University of Zagreb

Faculty of Science

Department of Biology

Paula Štancl

Influence of genomic characteristics on the optimization of models for tumor cell-of-origin prediction

DOCTORAL THESIS

Zagreb, 2024

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

Paula Štancl

Utjecaj genomskih karakteristika na optimizaciju modela za predviđanje ishodišne stanice tumora

DOKTORSKI RAD

Zagreb, 2024

This doctoral thesis was conducted within the Bioinformatics Group, under the supervision of Rosa Karlić, PhD, associate professor, as part of the project HRZZ-IP-2019-04-9308, A statistical modeling approach to predict the cell-of-origin and investigate mechanisms of cancer development (PREDI-COO). The research was completed as part of the Postgraduate Doctoral Programme in Biology at the Department of Biology, Faculty of Science, University of Zagreb.

Information about the mentor

Rosa Karlić is an associate professor in the Bioinformatics Group at the Department of Molecular Biology, Biology Department, Faculty of Science.

She is engaged in multiple projects that aim to relate chromatin structure and other epigenetic mechanisms to transcriptional regulation, evolution of regulatory elements and the development of complex diseases, especially cancer. Her focus lies in utilizing statistical and machine learning methods to analyze high-throughput genomic data in these contexts. She published her research in multiple Nature papers such as the developed models for prediction of cell-of-origin in multiple cancers. Moreover, she was working as a part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) which resulted in cutting-edge research and a valuable cancer database resource to study cancer genomics and epigenomics. In 2011, she was awarded with L'Oréal-UNESCO “Za žene u znanosti” stipend for her research about histone modification levels being predictive for gene expression. She is the leader of HRZZ-IP-2019-04-9308, A statistical modeling approach to predict the cell-of-origin and investigate mechanisms of cancer development (PREDI-COO). At the Faculty she teaches 9 graduate-level courses and 1 post-graduate course: Biostatistics. She mentored 2 Bachelor’s and 6 Master’s students. Among other teaching activities, she was involved in creating online courses as a part of Erasmus+ project PROMISE (Personalized Medicine Inquiry-Based Education).

She published over 40 scientific papers in international scientific journals with high-impact factors that were cited over 10000 times.

I would like to extend my heartfelt gratitude to my mentor Rosa for her invaluable guidance and unwavering support throughout my journey during my PhD thesis. Especially for the support you have shown for all my scientific, educational, and population science activities during my PhD. Your insightful advice, positive attitude, and patience have truly made a difference in my experience of PhD.

Special thanks goes to professor Kristian Vlahoviček without whom I would not started my bioinformatic journey.

This whole journey would not be possible without amazing colleagues; Maja, Dunja, and Filip. Thank you for everything you have taught me through the years and for all the advice and help when I need it. You are simply the best bioinfo people out there whose advice I will always cherish. Thanks to my collaborator Paz Polak for showing me the amazing world of cancer genomics.

I am deeply grateful for all the coffee sessions with my friends whose support has lifted my spirit during the PhD journey.

One special gratitude goes to my family, mother Andreja and brother Juraj, for their unwavering love, support, and understanding throughout this journey. I am immensely grateful for your constant presence, sacrifice, and encouragement even in the darkest times when nothing was working and constantly crashing. Also, to my beloved Kala and Oscar for always being there.

Lastly, I thank Sven for his support and for pushing me to be a better person in life and career. You make my everyday shine even brighter.

University of Zagreb Doctoral thesis
Faculty of Science
Department of Biology

INFLUENCE OF GENOMIC CHARACTERISTICS ON THE OPTIMIZATION OF MODELS FOR TUMOR CELL-OF-ORIGIN PREDICTION

PAULA ŠTANCL
Bioinformatics group, Department of Biology, Faculty of Science

Identifying the cell-of-origin (COO) for tumors of unknown primary site is crucial for selecting effective therapies, but the heterogeneity and unique genomic profiles of these cancers make accurate prediction by machine learning models challenging. In this study, I analyzed how various genomic features influence COO prediction by using different machine learning models based on the mutational landscapes and epigenetic characteristics of normal tissues in distinct genomic features. The analysis of breast, liver, and skin melanoma cancers revealed significant differences of their mutational landscapes, and their influence on COO model accuracy. APOBEC mutations in early-replicating regions associated with kataegis were linked to poorer COO predictions in breast cancer. Including protein-coding genes and topologically associated domains can also be used for reliable COO predictions and identification of under-predicted genes with potential roles in tumorigenesis. While whole-exome sequencing and RNA-seq mutations aid COO identification, they are less accurate than whole-genome sequencing. Overall, the tumor's complex genomic landscape significantly affects COO prediction accuracy.

(224 pages, 106 figures, 11 tables, 246 references, original in English)
Keywords: cancer genomics, mutational signatures, cell-of-origin, machine learning

Supervisor: Rosa Karlić, PhD, Associate professor

Reviewers: Kristian Vlahoviček, PhD, Full professor tenure
Inga Urlić, PhD, Full professor
Petar Ozretić, PhD, Senior research associate

Sveučilište u Zagrebu Doktorska disertacija
Prirodoslovno-matematički fakultet
Biološki odsjek

UTJECAJ GENOMSKIH KARAKTERISTIKA NA OPTIMIZACIJU MODELA ZA PREDVIĐANJE ISHODIŠNE STANICE TUMORA

PAULA ŠTANCL
Grupa za bioinformatiku, Biološki odsjek, Prirodoslovno-matematički fakultet

Određivanje ishodišne stanice tumora (engl. *cell-of-origin*, COO) ključno je za odabir učinkovitih terapija. Međutim, heterogenost i jedinstveni genomski profili ovih tumora čine njihovu preciznu identifikaciju pomoću modela strojnog učenja izazovnom. U ovom radu analizirala sam kako različite genomske značajke utječu na predviđanje COO koristeći različite modele strojnog učenja temeljene na mutacijskim krajolicima i epigenetskim karakteristikama normalnih tkiva. Analizom tumora dojke, jetre i melanoma kože otkrivene su značajne razlike u njihovim mutacijskim krajolicima te njihov utjecaj na točnost modela za predviđanje COO. APOBEC mutacije koje se pojavljuju u rano replicirajućim regijama i povezane sa žarištima mutacija, pokazale su lošiju točnost predviđanja COO kod tumora dojke. Geni za kodiranje proteina i topološki povezane domene također se mogu koristiti za pouzdana predviđanja COO, kao i za identifikaciju gena s premalim brojem predviđenih mutacija, koji imaju potencijalnu ulogu u tumorigenezi. Iako mutacije dobivene sekvenciranjem cijelog egzoma i transkriptoma pomažu u identifikaciji COO, one su manje točne od mutacija dobivenih sekvenciranjem cijelog genoma. Sveukupni rezultati ukazuju na to da kompleksni genomski krajolik tumora značajno utječe na točnost predviđanja COO.

(224 stranica, 106 slike, 11 tablica, 246 literaturnih navoda, jezik izvornika: engleski)
Keywords: genomika tumora, mutacijski potpisi, ishodišna stanica tumora, strojno učenje

Mentor: dr.sc. Rosa Karlić, izvanredni profesor
Ocjenitelji: dr.sc. Kristian Vlahoviček, redoviti profesor u trajnom zvanju
dr.sc. Inga Urlić, PhD, redoviti profesor
dr.sc. Petar Ozretić, viši znanstveni suradnik

1 Introduction

The latest worldwide cancer statistics from 2022. recorded close to 20 million new cases of cancer with almost 9.7 million deaths from cancer (Bray et al., 2024). The number of deaths is expected to rise to 16.3 million by 2040, while the number of new cases of cancer will increase by 28 million each year by 2040 (International Agency for Research on Cancer, 2024). Late diagnosis and lower standards of living and healthcare in low-income developing countries as well as more stressful and lower quality sedentary lifestyles with poor diet and infrequent exercise are certainly major contributors to the increase in the incidence of cancer (Beddoe et al., 2016; Friedenreich et al., 2021; Pisani, 2011). The most prevalent and fatal cancers among women include breast and cervical cancers, whereas among men, prostate cancer is the most common, followed by lung, liver, colorectal, and stomach cancers. The deadliest cancers in men are lung, prostate, and liver cancers, according to recent epidemiological studies (Bray et al., 2024).

One particular dangerous cancer type is the one whose primary origin is not known and those cancers are referred to as cancers of unknown primary origin tissues (CUPs). CUPs are especially dangerous and pose a challenge to identify since they lack certain morphological, histopathological and molecular characteristics of cancer with known primary origins. Patients with CUPs account for ~3-5% of all cancer diagnoses and they significantly suffer from the lack of therapeutic options as primary cancer type classification is still an important factor in directing treatment choices (Beauchamp et al., 2023; Greco & Hainsworth, 1992; Pavlidis et al., 2003; Pavlidis & Fizazi, 2009). Large-scale research projects, such as International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), generate vast datasets comprising of genomic, epigenomic, and transcriptomic profiles from numerous patients across various cancer types (Hudson et al., 2010; Weinstein et al., 2013). Their aim is to get better understanding of the underlying oncogenic mechanisms driving tumor development, and progression. Due to their enormous efforts, various machine learning methods have been developed to predict the cell-of-origin (COO) of cancers using genomes (Liu et al., 2020; Nguyen et al., 2022; Polak et al., 2015; S. Yang et al., 2023), exomes (Li & Luo, 2020), transcriptomes (Divate et al., 2022; Wei et al., 2014; Zhao et al., 2020) or methylome profiles (Zheng & Xu, 2020). Although various methods exist for predicting cancer origin of unknown primary (CUP), not all perform uniformly across cancer types, often relying on whole-genome sequencing (WGS) data for high accuracy. COO predicting models developed with more accessible technologies like whole-exome and RNA-sequencing tend to have lower COO prediction accuracy compared to WGS. This has led to an increase of developed machine learning models using WGS data compared to other technologies. Despite these advancements, the precise reasons for COO prediction failures, particularly for individual patients, remain poorly understood. This may be attributed to the diverse genetic mutational landscapes observed in different cancers, reflecting the complex interplay of genetic and environmental factors in tumorigenesis. For instance, the COO predicting model based on predicting the mutational density along the cancer genome from the profile of epigenomic modifications in normal cell type developed by Polak et al. (2015) showed that patients with lower number of mutations in general could not have their COO reliably detected. However, this was not the case for breast cancer patients which showed lower accuracy of the COO model despite the higher number of mutations, which opens up the question of whether the origin of mutations can have an overall impact on the prediction of COO. Origins of mutations can be determined by analyzing mutational signatures, which represent unique mutational patterns of cancer genomes due to mutational processes of different aetiologies being active during the course of cancer development (Alexandrov et al., 2020). Nguyen et al. (2022) showed that in some cancer types mutational signatures, such as SBS4 associated with smoking in lung cancer, are important for correct identification of COO. In breast cancer and some patients with non-small lung cell carcinoma, they detected apolipoprotein B mRNA-editing enzyme (APOBEC) mutational signatures as one of the most important features for predicting COO. APOBEC signatures, extremely commonly found in breast cancer, are associated with hotspots that are enriched for regulatory elements, coding elements, transcription factors binding sites and known tumor drivers (Wong et al., 2022). However, a more thorough investigation of the origin of mutations, based on mutational signature analysis in aggregated cancer profiles and individual cancer patients, is required to improve the existing models. While mentioned COO models utilize mutational counts per 1 Mb genomic regions, other genomic features such as topologically associated domains (TADs) and even genes have not been thoroughly investigated for COO predictions. TADs are fundamental units of chromatin organization in the genome. They represent functional architecture features that supports various genomic processes such as transcription, replication, and DNA repair unlike fixed 1 Mb genomic regions. Additionally, the impact of structural variants (SVs) on COO models have not been fully explored, despite their known importance in certain COO predictions of certain cancer types (Nguyen et al., 2022). SVs can lead to alterations in copy number or deletions within coding sequences, as well as disturbances in structural chromatin elements, particularly topologically associated domains (TADs). These disruptions have the potential to impact the precision of cell-of-origin prediction models.

1.1 Objectives

In this thesis I will perform an exploratory analysis of retrieved and annotated genetic variants derived from whole-genome, whole-exome, and RNA sequencing data obtained from melanoma, breast, and liver cancer samples. I selected those cancer types due to their ability to form secondary tissues, metastasis, whose cell-of-origin might be difficult to predict, as well as their distinct genomic features. I aim to discern those genomic features: mutational signatures, kataegis patterns, and structural variant (SV) hotspots within the genomes of these cancer types. To gain deeper insights how structural variants are associated with cell-of-origin, the identified variants, signatures, kataegis events, and SV hotspots features will be correlated with epigenomic attributes of normal cells which represent the cells-of-origin for these cancers. Subsequently, I will focus on the development and comparative assessment of diverse machine learning models for predicting the cell-of-origin of these cancers. Cell-of-original models will be extension of methods developed by Polak and colleges (2015). These models will be constructed using the identified tumor mutations and epigenomic characteristics derived from normal tissues in 1 Mb genomic regions, topologically associated domains and genes. Additionally, I will investigate the impact of various genomic features, as well as advancer machine learning methods, on the performance of cell-of-origin prediction models when applied to aggregated cancer profiles as well as individual cancer patients.

2 Literature overview

2.1 Sequencing technologies in clinical oncology

From 1980 there has been much research shedding the light on the importance of genes in cancer development. The majority of those studies focused on viral transforming genes or oncogenes, explaining and unraveling their role in cancer one at a time. Two types of viral transforming genes or oncogenes were identified: those that immortalize cells, and those that make them tumorigenic (Land et al., 1983). However, there was still a major knowledge gap in cancer research regarding the connection between oncogene activity and the cancer progression. Dulbecco (Dulbecco, 1986) highlighted this problem way back in 1983 and realized the value of having the whole genome sequence of a species as the foundation for studying cancers. This was accompanied with one of the biggest achievements in human history the Human Genome Project back in 1990. which took 13 years and 2.7 billion dollars to produce the first assembly of a human genome (Lander et al., 2001; National Human Genome Research Institute, 2024; Venter et al., 2001). The first human genome allowed us to begin understanding human biology, particularly diseases, shedding new light on the role of genes in cancer development and progression. The sequencing technologies have come a long way since then and with each technological advancement we have obtained a better assembled human genome. The reference human genome has become an irreplaceable and highly valuable resource for researchers to understand cancer. Nowadays, there is no need to perform *de novo* sequencing and tedious assembly of the human genome when interested in the profile of cancer patient's genome. Instead, resequencing of the whole genome or certain parts are conducted and compared to the reference genome as a method of detecting mutational changes leading to cancer progression.

Regardless of the sequencing technology being used, the experimental steps of all DNA or RNA sequencing experiments are the same. First, the DNA or RNA is extracted and purified from certain tissue or cells of interest, then fragmented into smaller pieces using ultrasonication or restriction digestion. Afterwards, DNA ends are covalently attached with adapters which serve various roles. Adapters can link the sequences to a flow cell or ensure compatibility with the specific sequencing platform being used, and they incorporate barcodes for sample identification, allowing for both target enrichment and sequencing multiplexing (Qin, 2019). Two prominent approaches of target enrichment are amplification with polymerase chain reaction-based (PCR, or amplicon-based) or direct capture using hybridization capture-based methodologies (Singh, 2022). In RNA sequencing experiments, there is an additional step of translating the isolated RNA to cDNA using reverse transcriptase before or after fragmentation step. In the following chapters I will go through the most used sequencing approaches and technologies in clinics.

2.1.1 First generation sequencing methods

First generation sequencing consists of two major methods: Maxam-Gilbert sequencing and Sanger sequencing. Maxam-Gilbert sequencing is a chemical method based on nucleobase-specific partial chemical modification using dimethyl sulfate and hydrazine followed by the cleavage of the DNA backbone at specific points (Maxam & Gilbert, 1977). The main drawbacks of this method are the speed, cleave reactions, gel electrophoresis and developing the film which limit the number of bases around 200-300 of DNA per day. On the other hand, Sanger sequencing relied on four polymerase chain reactions happening separately. In each reaction there were nucleotides radioactively labeled with chain-terminating dideoxynucleotide (ddNTP). Incorporation of those nucleotides during *in vitro* DNA replication would result in random fragments with varying length. Afterwards, gel electrophoresis would be used to arrange the fragments of varying lengths in each lane in gel so that the DNA sequence can be "read" (Sanger et al., 1977). Although both Walter Gilbert and Frederick Sanger were awarded The Nobel Prize in Chemistry in 1980 alongside Paul Berg for contributing to the determination of base sequences in nucleic acids (NobelPrize.org, 2024), only the Sanger method has stood the test of time. In the updated and automated Sanger method, the radioactively labeled ddNTPs were replaced with fluorescently labeled ddNTPs and capillary electrophoresis was introduced instead of gel electrophoresis (Smith et al., 1986). This significantly reduced the time and technical compilations when sequencing DNA. Even though the Sanger method was first used in the beginning to sequence the first human genome, it was replaced by more high-throughput sequencing approaches to increase time and cost efficiency. Nowadays, the Sanger method is only used for a single target, usually a gene or smaller gene sets of interest, due to its high robustness and precision (Vestergaard et al., 2021). Hence, why it is often referred to as targeted sequencing. Therefore, in clinical settings it is still considered a gold standard for targeted sequencing of clinically relevant germline mutations, specific pathogen identification and drug resistance mutations (Alhamlan et al., 2023; Mercier-Darty et al., 2019; Nicolussi et al., 2019).

2.1.2 Next-generation sequencing

Next-generation sequencing is also referred to as second generation sequencing. The second generation of sequencing has made tradeoffs for a bit lower precision and a shorter read length, typically 75 - 300 base pairs (bp) compared to ~1000 bp by Sanger, in order to sequence multiple DNA molecules in parallel, around hundreds of millions even up to 5 billion, in a single experiment at a time (Kchouk et al., 2017). This allowed for a much faster sequencing of multiple genes in panels used in clinics, while Sangers is still deemed too expensive and time consuming even on a small gene panel. Most commonly used technology of second-generation sequencing is the Illumina technology based on the sequencing-by-synthesis (SBS) principle. In Illumina, DNA fragments are attached to a solid surface and amplified to form clusters of up to 1,000 identical copies of each single fragment molecule in close proximity. This process is known as solid-phase amplification. Afterwards, each cluster undergoes multiple cycles of sequencing where a single fluorescently labeled deoxynucleoside triphosphate (dNTP) is added to the nucleic acid chain. The nucleotide label serves as a terminator for polymerization whose fluorescent dye is imaged to detect the incorporated base. Emitted signals from all incorporated nucleotides in each cluster are averaged during each cycle. After base calling, the dNTP is enzymatically cleaved to allow the next nucleotide to be incorporated. Although the natural competition of all four single, separated reversible terminator-bound dNTPs (A, C, T, G) minimizes incorporation bias, incorporation of wrong base still occurs (Illumina.Inc, 2010). Some of the molecules in clusters are out of phase, known as phasing, because either the dNTP was not removed or the new dNTP was not incorporated in one cycle which leads to the different dNTP incorporated in the next cycle. The incorporation of the correct base at later cycles deteriorates the overall averaged signal quality for later bases in the sequenced fragment. Regardless of this main error contributor in Illumina, it still produces highly accurate base calls with error rates of mostly substitutions typically below 1% (Pfeiffer et al., 2018).

Newer technologies, third-generation sequencing also known as long-read sequencing avoid the PCR-biases generated during the PCR amplification step by directly sequencing the DNA or RNA molecule (Xiao & Zhou, 2020). Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) were the first technologies showing the potential of long-read sequencing technologies in all fields of biology. ONT performs a nanopore-based sequencing (NS) with biological or synthetic nanopores that allow only one molecule to pass at a time. As the molecule passes through the pore, the ionic current is disrupted because different nucleotides have varying resistances to the flow of ions between outer and inner sides of membrane and pore. These fluctuations are read by the device which are base-called using machine-learning approaches (Jain et al., 2016). On other hand, PacBio platforms are based on a “single-molecule real-time” (SMRT) approach by immobilizing the polymerase and detecting incorporation of fluorescently labeled nucleotides (Eid et al., 2009). The most common types of errors generated with long-read sequencing are indels, which include insertions and deletions. These errors contribute to the high sequencing error rate, which is 10-15% in PacBio and 5-20% in ONT, especially in homopolymer regions of the genome (Amarasinghe et al., 2020; Xiao & Zhou, 2020).

Out of the two newer generations of sequencing technologies, Illumina is one which is mostly used in cancer genomics for genomic and transcriptomic characterization of patient’s tumor by whole-genome, whole-exome or RNA sequencing due its ability to sequence multiple genes with high accuracy. On the contrary, long-sequencing reads from ONT or PacBio enable for a reliable detection of SVs or fusion genes in cancer patients due its generation of much longer reads (Depledge et al., 2019; Ho et al., 2020). This was well applied in hematological cancers. Since both technologies have a higher error rate than second-generation sequencing technologies, they needed to be corrected using other more accurate NGS methods to reliably detect mutations in patient's tumors. Therefore, they still have not found its common and broad use in clinical cancer genomics like Illumina sequencing. However, these platforms have proven to be quite exceptional in detection of base modifications in native DNA, such as 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) (Flusberg et al., 2010) and even detection of post-transcriptional RNA modifications (Stephenson et al., 2022).

2.1.2.1 Whole-genome and exome sequencing

Whole-genome sequencing (WGS) enables reading the entire genomes, all of 3 billion bases in the human genome. The re-sequenced genome is then mapped onto a high-quality reference one for identification of mutations. The main advantage of the WGS method is its ability to cover entire human genomes, both coding and non-coding regions, allowing for identifications of novel and rare variants. On the other hand, whole-exome sequencing (WXS) covers only coding regions of the genome. The exome accounts for less than 1 to 2% of the human, but it contains 85% of documented disease-causing variants (Choi et al., 2009). In order to sequence only exomes, WXS requires two main processes; target-enrichment to select and capture exomes and sequencing. Main platforms for exome capture are NimbleGen, Agilent, and Illumina with varying designs and strengths. All three technologies rely on biotinylated oligonucleotide baits complementary to the exome targets to hybridize sequencing libraries prepared from fragmented genomic DNA (Clark et al., 2011). After hybridization, the bound fragments are pulled-down with magnetic streptavidin and sequenced. The differences between technologies are in their target choice, bait characteristics (length and density) and molecules for capture (Clark et al., 2011). NimbleGen and Illumina use DNA, while Agilent uses RNA bait (Warr et al., 2015). Out of the three, NimbleGen required the least amount of reads to sensitively detect small variants since it uses short (55−105 bp) overlapping baits that cover the bases it targets multiple times. Agilent uses longer RNA baits (114−126 bp) and the corresponding target sequences are adjacent to one another rather than overlapping. These longer baits allow for better identification of indels (Chilamakuri et al., 2014; Clark et al., 2011). Lastly, Illumina design of TruSeq Exome Enrichment Kit with 96-probes has the most reduced target efficiency due to high percentage of off-target enrichment. The kit uses paired-ends reads to extend outside the bait sequences and fill in gaps. It is the only one to be able to detect mutations in untranslated regions (UTRs) allowing researchers to explore those regions. All three exome enrichment technologies have different target regions/choices that are taken and combined from available reference databases of coding and non-coding RNA genes (RefSeq, UCSC KnownGenes and Ensembl) (Clark et al., 2011; Flicek et al., 2011; Hsu et al., 2006; Pruitt et al., 2009). Different numbers of transcripts, as well as the start and end positions of some transcripts, differ between the technologies. One study found that there is very little overlap between three technologies when targeting human exome, with just 26.2 Mb covered by all three target regions (Chilamakuri et al., 2014). Newer comparison of updated versions of mentioned technologies; Agilent and NimbleGen "+UTR" kits and Illumina with Nextera Expanded Exome; revealed that both updated Agilent and NimbleGen outperformed Illumina (Meienberg et al., 2015). Furthermore, Agilent managed to capture overall more coding exons with sufficient read depth than other two technologies. However, despite their strengths, all three platforms exhibit significant gaps in effective exome coverage, suggesting the need for further improvements. It is worth mentioning that these drawbacks are generally reduced over time as exome kit design and standardization of variant calling procedures are improved. Barbitoff et al. (2020) showed that WGS has more coverage and is more efficient by only 1-2% when compared to the best newer WXS platforms. They suggested that only a small fraction of already annotated pathogenic variants found in the ClinVar database (Landrum et al., 2014) are not targeted by WXS. Alongside its reduced cost compared to WGS and smaller data generation, WXS has therefore become a staple tool in clinical study in recent years. Both WGS and WXS are improving the medical healthcare and clinical oncology field by identifying new variants leading to tumor progression and proliferation.

2.2 Cancer genomics

Even though the word cancer (greek *carcinos* and *carcinoma*) inspired Greek physician Hippocrates to describe cancerous growths after a moving crab somewhere between 460-375 BC, the first written evidence describing what we know as cancer is much older (Hajdu, 2011). It is called the Edwin Smith Papyrus and it was written about 3000 BC ago in Egypt and brought by Smith in 1862. It was the first description of breast cancer as bulging tissue. One of the most striking parts of the writing states: "There is no treatment.", highlighting the graveness of the disease (Breasted, 1930). Nowadays, we are not only challenging cancers by creating more efficient treatments but also questioning the definition of cancer itself. There is no question that cancer is an aggressive and ever-evolving complex heterogenous disease that cannot be solely described based on its look and function. Therefore, the most recent proposed definition of cancer goes: “Cancer is a disease of uncontrolled proliferation by transformed cells subject to evolution by natural selection.” (Brown et al., 2023).

In general, we refer to the abnormal growth of cells or tissues in the body to as neoplasm (National Cancer Institute, 2024). Neoplasm can broadly be classified as benign tumors that stay in the primary site and not spreading through the body and as malignant (cancers) that divide uncontrollably and spread to nearby or distant sites in the body (Patel, 2020). To get a deeper understanding into the biology of cancer, Hanahan and Weinberg have provided us through the years with 14 structured functional pathways on how normal cells progressively transform into neoplasm and specifically malignant ones (Hanahan, 2022; Hanahan & Weinberg, 2000, 2011). Those pathways are known as The Hallmarks of Cancer and are continually updated as new research sheds light on tumor pathogenesis. The core hallmarks described in 2011. consist of the following biological capabilities for sustaining proliferative signaling, evading growth suppressors, avoiding immune destruction, replicative immortality, activating invasion and metastasis, resisting cell death, enabling inducing/accessing vasculature, reprogramming cellular metabolism, and two enabling characteristics; tumor-promoting inflammation and genome instability and mutation. The newly proposed hallmarks consist of two new emerging hallmarks; unlocking phenotypic plasticity and senescent cells; and two new enabling characteristics; nonmutational epigenetic reprogramming and polymorphic microbiomes. Enabling characteristics are responsible for the aberrant condition of neoplasm and enable the cancers cells to acquire all of the hallmark capabilities (Hanahan, 2022).

2.2.1 Cancer mutations

Out of all hallmarks, the most prominent one in development of cancer is the enabling characteristic of genomic instability which is facilitated by acquirement of mutation; from large chromosomal rearrangement to point mutations or indels (insertions or deletions). Mutations pose a significant risk as they can disrupt the normal function of a gene's protein product by altering its amino acid sequence or producing a stop codon leading to non-functional shorter protein. We distinguish two types of mutations; germline and somatic. While germline mutations are heritable that occur in germ cells, somatic mutations occur in somatic cells of tissues and cannot be inherited by offspring (Meyerson et al., 2020). Based on the “Knudson two hit hypothesis” model, dominantly inherited predisposition to cancer involves an initial germline mutation, while carcinogenesis or tumorigenesis requires a subsequent somatic mutation to occur. Similarly, non-hereditary cancer of the same type requires two somatic mutations in both alleles of a gene (Knudson & Knudson, 1996). Although this is generally applicable to majority of tumor suppressor genes, unlike oncogenes, there are always exceptions. For instance, some well-known tumor-suppressor genes, like *CDKN1B, TP53, DMP1*, *NF1*, and *PTEN*, have shown that when one copy of a gene is inactivated or deleted, the remaining functional copy of the gene fails to produce sufficient gene product to maintain normal function, known as haploinsufficiency (Inoue & Fry, 2017; National Cancer Institute, 2024). Out of the mentioned ones, the *TP53* gene was given the title "guardian of the genome" as it acts as an important G1 DNA check inhibitor preventing genomic instability that leads to tumor development (Lane, 1992). Among other function besides inhibiting cell-cycle, p53 induces DNA repair, apoptosis, autophagy, promotes senescence and many more to prevent tumorigenesis (Aylon & Oren, 2011; Kastan et al., 1995). The primary *TP53* mutations is the loss-of-wild-type p53 functions as a main driving force in prevention of cancer development. However, unlike majority of other tumor-suppressor genes, the majority of *TP53* mutations are predominantly missense mutations which cause single amino acid substitutions (Mantovani et al., 2019). It is commonly mutated in more than 50% of all human cancers (Hainaut & Pfeifer, 2016; Kandoth et al., 2013) making it up to be one of the most studied tumor suppressor genes in cancer field. Therefore, to effectively treat cancer and devise drugs tailored to target mutated gene products for each cancer's patient, it is important to identify the most prevalent alterations in genes, such as *TP53*, driving tumorigenesis.

2.2.1.1 Variant calling of different types of mutations

Most common form of mutations driving tumorigenesis are divided into three classes of genomic variations: single nucleotide variants (SNV) or single-base substitutions (SBS), short insertions or deletions, and large SVs. SNVs can be further classified as synonymous, non-synonymous and nonsense variants depending how the variants affect the biological function of the protein. While indels represent mutations affecting shorter parts of the genome usually between 1 and 50 bp, SVs are larger in size (typically above 1 kb) and include large deletions, tandem duplications, insertions, inversions and translocations (Feuk et al., 2006; Nesta et al., 2021; Tate et al., 2019). Large scale changes that result in different copy number of those larger regions either as deletions, duplications or even insertions are called copy number variants (CNVs). Deletions, inversions, and tandem duplications are categorized based on size into five ranges: 1-10Kb, 10-100Kb, 100Kb-1Mb, 1Mb-10Mb, and events exceeding 10Mb. Additionally, SVs are classified as clustered or non-clustered, determined by the distance between adjacent SVs, resulting in a total of 32 SV types (Tate et al., 2019). The most comprehensive information about somatic variants in human cancers responsible for tumorigenesis is found in an expert-curated database "Catalogue Of Somatic Mutations In Cancer" (COSMIC) (Tate et al., 2019). As already mentioned, not all NGS technologies are suitable for detection of every mutation type. Table 1. has summarized general information about the certain NGS strategy for decent or reliable detection of variants. Whole-genome sequencing is the only one able to reliably call all mutations types albeit by higher cost and lower coverage compared to exome and gene panel sequencing.

Table 1. *General characteristics of NGS technologies for detecting variants. Relative cost is represented by amount of dollar signs. The cost of gene panel can vary depending on the panel size. The empirical performance of each strategy for detecting variants of different classes is indicated as good (+), outstanding (++), or poor/absent (-). SNV - single nucleotide variant. CNV – copy number variant. SV – structural variant. Taken and adjusted from (Koboldt, 2020)*

Strategy	**Gene panel**	**Exome**	**Genome**
Size of target space	~0.5	~50	~3200
Average read depth	500-100X	100-150X	~30-60X
Relative cost	$	$$	$$$
SNV/indel detection	++	++	++
CNV detection	+	+	++
SV detection	-	-	++

The main goal of tumor sequencing in clinics is to identify clinically relevant mutations, that are somatic mutations despite the fact that ~10% of cancer patients harbor germline predisposition variant (Koboldt, 2020). The somatic mutation calling can be done either with “tumor-only” or “paired tumor-normal” mode. In paired tumor-normal approach, the DNA is isolated from both tumorigenic tissue and non-malignant “normal” cells. Use of non-malignant tissue enables the detection of truly somatic mutations by removing those germlines present in a matched normal tissue from the individual. While in tumor-only sequencing, it is impossible to distinguish germline from somatic ones. Non-malignant cells can either be peripheral blood, saliva, buccal swab, fibroblasts or even nail (Mandelker & Ceyhan-Birsoy, 2020), but the most commonly used matched normal sample is adjacent non-tumorigenic tissue or normal adjacent to tumor (NAT). Using NATs as a control allows for comparison between samples from the same individual reduces both individual-specific and anatomical site-specific effects. However, studies have shown that NAT has distinct characteristics from both health and tumor tissue, potentially leading to the omission or misinterpretation of new biomarkers and therapeutic targets when used as normal tissues in “paired tumor-normal” variant calling (Aran et al., 2017; Oh & Lee, 2023). Therefore, one of the main factors affecting the sensitivity and specificity of somatic mutation calling methods is the presence of normal control tissues. Other factors include the depth of sequence coverage in both the tumor and a matched normal sample, the local sequencing error rate of used NGS technology, the low allelic fraction of the mutation, and the evidence thresholds employed to identify a mutation (Cibulskis et al., 2013).

