Modeling the Tissue-Specific Somatic Mutation Rate Along the Genome Based on Genomic Features

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Corinna Lewis Schmalohr

aus Bonn

angenommen im Jahr 2025

Abstract

Somatic mutations, arising from unresolved DNA damage, play a critical role in driving cancer development and progression. Previous studies have demonstrated that mutation rates vary throughout the genome and are affected by large-scale genomic determinants. However, they frequently overlooked important genomic features or lacked the resolution to thoroughly examine changes that are particular to different tissues and mutation types.

To address these limitations, we developed a base-pair resolution model to predict somatic mutation rates in the exome. Using cancer mutation datasets and a diverse set of genomic features, we trained and compared several predictive models, including Random Forest, Generalized Linear Models, and LASSO with stability selection. Random Forest performed the best among them and was selected for the majority of analyses. Our findings highlight robust predictive performance, with improved accuracy for specific tissues and mutation types. Key predictors of mutation rate included GC content, replication timing, DNA methylation, histone marks (H3K27ac, H3K4me3, and H3K9ac), RNA expression, transcription factor binding site density, and eQTL annotations. These results underscore the central role of characteristics linked to transcriptional activity in determining local mutation rates.

Remarkably, our models showed a high degree of tissue similarity, and tissue-specific models could be transferred between tissues without losing their predictive power. This finding suggested that the same mutational mechanisms are at play across tissues, enabling the use of a single, generalized model to predict mutation rates effectively across tissues. Extending the approach to the whole genome demonstrated that intergenic areas are subject to the same mutational processes as exonic regions. The models were validated on data from healthy tissues, further supporting their broad applicability.

This study provides a detailed and comprehensive characterization of somatic mutational patterns, leveraging base-pair resolution and an extensive array of genomic predictors. These insights advance our understanding of mutation processes and have the potential to enhance tumor evolution models and driver mutation discovery.

Contents

1 Introduction

1.1 Somatic mutations

Somatic mutations are changes in the DNA sequence that occur in somatic tissues, including single nucleotide variants (SNVs), short insertion or deletions (indels), or even large structural variations such as copy number variations (CNVs). This thesis will focus on SNVs, which represent the most frequent type of somatic mutation (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Somatic mutations are distinguished from germline variants by the fact that they do not affect the germ cells and are thus not passed on to the next generation.

The net occurrence of somatic mutations in tissues is the result of the interplay between the rate of *de novo* mutation appearance versus their repair (Gonzalez-Perez et al., 2019). Furthermore, selection also plays a significant role in the number and type of somatic mutations that can be observed in any tissue (Section 1.7.11; Martincorena et al., 2017). Importantly, the rate of somatic mutation along the genome is not constant, meaning that some genomic regions are more prone to accumulate somatic changes than others (Supek and Lehner, 2019).

Somatic mutations play a major role in the emergence of cancer and thus have been extensively studied in recent years (Section 1.6). Furthermore, somatic mutations were also implicated to play a role in aging-associated degeneration and other diseases (Section 1.4; Vijg and Dong, 2020). Surprisingly, it was recently discovered that somatic mutations are also highly prevalent in normal, apparently healthy tissues, but their impact and potentially physiological role in healthy tissues remains poorly understood (Section 1.6; Moore et al., 2021; Li et al., 2021).

1.2 DNA damage and repair in somatic cells

The cells in our bodies are constantly subjected to DNA damage due to both endogenous and exogenous factors (Chatterjee and Walker, 2017). Accordingly, there are multiple molecular repair machineries in the cell that constantly screen the DNA for damage and try to repair it in order to maintain a functional genome. Failure to revert this DNA damage may lead to functional deficiencies and, in replicating cells, will lead to a permanent sequence change in the daughter cells. Below follows an overview of the different kinds of DNA damage that can arise, how they are typically repaired, and their effect when this repair is not successful (Figure 1). The contribution of each of these mutational processes differs between tissues and cell types, depending on environmental factors and, for example in the context of cancer, possibly disabled repair mechanisms (Alexandrov et al., 2020).

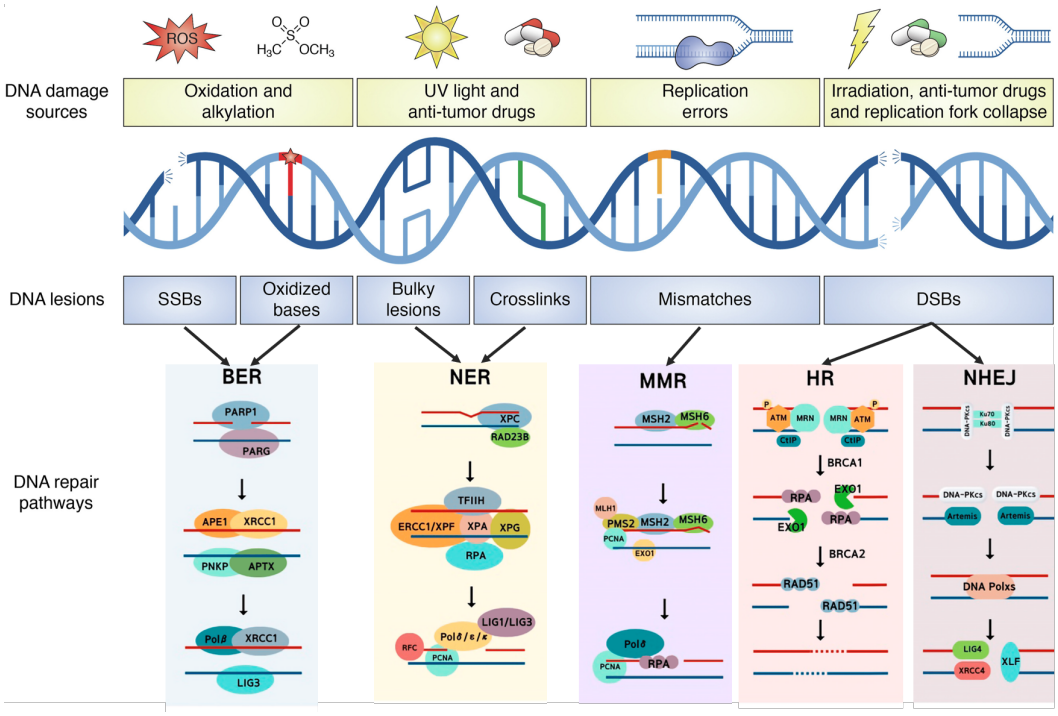

Figure 1: DNA damage and repair pathways. Schematic representation of various endogeneous and exogeneous sources of DNA damage and which DNA repair pathways are typically activated to remove them. Figure adapted from Dall'Agnese et al. (2023) and Moon et al. (2023). Abbreviations: reactive oxygen species (ROS), Ultraviolet (UV), single strand break (SSB), double strand break (DSB), base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), homologous recombination (HR), non-homologous end joining (NHEJ).

One of the most prevalent origin of endogenously caused mutations are errors during DNA replication (Seplyarskiy et al., 2019; Kunkel, 2009). When a cell divides, its DNA is duplicated by high-fidelity DNA polymerases. These polymerases have proof-reading capabilities to ensure that the correct base complementary to the template strand is incorporated (Kunkel, 2009). In addition, the mismatch repair (MMR) machinery corrects most of the replication errors missed by the polymerase proof-reading (Kunkel, 2011). Together, these mechanisms remove about 99% of DNA replication errors. Nevertheless, base substitutions and indels happen frequently at every cell division. The error rates for DNA replication were estimated to be about one wrong nucleotide out of 10^8 to 10^{10} replicated nucleotides, which corresponds to up to about 30 mutations in the human genome (about 3 Giga base pairs) at every cell division (Bębenek and Ziuzia-Graczyk, 2018).

Another major source of endogenously caused mutations is the spontaneous deamination of bases (De Bont and van Larebeke, 2004). The nucleobases adenine (A), cytosine (C), guanine (G) and methylated cytosine (5mC) lose

an amine group to become uracil (U), hypoxanthine, xanthine, and thymine (T), respectively, which, if not corrected, leads to single base substitutions in the DNA sequence. Spontaneous deamination of 5mC represents the largest contributor to SNV occurrence, leading to C>T base substitutions at CpG sites throughout the genome (Tomkova and Schuster-Böckler, 2018; Alexandrov et al., 2020).

Reactive oxygen species (ROS) are oxide radicals such as superoxide ($O_2^{\bullet -}$), hydroxyl radicals ($HO^{\bullet}$), and hydrogen peroxide (H_2O_2), that are present in all cells as byproducts of oxygen metabolism (Bayr, 2005). The amount of ROS in a cell is normally tightly controlled because they are highly reactive and can oxidize DNA or RNA, as well as lipid fatty acids and amino acids in proteins. ROS attack DNA double bonds, remove hydrogens from methyl groups, or attack sugar residues in the DNA backbone, leading to chemically modified bases, single strand breaks (SSBs) or bulky lesions such as adducts and crosslinks (Breen and Murphy, 1995). The most commonly oxidized base is 8-oxoG, which mispairs with adenine and thus can lead to a G>T base exchange during replication, if not corrected (Cheng et al., 1992).

Apurinic or apyrimidic sites occur when the N-glycosyl bond between nucleobase and the backbone breaks (Chatterjee and Walker, 2017). This can happen either through spontaneous hydrolyzation, or through cleavage by a DNA glycosylase, for example during the base excision repair (BER) pathway. Such abasic sites destabilize the DNA and can lead to SSBs (Boiteux and Guillet, 2004).

Another source of somatic mutation is APOBEC-associated mutagenesis. Proteins of the APOBEC (Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) family are cytidine deaminases that are responsible for C-to-U editing of mRNA during RNA processing and as a form of protection against viral RNA. However, they also sometimes erroneously deaminate transiently single-stranded DNA, leading to C>U changes, which result in C>T mutations upon DNA replication (Harris et al., 2002; Nik-Zainal et al., 2012; Kazanov et al., 2015). This APOBEC-driven mutation is highly prevalent in many cancer types, contributing more than ten mutations per Megabase for some cancer types (Alexandrov et al., 2020).

Aside from these endogeneous mutation sources, exogeneous mutagens act on our tissue and cause DNA damage. For instance, Ultraviolet (UV) light, which our skin is constantly exposed to, leads to covalent linkages between two adjacent pyrimidines, especially at two adjacent cytosines (Hu et al., 2017). These bulky, helix-distorting lesions can either be directly repaired by chemical reversal of the UV-damaged bases or through nucleotide excision repair (NER). Alternatively, specialized translesion synthesis polymerases can replicate past such bulky lesions with reduced fidelity (Vaisman et al., 2012),

leading to C/G>T/A, T/A>C/G, and, most commonly, CC>TT variants. Similarly, ionizing radiation, for example from X-rays, can damage the DNA in various ways, for example through the creation of free radicals, leading to mutations similar to those caused by ROS, as well as various chemical base changes or SSBs (Chatterjee and Walker, 2017).

Furthermore, our bodies are exposed to a wide variety of substances that have mutagenic effects. These include alkylating agents, aromatic amines, or polycyclic aromatic hydrocarbons, which are present in our diet (e.g., charred food, N-nitrosamines in preserved meats), in the air (e.g., tobacco smoke, air pollution), chemotherapeutics, or other environmental sources (Kucab et al., 2019). These chemicals intercalate, react with, and/or directly damage our DNA in various ways, leading to a wide array of chemically modified bases, bulky lesions and/or single and double strand breaks (DSBs) in our DNA. Many of these mutagens lead to characteristic mutations at specific sequence contexts if not properly repaired (Section 1.7; Kucab et al., 2019).

In summary, there is a myriad of endogenous and exogenous factors leading to DNA damage. Therefore, our cells constantly have to screen our DNA for damage, signal its presence and promote its repair via a pathway called the DNA damage response. Failure to repair the damage, for example in the case of excessive DNA damage, will lead to cell cycle arrest or even apoptosis (Chatterjee and Walker, 2017). There are five major DNA repair pathways: mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER), homologous recombination (HR), and non-homologous end joining (NHEJ) (Chatterjee and Walker, 2017).

Small base modifications, including deaminated or oxidized bases described above, as well as single strand breaks, are typically mended via the BER pathway (Dianov and Hübscher, 2013). Therein, glycosylases specifically recognize modified bases and remove them, leading to an abasic site. This is then repaired either in the so-called short patch repair, where the base-less backbone is removed and replaced with the correct nucleobase, or in the long patch repair, where, through displacing synthesis, a larger area around the lesion is re-synthesized, the old strand is removed, and the gap is ligated.

Bulky lesions and helix-distorting adducts or crosslinks are handled by NER. The damaged DNA is detected through two possible sub-pathways, global excision repair and transcription-coupled repair. The former senses transient single-strand DNA caused by disrupted base pairing due to the lesion, while the latter is activated by stalled RNA polymerase II. Then, the NER machinery unwinds the DNA helix around the lesion and removes the damaged strand. Finally, the gap is filled by DNA polymerase and ligase (Marteijn et al., 2014).

The MMR machinery is responsible for the correction of base mismatches and indels, especially those caused by faulty DNA replication. This is achieved through the specific recognition of mismatched bases by proteins from the MutS group of DNA MMR proteins, their excision (involving MutL proteins), and subsequent re-synthesis and re-ligation of the resulting gap (Kunkel and Erie, 2005).

DSBs are repaired via NHEJ or HR. In NHEJ, microhomologies between the overhanging loose ends of the two ends of double strand break are used to directly repair the break. In HR, the homologous sister chromatid is used as a template to re-synthesize the broken DNA strand (Weterings and Chen, 2008).

Unresolved DNA damage or incorrectly repaired DNA lesions ultimately lead to a somatic mutation. Thus, the collection of somatic mutations present in any cell is the result of the combination of DNA lesion accumulation and failure to repair this damage. The type of mutation depends on the type of lesion, with some mutagens and/or faulty repair processes resulting a specific signature of mutations (discussed below, Section 1.7). While smaller DNA lesions such as base modifications or base mismatches typically lead to SNVs, bulky lesions can also result in indels. Unresolved double strand breaks or errors during their repair can lead to mutations as drastic as larger insertions or deletions or chromosomal rearrangements.

1.3 Studying somatic mutations

Somatic mutations are nowadays commonly investigated using next-generation sequencing technologies (Dou et al., 2018). However, detecting somatic mutations is difficult because special care has to be taken in order to differentiate them from germline mutations. In order to discriminate somatic mutations from germline mutations, multiple samples from the same individual have to be investigated, ideally from different organs. Variants that are present in all samples are then assumed to be germline variants, based on the assumption that it is unlikely that the same mutation at the same position arises independently multiple times in the same individual. For instance, when studying cancer mutations, it has become common practice to sequence both the tumor biopsy as well as adjacent normal, non-cancerous tissue or a blood sample as a germline control (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).

In addition to the potential confusion with germline variants, there is also the problem of low clonality. Somatic mutations are not present in all cells of a biopsy. Therefore, the expectation of a variant allele frequency close to 0.5 among sequencing reads, which is the assumption made by germline mutation callers, does not apply to somatic mutations. When sequencing a

somatic biopsy sample, only a small proportion of sequencing reads will show the mutated allele, according to the proportion of cells in the biopsy that carry this mutation (Wood et al., 2015). Somatic mutations that are only present in a small clonal fraction or even in single cells are thus almost undetectable when sequencing bulk samples.

In addition, sequencing artifacts further complicate the detection of true somatic mutations. DNA damage that occurred during sample preparation, amplification bias, allelic imbalance, polymerase errors during DNA amplifications, errors during sequencing itself (e.g. during base calling), and biases during data processing (e.g. read mapping problems or uneven read depth) can lead to false positive somatic mutation calls (Costello et al., 2013; Minoche et al., 2011; Cline et al., 1996; Ma et al., 2019; Pfeiffer et al., 2018). Thus, it is hard to distinguish between sequencing errors and true somatic mutations, especially for low-clonal variants which are only represented in few sequencing reads.

For cancer samples, these problems are less pronounced, because somatic cells have often undergone expansive clonal expansions and thus the groups of cells carrying the same somatic mutations (i.e., clones) are relatively large (Greaves and Maley, 2012). In the context of cancer, researchers are often mostly interested in the mutations that are present in cells that have clonally expanded, since those mutations are more likely to provide a selective growth advantage and thus be the driver of cancer progression. However, if researchers are interested in mutations that are present only in few or even single cells, for example when studying healthy samples, the problems with calling mutations present at a low clonal prevalence are more pronounced (Dou et al., 2018). Thus, in order to identify somatic mutations of small clones, one has to study very small biopsies, sequence to a very high sequencing depth, and/or employ sequencing techniques designed to minimize sequencing errors, or at least make them distinguishable from true mutations (Dou et al., 2018; Abascal et al., 2021). For example, the implementation of unique molecular identifyers (UMIs) can reduce the impact of sequencing errors (Dai et al., 2021). Alternatively, in vitro clonal expansion of single cells, followed by whole genome sequencing of the clone, enables the characterization of the genome of a single cell (Youk et al., 2021). Additionally, somatic variants can be identified by sequencing of natural occurring clonal patches in healthy tissues (Ellis et al., 2021). Directly sequencing single cells is very error-prone due to the low DNA input amount, but new techniques tackling these problems are being developed (Abascal et al., 2021; Xing et al., 2021; Huang et al., 2022; Zhang et al., 2023).

Special analysis pipelines that are able to incorporate paired samples from the same donor are required in order to rule out germline mutations (Dai et al., 2021). Furthermore, population controls are also often used to further restrict

false positive somatic mutation calls (Kalatskaya et al., 2017; Benjamin et al., 2019). Variants that are observed at a certain frequency in this population of control samples are more likely to have a germline origin than to have arisen within the somatic tissue and are therefore often filtered out during somatic mutation calling. Algorithmic techniques such as haplotype phasing, identifying sequencing errors based on certain sequence characteristics, incorporation of special sequencing techniques such as UMIs or nanorate sequencing, as well as careful quality filters can mitigate the impact of sequencing errors (Dai et al., 2021; Abascal et al., 2021; Van der Auwera et al., 2013; Chen et al., 2020; Xu, 2018).

In summary, somatic mutations are ideally studied using deep, non-error-prone sequencing of small somatic biopsies with one or more paired biopsies from the same individual. Currently, the vast majority of available information about somatic mutations stems from cancer sequencing. However, there are various processes in cancer that may not apply to normal tissues. Thus, more studies on somatic mutations in healthy tissues are needed to fully understand the mutation compendium of somatic tissues. Due to the special experimental designs needed and the associated high sequencing costs, studies explicitly studying somatic mutations across multiple tissues are only just arising (Moore et al., 2021; Li et al., 2021).

1.4 Factors influencing somatic mutation burden

Since the cells in our bodies are in a constant battle against mutations by repairing DNA damage, it is mostly a matter of time until somatic mutations occur and accumulate. Indeed, it is well-established that the number of somatic mutations in normal tissues increases with age (Manders et al., 2021; Alexandrov et al., 2015; Milholland et al., 2015; Manders et al., 2021; Blokzijl et al., 2016). Accordingly, pediatric cancers such as leukemia or neuroblastoma typically have a much lower mutation burden compared to tumors that tend to occur later in life (Watson et al., 2013; Alexandrov et al., 2013). The mutation rate within tissues seems to be fairly constant during adult life. Exceptions from this continuous accumulation of mutations over time are phenomena such as chromotripsis or kataegis, where many chromosomal arrangements or point mutations, respectively, suddenly appear in a single catastrophic mutation event (Martincorena and Campbell, 2015). Thus, somatic mutations accumulate over time as a semi-stochastic process, depending on tissue, resulting in an ever-increasing likelihood of the development of cancer (Milholland et al., 2015). Accordingly, age is one of the most important cancer risk factors (DePinho, 2000).

There are large differences in the prevalence of somatic mutations between different tissues. For instance, mutational burdens range from fewer than one mutation per Megabase in some cancer types (e.g., kidney chromophobe carci-

noma, thyroid cancer, and various brain cancers), up to hundreds of mutations per Megabase in other cancer types (e.g., esophagus, lung, and skin cancers) (Alexandrov et al., 2020). In healthy donors, the number of detectable SNVs varied by three orders of magnitude between different tissues, and estimates for the rate of mutation accumulation for different tissues (excluding germ cells) ranged between 9 and 56 SNVs per cell per year (Ren et al., 2022; Manders et al., 2021). The reasons for these stark differences are presented in the following.

The exposure to mutagens plays a major, probably the largest, role in this difference in mutation burden between tissues (Rosendahl Huber et al., 2021). For instance, melanomas and lung cancers are the cancer types with the most somatic mutations, due to the frequent exposure of each tissue to UV light and tobacco smoke, respectively (Lawrence et al., 2013b).

Many types of DNA damage only lead to a sequence change through DNA replication during cell division (Section 1.2). Therefore, the somatic mutation rate of a tissue increases with the rate of cell proliferation. Accordingly, the (stem cell) proliferation rate of a tissue influences the rate of mutation and thus the likelihood of developing cancer in this tissue (Tomasetti and Vogelstein, 2015; Blokzijl et al., 2016). However, mutations also accumulate linearly with time in post-mitotic, non-replicating cells such as neurons (Luquette et al., 2021; Abascal et al., 2021; Lodato et al., 2018) or fully differentiated liver cells (Brazhnik et al., 2020). Some mutation types, reflected by the sequence context in which they appear (i.e., their signature, Section 1.7) tend to be especially associated with age, which is why they were termed clock-like mutations (Alexandrov et al., 2015).

The repertoire of somatic mutations present in any tissue is also the result of selection. In fact, in some regards, somatic tissues behave similarly to the evolutionary dynamics of natural populations. Variants that inhibit cell growth, damage essential cellular functions, or even lead to cell death are under negative selection and will disappear from a cell population. In contrast, variants that lead to a proliferative advantage compared to surrounding tissue will lead to its clonal expansion (positive selection). Variants that have no impact on cellular function are considered neutral mutations (Martincorena and Campbell, 2015). It is often assumed that the majority of mutations randomly occurring in a cell are neutral, since large parts of the genome would not be functionally impaired by smaller mutations (Martincorena et al., 2017). There is also competition between multiple clones within a tissue, so that the clonal composition of a tissue will change over time, depending on the varying fitness of each mutant clone and in response to changes in environment (Colom et al., 2021; Martincorena et al., 2018; Moore et al., 2021). For example, cancer therapy will lead to a selective advantage for

clones that are resistant to the treatment (Venkatesan et al., 2017). Positively selected mutations are, due to their clonal expansion, inherently easier to find than negatively selected mutations which, by definition, disappear from the population. Currently, negative selection is thought to play only a minor role in somatic tissues, especially compared to germline mutations, as there are only weak traces of such purifying selection in somatic tissues (Martincorena et al., 2017; Yadav et al., 2016). While these clonal dynamics have been extensively studied in cancer, recent research indicates that clonal expansion and selection is widespread in non-cancerous, normal tissues as well (Moore et al., 2021; Cooper et al., 2015; Martincorena et al., 2015; Yoshida et al., 2020; Brunner et al., 2019; Martincorena et al., 2018). It was even suggested that the clonal competition even acts as a physiological protection against the emergence of new, potentially cancerous, clones (Martincorena et al., 2018; Colom et al., 2021; Yokoyama et al., 2019).

1.5 (Patho)physiological consequences of somatic mutations

Somatic mutagenesis can serve specific physiological functions. For example, somatic hypermutation in immune cells contributes to immunological diversity and promotes the emergence of high-affinity antibodies. Somatic aneuploidy is widespread in liver, which was proposed to represent adaptation to chronic hepatic injury (Duncan et al., 2012). Furthermore, the widespread presence of somatic mutations in brain neurons has been proposed to reflect adaptation to specific neuronal functions and thus contribute to physiological diversity of the brain (Bushman and Chun, 2013).

In other cell types, somatic mutations tend to have more detrimental effects. The most well-known pathophysiological impact of somatic mutations is their role in cancer (Section 1.6). However, somatic mutations have also been implicated in other diseases related to somatic mosaicism in tissues, including neurodevelopmental disorders (Erickson, 2016), skin abnormalities, clonality of blood cells, X-linked disorders, as well as multiple syndromes involving tissue overgrowth (e.g., Proteus syndrome) (Erickson, 2014; Saini and Gordenin, 2018), most of which are caused by somatic mutations occurring during early development. However, it was recently discovered that many healthy tissues comprise of complex mixture of clonal growths that even carry typical cancer driver genes (Section 1.6). The enhanced growth capacities of these mutated clones may help to heal tissue injuries and may even protect against cancer by posing a competitive environment for new malignant lesions. (Colom et al., 2021; Zhu et al., 2019)

Somatic mutations have also been implicated in the context of aging-associated degenerative processes. For instance, several premature aging diseases are

caused by DNA repair deficiencies (Burtner and Kennedy, 2010). Genomic instability, telomere shortening, as well as epigenetic alterations are considered hallmarks of aging (López-Otín et al., 2013). For instance, somatic loss of the Y chromosome during aging is associated with shorter life span and several age-related diseases such as Alzheimer's disease, diabetes, or immune disorders (Dumanski et al., 2016; Loftfield et al., 2018). Furthermore, increasing immune senescence during aging may further propagate the occurrence of somatic mutations, because aberrant cells are not cleared as effectively. A common hypothesis is that the decay of genomic information in turn leads to a progressive loss of cellular function, for example via increasing transcriptional noise and destabilization of intracellular regulatory networks and pathways (Vijg, 2021). Indeed, transcriptional noise and gene disregulation have been shown to increase with age (Nikopoulou et al., 2019; Debès et al., 2023; Levy et al., 2020; Enge et al., 2017). Furthermore, clonal expansions of somatic mutant clones were shown to be widespread in multiple tissues. This takeover of tissue by individual, "selfish" clones may ultimately alter tissue homeostasis and thus lead to progressive loss of tissue function (Vijg and Dong, 2020). However, chromosomal organization, reflected by altered DNA methylation and histone modification, is often thought to be a better indicator of the increasing disregulation during ageing (Wang et al., 2022). Thus, it remains unclear if, and to what extent, somatic mutations play a causal role in ageing and pathophysiological defects, aside from cancer.

1.6 Somatic mutations in cancer

Over a hundred years ago, it was first proposed that mutations are the cause of cancer, based on the observation of abnormal chromosome structures in tumor cells and the observations that DNA-damaging substances cause cancer. This hypothesis was confirmed and further explored in a plethora of subsequent cancer genome studies (reviewed in Stratton et al. 2009). Since then, the somatic genomes of thousands of tumors have been studied, including large cancer genome sequencing cohort such as the The Cancer Genome Atlas (TCGA) program and the International Cancer Genome Consortium (ICGC). The goal of these studies was to identify the specific mutations that drive cells to uncontrolled growth. Such "driver" mutations are genomic variants that alter the function of specific genes, providing a survival/proliferative advantage of these cell compared to surrounding tissue. This is achieved through functional changes also termed the hallmarks of cancer, including prevention of cell death and growth suppression, modified pathways involved in metabolism and proliferation signaling, evasion of immune control, among others (Hanahan and Weinberg, 2011). For example, one of the most well-known tumor suppressor gene *TP53* (Tumor Protein P53) encodes for a transcription factor (TF) that controls cell proliferation and induces apoptosis upon DNA damage, thus acting as a DNA damage checkpoint. When *TP53* is mutated, this

checkpoint is disabled and cells can divide without restraint, despite profound DNA damage. Cancer mutations include CNVs, so either copy number loss of genes that suppress growth, or amplification of genes that promote proliferation. Furthermore, there are SNVs or indels that either completely disrupt the protein sequence (loss-of-function mutations) by introducing early stop codons, shifting the reading frame, or affecting splice sites, or by changing functional sites of the affected protein.

Cancers are subject to heavy clonal expansion and selection. Thus, observed somatic mutations in cancers do not reflect the pure mutation occurrence along the genome, but also the result from a history of clonal selection. However, it was previously estimated that the majority of cancer mutations are passengers (i.e., do not pose a selective advantage or disadvantage) and negative, purifying selection is very weak for somatic mutations (Yadav et al., 2016; Vogelstein et al., 2013; Martincorena et al., 2017). Therefore, the impact of selection on the global set of somatic mutations observed in cancer tissues is likely small.

Identifying the specific genes which, when mutated, provide a selective advantage (i.e., cancer genes) can also help to better understand cancer physiology and thus to develop treatment approaches. Interestingly, there are differences in the mutations that occur in different tumor types. While some genes are mutated in many cancer types (e.g., *KRAS* and *TP53*), other genes (e.g., *MYC* or *PTCH1*) are mainly associated with tumor emergence in certain tissues Bianchi et al. (2020); Martínez-Jiménez et al. (2020); Haigis et al. (2019). This difference between tissues can in part be explained by different expression levels between the cell types, or by varying regulatory functions. For example, genes involved in NER, which is important for the correction DNA damage caused by UV light, is often mutated in skin cancer (Schaefer and Serrano, 2016). Similarly, BRCA1/2, whose mutations are mainly associated with breast and ovary cancer emergence, seems to be associated with estrogen sensing (Wang and Di, 2014).

In recent years, it was shown that cancer genes are also often mutated in histologically normal, healthy tissues, even to a similar rate compared to tumors (Aghili et al., 2014; Martincorena et al., 2018, 2017; Moore et al., 2020; Lawson et al., 2020; Coorens et al., 2021). We now have first systematic catalogs of somatic mutations in healthy tissues (Li et al., 2021; Moore et al., 2021; Sun et al., 2022). The mutational patterns in healthy tissues appear to be very similar to those observed in cancer. If cancer drivers are present in normal tissue, there have to be other factors that act as the tipping point for the transformation of apparently harmless somatic clones carrying cancer driver mutations into tumors (Wood et al., 2015; Cooper et al., 2015; Martincorena et al., 2015; Yoshida et al., 2020; Brunner et al., 2019; Martincorena et al.,

2018; Millikan et al., 1995). Further investigation of the clonal dynamics in healthy tissues will help to understand the origin of cancer.

1.7 Varying somatic mutation rates along the genome

The frequency of mutations varies along the genome, some genomic regions are more prone to accumulating somatic variants than others. This variation in mutation rate along the genome has been associated to various genomic features acting at different scales of the genome (Figure 2). For example, on a large genomic scale, late-replicating, gene-poor heterochromatin tends to carry more mutations than early-replicating, gene-dense euchromatin. On a smaller scale, the DNA accessibility influences the probability of somatic mutation emergence. The DNA accessibility is determined by the density by which the DNA strands are wrapped around nucleosomes, and thus has an effect on differential DNA damage repair. Finally, at the smallest scale, the likelihood of mutation is determined by the sequence context at each position.

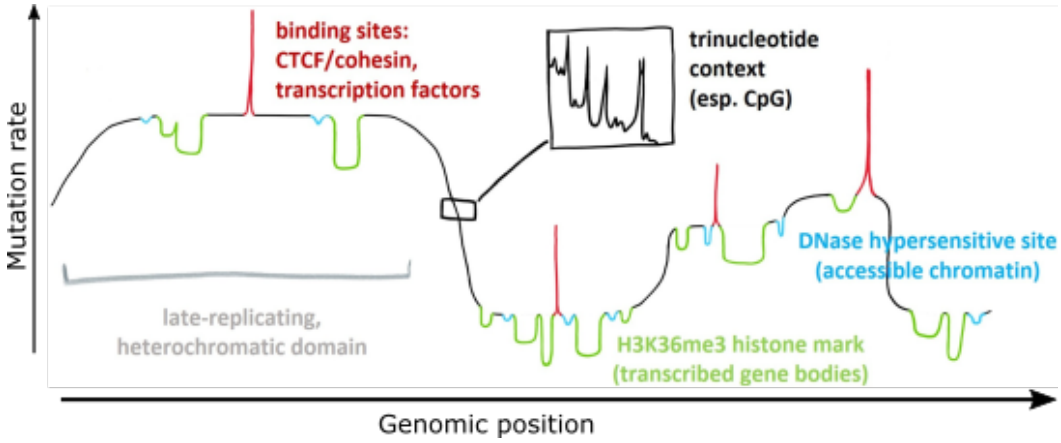

Figure 2: Variation in somatic mutation rates along the genome. Schematic representation of the genomic factors influencing varying somatic mutation probability. Figure adapted from (Supek and Lehner, 2019).

Recent studies have started to unravel many of the genomic features that are correlated with regional variation in mutation rate. The correlation of some of the genomic factors with somatic mutation rate is so weak that it may only be detectable when analyzing many such features in combination. For example, one can test for enrichment of mutations at certain genomic features compared to flanking regions, or, the other way around, compute the enrichment of specific genomic features at observed mutations compared to the rest of the genome (Gonzalez-Perez et al., 2019; Lim et al., 2017; Pich et al., 2018; Georgakopoulos-Soares et al., 2018). However, many of the genomic features might be confounded with each other, and interactions between genomic features in their influence on somatic mutation rate have also been reported. Thus, integrated analyses, allowing for the control of possibly confounded

genomic features are needed to determine which genomic features are likely to play a causal role, and which features are only correlated with somatic mutations emergence due to their confounding with said causal factors. Furthermore, it remains unclear how much each factor contributes to the *de novo* occurrence of genomic variants, or rather to differential DNA damage repair.

Below follows a review of genomic features that might be related to somatic mutation frequency variation along the genome. The next section (Section 1.7.11) presents existing insights into the relative importance of the various genomic factors for somatic mutation occurrence, as well as the impact of DNA damage versus its repair. Note that the mechanisms described here pertain to somatic mutations, since different, although sometimes related, mechanisms pertain to the emergence of germline mutations (Makova and Hardison, 2015; Hodgkinson and Eyre-Walker, 2011)).

1.7.1 Sequence context

As mentioned above, the sequence context represents the smallest genomic scale known to influence somatic mutation rate. That means that the likelihood and the type of mutation at a certain genomic position depend on the nucleotide at the position itself, as well as the immediate 5' and 3' neighboring nucleotides. This context dependence is often related to a specific mutagen acting on the DNA. For instance, skin tissues carry a lot of C to T mutations at dipyrimidine sequences, while G>T mutations at CG dinucleotides are often observed in the lungs of smokers. The spectrum of different, context-specific mutational processes acting on a somatic genome was termed mutational signatures. Alexandrov et al. (2013) were the first to use nonnegative matrix factorization to deconvolute the collection of observed somatic mutations into a combination of latent variables, each of which represent a mutagenic process. Thus, the collection of somatic mutations can be allocated to multiple mutational processes (i.e., mutation signatures) which act on different tissues with varying strength. Many of these signatures have been linked to specific mutagens or mutational processes (Kucab et al., 2019). For example, a signature that mainly appeared in bladder cancers could be tracked down to the consumption of aristolochic acid in herbal medicine (Poon et al., 2015). Thus, the spectrum of signatures varies strongly between tissues, reflecting the difference in mutagen exposure of each different tissue. However, the origin of many of these signatures remains unclear. A list with the repertoire of known mutational signatures, including single Base Substitution Signatures (SBSs), along with their potential sources, is maintained by the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Sondka et al., 2023).

1.7.2 Non-B DNA structures and Repeats

DNA normally takes the conformation of the canonical right-handed double helix, called B-DNA. However, other structures deviating from this canoni-

cal conformation also exist. These so-called non-B DNA structures form at certain repetitive DNA sequences (Bacolla and Wells, 2004; Figure 3). For example, hairpin (slipped DNA) structures form at regions with direct repeats. Related to them is cruciform DNA, where the two DNA strands go into hairpin loops due to inverted repeats, forming a cross-shaped structure. CG-repeats can lead to the formation of left-handed helix, called z-DNA. At DNA tracts with purine-pyrimidine repeat, three DNA strands can come together to form a triple helix (triplex DNA/H-DNA). G-quadruplexes represent an association of guanines from (typically) four different guanine-rich DNA regions into planar stacks. A-phased repeats, which are repeating segments of multiple Adenines, can lead to bends in the DNA strand (Guiblet et al., 2021). Furthermore, many other DNA conformations have been discovered. Since they typically form at specific repeat structures, their presence can be predicted based on DNA sequence (Cer et al., 2013).

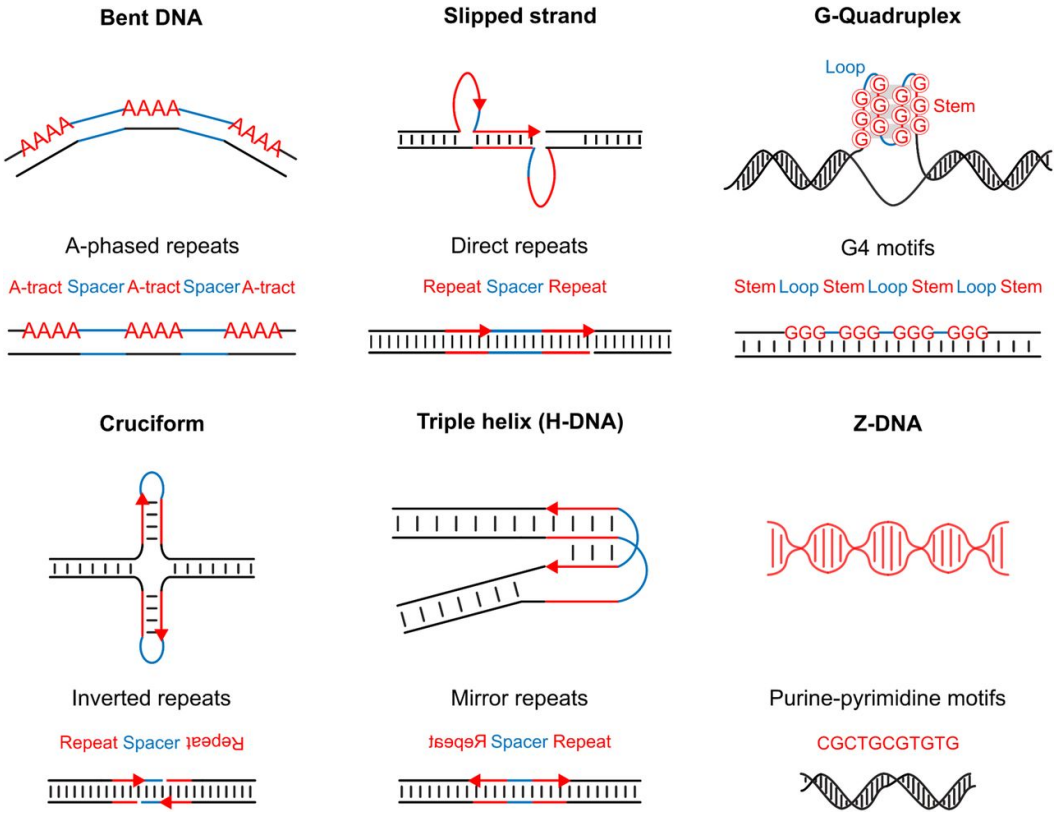

Figure 3: Non-B DNA structures. Schematic representation of abnormal DNA structures that can form at specific DNA sequence elements. Figure design adapted from Weissensteiner et al. (2023).

Non-B structures destabilize the DNA and affect transcription, DNA replication, and telomere maintenance (Zhao et al., 2010). Importantly, they are associated with increased mutation probability at both the germline and somatic

level, including SNVs, indels and chromosomal rearrangements (Bacolla and Wells, 2004; Georgakopoulos-Soares et al., 2018; Aghili et al., 2014). Repeats and the resulting DNA structures can affect somatic mutations by causing errors in DNA replication through polymerase stalling and slippage (Madireddy and Gerhardt, 2017). The somatic instability of repeat tracts is especially relevant in the context of neurodegenerative disorders such as Huntington disease (Kacher et al., 2021) or spinocerebellar ataxia type 1 (Chong et al., 1995), which are caused by destabilization of specific proteins due to the amplification of repeat segments beyond a specific threshold, which may also occur somatically. Further, mutations at simple repeat regions were associated with cancer formation (Ionov et al., 1993).

1.7.3 DNA accessibility

The DNA organization and compaction on the smallest level is realized by the wrapping of DNA around nucleosomes. Nucleosomes are mixed complexes of DNA and histones. Each nucleosome consists of two copies each of the four proteins H2A, H2B, H3, and H4, as well as a DNA stretch of about 147 base pairs (bps) that is wrapped around them 1.67 turns. Nucleosomes represent the basis for chromatin structure, thus playing an important role in genome organization and activity (Section 1.7.4). For example, the distance between nucleosomes, the so-called linker sequences can vary, with lower linker length corresponding to more densely compacted DNA and lower transcriptional activity (Voong et al., 2017; Schones et al., 2008). Furthermore, the histones proteins can carry a wide array of chemical modifications which act as markers of chromatin state including DNA compaction and transcriptional regulation (discussed below).

DNA accessibility is commonly assayed using DNAse hypersensitive sites sequencing (DNAse-seq) or Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq), where nuclear DNA is treated with DNase I or with Tn5 Transposase, respectively. Both enzymes cut the DNA at accessible stretches, and the Tn5 Transposase simultaneously inserts sequencing adaptors. Sequencing of the resulting DNA fragments will result in periodic patterns of accessible and non-accessible DNA regions, which give insight into the positioning of nucleosomes along the sequence (Zhong et al., 2016).

DNA accessibility is correlated with somatic mutation rate. For instance, nuclosome-wrapped DNA stretches accumulate more mutations compared to the linker regions, where repair complexes can more easily sense and remove DNA lesions (Hara et al., 2000; Yazdi et al., 2015; Tolstorukov et al., 2011; Sabarinathan et al., 2016). In addition, there is a 10 bp periodic oscillation of mutation frequency at nucleosome-bound DNA, corresponding to the distance of outward-facing minor groves of the DNA winding around the histones (Pich et al., 2018; Brown et al., 2018).

1.7.4 Genome organization and domains

The 23 chromosome pairs of the human genome encompass a total of more than 6000 Mega base pairs (Mbps), which have to fit inside a nucleus that is, on average, less than 10 micrometers wide (Piovesan et al., 2019). This is achieved through the ordered compaction of the DNA by nucleosomes. However, parts of the DNA still have to be accessible to the transcription machinery as well as various regulatory DNA-binding factors. Therefore, the compaction level of DNA is flexible and organized in a way to make its efficient readout possible. Chromatin exists in two general varieties, heterochromatin and euchromatin, which were originally distinguished by the color intensity upon staining of chromosomes during metaphase (Heitz, 1928). The lighter bands correspond to more open euchromatic regions, which are associated with active chromatin and higher gene density. By contrast, the darker, heterochromatic bands correspond to transcriptionally silenced regions (Penagos-Puig and Furlan-Magaril, 2020). More densely packed, heterochromatic regions accumulate more mutations, most likely due to differential MMR along the genome (Supek and Lehner, 2019). That means that chromatin density does not influence the rate of DNA damage, but rather the rate at which damage is removed.

Centromeres and Telomeres are special regions of the chromosomes. Centromeres are the genomic location that link sister chromatids together during cell division, while telomeres represent the ends of chromosomes that act as buffers to protect the DNA of continuous degradation due to failure to replicate the ends of chromosomes during DNA replication. Both represent special genomic features in that they consist of heterochromatic DNA with large arrays of repetitive DNA. Although neither of them were, to our knowledge, previously linked to somatic mutation rate, it is plausible that mutations are prone to happen in or close to these condensed, repetitive regions.

The 3D organization of chromatin is nowadays commonly studied on a genome-wide scale with high-throughput chromosome conformation sequencing (Hi-C) (Belton et al., 2012). In this technique, chromatin contacts are measured by crosslinking adjacent DNA strands and sequencing the resulting hybrid DNA fragments. The resulting interaction matrix is then analyzed to infer chromosomal organization, up to reconstructing the relative positioning of chromatin within the nucleus (Belton et al., 2012). The first principal component of intrachromosomal contact correlations divides the genome into two general compartments corresponding to open and closed chromatin, or eu- and heterochromatin (Belton et al., 2012; Dekker et al., 2013). Regions with exceptionally many chromatin contacts, termed frequently interacting regions (FIREs), correspond to important enhancer sites (Schmitt et al., 2016). While chromatin organization is mostly conserved, there are differences between tissues, reflecting tissue specific transcriptional regulation (Schmitt et al., 2016). Fur-

thermore, the intrachromosomal contacts are also often visualized in heatmaps of chromosomal contacts along the genome, where nested regional modules, called topologically associating domains (TADs) appear as square clusters with relatively higher within-region contact frequency (Figure 4). These TADs often correspond to functional modules formed by DNA loops, which are necessary for efficient transcriptional regulation, for example through the spacial aggregation of promoters and enhancers (Dekker et al., 2013). Accordingly, disruption of TAD boundaries was shown to impair transcriptional control, which can even lead to disease (Lupiáñez et al., 2016).

