bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis

Anna R Poetsch[1,2,3], Simon J Boulton[1+], Nicholas M Luscombe[1,2,3+]

[1] The Francis Crick Institute, 1 Midland Road, London, NW1 1AT, UK
[2] Okinawa Institute of Science & Technology Graduate University, Okinawa 904-0495, Japan
[3] UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK

[+]Corresponding authors: nicholas.luscombe@crick.ac.uk, simon.boulton@crick.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

DNA is subject to constant chemical modification and damage, which eventually results in variable mutation rates throughout the genome. Although detailed molecular mechanisms of DNA damage and repair are well-understood, damage impact and execution of repair across a genome remains poorly defined. To bridge the gap between our understanding of DNA repair and mutation distributions we developed a novel method, AP-seq, capable of mapping apurinic sites and 8oxoguanidine bases at high resolution on a genome-wide scale. We directly demonstrate that the accumulation rate of oxidative damage varies widely across the genome, with hot spots acquiring many times more damage than cold spots. Unlike SNVs in cancers, damage burden correlates with marks for open chromatin notably H3K9ac and H3K4me2. Oxidative damage is also highly enriched in transposable elements and other repetitive sequences. In contrast, we observe decreased damage at promoters, exons and termination sites, but not introns, in a seemingly transcription-independent manner. Leveraging cancer genomic data, we also find locally reduced SNV rates in promoters, genes and other functional elements. Taken together, our study reveals that oxidative DNA damage accumulation and repair differ strongly across the genome, but culminate in a previously unappreciated mechanism that safe-guards the regulatory sequences and the coding regions of genes from mutations.

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Introduction

The integrity of DNA is constantly challenged by damaging agents and chemical modifications. Base oxidation is a frequent insult that can arise from endogenous metabolic processes as well as from exogenous sources such as ionizing radiation. At background levels, a human cell is estimated to undergo 100 to 500 such modifications per day, most commonly resulting in 8-oxo-7,8-dehydroguanine (8OxoG) [1]. Left unrepaired, 8OxoG can compromise transcription, DNA replication, and telomere maintenance. Moreover, damaged sites provide direct and indirect routes to C-to-A mutagenesis.

Oxidative damage is reversed in a two-step process through the base excision repair (BER) pathway. The damaged base is first recognized and excised by 8Oxo-G-glycohydrolase 1 (OGG1), leaving an apurinic site (AP-site). Glycohydrolysis is highly efficient, with a half-life of 11 min [2]. In a second step, the AP-site is removed through backbone incision by AP-lyase (APEX1) and subsequently replaced with an undamaged nucleotide.

Though originally controversial [3,4], there is now broad acceptance that mutation rates vary across different regions within genomes. Background mutation rates in *Escherichia coli* genomes were shown to vary non-randomly between genes by an order of magnitude, with highly expressed genes displaying lower mutation rates [5]. In cancer genomes, single nucleotide variants (SNVs) tend to accumulate preferentially in heterochromatin [6,7]. More recently, it was reported that SNV densities in cancers are lower in regions surrounding transcription-factor-binding, but are elevated at the binding sites themselves and at sites with a high nucleosome occupancy [8–11]. Although these variabilities remain mechanistically unexplained, they likely arise through a combination of regional differences in damage sensitivity and the accessibility to the DNA repair machinery [12]. However, since mutations represent the endpoint of mutagenesis, it is impossible to tease apart the contributions from damage and repair using re-sequencing data alone.

To further our understanding of the molecular mechanisms underlying local heterogeneity of mutation rates, direct measurement of specific DNA-damage types is required at high resolution and on a genomic scale. Dissecting these mechanisms will help understand the local sensitivities of the genome and why certain regions appear to be protected.

2. A genome-wide map of oxidative damage

To measure oxidative damage across the genome, we developed an approach that detects AP-sites using a biotin-labelled aldehyde-reactive probe [13]; (Figure 1A and Supplementary Figure S1). After fragmentation, biotin-tagged DNA with the original damage sites was pulled down using streptavidin magnetic beads and prepared for high-throughput sequencing. The signal was quantified as the Relative Enrichment of the pull-down over the input DNA, with positive values indicating regions of damage accumulation.

Figure 1B provides the first high-resolution, genome-wide view of oxidative DNA damage. It immediately highlights the extreme variability in the density of AP-sites across the human genome after X-ray treatment in HepG2 cells: though the genome-wide mean Relative Enrichment is 0.1, local enrichments vary from less than -0.6 to more than 3.0. Hot and cold spots are found across all chromosomes and do not appear to follow a particular distribution pattern: whereas the entire chromosome 19 is enriched for damage, on chromosome 7 we observe pericentromeric hot spots. Figure 1C shows a more detailed profile of chromosome 16, including distributions for treated and untreated samples. The profiles of the X-ray treated samples indicate

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

an overall treatment-dependent accumulation of damage; however local distribution patterns are maintained, suggesting that hot spots gain the most additional damage. In Figure 1D, we zoom further into an 8kb region upstream of the MALT1 gene. Here, differences between the treated and untreated samples become apparent, with damage particularly accumulating on *Alu* transposable elements after X-ray exposure. These plots exemplify how variable damage enrichments can be, with hot and cold spots occurring from ~50-500bp to kilo base resolution.

To assess whether the distribution of AP-sites is representative of 8OxoG we applied recombinant OGG1 *in vitro* to the extracted DNA (Figure 1A); this additional treatment excises any remaining 8OxoG after DNA extraction and results in a set of secondary AP-sites. Any difference in enrichment between the original and OGG1-enriched samples indicates the presence of unprocessed 8OxoG *in vivo*. The control and X-ray treated samples are highly correlated overall (Figure 1E). Moreover, the OGG1-enriched samples are very similar to the unenriched, indicating that the AP-site pull-down provides a good measure of the *in vivo* 8OxoG distribution.

3. Genomic features shape distribution of oxidative damage

3.1 Damage accumulates preferentially in euchromatin but not heterochromatin

To identify potential causes of variation across the genome we compiled for the same HepG2 cell line a set of 18 genomic and epigenomic features associated with DNA damage, repair, and patterns of mutagenesis (Figure 2A). Previous studies reported that SNV densities in cancer genomes were positively correlated with heterochromatic markers (eg, H3K9me3) and negatively correlated with euchromatic ones (eg, H3K4me3, H3K9ac) [7]. Here, oxidative DNA damage displays the opposite trend, correlating with open chromatin and anticorrelating with closed chromatin. At first glance, it is surprising that SNVs and DNA damage should show opposing trends; open chromatin is probably more accessible to damage-causing agents, but is also more accessible for repair and it is more accurately replicated. The balance of these three mechanisms leads ultimately to decreased mutations in euchromatin despite the increased damage levels. Observations are upheld at higher resolutions for many features; for instance, the Spearman's correlation with H3K9me3 is -0.48 at 1Mb resolution, -0.34 at 100kb, -0.3 at 10kb, and -0.14 at 1kb resolution. For other features, these correlations break down; DNase I hypersensitivity correlates at low resolution (Spearman's r = 0.5 and 0.3 at 1Mb and 100kb respectively), but the relationship is lost at higher resolutions (r = 0.06 and -0.06 at 10kb and 1kb respectively). This suggests that more detailed genomic features and functional elements also play a role in shaping the local damage distributions.

3.2 Damage enrichment is GC-content dependent

As oxidative damage predominantly occurs on guanines, base content is expected to be a prime determinant of genome-wide distribution. The heatmap in Figure 2A shows that this is true in general, with average damage levels in 100kb windows correlating with GC content (Spearman's r = 0.37). However closer examination shows a more complex relationship: in Figure 2B, we plot average damage levels in 1kb windows against their GC content. While there is a clear increase in damage as GC content rises from 25% to 47%, this relation breaks down above 47% GC and damage levels drop sharply. This indicates that damage in regions of high GC content cannot be explained by base composition alone.

3.3 Gene promoters and bodies show selective protection from damage

Next, we interrogated damage distributions over coding regions by compiling a metaprofile for 23,056 protein-coding genes (Figure 2C). The analysis reveals rigid compartmentalisation, with

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

damage levels varying substantially between elements. Damage is dramatically reduced within genes compared with flanking intergenic regions (Relative Enrichment = 3.8), most prominently at the transcriptional start (Relative Enrichment = -8.0), 5'-UTRs (Relative Enrichment = -6.9), exons (Relative Enrichment = -6.1) and termination sites (Relative Enrichment = -5.8). In stark contrast, introns show high damage (Relative Enrichment = 0.4), though still below intergenic levels. Intron-exon junctions are accompanied by steep transitions in damage indicating the sharp distinction between coding, regulatory and non-coding regions (Relative Enrichment changes from -6.0 to -0.5 within 300bp around the 3'-exon junction). Damage levels rapidly rise again downstream of termination sites towards intergenic regions (Relative Enrichment shifts from -4.3 to 2.0 within 500bp).

