In-depth characterization of the cisplatin mutational signature in a human cell line and in esophageal and liver tumors

Arnoud Boot[1,2], Mi Ni Huang[1,2], Alvin W.T. Ng[2,3], Yoshiiku Kawakami[4], Kazuaki Chayama[4], Bin Tean Teh[5], Hidewaki Nakagawa[6], Steven G. Rozen[1,2]*

[1] Cancer and Stem Cell Biology, Duke-NUS Medical School, 8 College Road, 169857, Singapore

[2] Centre for Computational Biology, Duke-NUS Medical School, 8 College Road, 169857, Singapore

[3] NUS Graduate School for Integrative Sciences and Engineering, 28 Medical Drive, 117456, Singapore

[4] Department of Gastroenterology and Metabolism, Graduate School of Biomedical and Health Sciences, Hiroshima University, 1-2-3, Kasumi, Minami-ku, Hiroshima, 734-8551, Japan

[5] Division of Medical Sciences, National Cancer Centre Singapore.

[6] Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan.

Corresponding author:

*Steven G. Rozen, steve.rozen@duke-nus.edu.sg

Working title:

Cisplatin mutational signature in tumors

Keywords:

Experimental delineation of mutational signatures; mutational signatures of DNA-damaging agents; dinucleotide substitutions; hepatocellular carcinoma

Abstract

Background and aims

Cisplatin reacts with DNA, and thereby likely generates a characteristic pattern of somatic mutations, called a mutational signature. Despite the widespread use of cisplatin in cancer treatment and its role in contributing to secondary malignancies, its mutational signature has not been delineated. Delineation of the mutational signature of cisplatin would enable identification of cisplatin-induced secondary malignancies, and consequently improve screening for secondary malignancies after cisplatin chemotherapy.

Methods

We sequenced the whole genomes of 6 independent clones of cisplatin-exposed MCF-10A cells, and delineated the patterns of single- and di-nucleotide mutations in each clone in terms of flanking sequence context, transcription strand bias, and other characteristics. We used statistical tests and non-negative matrix factorization to search for these signatures in hepatocellular carcinomas and esophageal adenocarcinomas.

Results

All clones showed highly consistent patterns of single- and di-nucleotide substitutions in the contexts of immediately flanking bases. The proportion of dinucleotide substitutions was high: 7.3% of single nucleotide substitutions were part of dinucleotide substitutions, presumably due to cisplatin's propensity to form intra strand and inter-strand crosslinks between purine bases in DNA. Statistical and non-negative-factorization-based analyses identified likely cisplatin exposure in 8 hepatocellular carcinomas and 2 esophageal adenocarcinomas. All hepatocellular carcinomas for which clinical data were available and both esophageal cancers indeed had histories of prior cisplatin treatment.

Conclusions

We experimentally delineated the mutational signature of cisplatin based on the patterns of the single nucleotide and dinucleotide substitutions. This signature enabled us to detect previous cisplatin exposure in human hepatocellular carcinomas and esophageal adenocarcinomas with high confidence.

Introduction

For 40 years, cisplatin and its derivatives have been cornerstones of the treatment of almost every type of cancer (Dasari and Tchounwou 2014; Dugbartey et al. 2016). However, cisplatin treatment often causes numerous side effects, including hepatotoxicity (Waseem et al. 2015; Dugbartey et al. 2016), and it increases the risk of developing secondary malignancies. For example, cisplatin based treatments almost always cure testicular cancers, but increase the risk of developing a solid tumor later in life 1.8-fold (Travis et al. 2005), and cisplatin treatment of several types of cancers increases the incidence of secondary leukemia's (Ratain et al. 1987; Kushner et al. 1998). Cisplatin's therapeutic properties depend partly on its DNA damaging activity, and the risk of secondary malignancies presumably stems from the consequent mutagenesis (Choi et al. 2014). This highlights the importance of understanding cisplatin mutagenesis and how it promotes carcinogenesis. This also highlights the need for a biomarker to identify cisplatin-induced secondary malignancies.

The mechanisms of cisplatin induced DNA damage have been extensively studied. When cisplatin enters the cells, its two chloride atoms are hydrolyzed, resulting in two positive charges (Masters and Koberle 2003; Behmand et al. 2015). Although the hydrolyzed molecule presumably reacts with many molecules in the cell, its therapeutic cytotoxicity is generally considered to stem from reactions with the N7 atoms of purine bases in DNA (Harrington et al. 2010; Dasari and Tchounwou 2014; Behmand et al. 2015). Most cisplatin-DNA adducts are crosslinks between two adjacent guanines (GpG, 65%) or between an adenine and a guanine (5'-ApG-3', 25%). Mono-adducts and interstrand crosslinks are much rarer (Jamieson and Lippard 1999; Masters and Koberle 2003; Enoiu et al. 2012). Cisplatin induced DNA intrastrand crosslinks and mono-adducts are repaired through nucleotide excision repair (NER) (Zamble et al. 1996; Reardon et al. 1999; Hu et al. 2016). Interstrand crosslinks are the most difficult to repair and the most cytotoxic, because they covalently link the two strands of the DNA helix and consequently block transcription and replication

(Jamieson and Lippard 1999; Masters and Koberle 2003; Enoiu et al. 2012; Hashimoto et al. 2016; Roy and Scharer 2016). The mechanisms of interstrand-crosslink repair have not yet been fully elucidated but appear to be complicated (Hashimoto et al. 2016; Roy and Scharer 2016).

Cisplatin likely causes a characteristic pattern of single-nucleotide substitutions (SNSs), known as a mutational signature, along with possible additional features including fewer mutations on the transcribed strands of genes and association with small insertions and deletions (indels) or dinucleotide substitutions (DNSs) (Alexandrov et al. 2013a). Currently 30 mutational signatures are widely recognized, and they have a variety of known, suspected or unknown causes (Alexandrov et al. 2013a; Alexandrov et al. 2013b; Wellcome Trust Sanger Institute 2016). Mutational signatures can serve as biomarkers for endogenous mutagenic processes and exposures that led to the development of tumors. We hypothesize that cisplatin's mutational signature can serve as biomarker to identify cisplatin-induced secondary malignancies, to improve screening for secondary malignancies after cisplatin chemotherapy, and to identify which tissues are especially vulnerable to secondary malignancies after cisplatin treatment.

Two previous studies investigated the mutational signature of cisplatin, one in *Caenorhabditis elegans* and one in a chicken (*Gallus gallus*) B-cell cell line (Meier et al. 2014; Szikriszt et al. 2016). Although both studies reported mutational signatures with primarily C>A mutations, the SNS signature were otherwise dissimilar: the *C. elegans* signature was dominated by CCA>CAA and CCT>CAT mutations, while the chicken signature was dominated by NCC>NAC mutations (where N is any base). This lack of similarity may have been due to the different model systems used, to the low numbers of mutations in the *C. elegans* study, or to experimental differences between the studies. In any case, these studies failed to unequivocally elucidate the mutational signature of cisplatin.

Therefore, we studied cisplatin mutations in MCF-10A cells, a non-tumorigenic human breast epithelial cell line. Here we report the extensive characterization of the

cisplatin signature obtained, as well as its discovery in hepatocellular carcinomas and esophageal adenocarcinomas in patients previously exposed to cisplatin.

Results

Cisplatin's single-nucleotide substitution signature

We exposed two independent cultures of MCF-10A cells to 0.5 μM and 1 μM of cisplatin once a week for 8 weeks. Single cells were isolated and expanded for whole-genome sequencing and mutational analysis. We sequenced untreated MCF-10A and 3 cisplatin-exposed clones from each concentration, one exposed for 4 weeks and 2 exposed for 8 weeks. Mean coverage was >33x, and in total we identified 30,153 SNSs and 1,708 indels (Supplementary Table 1).

The SNS mutation spectra from all 6 clones were highly similar (Figure 1A, Supplementary Figure 1A, Supplementary Table 2, all Pearson correlations > 0.958 and cosine similarities > 0.971). The most prominent features were two C>T peaks (CCC>CTC and CCT>CTT) and four T>A peaks (CT>CA). There were also substantial numbers of C>A mutations (~22.8% of all mutations), and peaks at GCC>GAC and GCC>GGC. Figure 1B and Supplementary Figure 1B display the signatures as mutation rates per trinucleotide, which better reflects the sequence specificity of mutational processes because they are not affected by differences in trinucleotides abundances. For example, Figure 1B shows more prominent CCC>CTC peaks and reveals that the gap at CCG>CTG in Figure 1A reflects the low abundance of CCG trinucleotides in the genome rather than reduced mutagenicity.

In addition to consistent patterns of the bases immediately 5' and 3' of cisplatin SNSs, there were also many preferences 2 bp 5' and 3' of the SNSs (Figure 1C, Supplementary figure 2). For example, CT>CA mutations were usually preceded by an A (ACT>ACA). Similarly, CC>CT mutations were usually preceded by a pyrimidine (YCC>YCT). These and other preferences at the -2 bp or +2 bp positions were statistically significant (Supplementary figure 3).

