Published in final edited form as:

Nat Rev Genet. 2015 April ; 16(4): 213–223. doi:10.1038/nrg3890.

The effects of chromatin organization on variation in mutation rates in the genome

Kateryna D. Makova¹ and Ross C. Hardison²

¹Department of Biology, Huck Institute for Genome Sciences, The Pennsylvania State University, University Park, PA 16802 USA

²Department of Biochemistry and Molecular Biology, Huck Institute for Genome Sciences, The Pennsylvania State University, University Park, PA 16802 USA

Abstract

The variation in local rates of mutations can affect both the evolution of genes and their function in normal and cancer cells. Deciphering the molecular determinants of this variation will be aided by resolving distinct types of mutation, since they differ in regional preferences and in associations with genomic features. Chromatin organization contributes to regional variation in mutation rate, but differently among mutation types. In both germ-line mutations and somatic mutations, base substitutions are more abundant in regions of closed chromatin, perhaps reflecting error accumulation late in replication. In contrast, a distinctive mutational state with very high levels of indels and substitutions is enriched in regions of open chromatin. These associations illuminate an intricate interplay between the nucleotide sequence of DNA and its dynamic packaging into chromatin, and they have important implications for current biomedical research. This review focuses on the recent studies showing associations between chromatin state and mutation rates, including pairwise and multivariate investigations of germ-line and somatic (particularly cancer) mutations.

INTRODUCTION

Mutations are the foundation for evolution by providing raw material for selection and drift, and they have a central role in causing many human diseases including cancer. Therefore the knowledge about how mutations occur and how their frequency is affected by genomic landscape is paramount for understanding both the evolutionary process and human diseases. Mutations can be classified based on their effect on DNA structure and the number of nucleotides they effect (referred to as 'scale') (**Box 1**). In mammals, some of the most common mutations are base substitutions, small insertions and deletions (indels), transposable element (TE) insertions, and segmental duplications. Studies of individual genes had indicated that mutation rates are not uniform across the genome^{1,2}, and this rate heterogeneity was demonstrated unequivocally by analysis of alignments of genome

Correspondence to K.D.M. kdm16@psu.edu.

Competing interests statement

The authors declare no competing interests.

sequences of several mammalian genomes³⁻⁹. Mutation rates not only differ between autosomes and the two sex chromosomes due to male mutation bias¹⁰, but also vary along individual chromosomes, a phenomenon termed *regional variation in mutation rates* (RViMR³; reviewed in⁴). RViMR was originally demonstrated for base substitutions^{1-3,5}, but was soon extended to include small insertions⁶, small deletions^{3,6}, and TE insertions^{3,5,7,8,11} (**Figure 1**). Moreover, a substantial *co-variation* has been found among rates of different mutation types^{3,7,9}. In these studies, the mutation rate was inferred via comparison of neutrally evolving orthologous regions (synonymous sites, ancestral repeats, or noncoding nonrepetitive regions) of mammalian genomes, because in such regions, the mutation rate is equal to the gene substitution rate¹².

RViMR not only determines (at least in part) the patterns and rates of evolution, it also affects the location of genes along the genome, with more strongly conserved genes located in regions with lower mutation rates¹³. Furthermore, RViMR influences how genome sequence alignments can be used to predict function. For instance, frequently it is necessary to evaluate whether the alignment is ‘more conserved than can be expected by chance alone’¹⁴, e.g., when inferring selection. Such evaluation will greatly depend on the underlying local mutation rate.

Recent studies have shown that chromatin organization is frequently altered in cancer, and this can strongly affect mutation rates and patterns¹⁵. Indeed, while originally RViMR was discovered via studying non-cancerous germ-line mutations inferred from alignments of mammalian genomes (e.g.,³), the rates of somatic mutations are also non-uniform within cancer genomes⁴. A better understanding of the determinants of the rate variation could assist in interpreting the biological impacts of the large numbers of somatic mutations that accumulate in cancers.

Recently, *nucleosome occupancy* and other *epigenomic features* were found to be significant predictors of non-uniformity in mutation rates, both in normal and cancer cells^{16,17} (note that our use of the term “epigenomic” does not imply transgenerational inheritance). Moreover, epigenomic features including chromatin provide a link between the nucleotide sequence of the DNA and the dynamic changes in the packaging and expression of the DNA, a link that helps explain various aspects of human genetic diversity. For example, many human SNPs are located in regions of variably methylated DNA¹⁸, and trait-associated genetic variants are enriched in DNA packaged into chromatin with histone modifications and other features associated with regulation^{19,20}. Here, we review how genomic and epigenomic features characterizing chromatin organization affect rates and patterns of the most common mutations in mammalian genomes – base substitutions, small insertions and deletions, and TE insertions. We highlight the differences in potential determinants of the distinct classes of mutations, examining both germline and somatic mutations important for evolution and disease susceptibility. We focus on mammals because for them both RViMR and chromatin organization have been studied in most detail.

Genomic landscape features that contribute to regional variation in mutation rate

Several hypotheses have been proposed to explain RViMR, and most of them stem from an observation that local rates of different mutation types correlate with various features of

local genomic landscape. Such so called *genomic landscape features* characterize the genome at levels beyond the primary DNA sequence; they include GC content, recombination rates, proximity to the closest telomere, replication timing²¹, among many others^{3,22} (**Table 1**). The association between many genomic landscape features and mutation rate, which are usually measured within windows (that is, genomic intervals of a defined size), can be explained in some cases. On the one hand, the base substitution rate has a quadratic^{22,23} relationship with GC content. The elevated substitution rate at high GC content results from the increased frequency of CpG nucleotides, which, when methylated, become mutation hotspots and thus have elevated mutation rates^{3,22,24}. On the other hand, genomic regions with high AT content also have elevated substitution rates. Many AT-rich regions are depleted of genes and can be packaged into heterochromatin, which in turn has a high substitution rate (see below). Also, an increase in base substitution rates close to telomeres can be explained by altered repair in these regions of the genome²⁵. Altogether, multivariate analyses of genomic landscape features (e.g., GC content, exon density, location on autosomes vs. sex chromosomes, exon density, male recombination rates, and distance to telomere) explain 82% and 52% of the genome-wide variability in mutation rates at CpG and non-CpG sites, respectively²² (**Table 1**). Similarly, ~30% of the genome-wide variability in small indel rates can be attributed to the variation in many of the same genomic landscape features plus CpG island content and poly(A/T) content (**Table 1**)⁶. Insertion preferences of TEs are also determined in part by the genomic landscape. For instance, some of these same genomic landscape features plus recombination hotspot frequency, LINE target sequence frequency, and frequencies of the genome instability 13-mer and of the telomere hexamer (**Table 1**) can account for 20% and 41% of the variability in insertion rates for young (human-specific) *Alus* and *L1s*⁸, respectively. These landscape features also contribute to the preferences of *ex vivo* DNA transposon¹¹ and *Alu*²⁶ integrations.