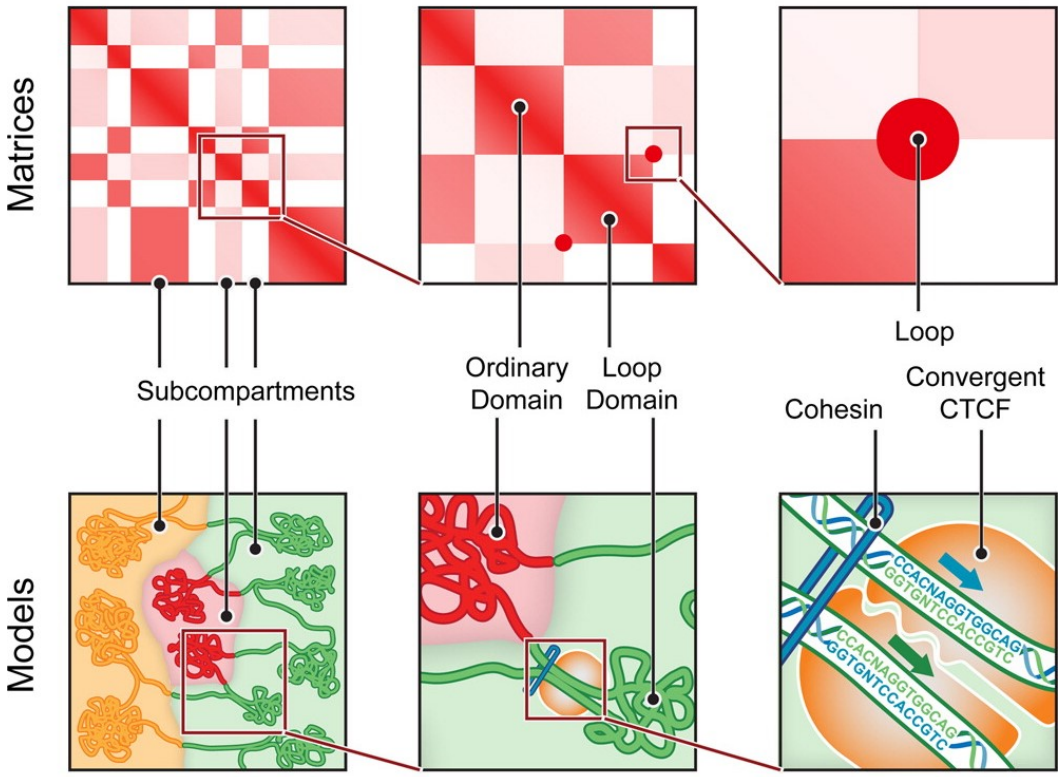

Figure 4: Chromosomal contact maps. Schematic representation of typical visualization of chromatin contact maps (upper panel) and corresponding organizational features. In a contact map, the color intensity represents the frequency of DNA-DNA contacts between two genomic regions. Domains, loops, and subcompartments appear as squares and points along the diagonal, which represent relatively high contact frequencies. Figure adapted from Rao et al. (2014).

The chromatin loops represented by TADs are formed and maintained by multiple proteins. For example, the protein CCCTC-Binding Factor (CTCF) is involved in transcriptional regulation through the formation and stabilization of chromatin loops and genome domains. CTCF was mostly known as a TF that can block or enhance the communication between enhancers and promoters, thus acting as an insulator. CTCF, together with other proteins,

is enriched at the boundaries of TADs, thus functionally and spatially separating genomic regions (Kim et al., 2015). Increasing evidence suggests that CTCF excerts this insulation function by acting as an anchor of DNA loop boundaries. More specifically, DNA domains are formed through loop extrusion, where the DNA strand is threaded trough the protein complex Cohesin until blocked by CTCF (Hansen, 2020).

Since genome organization is confounded with many other genomic factors, including DNA accessibility and replication timing, its role in somatic mutation occurrence remains unclear and has mostly been studied only indirectly. However, Lawrence et al. (2013b) found a negative correlation between Hi-C compartment and somatic mutation density, which means that closed DNA regions displayed more mutations. This is in line with the correlation of mutation rate with DNA accessibility described above (Section 1.7.3). Furthermore, the spacial clustering of DNA was shown to impact the occurrence of CNVs (De and Michor, 2011). Genomic regions that tend to localize to the nuclear periphery are associated with a higher somatic mutation rate (Smith et al., 2017). In addition, increased somatic mutation rates at TAD boundaries and CTCF binding sites have been reported in multiple cancer types, including melanoma and gastric cancer, likely due to impaired NER (Akdemir et al., 2020; Kaiser et al., 2016; Sivapragasam et al., 2021; Umer et al., 2016; Guo et al., 2018; Sabarinathan et al., 2016).

1.7.5 Transcription

In general, higher transcription levels are negatively correlated with mutation rates (Haradhvala et al., 2016; Moore et al., 2021). This contradicts the fact that transcribed DNA, due to the torsion stress during unwinding and due to the strands being exposed as single strands during transcription, is theoretically subjected to more DNA damage (Jinks-Robertson and Bhagwat, 2014). However, fewer somatic mutations occur at transcribed genes due to transcription-coupled NER, where the repair complex is recruited to the DNA upon RNA polymerase stalling due to a DNA lesion (Hanawalt and Spivak, 2008). Furthermore, exons exhibit fewer somatic mutations compared to introns. Interestingly, there are marked differences in mutation rate and spectrum on the transcribed versus the untranscribed strand (Alexandrov et al., 2013; Mao et al., 2020; García-Nieto et al., 2019; Tomkova et al., 2018), because lesions on the non-coding strand are detected less often and thus avoid repair. However, transcription-coupled damage on the untranscribed strand with unknown etiology was also reported (Haradhvala et al., 2016; Lodato et al., 2015; Tomkova et al., 2018). This could be due to the untranscribed strand being in a single-stranded state during transcription and thus prone to form abnormal DNA structures (as described above) compared to the coding strand which is bound by the transcription complex and newly synthesized

RNA (Jinks-Robertson and Bhagwat, 2014).

However, the observed reduction in mutation rate at expressed genes might also be due to the confounding with generally higher DNA accessibility of transcriptionally active DNA, as well as them lying in early-replicating genomic regions which are associated with lower mutation rate (Supek and Lehner, 2015). Thus, transcription is confounded with numerous other genomic features in their effect on somatic mutation occurrence.

1.7.6 DNA replication

When our cells divide, the DNA is replicated in a highly orchestrated and tightly controlled manner (Fragkos et al., 2015). DNA replication starts during the S-phase at so-called replication origins, where the replication complex assembles and the DNA is copied via bidirectional extension of replication bubbles. Since DNA polymerases can only synthesize DNA in one direction, one strand (the leading strand) is synthesized continuously, while the other strand is synthesized discontinuously in so-called Okazaki fragments, which then have to be connected to form a continuous strand (Lujan et al., 2016). Interestingly, it was shown that replication origins may co-locate with g-quadruplex structures, since the quadruplex folding may facilitate DNA duplex unwinding (Besnard et al., 2012).

Not all replication origins of the genome are activated at the same time, leading to variation in replication timing along the genome (Jeon et al., 2005; Fragkos et al., 2015). This replication timing can be assayed by synchronizing cells according to their cell cycle phase and then either sequencing newly synthesized DNA (Besnard et al., 2012) or by assaying DNA copy number states at different cell cycle time points and deducting timing of replication from this data (Hansen et al., 2010; Jeon et al., 2005). Replication timing is correlated with transcriptional regulation and chromatin structure (Hiratani et al., 2009; Gilbert, 2002; Hansen et al., 2010; De and Michor, 2011). More specifically, euchromatic, transcriptionally active regions tend to be replicated earlier than heterochromatin. In addition, replication timing was also associated with germline and somatic mutation frequency variation along the genome (Stamatoyannopoulos et al., 2009; Koren et al., 2012; Lawrence et al., 2013b; Liu et al., 2013). This might be caused by a deterioration of replication fidelity with progressing cell cycle phases, for example due to the depletion of the pool of free nucleotides or accumulation of single-stranded DNA at later-replicating regions (Stamatoyannopoulos et al., 2009). Furthermore, an asymmetry in mutation frequency and signatures was observed between the leading and lagging strands of DNA replication, especially for mutations associated with an APOBEC signature (Tomkova et al., 2018; Haradhvala et al., 2016).

1.7.7 Transcription factor binding

Transcription factors (TFs) are proteins that bind to the DNA at so-called transcription factor binding sites (TFBSs) to regulate the transcription of genes. There are more than a thousand different TFs with varying preferred binding sites and regulatory functions. Active TFBSs usually lie in accessible regions of the genome and thus are identified through a combination of Chromatin Immunoprecipitation sequencing (ChIP-Seq) and DNA accessibility data.

The rate of somatic mutations is particularly high at TFBSs (Melton et al., 2015), but the exact cause for this remains unclear. It may in part be due to increased accumulation of DNA damage at TFBSs. For example, it was shown that UV-induced damage is augmented at TFBSs, especially for TFs from the Erythroblast Transformation Specific (ETS)-domain TF family (Hu et al., 2017; Mao et al., 2018; Fredriksson et al., 2017). Furthermore, the increased mutation rate at TFBSs was also attributed to impaired DNA damage control via NER due to limited DNA accessibility at the TF-bound site (Perera et al., 2016; Sabarinathan et al., 2016). Since active TFBS lie in accessible DNA regions (see above), this results in a pattern of increased mutation rate at the binding sites themselves, combined with decreased mutation rates at flanking regions (Gonzalez-Perez et al., 2019).

1.7.8 Histone modifications

As described above, the histone proteins within nucleosomes can carry a wide array of different chemical modifications. Depending on the position and the type (e.g., methylation, acetylation, or phosphorylation) of the modification, these modifications confer varying regulatory information (Bannister and Kouzarides, 2011). They do this by either directly influencing chromatin structure through alteration of the tightness of nucleosome wrapping, or by recruiting other DNA-binding factors that exert regulatory functions. Many histone marks have been associated with specific regulatory regions and functions. For instance, H3K4 trimethylations are typical marks for active promoters, and H3K9 and H3K27 acetylations occur at active enhancers, while H3K9me3 and H3K27me3 are associated with repressive chromatin (Albini et al., 2019).

Moreover, certain chromatin marks are correlated with varying somatic mutation rate along the genome (Li et al., 2013; Supek and Lehner, 2017). For example, repressive marks such as H3K9me3 co-occur with high density of mutations (Schuster-Böckler and Lehner, 2012), while the presence of chromatin marks associated with transcriptional activation, such as H3K4me3 or H3K9ac, are negatively correlated with mutation occurrence (Schuster-Böckler and Lehner, 2012; Polak et al., 2015). Since the rotational orienta-

tion of DNA wrapped around nucleosomes impacts mutation occurrence, it does not seem inconceivable that modifications of the nucleosomes themselves in turn also affect mutation rate. However, since, by definition, chromatin marks are confounded with transcriptional activity and chromatin state, it is unclear whether the association with somatic mutation represents a causal relationship or simply correlation.

1.7.9 DNA methylation

DNA methylation represents an epigenetic DNA modification in the form of methylation of cytosines, which plays a central role in transcriptional regulation (Moore et al., 2013). More specifically, cytosines are methylated at the C5 carbon position, especially at CpG dinucleotide loci. DNA methylation is an important factor for transcription regulation, but also plays a role in genomic silencing of retroviral elements or X-chromosome inactivation (Moore et al., 2013; Duncan et al., 2018). The majority of CpG sites (about 75%) that are spread along the genome are methylated (Tost, 2010). Due to the spontaneous deamination occuring at methylated cytosines, leading to a C>T mutation, CpG motifs have been lost during evolution and are relatively depleted in mammalian genomes. The exception are so-called CpG islands, which are about 1 kilo base pair (kbp) stretches of DNA with higher CpG density. These CpG islands tend to be unmethylated and are associated with the maintenance of active chromatin state at promoters (Tost, 2010). Together with other epigenetic factors, methylation regulates embryogenesis, transcription, and genome organization.

DNA methylation can be assayed using bisulfite sequencing, where methylated cytosine residues are chemically converted into uracils, which will be identified as thymines during DNA sequencing (Li et al., 2018; Frommer et al., 1992). The percentage of reads with C>T mismatches therefore reflects the proportion of methylated cytosines in a tissue.

The potential impact of methylation on somatic mutation is obvious, considering the widespread deamination of methylated cytosines (Pfeifer, 2006). Indeed, the mutational signature attributed to methyl-cytosine deamination (SBS1, Figure S20) is prevalent in virtually all cell types investigated (Alexandrov et al., 2013).

1.7.10 GC content

The GC content describes the proportion of G and C nucleotides of a DNA sequence. It can be computed for different ranges of DNA, thus reflecting the local (for smaller ranger) up to a global GC content. The GC content of the entire human genome is about 40% (Piovesan et al., 2019), but the GC content varies along the genome, in part reflecting different functional genomic capacities (Arndt et al., 2005). For example, genes tend to lie in GC-rich re-

gions (Lercher et al., 2003). GC content is often used as a proxy to distinguish euchromatic, gene-rich regions (high GC content) from heterochromatic, gene-poor regions (low GC content). However, the concept of GC content defining eu- versus heterochromatic regions, or so-called isochores, has been discarded (Niimura and Gojobori, 2002). Still, the observed correlation of mutation rate with GC content along the genome is likely a mix between the presence of mutable methylated CpG sites as well as correlation with transcriptional activity (Fryxell and Moon, 2005; Arndt et al., 2005).

1.7.11 Integrated analyses of factors influencing somatic mutation rate along the genome

Many of the connections between genomic features and somatic mutation rate described above were investigated in isolation. However, the relative impact on mutation rate for the different genomic features has not been fully elucidated. In order to understand how much of the total variation in mutation rate along the genome can be attributed to each genomic factor, integrative studies have to be conducted, where possibly confounded factors are accounted for (Supek and Lehner, 2019). Indeed, multiple studies have investigated, or even attempted to model mutation rate variation, using multiple genomic features at once. The first studies to try to predict the somatic mutation rate along the genome based on large-scale genomic features (i.e., features that span large chromosomal regions, corresponding to low resolution) such as GC content, gene density, nucleosome occupancy, recombination rate, distance to telomeres, and replication timing (Hodgkinson et al., 2012; Woo and Li, 2012; Liu et al., 2013). They concluded that, at a Mbp scale, higher-order genomic organization, GC content and replication timing were among the most important determinants of mutation rate. However, Schuster-Böckler and Lehner (2012) compared various genomic features in their correlation with mutation rate on a Mbp scale, including, in addition to the features included in said studies, various histone modifications as well as RNA polymerase and CTCF binding. They found that multiple repressive and active histone marks, most of all H3K9me3, were correlated with mutation rate. Polak et al. (2015) modeled the Mbp-scale contribution of various histone marks measured in multiple tissues, as well as DNA accessibility, replication timing, gene density, expression, and nucleotide composition. Histone modifications marking closed chromatin were the most important predictors, while gene expression and nucleotide content provided no predictive performance advantage. Importantly, they found that tissue-specific histone marks were more predictive for the mutation rate of the corresponding tumor type. Moreover, they suggested that the tumor type can be predicted simply based on the distribution of somatic mutations along the genome, a concept that was later further explored to predict the tissue of origin of tumor samples (Kübler et al., 2019; Jiao et al., 2020). While all of these studies only investigated the mutation rate in larger (Mbp-scale) ge-

nomic windows, Bertl et al. (2018) predicted mutation rate on a single bp scale based using multinomial logistic regression. They included sequence context, GC content and CpG island annotations, expression, replication timing, DNA accessibility, conservation, as well as annotations of genomic element types as predictors and concluded that the site-specific conservation, replication timing and expression level were the most predictive features. Arnedo-Pac et al. (2023) focused on the hotspot propensity of mutation signatures as a proxy for the variability of mutation rates at single base resolution. For example, they found that methylation rate along the genome explained almost all (80–100%) of the SBS1 hotspot propensity, and that hotspot propensity along the genome was correlated with DNA accessibility, transcriptional activity, and replication timing.

Such models of mutation rate can be used to identify cancer drivers. More specifically, they can be used to find genomic regions or genes that are mutated in cancer more often than would be expected. This concept was first applied to the identification of mutated driver genes (Lawrence et al., 2013b; Martincorena et al., 2017). It was then extended to non-coding regions in order to identify cancer-driver mutations in functional regions outside of coding genes, which are inherently harder to study (Lochovsky et al., 2015; Juul et al., 2019; Shuai et al., 2020; Rheinbay et al., 2017). For example Juul et al. (2019) used the model developed in Bertl et al. (2018) as the baseline expected mutation rate to identify driver mutations in promoter elements, splice site, and untranslated regions (UTRs).

Despite all these studies, it remains unclear which genomic features have the largest influence on regional somatic mutation rates, although common trends became apparent. In most studies, replication timing and features reflecting on the general transcriptional activity were most predictive. However, many of the factors are confounded with each other, and thus it remains unclear which features are actually causal and which are just correlated (Supek and Lehner, 2019). In order to prove causality, carefully designed studies are needed where each genomic feature is perturbed and the impact on mutation rate is tested. For example, Avgustinova et al. (2018) tested the impact of H3K9 methylation on mutation burden by deleting the H3K9 methyltransferase G9a and concluded that it does not play a causal role. Analogously, the impact of various mutagens on mutational signatures was investigated by Kucab et al. (2019).

However, as stated above, most of these studies only looked at the density of mutations in larger genomic windows. Therefore, genomic features that vary on a smaller genomic scale, for example TFBSs or even the sequence context, are underrepresented. Supek and Lehner (2019) estimated that trinucleotide context, CTCF and TF binding, as well as replication timing were among

the factors most strongly associated with mutation rate. Current research suggests that many of the features that were shown to be able to predict regional mutation rates rather reflect local DNA repair efficiency as opposed to differential DNA damage rates (Supek and Lehner, 2017; Li et al., 2013; Yazdi et al., 2015; Supek and Lehner, 2015; Perera et al., 2016; Lim et al., 2017).

Taken together, it seems that on a small genomic scale (i.e., higher resolution) factors such as the sequence context and abnormal DNA structures due to repeats are the main determinants of the rate of emergence of new mutations, whereas differential DNA repair, for example due to variable DNA accessibility (nucleosomes, TFBS) is mainly responsible for the large-scale (i.e., low-resolution) variation in mutation rate along the genome (Supek and Lehner, 2015; Gonzalez-Perez et al., 2019).

2 Aims and Objectives

We need to understand where in the genome somatic mutations occur more often, in order to better understand the origins of cancer, assess selection processes, and identify forces that drive mutational processes. That requires a model that reflects the baseline somatic mutation rate along the genome. It is important to quantify mutation rates independent of their selective advantage for the tumor or somatic tissue.

Previous work has shown that is possible to predict large-scale mutation rates based on genomic features and that it is possible to create such models at single-nucleotide resolution (Bertl et al., 2018). We are going beyond that by using improved regression methods, including more genomic covariates, using tissue-specific data, while going down to single base pair resolution.

The first goal of this study is to create a model that robustly models the baseline mutation rate along the genome. To that end, we will identify, collect, and process all genomic features that might have an influence on somatic mutation processes, such as epigenetic features, 3D chromatin structure, replication timing, expression, and many more. This model will focus on the exome first. Since the impact of sequence context has already been extensively previously studied through the analysis of mutational signatures, we will disregard the local genetic sequence around variants. We will compare three machine learning models in order to identify the approach that is best suited for our data. This model will be trained on tumor data and predictions will be tested on independent data from healthy tissues. The model will have to be thoroughly tested to identify and exclude various systemic or technical biases.

Having established a model, we will use it to systematically characterize molecular genomic factors that act on mutation rates. We want to answer how much each chromatin-related factor contributes to the variation in the baseline mutation rate. How different or similar are the mechanisms influencing somatic mutations between tissues and what are the major determinants explaining differences between tissues? We will investigate how much different mutagen sources are reflected in our models and how they are related to tissue-specific differences.

Finally, we will extend our insights from the exome to the entire genome and create a model that predicts the baseline somatic mutation along the human genome.

In summary, this project aims to predict the occurrence of somatic mutations as a function of molecular and genomic co-factors, which will provide new insight into the mechanisms underlying somatic mutation variation.

3 Data and Methods

3.1 Software

All processing was performed using publically available software (Table 1). Analyses were performed in R (R Core Team, 2021) using publicly available R packages (Table 2). All scripts, functions, and essential raw data are accessible at https://github.com/Cschmalohr/MutationModel which can be accessed using the read-only ssh access token "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" (with the user name "Cschmalohr").

Table 1: Software used in this study.

Software and Version	**Source**
SIFT4G, GRCh37.74 database	https://sift.bii.a-star.edu.sg/sift4g/AnnotateVariants.html; https://sift.bii.a-star.edu.sg/sift4g/public/Homo_sapiens/GRCh37.74.zip; Vaser et al. (2016)
BEDOPS 2.4.36	https://bedops.readthedocs.io/en/latest/; Vaser et al. (2016)
R 4.3.1	https://www.r-project.org/; R Core Team (2021)
wiggletools 1.2.2	https://github.com/Ensembl/WiggleTools; Zerbino et al. (2014)
bedGraphToBigWig 2.8	http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/bedGraphToBigWig; Kent et al. (2010)
samtools 1.6-12-gc7b2f4f (htslib 1.6-38-g8003166)	http://www.htslib.org/; Danecek et al. (2021)
BigWigAverageOverBed v2	http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/BigWigAverageOverBed; Kent et al. (2010)
liftOver	http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver; Hinrichs et al. (2006)
EMBOSS 6.6.0.0	http://emboss.sourceforge.net/; Rice et al. (2000)
bedtools 2.27.1	https://bedtools.readthedocs.io; Quinlan and Hall (2010)

Table 2: R packages used in this study.

Package	Version	Citation
berryFunctions	1.19.1	Boessenkool (2020)
biomaRt	2.46.3	Durinck et al. (2009)
Biostrings	2.58.0	Pagès et al. (2020)
c060	0.3-0	Sill et al. (2014)
colorspace	2.1-1	Zeileis et al. (2020)
corrplot	0.84	Wei and Simko (2017)
data.table	1.14.0	Dowle and Srinivasan (2021)
dplyr	1.0.4	Wickham et al. (2021)
GenomicRanges	1.42.0	Lawrence et al. (2013a)
ggforce	0.4.2	Pedersen (2024)
ggplot2	3.3.3	Wickham (2016)
gplots	3.1.3.1	Warnes et al. (2024)
gridExtra	2.3	Auguie (2017)
Matrix	1.3-2	Bates et al. (2021)
matrixStats	1.4.1	Bengtsson (2024)
plotrix	3.8-1	Lemon (2006)
plyr	1.8.9	Wickham (2011)
ranger	0.12.1	Wright and Ziegler (2017)
RColorBrewer	1.1-2	Neuwirth (2014)
readxl	1.3.1	Wickham and Bryan (2019)
reshape2	1.4.4	Wickham (2007)
rhdf5	2.34.0	Fischer et al. (2020)
ROCR	1.0-11	Sing et al. (2005)
rtracklayer	1.50.0	Lawrence et al. (2009)
scales	1.3.1	Wickham et al. (2023)
sinaplot	1.1.0	Sidiropoulos et al. (2018)
strawr	0.0.1	Durand et al. (2016)
stringr	1.4.0	Wickham (2019)
tibble	3.0.6	Müller and Wickham (2021)
tidyr	1.1.2	Wickham (2020)
vcfR	1.12.0	Knaus and Grünwald (2017)
viridis	0.5.1	Garnier (2018)
xlsx	0.6.5	Dragulescu and Arendt (2020)
zoo	1.8-12	Zeileis and Grothendieck (2005)

3.2 Reference Genome Files

All analyses were based on human genome version GrCh37 and limited to the autosomes (Table 3). The genome fasta file was downloaded from Gencode, unzipped, and indexed with `samtools faidx`. The annotation gtf file was filtered for exons from known, protein-coding transcripts (according to `feature type = "exon"` and the tags `type = "protein_coding"`, `transcript_type = "protein_coding"` and `transcript_status = "KNOWN"`). In addition, we excluded potentially problematic genomic regions such as repeats, or regions prone to mapping problems using repeatMasker, Tandem Repeats Finder, 100mer Alignability, and Encode Consensus Excludable an-

notations (Table 3). These filtered exons were used for any analyses pertaining to the exome. Files were transformed from hg38 to hg19 using the liftOver tool (Hinrichs et al., 2006) where necessary (e.g., when annotations where not available for hg19).

Table 3: Reference Genome Files used in this study.

Files	Source
Gencode GRCh37 Reference Genome release 43	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_43/GRCh37.p11.genome.fa.gz; (Frankish et al., 2018)
Gencode GRCh37 Annotation release 43	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_43/GRCh37_mapping/gencode.v43lift37.basic.annotation.gtf.gz; (Frankish et al., 2018)
University of California Santa Cruz (UCSC) LiftOver chain files	https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gzhttp://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz (Hinrichs et al., 2006)
UCSC chromosome lengths	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes; (Raney et al., 2023)
UCSC repeatMasker	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.out.gz; (Raney et al., 2023)
Tandem Repeats Finder	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.trf.bed.gz; (Raney et al., 2023)
Mappability	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeCrgMapabilityAlign100mer.bigWig; (Raney et al., 2023)
ENCODE Consensus Excludable Regions	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz; (Raney et al., 2023)

3.3 Mutations

3.3.1 Exome cancer mutations

For the training of the exome mutation models, we used mutation calls from the Pan-Cancer Atlas (Ellrott et al., 2018), where they uniformly processed sequencing data from 33 tumor types profiled by TCGA. The file `mc3.v0.2.8.PUBLIC.maf.gz` was downloaded from http://api.gdc.cancer.gov/data/1c8cfe5f-e52d-41ba-94da-f15ea1337efc. We took the subset of mutations that corresponded to SNVs (column `Variant type = "SNP"`) and that lay in regions with sufficient coverage (column `n_depth>=8` and column `t_depth >= 14`). We then filtered out mutations with $> 1\%$ population frequency (in any of the ancestries `GMAF`, `AFR_MAF`, `AMR_MAF`, `EAS_MAF`, `EUR_MAF`, `SAS_MAF`, `AA_MAF`, or `EA_MAF`). Finally, we subset the variants to positions lying within non-problematic exonic regions as defined above (Section 3.2). For each tissue, we used the variants of each corresponding cancer type according to Table 4.

Finally, we excluded positions that were mutated more than once (i.e., in more than one patient), to minimize the influence of mutations under positive selection in tumors. The remaining positions were treated as true positive (TP) mutations during model training.

We generated true negative (TN) positions for model training by randomly sampling non-mutated positions from all exonic regions. The TNs were sampled in a way that the sequence context was balanced between TPs and TNs. To that end, for each TP, we sampled (without replacement) a random exonic position from the same chromosome with the same sequence context (i.e., pentamer) as the TP. That way there were equal numbers of TPs and TNs for each sequence context (the pentamer defined by the mutated position and the two preceding and following bases) on each chromosome. For the sequence context, reverse complementary pentamers were always consolidated since it is not possible to differentiate which of the two strands was mutated. Sampling was restricted to the filtered exonic regions (Section 3.2). All observed mutations (i.e. TPs before filtering) were excluded when sampling for TN positions.

Table 4: Cancer types from TCGA used for each tissue.

Tissue	Cancer type
Brain	Brain lower grade glioma (LGG)
Breast	Breast invasive carcinoma (BRCA)
Colon	Colon adenocarcinoma and Rectum adenocarcinoma (COADREAD)
Esophagus	Esophageal carcinoma (ESCA)
Kidney	Kidney renal clear cell carcinoma (KIRC)
Liver	Liver hepatocellular carcinoma (LIHC)
Lung	Lung adenocarcinoma (LUAD)
Ovary	Ovarian serous cystadenocarcinoma (OV)
Prostate	Prostate adenocarcinoma (PRAD)
Skin	Skin cutaneous melanoma (SKCM)

3.3.2 Whole genome cancer mutations

For the whole genome models, we used the mutation calls generated by the Pan-Cancer Analysis of Whole Genomes (PCAWG) project for ICGC cancer samples (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). We downloaded the file `final_consensus_passonly.snv_mnv_indel.icgc.public.maf.gz` from the ICGC Data Portal https://dcc.icgc.org/api/v1/download?fn=/PCAWG/consensus_snv_indel/final_consensus_passonly.snv_mnv_indel.icgc.public.maf.gz. As for the exome mutations, we subsetted the mutations to SNVs (column `Variant_Type = "SNP"`) from the autosomes. We also filtered for variants with sufficient support (at least two reads supporting the alternative allele in the cancer sample, a coverage of at least 14, and called by at least two mutation calling pipelines). We removed

variants with >1% population frequency in the 1000 genomes project (column `i_1000genomes_AF`). Again, we excluded mutations that were detected in more than one sample. Finally, we limited the analysis to "callable" regions, by excluding problematic genomic intervals defined by the ENCODE Consensus Excludable Regions (Table 3). In contrast to the exome data, we did not filter based on repeat or mappability annotations, since repeats can actually have a functional impact on mutation emergence and are much more prominent in non-coding regions. The remaining mutations were partitioned to each tissue (Table 5) and labeled as TPs. Just like for the exome data, we generated TN positions by randomly sampling non-mutated positions from the entire genome, such that there are equal numbers of TPs and TNs for each sequence context (pentamer) on each chromosome. Sampling was restricted to non-problematic regions according to the the ENCODE Consensus Excludable Regions and excluding all non-filtered TP mutation calls.

Table 5: Cancer types from TCGA used for each tissue.

Tissue	**Cancer type**
Brain	CNS-Medullo, CNS-PiloAstro
Breast	Breast-AdenoCa
Esophagus	Eso-AdenoCa
Kidney	Kidney-RCC
Liver	Liver-HCC
Ovary	Ovary-AdenoCA
Prostate	Prost-AdenoCA
Skin	Skin-Melanoma

3.3.3 Healthy tissue somatic mutations

Healthy tissue mutations were taken from SomaMutDB V1.4 (Sun et al., 2022). All mutations (hg19) for each tissue were accessed and downloaded manually from the Search function of the SomaMutDB website. The tissues iPSC, blood, and embryonic_stem_cell were removed because they included cell types that were not representative of normal somatic mutation accumulation (e.g., immune cells and stem cells). The data from the endometrium could not be used because there were no SNVs available, only indels. We generated one healthy tissue dataset for only the exome, and another covering the whole genome. The processing was done as similar as possible as for the training data. We removed indels, non-autosomal chromsomes, and positions that were mutated more than once. For the exome data, we limited TPs and TNs to the filtered exonic regions (Section 3.2), while for the genome data, we only removed the ENCODE Consensus Excludable Regions. Again, TN positions were randomly sampled such that the sequence context was balanced between TPs and TNs for each chromosome.

3.4 Predictors

We chose a wide range of genomic features that might influence, or be correlated with somatic mutation rate (Table 7). These were used as predictors in our models for somatic mutation probability. In order to streamline data preparation, predictors were pre-processed so that a single file encompasses the genome-wide signal of a tissue for each predictor, either in the form of a `BigWig` file (a binary file format for dense, continuous genomic data, Raney et al. 2023) or a `GenomicRanges` (`GR`) R object (Lawrence et al., 2009). For instance, replicates of histone modification ChIP-Seq measurements for a specific tissue were processed and combined into a single `BigWig` file. This `BigWig` file was then read out to extract the signal of this histone modification for specific positions. Exceptions were the features GC content and Coding effect score, which were computed in-place for each specific position. The sources and pre-processing steps for each predictor are described in the following.

Coding effect score

This feature represents the effect a mutation at a specific genomic position would likely have on the encoded transcript or protein. To that end, we used Sorting Intolerant From Tolerant (SIFT) (Vaser et al., 2016). The effect of a variant (i.e., whether it changes the protein sequence of the encoded gene or even leads to a stop-gain or stop-loss) depends on the exact type of mutation. For example an $A > T$ variant might be silent, while $A > G$ leads to a base exchange. Since the type of mutation was undefined for TNs, we considered every possible mutation for each query position. Therefore, for each query position, we generated three `vcf` entries, corresponding to the three possible nucleotide changes that could occur at this position. The resulting `vcf` file was used as input for SIFT. The column "VarType" of the SIFT output was then translated into a score: "NONCODING" mutations were assigned a score of 0, "SYNONYMOUS" mutations got a score of 1, "NONSYNONYMOUS" got a score of 2, and the remaining ("START-LOST", "STOP-GAIN", or "STOP-LOSS") were assigned a score of 3. The scores resulting from the three possible nucleotide changes for each position were then averaged and used as the effect score.

Non-B DNA

Non-B DNA describes DNA stretches that differ from the canonical B-DNA conformation. Annotation of regions in hg19 that were predicted to form non-B structures were downloaded from non-B DB (https://nonb-abcc.ncifcrf.gov; Cer et al., 2013). The database includes positions of sequence motifs that tend to form non-B conformations, including a-phased repeats, direct repeats, g-quadruplex forming repeats, inverted repeats, mirror repeats, short tandem

repeats, and Z-DNA motifs. For each of these features, we combined the annotations across all chromosomes into a `GR` object.

Conservation

A `BigWig` with conservation scores computed by UCSC based on multiple alignments of 99 vertebrate genomes to the human genome were downloaded from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP100way/hg19.100way.phyloP100way.bw) and used as-is.

GTEx eQTL

Expression quantitative trait loci (eQTLs) computed by Genotype-Tissue Expression (GTEx) v8 were downloaded from GTEx (GTEx Consortium, 2017). For each tissue, the `signif_variant_gene_pairs.txt.gz` file was loaded and converted into two `GR`s with either -log(nominal p-value) or the eQTL slope as the value column. Since v8 GTEx is based on hg38, positions were lifted over using the `rtracklayer` R package. For the brain predictors, all GTEx brain regions were merged together into one file for p-values and slope, respectively.

Transcription factor binding site annotations

Transcription factors (TFs) are proteins that bind to DNA and regulate gene expression. We used the UCSC TFBS track that represents TFBS clusters from ENCODE data (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredWithCellsV3.bed.gz) and converted it into a `GR` object with the number of cells in which each TFBS was identified as a meta column. Since ETS-type TFs were previously reported to be especially prone to mutation, we created a separate `GR` object with only the binding sites of TFs belonging to this family. The following TFs were considered to belong to this group: E1AF, EHF, ELF, ELF1, ELF2, ELF3, ELF4, ELF5, ELG, ELK1, ELK3, ELK4, ER71, ER81, ERF, ERG, ERM, ESE, ESE1, ESE2, ESE3, ESX, ETS, ETS1, ETS2, ETV1, ETV2, ETV3, ETV4, ETV5, ETV6, ETV7, FEV, FLI1, GABP, GABPA, MEF, NERF, NET, PDEF, PE-2, PE1, PEA3, PSE, PU.1, SAP-1, SAP1, SAP2, SPDEF, SPI, SPI1, SPIB, SPIC, TCF, TEL, and TEL2 (Gutierrez-Hartmann et al., 2007).

Expression

We created three separate tissue-specific predictors for RNA expression: expression in cancer, expression in normal tissue from cancer studies ("Normal expression"), and expression from healthy organs from GTEx.

The cancer expression data, based on RNA-seq provided by the Pan-Cancer

Atlas (PanCanAtlas) initiative (Hoadley et al., 2018), was downloaded from cBioPortal (Cerami et al., 2012; http://www.cbioportal.org/datasets) for each tissue (Breast: BRCA, Colorectum: COADREAD, Esophagus: ESCA, Brain: Glioblastoma multiforme (GBM) and LGG, Kidney: KIRC, Liver: LIHC, Lung: LUAD, Ovary: OV, Prostate: PRAD, Skin: SKCM). For each tissue type, we used the file `data_RNA_Seq_v2_expression_median.txt`. Since these files use Entrezgene IDs, the corresponding positions were taken from the ENSEMBL database for GrCh37 provided by `biomaRt`. For each tissue, we extracted the expression values across samples and calculated the log(median expression + 1). These were then saved as a `GR` object using the `biomaRt` gene coordinates. Note that annotation coordinates may overlap, which was accordingly handled downstream (i.e., if a query position overlapped with two gene annotations, their expression was averaged). Accordingly, for the "normal expression" (adjacent tissues from cancer studies), we used the file `data_RNA_Seq_v2_mRNA_median_normals.txt` for each tissue instead and processed them in the same way as the cancer expression.

For healthy expression, we downloaded median transcripts per million (TPM) per tissue from GTEx v7 (https://gtexportal.org) along with the gene annotation used in their pipeline, in order to get accurate gene coordinates for the expression values. These gene coordinates were used to create a `GR` object for each tissue with the gene positions as ranges an log(median expression+1) as value metacolumn. For brain, all GTEx tissues from the brain were taken together and averaged, since brain location information from GTEx was much more specific than for TCGA cancer samples. Again, coordinates may overlap, which was appropriately handled downstream.

Histone ChiP-seq

Suitable ChIP-Seq files for histone modifications were selected using the "Experiment search" feature on the ENCODE website (Luo et al., 2019), filtering for unperturbed, human samples and selecting appropriate tissues. The following histone modifications were available as ChIP-Seq targets: H3K27ac, H3K36me3, H3K9me3, H3K27me3, H3K4me3, H3K4me1, and H3K9ac. We downloaded the corresponding metafiles (Table SS1) further filtered them to exclude experiments with the keywords "extreme" or "severe" in columns `Audit.ERROR` or `Audit.NOT_COMPLIANT`. We only used files that were tagged as "released" (column `File.Status`) and that were either `bigWig` or `bed` files. The remaining experiments were then downloaded (Table SS2). `Bigwigs`, corresponding to the signal p-value along the genome, and `bed` files, corresponding to Irreproducible Discovery Rate (IDR) thresholded peaks, were handled separately. For the `bigWig` files, we first grouped the experiments based on their respective ChIP-Seq targets. Multiple files corresponding to the same target and tissue were combined by using `wiggletools` to compute

the mean p-value across the multiple `bigWig` files along the genome. The resulting `bedGraph` files (one for each ChIP-Seq target) were then lifted over from hg38 to hg19. After sorting and merging overlapping genomic intervals (using `bedtools merge`), the files were then converted back to a `bigWig` file using `bedGraphToBigWig`. The `bed` files containing the ChIP-Seq peak calls were combined as follows: they were first converted into `bedGraph` files, sorted and overlapping intervals were merged using `bedtools merge`. We then combined all files corresponding to the same target and tissue by counting how many of these files indicated a ChIP-Seq peak at each position (i.e., the peak "prevalence" across tissues), using `bedtools unionbedg`. The combined peaks were stored as a `GR` object, with the peak prevalence as a metadata column. These were then lifted over from hg38 to hg19 and sorted again. To summarize, for each tissue and each histone modification, we generated a `bigWig` file containing the averaged signal p-value, and a `GR` file containing the combined peak calls of all available ChIP-Seq experiments.

Transciption Factor ChiP-seq

We selected suitable files from the ENCODE website (Luo et al., 2019) by searching for experiments tagged as TF ChIP-Seq (assay_title = TF Chip-seq) and filtering for appropriate tissues (Table S3). The following transcriptional regulators were available as ChIP-Seq targets: CTCF, POLR2A, POLR2AphosphoS5, and EP300. The ChIP-Seq files were then processed exactly in the same way as the ENCODE histone ChIP-Seq experiments (Section 3.4). After further filtering of the metatable, we downloaded all `bigWig` or `bed` files, corresponding to signal p-value and peaks, respectively (Table SS4). In the end, for each tissue and each TF, we generated one `bigWig` file containing the averaged signal p-value, and on `GR` file containing the combined peak calls of all available ChIP-Seq experiments.

DNase-seq

Suitable tissue-specific DNA accessibility data based on DNAse-seq were selected in the "Experiment search" feature on the ENCODE website (Table S5). The DNAse-seq files were again processed in the same way as the ENCODE histone ChIP-Seq experiments (Section 3.4). This resulted in one `BigWig` and one `gr` file per tissue with average DNA accessibility p-value and combined peaks across ENCODE experiments, respectively (Table S6).

ATAC-seq

DNA accessibility data based on ATAC-seq were taken from ENCODE. Again, suitable datasets were selected in the "Experiment search" feature on the ENCODE website, filtering for unperturbed human ATAC-seq experiments from

appropriate tissues (Table S7). The filtered ATAC-seq files (Table S8) were also processed in the same manner as the ENCODE histone ChIP-Seq experiments (Section 3.4), leading to a `BigWig` file (DNA accessibility p-value) and a `gr` (combined peaks) file for each tissue.

DNA methylation

DNA methylation based on deep whole-genome bisulfite sequencing `wig` was taken from Loyfer et al. (2023). We downloaded the `bigWig` files for each tissue (Table S9). We then imported them into R and used a custom script to average the methylation proportions at the same positions over multiple experiments in the same tissue. The result was stored as a `GR` file for each tissue.

Hi-C data

Chromatin conformation data based on high-throughput chromosome conformation sequencing (Hi-C) (Belton et al., 2012) was taken from ENCODE. For each tissue, we selected suitable organ/tissue terms using the Experiment search feature on the ENCODE website (Supplementary Table SS10). We filtered the metatable further according to column `File.Status` = "released". We used Hi-C data in the form of genome compartment annotations (first principle component of a principle component analysis of chromosome-wide Hi-C measurements in `BigWig` format) as well as mapping quality thresholded contact matrix (pair-wise chromatin contacts between genomic regions in `hic` file format) (Table SS11).

For the genome compartments, we imported each `bigWig` file as a `GR` object using the `rtracklayer` package such that the normalized compartment score (i.e., the first principal component) was stored as a `GR` metadata column. The `GR`s were lifted over from hg38 to hg19 and sorted. Multiple `GR`s originating from the same tissue were combined into one `GRangesList` (`GRL`) object where each list item corresponded to one Hi-C experiment. Downstream, values were averaged over experiments for each position.

Replication timing

Replication timing was based on Repli-seq data from ENCODE, where replication timing was determined by sequencing newly replicated DNA from fractions of cells in specific cell cycle phases (G1/G1b, S1, S2, S3, S4, G2). Replication patterns can be visualized as a continuous function along the genome, based on the smoothed signal of the fraction profile ("Wave signal"). Peaks and valleys (local maxima and minima, respectively) in this curve represent replication initiation and termination zones, respectively. A metafile with available

replication timing datasets was downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeUwRepliSeq/files.txt). We then downloaded all files with the tags "WaveSignal", "Peaks", or "Valleys". The "Peaks" and "Valleys" annotations were `bed` files, which were converted into a `GR` object for each cell line. The "WaveSignal" files were `BigWigs` which were converted into a `GR` object with the WaveSignal as a metadata column. We assigned each tissue the cell line that most closely matched the corresponding tissue, based on cell type and origin tissue of the cell lines (Table 6).

Table 6: Cell lines used for each tissue for replication data.

Tissue	Cell line
Brain	SK-N-SH
Breast	MCF-7
Colon	NHEK
Esophagus	NHEK
Kidney	NHEK
Liver	HepG2
Lung	IMR90
Ovary	NHEK
Prostate	NHEK
Skin	NHEK

Distance to centromeres and telomeres

We downloaded genome assembly gaps and contigs for hg19 from UCSC (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/gap.txt.gz). There is a known error for the hg19 version of this file, so that telomeres are missing for chromosome 17. Those were added manually according to the positions of flanking non-telomere regions. Telomere and centromere positions were extracted (eighth column = "telomere" or "centromere", respectively), read into R and saved as `GR` objects.

GC content

We computed GC content by extracting the genomic sequence of a window around each variant from the reference genome fasta file using `bedtools slop` and computing GC proportion nucleotides within this window using `bedtools nuc` (column "X6_pct_gc" in the resulting output).

