Determinants of the local mutation rate variability along the genome

Joan Frigola Rissech

TESI DOCTORAL UPF / 2019

DIRECTORS DE LA TESI

Dra. Núria Lopez-Bigas & Dr. Abel González-Pérez

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Universitat Pompeu Fabra
Barcelona

Acknowledgments

There is a huge number of people that helped me during these last years. Among them, my mentors and supervisors Nuria and Abel deserve to be mentioned in the first place. Thank you for the guidance, supervision, enthusiasm and for showing me how fun learning can be. Also, thanks for giving me an opportunity. To Nuria for letting me join your laboratory as a master student (regardless of the pathetic interview I did, which I hope has been forgotten by now), and for allowing me to stay as a PhD student afterwards. To Abel, for being the first person to talk to me about doing a PhD. I never thought this was something I could do until that moment.

Of course, to the rest of the BBG family as well, starting with the rest of the PhD students. Thanks to Michi and Carlota for making me feel like a member of the lab from the very first day, and introducing me to your volleyball and football teams. To Oriol, our morning discussions where always the best moment of the day. And for our afternoon breaks as well: complaining about the world with you has been a great hobby. To Pepe, for all the work(out) we did together. Eventually I'll make it to the lab earlier than you. To Inés, we all need a mum that takes care of us. To Hania, I'm already looking forward to your next trip to Poland. By the way, my favorite sweets are the chocolate-covered marshmallows. To Winona, our party soul (does this exist in English?), and to Claudia, I've never seen you without a smile on your face. Also, to the postdocs, Sabari, David, Santi, Fran, Ferran for taking care of us, being a deep source of wisdom in the lab and an even deeper source of fun outside it. To the wet lab people: Erika, Meritxell, Victor, Nuria, Mari. I still don't see the bands in your gels (and I suspect you don't either), but you are amazing. To the software engineers: Jordi, Iker, Loris, who make us of all their huge computational knowledge and endless patience to save us. And to Martina. I would have probably been fired from the PhD programme a few times without you. I really hope I will keep seeing you all even if we don’t work together anymore.

Also, many people outside the laboratory has also helped me during this PhD. Especially my family: parents, sister and also Mireia. With you, problems are insignificant.

Finally, this thesis wouldn’t have been possible without the effort of many research groups and international consortia that have generated and shared their data with the rest of the scientific community. Therefore, I would like to thank all the people involved in the generation of the data we have analyzed: the TCGA consortium, Roadmap epigenomics project, the The International Biallelic Mismatch Repair Consortium, Dr. Sancar and colleagues, Dr. Wyrick and colleagues, and finally to the patients, who generously allowed their data to be used in research.

Abstract

The rate at which mutations accumulate along the genome is not uniform but influenced by factors such as chromatin compactness, replication time or transcription. These factors and others shaped the mutation rate along the genome at the large scale. In recent years, local mutational asymmetries spanning just a few base pairs have also been identified. This thesis focuses on the study of the mechanisms that explain the occurrence of two of these local mutational asymmetries in somatic cells. First, we describe a reduction in the number of exonic somatic mutations derived from DNA polymerase mismatches, which we attribute to a higher efficacy of the mismatch repair mechanism in these locations. Second we study the UV induced DNA damage formation and repair at transcription factor binding sites and assess the relative contribution of these two factors to the unexpected number of mutations of these areas across transcription factors families. The study of these local mutation rate variations helps us understand how DNA repair mechanisms work and interact with other cell mechanisms, and illustrate the difficulty of properly modeling the mutation rate, an important procedure in many cancer genomics and evolutionary studies.

Resum

La velocitat a la que les mutacions s’acumulen al llarg del genoma no és uniforme sinó que depèn de diversos factors. Alguns dels més coneguts són l’empaquetament de la cromatina, el moment de replicació o la transcripció. Aquests factors juntament amb d’altres creen variacions mutacionals que abarquen grans àrees del genoma. En els últims anys també s’ha identificat variabilitat en el ritme en que s’acumulen les mutacions a escala molt més petita, de només poques bases. Aquesta tesi es centra en l'estudi dels mecanismes que expliquen la presència de dues d'aquestes variacions locals en el ritme en que les mutacions tenen lloc en cèl·lules somàtiques. Primer, hem descrit una reducció en el número de mutacions somàtiques en els exons causades per errors de la ADN polimerasa, que hem atribuït a una major activitat del mecanisme encarregat aquest tipus d'errors en els exons. En segon lloc, hem estudiat com les lesions en l'ADN causades per la llum ultraviolada es generen i són reparades als llocs d'unió dels factors de transcripció i hem determinat fins a quin punt cada un d'aquests processos permeten explicar l'inesperat número de mutacions en aquestes regions. L'estudi d'aquestes variacions locals en la velocitat a la que s'acumulen mutacions al llarg del genoma ens ajuda a entendre com funcionen els mecanismes de reparació de l'ADN i com interaccionen amb altres mecanismes cel·lulars. A més, es posa de manifest la dificultat de modelar correctament aquest procés, un procediment central en molts estudis evolutius i de genòmica del càncer.

1. Introduction

Introduction

In 2013, Bianconi et al. estimated that, on average, a human body is composed of around 37 trillion (3.7 x 10^{13}) cells[1]. Despite the overwhelmingness of this number, our bodies remain in an exquisitely regulated state of homeostasis where cells replicate in an organized manner, perform the physiological functions they are intended to, and even die if required.

However, upon internal or external insults this state of homeostasis may be disrupted, potentially leading to disease. Some cells, for example, may start growing uncontrollably. These abnormal growths are called neoplasms or tumors when they form a mass[2], and can appear in virtually all human tissues.

Usually, neoplasms are divided into three categories based on their ability to spread and colonize other parts of the body. Thus, benign neoplasms lack spreading capacity and in many cases require no treatment, while neoplastic cells that have acquired the capability to invade surrounding tissues, and even enter the circulatory or lymphatic system to migrate and settle in distal organs are called malignant and are given the generic name of cancer. Finally, some non malignant neoplasms show potential to become so and, therefore, are labeled as pre-malignant or pre-cancerous[3].

The ability of malignant cells to travel and grow in locations different than the primary tumor site confers cancer most of its harmful potential. These new growths are called metastases and are responsible for the majority of the cancer associated deaths[4], since may condition the correct functioning of vital organs[5] [6] or weaken the patient’s immune system predisposing to severe infections[7]. Hence, due to its prognosis and high incidence, cancer is a major concern worldwide and poses a significant health risk for humans.

Cancer incidence

According to the World Health Organization (WHO), cancer is the second leading cause of death globally. Only in 2018, 18 million cases of cancer were diagnosed worldwide and 9.6 million cancer associated deaths were reported, most of which are related to bowel, lung or breast malignancies. Moreover, according to recent estimates, these numbers are expected to increase over the years: it has been predicted that, in 2040, 27.5 million new cancer will be developed. This represents an increase of over 60% compared to 2018, even though it could be partially attributed to the life expectancy increase and the reduction of mortality associated to other diseases.

Fortunately, advances both at diagnosis and treatment level are being made and the survival of this disease is incre of the main attributes of cancer cells is unrestrained growth. However, the carcinogenic phenotype is more complex and there are several complementary features cells acquire in tumorigenesis. These features are known as the hallmarks of cancer and were described in a landmark paper by Hanahan and Weinberg in the year 2000[8] and reviewed and expanded by the same authors in 2011[9] (Fig. 1.). These hallmarks of cancer are the following:

Sustainment of proliferative signaling: cancer cells may disrupt the growth signaling system in order to permanently receive proliferative stimuli, needed to maintain their abnormal growth rate. This can be achieved by overexpressing specific membrane receptors[10], acquiring the capability to secrete proliferative signals[11] or through many other alternative mechanisms.

Growth suppression evasion: complementary to the previous hallmark, cancer cells acquire the ability to ignore regulatory signals that would limit their growth. Key genes in these pathways act as tumor suppressors and have been found to be inactivated in many tumors[12].

Cell death resistance: upon damage or stress, cells are able to trigger a cascade of biochemical events that lead to their own death. This highly regulated process is called apoptosis and is a form of programmed cell death that may be activated under different circumstances and used to eliminate potentially carcinogenic cells[13]. However, in some occasions these cells find ways to bypass the control of the apoptotic programme and are able to develop regardless of this mechanism[14].

Replicative immortality: healthy cells are only able to replicate a limited number of times, which is determined among others by the telomeres length. However cancer cells escape this constraint and replicate perpetually. This capability can be achieved through upregulation of the telomerases, which are ribonucleoproteins responsible for telomere maintenance[15].

Angiogenesis induction: tumors, like any other human tissue, need a supply of oxygen and nutrients to survive and develop. These resources are present in the blood and, therefore, carcinogenic cells need to be able to stimulate the formation of blood vessels in order to irrigate the tumor mass as it expands.

This process is known as angiogenesis and is regulated by a collection of signaling proteins, some of which have been found to be permanently upregulated in tumors[16].

Invasion and metastasis activation: as mentioned in the first section of the introduction, carcinogenic cells are characterized by their potential to invade surrounding tissues and to migrate to distal locations and start growing there. One of the most common ways to achieve this is by losing E-cadherin, a molecule that participates in the creation of cell to cell joints[17]. Underexpression of this gene releases tumor cells from their spatial constraints, allowing them to spread.

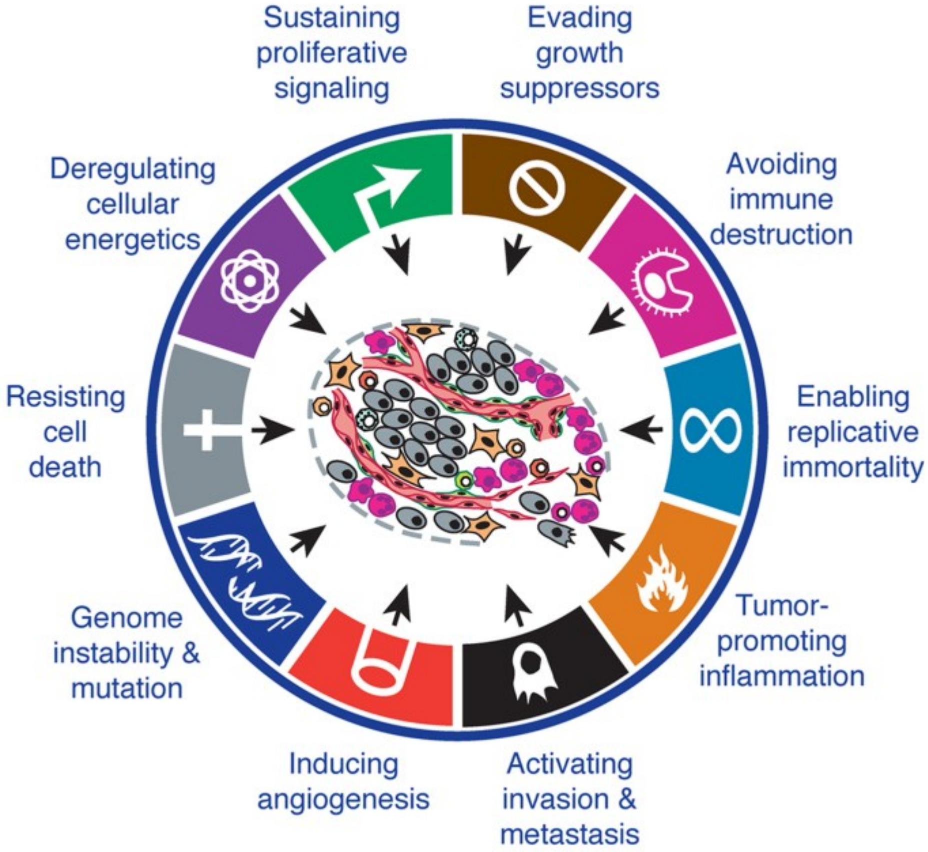

Sustaining proliferative signaling
Evading growth suppressors
Deregulating cellular energetics
Avoiding immune destruction
Resisting cell death
Enabling replicative immortality
Genome instability & mutation
Tumor-promoting inflammation
Inducing angiogenesis
Activating invasion & metastasis

Figure 1. The hallmarks of cancer (Hanahan & Weinberg, 2011).

In combination, these capabilities or hallmarks define what tumor cells are and how they behave. However, how are they acquired? What are the events that trigger such behavioural changes? How do healthy cells acquire these hallmarks of cancer?

Driver genomic events in cancer

Most cancers result from genetic changes that confer cells any of the traits previously described[18]. These changes are given the generic name of mutations and may appear at any time during human development and in any tissue. Thus, when they occur at germ cells, they are known as germline and might be inherited by the offspring[19]. However, in the context of cancer, mutations usually have a somatic origin, meaning that they have been originated during cell division in cells other than the germline and therefore will only be present in tissues derived from the mutated cell but won't be inherited by the descendants of the individual who acquired them[20].

Mutations can also be classified by the size of the genomic region they affect. When they alter the chromosomal structure, they are called structural variants. This includes deletions, duplications, insertions, inversions and translocations[21] and, despite being relevant for tumorigenesis[22], are out of the scope of this work and will not be further discussed. On the other hand, single nucleotide variants (SNVs) are substitutions affecting one single DNA position and from now on I will refer exclusively to them when using the term mutations.

The number of mutations per tumor largely depends on the tumor type, and ranges from an average of 0.05 mutations per megabase (Mb) in pilocytic astrocytoma to around 10 mutations/Mb in melanoma[23]. Considering that the human genome has 3200 Mb, this means from 160 to 32000 mutations per tumor, on average. However, most of these mutations do not confer tumorigenic capabilities and are known as passengers. Only a small subset may affect genes or regulatory regions modifying cellular activity in a specific way that could lead to the acquisition of any of the hallmarks of cancer. These mutations are known as drivers, and the genomic elements susceptible of acquiring them are known as driver elements[18]. Interestingly, it has been reported that tumors carry around 4 driver mutations on average. However, this value has been observed to vary across tumor types, ranging from 1 in testicular or thyroid cancers to more than 10 in endometrial or colorectal [24] [25].

During the last few years, the identification of cancer driver genes has been the subject of intense research by the cancer genomics community, believing that a better understanding of the driver alterations carried by a tumor could help to better assess patients prognosis and treatment. Thus, since the first cancer exomes where sequenced, a lot of effort has been put into trying to pinpoint these few cancer driver genes and mutations actually responsible for tumor initiation and growth among all tumor mutations. To do so, some concepts from the evolutionary field have been borrowed: in evolutionary biology, fitness is a measure of the reproductive success of an individual or group of individuals with a specific genotype or phenotype[26]. This same concept can be applied to cells within a tissue, given that cells acquire mutations and therefore genotypic and probably also phenotypic differences that could lead to differential proliferative capabilities are expected to be found. Thus, specific mutations may negatively impact their fitness, leading to reduced growth, apoptosis or senescence. However, others may provide them with a selective advantage over the surrounding cell population, potentially initiating a proliferative process that could lead to tumor formation.

This process involving a single cell overgrowing the rest of the population leading to tumor formation is called clonal expansion[27] and results in the mutations found in this founder cell being present in all the cells of the grown tumor. However, later appearing mutations will only be shared by a fraction of tumor cells, whose size will depend on the phase of tumor development these mutations occurred, but more importantly on the fitness increase these mutations confer to the cells, if any. When this happens, sub-populations of cells with different genotypic and phenotypic features may coexist within a tumor. These sub-populations are known as clones[27]. Moreover, clones with higher fitness may outcompete others. Therefore, tumors clonal structure and consequently tumor properties may evolve over time (Fig. 2.). This heterogeneity has direct implications in tumor studies, since tumor samples will most likely be heterogeneous and studying them as a hole does not allow to capture the heterogeneity of its forming cell populations. Also, it plays a role in treatment resistance, since antitumoral drugs may efficiently target and eliminate specific clones of a tumor but may be ineffective against others, leading to a change in the clonal structure of the tumor rather than its eradication[28].

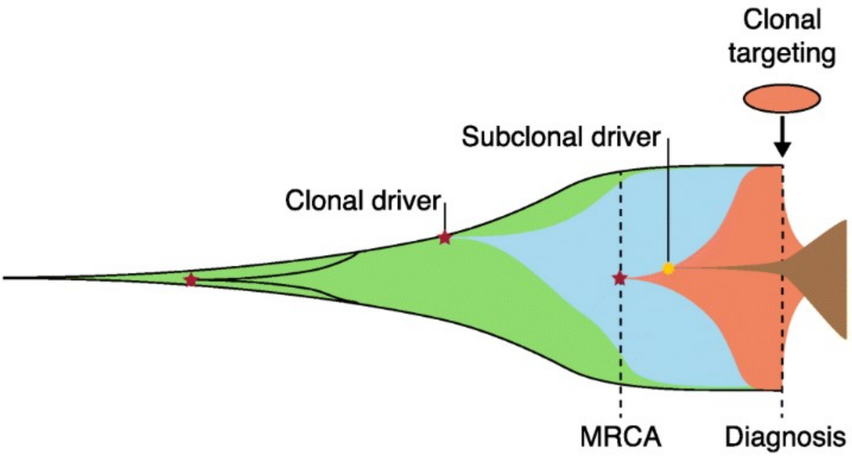


Figure 2. Evolution of a tumor cell populations over time. Colors belong to specific cell populations. Red stars and yellow dots represent clonal and subclonal driver mutations respectively. Dotted lines show the origin of the most recent common ancestor (MRCA) and the moment of diagnosis. Treatment administered after diagnosis targets a specific cell population, leading to its elimination and the posterior proliferation of the resistant one (Fittal & Van Loo, 2019).

Interestingly though, this evolutionary process cells undergo during tumor development leaves signals of positive selection in the pattern of somatic mutations observed in genes. These signals can be identified and used to differentiate these driver elements from the passenger ones, providing information about which specific alterations among all carried by tumors actually contributed to the disease.

Positive selection identification and cancer driver genes mode of action

In the context of cancer driver genes identification, looking for signals of positive selection can be understood as searching for patterns of mutations accumulation not expected to be observed by chance.

Multiple signals of positive selection may be studied but probably the most intuitive one is recurrent mutations accumulation in specific genes[29]. Let's imagine that a group of tumor samples from a specific cohort are biopsied, sequenced and their mutations called. Each tumor will carry a considerable number of mutations and, if interrogated individually with no previous knowledge, it won't be possible to pinpoint those responsible for tumor development. However, more information can be gathered through samples comparison, since observing specific genes mutated across a majority of tumor samples may be indicative of a putative driver effect. In other words, we would expect driver genes to appear recurrently mutated across patients, not because they are more likely to be mutated than other genes but because when this happens, cells develop into tumors and, therefore, are biopsied, sequenced and included in the cohort. However, establishing a threshold of recurrence to decide whether a gene is recurrently mutated in a cohort or not is a non trivial problem. Moreover, not all genes are equally likely to be mutated[30], even in a healthy setting, and therefore it is not equally surprising to find them mutated in a given number of patients. Long genes, for example, are more likely to be mutated just because there are more posities where mutations can happen. Therefore, precise driver identification algorithms should be able to estimate how likely is for a gene to be mutated under normal conditions and then, examine the excess of mutations not explained by any other factors than positive selection.

An additional signal of positive selection exploited for driver genes identification is the accumulation of mutations with high functional impact[31] [32]. In coding regions of the genome, for example, not all possible mutations are expected to have the same impact on the structure or function of the resulting protein. If no positive selection took place, the functional impact of all mutations across genes in a specific cancer cohort would follow a distribution that could be modeled and used as background. Then, genes with a bias towards accumulating mutations with higher functional impact than expected would be considered putative drivers. Some of these methods have been adapted to look for functional impact biases outside genes coding region[32]. This has been achieved by changing the way the functional impact is measured: in a non transcribed region like a promoter, it is meaningless to try to assess how a mutation would affect a hypothetical protein. Instead, it is relevant to determine its effect on gene regulation or protein binding affinity, for example.

Other driver identification methods do not focus on the total number of mutations genes accumulate but on how they are distributed along them. These are the so called clustering methods and seek to find positions within genes (or other regulatory elements) where mutations recurrently accumulate across samples[33]. These methods operate under the assumption that mutations in different parts of genes may alter the encoded proteins activity in different ways (or not at all), and because of this, only mutations in specific regions should be able to modify their behavior in a way that would stimulate cell proliferation or any other hallmark of cancer. This is applied mainly to genes that stimulate tumorigenesis by modifying their activity (called oncogenes), but not so much for those genes that become drivers upon inactivation (called tumor suppressors). This is so because there are many different mutations that could inactivate a gene, but only a few that could alter it's activity in a specific manner.

Finally, not all cancer driver genes are expected to present all signals of positive selection. This will depend on the way each gene needs to be altered in order to behave in a pro-tumorigenic manner. Therefore, different driver identification methods are complementary and have been used in combination in order to obtain a collection of cancer driver genes as complete as possible[34].

All these methods, though, incorporate information about how mutations accumulate in a neutral context and use it to create a mutational background used as reference in order to identify mutational biases potentially caused by positive selection. The creation of an accurate mutational background is a challenge by itself, since mutations do not occur in a homogeneous manner along the genome, and even more, the way mutations are differentially distributed along the genome varies from sample to sample as well[29].

DNA damage and repair

Mutations are a consequence of lesions inflicted to the DNA by a variety of extrinsic and intrinsic damaging agents cells are exposed to, or due to polymerase errors during replication, and are generated despite the activity of the collection of repair mechanisms cells possess, either because a fraction of these lesions are left unrepaired or because of some inaccuracy or malfunction of the repair process itself[35] (Fig. 3.).

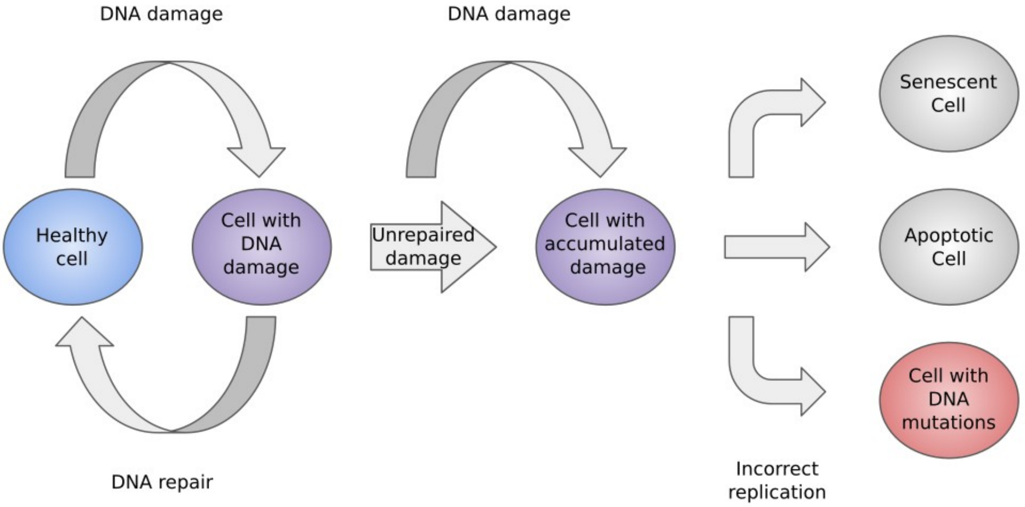


Figure 3. Schematic representation of the possible outcomes of DNA damage. Unrepaired DNA damage accumulates over time and may lead to apoptosis, senescence or to the generation of somatic mutations.

Each specific combination of a DNA damaging agent with potential to be mutagenic and the repair mechanisms (either proficient or deficient) responsible for dealing with its associated lesions leads to a mutational process and may cause different types of mutations[36]. Thus, based on the mutational processes to which a cell has been exposed to and the intensity of this exposure, the amount and type of mutations and also the way they are distributed along the genome may change.

Hence, in order to properly model how mutations occur along the genome, it is needed to understand the mutational processes to which genomes have been exposed and, therefore, comprehending how each specific damaging agent and repair mechanism behave.

DNA lesions

The repertoire of DNA damaging agents is broad (Fig. 4.). Some of them are exogenous to the cell, like UV[37] or ionizing radiation[38], while others are originated during normal intracellular functioning and are called endogenous. Well known examples of these are byproducts of oxygen metabolism[39]. Additionally, different DNA damaging agents may inflict different types lesions with specific cytotoxic potential.

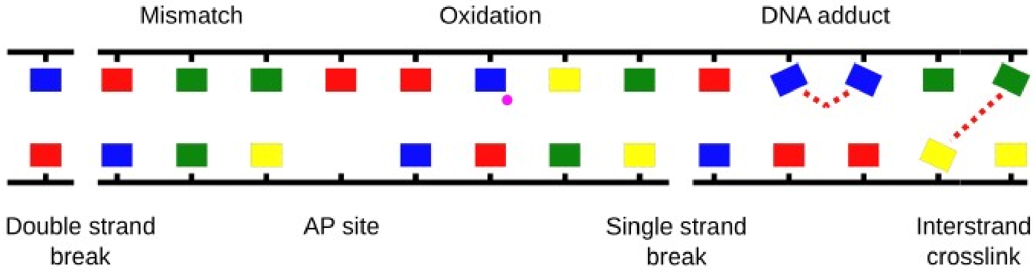


Figure 4. Segment of DNA with different types of lesions.

Single base alterations

Upon certain insults, the nucleotides forming the DNA sequence may be damaged. For example, abasic sites appear when a nucleotdie’s nitrogen base is removed, leaving the backbone formed by the sugar molecule and the phosphate group undamaged but losing the sequence information[40]. Depending on whether the excised base was a purine or a pyrimidine, abasic sites may be classified as apurinic or apyrimidinic. Because of this, they are often called apurinic/apyrimidinic sites or AP sites.

Bonds between the sugar subunit and the phosphate group are more vulnerable in abasic site positions and, if disrupted, may generate other types of DNA damage such as single strand breaks, which are reviewed below. Moreover, AP sites may block transcription and replication, which has cytotoxic effects[41].

Apart from AP sites, nucleotides may suffer other types of less drastic alterations such as alkylation (gain of an alkyl group)[42], oxidation (gain of an oxygen atom)[43] or deamination (loss of an amine group)[44], which are potentially mutagenic as well.

Bulky lesions

Some chemical compounds have the ability to covalently bind to the DNA, distorting the double helix structure by generating prominences[45] [46]. Bulky lesions may also be generated upon formation of abnormal covalent bonds between nearby nucleotides, altering the normal base-pairing of a given region, in a process that may be mediated by UV radiation[37]. Due to their voluminous nature, these bulky lesions have the potential to impair replication.

Replication errors

One of the features of the canonical DNA double helix is the complementarity of its two strands, allowing both of them to be used as template during replication. However, replication is not an error-free process and erroneous nucleotides may be introduced by DNA polymerases. Mismatches occur when a nucleotide is miss-incorporated, and therefore the complementarity in that specific position is lost. Moreover, one or more bases may be gained or lost during replication. These are known as insertions and deletions respectively[47].

Finally, base analogs are molecules structurally very similar to the four bases forming the canonical DNA[48]. During replication, they may be mistakenly used in the synthesis of the newly created DNA strand, becoming a potential source of errors if used as template in posterior rounds of DNA replication.

DNA strand breaks

Some damaging agents including byproducts of cell's metabolism or radiation are able not only to modify DNA nucleotides or promote the formation of abnormal bonds, but also to physically break one[49] or both[50] DNA strands.

In the case of single strand breaks, they may also cause double strand breaks if the replication machinery collapses when attempting to replicate through them[51].

Interstrand crosslinks

Abnormal covalent bonds may be formed between complementary strands of the DNA double helix[52]. This process may be mediated by either endogenous or exogenous agents and poses a serious difficulty to DNA replication since it impairs the separation of the two strands of the DNA double-helix[53].

Repair mechanisms

DNA lesions and mismatches are potentially cytotoxic and could induce cell death or senescence, but also mutagenesis. Therefore, in order to secure their normal performance and preserve genome integrity, cells have acquired a collection of repair mechanisms specialized in repairing each type of lesions[54].

Mismatch repair

DNA mismatch repair (MMR) is the pathway responsible for the repair of mismatches introduced during replication. This process begins when a mismatch is recognized by a dimeric protein formed by MSH2 and MSH3 or MSH6. In an ATP dependent manner, this dimer undergoes a conformational change that allows the recruitment of a second dimer comprising MLH1 and PMS2. The resulting complex is then able to slide along the DNA double strand and stimulate the activity of the Exo1 exonuclease that degrades a segment of the newly synthesized DNA strand, including the mismatch. Finally, the gap will be re-synthesized by the DNA Polymerase delta and sealed by DNA ligase I[55].

Given that this process of repair specializes in mispaired bases, being able to distinguish and target specifically the newly synthesized strand of DNA is of great importance, as it is the one carrying the misincorporated nucleotide. To do so, mismatch repair machinery recognizes the Okazaki fragments present in the lagging strand or the 3’ terminus of the leading strand[56].

Deficiencies in the mismatch repair process have been associated with the emergence of colorectal and other gastrointestinal cancers, as well as with other types of cancer such as endometrial, breast, bladder, thyroid or prostate adenocarcinomas.

Lynch syndrome is a disease caused by an inherited mismatch repair deficiency and is linked to increased susceptibility to colorectal cancer[57]. The patients tend to accumulate mutations all over the genome, and mainly at repetitive regions known as microsatellites. This phenomenon is known as microsatellite instability and is one of the hallmarks of mismatch repair deficiency. Also, loss of MMR reduces the chances of a cell to enter apoptosis when accumulating mismatches, conferring an enhanced survival capability and therefore a selective advantage[58] [59]. Additionally, mismatch repair deficiency and microsatellite instability can also be acquired somatically. For example, somatic mutations in MSH2 gene of non-hereditary endometrioid carcinomas have been described[60]. However, the more widespread mechanism of non-hereditary loss of mismatch repair capacity is the repression of MLH1 gene driven the hypermethylation of its promoter[61].

Interestingly, loss of mismatch repair activity has also been associated to chemotherapy resistance[62]. The proposed mechanistic explanation has to do with the cross-talk between this pathway and the programmed cell death process: apart from dealing with DNA lesions, mismatch repair also participates in the elimination of those cells whose damage burden is too high by triggering apoptosis. However, mismatch repair deficient tumor cells fail to enter apoptosis even when have received extensive chemotherapy induced DNA damage. On the contrary though, microsatellite instability has been shown to be a predictor of good response to anti-PD1 immunotherapy, probably because tumors with MSI present a greater number of neoantigens due to their increased mutation burden[63].

Nucleotide excision repair

Nucleotide excision repair (NER) is a mechanism specialized in repairing bulky DNA adducts like those induced by UV light or cisplatin. This pathway can be divided into two sub pathways that differ in the way they recognize the adducts. On one hand, there is a transcription dependent damage recognition mechanism carried out by the RNA polymerase itself. Consequently, only the template strand of the transcribed regions of the genome benefit from it. Because of this, it is known as transcription coupled repair (TC-NER). On the other, there is a second damage recognition mechanism with the ability to sense DNA adducts all along the genome. It requires the involvement of the DDB and XPC-Rad23B complexes and is called global genomic repair (GG-NER)[64].

Once the damage has been recognized, the two NER pathways converge: TFIIH is recruited to the adduct site and acts as a helicase, unwinding the DNA. Next, XPG and XPF-ERCC1 create incisions upstream and downstream the DNA adduct, releasing a segment of single-stranded DNA of 27-29 nucleotides including the lesion. Finally, the gap is resynthesized by a DNA polymerase (delta, epsilon and/or kappa) and the nicks are sealed by DNA ligase I or III[64].

As with MMR, congenital loss of NER activity causes severe disease: xeroderma pigmentosum is an autosomal recessive hereditary disease whose patients display extreme sensitivity to UV light, freckles, increased skin pigmentation and high risk of developing skin cancer. Mutations in up to 8 different genes from the NER pathway have been associated with this disease. Seven of these genes belong to the XP family (XPA to XPG), being involved in the GG-NER damage sensing phase and in the common part of the NER pathway and the eighth one being a DNA polymerase (POLH)[65].

Cockayne syndrome is another NER associated disease. In contrast with Xeroderma pigmentosum, genes involved in Cockayne syndrome (CSA and CSB), are specific to the TC-NER sub pathway. Its symptoms are, among others, microcephaly, short stature and delayed development. Interestingly, however, Cokayne syndrome is not linked to higher risk of developing cancer[66].

Base excision repair

Base excision repair (BER) is a mechanism specialized in repairing small lesions affecting single DNA bases induced by processes such as oxidation, alkylation or deamination. Its first step involves a DNA glycosylase recognizing and excising the nitrogen base at the damaged position. Glycosylases are divided into monofunctional and bifunctional and specialize in repairing different types of lesions by triggering distinct subpathways of BER. However, they all initiate the repair process by generating an AP site after excising the damaged base. This site is then processed by an AP endonuclease called AP1 in a process that involves the cleavage of the DNA backbone, generating a one nucleotide single strand gap flanked by a 3’-hydroxyl in one end and a 5’-deoxyribosephosphate (5’-dRP) in the other. Then, DNA polymerase β processes these ends by removing the 5’-dRP, creating a situation that allows the synthesis of the missing correct base by polymerase β itself. Also, there are other glycosilases that generate one nucleotide gaps to the DNA without the need of endonucleases. These gaps are flanked by two phosphate ends that also need to be processed to accommodate the activity of DNA polymerase β. This involves the removel of the 3' phosphate end by PNKP, creating a substrate that allows the DNA polymerase β to insert the correct missing base. Finally, a protein complex called XRCC1-Lig IIIα, which involves a DNA ligase, seals the nick[67].

In addition, there is another sub-pathway of base excision repair named long-patch BER that involves the excision of small DNA segments of 2 to 10bp[67].

The loss of any of the essential components of BER is embryonic lethal. Therefore no human diseases are associated with this phenotype[68]. However, mutations in some DNA glycosylases such as MUTYH appear not to be lethal, probably due to the partial functional redundancy with other glycosylases, and have been associated with higher colon cancer susceptibility[69].

Homology directed repair

Homology directed repair is a mechanism used by cells to repair DNA breaks affecting both strands of the double helix (DSB). These lesions may be caused by ionizing radiation[50], UV light[37] or genotoxic chemical compounds[70] among others, and are potentially very genotoxic. Homologous recombination allows cells to repair these DSB in an error free manner by using the undamaged homologous chromosome as a template[71].

Homology directed repair requires the presence of single stranded DNA ends around the DSB in order to proceed. These ends are achieved through a process called resection that consists in a 5’ to 3’ degradation of the broken DNA, generating a single stranded 3’ ends per side. This process is started by the MRN complex, which is able to recognize and bind to DSBs to initiate the repair process while pausing the cell cycle to avoid interference from the replication machinery. Then, MRN recruites the CtIP nuclease, responsible for initiating the resection process. Next, EXO1 and DNA2, together with several helicases from the RecQ family take over and complete the resection step, generating two 3’ overhanging single strands of DNA. These overhanging regions are covered by the RPA complex in order to avoid self-annealing or degradation. Then, with the help of Rad51 and other proteins such as BRCA1 and BRCA2, the process of searching and invading the homologous region, either from the sister chromatid during mitosis or from the homologous chromosome during meiosis, takes place. In either case, the homologous region is used as a template by a DNA polymerase, that will extend the end of the invading 3' strand. This process creates a cross-shaped structure called Holliday junction that encompasses four double strands of DNA from both the damaged double helix and the one used as a template[72].

From this point, several outcomes are possible: when the invading DNA strand is released after being extended and anneals back to the other 3’ overhanging strand from the damaged chromosome generated during the resection process, it can be used as a template to synthesize the resulting single strand gap and ligate the ends. This is known as the SDSA sub pathway. However, there is a second sub pathway called DSBR that is triggered when this other 3’ overhanging end that did not participate in the strand invasion also forms a Holliday junction with its homologue region, resulting in a double Holliday junction structure that needs to be resolved with the help of specific enzymes called nicking endonucleases. To do so, these enzymes perform single strand cuts in the DNA forming the Holliday junctions, and depending on the way these cuts are performed, crossover between the two homologous chromatids may happen or not[72].

Several diseases have been associated with deficient homology directed repair. Werner’s syndrome, Rothmund-Thomson syndrome or Bloom’s syndrome are diseases caused by mutations in different helicases from the RecQ family (more specifically WRN, RECQL4 and BLM respectively), and are characterized by inducing a progeroid phenotype, meaning that promote the early onset of age associated diseases, as well as increased cancer susceptibility[73].

Homology directed repair is also affected in individuals carrying mutations in either BRCA1 or BRCA2 genes[74]. These mutations are usually germline and therefore inheritable, and are associated with higher risk of breast and ovarian cancer[75].

Non-homologous end joining

An alternative mechanism for repairing DSB is the non-homologous end joining (NHEJ) pathway. This mechanism doesn’t rely on the presence of homologous DNA regions to be used as a template and, therefore, is error-prone[76].

Similarly to the homology directed repair, MRN complex is bound to the DSB soon after the lesion occurs. In NHEJ, however, a protein heterodimer called Ku is also recruited and will serve as an anchor point for other proteins involved in the repair. The first is a DNA kinase known as DNA-PKcs, that will phosphorylate itself and also the Artemis complex. Artemis has endo and exonuclease activity and, if needed, processes the broken DNA ends in order to allow their ligation, conducted by a complex involving XLF-XRCC4 and DNA ligase IV[77].

There is a crucial step that determines whether a DSB is repaired by homologous recombination or NHEJ: the 5' end resection. Since NHEJ doesn't use a template to operate, it needs the DNA ends to be as intact as possible in order to minimize the loss of information. Therefore, 5' resection of the DNA inhibits NHEJ[78]. Moreover, the choice between these two repair pathways is also dependent on the cell cycle phase, since homologous recombination is more likely to take place during the S and G2 phases, when a chromatid sister is available, while NHEJ can equally act throughout the cell cycle[79].

Additionally, there is a non-canonical mechanism of non-homologous end joining known as microhomology end joining (MMEJ) that, unlike NHEJ, takes place after 5' end resection, but only if this process exposes regions with microhomology at both resulting single-stranded overhangs. When this happens, these regions may anneal and the remaining DNA gap filled by a DNA polymerase and ligated by a ligase. However, this process implies the deletion of the DNA region between these two microhomologies[80].

Loss of NHEJ is associated with several human diseases. LIG4 syndrome, for example, is caused by mutations in the DNA ligase IV gene. Among LIG4 syndrome symptoms are severe immunodeficiency, microcephaly and other facial abnormalities, photosensitivity and higher risk of developing leukemias[81 82].

Another disease associated with malfunction of the DBSs repair machinery (not specific of NHEJ) is Ataxia-telangiectasia (A-T), caused by inactivation of the ATM gene[83]. As pointed out by its name, symptoms of this disease are ataxia (difficulty to coordinate movement) and telangiectasia (dilated eye blood vessels), but is also linked to increased cancer incidence, immune problems leading to higher risk of infections and others[84]. ATM gene encodes for a protein kinase that participates in multiple cell cycle checkpoints that may be activated in response to DSB, leading to cell cycle arrest and facilitating the repair of the DNA lesion before replication happens[83].

A phenotypically very similar syndrome is the Nijmegen Breakage syndrome. This disease is characterized by mutations in a specific component of the MRN complex: the Nbs1 subunit encoded by the NBS1 gene[85].

Fanconi anemia pathway

Fanconi Anemia pathway (FA) is a repair mechanism specialized in repairing interstrand crosslinks (ICL)[86], which are abnormal covalent bonds between opposite strands of the DNA double-helix that may impair strand separation during replication or transcription. Even though there are some proteins that are specific to the FA pathway, this mechanism mainly involves the coordinated action of several already described repair mechanisms including homologous recombination and nucleotide excision repair[87].

FA pathway commences with FANCM recognizing a stalled replication fork. Then, FANCM and other associated proteins recruit the FA core complex, formed by 10 different proteins of the FA family, which induces the monoubiquitylation of the FANCD2-I heterodimer. Next, FANCS evicts the helicase belonging to the stalled replication fork, and ERCC4 nuclease performs nicks at both sides of the crosslinked base in one of the two affected strands in a process called unhooking, releasing the ICL from one of the parental strands and leaving it attached to the other. This process generates two different products: on one side the nicked parental strand of DNA and its corresponding nascent strand that could not be completely synthesized due to the stalling of the replication fork will carry a double strand break. On the other side, the second parental DNA strand still containing the crosslinked nucleotide from the nicked strand attached to it, while its respective unfinished nascent DNA has a single strand gap. In this second scenario, the single strand gap is filled by a translesion polymerase such as REV1 or DNA polymerase ζ, that is able to do so despite the attached crosslinked nucleotide. Then, this newly synthesized DNA will be used as a template to repair the double strand break using the homologous recombination mechanism. Finally, the unhooked ICL is removed by NER[86].

Fanconi Anemia pathway owes its name to the Fanconi Anemia disease, caused by germline inactivation of one of the 19 genes belonging to this pathway. This disease is characterized by increased cancer susceptibility and bone marrow failure, together with developmental abnormalities[88].

Interestingly, two of the genes belonging to the FA family (FANCD1 and FANCS) were also described in the context of homologous recombination repair and happened to be the well known BRCA2 and BRCA1 genes[89].

Mutational processes

Mutagenesis is a multistep process that usually involves a damaging agent inducing DNA lesions (including mismatches) with the potential to be turned into mutations, either because of a fraction of damage remained unrepaired or because of inaccuracies during the repair or replication process. Each of these combinations of potentially mutagenic DNA lesions induced by a specific damaging agent and the repair mechanisms, either proficient or deficient, responsible to deal with them, are known as mutational process and may induce different types of mutations[23]. Below are a few examples of this:

Reactive oxygen species

Mitochondrial respiration or redox-reactions mediated by heavy metals that take place in aerobic organisms have the potential to generate reactive oxygen species (ROS) like hydroxyl radicals, superoxide anions or hydrogen peroxide. These species may also originate as a result of exogenous damaging agents such as ionizing radiation.

ROS are involved in many aspects of cell metabolism and have been reported to play a role in inflammation, apoptosis[90] and even memory formation[91]. However, ROS also have the potential to damage organic molecules including proteins, lipids, RNA and DNA[92].

ROS mediated DNA damage occurs when these species oxidize DNA bases. This is especially likely to happen in guanines, as they have lower redox potential compared to the rest of the nucleotides[93].

More than 20 different types of oxidative DNA lesions have been identified, the most common of which is the 8-oxo-G, which appears as a result of a two-step process that starts with the addition of an OH to the carbon 8 of a guanine, forming a 8-hydroxy-7,8-dihydroguan-8-yl radical, followed by a one electron oxidation, that creates an 8-oxo-G[94].

During replication, the DNA polymerase is not able to recognize 8-oxo-G lesions as guanines and introduces adenines opposite to them instead, generating pro-mutagenic mispairs. Additionally, 8-oxo-G transformation can happen in guanines belonging to the nucleotidic pool cells use to synthesize new DNA. Thus, 8-oxo-G may be used during this synthesis process, usually opposite to adenines[95].

Another common type of ROS-induced lesions are the FapyG. They originate like the 8-oxo-G but in different environmental conditions that do not favor the oxidation reaction necessary for the second step of the 8-oxo-G formation. If so, a reduction reaction takes place, finally leading to the breakage of the imidazole ring of the guanine[96]. FapyG molecules are potentially mutagenic in a way similar to 8-oxo-G[97].

Endogenous alkylating agents

Endogenous alkylating agents are small molecules with the potential to transfer alkyl groups (C_nH_{2n+1}) to other molecules, including DNA. One of the most well known alkylating agents is the S-Adenosyl methionine (SAM), acting as a methyl group donor (CH_3) in the methylation process[98]. This DNA methylation is known to play a role in gene expression regulation[99]. However, this process may also create adducts with mutagenic potential[100], the most abundant of which are 7-methylguanines (guanines whose N7 position has undergone methylation)[39]. 7-methylguanines are not harmful *per se*, but may destabilize the glycosyl bond between the nucleotide sugar and its base increasing the likelihood of a breakage to occur, generating an apurinic site in the DNA[101]. If replication occurs before these sites have been repaired, errors will be introduced.

