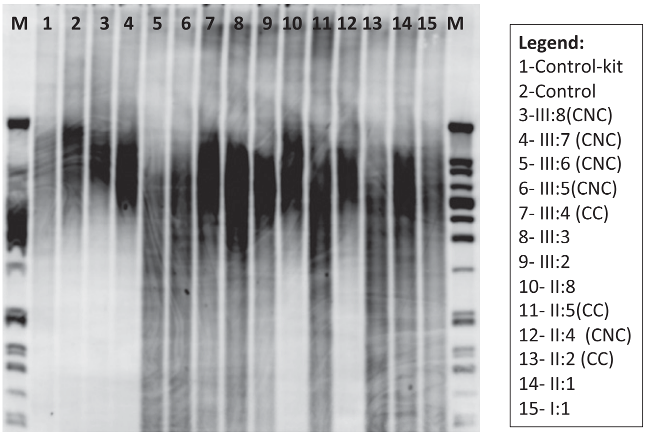

**Fig. S2.** Southern blot assessment of telomere length in members of the KA family using TeloTAGGG kit (Roche). CNC, carrier with no cancer; CC, carrier with cancer.

**Table S1. Haplotypes of *TP53* in family KA**

| | | | | Carriers with cancer | | Carriers without cancer | | Noncarriers | | Haplotype | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position | Ref. | Mother II:2 | Father II:1 | III:4 | III:1 | III:5 | III:6 | III:2 | III:3 | M1 | M2 | P1 | P2 |
| 7503347 | G | GG | GA | GG | GA | GG | GA | GA | GA | G | G | G | A |
| 7511548 | G | GG | GC | GG | GC | GG | GC | GC | GC | G | G | G | C |
| 7511654 | C | TT | CT | TT | CT | TT | CT | CT | CT | T | T | T | C |
| 7511655 | A | GG | AG | GG | AG | GG | AG | AG | AG | G | G | G | A |
| 7511681 | G | GG | GA | GG | GA | GG | GA | GA | GA | G | G | G | A |
| 7511703 | T | TT | TC | TT | TC | TT | TC | TC | TC | T | T | T | C |
| 7512177 | G | GG | GA | GG | GA | GG | GA | GA | GA | G | G | G | A |
| 7518132 | A | AC | AC | AA | AC | AA | AC | CC | CC | A | C | A | C |
| 7518152 | G | GA | GA | GG | GA | GG | GA | AA | AA | G | A | G | A |
| 7521849 | A | AG | AG | AA | AG | AA | AG | GG | GG | A | G | A | G |
| 7523808 | G | GA | GA | GG | GA | GG | GA | AA | AA | G | A | G | A |
| 7525125 | T | TC | TC | TT | TC | TT | TC | CC | CC | T | C | T | C |
| 7528624 | T | TC | TC | TT | TC | TT | TC | CC | CC | T | C | T | C |
| 7529195 | A | AT | AT | AA | AT | AA | AT | TT | TT | A | T | A | T |
| 7529285 | A | AG | AG | AA | AG | AA | AG | GG | GG | A | G | A | G |
| 7529913 | T | TC | TT | TT | TT | TT | TT | TC | TC | T | C | T | T |
| 7533098 | C | CG | CC | CC | CC | CC | CC | CG | CG | C | G | C | G |
| 7533285 | C | CT | CT | CC | CT | CC | CT | TT | TT | C | T | C | T |
| 7533505 | G | GA | GA | GG | GA | GG | GA | AA | AA | G | A | G | A |
| 7539985 | C | CG | CG | CC | CG | CC | CG | GG | GG | C | G | C | G |

Twenty SNPs of high confidence and informative for phasing of *TP53* region are shown with the genotypes of members in the family and their haplotypes. The TP53 mutation is found on haplotype M1.

**Table S2. Number of de novo and inherited deletions and duplications using different *P* value thresholds from CNValidator for CNV showing no significant difference between siblings in the KA family who are carriers with early cancer (III:1 and III:4), carriers without early cancer (III:5 and III:6), and noncarriers (III:2 and III:3)**

