Radiation-related genomic profile of papillary thyroid cancer after the Chernobyl accident

One Sentence Summary (125 character maximum): Post-Chernobyl papillary thyroid cancers demonstrate radiation dose-dependent increases in clonal DNA double-strand breaks.

Lindsay M. Morton[1,*], Danielle M. Karyadi[2,*], Chip Stewart[3,*], Tetiana I. Bogdanova[4,†], Eric T. Dawson[2,5,†], Mia K. Steinberg[6,†], Jieqiong Dai[6], Stephen W. Hartley[2], Sara J. Schonfeld[1], Joshua N. Sampson[7], Yosi Maruvka[3], Vidushi Kapoor[7], Dale A. Ramsden[8], Juan Carvajal-Garcia[9], Chuck M. Perou[10,11], Joel S. Parker[11], Marko Krznaric[12], Meredith Yeager[6], Joseph F. Boland[6], Amy Hutchinson[6], Belynda D. Hicks[6], Casey L. Dagnall[6], Julie M. Gastier-Foster[13,14], Jay Bowen[13], Olivia Lee[2], Mitchell J. Machiela[15], Elizabeth K. Cahoon[1], Alina V. Brenner[16], Kiyohiko Mabuchi[1], Vladimir Drozdovitch[1], Sergii Masiuk[17], Mykola Chepurny[17], Liudmyla Yu. Zurnadzhy[4], Maureen Hatch[1], Amy Berrington de Gonzalez[1], Gerry A. Thomas[12,‡], Mykola D. Tronko[18,‡], Gad Getz[3,19,20,‡], Stephen J. Chanock[2,‡]

[1]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, 20892, United States

[2]Laboratory of Genetic Susceptibility, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, 20892, United States

[3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, 02142, United States

[4]Laboratory of Morphology of the Endocrine System, V.P.Komisarenko Institute of Endocrinology and Metabolism of the National Academy of Medical Sciences of Ukraine, Kyiv, 04114, Ukraine

[5]Nvidia Corporation, Santa Clara, California, 95051, United States

[6]Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Bethesda, Maryland, 20892, United States

[7]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, 20892, United States

[8]Department of Biochemistry and Biophysics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, United States

[9]Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, United States

[10]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, United States

[11]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, United States

[12]Department of Surgery and Cancer, Imperial College London, Charing Cross Hospital, London, W6 8RF, United Kingdom

[13]Nationwide Children's Hospital, Biospecimen Core Resource, Columbus, Ohio, 43205, United States

[14]Departments of Pathology and Pediatrics, Ohio State University College of Medicine, Columbus, Ohio, 43210, United States

[15]Integrative Tumor Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, 20892, United States

[16]Radiation Effects Research Foundation, Hiroshima, 732-0815, Japan

[17]National Research Center for Radiation Medicine of the National Academy of Medical

Sciences of Ukraine, Kyiv, 04050, Ukraine

[18]Department of Fundamental and Applied Problems of Endocrinology, V.P.Komisarenko Institute of Endocrinology and Metabolism of the National Academy of Medical Sciences of Ukraine, Kyiv, 04114, Ukraine

[19]Center for Cancer Research and Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, 02114, United States

[20]Harvard Medical School, Boston, Massachusetts, 02115, United States

[*]These authors contributed equally to this work.

[†]These authors contributed equally to this work.

[‡]These authors jointly directed this work.

Correspondence to: chanocks@mail.nih.gov, mortonli@mail.nih.gov

Abstract (125 word maximum):

The 1986 Chernobyl nuclear power plant accident increased papillary thyroid cancer (PTC) incidence in surrounding regions, particularly for ^{131}I-exposed children. We analyzed genomic, transcriptomic, and epigenomic characteristics of 440 PTCs from Ukraine (359 with estimated childhood ^{131}I exposure and 81 unexposed children born after 1986). PTCs displayed radiation dose-dependent enrichment of fusion drivers, nearly all in the mitogen-activated protein kinase pathway, and increases in small deletions and simple/balanced structural variants that were clonal and bore hallmarks of non-homologous end-joining repair. Radiation-related genomic alterations were more pronounced for those younger at exposure. Transcriptomic and epigenomic features were strongly associated with driver events but not radiation dose. Our results point to DNA double-strand breaks as early carcinogenic events that subsequently enable PTC growth following environmental radiation exposure.

The accidental explosion in reactor 4 at the Chernobyl nuclear power plant in April 1986 resulted in the exposure of millions of inhabitants of the surrounding areas of Ukraine, Belarus, and the Russian Federation to radioactive contaminants (*1*). Epidemiologic and clinical research in the ensuing decades has demonstrated increased risk of papillary thyroid carcinoma (PTC) with increasing thyroid gland exposure to radioactive iodine (^{131}I) from fallout, which was deposited on pastures with grazing cows and ingested through milk, particularly during early childhood (*2*). Together with data from populations exposed to other types of radiation, compelling evidence indicates that PTC risk increases following childhood exposure to ionizing radiation, a recognized carcinogen (*2-5*).

Currently, there are no established molecular biomarkers for cancers induced by radiation, nor have there been large-scale analyses of the genomic landscape of human cancers occurring after a well-quantified radiation exposure. Classical cytogenetic studies have demonstrated radiation dose-associated increases in large chromosomal aberrations (such as inversions and translocations) that reflect DNA double-strand breaks and are the current standard for biodosimetry; however, these assays are typically performed in peripheral blood lymphocytes from individuals exposed to whole body irradiation and have not been directly linked to tumor characteristics (*6, 7*). Next-generation sequencing of 12 second primary tumors of various types that occurred within the field of previous therapeutic ionizing radiation suggested an excess of small deletions and balanced inversions (*8*), but radiation dose estimates were not available. RNA sequencing (RNA-seq) analyses of 65 PTCs (mean age at diagnosis=24.7 years) occurring after the Chernobyl accident identified that higher doses were associated with an increased likelihood of gene fusion drivers (*9*). In a genomic landscape analysis of 496 primarily unexposed PTCs (mean age at diagnosis=46.8 years; 16 patients with known prior radiation

exposure), The Cancer Genome Atlas (TCGA) reported a low density of somatic point mutations, especially for PTCs in younger patients, and a high frequency of activating somatic alterations in the mitogen-activated protein kinase (MAPK) pathway, including point mutations in *BRAF* (61.7%) and *NRAS/HRAS/KRAS* (12.9%) as well as fusions with *RET* (6.8%), *BRAF* (2.7%) and other MAPK-related genes (5.0%) (*10*).

Here, we report a comprehensive characterization of the genomic, transcriptomic, and epigenomic profile of PTC as well as non-tumor thyroid tissue and/or blood for 440 individuals from Ukraine who developed PTC after the Chernobyl accident (mean age at PTC=28.0 years), affording an opportunity to investigate the contribution of environmental radiation to PTC characteristics. The study analyzed a collection of pre-treatment fresh frozen tumor tissues with pathological confirmation of first primary PTC by an international panel of experts through the Chernobyl Tissue Bank (CTB) (*11*, *12*). Our study included 359 individuals with PTC with well-quantified ^{131}I-exposure before adulthood (≤18 years of age; mean=7.3 years) and, as controls, 81 ^{131}I-unexposed individuals with PTC born >9 months after the Chernobyl accident (all were born after March 1, 1987) (*13*).

RESULTS

Samples, Clinical Data, and Analytic Approach

Based on the availability of sufficient DNA and RNA extracted from CTB samples, we analyzed up to 440 individuals with whole genome sequencing (WGS) and/or mRNA-seq of pathologically confirmed tumor (374 both, 57 mRNA-seq only, 9 WGS only) (**Fig. S1-2**) (*13*). Matched normal tissue with WGS and/or mRNA-seq included non-tumor thyroid tissue and/or blood (233 both, 182 non-tumor tissue only, 16 blood only, 9 normal tissue not available). The

genomic landscape characterization was augmented by single nucleotide polymorphism (SNP) microarray genotyping (Illumina Infinium HumanOmniExpress-24 array) and relative telomere length quantification on all samples, and DNA methylation profiling (Illumina Infinium MethylationEPIC array) and microRNA (miRNA)-seq for PTC and non-tumor thyroid tissue (**Fig. S3**). A total of 357 individuals had tumor sample data across all platforms.

The majority of individuals with PTC were female (n=335, 76.1%), resided in the Kiev region at the time of the accident (n=286, 65.0%), and were diagnosed during young adulthood (mean=28.0 years, range: 10.0-45.6); ^{131}I-unexposed individuals with PTC were born at least 9 months after the accident and thus had a younger average age than the exposed individuals (mean: unexposed=20.7 years, exposed=29.7 years) (**Table S1, Fig. S4**). For ^{131}I-exposed individuals, mean age at exposure was 7.3 years (range: *in utero* to 18.9 years) and mean time from the accident to PTC diagnosis was 22.4 years (range: 12.5-29.9). Radiation doses to the thyroid were reconstructed by an international team of dosimetry experts (*14-16*). For 53 individuals, doses were estimated using detailed information derived from individual direct thyroid radioactivity measurements taken within 8 weeks of the accident, with (n=49) or without (n=4) personal interviews regarding residential history and dietary patterns. For the remaining individuals, dose estimates were derived from direct measurements taken for other individuals who lived in the same residential area (n=249), neighboring area (n=9), or other areas (n=39), or based on dose estimates to the mother for individuals who were *in utero* at the time of the accident (n=9). Mean estimated radiation dose was 250 mGy (range: 11-8,800) (**Fig. S4-S5**).

Our primary analyses investigated the relationship between ^{131}I dose and 68 PTC molecular characteristics derived from a comprehensive genomic landscape analysis (**Fig. 1**) using multivariable linear, proportional odds, or logistic regression models adjusted for sex and

age at PTC diagnosis (*13*). For associated variables (defined as $P<7.4\times10^{-4}$ based on a Bonferroni correction for 68 tests), further analyses were conducted by specific molecular characteristics, as well as by age at PTC, age at exposure, and time since exposure (latency) because these factors influence radiation-related thyroid cancer risk (*17*). In addition, we conducted sensitivity analyses to assess whether the results were consistent when we restricted the population to ^{131}I-exposed individuals with lower radiation dose (<500 mGy, resulting in n=326 with mean dose=110 mGy) (*18, 19*).

Simple somatic variants

WGS analysis of tumor/normal pairs (n=383; mean sequencing depth, tumor=89X, non-tumor thyroid tissue=33X, blood=33X; **Table S2**) revealed a low burden of simple somatic variants (SSV) (mean=0.27 nonsynonymous mutations per Mb) (**Fig. 1, Fig. S6-S7**), which was lower than in older TCGA PTC cases (0.41 nonsynonymous mutations per Mb) (*10*) and comparable to mutationally quiet tumors typically reported for pediatric cancers (*20*). A total of 318,957 SSVs were identified, the majority (93.3%) of which were single nucleotide variants (SNVs) (n=297,513 in 383 tumors; mean per tumor=776.8), whereas small insertions and deletions (indels) were less common (insertions: n=5,842, 1.8%, mean=37.2; deletions: n=14,231, 4.5%, mean=15.3), and doublet and triplet base substitutions were rare (dinucletodie polymorphism [DNP] or doublet: n=1,351, 0.4%, mean=3.5; trinucleotide polymorphism [TNP] or triplet: n=20, 0.006%) (**Table S3**). Among the 3,886 coding mutations (1.2% of total; 0.35/Mb), most were nonsynonymous (3,023/3,886=77.8%). Approximately one-third of mutations (n=114,898, 36.0%) were clonal (cancer cell fraction>0.9), regardless of mutation type (SNV=35.9%, insertions=36.7%, deletions=38.0%) (**Table S3, Fig. S8**) (*13*).

In multivariable analyses restricted to n=356 samples with both high tumor and normal tissue purity (**Fig. 1**, **Fig. S9**) (*13*), increased radiation dose was associated with an increase in small deletions ($P=8.0\times10^{-9}$) as well as the deletion:SNV ratio ($P=4.9\times10^{-21}$), but not other SSV types (**Fig. 2, Table S4**). In addition, we observed the expected increase in the burden of SNVs ($P=3.2\times10^{-6}$), doublet mutations ($P=2.7\times10^{-5}$), insertions ($P=1.5\times10^{-6}$), and deletions ($P=7.4\times10^{-16}$) with increasing age at PTC diagnosis (**Table S4**) (*10*). Few of these mutations were clustered (22 clusters [>2 mutations within 150 base pairs (bp)] in 18 cases; 83 clusters [>2 mutations within 1 kb] in 36 cases) and were not associated with radiation dose (P>0.3). In an analysis of the frequency and types of microsatellite indels in the tumors, detected using MSMuTect (*21*), all tumors were microsatellite stable (mean [range] per tumor, insertions=1.8 [0-7], deletions=7.3 [0-24]), and radiation dose was not significantly associated with the number of microsatellite insertions or deletions (**Table S4**).

An investigation of mutational processes in PTC was conducted using SigProfiler to determine both single base substitution (SBS) and small indel (ID) mutational signatures (*20, 22*). Comparing the PTC mutations with known signatures from the Catalogue of Somatic Mutations in Cancer (COSMIC v3, https://cancer.sanger.ac.uk/cosmic/signatures) (*20*), the majority of the SBS signatures (69.9%) were attributable to clock-like signatures (SBS1=9.8%, SBS5=60.2%), with smaller fractions due to APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) cytidine deaminase DNA-editing activity (SBS2=6.2%, SBS13=6.4%), damage from reactive oxygen species (SBS18=0.9%), and two signatures of unknown etiology (SBS8=15.1%, SBS23=1.6%) (mean cosine similarity between actual mutations and attributed patterns=0.94) (**Fig. 1**, **Table S5, Fig. S10-S12**). In multivariable analyses, no SBS signatures were significantly associated with radiation dose, whereas increased

age at PTC diagnosis was associated with an increase in clock-like SBS mutations ($P_{SBS1}=1.9\times10^{-7}$; $P_{SBS5}=6.8\times10^{-17}$) as well as SBS8 ($P=4.3\times10^{-11}$) and SBS18 ($P=2.0\times10^{-8}$) (**Table S4**).

The majority (54.0%) of indels were attributed to clock-like signatures (ID1=14.2%, ID5=39.8%), 21.2% to ID3 (tobacco smoking, which is not a major risk factor for PTC), 19.0% to repair of DNA double-strand breaks by end-joining mechanisms (ID6=3.3%, ID8=15.8%), and 5.8% to ID4 (unknown etiology) (mean cosine similarity=0.77) (**Fig. 1, Table S5, Fig. S10-S12**). In multivariable analyses, radiation dose was strongly associated with end-joining-related indel mutational patterns ($P=1.5\times10^{-10}$), particularly ID8 ($P=7.3\times10^{-9}$), and more weakly with the clock signature ID5 ($P=1.3\times10^{-4}$) (**Fig. 2**, **Table S4**). In comparison, increased age at PTC diagnosis was associated with significantly increased numbers of ID3 ($P=1.9\times10^{-7}$), ID5 ($P=1.9\times10^{-9}$), and ID8 ($P=6.6\times10^{-4}$) mutational patterns (**Table S4**). *De novo* signature extraction did not reveal a novel signature related to environmental exposure to ionizing radiation but identified 4 SBS (mean cosine similarity=0.96) and 2 ID (mean cosine similarity=0.83) signatures highly correlated with COSMIC signatures described above (**Table S4**, **Table S6, Fig. S10-S12**). Similarly, no novel signature was identified when we restricted the analysis to PTCs in individuals who received ≥200 mGy (**Table S5**).

Structural variation

Overall, 479 structural variants (SV) were identified in 356 tumors; approximately one-quarter of SVs (n=132, 27.6%) were simple/balanced events (balanced interchromosomal translocations and inversions), one-half (n=253, 52.8%) were simple/unbalanced (deletions, unbalanced interchromosomal translocations), and the remaining (n=94, 19.6%) were complex

(>2 breaks repaired together in a cluster) (**Fig. S13-14, Table S6**) (*13*). Approximately one-third of tumors (n=113, 31.7%) had no SV, one-third (n=126, 35.4%) had one SV, and the remaining third had two (n=61, 17.1%) or more (n=56, 15.7%) SVs. Two tumors had >10 SV events (**Fig. S15**), one of which was the only tumor with evidence for chromothripsis (age at PTC=30.2 years, age at exposure=1.3 years, dose=1000 mGy).

Multivariable analyses (n=354, excluding the 2 outliers) demonstrated that increasing radiation dose was significantly associated with increased SV count ($P=1.4\times10^{-8}$), particularly simple/balanced SVs ($P=1.2\times10^{-14}$) but not those classified as complex (P=0.52) or simple/unbalanced ($P=5.6\times10^{-3}$) (**Fig. 3A**, **Table S4**). Increasing radiation dose also was not associated with occurrence of chromoplexy (P=0.70) (**Table S4**), which was identified in 19 tumors (n=15 single event, n=4 two events) (*13, 23*), nearly all unexposed or in the lower dose groups ($n_{unexposed}=7$, $n_{1\text{-}99\ mGy}=8$, $n_{100\text{-}199\ mGy}=2$, $n_{\geq500\ mGy}=2$).

Somatic copy number alteration

A total of 40.3% (n=143/355) of the tumors evaluated for somatic copy number alterations (SCNA) had one (n=96, 27.0%) or more (n=47, 13.2%) such events (**Fig. S16**) (*13*). Four tumors had ≥20 SCNAs each (**Fig. S17**): the tumor with chromothripsis and three additional tumors (age at PTC=19.0-29.3 years, age at exposure=1.3-2.3 years, dose=125-175 mGy). Those three tumors predominantly had gains or copy neutral loss of heterozygosity (CNLOH) and were the only tumors with ploidy>2.5, with one of the three displaying extensive CNLOH (>20 arms), similar to previous reports for the rare thyroid Hürthle cell carcinoma (*24, 25*). Exclusion of the four tumors with ≥20 SCNAs each yielded 239 total SCNAs: 69 (28.9%) at

the chromosome level, of which 48 were deletions, and 170 (71.1%) sub-chromosomal, of which 106 were deletions (**Table S6**).

In multivariable models, radiation dose was related to the number of sub-chromosomal SCNAs ($P=3.5\times10^{-5}$), particularly deletions ($P=7.0\times10^{-4}$) but not gains (P=0.32) or CNLOH (P=0.52); radiation dose also was not related to the number of chromosomal SCNAs (P=0.20) (**Table S4**). The most frequent recurrent event was loss of 22q (n=47/353, 13.3%) (**Fig. S18-S19**), but occurrence of 22q deletions was not associated with radiation dose (P=0.37) (**Table S4**).

Drivers of PTC

We identified at least one candidate driver for 433 of 440 (98.4%) tumors (**Fig. S20, Table S7**) (*13*), with the majority (n=401; 92.6%) having a single candidate driver, underscoring the parsimony of events driving PTC carcinogenesis. We designated 429 (97.5%) drivers for analysis (**Fig. 1**). Over half the designated driver events (n=253, 59.0%) were mutations (SSVs; **Table S8**), predominantly activating point mutations in genes previously implicated in PTC. The most commonly mutated gene was *BRAF* (n=194, 45.2%), where all the mutations either were canonical $BRAF^{V600E}$ substitutions (n=190) or disrupted the V600 sequence context (n=4). *RAS* genes were the next most commonly mutated (n=44, 10.3%), specifically *NRAS* (n=20, 4.7%), *HRAS* (n=15, 3.5%), and *KRAS* (n=9, 2.1%). Additional mutation drivers were identified in *TSHR* (n=6, 1.4%), *DICER1* (n=3, 0.7%), *APC* (n=2, 0.5%), *TSC1/TSC2* (n=2, 0.5%), and *NFE2L2* (n=2, 0.5%). In TCGA, 9.4% of PTC harbored *TERT* promoter mutations, often in older individuals (*8*), but only one individual with a *TERT* promoter mutation was observed in our study (age at PTC diagnosis=40.7 years, designated driver=$BRAF^{V600E}$).

Fusion drivers accounted for 176 (41.0%) PTC cases (**Table S9**). The most frequently involved genes were *RET* (n=73, 17.0%) as well as other receptor tyrosine kinase (RTK) genes, specifically *NTRK3* (n=36, 8.4%), *NTRK1* (n=13, 3.0%), *ALK* (n=12, 2.8%), and *LTK* (n=3, 0.7%). Additional fusion drivers included *BRAF* (n=20, 4.7%) and *PPARG* (n=13, 3.0%), as well as SVs that resulted in overexpression of *IGF2* or *IGF2BP3* (n=6, 1.4%). Of the 23 chromoplexy events described above, 16 generated driver fusions. All 22q deletions co-occurred with known driver mutations, most frequently *RAS* mutations ($P_{heterogeneity}=2.8\times10^{-10}$; n=22/38, 56.4%) (**Table S10**).

In multivariable analyses, fusion drivers in PTC were more common in individuals exposed to higher radiation dose ($P=6.6\times10^{-8}$) and in those diagnosed at younger ages ($P=5.4\times10^{-9}$) relative to those with mutation drivers (**Fig. 3B, Table S4**). There was a suggestion of a heterogeneous effect of dose by specific gene fusion ($P_{heterogeneity}=0.020$), with higher doses on average for PTCs with *IGF2/IGF2BP3* or *BRAF* fusion drivers, whereas the dose distribution did not differ significantly among mutation drivers (*BRAF*, *RAS*, other; $P_{heterogeneity}=0.17$) (**Fig. 4**). We extended our observations by inclusion of 45 non-overlapping individuals with PTC (excluding the 20 individuals already in our analyses) drawn from a previous Chernobyl study with known drivers identified with RNAseq (*7*); over half had doses ≥500 mGy (mean age at PTC=24.2 years, mean age at exposure=7.2 years, mean dose=1050 mGy) (**Table S1**). That smaller sample set also suggested a radiation dose-related increase in fusion drivers (P=0.069), which was consistent with the results from our study ($P_{heterogeneity}=0.90$; pooled analysis of fusion vs. mutation driver, adjusting for age at PTC, sex, and study: $P=4.6\times10^{-9}$).

Gene expression and methylation patterns

We conducted several analyses to assess whether gene expression and methylation patterns were related to radiation dose. First, unsupervised clustering analyses restricted to PTC tumor tissue yielded 5 mRNA clusters, 5 miRNA clusters, and 3 methylation clusters (**Fig. S21-S22**) (*13*). None of these clusterings were associated with radiation dose (P_{mRNA}=0.85; P_{miRNA}=0.38; $P_{methylation=}$0.10), but each closely correlated with the driver gene pathway (P_{mRNA} =1.6×10^{-64}; P_{miRNA}=1.0×10^{-9}; $P_{methylation=}$$6.4\times10^{-43}$) (**Fig. 5A, Table S4, Table S11**, **Fig. S23**), supporting the overriding importance of the driver for RNA expression patterns (*10*). Second, we identified three transcriptional patterns important in PTC based on the TCGA analysis (*10*): the *BRAF*^{V600E}*-RAS* score (BRS), estimating the degree to which the mRNA, miRNA, and methylation profiles resemble either *BRAF*^{V600E} or *RAS*-mutated PTC; the thyroid differentiation score (TDS), based on expression of 16 thyroid metabolism and function genes; and the ERK-activity score of 52 expressed genes responsive to MEK inhibition (*13*). As expected, the mRNA, miRNA, and methylation BRS scores were highly correlated with one another (r=0.78-0.92) (**Table S12**). Consistent with TCGA (*10*), the three different mRNA-based scores also were significantly correlated, particularly the BRS with both the TDS (r=0.69) and the ERK score (r=-0.66). None of the scores were associated with radiation dose after correction for multiple testing ($P_{mRNA\text{-}BRS}$=2.1×10^{-3}; $P_{miRNA\text{-}BRS}$=5.5×10^{-3}; $P_{methylation\text{-}BRS}$=0.082; P_{ERK}=0.011, P_{TDS}=7.8×10^{-3}) (**Table S4, Fig. S24A**), whereas each was strongly related to driver gene pathway ($P<1.0\times10^{-30}$ for all scores) (**Fig. S24B**).

To confirm the lack of association between radiation exposure and gene expression patterns, we conducted exploratory analyses of the differential expression of specific genes and gene sets by dose (*13*). In multivariable linear regression models adjusted for age at PTC, sex,

and batch, the $P_{adjusted}$ for dose was <0.05 for five genes (**Fig. 5B, Table S13**), with the smallest P-value ($P_{adjusted}=8.0\times10^{-3}$; $\log_{10}$-fold expression change/100 mGy=0.059) for transfer RNA asparagine (anticodon GUU) (*TRNAN-GUU-2*), which is a target of the transcription factor *MYBL2*, a key regulator of cell cycle progression and apoptosis (*26, 27*). However, each of these associations were attenuated when the model was further adjusted for the driver gene pathway (**Table S13**). In contrast, over half the genes were differentially expressed ($P_{adjusted}<0.05$) among the different driver gene pathways (**Table S13**). Despite previous reports suggesting radiation dose could be linked to *CLIP2* expression (*28-30*), no such relationship was observed in our substantively larger study, which included 33 overlapping samples from the previously-published analyses (*29*) (**Fig. S25**), either in the overall set of PTC cases (P=0.42, **Fig. 5C**) or in subsets defined by early age at radiation exposure (**Fig. S26**). An exploration of expression signatures through gene set enrichment analyses was pursued in the Molecular Signatures Database (MSigDB; https://www.gsea-msigdb.org/gsea/msigdb) (*31*). For 3,213 gene sets (those related to hallmark biological processes, thyroid, radiation, and the genes included in germline analyses below) (*13*), multivariable regression models adjusting for age at PTC, sex, and batch revealed similar results as those above for single gene differential expression analyses, namely no gene set expression patterns were significantly associated with radiation dose ($P_{adjusted}<0.05$), whereas over half were strongly associated with the driver gene pathway (**Fig. S27, Table S14**).

Germline genetic variation

Possible contribution of germline genetic variation to radiation-related PTC was investigated in individuals of comparable Ukrainian ancestry (n=383 individuals, including 305 exposed, 78 unexposed) (**Fig. S28**). Twelve previously reported risk SNPs for sporadic PTC

were used to generate a polygenic risk score (PRS) (*32*). Multivariable analyses adjusting for population substructure revealed that unexposed individuals with PTC and those who received lower radiation doses were more likely to have higher genetic risk ($P=4.7\times10^{-4}$) (**Fig. 6**, **Table S4**). Analyses of the 12 individual SNPs, albeit underpowered, yielded three possible associations with radiation dose: rs1588635 (9q22.33; P=0.012), rs2289261 (15q22.33, *SMAD3*; P=0.030) and rs10069690 (5p15.33, *TERT*; P=0.054) (**Table S15**).

Investigation of rare potentially protein-damaging variants in genes and pathways related to thyroid or other cancer predisposition, clinical radiation sensitivity syndromes, and DNA damage response revealed no major differences in the burden of these variants among individuals who developed PTC after different radiation doses (**Table S4, Table S16-S17**). Only four individuals (n=2, 2.6% unexposed; n=2, 1.2% <100 mGy; 0, 0% ≥100 mGy) carried potentially protein-damaging variants in known thyroid cancer susceptibility genes (**Table S16-S17**).

Detailed analyses of molecular characteristics associated with radiation dose

Analyses by clonality for each of the deletion metrics (total deletion count, deletion:SNV ratio, and ID5 and ID8 mutational patterns) revealed that the radiation dose-related associations were restricted consistently to clonal rather than subclonal deletions (**Fig. 2, Table S18**). Similarly, analyses of SCNAs also demonstrated associations only with clonal but not subclonal sub-chromosomal deletions (**Table S18**). Because distinct repair mechanisms can generate deletions of different lengths (*33-35*), we further stratified the clonal deletion count by length (**Fig. S29**) and found the strongest association between radiation dose and ≥5 bp clonal deletions with patterns characteristic of end-joining repair ($P=4.9\times10^{-31}$) (**Fig. 3C, Table S18**). These

results are consistent with the ID8 mutational association and suggest a key role for end-joining mechanisms in repairing radiation-induced DNA double-strand breaks. Analyses of the ≥5 bp clonal deletions by the amount of microhomology at the deletion boundary revealed consistent associations between radiation dose and deletions with 0-1 bp microhomology as well as those with ≥2 bp microhomology (**Table S18**). These results implicate non-homologous end-joining (NHEJ) repair mechanisms, which are employed regardless of the amount of microhomology, whereas alternative end-joining (alt-EJ) repair mechanisms such as theta-mediated end-joining (TMEJ) typically generate deletions with ≥2 bp microhomology (*33*-*35*). In an ancillary analysis, we quantified in the small insertions the number of TINS (locally templated insertions), which are characteristic of TMEJ repair (*34*), and found that TINS were not associated with radiation dose (P=0.69) (**Fig. S30**). Further exploration of insertions and deletions by genomic sequence context (*8*) revealed only weak correlations for radiation dose with occurrence of deletions classified by flanking GC content (P=0.015), proximity to CPG islands (P=0.014), and the mean replication timing at the variant locus (P=0.010); in each case deletions in the higher radiation dose groups were more similar to a random background distribution (**Table S19)** (*13*). No such correlations between radiation dose and genomic sequence context were observed for insertions.

We undertook similar analyses of SVs after confirming each event (**Table S6**) (*13*), identifying those SVs with <20 bp of intervening loss/gain at the breakpoint, which indicates repair by end-joining mechanisms (*33*-*35*). Increased radiation dose was strongly associated with simple/balanced SVs that were clonal ($P=1.4\times10^{-16}$) but not subclonal (P=0.91), with a pronounced association for clonal simple/balanced SVs enriched for patterns characteristic of end-joining mechanisms ($P=5.5\times10^{-19}$) (**Fig. 3D**) versus other clonal SVs (P=0.41, **Fig. 3E**) (**Table S18**). Further analyses demonstrated consistent associations for radiation dose with clonal

simple/balanced/end-joining SVs with <4 bp and 4-<20 bp of intervening loss/gain (**Table S18**). Similar to our observations for small deletions, these results specifically implicate the importance of NHEJ repair, which accounts for almost all <4 bp events but which could contribute regardless of the amount of intervening loss/gain. By comparison, alt-EJ repair mechanisms primarily give rise to events with ≥4 bp of intervening loss/gain (*33*-*35*). Additional analyses of clonal simple/balanced/end-joining SVs by type revealed a strong association between radiation dose and inversions ($P=3.6\times10^{-14}$), consistent with a previous report (*8*), but also an association with translocations ($P=4.4\times10^{-4}$) (**Table S18**).

For each of the radiation dose-associated variables, the results were similar when we restricted the study population to exposed individuals (**Table S20**). Albeit based on limited statistical power, further restriction to individuals with exposures 1-<500 mGy revealed consistent associations for dose only with the clonal deletion:SNV ratio, enrichment of fusion drivers, and presence of clonal simple/balanced/EJ SVs (**Table S20**). Linear-quadratic and linear-exponential models of radiation dose generally did not improve the model fit compared with a linear model for any of the variables, except for clonal small deletions (total and restricted to ≥5 bp EJ deletions) (**Table S20**).

Notably, radiation dose-related increases in clonal deletions (particularly the deletion:SNV ratio, ID8, and the number of clonal ≥5 bp EJ small deletions), as well as fusion PTC drivers were substantially more pronounced for individuals exposed at younger ages (**Fig. 7, Table S21**-**S22**), albeit based on small numbers for certain analyses. In contrast, the radiation dose-related increase in SCNA clonal sub-chromosomal deletions was most pronounced at longer latencies (**Table S21**).

Radiation-related acceleration of PTC development

Exploratory analyses to address previous reports that ionizing radiation exposure accelerates aging and cancer development (*36, 37*) revealed no such evidence in our study population. First, we stratified analyses of the relationship of clock-like SBS and ID signatures with age at PTC and latency but found no effect modification by radiation dose (age at PTC: P_{SBS}=0.63, P_{ID}=0.93; latency: P_{SBS}=0.28, P_{ID}=0.21) (**Fig. S31-S32**). Additionally, radiation dose-dependent associations with key molecular characteristics did not appear to be strongly modified by latency, after accounting for age at exposure and age at PTC (**Table S21**). Analyses of relative telomere length demonstrated the expected association between decreased telomere length and increased age at PTC in blood ($P=3.2x10^{-5}$) but not in normal thyroid tissue (P=0.99) or PTC (P=0.81), and there was no association between relative telomere length and thyroid radiation dose (P>0.4 for all tissues). Methylation profiles were evaluated to estimate epigenetic age acceleration using two established metrics (*38, 39*). Regressing epigenetic age against chronological age in the non-tumor thyroid tissue and then comparing the residuals from this predicted age in the PTC tissue (*13*) revealed no association between age acceleration and radiation dose using either metric (P>0.1).

Discussion

Our large-scale integrated analyses of the genomic landscape of PTC that developed following the 1986 Chernobyl nuclear power plant accident provide consistent evidence that ionizing radiation-induced DNA damage, particularly double-strand breaks, represents an early carcinogenic event in thyroid tumorigenesis following radiation exposure. These findings

substantially extend preliminary reports of radiation-related human tumor characteristics (*8, 9*) by integrating data from multiple platforms with large sample size and detailed radiation dose data. Increasing radiation dose was strongly associated with increased likelihood of fusion versus point mutation drivers, simple/balanced SVs, and small deletions, with the strongest associations observed for those that bore hallmarks of NHEJ repair and were clonal, particularly for individuals exposed at a young age. However, no unique radiation-related biomarker was identified. Together, our results indicate that thyroid tumorigenesis following radiation exposure results from DNA double-strand breaks in the genome that have an impact on key thyroid cell growth and differentiation genes, which in turn drive the expression and epigenetic characteristics of individual PTCs.

Most tumors had evidence for only a single, known oncogenic driver, which involved the MAPK pathway in nearly all cases, which is consistent with previously published studies of sporadic PTCs (*10, 40, 41*). These findings combined with the low mutational burden in thyroid tumors emphasize the efficiency of driver mutations in thyroid tumorigenesis even following ionizing radiation exposure, in contrast to other environmentally-driven cancers, such as cigarette smoking and lung adenocarcinoma or ultraviolet light and melanoma, that often require multiple drivers and have multiple subclones together with substantial somatic burden (*23*).

Based on multiple lines of evidence, our study demonstrates striking radiation dose-related increases in DNA double-strand breaks in human thyroid cancers developing after the Chernobyl accident, extending results from in vitro and animal radiobiological experiments (*33-35*). In contrast, the PTCs did not have evidence of radiation-related specific base mutations or clustered mutations (*42*). Cells with DNA double-strand breaks can recruit various repair mechanisms, each of which leaves characteristic evidence in the repaired sequence. A series of

analyses consistently implicated NHEJ as the most important repair mechanism for the radiation dose-associated DNA double-strand breaks observed in the PTCs. While the importance of end-joining repair in human tumors has been reported previously (*8*), our detailed examination of the local sequence context for the SVs (including fusion drivers) and small deletions enabled us to identify that radiation dose was most clearly associated with NHEJ rather than alt-EJ or other repair mechanisms. The lack of association between radiation dose and TINS further demonstrated the lack of importance of alt-EJ mechanisms. The importance of NHEJ repair also was supported by the lack of significant association between radiation dose and mutation signatures of APOBEC, which preferentially targets intermediates in replication and repair by homologous recombination (*43*). Our results necessitate further research such as using genetically modified organoids (*44, 45*) to establish the causal role of radiation-related DNA double-strand breaks predominantly repaired by NHEJ in human carcinogenesis.

The role of radiation-related DNA damage as an early step in PTC carcinogenesis following the Chernobyl accident is further supported by the lack of association between radiation dose and PTC transcriptomic and epigenomic features, despite the use of various analytic approaches, including clustering, differential expression by gene or miRNA, and gene set enrichment analyses. With our large sample size, we did not confirm the previously reported association between radiation dose and *CLIP2* expression (*28-30*), even when we restricted our analyses to individuals exposed at younger ages. Notably, however, the PTC transcriptomic and epigenomic features differed strikingly by driver gene/pathway, supporting the importance of the specific driver in shaping the tumor profile (*10, 40, 41*). Utilization of both WGS and RNA-seq enabled us to identify a driver in 98% of the PTCs in our study. Deletion of chromosome 22q has been suggested as a driver for PTC, but all cases in our study with 22q deletions also had other

known PTC drivers, suggesting that 22q did not act independently in our set of individuals who developed PTC during young adulthood. Intriguingly, however, 22q deletions were strongly related to the driver pathway, occurring most commonly in *RAS*-mutated PTCs, suggesting that 22q deletion could provide a growth advantage or otherwise enhance the effect of certain MAPK drivers.

With our large sample size, we were able to explore the independent effects of radiation dose, age at PTC, age at exposure, and latency on PTC molecular characteristics. The pronounced evidence of radiation-related damage that we observed for individuals exposed at younger ages is consistent with epidemiologic analyses that have identified higher thyroid cancer risks with radiation exposure at younger ages (*17*). The relationship of a number of molecular characteristics, particularly total mutational burden and driver type, with age at PTC warrants further investigation across a broader age range (*10*). Additional research with detailed dose data is needed to understand whether our findings extend across a broader dose range, to other types of radiation, as well as to other tumor types, and whether radiation-related genomic characteristics have an impact on histopathological parameters (*46-48*). It has been hypothesized that ionizing radiation exposure could accelerate tumor development, and substantial evidence demonstrates that cancer survivors exposed to high-dose radiotherapy exhibit an aging phenotype (*36, 37*). However, exploratory analyses within our data did not support this hypothesis.

Our results have important implications for radiation protection and public health, particularly for low dose exposure, from two perspectives. First, the lack of a unique radiation-related pattern of molecular characteristics in the PTCs in our study, due in part to the random nature of ionizing radiation-related damage across the genome as well as the fact that other

mutagens can cause DNA double-strand breaks, suggests that we are yet to establish a reliable biomarker to distinguish tumors induced by radiation versus other causes. Nevertheless, the strong associations that we observed between radiation dose and molecular characteristics suggest that consideration of these factors could improve predictions of the probability that a specific thyroid tumor was caused by ^{131}I exposure (probability of causation [POC] (**Fig. S5**) (*49, 50*)), which is currently based on prior epidemiologic studies (*17*). Second, our data are consistent with a linear dose-response for the key molecular characteristics associated with radiation dose in the range examined in our analysis (≤1 Gy), which aligns with the extensive radiobiological literature and other epidemiologic evidence regarding DNA damage and cancer risk following ionizing radiation exposure (*51*, *52*).

Our study population included a substantial number of PTCs occurring after <100 mGy exposure, likely reflecting both the availability of remaining samples from the Chernobyl Tissue Bank as well as the increased detection of pre-existing PTC in the population that may not become clinically evident until later, if at all, due to intensive screening and heightened awareness of thyroid cancer risk in Ukraine. The increased genetic risk based on the PRS was notable among PTCs that occurred after lower doses despite limited statistical power to investigate germline genetic variants. The low overall mutational burden of early adulthood PTC, small sample sizes in certain population subgroups, and uncertainties in radiation dose estimates limited our statistical power to thoroughly investigate the shape of the dose-response curve, precisely identify the magnitude of radiation-related effects (as reflected by the wide confidence intervals for many effect estimates), or reliably characterize new radiation signatures.

In conclusion, we have characterized the genomic landscape of PTC, the most frequent cancer observed after the Chernobyl nuclear accident. Our results demonstrate a dose-dependent

carcinogenic effect of radiation derived primarily from DNA double-strand breaks repaired by NHEJ that initiate subsequent thyroid tumor growth, the patterns of which are shaped not by radiation exposure but rather by the specific driver gene. The consistency of the spectrum of PTC drivers in our study population compared with previous PTC series suggests that current therapeutic approaches for PTC are appropriate even for tumors that arise following radiation exposure (*53*). Our work provides a foundation for further investigation of radiation-induced cancer, particularly with respect to differences in risk as a function of both dose and age, and underscores the deleterious consequences of ionizing radiation exposure.

REFERENCES AND NOTES

1. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), "Sources and Effects of Ionizing Radiation: UNSCEAR 2008 Report to the General Assembly, with Scientific Annexes," (United Nations Publication, New York, 2010).
2. M. Tronko *et al.*, Thyroid neoplasia risk is increased nearly 30 years after the Chernobyl accident. *Int J Cancer* **141**, 1585-1588 (2017).
3. K. Furukawa *et al.*, Long-term trend of thyroid cancer risk among Japanese atomic-bomb survivors: 60 years after exposure. *Int J Cancer* **132**, 1222-1226 (2013).
4. L. H. Veiga *et al.*, Thyroid cancer after childhood exposure to external radiation: an updated pooled analysis of 12 studies. *Radiat Res* **185**, 473-484 (2016).
5. International Agency for Research on Cancer. (Lyon, France, 2012).
6. V. S. T. Goh *et al.*, Construction of fluorescence in situ hybridization (FISH) translocation dose-response calibration curve with multiple donor data sets using R, based on ISO 20046:2019 recommendations. *Int J Radiat Biol* **95**, 1668-1684 (2019).
7. Q. J. Liu *et al.*, Assessment of retrospective dose estimation, with fluorescence in situ hybridization (FISH), of six victims previously exposed to accidental ionizing radiation. *Mutat Res Genet Toxicol Environ Mutagen* **759**, 1-8 (2014).
8. S. Behjati *et al.*, Mutational signatures of ionizing radiation in second malignancies. *Nat Commun* **7**, 12605 (2016).
9. A. A. Efanov *et al.*, Investigation of the relationship between radiation dose and gene mutations and fusions in post-Chernobyl thyroid cancer. *J Natl Cancer Inst* **110**, 371-378 (2018).
10. Cancer Genome Atlas Research Network, Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159**, 676-690 (2014).
11. G. A. Thomas, The Chernobyl Tissue Bank: integrating research on radiation-induced thyroid cancer. *J Radiol Prot* **32**, N77-80 (2012).

12. G. A. Thomas, E. D. Williams, Thyroid tumor banks. *Science* **289**, 2283 (2000).
13. Materials and methods are available as supplementary materials at the Science website.
14. I. Likhtarov *et al.*, Thyroid cancer study among Ukrainian children exposed to radiation after the Chornobyl accident: improved estimates of the thyroid doses to the cohort members. *Health Phys* **106**, 370-396 (2014).
15. I. Likhtarov *et al.*, Reconstruction of individual thyroid doses to the Ukrainian subjects enrolled in the Chernobyl Tissue Bank. *Radiat Prot Dosimetry* **156**, 407-423 (2013).
16. I. Likhtarov *et al.*, Estimation of the thyroid doses for ukrainian children exposed in utero after the chernobyl accident. *Health Phys* **100**, 583-593 (2011).
17. E. Ron *et al.*, Thyroid cancer after exposure to external radiation: a pooled analysis of seven studies. *Radiat Res* **141**, 259-277 (1995).
18. A. Berrington de Gonzalez *et al.*, Epidemiological studies of low-dose ionizing radiation and cancer: rationale and framework for the monograph and overview of eligible studies. *J Natl Cancer Inst Monogr* **2020**, 97-113 (2020).
19. M. Hauptmann *et al.*, Epidemiological studies of low-dose ionizing radiation and cancer: summary bias assessment and meta-analysis. *J Natl Cancer Inst Monogr* **2020**, 188-200 (2020).
20. L. B. Alexandrov *et al.*, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
21. Y. E. Maruvka *et al.*, Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nat Biotechnol* **35**, 951-959 (2017).
22. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, M. R. Stratton, Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-259 (2013).
23. ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
24. R. K. Gopal *et al.*, Widespread chromosomal losses and mitochondrial DNA alterations as genetic drivers in Hurthle cell carcinoma. *Cancer Cell* **34**, 242-255 e245 (2018).
25. I. Ganly *et al.*, Integrated genomic analysis of Hurthle cell cancer reveals oncogenic drivers, recurrent mitochondrial mutations, and unique chromosomal landscapes. *Cancer Cell* **34**, 256-270 e255 (2018).
26. T. M. Lowe, S. R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
27. A. Lachmann *et al.*, ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438-2444 (2010).
28. J. Hess *et al.*, Gain of chromosome band 7q11 in papillary thyroid carcinomas of young patients is associated with exposure to low-dose irradiation. *Proc Natl Acad Sci U S A* **108**, 9595-9600 (2011).
29. M. Selmansberger *et al.*, CLIP2 as radiation biomarker in papillary thyroid carcinoma. *Oncogene* **34**, 3917-3925 (2015).
30. M. Selmansberger *et al.*, Dose-dependent expression of CLIP2 in post-Chernobyl papillary thyroid carcinomas. *Carcinogenesis* **36**, 748-756 (2015).

31. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
32. J. Gudmundsson *et al.*, A genome-wide association study yields five novel thyroid cancer risk loci. *Nat Commun* **8**, 14517 (2017).
33. H. H. Y. Chang, N. R. Pannunzio, N. Adachi, M. R. Lieber, Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol* **18**, 495-506 (2017).
34. J. Carvajal-Garcia *et al.*, Mechanistic basis for microhomology identification and genome scarring by polymerase theta. *Proc Natl Acad Sci U S A* **117**, 8476-8485 (2020).
35. B. Zhao, E. Rothenberg, D. A. Ramsden, M. R. Lieber, The molecular basis and disease relevance of non-homologous DNA end joining. *Nat Rev Mol Cell Biol* **21**, 765-781 (2020).
36. M. E. Sehl, J. E. Carroll, S. Horvath, J. E. Bower, The acute effects of adjuvant radiation and chemotherapy on peripheral blood epigenetic age in early stage breast cancer patients. *NPJ Breast Cancer* **6**, 23 (2020).
37. J. Tong, T. K. Hei, Aging and age-related health effects of ionizing radiation. *Radiat Med Protect* **1**, 15-23 (2020).
38. S. Horvath, DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).
39. G. Hannum *et al.*, Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* **49**, 359-367 (2013).
40. R. P. Tufano, G. V. Teixeira, J. Bishop, K. A. Carson, M. Xing, BRAF mutation in papillary thyroid cancer and its value in tailoring initial treatment: a systematic review and meta-analysis. *Medicine (Baltimore)* **91**, 274-286 (2012).
41. J. Liang *et al.*, Genetic landscape of papillary thyroid carcinoma in the Chinese population. *J Pathol* **244**, 215-226 (2018).
42. C. Turner *et al.*, Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Hum Genet* **112**, 303-309 (2003).
43. S. A. Roberts *et al.*, Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* **46**, 424-435 (2012).
44. J. Drost, H. Clevers, Organoids in cancer research. *Nat Rev Cancer* **18**, 407-418 (2018).
45. J. Drost *et al.*, Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234-238 (2017).
46. V. A. LiVolsi *et al.*, The Chernobyl thyroid cancer experience: pathology. *Clin Oncol (R Coll Radiol)* **23**, 261-267 (2011).
47. T. I. Bogdanova *et al.*, Papillary thyroid carcinoma in Ukraine after Chernobyl and in Japan after Fukushima: different histopathological scenarios. *Thyroid*, (2020).
48. E. D. Williams *et al.*, Thyroid carcinoma after Chernobyl latent period, morphology and aggressiveness. *Br J Cancer* **90**, 2219-2224 (2004).
49. D. C. Kocher *et al.*, Interactive RadioEpidemiological Program (IREP): a web-based tool for estimating probability of causation/assigned share of radiogenic cancers. *Health Phys* **95**, 119-147 (2008).

