UNIVERSITAT DE BARCELONA

Localized hypermutation and hypomutation in the genomes of human somatic cells

David Mas-Ponte

Barcelona, 2022

Localized hypermutation and hypomutation in the genomes of human somatic cells

Director
Fran Supek

Autor
David Mas-Ponte

Tutor
Modesto Orozco

INSTITUT DE RECERCA BIOMÈDICA (IRB)

UNIVERSITAT DE BARCELONA
Facultat de Química
Programa de Doctorat en Biomedicina

Als que ja no hi són,

The wind blew southward, through knotted forests,
over shimmering plains and towards lands unexplored.

This wind, it was not the ending. There are no endings,
and never will be endings, to the turning of the Wheel of Time.

But it was an ending.

Brandon Sanderson and Robert Jordan,
A Memory of Light,
The Wheel of Time Series

Abstract

Somatic cells accumulate mutations in their genome resulting from a set of exogenous and endogenous processes. The interplay of DNA lesions and the DNA repair mechanisms in each cell shape the genetic mosaicism that composes an adult tissue. Although many of these alterations have a neutral effect, some can eventually impede the correct physiological function of the tissue, causing cancer, and other diseases such as clonal hematopoiesis and repeat expansion disorders. Understanding the molecular mechanisms of how these somatic mutations are generated can thus help in the prevention and treatment of such diseases, and can help understand DNA replication and repair mechanisms operative in human cells.

In this thesis, we explore somatic mutation distributions from several perspectives, focusing on the genomic features that modulate the local rate at which mutations accumulate. First, we systematically study the mechanisms that generate APOBEC mutations in tumor samples; we describe a new mechanism of diffuse mutation clusters that are enriched in gene-rich domains of the human genome, consistent with a DNA repair-mediated mutagenesis. Next, we study various somatic mutation signatures across a wide range of human healthy tissues and compare them with their corresponding cancer types, reporting broad similarities. We also study the sub-gene resolution heterogeneity in mutation rates, revealing a gradient of mutation rate along the gene body and its interaction with other functional elements like promoters, enhancers, and loop anchors. Lastly, we detect and characterize distal mutation clusters in trans-interacting chromatin loci, which suggests a three-dimensional-acting mutagenesis mechanisms in human cells.

Overall, studies in this thesis highlight the variable activity of the endogenous sources of DNA mutations along the loci in the human genome, elucidate mechanisms and impact on accruing mutations in functional elements.

Resum

Les cèl·lules somàtiques acumulen mutacions en el seu genoma a partir d'un conjunt de processos exògens i endògens. La interacció de les lesions d'ADN i els mecanismes de reparació de l'ADN de cada cèl·lula configuren el mosaicisme genètic que compon un teixit adult. Tot i que moltes d'aquestes alteracions tenen un efecte neutre, algunes poden eventualment impedir la correcta funció fisiològica del teixit, provocant càncer i altres malalties com l'hematopoiesi clonal i les malalties d'expansió de seqüències repetida. Comprendre els mecanismes moleculars de com es generen aquestes mutacions somàtiques pot ajudar a la prevenció i el tractament d'aquestes malalties, i pot ajudar a comprendre la replicació i reparació de l'ADN a les cèl·lules humanes.

En aquesta tesi, explorem les distribucions de mutacions somàtiques des de diverses perspectives, centrant-nos en les característiques genòmiques que modulen la taxa local a la qual s'acumulen. En primer lloc, estudiem sistemàticament els mecanismes que generen mutacions derivades de l'activitat dels enzims APOBEC en mostres tumorals; descrivim un nou mecanisme de cúmuls de mutació difusa que s'enriqueixen en dominis genòmics rics en gens, aquest sistema es compatible amb una mutagènesi mediada per la reparació de l'ADN. A continuació, estudiem els patrons de mutació somàtica en una àmplia gamma de teixits humans sans i les comparem amb els seus tipus de càncer corresponents, detectant grans similituds. També estudiem l'heterogeneïtat de la resolució de subgens en les taxes de mutació, revelant un gradient a la taxa de mutació al llarg del cos del gen i la seva interacció amb altres elements funcionals com promotors, potenciadors i llaços d'ancoratge de la cromatina. Finalment, detectem i caracteritzem cúmuls de mutacions distals en loci de cromatina que interaccionen trans, cosa que suggereix mecanismes de mutagènesi d'acció tridimensional actius a les cèl·lules humanes.

En conjunt, els estudis d'aquesta tesi posen de manifest l'acumulació variable de les fonts endògenes de mutacions de l'ADN al llarg del genoma humà, dilucida els mecanismes de com s'originen i remarca l'impacte en l'acumulació de mutacions en els elements funcionals.

List of Figures

Contents

Chapter 1

Introduction

Coined by Hugo de Vries in 1903[1] , the term mutation implies a sudden change in the inherited material of a species. He first used the term to explain the unexpected new varieties that arise in his experimental gardens of evening primrose flowers (*Oenothera lamarckiana*). However, this experiment was later attributed to a recombination event in a balanced chromosome, which does not align with our current definition of a mutation. The first characterization of a DNA mutation, as we understand it today, was introduced 27 years later by Hermann Muller, an American scientist who increased the mutation rate in fruit flies by X-rays irradiation[2] . His experiments were crucial to prove that a chemical molecule encoded the inherited information of the cell.

1.1 Somatic mutagenesis

For humans and other higher eukaryotes and according to the Weisman's germ plasm theory[3] , we divide the mutations in the human genome in somatic and germline. Somatic mutations are defined as DNA variants that occur outside the germ cell lineage and thus are not inherited by the next generation. Contrary to these, germline variants occur prior to the zygote formation, are present in all cells of the organism and can be transmitted to the offspring.

At the somatic level, mutations are generated both from endogenous sources, such as DNA replication errors or cytosine deamination and exogenous sources, such as UV radiation or harmful chemicals. These lesions accumulate in our tissues through the lifespan of the organism. After the damage arises, DNA repair proteins act to revert to the ancestral state before replication occurs, but if these lesions are not corrected, mutations get fixed in further daughter cells. By definition, somatic mutations occur after the zygote formation, thus limiting the cell lineage that will inherit them. In practical terms, to detect mutations in a somatic tissue a clonal expansion of the mutation harboring cell is usually required so enough DNA material can be extracted and analyzed [a]. Thus, within this methodological framework, only mutations which are neutral or have a positive effect on the clonal growth, e.g. in cancer, will be detected by sequencing 1.1.

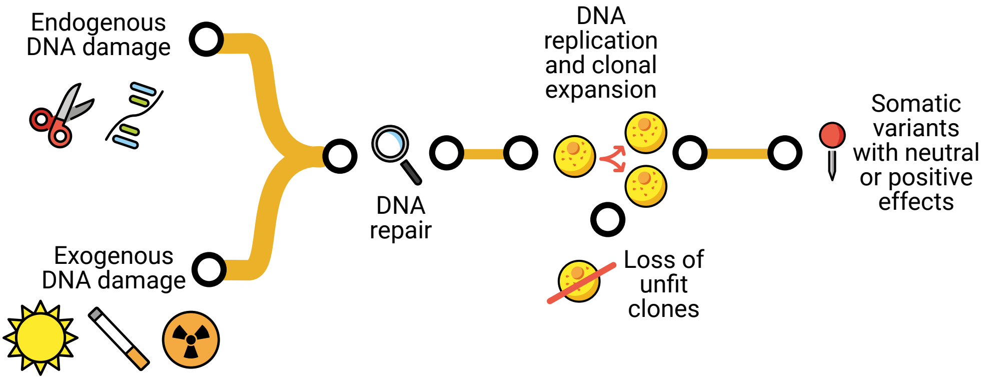

Endogenous DNA damage
Exogenous DNA damage
DNA repair
DNA replication and clonal expansion
Loss of unfit clones
Somatic variants with neutral or positive effects

Figure 1.1: Schematic showing how mutations accumulate in the genome of somatic cells. Either exogenous or endogenous mutagens generate lesions, which are later repaired by DNA repair. If the cell suffers from a clonal expansion (i.e. cancer) these mutations get amplified to the level that can be detected by sequencing.

1.1.1 Pre-genomic studies

The role of mutagenesis in the theory of evolution has always been of interest in the field of genetics and molecular biology[4] . In the pre-genomic era, mutation accumulation experiments were used to uncover and characterize the mutagenic processes of model organisms such as yeast, fruit fly and *C. elegans*, yielding an important understanding of how mutations occur[5,6] .

[a] We do note the recent developments that allow detection of somatic mutations with no clonal expansion, see section 1.1.3.1 for more detail.

1.1.1.1 Early definition of somatic mutagenesis

A substantial advancement in the study of mutagenesis was made after the discovery that human cancer was caused directly by DNA and its containing mutations[7,8] . After that discovery, the sequencing of cancer extracts started to lead to the identification of the first cancer causing somatic mutation[9,10] . Interestingly, the sequencing of cancer genes such as *HRAS* or *TP53* yielded a large accumulation of DNA sequences and the characterization of the molecular mechanisms of mutagenesis[11] . One of this processes was the spontaneous deamination of CpG islands which has been recently characterized and detected in almost every human tissue both for healthy and tumor samples[12,13] (see section 1.2) .

Later studies in DNA repair were based in simple model systems, like bacteria or yeast, where scientists reconstructed the main DNA repair pathways using recombinant strains[14,15] to study epistasis between constituent genes. From these experiments, three main DNA repair pathways emerged as the main guardians of the genome from point mutations: base excision repair (BER)[16] , mismatch repair (MMR)[17] and nucleotide excision repair (NER)[18] . Additional pathways mend DNA breaks and so protect against rearrangements.

The publication of the human genome sequence and the development of modern sequencing technologies allowed the field of cancer genomics to shift from a more targeted approached to now being able to sequence many genes for a limited number of patients[11] . These revelations, lead to the coordination and set up of a number of international consortia that lead to the sequencing and analysis of tumors form large sets of patients[11,19–23] . In parallel, significant efforts have also focused in the sequencing of germline variation, both in the population level and from trios, leading to the study of mutagenesis and selection in germline, which has differences compared with somatic mutagenesis[24–28] .

Only 13 years ago[11] the number of available somatic mutations was in the range of hundreds of thousands, while today, the multiple sequencing projects have yielded hundreds of millions of mutations, increasing the potential for novel discoveries in both cancer evolution, and the biology of mutagenesis in human.

1.1.2 Cancer as a model organism

Cancer is a somatic disease where cells grow and reproduce uncontrollably, outside the normal homeostasis of a given tissue. Cancer, however, can also be studied as an *in vivo* mutation accumulation experiment.

1.1.2.1 Cancer as a human disease

Cancer is the second most common disease in humans, with 18 million[new cases and 10 million deaths per year world wide (IARC[29]); 2 million new cases and 600 thousands deaths per year in the US (NIH-SEER[30]) and 280 thousand new cases in Spain (REDCAN[31]). At these rates, at least 40% of the population will be diagnosed with cancer during their life[30] .

Nowadays, cancer is treated with a specific set of treatments depending on the type of cancer, the tissue where it originates and the stage at which it is detected. Traditional therapies such as surgery, radiotherapy, and chemotherapy now coexist with a variety of novel, biological techniques. Examples of such are immune checkpoint blockade therapy, engineered cell therapy (CAR-T) and hormone therapy. Together with all of these techniques, the DNA and, sometimes, RNA sequencing of tumors is helping in prioritizing a given therapy to the molecular conditions of the tumor[32] (see section 1.2).

Together with these new therapies, much effort is directed at the early detection of the cancer, which has been demonstrated as a key predictor for better prognosis. An important genomic advance in this field is the development of the *liquid biopsy,* a technique that consists in the sequencing and characterization of blood circulating tumor cells, or tumoral DNA fragments. The genomic material that is leaked from the cancer cells to the bloodstream can be extracted and analyzed to obtain molecular information about multiple molecular and genomic features[33–37] . Newer reports are now starting to recapitulate mutational signatures from this data[38] . These data cannot only reveal, with surprising accuracy, the presence of a tumor but also lead to the prediction of the primary site[39] .

1.1.2.2 Evolutionary conservation of somatic mutagenesis

Cancer is not exclusive to humans, *neoplasia* has been identified in a wide variety of metazoans[40–42] and only some selected species such as the naked mole rat seems to be to some extent protected[40] from it. A recent report studying up to 191 of mammal species from zoos has highlighted the high prevalence of the disease, mirroring in some cases the ones in humans[43] .

This striking conservation of this disease across evolution also raises challenges in our understanding of the evolutionary dynamics of somatic tissues. Early models, for example, suggested that the higher body size of larger mammals would put them at a higher risk of cancer. However, early reports suggested that neither the developmental status or the size of the organism was relevant to the development of cancer[44,45] . This observation, known as Peto's paradox, has motivated research in the novel protective mechanisms that might exist in these species. A recent report sequenced the somatic tissues [b] of a total of 16 mammal species to quantify their somatic mutation rate[46] (see section 1.1.3.1) which manifested a overall highly conserved mutation processes but in a wide range of rates. Interestingly, mutation rate inversely correlated with lifespan but not size[c] hinting at a possible evolutionary mechanism to control cancer progression.

1.1.2.3 Role of somatic mutagenesis in cancer

The causal role of somatic mutations in carcinogenesis is widely studied due to two main reasons. First, strong evidence accumulated in pre-genomic reports (see section 1.1.1) about the correlation and likely causality of somatic mutations with the disease. Secondly, the cellular characteristics of the disease, mainly a clonal expansion from a single cell, represents a natural *in vivo* experiment for the efficient detection and study of mutations accumulated in the somatic tissues (figure 1.2).

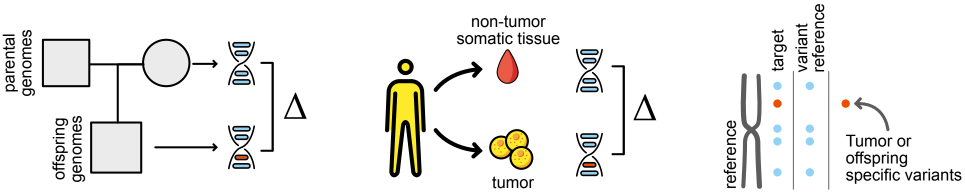


Figure 1.2: Schematic of identification of de novo mutations (left) and somatic mutation calling (center). For both techniques the reference variants (parents or blood) are compared against the target variants (offspring or tumor).

The first early studies of cancer genomes lead to the identification of specific coding somatic mutations that significantly re-occurred in multiple independent samples[19,47,48] . The detectable positive selection of these particular sites suggested their active role in the tumor progression and were termed consequentially, *driver mutations* . Although these mutations contain important information to understand the biology of a particular cancer, quantitatively, they represent a minority of all the somatic mutations that can be identified in a tumor. Recent studies quantifying the amount of positive selection in tumors report that the average sample will contain between 2-10 driver mutations[49] (see figure 1.3). The rest of the mutations termed by analogy passenger mutations are thought to carry either a neu-

[b] intestinal crypts extracted from microdissections

[c] Although the mutation rate also inversely correlated with adult mass (or size), when controlling for the correlation with lifespan, the variability in size was not informative.

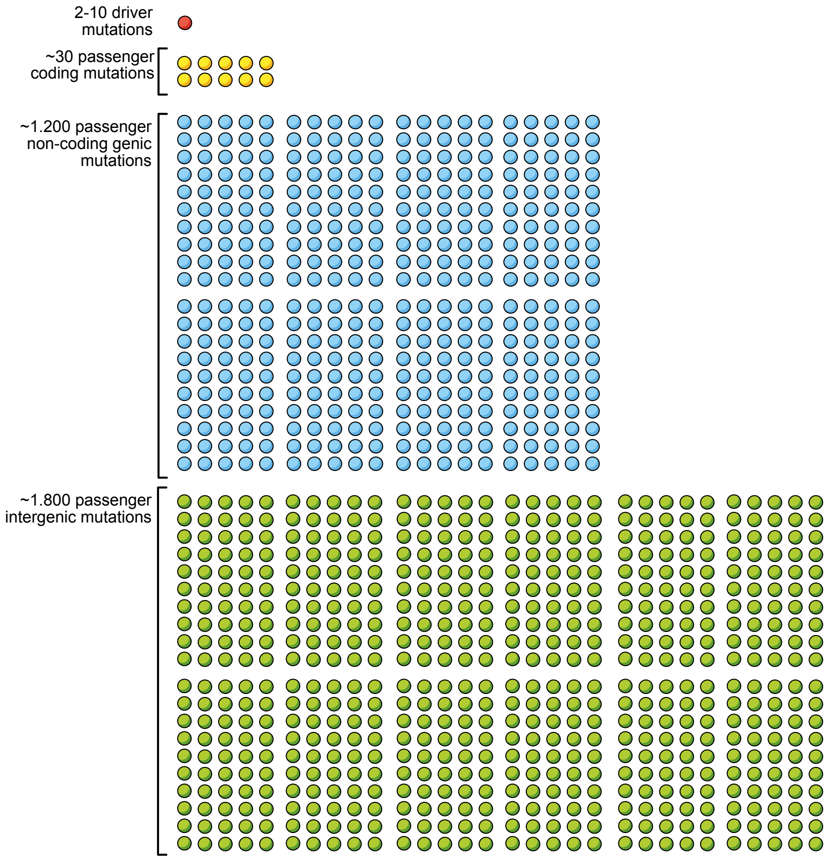

2-10 driver mutations
~30 passenger coding mutations
~1.200 passenger non-coding genic mutations
~1.800 passenger intergenic mutations

Figure 1.3: Schematic of the volume of passenger and driver mutations in an average cancer sample.

tral[49] or small deleterious[50,51] functional effect to the tumor fitness (see figure 1.3).

More interestingly, these passenger mutations are highly abundant in most cancer types, with the exception of pediatric and some blood cancers[52] , and because they are unaffected by selection, they accurately reflect the molecular characteristics of the mutagenic process that had caused them (see section 1.2) .

Overall, the molecular and cellular characteristics of tumors make them a great resource for the study of somatic mutagenic agents directly in humans bypassing the use of other common model organisms and in vitro cell lines.

1.1.3 Somatic mutagenesis in healthy tissues

Somatic mutagenesis is however not specific to cancer tissues. Normal cells [d] get mutations from most of the mutagens that can also be identified in cancer tissues, especially the ones associated with common exposures such as UV light in skin. Recent reports have shown the prevalence of somatic mutations of different types in multiple non-disease tissues. Brain is one of the particular organs with more reports where CNVs[53,54] , structural variants[55,56] and point mutations[57,58] have been identified and characterized. Other tissues like Skin[59,60] , Esophagus[61–63] , muscle cells[64] , kidney[65] , and colon crypts[46,66] are also other examples of the wide range of human tissues where these phenomena have been described.

Mutations in cancer genes [e] such as *NOTCH1* and *TP53*[59,61,63] were also detected with striking frequencies in healthy tissues. This finding represents a challenge of the general hypothesis about how a small set of selected mutations could be sufficient to cause the tumor growth[67] . These studies have reported a surprising accumulation of genetically diverse clonal subpopulations of cells which are detectable by directly sequencing a tissue sample[61,62] . A proposed physiological benefit for the existence of these healthy clones is the control of other carcinogenic clones arising in the same tissue[68] .

1.1.3.1 Detecting mutations in non-cancerous tissues

Contrary to germline variants, somatic mutagenesis in the adult tissues of complex organisms only occurs during and after the development of the tissues. This fact leads to a characteristic genetic mosaicism of mutations in every cell lineage[69] . Without the natural clonal expansion occurring in tumors (see section 1.1.2.3), the low allelic frequency of most somatic mutations makes the detection of genetic variants extremely complex compared to their germ line counterparts. Theoretically, the private DNA sequence of a single cell needs to be amplified in order to obtain sensible readings and sufficient coverage to identify a somatic variant[70] .

In recent years, a set of techniques and methodologies based on complementary cellular and molecular techniques have been developed to overcome this limitation. A first approach, Duplex sequencing, relies on the *in vitro* capture of fragmented somatic DNA and its amplification with traditional PCR machinery. The use of a randomized tag in each of the amplification primers can be used to detect if the sequenced mutations come from the original sample, thus being present in both amplified strands, or are artifacts of the PCR amplification, which are present in only one strand[71,72] . Various improvements to this technique have also been

[d] here defined as non-carcinogenic

[e] defined as genes harboring detected positive selection in cancer tissues

proposed and used to detect somatic mutation rates in humans and other organisms[73,74] . A second approach to detect somatic mutagenesis is to generate small microdissections from the tissue to obtain a small set of cells where relative allele frequencies can be sufficiently high and mutations can be reliably identified. These microdissections are generally performed with specialized equipment and power to call mutations is limited by the sequencing depth (i.e. cost) of the experiment. This technique has been used extensively in the recent years particularly in tissues with a natural clonal expansion such as colon crypts[46,66,75,76] but also in a wider range of tissues[59,61,63,77,78] . Finally, a more conventional, widely used approach is to sequence an *ex vivo* cell line culture that is derived from a primary tissue. Cells from the tissue are grown in a dish, isolated in single clones and expanded sufficiently to obtain enough DNA material. Then, each expanded clone culture is sequenced, and germline variants removed by comparing across clones or by comparing with the blood of the patient to extract somatic mutations, similarly as it done for tumors (see figure 1.2)[57,64,65,79] .

Each methodology has its own caveats and advantages. While the *in vitro* amplification of duplex sequencing is able to capture subclonal mutations at a extreme high resolution[f], the amount of genomic DNA that can be covered is generally small, usually not surpassing 1 Mb in size[72] . Some of the variations of the modified techniques also seemed to introduce some false positives around the fragment ends from the digestion of the fragmented DNA. More recent variations of this technology, named Nanoseq, seem to have solved this issue[74] . Microdissections cannot reach such levels of detection of low allele frequency mutations, but represent an improvement because of its spatial information and the ability to precisely analyze a relevant section of a given tissue of interest, i.e. the colon crypts. A caveat of this process is the costly equipment that is required, however each microdissection can be relatively labor- and cost-effective. Thus, these experiments normally contain a large number of clones. A caveat of this technology is that the clonality of the mutations is normally lower than what it can be detected, thus, a deep sequencing process needs to be used with coverages higher than 500X[80] . This sequencing need means that some of the early experiments which do not rely on natural expanding clones needed to focused on sequencing a limited set of genes[59] . Finally, the *ex vivo* approach also represents a simple, yet laborious and time consuming solution in order to accurately amplify the genetic material prior to sequencing. The limitations however come from the biology of the tissue that is lost to a certain degree when cultured in a dish, where only the stem cells subpopulations would get expanded more easily than differentiated cells introducing a biological bias. Other mutagenic processes related with the oxygen in the cell culture conditions can also contribute a significant amount of mutations

[f] a mutation present only once in a sample of 10^7 cells

to the sample[81] .

Overall, although mutation detection in healthy tissues remains a challenge, significant technological developments will result in a deeper understanding of the role of mutagenesis in other non-cancerous processes like aging and neurodegenerative disease.

1.1.3.2 Role of somatic mutations in aging and disease

The process of aging can be defined as the changes at the molecular and physiological level that our tissues suffer over time. The role of somatic mutations in the aging process was initially proposed[82,83] in the 60s during the initial developments in the nascent field of molecular genetics. Its proponents, mostly prominent physicists who had begun to study of DNA, were focused on the effects of radiation such as X-rays on the DNA molecule.

However, despite its early start, little is still known on the functional effect of somatic mutagenesis in determining the pace of aging. Most of the relevant advances in the field have focused on the role of cellular senescence[84] and in the study of epigenetic and chromosomal alterations[69,85,86] . Evidence from inherited accelerated aging syndromes like progeria or Cockayne syndrome are significantly enriched in DNA repair deficient genotypes[79,87] . It is still unclear though whether the actual mutations that accumulate in these samples are the causal effectors of aging or if another effect of these deficiencies, such as the induced apoptosis by DNA damage, might be more directly involved in aging. Other germline deficiencies in DNA repair genes, for instance genes in both the MMR and the BER pathway, do not cause premature aging even with a large amount of accumulated point mutations[75,76,88] (somatic rearrangements have been less studied in this property) . Of note, however, these inherited variants in MMR and BER do increase cancer rate significantly (reviewed in[89]), and increased cancer risk might be understood as a facet of aging. Overall, the lack of a good experimental model to measure normal aging limits our ability to determine the role of different molecular factors that may act as modulators of this process[79] .

Mutations however do accumulate with increased age. Data from the first cancer genomic studies revealed some mutagenic processes, signature 1 and 5 (see section 1.2), which strongly correlated with the age of the patients[90,91] suggestinga biological association. Interestingly, the study of the non-cancer tissues (mentioned above[57,64,66,77]) have also highlighted the pervasive nature of these same mutagenic processes and their association with the age of the individuals[79] even in healthy tissues. Due to this reported correlation with age[79,91] , the genomic instability of normal tissues is also considered one of the primary hallmarks of aging[92] .

Considering this evidence, the specific causal role of somatic mutagenesis in the aging process remains elusive and seems to be highly dependent on the definition of aging as a phenotype. Aging causes somatic mutations but it is less clear that somatic mutations cause aging.

Apart from cancer, genomic instability at the somatic level has also been proposed as a causal of other human diseases like Alzheimer disease or Parkinson. The fact that these conditions normally appear with age and have been linked to specific protein coding mutations seem to indicate that the accumulation of somatic mutations might plausible play a role. The most promising evidence is related with brain neurodegeneration and Alzheimer disease (AD) where recent reports have suggested not only the steady accumulation of somatic mutations in neurons but an increased mutation rate and the existence of a disease specific mutation process for AD[57,58] . Further evidence is required to establish a causal link between these increased mutagenesis and the pathology.

In this thesis, I have systematically analyzed the mutational processes that accumulate in a diverse set of human healthy samples amplified by the *ex vivo* methodology, compared to mutational processes in tumors of same tissues, and studied their genomic characteristics (see chapter 4).

1.2 Molecular mechanisms of somatic mutagenesis

A key feature of the observational study of genome sequences of tumoral samples is the ability to obtain insights into the biology of the mutagenic process. An important methodological advance was the development of DNA trinucleotide "mutational signatures" which can isolate biological relevant mutagenic processes (hereafter *mutational signatures*) and quantify their activity across individuals ("exposures")[90] .

1.2.1 Mathematical representations of mutational process

A mutational process can be defined as the distribution of specific DNA lesions generated by a mutagen which is either acted upon with a specific DNA repair pathway(s) or fixed into the genome through DNA replication across the lesion (see 1.1 and 1.5)[93] .

Mutations can accumulate from a wide variety of sources, including both exogenous and endogenous sources, lesion driven causes or DNA replication errors.

Thus, in general terms, a mutational process can be defined as a fixed combination of biological factors that influence the chemical and thus genomic features of the generated mutations.

1.2.1.1 Feature extraction from genomic features

Two clear examples of mutagenic processes with clearly defined genomic features are the deamination of the methylated cytosine at CpG sites (currently represented by the signature 1 or SBS1, where "SBS" stands for single base substitution)[12,13] and the accumulation of pyrimidine dimers in UV exposed cells (currently represented by the set of signature 7 or SBS7)[94] both generating C>T changes but in different contexts . These early studies were already able to determine a significant DNA sequence predisposition of these agents to the mutation risk, indicating that the lesion occurence and/or the subsequent repair had a particular chemical predisposition towards a given oligonucleotide context.

After these early analyses, the first genome sequences of human cancers (see section 1.1.1) yielded a set of clearly non-randomly mutated sequences[95] . In particular, these first reports[90,95] focused on the determination of mutagenic SNV (or single base substitutions, SBS). These processes could be characterized by considering the 5' and 3' of the mutated base together with the alternative (mutant) somatic allele. Because the genomic strand can only be measured relative to a local biological feature (such as replication or transcription) and so is by default undefined, mutations were collapsed strand-symmetrically and assigned to the pyrimidine base. Thus, from the 16 possible (A, C, G and T at each side) combinations of each mutation class (C>T, C>A, C>G, T>A, T>C and T>G) a total of 96 features were extracted and tallied in a set of human tumors (see 1.5).

Further studies have extended this initial classification of SNVs either by extending the sequence motif two extra bases (from trinucleotides to pentanucleotides[96]) and by introducing external genomic features (i.e. the direction of transcription, DNA replication strand, or the clustered nature of the mutations)[96–99] .

Although most of the research in this field is focused on SNVs due to its abundance in somatic tissues and the relative ease of their detection, other classes of mutations represent interesting sources of mutations in the soma, with a higher functional impact in the coding sequences. Small insertion and deletions (indels) represent the second most studied class in this field. The optimal feature classification for indels is not as clear as for SNVs but the features that are usually used comprise information about the size of the indel (how many base pairs are deleted or inserted) and the sequence context where they occur (i.e. occurring in a homopolymer or sequence repeat, and presenting microhomology at borders)[100] .

Other mutation types that have been identified to contain non-random accumulation, and thus potentially driven by a biologically relevant mutational process, are (i) clustered mutagenesis[96,97,99,101,102] (see section 1.4); (ii) double-base substitutions (DBS)[96,103] ; structural variants (i.e. rearrangements and fusions)[104,105] and related copy-number alterations[106–110] . Like in the indels, the feature characterization of these other mutational classes is still less standardized and multiple techniques are being developed and applied to genomes at the time of writing.

1.2.1.2 Detection, extraction, and fitting of mutational signatures

After the feature extraction and tally of the mutations, a factorization step is applied to deconvolute the set of mutagenic features into independent factors representing independent mutational processes.

The first reports[90,95,111] used the non-negative matrix factorization (NMF). NMF is used in a wide range of scientific fields like Astronomy, Image analysis, and gene expression[112] and it was also applied to detect factors coming from the mutational data. The methodology consists of a bootstrapped resampling set, a factorization step, and finally a clustering of the resulting solutions to generate a set of robust NMF factors; the clustering quality score (typically, the "silhouette index") can be used to determine the optimal number of signatures. In some implementations, a separate fitting step is applied to determine the exposures of the resulting mutation signatures to each sample (see figure 1.5).

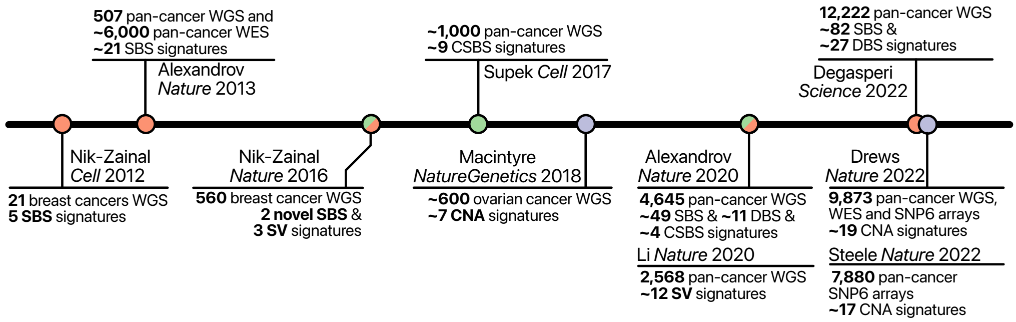


Figure 1.4: A timeline of the major advancements in mutational signatures, the number of samples used in each study and the number of signatures and type.

Simplified, the original mutational spectra (M), which contains the tallied mutation classes for each sample (genome) in a set of samples, is decomposed into 2 matrices that when multiplied, recover the original mutational spectra with an error component[111] . The decomposed matrices represent, for each mutagenic process (hereafter referred as signature) the feature (trinucleotide) weights or mutation spectrum (S) and the sample exposure or signature activity (E). The mutation profiles capture the sequence predisposition of a given mutational signature, helping in the identification of its source (see 1.1.1). The exposure matrix works as an estimate of the influence or weight of each mutagenic process in a particular sample, in a way, representing how much a particular sample has been 'exposed' to a particular mutagen (see figure 1.5). A multinomial resampling of mutation counts in every row in the original matrix is also performed. This incorporates a representation of the uncertainty present in the mutation spectra of samples with low mutation counts for which there is less numerical evidence for a specific profile.

The NMF algorithm used in the factorization step does require a priori knowledge of the number of factors (which will eventually represent the mutational signatures), but this information is normally not known for a generic somatic sequencing analysis. In order to infer this parameter, the (i) stability and (ii) the error minimization of a solution in multiple repeated factorizations is used to determine the optimal number of clusters (k), (see figure 1.5) . The minimization of the error component is normally performed by establishing a threshold (in the number of factors) where the error component (residual) is no longer notably reduced(manual inspection of an 'elbow' in the curve). The maximization of the stability is measured using the cosine similarity of the clusters obtained after the clustering of the multiple solutions obtained in each iteration. At higher silhouette index (lower distance between the clusters) the resulting solutions are identified recurrently and with similar profiles in multiple NMF runs, suggesting that they are robustly found in the input sample. Because each cluster of solutions represents a mutational signature the cosine similarity can be used in downstream analysis to measure the quality of each derived factor[111,113] .

Once the spectra of the mutation signatures are identified, the subsequent step consists in the estimation (or assignment) of exposures of these signatures for every sample. A common method used in the standard tools[113] consists in fitting, through a regression model, to each sample based on the extracted NMF profiles and the original mutational spectrum of the sample. This is normally performed using a Non-negative least squares (nnls) optimization.

Since the initial description of this methodology[90,111] multiple alternatives have been published modifying multiple individual steps of the process. In brief, approaches using Bayesian NMF[114] , Hidden Markov models[101] independent probabilistic modeling[115] , tensor tucker decomposition[99] topic models[116] or independent component analysis and unsuperviser neural networks (a variational autoencoder)[117] are some examples.

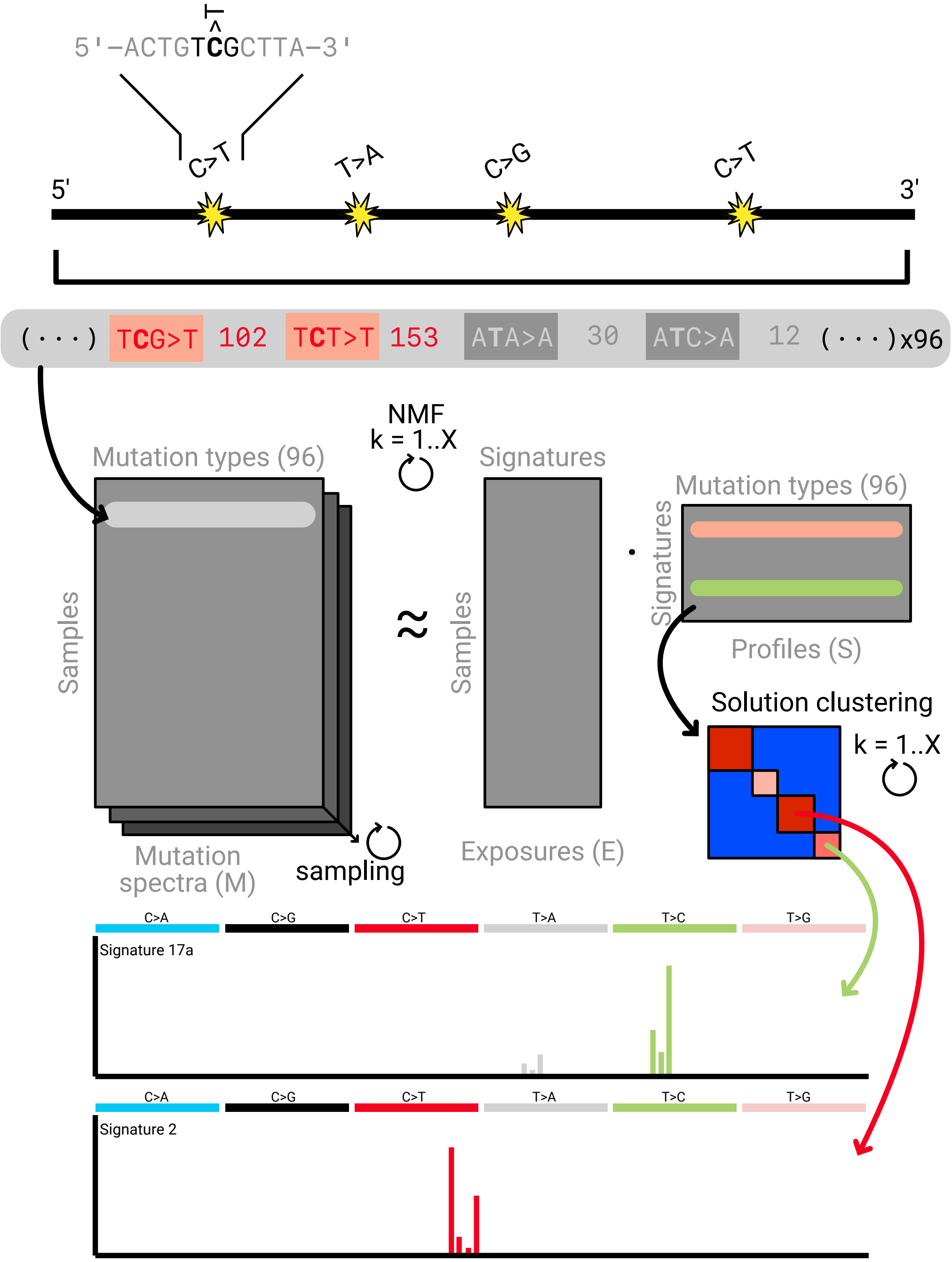


Figure 1.5: A diagram of the main steps in the mutation signature extraction.

1.2.1.3 Caveats of mutational signatures

Although nowadays mutational signature extraction is used pervasively in the field of cancer genomics and in the study of somatic non-cancerous mutagenesis, several variations and caveats have been identified and are worth mentioning to aid interpretation of these factors[93] .

The most common source of mutational signatures mis-quantification is called mutational bleeding. The similarity between some signature profiles (sometimes overlapping by many trinucleotide contexts) can generate inaccuracies in determining which signature better explains the observed profile of a sample[93,118,119] . This problem is particularly important in the hypermutated tumor samples, where a small error in the fitting can lead to a significant accumulation of the wrong mutational signature[93] . The use of sparse fitting solutions, like lasso regression, can help in diminishing the negative effects of this caveat[120] during the step of estimating exposures. Another popular solution is to only fit relevant mutational signatures for the tissue of interest, which ensures that only biologically pertinent signatures are allowed[75,93] .

Another common handicap in the extraction of mutational signatures is the bias in quantification between more and less sparse signatures[93,119] . Signature 1, 2 and 17 for instance, have sparse profiles making them easy to identify by a mathematical model. On the other hand, signatures like 3, 5 and 8 have more ‘flat’ (less sparse) profiles that impede their reproducible extraction across even the repeated NMF runs for the same datasets[93] . It’s important to note, that these ‘flatter’ (less sparse) signatures are normally less strongly associated with a known etiology, potentially due to the technical caveats in inferring the signatures.

Mutational processes with a highly concentrated localization pattern across the genome are also problematic for the correct extraction of its signature. An example of this is the activation-induced deaminase (AID) related mutagenesis in B-cells (see section 1.4.2.1) where the mutation spectra is defined mostly around the targeted immunoglobulin sites and, thus, when relying on the whole genome, the signature might get ‘diluted’ within the other less sparse mutational processes. A way to address such cases is normally by separating the targeted loci from the rest of the genome and perform inference[97,119,121] .

Finally, an important caveat of the mutational signature extraction method is the detection power in small sample sets. An example of this problem are rare chemotherapeutic treatments, for which not enough tumor samples are normally sequenced in the current datasets. When treated *globally*[96] or in a pan-cancer setting, these processes are sometimes less evident due to the reduced number of samples that carry them[93] . In order to solve this problem, a *local* extraction sacrifices the stability of a larger matrix to obtain a higher representation of the rarer processes. For this approach, samples are normally classified according to a meaningful biological feature, i.e. tissue of origin, and signatures are extracted independently of each set, to be later combined. This mode of action significantly increases the power to detect a larger and more diverse set of signatures some of which may be rare[23] , but may also split unique biological mutagenic processes into multiple signatures and increase the noise derived from the factorization of smaller sample sets. A related technique of "hierarchical extraction" repeats NMF iteratively while downweighting or removing samples that were already adequately described by existing signatures[113] .

1.2.1.4 Known etiologies of mutational signatures

The discovery and characterization of the mutational signature etiologies is an important field of research that helps elucidate the underlying molecular mechanisms and to predict and control the mutagenic processes, as well as its roles in cancer risk and evolution.

For some of the first mutational processes described[90] the mutation profiles obtained provided an indication to which element was responsible for the mutagenesis. Extensive research prior to the large genomic datasets (see 1.1.1) already highlighted sequence preferences for some of the most mutagenic agents in nature, i.e. UV damage focused at YY sites. A second line of evidence in the discovery of causality in a signature is the statistical association between the presence, or the rate, of the mutations, and the clinical metadata of the samples with a high 'exposure' to that signature. A clear example of this is the detection of the signature 31 which is only present in tumors of patients with[122,123] prior platinum drug treatment.

A more definitive line of evidence is to recapitulate the accumulation of a given signature in an experimental setting. A mutational accumulation assay can be used in combination with either an administration of a mutagen or the disruption of a relevant DNA repair gene. The resulting mutations are measured and compared with a signature catalog and linked to their experimental condition.

One remarkable example of this type of associations was the confirmation of the mutagenic mechanism of signature 2 and 13 (associated to APOBEC mutagenesis, see 1.4.1.2) where the gene encoding for the human protein was introduced in the yeast genome and selected via a mutation reporter[124] . The sequences of the selected yeast clones revealed a defined trinucleotide pattern for both APOBEC3A and APOBEC3B proteins which are considered to be the main mutagens in human cancers. Another example is the characterization of signature 14 which associates to a co-deficiency in the MMR pathway and the correction mechanism of POLE[125,126] in a genomic analysis. POLE deficient cell lines were edited using CRISPR to generate MMR deficient clones, yielding a significant shift in their mutational profile to confirm the interaction of the two processes generating the signature[125,126] .

The experimental validation of a signature can also be done more systematically by applying a set of genetic alteration or chemical exposures to the same biological system[81,103,127–129] . These types of validations were introduced first in worms (*C. elegans*) by sequencing the descendants of a self-fertilized multi-generation line[128,129] . These studies yielded experimental evidence for mutational patterns arising from the MMR deficiencies[128] such as signature 6, 15, 21 among others. In similar experiments, authors knocked out key repair genes and administered exogenous mutagenic chemicals to the worms[129] . The resulting mutations represented the interaction of both the damage caused by the chemical particularly present in the KO of the gene responsible for its repair[129] .

Further large scale experiments have also been conducted in human cell lines. In these examples, cell lines are either edited with specific KO[81,127,130,131] in a set of DNA repair genes or alternatively grown in a plate with a given administered genotoxin[103,132,133] . After a given time accumulating mutations, these cultures are seeded to extract single cell clones which are expanded to a sufficient DNA amount to be sequenced. The sequences of the daughter cells are then compared to the progenitor population, yielding the mutations accumulated during the assay. Although these systematic experiments yield valuable one-to-one associations of a given mutational pattern with its causal agent, they are also costly and time-consuming. The published datasets at the time of writing are expected to grow considerably in the coming years with the development of better tools for the gene editing of human cells, the better detection of somatic mutations, and automation in the cell culture techniques.

The experimental validation represents the empirical approach to the characterization of a mutational signature, however, it does have limitations. If used in a model organism, the genomic and molecular characteristics of the chosen model organism might modify the pattern at which specific trinucleotides are mutated. Human cell lines have also reported artefacts, such as the increased mutation rate in C>A mutations caused by the high oxygen percentage of the culture conditions[127,134] similar to the signatures 18 and 36.

Another methodology that can be used in order to elucidate the molecular etiology of a mutational signature is to detect and characterize causal germline associations with their rate. A classical example is the association of a polymorphic loci linked to the fusion of the APOBEC3B and APOBEC3A transcripts which yields a substantial increase in the signature 2 burden[135,136] , expanded in section 1.4.1.2 .

In some cases, mutations do not directly occur by the external environmental agent acting upon DNA, but from an intermediate chemical species generated by the mutagenic agent. This has been termed *secondary exposure*[93] and explains how two distinct etiologies might converge onto the same mutation signature. An exemplary case is Signature 17 which originally was attributed to the exposure to gastric acid[137,138] in esophagus cancers. Further studies also identified the chemotherapeutic agent, fluorouracil (5-FU) as a potential cause of these mutations with exactly the same signature[139] . Although more research is needed, both mechanisms seem to be compatible with an intermediate enrichment of the oxidized form of the free guanine nucleoside (8-oxo-dGTP) which is then wrongly incorporated into the nascent DNA molecule, pairing with T or A instead of C and resulting in a mismatch[93] .

Another complex etiology is time. Two mutational signatures in particular, signature 1 and 5, have a positive correlation with the age of the cancer patient[91] . These two signatures seem to also be pervasive in every tissue, including non-cancerous samples[66,77] and even other mammals[46] . For signature 1, the molecular mechanism that generates the mutation is likely deamination of the methylated cytosine[12,13] generating a thymine that creates a mismatch that eventually gets fixed through replication. Although the deamination of the cytosine should occur at a constant rate, the mutation fixation step depends on the division of the cell. Therefore, signature 1 is associated with age at different rates, with high dividing tissues, such as colon stem cells, exhibiting a faster accumulation compared to other tissues that divide more slowly, such as kidney or breast epithelium[77,140] . Signature 5, on the other hand, seems to be mechanistically more elusive. Its characteristic 'flat' trinucleotide profile represents a challenge in the determination of its source. Reports have suggested a variety of potential mechanisms[127,141,142] with a likely involvement of error-prone DNA polymerases via the REV1 scaffolding protein[143] . Thus, despite the numerical correlation with age, the lack of a plausible molecular mechanism represents a challenge in determining the true etiology of the very widespread and abundant mutational signature 5.

1.2.2 Clinical relevance of the identification of mutational processes

The use of (exclusively) genomic factors for the approval of a cancer treatment was first granted by the US Food and Drug Administration (FDA) in 2017 for the use of pembrolizumab in MSI-H[g] tumors and later in 2020 in high tumor mutational burden tumors (TMB-H)[144,145] . resulting from the deficiency in MMR and evident

[g] Resulting from the deficiency in the MMR pathway, see section 1.3.1.1

in the microsatellite instability (MSI) phenotype . The high mutagenic potential of POLE and POLD mutants can be used in order to select immunotherapies that rely on the increased generation of cancerous epitopes[146] .

Like the TMB of a sample, which broadly captures all mutation processes in a cancer, mutational signatures have the potential to provide a finer-grained classification of cancer patients with the potential to improve the classification of patients and provide more targeted treatments[11,147] .

A direct example of this approach is the use of mutational signatures to predict, statistically, if a given tumor sample is Homologous recombination (HR) deficiency and thus can benefit from PARP inhibitor treatment[93,148] . A more recent report has also used the information of mutational signatures extracted from cell line panels to systematically associate them with their response to an array of drugs highlighting hundreds of novel associations[149] .

Another interesting use of the genomic technologies for the detection and treatment of the cancer is the sequencing of circulating tumor DNA (ctDNA), that escapes from the tumor mass and that still carries information about the source tissue[33–39] (see also section 1.1.2.1). The use of somatic mutations in this ctDNA setting is still in its infancy[38] but prior work on the tissue of origin classification based on these and other features[150,151] has already provided conceptual frameworks for when more data sets become available.

1.3 Somatic modulators of mutation rate

In mutation accumulation experiments[6] in cell culture, the mutation rate is normally determined as the number of mutations over the covered genome and time (expressed in days or in generations). Other experimental settings, like trio (mother-father-offspring) sequencing, can also be used to obtain similar information as they yield the number and spectra of mutations accumulated in one generation.

In intact somatic tissues however, the determination of the mutation rate is more problematic as there is no clear factor which can determine when or for how long a mutational process has been active[h]. For tumor samples and healthy somatic tissues, the relative mutation frequency is normally used as a proxy for the mutation rate, assuming that various processes were active for similar fractions of the time elapsed.

[h] The main exceptions to this limitation of somatic mutations are the mutational signatures linked with age, such as signature 1 and 5

1.3.1 Global modulators of mutation rate

In human somatic tissues, the mutation rate varies within multiple orders of magnitude across tissues and individuals, which highlights the importance of biological modulators of mutation rate. These are particularly important in the accumulation of spontaneous replication errors, but also in the repair of DNA lesions caused by exogenous and endogenous factors.

1.3.1.1 The mismatch repair pathway and its role in the control of mutation rate

DNA mismatch repair (MMR) is one of the key regulators of mutation rate in a wide range of organisms[6] . Its main function is to detect and repair replication errors, both DNA mismatches and small DNA loops (replication DNA slippage products). However, it also corrects failed recombination events and can trigger DNA damage response signals eventually triggering apoptosis[152–154]. The MMR pathway can be divided into two main components, MutS and MutL. The MutS component detects the mismatches, while the MutL component is necessary for actual repair by stimulating excision. In humans, mismatches and small loops are detected by MutSα, a heterodimer formed by MSH2 and MSH6. Loops of a wider range of sizes, however, are detected by the MutSβ , an alternative heterodimer formed by MSH2 and MSH3. Thus, it is important to note that depletion of MSH3 or MSH6 would then have different genomic effects downstream. Indeed, in MSH3 deficient tumors more indels accumulate but not SNVs, compared to MSH6 deficient samples where more SNVs are detected[127,155] . The second component (MutL) is also divided into 3 different heterodimers, MutLα (MLH1 and PMS2) , MutLβ (MLH1 and PMS1) and MutLγ (MLH1 and MLH3). At the time of writing, only MutLα is known to have a significant effect in the control of mutations. Finally other external components participate in the MMR pathway like EXO1[156] an exonuclease performing the excision of the mismatched strand; and PCNA, a DNA replication protein which was reported to modulate different steps in the MMR pathway[157,158] At the germline level, the deficiency of MMR causes several disorders that are characterized by an increase in overall mutation rate and an increase in cancer risk, particularly in colon but also other cancer types such as uterus (endometrial), stomach or ovarian. Lynch syndrome or hereditary non-polyposis colorectal cancer syndrome (HNPCC) is the most common and studied MMR deficiency documented in humans[159,160] . The deficiency affects mostly the core genes of the MMR pathway ([IT MSH2] , *MLH1* , and less commonly *MSH6* and *PMS2*) but a considerable percentage of cases though remain orphan suggesting that other variants still need to be characterized[161–165] . Interestingly, the inactivating variants are heterozygous, requiring a somatic loss of heterozygosity (LOH) event[166] to increase the mutation rate and increase cancer risk. This *second-hit* inactivation can take different forms, but the most common seem to be somatic mutations[167] and promoter hypermethylation of MLH1[168] . Interestingly, these secondary alterations are also causal of the deficiency of MMR at the somatic level (see below). Another type of germline inherited MMR deficiency is constitutional MMR deficiency (CMMRD), which is responsible for an increased risk of early onset brain and blood cancers[88] . These patients have homozygous inactivating variants in the core MMR genes[155] thus causing an increased mutation rate earlier in life. Contrary to Lynch syndrome, however, the more commonly associated genes are PMS2 and MSH6[88] . Some of these patients also generate a particular hypermutator phenotype which arises from the combination of MMR loss and the somatic deficiency of the DNA replicative polymerases (either Polϵ andPolδ)[88,169] . In tumors, the somatic inactivation of MMR causes a characteristic mutational phenotype known as microsatellite instability (MSI) because of the accumulation of indel mutations in Micro Satellite (MS) loci due to replication DNA strand slippage. Although MS are hotspots of mutagenesis within populations, in a typical somatic sample they remain relatively stable (although still with mutation rates higher than non repetitive DNA). If MMR cedes, the indels occurring at those sites cannot be repaired and the number of copies of the repeat units in each MS becomes unstable. This characterization of the samples through computational analysis represents a powerful tool for the detection of MSI cancers in research, but still represents a costly endeavor in the clinic as the whole genome needs to be sequenced to reach significant accuracy. More recent publications use machine learning to classify if a sample is MSI or not based on mutational features like the type of mutations accumulated or the number of indels (see 1.2)[170–172] . Deficiencies in the MMR pathway are characteristic in certain tissues like colon, stomach, uterus, and therapy resistant gliomas[173] . However, the detection of MSI-H samples in a pan-cancer model indicates that other tissues might also contain a lower but significant percentage of MSI cases. A particularly relevant finding regarding this tissue specificity is the enrichment of mutations at certain MS loci depending on the tissue of interest[172] . This finding fits well with the observation that SNV mutation rate also correlates with tissues due to the differential regional activity of MMR in different tissues[150,174,175] . In addition to indels at MS sites, MSI samples also accumulate a substantial excess of SNVs. The study of mutational signatures (see section 1.2) has revealed several that present a significant association with the phenotype (signatures 6, 15, 21, 26 and 44; and 14 and 20 in association with DNA polymerase deficiencies). However, it is still not clear what molecular characteristics generate the distinction between them. The most direct evidence available comes from MMR deficiencies for specific components of the pathway. Cancer genomes which are deficient in the MutSα component accumulate more mutations in the C>T side of the spectrum with a particular enrichment at the CpG sites (similar to signature 1 and 6) while mutants in the MutLα have a more classical signature with also C>A and T>C mutations[128,155,176] (the role of MMR in CpG mutations is expanded in section 1.3.2.3). The analysis of MMR KOs in human cell lines, however, seems to reproduce these findings partially, while mutations in *MSH6, MSH2* and *MLH1* each generate a complete signature with C>A, C>T and T>C mutations (similar to signature 44), the *PMS2* KO preferentially accumulates T>C mutations[127] . At the moment, more evidence is needed to confirm the distinctive mechanisms of the different signatures associated with MSI and which technical conditions, such as the use of only whole genome sequences, allow a better estimation of the mechanisms underlying various MMR-associated mutational signatures.

1.3.1.2 Other germline alterations that modulate mutation rate

Germline deficiencies in the members of the Nucleotide excision repair (NER) pathway can also yield a substantial increase in mutation rate and cancer incidence[177] . Patients with these deficiencies often suffer from Xeroderma pigmentosum (XP) and Cockayne syndrome (CS) which are characteristic for its increased rate of skin cancer and neurologic abnormalities. Interestingly, deficiencies in the transcription associated subpathway NER are more prone to generate neurodegeneration in CS while deficiencies in the genome-wide subpathway are more likely to generate skin cancers[177] . These tissue specificities might be attributed to the role of NER as a main repair pathway responsible for exogenous mutagens like Ultraviolet radiation (UV)[129] . The study of XP deficient patients[178] has also been pivotal for the study of mutation accumulation in NER deficient conditions (see section 1.3.2.2). More recent studies that aim to sequence healthy tissues (see 1.1.3) have also focused on patients with DNA repair deficiencies. In particular, patients with MUTYH-Associated Polyposis (MAP), deficient in the MUTYH protein, part of the Base excision repair (BER)[75] and patients deficient in the DNA replicative polymerases (*POLE* and *POLD1*) responsible for proofreading-associated polyposis (PPAP)[76,179,180] . Both these cases yield a significant increase in mutation rate in healthy somatic cells that is comparable with human cancers. Other rare genetic diseases are also caused by deficiencies in DNA repair proteins or replication enzymes and affect the rate of structural and complex mutations. As with SNV rate associated genes, these deficiencies increase the genomic instability of the tissues and increase the rate of carcinogenesis. Examples of such conditions include

1.3.1.3 Other Somatic factors that increase mutation rate

Similar to the germline associated hypermutators, somatically-altered global modifiers of mutation rate also involve the core DNA repair pathways and the replicative DNA polymerases. The three more prevalent categories of hypermutators are the MSI (see section 1.3.1.1) cancers which have lost proficiency of MMR, the Polϵ proofreading domain deficient tumors, and the third group consists in a combination of these two deficiencies.[173] . In addition, extreme exposures to exogenous mutagens can result in very high mutation rates even with apparently proficient DNA repair. In tumor samples, the Pol ϵ deficient patients contain a clear enrichment for signature 10 (divided in 10a and 10b for *POLE* and 10c and 10d for *POLD1*)[96] . Signature 10a, the most abundant of the four, is characterized by C>A mutations at TCT trinucleotides. Signature 10b is characterized by numerous C>T mutations specifically at the TCG context. As with MMR, some reports have suggested an association with the DNA methylation status of the nucleotides[181,182] that will be expanded in the section 1.3.2.3 . POLD1 deficient tumors represent a smaller percentage of the cancers and are thus less prevalent in the global mutation signature extractions. These signatures (10c and 10d) are mostly enriched in C>A mutations, primarily in the TCW (where W is A or T) context[96] . The close relationship between MMR activity and replication makes the interaction of POLE mutants and MSI common and synergistic. In samples with germline MMR deficiency, in particular (CMMRD), some samples additionally acquire a somatic deficiency in the exonuclease domain of the *POLE* gene which further increases the mutation rate and increases its risk of cancer[88] . Some reports using conditional expression of MMR in human *POLE* deficient cell lines have also suggested that a fully functional MMR can compensate for the depletion of DNA polymerase proofreading ability[125,128,155] suggesting that even microsatellite-stable but hypermutating tumors may have some degree of MMR deficiency . However, more evidence is needed to assess this hypothesis directly in human tissues. Other, more rare, endogenous modulators of mutation rate are deficiencies in the BER pathway such as *MUTYH* and *NTHL1* mutants. Patients deficient in these genes have also been observed to have a higher risk of colorectal tumors[183,184] and are generally characterized with a C>A (signature 36) and a C>T (signature 30) predominant mutational signatures. Modulation of the mutation rate can also occur, not just by the lack of repair, but also through an excess of endogenous mutagenesis. DNA and RNA base editors like the APOBEC family of cytidine deaminases (see 1.4.1.1) has been characterized as a prevalent and common DNA mutator in multiple cancer types[90,95,185] . The AID protein, a member of this family, is also a known hypermutator which acts somatically within the physiological mutagenesis occurring during maturation of B-cells (see section 1.4.2.1). These mutations are characterized by a C>T spectrum in TCW contexts for APOBEC (signatures 2 and 3)[124] and WR$\underline{\text{C}}$YN contexts for AID (signature 85)[186] . Although the global mutation rate of these processes seems to be lower than POLE or MMR deficiencies, their localized nature predicts a strong functional impact.

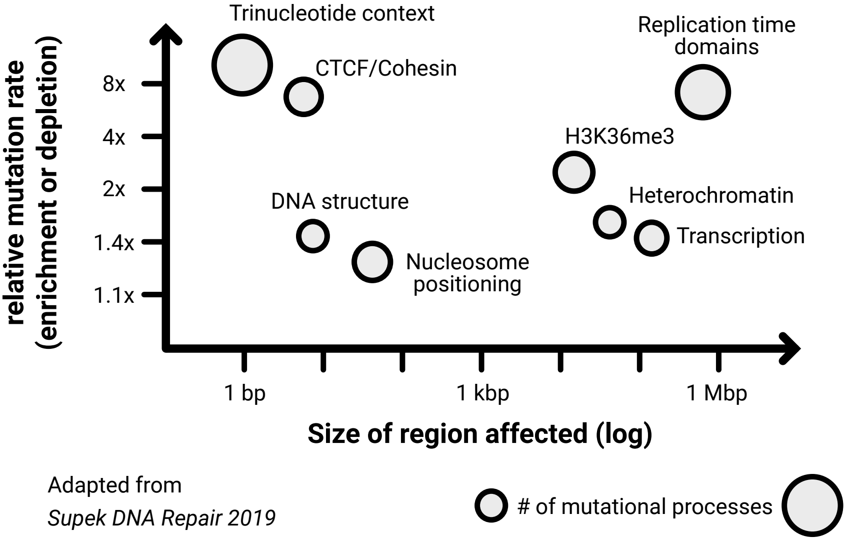


Figure 1.6: Schematic of the scale and relative enrichment of several genomic feature that can modulate the mutation rate locally. Adapted from[187]

1.3.2 Regional determinants of mutation rate

The variation of mutation rates is not only present between samples but has been also detected across the genome[175,187,188] . Thus, the genomic characteristics of a given genomic locus also play a role as a modulator of its mutation rate. These regional modulators can be divided into four main categories depending on their size and mechanism, (i) large megabase-sized domains, like replication time domains; (ii) short functional elements, like the binding sites of CTCF and cohesin; (iii) epigenetically modified loci, like the hypomethylation of the cytosine at CpG islands

1.3.2.1 Regional modulators of mutation rate at larger scales

The mutation rate variability at large megabase-sized domains was first explored after the first cancer genomes were sequenced. Mutations accumulated preferentially in heterochromatic regions (measured by levels of H3K9me3) while depleted in open and active chromatin[189] . Mutation rate also showed a correlation with other global genomic features like replication time, GC content and germline mutation rates[190,191] .

Initial reports suggested that one plausible mechanism was through the combined effect of multiple open chromatin[i] regions[174,192,193] where repair is normally more efficient. While these factors may be relevant at local scales, , replication time was suggested as the more probable causal factor in the determination of this mutation rate variability at the megabase-scale. The genomic resolution at which replication time fluctuates matches closely with the variation in mutation rates and the robust assessment of the correlation with RT, even when controlling for the aforementioned confounders, highlighted it as a more predictive factor[97,175,191,194] . The structure of the genome and the predictability of replication time from other epigenetic factors, like DHS[195] complicates the characterization of the proximal cause molecular mechanism of mutation rate variation.

Later reports, however, showed that the association between mutation rate and replication time was caused through the differential activity of MMR which preferentially targets the early-replicating section of the genome, thus reducing the mutation rates in these regions[175] . Samples with MMR deficiency (MSI) showed a flatter regional density profile and a reduced variability, directly linking the activity of this pathway to the phenotype. This preference of MMR towards early-replicating sections of the genome is conserved across multiple model organisms[6,196] ; the molecular mechanism underlying this process is, however, less understood. A potential explanation is the recruitment of MMR complexes directly toward euchromatic regions during S-phase through the binding of the H3K36me3 histone mark by the MSH6 protein[197] . Other mechanisms like the depletion of a required repair factor in the late stage of the replication or increased use of TLS polymerases or reduced accessibilty of heterchromatin to repair factors, or, more parsimoniously, the reduced time available for repair prior to mitosis have also been suggested as causes of increased mutation rates in late replicating domains[187] .

This model, however, does not explain how tissues associated with exogenous mutagens that generate bulky adducts, like UV and tobacco smoking in skin and lung cancers, also presents a strong mutation rate variance that also correlates strongly with replication time. Reports have also proposed that global activity of NER (the NER branch not associated with transcription), also shows a significant targeting for the early genomic regions[178] and when switched off, mutation densities tend to become flatter, thus different DNA repair pathways are enriched in early-replicating DNA.

Another regional modulator of mutation rate is transcription. Transcription Coupled Repair (TCR) is a branch of NER that gets coupled with the transcriptional activity of the RNA polymerase and quickly clears the lesions along the template strand that block the elongation of the RNA polymerase, disrupting transcription

[i] Open chromatin was defined as regions which are generally accessible to DNA repair genes and was measured with techniques such as DHS, ATAC-seq and ChIP-seq of the H3K4me3 mark

and increasing the rate of DSBs . This strand preference generates a strong imbalance in how many mutations occur in each strand and can be detected through measuring the mutational strand bias. Higher values of transcriptional strand bias are observed for multiple mutational processes where NER participates like signature 7 (UV damage) and 4 (tobacco smoking)[96,98] .

More generally though, highly transcribed genes are generally less mutated on either strand[198,199] but it is currently unclear the overall contribution of the several possible mechanisms. Transcription appears heavily confounded with multiple other mutation modulators such as replication. For instance, while transcription and replication timing are independent processes, more highly expressed genes are located in early replicating regions[200] . Another potential mechanism is the participation of the H3K36me3, an epigenetic mark that recruits MMR and is enriched within the gene body of genes (see section 1.3.2.4)[97,197] .

1.3.2.2 Local modulators of mutation rate

Another type of regional modulation of mutation rate is the binding of proteins to their sequence-determined target loci. Although the mutations in a single site of an individual tumor sample are still too sparse, an analysis pooling across sites and across samples can reveal a strong change in the mutation rate. In simple terms,, the probability of a mutation occurring in that binding site is higher.

One of the most studied of these phenomena is the high accumulation of mutations at CCCTC-binding factor (CTCF) sites. Reports show a sharp peak in mutation rate when pooling various CTCF loci, and centering around its binding domain[201–204] . This pattern is only clear in the functional sites which are defined as both CTCF and cohesin bound[203] . The CTCF and cohesin protein alone, however, cannot be the effector mechanism as their role in the loop extrusion mechanisms is required for the chromosome folding of the nucleus[205] , while the mutation enrichment shows a high tissue specificity and/or mutational signature specificity. It seems to be more common in certain cancer types like colon, stomach, liver cancer[201,202] and melanoma[203] but less obvious or marginal in others.

The suggested mechanism is based on regional impairment of MMR for colon cancers[201] and NER for Skin cancers[203] . The binding of CTCF and cohesin impedes the accessibility of DNA repair proteins and thus mutations accumulated. This model fits with the mutational signatures that occur at these sites, with mostly C>T mutations in the Melanoma, associated with signature 7, samples and T>G mutations in the colon and stomach, associated with signature 17. Other possible mechanisms that have also been highlighted in recent reports is the differential damage which can occur at the CTCF sites, particularly for cyclobutane pyrimidine dimer (CPD) UV lesions[206] .

Similarly to the CTCF loci, a sharp enrichment at Transcription Factor Binding Site (TFBS) was also detected in melanomas[207,208] and to a lesser extent in other cancer types like ovarian and Lung adenocarcinoma[208] . Within these cancer types, tumor genome samples with a high proportion of signature 7 and signature 4 showed a higher enrichment at these sites, suggesting a transcriptionally NER impairment mechanism. The required binding of the TF, the lack of enrichment in *XPC* -/- samples seem to support this hypothesis[208] . Other reports also suggest an enrichment of the UV damage formation, mainly CPD lesions, in a TTCCG motif which is highly conserved in the binding sites ETS family of TFs[209,210] . This enrichment, however, seems uncorrelated with the mutation rate observed in Melanoma[209] . The heterogeneity of the TFBS (including CTCF) sequences complicates elucidating the associated molecular mechanisms[211] .

Another interesting local modulator of mutation rates are chromatin loops or loop anchor point (LAP) which can be defined as two independent loci that interact within each other in trans. These loci are normally detected through high-throughput conformation capture (Hi-C) experiments[212,213] and are associated to the activity of CTCF sites and the loop extrusion mechanism. However, other sources of loop anchors include other protein insulators like YY1 or the activation of transcription through the interaction of enhancers and promoters[214] . Interestingly, both Cohesin bound CTCF loci and TFs normally occur at regulation clusters with high interactivity scores and which may correlate with LAPs[202,215] .

Contrary to the SNV hypermutation seen in the CTCF motif, the structural variant (rearrangement) mutation rates are instead increased at LAPs[216,217] potentially through the increased topological stress that the loop extrusion mechanism generates at these sites[218] . However, SNV rates may in fact be decreased at the regions. The difference in resolution, one at the motif level (11bp for CTCF) and the other spanning multiple kilobases (LAP), suggests that a distinct mechanism might be responsible for these patterns. The activity of AID (see 1.4.2.1) and in particular the off-target cancer related mutagenesis has also been linked to these trans-interacting loci where mutagenesis is targeted to both enhancer and promoter interacting loci[219] which mimic the on-target immunoglobulin sites. Overall, these chromatin loop associated mutation patterns are still underexplored mechanistically and will require further research to elucidate specific mechanisms.

1.3.2.3 Role of DNA methylation as a modulator of mutation rate

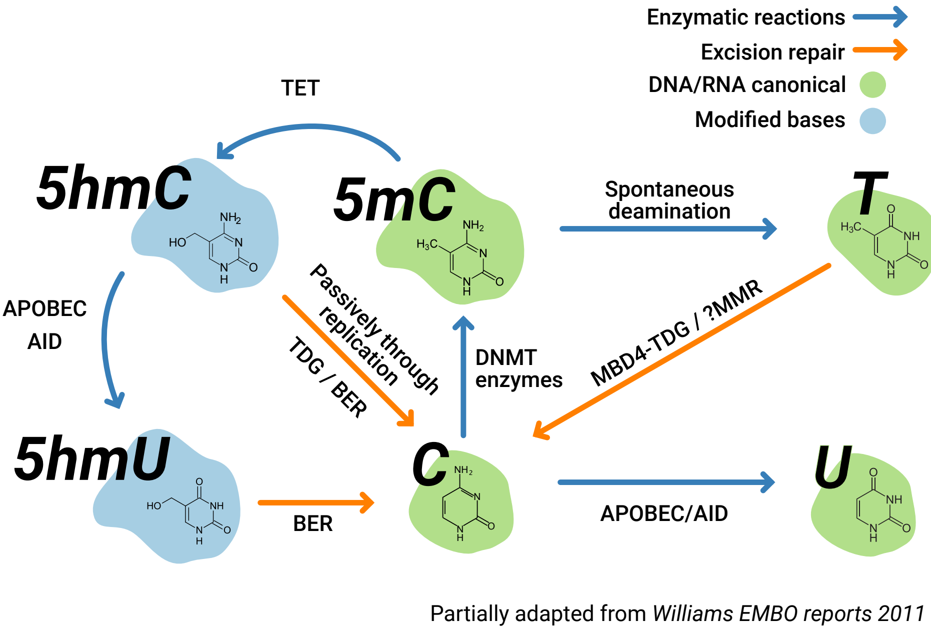
The methylation of the cytosine was first observed in the DNA of several animals and plants in 1950[220,221]. In humans, methylated cytosines occur normally in the CpG dinucleotide although in some tissues like the brain, alternatively the CpH dinucleotides can also be methylated[222] (where H is A, C or T) . The CpG sequences are found in the genome at a lower frequency compared to other dinucleotides, but are locally enriched near transcription start sites. These local accumulations of CpG sites are known as CpG islands and their main role is gene regulation; when the CpG island is methylated, transcription factors can usually bind less well to the promoter, and the gene is normally switched off. This strong silencing capacity makes this system commonly employed by mammals to regulate transcriptional programs related to development[223–225] .

Although in adult somatic tissues the majority of the CpG sites are methylated, certain large sections of the genome appear under constant hypomethylation during aging and cancer[220,226,227] . The current mechanistic hypothesis[226] is that while active and regulatory sites are epigenetically maintained with active methylation, late replicating and peripherally-located (nuclear lamina adjacent) regions seem to passively lose their methylation status over many cell divisions; these are named partially-methylated domains. Of note, this does not imply active removal: the lack of methylation maintenance by DNMT1 passively leads to a depletion of methylation through replication as the newly synthesized DNA strand is not correctly methylated[228,229] .

The first observations that both animal and plant genomes were relatively AT rich and particularly depleted in the CpG dinucleotide were already indicative of the possibility that the 5-Methylcytosine (5mC) may be more mutagenic than the unmodified cytosine, and thus rapidly lost in evolution. Later experimental evidence from hotspot mutation sites in reporter genes *E. coli* confirmed this hypothesis[12,220,230,231] . The methylated cytosine is 15-fold[232] more likely to deaminate directly to thymine, causing a T-G mismatch (see figure 1.3.2.3) .

The DNA repair enzymes responsible for the correction of these methylation related mismatches are MBD4 and TDG, both glycosylases and members of the BER pathway[233,234] . Both enzymes, when deficient, also caused an increase in C>T mutations and an increase in colon cancer risk in mice[234,235] .

While the transition from Cytosine to 5mC seems well understood and represents a straightforward enzymatic reaction through DNA methyltransferases such as DNMT1, DNMT3A or DNMT3B, the mechanism performing the reverse reaction is less obvious. The potential mechanisms are classified into (i) passive through the lack of maintenance (see above), and active, which is more targeted to specific sites and requires the direct involvement of enzymatic activity[224] . A mechanism of action seems to be through the activity of the TET enzymes, which oxydate the 5mC base to 5-hydroxymethylcytosine (5-hmC)[236] for a posterior repair through TDG . Recent reports also highlight the activity of the AID protein (see section 1.4.1.2) which is required for the removal of DNA methylation during mouse development[224,237] and iPS reprogramming[224,238,239] . Considered together, this evidence



Partially adapted from *Williams EMBO reports 2011*

Figure 1.7: Diagram summarizing common chemical transformations that centralize in the Cytosine nucleobase, adapted from[236]

suggests that the DNA repair machinery might be crucial also for the unmethylated genes. However, it remains unclear whether the involvement of base modifiers, like AID or TET genes, and DNA repair proteins, like TDG, leave a relevant mutational footprint in the somatic tissues where they participate.

In the more recent analysis of tumor genomes, signature 1 represents this process, with a sparse profile consisting of nearly exclusively NCG>T mutations (see section 1.2)[90,91]. Apart from the known role of BER enzymes, MMR has also been proposed to have a significant role in the repair of the T-G mismatches generated at methylated sites. The mutations in signature 1 have a strong correlation with replication time[100,182,240] being relatively more abundant in late replicating regions[j]. MSI tumor samples, deficient in MMR, lose this replication time gradient which represents direct evidence of the involvement of MMR (in particular the MutS α branch) in the detection or repair of the intermediate mismatches[175,176,181] . In experiments in worms, where there is no CpG methylation, the main difference in the mutational signature of MSI samples in humans was also the lack of NCG>T mutations[128] .

Apart from signature 1, other mutational signatures with mutagenic preference for any NCG context will likely be modulated by the methylation status of its substrate. A clear example seems to be signature 10, caused by the deficiency in the replicative DNA polymerase ϵ[181,182] . For this signature, the mechanism seems to rely on the incorrect incorporation of an adenine opposite to the 5mC by the defective polymerase, generating a 5mC : A mismatch resulting in a C>T mutation[232,241,242] .

Another example of this is the formation of UV di-pyrimidine dimers (CPD). In early studies of skin cancer cells, the mutation rate of CpG sites in sun-exposed cells was reported to increase significantly upon methylation[243] . The proposed mechanism for this observation is still debated and it is not clear whether more lesions are formed in methylated DNA or if the lesions deaminate faster[244,245] .

Finally, APOBEC and AID mutagenesis are other examples of mutation rate modulation by DNA methylation. AID seems to be less likely to mutate the methylated cytosine but seems able to mutate its alkylated form, 5hmC, generating a U-G mismatch which is then repaired to a unmethylated CpG site (see figure 1.3.2.3)[236,241] . These reports suggest that AID might play a role in the global genome demethylation which occurs in reprogrammed iPS cells and during embryogenesis[237–239] . The evidence for other members of the APOBEC family seems less consistent, with reports suggesting either a reduced[241,246–248] or equal[249,250] deamination activity on cytosines upon their methylation.

In this thesis, I have focused on the mutation rate changes in under-methylated

[j] Although the mutations show no correlation in absolute values, the early replicating parts of the genome contain the majority of CpGs.

regions (UMRs) and aimed to systematically quantify how each mutational signature is influenced by this mechanism. We explore which are the downstream effects of this local variability focusing on functional elements of the genome that overlap with UMRs like promoters, enhancers, or LAPs (see chapter 5). We further investigate gradients in DNA methylation along gene bodies and association with mutation rates for various mutagenic processes.

1.3.2.4 Other epigenetic associations with mutation rate

In addition to DNA methylation, there are other mutation rate modulating factors at the epigenetic level. Histone marks are traditionally used to determine the function of DNA regions, i.e. active transcription, enhancers, and others[251] . Currently, large amounts of Chromatin Immunoprecipitation Sequencing (ChIP-seq) data from the ROADMAP consortia and the ENCODE datasets are available[252,253] making the integration and joint analysis of somatic mutations and epigenetic data accessible.

The first studies in mutation rate variability across the genome yielded many associations with mutation rates and histone modifications[188–190] (see section 1.3.2.1). In brief, SNV rates in cancer are positively correlated with H3K9me3 and H4K20me3, both markers of heterochromatin[189] while negatively correlated with all other examined marks. These associations are also consistent with the correlation of mutation rates with a broader, domain scale feature, replication time[175] which itself correlates with various histone modifications (e.g. the heterochromatin mark H3K9me3 is highly enriched in later replicating DNA) .

One histone mark that is likely directly causal to mutation rates is H3K36 methylation. In particular, the H3K36me3 accumulates a few hundreds base pairs after the TSS and incrementally increases along the gene body[254] . The interaction with DNMT3B[220,255] protein seems to highlight its function by regulating the deposition of methylated groups in CpG dinucleotides in the gene body[224] .

Its described role in mutation rate is mainly mediated through the interaction with the MSH6, a core of MMR protein, during S-phase replication[197] . Analysis of mutation rate along those sites have reported a substantial reduction of mutation rates (up to 2-fold) even when controlling for alternative confounders[97] . This effect in mutation rates is likely caused by the ability of the mark to recruit MMR as this pattern disappears with MSI samples where the pathway is not functional[97,256] . On a related note, active transcription may also increase oxidative damage to gene body DNA, and the enrichment of H3K36me3 may also help counteract that effect[256] .

The interaction of this mark with MMR and DNA methylation, together with the unraveled evidence of the involvement of MMR in signature 1 converges to a model where an increased mutation rate in the gene bodies is molecularly counteracted by the increased recruitment of DNA repair machinery to these sites.

1.3.3 Mutation rate of other mutation types

Signatures of other types of mutational events such as structural variants, including Copy Number Alteration (CNA) and neutral SV, are generally less studied in cancer genomics due to the difficulties in their identification from genome sequencing and challenges in categorizing. Numerically, their frequencies are lower than SNVs reducing the power of most statistical genomic methods. However, these variants hold a strong potential for functional impact, as they can disrupt gene coding sequences and additionally act by changing gene dosage, or juxtaposing genes to functional elements. Gene fusions, for instance, represent one of the prototypical carcinogenic mechanisms of driver gene activation[257] and CNA driver mutations have also been observed in a large range of cancer types[258–261].

New bioinformatics analysis and the more abundant datasets are establishing variant classifications for structural variants. This classification allowed, for instance, the detection of Copy Number Variant (CNV) signatures showing evidence for multiple biologically regulated processes[106] . Currently, around 17 signatures[107–110] have been identified, and they seem to associate to orthogonal molecular traits such as the deficiency of HR through the inactivation of BRCA1 and BRCA2.

Along the genome, the association between local rates of structural variants and epigenomic regions remains unexplored, some reports have shown evidence for an enrichment of SVs in loop anchors[216,218] and promoters[262] . However, the impossibility to generate a sufficient baseline model for these types of mutations makes it difficult to statistically validate these associations.

It remains to be explored, then, if the observed complex structural mutagenic process will contain enough information to depict their molecular mechanisms and how they interact with other somatic processes that generate SNVs (see chapter 1.4.1.3).

1.4 Mutation clusters

We define a mutation cluster, or a local hypermutation event, as a group of 2 or more mutations which occur in close proximity to each other, suggesting that they were generated by the same event. The discovery and study of mutation clusters has been a small part of the discoveries from human tumor sequencing studies, however, due to its close association to the mechanism that generates them it holds a potential to provide substantial insights in the understanding of the molecular mechanisms of mutagenesis.

1.4.1 APOBEC mutation clusters

Discovered in early sequencing efforts[95] and showing a pervasive activity in multiple cancer types, APOBEC mutagenesis is one of the most studied processes in tumor cancers. The full understanding of its biology however remains elusive. In this section we review Its tight association with the generation of mutation clusters and its overall genomic characteristics, trying to understand a bit better the multiple factors that regulate its activity in human somatic tissues.

1.4.1.1 APOBEC/AID family of cytosine deaminases

The APOBEC/AID family of cytosine deaminases represent a diverse set of enzymes responsible for the deamination of a cytosine to a uracil (C-to-U edits in RNA and DNA). They are the most studied proteins in mammals that are capable of performing this reaction[263] , relevant in a surprisingly diverse array of physiological functions and some pathological ones. The **Apo**lipoprotein **B** mRNA **E**diting enzyme **C**atalytic subunit **1** (*APOBEC1,* first named *REPR* from RNA Editing PRotein) was the first member to be characterized in rat and later in human intestine cells[264,265] . The protein is responsible for the editing of a single base of the **Apo**lipoprotein-**B**, *APOB,* transcript, which is physiologically expressed in two isoforms depending on the tissue. The APOBEC1 editing introduces a C-to-U change in a glutamine codon (CAA) to a stop codon (UAA) which reduces the translated protein size from 100 amino acids to 48[263,266]. Due to its sequence similarity, most of the later-identified members of the family share the same nomenclature, although they do not participate in any way in the edition of the *APOB* mRNA.