Promoters and transcription start sites have the lowest damage levels of any functional element in the genome (average Relative Enrichment = -8.0 compared with intergenic average of 3.8). Unlike SNVs and other damage types, which decrease with rising expression levels, we do not detect an association between oxidative damage and expression (Figure 2D). There is a substantial GC content effect (Figure 2E); but in contrast to expectations from base composition alone, damage levels fall as GC content rises (Relative Enrichment = 1.1 at 45% GC and Relative Enrichment = -12.6 at > 64% GC).

3.4 Retrotransposons accumulate large amounts of damage

Retrotransposons provide a fascinating contrast to coding genes: Long Interspersed Nuclear Elements (*LINEs*) possess similar structures to genes with an RNA Pol II-dependent promoter and two open reading frames (ORFs), whereas Short Interspersed Nuclear Elements (*SINEs*) resemble exons in their nucleotide compositions and presence of cryptic splice sites. Unlike coding genes though, *LINEs* and *SINEs* accumulate staggeringly high levels of damage. *Alu* elements, the largest family among *SINEs*, show by far the highest damage levels of any annotated genomic feature: a metaprofile of >800,000 *Alu* elements in Figure 2F peaks at an average Relative Enrichment of 59, much higher than the genomic average of 0.1. The damage profile rises and falls within 500bp. Similarly, a metaprofile of >2,500 *LINE* elements in Figure 2G displays heterogeneous, but high levels of damage accumulation: like coding genes, there is reduced damage at the promoter (average minimum Relative Enrichment = -5.2), but in contrast to genes there is a gradual increase in damage from the 5' to 3'end, peaking at a Relative Enrichment of 26.9 near the end of the second ORF.

Retrotransposons, though usually silenced through epigenetic mechanisms, can be activated by DNA damage in general [14] and ionizing radiation in particular [15]. How DNA damage or repair affects such silencing mechanisms is currently unknown. One might speculate that DNA damage at these positions could lead to unwanted *LINE* transcription, for instance through repair-associated opening of the chromatin. These distinct and unique damage patterns of both protection and strong accumulation of damage within one functional element suggest the existence of targeted repair or protective mechanisms that are unique to retrotransposons.

3.5 Transcription factor-binding sites, G-quadruplexes and other regulatory sites

Finally, we examine the most detailed genomic features previously implicated in mutation rate changes. In Figure 3A-C we assess the impact of DNA-binding proteins: there is a universal U-shaped depletion of damage levels +/-500bp of the binding-site regardless of the protein involved, suggesting that the act of DNA-binding itself is a major factor. We find the greatest reduction in damage for actively used binding-sites that overlap with DNase-hypersensitive regions in the HepG2 cell line. However, a smaller reduction is also present for inactive sites,

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

indicating that the effects go beyond simple DNA-binding. It is notable that the binding-site effects override the contribution of the GC content to damage levels.

GC-rich features are particularly interesting because of the complex relationship between GC content, protein-binding and damage levels. CpG islands are frequently located in promoters and display reduced damage (Figure 3D). Most surprising is the dramatic reduction in damage in CpG islands outside promoters and DNase-hypersensitive regions, indicating that the localisation in promoters is not the main reason for damage reduction; in fact, it is possible that the reduction in damage for high-GC promoters might be explained by the presence of CpG islands and not vice versa.

Another feature of GC-rich sequences are G4-quadruplexes (G4 structures) formed by repeated oligo-G stretches. G4-quadruplexes are prevalent in promoters and telomeric regions, where they impact telomere replication and maintenance. A meta-profile for >350,000 predicted G4 structures displays a dramatic asymmetric reduction in damage, in which the minimum occurs just downstream of the G4-quadruplex centre (Figure 3E). G4 structures are also one of the few features in which we detect a difference between the 8OxoG and AP-site distributions with a particular enrichment at the centre of G4 structures. This finding is particularly relevant for telomeric repeats (Figure 3F), where oxidized bases impact on telomerase activity and telomere length maintenance[16]. These repeats are thought to form G4 structures, but in contrast to quadruplexes in general, telomeres present with a mild increase in AP-sites after X-ray treatment (average Relative Enrichment=1.1) and stronger enrichment of 8OxoG (average Relative Enrichment=2.3).

Micro-satellites are 3-6bp sequences that are typically consecutively repeated 5-50 times. Whereas GC-rich micro-satellite repeats show generally reduced damage, most simple repeats show an accumulation of damage; this is depicted for individual repeat sites at the *LINC00955* locus (Figures 3G). The motifs $(GAA)_n$, $(GGAA)_n$, and $(GAAA)_n$ accumulate the largest amounts of damage (Figure 3H). Interestingly, specific sequences display preferential damage enrichment in the OGG1-enriched samples, such as $(CCCA)_n$ and $(ATGGTG)_n$. Micro-satellites are capable of forming non-B-DNA structures, such as hairpins; we suggest that changes in the DNA's local structural properties impairs 8OxoG-processing on these genomic features with possible regulatory functionality.

4. SNVs in oxidative damage-dependent cancers reflect underlying damage profiles

Lastly, we address how the distribution of oxidative DNA damage is reflected in the landscape of SNVs in cancer genomic data. We compiled a dataset of 9.4 million C-to-A transversions, the major mutation-type caused by oxidative damage, from 2,702 cancer genomes [17]. Of these, 8 hypermutated tumours are defective in polymerase epsilon (Pol E) activity (total 3.4 million C-to-A SNVs). Under normal conditions, Pol E-proofreading prevents 8OxoG-A mismatches, but in the absence of this activity, a large proportion of mismatches is thought to result in C-to-A mutations. Thus, the distribution of SNVs in the absence of Pol E-proofreading is expected to follow the underlying oxidative damage pattern, reflecting local differences in damage susceptibility and repair preferences. We also identified 2,401 tumours with increasing proportions of C-to-A SNVs originating from the mutational process associated with the COSMIC Mutational Signature 18, which has been suggested to arise from oxidative damage [19,20].

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In most tumours, about 9% of C-to-A SNVs occur in regions of high GC content (Figure 4A); however, the proportion drops to just 3% among Pol E-defective tumours, in line with the unexpected depletion of oxidative damage in these genomic regions (Figure 2B). Similarly, tumours display decreasing proportions of SNVs with rising amounts of Signature 18 (Figure 4A), following the expected trend for oxidative damage. We also observed that damage is preferentially distributed in euchromatin at 100kb resolution, whereas SNVs tend to accumulate in heterochromatin; unsurprisingly at this resolution, the damage and SNV densities are anticorrelated (Spearman's r =-0.49 and -0.45 for proofreading-defective and control tumours respectively).

We focused on the proof-reading defective and control tumour samples for the high-resolution genomic features, as they contain the largest numbers of SNVs. In protein-coding genes, the SNV distribution for Pol E-defective tumours is remarkably similar to the damage profiles (Figure 4B): decreased rates at the TSS, 5'-UTR, exons, and increased rates in introns. The profile is lost in control tumours: we speculate that bulky adducts or strand breaks – a distinct form of damage – cause the accumulation of SNVs at the promoter. SNVs are also depleted from GC-rich genomic features in Pol E-defective tumours, including CTCF-binding sites, transcription factor binding sites, CpG islands and G4-quadruplexes. The patterns are lost in the controls (Figure 4C). The difference between the two tumour sets indicates that at high resolution, the distribution of distinct damage types dominates the ultimate SNV profiles. However, there is a striking divergence from damage distributions in retrotransposons (Figure 4D); whereas above we observed high levels of damage in *Alus* and *LINEs*, there appears to be increased safe-keeping, leading to lower levels of mutations. This pattern is lost in the control tumours.

5. Discussion

Our results demonstrate the feasibility of measuring oxidative damage across a genome at high resolution and specificity. In addition to the considerable feature-dependent variability in damage rates, we are able to relate them directly to patterns of SNV occurrences in cancer genomes. At the 100kb scale, euchromatin has increased exposure to oxygen radicals but also better accessibility for repair enzymes, leading to high damage levels but fewer SNVs; in heterochromatin with poorer relative access to repair, the trends are reversed. At the 10kb to 200bp resolution, we find reduced damage levels in functional elements such as coding sequences, promoters, and transcription factor binding sites, which correlate with SNV occurrences in cancers. The heterogeneity results from changes in the balance of damage susceptibility and repair rates at different genomic regions.