Lipid peroxidation products

Lipid peroxidation is an oxidative process that affects lipidic molecules in situations of oxidative stress. During this process, some byproducts with potential to damage DNA are generated. Malondialdehyde (MDA) is one of these byproducts with mutagenic potential[102]. MDA has been reported to increase the interstrand crosslinks formation, mainly in the presence of the sequence 5′-d(CG). This may finally lead to the formation of insertions and deletions[103] and also to the creation of DNA adducts that, when bypassed by error-prone DNA polymerases during replication may lead to G to A mutations[104].

Spontaneous deamination

Deamination is a process in which amine groups are removed from molecules including amino acids and nucleotides. This is usually catalyzed by a family of enzymes called deaminases and is involved in cell metabolism, being essential for protein catalisis, for example. However, deamination may also occur spontaneously, and when affecting DNA, may promote mutagenesis. DNA spontaneous deamination mainly takes place at cytosines and 5-methylcytosines[105]. Deaminated cytosines become uracils, which are normally not found in DNA. If this is not solved, adenines are inserted opposite to them during DNA replication, originating C to T mutations. However, repair of deaminated cytosines is efficient and quick.

In the case of 5-methylcytosines though, deamination occurs at a higher rate than in cytosines[106], and when this happens thymines are obtained, which may also generate C to T mutations if unrepaired. Additionally, the repair of these mispairs is much slower compared to unmethylated cytosines. Therefore, deaminated 5-methylcytosines are the main source of deamination induced mutations[105].

Replication errors

Even though replication is a highly accurate process, errors may still occur, and even though these errors may not be considered a regular damaging agent, they are an endogenous source of mutagenesis[107].

Mismatches are the result of nucleotides misincorporation during DNA synthesis. To minimize these mistakes, some DNA polymerases like polδ and polε have a proofreading subunit able to detect these bases. In humans, the proteins that perform this function are encoded by the POLD1 and POLE genes respectively. Mutations in these genes may impair this proofreading activity and lead to increased error-rate during replication, and therefore, to increased mutation rate as well. Loss of DNA polymerases proofreading capability has been observed in multiple cancer patients, mostly in colon and uterine malignancies[108].

Strikingly, some tumors present loss of polymerase proofreading activity together with mismatch repair deficiency: patients with biallelic mismatch repair deficiency (bMMRD) present inactivating germline mutations in both copies of mismatch repair genes[109]. This inactivation leads to increased mutation rate and patients develop tumors at a very early age. These tumors have, on average, 5 to 10 times more mutations than other pediatric tumors from the same tissue. Due to this increased mutation rate, these tumors rapidly acquire POLE or POLD mutations as well. When this happens, their mutation rate abruptly increases again, reaching levels up to 230-fold higher than in bMMRD tumors with proofreading capability[110].

UV radiation

Ultraviolet radiation is a well known cancer risk factor due to its DNA damaging capacity. It promotes the formation of ROS but is also able to stimulate adjacent pyrimidines (CC, TT, CT, or TC), triggering the formation of abnormal bonds between them. Based on the atoms participating in these bonds, two types of lesions may be formed: cyclobutane pyrimidine dimers (CPDs) and 6-4 pyrimidine-pyrimidone photoproducts (6-4PPs)[37]. These lesions pose several challenges to cells normal functioning: due to their bulky nature they may block replication, leading to replication fork collapse and double strand break formation[111] [112]. Some DNA polymerases such as polη have evolved to bypass these lesions in an error-free manner. However, other polymerases such as polQ tend to systematically introduce adenines opposite to CPDs, generating mispairs whenever cytosines are involved in the CPD formation. These mispairs could finally lead to C to T mutations in subsequent rounds of replication [113].

Moreover, cytosines in CPDs become more prone to being deaminated to uracil[114]. When this happens, adenines are also inserted opposite to them during replication, again generating mispairs that could lead to C to T mutations, even if replicated by error-free polymerases.

Ionizing radiation

Other types of radiation besides UV also have DNA damaging potential. This is the case of ionizing radiation, whose energy is high enough to detach electrons from the atoms it encounters in a process called ionization. There are several well known types of ionizing radiation: x-rays, gamma rays, alpha and beta particles and even some part of the ultraviolet spectrum.

Ionizing radiation is able to induce breaks to the DNA. These breaks may affect just one or both strands of the DNA, generating single or double strand breaks respectively. In addition, if radiation ionizes surrounding water molecules instead of the DNA itself, ROS may be generated as a side effect[115].

This damaging potential of ionizing radiation is exploited as cancer therapy (radiotherapy), which focuses on the use of X-rays and gamma radiation to induce cell death in the tumor[116].

Benzo(a)pyrene

Benzo(a)pyrene (BaP) is an aromatic hydrocarbon released during the combustion of organic matter. Therefore, BaP is generated in many different processes and can be found ubiquitously, being present, for example, in vehicle fumes[117] or grilled food[118]. However, one of the main sources of BaP exposure in humans is the tobacco smoke[119].

BaP itself is not mutagenic, however, it may be metabolized to BaP diol epoxide (BaPDE), which is able to covalently bind to DNA and form DNA adducts, mostly with guanines. Nucleotide excision repair pathway, is the mechanism cells use to repair these lesions. If unsuccessful, guanine-benzopyrene positions may be mistaken for thymines during DNA replication, and adenines introduced opposite to them in the newly synthesized DNA strand. As a consequence G-C to T-A mutations appear [120].

Cisplatin

Cisplatin is a drug that has been used for cancer treatment since 1978[121]. It targets rapidly proliferating body cells, including the ones of the tumor, by forming DNA adducts and, to a lower extent, interstrand crosslinks. These DNA adducts are commonly formed between adjacent guanines (G-G) and less frequently between adenines and guanines (A-G) or guanines separated by another base (G-X-G)[122 123 124].

Cisplatin induced mutations are mainly C to T, although C to A and T to A mutations also occur. Moreover, upon cisplatin exposure, cells show a remarkable increase of dinucleotide substitutions, preferentially at CC, CT, TC, or TG positions. These double substitutions are thought to be created during the unfaithfully repair of cisplatin induced interstrand crosslinks[125]. However, the exact mechanism of cisplatin induced mutagenesis is still unclear.

Base analogs

Base analogs are molecules structurally similar to DNA or RNA forming nucleotides. During replication, base analogs may be introduced instead of one of these nucleotides and, if used as template later on, lead to a mutation[48].

5-Bromouracil is a well known base analog that may be introduced during DNA replication instead of a thymine or a cytosine, pairing with the adenine or guanine from the opposite DNA strand. If used as a template in subsequent rounds of replication, an adenine or a guanine will be indistinctly introduced opposite to the 5-Bromouracil, leading to A to G or G to A mutations, depending on the original base in the position where 5-Bromouracil was introduced[126]. A derivative of 5-Bromouracil known as 5-bromo-2-deoxy-uridine has been used in cancer treatment[127].

Other base analogs also used for cancer therapy are the fluoropyrimidines. This family of drugs includes well known chemotherapeutic agents such as capecitabine and tegafur. Within the body, fluoropyrimidines are metabolized into 5-fluorouracil (5-FU), which impedes thymine synthesis by inhibiting the thymidylate synthase enzyme, ultimately leading to cell death[128]. Additionally, a 5-FU metabolite known as fluorouridine triphosphate may be incorporated into newly synthesized RNA instead of uridine triphosphate, blocking RNA processing and protein synthesis[129].

Intercalating agents

Intercalating agents are molecules that are able to fit between adjacent bases of the DNA double helix in a process is called intercalation[130].

Ethidium bromide is an intercalating agent that emits visible radiation when exposed to UV light. Thanks to this property, ethidium bromide is widely used to visualize DNA bands when performing agarose gel electrophoresis[131].

When intercalated to the DNA, ethidium bromide and other intercalating agents alter the separation between adjacent nucleotides. This may interfere with DNA replication, either by inhibiting it[132] or by leading to the introduction of small insertions or deletions[133]. Thus, intercalating agents are potential mutagens and like many other DNA damaging agents, have been exploited as anticancer drugs (i.e. Doxorubicin)[134].

Mutational signatures

By definition, mutational processes induce mutations to the DNA. However they differ from one another on the type of mutations they lead to. To better characterize this, the frequency of each mutation type, defined by the affected and the resulting nucleotides also considering the upstream and downstream bases are computed. This is the idea behind the concept of the mutational signatures[23].

Thus, mutational signatures may be represented as a combination of the probabilities of each of the 96 possible tri-nucleotide changes to occur, associated to a specific mutational process. UV light, for example, is associated to a mutational signature (signature 7 according to the COSMIC catalogue) characterized by high probability of observing C to T mutations, mainly when cytosines are preceded and/or followed by other pyrimidines[23], reflecting the fact that UV induced lesions occur at dipyrimidines[135]. (Fig. 5.)

This concept has had a big impact on the cancer genomics field, since it has allowed to decipher the mutational processes to which a given sample has been exposed to by studying the proportion of each of these 96 types of mutations. And not only this, but it has also allowed to discover new mutational processes

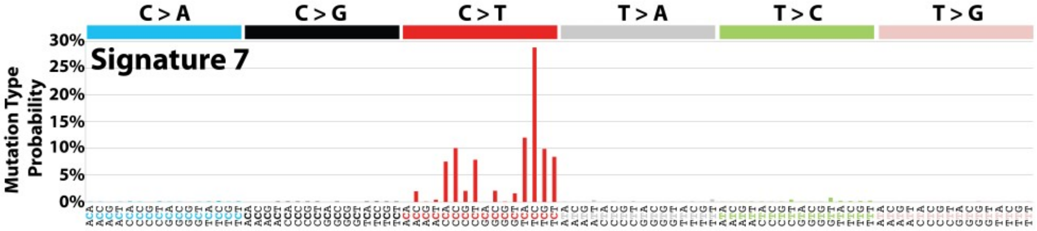


Figure 5. Signature 7 as described by COSMIC. The 96 columns represent all possible mutations when considering the altered nucleotide together with the upstream and downstream flanking bases. The height of the column represents the probability of each of the mutations to be observed.

when the mutational profile of a sample did not resemble any of the already described.

A relevant example of the type of information the analysis of mutational signatures provides has been described in a work by Helen, D. & Glodzik, D. et al. They identified a mutational signature associated with BRCA1/BRCA2 deficiency in breast cancer. This signature reflects the accumulation of mutations of a particular type due to the deficiency of these genes, that participate in the homologous recombination repair mechanism. Thus, by identifying patients with tumors presenting this mutational signature, the BRCA deficient phenotype can be inferred. Interestingly, by using this method they were able to identify patients with mutations in BRCA1 and BRCA2 genes, but also patients that did not carry these mutations, and that presented the BRCA deficient phenotype due to alterations in other genes of the homology directed repair pathway[136].

BRCA deficient tumors are sensitive to a type of drugs known as PARP inhibitors. PARP (Poly (ADP-ribose) polymerase) participates in the repair of single strand breaks. When inhibited, these breaks may turn into double strand breaks, which BRCA deficient tumors fail to repair because they lack the homology directed repair mechanism[137]. Therefore, being able to correctly identify those patients with BRCA deficient phenotype is relevant in order to adopt the most suitable therapeutic strategy.

Finally, the use of mutational signatures has proven useful to build mutation rate models. To do so, the probabilities of all 96 types of mutations are computed from a given sample with enough mutations. This probabilities are sample dependent since they depend on the mutational processes to which the sample has been exposed. Next, the expected number of mutations in a region of the genome can be estimated given its DNA sequence. This type of modeling has been widely employed in cancer driver identification methods[32] [25] as it allows to estimate a mutational background. Later these methods evaluate the differences in terms of signals of positive selection, between observed mutations and the background model, and label genes with statistically significant differences as cancer drivers.

Genomewide mutational heterogeneity

The expected mutation rate of a given genomic location can be modeled by determining the genomewide probabilities of each type of mutation to occur, taking into account the flanking upstream and downstream nucleotides, and then using these probabilities to re-distribute the observed mutations in the area. Sometimes though, these simulations do not reflect the reality, since the number of mutations a region of the genome acquires does not only depend on its DNA sequence.

Back in the year 1989, Kenneth H. Wolfe et al. Already described mutation rate variations across genes, and correlated it with their GC content. Interestingly though, they realized that this was not enough to explain all the variability, and proposed that replication time differences between genes could also have an impact on the rate at which they acquired mutations[30].

With the advent of whole genome sequencing technologies, the capability to study the distribution of mutations greatly increased. In 2012, a landmark study from Benjamin Schuster-Böckler & Ben Lehner used whole genome sequencing data generated from tumor samples to describe how the amount of mutations in different parts of their genomes positively or negatively correlated with a collection of genomic features, including histone modifications, genic density, replication time and others. Remarkably, they reported that H3K9me3, a histone modification associated with the presence of heterochromatin, positively correlated with the mutation rate and could explain up to 40% of the mutation rate variability along the genome[138].

Similarly, in 2013, Michael S. Lawrence, et al. Assessed the mutational heterogeneity along the genome, likely caused by the genomic features described above, and incorporated this information in their cancer driver genes identification pipeline in order to be able to separate genes with actual signals of positive selection from those with a mutational enrichment resulting from the intrinsic mutational heterogeneity of the genome[29].

Moreover, the fraction of the genome covered by each of these genomic features varies. Some of them may encompass large areas while others are very local, ranging from megabases to single nucleotides. Therefore, if their presence affects the mutation rate, we expect these effects to display scale differences as well. For example, heterochromatic areas tend to be several megabases big. Therefore, if chromatin compactness has an effect on modulating the mutation rate, we expect it to be evident when comparing large areas of the genome with different levels of compactness. A typical approach to study this phenomena is to divide the genome in 1Mb segments. However, this analysis will most likely mask the effect of other more local features, such as the effect of a specific protein binding to the DNA. In this case, a local analysis, comparing the binding site of this protein to the immediate flanks would be needed (Fig. 6.).

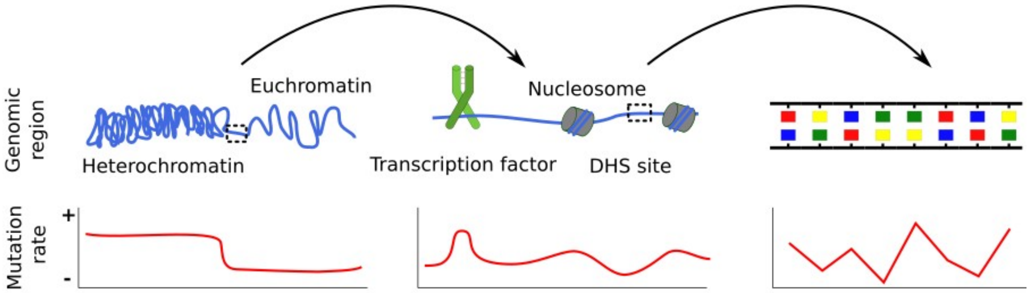


Figure 6. Mutational heterogeneity at different scales. The left panel shows the impact of the level of chromatin compactness in the number of mutations accumulated. The center panel depicts how the presence or absence of different proteins bound to the DNA leads to mutational heterogeneity. The right panel illustrates DNA sequence associated mutation variations.

Multiple cases of mutation rate abnormalities have been described, both by our laboratory and others. Some of them are reviewed below:

Nucleosomes

A well studied local structural variation of the DNA is the alternation between nucleosomes and linker regions along the genome. Several studies aiming to determine the influence of the nucleosomes on the mutation rate have been published, with apparently contradictory results. In 2011, for example, Tolstorukov, M. et al. reported an enrichment of mutations in the DNA wrapped around nucleosomes[139], while Chen, X. et al. reported a depletion of C - T mutations, in the same regions[140]. These results could be explained by the fact that nucleosomes may favour the activity of specific mutational processes, promoting the generation of mutations derived from them, while impairing some others. In this direction, Pich et al. reported that UV induced mutations were enriched in nucleosome-wrapped DNA compared to linkers, while smoking induced mutations behaved in the opposite manner. More specifically, they realized that, while UV damage was created uniformly regardless of the nucleosomes presence, NER repair activity was higher at linkers, leading to an accumulation of unrepaired UV damage in the nucleosome-wrapped DNA. However, they also saw that DNA damage created by tobacco preferentially accumulates in the linkers and not even NER activity being enhanced in these areas was enough to revert this trend, consequently leading to an increased number of mutations[141].

Interestingly, in this same manuscript they described another phenomena affecting the mutation rate at even smaller scale: within the DNA wrapped around each nucleosome. This arises from the fact that DNA minor grooves tend to face towards the nucleosome or the opposite direction in a periodic manner. Since in the context of the double helix structure of the DNA there is a minor groove every 10 bases approximately, this creates a periodicity of this same amplitude. Pich, et al. Described that the incidence of several DNA damaging agents and/or repair mechanisms also followed the same periodicity. Similarly to the nucleosomes vs linkers analysis, specific mutational processes were affected in a different way by this phenomena. For example, mutations caused by oxidative stress tended to be enriched in DNA minor grooves facing towards the nucleosome while UV induced mutations followed the opposite trend[141].

Transcription factors

Another striking local variation of the mutation rate has been observed at transcription factor binding sites. In 2016, two papers from Sabarinathan, R. et al. and Perera, D. et al. describing a very pronounced enrichment of mutations in the active transcription factor binding sites were published in a back-to-back[142] [143]. This phenomena, observed in several tumor types, was most pronounced in melanomas (dominated by UV induced mutations). Also, they reported a decrease in the activity of NER pathway in these areas, concluding that the mutational hotspots present in the transcription factor binding sites were a consequence of the impairment of the repair machinery due to the presence of the transcription factor (Fig. 7.).

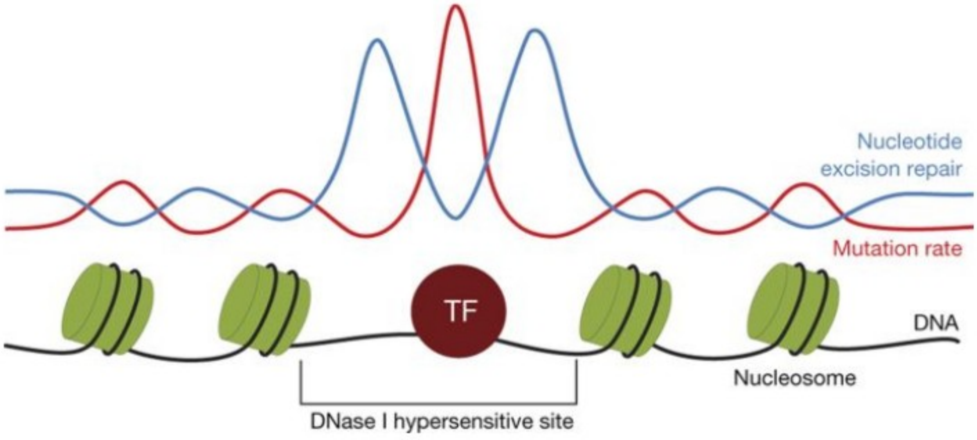


Figure 7. Impact of a transcription factor and flanking nucleosomes on the activity of the nucleotide excision repair machinery and the generation of mutations. Both the transcription factor and the nucleosomes create a decrease in the nucleotide excision repair activity which leads to an increase in the number of mutations in the DNA region they are bound to (Sabarinathan et al., 2016).

In 2018, though, two separate articles reported that in the ETS transcription factors family, the mutational hotspot present in their TFBS was not caused by an impairment of the repair machinery but because of increased susceptibility to UV induced DNA damage[144145].

Non-canonical DNA secondary structures

It has been described that the secondary structure of the DNA has an impact on the number of mutations acquired. Georgakopoulos-Soares, I. et al. Extensively evaluated the rate of mutations in seven different non-canonical DNA secondary structures: G-quadruplexes, Z-DNA, H-DNA, direct repeats , inverted repeats, mirror repeats and short tandem repeats, mostly analyzing mutations from breast tumors. In this context they observed that some non-canonical secondary structures such as H-DNA, Z-DNA and short tandem repeats had a mutation rate more than 1.5 fold higher than expected. This enrichment was even more striking when evaluating indels rather than single point mutations[146].

DHS sites

DNase hypersensitivity sites (DHS) are genomic regions with a relaxed chromatin structure that makes them highly accessible. In 2014, Polak, P. et al.

Described that the mutation rate was depleted in these areas. Interestingly, they also realized that gg-NER was highly active there[147]. Therefore, this was proposed as putative mechanistic explanation for the depletion of mutations. However, the fact that this depletion was present not only in melanoma but also in other tumor types such as colorectal and multiple myeloma, were the relative importance of NER is low, suggests that the efficacy of other repair mechanisms could also be enhanced at DHS regions.

DNA curvature

DNA double-helix may adopt several conformations, defined by the handedness, number of nucleotides needed in a 360º turn, etc. Being the B conformation the most common one[148]. However, DNA conformation is not rigid and variability in terms of thermodynamic features, groove characteristics and DNA shape may be observed along the genome. In 2018, Duat et al. Assessed the effect of this heterogeneity in the mutagenic process, and observed a strong correlation between the intrinsic DNA curvature and the rate of mutations generation. More specifically, they found that regions with less sharp DNA curvature tended to have higher mutation rates and proposed differences in damage sensitivity induced by differential curvature as mechanistic explanation[149].

Final remarks

Understanding the mutational heterogeneity of a genome is a rather complex problem. In order to properly model it, information relative to the mutational processes to which a sample has been exposed to, together with sequence information and all sorts of genomic features needs to be integrated. Moreover, these features may or may not be relevant for the model depending on the dominant mutational processes and the resolution of the model.

However, even though complicated, I believe it is a problem of great interest. An accurate whole genome model of the mutation rate of a tumor can be used as a background model to identify signals of positive selection as previously covered in this introduction. Also, in the context of germline mutations, being able to properly model the neutral rate of evolution of genes is relevant for evolutionary studies that aim to assess the selective pressure of genes. And beyond that, the presence of mutation rate abnormalities unrelated to a selective process may point towards a disruption or enhancement of a given DNA damage and/or repair mechanism and, by unraveling these phenomena, we may get new insights into the basic biology behind these processes.

Therefore, studying the mutation rate variation along the genome has been of great interest to our laboratory and elucidating some of the mechanisms behind it is the main objective of this thesis.

2. Objectives

The two research projects presented in this PhD thesis have the common goal of understanding how mutational processes shape the distribution of somatic mutations along the genome, mainly at a local scale. This general objective can be divided into several, more specific ones:

- Understand how mutations accumulate in exons and introns across tumor types.
- Dissect the reasons for these deviations by determining a differential DNA damage formation and/or DNA repair in these genomic elements.
- Understand how DNA modifications and chromatin conformation affect the occurrence of DNA damage and the activity of the DNA repair mechanisms.

Thus, the goals specific to each of the two research projects derive from this general ones. On the one hand, we aimed to understand the accumulation of UV induced mutations around active transcription factor binding sites. In this context, our objectives could be subdivided into:

- Assess the level of disagreement between the observed and expected number of mutations in active transcription factor binding sites.
- Understand how UV induced DNA damage is distributed in and around these areas.
- Evaluate how bound transcription factors interfere with the activity of the nucleotide excision repair machinery.
- Determine the degree of similarity across families of transcription factors in the way they impact on the UV induced DNA damage

formation and its repair, and ultimately in the way mutations accumulate in their active binding sites.

On the other hand, we planned to study how mutations accumulate in the exonic regions of genes and compare it to their intronic counterparts. More specifically our goal were:

- Identify discrepancies between the observed and expected mutation rate at exonic regions across tumor types.
- Determine the mutational processes more likely to be acting in a biased manner in these areas.
- Evaluate the relative contribution of both DNA damage formation and repair in the biased generation of mutations in exons by comparing tumors with and without inactivating alterations in the DNA repair mechanism of interest.
- Identify chromatin features associated to the observed variations on the DNA repair machinery.

3. Results

Reduced mutation rate in exons due to differential mismatch repair

The first chapter of the results section describes a research project whose aim is to identify differences in the rate at which mutations are generated between introns and exons in somatic cells. To do so we made use of publicly available tumors whole genome sequencing data, together with histone modifications chip-seq data. Tumor whole genome DNA sequencing data was obtained from TCGA[150] and from the International Biallelic Mismatch Repair Consortium[110], and allowed us to study the distribution of somatic mutations along the genome, while histone modifications chip-seq was generated by the Roadmap Epigenomics Project[151].

Thus, we first described the presence of a reduced exonic mutation rate, which had been observed before, but attributed to negative selection on exonic mutations. Nevertheless, we ruled out that differences in DNA or negative selection on mutations in exons are the cause of this decrease in somatic mutations. Next, and given that this reduced exonic mutation rate was especially evident in samples whose mutations had been generated by mutational processes involving mismatch repair, we proposed differences in its activity as mechanistic explanation. We validated this hypothesis by analyzing samples with inactive mismatch repair and observed that the reduced exonic mutation rate decreased or disappeared depending on how early in tumor development the inactivation of mismatch repair occurred. Finally, we proposed that enrichment of the H3K36me3 mark in exons is the reason behind this increased exonic mismatch repair activity.

First authorship of this work is shared between Sabarinathan, R. and myself. The two of us participated in most of the analyses performed during the process, many times working in parallel and conducting very similar experiments in order to validate each other’s results. The only section of the paper conducted exclusively by Sabarinathan, R. Is the one exploring the differences in the activity of the nucleotide excision repair (NER) pathway between introns and exons, illustrated in the Supplementary Figure 7. On the other hand, I was the main responsible for carrying out the modifications and additional analysis suggested by the reviewers during the publishing process.

Frigola J*, Sabarinathan R*, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. (2017) Reduced mutation rate in exons due to differential mismatch repair. Nature Genetics. 49(12):1684-1692
https://doi.org/10.1038/ng.3991

*Co-first authors

nature genetics

Reduced mutation rate in exons due to differential mismatch repair

Joan Frigola[1,2,4], Radhakrishnan Sabarinathan[1,2,4], Loris Mularoni[1,2], Ferran Muiños[1,2], Abel Gonzalez-Perez[1,2] & Núria López-Bigas[1–3]

While recent studies have identified higher than anticipated heterogeneity of mutation rate across genomic regions, mutations in exons and introns are assumed to be generated at the same rate. Here we find fewer somatic mutations in exons than expected from their sequence content and demonstrate that this is not due to purifying selection. Instead, we show that it is caused by higher mismatch-repair activity in exonic than in intronic regions. Our findings have important implications for understanding of mutational and DNA repair processes and knowledge of the evolution of eukaryotic genes, and they have practical ramifications for the study of evolution of both tumors and species.

Genetic variation in exonic regions is lower than in intronic ones both across species and within populations. This differential exon–intron variation rate is attributed to the action of stronger purifying selection on exonic nucleotide changes, whereas the rate of generation of variants—which precedes the effect of selection—is generally assumed to be overall homogeneous between these two genic regions. This assumption lies at the heart of evolutionary biology and cancer genomics approaches that compare the rates of intronic and exonic variation to estimate the strength of selection acting on coding genes[1–5].

Recent studies have shown that the rate of mutations across genomic regions is highly heterogeneous. Replication timing[6,7], the level of gene expression[8] and the degree of chromatin compaction[9,10] have been described as features that affect mutation rate at the megabase scale. Our group and others recently demonstrated that the local efficiency of DNA repair is influenced by factors that affect accessibility of the repair machinery[11–14].

The assumption that introns and exons have similar basal rates of mutation before the action of purifying selection is a reasonable one because exonic and intronic regions are replicated at the same time and transcribed equally. DNA repair mechanisms associated with the advance of the replication fork, as well as transcription-coupled repair, are therefore expected to have equivalent access to both regions. Nevertheless, several features of the chromatin structure—including some that have been related to the recruitment of DNA repair machineries[15–17]—vary widely between exons and introns[18,19]. This motivated us to question the long-standing assumption that introns and exons have similar rates of mutation before selection.

Somatic mutations detected in tumors[20] are an ideal ground to explore whether exonic and intronic variants appear at the same rate. Tumor cells, upon clonal expansion, accumulate somatic mutations at accelerated rates as compared to the germ line. We demonstrate here that, even in the absence of purifying selection, exons acquire fewer mutations than expected given their nucleotide composition. We show that this decreased exonic mutation burden is detectable across seven tumor types. We also demonstrate that the cause of this reduction is that the mismatch repair (MMR) system acts more efficiently in exons than in introns, and we propose that this differential repair is caused by the differential positioning of histone marks in these two genic regions.

These findings imply that the differential genetic variation in exonic and intronic regions across species and within populations is caused by a combination of different sequence context, rate of DNA repair and purifying selection. This has ramifications of a technical nature for evolutionary methods that rely on the calculation of intronic variation to estimate the strength of selection on genes or to detect cancer driver genes[1–3,5,21,22]. More generally, these findings have profound implications for knowledge of gene evolution and DNA repair mechanisms.

RESULTS

Differential distribution of chromatin features in exons and introns

We first sought to identify chromatin features with the most different distributions between exons and introns, using data generated by the Roadmap Epigenomics[23] and the Encyclopedia of DNA Elements (ENCODE)[24]. We analyzed 32 chromatin features—comprising 30 histone modifications, the presence of a histone variant (H2A.Z) and DNase I–hypersensitive sites (DHSs)—in 127 cell lines and primary cells from different tissue types and nucleosome density obtained in a lymphoblastoid cell line (**Supplementary Table 1**). We computed the coverage (fraction of bases overlapping peaks) of each feature on exons and introns located at different positions along the structure of genes (the results of this calculation for three chromatin features are shown in **Fig. 1a**; Online Methods). Then, we defined the difference

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [2]Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona, Spain. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [4]These authors contributed equally to this work. Correspondence should be addressed to N.L.-B. (nuria.lopez@irbbarcelona.org).

Received 30 September 2016; accepted 12 October 2017; published online 6 November 2017; doi:10.1038/ng.3991

in the exonic and intronic coverage of each mark in each cell type as the *P* value of the two-tailed Mann–Whitney test of their comparison (box plots in **Fig. 1a**). Several chromatin marks exhibited a significant overall difference in exonic and intronic coverage (**Fig. 1a**,**b**). In particular, nucleosome density and trimethylation of histone H3 at lysine 36 (H3K36me3) were significantly higher in exons than in introns across the gene, and H3K36me3 was the histone mark with higher coverage across all exons in the gene. This behavior of H3K36me3 was consistent across the majority of the 127 cell types in Roadmap Epigenomics (**Fig. 1b** and **Supplementary Tables 1** and **2**). Moreover, H3K36me3 coverage decreased steeply in flanking introns (**Fig. 1c**). Interestingly, the hMutSα protein of MMR machinery, involved in the recognition of mismatches, has recently been described as being recruited to the chromatin through interaction of its hMSH6 subunit with H3K36me3 (refs. 15,17).

We therefore hypothesized that the exonic enrichment of certain chromatin features, in particular H3K36me3, might result in increased recruitment of the MMR machinery to exons. This, in turn, would lead to a reduction in the quantity of exonic mutations with respect to the number of mismatches expected from the exonic sequence content alone.

Internal exons exhibit decreased exonic mutation burden in *POLE*-mutant tumors

POLE-mutant tumors, owing to the decreased proofreading capabilities of DNA polymerase ε, sustain a substantial number of mismatches during DNA replication, which make up a sizable part of their somatic mutations. Therefore, to determine whether the rate of somatic mutations caused by mismatches differs in exonic and intronic regions, we first explored the mutations detected across the whole genomes of six *POLE*-mutant colorectal tumors, sequenced by The Cancer Genome Atlas (TCGA). We stacked exon-centered 2,001-nucleotide (nt) sequences and computed the mutation burden at each position of this window as the number of mutations overlapping the position. This analysis showed that the mutation burden in positions dominated by exonic sequences was lower than that observed along flanking intronic regions (**Fig. 2a**, red line).

The mutation probability at individual DNA positions is influenced by sequence context. Therefore, differences in nucleotide composition between exons and introns could provide a plausible explanation for the observed difference between exonic and intronic mutation counts. To compute the expected mutation burden at each position of the 2,001-nt exon-centered window, we distributed the mutations observed in each sequence in the stack, taking into account the conditional probability that each of the 2,001 positions was mutated given the adjoining 5′ and 3′ bases. This sequence-wise distribution of expected mutations (details in the Online Methods) avoids potential biases resulting from aggregating genic regions with different mutation rates and exon/intron proportions (**Supplementary Fig. 1**). The distribution of these synthetically generated 'expected' mutations in *POLE*-mutant tumors across exons and their flanking introns showed that more mutations are expected in exons than in introns, as represented by the black line in **Figure 2a**.

We then set out to compare the number of observed exonic mutations to the expected quantity in *POLE*-mutant tumors and to assess the statistical significance of the deviation between the two (**Fig. 2b** and Online Methods). We carried out this comparison at the level of individual genes to guarantee that its results were free from the aforementioned caveat. (Known cancer driver genes[25,26] were excluded from this and subsequent analyses to eliminate any deviation due to positive selection.) First, we randomly distributed a number of mutations equal to that observed in each gene across its exons and introns, according to the probability of each nucleotide being mutated. A second method to obtain expected mutation burden based on permutations of observed mutations yielded similar results (Online Methods, **Supplementary Fig. 2** and **Supplementary Table 3**). We then computed the difference between the observed and expected mutation burdens for each gene (**Fig. 2c**). Most genes (77%) possessed fewer exonic mutations than expected from their sequence content, resulting in a negative difference. After aggregating the numbers of observed and expected mutations across all genes (**Fig. 2d**), we discovered that, whereas internal exons bore only 5,616 mutations in the six *POLE*-mutant tumors, 8,996 exonic mutations were expected, given (i) the total number of genic mutations, (ii) the nucleotide composition of exons and introns, and (iii) the mutational processes operating in these tumors. This represents a decrease of 37.6% for the observed exonic mutation burden with respect to the expected burden. Employing a likelihood-based statistical approach (Online Methods), we found this decrease to be statistically significant ($P < 0.0001$). We have named this phenomenon 'decreased exonic mutation burden', and we quantify it globally as the percentage decrease with respect to the expected mutation burden.

We next tested whether the decreased exonic mutation burden was due to increased selective pressure on exons resulting in purifying selection of mutations in these regions during tumor evolution. To determine the impact of purifying selection on the exonic mutation burden, we separated exonic mutations on the basis of their consequence types. We found that the 5,616 exonic mutations in the six *POLE*-mutant tumors corresponded to 950 synonymous and 4,666 nonsynonymous mutations. If the decreased exonic mutation burden were caused by purifying selection, we would expect it to consist mostly of a decrease in nonsynonymous mutations. Nevertheless, when redistributing genic mutations across intronic, synonymous and nonsynonymous sites according to their mutational probabilities, we found a 35.7% decrease in nonsynonymous mutations, along with a 45.4% decrease in synonymous mutations ($P < 0.0001$; **Fig. 2d**). On the other hand, when redistributing solely exonic mutations on the basis of their mutational probability, we found that the expected number of nonsynonymous mutations was very close to the actual number observed: 4,562 (with the remaining 1,054 expected to yield synonymous variants). The results of these two tests support the conclusion that the decrease in the exonic mutation burden is not due to negative selection (**Fig. 2d**). This result was maintained across bins of genes with different mutation rates and was observable for all individual *POLE*-mutant tumors (**Supplementary Tables 4** and **5**).

We then checked that the decreased exonic mutation rate was not driven by a subset of genes at either extreme of the mutation rate range. To do this, we binned the genes into ten groups of increasing mutation rate (**Fig. 2e**, top). We then aggregated the mutations of the genes in each bin and confirmed that the decreased exonic mutation burden remained around 40% across all bins. Finally, we found that very similar values of decreased exonic mutation burden were observed across groups of genes with increasing replication times, expression levels and H3K36me3 coverage, and also across exons at different positions along genes (**Fig. 2e**, second to bottom panel). Furthermore, the decrease in exonic mutation burden was not driven by one or few *POLE*-mutant tumors, as it was observable and significant for each of them (**Fig. 3a**; this analysis also included a *POLE*-mutant tumor of uterine adenocarcinoma origin).

In summary, we found a significant decrease in the exonic mutation rate in *POLE*-mutant tumors. This decrease is not due to sequence content and cannot be explained by negative selection acting on exonic mutations, and it is maintained across genes with all levels of mutation rate and across exons at different positions of the gene.

Decreased exonic mutation rate is caused by differential mismatch repair

We reasoned that the decreased exonic mutation burden observed in *POLE*-mutant tumors could be caused by elevated activity of MMR in exons with respect to their neighboring introns. MMR is

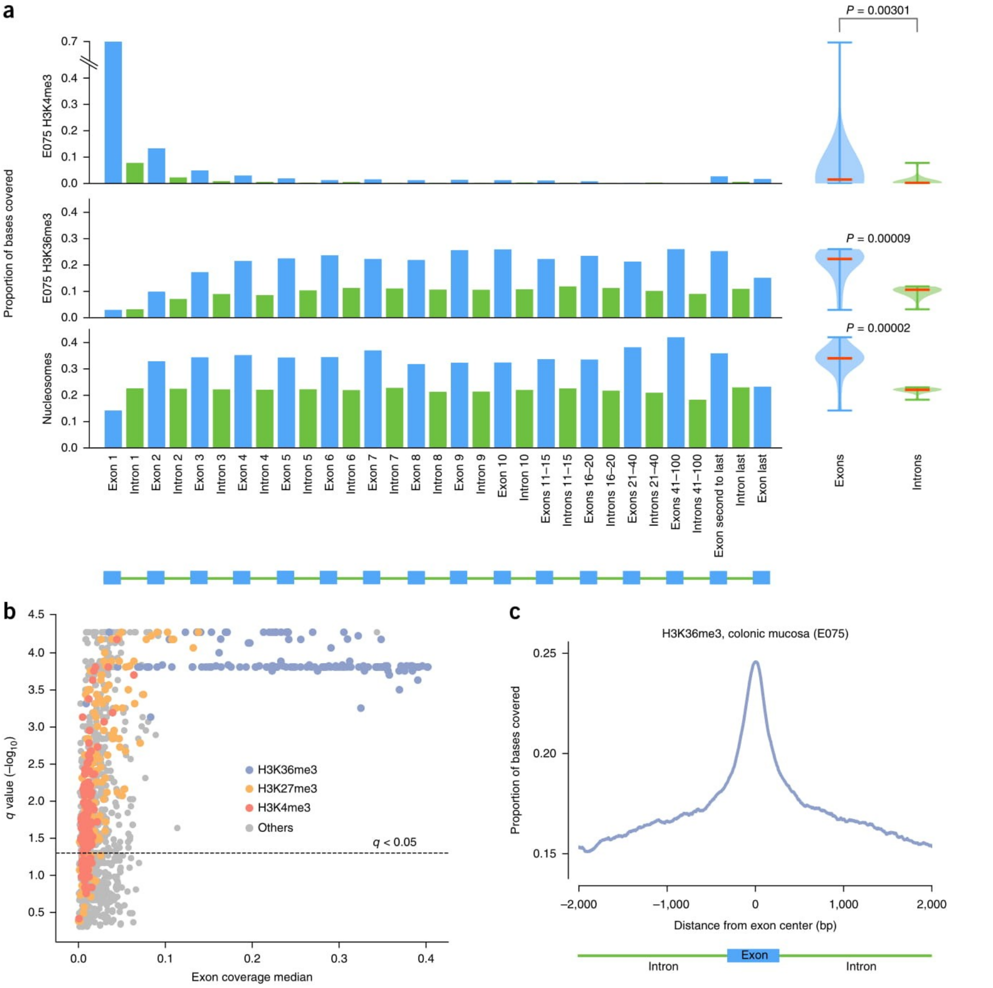


Figure 1 Exonic enrichment for several histone marks. (**a**) Exonic and intronic coverage of H3K4me3 and H3K36me3 peaks in primary cells of the normal colon mucosa (E075) and of nucleosome-covered regions in GM12878 (lymphoblastoid cell line). Each bar represents the coverage of the mark in exons or introns at different positions of genes, depicted by the schematic structure of the genes at the bottom of the figure. The distribution of the exonic and intronic coverage of each chromatin feature across the gene structure is represented by the box plots at the right of the panel. The *P* value from a two-tailed Mann–Whitney test comparing the two distributions is shown. (**b**) Scatterplot representing the difference in exonic and intronic coverage of each histone mark ($-\log_{10}$ two-tailed Mann–Whitney *P* value corrected for multiple testing) on the *y* axis and the median coverage across all exons of genes on the *x* axis. Each dot represents one chromatin mark in one cell type, colored according to the former. All data on exonic and intronic coverage of all marks across cell types is available in **Supplementary Tables 1** and **2**. (**c**) Proportion of bases covered by H3K36me3 at internal exons and flanking introns in primary cells of the normal colon mucosa (E075).

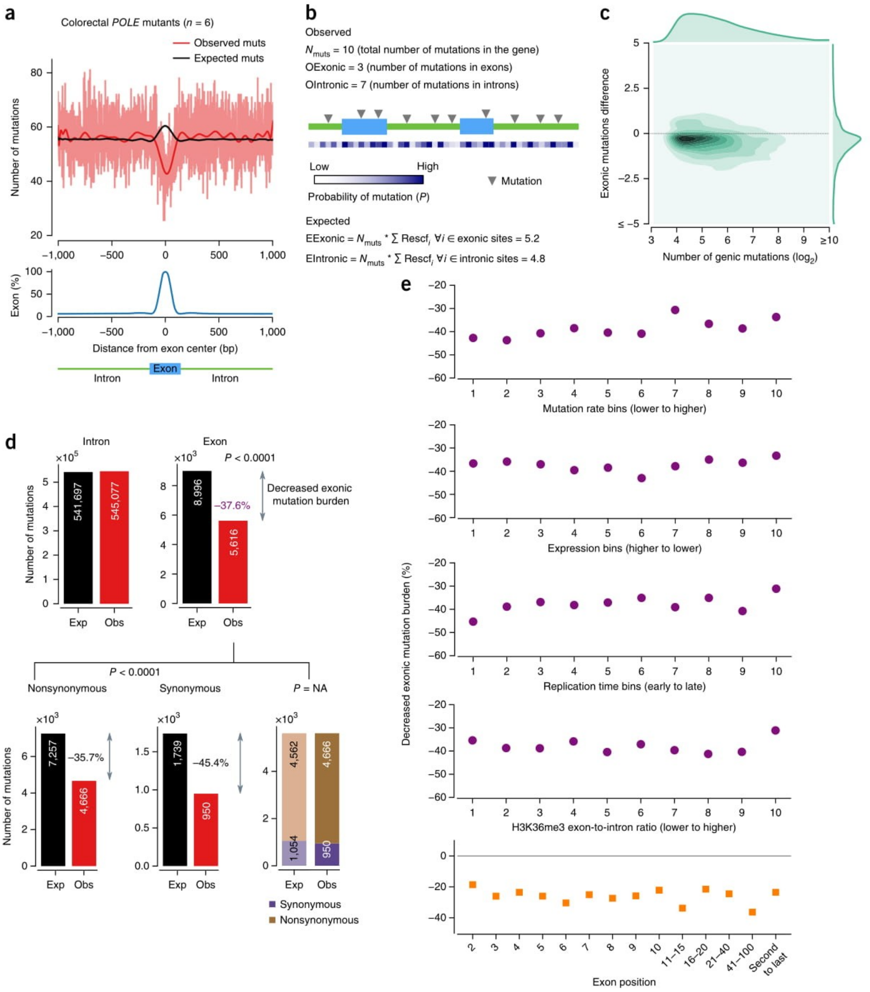


Figure 2 Decreased exonic mutation burden in *POLE*-mutant colorectal tumors. (**a**) Exon-centered 2,001-nt-wide observed and expected profiles of mutations in six *POLE*-mutant colorectal tumors. The light red line represents the actual number of mutations at each position, while the red and black lines represent smoothed mutation numbers using a polynomial fit. Bottom, distribution of the percentage of exonic bases at each position across the 2,001-nt window. (**b**) Schematic of the method used to compute observed and expected numbers of mutations in exons and introns at the gene level. Rescf_i represents the rescaled expected frequency of mutation of each nucleotide in the gene. The conditional probability of mutation of each site (*P* in the figure) is proportional to this quantity. EExonic and EIntronic represent the expected numbers of exonic and intronic mutations, respectively (see Online Methods for details). (**c**) Density plot representation of the distribution of gene-level differences in the numbers of observed and expected exonic mutations versus total mutations in *POLE*-mutant tumors. The distribution of exonic mutation difference (right-hand 1D density plot) is biased toward negative values, indicating that a majority of genes possess lower than expected numbers of mutations in exons. An analogous plot, restricted to genes with at least one expected exonic mutation (**Supplementary Fig. 2b**) shows similar results. (**d**) An overall highly significant 37.6% decreased in exonic mutation burden (3,380 fewer observed exonic mutations than the 8,996 expected) is observed in these tumors (top). Both synonymous and nonsynonymous mutations account for this decrease (bottom), with their observed values significantly below expected (left and middle plots), and no fewer nonsynonymous mutations than expected when the latter are computed solely from observed exonic mutations (right plot). (**e**) The decreased exonic mutation burden is maintained around the overall computed value (37.6%) across groups of genes with different mutation rate (first row), level of expression (second row), replication time (third row), the number of genic bases covered by H3K36me3 peaks (fourth row) and across exons at different positions in the gene (fifth row). The values in the abscissa in each graph represent the ordinal number of the bins of genes, sorted in the direction indicated in each panel.