2.2.1.1.1 Somatic single-nucleotide variant callers

Nowadays the gold standard for calling mutations either from DNA or RNA-sequencing is by following the Gene Analysis Toolkit (GATK) Best Practices guides (McKenna et al., 2010; Van der Auwera et al., 2013). The first step in best practice guides, after obtaining reads from WGS, WXS or RNA-sequencing technologies, is to map the reads onto a reference human genome. Since short-read technology sequencing with Illumina is the most used in detection of mutations, the shorts read have to be mapped onto the genome using short-read mappers such as BWA (Li & Durbin, 2009) or Bowtie2 (Langmead & Salzberg, 2012). To obtain reliable mapped reads onto genome to call variants from RNA-seq reads, STAR mapper in 2-pass mode is recommended to use (Dobin et al., 2012). Afterwards, the duplicates are marked in generated mapping output files in SAM or BAM format. Duplicates represent errors occurring during library preparation or sequencing, such as PCR duplicates, and by marking them the GATK tools will ignore them in downstream analysis of variant calling. In order to reliably detected indels and SNVs, local realignment must be performed around indels due to artifacts introduced by mapping. For instance, reads aligning to the edges of indels frequently exhibit mismatches with bases that may appear suggestive of SNVs, but are, in fact, mapping artifacts (Van der Auwera et al., 2013). The last pre-processing step prior to variant calling is base quality score recalibration. In this step machine learning models are used to correct for systematic error introduced by the sequencing machines that tend to leave over- or under-estimated base quality scores. In RNA-seq variant calling it is important to split the reads that may span over exon-exon junction or harbor splice junctions for correct variant calling.

After the data preprocessing, variants are identified using a variant calling tool of choice. Common tools include Mutect2, Strelka2, and VarScan2 for somatic variants, and HaplotypeCaller and FreeBayes for germline variants. These tools follow similar steps but may differ in statistical methods and thresholds for filtering low-quality reads or variants. They also implement approaches to reduce false-positive results by removing sequencing artifacts and other means. For example, the original Mutect had the following four steps: 1) removal of low-quality reads based on coverage, mapping quality and others factors; 2) calling variants using Bayesian classifier; 3) removal of false positive variants based on correlations with sequencing artifacts that are not present in the error model; 4) distinguishing between somatic and germline variant using second Bayesian classifier (Cibulskis et al., 2013). More detailed descriptions of Mutect variant calling steps are shown on Figure 1. The improved version of Mutect, called Mutect2, uses a different a Bayesian somatic genotyping model alongside the assembly-based machinery of HaplotypeCaller (Benjamin et al., 2019).

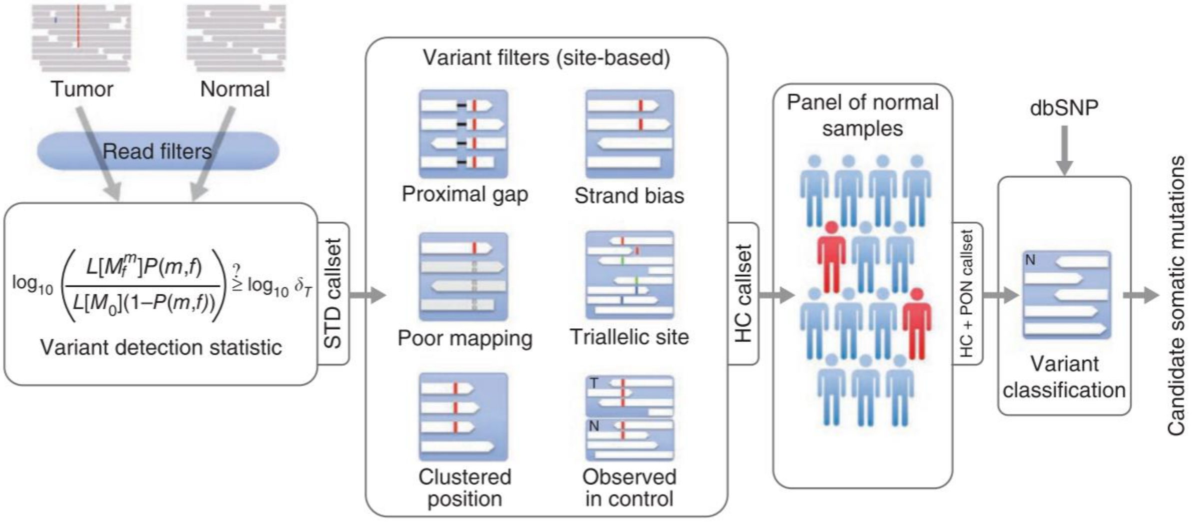


Figure 1. *The MuTect pipeline is employed for the identification of single nucleotide substitution mutations. Initially, next-generation sequencing (NGS) reads from tumor-normal paired samples are aligned and processed through MuTect. This involves the removal of low-quality reads and an assessment to discern variants from expected random sequencing errors. Subsequently, identified variants undergo filtration through six specific filters to eliminate artifacts. Following this, remaining false positives are further filtered using a panel of normal samples (PON). Leveraging matched normal samples, the status of variants as somatic or germline is determined. The "STD" stands for the standard, where no additional filters are applied post-variant calling, whereas "HC" signifies high-confidence, indicating the application of all six filters. Taken and adjusted from (Cibulskis et al., 2013)*

Alongside Mutect2, Strelka2 is the other tool that showed similar performance of variant calling (high precision rate) compared to the rest of tools (Cai et al., 2016; Chen et al., 2020; Kim et al., 2018). Strelka2 is much faster implementation of variant calling compared to Mutect2, but it has less flexibility since it cannot call variants in “tumor-only” mode. It is as well an upgrade from its older version, Strelka (Saunders et al., 2012), which utilizes a Bayesian approach. In Strelka, tumor and normal allele frequencies are considered as continuous variables. The normal sample is depicted as a mixture of diploid germline variation with noise, while the tumor sample is characterized as a mixture of the normal sample with somatic variation (Saunders et al., 2012). Moreover, it considers potential tumor cell contamination in the normal sample in somatic variant calling, thus leading to a significant improvement in variant analysis, particularly for liquid and late-stage tumors. Lastly, to obtain even greater precision, a final empirical variant re-scoring step using a random forest model trained on diverse call quality metrics is employed as the concluding step in Strelka2. Important features of the model are: 1) the genotype probability computed by the core variant probability model, 2) root-mean-square mapping quality, 3) strand bias, 4) the fraction of reads consistent with locus haplotype model, and 5) the complexity of the genome reference context characterized by homopolymer length and compressibility (S. Kim et al., 2018). Despite the improvement made in the algorithms for somatic variants, there is no perfect tool that can detect 100% of what we consider to be true mutations while accounting for all of the occurring biases. Therefore, it is recommended to use an ensemble approach by taking the overlap of mutations called by at least two or more mutations variant tools.

2.2.1.1.2 Structural variants and copy number variants callers

A gold standard for detecting structural variations in clinical settings is non-next-generation sequencing approaches, such as array-based comparative genome hybridization (array-CGH) (Shaw-Smith et al., 2004). These arrays use small DNA segments (genomic clones as BACs or PCR products, oligonucleotides, cDNA) as baits/targets. Then the basic principle is to fluorescently label the test (tumor) and reference (normal) DNA using differently fluorescent labels, mixed them, perform denaturation so they can hybridize with the baits on the array. After hybridization, the array is scanned to detect the signals emitted by the labeled DNA. The fluorescent ratio of test and reference hybridization signals at each probe is used to determine the copy number changes (Theisen, 2008). However, despite their precision and low cost, they cannot identify *de novo* SV especially in regions not targeted by the designed probes unlike NGS technologies. As it was already highlighted, long-read sequencing or third-generations is superior to detecting SVs compared to short-read technologies of second generation. Unfortunately, the cost of long-read technologies is still too high for it to be implemented routinely in clinics. Nevertheless, there are multiple tools available for reliably calling the SVs from other technologies.

The designed SV/CNV tools utilize one or more approaches: read-pairs (RP), read-depth (RD), split read (SR) and assembly (Baker, 2012) shown on Figure 2. The simplest method, read-depth, summarized the number of reads per targeted regions in the genome.

Duplicated regions exhibit twice the number of reads, while deleted regions will have none. Deletions also detected using split reads, where the reads map to two locations separated by the deleted region into the genome. If the reads have a high mapping quality, they become a valuable resource to precise detecting of SVs due to known length of the split read. On contrary, assembly methods provide identification of SVs by examining the differences of *de novo* assembled and the reference genome. Although they are more expensive and computationally demanding, they offer more reliable SV detection in complex regions with a lot of repetitive elements. However, the most powerful and standard approach at the moment in SV detection is paired-end whole-genome sequencing data where both ends of the fragments are sequenced. Similar to SR, if the mapped length between start position of forward and reverse reads is different from insert size, a certain SV is detected. Insert size is the length of the sequenced DNA fragment between when preparing the library for WGS DNA sequencing. Information about pair-ends reads allows for detection of complex SVs such as inversion, tandem duplications among others.

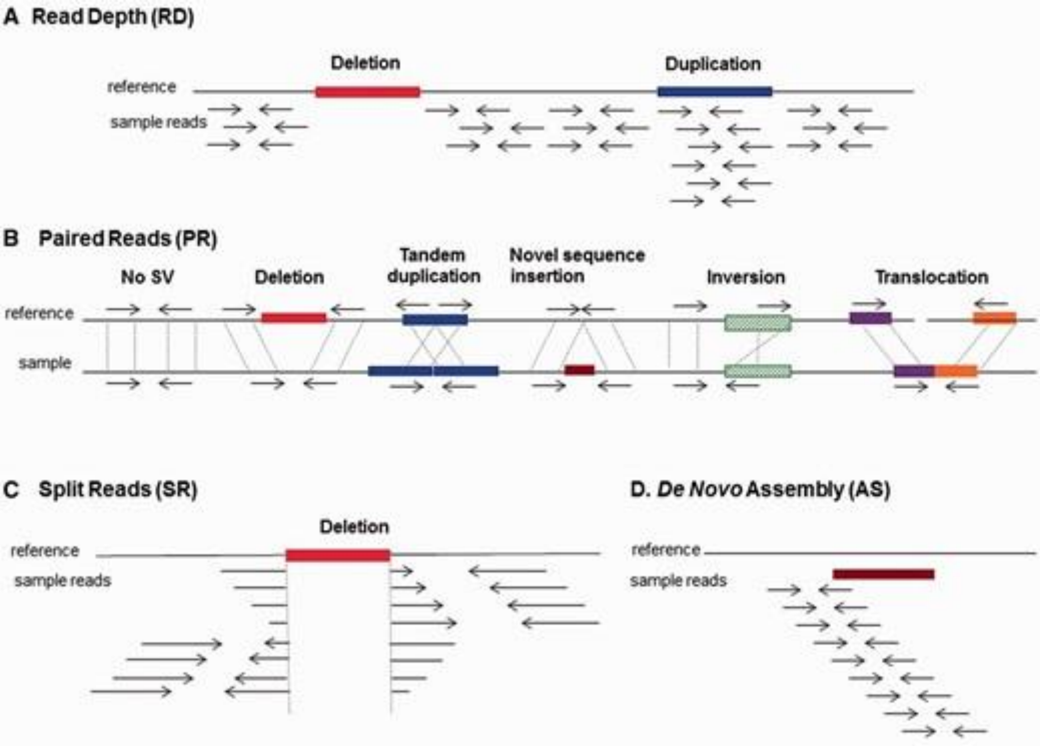


Figure 2. *Computational approaches for structural variant (SV) detection. Taken from (Escaramís et al., 2015)*

Gabrielaite et al. (2021) suggested to use multiple SV/CNV tools, similarly to ensemble approach with somatic mutations. They benchmarked 11 SV/CNV tools and found that the best tools when combining are: DELLY, GATK gCNV, Lumpy and cn.MOPS.

2.2.1.2 Driver mutations guide clinical therapy guidelines

Not all mutations will lead to cancer development. Therefore, it is important to distinguish “driver mutations” that confer growth advantage and has been positively selected from “passenger mutations” that do not contribute to carcinogenesis (Stratton et al., 2009). Driver mutations are often activating mutations such as missense or focal amplification in oncogenes, while they are inactivating in tumor suppressor genes, except in some genes like *TP53*. Both oncogenes and tumor suppressor genes harboring driver mutation are called “driver” or “cancer genes”. Usually, 4-9 driver mutations are required for cancer to develop (Stratton et al., 2009). Diagnostic, prognostic and treatment approaches are still guided by the identification of driver event in cancer genes.

One of the most common examples of personalized therapy is inactivation of the MAPK signaling pathways by identifying and targeting key driver genes with specific driver mutations. In malignant melanoma, *BRAF* proto-oncogene mutations are prevalent in over 66% of cases (Davies et al., 2002). The V600E missense mutation is a significant driver mutation, resulting in a substitution of valine to glutamine acid, which causes constitutive activation of RAS-RAF-MEK-ERK signaling pathway or shorter MAPK pathway. This variant is found in over 80% of all BRAF-mutated melanomas (Davies et al., 2002). Valine-to-lysine (V600K) and valine-to-arginine (V600R) mutations represent additional driver mutations in BRAF, accounting for 20% and 7% of all BRAF mutations, respectively (Manzano et al., 2016). The two gold standard therapeutic options for treating BRAFV600 mutant melanoma are combining a BRAF inhibitor with a MEK inhibitor and using immune checkpoint inhibitors (Flaherty et al., 2012; Reddy et al., 2017). In breast and ovarian cancers, the identification of bi-allelic mutations in *BRCA1*/*2* tumor suppressors genes is a routine genetic testing for patients with severe family histories of these cancers. *BRCA* genes are responsible for repair of double-stranded DNA (dsDNA) breaks via homology recombination repair (HRR) mechanism. Inactivation of both alleles of either *BRCA* gene, disables tumor capability to utilize this mechanism and forces the tumor to switch to non-homologous end joining (NHEJ) as an alternative DNA repair mechanism for dsDNA breaks. NHEJ pathways increases genomic instabilities by introducing additional mutations such as indels due to its working mechanism. This can be taken as an advantage in therapy with PARP (poly-ADP ribose polymerase) inhibitors that disable the repair of single-stranded DNA breaks which ultimately convert into dsDNA breaks. PARP inhibitors are forcing homology recombinant deficient (HRD) cancer cells to use NHEJ pathways which will lead them to the path of destruction by increase genomic instability and activating apoptosis. Other gene mutations associated with the HRD phenotype include mutations in *PALB2*, as well as promoter methylation of *BRCA1* and *RAD51C*, which account for the majority of HRD cases (Antoniou et al., 2014; Ruscito et al., 2014; Štancl et al., 2022). Since not all of the identified driver genes are druggable, the treatment options for certain cancers are slimmer than in others. An example is hepatocellular carcinoma (HCC) or liver cancer, where the best treatment is surgical removal or liver transplant especially in more advanced stages of HCC. There are tyrosine-kinase inhibitors (TKIs) which have proven to be successful in treating HCC like sorafenib in combination with other treatments (cytotoxic chemotherapy, immunotherapy, other TKIs) (da Fonseca et al., 2020). However specific drugs targeting HCC most mutable cancer genes, such as *TERT* and *TP53*, are still missing.

Hence, identification of driver variants in genes in pivot for personalized medicine as well as designing new therapeutics to fight cancer. However, identification of driver mutations has proven to be quite challenging due to low mutation rate of certain tumor types, high heterogeneity, significantly higher number of passengers than driver mutations and low number of samples. Therefore, numerous strategies and tools have been developed to identify driver mutations. They can be broadly grouped into three groups: frequency-based, function-based and combined frequency- and function-based approaches (Pon & Marra, 2015). Frequency-based methods suffer the most from the mentioned problems since they require a sufficient information to statistically identify driver mutations as ones with higher mutational rate frequency from background mutational frequency. Another challenge is the defining the background mutations. Multiple propositions have been made through the years from synonymous mutations in genome, synonymous one in introns and UTRs to modeling the frequencies using genomic features, but all of them have drawbacks discussed in detail by Pon and Marra (2015). Certain tools like MutSigCV tools estimates gene-specific background mutations rates considering genomic factors such as expression level and replication time, as well as corrects for variation with patient-specific mutation frequency and spectrum (Lawrence et al., 2013). On the contrary, only a function-based approach can predict driver mutations in one sample and is useful when interested in precise drivers' mutations affected by chemotherapy or other factors. However, they have extremely low accuracy and sensitivity (Gnad et al., 2013). There has been a huge surge of various driver mutations bioinformatic tools that combine these approaches. Integrative networks of gene expression, protein structure, multi-omics data, as well as supervised and unsupervised machine learning (ML) models have all contributed to improved identification of drivers and their role in carcinogenesis. As no single tool is perfect on its own, combining multiple approaches has emerged as a promising strategy for reliably identifying driver mutations and genes (Nourbakhsh et al., 2024).

2.2.2 Mutational signatures

While most tumor mutational landscapes comprise primarily of passenger mutations, their significance has often been overlooked in many previous cancer research. Despite not directly contributing to carcinogenesis, these mutations still provide valuable insights into the underlying mutational processes driving cancer development. Much like how we leave footprints in the sand, a variety of factors such as endogenous and exogenous mutagens leave distinct mutational footprints in the genome which we can identify through analysis of mutational signatures. Mutational signatures represent specific contexts of mutational combinations, defined by substitutions of pyrimidines within the Watson-Crick base pairs (C>A, C>G, C>T, T>A, T>C, T>G) along with their 5' and 3' flanking bases. Currently, single-base substitution (SBS) signatures are defined by 96 distinct trinucleotide contexts, resulting from the combination of six possible substitutions with 16 possible contexts each (Alexandrov et al., 2020). Other types of mutational signatures are double-base substitution (DBSs), small indels (IDs), rearrangement (RSs) and chromosome copy number changes (CNVs) signatures whose representation is shown on Figure 3.

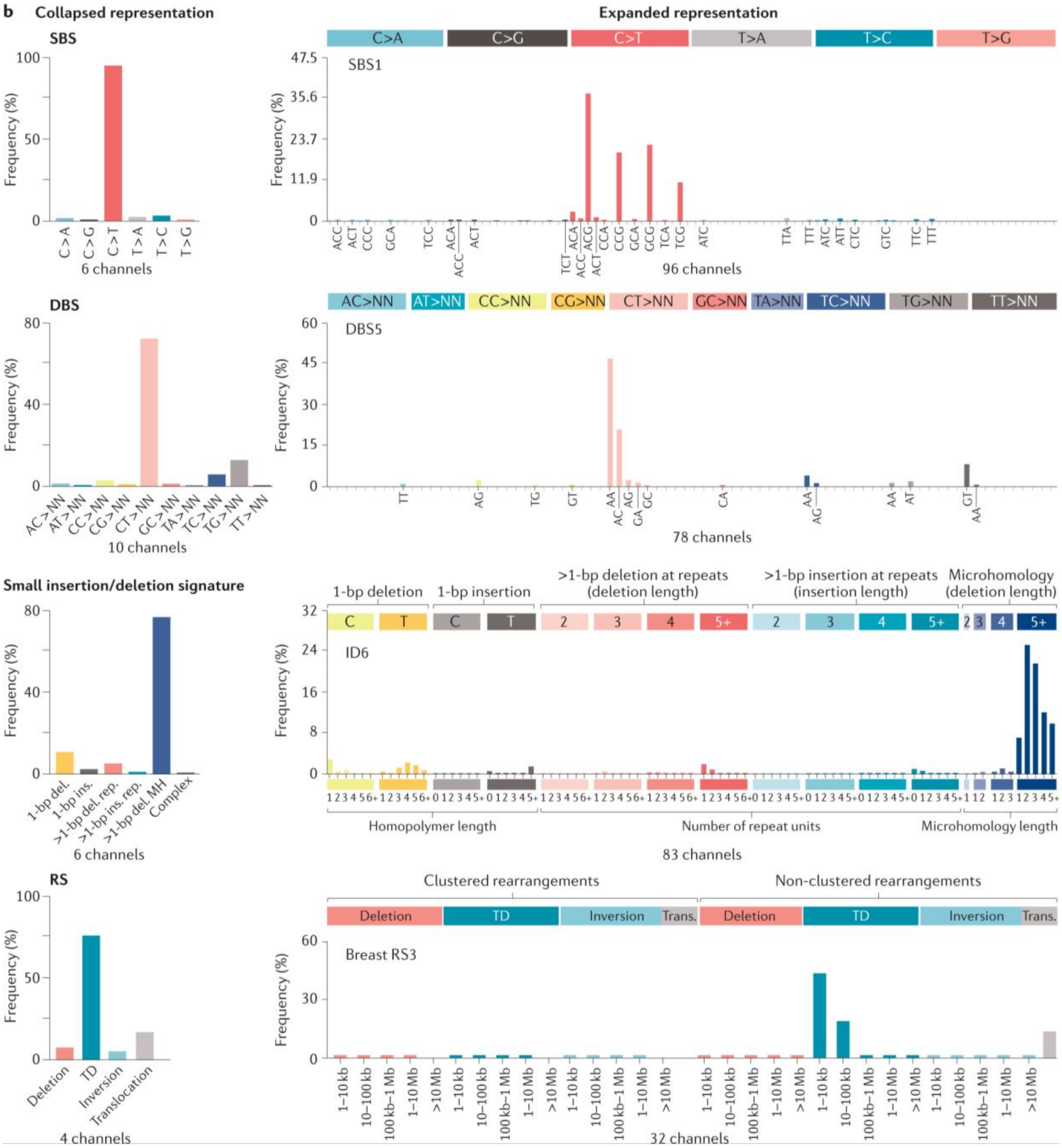


Figure 3. *Schematic representation of preferred method for presenting different mutational signatures. Single-base substitution mutational signatures (SBSs) are shown with 96 distinct trinucleotide contexts. Double-base substitution signatures (DBSs) are defined with 78 strand-agnostic DBS mutation types. Small indel signatures (IDs) are broadly categorized by type (insertion, deletion or complex). Single base indels are only classified as C or T. IDs are further classified based on the length of the mononucleotide repeat tract where they occur. Longer indels are classified by whether they occur at repeats or have microhomology at indel junctions. Rearrangement signatures (RSs) can be classified based on the four types of rearrangements and their regional clustering, with additional consideration given to the size of the rearranged fragment. Abbreviations used include del. for deletion, ins. for insertion, MH for microhomology, rep. for repeat, TD for tandem duplication, and trans. for translocation. Adjusted and taken from (Koh et al., 2021)*

While mutational signature analysis represents a hot new trend in cancer research, previous studies have hinted at a link between mutational patterns and underlying mutational processes. The first researches focused on studying patterns of somatic mutations found in single gene, such as *TP53*. Dangerous UV radiation has been linked to a high abundance of C>T mutations occurring exclusively at dipyrimidines (Brash et al., 1991). In lung cancer it was found that the *TP53* mutational patterns of smokers are significantly enriched in G to T transversions compared to non-smokers associated with exposure of benzo[a]pyrene (BaP), a famous tobacco carcinogen (Nik-Zainal et al., 2015; Pfeifer et al., 2002). Another example is G>T mutation at the third base of codon 249 of the *TP53* gene in hepatocellular carcinoma exposed to aflatoxin (Bressac et al., 1991). Due to rise of NGS technologies, the mutational patterns analysis was extended beyond single gene analysis. This resulted in plethora of studies that reinforced or discovered new mutational patterns from various factors; such as UV light of C>T, tobacco-smoke-associated damage of G>T, aristolochic acid of T>A at TpG and TpA dinucleotides, as well as distinct pattern of C>T and C>G mutations at TpC sites in a subset of breast cancers (Petljak & Alexandrov, 2016). All of the endogenous and exogenous processes can induce specific mutations in both normal and cancer cells over the cell's lifespan. Most famous SBS signature universally detected across all cancers and normal tissues is SBS1. This signature is associated with patient's age hence it is sometimes referred to as age-associated signature. SBS1 is characterized by C>T mutations at NCG sites due to deamination of 5-methylcytosine and represents the endogenous processes which generate mutations at a constant rate during cell's lifespan. Other signatures may only be active for a shorter period of time depending on the patient's lifestyle choices, as an example SBS4-tobbacoo related signature. The dynamic process of various mutational processes being active though the cell's lifespan leads to the accumulation of multiple mutational signatures with overlapping features (Koh et al., 2021; Van Hoeck et al., 2019). In order to successfully distinguish individual mutational processes from various mutational patterns present in the genomes, mathematical approaches had to be implemented to solve this problem. Only then it was possible to distinguish meaningful biological processes driving the carcinogenesis from artifacts caused by sequencing errors, mutational calling, and other sources of noise. For instance, some mutagens, notably exposure to UV radiation, have been associated with multiple mutational signatures, such as SBS7a, SBS7b, SBS7c, SBS7d, and recently annotated SBS38. In the COSMIC database, both SBS7a and SBS7b primarily exhibit C>T mutations, likely attributed to UV-induced photoproducts like cyclobutane pyrimidine dimers or 6-4 photoproducts. Conversely, SBS7c, characterized by T>A and T>C mutations, and SBS7d, with a predominance of T>C mutations, may result from translesion DNA synthesis by enzymes favoring T or G insertion over A respectively. On other hand, less studied SBS38 signature, exclusive to UV-associated melanomas, suggests potential indirect damage from UV light exposure (Tate et al., 2019). Hence, experimental validation is essential to attribute each mutational signature to its underlying cause.

There are two main methods for detecting mutational signatures: *de novo* signature extraction and signature refitting methods. Both methods produce a decomposition matrix $C \approx SE$, where C is the catalog matrix (mutation context per sample), S is the mutational signature matrix (mutation context per signature) and E is the exposure matrix, also termed "signature contribution" or "activity of a signature" (mutational signature per patient) (Maura et al., 2019). In the *de novo* method, the C matrix is utilized to derive S and E through modeling techniques that determine the optimal number of signatures. Conversely, in the fitting method, the catalog of already known signatures (S) is used, allowing algorithms to calculate the exposure matrix (E). *De novo* methods allow for unbiased identification of novel signatures and were first employed to characterize mutational processes found in 21 breast cancers(Nik-Zainal et al., 2012). Original method used nonnegative matrix factorization (NMF) to extract the signatures and whose results obtained with SigProfiler tool can be found in the COSMIC database. Other newer modification of the NMF, such as a Bayesian variant of NMF implemented in SignatureAnalyzer tool, showed similar performance when reconstructing known signatures (Alexandrov et al., 2020; J. Kim et al., 2016). The difference was mostly prominent in higher number of detected and assigned signature, especially to more problematic samples with hypermutations, with SignatureAnalyzer.

Multiple mutational fitting methods have been developed to determine the best fit of selected signatures (S) that closely reconstruct the original mutational context of a sample. These methods use either multiple linear regression, a non-negative least-squares constraints problem (NNLS), Bayesian inference or simulated annealing (SA) method to perform the fitting (Pandey et al., 2022). Since fitting methods are purely mathematical approaches aimed at fitting as many given signatures as possible, they tend to leading to a problem known as overfitting. Depending on the research objectives, it may be beneficial to apply stricter parameters in fitting or to use a pre-selected set of signatures to reduce the overfitting. One of the simplest ways to reduce overfitting is to perform fitting using bootstraps techniques. In each signature bootstrap, signature contribution is calculated as cosine similarity or mean squared error (MSE) of difference between the bootstrap estimates and original mutational context. The obtained distribution of calculated signature contributions is used to calculate the empirical probability of an exposure to be larger or equal to a given threshold (Huang et al., 2018). The main idea of this modified fitting methods is to remove those signatures with little or non-contribution that do not exceed a certain threshold. Another variation of this method is the strict fitting of signatures implemented in MutationalPatterns R package that moves in iterative fashion and stops removing signature once the difference in reconstruction error between two iterations is bigger than a set threshold (Manders et al., 2022). Although the main drawback of this methods is the arbitrary selection of the threshold values, these approaches enable identification of stable and unstable signatures allowing researches to call signatures more reliably. Removing certain signatures beforehand also reduces overfitting but introduces a bias. Therefore, it is also important to reconsider the initial set of mutational signatures (S) in fitting methods; whether to use tissue-specific or all mutational signatures (Maura et al., 2019). Tissue-specific sets of mutational signatures represent the up-to-date signatures that have been identified in certain cancers and reduce the number of unambiguous identification of signatures that represent artifacts in corresponding cancer. They are more informative when studying mutational processes in particular cancer to identify tissue-specific biomarkers and therapeutic targets, particularly when the tumor's cell-of-origin is known. Otherwise, in rare, less studied and/or unknown cancer types important signatures may go undetected due to their rarity or influence by factors beyond the examined tissue type.

2.2.2.1 Clinical implications with mutational signatures

Only in recent years have mutational signatures begun to demonstrate clinical importance, serving as novel biomarkers. Among various cancers, breast cancer mutational signatures had the most significant advancement in diagnostics field. Stratton’s group have identified specific mutational signature associated with *BRCA1/2* which they later identified as single-base substitution signatures SBS3 and SBS8 and rearrangement signature 3 and 5 (Nik-Zainal et al., 2012, 2016). Moreover, they detected that SBS3 often co-occurs with increased numbers of indels > 3 bp in length with overlapping microhomology at breakpoint junctions, consistent with double-strand break (DSB) repair by non-homologous end joining. Soon after, this was reinforced from study by Polak et al. (2017) who found that biallelic inactivation of *BRCA1/2* is associated with SBS3. They also highlighted that epigenetic silencing of *BRCA1* and *RAD51* through promotor methylation contributes to SBS3, alongside germline *PALB* mutation. Since epigenetic modifications are not typically included in routine diagnostic procedures, patients with these alterations that show the so called “BRCAness” phenotype may be excluded from potentially beneficial therapies. “BRCAness” or “homologous-recombinant deficient (HRD) phenotype” terms are often used to describe tumors without *BRCA* mutations who share similar clinicopathological and molecular characteristics to tumors harboring *BRCA1/2* mutations (Lord & Ashworth, 2016).

Hence, various tools have been developed utilizing the mutational signature to predict whether a patient has HRD phenotype in order to receive a proper treatment with PARP inhibitors or platinum-based chemotherapy. The most famous being HRDetect developed by Davies et al. (2017) and trained on *BRCA1/2*-null breast tumors. HRDetect is a weighted logistic regression model based on six input features: SBS signatures 3 and 8, and two rearrangement signatures 3 and 5, the proportion of small deletions with microhomology at the breakpoint junction, HRD index based on genomic scars. The tool achieved high sensitivity of 98.7% (AUC 0.98) for predicting HRD breast cancer patients. The population-based clinical study of whole-genome sequencing in triple-negative breast cancers demonstrated the effectiveness of HRDetect as an independent prognostic factor. Patients classified as HRD-high exhibited better outcomes on adjuvant chemotherapy for invasive disease-free survival and distant relapse-free interval compared to HRDetect-low individuals, regardless of the identified known cause of HRD (Staaf et al., 2019). Although HRDetect was initially developed using breast cancer data and demonstrated remarkable prognostic capabilities, its utility extends beyond breast cancer alone. Other cancer type patients, including ovarian, pancreatic, and metastatic prostate cancer patients, have also been shown to significantly benefit from HRDetect predictions for therapeutic choices.

In addition to the homologous recombination repair mechanism, deficiencies in other DNA repair mechanisms are known to produce specific mutational patterns and signatures in genomes. Cancers with defective DNA mismatch repair (MMR) exhibit a high frequency of indels, particularly at microsatellite unstable regions, known as microsatellite instability (MSI), and an increased mutational load of substitutions, predominantly C>T and C>A substitutions. Mutational signatures associated with MRR and MSI are SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and SBS44 (Tate et al., 2019). MMR deficiency has been extensively studied in colorectal cancer, where patients with MMR deficiency significantly benefit from immunotherapy. It has been demonstrated that MRR deficient patient from multiple cancers are sensitive to inhibitors of the programmed death 1 (PD1) immune checkpoint, such as pembrolizumab and nivolumab (Bouffet et al., 2016; Le et al., 2015, 2017). Deficiencies in base excision repair (BER) and nucleotide excision repair (NER) are associated with characteristic mutational signatures. BER deficiency is linked to signature SBS30, associated with inactivating mutations in *NTHL1*, while SBS18 has been linked to *MUTYH* mutations with possible etiology of reactive oxygen species (ROS) (Tate et al., 2019). Signatures similar to SBS5 has been linked to mutations in NER’s core protein ERCC2. Although patients with higher levels of this signature benefit from cisplatin therapy, it has limited diagnostic value as it is also considered an age-related signature with contamination from SBS16 (Tate et al., 2019; Van Hoeck et al., 2019).