Associations of cisplatin-induced single-nucleotide substitutions with genomic features

Many mutational processes cause fewer mutations due to damage on the transcribed strands of genes than on the non-transcribed strands. This is termed transcription strand bias and is due to transcription-coupled nucleotide excision repair (TC-NER) of adducted bases in the transcribed (antisense) strands. Since cisplatin forms adducts on purines, we would expect reduced numbers of mutations when G and A is on the transcribed strand (C and T are on the sense strand). As expected, C>A, C>T and T>A SBSs were strongly reduced on the non-transcribed strand (Supplementary Figure 4) (Fousteri and Mullenders 2008; Harrington et al. 2010; Dasari and Tchounwou 2014; Behmand et al. 2015; Hu et al. 2016). Also consistent with TC-NER, strand bias for C>A, C>T and T>A mutations was stronger in more highly expressed genes ($p = 2.20\times10^{-16}$, one-sided Chi-squared test for all MCF-10A clones combined, Figure 2A, Supplementary Figure 5). Finally, TC-NER efficiency decreases from the 5' to the 3' ends of transcripts (Conaway and Conaway 1999; Hu et al. 2015; Huang et al. 2017). Consistent with this, strand bias for C>A, C>T and T>A SNSs decreased toward the 3' ends of transcripts ($p = 2.46\times10^{-12}$, logistic regression for all MCF-10A clones combined, Figure 2B, Supplementary Figure 6).

For some mutational processes, the intensity of mutagenesis is associated with chromatin state (Polak et al. 2015; Seplyarskiy et al. 2015; Kaiser et al. 2016). Additionally, there is increased cisplatin adduct formation in open chromatin compared to closed chromatin (Hu et al. 2016). In the MCF-10A cells, regions with histone marks indicative of active promoters, enhancers and actively transcribed genes were less highly mutated , and regions with histone marks associated with heterochromatin and transcriptional repression were more highly mutated (Figure 2C).

DNSs in cisplatin signature

To investigate the presence of DNSs in the cisplatin genomes we selected all adjacent SNS, and verified that both SNS were on the same reads (see Materials and Methods). We identified 1,106 DNSs in the cisplatin genomes, of which most were mutations from CC, CT, TC and TG (Figure 3A, Supplementary Figure 7). We hypothesized that mutations from CC, CT, and TC are consequences of intrastrand crosslinks at GpG, ApG and GpA, and that mutations from TG were consequences of diagonally-offset interstrand guanine-adenine crosslinks $^{5'}\mathbf{TG}^{3'}_{3'}\mathbf{AC}_{5'}$ (crosslinked bases in bold). Mutations from AT, TA and TT were rare, which is consistent with previous reports that cisplatin does not induce adenine-adenine crosslinks (Supplementary Table 3) (Jamieson and Lippard 1999; Masters and Koberle 2003).

The proportion of SNSs involved in DNSs ranged from 6.2% to 8.5%. To relate this to other mutagenic processes known to be associated with DNSs, we examined the percentage of SNSs involved in DNSs associated with COSMIC Signatures 4 (smoking-related) and 7 (due to UV exposure) (Wellcome Trust Sanger Institute 2016). We studied Signature 4 in 24 lung adenocarcinomas (Imielinski et al. 2012) and Signature 7 in 112 melanomas (Zhang et al. 2011). In both tumor types, the percentage of SNSs involved in DNSs was significantly lower than in cisplatin (Figure 3B, mean 3.5%, sd=1.4%, p=6.7×10^{-7} and mean=3.3%, sd=1.6%, p=2.1×10^{-8} respectively, 2-sided t-tests versus cisplatin). We hypothesize that this high proportion of DNSs in cisplatin stems from cisplatin's propensity to form intrastrand crosslinks between adjacent bases and to form diagonally offset interstrand crosslinks.

To investigate possible sequence context preferences of cisplatin DNSs, we plotted 1bp contexts of each reference dinucleotide, irrespective of the mutant allele (Figure 3C, Supplementary Figure 8). There was strong enrichment for TC and TG DNSs in TCT and TGG contexts. Both TC and TG DNSs were further enriched for a 5' flanking purine (Supplementary Figure 8, 9). The strongest sequence context preference was for CC>NN

mutations, 49.8% of which occur in GCCT context (Supplementary Figure 8, 9). As methodological control, we also evaluated ±1bp sequence context for DNSs associated with COSMIC Signatures 4 and 7. DNSs associated with COSMIC Signature 7 showed strong sequence context preference for most mutation classes, including CC>NN, CT>NN and TT>NN (Supplementary Figure 10). The context preferences were very different however, from those of cisplatin DNSs. By contrast, DNSs associated with COSMIC Signature 4 had only weak sequence context preferences (Supplementary Figure 10).

To assess transcription strand bias in DNSs, we examined separately the mutations hypothetically involving interstrand purine-purine crosslinks, predominantly mutations from the $^{5'}\mathbf{TG}^{3'}_{3'}\mathbf{AC}_{5'}$ configuration, and the mutations hypothetically involving intrastrand purine-purine crosslinks (predominantly mutations from CC, CT, and TC). We observed transcription strand bias at the potential intrastrand crosslink sites other than TC in most of the MCF-10A clones. (Figure 3D, Supplementary Figure 11). There was no evidence of transcription strand bias at potential interstrand crosslink sites. As methodological control, we also evaluated transcription strand bias for DNSs associated with COSMIC Signatures 4 and 7, in which we also detected strand bias (Supplementary Figure 12).

Likely cisplatin mutational signature in human tumors

We examined publicly available human tumor mutation data for evidence of the experimental cisplatin signature. Notably, mutational signature W6, which was reported in the whole genome sequences of hepatocellular carcinomas (HCCs), resembles the experimental cisplatin signature (cosine similarity = 0.803, Supplementary Figure 13) (Fujimoto et al. 2016). Although the relative proportions of the major substitution classes (C>A, C>T and T>A) are rather different between Signature W6 and our experimental cisplatin signature, the profiles within each mutation class are similar (cosine similarities for C>A, C>T and T>A of 0.887, 0.921 and 0.980 respectively, Supplementary Figure 13).

Given this resemblance, we analyzed whole-genome trinucleotide spectra from Japanese and Hong Kong HCCs (Kan et al. 2013; Fujimoto et al. 2016), using the mSigAct signature presence test (see Materials & Methods). Out of 342 HCCs, 9 showed evidence of cisplatin exposure (Table 1, Figure 4A, Supplementary Figure 14, compare with Figure 1A). To further assess presence of cisplatin mutagenesis, we also examined the dinucleotide spectra of these samples (Figure 4B, Supplementary Figure 15, compare with Figure 3A). 7 of the 9 HCCs with the cisplatin SNS spectrum also had high cosine similarities between their DNS spectra and the cisplatin signature (Figure 4C) and high numbers of DNSs relative to their total SNS load (ranging from 2.9 to 6.2%, with the median of all HCCs at 1.6%, Supplementary Figure 16A).

We also analyzed the mutational spectra of 140 esophageal adenocarcinomas (ESADs), of which 68 had been treated with cisplatin prior to surgery (Noorani et al. 2017). SNS analysis suggested 3 of the cisplatin treated ESADs had the cisplatin signature, whereas we found no evidence of cisplatin mutagenesis in any of the untreated ESADs. Of the 3 ESADs identified in the SNS analysis, the DNS analysis supported likely cisplatin exposure in 2 (Table 1, Supplementary Figures 16B, 17, 18).

We further investigated whether DNS analysis could identify cisplatin-exposed tumors that were missed by the SNS analysis. We performed semi-supervised nonnegative matrix factorization (ssNMF) on all tumors with ≥25 DNSs, specifying the cisplatin DNS signature as one input signature and asking for discovery of 1 to 7 additional signatures (Materials & Methods, Supplementary Figures 19, 20). We recovered all 7 previously identified cisplatin-positive HCCs. Additionally, we identified RK140 to have >50% of DNSs attributed to cisplatin by ssNMF. Looking closer into this sample revealed a high cosine similarity between the DNS spectrum with that of the experimental data, as well as a relatively high proportion DNSs (Table 1, Supplementary Figure 16A). Although the SNS based *p* value was not significant after multiple-testing correction, we nevertheless concluded based on the combined SNS and DNS analyses that RK140 showed strong

evidence for cisplatin mutagenesis. ssNMF also identified several other HCCs with high proportions of cisplatin-associated DNSs, but neither mSigAct nor visual inspection of the SNS spectra warranted reclassifying these samples as cisplatin positive. Similarly, ssNMF identified high proportions of cisplatin-associated DNSs in several ESADs. These included the 2 identified in our initial analysis. Of the remainder, neither mSigAct nor visual inspection of the SNS spectra warranted reclassification as cisplatin positive. None of the chemotherapy naïve ESADs displayed signs of cisplatin mutagenesis.

Like the cisplatin exposed MCF-10A cells, most HCCs and ESADs showed strong transcription strand bias at CC and CT DNSs but not at TC DNSs (Figure 4D, Supplementary Figure 21). Similarly, none of the HCCs and ESADs had detectable transcription strand bias at potential interstrand crosslink sites (mainly TG DNSs, Supplementary Figure 21).

The clinical records of the Japanese HCCs (Fujimoto et al. 2016) confirmed cisplatin exposure of all of the 7 HCCs identified positive for the cisplatin mutational signature (Table 1). All 7 had received cisplatin-based DEB-TACE (transarterial chemoembolization using drug eluting beads) several months prior to surgical resection. In addition to TACE treatment for the sampled tumor, RK205, RK241 and RK256 also had had prior malignancies (Table 1). The variant allele frequencies of the cisplatin-associated DNSs were similar to the variant allele frequencies of all other SNSs, including those not likely due to cisplatin exposure (Supplementary Table 4). This suggested that the cisplatin was an early event in tumorigenesis, which is concordant with rapid clonal expansion after DEB-TACE treatment (Zen et al. 2011). Notably, the 2 HCCs we suspected to be false-positives based on the DNS spectra (RK047 and RK309) had no record of treatment with cisplatin prior to surgery.