The substantial co-variation in rates for substitutions, indels, and TE insertions deduced from whole-genome alignments between mammals in different orders suggested that genomic landscape features had a similar impact on these disparate types of mutations^{3-5,7}. However, studies built upon alignments among closely related primates showed distinct rates and patterns for different types of mutations^{6,8,9,22} (**Table 1**). Resolving these links between genomic features and mutation rates/patterns for distinct types of mutations assists in elucidating the causes of RViMR, thereby uncovering the intricacies of different mechanisms of mutagenesis. For instance, an investigation of regional rate variation suggested that recombination plays an important role in generation of small insertions, while replication errors contribute in a major way to small deletions⁶.

The greatest variation in RViMR has been observed at the smallest scales⁴, in particular, at the level of neighboring nucleotides where the CpG context is dominant, as methylated cytosines adjacent to guanines undergo spontaneous deamination to thymines at about ten times the rate of other C>T transitions²⁷. However the RViMR has been studied most extensively at the 1-Mb scale that arguably represents the “natural scale” of variation in mutation rates in mammals²⁸.

Chromatin structure affects mutation rate

Many genomic landscape features that are associated with RViMR – for example, GC content, CpG islands (CpG rich regions present at many promoters), and recombination rates – are largely static over the lifetime of an organism. However, the processes that “read” the genetic information and regulate that reading in time and space are highly dynamic and are affected by changes in epigenomic modifications. The first order of such epigenomic features can be considered “one-dimensional”, i.e. they are laid down along the DNA sequence. Methylation of the DNA is the prime example. While this modification is classically associated with gene repression, recent studies show that active promoters are demethylated, but the introns and exons of transcribed genes show increased methylation²⁹. The methylation relates directly to mutation rates, since as already noted, methylated CpG undergoes spontaneous oxidative deamination to form TpG at high rates.

Other epigenomic features affect the accessibility of DNA in chromatin. These include one-dimensional features such as the positioning of nucleosomes on specific sequences, which decreases access to DNA on the core of the nucleosome, and the remodeling of nucleosomes leading to their ejection, sliding, or restructuring³⁰ and increasing the local accessibility of DNA in *open chromatin* (**Figure 2**). Remodeling of nucleosomes can be detected by hypersensitivity of DNA in chromatin to DNases, revealing DNase hypersensitive sites (DHSs). Many DHSs are bound by transcription factors to form active promoters, enhancers, insulators, and other regulatory modules. Histone modifications are placed on the tails of histones in chromatin and removed in dynamic processes associated with transcription and other nuclear events. Thus the packaging of genomic DNA into chromatin with specific histone modifications can indicate candidate functions. For example, enhancers tend to be located in chromatin with monomethylation on lysine 4 and acetylation on lysine 27 of histone H3 (H3K4me1 and H3K27ac), active transcription start sites in promoters tend to be flanked by nucleosomes with trimethylation of lysine 4 of histone H3 (H3K4me3), and transcribed chromatin is marked by trimethylation of lysine 36 of histone H3 (H3K36me3) (reviewed in³¹; **Figure 2**). In contrast, repressed DNA can be in one of three chromatin states³², namely trimethylation of lysine 27 of histone H3 (H3K27me3), trimethylation of lysine 9 of histone H3 (H3K9me3), or markedly depleted of histone modifications (quiescent or dead zones; **Figure 2**).

The one-dimensional alignment of epigenomic features along the genomic DNA sequences only captures part of the chromatin conformation that could affect mutation rates. The three-dimensional structure of chromatin is not known in detail, but multiple lines of evidence indicate that specific, long-distance interactions affect gene function, such as contacts between active distal enhancers and promoters³³. These contacts in three-dimensional space between DNA sequences that are far apart on a linear genome sequence have been proposed to generate loops of the intervening DNA that could be sites of large-scale deletions or rearrangements. Indeed, independently derived large deletions (from different ethnic groups) causing beta-thalassemia can have similar end-points, suggesting that large deletions may occur by breaking and rejoining of DNA duplexes at the bases of the loops³⁴. Technologies such as those related to *chromosome conformation capture*³⁵ are providing large-scale, even genome-wide^{36,37}, maps of distal interactions. These comprehensive interaction maps

indicate that genomic DNA segments fall into one of two categories, either dominated by local interactions (closed chromatin) or with a higher frequency of distal interactions (open chromatin, **Figure 2**). Some of the closed chromatin may correspond to heterochromatin, a highly condensed form of chromatin that can be found at the nuclear periphery in contact with the nuclear lamina (**Figure 2**). Indeed, large numbers of nuclear lamina binding sites are associated with a form of silenced heterochromatin³⁸.

These epigenomic features affecting or associated with different degrees of accessibility to the underlying DNA might be expected to affect mutation rates, but the prediction of the outcome is dependent on whether the accessibility has a greater impact on mutagenic or repair processes. For example, increased access could lead to increased exposure to DNA damaging agents resulting in higher mutation rates. Conversely, improved access could lead to greater surveillance and correction of mutations by the cellular DNA repair enzymes, which would result in lower mutation rates. The next section explores the connections between chromatin organization and rates of different classes of mutations on a genome-wide scale.

Chromatin structure and interpecies divergence

Recent studies have indicated that chromatin organization is one of the causative agents of RViMR^{39,40,41,42}. Below, we explore the connections between chromatin organization and rates of different classes of mutations on a genome-wide scale.

Pairwise analyses—The overall mutation rate at methylated CpG dinucleotides is an order of magnitude higher than at other sites due to replication-independent spontaneous cytosine deamination⁴³. Interestingly, CpG islands (1-kb regions present in many promoters) are usually unmethylated, and thus have lower mutation rates, and this serves as one of the major mechanisms of maintaining high GC content in these regions which are frequently located in the vicinity of transcription start sites⁴⁴.

Several studies illustrate that nucleosome occupancy affects the rate of mutation by base substitution. For instance, in yeast, nematode, and medaka (commonly known as the Japanese rice fish), the C->T, G->T and A->T rates are reduced in DNA packaged in nucleosomes⁴⁰. This result can potentially be explained by the fact that DNA within nucleosomes undergoes less breathing⁴⁵ (spontaneous local conformational fluctuations within double-stranded DNA) and thus is less prone to cytosine deamination⁴⁰. A detailed study of human base substitution patterns indicated that the C->T rates are significantly depleted within nucleosome core regions, but are elevated at linker regions located 60-90 bps from the nucleosome dyad⁴⁶. The authors suggest that this substitution pattern may result from selection acting to maintain optimal GC composition in core and linker regions⁴⁶. In contrast, the T->C, A->G, G->T, C->A, T->A, and A->T mutation rates were elevated and associated with certain histone modifications in nucleosome-occupied regions in human⁴⁶. Moreover, a recent study of *de novo* mutations in autistic individuals also indicated that nucleosome occupancy was associated with suppressed substitution mutations⁴², however another study indicated that human SNPs are depleted around the nucleosomes with histone modifications but are enriched around the bulk chromosomes⁴⁷.