3.5 Data preparation

The predictors were mapped to query genomic positions (e.g., TPs and TNs) using a streamlined approach. As described above, for each predictor and each tissue, we created one file encompassing genome-wide values for this predictor, based on pre-processed and collated data from multiple files for each

tissue. Depending on the predictor, we used either GR/GRL objects or BigWig files, with exception of GC content and the effect score, which are computed for each position in-place. To create the dataset for model training and/or prediction, we simply read out the value corresponding to each query position (i.e. TP or TN positions) for each tissue from each GR or BigWig, using a custom R function. This function takes as input the file location of the bed file with the query positions, as well as a table indicating which predictors should be mapped in which way, thus acting as a configuration file (Tables SS13 and SS14). This table includes the file locations of the pre-processed predictor to be used, as well as further options for mapping. Relevant here are the following parameters: "range" indicates the size of the window around the query position that should be considered when reading out a predictor, from 1bp up to 1Mbp, while an "N/A" indicates only a 1bp window. The option "measure" defines the manner with which a predictor should be read out. Possible options are "distance" (distance from query position to closest feature in the GR object in bps, used for example to compute distance to telomeres and centromeres), "ifany" (indicating whether the query position overlaps any feature in the GR at all, resulting in a binary 0/1 predictor), "mean" (mean predictor value across a range around the positions, with regions that are not covered by the reference predictor file being ignored), "mean0" (same as "mean", only that non-covered regions are counted as 0 towards the mean), and "nHits" (number of GR features within the window around the query position). Finally, the option "transform" indicates whether the predictor should be further transformed by either taking the square root ("sqrt"), the logarithm ("log"), or used as-is (not applicable, N/A). The transform options were chosen to approximate the predictor distributions to a gaussian distribution and make them comparable across tissues and file sources.

The mapPredictors function extracts values from BigWigs using BigWigAverageOverBed, the "range" option is simply passed on to -sampleAroundCenter. GR objects were read out by creating a GR object from the query positions as well and extending the ranges in either direction according to the by range option. For the output options "mean" or "mean0", we used the function findOverlaps to find matches between query positions and predictor GR and took the mean of all overlap hits, counting non-covered positions as either 0 or not applicable (N/A) for "mean0" or "mean", respectively. When taking the mean, overlap width is taken into account (i.e., if only part of the range around a variant was covered in the GR object, only the regions which were actually overlapping were taken into account, the rest was considered non-covered). For output "measure" options "ifany", "nHits", or "distance", we used the output of the functions findOverlaps, countOverlaps, or distanceToNearest, respectively.

The output is a table with each row representing either a TP or TN position

and each column a predictor at a specific resolution (i.e. range, where applicable). These tables are stored alongside the source chromosome for each position, and used for model training and/or testing.

3.6 Modeling

Model training and prediction were performed based on a chromosome-wise cross-validation (CWCV) scheme: for each of the 22 autosomes, we trained models on all except this chromosome and then used these models to get predictions (and estimate performance) for the excluded chromosome. In addition, we created a "final" model including all chromosomes, which was often used for downstream analyses.

3.6.1 Random Forest

Random Forest (RF) is a machine learning technique for regression and classification, which consists of an ensemble of decision trees (Breiman, 2001). In each decision tree, the data is iteratively split based on predictor variables ("bagging") so that the variation of the outcome variable within the subgroups (nodes) is smaller than before the split. RF owes its robustness against overfitting and biases in the data to two separate random sampling steps: First, each tree is trained (grown) based on a bootstrap sample of the original data (data points that are outside of this bootstrap sample are called out-of-bag (OOB)); second, at each split only a random subset of predictors is considered for the split when choosing the variable that can decrease the outcome variable's variation the most. Due to the hierarchical organization of each tree, interactions between predictors are inherently modeled. Furthermore, feature selection is automatically included in the model training process, since the features that explain more outcome variable variation are selected as splitting variables in a node more often than uninformative features.

When used for prediction, new data points are run down each of the decision trees and assigned an outcome value based on the OOB data points falling into the same terminal nodes. The final prediction is then determined by taking the average of tree-wise predictions (in the case of regression), or considering the majority vote over all trees (for classification).

Furthermore, the relative relevance of each predictor for predictions can be extracted from the RF. For instance, the decrease in outcome variable variation (i.e., residual sum of squares, RSS, for regression, or gini impurity for classification) can be averaged over all splits on a specific variable in the entire RF. Alternatively, the prediction performance on the OOB data can be compared before and after permuting each individual predictor - the larger the decrease in prediction accuracy, the more important this feature is for correct prediction. However, importance values are biased towards predictors with many possible split levels - for example, continuous predictors will receive higher im-

portance values than binary variables even if both are uninformative, simply due to the fact that there are more splits possible for the continuous variable, thus increasing the probability of a favorable split by chance. This bias in predictor importance values can be corrected for based on data permutations, or by introducing random variable splits (see below).

In this study, we used the R package `ranger` for RF (Wright and Ziegler, 2017). RFs were grown with arguments `importance = 'permutation', respect.unordered.factors = 'partition', probability = TRUE, scale.permutation.importance = TRUE`, and defaults for all other parameters. In addition, we grew RFs with the argument `importance = 'impurity_corrected'`. The `'impurity_corrected'` importance measure is less biased for differing numbers of categories and category frequencies between predictors, which is achieved by comparing the gini importance of each variable with gini importance values achieved by randomly introduced splits on the same variable in the RF (Wright and Ziegler, 2017; Nembrini et al., 2018). Due to these randomly introduced splits, the corresponding RFs cannot be used for prediction, they were only used to estimate predictor importance.

3.6.2 Logistic regression

We used logistic regression as a parallel approach to RF. A generalized linear model (GLM) is applicable to a binomially distributed response variable. In short, it models the logarithm of the odds (log-odds) of the response variable being 1 (versus 0) based on the predictor variables, using the logit link function. We used the `glm` function in R with arguments `family = binomial(link = "logit")`. We computed predictor p-values were calculated using the `summary` function and then generated a second model with only the features that had a p-value<0.05. This two-step approach was done for each CWCV fold as well as for the final model.

3.6.3 Lasso with Stability Selection

Ss a complementary approach, we also implemented Least Absolute Shrinkage And Selection Operator (LASSO) combined with stability selection (SL). The LASSO is a regularization method that introduces an $l1$ penalty to the regression problem, encouraging sparsity in the coefficient estimates. The optimization problem for Lasso can be expressed as:

$$\hat{\beta} = \underset{\hat{\beta} \in \mathbb{R}}{\operatorname{argmin}} \, ||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

where y is the response vector, X is the predictor matrix, $\hat{\beta}$ are the model coefficients, p is the number of predictors, and λ is the regularization parameter controlling the trade-off between model fit and sparsity. This means that, de-

pending on the parameter λ, variable coefficients β will be sequentially set to zero, leading to a sparse model. While LASSO is effective at selecting a subset of predictors, its selection is sensitive to the choice of λ and may vary significantly with small changes in the data, especially in the presence of correlated predictors. Stability selection (Meinshausen and Bühlmann, 2010), addresses this sensitivity by introducing a resampling procedure. By evaluating the robustness of variable selection across multiple subsamples of the data (i.e. the stability), the influence of random fluctuations can be reduced. In short, a LASSO model is trained N times on random subsamples of the data and the rate of inclusion for each predictor (i.e., the fraction of iterations where each predictor had a non-zero parameter) is computed. For a pre-defined threshold π_{thr} with $0 < \pi_{\text{thr}} < 1$, the expected rate of falsely selected predictors, i.e. the per-comparison error rate (PCER), can be controlled over a space of regularization parameters Λ

$$\frac{E(V)}{p} \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{q_\Lambda^2}{p^2},$$

where $E(V)$ is the expected number of falsely selected predictors and q_Λ is the mean number of non-zero coefficients.

In this study, SL was performed using the R package `c060`, with $\pi_{\text{thr}} = 0.6$, and a PCER of 0.1.

3.6.4 Performance estimation

As stated above, performance estimations were always based on CWCV. We used the `ROCR` R package to compute receiver operating characteristic (ROC), area under the ROC (AUC), and precision-recall (PR) measures of prediction accuracy. The ROC curve visualizes the relationship between false positive rate (FPR, proportion of data points falsely classified as positives) and true positive rate (TPR or recall, i.e. the probability that a positive data point is correctly identified as positive) along all possible cutoff values of a predicting model. A ROC curve along the diagonal indicates that the model's predictions are equally informative as assigning random labels to each data point. The closer the ROC curve gets to the upper left corner, the better the prediction performance of the model is considered. The area under the ROC curve (AUC) is commonly used to summarize the performance estimate into a single value, with AUC values close to 0.5 indicating random-like performance and AUC values close to one representing a near-perfect prediction. A PR curve visualizes the relationship between precision (positive predictive value, i.e. the proportion of correctly classified positives among all positive predictions) and recall (TPR) along all possible cutoff values of the predictions. PR curves that run closer to the upper right corner indicate better model performance.

In some cases, we stratified the prediction by certain features (e.g., mutation type, patient of origin) before computing prediction performance. In such cases we faced the problem that these features were not defined for the TNs. Therefore, in those cases we randomly sampled TNs to match the sequence context distribution of the TPs. In other words, we paired each TP with a randomly selected TN with the same sequence context and put them in the same group for stratification.

3.6.5 Cross-tissue application

We applied the RF models trained in one tissue to the data of all other tissues and estimated the prediction performance. To that end, we used the CWCV fold which left out chromosome 1 during training, and predicted only on chromosome 1 of the foreign tissue. If a predictor was present in the tissue used for training, but missing in the foreign tissue, it was replaced with a dummy variable containing the mean value of the predictor in the training tissue.

4 Results

4.1 Summary of training data

In order to study somatic mutation occurrence, we used cancer mutations from the PanCanAtlas based on whole exome sequencing. As described above, the prevalence of positions under positive or negative selection can be assumed to be negligible (Section 1.6). Furthermore, we limited the impact of selection on our models by removing positions that were mutated more than once (Figure S1). We limited our analyses to the exome and to the autosomes due to better data availability in exonic regions for both mutations and predictors.

We wanted to evaluate the tissue-specificity of local mutation rates. To that end, we collected data from multiple tissues, including brain, breast, colon, esophagus, kidney, liver, lung, ovary, prostate, and skin. These tissues were chosen to include both fast-replicating tissues (e.g., colon and skin) and non-replicating tissues (e.g., brain). While mutations due to spontaneous CpG deamination (C>T mutations) predominate in all tissues, the type and absolute number of mutations between tissues reflects their exposures to environmental carcinogens (Figure 5). For example, skin is one of the tissues with the highest mutation burden due to its exposure to mutagenic UV light, leading to a particularly high abundance of C>T mutations at CC dimers. Similarly, the lung samples are characterized by C>A mutations associated with tobacco smoking. Thus, the data encompasses mutations from various mutagenic sources.

In order to predict somatic mutation frequency along the genome, we collated various genomic features which were previously shown to be correlated with somatic mutation rate, or which had a potential mechanistic connection with somatic mutation rate (Table 7). These genomic features, termed predictors in this study, were then used to predict the variation in somatic mutation rate occurrence along the genome. When possible, the predictors were tissue-specific, so that mutation rate differences between tissues might be attributed to a specific predictor's tissue-specificity.

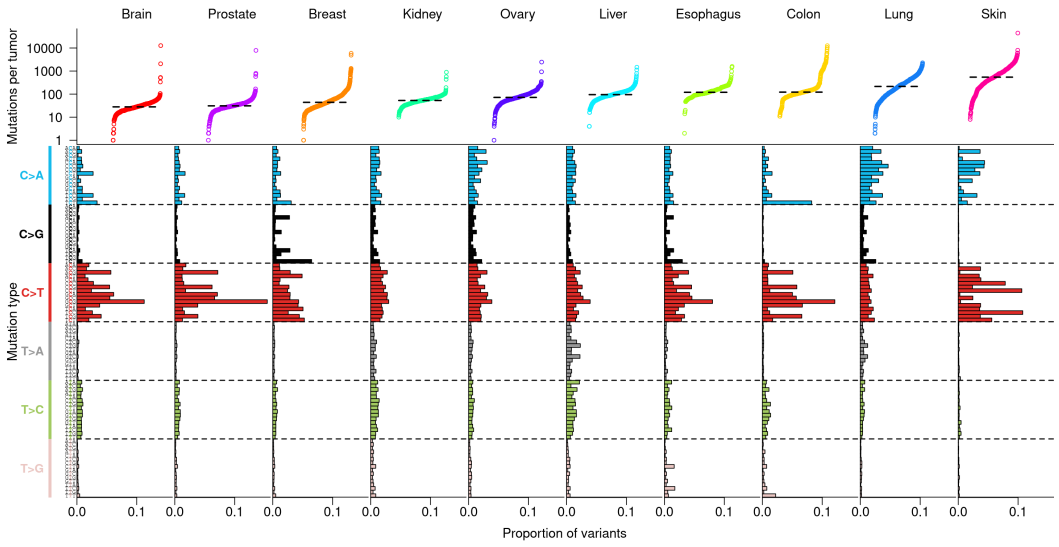

Figure 5: Overview of mutation data used in this study. Upper panel: number of mutations observed per patient for each cancer type. Horizontal lines represent the median. Lower panel: proportion of mutation types for each tissue. The six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Merging reverse complement equivalents results in the 96 possible mutation types depicted here. Thus, each bar represents the percentage of mutations that correspond to each mutation type. Figure design adapted from (Lawrence et al., 2013b).

Table 7: Overview of predictors included in the model. Abbreviations: Not applicable (N/A), Sorting Intolerant From Tolerant (SIFT), Genotype-Tissue Expression (GTEx), University of California Santa Cruz (UCSC), The Cancer Genome Atlas (TCGA), Base pair (bp), Kilo base pair (kbp), Mega base pair (Mbp), Expression quantitative trait locus (eQTL), Transcription factor binding site (TFBS), Chromatin Immunoprecipitation sequencing (ChIP-Seq), DNAse hypersensitive sites sequencing (DNAse-seq), Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq).

Predictor	**Description**		**Source**	**Tissue-specific**	**Window**
Coding effect	Score reflecting whether a mutation at a specific position is expected to be non-coding or silent (low score), lead to a mismatch (medium score), or is a nonsense mutation (stop-gain or similar, high score).		Reference genome/ SIFT	no	N/A
Non-B DNA	zDNA	Non-B DNA describes DNA regions that differ from the canonical B-DNA conformation.	non-B DB (Cer et al., 2013)	No	100bp
	Short tandem repeats				
	Mirror repeats				
	Inverted repeats				
	g-Quadruplex				
	Direct repeats				
	a-Phased repeats				
Conservation	Conservation over 99 vertebrate genomes		UCSC	No	N/A
eQTL	Slope	Effect size and (-log10) p-Values of eQTL loci in the vicinity of the variant, suggesting functional impact	GTEx	Yes	1kbp-1Mbp
	p-Value				
TF annotation	TFBS	Presence of a TFBS annotation	ENCODE	No	N/A
	TFBS density	Number of TFBSs annotations within the window			10kbp-1Mbp
	ETS TFBS	Presence of binding sites for TFs from the ETS family			N/A
	ETS TFBS density	Number of binding sites for TFs from the ETS family			10kbp-1Mbp
Expression	Cancer expression	RNA expression in cancer samples	TCGA	Yes	N/A
	Normal expression	Expression in samples from cancer-adjacent, histologically normal biopsies			
	Healthy expression	Expression in healthy tissues from donors without cancer	GTEx		
TF binding	PolR2A PolR2A-phospho	ChIP-Seq of RNA polymerase II and its phosphorylated counterpart, indicative of transcriptional activity. We used both calls as well as signal p-value along the genome.	ENCODE	Yes	signal: 1bp-1Mbp, peaks: N/A
	EP300	Histone acetyltransferase which acts as a transcription co-factor. Can facilitate transcription via chromatin remodeling.			
	CTCF	DNA-binding protein that is involved in control of the 3D structure of chromatin by forming chromatin loops, and thus influences gene expression.			
Histone marks	H3K9ac	ChIP-Seq of various histone modifications. We used both peak calls as well as signal p-value along the genome.	ENCODE	Yes	signal: 1bp-1Mbp, peaks: N/A
	H3K9me3				
	H3K4me3				
	H3K4me1				
	H3K36me3				
	H3K27me3				
	H3K27ac				
Methylation	Proportion of methylated CpGs within the window. DNA methylation acts as transcriptional repressor. Methylated CpG can spontaneously deaminate, leading to a C>T mutation.		(Loyfer et al., 2023)	Yes	100kbp
DNA accessibility	DNAse-seq	DNA accessibility based on genome-wide sequencing of regions sensitive to cleavage by DNase I. We used called peaks as well as signal p-values	ENCODE	Yes	signal: 1bp-1Mbp, peaks: N/A
	ATAC-seq	Identification of accessible DNA regions by probing open chromatin with Tn5 Transposase, followed by sequencing. We used called peaks as well as signal p-values			
Hi-C	Interactions	Number of chromatin contacts involving each position based on Hi-C.	ENCODE	Yes	N/A
	Compartment	First principal component of the Hi-C interaction matrix, which can be used to classify the genome into A/B compartments, corresponding to eu-/heterochromatin.			
Replication Timing	WaveSignal	Timing of DNA replication during cell division based on Repli-seq.	ENCODE	Yes (Cell lines)	N/A
	Peaks	Replication initiation zones			100bp
	Valleys	Replication termination zones			
Distance to centromere	Distance in bp to the nearest centromere.		Reference genome	No	N/A
Distance to telomere	Distance in bp to the nearest telomere.		Reference genome	No	N/A
GC content	Proportion of G or C nucleotides in a window around the variant.		Reference genome	No	10bp-100kbp

As training data, we created a dataset with equal numbers of TP and TN positions, which corresponded to mutated and non-mutated positions, respectively (Section 3.3). This approach was chosen as opposed to using the entire exome for training for two reasons: First, using the entire exome or genome for training would pose severe computational problems since the dataset would be too large for efficient training. Second, and more importantly, this way we address the problems that many machine learning models face with extremely imbalanced training data. In essence, most machine learning models try to minimize the overall prediction error. Thus, in the case of imbalanced data, always predicting the majority class (in this case non-mutated positions) would lead to a seemingly low prediction error. However, for many classification problems, like this study, the main interest lies in the minority case (in this case the mutated positions). By effectively down-sampling the majority class (TNs), we alleviate this class imbalance problem (Chen et al., 2004).

Aside from balancing the number of TPs and TNs, we also balanced the data with respect to sequence context pentamers (i.e., the two 5' and 3' nucleotides adjacent to each position). As a result, the proportion of each of the 512 possible pentamers (when merging reverse complement pentamers) was equivalent between the set of TPs and TNs. This was done to limit the effect of confounding by the sequence context with other genomic features.

We then annotated each of the training positions with each of the predictors using a streamlined approach where we created, for each tissue and predictor, a single file containing the genome-wide signal for this predictor, which was created by pre-processing and collating data from multiple files for each tissue (Section 3.4). Thus, predictors could be easily mapped to query positions (e.g., TP or TN positions) for each tissue (Section 3.5). Some predictors were read out in a window around the query position in order to even out potential noise in the predictor's data. For instance, ChIP-Seq measurements of histone modifications result in so-called peaks along the genome, which may vary in height and width due to sequencing artifacts. Furthermore, by design, histone modifications can only be detected in genomic regions that are associated with a nucleosome. However, the density of a certain histone mark over a larger genomic window, reflecting the more general chromatin state of this region (i.e. active/inactive), might be more relevant for the occurrence of somatic mutations. Thus, the predictors were not entirely independent from each other (Supplementary Figures S2-S11).

In summary, we created tissue-specific training data based on somatic mutations observed in cancers from 10 different tissues. In total, we included 1,170,928 mutations (TPs), ranging from 22,430 (kidney) to 456,455 (skin) mutations per tissue, and considered up to 146 different genomic features.

4.2 Comparison of modeling approaches

We applied and compared three machine learning techniques, RF, GLM (logistic regression), and LASSO with SL. These three methods represent a diverse spectrum of statistical and machine learning approaches for predictive modeling and feature selection. Logistic regression serves as a foundational statistical method that is widely used for binary classification due to its interpretability and well-established theoretical framework. RF, on the other hand, is a powerful ensemble machine learning technique that excels in handling nonlinear relationships and interactions among predictors, making it particularly robust in high-dimensional settings. Lastly, LASSO with SL was included due to its strength in high-dimensional data, where the number of predictors exceeds the number of observations. Its ability to perform simultaneous feature selection and regularization, coupled with the added robustness of SL, makes it a valuable method for identifying stable and interpretable predictors.

We used a chromosome-wise cross-validation (CWCV) approach to estimate model performance, where the data from each chromosome, in turn, was excluded during model training and then used as a test set to assay the model's performance (Section 3.6). This approach was chosen in order to prevent biases in performance estimation due to correlation between positions that are close to each other on the chromosome. For example, a more naive sampling approach such as randomly splitting TPs and TNs positions into training and test set irrespective of their chromosomal location, could lead to a setup where two TP positions that are close to each other on the chromosome are in the test and training set, respectively. These two positions, simply due to their proximity, will also have highly correlated genomic feature predictors and thus a model will predict the same mutation state. Thus, the test positions would not be independent of the training positions, leading to biased performance estimates. By using the CWCV approach, predictions are independent of such local effects and biases are avoided for performance estimation.

We compared the ability of RF, GLM and LASSO to distinguish mutated (TP) from non-mutated (TN) positions based on ROC and PR curves and the distribution of class distributions (Figures 6 and 7). The models were predictive (i.e., distinctly better than random predictions) with AUC estimates based on CWCV ranging between 0.54 and 0.64. The differences in performance between the tissues was bigger than that between the methods. The RF models performed slightly, but consistently, better than the other two approaches. This might be due to the fact that RF is able to capture non-linear associations of predictors with the mutation rate as well as interactions between predictors. Therefore, we decided to use the RF models for all further analyses.

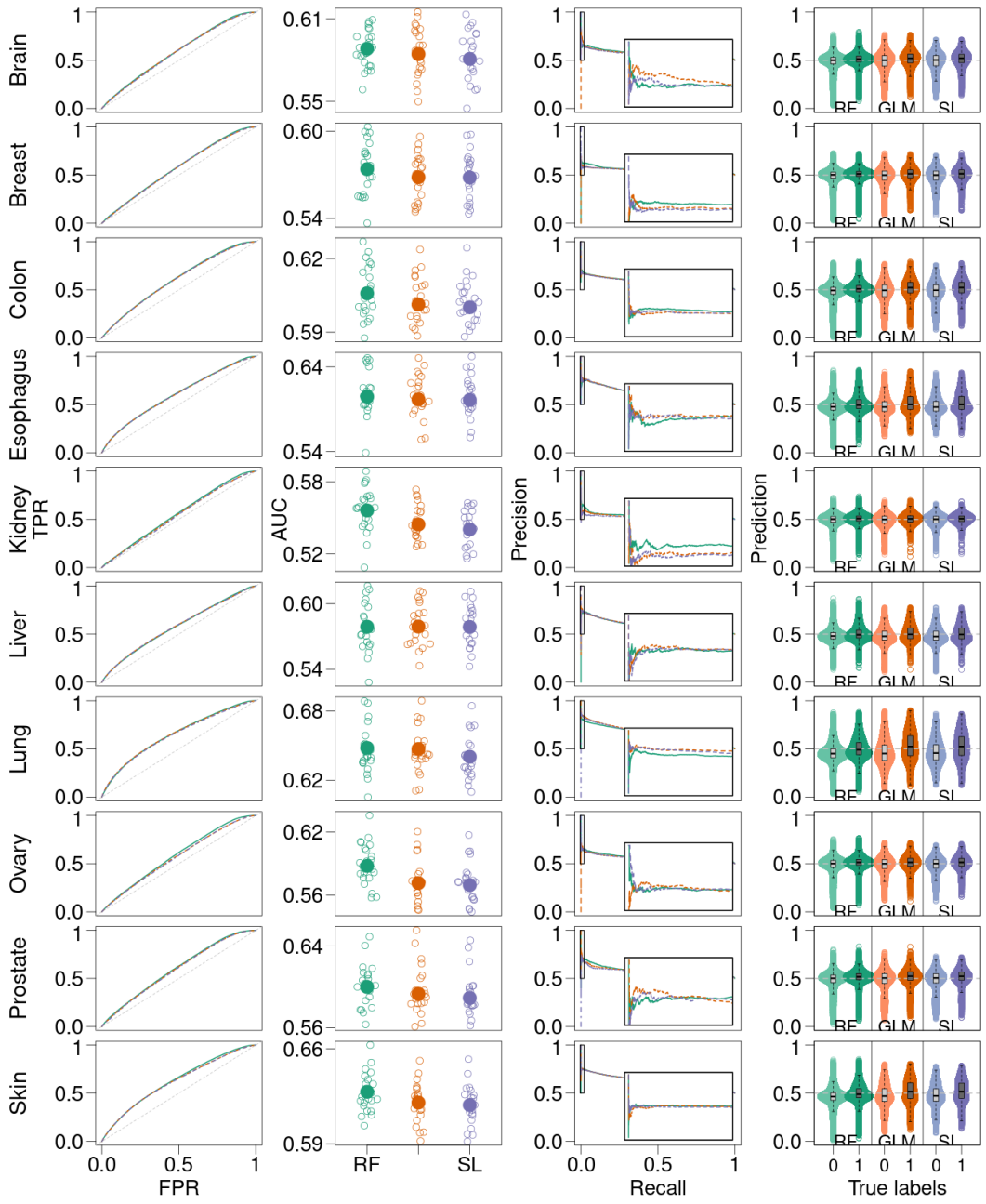

Figure 6: Performance comparison of model approaches. (**A**) ROC curves for Random Forest (RF), GLM, and LASSO with SL, based on the concatenation of chromosome-wise predictions for each tissue. (**B**) AUC estimates for each of the CWCV folds (light dots) and the estimate based on the entire exome (concatenation of CWCV folds, filled dots). (**C**) PR curves achieved by each method. The inset shows an magnification of the indicated left-most part of the curve in order to highlight the difference between the approaches. (**D**) Sina plot and overlayed boxplot of the prediction values stratified by the true TP/TN labels. The horizontal dashed line indicates a hypothetical cutoff at 0.5.

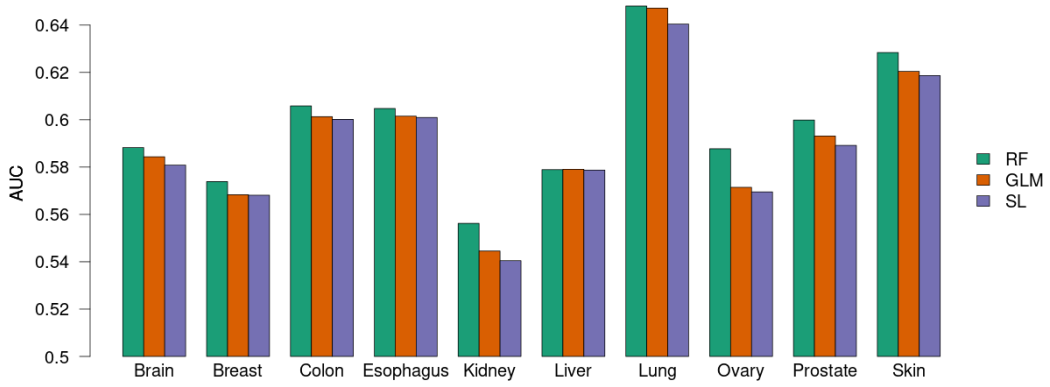

Figure 7: Comparison of AUC between model approaches. AUC was computed based on the concatenation of chromosome-wise predictions for each tissue and each method, RF, GLM, and LASSO with SL.

4.3 Model robustness

Before moving on to more detailed investigations of our models, we first wanted to ensure that they are robust and free of potential biases.

Test versus training performance

Our CWCV approach already ensures that the performance estimates are unbiased. Nevertheless, we confirmed that there were no signs of over-fitting by comparing the models' performance on the training data with the performance on the test data. There were negligible differences in performance between training and test data (Figure 8). Note that this analysis represents a rudimentary control for over-fitting, another test based on an independent dataset was performed as well (Section 4.7)

Consistency across cross-validation folds

There were only small differences in performance estimates between CWCV folds (Figure 9A-B). Notably, the difference in performance between samples was not due to different data size or other chromosomal features (Figure 9C). Furthermore, there were only minor differences in the predictor gini importance values across CWCV folds (Figure S12). Thus, there was no detectable bias introduced by our leave-one-chromosome-out approach.

Performance differences between tissues

There were clear differences in model performance between tissues (Figure 10A-B). Importantly, the tissues with larger datasets (i.e., Colon, Lung, and Skin) tended to achieve better performances (Figure 10C). There were stark

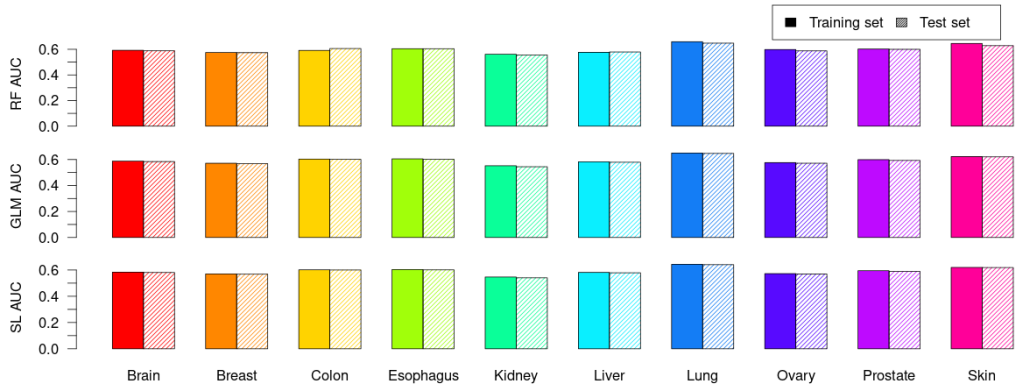

Figure 8: Test versus train performance. Comparison of training set performance with test set performance estimates of the concatenated CWCV iterations, for each tissue and each method. For each of the CWCV folds, we applied the model to the 21 chromosomes it was also trained on ("training" performance) as well as the left-out chromosome ("test" performance). The preictions over the CWCV folds were then concatenated and used to compute AUC.

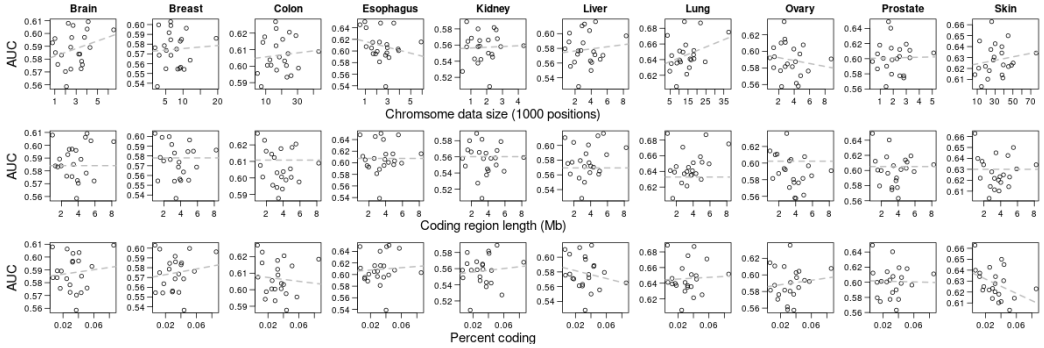

Figure 9: Evaluation of chromosome-wise cross-validation performance. Relationship between model performance (AUC) and either data size (number of TPs and TNs remaining after excluding each chromosome), total size of the coding regions of each chromosome, and the proportion of coding positions on each chromosome.

differences between tissues with respect to the number of mutations available for each tissue (Figure 10D), but also with respect to the number of samples included in the dataset (Figure 10E). Thus, the varying model performance could be due to the fact that machine learning models profit from larger datasets. In addition, the more mutations were observed in a tissue, the better the mutability along the genome can be understood: in tissues with few mutations, there might be lots of genomic positions that, although they theoretically have a high mutability, no mutation has occurred there. In contrast, tissues with more mutations are closer to a "saturation" of mutable positions.

To test whether the differences in model performance was due to data size, we

conducted two forms of downsampling analyses: First, we randomly sampled the mutations from each tissue to match the number of mutations available in the tissue with the fewest mutations (i.e., kidney). Second, we randomly sampled as many samples as there were in the tissue with the fewest available patients (i.e., breast). Note that in the second approach, the total amount of mutations may vary between tissues, since the sampled samples have differing numbers of mutations. Next, we created TN positions as before, with the modification that mutations which were removed during downsampling were not excluded during TN sampling. That means that some mutated positions that were removed during downsampling may be selected as TN. That way, we compensate for the potential advantage of tissues with more observed mutations. We then re-trained models and computed AUC based on CWCV. Downsampling did not have an effect on model performance (Figure 10F). Thus, there were likely other factors which determine the differences in predictive power across tissues, which were explored in a separate analysis (Section 4.5).

Impact of individual samples

As described above, specific mutation types are overrepresented in certain tissues compared to other tissues (Section 4.1). While these differences can be mainly attributed to tissue-specific mutational processes, it is also possible that there are samples with large numbers of atypical mutations that skew the mutation spectrum. So-called hypermutators are somatic clones that have exceptionally large amounts of mutations, for example due to intrinsic problems like defects in DNA polymerases or mismatch repair deficiencies, or external factors such as abnormally high mutation burden due to chemotherapies or other mutagens (Alexandrov et al., 2020).

In our training data, such samples sometimes contributed large proportions of mutations (Figure 11A). Notably, in brain and in prostate, there were hypermutator samples that respectively made up about a third of the data. We defined samples that contribute more than 5% of the mutations as potentially problematic hypermutators (Table 8). Some of these samples had a considerable effect on the observed mutation type spectrum (Figures S13-S19). Samples making up a large proportion of data in a tissue pose potential issues, since the model might learn mutational patterns of a hypermutator that might not generalize to other samples of the tissue. To test this, we computed the prediction performance after removing each of the top five hypermutator samples in each tissue from our data (Section 3.6.4, Figure 11B). Note that such an analysis is only possible for these samples with many mutations, since applying the models to patients with fewer mutations would result in unstable performance estimates. There was only one sample in brain (Patient1 in brain, TCGA-DU-6392) which noticeably influenced model performance.

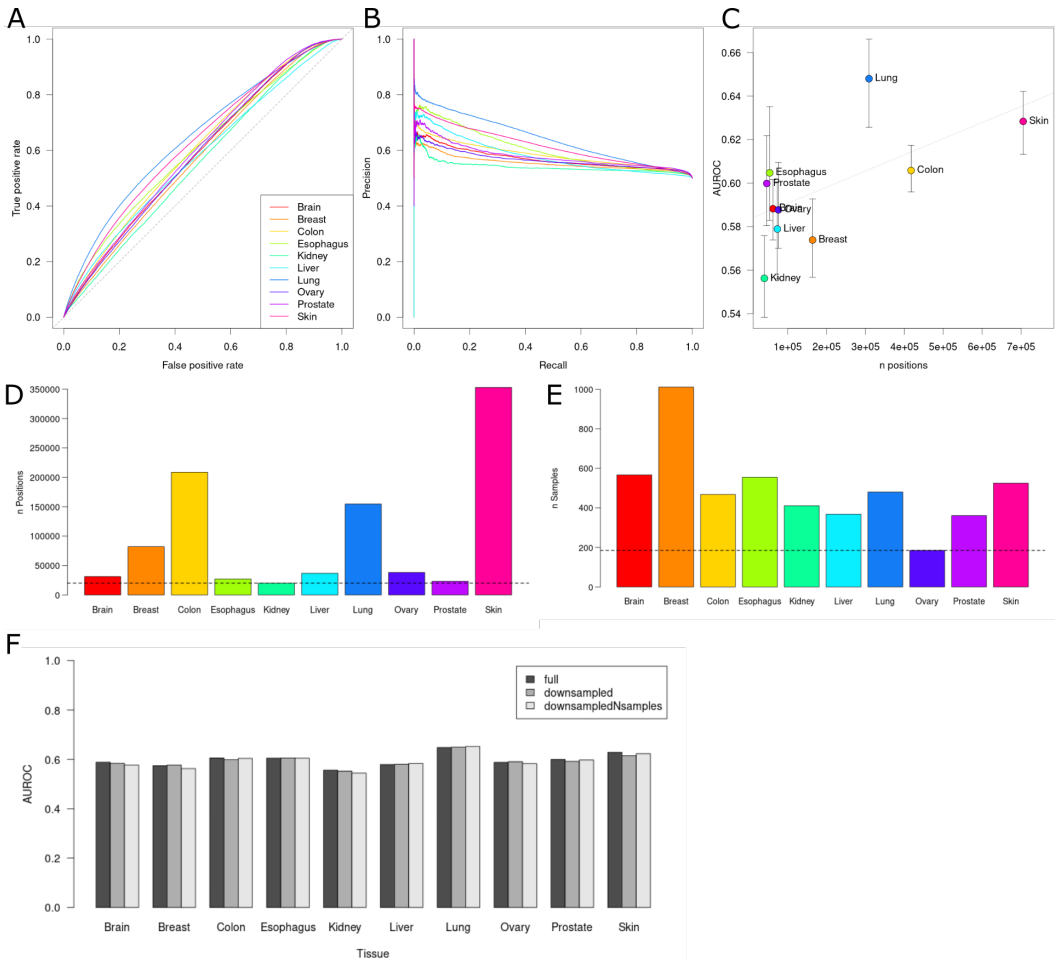

Figure 10: Difference of model performance between tissues. (**A**-**B**) ROC and PR curves achieved by each tissue. (**C**) Relationship between model performance (AUC) and data size (number of TPs and TNs positions) available for each tissue. Error bars represent the standard deviation of the AUC estimates across CWCV folds while the dots represent the estimate based on concatenated predictions. (**D**) Overview of the numbers of mutations available for each tissue in our data. The horizontal dashed line represents the lowest number of mutations available (i.e., kidney) which all other tissues were downsampled to. (**E**) Overview of the numbers of samples/patients making up the data for each tissue. The horizontal dashed line represents the lowest number of samples available across tissues (i.e., ovary) which was used to downsample all other tissues. (**F**) Prediction performance (AUC) of the models trained on either the full data for all tissues, data downsampled to match the number of mutations between tissues, or data downsampled to match the number of samples between tissues ("downsampledNsamples").

This sample seemed to carry mutations that are hard to predict and thus impairs model performance. However, it only made a difference when removed from the test data, re-training a model without this sample did not influence model performance (Figure 11C). Since this sample in our data, in the worst case, lead to an underestimation of model performance, we decided against excluding it for further analyses.

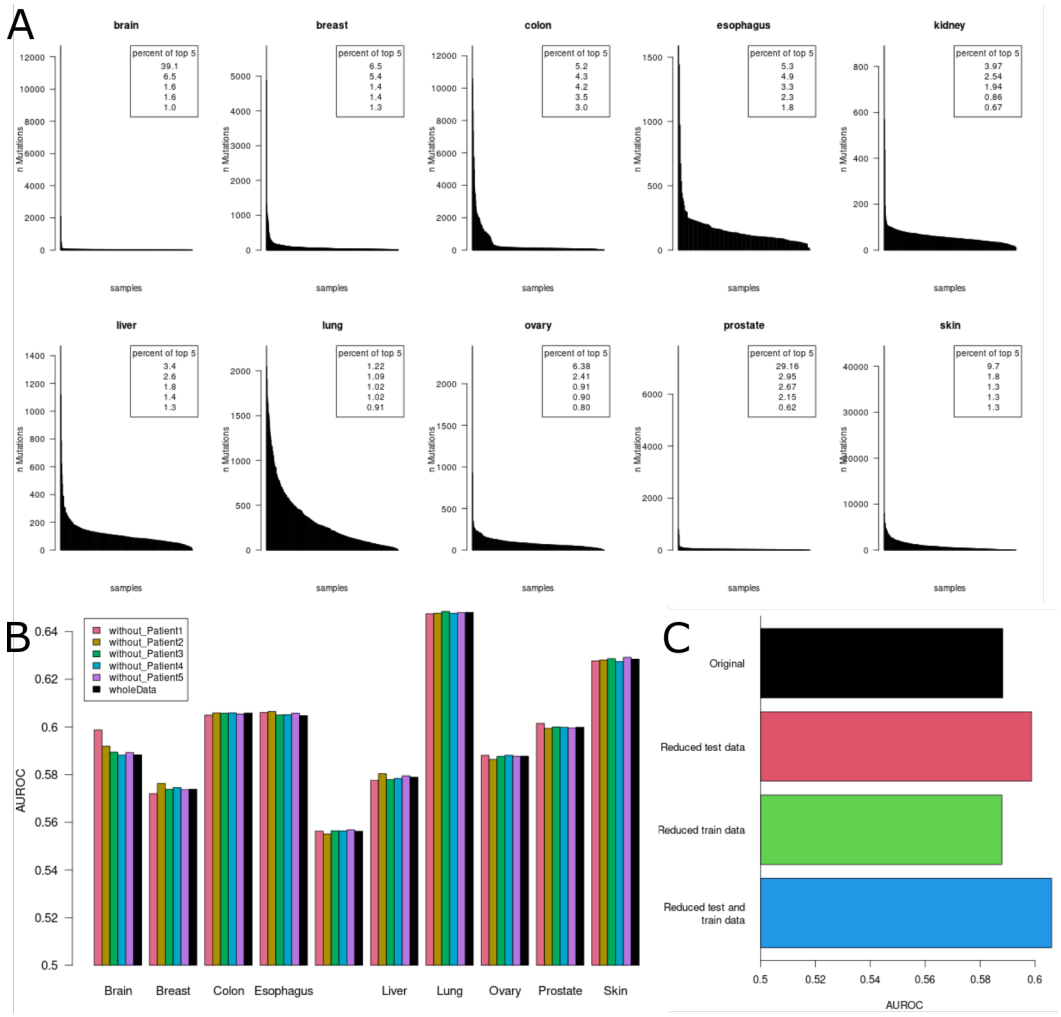

Figure 11: Effect of outlier samples. (**A**) Distribution of numbers of mutations across tissues for each tissue. The inset lists the proportion of data that the five samples with the most mutations ("hypermutators") contribute in each tissue. (**B**) Effect on prediction performance after removing each of the top five hypermutators from the test data. (**C**) Effect on model performance when removing brain sample TCGA-DU-6392 from the test data (i.e., model trained on all samples, but tested on all except TCGA-DU-6392), from the train data (i.e., model trained on all samples except TCGA-DU-6392, but tested on the full data), or both.

4.4 Model similarities between tissues

We used the RF gini importance to determine which genomic features were most influential in predicting the mutation rate along the genome (Figure 12). In general, the importance values were fairly similar between tissues. The features that consistently had high gini importance values across tissues were GC content, Replication WaveSignal, DNA methylation, H3K27ac peaks, H3K4me3 peaks, H3K9ac peaks, expression (all sources), TFBS density, and GTEx eQTL annotations. These findings are in line with previous

Table 8: List of hypermutator samples. All samples that made up more than 5% of the data for each tissue were defined as potentially problematic hypermutators.

Sample ID	Tissue	Percent mutations	Age	History of radiation therapy
TCGA-DU-6392	Brain	39.12	35	YES
TCGA-DU-6407	Brain	6.47	35	YES
TCGA-AN-A046	Breast	6.50	68	NO
TCGA-AC-A23H	Breast	5.41	90	NO
TCGA-AG-A002	Colon	5.18	35	NO
TCGA-L5-A4OI	Esophagus	5.35	79	NO
TCGA-09-2044	Ovary	6.38	77	NO
TCGA-XK-AAIW	Prostate	29.16	78	NO
TCGA-FW-A3R5	Skin	9.74	68	NO

investigations of mutation variation along the genome, where GC content, DNA replication timing, DNA methylation, histone marks, and TFBS were frequently determined as predictors of mutation rate (Section 1.7). The predictor importance of some genomic features in this exome-based model may be underestimated, since they play only a minor role in the exome. For instance, non-B DNA structures would hinder regular transcriptional activity and thus their presence in exonic regions is selected against. Similarly, most TFBSs lie outside of coding regions. This was the reason why we implemented many predictors with a genomic region. That way we could gauge their effect in smaller or larger genomic windows, even if we only observed very few such annotations in the exome.