Discussion

We have delineated the *in vitro* multidimensional mutational signature of cisplatin with extensive characterization of patterns of SNSs in tri- and pentanucleotide contexts and their associations with genomic features, as well as the patterns of DNSs and flanking bases. We began with *in vitro* delineation because it directly links mutational signatures to etiologies and because it generates signatures that are relatively unobscured by other mutational processes. We analyzed whole genome data because these provide >50 times more mutations than exomes and consequently greater stability and reproducibility of signatures. Indeed, whole genome data are practically essential for analysis of DNSs, which are rare compared to SNSs. Importantly, with the experimentally delineated SNSs and DNSs signatures in hand, we were able to detect cisplatin mutagenesis in HCCs and ESADs with high confidence. All hepatocellular carcinomas for which clinical data were available and both esophageal cancers indeed had histories of prior cisplatin treatment. We therefore conclude that the mutational signature established here serves as a biomarker for cisplatin mutagenesis that can detect cisplatin-induced secondary malignancies.

Prior to this study, 2 different experimentally elucidated mutational signatures of cisplatin were reported, one in *Caenorhabditis elegans*, and the other in cultured chicken B-cells (DT40) (Meier et al. 2014; Szikriszt et al. 2016). Both studies found primarily C>A mutations, but in terms of SNSs in trinucleotide context, the signatures bore no resemblance to each other or to the MCF-10A signature (Supplementary Figure 22). Because of the low mutation count in the *C. elegans* data, this was true for both the DNA repair proficient worms (N2) as well as for all worms combined. Like the treated MCF-10A cells, the exposed worms and DT40 cells had relatively high numbers of DNSs relative to SNSs, with many at potential intrastrand crosslink sites (ApG, GpA and GpG). However, in neither system was it possible to discern the MCF-10A cisplatin signature in the SNS mutation spectra, due to the high number of C>A mutations (Supplementary Figure 22). We also note that the C>A mutations in the treated worms and DT40 cells do not resemble any currently known mutational

signature or artefact.(Wellcome Trust Sanger Institute 2016) The differences between the MCF-10A cisplatin signature and the *C. elegans* and DT40 signatures might stem from the different model organisms used, which may differ in DNA damage susceptibility and characteristics of DNA repair and replication errors. In any case, the differences between the previously published cisplatin spectra and the MCF-10A cisplatin signature emphasize the need for standardization of *in vitro* mutational signature models. We propose that it is prudent to use human cell lines for experimental elucidation of mutational signature etiology, to avoid possible differences in translesion synthesis and DNA repair proficiencies between organisms.

Mutational processes reflect the cumulative effect of 3 steps: (i). DNA damage (for cisplatin, adduct formation), (ii) DNA repair (for cisplatin, NER), which may or may not correct the damage, and (iii) if DNA repair fails, translesion synthesis across the damaged base or bases may replicate the DNA correctly or incorrectly, in the latter instance creating a mutation.

In this study, while known patterns of adduct formation did not predict the patterns of substitutions (Figure 5), we can nevertheless postulate models that explain the observed mutations by combining our knowledge of adduct formation and models of how DNA replication and translesion synthesis might behave (II and III).

First, despite high proportions of DNSs relative to SNSs, SNSs still greatly outnumbered the DNSs (Figure 5A). We postulate that these SNSs are formed by correct translesion synthesis opposite one of the purines of the purine-purine intrastrand crosslinks, and misincorporation occurring opposite the other, as has been shown for UV-induced intrastrand crosslinks (McCulloch et al. 2004). This is supported by the high number of SNSs at potential intrastrand crosslink sites: 85% of the 30,153 SNSs are at GpG, GpA or ApG sites (Supplementary Figure 23).

Second, the relative abundance of the different types of DNSs did not correspond to the reported ratios of intrastrand and interstrand adducts at their respective dinucleotides

(compare the right pie-charts of Figures 5A,B, with graphical representations of the most prominent adducts in Figure 5C). For example, crosslinks at ApG are half as abundant than at GpG, but DNSs from AG:CT were 1.5-times more common than from GG:CC. This suggests that repair of GpG crosslinks is more efficient, or that translesion synthesis past these adducts is less error-prone. As another example, the 2-fold higher number of DNSs from AG:CT than GA:TC does not correspond to the relative frequencies of ApG and GpA adducts (Murray et al. 1992; Mantri et al. 2007). Furthermore, 28.2% of DNSs were in potential interstrand crosslink sites, while these represent <5% of cisplatin-adducts (Jamieson and Lippard 1999; Enoiu et al. 2012). However, the higher proportion of DNS putatively due to interstrand crosslinks is consistent with interstrand crosslinks being more damaging than intrastrand crosslinks (Andreassen and Ren 2009; Hashimoto et al. 2016; Roy and Scharer 2016).

In this study, combined SNS and DNS information was crucial for high-confidence detection of cisplatin mutagenesis in human tumors. SNS analysis alone would have identified 3 false-positives and missed RK140, and DNSs analysis alone would have identified 5 likely false positives among the ESADs. Ideally, the field of mutational signature analysis will move towards a standard of integrated SNS and DNS analysis. To enable this, a comprehensive catalogue of DNS signatures similar to that of SNS signatures (Wellcome Trust Sanger Institute 2016) would be required.

Materials & Methods

Cell line exposure

MCF-10A cells were obtained from the ATCC. Culturing was performed in DMEM/F12 medium supplemented with 10%,FBS, 10 ng/mL insulin, 20 ng/mL EGF, 0.5 µg/mL hydrocortisone, 50 ng/µL penicillin and 50 U/mL streptomycin. For cisplatin exposure, 60,000 cells/well were seeded at day 0 in a 6-wells plate. On day 1 cisplatin was added to final concentrations of 0.5 µM and 1 µM. At day 7, cells were trypsinized and counted, and

per population, 60,000 were seeded in a new 6-wells plate. This process was repeated 8 times. As mutagenesis requires DNA replication, the proliferation rate was monitored. The proliferation rate of 0.5 µM and 1 µM treated cells was 66% and 15% of the untreated population ($p < 0.001$, Supplementary Figure 24). After 4 weeks and 8 weeks, cells were expanded, and single cells were isolated through FACS-sorting directly into a 96-well plate with culture medium. These single cell clones were expanded for DNA isolation and whole-genome sequencing.

Whole-genome sequencing

The MCF-10A cell line was sampled at the start of the cisplatin exposure. DNA isolation was performed using the Wizard Genomic DNA Purification Kit (Promega, Madison, WI, USA) according to the manufacturer's instructions. Paired end sequencing was performed on a HiSeq 10x instrument with 150bp reads at Novogene Co., Ltd. (Beijing, China).

Alignment and variant calling

Read alignment to hs37d5 was done using BWA-MEM, followed by PCR duplicate removal and merging using Sambamba (v0.5.8) (Tarasov et al. 2015). Variant calling was performed using Strelka (v1.014) (Saunders et al. 2012). Variants in dbSNPv132, 1000 genomes (1000 Genomes Project Consortium 2015), segmental duplications, microsatellites and homopolymers, and the GL and decoy sequences were excluded. Additionally, variants were filtered for having at least: 20% variant allele frequency, 25x coverage in both treated and control sample and at least 4 reads supporting the variant. 0.4% and 0.2% of the variants were shared between the clones from the 0.5 µM and 1 µM treated cells. The variant allele frequency distribution is down Supplementary Figure 25.

DNSs were identified as 2 adjacent SNSs. As primary QC we checked that the variant allele frequencies of both SNSs were equal. Secondly, we re-called the genomes using Freebayes, which calls DNSs when the SNSs are in the same reads (Garrison 2012). Out of the 1,123 DNSs extracted from the Strelka calls, 1,093 were also called by Freebayes. Lastly, we checked the DNSs in IGV. All DNSs identified from the Strelka analysis were in the same DNA molecule. Focusing specifically on those DNSs that were not called by Freebayes, 10 were not called as DNSs by freebayes as they were close to a germline variant, and Freebayes called these as tri- or tetranucleotide substitutions. Beyond this, 12 putative DNSs were part of complex mutations that were not in fact DNSs. Of the remaining 8 putative DNSs not called by Freebayes, 5 were likely false-positives, as most were only present in one sequence read direction, in regions with low mapping quality, or located near the end of sequencing reads. Overall, we estimated the initial false-discovery rate of DNSs to be ~1.5% (17/1123) but after Freebayes and IGV inspection we estimate that the false-discovery rate is close to zero.

Statistical analysis of enrichment of mutations in pentanucleotide context

To statistically test for enrichment or depletion of SNSs in each pentanucleotide context we used a binomial test against the null hypothesis that the proportion of a given pentanucleotide that contained a given SNS was the same as the proportion of all pentanucleotides with that SNS. We take as an example A:T>C:G SNS at the center of pentanucleotide CCACC:GGTGG. There were a total of 5,639 T>A mutations in the sequenced portions of the genome, of which 10 were in a CCACC:GGTGG pentanucleotide. In total there were 1,491,086,541 pentanucleotide sites centered on A:T in the sequenced regions of the genome, of which 6,784,989 were CCACC:GGTGG. We then used the R function call binom.test(x = 10, n=5,639, p = (6,784,989 / 1,491,086,541)), which yielded $p < 7.12 \times 10^{-4}$. The alternative hypothesis was that the proportions were not equal.