Unlike mutational signatures that can serve as biomarkers for specific therapy selection, some signatures indicate resistance to certain treatments. Such examples are APOBEC signatures that have been associated with resistance to tamoxifen (Law et al., 2016; Sieuwerts et al., 2014). Apolipoprotein B mRNA-editing enzyme (APOBEC) consists of family of evolutionarily conserved cytidine deaminases with a role to protecting cells as a part of cellular immune response to viruses and retrotransposons. The major source of mutations is the *APOBEC3B* gene which is extensively studied in breast cancer (Petljak et al., 2022). Alongside other family members, it is responsible for driving tumor development, evolution and resistance to chemotherapy (Swanton et al., 2015). APOBEC mutations are responsible for kateagis regions, local hypermutated regions in the genome. In COSMIC database, APOBEC mutational signatures are SBS2 and SBS13. While SBS2 are predominantly C>T mutations at TCN trinucleotides, SBS13 are mostly C>G mutations at TCN sites (Petljak & Alexandrov, 2016).

2.3. Epigenetics and cancer

The second most important emerging hallmark of cancer, after genomic instability, that lead to cancer development is nonmutational epigenetic reprogramming. While mutations can alter protein function and even render its production completely, epimutations are considered a driving force of significantly changing the expression levels of cancer genes, both oncogenes and tumor suppressors. Epimutations can be reversed by nature, unlike regular genetic mutations, which makes them an excellent therapeutic target. Epigenome is responsible for condensing the ~2 m long genome into the nucleus of the cell, as well as regulating tissue-specific gene expression to maintain normal cellular identity and function. Main components of epigenome are DNA methylation, histone modification, chromatin remodeling, and non-coding RNAs. In the next few sections I will briefly go through normal function of each component, their disruption in cancer and potential therapeutic choices.

2.3.1 DNA methylation

One of the most common and well-studied epigenetic regulations in eukaryotes that chemically modify DNA sequence is the methylation of cytosine in CpG dinucleotides, resulting in 5-methylcytosine (5mC). The 5-methylcytosine (m5C) is a repressive mark of the epigenome, responsible for silencing gene expression through either preventing access for transcriptional factors (TFs) to bind on DNA or by recruiting regulative methyl-binding domain proteins (MBDs) along with chromatin remodelers (Lu et al., 2020). The majority of CpG sequences are methylated in human genome, except for CpG dinucleotides concentrated in short CpG-rich DNA stretches known as CpG islands. Hypomethylation of CpG islands is important for gene regulations as these regions are predominantly located at the 5’ end of human genes and occupy more than ~50% of promoters (Bird, 1986; Lister et al., 2009). DNA methyltransferases (DNMTs) are enzymes responsible for DNA methylation. DNMT1 is involved in maintaining DNA methylation patterns by methylating hemimethylated DNA during replication. On the other hand, DNMT3a and DNMT3b are *de novo* DNA methyltransferases, responsible for methylating unmethylated DNA independently of replication processes (Lu et al., 2020). The demethylation 5mC is either driven by ten-eleven translocation methyl-cytosine dioxygenases which produce 5-hydroxymethylcytosine (5-hmC) alongside other intermediate products, or by passive demethylation during DNA replication if DNMTs are absent (Costa et al., 2023). Disruption of the homeostasis between methylation and demethylation of the genome can lead to the development of various diseases, including cancer. Famous examples of hypermethylation of tumor suppressor genes include *BRCA1* in breast and ovarian cancers (Polak et al., 2017; Ruscito et al., 2014), and *PTEN* in melanoma (Mirmohammadsadegh et al., 2006). In contrast, hypomethylation-mediated upregulation of oncogenes is observed in various cancers, such as the expression of MAGE (melanoma-associated antigen) in melanoma (D. Wang et al., 2016) and TERT in hepatocellular carcinoma (H. Zhang et al., 2015). Therapeutic strategies for reversing aberrant DNA methylation of tumor suppressor genes involve the use of DNA methyltransferase inhibitors (DNMTIs), which can be categorized into two main groups: cytosine analogue inhibitors and non-nucleotide analogue inhibitors. These inhibitors function by either depleting the pool of active DNMT enzymes or interfering with their binding to DNA (Lu et al., 2020).

2.3.2 Histone modifications

Another frequently studied epigenetic mechanism regulating gene expression is histone modification. Histones are proteins responsible for packaging DNA into nucleosomes, which are the fundamental units of chromatin. Each nucleosome core consists of protein octamer defined with four histone dimers (H3, H4, H2A, H2B) linked by histone H1 and wraps up ~ 146 bp DNA (Luger et al., 1997). Histone proteins contain a globular C-terminal domain and an extended N-terminal domains which undergo various post-translational modifications (PTM) such as methylation, acetylation, phosphorylation and other modifications of a specific amino acid (Lu et al., 2020). Each modification has been associated with either transcriptional activation/euchromatin regions or repression/heterochromatin regions. In general, sumoylation, deamination and proline isomerization are detected in transcriptionally silent regions, while acetylation and phosphorylation are correlated with open chromatin and positive transcription (Kouzarides, 2007). On the other hand, methylation and ubiquitination can be associated with both transcriptional activation and repression. The most well studied modifications are methylation and acetylation. Histone methyltransferases (HMTs) are responsible for various histone methylation at different amino acid residues. Depending which residue was methylated, the biological consequence can significantly differ. For instance, the histone H3 lysine 27 trimethylation (H3K27me3) is associated with the repression of transcription elongation, while histone H3 lysine 4 trimethylation (H3K4me3) is associated with transcriptionally active/poised chromatin and histone H3 lysine 36 trimethylation (H3K36me3) with open chromatin through removal of histone acetylations in the wake of an elongating pol II (Karlić et al., 2010; Kouzarides, 2007). Based on these associations, Karlić et al. (2010) showed that only a small number of histone modifications can be used to predict the level of gene expression using quantitative models. Specifically, they found that H3K4me3 and H3K79me1 modifications were most important for predicting expression of genes with low CpG content promoters, whereas H3K27ac and H4K20me1 were more relevant for genes with high CpG content promoters.

Maintaining the equilibrium of histone modifications is pivotal for normal cellular function. Aberrant reprograming of histone modifications has been associated with various disease progression, including cancer. A common hallmark of cancer is loss of acetylation at Lys16 (H4K16ac) and trimethylation at Lys20 of histone H4 (H4K20me3) (Fraga et al., 2005). In prostate cancer, histone modifications associated with transcriptional activation, acetylation (H3K9, H3K18 and H4K12) and demethylation (H4R3 and H3K4), have shown to have prognostic value (Seligson et al., 2005). Moreover, aberrant methylation patterns of H3K9 and H3K27 have been shown to drives oncogenic transformation and chemoresistance (Costa et al., 2023; Sasidharan Nair et al., 2018; Sharma et al., 2009). The dysregulated distribution of histone modifications in cancer genomes often involves altered histone methyltransferases (HMTs) and histone deacetylases (HDACs). For instance, overexpression of H3K27 HMT (EZH2) and H3K9 HMT (G9a) have been detected in breast and liver cancer, respectively (Borkiewicz, 2021; Kondo et al., 2007). The most advanced clinical studies include the tazemetostat, an epidrug approved by the Food and Drug Administration (FDA) targeting EZH2 for advanced epithelioid sarcoma. However, there are much more FDA approved histone deacetylase inhibitors (HDACI), such as panobinostat and tucidinostat, that are considered to be one of the main epidrug alongside DNMTIs in clinics (Costa et al., 2023; Nepali & Liou, 2021).

2.3.3 Non-covalent modifications

Non-covalent modification of epigenome include nucleosome remodeling and non-coding RNAs (ncRNA). Nucleosome remodeling involves ATP-dependent remodeling enzymes that regulate gene expression by mobilizing nucleosome using energy from ATP hydrolysis. One of the examples is SWI/SNF complex which was found to be mutated nearly 25% of all cancers. Furthermore, dysregulated expression of SWI/SNF subunits has been closely associated with tumor initiation and progression (Zhang & Li, 2022). One of the main components of SWI/SNF complex, that is frequently mutated, is a tumor suppressor *ARID1A*. Alongside other parts of SWI/SNF complex, *ARID1A* when through a lot of preclinical and clinical studies with varying outcomes (Mittal & Roberts, 2020). Non-coding RNAs (ncRNAs) are broadly categorized into two main groups: long ncRNAs (>200 nucleotides) and small ncRNAs (<200 nucleotides). Both groups of ncRNAs can function as oncogenes or tumor suppressors by modulating gene expression post-transcriptionally through translation inhibition or mRNA degradation (Costa et al., 2023). Also, lncRNA can regulate gene expression in the nucleus such as guiding chromatin-modifying complexes like PRC2 to specific genomic loci, leading to the formation of inactive chromatin marked by H3K27me3, or by recruiting MLL histone methyltransferase complex to gene promoters, promoting active chromatin marked by H3K4me3, or they may act as decoys for histone deacetylases, maintaining activating chromatin modifications such as H3K9ac and H3K56ac (Ahmad et al., 2023). The most studied small ncRNAs are miRNAs, approximately 20-25 bp molecules targeting multiple genes, serving as potential biomarkers or therapeutic targets, although drug development based on miRNAs remains challenging, showing promising potential for cancer treatment (Kim & Croce, 2023).

2.4 Chromatin structure

Chromatin is compacted and organized into higher-order structures, which play significant roles in regulating gene expression and cellular function. Studies on chromatin conformation (Hi-C) have provided deeper insights into the 3D genome structure, revealing various active and inactive chromatin regions across different scales. In eukaryotes, chromosomes are spatially arranged within specific regions known as “chromosomal territories” (CT) (Cremer & Cremer, 2010) shown on Figure 4AB. Within these territories, two hierarchical structures are observed: A and B compartments at the megabase level, associated with euchromatin and heterochromatin, respectively (Figure 4CD). At the sub-megabase level, topologically associated domains (TADs) represent DNA interactions that occur more frequently within a given domain than with regions in other domains (Figure 4EF) (Akdemir, Le, Chandran, et al., 2020; Boltsis et al., 2021). TADs are formed by chromatin loop extrusion mechanism, during which DNA strands move within the cohesin or SMC complex until encountering bound CCCTC-binding factor (CTCF) (Rajderkar et al., 2023). Furthermore, they are organized into two basic features: the TAD, self-interacting loop domains where cis-regulatory elements and genes interact, and the TAD boundary, a region between TADs acting as insulators (McArthur & Capra, 2021). One such cis-regulatory elements which in TADs are responsible for co-regulating gene expressions are super-enhancers (SE). SE are large clusters of regulatory elements defined by unusually strong enrichment for the binding of transcriptional coactivators, specifically Med, EP300, BRD4 and CDK7, as well as their extremely high potential of activating transcription of target genes (Jia et al., 2019; Pott & Lieb, 2015). Therefore, TADs and TADs boundaries have been shown to be a crucial fundamental units in genome organization which regulated gene expression in tissue- and cell-specific manner (Boltsis et al., 2021; H. S. Long et al., 2022; Rajderkar et al., 2023; Schoenfelder & Fraser, 2019). Although multiple studies have shown that TADs are highly conserved among different cell types and across various species, newer research are defining that there maybe even ~50% TADs and up to 80% TADs boundaries which are different across cell types (Boltsis et al., 2021; McArthur & Capra, 2021).

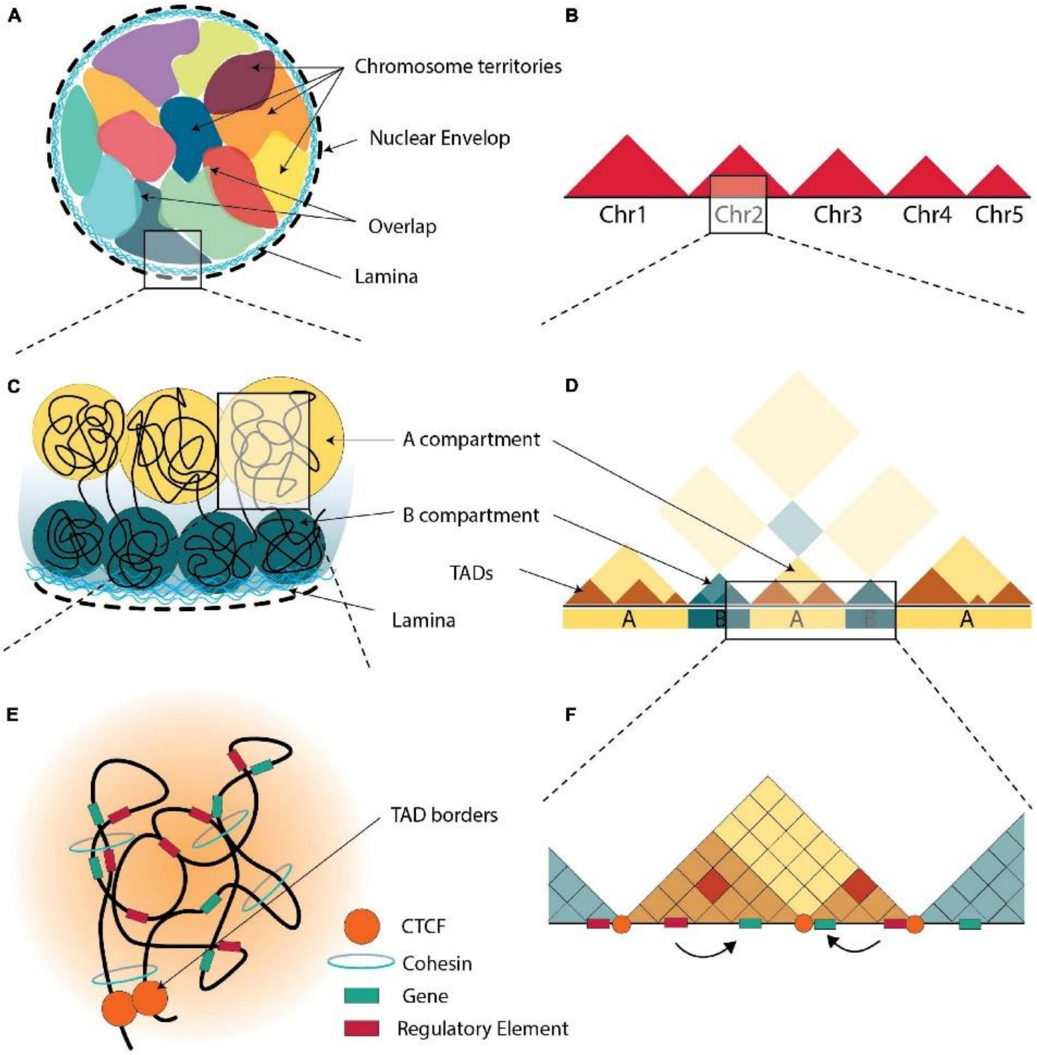


Figure 4. *Schematic representation of 3D spatial organization of chromatin in eukaryotes. A) Arrangement of chromosomes in the nucleus where all chromosomes are connected to nuclear lamina. Chromosomal territories (CT) are colored by distinct color. Overlapping areas of CTs are also highlighted. B) An illustration of Hi-C map showing the frequency of physical interactions between pairs of genomic regions on the chromosomal scale. C) "A" (yellow) and "B" (green) compartments of chromatin. D) An illustration of Hi-C map at the compartmental scale showing distinct plaid pattern from interactions of distal chromatin and compartments "A" and "B". E) Representation of topologically associated domains (TADs) and associated proteins necessary for loop extrusion mechanism. C) Hi-C map at sub-megabase scale showing TADs. Regulatory elements, such as super-enhancers, are typically located closer to gene promoters within topologically associated domains (TADs), facilitating their regulatory control over gene expression. Taken from (Boltsis et al., 2021)*

The disruption of TADs can contribute to the development of various diseases, including cancer, by dysregulating gene expression. One notable mechanism implicated in this process is "enhancer hijacking" or "enhancer adoption", where chromosomal rearrangements lead to the fusion of adjacent TADs (Boltsis et al., 2021). This fusion enables enhancers from neighboring TADs to inappropriately activate oncogenes thereby promoting tumorigenesis. One such examples is the activation of key oncogenic driver *TAL1* for T cell acute lymphoblastic leukemia (T-ALL) by site-specific deletion of a loop boundary CTCF site (Hnisz et al., 2016). A comprehensive pan-cancer study of 38 tumor types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) also detected numerous chromatin loops disrupted by structural variants across various cancers (Akdemir et al., 2020). For instance, a CTCF site proximal to *FOXC1* coincides with recurrent deletions observed in esophageal, gastric, and colon adenocarcinomas, while another CTCF site near *BCL6* in hepatocellular carcinoma and breast adenocarcinoma (Akdemir et al., 2020). Moreover, structural variants can also lead to TADs shuffling and formation of "neo-TADs". "Neo-TADs" represent a significant outcome of SVs affecting the genome, alongside "enhancer hijacking," resulting in the creation of novel chromosomal domains (Boltsis et al., 2021). This effect is particularly prominent in cancerous cells, where TADs are often observed to be shorter compared to their normal counterparts (McArthur & Capra, 2021). In prostate cancer it was shown that these new smaller TADs often reside within the older TADs instead of forming a completely new one, as well as keeping majority of TAD boundaries, ~98%, intact (Taberlay et al., 2016). The similar trend was observed between mammary epithelial and breast cancer cells with creation of multiple sub-TADs related to repression of WNT signaling and high number of conserved TAD boundaries (Akdemir et al., 2020; Barutcu et al., 2015). Although targeting chromatin interactions holds promise for precise gene expression control by perturbing promoter-enhancer interactions, challenges arise due to the involvement of CTCF, cohesin, and other transcription factors in multiple chromatin interactions and signaling pathways, leading to potential off-target effects (Boltsis et al., 2021).

2.5 Cell-of-origin of cancer

The primary choice of therapeutic options for oncologists is guided by the identification of a tumor's primary origin through histopathological examination and biomarker profiling using various NGS tools. However, carcinomas of CUP, which account for approximately 10-15% of diagnosed tumors, pose a unique challenge due to their unknown and diverse morphological, immunohistochemical, and molecular characteristics. Reliable identification of the primary origin of CUP tumors requires the integration of multiple techniques, ranging from immunohistochemistry to NGS (Beauchamp et al., 2023). When the tumor type cannot be initially identified based on morphological characteristics alone, the first step is to employ a basic initial immunohistochemistry (IHC) panel targeting broad cancer types. This may include markers such as S100 for advanced cutaneous melanoma, CD45 for lymphoma, and AE1/AE3, which are epithelial markers positive for carcinoma (Schofield et al., 2018). Subsequently, if carcinoma is detected, more specialized IHC markers are utilized to determine the specific subtype of carcinoma and/or adenocarcinoma. To enhance the accuracy of classification, it is recommended to use antibody cocktails on the same slide (Beauchamp et al., 2023). Furthermore, poorly or undifferentiated carcinomas may necessitate more detailed analysis using larger sequencing panels, as well as WGS or WXS if needed. Current understanding of the primary cell-of-origin in cancer relies on mouse models, primarily employing two approaches: 1) utilizing cell-specific promoters to drive expression of an oncogene or the Cre-mediated deletion of a tumor-suppressor gene within specific cell subsets *in vivo*; and 2) genetically manipulating *ex vivo* cells, which are subsequently orthotopically transplanted into mice to assess their predisposition to tumor initiation (Visvader, 2011). However, these models may not fully recapitulate the complex genetic, epigenetic, environmental, and stochastic processes observed in human cancers. Consequently, various approaches, including the examination of genomic, epigenomic, and gene expression profiles in cancers using machine-learning methods, have been developed to address these limitations.

2.5.1 Machine learning algorithms for determining cell-of-origin

Previous machine learning (ML) methods focused on identifying the most critical features associated with carcinogenesis. However, these studies often faced limitations, requiring a substantial number of samples to accurately identify the features necessary for correctly predicting the COO. Additionally, studies restricted to a single cancer type may not generalize well to other cancer types. In contrast, pan-cancer studies, while more challenging to interpret, have the potential to uncover universal biomarkers that can be applied across different cancer types for COO identification. In pan-cancer of RNA-seq based ML models there are small number of genes shared among other studies, questioning if the found genes and models can be reliably used for COO discovery (Štancl & Karlić, 2023). Whole-genome sequencing models, which incorporate multiple genomic features, tend to achieve higher accuracies than RNA-seq and/or whole-exome ML models. For example, a random forest COO classifier trained across 35 cancers achieved approximately 90% recall and precision based on cross-validation and test set predictions (Nguyen et al., 2022). This model utilized 511 features based on simple and complex somatic driver and passenger mutations, such as density (RMD) profiles, simple and complex rearrangements, mutational signatures, and gene gain/loss of function, among others.

The ML model for predicting the COO developed by Polak et al. (2015) overcomes these challenges by leveraging WGS mutations and epigenome of normal tissue cells across 1

MB genomic windows to be able predict the COO of individual patients. This approach capitalizes on the influence of chromatin structure, regulated by processes such as histone modifications, on the accumulation of background (passenger) mutations in a cell type-specific manner. Their study was based on two key observations: 1) mutations exhibit non-uniform distribution along chromosomes and across tumor types; 2) mutation densities correlate with regional histone modifications, DNA accessibility, and DNA replication timing in a tissue-specific manner (Kübler et al., 2019). The improved accuracy of the recently developed extension of this model, COOBoostR (S. Yang et al., 2023a), suggests that alterations in chromatin marks, particularly those occurring in tissue-specific enhancer regions, likely influence the somatic mutation density profile in these regions. This phenomenon has been observed in Barrett's metaplasia and esophageal adenocarcinoma. Further investigation into additional features and extensions of this model approach is needed.

3 Materials and methods

3.1 Publicly available data

3.1.1 Whole genome and whole-exome sequencing data

I used available whole genome and whole-exome sequencing datasets from ICGC/TCGA repository for liver, skin melanoma and breast cancer types. Summary of open and controlled access number of ICGC/TCGA donors is shown in Table 2.

Table 2. *Number of ICGC/TCGA donors of raw called single-nucleotide variants (SNVs) and indel variants for liver, skin melanoma and breast cancer*

Cancer cohorts		**WGS**		**WXS**	
		Open	**Controlled**	**Open**	**Controlled**
Liver	LICA-CN	112	0	288	0
	LICA-FR	49	6	234	0
	LIHC-US	54	54	362	375
	LIHM-FR	0	0	4	0
	LINC-JP	31	28	363	0
	LIRI-JP	258	251	0	
	LIAD-FR	5		30	
	TOTAL	**509**		**1281**	
Skin	SKCM-US	37	38	466	470
	MELA-AU	183	70	0	0
	SKCA-BR	100	0	0	0
	TOTAL	**320**		**466**	
Breast	BRCA-EU	569	78	0	0
	BRCA-US	91	92	1015	1 044
	BRCA-UK	45	45	117	0
	BRCA-KR	0	0	50	0
	BRCA-FR	72	0	0	0
	TOTAL	**777**		**1182**	

Due to the lack of the same donors in controlled data and to reduce bias from analyzing open and controlled data, I will take open access called raw SNVs and indel mutations for downstream analyses to analyze the variants.

3.1.2 Characteristics of individual patients

Clinical information of patients was obtained from the ICGC/TCGA platforms. The performance of patients' cell-of-origin models was analyzed in relation to their tumor histological types, which were annotated using the ICD-O (International Classification of Diseases for Oncology) shown in Table 3.

Table 3. *Annotated tumor histological type and tumor histological codes of ICD-O International Classification of Diseases for Oncology from ICGC/TCGA datasets for breast, liver and skin cancers. NOS means not otherwise specified. *LIAD-FR cohort samples are all hepatocellular adenoma*

Cancer type	Histological type	Histological code
Breast	Adenoid cystic carcinoma	8200/3
	Carcinoma with apocrine differentiation	8401/3
	Duct and lobular carcinoma	8520/3 and 8022/3
	Duct micropapillary carcinoma	8507/3
	Infiltrating duct carcinoma	8500/3
	Intraductal papillary adenocarcinoma with invasion	8503/3
	Lobular carcinoma	8520/3
	Pleomorphic carcinoma	8022/3
	Medullary carcinoma	8510/3
	Metaplastic carcinoma	8575/3
	Mucinous adenocarcinoma	8480/3
	Neuroendocrine carcinoma	8246/3
	Tubular and invasive Cribriform carcinoma	8211/3 and 8201/3
Liver	Cholangiocarcinoma	8160/3
	Combined hepatocellular + cholangiocarcinoma	8180/3
	Fibrolamellar hepatocellular carcinoma	8171/3
	Hepatocellular adenoma*	
	Hepatocellular carcinoma	8170/3
Skin	Acral lentiginous melanoma	8744/3
	Desmoplastic melanoma	8745/3
	Lentigo maligna melanoma	8742/3
	Malignant melanoma, NOS	8720/3
	Mucosal lentiginous melanoma	8746/3
	Nodular melanoma	8721/3

	Superficial spreading melanoma	8743/3

For breast cancer, the 86 patients were further annotated using Prediction Analysis of Microarray 50 (PAM50) metrics into Her2, Basal, LumA and LumB from available data upon request from Kübler et al. (2019). Moreover, I took published scores of HRD and CHORD tools from my paper (Štancl et al., 2022) for 371 breast cancer patients. I labeled each patient as homologous recombinant deficient if HRDetect score was above 0.7 or CHORD score was above 0.5.

3.1.3 1 Mb genomic regions

As previously described by Polak et al. (2015), I divided the human genome (GRCh37) into genomic regions of 1Mb window size. I excluded regions overlapping with centromeres and telomeres, as well as regions with a low fraction of uniquely mappable bases (<92% of bases within uniquely mapped 36-mers). In total, I was left with 2,128 1Mb genomic regions.

3.1.4 Topologically associated domains

I download publicly available topologically associated domains (TADs) from the TADBK (T. Liu et al., 2019) and 3D genome browser (Y. Wang et al., 2018) for normal tissues or cell lines. Specifically, liver tissue data was represented by a single TAD dataset from STL011. For melanocytes, no specific TAD data was available; therefore, I utilized TADs from the epidermal keratinocyte cell line NHEK as the closest proxy. For breast tissue, TADs from the mammary epithelial cell line HMEC were used. The characteristics of the selected TADs are detailed in Table 4.

Table 4. *Topologically-associated domains of closely related tissues and cell lines to breast, liver and skin melanoma cancers that was used in downstream development of cell-of-origin model*

Cell type	**TADs calling tool**	**Number of TADs**	**Median TAD length**	**Standard deviation of TAD length**
HMEC	DI 10kb	4363	430001	548835.7803
	DI 50kb	2621	800001	1006272.105
	GMAP 10kb	3069	710001	674360.2005
	GMAP 50kb	1851	1300001	816604.8367
	IS 10kb	4854	440001	729101.2602

	IS 50kb	3154	350001	660268.5424
	Lieberman	3235	600001	988088.7944
Liver	STL011	2066	720001	767317.5417
NHEK	DI 10kb	4213	470001	590566.1908
	DI 50kb	2734	800001	849001.1633
	GMAP 10kb	3238	700001	591058.8569
	GMAP 50kb	1876	1300001	819498.1334
	IS 10kb	5044	460001	539163.8059
	IS 50kb	3189	350001	659025.0822
	Lieberman	2832	725001	1011546.394

3.1.5 Super-enhancers

Tissue specific super-enhancers (SEs) were downloaded from SEdb2.0 (Y. Wang et al., 2023). For each cancer type I selected both normal and cancerous cell lines if they exist which are summarized in Table 5 alongside the total number of SEs found in each tissue and cell-line.

Table 5. *Tissue specific super-enhancers (SEs) of normal and cancerous cell lines for breast, liver and skin melanoma cancer types from SEdb2.0*

SE tissue	**Cell-line**	**Type**	**Count**
Breast epithelium	breast-epithelium	Tissue	923
Liver	hepatocyte	*In vitro* differentiated cell	1329
Liver	HepG2	Cell line	505
Liver	hepatocytes_d1	Primary cells	1263
Liver	hepatocytes_d3	Primary cells	1333
Liver	hepatocytes_d6	Primary cells	1139
Liver	HuH-7	Cell line	39
Liver	SMMC-7721	Cell line	742
Liver	liver	Tissue	724
Skin	keratinocyte	Primary cells	1033
Skin	BJ	Cell line	235
Skin	neonatal keratinocytes	Primary cells	913
Skin	SK-MEL-5 DMSO 6h	Cell line	1142
Melanoma	CJM	Cell line	821

Melanoma	COLO679	Cell line	829
Melanoma	LOX-IMVI	Cell line	914

3.1.6 Cancer-associated genes

I downloaded the Cancer Gene Census set of 736 genes from the COSMIC database (Tate et al., 2019) on December 23, 2022. Additionally, I obtained 164 canonical and 313 driver tumor immune microenvironment (TIME) genes from Misetic et al. (2023).

3.1.7 Chip-seq data of histone modifications

To develop the 1 Mb gene models, I utilized the raw count data from 98 normal tissues obtained from Kübler et al. (2019) which is available only upon request. To summarize the counts per topologically-associated domain and genes, I compiled a comprehensive set of 664 epigenomic ChIP-seq datasets. These datasets were used for chromatin feature selection, correlation analyses, and cell-of-origin prediction analyses. The data was sourced from the NIH Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics Consortium et al., 2015) and downloaded on January 31, 2021. The NIH Roadmap epigenomics data can be accessed through the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2013) under the accession number GSE18927. The ChIP-seq data were mapped and peaks were called on the human genome version hg19/GRCh37. I calculated the reads per kilobase per million mapped reads (RPKM) for histone modifications including H3K27ac (histone H3 lysine 27 acetylation), H3K27me3 (histone H3 lysine 27 trimethylation), H3K36me3 (histone H3 lysine 36 trimethylation), H3K4me1 (histone H3 lysine 4 monomethylation), H3K4me3 (histone H3 lysine 4 trimethylation), and H3K9me3 (histone H3 lysine 9 trimethylation), as well as the background 'Input' sample to obtain histone modification profiles across genes and topologically associated domains. No normalization was needed for summarizing across 1Mb genomic regions.

Only tissues with more than five modifications were included in the development of the COO prediction model, resulting in a total of 101 cell types. Certain cell types were grouped into histologically related categories as follows: (i) brain group consisting of cells from fetal brain, adult brain regions, and neurospheres; (ii) immune cells comprising all cells involved in immune response, such as CD4 naive primary, CD4+ CD25+ CD127- Treg primary, and others; (iii) bone marrow including chondrocytes and bone marrow-derived stem cells; (iv) gastrointestinal mucosa comprising stomach, colonic, rectal, and duodenal mucosa; (v) gastrointestinal muscle including stomach, rectal, duodenal, and colon smooth muscle; (vi) muscle including skeletal muscle and muscle satellite cultured cells.

3.2 Mutational landscape exploratory analysis

I calculated the number of SNVs, indels, and SV classes per patient within each cancer type and cohort, separately for WGS and WXS data. For SNVs, I calculated the number of transversions (A>C, A>T, G>C, G>T) and transitions (A>G, G>A) per patient within each cancer type and cohort. Indels were further analyzed and separated based on their type; insertion or deletion. Pearson's correlation coefficients between SNVs and indels from WGS and WXS data of the same patients were calculated. Statistical differences between cancer types, sequencing techniques, and cohorts were assessed using a two-sided Wilcoxon test and the Kruskal-Wallis test. Additionally, the chi-square test was employed to assess significant differences between specific types of indels and SV classes. Correction for multiple hypothesis testing was performed using Benjamini-Hochberg correction.