Locus-specific oxidative damage is distinct from damage types repaired by other pathways such as nucleotide excision repair (NER). For instance, oxidative damage levels are seemingly independent of gene expression, whereas nucleotide excision repair can be coupled to transcription. Moreover, for NER, Sabarinathan and Perera reported UV-dependent mutation hotspots around transcription factor binding sites explained by hindered access of the repair machinery. For oxidative damage, we observe the opposite: protection of the same regions from oxidative damage and its derived mutations. Such hotspots are probably prevented through inaccessibility of the DNA to oxygen radicals, which is not the case for UV light.

Intriguingly, though damage accumulates in *LINEs* and *Alus*, they are protected from mutations in cancer genomes; this suggests a specific mechanism for targeted repair at these features that was not reflected in the damage distribution or may be defective in the HepG2 cell line used here.

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Modulation of DNA damage at these sites would suggest not only effects on mutagenesis, but perhaps even an epigenetic regulatory mechanism through oxidative damage to silence retrotransposons; indeed, an epigenetic function for 8OxoG has been suggested at G4 structures [21]. At these sites and other potential non-B-DNA structures we detected elevated signals in the OGG1-enriched samples indicating the *in vivo* accumulation of 8OxoG; this suggests that 8OxoG-processing is impaired. It is interesting to speculate that these sites may have acquired a regulatory function beyond accumulating mutations.

In conclusion, we have established a robust method to measure oxidative damage in a genome-wide manner. With minor modifications, it will be suitable for detecting any base modification that can be excised with a specific glycohydrolase. Identifying the pathways that lead to selective repair fidelity and protection of functional elements will not only provide insights into basic mutagenesis but will also allow us to identify any regulatory characteristics of 8OxoG and AP-sites as epigenetic marks.

6. References

1. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709-715 (1993).
2. Hamilton, M. L. et al. A reliable assessment of 8-oxo-2-deoxyguanosine levels in nuclear and mitochondrial DNA using the sodium iodide method to isolate DNA. *Nucleic Acids Res* **29**, 2117-2126 (2001).
3. Cairns, J., Overbaugh, J. & Miller, S. The origin of mutants. *Nature* **335**, 142-145 (1988).
4. Lenski, R. E. & Mittler, J. E. The directed mutation controversy and neo-Darwinism. *Science* **259**, 188-194 (1993).
5. Martincorena, I., Seshasayee, A. S. & Luscombe, N. M. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* **485**, 95-98 (2012).
6. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-364 (2015).
7. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504-507 (2012).
8. Katainen, R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-821 (2015).
9. Perera, D. et al. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259-263 (2016).
10. Reijns, M. A. et al. Lagging-strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502-506 (2015).
11. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264-267 (2016).
12. Martincorena, I. & Luscombe, N. M. Non-random mutation: the evolution of targeted hypermutation and hypomutation. *Bioessays* **35**, 123-130 (2013).
13. Kubo, K., Ide, H., Wallace, S. S. & Kow, Y. W. A novel, sensitive, and specific assay for abasic sites, the most commonly produced DNA lesion. *Biochemistry* **31**, 3703-3708 (1992).
14. McClintock, B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36**, 344-355 (1950).
15. Farkash, E. A., Kao, G. D., Horman, S. R. & Prak, E. T. Gamma radiation increases endonuclease-dependent L1 retrotransposition in a cultured cell assay. *Nucleic Acids Res*

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

34, 1196-1204 (2006).
16. Fouquerel, E. et al. Oxidative guanine base damage regulates human telomerase activity. *Nat Struct Mol Biol* **23**, 1092-1100 (2016).
17. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *bioRxiv* (2017).
18. Lujan, S. A. et al. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res* **24**, 1751-1764 (2014).
19. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
20. Pan-cancer Analysis of Whole Genomes consortium -working group 7 (Mutation signatures and processes). *Manuscript in preparation* (2017).
21. Fleming, A. M., Ding, Y. & Burrows, C. J. Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. *Proc Natl Acad Sci U S A* **114**, 2604-2609 (2017).
22. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
23. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
24. Robinson, J. T. et al. Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).
25. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841-1842 (2009).
26. The ENCODE, Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
27. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol* **13**, R77 (2012).
28. Bedrat, A., Lacroix, L. & Mergny, J. L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* **44**, 1746-1759 (2016).
29. Feuerbach, L. et al. TelomereHunter: telomere content estimation and characterization from whole genome sequencing data. *Cold Spring Harbor Labs Journals*, 2016.
30. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).

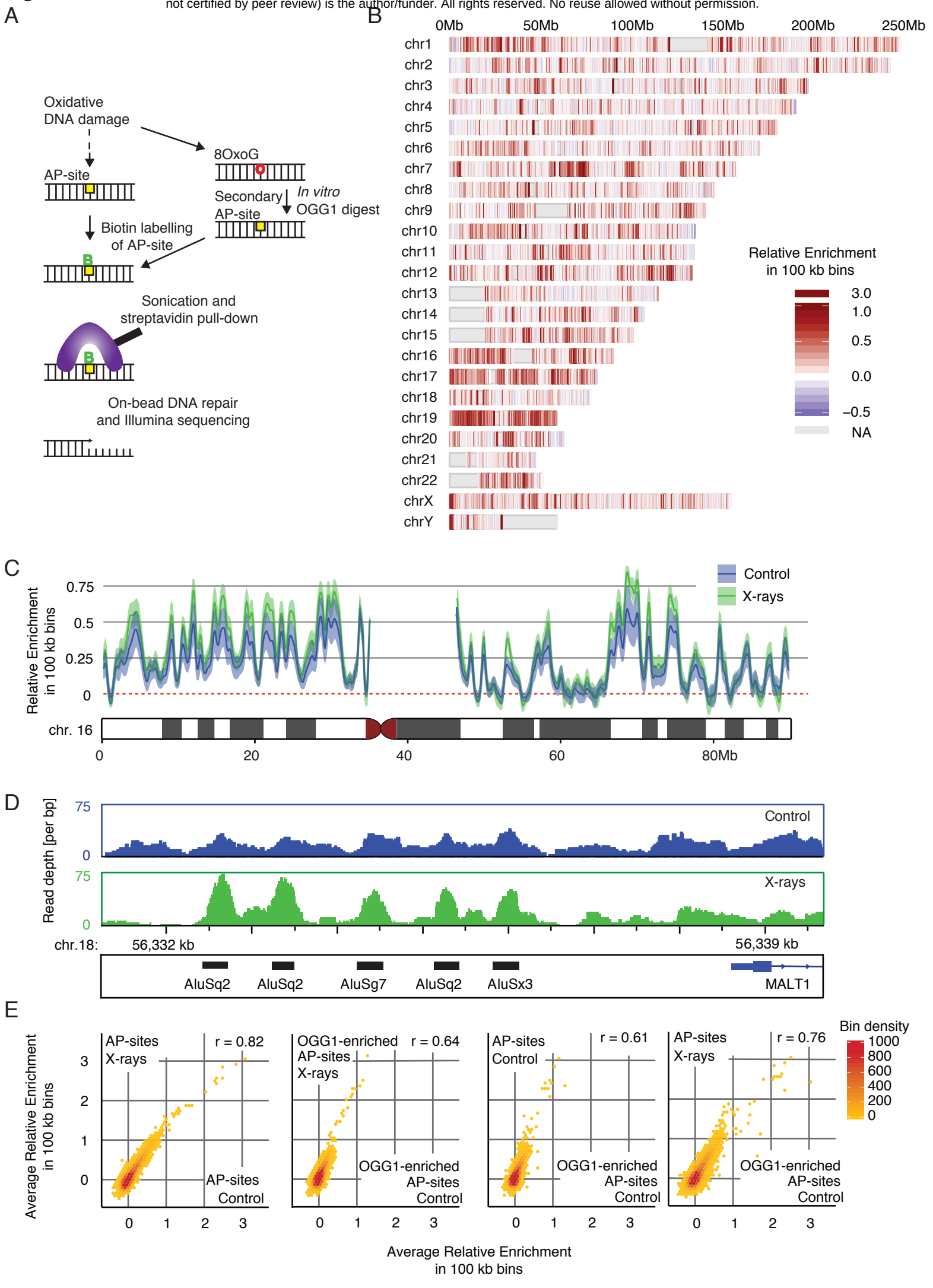
bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7. Figures

