

University of Massachusetts Medical School

eScholarship@UMMS

---

Open Access Articles

Open Access Publications by UMMS Authors

---

2014-11-20

## Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes


Christina L. Zheng

*Oregon Health & Science University*

*Et al.*

### Let us know how access to this document benefits you.

Follow this and additional works at: <https://escholarship.umassmed.edu/oapubs>

 Part of the [Bioinformatics Commons](#), [Cancer Biology Commons](#), [Computational Biology Commons](#), [Genetics Commons](#), [Genomics Commons](#), and the [Neoplasms Commons](#)

---

### Repository Citation

Zheng CL, Fudem GM, Purdom E, Cho RJ. (2014). Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. Open Access Articles. <https://doi.org/10.1016/j.celrep.2014.10.031>. Retrieved from <https://escholarship.umassmed.edu/oapubs/2581>

Creative Commons License

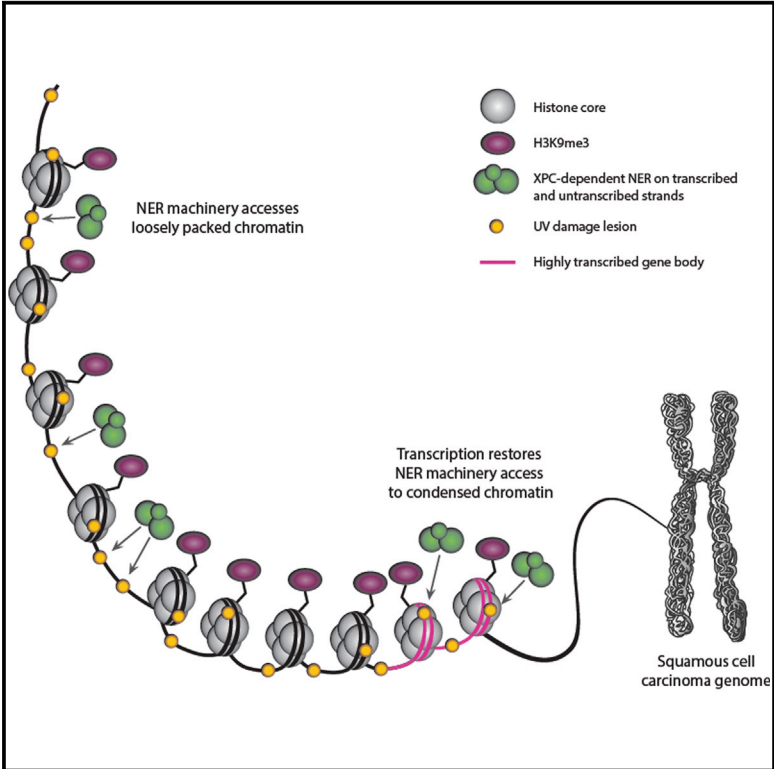


This work is licensed under a [Creative Commons Attribution 3.0 License](#).

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Open Access Articles by an authorized administrator of eScholarship@UMMS. For more information, please contact [Lisa.Palmer@umassmed.edu](mailto:Lisa.Palmer@umassmed.edu).

# Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes

## Graphical Abstract



## Authors

Christina L. Zheng, Nicholas J. Wang, ..., Elizabeth Purdom, Raymond J. Cho

## Correspondence

epurdom@stat.berkeley.edu (E.P.), chorj@derm.ucsf.edu (R.J.C.)

## In Brief

Zheng et al. report that variable mutation densities within cancer genomes result from differential access of DNA repair machinery, imposed by chromatin state. By showing that transcription restores DNA repair to tightly packaged DNA, their study reveals natural differences in expression level as a potentially important modulator of oncogene mutation rate.

## Highlights

Regional genomic mutational rates reflect differential access to DNA repair

Transcription restores DNA repair access to tightly packaged chromatin

We model gene mutation rate based on transcription level and chromatin state

# Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes

Christina L. Zheng,<sup>1,2</sup> Nicholas J. Wang,<sup>3</sup> Jongsuk Chung,<sup>4</sup> Homayoun Moslehi,<sup>5</sup> J. Zachary Sanborn,<sup>6</sup> Joseph S. Hur,<sup>7</sup> Eric A. Collisson,<sup>8</sup> Swapna S. Vemula,<sup>9</sup> Agne Naujokas,<sup>9</sup> Kami E. Chiotti,<sup>10</sup> Jeffrey B. Cheng,<sup>5</sup> Hiva Fassihi,<sup>11</sup> Andrew J. Blumberg,<sup>12</sup> Celeste V. Bailey,<sup>13</sup> Gary M. Fudem,<sup>14</sup> Frederick G. Mihm,<sup>15</sup> Bari B. Cunningham,<sup>16</sup> Isaac M. Neuhaus,<sup>5</sup> Wilson Liao,<sup>5</sup> Dennis H. Oh,<sup>5,17</sup> James E. Cleaver,<sup>5</sup> Philip E. LeBoit,<sup>9</sup> Joseph F. Costello,<sup>18</sup> Alan R. Lehmann,<sup>19</sup> Joe W. Gray,<sup>2,3</sup> Paul T. Spellman,<sup>2,10</sup> Sarah T. Arron,<sup>5</sup> Nam Huh,<sup>4</sup> Elizabeth Purdom,<sup>20,21,\*</sup> and Raymond J. Cho<sup>5,21,\*</sup>

<sup>1</sup>Division of Bioinformatics and Computational Biology, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Sciences University, Portland, OR 97239, USA

<sup>2</sup>Knight Cancer Institute, Oregon Health & Sciences University, Portland, OR 97239, USA

<sup>3</sup>Department of Biomedical Engineering, Oregon Health & Sciences University, Portland, OR 97239, USA