3.2.1 Mutational signature calling and evaluation

I calculated the absolute and relative contributions/exposures of mutational signatures for single-base substitutions (SBS) from individual patients, separately for breast, liver, and skin cancer. I used the 78 COSMIC SBS mutational signatures available in the R package Palimpsest(Shinde et al., 2018) and four different mutational signature tools with signature refitting methods: MutationalPatterns, signature.tools.lib, Palimpsest, and mutSigExtractor in R. For signature.tools.lib and mutSigExtractor, I used the default parameters, with the exception of mutSigExtractor, for which I applied strict signature refitting with a default cutoff value of 0.004. In the case of the Palimpsest tool, I used two functions with their default settings: the deconvolution_fit function for signature refitting and the signature_origins function for probabilistic assignment of mutational signature origins to mutations using simple Bayesian statistics (Letouzé et al., 2017). To evaluate the performance of each tool, I calculated the reconstruction error using COSMIC mutational signatures and exposures. The reconstruction error measures the quality of mutational profile reconstruction based on the mutational signatures identified through the learning process. I calculated the reconstruction error using cosine similarity and relative root-mean-square error (RMSE) metrics with the following equations:

a) Cosine similarity equation:

$$sim(A,B) = \alpha = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A is the original mutational profile vector, B is the reconstructed mutational profile vector and n is the number of mutation types defined as immediate $5'$ and $3'$ sequence context.

b) RMSE equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (A_i - B_i)^2}$$

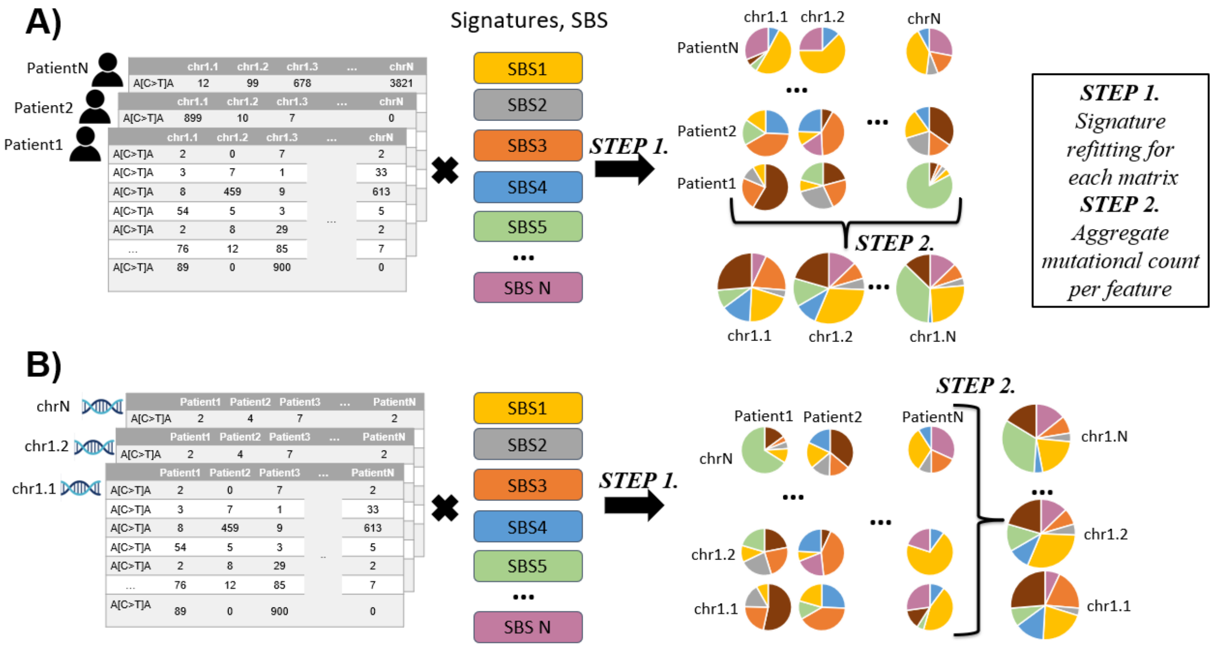
where A is the original mutational profile vector, B is the reconstructed mutational profile vector and n is the number of mutation types defined as immediate $5'$ and $3'$ sequence context.

The cosine similarity has a value between 0 and 1 because both A and B matrices are non-negative. If cosine similarity is 1, then two mutational profiles are identical, and if the cosine similarity is close to zero then two profiles are independent. On the other hand, RMSE can take any value from 0 to infinity and lower values indicate a better reconstruction. I calculated both reconstruction error metrics for the entire cohort of breast, liver, and skin cancer, as well as separately for each individual patient using WGS and WXS data. I considered the tool with the highest cosine similarity and lowest RMSE as the best-performing one. The results from the best-performing tools were visualized as the proportions of each mutational signature for aggregated breast, liver, and skin melanoma cancer WGS and WXS profiles. Additionally, the proportions of mutational signatures were analyzed for individual patients within each cancer type.

Indel mutational signatures were also calculated using 17 COSMIC indel signatures from R package Palimpsest and the signature refitting method using the best tool identified with calling SBS mutational signatures. Reconstruction error was calculated as cosine similarity between the original and reconstructed indel context.

3.2.1.1 Mutational signature calling per genomic feature

Since existing tools cannot calculate the absolute and relative contributions of mutational signatures for specific genomic features, I developed two additional approaches by modifying the original input matrix to calculate mutational signatures per 1 Mb genomic window, TADs, and genes using all tools. In the first approach, I analyzed individual patients one by one, generating a new matrix using the mutational context per genomic feature and calculating the mutational signature estimates per feature within each patient. This approach is referred to as method A (Figure 5A). In the second approach, referred to as method B (Figure 5B), I utilized the mutational context for specific features one by one across all patients and calculated the mutational signature estimates per patient for each individual feature. Mutational signatures were then summarized by genomic features of interest.



A)
PatientN
Patient2
Patient1
chr1.1 chr1.2 chr1.3 ... chrN
A[C>T]A
Signatures, SBS
SBS1
SBS2
SBS3
SBS4
SBS5
...
SBS N
STEP 1.
STEP 2.
chr1.1
chr1.2
chrN
chr1.N
STEP 1. Signature refitting for each matrix
STEP 2. Aggregate mutational count per feature
B)
chrN
chr1.2
chr1.1
Patient1 Patient2 Patient3 ... PatientN

Figure 5. Schematic representation of designed and adjusted mutational calling approaches to determine the mutational exposures on certain genomic features

I selected 2182 1 Mb genomic regions excluding blacklisted regions, as well as all genes and topologically associated domains from the HMEC cell line, identified using GMAP at a 50 kb resolution, to calculate mutational signatures for these features. For Method A, I calculated the reconstruction error for each patient and then determined the mean error for each feature by averaging the errors within each patient. For Method B, I calculated the reconstruction error for each feature and then determined the mean error for each patient by averaging the errors within each feature. I compared the reconstruction errors as cosine similarities between the best-performing tools: mutSigExtractor and Palimpsest (using the origin setting) with those of

Methods A and B. Significant differences between the tools were assessed using the Kruskal-Wallis test.

3.2.2 *De novo* structural variants (SV) signatures

Structural variant (SV) signatures were extracted using non-negative matrix factorization (NMF) with the Palimpsest tool for each cancer type separately. The parameters of Palimpsest were set to extract the optimal number of signatures using the "brunet" method, ranging from 1 to 10 signatures with 20 runs to ensure stability and avoid local minima for all SVs. Subsequently, I performed deconvolution with the optimal number of signatures determined by the "brunet" method to calculate the absolute number and proportion of each *de novo* SV signature in each patient.

Given that SV signatures are generally not well annotated or explored across different cancer types, I assessed the similarity of the extracted *de novo* SV signatures to the well-annotated rearrangement signatures from COSMIC. To facilitate this comparison, I modified the SV classification in the Palimpsest tool to include only 32 out of the 38 features present in the SV classification provided by COSMIC database. Cosine similarity was then calculated between the *de novo* SV signatures and the annotated signatures from COSMIC, as well as among *de novo* SV signatures across breast, liver, and skin melanoma cancer types.

To determine the most probable *de novo* SV signature origin for each structural variation, I used the Palimpsest mutational signature origin assignment. For calculating the percentages of SV signatures for each patient, I excluded samples with fewer than 10 SVs per patient to ensure robustness.

3.2.3 Kataegis and SV-hotspots

I identified local hypermutation sites, known as kataegis, by detecting SNVs with inter-mutational distances of less than 6 base pairs in both WGS and WXS datasets. This was achieved using the rainfallPlot function from the R package maftools (Mayakonda et al., 2018) on the human reference genome hg19. For each cancer type, I summarized the number of kataegis regions per patient and the number of mutations per kataegis region per patient. I calculated the proportion of mutation types, specifically transitions and transversions, detected in kataegis regions for each cancer type. Additionally, I analyzed the proportions of various mutational signatures assigned to mutations within kataegis regions, which were previously identified using the Palimpsest tool with the signature_origins function.

Structural variant hotspots were called with the SV-HotSpot tool (Eteleeb et al., 2020). SV-HotSpot uses a peak calling algorithm to identify regions with elevated frequency of SVs referred to as SV hotspots or peaks. It does that by counting samples harboring SVs overlapping with sliding windows over each chromosome. Next, peakPick peak calling algorithm is employed to detect windows ("peaks") where counts exhibit a significant increase compared to the surrounding windows. I set the default threshold of peaks occurring in at least 10% of SV samples to be identified as peaks. Subsequently, SV-HotSpot groups adjacent peaks with similar sample counts by applying a peak merging algorithm. First, the peak merging algorithm identifies clusters of adjacent peaks where any contiguous peaks are within a 10 kb distance per default. Afterwards, the algorithm selects the top peak within a cluster of peaks based on the highest sample count and proceeds to merge adjacent peaks upstream and downstream until it identifies k peaks displaying a significant change in sample counts compared to the top peak. The default parameter delta, set at 5%, is used for this purpose. The whole process is repeated until there are no more peaks in the cluster. The final peaks I used for downstream analysis are those merged peaks. I ran the SV-Hotpots tool on each separate group of SVs belonging to certain *de novo* identified SV clusters. The results from each individual run were then aggregated for downstream analyses.

3.3 Defining various gene subsets

I downloaded all human genes version GRCh37 from the Ensembl database using the Biomart (Durinck et al., 2005) R package. I consolidated all the gene models to eliminate alternative transcripts and establish a comprehensive set of non-overlapping exons for each gene. Subsequently, the regions between these exons were identified as introns for each gene. I calculated the proportion of SNVs in exons and introns separately for WGS and WXS data. For patients containing both types of sequencing data, I calculated Pearson's correlation coefficient between total number of WGS and WXS mutations. Fisher's exact test or Chi-square was used to test the significant difference in the enrichment of mutations between introns and exons, as well as between different cohorts, tissue types, and sequencing technologies.

3.3.1 Tissue-specific expressed genes in normal tissues

I determined a list of tissue-specific expressed genes for melanocytes, liver and breast cells using two main approaches; (a) predefined and literature-supported Gene Ontology (GO) terms that are known to only be active in a specific tissue, (b) applying tissue-specificity metrics on normal RNA-seq tissues. For the first approach, I used the R package clusterProfiler (Yu et al., 2012) to download relevant GO terms. Specifically, I identified 117 genes involved in xenobiotic metabolism (GO:0006805), which are known to be liver-specific. Additionally, I selected 14 genes involved in the melanin biosynthetic process (GO:0042438) and 23 genes involved in melanocyte differentiation (GO:0030318), both of which are specific to melanocytes. Furthermore, I included 28 genes involved in mammary gland development (GO:0030879), which are active only in breast tissue.

For the second approach, I calculated tissue-specificity metrics on the following normal RNA-seq tissues: publicly available GTEx RNA-seq (Lonsdale et al., 2013) files from 53 and 30 normal tissues which were TPM (Transcripts Per Kilobase Million) or RPKM (Reads Per Kilobase Million) normalized, respectively; and Fagerberg et al. (2014) dataset which contained 12 different normal tissues. The calculation was performed the same way in paper by Kryuchkova-Mostacci and Robinson-Rechavi (2017). Preprocessing of the RNA-seq data was done using the following steps:

1. All the genes with expression <1 RPKM or TPM were set as not expressed.

2. The RNA-seq data were $\log_2$-transformed.

3. After the $\log_2$ normalization, a mean value from all replicates for each tissue separately was calculated.

4. All genes that were not expressed in at least one tissue were removed.

After preprocessing the RNA-seq datasets separately, I calculated tissue-specificity metrics which are divided in two groups. The first group provides a single number to indicate whether a gene is tissue-specific or expressed ubiquitously (Tau, Gini, TSI, Counts, and Hg). In contrast, the second group presents information on the gene's specificity for each tissue individually (z-score, SPM, EE, and PEM). To facilitate comparisons with the first group, I utilize the maximum specificity value obtained from the second group the same way as Kryuchkova-Mostacci and Robinson-Rechavi (2017) did. The mentioned scores were calculated using the the following equations:

1. Tau index:

$$\tau = \frac{\sum_{i=1}^{n} (1 - \widehat{x_i})}{n-1} ; \widehat{x_i} = \frac{x_i}{max(x_i)}$$

2. The EE score:

$$EE = \frac{x_i}{\sum_{i=1}^{n} x_i * \frac{s_i}{\sum_{i=1}^{n} s_i}} = \frac{\sum_{i=1}^{n} s_i}{s_i} * \frac{x_i}{\sum_{i=1}^{n} x_i} ;$$

where s_i is the summary of the expression of all genes in i

3. The Gini coefficient:

$$Gini = \frac{n+1}{n} - \frac{2\sum_{i=1}^{n} (n+1-i)x_i}{n\sum_{i=1}^{n} x_i} ;$$

where x_i has to be ordered from least to greatest

4. Hg scores:

$$H_g = -\sum_{i=1}^{n} p_i * log_2(p_i) ; p_i = \frac{x_i}{\sum_{i=1}^{n} x_i}$$

5. The z-score was so only over-expressed genes are defined as tissue-specific which was then able to compare z-score with other methods.

$$z = \frac{x_i - \mu}{\sigma} ;$$

where μ is the mean of gene expression and σ is the standard deviation

6. PEM score:

$$PEM = log_{10}\left(\sum_{i=1}^{n} s_i * \frac{x_i}{\sum_{i=1}^{n} x_i}\right)_{\blacksquare} ;$$

where s_i is the summary of the expression of all genes in i

The output of all calculated scores was modified to the same scale from 0 (ubiquitous) to 1 (tissue-specific) to be able to compare them (Table 6). Four of the methods calculate specificity value for each tissue separately; for these methods, the largest (most specific) value among all tissues was assigned to the gene.

Table 6. *Tissue specificity parameters. N is the number of tissues in the data set. Taken and adjusted from Kryuchkova-Mostacci and Robinson-Rechavi (2017). X = max xi: is the maximal specificity value for a certain gene among all tissues.*

Methods	**Tissues**	**Ubiquitous**	**Specific**	**Transformation**
τ (tau)	all	0	1	-
Gini	all	0	(N - 1) / N	x * (N / (N - 1))
TSI	all	0	1	-
Counts	all	N	1	(1 - x / N) * (N / (N - 1))
EE_i	separately	0	> 5	X / max X
Hg	all	$\log_2 N$	0	1 - x / $\log_2 N$
Z score	separately	0	> 3	X / n − 1 / √N
PEM score	separately	0	~1	X / max X
SPM	separately	0	1	X

To validate the tissue-specificity determined in these datasets, I employed GO terms similar to those used by Kryuchkova-Mostacci and Robinson-Rechavi (2017). The GO terms were retrieved using the clusterProfiles package and included: spermatogenesis (GO:0007283), which is specific to testis and consists of 469 human genes; neurological system process (GO:0050877), specific to brain and neural tissues, including 1,338 genes; xenobiotic metabolic process (GO:0006805), specific to liver and kidney, with 163 genes; protein folding (GO:0006457), expected to be ubiquitous and involving 231 genes; membrane organization (GO:0061024), also ubiquitous, encompassing 607 genes; RNA splicing (GO:0008380), another ubiquitous process involving 383 genes; and additional melanin-producing genes (GO:0042438). These GO terms were used to assess and confirm the specificity of gene expression across corresponding tissues, thus ensuring the accuracy of the tissue-specific gene identification.

Afterward, I applied a more rigorous approach for assigning genes to multiple tissues for specific expression using extended Tau defined by Lüleci and Yılmaz (2022). The extended Tau method refines the traditional Tau score, which measures tissue specificity by normalizing gene expression levels. The Tau score was calculated using normalized gene expression values across tissues. The extended method integrates this score with a statistically significant distance from the maximum expression value, determined through the standard deviation of non-zero expression values and an optimized Z-value threshold. This statistical distance was calculated as the maximum expression value minus the product of the standard deviation and the Z-value shown by the following formula:

$$distss=xmax-\sigma\times zval$$

where xmax is the maximum expression value of a gene among all tissues, σ is the standard deviation of non-zero expression of a gene among all tissues and zval is optimized threshold as Z-value. This integration allows for a more accurate assignment of genes to multiple tissues by considering both the specificity and the statistical robustness of the expression data shown on schematic representation on Figure 6.

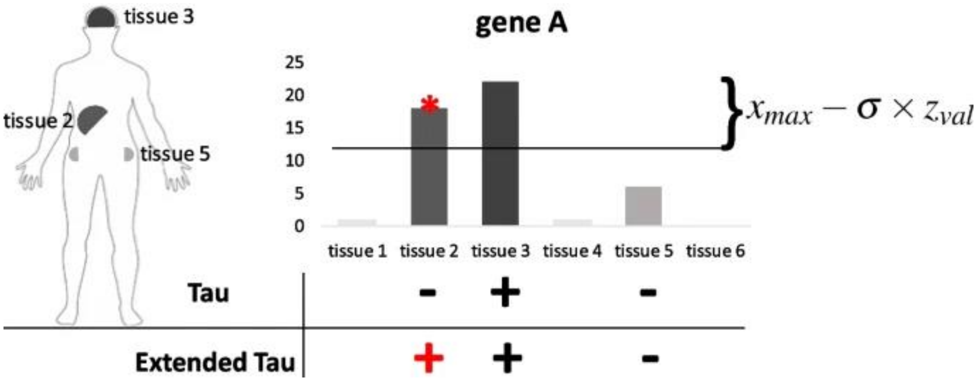


Figure 6. *Schematic representation of how extended Tau score is used to determine tissue specific genes. Taken from Lüleci and Yılmaz (2022).*

To further validate the obtained tissue-specific genes, I examined well-annotated liver-specific alpha-fetoprotein (*AFP*) and kidney-specific D-amino acid oxidase (*DAO*) in the obtained results. Genes found to be specifically expressed in breast, liver, and skin tissues across multiple datasets were used to define subsets of tissue-specific genes corresponding to each cancer tissue type. The overlap of the final defined tissue-specific gene sets for each cancer type was visualized using Upset plot.

3.3.2 Most mutated genes

To identify the genes with the highest mutation frequency, I defined them as those mutated in the majority of individual samples. I ranked these genes based on the number of samples in which mutations were detected and selected a specific percentage of the top-ranked genes. Various percentages were considered, such as 5%, 10%, 20%, 30%, 40%, and 50%, to identify the most frequently mutated genes. This selection process was performed for both all genes and exclusively protein-coding genes in breast, liver, and skin cancer types, as well as for each individual cohort. I then overlapped the top 40% of the most mutated genes found in each cancer type and identified the unique intersection of all three. For downstream analysis, I characterized all percentages of the top mutated genes found in each cancer type, cohort, and the overlapping intersection of the top 40% most mutated genes. Initially, I examined the biological pathways they were predominantly enriched for. I conducted an over-representation analysis of biological processes (BP) from Gene Ontology (GO) terms using the clusterProfiler R package. To reduce redundancy of enriched GO terms, I used the simplify function with the following settings: cutoff = 0.4, by = "qvalue", and select_fun = "min". Pairwise similarities of the reduced enriched GO terms were calculated using Jaccard's similarity index (JC) within the pairwise_termsim function, and the terms were hierarchically clustered with Ward's method using the treeplot function. Additionally, I analyzed the expression of overlapping top 40% protein-coding genes from all three cancers in normal tissues using GTEx TPM normalized data of 30 tissues and performed unsupervised hierarchical clustering with Euclidean distance of the top mutated genes.

3.4 Correlation analysis of mutations and chromatin features over genetic features

I counted the number of mutations in each genomic feature (gene, 1Mb region, TADs) using a custom R script. All counts were normalized to the feature length. As for histone modifications, I calculated RPKM for genes, 1 Mb regions and TADs using the following equations:

$$RPKM = \frac{C}{\boldsymbol{N*L}} * 10^9;$$

where *C* is the number of reads that map to a particular genomic feature (gene, 1 Mb region or TAD), *N* is the total number of mapped reads in the sample (in millions) and *L* is the length of the genomic feature in kilobases.

The correlation between various log normalized mutational profiles on genomic feature and all histone modifications of normal tissues were calculated using Spearman's rank correlation coefficient in all downstream analysis using the following equation:

$$\rho = 1 - \frac{6\sum\ d_i^2}{n(n^2-1)};$$

where d_i is the difference between the two ranks of each observation and n is the number of observations.

3.5 Machine learning methods for development of cell-of-origin model

I developed a model for predicting the cell-of-origin (COO) using mutational profiles and histone modifications of normal tissues across various genomic features: 1 Mb regions, TADs and genes. The COO model utilized 10-fold cross-validation to predict the mutational density profiles in various genomic features using chromatin profiles on aggregate mutational profiles per cancer type and per cohort and individual mutational profiles of patients. Mutation counts were normalized to feature length and log transformed alongside histone normalized values (RPKM) prior to modeling. I trained all distinct models for each mutation density profile, utilizing the chromatin profiles specific to tissue types. The model demonstrating the highest variability signifies the COO for a specific cancer type. The general predictive performance of each model was assessed by calculating the average R^2 value, which measures the similarity between predicted and observed profiles, across ten sets of windows. The modeling process is shown in Figure 7. I applied all of the different models using various genomic features to each individual sample and determined the proportion of individual patients where the best-matching model matched the presumed cell-of-origin of each cancer type.

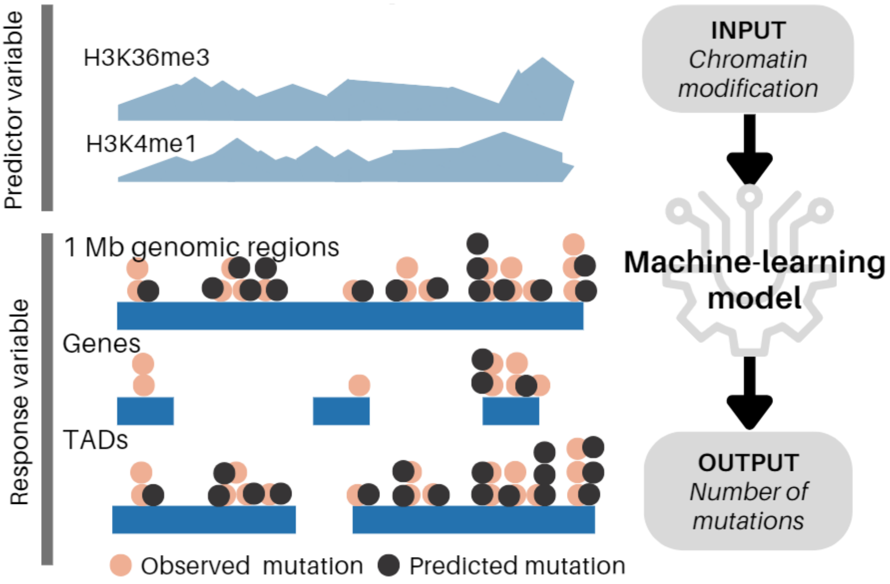


__Figure 7.__ Schematic representation of cell-of-origin (COO) model across various features. The response variable comprises aggregated mutation profiles from all patients within each cancer type, or the mutational profiles of individual patients. The predictor variables are RPKM values of histone modifications for specific genomic features within the normal tissue epigenome. This model leverages the correlation between the cancer mutational landscape and normal tissue epigenetic marks to predict the tissue of origin

To begin with the modeling process, I designed a preliminary model using multiple linear regression to identify the subsets of various genomic features affected by different genomic features, such as mutational signatures, kategis and others, for which model performance is the best. Additionally, for gene-based COO models, I developed three model settings scenarios: (a) coding sequence (CDS) model setting where the mutational counts were summed and normalized by CDS, (b) Gene model where I considered only exonic mutations and normalized by gene length, (c) Gene model with both exonic and intronic mutations normalized by gene length. By doing so, I tested for possible biases in prediction accuracies of gene-based models that depend on the intronic and exonic regions.

The best model performance was assessed based on the COO model's result of aggregated profiles and proportion of individual patients with correctly identified COO. For each genomic feature in the preliminary COO model, I calculated a standardized residual to identify outliers. Standardized residual was calculated as follows:

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{RSE\sqrt{1 - h_{ii}}} ;$$

where: e_i is the i$^{\text{th}}$ residual, RSE is the residual standard error of the model, h_{ii} is the leverage of the ith observation.

Outliers with standardized residual below -2 were annotated as over-predicted, while those with standardized residual above 2 were annotated under-predicted. Annotated outliers from breast, liver and skin cancer were overlapped using the UpsetR package. Identified outliers were tested for enrichment of different genomic features such as enriched regions or genes with kataegis and SV hotspots using Fisher’s exact or Chi-square test depending if the expected observations were all above 5. Significantly enriched outliers were then removed from the original model to see if the model's performance would improve. Also, I assessed the enrichment of outliers with genes from Cancer Gene Census, TIME and tissue-specific super-enhancer from both normal and tumorigenic tissue/cell-line from SEdb2.0. Over-representation analysis (ORA) of Gene Ontology (GO) terms, hallmarks of cancer and disease-association database (DisGeNET) (Piñero et al., 2017) was also used to characterize the outliers to identify pathways more associated with each specific group of outliers in its cancer type. As well I examined the expression of all and tissue-specific genes from GTEx TPM 30 database of appropriate tissues in each outlier group. Wilcoxon-test with Benjamini-Hochberg correction was used to test if there is a significant difference in expression of those genes between annotated outliers. As for SBS mutational signatures, I calculated Pearson’s correlation of the number of each signature per outlier and absolute value of standardized residual referred to as erroneous prediction rate. Calculated correlations were compared to the baseline Pearson’s correlation of all mutations and standardized residual. For TADs and genes, I calculated the overlap with of annotated TADs by Akdemir et al. 2020. based on active and inactive state (heterochromatin, low, low-active and repressed).

For each affected region with SV-hotspots and/or kataegis, as well as annotated outliers, I calculated the Spearman’s correlation of number of mutations to all histone marks from the expected or correct normal cell-of-origin tissue for each cancer type. Moreover, I did the same correlation analysis of normal epigenomes with SBS mutational signatures which were most abundant in cancer features, top 5, or were identified to have higher positive correlations with erroneous prediction rate.

Once I established a multiple linear regression model with the highest performance for genomic regions, genes and TADs separately for each cancer type, I applied random forest and extreme boosting machine learning methods with 10 cross-validation to see if the model's performance would significantly improve with the same features. Random forest model was done using 500 trees from ranger package (Wright & Ziegler, 2017), while extreme boosting was done with parameter settings/hypertuning of nrounds=200, max_depth=2, eta = 0.01, gamma = 0.01, colsample_bytree = 0.75, min_child_weight = 0, subsample = 0.9 using xgbTree. All of the machine learning methods were called with caret package (Kuhn, 2008) in R.

Afterwards, I took the best COO model based on the accuracy of prediction on aggregated and individual patients to assess how different patients’ genomic features, such as number of mutations, explained variance by the used COO model, number of kataegis and SV-hotspots per patient, affect the model. Moreover, I examined how patient characteristics, listed in chapter *3.1.2 Characteristics of individual patients*, affect the predictions by examining the correctly and incorrectly identified COO patients and their features.

3.6 Variant calling pipeline from RNA-seq

3.6.1 Datasets

I have downloaded the restricted access STAR-2 passed BAM file aligned to hg38 for 69 samples from only melanoma tumor skin in TCGA datasets. From Gene Expression Omnibus (GEO), I downloaded and analyzed 12 hepatocellular carcinoma patients with their non-adjacent normal tissue from GSE105130 and 8 breast cancer tissue alongside normal tissue from GSE229571. Only downloaded data from GEO had both tumor and normal tissue samples.

3.6.1 Pipeline for calling RNA-seq mutations

The GEO dataset quality was assessed using the FASTQ tool (v0.11.5) (Van der Auwera et al., 2013). Afterwards the reads were trimmed and filtered using the Trimmomatic tool (v0.32) (Bolger et al., 2014) with following parameters: Seq_adapters.fasta:2:30:10 SLIDINGWINDOW:5:10 MINLEN:60 HEADCROP:5. I aligned the trimmed reads to hg38 using a two-STAR pass mapper (v2.7.3) (Dobin et al., 2012). Sambamba tool (v0.6.1) was used to sort and index the BAM output.

Downloaded TCGA BAM files needed to be filtered to keep only autosomal and sex chromosomes which was done using custom bash scripts using GATK tool BedToIntervalList. Afterwards, I used the standard GATK pipeline (v4.3.0.0) to prepare the alignments for calling the mutations from both TCAG BAM and GEO generated BAM files. Duplicates were marked with Picard tool MarkDuplicates. I added read group information to each sample using Picard AddOrReplaceReadGroups, afterward the bam file was reindexed using Picard BuildBamIndex. For handling splicing events in RNA-seq data, I used GATK SplitNCigarReads. Final readjusting of scores and quality processes were done using BaseRecalibrator and ApplyBQSR from the GATK pipeline.

For TCGA skin melanoma RNA-seq, I used only Mutect2 tool that was a part of GATK tool v4.3.0.0 tumor only mode to call somatic mutation since normal tissues were not available. Breast and liver sample mutations were called using Varscan (v2.4.6), Strelka2 (v2.9.10) and Mutect2 tools. Mutect2 analysis included the use of The Panel of Normals (PoN) (1000g_pon.hg38.vcf.gz) and germline resources (af-only-gnomad.hg38.vcf.gz) following GATK Best Practices to keep somatic tumor mutations. For breast and liver samples, I only kept the mutations which were detected by at least two used tools for downstream analysis. Afterwards, the filtered SNVs were converted to hg19 using the R package liftOver (v1.26.0). Additional filtering of obtained variants was done by removing RNA editing sites listed in the DARNED (Kiran & Baranov, 2010) and RADAR (Ramaswami & Li, 2014) databases using custom script in R.

3.6.2 Comparison with WGS and WXS data

For RNA-seq-derived mutations across all cancer types, I compared the mutation profiles, including the abundance of mutations, transversions, and transitions, to those obtained from WGS and WXS. To evaluate the correlation between SNV counts detected by different technologies, I computed the differences in the number of mutations per patient across these technologies. Significant differences were assessed using the Wilcoxon test, with the Benjamini-Hochberg correction applied to account for multiple testing. Additionally, I analyzed patients with both WXS and RNA-seq mutation data for skin melanoma by calculating Pearson's correlation coefficients to examine the relationship between the SNV counts detected by each technology. The total number of transversions and transitions was calculated for each cancer type based on the mutations detected by the different NGS technologies. The differences in mutation types across technologies were evaluated using the Chi-square test.

I further analyzed the top N% of frequently mutated genes in the RNA-seq cancer mutational profiles. Due to the limited number of patients in the liver and breast cancer cohorts, I included all genes that were mutated in more than 50% of patients in these analyses. These genes were evaluated for their annotation in the Cancer Gene Census and Tumor Immune Microenvironment gene lists, as well as their presence among the top N% of mutated genes identified from WGS, and their association with super-enhancers. For skin melanoma, I identified the top 5%, 10%, 20%, 30%, 40%, and 50% most frequently mutated genes. I focused on the top 10% most mutated genes, conducting an over-representation analysis of Gene Ontology terms to determine the pathways in which these genes are involved. Additionally, I analyzed the overlap of the top 10% most mutated genes between RNA-seq and WGS to identify commonalities in the mutational landscapes captured by these technologies.

I investigated the types of genes most frequently affected by mutations identified through RNA-seq, comparing these findings with those derived from whole-genome WGS and WXS technologies. Additionally, I analyzed the enrichment of these mutations within intronic and exonic regions of the affected genes.

3.6.3 Cell-of-origin prediction of RNA-seq data

I used the most robust models developed from genome-wide mutational profiles to predict the COO for RNA-seq identified single-nucleotide variants, both at the level of aggregated profiles and for individual patients using Gene based (intron + exon) and CDS based normalizations. The gene subgroups employed for these RNA-seq predictions were consistent with those used in the WGS and WXS gene-based COO models. For the top N% of frequently mutated gene groups, I selected the genes identified from the WGS datasets to ensure comparability and accuracy in the analysis.