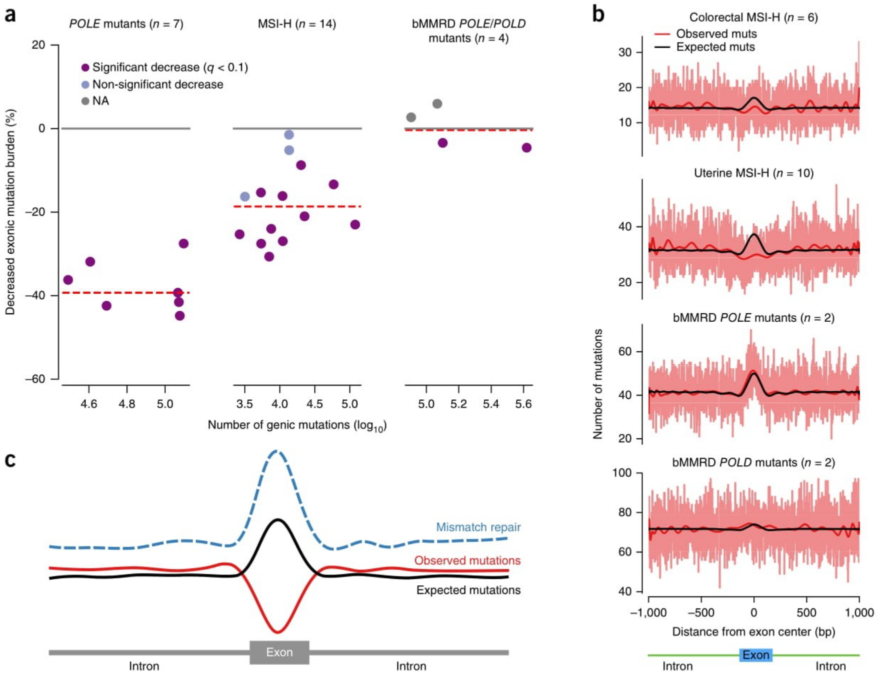


Figure 3 Decreased exonic mutation burden across scenarios of MMR activity. (**a**) Tumor-level decreased exonic mutation burden in *POLE*-mutant (left) and, MSI-H (center) colorectal and uterine tumors and bMMRD glioblastomas (right). Dots represent individual tumors. Broken red lines represent the median decreased exonic mutation burden of each group of tumors. (**b**) Exon-centered 2,001-nt-wide observed and expected profiles of mutations in six colorectal and ten uterine MSI-H tumors and two *POLE*-mutant and two *POLD*-mutant bMMRD glioblastomas. (**c**) Schematic showing the increased efficiency of MMR at exons and the decreased exonic mutation burden.

the main mechanism responsible for the repair of errors generated by the polymerase during DNA replication. Colorectal tumors, as well as other cancer types, acquire a microsatellite instability (MSI) phenotype when mismatches introduced by the DNA polymerase are not corrected, owing to deficiencies in the MMR system[27]. MSI tumors are normally classified on the basis of the level of five biomarkers into MSI-H (high, with over 40% of the biomarkers of MSI) and MSI-L (low, with less than 40%), although the latter have recently been shown to not significantly differ from microsatellite stable (MSS) tumors in numbers of gained microsatellite alleles[28]. Thus, if our hypothesis were true, we would expect tumors with an impaired MMR function (MSI-H) to show lower decreased exonic mutation burden than MMR-competent tumors, such as *POLE*-mutant or MSS tumors.

We proceeded to compute the decreased exonic mutation burden of six colorectal and eight uterine MSI-H tumors in the TCGA cohort. We found, as predicted by our hypothesis, that MSI-H tumors exhibited a decreased exonic mutation burden of around 20% (**Fig. 3a**,**b**), close to half of the decrease observed in the MMR-proficient *POLE*-mutant tumors. Several reasons may explain why the decreased exonic mutation burden did not disappear completely in MSI-H tumors. On the one hand, the impairment of the MMR system may not be complete and has probably not existed throughout the entire history of the tumor. On the other hand, alternative mutational processes may also contribute to the mutation load.

Then, we computed the decreased exonic mutation burden of two *POLE*-mutant and two *POLD*-mutant glioblastomas from children with inherited biallelic mismatch-repair deficiency (bMMRD) sequenced by the International BMMRD Consortium[29]. These tumors have been MMR-deficient throughout their entire history, and their *POLE* or *POLD* mutations guarantee a preponderance of mismatch-caused mutations. Their decreased exonic mutation burden was indeed close to zero (**Fig. 3a**,**b**), with independence of the mutation rate of genes (**Supplementary Fig. 3** and **Supplementary Table 4**). Mismatches in these tumors were generated at a rate comparable to that in previously analyzed *POLE*-mutant tumors. However, most of these mismatches remained uncorrected and turned into mutations. In other words, the mutations observed in these tumors follow the pattern of mismatch generation, corroborating our hypothesis that they appear with higher probability in exons than introns and that it is MMR, with its increased efficiency in the former, that causes the decreased exonic mutation burden.

In summary, the decreased exonic mutation burden differs between three different scenarios of MMR activity, with higher decrease in MMR-proficient tumors to none in MMR-deficient tumors. These results indicate that the increased activity of MMR in exons is the

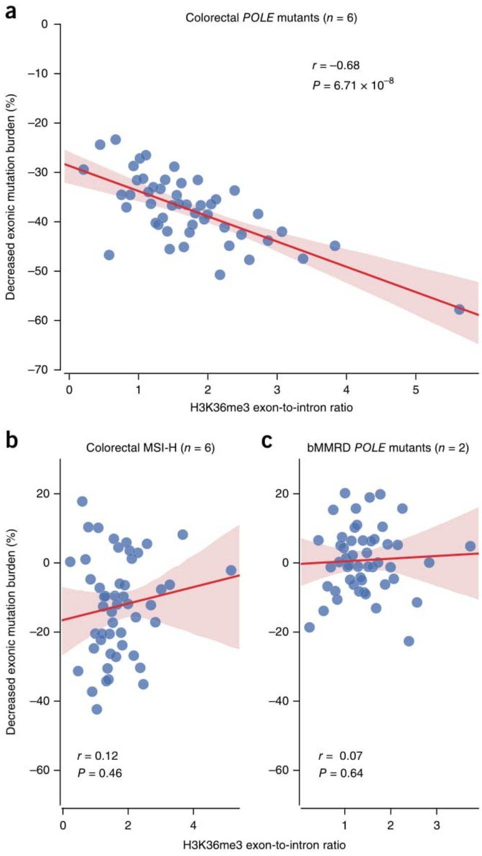


Figure 4 Decreased exonic mutation burden and H3K36me3 exon-to-intron ratio. (**a**–**c**) Decreased exonic mutation burden computed in *POLE*-mutant colorectal tumors (**a**), MSI-H colorectal tumors (**b**) and bMMRD glioblastoma tumors (**c**) for 50 groups of genes with increasing exon-to-intron ratios of H3K36me3 coverage (in the corresponding cell of origin; Online Methods). The trendline and its confidence interval in graphs was added using the seaborn package of Python, while the correlation coefficient and its significance were computed using an iteratively reweighted least-squares approach.

cause of the decrease in exonic mutation burden in *POLE*-mutant tumors (**Fig. 3c**).

A role for H3K36me3 in the differential activity of MMR in exons and introns

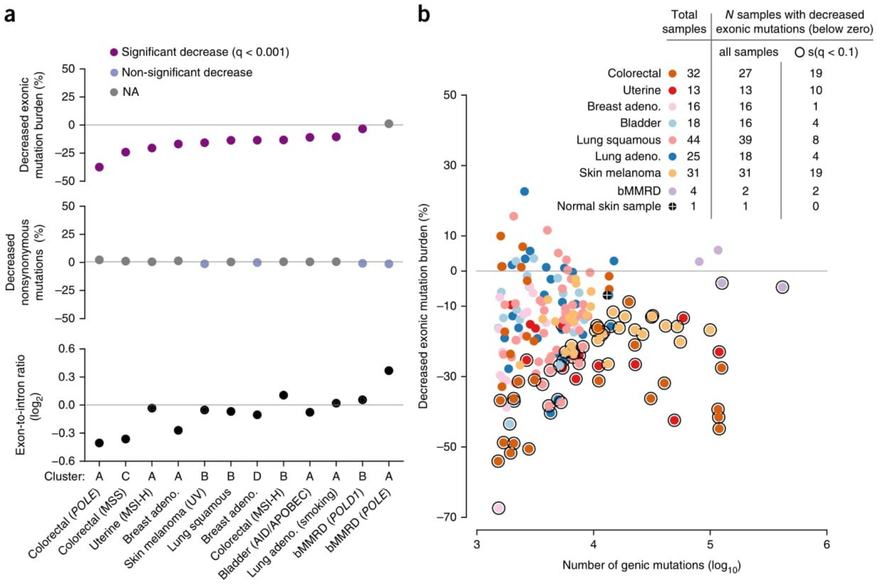
The results of the previous two sections demonstrate that the enhanced efficiency of the MMR system in exons is the cause of the observed decreased exonic mutation burden of colorectal *POLE*-mutant tumors. On the basis of formerly established mechanistic links between H3K36me3 and the recognition of mismatches, we then hypothesized that the decreased exonic mutation burden could be explained, at least in part, by the exonic enrichment of this histone mark in cells of the colon epithelium. If true, we should be able to observe the biggest decrease in exonic mutation burden in genes with the strongest exonic enrichment for H3K36me3 in MMR-proficient tumors. To test this, we first computed the exon-to-intron ratio of H3K36me3 read count for primary cells from the colonic mucosa (E075; **Fig. 4a**)[23]. Then, we grouped the genes into bins of increasing H3K36me3 exon-to-intron ratio, and we computed the aggregated decrease in exonic mutation burden of the genes in each bin for *POLE*-mutant colorectal tumors (**Fig. 4a** and **Supplementary Figs. 4** and **5**). As predicted by our hypothesis, we found a significant negative correlation between the H3K36me3 exon-to-intron ratio and the decrease in exonic mutation burden (correlation coefficient = −0.68, $P = 6.7 \times 10^{-8}$). A much weaker, non-significant correlation (**Supplementary Fig. 5**, bottom) was observed between the exon-to-intron ratio of nucleosomes and the decrease in exonic mutation burden. This suggests that the H3K36me3 histone mark and not just the presence of nucleosomes underlies the increased level of MMR in exons that results in the decreased exonic mutation burden. The correlation with other histone marks was also lower (**Supplementary Table 6**). On the other hand, the negative correlation between the H3K36me3 exon-to-intron ratio and the decreased exonic mutation burden was absent in MSI-H colorectal tumors (correlation coefficient = 0.12, $P = 0.46$) and bMMRD tumors (exon-to-intron H3K36me3 read count ratio computed from cells of the brain angular gyrus, E067; correlation coefficient = 0.07, $P = 0.64$) (**Fig. 4b**,**c**).

These results indicate that the exonic enrichment for H3K36me3, possibly in combination with other chromatin features, could act as a driver of the enhanced MMR activity in exons that ultimately results in the decreased exonic mutation rate of *POLE*-mutant tumors. When cells become MMR deficient, either during tumor evolution (MSI-H colorectal samples) or before its emergence (bMMRD glio blastomas), the link between the H3K36me3 exonic enrichment and the decreased exonic mutation burden is thus severed. This results in uncorrected mismatches accumulating and, ultimately, mutations appearing more frequently in exons.

Tumors of other cancer types also exhibit decreased exonic mutation rate

Our observations in previous sections have been limited to colorectal and uterine carcinomas, the mutational spectra of which are dominated by the interplay between the generation of mismatches in the course of DNA replication and their correction by the MMR machinery. The mutational processes of other somatic tissues are dominated by different types of damage dealt with by other DNA repair systems. Nevertheless, somatic cells in a human body, as well as the gametes, are the result of millions of cell divisions involved in organism development and tissue renewal. Therefore, MMR must have a role—although with different relative contribution—in shaping the mutational processes of all human tissues. We then asked whether tumors originated from other tissues exhibit a decreased exonic mutation rate. To do this, we first clustered the samples of eight tumor types on the basis of their mutational signatures (**Supplementary Fig. 6**).

For the tumors in each cluster, we next computed the decreased exonic mutation burden (**Fig. 5a**, top). All clusters except the one grouping *POLE*-mutant bMMRD glioblastomas exhibited significantly decreased exonic mutation burden. This global trend was corroborated for individual samples (**Fig. 5b**). Interestingly, we found that the decreased exonic mutation rate was apparent also in the somatic mutations detected in a normal skin sample (**Fig. 5b**, white-crossed



	Total samples	*N* samples with decreased exonic mutations (below zero)	
		all samples	○ s(q < 0.1)
Colorectal ●	32	27	19
Uterine ●	13	13	10
Breast adeno. ●	16	16	1
Bladder ●	18	16	4
Lung squamous ●	44	39	8
Lung adeno. ●	25	18	4
Skin melanoma ●	31	31	19
bMMRD ●	4	2	2
Normal skin sample ⊕	1	1	0

Figure 5 Decreased exonic mutation burden across cancer types. (**a**) Top, decreased exonic mutation burden of groups of tumors clustered according to their underlying mutational processes. Dots represent clusters of tumors denoted in the bottom panel. Middle, decreased nonsynonymous mutation burden of the same groups of tumors, computed as in **Figure 2d**. Bottom, ratio of exonic and intronic mutation rates of the same groups of tumors. Clusters of tumors in the three panels are sorted following their decreased exonic mutation burden. (**b**) Decreased exonic mutation burden of individual tumors versus their mutational burden. Dots representing individual tumors are colored according to their cancer type; dots of tumors with significantly decreased exonic mutation burden are encircled by a black ring. The table to the top right presents the total number of samples, the subset of them with decreased exonic mutation burden and the subset of these with a significant decrease.

black dot)[30], indicating that this phenomenon is not a pathological effect caused by tumorigenesis. In none of the clusters could the decreased exonic mutation burden be attributed to negative selection acting on exonic mutations (**Fig. 5a**, middle). We also computed the exon-to-intron mutation rate ratio as explained in the first section for the chromatin features (Online Methods). In coherence with the decreased exonic mutation burden, in most clusters, exons showed fewer mutations than their intronic counterparts (**Fig. 5a**, bottom).

Strikingly, even melanomas and lung carcinomas, whose mutations arise mostly as a consequence of DNA damage caused by UV light or tobacco, respectively, repaired via nucleotide-excision repair (NER)[31,32], exhibited a clearly decreased exonic mutation burden. Two explanations are plausible for the pervasive decreased exonic mutation rate identified. The first, as pointed out above, is that, although modest in relative terms, the MMR still has a role in DNA repair in these tumors. Nevertheless, a second intriguing possibility is that other DNA repair machineries, also acting with higher efficiency in exons, contribute to the reduced exonic mutation rate. Exploring this prospect in the case of NER in melanomas, we indeed found higher activity in exonic regions (**Supplementary Fig. 7**), although we cannot rule out the possibility that this is due to a higher exonic rate of UV-induced damage.

To summarize, the decrease in somatic mutation burden in exonic regions with respect to the expectations and to neighboring introns is apparent across cancer types. While we have demonstrated that MMR has a role in shaping this decrease, other DNA repair mechanisms may also contribute to it.

DISCUSSION

In this work, we provide, to the best of our knowledge, the first demonstration that the generation of somatic mutations—in the absence of negative selection—is lower in exons than expected given their nucleotide composition. In other words, somatic cells exhibit a decreased exonic mutation burden. We have also shown that the reason is that mismatches in exonic DNA are repaired more efficiently than their intronic counterparts. These results represent an important contribution to the body of research that in recent years has revealed higher than anticipated heterogeneity in the mutation rate across different regions of the genome. Several recent seminal studies exploiting whole-genome germline and somatic mutations and the availability of nucleotide-resolution maps of DNA repair[33,34] have provided glimpses at a complex relationship between chromatin conformation, basic cellular processes like gene expression, DNA replication, the binding of transcription factors, and DNA repair[5,7,10–12,14,32,35–39]. It is the complicated interplay between these processes that determines that mutations accumulate heterogeneously across the genome. The results we present here show that the interaction of the most basic structural feature of eukaryotic genes, namely their segmentation in exons and introns, and its correlative chromatin structural differences, results in these two regions being repaired at very different rates.

As a possible explanation of the mechanisms through which the segmented structure of genes influences the activity of the DNA repair machinery, we have shown a pervasive enrichment of the trimethylation of H3K36 in exons of normal tissues, which correlates with the decrease in exonic mutation burden in the corresponding tumors. As we show here, H3K36me3, possibly in combination with other chromatin features, may participate in shaping the observed depletion of exonic mutations. The enrichment of H3K36me3 for exonic regions, which appears in both germline and somatic tissues, has been proposed to be ultimately responsible for the correct recognition of exon–intron boundaries by the splicing machinery[18,19,40]. Nevertheless, H3K36me3 is bound by the MutSα protein via the PWWP domain of its MSH6 subunit[15]. A concomitant factor may thus have acted in the evolution of H3K36me3-enriched exons. Indeed, our results suggest that the increased recruitment of the MMR machinery to exonic regions as a result of higher levels of this histone mark would result in a reduction of the exonic mutation burden after DNA replication and, ultimately, in an increase of fitness.

Our results demonstrate that the decreased exonic mutation burden is not due to negative selection in the generation of cancer somatic mutations across all tumor types analyzed. This finding suggests that the mutational landscape of cancer-related genes is not strongly influenced by negative selection, in agreement with a recent report[41]. Nevertheless, we expect that, in the germ line, purifying selection has a predominant role, filtering out all variants that prevent the development of a viable individual[42]. Given that MMR components are highly conserved across evolution, and that the exonic enrichment for H3K36me3 and other chromatin marks has been observed across species[18,43], it is reasonable to assume that the enhanced exonic MMR observed in human somatic cells is also present in germline cells and in other organisms. Therefore, intronic regions could accumulate more nucleotide changes across evolution as a result not only of intense purifying selection on exonic variants but also of this differential repair. This, in turn, would bring into question the use of rates of intronic substitution (K_i) as a proxy for neutral evolution[44–46], with important implications for understanding of the evolution of genes. Further implications may be extracted for methods aimed at detecting cancer driver genes that model the background mutation rate of exonic elements from their surrounding areas to identify signals of positive selection in the coding region of genes. Some of these methods[5,21,22], which use intronic mutations as estimators of the exonic background mutation rate, may be strongly affected by the differential generation of mutations in these regions.

In summary, we demonstrate that the differential MMR in exons and introns in somatic cells causes the former to harbor fewer mutations than expected from their nucleotide composition. This finding advances knowledge of the interplay between mutational processes and the DNA repair machinery. Moreover, our results have important implications regarding the way we study the forces that shape the development of tumors and the evolution of the genome.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge funding from the Spanish Ministry of Economy and Competitiveness (SAF2015-66084-R, MINECO/FEDER, UE), La Fundació la Marató de TV3, EU H2020 Programme 2014-2020 under grant agreements 634143 (MedBioinformatics) and by the European Research Council (Consolidator Grant 682398). IRB Barcelona is the recipient of a Severo Ochoa Centre of Excellence Award from the Spanish Ministry of Economy and Competitiveness (MINECO; Government of Spain) and is supported by CERCA (Generalitat de Catalunya). R.S. is supported by an EMBO Long-Term Fellowship (ALTF 568-2014) cofunded by the European Commission (EMBOCOFUND2012, GA-2012-600394) with support from Marie Curie Actions. A.G.-P. is supported by a Ramón y Cajal contract from the Spanish Ministry of Economy and Competitiveness (RYC-2013-14554). We acknowledge the contribution of I. Reyes-Salazar to refactoring and cleaning all code produced in the study for publication. We are grateful to B. Campbell and U. Tabori for help in obtaining the mutation calls for bMMRD samples sequenced by the International BMMRD Consortium. The results published here are in part based upon data generated by the TCGA Research Network (http://cancergenome.nih.gov/).

AUTHOR CONTRIBUTIONS

J.F. and R.S. participated in the design and execution of analyses, produced the figures, participated in the interpretation of results and edited the manuscript. L.M. developed computational code employed in the analyses. F.M. developed the statistical framework to compute the significance of the decreased exonic mutation burden and its correlation with chromatin features. A.G.-P. participated in the design of analyses, the interpretation of results, the oversight of analyses, and drafted and edited the manuscript. N.L.-B. conceived the study, participated in the design of analyses, oversaw the study and the interpretation of results, and drafted and edited the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
2. Dulak, A.M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).
3. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
4. Li, J. *et al.* A dual model for prioritizing cancer mutations in the non-coding genome based on germline and somatic events. *PLoS Comput. Biol.* **11**, e1004583 (2015).
5. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
6. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
7. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
8. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).
9. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
10. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
11. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
12. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
13. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
14. 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).
15. Li, F. *et al.* The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell* **153**, 590–600 (2013).
16. Tatum, D. & Li, S. Evidence that the histone methyltransferase Dot1 mediates global genomic repair by methylating histone H3 on lysine 79. *J. Biol. Chem.* **286**, 17530–17535 (2011).
17. House, N.C.M., Koch, M.R. & Freudenreich, C.H. Chromatin modifications and DNA repair: beyond double-strand breaks. *Front. Genet.* **5**, 296 (2014).
18. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon–intron structure. *Nat. Struct. Mol. Biol.* **16**, 990–995 (2009).
19. Huff, J.T., Plocik, A.M., Guthrie, C. & Yamamoto, K.R. Reciprocal intronic and exonic histone modification regions in humans. *Nat. Struct. Mol. Biol.* **17**, 1495–1499 (2010).
20. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).

21. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
22. Lanzós, A. *et al.* Discovery of cancer driver long noncoding RNAs across 1112 tumour genomes: new candidates and distinguishing features. *Sci. Rep.* **7**, 41544 (2017).
23. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
24. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
25. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
26. Rubio-Perez, C. *et al. In silico* prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27**, 382–396 (2015).
27. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).
28. Hause, R.J., Pritchard, C.C., Shendure, J. & Salipante, S.J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22**, 1342–1350 (2016).
29. Shlien, A. *et al.* Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat. Genet.* **47**, 257–262 (2015).
30. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
31. Marteijn, J.A., Lans, H., Vermeulen, W. & Hoeijmakers, J.H.J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15**, 465–481 (2014).
32. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
33. Adar, S., Hu, J., Lieb, J.D. & Sancar, A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl. Acad. Sci. USA* **113**, E2124–E2133 (2016).
34. Hu, J., Adar, S., Selby, C.P., Lieb, J.D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29**, 948–960 (2015).
35. Haradhvala, N.J. *et al.* Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
36. Tolstorukov, M.Y., Volfovsky, N., Stephens, R.M. & Park, P.J. Impact of chromatin structure on sequence variability in the human genome. *Nat. Struct. Mol. Biol.* **18**, 510–515 (2011).
37. Francioli, L.C. *et al.* Genome-wide patterns and properties of *de novo* mutations in humans. *Nat. Genet.* **47**, 822–826 (2015).
38. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
39. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017).
40. Kim, S., Kim, H., Fong, N., Erickson, B. & Bentley, D.L. Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proc. Natl. Acad. Sci. USA* **108**, 13564–13569 (2011).
41. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* http://dx.doi.org/10.1016/j.cell.2017.09.042 (2017).
42. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. USA* **107**, 961–968 (2010).
43. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41**, 376–381 (2009).
44. Chamary, J.V., Parmley, J.L. & Hurst, L.D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108 (2006).
45. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
46. Hoffman, M.M. & Birney, E. Estimating the neutral rate of nucleotide substitution using introns. *Mol. Biol. Evol.* **24**, 522–531 (2007).

ONLINE METHODS

Whole-genome expression and mutation data. Whole-genome somatic mutations and expression data for 38 skin cutaneous melanomas (SKCM), 46 lung adenocarcinomas (LUAD), 45 lung squamous cell carcinomas (LUSC), 42 colorectal adenocarcinomas (CRC), 96 breast carcinomas (BRCA), 21 bladder carcinomas (BLCA), 47 uterine corpus squamous cell carcinomas (UCEC), 27 glioblastomas (GBM), 18 low-grade gliomas (LGG), 20 prostate adenocarcinomas (PRAD), 34 thyroid carcinomas (THCA) and 27 head and neck squamous cell carcinomas (HNSC) probed by TCGA were obtained from Fredriksson *et al.*[20]. Cohorts of tumors with fewer than 5,000 genic mutations or fewer than 1,000 exonic mutations (HNSC, GBM, KIRC, THCA, LGG and PRAD) were discarded from the analysis. The somatic mutations detected in four bMMRD pediatric glioblastomas sequenced by the International BMMRD Consortium[29] were obtained through personal communication from the authors. Finally, we obtained the somatic mutations detected across the whole genome of a normal skin sample from Martincorena *et al.*[30].

Genomic coordinates of exons and introns. GENCODE[47] v19 coordinates for 20,345 protein-coding genes were retrieved. Genes without introns, overlapping genes and cancer driver genes, according to the Cancer Gene Census and other sources[25,26], were discarded, leaving a filtered set of 12,104 genes. All transcripts per gene were merged, generating meta-exon and meta-intron coordinates. Finally, 5′ and 3′ exons were removed, as well as all UTRs (except for the analysis shown in **Fig. 1**), thus leaving only internal exons and their flanking introns. We then identified all genic regions where mutation calling would be technically challenging because of low sequence complexity, ambiguous mappability of sequencing reads or low sequencing coverage. Regions of low complexity or low mappability were obtained from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability). The former included repetitive regions defined by RepeatMasker, while the latter comprised regions with low unique mappability for 36-mer sequences (CRG Alignability 36′ track, score <1). Finally, regions covered by fewer than eight reads in any of five randomly selected tumor samples of each tumor type (the requirement to make somatic calling in Fredriksson *et al.*[20]) were considered of low coverage. Regions of any of these three types were removed from introns and exons.

Clusters of tumors with different somatic mutational processes. To group the tumors of each cancer type in the cohort according to their underlying mutational processes, we first built a matrix of the frequencies of the 96 trinucleotide changes across tumors, as in a previous work[12]. We carried out hierarchical clustering (using a Euclidean distance and average method to compute the similarity between clusters) of this matrix. We then manually separated the clusters of tumors and identified their underlying mutational processes through visual comparison with previously obtained[32] mutational signatures across cancer types. Clusters of tumors with fewer than 5,000 genic mutations or fewer than 1,000 exonic mutations were discarded for downstream analyses.

Chromatin features. We downloaded peak (narrow) coordinates and genome-wide read coverage for the 32 chromatin features presented in **Supplementary Table 1** across 127 cell lines and primary cell types from Roadmap Epigenomics[23] and the nucleosome density from ENCODE[24]. Peaks and reads (see below) obtained from http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak and http://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated, respectively, for each feature were mapped to intronic and exonic regions of genes. The primary cell types closest to colorectal tumors and glioblastomas in Roadmap Epigenomics were selected to represent the exon–intron distribution of chromatin features. Genome-wide nucleosome positioning signals (density graphs) for the ENCODE cell line GM12878 (lymphoblastoid cell line) were obtained via the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/). Further, by using the bwtool find program (with parameters local-extrema -maxima -min-sep = 150), nucleosome peak regions were identified across the genome, and the 146-bp regions flanking each peak (73 bp per side) were considered as regions covered by a nucleosome.

We numbered the exons and introns in each gene according to their positions with respect to the transcriptional start site (TSS). Exons and introns that occupied different positions in different transcripts and those in the lower quartile of length were discarded. We then stacked all exons and all introns separately and computed the aggregated coverage (fraction of bases covered by peaks for each mark) at the center of the stack corresponding to the number of bases of the shortest exon or intron remaining after filtering. Finally, the difference between exonic and intronic coverage was computed via the two-tailed Mann–Whitney *P* value for the comparison of the two distributions.

Classification of colorectal tumors according to MMR level. Colorectal samples were separated into four subtypes on the basis of MMR levels. MSI-H (n = 6), MSI-L (n = 4) and MSS (n = 26) groups were defined on the basis of clinical information from TCGA (https://portal.gdc.cancer.gov/query). The *POLE*-mutant group (n = 6) was defined by identifying samples with missense mutations of the *POLE* (DNA polymerase ε) gene.

Exon-centered mutational analyses. We stacked 2,001-nt sequences centered on the middle position of internal exons. In this analysis, we did not exclude regions that overlapped any of the three types of technically challenging regions mentioned above. Thus, we obtained a stack of 95,164 sequences centered on exons. We then counted the observed and expected (distributed across each sequence of the 2,001-nt window following the mutational probability for each nucleotide, as explained below for individual genes) mutations associated with each nucleotide of these sequences. With these counts across the selected windows, we produced exon-centered plots as shown in **Figures 2a** and **3b**.

Computing decrease in exonic mutation burden. We first computed the relative frequencies of the 192 possible trinucleotide changes, $f(A_iX_jC_k{\rightarrow}A_iX_lC_k)$, across each cluster of tumors as

$$f(A_iX_jC_k \rightarrow A_iX_lC_k) = \frac{N(A_iX_jC_k \rightarrow A_iX_lC_k)}{T}$$

where $N(A_iX_jC_k{\rightarrow}A_iX_lC_k)$ was the number of such changes among all mutations observed in the tumors and T was the total number of substitutions observed across tumors. Then, we made f relative to the abundance of each trinucleotide in the genome, $G(A_iX_jC_k)$.

$$\bar{f}(A_iX_jC_k \rightarrow A_iX_lC_k) = \frac{f(A_iX_jC_k \rightarrow A_iX_lC_k)}{G(A_iX_jC_k)}$$

Next, for each genic site, we summed the relative frequency of its three possible changes given its 5′ and 3′ flanking bases.

$$\overline{f_{\text{site}}}(A_iX_jC_k \rightarrow A_iX_lC_k) = \sum_{l=0}^{2} \bar{f}(A_iX_jC_k \rightarrow A_iX_lC_k)$$

We rescaled the relative frequency of change for each site to 1 by multiplying each frequency by factor k.

$$k = \frac{1}{\sum \overline{f_{\text{site}}}}$$

The rescaled frequency (Rescf) of each nucleotide in the gene is proportional to the conditional probability that the reference nucleotide changes to the alternative given its 5′ and 3′ nucleotides. Finally, for each independent gene, we redistributed all observed mutations (N_{muts}) across exonic and intronic sites following these summed rescaled frequencies of each site to be mutated.

$$\text{EExonic} = N_{\text{muts}} * \sum \text{Rescf}_i \;\; \forall\, i \in \text{exonic sites}$$

$$\text{EIntronic} = N_{\text{muts}} * \sum \text{Rescf}_i \;\; \forall\, i \in \text{intronic sites}$$

Note that this redistribution process could be performed equivalently for the mutations observed in one tumor (for single-tumor analysis; **Fig. 5b**) or across a group of tumors (for group or cluster analyses; **Figs. 2**, **3**, **4** and **5a**). The process yielded the number of expected exonic (EExonic) and intronic (EIntronic) mutations in the gene. (We employed a second method to compute the expected number of exonic mutations based on the average of 1,000 random permutations of the observed mutations in each gene following the probability of each site to acquire a mutation (**Supplementary Table 3**).)

Summing the observed and expected exonic mutations over all genes, we computed the difference between the observed and expected numbers of exonic mutations, which we refer to as the decrease in exonic mutation burden (as in most tumors there was a negative difference). Throughout the paper, we express this decrease as the percentage of the total number of observed exonic mutations.

To compute the significance of this decrease, we employed two tests: (i) a *G* test of independence comparing the numbers of observed and expected mutations in exons and introns, under the null hypothesis that the observed and theoretical distributions of the variables are equal, and (ii) for the expected number of exonic mutations computed using the permutations approach, we computed an empirical *P* value as the fraction of the iterations with fewer expected than observed exonic mutations.

Test for negative selection on exonic mutations. The consequence type of all observed exonic mutations was obtained using the Ensembl Variant Effect Predictor[48] (VEP; v.70). We subsequently separated exonic mutations into two groups: those with synonymous consequence and those with a consequence ranking higher than synonymous in the Ensembl Variation hierarchy (http://www.ensembl.org/info/genome/variation/predicted_data.html), which were collectively deemed nonsynonymous. All possible nucleotide changes in a gene were then divided into three categories: (i) synonymous; (ii) nonsynonymous (with the consequences defined above); and (iii) intronic. We redistributed the mutations observed in each gene across these three types of sites following the probability of occurrence of each change computed as explained above. Through the difference between observed and expected synonymous and nonsynonymous mutations, we were able to compute the decrease in the burden of both types of mutations (expressed as the percentage of the expected number, as explained above for all exonic mutations). Finally, a *G* test of independence was used on the null hypothesis that fewer nonsynonymous mutations should be observed than expected.

We also redistributed only the exonic mutations across synonymous and nonsynonymous sites according to the probability of change for each type of site. In this case, we used the *G* test of independence on the null hypothesis that the number of expected nonsynonymous mutations was not smaller than the observed number.

Stratification of genes by mutation rate and several covariates. The mutation rate of each gene was computed as the quotient between the number of observed mutations and the number of bases in the gene. Genes were subsequently grouped into ten bins according to their mutation rate.

We computed the 75th percentile of the expression of each gene across the tumors in each cohort. Genes with a 75th percentile of expression equal to 0 were considered to be non-expressed and were grouped together. All other genes were sorted on the basis of their previously computed expression percentile and divided into nine bins of equal size. Non-expressed genes were subsequently added as a tenth bin.

Replication time data across the human genome measured in lymphoblastoid cell lines were obtained from Koren *et al.*[6]. Using these data, the mean replication time per gene was computed. Next, genes were sorted on the basis of this value and divided into ten groups of equal size.

Finally, we also grouped the genes into ten bins according to H3K36me3 peak coverage.

Relationship between decreased exonic mutation rate and exonic enrichment of nucleosomes and histone marks. For each gene, we computed the read-count-based exonic enrichment for any chromatin feature as the ratio between the exonic and intronic read counts (total number of bases covered by reads of the chromatin feature). (This read-count-based exonic enrichment was used to compute the correlations shown in **Fig. 4** and **Supplementary Fig. 4**.) We computed the peak-based exonic enrichment of any chromatin feature as the ratio of exonic and intronic bases covered by peaks of the feature (to compute the correlation shown in **Supplementary Fig. 5**). The exonic and intronic numbers of bases covered by reads or peaks of the chromatin feature for colorectal and bMMRD glioblastoma tumors were computed from colonic mucosa (E075) and brain angular gyrus (E067) cells, respectively, both obtained from Roadmap Epigenomics. (In the case of nucleosomes, peaks were obtained from occupancy values as explained above.) Genes were grouped into 10, 25 or 50 bins according to their exonic H3K36me3 enrichment, and the aggregated decrease in the exonic mutation rate of the genes in each bin was computed as explained above for colorectal *POLE*-mutant, MSI-H and bMMRD tumors. We then computed the correlation between the median exonic chromatin feature enrichment and the decreased exonic mutation rate across the bins. The trendline and its confidence intervals were added to each plot using the bootstrapping functions of the Python seaborn package, which confers equivalent weights in the regression to all points. To guarantee that the trend was not the result of a few outliers, the correlation coefficient and its significance were computed using an iteratively reweighted least-squares approach, letting the variance in exonic H3K36me3 enrichment of the bins influence the weight of each point.

Exon-to-intron mutation rate ratio. As described above, we stacked all exon-centered and intron-centered sequences. Then, we averaged the total number of mutations observed at each of the 41 central positions of each stack. The selection of 41 central positions guaranteed both a vast majority of exonic sequences contributing mutations and enough mutations for calculation across all clusters at exon-centered stacks. The exon-centered and intron-centered mutation burden averages were then divided by the number of sequences included in each stack to make them comparable. Finally, we computed the exon-to-intron mutation rate ratio as the quotient between the corrected exon-centered and intron-centered mutation burden averages.

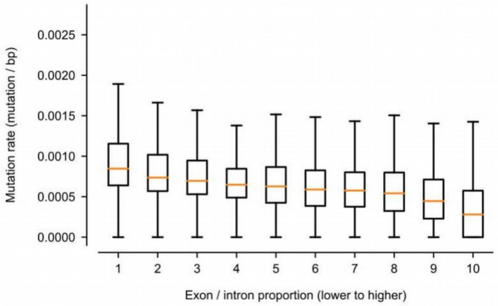
Computing the activity of NER from XR–seq data. Genome-wide maps of NER for two UV-induced photoproducts, namely cyclobutane pyrimidine dimers (CPDs) and pyrimidine–pyrimidone (6,4) photoproducts (PP64s), in irradiated skin fibroblast cell lines were obtained from Hu *et al.*[34]. This data set comprises NER maps for the following three cell lines: (i) wild-type NHF1 skin fibroblasts, which have active global and transcription-coupled repair mechanisms; (ii) XP-C mutants, which are deficient in the global repair mechanism; and (iii) CS-B mutants, which are deficient in transcription-coupled repair. For each of these cell lines, we extracted the sequencing reads, processed and mapped to the human genome, following the steps mentioned in Hu *et al.* Further, we selected the mapped reads that were 26 nt in size, which is the typical size of NER-excised oligomers, and classified the reads on the basis of the presence of dipyrimidines (TT, CT, TC and CC) at positions 19–20 or 20–21 of the reads. In addition, we recorded the mapped genomic locations of the nucleotides in positions 19–20 or 20–21 of the reads. This way, we could predict the damage site according to the excised fragments. We mapped this information to the XR–seq exon-centered plot together with the frequencies of the dipyrimidines observed in each column (**Supplementary Fig. 7**).

Statistics. We used the *G* test described above to compute *P* values to test the significance of decreased exonic mutation burden for groups of genes or all genes in a tumor or across groups or clusters of tumors. (All *P* values computed for all comparisons are provided in **Supplementary Table 3**, which also includes *P* values computed using a permutation-based test also described above.) When appropriate, *P* values computed with this test were corrected using the Benjamini–Hochberg approach. In **Figure 1**, we used the two-tailed Mann–Whitney test to compare the exonic and intronic distributions of chromatin features (and corrected them when appropriate). Above, we describe the approach employed to compute the correlation coefficient (and its associated *P* value) for the regression lines shown in **Figure 4** and **Supplementary Figures 4** and **5**.

Code availability. All code needed to reproduce the analyses described in the paper are available at the Bitbucket repository (https://bitbucket.org/bbglab/intron_exon_mutrate).

Data availability. Preprocessed data needed to reproduce all analyses described here are provided together with the code at https://bitbucket.org/bbglab/intron_exon_mutrate. A **Life Sciences Reporting Summary** is available.

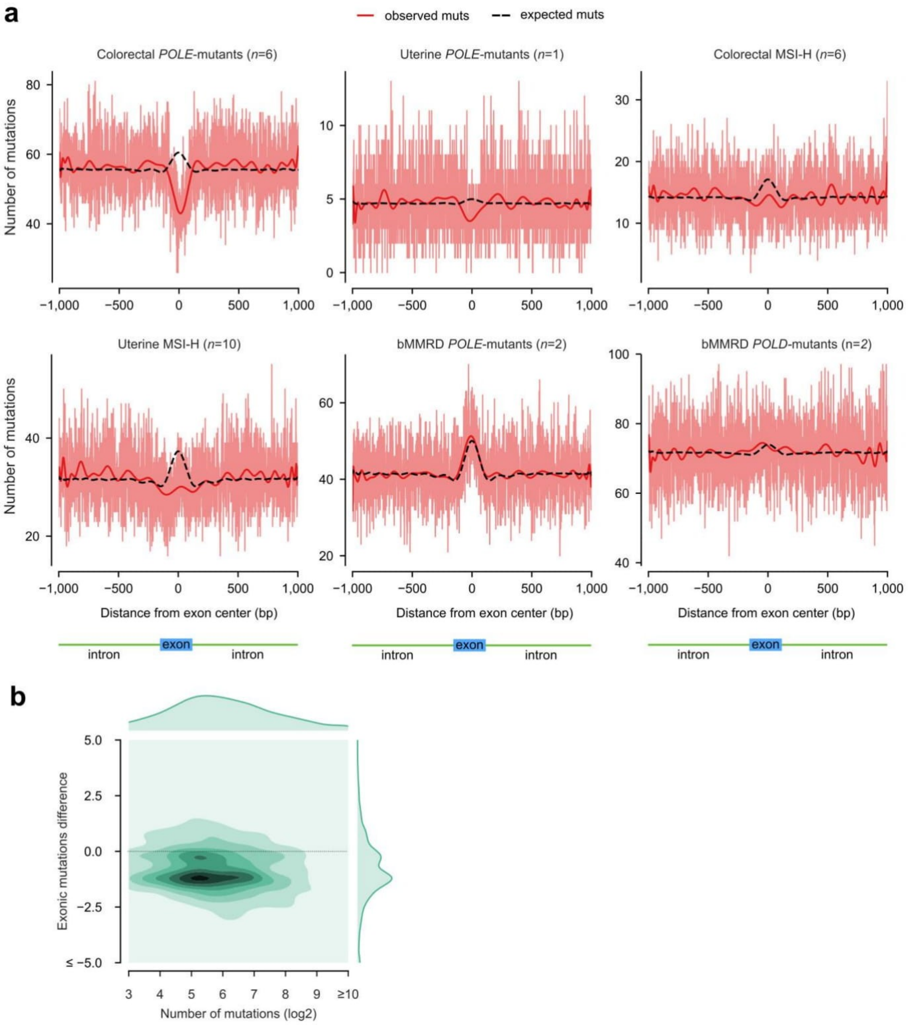
47. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
48. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).



Supplementary Figure 1

Relationship between the exon/intron proportion of genes and their mutation rate.