<sup>4</sup>Emerging Technology Research Center, Samsung Advanced Institute of Technology, Kyunggi-do 446-712, Korea

<sup>5</sup>Department of Dermatology, University of California, San Francisco, San Francisco, CA 94143, USA

<sup>6</sup>Five3 Genomics, LLC, Santa Cruz, CA 95060, USA

<sup>7</sup>Headquarters, Samsung Electronics, Seocho-gu, Seoul 137-857, Korea

<sup>8</sup>Department of Medicine, University of California, San Francisco, San Francisco, CA 94143, USA

<sup>9</sup>Department of Pathology, University of California, San Francisco, San Francisco, CA 94143, USA

<sup>10</sup>Department of Molecular and Medical Genetics, Oregon Health & Sciences University, Portland, OR 97239, USA

<sup>11</sup>National Xeroderma Pigmentosum Service, St John's Institute of Dermatology, Guy's and St Thomas' NHS Trust, London SE1 9RT, UK

<sup>12</sup>Department of Mathematics, University of Texas, Austin, Austin, TX 78712, USA

<sup>13</sup>UCSF Helen Diller Family Comprehensive Cancer Center, San Francisco, CA 94158, USA

<sup>14</sup>Department of Surgery, University of Massachusetts Medical School, Worcester, MA 01655, USA

<sup>15</sup>Department of Anesthesiology, Pain and Perioperative Medicine, Stanford University Medical Center, Stanford, CA 94305, USA

<sup>16</sup>Department of Dermatology, University of California, San Diego, La Jolla, CA 92093, USA

<sup>17</sup>Dermatology Research Unit, Veterans Affairs Medical Center, San Francisco, San Francisco, CA 94121, USA

<sup>18</sup>Department of Neurological Surgery, University of California, San Francisco, CA 94143, USA

<sup>19</sup>Genome Damage and Stability Centre, University of Sussex, Brighton BN1 9RH, UK

<sup>20</sup>Department of Statistics, University of California, Berkeley, Berkeley, CA 94720, USA

<sup>21</sup>Co-senior author

\*Correspondence: [epurdom@stat.berkeley.edu](mailto:epurdom@stat.berkeley.edu) (E.P.), [chorj@derm.ucsf.edu](mailto:chorj@derm.ucsf.edu) (R.J.C.)

<http://dx.doi.org/10.1016/j.celrep.2014.10.031>

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/3.0/>).

## SUMMARY

Somatic mutations in cancer are more frequent in heterochromatic and late-replicating regions of the genome. We report that regional disparities in mutation density are virtually abolished within transcriptionally silent genomic regions of cutaneous squamous cell carcinomas (cSCCs) arising in an *XPC*<sup>-/-</sup> background. *XPC*<sup>-/-</sup> cells lack global genome nucleotide excision repair (GG-NER), thus establishing differential access of DNA repair machinery within chromatin-rich regions of the genome as the primary cause for the regional disparity. Strikingly, we find that increasing levels of transcription reduce mutation prevalence on both strands of gene bodies embedded within H3K9me3-dense regions, and only to those levels observed in H3K9me3-sparse regions, also in an *XPC*-dependent manner. Therefore, transcription appears to reduce mutation prevalence

specifically by relieving the constraints imposed by chromatin structure on DNA repair. We model this relationship among transcription, chromatin state, and DNA repair, revealing a new, personalized determinant of cancer risk.

## INTRODUCTION

Somatic point mutations and chromosomal aberrations in cancer are not distributed uniformly throughout the genome (Alexandrov et al., 2013; Jäger et al., 2013; Lawrence et al., 2013; Polak et al., 2014). Despite the myriad mutational processes active in human cancers (Alexandrov et al., 2013), similar regional patterns of somatic mutation density are observed across many malignancy types, suggesting a common underlying mechanism (Hodgkinson et al., 2012; Lawrence et al., 2013). Chromatin organization heavily influences regional mutation rate, with higher densities of mutation observed in tightly packaged DNA, corresponding to late-replicating portions of the genome and genes with lower expression level (Liu et al., 2013; Schuster-Böckler

and Lehner, 2012). For example, more than 40% of mutation frequency variation is correlated with the heterochromatin-associated histone modification H3K9me3 in both solid and hematologic cancer types (Schuster-Böckler and Lehner, 2012).

The reason why chromatin density and replication timing predict regional heterogeneity in mutation prevalence is unclear. Mutation rate correlates most strongly with H3K9me3 and to a lesser degree with H4K20me3 and H3K79me3 (Schuster-Böckler and Lehner, 2012). All three marks correlate with constitutively closed chromatin states, cytogenetically recognized as heterochromatin (Barski et al., 2007; Mikkelsen et al., 2007), suggesting a specific chromatin conformation may underlie the variance. Higher transcription rates correlate with lower prevalence of mutations originating on transcribed strands of genes (Pleasant et al., 2010), but transcription-coupled nucleotide excision repair (TC-NER) explains only a fraction of observed regional heterogeneity. It has been speculated that late-replicating regions suffer from lower-fidelity DNA synthesis because of depletion of the free nucleotide pool (Liu et al., 2013; Stamatoyannopoulos et al., 2009). However, a direct functional effect of specific chromatin state or replication timing on NER has not been established in humans (Gospodinov and Herceg, 2013). Recently, some melanomas with acquired mutations in NER genes were shown to demonstrate weaker association of mutation density with transcription and DNase I hypersensitivity sites (Polak et al., 2014).