The latter study found indel depletion around positions occupied by nucleosomes⁴⁷. Similarly, periodic high indel rates (but low base substitution rates) were observed at positions 200, 400, and 600 bp downstream of transcription start sites -- regions between positioned nucleosomes -- in medaka⁴¹. In contrast, positioning of H2A.Z nucleosomes, which tend to be close to transcription start sites of genes, had no effect on *Drosophila* mutation rates⁴⁸. Also, in yeast, nucleosome-free “linker sequences” had lower substitution rates than nucleosomal DNA^{49,50}, arguing for species-specific patterns.

Nucleotide substitution rates at CpG sites in primates (human-chimpanzee comparisons) were found to be decreased in open chromatin regions of the genome likely because such regions experience lower rates of DNA damage and enhanced DNA repair³⁹. Additional evidence supporting this observation came from the analysis of DHSs that were shown to have less divergence in primates than the sequences surrounding them^{51,52}. Data from the Encyclopedia of DNA Elements (ENCODE) project demonstrated that DHSs exhibit lower nucleotide diversity in humans than fourfold degenerate sites, arguing that DHSs as a group are subject to purifying selection⁵³.

Alu transposable elements are enriched in GC-rich DNA of primates and rodents, and this can be explained by them landing in open chromatin (e.g. highly expressed regions of the genome)²⁶. The benefit is rapid expression under stress, as *Alu* repeat expression products have been implicated in stress-control of translation^{7,54}.

Yeast experiments demonstrated that *chromatin acetylation* protects DNA from spontaneous mutations by contributing to replication fidelity⁵⁵. In particular, acetylation of H3K56 by Hst3 and Hst4 is required to suppress multiple types of mutations, including base substitutions, small insertions and large-scale rearrangements. Likewise, the frequency of base substitution mutations was negatively correlated with H3K27ac for *de novo* germline mutations in autistic patients⁴² as well as for somatic mutations in cancer¹⁵.

In summary, pairwise studies paint a complex picture of correlation between chromatin and mutation rates, with some studies supporting a link between open chromatin and repressed mutations^{39-42,49-52} potentially due to enhanced repair, while others arguing for a link between closed chromatin and decreased mutations⁴² potentially due to lack of exposure to mutagens. Yet other studies highlight patterns that are base-specific^{40,46}, depend on epigenomic modifications in a genomic region⁴⁷, or are shaped by selection^{46,53}.

Multivariate and segmentation analyses—The analyses presented above explored pairwise relationships between one feature of chromatin organization and mutation rate of one type. However, many genomic features are correlated and so are the rates of different mutation types. Recently, to evaluate the contributions of genomic features to the co-variation in mutation rates more accurately, we conducted a genome-wide *canonical correlation analysis* (CCA)^{56,57} of human-orang-utan divergence variation for four mutation types, together with multiple genomic features characterizing human genome, in 1-Mb windows⁹. In this CCA, variation in the group of genomic features was analyzed along with variation in the group of mutation types, and combinations of features were discovered that correlated well between the two groups. These combinations provided insights into

particular genomic features that could explain variation in rates of specific types of mutations, but within a context that considered all genomic features and mutation types simultaneously. For example, the canonical correlation combination dominated by elevated base substitution rates, also contained genomic features such as, a high number of nuclear lamina binding sites⁴, leading to the inference that base substitution rates are elevated in closed chromatin. This result from the multivariate analysis corroborates some of the results of the pairwise analyses summarized in the previous section³⁹⁻⁴². In contrast, the canonical correlation combination dominated by a correlation in rates of substitution, insertion, and deletion mutations also contained indicators of open chromatin regions such as a low number of nuclear lamina binding sites⁹. This relationship suggests that open chromatin is more prone to these three mutation types when they co-occur in the genome. This example illustrates that, despite the value of pairwise correlations for finding major trends, they could present an oversimplified view of influence of chromatin on mutation rates. Multivariate approaches, which take into account correlations of features and mutation rates, provide a more complex but also more realistic picture of the dynamics of mutations in the genome. Indeed, such approaches are being increasingly applied to the analyses of genomics data – from gene expression analyses⁵⁸ and GWASs⁵⁹ to the integration of ‘omics’ data⁶⁰.

Another way to infer the influence of chromatin organization on mutation rates is via an Hidden Markov Model (HMM)⁶² segmentation of the genome based on rates of different mutation types⁶¹. HMMs have been used extensively in genomics to model stretches of DNA – the sequences of observations – in algorithms predicting genes (the hidden states)⁶³ and more recently to produce segmentations of the genome based on epigenomic signatures⁶⁴. For mutation rate studies, the object was to infer underlying states of high or low rates of various mutations from the observed differences in genome sequence alignments.

Specifically, to apply HMMs in a study of mutation rates, divergence estimates were computed for four mutation types – base substitutions, insertions, deletions, and microsatellite repeat number alterations – in non-overlapping windows of neutrally evolving sequences present in human–orang-utan genomic alignments⁶¹. Modeling the resulting observations with HMMs revealed distinct divergence states characterized by biologically meaningful combinations of elevated, average, or depressed divergence levels for the four mutation types, e.g., a state where all four mutation rates are elevated; one where only one mutation type is elevated, whereas the other divergence types are average, etc. One result of this analysis was a partitioning of the genome into chromosomal segments governed by six states characterized by the incidence of each of four types of mutation (the ‘mutation rate profile’) (**Table 2**). The DNA intervals assigned to each discrete state (mutation rate profile) were then examined for enrichment or depletion in several genomic and epigenomic features. The hot state, characterized by highly elevated insertion, deletion, and substitution rates (also known as IDS++), was found in regions with open chromatin, as such regions had few nuclear lamina sites, but were enriched in DHSs and H3K4me1 marks. This finding appears to contradict an association between elevated substitution rates and closed chromatin found in previous pairwise analyses³⁹⁻⁴² (see section 3.1). In contrast, the warm state characterized by mildly elevated deletion and substitution rates (also known as DS+)

was located in regions with closed chromatin, supporting previous pairwise analyses³⁹⁻⁴² (see section 3.1). Since the mildly elevated DS+ state makes up 18% of the genome while the highly elevated IDS++ state comprises only 8%, this suggests that pairwise results do not provide enough resolution to detect the less common but more nuanced associations. The insertion warm state, I+, was found in the open chromatin regions, while the state with low insertion, deletion, and substitution rates (IDS-) was located in closed chromatin⁶¹. The microsatellite state (M+) did not show any preference to a particular chromatin organization⁶¹. The segmentation analysis also placed different states “geographically” in the genome, with the IDS++ state tending to occupy the tips of the chromosomes while other states were situated closer to the middle of the chromosomes; the M+ state was interspersed (Figure 3).