After the detection of APOBEC1, other members of the family sharing a strong sequence similarity, particularly in the enzymatic domain, were identified and classified in different human tissues . *APOBEC2* was first identified in the skeletal and cardiac muscle[267] and the AID protein in B lymphocytes[268] (see section 1.4.2.1) . The subfamily of APOBEC3 was later genomically characterized as a recent amplified gene cluster in chromosome 22. Initially, no physiological function could be assigned to these genes, and they were hypothesized to act as pseudogenes[269,270] .

In humans, the whole family is thus formed by AID and APOBEC2 which are the most evolutionary ancient forms shared among vertebrates, APOBEC1 which is shared among tetrapods and finally the APOBEC3 gene cluster which appear more recently in placental mammals. Later in evolution the APOBEC3 gene has expanded independently in several branches such as bats and primates[263,271]. All members of the family share the ability to interact either with only DNA (specialists, comprised by AID and APOBEC2) or RNA and DNA (generalists, comprised by A3A, A3G and A1)[263,271].

The physiological function of the APOBEC3 subfamily is the defense against a wide range of viruses via the restriction of viral genomes[263] . In brief, they participate in the defense against retrovirus like the human immunodeficiency virus (HIV)[271–275], against DNA viruses like the Herpes B virus (HBV) or the Human Papillomavirus (HPV)[263,276,277] and there is also some evidence about their role in restricting single stranded RNA (ssRNA) virus like Rubella virus or Sars-Cov-2[263,278,279].

Notwithstanding all the functional diversity of this family, the focus of interest of this thesis, the most relevant role of the APOBEC/AID family in humans it is its capacity to edit or mutate DNA (and potentially RNA[280]) in human somatic cells including tumors.

1.4.1.2 Evidence of APOBEC and cluster mutagenesis in human tumors

The first evidence of a TCN-trinucleotide context mutational signature was observed in a systematic analysis targeted of human kinases in breast cancers and cell lines[281] although at the time no mechanism was proposed for this pattern. Seven years later, the first systematic analysis of breast cancer whole genomes, a total of 21 tumors[95,282], reported the factorization of the mutational spectra observed in these sequences in 5 mutational patterns or signatures (see section 1.2). Even with so few samples, two mutation patterns contained clearly defined and sparse profiles. The first was associated with the NCG dinucleotide (see section 1.3.2.3) and the second was a mutation process enriched in the TCW (where W is A or T) both generating C>T and C>G mutations. The same study[95] , also reported numerous similar mutations at TCW contexts located in close proximity and in DNA strand-coordinated groups. These mutations matched the previously reported sequence predisposition[283] of the APOBEC family of cytosine deaminases. The groups of mutations or mutation clusters were termed *kataegis* from the Greek word *thunderstorm* due to its similarity to the “rainfall plot” [k] and from the terminology used in the first report[284] on cluster mutations (see section 1.4.1.3).

In an independent study published at the same time, Roberts *et al*[185] used a double reporter mutant yeast strain to detect the presence of mutation clusters. Briefly,

[k] The rainfall plot represents the mutations in a somatic tissue with the mutation index or the chromosomal position in the X axis and the distance between mutation pairs in the Y index. Mutation clusters appear as sharp vertical lines while unclustered mutations occur as a cloud in the upper part of the plot

two reporter genes *CAN1* and *URA3* were moved to adjacent positions within the same chromosome. The observed mutation frequency was much larger than the one expected if events in each gene occurred independently. Sequencing the genome of the mutated clones revealed that while the genome-wide mutation rate was not highly increased, a cluster of strand coordinated mutations could be detected at the reporter genes. In the same report, the authors also observed mutations that accumulated in clusters in human tumor genomes. Following the observations from[95] these accumulations contained coordinated C mutations and were enriched in the TCW context. Although none of these studies contained direct empirical evidence of the role of the APOBEC family of deaminases, soon later, reports showed that A3G incorporated in recombinant yeast it was possible to obtain mutation clusters from the activity of the APOBEC protein[285] .

The first reports from the analysis of gene expression in breast cancer cell lines showed a positive correlation of the accumulation of APOBEC mutations with the expression of A3B[286–288]. Later studies in recombinant yeast[285,289] showed that based on the mutational signatures that could be extracted from A3A, A3B, A3C, A3G and AID, only A3A and A3B generated mutations enriched in a TCW context. Further extensions of the yeast experiments also focused on the extended mutational predisposition of the mutational signature. In particular, they found a significant change in the frequency of the first nucleotide of the pentanucleotide mutation context. They showed that while human A3A expressed in yeast had a particular preference for YTCAN contexts, the A3B enzyme preferred RTCAN contexts[124] . They also classified the available tumor samples according to this ratio suggesting that most samples had a A3A-like profile. Other evidence in favor of the A3A protein as the mutagenic element came from association studies of population polymorphisms[135,136]. A germline SNP (in linkage with the fusion polymorphism of the A3A gene body with the 3' UTR of the A3B gene[135]) was associated with a strong enrichment in TCW mutations in their somatic tissues. It is thus unlikely that the protein activity of A3B, which is deleted in these samples, might be the cause of the tumor mutational signature.

A consensus model seems to be an A3A protein with high mutagenic potential but sporadic expression and a A3B protein with less mutagenic potential but with more constant and/or frequent expression. In a recent report using whole genome sequences of a large panel of cancer cell lines[134] mutations at YTCAN contexts appeared sporadically in some clones while not in others. Overall, APOBEC mutations were uncorrelated with expression of the A3A or A3B genes[134] . The final confirmation of the role of A3A in this mechanism comes from the sequencing of KOs in human cells. While the A3A KO did not show TCW>K mutations or clusters, they were still present in the A3B KO[143].

As other members of the family, the mutations generated either by A3A or A3B are characterized by the deamination of the cytosine into uracil. This lesion is then either excised and repaired by the BER pathway or bypassed during replication. If directly replicated, the base pairing of uracil with an adenine in the complementary strand creates a C>T mutation. Due to the efficiency of the UNG1 glycosylase another mutagenic mechanism results from when the uracil nucleobase gets excised. Because the mutation occurs in a ssDNA stretch, the cell requires the use of a Translesion synthesis polymerases (TLS) polymerase to bypass the error. As there is no guide to copy from, the lesion is normally substituted by a random nucleoside. The TLS enzymes such as Pol ζ may incorporate preferentially an Adenine (also generating a C>T mutation) but can also incorporate a Cytosine (generating a C>G mutations)[290–293]. These mechanisms were first confirmed by using yeast strains and APOBEC transgenes[289] but have later been also confirmed partially in human cancer cell lines[134,143] (see figure 1.4.1.2).

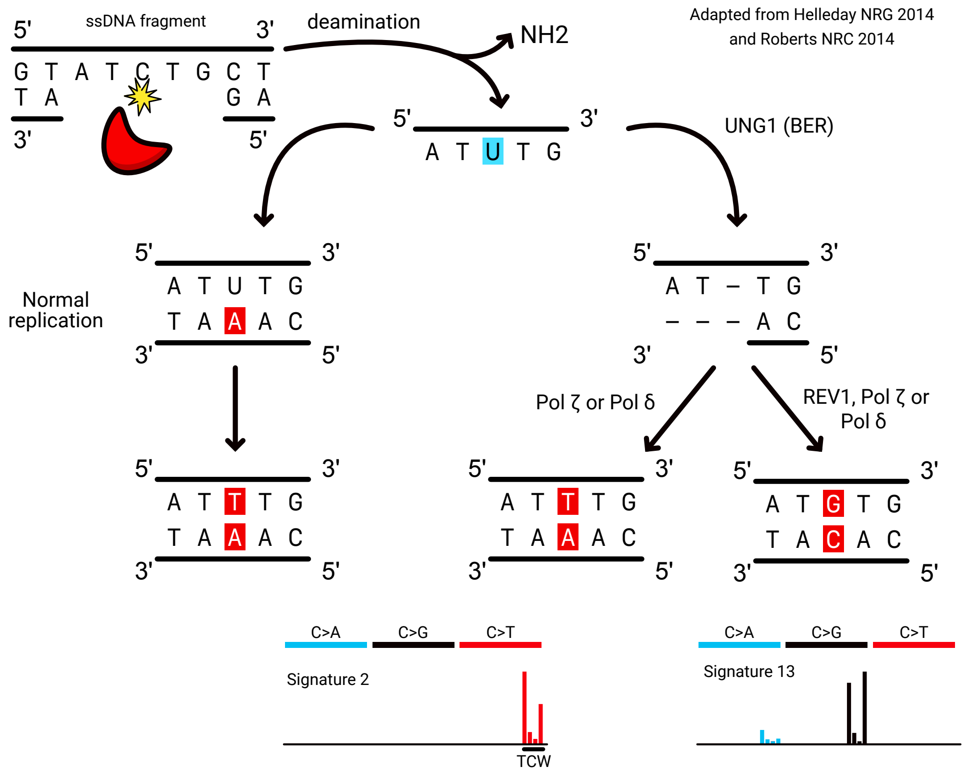


Figure 1.8: Summary of the molecular mechanism associated to the APOBEC mutational signatures (Signature 2 and 13). The initial deamination of a cytosine in a ssDNA fragment by A3A or A3B is then either fixed during replication (left) or excised by UNG1. The incorporation of either an adenine or a cytosine in the complementary strand is mediated by either the TLS polymerases Pol ζ and δ or by REV1.

1.4.1.3 Molecular mechanism of kataegis clusters

The first direct evidence for local hypermutation reported for higher organisms comes from the cluster of mutations observed in Big Blue mice, a mutational reporter assay, where they estimated that up to 1% of the mutations observed in this system were coming from chronocoordinated[l] events. They termed these mutations as *mutation showers*[284,292] . With the detection of APOBEC mutagenesis in human cancers[95,185,288] and generally due to the accessibility of extensive datasets from human tumors, and to some extent trio sequencing[294] , the increased num-

[l] At an equivalent time

ber of available mutations was sufficient to significantly expand the knowledge in mutation clusters.

The mutations associated with *kataegis* were quickly characterized for their tendency to co-localize within rearrangement breakpoint sites[95] . Due to the specificity of APOBEC to ssDNA as a substrate, the association with double-strand break (DSB) repair seems highly plausible. The repair pathways that participate in it, mainly HR, and related processes such as Break Induced Repair (BIR), generate large sections of ssDNA which could potentially be attacked by the protein. Only BIR has shown direct experimental evidence for the association with *kataegis* although it was tested in a yeast system with MMS chemical treatment, therefore, potentially different to to human APOBEC[295].

In cancer, structural variants (SV) are often used as a proxy for activity of these DSB repair pathways, which may occasionally result in erroneous rejoining and thus a SV. Consequently, mutational signatures extracted from SV have shown also significant correlations with *kataegis*[104,107] . Other reports in experimental systems also show a high activity of A3B mutational signature within chromothripsis, a large cluster of rearrangements that span multiple chromosomes in a single genome[296,297]. However, data from multiple cancer types revealed that while A3 kataegis can co-occur with chromotripsis, it does not seem common, with only 9.3% of the samples with chromotripsis displaying significant *kataegis* activity[298] . Thus, although the link with SV and DSB repair is clear, further direct experimental evidence would be needed to confirm that the ssDNA intermediate in the DSB repair is used as a substrate for APOBEC in human tumor cells.

In this section and throughout the literature, the term *kataegis* is mostly used as synonymous for *clustered APOBEC mutagenesis* . However, it can also be used for other mutational processes that generate large focal mutation clusters. One example are the MMS chemical exposures described above. A further example of this is Somatic Hypermutation (SHM), which will be expanded further in section 1.4.2 .

1.4.1.4 Genomic characteristics of somatic A3 mutations

Since its detection in human cancers, the study of A3 mutagenic properties has been of interest for many researchers in the field. The large number of mutations generated and their potential to drive tumor evolution makes the APOBEC mutagenesis an interesting druggable pathway.

A limiting factor in determining the APOBEC mutational processes is the availability of its substrate, ssDNA[263] . The interaction of other biochemical features like its different efficiency at methylated sites are reviewed in section 1.3.2.3 .

Early in the detection of APOBEC mutagenesis and their clustered pattern, the two main hypotheses for potential sources of ssDNA, apart from DSB repair, were proposed: transcription and DNA replication. During transcription, while the RNA polymerase copies fromthe template strand, the coding strand remains in ssDNA form. Highly transcribed genes would then expose significant portions of ssDNA that could in principle be mutated by APOBEC. Some evidence of this exists for AID, closely related to APOBEC, who targets transcriptionally related ssDNA generated at the immunoglobulin loci[292,299]. This hypothesis was, however, early discarded due to the lack of transcriptional strand bias in the analyzed tumor genomes and recombinant yeast[100,246,300–302].

Replication strand bias was, however, detected in early reports about the genomic properties of APOBEC mutations[246,301] and has been widely confirmed in other more systematic studies[99,100,104,240] . The replication strand bias suggests that APOBEC deaminates preferentially cytosines in the lagging strand compared to the ones in the leading strand. This effect generates a bias in the mutations observed in tumors when adjusting the reference base with respect to the replication direction[98] . This strong bias for the lagging strand suggested a hypothesis that APOBEC targeted preferentially the ssDNA sections in the Okazaki fragments during replication. Other evidence from yeast, also suggested that chemically and genetically induced replication stalling also increased the capacity of APOBEC to generate mutations[301].

Another feature that was early associated with APOBEC mutagenesis is its relative enrichment in early replication time and gene-rich regions of the genome[100,302]. APOBEC mutations presented either a flatter profile[100] or a direct enrichment in the early replication sections[246,302] . This slope was even more pronounced within genomes of tumor samples that were individually more enriched with APOBEC mutations and mutation clusters[246,302] proposing a direct link with APOBEC activity. Although this correlation with replication time strengthened the association with replication, it is not clear that it supports the causal link to Okazaki fragment mechanism.

A report from Chen et al[303] introduced an alternative source of ssDNA fragments for the activity of APOBEC: the intermediate DNA state of both MMR and BER pathways. They observed that when introducing an artificially induced mismatched sequence into mammalian cells, the flanking sites accumulated unexpected mutations in the strand where the mismatch was introduced. These flanking mutations were strongly enriched in the TCN context, suggesting the implication of either the A3A or A3B genes. Further genetic knock-down (using siRNA) confirmed that the activity of various A3 genes was responsible for this increment in mutation rate. Other knock-down experiments at BER and MMR genes also yielded a reduction in the mutagenesis in the flanking sites, confirming how the activity of both MMR and BER could induce mutagenesis *in vivo* . The reduction of mutations was consistent with the type of the introduced mismatch, higher for MMR genes in T/G mismatches and BER genes for U/G mismatches. This report introduced substantial evidence for the possibility of MMR to associate with APOBEC mutagenesis[304]. Interestingly, the genomic characteristics of MMR activity 1.3.2.1 , mainly enrichment in early replicating regions[175] and bias towards lagging strand[305] , fit well into a model where the intermediate ssDNA fragment during the repair of a mismatch could work as a source of ssDNA for the overall mutagenic event caused by A3 proteins in human tumors.

In this thesis, I have systematically quantified clustered mutations in somatic tumor datasets with improved statistical methodology to control for false discoveries. Focusing on APOBEC mutation clusters, I have described and characterized genomic footprints of a novel molecular mechanism that causes diffuse mutation clusters. The same mechanism may also be responsible for a substantial portion of the unclustered APOBEC mutations (see chapter 3), and generates mutations with unusually high functional impact.

1.4.2 Other sources of local clustered mutations

Since its discovery in APOBEC mutagenesis, other mutation processes generating clusters, like somatic hypermutation via AID, or usage of TLS (error-prone) DNA polymerases, have been extensively characterized now in human tumors. Most of these processes were already known to generate mutation clusters in model organisms or cell line models by prior research, but they still missed the observational evidence suggesting that they also occur in human tissues *in vivo* .

1.4.2.1 Mutations by AID and Somatic Hypermutation

Human antibody proteins are built from a heavy (encoded in the *IGH* gene) and a light chain (encoded by *IGK* gene for the κ type and *IGL* gene for the λ type). Each chain is formed by a constant and variable region. Within this variable region 3 types of gene segments (variable, diversity, and joining), are encoded sequentially in the genome sequence. After differentiation of the B or T cells, only one segment from each type will be included in the final transcript. This process is called V(D)J recombination and is mediated via the RAG proteins. Recent integration analysis of this pathway with chromosome folding studies seem to suggest that the loop extrusion mechanism, and thus CTCF and cohesin binding, seems to play an important role in this step[306] (see section 1.4.3.2 and 1.3.2.2).

In addition to this diversification process which randomizes the somatic genomic sequence of the antibody genes, an extra layer of diversity is included through the process of Somatic Hypermutation (or SHM). This process consists in the initial activity of the AID protein (see section 1.4.1.2) that targets the promoter of the immunoglobulin genes and deaminates a cytosine, with some preference to the WRCY tetranucleotide motif[186]. The lesion leads to its repair through BER and/or MMR[307,308] . Either the direct fixation of the uracil or the repair by short-patch BER seem to generate C>T mutations, which are characteristic in the AID signature, signature 84 in the Cosmic catalog. Alternatively, the lesion will be detected by MMR machinery, particularly by the MutSα complex[309], which then recruits a strandless and error-prone version of the rest of the pathway[310]. Although how MMR switches between these two modes is not fully understood, evidence suggests that PTMs in the PCNA protein, required during the re-synthesis of the gap, lead to the recruitment of TLS polymerases, mainly Pol η[311] . Thus, the DNA synthesis is extended by generating clustered mutations in A:T pairs around the immunoglobulin genes[97,307,308,310] . The study of blood tumors has revealed a significant amount of non-APOBEC kataegis events, both due to the activity of AID and pol η in proximity to the IGG loci or near known AID off-targets[99,119,312] .

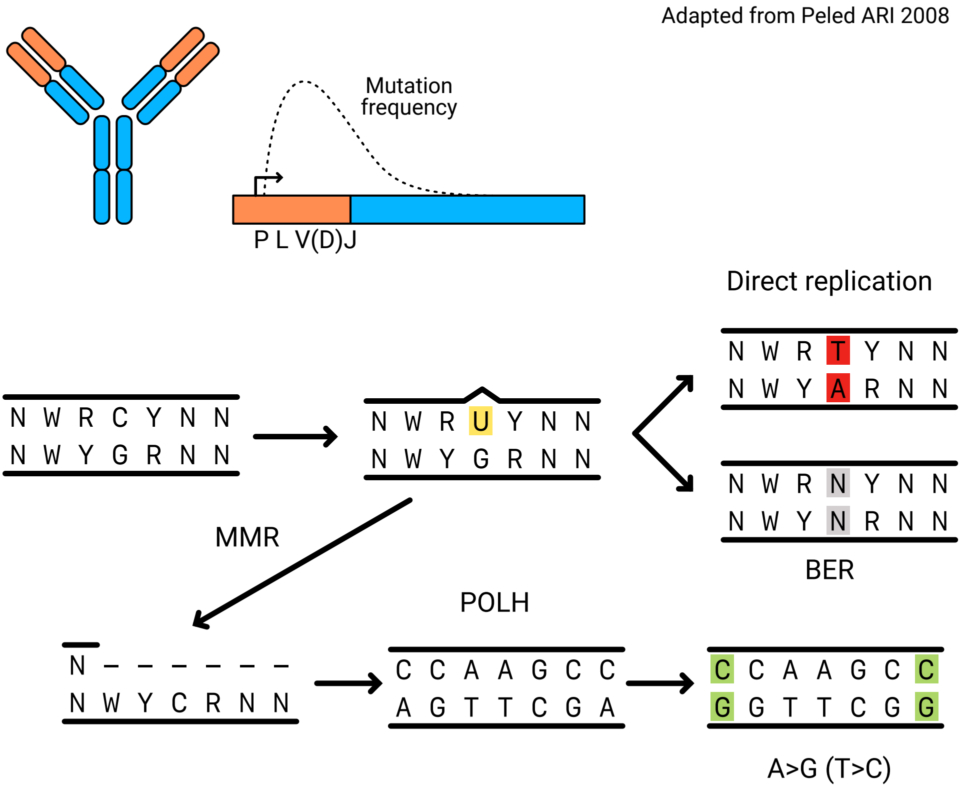


Figure 1.9: Summary of the molecular mechanism of the Somatic Hypermutaiton process happening in B and T cells during its differentiation. Initially, AID deaminates a cytosine to a Uracil (left), triggers its repair either through BER creating WRCYN>N mutations or through MMR which recruits the TLS polymerase Pol η that causes A>G cluster mutations. Adapted from ref[307]

1.4.2.2 Mutation clusters by TLS polymerases

The complete process of SHM seems, at the time of writing, limited to the lymphocyte differentiation. However, the use of an error-prone version of MMR seems to be more widespread in other tissues. For instance, treatment of cells with certain genomic stress chemicals such as alkylating or oxidative damage[310,313] seems to trigger this mutagenic MMR branch.

Some analysis of localized hypermutation in breast cancers revealed a small percentage of *kataegis* events with a significant enrichment of signature 9, possibly related to pol η[104] .

A systematic analysis of the clustered processes occurring in human tumors[97,314] led to the identification of a strongly clustered pol η mutational signature at A:T pairs, enriched in WAN>G (equivalent to NTW>C) motifs. Although this signature was mostly present in lymphomas, there was a significant contribution in a wide range of tissues, specifically, liver, melanoma, bladder, lung, stomach and esophageal tumors. In these solid tissues, these A>G mutations were not associated with promoter features as in blood tumors, but presented a strong association with H3K36me3 and other characteristics of MMR. The switch to this error-prone mode was associated with an increased exposure to carcinogenic elements such as alcohol for the liver and UV exposure for the skin[97] . One of the more important takes from this analysis was that not only this process was generating mutation clusters, but was also responsible for the introduction of significant numbers unclustered A>G changes in the rest of the genome in a *single mutation* clustered events.

Other mutational processes that generate clusters might still be identified as the amount of available data grows; recent efforts have reported tens of clustered signatures (9 from ref[97] and 9 from ref[99] with 5 overlapping) which contain plausibly new sources of non-classical mutation clusters. A set of plausible candidates might be associated to the activity of a wide variety of TLS enzymes with significant prior evidence in the germline[315,316] .

1.4.2.3 Cluster mutations in structural variants

Due to their complexity, mutations clusters of structural variants are less characterized; moreover they are significantly more scarce. In terms of mutation clusters, the best characterized example is the SNV cluster co-occurrence with APOBEC *kataegis* events (see section 1.4.1.3) and AID activity, where the SVs mark regions where presumably there was availability of ssDNA. Recent whole genome sequencing reports however have also suggested that the structural somatic variants can also occur in proximity to other structural variants; bioinformatics methods to identify and resolve such “complex SVs” (clusters of SVs) are rapidly evolving[317] . In Hadi *et al*[318] , the authors used a genome graph to redefine the topology of structural variants and detected 3 novel types of clustered structural variants. The first component is characteristic for small clustered insertions named *pyros* from the Greek word tower, a second component called *rigma* from the Greek word chasm which is characterized by large clustered deletions and finally a third process named *tyfonas* from the Greek word typhoons which represents large sections of the genome with a high number of copies.

Another process which might be considered as clustered structural variants is chromoplexy, first reported in prostate tumors[319] it describes multiple distant regions which are all disrupted at once, multiple DSB which are then re-joined outside their original source. Further experimentation is needed to determine if the process generating the breaks acts in a coordinated manner or just at a higher rate.

These clustered rearrangements are a good example that mutation clusters go beyond APOBEC mutagenesis. In the next section, we argue that the concept of localized hypermutation can be more generally defined in order to include other types of mutagens and mechanisms.

1.4.3 Generalization of mutation clusters

In the previously surveyed literature and generally through this thesis, mutation clusters are defined as a group of somatic mutations in proximity of each other in the one-dimensional DNA sequence, the reference genome. A more broad definition, however, might encompass other types of mutation clusters such as clustered mutations in the germline, or mutation clusters in trans-interacting genome loci i.e. those which are close in three-dimensional space due to chromatin folding.

1.4.3.1 Mutation clusters in the germline

In cancer and somatic tissues, the detection and classification of local hypermutation or mutation clusters represents a relatively easy task because of three main reasons. (i) The lack of recombination makes the InterMutational Distance (IMD) a direct proxy for the proximity of mutational events; (ii) the known mutational processes allows generating a robust baseline of somatic mutagenesis to compare against while this baseline is less clear for the germline, and (iii) the diverse set of mutational processes allow for extraction of informative mutational signatures. These 3 main conditions, however, are generally not met in the study of germline mutagenesis, heavily convoluting the study of clusters. The first studies on population genetics data[315] looking for germline clusters described a mutational signature associated with the activity of Polymerase ζ which was detected upon clusters spanning tens of nucletides between mutations. This signature is characterized by GA>TT and GC>AA mutations, which were previously identified to come from pol ζ in yeast experiments[320] . Because the detection of clusters in this analysis requires that the groups of mutations occur at perfect Linkage Disequilibrium (LD), the limited sample size used here represents a difficulty for the analysis. More recently, in the analysis of the TOPMed program dataset[25] , authors selected only singletons from unrelated individuals to reduce the effect of recombination and selection in their samples. After this strict filtering, they extracted multiple components from the IMD distribution using an exponential mixture model analysis. The first component is short (10bp) and its suggested mechanisms involve the activity of the TLS enzymes. They were also able to classify a second, longer (500–5,000 bp) process which is characterized by the enrichment of C>G mutations which is consistent with prior studies of *de novo* variants[321,322] . The last 2 components occupy large spans and their trinucleotide mutational profile is more flat. Thus, the possible molecular mechanism still remains unclear. Another report analyzing the same TOPMed data also verified this observation when extracting mutation signatures from rare population variants. In two out of the nine mutational processes, with the same characteristic C>G mutations, where mutation clusters could also be detected[323] . Another type of germline mutations which more directly represent the direct mutation predisposition of the germline are *de novo* mutations obtained by sequencing trios (see figure 1.2). In the way they are obtained, they are thought to contain a negligible selection component, they accurately represent only a single generation rather than a composite of many generations (as a population does). A handicap of this mutation class is potentially the sparseness, with orders of magnitude smaller sets than cancer genomes. The first studies[321,324] in local hypermutation for *de novo* mutation (DNM) detected a clear enrichment of C>G variants at shorter IMDs suggesting a novel mechanism of mutation accumulation. Further studies in a larger cohort[322] showed that these clusters were coming preferentially from the mother and that they correlated strongly with the mother's age at birth. Interestingly, certain regions of the chr2, 8 and 16 contained hotspots for these mutations. Finally, it was proposed[294] that the mutational process might be related to a DSB-induced mutation mechanism in dormant oocytes that is active during aging. The clustered C>G mutations were co-localized with meiotic gene conversion loci and de novo copy-number. Both the meiotic gene conversion loci and the copy number alterations are associated to the occurrence of DSB, hinting at a potential mechanism.

Early reports that focused in phylogenetic data[325] where they detected template switching events, a type of rearrangement, in highly homologous sequences. This mechanism was responsible for sets of cluster mutations that were previously thought to occur independently of each other but at a low distance.

Overall, the numerous prior evidence presented here shows that mutation clusters are also present within human germline mutations and highlight the role that these can have in shaping human population genome.

1.4.3.2 Mutation clusters in trans interacting sites of the genome

The approximately 2 meter long[m] unidimensional string of DNA is folded inside the nucleus, resulting in proximity interaction also in the three-dimensional space. Because mutational processes result from chemical reactions, the capacity of a mutational process to generate multiple mutations in proximity is not restricted to the one dimensional sequence, but may be able to occur in trans too. For instance, oxidative damage to DNA was reported to occur in clusters in human cells, possibly in relation with deficient DNA repair[327] , and ionizing radiation is widely appreciated to generate clustered DNA damage (reviewed in ref[328]). If DNA damages are clustered, plausibly, the resulting mutations sometimes can be so. We call the hypothesized mutation clusters which occur in proximity but far away from each other in the one dimensional sequence trans-clusters.

As expected, general chromosome folding features of the genome have been described to modulate significantly the local mutation rates in cancers. One example is the position of the chromosomal ‘territory’ in the nuclear space. Chromosome 18 which is relatively closer to the periphery of the nucleus accumulates up to 2 times more mutations than chromosome 19, of similar size but with a more central location[329] . Similar phenomena is observed with regards to Topologically associating domains (TAD). The boundaries between an active and an inactive TAD seem to be markers of the switch in mutation rate for a wide range of signatures[217] although the correspondence between TADs and replication time domains makes it difficult to ascertain a causal role of one or the other. For instance, Lamina associated domain (LAD) are domains that are located at the nuclear periphery and contain heterochromatic regions of the chromosomes, while genic and early replication sections seems to be located centrally[212,329,330] . These various overlapping genomic features are potentially the actual causal elements in the modulation of mutation rate, however because they are so strongly correlated it is difficult to pinpoint the causal ones. Other spatial features have also been reported to participate in the modulation of the damage accumulation in the nucleus. The periphery of the nucleus and LADs in particular tend to accumulate a greater amount of UV damage compared to the central sections[331] suggesting a potential role of these structures also in the modulation of mutation rates. Interestingly, the AID protein, in its physiological mutagenic role, known to target some of the highly active promoters and enhancers, appears to be targeting those that are also high interacting sites in 3D space. These interactions sometimes lead to the AID mutagenic mechanism to cause off-target hits in other expressed parts of the genome[219] .

From the existing literature and to our knowledge, though, there is no actual evidence supporting the existence of mutation trans-clusters as defined in this thesis.

[m] estimate based on 3.3Å per bp[326]

A potential reason for this is that genome-wider spatial genomics data e.g. Hi-C and Micro-C has been available only recently. Moreover, there is an issue with resolutions of current Hi-C studies, which focus at the 5kbp resolution[212–214] , which is relatively coarse compared to mutational data, which is normally obtained in a specific single base resolution. This disparity generates a significant amount of noise that make it challenging to capture robust signal in mutation enrichment and/or clustering.

Another limitation of such studies is the high variability of the interaction of two specific points within the cell population. Although the folding of the genome follows an active mechanism at loop anchors (extrusion by cohesin) , there is still a large amount of coverage which varies from cell to cell within pre-defined domains. Loop anchors, because of its active mechanism, are a good candidate for the detection of trans mutation pairs. A large set of recent studies of 3D genome conformation, using diverse methodologies are available for the study of how spatial organization of the chromosome may result in mutation clustering[212,213,332–334].

In this thesis, we explore the novel concept of mutational trans-clusters by systematically quantifying mutation pair occurrences in loop anchors, and describing potential mechanisms that may generate them. Some of them were anticipated, such as AID mutagenesis, while other mutational signature-like patterns in 3D space were additionally discovered (see chapter 6).

Chapter 2

Objectives

The recent studies in tumor and healthy somatic cell genomes highlight the power of the mutational data available to study the molecular mechanisms of mutagenesis and repair in human cells.

In this thesis, we aim to systematically characterize the patterns of local mutation rate variation, including mutation clusters (as an important example of local hypermutation) and coldspots (local hypomutation), and to apply systematic statistical analyses to uncover their underlying mechanisms.

The specific objectives of this thesis are:

1. The development of new methodology to explore, identify and quantify the local increase in mutation rate in human tumors.
2. The analysis and characterization of molecular mechanisms generating both an increased and decreased local mutation rates.
    (a) The study of various mechanisms of local hypermutation and clustered mutagenesis, focusing on APOBEC mutation patterns.

    (b) The study of mutagenic mechanisms contrasting healthy somatic tissues and tumors.

    (c) The study of local hypomutation across the human genome, mediated by hypomethylated DNA regions.
3. To measure the impact of such locally variable mutagenic mechanisms on the fitness and integrity of the genome.
    (a) To measure the effect of before-mentioned local mutagenesis mechanisms on functional elements, e.g., genic regions and chromatin loop

anchors.

(b) To study the influence of the newly characterized variability into existing methods to infer selection.

Chapter 3

DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers

The following chapter has been selected from the paper:

Mas-Ponte, David, and Fran Supek. ”DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers.” *Nature genetics* 52.9 (2020): 958-968.

The published document can be accessed at:

https://doi.org/10.1038/s41588-020-0674-6

DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers

David Mas-Ponte[1], Fran Supek[1, 2, *]

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain.

[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain.

[*] Correspondence: fran.supek@irbbarcelona.org

Abstract

Certain mutagens, including the APOBEC3 (A3) cytosine deaminase enzymes, can create multiple genetic changes in a single event. Activity of A3s results in striking ‘mutation showers’ occurring near DNA breakpoints, however less is known about mechanisms underlying the majority of A3 mutations. We classified the diverse patterns of clustered mutagenesis in tumor genomes, which identified a novel A3 pattern: nonrecurrent, diffuse hypermutation (*omikli*). This mechanism occurs independently of the known focal hypermutation (*kataegis*), and is associated with activity of the DNA mismatch repair (MMR) pathway, which can provide the single-stranded DNA substrate needed by A3 and contributes to a significant portion of A3 mutations genome-wide. Because MMR is directed towards early-replicating, gene-rich domains, A3 mutagenesis has a high propensity to generate impactful mutations, which exceeds other common carcinogens such as tobacco smoke and UV exposure. Cells direct their DNA repair capacity towards more important genomic regions, thus carcinogens that subvert DNA repair can be remarkably potent.

Introduction

Many types of mutation patterns in somatic cells are linked either with exposure to DNA damaging agents, or with genome instability resulting from failures of DNA repair. Both are causal factors for carcinogenesis due to increases in mutation rates. In addition, dysregulated activity of certain enzymes may be mutagenic. For example, many tumors as well as the human germline bear signatures of error-prone DNA polymerases[1–4]. However, the most striking example of endogenous mutagens is the APOBEC family of cytosine deaminases. They defend against viruses and retrotransposons by damaging their genetic material; additionally, APOBEC1 is an mRNA editing enzyme (reviewed in ref. [5]).

The protein products of *APOBEC3* (A3) paralogs were implicated as mutagens in many human cancer types[6–10]. This is consistent with their ability to deaminate DNA[11,12] when it is single-stranded (ss)[13,14]. Tumors have a highly variable burden of the A3 mutational spectrum, which is associated with differential A3 activity: an activating germline polymorphism in *APOBEC3A* and *APOBEC3B* genes results in a higher mutation burden[15], and there is some correlation thereof with tumoral mRNA expression level of *APOBEC3A* and *APOBEC3B*[4,7,16,17]. In addition to the A3 activity, the availability of its ssDNA substrate is a requirement for mutagenesis. One known source of such ssDNA are intermediates of DNA repair of double-stranded breaks[10,18,19], where A3 results in 'mutation showers' or *kataegis* (greek for thunderstorm), local hypermutation events that may consist of tens of mutations[8,10]. While *kataegis* is striking, it is not common: very few of the A3-signature mutations are accounted by the mutation showers[10,20]. Additionally, DNA secondary structures can generate A3 mutational hotspots [21], however, the processes that generate global, abundant ssDNA substrate for A3 mutagenesis need to be further explored.

Clues are provided by the peculiarities of the A3 mutation patterns. Most mutation types are enriched in late-replicating domains, because DNA mismatch repair (MMR) and possibly nucleotide excision repair are more active in early-replicating domains[22,23]. A3 signature mutations run counter to this trend[20]. Additionally the A3 mutations have a curiously strong DNA replication strand bias[24–26]. These biases, considered together with experimental evidence [27–29], suggest that A3 mutagenic activity is coupled to DNA replication. Expressing A3 enzymes in *E. coli* and yeast produced a mutational bias at replication origins [30,31], suggesting that ssDNA exposed during discontinuous DNA synthesis may be vulnerable to A3. In addition, another source of A3 substrate ssDNA was suggested by experiments in which the repair of a lesion-bearing DNA by base excision repair (BER) and MMR promoted A3 signature mutagenesis in flanking segments[32]. Identifying the mechanisms that allow access of A3s to nuclear DNA is important because A3 enzymes generate cancer driver mutations[21,33–35] and promote tumor heterogeneity[36–38].

Kataegis illustrates how mutation clustering patterns can be used to detect ssDNA generating mechanisms[10,18]. We introduce a sensitive statistical method to detect non-random mutation distribution that results from localized mutagenic events. Applying this to human cancer genomes uncovered a ubiquitous pattern of diffuse A3 mutation clusters, which we named *omikli* (greek: ομίχλη, meaning "fog"). This 'mutation fog', *omikli*, is more common than *kataegis*, however it occurs via a distinct mechanism. We present evidence that the activity of DNA mismatch repair (MMR) promotes A3 mutagenic activity, evident in the *omikli* pattern, and that the same process is responsible for the majority of unclustered A3 mutations. They are surprisingly likely to impact cancer genes – more so than the changes resulting from common external mutagens – because DNA repair directs A3 mutagenesis towards early-replicating, gene rich domains.

Results

Detection of two distinct types of local hypermutation

Our aim was to systematically characterize the different types of mutation clustering in human cancer cells. To this end, we developed a statistical approach (HyperClust) that has two distinguishing features (Fig. 1a; Extended Data Fig. 1a, b). Firstly, it accounts for the heterogeneity of mutation rates and of trinucleotide composition across chromosomal domains, which is an extension of our recent approach[4] with additional support for local false discovery rate (*lfdr*) statistics. Secondly, it draws on the signal present in allelic frequencies of mutations – serving as a proxy for mutation timing – to enforce that mutations constituting one clustered event must occur simultaneously (Methods). We tested these improvements in HyperClust using simulated data with spiked-in mutation clusters, generating precision-recall curves (Extended Data Fig. 1c-e), comparing HyperClust to two previous approaches for detecting clustered mutations [8,10,29]. Our simulation studies suggest that HyperClust compares favorably in calling shorter clusters consisting of two mutations (at various intermutational distance (IMD) distributions, Extended Data Fig. 1e). Therefore our method supports systematic studies of diverse types of clustered mutagenesis.

We used HyperClust to identify clustered somatic single-nucleotide variants in whole-genome sequences of 22 tumor types, detecting a total of 108,401 clustered mutations in 699 tumors (at a *lfdr*≤20%). Henceforth, we defined the A3 spectrum as C>T and C>G changes in a T<u>C</u>W context (W is A or T). Overall 45% of all clustered mutations are in A3 contexts, consistent with A3 enzymes being an important cause of local hypermutation, however 55% of mutation clusters are not in the canonical A3 context, supporting that additional processive agents including error-prone DNA polymerases commonly mutagenize human cells[1–4,39] (we note that A3 may also rarely generate C>A changes[40]). In contrast to prior heuristic rules [29,41,42] that required e.g. at least 5 mutations with an IMD ≤1kb, importantly, the majority of A3 clusters do not meet this definition and instead consist of pairs and triplets (Fig. 1b, c). The distribution of A3 mutation cluster lengths (number of consecutive mutations) was significantly better described by a mixture of two distributions than by a single distribution (Fig. 1d; Extended Data Fig. 1f, g). This suggests that there are at least two types of mutagenesis generating tracts of A3-context changes, which we estimate to have a mean length of 2.2 mutations and 7.1 mutations.

While the latter distribution neatly fits current notions of *kataegis*, the former one does not. We named this type of diffuse mutation clustering *omikli* (fog), by analogy to the focused *kataegis* (thunderstorm) events. Henceforth, we classify mutation clusters with 2, 3 or 4 variants as *omikli* (the short-tract Poisson mixture component predominates; Fig. 1d), and clusters with 5 or more single-nucleotide variants as *kataegis* (with ≥95% contribution of the component with long tracts; Fig. 1d). *Omikli* is ubiquitous, occuring in more tumors (76% tumors contain at least three A3 *omikli* mutations; by random expectation approx. 14% would do so; Fig. 1e) than A3 *kataegis* (48% samples with at least three A3 *kataegis* mutations). In tumors in which they occur, A3 *omikli* are similarly abundant per genome (Q_1-Q_3: 4-36 mutations) as A3 *kataegis* (6-36 mutations; Fig. 1f, Extended Data Fig. 1h).

Distinct mechanisms for *kataegis* and *omikli* A3 mutagenesis

Multiple lines of genomic evidence suggest that A3 *omikli* clusters are generated by a mechanism distinct from *kataegis*. First, *kataegis* is, expectedly [8,10], enriched near rearrangement breakpoints, a proxy for locations of chromosome breaks [43], but not so for *omikli* (Fig. 1g). Second, the burden of A3 *omikli* clusters appears uncoupled from *kataegis* across individual tumors and is weakly correlated (R^2=0.11) with long *kataegis* events (≥8 mutations; Fig. 1h), suggesting that short clusters derive from a different mechanism than the intermediate and long ones, which share a common mechanism (R^2=0.52; Fig. 1h). Third, correlation of A3 mutation burden with APOBEC3A and APOBEC3B mRNA levels is stronger for *omikli* (Spearman rho=0.31 and 0.45, respectively) than for *kataegis* (rho=0.04 and 0.14). This suggests that for *omikli* the A3 expression is commonly limiting, while for *kataegis* another factor becomes limiting, plausibly the source of ssDNA that is available only rarely, e.g. during repair of ds breaks [10,18,44]. Fourth, the 5' mutational context of A3 *omikli* mutations had a significant enrichment of the A3A-like context over the A3B-like context [45] in five cancer types, compared to *kataegis* (Extended Data Fig. 2a-c; the converse was not the case in any cancer type), thus A3A and A3B may have preferential roles in causing *omikli* and *kataegis*, respectively. We also note overall tissue-specific differences A3A-like *versus* A3B-like contexts, as reported [4,45] (Extended Data Fig. 2c). Fifth, the unclustered A3 mutation burden is highly correlated with *omikli* (rho=0.66) but less with *kataegis* (rho=0.27). The numerous unclustered A3 mutations can be seen as a mixture of three components: singletons created by the *omikli* process (henceforth, A3-O), singletons created by the *kataegis* process (A3-K), and the remainder (A3-X) would encompass mutations caused by A3s independently of *kataegis* and *omikli* mechanisms plus the TCW>K mutations not caused by A3s. Consistently, the distribution of the numbers of mutations per cluster in *omikli* (Fig. 1d; >98% are pairs or triplets) suggests that A3-O generates many A3 singletons while A3-K generates few.

Regional distribution of A3 clusters suggests a link to MMR

To gain insight into the process generating *omikli*, we studied its distribution across the genome. A3-context *omikli* mutations were strongly enriched in early-replicating regions (2.0-fold and 2.5-fold for C>T and C>G respectively, Fig. 2a, b), in contrast to unclustered TCW (0.54 and 0.72-fold) and to the control, non-A3 context (VCN, where V is not T; 0.56 and 0.47-fold). These latter enrichments are similar to various other unclustered mutation types (Extended Data Fig. 3a), which are known to be depleted from early-replicating domains[46–48]. Protection of early-replicating domains from mutations stems from the differential activity of DNA mismatch repair (MMR) [4,22,49]. The enrichment of diffuse clustered A3 mutations (*omikli*), uniquely, matches the genomic gradient of increasing MMR activity, rather than that of decreasing MMR activity, as for most other mutation types (this is not explained by the genomic distribution of the TCW trinucleotide; Extended Data Fig. 3b).

MMR is directed towards the regions bearing the H3K36me3 histone mark [50], which is enriched at gene bodies of expressed genes [51,52], lowering their mutation rates [4,53]. Consistently with higher MMR activity, we find a significant enrichment of A3 *omikli* clusters at H3K36me3 regions, after conditioning on replication time and gene expression levels (Fig. 2c; Methods). However, the mRNA level, after conditioning on H3K36me3 and replication time, was not associated with higher A3 *omikli* burden (Fig. 2c). This agrees with prior data [20,31] suggesting that transcription is not a common source of ssDNA substrate for A3 enzymes, even though ssDNA generated during transcription can be prone to mutagenic spontaneous deamination [54]. Regarding A3 *kataegis*, the enrichment in H3K36me3 regions (Extended Data Fig. 3c, d) might stem from recruitment of the homologous recombination machinery (that can generate ssDNA tracts) by this histone mark[55].

We further examined a set of regions proximal to CpG dinucleotides, proposed to be linked with differential MMR activity [56]. There were more A3 *omikli* clusters in the top genomic tertile by CpG density (Extended Data Fig. 3e). Consistently with MMR activity causing the mutations, this difference was more pronounced within early-replicating regions. The mutation rate of the control VCH context in CpG-dense regions was, in contrast, lowered (Extended Data Fig. 3e) [56].

Next, we examined the replication strand bias [24,25] of A3 clusters. The ratio of A3 *omikli* in the leading *versus* the lagging DNA strand closely matched that observed in MMR-deficient (microsatellite instable, MSI) tumors (1.006-fold difference, Fig. 2d), but was less compatible with strand bias associated with mutated proofreading domain of the leading strand-specific DNA polymerase epsilon (POLE, 0.81-fold difference). This suggests that the strand asymmetry of postreplicative MMR activity [57] rather than the asymmetry of DNA replication itself [58] underlies *omikli*; see Supplementary Note.

APOBEC mutagenesis hotspots can occur in DNA sequences that form hairpin secondary structures [21]. Our data do not reflect this: *omikli* after excluding hairpin loci maintained the early replication time enrichment at 2.16-fold.

Coupling of A3 mutagenic mechanisms with DNA replication.

We hypothesized a mechanism by which MMR promotes A3 mutagenesis. MMR generates a single-stranded (ss) DNA intermediate during excision of a mutated DNA segment [59,60]. This provides an opportunity for A3 enzymes to cause DNA damage that converts into clustered mutations, wherein such mutation tracts are short (*omikli*) because the ssDNA segments are short. The widespread occurrence of A3 *omikli* clusters is consistent with most tumors being largely MMR-proficient [61–63]. This is in contrast to *kataegis*, which is known to also stem from DNA repair intermediates, however, these longer segments result from processing of double-strand breaks [10,18,19,40]. The MMR mechanism would explain the enrichment of A3 diffuse clustered mutations in early-replicating domains, and also enrichment in the lagging DNA strand, both associated with higher MMR activity [22,57]. Because MMR is largely replication-coupled [64,65], the MMR-associated A3 mutagenesis is consistent with the greater vulnerability to A3 damage in dividing cells [27].

An additional hypothesis was proposed to explain the associations of A3 mutations with DNA replication-related genomic features [20,47]: ssDNA exposed during discontinuous synthesis of the lagging strand would be mutagenized by A3. This was proposed based on strand-biased mutations that result from expressing human A3s in *Escherichia coli* [30] and in yeast [31]. Because length of eukaryotic Okazaki fragments is known, and length of MMR intermediates has been characterized in eukaryotic systems reconstituted *in vitro*[66,67], we next examined the length distribution of inter-mutational distances (IMD) in the A3 clustered mutations.

The IMD distribution for A3 *omikli* has a global peak at 355 nt, closely matching the peak (378 nt) of a simulated IMD distribution resulting from 800 nt long ssDNA segments (Fig. 2e, Methods). The length of MMR excision tracts was estimated at 800 nt using *in vitro* studies of human and yeast MMR [66,68]. Additionally, we approximated the length of MMR tracts by an analysis of somatic hypermutation events in lymphomid genomes (Methods); this suggested an approx. 400-1000 nt length range (Extended Data Fig. 4a, b). In contrast, the global peak in *omikli* IMD was not compatible with the approx. 200 nt long Okazaki fragments [67], which would generate a peak at 96 nt (Fig. 2e). (Of note, in *kataegis* events, IMD are devoid of the peak corresponding to ~800 nt length tracts (Fig. 2e), thus *kataegis* would result independently of MMR). These data suggest that discontinuous lagging strand synthesis is not the main mechanism supplying ssDNA that yields A3 clustered mutations because the observed IMDs are too long. However the IMDs are compatible with MMR-supplied ssDNA. Moreover, the proposed mechanism agrees with the early replication time enrichment of A3 *omikli*, which is consistent with higher MMR activity.

We do not exclude however that the discontinuous synthesis of the lagging strand contributes to A3 mutagenesis because the *omikli* IMD distribution has a secondary peak corresponding to 200 nt segment lengths (Fig. 2e). Modelling the IMD as a mixture of gamma distributions (Fig. 2f) suggests that up to one-quarter of A3 clusters might be generated by a process corresponding to ~200 nt long segments (Extended Data Fig. 4c, d). Notably, the mixture modelling also suggests a minor component in *omikli* IMD at very short peak lengths (~25 nt, Fig. 2f). It is tempting to speculate that this reflects the binding of the ssDNA protective protein RPA, which has a 24-30 nt footprint [69,70]. A secondary IMD peak of this length is observed also in *kataegis* (Fig. 2e; see Methods for limitations of use of IMD measure for *kataegis* analyses).

MMR deficiencies are associated with lower A3 mutagenesis

We next examined the tumors exhibiting microsatellite instability (MSI), which are MMR deficient; we took care to adjust for different statistical power to detect clusters in these high mutation burden tumors (Extended Data Fig. 4e, f) making the following analyses conservative.

We compared the fraction of A3 *omikli* mutations in MSI and microsatellite stable (MSS, MMR-proficient) tumors of the matched cancer types (Fig. 3a). Supporting our hypothesis, the fraction of A3 *omikli* clusters in the MSI samples was significantly lower than in the MSS tumors (p<0.001 by Mann-Whitney test; 5.52-fold difference between the median of samples), but there was no significant difference in the non-A3-context (V<u>C</u>N>K) clusters (p=0.34, 1.2-fold difference; Fig. 3a). Of note, comparing absolute, i.e. not normalized to overall number of mutations, *omikli* A3 burdens were also lower in MSI (p<0.01, Extended Data Fig. 4g). Therefore, the depletion of A3 clusters is in contrast with the overall increase of mutation load in MSI tumors: MMR normally protects against many types of mutations but provides an opportunity for A3. The MSI-MSS difference is consistently observed across three cancer types (4.0, 3.7 and 12.1-fold enrichment of A3 *omikli* in MMR proficient MSS tumors, Fig. 3a) and the overall difference is significant after stratifying by cancer type (Fig. 3b, pooled p<0.001, Fisher's method for combining p-values).

The early replication enrichment of *omikli* is not observed in MSI (Fig. 3c), but instead a profile more similar to unclustered mutations is seen, further supporting that MMR directs the A3 mutagenesis. Consistently, A3 *omikli* burden associates with expression levels and copy number status of MMR genes *MSH6*, *MSH2* and *EXO1* (Fig. 3d, e; Extended Data Fig. 3f, g; discussed in Supplementary Note).

We have further validated findings on an independent set of 2,304 tumor whole genome sequences (WGS, Methods). This supported the dichotomy between A3 *kataegis* and *omikli* clustering in tract lengths (Extended Data Fig. 5a-c). The key evidence that links A3 mutagenesis to MMR activity validates: there is a strongly increased A3 *omikli* fraction in MSS *versus* MSI cancers, in a data set stratified by cancer type, here also including additional tissues such as prostate and breast; this difference is however modest in the control, non-A3 context (Extended Data Fig. 5d, e). Moreover, additional supporting evidence of MMR involvement validates in these data: significantly increased A3 *omikli* burdens in tumors with copy number gains in *MSH6* and *MSH2* and *EXO1* genes (Extended Data Fig. 5f), and the altered regional distribution of A3 *omikli* between MSS (enriched in early-replicating) and MSI cancers (less enriched) (Extended Data Fig. 5g). The IMD distributions of A3 *omikli* similarly have a peak corresponding to approx. 800 nt long vulnerable DNA segments (Fig. 2e; Extended Data Fig. 5h). Finally, an analysis of >3,000 whole-exome sequences showed a 3.02-fold excess of nearby T<u>C</u>W mutation pairs (within 1 kb), compared to more distant TCW pairs, in MSS over MSI samples; we also note the overall differences in T<u>C</u>W mutation burden in MSS *versus* MSI (Extended Data Fig. 5i, j). This further supports the association between A3 local hypermutation and MMR activity, which – as suggested by our IMD analysis – may stem from the ssDNA excision tracts generated during MMR. However other molecular mechanisms may similarly be able to explain the MMR-associated A3 mutagenesis, such as changes in replication fork dynamics.

Contribution towards the global A3 mutation burden

While *kataegis* and *omikli* clusters are informative markers of certain mutational processes, their numbers are low. We quantified the contribution of the two clustered A3 processes to the (much more abundant) unclustered mutational burden using a regression analysis, similar to ref. [4]; see Methods. Informally, a correlation between clustered burden of tumor samples and unclustered burden in the same mutational context suggests that the same process underlies the clustered and unclustered component (Fig. 4a shows A3 *omikli* and *kataegis* fits for lung adenocarcinoma; the former is a good fit, while the latter a poor one).

In the pan-cancer data, we estimated that the *omikli* process contributes approximately two-thirds of all A3 context mutations (A3-O, 66.4%, Fig. 4b), while the *kataegis* contribution is negligible (A3-K, ~0%) and an unknown process (or a mix thereof) contributes the remaining nearly one-third of A3 context mutations (A3-X, 32.4%; Fig. 4b). The lack of *kataegis* contribution is not unexpected, given that this process generates long tracts but almost never pairs or triplets (Fig. 1d) and thus by extension singletons would not be generated. The presence of mutations originating from the A3-X process, which is not associated with *omikli* and thus likely independent of MMR, suggests that the MMR hypothesis is one of the possible explanations for the mechanisms that generate the global pool of ssDNA vulnerable to A3.

We also considered cancer types individually (Extended Data Fig. 6), showing that the relative contribution of A3-O was strongly correlated with the absolute A3 mutation burden across cancer types (Fig. 4c). This further supported that a MMR-dependant, likely A3A-driven process which can be diagnosed via *omikli* is the major source of APOBEC mutagenesis in human cancer. This creates very high A3 mutation burdens in lung, breast, bladder and head-and-neck cancers (Fig. 4c), while other cancer types such as prostate – even though *kataegis* is known to occur therein – exhibit less *omikli* and lower overall A3 mutation burdens.

A3 mutagenesis has a high functional impact per mutation

Certain mutational processes – including A3 activity, MMR failures and use of translesion DNA polymerases – were reported to, atypically, produce many mutations in early-replicating, gene-rich chromosomal domains [4,26]. Such 'mutation redistribution' [71] means that at an equal global mutation burden, different mutagens may have different potential for affecting genes, thus having varied functional consequences. To quantify this, we introduce a concept of 'functional impact density' (FID) of a mutational process: the fraction of putatively impactful mutations among all mutations observed.

In case of cancer, a simple estimate of the oncogenic FID is the fraction of changes affecting coding regions of known cancer genes ('oncogenic mutations per thousand', henceforth OMPK; Methods). This is based on the reasonable assumption that many mutations occurring in a typical cancer gene are oncogenic and also that the set of 299 frequently mutated cancer genes [72] contains many of the driver mutations found in a tumor.

We examined the oncogenic FID of A3-O and A3-K mutations, as estimated from total A3 burden in tumors that harbor predominantly *omikli* or predominantly *kataegis* clusters (Methods). This was compared to common mutagenic processes[6] associated with tobacco smoking (C>A in lung), UV exposure (C>T in skin), exposure to gastric acid (A>C in stomach) and finally with aging (C>T changes at CpG dinucleotides). A3 mutations derived either from *omikli* or from *kataegis* processes have very high oncogenic FID: 0.47 and 0.46 OMPK, respectively (Fig. 5a, Methods), approximately twice that of common external mutagens: tobacco smoking and stomach acid-associated mutations, both at 0.24 OMPK, and of UV at 0.19 OMPK.

In addition to A3, another endogenous mutagenic process – the aging-associated C>T changes at CpG dinucleotides – also had high oncogenic FID per mutation (Fig. 5a). This is in line with a high frequency of CpG dinucleotides in coding regions in the human genome (Extended Data Fig. 7a); consistently, aging-related mutagenesis was suggested to have a higher risk of generating coding mutations than cancer chemotherapeutics did[73]. Of note, the A3 TCW context is not markedly enriched in coding regions so the high FID of A3 mutations is irrespective of trinucleotide composition therein.

We asked if the high FID of A3 mutagenesis stems from increased positive selection on oncogenic changes introduced by A3. Using intronic mutation rates as a baseline[74] (Methods), we find that selection on A3 mutations is not stronger than on external mutagen-induced changes (Extended Data Fig. 7b), which agrees with recent reports [33].

Instead, we hypothesized the higher FID of A3 results from the increased susceptibility of the affected genes to DNA repair as they are more often located in early-replicating euchromatic domains [22,23,25,75] than intergenic regions are. The high intronic/intergenic ratio shows that A3 mutagenesis is strongly redistributed towards genic DNA, compared to the various external mutagens (Extended Data Fig. 7b). The difference of FID of A3 processes *versus* external mutagens is exaggerated in cancer genes that reside in early-replicating regions (Extended Data Fig. 7c). This suggests that the *omikli*-driven A3 mutations are impactful due to an enrichment in gene-dense, early replicating domains, which are protected from many other mutation types. In addition to cancer genes, because somatic mutations might play a role in aging and neurodegeneration [76,77], we also examined a set of known essential genes, and a set of genes linked with neurodegeneration (Methods). Overall, we observed very similar results, with FID increases of A3 over the external mutagens ranging from 2 to 11-fold (Extended Data Fig. 7d, e).

A3 mutagenesis affects genes encoding chromatin modifiers

FID is a measure of the relative impact of a mutational process (expressed per mutation), however the absolute mutational burden of a process also needs to be considered. While tobacco smoking and UV mutations are less impactful, they are abundant. Aging-associated mutations are impactful per mutation but lowly abundant. The two A3 processes are however both impactful and abundant (Fig. 5a; error bars show variation across those tumors that were affected by a mutagenic process).

The absolute mutation burden strongly differentiates the *omikli* from the *kataegis* mutagenesis (A3-O and A3-K, respectively) even though their FID is similar. We estimate that the MMR-associated *omikli* process can generate, in tumors where it is highly active, approximately twice as many mutations with oncogenic potential (2.72 per tumor) than the DNA break repair-mediated *kataegis* process (1.32 per tumor) on average. Moreover, *omikli* generates twice as many oncogenic mutations as the aging-associated CpG mutagenesis. Notably, the A3 *omikli* process generates a comparable number of putatively oncogenic mutations per sample as the tobacco smoking (2.14 per tumor, in smokers' lung adenocarcinoma) and UV light (3.54 per tumor, in melanoma). This suggests that A3–considering jointly the (major) *omikli* and the (minor) *kataegis* components – may be an important carcinogen because, in exposed cells, it is able to create larger numbers of mutations in cancer genes than common external mutagens.

We observed a significant association between *omikli* burden and mutation occurrence (Methods) in 22 cancer genes at FDR<5%, and in 30 at FDR<10% (of 61 testable genes with ≥3 TCW>K coding mutations in our data; Fig. 5b; Supplementary Table 1). However, no genes were significantly associated with *kataegis* burden (Extended Data Fig. 8a), supporting that *omikli* is more oncogenic than *kataegis*. The genes linked with *omikli* are enriched in tumor suppressors (n=14, *versus* 5 oncogenes; Fig. 5c) and are commonly chromatin modifiers (e.g. K*MT2A/C/D*, *NCOR1*, *SETD2*, *MECOM*) or chromatin remodelers (e.g. *PBRM1*, *ARID2*) (Fig. 5c) which have a higher count of TCW motifs in the coding sequence (Extended Data Fig. 8b). These associations do not however show the direction of the effect. We thus examined the control VCN mutations, which were significantly associated in only 3 genes (Fig. 5b; Extended Data Fig. 8c). This suggests that the MMR-mediated A3 mutagenic pathway is an important source of cancer driver events. Consistently, cancer gene mutations in early-replicating regions are more strongly associated with overall *omikli* burden than those in late replicating regions (Extended Data Fig. 8d).

Discussion

Clustered mutations, even though rare, can occur in different types of clustering patterns, which serve as markers of different mutagenic processes. *Kataegis* originates from repair of double-stranded DNA breaks by the homologous recombination or break-induced replication pathways, which expose long tracts of ssDNA [18,40,78]. Here we propose that another DNA repair pathway – MMR –promotes A3 mutagenesis, generating *omikli* clusters and the bulk of A3 unclustered context mutations in human tumors. A different link of A3 with DNA repair was proposed recently, resulting from DNA lesions processed by the base excision repair (BER) pathway (abasic sites, uracils, or T:G mismatches), which generated A3-context mutations flanking the repaired site [32]. MMR was suggested to be able to 'hijack' the BER intermediates to provide additional ssDNA substrate for A3 [32]. Our data suggest that MMR may generate A3 substrate ssDNA more generally, which could occur by processing mismatches occurring during DNA replication. We do not exclude that BER-processed lesions result in A3 mutagenesis in cancer; indeed this may help explain the approximately one-third of the unclustered A3 mutations (A3-X) that we do not account for via *omikli*. Another likely contributor to this MMR-independent A3 mutation fraction is A3 activity at ssDNA occurring discontinuous synthesis of the lagging strand in DNA replication**[24,25,30,31]**, which finds some support in our IMD distribution analyses.

MMR activity preferentially protects early-replicating, euchromatic regions from mutations [22,79,80] and additionally transcribed gene bodies therein, because it is recruited by the H3K36me3 histone mark [4,53]. Therefore, mutagenic processes that subvert MMR would be particularly dangerous because they are directed to active genes. One example of this is non-canonical MMR that recruits the error-prone DNA polymerase η (POLH protein) [81,82], whose mutational signatures are seen across human tumors [2,4]. Here we provide another example of MMR activity leading to mutagenesis, in this case by promoting APOBEC activity. Based on the enrichment of MMR-associated A3-context mutations in early-replicating gene-rich chromosome domains, we propose that the MMR-A3A coupling has particularly high potential for generating impactful mutations, exceeding common exogenous mutagens. In addition to oncogenes and tumor suppressor genes, A3-context mutations were directed towards essential genes and neurological disease-associated genes, suggesting possible roles for APOBEC mutagenesis not only in cancer, but also more generally in aging-related pathologies.

Acknowledgments

We thank the members of the Genome Data Science group and Ben Lehner for comments and discussions. This work was funded by the ERC Starting Grant HYPER-INSIGHT (757700) and by the Spanish Ministry of Economy and Competitiveness (grant no. BFU2017-89833-P). F.S. was funded by the ICREA Research Professor programme. D.M.P. was funded by a Severo Ochoa FPI fellowship (MCIU/Fondo Social Europeo ref. BES-2017-079820). The authors acknowledge support of the Severo Ochoa Centres of Excellence programme to the IRB Barcelona.

The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This publication and the underlying research are partly facilitated by Hartwig Medical Foundation and the Center for Personalized Cancer Treatment (CPCT) which have generated, analysed and made available data for this research.

Author contributions

F.S. and D.M.P. conceptualized the study and devised the methodology. D.M.P. carried out the formal analysis, the investigation, operated the software and performed data visualization. D.M.P. and F.S. wrote and edited the draft manuscripts. F.S. acquired the funding and supervised the study.

Competing interests

The authors declare no competing interests.

References for main text

1. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res.* **24**, 1445–1454 (2014).
2. Rogozin, I. B. *et al.* DNA polymerase η mutational signatures are found in a variety of different types of cancer. *Cell Cycle* **17**, 1–31 (2018).
3. Seplyarskiy, V. B. *et al.* Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nat. Genet.* **51**, 36 (2019).
4. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534-547.e23 (2017).
5. Moris, A., Murray, S. & Cardinaud, S. AID and APOBECs span the gap between innate and adaptive immunity. *Front. Microbiol.* **5**, 534 (2014).
6. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
7. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013).
8. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
9. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
10. Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
11. Landry, S., Narvaiza, I., Linfesty, D. C. & Weitzman, M. D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO Rep.* **12**, 444–450 (2011).
12. Suspène, R. *et al.* Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc. Natl. Acad. Sci.* **108**, 4858–4863 (2011).

13. Byeon, I.-J. L. *et al.* NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. *Nat. Commun.* **4**, 1890 (2013).

14. Holtz, C. M., Sadler, H. A. & Mansky, L. M. APOBEC3G cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure. *Nucleic Acids Res.* **41**, 6139–6148 (2013).

15. Nik-Zainal, S. *et al.* Association of a germline copy number polymorphism of *APOBEC3A* and *APOBEC3B* with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.* **46**, 487–491 (2014).

16. Glaser, A. P. *et al.* APOBEC-mediated mutagenesis in urothelial carcinoma is associated with improved survival, mutations in DNA damage response genes, and immune response. *Oncotarget* **9**, 4537–4548 (2017).

17. Cortez, L. M. *et al.* APOBEC3A is a prominent cytidine deaminase in breast cancer. *PLOS Genet.* **15**, e1008545 (2019).

18. Sakofsky, C. J. *et al.* Break-induced replication is a source of mutation clusters underlying kataegis. *Cell Rep.* **7**, 1640–1648 (2014).

19. Sakofsky, C. J. *et al.* Repair of multiple simultaneous double-strand breaks causes bursts of genome-wide clustered hypermutation. *PLOS Biol.* **17**, e3000464 (2019).

20. Kazanov, M. D. *et al.* APOBEC-induced cancer mutations are uniquely enriched in early-replicating, gene-dense, and active chromatin regions. *Cell Rep.* **13**, 1103–1109 (2015).

21. Buisson, R. *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaaw2872 (2019).

22. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).

23. Zheng, C. L. *et al.* Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. *Cell Rep.* **9**, 1228–1234 (2014).

24. Haradhvala, N. J. *et al.* Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).

25. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).

26. Seplyarskiy, V. B. *et al.* APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* **26**, 174–182 (2016).

27. Green, A. M. *et al.* APOBEC3A damages the cellular genome during DNA replication. *Cell Cycle* **15**, 998–1008 (2016).

28. Kanu, N. *et al.* DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. *Genome Biol.* **17**, 185 (2016).

29. Nikkilä, J. *et al.* Elevated APOBEC3B expression drives a kataegic-like mutation signature and replication stress-related therapeutic vulnerabilities in p53-defective cells. *Br. J. Cancer* **117**, 113–123 (2017).

30. Bhagwat, A. S. *et al.* Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **113**, 2176–2181 (2016).

31. Hoopes, J. I. *et al.* APOBEC3A and APOBEC3B preferentially deaminate the lagging strand template during DNA replication. *Cell Rep.* **14**, 1273–1282 (2016).

32. Chen, J., Miller, B. F. & Furano, A. V. Repair of naturally occurring mismatches can induce mutations in flanking DNA. *eLife* **3**, e02001 (2014).

33. Cannataro, V. L. *et al.* APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma. *Oncogene* **38**, 3475 (2019).

34. Henderson, S., Chakravarthy, A., Su, X., Boshoff, C. & Fenton, T. R. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep.* **7**, 1833–1841 (2014).

35. Li, Z. *et al.* APOBEC signature mutation generates an oncogenic enhancer that drives *LMO1* expression in T-ALL. *Leukemia* **31**, 2057–2064 (2017).

36. Bruin, E. C. de *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).

37. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.* **7**, 283ra54-283ra54 (2015).

38. Ullah, I. *et al.* Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes. *J. Clin. Invest.* **128**, 1355–1370 (2018).

39. Reijns, M. A. M. *et al.* Lagging strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502–506 (2015).

40. Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* **2**, e00534 (2013).

41. D'Antonio, M., Tamayo, P., Mesirov, J. P. & Frazer, K. A. Kataegis expression signature in breast cancer is associated with late onset, better prognosis, and higher HER2 levels. *Cell Rep.* **16**, 672–683 (2016).

42. Petljak, M. *et al.* Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282-1294.e20 (2019).

43. Zhang, Y. *et al.* A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. *Cell Rep.* **24**, 515–527 (2018).

44. Yang, Y., Sterling, J., Storici, F., Resnick, M. A. & Gordenin, D. A. Hypermutability of damaged single-strand dna formed at double-strand breaks and uncapped telomeres in yeast *Saccharomyces cerevisiae*. *PLOS Genet.* **4**, e1000264 (2008).

45. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, ng.3378 (2015).

46. De, S. & Michor, F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.* **29**, 1103–1108 (2011).

47. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).

48. Woo, Y. H. & Li, W.-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.* **3**, 1004 (2012).

49. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1–16 (2018).

50. Li, F. *et al.* The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell* **153**, 590–600 (2013).

51. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).

52. Vavouri, T. & Lehner, B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome Biol.* **13**, 1–12 (2012).

53. Huang, Y., Gu, L. & Li, G.-M. H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. *J. Biol. Chem.* **293**, 7811–7823 (2018).

54. Mugal, C. F., von Grünberg, H.-H. & Peifer, M. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol. Biol. Evol.* **26**, 131–142 (2009).

55. Pfister, S. X. *et al.* SETD2-dependent histone H3K36 trimethylation is required for homologous recombination repair and genome stability. *Cell Rep.* **7**, 2006–2018 (2014).

56. Chen, J. & Furano, A. V. Breaking bad: The mutagenic effect of DNA repair. *DNA Repair* **32**, 43–51 (2015).

57. Andrianova, M. A., Bazykin, G. A., Nikolaev, S. I. & Seplyarskiy, V. B. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. *Genome Res.* **27**, 1336–1343 (2017).

58. Shinbrot, E. *et al.* Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res.* **24**, 1740–1750 (2014).

59. Jiricny, J. The multifaceted mismatch-repair system. *Nat. Rev. Mol. Cell Biol.* **7**, 335–346 (2006).

60. Tran, P. T., Erdeniz, N., Symington, L. S. & Liskay, R. M. EXO1-A multi-tasking eukaryotic nuclease. *DNA Repair* **3**, 1549–1559 (2004).

61. Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 15180 (2017).

62. 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).

63. Maruvka, Y. E. *et al.* Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nat. Biotechnol.* **35**, 951–959 (2017).

64. Hombauer, H., Srivatsan, A., Putnam, C. D. & Kolodner, R. D. Mismatch repair, but not heteroduplex rejection, is temporally coupled to DNA replication. *Science* **334**, 1713–1716 (2011).

65. Hombauer, H., Campbell, C. S., Smith, C. E., Desai, A. & Kolodner, R. D. Visualization of eukaryotic DNA mismatch repair reveals distinct recognition and repair intermediates. *Cell* **147**, 1040–1053 (2011).

66. Jeon, Y. *et al.* Dynamic control of strand excision during human DNA mismatch repair. *Proc. Natl. Acad. Sci.* **113**, 3281–3286 (2016).

67. Smith, D. J. & Whitehouse, I. Intrinsic coupling of lagging-strand synthesis to chromatin assembly. *Nature* **483**, 434–438 (2012).

68. Bowen, N. *et al.* Reconstitution of long and short patch mismatch repair reactions using *Saccharomyces cerevisiae* proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18472–18477 (2013).

69. Brosey, C. A. *et al.* A new structural framework for integrating replication protein A into DNA processing machinery. *Nucleic Acids Res.* **41**, 2313–2327 (2013).

70. Fan, J. & Pavletich, N. P. Structure and conformational change of a replication protein A heterotrimer bound to ssDNA. *Genes Dev.* **26**, 2337–2347 (2012).

71. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* **81**, 102647 (2019).

72. Bailey, M. H. *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371-385.e18 (2018).

73. Pich, O. *et al.* The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).

74. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).

75. Drost, J. *et al.* Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. *Science* **358**, 234–238 (2017).

76. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).

77. Verheijen, B. M., Vermulst, M. & van Leeuwen, F. W. Somatic mutations in neurons during aging and neurodegeneration. *Acta Neuropathol. (Berl.)* **135**, 811–826 (2018).

78. Lei, L. *et al.* APOBEC3 induces mutations during repair of CRISPR–Cas9-generated DNA breaks. *Nat. Struct. Mol. Biol.* **25**, 45 (2018).

79. Belfield, E. J. *et al.* DNA mismatch repair preferentially protects genes from mutation. *Genome Res.* (2017) doi:10.1101/gr.219303.116.

80. Lujan, S. A. *et al.* Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res.* **24**, 1751–1764 (2014).

81. Peña-Diaz, J. *et al.* Noncanonical mismatch repair as a source of genomic instability in human cells. *Mol. Cell* **47**, 669–680 (2012).

82. Zlatanou, A. *et al.* The hMSH2-hMSH6 complex acts in concert with monoubiquitinated PCNA and pol η in response to oxidative DNA damage in human cells. *Mol. Cell* **43**, 649–662 (2011).

Figure legends for main text

Figure 1. Two types of local hypermutation in human tumors. a, The HyperClust framework detects mutation clustering by accounting for heterogeneous mutation rates at the megabase scale, further stratifying mutations by type, and additionally by their approximate timing (clonal fraction). **b**, *Kataegis* (thunderstorm) and *omikli* (fog) mutation clusters in an example tumor genome segment (chromosome 8 of TCGA-DK-A1A6). Vertical lines are rearrangement loci. **c**, Distribution of the number of A3-context T<u>C</u>W>K mutations in *omikli* (bottom) and *kataegis* (top) of different sizes (number of mutations per cluster; callouts). **d**, Poisson mixture modelling of number of A3 context mutations per cluster. Solution with two distributions is shown (*kataegis*, teal and *omikli*, orange). Stacked bars show component proportions and curves are density estimates. Grey curve is the baseline solution with one component; p-values are from a two-sided bootstrap test; LL, log likelihood. **e**, Cumulative percentage of tumor samples that contain at least the given number of clustered mutations, either observed, or expected at random. **f**, Distribution of the burden of A3 context somatic mutations per tumor, across tumors; samples with no *omikli* or no *kataegis* mutations were not considered. **g**, Cumulative fraction of A3 mutations within the neighborhood (width on X-axis) of a rearrangement breakpoint. Error bars are 95% binomial C.I.; number of mutations listed in parenthesis. **h**, Pearson correlation between the burden of two-mutation *omikli* and of long *kataegis* events (left) and the correlation between burden of *kataegis* of different lengths (right). Significant difference by a two-tailed t-test on the Fisher-transformed correlation coefficients.

Figure 2. Association of A3 clustered mutation density with genomic features. a, Mutation rates in replication time (RT) quartiles, relative to the latest RT, for A3 mutation contexts (top) and control contexts (bottom). **b**, Mutation enrichment in the earliest *versus* latest RT quartile for A3 context clusters (top) and non-A3 context clusters (bottom). Cancer types are ordered by total A3 burden across all tumors (shading in top bar). Moderate/low-A3 burden cancer types are pooled into the group “other”. **c**, Relative density of A3 and non-A3 mutation types across genomic regions. All enrichments are relative to the lowest bin (the latest-replicating quartile for RT), which is not shown on figure. Points are coefficients from negative binomial regression, and error bars are 95% C.I. **d**, Replication strand bias (ratio of mutation count on the leading *versus* lagging DNA strand) of clustered T<u>C</u>W mutations. Error bars are binomial 95% C.I. As a control, the reciprocal of the strand bias for MSI-H (orange; 24 samples) and POLE-mutant (purple; 9 samples) tumors is shown as a dashed line. Values in parentheses are mutation counts used to estimate the ratios. **e**, Distributions of intermutation distances (IMD) in A3 context *kataegis* and *omikli* clusters (left). Expected IMD distributions from simulations using three different segment lengths (right). **f**, Gamma mixture modeling of the *omikli* IMD distribution using three components. Bar shows proportions of the three components, while curves show their densities at various IMDs.

Figure 3. MMR activity in tumors is associated with APOBEC mutagenesis. a, Proportion of *omikli* clusters in A3 (left) and control non-A3 contexts (right), comparing MMR deficient (MSI-H) samples with MMR-proficient (MSS) samples, in matched tissues (“MSI tissues”, COAD, STAD and UCEC, green) or in non-matched tissues (red). Significance by Mann-Whitney test, two-tailed; p < 0.001 (***); number of tumor samples listed in parenthesis. **b**, Same as (a) but broken down by tissue. UCEC, uterus; STAD, stomach; COAD, colon. Pooled

p-value (p < 0.001 for A3; p = 0.433 for control) from two-tailed Mann Whitney tests on stratified data. **c**, Enrichment of A3 *omikli* clusters and unclustered A3 mutations in various genome regions in MMR-deficient samples (MSI-H). Related to Fig. 2c. Coefficients of negative binomial regression are shown (as $\log_2$), indicating enrichments of mutation frequency in a genomic bin *versus* the lowest bin (in case of RT, latest-replicating), where enrichment would equal unity and is thus not shown. Error bars are 95% C.I. **d**, Correlation of the burden of A3-context (T$\underline{\text{C}}$W>K) *kataegis*, *omikli*, and unclustered mutations with mRNA levels of MMR genes and of *APOBEC3A* and *APOBEC3B* genes. Error bars are 95% C.I. **e**, Association of copy number alterations (CNA) in selected MMR genes with burden of A3 *omikli*. CNAs are represented as integer copy number differences (Methods); positive values are gains and negative losses. See also Extended Data Fig. 3g. Significance by Mann-Whitney test (two-tailed) comparing the neutral (0) *versus* the gain (+1 and +2) states considered jointly.

Figure 4. The *omikli* process generates the majority of unclustered A3 mutations across tissues. **a**, A regression analysis estimates the contributions of *omikli* and *kataegis* processes towards the unclustered A3 mutation burden, shown for lung adenocarcinoma (LUAD, other cancers in Extended Data Fig. 6) tumor samples (points). For clarity, data panels show combinations of two variables (*omikli versus* unclustered, center; *kataegis versus* unclustered, right), whereas the regression is performed on the three variables simultaneously (schematic in leftmost panel; Methods). Red line is the intersection of the fitted plane with the shown two-dimensional coordinate system. Error bars are 95% prediction intervals of the fit. Dotted line is the average of *omikli* (center) and *kataegis* (right) mutation burden across tumors. Bottom panels have same data as top panels, but zoomed in on the X-axis for clarity. **b**, Pan-cancer regression analysis provides estimates of the fraction of unclustered T$\underline{\text{C}}$W>K mutations contributed by processes that generate *omikli* (A3-O), that generate *kataegis* (A3-K) and a remainder (“intercept”) not explained by either process (A3-X). Error bars are standard errors (S.E.) of regression coefficients; n = 646 tumors. **c**, Relative contribution of the *omikli*-process to the unclustered A3 burden (Y-axis) of cancer types correlates with the overall burden of A3 mutations in that cancer type (X-axis) suggesting that differential activity of the *omikli* mechanism drives differences of A3 burden between tissues. Error bars are S.E. of regression coefficients. Shaded band is 95% C.I. of the linear fit.

Figure 5. APOBEC mutagenesis generates many impactful mutations. **a**, Functional impact density of mutational processes (slope of line), estimated as the number of mutations in coding regions of 299 cancer genes (Y-axis) normalized to the total mutation tally contributed by a process (X-axis). Bottom panel shows the number of mutations estimated to result from each process across tumor samples. Points in boxplots (lower panel) and on lines (upper panel) are the average mutation burden of that process in the affected samples (definition in Methods); error bars are S.E.M. **b**, Occurrence of A3 context mutations in many cancer genes is associated with the genomic burden of A3 *omikli* mutation clusters, suggesting that the *omikli* process generates driver mutations. FDRs are Benjamini-Hochberg adjusted p-values from a logistic regression to predict presence of a T$\underline{\text{C}}$W>K (A3 context, X-axis) or a V$\underline{\text{C}}$N>K (control non-A3 context, Y-axis) mutation in each driver gene. Red and gold, hits at stringent (5%) and permissive (10%) FDR thresholds in the A3 context; blue, hits in the control context (FDR <

5%) suggesting an indirect association with A3 *omikli* burden. Diagonal line denotes equal FDR between the A3 and the control contexts. FDRs were capped at 0.1%. **c**, Burden of A3 *omikli* mutations in tumors which are *wild-type* (teal) or which are mutated (orange) in the driver genes that were significantly associated in the logistic regression in panel **b**.