## RESULTS

We sought to understand whether observed differences in regional mutation frequency within cancer genomes were driven primarily by NER activity. We studied tumors from patients with xeroderma pigmentosum (XP), a spectrum of genetic disorders associated with defects in NER (Cleaver, 2005). Patients with loss of function in *XPC* are defective in global genome nucleotide excision repair (GG-NER) but proficient in TC-NER. If regional mutation frequency were caused by NER, in an *XPC*<sup>-/-</sup> background, we would expect regional disparities in mutation to persist within transcriptionally active portions of the genome, but not within transcriptionally silent regions. To test this hypothesis, whole-genome sequences were obtained from cSCCs arising in five patients with homozygous frameshift mutations (C<sub>940</sub>del-1) in the *XPC* gene (Cleaver et al., 2007), as well as from eight patients with no known major germline DNA repair deficiency (repair wild-type [WT]) (Table S1) (Durinck et al., 2011). A total of 3,543,126 point mutations were identified. As expected for skin cancers, transitions (C > T/G > A) typical of UV damage predominated among detected mutations, representing 76% of point mutations in WT cutaneous squamous cell carcinomas (cSCCs) and 86% in *XPC*<sup>-/-</sup> cSCCs. Mutation frequency, measured as transition mutations per kilobase, was explored in relation to chromatin structure, replication time, and gene expression using ENCODE data derived from keratinocytes (ENCODE Project Consortium, 2012).

### Regional Mutation Disparities in Cancer Genomes Result Primarily from DNA Repair

Consistent with recent work (Liu et al., 2013; Schuster-Böckler and Lehner, 2012), we report that mutation prevalence correlated

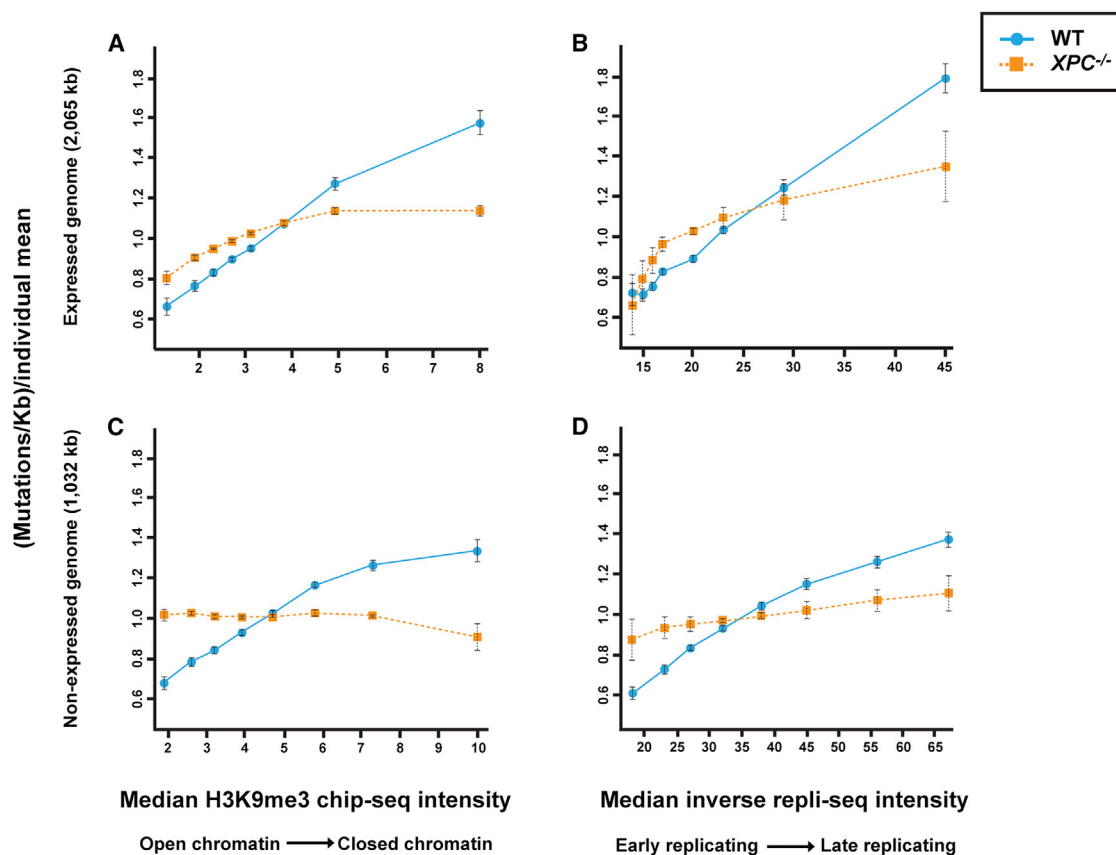
directly with both H3K9me3 density ( $p < 0.001$ ) and replication time ( $p < 0.001$ ), and anticorrelated with density of the repressive mark H3K27me3, within both expressed and nonexpressed portions of WT cancer genomes (Figure 1). Strikingly, in all five examined *XPC*<sup>-/-</sup> cancers, these associations were virtually abolished in nonexpressed portions of the genome, with mutation density at most 10% of that of WT cancers (Figure 1) and reduced to about half of that of WT cancers in expressed portions of the genome, where only TC-NER would be expected to remain active. Increased mutation density was also associated with sparser active histone marks such as H3K27ac and H3K4me1, and these relationships were once again absent within nonexpressed regions of *XPC*<sup>-/-</sup> cancers (Figure S1).