Chromatin structure and cancer

Chromatin affects mutation landscape in cancer—While the studies described above were based on the analysis of germ-line mutations as inferred from genomic comparisons of primates, strong regional variation was also observed for somatic mutation in cancer^{4,16,17}. Indeed, over five-fold differences in mutation rates were found across the genomes (within chromosomes) of individual tumors⁶⁵. In one of the first attempts to explain this variation, Hodgkinson and colleagues analyzed somatic mutation rates in lung cancer and melanoma and were able to explain approximately 40% of variance in mutation rates by variation in genomic landscape features¹⁶, including GC content, distance to telomere, gene density, nucleosome occupancy, and replication timing. However, this study included only one chromatin-related predictor, nucleosome occupancy (Table 2). The results suggested disease-specific differences: nucleosome occupancy displayed a positive association with mutation rates in melanoma, but a negative one in lung cancer.

In a more detailed study, Schuster-Bockler and Lehner included a large number of chromatin-related features -- 38 different histone modification marks, nucleosome occupancy, and a metric of long-range chromatin interactions as measured by a genome-wide method for chromosome conformation capture, Hi-C. They demonstrated that intra-chromosomal regional variation in mutation rates is affected by chromatin state for many cancer types¹⁵. The heterochromatin-associated repressive histone modification H3K9me3 accounted for >40% of variation. Altogether, chromatin-related features accounted for >55% of variation (Table 2). In fact, elevated mutation rates were strongly positively associated with indicators of closed chromatin for diverse cancers (leukemia, melanoma, small cell lung cancer, and prostate cancer) and diverse substitution mutations (transitions and transversions, and CpG mutations and non-CpG mutations). This corresponds to the segmentation analysis DS+ state that was also associated with closed chromatin⁶¹. To explain a positive association of closed chromatin with elevated substitution rates, several hypotheses were proposed, including differing accessibility to DNA repair complexes, variation in the ability to signal repair, and increased exposure to mutagens at the nuclear periphery¹⁵.

A recent study analyzed the correlation of variation in mutation rates in many cancers with epigenomic features in a large number of cell types, including the parental cells of origin of

the cancers¹⁷. The chromatin organization in the parental cells of origin was able to explain an even higher amount of the mutation rate variation (on average 70%), supporting a positive association between closed chromatin and increased base substitution rate. Furthermore, the authors demonstrated that the mutation profile can be used as a diagnostic feature to aid in the identification of the cell type of origin of the cancer.

Moderate¹⁶ to high⁶⁶ correlations were found between somatic mutation rates in cancer genomes and germline mutation rates approximated by human-chimpanzee divergence. However, chromatin marks play an important role in determining cell fates during embryonic development⁶⁷, but have not been studied in detail in reproductive tissues -- and such marks are particularly important for germ-line mutations. When such data become available, the role of chromatin in determining variation in germ-line mutations might become more prominent⁶¹, just as a recent cancer mutation study demonstrated the value of examining chromatin organization in the most relevant cell types¹⁷. However a simple one-to-one association is not always expected in the factors and patterns for somatic vs. germline mutations. One remarkable example of differences is seen with the X chromosome: it has low germ-line mutation rate because of male mutation bias⁶⁸, but an elevated somatic mutation rate in cancers⁶⁹.

Mutations affect chromatin states genome-wide and can cause cancer—

Although chromatin affects the mutation landscape in some types of cancers, in several other types of cancer, driver mutations with key roles in oncogenesis affect chromatin-remodeling enzymes, leading to a global change in chromatin organization and mutations as compared with that observed for normal cells. For example, genes encoding chromatin-state modifiers such as the histone demethylase *KDM6A* (also known as *UTX*) are frequently mutated in adenoid cystic carcinoma⁷⁰. This can subsequently lead to aberrant epigenomic regulation and affect cell growth⁷⁰. Chromatin modifiers are frequently mutated in medulloblastomas⁷¹ and in small cell lung cancer⁷². Similarly, pediatric glioblastomas⁷³ and some other brain tumors (e.g., diffuse intrinsic pontine glioma and high-grade astrocytoma) are characterized by recurrent driver mutations in *H3F3A* encoding replication-independent histone variant H3.3 (reviewed in⁷⁴). The same gene is mutated in chondroblastoma, chondrosarcoma, and osteosarcoma⁷⁴. It has been shown that mutations in H3.3 may alter either local or global histone methylation patterns (reviewed in⁷⁴). Furthermore, some cancers are characterized by recurrent mutations in genes encoding de novo DNA methyltransferases, DNMT3A and DNMT3B. For instance, mutations in DNMT3A arise early in acute leukemia evolution likely providing a selective advantage^{75,76}. Mutations in DNMT3B are associated with accelerated oncogenesis (e.g., also in acute leukemia) and are known to lead to global perturbed methylation profiles, e.g. hypomethylation⁷⁷. Moreover, rhabdoid tumors, which usually do not harbor many different somatic mutations, are characterized by inactivating mutations in *SMARCB1* encoding a core subunit of the SWI/SNF chromatin remodeling complex⁷⁸. Therefore, mutated chromatin remodeling complexes in some cancers can have the same effect (leading to malignancy) as high mutation rates in other cancers⁷⁹.

Chromatin, gene expression, and mutation

Transcription has been shown to induce strand-specific asymmetry in mutations, particularly in genes expressed at high levels in the germ line^{80,81}. Specifically, there is an excess of G +T over A+C on the coding strand of genes⁸⁰, and this leads to significant G over C and T over A biases in transcribed sequences of genes (importantly, including introns)⁸¹. This mutation signature has been attributed to transcription-coupled repair which may resolve mismatches in DNA prior to the next replication round⁸⁰. Since chromatin states influence transcription, may the process of transcription have some effect on the correlation between mutation and chromatin described above?