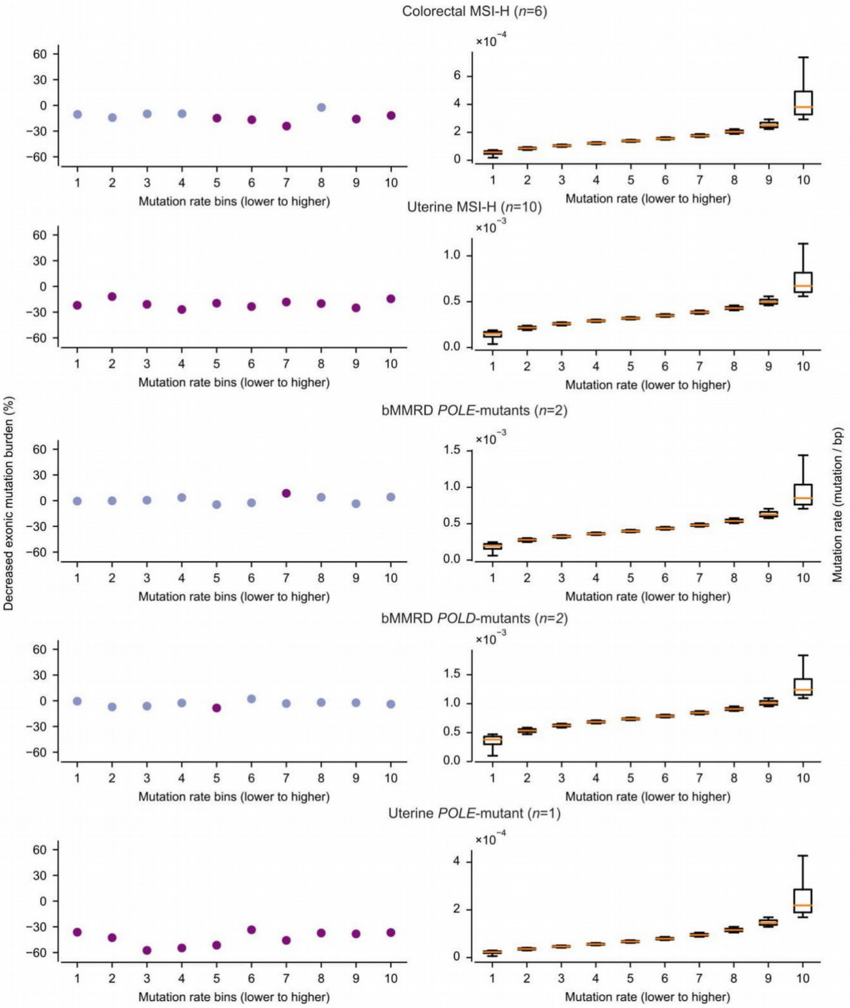
Each box plot represents the distribution of mutation rate (in colorectal *POLE*-mutant tumors) of genes in deciles of increasing exon to intron proportion. This relationship implies that computing a mutation rate in exons and introns by aggregating genes with different mutation rates and exon to intron proportions would create an artifact. To avoid this artifact, we compute the expected mutations in exons and introns for each gene (**Fig. 2**).



Supplementary Figure 2

Permutation-computed expected mutation rate.

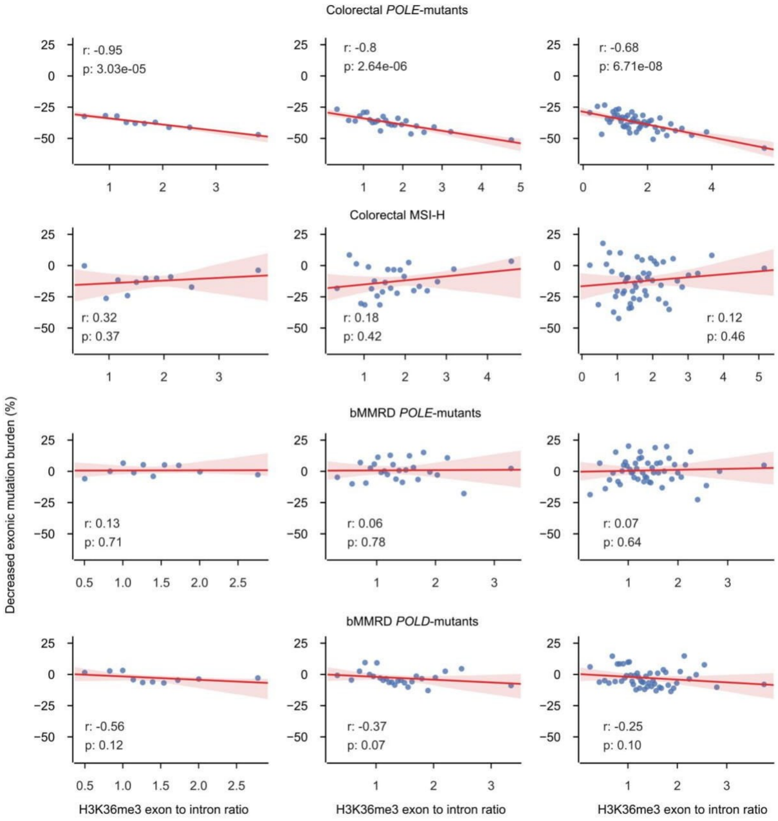
(**a**) Exon-centered plot of observed and expected mutation rates for several clusters of tumors: top row, colorectal and uterine *POLE* mutant tumors; top and bottom rows, colorectal and uterine MSI-H tumors; bottom row, bMMRD *POLE*- and *POLD*-mutant tumors (analogous to **Figs. 2a** and **3a**). The expected mutation rate was computed using 1,000 random permutations of mutations in each sequence of the stack, as described in the Online Methods. (**b**) Density plot representing the difference between the observed and expected number of exonic mutations in individual genes with at least one expected exonic mutation (similar to **Fig. 2c**).



Supplementary Figure 3

Decreased exonic mutation burden across groups of genes with increasing mutation rate.

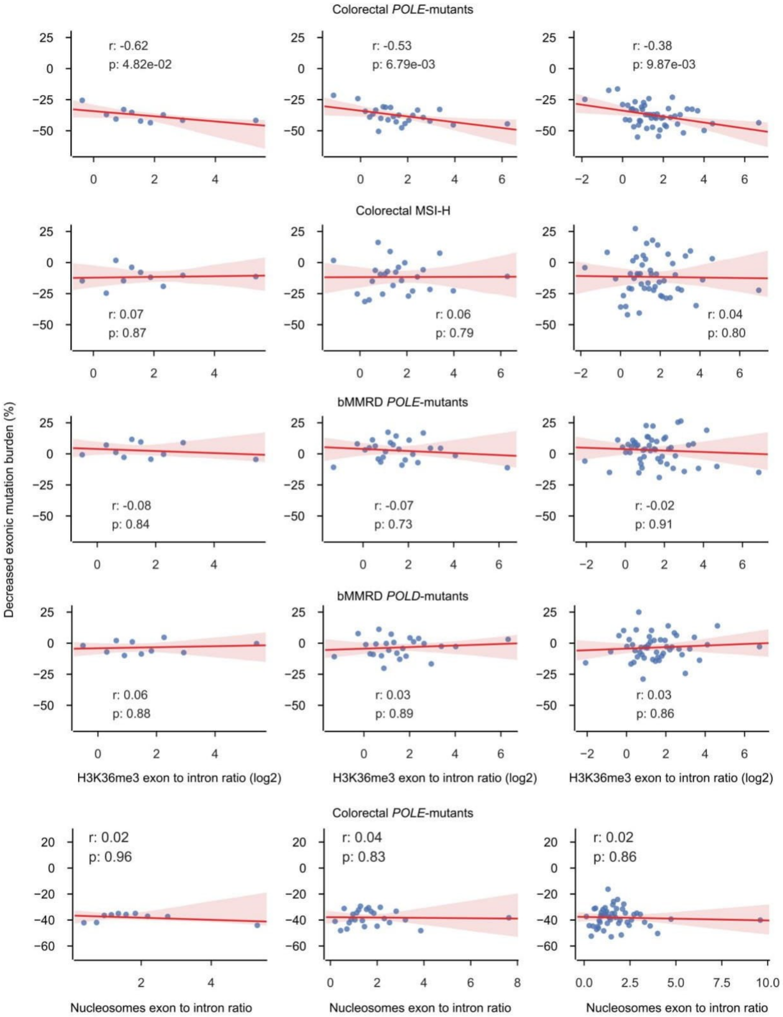
Graphs to the left present the decreased exonic mutation burden of groups of genes of increasing mutation rate (circles in the graphs). In each graph, groups of genes with significantly decreased exonic mutation burden are colored in purple, while those with non-significantly decreased exonic mutation burden are colored in blue. Graphs to the right present the distribution of mutation rates for each group of genes.



Supplementary Figure 4

Relationship between decreased exonic mutation burden and exonic enrichment of H3K36me3 (read count based) across clusters of tumors.

From top to bottom, the rows represent the relationship between exonic enrichment of H3K36me3 (read count based as explained in the Online Methods) and decreased exonic mutation burden in colorectal *POLE*-mutant tumors, colorectal MSI-H tumors, bMMRD *POLE*-mutant tumors and bMMRD *POLD*-mutant tumors. From left to right, the correlations have been computed grouping the genes into 10, 25 and 50 bins (similar to **Fig. 4**).

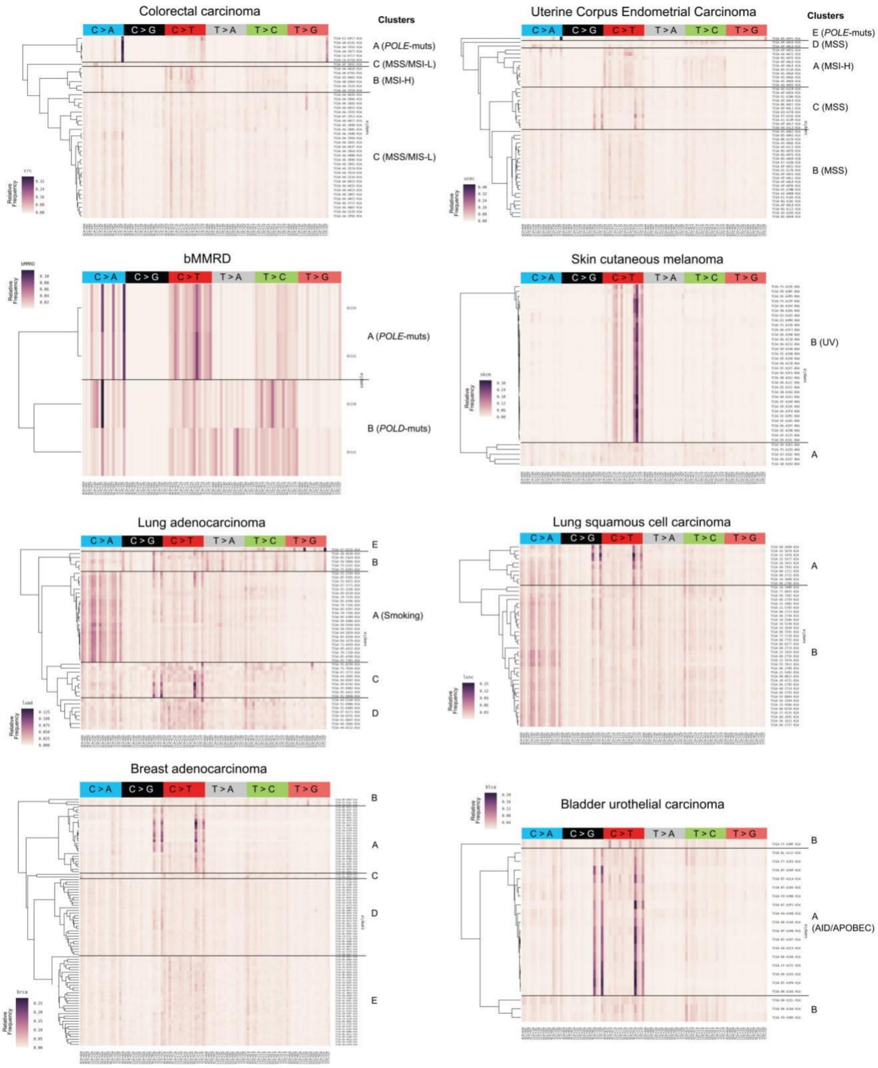


Supplementary Figure 5

Relationship between decreased exonic mutation burden and exonic enrichment of H3K36me3 (peak based) across clusters of

tumors.

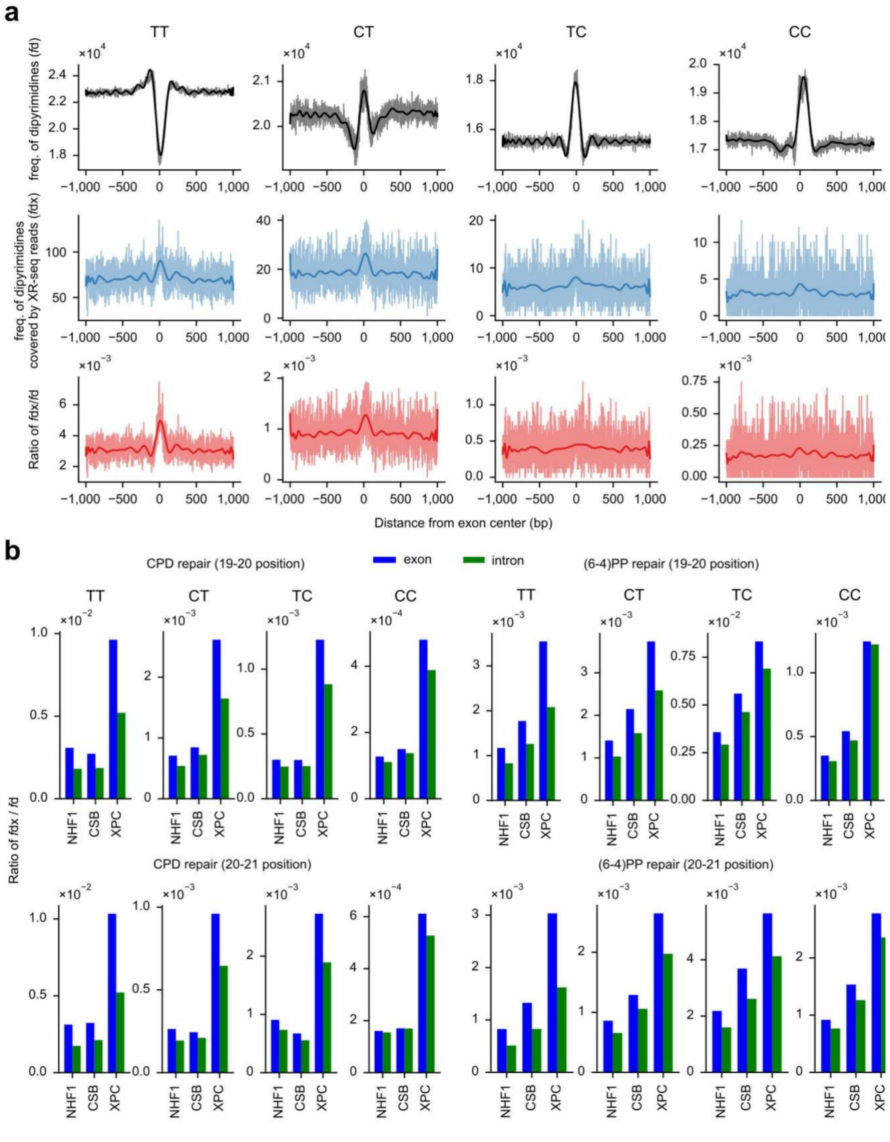
From top to bottom, the rows represent the relationship between exonic enrichment of H3K36me3 and decreased exonic mutation burden in colorectal *POLE*-mutant tumors, colorectal MSI-H tumors, bMMRD *POLE*-mutant tumors and bMMRD *POLD*-mutant tumors. The exonic enrichment for H3K36me3 (exon to intron ratio) has been calculated from H3K36me3 peaks detected in the closest representative in Roadmap Epigenomics to the cell of origin of each tumor. From left to right, the correlations have been computed grouping the genes into 10, 25 and 50 bins (similar to **Fig. 4**). We corroborate that the same negative correlation is obtained between exonic enrichment for H3K36me3 and decreased exonic mutation rate if peaks instead of ChIP–seq reads (as in **Fig. 4**) are used to compute the former. The three panels at the bottom represent the correlation between the exon to intron ratio of nucleosomes (computed from nucleosome occupancy; Online Methods) and the decreased exonic mutation rate in *POLE*-mutant tumors.



Supplementary Figure 6

Clustering of tumors of different cancer types according to their mutational signatures.

The approach to build the clusters is described in the Online Methods.


a
TT
CT
TC
CC
freq. of dipyrimidines (fd)
×10^4
freq. of dipyrimidines covered by XR-seq reads (fdx)
Ratio of fdx/fd
×10^-3
−1,000 −500 0 500 1,000
Distance from exon center (bp)
b
CPD repair (19-20 position)
exon
intron
(6-4)PP repair (19-20 position)
CPD repair (20-21 position)
(6-4)PP repair (20-21 position)
Ratio of fdx / fd
NHF1
CSB
XPC

Supplementary Figure 7

Exon-centered nucleotide-excision repair efficiency.

(**a**) The first row shows the exon-centered frequency of dipyrimidines (TT, TC, CT and CC, from left to right) favorable for UV-induced damage. The second row represents the frequency of dipyrimidines that are covered by at least one XR–seq read (with matching dipyrimdine at positions 19–20 of the reads) for CPD repair in NHF1 cell lines. The last row represents the ratio of the two previous rows. (**b**) Quantification of the ratio computed above for exonic and intronic dipyrimidines across three cell lines with different NER pathways at play, both for CPDs and 6,4 photoproducts (see Online Methods for details).

Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites

This second chapter of the results section describes a research project designed as a follow up of a previous study conducted in our laboratory in 2016 by Sabarinathan, R. et al. [142] showing that active binding sites of transcription factors are mutational hotspots in UV exposed samples because of impaired nucleotide excision repair activity in these areas. In this direction, two works [144] [145] describing increased UV damage susceptibility in the binding motifs of the ETS family transcription factors were published in 2018. These previous reports motivated us to carry out a detailed and comprehensive study of the transcription factor binding sites across all families of transcription factors to settle the contribution of both damage generation and repair on the observed increased mutation rate at these genomic regions. Therefore, we proceeded to evaluate the influence exerted by transcription factors of different families in both UV damage formation and repair to better understand the mutational biases observed in their binding sites and to which extent these biases are shared across transcription factor families.

This research project involves the analysis of whole genome somatic mutations from UV exposed melanomas[150]. This allowed us to study the distribution of somatic mutations in and around active transcription factor binding sites. The analysis of the UV induced DNA damage and its repair required the use of whole genome UV DNA damage maps generated by Dr. Sancar[152] and Dr. Wyrick[144] laboratories.

Frigola J, Gonzalez-Perez A, López-Bigas N. Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites. In preparation

Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites

Introduction

Living cells are exposed to agents with the potential to damage their DNA [1] [2] [3]. The lesions that result from this damage which are not dealt with by repair mechanisms at the time of replication may result in mutations [4] [5]. A repertoire of such repair programs are involved in the correction of different types of lesions [6] [7] [8]. For example, the exposure to UV light, a well known mutagen may cause adjacent pyrimidines to form abnormal covalent bonds creating dipyrimidine adducts, the most abundant type of which are cyclobutane pyrimidine dimers (CPDs [9]). These lesions are preferentially detected and repaired by the nucleotide excision repair machinery (NER [3]). Unrepaired bulky CPDs often block the advance of the replication fork, thus triggering the recruitment of translesion polymerases, capable of bypassing the block, sometimes at the cost of incorrect nucleotide insertion [10]. In the case of UV-induced damage, adenines are inserted opposite CPDs. Moreover, cytosines in CPDs are more prone to undergo spontaneous deamination than cytosines in B-DNA [11]. Both processes thus result in C-G>T-A mutations in cytosines within dipyrimidine contexts, leaving a distinctive mutational pattern known as the UV mutational signature (COSMIC signature 7 [12]).

The distribution of mutations along the genome that results from DNA-damage processes, such as exposure to UV light varies depending on both, large and small-scale genomic features. For example, chromatin regions with different replication time or genes with different level of expression [13] [14] receive different rates of mutations. Chromatin features that affect the mutation rate at the local scale include nucleosomes [15] and active transcription factor binding sites (TFBS)[16] among others. The rate of mutations in TFBS in certain cell types, such as melanocytes (and the derived tumor type melanomas), are abnormally high in comparison to that observed in their flanking regions [16]. Several studies have reported an unexpected increase of the mutations in active TFBS in melanomas linked to a reduced accessibility of the NER machinery to these regions [16] [17]. More recently an abnormal increase in the formation of CPDs in the active binding sites of the ETS TFs has been reported as a cause of the increase mutation rate in TFBS [18] [19]. The relative contribution of these two factors to the increase rate of mutations at the binding sites of different TFs, and across the whole TFBs genomic compartment in melanomas remains unclear. To answer this question, we employed single-nucleotide-resolution CPD formation data to study the formation of the damage, and its repair, within and around 64 types of binding motifs of 48 individual TFs of 10 different families. For each TF we have identified the contribution of the DNA damage and its repair to the distribution of mutations along their binding sites. Our results reveal a complex panorama of damage formation and repair across the regions defined by the binding of the TF, which varied between families of TFs.

Results

Variable mutation rate increase in the binding sites of TFs of different families

We first sought to explore the variability of mutation rate increase of binding sites for different families of TFs. We collected 598,987 TFBS grouped into 64 types of binding motifs of 48 TFs of 10 different families, active in melanocytes and fibroblasts. The binding sites were marked by specific TF ChIPseq peaks (GTRD database [20]) which overlapped DNAse hypersensitivity sites (DHS) detected in both tissue types. We determined the exact location of the binding motif (described by a JASPAR position weight matrix ref)[21] within each peak using the MOODS matching algorithm [22]. (All analyses described below were carried out for TF binding motifs with at least 5000 active instances.)

In order to study the influence of the TF on the distribution of UV-induced mutations across the entire nucleosome-free area, we first defined a region of DNA spanning 1000 nucleotides at both sides of the center of the TF binding motif (Fig. 1a). Within this 2001-nucleotides wide window, we distinguished four areas. The first region corresponding to the binding motif itself (hereinafter motif), comprised the 21 central nucleotides in direct contact with the DNA-binding domain of the TF. Secondly, we denoted the region spanning 50 nucleotides at each side of the center of the window (including the motif) as TFBS. We called the region immediately flanking the TFBS (200 nucleotides on each side), considered to consist mostly of protein-free DNA, DHS flanks. The remaining 1500bp outside the DHS flanks were labeled flanks. These four regions were thus defined in each of the 598,987 collected active TFBS.

From a cohort of 136 UV-exposed melanomas [23] we obtained the mutations overlapping these 598,987 2001-nucleotide wide regions. We stacked the regions corresponding to binding sites of the same type of motif of a given TF and thus summed the mutations observed across each of the 2001 positions (red line in Fig. 1a). Using all the mutations identified in the whole-genomes of the 136 melanomas, we also computed the frequency of each tri-nucleotide context change. We derived an expected mutational frequency at each position of the stacked sequences by randomizing the mutations observed in each of them following these tri-nucleotide substitution frequencies (black line in Fig. 1a). Figures 1a-d present the observed and expected mutation rates across the stacked 2001-nucleotide sequences of four different binding motifs of the TFs SP1, JUND, CTCF, and FLI1 (all analyzed TF motifs appear in Fig. S1). They represent a variety of patterns of mutation rate, including broad (extending beyond the TFBS) and very narrow higher-than-expected mutation rates, as well as lower-than-expected mutation rate (the binding motif of JUND).

The behaviour of the represented motifs of these four TFs is paradigmatic of their three families (SP1 and CTCF: C2H2 zinc fingers; JUND: basic leucine zipper; FLI1: tryptophan cluster; Fig. 1e). A general higher-than-expected (positive log2 fold-change in the figure) rate of mutations is observed across the motif of TFs of most families. The motifs of most basic leucine zipper and fork head TFs, on the other hand, exhibit significantly lower than expected mutation rate. In most instances, the excess of the mutation rate over the expected value in the TFBS is smaller than that observed at the motif, i.e., producing a very narrow peak. In some TFs, notably SP1, this reduction is less abrupt, corresponding to rather broad peaks of observed mutations. In the DHS flanks of TF motifs only smaller increases or decreases of the mutation rate with respect to the expectation are observed.

The increase (or decrease) of the mutation rate in the motifs, with respect to the expectation is hence a feature common to several TF families. We next zoomed-in on the motif area to explore whether the deviations of the mutation rate from the expectations are roughly distributed across dipyrimidine sites or driven by just a few of them, as recently observed for the ETS1 binding site [18] [19]. We focused on dipyirimidines that are present in at least 50% of the instances of each binding motif (Fig. 2). For both positions of these dipyrimidines, we computed the number of mutations observed across melanomas, and compared them to their expected numbers, based on random samples of the same tetranucleotides (the dipyrimidine and its immediate flanks) from the flanks of the corresponding 2001-nucleotide sequence (Methods). Finally, we subtracted the expected mutations at each position from their observed number, and made the difference relative to the expected (percent of mutations over expected).

First, we observed that for many TFs --as in the case of ETS1-- the increase or decrease in the number of UV-induced mutations within the motif with respect to the expectation is mainly driven by one or few dipyrimidines (Fig. 2a-d and Fig. S2). Interestingly, the range of variation in the number of observed mutations in dipyrimidines is larger than the range of variation of the number of expected mutations at the same positions (Fig. 2e). This means that features other than sequence context influence the rate of mutations at specific motif positions, pointing at the influence of bound TFs on the formation of UV-induced photoproducts and/or their repair.

In summary, there is a general increase in the mutation rate at the sites of bound TFs, with some exceptions that exhibit a decrease. The extent and direction of variation in mutation rate depends on the family of TFs. In some families the effect of this influence is not limited to deviations (positive or negative) of the mutation rate with respect to the expectations at the binding motif, but extends to the entire TFBS.

Increased CPDs formation within TFBS of tryptophan cluster family TFs

Deviations in the mutation rate at the binding sites of specific TFs with respect to the expectation may be driven by their influence on the formation of UV-induced lesions [18] [19] and/or on the efficiency of the repair of such lesions by NER [16] [17]. Thus, we next analyzed the influence of the TF on the formation of CPDs in TFBS. As with mutations, we mapped experimentally generated CPDs in the whole genome of irradiated fibroblasts [18] (Fig. S3) across the 2001-nucleotide wide sequences in the stack of each TF motif (ambar line in Fig. 3a). CPDs generated in the naked DNA of each binding site in a parallel experiment were used to compute their expected rate across each TF motif (dark gray line in Fig. 3a). To correct for experimental variability, the observed and expected CPDs at each individual position were normalized by the total number across the whole 2001-nucleotide window in either experiment.

Figures 3a-d (and Fig. S4) present the distribution of CPDs formation rate across binding motifs of TFs of different families. While higher-than-expected rate of CPDs are formed by UV light in the binding motif of FLI1, in the other three cases shown in Figures 3a-e, the rate of CPDs formed when the binding site is active is lower than for naked DNA. Indeed, as a rule TFs of the tryptophan cluster family exhibit significantly higher-than-expected CPDs formation within their binding motif (Fig. 3e). This is probably attributable to the influence of the TF bound to the motif, which has been demonstrated in the case of ETS1 [18]. On the other hand, TFs of other families show lower-than-expected (with many significant instances) CPDs formation within the motif. This suggests that in some instances the bound TF could have a protective effect on its binding sites.

Zooming-in on specific dipyrimidines in the motif (following the same approach as with mutations; Fig. S5), the deviations from the expected rate of CPDs was lower than that computed for the mutation rate. Only few motifs presented large higher- (some tryptophan cluster TFs) or lower-than-expected (some C2H2 zinc finger and basic leucine zipper TFs) rates of CPDs formation. No clear pattern of deviations of CPDs formation from the expectation across motifs were observed amongst the four dipyrimidine types (Fig. 3f,g).

Summing up, CPDs formation rate is higher than expected in the motifs of TFs of the tryptophan cluster family, which could explain their higher-than-expected rate of mutations. However, for the majority of TF motifs analyzed, the rate of CPDs formation is not above the expectation and thus, it does not explain their increased mutation rate.

Widespread CPDs repair decrease within TFBS

Following the reasoning that the distribution of mutations across a genomic area is determined by the interplay between the distribution of the lesions and the efficiency of their repair, we next focused on the efficiency of repair of CPDs and TFBS. Previous studies have shown that the efficiency of NER in the repair of CPDs within the TFBS is diminished with respect to their flanks, which has been attributed to reduced accessibility to the damage generated by the bound TF [16] [17].

To determine the efficiency of NER across the binding sites of the TFs under study, we used the distribution of CPDs generated immediately after radiation and those remaining 48h after exposure [24]. Thus, we mapped the CPDs experimentally generated at 0h and those remaining at 48h after UV exposure to the 2001-nucleotides wide sequences containing the TF binding motifs. Then, the activity of repair was computed as the percentage of CPDs repaired 48h after exposure (blue line in top panel of Fig. 4a) respect to the number recorded immediately after irradiation (ambar line in top panel of Fig. 4a), and is represented as a gray line in the bottom panel of Figure 4a. Hence, while at the flanks of the motif of SP1 illustrated in Figure 4a some 85% of the CPDs identified at 0h after irradiation have been repaired after 48 hours, at the center of the motif, the fraction of CPDs repaired is less than 30%. In two of the other three motifs represented in Figures 4b-d, the fraction of repaired CPDs at the center of the motif is smaller than at the flanks. The exception is the motif of FLI1 (Fig. 4d), with slightly higher repair rate within the motifs than at the flanks. (Fig. S6 presents the repair efficiency across more exemplary TF motifs.)

Taking then the flanks as the reference of repair activity (Fig. 4e) we are able to obtain an overall view of the efficiency of repair in different regions of the binding sites across TFs of the families under study. For most TFs, the motif is repaired at significantly lower rates than the flanks, with several tryptophan cluster and basic leucine zipper TFs as exceptions. The decreased repair efficiency also dominates in absolute magnitude over the increased repair efficiency of the aforementioned exceptions. If the DHS flanks --which comprise protein-free open chromatin-- are taken as reference of repair activity (Fig. 4f), the widespread decrease of repair activity across motifs and TFBS becomes more apparent. This supports the hypothesis that the decrease in repair efficiency observed for the majority of TFs is caused by the impairment in NER accessibility to the damage at the motif and the TFBS [16] [17].

A similar view is obtained from the zoomed-in analysis of the repair of CPDs at specific positions within motifs (Figs. 4g, S7). CPDs involving dipyrimidines within the motifs of tryptophan cluster TFs tend to be repaired at the same (or slightly-higher) rate than their counterparts at the flanks. On the other hand, some specific positions within the motifs of C2H2 zinc fingers TFs receive lower CPDs repair activity that their flanks. Intriguingly, we observed that CPDs involving CC dipyrimidines tend to occupy lowly-repaired positions (Fig. 4h), while those of TT or TC dipyrimidines were at hilghly repaired positions. This is true along the whole genomic sequence, and is therefore not caused by the binding of proteins to the DNA (Fig. S8). This observation suggests that different dipyrimidines --possible by virtue of differences in the torsion of the B-DNA helix-- are intrinsically repaired at different rates.

The decreased CPDs repair within TF motifs is, in conclusion, widespread across TFs, regardless of their family, with some exceptions.

The generation of UV-induced mutations in TFBS

On the basis of the results of the analyses outlined in the two previous sections, we hypothesized that the decrease of CPDs repair is the most general explanation of the increased UV-induced mutation rate within motifs observed across most TF families.

We thus used the rate of CPDs at 0h in the binding sites of different TFs, and the repair dynamics computed for each of them to predict the fraction of CPDs produced by the irradiation that persisted after 48h (Methods; Figs. 5a and b). (Due to the biases arising from the use of antibodies, we estimate the number of persistent CPDs using inferred repair activity rather than directly using CPDs mapped in the DamageSeq experiment.) If the observed distribution of mutations is in general shaped by the decreased repair, the number of persistent CPDs across TFBS would correlate better with the increased mutation rate than their observed numbers immediately after irradiation. In agreement with this reasoning, we observed a better correlation between the number of 48h (Fig. 5b) than 0h CPDs (Fig. 5a) in the TFBS relative to the flanks and the mutation rate of the motif relative to the flanks (Fig. S9). This supports the notion that a general impairment of NER caused by the protein bound to the DNA drives the observed increased of mutation rate across most TFBS.

If the succession of events (CPDs generation and CPDs repair) leading to the generation of UV-induced mutations on TFBS of different families are considered separately, three patterns with distinct dynamics can be distinguished (Fig. 5c). The first pattern corresponds to tryptophan cluster TFs, the binding sites of which exhibit a higher rate of CPDs formation (0h) than the flanks (upper half in Fig. 5a). The repair of the CPDs in these TFBS is higher than --or similar to-- that in their flanks (see Figs. 4d and e). As a result, after 48 hours higher rate of CPDs persist in the TFBS than in the flanks, leading to the higher mutation rate of the former. The second pattern presents the generation of mutations in TFBS of C2H2 zinc fingers. The rate of CPDs produced after irradiation in the motif (top panel) or the TFBS (bottom panel) are similar to those in the flank (close to x-axis in Fig. 5a). However, due to a decreased repair in comparison to the flanks (see Figs. 4a, c and e) most binding sites of TFs of this family exhibit higher rate of (predicted) CPDs persist at 48 hours after irradiation (upper half in Fig. 5b). This ultimately causes a broad pattern of higher rate of mutations in the TFBS than in the flanks. (Note that some members of this family constitute exceptions to this pattern.) In the third pattern, TFs of the basic leucine zipper family show a similar rate of CPDs formation in the TFBS as in the flanks (close to x-axis in Fig. 5a). The repair rate of these CPDs in the motif is similar to that of the flanks, and higher in the whole TFBS (see Figs. 4b and e). As a result of this increased repair, fewer mutations than in the flank are generated in the binding sites of these TFs (left half of Fig. 5b).

The dynamic interplay between damage and repair within TFBS, is thus different between families of TFs. While the interference of the TF with the rate of formation of CPDs after irradiation influences the resulting mutation rate, its role in obstructing their repair --in particular when the entire TFBS is analyzed-- is determinant in shaping the distribution of mutations.

Discussion

The observation that the burden of mutations in promoter regions --and in particular in TFBS-- in melanomas was unexpectedly high paved the way for the dissection of the detailed influence of the bound TF on the formation of UV-induced lesions [18] [19] and their repair [16] [17]. Different studies discovered that TFs bound may increase the rate of CPDs formation at certain positions of their binding motifs [18] [19] and that they interfere with the accessibility of the NER machinery to CPDs, thus hindering their repair at TFBS [16] [17]. This opportunity to carry out this careful dissection was created in the past few years by the ability to map UV-induced damage and its repair at single-nucleotide resolution, and the availability of the whole-genome sequence of UV-exposed skin tumors [23]. The main motivation of the present work consisted in unraveling the specific contribution of these two roles to the abnormally high burden of UV-induced mutations at TFBS for TFs of different families was the essential motivation of the work presented here.

Although, as aforementioned, we undertook this work with a comprehensive toolbox of maps of UV-induced damage, repair, and mutations, we must first acknowledge its intrinsic limitations. Using antibodies-based experiments in which CPDs are mapped across the genome at different times after irradiation we are able to infer the rate of repair at TFBS, but the maps are possibly biased against the recognition of certain types of CPDs [25] (Fig. S3). On the other hand, whole-genome maps of CPDs computed through a potentially less biased enzyme-based experiment are available only immediately after irradiation. We posit that any existing biases of the antibody-based damage maps are biased out when the CPDs from two time points of the same type of experiment are compared. We carried out the inference of the intensity of repair based on the subtraction of the damage at two time points rather than resorting to the direct mapping of the repair (XR-seq [26]), in order to be able to correct for the rate of the damage immediately after irradiation. This should result in more accurate estimates of repair intensity irrespective of the rate of CPDs generated by the exposure. In XR-seq segments, moreover, the exact location of the CPD is difficult to map.

Due to the possible aforementioned bias, we also indirectly estimate the number of CPDs in different areas of individual TFBS 48 hours after irradiation on the basis of their original numbers estimated through the enzymatic experiment, and the previously inferred repair rate. Whole-genome mapping of CPDs at late time points after irradiation using less biased experiments will support the direct calculation of CPDs rate in different TFBS. A second limitation of the analysis stems from the fact that while CPDs have been mapped in a fibroblasts cell line [24] [18], the source of mutations are melanomas [23], which derive from melanocytes, a different cell type. We face this hurdle by restricting the analysis to TFBS active in both cell types [27]. Finally, CPDs involving a methylated cytosine --which the techniques mapping the damage are unable to distinguish-- are more likely to yield mutations through spontaneous deamination of the base than others. These caveats aside, the estimated rate of CPDs after 48 hours is accurate enough to show a high correlation (r=0.872; Fig. 5b) with the mutation rate at the same sites.

Our observations showed that the UV mutational process in the TFBS --i.e., the distribution of mutations across different areas that results from the interplay between the formation of the CPDs and their repair-- is complex, and varies sharply between families of TFs. First, we ratified our previous observation [16] that the higher-than-expected mutation rate is widespread in melanomas across the binding motifs of TFs. However, with the systematic study carried here we have been able to determine that the binding sites of basic leucine zipper TFs actually exhibit lower-than-expected mutation rate. In the majority of TFs whose binding sites do exhibit an unexpected increase of mutations, the causes are complex and diverse. The deviation from the B-DNA structure that TFs of the tryptophan cluster family (such as ETS1)[18] [19] force on their binding motif results in higher-than-expected rate of CPDs on their dipyrimidines. This is probably the main driver of the unexpected mutational peak observed in melanomas in the binding sites of these TFs, reported elsewhere [18] [19]. On the other hand, the decreased repair is more likely caused by nonspecific impairment of NER accessibility. Indeed, in the binding sites of the TFs of several families (notably, C2H2 zinc fingers) the increase of mutations is essentially driven by the decreased accessibility of NER to the lesions determined by the bound TFs. The mutational peak that arises from the increase CPDs damage formation tends to be confined to one or few dipyrimidines within the motif that contact the protein. On the other hand, since the TF covers an area of DNA extending beyond the binding motif, the impaired accessibility of the repair machinery engenders broader peaks of mutations.

Besides these results, the study presented here possess some conceptual and methodological value that may be applicable to the analysis of the interaction between mutational processes and features of the chromatin. Using the amount of DNA damage observed in a genomic regions (TFBS in this case) and the estimated rate of repair during 48 hours, we predict the rate of damage that remains at that time. This idea could be applied to the estimation of other types of DNA damage across other genomic features. Furthermore, through the correlation between the damage at earlier and later time points and the mutation rate, we demonstrate the role of repair in the formation of mutations.

Finally, even though focused on TFBS, our work is an example of how the path leading from the creation of DNA lesions to mutations can be reconstructed and studied in a stepwise manner. Therefore, we believe a similar approach could be used to understand other mutational biases observed along the genome, either caused by UV or any other mutational processes.

References

1. Bergamini, C. M., Gambetti, S. & Cervellati, A. D. and C. Oxygen, Reactive Oxygen Species and Tissue Damage. *Curr. Pharm. Des.* **10**, 1611–1626 (2004).
2. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169–185 (2004).
3. Rastogi, R., Ashok Kumar, R., Tyagi, M. & Sinha, R. Molecular Mechanisms of Ultraviolet Radiation-Induced DNA Damage and Repair. *J. Nucleic Acids* **592980**, (2010).
4. Grollman, A. P. & Moriya, M. Mutagenesis by 8-oxoguanine: an enemy within. *Trends Genet.* **9**, 246–249 (1993).
5. Xie, Z. *et al.* Mutagenesis of benzo[a]pyrene diol epoxide in yeast: requirement for DNA polymerase zeta and involvement of DNA polymerase eta. *Biochemistry* **42**, 11253–11262 (2003).
6. Modrich, P. Methyl-directed DNA mismatch correction. *J. Biol. Chem.* **264**, 6597–6600 (1989).
7. Huang, J. C., Svoboda, D. L., Reardon, J. T. & Sancar, A. Human nucleotide excision nuclease removes thymine dimers from DNA by incising the 22nd phosphodiester bond 5’ and the 6th phosphodiester bond 3’ to the photodimer. *Proc. Natl. Acad. Sci.* **89**, 3664 (1992).
8. Karran, P., Lindahl, T. & Griffin, B. Adaptive response to alkylating agents involves alteration in situ of O 6 -methylguanine residues in DNA. *Nature* **280**, 76–77 (1979).
9. Setlow, R. B. Cyclobutane-Type Pyrimidine Dimers in Polynucleotides. *Science* **153**, 379 (1966).
10. Yoon, J.-H. *et al.* Error-Prone Replication through UV Lesions by DNA Polymerase θ Protects against Skin Cancers. *Cell* **176**, 1295-1309.e15 (2019).
11. Barak, Y., Cohen-Fix, O. & Livneh, Z. Deamination of Cytosine-containing Pyrimidine Photodimers in UV-irradiated DNA SIGNIFICANCE FOR UV LIGHT MUTAGENESIS. *J. Biol. Chem.* **270**, 24174–24179 (1995).

12. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
13. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
14. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer genes. *Nature* **499**, 214–218 (2013).
15. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).
16. 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).
17. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259 (2016).
18. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.* **9**, (2018).
19. Elliott, K. *et al.* Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLoS Genet.* **14**, e1007849 (2018).
20. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.* **45**, D61–D67 (2017).
21. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
22. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).
23. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).

24. Hu, J., Adebali, O., Adar, S. & Sancar, A. Dynamic maps of UV damage formation and repair for the human genome. *Proc. Natl. Acad. Sci.* 201706522 (2017) doi:10.1073/pnas.1706522114.
25. Mori, T. *et al.* Simultaneous establishment of monoclonal antibodies specific for either cyclobutane pyrimidine dimer or (6-4)photoproduct from the same mouse immunized with ultraviolet-irradiated DNA. *Photochem. Photobiol.* **54**, 225–232 (1991).
26. Hu, J., Adar, S., Selby, C. P., Lieb, J. D. & Sancar, A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **29**, 948–960 (2015).
27. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
28. Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684 (2017).

Figures

Figure 1.

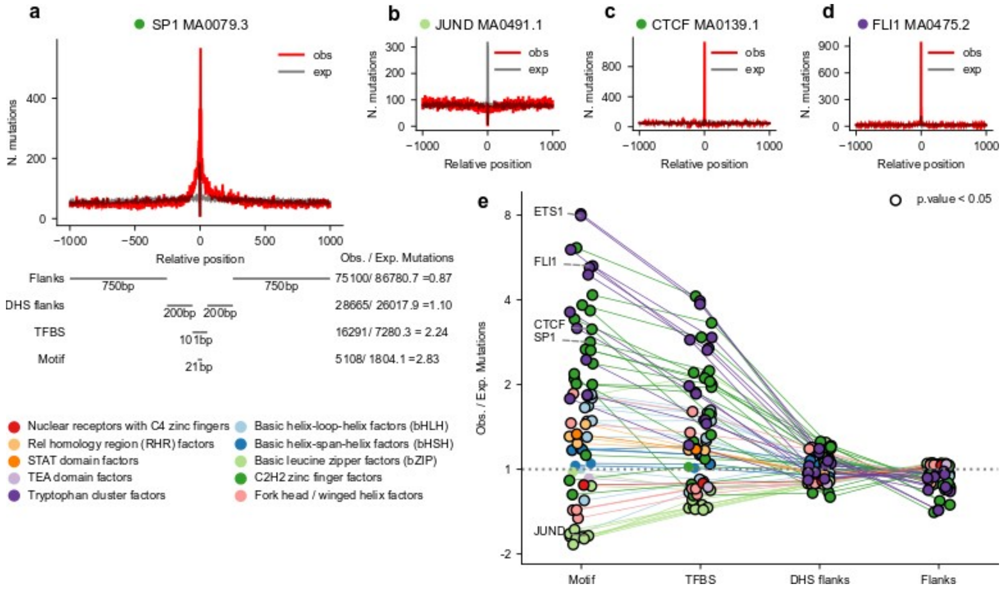


Figure 2.