Since we observed that the predictors were similarly important across tissues, we asked whether the relationship between the predictors and mutation rate were similar between tissues. To that end, we tested whether the models could be transferred between tissues by applying models trained in one tissue to the data of another tissue and computing the cross-tissue prediction performance (Section 3.6.5). The models could be applied to different tissues with only minor loss of accuracy (based on CWCV, Figure 13A). Thus, the tissue-specific models are in general transferable and predictive in other tissues. Surprisingly, some tissues were associated with better or poorer prediction accuracy independent on the origin tissue of the model. For instance, a model trained on breast tissue performed better on lung data than on the breast tissue data itself. This differs from the usual observation in machine learning that a model performs best on the data that it was trained on, and is less accurate on an independent dataset. Therefore, we were intrigued to understand the source of these performance discrepancies.

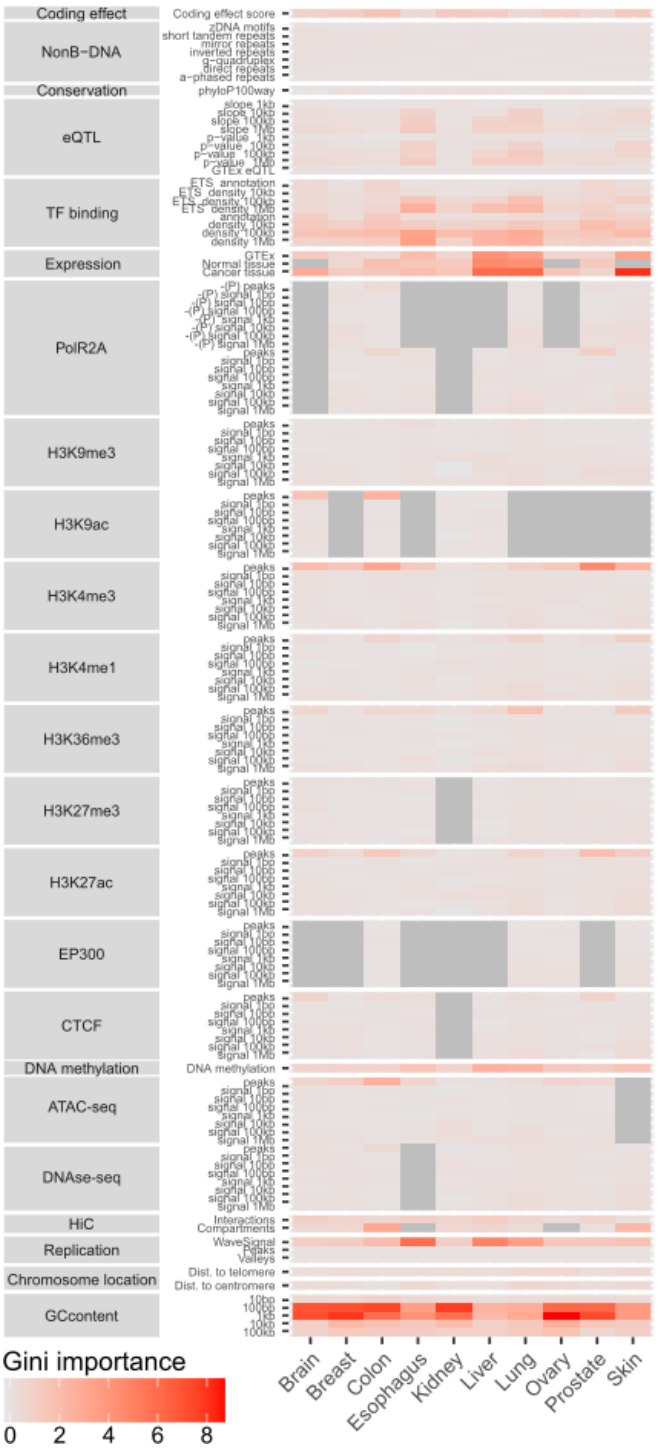

Figure 12: Random Forest Gini importance across tissues. Heatmap of gini importances values, scaled across tissues. Each row represents one genomic feature that acted as a predictor in our models. Grey fiels represent predictors that were not available in a tissue.

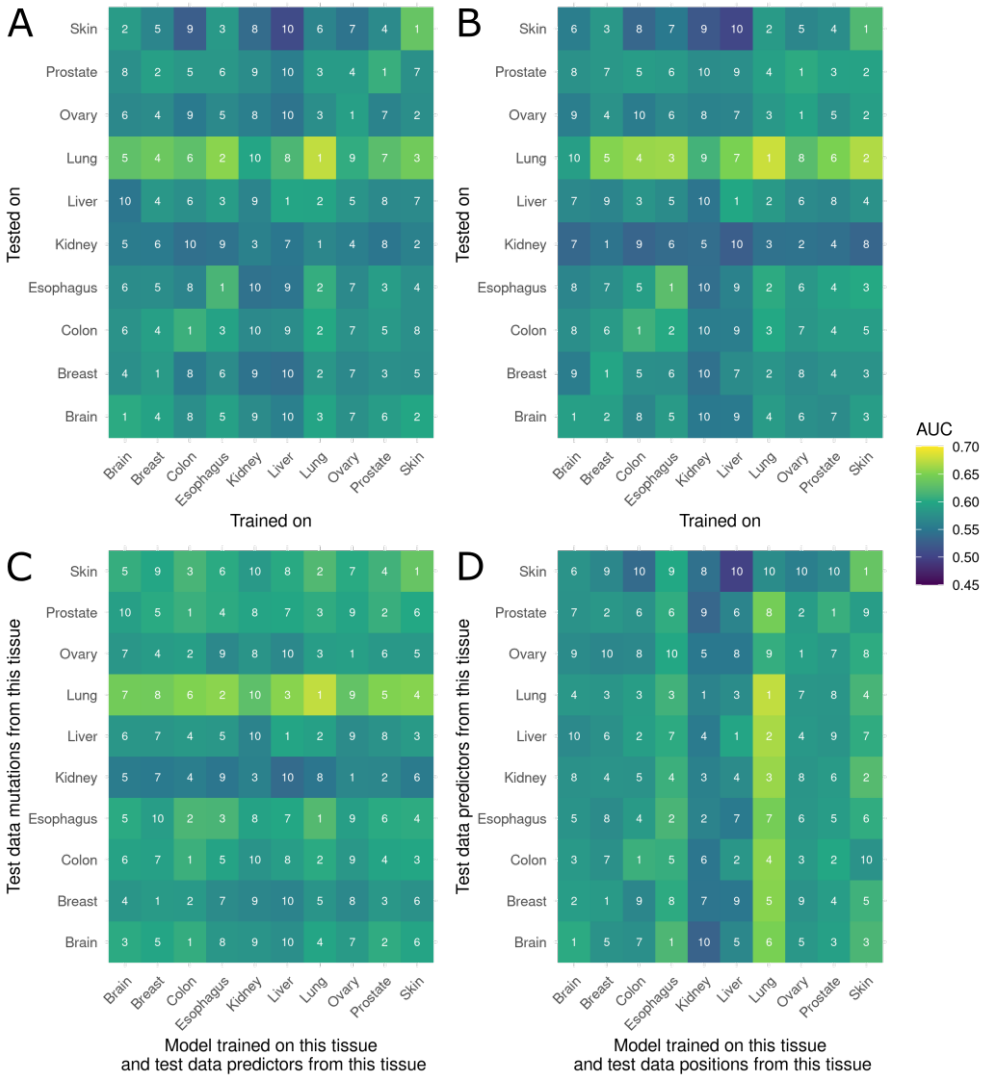

Figure 13: Cross-tissue application of the Random Forest models. (**A**) We applied RF models trained on one tissue (x-axis) to the data of all other tissues (y-axis), and estimated prediction performance (AUC based on CWCV), fill color). The numbers in the heatmap fields indicate the row-wise rank of the AUC value with lower numbers corresponding to best performance. (**B**) Same as (A), but after downsampling test and training data to match the tissue with the fewest mutations. (**C**) Each tissue-specific model (x-axis) was applied to the positions (i.e. TPs and TNs) from each tissue (y-axis), but using predictors originating from the same tissue that the model was trained on. The horizontal color bands indicate that the source of the mutations in the test data determine model performance. (**D**) Analogously to (C), the model from each tissue was applied to positions from the same tissue (x-axis), but using predictors from a foreign tissue (y-axis). The numbers in the heatmap fields indicate the column-wise rank of the AUC value with lower numbers corresponding to best performance. The vertical color bands indicate that the source of the predictors in the test data does not determine model performance.

First, we excluded that this phenomenon was simply due to the differing data size between tissues, by repeating the cross-tissue analysis using downsampled data (Figure 13B). That left us with two potential sources for the cross-tissue performance differences: First, there might be qualitative differences of the predictors between tissues. Although we took care to process each predictor in the same way between the various tissues, it is possible that there are inherent quality differences of the source data between tissues for some predictors. Second, there might be something inherent to the mutations in each tissue that determines how well it can be predicted. For example, maybe the tobacco-induced mutations that prevail in lung tend be more influenced by genomic features (e.g., have a stronger bias towards closed chromatin, etc.) compared to randomly occurring aging-associated mutations that prevail in other tissues. To test these two hypotheses, we performed similar cross-tissue analyses as above, only this time cross-combining predictors or mutation data between tissues. To that end, we prepared data where we mapped predictors to the training positions from each tissue, using the predictor files originally intended for the other tissues. For example, we mapped breast predictors to brain TP and TN positions. We then used these tissue-hybrid datasets to again test the cross-tissue prediction accuracy of our RF models (Figure 13C-D). The cross-tissue performance did not decline substantially when exchanging predictors, but was rather associated with the tissue source of the TP/TN positions. This analysis suggested that the mutations, not the predictors, were the determining factors in the varying tissue performance. Thus, in the following, we further explored how well varying mutation types can be predicted and how that might explain the performance differences between tissues.

4.5 Mutation type determines predictability

As described above, the lung mutations stood out against the other tissues by achieving good prediction performance, even when using models from a different tissue. Thus, we hypothesized that the mutation type determines how well mutations can be predicted. For example, mutations due to tobacco smoke (C>A, Figure 14A) might be more strongly influenced by genomic features compared to spontaneous clock-like mutations that are predominant in other tissues. The general relationship between each genomic feature and mutation rate variation along the genome would still be present, thus learnable, in all tissues, but it would be more pronounced in some tissues (e.g., lung) than in others. That would mean that we observe a better prediction performance in lung simply because this well-predictable mutation type is more prevalent (Figure 14B). We investigated this hypothesis by focusing the prediction performance evaluation to specific substitution types individually (Section 3.6.4). Subsetting the test data of each tissue to a specific substitution type highlighted mutation types that could be well predicted (Figure 14B). Indeed,

C>A substitutions can be well predicted, which is a mutation type that is especially frequent in lung samples (Figure 14C). Similarly, C>T mutations were relatively frequent in skin. They likely comprise of mutations due to UV light which are characterized by C>T transversions. Note that in this analysis, we cannot completely disentangle them from C>T mutations due to spontaneous deamination of methylated cytosine, which occur in all tissues. Another unexpected finding was an exceptionally high prediction accuracy for T>G mutations in esophagus, which likely correspond to Catalogue Of Somatic Mutations In Cancer (COSMIC) signature SBS17b (Figures 14B and S21, discussed in Section 5.1). Notably, extending this analysis to mutation types incorporating the 3' and 5' sequence context did not change any of the conclusions above (Figure S21).

Having established that certain mutation types can be better predicted by tissue-specific models than others, we now asked whether we can use this to explain the cross-tissue performance imbalance. Our hypothesis is that all models, no matter which tissue they were trained on, could perform relatively well on tobacco mutations. To that end, we combined the cross-tissue model application with the mutation type-specific performance estimation (Figure 14D). Note that C>A substitution are not only associated with tobacco exposure, but also other types of mutations. Vice versa, tobacco does not only cause C>A mutations, but also other substitutions, which makes it hard to disentangle tobacco mutations here (Figure S20). However, by subsetting our analysis to C>A substitutions, we enrich for tobacco mutations, and indeed, the increased prediction accuracy for lung data was especially pronounced when focusing on C>A mutations. In contrast, C>T mutations in skin can be well predicted especially by the skin model. This indicates that mutations due to UV light follow unique mutation patterns which the models trained on other tissues did not capture. Similarly, T>G mutations (potentially due to signature SBS17b, described above) were generally hard to predict except for the models trained on lung or esophagus.

Taken together, our analyses support our hypothesis that the types of mutations caused by tobacco (especially C>A) can be well predicted and that may be the reason why the models perform better on lung in our cross-tissue application.

4.6 All-tissue general model

We found that the tissue-specific models were relatively similar and that models could be transferred other tissues. However, some mutation types were rare in some tissues, which meant that their specific mutation patterns could not be captured in a tissue-specific model. Therefore, we decided to build one general, all-tissue model that captures the mutations of all tissues. To that end, we combined the mutation and predictor tables from all tissues (Figure S22).

In order to reduce the model complexity, we also removed predictors that had turned out to be uninformative for mutation prediction, namely ChIP-Seq-based predictors that corresponded to signal p-value along the genome. Instead, we only kept ChIP-Seq peak annotations. Furthermore, due to computational reasons, we had to reduce the predictor set to those that were available in all tissues (Figure 15A). We then trained new RF models on this all-tissue dataset and tested whether this general model was sufficient to predict tissue-specific mutation (Figure 15B). There was no substantial difference between the performance of the general model compared to the tissue-specific models. That means that, as intended, we have created a single model that captures the mutational patterns across tissues.

4.7 Validation on healthy tissues

In recent years it was discovered that healthy tissues may carry many somatic mutations (Sections 1.6), so we wondered whether the mutational patterns that were learned in cancer data also apply to healthy somatic tissues. Furthermore, we wanted to test our models on a completely independent data. To that end, we tested our models on somatic mutations in healthy tissues taken from SomaMutDB (Sun et al., 2022). Our tissue-specific models achieved relatively good prediction performances even in this independent dataset, which means that our model is valid beyond our training data (Figure 16).

Our all-tissue model seemed to perform equally well as the tissue-specific models in the ten tissues that were also included in the training data. Furthermore, our all-tissue general model was also applicable to other tissues that were not among the ten tissues in our training data. Exceptions were the tissues heart, kidney, and testicle, where we achieved lower prediction accuracy. For the heart, the AUC estimate is not reliable since there very few mutations in our test set (Figure S23). The lower prediction accuracy for kidney was observed both in our training data as well as the healthy tissue data, so it seems that its mutations are harder to predict. In general, the mutation landscape of kidney tumors is complex, since this tissue seems to be influenced by various mutagens and comprises of several mutational signatures. For example, SBS40a-c, a signature of still unknown aetiology with a diffuse mutational profile (Figure S20), contributes up to about 50% of mutations in renal tissues (Alexandrov et al., 2020; Senkin et al., 2024). Still, we achieve a better-than-random prediction accuracy in this tissue. Finally, the testes samples essentially correspond to sperm cells. Thus, they represent mutations of germ cells as opposed to somatic mutations. It is already known that germ cells follow a different mutation pattern, since DNA repair and replication checkmarks are much more active, in addition to the fact that sperm are haploid (Moore et al., 2021; Yang et al., 2021). Thus, it is not surprising that our model did not work there.

4.8 Model for whole genome

We extended the insights from the exome-based model to create a model for the entire genome. We used the same tissues, except colon and lung, because there was no whole genome mutation data available for them. The data processing and model architecture remained the same (Section 3.3). The only differences were the addition of predictors reflecting whether a position was coding/non-coding or annotated in a transcript or not. Furthermore, since the expression value for non-transcribed genomic regions is not defined, we added predictors where expression was computed in a window around each position (Table S14). This time, we only compared RF with logistic regression, which demonstrated that RF outperformed the GLM as for the exome data (Figure 17A).

Like in the exome model, we saw distinct performance differences between tissues, which coincided with data size (Figure 17C). The predictor importances were also similar between tissues (Figure 17D). Compared to the exome model, histone marks had lost relevance. This might be due to the fact that most histone modifications are associated with enhancers, promoters, or gene bodies. Since these regions make up only a small part of the entire genome, their effect seems to only be relevant when focusing on the coding genome. In summary, we have created a model that robustly models the somatic mutation along the human genome.

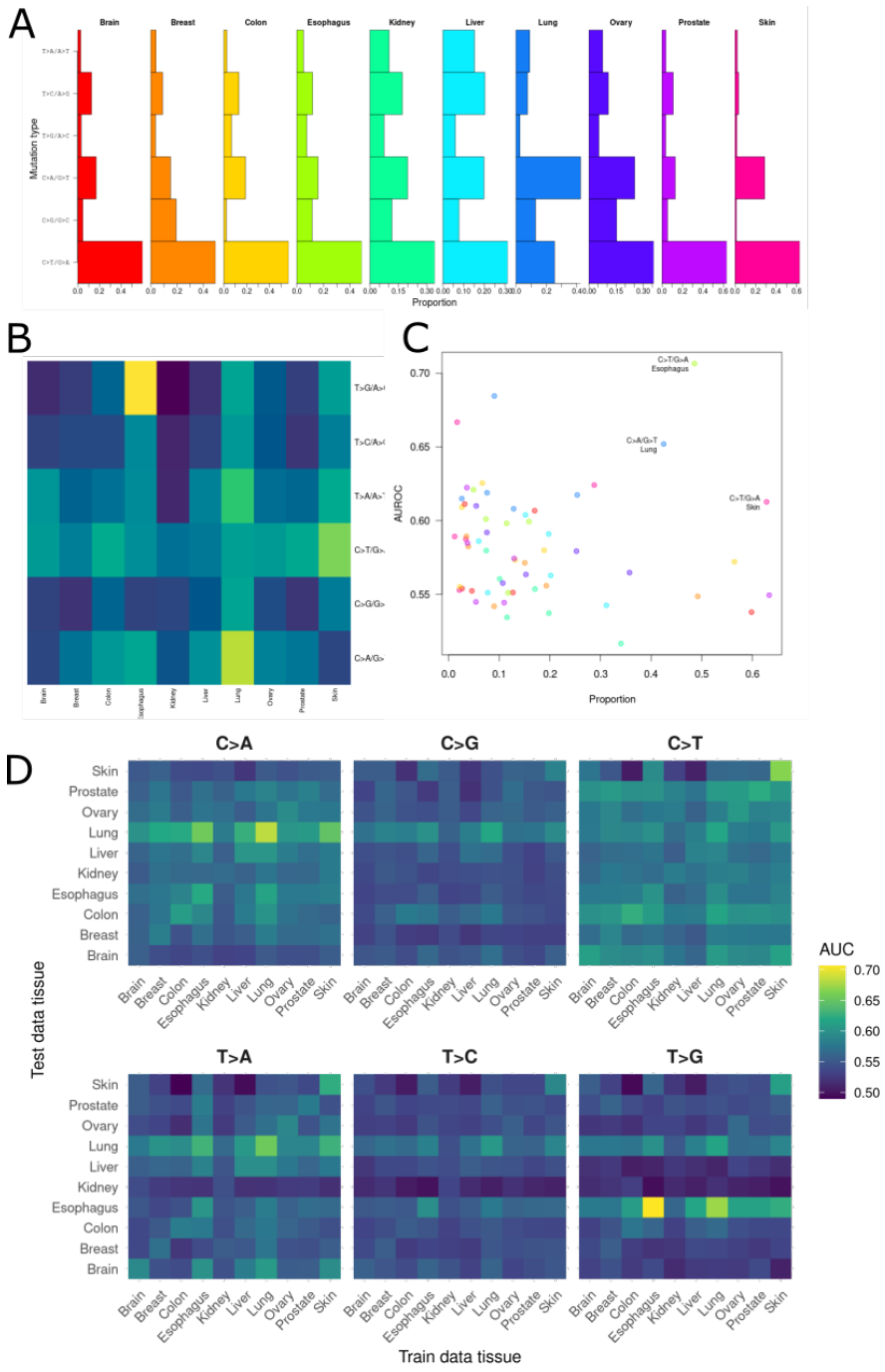

Figure 14: Effect of mutation type on model performance. (**A**) The composition of substitution types present in each tissue. (**B**) Each tissue-specific model (trained on entire data) was tested on the subset of positions corresponding to a certain substitution type and performance computed (AUC, fill color). (**C**) Relationship between the prediction performance shown in (B) with the proportion of mutations that the substitution type makes up in each tissue. Points outside the lower left quadrant (grey dashed line) were considered tissue-substitution type pairs that might explain the superior performance of some tissues. (**D**) RF models trained on one tissue (x-axis) were applied to the data of all other tissues (y-axis), respectively. We then stratified the predictions based on mutation type and computed prediction performance in the form of AUC (fill color, based on CWCV).

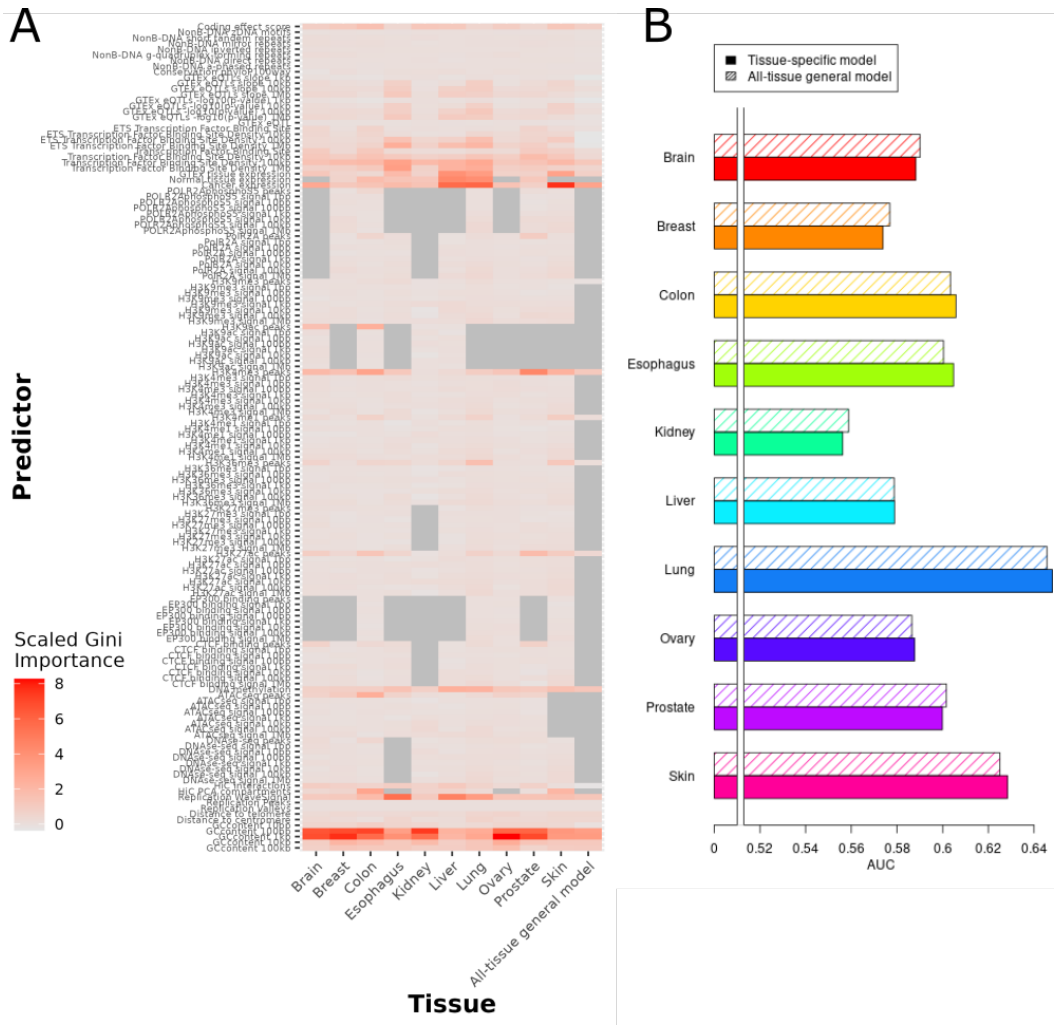

Figure 15: All-tissue general model. (**A**) RF gini importance values of the predictors in the tissue-specific models, compared to the all-tissue general model. (**B**) Prediction performance of the general model on each of the tissues, compared to the respective tissue-specific model.

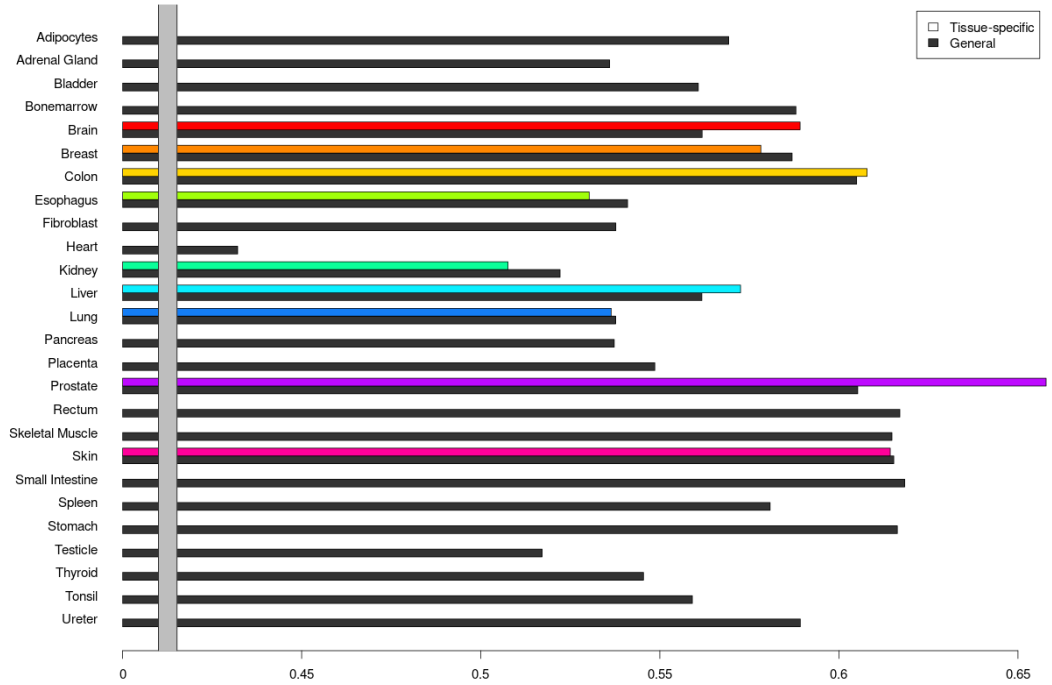

Figure 16: Model performance on healthy tissues. We applied the all-tissue general model and our tissue-specific models to mutation data from healthy tissues. Depicted are AUC values, comparing tissue-specific to general performance, where applicable.

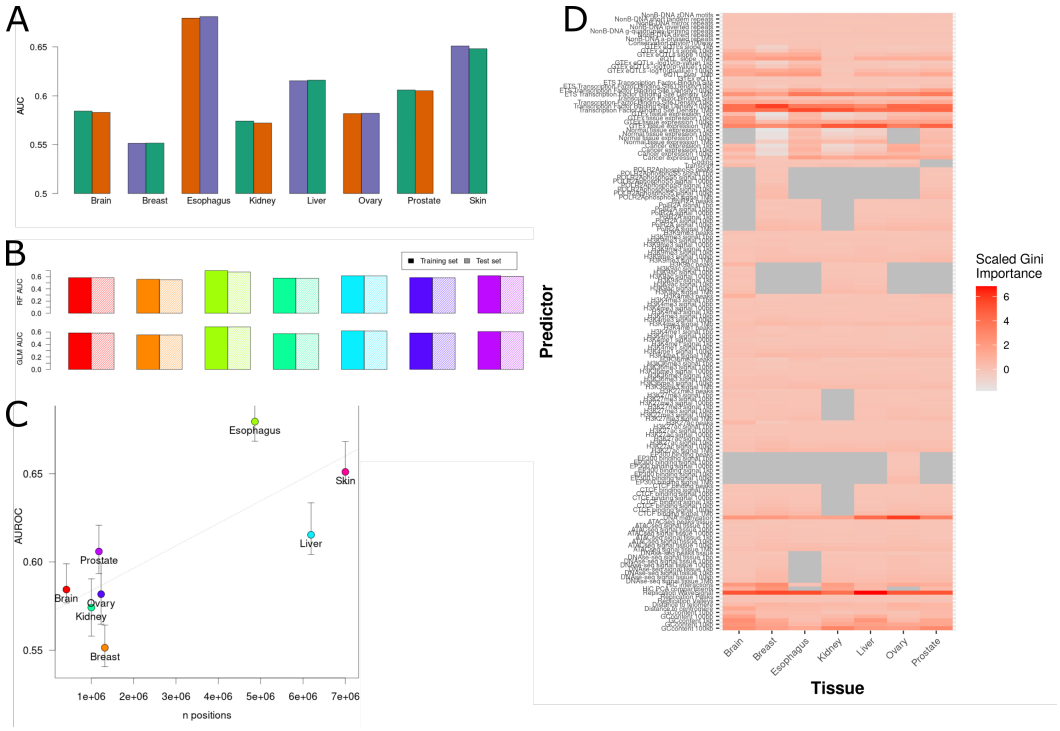

Figure 17: Model based on Whole Genome Data. (**A**) Comparison of CWCV-based performance between RF and logistic regression (GLM). (**B**) Comparison between test and training data performance of the RF whole genome model. (**C**) Dependency between model performance of each tissue and available data size. (**D**) We applied the whole genome model to the exome data and vise versa and computed prediction performance. For comparison, we also display AUC values achieved by each of the two models on the training data (with CWCV). (**E**) Scaled gini importance values across tissues.

5 Discussion

5.1 Conclusions

We collected tissue-specific data for a plethora genomic features which might influence somatic mutation rate and processed them into genome-wide reference files. These genomic feature files themselves represent a valuable resource for somatic mutations research. These genomic features were used to robustly differentiate mutated from non-mutated genomic positions in ten different tissues. We compared three machine learning techniques, namely RF, logistic regression (GLM) as well as LASSO with SL and found that RF slightly, but consistently outperformed the other two. We evaluated the method performance using CWCV as well a completely independent dataset. We were careful to evaluate and avoid potential biases due to selection (i.e., by removing recurrent mutations and using functional predictors as controls in our model), sequencing errors (i.e., by excluding known problematic regions), outlier hypermutator samples, as well as technical biases that are typical for machine learning, such as differing data size or over-fitting.

To our knowledge, the models in this study represent the most complete set of genomic features that may influence somatic mutation rate. Furthermore, the single-basepair resolution of the models presented here represent a considerable advancement to previous studies that mostly investigated somatic mutation rate variation along the genome on a large scale, often using Mbp-wide windows (Supek and Lehner, 2019; Polak et al., 2015; Sherman et al., 2022).

The best predictors of mutation rate according to RF gini importance were GC content, replication timing, DNA methylation, H3K27ac peaks, H3K4me3 peaks, H3K9ac peaks, RNA expression (all sources), TFBS density, and GTEx eQTL annotations. The influence of DNA methylation on somatic mutation occurrence is two-fold: firstly, spontaneous deamination of methylated Cytosine is one of the main sources of somatic mutation, so the abundance of such sites plays a major role. Secondly, large-scale genomic methylation rate is indicative of transcriptional regulation. Similarly, GC content and histone marks are proxies for gene activity, as well as, obviously, expression itself. This is in line with the observation that DNA accessibility, which is correlated with transcriptional activity, influences the effectiveness of DNA repair machinery. Furthermore, transcription-coupled repair limits the occurrence of mutations in active, thus relevant genes. Replication timing is also a commonly found feature that correlates with somatic mutation prevalence. Moreover, any studies have found various histone marks to be predictors of somatic mutation occurrence (Supek and Lehner, 2015; Schuster-Böckler and Lehner, 2012; Shuai et al., 2020; Polak et al., 2015; Juul et al., 2019), but

interestingly there is little overlap in which specific histone marks are most relevant. This is probably due to the fact that different histone marks are often correlated (positively and negatively). Furthermore, they are confounded because of the dependence on nucleosome occupancy: histone modifications can only be observed where DNA is wrapped around nucleosomes. The connection between TFs and mutation rate was also previously observed (Melton et al., 2015; Gonzalez-Perez et al., 2019). Finally, the presence of GTEx eQTL is another indicator of transcriptional activity, since local eQTL (e.g. eQTL that are close to the target gene) outnumber distant eQTL, and the association between a genetic variant and a transcript are much more likely to be significant for genes that are (highly) transcribed (GTEx Consortium, 2017). Surprisingly, non-B DNA structures had no predictive value in our model. This might be due to the fact that we excluded potentially problematic genomic regions, which included complex repeats. The same kind of repeats coincide with non-B DNA structures (Cer et al., 2013).

The fact that conservation and mutational impact (represented by SIFT score) were not relevant in our model reassured us that we had limited the impact of selection. We did not simply model which mutations occur due to their beneficial impact on cancer progression, but were closer to understanding how and where mutations happen *before* selection. In contrast, Bertl et al. (2018) had conservation as one of their predictive features. This is especially problematic when the model is used to detect cancer drivers, i.e. mutations that have a high functional impact.

One of the major findings of this study is the observation that models were relatively similar between tissues and could even be interchangeably be applied to each other. That suggests that the same underlying mechanisms act in all of the tissues. Furthermore, we created a general model that was applicable even to tissues that were not included in our training data. However, there are particular situations where having a tissue-specific model for the given tissue is advantageous. One such example are T>G mutations in the esophagus, which seemed to arise from a mutational process unique to the esophagus and thus could not be predicted by the other tissues' models. These mutations can be attributed to the COSMIC mutation signature SBS17b, which is of unknown aetiology (Figure S20). This signature is associated with precancerous lesions of Barrett's esophagus and seems to be specific to esophageal cancer and other tumors of the digestive tract (Busslinger, 2022; Alexandrov et al., 2020). It might be caused by oxidative damage after exposure to gastric acid or 5-fluorouracil treatment (Koh et al., 2021). Mutations from this signature have a tendency to occur in hotspots (Arnedo-Pac et al., 2023), which would explain why our model is so successful in predicting their occurrence. In contrast, lung mutations were consistently better predictable, even by tissue-foreign models. This seemed to be a result of a higher abundance of specific mutation types

(particularly C>A mutations) that tend to more strongly follow the patterns of mutation frequency along the genome. Since our lung data was based on Lung adenocarcinoma (LUAD), a cancer type strongly associated with a history of smoking, it stands to reason that the mutations due to tobacco are the ones that are responsible for the better predictability of lung mutations. This is supported by the fact that the healthy lung data (mostly non-smokers) suddenly did no longer display exceptional predictability.

Our models were applicable to healthy tissues, which indicates that they capture processes that are in operation before tumor emergence. Our models allow us to make predictions for pre-cancerous stages and mutational processes that are acting on normal tissues.

Finally, we used our insights to extend our approach from exome to the whole genome. That means that we present a genome-wide background model of mutation tendency across tissues. This model can be used to gain insights into the mutability of coding as well as non-coding genomic regions (Section 5.3).

5.2 Limitations of the Study

Any model can only be as good as the data it is based on. Consequently, we have to acknowledge that some of the predictors we have deemed uninformative in our study, might have simply not been captured due to poor data quality or inadequate processing. Our predictors come from various sources, which each bear the potential of small or large biases which could eventually have an impact on our model's accuracy. We had to limit the list of genomic features to those that were available from the same source for our wide range of tissues. Consequently, for example in the case of replication timing data, we resorted to using data based on cell lines. In order to reduce random variations and biases of individual experiments, we collated multiple datasets from the same tissues for each predictor, when available. While this approach limited the impact of outliers, it may have also artificially diluted real biological signals that were present in the data. For example, transient chromatin marks present only in a subset of the samples would have been diluted by averaging them with experiments where this peak was missing. This problem is one example of the trade-off between model generalization and interpretability. We combined multiple samples, which might actually follow completely different mutational patterns, genomic structures, or both. In other words, by combining data from many individuals, we no longer capture the individuality of patients. However, this generalization over samples was necessary due to the sparsity of somatic mutation data.

We based the selection of cancer mutation data for model training on the assumption that most mutations are neutral, i.e. do have not undergone pos-

itive or negative selection. Existing studies estimate that between 95-98% of all coding mutations are under neutral selection (Greenman et al., 2007; Martincorena et al., 2017). Especially in the context of cancers, this assumption has to be called into question, since tumors constantly undergo strong selective pressures in order to improve growth and to evade the immune system or cancer treatments, and circumvent cell death (Hanahan and Weinberg, 2011). However, separating the de novo occurrence of somatic mutations from selective pressures is almost impossible in somatic tissues. Even healthy tissues are also comprised of clonal expansions of somatically aberrant cells (Moore et al., 2021). The only way to separate mutation emergence from selection would by through DNA sequencing of single cells. To date, such data is still too scarce to provide sufficient data for studies such as this project.

In order to create balanced data for our machine learning models, we labeled observed mutations as TPs and other, randomly selected genomic positions as TN. This approach is only an approximation of the true mutational processes. Just because we have not observed a mutation at a certain position in our data, it does not mean that this position is not prone to somatic mutation, so labeling it a "True Negative" is technically not appropriate. In essence, stochasticity plays an considerable element in the emergence of somatic mutations, which we could only approximate with our model.

We compared three machine learning techniques in our study. Of course this represents only a small selection from the huge collection of machine learning models available. Logistic regression as a GLM is a relatively straightforward approach, which poses the main advantage that the model itself is easy to interpret and that the incorporation of formal statistics is straightforward. However, some of the assumptions of a GLM were potentially not completely met: we could not guarantee a linear relationship between predictors and the log-odds of the outcome variable. Furthermore, there was multicollinearity among predictors. Therefore, we also added the LASSO model combined with SL, which still renders an interpretable model, but deals better with correlated predictors. However, a RF model outperformed them both. While interpreting a RF model requires extra efforts (e.g., estimating predictor importance, effect directions or statistical significance), RF has the advantage that it makes no assumptions about data distribution or the form of relationship between predictors and response variable. Furthermore, it inherently captures interactions between predictors and can model complex causal relationships. Furthermore, RF has an in-built variable selection feature. As stated above, it is possible that application of another machine learning technique could further improve the prediction accuracy. One promising candidate would be deep neural networks, which have proven to perform well in mutation rate prediction (Sherman et al., 2022; Fang et al., 2022). Neural networks excel in capturing complex patterns and interactions in data, and are capa-

ble to interpret text-based data such as genomic sequence (Eraslan et al., 2019). However, neural networks, even more than RF models, require effort to interpret and require large amounts of data.

The sequence context and the type of mutation play a major factor in the mutability of a genomic position. The impact of the sequence context was previously extensively studied, for example through the analysis of mutational signatures. Therefore, we decided to disregarded the impact of sequence context and substitution type in our models. We also did this because evaluating each mutation type (i.e. substitution type along with sequence context) separately would have been problematic due to the resulting sparsity of data available for each mutation type. By analyzing all mutations together and balancing the training data with respect to sequence context, the different mutation types aided in learning the underlying mechanisms independent of sequence context. Indeed, we observed that our model was applicable across different mutation types, although with varying prediction accuracy.

5.3 Outlook

We have extensively analyzed our models in order to gain insights into somatic mutation mechanisms. Our model, along with the collection of tissue-specific, genome-wide genomic features, represent a valuable resource further further scientific questions. For example, we have opened up the possibility to explore how much our predictions vary on a small scale as compared to on a large genomic scale. Additionally, our models allow the investigation of how much mutation probability varies within one gene, or whether it rather affected across wider genomic regions.

As described above (Section 5.2), we have so far disregarded the influence of the sequence context on mutation rate. Combining our model with mutation probability estimates based on sequence context could thus boost the predictive power. This could be either achieved by creating a separate, signature-based model of mutation rate which is then integrated with our RF model. Alternatively, the possible mutation types could be modeled by a multinomial model, where we predict the possible mutation types (no mutation, C>A, C>T, A>G, ...). Furthermore, we do not capture the variation in mutation rate between individual patients or cancer sub-types. The implementation of patient-specific, cancer-type sensitive offsets could be used to address this.

Having established a robust model of base-pair resolution mutation susceptibility along the exome and the whole genome, this model can now act as a baseline for various applications. For example, it can be used as a background mutation rate for driver detection. Existing driver detection tools are based on background models of large-scale mutation variation (Sherman et al., 2022; Martincorena et al., 2017; Lawrence et al., 2013b; Shuai et al., 2020). Thus,

using a single-base-pair resolution may boost the detection of cancer drivers. Furthermore, we provide tissue-specific models of background mutation rate. For example, we could test whether a gene or even a specific genomic position is mutated more often than we would expect based on our model. We could then compare this approach with a more naive procedure where we assume equal mutation rates across the genome or for each gene. This is especially interesting in the context of tissue-specific cancer drivers: there are some genes that are consistently mutated in some tumor types, but never affected in other tissues (Martínez-Jiménez et al., 2020). While this is partly due to different selective pressures acting in various tissues, this phenomenon could, to date, not be completely explained. To address this, we could compare mutability and genomic predictors between tissues for specific cancer genes in order to try to understand why they are mutated in some tissues but not in others.

Similarly, our model could act as a baseline mutation rate for tumor evolution prediction. Previous attempts learned the evolution of a tumor through inferring the ordered sequence of mutations from sub-clones existing in a tumor sample (Auslander et al., 2019; Diaz-Colunga and Diaz-Uriarte, 2021; Li et al., 2024). Such phylogenetic models often assume a constant mutation rate along the genome, which, as outlined in this thesis, is not accurate. Thus, a tumor-specific model of the background mutation rate may improve the inference of tumor phylogeny (Goncearenco et al., 2017; Beichman et al., 2023). Indeed, incorporating genomic features was shown to be advantageous in the analysis of tumor progression (Ha et al., 2017; Li et al., 2024). This idea could even be extended to making predictions of prognosis, by estimating the likelihood of further detrimental mutations and their potential impact (Seifert et al., 2016; Schwartz and Schäffer, 2017).

Taken together, our model contributes to understand the origin of cancers, assess selective processes happening during tumor evolution, and estimate risks of somatic mutations.

References

Abascal, F., Harvey, L. M. R., Mitchell, E., et al. (2021). Somatic mutation landscapes at single-molecule resolution. *Nature*, 593(7859):405–410.

Aghili, L., Foo, J., DeGregori, J., and De, S. (2014). Patterns of somatically acquired amplifications and deletions in apparently normal tissues of ovarian cancer patients. *Cell Reports*, 7(4):1310–1319.

Akdemir, K. C., Le, V. T., Kim, J. M., et al. (2020). Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nature Genetics*, 52(11):1178–1188.

Albini, S., Zakharova, V., and Ait-Si-Ali, S. (2019). Chapter 3 - Histone Modifications. In Palacios, D., editor, *Epigenetics and Regeneration*, volume 11 of *Translational Epigenetics*, pages 47–72. Academic Press.

Alexandrov, L. B., , Kim, J., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101.

Alexandrov, L. B., Jones, P. H., Wedge, D. C., et al. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47(12):1402–1407.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.

Arndt, P. F., Hwa, T., and Petrov, D. A. (2005). Substantial Regional Variation in Substitution Rates in the Human Genome: Importance of GC Content, Gene Density, and Telomere-Specific Effects. *Journal of Molecular Evolution*, 60(6):748–763.

Arnedo-Pac, C., Muiños, F., Gonzalez-Perez, A., and Lopez-Bigas, N. (2023). Hotspot propensity across mutational processes. *Molecular Systems Biology*, 20(1):6–27.

Auguie, B. (2017). gridextra: Miscellaneous functions for "grid" graphics. *R package version 2.3*.

Auslander, N., Wolf, Y. I., and Koonin, E. V. (2019). In silico learning of tumor evolution through mutational time series. *Proceedings of the National Academy of Sciences*, 116(19):9501–9510.

Avgustinova, A., Symeonidi, A., Castellanos, A., et al. (2018). Loss of G9a preserves mutation patterns but increases chromatin accessibility, genomic instability and aggressiveness in skin tumours. *Nature Cell Biology*, 20(12):1400–1409.

Bacolla, A. and Wells, R. D. (2004). Non-B DNA Conformations, Genomic Rearrangements, and Human Disease. *Journal of Biological Chemistry*, 279(46):47411–47414.

Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395.