As a rule, open chromatin is associated with high transcription rates. Therefore, based on transcription-induced mutation strand bias, in regions with high transcription rates we expect (1) mutation pattern biases; and (2) overall lower substitution rates because transcription-coupled repair is an additional repair mechanism acting there (although selection can also contribute). While the mutation pattern biases have been studied in genes extensively⁸⁰⁻⁸³, they are yet to be investigated with respect to chromatin states on a genome-wide scale using multivariate analyses (e.g., HMMs) for both germ-line and somatic (i.e. cancer) data. The second expectation is consistent with the cancer data. In the two large cancer genome studies^{15,17}, closed chromatin was associated with higher base substitution rates, and conversely, open chromatin was associated with lower base substitution rates. While gene expression levels have not been analyzed, one study also found a negative association between somatic mutation rates and gene density¹⁵. The situation is less clear for germline mutations: despite the association of closed chromatin with higher substitution rates in the abundant DS+ state, coding exons are underrepresented in DNA segments in the DS+ state⁶¹. Gene expression levels need to be studied to arrive at a clearer conclusion regarding a potential link among chromatin, transcription, and mutation in the germ line.

Conclusions and perspectives

Chromatin structure and organization varies across the genome in patterns reflecting the expression and activity of the underlying DNA. Recent studies have shown that the chromatin organization is also strongly correlated with the variation in mutation rates, both germ-line and somatic. For some of the most frequently occurring mutations, such as single base substitutions, their rates are suppressed in open, transcriptionally active chromatin. This supports a model that DNA in more open chromatin is either mutated less frequently or repaired more frequently, perhaps associated with its earlier time of replication during the S phase of the cell cycle.

While less frequent mutation in open chromatin is a dominant pattern for some types of sequence alteration, it does not apply to the entire mutational spectrum. Genome-wide multivariate segmentation analyses of chromatin structure and mutation rates have revealed a more complex, nuanced, and multi-faceted picture. The association between chromatin structure and mutation rates, as assayed with multivariate analyses, shows striking heterogeneity along the genome, with some regions with elevated germline mutation rates associated with closed or open chromatin. It will be of great interest to examine whether

such heterogeneous associations exist also in cancer genomes by applying multivariate statistical analyses (e.g., CCA and HMMs) to the cancer mutation data.

Importantly, the association between chromatin structure and mutation rates point to candidates for key determinants of the mutation rate variation, such as alterations in enzymes modulating chromatin state in certain cancers. Experimental manipulations of these enzyme complexes can test these hypotheses. In general, more studies of mechanistic determinants of the mutation rate variation are needed.

The analyses of RViMR provide important information about genome function. Indeed, in the segmentation analysis of the genome, hot (IDS++) and insertion-warm (I+) states located in open chromatin were found to be significantly enriched in genes and regulatory elements, while depleted in functionally inactive regions (quiescent zones)⁶¹. Thus, mutationally active and transcriptionally active regions in the genome frequently coincide. This supports a recent assertion that the chromatin organization establishes functional and spatial biases on specific regions of the genome^{13,61}, and opens up a possibility that variation in mutation rate is adaptive¹³, although this has been debated⁴.

Current approaches and data sets do have deficiencies, and additional, broader studies are needed to ascertain how robust these results are. For instance, future studies should examine the influence of biased gene conversion²⁴ on the relationship between mutation rates and chromatin. However, it is already clear that the associations between mutation rates and chromatin point to a complex relationship between the DNA sequence and its dynamic packaging into nucleosomes and higher order structures¹⁸. Further study should illuminate how the packing and expression of DNA in the germline and during various stages of development influences the rates of all types of mutations. Such studies would benefit from a more finely grained and precise definition of chromatin states, just as investigation of different types of mutations has helped resolve complex relationships. More complete insights into the connections between chromatin states and mutation rate variation will improve our understanding of fundamental evolutionary processes in the genome and provide important information about disease processes, such as cancer and autism.

Box 1

Mutation types and their variation in rates/patterns

Depending on their *effect on DNA structure*, mutations can be classified into:

- **base substitutions, or nucleotide substitutions**: the replacement of one nucleotide by the other;
- **deletions**: the removal of one or several nucleotides;
- **insertions**: the addition of one or several nucleotides;
- **microsatellite mutations** (a subtype of indels): deletions or insertions of motif units within tandemly repeated DNA;
- **inversions**: DNA rearrangement in which a fragment changes its orientation by 180°

- **translocations**: the movement of a DNA fragment to another location in the genome.

Depending on their *scale*, mutations can be classified into

- **point mutations** (affecting one nucleotide, e.g., nucleotide substitutions and indels affecting a single nucleotide belong to this category);

- **small-scale mutations** (affecting several nucleotides);

- **large-scale mutations** (involving a larger chromosomal region);

- **aneuploidies** (insertions or deletions of a whole chromosome);

- **whole-genome polyploidies** (duplications involving a whole genome).

Transposable Element (TE) insertions and Segmental Duplications are common large-scale insertions, and are described in more detail below.

TE insertions represent a common type of mammalian mutations. These are insertions of SINEs (*Alus* in humans), LINEs, ERVs, and DNA transposons. The insertions of the former three groups of TEs are insertions of 300 bp – 9 kb, while insertions of DNA transposons technically should be considered translocations as they move by the cut-and-paste mechanism. In humans, only *Alus*, L1s, and ERVs are currently active^{8,26}.

Segmental duplications are another common type of mutations in mammalian genomes. The signature of such mutations is the presence of relatively large (1->200-kb) nearly identical DNA segments present in at least two copies in the genome that are thought to originate from duplicative transposition of DNA^{84,85}. The actual number of copies frequently varies among individuals leading to Copy Number Variants (CNVs). The RViMR for segmental duplications has been understudied. It is known that subtelomeric regions are enriched for segmental duplications²⁵. To our knowledge, no studies have been published on the effects of chromatin organization on the RViMR for segmental duplications.

Acknowledgements

KDM is supported by the US National Science Foundation grant DBI-0965596, and RCH is supported by the US National Institutes of Health grants R01DK065806, RC2HG005573, and U54HG006998. The authors are grateful to Prabhani Kuruppumullage Don for help with Figure 1.

Biographies

Kateryna Makova is the Pentz Professor in the Department of Biology at Penn State University. Her lab uses both computational and experimental approaches to study mutations. Of particular interest are topics on regional variation in mutation rates, mitochondrial mutations, microsatellites and genome stability, and mutations at sex chromosomes.

Ross Hardison is the T. Ming Chu Professor of Biochemistry and Molecular Biology at Penn State University. His laboratory studies gene regulation, with a special emphasis on

hematopoiesis, by applying genomic technologies to gain evolutionary insights from comparative genomics and mechanistic insights from functional genomics.