7.1 Figure 1. Oxidative damage is heterogeneously distributed at different scales of resolution. **(A)** Schematic of AP-seq, a new protocol to detect apurinic-sites (AP-sites) as a measure of oxidative damage in a genome. 8OxoG is excised by OGG1 in the first, rapid step of base excision repair, leaving an AP-site. DNA containing these sites are biotin-tagged using an aldehyde reactive probe (ARP), fragmented, and pulled-down with streptavidin. The enriched DNA is processed for sequencing and mapped to the reference genome. The damage level across the genome is quantified by assessing the number of mapped reads. To check for unprocessed 8OxoG, we perform an *in vitro* digest of extracted genomic DNA with OGG1 and repeat the AP-site pull-down. (**B**) Genome-wide map of AP-site distribution after X-ray treatment. The colour scale represents the Relative Enrichment of AP-sites in 100kb bins across the human genome, averaged across biological replicates. Damage levels are highly correlated between treatment conditions at 100kb resolution. **(C)** More detailed view of AP-site distribution on Chromosome 16. Plot lines depict the average Relative Enrichment for X-ray treated (green) and untreated (blue) samples. Shaded boundaries show standard error of the mean for the biological replicates. Untreated and X-ray treated samples display very similar damage profiles. **(D)** Genome browser views of damage distributions for untreated and X-ray treated samples across an 8kb region upstream of MALT1. Damage levels are represented by the read depths of the pooled biological replicates. At high resolution, the sharp relative increase in damage at *Alu* elements after X-ray treatment becomes apparent. **(E)** Scatterplots of the correlation in average Relative Enrichments of samples with differing treatment and OGG1-enrichment conditions. Damage levels are highly correlated across all conditions.

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

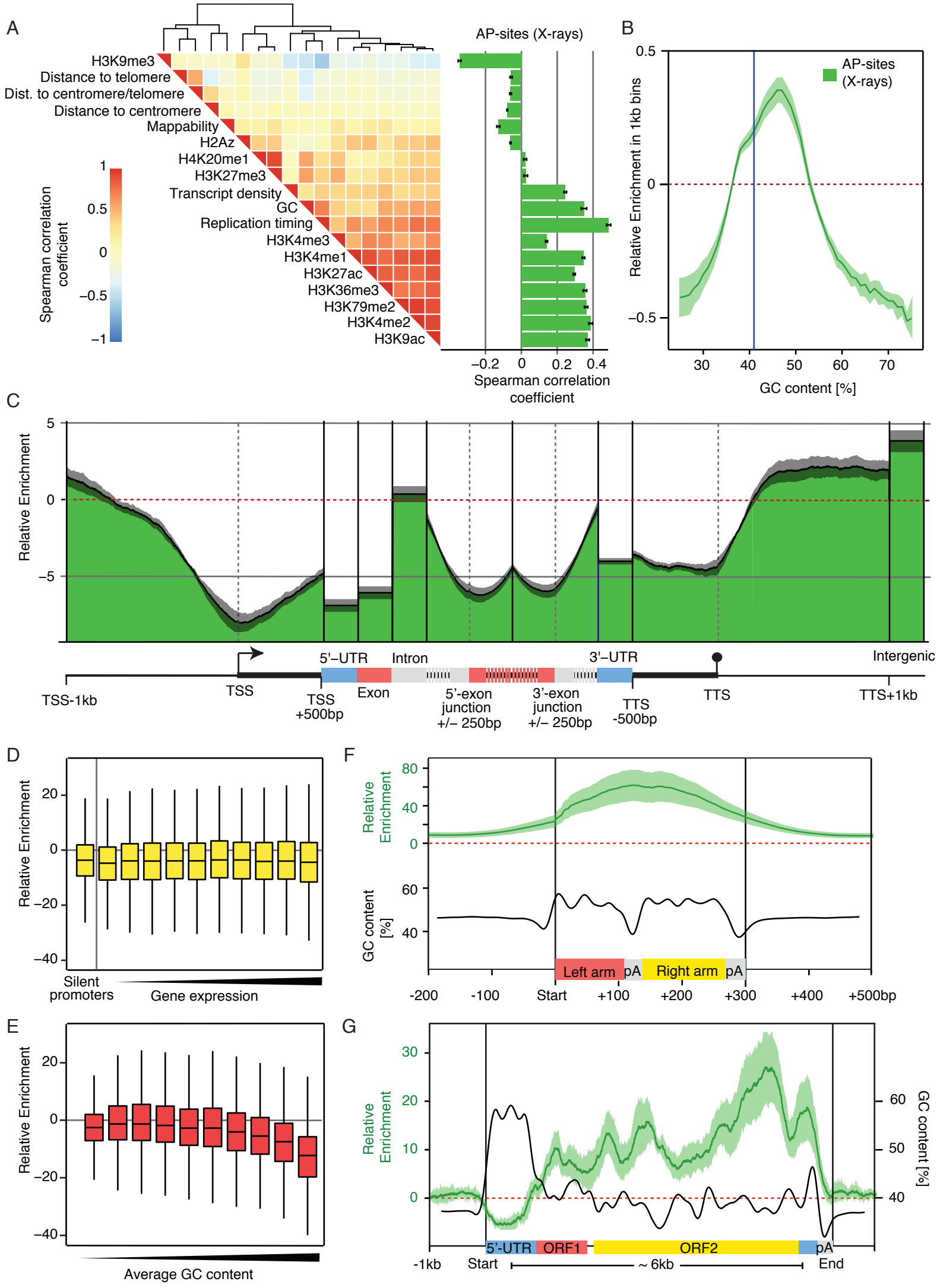
Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7.2 Figure 2. Oxidative damage distribution is associated with genomic features. **(A)** Bar plot displays the average correlation of damage levels with large-scale chromatin and other features in HepG2 cells at 100kb resolution. Damage correlates with euchromatic features and anticorrelates with heterochromatic ones, the opposite of that observed for cancer SNVs. The heatmap shows the relationship between the features, grouped using hierarchical clustering. **(B)** The plot shows dependence between Relative Enrichment of damage and genomic GC content at 1kb resolution. Damage levels increase with GC content and then surprisingly fall in high GC areas. The blue line marks the genomic average GC content of 41%. **(C)** Metaprofile of Relative Enrichment over ~23,000 protein-coding genes (n_{genes}=23,056, $n_{promoters}$=48,838, n_{5UTRs}=58,073, n_{exons}=214,919, $n_{introns}$=182,010, n_{3UTRs}=28,590, $n_{termination}$=43,736, $n_{intergenic}$=22,480). Damage levels for UTRs, exons, introns, and intergenic regions are averaged across each feature due to their variable sizes. Coding and regulatory regions are depleted for damage, whereas introns have near intergenic damage levels. **(D, E)** Boxplots depict damage levels at 48,838 promoters binned into unexpressed and expression deciles (D), and average GC content deciles (E). Promoters are defined as the transcriptional start sites +/- 1kb. Damage is not transcription-dependent, but reduces with increasing promoter GC content. **(F, G)** Metaprofiles of Relative Enrichments and average GC contents across 848,350 *Alu* and 2,533 *LINE* elements. There is a very large accumulation of damage inside these features. All panels display measurements for X-ray treated samples. Error bars and shaded borders show the standard error of mean across biological replicates.