50. Interactive RadioEpidemiological Program Information - Version 5.7. Available at: https://radiationcalculators.cancer.gov/irep/. (Accessed November 26, 2019).
51. National Research Council, Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2. Washington, DC: The National Academies Press. https://doi.org/10.17226/11340. (2006).
52. S. Barnard, S. Bouffler, K. Rothkamm, The shape of the radiation dose response for DNA double-strand break induction and repair. *Genome Integr* **4**, 1 (2013).
53. M. Schlumberger, S. Leboulleux, Current practice in patients with differentiated thyroid cancer. *Nat Rev Endocrinol* **17**, 176-188 (2021).
54. V. A. Stezhko *et al.*, A cohort study of thyroid cancer and other thyroid diseases after the chornobyl accident: Objectives, design and methods. *Radiation Research* **161**, 481-492 (2004).
55. M. Hatch *et al.*, Thyroid cancer and benign nodules after exposure in utero to fallout from Chernobyl. *J Clin Endocrinol Metab* **104**, 41-48 (2019).
56. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **1303.3997v2**, (2013).
57. G. A. Van der Auwera *et al.*, From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-11 10 33 (2013).
58. J. Dai, https://doi.org/10.5281/zenodo.4543812. (2021).
59. J. Dai, https://doi.org/10.5281/zenodo.4543818. (2021).
60. J. P. Fortin *et al.*, Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* **15**, 503 (2014).
61. R. M. Cawthon, Telomere measurement by quantitative PCR. *Nucleic Acids Res* **30**, e47 (2002).
62. S. M. Gadalla *et al.*, Effect of recipient age and stem cell source on the association between donor telomere length and survival after allogeneic unrelated hematopoietic cell transplantation for severe aplastic anemia. *Biol Blood Marrow Transplant* **22**, 2276-2282 (2016).
63. C. Birger *et al.*, FireCloud, a scalable cloud-based platform for collaborative genome analysis: Strategies for reducing and controlling costs. *bioRxiv*, 209494 (2017).
64. K. Cibulskis *et al.*, Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219 (2013).
65. D. Benjamin *et al.*, Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 861054 (2019).
66. C. T. Saunders *et al.*, Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-1817 (2012).
67. S. Kim *et al.*, Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594 (2018).
68. J. A. Wala *et al.*, SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28**, 581-591 (2018).
69. K. Cibulskis *et al.*, ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601-2602 (2011).

70. P. Danecek *et al.*, The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
71. A. H. Ramos *et al.*, Oncotator: cancer variant annotation tool. *Hum Mutat* **36**, E2423-2429 (2015).
72. J. T. Robinson, H. Thorvaldsdottir, A. M. Wenger, A. Zehir, J. P. Mesirov, Variant Review with the Integrative Genomics Viewer. *Cancer Res* **77**, e31-e34 (2017).
73. A. Smit, R. Hubley, P. Green. (RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>).
74. M. Costello *et al.*, Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* **41**, e67 (2013).
75. D. A. Landau *et al.*, Mutations driving CLL and their evolution in progression and relapse. *Nature* **526**, 525-530 (2015).
76. K. Ellrott *et al.*, Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281 e277 (2018).
77. E. Rheinbay *et al.*, Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102-111 (2020).
78. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
79. A. Taylor-Weiner *et al.*, DeTiN: overcoming tumor-in-normal contamination. *Nat Methods* **15**, 531-534 (2018).
80. Y. Drier *et al.*, Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* **23**, 228-235 (2013).
81. Y. Li *et al.*, Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121 (2020).
82. X. Chen *et al.*, Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
83. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
84. S. L. Carter *et al.*, Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413-421 (2012).
85. S. W. Hartley, https://doi.org/10.5281/zenodo.4535549. (2021).
86. M. S. Lawrence *et al.*, Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
87. C. H. Mermel *et al.*, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
88. E. T. Dawson, https://doi.org/10.5281/zenodo.4539528. (2021).
89. E. T. Dawson, https://doi.org/10.5281/zenodo.4539568. (2021).
90. E. T. Dawson, https://doi.org/10.5281/zenodo.4539636. (2021).
91. J. Kim *et al.*, Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat Genet* **48**, 600-606 (2016).
92. M. P. Purdue *et al.*, Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. *Nat Genet* **43**, 60-65 (2011).

93. H. M. Cann *et al.*, A human genome diversity cell line panel. *Science* **296**, 261-262 (2002).
94. Y. Jin, A. A. Schaffer, M. Feolo, J. B. Holmes, B. L. Kattman, GRAF-pop: a fast distance-based method to infer subject ancestry from multiple genotype datasets without principal components analysis. *G3 (Bethesda)* **9**, 2447-2461 (2019).
95. A. L. Price *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
96. L. M. Morton *et al.*, Subsequent neoplasm risk associated with rare variants in DNA damage response and clinical radiation sensitivity syndrome genes in the Childhood Cancer Survivor Study. *JCO Precis Oncol* **4**, (2020).
97. M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
98. Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
99. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80-92 (2012).
100. M. J. Landrum *et al.*, ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014).
101. S. Liyanarachchi *et al.*, Assessing thyroid cancer risk using polygenic risk scores. *Proc Natl Acad Sci U S A* **117**, 5997-6002 (2020).
102. M. K. Iyer, A. M. Chinnaiyan, C. A. Maher, ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27**, 2903-2904 (2011).
103. M. Benelli *et al.*, Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* **28**, 3232-3239 (2012).
104. B. J. Haas *et al.*, STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*, 120295 (2017).
105. K. Wang *et al.*, MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, e178 (2010).
106. S. Kumar, A. D. Vo, F. Qin, H. Li, Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Sci Rep* **6**, 21597 (2016).
107. S. Liu *et al.*, Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res* **44**, e47 (2016).
108. M. K. Steinberg, https://doi.org/10.5281/zenodo.4544030. (2021).
109. Z. Lai *et al.*, VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44**, e108 (2016).
110. M. K. Steinberg, https://doi.org/10.5281/zenodo.4544031. (2021).
111. J. Dai, https://doi.org/10.5281/zenodo.4543808. (2021).
112. J. Dai, https://doi.org/10.5281/zenodo.4543824. (2021).
113. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B* **57**, 289-300 (1995).
114. M. J. Aryee *et al.*, Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).

115. Y. Tian *et al.*, ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33**, 3982-3984 (2017).
116. C. A. Pratilas *et al.*, (V600E)BRAF is associated with disabled feedback inhibition of RAF-MEK signaling and elevated transcriptional output of the pathway. *Proc Natl Acad Sci U S A* **106**, 4519-4524 (2009).
117. J. Dai, https://doi.org/10.5281/zenodo.4543809. (2021).
118. P. T. C. Group *et al.*, Genomic basis for RNA alterations in cancer. *Nature* **578**, 129-136 (2020).
119. S. Hanzelmann, R. Castelo, J. Guinney, GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
120. M. Imielinski, G. Guo, M. Meyerson, Insertions and deletions target lineage-defining genes in human cancers. *Cell* **168**, 460-472 e414 (2017).
121. F. Panebianco *et al.*, THADA fusion is a mechanism of IGF2BP3 activation and IGF1R signaling in thyroid cancer. *Proc Natl Acad Sci U S A* **114**, 2307-2312 (2017).
122. UNSCEAR, "Effects of ionizing radiation: United Nations Scientific Committee on the Effects of Atomic Radiation - UNSCEAR 2006 Report, Volume 1 - Report to the General Assembly, with Scientific Annexes A and B," (United Nations, New York, 2008).
123. D. Preston, J. Lubin, D. Pierce, M. McConney, N. Shilnikova, Epicure Risk Regression and Person-Year Computation Software: Command Summary and User Guide. Ottawa: Risk Sciences International; 2015.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the commitment of the staff of the Laboratory of Morphology of Endocrine System and the staff of the Department of Surgery of Endocrine System of IEM, who prepared the pathological material for the study and operated on the patients; the confirmation of thyroid tumor diagnoses provided by the International Pathology Panel of the Chernobyl Tissue Bank, including Professors A. Abrosimov, T. Bogdanova, G. Fadda, J. Hunt, M. Ito, V. Livolsi, J. Rosai, and E.D. Williams, and Dr. N. Dvinskyh; Nathan Appel (Information Management Services, Inc., Calverton, MD) for programming support; Elizabeth C. Sasse (Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute) for support in creating figures and editorial review of the manuscript; Dr. Clara Bodelon (Integrative Tumor Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute) for consultation regarding methylation

age calculations; Dr. Jia Liu (Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research) for consultation regarding analyses of mRNA expression and assessment of cross-platform concordance; Dr. Bin Zhu (Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute) for consultation regarding analyses of mRNA expression; Cameron Palmer (Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research) for assistance computing principal components for ancestry analyses; Timothy Myers (Laboratory of Genetic Susceptibility, Division of Cancer Epidemiology and Genetics, National Cancer Institute) for independent review of SV sequence context; and the leadership of The Cancer Genome Atlas for conduct of an initial pilot study. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The authors acknowledge alternative spellings of Chornobyl and Kiev. The opinions expressed by the authors are their own, and this material should not be interpreted as representing the official viewpoint of the U.S. Department of Health and Human Services, the National Institutes of Health, or the National Cancer Institute.

Funding: This project was supported by the Intramural Program of the National Cancer Institute, National Institutes of Health. The Chernobyl Tissue Bank is supported by the National Cancer Institute (U24CA082102).

Author contributions: The following authors conceptualized the study: L.M.M, D.M.K., C.S., T.I.B., E.T.D., A.V.B., A.B.d.G., G.A.T., M.D.T., G.G., and S.J.C. The following authors designed the study methodology: L.M.M, D.M.K., C.S., T.I.B., E.T.D., M.K.S., J.D., S.W.H., S.J.S., J.N.S., Y.M., V.K., D.A.R., J.C.G., C.M.P., J.S.P., M.Y., O.L., M.J.M., A.V.B., V.D., S.M., M.C., L.Y.Z., G.A.T., G.G., and S.J.C. The following authors analyzed and synthesized

the data: L.M.M, D.M.K., C.S., E.T.D., M.K.S., J.D., S.W.H., S.J.S., Y.M., V.K., M.Y., O.L., M.J.M., V.D., A.B.d.G., G.A.T., G.G., and S.J.C. The following authors collected or generated the study data: C.S., T.I.B., E.T.D., M.K.S., J.D., M.Y., J.F.B., A.H., B.D.H., C.L.D., J.M.G.F., J.B., E.K.C., S.M., M.C., L.Y.Z., G.A.T., and M.D.T. The following authors contributed resources (study materials, patients, computing resources, or other analysis tools): T.I.B., M.K.S., J.D., M.K., A.V.B., K.M., V.D., M.H., A.B.d.G., G.A.T., M.D.T., G.G., and S.J.C. The following authors curated the research data: L.M.M, D.M.K., C.S., T.I.B., E.T.D., V.K., M.K., A.H., C.L.D., J.M.G.F., and J.B. The following authors wrote the original draft: L.M.M, D.M.K., C.S., E.T.D., M.K.S., J.D., S.W.H., S.J.S., M.Y., G.A.T., G.G., and S.J.C. The following authors visualized the data: L.M.M, D.M.K., C.S., T.I.B., E.T.D., M.K.S., J.D., V.K., O.L., and G.G. The following authors supervised or managed the research: L.M.M, D.M.K., C.S., G.A.T., M.D.T., G.G., and S.J.C. The following authors had responsibility for managing and coordinating the research activity: L.M.M, D.M.K., C.S., T.I.B., E.T.D., E.K.C., A.B.d.G., G.A.T., M.D.T., G.G., and S.J.C. The following authors acquired funding for the study: G.A.T., M.D.T., and S.J.C. All authors edited the final manuscript.

Competing interests: E.T.D. is an employee of Nvidia Corporation and owns stock in Nvidia, Illumina, and Pacific Biosciences. G.G. receives research funds from IBM and Pharmacyclics, and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect, MSMutSig, MSIdetect, POLYSOLVER and TensorQTL. G.G. is a founder, consultant and holds privately held equity in Scorpion Therapeutics. All other authors declare no competing interests.

Data and materials availability: Analytic data are available in Data S1-S3. The Materials and Methods text specifies code that has been posted to GitHub and is archived on Zenodo (*13*). Raw

molecular data are available from the Genomic Data Commons, accessed through the database of Genotypes and Phenotypes (dbGaP, accession phs001134; https://www.ncbi.nlm.nih.gov/gap/).

SUPPLEMENTARY MATERIALS

Materials and Methods

Figs. S1 to S32

Tables S1 to S22

References 53 to 123

Data S1-S3 and accompanying data dictionaries

MDAR Reproducibility Checklist

FIGURE TITLES AND LEGENDS:

Fig. 1. Landscape of somatic alterations in 440 papillary thyroid carcinomas, by radiation dose from ^{131}I exposure. Blank (white) spaces represent unavailable data due to lack of data from a specific platform (**Figs. S1-S3**). Signature analyses were restricted to high purity samples, defined as those with tumor purity >20% and no evidence of tumor contamination in the normal tissue.

Fig. 2. Relationship between radiation dose from ^{131}I exposure and small deletions. (**A**) Total small deletion count and restricted to (**B**) clonal and (**C**) subclonal small deletions. (**D**) Total deletion:SNV ratio and restricted to (**E**) clonal and (**F**) subclonal deletions and SNVs. (**G**) Total ID5 count and restricted to (**H**) clonal and (**I**) subclonal ID5. (**J**) Total ID8 count and restricted to (**K**) clonal and (**L**) subclonal ID8. β per 100 mGy and P-value were derived from

multivariable linear regression models adjusting for age at PTC and sex. Full model results are provided in Table S18.

Fig. 3. Relationship between radiation dose from ^{131}I exposure and selected SV metrics. (**A**) Number of simple/balanced SVs. (**B**) Likelihood of having a fusion versus mutation driver. (**C**) Number of clonal ≥5 bp EJ small deletions. (**D**) Number of confirmed clonal simple/balanced/end-joining SVs. (**E**) Number of confirmed clonal other SVs (E). Different scales are used for each panel to reflect the distributions and uncertainties of the EOR estimates. Referent group for categorical analyses: EOR=0 (which is equivalent to OR=1). EOR per 100 mGy and P-value were derived from multivariable proportional odds or logistic regression models adjusting for age at PTC and sex. Full model results are provided in Table S18.

Fig. 4. Distribution of radiation dose from ^{131}I exposure by driver type and pathway.

Fig. 5. Selected RNA-seq results. (**A**) Differential expression by driver and cluster, (**B**) Differential expression for all genes by radiation dose from ^{131}I exposure. (**C**) Differential expression of *CLIP2* by radiation dose from ^{131}I exposure.

Fig. 6. Relationship between radiation dose from ^{131}I exposure and PRS. Data for the 12 single nucleotide polymorphisms that comprise the PRS are provided in Table S15.

Fig. 7. Relationship between radiation dose from ^{131}I exposure and selected genomic characteristics by age at exposure. Clonal deletion:SNV ratio for (**A**) <5 years at exposure, (**B**) 5-9 years at exposure, and (**C**) ≥10 years at exposure. Number of clonal ID8 mtuations for (**D**) <5 years at exposure, (**E**) 5-9 years at exposure, and (**F**) ≥10 years at exposure. Number of clonal ≥5 bp EJ small deletions for (**G**) <5 years at exposure, (**H**) 5-9 years at exposure, and (**I**) ≥10 years at exposure. Likelihood of having a fusion versus mutation driver for (**J**) <10 years at exposure and (**K**) ≥10 years at exposure. All analyses exclude ^{131}I-unexposed individuals. β or EOR per 100 mGy and P-value were derived from multivariable linear, proportional odds, or logistic regression models adjusting for age at PTC and sex. Full model results are provided in Table S22.

* Models evaluating the effect of dose on driver type restricted to <5 years of age at exposure did not converge, so individuals exposed at <5 and 5-9 years were combined in panel J. EOR/100 mGy for 5-9 years alone=1.78, 95%CI=0.12-226.

^^ in Panel H indicates that the upper 95% CI exceeds the y-axis maximum value.

Supplementary Materials for

Molecular characterization of papillary thyroid cancer in relation to ionizing radiation dose following the Chernobyl accident

Lindsay M. Morton[*], Danielle M. Karyadi[*], Chip Stewart[*], Tetiana I. Bogdanova[†], Eric T. Dawson[†], Mia K. Steinberg[†], Jieqiong Dai, Stephen W. Hartley, Sara J. Schonfeld, Joshua N. Sampson, Yosi Maruvka, Vidushi Kapoor, Dale A. Ramsden, Juan Carvajal-Garcia, Chuck M. Perou, Joel S. Parker, Marko Krznaric, Meredith Yeager, Joseph F. Boland, Amy Hutchinson, Belynda D. Hicks, Casey L. Dagnall, Julie M. Gastier-Foster, Jay Bowen, Olivia Lee, Mitchell J. Machiela, Elizabeth K. Cahoon, Alina V. Brenner, Kiyohiko Mabuchi, Vladimir Drozdovitch, Sergii Masiuk, Mykola Chepurny, Liudmyla Yu. Zurnadzhy, Maureen Hatch, Amy Berrington de Gonzalez, Gerry A. Thomas[‡], Mykola D. Tronko[‡], Gad Getz[‡], Stephen J. Chanock[‡]

[*]These authors contributed equally to this work.
[†]These authors contributed equally to this work.
[‡]These authors jointly directed this work.

Correspondence to: chanocks@mail.nih.gov, mortonli@mail.nih.gov

This PDF file includes:

Materials and Methods
Figs. S1 to S32
Tables S1 to S12, S15-S16, S18-S22
Captions for Tables S13, S14, and S17
Captions for Data S1 to S3

Other Supplementary Material for this manuscript includes the following:

Tables S13, S14, S17 (Microsoft Excel)
Data S1 to S3 (text files) and accompanying data dictionaries (PDF)
MDAR Reproducibility Checklist (PDF)

Materials and Methods

Study Populations

The primary study population for this analysis included individuals who developed pathologically-confirmed papillary thyroid cancer (PTC) during young adulthood in the Ukraine following the April 1986 Chernobyl nuclear power plant accident and whose pre-treatment, fresh frozen tumor sample was collected by the Chernobyl Tissue Bank (CTB) (*11*) (**Table S1**). Eligibility criteria included: *in utero* or <19 years of age on April 26, 1986 (cases) or born >9 months after the accident (comparison population of unexposed individuals); residence in one of the most contaminated territories (oblasts or states) of Ukraine, specifically, Zhytomyr, Kyiv, or Chernihiv; histopathologically confirmed diagnosis of a first primary papillary thyroid cancer based on review of tumor tissue by an international panel of experts and medical record data documenting no prior cancer history (*11*); availability of ^{131}I dose estimates (see details below); availability of fresh frozen thyroid tissue with sufficient, high quality nucleic acids (see details below); and availability of non-tumor thyroid tissue and/or a blood sample (collected years later). **Fig. S1** provides the distribution of available tumor tissue, non-tumor thyroid tissue, and blood in the final dataset of 440 individuals with PTC (n=359 ^{131}I-exposed, n=81 unexposed). Participants (or guardians for minors) provided informed consent for donation and broad research use of their materials through the CTB, and the study was approved by Institutional Review Boards at the tissue collection center (Institute of Endocrinology in Kiev, Ukraine), the CTB coordination center (Imperial College of London, UK), and the United States National Cancer Institute.

Among the 440 individuals with PTC included in our final analysis, the majority (n=365, 83.0%) underwent a total thyroidectomy and had lesions <2.0 cm (n=247, 56.1%) (**Table S1**). Among those with available non-tumor thyroid tissue that passed quality control assessment (n=416), the sample was taken from the contralateral lobe/side for most individuals (n=359, 86.3%). Non-tumor thyroid tissue from the same (ipsilateral) side (n=56, 13.5%) was taken as far from the tumor as possible. To our knowledge, no study participants had concurrent thyroid disease.

For investigation of the relationship between radiation dose and the PTC driver type (described below), we also included data from 45 non-overlapping individuals with PTC occurring after the Chernobyl accident whose PTC drivers were identified in a previously-published RNA sequencing (mRNA-seq) analysis (**Table S1**) (*9*).

^{131}I Dose Estimation

For ^{131}I-exposed individuals with PTC (individuals who were *in utero* or <19 years of age on April 26, 1986), radiation doses to the thyroid were reconstructed by an international team of dosimetry experts using well-established methods that have been described in detail previously for studies of Chernobyl-related cancer risks (*14*, *15*). Briefly, the dose reconstruction approach varied based on the type of information available for each individual: presence or absence of direct thyroid radioactivity measurements, presence or absence of personal interview, and information on place of residence (oblast/state, raion/county, settlement) in 1986. Among the 359 ^{131}I-exposed individuals included in the primary analytic population, 49 were part of a large-scale thyroid cancer screening cohort (*54*), and their doses were estimated based on individual thyroid radioactivity measurements taken in May-June 1986, personal interview data concerning residential history and intake of milk and green leafy vegetables (the main source of ^{131}I exposure) after the accident, and results of radio-ecological modeling (**Table S1**). For an

additional 4 individuals, dose estimates were calculated based on individual thyroid radioactivity measurements, but not on personal interview. Dose estimates for the largest group of individuals (n=297) were based on measurements made on different individuals who lived in the same residential area. Finally, for the 9 individuals who were *in utero* at the time of the accident, estimated doses were based on the mother's exposure (*55*). Among the 45 ^{131}I-exposed individuals included in analyses of the oncogenic PTC driver from the previously published report, doses were estimated based on direct thyroid radioactivity measurement with personal interview data for 39, and the dose estimation method was unknown for the remaining 6 individuals (**Table S1**) (*9*).

Laboratory Methods

Sample Handling and Nucleic Acid Extraction

Fig. S1 provides a schematic of the study procedures and sample sizes for laboratory work and statistical analyses. Frozen tissue specimens (535 PTC, 522 non-tumor thyroid tissue) were received from CTB and processed by Nationwide Children's Hospital (NCH) Biospecimen Core Resource (BCR) (Columbus, Ohio). A 5 μm section of tissue was cut from the provided block for creation of an H&E slide for pathology review. Specimens were forwarded for total nucleic acid extraction if the tissue necrosis <20% and if the PTC tissue >50% tumor nuclei with matching non-tumor thyroid tissue with <10% tumor nuclei and/or blood available. Specimens for 72 individuals did not pass this inclusion criteria based on 66 PTC failures and 6 non-tumor thyroid failures without blood available, resulting in the removal of 72 PTC, 71 non-tumor thyroid, and 21 blood specimens. A two-column approach was utilized for extraction of DNA and RNA from approximately 10-40 mg of tissue, depending on tissue availability. Tissue was homogenized using the Qiagen Qiagen TissueLyser II. Dual DNA and RNA Extraction was performed utilizing the AllPrep DNA/RNA Mini Kit (Qiagen) for DNA and mirVana miRNA Isolation Kit (Applied Biosystems) for total RNA and small RNA. Purified nucleic acids were quantified and quality checked post-extraction. RNA was tested utilizing the Agilent 2100 Bioanalyzer with the RNA 6000 NanoChip (Agilent). DNA was quantified with the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher). DNA molecular weight was evaluated by E-Gel 48 Agarose Gels, 1% (ThermoFisher). Tumor and normal tissue DNA samples were identity matched via SNP analysis MassARRAY system and iPLEX chemistry (Sequenom) across 50 loci.

Whole blood samples (n=326) collected in standard EDTA vacutainers were received from CTB and processed at the Cancer Genomics Research Laboratory (CGR) of the National Cancer Institute. Blood samples were extracted on the QIAsymphony SP® instrument utilizing manufacturer-supplied reagents and protocols (QIAGEN). Double-stranded DNA was quantified with QuantiFluor® dsDNA System (Promega Corporation) for use in downstream assays.

Sample quality and quantity assessment and final inclusion criteria

DNA/RNA yield and the RNA Integrity Number (RIN; Agilent Bioanalyzer) identified that 11 individuals had insufficient yield and/or RNA quality for WGS or RNA-seq analysis and were excluded. Data from the Illumina Infinium OmniExpress SNP array and/or mRNA-seq were used as preliminary assessment of sample quality and to check for sex and/or intra-individual discordance, which identified 12 individuals for exclusion (11 for discordance among matched samples and one for DNA/RNA tumor contamination). The non-tumor thyroid tissue pathology

was reassessed and only tissues with no evidence for tumor nuclei were processed further. Following these exclusions, 440 individuals with 440 PTC (23 excluded), 416 non-tumor thyroid (35 excluded), and 290 blood (15 excluded) samples remained for further laboratory processing.

All available tissue types (PTC, non-tumor thyroid, and blood) that passed the above quality control metrics were attempted on all six platforms (WGS, mRNA-seq, miRNA-seq, SNP array, methylation, and relative telomere length) when possible. PTC WGS required a paired normal tissue sample (non-tumor thyroid or blood) for analysis, thus PTC samples were excluded from the WGS platform if no normal tissue was available. Similarly, NT WGS (non-tumor thyroid tissue) required both a non-tumor thyroid tissue and blood sample. Sufficient DNA/RNA was available to attempt all six platforms for 369/440 PTC samples, 352/416 non-tumor thyroid tissues, and 232/290 blood samples, whereas only a subset of platforms was attempted for the remaining 71/440 PTC samples, 64/416 non-tumor thyroid tissues, and 58/290 blood samples. **Fig. S1** provides further details regarding the number of attempted assays, the number of failed assays per platform, and the final number of samples per platform by sample type.

Whole Genome Sequencing

All WGS library preparation and sequencing was done at the Broad Institute. Libraries were constructed and sequenced on the Illumina HiSeqX with the use of 151 base pair (bp) paired-end reads for whole-genome sequencing. Output from Illumina software was processed by the Picard data-processing pipeline to yield BAM files containing well-calibrated, aligned reads. All sample information tracking was performed by automated LIMS messaging.

WGS Library Construction

Initial genomic DNA input for shearing was reduced from 3 μg to 350ng in 50 μL of solution. In addition, for adapter ligation, Illumina paired end adapters were replaced with palindromic forked adapters with unique 8 base index sequences embedded within the adapter.

Library preparation and sequencing was done in two batches with slightly modified protocols. The first 919 samples were sequenced with the following library preparation protocol: Aliquots of genomic DNA (350ng in 50 μL) underwent fragmentation by means of acoustic shearing using Covaris focused-ultrasonicator, targeting 385 bp fragments. Following fragmentation, additional size selection was performed using a SPRI cleanup. Library preparation was performed using a commercially available kit provided by KAPA Biosystems (product KK8202, name: KAPALIBPREPKT) and with palindromic forked adapters with unique 8 base index sequences embedded within the adapter (purchased from IDT). Following sample preparation, libraries were quantified using quantitative PCR (kit purchased from KAPA biosystems) with probes specific to the ends of the adapters. This assay was automated using the Agilent Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 1.7nM. Samples were then pooled into 24-plexes, and the pools were once again qPCRed. Samples were combined with HiSeq X Cluster Amp Reagents EPX1, EPX2 and EPX3 into single wells on a strip tube using the Hamilton Starlet Liquid Handling system. Cluster amplification of the templates was performed according to the manufacturer's protocol (Illumina) using the Illumina cBot. Flowcells were sequenced either on HiSeqX Sequencing-by- Synthesis Kits to produce 151 bp paired-end reads, then analyzed using RTA2. Output from Illumina software was processed

by the Picard data-processing pipeline to yield CRAM or BAM files containing demultiplexed, aggregated aligned reads.

The final 224 samples were sequenced with the following minor changes in protocol: Library preparation was performed using a commercially available kit provided by KAPA Biosystems (KAPA Hyper Prep without amplification module, product KK8505), and with palindromic forked adapters with unique 8-base index sequences embedded within the adapter (purchased from Roche). Following sample preparation and qPCR quantification, libraries were normalized to 2.2nM and pooled into 24-plexes. All other aspects of the protocol remained the same as described in the previous paragraph.

Sequencing data (FASTQ files) were aligned to the reference genome (hg19) using BWA-MEM (*56*). Duplicated fragments of DNA were marked to avoid 'double counting' evidence for somatic events using Picard MarkDuplicates (version 1.1310 broadinstitute.github.io/picard) and re-calibrated the base quality scores to more accurately reflect the probability of error using BQSR (broadinstitute.github.io/picard). Since alignments around indels can be different among reads, which affects downstream indel and mutation calling, we performed local realignment on both the tumor and normal for each patient ("co-cleaning" GATK3 IndelRealigner version nightly-2015-07-31-g3c929b0) (*57*). At the end of this step, we obtain hg19 binary alignment files (BAM files) for each of the tumor and normal samples.

Each WGS sample was assessed for coverage (**Table S2**), library complexity, fingerprinting across tumor and normal samples, sequencing error rates, fragment length, chimeric fragment rate, and DNA oxidative damage using a suite of Picard tools (Picard CollectWGSmetrics, CollectSequencingArtifactMetrics for 8-oxoG damage, and CrosscheckFingerprints).

RNA Library Preparation and Sequencing
Quality control was performed on total RNA using the Qubit BR RNA assay (ThermoFisher Scientific) and Agilent RNA 6000 Pico Kit (Agilent Technologies). 100-500 ng of total RNA in 10 µl was used as input. Libraries were prepared using the Kapa RNA HyperPrep Kit with RiboErase per manufacturer's instructions with slight modifications (Kapa Biosystems, part# KK8561). Briefly, following depletion of ribosomal RNA, fragmentation was performed at 85ºC for 4.5 minutes for good quality samples, 85ºC for 4 minutes for average quality, and 65ºC for 1-1.5 min for the lowest quality and/or lowest input samples. Following first and second strand cDNA synthesis, 1.5 µM NEXTflex-96 DNA Barcodes (BioO Scientific) were ligated to the A-tailed cDNA. Eight cycles of library amplification were performed, followed by two rounds of 1x Ampure purification. Final libraries were quantified by Qubit, equal molar pooled at 16 samples per pool and sequenced on the Illumina HiSeq 2500 using the SBS 2x125 bp and Cluster v4 kits. Runs generated an average of 125,000,000 reads/sample, with 44,000 transcripts and 19,000 genes detected. The sequenced paired-end reads were trimmed using trimmomatic/0.36. Trimmed reads were aligned to the GRCh38 human reference genome (illumine iGenomes NCBI GRCh38) using STAR/2.5.4a, and gene reads count was quantified according to illumine iGenomes NCBI GRCh38 annotation file.

For miRNA sequencing, the Illumina Ribo-Zero Gold rRNA Removal Kit (Illumina, part# MRZG12324) was used for ribosomal RNA depletion from 500 ng purified total RNA per

manufacturer's instructions. 500ng of the rRNA depleted sample was then processed as described in the NEBNext Multiplex Small RNA Library Prep Set for Illumina kit (NEB, part #E7300L) with minor modifications. Following first strand synthesis, Illumina indexes were added using 12 cycles of PCR amplification, the products purified with 1.8x Ampure XP beads and recovered in 15 µL low TE buffer. Size selection of the miRNA libraries was performed on Blue Pippin using 3% Agarose, dye free gel with internal standards selecting for 105-190 bp products (Sage Sciences, part# BDQ3010). Quantitation and quality control check was performed using Qubit dsDNA HS assay and on Agilent 2100 Bioanalyzer system with the Agilent HS DNA assay. Libraries were equal molar pooled at 16 samples per pool and sequenced on the Illumina HiSeq 2500 using the 50 cycle single read SBS and Cluster v4 kits according to the manufacturer's instructions. The sequenced single-end reads were processed according to the ENCODE microRNA-seq pipeline (https://www.encodeproject.org/microrna/microrna-seq/#references). Briefly, reads were trimmed using Cutadapt/1.18, and after adaptor trimming and quality filtering only those with lengths of 15-31 nt were kept. Trimmed reads were aligned to the GRCh38 human reference genome (illumine iGenomes NCBI GRCh38) using STAR/2.5.4a, and microRNA reads count was quantified according to GENCODE V24 genome annotation file which was the microRNA subset from comprehensive GENCODE annotations.

FASTQC/0.11.5 was used for the quality control analysis of post-trimmed RNA-seq and miRNA-seq reads. Samples that received warning or fail messages on the analysis of Mean Quality Scores, per sequence quality scores, per base N content, or adapter content were filtered. STAR alignment scores were used for alignment quality control analysis. Samples over 60% of any type of unmapped reads were filtered. Code for RNA-seq quality control and alignment is available at: https://github.com/NCI-CGR/ChernobylThyroidCancer-RNAseq and https://github.com/NCI-CGR/ChernobylThyroidCancer-miRNAseq and archived on Zenodo (*58, 59*).

Illumina Infinium OmniExpress Single Nucleotide Polymorphism (SNP) Arrays
High-throughput, genome-wide SNP genotyping, using Infinium BeadChip technology (Illumina Inc.), was performed at the Cancer Genomics Research Laboratory (CGR) of the National Cancer Institute. Genotyping was performed per the manufacturer's guidelines using the Infinium automated HTS Assay protocol. Briefly, 200 ng genomic DNA, quantitated using Quant-iT PicoGreen dsDNA Reagent (ThermoFisher Scientific) was denatured and neutralized, then isothermally amplified by whole-genome amplification. The amplified product was enzymatically fragmented, then precipitated and re-suspended. Resuspended samples were denatured, then hybridized to locus-specific 50-mer oligonucleotides, which were attached to 1-micron beads on the BeadChip. These 50-mer probes stopped one base before the location of interest. Enzymatic single-base extension of the oligos on the BeadChip, using the captured DNA as a template, incorporated tagged nucleotides on the BeadChip, which were subsequently fluorophore-labeled during staining. The fluorescent label determined the genotype call for the sample. The Illumina iScan scanned the BeadChips at two wavelengths to detect the fluorescent label, creating image files that were converted into genotype calls based on the detected fluorescence.

All samples in this study were scanned on the Infinium HumanOmniExpress-24 (v1.1, v1.2) array using the standard Illumina microarray data analysis workflow

(https://www.illumina.com/documents/products/technotes/technote_array_analysis_workflows.pdf) and data quality check was performed using PLINK (v.1.90b5.4). A two-stage filter by completion rate threshold of >0.8 for samples and >0.8 for loci, followed by >0.95 for samples and >0.95 for loci was performed. Sample contamination was examined by running the tool VerifyIDintensity(https://genome.sph.umich.edu/wiki/VerifyIDintensity) on each sample that passed completion rate filters or had a median raw intensity >6000. Additionally, sex verification, autosomal heterozygosity, and genotype concordance checks across all samples were performed to identify any problematic samples.

Methylation EPIC Array

400 ng of sample DNA, according to Quant-iT PicoGreen dsDNA quantitation (ThermoFisher Scientific), was treated with sodium bisulfite using the EZ-96 DNA Methylation MagPrep Kit (Zymo Research) according to manufacturer-provided protocol. Bisulfite conversion modifies non-methylated cytosines into uracil, leaving 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) unchanged. For every 95 samples, an internal control, NA07057 (Coriell Cell Repositories), was utilized to confirm the efficiency of bisulfite conversion during subsequent methylation analysis.

High-throughput epigenome-wide methylation analysis, using the Infinium MethylationEPIC BeadChip (Illumina Inc.), which uses both Infinium I and II assay chemistry technologies, was performed according to manufacturer's protocol. Bisulfite-treated samples were denatured and neutralized and whole genome amplified isothermally to increase the amount of DNA template. The amplified product was enzymatically fragmented, precipitated and resuspended in hybridization buffer. Twelve samples were applied to each BeadChip and hybridized overnight where fragmented DNA samples anneal to locus-specific 50mers (covalently linked to bead types for more than 850,000 methylation sites). Two bead types correspond to each CpG locus for Infinium I assays: one bead type corresponds to methylated, another bead type to the unmethylated state of the CpG site, while one bead type corresponds to each CpG locus (both methylated and unmethylated) for Infinium II assays. Single-base extension of the oligos on the BeadChip, using the captured DNA as template, incorporates tagged nucleotides on the BeadChip, which are subsequently fluorophore-labeled during staining. The Illumina iScan scanned the BeadChips at two wavelengths to create image and intensity files.

The intensity files from the Illumina methylation assay on the MethylationEPIC BeadChip were processed and analyzed with the R programming language using the R package "minfi". Briefly, raw intensity file (idats) were loaded into R using minfi. Samples were excluded if the percent of probes with detection p value >0.01 was greater than 4%, where the detection p-value=1-p-value as computed from the background model characterizing the chance that the target sequence signal was distinguishable from the negative controls. Concordance was checked for both expected and unexpected replicates using the ~60 polymorphic SNPs on the array. Raw methylation beta values were normalized according to previously published methods (*60*).

Relative Telomere Length Measurement

Relative telomere length measurement was performed on DNA using a qPCR assay adapted from Cawthon's originally published protocol which is briefly summarized below (*61, 62*). Relative telomere length determination by qPCR measured the ratio of telomere (T) signals,

specific to the telomere hexamer repeat sequence TTAGGG, to autosomal single copy gene (S) signals. This ratio was standardized by internal control DNA samples to yield relative standardized T/S ratios proportional to average telomere length. In this technique, reactions for each assay were performed independently, in triplicate, so a standard curve of pooled gDNA samples was utilized to assess the amount of each signal, while compensating for inter-plate variations in PCR efficiency.

In this dataset, the mean coefficient of variation (CV) of triplicates for the telomere assay was 0.59% and for the 36B4 assay (single copy gene, S) was 0.32%. The CV for the internal controls (n=85) standardized T/S ratio across all plates was 5.14%. In this dataset, the intraclass correlation coefficient (ICC) was 0.958 (95% confidence interval: 0.947, 0.967). The mean CV for the standardized T/S ratio of technical replicates was 4.9% based on 61 subjects.

Bioinformatic Analysis

The bulk of WGS processing was done within the Terra workflow framework (Terra.bio) (*63*). All workflows described below are available from a public workspace REBC_methods_only. Some of the input parameters for the workflows (particularly the Panels-of-Normals [PoN] - used for mutation, structural variation, and copy number detection) are protected data which requires dbGaP application (accession #phs001134 / PRJNA324143) approval for access. The same pipeline as described below was run on the cohort of WGS tumor-normal pairs as well as the WGS non-tumor thyroid tissue-blood normal pairs. Additionally, the pipeline was run on 42 individuals without known radiation exposure who had available tumor-normal paired WGS data from The Cancer Genome Atlas (TCGA) data set according to current quality control criteria (*10*). Note that the subset of individuals selected for WGS from the full TCGA analysis of PTC were more likely to have an unidentified driver (*10*) due to sampling of WGS for this reason.

Mutation Detection, Merging, and Filtering Pipeline

A consensus calling approach for somatic mutations combined evidence from multiple detection algorithms (**Fig. S6**). Single nucleotide variants (SNVs) were detected by MuTect 1.0 version GATK3 v1.1.6 (*64*) (Terra task wgs_pip_m1_fragcounter_oxoq_1), MuTect2.0 version GATK3 "3.6-97-g881c5e9" (*65*) (Terra task wgs_pip_m2_64core), Strelka 1 version 1.0.11 (*66*) (Terra task strelka), and Strelka 2 version 2.8.3 (*67*) (Terra task strelka2), all with default parameters except as specified in Terra workspace REBC_methods_only. Insertion/deletion variants (INDELs) were detected by MuTect2.0, Strelka 1, and Strelka 2, as well as PCAWG_snowman version 1.0 (*68*) (Terra task pipette_wgs_SV), and SvABA version 134 (*68*) (Terra task SvABA_xtramem) all with default parameters except as specified in Terra workspace REBC_methods_only.

To account for algorithms that were based on similar approaches and/or development teams (Strelka version 1 and Strelka version 2, PCAWG snowman and SvABA), a consensus "voting" scheme was used to select the final calls. Specifically, the consensus voting scheme (Terra task rebc_consensus_maf) gave MuTect 1.0 and 2.0 one vote each because of their different approaches to mutation calling, whereas Strelka 1 and 2 were combined to provide one vote if either algorithm detected an event, and PCAWG snowman and SvABA were combined to provide one vote if either algorithm detected an event. The post-process consensus filter required at least 2 votes from all the algorithms and was applied following the artifact filters described

below (**Fig. S7**). Mutations passing all filter criteria were included in the next layer of mutation analysis (MutSig2CV, SigProfiler signature analysis, ABSOLUTE purity and ploidy estimation, clonality estimation, and association analyses).

The first computational step in MuTect (1.0 and 2.0) was to estimate sample DNA contamination from other individuals (ContEst version Queue-1.4-437-g6b8a9e1-svn-35362.jar (*69*), Terra task wgs_pip_m1_fragcounter_oxoq_1) which was used internally within MuTect to set somatic evidence thresholds. The ContEst tumor sample contamination estimate was also used by a post-process filter (described below) on the consensus mutation calls so that algorithms other than MuTect were also desensitized to tumor sample contamination. Mutations that passed somatic mutation criteria for each algorithm were converted from their native format (usually VCF format (*70*)) to a common "maf-lite" format (*71*) (Terra task maflite_merge_maf_workflow) so that algorithms detecting a given variant could be merged into a common format before being annotated and converted to the standard MAF format by oncotator (*71*).

Among the decisions imposed by the merging algorithm, multiple algorithms often have differences in allele counts for the same variant. This was resolved by ordering the algorithms (MuTect 1, MuTect 2, PCAWG Snowman, Strelka 1, Strelka 2, SvABA) such that the first algorithm in the ordered list with a call at the specific allele would be the algorithm to provide the merged allele counts. Indels were merged when the indel start occurred within a window of three bases. The merged mutations were then filtered (Terra task contest_oxoG_PoN_blat_Filter_Workflow) for possible sources of artifacts including 8-oxoG damage, PoN, local realignment around candidate mutations, tumor sample contamination, a hard threshold against more than one alternate supporting read in the normal, consensus among algorithms (**Fig. S7**). At each stage in the analysis a large fraction of the mutation candidates were manually reviewed in IGV (*72*), particularly those found to be outliers in terms of overlap with repeat regions (*73*), coverage spikes (>200x), low evidence (number of ALT reads < 5), or evidence of the ALT allele a normal (either by the matched normal or the PoN). Conclusions from manual review were applied to the list of candidate mutation in terms of modifications to the filter rules uniformly applied to all candidate mutations rather than as separate manual review judgements for particular mutations. The modification of rules included filter threshold adjustments for sample contamination, number of ALT alleles in the normal, and the consensus voting (described below).

OxoG filter: The deToxoG filter (*74*) was used to remove candidate SNV that were likely to arise from 8-oxoG DNA damage in library prep. The OxoG filter was included in the Terra task contest_oxoG_PoN_blat_Filter_Workflow. Briefly, the deToxoG filter is based on the balance of alternate allele counts of two possible fragment "orientations" (F1R2 and F2R1 in which read 1 of a pair is mapped forward - F1R2 - or reverse - F2R1). The artifact null model expects that the F1R2 and F2R1 should be equally balanced while mutations arising from 8-oxoG damage are highly skewed (~96%) in favor of F1R2 for G>T and F2R1 for C>A mutations mapped to the reference. The filter calculates a p-value that a given SNV is balanced, and any mutation with a Benjamini-Hochberg FDR less than 1% was removed, which is designed to leave a pool of candidate mutations with less than 1% arising from 8-oxoG damage.

PoN filters: There were several rounds of PoN filters applied. Several algorithms had their own internal PoN filters (eg. MuTect1, MuTect2, SvABA) which were either in VCF or bed format and were used as a blacklist to remove common germline variants or sites of artifacts. The merged MAF was subsequently filtered based on "Token PoN" evidence. The Token PoN filter method has been used in many previous cancer projects (*10, 74-77*), the only difference being that one of the PoNs was specifically constructed from the CTB WGS cohort. A Token PoN consists of counts from each sample in the PoN at each site of the genomic classified into 8 bins according to the reference and non-reference coverage with Variant Allele Fractions (VAF) of the sample at that site. The bins for each site were defined as:

1. Count samples with sequencing depth < 8x
2. Count samples with sequencing depth ≥ 8x, and not included in bins 3-8
3. Count samples with depth ≥ 8x, alternate alleles > 1, and 0.1% < VAF ≤ 0.3%
4. Count samples with depth ≥ 8x, alternate alleles > 2, and 0.3% < VAF ≤ 1.0%
5. Count samples with depth ≥ 8x, alternate alleles > 3, and 1.0% < VAF ≤ 3.0%
6. Count samples with depth ≥ 8x, alternate alleles > 3, and 3.0% < VAF ≤ 20.0%
7. Count samples with depth ≥ 8x, 3 ≤ alternate alleles ≤ 10, and VAF > 20.0%
8. Count samples with depth ≥ 8x, alternate alleles >10, and VAF > 20.0%

Counts of samples in these bins were used to approximate a PoN likelihood vector for each site in VAF bins 0-0.1%, 0.1-0.3%, 0.3-1%, 1-3%, 3-20%, and 20%-100%. For each candidate mutation a corresponding data likelihood was constructed based on the beta distribution of VAF for the observed alternate and reference counts of supporting reads. The joint likelihood between the PoN and data is the bin-by-bin sum of the product of the two likelihood vectors. A candidate mutation fails the token PoN filter when log10(PoN joint likelihood) > -2.5, a threshold which was empirically determined from this and previous projects with combinations of validation data and manual review.

There were two independent PoNs used for mutation filtering included in the Terra task contest_oxoG_PoN_blat_Filter_Workflow. One PoN was constructed from 248 normal blood samples with TIn=0 (see "deTiN" below) from this cohort (dbGaP accession # phs001134 / PRJNA324143). The other token PoN was derived from PCAWG normal blood samples (dbGaP accession # phs000178.v1.p1, EGA accession # EGAS00001001692). Each PoN was used as a statistically separate token PoN filter, although in most cases if a mutation failed one of the PoN filters it would also fail the other.