Online methods

Data sources

Mutation calls for TCGA-WGS were obtained as in ref. [22]. In brief, BAM files were downloaded from the cgHub repository (now superseded by the NCI Genomic Data Commons) for normal and tumor samples, and somatic single-nucleotide variants were called with Strelka 1.0.6 [83]. Also as previously [4,22] we excluded mutations in blacklisted regions by UCSC (Duke and DAC) and in difficult-to-align genomic regions by the 'CRG Alignability 36' criterion, meaning we required genomic 36-mers to be unique in the hg19 genome assembly (even after allowing up to two mismatches).

SNP6 Affymerix microarray data were downloaded from the GDC legacy portal (portal.gdc.cancer.gov/legacy-archive) for matched donors, with both normal and tumor data available. The final dataset contained 699 TCGA samples with WGS mutations and SNP6 array data available. One of the donors (TCGA-CZ-5454) was excluded from those analyses that required external metadata as two different aliquots were available and metadata could not be unambiguously matched. This change makes the number of total samples equal to 697 in some analyses.

MSI status and other metadata for hypermutated tumors (i.e. POLE status) was obtained as described in ref. [22]. In total, our TCGA-WGS dataset contained 24 MSI samples (Supplementary Table 2).

An additional dataset, comprising WGS single nucleotide variants, purity estimates, and copy number alterations was obtained from the Hartwig Medical Foundation[84], was used for validation analyses in Extended Data Fig. 5a-h. This dataset has been processed similarly to our TCGA WGS (Strelka version 1.0.14 was used to call single-nucleotide variants) and additionally the Purple tool was used to infer purity and obtain CNA estimates[84] (Supplementary Table 3).

Inferred MSI/MSS labels[85] were obtained from the supplementary data of the corresponding publication[84]. We additionally discarded samples (n = 53) that were treated with temozolomide (TMZ), which is known to positively select for MMR deficient cells in brain tumors[86].

For the functional impact of UV mutations we additionally obtained WGS variant calls of 70 melanomas tumors from the MELA-AU study [87] within PCAWG. For the somatic hypermutation analyses, we additionally obtained WGS variant calls of blood tumors CLLE-ES and MALY-DE from the PCAWG dataset[88] available as controlled files in the ICGC data portal (https://dcc.icgc.org/pcawg). We selected the SANGER pipeline calls (Supplementary Table 4).

We obtained exonic mutations from the TCGA mc3 dataset, available at (https://gdc.cancer.gov/about-data/publications/mc3-2017)[89]. This dataset contains unified somatic mutation calls for approximately 10,000 whole-exome sequences (WES). We selected cancer types that had at least one sample classified as MSI (see below), therefore the subset used in this analysis comprised 5,831 tumors from 16 cancer types. Only 6% of the WES samples overlap with the WGS cohort. We obtained the MSI status from ref. [61], which contains experimentally determined MSI labels (for ESCA, UCEC, COAD, READ and STAD) and additionally inferred MSI status labels at 80% confidence level that covered additionally 11 cancer types (Supplementary Table 5).

The acronyms used for cancer types in this analysis are as listed in the ICGC Project portal page (https://docs.icgc.org/submission/projects/).

HyperClust, a randomization-based FDR estimation for local hypermutation detection.

The process of detecting local hypermutation (or mutation clusters) aims to distinguish those pairs of mutations that occurred in the same event from those that occurred independently. The classification is based primarily on intermutational distances (IMD) on the genomic sequence but other sources of information can be used such as the allelic fraction of the mutations.

We developed HyperClust building upon our recent approach[4] which employs a trinucleotide context-preserving randomization of mutations within megabase-sized chromosomal domains, obtaining a baseline frequency of mutation cluster occurrence at a certain IMD (Extended Data Fig. 1a). While the original approach applied a single IMD threshold at which every genome was evaluated, in HyperClust we compute significance estimates at the level of each mutation, meaning that many more samples could be analyzed while retaining acceptable false discovery rates.

HyperClust provides a rigorous estimate of the local FDR (*lfdr*) for each clustered mutation event, given its IMD and the baseline distribution of IMDs in that genome. It is also possible to stratify mutations pairs in each tumor sample into smaller sets according to different features. Because A3 mutagenesis occurs primarily in coordinated cytosines within ssDNA fragments[8,10], we stratified of mutation pairs according to base types (C:G and A:T) and to strand-coordinated bases. We additionally stratified by mutation clonal fraction, as it should be shared by the mutations occuring contemporaneously in a cluster (Supplementary Note).

We evaluated the different stratification features of HyperClust together with other local hypermutation detection approaches from the literature using 48 randomized tumor samples with simulated spiked-in mutation clusters. The stratification with both the strand-coordinated base types and clonal fraction of the mutations outperforms the other tested set ups and was therefore used to obtain mutations for the rest of the analysis (Supplementary Note).

Our method is designed to test pairs of mutations, instead of on larger groups, which leads to balanced power of detection for shorter clusters and longer clusters (*kataegis*-like), while previous methods tend to be better adapted to calling the latter.

Poisson mixture modelling of number of mutations per tract.

The aim of this analysis is to examine whether there exist multiple mechanisms generating clustered mutations, resulting in tracts of different lengths. The number of mutations per cluster can be modeled with a Poisson distribution. We considered only clustered events consisting of two or more mutations at T$\underline{C}$W>K, which are likely to be a highly pure set of the A3 mutations. Then, we modeled the probability that *x* mutations occur in a fragment of ssDNA when two mutations are already present $P(x| x = 2) = Pois(\lambda)$, meaning that 0 represents a cluster pair, 1 represents a triplet etc. If more than one biological mechanism generates clustered mutations at different tract lengths (number of mutations), the observed distribution would be better modeled as a mixture of two or more Poisson distributions, than by a single Poisson distribution.

We used the R package *flexmix*[90] to fit a mixture model, testing the range of components from 1 to 5. We transformed the Akaike Information Criterion (AIC) values extracted from the models to relative likelihoods by calculating the exponential of the difference between each AIC value and the minimum AIC (Extended Data Fig. 1f).

We performed a bootstrap likelihood test (*LR_test* function in *flexmix*) with 500 iterations. This test yields a p-value for the difference of the log-likelihood distributions between the selected model and one more or one less component.

The λ of each Poisson component is the exponential of the fitted intercept in the regression. The confidence intervals of the λ values were obtained by transforming the standard error of that value at C.I. = 95%. We used the λ values to compute density distributions of each component.

We then used the posterior probabilities to obtain the proportion of events with a given track length that can be attributed to each Poisson component (relevant for Fig. 1d, bars). We also obtained a random Poisson distribution for each component based on the λ (relevant for Fig. 1d, lines).

Samples from skin cancer (SKCM) and B-cell lymphoma (DLBC) were excluded from this analysis as they contain particular mutation properties that may confound our analysis. Skin cancer has a high percentage UV signature mutations which overlap with the APOBEC TCW>T context. Somatic hypermutation (SHM) is common in lymphomas and some mutations therein may present a similar profile to the APOBEC mutagenesis.

Association of increased A3 clustered burden with various genomic regions.

Genomic segments and bins extracted from chromatin marks were computed as in ref. [4]. In brief, data for epigenetic marks (H3K36me3) were downloaded from the Roadmap Epigenomics repository, stratified according to the fold-enrichment (FE) of that mark over the input, into three equal-sized bins where the FE>1, and additionally the bin 0, which correspond to regions with FE<1. Expression values were obtained from Roadmap Epigenomics for genic and intergenic regions and processed in a similar manner to the ChipSeq data. Replication time bins were computed from wavelet-smoothed RepliSeq signal tracks from the ENCODE dataset. Again, we binned the genome into equal-frequency bins where bin 1 is the latest-replicating quartile, and bin 4 is earliest-replicating quartile. These data were averaged over the 8 cell lines, as in ref. [4].

To detect significant associations of mutations in specific regions of the genome we used a negative binomial regression[4] (*glm.nb* from the *MASS* R package). In brief, combinatorial intersections between the genomic region sets were computed, 4 bins for each feature. In each set, the number of TCW>K mutations were stratified by the four A3 mutation types (T$\underline{\text{C}}$A>T, TCA>G, TCT>T and TCT>G). These values (mutation counts stratified by mutation type) are used as the dependent variable in the regression and has a total length of 256, corresponding to 64 x 4 mutation types. The number of susceptible genomic sites in 64 bins was also computed and multiplied by the number of samples, thus representing the exposure variable. The three independent variables were the genomic bins of each feature, encoded as factors. This same approach was used for the control contexts (VCN>T). The 95% confidence intervals of the regression coefficient were computed with the *confint* function in R.

For this analysis, we excluded the DLBC (lymphoma) dataset and we discarded mutations in the somatic hypermutation (SHM) off-targets extracted from ref. [91] which might derive from tumor-infiltrated lymphocytes. .

Determining IMD distributions of mutation tracts by simulation.

The IMD distribution of a clustered mutational process will be dependent on the length of the vulnerable DNA segment (for A3, the length of the ssDNA). To determine the expected IMD distribution we randomly sampled with replacement 1,000 times from a set of possible positions and computed the distance between random pairs. We used three sets representing three lengths of ssDNA fragments: short (25 bp), mid-length (200 bp) meant to represent the approximate length of ssDNA between Okazaki fragments in eukaryotes [67] and a long ssDNA (800 bp) meant to represent the ssDNA segments generated during the MMR process [66]. We note that, in order to draw conclusions about ssDNA tract lengths underlying *kataegis*, the cluster span (distance from the first to the last mutation) would be a more appropriate measure. However in case of *omikli*, which consists predominantly of two-mutation clusters, the IMD measure can for practical purposes be considered equivalent to the cluster span measure. For this analysis we considered samples in the APOBEC-prone cancer types in our TCGA dataset: bladder, breast, lung (LUAD and LUSC), cervical, head-and-neck and mismatch repair proficient uterus cancers.

Gamma mixture modelling of IMD distributions.

It is expected the distance between 2 mutations occuring in a single hypermutation event will follow a gamma distribution. Thus, to quantify different mechanisms generating clustered mutations we modelled the observed IMD distributions as a gamma mixture.

We selected only the TCW>K mutations with IMD lower than 1 kb. We also required TCW coordination, meaning that at least 70% of the mutations in that clustered event must have occurred at TCW sites.

We used the R package *mixtools* (*gammamixEM*) that implements an Expectation Maximization (EM) based algorithm for the detection of different components. We obtained estimates for mixtures that ranged from 1 up to 8 components. As initial parameters, we used alpha = 0.2, 100 maximum iterations and an epsilon (convergence difference) of 0.01. We re-simulated the original IMD distributions (see above) for 10,000 iterations and re-computed the parameters. Based on the log-likelihood and the matching shape parameters of the distributions we extracted a total of three components, because the log-likelihood value suggests a strong increase from 1 to 2, and from 2 to 3 components, while the increase from 3 to 4 is more modest; we cannot however rule out a four-component model based on these data. Next, we computed the density of the components using the extracted parameters and the proportions of each component.

Same as the IMD distribution analysis we used samples in the APOBEC prone cancer types, bladder, breast, lung (LUAD and LUSC), cervical, head and neck and mismatch repair proficient uterus cancers.

Contribution of A3 clustered mutagenic process to the unclustered mutation burden.

In order to estimate how much the clustered processes contributed to the unclustered burden, which is the main contributor to the overall tumor mutation burden (TMB), we adapted a method that we recently introduced[4]. In brief, we used a robust linear regression (*rlm* function in the R MASS package) to predict the overall unclustered burden in the TCW>K context (dependent variable) from the counts of each clustered process (TCW>K *kataegis* and *omikli* burden, as separate independent variables (predictors), and additionally an interaction term.

From the fitted model, the intercept is the number of unclustered mutation that cannot be explained by the presence of either *omikli* or *kataegis* clusters, thus, these mutations likely occur independently from the mechanisms that generate either *omikli* or *kataegis*. We named this mutational process A3-X. Similarly, we obtained estimates of the average unclustered mutation burden when one of the two types of clusters (either *omikli* or *kataegis*) is not present but the other type is. These estimates represent the contribution of the *omikli* (A3-O) and *kataegis* (A3-K) processes to the unclustered A3 mutation burden. By adjusting for the total predicted unclustered mutations we can obtain estimates of the contribution of *kataegis* and *omikli* to unclustered burden. Note that because the A3 trinucleotide context (here defined as T<u>C</u>W>K) overlaps with signatures of certain other mutagens, presence of these non-A3-derived unclustered mutations may inflate the estimate of the intercept in the fits (Fig. 4a), causing a downward bias in the estimated *omikli* contribution to global A3 burden (A3-O). For further details, see Supplementary Note.

Parsimony suggests that unclustered (singleton) mutations are generated by the clustered processes of the same mutational context (T<u>C</u>W>K). However, we cannot rule out the possibility that the two processes (*omikli* and unclustered) are mechanistically distinct but tightly co-regulated thus co-occuring in the same tumor samples.

We extracted the 95% prediction intervals of the unclustered values (representing the number of mutations at the average value of each variable) by the R function *predict*. We then used the upper and lower ends of the interval to compute upper and lower bounds of the contribution in percentage. Error bars (Fig. 4 a-c) represent the SEM extracted from this interval.

Functional impact density of mutational processes.

We define the functional impact density (FID) as the putative functionally relevant mutations that occur in a certain set of genes which are associated with a selected mutational process. For a set of genes *G* and a mutational process S, the FID is computed as the number of mutations falling in the coding sequences (CDS) of *G* divided by the total number of mutations from *S*. For sake of clarity, this value can be represented as the number of mutations that fall in a gene coding sequence per thousand mutations.

This measure reports the joint effect of the mutational spectrum, the trinucleotide composition of the gene coding sequence (CDS) and, importantly for the A3 example, the regional preferences of the mutational process. For instance, if the trinucleotide composition of *G* matches with the trinucleotide propensity of *S* it will increase the FID. Also, if *S* is enriched in certain parts of the genome where *G* is also enriched, it will also yield a higher FID.

We selected three disease associated gene sets from the literature, (i) a set of 299 cancer genes, including tumor suppressor genes and oncogenes, which were recurrently mutated in TCGA cancer genomes [72], (ii) a set of genes associated with neurodegenerative disease (n = 39) [92], and finally (iii) a set of cell essential genes extracted from CRISPR/Cas9 genetic screens (n = 683) [93].

In order to obtain mutations that are putatively generated by a given mutational process, we selected those mutations matching the susceptible trinucleotides in a set of tumor samples where the mutational process was reported to occur. In total, we defined four mutational processes: (i) the aging associated process, (ii) “smoking”, (iii) “UV” and (iv) Signature 17. For the ageing process the trinucleotide set was NCG>T and the sample set was comprised by all samples (*n* = 697). For the “smoking” process the trinucleotide subset was NCN>A and the sample set was comprised by lung (LUAD and LUSC) tumor patients with at least three years of tobacco smoking[94] (self-reported data; *sub* 21). For the “UV” process the trinucleotide subset was TCC>T (thus minimizing overlap with other mutational processes) and the sample sets were the skin cancer patients from the TCGA (n = 13) and a set of melanomas PCAWG dataset (MELA-AU, *n* = 70) that were included to increase the number of mutations. For the Signature 17 process the trinucleotide subset was defined as AAN>C and the sample set was the stomach cancers available in our TCGA-WGS data (*n* = 20).

Note that estimates from this analysis are likely conservative because we use a stringent A3 trinucleotide context of TCW>K, and moreover because we examined only unclustered A3 mutations but did not explicitly consider the A3 clustered *omikli* and *kataegis* events in this analysis, on the basis of their lower abundance (Fig. 1f) relative to the unclustered A3 mutations.

Logistic regression approach to determine susceptibility in cancer genes.

We used a logistic regression to determine if the occurrence of a mutation in a cancer gene was associated with a higher burden of either *omikli* or *kataegis*. We examined the set of 299 cancer genes[72] and selected mutations in their coding sequence (CDS) matching the A3 context TCW>K (W is A or T; K is T or G). If a gene contained at least one of these mutations in the CDS it was classified as mutated by an A3 process. We tested only the 61 cancer genes (Supplementary Table 1) that bore A3 context mutations in at least 3 samples from the TCGA-WGS dataset. As negative control we also counted mutations in the cancer genes at the non-A3 context VCN>K (V is not T).

Next, we performed a multiple logistic regression using the square-rooted burdens of *omikli* and *kataegis* as independent variables to predict the mutation status of the gene (dependent variable). The independent variables were always restricted to the A3 (TCW>K) context to represent the A3 activity of either *omikli* or *kataegis*. The mutation status was tested both with genes harboring A3 mutations and the control context (VCN>K). The p-values for each gene were FDR adjusted using the Benjamini-Hochberg correction.

We also divided the CDS fragments from the cancer genes according to their replication time and then used logistic regression to predict if any of the CDS located in that specific replication time bin was mutated. We used the number of *omikli* mutations (square-rooted) as predictor.

Statistics

If not stated otherwise, the comparison of two distributions of continuous values was tested with a two-tailed Mann-Whitney U test. Pooling p-values obtained from stratified data groups was performed with the Fisher's method for combining P-values. P values are shown as exact values or otherwise referenced as symbol according to this scale: *** < 0.001, ** < 0.01, * < 0.05, "." < 0.1.

All boxplots used in the current analysis are represented according to the standard boxplot notation in the R statistical environment (*ggplot2* package): the central box represents the inter quartile range (IQR), the central line is the median value of the distribution, the outlier points are instances higher or lower than 1.5 times the IQR from the median value and the whiskers are the lowest and highest points of the distribution after removing the outliers. If the boxplot has notches, the notch width is 1.58 times the IQR divided by the square root of the sample size, which is an estimate of the 95% C.I. of the median.

Data availability statement

For the current study we used publicly available data described in the Methods. In brief, we used a set of whole genome sequences from TCGA available through cgHub repository (superseded by the NCI Genomic Data Commons, https://gdc.cancer.gov/). SNP arrays for the same data set were downloaded from the GDC legacy portal (portal.gdc.cancer.gov/legacy-archive). We used two validation sets: (i) the whole genome tumor cohort from the Hartwig Medical Foundation available at hartwigmedicalfoundation.nl (DR-069) upon request and (ii) the whole exome TCGA cohort through the MC3 dataset available at https://gdc.cancer.gov/about-data/publications/mc3-2017. Data generated by the analyses in this study are available in the Supplementary Tables.

Code availability

Code to generate clustered mutation calls was implemented in Python (version 3.6) and R environments (version 3.6). Relevant packages are biopython (version 1.73) and numpy (version 1.15.4) for Python, and Biostrings (2.52.0), VariantAnnotation (1.30.1) and GenomicRanges (1.36.0) for R. Code is available at https://github.com/davidmasp/hyperclust.

Statistical analysis of the data was performed using custom scripts in R (version 3.6); relevant packages are mclust (version 5.4.4), mixtools (version 1.1.0), MASS (version 7.3-51.4) and flexmix (version 2.3-15).

Reporting Summary

Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Methods references

83. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

84. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).

85. Huang, M. N. *et al.* Msiseq: software for assessing microsatellite instability from catalogs of somatic mutations. *Sci. Rep.* **5**, 1–10 (2015).

86. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).

87. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).

88. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

89. Ellrott, K. *et al.* Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271-281.e7 (2018).

90. Grün, B. & Leisch, F. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. *J. Stat. Softw.* **28**, 1–35 (2008).

91. Khodabakhshi, A. H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308–1319 (2012).

92. Krüger, S. *et al.* Rare variants in neurodegeneration associated genes revealed by targeted panel sequencing in a german ALS cohort. *Front. Mol. Neurosci.* **9**, (2016).

93. Hart, T. *et al.* Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 Genes Genomes Genet.* **7**, 2719–2727 (2017).

94. Liu, J. *et al.* An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400-416.e11 (2018).

Figure 1: Two types of local hypermutation in human tumors.

a, The HyperClust framework detects mutation clustering by accounting for heterogeneous mutation rates at the megabase scale, further stratifying mutations by type, and additionally by their approximate timing (clonal fraction). b, Kataegis (thunderstorm) and omikli (fog) mutation clusters in an example tumor genome segment (chromosome 8 of TCGA-DK-A1A6). Vertical lines are rearrangement loci. c, Distribution of the number of A3-context TCW>K mutations in omikli (bottom) and kataegis (top) of different sizes (number of mutations per cluster; callouts). d, Poisson mixture modeling of the number of A3-context mutations per cluster. A solution with two distributions is shown (teal: kataegis; orange: omikli). The stacked bars show component proportions and the curves are density estimates. The gray curve is the baseline solution with one component. The P values are from a two-sided bootstrap test. LL, log likelihood. e, Cumulative percentage of tumor samples that contained at least the given number of clustered mutations, either observed or expected at random. f, Distribution of the burden of A3-context somatic mutations per tumor, across tumors. Samples with no omikli mutations or no kataegis mutations were not considered. g, Cumulative fraction of A3 mutations within the neighborhood (width on x axis) of a rearrangement breakpoint. Error bars are 95% binomial CIs. Numbers of mutations are listed in parentheses. h, Pearson's correlations between the burden of two-mutation omikli and of long kataegis events (left) and between the burden of kataegis of different lengths (right). Statistical significance was determined by two-tailed t-test on the Fisher-transformed correlation coefficients.

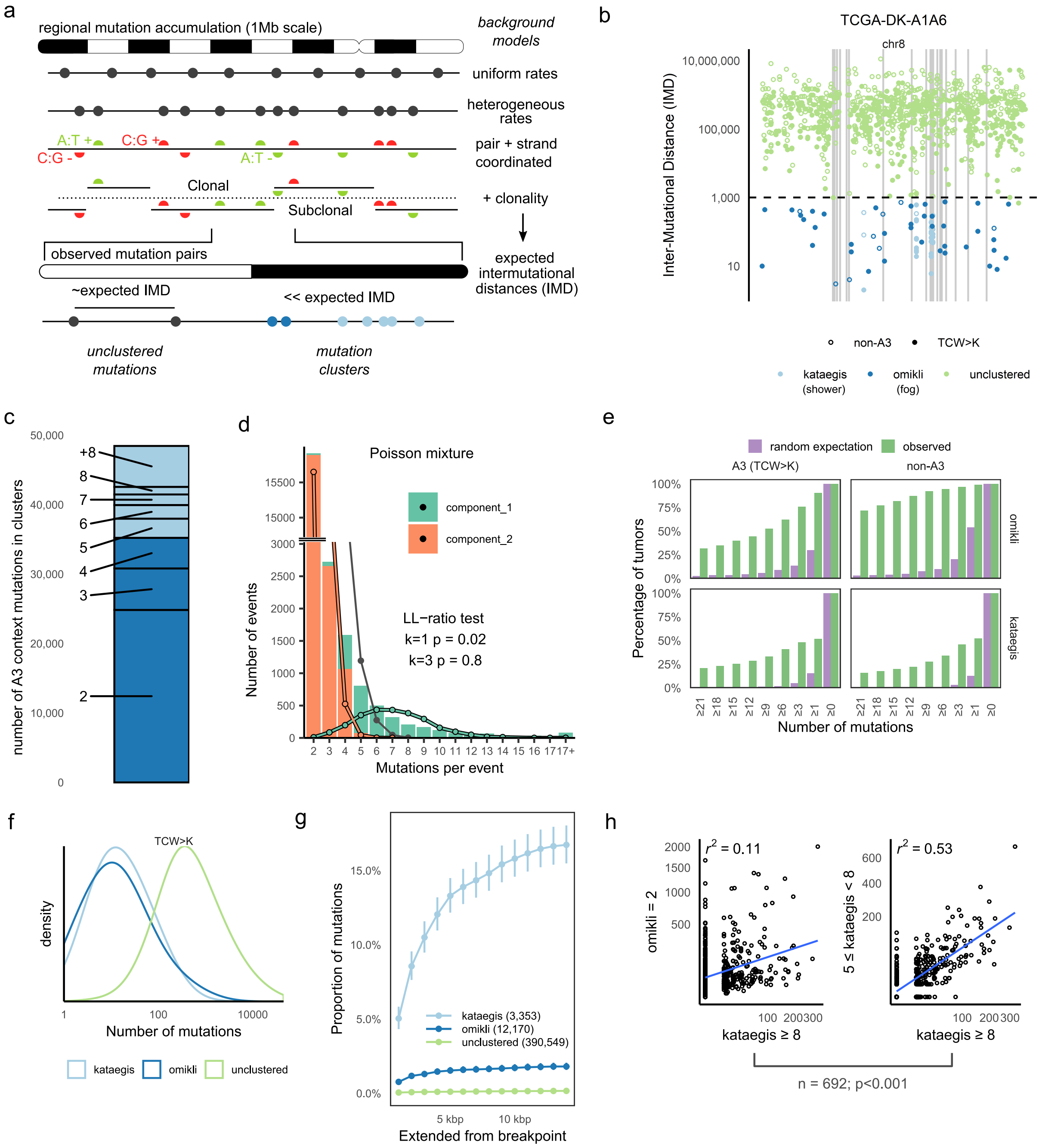

a
regional mutation accumulation (1Mb scale)
background models
uniform rates
heterogeneous rates
A:T +
C:G +
C:G -
A:T -
pair + strand coordinated
Clonal
Subclonal
+ clonality
observed mutation pairs
expected intermutational distances (IMD)
~expected IMD
<< expected IMD
unclustered mutations
mutation clusters
b
TCGA-DK-A1A6
chr8
Inter-Mutational Distance (IMD)
10,000,000
100,000
1,000
10
non-A3
TCW>K
kataegis (shower)
omikli (fog)
unclustered
c
number of A3 context mutations in clusters
50,000
40,000
30,000
20,000
10,000
0
+8
8
7
6
5
4
3
2
d
Poisson mixture
component_1
component_2
LL−ratio test
k=1 p = 0.02
k=3 p = 0.8
Number of events
15500
15000
3000
2500
2000
1500
1000
500
0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 17+
Mutations per event
e
random expectation
observed
A3 (TCW>K)
non-A3
omikli
kataegis
Percentage of tumors
100%
75%
50%
25%
0%
≥21 ≥18 ≥15 ≥12 ≥9 ≥6 ≥3 ≥1 ≥0
Number of mutations
f
TCW>K
density
1
100
10000
Number of mutations
kataegis
omikli
unclustered
g
Proportion of mutations
15.0%
10.0%
5.0%
0.0%
kataegis (3,353)
omikli (12,170)
unclustered (390,549)
5 kbp
10 kbp
Extended from breakpoint
h
r^2 = 0.11
r^2 = 0.53
omikli = 2
2000
1500
1000
500
5 ≤ kataegis < 8
600
400
200
100 200300
kataegis ≥ 8
n = 692; p<0.001

Figure 2: Association of A3 clustered mutation density with genomic features.

a, Mutation rates in replication time (RT) quartiles, relative to the latest RT quartile, for A3 mutation trinucleotide contexts (top) and control contexts (bottom). b, Mutation enrichment in the earliest versus latest RT quartile for A3-context clusters (top) and non-A3-context clusters (bottom). Cancer types are ordered by total A3 burden across all tumors (shading in top bar). Moderate/low A3 burden cancer types are pooled into the group ‘other’. c, Relative density of A3 and non-A3 mutation types across genomic regions. All enrichments are relative to the lowest bin (the latest RT quartile for replication time), which is not shown. Points are coefficients from negative binomial regression. Error bars are 95% CIs. d, Replication strand bias (ratio of the mutation count on the leading versus the lagging DNA strand) of clustered TCW mutations. Error bars are binomial 95% CIs. As a control, the reciprocal of the strand bias for MSI-H (orange; 24 samples) and POLE mutant (purple; nine samples) tumors is shown as a dashed line. Values in parentheses are mutation counts used to estimate the ratios. MSI-H, microsatellite instability-high. e, Left: distributions of IMD in A3-context kataegis and omikli clusters. Right: expected IMD distributions from simulations using three different segment lengths. f, Gamma mixture modeling of the omikli IMD distribution using three components. The bar shows the proportions of the components. The curves show their densities at various IMDs. BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma; HNSC, head and neck squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; SARC, sarcoma; UCEC, uterine corpus endometrial carcinoma.

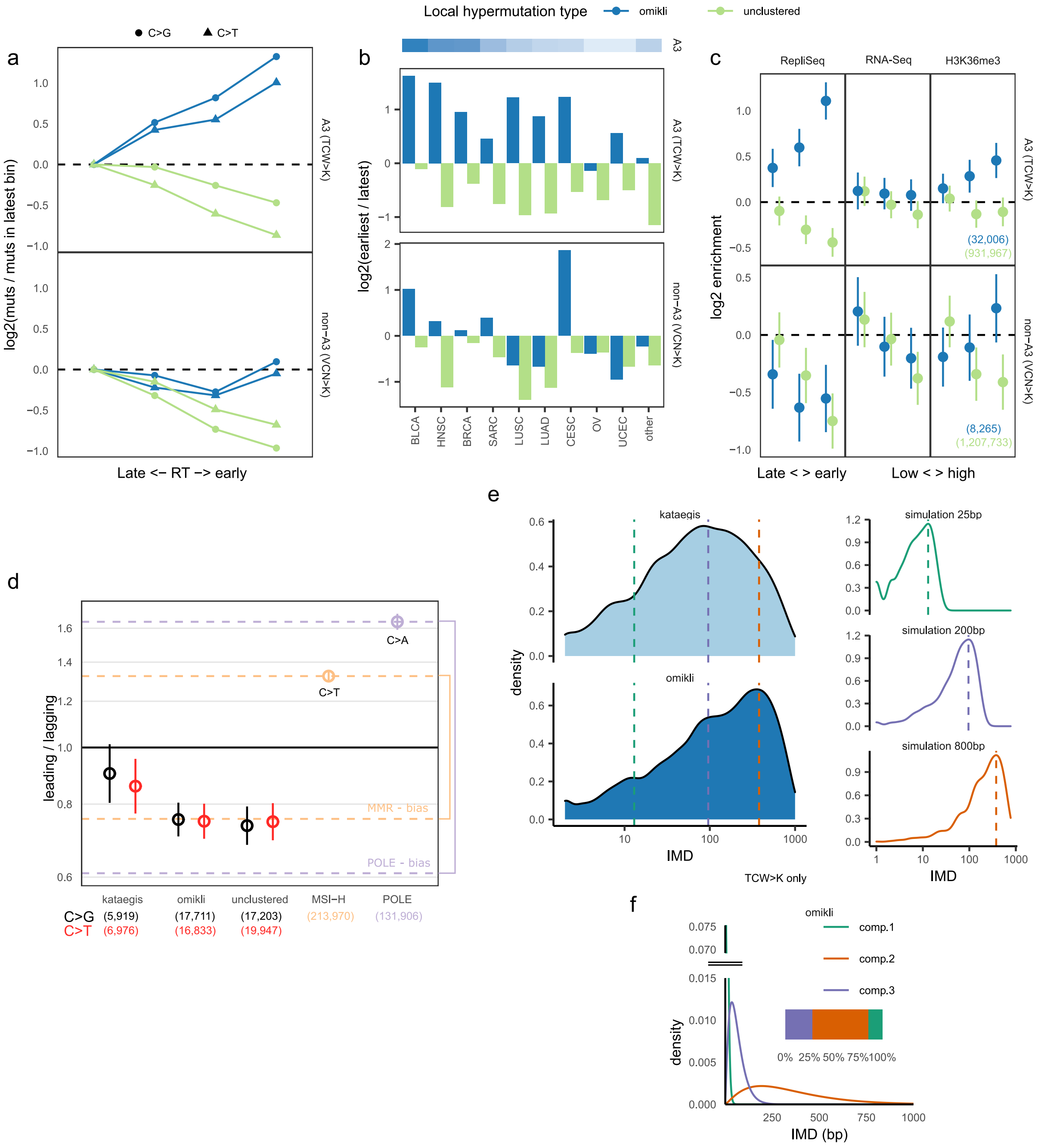

Local hypermutation type
omikli
unclustered
a
● C>G
▲ C>T
log2(muts / muts in latest bin)
1.0
0.5
0.0
−0.5
−1.0
A3 (TCW>K)
non−A3 (VCN>K)
Late <− RT −> early
b
A3
log2(earliest / latest)
1
0
−1
2
A3 (TCW>K)
non−A3 (VCN>K)
BLCA
HNSC
BRCA
SARC
LUSC
LUAD
CESC
OV
UCEC
other
c
RepliSeq
RNA-Seq
H3K36me3
log2 enrichment
1.0
0.5
0.0
−0.5
−1.0
A3 (TCW>K)
non−A3 (VCN>K)
(32,006)
(931,967)
(8,265)
(1,207,733)
Late < > early
Low < > high
d
leading / lagging
1.6
1.4
1.2
1.0
0.8
0.6
C>A
C>T
MMR - bias
POLE - bias
kataegis
omikli
unclustered
MSI−H
POLE
C>G
C>T
(5,919)
(6,976)
(17,711)
(16,833)
(17,203)
(19,947)
(213,970)
(131,906)
e
kataegis
omikli
density
0.6
0.4
0.2
0.0
10
100
1000
IMD
TCW>K only
simulation 25bp
simulation 200bp
simulation 800bp
1.2
0.9
0.6
0.3
0.0
1
10
100
1000
IMD
f
omikli
comp.1
comp.2
comp.3
0.075
0.070
0.015
0.010
0.005
0.000
density
0%
25%
50%
75%
100%
250
500
750
1000
IMD (bp)

Figure 3: MMR activity in tumors is associated with APOBEC mutagenesis.

a, Proportion of omikli clusters in A3 (left) and control non-A3 contexts (right), comparing MMR-deficient (MSI-H) samples with MMR-proficient (MSS) samples in either matched tissues (that is, colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD) and UCEC) or non-matched tissues. Significance was determined by two-tailed Mann–Whitney U-test. Numbers of tumor samples are listed in parentheses. b, Same as a, but broken down by tissue. Pooled P values: $P < 0.001$ for A3; $P = 0.433$ for the control. Statistical significance was determined by two-tailed Mann–Whitney U-test on stratified data. Black horizontal lines are medians of the distributions. c, Enrichment of A3 omikli clusters and unclustered A3 mutations in various genome regions in MMR-deficient samples (MSI-H). This panel is related to Fig. 2c. Coefficients of negative binomial regression are shown (as log2), indicating enrichments of mutation frequency in a genomic bin versus the lowest bin (in the case of replication time, latest replicating), where enrichment would equal unity and is thus not shown. Error bars are 95% CIs. d, Correlation of the burden of A3-context (TCW>K) kataegis, omikli and unclustered mutations with mRNA levels of MMR genes and of APOBEC3A and APOBEC3B genes. Error bars are 95% CIs. e, Association of CNAs in selected MMR genes with burden of A3 omikli. CNAs are represented as integer copy number differences (Methods). Positive values are gains and negative values are losses. See also Extended Data Fig. 3g. Significance was determined by two-tailed Mann–Whitney U-test comparing the neutral (0) versus the gain (+1 and +2) states considered jointly. **$P < 0.01$; ***$P < 0.001$. See the ‘Statistics’ section of the Methods for interpretation of the box plots.

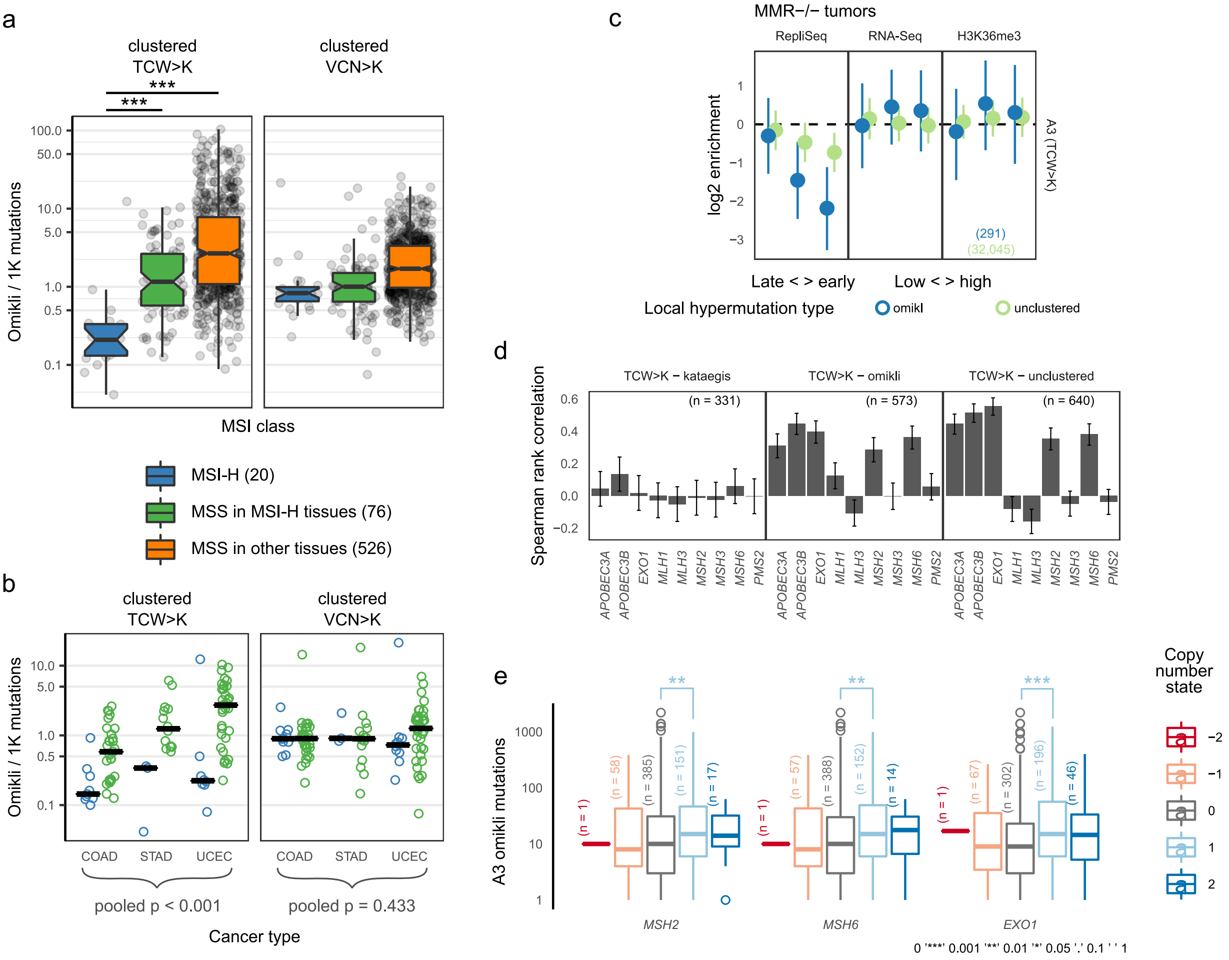

a
clustered TCW>K
clustered VCN>K
***
***
Omikli / 1K mutations
100.0
50.0
10.0
5.0
1.0
0.5
0.1
MSI class
MSI-H (20)
MSS in MSI-H tissues (76)
MSS in other tissues (526)
b
clustered TCW>K
clustered VCN>K
Omikli / 1K mutations
10.0
5.0
1.0
0.5
0.1
COAD
STAD
UCEC
COAD
STAD
UCEC
pooled p < 0.001
pooled p = 0.433
Cancer type
c
MMR−/− tumors
RepliSeq
RNA-Seq
H3K36me3
log2 enrichment
1
0
−1
−2
−3
A3 (TCW>K)
(291)
(32,045)
Late < > early
Low < > high
Local hypermutation type
omikl
unclustered
d
Spearman rank correlation
TCW>K − kataegis
TCW>K − omikli
TCW>K − unclustered
(n = 331)
(n = 573)
(n = 640)
0.6
0.4
0.2
0.0
−0.2
APOBEC3A
APOBEC3B
EXO1
MLH1
MLH3
MSH2
MSH3
MSH6
PMS2
e
A3 omikli mutations
1000
100
10
1
**
**
***
(n = 1)
(n = 58)
(n = 385)
(n = 151)
(n = 17)
(n = 1)
(n = 57)
(n = 388)
(n = 152)
(n = 14)
(n = 1)
(n = 67)
(n = 302)
(n = 196)
(n = 46)
MSH2
MSH6
EXO1
Copy number state
−2
−1
0
1
2
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 4: The omikli process generates the majority of unclustered A3 mutations across tissues.

a, A regression analysis estimates the contributions of omikli and kataegis processes towards the unclustered A3 mutation burden. The results for LUAD tumor samples (points) are shown (other cancers are shown in Extended Data Fig. 6). For clarity, combinations of two variables are shown (center: omikli versus unclustered; right: kataegis versus unclustered), even though the regression was performed on the three variables simultaneously (schematic in leftmost panel; Methods). The red line is the intersection of the fitted plane with the shown two-dimensional coordinate system. Error bars are 95% prediction intervals of the fit. The dotted line is the average omikli (center) and kataegis (right) mutation burden across tumors. Dashed lines are the estimated contributions for each process (also shown as bars on the right part of the plot). The bottom panels show the same data as the top panels, but zoomed in on the x axis for clarity. b, Pan-cancer regression analysis provides estimates of the fraction of unclustered TCW>K mutations contributed by processes that generate omikli, kataegis and a remainder (intercept of regression fit) not explained by either process. Error bars show s.e. of regression coefficients (n = 646 tumors). c, The relative contribution of the omikli process to the unclustered A3 burden (y axis) of cancer types correlates with the overall burden of A3 mutations in that cancer type (x axis), suggesting that differential activity of the omikli mechanism drives differences in A3 burden between tissues. Error bars show s.e. of regression coefficients. The shaded band is the 95% CI of the linear fit. GBM, glioblastoma multiforme; KICH, kidney chromophobe cancer; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, brain lower grade glioma; LIHC, liver hepatocellular carcinoma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; THCA, thyroid carcinoma.

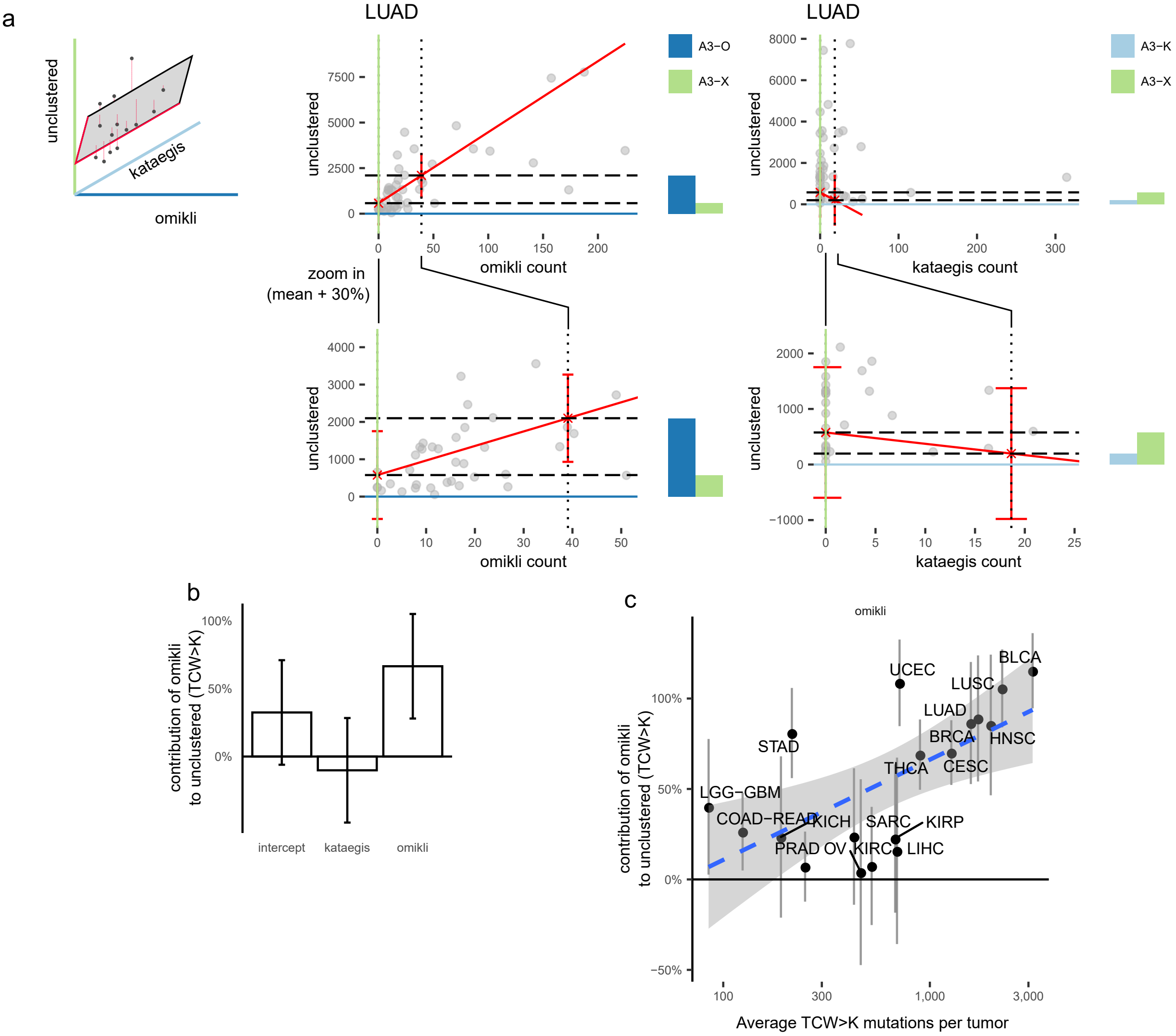
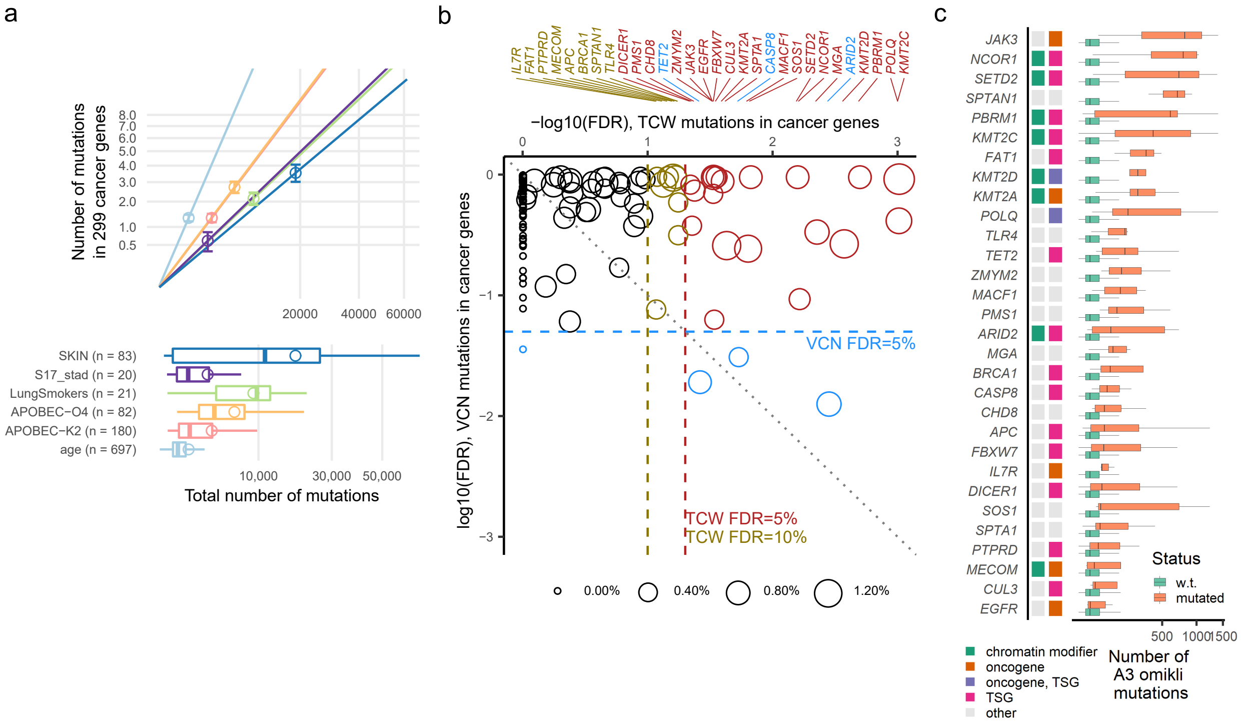

a
unclustered
kataegis
omikli
LUAD
unclustered
omikli count
A3−O
A3−X
LUAD
unclustered
kataegis count
A3−K
A3−X
zoom in
(mean + 30%)
unclustered
omikli count
unclustered
kataegis count
b
contribution of omikli
to unclustered (TCW>K)
100%
50%
0%
intercept
kataegis
omikli
c
omikli
contribution of omikli
to unclustered (TCW>K)
100%
50%
0%
−50%
100
300
1,000
3,000
Average TCW>K mutations per tumor
UCEC
BLCA
LUSC
LUAD
BRCA
HNSC
STAD
THCA
CESC
LGG−GBM
COAD−READ
KICH
SARC
KIRP
PRAD
OV
KIRC
LIHC

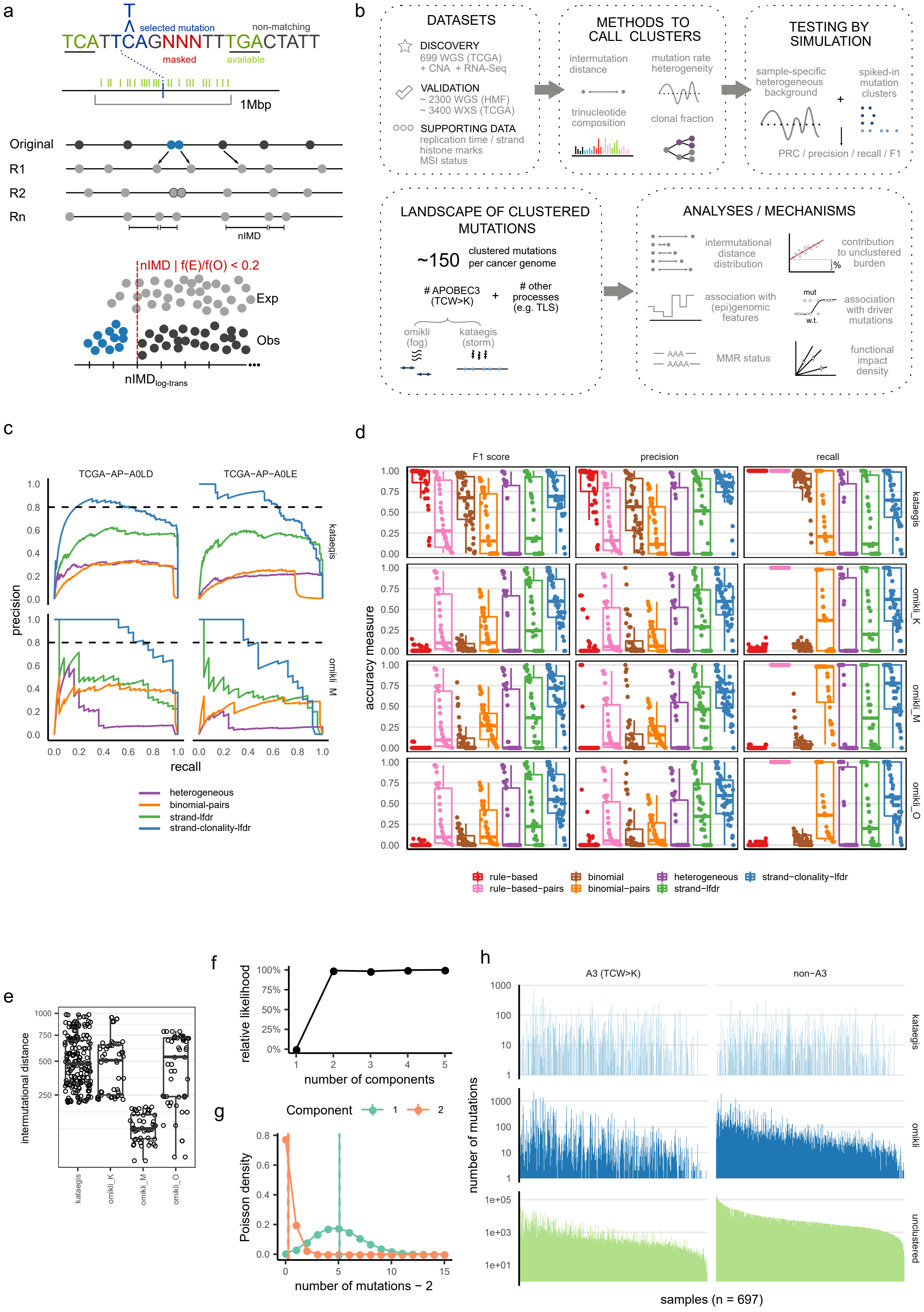
Figure 5: APOBEC mutagenesis generates many impactful mutations.

a, Top: the functional impact density (FID) of mutational processes (slope of line), estimated as the number of mutations in coding regions of 299 cancer genes (y axis) normalized to the total mutation tally contributed by a process (x axis). Bottom: number of mutations estimated to result from each process across tumor samples. Hollow circles in box plots (bottom panel) and on lines (top panel) are the average mutation burden of that process in the affected tumor samples (definition in Methods). APOBEC-O4, A3 mutagenesis in omikli-rich tumors; APOBEC-K2, A3 mutagenesis in kataegis-rich tumors; S17_stad, Signature 17 mutagenesis in stomach adenocarcinomas; SKIN, UV mutagenesis in melanoma; age, aging-associated mutagenesis (details in Methods). Error bars are s.e.m. b, The occurrence of A3-context mutations in many cancer genes is associated with the genomic burden of A3 omikli mutation clusters, suggesting that the omikli process generates driver mutations. FDRs are Benjamini–Hochberg adjusted P values from a logistic regression to predict the presence of a TCW>K (A3-context; x axis) or VCN>K (control non-A3-context; y axis) mutation in each driver gene. The red and gold dashed lines, respectively, represent stringent (5%) and permissive (10%) FDR thresholds for the A3 context. The blue dashed line represents the (5%) FDR threshold in the control context, suggesting an indirect association with A3 omikli burden. The diagonal line denotes equal FDR between the A3 and control contexts. FDRs were capped at 0.1%. c, Burden of A3 omikli mutations, in wild-type and mutated tumors, in the driver genes that were significantly associated in the logistic regression in b. See the ‘Statistics’ section of the Methods for interpretation of the box plots. TSG, tumor suppressor gene.


a
Number of mutations
in 299 cancer genes
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.5
20000
40000
60000
SKIN (n = 83)
S17_stad (n = 20)
LungSmokers (n = 21)
APOBEC−O4 (n = 82)
APOBEC−K2 (n = 180)
age (n = 697)
10,000
30,000
50,000
Total number of mutations
b
IL7R
FAT1
PTPRD
MECOM
APC
BRCA1
SPTAN1
TLR4
DICER1
PMS1
CHD8
TET2
ZMYM2
JAK3
EGFR
FBXW7
CUL3
KMT2A
SPTA1
CASP8
MACF1
SOS1
SETD2
NCOR1
MGA
ARID2
KMT2D
PBRM1
POLQ
KMT2C
−log10(FDR), TCW mutations in cancer genes
0
1
2
3
log10(FDR), VCN mutations in cancer genes
0
−1
−2
−3
VCN FDR=5%
TCW FDR=5%
TCW FDR=10%
0.00%
0.40%
0.80%
1.20%
c
JAK3
NCOR1
SETD2
SPTAN1
PBRM1
KMT2C
FAT1
KMT2D
KMT2A
POLQ
TLR4
TET2
ZMYM2
MACF1
PMS1
ARID2
MGA
BRCA1
CASP8
CHD8
APC
FBXW7
IL7R
DICER1
SOS1
SPTA1
PTPRD
MECOM
CUL3
EGFR
Status
w.t.
mutated
500
1000
1500
Number of
A3 omikli
mutations
chromatin modifier
oncogene
oncogene, TSG
TSG
other

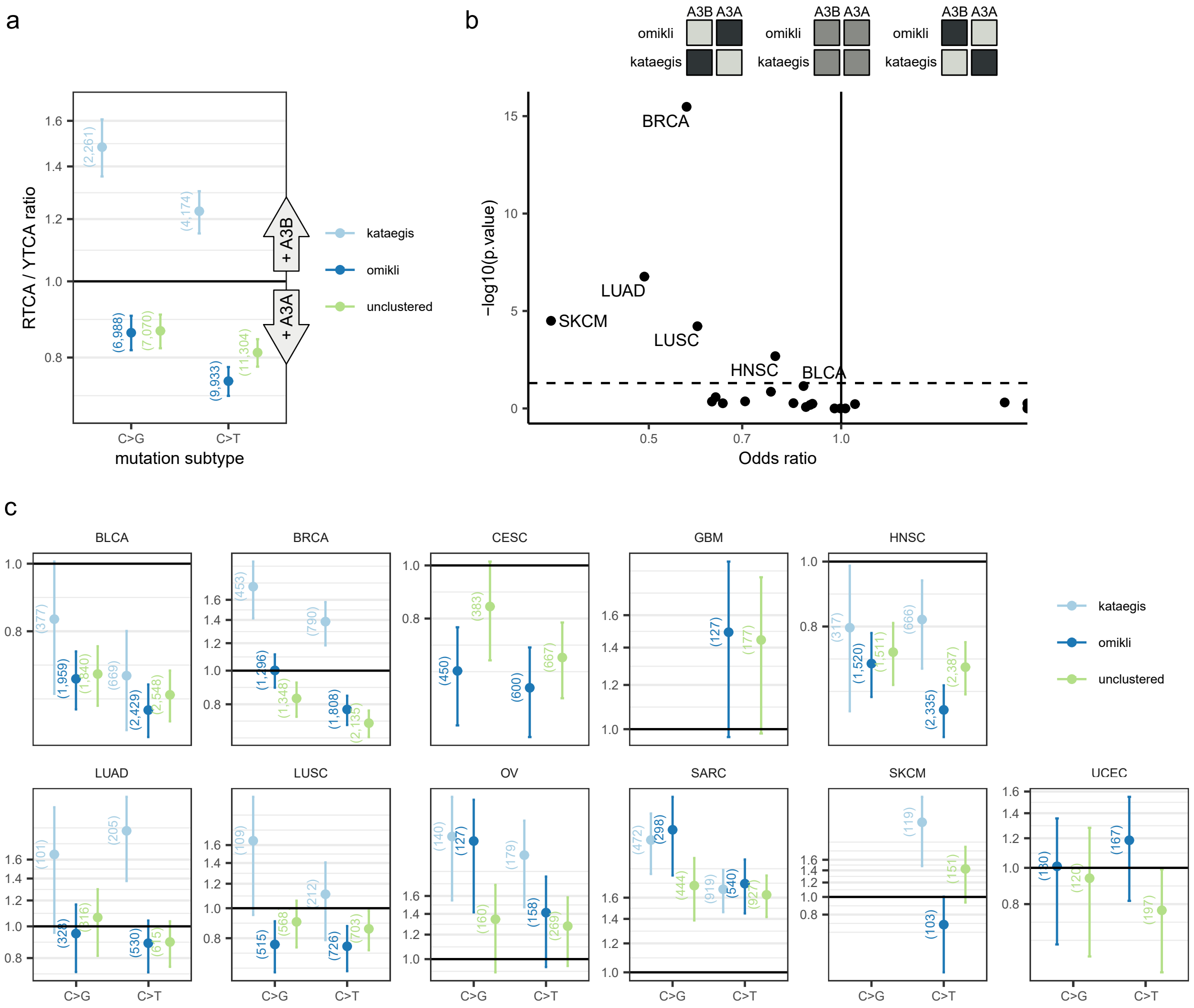
Extended Data Fig. 1: Detecting clustered mutations and simulating processes that generate clustered mutations.

a, Method to determine significant mutation clustering using HyperClust. A baseline distribution is generated by shuffling mutations within 1 Mbp windows multiple times (R1, R2, …, Rn) to loci with matching trinucleotide contexts. For every mutation, the observed intermutational distance to its nearest neighbour (nIMD) is compared with distributions of expected IMDs (from randomized data) to determine a local FDR (lfdr). Thresholding by lfdr yields clustered mutation calls (blue). b, Overview of study. c, Precision-recall curves for models in Fig. 1a, derived from simulated data with spiked-in mutation clusters: kataegis (top; with five mutations per cluster at an average 600 bp pairwise distance) or omikli_M (bottom; two mutations at 101 bp). Two examples of high mutation burden tumors (TCGA-AP-A0LD, TCGA-AP-A0LE) were used to generate the background mutation distributions. d, e, Testing accuracy of mutation cluster calling methods using simulated data. Points represent randomized tumor samples into which spiked-in mutation clusters were introduced. Samples are ordered according to total mutation burden (panel d). Columns show different performance metrics: F1 score, precision, and recall, all at lfdr=20%. Rows represent different types of spiked-in mutation clusters (IMD distributions plotted in panel e, where kataegis have five mutations and omikli_K/M/O two mutations. Boxplots compare cluster calling methods, including implementations of some previous methodologies (details in Methods). The “strand-clonality-lfdr” (blue) is the HyperClust method used throughout our work. f, g, Poisson mixture modelling (related with Fig. 1d) of the number of mutations per cluster, showing relative likelihood (panel f) of models with increasing number of components and the density functions (panel g) of a model with two Poisson components. solid line represents mean and dashed lines the 95% C.I. h, Number of mutation events per tumor sample (x axis, n) per local hypermutation type (rows), either the A3 context TCW>K mutations, or the remaining mutations (columns).


a
T
selected mutation
non-matching
TCATTCAGNNNTTTGACTATT
masked
available
1Mbp
Original
R1
R2
Rn
nIMD
nIMD | f(E)/f(O) < 0.2
Exp
Obs
nIMD$_{log-trans}$
b
DATASETS
DISCOVERY
699 WGS (TCGA)
+ CNA + RNA-Seq
VALIDATION
~ 2300 WGS (HMF)
~ 3400 WXS (TCGA)
SUPPORTING DATA
replication time / strand
histone marks
MSI status
METHODS TO CALL CLUSTERS
intermutation distance
mutation rate heterogeneity
trinucleotide composition
clonal fraction
TESTING BY SIMULATION
sample-specific heterogeneous background
+
spiked-in mutation clusters
PRC / precision / recall / F1
LANDSCAPE OF CLUSTERED MUTATIONS
~150
clustered mutations per cancer genome
# APOBEC3 (TCW>K)
+
# other processes (e.g. TLS)
omikli (fog)
kataegis (storm)
ANALYSES / MECHANISMS
intermutational distance distribution
contribution to unclustered burden
%
association with (epi)genomic features
mut
w.t.
association with driver mutations
AAA
AAAA
MMR status
functional impact density
c
TCGA−AP−A0LD
TCGA−AP−A0LE
precision
recall
kataegis
omikli_M
1.0
0.8
0.6
0.4
0.2
0.0
0.0 0.2 0.4 0.6 0.8 1.0
heterogeneous
binomial-pairs
strand-lfdr
strand-clonality-lfdr
d
F1 score
precision
recall
accuracy measure
1.00
0.75
0.50
0.25
0.00
kataegis
omikli_K
omikli_M
omikli_O
rule−based
rule−based−pairs
binomial
binomial−pairs
heterogeneous
strand−lfdr
strand−clonality−lfdr
e
intermutational distance
1000
750
500
250
kataegis
omikli_K
omikli_M
omikli_O
f
relative likelihood
100%
75%
50%
25%
0%
1 2 3 4 5
number of components
g
Component
1
2
Poisson density
0.8
0.6
0.4
0.2
0.0
0 5 10 15
number of mutations − 2
h
A3 (TCW>K)
non−A3
number of mutations
1000
100
10
1
kataegis
omikli
1e+05
1e+03
1e+01
unclustered
samples (n = 697)

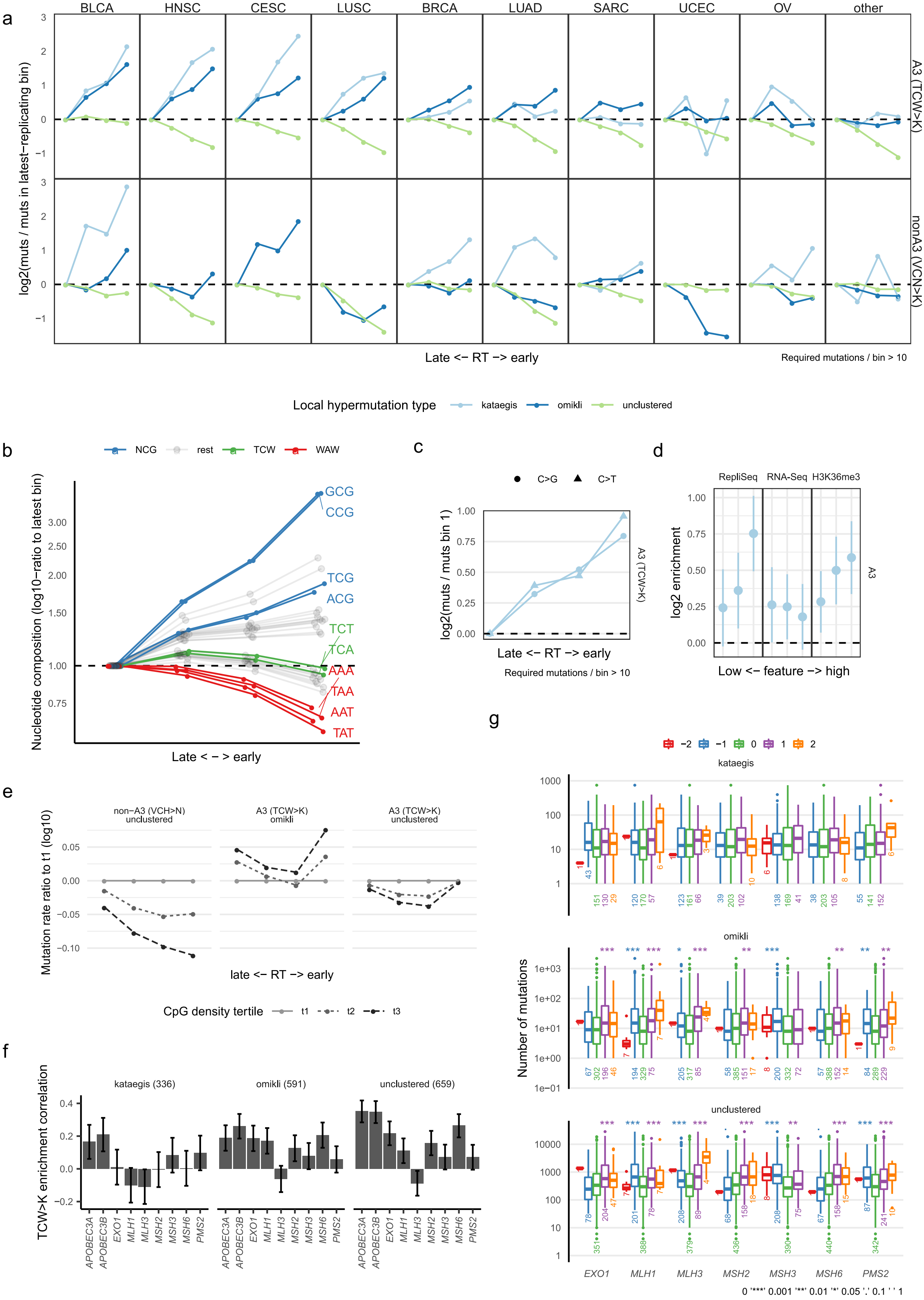
Extended Data Fig. 2: Tetranucleotide context suggests a role for the A3A enzyme in generating omikli and A3B in kataegis mutations.

a, c, Ratios of the YTCA (A3A-like) and RTCA (A3B-like) mutation frequencies suggest differential mutagenic activity of A3A versus A3B enzymes in cancer samples. The C>T and the C>G changes in the two A3 contexts are shown in a pan-cancer analysis (panel a) and broken down by cancer type (panel c). At least 100 TCW mutations of a certain type across all tumor samples in a tissue were required to perform analyses on that tissue (number of mutations in brackets). Error bars are the bootstrap 95% C.I. of the ratio. KICH and THCA cancer types are not shown due to low overall number of A3-context mutations. b, Across multiple cancer types, omikli shows a tendency towards A3A-like, lower RTCA/YTCA-ratios than does kataegis. Difference tested by Fisher's exact test (per tumor type), two-tailed; p-values were adjusted for multiple testing. Dashed line is FDR=20%. Lower odds ratios (<1) denote relative enrichment of YTCA (A3A-like) mutations in omikli compared to kataegis; see schematic above plot.


a
RTCA / YTCA ratio
1.6
1.4
1.2
1.0
0.8
(2,261)
(4,174)
(6,988)
(7,070)
(9,933)
(11,304)
+ A3B
+ A3A
C>G
C>T
mutation subtype
kataegis
omikli
unclustered
b
A3B A3A
omikli
kataegis
−log10(p.value)
15
10
5
0
BRCA
LUAD
SKCM
LUSC
HNSC
BLCA
0.5
0.7
1.0
Odds ratio
c
RTCA / YTCA ratio
BLCA
(377)
(1,959)
(1,840)
(669)
(2,429)
(2,548)
BRCA
(453)
(790)
(1,296)
(1,348)
(1,808)
(2,135)
CESC
(450)
(383)
(600)
(667)
GBM
(127)
(177)
HNSC
(317)
(1,520)
(511)
(666)
(2,335)
(2,387)
LUAD
(101)
(328)
(316)
(205)
(530)
(615)
LUSC
(109)
(515)
(568)
(212)
(726)
(703)
OV
(140)
(127)
(160)
(179)
(158)
(269)
SARC
(472)
(298)
(444)
(919)
(540)
(927)
SKCM
(119)
(103)
(151)
UCEC
(130)
(120)
(167)
(197)
C>G
C>T
kataegis
omikli
unclustered

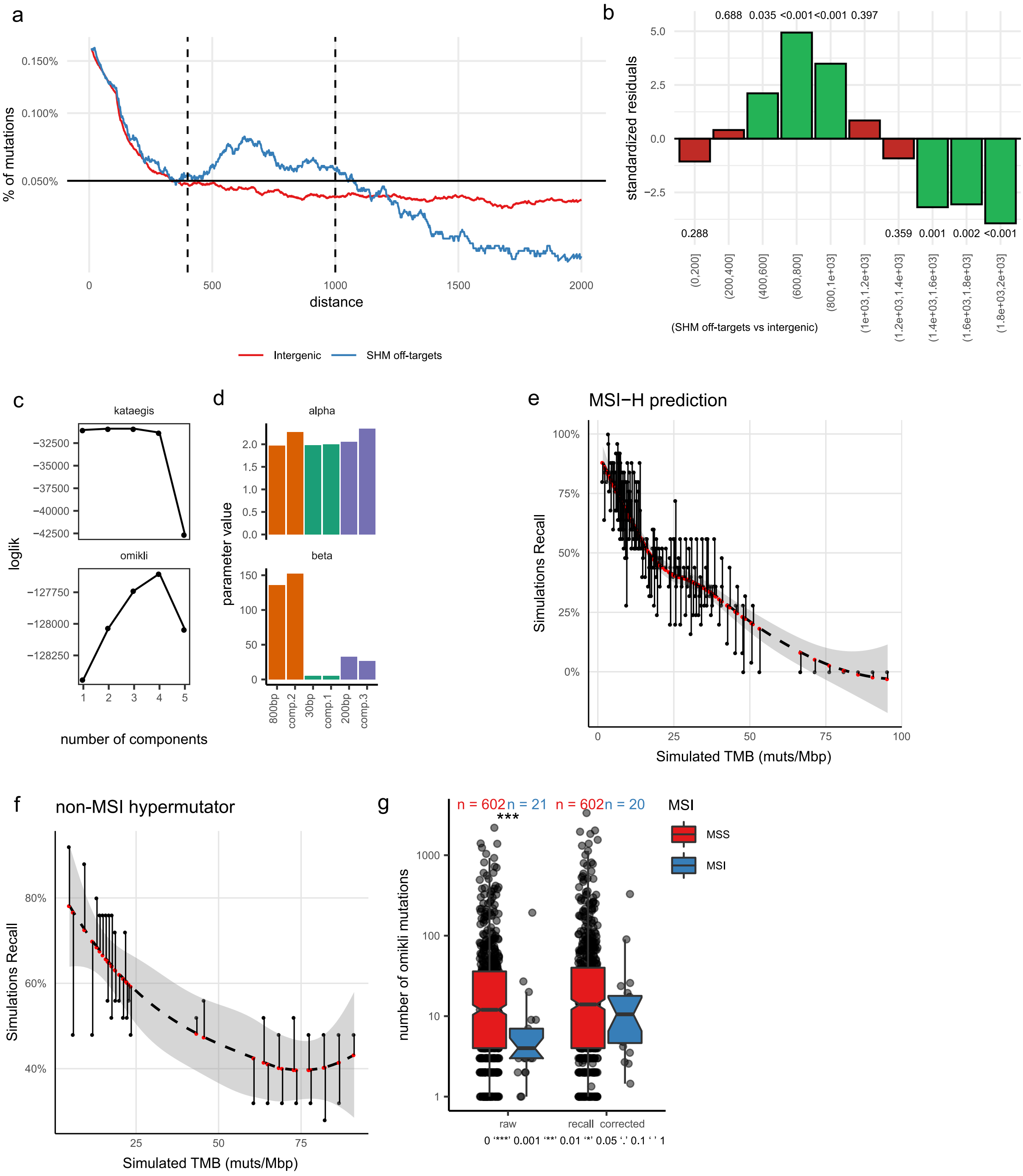
Extended Data Fig. 3: Association of clustered mutation rates with replication time (RT).

a, RT association per cancer type. Number of mutations per RT bin: A3 context (top row) and the non-A3 control context at C:G nucleotide pairs (bottom row). RT bins are ordered from the latest-replicating quartile to the earliest-replicating quartile; mutation rates are shown relative to the latest RT bin. Enrichments are not shown when the mutation count was lower than 10. b, Trinucleotide composition of the human reference genome in four RT bins, normalized to the latest RT quartile (leftmost point). The A3 trinucleotide contexts (TCW, green) are similarly abundant in the late and in the early-replicating regions of the genome. c, d, Enrichment of A3-context kataegis clusters, considering only RT (c), or jointly considering RT, mRNA levels and the H3K36me3 histone mark levels (d); points are coefficients from negative binomial regression, and error bars are 95% C.I. e, Mutation rates in genomic bins with different CpG density (determined per 10 kb segment), stratified by RT quartiles. y axis shows mutation densities relative to the first bin (‘t1’, lowest tertile by CpG content). f, Spearman correlation between mRNA expression of A3A, A3B and MMR genes, and the TCW context enrichment of clustered mutations in a tumor. Error bars are 95% C.I. from the Fisher transformation of the correlation coefficient. g, Association of A3 mutation burden (clustered and unclustered) with copy number alterations of MMR genes. Significance by a two-tailed Mann-Whitney test, comparing tumor samples with neutral (0) versus gain/amplification (+1 and +2) states (blue stars, showing p-values according to legend), and independently, comparing samples with neutral (0) versus loss (−1 and −2) states (purple stars). P-values were not adjusted.


a
BLCA
HNSC
CESC
LUSC
BRCA
LUAD
SARC
UCEC
OV
other
A3 (TCW>K)
nonA3 (VCN>K)
log2(muts / muts in latest−replicating bin)
3
2
1
0
−1
Late <− RT −> early
Required mutations / bin > 10
Local hypermutation type
kataegis
omikli
unclustered
b
NCG
rest
TCW
WAW
Nucleotide composition (log10−ratio to latest bin)
3.00
2.50
2.00
1.50
1.00
0.75
GCG
CCG
TCG
ACG
TCT
TCA
AAA
TAA
AAT
TAT
Late < − > early
c
C>G
C>T
log2(muts / muts bin 1)
1.00
0.75
0.50
0.25
0.00
A3 (TCW>K)
Late <− RT −> early
Required mutations / bin > 10
d
RepliSeq
RNA-Seq
H3K36me3
log2 enrichment
1.00
0.75
0.50
0.25
0.00
A3
Low <− feature −> high
e
non−A3 (VCH>N)
unclustered
A3 (TCW>K)
omikli
A3 (TCW>K)
unclustered
Mutation rate ratio to t1 (log10)
0.05
0.00
−0.05
−0.10
late <− RT −> early
CpG density tertile
t1
t2
t3
f
kataegis (336)
omikli (591)
unclustered (659)
TCW>K enrichment correlation
0.4
0.2
0.0
−0.2
APOBEC3A
APOBEC3B
EXO1
MLH1
MLH3
MSH2
MSH3
MSH6
PMS2
g
−2
−1
0
1
2
kataegis
omikli
unclustered
Number of mutations
1000
100
10
1
1e+03
1e+02
1e+01
1e+00
1e−01
10000
***
**
*
EXO1
MLH1
MLH3
MSH2
MSH3
MSH6
PMS2
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

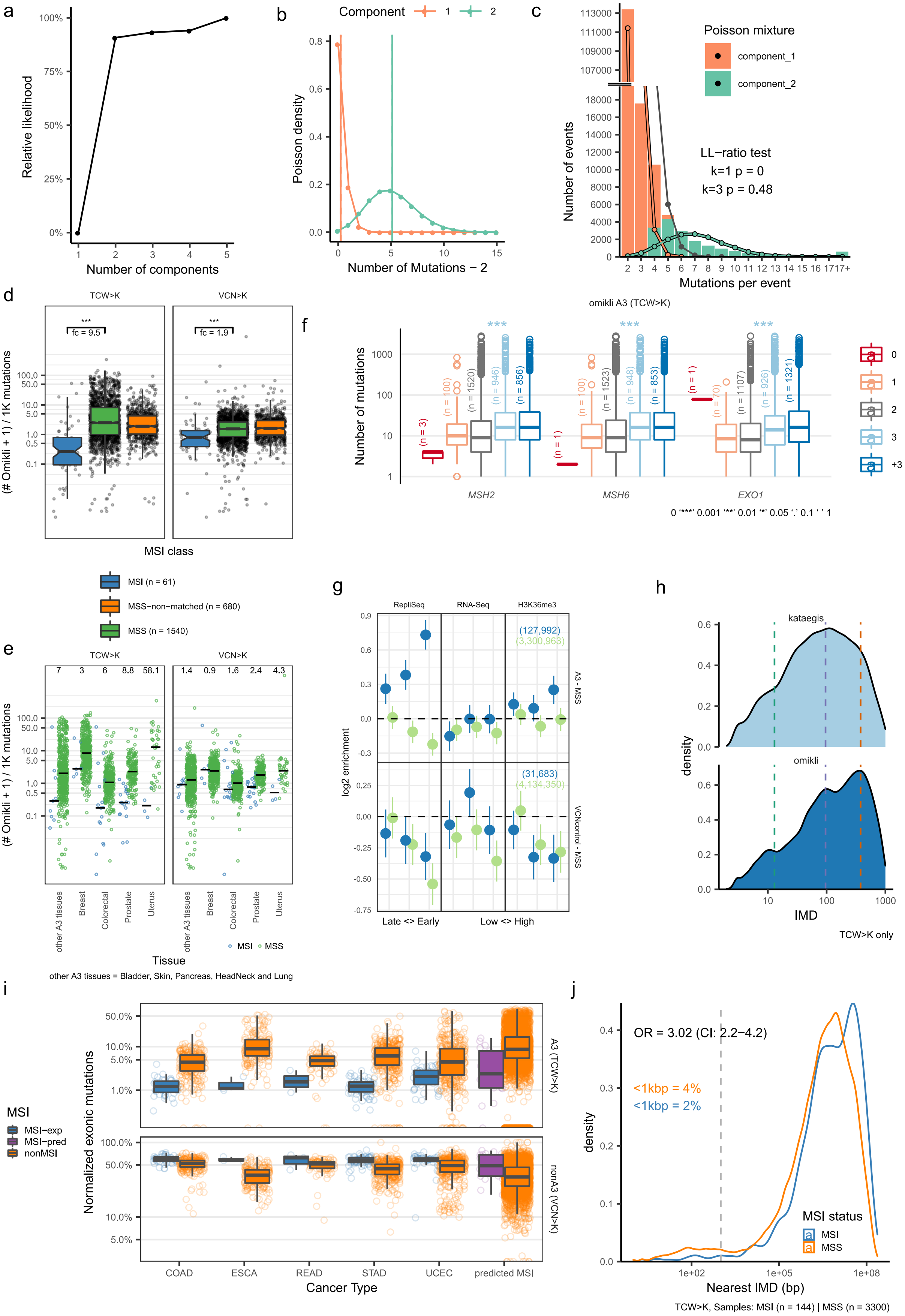
Extended Data Fig. 4: Simulations estimate power to detect mutation clusters and deconvolute their IMD distributions.