Figure 2 bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

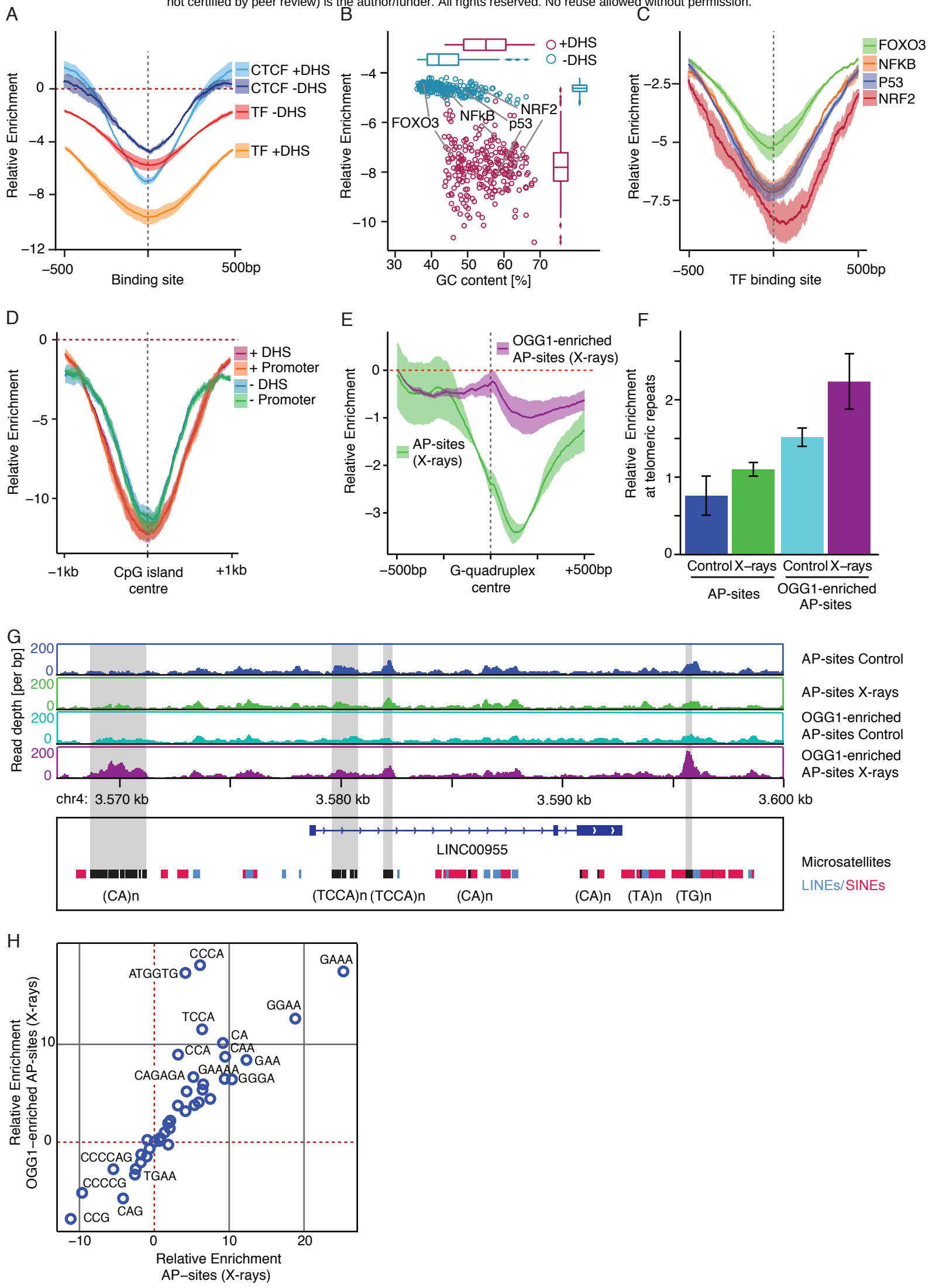


bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7.3 Figure 3. Oxidative damage distribution is associated with regulatory sites and repeats. (**A**) Metaprofiles of Relative Enrichments centred on CTCF- and DNA-binding sites within and outside DNase hypersensitive regions (DHS; $n_{CTCFinDHS}$=37,763, $n_{CTCFnotDHS}$=10,908, $n_{TFbsInDHS}$=253,613, $n_{TFbsNotDHS}$=5,463,612). Damage levels are reduced around binding sites. Shaded borders show the standard error of mean across biological replicates. **(B)** Scatter plot of average Relative Enrichments and GC contents +/-500bp of binding sites for each transcription factor. Binding sites are separated into within and outside DNase hypersensitive sites. Damage levels are universally reduced regardless of transcription factor, with particularly lowered levels for actively used sites in DHS regions. **(C)** Metaprofiles centred on binding sites for 4 selected transcription factors. **(D)** Metaprofiles centred on CpG islands, within and outside promoters and DHS regions (n_{DHS}=17,565, n_{NotDHS}=9878, $n_{Promoter}$=14850, $n_{NotPromoter}$=12,593). Damage levels are reduced regardless of location and accessibility. **(E)** Metaprofiles centred on predicted G4-quadruplex structures (n=359,449). There are asymmetrically reduced damage levels for AP-sites, but not for OGG1-enriched AP-sites. **(F)** Bar plots of average Relative Enrichments in G4-quadruplexes at telomeric repeats across the 4 treatment and processing conditions. Damage levels are increased in OGG1-enriched samples. Error bars show the standard error of mean across biological replicates. **(G)** Genome browser views of damage levels in ~30kb locus surrounding LINC00955, including microsatellite repeats. Some groups of microsatellites accumulate large amounts of damage and reduced 8OxoG processing. **(H)** Scatter plot displaying average damage levels in different microsatellites types for the AP-site and OGG1-enriched samples. Most types display similar damage levels in the two processing conditions; however, several display elevated damage in the OGG1-enriched sample. All panels display measurements for X-ray treated samples, unless indicated otherwise.

Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

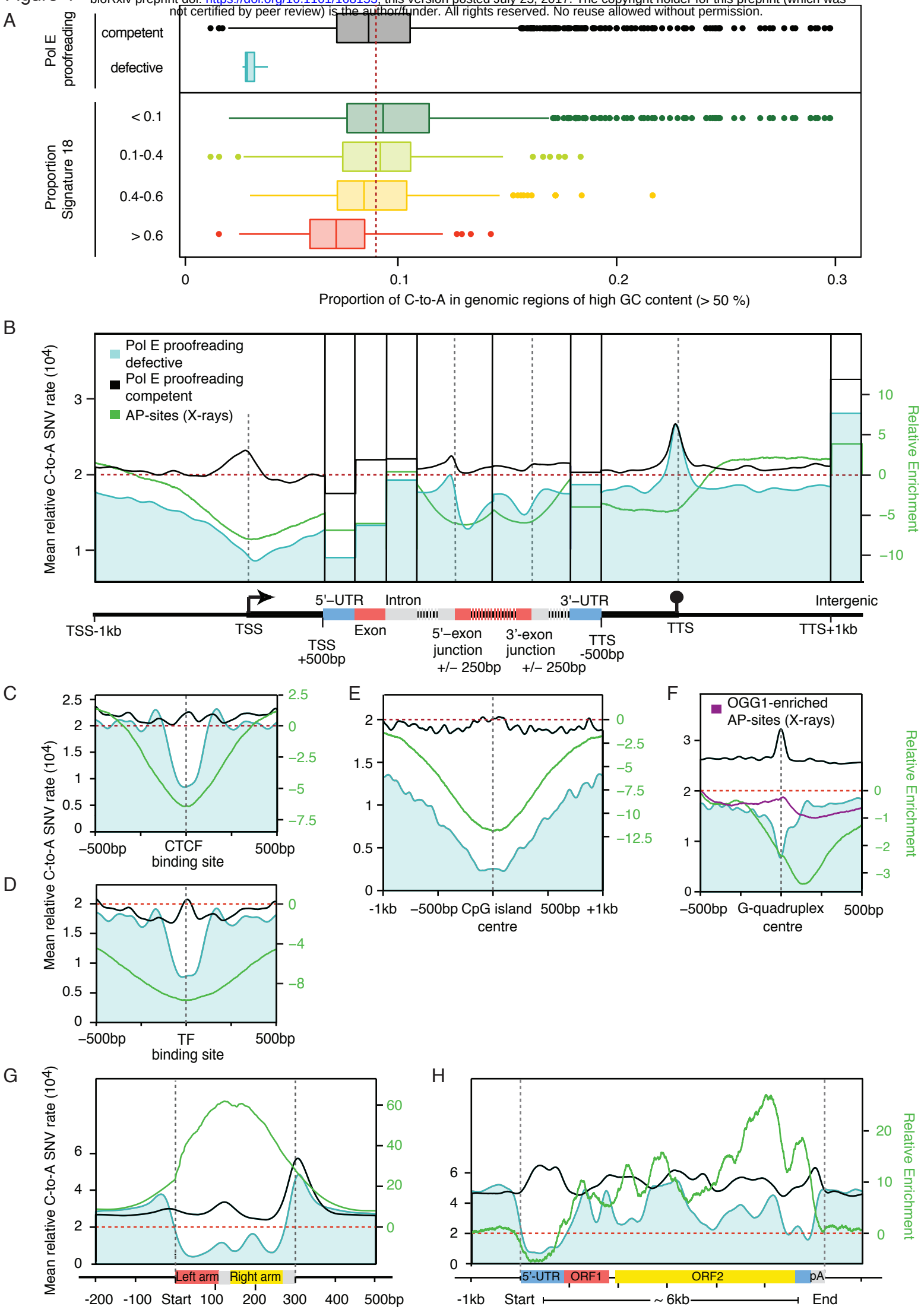


bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7.4 Figure 4. Oxidative damage patterns are reflected in cancer mutagenesis. **(A)** Boxplots of the proportion of C-to-A SNVs in genomic regions of high GC content. Tumour samples are separated into those that are Pol E-proofreading defective (n=8) and to all other tumours (n=2,694), and into 4 groups according to Mutational Signature 18 contributions ($n_{<0.1}$=1398, $n_{0.1\text{-}0.4}$=322, $n_{0.4\text{-}0.6}$=540, $n_{>0.6}$=141). Tumours that are proofreading defective and high in Signature 18 display lower proportions of SNVs in GC-rich regions. **(B)** Metaprofile of SNV rates over ~23,000 protein-coding genes in proofreading defective and control tumours. The damage profile is overlaid for comparison. The oxidative damage-dependent SNV profiles in proofreading-defective tumours show similar distributions to AP-sites, whereas the pattern is lost in control tumours. **(C-F)** Metaprofiles of SNV rates centred on CTCF-binding sites (n=48,671), transcription factor-binding sites in DHS regions (n=253,613), CpG islands (n=27,443), and G-quadruplex structures (n= 359,449). SNV profiles in proofreading defective tumours mimic the damage profiles. **(G, H)** Metaprofiles across 848,350 *Alu* and 2,533 *LINE* elements. SNV rates in proofreading defective tumours are reduced compared with damage profiles.

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4

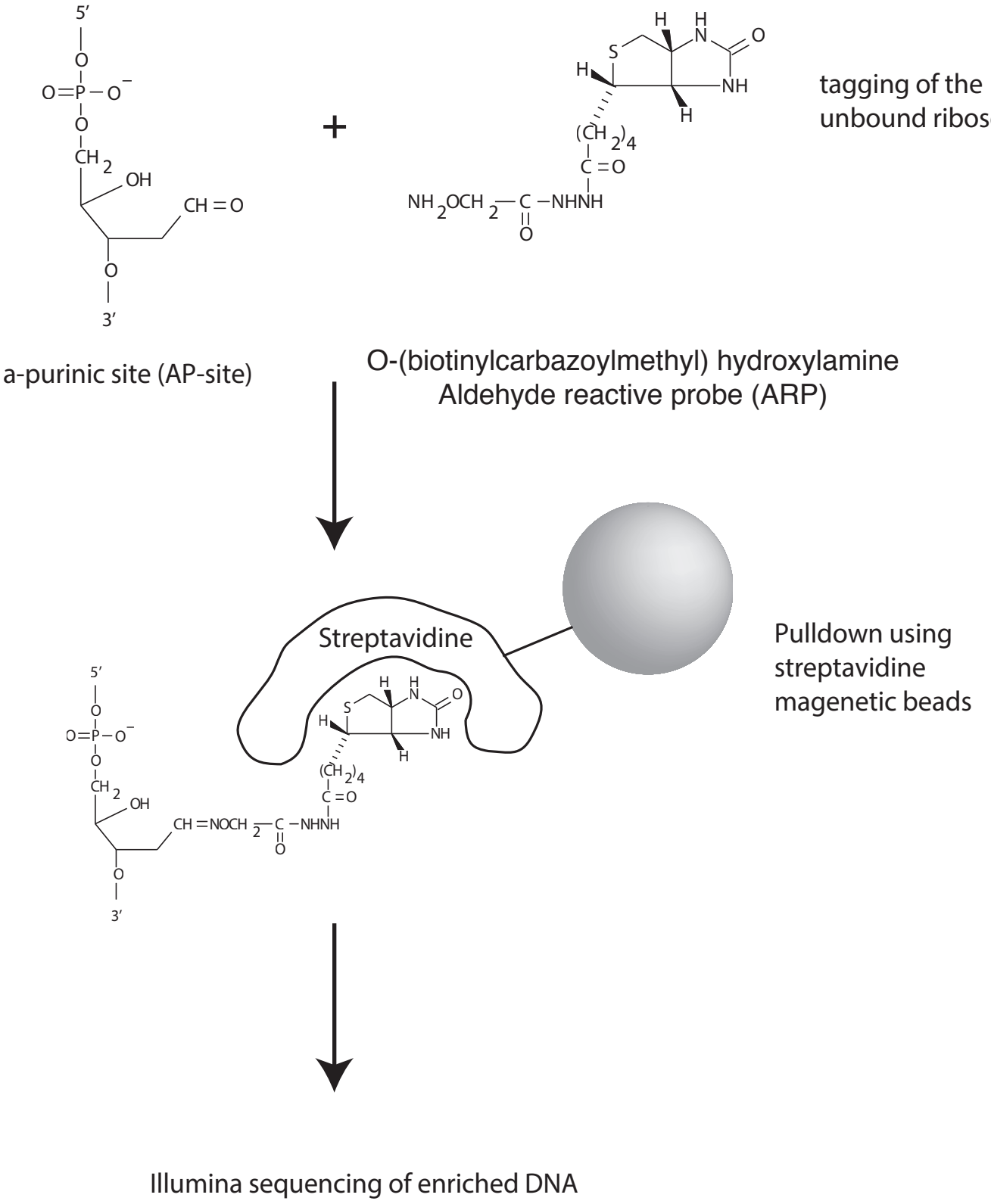


bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7.5 Supplementary Figure S1. Schematic diagram of the chemical enrichment process of AP-sites using an aldehyde reactive probe.

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Figure S1



bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

8. Methods

8.1 Cell culture and X-ray treatment

HepG2 cells were cultivated at 37°C and 5% CO2 in Dulbecco's Modified Eagle Medium (DMEM; Invitrogen) supplemented with 1% essential amino acids, 1% pyruvate, 2% penicillin/streptavidin and 10% heat-inactivated fetal bovine serum (FBS). ~1x10^6 cells were exposed to 6Gy X-ray using a SOFTEX M-150WE in triplicates. Triplicate samples of untreated control cells were processed in parallel, excluding irradiation. Cells were harvested 30 minutes post-treatment.

8.2 AP-Seq

Total genomic DNA was extracted using a Blood and Tissue Kit (Qiagen, catalogue number 69506) and genomic DNA was kept on ice during the process. 5.7µg of genomic DNA was tagged with biotin using 5mM Aldehyde Reactive Probe [13] (ARP; Life Technologies, catalogue number A10550) in phosphate buffered saline (PBS) for 2h at 37°C. Genomic DNA was then purified using AMPure beads as described above and was fractionated using a Covaris fractionator in 130µl for a mean fragment length of 300bp. After separating 30µl for sequencing as the input sample, the remaining DNA was used for biotin-streptavidin pulldown, using MyOne Dynabeads (Life Technologies, catalogue number 65601). 120µl beads (10µl per sample) were washed 3x with 1ml 1M NaCl in Tris-EDTA buffer (TE buffer) and re-suspended in 100µl 2M NaCl in TE and then added to 100µl of the sonicated DNA. Samples were rotated at room temperature for 10h. Subsequently the beads were washed 3x with 1M NaCl in TE and finally re-suspended in 50µl TE for library preparation.

For the *in vitro* OGG1-enrichment, 10µg of genomic DNA was digested with recombinant OGG1 (New England Biolabs, catalogue number M0241L). 0.1µg enzyme was taken for 1µg of genomic DNA in New England Biolabs (NEB)-buffer 2 and bovine serum albumin (BSA) for 1h, 37°C. Digested DNA was subsequently purified using AMPure beads (Agencourt, catalogue number A63882) with 1.8x bead solution, 2x 70% ethanol washing; beads were not allowed to dry to prevent DNA from sticking. The DNA was subsequently tagged with ARP as described above.

8.3 Library preparation and sequencing

Both the damage-enriched and input DNA were *in vitro* repaired using PreCR (NEB catalogue number M0309L). The input DNA and supernatant of the pull-down were purified using AMPure beads. The purified pull-down was recombined with the beads and library preparation was performed on the re-pooled sample using a 125bp paired-end ChIP-Seq library preparation kit (KAPA Biosystems catalogue number KK8504) and sequenced using an Illumina HiSeq 2000 on first a rapid and then a high-output run (catalogue number FC-401-4002). The resulting data were subsequently combined.

8.4 Read processing library normalisation and damage quantification

Unless stated, data-processing was performed using R 3.4.0 and Bioconductor 3.5.