Realignment filter: Alternate reads from each candidate mutation were realigned to the hg19 reference using BLAT (*78*). This filter was part of the Terra task contest_oxoG_PoN_blat_Filter_Workflow. Any mutations with 3 or more reads with an alignment score to another mapped position greater than or equal to the alignment score at the position supporting the alternate allele were rejected.

Contamination filter: The presence of germline mutations from another individual DNA contaminating the tumor sample can lead to false detection. MuTect 1.0 and MuTect 2.0 explicitly take this into account in their statistical model detection threshold, but Strelka (1 or 2), SvABA and PCAWG_snowman do not, so a post-process contamination filter (Terra task

postFilter_contEst) was applied to the merged set of mutation calls. The contamination filter calculates the VAF beta cumulative distribution for each mutation's count of reference and alternate supporting reads integrating from 0 to the contEst contamination level for that tumor sample. The VAF beta cdf was compared to a threshold of 1e-5 and any mutation exceeding the threshold was rejected as a likely consequence of sample contamination.

NALT≤1 filter and deTiN: The final post-process filter (Terra task filter_NALT01_maf) removed any mutation with more than 1 alternate allele supporting read in the normal sample. This would potentially contradict the effect of the deTiN (*79*) algorithm, which first estimated the level of tumor DNA in the matched normal sample (TiN) and restored mutations that were previously rejected due to evidence of the alternate allele in the normal that are consistent with the TiN estimate. The deTiN algorithm was slightly modified in that the input candidate mutations from the MuTect1 "call_stats" file was filtered prior to deTiN using the PoN filters described above, modified so that it could operate on call_stats files in addition to maf formatted files as input. This step in the pipeline was implemented in Terra as Filter_call_stats. After deTiN processing we found that the bulk of mutations with 2 or more alternate allele supporting reads occurred in samples with Tin=0, so we excluded all such mutations.

TERT mutation detection: To identify tumors with non-coding *TERT* hotspot mutations (C228T at hg19 chr5:1295228 and C250T at hg19 chr5:1295250) we used MuTect 1.0 in "forcecall" mode (Terra task Mutect1_ForceCall) across the *TERT* UTR region from chr5:1295103-1295257, which reported coverage and alternate allele counts across the region.

Microsatellite Insertions and Deletions

Somatic microsatellite insertions and deletions (indels) were ascertained using MSMuTect, as described previously (*21*). Briefly, microsatellite loci (sequences with at least five successive motifs, regardless of the motif size) and alleles were identified separately in tumor and normal samples. Microsatellite loci with inferred alleles that differed between tumor and normal samples were considered potential somatic mutations. PoN filtering to account for either missed germline events or sequencing artifacts excluded loci that were identified in >2 samples from a set of 210 individuals without cancer from the Ukraine collected for another study. Analyses assessed the total number of microsatellite indels as well as insertions and deletions separately.

Structural Variant Detection and Filtering Pipeline

Like the mutation detection pipeline, the SV detection pipeline was based on the consensus of calls among four structural variation detection algorithms (**Fig. S13**). The four algorithms were dRanger/Breakpointer (*80*), SvABA (*68*), PCAWG_snowman (*81*), and Manta (*82*) (Terra task manta). Default parameters were used for all algorithms as specified in the configuration files in Terra workspace REBC_methods_only. Output SV events from each algorithm were converted into a common format (tab delimited list of events patterned after the Breakpointer format) and filtered using a PoN based on PCAWG data (dbGaP accession # phs000178.v1.p1, EGA accession # EGAS00001001692). The conversion was done by Terra tasks svaba_snowmanvcf2dRangerForBP, mantavcf2dRangerForBP, extract_dRanger_intermediates, and pcawg_snowmanvcf2dRangerForBP (with the relevant algorithm in the task names). The merged set of SV events across the four algorithms was constructed by clustering the breakpoints and strands of each event within a window of 350 bp (approximate fragment length) by the Terra

task SV_cluster_forBP. The consensus SV calls generally had slightly different breakpoints and allele counts from the various algorithms, so the selected variant breakpoints and allele counts were set according to the algorithm order dRanger, Manta, SvABA, and PCAWG snowman. After the set of calls from the algorithms were combined into a single call set, a final Breakpointer job (Terra task breakpointer) detected the consensus breakpoint positions, counts of alternate and reference supporting split-reads and read-pairs, and identified any sequence homology or novel sequence at the breakpoints.

The consensus SV filter (Terra task REBC_SV_consensus_filter_v3) required at least two algorithms to detect a mutation for the SV event to pass to the next layer of analysis. Each algorithm was given one vote in the consensus. There was no need for a more sophisticated consensus scheme since the contribution of calls based only on PCAWG snowman and SvABA was negligible (2 events) after filtering (**Fig. S14**), while dRanger and Manta were considered as "independent" algorithms. SV post-process filters demanded at least four alternate allele supporting reads (split-reads or read pairs, with each read pair counted at most once) from the tumor sample and at most one alternate supporting read in the normal sample. The Variant Allele Fraction of a given SV was estimated as

$$VAF = TALT/(TALT + TREF/2)$$

where *TALT* and *TREF* are the counts of supporting pairs (split-read or read pairs, but each fragment counts at most once for fragments with both read pair and split read support for the alternate allele). SV calls with VAF<0.1were rejected by the filtering task. The SV filter also excluded of SVs with breakpoints within centromere or telomere regions.

At a later stage in the analysis, we found that 10% of the candidate SV events had the same precise breakpoints in multiple tumor samples, adjacent tissue samples, and/or unmatched normal samples. These "hotspot" SVs had low counts of supporting reads and tended to occur in in regions of germline structural variation and were interpreted as hotspot somatic artifacts. None of these hotspot SVs occurred in SV driver genes. These hotspot SV events were rejected as artifacts using a custom matlab script REBC_SV_hotspots.4May2020.m outside of the Terra framework (https://github.com/getzlab/REBC_tools/releases/tag/1.0)

The final step in SV processing was to identify simple/balanced events, simple/unbalanced and complex clusters of SV events based on breakpoint proximity of the SV calls, similar to the recent PCAWG approach (*83*). An SV with both breakpoints within 1kb of another SV in the same sample with opposite strands (+- and -+) were classified as simple/balanced SVs, which represent balanced interchromosomal translocations, balanced intrachromosomal translocations, and inversions. SVs with a breakpoint within 50kb of another SV in the same sample regardless of strand were considered clustered SVs consistent with possible chromoplexy. SVs without reciprocal SVs identified were considered simple/unbalanced and represent deletions, tandem duplications, and undetected balanced partner SVs (inversions or translocations). To improve sensitivity to finding SV clusters, SVs that failed post-process filters and were not called by PCAWG snowman or SvABA only were included as potential "links" in the chain. The inclusion of SVs that had failed filters only marginally expanded the list of clustered SVs by less than 10%. The classification of SVs into simple/balanced, simple/unbalanced, and complex chains

was done using a custom matlab script REBC_balanced_inv2_clusters_unfiltered_27Jun2020.m outside of the Terra framework (https://github.com/getzlab/REBC_tools/releases/tag/1.0).

Chromothripsis and Chromoplexy
Chromothripsis was defined as SV complex clusters with uninterrupted oscillations between two copy number states in at least seven adjacent segments, according to the recent PCAWG approach (*83*). Only one of the 383 PTC tumors was found to have chromothripsis, with the cluster involving chromosomes 11 and 22, which resulted in an *IGF2* overexpression driver.

Chromoplexy events were defined as SV complex clusters with ≥3 DNA breaks involving ≥2 chromosomes that create a loop, similar to the recent PCAWG approach (*83*). Loops were connected if the gap was <500kb or if there was a corresponding SCNA for gaps ≥500kb. Chromoplexy events were manually checked to confirm if the cluster was indeed a loop, except for the chromoplexy events with ≥30 SV calls (n=2 in two different PTC tumors). Chromothripsis and chromoplexy were considered mutually exclusive; the one tumor with chromothripsis was not assessed for chromoplexy.

SCNA Detection and Filtering Pipeline
Copy number detection was performed on the WGS data using the GATK 4.1.4 CNV workflows (**Fig. S16**). The first step was to build a PoN (Terra task 1-CNV_Somatic_Panel), followed by detection of allele copy number alterations (Terra task 2-CNV_Somatic_Pair). Both of these GATK4.1.4 method configurations were based on workflows from the public Terra workspace help-gatk/Somatic-CNVs-GATK4. The PoN consisted of 423 CTB normal samples that showed no sign of Tumor-in-Normal (TiN) contamination. The combination of total copy number based on normalized read coverage and germline heterozygous SNP allele fraction shifts allowed an estimate of the allelic copy number ratio across the genome and to assign allelic copy number ratios to discrete segments. Following the 2-CNV_Somatic_Pair workflow, germline copy number polymorphisms were removed and the allelic copy number information was transformed to AllelicCapSeg (*84*) format using the Terra task Model_Segments_PostProcessing_canonical_gatk414. A preliminary collapsing step combined allelic copy ratio segments that were within the estimated error on the mean for the respective segments, which produced a substantial reduction in noise in the segments while preserving arm-level and focal events. The preliminary collapsing step was implemented in Terra as task MAF_AC_PP_CCF_fit_v3.

A further collapsing step was implemented (Available at: github.com/hartleys/REBC_SCNA_utils and archived on Zenodo (*85*)) to group SCNA events of the same type across intervening segments without an SCNA designation (untyped event, see below), which most often occurred when segments had no heterozygous SNPs. For example, a series of DEL-Untyped-DEL segments would count as one DEL compared to DEL-GAIN-DEL segments, which would count as two DELs and one GAIN. Merged events were marked clonal if at least one of the component events was clonal (defined as CCF ≥ 0.75).

For analyses testing the association of radiation exposure with SCNA events, rather than utilize the typical arm-level designation, we separated SNCAs that result from aneuploidy (chromosome level) from those that were likely the result of DNA double-strand breaks (sub-chromosome

level). Chromosome level SCNA events were counted when the entire chromosome had only one merged event and that event spanned the entire chromosome. This required the absence of any "non-event" segments. Sub-chromosome level events were SCNA events that did not span the whole chromosome, with either sections of normal copy number or multiple event types.

Despite best efforts to optimize parameters of GATK4.14 CNV detection, one sample suffered from excessive coverage dropout and copy ratio noise using GATK4.1.4 CNV and was dropped from the SCNA radiation exposure association analyses. Only for that sample an alternative CNV pipeline was used that was based on the AllelicCapSeg algorithm (Terra tasks wgs_pip_m1_fragcounter_oxoq_1 and allelic_Capseg_WGS). The output formats of both allelic_Capseg_WGS and Model_Segments_PostProcessing_canonical_gatk414 were equivalent so the downstream copy number analysis (GISTIC2.0 and ABSOLUTE) were the same for all samples or cohorts.

Purity and Ploidy Assessment
ABSOLUTE (*84*) was run on all pairs, however only 20% of the tumors showed sufficient allelic copy number signal to make tumor purity and ploidy estimates based on ABSOLUTE manual review of all tumors. To estimate purity for the remainder of tumors a matlab script (alleleFraction_clonal_hets_Purity_estimate.m, and REBC_consensus_mutations.m https://github.com/getzlab/REBC_tools/releases/tag/1.0) was developed to estimate purity based on identifying the somatic clonal allele fraction peak of multiplicity 1 mutations. Briefly, the beta distribution for the variant allele fraction:

$$p(VAF|TALT, TREF)=\beta(VAF, TALT+1, TREF+1)$$

was calculated for each mutation to model the variant allele fraction (VAF) probability distribution for each mutation with TALT alternate allele supporting reads, and TREF reference supporting reads. Beta distributions in bins of 0.002 VAF units for mutations passing "high stringency" filters (eg. Only mutations with no ALT alleles in the normal sample, NALT=0, and excluding regions of copy number alterations) were summed into a combined VAF distribution and the highest peak with a VAF<0.5 was taken as an initial estimate of the clonal VAF. Mutations consistent with this clonal VAF (p-value>0.05) were then selected as the mutations most likely to be clonal and the joint likelihood VAF was constructed across all of these "clonal het" mutations. The VAF distribution defines the clonal VAF confidence interval and the median of the VAF distribution was the estimated clonal VAF value. The tumor purity is then estimated as 2*VAF and the confidence interval of the purity estimate is the 68% (1-sigma) or 95% (2-sigma) width of the 2 x VAF joint likelihood distribution.

Final estimate of purity considered both ABSOLUTE and mutation VAF methods, but in most cases the purity estimate was based on mutation VAFs since relatively few tumors had sufficient copy number variation. A potential limitation of the mutation VAF method was the lack of constraint on tumor ploidy, but this was not a practical limitation because the samples that had exhibited copy number alternations used the ABSOLUTE algorithm, which did include a ploidy estimate. Only seven tumors had a tumor ploidy estimate inconsistent with 2.0.

For the purposes of most statistical analyses, we excluded samples with tumor purity <20% or with evidence of tumor contamination in the normal tissue (**Fig. S1**). We further excluded two tumors as low purity samples because the fraction of clonal mutations used to predict the purity at the time of final sample eligibility determination was <1% of the total variants (review of the VAF distribution revealed that these outlier variants were inflating the purity estimate).

Variant Clonality (Cancer Cell Fraction) Estimation
SCNA CCFs: The ABSOLUTE algorithm (*84*) includes an estimate of the cancer cell fraction (CCF) and alternate allelic copy number for each allelic copy ratio segment, however since the bulk of the tumors did not exhibit sufficient copy number to constrain ABSOLUTE purity and ploidy solutions with corresponding allelic Somatic Copy Number Alterations (aSCNA) and mutation CCFs so a custom matlab script MAF_AC_PP_CCF_fit_v3.m (https://github.com/getzlab/REBC_tools/releases/tag/1.0) was implemented as Terrra task MAF_AC_PP_CCF_fit_v3. The logic of this task was largely based on the logic described in the ABSOLUTE paper (*84*) with an additional constraint penalizing solution in which the fitted absolute allelic copy number fluctuates slightly above an integer value, leading to a sharp discontinuity in the SCNA CCF. A simplifying assumption in both ABSOLUTE and MAF_AC_PP_CCF_fit_v3 is that each aSCNA segment represents a combination of the WT allele (copy number 1 for each allele in the autosomes), copy number NA on the A allele (minor lower CN allele), and NB on the B (major allele) variant alleles in fraction (CCFA and CCFB) of tumor cells. The basic logic is to fit the allelic copy number (*NA* and *NB* for the minor A and major B alleles) with CCFs (*CCFA* and *CCFB*) to the observed allelic copy ratios:

$$CRA \sim (CCFA*\alpha*NA + (1-CCFA)*\alpha + (1-\alpha))$$
$$CRB \sim (CCFB*\alpha*NB + (1-CCFB)*\alpha + (1-\alpha))$$

for each allelic copy ratio segment, where α is the tumor purity. NA and NB were by definition alternate alleles (not 1 on the autosomes) so segments of normal copy number had CCFs equal to zero, or very close to zero due to the uncertainty in the copy ratio measurement.

Mutation CCF: ABSOLUTE (*84*) includes an estimate of the cancer cell fraction (CCF) and multiplicity (m) for each mutation, however as in aSCNA CCF estimation the bulk of the tumors did not exhibit sufficient copy number, so mutation CCFs and multiplicities were estimated by the custom matlab script MAF_AC_PP_CCF_fit_v3.m (https://github.com/getzlab/REBC_tools/releases/tag/1.0, Terra task MAF_AC_PP_CCF_fit_v3). The modified aSCNA constraint also helps to constrain mutation CCF and multiplicity for those mutations within problematic SCNA regions. The CCF and multiplicity of a given mutation is fit based on the observed ALT and REF allele counts, the local copy number state (NA, NB, CCFA, CCFB), and tumor purity α:

$$VAF = (CCF*\alpha*m) / TotDNA$$

$$TotDNA = \alpha*(NA*CCFA + NB*CCFB + (1-CCFA) + (1-CCFB)) + (1-\alpha)$$

The multiplicity *m* of a mutation depends on the substrate of allelic copy number (NA and NB) at the site of the mutation, where the mutation may occur on either A or B copy number alleles

(unless the allele is a deletion) or occurs in tumor cells that have normal copy number. The likelihood fit was done for each possible copy number substrate and the solution with the highest likelihood was selected. The six possible configurations for a mutation to occur on a copy substrate of NA, CCFA and NB, CCFB were:

1) Mutation on allele A (NA, CCFA) with m=1
2) Mutation on allele B (NB, CCFB) with m=1
3) Mutation on allele A (normal CN, 1-CCFA), m=1
4) Mutation on allele B (normal CN, 1-CCFB), m=1
5) Mutation on allele A (NA, CCFA) with m=NA
6) Mutation on allele B (NB, CCFB) with m=NB

The mutation VAF was calculated for each configuration over the full range of mutation CCFs, and the likelihood was estimated by the binomial pdf using the mutation ALT counts, sequencing depth, and VAF. The maximum likelihood over mutation CCFs and configuration set the final mutation CCF and CCF confidence interval. The posterior likelihood included a small prior (1%) favoring higher CCF solutions was used to break ties between different solutions. Note that the overwhelming bulk of mutations in this cohort occurred in regions without measurable copy number alternations such that configuration 3 (the same as 4 since CCFA=CCFB=0) was the only possible configuration.

SV CCF: Estimating the CCF of SV events is similar to the method used for mutations, except that SVs incur additional complications. The first complication is that each breakpoint could have its own copy number substrate such that the variant allele fractions for the two breakpoints could be different, although the cancer cell fractions for both the breaks must be the same. The second complication is that the observed counts of ALT and REF supporting is more prone to bias against the ALT allele than is the case for mutations.

For each breakpoint, the somatic copy number substrate is determined by matching the side of the break containing the ALT supporting reads to the local copy number segment NA, NB, CCFA, and CCFB. Since SV breakpoints can also be copy number breaks it is important to match the local number segment on the side of the break with the supporting reads. Once the copy number state is found for each breakpoint, the SV CCF and multiplicity for each break is found using the same procedure as for mutations, providing two estimates of the SV CCF with confidence intervals. In some cases, a breakpoint occurs in a region without a measured copy number segment (eg. noisy SCNA region or a segment not containing any germline hets used to estimate allelic copy ratios). No breakpoint CCFs are estimated without a corresponding copy number state. The SV CCF distribution is estimated as the average of the breakpoint CCF distributions. The SV CCF is the median of the CCF distribution and the confidence intervals (95% CI and 68% CI) are calculated from the CCF distribution. The SV CCFs were calculated by the Terra task SV_CCF_v3, which was based on the matlab script by the same name (https://github.com/getzlab/REBC_tools/releases/tag/1.0)

CCF Clonal and Subclonal Thresholds: For the SSV mutations, clonal variants were defined as those with a CCF hat >0.9 in order to generate an ultra-clean clonal set (**Fig. S8**). Subclonal SSV variants were those with a CCF hat <0.6. For both the SV and SCNA variants, clonal variants were defined as those with a CCF hat ≥ 0.75 with subclonal variants having a CCF hat <0.75.

Cohort-Level Analysis
Mutsig2CV: MutSig2CV performs an unbiased assessment of all somatic mutations, identifying candidate driver genes as those with recurrent mutations occurring significantly more often than expected from the null model (**Fig. S20**) (*86*) (Terra task Mutation_MutSig2CV_hg19_consensus_fix_NALT01_maf). MutSig2CV identified eight genes as significantly mutated with an FDR<0.1: *BRAF*, *NRAS*, *HRAS*, *TG*, *KRAS*, *TSHR*, *CR2*, and *DICER1* (**Fig. S20**).

GISTIC2.0: Significantly recurrent copy number regions were detected using GISTIC2.0 (*87*) (Terra task CopyNumber_Gistic2_hg19_GATK414) for both chromosome arms and focal peaks. GISTIC identified 9 significant arm-level results, 2 significant focal amplifications, and 7 significant focal deletions (**Fig. S19**). Three of the seven recurrent deletion peaks were associated with driver fusions (*RET*, *NTRK3*, and *ALK*), which was interpreted as additional evidence for these fusion partners.

Mutational Signature Analysis
Mutational signature classification was performed on total, clonal, and subclonal single base substitutions (SBS) and small indel (ID) variants separately using SigProfiler (*20*, *22*), combining samples from our study with the 42 individuals without known radiation exposure who had available tumor-normal paired WGS data from the TCGA analysis (*10*). Briefly, SBS were classified according to the 6 possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G, considering the pyrimidine of the mutated bp), plus the flanking 5´ and 3´ bases. Indels were first classified as deletions or insertions. Single base indels were then classified by the homopolymer length (length of the mononucleotide repeat tract in which they occurred; 1, 2, 3, 4, 5, or 6+ for single base deletions; 0, 1, 2, 3, 4, 5+ for single base insertions). Indels >1 bp were then classified by the whether they occurred at repeats or with overlapping microhomology at deletion boundaries, and finally by the indel size. We identified these SBS and ID patterns using SigProfilerMatrixGenerator (version 1.1.0), with modification using presig, version 0.0.1 to identify the single base indels using our untrimmed MAF files as inputs (Available at: https://github.com/edawson/presig/releases/tag/0.1.0 and archived on Zenodo (*88*)).

We then performed nonnegative matrix factorization-based signature extraction using SigProfilerExtractor version 1.0.3 via the SigProfilerHelper wrapper (Available at: https://github.com/edawson/SigProfilerHelper and archived on Zenodo (*89*)), performing 1000 iterations to optimally attribute the mutational patterns to one of the known 96 SBS or 83 ID signatures (signature deconvolution) from the Catalogue of Somatic Mutations in Cancer (COSMIC v3, https://cancer.sanger.ac.uk/cosmic/signatures) (*20*) or to identify *de novo* signatures. SBS and ID signature analyses were run on the set of samples from the CTB and TCGA combined to facilitate comparison of results among the study populations (**Fig. S10-S12**). **Table S5** shows the fraction of mutations attributable to each known COSMIC or *de novo* signature and the correlation amongst different signatures in the full study population. Signature decomposition analyses restricted only to those from the CTB overall, those who were exposed to any radiation, or those with an estimated thyroid radiation dose >200 mGy were highly correlated with those from the combined analysis with the TCGA samples and thus are not presented (data not shown). Results were plotted using tidysig, version 0.0.1 (Available at:

https://github.com/edawson/tidysig/releases/tag/0.0.1 and archived on Zenodo (*90*)), an R package developed to maintain a consistent style among the plots and supporting normalized counts across signatures in the plots.

For further evaluation of the deconvolution SBS signatures, we also utilized SignatureAnalyzer, which has a slightly different approach to nonnegative matrix factorization compared with SigProfiler (*91*). As shown in previous studies (*20*), results from the two approaches were highly correlated and yielded similar results; herein we present data from SigProfiler.

SNP Array Data Processing and Subject Ancestry

In order to understand the underlying population genetic structure of study population, we used genotypes drawn from a study of 7,544 individuals from Europe (*92*), which included countries adjacent to Ukraine in Eastern Europe. Additionally, 157 European individuals from the Human Genome Diversity Project (*93*) also were included from France, Italy, Orkney Islands, and Russia. The total number of individuals by country were Czech Republic, n=902; Denmark, n=137; Finland, n=1935; France, n=300; Germany, n=108; Greece, n=36; Hungary, n=259; Italy, n=139; Norway, n=521; Orkney Islands, n=15; Poland, n=1075; Romania, n=228; Russia, n=1593; Slovakia, n=209; Spain, n=55; Sweden, n=130, and the Netherlands, n=59.

Subject genotypes were cleaned using PLINK 1.90 (https://www.cog-genomics.org/plink/1.9/general_usage#cite) with the following thresholds: SNP-level missingness <2%, Hardy-Weinberg equilibrium exact test P>0.001, and minor allele frequency >0.01. Variants were pruned for linkage disequilibrium with PLINK 1.9 --indep and --indep-pairwise, and heterozygosity outliers were screened at |F|>0.2. Subject ancestry was computed with GRAF (*94*). Variants with informative missingness as computed with PLINK 1.9 --test-missing (permutation p<0.05) or --test-mishap ($p<1\times10^{-5}$) were removed. PCA was run using smartpca (from EIGENSOFT v6.1.4) (*95*) with outliers flagged at >6 standard deviations (n=8). For analyses of germline genetic variants, the top 10 principal components were generated for inclusion in multivariable models, excluding these 8 outlier individuals, as well as 3 individuals with estimated East Asian ancestry >10%, and 1 individual in the set of twins.

Germline Mutation Detection and Filtering Pipeline

Germline variant calling from the WGS data was performed using HaplotypeCaller (docker version broadinstitute/gatk:4.0.6.0, Terra task haplotypecaller-gvcf-gatk4), FreeBayes (v1.1.0-46-g8d2b3a0), and Strelka (version 2.8.3). For HaplotypeCaller, the GATK combineGVCFs (v3.8.0) utility was used to joint call all samples together. For FreeBayes and Strelka variants were called individually for each sample. All individual sample-level variant files were left aligned and trimmed, and multiallelic variants were split apart using vArmyKnife (v2.2.167, concordanceCaller utility) and vcftools (v0.1.16). GATK CombineVariants (v3.8.0) was used to create a merged, multi-sample variant file.

Primary identification of genotypes was based on HaplotypeCaller, with the secondary callers (Strelka and/or FreeBayes) used to confirm the existence of each variant. We filtered poor quality data based on caller concordance and hetAB fraction (the ratio of the alternative allele read depth versus the total read depth among heterozygotes). Variants were required to be detected by HaplotypeCaller and Strelka and have hetAB>0.2 at a minimum, as well as fulfilling

one of the following: 1) called by all three callers, 2) hetAB>0.3, or 3) <3 observed heterozygotes and hetAB>0.25.

From these filtered germline mutations, candidate gene investigation was assessed for potentially protein damaging variants in genes associated with thyroid cancer predisposition specifically, cancer predisposition more generally, and/or DNA damage response (*96*). Variants were annotated from external variant databases using the SnpSift software library, accessed via vArmyKnife, filtering out those variants that had a frequency of ≥1% in any ancestral population in the Exome Aggregation Consortium (ExAC; excluding TCGA) (*97*), 1000 Genomes Project (*98*), Exome Sequencing Project (ESP; Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA; http://evs.gs.washington.edu/EVS/), gnomAD exomes (excluding TCGA) or gnomAD genomes. Variants were filtered if >10% of the CTB sample set had one or more alternative alleles. Variants with set frequencies between 1 and 10% were flagged and manually reviewed by visualization in Integrative Genomics Viewer (IGV). Finally, we identified potentially protein-damaging variants in the genes of interest, defined as either SnpEff (*99*) classic high-impact variants without a ClinVar (*100*) designation of B/LB, or SnpEff moderate- or low-impact variants with a ClinVar designation of P/LP.

SNPs were extracted from the WGS data to construct a polygenic risk score (PRS) from 12 SNPs previously reported to be associated with thyroid cancer (**Table S15**) (*101*). The previously reported meta-analysis ORs range from 1.20 to 1.71 (*101*). The risk allele frequencies in gnomAD non-Finnish Europeans ranged from 3% to 75%, with five risk alleles having ≥45% frequency.

RNA-Based Fusion Detection

mRNA-seq was analyzed with four fusion detection programs: ChimeraScan (*102*), EricScript (*103*), STAR-Fusion (*104*), and MapSplice 2 (*105*). EricsScript and Mapsplice 2 were selected based on previous reviews and reports on accuracy and performance (*106, 107*). Potential fusions were ranked and scored by consensus detection and the number of supporting reads. A fusion score of 4.0 would indicate that the fusion was the top-ranked fusion for all four programs. For lower values, multiple scenarios could result in the same score. For a fusion score of 3.0, the fusion could be the top-ranked fusion in three callers but not detected (or ranked low) by the fourth caller (1+1+1+0) or could have been the top-ranked fusion in two callers and the median ranked fusion in the other two callers (1+1+0.5+0.5). Similarly, a score of 2.0 could indicate that a fusion was the top-ranked fusions for two programs and not detected (or ranked low) for the other two programs (1+1+0+0) or that the fusion was the median ranked fusion for all four programs (0.5+0.5+0.5+0.5). Code for RNA fusion calling is available at: https://github.com/NCI-CGR/ChernobylThyroidCancer-RNAFusionCalling and archived on Zenodo (*108*).

RNA-Based Mutation Detection

mRNA-seq was analyzed with VarDict to identify SSVs (*109*). Variant processing was carried out with vArmyKnife (v2.2.305). Variants were left aligned and trimmed and annotation was performed using SnpEff v4.3t. Code for RNA variant calling is available at: https://github.com/NCI-CGR/ChernobylThyroidCancer-RNAVariantCalling and archived on Zenodo (*110*).

mRNA, miRNA, and Methylation Clustering Analyses
We performed unsupervised consensus clustering using ConsensusClusterPlus. Input data included the variance stabilizing transformation (vst) expression of the 1000 most variably expressed genes for mRNA analysis, the vst expression of the 100 most variably expressed miRNAs for miRNA analysis, and the normalized beta value of 3000 most variably methylated CpG islands for methylation analyses. Code for clustering analyses is available at: https://github.com/NCI-CGR/ChernobylThyroidCancer-Clustering and archived on Zenodo (*111*). Code for detection of differentially methylated cpg islands is available at: https://github.com/NCI-CGR/ChernobylThyroidCancer-Methylation and archived on Zenodo (*112*).

The appropriate cluster number (k) was determined by identifying the largest cluster with a delta area value >0.3. Because PTC formed a distinctive cluster from non-tumor thyroid tissue in initial mRNA clustering analyses including both sample types (**Fig. S21**), subsequent clustering analyses for mRNA, miRNA, and methylation was restricted to PTC samples only (**Fig. S22-S23**).

Signaling and Thyroid Differentiation Scores
Based on the TCGA analysis (*10*), we constructed three scores to reflect transcriptional patterns related to key signaling and thyroid differentiation pathways (**Fig. S24; Table S12**). Since nearly all PTC oncogenic drivers are members of the MAPK pathway, we constructed a score to distinguish the transcriptional patterns of tumors with the two most common activating point mutations, *BRAF*^{V600E} and *RAS* (*BRAF*^{V600E}*-RAS* score [BRS]). We constructed the BRS score using mRNA, miRNA, and methylation data. Based on mRNA data, the 100 significantly differentially expressed genes between *BRAF*^{V600E} and *RAS* tumor-normal paired samples were determined, based on a DESeq2 q-value (adjusted p-value (*113*)) ≤0.01 and fold change ≥2. The variance stabilizing transformed (vst) expression of these 100 genes was used to quantify the degree to which each tumor's mRNA profile resembled that of either *BRAF*^{V600E} or *RAS*-mutated PTC, as described previously (*10*). A miRNA BRS was constructed, based on 77 miRNAs significantly differentially expressed between *BRAF*^{V600E} and *RAS* tumor-normal paired samples (DESeq2 q-value≤0.05 and fold change ≥1.5) in our study, and a methylation BRS, based on the normalized beta value of 1000 significantly differential methylated cpg islands between *BRAF*^{V600E} and *RAS* PTC samples, as determined by the overlap of results from two tools: minfi (*114*) and ChAMP (*115*) (q-value≤0.01 and fold change ≥2.0). To confirm the validity of the mRNA BRS score generated using our internal gene expression data, we also reconstructed the mRNA BRS score using data for 70 of the 71 genes differentially expressed between *BRAF*^{V600E} or *RAS*-mutated PTC in the TCGA analysis (*10*) (excluding *ANXA2P2*, a pseudogene not included in our GRCh38 gene annotation file). **Table S12** shows the high correlation (Pearson correlation, r=0.98) between the internal- and TCGA-based mRNA BRS scores, as well as among the mRNA, miRNA, and methylation BRS (r=0.78-0.92).

An ERK-activity score was constructed to estimate the expression profiles for genes that were shown in *BRAF*^{V600E}-mutated melanoma cell lines to be responsive to MEK inhibition (*116*), as per the previous TCGA analysis (*10*). Briefly, we compared the vst expression of 52 genes in tumors, standardized based on the expression of each gene in *BRAF*^{V600E} -like (BRS<0) or *RAS*-

like (BRS>0) tumors. Finally, we constructed a thyroid differentiation score (TDS), also according to the previous TCGA analysis (*10*), whereby we calculated the mean fold change in vst expression level across 16 thyroid function genes. Code for generating all scores is available at: https://github.com/NCI-CGR/ChernobylThyroidCancer-Scores and archived on Zenodo (*117*).

RNA-Seq Differential Expression and Gene Set Enrichment Analysis
We conducted exploratory analyses evaluating whether radiation dose was related to differential expression of specific genes and gene sets, following the approach used by PCAWG (*118*). Briefly, for analyses of individual genes, normalization factors for gene-level read-pair counts (generated as described above) were calculated using the "UQCT" method, which takes the 75 percentile of the counts for each sample, excluding zero-count genes. The raw counts were increased by 1 (to eliminate log-0 errors), divided by the normalization factor, and then $\log_2$-transformed. These log-normalized counts were used to fit a standard multivariable linear regression, as described below in the *Statistical Analysis*, with additional adjustment for experiment batch (phase) in addition to age at PTC and sex.

Patterns of RNA-Seq expression in previously identified pathways and gene sets were extracted from the Molecular Signatures Database (MSigDB v7.1; https://www.gsea-msigdb.org/gsea/msigdb) (*31*). All gene sets related to hallmark biological processes, thyroid, radiation, and the genes from our germline analyses were included, resulting in 3,213 gene sets. GSVA, an R package that performs "Gene Set Variation Analysis," was used to collapse expression information across gene sets. This tool takes log-normalized counts for each gene and set of genes, outputting a single value for each geneset/sample combination (specifically the Kolmogorov-Smirnov-like rank statistic) (*119*). Linear regression analyses were performed on these statistics, as described below.

Methylation Age
Exploratory analyses were conducted to evaluate whether radiation dose was associated with epigenetic age acceleration, estimated from tumor methylation profiles using two well-established approaches (*38, 39*). Briefly, we regressed epigenetic age against chronological age in the non-tumor thyroid tissue and then compared the residuals from this predicted age in the PTC tissue. These residuals were utilized as outcomes in linear regression models described below.

PTC Driver Identification
Candidate mutation and fusion drivers were identified in the 440 PTC tumors by interrogating both the WGS and mRNA-seq with a comprehensive candidate driver gene list. For the simple somatic variants or mutation drivers, candidates included genes significantly mutated in the MutSig2CV analysis of 383 WGS tumors (**Fig. S20**), genes previously reported as mutation drivers in the TCGA analysis (*10*), and genes reported in the COSMIC Cancer Gene Census (CGC) v90 (https://cancer.sanger.ac.uk/census) with frameshift, missense, nonsense, or splice site mutation types. For the SV or fusion drivers, candidates included previously identified fusion drivers or focally deleted genes (*BRAF* and *PTEN*) in the TCGA analysis (*10*) and genes reported as oncogenes or tumor suppressors in the COSMIC CGC with fusion indicated for

either the “role in cancer” or “mutation type.” Fusions involving *HRAS*, *KRAS*, and *NRAS* were included because of the substantial contribution of RAS-activating mutations in thyroid cancer.

For mutation drivers, we identified somatic protein altering variants that had a corresponding match in the COSMIC CGC v90 database; additionally, for *CR2*, which did not have a COSMIC CGC v90 entry, we considered the variants identified by MutSig2CV (described above). Mutation identification was carried out first in the WGS using the final somatic variant calls, which have inherent matched normal sample, and PoN filters (see *Mutation Detection, Merging, and Filtering Pipeline* section above). Since mRNA-seq variant calling is not as robust as WGS calling, additional rules were applied to exclude variants meeting any of the following criteria: 1) evidence in a PoN comparison set, created using the 392 high-quality non-tumor thyroid tissue samples that had a corresponding PTC mRNA-seq sample; 2) alternate allele read depth ≤10; 3) splicing variants; or 4) variants in genes not identified as drivers in the WGS data. Recurrent *BRAF* and *RAS* variants in the WGS dataset were force called to avoid the problem of recurrent drivers being present in the non-tumor thyroid tissue samples. For both WGS and mRNA-seq, we also manually reviewed doublet and indel mutations near *BRAF* p.V600 because of the strong supporting evidence for this locus in driving PTC. Two indels near *BRAF* p.V600 were identified and determined to be mutation drivers: one resulted in p.600_601VK>E and had a direct COSMIC v90 database match, and other resulted in the insertion of six amino acids and was similar to another COSMIC v90 database entry. Because annotated databases are often SNV-biased and do not always match on doublets appropriately, we interrogated all nine WGS doublet calls individually and determined that three were drivers because the corresponding protein changes had a COSMIC v90 database match. Finally, *TG* variants were excluded as driver mutations because most of the *TG* variants were noncoding (1:17 coding:noncoding in *TG* versus 10:1 in *BRAF*), which is a hallmark of hypermutation due to lineage-specific overexpression (*120*).

For fusion drivers, we queried the WGS SV and mRNA-seq gene fusion calls to identify events that occurred in both the WGS (final SV calls) and mRNA-seq (RNA fusion score ≥2). Events occurring in only WGS or mRNA-seq utilized more stringent criteria, requiring SV calls to have 1) tumor alt reads ≥10, 2) normal alt reads=0, 3) number of SV callers (dRanger + max(pcawg_snowman, SvABA) + Manta ≥ 2, and 4) breakpointer type contains fusion or deletion and is not antisense, transcript fusion or out of frame protein fusion. mRNA-seq-only calls required RNA fusion score ≥3. Our approach allowed for identification of high-confidence fusion drivers in the set of tumors with both WGS and mRNA-seq (n=374) as well as those with only WGS (n=9) or only mRNA-seq (n=57). The approach ensured identification of fusion drivers resulting from a complex set of multiple SV events, in which mRNA-seq identified the final gene fusion product but the WGS identified multiple intermediate steps, for example, *NCOA4-RET* versus *NCOA4-SHOC2*, and *SHOC2-RET*. After the fact, we performed an unbiased assessment of the confirmed SVs utilizing the stringent WGS fusion criteria without restricting the analysis to the fusion candidate gene list and did not identify additional recurrent fusion driver candidates.

For tumors without an identified driver based on the above approach, we utilized PTC mRNA expression (clustering, overexpression outliers) and SCNAs to identify candidate drivers. Unsupervised mRNA clustering of tumors identified 5 clusters (as described above), which were

correlated with the already-identified drivers (**Table S11**). Cluster 5 included two tumors, one of which had both a frameshift and a stop-gained *TSC2* variant suggestive of biallelic loss of the tumor suppressor *TSC2*, which forms a complex with *TSC1*. Further review of the other tumor revealed a stop-gained germline *TSC1* variant with an SCNA loss over the *TSC1* coding region, suggesting biallelic loss of *TSC1* as the candidate driver. For the two tumors in cluster 3, further review of high impact (frameshift and stop-gained) somatic and germline variants revealed biallelic loss of *APC* as the candidate driver, with one tumor having a somatic *APC* frameshift variant (no COSMIC v90 database match) and an SCNA loss over the *APC* coding region, and other having a germline stop-gained pathogenic variant (based on ClinVar) and an SCNA over *APC*.

THADA-fusions are known as drivers in thyroid cancer (*10*) and have been found to result in overexpression of *IGF2BP3* (*121*). Three tumors with *THADA* fusions and corresponding *IGF2BP3* overexpression (vst expression >6 standard deviations from the mean) were identified with the rules above. We designated *IGF2B3* as the candidate driver for the one additional tumor with similar overexpression. Overexpression of *IGF2BP3* leads to an increase in IGF2 protein levels and activation of the PI3K and MAPK pathways (*121*), thus we also checked for *IGF2* overexpression outliers. We identified two tumors (one with mRNA-seq only, one with multiple SVs near *IGF2*) without a candidate driver based on the rules above with *IGF2* overexpression, which we designated as the candidate driver.

All candidate drivers are listed in **Table S7**. For **Fig. 1**, driver genes were displayed for oncogenes if they were recurrent within the dataset and for tumor suppressor genes if they were recurrent and had evidence for biallelic loss. For all driver analyses, final driver designations were made according to the following rules: 1) if individual has only one candidate driver that is recurrent (observed in at least one other sample); or 2) if >1 recurrent candidate driver, all drivers with ≥ 5 recurrences were outputted and further evaluated (n=3 tumors). For the tumor with a *PPARG* fusion and *NRAS* mutation, the *NRAS* mutation was designated as the final driver because it had higher CCF, whereas the CCF of the *PPARG* fusion was low and the tumor did not cluster with the other 13 *PPARG* fusions in the PTC mRNA cluster analysis. For the tumor with a *BRAF* p.V600E and *THSR* with similar CCF, the *BRAF* mutation was designated as the final driver due to its overwhelming recurrence rate. For the tumor with a non-V600E *BRAF* and *KRAS* mutations with similar CCFs, no final driver was designated.

For the 351 individuals that had both a high purity (>20% purity) WGS and mRNA PTC sample, designated mutation driver concordance was 100% (n=215 mutations), and designated fusion driver concordance was 100% (n=130 fusions), after accounting for the final fusion being identified in mRNA-seq and fusion intermediates in WGS, for example, *NCOA4-RET* versus *NCOA4-SHOC2*, and *SHOC2-RET*.

Analyses by the driver pathway combined drivers by biological function for those identified in small numbers of samples, specifically other RTK (receptor tyrosine kinase) fusions (*ALK*, *LTK*, *NTRK1*, and *NTRK3* fusions), *RAS* mutations (*HRAS*, *KRAS* and *NRAS* mutations), and *IGF2* and *IGF2BP3* overexpression drivers. Finally, the small number of other mutation drivers (*APC*, *DICER1*, *NFE2L2*, *TSC1*, and *TSC2*) were combined for model stability.

Identification of DNA Repair Mechanisms for SVs and Small Deletions
DNA double-strand breaks—the most important type of radiation-related DNA damage—can be repaired by various mechanisms, including two key types of end-joining repair: non-homologous end-joining (NHEJ) and alternative end-joining (alt-EJ or theta-mediated end-joining [TMEJ]) (*33-35*). We performed a series of analyses to classify the DNA repair mechanisms that generated the SVs and small indels in our dataset.

SV confirmation was based on a series of steps. Among individuals with a WGS sample (n=383), manual review in IGV was conducted for 145/150 designated fusion drivers (excluding four drivers without an available SV call and one where the breakpoint could not be evaluated due to the repetitive nature of the local sequence). During manual review, each fusion was categorized by the number of DNA breaks (2 [simple] versus >2 or local sequence bridge repair [complex]); SV type (deletion, inversion, interchromosomal translocation, intrachromosomal translocation, or complex); balanced (<1kb gained/lost at breakpoint) versus unbalanced (≥1kb gained/lost at breakpoint); end-joining (<20 bp for both breakpoints) versus non-end-joining/other (≥20 bp at one breakpoint); and microhomology at the breakpoint, which was determined utilizing blast within the UCSC genome browser (http://www.genome.ucsc.edu/). Agreement was acceptable between the automated and manual calls for all metrics except microhomology (22/226 or 9.7% discordant breakpoints) and accordingly, microhomology was not included in the final categorization. Notably, the review of fusion drivers indicated that certain variant groups (simple/unbalanced SVs and complex SVs from individuals with multiple complex clusters) were most likely to be false positives, thus all SVs in these categories were manually checked in IGV. The discordant events and false positives were corrected or removed to generate the final SV dataset (**Table S6**).

Following SV confirmation, SVs were first categorized as simple and balanced and end-joining (simple/balanced/end-joining) versus complex, unbalanced, or non-end-joining (complex/unbalanced/other). Simple/balanced/end-joining SV events with <4 bp of intervening loss/gain at both breakpoints, representing enrichment for NHEJ, were then separated from SV events with 4-19 bp of intervening loss/gain for at least one breakpoint, representing both NHEJ and alt-EJ/TMEJ repair mechanisms.

For small deletions called as indels (**Fig. S29**), direct counting of NHEJ and alt-EJ/TMEJ features within the 83 ID mutational signature classifications was performed to circumvent potential misattribution of the indels to ID6 and ID8, both of which have been ascribed to end-joining repair. In contrast to SVs, end-joining repair is most enriched when assessing deletions ≥5 bp due to noise from other small indel generating mechanisms with deletions <5 bp in length. The patterns consistent with end-joining repair were defined as deletions ≥5 bp that are not within repeats and have any amount of microhomology (≥0 bp of microhomology.) The amount of microhomology was utilized to further group the deletions by events enriched for NHEJ (0-1 bp microhomology at the deletion boundary) and those that are a mix of NHEJ and alt-EJ/TMEJ mechanisms (≥2 bp of microhomology at the deletion boundary). Locally templated insertion events (TINS), which are characteristic only of TMEJ repair, were identified as described previously (*34*). Briefly, small insertions ≥3 bp were interrogated for whether either the direct or reverse complement sequence was present within 50 bp either upstream or downstream of the indel breakpoint.

Identification of Genomic Locus-Based Attributes for Small Insertions and Deletions
Following the approach described previously (*8*), we annotated small insertions and deletions according to a series of genomic locus-based attributes:
Continuous metrics:

- Distance from the centromere ($\log_{10}$)
- Distance from the nearest telomere ($\log_{10}$)
- Fraction of G/C bases (flanking 100 bp in either direction)
- Trinucleotide complexity (flanking 50 bp in either direction; the sum of the squares of the occurrence rate for each possible trinucleotide combination. Higher values indicate lower complexity).
- Chromatin state (selecting one of the 25 chromatin states [heterochromatin] from ROADMAP and taking consensus across all 127 cell types because no data specific to the thyroid were available)
- Replication timing (average of the wavelet-smoothed replication timing signals in 1 kb bins across the genome, in three samples: NHEK [normal skin], GM12878 [normal blood], and IMR90 [normal lung]
- LADs (Tig3ET normal human embryonic lung fibroblasts)

Dichotomous metrics indicating whether the variant is within 100 bp of the given event:

- Genes (known protein coding gene found in GENCODE)
- CpG Islands (known CpG island)
- Direct Repeats
- G-quadruplexes
- Cruciform inverted repeat
- Triplex mirror repeat
- Short tandem repeat
- z-DNA motif
- ALU repeats
- MIR repeats
- L1 repeats
- L2 repeats
- LTR repeats
- DNA repeats
- Simple repeats

To compare the distribution of the small insertions and deletions observed in our study compared with a random background distribution, we randomly selected 100,000 loci from across the genome and annotated them using identical methods.