a, b, An analysis of somatic hypermutation (SHM) events in lymphoid cancers suggests length of MMR excision tracts in human cells. The distance from the initiating AID mutation (here, WNCYN>N context) to the flanking mutation introduced by error-prone MMR (here, any mutation at a A:T pair) is plotted, in known SHM off-target regions (blue) and, as a control, in intergenic regions (red) (panel a). A statistically significant enrichment is seen in the bins of the distance to central AID mutation (x axis) between 400–1000 nt (panel b). Numbers above/below bars are p-values by Chi-square test on the standardized residuals. c, Gamma mixture modelling of the IMD distributions. Log-likelihood values for different number of components when modelling IMD of the A3 kataegis and omikli mutations. d, The alpha and beta parameters of the three fitted gamma distributions (‘comp.1’, ‘comp.2’ and ‘comp. 3’) approximately match the alpha and beta parameters expected from simulated distributions with IMD at 30 bp, 800 bp and 200 bp, respectively. e, f, Simulations using spiked-in clustered mutations into genomes obtained by randomizing and subsampling mutations from MSI-H hypermutated tumors (panel e) and other hypermutators (panel f), with the goal of determining the recall (or sensitivity; y axis) of recovering mutation clusters at various global mutation burdens (x axis). Dashed line is a loess fit and shaded area is its 95% C.I. Vertical lines are residuals of the fit. g, Difference between MSI and MSS tumor samples in the absolute burden of clustered A3 omikli mutations; significance by Mann-Whitney test (two-tailed).


a
% of mutations
0.150%
0.100%
0.050%
0
500
1000
1500
2000
distance
Intergenic
SHM off-targets
b
standardized residuals
5.0
2.5
0.0
−2.5
0.688 0.035 <0.001 <0.001 0.397
0.288
0.359 0.001 0.002 <0.001
(0,200]
(200,400]
(400,600]
(600,800]
(800,1e+03]
(1e+03,1.2e+03]
(1.2e+03,1.4e+03]
(1.4e+03,1.6e+03]
(1.6e+03,1.8e+03]
(1.8e+03,2e+03]
(SHM off-targets vs intergenic)
c
kataegis
−32500
−35000
−37500
−40000
−42500
omikli
−127750
−128000
−128250
loglik
1
2
3
4
5
number of components
d
alpha
2.0
1.5
1.0
0.5
0.0
beta
150
100
50
0
parameter value
800bp
comp.2
30bp
comp.1
200bp
comp.3
e
MSI−H prediction
Simulations Recall
100%
75%
50%
25%
0%
0
25
50
75
100
Simulated TMB (muts/Mbp)
f
non-MSI hypermutator
Simulations Recall
80%
60%
40%
25
50
75
Simulated TMB (muts/Mbp)
g
n = 602 n = 21 n = 602 n = 20
***
number of omikli mutations
1000
100
10
1
raw
recall corrected
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
MSI
MSS
MSI

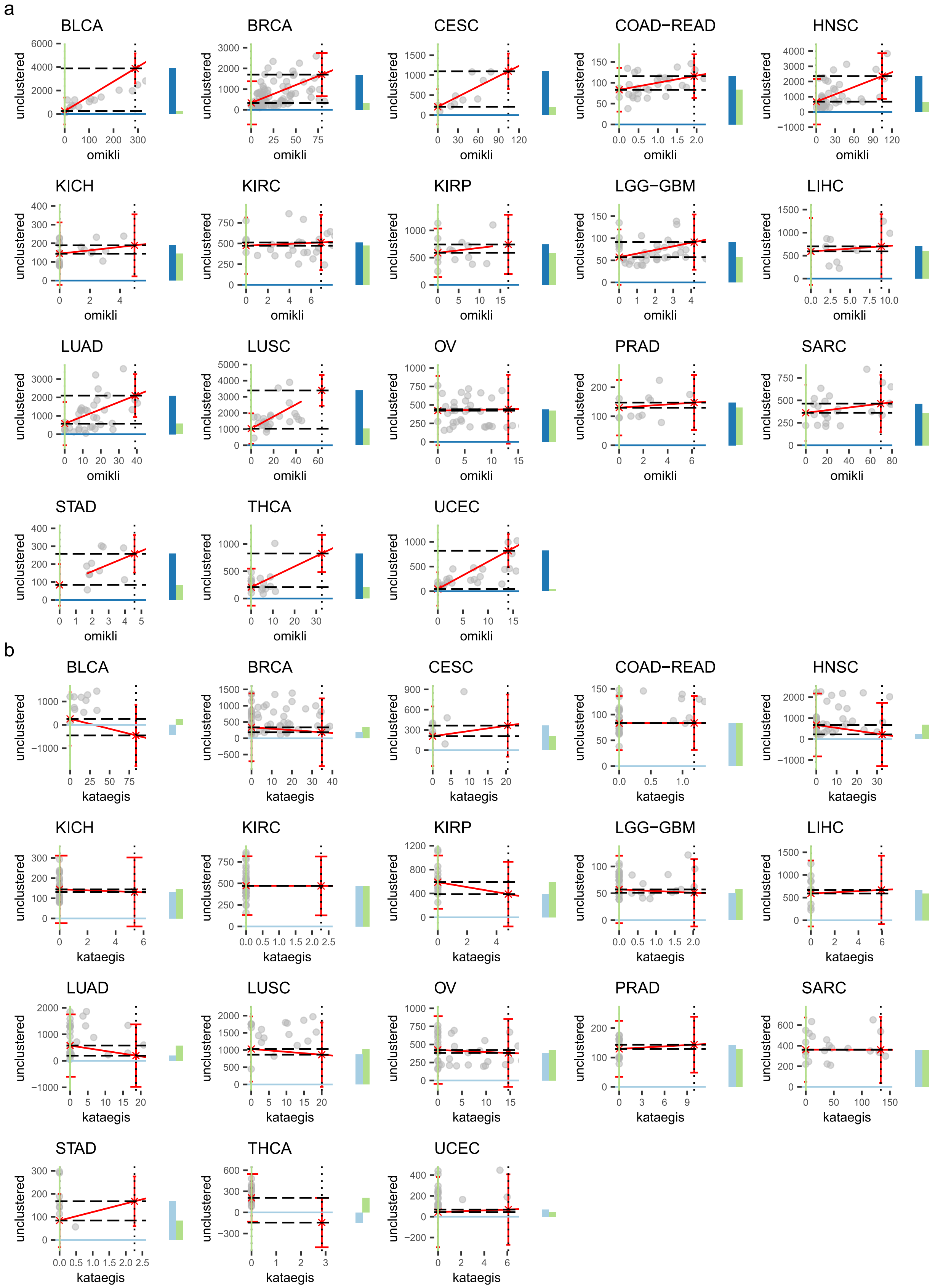
Extended Data Fig. 5: Validation analyses using independent genomic data sets.

a–c, Fitting a Poisson distribution mixture to the number of mutations per cluster in the Hartwig Medical Foundation (HMF) dataset. The near-maximum log likelihood (LL) is obtained with two components (panel c) and the increase to three components is not statistically supported; p-values are from a two-sided bootstrap test. d, e, The relative density of A3 context (left) clustered mutations is higher in MSS (MMR-proficient) than in MSI (MMR-deficient) samples of the same tumor type (left column) in the HMF data. The difference is smaller for the non-A3, control context (right). Significance by Mann-Whitney (two-tailed), n is the number of samples, *** is $p < 0.001$. Numbers show fold-difference between MSS and MSI samples. The 'other A3 tissues' are lung, head-and-neck, skin, pancreas and bladder cancer. f, In HMF data, the A3-context omikli clustered mutations are enriched in tumors with amplified MMR genes; significance by Mann-Whitney test (two-tailed) comparing the neutral (0) versus the gain states (+1 and +2, considered jointly); n is the number of samples. g, In HMF data, A3-context omikli are enriched in early replicating, H3K36me3-marked genomic regions; error bars are 95% C.I. h, Intermutational distance distributions for kataegis (top) and omikli (bottom) A3 context mutations in the HMF data. Dashed lines show peaks of the simulated distributions (Fig. 2) with segment lengths of 25 bp (green), 200 bp (purple) and 800 bp (orange). i, j, Whole-exome sequences in the TCGA data show an excess of A3 context (TCW) mutation fraction in MSS compared to MSI cancers (panel i), and an excess of TCW mutations at distances <1000 bp, normalized to longer distances, in MSS over MSI samples (panel j). 'MSI-exp' (n = 152) denotes the experimentally established MSI-H status while 'MSI-pred' (n = 18) is the MSI status predicted using machine learning (ref. 61), 'nonMSI' (n = 5,661) is neither of these cases.


a
Relative likelihood
100%
75%
50%
25%
0%
1 2 3 4 5
Number of components
b
Component 1 2
Poisson density
0.8
0.6
0.4
0.2
0.0
0 5 10 15
Number of Mutations − 2
c
Poisson mixture
component_1
component_2
LL−ratio test
k=1 p = 0
k=3 p = 0.48
Number of events
113000
111000
109000
107000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 17+
Mutations per event
d
TCW>K
VCN>K
***
fc = 9.5
***
fc = 1.9
(# Omikli + 1) / 1K mutations
100.0
50.0
10.0
5.0
1.0
0.5
0.1
MSI class
MSI (n = 61)
MSS−non−matched (n = 680)
MSS (n = 1540)
f
omikli A3 (TCW>K)
Number of mutations
1000
100
10
1
***
***
***
(n = 3)
(n = 100)
(n = 1520)
(n = 946)
(n = 856)
(n = 1)
(n = 100)
(n = 1523)
(n = 948)
(n = 853)
(n = 1)
(n = 70)
(n = 1107)
(n = 926)
(n = 1321)
MSH2
MSH6
EXO1
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
0
1
2
3
+3
e
TCW>K
VCN>K
7 3 6 8.8 58.1
1.4 0.9 1.6 2.4 4.3
(# Omikli + 1) / 1K mutations
100.0
50.0
10.0
5.0
1.0
0.5
0.1
other A3 tissues
Breast
Colorectal
Prostate
Uterus
MSI
MSS
Tissue
other A3 tissues = Bladder, Skin, Pancreas, HeadNeck and Lung
g
RepliSeq
RNA-Seq
H3K36me3
log2 enrichment
0.9
0.6
0.3
0.0
-0.3
0.25
0.00
-0.25
-0.50
-0.75
(127,992)
(3,300,963)
(31,683)
(4,134,350)
A3 - MSS
VCNcontrol - MSS
Late <> Early
Low <> High
h
kataegis
omikli
density
0.6
0.4
0.2
0.0
10
100
1000
IMD
TCW>K only
i
MSI
MSI−exp
MSI−pred
nonMSI
Normalized exonic mutations
50.0%
10.0%
5.0%
1.0%
100.0%
A3 (TCW>K)
nonA3 (VCN>K)
COAD
ESCA
READ
STAD
UCEC
predicted MSI
Cancer Type
j
OR = 3.02 (CI: 2.2−4.2)
<1kbp = 4%
<1kbp = 2%
density
0.4
0.3
0.2
0.1
0.0
1e+02
1e+05
1e+08
Nearest IMD (bp)
MSI status
MSI
MSS
TCW>K, Samples: MSI (n = 144) | MSS (n = 3300)

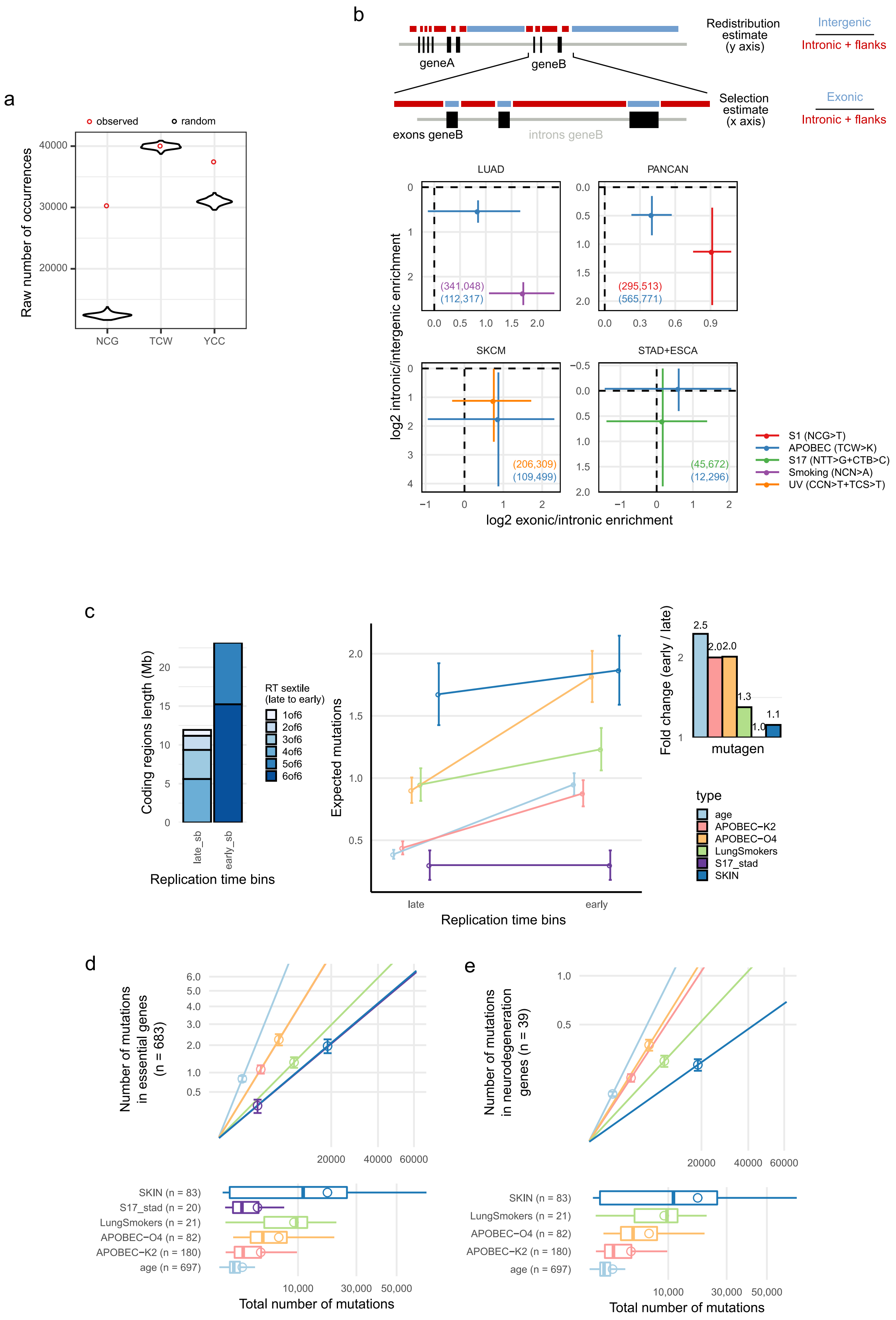
Extended Data Fig. 6: Contribution of the omikli and the kataegis mechanisms to the unclustered A3 mutation burden in various tissues.

a, The omikli mechanism generates many unclustered mutations (‘A3-O’) in various cancer types. b, The kataegis mechanism generates comparatively few unclustered mutations (‘A3-K’). Panels show the fit (red line) of the unclustered A3 burden (y axis) to the clustered A3 burden (x axis), (see Methods). Error bars are 95% prediction intervals at x=0, and at x = mean burden of A3 clustered mutations for that cancer type. Horizontal dashed lines are the predicted numbers of unclustered A3 mutations at those two points (for clarity also shown in blue/green bars next to each plot). Fits use robust regression (rlm function in R). For visual clarity, only the part of the plot up to the mean of unclustered mutation burden plus a margin is shown, however the fit uses all data points (that is tumor samples) including ones not visualized.


a
BLCA
BRCA
CESC
COAD−READ
HNSC
KICH
KIRC
KIRP
LGG−GBM
LIHC
LUAD
LUSC
OV
PRAD
SARC
STAD
THCA
UCEC
unclustered
omikli
b
BLCA
BRCA
CESC
COAD−READ
HNSC
KICH
KIRC
KIRP
LGG−GBM
LIHC
LUAD
LUSC
OV
PRAD
SARC
STAD
THCA
UCEC
unclustered
kataegis

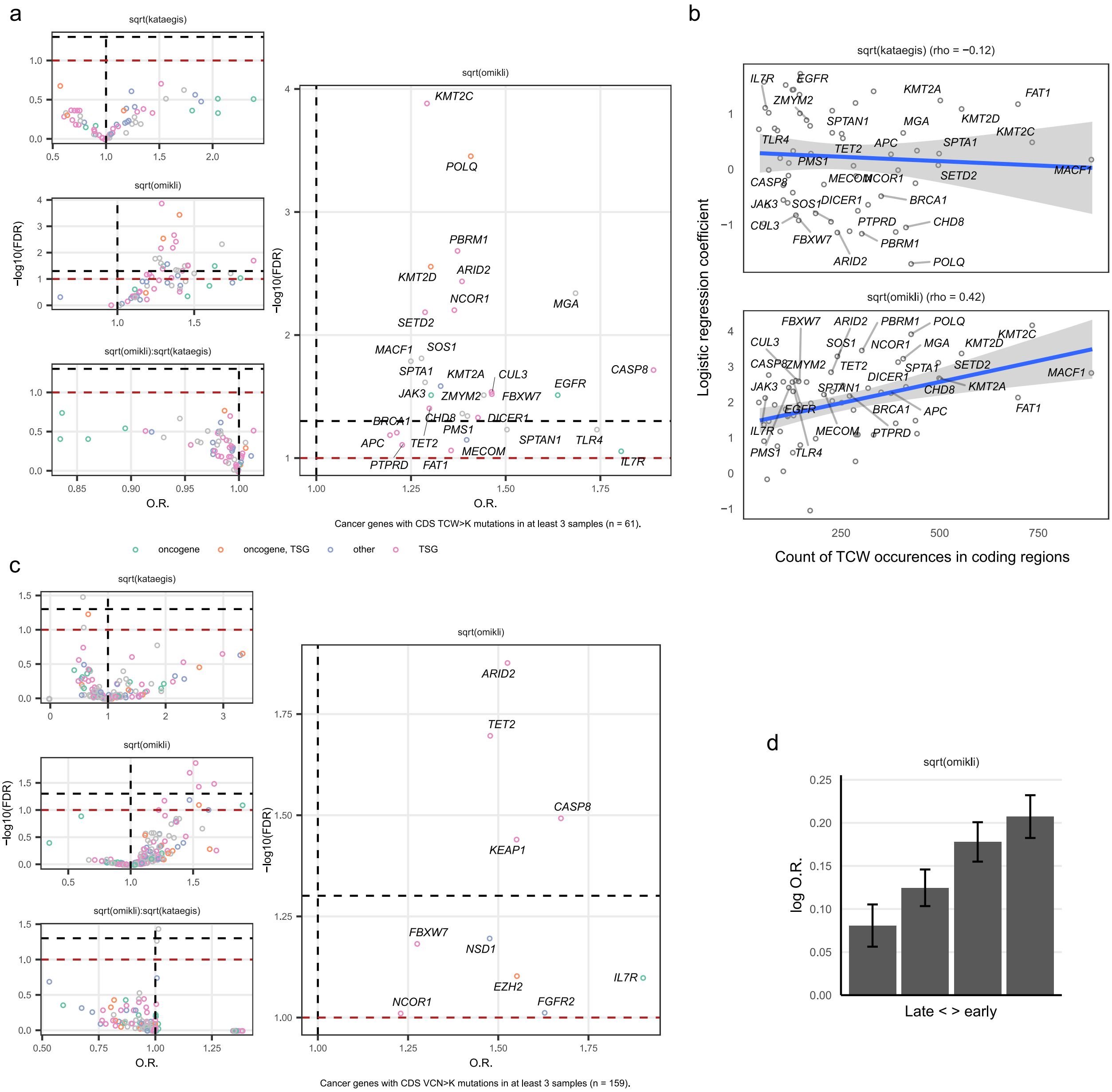
Extended Data Fig. 7: Mechanisms underlying A3 clustered mutations generate many impactful changes, affecting disease genes.

a, Coding regions in the human genome are enriched for CpG dinucleotides (NCG), but not with the A3-context TCW trinucleotides, compared to random expectation. b, Enrichment of mutations in exons versus introns (estimate of selection strength, x axis) and the enrichment in intergenic regions versus introns (estimate of redistribution of mutations towards regions containing genic DNA, y axis; flipped). The comparison of mutagenic agents against APOBEC was performed for selected tissues, matching the relevant tissue with the particular mutagen (tumor samples listed in Supplementary Table 7). Error bars are 95% C.I. from negative binomial regression; numbers in parenthesis are the tally of mutations. c, The differential functional impact of the tested mutagens across replication time (RT) bins. Left: total length of coding sequences (CDS) in the late and early RT bins, shaded by the RT sextiles that were merged to create the two bins (where 1 is the latest and 6 is the earliest RT). Middle: expected number of cancer gene CDS-affecting mutations in an average tumor sample (same sets of samples, genes and mutations as in Fig. 5a; y axis) for the late versus early RT bin (x axis), for various mutagens (colors); error bars are s.e.m. Right: fold-difference between the functional impact at the late versus early bin, for various mutagen types. d, e, The functional impact density (FID) of various mutational processes in a set of cell-essential genes (panel d) and neurodegenerative disease-associated genes (panel e). Slope shows the fraction of impactful genetic changes i.e. those affecting the CDS of at least one gene in the set. Points show the expected number of impactful changes resulting from a mutational process, on average, in a tumor genome affected by that mutational process. Error bars are s.e.m. ‘APOBEC-O4’ is A3 mutagenesis in omikli-rich tumors. ‘APOBEC-K2’ is A3 mutagenesis in kataegis-rich tumors.


a
observed
random
Raw number of occurrences
40000
30000
20000
NCG
TCW
YCC
b
geneA
geneB
Redistribution estimate (y axis)
Intergenic
Intronic + flanks
exons geneB
introns geneB
Selection estimate (x axis)
Exonic
Intronic + flanks
LUAD
PANCAN
(341,048)
(112,317)
(295,513)
(565,771)
SKCM
STAD+ESCA
(206,309)
(109,499)
(45,672)
(12,296)
log2 intronic/intergenic enrichment
log2 exonic/intronic enrichment
S1 (NCG>T)
APOBEC (TCW>K)
S17 (NTT>G+CTB>C)
Smoking (NCN>A)
UV (CCN>T+TCS>T)
c
Coding regions length (Mb)
RT sextile (late to early)
1of6
2of6
3of6
4of6
5of6
6of6
late_sb
early_sb
Replication time bins
Expected mutations
late
early
Replication time bins
Fold change (early / late)
2.5
2.0
2.0
1.3
1.0
1.1
mutagen
type
age
APOBEC−K2
APOBEC−O4
LungSmokers
S17_stad
SKIN
d
Number of mutations in essential genes (n = 683)
6.0
5.0
4.0
3.0
2.0
1.0
0.5
20000
40000
60000
SKIN (n = 83)
S17_stad (n = 20)
LungSmokers (n = 21)
APOBEC−O4 (n = 82)
APOBEC−K2 (n = 180)
age (n = 697)
10,000
30,000
50,000
Total number of mutations
e
Number of mutations in neurodegeneration genes (n = 39)
1.0
0.5
20000
40000
60000
SKIN (n = 83)
LungSmokers (n = 21)
APOBEC−O4 (n = 82)
APOBEC−K2 (n = 180)
age (n = 697)
10,000
30,000
50,000
Total number of mutations

Extended Data Fig. 8: Associations between genic mutations and global burden of clustered mutations.

a, Associations between A3-context TCW>K mutations in coding regions of each cancer gene, and the global burden of A3 kataegis (top left) or omikli (middle left) and their interaction term (bottom left). Right panel is same as middle-left panel, but showing only the significant genes, with labels. Volcano plots show logistic regression coefficients (transformed to odds ratio) on the x axis and the log FDR on the y axis. Genes that bore coding mutations in at least three tumor samples were tested. b, Number of TCW sites in a gene coding sequence (CDS; x axis) predicts the association of cancer gene mutations (y axis) with A3 omikli burden (bottom) but not with A3 kataegis burden (top). Error bands are 95% C.I. of the linear fit. c, Same association analysis as panel a but for the control, non-A3 context VCN>K mutations in the gene CDS. d, Early RT cancer genes are more affected by A3 mutagenesis. Cancer genes were stratified into RT quartiles (x axis) and logistic regression coefficient (log odds ratio, y axis) linking A3 omikli burden with the presence of a mutation in the CDS of any cancer gene in that RT bin was determined. Error bars are 95% C.I. from logistic regression (on n=593 tumor samples).



Cancer genes with CDS TCW>K mutations in at least 3 samples (n = 61).

Cancer genes with CDS VCN>K mutations in at least 3 samples (n = 159).

Chapter 4

Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type

The following chapter has been selected from the paper:

Franco, Irene; Helgadottir, Hafdis T.; Moggio, Aldo; Larsson, Malin; Vrtačnik, Peter; Johansson, Anna; Norgren, Nina; Lundin, Pär; Mas-Ponte, David; Nordström, Johan; Lundgren, Torbjörn; Stenvinkel, Peter; Wennberg, Lars; Supek, Fran; Eriksson, Maria. ”Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type” *Genome Biology* 20.1 (2019): 1-22.

The published document can be accessed at:

https://doi.org/10.1186/s13059-019-1892-z

Genome Biology

RESEARCH Open Access

Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type

Irene Franco[1*†], Hafdis T. Helgadottir[1†], Aldo Moggio[2], Malin Larsson[3], Peter Vrtačnik[1], Anna Johansson[4], Nina Norgren[5], Pär Lundin[1,6], David Mas-Ponte[7], Johan Nordström[8], Torbjörn Lundgren[8], Peter Stenvinkel[9], Lars Wennberg[8], Fran Supek[7,10] and Maria Eriksson[1*]

Abstract

Background: The lifelong accumulation of somatic mutations underlies age-related phenotypes and cancer. Mutagenic forces are thought to shape the genome of aging cells in a tissue-specific way. Whole genome analyses of somatic mutation patterns, based on both types and genomic distribution of variants, can shed light on specific processes active in different human tissues and their effect on the transition to cancer.

Results: To analyze somatic mutation patterns, we compile a comprehensive genetic atlas of somatic mutations in healthy human cells. High-confidence variants are obtained from newly generated and publicly available whole genome DNA sequencing data from single non-cancer cells, clonally expanded in vitro. To enable a well-controlled comparison of different cell types, we obtain single genome data (92% mean coverage) from multi-organ biopsies from the same donors. These data show multiple cell types that are protected from mutagens and display a stereotyped mutation profile, despite their origin from different tissues. Conversely, the same tissue harbors cells with distinct mutation profiles associated to different differentiation states. Analyses of mutation rate in the coding and non-coding portions of the genome identify a cell type bearing a unique mutation pattern characterized by mutation enrichment in active chromatin, regulatory, and transcribed regions.

Conclusions: Our analysis of normal cells from healthy donors identifies a somatic mutation landscape that enhances the risk of tumor transformation in a specific cell population from the kidney proximal tubule. This unique pattern is characterized by high rate of mutation accumulation during adult life and specific targeting of expressed genes and regulatory regions.

Keywords: Somatic mutations, Aging, Kidney cancer, Proximal tubule, kidney progenitors

Background

Over a lifetime, the human body is vulnerable to a vast number of mutagenic forces that collectively lead to loss of genome integrity and subsequently cellular aging and cancer initiation [1]. Sequencing studies have revealed genetic variations among cells within an individual, referred to as "somatic variance." This information can be used to study the genome evolution during the lifespan of an individual [2] and outline specific mutagenic processes that promote the transition from a normal to a cancer cell [3]. Variants that are exclusively detected in the clonal-cell population of a tumor are believed to represent the mutations that occurred in the cell prior to the initiation of cancer [4] and are widely used to study mutational processes in normal tissues. However, inherent within cancer clones are characteristics (increased genomic instability and selective advantage), which can

* Correspondence: irene.franco@ki.se; Maria.Eriksson.2@ki.se
† Irene Franco and Hafdis T. Helgadottir contributed equally to this work.
[1] Department of Biosciences and Nutrition, Center for Innovative Medicine, Karolinska Institutet, Huddinge, Sweden
Full list of author information is available at the end of the article

present a conundrum in understanding the etiology of somatic mutations in normal tissues. The elimination of confounding factors can be achieved by studying mutations in *non-cancerous cells*, thus allowing a direct assessment of genomic changes occurring with typical aging of organ systems. Whole genome sequencing (WGS) of a high number of single cells would be the most informative method. However, there are technical challenges associated with single-cell WGS and these have impeded massive analysis of somatic variance in normal cells [5, 6]. An alternative strategy is the bulk sequencing of non-cancer human tissues [7–10]. This approach provides only selected variants, i.e., variants contained in the genome of cells that clonally expanded in the normal tissues and contributed a detectable number of copies. But, similar to what observed for cancer, detectable variants may not be fully representative of the common mutational processes. In addition, bulk data are not ideal for analyses that compare the frequency of mutations in specific genomic regions or for exploring the non-coding portion of the genome [7–10]. It is possible to obtain WGS data relative to a single genome while avoiding single cell sequencing. This method requires in vitro clonal expansion of a single cell prior to sequencing, and a specific processing of data, in order to select the somatic variants that were present in vivo and eliminate those that occurred during culture [2, 6]. This strategy has some limitations. For example, it is necessarily restricted to cells that are able to proliferate in vitro (e.g., stem/progenitor cells or reprogrammed cells), and the culturing procedure is demanding and not suitable for the analysis of a large number of cells. Despite these limitations, the strategy has been successfully applied to the analysis of skeletal muscle progenitors [11]; intestine, colon, and liver stem cells [12]; blood stem and progenitor cells [13, 14]; and reprogrammed skin fibroblasts [15].

Results generated from clonally expanded, normal cells demonstrate that aging is correlated with a linear increase of somatic mutations and specific mutation patterns and distributions. These features appear very consistent among different cells of the same tissue, even when obtained from different individuals. Therefore, despite the low number of genomes analyzed per tissue, important general conclusions regarding the rate of occurrence and the main features of somatic mutations have been drawn for skeletal muscle, liver and intestinal stem cells, and blood cells during aging [11, 12, 14]. Importantly, information can be gleaned from these data and used to build an understanding of cellular and genomic activities prior to the appearance of mutations. A catalogue of somatic mutations can be deconstructed into distinct components or mutational signatures, through non-negative matrix factorization (NMF) [16]. In multiple cases, mutational signatures obtained through the analysis of thousands of cancer genomes have efficiently been attributed to a specific etiology [17] (http://cancer.sanger.ac.uk/cosmic/signatures). This is the case of signature 7, which is found predominantly in cancers derived from the skin and is consistent with the chemical modifications of DNA expected after sunlight UV exposure [17]. Unfortunately, the mechanisms underlying other signatures remain unknown. For example, the single base substitution signature (SBS)40 was recently separated from signature 5 and shown to induce a large number of mutations in cancer samples, especially those derived from the kidney [18]. While the etiology of signature 5 seems to be related to uncorrected errors [19, 20], the etiology of SBS40 is unexplored. Another strategy to identify the mutagens that shape a given genome is to study regional differences in the distribution of somatic mutations [21]. Genomic features that determine the non-random localization of mutations are (1) DNA replication timing [22], (2) chromatin organization [11, 23, 24], and (3) the levels of active transcription [25]. Consequently, these features influence DNA exposure to both extrinsic (genotoxic compounds and radiations) and intrinsic (DNA synthesis and repair mechanisms) mutagens [21–23, 25] and are thought to be dependent on the organ or tissue. Taken together, it is the current belief that the development of somatic mutations in healthy tissues occurs as tissue-specific somatic mutagenesis [12, 14, 17, 26].

The findings derived from our atlas of somatic mutations in healthy tissues do not support a simple association of each tissue to a specific somatic mutation pattern. In contrast, we identify a stereotypical, mutational pattern across progenitor cells from a variety of tissues and two distinct mutation profiles in the same tissue portion, indicating that mutagen exposure is modulated by multiple factors in addition to tissue type. In particular, we identify cell differentiation state and cell-type-specific activities as critical determinants of mutagenesis. Importantly, our high coverage WGS data allowed us to define that the landscape of somatic mutations in different cell types is different in terms of mutational signatures, but also genomic distribution of mutations. Our analyses, based on single genome data from the kidney, skin, subcutaneous, and visceral fat cells from healthy donors, and complemented with a meta-analysis of somatic mutations from healthy ($N = 161$) and tissue-matched cancer genomes ($N = 192$), identify a unique mutation pattern in a population of proximal tubule (PT) cells. This population expresses the distinguishing markers of a PT cell type previously identified as the cell of origin of the most common kidney cancer subtypes [27]. Its unique mutation pattern is characterized by high rate of mutation acquisition during adult life and mutation enrichment in regulatory regions and expressed genes, ultimately resulting in a higher risk of a transition to cancer. Overall, our work constitutes the proof of principle for exploiting somatic mutation data from healthy cells to tailor cell-type-specific approaches of cancer prevention.

Results

Detection of mutations in different tissues from the same individual

To explore differences in mutagenic processes occurring in adult human tissues, we analyzed the somatic variation in human kidney tubules (KT), epidermis (EP), and subcutaneous and visceral adipose tissue (SAT and VAT, respectively) from healthy individuals of different ages. These tissues are subjected to extensive morphological changes during aging, including loss of regenerative potential and atrophy in the case of kidney tubules, epidermis, and subcutaneous fat and progressive hypertrophy in the case of visceral fat [28, 29]. Genomic alterations, for example those connected with premature-aging syndromes, have been associated to kidney, skin, and fat changes [30–32], and our analysis aims to better establish a link between loss of genome integrity and specific morphological modifications in these tissues.

Genomic data were obtained by WGS of single cells freshly isolated from tissue biopsies and clonally expanded in vitro (Fig. 1a). This strategy allowed the survey of ~ 92% of the genome at a minimum coverage of 15x and the discovery of somatic mutations present in the single cell at the moment of isolation from the tissue. A stringent filtering on the allele frequency (AF), allowing only variants with AF comprised between 0.4 and 0.6, efficiently discarded somatic variants acquired during in vitro culture (see the "Methods" section). A well-controlled comparison of tissue-specific differences was achieved through the analysis of cells derived from multiple tissues from the *same* individual (Fig. 1a, b). Multi-tissue biopsies were obtained from three living, kidney donors of younger age (30, 31, 38 years) and three donors of older age (63, 66, 69 years). Characteristics of the donor pool were as follows: (1) provided an extensive, clinical evaluation before surgery; (2) no history of cancer, only two donors reported forms of benign hyperplasia that are very common in the population; (3) a body mass index ranging from 20 to 30 kg/m^2; and (4) normal kidney function (Additional file 1: Table S1A). None of the donors carried a genetic predisposition to cancer, according to our analysis of germline mutations in 47 known cancer genes (Additional file 1: Table S1B).

Specific cell types were cultured from all tissues tested: kidney tubule cells from the kidney, pre-adipocytes from fat, and keratinocytes from the skin (Additional file 1: Figure S1). Cells were sequenced only if they were able to attach and proliferate as a colony for 17–20 divisions (Additional file 1: Table S1C). Based on these unique properties of colony formation and long-term proliferation, we named our samples as *progenitors* from KT, EP, SAT, and VAT.

Our newly generated data comprises a total of 69 single genomes (Fig. 1b, Additional file 1: Table S1D). From one donor (a 69-year-old woman), we obtained multiple, progenitor clones from four tissues. From the other individuals, we sequenced multiple KT clones and, in most cases, also multiple SAT and VAT clones (Fig. 1b). The sequencing data yielded information on single nucleotide variants (SNVs) and small insertion/deletions (InDels) (Additional file 1: Table S1D and Additional file 2) that were validated using a technical replicate. The validation rate was 99 and 97% for SNVs and InDels, respectively (Additional file 1: Table S1E). This validation confirmed that our pipeline could recover a set of high-confidence somatic variants and exclude variants that occurred during cell culture, as demonstrated in our previous publication [11]. The false-negative rate is also expected to be the same (0.41) [11].

The data have been used in either tissue- or age-focused analyses in order to explore both the tissue-specific differences of somatic mutation accumulation and the age-related genome modifications common among tissues (Fig. 1b).

The tissue of origin of a cell is not the only determinant of the somatic mutation profile

To understand somatic mutagenesis in different tissues, we compared the spectrum of somatic mutations recovered in each sample. Somatic SNVs were organized in 96 classes based on the type of base substitution and its trinucleotide context. This classification yielded a somatic mutation profile that was used to cluster samples (Fig. 2a). As expected, EP samples, rich with UV-induced C > T transitions, separated from all the others (first cluster to the left). Unexpectedly, the other samples did not cluster according to the tissue of origin, but created two main subgroups. The largest group (right) included all SAT and VAT clones and some of the KT samples (KT1). The other cluster (center) consisted of the remaining KT samples (KT2; 54% of KT clones). All but one biopsy showed the concomitant presence of KT1 and KT2 cells (Fig. 2b). The KT2-mutation profile characterized all the clones with the highest numbers of

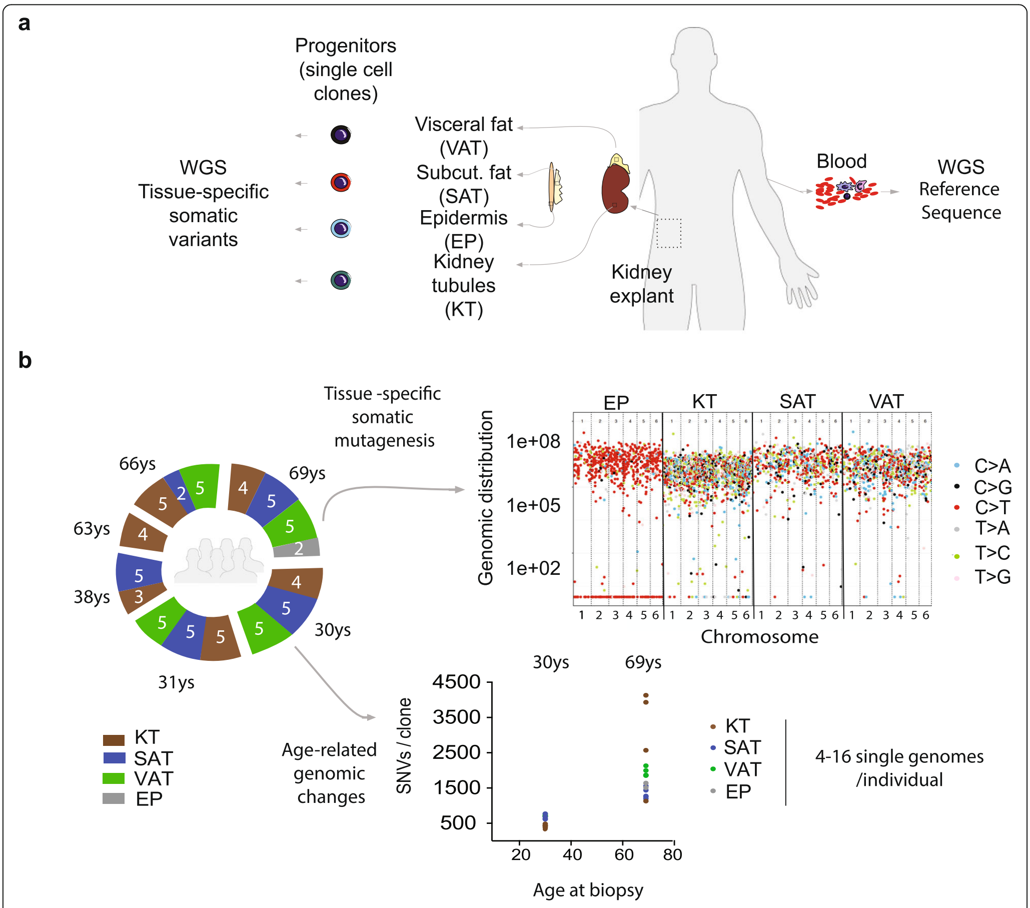


Fig. 1 Somatic mutation detection in single genomes from different tissues of the same individual. **a** Experimental strategy for single genome analysis of progenitor cells from multiple tissues from the same healthy individual. Blood, kidney, subcutaneous fat (SAT), visceral fat (VAT), and skin biopsies were obtained from living kidney donors undergoing surgery. The blood tissue was whole genome sequenced (WGS) as a bulk to obtain the individual's reference sequence. The kidney tubule (KT) and epidermis (EP) portions were separated from the kidney and skin biopsies, respectively. Single progenitor cells were isolated from KT, SAT, VAT, and EP and clonally expanded in culture to obtain WGS data. These data were filtered using the individual's reference sequence to obtain the catalogue of somatic variants for every clone. **b** Schematic summary of sequenced samples and analysis strategy. Two to five single genomes per biopsy were sequenced (white numbers in the round plot) from six individuals of either younger (30–38) or older (63–69) age. KT progenitors were sequenced for all six individuals, while SAT, VAT, and EP progenitors were sequenced in a subset of the donors. Somatic mutation data were used to study either the tissue or the age effect on mutation accumulation. An example of tissue-related differences found in the study is provided (top right): somatic SNVs found in 4 clones from different tissues of the same individual were plotted according to their genomic position and in different colors according to the type of base substitution. An example of age-related changes is provided (bottom right): total amount of SNVs in the genome of each sequenced clone from two selected individuals of either younger (30 years) or older (69 years) age

variants, both SNVs and InDels (Fig. 2c, d, respectively). In agreement, KT2 clones showed higher, yearly increase of mutations (56.6 SNVs and 8.0 InDels per genome per year), compared to the other cell types (KT1 clones 11.7 SNVs and 1.4 InDels; SAT 17.5 SNVs and 0.9 InDels; VAT 27.2 SNVs and 1.4 InDels) (Fig. 2e, f).

In summary, we identify a stereotyped mutation spectrum in multiple, different tissues (KT, SAT, VAT)

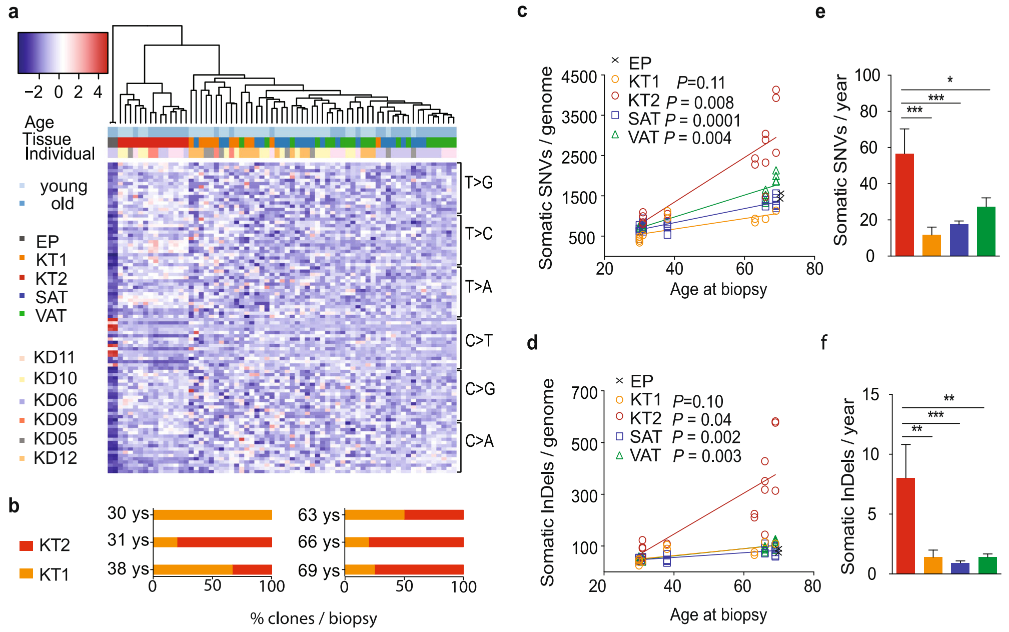


Fig. 2 Clustering of samples on the base of mutation types defines similarities between different tissues and two subsets of KT cells. **a** Mutation pattern of 69 single genomes obtained from different human tissues of six healthy individuals of either younger (30–38) or older (63–69) age (horizontal). SNVs were subdivided in 96 classes based on the single base substitution types and their trinucleotide context (vertical) and the relative amount of mutations for each class were plotted as a heatmap. Hierarchical clustering of the samples based on the mutation pattern is shown on top of the heatmap. **b** Percentage of kidney-tubule-derived cells clustering in the KT1 or KT2 subset per biopsy. Each biopsy is defined by the age of the donor (30 years $N = 4$; 31 years $N = 5$; 38 years $N = 3$; 63 years $N = 4$; 66 years $N = 5$; 69 years $N = 4$ clones). **c**, **d** Number of somatic single nucleotide variants (SNVs, **c**) and small insertions/deletions (InDels, **d**) found in single genomes of multiple progenitors from 6 individuals of different ages. (*x* axis) The numbers of somatic variants per clone were normalized to the percentage of autosomes covered by the sequencing. Linear regression curves and *P* values calculated with the linear mixed models are shown for each tissue. **e**, **f** Average yearly increase of somatic SNVs (**e**) and InDels (**f**) per tissue. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, one-way ANOVA and multiple comparisons tests. EP epidermis, KT1 kidney tubule 1, KT2 kidney tubule 2, SAT subcutaneous fat, VAT visceral fat

and two distinct spectra in the same tissue (KT1 and KT2), suggesting that the tissue of origin is not the main determinant of somatic mutation accumulation in this sample set.

An atlas of somatic mutagenesis in healthy tissues distinguishes basal and mutagen-driven processes

In order to build a more comprehensive atlas of somatic mutation landscapes in human tissues, we extended our analysis to public datasets of somatic mutations from WGS of clonally expanded non-cancer cells. The cell types in this meta-analysis include skin fibroblasts (SkinFB) [15]; stem cells from the liver, intestine, and colon [12]; and progenitor cells from skeletal muscle (SkM) [11] and blood [13] (Additional file 1: Table S2). A total of 92 genomes were analyzed, in addition to our 69 genomes, and the samples subjected to unsupervised clustering on the base of their trinucleotide spectra (Fig. 3a). The groups defined in our initial clustering (Fig. 2a) were mostly maintained. Interestingly, the cluster including cells from multiple tissues (KT1, SAT, VAT) was confirmed and two more cell types, the SkM and blood progenitors, overlapped with it in the center of the plot. This cluster was called the “common progenitors” (Fig. 3a).

To understand the main factors driving the sample clustering (Fig. 3a), mutational signatures were analyzed (Fig. 3b–d and Additional file 1: Figure S2–S5). To increase the power, the WGS of 192 tissue-matched tumor samples were analyzed along with the 161 healthy samples (Additional file 1: Table S2). Eight signatures were obtained by NMF and named after the most similar, single base substitution (SBS) signature from the catalogue of signatures observed in cancer [18] (Additional file 1: Figure S2). The relative exposure of each signature in different normal and cancer types was analyzed in order

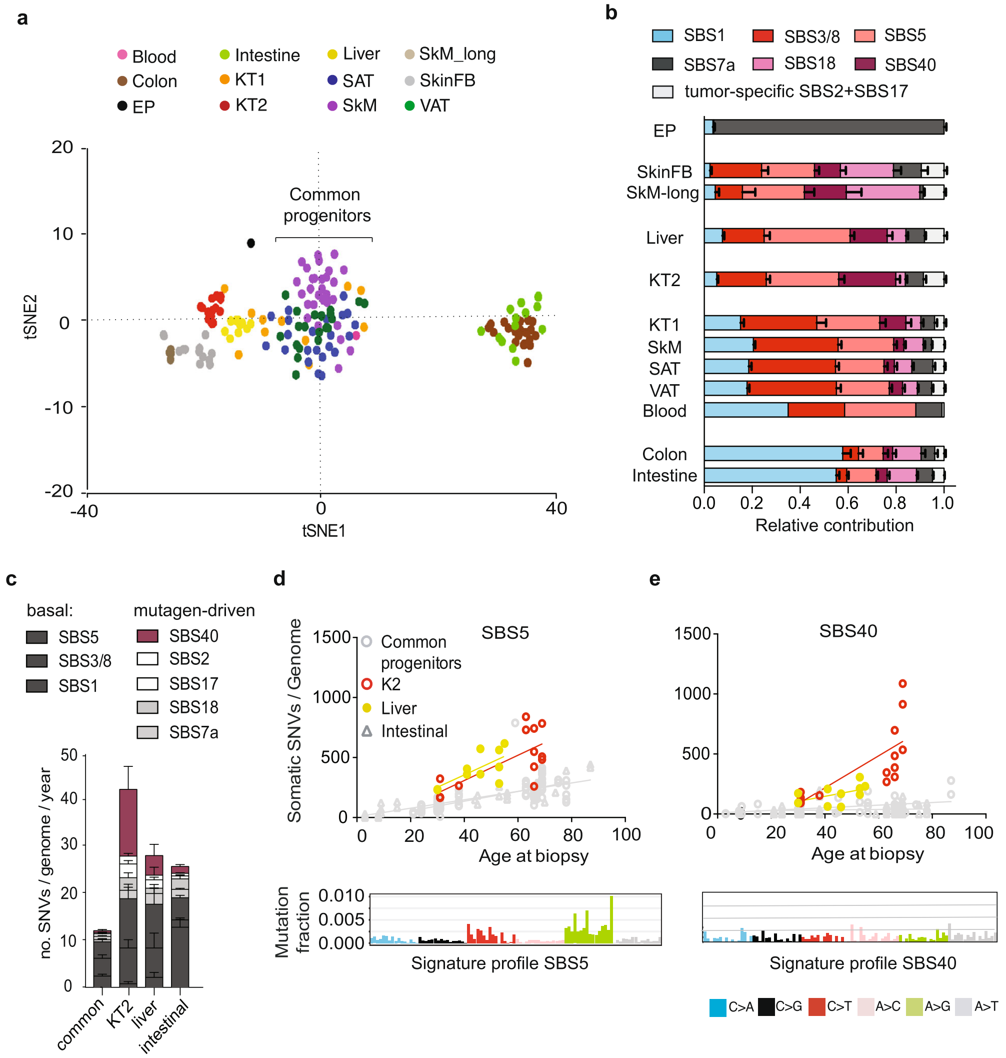


Fig. 3 Meta-analysis of somatic mutation data from healthy donors defines basal and mutagen-driven mutagenesis in adult tissues. Sixty-nine single genomes from epidermis (EP), kidney tubule 1 (KT1), kidney tubule 2 (KT2), subcutaneous fat (SAT), and visceral fat (VAT) were analyzed together with public datasets of somatic mutations from WGS of clonally expanded non-cancer cells, including skin fibroblasts (SkinFB) [15]; liver, intestine, and colon stem cells [12]; skeletal muscle progenitors (SkM) [11]; and blood progenitors [13]. **a** tSNE plot of the trinucleotide profile of somatic SNVs. Multiple tissues displaying a common mutation profile (SkM, SAT, VAT, KT1, and blood) were named "common progenitors." **b** Relative contribution of the eight mutational signatures identified in healthy cells via non-negative matrix factorization. Each signature was named after the most similar single base substitution (SBS) signature from [18]. **c** Average yearly increase of somatic SNVs obtained by linear fit of mutations with age in the common progenitors, KT2, liver stem cells, and intestinal stem cell (intestine and colon) groups. *P* values from linear mixed models are shown in Additional file 1: Table S3a. **d**. **e** Linear increase of mutations with age and signature profile of SBS5 (**d**) and SBS40 (**e**) in KT2 (red), liver (yellow), and common progenitors and intestine-derived (colon and intestine stem cells) samples (gray). SBS5 and SBS40 showed similar profiles (bottom), but different tissue distribution

to identify cell types with significantly higher exposure to specific signatures (Additional file 1: Figure S3 and Table S3). Two signatures, SBS2 (APOBEC) and SBS17b, appeared largely tumor-specific in the sample set examined here and were found at high levels in sparse cancer genomes and at negligible levels in healthy samples (Additional file 1: Figure S3). Apart from these signatures, the somatic mutation profiles found in cancer samples broadly supported the results found in the corresponding healthy samples (Additional file 1: Figure S3 and S4a).

Overall, our analysis shows that signatures SBS1, 3/8, and 5 were found ubiquitously (Additional file 1: Figure S3) and linearly increased with age (Additional file 1: Table S4). The common progenitors (SAT, VAT, KT1, SkM, and blood) presented the lowest yearly increase of mutations among the cell types analyzed, and the majority of these mutations could be attributed to SBS1, SBS3/8, and SBS5 (Fig. 3c). These evidences suggest that the signature combination comprised of SBS1, SBS3/8, and SBS5 is the unavoidable product of core cellular processes. Therefore, we define it as "basal mutagenesis." Consistent with this concept, cell types that were not common progenitors had higher exposure to additional signatures that are associated with specific, mutagen exposure. Examples are (1) EP samples showing high levels of SBS7a, a signature induced by UV light exposure, and (2) the SkM cells used as a control for culture-induced mutagenesis in our previous study [11] (SkM-long), which showed SBS18, a signature linked to in vitro culture stress [20, 33] and consequent production of intracellular reactive oxygen species [34] (Fig. 3b). These samples were used as positive controls for prolonged exposure to a mutagen.

KT2 and liver stem cells generated two specific clusters, adjacent to each other (Fig. 3a). This similarity matched the higher rate of age-related accumulation of SBS5 seen in KT2 and liver samples (Fig. 3d). However, this increase did not seem to be the consequence of a major defect of nucleotide excision repair (NER) [19] because SBS5 was 15-fold lower in liver and KT2 cells compared to our positive controls for NER deficiency, the *ERCC2*-null tumors (Additional file 1: Figure S4b-c). In contrast to SBS5, SBS40 increased with aging mainly in KT2 cells (Fig. 3c, e). Among analyzed samples, SBS40 was stronger in KT2 and two types of kidney cancer, clear cell and papillary renal cell carcinomas (KIRC and KIRP, respectively) (Additional file 1: Figure S3). Like KT2, these tumor types demonstrated a rise in SBS40 with aging (Additional file 1: Figure S4d-e), suggesting that signature SBS40 is the result of a mutagen active in the kidney. Interestingly, the chromophobe subset of kidney carcinoma (KICH) and KT1 showed low SBS40 contribution (Additional file 1: Figure S3 and S4d-e), indicating that only specific subsets of kidney cells are exposed to the mutagenic process eliciting this signature. To obtain insight into possible mutagens active in these cells, the mutation profiles of 161 normal and 192 tissue-matched tumor samples were compared to the spectrum induced by 53 genotoxic compounds in a clonal population of iPSCs [33]. The spectrum of mutations found in KT2 and kidney tumors KIRC and KIRP (Additional file 1: Figure S5b) was similar to that generated by exposure to formaldehyde and alkylating agents, suggesting that these specific cell types in the kidney might be exposed to these mutagens, more likely derived by endogenous chemical reactions [35].

Taken together, results indicate that a group of cells from different tissues (common progenitors) provide a model of minimal mutagenesis, which we named "basal mutagenesis." Relative to these cells, all other cell types show signs of exposure to additional extrinsic (UV light in EP, in vitro culture stress in SkM-long), intrinsic (high SBS1, probably caused by higher proliferation rate in intestinal stem cells [12]), or endogenously produced (KT2) mutagens.

KT2 are damaged cells from the proximal tubule

To better understand mutagen exposure in KT cells, the similarities between normal kidney cells and different subsets of kidney cancer were further explored. A comparison of somatic mutation profiles showed that KT1 cells did not overlap with any kidney cancer type, but were intermixed with the common progenitor group (Fig. 4b). Conversely, the KT2 mutational profile was similar to KIRPs and KIRCs and very distant from the distal-tubule-derived KICH (Fig. 4b). The different subsets of kidney tumors show specific genetic, epigenetic, and transcriptional profiles [27, 36, 37], due to their origin from distinct cell types within the kidney (Fig. 4a). KIRCs and KIRPs originate from the proximal tubule (PT) [27, 36], where the epithelial layer is exposed to a continuous flow of potentially mutagenic compounds either reabsorbed from or excreted into the urine (Fig. 4a). A specific population of epithelial cells from the convoluted PT (named PT1) was recently identified as the more likely precursor of ccRCC and pRCC tumors on the base of scRNA seq data [27]. Given the similarities between KT2 and ccRCC/pRCC at the somatic mutation level, we hypothesized that KT2 clones may overlap with the PT1 population and tested the expression of a number of markers by FACS and qPCR (see the "Methods" section and Table 1). Selected KT1 and KT2 clones were tested and found positive for markers of kidney progenitors, while most markers of differentiated cells were not expressed, suggesting that both populations are in an

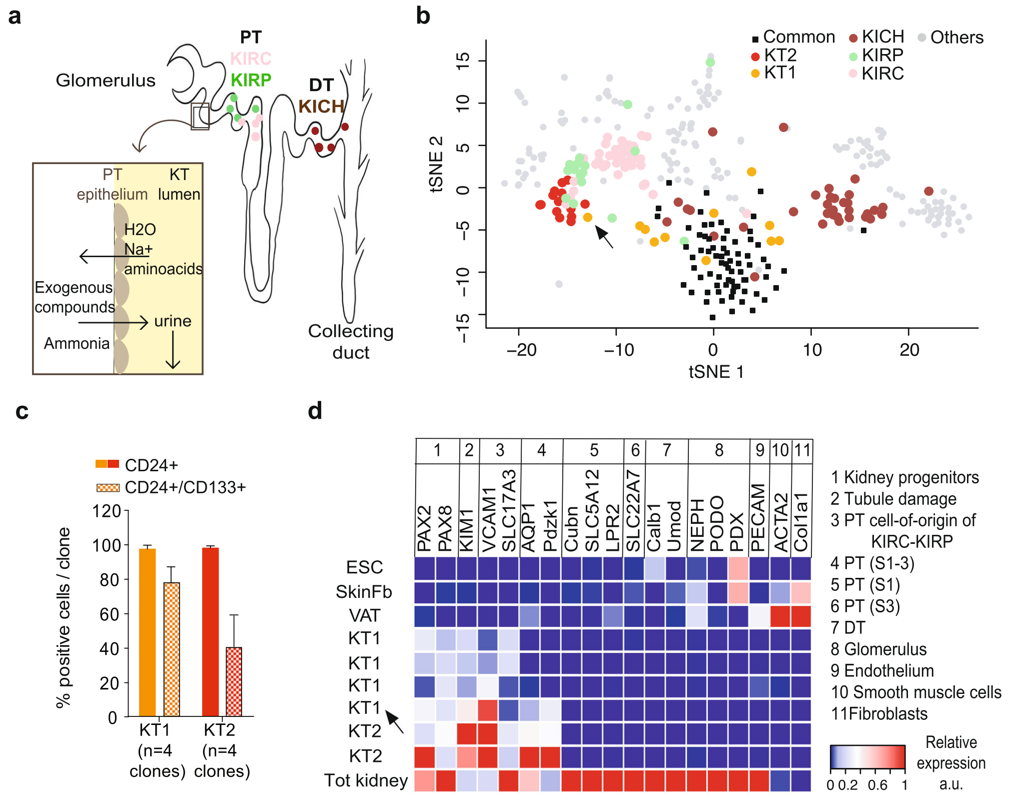


Fig. 4 KT2 cells are proximal tubule cells exposed to mutagens. **a** Cartoon representing a kidney nephron and the location of the different tumor samples included in the analyses (according to [27, 36]). A section of proximal tubule (PT) is enlarged to show the trafficking of water, solutes, and other compounds across the PT epithelium. **b** tSNE plot of the trinucleotide profile of somatic SNVs in healthy (*n* = 161) and tumor (*n* = 192) samples. The common progenitors (SAT, VAT, SkM, and blood) and kidney-derived healthy and tumor genomes are highlighted with specific colors, while all other samples are shown in gray. **c** FACS analysis of the kidney progenitor markers CD133 and CD24 in selected KT1 and KT2 clones (*n* = 4). The average percentage of double- or single (CD24)-positive cells per clone is shown. **d** Heatmap showing the relative expression of markers of undifferentiated and differentiated kidney cells in single clones (subdivided in 11 categories described in the legend on the right) from either the KT1 (*n* = 4) or the KT2 (*n* = 2) group, tested by qPCR. Human embryonic stem cells (ESC bulk) and skin fibroblasts (SkFB bulk) were included as negative controls, together with a VAT clone. RNA extracted from a fresh kidney biopsy was included as positive control. The same KT1 clone is marked with an arrow in **b** and **d**, to highlight its intermediate KT1/KT2 phenotype at both somatic mutation (**b**) and gene expression (**d**) levels. KT1 and KT2, healthy kidney-tubule-derived cells; KIRC, clear cell renal cell carcinoma; KIRP, papillary renal cell carcinoma; KICH, chromophobe renal cell carcinoma; PT, proximal tubule; DT, distal tubule; S1, first segment of PT, convoluted; S3, last segment of PT, straight

undifferentiated state. Despite this, KT2 also expressed VCAM1/CD106 and SLC17A3, the markers that define the PT1 population found by Young et al. In addition, KT2 expressed AQP1 and PDZK1, two PT markers, and KIM1, a marker of tubule damage. The same markers were absent or expressed at lower levels in KT1 clones, except for a clone that showed a mutation spectrum very close to KT2 and alkylating agent exposure (marked with an arrow in Fig. 4a, d; Additional file 1: Figure S5b). Overall, these data suggest that KT2 cells can originate from the PT1 population, but are found in a less differentiated state. Indeed, our cell culture procedure selects for proliferating cells and KT epithelial cells are known to reacquire proliferative capacities after de-differentiation in response to tubule damage [38]. Conversely, the KT1 population expression profile is overall consistent with a previously characterized population of scattered kidney tubule progenitors [39].

Somatic mutagenesis in the kidney proximal tubule predisposes to the acquisition of driver mutations

Tumors derived from the PT (KIRC and KIRP) constitute the vast majority of tumors diagnosed in the kidney (Fig. 5a) [40], supporting the hypothesis that somatic mutagenesis in the PT favors tumorigenic transformation. Since KT2 are non-cancer clones from the PT of healthy kidneys, we studied these cells as a model of mutagenesis in the PT, prior to cancer initiation.

First, we confirmed that KT2 were not cancer clones at the moment of isolation from the tissue by analyzing the possible presence of the genetic lesions that commonly drive cancer initiation in KIRC and KIRP [41]. KT2 showed lower mutation burden compared to KIRC and KIRP (Fig. 5b) and did not display the typical kidney cancer genetic lesions, nor mutations in *TP53*, a tumor suppressor often mutated in pre-cancer clones in human tissues [7, 8, 10] (Additional file 1: Table S5). Yet, the mutation burden in cells from 63- to 69-year-old donors was higher in KT2 compared to other kidney cells (KT1; Fig. 5b) and the specific mode of somatic mutation accumulation in the PT could facilitate the acquisition of driver mutations and ultimately promote tumor initiation.

Kidney tumors are very rare at 30 years of age, but the incidence increases constantly and peaks in the 8th decade of life [40]. To model driver mutations, we selected the somatic mutations predicted to have a functional effect on a gene that is actually expressed in the tissue of origin. We defined these variants as potentially pathogenic mutations and determined their age-related increase (Fig. 5e, f). KT2 cells acquired higher numbers of potentially pathogenic mutations compared to other cell types from the same donors (KT1-SAT-VAT, Fig. 5e, f). The yearly increase was 5.7-fold higher in KT2 compared to KT1-SAT-VAT (Fig. 5f). From these data, we estimate that each PT cell accumulates an average of 86.5 potentially pathogenic mutations by the age of 70. A higher rate of accumulation of potentially pathogenic mutations makes the acquisition of cancer driver mutations in PT cells a more likely event compared to other cell types. These data are in agreement with the overall higher somatic mutation burden in KT2 (Fig. 2c-f). However, we also noticed that the mutation load in introns and exons of transcribed genes was higher than expected by random distribution and higher compared to non-expressed introns and exons (Fig. 5d). Conversely, the other cell types from the same donors (KT1-SAT-VAT, Fig. 5 d and Additional file 1: Figure S6) showed mutation depletion in these regions, in agreement with previous reports [11, 12]. Similarly, conserved regions were protected from mutations in KT1-SAT-VAT and enriched in KT2 (Fig. 5e). Finally, KT2 showed a particularly strong enrichment of mutations in regulatory regions (Fig. 5d). Overall, our somatic mutation analysis of non-cancer cells points to substantial differences in the genomic distribution of mutations depending on the cell of origin. These differences make specific cell types more vulnerable to the acquisition of mutations that affect the function of important genes, and this feature correlates with increased chances of a transition to cancer.

Different efficiency of DNA repair in cells exposed to basal mutagenesis or additional mutagens

The regional pattern of distribution of mutations across the genome is shaped not only by mutagen exposure, but also by DNA repair. In fact, transcribed DNA is generally depleted of mutations due to the activity of the transcription-coupled NER (TC-NER) [25, 42]. In addition, mismatch repair (MMR) more efficiently protects from mutations the early-replicating and H3K36me3-rich DNA [21, 43]. Transcribed genes are usually located in early-replicating and H3K36me3-rich chromatin and benefit of both high TC-NER and MMR activities. Specific alterations in the pattern of regional differences of mutation accumulation are signs of TC-NER and MMR defects [21, 25, 42–44]. Therefore, we analyzed these patterns in our catalogue of healthy genomes.

Figure 6a shows the specific contribution of early/late DNA replication timing (RT), abundance of H3K36me3 marks, and transcription levels to the enrichment/depletion of mutations in different cell types. The group of common progenitors, including SAT, VAT, SkM, and blood, but not KT1, showed the expected depletion of mutations with earlier RT, higher H3K36me3 abundance and higher transcription levels (Fig. 6a and Additional file 1: Figure S7a-b). This pattern indicates that the basal mutagenesis is actively counteracted by MMR and/or TC-NER. However, EP, KT2, KT1, liver, SkM-long, and SkinFB deviated from the pattern seen for common progenitors and showed a loss of association of mutation rates with RT and H3K36me3 (Fig. 6a and Additional file 1: Figure S7c).

KT2 showed a severely affected RT and H3K36me3 pattern (Fig. 6a), thus suggesting that many mutations escaped MMR activity. While an increased proportion of InDels compared to SNVs in KT2 genomes was consistent with MMR defects (Fig. 6b), no evidence of a classical form of microsatellite instability (MSI) was detectable (Fig. 6c). These data suggest that some form of MMR is likely operative in these cells. Interestingly, KT2 were the only cell types displaying higher amounts of mutations in highly transcribed regions, while in all other cell types transcription protected from mutations

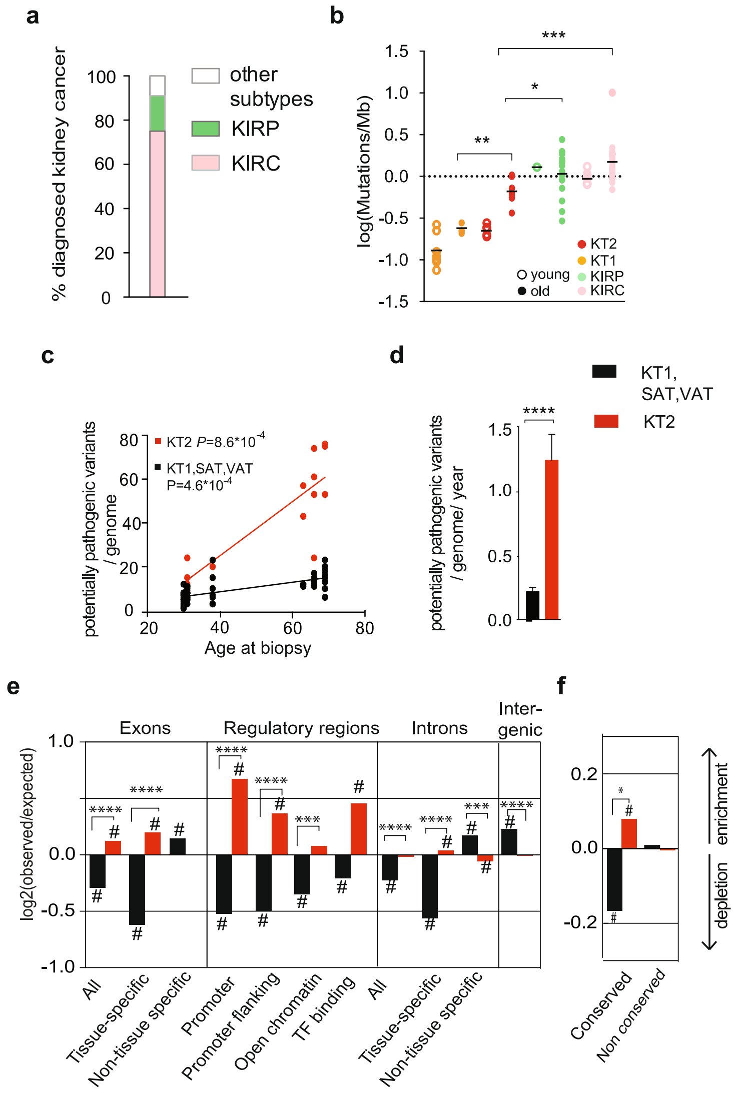


Fig. 5 (See legend on next page.)

(See figure on previous page.)
Fig. 5 Kidney PT shows a unique somatic mutation pattern that confers high risk for tumor transformation. **a** Epidemiologic data showing the percentage of kidney tumors either derived from the proximal tubule, such as KIRC (clear cell renal cell carcinoma) and KIRP (papillary cell renal cell carcinoma), or from other kidney structures (other subtypes). **b** Somatic mutation burden in KT1, KT2, KIRP, and KIRC of either a younger (30–40) or older (60–70) age range. Significance among older groups was measured by one-way ANOVA. **c**, **d** Linear fit with age (**c**) and yearly increase (**d**) of potentially pathogenic variants in KT2 vs KT1-SAT-VAT clones. Potentially pathogenic variants are defined as follows: all variants were annotated with CADD (Combined Annotation Dependent Depletion; https://cadd.gs.washington.edu/). SNVs and InDels predicted to affect the coding sequence (presenting CADD score > 15) were selected and subsequently filtered on expression data in order to select only variants affecting a gene actually expressed in the tissue of origin of the clone. Tissue-specific and non-tissue-specific genes correspond to the expressed and non-expressed genes in the corresponding tissue according to the Human Protein Atlas (http://proteinatlas.com). Adjusted *P* values of the linear fit are calculated with the linear mixed model (**c**) or two-sided *t* test (**d**). **e** Enrichment (upward bars) or depletion (downward bars) of somatic mutations in indicated genomic features. The log2 ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region. Log2 = 0 corresponds to a number of observed mutations equal to the number expected by random distribution. **f** Enrichment (upward bars) or depletion (downward bars) of somatic mutations in conserved and non-conserved regions of the genome. $^{\#}P < 0.05$, one-sided binomial test. $^{***}P < 0.001$, $^{****}P < 0.0001$ two-sided *t* test of log2 ratios for either KT2 or KT1-SAT-VAT in specified genomic regions. EP epidermis, KT1 kidney tubule 1, KT2 kidney tubule 2, SAT subcutaneous fat, VAT visceral fat

(Fig. 6a, right). This suggests that a transcription-coupled mutagenic process [45] may be active in KT2 cells, supported by a striking, altered pattern of transcription-strand asymmetry of the different substitution types (Fig. 6d).

Overall, these results indicate a mechanism in cells that are exposed only to basal mutagenesis for sparing early-replicating-, H3K36me3-rich and highly transcribed regions from mutations. This occurs in diverse tissue types and is consistent with previous evidence of a more efficient activity of MMR and NER pathways directed towards active chromatin [22, 42]. In cells putatively exposed to a mutagen (EP, KT2, KT1, liver, SkM-long, and SkinFB), the altered, mutation-depletion pattern suggests that NER- and/or MMR-mediated protection is not as effective. KT2 cells show a unique pattern of mutation distribution that explains the higher mutation rate in transcribed genes (Fig. 5e).

Aging affects the efficiency of MMR and NER

Finally, we focused on non-tissue-specific effects of aging. Chromosomal instability is known to increase with age in normal tissues [2, 46]. Sequencing data from the 69 genomes from KT, SAT, VAT, and EP samples from 6 healthy kidney donors and 29 SkM progenitor genomes from 7 healthy donors from [11] were used to detect large chromosomal aberrations (Additional file 1: Table S6). These aberrations were recovered in three different tissues, i.e., skeletal muscle, VAT, and kidney tubules (both KT1 and KT2 cell types), but only in association with aging (Fig. 7a, b), supporting a general age-related increase of chromosomal instability.

The number of SNVs and InDels per genome also increased in all surveyed tissues with aging (Fig. 2c, d). To explore whether an age-related decline in DNA repair could contribute to somatic mutation accumulation, we selected cell types showing the more effective MMR and NER activities (Fig. 6a and Additional file 1: Figure S7a-c) and analyzed differences in mutation distribution and spectra in different age groups. Older genomes showed a weakened association of mutations with RT compared to younger ones, indicating a partial loss of MMR activity (Fig. 7c and Additional file 1: Figure S8a). The effect size of this defect was approximately one third of that observed in tumors with known MMR loss (MSI-H) (Additional file 1: Figure S8b), suggesting that aged, healthy cells acquire an early-stage mutator phenotype. MSI tumors were also found to lack mutations in binding sites for CTCF and Cohesin, in agreement with the requirement of a functional MMR to produce mutations at these sites [47]. Relative amount of mutations at CTCF/Cohesin peaks was lower in old vs young genomes. This result constitutes a further proof in support of a partial defect of MMR activity in old cells.

To investigate if defects extend to other pathways, we analyzed the age-related increase of SBS5, known to be associated with NER inactivation [19]. Results show that the fraction of SBS5 mutations per genome increases with age progression (Fig. 7d). This age-related expansion was specific for SBS5 and not detectable for the other ubiquitous signatures SBS1 and SBS3/8 (Additional file 1: Table S3b); this supports the hypothesis that NER weakens with advancing age. In summary, evidence demonstrates the decline of both MMR and NER in the genome of healthy cells as they age. This phenomenon is conserved across different tissues and occurs in cells that did not show genomic evidence of exposure to extrinsic mutagens.

Discussion

We present here the basis of a somatic mutation atlas that can systematically guide the identification of cancer-prone cell types and high-risk somatic mutation processes. This collection exclusively includes whole genome data and high-confidence somatic variants obtained from single human cells, clonally expanded in vitro. Our newly generated data from the kidney, epidermis, subcutaneous fat, and visceral fat are based on

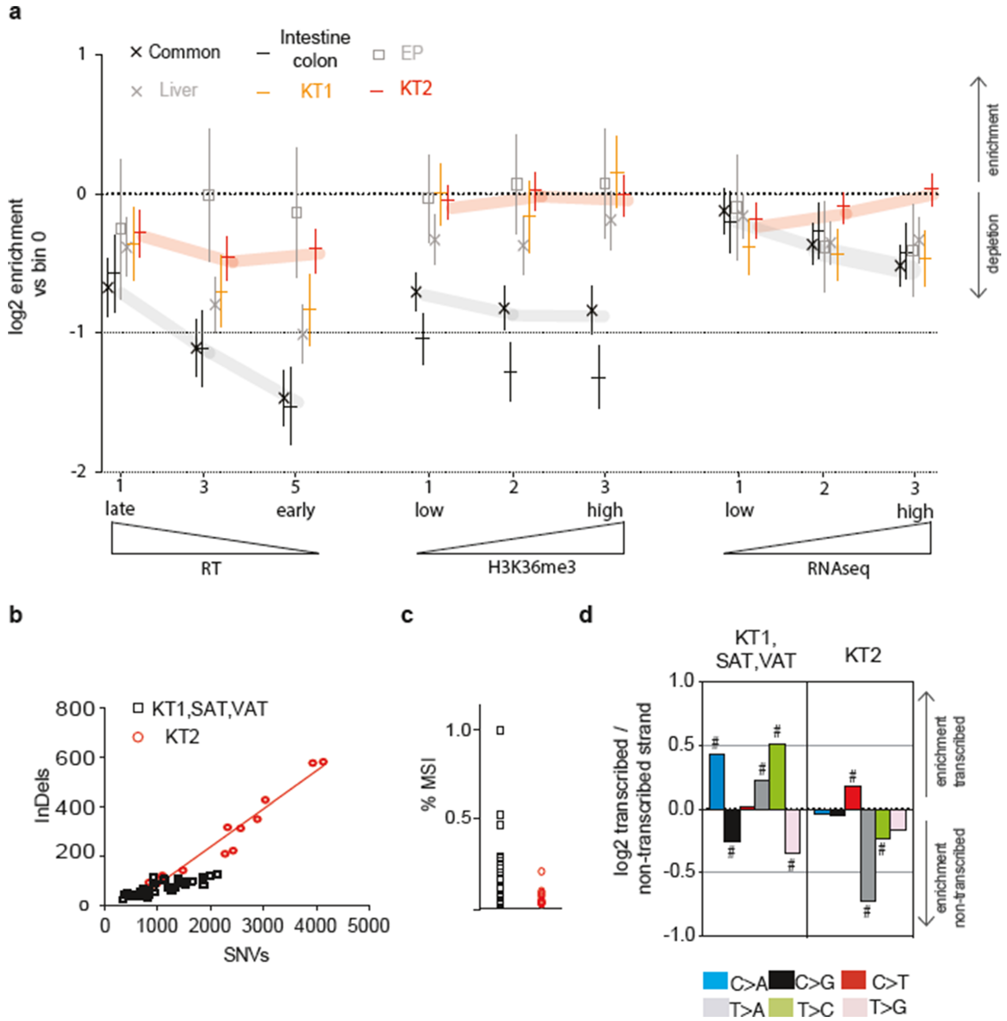


Fig. 6 Mutation enrichment in specific genomic regions provides information on DNA repair efficiency and mutagen exposure in different cell types. **a** Enrichment/depletion of mutations in specific genomic regions. The genomes were divided in multiple sectors (bins) according to decreasing DNA replication time (RT, bins 0 to 5. For clarity, only bins 1, 3, and 5 are shown), increasing abundance of the histone mark H3K36me3 (bins 0–3), and increasing transcriptional levels (RNA-seq, bins 0–3). The relative abundance of mutations in each bin vs bin 0 for every tissue (EP, liver, KT1, KT2) or tissue group (common progenitors: SAT, VAT, SkM, blood; intestine-colon) is estimated as the coefficient in negative binomial regression (expressed as $\log_2$), where error bars show its 95% C.I. **b** Linear regression of SNVs and InDels per genome in the KT2 vs KT1-SAT-VAT group. **c** Percentage of sites subjected to microsatellite instability (MSI) in each genome of either the KT2 or the KT1-SAT-VAT group. **d** Enrichment of the six classes of substitution types in either transcribed or non-transcribed strand of genes. The log2 ratio of the number of observed and expected point mutations indicates the effect size of the enrichment in the transcribed (upper) or non-transcribed (lower) strand. $^{\#}P < 0.05$, one-sided binomial test

samples derived from multiple tissues from the same individual. This strategy provides the advantage of a reliable comparison of tissue-specific differences, excluding the variability derived from different genetic backgrounds and environmental exposure. Newly generated data are complemented and compared with publicly

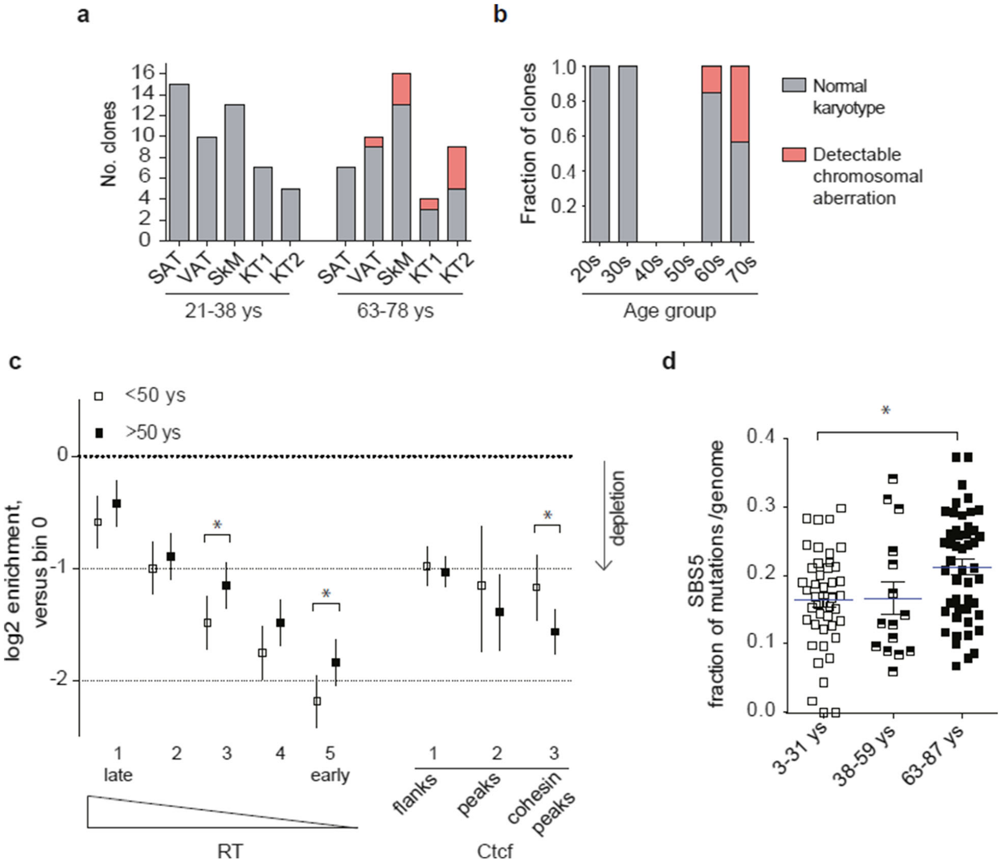


Fig. 7 Genomic instability and weakening of DNA repair with aging. **a** Number of clones showing large chromosomal aberrations per tissue and age group. Young 21–38, old 63–78. **b** Fraction of genomes showing large chromosomal aberrations in the samples analyzed in **a**, but divided in tighter age groups (10 year-span). **c** Enrichment/depletion of mutations according to DNA replication timing (RT) while controlling for CTCF binding sites in either younger (< 50 years old, $N = 52$) or older (> 50 years, $N = 54$) genomes from the tissues not showing signs of exposure to external mutagens (SkM, SAT, VAT, intestine, and colon, according to the analyses in Figs. 3 and 6). Enrichments are coefficients from negative binomial regression (as log2), and error bars are their 95% C.I. Significance of young-vs-old differences was tested via a *Z*-test on the interaction term between age and replication time bin **d**. Fraction of SBS5 mutations per genome in different age groups of SkM, SAT, VAT, blood, intestine, and colon cells. **P* < 0.05, one-way ANOVA and multiple comparison tests

available data sets from either healthy donors [11–13, 15] or tissue-matched cancer samples from TCGA and ICGC, for a final catalogue of 353 genomes and 12 different healthy cell types.