| Child | De novo deletion | De novo duplication | Inherited deletion | Inherited duplication | Total CNV |
|---|---|---|---|---|---|
| | | | $P < 0.01$ | | |
| III:1 | 1 | 0 | 67 | 55 | 123 |
| III:2 | 0 | 1 | 58 | 59 | 118 |
| III:3 | 0 | 0 | 58 | 56 | 114 |
| III:4 | 0 | 0 | 71 | 66 | 137 |
| III:5 | 0 | 0 | 64 | 61 | 125 |
| III:6 | 1 | 0 | 68 | 53 | 122 |
| | | | $P < 0.001$ | | |
| III:1 | 1 | 0 | 60 | 17 | 78 |
| III:2 | 0 | 0 | 56 | 23 | 79 |
| III:3 | 0 | 0 | 56 | 18 | 74 |
| III:4 | 0 | 0 | 66 | 24 | 90 |
| III:5 | 0 | 0 | 62 | 22 | 84 |
| III:6 | 0 | 0 | 62 | 19 | 81 |
| | | | $P < 0.0001$ | | |
| III:1 | 0 | 0 | 56 | 8 | 64 |
| III:2 | 0 | 0 | 54 | 7 | 61 |
| III:3 | 0 | 0 | 50 | 4 | 54 |
| III:4 | 0 | 0 | 62 | 3 | 65 |
| III:5 | 0 | 0 | 58 | 5 | 63 |
| III:6 | 0 | 0 | 59 | 1 | 60 |

**Table S3. Candidate genetic modifiers for anticipation in family KA**

| Variant | Nucleotide change | Amino acid change | Accession no. | Encoded protein | Function | p53 interaction |
|---|---|---|---|---|---|---|
| RPS6KA1 | c.G1661A | p.R554H | NM_002953 | Ribosomal protein S6 kinase alpha-1 | Control of cell growth and differentiation | No known interaction |
| SCMH1 | c.A45C | p.K15N | NM_001172219 | Human sex comb on mid-leg homolog 1 | DNA binding and sequence-specific DNA binding transcription factor activity | No direct interaction, but forms part of the PRC1 complex regulated by p53 |
| CACNA1E | c.G6347A | p.R2116H | NM_001205293 | Calcium channel, voltage-dependent, R type, alpha 1E subunit | Calcium-dependent processes including gene expression, cell division and cell death | No direct protein–protein interaction but CACNA1E contains p53 binding site |
| LOC729059 | c.G189T | p.L63F | NM_001242521 | Novel uncharacterized protein | Novel uncharacterized protein | Novel uncharacterized protein |
| TTC12 | c.G585C | p.E195D | NM_017868 | Tetratricopeptide repeat domain 12 | Protein binding | No known interaction |
| ACSS3 | c.T482C | p.V161A | NM_024560 | Acyl-CoA synthetase short-chain family member 3 | Activation of acetate for lipid synthesis or energy generation | No known interaction |
| PAPD4 | c.G1178T | p.G393V | NM_001114394 | PAP associated domain containing 4 | Cytoplasmic poly(A) RNA polymerase and metal ion binding | No direct interaction but regulates miR-122 stability to control p53 protein levels |
| EPHA5 | c.C712T | p.R238X | NM_004439 | Ephrin type-A receptor 5 | Receptor tyrosine kinase implicated in regulating development | Direct protein–protein interaction |
| ZFYVE16 | c.C565G | p.Q189E | NM_014733 | Zinc finger, FYVE domain containing 16 | Regulate endosomal membrane trafficking and forms part of various signaling pathways | Direct protein–protein interaction |
| LRIG3 | c.A364G | p.N122D | NM_153377 | Leucine-rich repeats and Ig-like domains 3 | DNA ligase integral to DNA replication and repair | Correlated expression in cancers but no known interaction |
| CEP350* (de novo variant found in III:4) | c.T1404G | p.I468M | NM_014810 | Centrosomal protein 350 kDa | Required for anchoring of microtubules to centrosomes | Direct protein–protein interaction |

Rare paternally inherited variants found in the two carrier children with cancer (III:1 and III:4) but not in the carrier children without cancer (III:5 and III:6). One de novo mutation found in III:4. All variants were validated by Sanger sequencing. All these variants have not been reported as polymorphisms in the dbSNP and 1000 Genomes Project databases.

**Dataset S1. CNVs of members of the KA family from aCGH**

Dataset S1

**Dataset S2. CNVs of members of the MM family from aCGH**

Dataset S2

**Dataset S3.** **Genotypes and CNVs of mice ordered by their batch and sample number. Parents followed by their offsprings are grouped together, and their genotypes and age at mating (for parents) are shown. DNA of parents were taken from liver whereas DNA of offsprings were extracted from whole embryos at 17.5 d old. CNVs of mice are ordered by batch and sample number. *Upper* shows the B allele frequency (BAF) whereas *Lower* shows the log R ratio (LRR). Chromosomes are ordered in ascending numerical order and shown in different colors**

Dataset S3