Statistical Analysis
Multivariable regression models were employed to assess the relationship between radiation dose and PTC molecular characteristics. Unless otherwise specified (**Table S4**), models were mutually adjusted for sex, age at PTC (continuous), and radiation dose (continuous, with doses >1000 mGy truncated [assigned the value of 1000 mGy] to reduce their influence on the

estimated model coefficients) (**Fig. S4**). Two-sided P-values were generated using likelihood ratio tests, comparing model fit with and without the variable of interest.

The regression model type depended on the distribution of the molecular characteristic and is specified for each characteristic in the supplementary tables that present model results (**Table S4, Table S18, Table S20-S22**). For continuous characteristics, such as the number of deletions, we utilized linear regression to estimate the expected value (E[Y]) as:

$$E[Y] = \beta_0 + \beta_1 * Dose + \beta_2 * Age_{PTC} + \beta_3 * Sex$$

where β_1 represents the change in the value of the molecular characteristic per unit dose, which was expressed per 100 mGy. Normality of the residuals was confirmed visually. Analyses were conducted using SAS version 9.4 (Cary, NC) and R (Foundation for Statistical Computing, Vienna, Austria).

For molecular characteristics that were dichotomous, such as presence or absence of chromoplexy, we utilized logistic regression to estimate the odds ratio (OR; the odds of the characteristic being present) as:

$$OR = \exp(\alpha_1 * Age_{PTC} + \alpha_2 * Sex)(1 + \beta_1 * Dose)$$

where β_1 represents the excess OR (EOR) per 100 mGy so that the effect of dose is linear (rather than log-linear), the standard approach in radiation epidemiology (*122*). For molecular characteristics with discrete counts over a limited range where the residuals were not normally distributed in linear regression models, such as the number of SVs, we utilized proportional odds models, collapsing the highest values into a maximum category that included at least 5% of the total study population. For "n" categories of each characteristic with this distribution, we created n-1 replications of the dataset; within each replication (i), we defined a new outcome with a value of 1 if the original outcome was $\geq$i or 0 if <i (missing values retained a missing value). We then included in the logistic regression model the index as a categorical variable in the log-linear term (along with other covariates such as sex and age at PTC), while modeling dose in the linear term, as in the standard dichotomous model described above. This proportional odds model assumes that the EOR/100 mGy is consistent across the categories of the molecular characteristic outcome strata. We confirmed the validity of this assumption visually for radiation-associated characteristics (as defined below and shown in **Table S20**) by fitting a polytomous model. Polytomous models were utilized for molecular characteristics without inherent ordering, such as mRNA clusters or driver gene, whereby we included the index term as a categorical variable crossed with every other term in the model (index*age at PTC and index*sex in the log-linear term, and index*dose in the linear term), thereby fitting a separate EOR/100 mGy for each level of the characteristic relative to the referent category. Logistic, proportional odds, and polytomous regression analyses were conducted using the GMBO module of Epicure, version 2.0 (Risk Sciences International, Ottawa, Canada) (*123*).

Our primary analyses included 68 PTC molecular characteristics derived from a comprehensive landscape analysis (**Fig. 1, Table S4**). We defined radiation dose-associated molecular characteristics based on $P<7.4\times10^{-4}$, which represents a Bonferroni correction for multiple

testing (α=0.05/68). For associated variables, we conducted analyses of more specific molecular characteristics as well as stratifying the study population into tertiles by age at PTC, age at exposure, and time since exposure (latency) because these factors influence radiation-related thyroid cancer risk (**Table S1**) (*17*). We tested for heterogeneity (modification) of the radiation dose effect on the molecular characteristic across these factors using continuous variables, comparing the model fit with and without an interaction term (Dose*age at PTC, Dose*age at exposure, or Dose*latency) using likelihood ratio tests. The models evaluating modification of the radiation dose effect by age at exposure and latency also included main effects of these variables, as follows:

Linear regression:

Modification by age at PTC:

$E[Y] = \beta_0 + \beta_1*Dose + \beta_2*Age_{PTC} + \beta_3*Sex + \beta_4*Dose*Age_{PTC}$

Modification by age at exposure or latency (Age/Lat):

$E[Y] = \beta_0 + \beta_1*Dose + \beta_2*Age_{PTC} + \beta_3*Sex + \beta_4*(Age/Lat) + \beta_5*Dose*(Age/Lat)$

Logistic and proportional odds regression:

Modification by age at PTC:

$OR=\exp(\alpha_1*Age_{PTC} + \alpha_2*Sex)(1 + [\beta_1*Dose \times \exp(\beta_2*Age_{PTC})]$

Modification by age at exposure or latency (Age/Lat):

$OR=\exp(\alpha_1*Age_{PTC} + \alpha_2*Sex + \alpha_3*Age/Lat)(1 + [\beta_1*Dose \times \exp(\beta_2*Age/Lat)])$

We conducted sensitivity analyses to assess whether the results were consistent when we restricted the population to individuals with lower dose exposure (1-<500 mGy, resulting in n=326 with mean estimated dose=110 mGy), a critical question in radiation epidemiology (*18, 19*). Finally, we evaluated whether there was a statistically significant departure from a linear dose model by fitting linear quadratic (Dose squared) and linear exponential [exp(Dose)] models, comparing the fit of these models to that of a linear model using likelihood ratio tests, as follows:

Linear regression:

Linear quadratic:

$E[Y] = \beta_0 + \beta_1*Dose + \beta_2*Age_{PTC} + \beta_3*Sex + \beta_4*Dose^2$

Linear-exponential:

$E[Y] = \beta_0 + \beta_1*Dose + \beta_2*Age_{PTC} + \beta_3*Sex + \beta_4*\exp(Dose)$

Logistic and proportional odds regression:

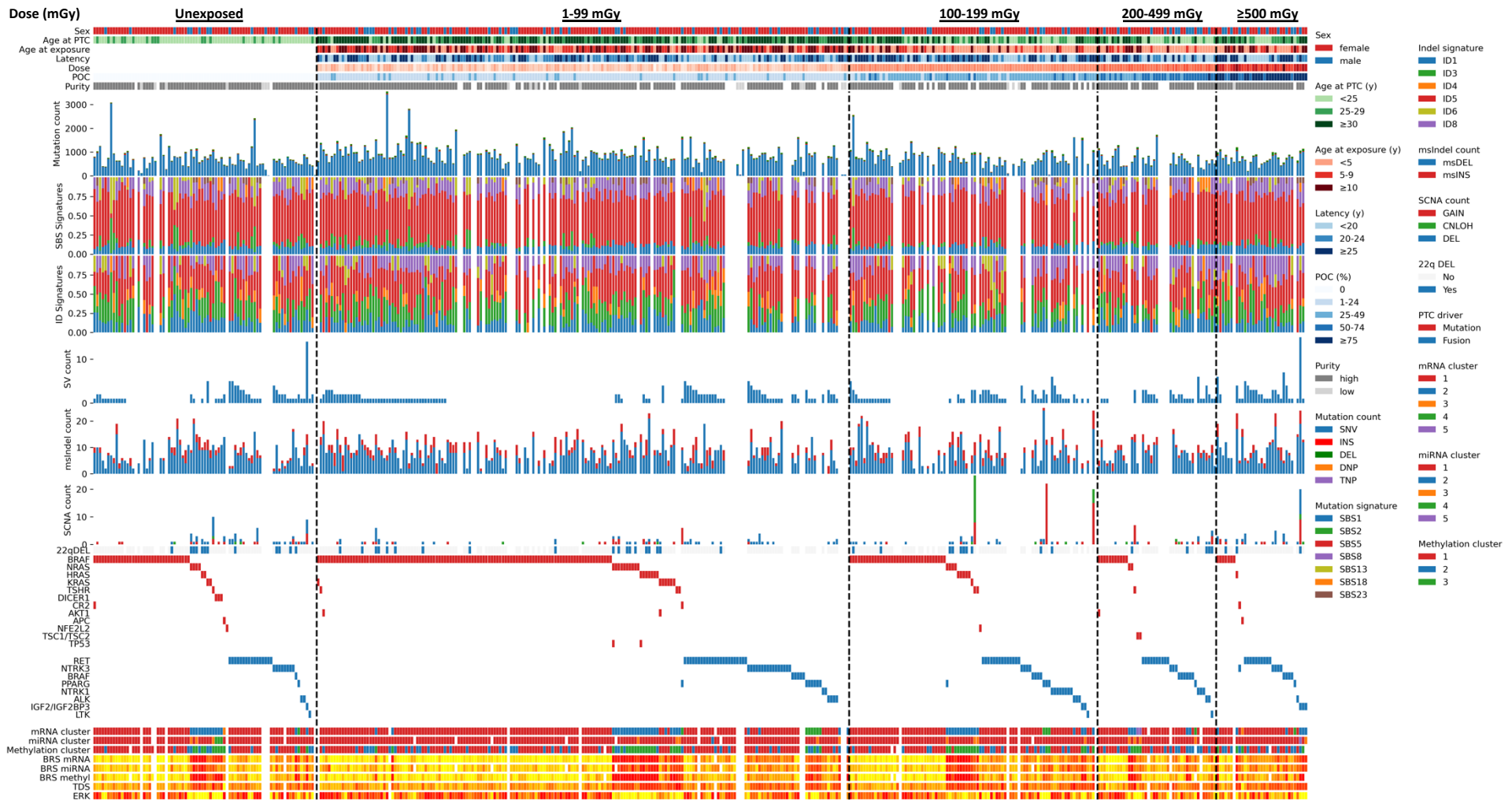
Linear quadratic:

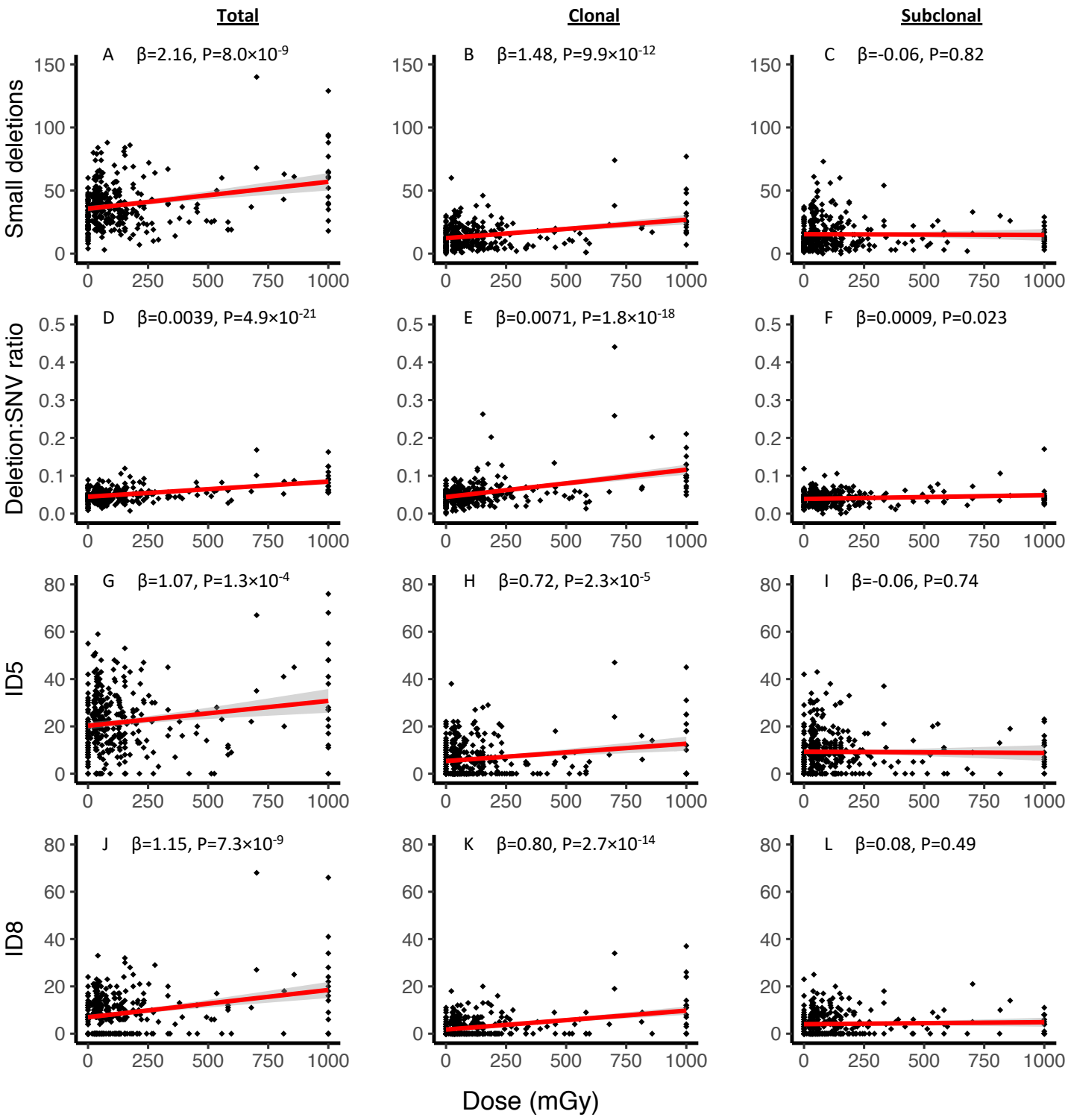
$OR=\exp(\alpha_1*Age_{PTC} + \alpha_2*Sex)(1 + \beta_1*Dose + \beta_2*Dose^2)$

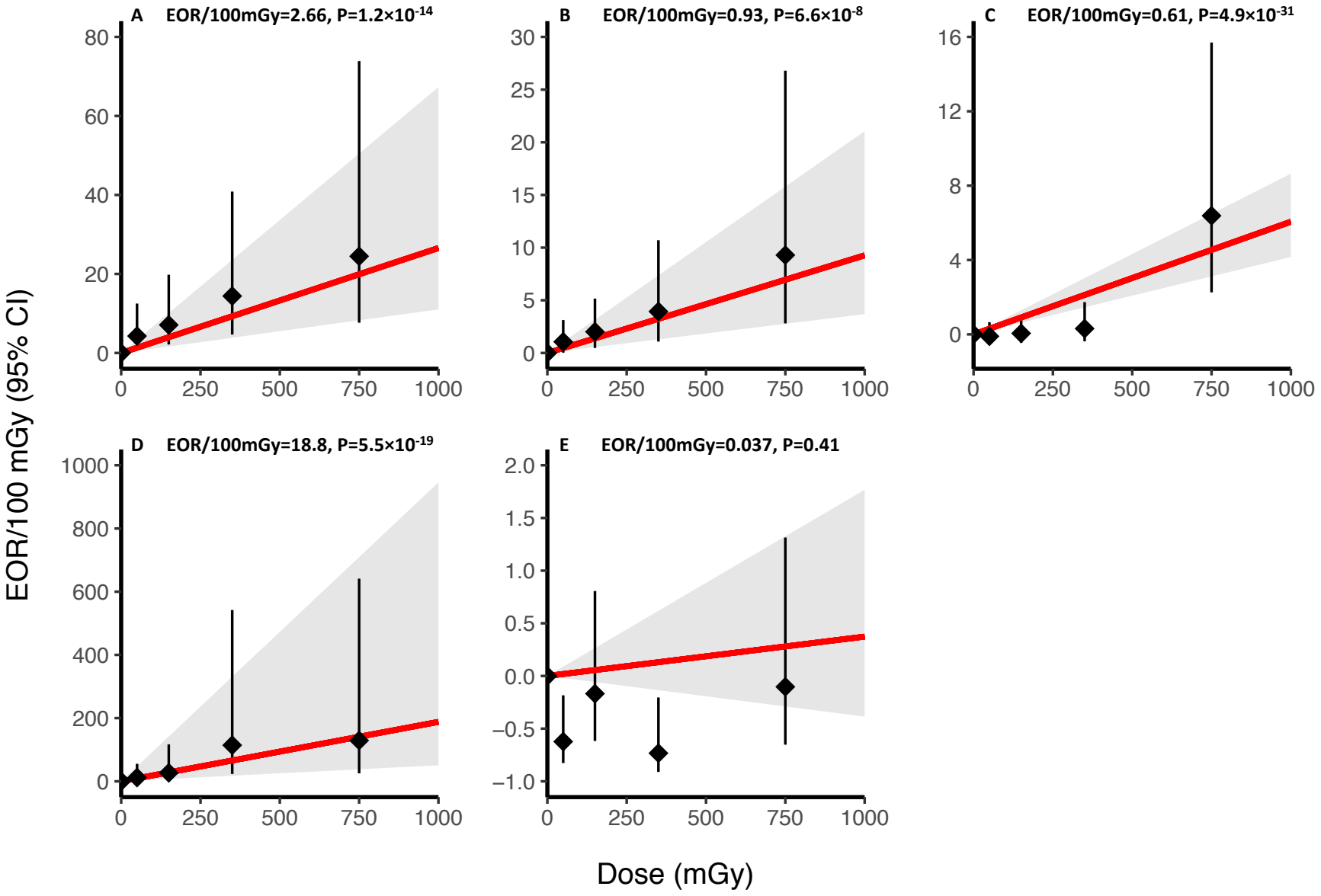
Linear-exponential:

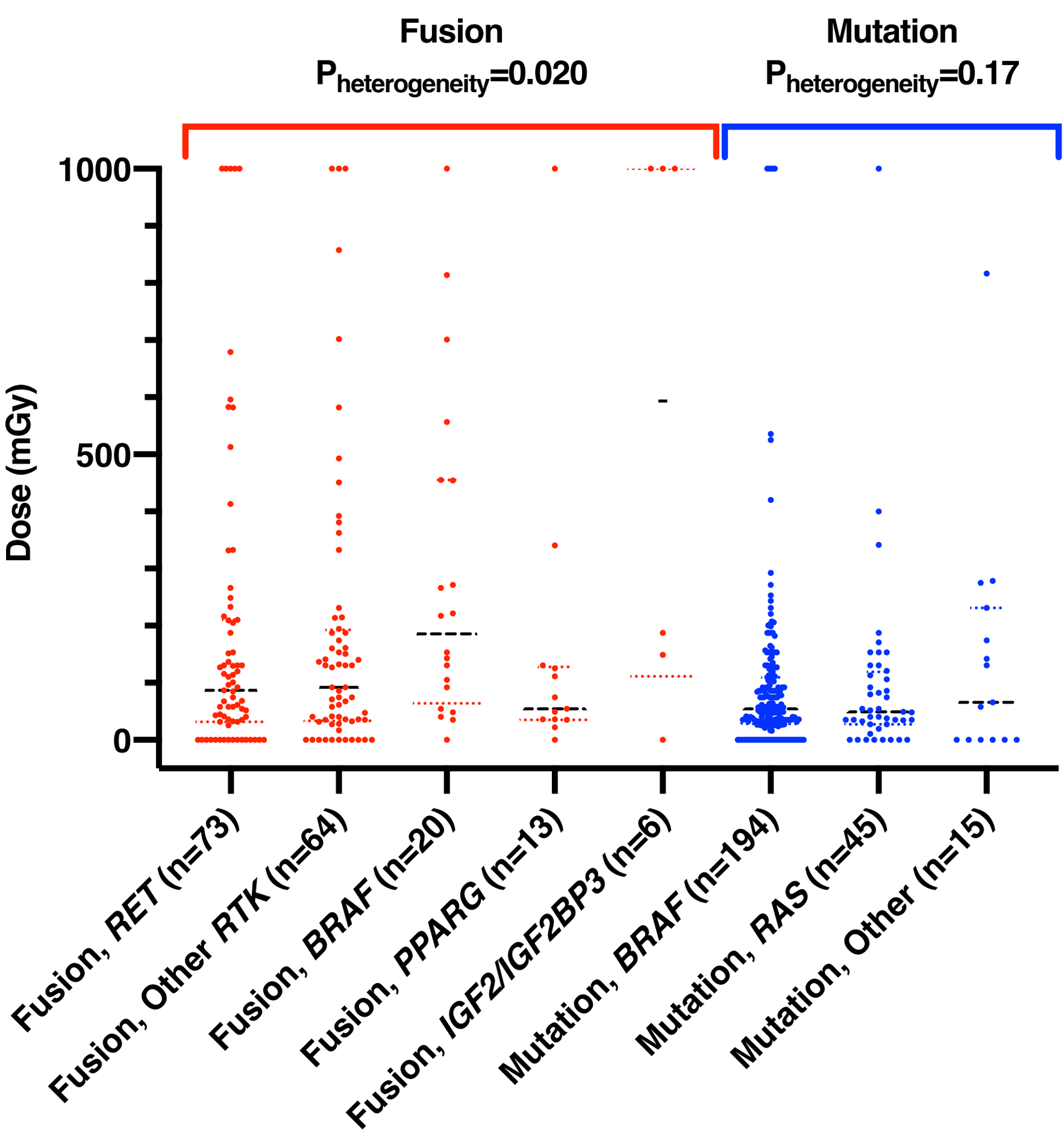
$OR=\exp(\alpha_1*Age_{PTC} + \alpha_2*Sex)(1 + [\beta_1*Dose \times \exp(\beta_2*Dose)])$

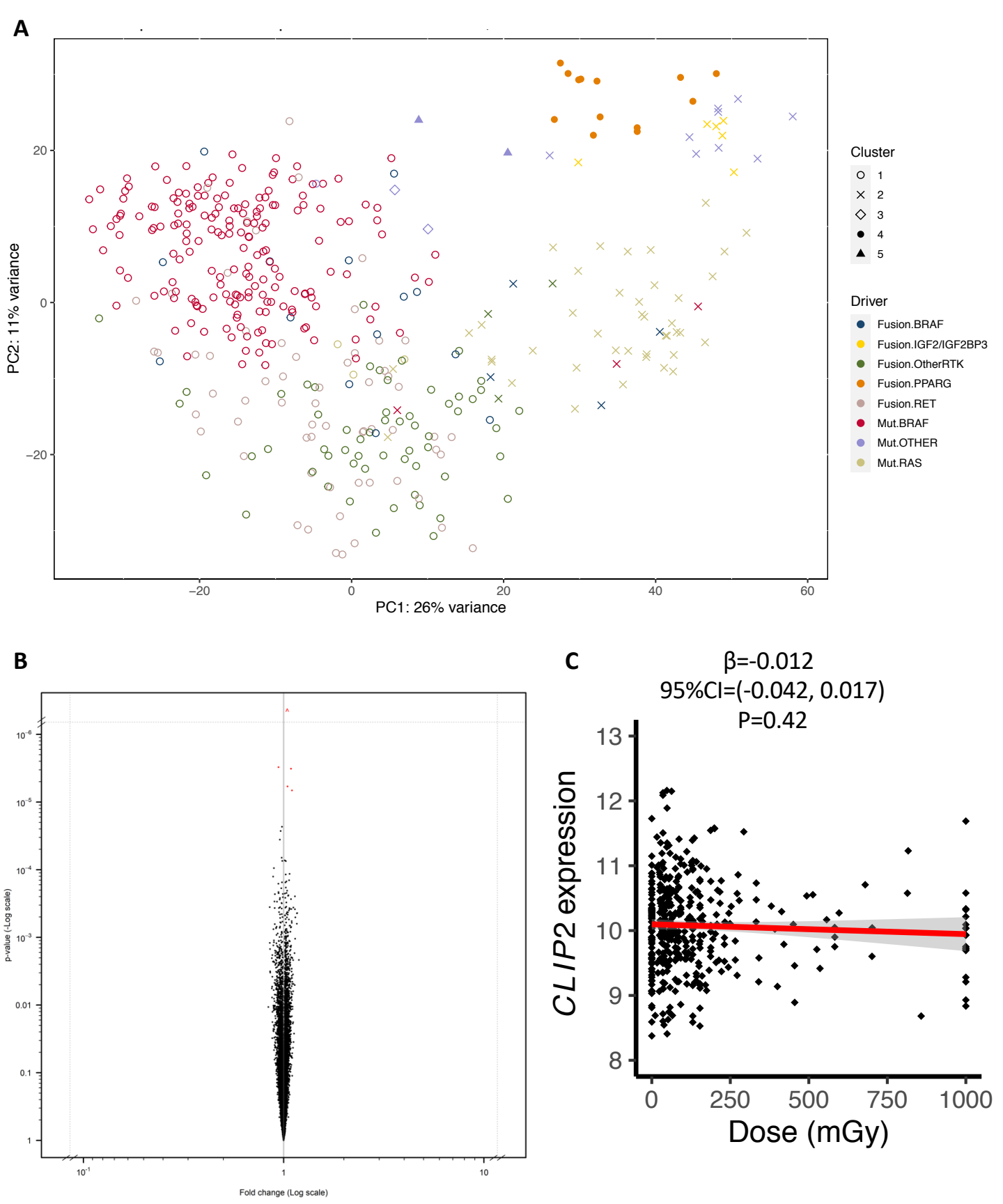
For the analyses of genomic locus-based attributes for small insertions and deletions, we conducted two analyses. First, we stratified the study population by ^{131}I-exposure category (unexposed and 1-99, 100-199, 200-499, and ≥500 mGy) and performed dichotomous logistic regression, modeling the observed versus background locus as the response variable and all genomic characteristics as independent variables. Predicted values for each of the 100,000 randomly selected loci across the genome were calculated using the results of each logistic regression, representing the estimated variant density at each locus based on the linear predictors (**Fig. S31**). Second, to test for the effect of radiation dose, we conducted multivariable linear regression models with dose as the response variable and age at surgery and all genomic characteristics as independent variables.

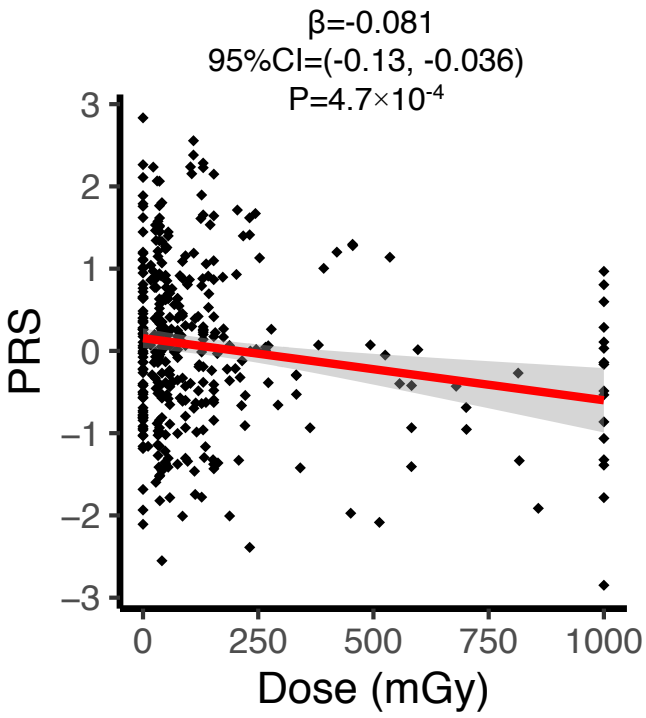

Dose (mGy)
Unexposed
1-99 mGy
100-199 mGy
200-499 mGy
≥500 mGy
Sex
Age at PTC
Age at exposure
Latency
Dose
POC
Purity
Mutation count
SBS Signatures
ID Signatures
SV count
msIndel count
SCNA count
22qDEL
BRAF
NRAS
HRAS
KRAS
TSHR
DICER1
CR2
AKT1
APC
NFE2L2
TSC1/TSC2
TP53
RET
NTRK3
BRAF
PPARG
NTRK1
ALK
IGF2/IGF2BP3
LTK
mRNA cluster
miRNA cluster
Methylation cluster
BRS mRNA
BRS miRNA
BRS methyl
TDS
ERK
Sex
female
male
Age at PTC (y)
<25
25-29
≥30
Age at exposure (y)
<5
5-9
≥10
Latency (y)
<20
20-24
≥25
POC (%)
0
1-24
25-49
50-74
≥75
Purity
high
low
Mutation count
SNV
INS
DEL
DNP
TNP
Mutation signature
SBS1
SBS2
SBS5
SBS8
SBS13
SBS18
SBS23
Indel signature
ID1
ID3
ID4
ID5
ID6
ID8
msIndel count
msDEL
msINS
SCNA count
GAIN
CNLOH
DEL
22q DEL
No
Yes
PTC driver
Mutation
Fusion
mRNA cluster
1
2
3
4
5
miRNA cluster
1
2
3
4
5
Methylation cluster
1
2
3

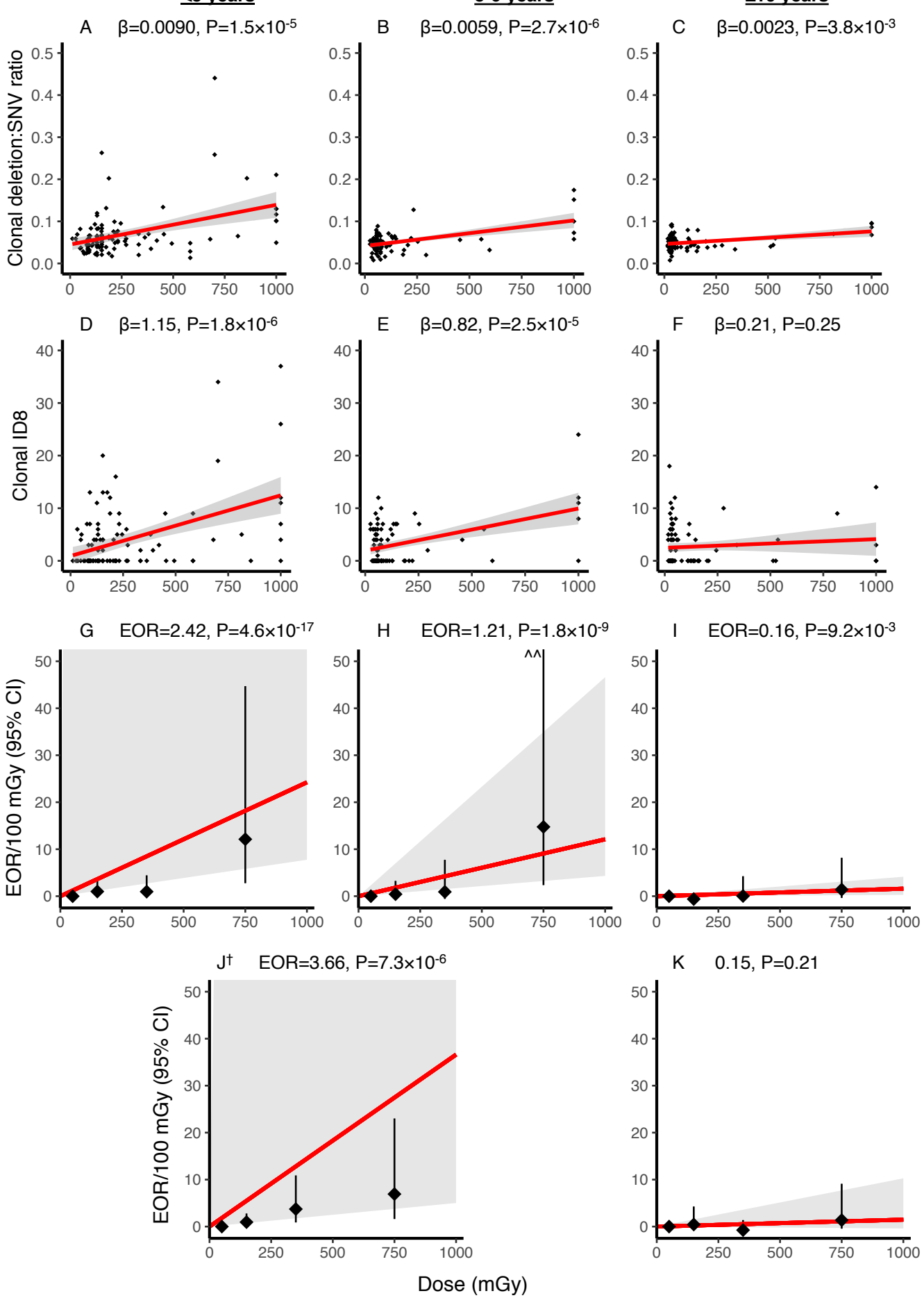

Total
Clonal
Subclonal
Small deletions
Deletion:SNV ratio
ID5
ID8
A β=2.16, P=8.0×10-9
B β=1.48, P=9.9×10-12
C β=-0.06, P=0.82
D β=0.0039, P=4.9×10-21
E β=0.0071, P=1.8×10-18
F β=0.0009, P=0.023
G β=1.07, P=1.3×10-4
H β=0.72, P=2.3×10-5
I β=-0.06, P=0.74
J β=1.15, P=7.3×10-9
K β=0.80, P=2.7×10-14
L β=0.08, P=0.49
Dose (mGy)

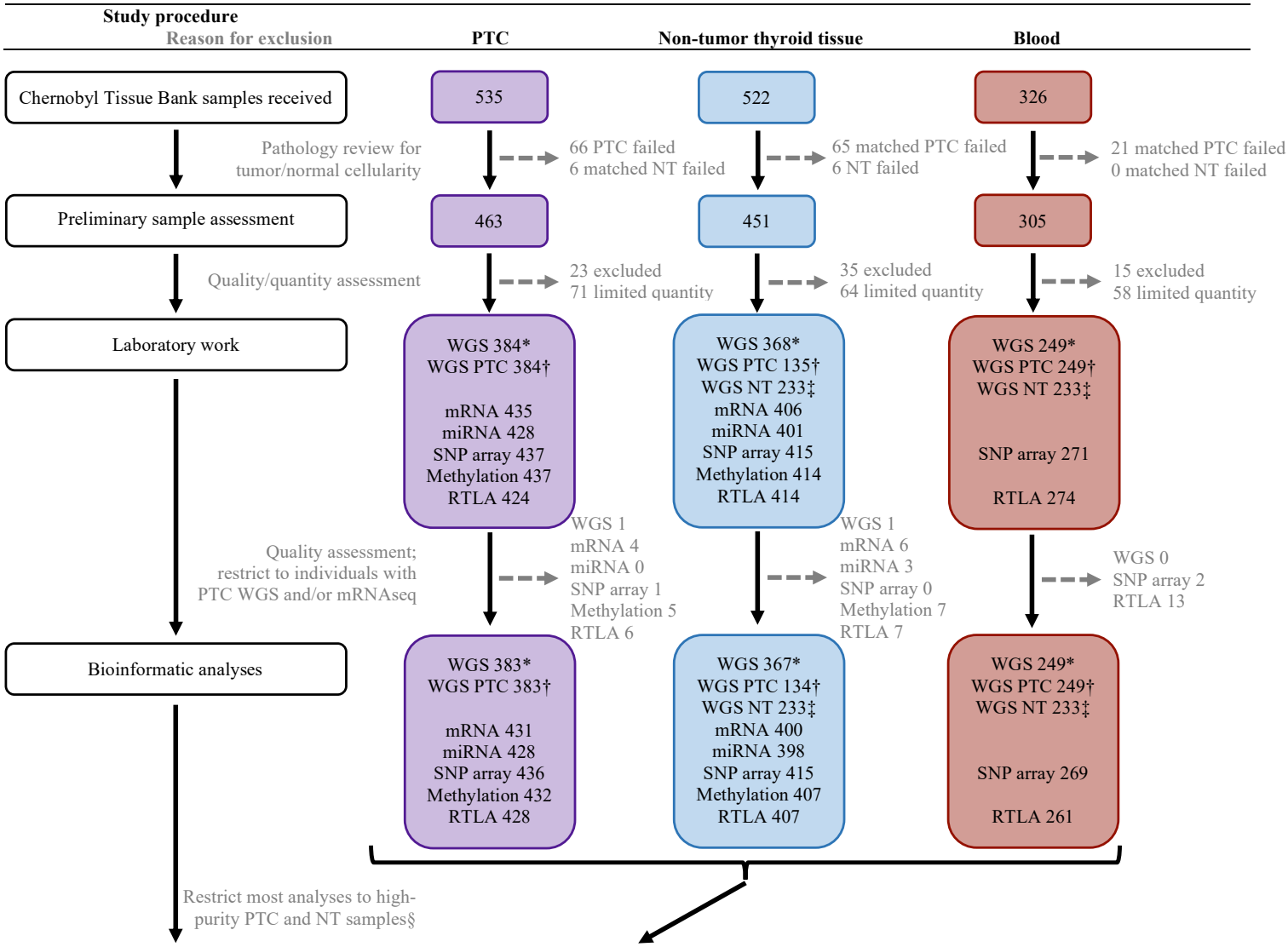

A EOR/100mGy=2.66, P=1.2×10-14
B EOR/100mGy=0.93, P=6.6×10-8
C EOR/100mGy=0.61, P=4.9×10-31
D EOR/100mGy=18.8, P=5.5×10-19
E EOR/100mGy=0.037, P=0.41
EOR/100 mGy (95% CI)
Dose (mGy)


Fusion
P heterogeneity=0.020
Mutation
P heterogeneity=0.17
1000
500
0
Dose (mGy)
Fusion, RET (n=73)
Fusion, Other RTK (n=64)
Fusion, BRAF (n=20)
Fusion, PPARG (n=13)
Fusion, IGF2/IGF2BP3 (n=6)
Mutation, BRAF (n=194)
Mutation, RAS (n=45)
Mutation, Other (n=15)


A
PC2: 11% variance
PC1: 26% variance
Cluster
1
2
3
4
5
Driver
Fusion.BRAF
Fusion.IGF2/IGF2BP3
Fusion.OtherRTK
Fusion.PPARG
Fusion.RET
Mut.BRAF
Mut.OTHER
Mut.RAS
B
p-value (-Log scale)
Fold change (Log scale)
C
β=-0.012
95%CI=(-0.042, 0.017)
P=0.42
CLIP2 expression
Dose (mGy)


β=-0.081
95%CI=(-0.13, -0.036)
P=4.7×10-4
PRS
3
2
1
0
−1
−2
−3
0
250
500
750
1000
Dose (mGy)

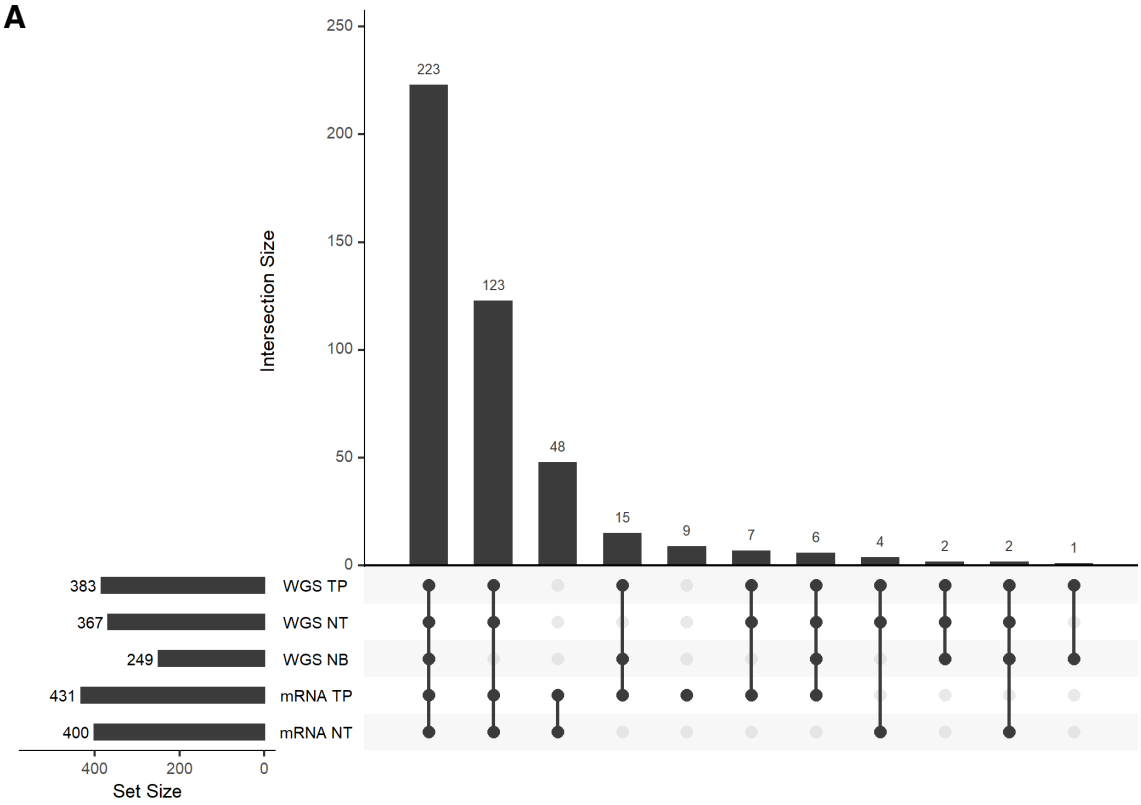

<5 years
5-9 years
≥10 years
A β=0.0090, P=1.5×10-5
B β=0.0059, P=2.7×10-6
C β=0.0023, P=3.8×10-3
Clonal deletion:SNV ratio
D β=1.15, P=1.8×10-6
E β=0.82, P=2.5×10-5
F β=0.21, P=0.25
Clonal ID8
G EOR=2.42, P=4.6×10-17
H EOR=1.21, P=1.8×10-9
^^
I EOR=0.16, P=9.2×10-3
EOR/100 mGy (95% CI)
J† EOR=3.66, P=7.3×10-6
K 0.15, P=0.21
EOR/100 mGy (95% CI)
Dose (mGy)



	Data source for each analysis						
Statistical analyses	WGS	mRNA-seq	miRNA-seq	SNP array	Methylation	RTL	n
PTC driver±	X	X					440
Mutations & mutational patterns	X						356
Structural variants	X						356
SCNAs	X						355
PTC mRNA expression patterns		X					409
PTC miRNA expression patterns			X				405
PTC methylation patterns					X		410
Relative telomere length PTC						X	406
Relative telomere length NT						X	399
Relative telomere length NB						X	261
Germline variation	X						383
Population substructure±				X			440
Non-tumor thyroid tissue	X						218

Fig. S1. Schematic of study procedures and sample sizes for laboratory work and statistical analyses

Gray font indicates samples for exclusion. Reasons for exclusion and thresholds for exclusion metrics are provided in the methods.
Abbreviations: normal blood (NB), non-tumor thyroid tissue (NT), papillary thyroid carcinoma (PTC), relative telomere length (RTL).
* WGS: total count of WGS for that tissue type.
† WGS PTC: count of that tissue type utilized for analyses of PTC-matched normal tissue pairs.
‡ WGS NT: count of that tissue type utilized for analyses of NT-matched blood pairs.
§ Most statistical analyses were restricted to high purity samples, defined as those with tumor purity >20% and no evidence of tumor contamination in the normal tissue.
± Analyses that include all individuals with available data (not restricted to high purity samples).



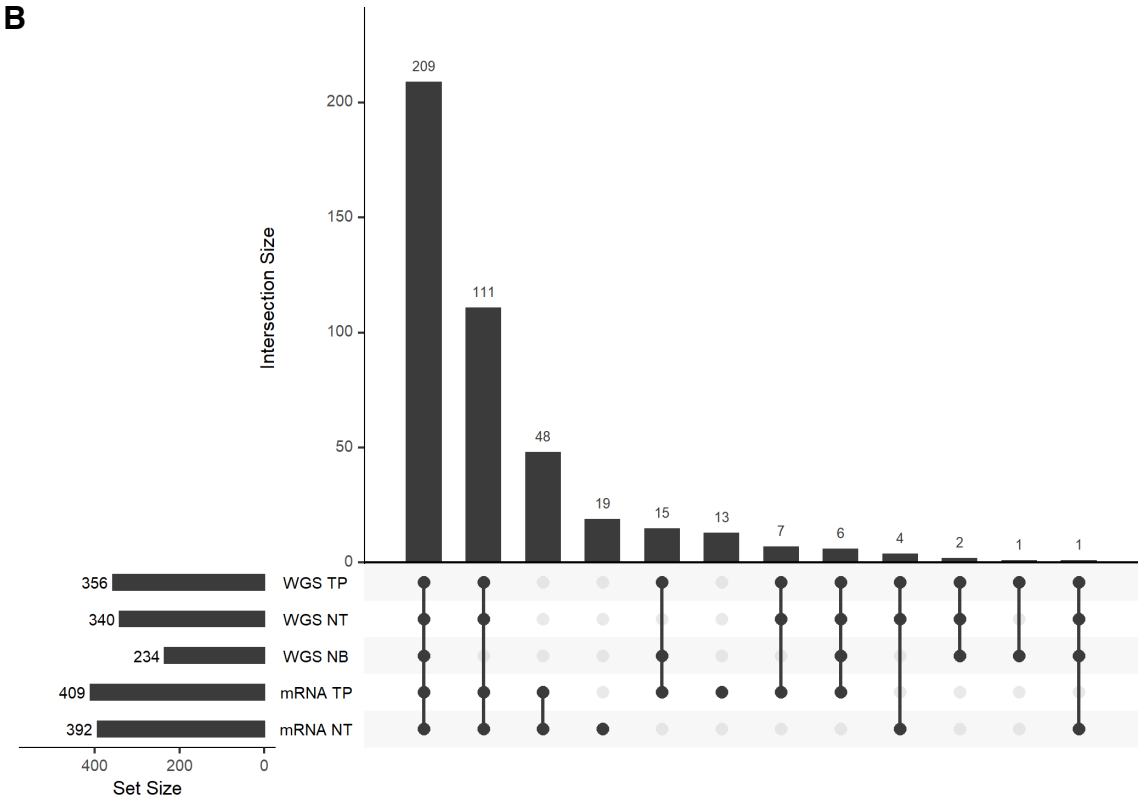


Fig. S2. Distribution of samples passing quality control metrics for WGS and/or mRNAseq.
(**A**) All samples. (**B**) Restricted to high purity samples (tumor purity >20% and no evidence of tumor contamination in the normal tissue).