4 Results

4.1 Mutational landscape obtained by whole-genome and whole-exome sequencing

In this chapter, I analyzed the general characteristics of the mutational landscape in patients with breast, liver, and skin cancers. I investigated the overall distributions of SNVs, indels, and SVs per patient across independent cohorts and combined for each cancer type. Utilizing mutational signature software in R, I identified mutational signatures for single-base substitutions in individual patients and various smaller genomic features, including 1 Mb genomic windows, topologically associated domains and genes. Kataegis regions, clusters of hypermutations, were annotated based on SNV profiles and characterized according to their associated mutational signatures. Structural variants were also annotated by their mutational signatures, and specific SV-hotspots were identified and analyzed for their contribution to the overall mutational landscape.

4.1.1 Characterization of single-nucleotide variants (SNVs)

I characterized the breast, liver, and skin cancer cohorts by analyzing their SNV counts. Table 7 provides a summary of the total number of mutations detected for each cancer type. WGS data revealed that skin cancer exhibited the highest number of SNVs, approximately 30 million, while liver cancer had around 6 million SNVs, and breast cancer showed the lowest count with 4 million SNVs. In contrast, WXS data indicated that liver cancer had the highest number of SNVs among the three cancer types. Notably, the SNVs detected through WXS comprised only about 4.4% of those detected by WGS. This discrepancy highlights the more comprehensive mutational landscape captured by WGS compared to WXS, which is limited to coding regions.

Table 7. *Total number of SNVs detected in whole-genome (WGS) and whole-exome (WXS) sequencing for breast, liver and skin cancer*

Cancer type	Sequencing technology	Total number of SNVs
Breast	WGS	4585045
	WXS	324882
Liver	WGS	6514226
	WXS	993646
Skin	WGS	30590540
	WXS	542854

The highest number of single-nucleotide variants per patient, as detected by whole-genome sequencing, was observed in skin cancer, with a mean of 95,595 and a standard deviation of 159,113. This was followed by liver cancer, which had a mean of 12,798 SNVs and a standard deviation of 29,389, and breast cancer, which had the lowest mean SNV count of 5,901 with a standard deviation of 9,358. Notably, cohorts from the TCGA, including BRCA-US, LIHC-US, and SKCM-US, exhibited significantly lower SNV counts per patient compared to other cohorts (Figure 8).

Intra-cancer type analysis revealed significant differences in SNV counts across cohorts for each sequencing technology. However, these differences were least pronounced in breast cancer WXS cohorts (Kruskal-Wallis test, p=0.044). All comparisons between individual cohorts within each sequencing technology demonstrated statistically significant differences for both WGS and WXS data (p-value < 0.05, two-sided Wilcoxon test, Benjamini-Hochberg correction for multiple hypothesis testing).

The SNV counts per patient obtained through WXS were significantly lower than those obtained via WGS, with the exception of TCGA cohort SKCM-US (p-value=0.316, Wilcoxon test). Pearson's correlation coefficients between WGS and WXS data from the same patients were significantly positive across all cancer types, indicating a strong concordance: 0.987 (p-value=$1.6*10^{-67}$) for liver cancer, 0.997 (p-value=$1.3*10^{-118}$) for breast cancer, and 0.907 (p-value=$9.7*10^{-15}$) for skin cancer.

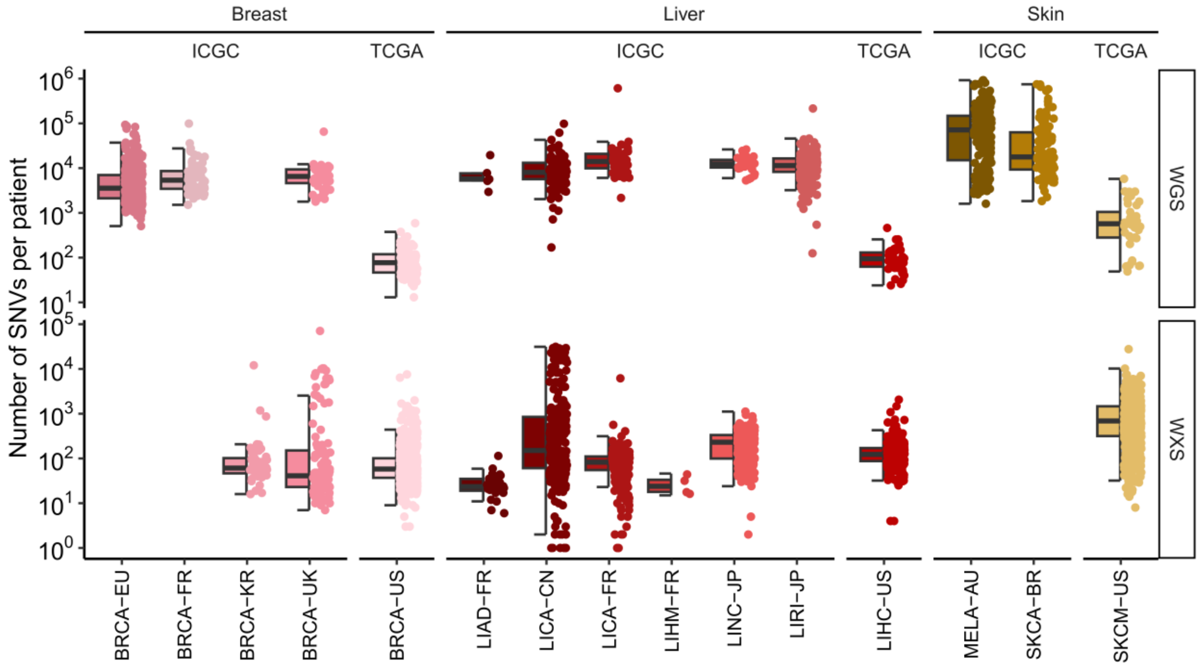


Figure 8. *Distribution of single-nucleotide variants (SNVs) per patient across independent breast, liver and skin cancer cohorts obtained by whole-genome (WGS) or whole-exome sequencing (WXS). Plot shows log_{10} transformed values of mutations. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

The analysis of transversions (Tv) and transitions (Ti) per patient revealed consistent trends across different cancer types, with skin cancer exhibiting the highest counts for both types of mutations (Figure 9A). This trend was particularly evident in C>T and G>A transversions, where skin cancer displayed significantly higher mutation counts than other cancer types, as observed in both WGS (Kruskal-Wallis test, p-value = $1.9*10^{-163}$) and WXS data (Kruskal-Wallis test, p-value = $7.4*10^{-313}$). Overall, breast cancer showed the lowest number of transitions and transversions per patient among the analyzed cancer types, while skin cancer had the highest (Figure 9B). Transitions were more abundant than transversions across all cancer types and sequencing technologies, reflecting a common mutational pattern in cancer genomes.

Further examination of cohort-specific differences within each cancer type revealed that cohorts associated with TCGA and those analyzed using WXS demonstrated significantly lower counts of transitions and transversions (Supplementary Figure 1). This reduction aligns with the overall smaller number of SNVs observed in these cohorts.

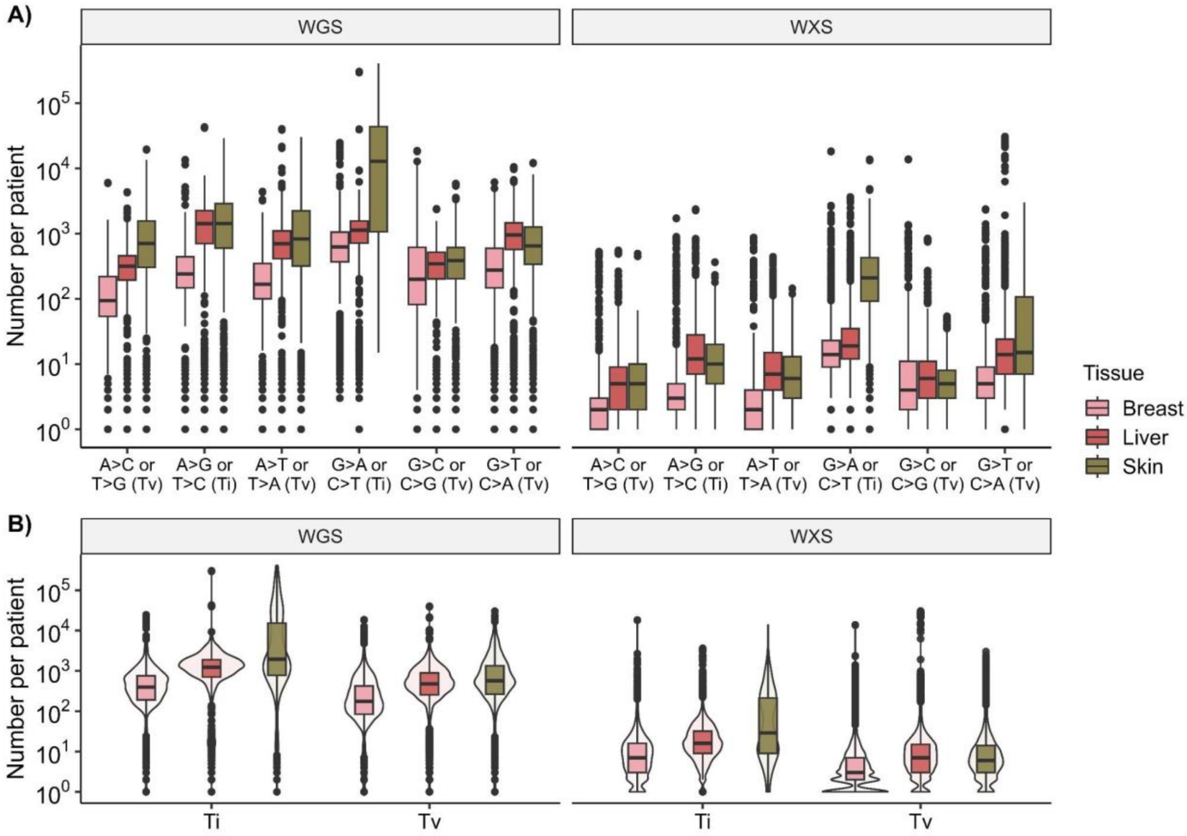


Figure 9. *A) Distribution of various transversions (Tv) and transitions (Ti) in aggregated mutational profiles breast, liver and skin cancer types. B) Distribution of aggregated Tv and Ti. Plots show log*$_{10}$ *transformed values of mutations. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

4.1.1.1 Single-base substitution mutational signatures

To determine mutational signatures in the samples, I used four different tools implemented in R: signature.tools.lib, Palimpsest, mutSigExtractor, and MutationalPatterns. The reconstruction error, calculated as cosine similarity per cancer type, indicated that all the tools exhibited high cosine similarity values above 0.985, suggesting low reconstruction error (Figure 10A). Among the different cancer types, breast cancer data obtained with WGS displayed the lowest cosine similarity in most tools, while the highest similarity was observed in WXS data for breast cancer across all tools.

Liver and skin cancer data obtained via WXS exhibited lower cosine similarities compared to breast cancer, but the values remained relatively high, all above 0.92. Overall, the cosine similarity was notably higher in WGS-generated data than in WXS data, with all WGS values well above 0.985 when examining the calculated score per cohort. The tool with the poorest performance was Palimpsest, particularly its signature refitting method and probabilistic assignment of mutational signature origins (Palimpsest_origin), where breast cancer showed the greatest increase in reconstruction error.

Upon examining the distribution of calculated cosine similarities for each individual patient, I found that WGS data consistently had better scores compared to WXS data (Figure 10B), with median values closer to 1. Once again, Palimpsest demonstrated the worst reconstruction error, with the lowest median and a standard deviation of 0.99152 ± 0.00562 for the default settings, and 0.99165 ± 0.00312 for the Palimpsest origin setting. On the other hand, mutSigExtractor displayed higher cosine similarity than MutationalPatterns and signature.tools.lib, particularly with stricter refitting criteria.

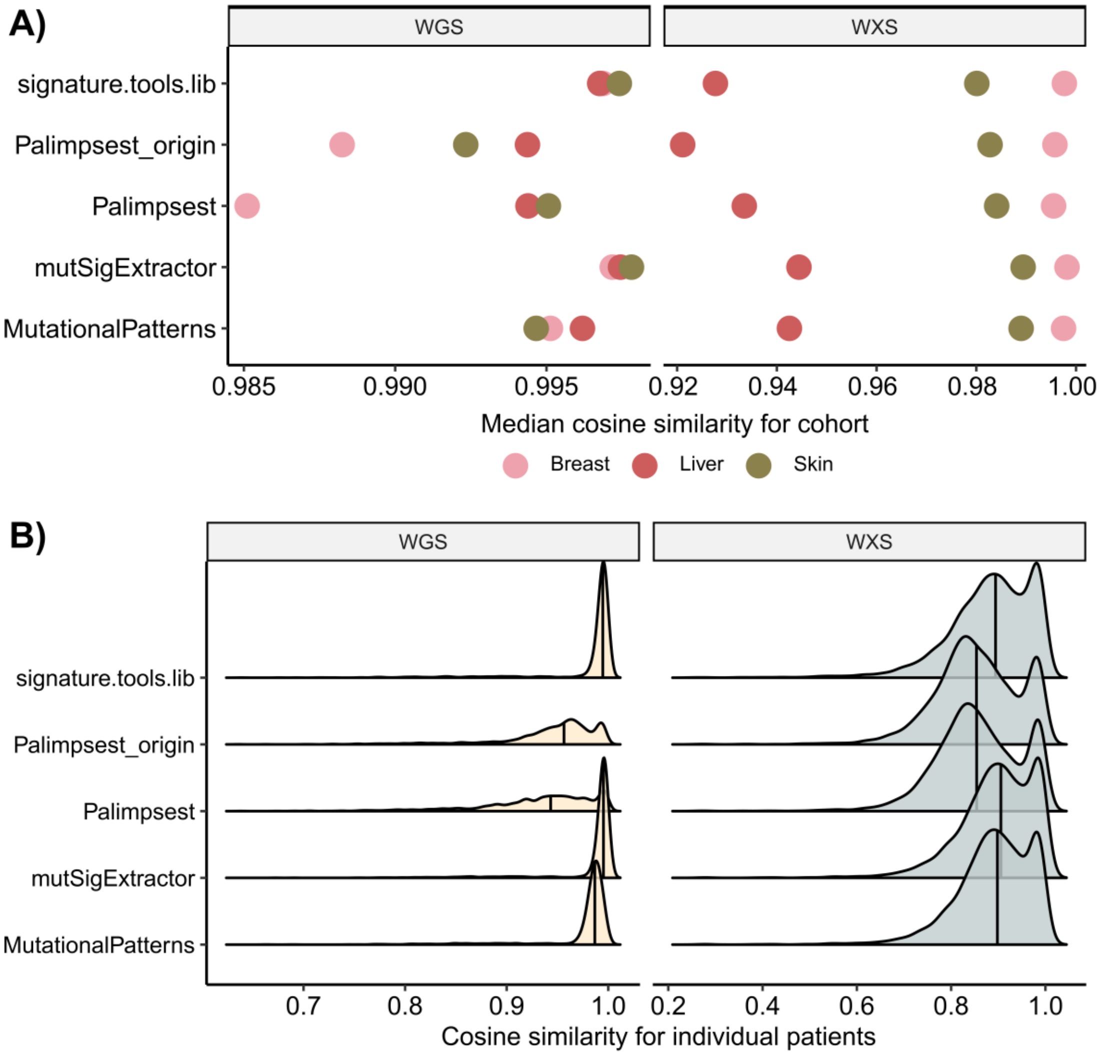


Figure 10. *A) Comparison of tools for calling mutational signatures (signature.tools.lib, Palimpsest origin and signature refitting, mutSigExtractor, MutationalPatterns) based on reconstruction error calculated as cosine similarity separated by cancer type. B) Distribution of cosine similarity per individual patient of 5 different mutation calling tools separated by WGS and WXS data shown as a density function.*

The root mean square error (RMSE) as an additional reconstruction error metric confirmed similar findings as cosine similarity, where Palimpsest was the worst performer with a mean and standard deviation of 154 ± 496. In contrast, mutSigExtractor emerged as the best-performing mutational signature calling tool, with an RMSE of 55 ± 225. These results highlight the superior performance of WGS data in capturing accurate mutational signatures and underscore the variability in the effectiveness of different tools, with mutSigExtractor demonstrating particularly robust performance.

Since the mutSigExtractor tool showed the best performance calling the mutational signatures, I conducted all of the downstream analysis describing the mutational signature compositions of each cancer type using this tool. For breast cancer, the most prevalent mutational signatures identified in both WGS and WXS data were SBS13 and SBS2 (Figure 11A), contributing approximately 30% and 36% of the total mutations, respectively. In skin melanoma, more than 88% of mutations in WGS data were attributed to well-annotated UV-induced signatures, SBS7a-d, with SBS7a alone dominating the mutational landscape at around 64%. Even though the WXS data for skin melanoma showed a reduction in the detection of these signatures compared to WGS data, UV signatures still constituted over 57% of the mutational profile.

For liver cancer, the WGS data revealed a diverse mutational landscape where signatures SBS12, SBS40, SBS23, SBS93, SBS8, SBS16, and SBS24 collectively comprised approximately 52% of the mutational signatures. In contrast, SBS29 was the most predominant in liver WXS data, accounting for about 50% of the mutations. Notably, liver WXS data exhibited the most significantly different mutational profile compared to liver WGS data when compared to the differences observed in breast and skin cancer datasets (Chi-square test, p-value < 0.05).

When analyzing the proportions of mutational signatures in each cancer type on a per-patient basis, I observed that patients with a higher mutational burden had an increased number and percentage of specific mutational signatures (Supplementary Figure 2A-F). In breast cancer, both WGS and WXS data showed that higher mutation counts were associated with increased proportions of SBS2 and SBS13. For liver cancer, SBS29 was notably more prevalent in WXS data, while SBS12 was dominant in WGS data. Skin melanoma patients, regardless of sequencing technology, exhibited higher proportions of UV-induced signatures SBS7a and SBS7b.

Breast cancer patients had higher proportions of the age-related mutational signature SBS1 compared to liver or skin cancer patients (Figure 11B). Additionally, SBS2 and SBS13 were significantly more prevalent in breast cancer patients, as determined by post-hoc Dunn's test with Benjamini-Hochberg corrected p-values below 0.05. In skin melanoma, the majority of patients had very high proportions of SBS7a and SBS7b, consistent with the predominant UV-signature mutation patterns typically associated with this cancer type.

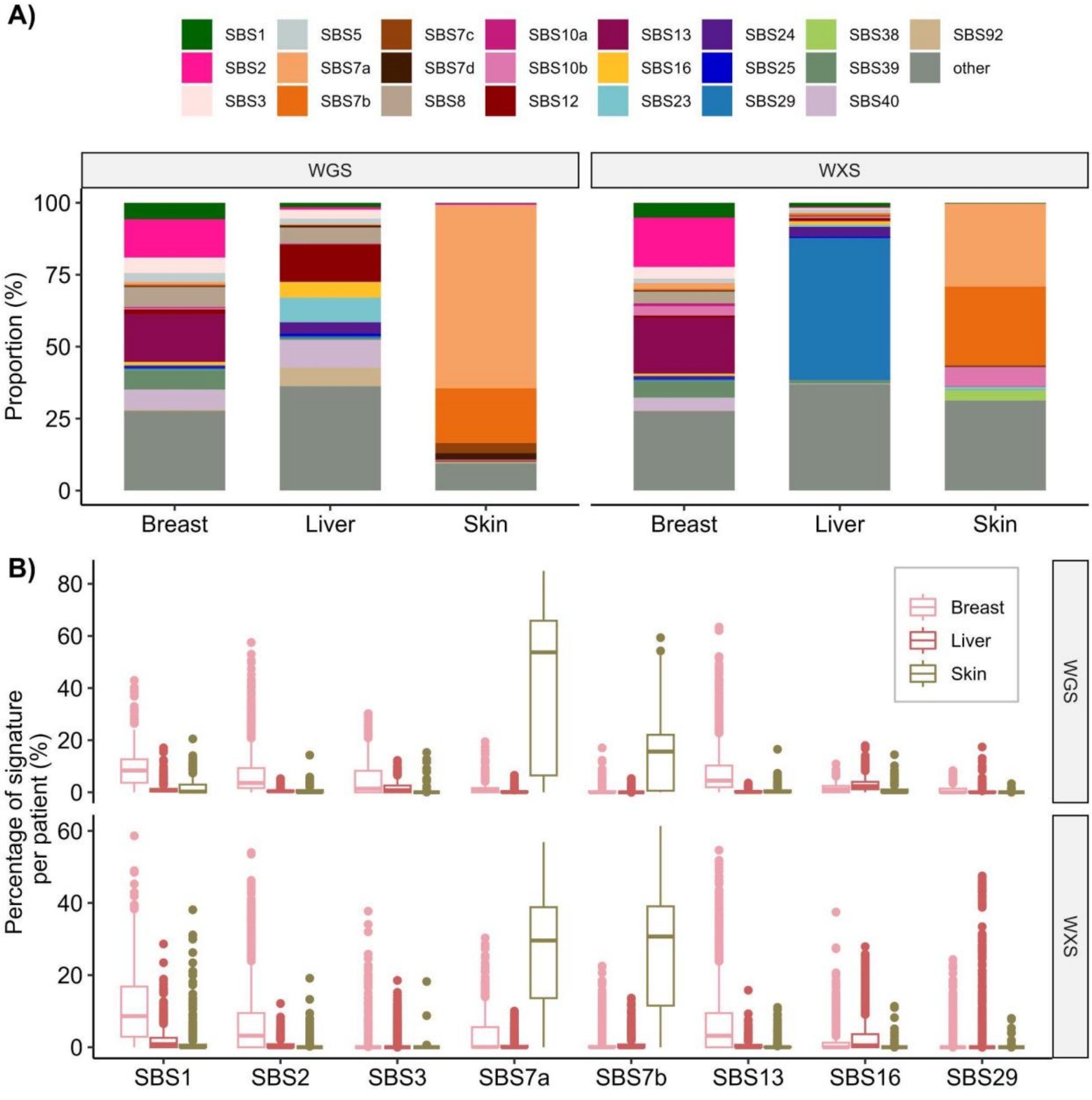


Figure 11. *A) Proportion of mutational signature determined with mutSigExtractor tool for breast, liver and skin cancer types separated by WGS and WXS data. Gray colors indicate signatures whose contribution to a particular cancer type was below 4000 mutations. B) Distribution of percentage of mutational signatures per individual patient. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

4.1.1.2 Evaluation of calling mutational signatures per genomic feature

Mutational signature calling software generally estimates the number of mutations per patient but often does not attribute specific mutations to distinct mutational signatures. An exception to this is the Palimpsest tool in its special setting mode, which attempts to assign a mutational signature to individual mutations. Although the Palimpsest origin setting exhibited a lower cosine similarity compared to other tools, the values were still notably high, exceeding 0.9, which suggests an adequate refitting of the mutational signatures. Despite this, I modified the mutSigExtractor tool, which demonstrated superior performance relative to the Palimpsest origin setting, to call mutations within smaller genomic regions to better capture mutational profiles on a finer scale. The modifications to mutSigExtractor were designed to enable mutational signature calling in two distinct ways: A) Calling mutations using the mutational profile of one patient whose profile was split into genomic features: 1 Mb genomic region, gene or TADs. B) Taking a mutational profile of all patients per cancer type and calculating the mutational signatures for each genomic feature separately.

Both modified approaches, calling mutational signatures using mutSigExtractor per specific genomic feature, resulted in significantly lower cosine similarity compared to the original mutSigExtractor calls and the Palimpsest origin setting (Figure 12). This indicates a higher reconstruction error when analyzing smaller genomic segments. Notably, calling signatures by genes yielded the lowest cosine similarity, suggesting it had the poorest reconstruction accuracy of all methods tested. In contrast, the 1 Mb regions and TADs showed more comparable reconstruction errors to the gene method setting, highlighting some consistency in their performance.

Additionally, methods A and B showed similar performance within each genomic feature, as evidenced by Wilcoxon test p-values of 0.17, 0.46, and 0.89 for 1 Mb regions, TADs, and genes respectively, when assessing aggregated cancer cosine similarity per genomic feature.

For subsequent analyses, I chose to utilize the mutational signatures assigned to individual mutations determined by the Palimpsest origin setting for quantifying signatures within specific genomic features. Despite its higher reconstruction error compared to calling signatures per genomic feature directly, the Palimpsest origin setting still provided a more accurate assignment of signatures to specific mutations, thereby offering a better framework for detailed mutational analysis in downstream analysis.

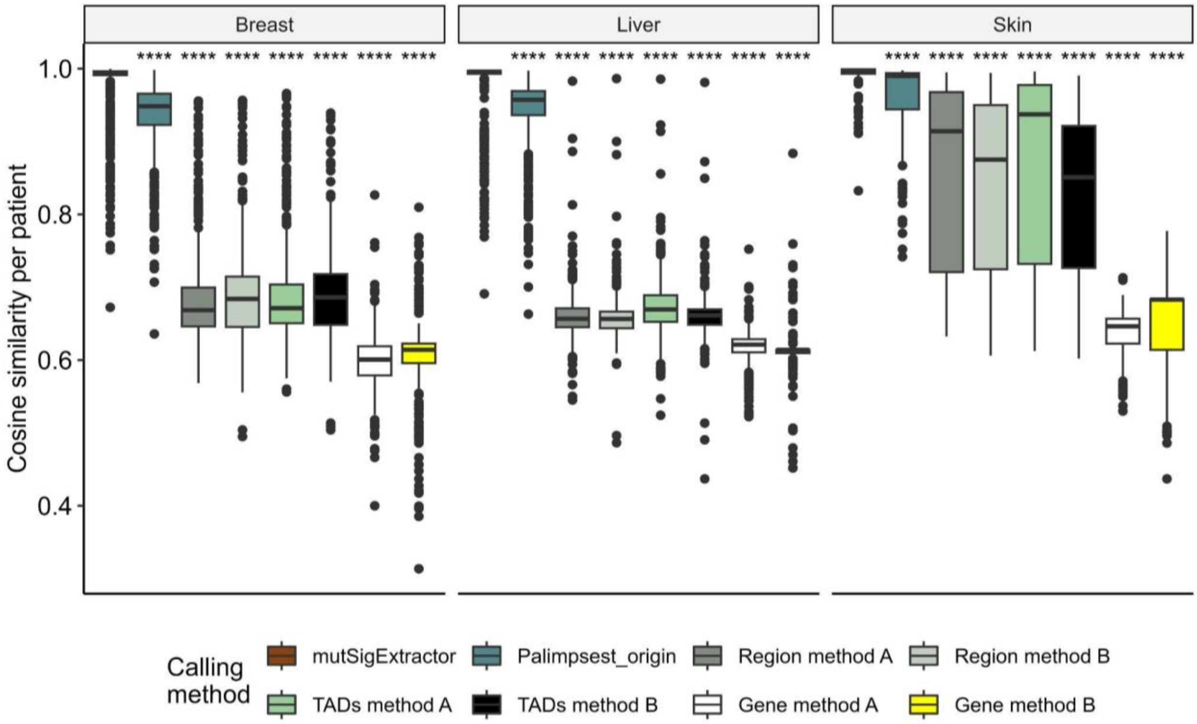


Figure 12. *Distribution of reconstruction error, cosine similarity, per each patient colored by each mutation signature calling method for breast, liver and skin cancer. Method A involves mutation calling by utilizing the mutational profile of a single patient, which was divided into specific genomic features. Method B involves calling the mutational signatures for each genomic feature individually. Patients' reconstruction error was calculated as the median of all errors per genomic feature. One-sided Wilcoxon test to cosine similarity of mutSigExtractor tools results as reference. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.1.1.3 Kataegis

I identified kataegis regions, which are characterized by clusters of SNVs where the intra-mutational distance is less than 6 bp, in both WGS and wWXS datasets. Skin melanoma exhibited the highest number of detected kataegis regions per patient in both WGS and WXS data (Figure 13A), followed by liver cancer in WGS data. The maximum number of detected kataegis regions in a single patient was 21,454 in skin melanoma WGS, 3,187 in liver cancer WGS, and 191 in breast cancer WGS. In the WXS datasets, the highest counts were 207 for skin melanoma, 133 for liver cancer, and 181 for breast cancer. The median number of detected kataegis regions per patient was consistently around 7-8 across all cancer types and sequencing technologies. Notably, kataegis regions with a higher number of mutations, identified as outliers in Figure 13A, were more prevalent in skin melanoma compared to liver and breast cancers. Overall, in WGS and WXS datasets, skin melanoma had a total of 226,774 and 311 kataegis regions respectively, liver cancer had 3,898 and 1,243, and breast cancer had 3,430 and 343. Kataegis regions were predominantly located on autosomal chromosomes (Figure 13B). The correlation between chromosome length and the number of kataegis regions was found to be significantly positive in liver and skin cancer WGS datasets, with Pearson's correlation coefficients of approximately 0.80 and p-values less than 0.05. In breast cancer WGS data, the correlation was also positive with a coefficient of approximately 0.42 and a p-value of 0.054. In contrast, the correlations in WXS data were lower, ranging from approximately 0.32 to 0.48, all with p-values less than 0.05. No kataegis regions were identified on the Y chromosome in the WGS and WXS datasets for breast cancer, as is expected, nor in the WXS dataset for skin melanoma, indicating a lack of such mutational clusters on this chromosome in these specific datasets.

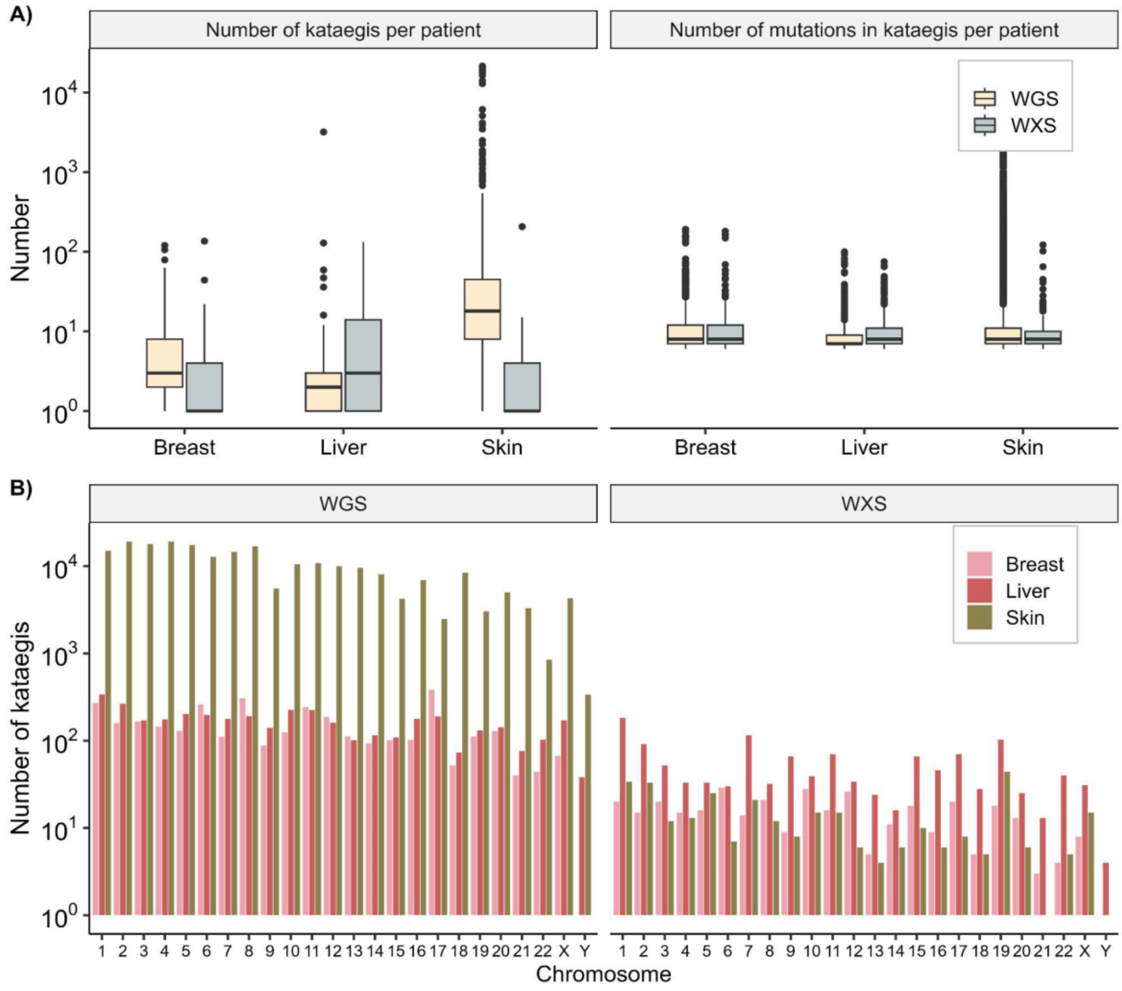


Figure 13. *A) Number of kataegis per patient (left panel) alongside the number of kataegis mutations per patient (right panel) per cancer type separated by sequencing technologies (WGS, WXS) B) Number of kataegis regions per chromosome per cancer type. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

Further characterization of the identified kataegis regions revealed that the majority of mutations within these regions were C>T transitions across all three cancer types. Notably, more than 75% of kataegis regions in liver and skin cancers exhibited C>T transitions, highlighting a prominent mutational pattern in these cancers (Figure 14A). In contrast, breast cancer showed a relatively higher proportion of C>G transitions compared to liver and skin cancers in both WGS and WXS data. When examining the mutational signatures assigned to mutations within kataegis regions (Figure 14B), I observed that in breast cancer, a significant proportion, approximately 70% of all mutations were attributed to mutational signatures SBS2, SBS13, and SBS40. In skin cancer, the kataegis regions were predominantly characterized by mutational signature SBS7a, accounting for 50-60% of the mutations in both WGS and WXS data.