Analysis of association between cisplatin mutations and genomic features

As histone ChIP-seq data for MCF-10A was not available, we obtained processed ChIP-seq datasets for normal human mammary epithelial cells (HMEC) for H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me3, CTCF and EZH2 (www.ncbi.nlm.nih.gov/geo/, accession GSE29611). We obtained MCF-10A expression data from (www.ncbi.nlm.nih.gov/geo/, accession GSM1100206).

Sources of publicly available sequencing data

This study used whole genome sequencing data from 264 HCCs from Japan (Fujimoto et al. 2016) and 78 from Hong Kong (Kan et al. 2013) and 140 ESADs (Noorani et al. 2017). Additionally, we used whole genome sequencing data of 24 lung adenocarcinomas (Imielinski et al. 2012) and 112 melanomas. For the HCCs, ESADs and melanomas, simple somatic mutation data was downloaded from the ICGC data portal (https://dcc.icgc.org/, release 18, March, 2015). The 78 Hong Kong HCCs were re-analyzed as described previously (Huang et al. 2017).

Analysis of the SNS cisplatin signature exposure in tumors

We used the mSigAct signature.presence.test function to assess possible presence of the experimental cisplatin signature in the publicly available mutational spectra of HCCs and ESADs listed above, as specified in Supplementary Code 1. We defined the "SNS experimental cisplatin signature" as the sum of the SNS spectra of all MCF-10A cisplatin clones divided by the count of all SNSs in the clones. The software is available from the URL https://zenodo.org/record/843773#.WZQQE1EjHRZ as the following doi: 0.5281/zenodo.843773.

NMF on DNS spectra

To assess the effect of cisplatin on primary tumors based on DNSs, we developed a customized semi-supervised NMF (ssNMF) method that incorporated the method from (Schmidt 2007) into the NMF code from (Alexandrov et al. 2013b); Supplementary Code 2 provides the patch file. We use a customary notation for NMF, $V \approx WH$, in which V is the matrix of observed mutational spectra, W, is the matrix of mutational signatures, and H is the matrix of "exposures". ssNMF treats W, the signature matrix, as composed of two segments: W_f, which specifies the known, fixed signatures, and W_u, which is computed by NMF. ssNMF updates only W_u and H. We ran ssNMF separately on (i) V_{ESAD}, the ESAD spectra plus the MCF-10A spectra and (ii) V_{HCC}, the combined spectra from the HCCs, lung adenocarcinomas, and the MCF-10A treated cells. We ran ssNMF on each of V_{ESAD} and V_{HCC}, asking for 2, 3, 4, 5, 6, 7 and 8 signatures (i.e. number of columns of W_u). In both cases W_f consisted of a fixed signature that was the sum of the DNS spectra of all MCF-10A divided by the total DNS count of all spectra. Using the signature stability and average Frobenius reconstruction error approach described in (Alexandrov et al. 2013b), we chose 3 signatures for both of V_{ESAD} and V_{HCC} (Supplementary Figures 19, 20).

Abbreviations:

DNS: dinucleotide substitution

ESAD: esophageal adenocarcinoma

HCC: hepatocellular carcinoma

NER: nucleotide excision repair

SNS: single nucleotide substitution

TC-NER: Transcription coupled nucleotide excision repair

Data availability

Sequencing reads for the cisplatin exposed MCF-10A clones are available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB21971.

Acknowledgements

We thank Willie Yu for technical assistance.

Financial support

This study was funded by NMRC/CIRG/1422/2015 to SGR.

Conflict of interest statement

The authors declare no conflicts of interest

Author contributions

AB and SGR designed the study. AB performed *in vitro* experiments. AB and MNH carried out electronic analyses. AWTN provided analysis tools and technical support. YK, KC and HN provided clinical information. AB and SGR drafted the manuscript and prepared figures. AB, BTT and SGR edited the manuscript. All authors read and approved the final manuscript.

References

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

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013a. Signatures of mutational processes in human cancer. *Nature* **500**: 415-421.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**: 246-259.

Andreassen PR, Ren K. 2009. Fanconi anemia proteins, DNA interstrand crosslink repair pathways, and cancer therapy. *Curr Cancer Drug Targets* **9**: 101-117.

Baik MH, Friesner RA, Lippard SJ. 2003. Theoretical study of cisplatin binding to purine bases: why does cisplatin prefer guanine over adenine? *J Am Chem Soc* **125**: 14082-14092.

Behmand B, Marignier JL, Mostafavi M, Wagner JR, Hunting DJ, Sanche L. 2015. Radiosensitization of DNA by Cisplatin Adducts Results from an Increase in the Rate Constant for the Reaction with Hydrated Electrons and Formation of Pt(I). *J Phys Chem B* **119**: 9496-9500.

Choi DK, Helenowski I, Hijiya N. 2014. Secondary malignancies in pediatric cancer survivors: perspectives and review of the literature. *Int J Cancer* **135**: 1764-1773.

Conaway JW, Conaway RC. 1999. Transcription elongation and human disease. *Annu Rev Biochem* **68**: 301-319.

Dasari S, Tchounwou PB. 2014. Cisplatin in cancer therapy: molecular mechanisms of action. *Eur J Pharmacol* **740**: 364-378.

Dugbartey GJ, Peppone LJ, de Graaf IA. 2016. An integrative view of cisplatin-induced renal and cardiac toxicities: Molecular mechanisms, current treatment challenges and potential protective measures. *Toxicology* **371**: 58-66.

Eastman A. 1983. Characterization of the adducts produced in DNA by cis-diamminedichloroplatinum(II) and cis-dichloro(ethylenediamine)platinum(II). *Biochemistry* **22**: 3927-3933.

Enoiu M, Jiricny J, Scharer OD. 2012. Repair of cisplatin-induced DNA interstrand crosslinks by a replication-independent pathway involving transcription-coupled repair and translesion synthesis. *Nucleic Acids Res* **40**: 8953-8964.

Fichtinger-Schepman AM, Vendrik CP, van Dijk-Knijnenburg WC, de Jong WH, van der Minnen AC, Claessen AM, van der Velde-Visser SD, de Groot G, Wubs KL, Steerenberg PA et al. 1989. Platinum concentrations and DNA adduct levels in tumors and organs of cisplatin-treated LOU/M rats inoculated with cisplatin-sensitive or -resistant immunoglobulin M immunocytoma. *Cancer Res* **49**: 2862-2867.

Fousteri M, Mullenders LH. 2008. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res* **18**: 73-84.

Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M et al. 2016. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet* **48**: 500-509.

Garrison E, Marth, G. 2012. Haplotype-based variant detection from short-read sequencing. doi:arXiv:1207.3907, arXiv:1207.3907.

Harrington CF, Le Pla RC, Jones GD, Thomas AL, Farmer PB. 2010. Determination of cisplatin 1,2-intrastrand guanine-guanine DNA adducts in human leukocytes by high-performance liquid chromatography coupled to inductively coupled plasma mass spectrometry. *Chem Res Toxicol* **23**: 1313-1321.

Hashimoto S, Anai H, Hanada K. 2016. Mechanisms of interstrand DNA crosslink repair and human disorders. *Genes Environ* **38**: 9.

Hu J, Adar S, Selby CP, Lieb JD, Sancar A. 2015. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev* **29**: 948-960.

Hu J, Lieb JD, Sancar A, Adar S. 2016. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **113**: 11507-11512.

Huang MN, Yu W, Teoh WW, Ardin M, Jusakul A, Ng A, Boot A, Abedi-Ardekani B, Villar S, Myint SS et al. 2017. Genome-Scale Mutational Signatures Of Aflatoxin In Cells, Mice And Human Tumors. *bioRxiv* doi:10.1101/130179.

Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A et al. 2012. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**: 1107-1120.

Jamieson ER, Lippard SJ. 1999. Structure, Recognition, and Processing of Cisplatin-DNA Adducts. *Chem Rev* **99**: 2467-2498.

Kaiser VB, Taylor MS, Semple CA. 2016. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet* **12**: e1006207.

Kan Z, Zheng H, Liu X, Li S, Barber TD, Gong Z, Gao H, Hao K, Willard MD, Xu J et al. 2013. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res* **23**: 1422-1433.

Kushner BH, Cheung NK, Kramer K, Heller G, Jhanwar SC. 1998. Neuroblastoma and treatment-related myelodysplasia/leukemia: the Memorial Sloan-Kettering experience and a literature review. *J Clin Oncol* **16**: 3880-3889.

Mantri Y, Lippard SJ, Baik MH. 2007. Bifunctional binding of cisplatin to DNA: why does cisplatin form 1,2-intrastrand cross-links with ag but not with GA? *J Am Chem Soc* **129**: 5023-5030.

Masters JR, Koberle B. 2003. Curing metastatic cancer: lessons from testicular germ-cell tumours. *Nat Rev Cancer* **3**: 517-525.

McCulloch SD, Kokoska RJ, Masutani C, Iwai S, Hanaoka F, Kunkel TA. 2004. Preferential cis-syn thymine dimer bypass by DNA polymerase eta occurs with biased fidelity. *Nature* **428**: 97-100.

Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, Raine K, Maddison M, Anderson E, Stratton MR et al. 2014. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res* **24**: 1624-1636.