The comparison of somatic mutation landscapes in different cell types enables the identification of cells more susceptible to somatic mutagenesis and consequent cancer initiation [3]. This knowledge is expected to promote significant therapeutic advantages, including more targeted and efficient means of cancer prevention [3]. A major result of our analysis is recognizing that mutagen exposure can be very different even within the same tissue, and this correlates with different susceptibility to cancer initiation. It is possible that analysis of great numbers of genomes will uncover the concomitant presence of multiple cell subsets showing distinct mutation spectra in most tissues. We provide here the proof of principle by characterizing two populations of proliferating cells residing in the kidney tubule, one likely derived from de-differentiated epithelial cells of the proximal tubule (PT) and the other presenting features of undifferentiated kidney tubule progenitors. The somatic mutation spectrum of PT-derived cells presents unique characteristics that could not be identified in any other kidney or non-kidney cell. PT-derived cells showed the highest yearly increase of mutations among the cell types analyzed and a high incidence of the signature SBS40. The only samples that showed similar levels of SBS40 were kidney cancers derived from the PT, namely the clear cell and papillary cell RCCs (KIRC and KIRP, respectively). This analogy suggests that there is a specific process ongoing in the kidney PT and this process underlies the signature SBS40. Unfortunately, the etiology of this signature has not yet been determined. However, the extensive screening of cancer samples that identified SBS40 highlighted its predominance in kidney cancer [18]. Nonetheless, high levels of this signature have also been found in sporadic cases of tumors derived from multiple tissues, including the lung, skin, esophagus, bladder, head, intestine, stomach, liver, and ovary carcinoma, thus supporting the hypothesis that the mutagen causing SBS40 is more common, but not exclusively present in the kidney [18]. PT cells also displayed a unique distribution of mutations across the genome. The regions that are commonly spared from mutations as a consequence of more intense MMR and NER activity [21, 25, 42, 43] presented equal or higher mutation load compared to the rest of the genome. In particular, highly transcribed genes were enriched of mutations and the distribution of the different substitution types on the transcribed and non-transcribed strand was altered. These data indicate not only inefficient DNA repair, but also the presence of a mutagenic process that is more active on transcribed DNA. An important consequence of this unique mutation pattern was a mutation enrichment in functional genes and an age-related accumulation of high-risk mutations that was 5.7-fold faster in PT cells, compared to other cells from the same individuals. We estimated the presence of 86 mutations altering the protein sequence of expressed genes in every PT cell of 70-year-old individuals. Absolute numbers and other estimates of age-related increase of mutations presented in this work will be more accurate when a larger number of cells, distributed along the whole spectrum of ages, are analyzed. In addition, our numbers are certainly an underestimation, since our somatic mutation detection has a false-negative rate of 0.41 and does not allow the detection of all the variants present in a clone. However, our estimates support a strong acceleration in the appearance of pathogenic mutations in the genome of PT-derived cells. Mutations in the non-coding portion of the genome are also expected to affect the function of the cell, and we detected an enrichment of mutations in regulatory regions which is expected to significantly impact on overall gene expression. The high-risk somatic mutation landscape that we describe in PT cells predicts an elevated rate of tumorigenic transformation in this portion of the nephron. In agreement, somatic mutagenesis is recognized as a major tumorigenic mechanism in the kidney [41, 48, 49] and the PT-derived tumors KIRC and KIRP constitute up to 95% of all cancers diagnosed in this organ [36, 40]. Therefore, our analysis points to PT cells as a cell type at particularly high risk of tumor transformation. A clear understanding of the underlying mutational mechanisms can be exploited to slow down mutation accumulation and kidney cancer incidence.

The comparison of mutational profiles observed in healthy cells with the landscape of mutations observed after in vitro exposure to common mutagens [33] provides interesting hypotheses about the mutagens active in the kidney PT. The genomic modifications observed in healthy PT cells or tumors derived from the PT were similar to those induced by formaldehyde and alkylating agents [33]. Alkylating agents used in [33] are chemotherapeutic drugs, such as 1,2-dimethylhydrazine and diethyl sulfate. The healthy kidney donors from which cells were isolated were never treated with those agents nor exposed to formaldehyde. Therefore, we hypothesize that the mutation spectrum might be due to the action of endogenously produced compounds that interact with the DNA in a similar way as the synthetic drugs [35]. Indeed, the epithelial layer of the kidney PT presents a complex chemical environment that is the consequence of ongoing physiological activities, such as ammonia production and excretion, amino acid reabsorption and modification, and transformation and excretion of xenobiotics [50]. Further analyses might support a link between the presence of these compounds in the kidney PT and enhanced mutagenesis in this specialized epithelium.

The kidney PT is an example of particularly high and specific mutagen exposure. However, our analysis also found cell types that are broadly protected from mutagens and constitute a model of minimal or “basal” mutagenesis. These cells are progenitors from multiple, unrelated tissues, namely skeletal muscle, kidney tubules, blood, and both subcutaneous and visceral fat. Unexpectedly, these different cell types present a somatic mutation profile that is strikingly similar. This finding is in contrast to the hypothesis of a tissue-specific mutation profile consequent to different activities and mutagen exposure in each tissue [2, 17]. The absence of tissue-specific mutagen exposure constitutes a simple way to explain how different cell types can share the same mutation profile. In this perspective, mutations observed in skeletal muscle, kidney tubules, blood, and fat progenitors are necessarily the consequence of common cellular activities, such as “house-keeping” activities. In support of this hypothesis, this group of cells, which we named "common progenitors," displays the lowest age-related increase of mutations among the cells analyzed. In addition, the signatures characterizing the common profile are found ubiquitously, but most cell types accumulate other tissue-specific mutations in addition to the common profile.

The lack of exposure to tissue-specific mutagens in the common progenitors is not surprising since tissues, like the skeletal muscle and blood, have stem/progenitor cells that reside in a protected microenvironment and are shielded from damage [51]. Somatic mutation profiles are a record of the cell lineage and activities during an individual's lifetime. Therefore, somatic mutation data can be used to address unsolved questions about stem cell hierarchy and tissue architecture [13, 52]. In the kidney, the existence of resident stem cells is controversial and the presence of a potential, protective niche is debatable [53]. Presently, the regeneration of damaged KTs appears to be mediated by (1) resident progenitors [39] and (2) tubule-epithelial cells that lose their differentiation and reacquire proliferative capacities [38]. Our analysis of the somatic mutation landscape supports both types of progenitors. Cells with in vitro proliferative capacities derived from human KTs showed either a mutation profile similar to the resident progenitors of fat and SkM (consistent with a resident KT stem cell) or a profile similar to PT-derived tumors and signs of cellular damage at both DNA and RNA level (consistent with a de-differentiated cell). The two populations do not seem completely separated. In agreement, we found a genome from a 38-year-old individual that showed an intermediate mutational and expression profile. The population of uncommitted KT progenitors also showed signs of mutagen exposure when we explored the distribution of mutations. This is consistent with their location in an environment that is not completely protected. We hypothesize that they reside in the PT, but are not part of the epithelial layer. Finally, our analyses also explored potential differences between adipose tissue progenitors residing either in the subcutaneous or visceral fat. SAT and VAT are considered two different tissues and show important differences, especially concerning the morphological changes occurring with aging [29]. However, our somatic mutation data do not support specific differences in mutagen exposure in progenitor cells from the two different types of fat during aging.

The finding and characterization of an age-related process that most likely occurs in every cell throughout the human body is a major finding of this study. This phenomenon has been termed here as "basal mutagenesis." Somatic mutation analysis in cancer genomes has identified two signatures that present clock-like features, i.e., inevitable increase in all cells as the human body ages [54]. These signatures are considered to be the products of core cellular processes, such as spontaneous deamination of methyl-cytosines (signature 1) and polymerase errors that escape the DNA repair system (signature 5) [17, 19, 20]. Results from our study expand the clock-like concept and define basal mutagenesis directly in non-cancer genomes from healthy, human tissues. Besides signatures SBS1 and 5, basal mutagenesis includes a signature that is similar but does not completely overlap with SBS3 and SBS8. In addition, we propose that SBS5 increases in a clock-like way in most cell types, but can also be enhanced by specific mutagenic processes, as observed in liver stem and kidney PT cells.

Our characterization of basal mutagenesis also includes the distribution of mutations in relation to specific, genomic features and the impact on DNA repair over time. Thanks to the comparison of older vs younger samples from multiple tissues, we are able to determine a loss of efficiency of MMR coupled with aging. In particular, the MMR-mediated protection of early-replicating DNA deteriorates with aging. We estimate that the effect size of this defect is one third of that observed in tumors with a complete MMR deficiency. These results show that the rate of somatic mutagenesis increases with aging especially in the gene-rich, early-replicating DNA, overall increasing the chances of acquiring cancer driver mutations. In addition, we found that samples from aged individuals were subjected to a relative expansion of mutations attributed to SBS5, a signature that is enhanced by another DNA repair pathway, NER. Overall, these findings suggest that the efficiency of DNA repair, in particular the MMR and NER pathways, is decreased in aged cells. These evidences point to the loss of DNA repair as an accelerating factor in cellular aging and open the door to innovations in pharmacology.

Conclusions

We provide a comprehensive genome-wide analysis of somatic mutagenesis in human cells. Our model of basal mutagenesis offers an enhanced understanding of the unavoidable loss of genome integrity and the protective forces that counteract this process, including the stem-cell niche and DNA repair. The finding of cell-type-specific mutagen exposures and consequences on cell fate in the kidney are a proof of principle supporting the importance of understanding mutational processes active in healthy human cells to understand cancer. WGS data from single genomes constitute a precious tool for achieving the goal because they allow the analysis of the non-coding portion of the genome. Overall, our comprehensive classification of mutagenic processes introduces a novel perspective for clinical advancements in preventing cancer- and age-related diseases.

Methods

Clonal cultures from multi-organ biopsies from kidney donors

Human biopsies were obtained intra-operatively from healthy living kidney donors, according to Ethical Permit Dnr 2015/1115-31. From the explanted kidney of each donor, a needle biopsy from the kidney cortex and a piece of suprarenal fat were obtained. In addition, a piece of skin with annexed subcutaneous fat was obtained. Tissues were preserved in cold PBS and immediately processed for cell isolation.

Isolation and clonal expansion of tubular progenitors from human kidney biopsies

Using a needle biopsy (1 mm diameter/10 mm height), 7–8 mg of tissue from the kidney cortex of the explanted kidney were obtained intra-operatively. The protocol for cell isolation and culturing was adapted from [55, 56]. Tissue was minced in tiny pieces with a scalpel. Around 1/5 of the biopsy was used for direct DNA/RNA extraction from whole kidney tissue. The rest was resuspended in medium and passed through tissue strainers with mesh sizes of 100 and 70 μm, thereby excluding glomeruli from the preparation. The tubular portion, which had passed through the cell strainers, was pelleted, then treated with 1× trypsin–EDTA for 5 min at 37 °C and gentle agitation, then mixed with medium and passed through a 40-μm strainer to obtain a single cell suspension. FACS sorting of CD133+ cells and single cell clonal expansion in 96-well plates was attempted ($n = 4$ biopsies) using the clone AC133 antibody (Milteny biotec, Bergisch Gladbach, Germany), but was unsuccessful. To obtain clone growth, single cell suspensions were directly plated in 6–8 wells of 6-well microtiters at 37 °C and 5% CO_2. Culture dishes were fibronectin coated (Sigma-Aldrich) and culture medium was EBM + EGM-2 MV BulletKit (Lonza, Basel, Switzerland). Twenty-four hours after plating, the medium was changed. First, the plating medium was collected and re-plated in a new 6-well microtiter to allow further attachment of kidney progenitors. One week after plating, 1–20 colonies per/well were distinguishable. Colonies with round shape and tight cell-cell contacts were considered for further culture, while scattered cells were discarded (Additional file 1: Figure S1b). When reaching ≈ 1000 cells, colonies were detached with trypsin, manually picked, and moved to new fibronectin coated 6-well microtiters, one colony per well. The whole procedure was performed under stereomicroscope inspection. Colonies were grown until confluence and used for DNA extraction. Clones that reached confluence within 1 week were moved to 10-cm-diameter petri dishes. Mean time in culture was 27.9 ± 0.8 days ($n = 26$ clones from 6 biopsies).

To assess the effectiveness of the culturing strategy, a selection of clones was subjected to FACS analysis of tubular progenitor markers [39] and qPCR analysis for markers of different kidney cell types. One hundred thousand cells per clone were stained for the kidney tubule progenitor markers CD133 (clone AC133) and CD24 (clone 32D12, both from Milteny biotec, Bergisch Gladbach, Germany) and analyzed with FACS (FACSCalibur™ - BD Biosciences). The percent of double positive cells was calculated by comparison with cells from the same clone stained with matching control IgGs (Milteny biotec) (see also Additional file 1: Figure S1c). A subset of sequenced and non-sequenced clones was also tested for the expression of transcripts considered markers of different cell types present in the kidney (see Additional file 1: Figure S1e and the section "RNA extraction and qPCR" in the "Methods" section). FACS and qPCR analyses of expression of kidney cell markers in KT clones were performed after 3–5 weeks in culture. To avoid loss of cells from clones meant for sequencing, only selected sequenced clones were inspected for the expression of kidney markers: P4903_104; P4903_117, P4903_118, P4903_119, P4903_131, P4903_132, tested by FACS; P4206_106; P4206_107; P4206_122; P4903_102, tested by qPCR; and P4903_128 and P4903_131, tested by both FACS and qPCR. The analyses were extended to clones not used for the sequencing (non-sequenced clones). These clones either came from a test biopsy ($n = 7$, female individual, age 57) or were selected among non-sequenced clones from individuals KD10 ($n = 3$), KD11 ($n = 4$), and KD12 ($n = 11$).

Clonal expansion of fat progenitors from human biopsies

One to ten grams of abdominal subcutaneous (external to the fascia superficialis) and visceral (peri-renal) fat were obtained from kidney donors undergoing surgery according to Ethical Permit Dnr 2015/1115-31. Part of the tissue was frozen for direct DNA/RNA extraction. The rest was accurately rinsed, cleaned of visible vessels, and minced with a scalpel. Tissue was placed in 30–50 ml of Hank's balanced salt solution (HBSS) containing 1 mg/ml collagenase (Collagenase A, Roche, Basel, Switzerland) in a 37 °C shaking incubator until complete digestion (30–40 min). To separate the stromal vascular fraction (SVF) from mature adipocytes, the digested tissue was centrifuged at 500*g* for 10 min and the supernatant discarded. The SVF pellet was resuspended in 1 ml of erythrocyte lysis buffer (RBC lysis solution, Qiagen) at room temperature for 5 min. To stop the lysis, cells were pelleted by centrifugation at 500*g* for 5 min and supernatant discarded. SVF was resuspended in medium and filtered through a 40-μm strainer, then plated in a 10-cm-diameter culture dish with low-serum plating medium (Dulbecco's modified Eagle's medium

(DMEM)/Ham's F-12, Life Technologies that contained 0.5% bovine serum). After 12 h in a 37 °C and 5% CO_2 incubator, non-adherent cells were carefully washed away and adherent pre-adipocytes were detached by 3–5 min of trypsinization. Cells were rinsed and stained for the hematopoietic marker CD45-APC (clone HI30, BD Biosciences, USA) and the endothelial marker CD31-PE (clone L133.1, BD Biosciences). CD45^{neg} CD31^{neg} fat progenitors were FACS sorted using a BD FACSAria™ Mu cell sorter (BD Biosciences) (see Additional file 1: Figure S1f) and single cell plated in uncoated 96-well culture plates, one plate/biopsy. Additional cells were sorted in 6-well plates as a population of 10,000–30,000 pre-adipocytes, 1 well/biopsy, and grown for 1 week before freezing. The plating medium (DMEM F12 10% FBS) of single cell cultures was changed every 2 days. The number of colonies was scored at 2 weeks after plating. At confluence (around 3 weeks), cells were trypsinized and moved to 24-well plates. Depending on the cell confluency, the colonies were then moved to 6-multiwell plates. After an average of 46.2 ± 1.3 and 48.0 ± 1.5 days in culture for subcutaneous and visceral fat, respectively, the colonies were confluent and used for DNA extraction.

Clonal expansion of epithelial progenitors from human biopsies

Skin biopsies from the lower abdomen were obtained from kidney donors undergoing surgery. The tissue was placed in cold HBSS without Ca^{2+}and Mg^{2+}(Life Technologies) containing antibiotics and antimycotics (Anti-anti, Gibco, Life Technologies) and kept at 4 °C for 4–6 h. Subcutaneous fat and loose connective tissues (hypodermis) were carefully removed. The tissue was flattened and cut into strips about 3–4 mm wide. The pieces were placed with the dermal side down in a dish containing HBSS with antibiotics and dispase (Corning, USA) and kept at 4 °C overnight. The digested epidermis was peeled from the dermal side, minced, and trypsinized with TrypLE Select (Gibco, Life Technologies) at 37 °C for 30–40 min. The digested tissue was passed through a 70-μm mesh filter, collected in a new tube containing medium and centrifuged. Pellet was resuspended in Epi-Life medium, filtered through a 40-μm strainer and plated in 4 wells of a 6-well multiwell coated with collagen (5 μg/cm^2 of Collagen I bovine protein, Gibco, following the "thin coating procedure"). Growth medium was EpiLife medium (Gibco, Life Technologies), no serum. The procedure did not produce any colonies for individuals KD05, KD09, KD10, KD11, and KD12. The culture of the epidermis from individual KD06 produced 2 colonies. Colonies of small, tight, and fast proliferating cells were visible on the extremities of the dish starting from 2 weeks after plating. When reaching ≈ 1000 cells, colonies were detached with trypsin, manually picked, and moved to new collagen-coated 6-well microtiters, one colony per well. The whole procedure was performed under stereomicroscope inspection. The cells tended to differentiate into mature large keratinocytes (see the picture in Additional file 1: Figure S1a), but a portion of cells kept small size and very high proliferative capacity for multiple passages. DNA was extracted 34 days after initial plating.

DNA extraction

DNA was extracted from the confluent wells of the 6-multiwell plate using the Gentra Puregen Kit, Qiagen. DNA was extracted from tissue biopsies using the Gentra Puregen Kit, supplemented with a lysis buffer containing Proteinase K as recommended by the supplier. DNA was extracted from 3 ml of total blood that was collected in EDTA as recommended by the instructions of the Gentra Puregen Blood Kit.

Sequencing

The library preparation and sequencing were carried out at NGI Sweden, Science for Life Laboratories, Stockholm, following standard methods. For cell clones, the library preparation was performed by a semiautomatic NeoPrep station using the Illumina TruSeq Nano Kit (350 bp average insert size) and 25 ng of DNA as starting material. The libraries of the bulk blood samples were prepared with Illumina TruSeq PCR-free library preparations (350 bp average insert size). Sequencing was performed on Illumina HiSeq X, PE 2 × 150 bp.

Somatic variant calling

Raw reads were aligned to the human reference genome (GRCh37/hg19 assembly version), using bwa mem 0.7.12 [57]. Alignments were sorted and indexed using samtools 0.1.19 [58]. Alignment quality control statistics were gathered using qualimap v2.2 [59]. The raw alignments were then processed following the GATK best practice [60] with version 3.3 of the GATK software suite. Alignments were realigned around InDels using GATK RealignerTargetCreator and IndelRealigner, duplicates were marked using Picard MarkDuplicates 1.120, and base quality scores were recalibrated using GATK BaseRecalibrator. Finally, genomic VCF files were created using the GATK HaplotypeCaller 3.3. Reference files from the GATK 2.8 resource bundle were used. All above steps were coordinated using Piper v1.4.0 (www.github.com/NationalGenomicsInfrastructure/piper).

Somatic variants were defined as heterozygous in the single cell clone and either absent or very rare in an unrelated tissue (blood), sequenced as a bulk. To identify somatic variants, a specific pipeline was developed. For each clone, variants were initially called with HaplotypeCaller (GATK) [61], MuTect2 (GATK 3.5.0), and

FermiKit version r178 [62]. The union of these three sets of variants was subjected to further filtering steps in order to exclude (1) sequencing artifacts, (2) germline variants (detected both in the clone and blood bulk), and (3) variants that occurred during the in vitro culture of the clone (found only in a subset of cells of the clone, therefore showing low AF). To this aim, the AF of each variant was derived from the .bam files and matched to the relative blood bulk sequencing. Somatic variants were defined as follows: the read fraction supporting the alternative allele was comprised between 0.4 and 0.6 in the clone sequence, a minimum of 3 reads supported the variant, the read fraction in the blood was low (alternative < 0.1), and the coverage in both the clone and blood was at least 15X. Chromosomes X and Y were excluded from the analyses (however, variants recovered on the X chromosomes of female donors can be found in Additional file 3). Additional quality filters were applied as follows: the reads supporting the variants were on both strands, the maximum coverage was 1000X, and the variants that were located in problematic regions [63, 64] were removed. Variants common to more than one sample were considered artifacts and removed. Variant validation was performed to ensure that our lists of somatic mutations only contained somatic variants that were present in the cell before in vitro culturing (see the section "Variant validation" in the "Methods" section). Comparison of variants recovered in DNA from a clone derived from the same ancestor cell, but cultured in 2 different wells and independently sequenced, shows high validation rate (99 and 97% for SNVs and InDels, respectively, Additional file 1: Table S1e) and supports low levels of culture-induced variants in our lists. However, we cannot exclude the presence of non-neutral, positively selected variants that might have occurred in vitro. Variants were annotated using the Ensembl Variant Effector Predictor from [65]. Frequency of detected somatic SNVs in the Swedish population (germline variants) was annotated in Additional file 2 and Additional file 3 using SweGen [66] version 20180409.

Variant validation

The variant validation was performed on a technical replicate of WGS. Two clones derived from the same ancestor cell (P4206_128 and P4206_130) were independently grown in culture. The DNA was extracted and sequenced independently, but clone P4206_130 was not included in the study. Variants were called in clones P4206_128 (discovery set) according to our somatic variant calling pipeline. Called variants that had a minimum coverage of 10x in both the discovery and the validation sets were used for the validation. In total, 870 SNVs and 71 InDels were tested. Variants were considered validated when at least 3 reads supporting the alternative alleles were present in the validation set. As a control for the background signal, we validated the variants in unrelated clones, e.g., clones derived from a different founder cell obtained from the same or a different biopsy. Additional validation and discussion of our somatic mutation calling strategy are available at [11].

Microsatellite instability

Microsatellite instability was assessed using MSIsensor v.0.5 [67] where every cell clone and representative blood bulk were analyzed and the msi score calculated.

Copy number variation

Copy number variation was detected in clonally expanded cells using Ascat [68]. Ascat detects allele-specific copy number variation in a tumor sample using Log R and B allele frequency (BAF) information at specific SNP loci in the tumor sample and a matched germline sample from the same individual. We used the loci of all bi-allelic SNPs in 1000 Genomes phase 3, release date 20130502 [69] with minor allele frequency > 0.3 to calculate Log R and BAF data in the clonally expanded cells and the matched blood samples. The software AlleleCount (https://github.com/cancerit/alleleCount) was used to generate the number of reads in the bam files supporting the two alleles of the SNPs. BAF and LogR was then calculated at all SNP loci according to:

$$\mathrm{BAF}_i^c = \frac{CountsB_i^c}{CountsA_i^c + CountsB_i^c}$$

$$\mathrm{BAF}_i^b = \frac{CountsB_i^b}{CountsA_i^b + CountsB_i^b}$$

$$\mathrm{LogR}_i^c = \log_2 \frac{CountsA_i^c + CountsB_i^c}{CountsA_i^b + CountsB_i^b} - \mathrm{median}\left(\log_2 \frac{CountsA^c + CountsB^c}{CountsA^b + CountsB^b} \right)$$

$$\mathrm{Log}R_i^b = 0$$

where i is a specific SNP locus, c is the clonally expanded sample, b is the blood sample, *CountsA* is the number of reads supporting one of the alleles of the SNP, and *CountsB* is the number of reads supporting the other allele of the SNPs.

Ascat was run with parameter gamma set to 1. We report only large copy number aberrations that were detectable by visual inspection of the ASPCF.png and ASCATprofile.png images generated by Ascat for each sample. Execution of Ascat and the generation of Log R and BAF was coordinated using Sarek release v2.1.0 [70].

Meta-analysis

Newly generated and publicly available somatic SNVs from normal and cancer samples underwent a common filtering step to exclude variants from the repeat-masked hg19 genome assembly. In particular, we excluded regions with CRG Alignability-75 score [71] below the maximum (< 1.0) and additionally the UCSC Browser blacklisted regions (DAC and Duke) were excluded; this step retained 2393.43 Mb of the genome. Furthermore, we excluded from all analyses the regions with low genomic coverage in our data (< 15 reads in WGS of > 5% of the samples), retaining 2094.95 Mb of the hg19 genome for the final analysis.

Mutational signature inference

Analysis of mutational signatures was performed as described in [21]. Briefly, the SNVs from the healthy samples and the tumor samples were analyzed jointly, where a NMF (non-negative matrix factorization) analysis was applied to matrices of mutation counts across the 96 mutational contexts, as customary (see, e.g., [16]). Upon repeated runs ($n = 200$) of the NMF procedure (function *nmf* in the *R* package *NMF*, using the default "Brunet" algorithm) on the bootstrap-resampled mutation count data, the 200 NMF results were clustered using k-medoids algorithm (function *pam* in R package *cluster*) to obtain the final set of mutational signatures and their contributions (exposures) in every sample. The signature profiles obtained from this NMF analysis were compared using cosine similarity to the known mutational signatures (http://cancer.sanger.ac.uk/cosmic/signatures and [18]).

Genomic distribution of mutations

Analysis of enrichment or depletion of mutations in exons, introns, regulatory, and conserved regions was carried on using the R package *MutationalPatterns* [72]. Tissue-specific genes were obtained from the Human Protein Atlas (http://proteinatlas.com). The genes that had the annotation "elevated in …," "expressed in all," and "mixed expression pattern" were considered tissue-specific gene for that tissue. To define the conserved regions, PhastConsElements46way data was used and downloaded from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons46way/.

The association of mutation enrichment/depletion with specific genomic features was performed as described in [21, 44]. In brief, regression analysis was performed to examine the relationship between the mutations and the covariates (replication timing, H3K36me3, transcriptional levels, CTCF motif) individually while controlled for others. The replication timing (RT) data was obtained from the ENCODE project (RepliSeq) and divided into six bins ranging from latest replicating (bin 0) to earliest replicating (bin 5); values are averages over eight diverse cell types (source file names in the form "wgEncodeUwRepliSeq_____WaveSignalRep1.bigWig" where the gap contains cell line names: Helas3, Hepg2, Huvec, Nhek, Bj, Imr90, Mcf7, Sknsh). The RNA-seq levels and H3K36me3 histone mark were collected from Roadmap Epigenomics project and averaged over eight diverse cell types (for H3K36me3: E017 LNG.IMR90, E114 A549, E117 CRVX.HELAS3.CNCR, E118 LIV.HEPG2.CNCR, E119 BRST.HMEC, E127 SKIN.NHEK, E125 BRN.NHA, E122 VAS.HUVEC; for RNA-seq, these same cell types except that we substituted E096 and E071 for E017 and E125 because of data availability). The RNA-seq was divided into four bins where non-expressed regions were in bin 0 and expressed regions were in bins 1 (low expression) to 3 (high expression). The H3K36me3 was divided into four bins, with bin 0 as absent from H3K36me3 (fold-enrichment versus ChIP-seq "input" ≤1.0) and ranging up to bin 3 with the highest abundance.

Predicted pathogenic variants

To obtain the number of potentially pathogenic mutations in each clone, SNVs and InDels were annotated with CADD (Combined Annotation Dependent Depletion) [73]. Mutations that obtained a PHRED score higher than 15 were selected and filtered on gene expression (obtained from Human Protein Atlas, as described in the section "Genomic distribution of mutations"). Variants with CADD score higher than 15, but no gene annotation were excluded, as well as variants affecting the sequence of a gene not expressed in the tissue of origin of the clone.

RNA extraction and qPCR

RNA from KT clones was extracted from plated cells, previously snap-frozen in their tissue culture plates, using the RNeasy Mini kit (Qiagen), according to the manufacturer's instructions. RNA from total kidney was obtained from a needle biopsy from a healthy kidney not included in the study (female, age 38) undergoing explant for kidney donation. The fresh biopsy was minced in tiny pieces, and around 1/5 of the material was snap-frozen for RNA extraction. The rest of the biopsy was used for KT progenitor culture. RNA from the biopsy was extracted using the RNeasy Mini kit (Qiagen) and homogenized with a syringe. RNA from all samples used in the qPCR analyses were extracted at the same time. cDNA synthesis was performed using random hexamers and SuperScript Reverse Transcriptase (Invitrogen). Quantitative RT-PCR was performed using either a TaqMan gene expression assay from Applied Biosystems (Podocalyxin, PDX, Hs00193638-m1) or SYBRgreen using the set of primers specified (Table 1).

Table 1 QPCR primers for gene expression analysis

	Forward	Reverse
ACTA2	acaggaatacgatgaagccg	gctttggctaggaatgatttgg
AQP1	ggaccggcagagctctacag	acgtcttctggacccatgct
CALB1	ttacctggaaggaaaggagctgca	tcttctgtgggtaatacgtgagcc
COL1A1	atgaccgagacgtgtggaaa	tttcttggtcggtgggtgact
CUBN	tgtttcttacggggtctgctca	gcagaccaattgcactcccttt
KIM1 (HAVCR1)	cgtgggtggttcaatgacatga	tgacggttggaacagttgtgac
LPR2	ccaaagactgttcagatgacgc	ctgagccatcatcacagtcttg
Nephrin (NPHS1)	cacacggtcagcacaacagagg	gaaacctcgggaataagacacct
PAX2	caaagttcagcagcctttcc	tcaccattggagcgaggaat
PAX8	atccggcctggagtgatagg	tggcgtttgtagtccccaatc
PDZK1	ccctgtgatgaatggaggtgt	tcatagccacaccttgaggtgt
PECAM1	ttcaagccttgagggtcaag	tgtaaaacagcacgtcatcctt
Podocin (NPHS2)	taccaaatcctccggcttagg	tttggctcttccaggaagcaga
SLC5A12	ttgtgggcttcttaacggttc	cgcctgagaggatctacatca
SLC9A3	ttgaggaggtccatgtcaacg	gcgccacgaaagattcaaaca
SLC17A3	aagaacgcacaagatatgcaagt	tgtaagacgagggctattccat
SLC22A7	actttcttcttcgccggtgt	attacatagctgacggaggctg
UMOD	actacgtctacaacctgacagc	tctatactgcactcctcacacg
VCAM1/CD106	cagtaaggcaggctgtaaaaga	tggagctggtagaccctcg

Statistical analyses

Unless otherwise indicated, the P values were calculated using either two-tailed distribution, two-sample unequal variance Student's t tests (when comparing two groups), or one-way ANOVA with multiple comparison post hoc test. Significance was defined as $P < 0.05$ ($^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0005$). The results are presented as the mean ± standard error of the mean (SEM). All calculations were performed using GraphPad Prism software. The linear fits between mutation numbers and age were obtained using a linear mixed-effects model where the dependent variable is the number of mutations or a given mutational signature, the fixed effect is age, and the random effect is the individual. Bonferroni correction was used to adjust for multiple testing. Analyses were performed in R. T-SNE analysis was performed using *tsne* package in R, and clustering showed in Fig. 2a was performed using *heatmap3* package in R.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1892-z.

Additional file 1. Supplementary figures (Figure S1-S8) and tables (Table S1-S6)

Additional file 2. Lists of somatic mutations detected on autosomes of 69 clones from 6 healthy donors and grouped per tissue

Additional file 3. Lists of somatic mutations detected on the X chromosome of female donors (KD05, KD06 and KD11). Not used in the analyses.

Additional file 4. Review history

Acknowledgements
We acknowledge the healthy donors who contributed tissue samples for the study. The authors wish to thank Christine Harling-Berg, Staffan Strömblad, and Letizia De Chiara for critical reading of the manuscript and Elvira Lundin for graphical contribution to the figures. Open access funding provided by Karolinska Institute.

Review history

The review history is available as Additional file 4.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

IF and ME designed the study. JN, TL, PS, and LW collected the samples. IF, AM, and PV performed the experimental work. IF, HTH, AM, ML, AJ, NN, PL, DMP, TL, PS, LW, FS, and ME analyzed and interpreted the data. HTH, DMP, and FS performed the statistical analysis. IF wrote the first draft of the manuscript, and all authors critically revised the manuscript. IF and ME obtained the funding. All authors read and approved the final manuscript.

Funding
This study was supported by grants to M.E. from the Swedish Research Council, the Center for Innovative Medicine, and the Olle Engkvist Byggmästare Foundation; to I.F. from Jeanssons, Hagelen and Osterman Foundations, Svenska Läkaresällskapet, and the Rönnbergs scholarship from Riksbankens Jubileumsfond; D.M.P. and F.S. were funded by the ERC Starting Grant HYPER-INSIGHT (757700), by the Spanish Ministry of Economy and

Competitiveness (BFU2017-89833-P) and by the Severo Ochoa Centre of Excellence program to the IRB Barcelona. F.S. is further supported by the ICREA Research Professor program. A.J., M.L., and N.N. are supported by a grant from the Wallenberg Foundation to the Wallenberg Advanced Bioinformatics Infrastructure. The authors acknowledge support from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure funded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure.

Availability of data and materials

Sequencing data generated during the current study are not publicly available due to the European General Data Protection Regulation (GDPR) to protect patients' privacy, but are available from the corresponding author on reasonable request. Aggregated lists of somatic variants recovered in all clones of all donors are available as Additional file 2 (autosomes) and Additional file 3 (X chromosomes of female donors). Lists of somatic variants used in the meta-analysis are either accessible from the original publication (listed in Additional file 1: Table S2) or available upon request from the GDC Data Portal (for the TCGA data set samples) and the ICGC Data Portal (sample IDs are listed in Additional file 1: Table S2).

Ethics approval and consent to participate

Human biopsies were obtained from living kidney donors that gave written consent, according to ethical permit Dnr 2015/1115-31. Experimental methods used in the study comply with the Helsinki Declaration.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biosciences and Nutrition, Center for Innovative Medicine, Karolinska Institutet, Huddinge, Sweden. [2]Department of Medicine Huddinge, Integrated Cardio Metabolic Center, Karolinska Institutet, Huddinge, Sweden. [3]Science for Life Laboratory, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden. [4]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. [5]Science for Life Laboratory, Department of Molecular Biology, Umeå University, Umeå, Sweden. [6]Science for Life Laboratory, Department of Biochemistry and Biophysics (DBB), Stockholm University, Stockholm, Sweden. [7]Genome Data Science, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain. [8]Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet, Division of Transplantation Surgery, Karolinska University Hospital, Huddinge, Sweden. [9]Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet, Division of Renal Medicine, Karolinska University Hospital, Huddinge, Sweden. [10]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Received: 14 June 2019 Accepted: 18 November 2019
Published online: 18 December 2019

References

1. Vijg J, Suh Y. Genome instability and aging. Annu Rev Physiol. 2013;75: 645–68.
2. Zhang L, Vijg J. Somatic mutagenesis in mammals and its implications for human disease and aging. Annu Rev Genet. 2018;52:397–419.
3. Chanock SJ. The paradox of mutations and cancer. Science. 2018;362:893–4.
4. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, Wartman LD, Lamprecht TL, Liu F, Xia J, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012;150:264–78.
5. Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, Vijg J. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat Methods. 2017;14:491–3.
6. Dou Y, Gold HD, Luquette LJ, Park PJ. Detecting somatic mutations in normal cells. Trends Genet. 2018;34(7):545–57.
7. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, Cagan A, Murai K, Mahbubani K, Stratton MR, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362:911–7.
8. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348:880–6.
9. Yizhak K, Aguet F, Kim J, Hess JM, Kubler K, Grimsby J, Frazer R, Zhang H, Haradhvala NJ, Rosebrock D, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. Science. 2019;364.
10. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, Shiozawa Y, Sato Y, Aoki K, Kim SK, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565:312–7.
11. Franco I, Johansson A, Olsson K, Vrtacnik P, Lundin P, Helgadottir HT, Larsson M, Revechon G, Bosia C, Pagnani A, et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. Nat Commun. 2018;9:800.
12. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, Huch M, Boymans S, Kuijk E, Prins P, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538:260–4.
13. Lee-Six H, Obro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, Osborne RJ, Huntly BJP, Martincorena I, Anderson E, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;561:473–8.
14. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, de la Fonteijne L, Varela I, Camargo FD, van Boxtel R. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. Cell Rep. 2018;25:2308–16 e2304.
15. Abyzov A, Tomasini L, Zhou B, Vasmatzis N, Coppola G, Amenduni M, Pattni R, Wilson M, Gerstein M, Weissman S, et al. One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. 2017;27:512–23.
16. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 2013;3:246–59.
17. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15:585–98.
18. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. The repertoire of mutational signatures in human cancer. bioRxiv 2019:322859 . Available from: https://www.biorxiv.org/content/10.1101/322859v2.
19. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Kwiatkowski DJ, Rosenberg JE, Van Allen EM, D'Andrea A, Getz G. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat Genet. 2016;48:600–6.
20. Zou X, Owusu M, Harris R, Jackson SP, Loizou JI, Nik-Zainal S. Validating the concept of mutational signatures with isogenic cell models. Nat Commun. 2018;9:1744.
21. Supek F, Lehner B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. Cell. 2017;170:534–47 e523.
22. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;521:81–4.
23. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488:504–7.
24. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence M, Reynolds A, Rynes E, Vlahovicek K, Stamatoyannopoulos JA, Sunyaev SR. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4.
25. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010;463:191–6.
26. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, Pletikos M, Pattni R, Chen BJ, Venturini E, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science. 2018;359:550–5.
27. Young MD, Mitchell TJ, Vieira Braga FA, Tran MGB, Stewart BJ, Ferdinand JR, Collord G, Botting RA, Popescu DM, Loudon KW, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. Science. 2018;361:594–9.
28. Bolignano D, Mattace-Raso F, Sijbrands EJ, Zoccali C. The aging kidney revisited: a systematic review. Ageing Res Rev. 2014;14:65–80.
29. Sepe A, Tchkonia T, Thomou T, Zamboni M, Kirkland JL. Aging and regional differences in fat cell progenitors - a mini-review. Gerontology. 2011;57:66–75.
30. McKenna T, Sola Carvajal A, Eriksson M. Skin disease in laminopathy-associated premature aging. J Invest Dermatol. 2015;135:2577–83.
31. Revechon G, Viceconte N, McKenna T, Sola Carvajal A, Vrtacnik P, Stenvinkel P, Lundgren T, Hultenby K, Franco I, Eriksson M. Rare progerin-expressing

preadipocytes and adipocytes contribute to tissue depletion over time. Sci Rep. 2017;7:4405.
32. Shiels PG, McGuinness D, Eriksson M, Kooman JP, Stenvinkel P. The role of epigenetics in renal ageing. Nat Rev Nephrol. 2017;13:471–82.
33. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP, et al. A compendium of mutational signatures of environmental agents. Cell. 2019;177:821–36 e816.
34. Rouhani FJ, Nik-Zainal S, Wuster A, Li Y, Conte N, Koike-Yusa H, Kumasaka N, Vallier L, Yusa K, Bradley A. Mutational history of a human cell lineage from somatic to induced pluripotent stem cells. PLoS Genet. 2016;12:e1005932.
35. Nakamura J, Mutlu E, Sharma V, Collins L, Bodnar W, Yu R, Lai Y, Moeller B, Lu K, Swenberg J. The endogenous exposome. DNA Repair (Amst). 2014;19:3–13.
36. Lindgren D, Eriksson P, Krawczyk K, Nilsson H, Hansson J, Veerla S, Sjolund J, Hoglund M, Johansson ME, Axelson H. Cell-type-specific gene programs of the normal human nephron define kidney cancer subtypes. Cell Rep. 2017;20:1476–89.
37. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, Buhay C, Kang H, Kim SC, Fahey CC, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014;26:319–30.
38. Kusaba T, Lalli M, Kramann R, Kobayashi A, Humphreys BD. Differentiated kidney epithelial cells repair injured proximal tubule. Proc Natl Acad Sci U S A. 2014;111:1527–32.
39. Angelotti ML, Ronconi E, Ballerini L, Peired A, Mazzinghi B, Sagrinati C, Parente E, Gacci M, Carini M, Rotondi M, et al. Characterization of renal progenitors committed toward tubular lineage and their regenerative potential in renal tubular injury. Stem Cells. 2012;30:1714–25.
40. Shuch B, Amin A, Armstrong AJ, Eble JN, Ficarra V, Lopez-Beltran A, Martignoni G, Rini BI, Kutikov A. Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity. Eur Urol. 2015;67:85–97.
41. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, Bowlby R, Gibb EA, Akbani R, Beroukhim R, et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. Cell Rep. 2018;23:313–26 e315.
42. Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, Collisson EA, Vemula SS, Naujokas A, Chiotti KE, et al. Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. Cell Rep. 2014;9:1228–34.
43. Huang Y, Li GM. DNA mismatch repair preferentially safeguards actively transcribed genes. DNA Repair (Amst). 2018;71:82–6.
44. Avgustinova A, Symeonidi A, Castellanos A, Urdiroz-Urricelqui U, Sole-Boldo L, Martin M, Perez-Rodriguez I, Prats N, Lehner B, Supek F, Benitah SA. Loss of G9a preserves mutation patterns but increases chromatin accessibility, genomic instability and aggressiveness in skin tumours. Nat Cell Biol. 2018;20:1400–9.
45. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ, et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell. 2016;164:538–49.
46. Forsberg LA, Rasi C, Razzaghian HR, Pakalapati G, Waite L, Thilbeault KS, Ronowicz A, Wineinger NE, Tiwari HK, Boomsma D, et al. Age-related somatic structural changes in the nuclear genome of human blood cells. Am J Hum Genet. 2012;90:217–28.
47. Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T, et al. CTCF/cohesin-binding sites are frequently mutated in cancer. Nat Genet. 2015;47:818–21.
48. Cancer Genome Atlas Research N, Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, Davis C, Wheeler DA, Murray BA, Schmidt L, et al. Comprehensive molecular characterization of papillary renal-cell carcinoma. N Engl J Med. 2016;374:135–45.
49. Mitchell TJ, Turajlic S, Rowan A, Nicol D, Farmery JHR, O'Brien T, Martincorena I, Tarpey P, Angelopoulos N, Yates LR, et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal. Cell. 2018;173:611–23 e617.
50. Vallon V. Tubular transport in acute kidney injury: relevance for diagnosis, prognosis and intervention. Nephron. 2016;134:160–6.
51. Schultz MB, Sinclair DA. When stem cells grow old: phenotypes and mechanisms of stem cell aging. Development. 2016;143:3–14.
52. Franco I, Fernandez-Gonzalo R, Vrtacnik P, Lundberg TR, Eriksson M, Gustafsson T. Healthy skeletal muscle aging: the role of satellite cells, somatic mutations and exercise. Int Rev Cell Mol Biol. 2019;346:157–200.
53. Kramann R, Kusaba T, Humphreys BD. Who regenerates the kidney tubule? Nephrol Dial Transplant. 2015;30:903–10.
54. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47:1402–7.
55. Bussolati B, Bruno S, Grange C, Buttiglieri S, Deregibus MC, Cantino D, Camussi G. Isolation of renal progenitor cells from adult human kidney. Am J Pathol. 2005;166:545–55.
56. Bussolati B, Moggio A, Collino F, Aghemo G, D'Armento G, Grange C, Camussi G. Hypoxia modulates the undifferentiated phenotype of human renal inner medullary CD133+ progenitors through Oct4/miR-145 balance. Am J Physiol Renal Physiol. 2012;302:F116–28.
57. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
59. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32:292–4.
60. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11–33.
61. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
62. Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics. 2015;31:3694–6.
63. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12:966–8.
64. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–51.
65. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. Genome Biol. 2016;17:122.
66. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, Viklund J, Kahari AK, Lundin P, Che H, et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur J Hum Genet. 2017;25:1253–60.
67. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics. 2014;30:1015–6.
68. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, Weigman VJ, Marynen P, Zetterberg A, Naume B, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010;107:16910–5.
69. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
70. Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, DiLorenzo S, Sandgren J, Diaz de Ståhl T, Wirta V, Nistér M, Nystedt B, Käller M. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. bioRxiv. 2018:316976. Available from: https://www.biorxiv.org/content/10.1101/316976v1.
71. Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, Ribeca P. Fast computation and applications of genome mappability. PLoS One. 2012;7:e30377.
72. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;10:33.
73. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.

Publisher's Note

Additional file 1

Whole genome DNA sequencing of healthy human cells provides an atlas of somatic mutagenesis and identifies a tumor-prone cell type

Irene Franco[1*], Hafdis T. Helgadottir[1*], Aldo Moggio[2], Malin Larsson[3], Peter Vrtačnik[1], Anna Johansson[4], Nina Norgren[5], Pär Lundin[1,6], David Mas Ponte[7], Johan Nordström[8], Torbjörn Lundgren[8], Peter Stenvinkel[9], Lars Wennberg[8], Fran Supek[7,9] and Maria Eriksson[1]

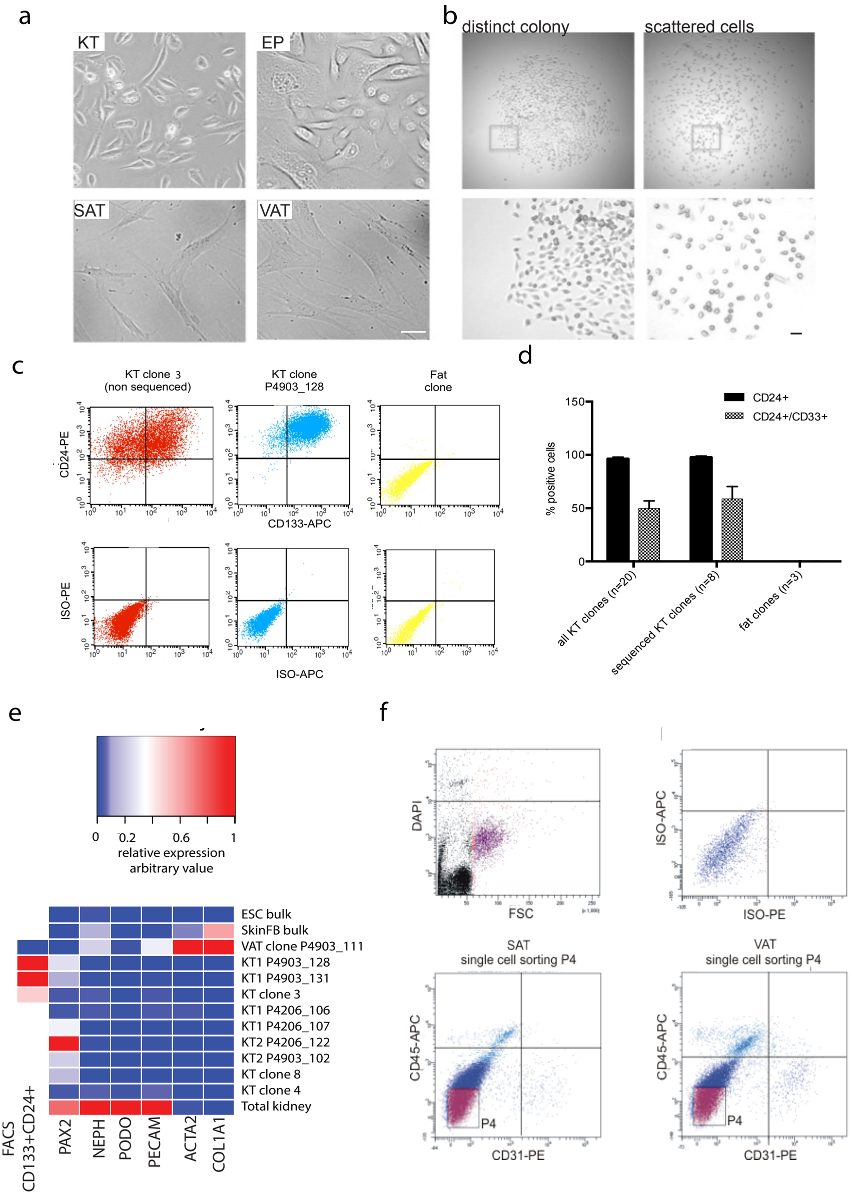
Supplementary figures

Supplementary Tables

Supplementary bibliography

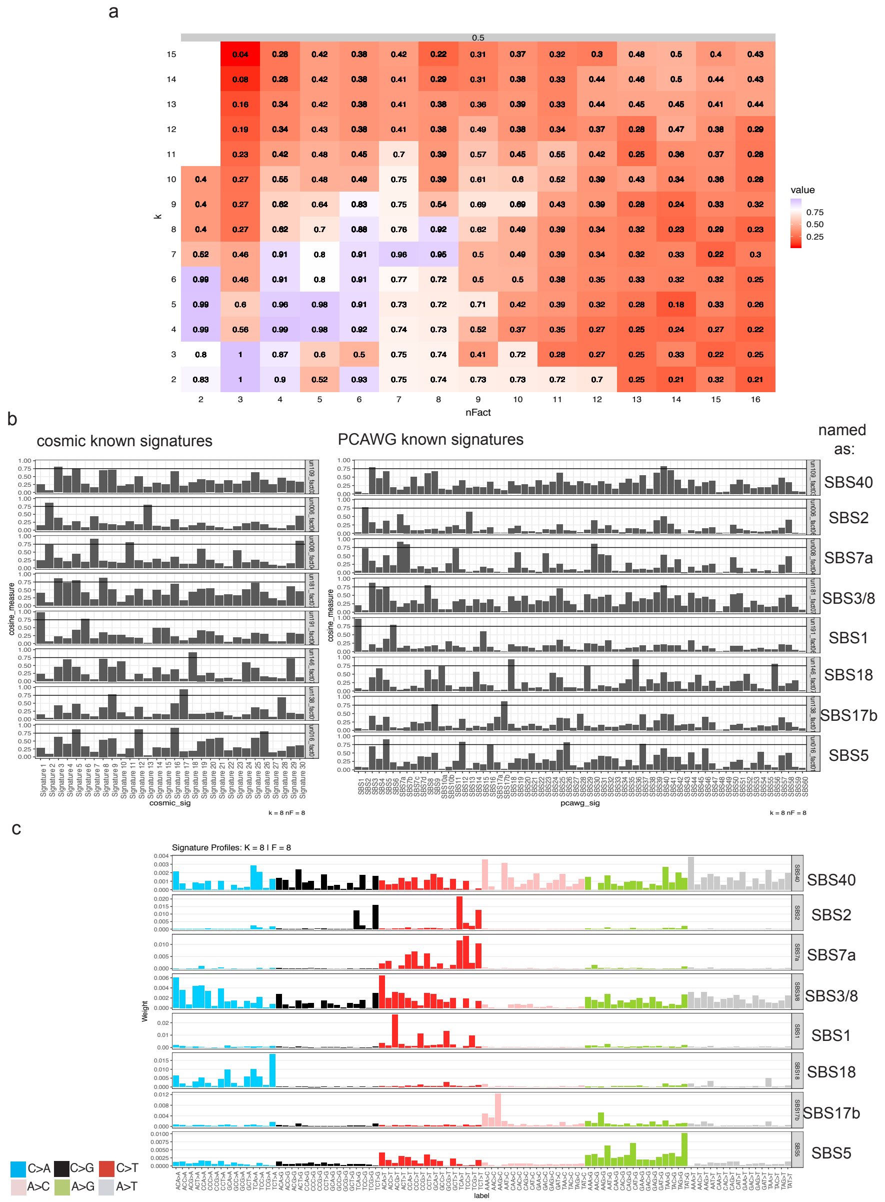
Supplementary figures

Fig. S1. Characterization of clonally expanded progenitors from human kidney tubules, fat and epidermis



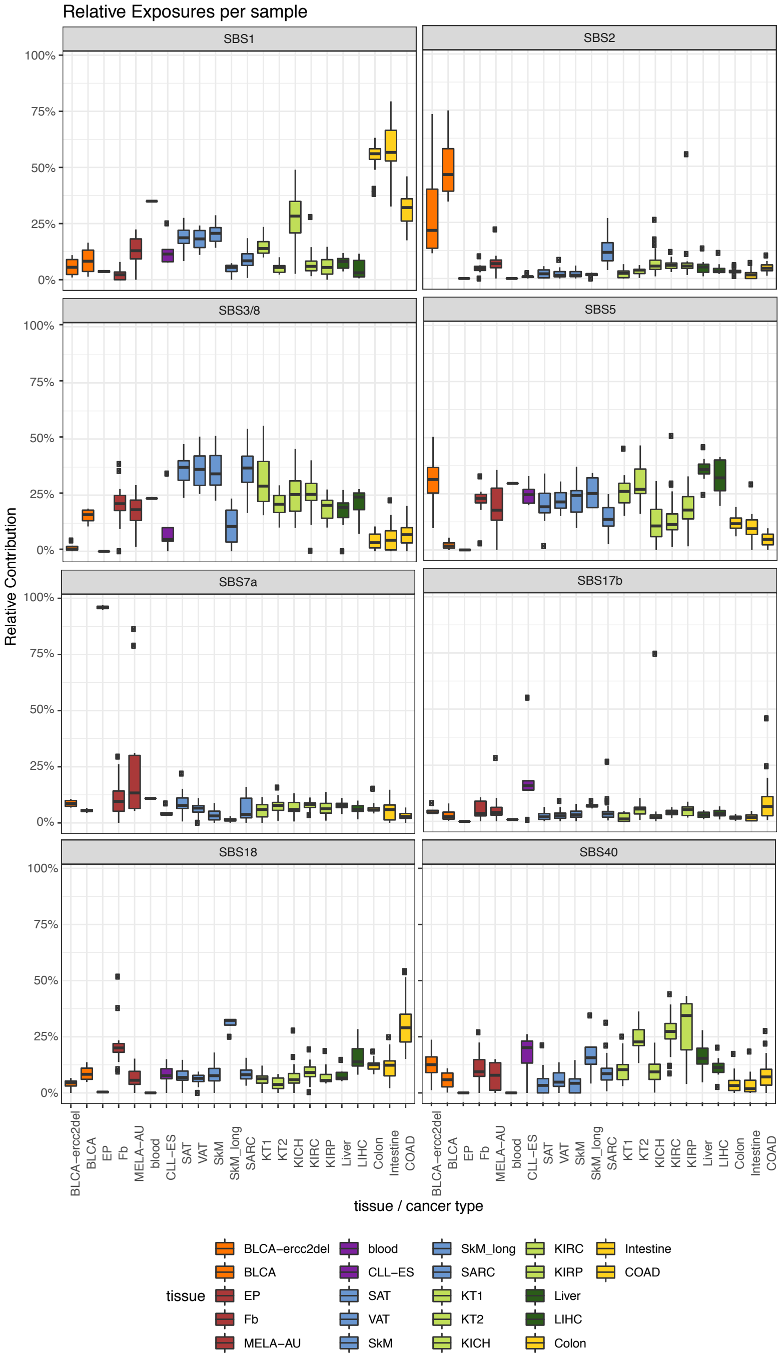
Representative micrographs of single cell clones from human biopsies used in the study. Kidney tubule (KT, top left), epidermis (EP, top right), sub-cutaneous adipose tissue (SAT, bottom left) and visceral adipose tissue (VAT, bottom right) progenitors were expanded in culture for 3 to 6 weeks, then used for DNA extraction and sequencing. The presented pictures correspond to the final stages of the cell culture. The cell morphology was checked and used for selecting suitable clones for sequencing. **b.** Representative images and 5x magnifications (bottom) of colonies from KT cultures and criteria for selection of KT progenitor colonies on the base of morphology. Cultures from KT cell suspensions were inspected daily to follow the growth of distinct colonies. Ten to 15 days after plating, one colony per well was selected, detached and moved to a new plate. Only colonies with round shape and tight cell-cell contacts (left panels) were considered for further culture, while colonies composed of scattered cells (right panels) were discarded. Bars=50 μm **c.-e.** FACS and qPCR assessment of expression of kidney cell markers in KT clones after 3-5 weeks in culture. Due to the reduced amount of material obtained from the clonal culture, only a portion of the KT clones included in the somatic mutation analysis could also be tested for the expression of kidney markers. To extend the characterization, FACS and qPCR analyses were performed on clones not used for the sequencing (non sequenced clones), but cultured at the same time as the ones chosen for DNA extraction. Overall, all tested KT clones (n=20) expressed the markers of kidney progenitors CD24 and CD133, while fat clones used as negative controls were completely negative **(c.-d.).** In KT clones, CD24 was expressed by nearly all the cells within the clonal population, while the levels of CD133 were more variable (**c.**). Expression of the kidney progenitor marker *PAX2* was detectable in most KT clones at the RNA level **(e)**. Conversely, KT clones were always negative for markers of non-tubular cells, like *NEPH* and *PODO* (glomeruli), *PECAM* (endothelium), *ACTA2* (smooth muscle cells) and *COL1A1* (fibroblasts) **(e)**. A portion of a healthy kidney biopsy (Total kidney), a VAT clone and non-clonal populations of either embryonic stem cells (ESC bulk) or skin fibroblasts (SkinFb bulk) were included in the qPCR analysis as positive and negative controls **(e)**. Three clones from biopsy KD12 were tested by both FACS and qPCR **(c and e)**: the non-sequenced KT clone 3 (a representative clone that was excluded from sequencing on the base of morphological appearance of the cells, as described in **(a))** and two sequenced clones, P4903_128 and P4903_130. **f.** Representative dot-plots of FACS analyses and single cell sorting of fat samples. For every fat biopsy, the stromal vascular fraction was plated for 12 h in low serum conditions. Adherent cells were detached by quick trypsinization to obtain a cell preparation enriched for adipocyte progenitors. Dot plot of a representative DAPI staining to assess the numbers of living cells in the preparation is shown (top left). The treatment ensured a very high viability. An ISO-IgG control staining was performed to assess antibodies reactivity (top right). Pre-adipocytes isolated from SAT and VAT were stained with the hematopoietic-cell marker CD45 and the endothelial cell marker CD31 and selected from the double negative population as indicated in the gate P4. The double negative population was predominant in all biopsies (n= 9). However, the percentage of CD45 and CD31 positive cells was variable across samples, as can be appreciated in the SAT and VAT samples from the same donor that are shown in panel **f,** bottom left and right, respectively**.**

Fig. S2. Non-negative matrix factorization and comparison of extracted signatures to COSMIC cancer signatures and PCAWG single base signatures (SBS)



(**a**) Eight mutational signatures obtained from NMF of somatic mutation catalogues from healthy (n=161) and tumors (n=192) samples. For each combination of k (number of possible clusters, rows) and nFact (rank from NMF, columns) the general silhouette index (SI representing high reproducibility, values within the heatmap) was obtained. The chosen parameters were at nFact=8 and k=8 with SI=0.92. **b**) The 8 de novo signature profiles obtained from the NMF analysis were compared to already characterized signatures from COSMIC (30 signatures; http://cancer.sanger.ac.uk/cosmic/signatures) and PCAWG (60 signatures;[1]). Cosine measurements, indicating what COSMIC/PCAWG signature fits best with the de novo signatures, are provided. **c**) The mutational profiles of the 8 de novo signatures named after the most similar single base signature (SBS) from PCAWG

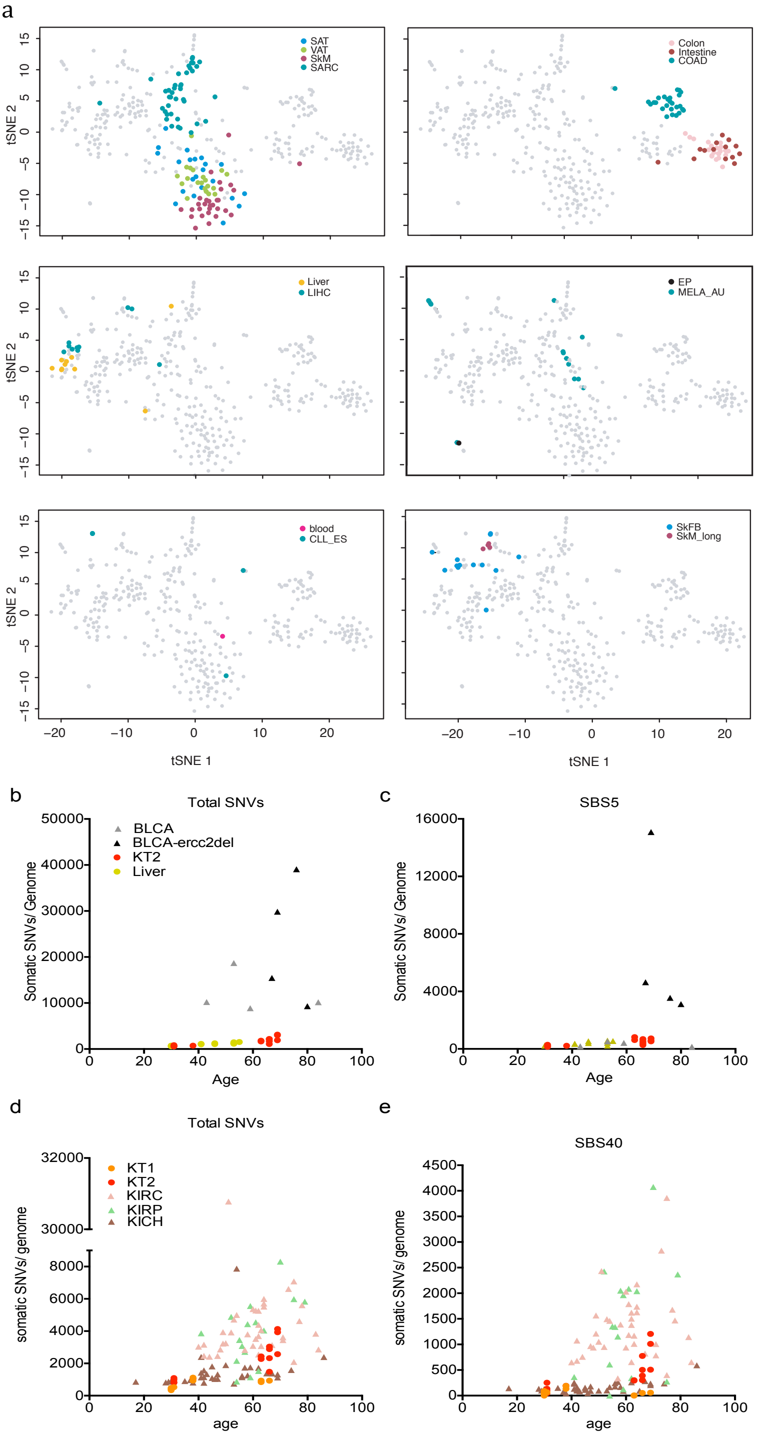
Fig. S3. Relative contribution of extracted signatures to healthy tissues and tissue-matched tumors



Relative contribution of the 8 mutational signatures to the somatic mutation catalogues of healthy (n=161) and tissue-matched tumor samples (n=192). Results of a statistical test (Mann-Whitney U test; *** FDR<1%,) testing enrichment of the exposure of the signature in one tissue compared to the same signature in all other tissues are shown in Table S3. Overall, our analysis shows that signatures SBS1, 3/8 and 5 were found ubiquitously and we defined this combination of signatures as "basal mutagenesis". Consistent with this concept, cell types that were not common progenitors, had additional signatures that are associated with specific, mutagen exposure. Examples are 1) EP samples showing high levels of SBS7a, a signature induced by UV-light exposure, 2) the SkM cells used as a control for culture-induced mutagenesis in our previous study [2] (SkM-long), which showed SBS18, a signature linked to *in vitro*-culture stress [3, 4] and consequent production of intracellular reactive-oxygen species [5]. These samples were used as positive controls for prolonged exposure to a mutagen. All groups of cells were compared to these controls for either basal or mutagen-driven mutagenesis. SkinFB clustered in close proximity to the SkM-long samples (Figure 3a) and showed high levels of SBS18, consistent with the long *in vitro* culture required for the reprogramming protocol [6]. The SkinFB also showed the second highest SBS7a contribution after EP (Figure 3b). Intestine and colon stem cells formed a distinct cluster and were characterized by very high SBS1 contribution, previously explained with a high replication rate of these cells [7].

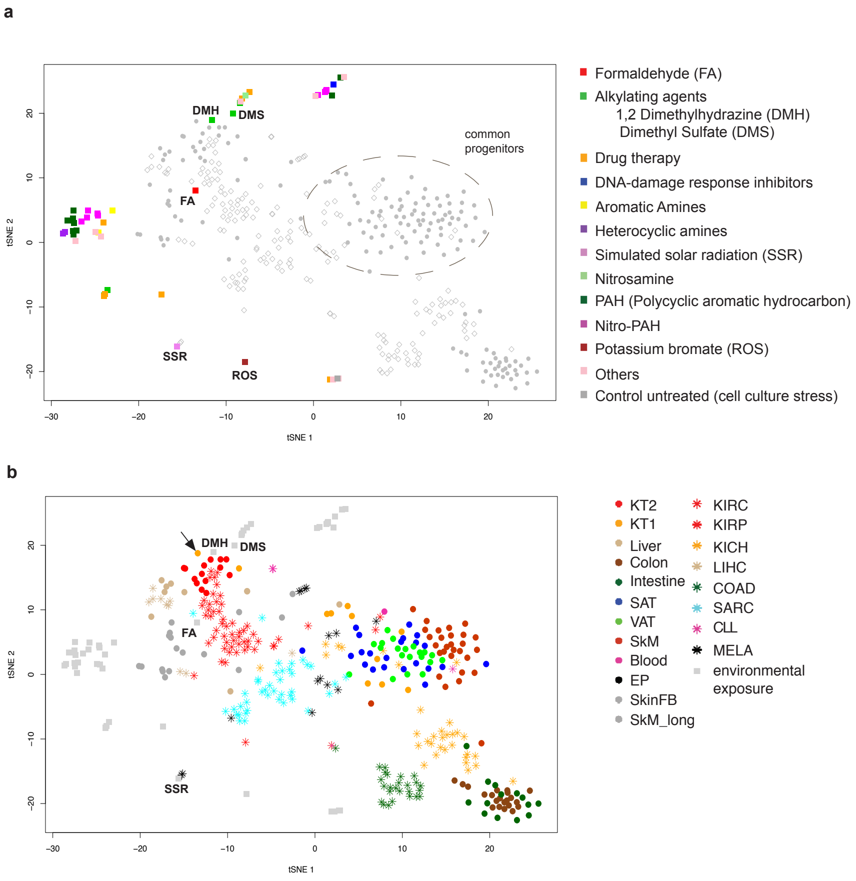
BLCA-ercc2del: bladder urothelial carcinoma with *ERCC2* knock out, BLCA: bladder urothelial carcinoma, EP: epidermis, Fb: skin fibroblasts, MELA-AU: melanoma, CLL-ES: chronic lymphocytic leukemia, SAT: subcutaneous fat, VAT: visceral fat, SkM: skeletal muscle, SkM_long: long-culture SkM cells, SARC: sarcoma, KT1: kidney tubule 1, KT2: kidney tubule 2, KICH: kidney chromophobe, KIRC: kidney renal clear cell carcinoma, KIRP: kidney renal papillary cell carcinoma, LHC: liver hepatocellular carcinoma, COAD: colon adenocarcinoma.

Fig. S4. Comparisons of cancer and normal samples



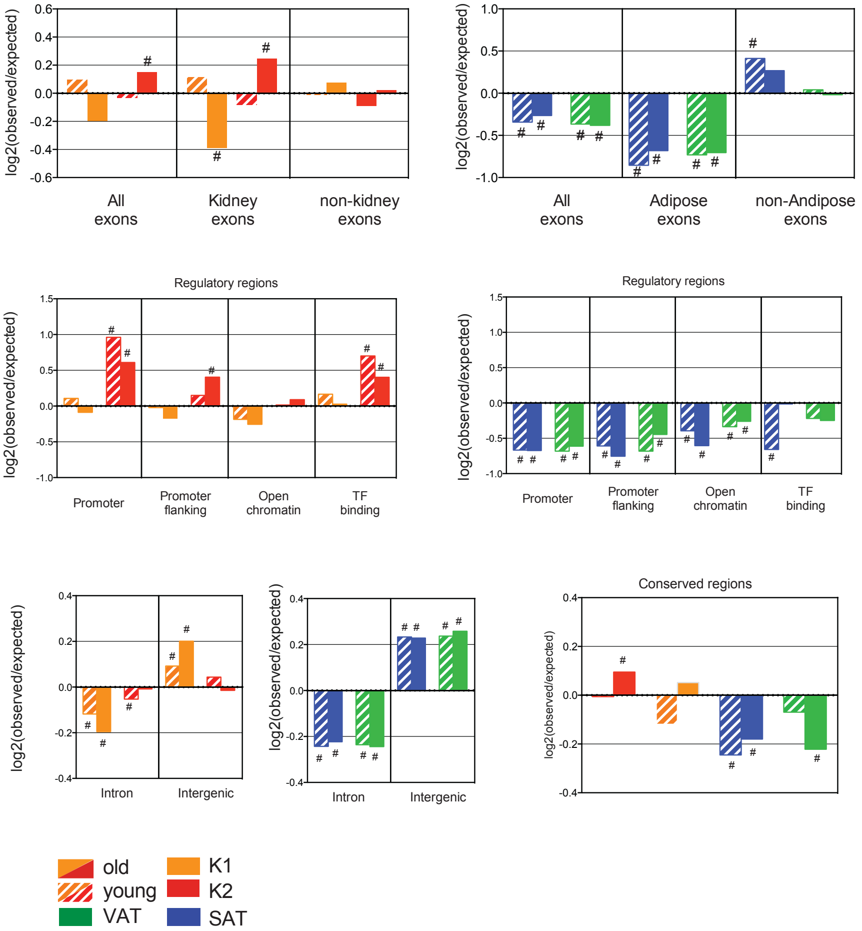
a. Comparison of somatic mutation profiles in tissue-matched healthy and cancer samples. The clustering (tSNE) based on the trinucleotide profile of somatic SNVs in the genome of healthy (n=161) and tumor (n=192) samples is shown. For each panel, different healthy and cancer samples are highlighted with specific colors (see legend), while all other samples are shown in grey. Cancer samples usually cluster in proximity of the tissue-matched healthy samples, but cancer and normal do not overlap. Bottom right panel shows the matching of two groups of healthy samples that shared a long culturing protocol: reprogrammed skin fibroblasts (SkinFB) and long-culture skeletal muscle progenitors (SkM-long). **b.-e.** Number of SNVs per genome, plotted according to age. Mutation burden (**b.)** and number of SBS5 mutations (**c.**) in normal kidney (KT2) and liver samples compared to cancer samples (bladder urothelial cell carcinoma) either NER proficient (BLCCs) or deficient (BLCC-*ERCC*del). The ERCCdel tumors were used as a control for SBS5 mutations induced by NER deficiency. Mutation burden (**d.**) and number of SBS40 mutations (**e.**) in normal kidney (KT1 and KT2) compared to kidney cancer samples of different subtypes: KICH (kidney chromophobe adenocarcinoma) KIRC (kidney clear cell renal cell carcinoma) KIRP (kidney renal papillary cell carcinoma).

Fig. S5. Comparison with mutation spectra determined by *in vitro* exposure to environmental agents



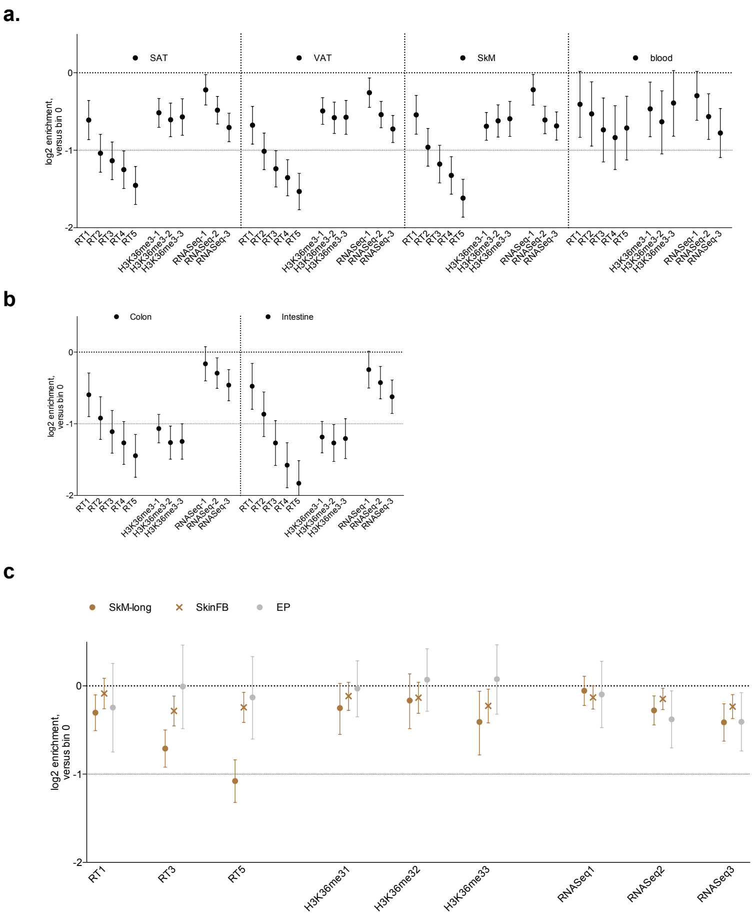
tSNE plot of the trinucleotide profile of somatic SNVs recovered in the genome of healthy cells (n=161), tumors (n=192) or an iPSC clone exposed to different environmental agents *in vitro* [3] (n=54). **a**. Environmental agents are highlighted with colors representing the different compound classes, while all other samples are shown in grey (normal: full dots, tumor: empty squares). A dashed line roughly describes the area occupied by common progenitors. The mutation spectrum of common progenitors does not show similarities with any spectra caused by environmental agent exposure, supporting the concept of basal mutagenesis. **b**. Same plot as in **a.**, but environmental agents are shown in grey, while normal (squares) and tumor (asterisks) genomes are shown in different colors according to the tissue of origin. The vast majority of spectra from treated cells located at the periphery of the plot and did not overlap with any normal or cancer genome (Figure S5a). Exceptions were 1) simulated solar radiation that perfectly overlapped with EP samples and one melanoma sample, 2) formaldehyde and alkylating agents, which located in proximity of KT2 and kidney tumors KIRC and KIRP (Figure S5b). Formaldehyde and multiple compounds with alkylating activity can be produced endogenously by human cells [8]. Therefore, the spectra of KT2, KIRP and KIRC might reflect the exposure of some kidney cells to endogenous formaldehyde and alkylating agents.