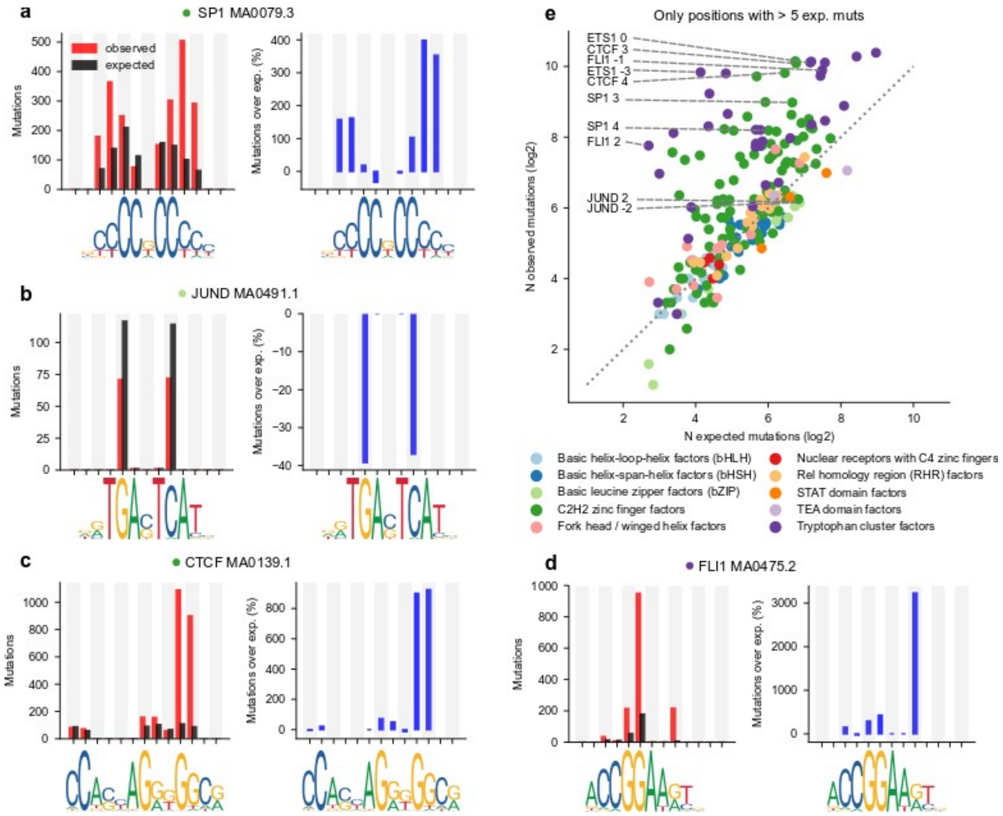

a
● SP1 MA0079.3
observed
expected
Mutations
Mutations over exp. (%)
b
● JUND MA0491.1
Mutations
Mutations over exp. (%)
c
● CTCF MA0139.1
Mutations
Mutations over exp. (%)
d
● FLI1 MA0475.2
Mutations
Mutations over exp. (%)
e
Only positions with > 5 exp. muts
ETS1 0
CTCF 3
FLI1 -1
ETS1 -3
CTCF 4
SP1 3
SP1 4
FLI1 2
JUND 2
JUND -2
N observed mutations (log2)
N expected mutations (log2)
Basic helix-loop-helix factors (bHLH)
Basic helix-span-helix factors (bHSH)
Basic leucine zipper factors (bZIP)
C2H2 zinc finger factors
Fork head / winged helix factors
Nuclear receptors with C4 zinc fingers
Rel homology region (RHR) factors
STAT domain factors
TEA domain factors
Tryptophan cluster factors

Figure 3.

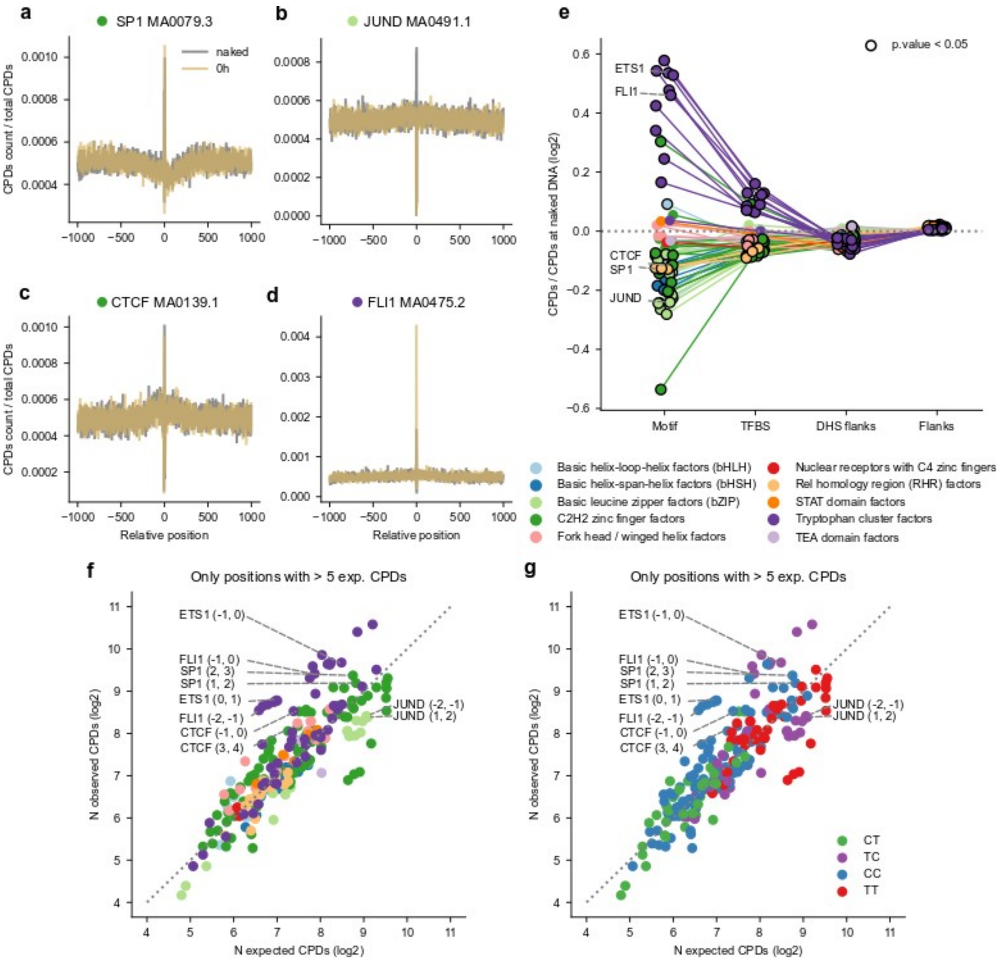

a
SP1 MA0079.3
naked
0h
CPDs count / total CPDs
0.0010
0.0008
0.0006
0.0004
−1000 −500 0 500 1000
b
JUND MA0491.1
0.0008
0.0006
0.0004
0.0002
0.0000
−1000 −500 0 500 1000
c
CTCF MA0139.1
CPDs count / total CPDs
0.0010
0.0008
0.0006
0.0004
0.0002
−1000 −500 0 500 1000
Relative position
d
FLI1 MA0475.2
0.004
0.003
0.002
0.001
0.000
−1000 −500 0 500 1000
Relative position
e
p.value < 0.05
CPDs / CPDs at naked DNA (log2)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
ETS1
FLI1
CTCF
SP1
JUND
Motif
TFBS
DHS flanks
Flanks
Basic helix-loop-helix factors (bHLH)
Basic helix-span-helix factors (bHSH)
Basic leucine zipper factors (bZIP)
C2H2 zinc finger factors
Fork head / winged helix factors
Nuclear receptors with C4 zinc fingers
Rel homology region (RHR) factors
STAT domain factors
Tryptophan cluster factors
TEA domain factors
f
Only positions with > 5 exp. CPDs
N observed CPDs (log2)
ETS1 (-1, 0)
FLI1 (-1, 0)
SP1 (2, 3)
SP1 (1, 2)
ETS1 (0, 1)
FLI1 (-2, -1)
CTCF (-1, 0)
CTCF (3, 4)
JUND (-2, -1)
JUND (1, 2)
N expected CPDs (log2)
4 5 6 7 8 9 10 11
g
Only positions with > 5 exp. CPDs
N observed CPDs (log2)
ETS1 (-1, 0)
FLI1 (-1, 0)
SP1 (2, 3)
SP1 (1, 2)
ETS1 (0, 1)
FLI1 (-2, -1)
CTCF (-1, 0)
CTCF (3, 4)
JUND (-2, -1)
JUND (1, 2)
CT
TC
CC
TT
N expected CPDs (log2)
4 5 6 7 8 9 10 11

Figure 4.

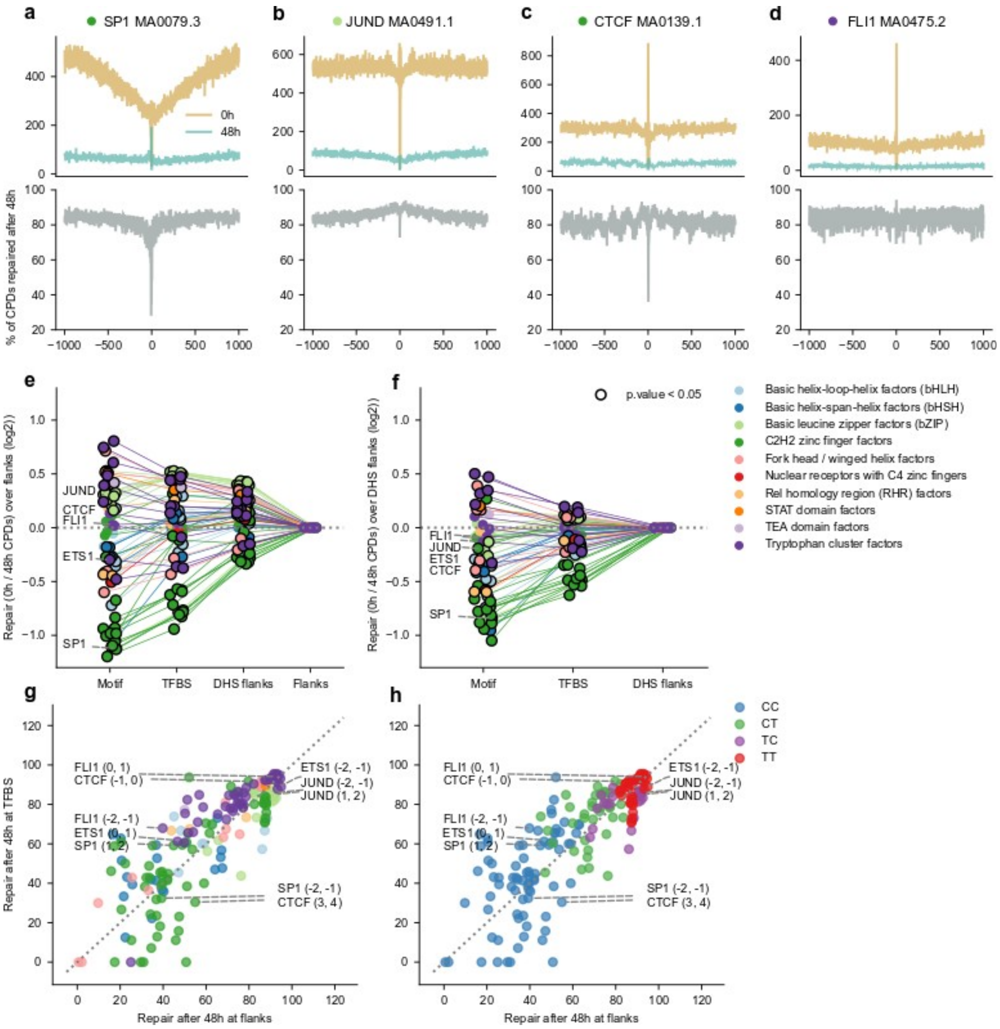

a
SP1 MA0079.3
b
JUND MA0491.1
c
CTCF MA0139.1
d
FLI1 MA0475.2
0h
48h
% of CPDs repaired after 48h
−1000
−500
0
500
1000
e
f
Repair (0h / 48h CPDs) over flanks (log2))
Repair (0h / 48h CPDs) over DHS flanks (log2))
JUND
CTCF
FLI1
ETS1
SP1
Motif
TFBS
DHS flanks
Flanks
p.value < 0.05
Basic helix-loop-helix factors (bHLH)
Basic helix-span-helix factors (bHSH)
Basic leucine zipper factors (bZIP)
C2H2 zinc finger factors
Fork head / winged helix factors
Nuclear receptors with C4 zinc fingers
Rel homology region (RHR) factors
STAT domain factors
TEA domain factors
Tryptophan cluster factors
g
h
Repair after 48h at TFBS
Repair after 48h at flanks
FLI1 (0, 1)
CTCF (-1, 0)
FLI1 (-2, -1)
ETS1 (0, 1)
SP1 (1, 2)
ETS1 (-2, -1)
JUND (-2, -1)
JUND (1, 2)
SP1 (-2, -1)
CTCF (3, 4)
CC
CT
TC
TT

Figure 5.

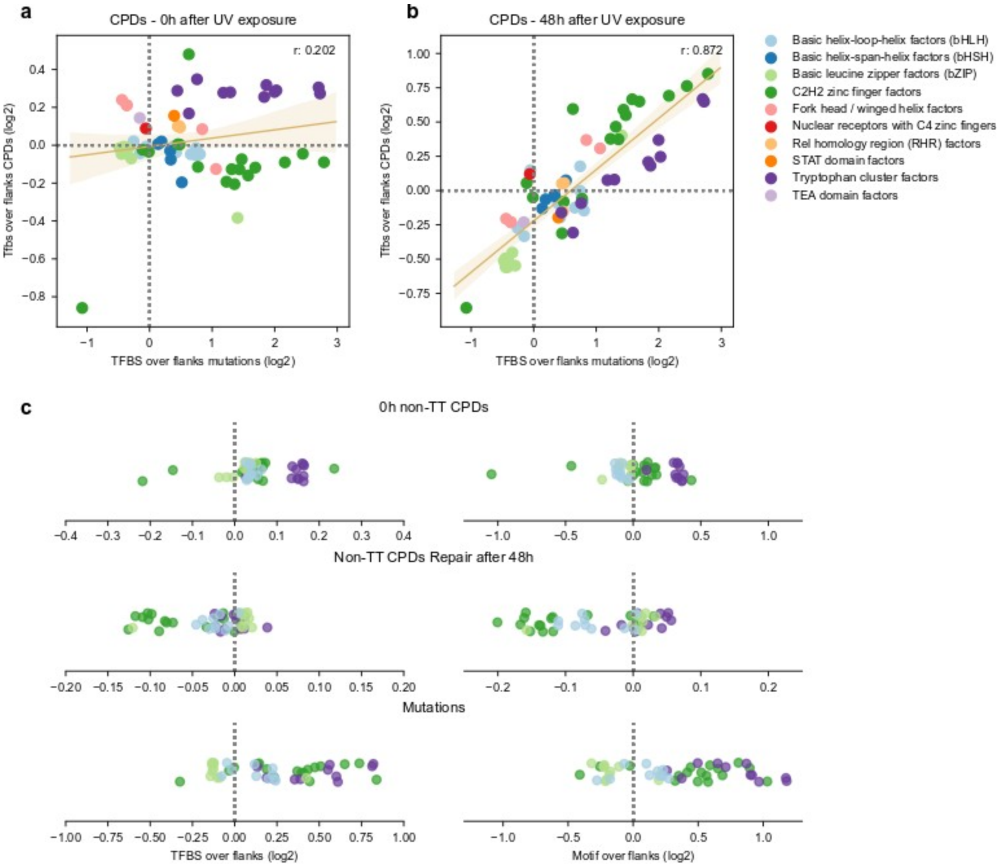

a
CPDs - 0h after UV exposure
r: 0.202
Tfbs over flanks CPDs (log2)
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1
0
1
2
3
TFBS over flanks mutations (log2)
b
CPDs - 48h after UV exposure
r: 0.872
Tfbs over flanks CPDs (log2)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1
0
1
2
3
TFBS over flanks mutations (log2)
● Basic helix-loop-helix factors (bHLH)
● Basic helix-span-helix factors (bHSH)
● Basic leucine zipper factors (bZIP)
● C2H2 zinc finger factors
● Fork head / winged helix factors
● Nuclear receptors with C4 zinc fingers
● Rel homology region (RHR) factors
● STAT domain factors
● Tryptophan cluster factors
● TEA domain factors
c
0h non-TT CPDs
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
−1.0
−0.5
0.0
0.5
1.0
Non-TT CPDs Repair after 48h
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
−0.2
−0.1
0.0
0.1
0.2
Mutations
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
−1.0
−0.5
0.0
0.5
1.0
TFBS over flanks (log2)
Motif over flanks (log2)

Figure legends

Figure 1. The mutation rate in TFBS in melanomas

(a) 2001-nucleotide sequences centered at the middle point of active TFBS are extracted from the genomic sequence. Within each sequence four areas are delimited: (from the center to the periphery) motif (21 bps), TFBS (101 bps), DHS flanks (400 bps), and flanks (1500 bps). The somatic mutations identified in a cohort of 136 melanomas are mapped to these sequences. Then, all 2001-nucleotide sequences containing the same type of binding motif of a TF are stacked. The number of observed mutations at each position of the stack is calculated, and the expected distribution of mutations across the 2001-nucleotide stack is computed from profile of tri-nucleotide whole-genome substitution frequencies observed across the cohort.

(a-d) Smoothed observed (red) and expected (gray) distributions of mutations in the stack of 2001-nucleotide sequences centered around MA1107.1 binding motif of KLF9 (a), MA0491.1 binding motif of JUND (b), MA0139.1 binding motif of CTCF (c), and MA 0475.2 binding motif of FLI1.

(e) Ratio of observed to expected mutations (in log2 scale) within the four regions defined across the stacks of 2001-nucleotide sequences centered across 64 types of TF binding motifs with more than 5000 sequences and a median number of mutations across all positions greater than 2. Positive values correspond to higher-than-expected mutation rates at each region, whereas lower-than-expected mutation rates possess negative values. Points that correspond to instances of significant deviation from the expectation (G-test p-value<0.05) are encircled in black. The thin straight lines join the values computed for the regions of a given type of motif. The circles corresponding to each type of motif are colored according to the family of the corresponding TF, following the legend presented next to the panel.

Figure 2. The mutation rate at specific positions within the TF binding motif

(a-d) The left graph in each panel presents the observed (red bar) and expected (black bar) number of mutations at each mutated dipyrimidine position of the four types of binding motifs shown in Figures 1a-d. To compute the expected mutations at each dipyrimidine position the tetramer containing the dipyrimidine was sampled from the Flanks of the same 2001-nucleotides sequence, and their observed number of mutations averaged. The right graph presents the corresponding percentage of increase of the rate of mutations over that expected (blue bar) at each mutated dipyrimidine position. Positive values thus correspond to increased mutation rate, while negative values occur at positions with decreased mutation rate.

(e) Scatterplot representing the relationship between the number of observed and expected mutations (in log2 scale) at all dipyrimidine positions within all types of TF binding motifs included in the study. Each dot, hence corresponds to an individual dipyrimidine position in a binding motif colored following the family of the corresponding TF, according to the legend below the panel. Dipyrimidine positions with significant increased or decreased number of mutations with respect to the expectation (G-test p-value<0.05) are encircled in black.

Figure 3. The formation of CPDs in TFBS

(a-d) Distribution of the rate of CPDs formed upon irradiation within 2001-nucleotides sequences centered at the middle position of TFBS occupied by the TF (ambar) or naked (dark gray), which reflect, respectively, the observed and expected distribution of CPDs along this window. The graphs present the distribution of observed and expected CPD rates along the 2001-nucleotides stack of the sequences of the types of motifs shown in Figures 1 and 2.

(e) Ratio of observed to expected CPDs formation rates (in log2 scale) within the four regions defined across the 2001-nucleotide stacks of sequences centered across 64 TF binding motifs under study. Positive values correspond to higher-than-expected CPDs formation rates at each region, whereas lower-than-expected mutation rates possess negative values. Points representing instances with significant deviations from the expectation (G-test p-value<0.05) are encircled in black. Thin straight lines join the values computed for the regions of a given type of motif. The circles computed for each region of each type of motif are colored following the family of the corresponding TF, according to the legend presented below the panel.

(f) Scatterplots representing the relationship between the number (log2) of observed and expected CPDs formed at specific dipyrimidine positions within each type of motif. In the left plot, the dots representing the motifs are colored following TF families, while in the right plot, their colors correspond to the type of dipyrimidine where they occur.

Figure 4. The repair of CPDs in TFBS

(a-d) The activity of repair of CPDs in four exemplary types of TF binding motifs. The top panels represent the rate of CPDs experimentally generated immediately after irradiation (0h, ambar) and the rate of CPDs persisting 48 hours (blue) after irradiation. The subtraction of the latter from the former, nucleotide-by-nucleotide yields the rate of repair across each type of binding motif. Different shapes of repair signals are visible. While in the first three examples, the activity of repair is clearly decreased within the TFBS, in the FLI1 motif, no decrease of repair may be appreciated. In the CTCF motif, decreased repair overlapping well-positioned nucleosomes flanking the TFBS is appreciable [16] [15]).

(e) Ratio of the repair rates of CPDs at motifs, TFBS or DHS flanks with respect to flanks (in log2 scale). Positive values correspond to larger CPDs repair rates at either region than at the flanks, whereas lower-than-flanks mutation ratios possess negative values. Points encircled in black represent instances of significant deviation from the expectation (G-test p-value<0.05). Thin straight lines join the values computed for the regions of a given type of motif. The points representing the ratios computed for each region of each type of motif are colored according to the family of the corresponding TF, following the legend presented next to the panel.

(f) Ratio of the repair rates of CPDs at motifs or TFBS with respect to DHS flanks (in log2 scale). Notation as in panel (e)

(g) Scatterplots representing the relationship between the percent of repaired CPDs at specific dipyrimidine positions within individual types of motifs (after 48 hours) and the average repair rate for the same type of dipyrimidine in the same context (tetranucleotide) within the flanks. Each dot corresponds to an individual dipyrimidine. Dots above the diagonal represent dipyrimidines whose CPDs are repaired at higher rate than those of equivalent dipyrimidines in the flanks. On the other hand, dots below the diagonal correspond to dipyrimidines at positions that experience lower repair rate. In the left graph, the dots are colored following the family of the corresponding TF, while in the right graph, they are colored according to the type of dipyrimidine.

Figure 5. The UV mutational process in TFBS

(a) Relationship between the ratio (log2) of TFBS-to-flanks CPDs (y-axis) computed immediately after irradiation and the ratio (log2) of TFBS-to-flanks mutations (x-axis). Each dot corresponds to a type of motif of a specific TF, and they are colored according to the TF family. The trendline and the Pearson’s correlation coefficients computed for both relationships are presented in the graph.

(b) Relationship between the ratio (log2) of TFBS-to-flanks CPDs (y-axis) computed 48 hours after irradiation and the ratio (log2) of TFBS-to-flanks mutations (x-axis). Each dot corresponds to a type of motif of a specific TF, and they are colored according to the TF family. The trendline and the Pearson’s correlation coefficients computed for both relationships are presented in the graph.

(c) Toy models explaining the interplay between UV-induced damage and repair within the binding site of TFs of three different families. The rate of CPDs, repair by NER and resulting mutations are represented respectively by amber, gray and red lines.

Methods

Genomic location of transcription factor binding sites

Transcription factors chip-seq data was obtained from the GTRD database [20]. Peaks called with MACS algorithm were used. Position frequency matrices (PFM) describing the sequence of each binding motif were downloaded from JASPAR database [21]. Then, PFM were mapped within the transcription factor peaks of their corresponding TF using MOODS suite [22]. DNase I hypersensitive sites (DHS) coordinates from melanocytes and skin fibroblasts were downloaded from the Epigenome Roadmap Project [27]. Only those binding motifs overlapping with DHS sites in both cell types were considered as active and therefore included in our analysis.

Whole genome mutation data

Whole-genome somatic mutations from 183 skin cutaneous melanoma (SKCM) samples belonging to the MELA-AU cohort [23] were retrieved. Proportion of C-T mutations was used to assess the extent of UV exposure per sample. Only the 136 samples with at least 70% of mutations being C-T were selected. Moreover, non C-T mutations were discarded.

Whole genome CPD maps

Genome-wide distribution of CPDs immediately after UV exposure, both at normal and naked DNA was obtained from Mao, P. et al 2018 [18]. CPDs distribution along the genome measured 0h and 48h after UV exposure was obtained from Hu, J. et al. 2017 [24]. Both experiments had single-nucleotide resolution and were performed in skin fibroblast NHF1 cells.

Expected mutation rate computation

The probability of a trinucleotide to be mutated was determined by using the whole genome mutations of all 136 UV induced melanomas, normalized by the abundance of each trinucleotide in the genome. This probabilities where then used to compute the expected mutation rate of any given DNA segment, as detailed in [28].

Transcription factor binding site centering

For any transcription factor included in the analysis, all active binding sites along the genome were identified. Then, a region spanning 1000 nucleotides per side from the center of the binding motif was defined. The resulting 2001bp windows were then aligned using as reference their middle position. Then, for each position, we counted the number of observed and expected mutations, and also CPDs at different timepoints after UV exposure. In those cases where the binding motif was found to be in the negative strand, the coordinates of the CPDs and mutations found in the 2001bp window around it were inverted.

Only those transcription factors presenting at least 5000 active binding motifs along the genome and whose median number of mutations per position after stacking all instances was greater than 2 were analyzed.

Motif zoom-in analysis

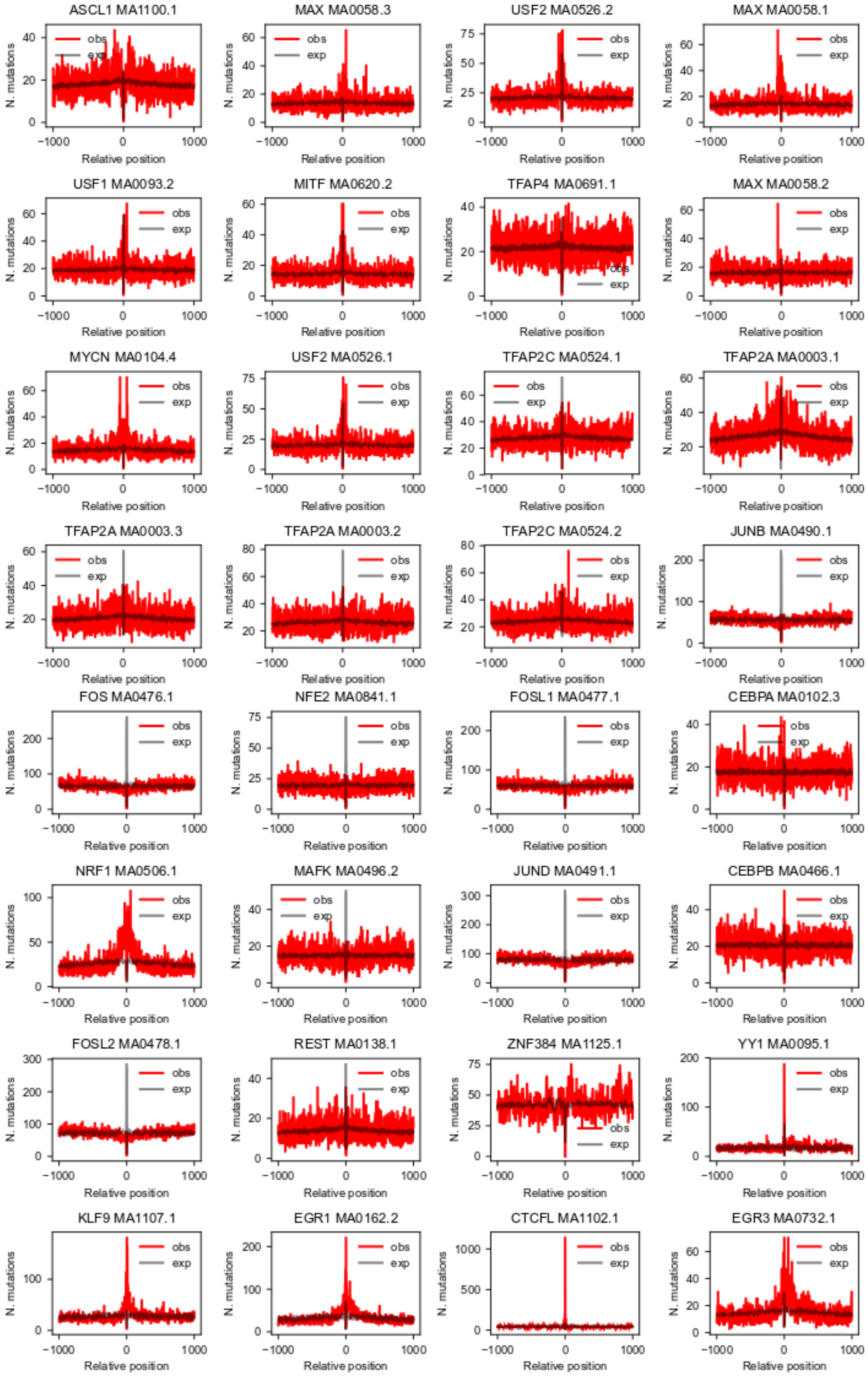
For each specific transcription factor binding motif, those dipyrimidines conserved in at least 50% of all instances of the motif were selected. Then, the number of mutations, CPDs at 0 and 48h after UV exposure and repair activity were assessed. Next, dipyrimidines of the same type with identical flanking bases were sampled from their flanks, whose mutation, and CPDs load were used as a reference to compare with. The sampling process was performed 50 times per dipyrimidine and the results averaged.

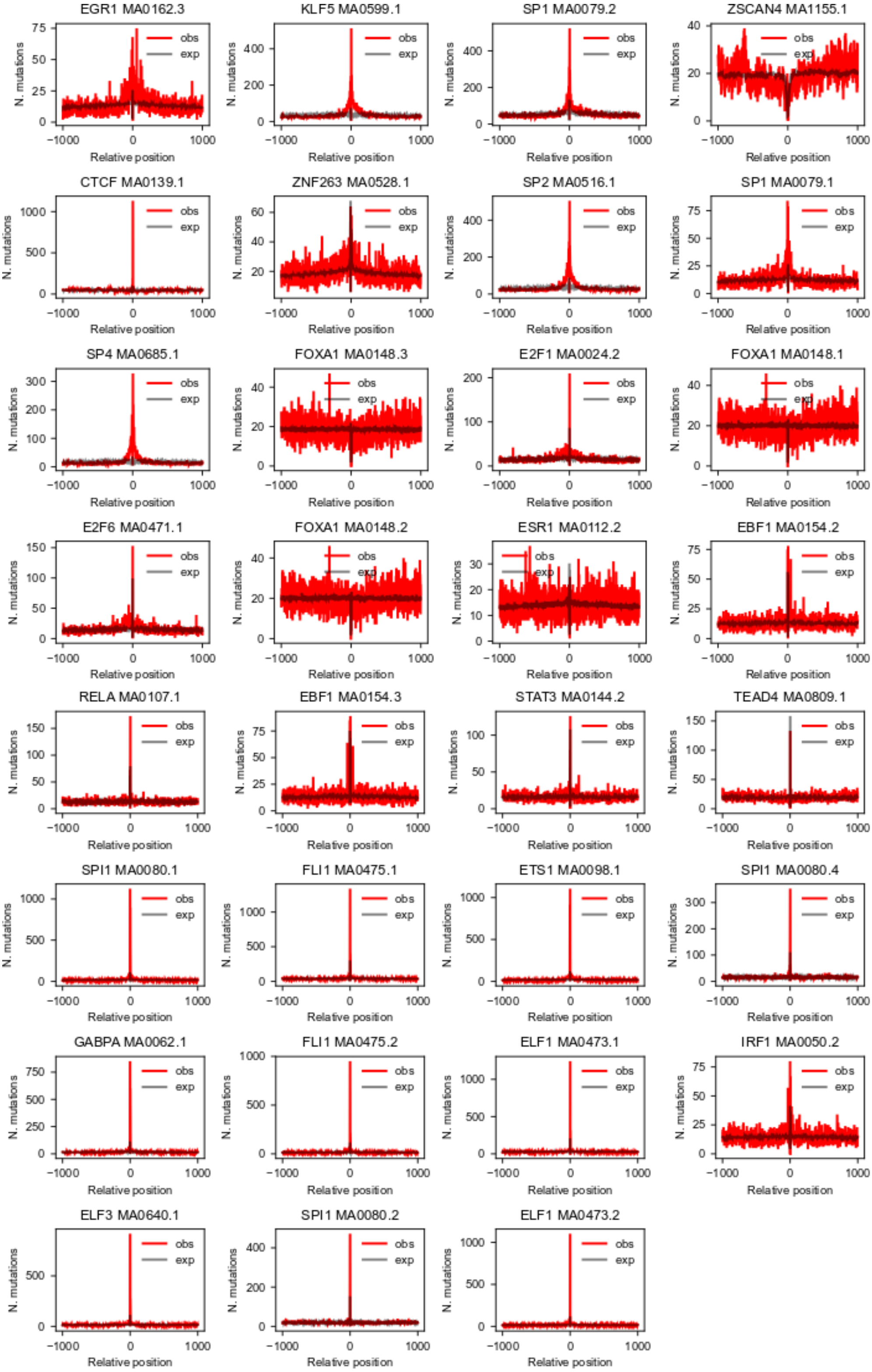
Statistics

When comparing the observed number of mutations or CPDs in a specific region with an expected distribution, being the simulated number of mutations or the CPDs in the naked DNA respectively, a chi-squared goodness of fit test was used. Alternatively, when evaluating the differences in the CPDs distribution at different timepoints after UV exposure, a standard chi-squared test was employed.

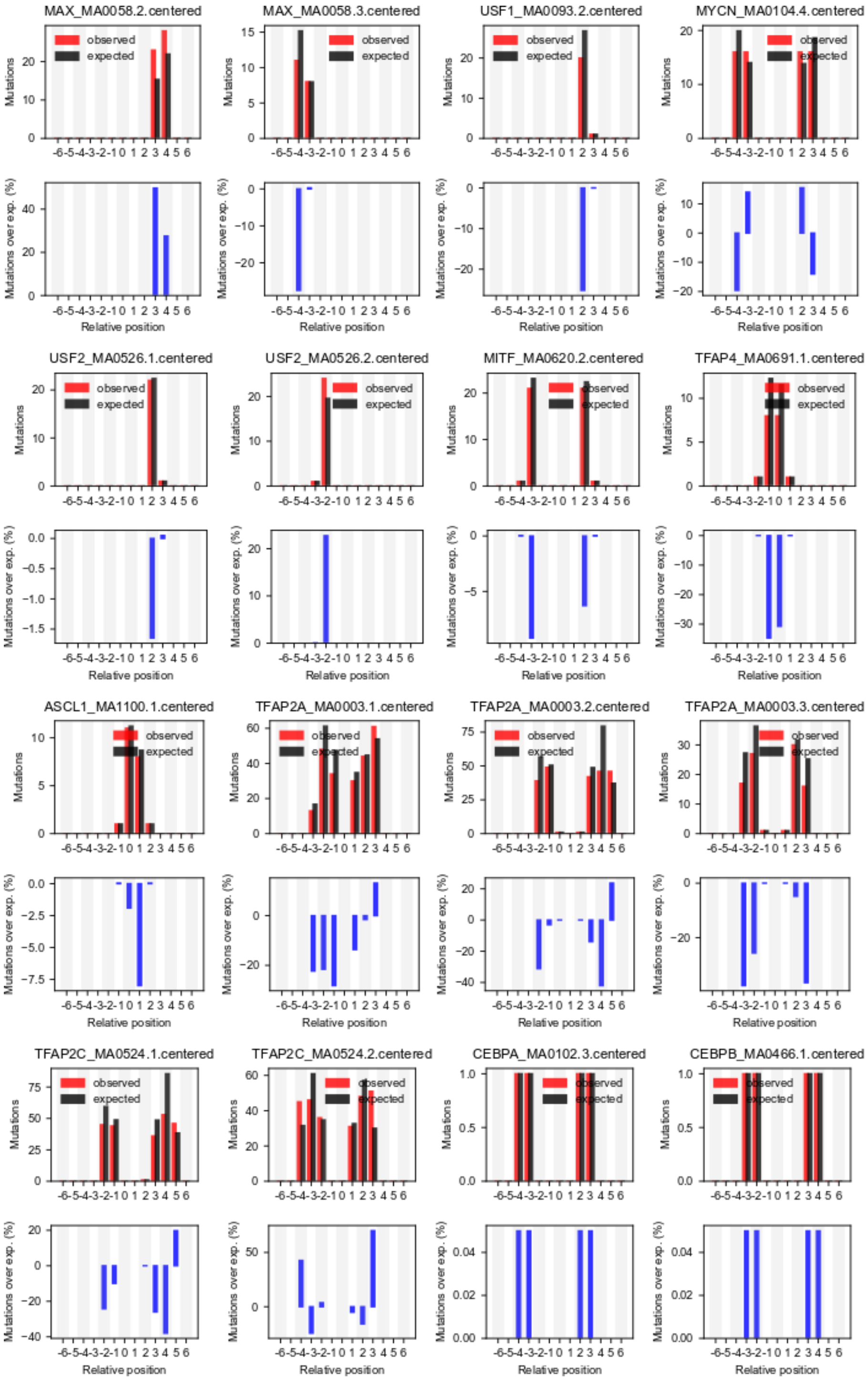
Supplementary Figures

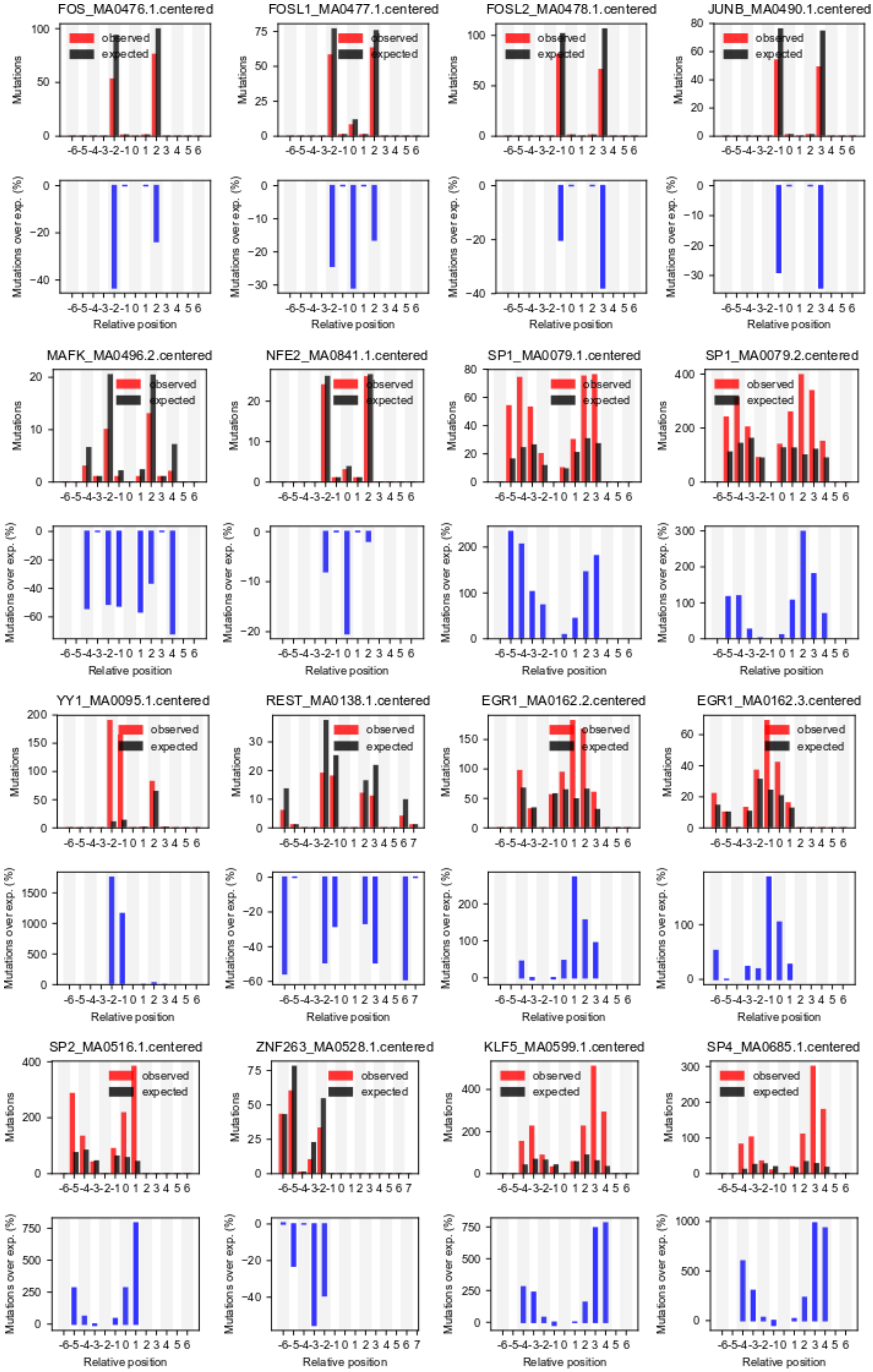
Supplementary Figure 1.