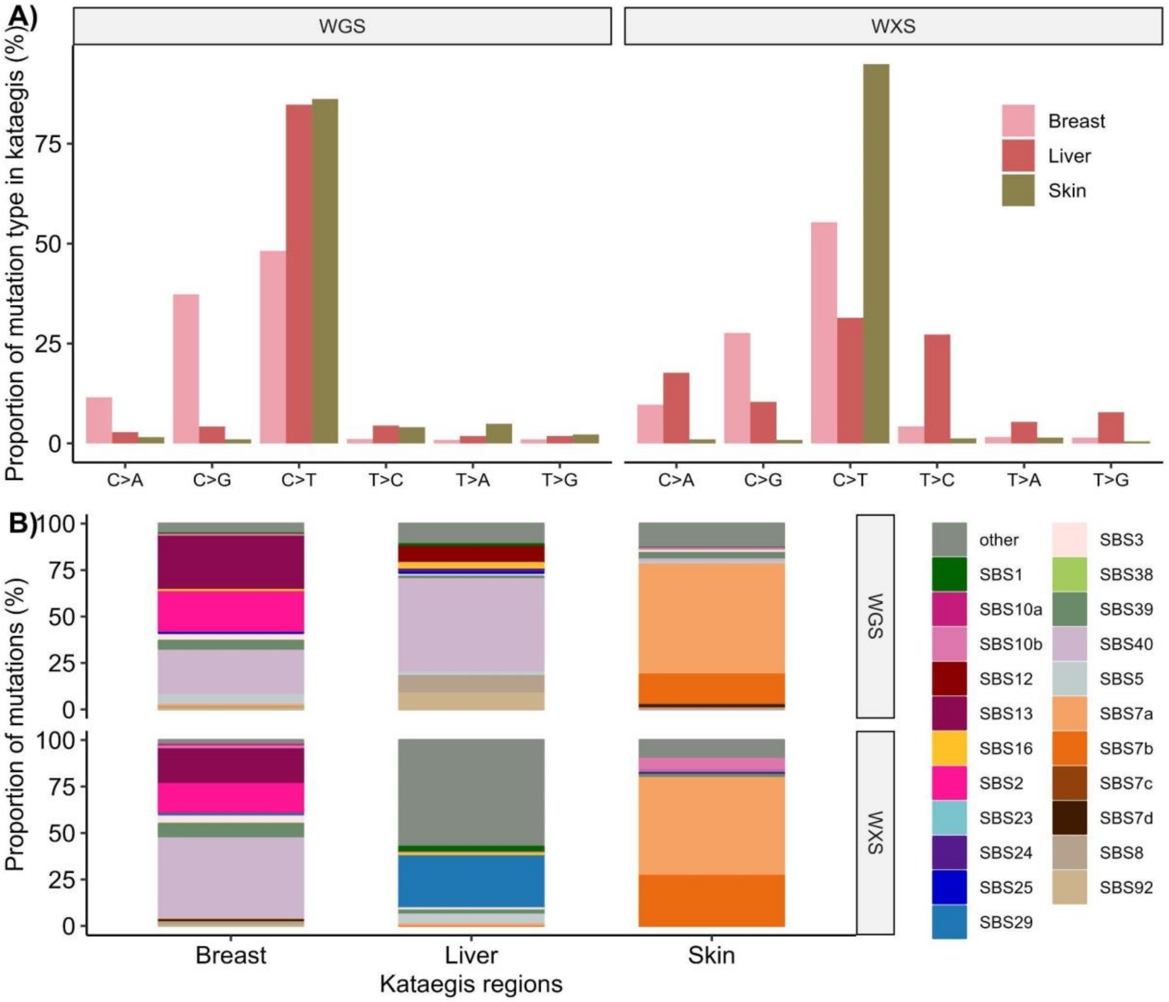


Figure 14. *A) Proportion of mutation types within kataegis regions across various cancer types (breast, liver, and skin cancer), differentiated by sequencing technologies (whole-genome sequencing, WGS, and whole-exome sequencing, WXS). B) Distribution of mutational signatures within kataegis regions, classified by their single-base substitution (SBS) origins as assigned by the Palimpsest tool, for each cancer type and sequencing technology used.*

In liver cancer, the kataegis regions exhibited a different profile depending on the sequencing technology. For WXS data, the most prevalent mutational signature was SBS29, comprising about 29% of the mutations, along with a diverse array of other signatures, each contributing no more than 4,000 mutations and collectively representing approximately 57% of the kataegis mutations. In contrast, the WGS data for liver cancer revealed a dominance of mutational signatures SBS40 (~50%), SBS12 (~10%), SBS92 (~10%), and SBS8 (~8%). Interestingly, the liver-specific signature SBS16 was also present in the WGS kataegis regions, although it constituted a relatively small proportion, approximately 2%.

4.1.2 Characterization of indels

I evaluated the different indel type counts in breast, liver, and skin cancer groups to characterize the mutational landscape of these cancer types. Table 8 provides a comprehensive overview of the total indels identified in each cancer category. Skin cancer exhibited the highest number of indels, with approximately 521,000 events, followed by liver cancer with around 477,000, and breast cancer with the lowest count at 398,000. In terms of WXS, liver cancer demonstrates the highest indels count among the three tumor types. Moreover, only about ~1% on average of whole-genome sequencing indels are accounted for by whole-exome sequencing indels when taking all of the cancer types into account. In both sequencing technologies of analyzed cancers the deletions made up more than ~55% of total indels except in skin melanoma WXS where only insertions were detected. The highest difference between identified indel types was between WGS and WXS of breast cancer where deletions in breast made up more than ~74% while in WXS are only around ~55% (Chi-square test, p-value=$3.8*10^{-154}$).

Table 8. *Total number of indels detected in whole-genome (WGS) and whole-exome (WXS) sequencing for breast, liver and skin cancer. Chi-square test with Benjamini-Hoechberg corrected p-values to test difference between WGC and WGS indel types*

Cancer type	**Sequencing technology**	**Total number of indels**	**Total number and proportion of insertions**	**Total number and proportion of deletions**	**Chi-square test**
Breast	WGS	396569	102206 (25.8%)	294363 (74.2%)	$3.8*10^{-154}$
	WXS	4004	1771 (44.2%)	2233 (55.8%)	
Liver	WGS	477251	144709 (30.3%)	332542 (69.7%)	$1.6*10^{-07}$
	WXS	8880	2922 (32.9%)	5958 (67.1%)	
Skin	WGS	521243	218177 (41.9%)	303066 (58.1%)	$1.1*10^{-153}$
	WXS	505	505 (100%)	0 (0%)	

Not all patients had detected indels across all cancer types. In breast cancer there were 9.3% (72/777, CI 0.07-0.12) WGS and 48.1% (568/1182, CI 0.45-0.51) WXS patients missing indels mutations. On the other hand, in skin melanoma there were 10.6% (34/320, CI 0.08-0.14) for WGS and 40.8% (190/466, CI 0.36-0.45) for WXS patients. Lastly, missing patients with indels in liver cancer were 4.7% (24/509, CI 0.03-0.07) for WGS and 21.2% (271/1281, CI 0.19-0.23) for WXS.

Out of the patients whose indels I had analyzed, I detected the highest number of indels per patient sequencing with WGS was in skin cancer with a mean and standard deviation of 1823 ± 2394, followed by liver cancer with 984 ± 6120 and breast cancer with 563 ± 6120. Similar to the analysis of SNVs, the TCGA cohorts (BRCA-US, LIHC-US, SKCM-US) have significantly fewer indels per patient compared to other cohorts especially in WGS data (Figure 15). Once again, I observed a similar trend as SNV where there was substantial variance between cohorts of the same cancer type within each cancer type and sequencing technology. Again, the smallest difference was detected in breast WXS cohorts (Kruskal-Wallis test, p=8.8e-18). There were a lot of patients with very small numbers of indels. Breast and skin WXS data had more than 94% (580/614, CI 0.92-0.96) and 98% (271/276, CI 0.96-0.99) patients with less than 5 indels. While in WGS data all the proportions were below ~8% for all cancer types. Pearson's correlations of WGS and WXS from the same patients was significantly positive in all cancer types, although not significant in skin melanoma; 0.850 (p-value=$5.4*10^{-21}$), 0.81 (p-value=$1.8*10^{-16}$) and 0.100 (p-value=0.66) for breast, liver, and skin cancer respectively. When examining the indel type, I identified that deletions were more prevalent than insertions in all cohorts (Supplementary Figure 3). Also, the TCGA WXS cohorts had only detected insertions.

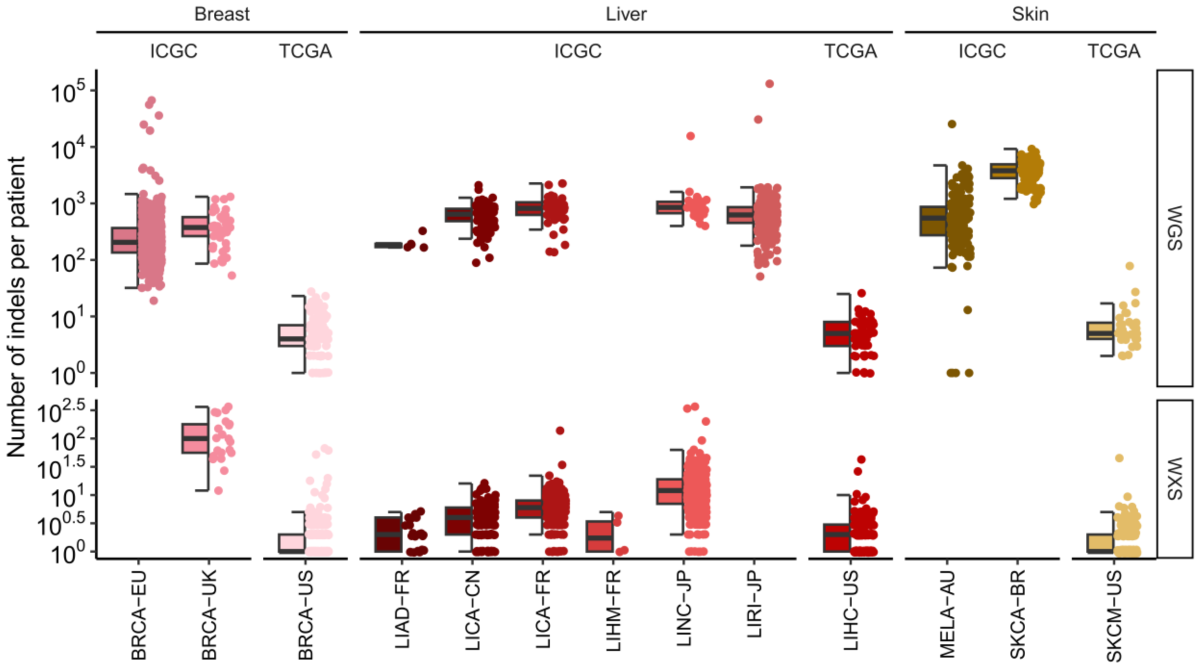


Figure 15. *Distribution of indels per patient across independent breast, liver and skin cancer cohorts obtained by whole-genome (WGS) or whole-exome sequencing (WXS). Plot shows log$_{10}$ transformed values of mutations. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

4.1.2.1 Indel mutational signatures

Indel mutational signatures were determined with the mutSigExtractor tool that had the best reconstruction error of cosine similarity 0.99. In total there were 17 analyzed indel mutational signatures from which ID2, ~47 and ~32%, and ID1, ~19 and ~20%, were the most abundant ones in WGS data from breast and liver cancer (Figure 16A). Skin melanoma WGS data had the most abundant ID8, ~30%, followed by ID4, ~16%, and ID13, ~10. WXS indel mutational landscapes differed significantly from the WGS ones (Chi-square test, p-value < 0.05). ID8, ID11, ID9 and ID5 contribute more to the overall indel landscape of breast and liver cancers. Skin melanoma WXS data was mostly, ~90%, made of ID1, ID11, ID16, ID17 and ID10. When examining the proportion of each indel signature per patient (Figure 16B), I determined that from WGS data breast cancer patients have a higher proportion of ID1, liver patients had ID3 and ID5, while skin patients had ID13 and ID8. From WXS data, the most noticeable difference is higher ID8 in liver cancer.

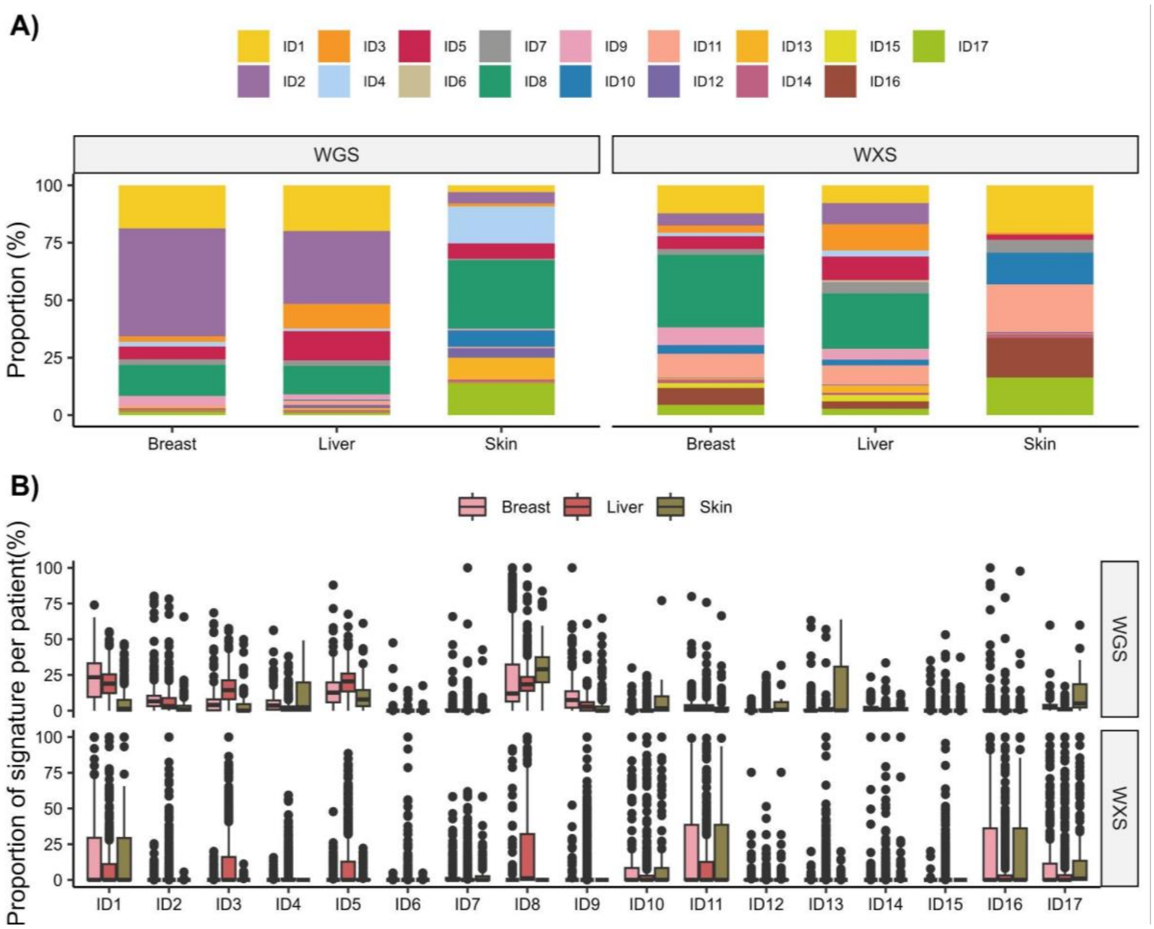


Figure 16. *A) Proportion of indel mutational signature determined with mutSigExtractor tool for breast, liver and skin cancer types separated by WGS and WXS data. B) Distribution of percentage of mutational signatures per individual patient. The plot shows log$_{10}$ transformed values of mutational signature counts. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

4.1.3 Characterization of structural variants (SV)

Structural variants are defined as larger DNA regions of length 1 kb or more involved in inversions and balanced translocations or genomic imbalances (insertions and deletions). In all cancer types inversion were in top 2 of most abundant types of SV in analyzed cancers regardless of their clustered profile. The highest number of structural variants was detected in breast, 98 853, followed by skin with 31 452 and liver with 27 022 (Supplementary Figure 4). Chi-square test showed that there is a significant difference in SV classes between cancer types (p-value < $2.2*10^{-308}$). When examining the SV profiles, I detected that clustered types of SV showed similar absolute counts of SV class profiles in breast, liver and skin cancer (Figure 17A). In all cancer types clustered inversions, 1 to 10 MB for breast and more than 10 MB for liver and skin, were most abundant out of all clustered SV classes. Out of three cancers, the liver had the least inversions and SV classes in general. On the other hand, the non-clustered SVs showed distinctly different SV mutational profiles. Skin melanoma had a higher number of deletions up to 1 kb in total, 2493, compared to other types of deletions. Other two cancer types were more abundant in deletions from 1 to 10 kb. Breast cancer had the highest number, 7550, of tandem-duplications 1-10 kb length. While in liver cancer inversions of 1-10 kb, 2455, were most abundant after translocations. After calculating the proportion of each SV class per patient (Figure 17B), I detected that breast and skin cancers had higher proportions of clustered SV classes than liver cancer. In non-clustered SV classes, breast cancer had the lowest proportions of deletions compared to other two cancer types. Other non-clustered SV classes per patient were more or less similar in all of the cancer types.

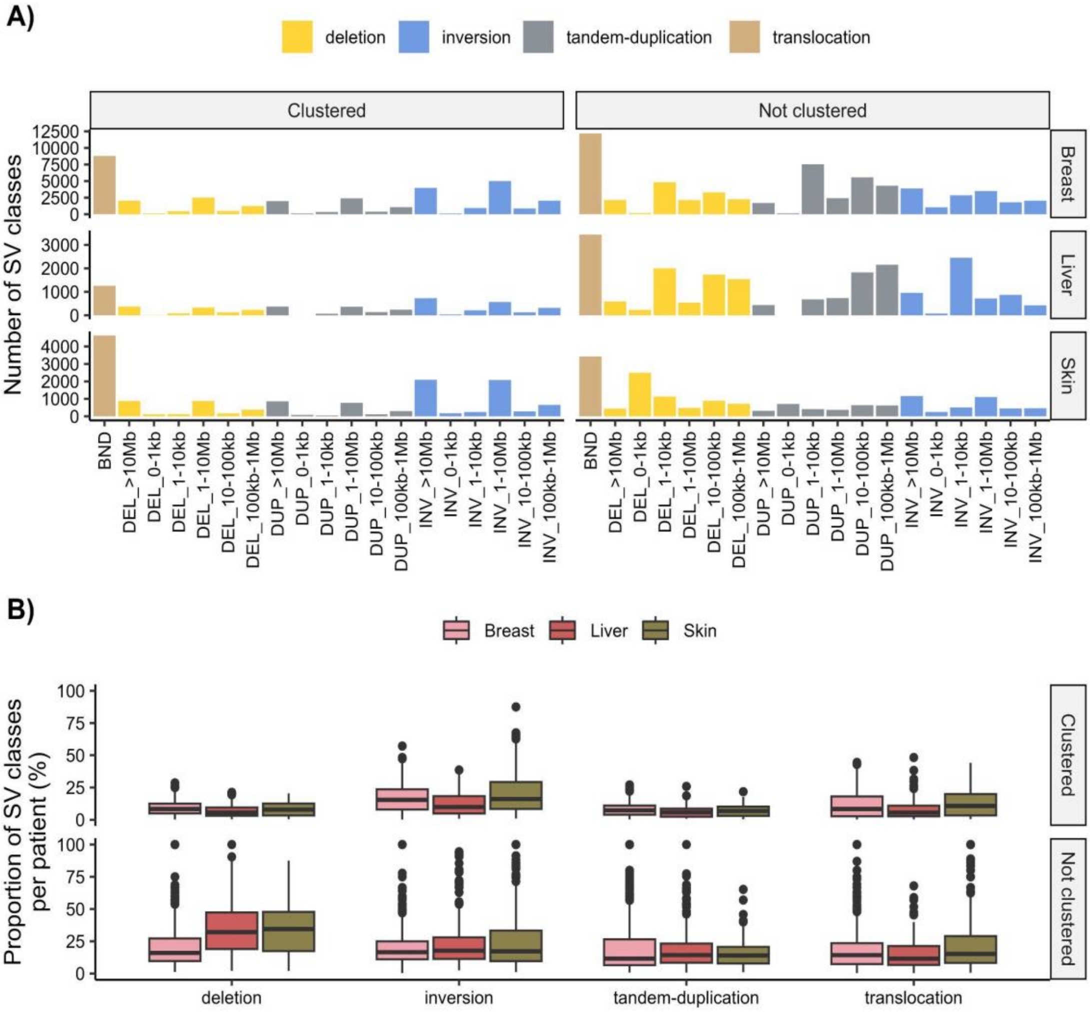


*__Figure 17.__ A) Number of annotated SV classes in breast, liver, and skin cancer depending on the length and clustered status. B) Proportion of annotated clustered and not clustered SV classes per patient in all three cancer types. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

4.1.3.1 SV mutational signatures

De novo SV mutation signature calling using deconvolution methods from the Palimpsest tool of generated SV mutational profiles resulted in 7 breast, 6 liver and 6 skin *de novo* SV mutational signatures (Figure 18). Most *de novo* SV signatures specific for each cancer were characterized by a higher proportion of non-clustered SV classes. Deletions and tandem-duplications were especially more abundant in non-clustered SV classes of SV signatures like breast SV1, SV2 and SV6; liver's SV2, SV5 and SV6 and skin's SV1 and SV2. Non-clustered inversions were more abundant in breast SV4, liver SV1 and skin SV6 and in a smaller percentage in skin's SV5. On the other hand, clustered inversion had higher percentages in the breast SV2, liver SV3 and skin SV3 and SV6.

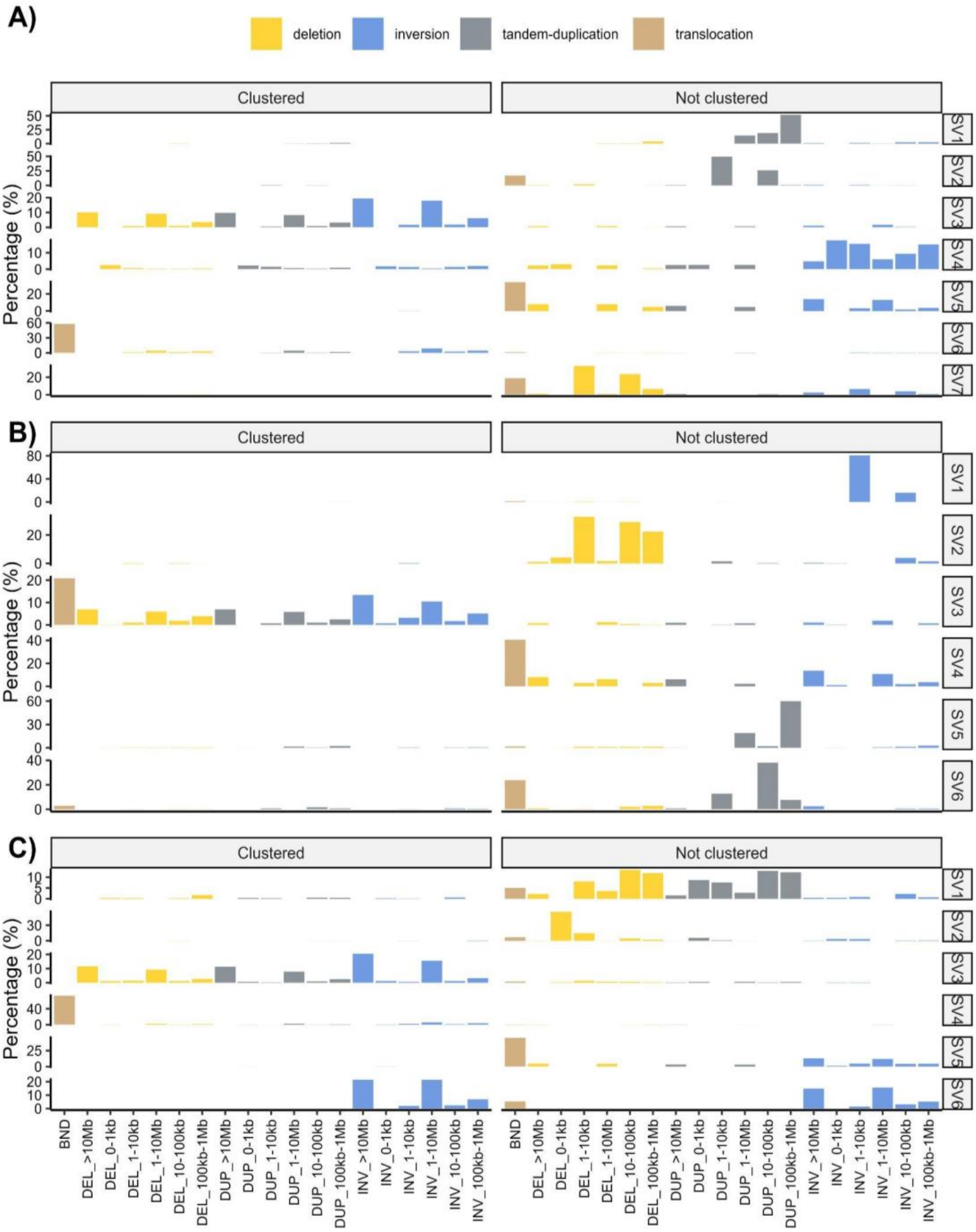


Figure 18. *De novo structural variant (SV) mutational signatures profile showing the percentage of clustered and non-clustered SV types (deletion, inversion, tandem-duplications and translocations) for A) breast, B) liver and C) skin melanoma cancers called using Palimpsest tool.*

To characterize the *de novo* SV signatures I calculated the cosine similarity of their generated profiles with each other and with annotated COSMIC SV signatures. Signature pairs of cancer-specific *de novo* SV signatures with the cosine similarity above 0.9 were breast SV6 and skin SV4, ~0.99; breast SV3 and skin SV3, ~ 0.98; breast SV5 and skin SV5, ~0.97 (Figure 19A). Next pairwise comparison SV signatures with cosine similarity above 0.8 were breast SV7 and liver SV2, ~0.84; followed by skin SV5 and liver SV4, ~0.84. Comparison with annotated SV signatures from COSMIC showed that extracted *de novo* SV signatures had lower similarities to annotated COSMIC ones (Figure 19B). The highest cosine similarity was obtained for liver *de novo* SV5 and COSMIC SV1 of ~0.77 and liver *de novo* SV6 and COSMIC SV3 of 0.63. Other signatures with cosine similarity over 0.5 were breast SV1 and COSMIC SV1, ~0.58; breast SV2 and COSMIC SV7, ~0.54; and breast SV4 and SV10, ~0.54. In skin melanoma, the highest cosine similarity, though still below 0.5, was observed between the skin SV3 and the COSMIC SV4, as well as between the skin SV2 and COSMIC SV6.

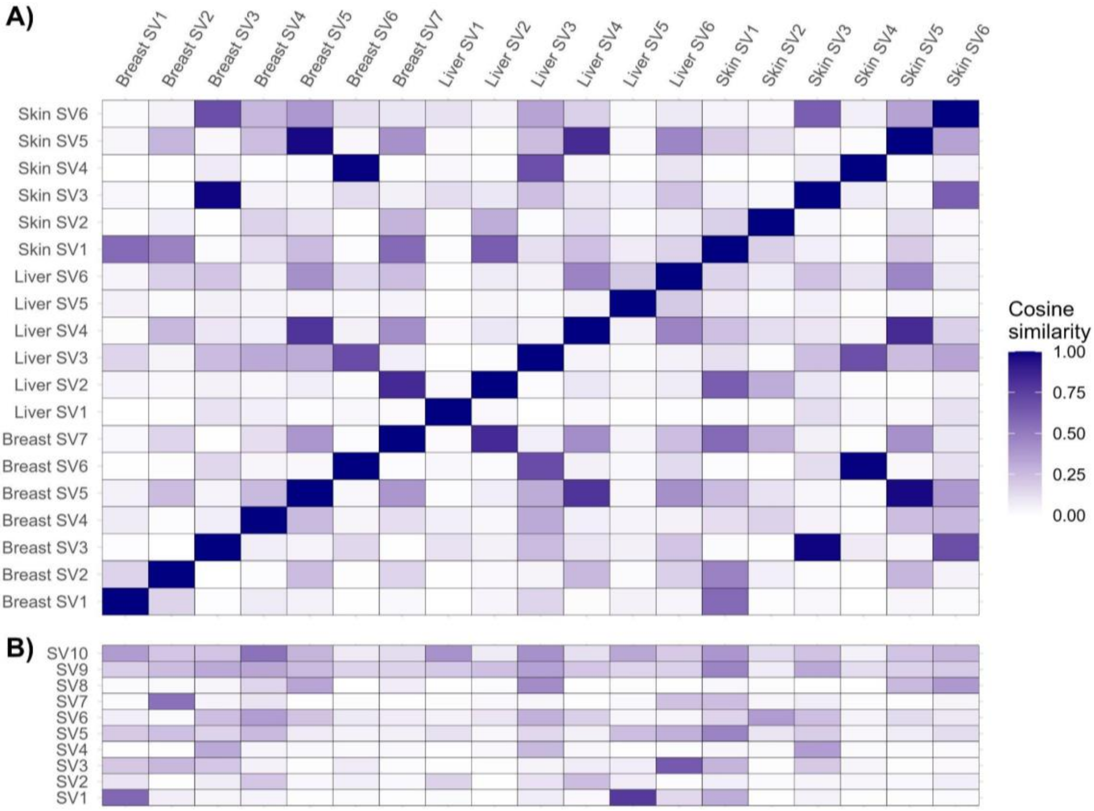


Figure 19. *A) Cosine similarity between de novo SV signatures in breast, liver and skin cancer B) Cosine similarity between de novo SV signatures and annotated SV signatures from COSMIC database determined and analyzed in Everall et al. 2023*

There was a significant difference in the proportions of *de novo* SV mutational signatures per patient in each cancer type (Figure 20A, Kruskal-Wallis, p-value <0.05). Breast and liver *de novo* SV2, SV3 and SV4 signatures were higher than the others. In skin melanoma *de novo* SV1, SV5 and SV6 dominated the landscape when looking at the proportions of SV signatures per patient. In the breast cancer *de novo* SV mutational signature landscape, SV5, SV3 and SV2 dominated the landscape (Figure 20B). As for liver cancer, *de novo* SV2, SV3 and SV4 were the most abundant out of all signatures. Lastly, in skin melanoma *de novo* SV5, SV3 and SV4 dominated the skin *de novo* SV mutational landscape. The difference in *de novo* SV mutational landscape was significant (Chi-square test, p-value < 0.05).