Murray V, Motyka H, England PR, Wickham G, Lee HH, Denny WA, McFadyen WD. 1992. The use of Taq DNA polymerase to determine the sequence specificity of DNA damage caused by cis-diamminedichloroplatinum(II), acridine-tethered platinum(II) diammine complexes or two analogues. *J Biol Chem* **267**: 18805-18809.

Noorani A, Bornschein J, Lynch AG, Secrier M, Achilleos A, Eldridge M, Bower L, Weaver JMJ, Crawte J, Ong CA et al. 2017. A comparative analysis of whole genome sequencing of esophageal adenocarcinoma pre- and post-chemotherapy. *Genome Res* **27**: 902-912.

Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahovicek K, Stamatoyannopoulos JA et al. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**: 360-364.

Ratain MJ, Kaminer LS, Bitran JD, Larson RA, Le Beau MM, Skosey C, Purl S, Hoffman PC, Wade J, Vardiman JW et al. 1987. Acute nonlymphocytic leukemia following etoposide and cisplatin combination chemotherapy for advanced non-small-cell carcinoma of the lung. *Blood* **70**: 1412-1417.

Reardon JT, Vaisman A, Chaney SG, Sancar A. 1999. Efficient nucleotide excision repair of cisplatin, oxaliplatin, and Bis-aceto-ammine-dichloro-cyclohexylamine-platinum(IV) (JM216) platinum intrastrand DNA diadducts. *Cancer Res* **59**: 3968-3971.

Roy U, Scharer OD. 2016. Involvement of translesion synthesis DNA polymerases in DNA interstrand crosslink repair. *DNA Repair (Amst)* **44**: 33-41.

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

Schmidt MN, Larsen, J., Hsiao F.T. 2007. Wind Noise Reduction using Non-Negative Sparse Coding. In *2007 IEEE Workshop on Machine Learning for Signal Processing*, doi:10.1109/MLSP.2007.4414345, pp. 431-436, Thessaloniki.

Seplyarskiy VB, Bazykin GA, Soldatov RA. 2015. Polymerase zeta Activity Is Linked to Replication Timing in Humans: Evidence from Mutational Signatures. *Mol Biol Evol* **32**: 3158-3172.

Szikriszt B, Poti A, Pipek O, Krzystanek M, Kanu N, Molnar J, Ribli D, Szeltner Z, Tusnady GE, Csabai I et al. 2016. A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol* **17**: 99.

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

Travis LB, Fossa SD, Schonfeld SJ, McMaster ML, Lynch CF, Storm H, Hall P, Holowaty E, Andersen A, Pukkala E et al. 2005. Second cancers among 40,576 testicular cancer patients: focus on long-term survivors. *J Natl Cancer Inst* **97**: 1354-1365.

Waseem M, Bhardwaj M, Tabassum H, Raisuddin S, Parvez S. 2015. Cisplatin hepatotoxicity mediated by mitochondrial stress. *Drug Chem Toxicol* **38**: 452-459.

Wellcome Trust Sanger Institute. 2016. COSMIC, Catalog of Somatic Mutations in Cancer - Signatures of Mutational Processes in Human Cancer, http://cancer.sanger.ac.uk/cosmic/signatures.

Zamble DB, Mu D, Reardon JT, Sancar A, Lippard SJ. 1996. Repair of cisplatin--DNA adducts by the mammalian excision nuclease. *Biochemistry* **35**: 10004-10013.

Zen C, Zen Y, Mitry RR, Corbeil D, Karbanova J, O'Grady J, Karani J, Kane P, Heaton N, Portmann BC et al. 2011. Mixed phenotype hepatocellular carcinoma after transarterial chemoembolization and liver transplantation. *Liver Transpl* **17**: 943-954.

Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B et al. 2011. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database (Oxford)* **2011**: bar026.

Figure legends

Figure 1: Cisplatin mutational signature. Trinucleotide-context mutational spectra shown as (**A**) raw counts and (**B**) rate of mutations per million trinucleotides for all MCF-10A clones combined. In (**A**), the number of mutations per SNS type is shown above the corresponding bars. (**C**) Pentanucleotide sequence contexts for all MCF-10A samples combined, normalized by pentanucleotide occurrence in the genome. Note prominence of CC<u>C</u>>CC<u>T</u>, TC<u>C</u>>TC<u>T</u>, and AC<u>T</u>>AC<u>A</u> mutations (See also Supplementary Figure 1).

Figure 2: Associations between cisplatin mutagenesis intensity and genomic features. (**A**) Transcription strand bias is more prominent in highly expressed genes for C>A, C>T and T>A mutations. See also Supplementary Figure 5. (**B**) Transcription strand bias of decreases with increasing distance from the transcription start site. See also Supplementary Figure 6. Mutations were binned per 100,000bp (i.e. the first bars are the numbers of mutations within the first 100,000bp from the TSS, then 100,001 to 200,000bp from the TSS and so on). (**C**): Mutation density in regions with histone modifications, relative to the mutation density of each respective sample (Supplementary Table 1).

Figure 3: Cisplatin induced dinucleotide substitutions (DNSs). (**A**) DNS mutation spectra of all MCF-10A clones combined, displayed as mutations per million dinucleotides (i.e. normalized for dinucleotide abundance in the genome). (**B**) Cisplatin induces higher numbers of DNSs than other mutational processes associated with dinucleotide substitutions such as UV (melanoma) and smoking (lung). (**C**) ±1bp sequence context preferences for the most prominent DNS mutation classes (CC>NN, CT>NN, TC>NN and TG>NN). The total number of DNSs per mutation class is indicated in parentheses. Color intensity is relative to the number of mutations with that sequence context, normalized for tetranucleotide abundance in the genome. The vertical axis is the preceding base, the horizontal axis is the following base. Some prominent enrichments in sequence context are indicated

(GCCT>GNNT, NTCT>NNNT and NTGG>NNNG). The full sequence context preference plots, both raw counts and normalized for tetranucleotide abundance in the genome are shown in Supplementary Figure 8. (**D**) Transcription strand bias of dinucleotide substitutions. Potential intrastrand crosslink sites are shown in blue, potential interstrand crosslink sites are shown in red.

Figure 4: Cisplatin mutational signature in human hepatocellular carcinomas (HCCs). (**A**) Example SNS and (**B**) DNS mutational spectra of a tumor that tested positive for the cisplatin signature in the SNS analysis (HK034). In (**A**) and (**B**), numbers of mutations in each mutation class are indicated. (**C**) DNS cosine similarities between our cisplatin signature and HCCs samples, grouped on whether they were negative (left) or positive (right) for cisplatin mutagenesis in the SNS analysis. Samples that were found positive for cisplatin mutagenesis in the SNS analysis but did not show the cisplatin DNS signature (false-positives) are shown in red. RK140 did not display significant evidence of cisplatin mutagenesis in the SNS analysis, but was later identified in the DNS analysis as cisplatin positive.

Figure 5: Comparison of cisplatin-induced substitutions and reported cisplatin adduct ratios. (**A**) Relative abundance of cisplatin-induced base substitutions in the *in vitro* signature. TNS = trinucleotide substitutions. (**B**) Relative abundance of cisplatin-adducts as extracted from literature. Relative abundances of adducts were averaged from (Fichtinger-Schepman et al. 1989; Jamieson and Lippard 1999; Enoiu et al. 2012), proportions of adenine to guanine mono-adducts and ApG to GpA intrastrand adducts were extracted from (Eastman 1983; Baik et al. 2003). Colors in (**A**) correspond to colors of the adducts they are expected to be caused by (in **B**). (**C**) Schematic representation of adducts (from **B**) related to cisplatin-induced substitutions (in **A**). The borders of the schematic adduct representation correspond to the colors used in the zoomed-in section of the pie-charts in (**A**) and (**B**).

Supplementary information

Supplementary Figure 1: Mutational spectra of cisplatin treated MCF-10A clones as (**A**) counts and (**B**) mutations per million trinucleotides.

Supplementary Figure 2: Pentanucleotide sequence contexts for all cisplatin treated MCF-10A clones (**A-F**), normalized for pentanucleotide occurrence in the genome.

Supplementary Figure 3: Graphical display of statistically significant enrichments and depletions of SNSs in pentanucleotide contexts. Binomial tests were performed as described in Materials and Methods. Results were summarized as either enriched for mutations (blue) or depleted for mutations (red) or no difference from expected (white). Dark green and red indicate sequence contexts that were significant after Bonferroni multiple testing correction (i.e. $p < 0.05/1536$). Light blue and pink indicate sequence contexts that were not significant after correction for multiple testing (i.e. $0.05 > p > 0.05/1536$).

Supplementary Figure 4: Transcription strand bias for each of the 96-channels of the mutation spectrum, both displayed as raw counts (**A**) and mutations per million trinucleotides (**B**). Dark = antisense (transcribed) strand, light = sense (untranscribed) strand.

Supplementary Figure 5: Transcription strand bias as a function of gene expression for each of the cisplatin treated MCF-10A clones. Genes were divided in either low or highly expressed (using the median). Transcription strand bias was plotted for each of the mutation classes. Dark = sense (untranscribed) strand, light = antisense (transcribed) strand.

Supplementary Figure 6: Transcription strand bias as a function of distance to the transcription start site in cisplatin treated MCF-10A clones. Mutations were binned according to their distance to the respective TSS of their respective genes. Bin 1 is 0 to 100,000 bp away from the TSS, bin 2 is 100,001 to 200,000 bp away from the TSS and so forth. Dark = sense (untranscribed) strand, light = antisense (transcribed) strand.