GLOSSARY

Canonical correlation analysis	Statistical analysis considering two groups of variables simultaneously and finding significant linear combinations between them that have maximum correlations with each other
Chromatin acetylation	Chromatin in which specific lysine residues within the N-terminal tails of histones have been covalently modified by the addition of an acetyl group
Chromosome conformation capture	A method for quantitatively estimating the frequency of interaction between two different genomic regions using a crosslinking and intermolecular ligation assay to identify interacting sites
CpG dinucleotides	Positions in the DNA sequence in which a cytosine nucleotide (C) is followed by a guanine nucleotide (G)
Epigenomic features	Biochemical features that are associated with genomic DNA sequences but are not the sequence themselves; examples include DNA methylation, histone modifications in the chromatin packaging the DNA, nuclease accessibility, and transcription factor binding
Genomic landscape features	Features that characterize the genome at levels beyond the simple primary DNA sequence and include GC content, recombination rates, proximity to the closest telomere, replication timing, etc
Hidden Markov Models	Statistical models that analyze a sequence of observations defined by underlying states that are not observable (“hidden”) but can be inferred from the data; these states alternate along the sequence following a Markovian structure, i.e., the state defining a given observation depends on the state governing the preceding observation
Indels	Insertions and deletions
Open chromatin	Chromatin in which the DNA is readily accessible to enzymes in the cell; it can be interpreted as regions with less compaction than bulk nucleosomes, depleted of nucleosomes, or having highly remodeled nucleosomes
Nucleosome occupancy	Packaging of DNA into a nucleosome, with the DNA wrapped tightly around a core of eight histone molecules; such DNA wrapped around the histone core is considered “occupied” by a nucleosome

**Regional variation
in mutation rates
(RViMR)**

The phenomenon of rate of mutation changing along individual chromosomes

REFERENCES

1. Wolfe KH, Sharp PM, Li WH. Mutation rates differ among regions of the mammalian genome. *Nature*. 1989; 337:283–285. [PubMed: 2911369]
2. Makalowski W, Boguski MS. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J Mol Evol*. 1998; 47:119–121. [PubMed: 9694659]
3. Hardison RC, et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res*. 2003; 13:13–26. An early genome-wide study illustrating not only regional variation but also regional co-variation among mutation rates of different types. [PubMed: 12529302]
4. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. *Nature reviews*. 2011; 12:756–766. An excellent review on regional variation in mutation rates.
5. Chiaromonte F, et al. Association between divergence and interspersed repeats in mammalian noncoding genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America*. 2001; 98:14503–14508. [PubMed: 11717405]
6. Kvikstad EM, Tyekucheva S, Chiaromonte F, Makova KD. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS computational biology*. 2007; 3:1772–1782. [PubMed: 17941704]
7. Yang S, et al. Patterns of insertions and their covariation with substitutions in the rat, mouse, and human genomes. *Genome Res*. 2004; 14:517–527. [PubMed: 15059992]
8. Kvikstad EM, Makova KD. The (r)evolution of SINE versus LINE distributions in primate genomes: sex chromosomes are important. *Genome Res*. 2010; 20:600–613. [PubMed: 20219940]
9. Ananda G, Chiaromonte F, Makova KD. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome biology*. 2011; 12:R27. [PubMed: 21426544]
10. Wilson Sayres MA, Makova KD. Genome analyses substantiate male mutation bias in many species. *Bioessays*. 2011; 33:938–945. [PubMed: 22006834]
11. Campos-Sanchez R, Kapusta A, Feschotte C, Chiaromonte F, Makova KD. Genomic Landscape of Human, Bat, and Ex Vivo DNA Transposon Integrations. *Molecular biology and evolution*. 2014
12. Kimura M. Evolutionary rate at the molecular level. 1968; 217:624–626. 1968.
13. Chuang JH, Li H. Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS biology*. 2004; 2:E29. [PubMed: 14966531]
14. Li J, Miller W. Significance of interspecies matches when evolutionary rate varies. *J Comput Biol*. 2003; 10:537–554. [PubMed: 12935343]
15. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*. 2012; 488:504–507. This comprehensive analysis links chromatin states with base substitution mutations in cancer genomes. [PubMed: 22820252]
16. Hodgkinson A, Chen Y, Eyre-Walker A. The large-scale distribution of somatic mutations in cancer genomes. *Hum Mutat*. 2012; 33:136–143. [PubMed: 21953857]
17. Polak P, et al. Cell type of origin chromatin organization shapes the mutational landscape of cancer. *Nature*. 2014 (in press). This is a large-scale analysis of the association between mutation rates and chromatin states in multiple cancers and cell types.
18. Hellman A, Chess A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. *Epigenetics Chromatin*. 2010; 3:11. [PubMed: 20497546]
19. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
20. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012; 337:1190–1195. [PubMed: 22955828]

21. Stamatoyannopoulos JA, et al. Human mutation rate associated with DNA replication timing. *Nat Genet.* 2009; 41:393–395. [PubMed: 19287383]
22. Tyekucheva S, et al. Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome biology.* 2008; 9:R76. [PubMed: 18447906]
23. Hellmann I, et al. Why do human diversity levels vary at a megabase scale? *Genome Res.* 2005; 15:1222–1231. [PubMed: 16140990]
24. Schaibley VM, et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res.* 2013; 23:1974–1984. [PubMed: 23990608]
25. Linardopoulou EV, et al. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature.* 2005; 437:94–100. [PubMed: 16136133]
26. Wagstaff BJ, et al. Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. *PLoS genetics.* 2012; 8
27. Ehrlich M, Wang RY. 5-Methylcytosine in eukaryotic DNA. *Science.* 1981; 212:1350–1357. [PubMed: 6262918]
28. Gaffney DJ, Keightley PD. The scale of mutational variation in the murid genome. *Genome Res.* 2005; 15:1086–1094. [PubMed: 16024822]
29. Laurent L, et al. Dynamic changes in the human methylome during differentiation. *Genome Res.* 2010; 20:320–331. [PubMed: 20133333]
30. Cairns BR. Chromatin remodeling: insights and intrigue from single-molecule studies. *Nat Struct Mol Biol.* 2007; 14:989–996. [PubMed: 17984961]
31. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews.* 2011; 12:7–18.
32. Beisel C, Paro R. Silencing chromatin: comparing modes and mechanisms. *Nature reviews.* 2011; 12:123–135.
33. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. *Nature.* 2007; 447:413–417. [PubMed: 17522674]
34. Vanin EF, Henthorn PS, Kioussis D, Grosveld F, Smithies O. Unexpected relationships between four large deletions in the human beta-globin gene cluster. *Cell.* 1983; 35:701–709. [PubMed: 6652684]
35. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science.* 2002; 295:1306–1311. [PubMed: 11847345]
36. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 2009; 326:289–293. This paper presents genome-wide mapping of DNA interaction frequencies. [PubMed: 19815776]
37. Kieffer-Kwon KR, et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell.* 2013; 155:1507–1520. [PubMed: 24360274]
38. Filion GJ, et al. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. *Cell.* 2010; 143:212–224. [PubMed: 20888037]
39. Prendergast JG, et al. Chromatin structure and evolution in the human genome. *BMC evolutionary biology.* 2007; 7:72. [PubMed: 17490477]
40. Chen X, et al. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science.* 2012; 335:1235–1238. [PubMed: 22403392]
41. Sasaki S, et al. Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science.* 2009; 323:401–404. [PubMed: 19074313]
42. Michaelson JJ, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell.* 2012; 151:1431–1442. [PubMed: 23260136]
43. Taylor J, Tyekucheva S, Zody M, Chiaromonte F, Makova KD. Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Molecular biology and evolution.* 2006; 23:565–573. [PubMed: 16280537]
44. Cohen NM, Kenigsberg E, Tanay A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell.* 2011; 145:773–786. [PubMed: 21620139]
45. Fei J, Ha T. Watching DNA breath one molecule at a time. *Proceedings of the National Academy of Sciences of the United States of America.* 2013; 110:17173–17174. [PubMed: 24096577]