In WT cSCC genomic regions with the lowest H3K9me3 density and highest transcription levels, our measure of TC-NER (the reduction of mutation density resulting from lesions on the transcribed strand, as a proportion of all expected mutations) was 29%–34% (Table S2). Interestingly, in regions with the highest H3K9me3 density and highest transcription levels, this reduction was only 16%–25%, suggesting that exclusion of TC-NER machinery within tightly packaged DNA may decrease its activity. In WT cancers, differences in TC-NER comprised on average only 1.4% of differences between the highest and lowest H3K9me3 densities, at the 70<sup>th</sup> percentile of most highly expressed genome (Table S3). In contrast, in *XPC*<sup>-/-</sup> cancers, 44% of the differences in mutation prevalence between the highest and lowest H3K9me3 levels could be ascribed to differences in TC-NER. Because TC-NER is not affected by loss of function in *XPC*, it is expected that TC-NER would be responsible for a greater proportion of residual disparities in mutation density in *XPC*<sup>-/-</sup> cancers (van Hoffen et al., 1995). Collectively, these findings reveal that the primary cause of regional disparities in mutation prevalence is differential access of DNA repair proteins imposed by chromatin state, specifically NER in cSCCs. Because global patterns of H3K9me3 density correlate with mutation prevalence across many different cancer types (Polak et al., 2014; Schuster-Böckler and Lehner, 2012), it is possible that this mechanism is active in other neoplasms and forms of mutagenesis.

### Transcription Enhances DNA Repair Only in Chromatin-Dense Portions of the Genome

We further analyzed the quantitative effects of GG-NER and TC-NER on mutation density in cancer genomes. In WT cSCCs, regions with greater expression levels showed a significantly decreased density of mutation originating both on the transcribed and untranscribed strands. The magnitude of this effect increased with greater H3K9me3 density and replication time (Figures 2 and S2). Notably, in *XPC*<sup>-/-</sup> cancers, higher expression levels only reduced the frequency of mutations resulting from lesions on the transcribed strand, an effect that can be attributed to TC-NER. However, the transcription-dependent (but TC-NER-independent) DNA repair observed on the untranscribed strand of WT cSCCs is possibly identical to an *XPC*-dependent phenomenon termed transcription domain-associated repair (DAR) (Nospikel and Hanawalt, 2000; Nospikel et al., 2006), which affects both strands in expressed regions.

Although DAR is active on both strands of expressed genes, a representative measure of DAR activity is limited to



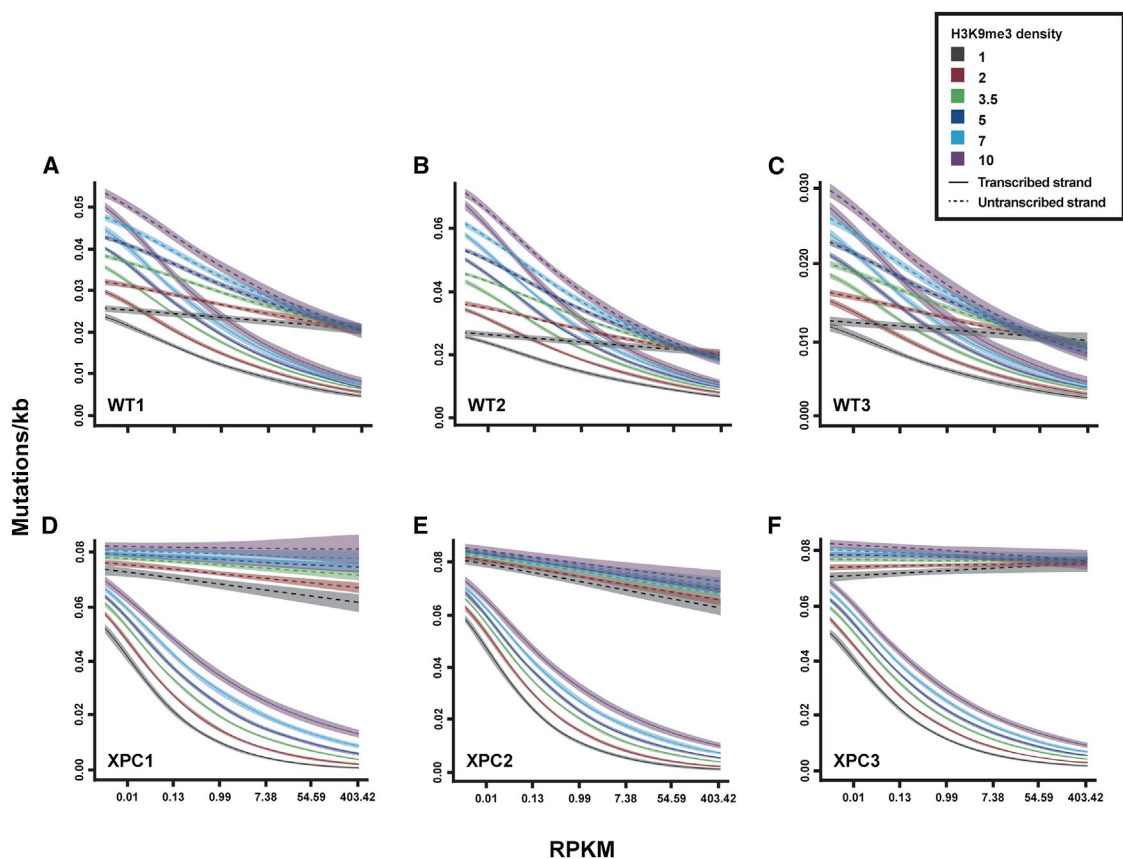
**Figure 1. Regional Disparities in Mutation Density Are Absent in Nonexpressed Portions of the Genome of Germline  $XPC^{-/-}$  Squamous Cell Carcinomas**