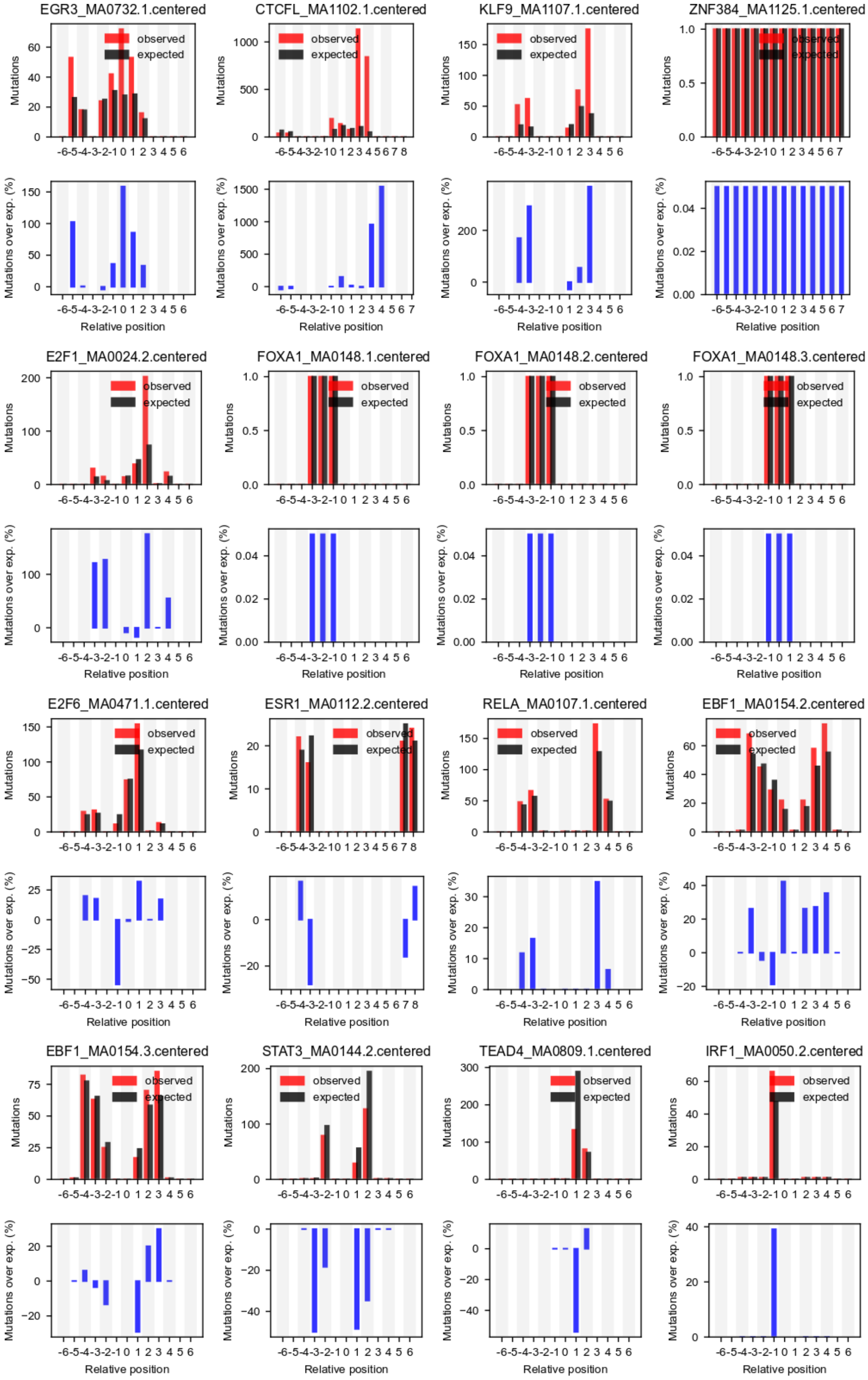


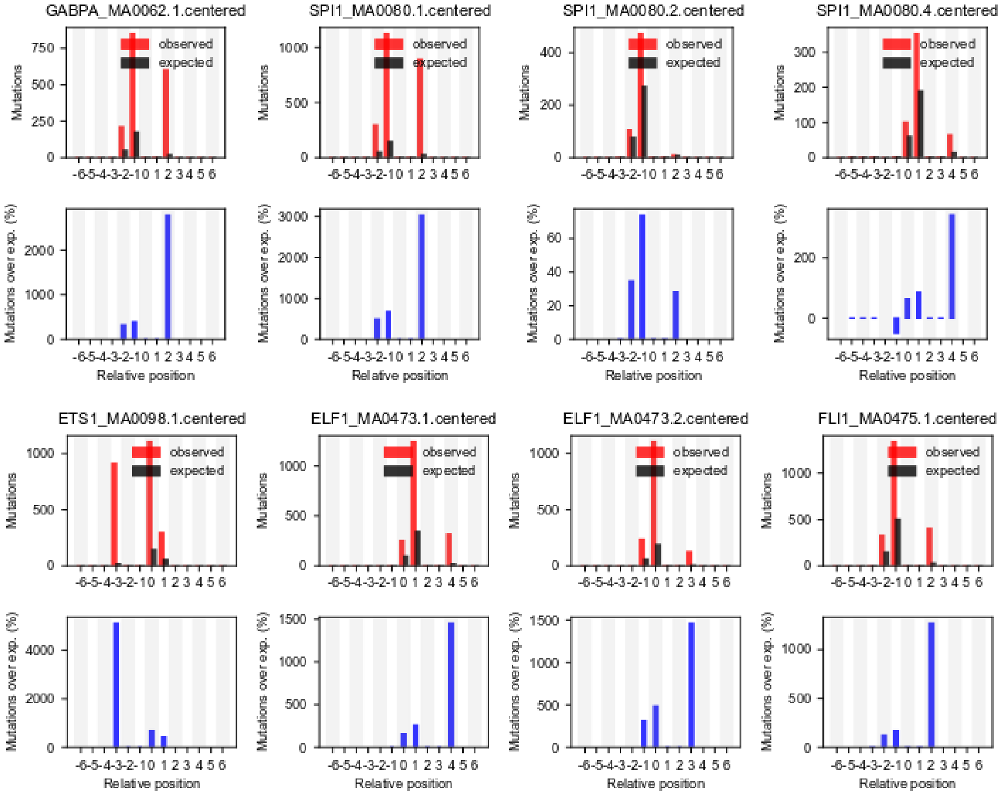

EGR1 MA0162.3
KLF5 MA0599.1
SP1 MA0079.2
ZSCAN4 MA1155.1
CTCF MA0139.1
ZNF263 MA0528.1
SP2 MA0516.1
SP1 MA0079.1
SP4 MA0685.1
FOXA1 MA0148.3
E2F1 MA0024.2
FOXA1 MA0148.1
E2F6 MA0471.1
FOXA1 MA0148.2
ESR1 MA0112.2
EBF1 MA0154.2
RELA MA0107.1
EBF1 MA0154.3
STAT3 MA0144.2
TEAD4 MA0809.1
SPI1 MA0080.1
FLI1 MA0475.1
ETS1 MA0098.1
SPI1 MA0080.4
GABPA MA0062.1
FLI1 MA0475.2
ELF1 MA0473.1
IRF1 MA0050.2
ELF3 MA0640.1
SPI1 MA0080.2
ELF1 MA0473.2
obs
exp
N. mutations
Relative position
−1000
0
1000

Supplementary Figure 2.

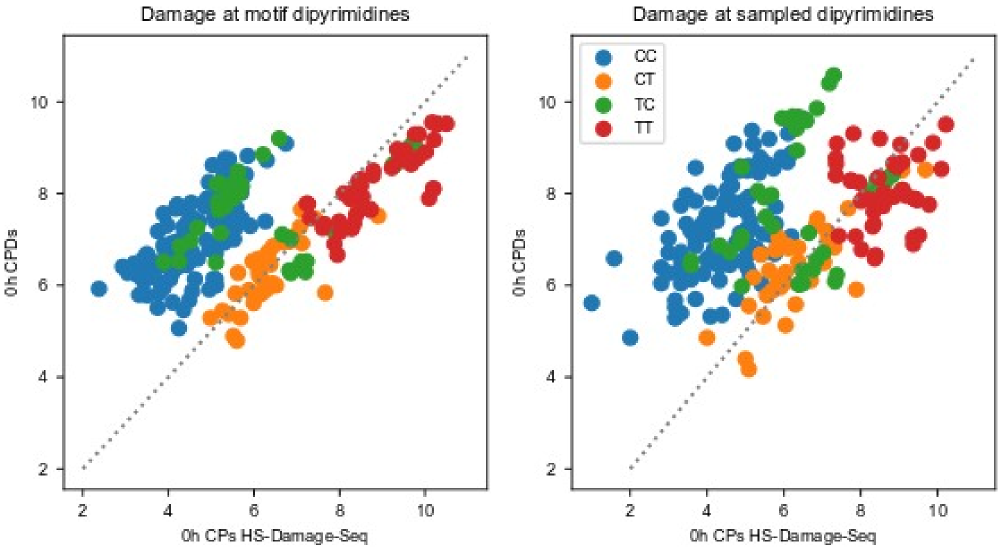



FOS_MA0476.1.centered
FOSL1_MA0477.1.centered
FOSL2_MA0478.1.centered
JUNB_MA0490.1.centered
MAFK_MA0496.2.centered
NFE2_MA0841.1.centered
SP1_MA0079.1.centered
SP1_MA0079.2.centered
YY1_MA0095.1.centered
REST_MA0138.1.centered
EGR1_MA0162.2.centered
EGR1_MA0162.3.centered
SP2_MA0516.1.centered
ZNF263_MA0528.1.centered
KLF5_MA0599.1.centered
SP4_MA0685.1.centered
observed
expected
Mutations
Mutations over exp. (%)
Relative position
-6-5-4-3-2-1 0 1 2 3 4 5 6
-6-5-4-3-2-1 0 1 2 3 4 5 6 7

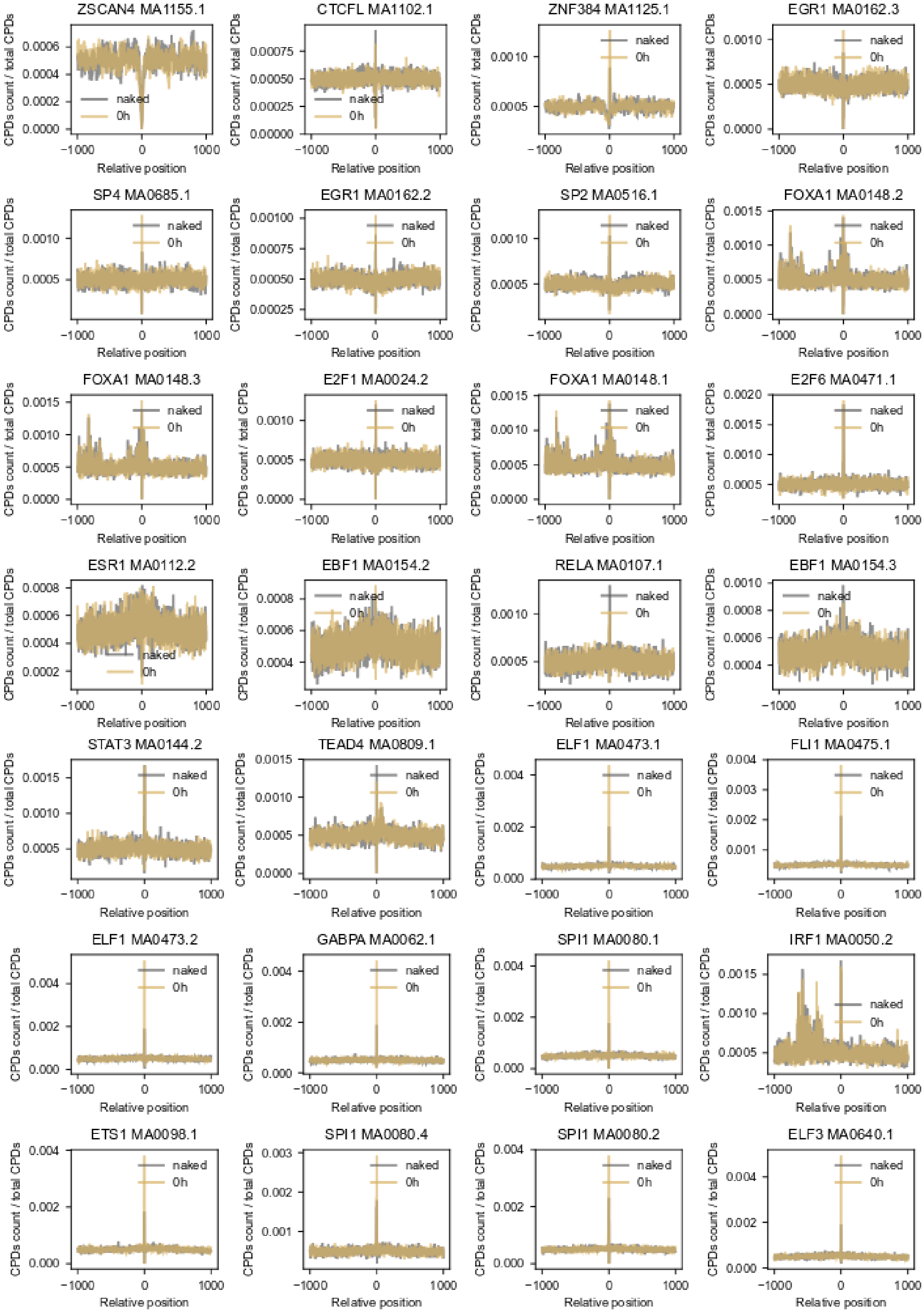

EGR3_MA0732.1.centered
CTCFL_MA1102.1.centered
KLF9_MA1107.1.centered
ZNF384_MA1125.1.centered
E2F1_MA0024.2.centered
FOXA1_MA0148.1.centered
FOXA1_MA0148.2.centered
FOXA1_MA0148.3.centered
E2F6_MA0471.1.centered
ESR1_MA0112.2.centered
RELA_MA0107.1.centered
EBF1_MA0154.2.centered
EBF1_MA0154.3.centered
STAT3_MA0144.2.centered
TEAD4_MA0809.1.centered
IRF1_MA0050.2.centered
observed
expected
Mutations
Mutations over exp. (%)
Relative position
-6-5-4-3-2-1 0 1 2 3 4 5 6

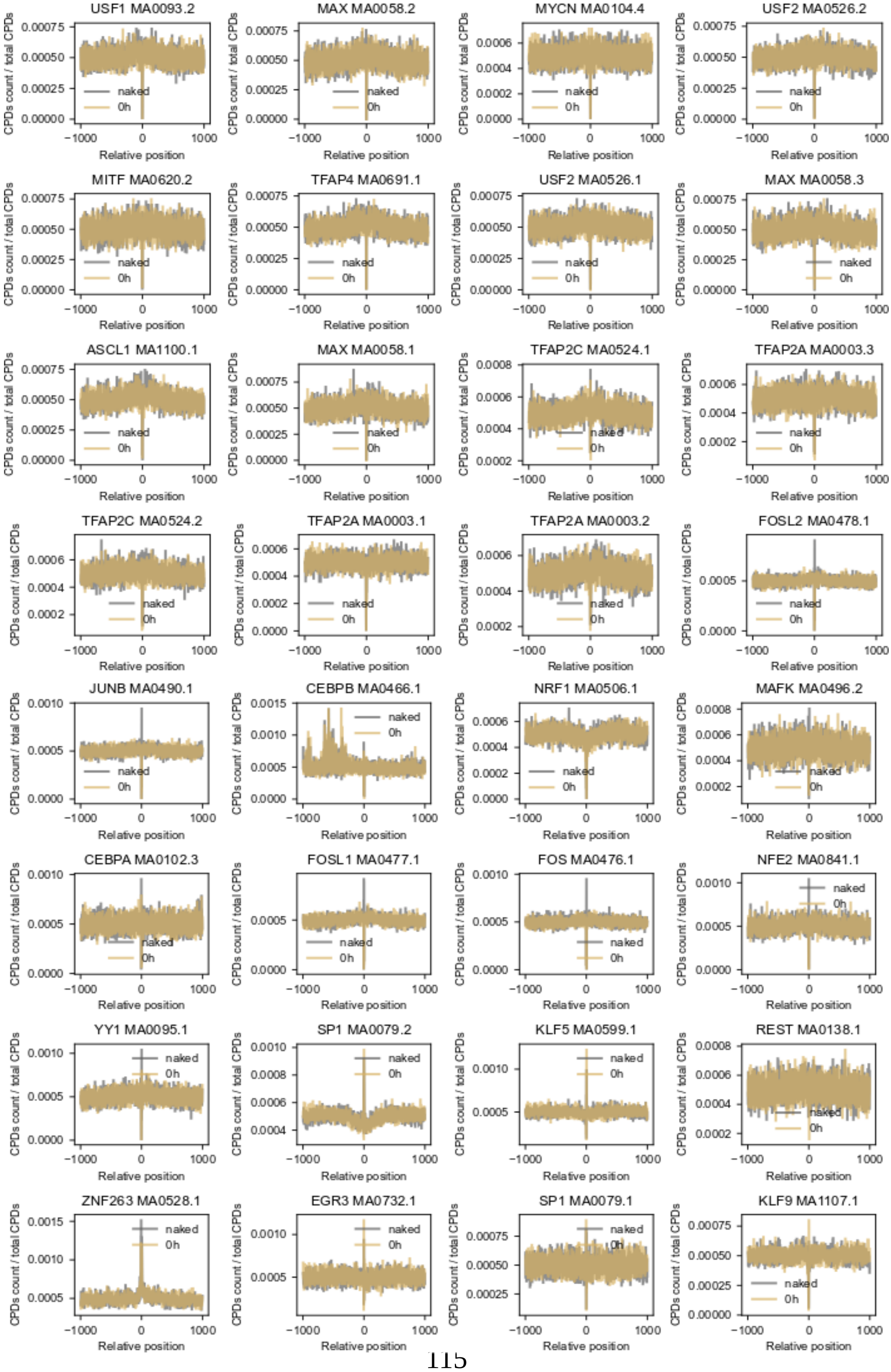

GABPA_MA0062.1.centered
SPI1_MA0080.1.centered
SPI1_MA0080.2.centered
SPI1_MA0080.4.centered
ETS1_MA0098.1.centered
ELF1_MA0473.1.centered
ELF1_MA0473.2.centered
FLI1_MA0475.1.centered
observed
expected
Mutations
Mutations over exp. (%)
-6-5-4-3-2-1 0 1 2 3 4 5 6
Relative position

Supplementary Figure 3.

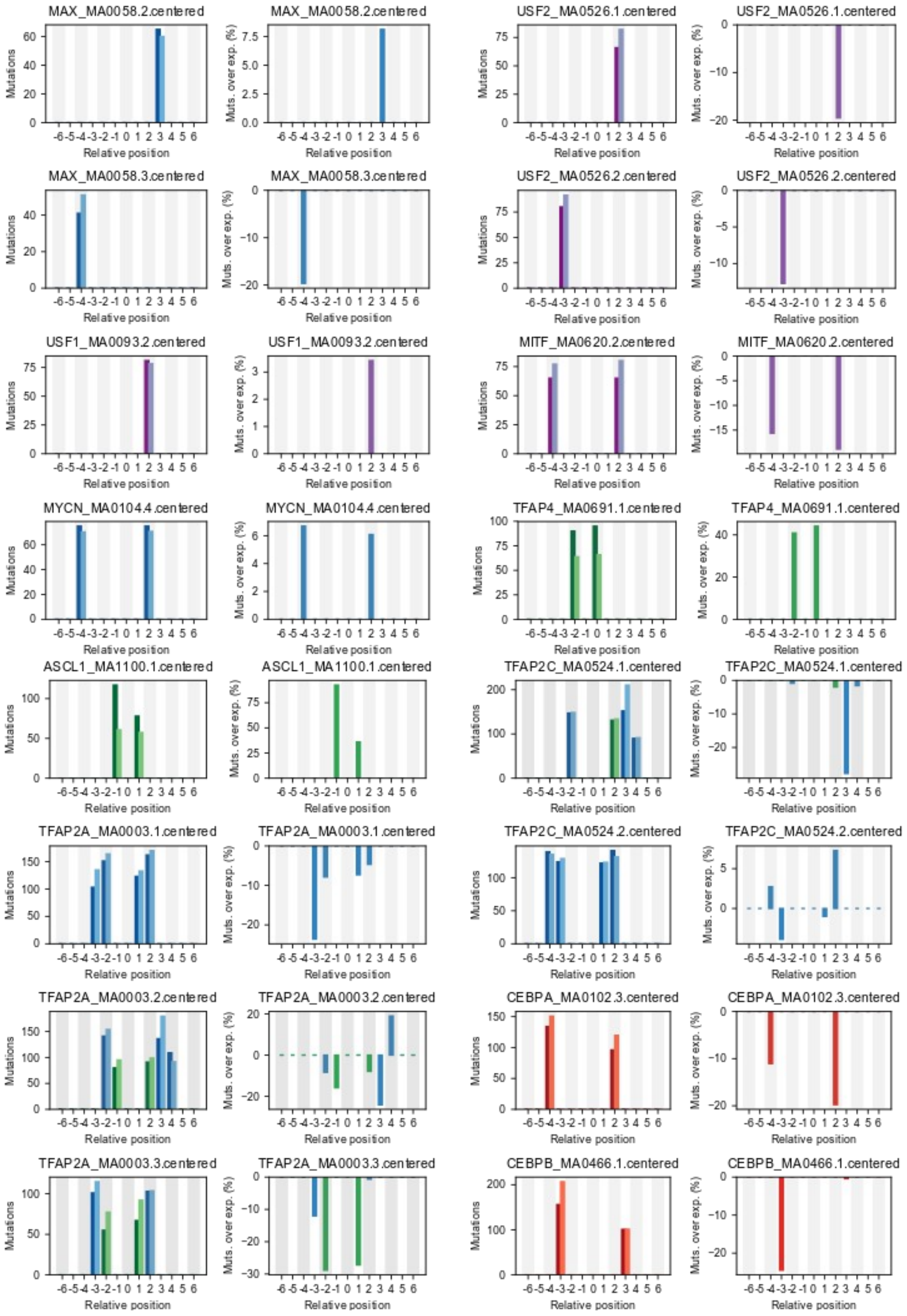


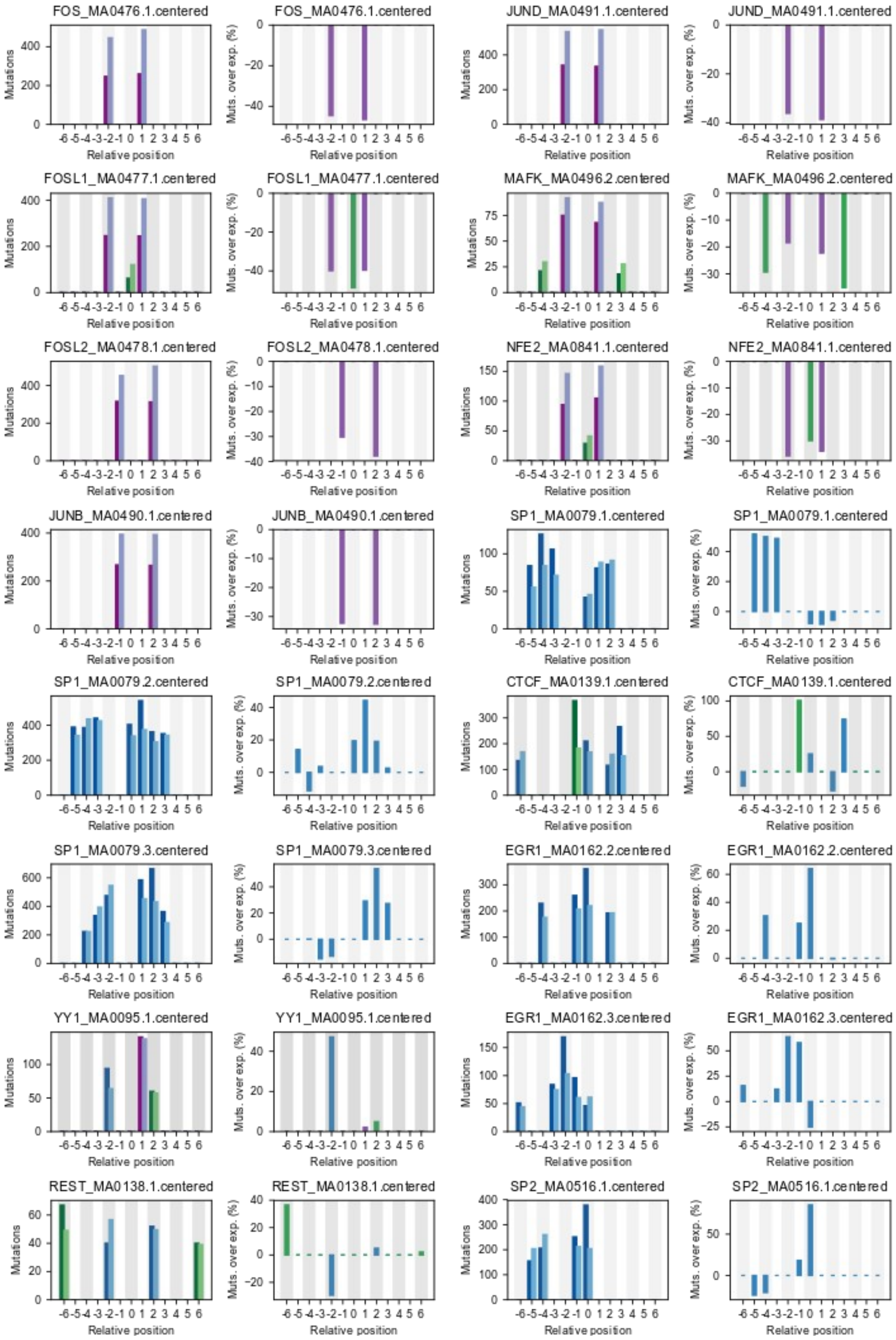
Supplementary Figure 4.

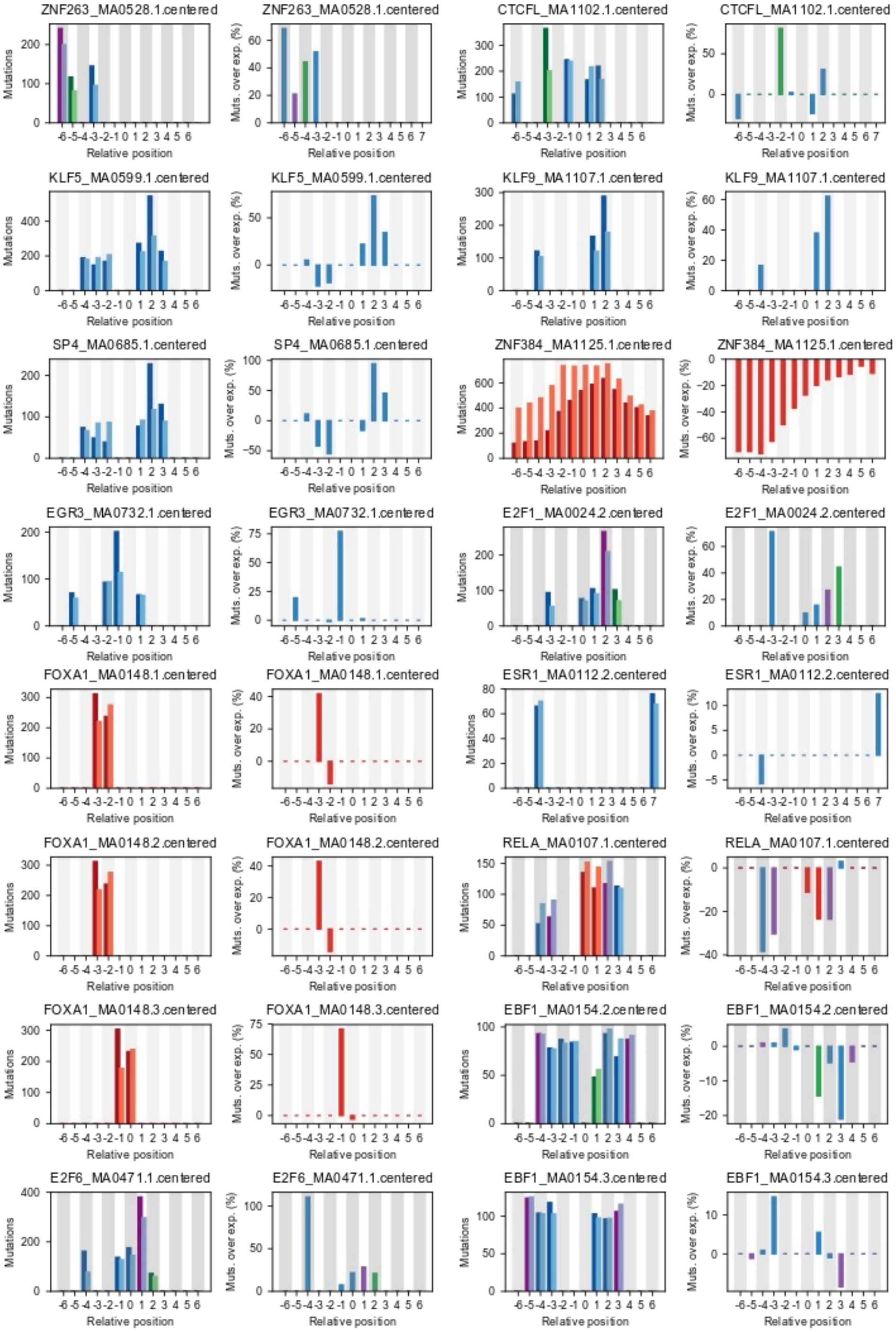


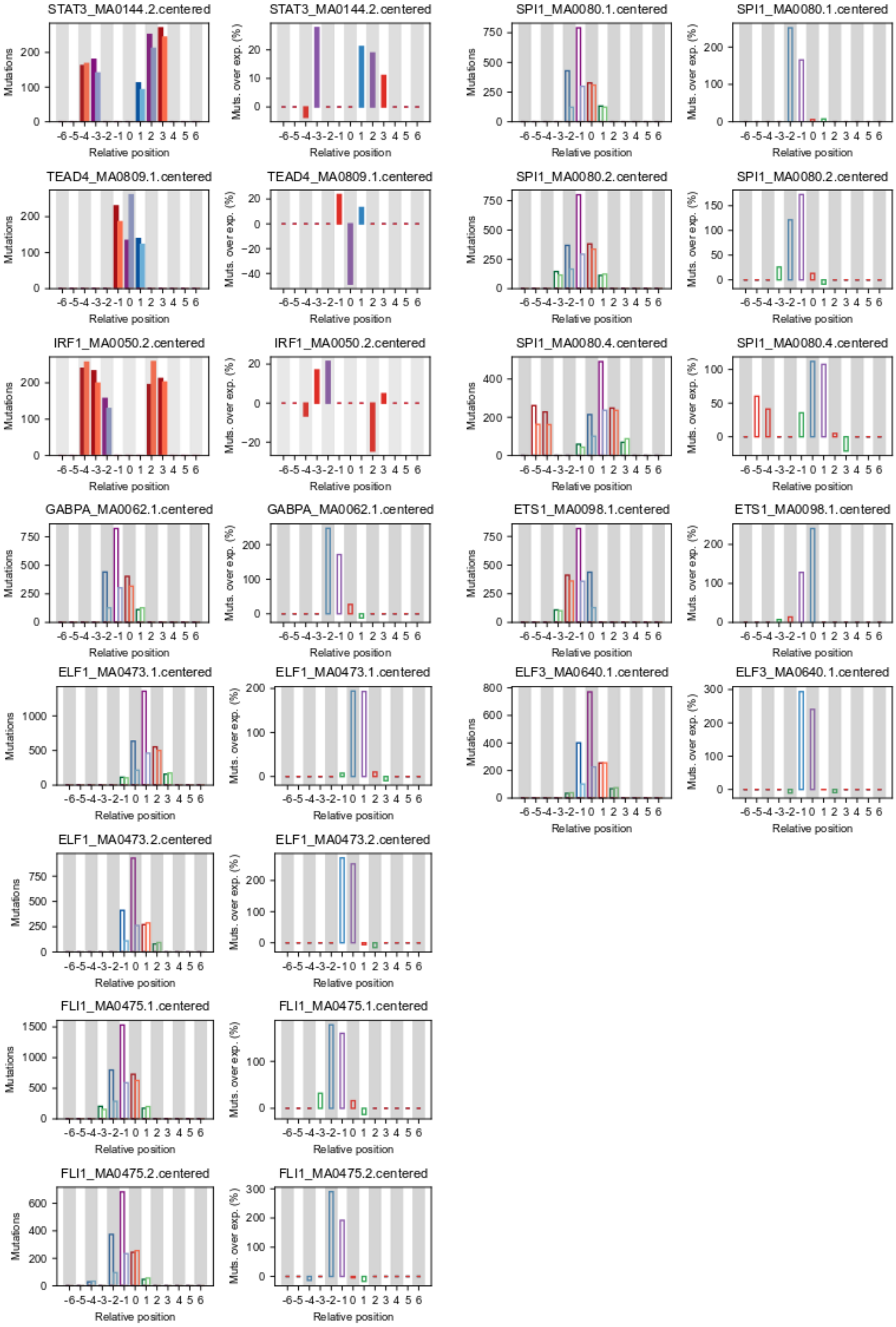

USF1 MA0093.2
MAX MA0058.2
MYCN MA0104.4
USF2 MA0526.2
MITF MA0620.2
TFAP4 MA0691.1
USF2 MA0526.1
MAX MA0058.3
ASCL1 MA1100.1
MAX MA0058.1
TFAP2C MA0524.1
TFAP2A MA0003.3
TFAP2C MA0524.2
TFAP2A MA0003.1
TFAP2A MA0003.2
FOSL2 MA0478.1
JUNB MA0490.1
CEBPB MA0466.1
NRF1 MA0506.1
MAFK MA0496.2
CEBPA MA0102.3
FOSL1 MA0477.1
FOS MA0476.1
NFE2 MA0841.1
YY1 MA0095.1
SP1 MA0079.2
KLF5 MA0599.1
REST MA0138.1
ZNF263 MA0528.1
EGR3 MA0732.1
SP1 MA0079.1
KLF9 MA1107.1
CPDs count / total CPDs
Relative position
−1000
0
1000
naked
0h

Supplementary Figure 5.

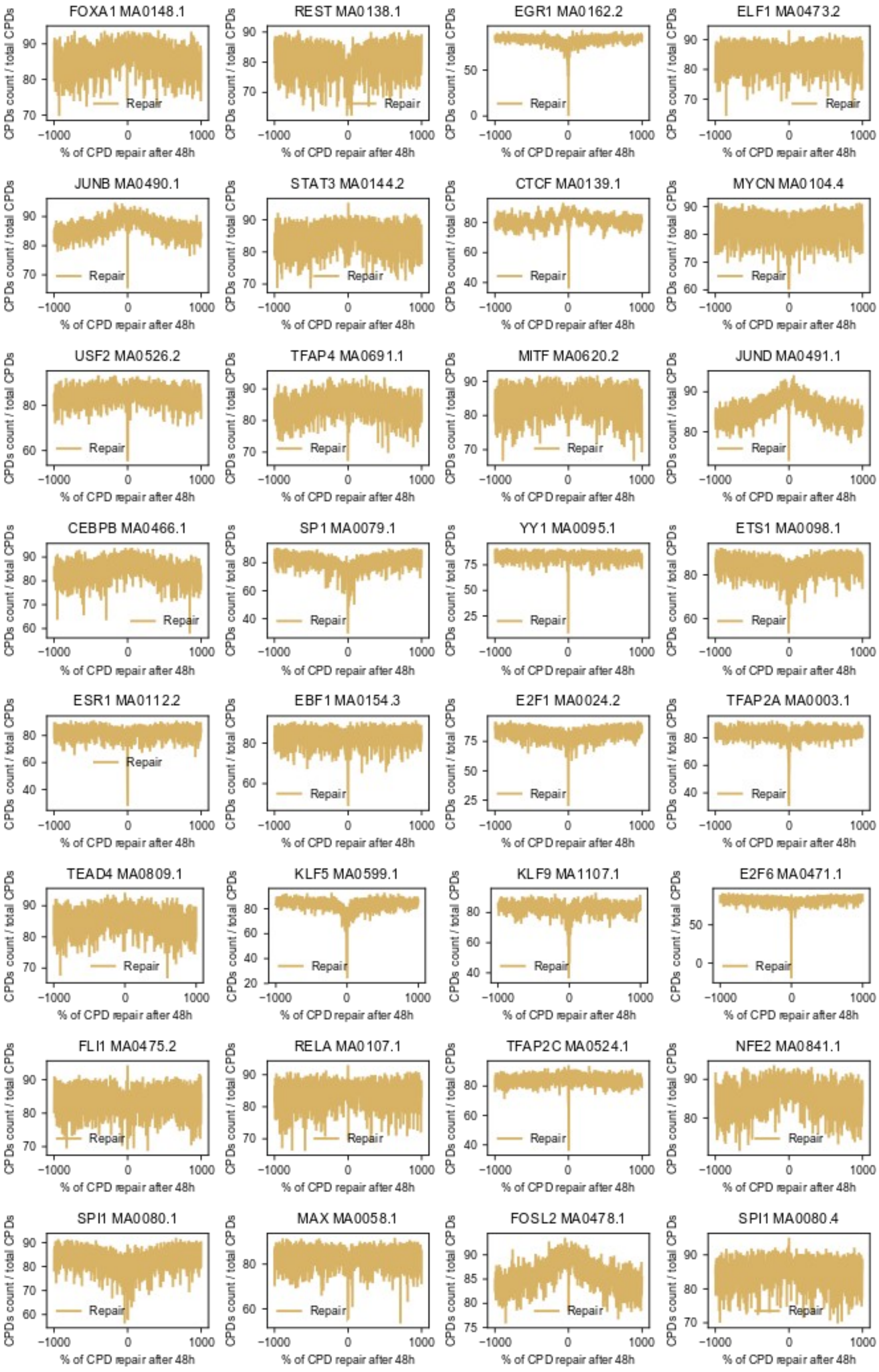


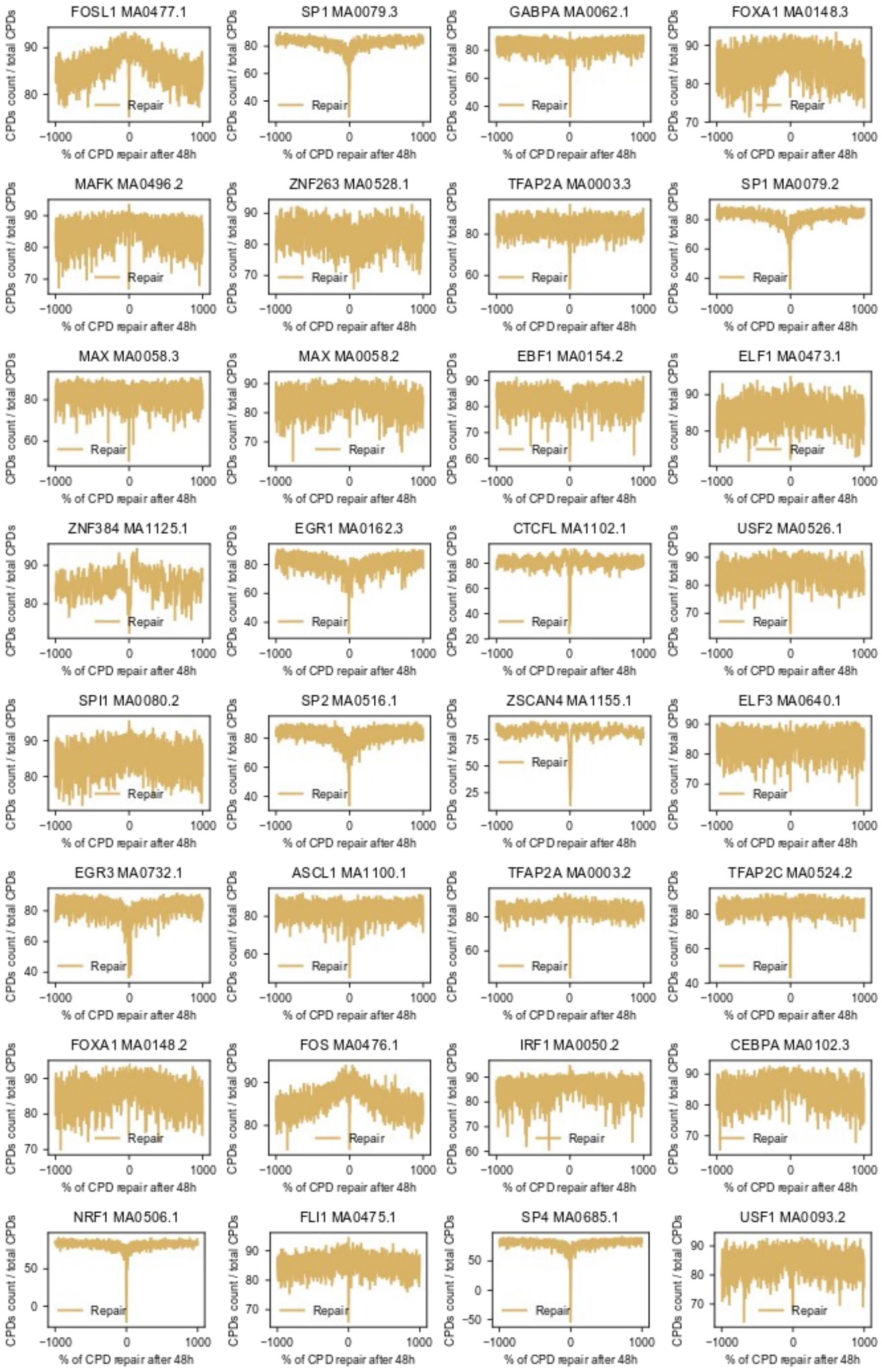

FOS_MA0476.1.centered
JUND_MA0491.1.centered
FOSL1_MA0477.1.centered
MAFK_MA0496.2.centered
FOSL2_MA0478.1.centered
NFE2_MA0841.1.centered
JUNB_MA0490.1.centered
SP1_MA0079.1.centered
SP1_MA0079.2.centered
CTCF_MA0139.1.centered
SP1_MA0079.3.centered
EGR1_MA0162.2.centered
YY1_MA0095.1.centered
EGR1_MA0162.3.centered
REST_MA0138.1.centered
SP2_MA0516.1.centered
Mutations
Muts. over exp. (%)
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6


ZNF263_MA0528.1.centered
Mutations
Relative position
ZNF263_MA0528.1.centered
Muts. over exp. (%)
Relative position
CTCFL_MA1102.1.centered
Mutations
Relative position
CTCFL_MA1102.1.centered
Muts. over exp. (%)
Relative position
KLF5_MA0599.1.centered
Mutations
Relative position
KLF5_MA0599.1.centered
Muts. over exp. (%)
Relative position
KLF9_MA1107.1.centered
Mutations
Relative position
KLF9_MA1107.1.centered
Muts. over exp. (%)
Relative position
SP4_MA0685.1.centered
Mutations
Relative position
SP4_MA0685.1.centered
Muts. over exp. (%)
Relative position
ZNF384_MA1125.1.centered
Mutations
Relative position
ZNF384_MA1125.1.centered
Muts. over exp. (%)
Relative position
EGR3_MA0732.1.centered
Mutations
Relative position
EGR3_MA0732.1.centered
Muts. over exp. (%)
Relative position
E2F1_MA0024.2.centered
Mutations
Relative position
E2F1_MA0024.2.centered
Muts. over exp. (%)
Relative position
FOXA1_MA0148.1.centered
Mutations
Relative position
FOXA1_MA0148.1.centered
Muts. over exp. (%)
Relative position
ESR1_MA0112.2.centered
Mutations
Relative position
ESR1_MA0112.2.centered
Muts. over exp. (%)
Relative position
FOXA1_MA0148.2.centered
Mutations
Relative position
FOXA1_MA0148.2.centered
Muts. over exp. (%)
Relative position
RELA_MA0107.1.centered
Mutations
Relative position
RELA_MA0107.1.centered
Muts. over exp. (%)
Relative position
FOXA1_MA0148.3.centered
Mutations
Relative position
FOXA1_MA0148.3.centered
Muts. over exp. (%)
Relative position
EBF1_MA0154.2.centered
Mutations
Relative position
EBF1_MA0154.2.centered
Muts. over exp. (%)
Relative position
E2F6_MA0471.1.centered
Mutations
Relative position
E2F6_MA0471.1.centered
Muts. over exp. (%)
Relative position
EBF1_MA0154.3.centered
Mutations
Relative position
EBF1_MA0154.3.centered
Muts. over exp. (%)
Relative position

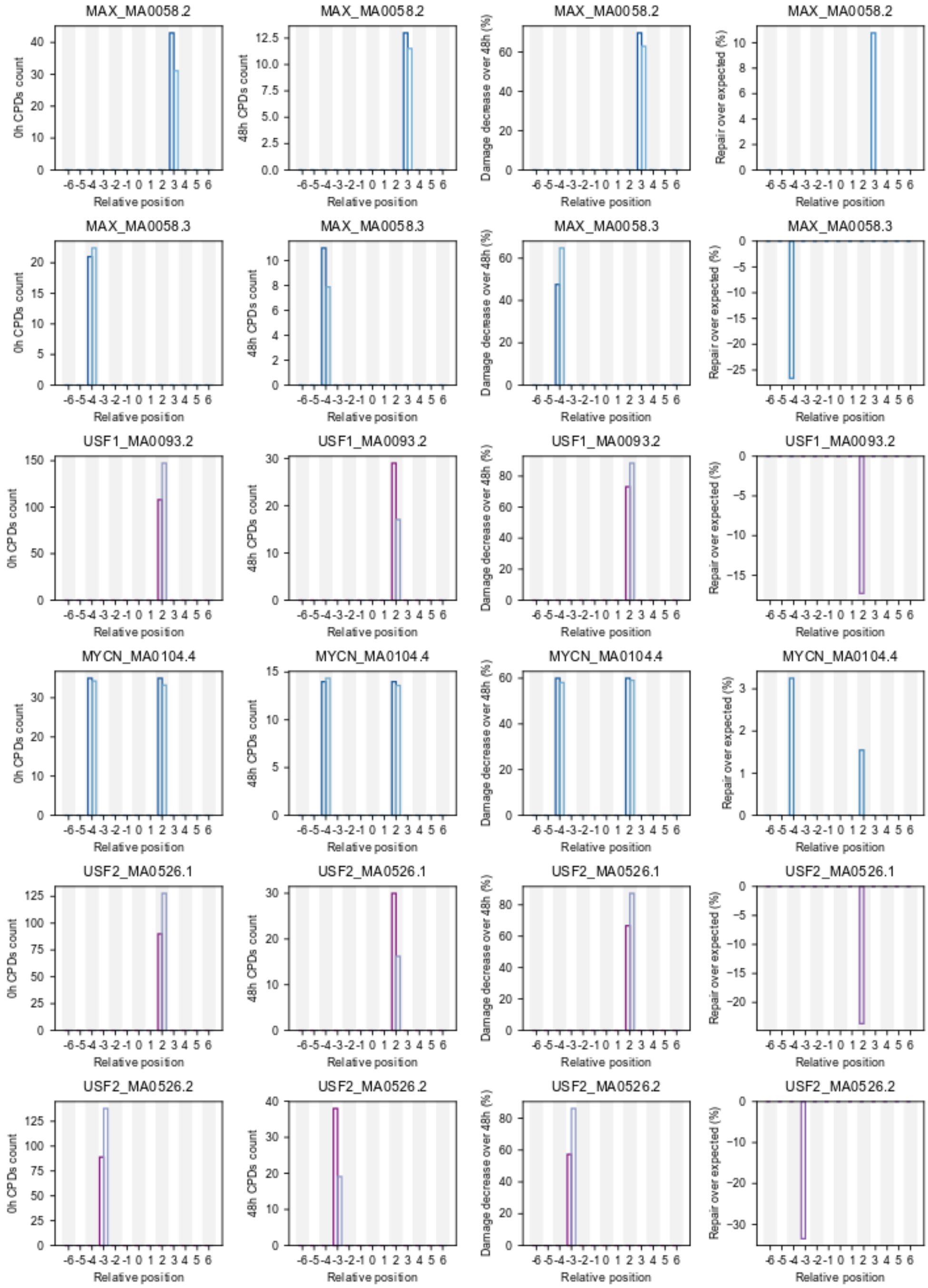

STAT3_MA0144.2.centered
STAT3_MA0144.2.centered
SPI1_MA0080.1.centered
SPI1_MA0080.1.centered
TEAD4_MA0809.1.centered
TEAD4_MA0809.1.centered
SPI1_MA0080.2.centered
SPI1_MA0080.2.centered
IRF1_MA0050.2.centered
IRF1_MA0050.2.centered
SPI1_MA0080.4.centered
SPI1_MA0080.4.centered
GABPA_MA0062.1.centered
GABPA_MA0062.1.centered
ETS1_MA0098.1.centered
ETS1_MA0098.1.centered
ELF1_MA0473.1.centered
ELF1_MA0473.1.centered
ELF3_MA0640.1.centered
ELF3_MA0640.1.centered
ELF1_MA0473.2.centered
ELF1_MA0473.2.centered
FLI1_MA0475.1.centered
FLI1_MA0475.1.centered
FLI1_MA0475.2.centered
FLI1_MA0475.2.centered
Mutations
Muts. over exp. (%)
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6

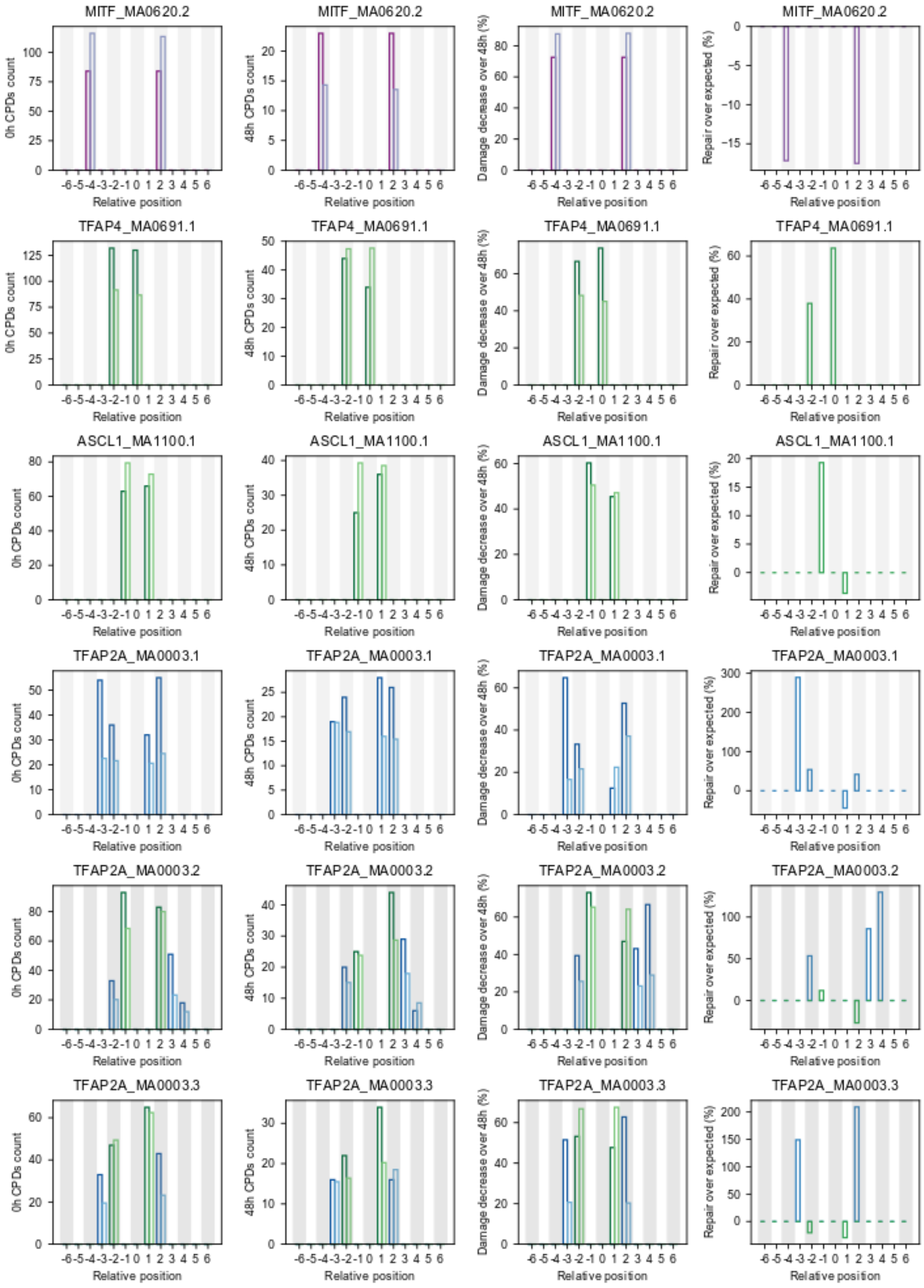
Supplementary Figure 6.