Supplementary Figure 7: DNS mutational spectra of cisplatin treated MCF-10A clones as mutations per million dinucleotides in the genome.

Supplementary Figure 8: ±1bp sequence context preference of DNSs in the cisplatin treated MCF-10A clones. Sequence context preference is displayed both as raw counts, and normalized for tetranucleotide abundance in the genome. The total number of DNSs per mutation class is shown between brackets above each plot. Color intensity is relative to the number of mutations with that sequence context, normalized for tetranucleotide abundance in the genome. The vertical axis is the preceding base; the horizontal axis is the following base.

Supplementary Figure 9: Graphical display of statistically significant enrichments and depletions of DNSs in tetranucleotide contexts. Binomial tests were performed as described in Materials and Methods. Results were summarized as either enriched for mutations (blue) or depleted for mutations (red) or no difference from expected (white). Dark green and red indicate sequence contexts that were significant after Bonferroni multiple testing correction (i.e. $p < 0.05/136$). Light blue and pink indicate sequence contexts that were not significant after correction for multiple testing (i.e. $0.05 > p > 0.05/136$).

Supplementary Figure 10: Example plots of DNS sequence context preference in 3 lung adenocarcinomas and 3 melanomas. The lung adenocarcinomas did not display very strong preference sequence context of DNSs. Conversely, in the melanomas most dinucleotides showed sequence context preference. For example, 52.8% of all CC mutations occurred TCCN context. Similarly, TT mutations preferentially occur in NTTA context (52.1%), and CT mutations showed extremely strong preference for a T either preceding or following the mutated dinucleotide (79.5%). The strongest tetranucleotide context preference was observed for TA mutations, which prefer ATAA context (46.9%), and CG mainly occur in TCGA context (50.4%).

Supplementary Figure 11: Transcription strand bias of DNSs in cisplatin treated MCF-10A clones. The total number of dinucleotides eligible for transcription strand bias analysis is displayed in parentheses.

Supplementary Figure 12: Transcription strand bias of DNSs in 3 lung adenocarcinomas and 3 melanomas. Contrary to the cisplatin DNSs, both the smoking and UV associated DNSs displayed transcription strand bias. The UV associated dinucleotides showed a decrease of CC>TT mutations on the transcribed strand, as CC crosslinks induced by UV are repaired by TC-NER. Similarly, smoking associated dinucleotides showed transcription strand bias with a decrease of CC>AA mutations on the untranscribed strand. This fits the prior knowledge that smoking causes GG intrastrand crosslinks, which are repaired by TC-NER if they are located on the transcribed strand. As we display the DNSs as CC>AA, the strand bias is reversed.

Supplementary Figure 13: Comparison of signature W6 with the cisplatin mutational signature.

Supplementary Figure 14: SNS mutation spectra for HCCs that were identified to be positive for the cisplatin mutational signature in the SNS analysis.

Supplementary Figure 15: DNS mutation spectra for HCCs that were identified to be positive for the cisplatin mutational signature in the SNS analysis, displayed as mutations per million dinucleotides.

Supplementary Figure 16: Clustering of patients according to the percentage of SNSs involved in DNSs and the cosine similarity of their DNS spectrum with that of the experimental cisplatin DNS signature. Samples that were identified to be positive for the cisplatin mutational signature in the SNS analysis are displayed in red. **A**: HCCs, **B**: ESADs. For the ESADs, samples with known prior exposure to platinum-based chemotherapeutics are shown as circles; samples without prior platinum-based treatment are shown as crosses.

Supplementary Figure 17: SNS mutation spectra for ESADs that were identified to be positive for the cisplatin mutational signature in the SNS analysis.

Supplementary Figure 18: DNS mutation spectra for ESADs that were identified to be positive for the cisplatin mutational signature in the SNS analysis, displayed as mutations per million dinucleotides.

Supplementary Figure 19: NMF analysis of DNS spectra of all HCCs with at least 25 DNSs.

Supplementary Figure 20: NMF analysis of DNS spectra of all ESADs with at least 25 DNSs.

Supplementary Figure 21: DNS transcription strand bias for tumors positive for cisplatin mutagenesis. The total number of DNSs eligible for transcription strand bias analysis is displayed in parentheses.

Supplementary Figure 22: Comparison of cisplatin mutational signature with previously published cisplatin signatures. *C. elegans* (all worms combined) is the combined mutation spectrum of all mutations of worms treated with cisplatin, regardless of DNA repair deficiency (total 681 mutations). *C. elegans* (N2-worms only) is the mutation spectrum of the mutations of N2 (DNA repair proficient) worms treated with cisplatin (total 51 mutations). *G. gallus* (DT40) is the mutation spectrum of the 3 DT40 clones sequenced (total 2436 mutations). *H. sapiens* (MCF-10A) is the mutation spectrum of all MCF-10A clones combined (30,153 mutations). Spectra are plotted as mutations per million trinucleotides in the respective genomes.

Supplementary Figure 23: Schematic display of which peaks from the SNS mutation spectra are inside GpG, GpA or ApG dinucleotides, or in any of these 3 (Total). All peaks observed in the cisplatin mutational signature occur in trinucleotides containing one of these 3 potential cisplatin intrastrand crosslink sites.

Supplementary Figure 24: Proliferation rate during cisplatin exposure.

Supplementary Figure 25: Variant allele frequency distribution of the cisplatin mutations. The tail towards the lower end of the histogram represents variants inside regions of the MCF-10A genome displaying aneuploidy.

Supplementary Table 1: Sequencing statistics and variants detected in each of the cisplatin treated MCF-10A clones.

Supplementary Table 2: Cisplatin SNS signature; weighted average of the 6 MCF-10A clones.

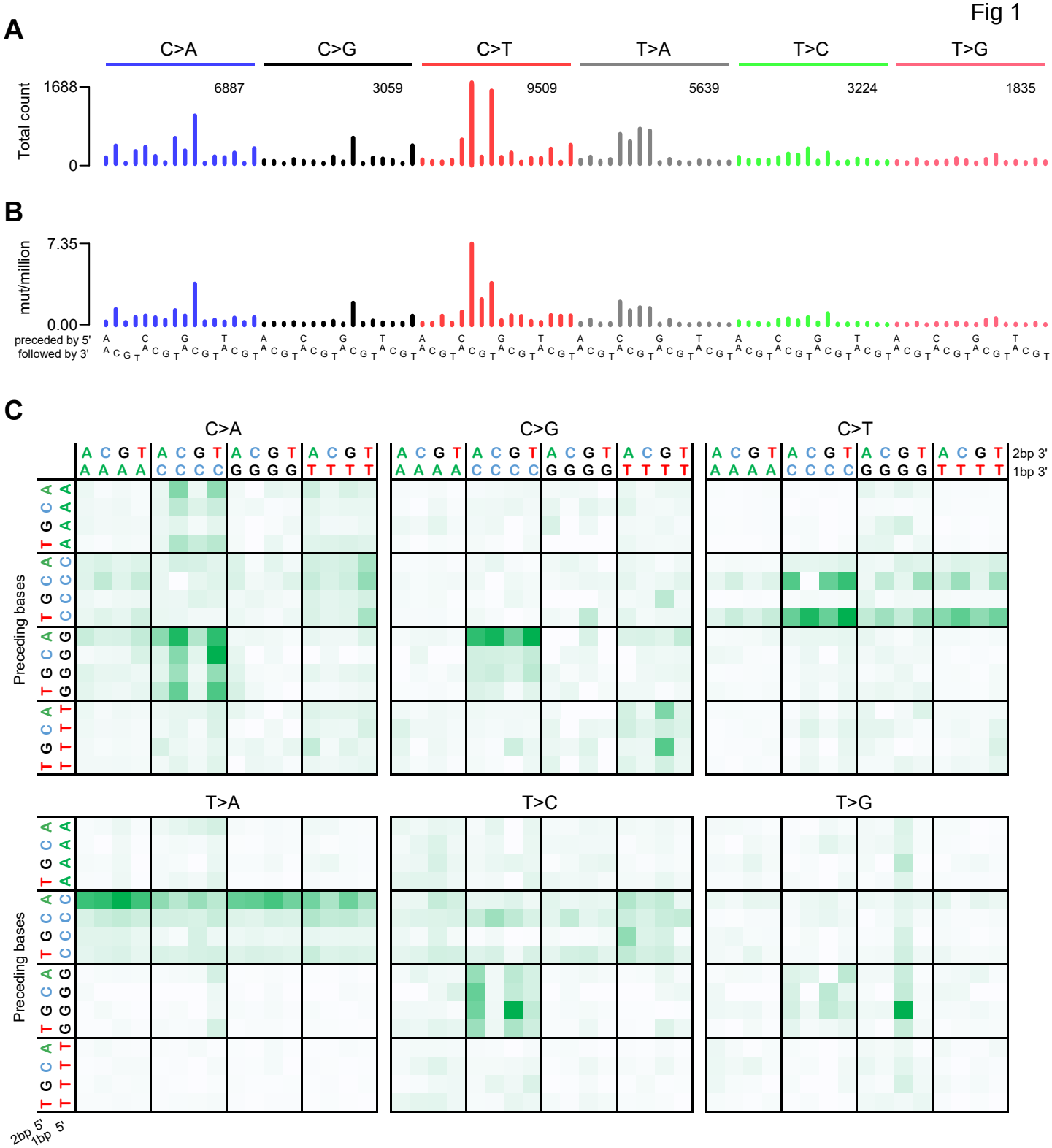
Supplementary Table 3: Cisplatin DNS signature; weighted average of the 6 MCF-10A clones.