The x axis of each graph shows increasing ChIP intensity of the heterochromatin-associated histone mark H3K9me3 (ENCODE data, Broad Institute) (A and C) and increasing inverse median RepliSeq values representing later replication time (ENCODE data, University of Washington) (B and D). The y axis represents the mutation density per kb divided by the individual mean. Plotted are values for either eight aggregated repair wild-type (WT) cancers (solid blue line) or five aggregated  $XPC^{-/-}$  cancers (broken orange line) for 8 equally sized genomic bins covering approximately 2Gb of expressed genome and 1Gb of nonexpressed genome ( $\pm$ SD). Whereas mutation density correlates positively with increasing H3K9me3 and later replication time for expressed regions in repair WT cancers, these associations are diminished in  $XPC^{-/-}$  samples (A and B). In nonexpressed portions of the genome, regional disparities in mutation density are almost completely abolished in  $XPC^{-/-}$  samples (C and D), indicating loss in the absence of GG-NER. See Figure S1 for additional data with sparser active marks H3K27ac and H3K4me1 and Table S1 for additional information on tumor samples.

the untranscribed strand where TC-NER is absent. In the WT cSCC genome, the impact of DAR, measured as decreasing mutation frequency from lesions on the untranscribed strand with increasing expression, was substantial. For example, within nonexpressed portions of WT cSCC genomes (RPKM [reads assigned per kilobase of target per million mapped reads] < 0.01), mutation frequencies in regions with high H3K9me3 levels were approximately 3-fold greater than those with low H3K9me3 levels, consistent with recent estimates (Lawrence et al., 2013). In contrast, for highly expressed genes (e.g., RPKM = 400), this difference disappeared, with frequency of mutations originating on the untranscribed strand of all regions approaching that of DNA with low chromatin levels (gray dashed line at H3K9me3 = 1; Figure 2). This effect was also seen in three WT basal cell carcinomas (Figures 3F–3H). However, expression levels showed no effect on mutation frequencies in genomic regions with the lowest H3K9me3 levels.

### Proto-Oncogene Transcription Level Significantly Influences Mutation Frequency

We noted that the differences in mutation frequency associated with both transcription and chromatin state were of comparable magnitude to those caused by  $XPC$  loss of function. On average,  $XPC^{-/-}$  tumors harbor about a 5-fold greater mutation burden compared to WT cancers in transcribed regions (Table S4), illustrating how modest differences in mutation frequency can confer a large increase in cancer susceptibility. For reference, if five to six independent mutations were required for cSCC formation, a 5-fold increase in frequency of each mutation would raise the cancer rate by about 4,000-fold, approximately the observed increase in  $XPC$  patients (DiGiovanna and Kraemer, 2012; Lehmann et al., 2011). For genes in regions of the greatest H3K9me3 density in WT cancers, overall mutation density was lowered up to 4.7-fold as a result of higher expression, resulting from combined activities of GG-NER (in the form of DAR) and



**Figure 2. Domain-Associated Repair Restores Low Mutation Rate Only to Highly Transcribed Genes in Tightly Packaged DNA**

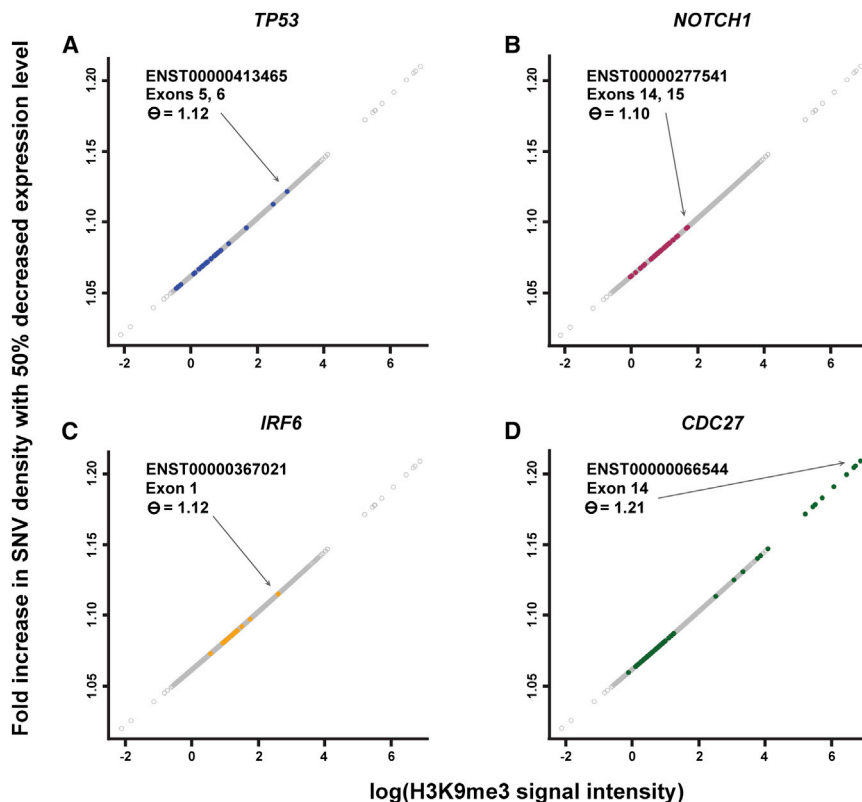
The x axis denotes increasing expression in NHEK, measured in RPKM (plotted on a log scale). On the y axis is the mutation density per kb. Values are plotted for three independent WT cSCCs (A–C) and three independent *XPC*<sup>−/−</sup> cSCCs (D–F). The plots show six different H3K9me3 densities representing different chromatin levels, represented by distinct colors, for the transcribed (solid line) and untranscribed (broken line) strands. The shaded area behind each line represents 95% confidence bands of the plotted line. In WT cancers, both strands show decreasing mutation density in tightly packaged DNA, illustrating robust domain-associated repair (DAR). DAR restores mutation rate in the most heterochromatic genomic regions to that of euchromatic regions, evidencing a dominant effect over chromatin state, but negligible additional impact in euchromatin (low H3K9me3). Even lower mutation density is seen from lesions on the transcribed strand, presumably representing TC-NER. In contrast, the *XPC*<sup>−/−</sup> cancers show an absence of DAR, represented by an absence of transcription-dependent repair on the untranscribed strand, but intact TC-NER. See Figure S2 for additional samples and Tables S2, S3, S4, S5, S6 for more detailed mutation density information.