Fig. S6. Mutation enrichment in specific genomic regions in KT1, KT2, SAT, VAT samples and age-related differences



Enrichment (upward bars) or depletion (downward bars) of somatic mutations in indicated VEP genomic features or conserved regions in different tissue and age-groups. Kidney-1 (KT-1). kidney-2 (KT-2). subcutaneous fat (SAT) and visceral fat (VAT)

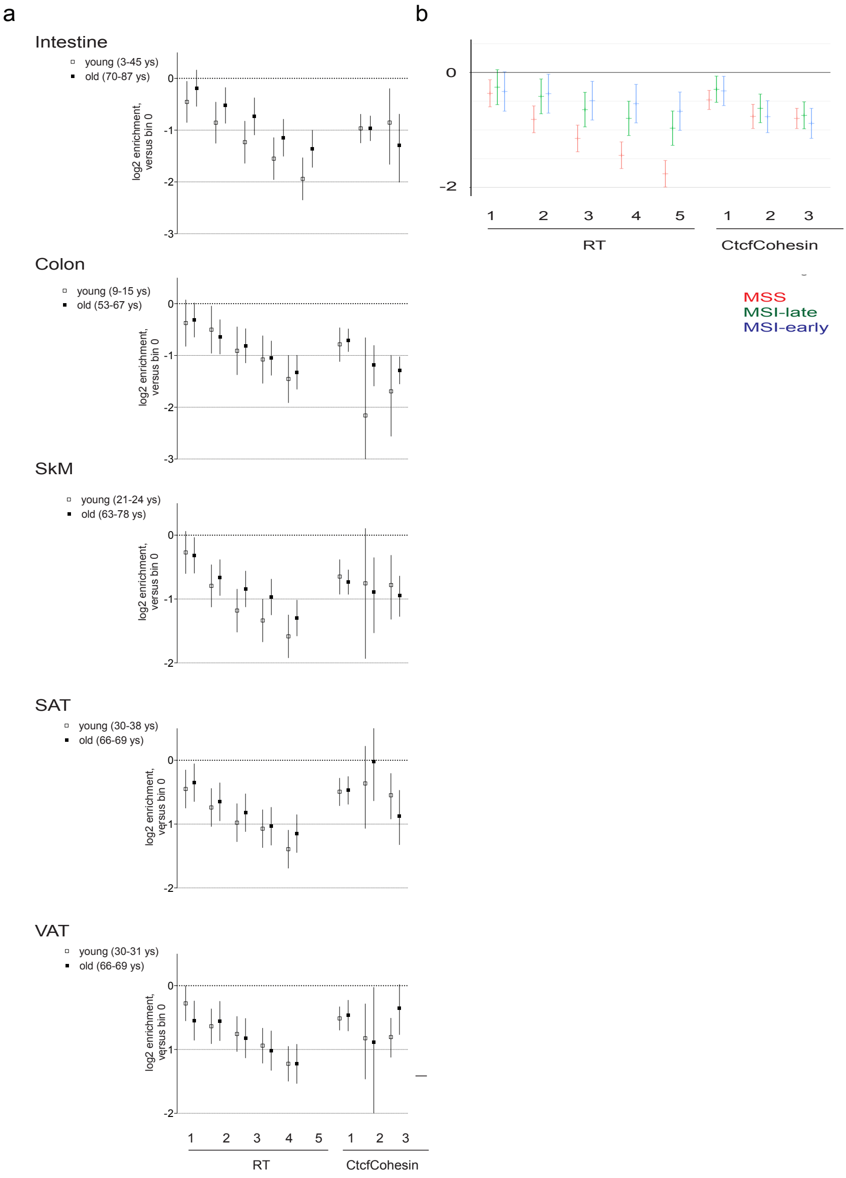
Fig. S7. Analysis of regional enrichment/depletion of mutations in different tissues



a.-c. Enrichment/depletion of mutations in specific genomic regions, as shown in figure 5a, but providing values either calculated separately for each tissue (**a.** common progenitors: SAT (N=22), VAT (N=20), SkM (N=29), blood (1 catalogue of mutations derived from randomly selected SNVs from multiple cell clones from the same individual)**-b**. intestinal stem cells: colon (N=21) intestine (N=14)) or from sample groups not shown in figure 5a (**c.** SkM-long (N=4), SkinFB (N=13)).The genomes were divided in multiple sectors (bins) according to decreasing DNA replication time (RT, bins 0 to 5, only bins 1, 3 and 5 are shown for clarity), increasing abundance of the histone mark H3K36me3 (bins 0-3), and increasing transcriptional levels (RNAseq, bins 0-3). The relative abundance of mutations in each bin vs bin 0 is estimated as the coefficient in negative binomial regression (expressed as $\log_2$), where error bars show its 95% C.I.

Common progenitors, including SAT, VAT, SkM and blood, but not KT1, showed the expected depletion of mutations with earlier RT, higher H3K36me3 abundance and higher transcription levels. This pattern indicates that the basal mutagenesis is actively counteracted by MMR and/or TC-NER. However, EP, KT2, KT1, liver, SkM-long and SkinFB deviated from the pattern seen for common progenitors and showed a loss of association of mutation rates with RT and H3K36me3. Therefore, in samples that appear to be exposed to a putative mutagen in addition to basal mutagenesis (Figure 3a and b), the early-replicating, active chromatin is less protected. These samples included the KT1 group, which showed a mutation profile similar to the common progenitors (Figure 4a), but also signs of cell damage (Figure 4f). Conversely, the intestinal stem cells (intestine and colon) showed regular association of mutations with RT and even stronger protection of H3K36me3-rich regions compared to common progenitors, suggesting that mutations in the active chromatin that are due to high proliferation are recognized by MMR.

Fig. S8. Association of mutations with replication timing in young and old genomes of healthy samples and MMR-proficient (MSS) or deficient (MSI) tumors



a. Enrichment/depletion of mutations according to DNA replication timing (RT) while controlling for CTCF ChipSeq peaks in either younger or older genomes as shown in figure 6c, but providing values calculated separately for each tissue. Enrichments are coefficients from negative binomial regression (as log2) and error bars are their 95% C.I **b.** Enrichment/depletions as in **a.** for 3 different groups of tumors (derived from colon, uterus, stomach) according to microsatellite stability. MSS= micro satellite stable, normal MMR function; MSI= micro satellite instability due to mutations in MMR genes which occurred with either early or late onset in the life of the patient. Fold-difference in depletion of mutations according to RT were 1.73 for MSS vs MSI-late and 2.13 for MSS vs MSI-early, showing that inactivation of MMR induces accumulation of mutations in early-replicating DNA that increases with time. These tumors were used as a control of the effect size of MMR-loss in causing mutations in early-replicating DNA. The fold-difference in young vs old healthy genomes (pulling together all tissues as in figure 6c) was 1.21, lower than that observed in MSI tumors, in agreement with only partial loss of MMR function with aging.

Chapter 5

Variable DNA methylation underlies mutation rate variability at the mesoscale in human somatic cells

Variable DNA methylation underlies mutation rate variability at the mesoscale in human somatic cells

David Mas-Ponte[1] and Fran Supek[1,2]

1) Institute for Research in Biomedicine (IRB Barcelona)
2) Catalan Institution for Research and Advanced Studies (ICREA)

The cytosine methylation in CpG dinucleotides is pervasive in mammalian genomes and its variability across regions can regulate gene expression and define cell differentiation. Although the role of DNA methylation in gene regulation is well understood, how the local variation in DNA methylation shapes somatic mutation rates is less well explored. Here, we show that hypomethylated (UMR) regions are also generally hypomutated in a wide range of human tumors and healthy somatic tissues. Remarkably, the exposure of the tissue to various mutational processes shapes its predisposition to this effect: while there is depletion in the mutation rates resulting from signatures of deamination of methylated cytosines, UV light, *POLE* and MMR deficiency, there is an increase in mutation rates from signatures of AID/APOBEC cytosine deaminase enzymes in the UMRs. Therefore, hypomethylated DNA loci can be either mutational coldspots or hotspots, depending on the mutagen exposure history of a particular cell. In addition to these genome-wide distributed UMRs we also identify several kilobases at the 5' ends of gene bodies as commonly hypomethylated and thus hypomutated. Clustering genes by methylation profiles also yielded variability in their mutation rate gradients along the gene body. Interestingly, lowly expressed genes have a less steep gradient due to a higher relative methylation of their 5' end, and polycomb repressed genes also show no relative hypomutation due to the lack of methylation at their gene body. Overall, we suggest DNA methylation is an important determinant of mesoscale, sub-genic, resolution mutation rate variability in human somatic tissues.

Introduction

In humans, CpG dinucleotides in DNA are usually methylated at the cytosine nucleobase and have, globally, a low frequency in the genome. However, they are particularly enriched near transcription start site (TSS) and other functional elements. The accumulations of these CpG loci are known as CpG islands (CpGi), and they play an important role in the regulation of the adjacent gene where they are located. When the CpG island is methylated, transcription factors binding to the promoter is altered, and often reduced, effectively switching off the gene's expression. This mechanism has a strong silencing capacity and is commonly used in mammals to regulate the expression of developmental genes[1−3].

The genome can be segmented according to the methylation level of the CpG dinucleotides in multiple ways. A parsimonious segmentation, such as by the Methylseeker algorithm, classifies the genomes in unmethylated (UMR), low-methylated (LMR) and fully methylated regions, or the rest of the genome[4]. The UMRs are high density CpG loci which are completely unmethylated while LMRs maintain a medium level methylation (~30%) and present a lower concentration of CpGs in their sequence. While the UMRs are specifically associated to the promoter regions of genes, LMRs are more intergenic and enriched in enhancer marks such as H3K4me1[4,5]. Other definitions of the undermethylation in the genome offer different classifications in how methylation is regulated. For instance, strong DNA hypomethylation can also be detected in large sections (bigger than > 3.5Kbp), termed canyon UMRs (cUMRs) are associated with developmental genes like the Homeobox family[6]. Other reports suggest that in both ageing tissues and cancer cells large domains (in the megabase scale) also lose their normal methylation. These domains are named partially methylated domains (PMDs), overlap late-replicating DNA domains, and they are thought to lose their methylation passively due to the imperfect methylation maintenance[7].

The interaction between mutations and DNA methylation was identified early with the first sequenced human genomes[8]. Cosmic signature 1, or SBS1, was the first identified mutational signature, proposed to result from deamination of the methylated cytosine at CpG sites[9,10] primarily due to its sharp profile at NCG>T contexts. Signature 1 also accumulates with age, is present in most healthy tissues[9,11] and is also commonly observed in *de novo* germline mutations[12], highlighting its pervasive implication in the genomic integrity of the human genome. Other mutational processes have also been associated previously with DNA methylation in cancer. In particular, the mutations resulting from deficiency of DNA polymerase ε and the deficiency of mismatch repair (MMR) activity have both shown associations with the methylation status of the mutated regions[13,14]. Contrary to the mechanism of signature 1, these mutagenic processes upon DNA repair failures are thought to be associated with the misincorporation of nucleotide bases in methylated sites during DNA replication[15]. A clear evidence of this role is the characteristic replication strand bias of signatures 10b and 15, which are associated with pol ε and MMR deficiencies respectively. Other epigenetic modifications in the CpG dinucleotides also modify the mutation rate in different ways, for instance, stable hydroxymethylated (5hmC) loci show an strong depletion in C>T mutation accumulation particularly for somatic tissues and increase C>G rates [16,17]. 5hmC is considered an intermediate in the process of demethylation of the CpG, which transforms a 5mC base to multiple oxydised modifications mediated by the TET enzymes.

Thus, combining the DNA methylation-aware genome segmentation and the known modulation of the mutation rate in tumors we hypothesize that there may be a yet uncharacterized variability in somatic mutation rates at the kilobase scale with a strong overlap with genes and regulatory elements. Here, we perform a systematic analysis of the mutation rate variation along UMRs, LMRs and gene bodies in order to quantify the role of DNA methylation in generating genome-wide mutational gradients, which differ across mutational signatures. We also quantify the role of DNA methylation in other functional elements, such as enhancers and chromatin loop anchors, that while not associated with genes, also exhibit hypomethylation and consequently lower mutation rates from selected mutational processes.

Results

Sub-genic mutation rate gradients originated mostly from DNA methylation associated signatures

In order to systematically analyze the sub-gene resolution variability of mutation rate in genes, we calculated the mutation rate for each mutational signatures across segments of genes covering both gene ends and an extended region flanking them. Each signature was divided by the tissue of origin and genes where further stratified into three bins by their average expression levels (see supplementary methods)., We estimated the mutation rate, controlling for trinucleotide composition of different regions using a negative binomial regression (see methods) and extracted the dominant patterns using a principal component analysis (Fig. 1A,B. The first principal component accounted for 38% of the systematic variability (Supp.Fig. 1A) and its profile along the gene body presented a sharp increase at the TSS (Fig. 1C). The second component explains substantially less variability (6%) and is less enriched at the TSS, but more so consistently enrichmed along the gene body and until the transcription end site (TES) (Fig. 1C).

The first component is characterized by a lower mutation burden from from signatures SBS1, SBS15 and SBS10b (Fig. 1B). Each of these signatures contains a significant NCG>T component in its trinucleotide profile, and each has been previously associated to the role of DNA methylation, either genome-wide for signature 1, or along the gene promoters form the dMMR-associated SBS15 mutations[13] (Fig. 1B, Supp. Fig. 1B). An association with DNA methylation would also fit with the difference observed between gene expression bins, higher expressed genes showing higher values and positive correlation (Fig. 1B and Supp. Fig. 1B,C). If the observed gene gradient of mutation rates summarized in PC1 was generated via the hypomethylation of the promoter in the promoter region, expressed genes which show a more evident hypomethylation would effectively also show a stronger mutation depletion. This is also consistent with highly expressed genes being more enriched in CpG island type promoters[18], which are more commonly unmethylated.

Overall, the result of this systematic analysis suggests that DNA methylation associates with the mutation rate gradient along gene bodies, specifically for mutational signatures with clear components of CpG dinucleotide mutagenesis.

Unmethylated regions show consistent hypomutation in multiple tissues

To characterize the role DNA methylation in the modulation of mutation rates in various genomic loci we focused on the genomic segments that are consistent hypomethylation. Consistent DNA hypomethylation can be detected in the unmethylated regions (UMR)s, with a complete lack of DNA methylation, and low methylatd regions (LMR), with lowly methylated regions[4,5] (Fig. 1D).

We curated a set of hypomethylated regions in the human genome from previous publications[5,6] (see Supplementary Table 1). Additionally, we collected genome-wide methylation data from WGBS experiments available in public repositories (Roadmap and Encode). From the downloaded WGBS experiments, we called UMR and LMR loci using the same methodology as in ref[5] (see methods and Supplementary Table 1). While the published datasets[5] contained 18 tissues and represented mostly stem cells and blood cell lines, here we focused on 34 diverse solid tissues, 6 blood and 4 brain tissues that will represent better the methylation patterns in most sequenced tumors (see methods; the solid, blood and brain tissue groups are treated separately) . In total, the union of all obtained sets of hypomethylated regions covered 40Mbp (Supp. Fig. 1D).

We measured the mutation rate in these regions across different tissue types for a set of tumor and healthy samples (see mehtods) from the PCAWG dataset[19] and other sources from the literature[20–22]. The majority of surveyed tissues, except the urinary tract and the lymphatic blood, showed a significant reduction of the mutation rate at UMRs and LMRs, with an average depletion across tissues of 25% (Fig. 1E). This reduction was substantial for tissues with a high proportion of SBS1 mutations, like colon and brain[10]. Skin cancers also showed a significant reduction in mutation rate, consistent with a previously proposed role of DNA methylation in the predisposition of UV damage mutations (Fig. 1E)[23]. These associations were highly correlated when tested on different sets of UMRs, both the ones obtained from the literature and the ones computed in this study (Supp.Fig. 1E,F).

Considering the signature-classified mutations, in a pan-cancer setting, mutational signatures SBS1, 10b and 15 decreased the most, mirroring previous analyses[13,14]. UMRs contained on average 75%, 65% and 55% less mutations than expected by trinuclotide composition, for SBS10b, 1 and 15, respectively. Other mutational signatures like SBS6, related to MMR deficiency, and SBS7a also showed a high reduction of mutations (Fig. 1F).

Surprisingly, certain signatures showed an increased mutation rate at UMRs. The most anticipated case from these was SBS84, associated with the activity of the Activation-Induced cytidine

Deaminase (AID) in the somatic hypermutation process at immunoglobulin sites[24]. AID mutations showed an increase equivalent to 4x times over the expected values. Three other signatures, SBS9 (also associated to SHM in lymphoid tissues, possibly in part reflecting the activity of polymerase η), and SBS2 and SBS13 (associated to APOBEC3 mutagenesis) also showed a moderate enrichment in the UMRs (~19%) (Fig. 1F).

In order to verify that the mutation reduction was directly caused by the drop in the methylation level, we used a set of UMRs, which contained specific sites enriched only in a given set of tissues, comparing with tissue-specific hypomutation at these sites. Although the separation of tissue specific UMRs was not very specific (Supp.Fig. 1G), potentially due to the heterogeneity of the selected tissue groups, our samples showed a significant depletion of methylation for the corresponding tissue set where the cancer sample was originally coming from (Fig. 1G) . For instance, the depletion of mutations in UMRs specifically extracted from solid tissues was of 30% for colon cancers and blood while it was reduced to no change for brain. Similarly, the reduction of mutation rates in the brain specific UMRs was 18% in brain tumors but only 12% and 6% for colon and blood myeloid.

Overall, the reduced methylation level at UMRs seems to be responsible for a reduction of mutations in a wide range of signatures but can be also associated with an increase for others. The observed variability at the tissue level, thus, might be explained by to what signatures the tissue is normally exposed.

Interaction of mutation rate at functional elements

Due to the characteristic hypomethylation of multiple regulatory elements like promoters, enhancers and loop anchors, we used these annotations to classify the extracted UMR sets to ask whether the methylation effect on mutation rate is different across functional elements (Fig. 2A). As expected from prior work, UMRs were enriched in promoters while LMRs showed a bigger predisposition to enhancers, measured as the odds ratio (Fig. 2B). Additionally, we find that chromatin loop anchors are also often hypomethylated, and that this effect is independent of them containing a known promoter or enhancer. Prior UMR sets showed very similar associations to these functional elements as the ones called in this study, being consistent between tissues and methodologies.

The highest number of UMRs was explained by promoters and 5' gene body ends. However, a total of 1,925 UMRs (or 10% of the total set) did not overlap with any of the functional element tested (Fig. 2C). For LMRs, this value was higher and up to 52% of the instances did not overlap with any functional element (Supp. Fig. 2A). These values are overall consistent with previous estimates for each class of segment[4].

We then asked if the reduction in mutation rate seen above analysis was, in part, due to these associated functional elements, rather than hypomethylation itself. For every tissue, we selected the UMRs that overlapped with either loop anchors or by the region around the TSS (defined as 2kb upstream and 1kb upstream) and removed them from the UMR set of interest. Although the reduction of mutations was less pronounced in UMRs not overlapping promoter/enhancer/LAP, the overall trend of hypomutation was still evident both across tissues and signatures, suggesting that the mutational effect of DNA methylation is independent of its overlap with promoters or LAPs (Fig. 2D,E).

In the converse analysis, measuring mutation rates in promoters with and without an associated UMR, however, the relative mutation rate showed a clear dependence on DNA hypomethylation. Only the promoters that overlapped significantly with an UMR showed a substantial mutation rate depletion. In brief, mutations were reduced up to 40% when considering all promoters in colon and skin cancers (Supp. Fig. 2C). Of note, this reduction was not as striking as when measuring the UMR alone, potentially due to only a partial matching of the actual unmethylated loci with the annotated promoters. When considering mutation rates in promoters that did not overlap with UMRs the mutation rate was not reduced (Fig. 2C). This observation highlights the direct role of DNA methylation in the determination of mutation rate at these sites. To explore if the effect of the UMR on mutation rates was indirect and resulted from the increased expression of genes with a UMR, we repeated this analysis after stratifying genes by expression tertiles (Supp. Fig. 2D). However, for colorectal and skin tissues, which contained sufficient mutation counts, the mutation rate in genes with high expression values but without overlapping UMR was (not reduced), suggesting transcription is not responsible for the mutation rate decrease. The relative mutation rate in the two highest expressed bins (Eq2 and Eq3) was equivalent and significantly reduced compared to their UMR-less counterparts of same expression level (Supp. Fig. 2D)., supporting the known effects of transcription on reduced mutation rates independently of DNA methylation. Also of note, some tissues like liver (Supp. Fig. 2D) did show reduction of mutation rate in higher expression bins, suggesting a role of transcription-coupled mutational processes, in this instance probably transcription-coupled mutagenesis as reported for liver[25]. Even with this strong role of transcription in the liver, mutations were still reduced in UMR overlapping promoters (Supp. Fig. 2D). In summary, DNA hypomethylation affects mutation rates in a manner independent of other features that may be present at regulatory elements and independent of transcription levels.

Epigenetic types of UMRs highlight different mechanisms of mutation rate control

In order to examine the role of other molecular factors that are known to associate with mutation rate we classified the pooled UMR dataset according to the accumulation of certain histone modifications, henceforth epigenetic profiles. This classification of UMRs represents an annotation-free classification and can clarify the mechanisms related to the mutation rate depletion.

The histone mark classification of the UMRs yielded two groups (Fig. 3A and Supp. Fig. 3A, one associated with increased H3K4me3 and reduced H3K36me3, consistent with a active promoter marks and one associated with H3K27me3 consistent with polycomb repression. A 33% of the UMRs was classified in the active group while the rest was classified as repressed (H3K27me3-enriched). The methylation levels in the active promoter-like UMRs contained a stronger hypomethylation while the H3K27me3-enriched showed more moderate hypomethylation (Fig. 3B). This difference in methylation between the 2 groups could be explained either by the overall increase of the methylation level across samples. Mutation reduction followed the same trend as the methylation levels, with a stronger depletion for the active promoter-like UMRs (Fig. 3C and (Supp. Fig. 3B).

Gene stratification according to methylation levels reveal differential mutational gradients

In order to systematically test if the hypomethylation, and the consequent hypomutation, would be relevant for the estimation of the mutation burden in genes

Because of the overlap of the hypomethylated segments genome-wide with the promoter regions and the 5’ ends of genes, we hypothesized that different groups of genes might show distinct patterns in their methylation levels and thus contain different mutation burdens across their gene body. To test this, we used the same DNA methylation data averaged along multiple solid tissues (see methods and Supp. Table 1) to profile the methylation levels along each gene body, and then cluster genes by the shape of DNA methylation profiles. In brief, gene bodies were segmented in 50bp bins extending the TSS and TES within-gene for 5kb and outside-gene extending for 2kb. For each gene, methylation level was averaged across every bin. The resulting profiles were then analyzed using a PCA (see methods, (Supp. Fig. 4A). Expectedly, the resulting principal components correlated to some extent with the average expression (Fig. 4A and (Supp. Fig. 4B). We used the three first components of the PCA (together accounting for 27% of the variability) to cluster genes into five groups. These three principal components represented the methylation levels globally in the gene body (Dim.1) the TSS methylation status (Dim.2) and the upstream and downstream methylation levels outside the gene (Dim.3) (Fig. 4B,C).

The obtained gene clusters were characterized by distinct genomic characteristics (Fig. 4E and (Supp. Fig. 4C,D). Cluster 1 (c1) and to some extent cluster 2 (c2) contained genes with a methylated promoter and were overall repressed. The main difference between these two clusters of genes was their average expression, with a lower median expression for c1. Cluster 5 (c5) contained generally short genes with and overall unmethylated gene body, they were enriched in polycomb marks like H3K27me3 (Supp.Fig. 4E,F). The homeobox genes, which have been previously described as a set of unmethylated developmental genes with roles in cancer[6] were included in this cluster (Supp.Fig. 4G).

Cluster 3 and 4 represent each a set of highly expressed genes with strong hypomethylation in the promoter region, as expected, however we here note also that hypomethylation extends into the 5' end of the gene body, approximately 1.5kb (Fig. 4D). The main differences between these groups are the extent and the position of the unmethylated region. C2 has a narrow unmethylated segment (~1kb) while c3 extends it downstream towards the gene body (up to a total of ~3kb), c4 has an extended hypomethylated region directed at both upstream and downstream sections of the TSS marking an overall wider promoter region (Fig. 4D and (Supp. Fig. 4E,F).

To further characterize these genes, we measured their overlap with chromatin states (according to ChromHMM, see methods), the existence of CpGi[18] and the normalized CpG content in their promoters[26], similar to the definition of CpGi (Supp.Fig. 4C). C1 was the only group underrepresented in the active transcription segments and showed a clear enrichment in polycomb repressed genes and in H3K9me3 heterochromatin (Fig. 4E). While c2, c3 and c4 did not show strong enrichment for any chromatin states, c2 was characterized for a depletion of genes with CpG islands (nor genes with a strong enrichment of CpG dinucleotides in their promoter region) while c3 and c4 were enriched in these CpG island categories. C5 showed a strong enrichment in the bivalent transcription chromatin (Supp.Fig. 4E).

We also measured the averaged histone profiles of each gene category (Supp. Fig. 4F) observing a strong increase of promoter marks (H3K4me3 and H3K27ac) for c3, c4 and c5 and to a lesser extend c2. H3K27me3 was particularly enriched in c5, consistent with the bivalent transcription enrichment in the chromatin states analysis. Based on this histone profiling data analysis and the overlap with nascent transcription (suggesting enhancer activity), we infer that the main distinguishing feature of the c3, c4 and to some extent c5 gene body methylation clusters is the overlap with enhancer features. This suggests that gene body hypomethylation profiles are commonly shaped by the existence of genic enhancers.

While gene cluster C4 contained a significant enrichment of enhancer nascent-transcription signal both upstream and downstream the TSS, in c3 only covered the downstream enrichment (within the gene body) (Supp. Fig. 4E). The accumulation of these genic enhancers might thus, as in the c3 group, cause the unmethylated region to extend uniquely in a single direction towards the gene body. The local accumulation of H3K4me1 (Supp. Fig. 4F) in these groups was also consistent with this classification.

Overall, the methylation profiling of genes yielded 5 distinct groups with specific epigenomic characteristics. C1 cluster contains the ‘classical’ repressed genes with a methylated promoter; c2 genes contain a short unmethylated region in the TSS and are generally less enriched in CpG islands; c3 genes contain a wider unmethyalted region that extends downstream of the TSS potentially due to genic enhancers; c4 genes contain that and also a wider unmethylated region at the TSS, which extends both downstream and upstream of the TSS potentially due to the overlap with a broader enhancer region and a partial bidirectional transcription; c5 genes represent the least numerous group and contain generally unmethylated short genes with enrichment in polycomb marks.

Subgenic mutation rate gradients in methylation based subgroups

It is interesting to jointly consider the association between DNA methylation and mutation rates of selected signatures shown above, and the stratification of gene populations according to their methylation profiles. Based on this, we hypothesized that the mutational gradients along the gene body and the TSS would not just depend on the mutational signature, but also the shape of the DNA methylation profile in the gene. We therefore repeated the mutation rate analysis along the gene bodies, asking if this differs for genes in the different methylation profile clusters. The genes with active demethylation at or nearby their promoters -those in clusters c2, c3 and c4- showed a stronger depletion of those signatures associated to mutation rate depletion at UMRs, mostly SBS1, SBS10b, and SBS15. Conversely, also mutation signatures that favor hypomethylation at UMRs, SBS2, SBS13 and SBS9, also showed an increase rate around promoters. This was, however, more moderate (Fig. 5B and Supp. Fig. 5A). Mutation rate was constant across gene bodies for the c1 group, consistent with the constant methylation levels across the promoter section of c1 genes (Fig. 5A).

Overall, differential enrichment of the mutational signature along the gene body considering grouped genes by methylation clusters was similar to the initial, unsupervised gene profile analysis (Fig. 1A) . This suggests that the main determinant of the variability in mutation rate along the gene bodies is DNA methylation but that it does not uniformly affect genes or mutations. More highly expressed genes, and genes with intragenic enhancers, will have more prominent and wider mutational coldspots at their 5' ends, respectively, when considering common mutational processes such as aging-associated SBS1. These trends are reversed for AID/APOBEC mutagenic signatures, which are enriched at hypomethylated promoters and adjacent intragenic enhancers.

Methylation based gene stratification can prune baseline mutation rates

Methods to detect signatures of selection on somatic mutations rely on an accurate baseline of regional mutation rates, to be able to establish whether there is an excess or dearth of mutations over that baseline, signifying positive or negative selection, respectively.

Gene methylation profiles and mutation signatures could be considered in order to establish better and more accurate baselines for mutation rates that account for the sub-gene-resolution variation in mutation rates. In order to test effects of methylation-aware baselines for mutation rates, we built a model to predict the mutation burden of a gene from the TCGA exome data. Because mutation rates at genes are known to be heavily influenced by the epigenetic state and the replication domain where they are located[27], we predicted mutation rates from the epigenomic covariates from dNdScv method[28] as a base model. We then compared this base model with one containing the methylation gene clusters defined above, and as negative control on where these gene clusters were randomized (Fig. 5C, Supp. Fig. 5B and methods). Calculating the goodness of fit of the model by the average root mean square error (RMSE) of 5 k-fold cross validation runs showed a decrease in the error measure for the methylation-aware model compared to both the base (covariate-only) model and the shuffled feature (Fig. 5D). Using the predicted number of mutations from this model, we can calculate the excess of mutation burden of every gene, (Supp. Fig. 5D) which is a measure of positive selection. As expected, the mutation excess in the cancer driver genes, labeled as positive, was significantly higher than in the non-cancer genes, when measured in the testing set. Reassuringly, there was no significant change in mutation rates however between the different models (Fig. 5E) when considering non-cancer related genes (most of which are not selected).

The expected mutation burden however differed significantly when considering the methylation gene clusters as different groups (Fig. 5F). The mutational burdens were corrected towards lower values for genes in the c3 group while they were corrected towards higher values for genes in cluster c1 and c2. Our model is able to capture this information and consequently correct the estimated expected burdens. Overall, we suggest that shapes of DNA methylation profile should be formally included in models for testing selection on somatic mutations.

Discussion

This study highlights the role of locally variable DNA methylation in the modulation of mutation rates, particularly, around hypomethylated regions, such as UMRs and LMRs. For many mutagenic processes, such as the ubiquitous cell division-associated (and thus aging-associated) C>T process dependent on spontaneous cytosine deamination, these generate mutation coldspots. However for AID/APOBEC mutagenesis, the local hypomethylation instead generates mutation hotspots.

Due to their overlap with the TSS and, often, the 5' end of the gene, this local hypomethylation can also represent an important determinant of the overall mutation burden of a gene, as well as of other functional genomic elements such as enhancers and loop anchors. Due to this effect, incorporating information on differential methylation profiles of genes (here, implemented via clustering), or explicilty considering the methylation status of a genomic region-of-interest may provide a better estimation of their baseline mutation rates. We suggest DNA methylation can complement existing covariates used to predict mutation rates, mainly based on coarse-resolution features such as replication time, or heterochromatin status, or expression level of the gene.

Generally in UMRs and to a lesser extent in LMRs, we find strong associations of the methylation status of the CpG dinucleotides with the mutation burdens of signatures, SBS1, SBS15 and SBS10b, as anticipated[13,14], and to a certain extent also associates with other signatures like UV-induced SBS7 (negatively), and AID/APOBEC associated SBS2, SBS13, SBS9 and SBS84 (positively). Mechanistically, the mutation rate association in polymerase ε and MMR-deficient tumors was suggested to derive from an incorrect incorporation of the corresponding nucleotide when methylated[13,15]. On the other hand, the SBS1 signature mechanism, widespread in most healthy and cancerous tissues, is associated with the increased spontaneous deamination rate[2,29,30] when methylated and/or by the more difficult repair of the deaminated cytosines if they are methylated.

The mechanism underlying SBS7, UV-mediated damage formation, has been reported to interact with DNA methylation in a diverse set of mechanisms, from the increased lesion formation in methylated DNA[23,31] to the faster deamination of the dipyrimidine lesion. Mutations in melanoma skin cancer, usually predominantly from SBS7, associated non-linearly in genome-wide correlations[13] with DNA methylation and are known to be modulated by other factors confounded with promoter hypomethlyation, such as transcription coupled repair, and also chromatin accessibility promoting repair[13,32,33]. Our approach, focusing on regions with significant methylation depletion, shows a depletion of UV-associated mutagenesis in UMRs of 45% over the expected rate. Importantly, we find this UV hypomutation is likely due to hypomethylation rather than other genomic features associated with it, for instance higher mRNA levels (and presumably higher transcription rates of the promoter and gene body). Because of known ability of TC-NER in clearing UV damage, we checked the hypomutation in promoters and 5' gene ends with UMRs, stratifying by different expression levels (Fig. 1). In skin, this revealed a similar pattern as the one seen in colorectal cancers (enriched in SBS1 but no UV damamge), where both promoters of both the lowly and the highly expressed genes showed similar levels of hypomutation, suggesting that the hypomethylation rather than transcription underlies the reduced UV mutagenesis at promoters. (We note that in certain, narrow loci within some promoters, which binding the AP-1 family transcription factors, there is increased UV mutagenesis due to increased damage accumulation[34,35]).

Contrary to cell cycling-associated SBS1 signature, and UV-associated SBS7, certain other mutational processes showed increased mutation burdens in hypomethylated regions. The APOBEC mutational signatures SBS2 and SBS13 showed an increased mutagenesis of ~19%.Its interaction with DNA methylation was proposed[36] consistent with our observation. The SBS9 association may be mechanistically linked to the somatic hypermutation process, which involves AID followed by error-prone repair, and predominantly targets promoters of immunoglobulins and, as off targets, a subset of other high expressed genes, and would be thus associated -- directly or indirectly -- with demethylated sites as well. A further explanation is suggested by the enrichment of SBS84 signature, which is characteristic for the AID mutagenesis[37]. The AID protein participates as the first step in the SHM process in B cells. Interestingly, however, AID was also suggested to participate in an active DNA demethylation mechanism[38], where AID damage can trigger eventual repair back to an unmethylated C[39]. This mechanism would be consistent with the strong correlation between AID and UMRs reported in this study.

In conclusion, different mutation signatures have unique interactions with local methylcytosine , causing either an increase or a decrease of mutation rate at unmethylated sites, depending on the signature. The variability of effects in DNA methylation observed across tissues (Fig. 1E) may therefore be generated in part both by tissue-specific DNA methylation patterns, and also by the differential exposure to mutational signatures in different tissues (Fig. 1F).

Because of the high enrichment of UMRs in active gene promoters and in 5' ends of active genes (FIG), the reduction of mutation rates at these sites can affect the estimation of the baseline mutation rate in genes. Current approaches to the detection of selection in genes are based on the estimation of a mutation baseline from various covariates (replication time, gene expression and others) which is then compared against either the distribution of the observed mutation density[40], the mutation spectra[41] or the type of aminoacid substitution[28]. In either case, baselines are typically established at the gene level and do not consider variation within the gene body. Here, we show that mutation rates change within the gene body, in function of the methylation level particularly in the TSS and the downstream region (FIG 1). Importantly, this gradient of mutations occurs differentially according to every gene category, with higher expressed genes showing a stronger depletion (FIG). Based on our findings of the role of local hypomethylation in mutation rates, we classified genes according to their gene-body methylation profiles into 5 clusters. The first two groups, c1 and c2 contained lowly expressed genes with a shorter (or absent) unmethylated section around the TSS, and consistently we also observed no mutation rate depletion in TSS and adjacent 5' gene regions. In contrast, the highly expressed gene clusters c3 and c4, with wider unmethylated 5' end regions showed an enrichment in active chromatin marks and stronger CpGi. For both c3 and c4 genes, the mutation rate reduction was more pronounced. A fifth group, c5, was composed by shorter genes that showed, interestingly, relative reduced methylation levels along the gene body (FIG). These genes were enriched in H3K27me3, a polycomb mark, which has also been reported to interact with DNA methylation through the H3K27me3 mark being mutually exclusive with the DNA-methyltransferase recruiting, active transcription mark H3K36me3[42]. The majority of Homeobox genes, a class of developmental associated genes were classified as c5 (FIG);

interestingly these genes are also reported to participate in cancer progression through the hypermethylation of its gene body[6].

An important practical use of the sub-gene mutation rate gradient prediction is in methods that test selection. Overall, when predicting the mutation burden of neutral genes from exonic data, a model that included the methylation aware clusters had higher accuracy than the base model. The increase in accuracy is modest, probably due to the fact that the histone mark information present in the base model (covariates used in dNdScv) can to some extent predict our methylation gene clusters. For instance, highly expressed genes share a both specific DNA methylation profile, and also a specific histone mark profile, where the latter may serve as a proxy to the former. However predicted mutation rates suggest that the mutation rate can be estimated with more detail if using the clusters. We propose that DNA methylation profiles should be incorporated into methods for detection of somatic selection. Particularly the methods that rely on the accumulation of positively selected hotspots in certain gene regions would benefit from more careful modeling of mutation rates on a sub-gene level, due to different DNA methylation and potentially also other factors.

References

1. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: in the right place at the right time. *Science* **361**, 1336–1340 (2018).
2. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
3. Greenberg, M. V. C. & Bourchis, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
4. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
5. Burger, L., Gaidatzis, D., Schübeler, D. & Stadler, M. B. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res.* **41**, e155 (2013).
6. Su, J. *et al.* Homeobox oncogene activation by pan-cancer DNA hypermethylation. *Genome Biol.* **19**, 108 (2018).
7. Zhou, W. *et al.* DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
8. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
9. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
10. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
11. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).

12. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
13. Poulos, R. C., Olivier, J. & Wong, J. W. H. The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Res.* **45**, 7786–7795 (2017).
14. Tomkova, M., McClellan, M., Kriaucionis, S. & Schuster-Böckler, B. DNA Replication and associated repair pathways are involved in the mutagenesis of methylated cytosine. *DNA Repair* **62**, 1–7 (2018).
15. Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat. Rev. Genet.* **22**, 672–686 (2021).
16. Tomkova, M., McClellan, M., Kriaucionis, S. & Schuster-Boeckler, B. 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA. *eLife* https://elifesciences.org/articles/17082 (2016) doi:10.7554/eLife.17082.
17. Supek, F., Lehner, B., Hajkova, P. & Warnecke, T. Hydroxymethylated cytosines are associated with elevated C to G transversion rates. *PLoS Genet.* **10**, e1004585 (2014).
18. Vavouri, T. & Lehner, B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome Biol.* **13**, 1–12 (2012).
19. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
20. Pleasance, E. *et al.* Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
21. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
22. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
23. Cannistraro, V. J., Pondugula, S., Song, Q. & Taylor, J.-S. Rapid Deamination of Cyclobutane Pyrimidine Dimer Photoproducts at TCG Sites in a Translationally and Rotationally Positioned Nucleosome in Vivo. *J. Biol. Chem.* **290**, 26597–26609 (2015).
24. Álvarez-Prado, Á. F. *et al.* A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *J. Exp. Med.* **215**, 761–771 (2018).
25. Haradhvala, N. J. *et al.* Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
26. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci.* **103**, 1412–1417 (2006).
27. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
28. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
29. Mattei, A. L., Bailly, N. & Meissner, A. DNA methylation: a historical perspective. *Trends Genet.* **38**, 676–707 (2022).
30. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
31. Rochette, P. J. *et al.* Influence of cytosine methylation on ultraviolet-induced cyclobutane pyrimidine dimer formation in genomic DNA. *Mutat. Res.* **665**, 7–13 (2009).
32. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
33. 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).
34. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat. Commun.* **9**, 2626 (2018).
35. 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).

36. Seplyarskiy, V. B. *et al.* APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* **26**, 174–182 (2016).
37. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun.* **10**, 2969 (2019).
38. Bhutani, N., Burns, D. M. & Blau, H. M. DNA Demethylation Dynamics. *Cell* **146**, 866–872 (2011).
39. Williams, K., Christensen, J. & Helin, K. DNA methylation: TET proteins-guardians of CpG islands? *EMBO Rep.* **13**, 28–35 (2011).
40. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinforma. Oxf. Engl.* **35**, 4788–4790 (2019).
41. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
42. Manzo, M. *et al.* Isoform-specific localization of DNMT3A regulates DNA methylation fidelity at bivalent CpG islands. *EMBO J.* **36**, 3421–3434 (2017).
43. Holland, P. W., Booth, H. A. F. & Bruford, E. A. Classification and nomenclature of all human homeobox genes. *BMC Biol.* **5**, 47 (2007).
44. Tung, K.-F., Pan, C.-Y., Chen, C.-H. & Lin, W. Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset. *Sci. Rep.* **10**, 16245 (2020).
45. Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534-547.e23 (2017).
46. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
47. Li, S., Crawford, F. W. & Gerstein, M. B. Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. *Nat. Commun.* **11**, 3575 (2020).
48. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16**, 22 (2015).
49. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
50. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinforma. Oxf. Engl.* **27**, 1571–1572 (2011).

Figures

Figure 1

Mutation gradients in genes and role of DNA methylation in mutation rates: (a) Diagram of the analysis of mutation rate gradients process. The genes are divided in 250bp long bins for which the mutation rate is calculated. The mutation rates at each bin is measured with a negative binomial regression and the output is factorized using a PCA. (b) PCA coordinates of the instances included in the regression, here 512 points representing each combination of expression bin, signature and tissue of origin. (c) Profile weights of Dimension 1 and 2 along the gene body. (d) Methylation profiles, measured as the median methylation level in each bin, for both UMRs and LMRs. Shadow area represents the 95% confidence interval of the median value across all regions. (e) Coefficients representing the relative mutation rate change for the UMR or LMR regions versus flanks. Each regression includes all mutations for a given tissue. (f) Same as in e but for the assigned mutational signatures. (g) Coefficients measuring the relative mutation rate change in tissue specific UMRs and LMRs versus flanks.

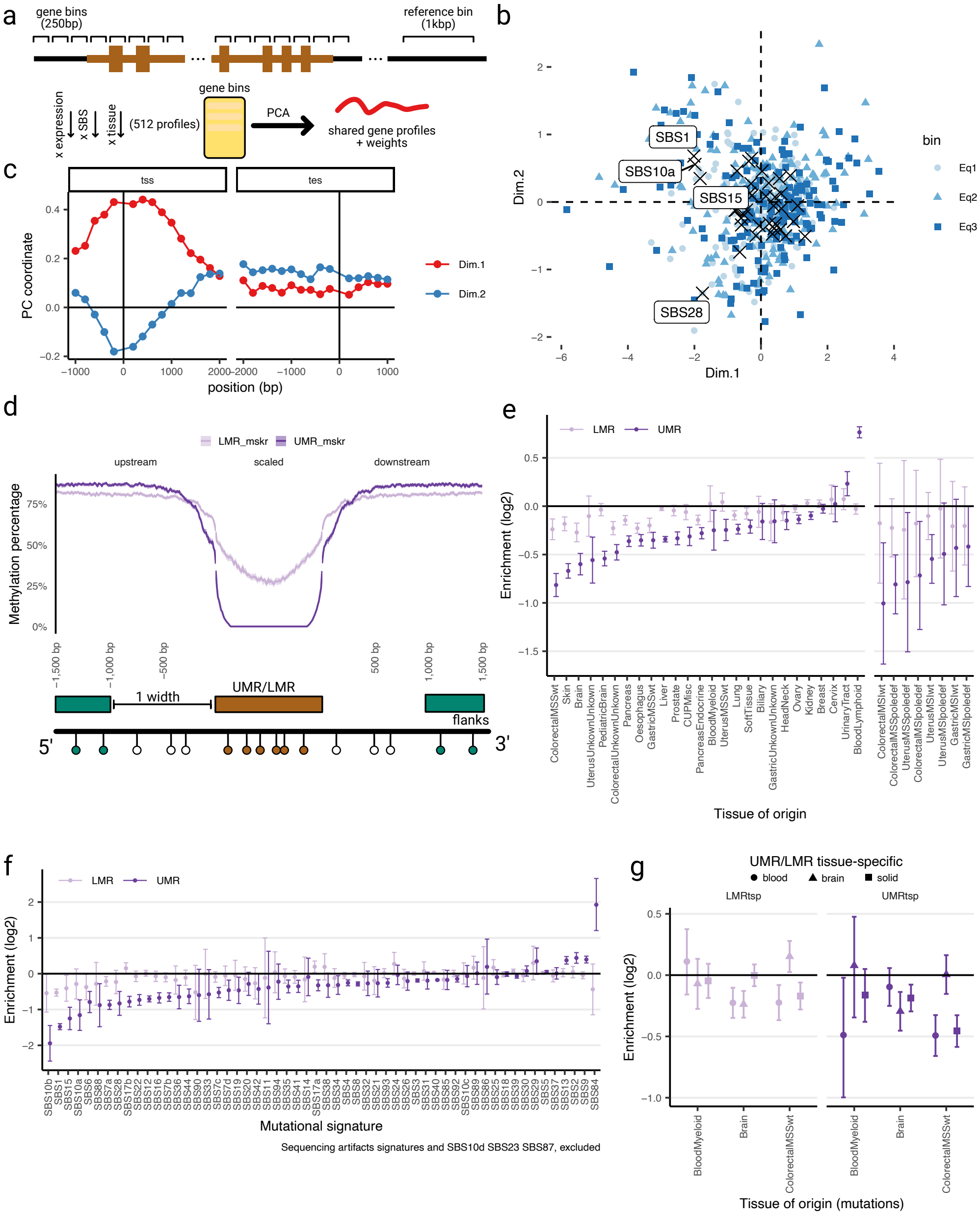

a
gene bins (250bp)
reference bin (1kbp)
x expression
x SBS
x tissue
(512 profiles)
gene bins
PCA
shared gene profiles + weights
b
SBS1
SBS10a
SBS15
SBS28
Dim.1
Dim.2
bin
Eq1
Eq2
Eq3
c
tss
tes
PC coordinate
position (bp)
Dim.1
Dim.2
d
LMR_mskr
UMR_mskr
upstream
scaled
downstream
Methylation percentage
UMR/LMR
1 width
flanks
5'
3'
e
LMR
UMR
Enrichment (log2)
Tissue of origin
ColorectalMSSwt
Skin
Brain
UterusUnkownUnkown
PediatricBrain
ColorectalUnkownUnkown
Pancreas
Oesophagus
GastricMSSwt
Liver
Prostate
CUPMisc
PancreasEndocrine
BloodMyeloid
UterusMSSwt
Lung
SoftTissue
Biliary
GastricUnkownUnkown
HeadNeck
Ovary
Kidney
Breast
Cervix
UrinaryTract
BloodLymphoid
ColorectalMSIwt
ColorectalMSSpoledef
UterusMSSpoledef
ColorectalMSIpoledef
UterusMSIwt
UterusMSIpoledef
GastricMSIwt
GastricMSIpoledef
f
LMR
UMR
Enrichment (log2)
Mutational signature
Sequencing artifacts signatures and SBS10d SBS23 SBS87, excluded
g
UMR/LMR tissue-specific
● blood ▲ brain ■ solid
LMRtsp
UMRtsp
Enrichment (log2)
BloodMyeloid
Brain
ColorectalMSSwt
Tissue of origin (mutations)

Figure 2

Functional elements associated to UMRs and LMRs: (a) Diagram of the set of the relevant functional elements represented in this figure. (b) Odds ratio enrichment of the overlap of a given functional element either with the UMR or the LMR. (c) Upset plot showing all possible intersections of UMRs with the functional elements depicted in a. In this panel, the 5’ end of the gene body and the promoter is mixed in a single group. (d) Mutation rate enrichment for UMRs that do not present an overlap with functional promoters or loop anchors. (lapless -> no LAP overlap; proless -> no promoter overlap; lapproless -> either a promoter or a LAP overlap). (e) Same as in e but for the stratified mutational signatures.

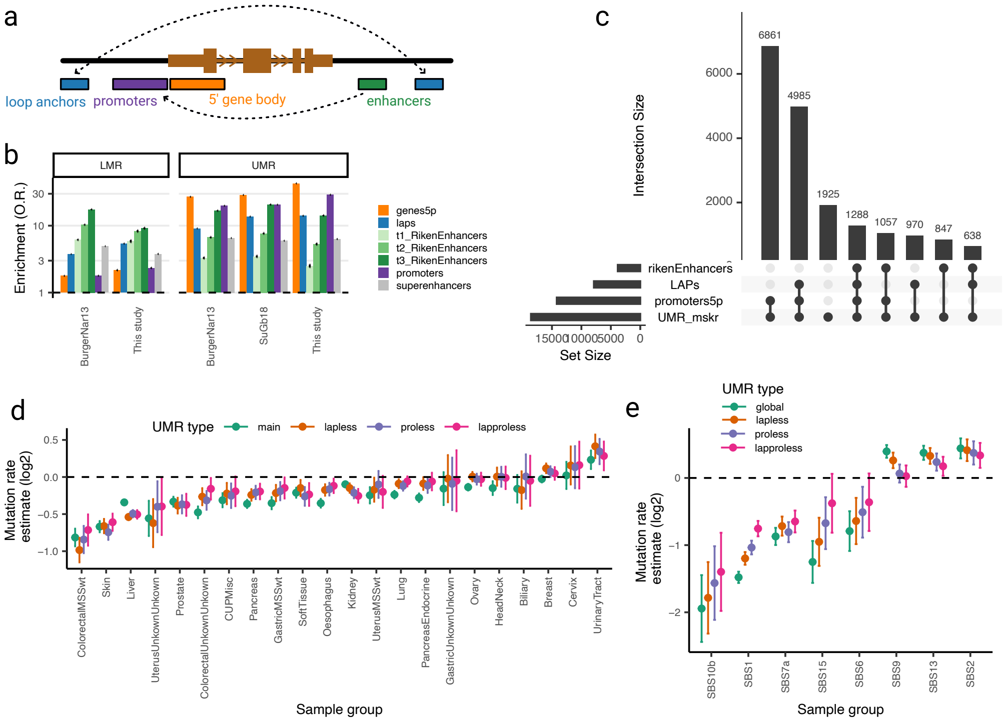

a
loop anchors
promoters
5' gene body
enhancers
b
LMR
UMR
Enrichment (O.R.)
30
10
3
1
BurgerNar13
This study
BurgerNar13
SuGb18
This study
genes5p
laps
t1_RikenEnhancers
t2_RikenEnhancers
t3_RikenEnhancers
promoters
superenhancers
c
Intersection Size
6000
4000
2000
6861
4985
1925
1288
1057
970
847
638
rikenEnhancers
LAPs
promoters5p
UMR_mskr
15000
10000
5000
0
Set Size
d
UMR type
main
lapless
proless
lapproless
Mutation rate estimate (log2)
0.5
0.0
−0.5
−1.0
ColorectalMSSwt
Skin
Liver
UterusUnkownUnkown
Prostate
ColorectalUnkownUnkown
CUPMisc
Pancreas
GastricMSSwt
SoftTissue
Oesophagus
Kidney
UterusMSSwt
Lung
PancreasEndocrine
GastricUnkownUnkown
Ovary
HeadNeck
Biliary
Breast
Cervix
UrinaryTract
Sample group
e
UMR type
global
lapless
proless
lapproless
Mutation rate estimate (log2)
0
−1
−2
SBS10b
SBS1
SBS7a
SBS15
SBS6
SBS9
SBS13
SBS2
Sample group

Figure 3

Epigenetic characterization of UMRs: (a) Histone accumulation profiles along UMRs clustered in two distinct groups, histones marks used are H3K27me, H3K36me3, H3K4me3 (depicted in the figure) and H3K27ac, H3K9me3 and H3K4me1 (depicted in Supp. Fig. 3). (b) Methylation median as in (Fig. 1D) for the two UMR methylation clusters. (c) Mutation rate estimates for SBS1, SBS15 and SBS7a for the appropriate tissues in both epigenetic UMR classes.

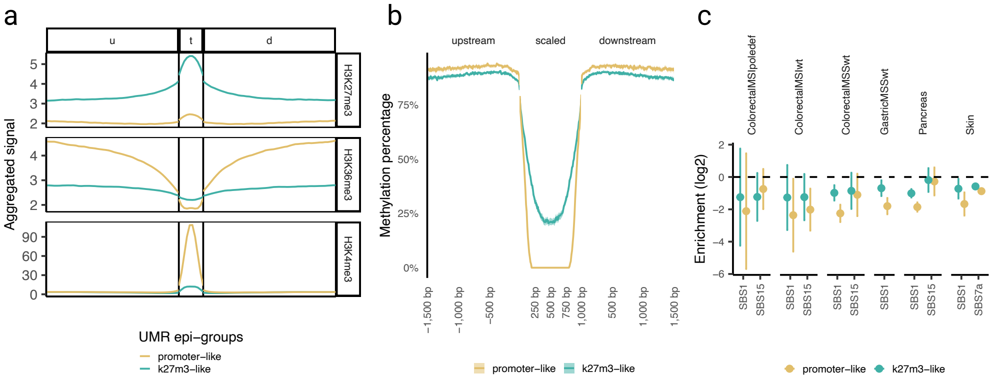

a
b
c
u
t
d
H3K27me3
H3K36me3
H3K4me3
Aggregated signal
UMR epi−groups
promoter−like
k27m3−like
upstream
scaled
downstream
Methylation percentage
75%
50%
25%
0%
−1,500 bp
−1,000 bp
−500 bp
250 bp
500 bp
750 bp
1,000 bp
500 bp
1,000 bp
1,500 bp
ColorectalMSIpoledef
ColorectalMSIwt
ColorectalMSSwt
GastricMSSwt
Pancreas
Skin
Enrichment (log2)
SBS1
SBS15
SBS7a

Figure 4

Clustering of genes according to their methylation profile: (a) PCA coordinates of each gene from the factorization of methylation profiles. (b) PCA weights for the three first components used in the clustering of the methylation profiles. (c) PCA coordinate distribution of each gene cluster for the first three principal components. (d) Median methylation level for all genes in a given cluster. Area represent the 95% confidence interval of the median across all genes in each group. (e) Overlap enrichment measured with a chi.sq test. Significant values are shown as numbers. Colors represent the logarithm in base 10 of the O/E score. Numeric values represent the raw O/E value.

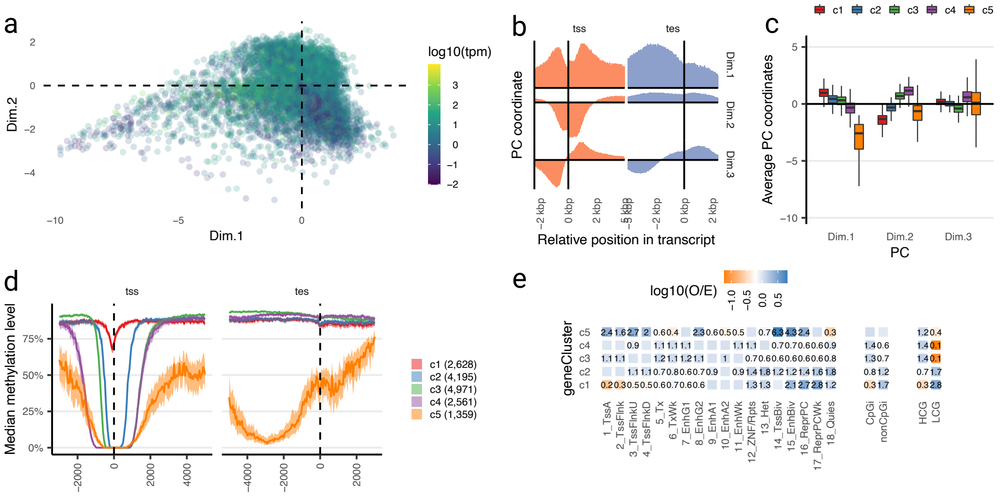
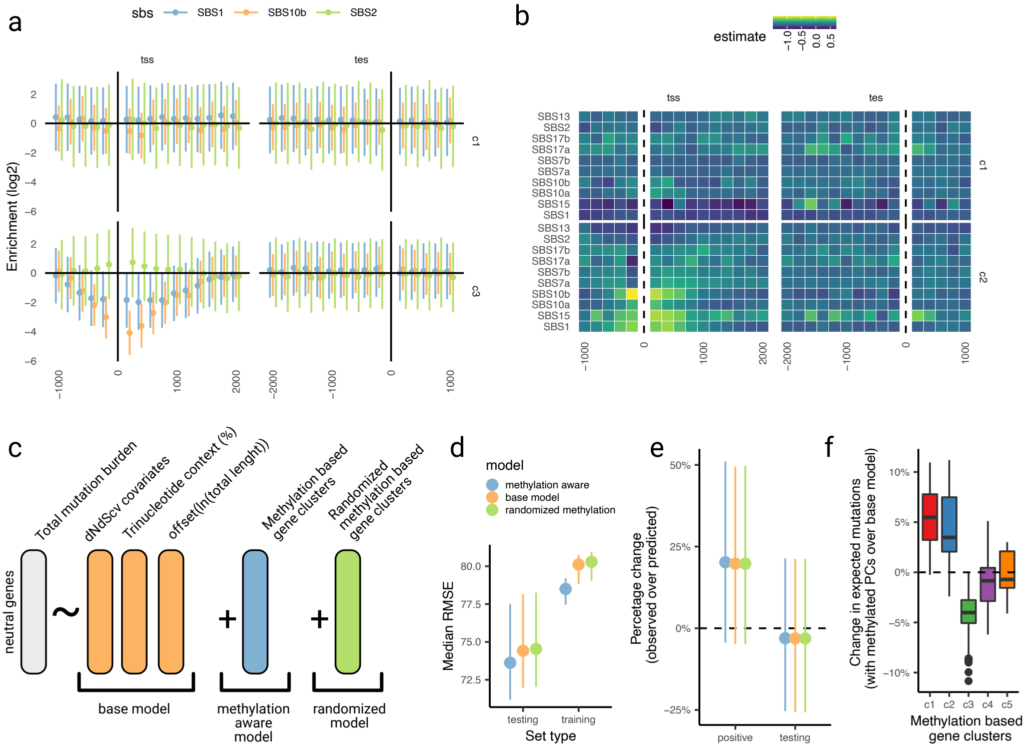

a
Dim.1
Dim.2
−10
−5
0
2
0
−2
−4
log10(tpm)
3
2
1
0
−1
−2
b
tss
tes
Dim.1
Dim.2
Dim.3
PC coordinate
−2 kbp
0 kbp
2 kbp
5 kbp
−5 kbp
Relative position in transcript
c
c1
c2
c3
c4
c5
Average PC coordinates
5
0
−5
−10
Dim.1
Dim.2
Dim.3
PC
d
tss
tes
Median methylation level
75%
50%
25%
0%
−2000
0
2000
4000
−4000
c1 (2,628)
c2 (4,195)
c3 (4,971)
c4 (2,561)
c5 (1,359)
e
log10(O/E)
−1.0
−0.5
0.0
0.5
geneCluster
c5 2.4 1.6 2.7 2 0.6 0.4 2.3 0.6 0.5 0.5 0.7 6.3 4.3 2.4 0.3 1.2 0.4
c4 0.9 1.1 1.1 1.1 1.1 1.1 0.7 0.7 0.6 0.6 0.9 1.4 0.6 1.4 0.1
c3 1.1 1.1 1.2 1.1 1.2 1.1 1 0.7 0.6 0.6 0.6 0.6 0.6 0.8 1.3 0.7 1.4 0.1
c2 1.1 1.1 0.7 0.8 0.6 0.7 0.9 0.9 1.4 1.8 1.2 1.2 1.4 1.6 1.8 0.8 1.2 0.7 1.7
c1 0.2 0.3 0.5 0.5 0.6 0.7 0.6 0.6 1.3 1.3 2.1 2.7 2.8 1.2 0.3 1.7 0.3 2.8
1_TssA
2_TssFlnk
3_TssFlnkU
4_TssFlnkD
5_Tx
6_TxWk
7_EnhG1
8_EnhG2
9_EnhA1
10_EnhA2
11_EnhWk
12_ZNF/Rpts
13_Het
14_TssBiv
15_EnhBiv
16_ReprPC
17_ReprPCWk
18_Quies
CpGi
nonCpGi
HCG
LCG

Figure 5

Mutation enrichment in gene bodies of methylation aware gene classes: (a) Mutation enrichment in 250bp long bins (similar from Fig.1A) for every gene in the c1 and c3 clusters defined in Fig. 4. (b) Mutation rate enrichment for a set of relevant signatures for cluster c1 and c2 as defined in Fig. 4.. (c) Diagram depicting a model to predict the mutation rate of genes according to dNdScv covariates, the context composition of the gene and the length as a offset. To this base model, the methylation-aware gene classes are added together with a randomized version of the gene clusters. (d) Root mean square error for the prediction of mutation rates by each model. (e) Percentatge change of predicted mutations in the positive set (cancer genes with positive selection) and the testing set (genes that are used to evaluate the performance of each CV round). (f) Changes in the predicted mutations of genes for each gene cluster as defined in Fig. 4.

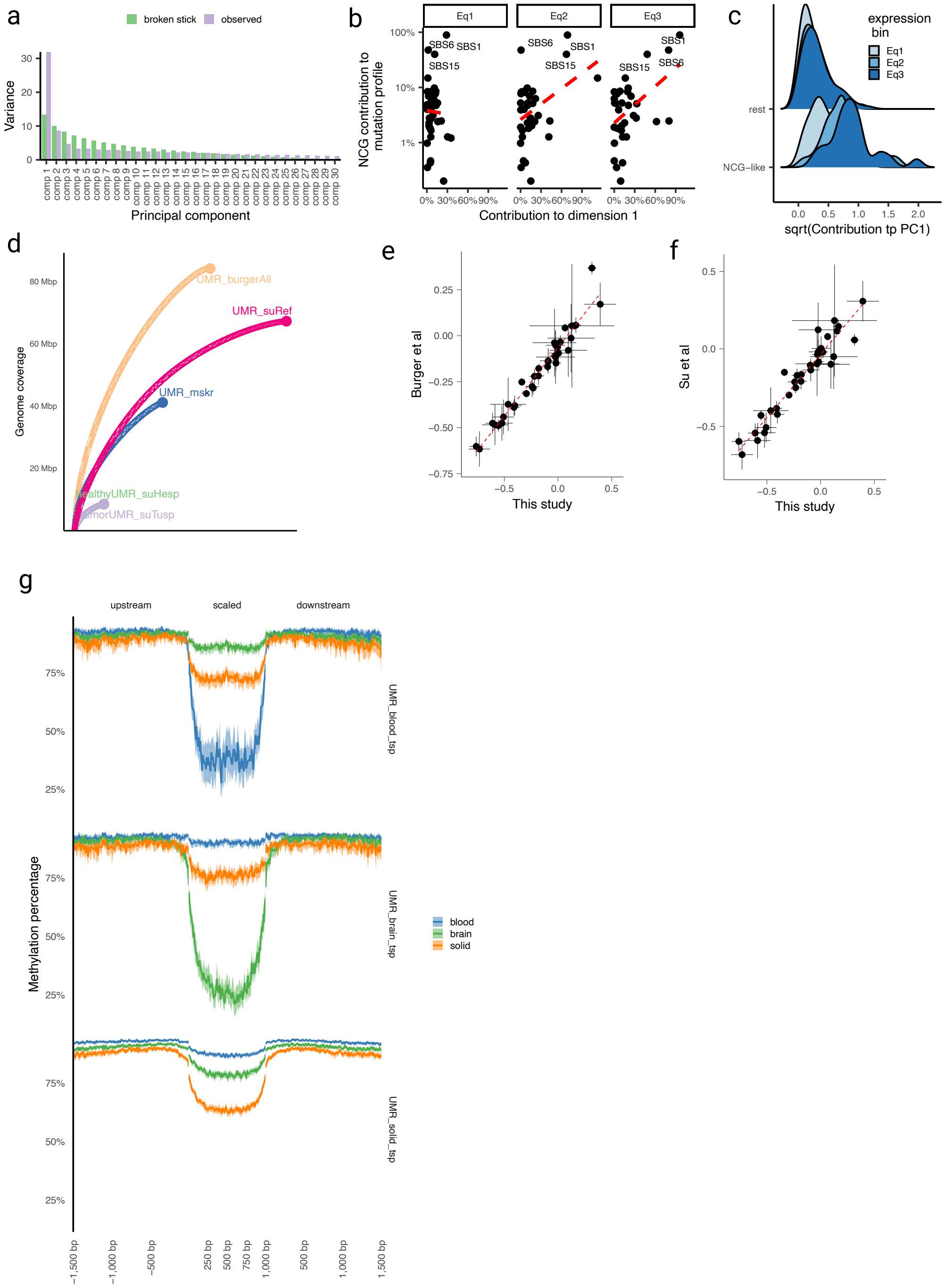


a
sbs
SBS1
SBS10b
SBS2
tss
tes
Enrichment (log2)
2
0
−2
−4
−6
c1
c3
−1000
0
1000
2000
−2000
b
estimate
−1.0
−0.5
0.0
0.5
SBS13
SBS2
SBS17b
SBS17a
SBS7b
SBS7a
SBS10b
SBS10a
SBS15
SBS1
c2
c
Total mutation burden
dNdScv covariates
Trinucleotide context (%)
offset(ln(total lenght))
Methylation based gene clusters
Randomized methylation based gene clusters
neutral genes
~
+
base model
methylation aware model
randomized model
d
model
methylation aware
base model
randomized methylation
Median RMSE
80.0
77.5
75.0
72.5
testing
training
Set type
e
Percetage change (observed over predicted)
50%
25%
0%
−25%
positive
testing
f
Change in expected mutations (with methylated PCs over base model)
10%
5%
0%
−5%
−10%
c1
c2
c3
c4
c5
Methylation based gene clusters

Supplementary Figures

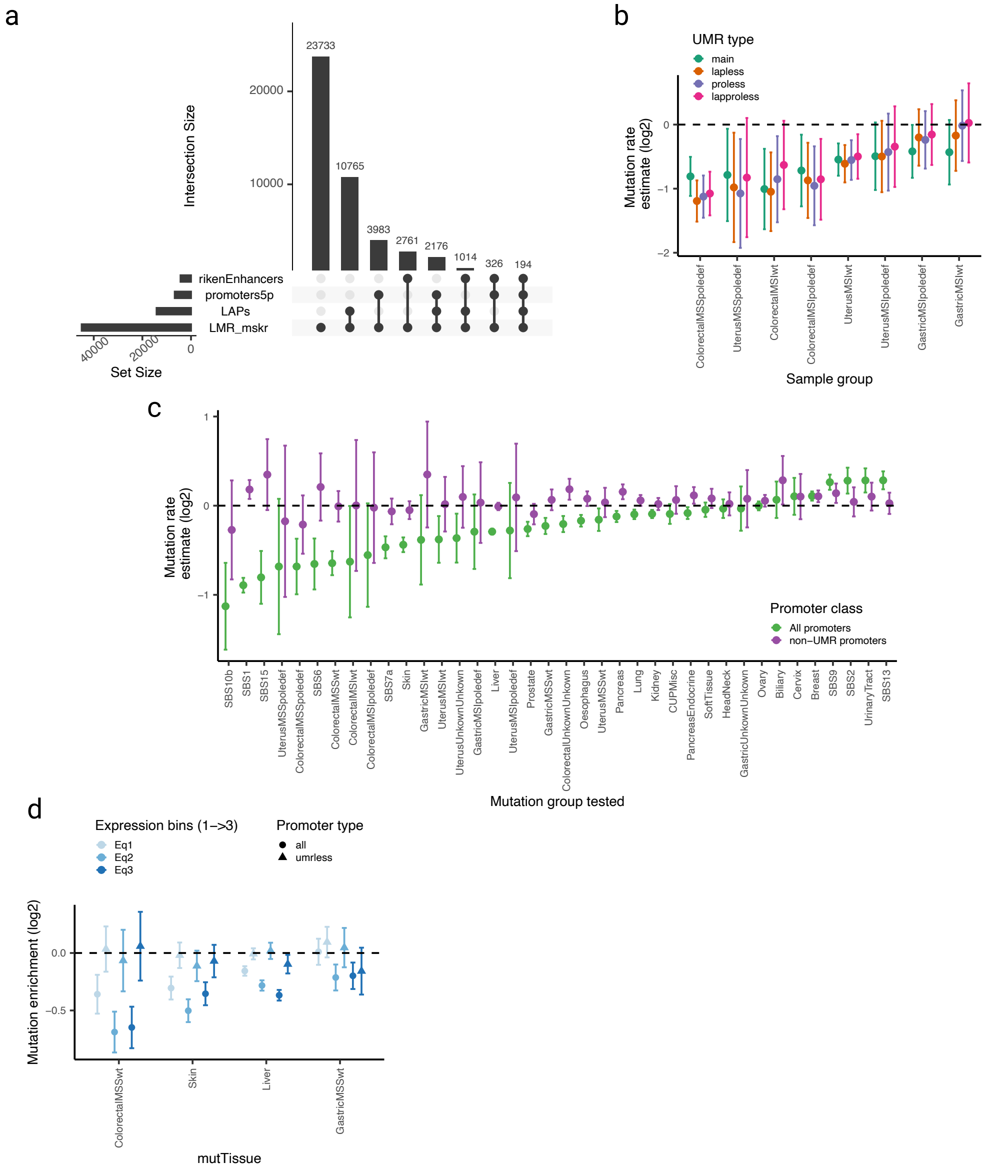
Supp. Figure 1

(Extended) Mutation gradients in genes and role of DNA methylation in mutation rates: (a) Scree plot from the gene gradient mutation rate factorization. (b) Correlation of the percentage of CG trinucleotides in each signature compared to the total contribution to the first principal component. (c) Same as in (b) but instances are stratified by gene expression and signatures are classified in CG-like or rest according to the CG percentage in their profiles. (d) Genomic coverage of the selected UMRs. (e-f) Correlation of the mutation rate estimations in different UMR sets. (g) Methylation levels in tissue specific UMRs.


a
broken stick
observed
Variance
0
10
20
30
comp 1
comp 2
comp 3
comp 4
comp 5
comp 6
comp 7
comp 8
comp 9
comp 10
comp 11
comp 12
comp 13
comp 14
comp 15
comp 16
comp 17
comp 18
comp 19
comp 20
comp 21
comp 22
comp 23
comp 24
comp 25
comp 26
comp 27
comp 28
comp 29
comp 30
Principal component
b
Eq1
Eq2
Eq3
NCG contribution to mutation profile
100%
10%
1%
SBS6
SBS1
SBS15
0% 30%60%90%
Contribution to dimension 1
c
expression bin
Eq1
Eq2
Eq3
rest
NCG–like
0.0
0.5
1.0
1.5
2.0
sqrt(Contribution tp PC1)
d
Genome coverage
80 Mbp
60 Mbp
40 Mbp
20 Mbp
UMR_burgerAll
UMR_suRef
UMR_mskr
healthyUMR_suHesp
tumorUMR_suTusp
e
Burger et al
0.25
0.00
−0.25
−0.50
−0.75
−0.5
0.0
0.5
This study
f
Su et al
0.5
0.0
−0.5
−0.5
0.0
0.5
This study
g
upstream
scaled
downstream
Methylation percentage
75%
50%
25%
UMR_blood_tsp
UMR_brain_tsp
UMR_solid_tsp
blood
brain
solid
−1,500 bp
−1,000 bp
−500 bp
250 bp
500 bp
750 bp
1,000 bp
500 bp
1,000 bp
1,500 bp

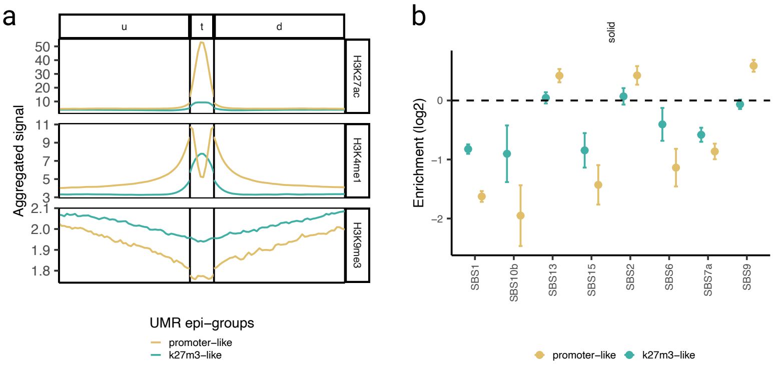
Supp. Figure 2

(Extended) Functional elements associated to UMRs and LMRs: (a) Upset plot representing the overlap with functional elements in LMRs. (b) Mutation estimates in functional element free UMRs for DNA repair deficient tissues. (c) Mutation rate estimates in promoters that significantly overlap with a UMR (> 200bp) and all promoters. (d) Same as in c but for selected tissues and stratifying the promoters according to the expression bins.


a
Intersection Size
20000
10000
23733
10765
3983
2761
2176
1014
326
194
rikenEnhancers
promoters5p
LAPs
LMR_mskr
40000
20000
0
Set Size
b
UMR type
main
lapless
proless
lapproless
Mutation rate estimate (log2)
0
−1
−2
ColorectalMSSpoledef
UterusMSSpoledef
ColorectalMSIwt
ColorectalMSIpoledef
UterusMSIwt
UterusMSIpoledef
GastricMSIpoledef
GastricMSIwt
Sample group
c
Mutation rate estimate (log2)
1
0
−1
Promoter class
All promoters
non−UMR promoters
SBS10b
SBS1
SBS15
UterusMSSpoledef
ColorectalMSSpoledef
SBS6
ColorectalMSSwt
ColorectalMSIwt
ColorectalMSIpoledef
SBS7a
Skin
GastricMSIwt
UterusMSIwt
UterusUnkownUnkown
GastricMSIpoledef
Liver
UterusMSIpoledef
Prostate
GastricMSSwt
ColorectalUnkownUnkown
Oesophagus
UterusMSSwt
Pancreas
Lung
Kidney
CUPMisc
PancreasEndocrine
SoftTissue
HeadNeck
GastricUnkownUnkown
Ovary
Biliary
Cervix
Breast
SBS9
SBS2
UrinaryTract
SBS13
Mutation group tested
d
Expression bins (1−>3)
Eq1
Eq2
Eq3
Promoter type
all
umrless
Mutation enrichment (log2)
0.0
−0.5
ColorectalMSSwt
Skin
Liver
GastricMSSwt
mutTissue

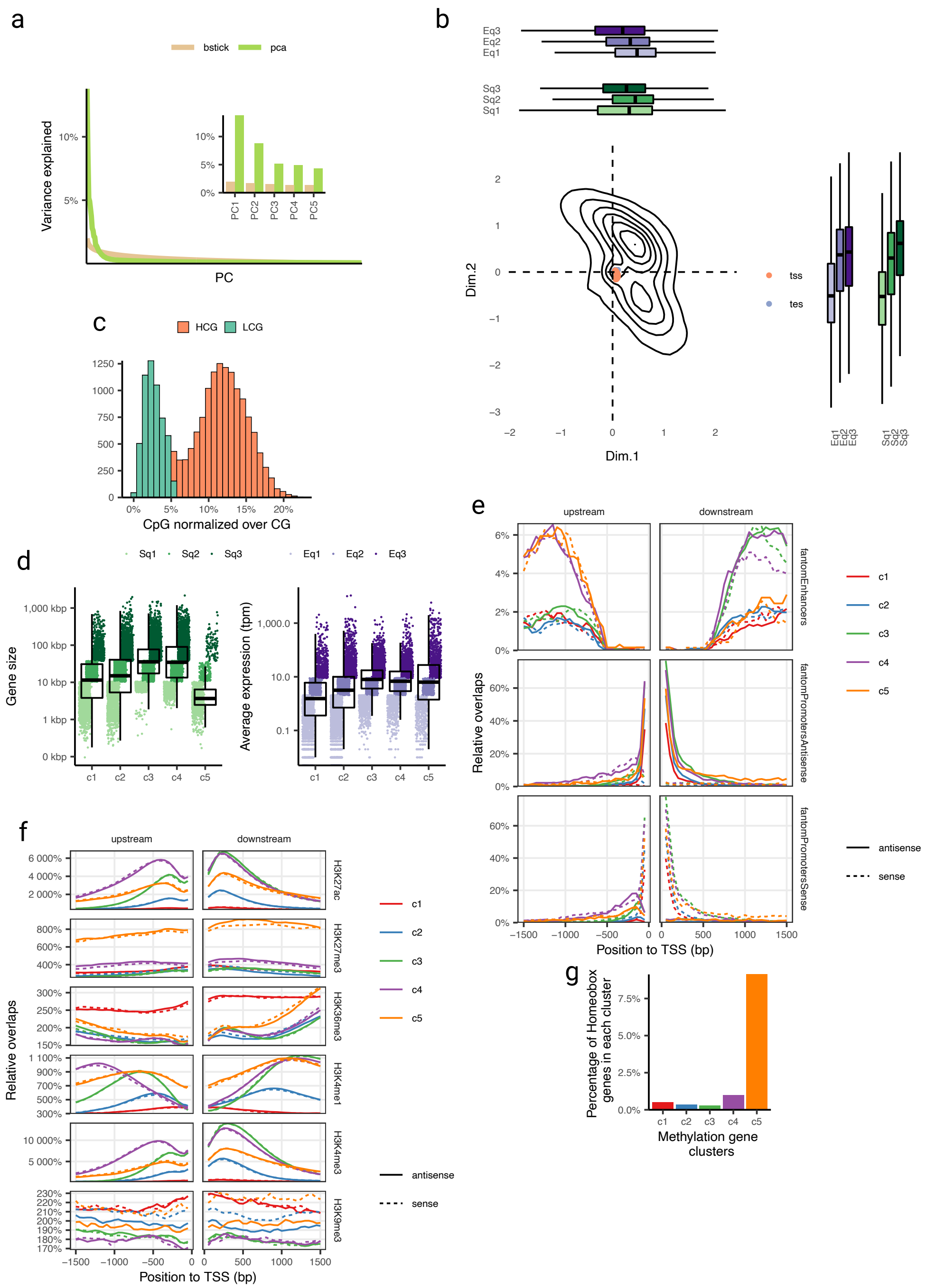
Supp. Figure 3

(Extends) Epigenetic characterization of UMRs: (a) Histone profile of H3K27ac, H3K4me1 and H3K9me3 around epigenetic defined clustering of UMRs. (b) Mutation rate estimation in each UMR class according to the mutational signature.


a
u
t
d
50
40
30
20
10
H3K27ac
Aggregated signal
11
9
7
5
3
H3K4me1
2.1
2.0
1.9
1.8
H3K9me3
UMR epi−groups
promoter−like
k27m3−like
b
solid
Enrichment (log2)
0
−1
−2
SBS1
SBS10b
SBS13
SBS15
SBS2
SBS6
SBS7a
SBS9
promoter−like
k27m3−like

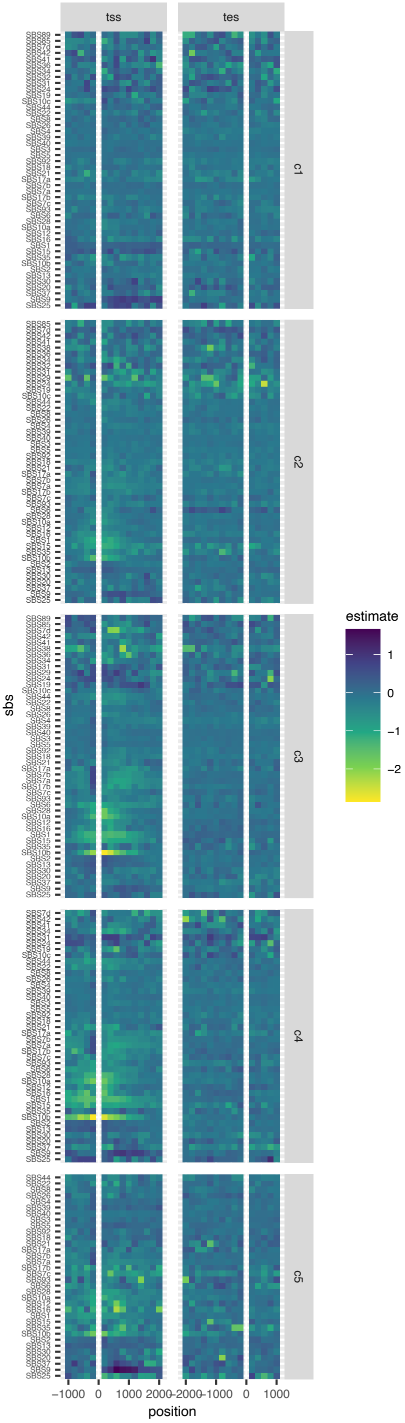
Supp. Figure 4

(Extends) Clustering of genes according to their methylation profile: (a) Scree plot of the methylation profile PCA used to cluster genes. (b) PCA coordinates of each gene (represented as a 2D density plot) with the expression and size distribution for each principal component represented in boxplots. (c) Definition of the HCG genes according to their normalized CG values. A mixture modeling is used to define the threshold. (d) Expression and Size bins of each gene methylation class. (e) Enrichment of FANTOM nascent transcription associated to promoters (middle and bottom) and enhancers. Promoters and genes are divided in sense and antisense. (g) Proportion of Homeobox genes, as defined in ref[43] , for each methylation aware cluster.


a
bstick
pca
Variance explained
10%
5%
PC
PC1
PC2
PC3
PC4
PC5
b
Eq3
Eq2
Eq1
Sq3
Sq2
Sq1
Dim.2
Dim.1
tss
tes
c
HCG
LCG
CpG normalized over CG
1250
1000
750
500
250
0
0%
5%
10%
15%
20%
d
Sq1
Sq2
Sq3
Eq1
Eq2
Eq3
Gene size
1,000 kbp
100 kbp
10 kbp
1 kbp
0 kbp
Average expression (tpm)
1,000.0
10.0
0.1
c1
c2
c3
c4
c5
e
upstream
downstream
fantomEnhancers
fantomPromotersAntisense
fantomPromotersSense
Relative overlaps
Position to TSS (bp)
antisense
sense
f
H3K27ac
H3K27me3
H3K36me3
H3K4me1
H3K4me3
H3K9me3
Relative overlaps
Position to TSS (bp)
g
Percentage of Homeobox genes in each cluster
7.5%
5.0%
2.5%
0.0%
Methylation gene clusters

Supp. Figure 5

(Extends) Mutation enrichment in different gene bodies: Mutation enrichment of each mutation signature (in rows) for each gene bin (in columns) of 250bp. Mutation rate estimates are represented as coefficients in natural logarithm.


tss
tes
c1
c2
c3
c4
c5
sbs
estimate
1
0
−1
−2
−1000
0
1000
2000
-2000
-1000
0
1000
position

Supplementary Tables

Supplementary Table 1: List of methylation datasets used to define the UMRs and LMRs in this study.

code	group	tissue	source	inclusion	FDRper	Coverage
E058	solid	skin	ROADMAP	excluded	6.10%	108,290,849
E054	brain	brain_ganglion	ROADMAP	included	1.20%	78,967,948
E053	brain	brain_cortex	ROADMAP	included	1.60%	85,146,106
E071	brain	brain_hippocampus	ROADMAP	included	0.70%	79,545,018
E070	brain	brain_matrix	ROADMAP	included	1.90%	98,948,007
E100	solid	muscle	ROADMAP	included	3.00%	114,119,567
E095	solid	heart	ROADMAP	included	1.00%	71,085,329
E109	solid	intestine	ROADMAP	included	1.20%	72,889,249
E079	solid	esophagusgut	ROADMAP	included	1.20%	78,609,887
E094	solid	stomach	ROADMAP	included	1.90%	103,815,766
E066	solid	liver	ROADMAP	included	0.80%	65,726,414
E096	solid	lung	ROADMAP	included	1.00%	73,087,658
E113	blood	spleen	ROADMAP	included	1.20%	72,300,133
E085	solid	intestine	ROADMAP	included	0.70%	69,148,189
E084	solid	intestine	ROADMAP	included	0.80%	80,254,187
E106	solid	colon	ROADMAP	included	0.90%	77,950,142
E112	blood	thymus	ROADMAP	included	0.30%	61,539,099
E050	blood	hsc	ROADMAP	included	0.60%	72,292,167
E008	stemcells	esc	ROADMAP	included	0.20%	27,273,886
E016	stemcells	esc	ROADMAP	included	0.10%	29,248,304
E024	stemcells	esc	ROADMAP	manually_excluded	0.30%	54,758,498
E021	stemcells	ips	ROADMAP	included	0.20%	47,685,111
E022	stemcells	ips	ROADMAP	included	0.20%	51,095,932
E007	stemcells	escd	ROADMAP	included	0.10%	39,496,379
ENCFF491ZQM	blood	natural killer cell	ENCODE	excluded	0.90%	50,057,397
ENCFF867JRG	blood	K562	ENCODE	manually_excluded	1.00%	1,554,341,638
ENCFF279HCL	blood	GM12878	ENCODE	excluded	115.90%	843,330,637
ENCFF355UVU	blood	T-cell	ENCODE	included	1.10%	51,246,932
ENCFF774VLD	blood	B cell	ENCODE	included	0.90%	64,729,153
ENCFF451WIY	blood	CD14-positive monocyte	ENCODE	included	1.50%	79,341,328
ENCFF489CEV	solid	stomach	ENCODE	included	2.00%	84,591,859
ENCFF577TCU	solid	gastroesophageal sphincter	ENCODE	excluded	4.80%	75,896,721
ENCFF844EFX	solid	stomach	ENCODE	included	2.70%	75,548,982
ENCFF923CZC	solid	large intestine	ENCODE	included	0.90%	75,162,056

ENCFF521DHD	solid	small intestine	ENCODE	excluded	0.90%	66,285,068
ENCFF424XKF	solid	transverse colon	ENCODE	excluded	3.40%	54,917,634
ENCFF811QOG	solid	stomach	ENCODE	included	2.90%	81,640,221
ENCFF241AQC	solid	small intestine	ENCODE	included	0.90%	52,559,972
ENCFF266NGW	solid	small intestine	ENCODE	included	1.20%	63,595,672
ENCFF534RNT	solid	stomach	ENCODE	included	1.20%	70,856,980
ENCFF455TQO	solid	sigmoid colon	ENCODE	included	2.00%	86,007,640
ENCFF435SPL	solid	stomach	ENCODE	included	2.10%	92,539,941
ENCFF122LEF	solid	small intestine	ENCODE	included	2.40%	89,505,578
ENCFF497YOO	solid	stomach	ENCODE	included	1.90%	94,022,932
ENCFF157POM	solid	sigmoid colon	ENCODE	included	0.60%	49,059,946
ENCFF366UWF	solid	hepatocyte	ENCODE	manually_excluded	1.00%	70,746,855
ENCFF847OWL	solid	HepG2	ENCODE	excluded	254.50%	1,398,562,294
ENCFF390OZB	solid	HepG2	ENCODE	excluded	219.80%	1,486,267,948
ENCFF487XOB	solid	hepatocyte	ENCODE	manually_excluded	1.00%	68,617,010
ENCFF577VGR	solid	right lobe of liver	ENCODE	included	1.60%	65,336,025
ENCFF064GJQ	solid	HepG2	ENCODE	excluded	250.40%	1,401,851,624
ENCFF369YQW	solid	HepG2	ENCODE	excluded	249.50%	1,416,591,817
ENCFF005TID	solid	A549	ENCODE	excluded	169.40%	615,993,233
ENCFF842MHJ	solid	upper lobe of left lung	ENCODE	included	1.20%	63,160,587
ENCFF937OSM	solid	IMR-90	ENCODE	included	3.10%	78,849,792
ENCFF003JVR	solid	A549	ENCODE	excluded	175.40%	647,542,688
ENCFF477AUC	solid	lung	ENCODE	included	0.60%	65,516,632
ENCFF733EFJ	solid	upper lobe of left lung	ENCODE	included	1.50%	62,959,700
ENCFF039JFT	solid	lung	ENCODE	included	0.90%	62,843,505
ENCFF288YTY	solid	IMR-90	ENCODE	excluded	64.30%	85,950,357
ENCFF254DBF	solid	IMR-90	ENCODE	excluded	78.20%	176,853,985
ENCFF714SUO	solid	GM23248	ENCODE	excluded	7.30%	107,515,043
ENCFF959WCA	solid	GM23248	ENCODE	excluded	7.20%	111,843,192
ENCFF116DGM	solid	GM23248	ENCODE	excluded	7.30%	125,812,390
ENCFF219GCQ	solid	lower leg skin	ENCODE	included	1.60%	74,565,673
ENCFF752NXS	solid	GM23248	ENCODE	excluded	7.40%	127,301,933
ENCFF121VIX	solid	lower leg skin	ENCODE	included	1.70%	73,448,503
ENCFF517AOL	solid	iPS DF 19.11	ENCODE	excluded	0.90%	14,104,156
ENCFF545MIY	solid	iPS DF 6.9	ENCODE	excluded	0.10%	35,906,952
ENCFF186EKM	solid	iPS DF 19.11	ENCODE	excluded	0.10%	39,896,920
ENCFF774GXJ	solid	skeletal muscle myoblast	ENCODE	manually_excluded	4.50%	94,999,394
ENCFF588ETU	solid	muscle of leg	ENCODE	included	2.70%	78,409,570
ENCFF837SXM	solid	skeletal muscle myoblast	ENCODE	manually_excluded	4.40%	96,476,185
ENCFF645AZF	solid	muscle of trunk	ENCODE	included	2.90%	83,324,773
ENCFF672QKY	solid	smooth muscle cell	ENCODE	manually_excluded	1.10%	77,304,925
ENCFF297CJG	solid	smooth muscle cell	ENCODE	manually_excluded	1.10%	76,185,773

ENCFF588IUK	solid	smooth muscle cell	ENCODE	manually_excluded	1.10%	79,993,937
ENCFF315ZJB	solid	smooth muscle cell	ENCODE	manually_excluded	1.10%	81,827,708
ENCFF913UZU	solid	psoas muscle	ENCODE	included	2.60%	94,696,041
ENCFF121ZES	solid	psoas muscle	ENCODE	included	1.20%	65,736,441
ENCFF940XWW	brain	SK-N-SH	ENCODE	excluded	75.90%	484,288,986
ENCFF179VKR	brain	SK-N-SH	ENCODE	excluded	96.70%	582,834,973

Online Methods

Reference region sets

ChromHMM states were downloaded as a bed file from the Roadmap data portal at egg2.wustl.eduroadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel. The core_K27ac model was selected for sample E017 (IMR90) and used throughout all the analysis.