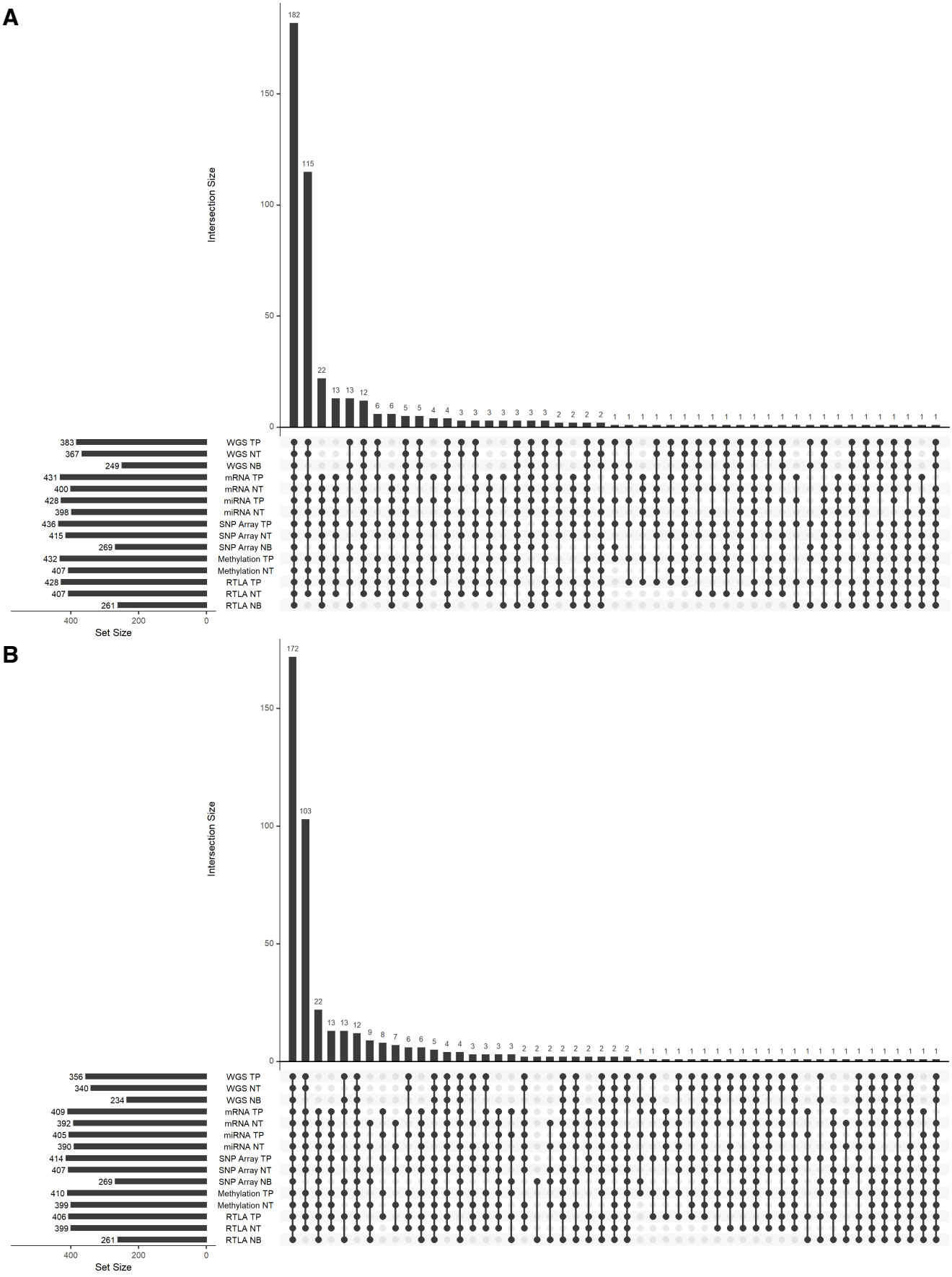


Fig. S3. Distribution of samples passing quality control metrics across all platforms.
(**A**) All samples. (**B**) Restricted to high purity samples (tumor purity >20% and no evidence of tumor contamination in the normal tissue).

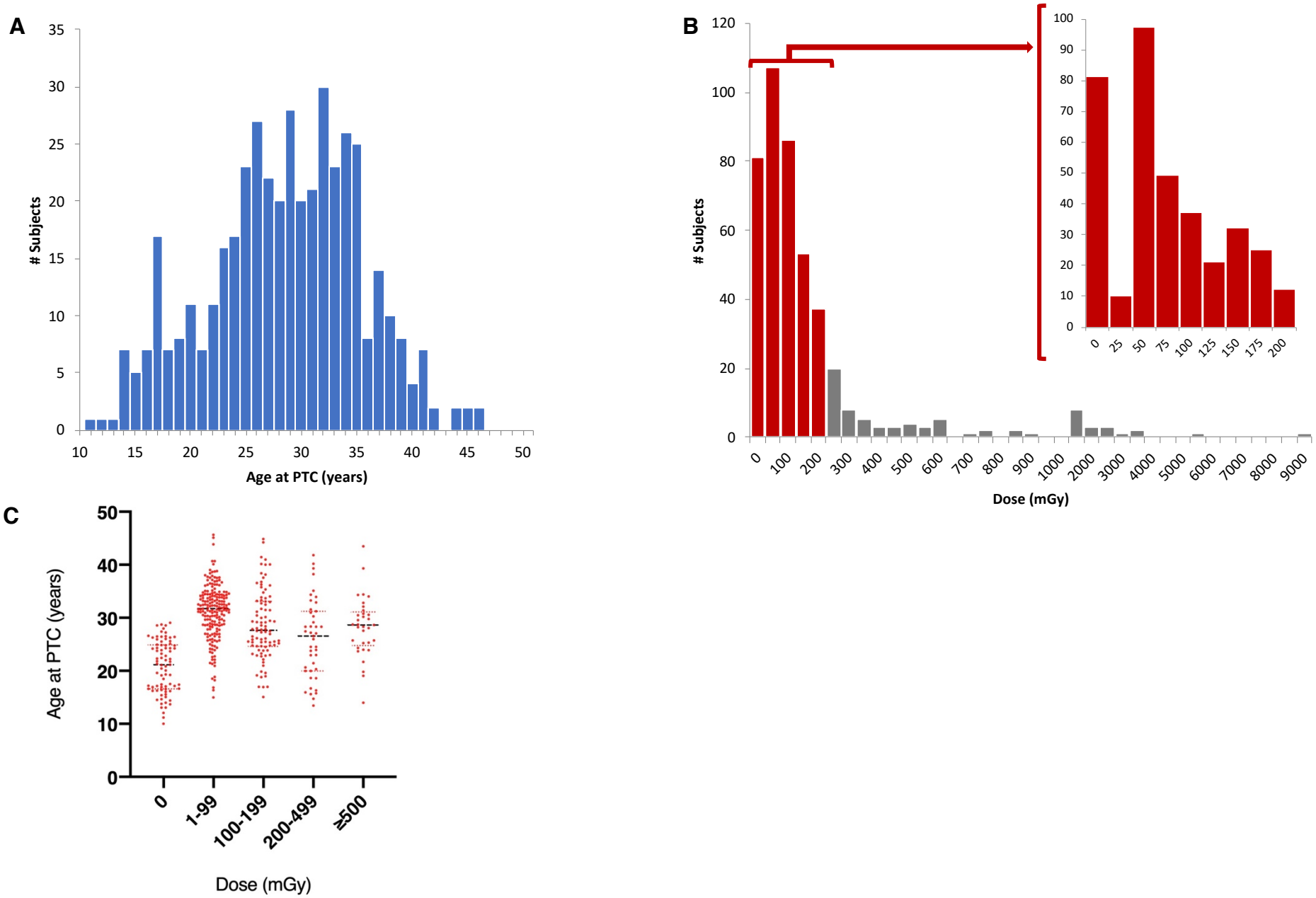


Fig. S4. Detailed distributions of age at PTC diagnosis and radiation dose from ^{131}I exposure.
(**A**) Age at PTC diagnosis. (**B**) Radiation dose from ^{131}I exposure, with an inset panel for doses <200 mGy because this dose range accounts for the majority of our study population. (**C**) Joint distribution of age at PTC and radiation dose to demonstrate slight differences in the age distribution among dose groups.

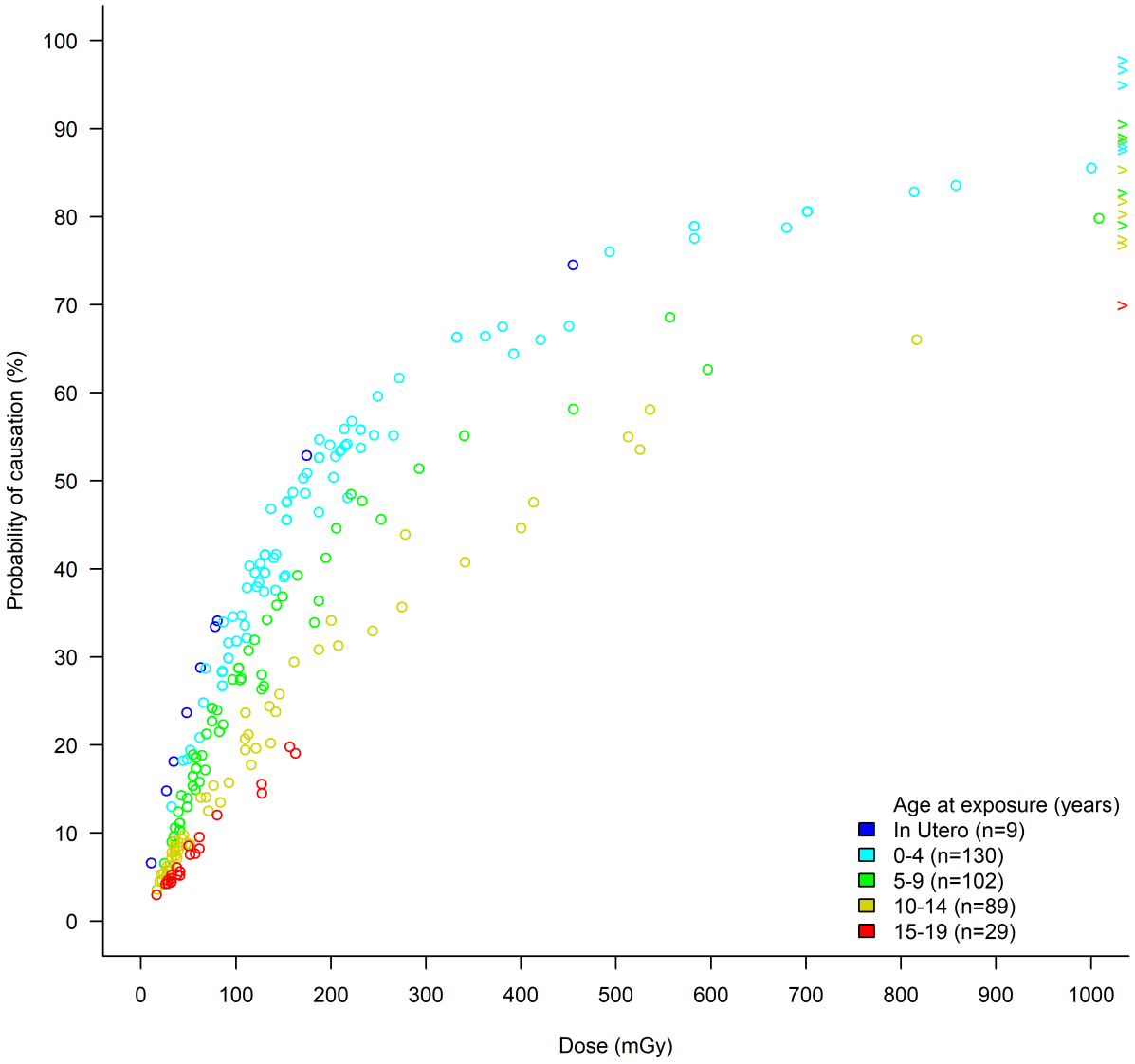


Fig. S5. Relationship between radiation dose and the probability of causation, by age at ^{131}I exposure, for the 440 individuals in our study population.

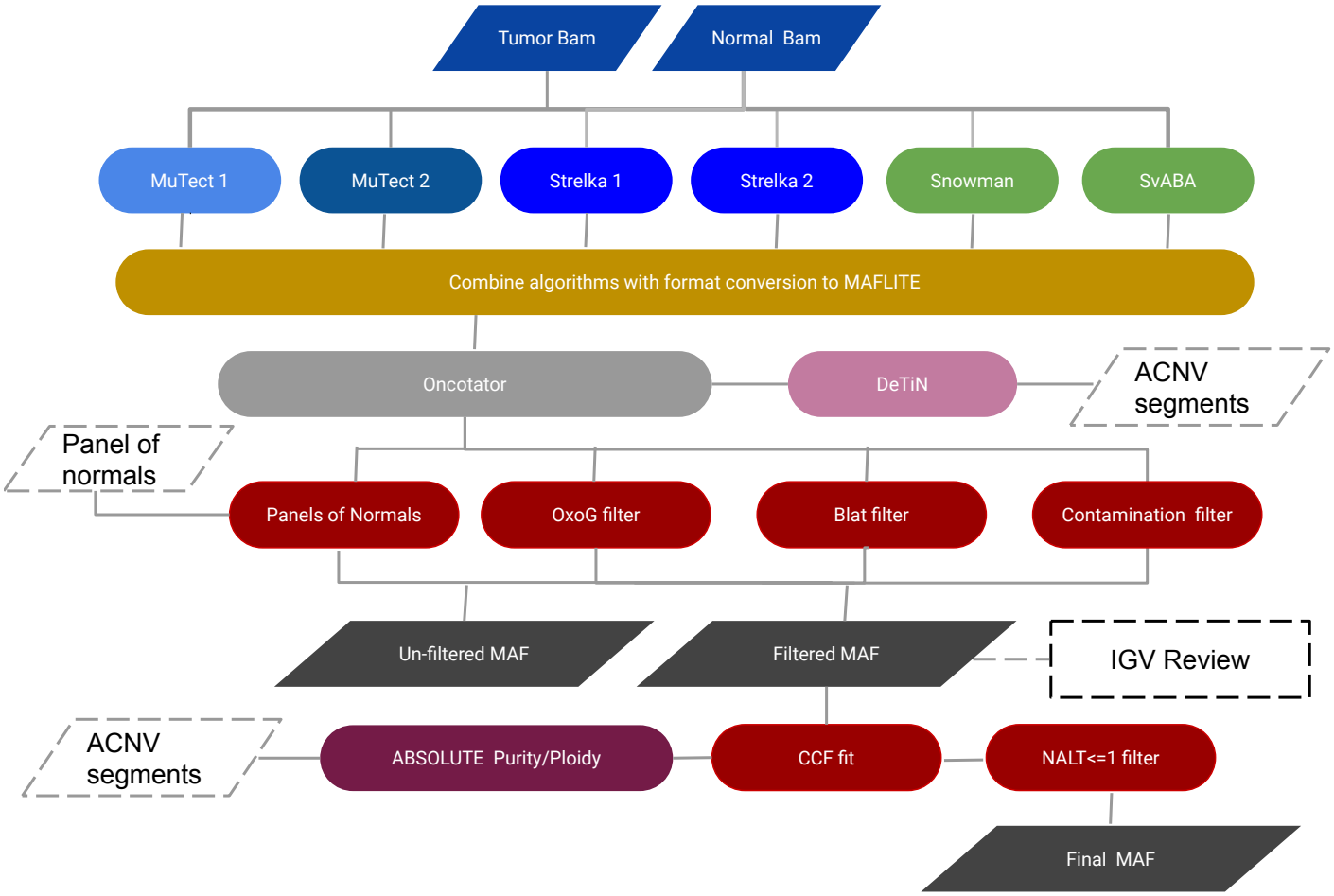


Fig. S6. Mutation analysis pipeline

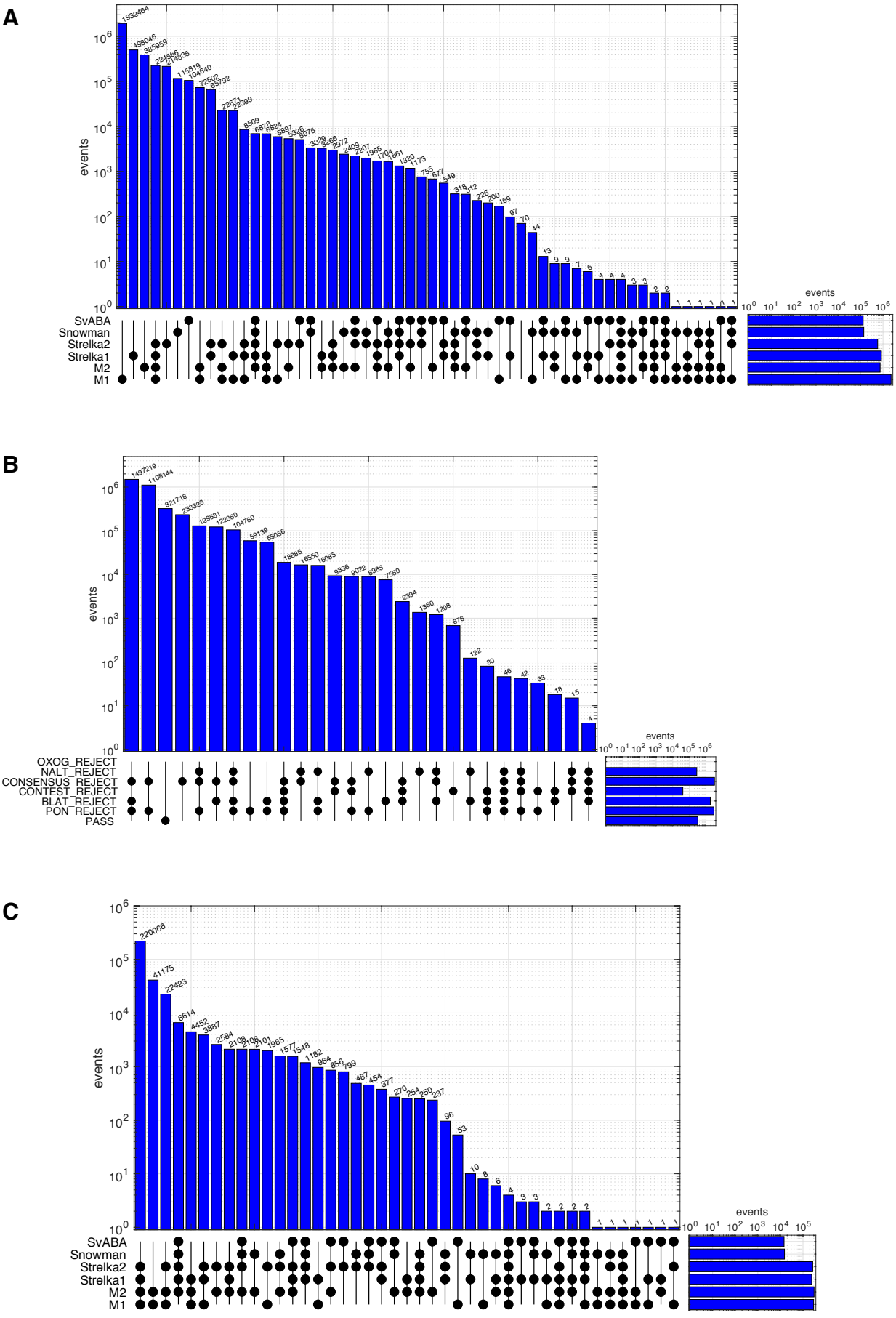


Fig. S7. Distribution of mutation filters for PTC samples.
(**A**) Pre-filtered calls. (**B**) Filtered calls. (**C**) Final mutational call algorithm.

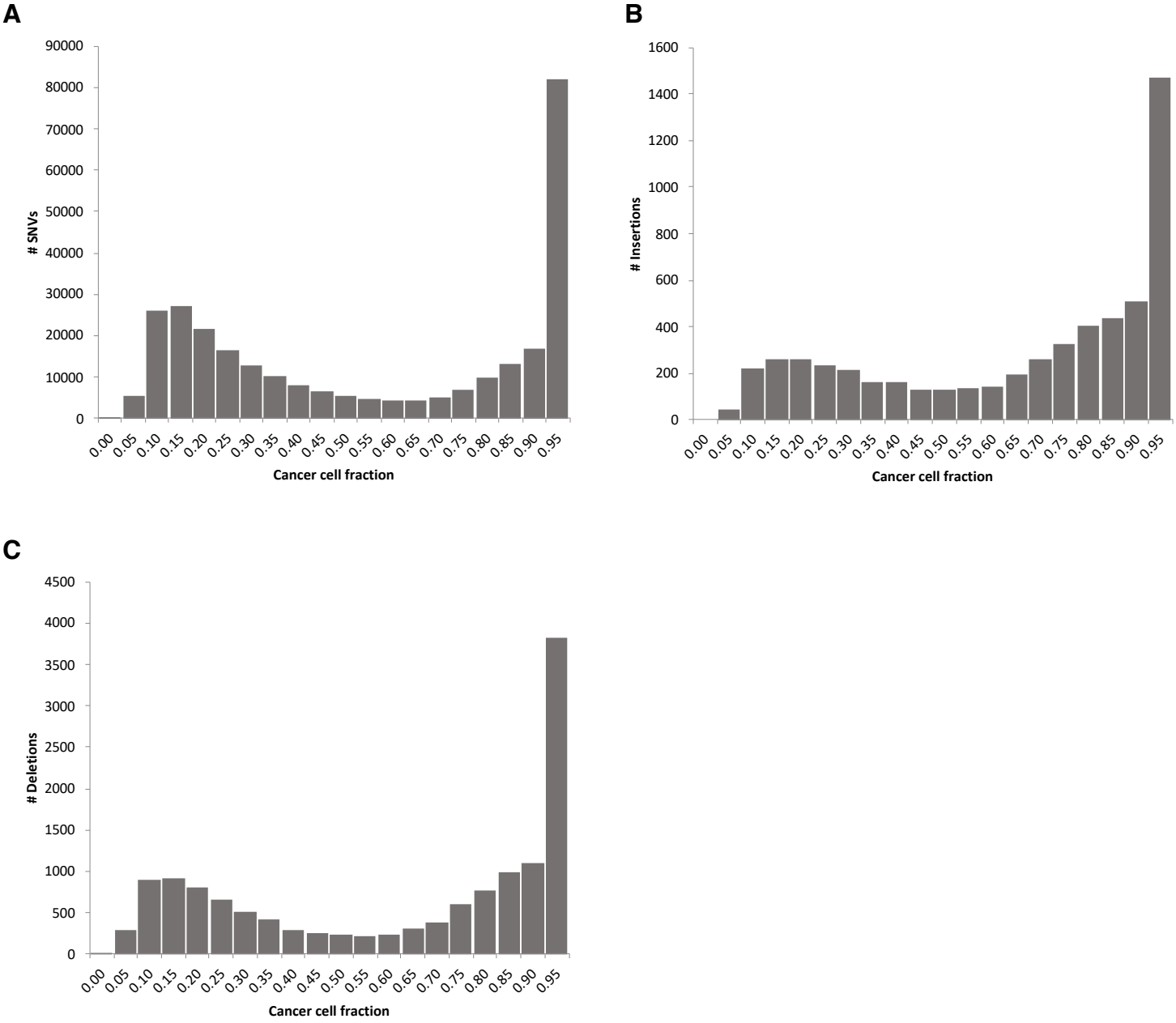


Fig. S8. Distribution of cancer cell fraction.
(A) SNVs (n=286,534). (B) Insertions (n=5699). (C) Deletions (n=13,706). Counts are restricted to n=356 individuals with high purity samples. As described in the Materials and Methods, cancer cell fraction reflects the variant allele fraction, accounting for the purity of the tumor sample. Clonal mutations were defined as those with cancer cell fraction >0.9, whereas subclonal mutations were defined as those with cancer cell fraction <0.6. Mutations with cancer cell fraction 0.6-0.9 had undetermined clonality. Counts by mutation type and clonality are provided in Table S3.

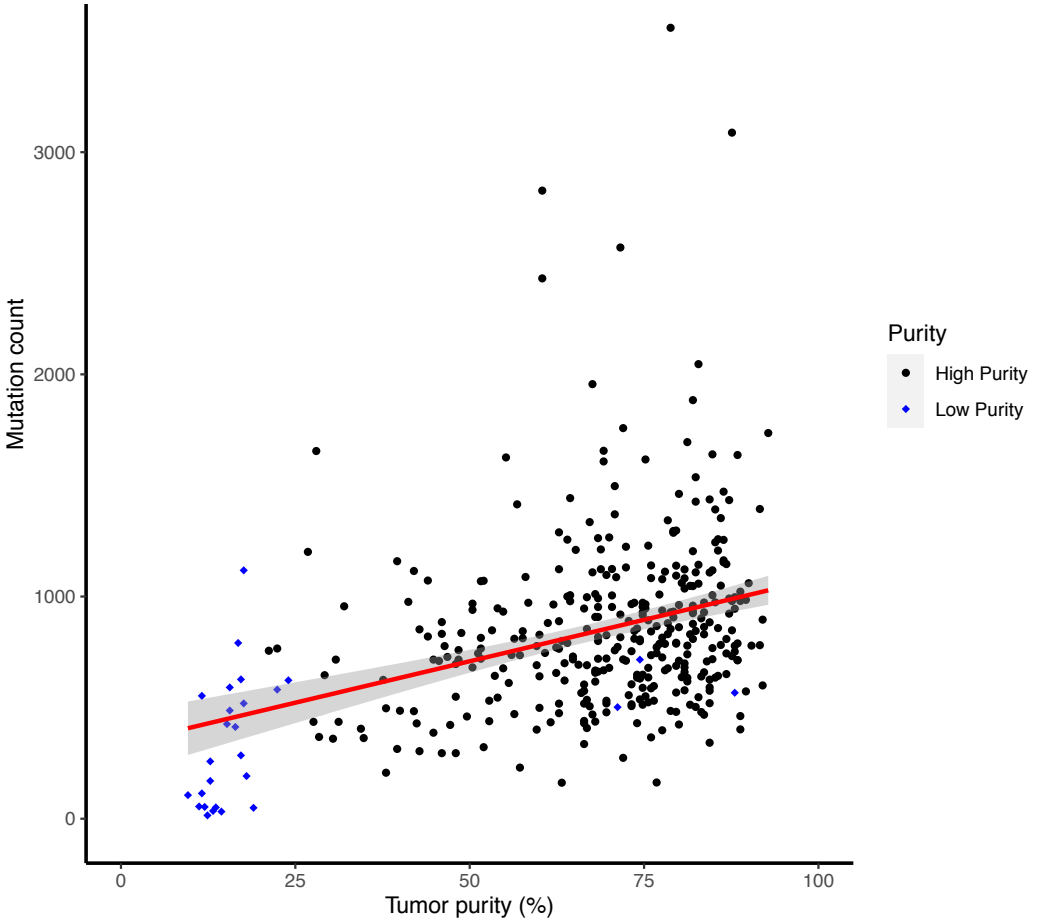


Fig. S9. Distribution of the total number of single nucleotide variants by tumor purity for 383 tumors with whole genome sequencing data

High purity samples were defined as those with tumor purity >20% and no evidence of tumor contamination in the normal tissue (n=356). Individuals with low purity samples (n=22 with tumor purity <20%; n=2 low purity tumors where outlier variants were inflating the purity estimate; n=3 with evidence of tumor contamination in the normal tissue) were excluded from selected analyses (Fig. S1).

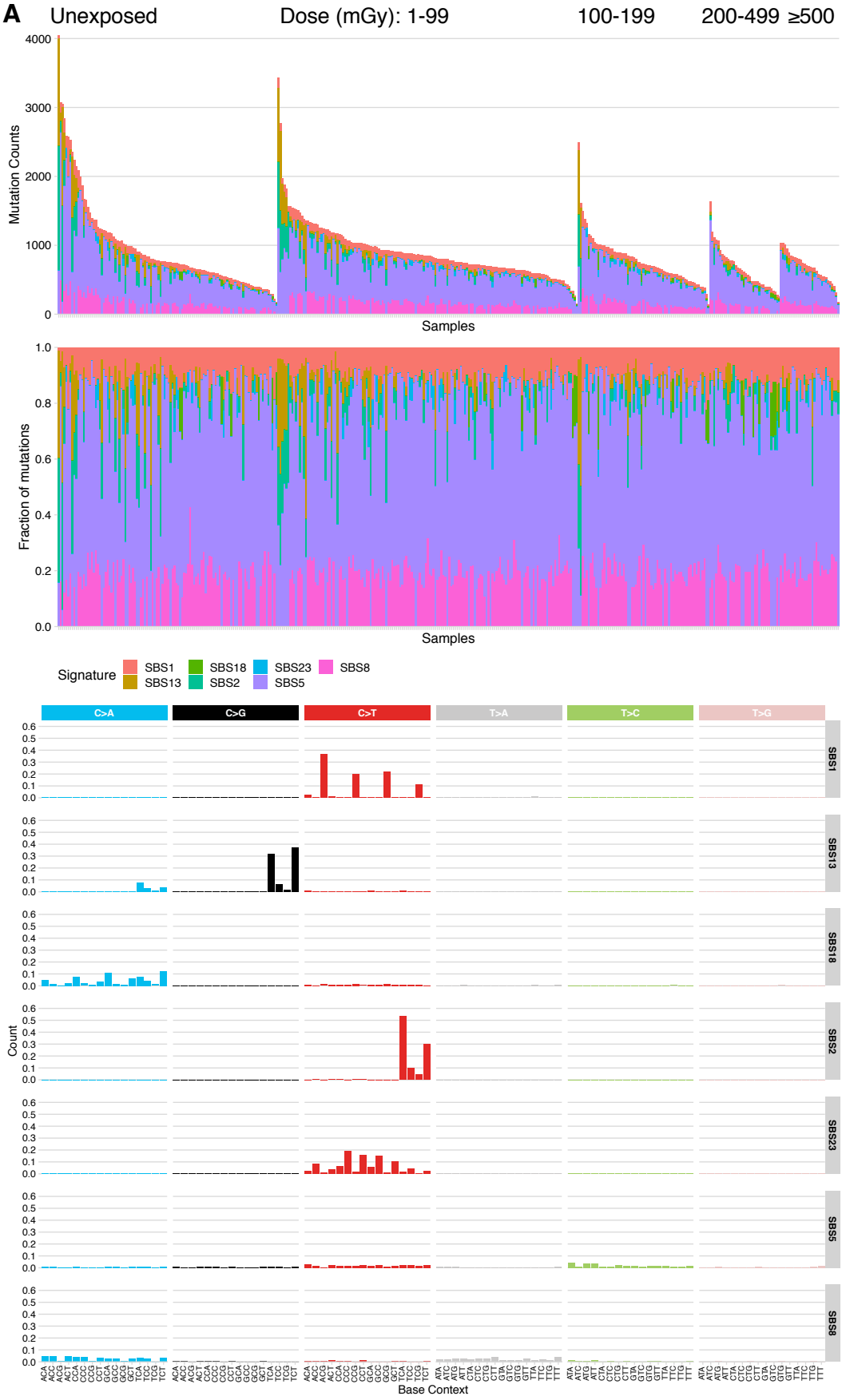

A
Unexposed
Dose (mGy): 1-99
100-199
200-499
≥500
Mutation Counts
Samples
Fraction of mutations
Samples
Signature
SBS1
SBS13
SBS18
SBS2
SBS23
SBS5
SBS8
C>A
C>G
C>T
T>A
T>C
T>G
Count
Base Context

Fig. S10. Distribution of SBS signatures by radiation dose from ^{131}I exposure and mutation count.
(**A**) Decomposed into known COSMIC signatures. (**B**) Identified de novo.


A
Unexposed
Dose (mGy): 1-99
100-199
200-499
≥500
Mutation Counts
Samples
Fraction of mutations
Samples
Signature
ID1
ID3
ID4
ID5
ID6
ID8
1bp Del, C
1bp Del, T
1bp Ins, C
1bp Ins, T
2bp Del in Repeat
3bp Del in Repeat
4bp Del in Repeat
5+bp Del in Repeat
2bp Ins in Repeat
3bp Ins in Repeat
4bp Ins in Repeat
5+bp Ins in Repeat
2bp Del w/ Microhomology
3bp Del w/ Microhomology
4bp Del w/ Microhomology
5+bp Del w/ Microhomology
Proportion
Motif Length or Count

Fig. S11. Distribution of ID signatures by radiation dose from ^{131}I exposure and mutation count.
(**A**) Decomposed into known COSMIC signatures. (**B**) Identified de novo.

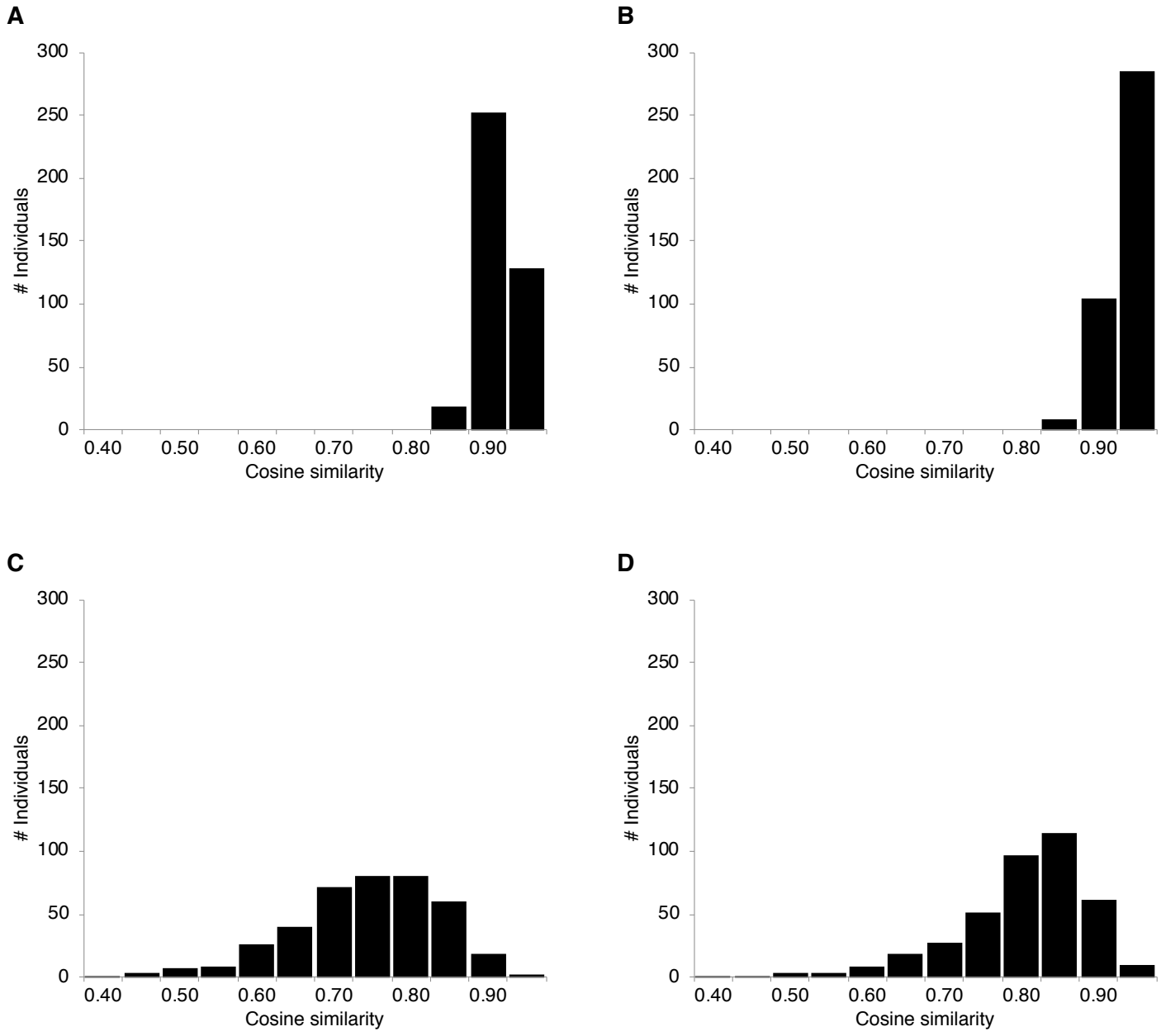


Fig. S12. Distribution of cosine similarities between actual mutation counts and those extracted using SigProfiler.

(**A**) SBS decomposition, Mean=0.94. (**B**) SBS de novo, Mean=0.96. (**C**) ID decomposition, Mean=0.77. (**D**) ID de novo, Mean=0.83.

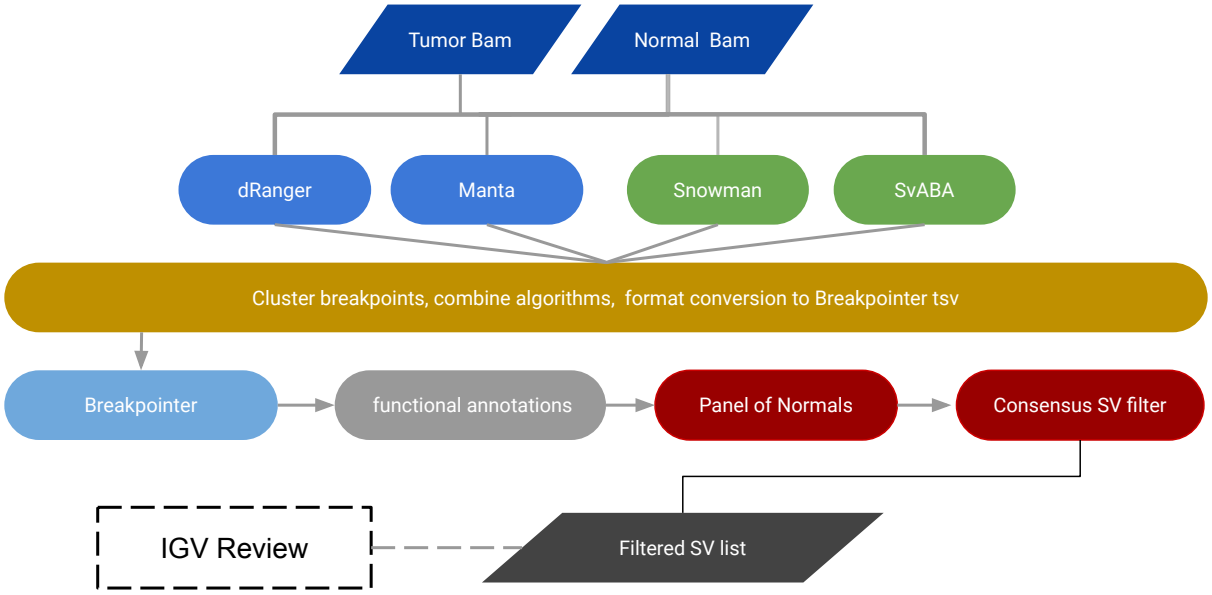


Fig. S13. Structural variant calling pipeline.

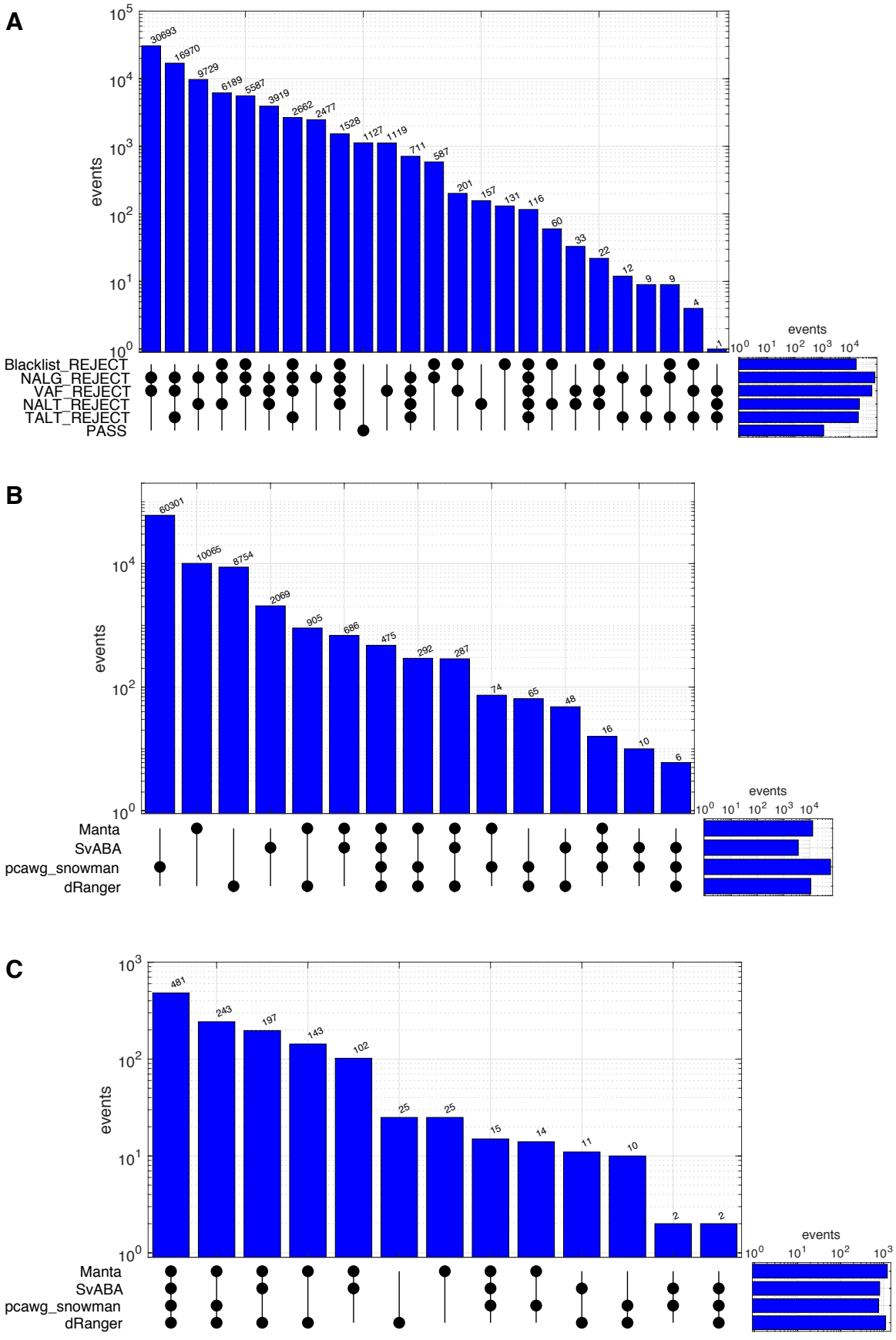


Fig. S14. Distribution of structural variant filters and calling algorithms.
(**A**) Filters. (**B**) Raw calls. (**C**) Final calling algorithm.

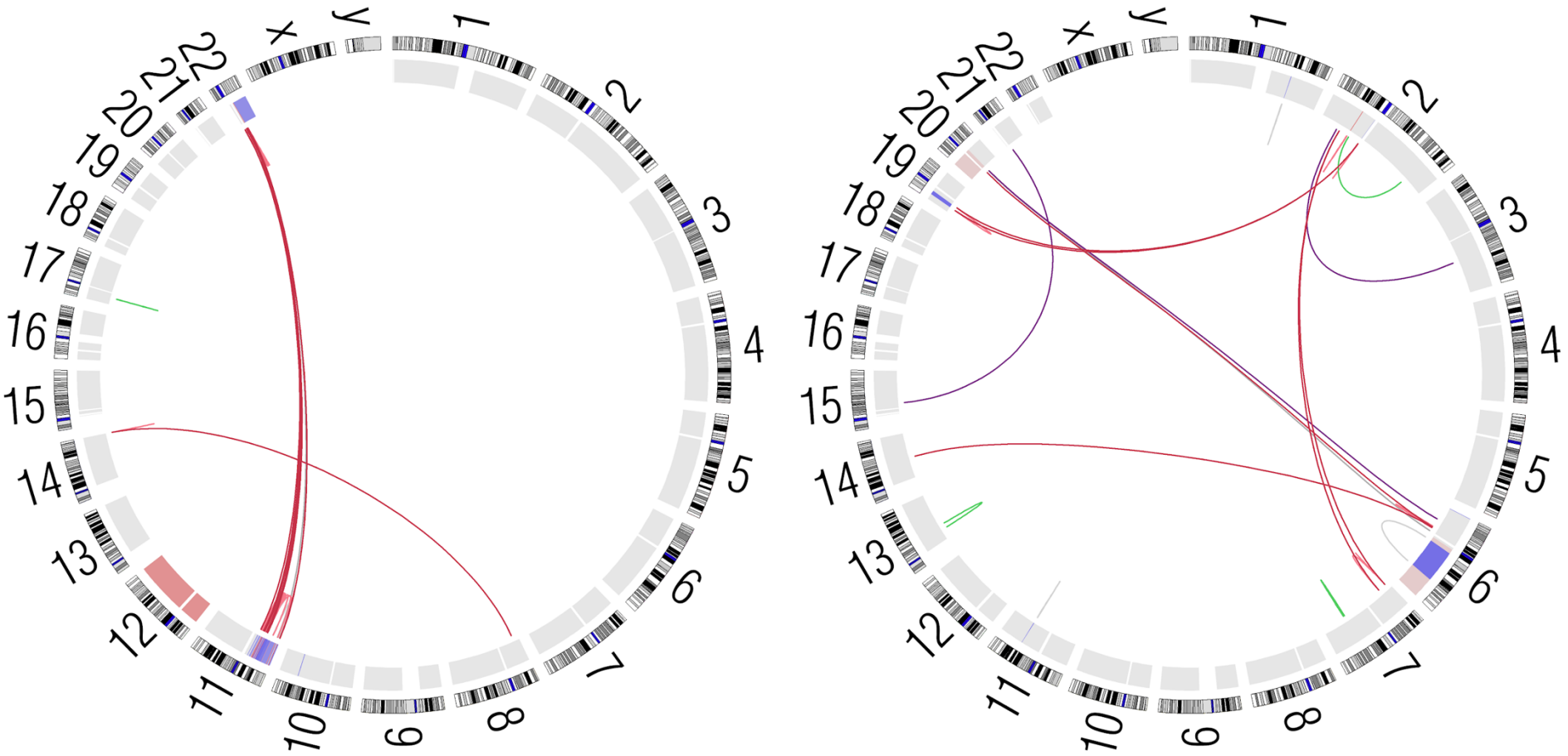


Fig. S15. Circos plots depicting tumors with >10 structural variants each.
The tumor depicted on the left also had evidence for chromothripsis and had ≥20 somatic copy number alterations. The outer ring shows cytoband structure within each chromosome. The inner ring shows the copy ratio (blue deletion, red amplification) across each chromosome. Red lines indicate SVs with breakpoints are clustered within 50kb of another SV to form chains of events. Purple lines indicate balanced translocations. Green indicate SVs forming inversions in which both inversion breakpoints were detected. Gray lines are isolated SVs with breakpoints more than 50 kb from other SVs.