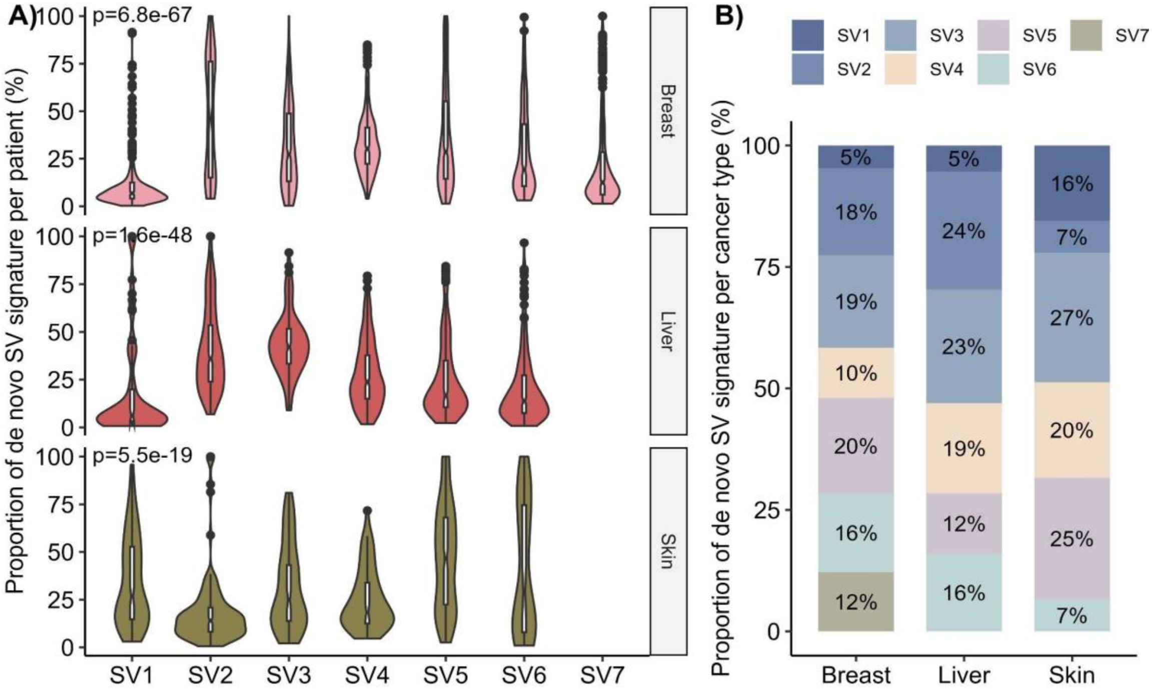


Figure 20. *A) Proportion of cancer-specific de novo SV mutational signature per individual patient calculated by keeping patients that have more mutations than the median in each cancer type (Kruskal-Wallis test). B) Proportion of de novo SV signatures per cancer type. Chi-square test (p<0.05). Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

4.1.3.2 SV-hotspots

SV-HotSpot tool was applied to each SV of a specific mutational signature type separately for different cancers, resulting in 2211 SV hotspots for breast cancer, 285 for liver cancer, and 504 for skin cancer. Different cancer types showed a significantly different enrichment of SV-hotspots signatures (Figure 21A; Chi-square test, p-value 4.1e-292), with liver cancer SV-hotspots resulting only from liver's *de novo* SV3 signature, most breast cancer SV-hotspots originating from breast's *de novo* SV5 and SV3 signatures, and skin melanoma SV-hotspots predominantly caused by skin melanoma *de novo* SV3 signature. Most breast and liver SV-hotspots were detected on chromosomes 8 and 1 (Figure 21BC), while chromosome 11 had the most skin melanoma SV-hotspots (Figure 21D). A significantly higher percentage of patients per each SV-hotspot were detected on chromosomes 17 and 4 in breast, chromosomes 5 and 11 in skin, and chromosomes 11 and 8 in liver cancer.

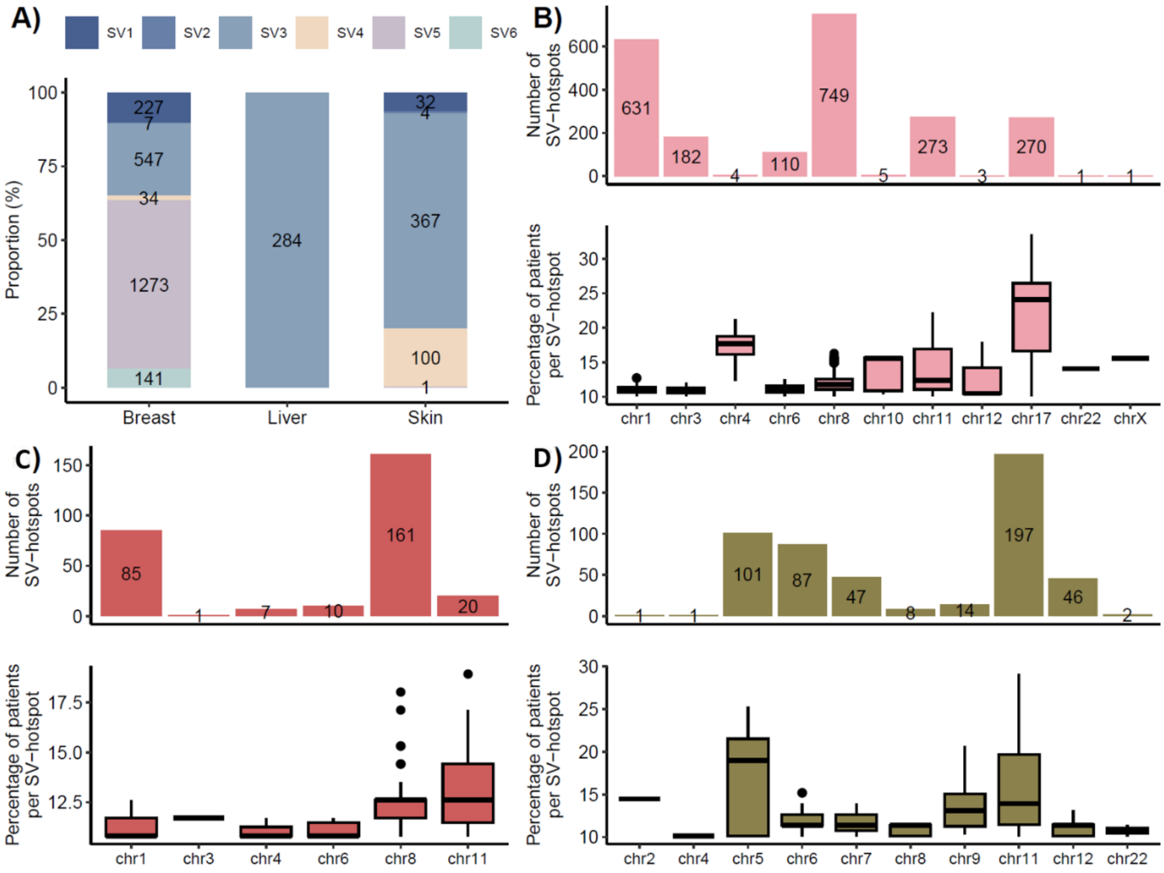


Figure 21. *A) Proportion of cancer-specific de novo SV mutational signatures determined SV-hotspots with SV-HotSpot tool in breast, liver and skin cancer. Number of detected SV hotspots per chromosome in B) breast, C) liver and D) skin cancer in the upper panels. Lower panels display percentages of patients per each SV-hotspot. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

Each chromosome had a different composition of *de novo* SV signature hotspots (Figure 22). In liver cancer as I detected only SV3 hotspots, it was expected that all of the SV-hotspots were of that type. Other two cancer types had a statistically significant SV-hotspots composition on chromosomes (Chi-square, p-value < 0.05). Majority of the more abundant SV-hotspots in breast cancer were enriched with *de novo* SV5 hotspots; chromosomes 1 with 79%, 3 with ~96%, 6 with ~99% and 8 with ~54%. Breast chromosome 8 had the most diverse SV-hotspot landscape out of the mentioned ones with ~30% SV3, 12% SV1, 5% SV6 and less than 1% of SV2 hotspots. Furthermore, breast chromosome 11 had a similar proportion of *de novo* SV3, SV5 and SV6 hotspots, around ~30%, while chromosome 17 had ~83% SV3 hotspots. Skins most abundant SV-hotspots chromosomes were enriched with skins *de novo* SV3 signature; chromosome 5 with ~97%, 6 and 7 with 100%. Meanwhile, chromosome 11 had ~44% of SV4 and SV6 hotspots. Lastly, chromosome 12 had ~78% SV3 and ~22% SV4 hotspots.

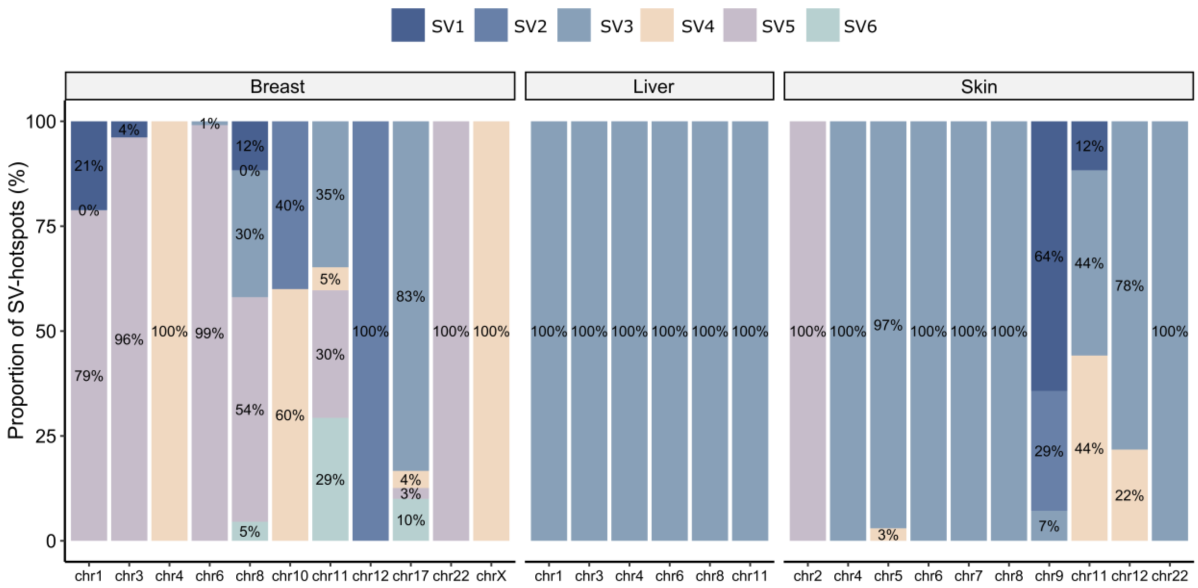


Figure 22. *Proportion of cancer-specific de novo SV signature SV-hotspots per chromosome for breast, liver and skin cancer.*

4.2 Identification of tissue-specific expressed genes in normal tissues

Through this chapter, I have evaluated different tissue-specific metrics using RNA-expression data from publicly available sources to obtain a comprehensive list of tissue-specific genes for various breast, liver and skin normal tissues. The list was later used to determine the ability of tissue-specific genes to predict the cell-of-origin in the developed models using mutation count and histone marks.

4.2.1 Evaluation metrics for tissue specificity and filtering parameters

In total, I used 9 datasets (8 GTEx datasets with different tissue counts and normalizations and one publicly available dataset), to determine the best tissue-specificity metric. Out of all calculated tissue-specific metrics, the Tau index stood out as the one with the less skewed distribution in the Fagerberg dataset, while more skewed towards tissue-specific genes with Tau closer to 1 in GTEx datasets (Figure 23). Other tissue-specific metrics, Zscore, Gini index and Tsi, metrics also showed similar distributions but Tau index was the one with the most tissue-specific genes since the cut-off is usually above 0.8.

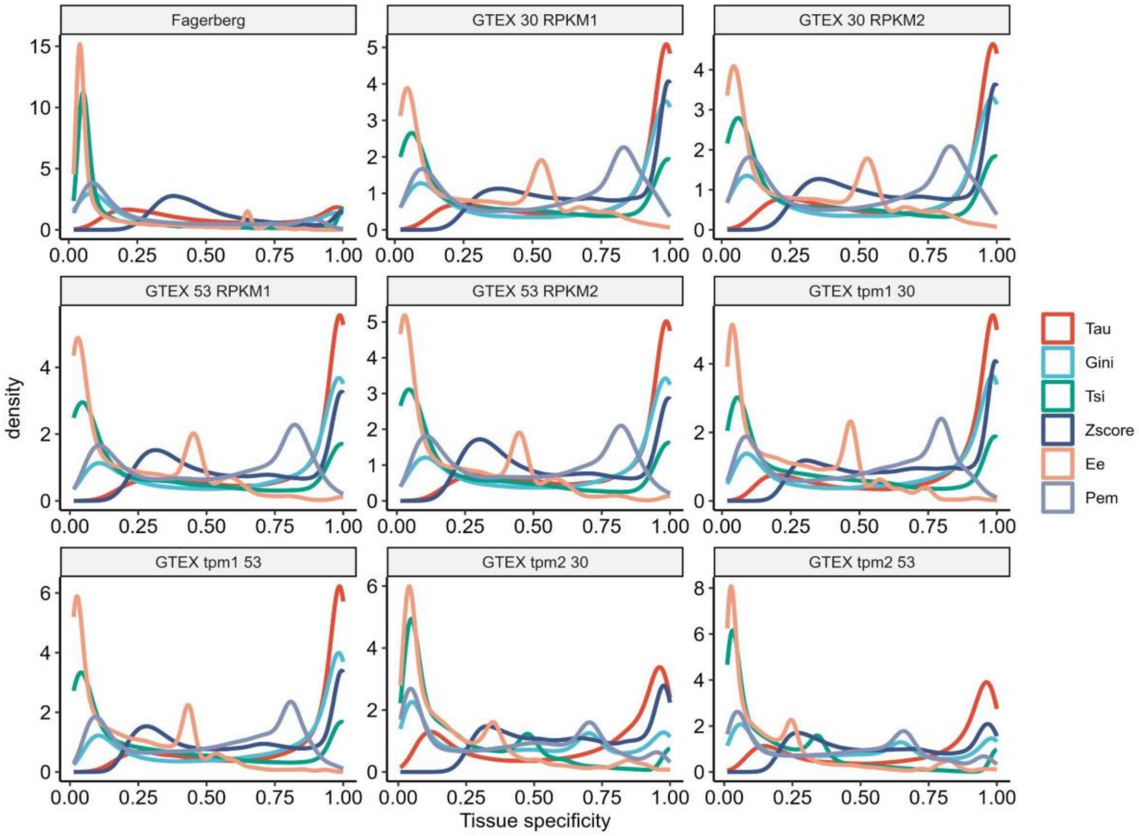


Figure 23. *Distribution of tissue-specificity parameters (Tau, Gini, Tsi, Zscore, Ee, Pem) with data for human RNA-seq of 9 different datasets shown as a density function.*

To further evaluate the specificity and performance of tissue-specific metrics, I looked at sets of well-defined tissue-specific genes (xenobiotic metabolism, melanin production, spermatogenesis) and other broadly expressed genes (membrane organization, RNA splicing, protein folding) (Figure 24). Only the Tau index was able to correctly capture distribution shifted to 1 of known expected tissue-specific genes and it was constant across all analyzed datasets (Supplementary Figure 5). After Tau index, Gini index and Zscore were the second-best tissue-specific metrics.

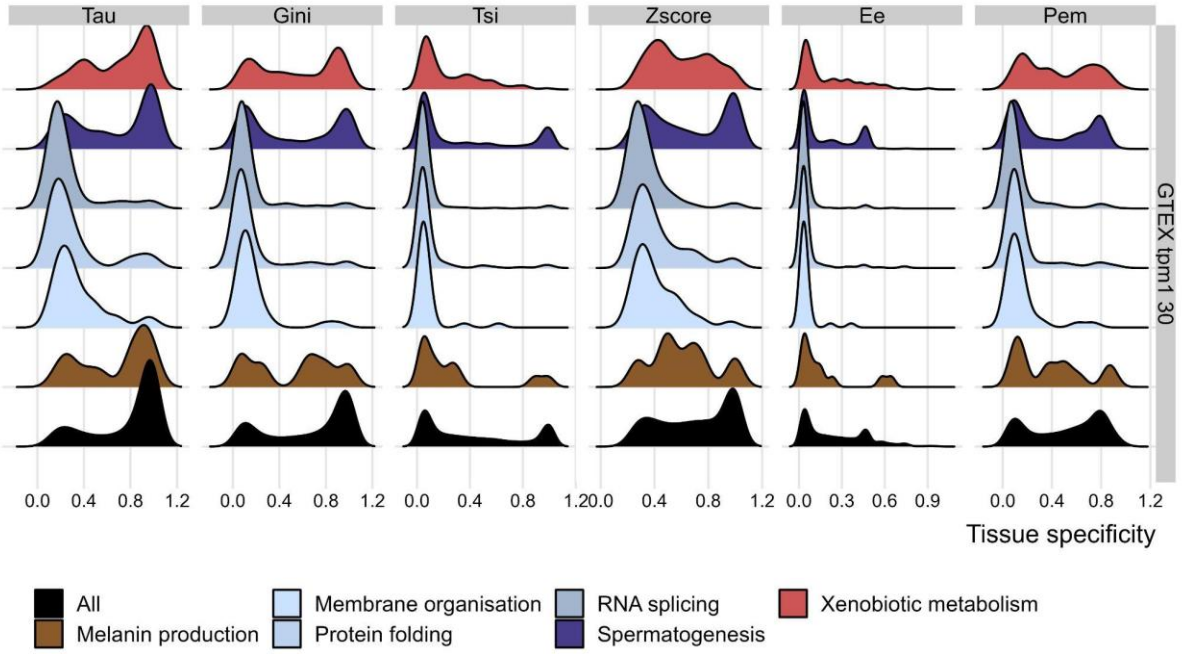


Figure 24. *Tissue-specific metrics on GTEX TPM1 dataset with 30 tissue types. Tissue-specificity parameters of subsets of genes which are expected to be tissue-specific (xenobiotic metabolism, spermatogenesis and melanin production) or broadly expressed (membrane organization lines, protein folding and RNA splicing), based on associated GO terms. The black distribution represents the distribution for all genes, including those not associated with any of these GO terms.*

As the Tau index clearly showed the best separation of expected tissue-specific and broadly expressed genes, I calculated extended Tau to assign each tissue-specific gene to a certain tissue. When defining specific genes with a Tau index larger than 0.8, the highest number of tissue-specific genes was detected in GTEx data normalized with TPM above 31700, while the Fageberg dataset had the lowest number of specific genes of 5515 (Figure 25A). In general, normalization with TPM resulted in a larger number of tissue-specific genes compared to RPKM normalization. When I calculated the extended Tau with whom I assigned tissue-specific genes with Tau > 0.8 to multiple tissues, I detected many genes, ~13000, assigned to the testis organ, followed by around 4000 genes to brain and nerve tissues using GTEX data of 30 tissues normalized by TPM (Figure 25B). The heart and pancreas were tissues with the least assigned tissue-specific genes which was mainly consistent across all of the analyzed RNA-seq datasets (Supplementary Figure 6).

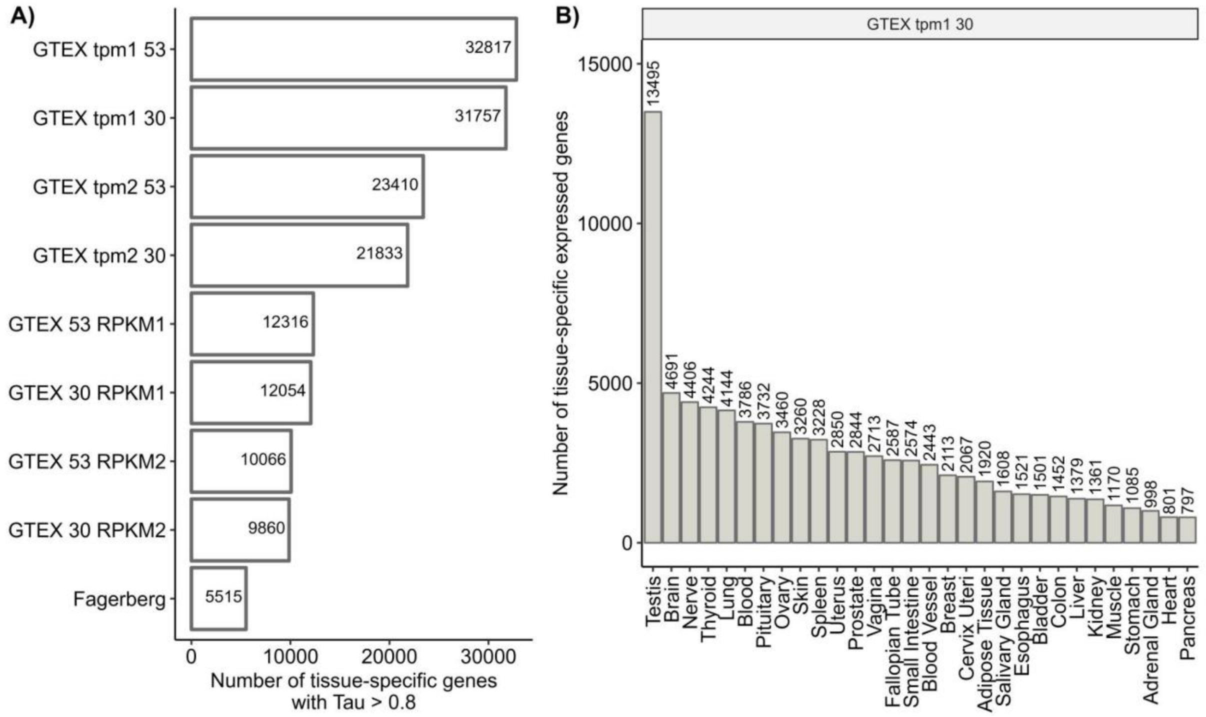


Figure 25. *A) Number of tissue-specific genes with Tau index larger than 2 in 9 different datasets RNA-seq datasets. B) Number of tissue-specific genes for each tissue of GTEX tpm1 of 30 tissues based on the extended Tau index.*

4.2.2 Breast, liver and skin tissue-specific genes

Because I used multiple datasets that I analyzed separately, I needed to determine whether to include all the assigned tissue-specific genes for a particular tissue within each dataset as that tissue. To validate the merging of all tissue-specific genes into one specific tissue, I assessed two genes to see if important genes were captured in all datasets. I found that the well-annotated liver-specific gene alpha-fetoprotein (*AFP*) was consistently identified as liver-specific across all utilized datasets (Figure 26A). Notably, the Fageberg dataset also classified *AFP* as kidney-specific, based on the extended Tau metric. Another tissue-specific gene, D-amino acid oxidase (*DAO*), traditionally recognized as kidney-specific, was detected not only in the kidney but also in the brain and liver when analyzing GTEx data normalized with TPM. Interestingly, the GTEx RPKM dataset did not include the *DAO* gene, and neither the Fageberg dataset nor GTEx TPM data with 30 tissues identified *DAO* as brain-specific. Given these multiple lines of evidence suggesting that certain genes may exhibit tissue-specificity beyond their traditionally recognized tissues, I decided to include all genes identified as tissue-specific in at least one dataset. This comprehensive approach ensures that potentially significant tissue-specific gene associations are not overlooked. Since there are multiple subtypes of skin and breast tissue available in different datasets, I attributed the tissue-specific genes for each of them (Figure 26B). The highest number of 5917 tissue-specific genes was assigned to the skin. Separated skin tissues that were exposed to the sun and not in GTEx datasets had lower numbers of genes, around ~2600, and ~1200. Both breast and liver tissue-specific genes resulted in a high number of genes around ~3500. Since one gene can overlap in many tissues I analyzed the overlapping genes in defined tissue-specific sets (Figure 26C). As expected, highest overlaps were detected within each major tissue/organ group, where breast and breast mammary tissue annotations shared the highest number of 960 genes. Followed by different combinations of shared genes between defined skin tissues. A total of 436 genes were shared among breast, liver and skin normal tissues. These results imply that the tissue-specific genes were indeed correctly annotated within each tissue and sub-tissue types.

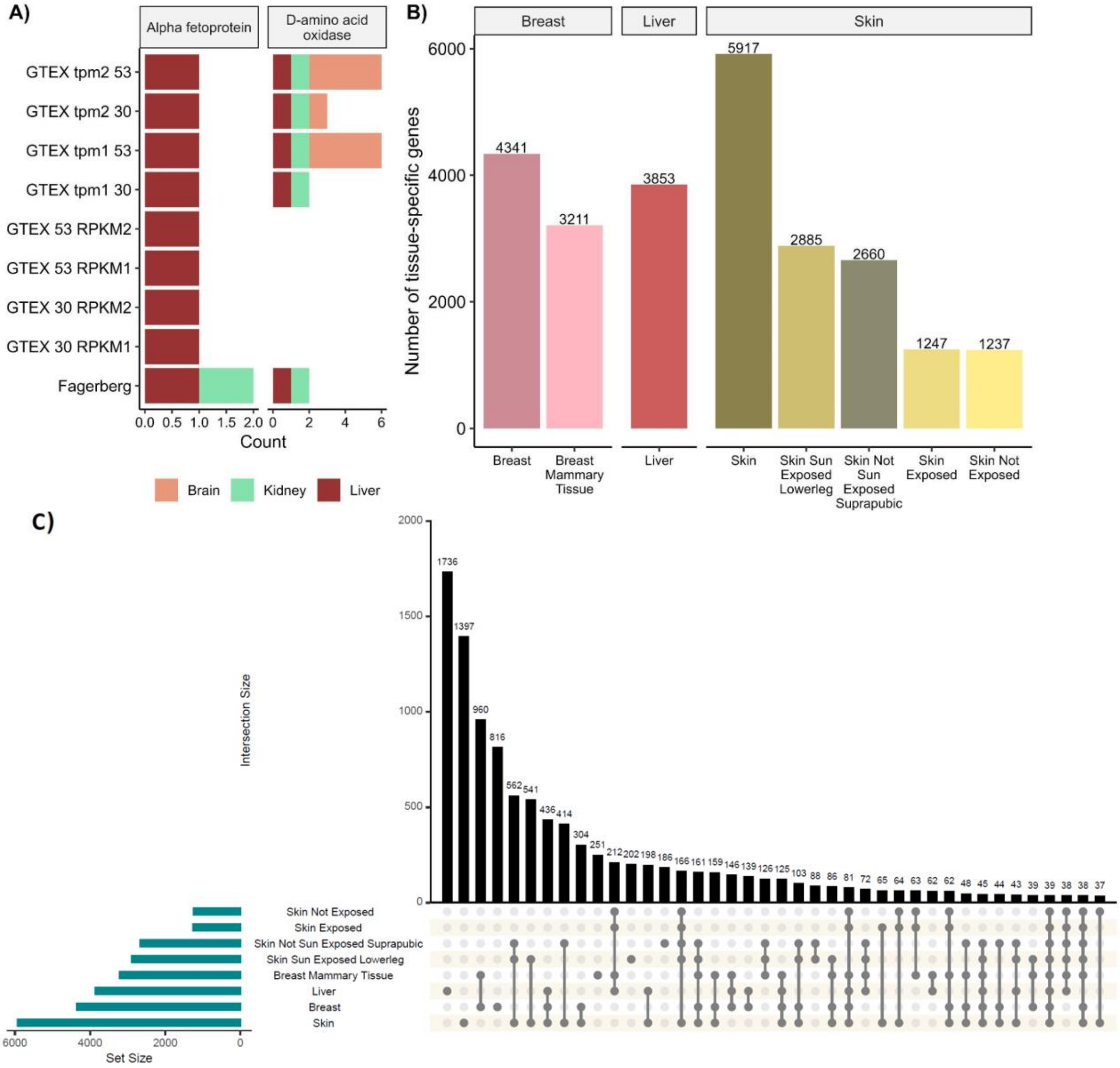


Figure 26. *A) Liver-specific alpha-fetoprotein (AFP) and kidney-specific D-amino acid oxidase (DAO) detected across 9 datasets in different tissues B) Total number of tissue-specific genes for breast, liver and skin tissues determined by taking into account all tissue-specific genes based on extended Tau index across all datasets C) UpsetR plot showing the overlapping tissue-specific genes.*

4.3 Machine learning models for predicting cell-of-origin

Through this chapter, I had developed a preliminary cell-of-origin prediction model using multiple linear regression of all mutations and histone modifications of normal tissues summarized and normalized to different genomic features; 1 Mb genomic regions, topologically-associated domains and genes. Erroneous prediction rates, defined by standard residuals of each feature, were assessed and associated with various genomic features such as mutational signatures, SV-hotspots, kataegis, tissue specific super-enhancers and others, to test whether these features influence the prediction accuracy of the models. To better understand the COO prediction results, I performed the correlation analysis of a specific mutational profile of a feature within each cancer type to the epigenome of a normal potential cell-of-origin of that cancer. Based on the obtained findings, I developed random forest and gradient extreme boosting models to potentially increase the prediction accuracy by employing a more complex machine learning model.

4.3.1 1 Mb genomic region cell-of-origin (COO) predictive models

I developed a preliminary model using multiple linear regression of aggregated tumor SBS mutations and epigenomes of normal tissues to correctly identify the COO for both WGS and WXS datasets using 2128 1 Mb genomic regions (predicted COO corresponds to the model with the highest prediction accuracy). As seen in Figure 27A, the best models with the highest variance explained (accuracy) from WGS datasets of different cancer types match the correct COO of that cancer type. Out of all three WGS cancer datasets, breast cancer had the lowest accuracy, 50.51% for breast luminal cells mature, followed by 81.24% for melanocytes in skin melanoma and the highest obtained value for liver cancer, 85.25% for liver tissue. All top WGS models showed significantly higher explained variance compared to its second-best non-cell-of-origin model (one-sided Wilcoxon test, p-values 0.001, 8.115×10^{-6}, 8.115×10^{-6} for breast, liver and skin cancer respectively). On the other hand, none of the top WXS cancer dataset models have been able to correctly identify the COO of corresponding cancer type. For both skin melanoma and liver cancer, the best COO model was the breast basal, 51.50%, and 56.47% respectively, while for breast cancer the best models were brain tissues, 44.53%. The best WXS cancer dataset models showed significantly lower accuracy than best correct WGS models of corresponding cancer (one-sided Wilcoxon test, p-values 0.0144, 8.115×10^{-6}, 8.115×10^{-6} for breast, liver and skin cancer respectively). When examining the model accuracies on individual patients from WGS and WXS datasets (Figure 27B), I detected that less than ~4% of patients had correctly identified COO from WXS dataset. In the WGS cancer dataset models, breast cancer had the lowest percentage of correctly identified patients (~56%), followed by melanoma (~79%) and liver (~86%). The WGS data COO models results were reproducible across aggregated profiles based on a specific cohort (Supplementary Figure 7A), except for TCGA cohorts (BRCA-US, LIHC-US and SKCM-US) where correct COO was not determined as the best model. While only for breast WXS data independent cohorts was the COO correctly identified. As seen in Supplementary Figure 7B majority of wrongly identified COO of individual patients originated from the TCGA cohorts. Out of all cancer types, melanoma had the highest mutational count per 1 Mb regions, followed by liver cancer (Figure 27C).