46. Prendergast JG, Sempile CA. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res.* 2011; 21:1777–1787. [PubMed: 21903742]
47. Tolstorukov MY, Volfovsky N, Stephens RM, Park PJ. Impact of chromatin structure on sequence variability in the human genome. *Nat Struct Mol Biol.* 2011; 18:510–515. [PubMed: 21399641]
48. Tang Y, et al. H2A.Z nucleosome positioning has no impact on genetic variation in *Drosophila* genome. *PLoS One.* 2013; 8:e58295. [PubMed: 23472174]
49. Warnecke T, Batada NN, Hurst LD. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS genetics.* 2008; 4:e1000250. [PubMed: 18989456]
50. Washietl S, Machne R, Goldman N. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.* 2008; 24:583–587. [PubMed: 18951646]
51. Ying H, Epps J, Williams R, Huttley G. Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Molecular biology and evolution.* 2010; 27:637–649. [PubMed: 19843619]
52. Ying H, Huttley G. Exploiting CpG hypermutability to identify phenotypically significant variation within human protein-coding genes. *Genome Biol Evol.* 2011; 3:938–949. [PubMed: 21398426]
53. Thurman RE, et al. The accessible chromatin landscape of the human genome. *Nature.* 2012; 489:75–82. [PubMed: 22955617]
54. Lander ES, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001; 409:860–921. [PubMed: 11237011]
55. Kadyrova LY, et al. A reversible histone H3 acetylation cooperates with mismatch repair and replicative polymerases in maintaining genome stability. *PLoS genetics.* 2013; 9:e1003899. [PubMed: 24204308]
56. Everitt, BS. *An R and S-Plus Companion to Multivariate Analysis.* Springer; 2005.
57. Mardia, KV.; Kent, JT.; Bibby, JM. *Multivariate analysis.* Academic Press; 1979.
58. Soneson C, Lilljebjorn H, Fioretos T, Fontes M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC bioinformatics.* 2010; 11:191. [PubMed: 20398334]
59. van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS genetics.* 2013; 9:e1003235. [PubMed: 23359524]
60. Gonzalez I, Cao KA, Davis MJ, Dejean S. Visualising associations between paired 'omics' data sets. *BioData Min.* 2012; 5:19. [PubMed: 23148523]
61. Kuruppumullage Don P, Ananda G, Chiaromonte F, Makova KD. Segmenting the human genome based on states of neutral genetic divergence. *Proceedings of the National Academy of Sciences of the United States of America.* 2013; 110:14699–14704. This study presents segmentation of the human genome based on states of neutral genomic divergence used as a proxy of germ-line mutation rate. [PubMed: 23959903]
62. Eddy SR. What is a hidden Markov model? *Nature Biotechnology.* 2004; 22:1315–1316.
63. Majoros WH, Pertea M, Antonescu C, Salzberg SL. *Nucleic acids research.* 2003; 31:3601–3604. [PubMed: 12824375]
64. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature.* 2011; 473:43–U52. [PubMed: 21441907]
65. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013; 499:214–218. [PubMed: 23770567]
66. Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun.* 2013; 4:1502. [PubMed: 23422670]
67. Cheedipudi S, Genolet O, Dobrev G. Epigenetic inheritance of cell fates during embryonic development. *Front Genet.* 2014; 5:19. [PubMed: 24550937]
68. Wilson Sayres MA, Venditti C, Pagel M, Makova KD. Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution.* 2011; 65:2800–2815. [PubMed: 21967423]

69. Davoli T, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*. 2013; 155:948–962. [PubMed: 24183448]
70. Ho AS, et al. The mutational landscape of adenoid cystic carcinoma. *Nat Genet*. 2013; 45:791–798. [PubMed: 23685749]
71. Jones DT, et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature*. 2012; 488:100–105. [PubMed: 22832583]
72. Peifer M, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat Genet*. 2012; 44:1104–1110. [PubMed: 22941188]
73. Schwartzentruber J, et al. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature*. 2012; 482:226–231. [PubMed: 22286061]
74. Lan F, Shi Y. Histone H3.3 and cancer: A potential reader connection. *Proceedings of the National Academy of Sciences of the United States of America*. 2014
75. Im AP, et al. DNMT3A and IDH mutations in acute myeloid leukemia and other myeloid malignancies: associations with prognosis and potential treatment strategies. *Leukemia*. 2014; 28:1774–1783. [PubMed: 24699305]
76. Shlush LI, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*. 2014; 506:328–333. [PubMed: 24522528]
77. Walton EL, Francastel C, Velasco G. Dnmt3b Prefers Germ Line Genes and Centromeric Regions: Lessons from the ICF Syndrome and Cancer and Implications for Diseases. *Biology (Basel)*. 2014; 3:578–605. [PubMed: 25198254]
78. Lee RS, Roberts CW. Rhabdoid tumors: an initial clue to the role of chromatin remodeling in cancer. *Brain Pathol*. 2013; 23:200–205. [PubMed: 23432645]
79. Lee RS, et al. A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *J Clin Invest*. 2012; 122:2983–2988. [PubMed: 22797305]
80. Green P, Ewing B, Miller W, Thomas PJ, Green ED. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet*. 2003; 33:514–517. [PubMed: 12612582]
81. Louie E, Ott J, Majewski J. Nucleotide frequency variation across human genes. *Genome Res*. 2003; 13:2594–2601. [PubMed: 14613976]
82. Polak P, Arndt PF. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res*. 2008; 18:1216–1223. [PubMed: 18463301]
83. Mugal CF, von Grunberg HH, Peifer M. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Molecular biology and evolution*. 2009; 26:131–142. [PubMed: 18974087]
84. Bailey JA, et al. Recent segmental duplications in the human genome. *Science*. 2002; 297:1003–1007. [PubMed: 12169732]
85. Samonte RV, Eichler EE. Segmental duplications and the evolution of the primate genome. *Nature reviews*. 2002; 3:65–72.
86. Kelkar Y, Tyekucheva S, Chiaromonte F, Makova KD. The genome-wide determinants of microsatellite evolution. *Genome Res*. 2008; 18:30–38. [PubMed: 18032720]