Bates, D., Maechler, M., Jagan, M., and Davis, T. A. (2021). Matrix: Sparse and dense matrix classes and methods. *R package version 1.3-2*.

Bayr, H. (2005). Reactive oxygen species. *Critical Care Medicine*, 33(12):S498.

Bębenek, A. and Ziuzia-Graczyk, I. (2018). Fidelity of dna replication - a matter of proofreading. *Current Genetics*, 64(5):985–996.

Beichman, A. C., Robinson, J., Lin, M., et al. (2023). Evolution of the mutation spectrum across a mammalian phylogeny. *Molecular Biology and Evolution*, 40(10).

Belton, J.-M., McCord, R. P., Gibcus, J. H., et al. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276.

Bengtsson, H. (2024). matrixstats: Functions that apply to rows and columns of matrices (and to vectors). *R package version 1.4.1*.

Benjamin, D., Sato, T., Cibulskis, K., et al. (2019). Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 861054.

Bertl, J., Guo, Q., Juul, M., et al. (2018). A site specific model and analysis of the neutral somatic mutation rate in whole-genome cancer data. *BMC bioinformatics*, 19(1):147.

Besnard, E., Babled, A., Lapasset, L., et al. (2012). Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature Structural & Molecular Biology*, 19(8):837–844.

Bianchi, J. J., Zhao, X., Mays, J. C., and Davoli, T. (2020). Not all cancers are created equal: Tissue specificity in cancer genes and pathways. *Current opinion in cell biology*, 63:135–143.

Blokzijl, F., de Ligt, J., Jager, M., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, 538(7624):260–264.

Boessenkool, B. (2020). berryfunctions: Function collection related to plotting and hydrology. *R package version 1.19.1*.

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

Brazhnik, K., Sun, S., Alani, O., et al. (2020). Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. *Science Advances*, 6(5):eaax2659.

Breen, A. P. and Murphy, J. A. (1995). Reactions of oxyl radicals with DNA. *Free Radical Biology and Medicine*, 18(6):1033–1077.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brown, A. J., Mao, P., Smerdon, M. J., Wyrick, J. J., and Roberts, S. A. (2018). Nucleosome positions establish an extended mutation signature in melanoma. *PLOS Genetics*, 14(11):e1007823.

Brunner, S. F., Roberts, N. D., Wylie, L. A., et al. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*, 574(7779):538–542.

Burtner, C. R. and Kennedy, B. K. (2010). Progeria syndromes and ageing: What is the connection? *Nature Reviews Molecular Cell Biology*, 11(8):567–578.

Bushman, D. M. and Chun, J. (2013). The genomically mosaic brain: Aneuploidy and more in neural diversity and disease. *Seminars in Cell and Developmental Biology*, 24(4):357–369.

Busslinger, G. A. (2022). Barrett's esophagus stages: their correlation with sbs17-associated dna mutations and the identification of histological marker genes. *Molecular & Cellular Oncology*, 9(1).

Cer, R. Z., Donohue, D. E., Mudunuri, U. S., et al. (2013). Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Research*, 41(Database issue):D94–D100.

Cerami, E., Gao, J., Dogrusoz, U., et al. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404.

Chatterjee, N. and Walker, G. C. (2017). Mechanisms of DNA damage, repair and mutagenesis. *Environmental and molecular mutagenesis*, 58(5):235–263.

Chen, C., Liaw, A., Breiman, L., et al. (2004). Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1-12):24.

Chen, Z., Yuan, Y., Chen, X., et al. (2020). Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports*, 10(1):3501.

Cheng, K. C., Cahill, D. S., Kasai, H., Nishimura, S., and Loeb, L. A. (1992). 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G-T and A-C substitutions. *The Journal of Biological Chemistry*, 267(1):166–172.

Chong, S. S., McCall, A. E., Cota, J., et al. (1995). Gametic and somatic tissue–specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. *Nature Genetics*, 10(3):344–350.

Cline, J., Braman, J. C., and Hogrefe, H. H. (1996). PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Research*, 24(18):3546–3551.

Colom, B., Herms, A., Hall, M. W. J., et al. (2021). Mutant clones in normal epithelium outcompete and eliminate emerging tumours. *Nature*, 598(7881):510–514.

Cooper, C. S., Eeles, R., Wedge, D. C., et al. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature Genetics*, 47(4):367–372.

Coorens, T. H., Oliver, T. R., Sanghvi, R., et al. (2021). Inherent mosaicism and extensive mutation of human placentas. *Nature*, 592(7852):80–85.

Costello, M., Pugh, T. J., Fennell, T. J., et al. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research*, 41(6):e67.

Dai, P., Wu, L. R., Chen, S. X., et al. (2021). Calibration-free NGS quantitation of mutations below 0.01% VAF. *Nature Communications*, 12(1):6123.

Dall'Agnese, G., Dall'Agnese, A., Banani, S. F., et al. (2023). Role of condensates in modulating dna repair pathways and its implication for chemoresistance. *Journal of Biological Chemistry*, 299(6):104800.

Danecek, P., Bonfield, J. K., Liddle, J., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008.

De, S. and Michor, F. (2011). DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nature Biotechnology*, 29(12):1103–1108.

De Bont, R. and van Larebeke, N. (2004). Endogenous DNA damage in humans: A review of quantitative data. *Mutagenesis*, 19(3):169–185.

Debès, C., Papadakis, A., Grönke, S., et al. (2023). Ageing-associated changes in transcriptional elongation influence longevity. *Nature*, 616(7958):814–821.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403.

DePinho, R. A. (2000). The age of cancer. *Nature*, 408(6809):248–254.

Dianov, G. L. and Hübscher, U. (2013). Mammalian Base Excision Repair: The Forgotten Archangel. *Nucleic Acids Research*, 41(6):3483–3490.

Diaz-Colunga, J. and Diaz-Uriarte, R. (2021). Conditional prediction of consecutive tumor evolution using cancer progression models: What genotype comes next? *PLOS Computational Biology*, 17(12):e1009055.

Dou, Y., Gold, H. D., Luquette, L. J., and Park, P. J. (2018). Detecting somatic mutations in normal cells. *Trends in genetics : TIG*, 34(7):545–557.

Dowle, M. and Srinivasan, A. (2021). data.table: Extension of 'data.frame'. *R package version 1.14.0*.

Dragulescu, A. and Arendt, C. (2020). xlsx: Read, write, format excel 2007 and excel 97/2000/xp/2003 files. *R package version 0.6.5*.

Dumanski, J. P., Lambert, J.-C., Rasi, C., et al. (2016). Mosaic Loss of Chromosome Y in Blood Is Associated with Alzheimer Disease. *American Journal of Human Genetics*, 98(6):1208–1219.

Duncan, A. W., Hanlon Newell, A. E., Bi, W., et al. (2012). Aneuploidy as a mechanism for stress-induced liver adaptation. *The Journal of Clinical Investigation*, 122(9):3307–3315.

Duncan, C. G., Grimm, S. A., Morgan, D. L., et al. (2018). Dosage compensation and DNA methylation landscape of the X chromosome in mouse liver. *Scientific Reports*, 8(1):10138.

Durand, N. C., Robinson, J. T., Shamim, M. S., et al. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, 3(1):99–101.

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, 4(8):1184–1191.

Ellis, P., Moore, L., Sanders, M. A., et al. (2021). Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nature Protocols*, 16(2):841–871.

Ellrott, K., Bailey, M. H., Saksena, G., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Systems*, 6(3):271–281.e7.

Enge, M., Arda, H. E., Mignardi, M., et al. (2017). Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell*, 171(2):321–330.e14.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403.

Erickson, R. P. (2014). Recent advances in the study of somatic mosaicism and diseases other than cancer. *Current Opinion in Genetics & Development*, 26:73–78.

Erickson, R. P. (2016). The importance of de novo mutations for pediatric neurological disease—It is not all in utero or birth trauma. *Mutation Research/Reviews in Mutation Research*, 767:42–58.

Fang, Y., Deng, S., and Li, C. (2022). A generalizable deep learning framework for inferring fine-scale germline mutation rate maps. *Nature Machine Intelligence*, 4(12):1209–1223.

Fischer, B., Smith, M., and Pau, G. (2020). rhdf5: R interface to hdf5. *R package version 2.34.0*.

Fragkos, M., Ganier, O., Coulombe, P., and Méchali, M. (2015). DNA replication origin activation in space and time. *Nature Reviews Molecular Cell Biology*, 16(6):360–374.

Frankish, A., Diekhans, M., Ferreira, A.-M., et al. (2018). Gencode reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773.

Fredriksson, N. J., Elliott, K., Filges, S., et al. (2017). Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLOS Genetics*, 13(5):e1006773.

Frommer, M., McDonald, L. E., Millar, D. S., et al. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831.

Fryxell, K. J. and Moon, W.-J. (2005). CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. *Molecular Biology and Evolution*, 22(3):650–658.

García-Nieto, P. E., Morrison, A. J., and Fraser, H. B. (2019). The somatic mutation landscape of the human body. *Genome Biology*, 20(1):298.

Garnier, S. (2018). viridis: Default color maps from 'matplotlib'. *R package version 0.5.1*.

Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M., and Nik-Zainal, S. (2018). Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Research*, 28(9):1264–1271.

Gilbert, D. M. (2002). Replication timing and transcriptional control: Beyond cause and effect. *Current Opinion in Cell Biology*, 14(3):377–383.

Goncearenco, A., Rager, S. L., Li, M., et al. (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Research*, 45(W1):W514–W522.

Gonzalez-Perez, A., Sabarinathan, R., and Lopez-Bigas, N. (2019). Local Determinants of the Mutational Landscape of the Human Genome. *Cell*, 177(1):101–114.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.

Greenman, C., Stephens, P., Smith, R., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213.

Guiblet, W. M., Cremona, M. A., Harris, R. S., et al. (2021). Non-b dna: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Research*, 49(3):1497–1516.

Guo, Y. A., Chang, M. M., Huang, W., et al. (2018). Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature Communications*, 9(1):1520.

Gutierrez-Hartmann, A., Duval, D. L., and Bradford, A. P. (2007). Ets transcription factors in endocrine systems. *Trends in Endocrinology and Metabolism*, 18(4):150–158.

Ha, K., Kim, H.-G., and Lee, H. (2017). Chromatin marks shape mutation landscape at early stage of cancer progression. *npj Genomic Medicine*, 2(1).

Haigis, K. M., Cichowski, K., and Elledge, S. J. (2019). Tissue-specificity in cancer: The rule, not the exception. *Science*, 363(6432):1150–1151.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674.

Hanawalt, P. C. and Spivak, G. (2008). Transcription-coupled DNA repair: Two decades of progress and surprises. *Nature Reviews. Molecular Cell Biology*, 9(12):958–970.

Hansen, A. S. (2020). CTCF as a boundary factor for cohesin-mediated loop extrusion: Evidence for a multi-step mechanism. *Nucleus*, 11(1):132–148.

Hansen, R. S., Thomas, S., Sandstrom, R., et al. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144.

Hara, R., Mo, J., and Sancar, A. (2000). DNA Damage in the Nucleosome Core Is Refractory to Repair by Human Excision Nuclease. *Molecular and Cellular Biology*.

Haradhvala, N. J., Polak, P., Stojanov, P., et al. (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell*, 164(3):538–549.

Harris, R. S., Petersen-Mahrt, S. K., and Neuberger, M. S. (2002). RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. *Molecular Cell*, 10(5):1247–1253.

Heitz, E. (1928). "Das" Heterochromatin der Moose. *Jahrbücher für wissenschaftliche Botanik*, 69:762–818.

Hinrichs, A. S., Karolchik, D., Baertsch, R., et al. (2006). The ucsc genome browser database: update 2006. *Nucleic acids research*, 34(suppl_1):D590–D598.

Hiratani, I., Takebayashi, S.-i., Lu, J., and Gilbert, D. M. (2009). Replication timing and transcriptional control: Beyond cause and effect. Part II. *Current opinion in genetics & development*, 19(2):142–149.

Hoadley, K. A., Yau, C., Hinoue, T., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304.e6.

Hodgkinson, A., Chen, Y., and Eyre-Walker, A. (2012). The large-scale distribution of somatic mutations in cancer genomes. *Human Mutation*, 33(1):136–143.

Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature Reviews Genetics*, 12(11):756–766.

Hu, J., Adebali, O., Adar, S., and Sancar, A. (2017). Dynamic maps of UV damage formation and repair for the human genome. *Proceedings of the National Academy of Sciences*, 114(26):6758–6763.

Huang, Z., Sun, S., Lee, M., et al. (2022). Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nature Genetics*, 54(4):492–498.

Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, 363(6429):558–561.

Jeon, Y., Bekiranov, S., Karnani, N., et al. (2005). Temporal profile of replication of human chromosomes. *Proceedings of the National Academy of Sciences*, 102(18):6419–6424.

Jiao, W., Atwal, G., Polak, P., et al. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications*, 11(1):728.

Jinks-Robertson, S. and Bhagwat, A. S. (2014). Transcription-associated mutagenesis. *Annual Review of Genetics*, 48:341–359.

Juul, M., Madsen, T., Guo, Q., et al. (2019). ncdDetect2: Improved models of the site-specific mutation rate in cancer and driver detection with robust significance evaluation. *Bioinformatics*, 35(2):189–199.

Kacher, R., Lejeune, F.-X., Noël, S., et al. (2021). Propensity for somatic expansion increases over the course of life in Huntington disease. *eLife*, 10:e64674.

Kaiser, V. B., Taylor, M. S., and Semple, C. A. (2016). Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLOS Genetics*, 12(8):e1006207.

Kalatskaya, I., Trinh, Q. M., Spears, M., et al. (2017). ISOWN: Accurate somatic mutation identification in the absence of normal tissue controls. *Genome Medicine*, 9(1):59.

Kazanov, M. D., Roberts, S. A., Polak, P., et al. (2015). APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. *Cell Reports*, 13(6):1103–1109.

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207.

Kim, S., Yu, N.-K., and Kaang, B.-K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental & Molecular Medicine*, 47(6):e166–e166.

Knaus, B. J. and Grünwald, N. J. (2017). Vcfr: A package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, 17(1):44–53.

Koh, G., Degasperi, A., Zou, X., Momen, S., and Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. *Nature Reviews Cancer*, 21(10):619–637.

Koren, A., Polak, P., Nemesh, J., et al. (2012). Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *American Journal of Human Genetics*, 91(6):1033–1040.

Kübler, K., Karlić, R., Haradhvala, N. J., et al. (2019). Tumor mutational landscape is a record of the pre-malignant state. *bioRxiv*, 517565.

Kucab, J. E., Zou, X., Morganella, S., et al. (2019). A Compendium of Mutational Signatures of Environmental Agents. *Cell*, 177(4):821–836.e16.

Kunkel, T. A. (2009). Evolving Views of DNA Replication (In)Fidelity. *Cold Spring Harbor Symposia on Quantitative Biology*, 74:91–101.

Kunkel, T. A. (2011). Balancing eukaryotic replication asymmetry with replication fidelity. *Current Opinion in Chemical Biology*, 15(5):620–626.

Kunkel, T. A. and Erie, D. A. (2005). Dna Mismatch Repair. *Annual Review of Biochemistry*, 74(1):681–710.

Lawrence, M., Gentleman, R., and Carey, V. (2009). Rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics*, 25(14):1841–1842.

Lawrence, M., Huber, W., Pagès, H., et al. (2013a). Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, 9(8):e1003118.

Lawrence, M. S., Stojanov, P., Polak, P., et al. (2013b). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

Lawson, A. R., Abascal, F., Coorens, T. H., et al. (2020). Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science*, 370(6512):75–82.

Lemon, J. (2006). Plotrix: a package in the red light district of r. *R-News*, 6(4):8–12.

Lercher, M. J., Urrutia, A. O., Pavlíček, A., and Hurst, L. D. (2003). A unification of mosaic structures in the human genome. *Human Molecular Genetics*, 12(19):2411–2415.

Levy, O., Amit, G., Vaknin, D., et al. (2020). Age-related loss of gene-to-gene transcriptional coordination among single cells. *Nature Metabolism*, 2(11):1305–1315.

Li, F., Mao, G., Tong, D., et al. (2013). The Histone Mark H3K36me3 Regulates Human DNA Mismatch Repair through Its Interaction with MutSα. *Cell*, 153(3):590–600.

Li, L., Xie, W., Zhan, L., et al. (2024). Resolving tumor evolution: a phylogenetic approach. *Journal of the National Cancer Center*, 4(2):97–106.

Li, R., Di, L., Li, J., et al. (2021). A body map of somatic mutagenesis in morphologically normal human tissues. *Nature*, 597(7876):398–403.

Li, R., Liang, F., Li, M., et al. (2018). MethBank 3.0: A database of DNA methylomes across a variety of species. *Nucleic Acids Research*, 46(D1):D288–D295.

Lim, B., Mun, J., Kim, Y. S., and Kim, S.-Y. (2017). Variability in Chromatin Architecture and Associated DNA Repair at Genomic Positions Containing Somatic Mutations. *Cancer Research*, 77(11):2822–2833.

Liu, L., De, S., and Michor, F. (2013). DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nature Communications*, 4:1502.

Lochovsky, L., Zhang, J., Fu, Y., Khurana, E., and Gerstein, M. (2015). LARVA: An integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Research*, 43(17):8123–8134.

Lodato, M. A., Rodin, R. E., Bohrson, C. L., et al. (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, 359(6375):555–559.

Lodato, M. A., Woodworth, M. B., Lee, S., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–98.

Loftfield, E., Zhou, W., Graubard, B. I., et al. (2018). Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Scientific Reports*, 8(1):12316.

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The Hallmarks of Aging. *Cell*, 153(6):1194–1217.

Loyfer, N., Magenheim, J., Peretz, A., et al. (2023). A dna methylation atlas of normal human cell types. *Nature*, 613(7943):355–364.

Lujan, S. A., Williams, J. S., and Kunkel, T. A. (2016). DNA polymerases divide the labor of genome replication. *Trends in cell biology*, 26(9):640–654.

Luo, Y., Hitz, B. C., Gabdank, I., et al. (2019). New developments on the encyclopedia of dna elements (encode) data portal. *Nucleic Acids Research*, 48(D1):D882–D889.

Lupiáñez, D. G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics*, 32(4):225–237.

Luquette, L. J., Miller, M. B., Zhou, Z., et al. (2021). Ultraspecific somatic SNV and indel detection in single neurons using primary template-directed amplification. *bioRxiv*, 2021.04.30.442032.

Ma, X., Shao, Y., Tian, L., et al. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biology*, 20(1):50.

Madireddy, A. and Gerhardt, J. (2017). Replication through repetitive dna elements and their role in human diseases. *DNA Replication: From Old Principles to New Discoveries*, pages 549–581.

Makova, K. D. and Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, 16(4):213–223.

Manders, F., van Boxtel, R., and Middelkamp, S. (2021). The Dynamics of Somatic Mutagenesis During Life in Humans. *Frontiers in Aging*, 2.

Mao, P., Brown, A. J., Esaki, S., et al. (2018). ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nature Communications*, 9(1):2626.

Mao, P., Smerdon, M. J., Roberts, S. A., and Wyrick, J. J. (2020). Asymmetric repair of UV damage in nucleosomes imposes a DNA strand polarity on somatic mutations in skin cancer. *Genome Research*, 30(1):12–21.

Marteijn, J. A., Lans, H., Vermeulen, W., and Hoeijmakers, J. H. J. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology*, 15(7):465–481.

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489.

Martincorena, I., Fowler, J. C., Wabik, A., et al. (2018). Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917.

Martincorena, I., Raine, K. M., Gerstung, M., et al. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041.e21.

Martincorena, I., Roshan, A., Gerstung, M., et al. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

Martínez-Jiménez, F., Muiños, F., Sentís, I., et al. (2020). A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20(10):555–572.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

Melton, C., Reuter, J. A., Spacek, D. V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, 47(7):710–716.

Milholland, B., Auton, A., Suh, Y., and Vijg, J. (2015). Age-related somatic mutations in the cancer genome. *Oncotarget*, 6(28):24627–24635.

Millikan, R., Hulka, B., Thor, A., et al. (1995). p53 mutations in benign breast tissue. *Journal of clinical oncology*, 13(9):2293–2300.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, 12(11):R112.

Moon, J., Kitty, I., Renata, K., et al. (2023). Dna damage and its role in cancer therapeutics. *International Journal of Molecular Sciences*, 24(5):4741.

Moore, L., Cagan, A., Coorens, T. H. H., et al. (2021). The mutational landscape of human somatic and germline cells. *Nature*, 597(7876):381–386.

Moore, L., Leongamornlert, D., Coorens, T. H., et al. (2020). The mutational landscape of normal human endometrial epithelium. *Nature*, 580(7805):640–646.

Moore, L. D., Le, T., and Fan, G. (2013). DNA Methylation and Its Basic Function. *Neuropsychopharmacology*, 38(1):23–38.

Müller, K. and Wickham, H. (2021). tibble: Simple data frames. *R package version 3.0.6*.

Nembrini, S., König, I. R., and Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, 34(21):3711–3718.

Neuwirth, E. (2014). Rcolorbrewer: Colorbrewer palettes. *R package version 1.1-2*.

Niimura, Y. and Gojobori, T. (2002). In silico chromosome staining: Reconstruction of Giemsa bands from the whole human genome sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):797–802.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., et al. (2012). Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, 149(5):979–993.

Nikopoulou, C., Parekh, S., and Tessarz, P. (2019). Ageing and sources of transcriptional heterogeneity. *Biological Chemistry*, 400(7):867–878.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2020). Biostrings: Efficient manipulation of biological strings. *R package version 2.58.0*.

Pedersen, T. L. (2024). ggforce: Accelerating 'ggplot2'. *R package version 0.4.2*.

Penagos-Puig, A. and Furlan-Magaril, M. (2020). Heterochromatin as an important driver of genome organization. *Frontiers in Cell and Developmental Biology*, 8.

Perera, D., Poulos, R. C., Shah, A., et al. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*, 532(7598):259–263.

Pfeifer, G. (2006). Mutagenesis at methylated cpg sequences. *DNA methylation: basic mechanisms*, pages 259–281.

Pfeiffer, F., Gröber, C., Blank, M., et al. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, 8(1):10950.

Pich, O., Muiños, F., Sabarinathan, R., et al. (2018). Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell*, 175(4):1074–1087.e18.

Piovesan, A., Pelleri, M. C., Antonaros, F., et al. (2019). On the length, weight and GC content of the human genome. *BMC Research Notes*, 12(1):106.

Polak, P., Karlić, R., Koren, A., et al. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364.

Poon, S. L., Huang, M. N., Choo, Y., et al. (2015). Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Medicine*, 7(1):38.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raney, B. J., Barber, G. P., Benet-Pagès, A., et al. (2023). The ucsc genome browser database: 2024 update. *Nucleic Acids Research*, 52(D1):D1082–D1088.

Rao, S., Huntley, M., Durand, N., et al. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680.

Ren, P., Dong, X., and Vijg, J. (2022). Age-related somatic mutation burden in human tissues. *Frontiers in Aging*, 3.

Rheinbay, E., Parasuraman, P., Grimsby, J., et al. (2017). Recurrent and functional regulatory mutations in breast cancer. *Nature*, 547(7661):55–60.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in genetics: TIG*, 16(6):276–277.

Rosendahl Huber, A., Van Hoeck, A., and Van Boxtel, R. (2021). The Mutagenic Impact of Environmental Exposures in Human Cells and Cancer: Imprints Through Time. *Frontiers in Genetics*, 12.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*, 532(7598):264–267.

Saini, N. and Gordenin, D. A. (2018). Somatic mutation load and spectra: A record of DNA damage and repair in healthy human cells. *Environmental and Molecular Mutagenesis*, 59(8):672–686.

Schaefer, M. H. and Serrano, L. (2016). Cell type-specific properties and environment shape tissue specificity of cancer genes. *Scientific Reports*, 6:20707.

Schmitt, A. D., Hu, M., Jung, I., et al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*, 17(8):2042–2059.

Schones, D. E., Cui, K., Cuddapah, S., et al. (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, 132(5):887–898.

Schuster-Böckler, B. and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, 488(7412):504–507.

Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229.

Seifert, M., Friedrich, B., and Beyer, A. (2016). Importance of rare gene copy number alterations for personalized tumor characterization and survival analysis. *Genome Biology*, 17(1).

Senkin, S., Moody, S., Díaz-Gay, M., et al. (2024). Geographic variation of mutagenic exposures in kidney cancer genomes. *Nature*, 629(8013):910–918.

Seplyarskiy, V. B., Akkuratov, E. E., Akkuratova, N., et al. (2019). Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nature Genetics*, 51(1):36–41.

Sherman, M. A., Yaari, A. U., Priebe, O., et al. (2022). Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. *Nature Biotechnology*, 40(11):1634–1643.

Shuai, S., Gallinger, S., and Stein, L. (2020). Combined burden and functional impact tests for cancer driver discovery using DriverPower. *Nature Communications*, 11(1):734.

Sidiropoulos, N., Sohi, S. H., Pedersen, T. L., et al. (2018). SinaPlot: An Enhanced Chart for Simple and Truthful Representation of Single Observations Over Multiple Classes. *Journal of Computational and Graphical Statistics*, 27(3):673–676.

Sill, M., Hielscher, T., Becker, N., and Zucknick, M. (2014). c060: Extended inference with lasso and elastic-net regularized cox and generalized linear models. *Journal of Statistical Software*, 62(5):1–22.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, 21(20):3940–3941.

Sivapragasam, S., Stark, B., Albrecht, A. V., et al. (2021). CTCF binding modulates UV damage formation to promote mutation hot spots in melanoma. *The EMBO journal*, 40(20):e107795.

Smith, K. S., Liu, L. L., Ganesan, S., Michor, F., and De, S. (2017). Nuclear topology modulates the mutational landscapes of cancer genomes. *Nature Structural & Molecular Biology*, 24(11):1000–1006.

Sondka, Z., Dhir, N. B., Carvalho-Silva, D., et al. (2023). Cosmic: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*, 52(D1):D1210–D1217.

Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., et al. (2009). Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41(4):393–395.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

Sun, S., Wang, Y., Maslov, A. Y., Dong, X., and Vijg, J. (2022). Somamutdb: a database of somatic mutations in normal human tissues. *Nucleic acids research*, 50(D1):D1100–D1108.

Supek, F. and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature*, 521(7550):81–84.

Supek, F. and Lehner, B. (2017). Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell*, 170(3):534–547.e23.

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

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93.

Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M., and Park, P. J. (2011). Impact of chromatin structure on sequence variability in the human genome. *Nature Structural & Molecular Biology*, 18(4):510–515.

Tomasetti, C. and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81.

Tomkova, M. and Schuster-Böckler, B. (2018). DNA Modifications: Naturally More Error Prone? *Trends in Genetics*, 34(8):627–638.

Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Böckler, B. (2018). Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biology*, 19(1):129.

Tost, J. (2010). DNA Methylation: An Introduction to the Biology and the Disease-Associated Changes of a Promising Biomarker. *Molecular Biotechnology*, 44(1):71–81.

Umer, H. M., Cavalli, M., Dabrowski, M. J., et al. (2016). A Significant Regulatory Mutation Burden at a High-Affinity Position of the CTCF Motif in Gastrointestinal Cancers. *Human Mutation*, 37(9):904–913.

Vaisman, A., McDonald, J. P., and Woodgate, R. (2012). Translesion DNA synthesis. *EcoSal Plus*, 5(1):10.1128/ecosalplus.7.2.2.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., et al. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43:11.10.1–33.

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nature Protocols*, 11(1):1–9.

Venkatesan, S., Swanton, C., Taylor, B. S., and Costello, J. F. (2017). Treatment-induced mutagenesis and selective pressures sculpt cancer evolution. *Cold Spring Harbor Perspectives in Medicine*, 7(8):a026617.

Vijg, J. (2021). From dna damage to mutations: All roads lead to aging. *Ageing Research Reviews*, 68:101316.

Vijg, J. and Dong, X. (2020). Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell*, 182(1):12–23.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., et al. (2013). Cancer Genome Landscapes. *Science*, 339(6127):1546–1558.

Voong, L. N., Xi, L., Wang, J.-P., and Wang, X. (2017). Genome-wide Mapping of the Nucleosome Landscape by Micrococcal Nuclease and Chemical Mapping. *Trends in Genetics*, 33(8):495–507.

Wang, K., Liu, H., Hu, Q., et al. (2022). Epigenetic regulation of aging: implications for interventions of aging and diseases. *Signal Transduction and Targeted Therapy*, 7(1).

Wang, L. and Di, L.-J. (2014). BRCA1 And Estrogen/Estrogen Receptor In Breast Cancer: Where They Interact? *International Journal of Biological Sciences*, 10(5):566–575.

Warnes, G. R., Bolker, B., Bonebakker, L., et al. (2024). gplots: Various r programming tools for plotting data. *R package version 3.1.3.1*.

Watson, I. R., Takahashi, K., Futreal, P. A., and Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nature Reviews Genetics*, 14(10):703–718.

Wei, T. and Simko, V. (2017). R package "corrplot": Visualization of a correlation matrix. *R package version 0.84*.

Weissensteiner, M. H., Cremona, M. A., Guiblet, W. M., et al. (2023). Accurate sequencing of dna motifs able to form alternative (non-b) structures. *Genome Research*, 33(6):907–922.

Weterings, E. and Chen, D. J. (2008). The endless tale of non-homologous end-joining. *Cell Research*, 18(1):114–124.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wickham, H. (2019). stringr: Simple, consistent wrappers for common string operations. *R package version 1.4.0*.

Wickham, H. (2020). tidyr: Tidy messy data. *R package version 1.1.2*.

Wickham, H. and Bryan, J. (2019). readxl: Read excel files. *R package version 1.3.1*.

Wickham, H., François, R., Henry, L., and Müller, K. (2021). dplyr: A grammar of data manipulation. *R package version 1.0.4*.

Wickham, H., Pedersen, T. L., and Seidel, D. (2023). scales: Scale functions for visualization. *R package version 1.3.0*.

Woo, Y. H. and Li, W.-H. (2012). DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nature Communications*, 3(1):1004.

Wood, H. M., Conway, C., Daly, C., et al. (2015). The clonal relationships between pre-cancer and cancer revealed by ultra-deep sequencing. *The Journal of Pathology*, 237(3):296–306.

Wright, M. N. and Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77:1–17.

Xing, D., Tan, L., Chang, C.-H., Li, H., and Xie, X. S. (2021). Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proceedings of the National Academy of Sciences of the United States of America*, 118(8):e2013106118.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24.

Yadav, V. K., DeGregori, J., and De, S. (2016). The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Research*, 44(5):2075–2084.

Yang, X., Breuss, M. W., Xu, X., et al. (2021). Developmental and temporal characteristics of clonal sperm mosaicism. *Cell*, 184(18):4772–4783.e15.

Yazdi, P. G., Pedersen, B. A., Taylor, J. F., et al. (2015). Increasing Nucleosome Occupancy Is Correlated with an Increasing Mutation Rate so Long as DNA Repair Machinery Is Intact. *PLOS ONE*, 10(8):e0136574.

Yokoyama, A., Kakiuchi, N., Yoshizato, T., et al. (2019). Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*, 565(7739):312–317.

Yoshida, K., Gowers, K. H. C., Lee-Six, H., et al. (2020). Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*, 578(7794):266–272.

Youk, J., Kwon, H. W., Kim, R., and Ju, Y. S. (2021). Dissecting single-cell genomes through the clonal organoid technique. *Experimental and Molecular Medicine*, 53(10):1503–1511.

Zeileis, A., Fisher, J. C., Hornik, K., et al. (2020). colorspace: A toolbox for manipulating and assessing colors and palettes. *Journal of Statistical Software*, 96(1):1–49.

Zeileis, A. and Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.

Zerbino, D. R., Johnson, N., Juettemann, T., Wilder, S. P., and Flicek, P. (2014). WiggleTools: Parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics*, 30(7):1008–1009.

Zhang, L., Lee, M., Maslov, A. Y., et al. (2023). Analyzing somatic mutations by single-cell whole-genome sequencing. *Nature Protocols*, 19(2):487–516.

Zhao, J., Bacolla, A., Wang, G., and Vasquez, K. M. (2010). Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences*, 67(1):43–62.

Zhong, J., Luo, K., Winter, P. S., et al. (2016). Mapping nucleosome positions using DNase-seq. *Genome Research*, 26(3):351–364.

Zhu, M., Lu, T., Jia, Y., et al. (2019). Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. *Cell*, 177(3):608–621.e12.

Appendix

Supplementary Figures

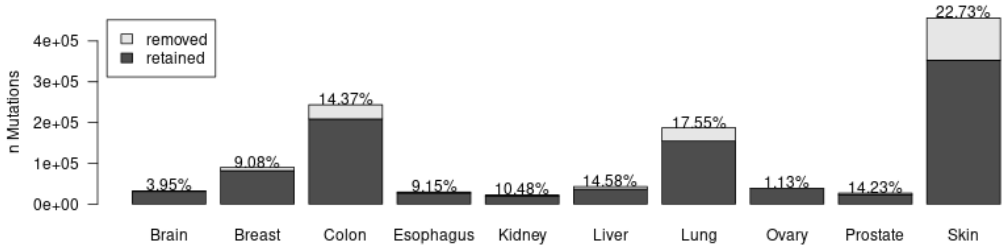

Figure S1: Proportion of positions that were removed due to recurrent mutations. We removed all mutations that were found in more than one sample, in order to limit the impact of positively selected mutations. The number of recurrent mutations is proportional with available data size. Thus, the proportion of recurrent mutations ranged between 1.13 and 22.73 percent.

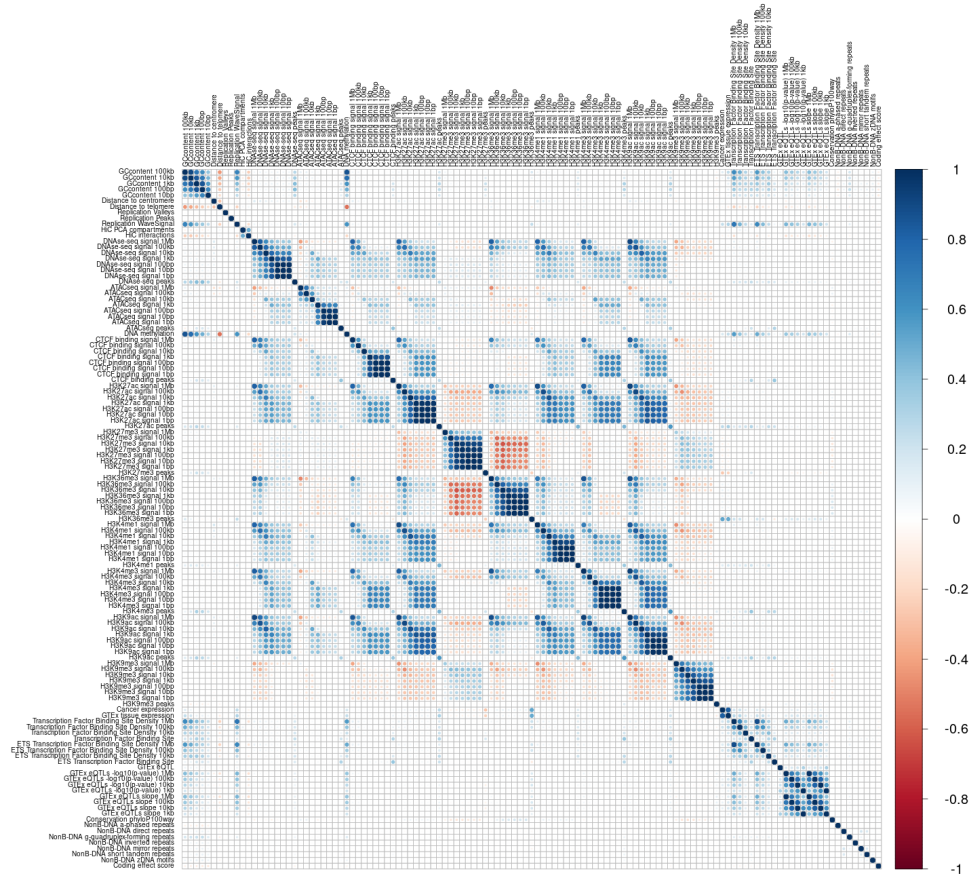

Figure S2: Correlation between predictor variables of brain training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

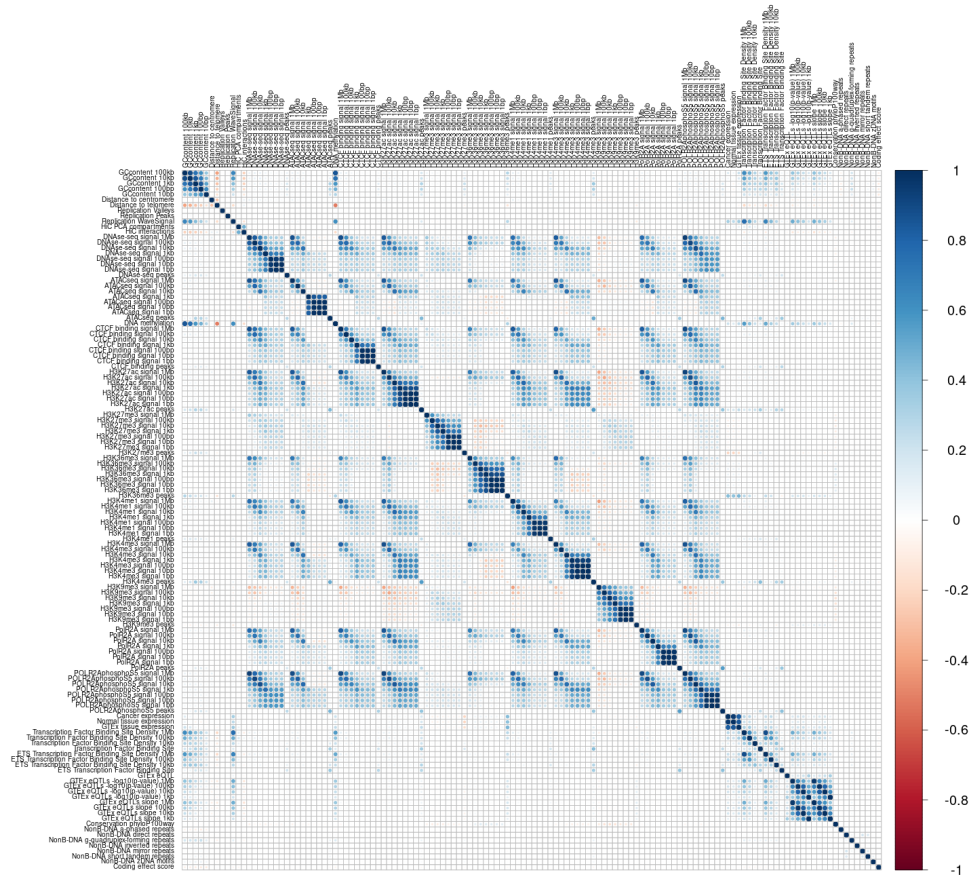

Figure S3: Correlation between predictor variables of breast training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

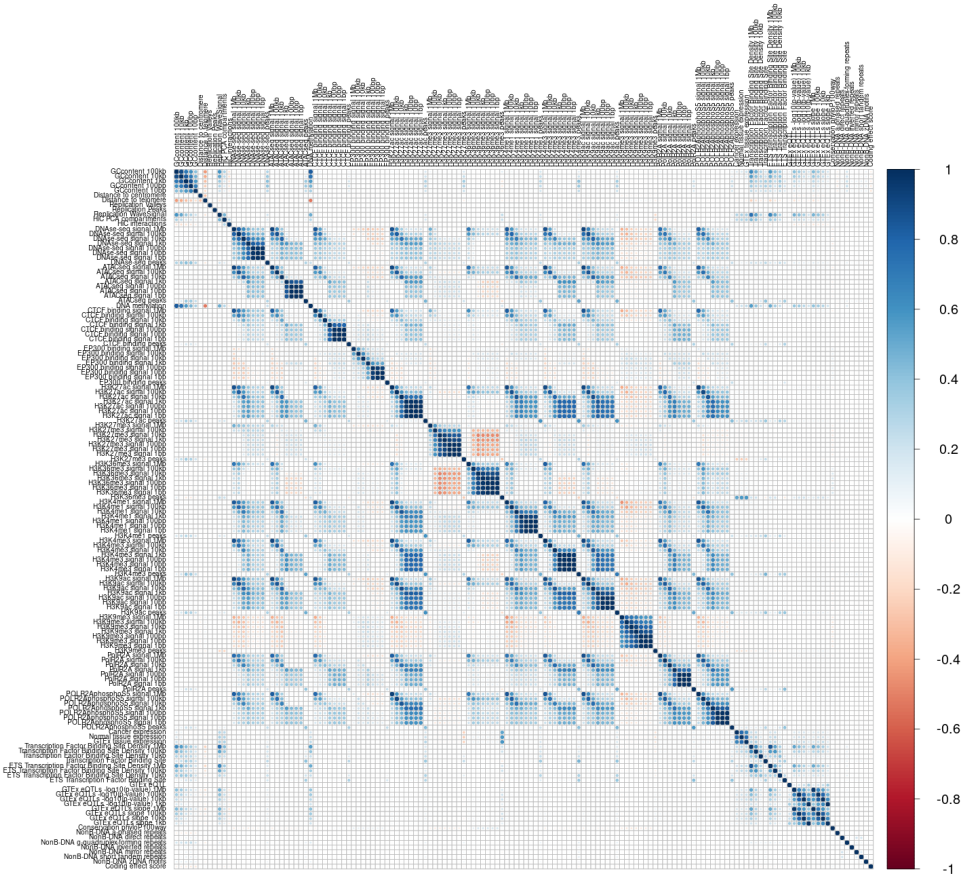

Figure S4: Correlation between predictor variables of colon training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

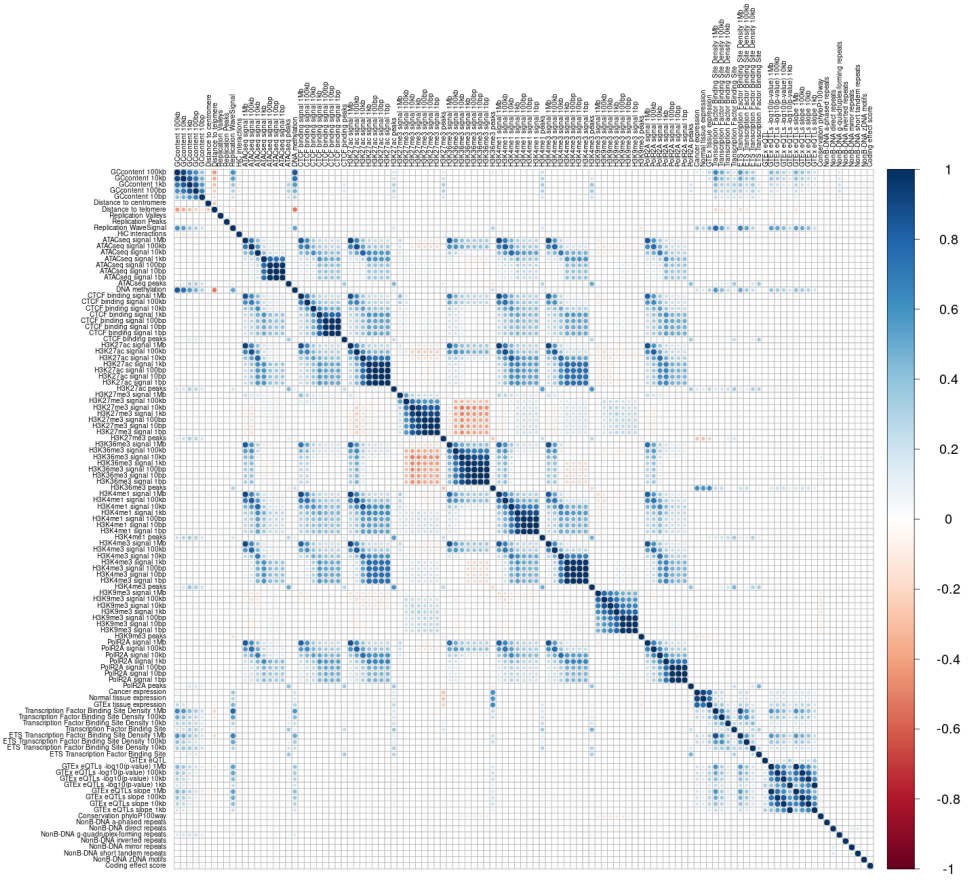

Figure S5: Correlation between predictor variables of esophagus training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