Gene models for assembly GRCh37 were downloaded from the GENCODE release website (https://www.gencodegenes.org/human/) for the version 19. For each gene, a single transcript was used, if not stated otherwise. TSS, TES and gene length were derived from this annotation if not stated otherwise. These transcripts were selected according to he TREGT gene list that uses a combination of CDS gene length and expression level to select the most appropriate isoform. The list is available in (tregt.ibms.sinica.edu.tw) and in ref[44]. Transcription levels for all genes were downloaded from GTEX website (version V8) in TPMs and averaged globally for all samples yielding an average value for each gene.

SomaticHypermutation (SHM) on target and off target regions were defined similarly as in ref[45]. In brief, on-target regions were defined as genomic regions for the immunoglobulin genes: IGH, IGL and IGK were retrieved and extended 10Kbp upstream, downstream and reduced. Mutations in those regions were filtered out when appropriate. Off-target regions were extracted from AID activity in mouse B-lymphocytes which was then translated (liftOver) to hg19.

Somatic mutations

In order to detect samples with deficient mutations in DNA methylation related genes we annotated both SNVs and indel somatic variants with annovar[46] using the ensGene database. We considered as deficient mutation any mutation in a coding sequence which was not classified as synonymous. For each selected genes, we stratified samples by their tissue of origin and by their MSI status. From each category, we selected a random set of samples to match the ones with deficient mutations. This set of random samples was used as a control in further analysis.

Mutational signature assignment

Mutation calls for SNVs were tallied and classified according to their trinucleotide context and their alternative base. COSMIC signature profiles and tissue exposures (V3.3) we downloaded directly from the cosmic website at (cancer.sanger.ac.uk/signatures/sbs/) . A mutational signature was assigned to a tissue if at least 1 sample in the cosmic dataset contained that signature. Of note, some of the samples in the cosmic signatures dataset are also included in our set, but their direct exposures were not taken. Signature 1 and 5 were assigned to all tissues. MSI and POLE deficient samples were treated independently within their tissues of origins and signatures associated with their phenotype were included, in brief, for MSI samples we included signatures 6, 15, 21, 26, 44, 14, and 20; and for POLE deficient samples we included 10 (a, b, c and d), 14 and 20. For each tissue, the matrix with the mutational profile of each sample was computed and fitted to the assigned cosmic signatures via SigLasso[47] which implements a lasso regression fitting that forces sparsity in the signature assignment. Results from the lasso fitting were then used as exposures for the rest of the analysis.

For every sample in our dataset, we used the signature exposures obtained from SigLasso fitting in order to obtain the probability of a given mutation to be caused by a given mutational signature. In brief, the exposure in a given sample was split to the 96 mutation categories according to the original mutational profile (weight of every mutation category) and afterwards each feature was normalized within every sample so that every mutation class had a given probability to be associated to any of the used mutational signatures. Thus, using this approach, we could estimate the probability of a mutation of class i to be associated with a given signature. If the signature was not present in a sample the probability was then zero.

Then, to classify the raw mutation calls, we used these probabilities to sample a single signature and assign it to a given mutation. This process allows us to classify raw mutation calls to distinct mutational signatures and allows us to pool mutations generated by the same process across different samples and different tissues.

DNA methylation data and analysis

Tissue specific data for the selected tissue groups (solid, blood and brain) were downloaded from the ENCODE main data portal (https://www.encodeproject.org/). From each of the selected groups of tissues we obtained 3 reference experiments (reference epigenomes). If available, data from primary tissues was obtained. If not available, data from cell lines and primary cell cultures was used.

We obtained a total of six histone mark signal for every experiment: (i) H3K4me3 for TSS and promoters; (ii) H3K4me1 for enhancers; (iii) H3K27ac for active promoters and enhancers; (iv) H3K9me3 for heterochromatin; (v) H3K36me3 for gene bodies of expressed genes and (vi) H3K27me3 for bivalent transcription and polycomb marked genes. The signal obtained measured fold change over control which is equivalent to the chip-seq signal value over the input in the experiment.

For the 3 samples included in each group, we averaged the signal using ucsc tools (bigWigMerge). We then combined the averaged signals with the different UMR types and run computeMatrix in scale-region mode from the deeptools toolset in order to obtain a meta profile scaled to the corresponding UMR region.

The metaprofiles of every selected histone mark for every selected UMR were clustered together using k-means for k 2 to 10. The resulting clusters were selected based on the total sum of squares within each cluster and after inspection of the resulting profiles for biological coherence. Two clusters were finally selected.

Functional element enrichment in UMRs

Enhancer data based on CAGE data was obtained from the FANTOM dataset [48], version V5 (https://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers). They were posteriorly divided into terciles using the predefined categories in the downloaded data, with t3 indicating a higher expression level (in TPMs) and t1 indicating the lowest. As in ref[6], superenhancers were downloaded from the supplementary material in ref[49]. From the available sets we used primarily the superenhancer track marked in red. The UCSC gene model, available in the bioconductor package TxDb.Hsapiens.UCSC.hg19.knownGene , was used to define promoters and the 5’ genic sections. Promoters were defined as the 2kbp upstream of the TSS with no upstream section, and the 5’ genic sections were defined as the 2kbp downstream of the TSS.

These functional elements were compared against different sets of UMRs for three different sources of methylation data: from ref[5,6] and the set gathered in this study. The enrichment measurement is based on a fisher exact test of the overlapping bp between 2 types of regions. Thus, if a feature is less specifically overlapped against another, the odds' ratio will decrease even if many of the sparser one are covered.

Methylation data sources

To maximize the genomic coverage of the DNA methylation data, we gathered whole genome bisulphite sequencing (WGBS) from publicly available datasets, in brief, the Roadmap epigenome project (see https://egg2.wustl.edu/roadmap/web_portal/) and the ENCODE data portal (see https://www.encodeproject.org/).

Data from the Roadmap project consisted in all sets with available WGBS data. They can be accessed in Supp. Table 1. Downloaded data consisted in fractional methylation data (FractionalMethylation) which contains information about the methylation of each sufficiently covered CpG in a percentage value. We also downloaded files containing genomic coverage of each CpG.

Similarly, all WGBS available data from ENCODE was downloaded. All files were in the bedMethyl format derived from the output of Bismark[50] in the ENCODE main processing pipeline. This format also contains the methylation levels of all sufficiently covered CpG in a percentage. In addition, the same format also contains information about the coverage of each CpG dinucleotide. Accession codes from these files are available in table. The ENCODE datasets were only available in the hg38 assembly and were translated to hg19 (using liftOver) to match the rest of the analysis. LiftOver statistics can also be found in Supp. Table 1.

Methylation data processing

In order to call significant unmethylated regions (UMR) we used *MethylSeekR* from bioconducor[5] implementing the default processing workflow suggested by the authors in the vignette. In brief, SNP positions are first removed from the set (see Supp. Table 1). PMDs were detected by using the shortest chromosome with at least 150 probes as a training set. CpG islands were downloaded from UCSC table query. These datasets were then used to calculate the FDRs for the detected UMR segments. A threshold of 4 CpG positions in each segment and at least a smaller than 50% methylation value was required. If the FDR value at these conditions was lower than 5%, the samples were automatically discarded. If the total number of CpG islands considered was smaller than 25M the samples were also discarded. Non-autosomal chromosomes were removed (Supp. Table 1).

This process was run for every sample in our dataset individually. UMRs extracted from each set were then translated in a matrix format, containing a binary encoding (1 or 0) if a specific locus was included or not in that sample. This matrix was factorized using tSNE (from the Rtsne package) with 25 perplexity. The resulting grouping was inspected for biological coherence, samples that were not grouped with its tissue group were manually excluded for further analysis (see Supp. Table 1).

For each tissue group (solid, brain, and blood), individually detected UMRs were pooled into a union set which contained all UMR loci from every experiment and then reduced to avoid overlaps. If not stated otherwise, these are the sets used for all analysis when compared to mutation calls. A full union set was also generated from the union of all sets together. Each union set for every tissue was then used to compare with the other tissue groups and the UMRs which were specific to that tissue group, not present in others, were selected as tissue-specific.

UMRs from other studies were also downloaded to be used as reference sets in this analysis. UMR calls from ref[5] were downloaded from the supplementary material and were pooled for both UMR and LMR classes. These experiments included mostly cell lines from blood tissues or reprogrammed cells. Other samples included adipose tissue and fibroblasts. This dataset was originally downloaded in hg18 and then translated into hg19 with liftOver. Of note, software used to call UMRs in these datasets was the same as the one used for the downloaded WGBS data. Data from ref[6] was also downloaded from the supplementary material and pooled across different available datasets. The UMRs were divided into Canyons, cUMR (conserved UMRs) and either healthy or tumor specific UMRs. If not stated otherwise, the conserved UMR dataset was used for all the analysis in this study.

Clustering of methylation profiles in gene bodies

From the downloaded WGBS datasets the average methylation value for every available CpG dinucleotide was computed within tissue groups (brain, blood and solid). The solid average values were used for this analysis.

Gene bodies were extracted from TSS to TES , thus including 3' UTRs, coding sequences, introns and 5' UTRs. For each gene body, the analyzed regions were located around either the ends. These ends were expanded 3kb outward, upstream for the TSS and downstream for the TES, and 5kb inward, in reverse order. These sections were divided in 50bp sections. If genes were shorter than 5kb (X%), the bins were further expanded from each direction. For the scaled genes analysis, each gene body was scaled to match an average sized gene (20kb) and extended unscaled with 3kb. Methylation averages were then extracted from each bin using the calculateMatrix tool in deeptools generating a matrix with TSS and TES concatenated bins as columns and genes as rows. The scaled analysis also followers a similar approach with bins in columns and genes as rows.

The resulting matrix was factorized using a PCA (from FactoMineR package) with no scaling. The NA values in the matrix, representing bins with no methylation signal, were imputed automatically using the mean value of the column. Per gene, the average number of NA values was . Significance for the number of principal components was extracted comparing to a broken stick model (from the vegan package), which simulates a non-signal scenario. The resulting coordinates of each gene for the top three principal components were grouped using medoids clustering (function cluster::pam in R). The number of clusters selected (k = 5) was chosen from a range (2 to 7) after visual inspection of the resulting methylation profiles and genomic characterization. Although a selection process based on silhouette index and sum squared of the residuals was also performed, the continuous characteristics of the clustering and the lack of defined numerical limits made this approach too conservative. The reader might interpret these clusters as data driven blocks.

To extract the methylation profile of every gene cluster, genes were grouped according to their assigned cluster and the average value was computed for each bin. This profile is indicative of the different methylation profiles in each group. Meta profiles of the methylation along the gene body were computed using the computeMatrix utility from deeptools in reference point mode. Plotting profiles were performed using *in house* scripts which also included the measure of a confidence interval. The confidence interval of the median is measured using the indices of a binomial distribution with the given sample size equal to the amount of rows tested, here, the number of genes in a specific cluster. Confidence interval levels are always 95% two-tailed if not stated otherwise.

For the genomic characterization of the profiles, genes were tested for local enrichment of histone marks, promoters, and enhancers and chromatin states. Histone marks used to characterize the gene clusters were obtained. Promoters and enhancers were downloaded from the FANTOM dataset but pooled across all expression levels. Chromatin states were downloaded as above. The division of genes categories according to the CpG content in their promoters was extracted from the supplementary material of ref[18] for CpGi genes and was calculated as in ref[26] for the HCG genes. In brief, CpG instances were tallied in each promoter and normalized against its CG content. This measure was then modeled by a Gaussian mixture model (using mclust package) with two components.

While the test for promoters and histone marks followed a similar methodology that the methylation meta profiles of the clusters, the overlap with chromatin states was computed using a co-occurence test. The enrichment of each cluster with the intersected classes was measured by dividing the observed and expected values in the matrices used by the chi square test. The individual p value of every cell was calculated using pair-wise fisher exact test.

Mutation rates estimation using Negative Binomial regression

The estimation of the mutation rate was performed using a Negative Binomial regression.

For the mutation rate at UMR or LMRs we compared the mutation accumulation at the region of interest (ROI) against their flanks. We defined flanks as the regions separated from the ROI by 1 width. Each flank had half of the width of the original ROI. This essentially translates to splitting the UMR/LMR in two halves and moving each section one width in the corresponding direction. Mutation rates are always represented as the ROI over flanks. Using this design, both the null and the ROI regions are likely in the same replication time domain minimizing the need to control for this co-factor. At the same time, separating these regions by one width allows us to detect clean signals which can not be underestimated due to loose ends when detecting the undermethylated region.

Mutations were stratified according to their trinucleotide content and according to their overlap with a ROI or a flank. After, mutations were tallied over those feature effectively pooling across types of regions. Likewise trinucleotides of the reference sequence were also tallied in the ROIs and flanks to determine the nucleotides at risk for each context. These values were used as an offset in the regression allowing us to control for the sequence context both at the ROI and the null regions.

The function MASS::nb.glm is then used to perform the negative binomial regression over the data table. The total number of rows is equal to the number of contexts used (96) multiplied by the region channels (2). This step leads to a formula such as:

$$\text{Mutations} \sim \text{ROI} + \text{offset}(\ln(\text{ntp\_at\_risk}))$$

Throughout the analysis of this study other features can also be controlled for by removing the ROI which overlap with a given external feature. While this reduces the number of available mutations the same methodology is used. If not stated otherwise, mutation rate estimates measured with external confounded features use this approach.

Alternatively and when explicitly stated in the results or figures, control for other features can also be performed within the same regression. The process is similar but includes an intersection step before the mutations are tallied over the region types. Different regional channels (essentially types of ROI) are intersected together to generate all possible combinations. Mutations occurring outside the intersection of two channels will be discarded. Mutations are then tallied according to the trinucleotide mutation type and each categorical interactions of the sites and the regression will be performed as above by adding the second channel in the regression formula such as:

$$\text{Mutations} \sim \text{ROI1} + \text{ROI2} + \text{offset}(\ln(\text{ntp\_at\_risk}))$$

The resulting estimates are the coefficients of each ROI feature and they represent its mutation rate of each channel against its null or reference section. For the UMR basic mutation rate estimates, the reference value are the flanking regions. The estimate is given as the natural logarithm of the odds ratio which can then be later transformed to logarithm in base 2 or as a percentage change. If not stated otherwise, mutation rate enrichments on figures are displayed as a natural logarithm.

Tissue specific analysis of the mutation rate

To differentiate mutation rates in different classes of UMRs we stratified them according to the overal with several functional features. UMRs for specific tissues were extracted as above and then used for estimation of mutation rates against all tissues, both matching and non-matching. Thus, all tissue specific UMRs were tested against all cancer types.

Chapter 6

Three-dimensional chromatin foci of mutational processes in human tumor genomes

Three-dimensional chromatin foci of mutational processes in human tumor genomes

David Mas-Ponte[1] and Fran Supek[1,2]

1) Institute for Research in Biomedicine (IRB Barcelona)
2) Catalan Institution for Research and Advanced Studies (ICREA)

The three-dimensional chromatin conformation of the genome has been associated with the variability of mutation rate at the coarse, megabase scale, where lamina-associated domains, and the TADs associated with late replication time present higher mutation rates. This suggests the spatial organization of chromatin can affect domain-scale mutation processes, and we asked if there exist finer-scale hypomutated or hypermutated chromatin spatial regions in human cells. Here, we present a systematic analysis of the mutational processes in the three-dimensional chromatin organization, by considering local mutation rate variability at chromatin loop anchors, loci that are in spatial contact with another distal locus. Loop anchors are protected from mutations from a diverse set of mutational signatures, most prominently the widespread signature of cytosine deamination, signature 1 and the UV DNA damage, signature 7a, which show a clear depletion at these loci. In contrast, some mutational signatures, like the AID-associated mutagenic activity, which shows an enrichment, possibly stemming from AID targeting in the somatic hypermutation in B-cell lymphocytes. In order to elucidate mechanisms of the mutation depletion in chromatin loop anchors seen in SBS1 and SBS7, we analyzed the role of multiple overlapping epigenetic features. DNA methylation for signature 1 and the chromatin states and DHS regions for SBS7a were able to explain a large proportion of the mutation rate variability, suggesting causal roles of the epigenetic features rather than chromatin folding per se. Finally, we implemented a methodology to detect clusters of mutations in trans, i.e. those distal in the one-dimensional DNA sequence but proximal in three-dimensional space. This method rigorously accounted for the particular mutation rate constraints that we observed across these chromatin looping sites. This analysis reveals a significant enrichment of spatially clustered mutation pairs in lymphoid tumors, bearing a characteristic mutational spectrum of AID activity, suggesting that AID forms spatial mutagenic foci in chromatin. Together, these analyses highlight the variability of mutation rate at a medium scale in three-dimensional chromatin organization. This is in large part explained by a set of epigenetic features that associate with loop anchors, converging onto a mutation protective chromatin environment. We also show the existence of a localized hypermutation in the three-dimensional nuclear space in human cells.

Introduction

The sequencing of human tumors and healthy somatic tissues has revealed a large set of mutagenic processes acting in somatic human cells. Distinct genomic and epigenomic features can influence the mutation rate at different scales, from the trinucleotide content[1,2] to large replication time domains[3,4]. Chromatin folding, or more generally the three-dimensional organization of the genome can also influence the mutation processes that are active locally, with DNA located at the nuclear periphery and in lamina associated domains harboring more mutations due to both increased DNA damage[5] and reduced repair[6]. Active and inactive topological associated domains (TAD)s accumulate less and more mutations, respectively[7], which may stem from their correspondence with early-replicating and late-replicating domains[8].

In addition to chromatin organization, somatic mutation rates are also heterogeneous at the sequence level, for instance, generating mutation groups or clusters of closely spaced mutations that share the same molecular event-of-origin. Mutation clusters were previously identified as a result of the activity of the APOBEC family of cytosine deaminase enzymes and also of methylating DNA agents[2,9] in ssDNA, generating DNA strand coordinated mutations. Here, we hypothesized that there are certain mutagenesis mechanisms particularly relevant for distal DNA loci that are in contact in 3D space. In particular, to test this hypothesis we consider chromatin LAPs and generalize the methodologies for detection of mutation clusters towards the 3D chromatin interaction map of the genome. Firstly, we report a characteristic hypomutation around 10Kb adjacent to LAPs, for specific mutational signatures like that resulting from spontaneous deamination of methylated CpG sites (SBS1), plausibly due to the reduced DNA methylation levels in LAPs. Secondly, taking this local hypomutation of LAPs into account, we devise a method to quantify the excess of mutations co-occuring in the trans-interacting loci, and mutational signatures thereof . We detect a significant enrichment of the AID mutagenic process only in SHM-positive lymphoid cells. This enrichment, thus, suggests that the activity of AID can cause 3D clusters of mutations situated in distal regions of DNA. Our analyses also suggest the possibility of additional 3D clustered mutational signatures.

Results

We first compiled a large set of chromatin loop anchors from the literature and additionally by identifying them with specialized software from published 3D genomic datasets (Supplementary Table 1)[10–13]. In brief, our final dataset comprised loop anchors from: (i) ChIA-PET experiments, targeting cohesin, CTCF and RNA polymerase II,[10,12] (ii) HiC *in situ* experiments[11] and (iii) micro-C experiments[13] in human cells, in total 20 datasets. The ChIA-PET loops were all obtained from the literature while the micro-C and HiC loops were called *de novo* or extracted from the literature. We explored genomic characteristics of each set of loop anchors to determine if sets were comparable and were representative of the sample (Fig. 1A). The chromosomal loops extracted from ChIA-PET experiments, similar to the low resolution *in situ* HiC loops, exhibited a strong association with canonical insulator motifs with cohesin binding and enriched CTCF motif directionality (Fig. 1B,D). Loops extracted from the micro-C experiments showed less canonical loops but the ones that were detected still showed a substantial enrichment in CTCF directionality (Fig. 1C).

The other sets of loop anchors were more heterogeneous and varied in size and in the association with epigenetic factors (Supp. Fig. 1,2). We then performed a filtering step to retain only loops observed in multiple experiments, after applying this requirement, the homogeneity of the sets was significant and the enrichment of canonical CTCF motifs was similar to previously published individual high-quality datasets[11] (Supp. Fig. 3). We also divided the extracted loops in sets according to multiple characteristics, like the chromatin states or the chromosomal compartments (see Methods).

For this global set of chromatin loop anchors, we next explored the mutation rate profiles in the loci they span. Loop anchors defined from HiC or micro-C were arbitrarily generated from the bins of the interaction map, while ChIA-PET loops are more precisely located around the DNA bound to cohesin (or protein of interest). Mutations were significantly reduced at these sites for a window around ~10kbp for signature 1 and 7a (Fig. 2A,B). To systematically characterize the mutation rate change in the anchors, we stratified the mutations according to the mutational signatures (Methods) and calculated the odds ratio of every mutational signature comparing the observed and expected accumulation in the anchors and flanks Fig. 2C). By comparing the resulting odds ratio from all loop sets and all mutational signatures we can see a general trend of relative reduction of mutation rate at the anchors Fig. 2D). In particular mutations assigned to SBS1 and SBS7 showed the greatest reduction Fig. 2E). In contrast, SBS9 showed a positive enrichment in the anchors Fig. 2E). We analyzed the odds ratio of every signature in a PCA, which yielded two principal components associated with the mutation rate depletion at chromatin loop anchors Fig. 2F). Although most signatures contained a slight depletion (Fig. 2D), the SBS1 and SBS7 mutational signatures showed a stronger effect. The mutation rate profiles for SBS1 (Fig. 2A) showed that while the mutation rate appeared approximately flat (uniform), the mutations in the functional portion were expected to increase based on a trinucleotide aware randomization suggesting that the mutation risk is in fact overall reduced at LAPs.

In order to elucidate molecular mechanisms relevant to the reduction of mutations as LAPs we next fit a negative binomial regression model to compare the variability of the LAPs with other overlapping genomic or epigenomic features. We segmented the loop anchor into upstream and downstream sections and compared them with the central region, and the flanking regions in their vicinity (Fig. 3A). In this analysis we also included the rest of the genome so we can intersect these regions with other genomic features. The introduction of other known regions that correlate with the local mutation rate in a joint model will account for the local influence in their overlapped in measuring the mutation depletion seen in anchors. Then, if the reduction of mutation rate is explained by another factor, the difference of the anchor and flanking regions would be diminished. The set of features tested are DNA methylation level, chromatin states (ChromHMM), DHS levels and DNA replication time domains.

For the mutation depletion in anchors associated with SBS1 we saw that DNA methylation levels were almost completely responsible for the local mutation rate depletion of loop anchors (Fig. 3B). Our model predicted a 50% mutation depletion associated with the loop anchor sites for SBS1 when not controlling for DNA methylation, while this value was reduced to only 16% when including the DNA methylation bins into the joint model. In other words, reduced DNA methylation levels can explain most of the mutation rate depletion at loop anchors (Supp. Fig. 4). Other local factors were also relevant to explain the hypomutation at anchors, with chromatin states reducing the depletion to 34% and DHS reducing it even lower to 30%. We expect the hypomethylated DNA fragments to overlap both with DHS and active promoters, limiting our ability to fully disentangle the mechanism of mutation reduction (Fig. 3B). Expectedly, controlling for replication time was not sufficient to remove the association of LAPs with mutation rates, because RT is variable only at much coarser genomic scales (hundreds of kilobases) than the width of loop anchors. We note that RT was however important to explain for the change in mutation rates between the anchors and their flanking regions, comparing against the rest of the genome (Fig. 3B). This observation was evident for multiple signatures and potentially reflects the enrichment of loop anchors in early replicating time regions.

We also tested the depletion of mutations in SBS7a (UV mutagenesis), which showed a similar reduction when incorporating the local covariates as SBS1. In the case of SBS7a, however, the factor which reduced the mutation rate more strongly were the chromatin states (Fig. 3C), which incorporates information on the transcriptional status of the region. The effect of DHS and DNA methylation also reduced significantly the observed depletion in anchors (Fig. 3C), suggesting that chromatin marks, DHS (chromatin accessibility) and DNA methylation can jointly determine the UV mutagenesis at chromatin loop anchors. However, while DHS had an important effect on mutation rates[14] (Supp. Fig. 5), DNA methylation only had a moderate effect (Supp. Fig. 5), suggesting that the hypomethylation of the anchor plays a lesser role in the reduction of mutations derived from UV, in contrast with the aging-associated SBS1.

Together, the different chromatin features accumulated in loop anchors, particularly DNA methylation and DHS, might be the cause of the observed protection of LAPs from mutation, rather than some intrinsic 3D folding property of the LAP.

With a better understanding of how the different mutational processes generate mutations at loop anchors, we used this as a baseline expectation to derive a methodology to detect enrichment of pairs of mutations bridging the LAP. Essentially, these pairs of mutations are far on the 1D sequence, but close in 3D space, constituting mutation trans-clusters. We calculate the number of 3D clustered mutation pairs i.e. loop anchors with mutations in both ends, and compared this with the expected number obtained from randomly paired anchors. The upstream part of an anchor was paired with the downstream part of another anchor within the same replication domain, at most up to 100kbp distance from the original one. The resulting mutation pairs in both sets were tallied across samples (Fig. 4A). We obtained observed versus expected ratios (O/E) for our set of samples (see Methods). Overall, there was no clear deviation from the expected values and the majority of samples showed values close to 1, thus similar values for observed and expected pairs (Fig. 4B). This result suggests that either mutational trans-clusters are rare in cancer genomes or that the analysis is heavily under-powered, due to the low genomic coverage of these anchor sites and/or low number of tumor samples. More WGS sequenced tumors or more sensitive loop detection algorithms or higher-resolution Hi-C datasets might improve these results and highlight other mutational processes with 3D activity.

When considering specific tissues, however, the set of blood tumors did contain a consistent positive enrichment in the OE ratio (Fig. 4C), implying 3D mutation clustering. This subset of blood samples showed up to a 5x enrichment compared with the neutral values. Interestingly, the mutational spectra of this enrichment shows a high cosine similarity with SBS84, a mutational signature resulting from AID mutagenesis (Fig. 4D). Consistently with our previous result, we saw that when other leukemia samples contained mutations at the immunoglobulin loci, considered then as mature B-cells, they also showed an enrichment in mutation pairs (Supp. Fig. 6). This association strongly points toward the SHM process (which includes the activity of AID) as a strong candidate for the observed 3D mutational clusters. Mutations coming from this process are known to cause hypermutation (large groups of mutations) in the one-dimensional DNA sequence[15] and these groups were reported to be unusually common in promoters/enhancers that make many 3D chromatin contacts[16–18]. Here, we show that the SHM process in lymphocytes also likely generates DNA damage in spatial hotspots in the nucleus; these 3D mutation clusters arise in a coordinated manner on both ends of the interacting DNA in three-dimensional space.

Discussion

In summary, we systematically quantified the mutagenesis at loop anchor points (LAPs) and showed a consistent depletion for most mutational signatures, while some like the SHM-associated

SBS9 show an enrichment in LAP loci. This is consistent with recent reports, which proposed an enrichment of structural variants but also a depletion of point mutations both in anchors and in TAD borders in a pan-cancer analysis[7,19]. Building upon those reports, however, we suggest that this depletion cannot only be explained by the replication time of these sites alone, but that they are protected due to a spectrum of distinct (epi)genomic features that co-exist in loop anchors (Fig. 3). For SBS1 mutations, generated from the deamination of the methylated cytosine at CpG dinucleotides, we show that the mutation reduction is probably caused exclusively by a hypomethylation of DNA at these sites. A Our report of widespread DNA hypomethylation at chromatin loop anchors, is consistent with prior reports that demethylation of the DNA might be required for some CTCF-mediated insulator loci[20–24] and that large demethylation domains can contribute to long-range 3D contact interactions[25] (Fig. 3B). Specifically for the special case SBS7 mutations, resulting from UV DNA damage, however, other features like DHS (accessible chromatin, promoting nucleotide excision repair[14], or active transcription seem to be more likely cause for the hypomutation. We suggest that a combination of features that occur at LAPs[26] influences different mutational signatures in different ways, converging onto hypomutation gradients at a similar genomic kilobase scale (Fig. 2A,B).

Specifically the AID/APOBEC cytosine deaminase mutational signatures like SBS9 and the related SBS84 show, contrary to other signatures, an enrichment in LAPs (Fig. 2F, 4D). This enrichment is likely linked with their role in SHM, a process of antibody diversification in B-cells, which has also shown significant off-target activity meaning it affects many other loci in addition to antibody genes themselves[27]. Prior reports already showed that AID targeted preferentially 3D interacting regions[16,17]. Consistent with this targeting of sites with high propensity to interact, we find evidence for an excess of 3D mutation clusters in loop anchors precisely for AID mutations in blood cancers (Fig. 4C). Importantly, this process was more pronounced in DLBCL and in the SHM-positive subset of lymphocytic leukemias providing a strong causal link to AID (Supp. Fig. 6). Other mutational signatures also present an excess of paired trans-clusters of mutations, but the size of our current dataset seems to limit the statistical power to identify these signatures, limiting to those with the highest burdens (Fig. 4C,E).

Overall, this study highlights the heterogeneous rates of mutational accumulation in trans-interacting loci such as chromatin loop anchors, providing a better baseline mutation rate profile for these sites that often overlap with functional elements, and allowing identification of 3D mutation clustering in the human genome.

References

1. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
2. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
3. Supek, F. & Lehner, B. Scales and mechanisms of somatic mutation rate variation across the human genome. *DNA Repair* **81**, 102647 (2019).
4. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
5. García-Nieto, P. E. *et al.* Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *EMBO J.* **36**, 2829–2843 (2017).
6. Smith, K. S., Liu, L. L., Ganesan, S., Michor, F. & De, S. Nuclear topology modulates the mutational landscapes of cancer genomes. *Nat. Struct. Mol. Biol.* **24**, 1000–1006 (2017).
7. Akdemir, K. C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.* **52**, 1178–1188 (2020).
8. Marchal, C., Sima, J. & Gilbert, D. M. Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **20**, 721–737 (2019).
9. Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
10. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).
11. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
12. Grubert, F. *et al.* Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* **583**, 737–743 (2020).
13. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol. Cell* **78**, 554-565.e7 (2020).
14. 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).
15. Ye, X. *et al.* Genome-wide mutational signatures revealed distinct developmental paths for human B cell lymphomas. *J. Exp. Med.* **218**, e20200573 (2021).
16. Qian, J. *et al.* B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. *Cell* **159**, 1524–1537 (2014).
17. Hübschmann, D. *et al.* Mutational mechanisms shaping the coding and noncoding genome of germinal center derived B-cell lymphomas. *Leukemia* **35**, 2002–2016 (2021).
18. Casellas, R. *et al.* Mutations, kataegis, and translocations in B lymphocytes: towards a mechanistic understanding of AID promiscuous activity. *Nat. Rev. Immunol.* **16**, 164–176 (2016).
19. Kaiser, V. B. & Semple, C. A. Chromatin loop anchors are associated with genome instability in cancer and recombination hotspots in the germline. *Genome Biol.* **19**, 101 (2018).
20. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
21. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
22. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482–485 (2000).
23. Maurano, M. T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Rep.* **12**, 1184–1195 (2015).
24. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
25. Zhang, X. *et al.* Large DNA Methylation Nadirs Anchor Chromatin Loops Maintaining Hematopoietic Stem Cell Identity. *Mol. Cell* **78**, 506-521.e6 (2020).
26. Yan, J. *et al.* Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell* **154**, 801–813 (2013).
27. Khodabakhshi, A. H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308–1319 (2012).

Figures

Figure 1

Characterization of the loop anchor sets obtained for this study. (a) In the left number of loops in each reference union, meaning mutations contained in any of the called sets of that category; in the right, distribution of loop sizes in each category. (b-c) show proportion of loop anchors that overlaps with a CTCF motif with binding evidence for CTCF and cohesin, hence “selected”. (b) shows non-tissue specific cohesin loops. (c) shows micro-C loops from H1. (d) shows HiC loops extracted from ref[11] for the GM12878 cell line.

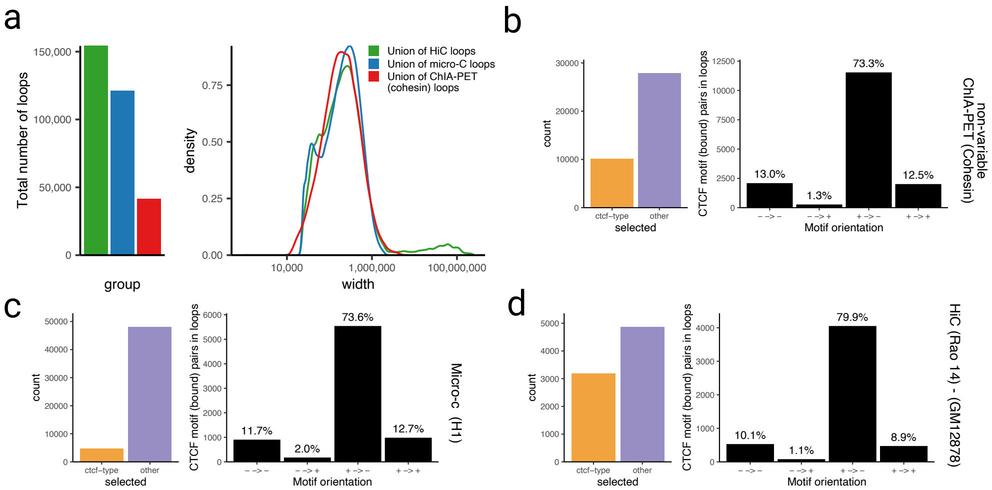


Figure 2

Mutation rate depletion around loop anchors. (a-b) Mutation rate profiles measured at loop anchors comparing the observed mutations (in a solid line) against a randomized baseline (in a dashed line) for signature stratified mutation calls. (c) Diagram representing the methodology to compute odds ratio of at the anchor sites. (d) Odds ratio analysis showing mutation rate depletion when comparing the observed loops against a shuffled set. (e) Odds ratio values for signatures 1, 7a, 7b, 5, 2 and 9 (colors) in 4 types of loop anchors (size bins 2, medium and 3, large) and transcription and enhancer overlapping anchors. (f) PCA from the Odds ratio analysis of all mutational signatures. In the plot, the correlation with PC1 and PC2 is shown as an arrow. Point represent each instance in the PCA and is equivalent with the sets in e.

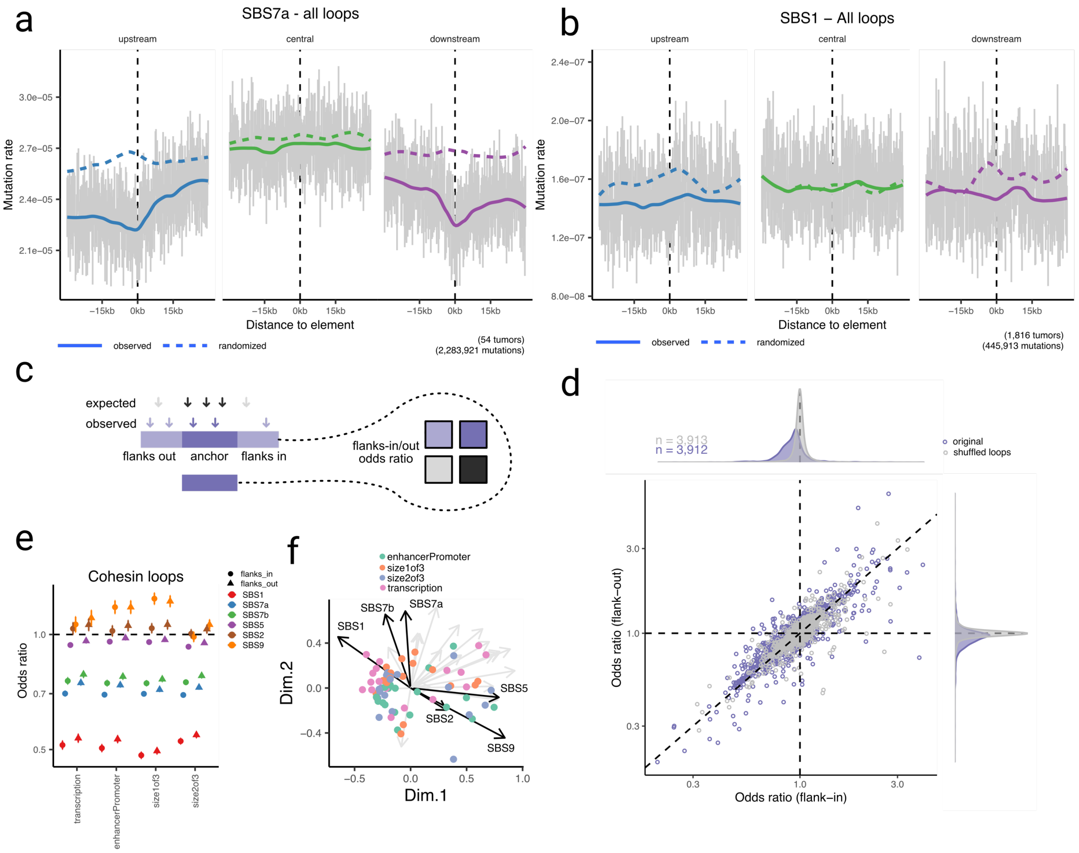


Figure 3

Modeling of mutation rates and overlapping covariates through a negative binomial regression: (a) Diagram of the negative binomial regression model used to determine the relative mutation enrichments in each segment of the loop anchor. Each extra column represent the base model with the addition of an extra feature (b-c) Coefficient for each regression focusing in the segments around the loop anchor. Y axis represents the enrichment of mutations in base 2 logarithm. Each color represents one regression with the base model depicted in (a) and the addition of the extra feature, color coded.

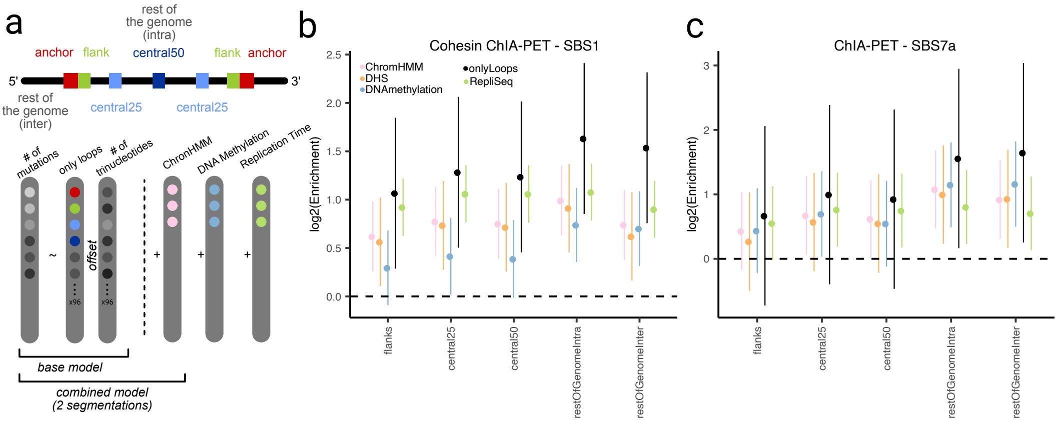
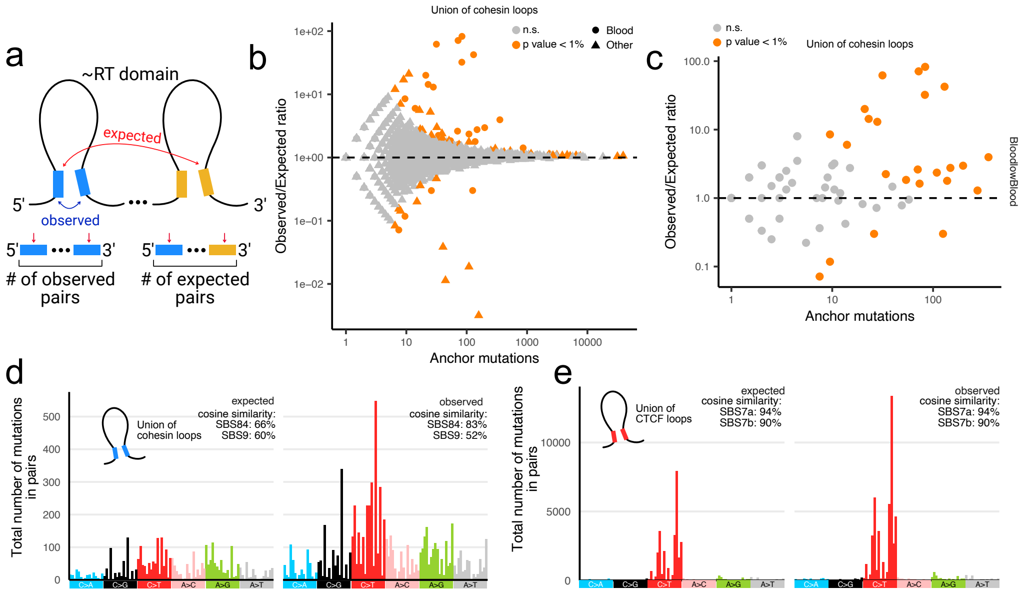


Figure 4

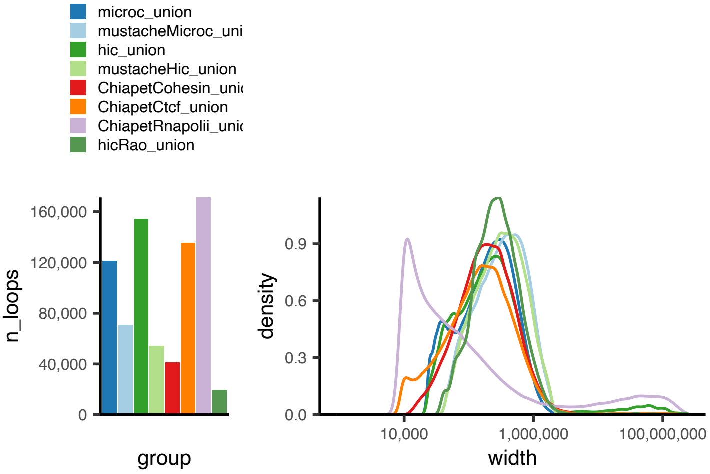
Mutation trans-cluster detection in human tumors: (a) Diagram of the method used for the detection of mutation cluster pair enrichment. In brief, the loop anchor pairing is randomized within 100kb of the original pair creating an expected set of loop anchors. Mutations are then tallied in both sets and the enrichment is calculated as the ratio of both figures. (b) O/E ratios for all samples showing no overall mutation enrichment in the cohesion union loop set. In orange samples which show a significance lower than 1% in a poisson ratio test. (c) same as in b but only for Blood samples. (d-e) Mutation profiles (trinucleotide counts) of the mutations in expected and observed loop sets for Blood samples in cohesin union loops (d) and Skin hypermutated samples for CTCF union loops (e).



Supplementary Figures

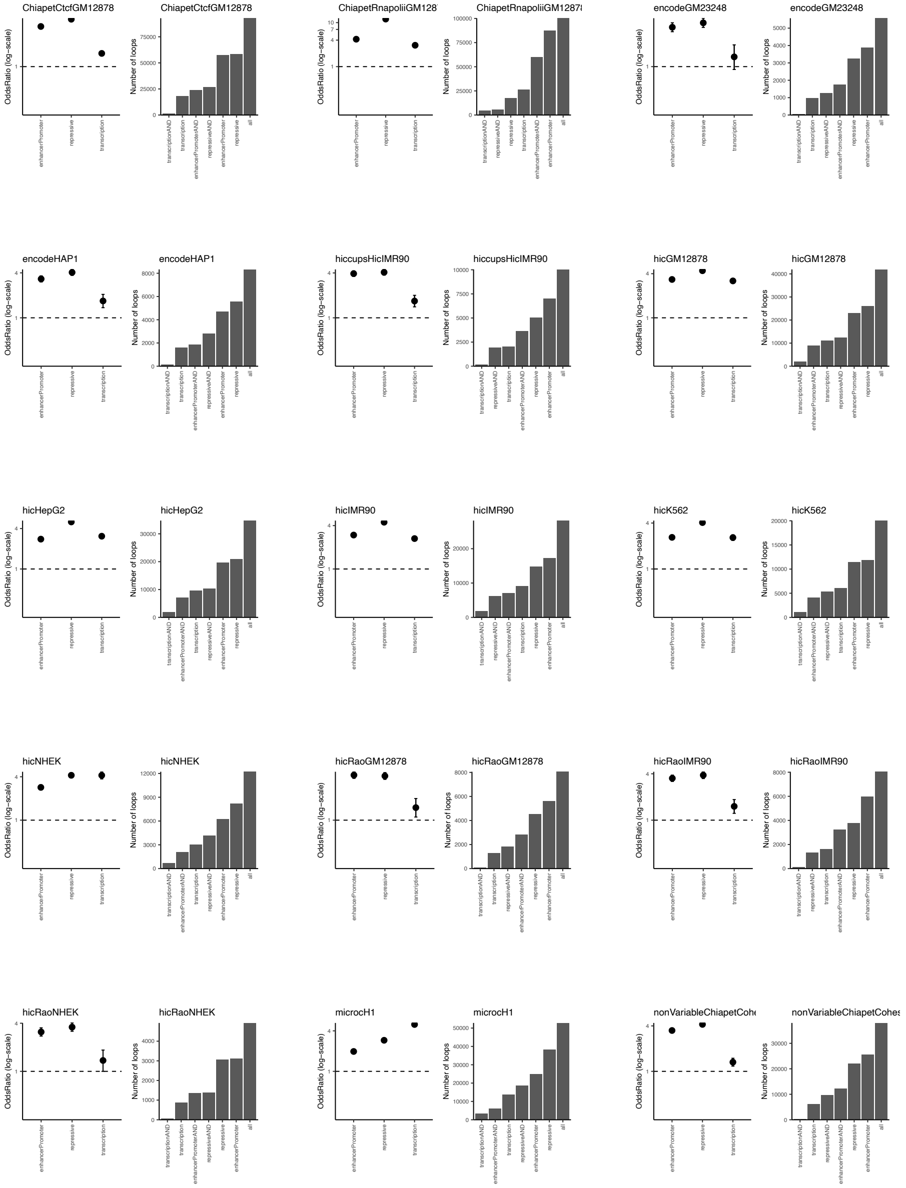
Supplementary Figure 1

Properties of the different set of loop anchors detected in this study. Colors represent the dataset source and type. Left panel shows total number of loops extracted in each category. Right panel shows the distribution of lengths of each loop class.



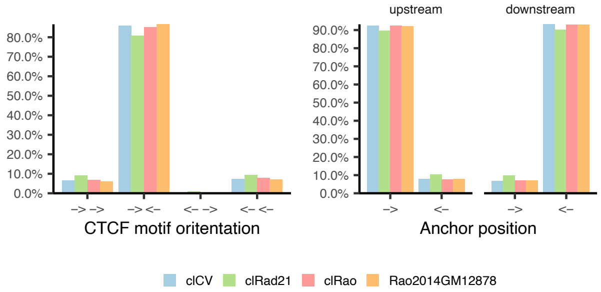
Supplementary Figure 2

Genomic characteristics and chromatin enrichment of a representative set of loops for each experiment type.



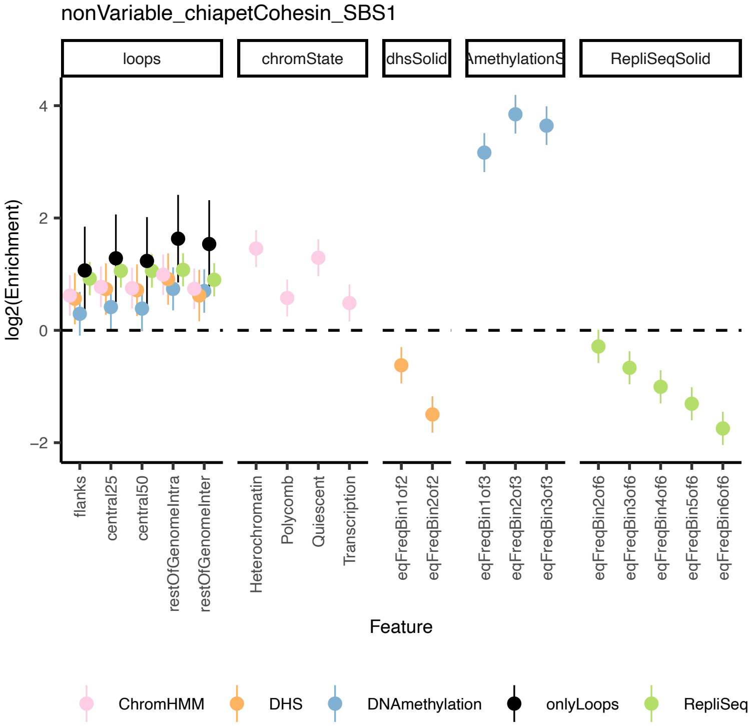
Supplementary Figure 3

CTCF motif directionality scores after filtering for the motifs only co-occuring in multiple datasets.



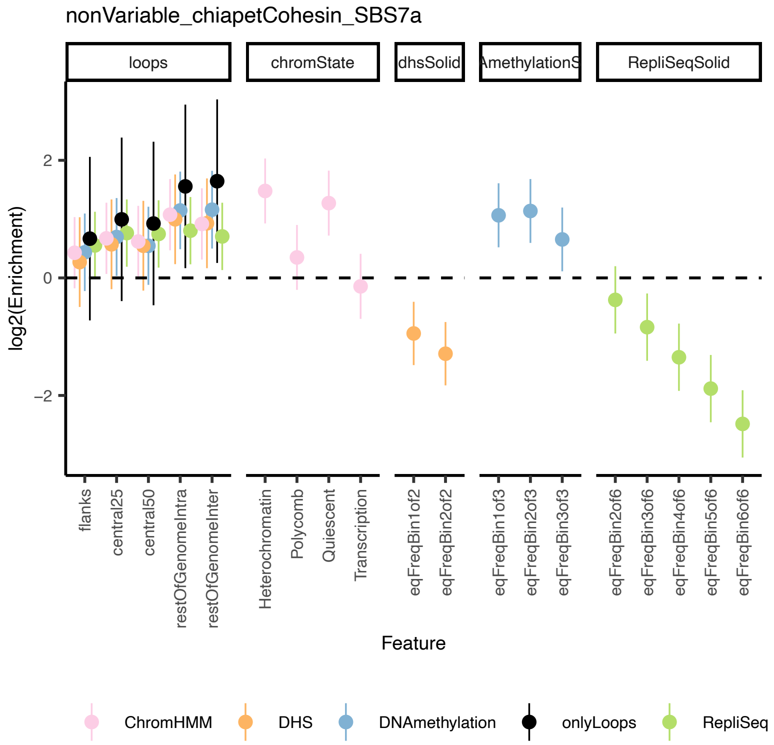
Supplementary Figure 4

Relative to Fig. 3b. Coefficients of the negative binomial regression to measure mutation rate estimates along the loop anchors. Each box includes the set of coefficients in each segmentation. For DHS, methylation and replication time, bins represent equally covered sections of the genome with equivalent signal. Lower bins contain less signal, thus, 1ofX represents the lowest value while XofX the highest, for replication time, higher values represent earlier replication times. The reference bin for the DHS and for the methylation segmentation is the rest of the genome. The reference bin for the chromatin states segmentation is the promoter region.



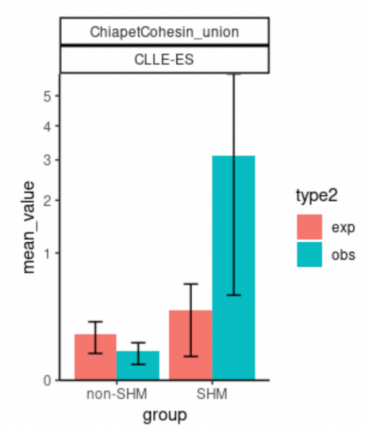
Supplementary Figure 5

Relative to Fig. 3c and equivalent to Supp.Fig. 4.



Supplementary Figure 6

Average mutations found in observed versus expected loops for the CLLE-ES dataset. Samples are stratified according to their SHM status. SHM+ (SHM), which indicates that IGG loci contained A>G mutations, and SHM- (non-SHM), which indicates that no mutations were found in the IGG loci.



Supplementary Tables

Supplementary Table 1

List of interaction maps used in this analysis, datasets marked as raw were processed with an *in house* to obtain loop anchors.

Cell line	Type	File type	Code	Source
microcH1	micro-C	raw	4DNFI2TK7L2F	4DN
microcHFF	micro-C	raw	4DNFIPC7P27B	4DN
hicIMR90	in situ HiC	raw	4DNFIH7TH4MF	4DN
hicGM12878	in situ HiC	raw	4DNFI1UEG1HD	4DN
hicNHEK	in situ HiC	raw	4DNFIL9M97T2	4DN
hicHepG2	in situ HiC	raw	4DNFICSTCJQZ	4DN
K562	in situ HiC	raw	4DNFITUOMFUQ	4DN
HelaUnS	in situ HiC	raw	4DNFIE7V3DN9	4DN
HelaSync	in situ HiC	raw	4DNFI7OMRYXC	4DN
KBM7	in situ HiC	raw	4DNFIT96Z365	4DN
GM23248	in situ HiC	loops	ENCFF432KUX	ENCODE
HAP-1	in situ HiC	loops	ENCFF817TXQ	ENCODE
GM12878	in situ HiC	loops	GSE63525	Rao Cell 2014
IMR90	in situ HiC	loops	GSE63525	Rao Cell 2014
NHEK	in situ HiC	loops	GSE63525	Rao Cell 2014
GM12878	ChIA-PET (CTCF)	loops	GSM1872886	Tang Cell 2015
GM12878	ChIA-PET (RNApol II)	loops	GSM1872887	Tang Cell 2015
HeLa	ChIA-PET (CTCF)	loops	GSM1872888	Tang Cell 2015
HeLa	ChIA-PET (RNApol II)	loops	GSM1872889	Tang Cell 2015
Multiple tissues	ChIA-PET (Cohesin)	loops	Supplementary	Grubert Nature 2020

Chapter 7

Results Summary

The first Results chapter of this thesis, chapter 3 , summarizes the development of new statistical tools to identify local hypermutation events from somatic mutation data and the application of this new methodology to identify a common mechanism generating diffuse, short mutation clusters associated with APOBEC and mediated by the activity of MMR.

First, we aimed to further characterize the landscape of the mutation spectra of the clustered mutational processes and to overcome the limitations of previous methods. We built upon previous work[97] to improve the systematic detection of mutation clusters. Although we focused in mutational clusters generated by APOBEC enzymes (see section 1.4.1.2), our methodology does work more generally and is able to detect various types of clustered processes.

We combined a trinucleotide-aware genomic randomization algorithm with an improved statistical significance assessment based on the *local-fdr*[335] that allowed us to estimate a threshold for significant clustering, even in hypermutated tumor genomes.

We used the inter-mutational distance of the adjacent mutation and compared them with the randomized set. We also included other additional features in order to maximize the power of our methodology. In brief, we classified the mutation calls according to their clonal fraction, derived from the estimated cancer cell fraction, and we enforced strand-coordination between the clusters. Once extracted, we were able to recover the APOBEC mutation clusters in 76% of our available samples.

From the identified events, we fitted a Poisson mixture model to the distribution of event counts and obtained a solution with 2 significant components. These components consisted in long-runs of 5 or more mutations, the previously reported kataegis[90,95,185] and short pairs or triplets which we termed *omikli* from the Greek word for fog.

We characterized the processes by measuring the components' genomic characteristics like the distribution around the genome and pentanucleotide predisposition. We find that while APOBEC *kataegis* is enriched around break-points and for A3B-like pentanucleotides (see section 1.4.1.2), the APOBEC omikli mutations show an enrichment in early replicating sections of the genome and for A3A-like sequences. Overall, *omikli* mutations correlated strongly with the unclustered portion of A3 mutations, while *kataegis* presented a weaker association. These characteristics suggested that *omikli* and *kataegis* occurred by independent mechanisms and that *omikli* clusters and the bulk unclustered mutagenesis partially shared the same mechanism.

A further characterization of the genomic properties of *omikli* mutations suggested DNA replication time associations and the distribution of intermutational distances, we gathered evidence that suggest the main source of this new mechanism are the ssDNA intermediates occurring in the DNA mismatch repair pathway. Our data suggest that this mechanism also plays a role in the generation of the majority of APOBEC mutations, mostly unclustered, thus contributing to a substantial proportion of the mutational accumulation in various cancers genome wide.

Because the MMR pathway targets the early replicating and gene rich portions of the genome[175] , A3 *omikli* mutagenesis is also directed towards those regions. Therefore, the overall A3 mutation burden has a high power to generate impactful mutations as it is partially directed to active regions.

Using a simple model for the prediction of mutations in driver genes, the mutagenic potential of A3s exceeds some common carcinogens like tobacco smoking or UV light, and is commonly directed towards certain cancer genes, e.g. chromatin modifiers.

Together, in this chapter (3) we developed a rigorous, sensitive statistical methodology for identifying mutation clustering, and applied it to cancer genomes to identify a new and prevalent type of mutation clustering (omikli), and one mechanism that can generate these mutation clusters by combining activity of APOBEC and the DNA mismatch repair.

Chapter 4 summarizes our contribution in the detection and characterization of mutation mechanisms present in a range of clonally expanded single cells derived from healthy, noncancerous tissues.

In this study, we interrogated both data obtained by our collaborators from in vitro single cell primary tissue expansions from muscle, kidney, fat and skin, as well as other previously published datasets. In the analysis of the somatic mutation profiles, a clear baseline process arises, which is present in any tissue (cancerous or healthy) and accumulates with age. Multiple tissues share a common main mutagenic mechanism that can be derived from the combination of CpG deamination, signature 1, and signature 5 of still unknown etiology. We performed a systematic comparison of activity of various mutational processes between healthy cell genomes, and the tissue type-matched cancer genomes. This analysis revealed that the activity of various mutational processes are overall similar in normal and cancer cell genomes. Notably, the APOBEC mutational processes were less commonly found in healthy cells.

Interestingly, we uncovered a subset of clones in the kidney samples which harbored an excess mutagenesis, with a profile similar to the previously identified Signature 40. This accumulation was also heavily dependent on the age of the donors. Cell clones with high exposure to this signature also expressed molecular markers from the proximal tubule section of the kidney, suggesting that a physiological characteristic of this set of cells might be responsible for the signature. Mutations in these cells targeted promoters and transcription binding sites, suggesting a high mutagenic potential. By comparing the mutation exposures in the healthy tissues with available tumor samples, we propose that the newly identified cell population in the kidney might give rise to the clear cell and papillary renal cell carcinomas subtypes, but not the chromophobe cell subtype.

Finally, we also focused on the differences between young and old donors. In older patients, we detected a modest loss of association with known markers of functional MMR, such as the steep gradient across replication time domains32, and its role in the accumulation of mutation peaks at CTCF/Cohesin binding sites.

Together, these results suggested a partial depletion of the repair capacity of healthy cells with age, a basal age-associated source of mutagenesis across tissues, but also the existence of a cell-type specific accumulation of impactful aging mutations in the kidney.

Chapter 5 summarizes our studies in the characterization of the role of local DNA methylation variability as a molecular mechanism that modifies the mutation density in human tumors. Although the role of DNA methylation in gene regulation is well understood, how the local variation in DNA methylation shapes somatic mutation rates is less well explored.

In this study, we show that unmethylated (UMR) regions are also generally hypomutated in a wide range of human tumors and healthy somatic tissues. The exposure of the tissue to various mutational processes shapes its predisposition to this effect: while there is depletion in the mutation rates resulting from signatures of deamination of methylated cytosines, UV light, POLE deficiency, and MMR deficiency, there is an increase in mutation rates from signatures of AID or APOBEC cytosine deaminase enzymes in the UMRs. Therefore, hypomethylated DNA loci can be either mutational coldspots or hotspots, depending on the mutagen exposure history of a particular cell.

We also characterized the UMRs by the overlap with multiple functional elements, such as promoters, enhancers and chromatin loop anchors, and observed similar characteristics within the different classes and even at UMRs outside any of these elements. This highlights the universal role of DNA methylation in the direct determination of mutation occurrence. In addition to these genome-wide distributed UMRs, we also identified DNA methylation gradients in gene bodies. Several kilobases at the 5' ends of gene bodies were commonly hypomethylated and thus hypomutated. Clustering genes by DNA methylation profiles also yielded variability in their mutation rate gradients: lowly expressed genes have a less steep gradient due to a higher relative methylation of their 5' end, and polycomb repressed genes show no relative hypomutation due to the lack of DNA methylation at their gene body.

Overall, we suggest DNA methylation is an important determinant of mesoscale, sub-genic, resolution mutation rate variability in human somatic tissues.

Chapter 6 summarizes our efforts in expanding the definition of 3D spatial local hypermutation using genomic folding estimated via the interaction frequency derived from HiC contact maps.

In this study, we curated an extensive set of CTCF/Cohesin bound set of loop anchors that were derived from a large set of developmentally independent tissues. Additionally, the compiled set of 3D maps also includes a diverse set of both molecular (Hi-C, Micro-C, Chia-PET) and bioinformatic techniques available to date (same data was characterized using multiple tools).

We then, characterized the mutation patterns enriched around the loop anchors, and designed a methodology to systematically detect significant 3D mutation clusters. We detected a general reduction of mutations in large domains (2kb) within loop anchors. This was opposite to the previously reported hypermutation in the specific binding site of the CTCF protein. We applied a systematic analysis of the mutational signatures that participated in this process, revealing heterogeneity in the effect of different signatures. The main signal focused on the mutations associated with deamination of CpG sites, Signature 1, and the mutations resulting from UV light damage, Signature 7. Consistent with this heterogeneity in the mutational processes involved, we report that the lower DNA methylation of the loop anchor sites, as well as its co-localization with DHS (DNAse hypersensitive sites) can explain the observed decreased mutation rate.

Rigorously accounting for these locally lowered mutation rates, we developed a statistical method to detect a significant enrichment of 3D-proximal (but 1D-distal) mutation pairs, "trans-clusters". Our method uses a randomization of the loop anchor pairings to measure an expected baseline. We could identify a positive enrichment of trans-clusters in a subset of B-cell lymphoma cancers, where the subtype suggested a mature stage of B-cell differentiation. Thus, in addition to clusters at the 1D sequence level, the AID enzyme mutagenesis seems to generate 3D mutation clusters in spatially interacting DNA strands, providing data to support prior hypotheses.

Together, these results show how the chromatin folding components may modulate the accumulation of certain mutation mechanisms, and demonstrate the existence of a previously uncharacterized type of local hypermutation in the 3D space.

Collectively, our results have focused mainly on the local variation in mutagenic potential of endogenous mutagenic processes, such as the methylated CpG deamination and the APOBEC mutagenesis, which contribute to substantial mutation burdens to both healthy and cancerous tissues. Although extensive work has been performed in the characterization of local variability in DNA repair pathways[175,178,192,256] , results presented in this thesis highlight that the local DNA damage distribution, either by APOBEC deamination or through the damage-promoting DNA methylation can also represent important determinants of the variability in local mutation rates.

We also highlight the disruptive potential of the studied processes by assessing the burden of (predicted) functional effects on genic sequences. In the case of APOBEC mutagenesis and in the aging-associated mutation processes in human tissues ([CREF chap:ng,chap:franco]), we report how the redistribution of mutations towards the early-replicating, gene-rich parts of the genome can increase the mutation rate in coding regions and generate pathogenic mutations such as cancer drivers. In our results, we further focus on the interaction of the methylation levels and local hypomutation, as observed in promoters and in loop anchors, chapters 5 and 6 we noted that there exists a sub-gene resolution mutation rate variability along gene bodies. This may be caused by for instance presence of intragenic promoters, or by silencing by facultative heterochromatin, which associate with hypomethylation of some parts in gene bodies. Some mutational signatures, like the common SBS1 and the ultramutating SBS10b, will be depleted at these subregions. Interestingly, however, mutations from APOBEC and AID signatures are enriched at these regions. This modulation of mutation rates due to DNA methylation gradients within-genes represents an important characteristic that might need to be taken into account when estimating selection.

These additional insights into mutation risk heterogeneity described in this thesis highlight how understanding of processes that shape the mutation burdens at various genomic loci can provide a complete picture of genome (in)stability in human tissues. We believe that the studies contained within this thesis contribute to the understanding of the mutational processes in the human somatic genome.

Chapter 8

Discussion

In this thesis, we present a systematic analysis of the patterns of mutagenesis from endogenous processes and their local variability, either through hypermutation or hypomutation. Further, we identify the plausible mutational mechanisms that causes the local hotspots or coldspots, and considered the functional impact that these mutation processes can have on genes. In particular, we report the role of DNA mismatch repair activity in the generation of APOBEC mutation clusters and also unclustered mutations, detect mutational signatures that occur in both healthy and also tumor somatic cells, quantify the role of DNA methylation in the local modulation of mutation rates and detect the 3D clustered mutagenesis resulting from AID activity upon trans-interacting chromatin regions. Furthermore, we also aim to characterize the role of how these process might contribute to functional mutations, highlighting the role of APOBEC3A mutagenesis as a strong generator of impactful mutations in various cancer genes, and revealing a sub-genic mutational gradient linked to the methylation levels across genes, which can affect differential mutation supply to various gene regions.

An important focus of this thesis was on determining molecular mechanisms associated with the detected mutational processes of local hypermutation and hypomutation. In particular, we have made contributions in describing a novel mutagenic mechanism for APOBEC clustered mutagenesis as a byproduct of the MMR pathway activity, and on the mutation rate gradients around hypomethylated regions likely to be directly caused by the lack of methylation itself.

The mutation patterns associated with APOBEC activity were detected early, during the analysis of the first sequenced cancer genomes[95] : mutation showers (groups of clustered and strand coordinated mutations) were observed in these tumors[95,185] . These clustered mutations were termed *kataegis* and were suggested to originate in long stretches of single-stranded DNA present in the intermediate states of DNA repair pathways like HR or BIR[295] . Because the activity of APOBEC needs to target ssDNA, these are indeed prime opportunities for the generation of the APOBEC mutation showers. This association was clear from the enriched mutation burden around structural variants[104] . However, the majority of mutations in the APOBEC enriched contexts were not in *kataegis* events, which are very rare. Contrary to *kataegis* , the mutational mechanism presented in this thesis, *omikli* , generates short diffuse clusters of APOBEC mutations, which are not enriched around rearrangement points, which are common and observed independently of *kataegis* , and which probably also contribute to global unclustered APOBEC burden . Furthermore, we showed that while mutations in *kataegis* events were likely caused by the activity of APOBEC3B, the APOBEC3A was the source of both *omikli* and unclustered mutations. This particular observation has been recently confirmed in human cell lines with knock-outs in APOBEC3A, 3B and related genes[143] .

The originally suggested mechanism of action for unclustered APOBEC mutagenesis was based on the relative strand bias associated with APOBEC mutations and proposed the ssDNA at the lagging strand during DNA replication as a substrate[301] , which was supported in experiment expressing human APOBEC in *E. coli*[300] . For *omikli* mutations however, the data in human cancer genomes was not consistent with this mechanism: i.e. focused on the genomic distribution of the mutations and showed a strong enrichment in early replication time, which suggested a role of MMR as their mechanism; note that the replicative strand bias of APOBEC mutations is consistent with the replicative strand bias of the MMR activity. Previous reports had already reported a similar enrichment of APOBEC mutations in early replication time[246,302] but with an unclear mechanism. The intermutational distance of the *omikli* clusters was also compatible with the ssDNA intermediate generated in the MMR activity[336,337] however not with the Okazaki fragment length in lagging strand synthesis. The depletion of these mutations in MSI tumors, deficient in MMR, further represented evidence to link the generation of APOBEC mutations to the activity of MMR. Although the current data presented in this thesis is purely observational, a previous report in human cells detected the interaction of both BER and MMR in the generation of APOBEC mutation clusters against an artificially incorporated mismatch[303,304] . In brief, when a mismatch containing plasmid was introduced in a mammalian cell, APOBEC-like mutations arose in the vicinity of the mutation, likely caused by the activity of the cell APOBEC in the ssDNA flanks; knocking down MMR reduced the APOBEC mutagenesis in that study[303] .

The role of DNA methylation in the somatic mutation rate of tumors was proposed in the first reports on landscapes of mutational signatures[90,91,95] . C>T mutations at the NCG trinucleotide, so-called Signature 1 or SBS1, were strongly suggestive of a previously described mutational mechanism[12] , the spontaneous deamination of the methylated cytosine. Later reports that specifically studied these mutations in DNA polymerase ϵ deficient and MMR-deficient tumor genomes detected a DNA replication strand bias, suggesting that methylated cytosines may promote errors in DNA copying, and also apparently lower mutation rates at gene promoters in colon cancers (a tumor type with high levels of Signature 1) consistent with the known low methylation at promoters[181,338] . In this thesis, we generalize these findings by systematically analyzing mutation rate gradients across gene bodies, separately for all cancer types and mutation signatures. A main statistical trend in mutation risk gradients was evident in several signatures including most prominently Signature 1, and tracks the typical gradient in DNA methylation across expressed genes. We build upon this finding by analyzing the patterns of mutations genome-wide (i.e. in gene bodies or elsewhere) specifically at UMRs and LMRs, segments of the genome that present a complete or partial hypomethylation, respectively. Consistent with previous reports[90,181,338] , the previously identified signatures with strand biases and depletion at promoters in colon cacncer have in our work presented a depletion genome-wide at the hypomethylated sites, and in many cancer types (see chapter 5 Fig. 1A). This effect is maintained across all kinds of hypomethylated functional elements such as promoters, enhancers (which may be intragenic) and loop anchors (see chapter 5 Fig. 2B). Also, some additional sites, without a known functional element, are hypomethylated and consistently hypomutated; some of these might be explained by polycomb silencing in facultative heterochromatin genes, which also seem to have hypomutated gene bodies. Our results are, therefore, consistent with the methylation of the cytosine being the causal determinant of the local, sub-gene-resolution mutation rate variation in multiple genomic contexts. In our model, a shared mechanism of both replication-associated mutagenesis (through the misincorporation of an adenine opposite to the 5mC) and through the spontaneous mutagenesis (thus replication independent, deamination of the 5mC to thymine)seem to coexist and both vary across loci.

Another highlighted signature in our analysis is signature 7, resulting from UV DNA damage. The role of DNA methylation at these UV damaged sites is more complex, as previous *in vitro* approaches are not clear about their potential mechanism; it is possible that the UV damage accelerates the cytosine deamination within the lesion, and also that the methylation facilitates forming of the damage[244,245] .

Perhaps more interestingly, we find that some mutational signatures show an enrichment in UMRs, thus DNA methylation can both lower and increase mutation risk depending on the exposure of each particular cell. These enriched signatures seem to be related to AID mutagenesis, signature 84, and APOBEC signatures 2 and 13. In the case of AID mutations, it is likely that their accumulation might be associated with the known, physiological AID targeting toward promoter regions in the somatic hypermutation process of B-lymphocytes[307] . The scarcity of this signature, unfortunately, prevents us to further characterize the fine-scale genomic distribution characteristics. In the same vein, signature 9 is a SHM-associated process, which occurs downstream of AID and also presents a positive association with UMRs and consistently is explained by the interaction of with known promoters (see chapter 5 Supp. Fig. 2C). The global association of APOBEC mutation risk and DNA methylation was previously reported[246] at the genome-wide level, showed an increased mutation rate for unmethylated cytosines. *In vitro* reports and other experimental data seem to corroborate a possible positive correlation[246] however others reported no correlation[250] . Our data considers local effects of methylation variation on mutation risk, and strongly supports that in the hypomethylated sites, the APOBEC-induced mutations show an enrichment.