TC-NER (Table S5). Furthermore, in WT tumors, we found a 3- to 4-fold reduction in mutation prevalence resulting from TC-NER of lesions on the transcribed strand (this reduction is 30-fold in *XPC*<sup>−/−</sup> tumors, possibly as the result of TC-NER acting in a compensatory role) (Table S6).

These observations led us to explore the possibility that natural variation in mRNA expression levels could exert an important influence on the mutation frequency of oncogenes located in tightly packaged DNA. In expression data obtained from the Genotype-Tissue Expression database (GTEx Consortium, 2013), we found that 72% of genes expressed in skin samples showed a 2-fold or greater variation in expression within a group of about 150 individuals. Similarly within ~660 lymphocytic cell lines in the 1000 Genomes Project (Lappalainen et al., 2013), ~80% of genes demonstrated at least a 2-fold difference. Thus, we assessed the potential impact of a 2-fold expression variance in our model. First, the variable  $\theta$  was modeled: the fold increase

in mutation frequency resulting from a 50% decrease in expression level, for a given H3K9me3 level, based on our data in WT cSCCs (Supplemental Experimental Procedures). We then examined  $\theta$  for 261 genes recently identified in a meta-analysis as recurrently mutated in human cancers (Lawrence et al., 2013). Genes were divided into 20,841 1 kb genomic segments for analysis (Figure 3; Table S7).

Our estimates of  $\theta$  predict that a 50% reduction in expression level would increase mutation frequency by 10%–20% or more for multiple exons in the SCC tumor suppressor genes *TP53*, *NOTCH1*, and *IRF6* (Agrawal et al., 2011; Wang et al., 2011). The exon with the highest  $\theta$  in this set, 1.21, belongs to *CDC27*, a gene demonstrating a 2% mutation frequency in head and neck SCCs and 4% in melanomas (Cerami et al., 2012), cancers whose tissues of origin depend on NER to control mutation frequency. The clinical impact of such effects in a population could be evaluated by determining both gene mutation



**Figure 3. Gene Expression Significantly Alters Tumor Suppressor Mutation Rates**

The x axis shows increasing H3K9me3 intensity, representing a more repressive chromatin state. The y axis shows the fold increase of the probability of a mutation, given a 50% decrease in expression level, referred to here as  $\theta$ . Plotted is  $\theta$  for 20,841 1 kb segments covering transcribed portions of 261 genes recently identified as recurrently mutated in human cancers. Highlighted are 1 kb fragments containing exons for the SCC tumor suppressors *TP53* (A), *NOTCH1* (B), and *IRF6* (C), as well as for the gene with exons of greatest average level of such mutation variance, *CDC27* (D), which has been shown to be mutated at about 4% in melanomas and 2% in head and neck SCCs. Exons with the highest variance and its corresponding  $\theta$  are indicated. See [Table S7](#) for  $\theta$  for all 20,841 1 kb segments.

density and expression level in a large series of tumors. We estimate that to have an 80% chance of detecting a 15% increase in mutation density in a gene within highly mutated cancers such as ours, a study would need to be powered with a minimum of approximately 600 samples at each transcript level ([Supplemental Experimental Procedures](#)).

## DISCUSSION

Our data show that in the germline absence of GG-NER, regional disparities in mutation density associated with chromatin-dense regions are virtually abolished in nonexpressed portions of cancer genomes, while the residual differences in mutation prevalence within expressed portions of the genome predominantly arise from disparities in TC-NER. Therefore, we establish that DNA repair efficiency is the main source of regional disparities in mutation density in cSCCs.

Unexpectedly, we also find that transcription and chromatin state do not influence mutation density independently. The decrease in mutation prevalence resulting from increasing levels of transcription was found to be correlated with H3K9me3 density and is in fact absent at the lowest H3K9me3 levels. A parsimonious interpretation of these data is that DAR acts solely and dominantly to restore GG-NER to expressed areas within tightly packaged DNA, perhaps as a result of transcriptional complexes increasing DNA accessibility to damage sensors such as XPC, rather than by a directed process such as TC-NER. This hypothesis agrees with previous observations that DAR proceeds even in the presence of RNA polymerase II inhibitors ([Noussipikel et al.,](#)

2006) and suggests a mechanism by which active genes maintain lower mutation frequencies, even in heterochromatic regions with reduced access to NER machinery. Highly expressed gene segments with greater H3K9me3 density have these marks concentrated in gene bodies, but not promoters. We therefore conclude that gene expression plays a critical role by relieving the structural

constraints imposed by densely packed chromatin on DNA repair machinery, rather than simply influencing mutation density alongside chromatin state either independently or in correlation. We further establish that the natural variation in transcription level of proto-oncogenes, between individuals, is sufficient to significantly influence their mutation rate. Our results therefore not only reveal a mechanistic basis for variable mutation density within cancer genomes but also show how to estimate proto-oncogene mutation rates of individuals in a population based on gene expression and chromatin state. These differences reveal an individual-specific modulator of risk for specific cancers, deserving further investigation in population-based studies.