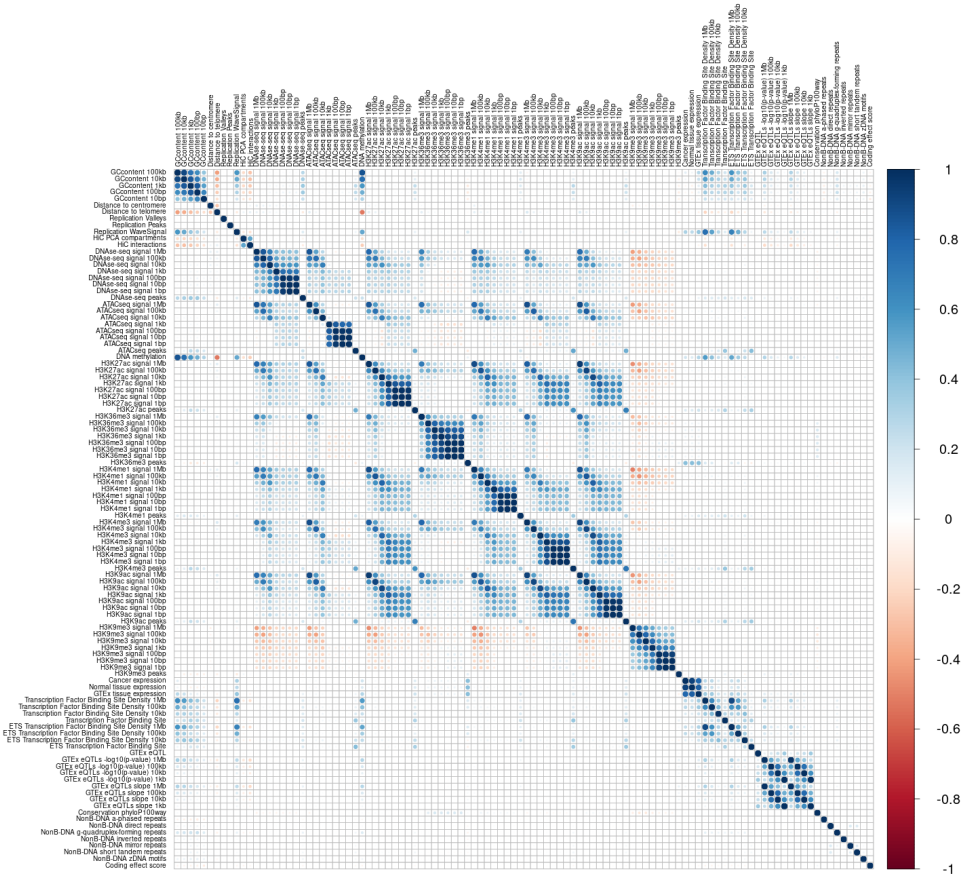

Figure S6: Correlation between predictor variables of kidney training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

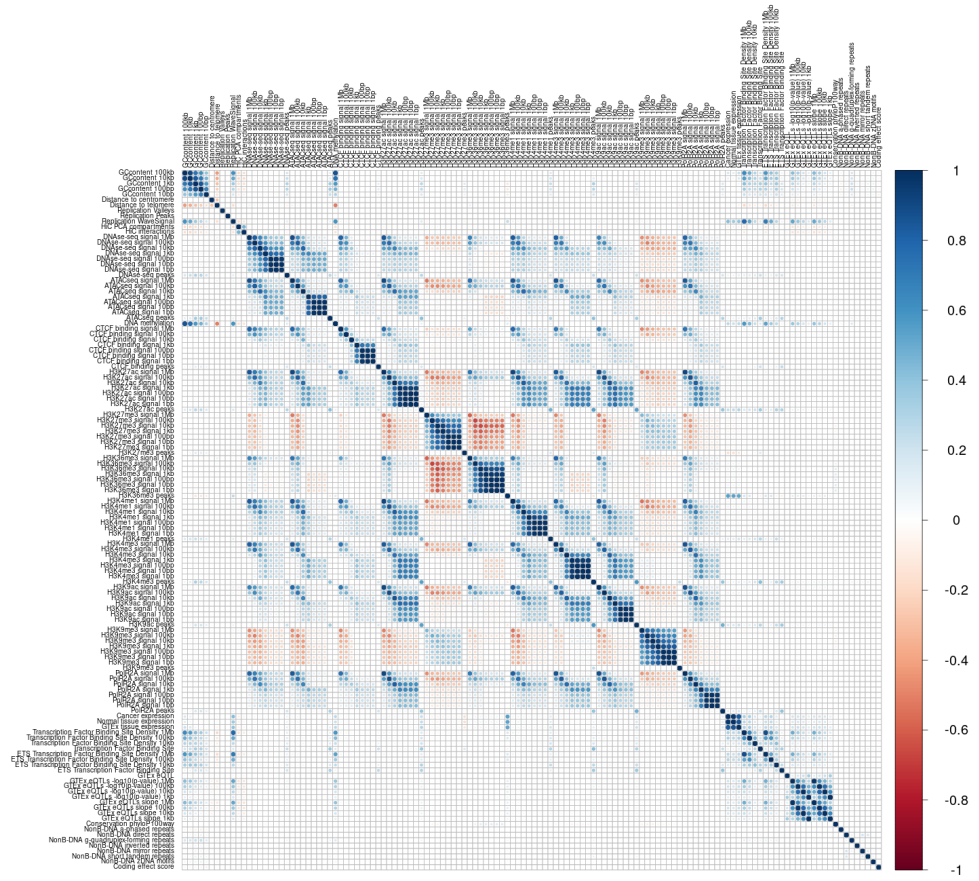

Figure S7: Correlation between predictor variables of liver training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

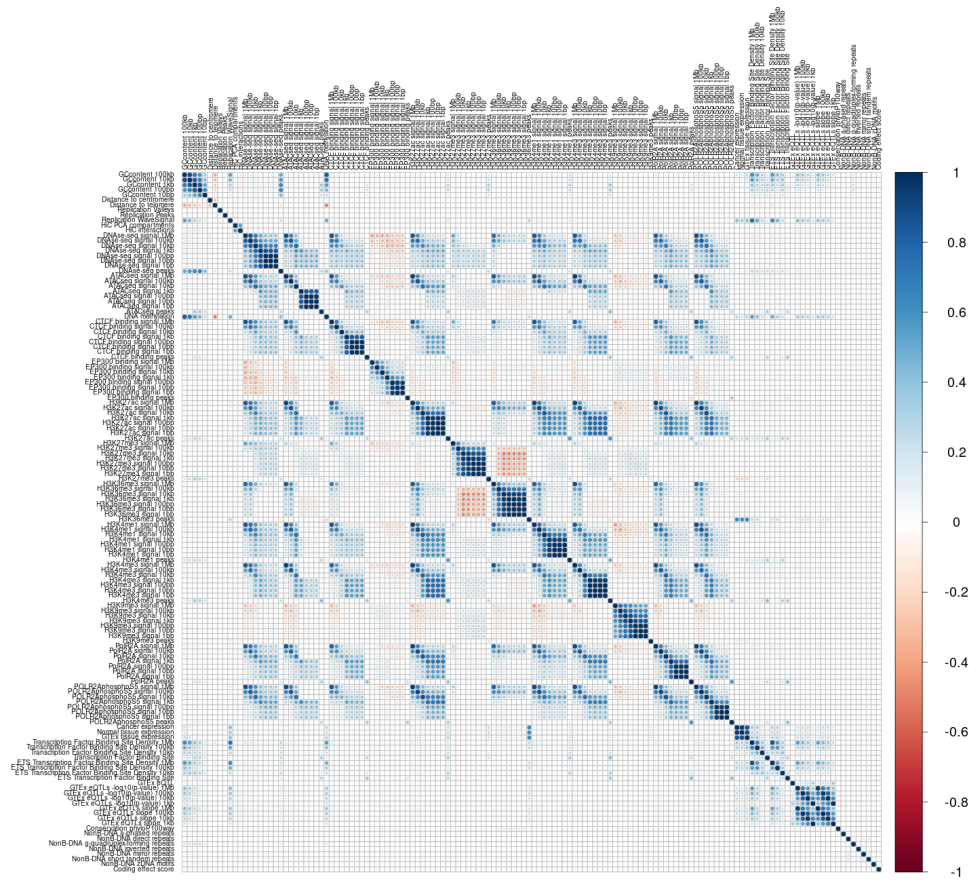

Figure S8: Correlation between predictor variables of lung training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

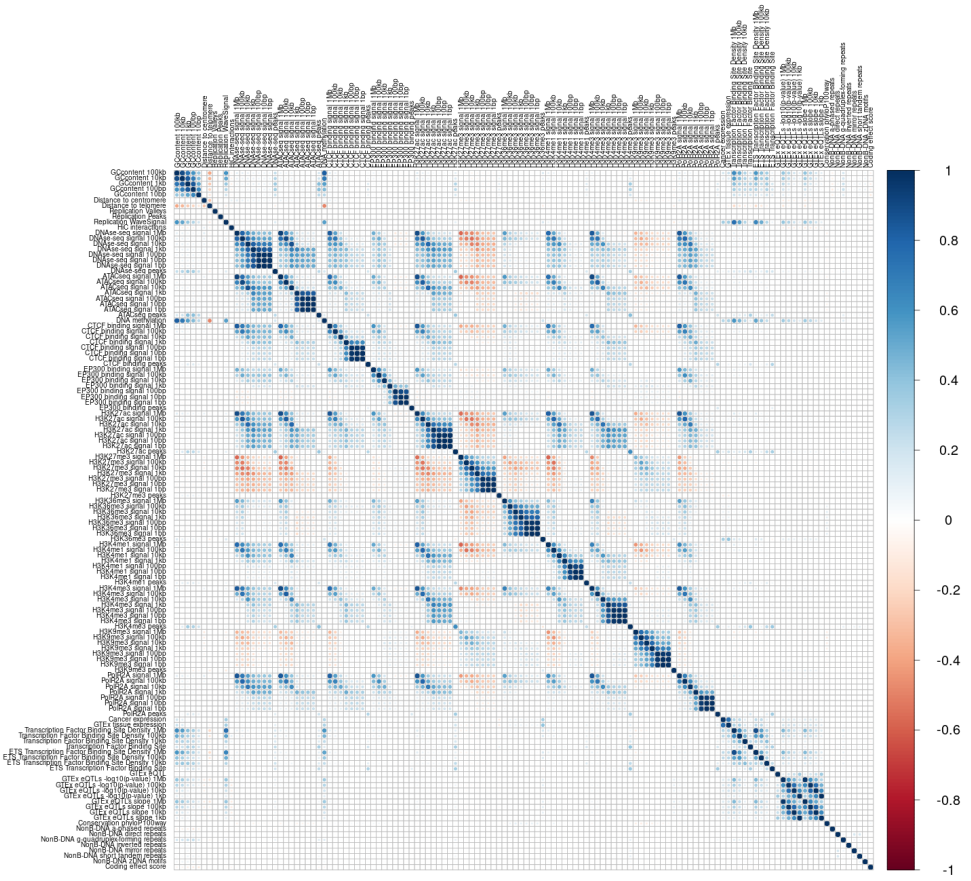

Figure S9: Correlation between predictor variables of ovary training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

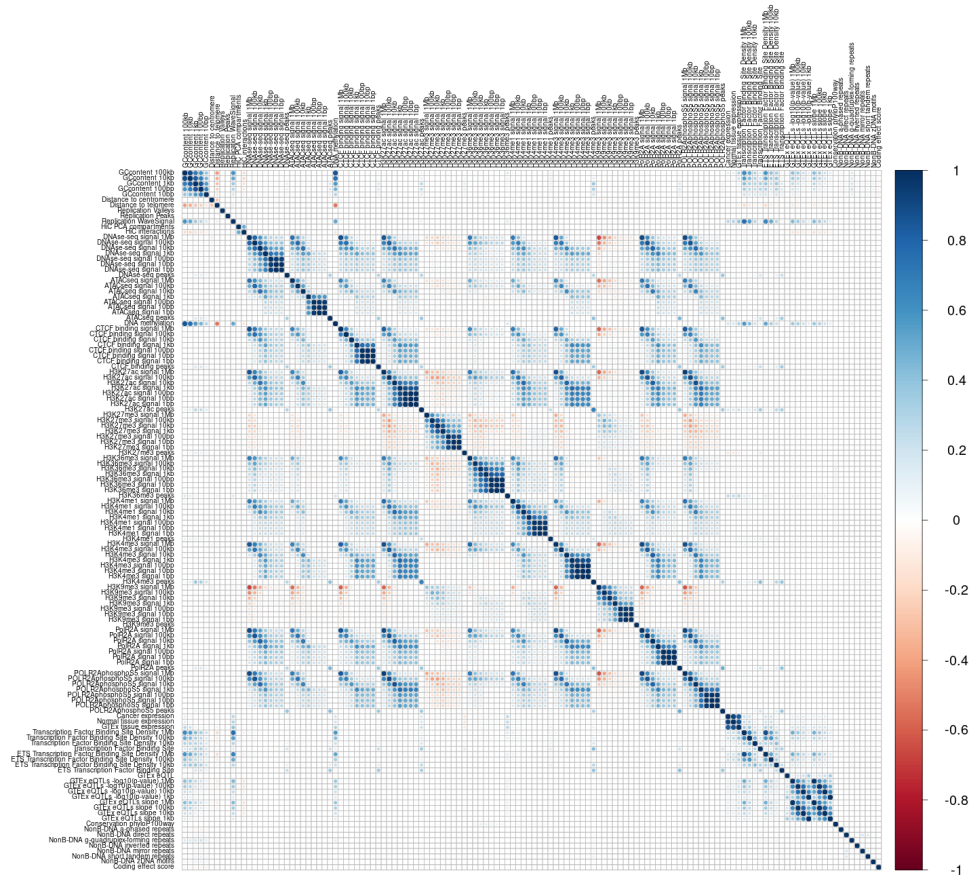

Figure S10: Correlation between predictor variables of prostate training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

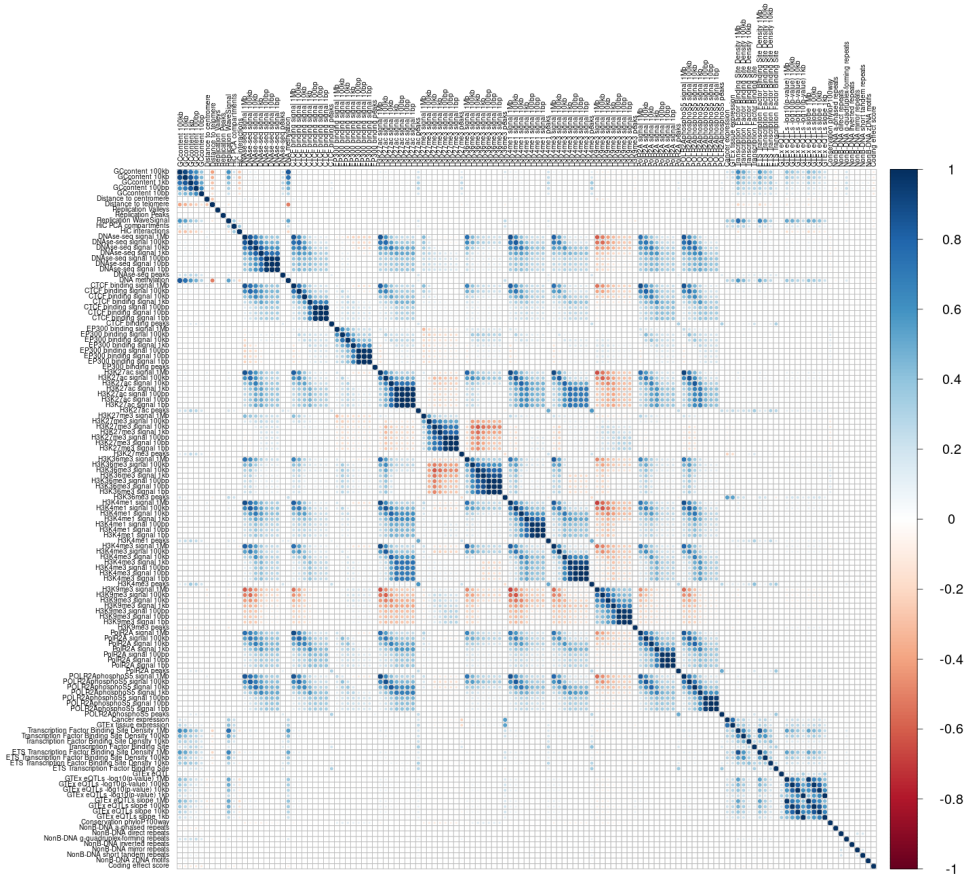

Figure S11: Correlation between predictor variables of skin training data. Pearson correlation coefficient between each predictor pair is represented by dot size and color.

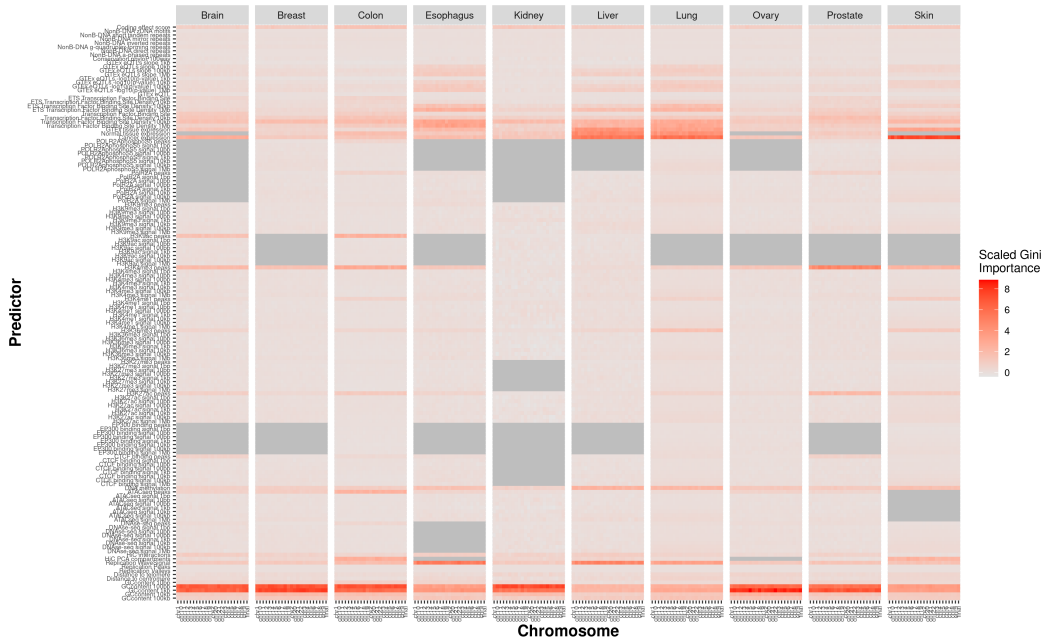

Figure S12: Comparison of chromosome-wise cross-validation predictor importance values. Heatmap visualization of RF gini importances measures across CWCV iterations. In each of the tissue panels, each row represents one predictor and each column one of the 22 CWCV iteration. The last column always represents the final model that includes all chromosomes. Importance measures where scaled between tissues for better comparison. Grey areas indicate predictors that were not available for a tissue.

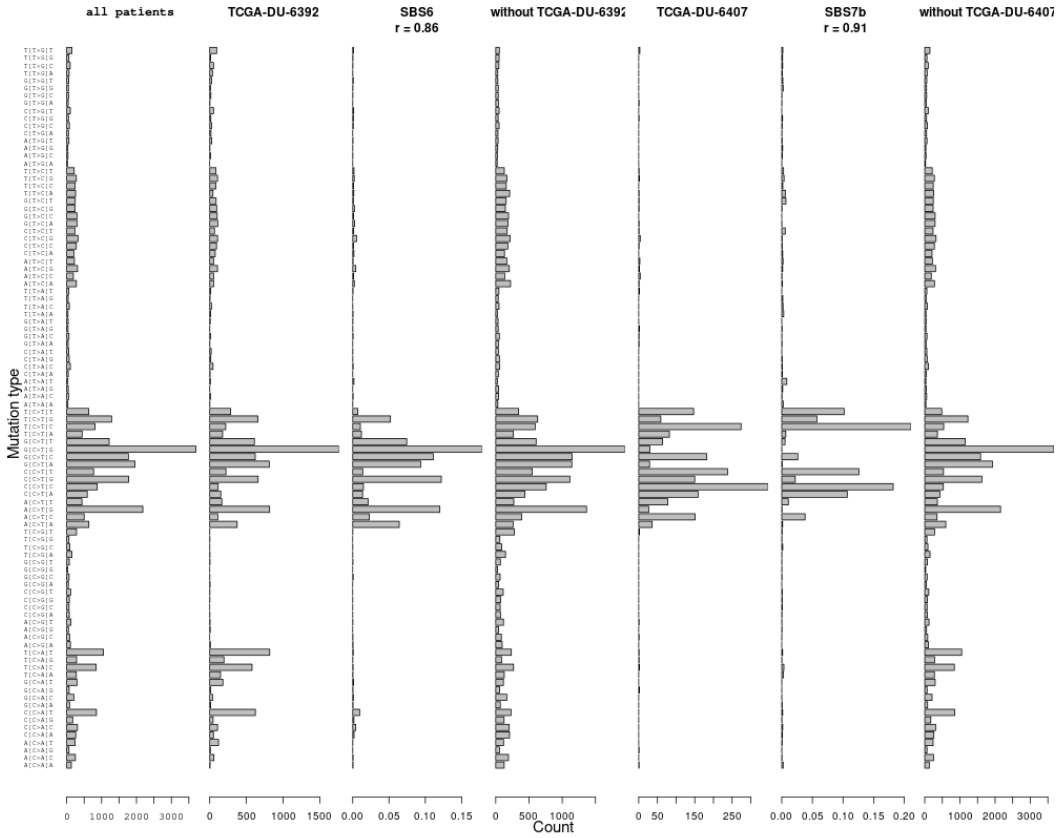

Figure S13: Mutation type distribution of hypermutators in brain. This figure depicts the effect of hypermutators (making up at least 5% of all mutations in this tissue) on the mutation type distribution. The mutation type distribution is defined as the frequency of the six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) which were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Depicted are, from left to right: the distribution of the entire tissue data, and then, for the two hypermutated samples TCGA-DU-6392 and TCGA-DU-6407, the distribution of only the samples, the COSMIC signature that most correlated with the mutation spectrum of this samples (with Pearson's R), and the distribution of the brain data after removing each samples, respectively.

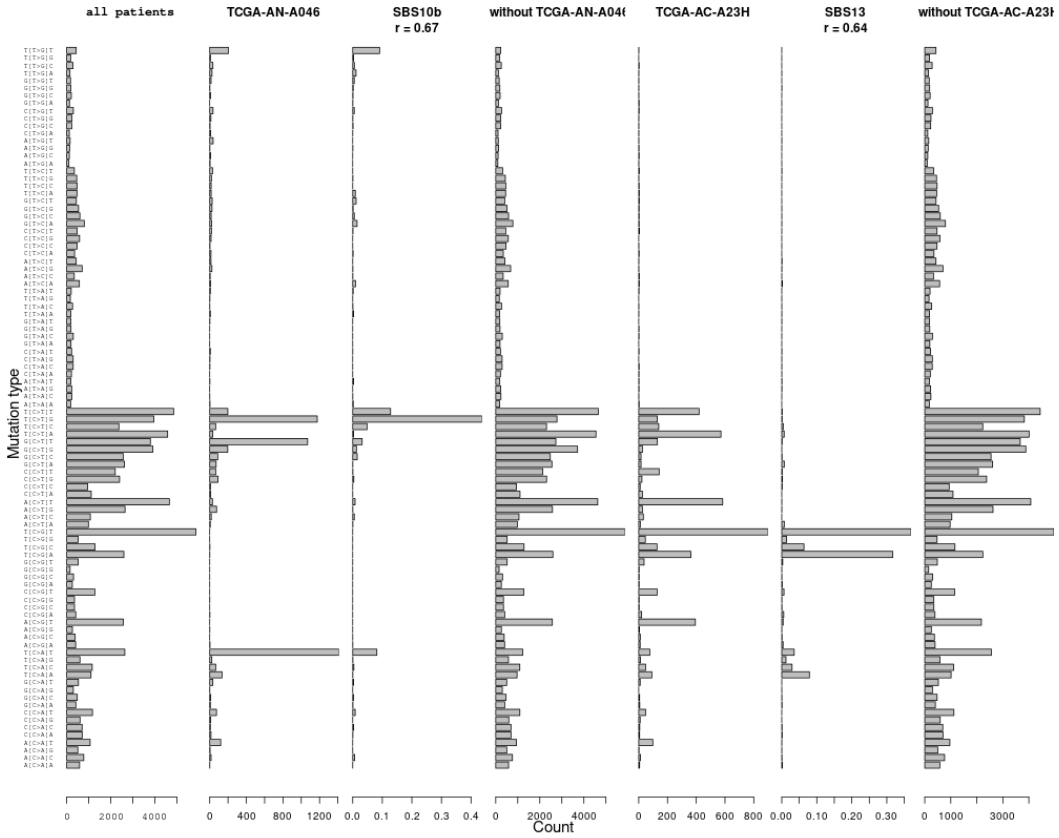

Figure S14: Mutation type distribution of hypermutators in breast. This figure depicts the effect of hypermutators (making up at least 5% of all mutations in this tissue) on the mutation type distribution. The mutation type distribution is defined as the frequency of the six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) which were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Depicted are, from left to right: the distribution of the entire tissue data, and then, for the two hypermutated samples TCGA-AN-A064 and TCGA-AC-A23H, the distribution of only the samples, the COSMIC signature that most correlated with the mutation spectrum of this samples (with Pearson's R), and the distribution of the breast data after removing each samples, respectively.

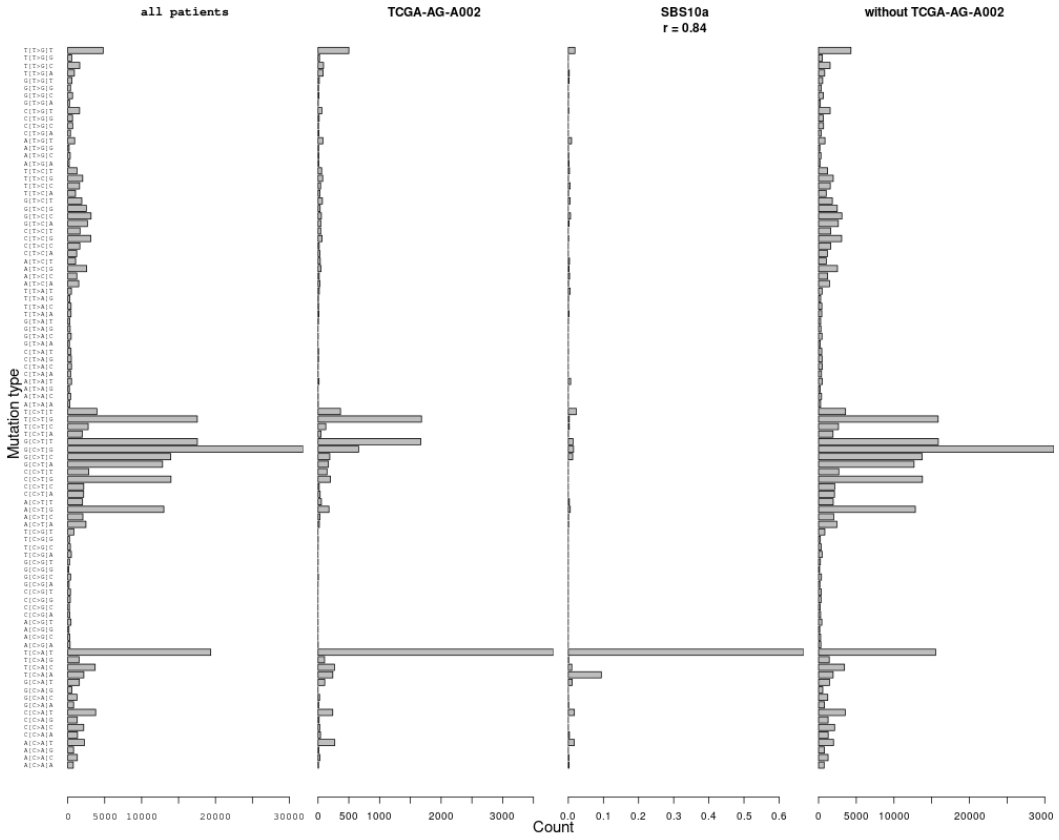

Figure S15: Mutation type distribution of hypermutators in colon. This figure depicts the effect of hypermutators (making up at least 5% of all mutations in this tissue) on the mutation type distribution. The mutation type distribution is defined as the frequency of the six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) which were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Depicted are, from left to right: the distribution of the entire tissue data, the distribution of only the hypermutated sample TCGA-AG-A002, the COSMIC signature that most correlated with the mutation spectrum of this sample (with Pearson's R), and the distribution of the colon data after removing this sample.

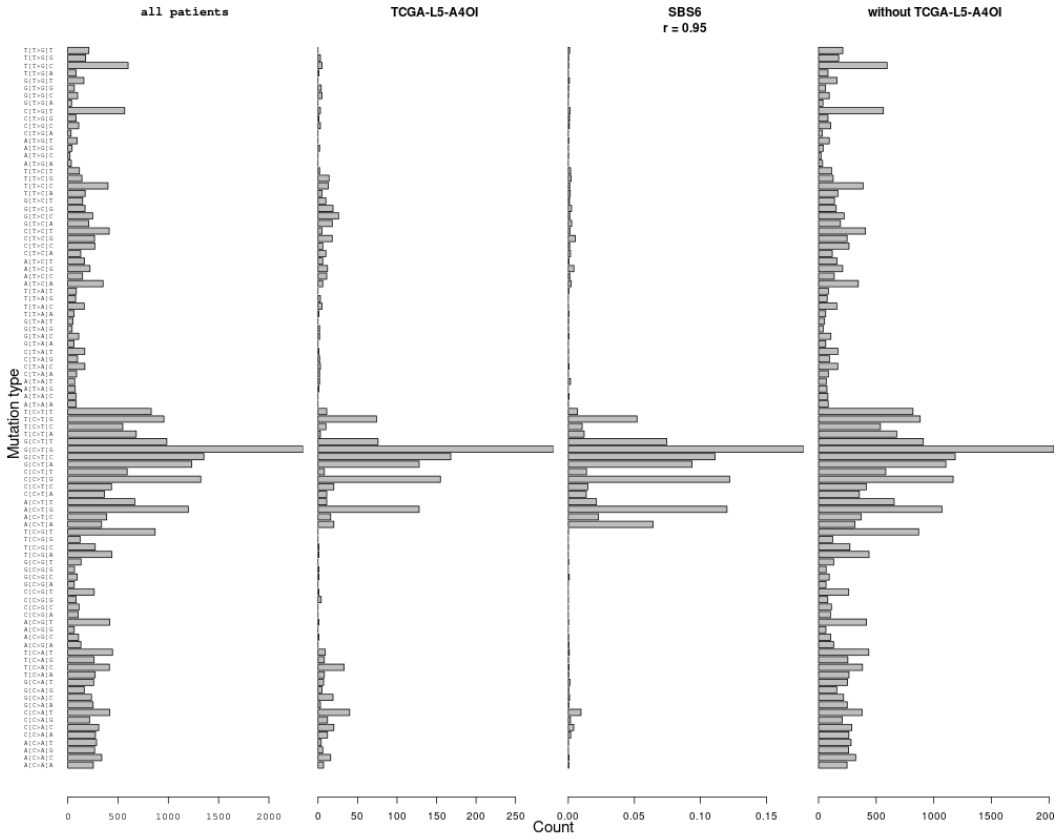

Figure S16: Mutation type distribution of hypermutators in esophagus. This figure depicts the effect of hypermutators (making up at least 5% of all mutations in this tissue) on the mutation type distribution. The mutation type distribution is defined as the frequency of the six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) which were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Depicted are, from left to right: the distribution of the entire tissue data, the distribution of only the hypermutated sample TCGA-L5-A4OI, the COSMIC signature that most correlated with the mutation spectrum of this sample (with Pearson's R), and the distribution of the esophagus data after removing this sample.

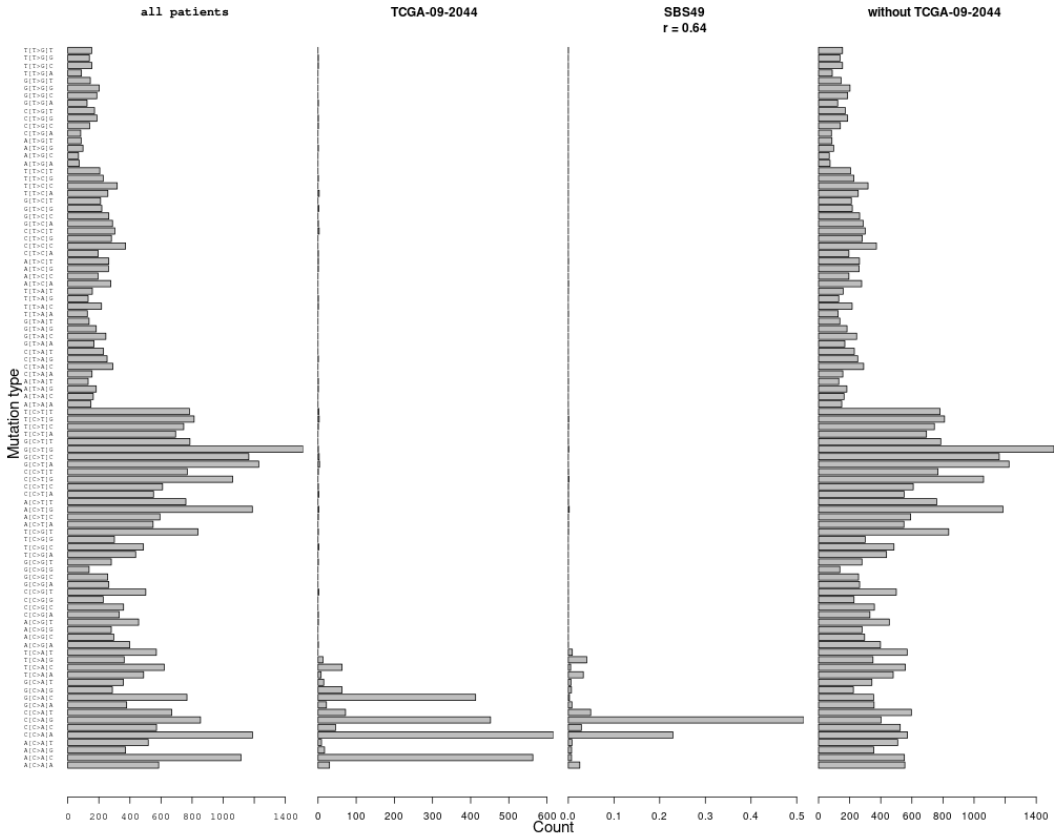

Figure S17: Mutation type distribution of hypermutators in ovary. This figure depicts the effect of hypermutators (making up at least 5% of all mutations in this tissue) on the mutation type distribution. The mutation type distribution is defined as the frequency of the six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) which were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Depicted are, from left to right: the distribution of the entire tissue data, the distribution of only the hypermutated sample TCGA-09-2044, the COSMIC signature that most correlated with the mutation spectrum of this sample (with Pearson's R), and the distribution of the ovary data after removing this sample.

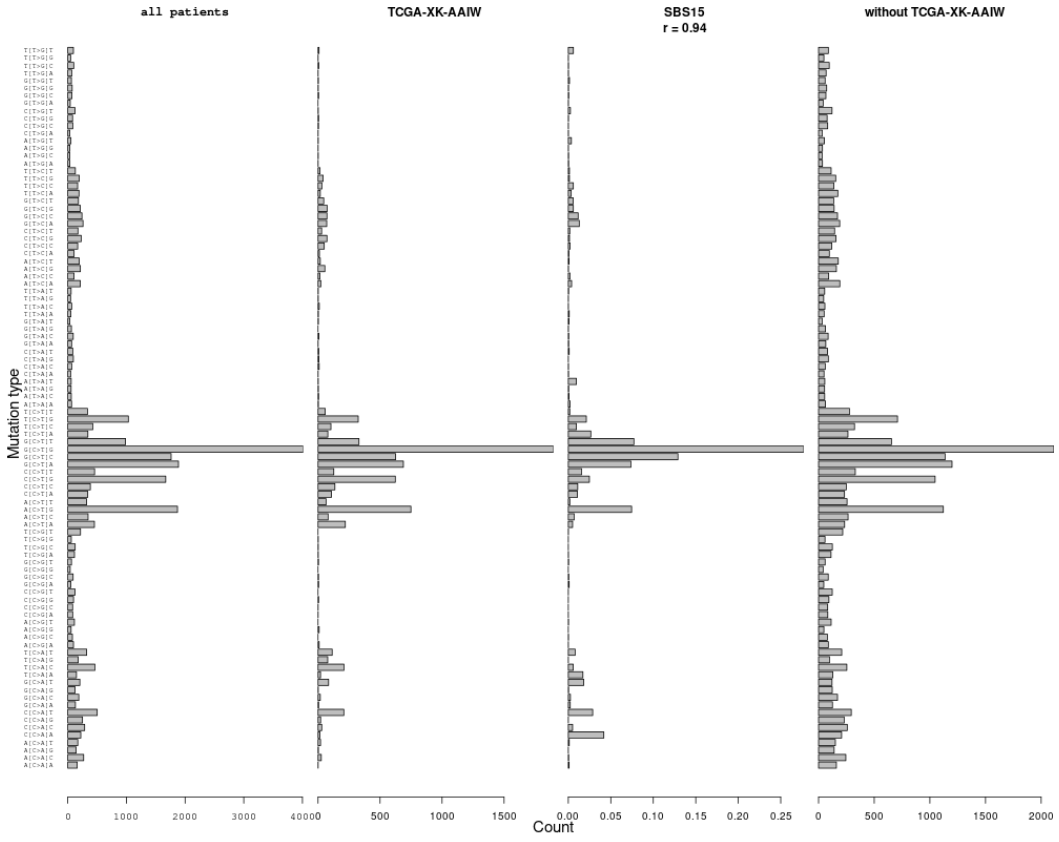

Figure S18: Mutation type distribution of hypermutators in prostate. This figure depicts the effect of hypermutators (making up at least 5% of all mutations in this tissue) on the mutation type distribution. The mutation type distribution is defined as the frequency of the six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) which were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Depicted are, from left to right: the distribution of the entire tissue data, the distribution of only the hypermutated sample TCGA-XK-AAIW, the COSMIC signature that most correlated with the mutation spectrum of this sample (with Pearson's R), and the distribution of the prostate data after removing this sample.

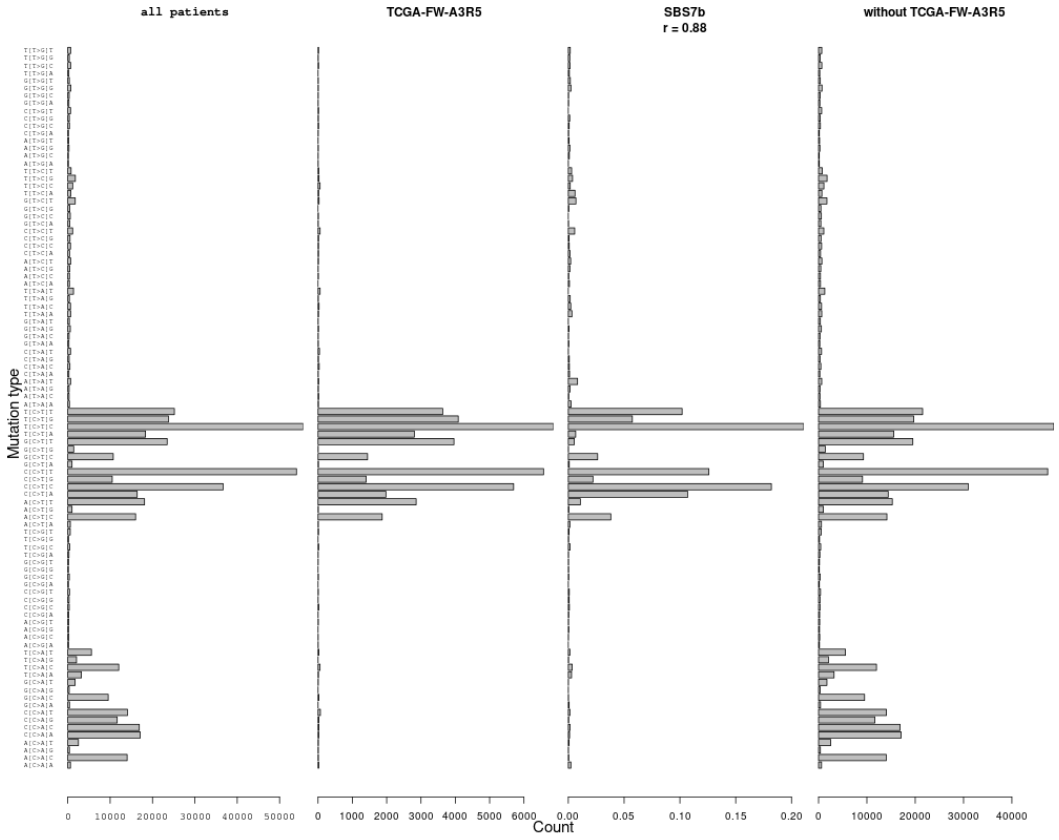

Figure S19: Mutation type distribution of hypermutators in skin. This figure depicts the effect of hypermutators (making up at least 5% of all mutations in this tissue) on the mutation type distribution. The mutation type distribution is defined as the frequency of the six possible classes of nucleotide substitution (C>A, C>G, C>T, T>A, T>C, and T>G) which were further stratified by trimer sequence context (i.e., the immediate 5' and 3' nucleotide neighbors). Depicted are, from left to right: the distribution of the entire tissue data, the distribution of only the hypermutated sample TCGA-FW-A3R5, the COSMIC signature that most correlated with the mutation spectrum of this sample (with Pearson's R), and the distribution of the skin data after removing this sample.