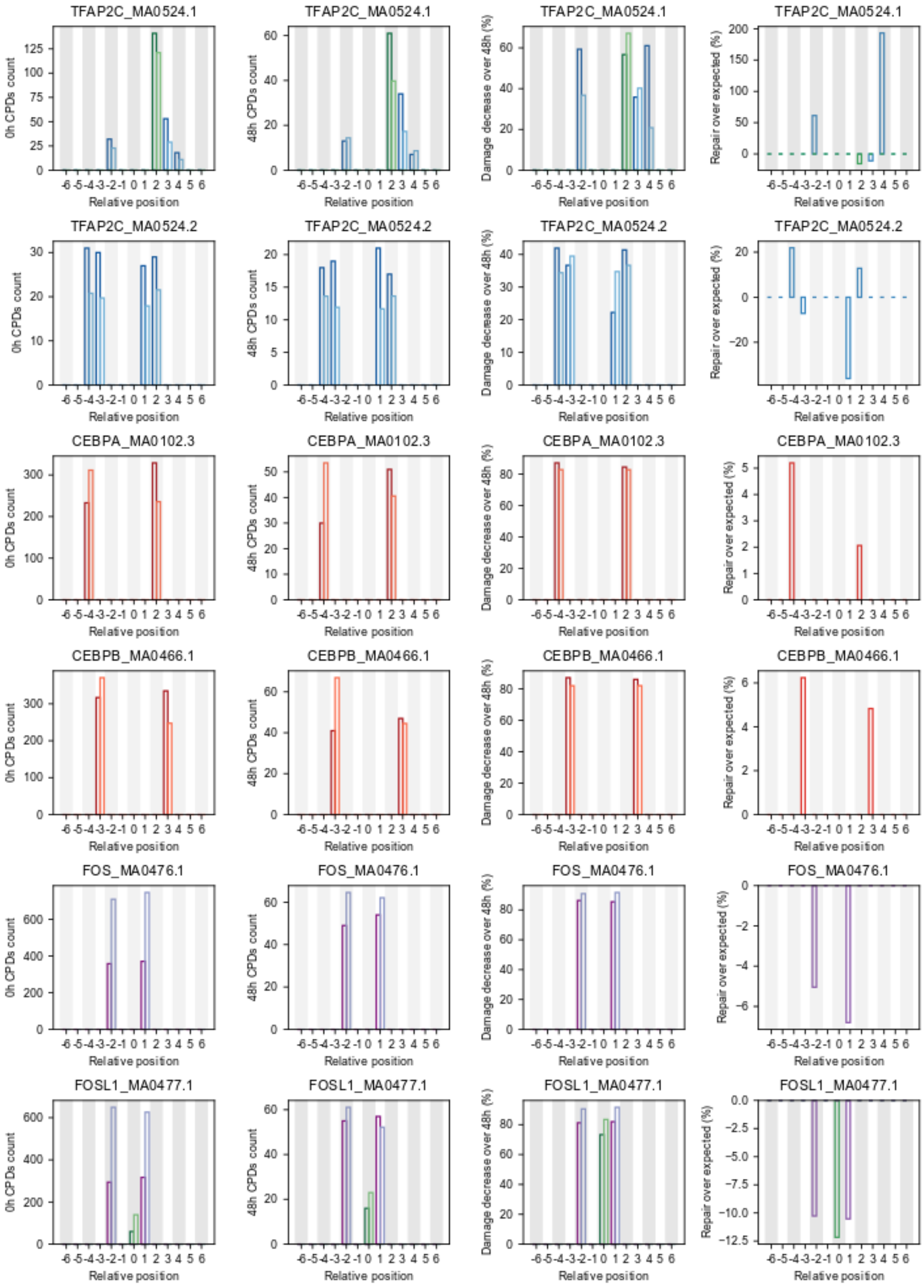



FOSL1 MA0477.1
SP1 MA0079.3
GABPA MA0062.1
FOXA1 MA0148.3
MAFK MA0496.2
ZNF263 MA0528.1
TFAP2A MA0003.3
SP1 MA0079.2
MAX MA0058.3
MAX MA0058.2
EBF1 MA0154.2
ELF1 MA0473.1
ZNF384 MA1125.1
EGR1 MA0162.3
CTCFL MA1102.1
USF2 MA0526.1
SPI1 MA0080.2
SP2 MA0516.1
ZSCAN4 MA1155.1
ELF3 MA0640.1
EGR3 MA0732.1
ASCL1 MA1100.1
TFAP2A MA0003.2
TFAP2C MA0524.2
FOXA1 MA0148.2
FOS MA0476.1
IRF1 MA0050.2
CEBPA MA0102.3
NRF1 MA0506.1
FLI1 MA0475.1
SP4 MA0685.1
USF1 MA0093.2
CPDs count / total CPDs
% of CPD repair after 48h
Repair
−1000
0
1000

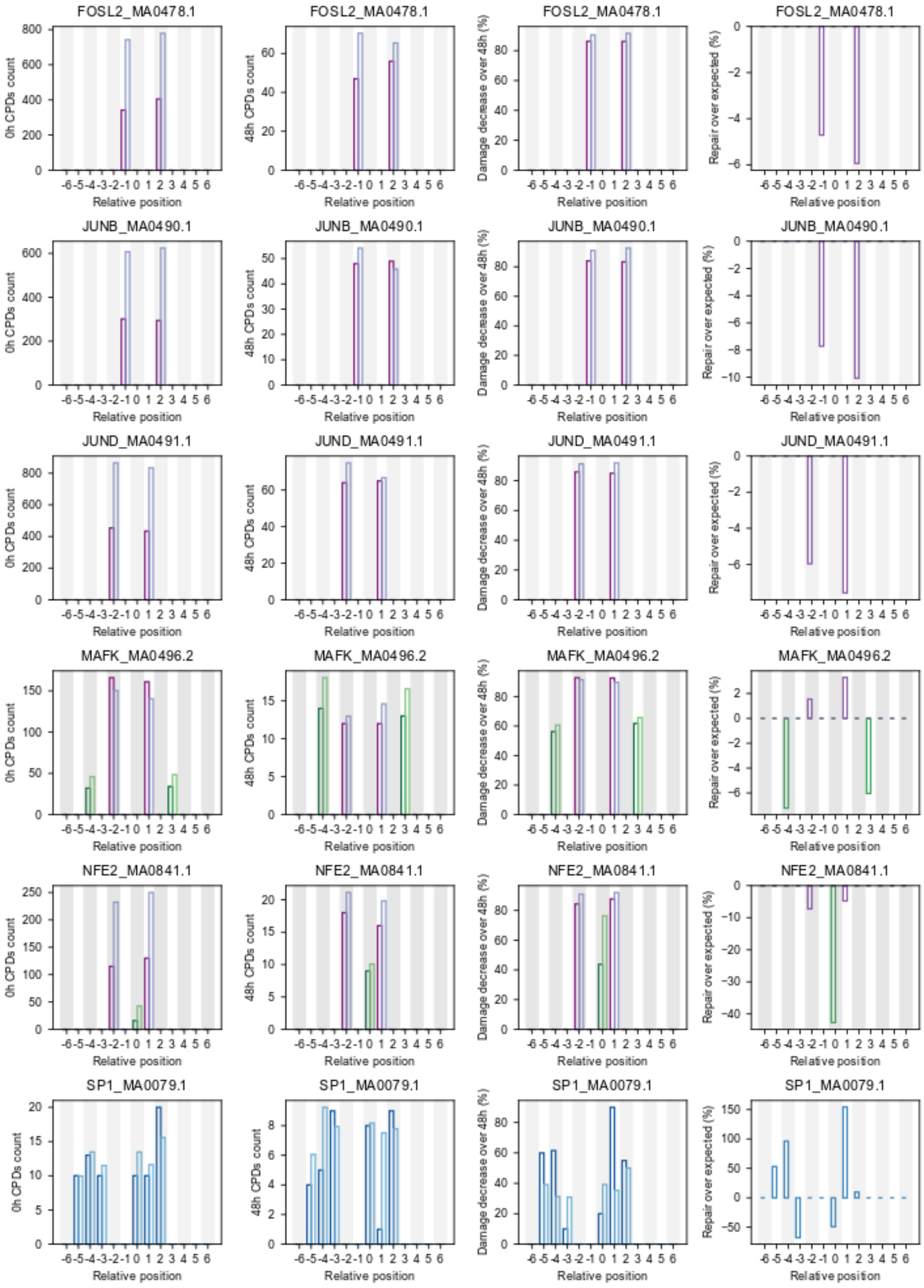
Supplementary Figure 7.

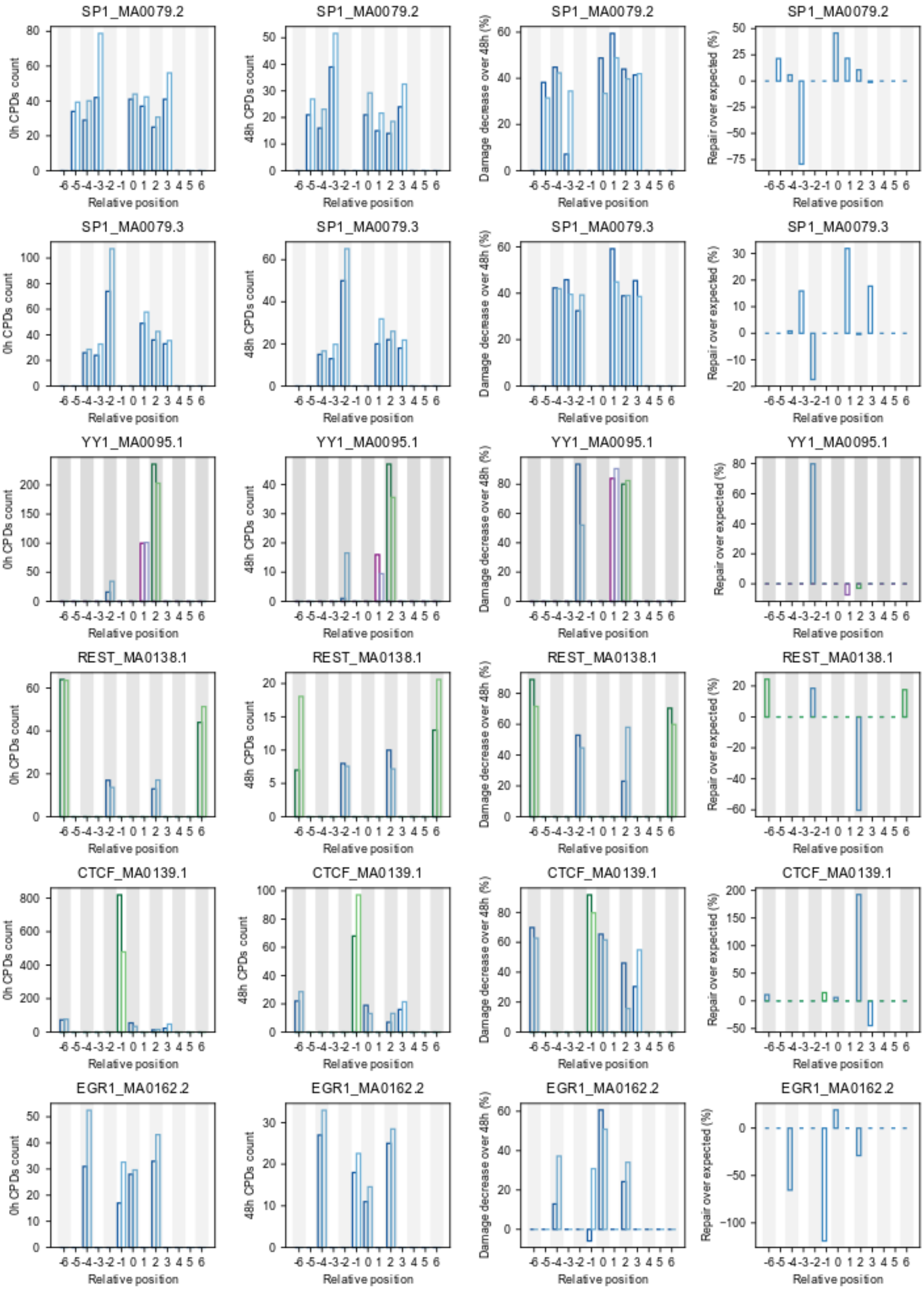


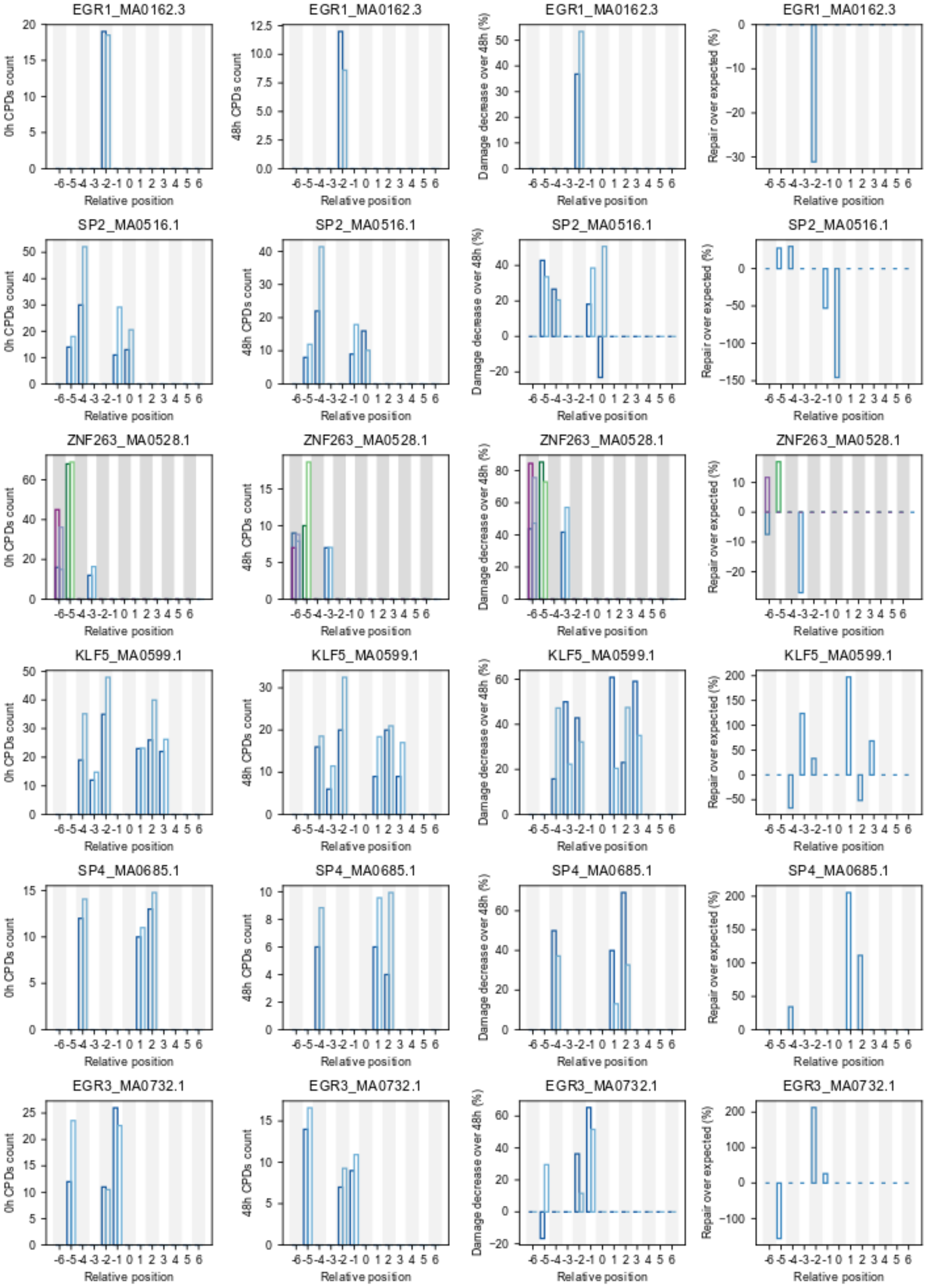

MITF_MA0620.2
0h CPDs count
48h CPDs count
Damage decrease over 48h (%)
Repair over expected (%)
TFAP4_MA0691.1
ASCL1_MA1100.1
TFAP2A_MA0003.1
TFAP2A_MA0003.2
TFAP2A_MA0003.3
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6

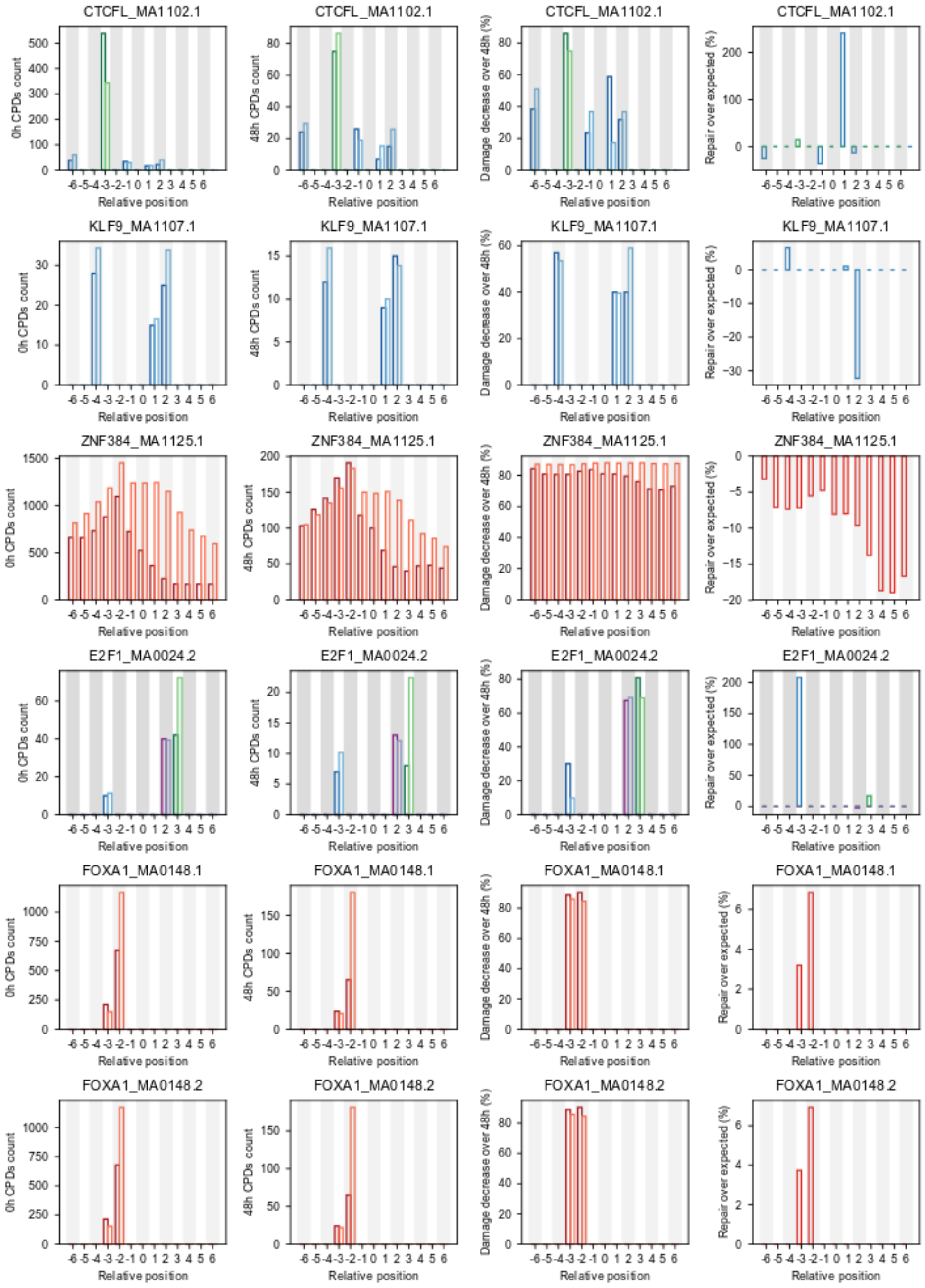


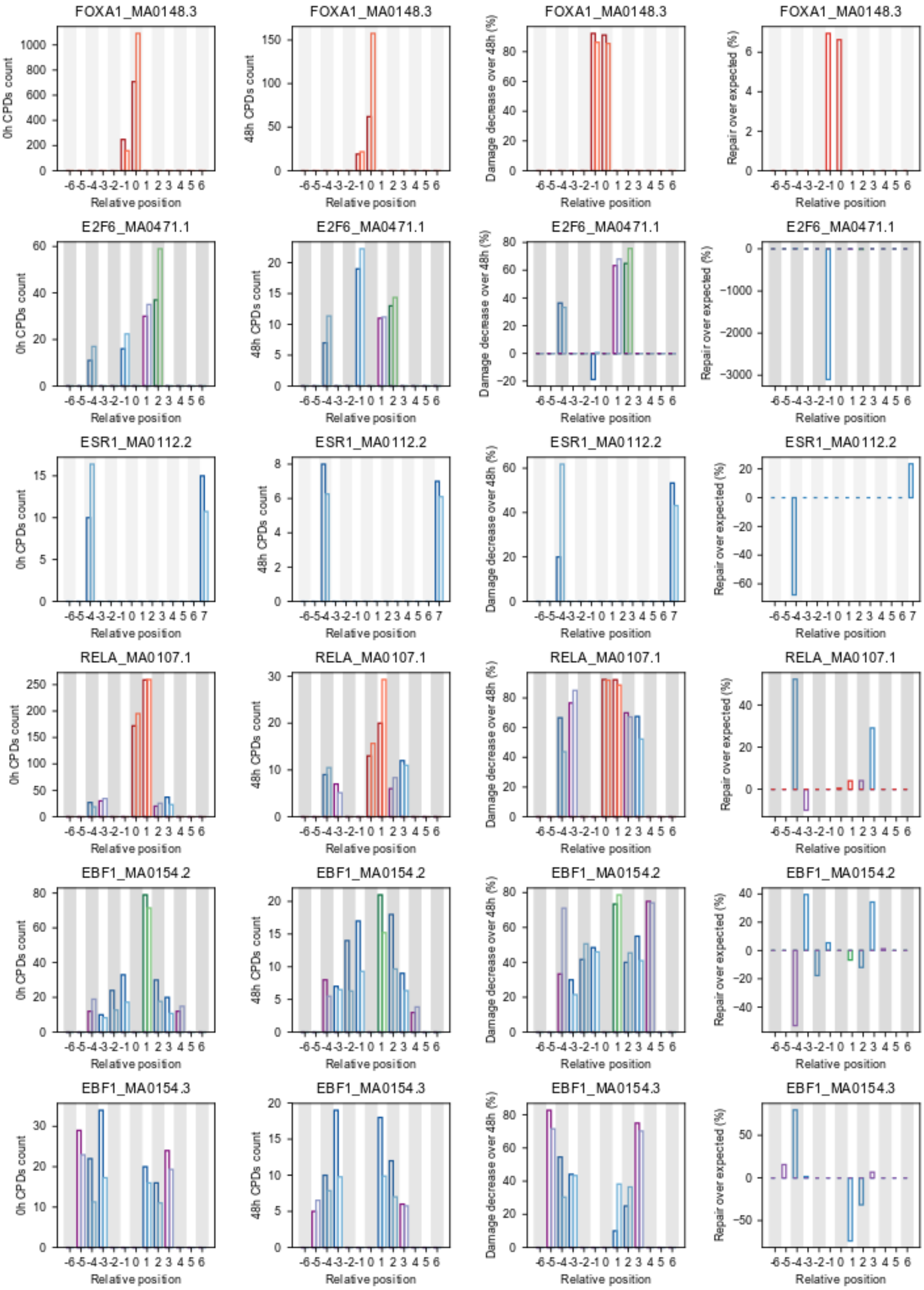
TFAP2C_MA0524.1
TFAP2C_MA0524.2
CEBPA_MA0102.3
CEBPB_MA0466.1
FOS_MA0476.1
FOSL1_MA0477.1
0h CPDs count
48h CPDs count
Damage decrease over 48h (%)
Repair over expected (%)
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6


FOSL2_MA0478.1
0h CPDs count
0
200
400
600
800
48h CPDs count
0
20
40
60
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
0
-2
-4
-6
-6-5-4-3-2-1 0 1 2 3 4 5 6
Relative position
JUNB_MA0490.1
0h CPDs count
0
200
400
600
48h CPDs count
0
10
20
30
40
50
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
0
-2
-4
-6
-8
-10
JUND_MA0491.1
0h CPDs count
0
200
400
600
800
48h CPDs count
0
20
40
60
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
0
-2
-4
-6
MAFK_MA0496.2
0h CPDs count
0
50
100
150
48h CPDs count
0
5
10
15
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
2
0
-2
-4
-6
NFE2_MA0841.1
0h CPDs count
0
50
100
150
200
250
48h CPDs count
0
5
10
15
20
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
0
-10
-20
-30
-40
SP1_MA0079.1
0h CPDs count
0
5
10
15
20
48h CPDs count
0
2
4
6
8
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
150
100
50
0
-50

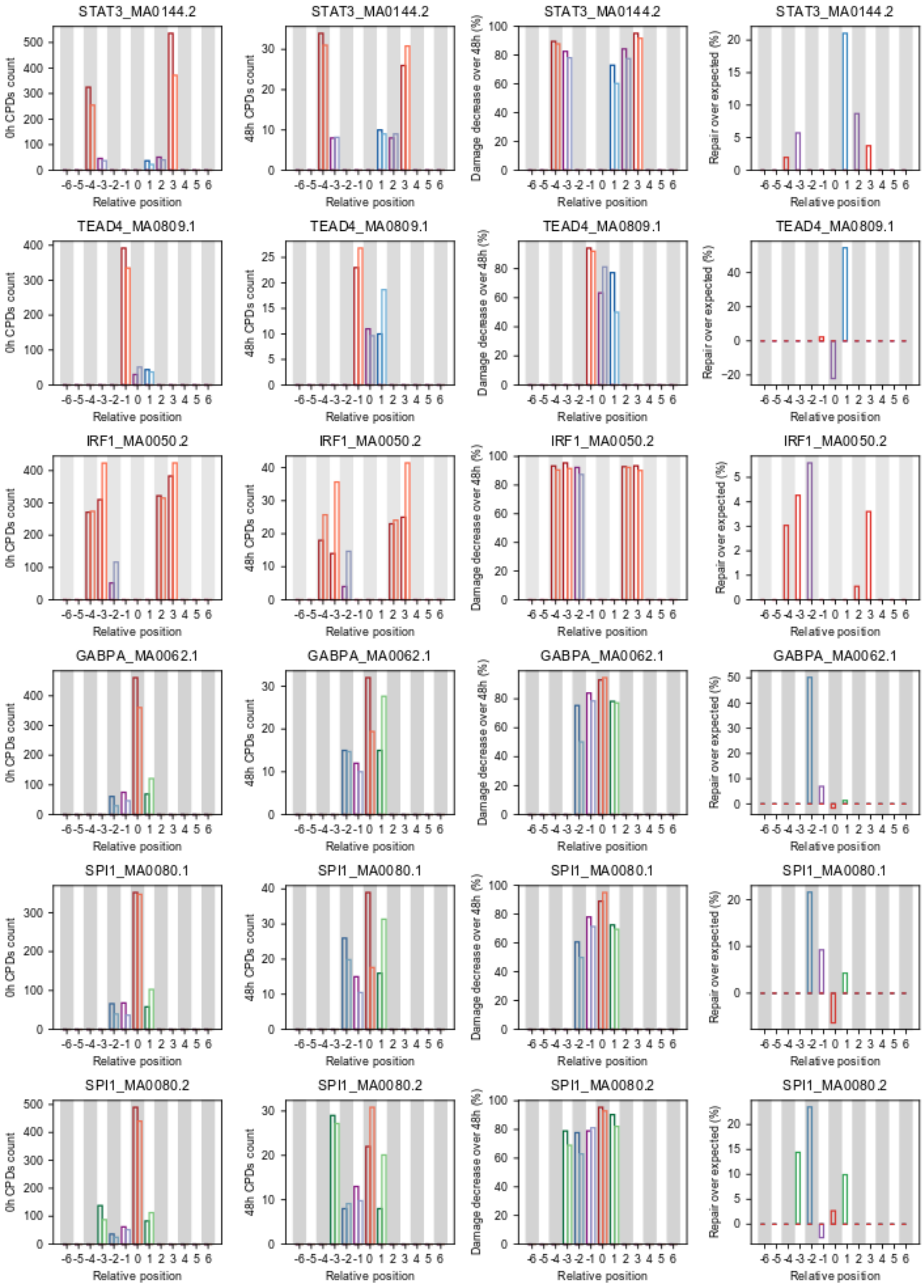

SP1_MA0079.2
0h CPDs count
0 20 40 60 80
48h CPDs count
0 10 20 30 40 50
Damage decrease over 48h (%)
0 20 40 60
Repair over expected (%)
−75 −50 −25 0 25 50
SP1_MA0079.3
0h CPDs count
0 20 40 60 80 100
48h CPDs count
0 20 40 60
Damage decrease over 48h (%)
0 20 40 60
Repair over expected (%)
-20 -10 0 10 20 30
YY1_MA0095.1
0h CPDs count
0 50 100 150 200
48h CPDs count
0 10 20 30 40
Damage decrease over 48h (%)
0 20 40 60 80
Repair over expected (%)
0 20 40 60 80
REST_MA0138.1
0h CPDs count
0 20 40 60
48h CPDs count
0 5 10 15 20
Damage decrease over 48h (%)
0 20 40 60 80
Repair over expected (%)
-60 -40 -20 0 20
CTCF_MA0139.1
0h CPDs count
0 200 400 600 800
48h CPDs count
0 20 40 60 80 100
Damage decrease over 48h (%)
0 20 40 60 80
Repair over expected (%)
-50 0 50 100 150 200
EGR1_MA0162.2
0h CPDs count
0 10 20 30 40 50
48h CPDs count
0 10 20 30
Damage decrease over 48h (%)
0 20 40 60
Repair over expected (%)
−100 −50 0
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Relative position

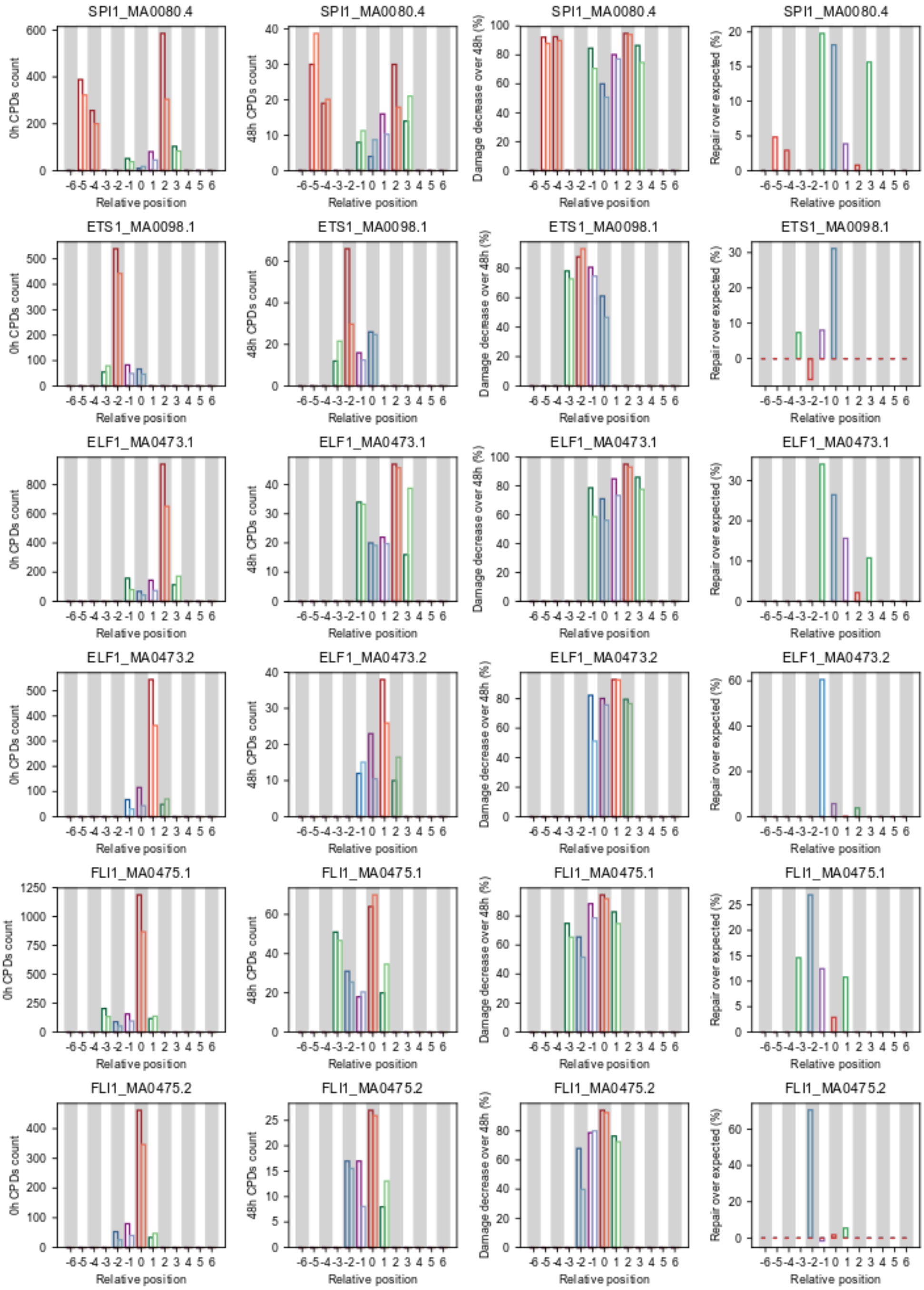

EGR1_MA0162.3
SP2_MA0516.1
ZNF263_MA0528.1
KLF5_MA0599.1
SP4_MA0685.1
EGR3_MA0732.1
0h CPDs count
48h CPDs count
Damage decrease over 48h (%)
Repair over expected (%)
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6

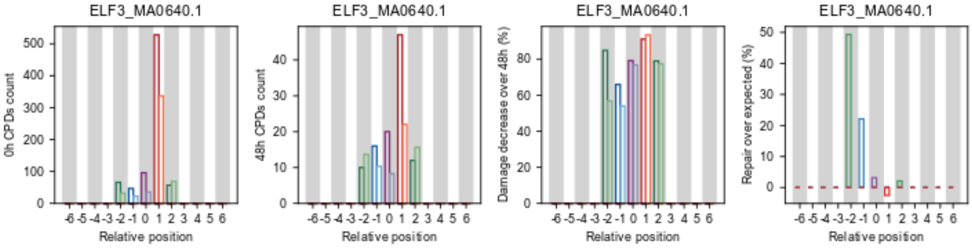

CTCFL_MA1102.1
KLF9_MA1107.1
ZNF384_MA1125.1
E2F1_MA0024.2
FOXA1_MA0148.1
FOXA1_MA0148.2
0h CPDs count
48h CPDs count
Damage decrease over 48h (%)
Repair over expected (%)
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6



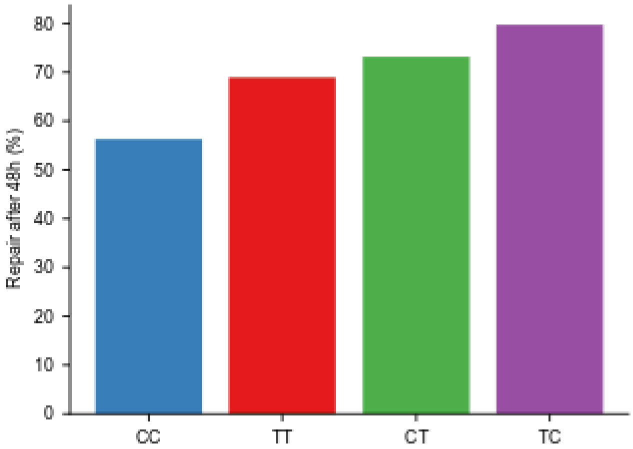
FOXA1_MA0148.3
0h CPDs count
0
200
400
600
800
1000
48h CPDs count
0
50
100
150
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
0
2
4
6
E2F6_MA0471.1
0h CPDs count
0
20
40
60
48h CPDs count
0
5
10
15
20
Damage decrease over 48h (%)
−20
0
20
40
60
80
Repair over expected (%)
−3000
−2000
−1000
0
ESR1_MA0112.2
0h CPDs count
0
5
10
15
48h CPDs count
0
2
4
6
8
Damage decrease over 48h (%)
0
20
40
60
Repair over expected (%)
−60
−40
−20
0
20
RELA_MA0107.1
0h CPDs count
0
50
100
150
200
250
48h CPDs count
0
10
20
30
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
0
20
40
EBF1_MA0154.2
0h CPDs count
0
20
40
60
80
48h CPDs count
0
5
10
15
20
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
−40
−20
0
20
40
EBF1_MA0154.3
0h CPDs count
0
10
20
30
48h CPDs count
0
5
10
15
20
Damage decrease over 48h (%)
0
20
40
60
80
Repair over expected (%)
−50
0
50
-6-5-4-3-2-1 0 1 2 3 4 5 6
-6-5-4-3-2-1 0 1 2 3 4 5 6 7
Relative position


STAT3_MA0144.2
0h CPDs count
48h CPDs count
Damage decrease over 48h (%)
Repair over expected (%)
TEAD4_MA0809.1
IRF1_MA0050.2
GABPA_MA0062.1
SPI1_MA0080.1
SPI1_MA0080.2
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
0
100
200
300
400
500
10
20
30
40
60
80
5
15
25
50
-20
1
2
3
4

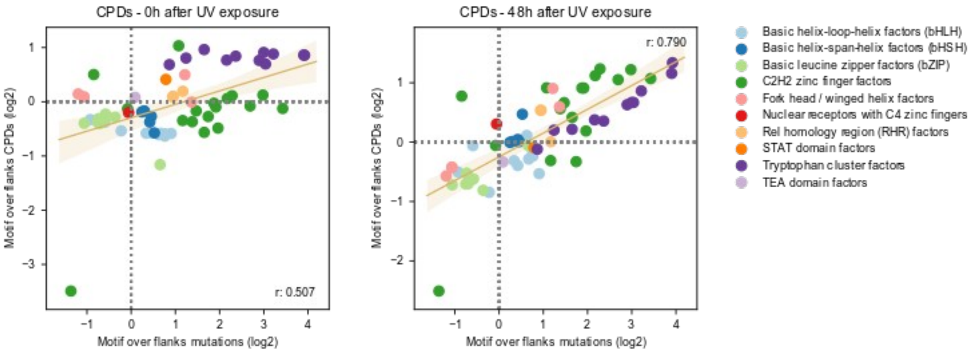

SPI1_MA0080.4
ETS1_MA0098.1
ELF1_MA0473.1
ELF1_MA0473.2
FLI1_MA0475.1
FLI1_MA0475.2
0h CPDs count
48h CPDs count
Damage decrease over 48h (%)
Repair over expected (%)
Relative position
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6


ELF3_MA0640.1
0h CPDs count
0
100
200
300
400
500
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Relative position
ELF3_MA0640.1
48h CPDs count
0
10
20
30
40
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Relative position
ELF3_MA0640.1
Damage decrease over 48h (%)
0
20
40
60
80
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Relative position
ELF3_MA0640.1
Repair over expected (%)
0
10
20
30
40
50
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Relative position

Supplementary Figure 8.