## EXPERIMENTAL PROCEDURES

### Study Design, Tumor Samples, and DNA Sequencing

Tumor samples were obtained for five germline *XPC*<sup>-/-</sup> cSCCs following a UCSF Committee on Human Research protocol addressing isolation of these tumors during surgery. At least 5  $\mu$ g of DNA was collected from dissected tissue or peripheral blood and sequenced using the Illumina HiSeq2000 systems. More than 85% of targeted regions received 70-fold coverage at >90% of bases. Processing of raw sequencing data was performed using BWA ([Li et al., 2008](#)), samtools ([Li et al., 2009](#)), and GATK software packages (<http://www.broadinstitute.org/gatk/>). A detailed description of these methods is provided in [Supplemental Experimental Procedures](#).

### Processing of NHEK Chromatin and Replication Time Data

Hg19 chromatin immunoprecipitation (ChIP) sequencing signal intensity of H3k4me1, H3k9me3, and H3k27ac ([Ram et al., 2011](#)) and percentage-normalized signal Repli-seq data ([Hansen et al., 2010](#)) were obtained from ENCODE for normal human epidermal keratinocytes (NHEK) ([ENCODE Project](#)

Consortium, 2012). Signal intensities were averaged for 1 kb intervals across the genome. Genomic regions overlapping gaps (e.g., centromeres, telomeres) within the genomic assembly (Hg19 gap track from the UCSC Genome Browser) were excluded.

### Identification of Expressed and Nonexpressed Genomic Regions in NHEK

Whole-NHEK-cell long poly(A) and non-poly(A) RNA fastq files generated by CSHL were downloaded from ENCODE and aligned to Hg19 using STAR V2.3.0.e (Dobin et al., 2013) with default parameters except for allowing for a maximum of two mismatches. Nonexpressed genomic regions were then identified as 1 kb regions with zero reads mapped to that region and expressed genomic regions were identified as 1 kb regions with one or more read mapped to that region. Genomic regions encompassed by spliced reads (e.g., introns) were included as expressed genomic regions. As defined by these parameters, 2,065,687 1 kb regions were identified as expressed, and 1,032,437 1 kb regions were identified as nonexpressed.

### Processing of NHEK Expression Data

NHEK RPKM data for Hg19 Gencode v.10 annotated genes were downloaded from the Encode RNA Dashboard ([http://genome.crg.es/encode\\_RNA\\_dashboard/hg19/](http://genome.crg.es/encode_RNA_dashboard/hg19/)). RPKM values were assigned to 1 kb genomic intervals spanning the length of the entire gene, including introns.

### Modeling of Mutation Density versus Genomic Feature

For each 1 kb region of the genome, the numbers of mutations were calculated as well as the total number of “callable” nucleotide positions, i.e., that met the criteria for being called mutated (at least 8x coverage for the tumor and 4x coverage for the normal). Additionally, for the 1 kb bins within annotated gene regions (Hg19 Gencode v.10), the number of mutations on the transcribed and untranscribed strands was counted separately. The expressed/nonexpressed portions of the genome were divided into eight equal bins with respect to increasing intensities of individual histone density signals based on ChIP from ENCODE data. For replication timing, expressed and nonexpressed portions of the genome were divided into eight equal bins based on  $(1/\text{Repli-seq intensity}) \times 10^3$ . Mutations per megabase within each bin were calculated as the total number of mutations normalized by the total number of callable bases. To aggregate WT and *XPC*<sup>-/-</sup> samples, respectively, mutations per megabase for each sample were normalized by the mean mutation rate of each sample.

### Relationship of Mutation Density to Histone and Expression Levels

For analysis, we considered only regions with (1) at least 100 bp of callable positions, (2) nonmissing RPKM and histone values, and (3) annotated genes (Hg19 Gencode v.10). This resulted in a total of 1,160,378 1 kb regions. We fit a generalized linear model (GLM) separately to the transcribed and untranscribed counts for each sample in order to estimate the proportion of bases mutated as a function of the RPKM and histone values at a base. The probability of a mutation at a position *i*, denoted as  $p_i$ , was modeled as a function of its RPKM value ( $RPKM_i$ ) and histone value ( $Histone_i$ ). We used a standard GLM model for binomial data:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_{RPKM} (\log RPKM_i + e^{-5}) + \beta_{Histone} \log Histone_i + \beta_{int} (\log RPKM_i + e^{-5}) (\log Histone_i).$$

In order to work on the log scale and handle zero-valued data, RPKM values were shifted by  $\exp(-5) \approx 0.006$ , as noted in the above equation. The input to the model was the number of mutated positions and the total number of callable positions, per 1 kb region. We fit this model using the `glm` function in R, allowing for overdispersion in the data via the standard quasi-likelihood option for the binomial family. Confidence intervals for the fitted model were provided via the `predict` function in R.

### Relationship of Mutation Rate to Histone and Replication Time for Nonexpressed Regions

A similar strategy was performed for calculating the relationship between histone and mutation rate for nonexpressed regions. In this analysis, only regions

with at least 100 bp of callable positions and that were manually identified as nonexpressed in the sample as described above, were included, a number that varied per sample. The GLM model was the same as above, only without the terms involving RPKM:

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_{Histone} \log Histone_i.$$

For replication timing, the same model was used, only  $\log Histone_i$  was replaced by  $1/\text{Reptime}_i$ , where  $\text{Reptime}_i$  refers to Repli-seq intensity for the region. The reported p value for the significance of histone or replication in predicting mutation rate was the p value determined by testing the null hypothesis that  $\beta_{Histone} = 0$  or  $\beta_{Reptime} = 0$ , respectively.