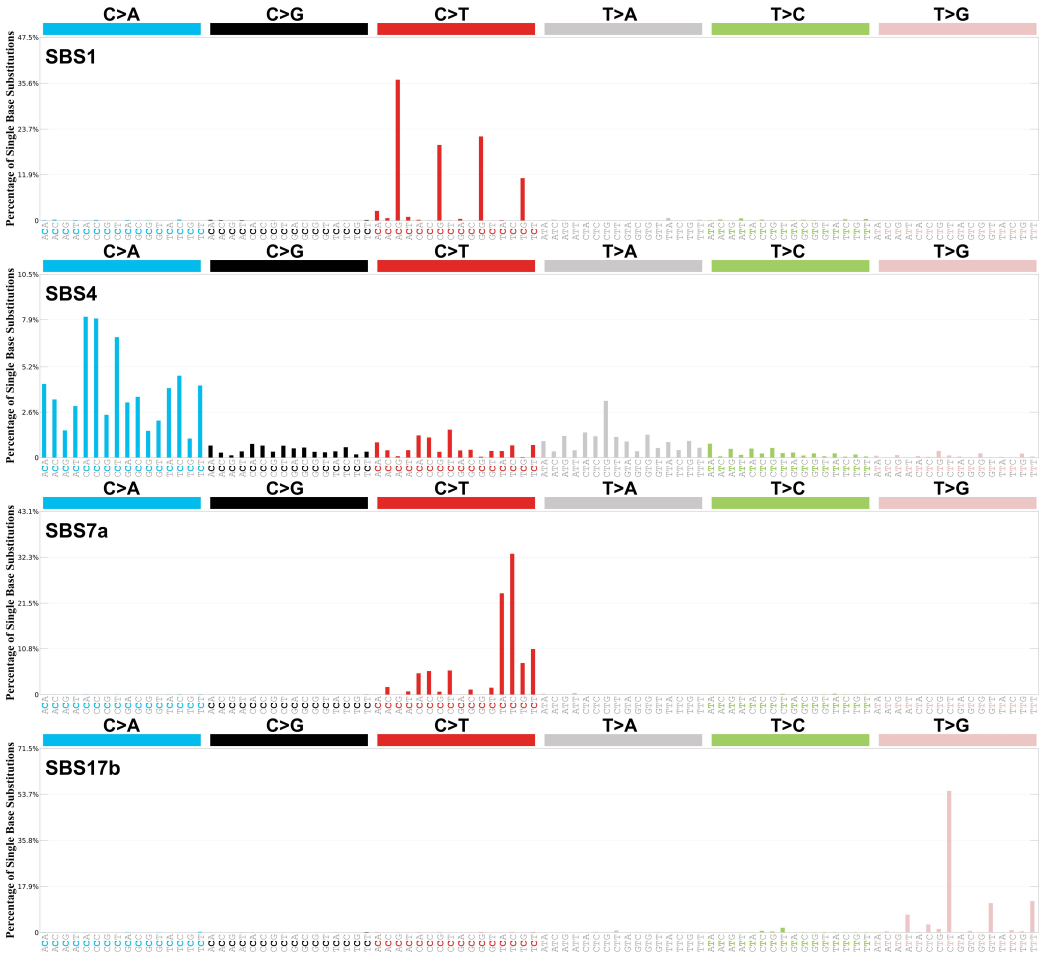

Figure S20: COSMIC mutational signatures relevant to this study. Depicted are the mutational profiles based on the six substitution subtypes (C>A, C>G, C>T, T>A, T>C, and T>G) combined with the nucleotides immediately 5' and 3' to the mutation. Signature SBS1 is often termed a clock-like signature, since it occurs in all tissue types and continuously increases with age. SBS4 is associated with exposure to tobacco smoke. SBS7a (and ist almost-twin SBS7b) can be attributed to UV light. The signature SBS17b is a signature of unknown origin, which seems to play a role in esophageal cancer, among others. Signatures SBS40a-c are of unknown aetiology and are typically found in renal cancers. Figures adapted from the COSMIC website (https://cancer.sanger.ac.uk/signatures/sbs/, accessed Dec 03, 2024; Alexandrov et al., 2020).

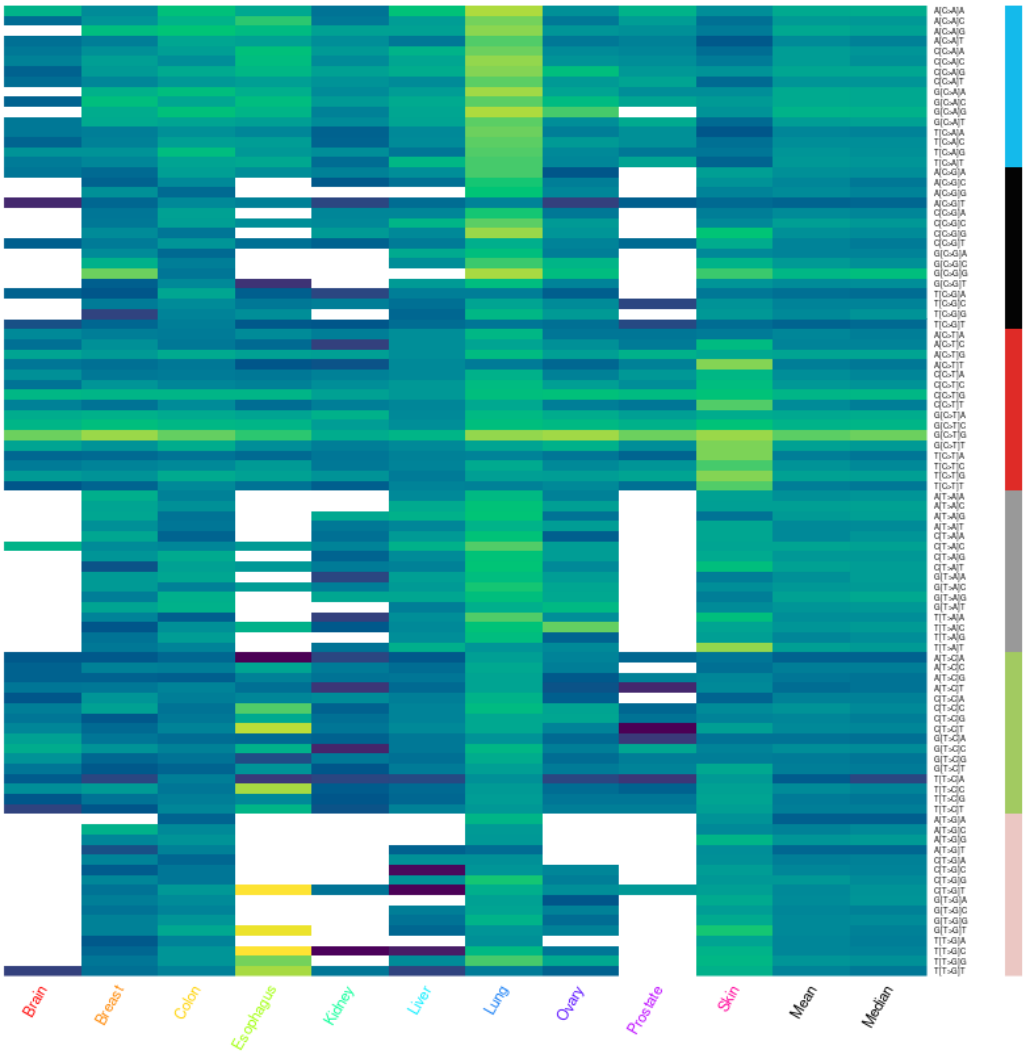

Figure S21: Dependence of model performance on mutation context. Each tissue-specific model was testen on the subset of positions corresponding to a certain mutation type (i.e., substitution type combined with the sequence context comprised of the 3' and 5' neighbors). Each row corresponds to one such mutation type, each column to one tissue. The last two columns indicate the row-wise mean and median, respectively, for easier interpretation. White fields indicate tissue-mutation type pairs with insufficient data to compute stable AUC (i.e., fewer than 100 data points).

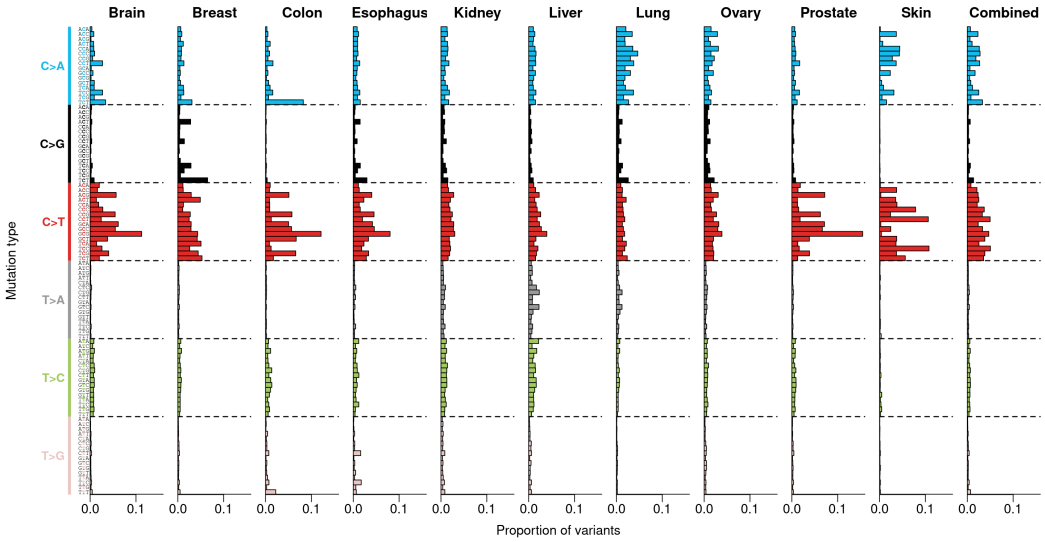

Figure S22: Mutation spectrum of combined tissue data. Depicted are the mutation spectra of each individual tissue as well as the combination of all tissue-specific datasets. The mutational profiles count the proportions of the six substitution subtypes (C>A, C>G, C>T, T>A, T>C, and T>G) combined with the nucleotides immediately 5' and 3' to the mutation.Figure design adapted from (Lawrence et al., 2013b).

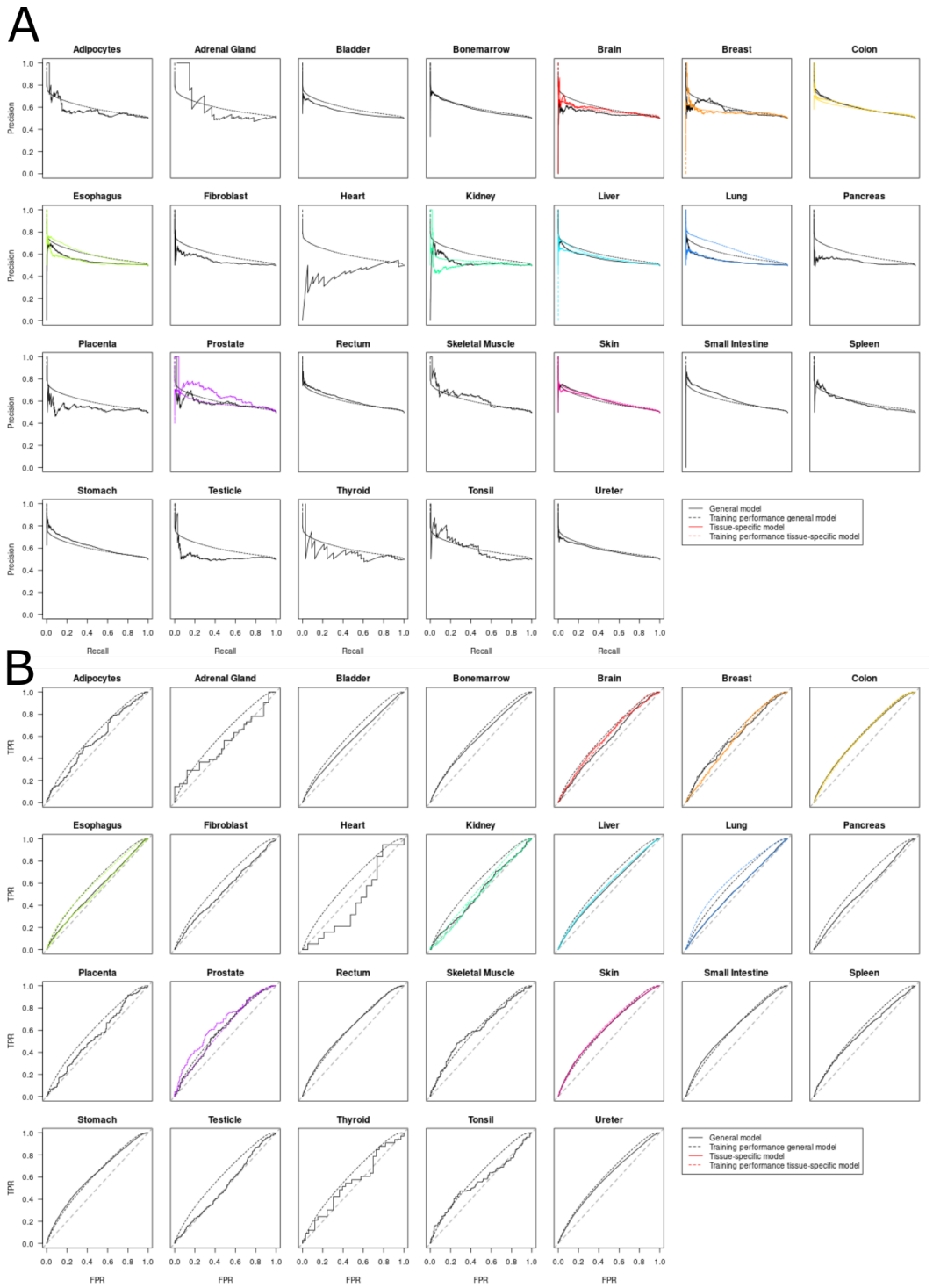

Figure S23: Model performance on healthy tissues. (**A**) Precision recall curves for each tissue of the all-tissue general model applied to the healthy tissue dataset, applied to the all-tissue training data, as well as (where applicable) the performance of the tissue-specific models on healthy and training data, respectively. (**B**) Same as (A), but showing the ROC curves. The dashed grey line indicates the diagonal.

Supplementary Tables

Table S1: Histone Chip-seq ENCODE metafile URLs. Download links to the ENCODE metafile tables documenting the histone modification ChIP-Seq experiments used in this study.

Tissue	Metafile URL
Brain	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&biosample_ontology.organ_slims=brain&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&type=Experiment&files.analyses.status=released&files.preferred_default=true
Breast	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&biosample_ontology.organ_slims=breast&type=Experiment&files.analyses.status=released&files.preferred_default=true
Colon	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&biosample_ontology.organ_slims=colon&biosample_ontology.term_name%21=muscle+layer+of+colon&type=Experiment&files.analyses.status=released&files.preferred_default=true
Esophagus	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&biosample_ontology.organ_slims=esophagus&biosample_ontology.term_name%21=esophagus+muscularis+mucosa&type=Experiment&files.analyses.status=released&files.preferred_default=true
Kidney	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&biosample_ontology.organ_slims=kidney&type=Experiment&files.analyses.status=released&files.preferred_default=true
Liver	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&biosample_ontology.organ_slims=liver&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S1: (continued)

Lung	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&biosample_ontology.organ_slims=lung&type=Experiment&files.analyses.status=released&files.preferred_default=true
Ovary	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&biosample_ontology.organ_slims=ovary&type=Experiment&files.analyses.status=released&files.preferred_default=true
Prostate	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&biosample_ontology.organ_slims=prostate+gland&type=Experiment&files.analyses.status=released&files.preferred_default=true
Skin	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assay_title=Histone+ChIP-seq&assay_title=Mint-ChIP-seq&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&biosample_ontology.organ_slims=skin+of+body&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S2: Histone Chip-seq experiments. List of ENCODE IDs corresponding to histone modification ChIP-Seq experiments used in this study. For each tissue and each ChIP-Seq target, we used bed files encompassing binding peaks as well as bigWig files encompassing signal p-value along the genome.

Tissue	Target	File type	ENCODE IDs
Brain	H3K27ac	bigWig	ENCFF496HWP, ENCFF818EXD, ENCFF403XAR, ENCFF578UND, ENCFF272ZKG, ENCFF999QTF, ENCFF834VRT, ENCFF654CFQ, ENCFF140CPL, ENCFF894IET, ENCFF965WPK, ENCFF726XBL, ENCFF180TKJ, ENCFF720KFF, ENCFF481OWU, ENCFF219VUF, ENCFF737XXA, ENCFF298WAS, ENCFF675ZSU, ENCFF866IWY, ENCFF794KRV, ENCFF540ROF, ENCFF853GMX, ENCFF456KFU, ENCFF560FDE, ENCFF848BKA, ENCFF684DXK, ENCFF924EWG, ENCFF242SBC, ENCFF497IVL, ENCFF706QDU, ENCFF561PQW, ENCFF702SZB, ENCFF429THJ
Brain	H3K27ac	bed	ENCFF236DDT, ENCFF150MNN, ENCFF551LSF, ENCFF066PAQ, ENCFF684COM, ENCFF444CZR, ENCFF461TWG, ENCFF311BUM, ENCFF712CHB, ENCFF955SXY, ENCFF721ZGP, ENCFF072GKG, ENCFF681YKY, ENCFF933DUA, ENCFF508PQT, ENCFF229NLK, ENCFF610ZPW, ENCFF272ROP, ENCFF761QUK, ENCFF703WCD, ENCFF937WTH, ENCFF581YQU, ENCFF105SAZ, ENCFF650YIZ, ENCFF169QSG, ENCFF288NLR, ENCFF089YVE, ENCFF216XRU, ENCFF337DEA, ENCFF860MVH, ENCFF100WAF, ENCFF595CLG, ENCFF289MDK, ENCFF676LJR
Brain	H3K36me3	bigWig	ENCFF088LVM, ENCFF254GZF, ENCFF821UFH, ENCFF783TGQ, ENCFF612KXC, ENCFF842CAN, ENCFF408NLE, ENCFF761KND, ENCFF791FNY, ENCFF072RDG, ENCFF116GSZ, ENCFF445KNC, ENCFF723FJT, ENCFF065APW, ENCFF149DDW
Brain	H3K36me3	bed	ENCFF518APA, ENCFF609DHQ, ENCFF922IYA, ENCFF912TFT, ENCFF645ALW, ENCFF442BXP, ENCFF526TTE, ENCFF613UPN, ENCFF727KZF, ENCFF084VLV, ENCFF945TJD, ENCFF236EMJ, ENCFF301UFB, ENCFF350XHO, ENCFF916HPS
Brain	H3K9me3	bigWig	ENCFF946RTU, ENCFF352GLF, ENCFF407KLN, ENCFF265ZLK, ENCFF348PTZ, ENCFF823QLT, ENCFF186SQL, ENCFF166DOF, ENCFF248NGW, ENCFF784SSN, ENCFF280KWF, ENCFF655AQX, ENCFF393BIE, ENCFF158MOD, ENCFF223IHB
Brain	H3K9me3	bed	ENCFF679LQD, ENCFF827KSG, ENCFF265CFH, ENCFF046SQR, ENCFF398NVA, ENCFF886QLM, ENCFF025HQM, ENCFF249HFV, ENCFF848UST, ENCFF610ZKG, ENCFF058RNM, ENCFF963WLO, ENCFF835ZYG, ENCFF477FPU, ENCFF966IRX
Brain	H3K27me3	bigWig	ENCFF598ICL, ENCFF185YGA, ENCFF973LTF, ENCFF784DSK, ENCFF109DUV, ENCFF094QHD, ENCFF765QGA, ENCFF887SYJ, ENCFF162GEQ, ENCFF328XVK, ENCFF859SEG, ENCFF902TZJ, ENCFF167ASN, ENCFF145VYV, ENCFF241CCA, ENCFF644DJO, ENCFF987VYP, ENCFF055LUU, ENCFF329QOB, ENCFF710WDC, ENCFF653WGR, ENCFF832VNW, ENCFF202HUV, ENCFF587DBY, ENCFF552IAL, ENCFF205RPD, ENCFF018IDJ, ENCFF796TKW, ENCFF588XAS, ENCFF942JQM, ENCFF016NPV, ENCFF689GBH
Brain	H3K27me3	bed	ENCFF998LIO, ENCFF017GGR, ENCFF878XDQ, ENCFF810LJW, ENCFF718SCV, ENCFF974SWY, ENCFF556XGB, ENCFF570UGK, ENCFF591UDY, ENCFF748JKZ, ENCFF215KEC, ENCFF369ASW, ENCFF518FDA, ENCFF098DXX, ENCFF321ZZK, ENCFF903CDW, ENCFF877AFT, ENCFF710FUW, ENCFF824JRL, ENCFF149ZZG, ENCFF756BDJ, ENCFF079QBO, ENCFF026LAQ, ENCFF653YAM, ENCFF864YZX, ENCFF259PMU, ENCFF745PTT, ENCFF797WQN, ENCFF729EZH, ENCFF008NWY, ENCFF510CIP, ENCFF042NGO

Table S2: (continued)

Tissue	Target	File type	ENCODE IDs
Brain	H3K4me3	bigWig	ENCFF465HLY, ENCFF449TQG, ENCFF214WDH, ENCFF730WMQ, ENCFF499KGL, ENCFF348TJB, ENCFF825COY, ENCFF557HYU, ENCFF507RVO, ENCFF335RNZ, ENCFF875YZF, ENCFF251SLL, ENCFF041TFY, ENCFF026MPK, ENCFF352KXC, ENCFF966LNR, ENCFF123QKH, ENCFF941KDP, ENCFF157FAL, ENCFF475UPQ, ENCFF204UIM, ENCFF125XSS, ENCFF752EVS, ENCFF462MRQ, ENCFF451OPA, ENCFF583MTT, ENCFF250PDX, ENCFF294RCE, ENCFF414APF, ENCFF522WRJ, ENCFF758OPC, ENCFF757OCX, ENCFF474LRE, ENCFF462JDJ, ENCFF016RMR, ENCFF680BLP, ENCFF338YMX
Brain	H3K4me3	bed	ENCFF615SLX, ENCFF451IGE, ENCFF285FJZ, ENCFF456DDO, ENCFF471WRX, ENCFF298IQY, ENCFF788EOD, ENCFF184MEF, ENCFF771USJ, ENCFF180SKE, ENCFF491IPL, ENCFF383THB, ENCFF397GYA, ENCFF859JHN, ENCFF525EJT, ENCFF453HCX, ENCFF686AFC, ENCFF963UEV, ENCFF384NUC, ENCFF724XKK, ENCFF324FFN, ENCFF629GOS, ENCFF410RJE, ENCFF350LHB, ENCFF196MXR, ENCFF107KDQ, ENCFF216OMD, ENCFF753BHM, ENCFF978BSU, ENCFF009ORT, ENCFF344TGW, ENCFF095HVI, ENCFF180XUG, ENCFF648FWC, ENCFF014FUG, ENCFF695WNQ, ENCFF706DGK
Brain	H3K4me1	bigWig	ENCFF758LAA, ENCFF436LHA, ENCFF326XDV, ENCFF076RXN, ENCFF531RSE, ENCFF578WHD, ENCFF433VSA, ENCFF241REN, ENCFF866WNB, ENCFF906XBM, ENCFF530ZZZ, ENCFF069PRS, ENCFF548TVE, ENCFF251EUG, ENCFF695JPH
Brain	H3K4me1	bed	ENCFF592RHP, ENCFF316XUU, ENCFF530ADZ, ENCFF328YWW, ENCFF598SLA, ENCFF893UEN, ENCFF466ZDI, ENCFF879HER, ENCFF377ASU, ENCFF106NZT, ENCFF355YCG, ENCFF598ZAY, ENCFF651TUA, ENCFF040KUL, ENCFF600AYY
Brain	H3K9ac	bigWig	ENCFF003IQI, ENCFF048JVO, ENCFF290SQY, ENCFF234RPA, ENCFF810WOP, ENCFF089YQZ, ENCFF297TFG
Brain	H3K9ac	bed	ENCFF769GHM, ENCFF541GIP, ENCFF703HPP, ENCFF146VKE, ENCFF175AYH, ENCFF802DMZ, ENCFF507ZNV
Breast	H3K9me3	bigWig	ENCFF607AGI, ENCFF679HFA
Breast	H3K9me3	bed	ENCFF634WCO, ENCFF446HVX
Breast	H3K4me3	bigWig	ENCFF621MFQ, ENCFF449LNR, ENCFF403VNT
Breast	H3K4me3	bed	ENCFF212SAT, ENCFF590DGH, ENCFF562WCY
Breast	H3K4me1	bigWig	ENCFF243OFE, ENCFF901FSR
Breast	H3K4me1	bed	ENCFF698LQG, ENCFF812HAE
Breast	H3K27me3	bigWig	ENCFF692YCJ, ENCFF851ZIJ
Breast	H3K27me3	bed	ENCFF520QGK, ENCFF962YZN
Breast	H3K27ac	bigWig	ENCFF750PLF, ENCFF455HUA
Breast	H3K27ac	bed	ENCFF988ZYW, ENCFF507WEL
Breast	H3K36me3	bigWig	ENCFF416LTT, ENCFF892SXS
Breast	H3K36me3	bed	ENCFF540RWM, ENCFF648VTD
Colon	H3K9me3	bigWig	ENCFF150JBJ, ENCFF946VNR, ENCFF098UTM, ENCFF549XKP, ENCFF523HVU, ENCFF954ZGC, ENCFF407OHR, ENCFF979IGD, ENCFF454ULG, ENCFF661OOO, ENCFF953KYX
Colon	H3K9me3	bed	ENCFF885GZB, ENCFF737ZJC, ENCFF491NVR, ENCFF416DSY, ENCFF732XUJ, ENCFF559NDK, ENCFF411WLB, ENCFF550CSC, ENCFF153MLO, ENCFF866NSS, ENCFF596HEB
Colon	H3K36me3	bigWig	ENCFF356IGI, ENCFF090EKJ, ENCFF949SMS, ENCFF917WXV, ENCFF004TYK, ENCFF361OGU, ENCFF330SQW, ENCFF039QPG, ENCFF393FYI, ENCFF325CMQ, ENCFF926EQN, ENCFF886GOR, ENCFF063ARQ
Colon	H3K36me3	bed	ENCFF186VPP, ENCFF990VOU, ENCFF650OEI, ENCFF913SGA, ENCFF732NMQ, ENCFF406TXJ, ENCFF709PXL, ENCFF406OXE, ENCFF394DNK, ENCFF900JIL, ENCFF382XMT, ENCFF129VAT, ENCFF893YTR
Colon	H3K4me1	bigWig	ENCFF058DKZ, ENCFF997VIZ, ENCFF139KGX, ENCFF156IHC, ENCFF226SJT, ENCFF176HCT, ENCFF181JAH, ENCFF457MRI, ENCFF149ODX, ENCFF225IMX, ENCFF719DXC, ENCFF918PSP, ENCFF194YJD

Table S2: (continued)

Tissue	Target	File type	ENCODE IDs
Colon	H3K4me1	bed	ENCFF951XHW, ENCFF693QVL, ENCFF328FCD, ENCFF489QOV, ENCFF556UDU, ENCFF341WYO, ENCFF025JKR, ENCFF006OWZ, ENCFF184JAS, ENCFF187OLQ, ENCFF052XGA, ENCFF489WZM, ENCFF081EIZ
Colon	H3K27ac	bigWig	ENCFF524LKF, ENCFF402LLT, ENCFF992YVV, ENCFF058WBO, ENCFF928WZU, ENCFF718FVY, ENCFF133FOG, ENCFF505LDB, ENCFF502JLH, ENCFF959PXE, ENCFF956OKD, ENCFF476UDF
Colon	H3K27ac	bed	ENCFF368OSK, ENCFF266NWL, ENCFF294GNO, ENCFF759DJM, ENCFF414SKU, ENCFF549AXK, ENCFF341BPG, ENCFF089APD, ENCFF208DJA, ENCFF142COP, ENCFF111SYA, ENCFF014ZMR
Colon	H3K9ac	bigWig	ENCFF665JEQ, ENCFF631SZJ
Colon	H3K9ac	bed	ENCFF486TGX, ENCFF023PFF
Colon	H3K4me3	bigWig	ENCFF219ZIO, ENCFF532SBD, ENCFF059CCG, ENCFF190OLZ, ENCFF582NYY, ENCFF284XHG, ENCFF526GOB, ENCFF401SNQ, ENCFF505MXJ, ENCFF542NYI, ENCFF848BPB, ENCFF364ITB, ENCFF051FST
Colon	H3K4me3	bed	ENCFF799XQH, ENCFF873QDA, ENCFF100ICD, ENCFF236IKZ, ENCFF755OXH, ENCFF409WQJ, ENCFF846YGO, ENCFF854FZP, ENCFF541NVY, ENCFF250ZHO, ENCFF057RCQ, ENCFF530UBK, ENCFF068YGR
Colon	H3K27me3	bigWig	ENCFF904OGD, ENCFF395ZNW, ENCFF754TYX, ENCFF983WQI, ENCFF535OFR, ENCFF584AMT, ENCFF612HRT, ENCFF349THP, ENCFF882AKN, ENCFF853GRR
Colon	H3K27me3	bed	ENCFF228JQE, ENCFF310IST, ENCFF728TNF, ENCFF644XAC, ENCFF912DUE, ENCFF958KOY, ENCFF411ZRA, ENCFF105MAV, ENCFF524VIK, ENCFF038LOI
Esophagus	H3K27ac	bigWig	ENCFF947UUA, ENCFF457MBW, ENCFF313TCZ, ENCFF424LHS, ENCFF521ARN, ENCFF129EVK
Esophagus	H3K27ac	bed	ENCFF592GQB, ENCFF445XCF, ENCFF057CRD, ENCFF253VVF, ENCFF847EPP, ENCFF673PCP
Esophagus	H3K4me1	bigWig	ENCFF442KVB, ENCFF314RQP, ENCFF495IVY, ENCFF690CTT, ENCFF613WRM
Esophagus	H3K4me1	bed	ENCFF706MJS, ENCFF685YIN, ENCFF977NUM, ENCFF527PUF, ENCFF828YQS
Esophagus	H3K4me3	bigWig	ENCFF658KLF, ENCFF662KUD, ENCFF421MWY, ENCFF378UIM, ENCFF026HSW, ENCFF433VPO
Esophagus	H3K4me3	bed	ENCFF714MLW, ENCFF933THO, ENCFF442NKL, ENCFF313UUQ, ENCFF470ELM, ENCFF651ZRX
Esophagus	H3K9me3	bigWig	ENCFF427KCB, ENCFF481PEG, ENCFF236BZU, ENCFF127VQG, ENCFF757KYK
Esophagus	H3K9me3	bed	ENCFF465IVD, ENCFF829WQM, ENCFF550NKZ, ENCFF380KRG, ENCFF854FHY
Esophagus	H3K36me3	bigWig	ENCFF432ZKH, ENCFF753ZLK, ENCFF147QZZ, ENCFF117FYK, ENCFF280OYS
Esophagus	H3K36me3	bed	ENCFF882FKD, ENCFF389ELB, ENCFF119RXX, ENCFF885BKA, ENCFF576MHH
Esophagus	H3K27me3	bigWig	ENCFF424QFB, ENCFF659WMK, ENCFF508PXP, ENCFF300RFW, ENCFF922KAK, ENCFF218WHB
Esophagus	H3K27me3	bed	ENCFF834PLN, ENCFF146BLW, ENCFF211CXS, ENCFF587TBQ, ENCFF905MKP, ENCFF458EVO
Kidney	H3K9ac	bigWig	ENCFF206NNF, ENCFF237XUE
Kidney	H3K9ac	bed	ENCFF280HID, ENCFF469MZT
Kidney	H3K27ac	bigWig	ENCFF829DEJ, ENCFF839HMN
Kidney	H3K27ac	bed	ENCFF737HGQ, ENCFF418XDA
Kidney	H3K9me3	bigWig	ENCFF477UKT, ENCFF135FMF
Kidney	H3K9me3	bed	ENCFF727EJL, ENCFF234OMD
Kidney	H3K36me3	bigWig	ENCFF042ZOW, ENCFF978ZRB
Kidney	H3K36me3	bed	ENCFF310IRC, ENCFF688GQP
Kidney	H3K4me3	bigWig	ENCFF480XFJ, ENCFF493NDW
Kidney	H3K4me3	bed	ENCFF349XHZ, ENCFF430ZMN
Kidney	H3K4me1	bigWig	ENCFF408FAL, ENCFF592MFL
Kidney	H3K4me1	bed	ENCFF249OJK, ENCFF955PYL
Liver	H3K9ac	bigWig	ENCFF805VZP, ENCFF775YWU

Table S2: (continued)

Tissue	Target	File type	ENCODE IDs
Liver	H3K9ac	bed	ENCFF344IUW, ENCFF766JJU
Liver	H3K4me1	bigWig	ENCFF425KUU, ENCFF092AOF, ENCFF702XFW, ENCFF926QRN, ENCFF564EMQ, ENCFF774JAU
Liver	H3K4me1	bed	ENCFF953NPP, ENCFF872NIB, ENCFF452XJY, ENCFF003KVZ, ENCFF608RNK, ENCFF189UID
Liver	H3K9me3	bigWig	ENCFF371LJV, ENCFF717HQL, ENCFF650IHL, ENCFF656QXE, ENCFF714DAV, ENCFF292JGY
Liver	H3K9me3	bed	ENCFF524VPY, ENCFF003IMJ, ENCFF093NPF, ENCFF789WSF, ENCFF681HRQ, ENCFF069MMP
Liver	H3K27me3	bigWig	ENCFF513NLK, ENCFF538BJP, ENCFF972NID, ENCFF450IZQ, ENCFF635OEL
Liver	H3K27me3	bed	ENCFF210DQA, ENCFF100SGX, ENCFF977LFS, ENCFF877YOY, ENCFF471HUJ
Liver	H3K27ac	bigWig	ENCFF106PVY, ENCFF555QGS, ENCFF914YXU, ENCFF566EVM, ENCFF150WYF
Liver	H3K27ac	bed	ENCFF805YRQ, ENCFF193GDV, ENCFF837ELM, ENCFF382EMP, ENCFF271HJI
Liver	H3K36me3	bigWig	ENCFF335LNH, ENCFF169JTR, ENCFF758WKE, ENCFF342EVX, ENCFF356MEX, ENCFF515OJV
Liver	H3K36me3	bed	ENCFF220MSF, ENCFF342YOA, ENCFF050ODT, ENCFF191SKQ, ENCFF808ULR, ENCFF734MXG
Liver	H3K4me3	bigWig	ENCFF763IFR, ENCFF215LWY, ENCFF089EHD, ENCFF790UTT, ENCFF307UZR
Liver	H3K4me3	bed	ENCFF416SIL, ENCFF178DYP, ENCFF504QPL, ENCFF522CRH, ENCFF830CCA
Lung	H3K27me3	bigWig	ENCFF157YIA, ENCFF128ZRT, ENCFF082NAO, ENCFF351LVZ, ENCFF891GDP, ENCFF566GDH, ENCFF440HPT, ENCFF846WPB, ENCFF875KAI, ENCFF839EGQ
Lung	H3K27me3	bed	ENCFF901WEE, ENCFF068NOV, ENCFF256JYB, ENCFF725NIC, ENCFF367LMT, ENCFF248IPH, ENCFF112CDW, ENCFF608JQT, ENCFF685CNW, ENCFF591PDC
Lung	H3K27ac	bigWig	ENCFF294WTF, ENCFF333RHP, ENCFF568HPI, ENCFF528NHC, ENCFF028TDS, ENCFF189JKC, ENCFF314IJY, ENCFF239QNU, ENCFF249RKQ, ENCFF303ERR
Lung	H3K27ac	bed	ENCFF882VLE, ENCFF200VYQ, ENCFF496RDA, ENCFF381SFJ, ENCFF314QTE, ENCFF299FWI, ENCFF683QWQ, ENCFF149QUM, ENCFF014OZD, ENCFF448PFH
Lung	H3K36me3	bigWig	ENCFF496ZLH, ENCFF084YHY, ENCFF750XLK, ENCFF862FBH, ENCFF655YLP, ENCFF574SRD, ENCFF631OAP, ENCFF266GKC, ENCFF042SAY, ENCFF302WVU
Lung	H3K36me3	bed	ENCFF671KBQ, ENCFF294NNK, ENCFF459IYH, ENCFF078AMJ, ENCFF089HKW, ENCFF597AKE, ENCFF121NYY, ENCFF829QFZ, ENCFF776LXR, ENCFF616HZY
Lung	H3K4me1	bigWig	ENCFF658BBD, ENCFF687XMK, ENCFF781YFY, ENCFF288VKE, ENCFF525VKS, ENCFF896NVM, ENCFF850BAE, ENCFF580SIO, ENCFF666SBO, ENCFF096JSH
Lung	H3K4me1	bed	ENCFF572UTB, ENCFF516LPC, ENCFF926YKP, ENCFF892HBS, ENCFF141WFE, ENCFF908EUN, ENCFF880EPD, ENCFF908UYY, ENCFF239QVJ, ENCFF641CQH
Lung	H3K9me3	bigWig	ENCFF409QFE, ENCFF527PZR, ENCFF370IDE, ENCFF303PQE, ENCFF189KIW, ENCFF810QZK, ENCFF521VBA, ENCFF970KWF, ENCFF813IBG, ENCFF117QYF
Lung	H3K9me3	bed	ENCFF706CZX, ENCFF357WOH, ENCFF080ZRB, ENCFF632XWZ, ENCFF185EKW, ENCFF699RLH, ENCFF580CFB, ENCFF284QKU, ENCFF937TOI, ENCFF082HMC
Lung	H3K4me3	bigWig	ENCFF993ISE, ENCFF497IMT, ENCFF746KCC, ENCFF466ADC, ENCFF434JKJ, ENCFF880VPE, ENCFF596VKC, ENCFF871GTV, ENCFF379BAB, ENCFF074QVI
Lung	H3K4me3	bed	ENCFF843JPR, ENCFF519JBU, ENCFF742AVP, ENCFF045ZYD, ENCFF254OZF, ENCFF926KAB, ENCFF917VSO, ENCFF937RVI, ENCFF321BAQ, ENCFF143ASW
Ovary	H3K36me3	bigWig	ENCFF642TNX, ENCFF718CFZ, ENCFF177HKS
Ovary	H3K36me3	bed	ENCFF478JCP, ENCFF102MIZ, ENCFF712QEQ
Ovary	H3K9me3	bigWig	ENCFF979RTW, ENCFF602MBJ

Table S2: (continued)

Tissue	Target	File type	ENCODE IDs
Ovary	H3K9me3	bed	ENCFF852ZHM, ENCFF483FPH
Ovary	H3K4me3	bigWig	ENCFF971BNY, ENCFF943GUN
Ovary	H3K4me3	bed	ENCFF985MLM, ENCFF112KXI
Ovary	H3K27me3	bigWig	ENCFF306ZAC, ENCFF383MSI
Ovary	H3K27me3	bed	ENCFF503FCF, ENCFF767BVL
Ovary	H3K27ac	bigWig	ENCFF650ZIC, ENCFF311RDZ, ENCFF262TJJ
Ovary	H3K27ac	bed	ENCFF554TER, ENCFF285RPL, ENCFF036IXW
Ovary	H3K4me1	bigWig	ENCFF866KEX
Ovary	H3K4me1	bed	ENCFF030RVT
Prostate	H3K4me1	bigWig	ENCFF990EER, ENCFF749VXS
Prostate	H3K4me1	bed	ENCFF275KSR, ENCFF404CWJ
Prostate	H3K27me3	bigWig	ENCFF859FTI
Prostate	H3K27me3	bed	ENCFF198QRW
Prostate	H3K36me3	bigWig	ENCFF825ELC, ENCFF149GHN
Prostate	H3K36me3	bed	ENCFF139GKT, ENCFF028CPL
Prostate	H3K27ac	bigWig	ENCFF614EBT, ENCFF769IGG
Prostate	H3K27ac	bed	ENCFF201VZW, ENCFF551MWX
Prostate	H3K9me3	bigWig	ENCFF726YZU
Prostate	H3K9me3	bed	ENCFF311CMQ
Prostate	H3K4me3	bigWig	ENCFF300OWT, ENCFF937OWO
Prostate	H3K4me3	bed	ENCFF945EUR, ENCFF155ZYZ
Skin	H3K27ac	bigWig	ENCFF915CZZ, ENCFF667OTY, ENCFF163EQY, ENCFF674HIP, ENCFF116SYZ, ENCFF394OCM, ENCFF807ITR, ENCFF168NAA, ENCFF109FWI, ENCFF564BHK, ENCFF249DHH, ENCFF197ZLE
Skin	H3K27ac	bed	ENCFF016AAS, ENCFF143DRH, ENCFF924OGR, ENCFF462XVH, ENCFF346GZD, ENCFF087WUV, ENCFF289XOR, ENCFF379RWI, ENCFF240EGF, ENCFF895PGR, ENCFF787FST, ENCFF398EEO
Skin	H3K4me3	bigWig	ENCFF851GVF, ENCFF697CGX
Skin	H3K4me3	bed	ENCFF748OYQ, ENCFF240CFZ
Skin	H3K4me1	bigWig	ENCFF436HQD
Skin	H3K4me1	bed	ENCFF800MDH
Skin	H3K36me3	bigWig	ENCFF232OVE
Skin	H3K36me3	bed	ENCFF804VWO
Skin	H3K9me3	bigWig	ENCFF348IXN
Skin	H3K9me3	bed	ENCFF453KLD
Skin	H3K27me3	bigWig	ENCFF733IEZ
Skin	H3K27me3	bed	ENCFF939OEI

Table S3: Transcription Factor Chip-seq ENCODE metafile URLs. Download links to the ENCODE metafile tables documenting the TF ChIP-Seq experiments used in this study.

Tissue	Metafile URL
Brain	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=brain&type=Experiment&files.analyses.status=released&files.preferred_default=true
Breast	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=breast&type=Experiment&files.analyses.status=released&files.preferred_default=true
Colon	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=colon&type=Experiment&files.analyses.status=released&files.preferred_default=true
Esophagus	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=esophagus&biosample_ontology.term_name%21=esophagus+muscularis+mucosa&type=Experiment&files.analyses.status=released&files.preferred_default=true
Kidney	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=kidney&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S3: (continued)

Tissue	Metafile URL
Liver	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=liver&type=Experiment&files.analyses.status=released&files.preferred_default=true
Lung	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=lung&type=Experiment&files.analyses.status=released&files.preferred_default=true
Ovary	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=ovary&type=Experiment&files.analyses.status=released&files.preferred_default=true
Prostate	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=prostate+gland&type=Experiment&files.analyses.status=released&files.preferred_default=true
Skin	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=TF+ChIP-seq&biosample_ontology.organ_slims=skin+of+body&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S4: Transcription factor Chip-seq experiments. List of ENCODE IDs corresponding to TF ChIP-Seq experiments used in this study. For each tissue and each ChIP-Seq target, we used bed files encompassing binding peaks as well as bigWig files encompassing signal p-value along the genome.