In this thesis, we also discuss the impact the above-mentioned mutational mechanisms (such as the ubiquitous, abundant Signature 1) can have on gene coding regions, which are the functional elements in the genome most likely to get disrupted by causal somatic mutations.

The mechanism that we describe for APOBEC mutagenesis (see above and in chapter 3) shows an interesting association between a mutagenic process and a DNA repair pathway. Because the activity of MMR is focussed on more actively protecting the early-replicating regions[175,197] . which are generally enriched in genes, our analysis yields a remarkably strong functional impact potential (considered per mutation) for APOBEC (see chapter 3 Fig. 5a). Only mutations from aging-associated signature 1, with a genomic context highly enriched within genes (these have a higher frequency of the CpG dinucleotide) have a higher relative potential however their total burden is lower compared with APOBEC mutations. Therefore, in absolute terms of cumulative functional impact, the mutations from this mechanism represent a very strong genic region-targeting mutators in human tumors, with values similar to those from the UV damage (and in relative terms *per mutation* far exceeding UV damage). Although UV generates substantially more mutations than APOBEC in skin cancers, the UV mutations are preferentially corrected in the gene-rich chromosomal domains and thus represent a lower functional impact risk for the cell function. In addition to MMR likely driving APOBEC mutagenesis towards early-replicating DNA, other mechanisms might additionally explain this increased mutation potential in genes, possibly related with the role of hypomethylation of some segments in active genes (e.g. intragenic enhancers) in promoting APOBEC mutagenesis (see below, and chapter 5).

The role of local DNA methylation in mechanisms regulating activity of promoters and enhancers is widely known[220] , however, the extent of how this variability bears on local mutation rates remained less explored. Prior reports generally assessed this question[181] finding a strong correlation of the methylation and hypomutation at gene promoters, focussing colon cancer genomes with DNA repair deficiencies (we also note that promoters in e.g. skin cancers are actually hypermutated rather than hypomutated, due to increased UV damage and/or reduced NER activity[209,210]). In our chapter 5 , we extend this by systematically classifying DNA methylation profiles along gene bodies across all human genes, and report categories of genes that exhibit a distinct DNA methylation profile and also differ in their epigenomic characteristics. Interestingly, when measuring mutation burden in the different groups, the resulting hypomutation gradients are only detected in the gene categories with an extensive hypomutation at TSS, meaning that the main gradient in somatic mutation rates along gene bodies likely stems from variable DNA methylation. Consistently, repressed genes, which show an overall methylated promoter, show no discernable gradient of mutation rate along the gene body. We believe that these findings represent contribution in the characterization of mutation variability in the gene-level and sub-gene level and that can be useful, as suggested by our selection analysis, in the better estimation of a mutation rate estimate for genes and other functional elements affected by differential DNA methylation such as promoters and enhancers.

A general limitation of the presented work, also common in other cancer genomic studies, is the use of mostly observational mutation data from human tumors, meaning that the causes of mutagenesis were not strictly controlled. While this provides the advantage of working directly with genomes of relevant human cell types, the lack of empirical evidence for causal effects (which can be modelled in model organisms or cancer cell lines[127]) represents a limitation of any cancer genomics study. In the same vein as the previous limitation, the power of any observational study relies on the sample size, which is an important limiting factor in finding modest effect size associations. We believe that for most analyses presented in this study, the amount of mutational data has been sufficient to sustain our claims, however, some analysis might improve substantially with an increased sample size. The analysis related to the detection of mutational clusters, which relies on a rare event (clustering) and in particular analysis of three-dimensional trans-clusters, which draws on narrowly sized loci (loop anchors) would benefit from an increase in the amount of tumor WGS data available. The scarcity of these mutational events represents a challenge in the dissection and detection of global trends evident in rare events or only in particular loci. Also, more generally, mutational signature deconvolution benefits from increased samples sizes when detecting less common mutagenic processes[23] . New tumor sequencing projects have recently increased the amount of sequenced tumor samples at a rapid pace, allowing future studies related to this work to overcome the aforementioned statistical power limitations.

Another important methodological challenge for any analysis in cancer genomics is the integration of multiple sources of epigenetic data, and matching the cell type to the cell type that generated the tumor (which may, in many cases, not be known). Because of their role in modulating mutagenesis and interacting with DNA repair pathways[97,189,192,197] , the integration of histone marks, DNA replication time measurements and DNA methylation to model local mutation rates is a significant feature of this work. While mutations are extracted from tumor biopsies, the epigenetic information is normally obtained in bulk from cell line experiments (either cancerous cells but also primary cells, or ESC/IPSC) experiments; intact tissue epigenome data exists but is very rarely from cancerous tissues and almost always contains a mix of cell types, which is suboptimal. This complexity generates an inconsistency where the epigenomic data is not necessarily well matched to the corresponding tumor cell type of origin. In the future, however, the fast-paced development of single-cell epigenetic techniques, i.e. scATAC-seq, together with the improvement in accuracy for whole genome/exome sequencing ([339]) to determine mutation patterns at the single cell level will represent a solution to this issue of matching mutational and epigenomic data to establish correlations better.

Globally, the work presented in this thesis deepens our understanding of the local mutation rate variation in the somatic human genome and highlights the functional impact potential of the presented mutational mechanisms.

Chapter 9

Conclusions

- A methodology for trinucleotide-aware mutation randomization, combined with a definition of the local False Discovery Rate, was developed and applied to human tumor genomes to robustly detect mutation clusters.
- The accumulation of diffuse and short mutation clusters, which we named *omikli* or *mutation fog* , is the result of a previously poorly characterized clustered mutagenic process, associated with APOBEC3A mutagenesis.
- The activity of the DNA mismatch repair pathway is a source of the *omikli* mutations. This mechanism is responsible for approximately two-thirds of the unclustered APOBEC mutation burden in human tumors.
- The association with MMR drives the generated APOBEC mutations towards early replicating domains of the genome, where the majority of active genes reside. The expected functional impact potential of this mechanism exceeds that of mutagenesis by UV damage and tobacco smoking in an average affected tumor.
- Mutations in healthy tissues can be reliably extracted from single clone *in vitro* expansion and can be used to model mutagenic processes using non-negative matrix factorization.
- The extraction of mutational processes in healthy clones reveal a basal mutational spectrum common in multiple human tissues and additionally a set of tissue-specific processes, some of which correlate with known exposures.
- The extraction of mutational signatures in a combined analysis of healthy and tumor samples of the same tissues suggests a remarkable consistency. Thus the tumor mutation spectrum can reveal the cell type of origin of a

given cancer subtype.

- A subtype of the kidney tubule cells with distinct mutational patterns observed in healthy kidney cell clones, tentatively labelled “KT2”, may be the cell-of-origin for the commonly occurring kidney cancers: clear cell and papillary renal cell carcinomas.
- A diverse set of mutational processes show a strongly reduced activity in unmethylated short segments in DNA (UMRs), which are commonly observed in promoters, some enhancers, chromatin loop anchors, and elsewhere. In particular, aging-associated signature 1, DNA repair deficiency signatures 10 and 15 and to a lesser extent, UV DNA damage associated signature 7 show significant hypomutation at UMRs.
- Unmethylated DNA segments also show an increase of mutagenic processes that derive from the activity of APOBEC and AID cytosine deaminases. The enrichment associates with the methylation status and/or the co-occurence with other functional elements.
- DNA methylation profiles provide an informative clustering of human genes, revealing epigenomically relevant groups. These gene groups present differential gradients in the mutation rates along their gene body, plausibly due to hypomethylation associated with intragenic enhancers and with polycomb histone marks.
- Taking into account these gene groups with variable intra-gene mutation rate gradients can better estimate the baseline mutation rate, aiding in the identification and detection of selection in genic regions and potentially promoters.
- Chromatin loop anchors (sites with high density of contacts in 3D genomic experiments) represent another coldspot of mutagenesis, and are protected from multiple but not all processes, in particular, from aging-associated signature 1 and UV damage signature 7.
- The genomic characterization of anchor sites reveals that multiple overlapping molecular features modulate this reduction of mutation rate. For signature 1, the DNA hypomethylation is the most plausible mechanism, while for signature 7 the combination of chromatin accessibility (DHS) and transcription may be the causal factor.
- The correct expectation baseline models of mutation rates at anchors can be used to identify three-dimensional mutation clusters (trans-clusters), consisting of pairs of mutations occurring in distal but interacting regions.
- The AID signature is enriched in three-dimensional mutation clusters for B

lymphocytes, suggesting that the AID may act at foci in 3D space, targeting interacting loci generating groups of mutation in a single event.

Bibliography

1. Vries, Hugo de. *Die Mutationstheorie* (1903).
2. Muller, H. J. Artificial Transmutation of the Gene. *Science* **66,** 84–87. ISSN: 0036-8075 (July 1927).
3. Weismann, A. *Das Keimplasma* (1892).
4. Fisher, R. A. *The Genetical Theory of Natural Selection* ISBN: 978-1-176-62502-0 (1930).
5. Katju, V. & Bergthorsson, U. Old Trade, New Tricks: Insights into the Spontaneous Mutation Process from the Partnering of Classical Mutation Accumulation Experiments with High-Throughput Genomic Approaches. *Genome Biology and Evolution* **11,** 136–165. ISSN: 1759-6653 (Jan. 2019).
6. Lujan, S. A. & Kunkel, T. A. Stability across the Whole Nuclear Genome in the Presence and Absence of DNA Mismatch Repair. *Cells* **10,** 1224 (May 2021).
7. Krontiris, T. G. & Cooper, G. M. Transforming Activity of Human Tumor DNAs. *Proc Natl Acad Sci U S A* **78,** 1181–1184. ISSN: 0027-8424 (Feb. 1981).
8. Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming Genes of Carcinomas and Neuroblastomas Introduced into Mouse Fibroblasts. *Nature* **290,** 261–264. ISSN: 0028-0836 (Mar. 1981).
9. Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene. *Nature* **300,** 149–152. ISSN: 0028-0836 (Nov. 1982).
10. Tabin, C. J. *et al.* Mechanism of Activation of a Human Oncogene. *Nature* **300,** 143–149. ISSN: 1476-4687 (Nov. 1982).
11. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The Cancer Genome. *Nature* **458,** 719–724. ISSN: 1476-4687 (Apr. 2009).

12. Duncan, B. K. & Miller, J. H. Mutagenic Deamination of Cytosine Residues in DNA. *Nature* **287,** 560–561. ISSN: 1476-4687 (Oct. 1980).

13. Shen, J. C., Rideout, W. M. & Jones, P. A. High Frequency Mutagenesis by a DNA Methyltransferase. *Cell* **71,** 1073–1080. ISSN: 0092-8674 (Dec. 1992).

14. Dzantiev, L. *et al.* A Defined Human System That Supports Bidirectional Mismatch-Provoked Excision. *Mol Cell* **15,** 31–41. ISSN: 1097-2765 (July 2004).

15. Mu, D. *et al.* Reconstitution of Human DNA Repair Excision Nuclease in a Highly Defined System. *J Biol Chem* **270,** 2415–2418. ISSN: 0021-9258 (Feb. 1995).

16. Lindahl, T. Instability and Decay of the Primary Structure of DNA. *Nature* **362,** 709–715. ISSN: 1476-4687 (Apr. 1993).

17. Modrich, P. Mechanisms in Eukaryotic Mismatch Repair. *Journal of Biological Chemistry* **281,** 30305–30309. ISSN: 0021-9258 (Oct. 2006).

18. Sancar, A. Mechanisms of DNA Excision Repair. *Science* **266,** 1954–1956. ISSN: 0036-8075 (Dec. 1994).

19. McLendon, R. *et al.* Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways. *Nature* **455,** 1061–1068. ISSN: 1476-4687 (Oct. 2008).

20. Priestley, P. *et al.* Pan-Cancer Whole-Genome Analyses of Metastatic Solid Tumours. *Nature* **575,** 210–216. ISSN: 1476-4687 (Nov. 2019).

21. Pleasance, E. *et al.* Pan-Cancer Analysis of Advanced Patient Tumors Reveals Interactions between Therapy and Genomic Landscapes. *Nat Cancer* **1,** 452–468. ISSN: 2662-1347 (Apr. 2020).

22. Campbell, P. J. *et al.* Pan-Cancer Analysis of Whole Genomes. *Nature* **578,** 82–93. ISSN: 1476-4687 (Feb. 2020).

23. Degasperi, A. *et al.* Substitution Mutational Signatures in Whole-Genome–Sequenced Cancers in the UK Population. *Science* **376,** abl9283 (Apr. 2022).

24. International HapMap Consortium *et al.* A Second Generation Human Haplotype Map of over 3.1 Million SNPs. *Nature* **449,** 851–861. ISSN: 1476-4687 (Oct. 2007).

25. Taliun, D. *et al.* Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program. *Nature* **590,** 290–299. ISSN: 1476-4687 (Feb. 2021).

26. Auton, A. *et al.* A Global Reference for Human Genetic Variation. *Nature* **526,** 68–74. ISSN: 1476-4687 (Oct. 2015).

27. Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The International Genome Sample Resource (IGSR) Collection of Open Human Genomic Variation Resources. *Nucleic Acids Research* **48,** D941–D947. ISSN: 0305-1048 (Jan. 2020).

28. Halldorsson, B. V. *et al.* The Sequences of 150,119 Genomes in the UK Biobank. *Nature* **607,** 732–740. ISSN: 1476-4687 (July 2022).

29. On Cancer (IARC), T. I. A. f. R. *Global Cancer Observatory* https://gco.iarc.fr/.

30. *Surveillance, Epidemiology, and End Results Program* https://seer.cancer.gov.

31. *Spanish Network of Cancer Registries* https://redecan.org.

32. FDA. *FDA Approves First-Line Immunotherapy for Patients with MSI-H/dMMR Metastatic Colorectal Cancer* https://www.fda.gov/news-events/press-announcements/fda-approves-first-line-immunotherapy-patients-msi-hdmmr-metastatic-colorectal-cancer. Tue, 06/30/2020 - 15:31.

33. Herberts, C. *et al.* Deep Whole-Genome ctDNA Chronology of Treatment-Resistant Prostate Cancer. *Nature* **608,** 199–208. ISSN: 1476-4687 (Aug. 2022).

34. Katsman, E. *et al.* Detecting Cell-of-Origin and Cancer-Specific Methylation Features of Cell-Free DNA from Nanopore Sequencing. *Genome Biology* **23,** 158. ISSN: 1474-760X (July 2022).

35. Kurtz, D. M. *et al.* Enhanced Detection of Minimal Residual Disease by Targeted Sequencing of Phased Variants in Circulating Tumor DNA. *Nat Biotechnol* **39,** 1537–1547. ISSN: 1546-1696 (Dec. 2021).

36. Ulz, P. *et al.* Inference of Transcription Factor Binding from Cell-Free DNA Enables Tumor Subtype Prediction and Early Detection. *Nat Commun* **10,** 4666. ISSN: 2041-1723 (Oct. 2019).

37. Abbosh, C. *et al.* Phylogenetic ctDNA Analysis Depicts Early-Stage Lung Cancer Evolution. *Nature* **545,** 446–451. ISSN: 1476-4687 (Apr. 2017).

38. Wan, J. C. M. *et al.* Genome-Wide Mutational Signatures in Low-Coverage Whole Genome Sequencing of Cell-Free DNA. *Nat Commun* **13,** 4953. ISSN: 2041-1723 (Aug. 2022).

39. Cohen, J. D. *et al.* Detection and Localization of Surgically Resectable Cancers with a Multi-Analyte Blood Test. *Science,* eaar3247. ISSN: 0036-8075, 1095-9203 (Jan. 2018).

40. Seluanov, A., Gladyshev, V. N., Vijg, J. & Gorbunova, V. Mechanisms of Cancer Resistance in Long-Lived Mammals. *Nat Rev Cancer* **18,** 433–441. ISSN: 1474-1768 (July 2018).

41. Leroi, A. M., Koufopanou, V. & Burt, A. Cancer Selection. *Nat Rev Cancer* **3,** 226–231. ISSN: 1474-1768 (Mar. 2003).

42. Tollis, M., Boddy, A. M. & Maley, C. C. Peto's Paradox: How Has Evolution Solved the Problem of Cancer Prevention? *BMC Biology* **15,** 60. ISSN: 1741-7007 (July 2017).

43. Vincze, O. *et al.* Cancer Risk across Mammals. *Nature* **601,** 263–267. ISSN: 1476-4687 (Jan. 2022).

44. Peto, R., Roe, F. J., Lee, P. N., Levy, L. & Clack, J. Cancer and Ageing in Mice and Men. *Br J Cancer* **32,** 411–426. ISSN: 0007-0920 (Oct. 1975).

45. Peto, R. Epidemiology, Multistage Models, and Short-Term Mutagenicity Tests 1. *International Journal of Epidemiology* **45,** 621–637. ISSN: 0300-5771 (June 2016).

46. Cagan, A. *et al.* Somatic Mutation Rates Scale with Lifespan across Mammals. *Nature* **604,** 517–524. ISSN: 1476-4687 (Apr. 2022).

47. Chapman, M. A. *et al.* Initial Genome Sequencing and Analysis of Multiple Myeloma. *Nature* **471,** 467–472. ISSN: 1476-4687 (Mar. 2011).

48. Lawrence, M. S. *et al.* Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes. *Nature* **499,** 214–218. ISSN: 0028-0836 (July 2013).

49. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171,** 1029–1041.e21. ISSN: 0092-8674, 1097-4172 (Nov. 2017).

50. McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. Impact of Deleterious Passenger Mutations on Cancer Progression. *Proceedings of the National Academy of Sciences* **110,** 2910–2915 (Feb. 2013).

51. Tilk, S., Curtis, C., Petrov, D. A. & McFarland, C. D. Most Cancers Carry a Substantial Deleterious Load Due to Hill-Robertson Interference. *bioRxiv,* 764340 (Feb. 2021).

52. Sweet-Cordero, E. A. & Biegel, J. A. The Genomic Landscape of Pediatric Cancers: Implications for Diagnosis and Treatment. *Science* **363,** 1170–1175 (Mar. 2019).

53. McConnell, M. J. *et al.* Mosaic Copy Number Variation in Human Neurons. *Science* **342,** 632–637 (Nov. 2013).

54. Cai, X. *et al.* Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Reports* **8,** 1280–1289. ISSN: 2211-1247 (Sept. 2014).

55. Evrony, G. D. *et al.* Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. *Cell* **151,** 483–496. ISSN: 0092-8674, 1097-4172 (Oct. 2012).

56. Baillie, J. K. *et al.* Somatic Retrotransposition Alters the Genetic Landscape of the Human Brain. *Nature* **479,** 534–537. ISSN: 1476-4687 (Nov. 2011).

57. Lodato, M. A. *et al.* Aging and Neurodegeneration Are Associated with Increased Mutations in Single Human Neurons. *Science* **359,** 555–559. ISSN: 1095-9203 (Feb. 2018).

58. Miller, M. B. *et al.* Somatic Genomic Changes in Single Alzheimer's Disease Neurons. *Nature* **604,** 714–722. ISSN: 1476-4687 (Apr. 2022).

59. Martincorena, I. *et al.* High Burden and Pervasive Positive Selection of Somatic Mutations in Normal Human Skin. *Science* **348,** 880–886 (May 2015).

60. Abyzov, A. *et al.* Somatic Copy Number Mosaicism in Human Skin Revealed by Induced Pluripotent Stem Cells. *Nature* **492,** 438–442. ISSN: 1476-4687 (Dec. 2012).

61. Martincorena, I. *et al.* Somatic Mutant Clones Colonize the Human Esophagus with Age. *Science* **362,** 911–917. ISSN: 0036-8075, 1095-9203 (Nov. 2018).

62. Kakiuchi, N. & Ogawa, S. Clonal Expansion in Non-Cancer Tissues. *Nat Rev Cancer* **21,** 239–256. ISSN: 1474-1768 (Apr. 2021).

63. Yokoyama, A. *et al.* Age-Related Remodelling of Oesophageal Epithelia by Mutated Cancer Drivers. *Nature* **565,** 312. ISSN: 1476-4687 (Jan. 2019).

64. Franco, I. *et al.* Somatic Mutagenesis in Satellite Cells Associates with Human Skeletal Muscle Aging. *Nature Communications* **9.** ISSN: 2041-1723 (Dec. 2018).

65. Franco, I. *et al.* Whole Genome DNA Sequencing Provides an Atlas of Somatic Mutagenesis in Healthy Human Cells and Identifies a Tumor-Prone Cell Type. *Genome Biol.* **20,** 285. ISSN: 1474-760X (Dec. 2019).

66. Lee-Six, H. *et al.* The Landscape of Somatic Mutation in Normal Colorectal Epithelial Cells. *Nature* **574,** 532–537. ISSN: 1476-4687 (Oct. 2019).

67. Wijewardhane, N., Dressler, L. & Ciccarelli, F. D. Normal Somatic Mutations in Cancer Transformation. *Cancer Cell* **39,** 125–129. ISSN: 1535-6108 (Feb. 2021).

68. Colom, B. *et al.* Mutant Clones in Normal Epithelium Outcompete and Eliminate Emerging Tumours. *Nature* **598,** 510–514. ISSN: 1476-4687 (Oct. 2021).

69. Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in Health and Disease — Clones Picking up Speed. *Nat Rev Genet* **18,** 128–142. ISSN: 1471-0064 (Feb. 2017).

70. Cibulskis, K. *et al.* Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples. *Nat Biotechnol* **31,** 213–219. ISSN: 1546-1696 (Mar. 2013).

71. Schmitt, M. W. *et al.* Detection of Ultra-Rare Mutations by next-Generation Sequencing. *Proceedings of the National Academy of Sciences* **109,** 14508–14513 (Sept. 2012).

72. Kennedy, S. R. *et al.* Detecting Ultralow-Frequency Mutations by Duplex Sequencing. *Nat Protoc* **9,** 2586–2606. ISSN: 1750-2799 (Nov. 2014).

73. Hoang, M. L. *et al.* Genome-Wide Quantification of Rare Somatic Mutations in Normal Human Tissues Using Massively Parallel Sequencing. *Proceedings of the National Academy of Sciences* **113,** 9846–9851 (Aug. 2016).

74. Abascal, F. *et al.* Somatic Mutation Landscapes at Single-Molecule Resolution. *Nature* **593,** 405–410. ISSN: 1476-4687 (May 2021).

75. Robinson, P. S. *et al.* Inherited MUTYH Mutations Cause Elevated Somatic Mutation Rates and Distinctive Mutational Signatures in Normal Human Cells. *Nat Commun* **13,** 3949. ISSN: 2041-1723 (July 2022).

76. Robinson, P. S. *et al.* Increased Somatic Mutation Burdens in Normal Human Cells Due to Defective DNA Polymerases. *Nat Genet* **53,** 1434–1442. ISSN: 1546-1718 (Oct. 2021).

77. Moore, L. *et al.* The Mutational Landscape of Human Somatic and Germline Cells. *Nature* **597,** 381–386. ISSN: 1476-4687 (Sept. 2021).

78. Lee, B. C. H. *et al.* Mutational Landscape of Normal Epithelial Cells in Lynch Syndrome Patients. *Nat Commun* **13,** 2710. ISSN: 2041-1723 (May 2022).

79. Franco, I., Revêchon, G. & Eriksson, M. Challenges of Proving a Causal Role of Somatic Mutations in the Aging Process. *Aging Cell* **21,** e13613. ISSN: 1474-9726 (May 2022).

80. Martincorena, I. & Campbell, P. J. Somatic Mutation in Cancer and Normal Cells. *Science* **349,** 1483–1489. ISSN: 0036-8075, 1095-9203 (Sept. 2015).

81. Zou, X. *et al.* Validating the Concept of Mutational Signatures with Isogenic Cell Models. *Nature Communications* **9,** 1–16. ISSN: 2041-1723 (May 2018).

82. Szilard, L. On the Nature of the Aging Process. *Proc Natl Acad Sci U S A* **45,** 30–45. ISSN: 0027-8424 (Jan. 1959).

83. Hamilton, W. D. The Moulding of Senescence by Natural Selection. *Journal of Theoretical Biology* **12,** 12–45. ISSN: 0022-5193 (Sept. 1966).

84. Muñoz-Espín, D. & Serrano, M. Cellular Senescence: From Physiology to Pathology. *Nat Rev Mol Cell Biol* **15,** 482–496. ISSN: 1471-0080 (July 2014).

85. Horvath, S. DNA Methylation Age of Human Tissues and Cell Types. *Genome Biology* **14,** 3156. ISSN: 1474-760X (Dec. 2013).

86. Sano, S. *et al.* Hematopoietic Loss of Y Chromosome Leads to Cardiac Fibrosis and Heart Failure Mortality. *Science* **377,** 292–297 (July 2022).

87. Rieckher, M., Garinis, G. A. & Schumacher, B. Molecular Pathology of Rare Progeroid Diseases. *Trends Mol Med* **27,** 907–922. ISSN: 1471-499X (Sept. 2021).

88. 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. ISSN: 1546-1718 (Mar. 2015).

89. Mas-Ponte, D., McCullough, M. & Supek, F. Spectrum of DNA Mismatch Repair Failures Viewed through the Lens of Cancer Genomics and Implications for Therapy. *Clinical Science* **136,** 383–404. ISSN: 0143-5221 (Mar. 2022).

90. Alexandrov, L. B. *et al.* Signatures of Mutational Processes in Human Cancer. *Nature* **500,** 415–421. ISSN: 0028-0836 (Aug. 2013).

91. Alexandrov, L. B. *et al.* Clock-like Mutational Processes in Human Somatic Cells. *Nature Genetics* **47,** 1402–1407. ISSN: 1546-1718 (Dec. 2015).

92. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153,** 1194–1217. ISSN: 0092-8674 (June 2013).

93. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational Signatures: Emerging Concepts, Caveats and Clinical Applications. *Nat Rev Cancer* **21,** 619–637. ISSN: 1474-1768 (Oct. 2021).

94. Besaratinia, A. *et al.* DNA Lesions Induced by UV A1 and B Radiation in Human Cells: Comparative Analyses in the Overall Genome and in the P53 Tumor Suppressor Gene. *Proceedings of the National Academy of Sciences* **102,** 10058–10063 (July 2005).

95. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149,** 979–993. ISSN: 00928674 (May 2012).

96. Alexandrov, L. B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *Nature* **578,** 94–101. ISSN: 1476-4687 (Feb. 2020).

97. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal That Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170,** 534–547.e23. ISSN: 00928674 (July 2017).

98. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164,** 538–549. ISSN: 0092-8674, 1097-4172 (Jan. 2016).

99. Vöhringer, H., Hoeck, A. V., Cuppen, E. & Gerstung, M. Learning Mutational Signatures and Their Multidimensional Genomic Properties with TensorSignatures. *Nat Commun* **12,** 3628. ISSN: 2041-1723 (June 2021).

100. Morganella, S. *et al.* The Topography of Mutational Processes in Breast Cancer Genomes. *Nature Communications* **7,** 11383. ISSN: 2041-1723 (May 2016).

101. Wojtowicz, D. *et al.* Hidden Markov Models Lead to Higher Resolution Maps of Mutation Signature Activity in Cancer. *Genome Medicine* **11,** 49. ISSN: 1756-994X (July 2019).

102. Bergstrom, E. N. *et al.* Mapping Clustered Mutations in Cancer Reveals APOBEC3 Mutagenesis of ecDNA. *Nature* **602,** 510–517. ISSN: 1476-4687 (Feb. 2022).

103. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177,** 821–836.e16. ISSN: 00928674 (May 2019).

104. Nik-Zainal, S. *et al.* Landscape of Somatic Mutations in 560 Breast Cancer Whole-Genome Sequences. *Nature* **534,** 47–54. ISSN: 1476-4687 (June 2016).

105. Li, Y. *et al.* Patterns of Somatic Structural Variation in Human Cancer Genomes. *Nature* **578,** 112–121. ISSN: 1476-4687 (Feb. 2020).

106. Macintyre, G. *et al.* Copy Number Signatures and Mutational Processes in Ovarian Carcinoma. *Nature Genetics* **50,** 1262–1270. ISSN: 1546-1718 (Sept. 2018).

107. Drews, R. M. *et al.* A Pan-Cancer Compendium of Chromosomal Instability. *Nature* **606,** 976–983. ISSN: 1476-4687 (June 2022).

108. Steele, C. D. *et al.* Undifferentiated Sarcomas Develop through Distinct Evolutionary Pathways. *Cancer Cell* **35,** 441–456.e8. ISSN: 1878-3686 (Mar. 2019).

109. Steele, C. D. *et al.* Signatures of Copy Number Alterations in Human Cancer. *Nature* **606,** 984–991. ISSN: 1476-4687 (June 2022).

110. Wang, S. *et al.* Copy Number Signature Analysis Tool and Its Application in Prostate Cancer Reveals Distinct Mutational Processes and Clinical Outcomes. *PLoS Genet* **17,** e1009557. ISSN: 1553-7404 (May 2021).

111. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3,** 246–259. ISSN: 2211-1247 (Jan. 2013).

112. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *PNAS* **101,** 4164–4169. ISSN: 0027-8424, 1091-6490 (Mar. 2004).

113. Islam, S. M. A. *et al.* Uncovering Novel Mutational Signatures by de Novo Extraction with SigProfilerExtractor. *bioRxiv,* 2020.12.13.422570 (Dec. 2020).

114. Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: An Empirical Bayesian Approach to Mutational Signature Discovery. *Bioinformatics* **33,** 8–16. ISSN: 1367-4803 (Jan. 2017).

115. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: Probabilistic Inference of Mutational Processes and Their Localization in the Cancer Genome. *Genome Biology* **14,** R39. ISSN: 1474-760X (Apr. 2013).

116. Funnell, T. *et al.* Integrated Structural Variation and Point Mutation Signatures in Cancer Genomes Using Correlated Topic Models. *PLOS Computational Biology* **15,** e1006799. ISSN: 1553-7358 (Feb. 2019).

117. Vali-Pour, M., Lehner, B. & Supek, F. The Impact of Rare Germline Variants on Human Somatic Mutation Processes. *Nat Commun* **13,** 3724. ISSN: 2041-1723 (June 2022).

118. Degasperi, A. *et al.* A Practical Framework and Online Tool for Mutational Signature Analyses Show Intertissue Variation and Driver Dependencies. *Nat Cancer* **1,** 249–263. ISSN: 2662-1347 (Feb. 2020).

119. Maura, F. *et al.* A Practical Guide for Mutational Signature Analysis in Hematological Malignancies. *Nat Commun* **10,** 2969. ISSN: 2041-1723 (Dec. 2019).

120. Li, S., Crawford, F. W. & Gerstein, M. B. Using sigLASSO to Optimize Cancer Mutation Signatures Jointly with Sampling Likelihood. *Nat Commun* **11,** 3575. ISSN: 2041-1723 (July 2020).

121. Kasar, S. *et al.* Whole-Genome Sequencing Reveals Activation-Induced Cytidine Deaminase Signatures during Indolent Chronic Lymphocytic Leukaemia Evolution. *Nat Commun* **6,** 8866. ISSN: 2041-1723 (Dec. 2015).

122. Boot, A. *et al.* In-Depth Characterization of the Cisplatin Mutational Signature in Human Cell Lines and in Esophageal and Liver Tumors. *Genome Res* **28,** 654–665. ISSN: 1088-9051 (May 2018).

123. Pich, O. *et al.* The Mutational Footprints of Cancer Therapies. *Nature Genetics* **51,** 1732–1740. ISSN: 1546-1718 (Dec. 2019).

124. Chan, K. *et al.* An APOBEC3A Hypermutation Signature Is Distinguishable from the Signature of Background Mutagenesis by APOBEC3B in Human Cancers. *Nature Genetics* **47,** ng.3378. ISSN: 1546-1718 (Aug. 2015).

125. Hodel, K. P. *et al.* Explosive Mutation Accumulation Triggered by Heterozygous Human Pol ϵ Proofreading-Deficiency Is Driven by Suppression of Mismatch Repair. *eLife* **7** (ed van Oijen, A. M.) e32692. ISSN: 2050-084X (Feb. 2018).

126. Hodel, K. P. *et al.* POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. *Molecular Cell* **78,** 1166–1177.e6. ISSN: 1097-2765 (June 2020).

127. Zou, X. *et al.* A Systematic CRISPR Screen Defines Mutational Mechanisms Underpinning Signatures Caused by Replication Errors and Endogenous DNA Damage. *Nat Cancer* **2,** 643–657. ISSN: 2662-1347 (June 2021).

128. Meier, B. *et al.* Mutational Signatures of DNA Mismatch Repair Deficiency in C. Elegans and Human Cancers. *Genome Res.* **28,** 666–675. ISSN: 1088-9051, 1549-5469 (May 2018).

129. Volkova, N. V. *et al.* Mutational Signatures Are Jointly Shaped by DNA Damage and Repair. *Nat Commun* **11,** 2169. ISSN: 2041-1723 (May 2020).

130. Zámborszky, J. *et al.* Loss of BRCA1 or BRCA2 Markedly Increases the Rate of Base Substitution Mutagenesis and Has Distinct Effects on Genomic Deletions. *Oncogene* **36,** 746–755. ISSN: 1476-5594 (Feb. 2017).

131. Póti, Á., Szikriszt, B., Gervai, J. Z., Chen, D. & Szüts, D. Characterisation of the Spectrum and Genetic Dependence of Collateral Mutations Induced by Translesion DNA Synthesis. *PLOS Genetics* **18,** e1010051. ISSN: 1553-7404 (Feb. 2022).

132. Nik-Zainal, S. *et al.* The Genome as a Record of Environmental Exposure. *Mutagenesis* **30,** 763–770. ISSN: 0267-8357 (Nov. 2015).

133. Szikriszt, B. *et al.* A Comprehensive Survey of the Mutagenic Impact of Common Cancer Cytotoxics. *Genome Biology* **17,** 99. ISSN: 1474-760X (May 2016).

134. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176,** 1282–1294.e20. ISSN: 00928674 (Mar. 2019).

135. Middlebrooks, C. D. *et al.* Association of Germline Variants in the *APOBEC3* Region with Cancer Risk and Enrichment with APOBEC-signature Mutations in Tumors. *Nature Genetics* **48,** 1330–1338. ISSN: 1546-1718 (Nov. 2016).

136. Nik-Zainal, S. *et al.* Association of a Germline Copy Number Polymorphism of *APOBEC3A* and *APOBEC3B* with Burden of Putative APOBEC-dependent Mutations in Breast Cancer. *Nature Genetics* **46,** 487–491. ISSN: 1546-1718 (May 2014).

137. Secrier, M. *et al.* Mutational Signatures in Esophageal Adenocarcinoma Define Etiologically Distinct Subgroups with Therapeutic Relevance. *Nat Genet* **48,** 1131–1141. ISSN: 1546-1718 (Oct. 2016).

138. 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. ISSN: 1546-1718 (May 2013).

139. Christensen, S. *et al.* 5-Fluorouracil Treatment Induces Characteristic T>G Mutations in Human Cancer. *Nat Commun* **10,** 4571. ISSN: 2041-1723 (Oct. 2019).

140. Martínez-Jiménez, F. *et al.* Pan-Cancer Whole Genome Comparison of Primary and Metastatic Solid Tumors. *bioRxiv,* 2022.06.17.496528 (June 2022).

141. Kim, J. *et al.* Somatic ERCC2 Mutations Are Associated with a Distinct Genomic Signature in Urothelial Tumors. *Nat Genet* **48,** 600–606. ISSN: 1546-1718 (June 2016).

142. Rahbari, R. *et al.* Timing, Rates and Spectra of Human Germline Mutation. *Nat Genet* **48,** 126–133. ISSN: 1546-1718 (Feb. 2016).

143. Petljak, M. *et al.* Mechanisms of APOBEC3 Mutagenesis in Human Cancer Cells. *Nature* **607,** 799–807. ISSN: 1476-4687 (July 2022).

144. FDA. *FDA Grants Accelerated Approval to Pembrolizumab for First Tissue/Site Agnostic Indication* https://www.fda.gov/drugs/resources-information-approved-drugs/fda-grants-accelerated-approval-pembrolizumab-first-tissuesite-agnostic-indication. 2017.

145. FDA. *FDA Approves Pembrolizumab for Adults and Children with TMB-H Solid Tumors* https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-adults-and-children-tmb-h-solid-tumors. 2020.

146. Ma, X. *et al.* Functional Landscapes of POLE and POLD1 Mutations in Checkpoint Blockade-Dependent Antitumor Immunity. *Nat Genet* **54,** 996–1012. ISSN: 1546-1718 (July 2022).

147. Stratton, M. R. Exploring the Genomes of Cancer Cells: Progress and Promise. *Science* **331,** 1553–1558. ISSN: 1095-9203 (Mar. 2011).

148. Davies, H. *et al.* HRDetect Is a Predictor of *BRCA1* and *BRCA2* Deficiency Based on Mutational Signatures. *Nature Medicine* **23,** 517–525. ISSN: 1546-170X (Apr. 2017).

149. Levatić, J., Salvadores, M., Fuster-Tormo, F. & Supek, F. Mutational Signatures Are Markers of Drug Sensitivity of Cancer Cells. *Nat Commun* **13,** 2926. ISSN: 2041-1723 (May 2022).

150. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger Mutations Accurately Classify Human Tumors. *PLOS Computational Biology* **15,** e1006953. ISSN: 1553-7358 (Apr. 2019).

151. Jiao, W. *et al.* A Deep Learning System Accurately Classifies Primary and Metastatic Cancers Using Passenger Mutation Patterns. *Nat Commun* **11,** 728. ISSN: 2041-1723 (Feb. 2020).

152. Hsieh, P. & Yamane, K. DNA Mismatch Repair: Molecular Mechanism, Cancer, and Ageing. *Mechanisms of Ageing and Development. DNA Damage, Repair, Ageing and Age-Related Disease* **129,** 391–407. ISSN: 0047-6374 (July 2008).

153. Iyer, R. R., Pluciennik, A., Burdett, V. & Modrich, P. L. DNA Mismatch Repair: Functions and Mechanisms. *Chem. Rev.* **106,** 302–323. ISSN: 0009-2665 (Feb. 2006).

154. Kunkel, T. A. & Erie, D. A. DNA Mismatch Repair. *Annu Rev Biochem* **74,** 681–710. ISSN: 0066-4154 (2005).

155. Sanders, M. A. *et al.* Life without Mismatch Repair. *bioRxiv,* 2021.04.14.437578 (Apr. 2021).

156. Wei, K. *et al.* Inactivation of Exonuclease 1 in Mice Results in DNA Mismatch Repair Defects, Increased Cancer Susceptibility, and Male and Female Sterility. *Genes Dev* **17,** 603–614. ISSN: 0890-9369 (Mar. 2003).

157. Iyer, R. R. *et al.* The MutSα-Proliferating Cell Nuclear Antigen Interaction in Human DNA Mismatch Repair. *J Biol Chem* **283,** 13310–13319. ISSN: 0021-9258 (May 2008).

158. Jiricny, J. The Multifaceted Mismatch-Repair System. *Nat Rev Mol Cell Biol* **7,** 335–346. ISSN: 1471-0072 (May 2006).

159. Lynch, H. T. *et al.* Phenotypic and Genotypic Heterogeneity in the Lynch Syndrome: Diagnostic, Surveillance and Management Implications. *Eur J Hum Genet* **14,** 390–402. ISSN: 1476-5438 (Apr. 2006).

160. de la Chapelle, A. Genetic Predisposition to Colorectal Cancer. *Nat Rev Cancer* **4,** 769–780. ISSN: 1474-1768 (Oct. 2004).

161. Schubert, S. A., Morreau, H., de Miranda, N. F. C. C. & van Wezel, T. The Missing Heritability of Familial Colorectal Cancer. *Mutagenesis* **35,** 221–231. ISSN: 0267-8357 (July 2020).

162. Bellido, F. *et al.* POLE and POLD1 Mutations in 529 Kindred with Familial Colorectal Cancer and/or Polyposis: Review of Reported Cases and Recommendations for Genetic Testing and Surveillance. *Genet Med* **18,** 325–332. ISSN: 1530-0366 (Apr. 2016).

163. Buchanan, D. D., Rosty, C., Clendenning, M., Spurdle, A. B. & Win, A. K. Clinical Problems of Colorectal Cancer and Endometrial Cancer Cases with Unknown Cause of Tumor Mismatch Repair Deficiency (Suspected Lynch Syndrome). *Appl Clin Genet* **7,** 183–193. ISSN: 1178-704X (Oct. 2014).

164. Bucksch, K. *et al.* Cancer Risks in Lynch Syndrome, Lynch-like Syndrome, and Familial Colorectal Cancer Type X: A Prospective Cohort Study. *BMC Cancer* **20,** 460. ISSN: 1471-2407 (May 2020).

165. Rodríguez–Soler, M. *et al.* Risk of Cancer in Cases of Suspected Lynch Syndrome Without Germline Mutation. *Gastroenterology* **144,** 926–932.e1. ISSN: 0016-5085 (May 2013).

166. Hemminki, A. *et al.* Loss of the Wild Type MLH1 Gene Is a Feature of Hereditary Nonpolyposis Colorectal Cancer. *Nat Genet* **8,** 405–410. ISSN: 1546-1718 (Dec. 1994).

167. Liu, B. *et al.* Mismatch Repair Gene Defects in Sporadic Colorectal Cancers with Microsatellite Instability. *Nat Genet* **9,** 48–55. ISSN: 1546-1718 (Jan. 1995).

168. Kane, M. F. *et al.* Methylation of the hMLH1 Promoter Correlates with Lack of Expression of hMLH1 in Sporadic Colon Tumors and Mismatch Repair-Defective Human Tumor Cell Lines. *Cancer Res* **57,** 808–811. ISSN: 0008-5472 (Mar. 1997).

169. Chung, J. *et al.* DNA Polymerase and Mismatch Repair Exert Distinct Microsatellite Instability Signatures in Normal and Malignant Human Cells. *Cancer Discov* **11,** 1176–1191. ISSN: 2159-8274, 2159-8290 (May 2021).

170. Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M. & Park, P. J. A Molecular Portrait of Microsatellite Instability across Multiple Cancers. *Nature Communications* **8,** 15180. ISSN: 2041-1723 (June 2017).

171. Maruvka, Y. E. *et al.* Analysis of Somatic Microsatellite Indels Identifies Driver Events in Human Tumors. *Nature Biotechnology* **35,** 951–959. ISSN: 1546-1696 (Oct. 2017).

172. Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and Characterization of Microsatellite Instability across 18 Cancer Types. *Nature Medicine* **22,** 1342–1350. ISSN: 1546-170X (Nov. 2016).

173. Campbell, B. B. *et al.* Comprehensive Analysis of Hypermutation in Human Cancer. *Cell* **171,** 1042–1056.e10. ISSN: 0092-8674, 1097-4172 (Nov. 2017).

174. Polak, P. *et al.* Cell-of-Origin Chromatin Organization Shapes the Mutational Landscape of Cancer. *Nature* **518,** 360–364. ISSN: 0028-0836 (Feb. 2015).

175. Supek, F. & Lehner, B. Differential DNA Mismatch Repair Underlies Mutation Rate Variation across the Human Genome. *Nature* **521,** 81–84. ISSN: 0028-0836 (May 2015).

176. Fang, H. *et al.* Deficiency of Replication-Independent DNA Mismatch Repair Drives a 5-Methylcytosine Deamination Mutational Signature in Cancer. *Science Advances* **7,** eabg4398 (Nov. 2021).

177. Cleaver, J. E. Cancer in Xeroderma Pigmentosum and Related Disorders of DNA Repair. *Nat Rev Cancer* **5,** 564–573. ISSN: 1474-1768 (July 2005).

178. Zheng, C. L. *et al.* Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. *Cell Reports* **9,** 1228–1234. ISSN: 22111247 (Nov. 2014).

179. Palles, C. *et al.* Germline Mutations Affecting the Proofreading Domains of POLE and POLD1 Predispose to Colorectal Adenomas and Carcinomas. *Nat Genet* **45,** 136–144. ISSN: 1546-1718 (Feb. 2013).

180. Mur, P. *et al.* Role of POLE and POLD1 in Familial Cancer. *Genet Med* **22,** 2089–2100. ISSN: 1530-0366 (Dec. 2020).

181. Poulos, R. C., Olivier, J. & Wong, J. W. The Interaction between Cytosine Methylation and Processes of DNA Replication and Repair Shape the Mutational Landscape of Cancer Genomes. *Nucleic Acids Research* **45,** 7786–7795. ISSN: 0305-1048 (July 2017).

182. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational Signature Distribution Varies with DNA Replication Timing and Strand Asymmetry. *Genome Biology* **19,** 129. ISSN: 1474-760X (Dec. 2018).

183. Pilati, C. *et al.* Mutational Signature Analysis Identifies MUTYH Deficiency in Colorectal Cancers and Adrenocortical Carcinomas. *The Journal of Pathology* **242,** 10–15. ISSN: 1096-9896 (2017).

184. Drost, J. *et al.* Use of CRISPR-modified Human Stem Cell Organoids to Study the Origin of Mutational Signatures in Cancer. *Science* **358,** 234–238. ISSN: 0036-8075, 1095-9203 (Oct. 2017).

185. Roberts, S. A. *et al.* Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Molecular Cell* **46,** 424–435. ISSN: 1097-2765 (May 2012).

186. Álvarez-Prado, Á. F. *et al.* A Broad Atlas of Somatic Hypermutation Allows Prediction of Activation-Induced Deaminase Targets. *J Exp Med* **215,** 761–771. ISSN: 0022-1007 (Jan. 2018).

187. Supek, F. & Lehner, B. Scales and Mechanisms of Somatic Mutation Rate Variation across the Human Genome. *DNA Repair. Cutting-Edge Perspectives in Genomic Maintenance VI* **81,** 102647. ISSN: 1568-7864 (July 2019).

188. Hodgkinson, A. & Eyre-Walker, A. Variation in the Mutation Rate across Mammalian Genomes. *Nat Rev Genet* **12,** 756–766. ISSN: 1471-0064 (Nov. 2011).

189. Schuster-Böckler, B. & Lehner, B. Chromatin Organization Is a Major Influence on Regional Mutation Rates in Human Cancer Cells. *Nature* **488,** 504–507. ISSN: 1476-4687 (Aug. 2012).

190. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The Large-Scale Distribution of Somatic Mutations in Cancer Genomes. *Hum Mutat* **33,** 136–143. ISSN: 1098-1004 (Jan. 2012).

191. Woo, Y. H. & Li, W.-H. DNA Replication Timing and Selection Shape the Landscape of Nucleotide Variation in Cancer Genomes. *Nature Communications* **3,** 1004. ISSN: 2041-1723 (Aug. 2012).

192. Polak, P. *et al.* Reduced Local Mutation Density in Regulatory DNA of Cancer Genomes Is Linked to DNA Repair. *Nat Biotechnol* **32,** 71–75. ISSN: 1546-1696 (Jan. 2014).

193. Ocsenas, O. & Reimand, J. Chromatin Accessibility of Primary Human Cancers Ties Regional Mutational Processes and Signatures with Tissues of Origin. *PLoS Comput Biol* **18,** e1010393. ISSN: 1553-7358 (Aug. 2022).

194. Avgustinova, A. *et al.* Loss of G9a Preserves Mutation Patterns but Increases Chromatin Accessibility, Genomic Instability and Aggressiveness in Skin Tumours. *Nature Cell Biology* **20,** 1400–1409. ISSN: 1476-4679 (Dec. 2018).

195. Gindin, Y., Valenzuela, M. S., Aladjem, M. I., Meltzer, P. S. & Bilke, S. A Chromatin Structure-Based Model Accurately Predicts DNA Replication Timing in Human Cells. *Mol Syst Biol* **10,** 722. ISSN: 1744-4292 (Mar. 2014).

196. Lujan, S. A. *et al.* Heterogeneous Polymerase Fidelity and Mismatch Repair Bias Genome Variation and Composition. *Genome Res.* **24,** 1751–1764. ISSN: 1088-9051, 1549-5469 (Nov. 2014).

197. Li, F. *et al.* The Histone Mark H3K36me3 Regulates Human DNA Mismatch Repair through Its Interaction with MutSα. *Cell* **153,** 590–600. ISSN: 0092-8674 (Apr. 2013).

198. Pleasance, E. D. *et al.* A Small-Cell Lung Cancer Genome with Complex Signatures of Tobacco Exposure. *Nature* **463,** 184–190. ISSN: 1476-4687 (Jan. 2010).

199. Pleasance, E. D. *et al.* A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome. *Nature* **463,** 191–196. ISSN: 1476-4687 (Jan. 2010).

200. Schwaiger, M. & Schübeler, D. A Question of Timing: Emerging Links between Transcription and Replication. *Current Opinion in Genetics & Development. Chromosomes and Expression Mechanisms* **16,** 177–183. ISSN: 0959-437X (Apr. 2006).

201. Katainen, R. *et al.* CTCF/Cohesin-Binding Sites Are Frequently Mutated in Cancer. *Nat Genet* **47,** 818–821. ISSN: 1061-4036 (July 2015).

202. Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLOS Genetics* **12,** e1006207. ISSN: 1553-7404 (Apr. 2016).

203. Poulos, R. C. *et al.* Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. *Cell Reports* **17,** 2865–2872. ISSN: 2211-1247 (Dec. 2016).

204. Lee, C. A., Abd-Rabbo, D. & Reimand, J. Functional and Genetic Determinants of Mutation Rate Variability in Regulatory Elements of Cancer Genomes. *Genome Biology* **22,** 133. ISSN: 1474-760X (May 2021).

205. Davidson, I. F. & Peters, J.-M. Genome Folding through Loop Extrusion by SMC Complexes. *Nat Rev Mol Cell Biol* **22,** 445–464. ISSN: 1471-0080 (July 2021).

206. Sivapragasam, S. *et al.* CTCF Binding Modulates UV Damage Formation to Promote Mutation Hot Spots in Melanoma. *EMBO J* **40,** e107795. ISSN: 1460-2075 (Oct. 2021).

207. 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. ISSN: 0028-0836 (Apr. 2016).

208. Perera, D. *et al.* Differential DNA Repair Underlies Mutation Hotspots at Active Promoters in Cancer Genomes. *Nature* **532,** 259–263. ISSN: 0028-0836 (Apr. 2016).

209. Elliott, K. *et al.* Elevated Pyrimidine Dimer Formation at Distinct Genomic Bases Underlies Promoter Mutation Hotspots in UV-exposed Cancers. *PLOS Genetics* **14,** e1007849. ISSN: 1553-7404 (Dec. 2018).

210. Mao, P. *et al.* ETS Transcription Factors Induce a Unique UV Damage Signature That Drives Recurrent Mutagenesis in Melanoma. *Nat Commun* **9,** 2626. ISSN: 2041-1723 (July 2018).

211. Frigola, J., Sabarinathan, R., Gonzalez-Perez, A. & Lopez-Bigas, N. Variable Interplay of UV-induced DNA Damage and Repair at Transcription Factor Binding Sites. *Nucleic Acids Research* **49,** 891–901. ISSN: 0305-1048 (Jan. 2021).

212. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–1680. ISSN: 0092-8674 (Dec. 2014).

213. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171,** 305–320.e24. ISSN: 0092-8674, 1097-4172 (Oct. 2017).

214. Jerković, I. & Cavalli, G. Understanding 3D Genome Organization by Multidisciplinary Methods. *Nature Reviews Molecular Cell Biology,* 1–18. ISSN: 1471-0080 (May 2021).

215. Yan, J. *et al.* Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell* **154,** 801–813. ISSN: 0092-8674 (Aug. 2013).

216. Kaiser, V. B. & Semple, C. A. Chromatin Loop Anchors Are Associated with Genome Instability in Cancer and Recombination Hotspots in the Germline. *Genome Biology* **19,** 101. ISSN: 1474-760X (July 2018).

217. Akdemir, K. C. *et al.* Somatic Mutation Distributions in Cancer Genomes Vary with Three-Dimensional Chromatin Structure. *Nat Genet* **52,** 1178–1188. ISSN: 1546-1718 (Nov. 2020).

218. Canela, A. *et al.* Genome Organization Drives Chromosome Fragility. *Cell* **170,** 507–521.e18. ISSN: 0092-8674 (July 2017).

219. Qian, J. *et al.* B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. *Cell* **159,** 1524–1537. ISSN: 0092-8674, 1097-4172 (Dec. 2014).

220. Mattei, A. L., Bailly, N. & Meissner, A. DNA Methylation: A Historical Perspective. *Trends in Genetics* **38,** 676–707. ISSN: 0168-9525 (July 2022).

221. Wyatt, G. R. Occurrence of 5-Methyl-Cytosine in Nucleic Acids. *Nature* **166,** 237–238. ISSN: 1476-4687 (Aug. 1950).

222. Lister, R. *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* **341,** 1237905 (Aug. 2013).

223. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA Methylation: In the Right Place at the Right Time. *Science* **361,** 1336–1340. ISSN: 0036-8075 (Sept. 2018).

224. Jones, P. A. Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond. *Nat Rev Genet* **13,** 484–492. ISSN: 1471-0064 (July 2012).

225. Greenberg, M. V. C. & Bourchis, D. The Diverse Roles of DNA Methylation in Mammalian Development and Disease. *Nat Rev Mol Cell Biol* **20,** 590–607. ISSN: 1471-0080 (Oct. 2019).

226. Zhou, W. *et al.* DNA Methylation Loss in Late-Replicating Domains Is Linked to Mitotic Cell Division. *Nat Genet* **50,** 591–602. ISSN: 1546-1718 (Apr. 2018).

227. Berman, B. P. *et al.* Regions of Focal DNA Hypermethylation and Long-Range Hypomethylation in Colorectal Cancer Coincide with Nuclear Lamina–Associated Domains. *Nat Genet* **44,** 40–46. ISSN: 1546-1718 (Jan. 2012).

228. Leonhardt, H., Page, A. W., Weier, H. U. & Bestor, T. H. A Targeting Sequence Directs DNA Methyltransferase to Sites of DNA Replication in Mammalian Nuclei. *Cell* **71,** 865–873. ISSN: 0092-8674 (Nov. 1992).

229. Sen, M. *et al.* Strand-Specific Single-Cell Methylomics Reveals Distinct Modes of DNA Demethylation Dynamics during Early Mammalian Development. *Nat Commun* **12,** 1286. ISSN: 2041-1723 (Feb. 2021).

230. Ehrlich, M., Norris, K. F., Wang, R. Y., Kuo, K. C. & Gehrke, C. W. DNA Cytosine Methylation and Heat-Induced Deamination. *Biosci Rep* **6,** 387–393. ISSN: 0144-8463 (Apr. 1986).

231. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular Basis of Base Substitution Hotspots in Escherichia Coli. *Nature* **274,** 775–780. ISSN: 1476-4687 (Aug. 1978).

232. Seplyarskiy, V. B. & Sunyaev, S. The Origin of Human Mutation in Light of Genomic Data. *Nat Rev Genet* **22,** 672–686. ISSN: 1471-0064 (Oct. 2021).

233. Wiebauer, K. & Jiricny, J. In Vitro Correction of G.T Mispairs to G.C Pairs in Nuclear Extracts from Human Cells. *Nature* **339,** 234–236. ISSN: 0028-0836 (May 1989).

234. Bellacosa, A. & Drohat, A. C. Role of Base Excision Repair in Maintaining the Genetic and Epigenetic Integrity of CpG Sites. *DNA Repair. Cutting-Edge Perspectives in Genomic Maintenance II* **32,** 33–42. ISSN: 1568-7864 (Aug. 2015).

235. Millar, C. B. *et al.* Enhanced CpG Mutability and Tumorigenesis in MBD4-deficient Mice. *Science* **297,** 403–405. ISSN: 1095-9203 (July 2002).

236. Williams, K., Christensen, J. & Helin, K. DNA Methylation: TET Proteins-Guardians of CpG Islands? *EMBO Rep* **13,** 28–35. ISSN: 1469-3178 (Dec. 2011).

237. Popp, C. *et al.* Genome-Wide Erasure of DNA Methylation in Mouse Primordial Germ Cells Is Affected by AID Deficiency. *Nature* **463,** 1101–1105. ISSN: 1476-4687 (Feb. 2010).

238. Bhutani, N. *et al.* Reprogramming towards Pluripotency Requires AID-dependent DNA Demethylation. *Nature* **463,** 1042–1047. ISSN: 1476-4687 (Feb. 2010).

239. Bhutani, N., Burns, D. M. & Blau, H. M. DNA Demethylation Dynamics. *Cell* **146,** 866–872. ISSN: 0092-8674 (Sept. 2011).

240. Otlu, B. *et al.* Topography of Mutational Signatures in Human Cancer. *bioRxiv,* 2022.05.29.493921 (May 2022).

241. Tomkova, M. & Schuster-Böckler, B. DNA Modifications: Naturally More Error Prone? *Trends in Genetics* **34,** 627–638. ISSN: 0168-9525 (Aug. 2018).

242. Korona, D. A., Lecompte, K. G. & Pursell, Z. F. The High Fidelity and Unique Error Signature of Human DNA Polymerase Epsilon. *Nucleic Acids Res* **39,** 1763–1773. ISSN: 1362-4962 (Mar. 2011).

243. Tommasi, S., Denissenko, M. F. & Pfeifer, G. P. Sunlight Induces Pyrimidine Dimers Preferentially at 5-Methylcytosine Bases1. *Cancer Research* **57,** 4727–4730. ISSN: 0008-5472 (Nov. 1997).

244. Rochette, P. J. *et al.* Influence of Cytosine Methylation on Ultraviolet-Induced Cyclobutane Pyrimidine Dimer Formation in Genomic DNA. *Mutation Research* **665,** 7–13. ISSN: 0027-5107 (June 2009).

245. Cannistraro, V. J., Pondugula, S., Song, Q. & Taylor, J.-S. Rapid Deamination of Cyclobutane Pyrimidine Dimer Photoproducts at TCG Sites in a Translationally and Rotationally Positioned Nucleosome in Vivo. *Journal of Biological Chemistry* **290,** 26597–26609. ISSN: 0021-9258, 1083-351X (Oct. 2015).

246. Seplyarskiy, V. B. *et al.* APOBEC-induced Mutations in Human Cancers Are Strongly Enriched on the Lagging DNA Strand during Replication. *Genome Res.* **26,** 174–182. ISSN: 1088-9051, 1549-5469 (Jan. 2016).

247. Schutsky, E. K., Nabel, C. S., Davis, A. K. F., DeNizio, J. E. & Kohli, R. M. APOBEC3A Efficiently Deaminates Methylated, but Not TET-oxidized, Cytosine Bases in DNA. *Nucleic Acids Research* **45,** 7655–7665. ISSN: 0305-1048 (July 2017).

248. Carpenter, M. A. *et al.* Methylcytosine and Normal Cytosine Deamination by the Foreign DNA Restriction Enzyme APOBEC3A. *J Biol Chem* **287,** 34801–34808. ISSN: 1083-351X (Oct. 2012).

249. Suspène, R., Aynaud, M.-M., Vartanian, J.-P. & Wain-Hobson, S. Efficient Deamination of 5-Methylcytidine and 5-Substituted Cytidine Residues in DNA by Human APOBEC3A Cytidine Deaminase. *PLoS One* **8,** e63461. ISSN: 1932-6203 (2013).

250. DeWeerd, R. A. *et al.* Prospectively Defined Patterns of APOBEC3A Mutagenesis Are Prevalent in Human Cancers. *Cell Reports* **38,** 110555. ISSN: 2211-1247 (Mar. 2022).

251. Ernst, J. & Kellis, M. Chromatin-State Discovery and Genome Annotation with ChromHMM. *Nat Protoc* **12,** 2478–2492. ISSN: 1750-2799 (Dec. 2017).

252. Kundaje, A. *et al.* Integrative Analysis of 111 Reference Human Epigenomes. *Nature* **518,** 317–330. ISSN: 1476-4687 (Feb. 2015).

253. Moore, J. E. *et al.* Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. *Nature* **583,** 699–710. ISSN: 1476-4687 (July 2020).

254. Vavouri, T. & Lehner, B. Human Genes with CpG Island Promoters Have a Distinct Transcription-Associated Chromatin Organization. *Genome Biology* **13,** 1–12. ISSN: 1474-760X (Nov. 2012).

255. Baubec, T. *et al.* Genomic Profiling of DNA Methyltransferases Reveals a Role for DNMT3B in Genic Methylation. *Nature* **520,** 243–247. ISSN: 1476-4687 (Apr. 2015).

256. Huang, Y., Gu, L. & Li, G.-M. H3K36me3-mediated Mismatch Repair Preferentially Protects Actively Transcribed Genes from Mutation. *J. Biol. Chem.* **293,** 7811–7823. ISSN: 0021-9258, 1083-351X (May 2018).

257. Rheinbay, E. *et al.* Analyses of Non-Coding Somatic Drivers in 2,658 Cancer Whole Genomes. *Nature* **578,** 102–111. ISSN: 1476-4687 (Feb. 2020).

258. Beroukhim, R. *et al.* The Landscape of Somatic Copy-Number Alteration across Human Cancers. *Nature* **463,** 899–905. ISSN: 1476-4687 (Feb. 2010).

259. Chen, Y. *et al.* Identification of Druggable Cancer Driver Genes Amplified across TCGA Datasets. *PLOS ONE* **9,** e98293. ISSN: 1932-6203 (May 2014).

260. Mermel, C. H. *et al.* GISTIC2.0 Facilitates Sensitive and Confident Localization of the Targets of Focal Somatic Copy-Number Alteration in Human Cancers. *Genome Biology* **12,** R41. ISSN: 1474-760X (Apr. 2011).

261. Zack, T. I. *et al.* Pan-Cancer Patterns of Somatic Copy Number Alteration. *Nat Genet* **45,** 1134–1140. ISSN: 1546-1718 (Oct. 2013).

262. Schwer, B. *et al.* Transcription-Associated Processes Cause DNA Double-Strand Breaks and Translocations in Neural Stem/Progenitor Cells. *Proceedings of the National Academy of Sciences* **113,** 2258–2263 (Feb. 2016).

263. Pecori, R., Di Giorgio, S., Paulo Lorenzo, J. & Nina Papavasiliou, F. Functions and Consequences of AID/APOBEC-mediated DNA and RNA Deamination. *Nat Rev Genet* **23,** 505–518. ISSN: 1471-0064 (Aug. 2022).

264. Teng, B., Burant, C. F. & Davidson, N. O. Molecular Cloning of an Apolipoprotein B Messenger RNA Editing Protein. *Science* **260,** 1816–1819. ISSN: 0036-8075 (June 1993).

265. Espinosa, R., Funahashi, T., Hadjiagapiou, C., Le Beau, M. M. & Davidson, N. O. Assignment of the Gene Encoding the Human Apolipoprotein B mRNA Editing Enzyme (APOBEC1) to Chromosome 12p13.1. *Genomics* **24,** 414–415. ISSN: 0888-7543 (Nov. 1994).

266. Higuchi, K. *et al.* Human Apolipoprotein B (apoB) mRNA: Identification of Two Distinct apoB mRNAs, an mRNA with the apoB-100 Sequence and an apoB mRNA Containing a Premature in-Frame Translational Stop Codon, in Both Liver and Intestine. *Proc Natl Acad Sci U S A* **85,** 1772–1776. ISSN: 0027-8424 (Mar. 1988).

267. Liao, W. *et al.* APOBEC-2, a Cardiac- and Skeletal Muscle-Specific Member of the Cytidine Deaminase Supergene Family. *Biochem Biophys Res Commun* **260,** 398–404. ISSN: 0006-291X (July 1999).

268. Muramatsu, M. *et al.* Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. *Cell* **102,** 553–563. ISSN: 0092-8674, 1097-4172 (Sept. 2000).

269. Jarmuz, A. *et al.* An Anthropoid-Specific Locus of Orphan C to U RNA-editing Enzymes on Chromosome 22. *Genomics* **79,** 285–296. ISSN: 0888-7543 (Mar. 2002).

270. Madsen, P. *et al.* Psoriasis Upregulated Phorbolin-1 Shares Structural but Not Functional Similarity to the mRNA-editing Protein Apobec-1. *J Invest Dermatol* **113,** 162–169. ISSN: 0022-202X (Aug. 1999).

271. Ito, J., Gifford, R. J. & Sato, K. Retroviruses Drive the Rapid Evolution of Mammalian APOBEC3 Genes. *Proceedings of the National Academy of Sciences* **117,** 610–618 (Jan. 2020).

272. Harris, R. S. *et al.* DNA Deamination Mediates Innate Immunity to Retroviral Infection. *Cell* **113,** 803–809. ISSN: 0092-8674 (June 2003).

273. Mangeat, B. *et al.* Broad Antiretroviral Defence by Human APOBEC3G through Lethal Editing of Nascent Reverse Transcripts. *Nature* **424,** 99–103. ISSN: 1476-4687 (July 2003).

274. Sheehy, A. M., Gaddis, N. C., Choi, J. D. & Malim, M. H. Isolation of a Human Gene That Inhibits HIV-1 Infection and Is Suppressed by the Viral Vif Protein. *Nature* **418,** 646–650. ISSN: 0028-0836 (Aug. 2002).

275. Yu, Q. *et al.* APOBEC3B and APOBEC3C Are Potent Inhibitors of Simian Immunodeficiency Virus Replication. *J Biol Chem* **279,** 53379–53386. ISSN: 0021-9258 (Dec. 2004).

276. Bonvin, M. *et al.* Interferon-Inducible Expression of APOBEC3 Editing Enzymes in Human Hepatocytes and Inhibition of Hepatitis B Virus Replication. *Hepatology* **43,** 1364–1374. ISSN: 0270-9139 (June 2006).

277. Warren, C. J. *et al.* APOBEC3A Functions as a Restriction Factor of Human Papillomavirus. *J Virol* **89,** 688–702. ISSN: 1098-5514 (Jan. 2015).

278. Di Giorgio, S., Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for Host-Dependent RNA Editing in the Transcriptome of SARS-CoV-2. *Science Advances* **6,** eabb5813 (June 2020).

279. Klimczak, L. J., Randall, T. A., Saini, N., Li, J.-L. & Gordenin, D. A. Similarity between Mutation Spectra in Hypermutated Genomes of Rubella Virus and in SARS-CoV-2 Genomes Accumulated during the COVID-19 Pandemic. *PLOS ONE* **15,** e0237689. ISSN: 1932-6203 (Oct. 2020).

280. Jalili, P. *et al.* Quantification of Ongoing APOBEC3A Activity in Tumor Cells by Monitoring RNA Editing at Hotspots. *Nat Commun* **11,** 2971. ISSN: 2041-1723 (June 2020).

281. Stephens, P. *et al.* A Screen of the Complete Protein Kinase Gene Family Identifies Diverse Patterns of Somatic Mutations in Human Breast Cancer. *Nat Genet* **37,** 590–592. ISSN: 1546-1718 (June 2005).

282. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149,** 994–1007. ISSN: 0092-8674 (May 2012).

283. Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. *Mol Cell* **10,** 1247–1253. ISSN: 1097-2765 (Nov. 2002).

284. Wang, J. *et al.* Evidence for Mutation Showers. *Proceedings of the National Academy of Sciences* **104,** 8403–8408 (May 2007).

285. Chan, K. *et al.* Base Damage within Single-Strand DNA Underlies In Vivo Hypermutability Induced by a Ubiquitous Environmental Agent. *PLOS Genetics* **8,** e1003149. ISSN: 1553-7404 (Dec. 2012).

286. Burns, M. B. *et al.* APOBEC3B Is an Enzymatic Source of Mutation in Breast Cancer. *Nature* **494,** 366–370. ISSN: 0028-0836, 1476-4687 (Feb. 2013).

287. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B Mutagenesis in Multiple Human Cancers. *Nature Genetics* **45,** 977–983. ISSN: 1546-1718 (Sept. 2013).

288. Roberts, S. A. *et al.* An APOBEC Cytidine Deaminase Mutagenesis Pattern Is Widespread in Human Cancers. *Nature Genetics* **45,** 970–976. ISSN: 1546-1718 (Sept. 2013).

289. Taylor, B. J. *et al.* DNA Deaminases Induce Break-Associated Mutation Showers with Implication of APOBEC3B and 3A in Breast Cancer Kataegis. *eLife* **2** (ed Stamatoyannopoulos, J.) e00534. ISSN: 2050-084X (Apr. 2013).

290. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms Underlying Mutational Signatures in Human Cancers. *Nat Rev Genet* **15,** 585–598. ISSN: 1471-0056 (Sept. 2014).

291. Krokan, H. E., Drabløs, F. & Slupphaug, G. Uracil in DNA – Occurrence, Consequences and Repair. *Oncogene* **21,** 8935–8948. ISSN: 1476-5594 (Dec. 2002).

292. Roberts, S. A. & Gordenin, D. A. Hypermutation in Human Cancer Genomes: Footprints and Mechanisms. *Nature Reviews Cancer* **14,** 786–800. ISSN: 1474-175X, 1474-1768 (Nov. 2014).

293. Strauss, B. S. The "A" Rule Revisited: Polymerases as Determinants of Mutational Specificity. *DNA Repair* **1,** 125–135. ISSN: 1568-7864 (Feb. 2002).

294. Goldmann, J. M. *et al.* Germline de Novo Mutation Clusters Arise during Oocyte Aging in Genomic Regions with High Double-Strand-Break Incidence. *Nature Genetics* **50,** 487. ISSN: 1546-1718 (Apr. 2018).

295. Sakofsky, C. J. *et al.* Break-Induced Replication Is a Source of Mutation Clusters Underlying Kataegis. *Cell Reports* **7,** 1640–1648. ISSN: 2211-1247 (June 2014).

296. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* **163,** 1641–1654. ISSN: 1097-4172 (Dec. 2015).

297. Maciejowski, J. *et al.* APOBEC3-dependent Kataegis and TREX1-driven Chromothripsis during Telomere Crisis. *Nat Genet* **52,** 884–890. ISSN: 1546-1718 (Sept. 2020).

298. Cortés-Ciriano, I. *et al.* Comprehensive Analysis of Chromothripsis in 2,658 Human Cancers Using Whole-Genome Sequencing. *Nat Genet* **52,** 331–341. ISSN: 1546-1718 (Mar. 2020).

299. Sollier, J. & Cimprich, K. A. Breaking Bad: R-loops and Genome Integrity. *Trends in Cell Biology* **25,** 514–522. ISSN: 0962-8924 (Sept. 2015).

300. Bhagwat, A. S. *et al.* Strand-Biased Cytosine Deamination at the Replication Fork Causes Cytosine to Thymine Mutations in *Escherichia Coli*. *Proceedings of the National Academy of Sciences* **113,** 2176–2181. ISSN: 0027-8424, 1091-6490 (Feb. 2016).

301. Hoopes, J. I. *et al.* APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication. *Cell Reports* **14,** 1273–1282. ISSN: 2211-1247 (Feb. 2016).

302. Kazanov, M. D. *et al.* APOBEC-induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. *Cell Reports* **13,** 1103–1109. ISSN: 2211-1247 (Nov. 2015).

303. Chen, J., Miller, B. F. & Furano, A. V. Repair of Naturally Occurring Mismatches Can Induce Mutations in Flanking DNA. *eLife* **3,** e02001. ISSN: 2050-084X (Apr. 2014).

304. Chen, J. & Furano, A. V. Breaking Bad: The Mutagenic Effect of DNA Repair. *DNA Repair. Cutting-Edge Perspectives in Genomic Maintenance II* **32,** 43–51. ISSN: 1568-7864 (Aug. 2015).

305. Andrianova, M. A., Bazykin, G. A., Nikolaev, S. I. & Seplyarskiy, V. B. Human Mismatch Repair System Balances Mutation Rates between Strands by Removing More Mismatches from the Lagging Strand. *Genome Res.* **27,** 1336–1343. ISSN: 1088-9051, 1549-5469 (Aug. 2017).

306. Peters, J.-M. How DNA Loop Extrusion Mediated by Cohesin Enables V(D)J Recombination. *Current Opinion in Cell Biology* **70,** 75–83. ISSN: 0955-0674 (June 2021).

307. Peled, J. U. *et al.* The Biochemistry of Somatic Hypermutation. *Annu Rev Immunol* **26,** 481–511. ISSN: 0732-0582 (2008).

308. Casali, P., Pal, Z., Xu, Z. & Zan, H. DNA Repair in Antibody Somatic Hypermutation. *Trends Immunol* **27,** 313–321. ISSN: 1471-4906 (July 2006).

309. Wiesendanger, M., Kneitz, B., Edelmann, W. & Scharff, M. D. Somatic Hypermutation in Muts Homologue (Msh)3-, Msh6-, and Msh3/Msh6-Deficient Mice Reveals a Role for the Msh2–Msh6 Heterodimer in Modulating the Base Substitution Pattern. *Journal of Experimental Medicine* **191,** 579–584. ISSN: 0022-1007 (Feb. 2000).

310. Peña-Diaz, J. *et al.* Noncanonical Mismatch Repair as a Source of Genomic Instability in Human Cells. *Molecular Cell* **47,** 669–680. ISSN: 1097-2765 (Sept. 2012).

311. Masuda, K. *et al.* DNA Polymerase Eta Is a Limiting Factor for A:T Mutations in Ig Genes and Contributes to Antibody Affinity Maturation. *Eur J Immunol* **38,** 2796–2805. ISSN: 0014-2980 (Oct. 2008).

312. Beekman, R. *et al.* The Reference Epigenome and Regulatory Chromatin Landscape of Chronic Lymphocytic Leukemia. *Nat Med* **24,** 868–880. ISSN: 1546-170X (June 2018).

313. Zlatanou, A. *et al.* The hMsh2-hMsh6 Complex Acts in Concert with Monoubiquitinated PCNA and Pol η in Response to Oxidative DNA Damage in Human Cells. *Mol Cell* **43,** 649–662. ISSN: 1097-4164 (Aug. 2011).

314. Rogozin, I. B. *et al.* DNA Polymerase η Mutational Signatures Are Found in a Variety of Different Types of Cancer. *Cell Cycle* **17,** 1–31. ISSN: 1538-4101 (Feb. 2018).

315. Harris, K. & Nielsen, R. Error-Prone Polymerase Activity Causes Multinucleotide Mutations in Humans. *Genome Res.* **24,** 1445–1454. ISSN: 1088-9051, 1549-5469 (July 2014).

316. Seplyarskiy, V. B., Bazykin, G. A. & Soldatov, R. A. Polymerase ζ Activity Is Linked to Replication Timing in Humans: Evidence from Mutational Signatures. *Mol Biol Evol* **32,** 3158–3172. ISSN: 0737-4038 (Dec. 2015).

317. Shale, C. *et al.* Unscrambling Cancer Genomes via Integrated Analysis of Structural Variation and Copy Number. *Cell Genomics* **2,** 100112. ISSN: 2666-979X (Apr. 2022).

318. Hadi, K. *et al.* Distinct Classes of Complex Structural Variation Uncovered across Thousands of Cancer Genome Graphs. *Cell* **183,** 197–210.e32. ISSN: 0092-8674 (Oct. 2020).

319. Baca, S. C. *et al.* Punctuated Evolution of Prostate Cancer Genomes. *Cell* **153,** 666–677. ISSN: 0092-8674 (Apr. 2013).

320. Stone, J. E., Lujan, S. A. & Kunkel, T. A. DNA Polymerase Zeta Generates Clustered Mutations during Bypass of Endogenous DNA Lesions in Saccharomyces Cerevisiae. *Environmental and Molecular Mutagenesis* **53,** 777–786. ISSN: 1098-2280 (Dec. 2012).

321. Goldmann, J. M. *et al.* Parent-of-Origin-Specific Signatures of *de Novo* Mutations. *Nature Genetics* **48,** ng.3597. ISSN: 1546-1718 (June 2016).

322. Jónsson, H. *et al.* Parental Influence on Human Germline de Novo Mutations in 1,548 Trios from Iceland. *Nature* **549,** 519–522. ISSN: 1476-4687 (Sept. 2017).

323. Seplyarskiy, V. B. *et al.* Population Sequencing Data Reveal a Compendium of Mutational Processes in the Human Germ Line. *Science* **373,** 1030–1035 (Aug. 2021).

324. Francioli, L. C. *et al.* Genome-Wide Patterns and Properties of *de Novo* Mutations in Humans. *Nature Genetics* **47,** 822–826. ISSN: 1546-1718 (July 2015).

325. Walker, C. R., Scally, A., Maio, N. D. & Goldman, N. Short-Range Template Switching in Great Ape Genomes Explored Using Pair Hidden Markov Models. *PLOS Genetics* **17,** e1009221. ISSN: 1553-7404 (Mar. 2021).

326. Mandelkern, M., Elias, J. G., Eden, D. & Crothers, D. M. The Dimensions of DNA in Solution. *Journal of Molecular Biology* **152,** 153–161. ISSN: 0022-2836 (Oct. 1981).

327. Bennett, P. V., Cintron, N. S., Gros, L., Laval, J. & Sutherland, B. M. Are Endogenous Clustered Dna Damages Induced in Human Cells? *Free Radical Biology and Medicine* **37,** 488–499. ISSN: 0891-5849 (Aug. 2004).

328. Lomax, M. E., Folkes, L. K. & O'Neill, P. Biological Consequences of Radiation-induced DNA Damage: Relevance to Radiotherapy. *Clinical Oncology. Advances in Clinical Radiobiology* **25,** 578–585. ISSN: 0936-6555 (Oct. 2013).

329. Smith, K. S., Liu, L. L., Ganesan, S., Michor, F. & De, S. Nuclear Topology Modulates the Mutational Landscapes of Cancer Genomes. *Nature Structural & Molecular Biology* **24,** 1000–1006. ISSN: 1545-9985 (Nov. 2017).

330. Guelen, L. *et al.* Domain Organization of Human Chromosomes Revealed by Mapping of Nuclear Lamina Interactions. *Nature* **453,** 948–951. ISSN: 1476-4687 (June 2008).

331. García-Nieto, P. E. *et al.* Carcinogen Susceptibility Is Regulated by Genome Architecture and Predicts Cancer Mutagenesis. *EMBO J* **36,** 2829–2843. ISSN: 0261-4189, 1460-2075 (Oct. 2017).

332. Grubert, F. *et al.* Landscape of Cohesin-Mediated Chromatin Loops in the Human Genome. *Nature* **583,** 737–743. ISSN: 1476-4687 (July 2020).

333. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell* **78,** 554–565.e7. ISSN: 1097-2765 (May 2020).

334. Hsieh, T.-H. S. *et al.* Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell* **78,** 539–553.e8. ISSN: 1097-2765 (May 2020).

335. Efron, B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* First Edition edition. ISBN: 978-0-521-19249-1 (Cambridge University Press, Cambridge ; New York, Sept. 2010).

336. Jeon, Y. *et al.* Dynamic Control of Strand Excision during Human DNA Mismatch Repair. *PNAS* **113,** 3281–3286. ISSN: 0027-8424, 1091-6490 (Mar. 2016).

337. Bowen, N. *et al.* Reconstitution of Long and Short Patch Mismatch Repair Reactions Using *Saccharomyces Cerevisiae* Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 18472–18477. ISSN: 1091-6490 (Nov. 2013).

338. Tomkova, M., McClellan, M., Kriaucionis, S. & Schuster-Böckler, B. DNA Replication and Associated Repair Pathways Are Involved in the Mutagenesis of Methylated Cytosine. *DNA Repair* **62,** 1–7. ISSN: 1568-7864 (Feb. 2018).

339. Aska, E.-M., Dermadi, D. & Kauppi, L. Single-Cell Sequencing of Mouse Thymocytes Reveals Mutational Landscape Shaped by Replication Errors, Mismatch Repair, and H3K36me3. *iScience* **23,** 101452. ISSN: 2589-0042 (Aug. 2020).