ONLINE SUMMARY

1. Regional variation in mutation rates is an important phenomenon affecting genome evolution. It is determined by features of genomic landscape, with chromatin having a significant influence.
2. Pairwise studies have painted a complex picture of correlation between chromatin and mutation rates, with some studies supporting a link between open chromatin and repressed mutations, while others argue for a link between closed chromatin and decreased mutations. Still other studies highlight patterns that are base-specific, depend on epigenomic modifications in a genomic region or are shaped by selection.
3. Because features characterizing chromatin states are correlated among themselves and with other genome landscape features, multivariate segmentation (HMM) analyses are providing a more nuanced depiction of the relationship between chromatin and germline mutation rates. Specifically, a prevalent genomic state with moderately high substitution and deletion rates is located in regions with closed chromatin, while a less abundant state with very high substitution, insertion, and deletion rates is located in regions with open chromatin.
4. Several recent studies indicate a positive association between elevated somatic mutation rates and closed chromatin in cancer genomes.
5. In several types of cancer, driver mutations are located in genes regulating chromatin, leading to the hypothesis that consequent global or local chromatin remodeling results in malignancy.
6. Transcription of genes is influenced by chromatin state and leads to a biased substitution pattern likely due to transcription-coupled repair.

Figure 1. Variability in rates of base substitutions, small insertions, and small deletions (as inferred from human-orang-utan genomic alignments), plotted together with densities of *Alu* and L1 elements, along human chromosome 1

The Y axis is the number of small (<30-bp) insertions per site, the number of small (<30-bp) deletions per site, the number of base substitutions per site, the number of *Alus*, and the number of L1s, respectively, measured for 1-Mb windows.

Figure 2. Aspects of chromatin organization that can impact evolutionary rates

A portion of a chromatin fiber is shown to illustrate closed *versus* open chromatin and the different types of mutations that occur at higher or lower rates in each. The closed state can represent quiescent chromatin with little dynamic histone modification or with the repressive histone modifications H3K9me3 (associated with heterochromatin) or H3K27me3. Actively transcribed and regulated DNA tends to be in open chromatin marked by DNase hypersensitive sites, transcription factor occupancy, and activating histone modifications such as H3K4me1 (associated with enhancers), H3K4me3 (associated with promoters), H3K36me3 (associated with transcribed chromatin), H3K27ac and H4ac (both associated with enhancers and promoters). D, deletion; I, insertion; indel, insertion and deletion; Pol II, RNA polymerase II; S, substitution.

Figure 3. The distribution of mutation rate variation states across a typical autosome and genomic landscape features that characterize chromatin

Distribution of mutation rate variation states are obtained from REF. 61. D, deletion; DHS, DNase-hypersensitive site; H3K4me1, histone H3 lysine 4 monomethylation; I, insertion; S, substitution.

Table 1
Associations between different mutation types and genomic landscape features in interspecies comparisons

Mutation types (species compared)	G C co nt en t	X vs A [*]	Ex on de nsi ty	SN P de nsi ty	Ma le Re Xn	Fe ma le Re Xn	ReX ho ts po t	Dis tan ce to TE L	S I N E	L I N E	Cp G Is d	Pol y (A/ T)	Rep eat mbr	Gen ome inst abi lity 13-mer	TE L 6- mer	Total variabil ity explan ed (%)	Refs
Substitutions at CpG (human-macaque)	+	+														82	22
Substitutions at non CpG (human- macaque)	+	+	+	+	+			+								52	22
Deletions (human- chimpanzee)	+	+			+			+	+	+	+					32	6
Insertions (human- chimpanzee)	+	+	+	+	+	+		+	+	+	+	+				27	6
Young Alu repeats (human- chimpanzee-orang- utan-macaque)	+						+	+	+	+				+	+	20	8
Young LINEs (human- chimpanzee-orang- utan-macaque)	+						+			+	+				+	41	8
Microsatellite mutations (human- chimpanzee)													+			89	86

Plus marks indicate significant associations as found from multiple regressions (see original publications for *P* values). LINE, long interspersed element; SINE, short interspersed element; SNP, single-nucleotide polymorphism.

^{*} Location on the X chromosome versus autosomes (a categorical predictor).

Table 2

Associations among genomic and epigenomic features, and mutation rate profiles

	Germline mutations						Somatic mutations in cancer		
	IDS ⁺⁺	DS ⁺	I ⁺	IDS ⁻	IDS ⁻⁻	M ⁺	S	S	S
Proportion of genome (%)	8.0	18.1	27.8	36.3	4.4	5.4	NA	NA	NA
Chromosomal location	Subtelomeric	Subtelomeric	Subtelomeric	Internal	X chromosome	Interspersed	NA	NA	NA
GC content	Very high (45%)	Low (38%)	High (42%)	Avg (38%)	Avg (39%) (avg)	Avg (40%)	Quadratic relationship	Neg	NA
Replication timing	Early	Late	Early	Avg	Avg	Avg	Pos	Neg	Late
Recombination[*]	High; male>>>female	High; male>female	High; female>male	Low	Low; female only	Avg	NA	Neg	NA
Open chromatin	Pos	Neg	Pos	Neg	Neg	Avg	Neg and pos [‡]	Neg	Neg
Closed chromatin	Neg	Pos	Neg	Pos	Pos	Avg	Neg and pos [‡]	Pos	Pos
H3K4me1	High	Low	High	Avg	Low	Avg	NA	Neg	Neg
DHSs	High	Low	High	Avg	Low	Avg	NA	NA	Low
Nuclear lamina	Low	High	Low	Avg	Avg	High	NA	NA	NA
Methyl- CpG	Low	High	Low	High	Avg	Avg	NA	NA	NA
Transcription potential (exons)	High	Low	High	Avg	Moderately low	Avg	Neg	Neg	NA
Total variability explained (%)	NA	NA	NA	NA	NA	NA	40	>55	74–86
Refs	61	61	61	61	61	61	16	15	17

Avg, average; D, deletion; DHS, DNase-hypersensitive site; DS⁺, mutational ‘warm’ state with mildly elevated deletion and substitution rates; H3K9me3, histone H3 lysine 3 trimethylation; I, insertion; I⁺, insertion warm state; IDS⁺⁺, ‘hot’ state with highly elevated insertion, deletion and substitution rates; IDS⁻, mutational state with low insertion, deletion and substitution rates; NA, not applicable; M⁺⁺, microsatellite state; neg, negative association; pos, positive association; S, base substitution.

^{*} Male and female indicate male- and female-specific recombination rates, respectively.

[‡] Depends on the cancer type.