Tissue	Target	File type	ENCODE IDs
Brain	CTCF	bigWig	ENCFF702KLV, ENCFF263PRD, ENCFF063NKV, ENCFF151KEM, ENCFF298FKN, ENCFF691VIC, ENCFF211KYR, ENCFF154HTF, ENCFF409ZVA, ENCFF744MME, ENCFF057RPH, ENCFF263QKY, ENCFF720LRO, ENCFF506OJS, ENCFF109VST, ENCFF963JJJ, ENCFF150CFZ, ENCFF187ZZX, ENCFF491NJI, ENCFF216SWA, ENCFF848JZX
Brain	CTCF	bed	ENCFF427DTQ, ENCFF306BLG, ENCFF816RIB, ENCFF422EHF, ENCFF512PKN, ENCFF292VJF, ENCFF246XGW, ENCFF695XWD, ENCFF260CWT, ENCFF141KSG, ENCFF659ZSO, ENCFF238ZLG, ENCFF132UGC, ENCFF835BZY, ENCFF477QFS, ENCFF821ZJF, ENCFF534RYO, ENCFF108PPI, ENCFF881GDD, ENCFF535ATM, ENCFF201FRB
Breast	POLR2AphosphoS5	bigWig	ENCFF149MWU, ENCFF811OSL, ENCFF691WWI
Breast	POLR2AphosphoS5	bed	ENCFF269FRP, ENCFF267PHM, ENCFF221FKW
Breast	POLR2A	bigWig	ENCFF027JBQ, ENCFF212JWB
Breast	POLR2A	bed	ENCFF555UBC, ENCFF538YRS
Breast	CTCF	bigWig	ENCFF444GVF, ENCFF375RUG, ENCFF170QFS
Breast	CTCF	bed	ENCFF910ACN, ENCFF758FTM, ENCFF530ARF
Colon	CTCF	bigWig	ENCFF238UUV, ENCFF190IDG, ENCFF637OTM, ENCFF992JAI, ENCFF168VKA, ENCFF213WDF, ENCFF370OBU, ENCFF816TDF, ENCFF125KTC, ENCFF524MDB, ENCFF654FSU, ENCFF606UXM, ENCFF818RUE, ENCFF984JCD
Colon	CTCF	bed	ENCFF484VPU, ENCFF351URT, ENCFF480QYI, ENCFF015NKY, ENCFF860GRV, ENCFF613UQP, ENCFF292GYB, ENCFF681RYK, ENCFF829XMK, ENCFF324DOU, ENCFF115EJR, ENCFF827OLU, ENCFF663OEV, ENCFF168XIT
Colon	POLR2A	bigWig	ENCFF816DQM, ENCFF540ALG, ENCFF976VVU, ENCFF135SVJ, ENCFF670UMV, ENCFF509WCN, ENCFF455IVD, ENCFF353NLM
Colon	POLR2A	bed	ENCFF469KDX, ENCFF993GPP, ENCFF838LKZ, ENCFF536KQX, ENCFF061DKX, ENCFF836KEF, ENCFF890XQW, ENCFF750UAQ
Colon	EP300	bigWig	ENCFF800IMF, ENCFF529SJY, ENCFF373GRU, ENCFF093NJN, ENCFF861SLP
Colon	EP300	bed	ENCFF957TMT, ENCFF740VFC, ENCFF758MFR, ENCFF754LGO, ENCFF992RST
Colon	POLR2AphosphoS5	bigWig	ENCFF510FHL, ENCFF194WGL, ENCFF704WEB, ENCFF720LYZ, ENCFF956SUX, ENCFF499TVH, ENCFF637HLE
Colon	POLR2AphosphoS5	bed	ENCFF582ZOJ, ENCFF759CQD, ENCFF988SVM, ENCFF683IRK, ENCFF023HDE, ENCFF347VIP, ENCFF688OHP
Esophagus	POLR2A	bigWig	ENCFF051LIT, ENCFF182XEN, ENCFF227ZFC, ENCFF790OFR
Esophagus	POLR2A	bed	ENCFF872EBX, ENCFF254MJA, ENCFF473YGW, ENCFF673VSN
Esophagus	CTCF	bigWig	ENCFF927KRC, ENCFF795IAK, ENCFF874QMW, ENCFF644ESC, ENCFF875HLJ, ENCFF515HEW
Esophagus	CTCF	bed	ENCFF523MWC, ENCFF129JPF, ENCFF384CCC, ENCFF848GUU, ENCFF548ULK, ENCFF594HEK
Liver	POLR2A	bigWig	ENCFF506OAE
Liver	POLR2A	bed	ENCFF064ASM
Liver	CTCF	bigWig	ENCFF184SOF, ENCFF478LYE, ENCFF309RUL
Liver	CTCF	bed	ENCFF046XEZ, ENCFF870ZRR, ENCFF585ZAP
Lung	CTCF	bigWig	ENCFF490HYH, ENCFF946JYR, ENCFF881JZP, ENCFF586CIG, ENCFF983JGW, ENCFF828GRP, ENCFF295VVB, ENCFF494UNI, ENCFF865MBV, ENCFF683EZS, ENCFF404FYD, ENCFF975YTY
Lung	CTCF	bed	ENCFF629UDP, ENCFF529BHM, ENCFF690VRK, ENCFF647HXE, ENCFF061UVF, ENCFF354FKN, ENCFF786YIA, ENCFF887BKS, ENCFF238PEB, ENCFF992VPQ, ENCFF183WDD, ENCFF711KWX
Lung	POLR2A	bigWig	ENCFF977BOZ, ENCFF117QIN, ENCFF802VYO, ENCFF353VIA
Lung	POLR2A	bed	ENCFF082DKP, ENCFF843PJA, ENCFF389ZBU, ENCFF482BDW
Lung	POLR2AphosphoS5	bigWig	ENCFF875XRN, ENCFF387AVD
Lung	POLR2AphosphoS5	bed	ENCFF506HXM, ENCFF311CIX

Table S4: (continued)

Tissue	Target	File type	ENCODE IDs
Lung	EP300	bigWig	ENCFF810GBA, ENCFF360HGF, ENCFF251RNH, ENCFF930IZX
Lung	EP300	bed	ENCFF833AES, ENCFF976JNG, ENCFF689DSV, ENCFF576FCY
Ovary	EP300	bigWig	ENCFF858IGH
Ovary	EP300	bed	ENCFF917NDE
Ovary	POLR2A	bigWig	ENCFF349SSW
Ovary	POLR2A	bed	ENCFF105LDY
Ovary	CTCF	bigWig	ENCFF437GZJ, ENCFF546NJL, ENCFF481DRE
Ovary	CTCF	bed	ENCFF130ADS, ENCFF782GIY, ENCFF483KVM
Prostate	POLR2AphosphoS5	bigWig	ENCFF733MOG, ENCFF845ZWO
Prostate	POLR2AphosphoS5	bed	ENCFF696QPX, ENCFF581YZY
Prostate	CTCF	bigWig	ENCFF430NIU, ENCFF163FVP, ENCFF563VCW, ENCFF697UNS
Prostate	CTCF	bed	ENCFF703KDQ, ENCFF351NKT, ENCFF771ZYH, ENCFF400NQB
Prostate	POLR2A	bigWig	ENCFF100FTA, ENCFF152LGU
Prostate	POLR2A	bed	ENCFF151LHR, ENCFF558UJR
Skin	POLR2A	bigWig	ENCFF905LFM, ENCFF064TYZ, ENCFF767JSK, ENCFF686APW, ENCFF714IUH
Skin	POLR2A	bed	ENCFF616GPO, ENCFF885ZWB, ENCFF012VQA, ENCFF673HLW, ENCFF869HZM
Skin	CTCF	bigWig	ENCFF739PCC, ENCFF422IST, ENCFF693XPS, ENCFF876NXN, ENCFF714ZOY, ENCFF600UVT, ENCFF303VDR
Skin	CTCF	bed	ENCFF147HSY, ENCFF985QNV, ENCFF525JGP, ENCFF503GIR, ENCFF405VLP, ENCFF292NVM, ENCFF490CTJ
Skin	POLR2AphosphoS5	bigWig	ENCFF830GHT, ENCFF086KFQ, ENCFF575XET
Skin	POLR2AphosphoS5	bed	ENCFF190ASU, ENCFF227SQC, ENCFF576TSY
Skin	EP300	bigWig	ENCFF052PKX
Skin	EP300	bed	ENCFF741MWZ

Table S5: DNAse-seq ENCODE metafile URLs. Download links to the ENCODE metafile tables documenting the DNAse-seq experiments used in this study.

Tissue	Metafile URL
Brain	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=brain&type=Experiment&files.analyses.status=released&files.preferred_default=true
Breast	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=breast&type=Experiment&files.analyses.status=released&files.preferred_default=true
Colon	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=colon&type=Experiment&files.analyses.status=released&files.preferred_default=true
Esophagus	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=esophagus&biosample_ontology.term_name%21=esophagus+muscularis+mucosa&type=Experiment&files.analyses.status=released&files.preferred_default=true
Kidney	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=kidney&biosample_ontology.term_name%21=ureter&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S5: (continued)

Tissue	Metafile URL
Liver	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=liver&type=Experiment&files.analyses.status=released&files.preferred_default=true
Lung	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=lung&type=Experiment&files.analyses.status=released&files.preferred_default=true
Ovary	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=ovary&type=Experiment&files.analyses.status=released&files.preferred_default=true
Prostate	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&biosample_ontology.organ_slims=prostate+gland&type=Experiment&files.analyses.status=released&files.preferred_default=true
Skin	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&biosample_ontology.organ_slims=skin+of+body&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=DNase-seq&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S6: DNAse-seq experiments. List of ENCODE IDs corresponding to DNAse-seq experiments used in this study. For each tissue and each ChIP-Seq target, we used bed files encompassing binding peaks as well as bigWig files encompassing signal p-value along the genome.

Tissue	File type	ENCODE IDs
Brain	bigWig	ENCFF520MAO, ENCFF270GJA, ENCFF243QQE, ENCFF722YGX, ENCFF277PYT, ENCFF058QYM, ENCFF614XBI, ENCFF670TZZ, ENCFF302EVI, ENCFF018BZK, ENCFF449WDF, ENCFF048ZLO, ENCFF427FGG, ENCFF604AYM, ENCFF520DAO, ENCFF549PNN, ENCFF753DPM, ENCFF940HNG, ENCFF648JOD, ENCFF571QPS, ENCFF634CTV, ENCFF878ZQA, ENCFF057JIJ, ENCFF157NVB, ENCFF836SMO, ENCFF520QRU, ENCFF487QRP, ENCFF052CPA, ENCFF011RNN, ENCFF434ZTO, ENCFF143GBL, ENCFF436OUJ, ENCFF293YDF, ENCFF163NDW, ENCFF023UNU, ENCFF715WQG, ENCFF295JAT, ENCFF298BKQ, ENCFF586PFT, ENCFF477CEG, ENCFF812WBM, ENCFF258AWM, ENCFF497CVA, ENCFF354GAT, ENCFF301CMS, ENCFF900YTP, ENCFF566NSN, ENCFF130TYS, ENCFF497DIH, ENCFF158ZYM, ENCFF036SMP, ENCFF855MTJ, ENCFF289FQY, ENCFF636JMM, ENCFF733ZHD, ENCFF813UXN, ENCFF713MFZ, ENCFF243RUA, ENCFF153ZPN, ENCFF394GBX, ENCFF372RPI, ENCFF286BFK, ENCFF769AFQ, ENCFF796XMI, ENCFF636BNY, ENCFF179VUT, ENCFF505HNJ, ENCFF287RNZ, ENCFF514GFX, ENCFF775TPT, ENCFF823JWZ, ENCFF283AUC, ENCFF874WYJ, ENCFF941ZIR, ENCFF905SCU, ENCFF459TMW, ENCFF980SJY, ENCFF284PMB, ENCFF184JSH, ENCFF148TKD, ENCFF936VRA
Brain	bed	ENCFF109NJY, ENCFF400QKK, ENCFF635VUD, ENCFF880TFT, ENCFF219JBP, ENCFF049FKM, ENCFF359HZJ, ENCFF691UUG, ENCFF556NPT, ENCFF328UXU, ENCFF127HQV, ENCFF102QMS, ENCFF433LZG, ENCFF421GQV, ENCFF947BGT, ENCFF692HIA, ENCFF491YIR, ENCFF222AXX, ENCFF661CDH, ENCFF521CSC, ENCFF083DFD, ENCFF396EDV, ENCFF142UYG, ENCFF980SGP, ENCFF488OCQ, ENCFF211WGV, ENCFF037ADM, ENCFF351AFQ, ENCFF084OJU, ENCFF406DHI, ENCFF280SAO, ENCFF825JOD, ENCFF319VYE, ENCFF620YIX, ENCFF721HSD, ENCFF721JCD, ENCFF900IJE, ENCFF067LRX, ENCFF782XPH, ENCFF100UJA, ENCFF022TLX, ENCFF066MUO, ENCFF264NFU, ENCFF101IGO, ENCFF570VBF, ENCFF124CVB, ENCFF123RHR, ENCFF099FUV, ENCFF493KKC, ENCFF060BYF, ENCFF750DUU, ENCFF922ZXC, ENCFF299CHI, ENCFF008WSU, ENCFF247GJR, ENCFF981IQP, ENCFF854UOQ, ENCFF006ZAY, ENCFF400BFG, ENCFF234IMJ, ENCFF580YEP, ENCFF853PIR, ENCFF750SYW, ENCFF443KKI, ENCFF261ORB, ENCFF295XKP, ENCFF822HOO, ENCFF092MVE, ENCFF211HMH, ENCFF183VMB, ENCFF338NXN, ENCFF355JHP, ENCFF079GMX, ENCFF290LOQ, ENCFF491FEU, ENCFF374LXL, ENCFF267FDQ, ENCFF864LVN, ENCFF887QIZ, ENCFF358SPE, ENCFF792FEC
Breast	bigWig	ENCFF549MXK, ENCFF111NRY
Breast	bed	ENCFF594NFE, ENCFF557QSP
Colon	bigWig	ENCFF613BDA, ENCFF755FPE, ENCFF305IPH, ENCFF528JMP, ENCFF753MXL, ENCFF299OOV, ENCFF452PQB, ENCFF841ILF, ENCFF291LZE, ENCFF405NTZ
Colon	bed	ENCFF341AEQ, ENCFF157TDQ, ENCFF796LSS, ENCFF515TNB, ENCFF503AZD, ENCFF274XSP, ENCFF042YMQ, ENCFF769AMF, ENCFF903UTH, ENCFF690KZJ
Kidney	bigWig	ENCFF485CJZ
Kidney	bed	ENCFF132KNJ
Liver	bigWig	ENCFF412PVV, ENCFF020EPF, ENCFF634BMR, ENCFF606YDZ, ENCFF082KAC
Liver	bed	ENCFF644RPH, ENCFF905THS, ENCFF629FBJ, ENCFF096WOZ, ENCFF912XTH
Lung	bigWig	ENCFF387FUO, ENCFF831NQC, ENCFF162GDL, ENCFF883DDK, ENCFF615BRP, ENCFF960WAV, ENCFF807ZJE, ENCFF421SBY, ENCFF279ZNA, ENCFF674RXU, ENCFF505TAB, ENCFF787QGK, ENCFF918MVE, ENCFF990HTO
Lung	bed	ENCFF889ADL, ENCFF751WMO, ENCFF151SXP, ENCFF752THQ, ENCFF977IJW, ENCFF293PAS, ENCFF226HAT, ENCFF190WKI, ENCFF041XPT, ENCFF993CLU, ENCFF155DNH, ENCFF136IYG, ENCFF169ADV, ENCFF349QTP
Ovary	bigWig	ENCFF764IRG, ENCFF971KWT, ENCFF276GVH, ENCFF635IQN, ENCFF406GUT, ENCFF792AMR, ENCFF779FIH, ENCFF887BVC
Ovary	bed	ENCFF663IMO, ENCFF115GAD, ENCFF608SOD, ENCFF199AYL, ENCFF418PRB, ENCFF172OPL, ENCFF816OUE, ENCFF158YTV
Prostate	bigWig	ENCFF865IXT
Prostate	bed	ENCFF233XNV
Skin	bigWig	ENCFF642VOM, ENCFF241LIT, ENCFF548VEV, ENCFF654RET
Skin	bed	ENCFF780WOI, ENCFF116DZJ, ENCFF577YQU, ENCFF669VJK

Table S7: ATAC-seq ENCODE metafile URLs. Download links to the ENCODE metafile tables documenting the ATAC-seq experiments used in this study.

Tissue	Metafile URL
Brain	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.classification=tissue&biosample_ontology.organ_slims=brain&type=Experiment&files.analyses.status=released&files.preferred_default=true
Breast	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.classification=tissue&biosample_ontology.organ_slims=breast&type=Experiment&files.analyses.status=released&files.preferred_default=true
Colon	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.classification=tissue&biosample_ontology.organ_slims=colon&type=Experiment&files.analyses.status=released&files.preferred_default=true
Esophagus	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.classification=tissue&biosample_ontology.organ_slims=esophagus&biosample_ontology.term_name%21=esophagus+muscularis+mucosa&type=Experiment&files.analyses.status=released&files.preferred_default=true
Kidney	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.organ_slims=kidney&biosample_ontology.term_name%21=ureter&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S7: (continued)

Tissue	Metafile URL
Liver	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.organ_slims=liver&type=Experiment&files.analyses.status=released&files.preferred_default=true
Lung	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.organ_slims=lung&type=Experiment&files.analyses.status=released&files.preferred_default=true
Ovary	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.organ_slims=ovary&type=Experiment&files.analyses.status=released&files.preferred_default=true
Prostate	https://www.encodeproject.org/metadata/?replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released&biosample_ontology.classification=tissue&control_type%21=%2A&replicates.library.biosample.life_stage=adult&perturbed=false&replicates.library.biosample.disease_term_name%21=Alzheimer%27s+disease&replicates.library.biosample.disease_term_name%21=mild+cognitive+impairment&replicates.library.biosample.disease_term_name%21=Cognitive+impairment&replicates.library.biosample.disease_term_name%21=nonobstructive+coronary+artery+disease&replicates.library.biosample.disease_term_name%21=squamous+cell+carcinoma&replicates.library.biosample.disease_term_name%21=basal+cell+carcinoma&assay_title=ATAC-seq&biosample_ontology.organ_slims=prostate+gland&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S8: ATAC-seq experiments. List of ENCODE IDs corresponding to ATAC-seq experiments used in this study. For each tissue and each ChIP-Seq target, we used bed files encompassing binding peaks as well as bigWig files encompassing signal p-value along the genome.

Tissue	File type	ENCODE IDs
Brain	bigWig	ENCFF160VHY, ENCFF143FWG
Brain	bed	ENCFF081WNU, ENCFF129OYC
Breast	bigWig	ENCFF421QRX, ENCFF645PIY, ENCFF846PKO
Breast	bed	ENCFF934JPB, ENCFF402SHW, ENCFF749WUW
Colon	bigWig	ENCFF163IDS, ENCFF355IHN, ENCFF275VQX, ENCFF449DFP, ENCFF243ZNH, ENCFF299MFR, ENCFF807CHN, ENCFF628ARL, ENCFF469WCM, ENCFF686EAS, ENCFF575JBQ, ENCFF969UDD
Colon	bed	ENCFF721WEZ, ENCFF591HMI, ENCFF563UYM, ENCFF081FUV, ENCFF355GMG, ENCFF191ZOC, ENCFF846PJS, ENCFF270JNZ, ENCFF054QTC, ENCFF452WBJ, ENCFF009YES, ENCFF018EMP
Esophagus	bigWig	ENCFF413HMJ, ENCFF218UBX
Esophagus	bed	ENCFF068ISD, ENCFF181DWD
Kidney	bigWig	ENCFF562QRD
Kidney	bed	ENCFF845OZO
Liver	bigWig	ENCFF539JSQ, ENCFF258LPU, ENCFF318PKW, ENCFF295COO, ENCFF435JGG, ENCFF275TEH, ENCFF341RHY
Liver	bed	ENCFF631JIS, ENCFF415YNL, ENCFF188RKV, ENCFF318SNW, ENCFF488BRH, ENCFF854ULK, ENCFF360SWK
Lung	bigWig	ENCFF410SKR, ENCFF318GRT, ENCFF512MYF, ENCFF210HIS, ENCFF877MLL, ENCFF831ZWL, ENCFF312PDG
Lung	bed	ENCFF901RIY, ENCFF537TOU, ENCFF189LZA, ENCFF818EYD, ENCFF906HOT, ENCFF899TQV, ENCFF348XSG
Ovary	bigWig	ENCFF752AVC, ENCFF460RIA, ENCFF855DNF, ENCFF930UPB, ENCFF167GOY, ENCFF308DTK
Ovary	bed	ENCFF871DTB, ENCFF380AEO, ENCFF497SGQ, ENCFF586QDS, ENCFF985FDY, ENCFF596YVQ
Prostate	bigWig	ENCFF358ZBP
Prostate	bed	ENCFF030AIT

Table S9: [Methylation data. Data used as reference methylome for each tissue. Data was taken from the study from (Loyfer et al., 2023).

Tissue	File names
Brain	GSM5652219_Oligodendrocytes-Z000000TK.bigwig,
	GSM5652220_Oligodendrocytes-Z0000042E.bigwig,
	GSM5652221_Oligodendrocytes-Z0000042L.bigwig,
	GSM5652222_Oligodendrocytes-Z0000042N.bigwig,
	GSM5652223_Cortex-Neuron-Z000000TF.bigwig, GSM5652224_Neuron-Z000000TH.bigwig,
	GSM5652225_Cortex-Neuron-Z0000042F.bigwig, GSM5652226_Cortex-Neuron-Z0000042H.bigwig,
	GSM5652227_Cortex-Neuron-Z0000042J.bigwig, GSM5652228_Cortex-Neuron-Z0000042M.bigwig,
	GSM5652229_Cortex-Neuron-Z0000042P.bigwig, GSM5652230_Cortex-Neuron-Z0000042K.bigwig,
	GSM5652231_Cerebellum-Neuron-Z000000TB.bigwig,
	GSM5652232_Cortex-Neuron-Z000000TD.bigwig
Breast	GSM5652347_Breast-Luminal-Epithelial-Z000000V2.bigwig,
	GSM5652348_Breast-Luminal-Epithelial-Z000000VJ.bigwig,
	GSM5652349_Breast-Luminal-Epithelial-Z000000VN.bigwig,
	GSM5652350_Breast-Basal-Epithelial-Z000000V6.bigwig,
	GSM5652351_Breast-Basal-Epithelial-Z000000VG.bigwig,
	GSM5652352_Breast-Basal-Epithelial-Z000000VL.bigwig,
	GSM5652353_Breast-Basal-Epithelial-Z0000043E.bigwig
Colon	GSM5652370_Colon-Right-Epithelial-Z000000V0.bigwig,
	GSM5652371_Colon-Right-Epithelial-Z000000V8.bigwig,
	GSM5652372_Colon-Right-Endocrine-Z0000044S.bigwig,
	GSM5652373_Colon-Left-Epithelial-Z000000VA.bigwig,
	GSM5652374_Colon-Left-Endocrine-Z0000044J.bigwig,
	GSM5652375_Colon-Left-Endocrine-Z0000044T.bigwig,
	GSM5652376_Colon-Left-Epithelial-Z0000043B.bigwig,
	GSM5652377_Colon-Left-Epithelial-Z0000043C.bigwig
Esophagus	GSM5652332_Esophagus-Epithelial-Z000000PZ.bigwig,
	GSM5652333_Esophagus-Epithelial-Z00000426.bigwig
Kidney	GSM5652187_Kidney-Tubular-Endothel-Z000000PX.bigwig,
	GSM5652188_Kidney-Tubular-Endothel-Z000000Q3.bigwig,
	GSM5652189_Kidney-Tubular-Endothel-Z0000042R.bigwig,
	GSM5652258_Kidney-Tubular-Epithelial-Z000000QH.bigwig,
	GSM5652259_Kidney-Tubular-Epithelial-Z0000043Z.bigwig,
	GSM5652260_Kidney-Tubular-Epithelial-Z00000440.bigwig
Liver	GSM5652233_Liver-Hepatocytes-Z000000R3.bigwig,
	GSM5652234_Liver-Hepatocytes-Z000000T3.bigwig,
	GSM5652235_Liver-Hepatocytes-Z0000043Q.bigwig,
	GSM5652236_Liver-Hepatocytes-Z0000044H.bigwig,
	GSM5652237_Liver-Hepatocytes-Z0000044M.bigwig,
	GSM5652238_Liver-Hepatocytes-Z00000431.bigwig
Lung	GSM5652335_Lung-Bronchus-Epithelial-Z000000QD.bigwig,
	GSM5652336_Lung-Bronchus-Epithelial-Z000000RZ.bigwig,
	GSM5652337_Lung-Bronchus-Epithelial-Z000000S5.bigwig,
	GSM5652354_Lung-Alveolar-Epithelial-Z000000T1.bigwig,
	GSM5652355_Lung-Alveolar-Epithelial-Z000000VC.bigwig,
	GSM5652356_Lung-Alveolar-Epithelial-Z000000VE.bigwig
Ovary	GSM5652270_Ovary-Epithelial-Z000000QT.bigwig
Prostate	GSM5652338_Prostate-Epithelial-Z000000RV.bigwig,
	GSM5652339_Prostate-Epithelial-Z000000S3.bigwig,
	GSM5652340_Prostate-Epithelial-Z0000045F.bigwig,
	GSM5652341_Prostate-Epithelial-Z0000045G.bigwig
Skin	GSM5652321_Epidermal-Keratinocytes-Z00000424.bigwig

Table S10: Hi-C ENCODE metafile URLs. Download links to the ENCODE metafile tables documenting the Hi-C experiments used in this study.

Tissue	Metafile URL
Brain	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&biosample_ontology.organ_slims=brain&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.classification%21=in+vitro+differentiated+cells&type=Experiment&files.analyses.status=released&files.preferred_default=true
Breast	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.classification%21=in+vitro+differentiated+cells&biosample_ontology.organ_slims=mammary+gland&type=Experiment&files.analyses.status=released&files.preferred_default=true
Colon	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.classification%21=in+vitro+differentiated+cells&biosample_ontology.organ_slims=colon&biosample_ontology.classification=tissue&type=Experiment&files.analyses.status=released&files.preferred_default=true
Esophagus	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.organ_slims=esophagus&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S10: (continued)

Tissue	Metafile URL
Kidney	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.organ_slims=kidney&type=Experiment&files.analyses.status=released&files.preferred_default=true
Liver	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.organ_slims=liver&type=Experiment&files.analyses.status=released&files.preferred_default=true
Lung	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.organ_slims=lung&type=Experiment&files.analyses.status=released&files.preferred_default=true
Ovary	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.organ_slims=ovary&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S10: (continued)

Tissue	Metafile URL
Prostate	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.organ_slims=prostate+gland&type=Experiment&files.analyses.status=released&files.preferred_default=true
Skin	https://www.encodeproject.org/metadata/?control_type%21=%2A&status=released&perturbed=false&assay_slims=3D+chromatin+structure&assay_title=in+situ+Hi-C&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&replicates.library.biosample.treatments.treatment_term_name%21=anti-CD3+and+anti-CD28+coated+beads&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-4&replicates.library.biosample.treatments.treatment_term_name%21=Interferon-gamma+antibody&replicates.library.biosample.treatments.treatment_term_name%21=Interleukin-2&replicates.library.biosample.treatments.treatment_term_name%21=granulocyte-macrophage+colony-stimulating+factor&replicates.library.biosample.treatments.treatment_term_name%21=phorbol+12-myristate+13-acetate&replicates.library.biosample.treatments.treatment_term_name%21=retinoic+acid&assembly=GRCh38&assay_title=intact+Hi-C&replicates.library.biosample.life_stage=adult&replicates.library.biosample.life_stage=unknown&biosample_ontology.organ_slims=skin+of+body&type=Experiment&files.analyses.status=released&files.preferred_default=true

Table S11: ATAC-seq experiments. List of ENCODE IDs corresponding to Hi-C experiments used in this study. For each tissue, we accessed genome compartment annotations as well as mapping quality thresholded contact matrices, depending on availability.

Tissue	File type	ENCODE IDs
Brain	genome compartments	ENCFF808XAN, ENCFF419EFH, ENCFF245BDA
Brain	mapping quality thresholded contact matrix	ENCFF925QIF, ENCFF661EVS, ENCFF711XSR, ENCFF163YBP, ENCFF594KOM, ENCFF068LTZ
Breast	genome compartments	ENCFF168IVG, ENCFF863UVK, ENCFF327LIB
Breast	mapping quality thresholded contact matrix	ENCFF420JTA, ENCFF512PQA, ENCFF541RAA, ENCFF977XWK, ENCFF943JRY, ENCFF832RTC
Colon	genome compartments	ENCFF379IPO, ENCFF782IKO, ENCFF450RDW, ENCFF149MVD
Colon	mapping quality thresholded contact matrix	ENCFF579CAR, ENCFF660JWA, ENCFF035BLF, ENCFF309UNV, ENCFF897KRI, ENCFF355NFJ, ENCFF446NXA, ENCFF149IHN, ENCFF654YIQ, ENCFF059WMS
Esophagus	mapping quality thresholded contact matrix	ENCFF458VBB, ENCFF224HKR
Kidney	genome compartments	ENCFF818XKK, ENCFF057DGR
Kidney	mapping quality thresholded contact matrix	ENCFF954SBX, ENCFF417GBZ, ENCFF528JJC
Liver	genome compartments	ENCFF260UAB
Liver	mapping quality thresholded contact matrix	ENCFF952JZV, ENCFF705YZH, ENCFF861EKC
Lung	genome compartments	ENCFF079BUT, ENCFF575JLT
Lung	mapping quality thresholded contact matrix	ENCFF237UKR, ENCFF896OFN, ENCFF304HMS, ENCFF649OHR, ENCFF395INO, ENCFF181ROW, ENCFF676IVH, ENCFF689CUX
Ovary	mapping quality thresholded contact matrix	ENCFF496GEU, ENCFF156GGD, ENCFF700CYI, ENCFF019YVI
Prostate	genome compartments	ENCFF445CHI
Prostate	mapping quality thresholded contact matrix	ENCFF515ZBF, ENCFF678ZLX
Skin	genome compartments	ENCFF303AUY, ENCFF419POH, ENCFF206ZKC
Skin	mapping quality thresholded contact matrix	ENCFF702IFC, ENCFF628LSG, ENCFF557CTN

Table S12: Repli-seq data. List of Repli-seq experiments used in this study. We used `bigWig` files containing the replication signal ("WaveSignal") along the genome as well as bed files with "Peaks" and "Valleys" representing replication origin and termination sites.

Cell line	View	dccAccession	File name
BG02ES	Peaks	wgEncodeEH002250	wgEncodeUwRepliSeqBg02esPkRep1.bed.gz
BG02ES	Valleys	wgEncodeEH002250	wgEncodeUwRepliSeqBg02esValleysRep1.bed.gz
BG02ES	WaveSignal	wgEncodeEH002250	wgEncodeUwRepliSeqBg02esWaveSignalRep1.bigWig
BJ	Peaks	wgEncodeEH002236	wgEncodeUwRepliSeqBjPkRep1.bed.gz
BJ	Peaks	wgEncodeEH002236	wgEncodeUwRepliSeqBjPkRep2.bed.gz
BJ	Valleys	wgEncodeEH002236	wgEncodeUwRepliSeqBjValleysRep1.bed.gz
BJ	Valleys	wgEncodeEH002236	wgEncodeUwRepliSeqBjValleysRep2.bed.gz
BJ	WaveSignal	wgEncodeEH002236	wgEncodeUwRepliSeqBjWaveSignalRep1.bigWig
BJ	WaveSignal	wgEncodeEH002236	wgEncodeUwRepliSeqBjWaveSignalRep2.bigWig
GM06990	Peaks	wgEncodeEH002237	wgEncodeUwRepliSeqGm06990PkRep1.bed.gz
GM06990	Valleys	wgEncodeEH002237	wgEncodeUwRepliSeqGm06990ValleysRep1.bed.gz
GM06990	WaveSignal	wgEncodeEH002237	wgEncodeUwRepliSeqGm06990WaveSignalRep1.bigWig
GM12801	Peaks	wgEncodeEH002238	wgEncodeUwRepliSeqGm12801PkRep1.bed.gz
GM12801	Valleys	wgEncodeEH002238	wgEncodeUwRepliSeqGm12801ValleysRep1.bed.gz
GM12801	WaveSignal	wgEncodeEH002238	wgEncodeUwRepliSeqGm12801WaveSignalRep1.bigWig
GM12812	Peaks	wgEncodeEH002239	wgEncodeUwRepliSeqGm12812PkRep1.bed.gz
GM12812	Valleys	wgEncodeEH002239	wgEncodeUwRepliSeqGm12812ValleysRep1.bed.gz
GM12812	WaveSignal	wgEncodeEH002239	wgEncodeUwRepliSeqGm12812WaveSignalRep1.bigWig
GM12813	Peaks	wgEncodeEH002240	wgEncodeUwRepliSeqGm12813PkRep1.bed.gz
GM12813	Valleys	wgEncodeEH002240	wgEncodeUwRepliSeqGm12813ValleysRep1.bed.gz
GM12813	WaveSignal	wgEncodeEH002240	wgEncodeUwRepliSeqGm12813WaveSignalRep1.bigWig
GM12878	Peaks	wgEncodeEH002241	wgEncodeUwRepliSeqGm12878PkRep1.bed.gz
GM12878	Valleys	wgEncodeEH002241	wgEncodeUwRepliSeqGm12878ValleysRep1.bed.gz
GM12878	WaveSignal	wgEncodeEH002241	wgEncodeUwRepliSeqGm12878WaveSignalRep1.bigWig
HeLa-S3	Peaks	wgEncodeEH002243	wgEncodeUwRepliSeqHelas3PkRep1.bed.gz
HeLa-S3	Valleys	wgEncodeEH002243	wgEncodeUwRepliSeqHelas3ValleysRep1.bed.gz
HeLa-S3	WaveSignal	wgEncodeEH002243	wgEncodeUwRepliSeqHelas3WaveSignalRep1.bigWig
HepG2	Peaks	wgEncodeEH002244	wgEncodeUwRepliSeqHepg2PkRep1.bed.gz
HepG2	Valleys	wgEncodeEH002244	wgEncodeUwRepliSeqHepg2ValleysRep1.bed.gz
HepG2	WaveSignal	wgEncodeEH002244	wgEncodeUwRepliSeqHepg2WaveSignalRep1.bigWig
HUVEC	Peaks	wgEncodeEH002242	wgEncodeUwRepliSeqHuvecPkRep1.bed.gz
HUVEC	Valleys	wgEncodeEH002242	wgEncodeUwRepliSeqHuvecValleysRep1.bed.gz
HUVEC	WaveSignal	wgEncodeEH002242	wgEncodeUwRepliSeqHuvecWaveSignalRep1.bigWig
IMR90	Peaks	wgEncodeEH002245	wgEncodeUwRepliSeqImr90PkRep1.bed.gz
IMR90	Valleys	wgEncodeEH002245	wgEncodeUwRepliSeqImr90ValleysRep1.bed.gz
IMR90	WaveSignal	wgEncodeEH002245	wgEncodeUwRepliSeqImr90WaveSignalRep1.bigWig
K562	Peaks	wgEncodeEH002246	wgEncodeUwRepliSeqK562PkRep1.bed.gz
K562	Valleys	wgEncodeEH002246	wgEncodeUwRepliSeqK562ValleysRep1.bed.gz
K562	WaveSignal	wgEncodeEH002246	wgEncodeUwRepliSeqK562WaveSignalRep1.bigWig
MCF-7	Peaks	wgEncodeEH002247	wgEncodeUwRepliSeqMcf7PkRep1.bed.gz
MCF-7	Valleys	wgEncodeEH002247	wgEncodeUwRepliSeqMcf7ValleysRep1.bed.gz
MCF-7	WaveSignal	wgEncodeEH002247	wgEncodeUwRepliSeqMcf7WaveSignalRep1.bigWig
NHEK	Peaks	wgEncodeEH002249	wgEncodeUwRepliSeqNhekPkRep1.bed.gz
NHEK	Valleys	wgEncodeEH002249	wgEncodeUwRepliSeqNhekValleysRep1.bed.gz
NHEK	WaveSignal	wgEncodeEH002249	wgEncodeUwRepliSeqNhekWaveSignalRep1.bigWig
SK-N-SH	Peaks	wgEncodeEH002384	wgEncodeUwRepliSeqSknshPkRep1.bed.gz
SK-N-SH	Valleys	wgEncodeEH002384	wgEncodeUwRepliSeqSknshValleysRep1.bed.gz
SK-N-SH	WaveSignal	wgEncodeEH002384	wgEncodeUwRepliSeqSknshWaveSignalRep1.bigWig

Table S13: Exome predictor mapping configuration file. Table with configurations used for the mapping of predictors to query exonic positions. The contents of the columns are as follows. Name: Name of the predictor; Type: Code indicating what file type the predictor is, with possible values "BW" (`BigWig`), "GR" (genomic ranges object), "GRL" (genomic ranges list), "GC" (GC content), "ST" (variant effect score); Range: window in base pair around the query position considered for the predictor; Measure: indicator how the predictor should be read out, with options "distance" (distance of query position to closest feature in the `GR` object in bps), "ifany" (binary indicator whether the query positions overlaps a feature of the predictor), "mean" (mean predictor score in the window around the position, with regions not covered in the `GR` or `BigWig` being ignored), "mean0" (same as mean, only non-covered regions are counted as 0 towards the mean), "nHits" (number of `GR` features within a window around the query position); Tissue-specific: indicating whether this predictor uses tissue-specific data; Transform: indicator whether the predictor should be transformed by taking the square root ("sqrt"), the logarithm ("log"), or used as-is ("NA").

Name	Type	Range	Measure	Tissue-specific	Transform
GCcontent	GC	10bp-100kbp	N/A	no	N/A
Distance to centromere	GR	N/A	distance	no	sqrt
Distance to telomere	GR	N/A	distance	no	sqrt
Replication Valleys	GR	100bp	ifany	yes	N/A
Replication Peaks	GR	100bp	ifany	yes	N/A
Replication WaveSignal	GR	N/A	mean	yes	N/A
Hi-C PCA compartments	GRL	N/A	mean0	yes	N/A
Hi-C interactions	GRL	N/A	mean0	yes	N/A
DNAse-seq signal	BW	1bp-1Mbp	mean	yes	log
DNAse-seq peaks	GR	N/A	ifany	yes	N/A
ATAC-seq signal	BW	1bp-1Mbp	mean	yes	log
ATAC-seq peaks	GR	N/A	ifany	yes	N/A
DNA methylation	GR	100kbp	mean	yes	N/A
CTCF binding signal	BW	1bp-1Mbp	mean	yes	log
CTCF binding peaks	GR	N/A	ifany	yes	N/A
EP300 binding signal	BW	1bp-1Mbp	mean	yes	log
EP300 binding peaks	GR	N/A	ifany	yes	N/A
H3K27ac signal	BW	1bp-1Mbp	mean	yes	log
H3K27ac peaks	GR	N/A	ifany	yes	N/A
H3K27me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K27me3 peaks	GR	N/A	ifany	yes	N/A
H3K36me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K36me3 peaks	GR	N/A	ifany	yes	N/A
H3K4me1 signal	BW	1bp-1Mbp	mean	yes	log
H3K4me1 peaks	GR	N/A	ifany	yes	N/A
H3K4me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K4me3 peaks	GR	N/A	ifany	yes	N/A
H3K9ac signal	BW	1bp-1Mbp	mean	yes	log
H3K9ac peaks	GR	N/A	ifany	yes	N/A
H3K9me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K9me3 peaks	GR	N/A	ifany	yes	N/A
PolR2A signal	BW	1bp-1Mbp	mean	yes	log
PolR2A peaks	GR	N/A	ifany	yes	N/A
POLR2AphosphoS5 signal	BW	1bp-1Mbp	mean	yes	log
POLR2AphosphoS5 peaks	GR	N/A	ifany	yes	N/A
Cancer expression	GR	N/A	mean	yes	N/A
Normal tissue expression	GR	N/A	mean	yes	N/A
GTEx tissue expression	GR	N/A	mean	yes	N/A
Transcription Factor Binding Site Density	GR	10kbp-1Mbp	nHits	no	log
Transcription Factor Binding Site	GR	N/A	ifany	no	N/A
ETS Transcription Factor Binding Site Density	GR	10kbp-1Mbp	nHits	no	log
ETS Transcription Factor Binding Site	GR	N/A	ifany	no	N/A
GTEx eQTL	GR	100bp	ifany	yes	N/A
GTEx eQTL -log10(p-value)	GR	1kbp-1Mbp	mean0	yes	log
GTEx eQTL slope	GR	1kbp-1Mbp	mean0	yes	log

Table S13: (continued)

Name	Type	Range	Measure	Tissue-specific	Transform
Conservation phyloP100way	BW	N/A	mean	no	N/A
NonB-DNA a-phased repeats	GR	100bp	ifany	no	N/A
NonB-DNA direct repeats	GR	100bp	ifany	no	N/A
NonB-DNA g-quadruplex-forming repeats	GR	100bp	ifany	no	N/A
NonB-DNA inverted repeats	GR	100bp	ifany	no	N/A
NonB-DNA mirror repeats	GR	100bp	ifany	no	N/A
NonB-DNA short tandem repeats	GR	100bp	ifany	no	N/A
NonB-DNA zDNA motifs	GR	100bp	ifany	no	N/A
Coding effect score	ST	N/A	N/A	no	N/A

Table S14: Whole genome predictor mapping configuration file. Table with configurations used for the mapping of predictors to query genomic positions. The contents of the columns are as follows. Name: Name of the predictor; Type: Code indicating what file type the predictor is, with possible values "BW" (`BigWig`), "GR" (genomic ranges object), "GRL" (genomic ranges list), "GC" (GC content), "ST" (variant effect score); Range: window in base pair around the query position considered for the predictor; Measure: indicator how the predictor should be read out, with options "distance" (distance of query position to closest feature in the `GR` object in bps), "ifany" (binary indicator whether the query positions overlaps a feature of the predictor), "mean" (mean predictor score in the window around the position, with regions not covered in the `GR` or `BigWig` being ignored), "mean0" (same as mean, only non-covered regions are counted as 0 towards the mean), "nHits" (number of `GR` features within a window around the query position); Tissue-specific: indicating whether this predictor uses tissue-specific data; Transform: indicator whether the predictor should be transformed by taking the square root ("sqrt"), the logarithm ("log"), or used as-is ("NA").

Name	Type	Range	Measure	Tissue-specific	Transform
GCcontent	GC	10bp-100kbp	N/A	no	N/A
Distance to centromere	GR	N/A	distance	no	sqrt
Distance to telomere	GR	N/A	distance	no	sqrt
Replication Valleys	GR	100bp	ifany	yes	N/A
Replication Peaks	GR	100bp	ifany	yes	N/A
Replication WaveSignal	GR	N/A	mean	yes	N/A
Hi-C PCA compartments	GRL	N/A	mean0	yes	N/A
Hi-C interactions	GRL	N/A	mean0	yes	N/A
DNAse-seq signal tissue	BW	1bp-1Mbp	mean	yes	log
DNAse-seq peaks tissue	GR	N/A	ifany	yes	N/A
ATAC-seq signal tissue	BW	1bp-1Mbp	mean	yes	log
ATAC-seq peaks tissue	GR	N/A	ifany	yes	N/A
DNA methylation	GR	10kbp-10kbp	mean	yes	N/A
CTCF binding signal	BW	1bp-1Mbp	mean	yes	log
CTCF binding peaks	GR	N/A	ifany	yes	N/A
EP300 binding signal	BW	1bp-1Mbp	mean	yes	log
EP300 binding peaks	GR	N/A	ifany	yes	N/A
H3K27ac signal	BW	1bp-1Mbp	mean	yes	log
H3K27ac peaks	GR	N/A	ifany	yes	N/A
H3K27me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K27me3 peaks	GR	N/A	ifany	yes	N/A
H3K36me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K36me3 peaks	GR	N/A	ifany	yes	N/A
H3K4me1 signal	BW	1bp-1Mbp	mean	yes	log
H3K4me1 peaks	GR	N/A	ifany	yes	N/A
H3K4me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K4me3 peaks	GR	N/A	ifany	yes	N/A
H3K9ac signal	BW	1bp-1Mbp	mean	yes	log
H3K9ac peaks	GR	N/A	ifany	yes	N/A
H3K9me3 signal	BW	1bp-1Mbp	mean	yes	log
H3K9me3 peaks	GR	N/A	ifany	yes	N/A
PolR2A signal	BW	1bp-1Mbp	mean	yes	log
PolR2A peaks	GR	N/A	ifany	yes	N/A
POLR2AphosphoS5 signal	BW	1bp-1Mbp	mean	yes	log
POLR2AphosphoS5 peaks	GR	N/A	ifany	yes	N/A
Transcript	GR	N/A	ifany	N/A	N/A
Coding	GR	N/A	ifany	N/A	N/A
Cancer expression 1bp	GR	N/A	mean0	yes	N/A
Normal tissue expression 1bp	GR	N/A	mean0	yes	N/A
GTEx tissue expression 1bp	GR	N/A	mean0	yes	N/A
Cancer expression	GR	1kbp-1Mbp	mean	yes	N/A
Normal tissue expression	GR	1kbp-1Mbp	mean	yes	N/A
GTEx tissue expression	GR	1kbp-1Mbp	mean	yes	N/A
Transcription Factor Binding Site Density	GR	10kbp-1Mbp	nHits	no	log
Transcription Factor Binding Site	GR	N/A	ifany	no	N/A

Table S14: (continued)

Name	Type	Range	Measure	Tissue-specific	Transform
ETS Transcription Factor Binding Site Density	GR	10kbp-1Mbp	nHits	no	log
ETS Transcription Factor Binding Site	GR	N/A	ifany	no	N/A
GTEx eQTL	GR	100bp	ifany	yes	N/A
GTEx eQTL -log10(p-value)	GR	1kbp-100kb	mean0	yes	log
GTEx eQTL slope	GR	1kbp-100kbp	mean0	yes	log
Conservation phyloP100way	BW	N/A	mean	no	N/A
NonB-DNA a-phased repeats	GR	100bp	ifany	no	N/A
NonB-DNA direct repeats	GR	100bp	ifany	no	N/A
NonB-DNA g-quadruplex-forming repeats	GR	100bp	ifany	no	N/A
NonB-DNA inverted repeats	GR	100bp	ifany	no	N/A
NonB-DNA mirror repeats	GR	100bp	ifany	no	N/A
NonB-DNA short tandem repeats	GR	100bp	ifany	no	N/A
NonB-DNA zDNA motifs	GR	100bp	ifany	no	N/A

List of Figures

List of Tables

Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder singemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

Köln, 13.01.2025

Corinna Lewis Schmalohr