### Fold Increase in Mutation Density Resulting from 50% Decrease in Expression

The overall mutation rate (resulting from lesions on the transcribed and untranscribed strands combined) was modeled as a function of histone and RPKM on the log scale using a GLM, in the same manner as described above. Using this model, we calculated the rate of change of the log-odds of mutation rate as a function of RPKM for a fixed level of histone. This was computed as the partial derivative of the log-odds of a mutation with respect to the log of RPKM, which simplifies to

$$d(\log - \text{odds}) = (\beta_{RPKM} + \beta_{int} \log(Histone_i)) d(\log RPKM).$$

Because the probability of a mutation is very small, the log-odds are approximately equivalent to the log of the mutation probability. Then for a change in RPKM of  $X_1$  to  $X_2$ , and histone levels held constant, we can approximate the fold change in the mutation rate as

$$\frac{p_1}{p_2} \approx \left( \frac{X_1}{X_2} \right)^{\beta_{RPKM} + \beta_{int} \log(Histone)}$$

### ACCESSION NUMBERS

The sequencing data have been deposited to dbGAP under accession number phs000830.v1.p1.

### SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, two figures, and seven tables and can be found with this article online at <http://dx.doi.org/10.1016/j.celrep.2014.10.031>.

### ACKNOWLEDGMENTS

We thank Peggy Tuttle and Maria Damen for help in sample procurement and Dr. Leslie Cope for manuscript comments. This work was supported by the Well Aging Research Center, Samsung Advanced Institute of Technology, under the auspices of Professor Sang Chul Park, the Dermatology Foundation, NIH, National Cancer Institute grants K08 CA169865 (R.J.C.) and U54 CA112970 and by the OHSU Knight Cancer Institute (J.W.G.). We appreciate assistance with artwork from Sarah Pyle and Eli Blair Media.

Received: June 20, 2014

Revised: August 26, 2014

Accepted: October 11, 2014

Published: November 13, 2014

### REFERENCES

Agrawal, N., Frederick, M.J., Pickering, C.R., Bettgowda, C., Chang, K., Li, R.J., Fakhry, C., Xie, T.-X., Zhang, J., Wang, J., et al. (2011). Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* 333, 1154–1157.



- Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.-L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421.
- Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837.
- Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404.
- Cleaver, J.E. (2005). Cancer in xeroderma pigmentosum and related disorders of DNA repair. *Nat. Rev. Cancer* 5, 564–573.
- Cleaver, J.E., Feeney, L., Tang, J.Y., and Tuttle, P. (2007). Xeroderma pigmentosum group C in an isolated region of Guatemala. *J. Invest. Dermatol.* 127, 493–496.
- DiGiovanna, J.J., and Kraemer, K.H. (2012). Shining a light on xeroderma pigmentosum. *J. Invest. Dermatol.* 132, 785–796.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.
- Durinck, S., Ho, C., Wang, N.J., Liao, W., Jakkula, L.R., Collisson, E.A., Pons, J., Chan, S.W., Lam, E.T., Chu, C., et al. (2011). Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.* 1, 137–143.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.
- Gospodinov, A., and Herceg, Z. (2013). Shaping chromatin for repair. *Mutat. Res.* 752, 45–60.
- GTEX Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580–585.
- Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* 107, 139–144.
- Hodgkinson, A., Chen, Y., and Eyre-Walker, A. (2012). The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* 33, 136–143.
- Jäger, N., Schlesner, M., Jones, D.T.W., Raffel, S., Mallm, J.-P., Junge, K.M., Weichenhan, D., Bauer, T., Ishaque, N., Kool, M., et al. (2013). Hypermutation of the inactive X chromosome is a frequent event in cancer. *Cell* 155, 567–581.
- Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., González-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al.; Geuvadis Consortium (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511.
- Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218.
- Lehmann, A.R., McGibbon, D., and Stefanini, M. (2011). Xeroderma pigmentosum. *Orphanet J. Rare Dis.* 6, 70.
- Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
- Liu, L., De, S., and Michor, F. (2013). DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.* 4, 1502.
- Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560.
- Nouspikel, T., and Hanawalt, P.C. (2000). Terminally differentiated human neurons repair transcribed genes but display attenuated global DNA repair and modulation of repair gene expression. *Mol. Cell. Biol.* 20, 1562–1570.
- Nouspikel, T.P., Hyka-Nouspikel, N., and Hanawalt, P.C. (2006). Transcription domain-associated repair in human cells. *Mol. Cell. Biol.* 26, 8722–8730.
- Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.-L., Ordóñez, G.R., Bignell, G.R., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463, 191–196.
- Polak, P., Lawrence, M.S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R.E., Garraway, L.A., Mirkin, S., Getz, G., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2014). Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* 32, 71–75.
- Ram, O., Goren, A., Amit, I., Shoshitaishvili, N., Yosef, N., Ernst, J., Kellis, M., Gymer, M., Issner, R., Coyne, M., et al. (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* 147, 1628–1639.
- Schuster-Böckler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 488, 504–507.
- Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M., and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. *Nat. Genet.* 41, 393–395.
- van Hoffen, A., Venema, J., Meschini, R., van Zeeland, A.A., and Mullenders, L.H. (1995). Transcription-coupled repair removes both cyclobutane pyrimidine dimers and 6-4 photoproducts with equal efficiency and in a sequential way from transcribed DNA in xeroderma pigmentosum group C fibroblasts. *EMBO J.* 14, 360–367.
- Wang, N.J., Sanborn, Z., Arnett, K.L., Bayston, L.J., Liao, W., Proby, C.M., Leigh, I.M., Collisson, E.A., Gordon, P.B., Jakkula, L., et al. (2011). Loss-of-function mutations in Notch receptors in cutaneous and lung squamous cell carcinoma. *Proc. Natl. Acad. Sci. USA* 108, 17761–17766.