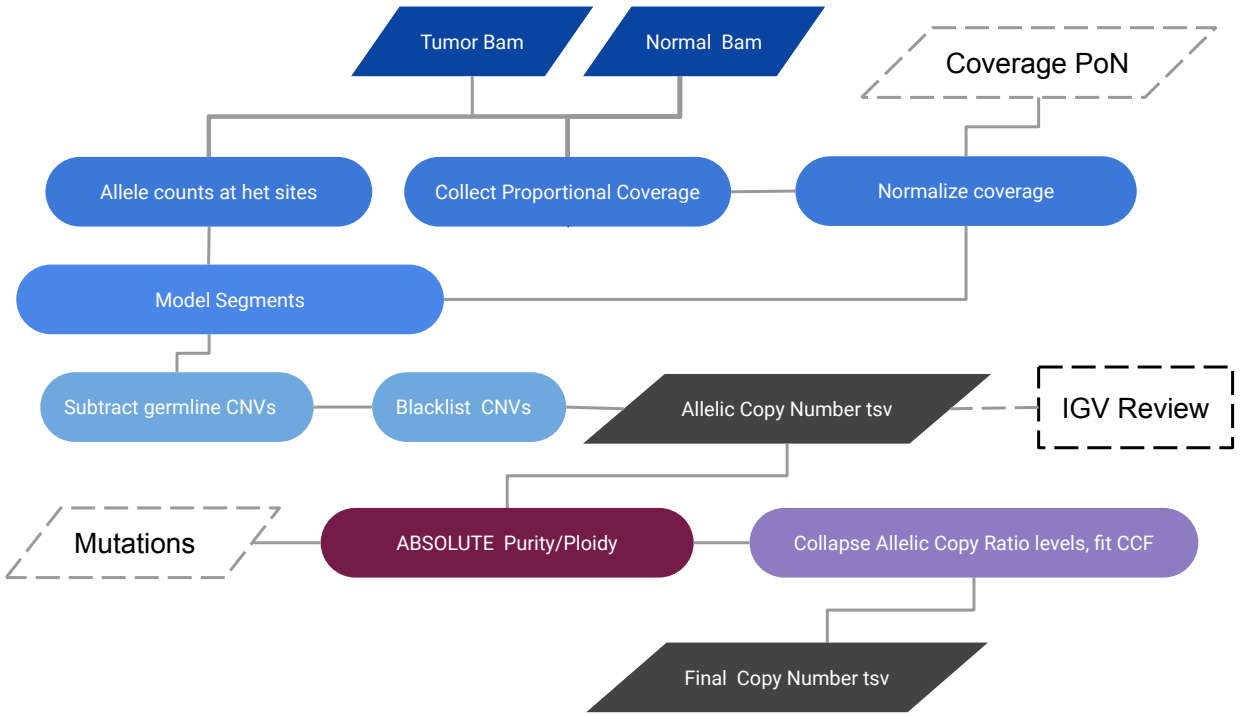


Fig. S16. Somatic copy number alteration calling pipeline.

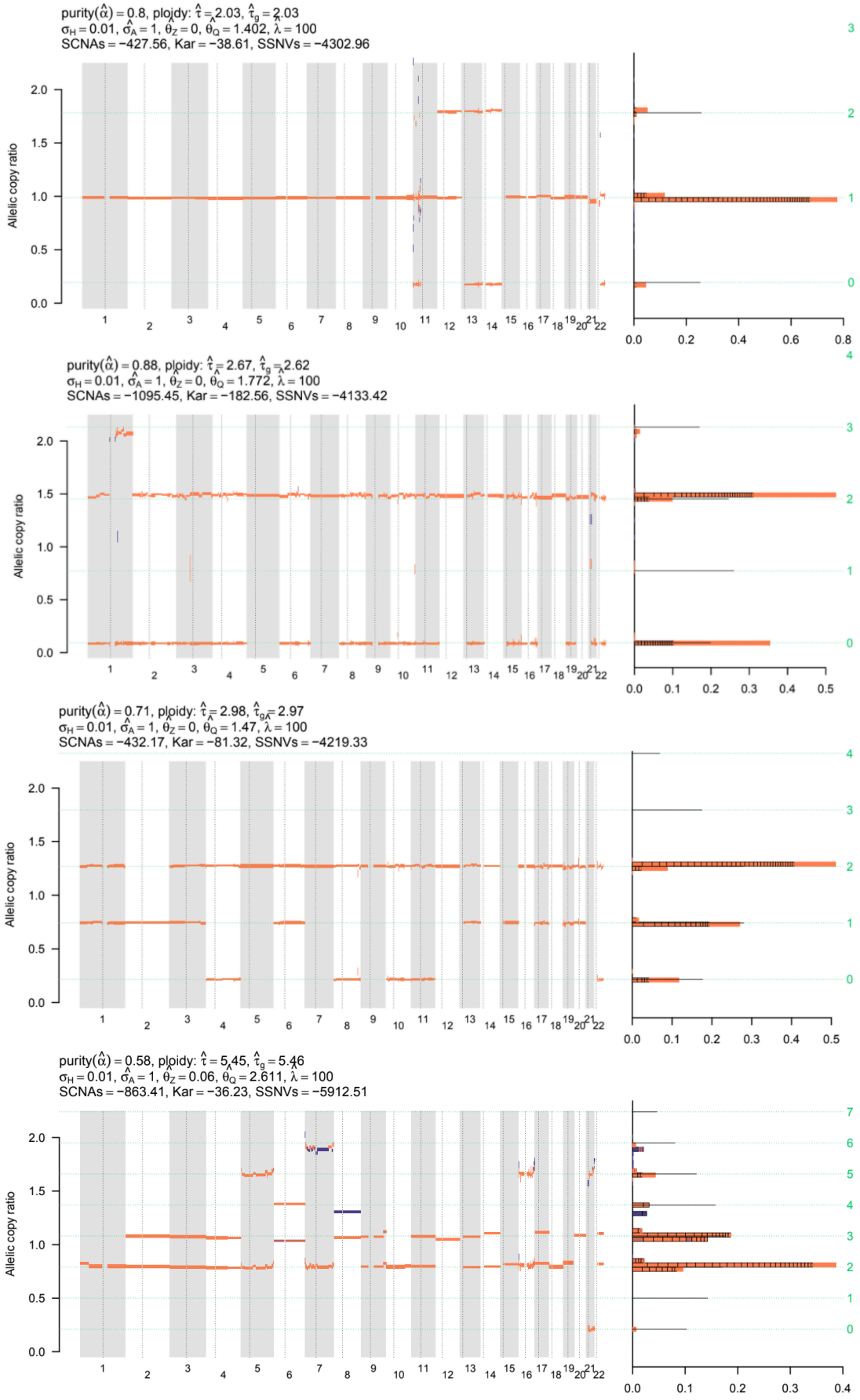


Fig. S17. Tumors with ≥20 somatic copy number alterations.
The tumor depicted in the top panel also had evidence for chromothripsis and had >10 structural variants.

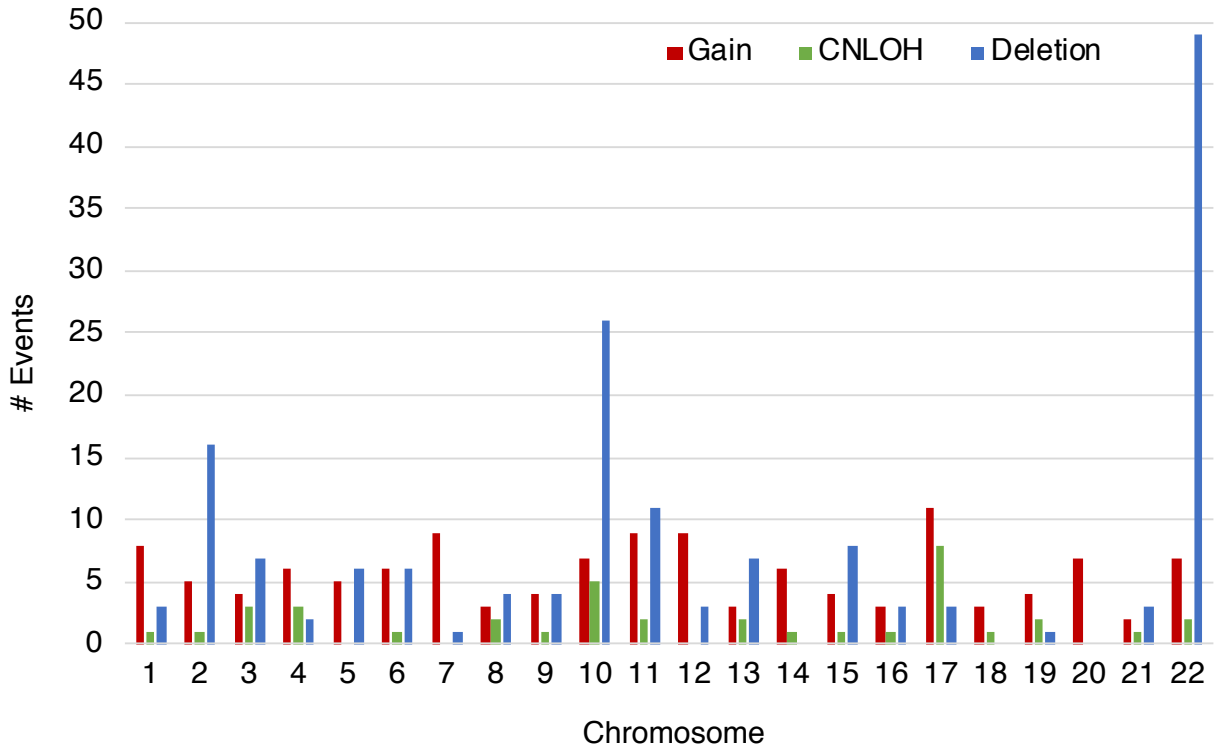


Fig. S18. Distribution of SCNAs, by chromosome number and SCNA type.

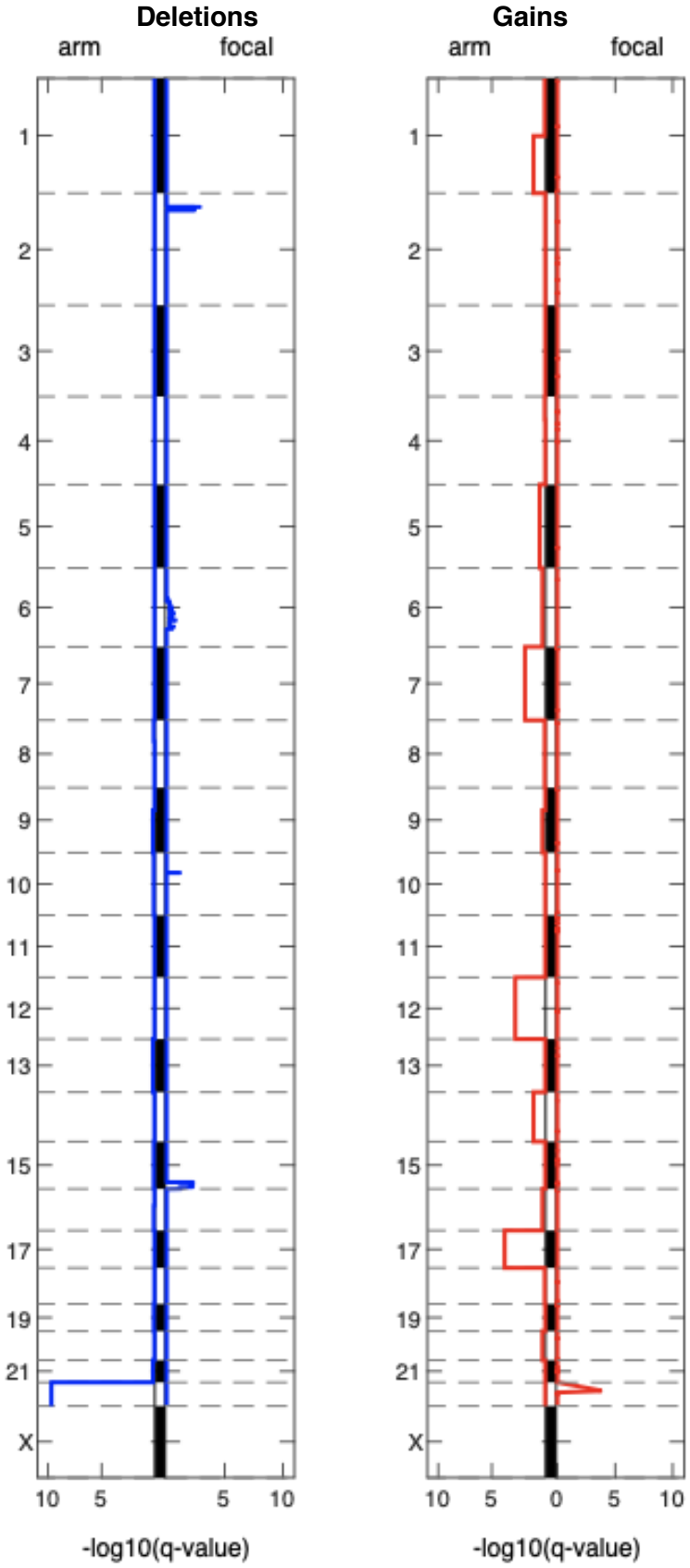


Fig. S19. Results from GISTIC2.0 for the 383 tumors with WGS data.
The horizontal scale shows the Benjamini-Hochberg corrected FDR (q-value) estimates for arm-level (SCNAs that cover more than half of an arm) and focal SCNAs. The FDR for 22q deletion is shown truncated to 1.0×10^{-10}. FDR<0.1 can be considered as significantly recurrent relative to the null model.

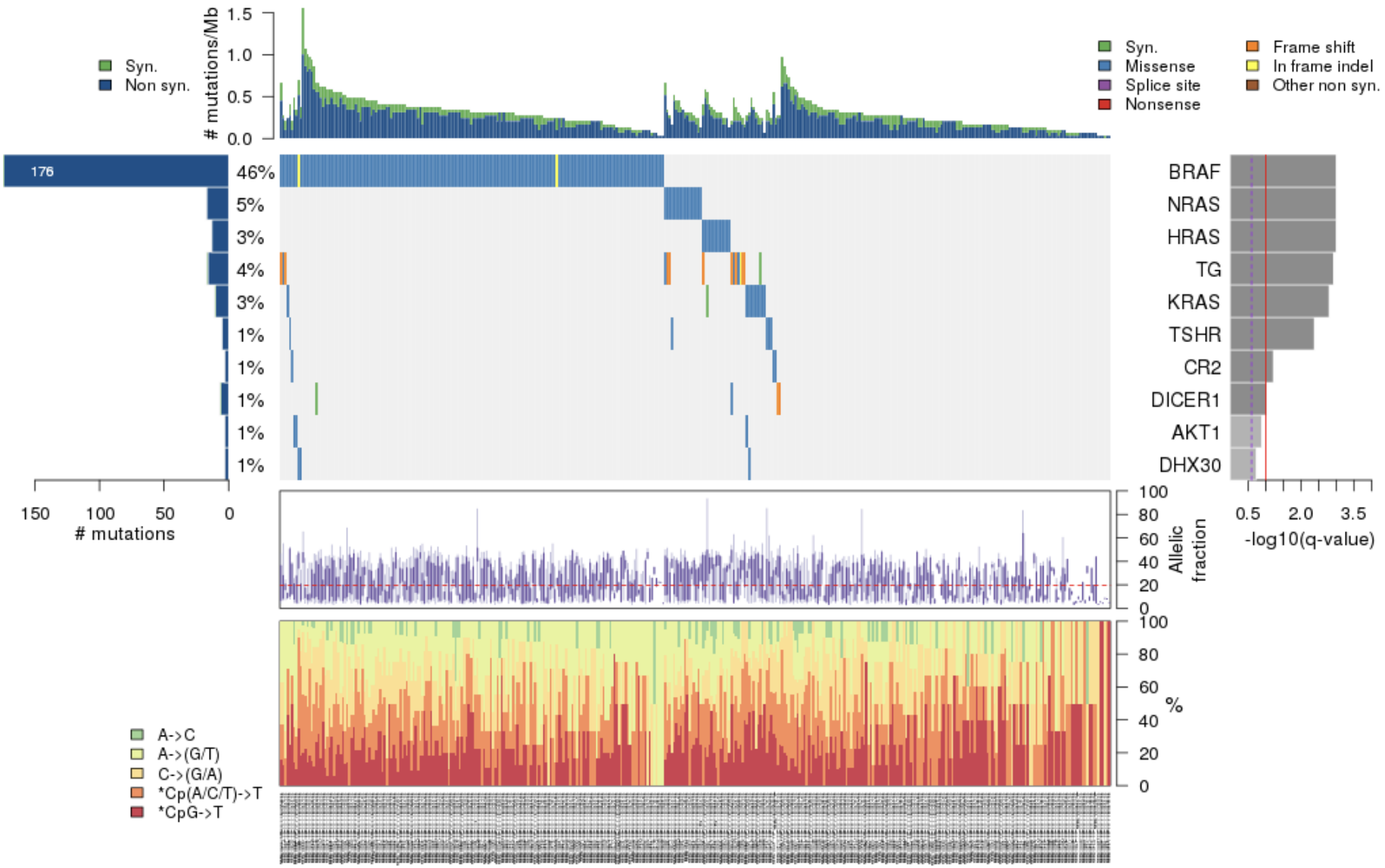


Fig. S20. Results from MutSig2CV for the 383 tumors with WGS data.
The MutSig2CV FDR is shown on the right side, with *BRAF, NRAS, HRAS, KRAS, TSHR, CR2, DICER1*, and *TG* found as significantly recurrent with an FDR<0.1. *TG* was dismissed as a driver gene due to a preponderance of non-coding mutations.

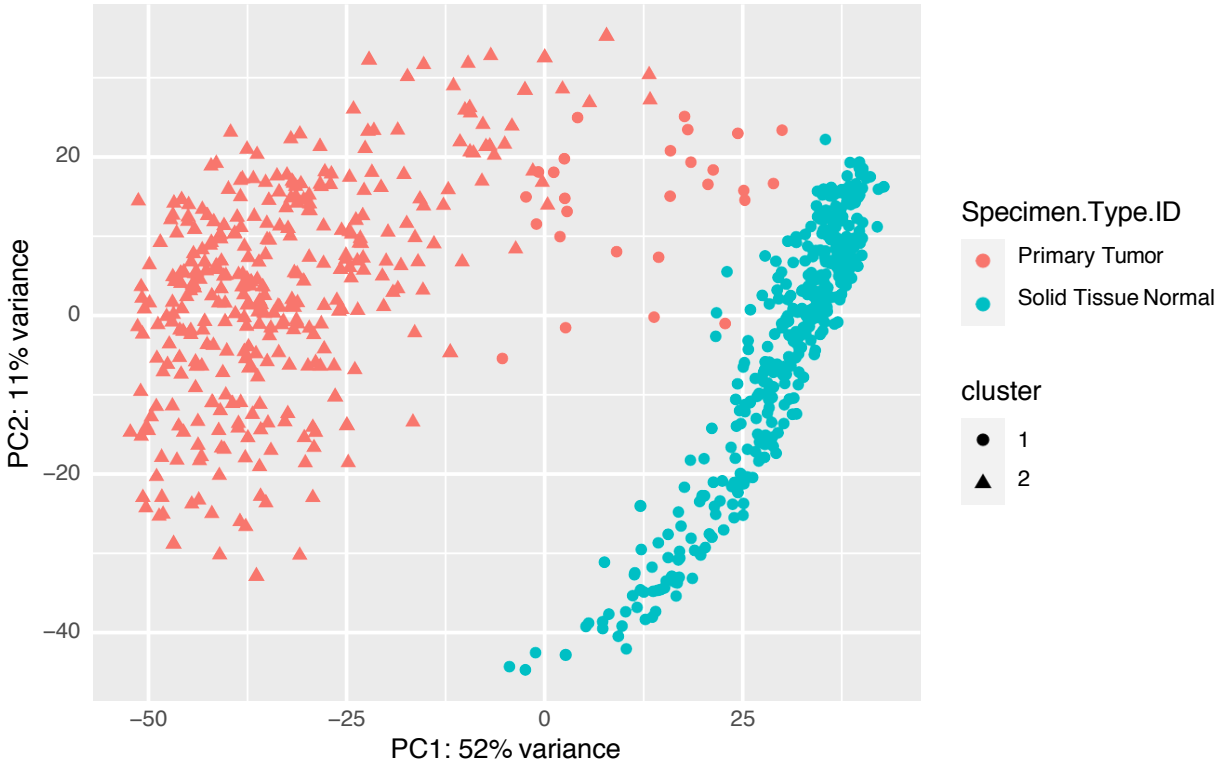


Fig. S21. Unsupervised mRNA clustering analyses of paired PTC and non-tumor thyroid tissue revealed two clusters strongly correlated with tissue type.

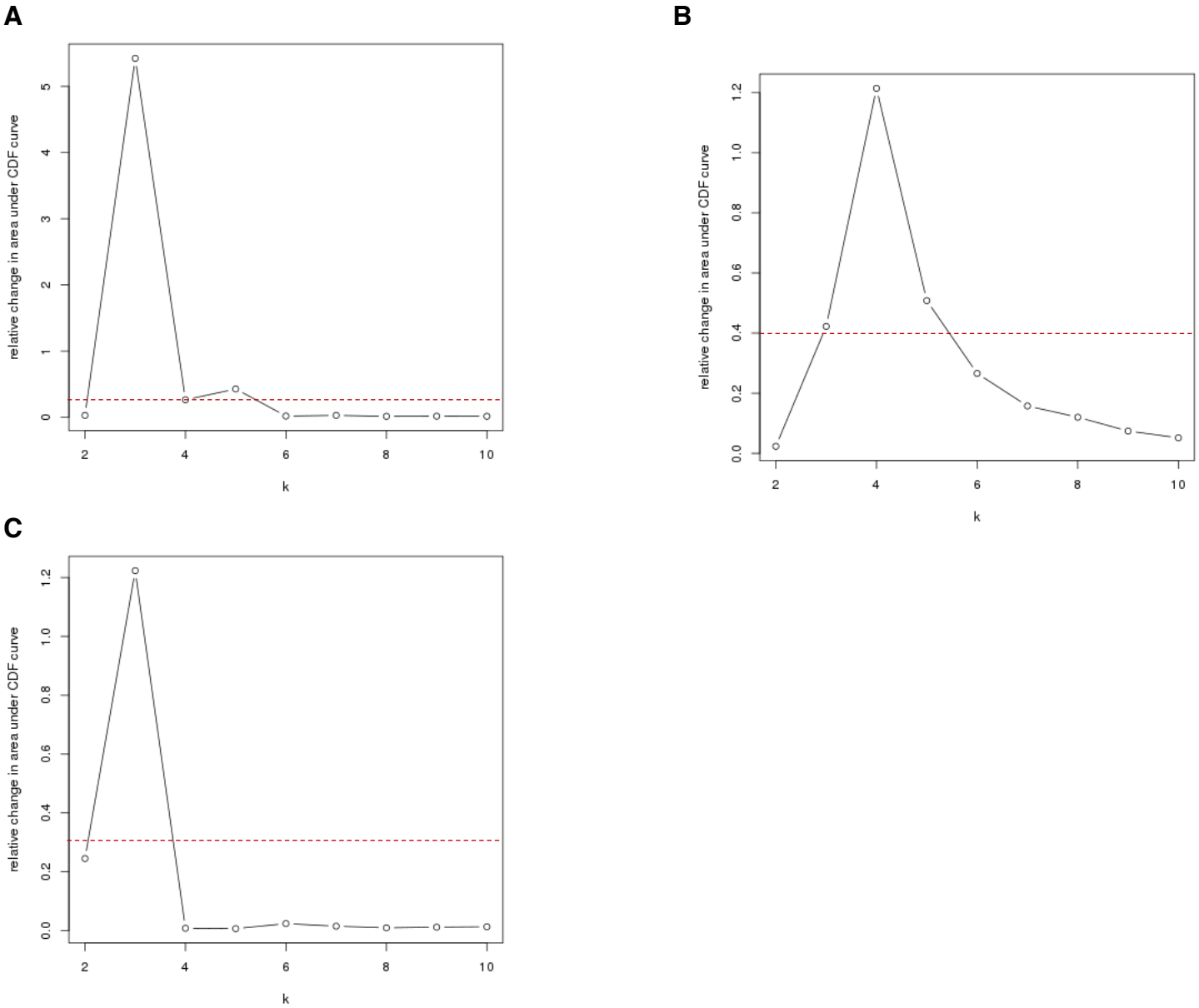
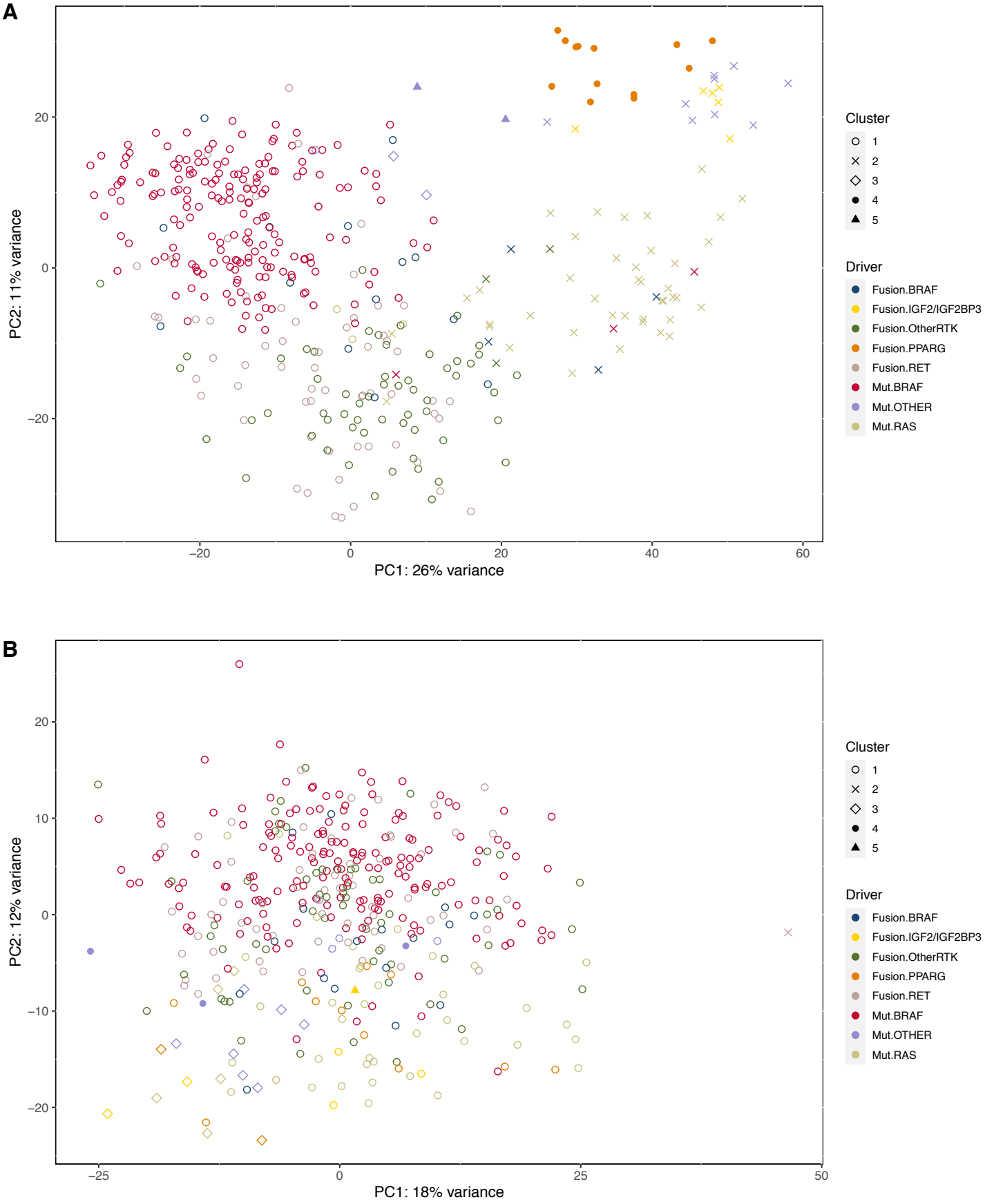


Fig. S22. Unsupervised clustering analysis of PTC only using ConsensusClusterPlus.
(A) mRNA, k=5. (B) miRNA, k=5. (C) Methylation, k=3. The appropriate cluster number (k) was determined by identifying the largest cluster with a delta area value >0.3, denoted with the dashed red line.


A
PC2: 11% variance
PC1: 26% variance
20
0
−20
−20
0
20
40
60
Cluster
1
2
3
4
5
Driver
Fusion.BRAF
Fusion.IGF2/IGF2BP3
Fusion.OtherRTK
Fusion.PPARG
Fusion.RET
Mut.BRAF
Mut.OTHER
Mut.RAS
B
PC2: 12% variance
PC1: 18% variance
20
10
0
−10
−20
−25
0
25
50
Cluster
1
2
3
4
5
Driver
Fusion.BRAF
Fusion.IGF2/IGF2BP3
Fusion.OtherRTK
Fusion.PPARG
Fusion.RET
Mut.BRAF
Mut.OTHER
Mut.RAS

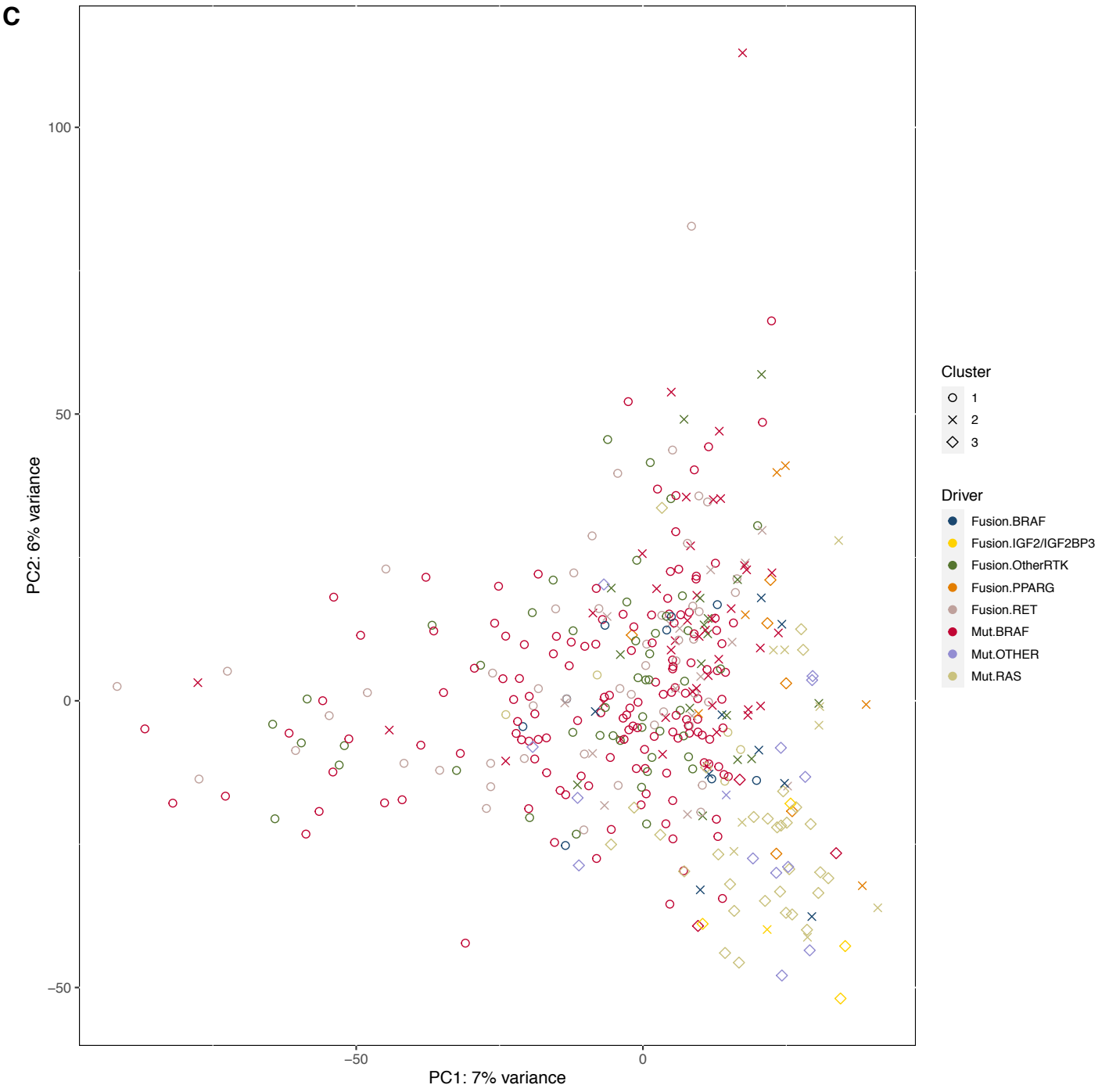


Fig. S23. Relationship of clusters to PTC driver.
(**A**) mRNA. (**B**) miRNA. (**C**) Methylation. Table S11 provides the distribution of clusters by driver

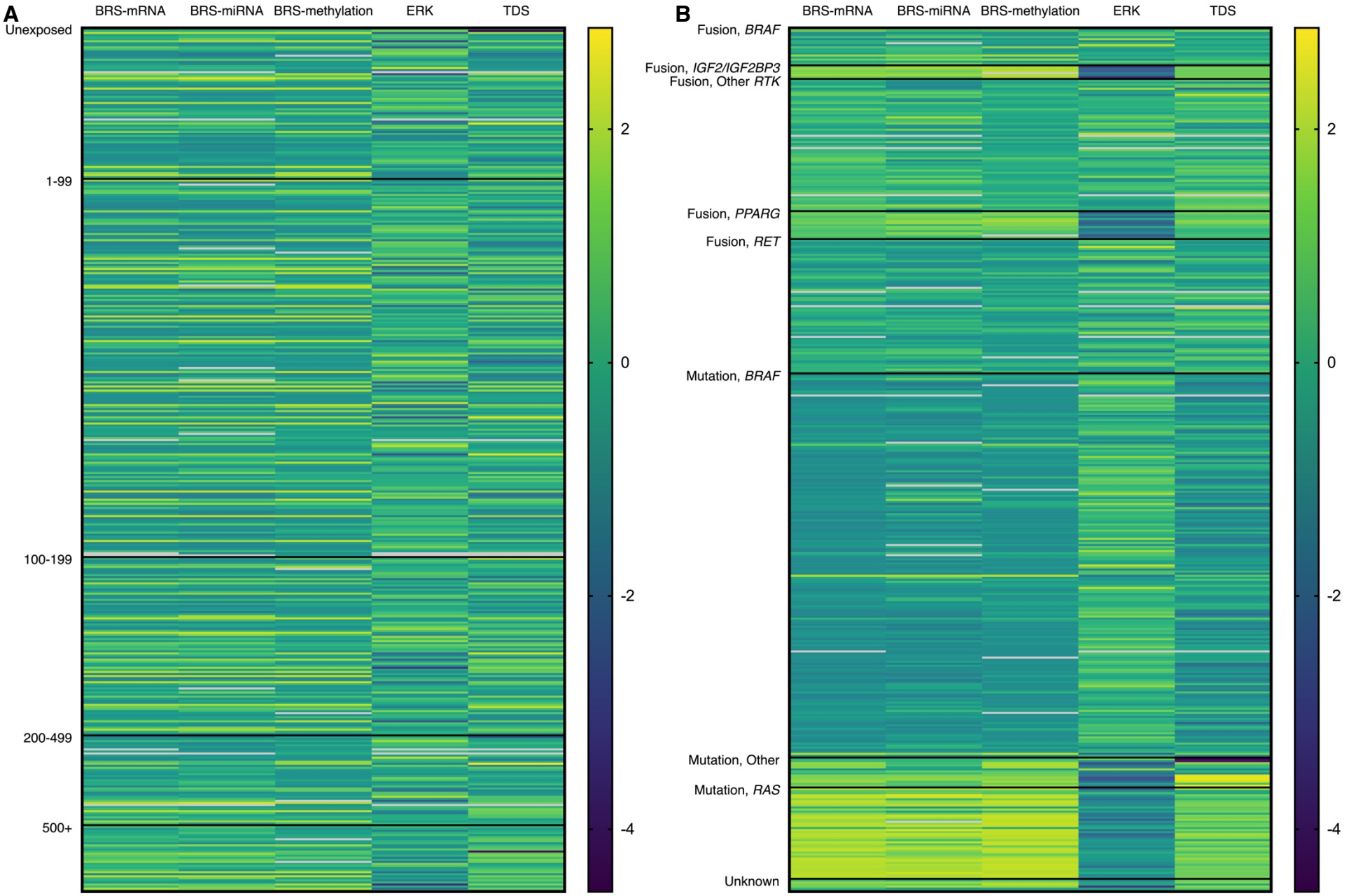


Fig. S24. Distribution of the *BRAF*^{V600E}-*RAS* (BRS) score from mRNA, miRNA, and methylation; ERK-activity score; and thyroid differentiation score.

(**A**) By radiation dose. (**B**) By PTC driver. Scores were standardized (mean=0, standard deviation=1) to facilitate comparison across the different measures.

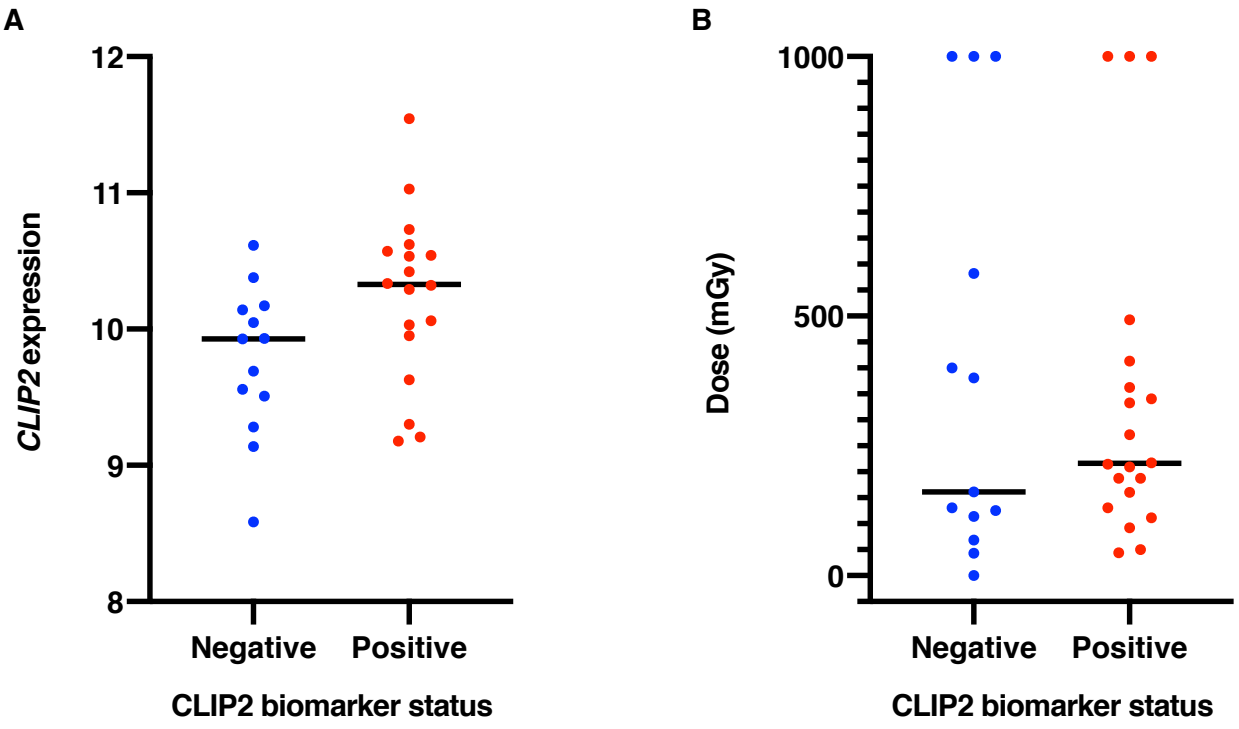


Fig. S25. Distribution of CLIP2 biomarker status.
In 33 overlapping samples between our study and a previously-published analysis (*25*), the CLIP2 biomarker status identified in that study correlated with *CLIP2* mRNA expression as measured by RNA-seq in our study (**A**) but did not correlate with radiation dose (**B**).

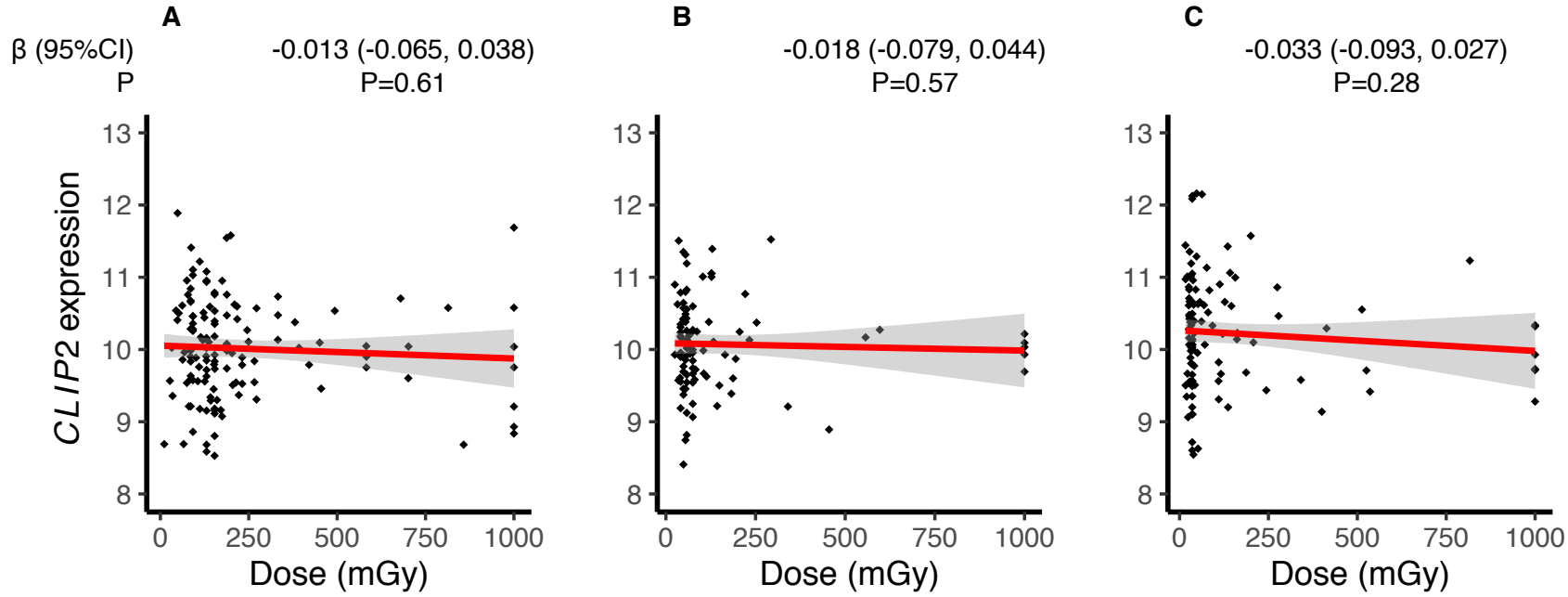


Fig. S26. Relationship between radiation dose from ^{131}I exposure and expression of *CLIP2* by age at exposure.
(**A**) <5 years at exposure, (**B**) 5-9 years at exposure, and (**C**) ≥10 years at exposure. Results for the total study population are provided in Fig. 5C.