Supplementary Table 4: Variant allele frequencies of SNSs and DNSs in cisplatin positive HCCs from Japan.

Supplementary Code 1 Detection of the cisplatin signature in HCC and ESAD SNS mutation spectra with mSigAct.

Supplementary Code 2: Patch-file to update the NMF-code (Alexandrov et al. 2013b) to allow semi-supervised NMF.

Fig 1



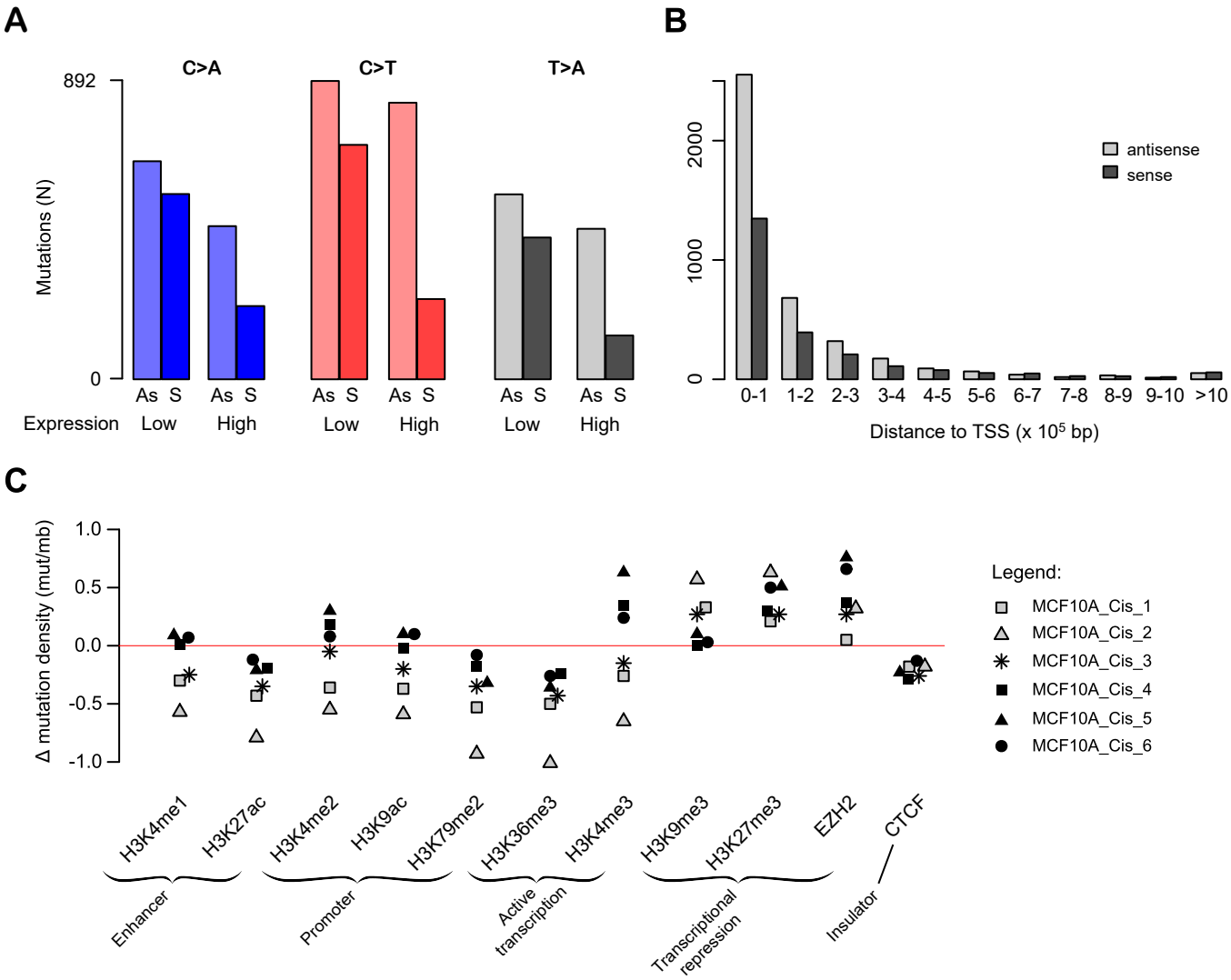


Fig 3

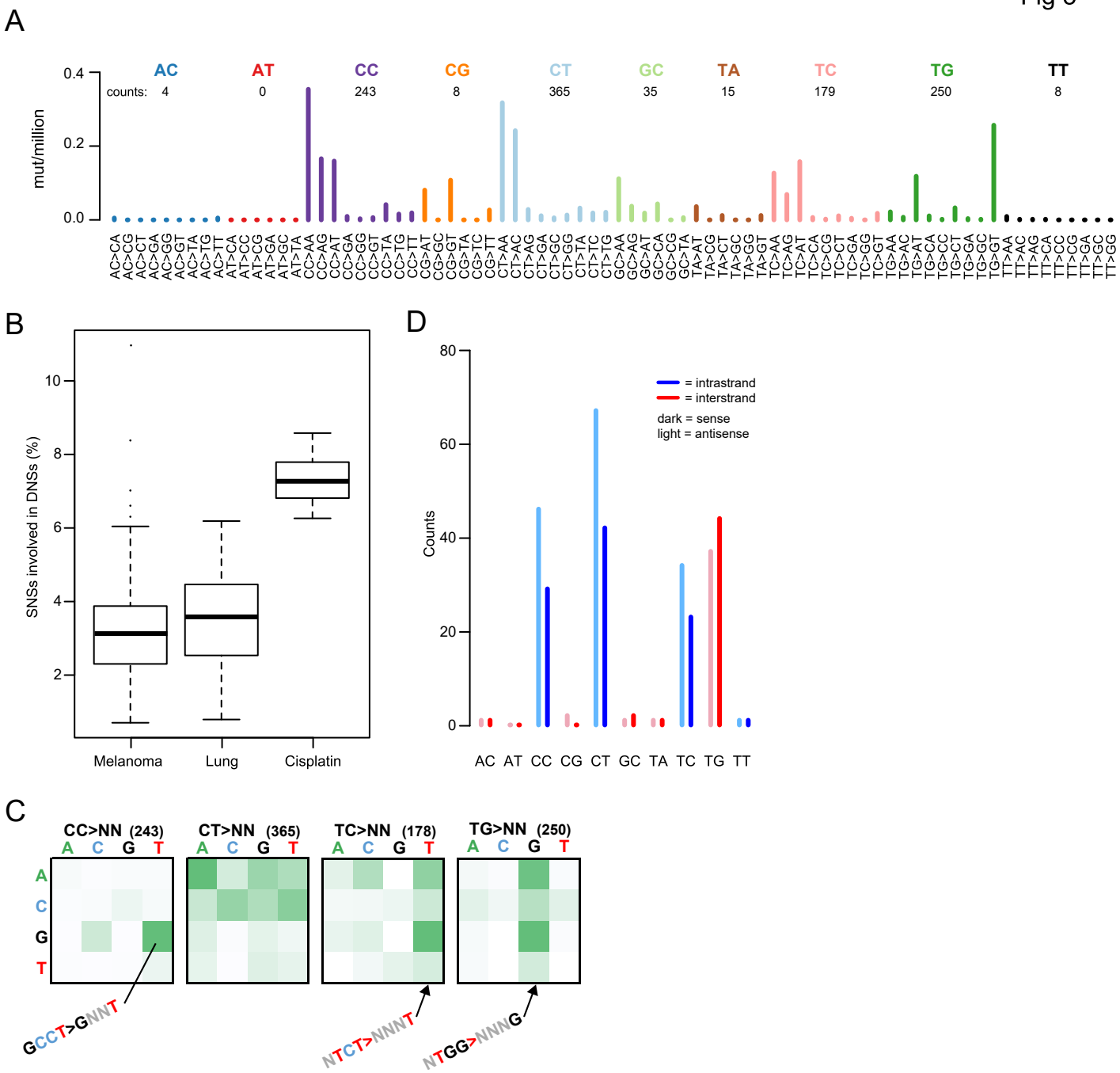


Fig 4

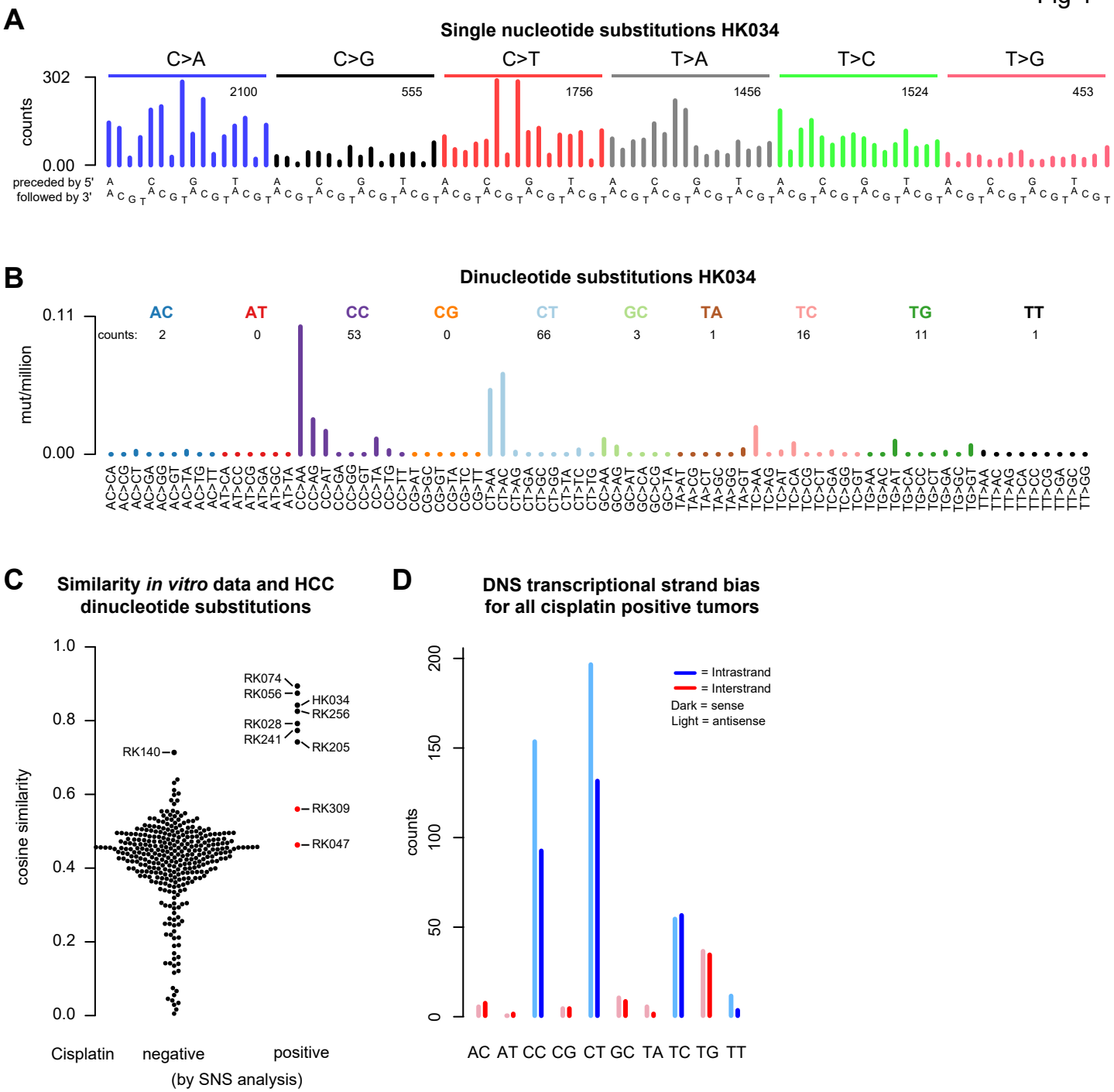


Fig 5

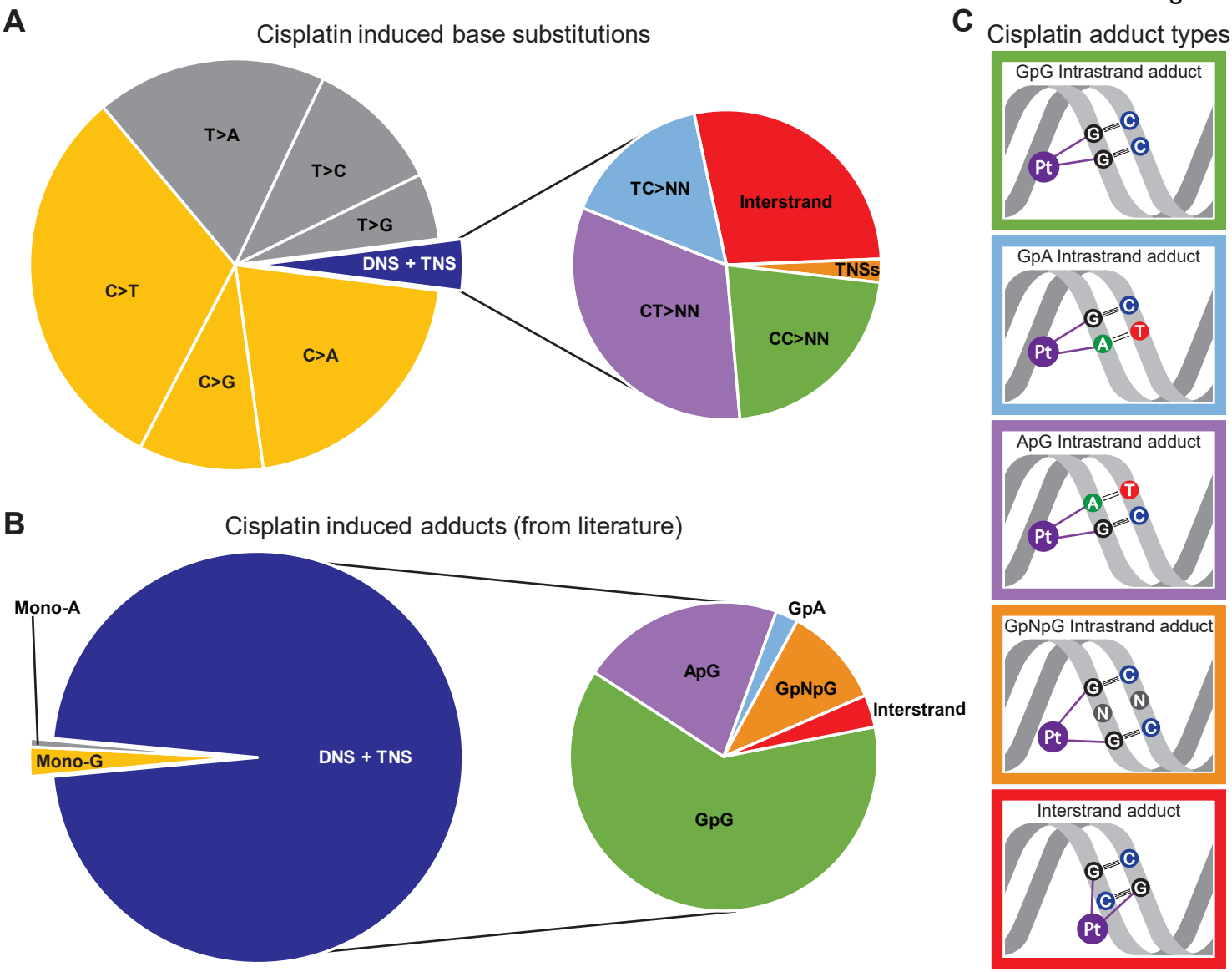


Table 1: HCCs and ESADs with cisplatin-associated mutagenesis

Tumor ID	Cancer type	Total SNSs	Cisplatin SNSs	mSigAct P	Bonferroni corrected P	DNSs	DNS cosine similarity to experimental signature	DNSs due to cisplatin*	Conclusion based on SNS and DNS analysis	Patient history**	Reference
HK034	HCC	7844	2274	1.3E-11	4.6E-09	153	0.841	130	Cisplatin positive	NA	Kan et al., 2013
RK028	HCC	20792	6974	1.3E-19	4.5E-17	642	0.791	533	Cisplatin positive	cisplatin DEB-TACE	Fujimoto et al., 2016
RK047	HCC	8345	1096	8.7E-05	3.0E-02	73	0.462	8	Negative	No neo-adjuvant chemotherapy	Fujimoto et al., 2016
RK056	HCC	17085	6686	7.3E-25	2.5E-22	479	0.874	426	Cisplatin positive	cisplatin DEB-TACE	Fujimoto et al., 2016
RK074	HCC	22406	6986	1.5E-19	5.2E-17	476	0.893	415	Cisplatin positive	cisplatin DEB-TACE	Fujimoto et al., 2016