The quality of damage-enriched AP-seq samples (n=12) and corresponding input samples were checked using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/); the quality was sufficient that no further filtering was required before alignment. The reads were mapped to the reference human genome (version hg19) using the Bowtie2 algorithm (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) [22] with standard settings, allowing for 2 mismatches and random assignment of non-uniquely mapping reads. To confirm the robustness of key results,

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

analyses were repeated excluding non-uniquely mapped reads (reads with FLAG 3 filtered using SAMtools; http://samtools.sourceforge.net/) [23]. Data were visualised with the Integrative Genomics Viewer version 2.3.92 (http://software.broadinstitute.org/software/igv/) [24].

Paired reads were imported into R using the "GenomicAlignments" and "rtracklayer" [25] packages. Paired reads mapping more than 1kb apart were discarded. Filters were applied to assess read duplication, reads mapping to the Broad Institute blacklist regions (ftp://encodeftp.cse.ucsc.edu/users/akundaje/rawdata/blacklists/hg19/wgEncodeHg19ConsensusSignalArtifactRegions.bed.gz) [26], and whether reads overlap with repeats annotated in the UCSC RepeatMasker track from the UCSC Table Browser (rrmsk_hg19.bed). The main analysis was performed without applying these filters, but the robustness of key results was confirmed by repeating analyses with the filters.

Inter-library normalization was performed using only genomic areas of low damage. It was necessary to consider that increased exposure to DNA damage leads to increased library sizes. A global scaling factor was calculated as the mean read coverage in a low-damage subset (10 %) of 100kb bins, which were identified by their read coverage as the lowest decile of 100kb bins over the mean of all samples.

Relative Enrichment of DNA damage was assessed through the normalised log2 fold-change of the enriched sample over input (termed Relative Enrichment). Analyses were restricted to Chromosomes 1 to 22 and X, except for the 100kb damage distribution map which includes the Y chromosome (Figure 1B).

All analyses were performed using the average Relative Enrichment in appropriate bin sizes tiled across the genome or covering genomic elements. Genome browser images were generated using absolute read counts pooled over replicates.

8.5 Analysis on local oxidative damage distribution

The karyogram map was compiled using the mean of the replicates at 100kb resolution with "ggbio" [27] karyogram plot fixing the colour scale to a Relative Enrichment of -1 to 1. Enrichment over chromosomes was also depicted with 100kb resolution for the mean of the replicates with shades depicting the standard error of the mean of triplicates. For illustration purposes data were smoothed with a Gaussian smooth over 10 bins, using the smth.gaussian function of the "smoother" package. Correlations at 100kb resolution were performed using Spearman correlation. Fine resolution images were depicted using the IGV browser without any additional smoothing applied.

8.6 Epigenome and feature analysis

Genome-wide feature sets were obtained from the UCSC Genome Browser. Chromatin features for HepG2 cells were retrieved from the data repository generated in the context of the ENCODE consortium and obtained through https://www.encodeproject.org/ [26]. Where applicable, datasets were pooled. Accession numbers are listed in Table 1.

Transcript density was calculated through the genome coverage with any one transcript as defined by UCSC. Distance to telomeres and centromeres was calculated as the absolute base pair distance to annotated telomeres and centromeres.

Genomic and chromatin features were calculated as mean values in 100kb bins over the genome and clustered using hierarchical clustering of Spearman's correlation coefficients. Features were

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

then correlated (also Spearman) to the individual DNA damage levels. Data points represent the mean of the correlation coefficients with the standard error of the mean over replicates.

Feature	ENCODE accession numbers, URL, or UCSC table browser ID
DNase hypersensitivity	ENCFF774LVT
H3K4me3	ENCFF000BGT
H3K4me2	ENCFF000BFV
H3K4me1	ENCFF000BFC
H3K27me3	ENCFF001FLH, ENCFF001FLI
H3K9me3	ENCFF000BEW
H2Az	ENCFF000BEK
H4K20me1	ENCFF000BFJ
H3K36me3	ENCFF001FLR, ENCFF001FLS
H3K79me2	ENCFF000BGB
H3K27ac	ENCFF000BGH
H3K9ac	ENCFF000BGM
RNA-Seq	ENCFF000DPL, ENCFF000DPM, ENCFF000DPN, ENCFF000DPO
Replication timing	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqHepg2WaveSignalRep1.bw
Mappability	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/wgEncodeCrgMapabilityAlign100mer.bigWig
Transcription factor binding sites	http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/tfbsConsSites.txt
CTCF binding sites	ENCFF661OYF
DNase hypersensitivity sites	wgEncodeAwgDnaseUwdukeHepg2UniPk.bed

Table 1: HepG2 specific datasets obtained from ENCODE and genomic annotation datasets obtained from the UCSC browser

6.7 GC content analysis

GC content preference of DNA damage distribution was assessed at 1kb resolution. For each 1kb bin in the genome, GC content was calculated and rounded to the closest percentage. Bins with

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

more than 10% undefined sequence were censored. For all bins falling into a particular percentage range, mean Relative Enrichment was calculated with also the standard error of the mean for biological replicates. For display, a Gaussian smooth was applied reaching over 10% GC content range.

6.8 DNA damage distribution over gene profile

Metaprofiles over coding genes were compiled using the UCSC transcript annotation. The mean was taken for different elements of the genes, which are comprised of a total of 26,860 transcripts. Gene elements were either centred around an appropriate centre point, in which case the mean Relative Enrichment was calculated for each base pair in the respective region. For gene elements of different sizes the mean over the gene element was taken. Independent of their size they were weighted as equal in subsequent analyses. The metaprofile was then compiled with the different gene elements in the following order: 48,838 promoters were centred around the transcriptional start site with 1kb sequence in 5' direction and 500bp in 3'. 58,073 5'-UTRs, 214,919 exons, and 182,010 introns were addressed as a scaled mean. In addition, exons and introns were addressed through the exon-intron junction, both 5' and in 3' of the exon +/- 250bp. Given the small sizes of exons, 250bp partially also contains following gene elements. The end of genes is represented through the means of 28,590 3'-UTRs and 43,736 transcription termination sites with 500bp in 5' direction and 1kb in 3'. 22,480 intergenic regions were addressed as the mean of each region. Shades represent the standard error of the mean over biological replicates.

8.9 GC content and transcription dependent promoter analysis

Gene transcription was assessed using RNA-Seq data for HepG2 cells from the ENCODE consortium (Table 1). Replicates were pooled and RNA-Seq coverage was calculated for each unique UCSC defined transcript (n = 57,564). Promoters, i.e. the transcriptional start sites +/- 1kb for each transcript were grouped into 11,058 silent promoters and the remaining 46,506 into deciles of increased transcriptional use. In parallel, the mean GC content for each promoter was calculated, which were then also grouped into deciles based on their GC content. Mean damage was assessed for each promoter in these groups.

8.10 Retrotransposon analysis

Retrotransposon information was obtained from the UCSC repeat masker. For repetitive sequences, there is a risk of mapping issues and errors of annotation. Therefore, retrotransposon analysis was limited to families of these repeats, where location issues should not arise and mis-estimation of total repeat numbers should largely be balanced out through the IP vs. input comparison. Analyses for particular locations were confirmed by excluding ambiguous mapping.

LINE elements were defined as belonging to *LINE* element families of L1PA7 or newer and only considered, if the size fell between 5.9 and 6.1kb (n=2,533). Alus were considered when 270 to 330bp in size (n=848,350). Retrotransposons were anchored to their start sites and addressed with flanking regions from the start –1kb to +7kb for *LINE* elements and -200bp to +500bp for *Alu* elements. Metaprofiles were compiled as the mean Relative Enrichment over the respective region. GC content was assessed as the mean GC content at the particular site and smoothed using Gaussian smoothing in windows of 5% of feature length.

8.11 Transcription factor binding sites, CpG islands and G4-quadruplex structure analysis

Transcription factor binding sites were obtained as the consensus set from ENCODE (Table 1), which is cell line unspecific. (n=5,717,225). HepG2 cell specific CTCF binding sites (n=48,671) and DNase hypersensitivity sites (n=192,735) were obtained through ENCODE and UCSC respectively (Table 1). G4-quadruplex (G4) structures were obtained using the G4Hunter method

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[28], utilising directly the reference file QP37_hg19_ref.RData provided with the associated R package (n=359,446) with the exception of telomeric G4 structures with the centre less than 500bp away from the chromosome end (n=3). CpG islands were defined through UCSC (n=27,443). Features were considered to be in a promoter, if they overlap with the region of a transcriptional start site +/-1kb. They were considered to overlap with DNase hypersensitivity only when the feature itself overlaps with a DNase hypersensitivity site. For metaprofiles the centres of the features were considered and mean Relative Enrichment of damage levels assessed relative to the centre point. For quantification of mean damage at a given feature site, only the feature itself was addressed and quantified as the mean Relative Enrichment over the region. The GC content of transcription factor binding site was however calculated as the mean over the region (+/-500bp) around the transcription factor binding site. Groups of features were summarised using the median.