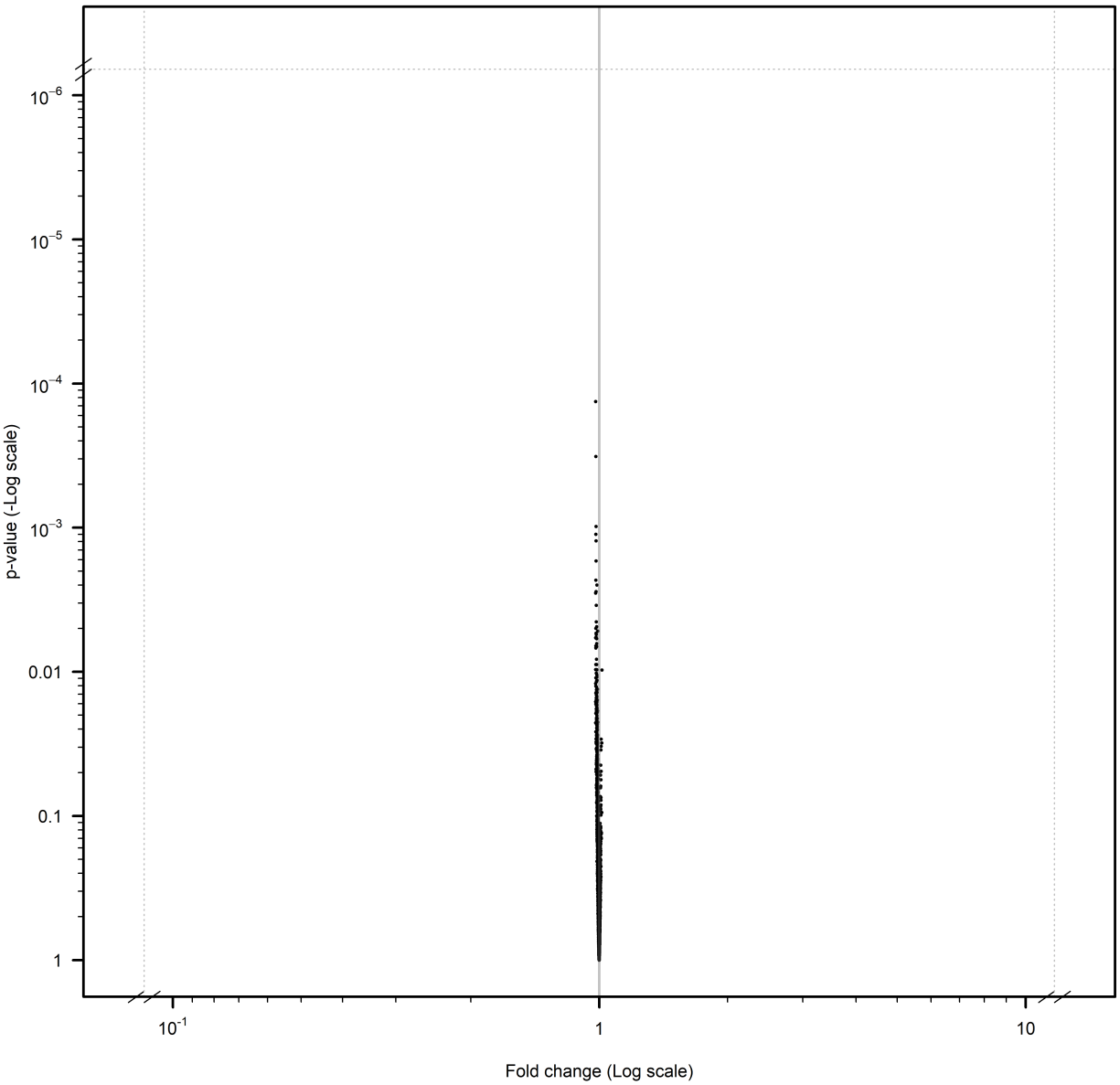
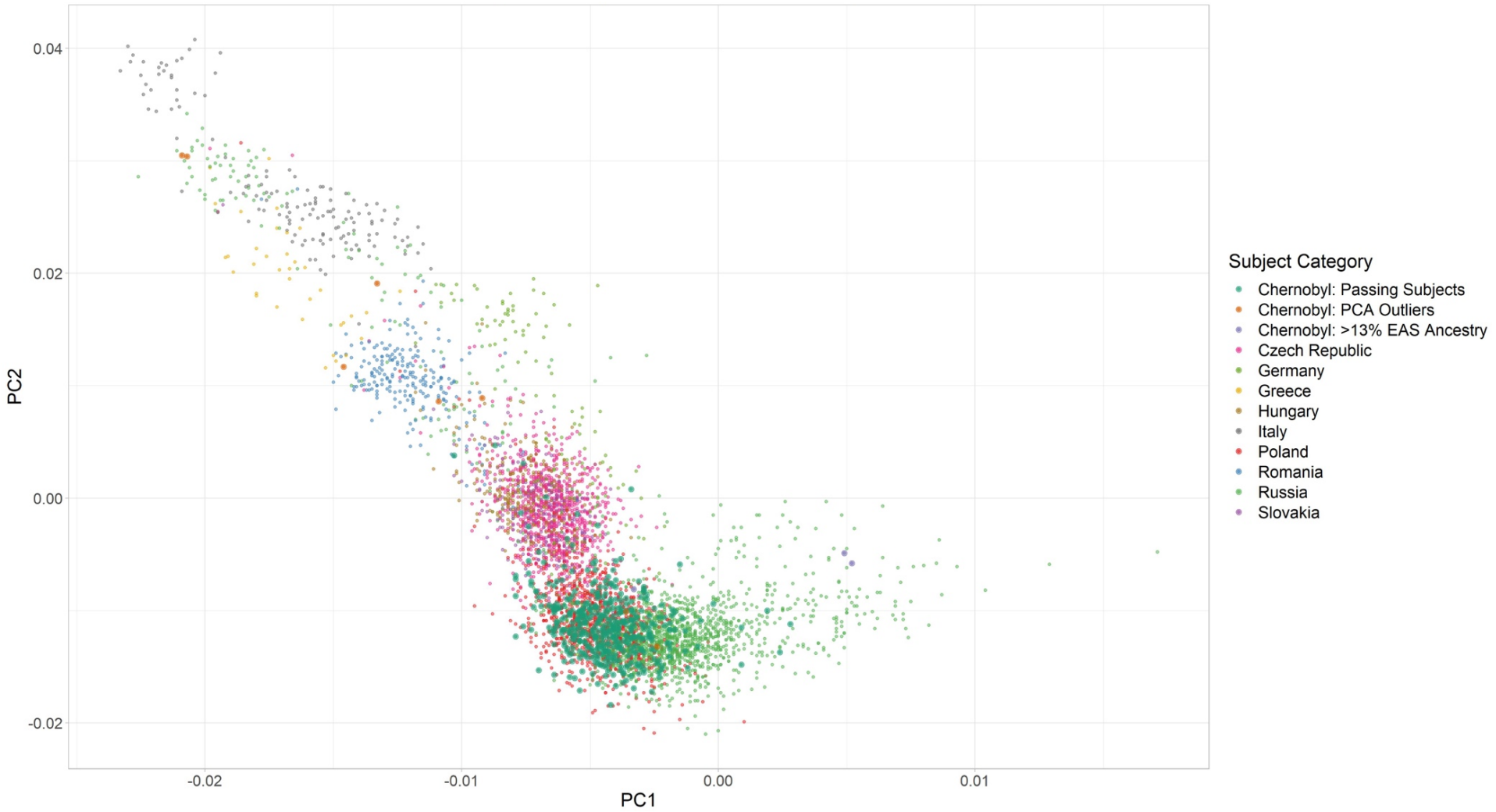


Fig. S27. Lack of association of radiation dose with any of 3,213 gene sets from the Molecular Signatures Database.

A


0.04
0.02
0.00
-0.02
PC2
-0.02
-0.01
0.00
0.01
PC1
Subject Category
Chernobyl: Passing Subjects
Chernobyl: PCA Outliers
Chernobyl: >13% EAS Ancestry
Czech Republic
Germany
Greece
Hungary
Italy
Poland
Romania
Russia
Slovakia

B

Fig. S28. Population substructure based on SNP array data.
(**A**) All ancestries combined in a single panel. (**B**) Each ancestry category in separate panels. Abbreviation: EAS=East Asian.

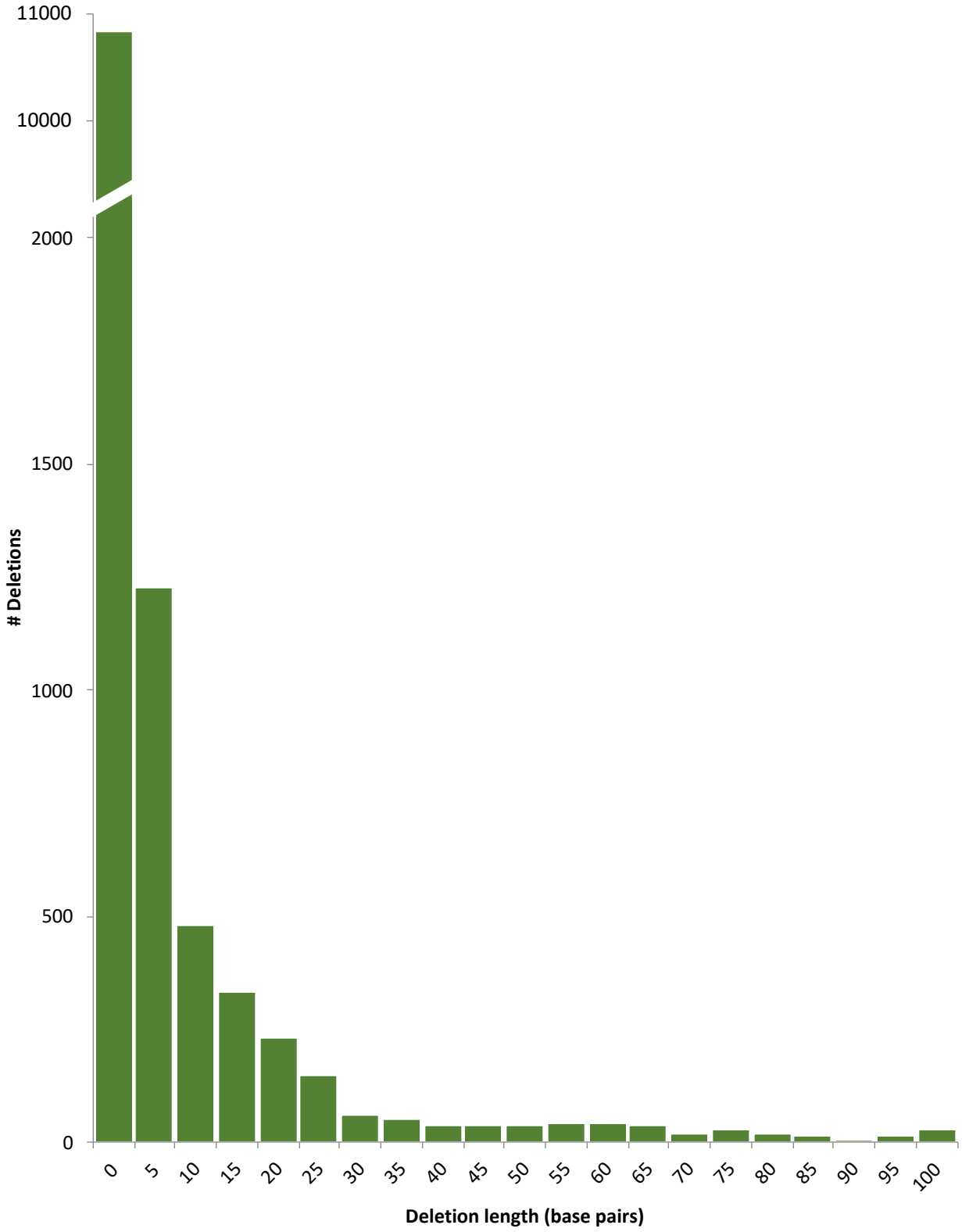


Fig. S29. Distribution of small deletions by length in n=356 individuals with high purity samples.

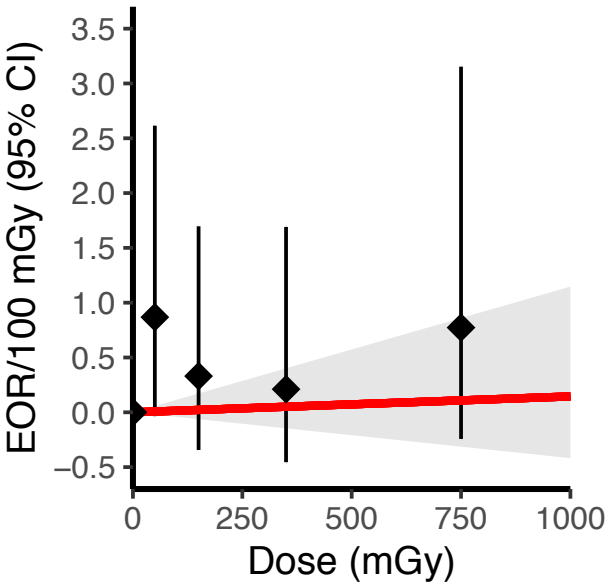


Fig. S30. Association between count of locally templated insertions (TINS) and radiation dose from ^{131}I exposure.

TINS are characteristic of theta-mediating end-joining (TMEJ) repair of DNA double-strand breaks (EOR/Gy=0.037, 95% confidence interval= -0.039-0.18; P=0.69)

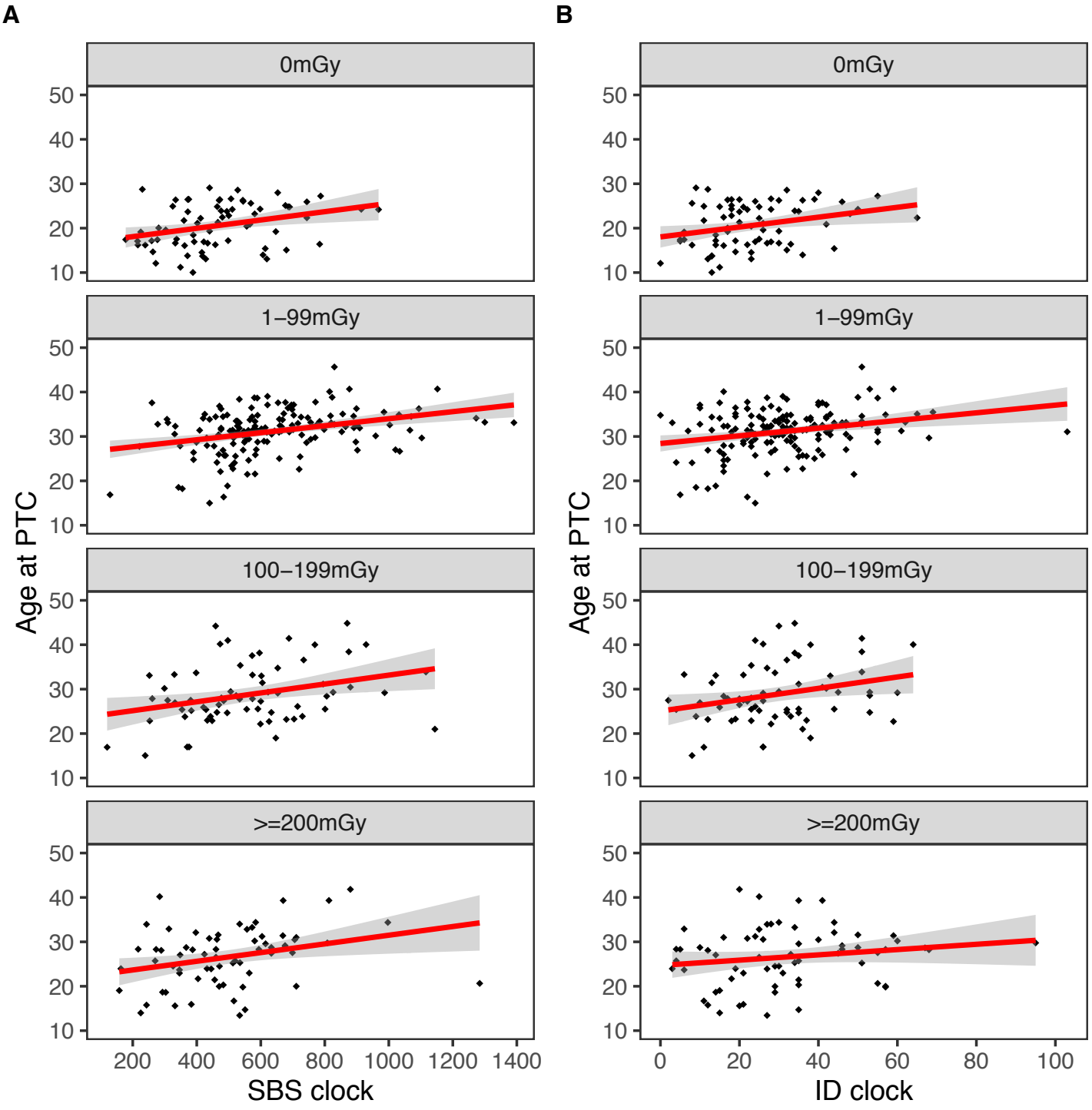


Fig. S31. Association for age at papillary thyroid cancer (years) with number of clock-like SBS and ID signatures, by radiation dose from ^{131}I exposure.

(A) SBS clock-like signatures (SBS1, SBS5), $P_{heterogeneity}$=0.63. (B) ID clock-like signatures (ID1, ID5), $P_{heterogeneity}$=0.93. $P_{heterogeneity}$ was derived from a generalized linear regression model, comparing model fit with and without an interaction term between radiation dose (continuous, truncated at 1 Gy) and age at PTC (continuous).

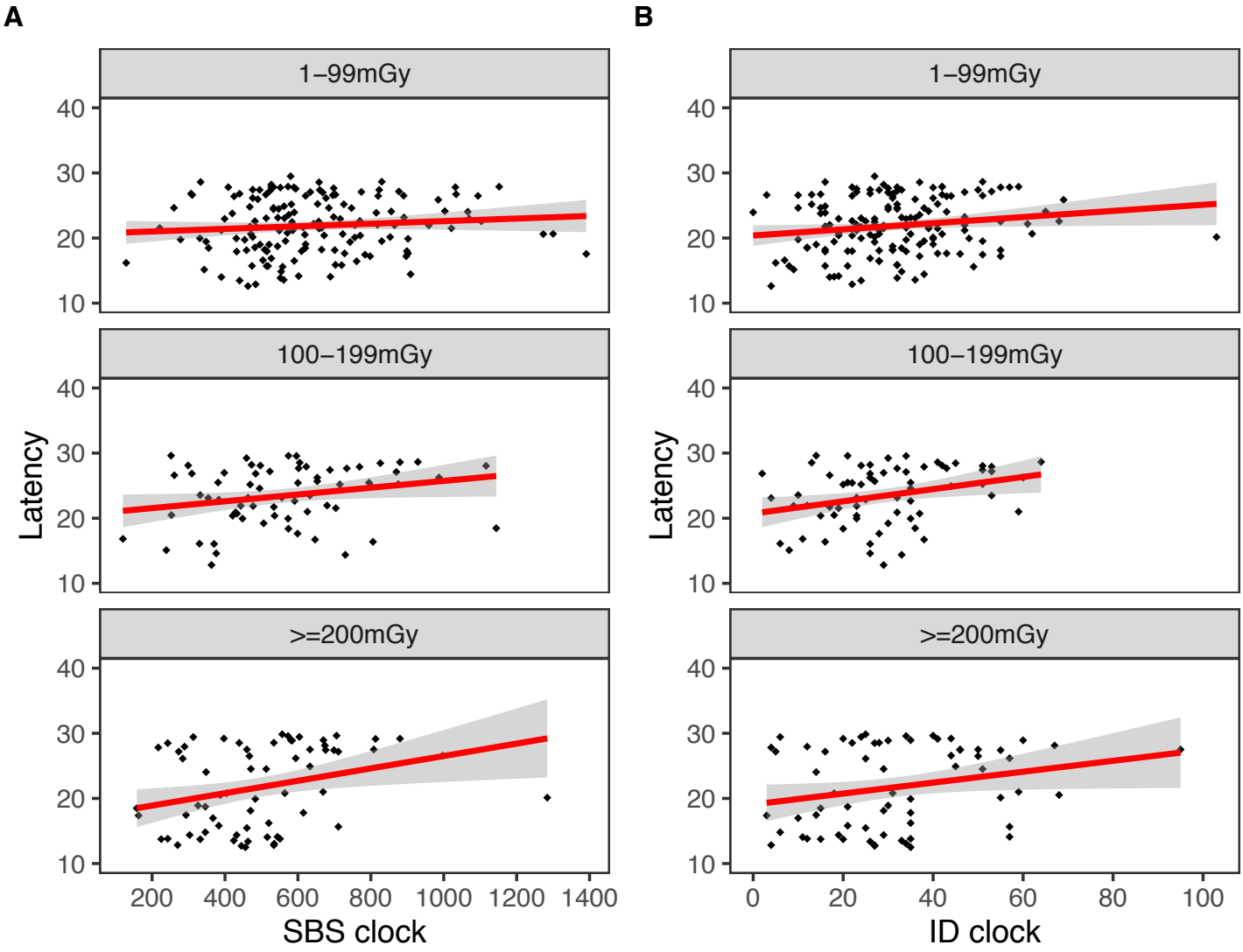


Fig. S32. Association for latency (years) with number of clock-like SBS and ID signatures, by radiation dose from ^{131}I exposure.

(A) SBS clock-like signatures (SBS1, SBS5), $P_{heterogeneity}$=0.28. (B) ID clock-like signatures (ID1, ID5), $P_{heterogeneity}$=0.21. $P_{heterogeneity}$ was derived from a generalized linear regression model, comparing model fit with and without an interaction term between radiation dose (continuous, truncated at 1 Gy) and age at PTC (continuous). Analyses were restricted to ^{131}I-exposed individuals.

Table S1. Demographic characteristics of the study populations.

Characteristic	^{131}I Exposed, Chernobyl Tissue Bank (n=359)	^{131}I Unexposed, Chernobyl Tissue Bank (n=81)	^{131}I Exposed, Efanov et al. (9) (n=45)
Sex			
Female	269 (74.9%)	66 (81.5%)	27 (60.0%)
Male	90 (25.1%)	15 (18.5%)	18 (40.0%)
Age at PTC (years)			
<15	4 (1.1%)	11 (13.6%)	1 (2.2%)
15-19	25 (7.0%)	25 (30.9%)	11 (24.4%)
20-24	48 (13.4%)	26 (32.1%)	12 (26.7%)
25-29	98 (27.3%)	19 (23.5%)	13 (28.9%)
30-34	125 (34.8%)	0 (0.0%)	8 (17.8%)
35-39	43 (12.0%)	0 (0.0%)	0 (0.0%)
40-44	14 (3.9%)	0 (0.0%)	0 (0.0%)
≥45	2 (0.6%)	0 (0.0%)	0 (0.0%)
Mean (± SD)	*29.7 (± 6.1)*	*20.7 (± 4.9)*	*24.2 (± 5.4)*
Radiation dose (mGy) †			
0	0 (0.0%)	81 (100.0%)	0 (0.0%)
1-99	193 (53.8%)	0 (0.0%)	8 (17.8%)
100-199	90 (25.1%)	0 (0.0%)	6 (13.3%)
200-499	43 (12.0%)	0 (0.0%)	6 (13.3%)
≥500	33 (9.2%)	0 (0.0%)	25 (55.6%)
Mean (± SD)	*247 (± 665)*	na	*1045 (± 1126)*
Age at exposure (years)			
<5	139 (38.7%)	na	17 (37.8%)
5-9	102 (28.4%)	na	14 (31.1%)
≥10	118 (32.9%)	na	14 (31.1%)
Mean (± SD)	*7.3 (± 5.1)*	na	*7.2 (± 4.5)*
Time since exposure (years)			
<20	119 (33.1%)	na	38 (84.4%)
20-24	109 (30.4%)	na	7 (15.6%)
≥25	131 (36.5%)	na	0 (0.0%)
Mean (± SD)	*22.4 (± 4.9)*	na	*17.0 (± 2.5)*
Probability of causality (%)			
0	0 (0.0%)	81 (100.0%)	0 (0.0%)
1-24	178 (49.6%)	0 (0.0%)	8 (17.8%)
25-49	108 (30.1%)	0 (0.0%)	8 (17.8%)
50-74	46 (12.8%)	0 (0.0%)	6 (13.3%)
≥75	27 (7.5%)	0 (0.0%)	23 (51.1%)
Dose reconstruction data source			
Direct measurement, with interview	49 (13.6%)	na	39 (86.7%)
Direct measurement only	4 (1.1%)	na	0 (0.0%)
Area measurements	297 (82.7%)	na	0 (0.0%)
In utero	9 (2.5%)	na	0 (0.0%)
Unknown	0 (0.0%)	na	6 (13.3%)
Residence at the time of the accident			
Chernigov	53 (14.8%)	8 (9.9%)	25 (55.6%)
Kiev	222 (61.8%)	64 (79.0%)	10 (22.2%)
Kirovograd	1 (0.3%)	0 (0.0%)	0 (0.0%)
Zhytomyr	83 (23.1%)	9 (11.1%)	10 (22.2%)
Thyroid surgery volume			
Hemithyroidectomy	46 (12.8%)	12 (14.8%)	0 (0.0%)
Total thyroidectomy	298 (83.0%)	67 (82.7%)	0 (0.0%)
Unknown	15 (4.2%)	2 (2.5%)	45 (100.0%)
Lesion size (cm)			
<1.0	47 (13.1%)	8 (9.9%)	4 (8.9%)
1.0-1.9	157 (43.7%)	35 (43.2%)	32 (71.1%)
2.0-2.9	66 (18.4%)	14 (17.3%)	2 (4.4%)
3.0-3.9	34 (9.5%)	11 (13.6%)	1 (2.2%)
≥4.0	42 (11.7%)	5 (6.2%)	0 (0.0%)
Unknown	13 (3.6%)	8 (9.9%)	6 (13.3%)
Mean (± SD)	*2.1 (± 1.3)*	*2.0 (± 1.1)*	*1.4 (± 0.5)*

Abbreviation: not applicable (na).

Table S2. Whole genome sequencing coverage, by study population and sample type.

	Primary tumor			Primary tumor control (blood or non-tumor thyroid)			Non-tumor thyroid			Non-tumor thyroid control (blood)		
Study population	n	Mean (± SD)	Range (min-max)	n	Mean (± SD)	Range (min-max)	n	Mean (± SD)	Range (min-max)	n	Mean (± SD)	Range (min-max)
Chernobyl Tissue Bank	383	88.8 (± 12.3)	(50.2 - 144.5)	383	33.2 (± 4.5)	(23.3 - 51.5)	233	33.3 (± 4.8)	(24.1 - 67.1)	233	33.2 (± 4.3)	(25.5 - 51.5)
^{131}I Exposed	305	91.2 (± 12.4)	(50.2 - 144.5)	305	33.2 (± 4.6)	(23.3 - 51.5)	178	33.6 (± 4.7)	(25.4 - 67.1)	178	33.2 (± 4.4)	(25.5 - 51.5)
^{131}I Unexposed	78	79.6 (± 6.0)	(71.1 - 98.1)	78	32.9 (± 4.2)	(24.5 - 43.9)	55	32.3 (± 5.1)	(24.1 - 45.5)	55	33.4 (± 3.9)	(26.2 - 43.9)

Table S3. Distribution of simple somatic variants by type and clonality in 383 tumors with whole genome sequencing data.

	Total number of mutations				Mutations per tumor					
		By clonality*					By clonality*			
	Total	Clonal	Subclonal	Undetermined	Total		Clonal		Subclonal	
Simple somatic variant	n (column %)	n (row %)	n (row %)	n (row %)	Mean (± SD)	Range	Mean (± SD)	Range	Mean (± SD)	Range
Total	318957 (100.0%)	114898 (36.0%)	153803 (48.2%)	50256 (15.8%)	832.8 (± 417.5)	(15 - 3560)	300.0 (± 196.2)	(14 - 2256)	401.6 (± 351.1)	(0 - 3238)
By type										
SNV	297513 (93.3%)	106881 (35.9%)	145665 (49.0%)	44967 (15.1%)	776.8 (± 401.0)	(14 - 3427)	279.1 (± 188.5)	(13 - 2238)	380.3 (± 336.9)	(0 - 3123)
Doublet	1351 (0.4%)	456 (33.8%)	680 (50.3%)	215 (15.9%)	3.5 (± 2.6)	(0 - 14)	1.2 (± 1.4)	(0 - 8)	1.8 (± 1.9)	(0 - 13)
Triplet	20 (0.006%)	8 (40.0%)	5 (25.0%)	7 (35.0%)	0.1 (± 0.2)	(0 - 2)	0.0 (± 0.1)	(0 - 1)	0.0 (± 0.1)	(0 - 2)
Small insertions	5842 (1.8%)	2143 (36.7%)	1951 (33.4%)	1748 (29.9%)	37.2 (± 18.7)	(0 - 140)	14.1 (± 10.1)	(0 - 77)	5.1 (± 5.1)	(0 - 38)
Small deletions	14231 (4.5%)	5410 (38.0%)	5502 (38.7%)	3319 (23.3%)	15.3 (± 8.7)	(0 - 89)	5.6 (± 3.9)	(0 - 31)	14.4 (± 11.8)	(0 - 73)

* As described in the Materials and Methods, cancer cell fraction reflects the variant allele fraction, accounting for the purity of the tumor sample. Clonal mutations were defined as those with cancer cell fraction >0.9, whereas subclonal mutations were defined as those with cancer cell fraction <0.6. Distributions of cancer cell fraction by mutation type are provided in Fig. S7.

Table S4. Relationship for selected molecular characteristics with radiation dose from ^{131}I exposure, age at papillary thyroid cancer, and sex.

Molecular characteristics †	Regression model*	Dose	Age at PTC	Sex
Simple somatic variants				
Total	Linear regression	5.4E-02	6.9E-07 ‡	5.9E-01
SNV	Linear regression	2.3E-02	3.2E-06 ‡	5.8E-01
Doublet	Linear regression	1.4E-01	2.7E-05 ‡	1.0E+00
Triplet	Logistic regression	6.8E-01	7.0E-01	6.9E-01
Small indels	Linear regression	2.6E-06 ‡	3.1E-15 ‡	9.2E-01
Small insertions	Linear regression	4.0E-01	1.5E-06 ‡	9.2E-01
Small deletions	Linear regression	8.0E-09 ‡	7.4E-16 ‡	1.0E+00
Indel:SNV ratio	Linear regression	9.5E-19 ‡	4.8E-02	1.3E-01
Insertion:SNV ratio	Linear regression	2.4E-03	5.9E-01	3.8E-01
Deletion:SNV ratio	Linear regression	4.9E-21 ‡	6.7E-03	1.6E-01
Microsatellite indels				
Total	Linear regression	9.8E-02	6.1E-02	1.2E-01
Insertion	Linear regression	8.2E-01	1.5E-01	2.9E-01
Deletion	Linear regression	2.1E-02	1.5E-01	2.3E-01
Total:SNV ratio	Linear regression	9.7E-03	3.5E-04 ‡	5.5E-01
Insertion:SNV ratio	Linear regression	1.5E-03	3.7E-04 ‡	5.9E-01
Deletion:SNV ratio	Linear regression	8.1E-01	1.7E-01	7.6E-01
SBS signatures				
Clock (SBS1,SBS5)	Linear regression	3.9E-02	9.1E-17 ‡	9.2E-01
SBS1	Linear regression	5.3E-01	1.9E-07 ‡	5.6E-01
SBS5	Linear regression	2.9E-02	6.8E-17 ‡	8.4E-01
APOBEC (SBS2,SBS13)	Proportional odds	8.9E-04	1.7E-01	8.0E-01
SBS2	Proportional odds	1.9E-03	3.8E-01	7.4E-01
SBS13	Proportional odds§	5.3E-02	3.4E-01	8.5E-01
SBS8	Linear regression	4.1E-01	4.3E-11 ‡	1.4E-01
SBS18	Proportional odds	4.8E-01	2.0E-08 ‡	5.6E-01
SBS23	Proportional odds	3.9E-01	3.5E-02	6.5E-01
SBS *de novo* A	Linear regression	9.6E-03	3.0E-19 ‡	8.9E-01
SBS *de novo* B	Linear regression	2.8E-01	2.2E-04 ‡	4.7E-01
SBS *de novo* C	Proportional odds§	2.2E-02	4.8E-01	9.6E-01
SBS *de novo* D	Proportional odds	2.0E-01	2.4E-05 ‡	9.9E-01
Indel signatures				
Clock (ID1,ID5)	Linear regression	1.6E-03	5.8E-09 ‡	1.0E+00
ID1	Linear regression	1.0E+00	6.9E-02	8.4E-01
ID5	Linear regression	1.3E-04 ‡	1.9E-09 ‡	9.2E-01
End-joining DNA repair (ID6,ID8)	Linear regression	1.5E-10 ‡	6.9E-05 ‡	8.9E-01
ID6	Proportional odds	9.4E-01	4.8E-01	3.0E-01
ID8	Linear regression	7.3E-09 ‡	6.6E-04 ‡	8.2E-01
ID3	Linear regression	7.8E-01	1.9E-07 ‡	1.0E+00
ID4	Proportional odds	5.9E-01	3.1E-02	6.2E-01
Indel *de novo* A	Linear regression	4.5E-07 ‡	9.4E-13 ‡	8.4E-01
Indel *de novo* B	Linear regression	7.9E-01	2.0E-02	7.1E-01
Fusion vs. mutation PTC driver	Logistic regression	6.6E-08 ‡	5.4E-09 ‡	2.0E-01
Structural variants				
Total	Proportional odds	1.4E-08 ‡	4.9E-04 ‡	7.2E-01
Simple/balanced	Proportional odds	1.2E-14 ‡	4.2E-06 ‡	1.9E-02
Complex	Proportional odds	5.2E-01	7.5E-01	7.3E-02
Simple/unbalanced	Proportional odds	5.6E-03	4.2E-01	1.5E-01
Chromoplexy	Logistic regression	7.0E-01	2.3E-01	2.9E-01
Somatic copy number alterations				
Total	Proportional odds	2.0E-03	4.7E-02	3.3E-02
Chromosomal (aneuploidy)	Logistic regression	2.0E-01	9.6E-01	7.1E-01

Sub-chromosome, total	Proportional odds	3.5E-05 ‡	4.8E-02	1.2E-02
Deletions	Proportional odds	7.0E-04 ‡	2.2E-02	6.6E-03
Gains	Logistic regression	3.2E-01	4.7E-01	4.3E-01
CNLOH	Logistic regression	5.2E-01	2.4E-01	3.1E-01
22q deletion	Logistic regression	3.7E-01	7.6E-01	7.4E-01
Unsupervised clustering analyses				
mRNA (k=5)‖	Polytomous regression	8.5E-01	9.6E-01	1.8E-01
miRNA (k=5)‖	Polytomous regression	3.8E-01	6.7E-01	8.8E-01
Methylation (k=3)‖	Polytomous regression	1.0E-01	8.1E-01	‖
BRAF-RAS score (BRS)				
mRNA BRS	Linear regression	2.1E-03	1.6E-01	4.4E-01
miRNA BRS	Linear regression	5.5E-03	8.4E-01	4.1E-01
Methylation BRS	Linear regression	8.2E-02	1.0E+00	4.3E-01
ERK-activity score	Linear regression	1.1E-02	5.9E-01	4.7E-01
Thyroid differentiation score	Linear regression	7.8E-03	6.8E-01	2.2E-01
Germline variants ¶				
Nature 2014 CSGs	Logistic regression	6.0E-01	9.4E-01	4.9E-01
Clinical radiation sensitivity syndromes	Logistic regression	7.6E-01	7.8E-01	6.5E-01
Any DNA damage response gene	Logistic regression	2.5E-01	2.9E-01	7.0E-01
Single strand break repair	Logistic regression	4.8E-01	3.5E-01	5.2E-01
Double strand break repair	Logistic regression	1.1E-01	6.1E-01	8.0E-01
Fanconi anemia	Logistic regression	7.1E-01	4.8E-01	9.8E-01
Homologous recombination	Logistic regression	6.1E-01	5.1E-01	5.4E-01
Polygenic risk score	Linear regression	4.7E-04 ‡	9.0E-02	7.3E-02

Abbreviation: cancer susceptibility gene (CSG).

* P-values were generated using likelihood ratio tests, comparing model fit with and without the variable of interest (dose, age at PTC, sex), based on multivariable generalized linear, proportional odds, or logistic regression models, depending on the distribution of the molecular characteristic. All models were mutually adjusted for radiation dose (linear, truncated at 1 Gy), age at PTC, and sex, unless otherwise noted below.

† Methods for defining each characteristic are described in the Materials and Methods.

‡ Associations of selected molecular characteristics with radiation dose with $P<7.4\times10^{-4}$ were considered statistically significant based on a Bonferroni correction for the 68 characteristics listed above.

§ Dose was modeled as a ordinal variable based on the means of the categories shown in Table S1 because models with the linear dose variable did not converge.

‖ Models were restricted to strata with >10 cases (mRNA: N=3, miRNA: N=2; see Table S11). Models of methylation clusters did not converge with sex in the model and thus included only radiation dose and age at PTC.

¶ Details regarding the germline variants are provided in the Materials and Methods and Tables S16-S17. Selected variables include those with >25 carriers of potentially protein damaging rare variants. All models of germline variants were additionally adjusted for population substructure (top 10 principal components).

Table S5. Fraction of mutations attributed to specific ID or SBS signatures and correlation amongst different signatures among N=356 tumors with high purity and available whole genome sequencing data (n=66 when restricted to dose ≥200 mGy).

Pearson correlation (black font), P (gray font); light red shading indicates |r|>0.4

Signature	Mutation count	Mutation fraction	Known ID signatures from COSMIC v3*								de novo ID signatures		de novo ID signatures (dose ≥200 mGy)		Known SBS signatures from COSMIC v3*									de novo SBS signatures				de novo SBS signatures (dose ≥200 mGy)
			ID3	ID4	Clock	ID1	ID5	End joining	ID6	ID8	A	B	A	B	SBS8	SBS18	SBS23	Clock	SBS1	SBS5	APOBEC	SBS2	SBS13	A	B	C	D	A
Known ID signatures from COSMIC v3*																												
ID3	4138	21.2%																										
ID4	1135	5.8%	0.17																									
			1.33E-03																									
Clock	10535	54.0%	0.01	-0.08																								
			0.825	*0.136*																								
ID1	2773	14.2%	0.12	0.00	0.56																							
			0.023	*0.982*	*6.33E-31*																							
ID5	7762	39.8%	-0.05	-0.10	0.89	0.13																						
			0.336	*0.072*	*8.47E-126*	*0.013*																						
EJ	3718	19.0%	0.14	0.08	0.39	0.08	0.42																					
			0.008	*0.149*	*2.60E-14*	*0.112*	*1.13E-16*																					
ID6	639	3.3%	0.04	0.03	0.05	0.14	-0.01	0.11																				
			0.509	*0.525*	*0.326*	*0.009*	*0.820*	*0.031*																				
ID8	3079	15.8%	0.12	0.06	0.35	0.02	0.40	0.90	-0.33																			
			0.025	*0.274*	*1.62E-11*	*0.708*	*1.75E-15*	*2.72E-130*	*2.20E-10*																			
de novo ID signatures																												
A	12536	64.2%	0.41	0.23	0.49	-0.19	0.69	0.70	-0.01	0.67																		
			3.90E-16	*1.21E-05*	*9.48E-23*	*3.79E-04*	*7.04E-51*	*2.17E-54*	*0.858*	*2.90E-48*																		
B	6990	35.8%	0.18	0.08	0.56	0.97	0.14	0.11	0.17	0.03	-0.16																	
			4.97E-04	*0.118*	*3.33E-30*	*2.39E-211*	*0.007*	*0.046*	*0.002*	*0.600*	*0.003*																	
de novo ID signatures (restricted to cases with dose ≥200 mGy)																												
A	2426	62.2%	0.26	0.01	0.67	-0.10	0.76	0.88	-0.08	0.88	0.94	-0.11																
			0.036	*0.906*	*6.28E-10*	*0.410*	*1.32E-13*	*1.15E-22*	*0.505*	*1.45E-22*	*4.90E-32*	*0.364*																
B	1476	37.8%	0.27	0.10	0.10	0.73	-0.11	-0.31	0.21	-0.37	-0.29	0.75	-0.52															
			0.027	*0.402*	*0.439*	*2.61E-12*	*0.398*	*0.011*	*0.087*	*0.002*	*0.020*	*2.81E-13*	*8.03E-06*															
Known SBS signatures from COSMIC v3*																												
SBS8	43397	15.1%	0.38	0.01	0.42	0.26	0.36	0.22	0.06	0.18	0.37	0.28	0.33	0.26														
			1.86E-13	*0.795*	*0.000*	*4.80E-07*	*1.19E-12*	*3.97E-05*	*0.270*	*0.001*	*6.54E-13*	*6.68E-08*	*0.007*	*0.032*														
SBS18	2632	0.9%	-0.02	0.03	-0.02	0.05	-0.06	-0.01	0.04	-0.03	-0.07	0.07	-0.14	0.04	-0.33													
			0.745	*0.533*	*0.655*	*0.354*	*0.300*	*0.855*	*0.467*	*0.623*	*0.200*	*0.215*	*0.271*	*0.758*	*2.04E-10*													
SBS23	4532	1.6%	0.13	0.04	0.18	0.18	0.12	0.10	-0.02	0.11	0.10	0.20	0.04	0.17	0.29	0.03												
			0.016	*0.485*	*0.001*	*0.001*	*0.021*	*0.055*	*0.668*	*0.044*	*0.071*	*9.88E-05*	*0.728*	*0.172*	*1.80E-08*	*0.564*												
Clock	201505	69.9%	0.44	0.18	0.56	0.26	0.53	0.37	0.07	0.32	0.58	0.31	0.44	0.25	0.43	-0.06	0.02											
			3.79E-18	*0.001*	*2.14E-30*	*5.91E-07*	*9.96E-27*	*4.35E-13*	*0.198*	*4.20E-10*	*2.99E-33*	*2.67E-09*	*2.60E-04*	*0.039*	*1.31E-17*	*0.242*	*0.700*											
SBS1	28137	9.8%	0.40	0.13	0.56	0.42	0.44	0.26	0.05	0.23	0.40	0.46	0.29	0.42	0.52	0.06	0.24	0.71										
			8.72E-15	*0.011*	*1.07E-30*	*1.82E-16*	*1.18E-18*	*4.34E-07*	*0.309*	*1.47E-05*	*6.50E-15*	*2.97E-20*	*0.019*	*0.0005*	*2.54E-26*	*0.297*	*2.92E-06*	*2.57E-55*										
SBS5	173368	60.2%	0.42	0.18	0.53	0.22	0.51	0.37	0.07	0.32	0.58	0.27	0.44	0.22	0.39	-0.08	-0.02	0.99	0.62									
			9.24E-17	*0.001*	*9.33E-27*	*2.55E-05*	*4.78E-25*	*6.59E-13*	*0.207*	*5.41E-10*	*5.34E-33*	*3.24E-07*	*0.0002*	*0.077*	*1.23E-14*	*0.142*	*0.745*	*0.00E+00*	*2.32E-39*									
APOBEC	36115	12.5%	0.08	0.13	-0.06	-0.07	-0.04	0.01	0.02	-0.004	0.06	-0.06	0.05	-0.04	-0.29	-0.08	-0.13	0.12	0.04	0.13								
			0.146	*0.013*	*0.243*	*0.199*	*0.481*	*0.905*	*0.668*	*0.941*	*0.242*	*0.294*	*0.677*	*0.762*	*3.32E-08*	*0.130*	*0.015*	*0.020*	*0.488*	*0.014*								
SBS2	17780	6.2%	0.08	0.14	-0.05	-0.06	-0.02	0.02	0.02	0.01	0.07	-0.04	0.09	-0.04	-0.28	-0.08	-0.13	0.14	0.05	0.14	0.98							
			0.141	*0.008*	*0.380*	*0.280*	*0.640*	*0.715*	*0.710*	*0.854*	*0.167*	*0.398*	*0.495*	*0.730*	*1.33E-07*	*0.129*	*0.013*	*0.010*	*0.342*	*0.007*	*9.18E-254*							
SBS13	18335	6.4%	0.07	0.12	-0.07	-0.08	-0.05	-0.01	0.02	-0.02	0.05	-0.06	0.00	-0.02	-0.29	-0.08	-0.12	0.11	0.02	0.11	0.99	0.93						
			0.164	*0.022*	*0.166*	*0.154*	*0.373*	*0.920*	*0.642*	*0.766*	*0.344*	*0.232*	*0.974*	*0.843*	*2.67E-08*	*0.143*	*0.021*	*0.043*	*0.656*	*0.030*	*1.52E-274*	*8.38E-160*						
de novo SBS signatures																												
A	140368	48.7%	0.38	0.14	0.45	0.15	0.46	0.33	0.07	0.29	0.54	0.19	0.41	0.15	0.48	-0.17	0.06	0.90	0.43	0.92	0.10	0.10	0.10					
			5.11E-14	*0.009*	*2.28E-19*	*0.005*	*3.87E-20*	*1.32E-10*	*0.212*	*3.66E-08*	*2.54E-28*	*4.34E-04*	*0.001*	*0.217*	*3.49E-22*	*0.001*	*0.222*	*6.67E-128*	*1.46E-17*	*1.34E-150*	*0.065*	*0.069*	*0.070*					
B	97010	33.7%	0.35	0.06	0.49	0.41	0.37	0.22	0.04	0.19	0.31	0.44	0.26	0.45	0.63	0.18	0.32	0.47	0.83	0.38	-0.38	-0.36	-0.38	0.20				
			8.67E-12	*0.236*	*3.63E-33*	*1.30E-15*	*5.15E-13*	*2.92E-05*	*0.431*	*3.00E-04*	*3.07E-09*	*1.79E-18*	*0.035*	*0.0002*	*3.13E-40*	*0.001*	*6.50E-10*	*8.51E-21*	*9.37E-92*	*6.37E-14*	*1.28E-13*	*1.30E-12*	*1.06E-13*	*1.49E-04*				
C	25527	8.9%	0.07	0.12	-0.07	-0.08	-0.04	0.002	0.03	-0.01	0.06	-0.07	0.09	-0.13	-0.29	-0.08	-0.11	0.11	0.02	0.12	0.98	0.93	1.00	0.11	-0.39			
			0.174	*0.020*	*0.194*	*0.132*	*0.458*	*0.969*	*0.627*	*0.861*	*0.269*	*0.210*	*0.494*	*0.305*	*2.63E-08*	*0.126*	*0.030*	*0.031*	*0.764*	*0.019*	*1.21E-251*	*4.05E-153*	*0.00E+00*	*0.037*	*1.61E-14*			
D	25276	8.8%	0.13	0.16	0.04	-0.01	0.05	0.08	0.02	0.06	0.15	0.01	0.22	0.07	-0.17	-0.08	-0.09	0.23	0.15	0.23	0.89	0.95	0.81	0.16	-0.26	0.79		
			0.015	*0.003*	*0.491*	*0.842*	*0.350*	*0.143*	*0.644*	*0.235*	*0.005*	*0.899*	*0.079*	*0.595*	*0.001*	*0.151*	*0.106*	*8.89E-06*	*0.006*	*7.79E-06*	*1.12E-122*	*1.08E-183*	*2.54E-83*	*0.002*	*7.60E-07*	*1.58E-78*		
de novo SBS signatures (restricted to cases with dose ≥200 mGy)																												
A	37095	84.0%	0.54	0.11	0.53	0.35	0.47	0.29	0.04	0.27	0.50	0.36	0.36	0.29	0.76	-0.07	0.16	0.74	0.73	0.72	-0.42	-0.42	-0.37	0.69	0.76	-0.35	-0.09	
			2.59E-06	*0.382*	*5.44E-06*	*0.004*	*6.07E-05*	*0.017*	*0.728*	*0.029*	*1.69E-05*	*0.003*	*0.003*	*0.017*	*1.78E-13*	*0.590*	*0.195*	*9.06E-13*	*3.88E-12*	*1.46E-11*	*4.35E-04*	*0.001*	*0.002*	*1.06E-10*	*1.14E-13*	*0.004*	*0.475*	
B	7055	16.0%	-0.11	-0.16	0.14	0.07	0.13	0.15	0.03	0.13	0.07	0.04	0.08	-0.02	-0.11	-0.15	-0.22	0.28	0.04	0.30	0.98	0.98	0.84	0.27	-0.15	0.70	0.85	-0.40
			0.393	*0.202*	*0.279*	*0.589*	*0.308*	*0.239*	*0.821*	*0.287*	*0.572*	*0.726*	*0.511*	*0.875*	*0.360*	*0.224*	*0.076*	*0.025*	*0.768*	*0.014*	*1.35E-46*	*5.32E-50*	*8.08E-19*	*0.028*	*0.222*	*4.14E-11*	*1.32E-19*	*0.001*

* Catalogue of Somatic Mutations in Cancer (COSMIC) v3, https://cancer.sanger.ac.uk/cosmic/signatures

Table S6. Distribution of structural variants and somatic copy number alterations by type and clonality

		By clonality	
	Total	Clonal	Subclonal
Event	n (column %)	n (row %)	n (row %)
Structural variant events (Total)			
Total	479 (100.0%)	259 (54.1%)	220 (45.9%)
Simple/balanced	132 (27.6%)	106 (80.3%)	26 (19.7%)
Complex	94 (19.6%)	66 (70.2%)	28 (29.8%)
Simple/unbalanced	253 (52.8%)	87 (34.4%)	166 (65.6%)
Structural variant events (excluding 2 PTC cases with >10 SVs)			
Total	450 (100.0%)	240 (53.3%)	210 (46.7%)
Simple/balanced	125 (27.8%)	100 (80.0%)	25 (20.0%)
Complex	83 (18.4%)	60 (72.3%)	23 (27.7%)
Simple/unbalanced	242 (53.8%)	80 (33.1%)	162 (66.9%)
Confirmed structural variant events (Total)			
Total	309 (100.0%)	227 (73.5%)	82 (26.5%)
Simple/balanced	139 (45.0%)	110 (79.1%)	29 (20.9%)
<20 bp intervening loss/gain	91 (29.4%)	88 (96.7%)	3 (3.3%)
<4 bp intervening loss/gain	45 (14.6%)	43 (95.6%)	2 (4.4%)
4-<20 bp intervening loss/gain	46 (14.9%)	45 (97.8%)	1 (2.2%)
Inversion	52 (30.2%)	50 (96.2%)	2 (3.8%)
Translocation	39 (22.7%)	38 (97.4%)	1 (2.6%)
Complex	90 (29.1%)	61 (67.8%)	29 (32.2%)
Simple/unbalanced	80 (25.9%)	56 (70.0%)	24 (30.0%)
Confirmed structural variant events, excluding SV/fusion drivers			
Total	172 (100.0%)	99 (57.6%)	73 (42.4%)
Simple/balanced	58 (33.7%)	34 (58.6%)	24 (41.4%)
<20 bp intervening loss/gain	26 (15.1%)	23 (88.5%)	3 (11.5%)
<4 bp intervening loss/gain	15 (8.7%)	13 (86.7%)	2 (13.3%)
4-<20 bp intervening loss/gain	11 (6.4%)	10 (90.9%)	1 (9.1%)
Inversion	19 (11.0%)	17 (89.5%)	2 (10.5%)
Translocation	7 (4.1%)	6 (85.7%)	1 (14.3%)
Complex	45 (26.2%)	18 (40.0%)	27 (60.0%)
Simple/unbalanced	69 (40.1%)	47 (68.1%)	22 (31.9%)
Confirmed structural variant events, restricted to SV/fusion drivers			
Total	137 (100.0%)	128 (93.4%)	9 (6.6%)
Simple/balanced	81 (59.1%)	76 (93.8%)	5 (6.2%)
<20 bp intervening loss/gain	65 (47.4%)	65 (100.0%)	0 (0.0%)
<4 bp intervening loss/gain	30 (21.9%)	30 (100.0%)	0 (0.0%)
4-<20 bp intervening loss/gain	35 (25.5%)	35 (100.0%)	0 (0.0%)
Inversion	33 (19.2%)	33 (100.0%)	0 (0.0%)
Translocation	32 (18.6%)	32 (100.0%)	0 (0.0%)
Complex	45 (32.8%)	43 (95.6%)	2 (4.4%)
Simple/unbalanced	11 (8.0%)	9 (81.8%)	2 (18.2%)
Somatic copy number alterations (Total)			
Total	326 (100.0%)	248 (76.1%)	78 (23.9%)
Chromosome level (aneuploidy)			
Total	132 (40.5%)	113 (85.6%)	19 (14.4%)
Deletions	48 (14.7%)	33 (68.8%)	15 (31.3%)
Gains	65 (19.9%)	61 (93.8%)	4 (6.2%)

CNLOH	19 (5.8%)	19 (100.0%)	0 (0.0%)
Sub-chromosome level (DNA double strand breaks)			
Total	194 (59.5%)	135 (69.6%)	59 (30.4%)
Deletions	115 (35.3%)	96 (83.5%)	19 (16.5%)
Gains	60 (18.4%)	28 (46.7%)	32 (53.3%)
CNLOH	19 (5.8%)	11 (57.9%)	8 (42.1%)
Somatic copy number alterations (excluding 4 PTC cases with >20 SCNAs each)			
Total	239 (100.0%)	165 (69.0%)	74 (31.0%)
Chromosome level (aneuploidy)			
Total	69 (28.9%)	50 (72.5%)	19 (27.5%)
Deletions	48 (20.1%)	33 (68.8%)	15 (31.3%)
Gains	20 (8.4%)	16 (80.0%)	4 (20.0%)
CNLOH	1 (0.4%)	1 (100.0%)	0 (0.0%)
Sub-chromosome level (DNA double strand breaks)			
Total	170 (71.1%)	115 (67.6%)	55 (32.4%)
Deletions	106 (44.4%)	87 (82.1%)	19 (17.9%)
Gains	51 (21.3%)	23 (45.1%)	28 (54.9%)
CNLOH	13 (5.4%)	5 (38.5%)	8 (61.5%)

Table S7. Designated and candidate drivers for N=440 PTCs included in the study.