Supplementary Figure 9.



Supplementary figure legends

Supplementary Figure 1.
Observed and expected mutation rates across the stacked 2001-nucleotide sequences across all transcription factors analyzed.

Supplementary Figure 2.
Observed and expected number of UV-induced mutations within the motif and the percentage of increase or decrease with respect to the expectation (in blue).

Supplementary Figure 3.
Correlation between the number of CPDs of each type measured by Mao et al. (y axis) and Hu et al. (x axis), both within the binding motif (left) and at the flanks (right).

Supplementary Figure 4.
Distribution of CPDs formation rate across binding motifs all TFs analyzed, when the TF is bound to the DNA and in naked DNA.

Supplementary Figure 5.
CPDs formed at each position within the binding motifs (dark) and in the flanks (light), when evaluating positions with identical sequence context, and percentage of increase or decrease with respect to this expectation .

Supplementary Figure 6.
Percentage of CPDs repaired 48h after UV exposure across the stacked 2001-nucleotide sequences across all transcription factors analyzed.

Supplementary Figure 7.
From left to right, CPDs formed at each position within the binding motifs and in the flanks, CPDs remaining at each position within the binding motif 48h after UV exposure, percentage of repair at each position within the binding motif 48h after UV exposure and percentage of repair increase or decrease with respect to the flanks.

Supplementary Figure 8.

Percentage of CPDs of each type repaired at 48h after UV exposure at whole genome level.

Supplementary Figure 9.

Relationship between the ratio (log2) of motif-to-flanks CPDs (y-axis) computed immediately after irradiation and the ratio (log2) of motif-to-flanks mutations (x-axis), both 0h (left) and 48h (right) after UV exposure.

4. Discussion

Cancer is characterized by genetic alterations that lead to uncontrolled cell growth and tumor formation. Given its high incidence and poor prognosis, enormous amounts of resources and manpower have been invested in cancer research. Among the objectives have been the identification of the alterations responsible for driving the tumorigenic process, the discovery of new therapeutic opportunities and the identification of markers that could help predict patient's prognosis and drug response. In this process, large volumes of data have been generated, including DNA sequencing data coming from panels of selected genes, exomes and in recent years, whole genomes as well. Moreover, this data is usually generated by large international consortia and publicly available.

Besides being abundant and accessible, DNA sequencing data from tumors has other very unique features. In the first place, tumor cells tend to have higher proliferation rates than their healthy equivalents[9]. In addition, like all cells in the human organism, tumor cells are subject to mutational processes[23]. However, during tumorigenesis, the clonal expansion amplifies the mutations present in the founder cell[153]. As a result, tumors mutations contain a record of all mutations experienced by that founder cell both before and after its transformation. The sequencing a tumor sample is usually complemented by sequencing a patient’s blood sample as well, and by comparing the mutations called in these two samples, germline variants can be identified and filtered out. Altogether, these characteristics make tumors a very valuable resource for the study of somatic mutations and mutational processes in general, and may reveal patterns that can be extrapolated to non malignant cells as well. Moreover, alterations of a wide number of pathways including DNA repair mechanisms have been found across tumor cohorts, either due to somatic mutations or other mechanisms such as epigenetic modifications and germline mutations[154] [155], which can be used for their study.

In summary, cancer genomics data is an extremely useful resource for the study of mutational processes and cellular behavior, and despite being usually generated with the aim of understanding tumorigenesis, it has also been used by the scientific community to tackle questions that go beyond cancer biology and I believe this thesis is an example of it.

During my PhD I have worked towards trying to understand how mutations accumulate along the genome. Being a cancer genomics laboratory, our initial motivation to do so was to improve the performance of cancer driver identification methods and the interpretation of their results. However, early after initiating this work, we developed a genuine interest in understanding the intricacies of the mutational processes, in an effort to elucidate the basic biology behind DNA damage and repair mechanisms. In this regard, we have studied the impact of local genomic features to the generation of DNA damage and the activity of the associated repair mechanisms. More specifically, this thesis includes two research projects where we have described local mutational biases at exonic regions and transcription factor binding sites and provided mechanistic explanation for them.

In the first of the projects aforementioned, we showed that most tumors tend to accumulate fewer exonic somatic mutations than expected given their DNA sequence. A straightforward explanation for this could be that mutations at exons are negatively selected due to their highest harmful potential. However, this was discarded after observing no differences between the observed and expected proportions of synonymous and non synonymous mutations. We believe this to be an interesting result by itself, since it indicates that negative selection in cancer cells is not a widespread phenomenon. Consequently, we determined that differences in the likelihood of mutations to occur between exons and introns could be the cause of the observed reduced exonic mutation rate. Given that this phenomenon was most evident in samples where the proofreading subunit of the DNA polymerase had been inactivated, leading to increased frequency of mismatches generation, this mutational process became the focus of our study.

Differences between the rate of observed and expected mutations along the genome may be caused either by deviations from the expectation in damage susceptibility or in repair efficiency. No data regarding mismatches introduction rate or mismatch repair activity along the genome was available at the time this project took place. Thus, in order to discern between these two processes, mutations from patients with high rate of mismatches generation caused by inactivation of the proofreading subunit of a DNA polymerase were compared with mutations from patients with this same type of alterations also carrying a germline inactivation of the mismatch repair machinery. In this second group of tumors, mismatches are not not repaired and therefore somatic mutations reflect how mismatches have been generated, allowing to study this process separately from the repair. We found that tumors with inactivated mismatch repair showed no reduced exonic mutation rate, pointing towards it being caused by differences in repair efficacy. This is an example of using tumors with inactivation of specific pathways to gain insights into their performance, in this case inferring the mismatch repair activity.

Moreover, we were able to relate the repair differences to an exonic enrichment of the H3K36me3 histone modification, which has been shown to be involved in the recruitment of the mismatch repair machinery [156]. Thus, in the mechanistic model we envision, the mismatch repair machinery is preferentially located at the H3K36me3 rich exons before replication initiation and therefore before mismatches have been generated. Next, when replication starts, nucleosomes are disassembled from the DNA, together with the mismatch repair machinery. Then, if mismatches are introduced during this process, the exonic ones are preferentially repaired since mismatch repair machinery is found in the nearby area.

We believe increased repair activity in exons to be intrinsic to mismatch repair pathway and therefore conserved across tumor types and healthy cells. However, cells suffer DNA lesions different than mismatches that are repaired by other repair mechanisms apart from mismatch repair, whose activity may or may not be altered in exonic regions. Thus, we expect the reduced exonic mutation rate effect to be more clearly appreciable in tissues in which the vast majority of mutations are the consequence of unrepaired mismatches, while other tissues where the majority of mutations are contributed by other mutational processes, the effect of over repaired exonic mismatches may be diluted.

The fact that mutations are not equally likely to occur in introns and exons even in a context of no selective pressure may have implications both at the level of cancer genomics and evolution studies. For instance, some methods for the identification of cancer driver genes use the intronic mutation rate as background model to which compare the exonic mutations and therefore assess positive selection [157]. According to our results, variation in terms of mutation burden between introns and exons could be attributed not only to positive or negative selection but also to differences in the activity of certain mutational processes between these two regions. Therefore, we believe that modeling and including this phenomena to the driver detection methods mentioned above could improve their performance and yield more accurate results.

Given that tumor cells presenting a reduced exonic mutation rate have an unaltered mismatch repair mechanism, we believe its increased repair efficiency in exons to occur in healthy cells as well. Thus the reduced exonic mutation rate could perhaps be observed in germline mutations as well. If increased mismatch repair efficacy contributed to the depletion of exonic germline mutations, this would question the practice of using the intronic mutation rate to estimate the neutral rate of evolution used as a reference when determining selective pressure in genes [158] [159].

To conclude, it is tempting to speculate about the reasons why mismatch repair machinery seems to be more efficient at repairing exonic DNA. Could this trait have been selected throughout evolution in order to protect those genomic regions where mutations could be more harmful?

In the second project of this thesis, we evaluated the influence exerted by the presence of transcription factors to the way mutations accumulate in their binding sites. This project was a follow-up of a previous study conducted in our laboratory by Sabarinathan, R. et al, [142] where they described mutational hotspots at active transcription factor binding sites, mainly in UV exposed melanoma samples, and attributed them to reduced activity of the nucleotide excision repair caused by the physical presence of the transcription factor bound to the DNA. Additionally, two papers by Mao, P. et al. And Elliot, K. et al. Were published by the end of 2018 reporting that the mutational hotspots present in the binding motifs of the ETS family of transcription factors were caused by increased sensibility to UV damage rather than impaired repair activity [144] [145], suggesting that different families of TF could differently alter the way UV damage is created and repaired. Based on this, we devised a project with the aim to classify the transcription factors by their type of binding domain and search for differences in the way they impact UV damage formation and/or repair in order to create a comprehensive mechanistic landscape.

We observed that the generation of UV induced mutations in the TFBS is a complex process, and that the way TFs affect both UV damage formation and its repair is highly variable across families. We identified TFs whose binding motifs showed increased mutation rate because of higher damage formation. Increased DNA UV damage susceptibility is thought to be caused by conformational changes in the DNA double helix [160] [161], which in this case would be triggered by the binding of the TF. Hence, if specific DNA conformations become more susceptible to UV damage, it is not unreasonable to think that other DNA conformations could exert the opposite effect. Indeed, we could identify a family of TFs whose binding motifs showed decreased UV damage formation, which correlated with reduced mutation rate, hinting towards a putative protective effect of these TFs. This could be tested by analyzing the way TF modify the distance and torsion angle between the atoms from adjacent pyrimidines involved in the UV lesions formation. However, we dropped this idea due to the scarcity of structural data available at the Protein Data Bank (PDB) [162]. Additionally, we could observe that most of the TF families presenting mutational hotspots in their binding sites had reduced nucleotide excision repair activity in these areas. In summary, we identified two different mechanisms responsible for the mutation rate enrichment in transcription factor binding sites of different families: one related to increased UV damage sensibility and another characterized by reduced repair efficiency. Moreover, we identified a fraction of TFs that exerted a protective effect from the UV damage in their binding sites, which led to fewer mutations than expected.

Thus, we envision the influence of TF binding on UV damage formation to be consequence of conformational changes in the DNA double-helix structure that make damage formation more or less favorable. Regarding the TF influence on repair though, we imagine it to be a less specific interaction, triggered by the physical interference between the repair machinery and the bound transcription factor. The fact that not all TFs impair nucleotide excision repair activity could be due to differences in the size or DNA binding mechanism of the TFs, or even to the fact that some UV lesions could prevent the binding of specific TF, therefore eliminating any potential interferences between the protein bound to the DNA and the nucleotide excision repair mechanism, and therefore allowing repair to occur in a normal manner.

A novel approach used in this project was to learn the intensity of nucleotide excision repair over time at different regions of the genome and use it to predict the proportion of damage remaining after repair using data from a second dataset containing only UV damage measured right after exposure. I believe this could be used to predict the evolution of lesions caused by other damaging agents also repaired by nucleotide excision repair for which data regarding damage distribution immediately after exposure is available, such as cisplatin [163].

Throughout this work we focused on two main factors that could trigger mutational biases: differential UV damage formation and differential NER activity. However, there is a third step that could also be considered: not all unrepaired UV lesions are equally likely to become mutations. Error-prone polymerases tend to introduce adenines opposite to UV induced DNA adducts. Thus, only when these lesions are formed by cytosines, mispairs may be introduced. Additionally, these cytosines have increased likelihood of being deaminated to uracil, leading to the insertion of adenines opposite to them during replication, generating mispairs as well [164]. Moreover, the deamination rate seems to depend on their methylation status, even though contradictory results have been published in this regard [165]. Therefore, information regarding the methylation status of cytosines or other features could, perhaps, provide additional information about the likelihood of unrepaired UV damage to promote the appearance of mispairs during replication, which would finally become mutations. Taking this into account, we could probably improve the accuracy of our model, especially when establishing the relationship between unrepaired UV damage and mutations.

Moreover, it is worth highlighting that our project has been focused on the analysis of a specific type of damage, the CPDs, which are the most abundant UV induced lesions. However, other types of UV induced damage exist, such as the 6-4PP [166]. Unfortunately, even though whole genome maps of 6-4PP have been generated [152], not all the data needed to run our analysis is available for this type of lesions. However, if this data was ever obtained, it would be of interest to compare the effect exerted by the transcription factors presence in the generation of both CPDs and 6-4PP. Given that the atoms involved in the abnormal covalent bonds created in these two types of DNA adducts are different [167], it may be possible that the binding of a given transcription transcription could stimulate the formation of one type of DNA damage while impairing the other. Differences in the effect of the transcription factor on the repair activity between these two types of damage could also be studied even though, in my opinion, it is unlikely to observe any, since the repair mechanism for repairing both types of lesion is common. However, differences in their repair between these two types of UV damage may exist, since CPDs are known to be repaired slower than 6-4PP [168].

As mentioned at the beginning of this section, our laboratory developed its interest in understanding how mutations accumulate along the genome with the objective of improving the performance and results interpretation of our cancer driver identification methods. More specifically, Sabarinathan et al. began studying the accumulation of mutations at TFBS of UV exposed melanomas after observing that some of these methods tended to output a huge excess of statistically significant cancer driver TFBS in these samples caused by an excess of mutations. They discovered that this was caused by a repair deficiency in these areas rather than any kind of positive selection and therefore concluded that the list of potential cancer driver TFBS obtained was not trustworthy. However, it is still likely that mutations at specific TFBS have cancer driver activity, and I believe a nice follow-up of our project could be to try to analyze the cohorts of UV exposed melanomas in order to identify driver TFBS, considering that mutations do not accumulate as expected in these areas. To do so, it would be necessary to update the way the mutational background of the TFBS is computed in order to take into account the biases both in UV damage formation and repair that we have learned occur in there. Since not all positions of TF binding motifs are equally relevant for their recognition, it could also be of interest to incorporate information about entropy changes in the motif triggered by each mutation. Perhaps signals of positive selection could be detected by using this measure as a score. This could probably allow to differentiate those cases where a binding motif is disrupted upon mutation from those where it is modified in a way that increases binding affinity or it becomes similar to a binding motif from a different TF. Similarly, this approach could also be used to identify TF binding motifs generated de novo as a result of somatic mutations happening in DNA regions with sequences similar to known binding motifs, perhaps using as measure this entropy variation score.

In summary, during the two projects encompassed by this thesis, we have used cancer genomics data to evaluate the activity of specific mutational processes in exons and in transcription factor binding sites, both at damage formation and repair levels, with the goal to understand how mutations are generated in these areas. In the second of these projects, we made use of damage distribution data measured at several time points after induction. This allowed us to evaluate how UV damage was generated and also to infer the repair activity by comparing time-points. In the first project though, this data was not available. Therefore we focused our analysis on comparing the mutations distribution in tumors where the repair mechanism of interest was active against others where it was not. This way, we could infer differences in repair intensity by identifying discrepancies between these two groups of samples in the way mutations had accumulated.

As mentioned at the beginning of this discussion, most of our analysis were performed using cancer genomics data. However, I believe our observations to be extendable to non malignant cells as well, and even though our motivation to start these projects was to improve the results and output interpretation of cancer driver genes identification methods, I believe that the main contribution of this thesis has been to emphasize the high degree of variation in the way somatic mutations accumulate along the genome, and the relevance of comprehending the mutational processes cells are exposed to in order to understand this heterogeneity.

5. Conclusions

A shared goal between the two research projects included in this thesis has been to study how somatic mutations accumulate along the genome. Our research in this direction has allowed us to conclude that:

- Somatic mutations are not uniformly generated along the genome.
- Heterogeneity in the formation of somatic mutations can be observed at different scales.
- The way somatic mutations are distributed along a tumor genome is determined by the mutational processes that its cells have been exposed to.

Additionally, conclusions specific to each project can be extracted as well. In the case of the study titled “Reduced mutation rate in exons due to differential mismatch repair”, these are:

- Exons have fewer mutations than expected given their nucleotide sequence.
- Reduced exonic mutation rate is evident in those mutations originated from polymerase mismatches.
- Polymerase mismatches don’t seem to be less likely to occur in exonic regions.
- DNA mismatch repair exerts a more efficient repair in exons.
- Increased DNA mismatch repair activity is driven by an enrichment of H3K36me3 in exons.

In the project named “Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites” we concluded that:

- Active transcription factor binding sites tend to be mutational hotspots, although the binding sites of at least one family of transcription factors are depleted of somatic mutations in UV exposed samples.
- UV induced DNA damage formation may be increased or depleted as a consequence of the transcription factor protein binding to its motif, in a family specific manner.
- As a general rule, bound transcription factors reduce the activity or efficacy of the nucleotide excision repair machinery across most transcription factor families.

6. Bibliography

1. Bianconi, E., Piovesan, A., Facchin, F. & Beraudi, A. An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, (2013).
2. Younger, P. Stedman's Medical Dictionary, 28th ed. *Ref. Rev.* (2007) doi:10.1108/09504120710719671.
3. Goldblum, J., Weiss, S. & Folpe, A. *Enzinger and Weiss's Soft Tissue Tumors*. (Elsevier, 2013).
4. Chaffer, C. L. & Weinberg, R. A. A Perspective on Cancer Cell Metastasis. *Science* **331**, 1559–1564 (2011).
5. McFadden, M. E. & Sartorius, S. E. Multiple systems organ failure in the patient with cancer. Part I: pathophysiologic perspectives. *Oncol. Nurs. Forum* **19**, 719–724 (1992).
6. Darmon, M., Ciroldi, M., Thiery, G., Schlemmer, B. & Azoulay, E. Clinical review: specific aspects of acute renal failure in cancer patients. *Crit. Care Lond. Engl.* **10**, 211 (2006).
7. Williams, M. D. *et al.* Hospitalized cancer patients with severe sepsis: analysis of incidence, mortality, and associated costs of care. *Crit. Care* **8**, R291 (2004).
8. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, (2000).
9. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

10. Ullrich, A. *et al.* Human epidermal growth factor receptor cDNA sequence and aberrant expression of the amplified gene in A431 epidermoid carcinoma cells. *Nature* **309**, 418–425 (1984).

11. Simon, M.-P. *et al.* Deregulation of the platelet-derived growth factor β-chain gene via fusion with collagen gene COL1A1 in dermatof ibrosarcoma protuberans and giant-cell fibroblastoma. *Nat. Genet.* **15**, 95–98 (1997).

12. Friend, S., Bernanrds, R., Rogelj, S., Weinberg, R. A. & Rapaport, J. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643–646 (1986).

13. Kerr, J. F. R., Wyllie, A. H. & Currie, A. R. Apoptosis: A Basic Biological Phenomenon with Wide-ranging Implications in Tissue Kinetics. *Br. J. Cancer* **26**, 239–257 (1972).

14. Korsmeyer, S. J. Chromosomal translocations in lymphoid malignancies reveal novel proto-oncogenes. *Annu. Rev. Immunol.* **10**, 785–807 (1992).

15. Nam, K., Piatyszek, M., Prowse, K. & Harley, C. Specific association of human telomerase activity with immortal cells and cancer. *Science* **266**, (1994).

16. Baeriswyl, V. & Christofori, G. The angiogenic switch in carcinogenesis. *Semin. Cancer Biol.* **19**, 329–337 (2009).

17. Christofori, G. & Semb, H. The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends Biochem. Sci.* **24**, 73–76 (1999).

18. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

19. Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. Somatic versus germinal mutation. *Introd. Genet. Anal. 7th Ed.* (2000).

20. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).

21. Feuk, L., Carson, A. & Scherer, S. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).

22. Shlien, A. & Malkin, D. Copy number variations and cancer. *Genome Med.* **1**, 62 (2009).

23. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

24. Sabarinathan, R. *et al.* The whole-genome panorama of cancer drivers. *bioRxiv* 190330 (2017) doi:10.1101/190330.

25. Martincorena, I., Raine, K., Gerstung, M. & Dawson, K. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041 (2017).

26. Orr, H. A. Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.* **10**, 531–539 (2009).

27. Greaves, M. & Maley, C. C. CLONAL EVOLUTION IN CANCER. *Nature* **481**, 306–313 (2012).

28. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).

29. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer genes. *Nature* **499**, 214–218 (2013).

30. Wolfe, K. H., Sharp, P. M. & Li, W.-H. Mutation rates differ among regions of the mammalian genome. *Nature* **337**, 283–285 (1989).

31. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).

32. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).

33. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinforma. Oxf. Engl.* **29**, 2238–2244 (2013).

34. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).

35. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071–1078 (2009).

36. Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome Med.* **2**, 54 (2010).

37. Rastogi, R., Ashok Kumar, R., Tyagi, M. & Sinha, R. Molecular Mechanisms of Ultraviolet Radiation-Induced DNA Damage and Repair. *J. Nucleic Acids* **592980**, (2010).

38. Lomax, M. E., Folkes, L. K. & O'Neill, P. Biological Consequences of Radiation-induced DNA Damage: Relevance to Radiotherapy. *Clin. Oncol.* **25**, 578–585 (2013).

39. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169–185 (2004).

40. Talpaert-Borlè, M. Formation, detection and repair of AP sites. *Mutat. Res. Mol. Mech. Mutagen.* **181**, 45–56 (1987).

41. Boiteux, S. & Guillet, M. Abasic sites in DNA: repair and biological consequences in Saccharomyces cerevisiae. *DNA Repair* **3**, 1–12 (2004).

42. Soll, J. M., Sobol, R. W. & Mosammaparast, N. Regulation of DNA Alkylation Damage Repair: Lessons and Therapeutic Opportunities. *Trends Biochem. Sci.* **42**, 206–218 (2017).

43. Cadet, J. & Wagner, J. R. DNA Base Damage by Reactive Oxygen Species, Oxidizing Agents, and UV Radiation. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

44. Kow, Y. W. Repair of deaminated bases in DNA. *Free Radic. Biol. Med.* **33**, 886–893 (2002).

45. Woźniak, K. & Błasiak, J. Recognition and repair of DNA-cisplatin adducts. *Acta Biochim. Pol.* **49**, 583–596 (2002).

46. Stowers, S. J. & Anderson, M. W. Formation and persistence of benzo(a)pyrene metabolite-DNA adducts. *Environ. Health Perspect.* **62**, 31–39 (1985).

47. Kunkel, T. A. & Bebenek, K. DNA Replication Fidelity. *Annu. Rev. Biochem.* **69**, 497–529 (2000).

48. Negishi, K., Bessho, T. & Hayatsu, H. Nucleoside and nucleobase analog mutagens. *Mutat. Res. Genet. Toxicol.* **318**, 227–238 (1994).

49. Caldecott, K. Single-strand break repair and genetic disease. *Nat. Rev. Genet.* **9**, 619–631 (2008).

50. Cannan, W. J. & Pederson, D. S. Mechanisms and Consequences of Double-strand DNA Break Formation in Chromatin. *J. Cell. Physiol.* **231**, 3–14 (2016).

51. Kuzminov, A. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 8241–8246 (2001).

52. Noll, D. M., Mason, T. M. & Miller, P. S. Formation and Repair of Interstrand Cross-Links in DNA. *Chem. Rev.* **106**, 277–301 (2006).

53. Vare, D. *et al.* DNA interstrand crosslinks induce a potent replication block followed by formation and repair of double strand breaks in intact mammalian cells. *DNA Repair* **11**, 976–985 (2012).

54. Sancar, A., Lindsey-Boltz, L. A., Ünsal-Kaçmaz, K. & Linn, S. Molecular Mechanisms of Mammalian DNA Repair and the DNA Damage Checkpoints. *Annu. Rev. Biochem.* **73**, 39–85 (2004).

55. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res.* **18**, 85–98 (2008).

56. Wang, H. & Hays, J. B. Human DNA mismatch repair: coupling of mismatch recognition to strand-specific excision. *Nucleic Acids Res.* **35**, 6727–6739 (2007).

57. Lynch, H. *et al.* Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin. Genet.* **76**, 1–18 (2009).

58. Koi, M. *et al.* Human chromosome 3 corrects mismatch repair deficiency and microsatellite instability and reduces N-methyl-N'-nitro-N-nitrosoguanidine tolerance in colon tumor cells with homozygous hMLH1 mutation. *Cancer Res.* **54**, 4308–4312 (1994).

59. Drummond, J. T., Anthoney, A., Brown, R. & Modrich, P. Cisplatin and adriamycin resistance are associated with MutLalpha and mismatch repair deficiency in an ovarian tumor cell line. *J. Biol. Chem.* **271**, 19645–19648 (1996).

60. Fujita, M. *et al.* Microsatellite instability and alterations in the hMSH2 gene in human ovarian cancer. *Int. J. Cancer* **64**, 361–366 (1995).

61. Esteller, M., Levine, R., Baylin, S. B., Ellenson, L. H. & Herman, J. G. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. *Oncogene* **17**, 2413–2417 (1998).

62. Aebi, S. *et al.* Loss of DNA mismatch repair in acquired resistance to cisplatin. *Cancer Res.* **56**, 3087–3090 (1996).

63. Dung, L., Uram, J., Wang, H. & Bartlett, B. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

64. Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.* **15**, 465–481 (2014).

65. Lehmann, A. R., McGibbon, D. & Stefanini, M. Xeroderma pigmentosum. *Orphanet J. Rare Dis.* **6**, 70 (2011).

66. Nance, M. A. & Berry, S. A. Cockayne syndrome: review of 140 cases. *Am. J. Med. Genet.* **42**, 68–84 (1992).

67. Krokan, H. & Bjoras, M. Base Excision Repair. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

68. Tebbs, R. S. *et al.* Requirement for the Xrcc1 DNA base excision repair gene during early mouse development. *Dev. Biol.* **208**, 513–529 (1999).

69. Farrington, S. M. *et al.* Germline susceptibility to colorectal cancer due to base-excision repair gene defects. *Am. J. Hum. Genet.* **77**, 112–119 (2005).

70. Povirk, L. F. DNA damage and mutagenesis by radiomimetic DNA-cleaving agents: bleomycin, neocarzinostatin and other enediynes. *Mutat. Res. Mol. Mech. Mutagen.* **355**, 71–89 (1996).

71. Liang, F., Han, M., Romanienko, P. J. & Jasin, M. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc. Natl. Acad. Sci.* **95**, 5172–5177 (1998).

72. Jasin, M. & Rothstein, R. Repair of Strand Breaks by Homologous Recombination. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

73. Croteau, D. L., Popuri, V., Opresko, P. L. & Bohr, V. A. Human RecQ Helicases in DNA Repair, Recombination, and Replication. *Annu. Rev. Biochem.* **83**, 519–552 (2014).

74. Stark, J. M., Pierce, A. J., Oh, J., Pastink, A. & Jasin, M. Genetic Steps of Mammalian Homologous Repair with Distinct Mutagenic Consequences. *Mol. Cell. Biol.* **24**, 9305–9316 (2004).

75. Venkitaraman, A. R. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**, 171–182 (2002).

76. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA Repair* **7**, 1765–1771 (2008).

77. Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.* **18**, 495 (2017).

78. Tomimatsu, N. *et al.* Phosphorylation of EXO1 by CDKs 1 and 2 regulates DNA end resection and repair pathway choice. *Nat. Commun.* **5**, 3561 (2014).

79. Karanam, K., Kafri, R., Loewer, A. & Lahav, G. Quantitative live cell imaging reveals a gradual shift between DNA repair mechanisms and a maximal use of HR in mid S phase. *Mol. Cell* **47**, 320–329 (2012).

80. McVey, M. & Lee, S. E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet. TIG* **24**, 529–538 (2008).

81. O'Driscoll, M. *et al.* DNA ligase IV mutations identified in patients exhibiting developmental delay and immunodeficiency. *Mol. Cell* **8**, 1175–1185 (2001).

82. Ben-Omran, T. I., Cerosaletti, K., Concannon, P., Weitzman, S. & Nezarati, M. M. A patient with mutations in DNA Ligase IV: clinical features and overlap with Nijmegen breakage syndrome. *Am. J. Med. Genet. A.* **137A**, 283–287 (2005).

83. Lavin, M. F. Ataxia-telangiectasia: from a rare disorder to a paradigm for cell signalling and cancer. *Nat. Rev. Mol. Cell Biol.* **9**, 759 (2008).

84. Rothblum-Oviatt, C. *et al.* Ataxia telangiectasia: a review. *Orphanet J. Rare Dis.* **11**, 159 (2016).

85. Chrzanowska, K. H., Gregorek, H., Dembowska-Bagińska, B., Kalina, M. A. & Digweed, M. Nijmegen breakage syndrome (NBS). *Orphanet J. Rare Dis.* **7**, 13 (2012).

86. Moldovan, G.-L. & D'Andrea, A. D. How the Fanconi Anemia pathway guards the genome. *Annu. Rev. Genet.* **43**, 223–249 (2009).

87. Niedzwiedz, W. *et al.* The Fanconi anaemia gene FANCC promotes homologous recombination and error-prone DNA repair. *Mol. Cell* **15**, 607–620 (2004).

88. Tischkowitz, M. D. & Hodgson, S. V. Fanconi anaemia. *J. Med. Genet.* **40**, 1–10 (2003).

89. D'Andrea, A. & Grompe, M. The Fanconi anaemia/BRCA pathway. *Nat. Rev. Cancer* **3**, 23–24 (2003).

90. Ray, P. D., Huang, B.-W. & Tsuji, Y. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. *Cell. Signal.* **24**, 981–990 (2012).

91. Kishida, K. & Klann, E. Sources and Targets of Reactive Oxygen Species in Synaptic Plasticity and Memory. *Antioxid. Redox Signal.* **9**, 233–244 (2007).

92. Bergamini, C. M., Gambetti, S. & Cervellati, A. D. and C. Oxygen, Reactive Oxygen Species and Tissue Damage. *Current Pharmaceutical Design* http://www.eurekaselect.com/62681/article (2004).

93. Neeley, W. & Essigmann, J. Mechanisms of Formation, Genotoxicity, and Mutation of Guanine Oxidation Products. *Chem. Res. Toxicol.* **19**, 491–505 (2006).

94. Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **17**, 1195–1214 (2003).

95. Grollman, A. P. & Moriya, M. Mutagenesis by 8-oxoguanine: an enemy within. *Trends Genet.* **9**, 246–249 (1993).

96. Delaney, S., Jarem, D. A., Volle, C. B. & Yennie, C. J. Chemical and Biological Consequences of Oxidatively Damaged Guanine in DNA. *Free Radic. Res.* **46**, 420–441 (2012).

97. Ranjan Jena, N. & Chand Mishra, P. Is FapyG Mutagenic?: Evidence from the DFT Study. *ChemPhysChem* **14**, 3263–3270 (2013).

98. Rydberg, B. & Lindahl, T. Nonenzymatic methylation of DNA by the intracellular methyl group donor S-adenosyl-L-methionine is a potentially mutagenic reaction. *EMBO J.* **1**, 211–216 (1982).

99. Holliday, R. & Ho, T. Evidence for gene silencing by endogenous DNA methylation. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8727–8732 (1998).

100. Stern, L. L., Mason, J. B., Selhub, J. & Choi, S. W. Genomic DNA hypomethylation, a characteristic of most cancers, is present in peripheral leukocytes of individuals who are homozygous for the C677T polymorphism in the methylenetetrahydrofolate reductase gene. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **9**, 849–853 (2000).

101. Tudek, B., Bioteux, S. & Laval, J. Biological properties of imidazole ring-opened N7-methylguanine in M13mp18 phage DNA. *Nucleic Acids Res.* **20**, 3079–3084 (1992).

102. Esterbauer, H. & Cheeseman, K. H. Determination of aldehydic lipid peroxidation products: malonaldehyde and 4-hydroxynonenal. *Methods Enzymol.* **186**, 407–421 (1990).

103. Niedernhofer, L. J., Daniels, J. S., Rouzer, C. A., Greene, R. E. & Marnett, L. J. Malondialdehyde, a product of lipid peroxidation, is mutagenic in human cells. *J. Biol. Chem.* **278**, 31426–31433 (2003).

104. Maddukuri, L. *et al.* In Vitro Bypass of the Major Malondialdehyde- and Base Propenal-Derived DNA Adduct by Human Y-family DNA Polymerases κ, ι, and Rev1. *Biochemistry* **49**, 8415–8424 (2010).

105. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).

106. Karran, P., Lindahl, T. & Griffin, B. Adaptive response to alkylating agents involves alteration in situ of O 6 -methylguanine residues in DNA. *Nature* **280**, 76–77 (1979).

107. Pray, L. DNA Replication and Causes of Mutation. *Nat. Educ.* **1**, 214 (2008).

108. Rayner, E. *et al.* A panoply of errors: polymerase proofreading domain mutations in cancer. *Nat. Rev. Cancer* **16**, 71 (2016).

109. Wimmer, K. *et al.* Diagnostic criteria for constitutional mismatch repair deficiency syndrome: suggestions of the European consortium ‘care for CMMRD’ (C4CMMRD). *J. Med. Genet.* **51**, 355–365 (2014).

110. Shlien, A. *et al.* Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermutated cancers. *Nat. Genet.* **47**, 257 (2015).

111. Bonura, T. & Smith, K. C. Enzymatic production of deoxyribonucleic acid double-strand breaks after ultraviolet irradiation of Escherichia coli K-12. *J. Bacteriol.* **121**, 511–517 (1975).

112. Thoms, B. & Wackernagel, W. Interaction of RecBCD Enzyme with DNA at Double-Strand Breaks Produced in UV-Irradiated Escherichia coli: Requirement for DNA End Processing. *J. Bacteriol.* **180**, 5639–5645 (1998).

113. Yoon, J.-H. *et al.* Error-Prone Replication through UV Lesions by DNA Polymerase θ Protects against Skin Cancers. *Cell* **176**, 1295-1309.e15 (2019).

114. Barak, Y., Cohen-Fix, O. & Livneh, Z. Deamination of Cytosine-containing Pyrimidine Photodimers in UV-irradiated DNA SIGNIFICANCE FOR UV LIGHT MUTAGENESIS. *J. Biol. Chem.* **270**, 24174–24179 (1995).

115. Reisz, J. A., Bansal, N., Qian, J., Zhao, W. & Furdui, C. M. Effects of Ionizing Radiation on Biological Molecules—Mechanisms of Damage and Emerging Methods of Detection. *Antioxid. Redox Signal.* **21**, 260–292 (2014).

116. Baskar, R., Lee, K. A., Yeo, R. & Yeoh, K.-W. Cancer and Radiation Therapy: Current Advances and Future Directions. *Int. J. Med. Sci.* **9**, 193–199 (2012).

117. Tancell, P. J., Rhead, M. M., Trier, C. J., Bell, M. A. & Fussey, D. E. The sources of benzo[a]pyrene in diesel exhaust emissions. *Sci. Total Environ.* **162**, 179–186 (1995).

118. Aygün, S. F. & Kabadayi, F. Determination of benzo[a]pyrene in charcoal grilled meat samples by HPLC with fluorescence detection. *Int. J. Food Sci. Nutr.* **56**, 581–585 (2005).

119. Hecht, S. Tobacco Smoke Carcinogens and Lung Cancer. *J. Natl. Cancer Inst.* **19**, 1194–1210 (1999).

120. Xie, Z. *et al.* Mutagenesis of benzo[a]pyrene diol epoxide in yeast: requirement for DNA polymerase zeta and involvement of DNA polymerase eta. *Biochemistry* **42**, 11253–11262 (2003).

121. Dasari, S. & Tchounwou, P. B. Cisplatin in cancer therapy: molecular mechanisms of action. *Eur. J. Pharmacol.* **0**, 364–378 (2014).

122. Wang, D. & Lippard, S. J. Cellular processing of platinum anticancer drugs. *Nat. Rev. Drug Discov.* **4**, 307–320 (2005).

123. Hall, M. D., Okabe, M., Shen, D.-W., Liang, X.-J. & Gottesman, M. M. The role of cellular accumulation in determining sensitivity to platinum-based chemotherapy. *Annu. Rev. Pharmacol. Toxicol.* **48**, 495–535 (2008).

124. Jamieson, E. R. & Lippard, S. J. Structure, Recognition, and Processing of Cisplatin-DNA Adducts. *Chem. Rev.* **99**, 2467–2498 (1999).

125. Boot, A., Ni Huang, M., Ng, A. & Ho, S.-C. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.* **28**, 654–665 (2018).

126. Kaufman, E. R. Replication of DNA containing 5-bromouracil can be mutagenic in Syrian hamster cells. *Mol. Cell. Biol.* **4**, 2449–2454 (1984).

127. Prados, M. D. *et al.* Phase III randomized study of radiotherapy plus procarbazine, lomustine, and vincristine with or without BUdR for treatment of anaplastic astrocytoma: final report of RTOG 9404. *Int. J. Radiat. Oncol. Biol. Phys.* **58**, 1147–1152 (2004).

128. Longley, D. B., Harkin, D. P. & Johnston, P. G. 5-fluorouracil: mechanisms of action and clinical strategies. *Nat. Rev. Cancer* **3**, 330–338 (2003).

129. Daher, G. C., Harris, B. E. & Diasio, R. B. Metabolism of pyrimidine analogues and their nucleosides. *Pharmacol. Ther.* **48**, 189–222 (1990).

130. Lerman, L. S. Structural considerations in the interaction of DNA and acridines. *J. Mol. Biol.* **3**, 18-IN14 (1961).

131. Sharp, P., Sugden, B. & Sambrook, J. Detection of two restriction endonuclease activities in Haemophilus parainfluenzae using analytical agarose-ethidium bromide electrophoresis. *Biochemistry* **12**, 3055–3063 (1973).

132. Koster, D., Palle, K., Bot, E., Bjornsti, M.-A. & Dekker, N. Antitumour drugs impede DNA uncoiling by topoisomerase I. *Nature* **448**, 1476–4687 (2007).

133. Ferguson, L. & Denny, W. Genotoxicity of non-covalent interactions: DNA intercalators. *Mutat. Res. Mol. Mech. Mutagen.* **623**, 14–23 (2007).

134. Wheate, N. J., Brodie, C. R., Collins, J. G. & Aldrich-Wright, S. K. and J. R. DNA Intercalators in Cancer Therapy: Organic and Inorganic Drugs and Their Spectroscopic Tools of Analysis. *Mini-Reviews in Medicinal Chemistry* http://www.eurekaselect.com/78376/article (2007).

135. Pfeifer, G. P., You, Y.-H. & Besaratinia, A. Mutations induced by ultraviolet light. *Mutat. Res.* **571**, 19–31 (2005).

136. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational-signatures. *Nat. Med.* **23**, 517–525 (2017).

137. Lord, C. J. & Ashworth, A. PARP inhibitors: Synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017).

138. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

139. Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M. & Park, P. J. Impact of chromatin structure on sequence variability in the human genome. *Nat. Struct. Mol. Biol.* **18**, 510–515 (2011).

140. Chen, X. *et al.* Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* **335**, 1235–1238 (2012).

141. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).

142. 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).

143. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259 (2016).

144. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.* **9**, (2018).

145. Elliott, K. *et al.* Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers. *PLoS Genet.* **14**, e1007849 (2018).

146. Georgakopoulos-Soares, I., Morganella, S., Jain, N. & Hemberg, M. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* **28**, 2–8 (2018).

147. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).

148. Dickerson, R. E. DNA structure from A to Z. *Methods Enzymol.* **211**, 67–111 (1992).

149. Duan, C. *et al.* Reduced intrinsic DNA curvature leads to increased mutation rate. *Genome Biol.* **19**, 132 (2018).

150. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).

151. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

152. Hu, J., Adebali, O., Adar, S. & Sancar, A. Dynamic maps of UV damage formation and repair for the human genome. *Proc. Natl. Acad. Sci.* 201706522 (2017) doi:10.1073/pnas.1706522114.

153. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).

154. Foulkes, W. D. Inherited Susceptibility to Common Cancers. *N. Engl. J. Med.* **359**, 2143–2153 (2008).

155. Li, S. K. H. & Martin, A. Mismatch Repair and Colon Cancer: Mechanisms and Therapies Explored. *Trends Mol. Med.* **22**, 274–289 (2016).

156. Li, F. *et al.* The Histone Mark H3K36me3 Regulates Human DNA Mismatch Repair through its Interaction with MutSα. *Cell* **153**, 590–600 (2013).

157. Lanzós, A. *et al.* Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Sci. Rep.* **7**, 1–16 (2017).

158. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108 (2006).

159. Hoffman, M. M. & Birney, E. Estimating the neutral rate of nucleotide substitution using introns. *Mol. Biol. Evol.* **24**, 522–531 (2007).

160. Yuan, S. *et al.* Detailed mechanism for photoinduced cytosine dimerization: a semiclassical dynamics simulation. *J. Phys. Chem. A* **115**, 13291–13297 (2011).

161. Schreier, W. J., Gilch, P. & Zinth, W. Early events of DNA photodamage. *Annu. Rev. Phys. Chem.* **66**, 497–519 (2015).

162. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

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

164. Ikehata, H. & Ono, T. The Mechanisms of UV Mutagenesis. *J. Radiat. Res. (Tokyo)* **52**, 115–125 (2011).

165. Tomkova, M. & Schuster-Böckler, B. DNA Modifications: Naturally More Error Prone? *Trends Genet.* **34**, 627–638 (2018).

166. Mitchell, D. L. & Nairn, R. S. The Biology of the (6–4) Photoproduct. *Photochem. Photobiol.* **49**, 805–819 (1989).

167. Yokoyama, H. & Mizutani, R. Structural Biology of DNA (6-4) Photoproducts Formed by Ultraviolet Radiation and Interactions with Their Binding Proteins. *Int. J. Mol. Sci.* **15**, 20321–20338 (2014).

168. Mitchell, D. L., Haipek, C. A. & Clarkson, J. M. (6–4)Photoproducts are removed from the DNA of UV-irradiated mammalian cells more efficiently than cyclobutane pyrimidine dimers. *Mutat. Res. Lett.* **143**, 109–112 (1985).