8.12 Telomere analysis

Due to expected mapping artefacts at telomeric repeats, telomeres were addressed separately not using the aligned sequence. Instead, Telomere hunter version 1.0.4. (https://www.dkfz.de/en/applied-bioinformatics/telomerehunter/telomerehunter.html) [29] was used to filter out reads that map to telomeric repeats. These were reassigned to intratelomeric and subtelomeric regions or other locations. Of these, only the intratelomeric repeats were considered. Normalisation between libraries was performed not within the Telomerehunter package but separately with the global scaling factor as described above. Mean Relative Enrichment between biological replicates was calculated with the standard error of the mean.

8.13 Microsatellite analysis

Microsatellites were defined through the UCSC repeat masker as the "Simple_repeat" class. For quantification purposes, reverse complement repeat classes were combined. Only microsatellite sequences that are represented >1,000 times in the genome were considered. This leaves 39 repeat types, which are represented by a total of 388,350 repeats. Median Relative Enrichment of damage was quantified over each microsatellite type.

8.14 Patient selection for mutation analysis

Data for mutations in cancer were obtained from the Pan-cancer Analysis of Whole Genomes consortium [17]. Contributions of mutational signatures were provided by PCAWG working group 7. [20]

The data set is comprised of 2,702 tumour-normal pairs for 39 cancer types. From this dataset, we obtained all data on mutation rates and mutation signature contributions, as well as clinical metadata. The analysis was restricted to chromosomes 1 to 22 and X.

For the mutations dependent on oxidative damage, 8 samples were selected that have a polymerase epsilon proofreading defect as determined by a hypermutator phenotype (C-to-A >100,000) with prominence of Signature 10. In total, these samples contain 3,436,531 mutations. For information to individual patients see Table 2.

For comparison, all other 2,695 tumour samples were taken with a total of 6,008,940 C-to-A mutations.

Tumour samples with mutations in direct processing of 8OxoG or AP-sites were identified through assessing, whether mutations fall into the coding sequence of OGG1 (n=7) or APEX1 (n=3). Mutations were considered, if their effect determined by the ensembl VEP tool (http://www.ensembl.org/Multi/Tools/VEP) [30] identified them as missense variants, stop codon

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

gained, frameshift variants, or splice donor variant. They were not considered, if there was an underlying hypermutator phenotype of >100,000 C-to-A mutations (n=2). Information to individual patients can be found in Table 3.

Patients with oxidative damage induced mutations beyond polymerase epsilon proofreading defects were separated based on the proportion contribution of Signature 18 to C-to-A mutations. Patients were censored that have a hypermutator phenotype (C-to-A >100,000, which includes Pol E proofreading deficient tumour samples) or coding mutations in 8OxoG or AP-site processing as identified above. In addition, patients were also censored based on documented smoking history or previous exposure to chemotherapy/radiotherapy. A total of 2,401 samples were used for analysis. They were grouped into Signature 18 based groups of <10% (n=1,398), 10% to 40% (n=322), 40% to 60% (n=540), and >60% (n=141).

Sample IDs [tumor_wgs_aliquot_id]	Total C-to-A	Proportion C-to-A	Total mutation count	Cancer type	Proportion Signature 10
00aa769d-622c-433e-8a8a-63fb5c41ea42	115,337	0.46	252,195	ColoRect-AdenoCA	0.65
0980e7fd-051d-45e9-9ca6-2baf073da4e8	396,377	0.44	907,411	ColoRect-AdenoCA	0.60
14c5b81d-da49-4db1-9834-77711c2b1d38	989,958	0.40	2,502,427	ColoRect-AdenoCA	0.57
154f80bd-984c-4792-bb89-20c4da0c08e0	126,870	0.45	280,527	ColoRect-AdenoCA	0.63
2df02f2b-9f1c-4249-b3b4-b03079cd97d9	964,307	0.38	2,570,161	ColoRect-AdenoCA	0.41
6ca5c1bb-275b-4d05-948a-3c6c7d03fab9	243,272	0.28	871,206	ColoRect-AdenoCA	0.55
93ff786e-0165-4b02-8d27-806d422e93fc	436,686	0.43	1,024,918	ColoRect-AdenoCA	0.50
b0a83df8-dd2c-4c1b-b238-9081d2c22258	163,724	0.54	303,201	Uterus-AdenoCA	0.65

Table 2: Selected tumour samples with polymerase epsilon proofreading defect.

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sample IDs [tumor_wgs_aliquot_id]	Total C-to-A	Proportion C-to-A	Total mutation count	Cancer type	Coding Mutation
42f88b95-fa12-47c7-93f1-cf72f207291c	1,308	0.16	7,945	Kidney-RCC	OGG1
4a1ad661-f6ae-44e8-b50b-72ff658ff22b	1,480	0.19	7,872	CNS-GBM	OGG1
7456abd5-303e-4e6f-bf4e-47efefc7310f	913	0.16	5,729	Breast-AdenoCA	OGG1
dc4ba4bc-6333-4fe9-8805-e058cc9e6e18	1,963	0.17	11,509	Panc-Endocrine	OGG1
e6801359-d1d7-4871-b2fb-180674a2e469	1,506	0.16	9,141	Kidney-RCC	OGG1
f7e7d61f-e2dc-b523-e040-11ac0c482000	477	0.16	3,011	Breast-AdenoCA	OGG1
fc5dc6d8-62d2-76d8-e040-11ac0d4863c3	1,494	0.17	8,637	Breast-AdenoCA	OGG1
45a7949d-e63f-4956-866c-df51257032de	2,631	0.10	25,181	Bladder-TCC	APEX1
9ebac79d-8b38-4469-837e-b834725fe6d5	2,594	0.16	16,191	Panc-AdenoCA	APEX1
bf91afc4-aa2b-4365-80c5-b98c9d118e10	333	0.13	2,543	Panc-Endocrine	APEX1

Table 3: Selected tumour samples with coding mutations in OGG1 and APEX1.

8.15 GC content preferences of mutation rates

For each 1kb bin in the genome, GC content was calculated and rounded to the closest percentage. Bins with more than 10% undefined sequence were censored. Mutations falling into bins of 50% GC content or higher was calculated as proportion of the total C-to-A mutation counts. Assuming equal distribution dependent only on base content, a total of 9% of C-to-A mutations would be expected to fall into such high GC content areas of the genome.

8.16 Genomic features analysis

Metaprofiles over genomic features were calculated for the features with the same selection strategy as described above. For this, C-to-A mutations were pooled for each patient group. Mean relative mutation rates over features were calculated as relative C-to-A mutation density normalized to 1,000,000 C-to-A mutations per patient group. The mean over the features was normalized for sequence content of the particular location by dividing with a factor of the local GC content divided by the average of 41%. For display purposes, data were smoothed using a Gaussian smooth spreading over 100bp for the gene body profile, *Alus*, protein binding sites,

bioRxiv preprint doi: https://doi.org/10.1101/168153; this version posted July 25, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CpG islands, and G4 structures. *LINE* elements were smoothed using Gaussian smoothing over 200bp to account for the increased noise originating from the lower frequency of this particular feature.

9. Acknowledgements

We are most grateful to Peter Van Loo, Kerstin Haase, Clemency Jolly, Jonas Demeulenmeester, and Maxime Tarabichi for their advice and technical assistance for analysing cancer genomics data. For technical assistance with experiments and sequencing we would like to thank Hiroki Goto, Shinichi Yamasaki, Keigo Hikishima, Rehab Abdelhamid, Lorea Blazquez, and Mary Bronks. For advice on data analysis we would like to thank the Crick Bioinformatics Science Technology Platform, in particular Harshil Patel. This work was supported by the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001110), the UK Medical Research Council (FC001110), and the Wellcome Trust (FC001110). ARP and NML were also supported by funding from the Okinawa Institute of Science & Technology Graduate University and a Wellcome Trust Investigator Award. ARP was funded by a postdoctoral fellowship from the Peter and Traudl Engelhorn Foundation. We thank Peter Van Loo and Jernej Ule for helpful advice and discussions throughout the project.