RK140	HCC	10132	986	3.2E-04	1.1E-01	125	0.713	73	Cisplatin positive	cisplatin DEB-TACE, 4 years, 1 year and 6 months prior	Fujimoto et al., 2016
RK205	HCC	10406	1668	1.4E-06	4.9E-04	158	0.741	150	Cisplatin positive	cisplatin DEB-TACE, prior history of HCC, reseceted 27 months ago	Fujimoto et al., 2016
RK241	HCC	10610	3016	7.7E-15	2.6E-12	235	0.772	177	Cisplatin positive	cisplatin DEB-TACE, prior history of colorectal cancer	Fujimoto et al., 2016
RK256	HCC	11240	2319	4.5E-09	1.5E-06	167	0.825	142	Cisplatin positive	cisplatin DEB-TACE, prior history of HCC, reseceted 37 and 18 months ago	Fujimoto et al., 2016
RK309	HCC	4785	1311	1.0E-05	3.4E-03	12	0.489	--	Negative	No neo-adjuvant chemotherapy	Fujimoto et al., 2016
SA594320	ESAD	24423	4692	2.8E-17	4.2E-15	313	0.851	210	Cisplatin positive	Cisplatin treated	Noorani et al., 2017
SA594557	ESAD	7648	637	4.4E-05	6.6E-03	63	0.809	33	Cisplatin positive	Cisplatin treated	Noorani et al., 2017
SA594775	ESAD	16323	1608	2.4E-05	3.6E-03	124	0.678	41	Cisplatin negative	Cisplatin treated	Noorani et al., 2017

* DNS assignment by ssNMF. The cisplatin DNS signature was given as input, and 2 other DNS signatures were requested. Reported here is the number of DNSs assigned to the cisplatin DNS signature. Tumors with <25 DNSs were excluded from this analysis

** NA denotes data not available