Driver gene	Designated driver*	Additional candidate driver	n
Mutations			
BRAF	*BRAF* ^{V600E}		175
BRAF	*BRAF* ^{V600E}	*AKT1*	2
BRAF	*BRAF* ^{V600E}	*PIK3R1*	2
BRAF	*BRAF* ^{V600E}	*ARID2*	1
BRAF	*BRAF* ^{V600E}	*BCOR*	1
BRAF	*BRAF* ^{V600E}	*CR2*	1
BRAF	*BRAF* ^{V600E}	*FAT3*	1
BRAF	*BRAF* ^{V600E}	*GPC3*	1
BRAF	*BRAF* ^{V600E}	*MTOR*	1
BRAF	*BRAF* ^{V600E}	*PIK3CA*	1
BRAF	*BRAF* ^{V600E}	*RET*	1
BRAF	*BRAF* ^{V600E}	*RGS7*	1
BRAF	*BRAF* ^{V600E}	*SETD2*	1
BRAF	*BRAF* ^{V600E}	*TSHR*	1
BRAF	*BRAF* $^{Non-V600E}$		3
BRAF	*BRAF* $^{Non-V600E}$	*PTEN-PXT1* fusion	1
NRAS	*NRAS*		15
NRAS	*NRAS*	*SETD2*	1
NRAS	*NRAS*	*SF3B1*	1
NRAS	*NRAS*	*TP53*	1
NRAS	*NRAS*	*NEBL-TCF7L2* fusion	1
NRAS	*NRAS*	*PAX8-PPARG* fusion	1
HRAS	*HRAS*		14
HRAS	*HRAS*	*TP53*	1
KRAS	*KRAS*		8
KRAS	*KRAS*	*AKT1*	1
TSHR	*TSHR*		6
DICER1	*DICER1*		3
APC	*APC*		2
NFE2L2	*NFE2L2*		1
NFE2L2	*NFE2L2*	*GPHN-RAF1* fusion	1
TSC1	*TSC1*		1
TSC2	*TSC2*		1
Fusions			
RET	*CCDC6-RET*		39
RET	*CCDC6-RET*	*PIK3R1*	1
RET	*NCOA4-RET*		14
RET	*NCOA4-RET*	*SETD2*	1
RET	*RET* -Other		18
NTRK3	*ETV6-NTRK3*		31
NTRK3	*ETV6-NTRK3*	*CR2*	1
NTRK3	*NTRK3* -Other		3
NTRK3	*NTRK3* -Other	*BCOR*	1
BRAF	*AGK-BRAF*		6
BRAF	*BRAF-SND1*		4
BRAF	*BRAF* -Other		9
NTRK1	*NTRK1-TPR*		5
NTRK1	*NTRK1-TPM3*		4
NTRK1	*NTRK1* -Other		4
PPARG	*CREB3L2-PPARG*		6
PPARG	*PAX8-PPARG*		5
PPARG	*PAX8-PPARG*	*CR2*	1
PPARG	*PPARG* -Other		1
ALK	*ALK-STRN*		6
ALK	*ALK-STRN*	*H3F3B*	1
ALK	*ALK-STRN*	*PTPRT*	1
ALK	*ALK* -Other		4
LTK	*LTK* -Other		3
Structural variants			
BRAF	*BRAF*		1
IGF2	*IGF2*		2
IGF2BP3	*IGF2BP3*		4
Not designated			
~	~	*BRAF,* Other*; KRAS*	1
~	~		7
~	~	*CTTNBP2-MET* fusion	1
~	~	*CCDC30-ROS1* fusion	1
~	~	*NUTM1-WHSC1L1* fusion	1

~ Indicates not applicable.

* Rules for specifying designated and candidate drivers are provided in the Materials and Methods.

Table S8. Distribution of designated mutation drivers, by mutation type and gene.

	Oncogenic mutation driver type				Tumor suppressor mutation driver types			
Gene	SNV	Doublet	Insertion	Deletion	SNV+SCNA	SNV+Deletion	Deletion+SCNA	Total
Oncogenes								
BRAF ^{V600E}	190	0	0	0				190
BRAF $^{Non\text{-}V600E}$	2	0	1	1				4
NRAS	20	0	0	0				20
HRAS	14	1	0	0				15
KRAS	3	6	0	0				9
TSHR	6	0	0	0				6
NFE2L2	2	0	0	0				2
Tumor suppressor genes								
DICER1					0	3	0	3
APC					1	0	1	2
TSC1					0	0	1	1
TSC2					0	1	0	1
Total	237	7	1	1	1	4	2	253

Table S9. Distribution of designated fusion or structural variant drivers, by oncogene and fusion partner.

Oncogene	Fusion*	n
Fusions		
RET	*CCDC6-RET (RET-PTC1)*	40
RET	*NCOA4-RET (RET-PTC3)*	15
RET	*RET* -Other*	18
NTRK3	*ETV6-NTRK3*	32
NTRK3	*NTRK3* -Other*	4
BRAF	*AGK-BRAF*	6
BRAF	*BRAF-SND1*	4
BRAF	*BRAF* -Other*	9
PPARG	*PAX8-PPARG*	6
PPARG	*CREB3L2-PPARG*	6
PPARG	*PPARG* -Other*	1
NTRK1	*NTRK1-TPR*	5
NTRK1	*NTRK1-TPM3*	4
NTRK1	*NTRK1* -Other*	4
ALK	*ALK-STRN*	8
ALK	*ALK* -Other*	4
LTK	*LTK* -Other*	3
Structural variants		
BRAF	*BRAF* , Large deletion	1
IGF2BP3	*IGF2BP3*	4
IGF2	*IGF2*	2

* All recurrent fusions are listed separately; a full list of fusion partners, including those for fusions that occurred only once in our dataset, is provided in Data S1.

Table S10. Distribution of 22q deletions by driver type and pathway.

	22q deletion		
	No	Yes	
	n (%)	n (%)	$P_{heterogeneity}$*
Driver type			
Fusion	125 (83.3%)	13 (8.7%)	0.07
Mutation	178 (79.1%)	34 (15.1%)	
Driver pathway			
Fusion, *RET*	50 (82.0%)	2 (3.3%)	2.8E-10
Fusion, Other *RTK*	49 (87.5%)	7 (12.5%)	
Fusion, *BRAF*	15 (75.0%)	3 (15.0%)	
Fusion, *PPARG*	8 (88.9%)	0 (0.0%)	
Fusion, *IGF2/IGF2BP3*	3 (75.0%)	1 (25.0%)	
Mutation, *BRAF*	152 (86.9%)	12 (6.9%)	
Mutation, *RAS*	16 (41.0%)	22 (56.4%)	
Mutation, Other	10 (90.9%)	0 (0.0%)	

*$P_{heterogeneity}$ computed using a likelihood ratio test, comparing logistic regression model fit with and without the variable of interest.

Table S11. Distribution of mRNA, miRNA, and methylation clusters from unsupervised clustering analyses, by radiation dose and PTC driver.

		By radiation dose (mGy)					Total with designated driver	By PTC driver							
Cluster number	Total n (%)	0 n (%)	1-99 n (%)	100-199 n (%)	200-499 n (%)	≥500 n (%)	n (%)	Fusion, *RET* n (%)	Fusion, Other *RTK* n (%)	Fusion, *BRAF* n (%)	Fusion, *PPARG* n (%)	Fusion, *IGF2/IGF2BP3* n (%)	Mutation, *BRAF* n (%)	Mutation, *RAS* n (%)	Mutation, Other n (%)
mRNA															
1	323 (79.0%)	55 (77.5%)	145 (80.6%)	67 (77.9%)	32 (80.0%)	24 (75.0%)	320 (79.4%)	62 (100.0%)	58 (95.1%)	14 (77.8%)	0 (0.0%)	0 (0.0%)	182 (98.4%)	3 (6.8%)	1 (7.1%)
2	69 (16.9%)	14 (19.7%)	28 (15.6%)	16 (18.6%)	5 (12.5%)	6 (18.8%)	66 (16.4%)	0 (0.0%)	3 (4.9%)	4 (22.2%)	0 (0.0%)	6 (100.0%)	3 (1.6%)	41 (93.2%)	9 (64.3%)
3	2 (0.5%)	1 (1.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.1%)	2 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (14.3%)
4	13 (3.2%)	1 (1.4%)	7 (3.9%)	3 (3.5%)	1 (2.5%)	1 (3.1%)	13 (3.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	13 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
5	2 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (5.0%)	0 (0.0%)	2 (0.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (14.3%)
Total	409 (100.0%)	71 (100.0%)	180 (100.0%)	86 (100.0%)	40 (100.0%)	32 (100.0%)	403 (100.0%)	62 (100.0%)	61 (100.0%)	18 (100.0%)	13 (100.0%)	6 (100.0%)	185 (100.0%)	44 (100.0%)	14 (100.0%)
miRNA															
1	383 (94.6%)	65 (91.5%)	170 (96.6%)	81 (95.3%)	37 (90.2%)	30 (93.8%)	378 (94.7%)	62 (98.4%)	61 (100.0%)	17 (100.0%)	11 (84.6%)	3 (50.0%)	182 (100.0%)	38 (88.4%)	4 (28.6%)
2	1 (0.2%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.3%)	1 (1.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
3	17 (4.2%)	3 (4.2%)	5 (2.8%)	4 (4.7%)	4 (9.8%)	1 (3.1%)	16 (4.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (15.4%)	2 (33.3%)	0 (0.0%)	5 (11.6%)	7 (50.0%)
4	3 (0.7%)	3 (4.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (0.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (21.4%)
5	1 (0.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.1%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (16.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Total	405 (100.0%)	71 (100.0%)	176 (100.0%)	85 (100.0%)	41 (100.0%)	32 (100.0%)	399 (100.0%)	63 (100.0%)	61 (100.0%)	17 (100.0%)	13 (100.0%)	6 (100.0%)	182 (100.0%)	43 (100.0%)	14 (100.0%)
Methylation															
1	252 (61.5%)	48 (66.7%)	115 (63.2%)	51 (60.7%)	24 (57.1%)	14 (46.7%)	250 (61.9%)	49 (76.6%)	47 (73.4%)	9 (50.0%)	0 (0.0%)	0 (0.0%)	140 (76.5%)	5 (11.4%)	0 (0.0%)
2	101 (24.6%)	13 (18.1%)	45 (24.7%)	17 (20.2%)	14 (33.3%)	12 (40.0%)	98 (24.3%)	15 (23.4%)	17 (26.6%)	9 (50.0%)	6 (50.0%)	1 (20.0%)	40 (21.9%)	9 (20.5%)	1 (7.1%)
3	57 (13.9%)	11 (15.3%)	22 (12.1%)	16 (19.0%)	4 (9.5%)	4 (13.3%)	56 (13.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (50.0%)	4 (80.0%)	3 (1.6%)	30 (68.2%)	13 (92.9%)
Total	410 (100.0%)	72 (100.0%)	182 (100.0%)	84 (100.0%)	42 (100.0%)	30 (100.0%)	404 (100.0%)	64 (100.0%)	64 (100.0%)	18 (100.0%)	12 (100.0%)	5 (100.0%)	183 (100.0%)	44 (100.0%)	14 (100.0%)

Table S12. Pearson correlation coefficients among the scores constructed based on the previous TCGA analysis (10) to reflect transcriptional patterns related to key signaling and thyroid differentiation pathways.

	Pearson correlation (black font), P (gray font), sample size (N, black font)				
	mRNA BRS, TCGA genes	mRNA BRS	miRNA BRS	Methylation BRS	ERK-activity score
mRNA BRS	0.98 2.9E-277 n=409				
miRNA BRS	0.83 4.0E-103 n=402	0.86 9.6E-117 n=402			
Methylation BRS	0.94 2.7E-191 n=402	0.92 3.5E-166 n=402	0.78 2.5E-82 n=398		
ERK-activity score	-0.66 8.1E-52 n=409	-0.61 1.7E-43 n=409	-0.54 2.9E-31 n=402	-0.62 3.0E-44 n=402	
Thyroid differentiation score	0.69 5.7E-60 n=409	0.73 5.0E-68 n=409	0.61 1.8E-42 n=402	0.64 1.3E-47 n=402	-0.56 2.3E-35 n=409

Table S13. Results from differential expression modeling of individual genes.

Provided as a separate Excel file available on the *Science* website.

Table S14. Results from differential expression modeling for gene sets from the Molecular Signatures Database.

Provided as a separate Excel file available on the *Science* website.

Table S15. Relationship of germline SNPs with radiation dose from ^{131}I exposure, including the 12 SNPs that comprised the PRS.

			gnomAD genomes (n=12,857 individuals)					Meta-analysis results for PRS construction				Risk allele frequency								
			Alleles		Alternative allele frequency							Total	Radiation dose (mGy)					Effect size per 100 mGy		
SNP	Chr	Position‡	Ref	Alt	NFE	EAS	AFR	Risk allele	Frequency	OR	(95% CI)	(n=383)	0 (n=78)	1-99 (n=165)	100-199 (n=73)	200-499 (n=36)	≥500 (n=31)	EOR	(95% CI)	P*
rs12129938	1	233412561	A	G	0.22	0.13	0.18	A	0.80	1.32	(1.20 , 1.43)	0.82	0.84	0.82	0.83	0.83	0.81	0.0004	(-0.06 , 0.14)	0.99
rs11693806	2	218292158	C	G	0.75	0.36	0.53	C	0.32	1.43	(1.33 , 1.54)	0.26	0.27	0.25	0.29	0.28	0.18	-0.05	(-0.08 , 0.02)	0.11
rs6793295	3	169518455	T	C	0.27	0.68	0.39	T	0.76	1.23	(1.15 , 1.33)	0.72	0.70	0.74	0.66	0.75	0.74	0.02	(-0.05 , 0.16)	0.70
rs10069690	5	1279790	C	T	0.27	0.15	0.59	T	0.28	1.20	(1.12 , 1.29)	0.29	0.31	0.28	0.32	0.29	0.21	-0.06	(-0.09 , 0.00)	0.054
rs73227498	5	111485904	A	T	0.14	0.02	0.08	A	0.87	1.37	(1.23 , 1.49)	0.91	0.92	0.92	0.90	0.93	0.90	-0.04	(-0.08 , 0.12)	0.46
rs2466076	8	32432796	G	T	0.53	0.85	0.54	G	0.48	1.32	(1.23 , 1.41)	0.56	0.56	0.56	0.55	0.67	0.48	-0.04	(-0.08 , 0.03)	0.20
rs1588635*	9	100537802	A	C	0.65	0.89	0.81	A	0.40	1.70	(1.59 , 1.82)	0.50	0.52	0.52	0.55	0.42	0.34	-0.11	(-0.18 , -0.02)	0.012
rs7902587	10	105694301	C	T	0.10	0.00	0.17	T	0.11	1.41	(1.27 , 1.56)	0.10	0.11	0.10	0.12	0.13	0.06	-0.05	(-0.09 , 0.08)	0.31
rs368187	14	36532576	G	C	0.45	0.55	0.76	G	0.58	1.39	(1.30 , 1.47)	0.58	0.59	0.59	0.56	0.61	0.53	-0.02	(-0.06 , 0.06)	0.57
rs116909374*	14	36738361	C	T	0.03	0.00	0.01	T	0.04	1.71	(1.47 , 2.00)	0.03	0.02	0.03	0.03	0.01	0.02	-0.10	(-0.16 , 0.19)	0.28
rs56062135	15	67455630	C	T	0.23	0.02	0.09	T	0.25	1.24	(1.16 , 1.34)	0.28	0.28	0.28	0.32	0.25	0.24	-0.04	(-0.07 , 0.04)	0.28
rs2289261	15	67457485	G	C	0.64	0.35	0.60	C	0.68	1.23	(1.15 , 1.32)	0.63	0.65	0.62	0.68	0.56	0.53	-0.06	(-0.08 , -0.01)	0.030

Abbreviations: African ancestry (AFR); alternate allele (alt); chromosome (chr); East Asian ancestry (EAS); non-Finnish European ancestry (NFE); referent allele (ref).

* Position from UCSC human genome assembly hg19.

† P-values were generated using likelihood ratio tests, comparing model fit with and without each SNP (modeled assuming an additive genetic effect), based on multivariable proportional odds models mutually adjusted for radiation dose (linear, truncated at 1 Gy), age at PTC, sex, and 10 principal components. Exceptions to this approach included use of logistic regression for rs116909374 because no individuals in our study were homozygous for the effect allele, and use of an ordinal dose variable (based on the means of the categories above) for rs1588635 because the linear EOR/Gy model would not converge.

Table S16. Distribution of rare potentially protein-damaging variants in selected gene sets or pathways in our study population, overall and by radiation dose, and in external reference population data from gnomAD.

	Number of alleles			Number of individuals					
	gnomAD NFE exomes (n=51,377 individuals)*	gnomAD NFE genomes (n=7718 individuals)*	Chernobyl Tissue Bank, Total (n=383 individuals)	Chernobyl Tissue Bank, Total (n=383)	Radiation dose (mGy) 0 (n=78)	1-99 (n=165)	100-199 (n=73)	200-499 (n=36)	≥500 (n=31)
Gene set or pathway†	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
Thyroid susceptibility	156 (0.2%)	37 (0.2%)	4 (0.5%)	4 (1.0%)	2 (2.6%)	2 (1.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Cancer susceptibility	7509 (7.3%)	1248 (8.1%)	57 (7.4%)	50 (13.1%)	9 (11.5%)	22 (13.3%)	10 (13.7%)	3 (8.3%)	6 (19.4%)
Autosomal dominant cancer susceptibility	1520 (1.5%)	303 (2.0%)	15 (2.0%)	15 (3.9%)	4 (5.1%)	7 (4.2%)	2 (2.7%)	1 (2.8%)	1 (3.2%)
Clinical radiation sensitivity	4063 (4.0%)	689 (4.5%)	27 (3.5%)	26 (6.8%)	8 (10.3%)	8 (4.8%)	5 (6.8%)	3 (8.3%)	2 (6.5%)
DNA damage response, total	29952 (29.1%)	3822 (24.8%)	228 (29.8%)	172 (44.9%)	42 (53.8%)	65 (39.4%)	33 (45.2%)	21 (58.3%)	11 (35.5%)
Single strand break repair	13824 (13.5%)	1066 (6.9%)	43 (5.6%)	39 (10.2%)	8 (10.3%)	16 (9.7%)	5 (6.8%)	6 (16.7%)	4 (12.9%)
Base excision repair	11167 (10.9%)	567 (3.7%)	22 (2.9%)	19 (5.0%)	4 (5.1%)	8 (4.8%)	2 (2.7%)	3 (8.3%)	2 (6.5%)
Nucleotide excision repair	2183 (2.1%)	401 (2.6%)	16 (2.1%)	14 (3.7%)	3 (3.8%)	6 (3.6%)	2 (2.7%)	1 (2.8%)	2 (6.5%)
Mismatch repair	1056 (1.0%)	208 (1.3%)	6 (0.8%)	6 (1.6%)	2 (2.6%)	2 (1.2%)	2 (2.7%)	0 (0.0%)	0 (0.0%)
Direct repair	432 (0.4%)	58 (0.4%)	6 (0.8%)	6 (1.6%)	0 (0.0%)	2 (1.2%)	1 (1.4%)	2 (5.6%)	1 (3.2%)
Double strand break repair	9548 (9.3%)	1581 (10.2%)	101 (13.2%)	91 (23.8%)	22 (28.2%)	35 (21.2%)	22 (30.1%)	7 (19.4%)	5 (16.1%)
Fanconi anemia pathway	3545 (3.4%)	666 (4.3%)	44 (5.7%)	44 (11.5%)	12 (15.4%)	13 (7.9%)	12 (16.4%)	4 (11.1%)	3 (9.7%)
Homologous recombination	6245 (6.1%)	992 (6.4%)	54 (7.0%)	51 (13.3%)	14 (17.9%)	23 (13.9%)	6 (8.2%)	4 (11.1%)	4 (12.9%)
Non-homologous end joining	1541 (1.5%)	235 (1.5%)	15 (2.0%)	14 (3.7%)	5 (6.4%)	3 (1.8%)	5 (6.8%)	1 (2.8%)	0 (0.0%)
Translesion synthesis	1855 (1.8%)	309 (2.0%)	21 (2.7%)	21 (5.5%)	4 (5.1%)	8 (4.8%)	5 (6.8%)	1 (2.8%)	3 (9.7%)
Checkpoint factor	2390 (2.3%)	406 (2.6%)	27 (3.5%)	25 (6.5%)	6 (7.7%)	6 (3.6%)	7 (9.6%)	3 (8.3%)	3 (9.7%)
Chromatin remodeling	458 (0.4%)	75 (0.5%)	4 (0.5%)	4 (1.0%)	3 (3.8%)	1 (0.6%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Chromosome segregation	611 (0.6%)	120 (0.8%)	10 (1.3%)	9 (2.3%)	1 (1.3%)	5 (3.0%)	1 (1.4%)	2 (5.6%)	0 (0.0%)
Nucleotide pools	122 (0.1%)	26 (0.2%)	2 (0.3%)	2 (0.5%)	0 (0.0%)	1 (0.6%)	0 (0.0%)	0 (0.0%)	1 (3.2%)
p53 pathway	115 (0.1%)	15 (0.1%)	1 (0.1%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	1 (1.4%)	0 (0.0%)	0 (0.0%)
Telomere maintenance	1343 (1.3%)	204 (1.3%)	14 (1.8%)	13 (3.4%)	2 (2.6%)	6 (3.6%)	1 (1.4%)	4 (11.1%)	0 (0.0%)
Topoisomerase and topoisomerase damage reversal	114 (0.1%)	20 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Ubiquitin response	525 (0.5%)	65 (0.4%)	6 (0.8%)	6 (1.6%)	2 (2.6%)	3 (1.8%)	0 (0.0%)	0 (0.0%)	1 (3.2%)

* Carrier frequencies from the gnomAD whole exome and whole genome sequencing are provided for reference only. Although the gnomAD data were restricted to individuals of non-Finnish European (NFE) descent, the methods are not directly comparable to the Chernobyl Tissue Bank samples due to likely differences in both the sequencing methods (e.g., capture kit, sequencing chemistries) and bioinformatics analysis (e.g., variants in gnomAD were only called using HaplotypeCaller and utilized different quality control filters). Additionally, note that gnomAD data are provided at the allelic rather than individual level.

† Details on the gene sets and pathways are provided in the Materials and Methods and Table S17.

Table S17. Distribution of rare potentially protein-damaging variants in selected genes in our study population, overall and by radiation dose, and in external reference population data from gnomAD.

Provided as a separate Excel file available on the *Science* website.

Table S18. Relationship between detailed molecular characteristics and radiation dose from ^{131}I exposure, separately for characteristics modeled using linear regression and logistic or proportional odds regression.

Molecular characteristics	Regression model*	Effect size (per 100 mGy) β	(95% CI)	P*
Small deletions				
Total	Linear regression	2.16	(1.44 , 2.87)	8.0E-09
Subclonal	Linear regression	-0.06	(-0.57 , 0.45)	8.2E-01
Clonal	Linear regression	1.48	(1.06 , 1.89)	9.9E-12
Length: 1-4 base pairs	Linear regression	0.0094	(0.0062 , 0.013)	1.9E-08
Length: ≥5 base pairs (see proportional odds models below)				
Deletion:SNV ratio				
Total	Linear regression	0.0039	(0.0032 , 0.0047)	4.9E-21
Subclonal	Linear regression	0.0009	(0.0001 , 0.0017)	2.3E-02
Clonal	Linear regression	0.0071	(0.0056 , 0.0086)	1.8E-18
ID5				
Total	Linear regression	1.07	(0.53 , 1.61)	1.3E-04
Subclonal	Linear regression	-0.06	(-0.43 , 0.31)	7.4E-01
Clonal	Linear regression	0.72	(0.39 , 1.06)	2.3E-05
ID8				
Total	Linear regression	1.15	(0.77 , 1.54)	7.3E-09
Subclonal	Linear regression	0.08	(-0.14 , 0.29)	4.9E-01
Clonal	Linear regression	0.80	(0.60 , 1.00)	2.7E-14
		EOR	(95% CI)	
PTC driver type				
Fusion vs. mutation PTC driver	Logistic regression	0.93	(0.37 , 2.11)	6.6E-08
Sub-chromosome somatic copy number deletions				
Total	Proportional odds	0.23	(0.07 , 0.51)	7.0E-04
Subclonal	Logistic regression	0.29	(-0.05 , 2.39)	1.7E-01
Clonal	Proportional odds	0.19	(0.03 , 0.48)	9.7E-03
Confirmed SVs†				
Total simple/balanced	Proportional odds	2.09	(0.89 , 5.07)	2.5E-13
Subclonal simple/balanced	Logistic regression	0.01	(-0.08 , 0.31)	9.1E-01
Clonal simple/balanced	Proportional odds	4.65	(1.81 , 12.98)	1.4E-16
Clonal simple/balanced/EJ	Logistic regression	18.81	(5.07 , 94.61)	5.5E-19
<4 bp intervening loss/gain	Logistic regression‡	7.84	(>0 , 150)	2.0E-08
≥4 bp intervening loss/gain	Logistic regression	3.61	(0.86 , 24.44)	6.1E-08
Inversion	Logistic regression	21.42	(3.98 , 400)	3.6E-14
Translocation	Logistic regression	0.74	(0.17 , 3.24)	4.4E-04
Clonal non-simple/balanced/EJ	Proportional odds	0.04	(-0.04 , 0.18)	4.1E-01
Clonal deletions, Length: ≥5 bp	Proportional odds	0.56	(0.38 , 0.79)	5.1E-29
Patterns consistent with EJ				
Total	Proportional odds	0.61	(0.42 , 0.86)	4.9E-31
0-1 bp microhomology at the deletion boundary	Proportional odds	0.54	(0.33 , 0.84)	1.3E-18
≥2 bp microhomology at the deletion boundary	Proportional odds	0.48	(0.27 , 0.80)	3.6E-13

Clonality of ID-EJ (ID6, ID8 combined) was not considered because no ID6 mutations were identified when the analyses were restricted to clonal mutations.

* P-values and effect estimates were derived from multivariable generalized linear, proportional odds, or logistic regression models, depending on the distribution of the molecular characteristic. All models were mutually adjusted for radiation dose (linear, truncated at 1 Gy), age at PTC, and sex.

† Analyses were restricted to confirmed SVs, as described in the Materials and Methods and Table S6.

‡ Lower bound would not converge.

Table S19. Analyses of genomic locus-based attributes for small insertions and deletions.

Attribute	Unexposed		1-99 mGy		100-199 mGy		200-499 mGy		≥500 mGy		
	β	P	β	P	β	P	β	P	β	P	P_{trend}
Insertions											
Replication timing	-0.005	3.79E-03	-0.006	1.78E-07	-0.006	1.22E-03	-0.003	0.30	-0.009	7.75E-04	0.40
GC content	-3.053	2.92E-12	-3.499	8.61E-40	-3.161	7.30E-14	-2.486	9.56E-06	-3.107	3.39E-07	0.33
Trinucleotide complexity	22.997	8.72E-06	22.661	5.14E-09	24.696	2.40E-06	23.320	7.96E-05	25.079	1.89E-05	0.15
ALU repeats	0.193	0.044	0.126	0.032	0.179	0.055	0.256	0.036	0.055	0.69	0.52
MIR repeats	0.004	0.98	-0.177	0.040	-0.187	0.19	0.122	0.46	-0.283	0.19	0.88
L1 repeats	0.145	0.058	0.015	0.75	0.022	0.77	-0.004	0.97	0.172	0.10	0.63
L2 repeats	-0.165	0.27	-0.038	0.65	-0.201	0.17	-0.219	0.27	0.149	0.42	0.39
LTR repeats	-10.032	0.95	-10.125	0.92	-10.107	0.95	-9.517	0.96	-9.312	0.96	na
DNA repeats	0.037	0.78	-0.024	0.77	-0.149	0.28	-0.318	0.11	0.275	0.10	0.16
Simple repeats	-0.139	0.37	-0.105	0.27	-0.114	0.46	-0.233	0.26	0.280	0.18	0.34
Genes	0.122	0.073	0.083	0.042	0.005	0.94	-0.007	0.93	-0.033	0.72	0.23
CpG islands	0.649	0.10	0.827	2.28E-04	1.073	6.48E-04	0.821	0.062	0.085	0.91	0.34
Direct repeats	0.153	0.28	-0.025	0.77	-0.254	0.085	0.116	0.54	-0.423	0.048	0.12
G-quadruplexes	-0.234	0.39	-0.220	0.18	-0.154	0.54	-0.094	0.77	-0.032	0.93	0.61
Cruciform inverted repeats	0.014	0.84	-0.002	0.97	0.055	0.40	0.131	0.14	-0.059	0.53	0.53
MIR repeats	-0.083	0.46	-0.010	0.88	0.159	0.12	-0.083	0.58	0.075	0.62	0.67
Short tandem repeats	0.414	1.82E-06	0.442	1.98E-17	0.433	2.29E-07	0.440	9.59E-05	0.625	6.83E-08	0.45
z-DNA motif	-0.280	0.20	0.067	0.55	0.059	0.75	-0.261	0.35	0.007	0.98	0.72
LADs	0.028	0.70	-0.038	0.37	-0.064	0.36	0.100	0.28	-0.165	0.10	0.53
Distance from the centromere	0.346	2.77E-05	0.333	1.43E-11	0.320	5.88E-05	0.445	6.03E-05	0.358	1.74E-03	0.77
Distance from the nearest telomere	0.138	2.78E-05	0.089	1.12E-07	0.112	1.24E-04	0.123	2.46E-03	0.162	1.45E-03	0.45
Chromatin state (heterochromatin)	4.460	5.01E-04	-9.592	0.98	-9.664	0.99	-9.172	0.99	-8.688	0.99	0.52
Deletions											
Replication timing	-0.007	6.19E-08	-0.005	2.58E-11	-0.007	5.81E-09	-0.003	0.06	-0.001	0.37	0.010
GC content	-2.634	4.42E-19	-2.220	5.05E-39	-1.862	6.94E-13	-2.277	1.28E-09	-1.341	5.33E-05	0.015
Trinucleotide complexity	21.906	1.92E-07	24.192	4.77E-17	18.986	1.85E-06	20.155	4.50E-05	14.530	4.86E-03	0.050
ALU repeats	-0.040	0.58	-0.142	6.42E-04	-0.214	9.45E-04	0.021	0.81	-0.074	0.35	0.24
MIR repeats	-0.117	0.22	-0.054	0.31	-0.116	0.17	-0.105	0.40	-0.115	0.29	0.82
L1 repeats	-0.066	0.24	-0.066	0.041	-0.056	0.26	-0.058	0.42	-0.096	0.14	0.46
L2 repeats	0.077	0.40	-0.057	0.30	-0.088	0.31	0.072	0.54	0.002	0.99	0.94
LTR repeats	-9.828	0.93	0.468	0.45	0.246	0.81	-9.262	0.93	-9.562	0.93	0.64
DNA repeats	-0.138	0.16	-0.108	0.053	-0.076	0.37	-0.018	0.88	-0.076	0.48	0.92
Simple repeats	0.194	0.070	0.237	1.24E-04	-0.049	0.61	0.081	0.56	0.130	0.30	0.57
Genes	0.019	0.69	-0.024	0.39	-0.025	0.55	0.076	0.22	-0.072	0.18	0.22
CpG islands	1.458	5.74E-17	1.146	1.46E-25	1.270	2.69E-16	1.124	9.31E-07	0.581	0.017	0.014
Direct repeats	0.028	0.79	-0.047	0.43	0.145	0.10	-0.138	0.31	0.014	0.91	0.71
G-quadruplexes	0.134	0.37	0.125	0.14	0.084	0.51	0.379	0.026	-0.083	0.64	0.29
Cruciform inverted repeats	0.101	0.029	0.081	2.67E-03	0.081	0.052	0.209	4.58E-04	0.172	1.21E-03	0.065
MIR repeats	-0.098	0.22	-0.016	0.73	-0.005	0.94	0.181	0.062	-0.012	0.89	0.88
Short tandem repeats	0.029	0.67	0.111	3.65E-03	0.244	1.71E-05	0.084	0.32	-0.044	0.58	0.33
z-DNA motif	0.062	0.65	-0.272	2.15E-03	-0.227	0.082	0.253	0.11	-0.085	0.62	0.74
LADs	-0.080	0.10	0.006	0.84	-0.072	0.10	-0.066	0.30	-0.012	0.84	0.70
Distance from the centromere	0.423	2.45E-13	0.374	4.16E-30	0.407	1.95E-15	0.390	1.96E-07	0.464	6.19E-12	0.34
Distance from the nearest telomere	0.086	8.69E-06	0.082	2.93E-14	0.075	4.71E-06	0.051	0.017	0.083	9.33E-05	0.96
Chromatin state (heterochromatin)	-9.069	0.98	-7.283	0.93	-6.422	0.94	-8.795	0.98	-8.995	0.98	na

Abbreviation: not applicable (na). No indel was observed in our study for that genomic feature, thus the models testing for trend in radiation dose excluded those parameters.

Table S20. Sensitivity analyses restricted to exposed (total) and those with <500 mGy exposure, and tests for departure from a linear radiation dose-response relationship.

		Linear dose model			Tests for departure from linearity*	
Molecular characteristics	Regression model*	Effect size (per 100 mGy)*		P*	Linear quadratic P*	Linear exponential P*
		β	(95% CI)			
Polygenic risk score						
Total	Linear regression	-0.08	(-0.13 , -0.04)	4.7E-04	0.99	0.54
Exposed		-0.08	(-0.12 , -0.03)	2.1E-03		
Exposed, <500 mGy		0.00	(-0.15 , 0.15)	9.9E-01		
Clonal small deletions						
Total	Linear regression	1.48	(1.06 , 1.89)	9.9E-12	6.63E-05	1.77E-03
Exposed		1.61	(1.14 , 2.08)	1.0E-10		
Exposed, <500 mGy		-0.65	(-1.88 , 0.58)	3.0E-01		
Clonal deletion:SNV ratio						
Total	Linear regression	0.0071	(0.0056 , 0.0086)	1.8E-18	0.86	0.25
Exposed		0.0071	(0.0053 , 0.0088)	5.0E-14		
Exposed, <500 mGy		0.0046	(0.0009 , 0.0084)	1.5E-02		
Clonal ID5						
Total	Linear regression	0.72	(0.39 , 1.06)	2.3E-05	0.03	0.21
Exposed		0.80	(0.42 , 1.18)	3.7E-05		
Exposed, <500 mGy		-0.21	(-1.23 , 0.81)	6.8E-01		
Clonal ID8						
Total	Linear regression	0.80	(0.60 , 1.00)	2.7E-14	0.08	0.18
Exposed		0.82	(0.59 , 1.05)	1.5E-11		
Exposed, <500 mGy		0.22	(-0.32 , 0.77)	4.2E-01		
		EOR	(95% CI)			
Fusion vs. mutation PTC driver						
Total	Logistic regression	0.93	(0.37 , 2.11)	6.6E-08	0.54	0.56
Exposed		0.58	(0.16 , 1.89)	8.0E-05		
Exposed, <500 mGy		0.59	(0.04 , 2.66)	2.6E-02		
Clonal sub-chromosome somatic copy number deletions						
Total	Proportional odds	0.19	(0.03 , 0.48)	9.7E-03	0.01	†
Exposed		0.59	(0.15 , 2.35)	1.7E-04		
Exposed, <500 mGy		0.01	(-0.21 ,1.10)	9.8E-01		
Clonal simple/balanced/EJ SVs						
Total	Logistic regression	18.81	(5.07 , 94.61)	5.5E-19	0.60	0.60
Exposed		4.37	(0.76 , 476)	4.7E-08		
Exposed, <500 mGy		12.38	(0.90 , 1894)	4.6E-05		
Clonal deletions, Length: ≥5 bp, EJ						
Total	Proportional odds	0.61	(0.42 , 0.86)	4.9E-31	8.4E-09	†
Exposed		0.73	(0.46 , 1.18)	9.0E-28		
Exposed, <500 mGy		0.13	(-0.03 , 0.40)	1.4E-01		

* P-values were generated using likelihood ratio tests, comparing model fit with and without the variable of interest (dose, dose squared, or exp(dose)), based on multivariable generalized linear, proportional odds, or logistic regression models, depending on the distribution of the molecular characteristic. All models were mutually adjusted for radiation dose (linear, truncated at 1 Gy), age at PTC, and sex, unless otherwise noted below.

† Model did not converge.

Table S21. Tests for departure from a multiplicative relationship between radiation dose from ^{131}I exposure and age at PTC, age at exposure, and latency, among exposed individuals.

Molecular characteristics †	Regression model*	Dose × Age at PTC P*	Dose × Age at exposure P*	Dose × Latency P*
Polygenic risk score	Linear regression	7.8E-01	2.2E-01	5.2E-01
Clonal small deletions	Linear regression	5.4E-01	2.9E-02 ‡	2.4E-02 ‡§
Clonal deletion:SNV ratio	Linear regression	2.3E-01	1.8E-04 ‡	1.2E-01
Clonal ID5	Linear regression	6.8E-01	1.1E-01	4.5E-01
Clonal ID8	Linear regression	3.6E-01	1.4E-03 ‡	2.8E-03 ‡§
Fusion vs. mutation PTC driver	Logistic regression	2.4E-02 ‡‖	2.4E-02 ‡	9.2E-01
Clonal sub-chromosome somatic copy number deletions	Proportional odds	6.1E-02	1.6E-01	2.6E-02 ‡
Clonal ≥5 bp EJ deletions	Proportional odds	4.9E-01	1.1E-04 ‡	1.6E-02 ‡§

* P derived from likelihood ratio tests comparing model fit with and without the interaction term of interest. Multivariable models included age at PTC (continuous), dose (continuous, truncated at 1 Gy), and sex. Models testing the multiplicative relationship of dose × age at exposure (continuous) or dose × latency (continuous) also included those main effects.

† Clonal simple/balanced/EJ SVs were excluded from these analyses because models did not converge.

‡ Indicates $P<0.05$.

§ Because all models adjusted for age at PTC, the latency effects for these variables appeared to actually reflect effects of age at exposure (because age at PTC=age at exposure + latency). Additional modeling evaluating interactions for dose with both age at exposure and latency (excluding age at PTC) demonstrated stronger interaction for age at exposure.

‖ Additional modeling of the driver type considering both age at PTC and age at exposure suggested a stronger interaction for age at exposure.

Table S22. Relationship between detailed molecular characteristics and radiation dose from ^{131}I exposure, by age at exposure, separately for characteristics modeled using linear regression and logistic or proportional odds regression.

Molecular characteristics Age at exposure	Regression model*	Effect size (per 100 mGy)		P*
Clonal small deletions	Linear regression			
<5 years		2.19	(1.34 , 3.04)	1.3E-06
5-9 years		1.36	(0.66 , 2.07)	2.3E-04
≥10 years		0.86	(0.04 , 1.68)	3.9E-02
Clonal deletion:SNV ratio	Linear regression			
<5 years		0.0090	(0.005 , 0.013)	1.5E-05
5-9 years		0.0059	(0.0035 , 0.0082)	2.7E-06
≥10 years		0.0023	(0.00076 , 0.0039)	3.8E-03
Clonal ID5	Linear regression			
<5 years		1.23	(0.58 , 1.88)	2.9E-04
5-9 years		0.46	(-0.22 , 1.13)	1.8E-01
≥10 years		0.40	(-0.31 , 1.10)	2.7E-01
Clonal ID8	Linear regression			
<5 years		1.15	(0.70 , 1.61)	1.8E-06
5-9 years		0.82	(0.45 , 1.18)	2.5E-05
≥10 years		0.21	(-0.15 , 0.57)	2.5E-01
		EOR	(95% CI)	
Fusion vs. mutation PTC driver†	Logistic regression			
<10 years		3.66	(0.50 , 308)	7.3E-06
≥10 years		0.15	(-0.04 , 1.03)	2.1E-01
Clonal ≥5 bp EJ deletions	Proportional odds			
<5 years		2.42	(0.78 , 492)	4.6E-17
5-9 years		1.21	(0.43 , 4.66)	1.8E-09
≥10 years		0.16	(0.03 , 0.41)	9.2E-03

* P-values and effect estimates were derived from multivariable generalized linear, proportional odds, or logistic regression models, depending on the distribution of the molecular characteristic. All models were stratified (run separately) by the age at exposure groupings specified in the table, with each model mutually adjusting for radiation dose (linear, truncated at 1 Gy), age at PTC, and sex.

† Models evaluating the effect of dose on driver type restricted to <5 years of age at exposure did not converge, so we combined individuals exposed at <5 and 5-9 years. EOR/100 mGy for 5-9 years alone=1.78, 95%CI=0.12-226.