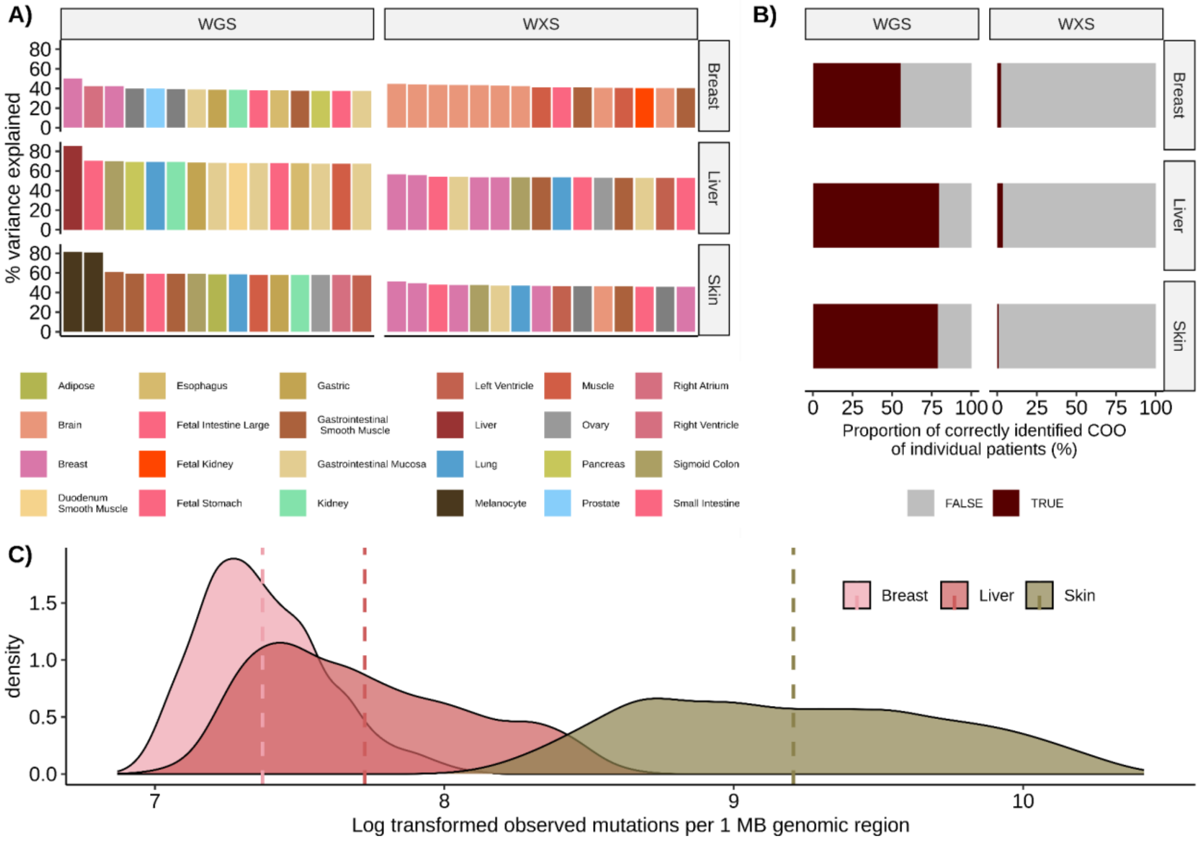


Figure 27. *A) Multiple linear regression models for the prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WGS and WXS were trained on an extended set of 101 tissue sets but showing only the top 15 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation. B) Proportion of correctly and incorrectly identified COO for individual patients based on results of corresponding COO models in A). C) Distribution of aggregated observed mutations per 1 Mb region in each cancer type displayed as density plot alongside vertical lines representing the medians*

Moreover, taking only indel mutations instead of SNVs did not result in correct COO identification in WGS data except for skin melanoma WGS data (Supplementary Figure 8). However, the explained variance of the indel COO model was significantly lower with only ~30% of explained variance than the SNV COO model (Wilcoxon test, p-value = 0.001). There was a significant drop in correct COO identifications of individual patients where the highest percentage of correctly identified COO was only ~21% (61/286, CI 0.17-0.26) for skin melanoma with indel mutations.

WXS datasets for liver and skin melanoma, unlike WGS datasets, contain regions with zero mutations per 1 Mb genomic regions, 50 and 239 respectively. Regions with 0 mutations had bigger error prediction rates in expected correct COO model for certain cancer type, median and standard deviation of 4.90±0.27 for liver and 2.23±0.46 for skin melanoma, compared to other regions, median and standard deviation of 0.36±0.44 for liver and 0.39+-0.45 for skin melanoma. These erroneous regions with no mutations form distinct clusters when examining the correlation of standardized residuals of WGS and WXS 1 Mb regions results (Figure 28). In both best and correct COO, breast and skin cancer have significant positive correlation, while liver cancer has significant slight negative correlation. These results imply the more or less the same direction of erroneous prediction rates of regions regardless with both WGS and WXS data from all cancers.

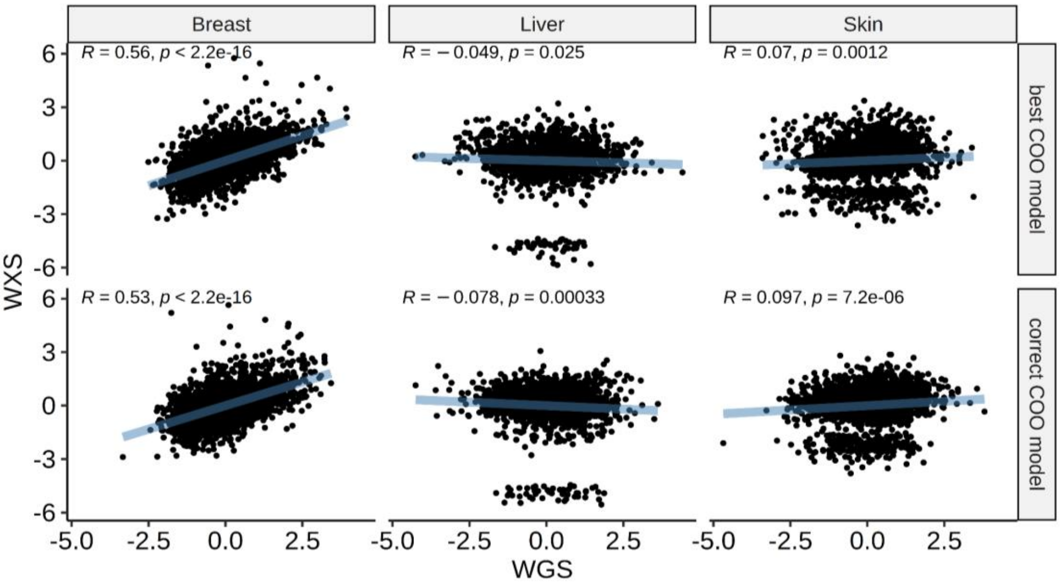


***Figure 28**. A) Pearsons's correlation of standardized residuals between WXS and WGS from the best obtained and correct COO models*

For downstream analysis, I focused the analysis on only WGS cancer dataset models due to better overall performance of COO models. More erroneous regions or outliers, which represent genomic regions where the relationship of histone modifications and mutations differs from the average, were defined as over-predicted regions with standard residuals lower than -2 and under-predicted regions with standard residuals higher than 2. Outliers were most abundant in breast cancer with a total of 111 outliers, 15 over-predicted and 96 under-predicted. Skin melanoma had 88 outliers, 53 over- and 35 under-predicted, while liver cancer had 82 outliers, 37 over- and 45 under-predicted (Figure 29A). Breast under-predicted outliers showed enriched origin from chromosome 8, 17 and 1; while other outliers of models trained in other cancer types did not show enrichment in any of the chromosomes (Figure 29B). I found that 1867 out of 2128 regions (88%, CI 0.86-0.89) were not outliers in all three cancer types (Figure 29C). Very few regions were annotated as the same outlier type by models trained in different cancer types, suggesting a distinct trait of each outlier in its corresponding cancer type. The highest overlap was 12 under-predicted outliers found both in liver and breast cancer.

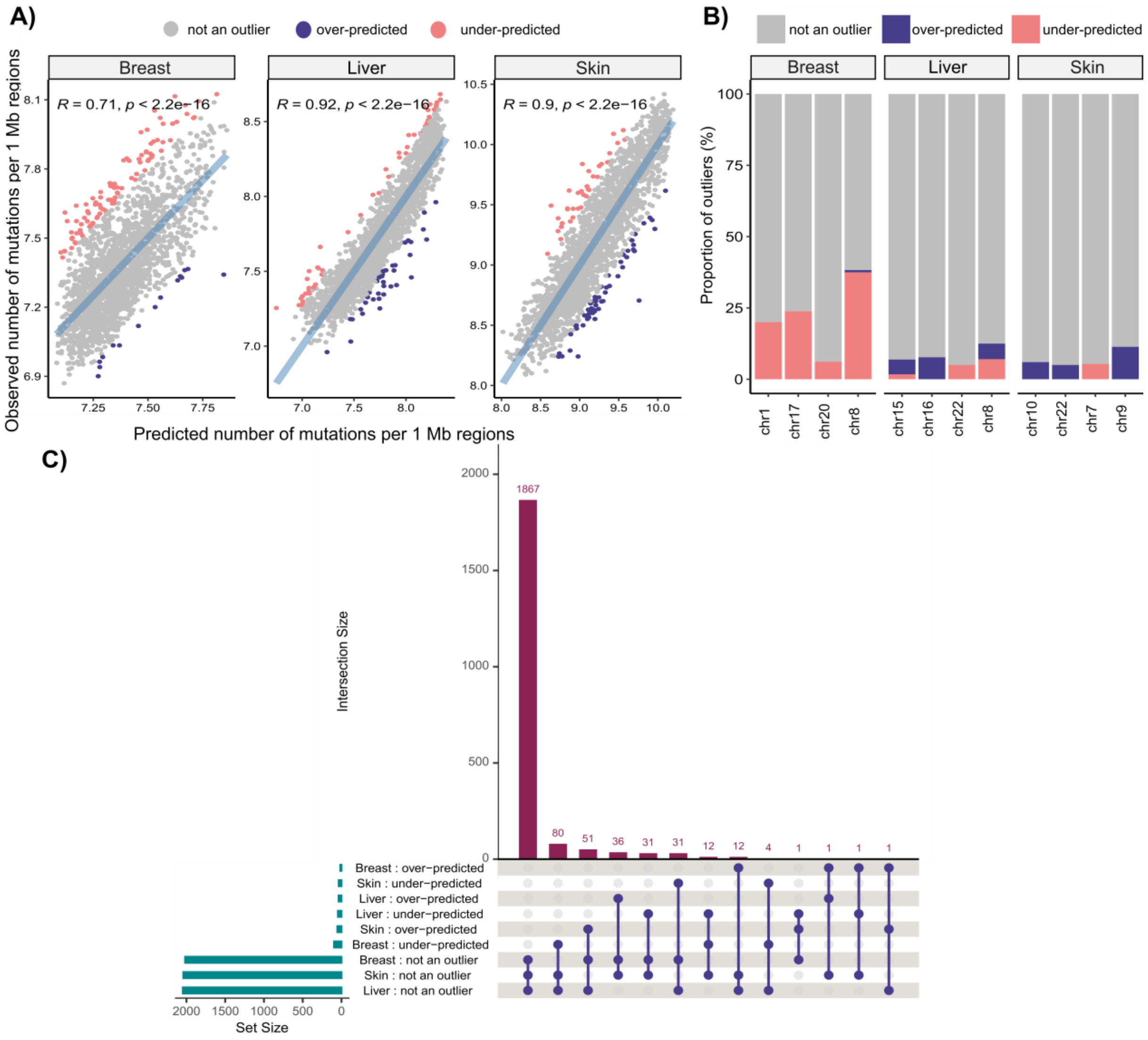


Figure 29. *A) Predicted vs observed mutations of aggregated tumor profiles in 1 Mb genomic regions of breast, liver and skin cancer from the best multiple linear regression models of WGS data. Colors denote different outlier classes B) Proportion of annotated outliers based on their location in the genome (if more than 15% are outliers) in each cancer C) UpsetR showing the overlaps of annotated regions as over-, under-outliers or not an outlier between different cancer types.*

4.3.1.1. Analysis of erroneous 1 Mb genomic regions

Next, I analyzed the gene composition and regulatory features in 1MB genomic regions based on their erroneous prediction rate, to determine whether they influence the error rates. First, I looked at the enrichment of genes associated with development and progression of cancer from Cancer Gene Consensus (CGC), tumor immune microenvironment (TIME) and enrichment of tissue specific super-enhancers (SE) from both normal and tumorigenic cell-lines or tissues from the SEdb2.0 database. As shown in Figure 30A, in breast and liver cancer the CGC were more enriched in under-predicted outlier regions, while they were more present in over-predicted regions of skin melanoma. The similar observations were also observed with TIME driver genes (Figure 30B). However only TIME drivers for breast cancer showed significantly different enrichment between all outliers (Fisher's test, p=0.0003). Tissue-specific super-enhancers were generally more enriched in under-predicted outliers across all normal and cancerous tissues/cell-lines in all three cancer types (Figure 30C). The only two cancerous cell-lines, HuH-7 for liver and BJ for skin melanoma, had the lowest number of SE-affected regions. Moreover, skin melanoma CJM had similar proportions in all three outlier groups, ~20%, while COLO679 had more SE in over-predicted outliers.

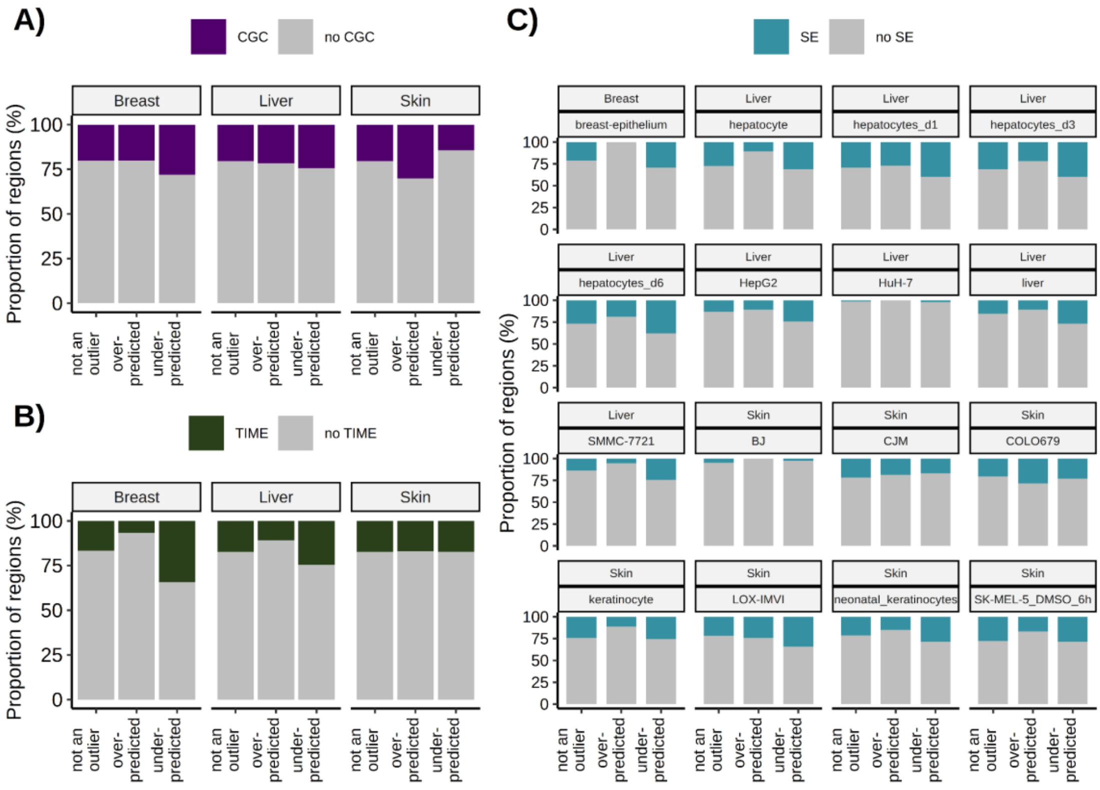


__Figure 30.__ Proportion of annotated 1 Mb genomic regions based on their erroneous status that are affected with A) Cancer Gene Consensus (CGC), B) tumor immune microenvironment (TIME) and C) tissue specific super-enhancers (SE) from both normal and tumorigenic cell-lines or tissues from the SEdb2.0 database

Moreover, I performed an over-representation analysis (ORA) of Hallmarks of cancer and Gene Ontology (GO) terms for all genes found over-, under- and non-outlier 1MB regions. Most common significant enriched pathways in non-outlier regions in all cancers based on ORA of GO terms are pattern specification process, epithelial cell proliferation, homophilic cell adhesion via plasma membrane adhesion molecules and pathways involved in Wnt signaling (Figure 31A). Since there are only 15 over-predicted regions in breast cancer, I did not find any over-represented GO term. However, only significant Hallmark of cancer I found in outlier regions in all cancer was P53 PATHWAY with genes *FBXW7* (ENSG00000109670) and *ANKRA2* (ENSG00000164331) in breast cancer over-predicted outliers. Most significant biological processes (BP) in over-predicted outliers in skin melanoma were response to positive regulation of peptidyl-serine phosphorylation of STAT protein, exogenous dsRNA and NK cell activation in immune response. On the other hand, most significant BP in over-predicted outliers in breast cancer were G protein-coupled purergic nucleotide receptor signaling, regulation of sensory perception of pain and biomineral tissue development. As for under-predicted outliers, those were enriched with keratinization, intermediate filament organization in breast; growth receptor signaling pathway in breast and liver; while in skin melanoma I detected only sensory perception of taste as enriched GO term. When examining the expression of genes in normal tissues corresponding to COO of each cancer type, I found that under-predicted regions had significantly higher gene expression compared to over-predicted ones in all cancers, while over-predicted regions had significantly lower expression than non-outliers (Figure 31B). Similar trend was observed when examining the tissue-specific expression of normal tissues in those erroneous regions but without the same significance levels (Figure 31C).

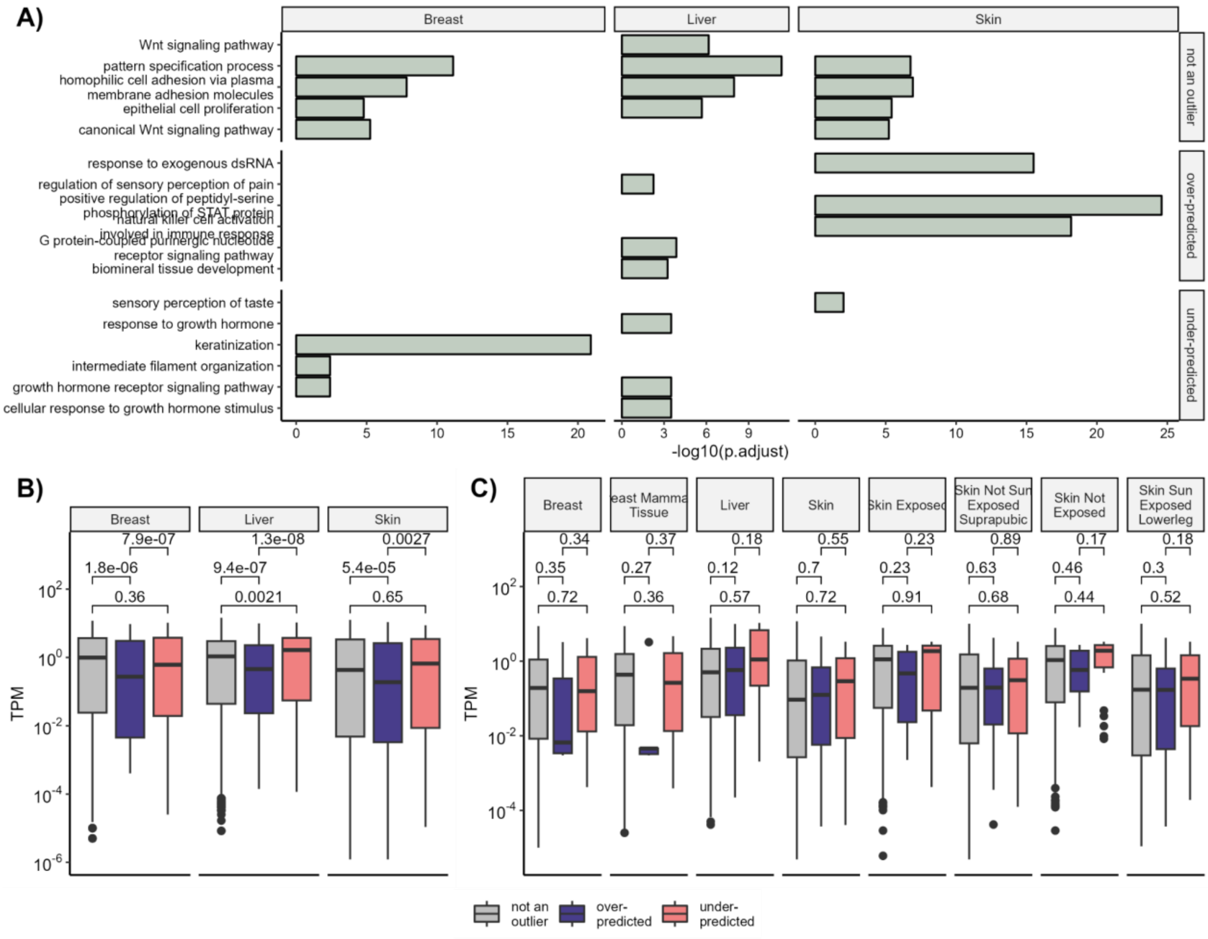


Figure 31. *A) Over-representation analysis of GO terms involved in biological processes. Top 3 most significant by filtering Benjamini-Hochberg p-value of 0.05 are shown for each category of annotated 1Mb genomic regions based on their standard residual (erroneous prediction rate) B) TPM (transcripts per million) expression of genes in normal tissues and C) expression of tissue-specific genes from normal tissues of corresponding cancer type (breast, liver and skin cancer) from GTEX TPM 30 database separated by residual annotation. (Wilxonov-test)*

To further characterize the outliers, I examined how enriched they are with SV-hotspots and kataegis regions. Under-predicted outliers were shown to be significantly enriched in SV-hotspots compared to non-outliers, especially in breast cancer (Pearson's Chi-squared test, p-value < $2.2*10^{-16}$, 0.00196 and 0.01995 for breast, liver and skin cancer) (Figure 32A). Skin *de novo* SV signature 2 hotspots were more abundant in over-predicted outliers, while breast *de novo* SV signature 4 hotspots had more under-predicted outliers (Figure 32B). Only liver *de novo* SV signature 3 hotspots were present in liver as it is the only signature I have detected for liver SV-hotspot signature. Similar observations were detected with kataegis regions where under-predicted outliers in all cancer types had significant enrichment in kataegis regions compared to the others (Figure 32C). The severity of kataegis regions, measured by the number of mutations detected in those regions, was positively correlated with the error rate in kataegis regions in breast cancer (Figure 32D), while a similar trend is not observed for non-kataegis regions.

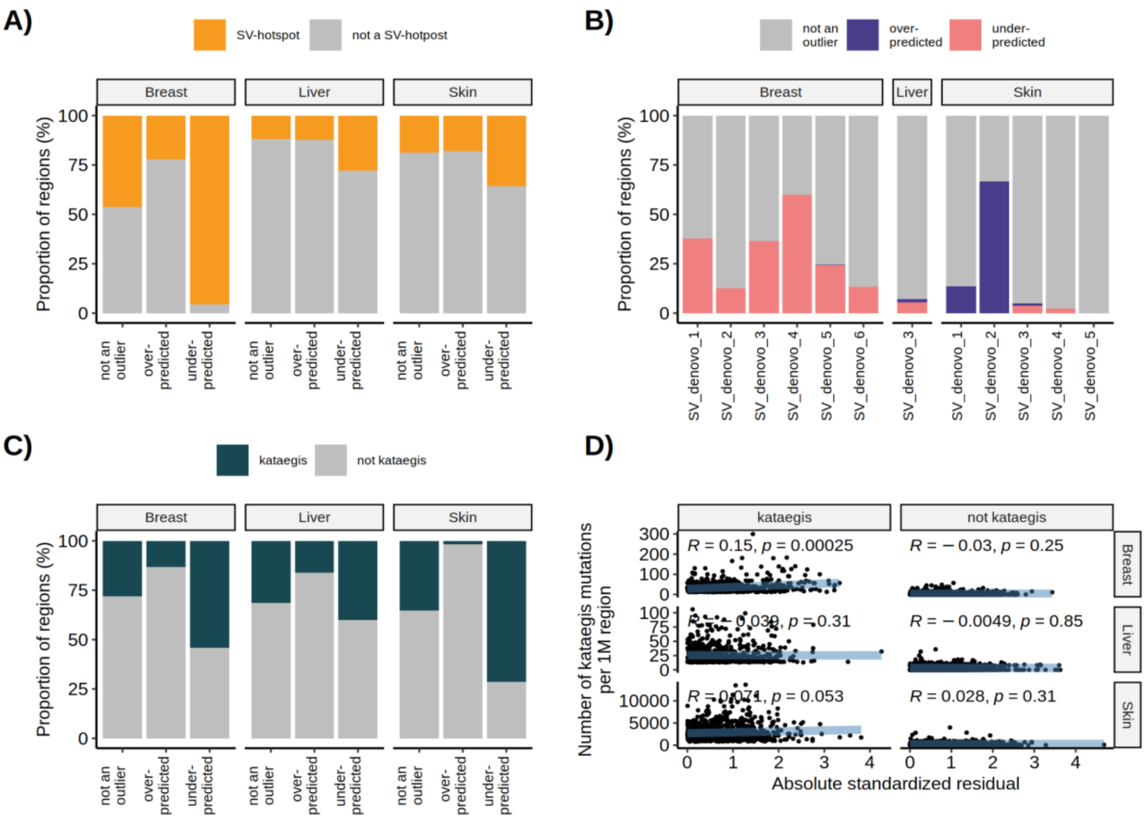


Figure 32. *A) Proportion of 1 Mb regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier B) Proportion of residual annotated regions based on their SV-hotspot mutational signature C) Proportion of regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier D) Spearman correlation between mutations affected by kataegis and erroneous prediction rate, as well as between non-kataegis mutations and erroneous prediction rate*

The Fisher exact test or Pearson's Chi-square test results imply that there was no statistically significant difference between the presence of SV-hotspot and kataegis regions in any defined outlier and non-outlier group, except for non-outliers in breast cancer where p-value was below 0.05 (Figure 33). Although the p-value is not below 0.05, significance is higher in non-outlier regions of all cancer types and in under-predicted outliers of breast cancer where regions affected by both SV-hotspots and kataegis make around 42% (CI 0.32-0.51).

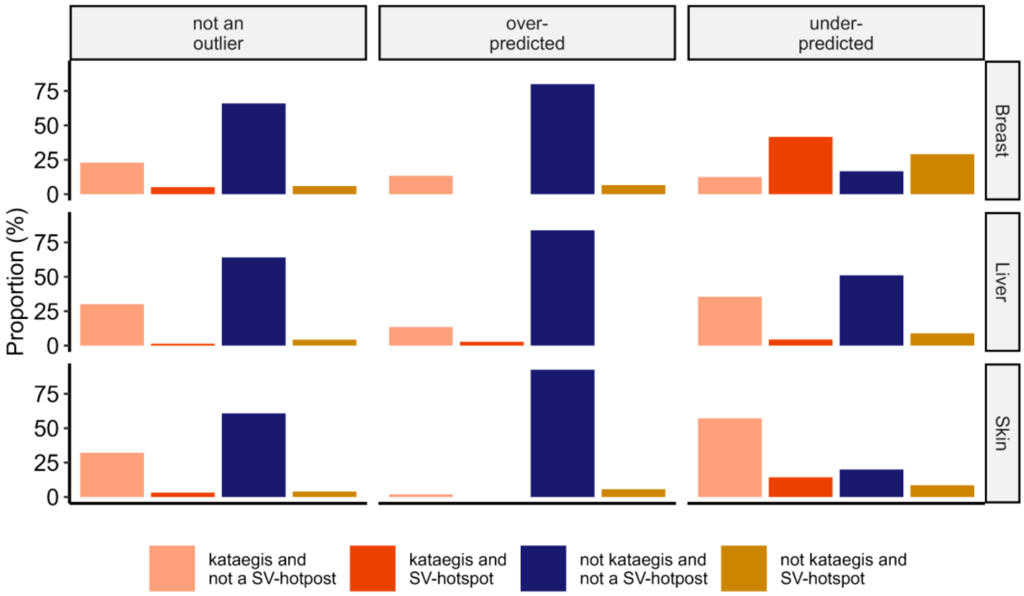


Figure 33. *Count matrix of SV-hotspots and kataegis affected 1 Mb genomic regions of over-predicted, under-predicted and non-outlier regions in breast, liver and skin cancer. Each tested group is colored by Benjamini-Hoechberg corrected p-value of Fisher exact test or Pearson's Chi-square test results*

To conclude whether certain SBS mutational signatures contribute to higher error prediction rate in 1 Mb genomic region COO model, I calculated Pearson's correlation of summarized SBS signatures in regions and absolute standard residual (Table 9). After filtering for significance, only signatures from breast and liver cancer remained. Lack of significant correlations of SBS signatures and absolute standard residual may indicate that in skin melanoma the mutational signatures do not contribute directly to more erroneous prediction rates. In both breast and skin cancer, all mutations showed significant positive correlations with erroneous prediction rates. When I examined the top 5 SBS signatures from both cancer types, I also detected positive correlations of signatures and erroneous prediction rates. The correlations were overall higher in breast than in liver cancer. However, only SBS13, a very well-known annotated APOBEC-related signature, had a higher correlation with ~0.31 than all mutations with ~0.28. As well, this correlation was found to be the overall highest and significant with lowest p-value where SBS13 mutations made up around 19% of total mutations in analyzed 1 Mb genomic regions. The second most positively correlated SBS signature in breast cancer is SBS40 but with lower correlation than all mutations. This signature was also found in liver cancer and contributes to more than 14% and 20% mutations in these cancers. In liver cancer, I found that SBS4 signature is the most significantly positively correlated signature, while SBS36 despite its significant positive correlations contributes to only ~0.06 of all mutations. Moreover, SBS3 signature was found to be in the top 5 most correlated signatures in both cancers.

Table 9. *Top 5 strongest Pearson's correlations of mutational signature or all mutations and absolute standard residual per 1 Mb genomic region for breast and liver cancer alongside the correlation of all mutations with Benjamini-Hochberg corrected p-values lower than 0.05. Table also contains proportion of mutations belonging to each signature in aggregated tumor profiles*

Cancer type	SBS mutational signature	Pearson's correlation	p-value	Proportion of mutations (%)
Breast	All mutations	0.28186263	$2.451223*10^{-38}$	100
	SBS13	0.3112222	$6.788302*10^{-47}$	18.47
	SBS40	0.2690914	$5.603827*10^{-35}$	14.16
	SBS3	0.2599621	$1.086181*10^{-32}$	8.11
	SBS39	0.2457532	$3.214680*10^{-29}$	9.64
	SBS5	0.2226095	$5.637889*10^{-24}$	6.57
Liver	All mutations	0.13074172	$9.840241*10^{-09}$	100
	SBS4	0.1637104	$3.584317*10^{-13}$	4.66
	SBS8	0.1517574	$1.848165*10^{-11}$	10.99
	SBS36	0.1506841	$2.735540*10^{-08}$	0.06
	SBS40	0.1435681	$2.362203*10^{-10}$	20.80
	SBS3	0.1419808	$3.674638*10^{-10}$	1.61

Since 1 Mb genomic regions with higher prediction error rates were associated with SV-hotspots, kataegis, and certain mutational signatures, I calculated the correlation between mutations and histone modifications from normal tissues of the cell-of-origin for each cancer type, grouped according to the influence of different genomic feature. The highest correlations of all mutations or certain mutational signatures with normal epigenome were observed in liver cancer (Figure 34). When examining the correlation of all mutations in all regions, I observed a strong negative correlation with activating histone modifications H3K36me3, H3K4me1, H3K4me3 in all cancers and H3K27ac in breast cancer as expected (Figure 34A). On the other hand, repressive histone modifications H3K27me3 and H3K9me3 were strongly positive in liver cancer, while slightly negative or slightly positive in breast and skin cancers. This trend was observed in all separation groups where regions were separated based on the presence of

SV-hotspots, kataegis or both. However, the correlations of mutations and non-affected regions by SV-hotspots and/or kataegis were higher compared to the ones affected by those genomic characteristics in all cancer types. In case of skin melanoma, the direction of correlation of repressive histone H3K27me3 modification was then changed into the expected direction, from negative to positive. Over- or under-predicted outliers seemed to have weaker or stronger correlations in different cancer types. In breast cancer, the over-predicted cancer types have a much stronger positive association of repressive histone H3K9me3 of 0.754, while in other categories of grouped regions the correlation was weaker around ~0.2. However, the correlations with breast over-predicted outliers were not found to be significant since a low number of regions, 15, was driving the apparently strong correlations.

As far as the correlation of SBS mutational signature and normal epigenomes of cell-of-origin, the age-related SBS1 signature was found to have weaker correlation and even wrong direction of correlation than expected in all cancers (Figure 34B). The SBS1 signature was present in the lowest amount of all other analyzed signatures, especially in skin melanoma. Another signature which had extremely positive association with both active and repressive histone modifications was SBS43 in skin melanoma. Also, in skin melanoma UV-related signatures, SBS7a and SBS7b, had more of the expected directions of correlation with repressive modifications than SBS7c and SBS7d which had overall stronger negative correlations with all modifications. In liver cancer, the mutational signature with much weaker correlations with modifications than all mutations or other signatures were SBS16, SBS23 and SBS36. While in breast cancer, the weakest overall correlation was identified with SBS13 signature. In general, more non-significant correlations were observed with repressive histone marks than with active ones regardless.

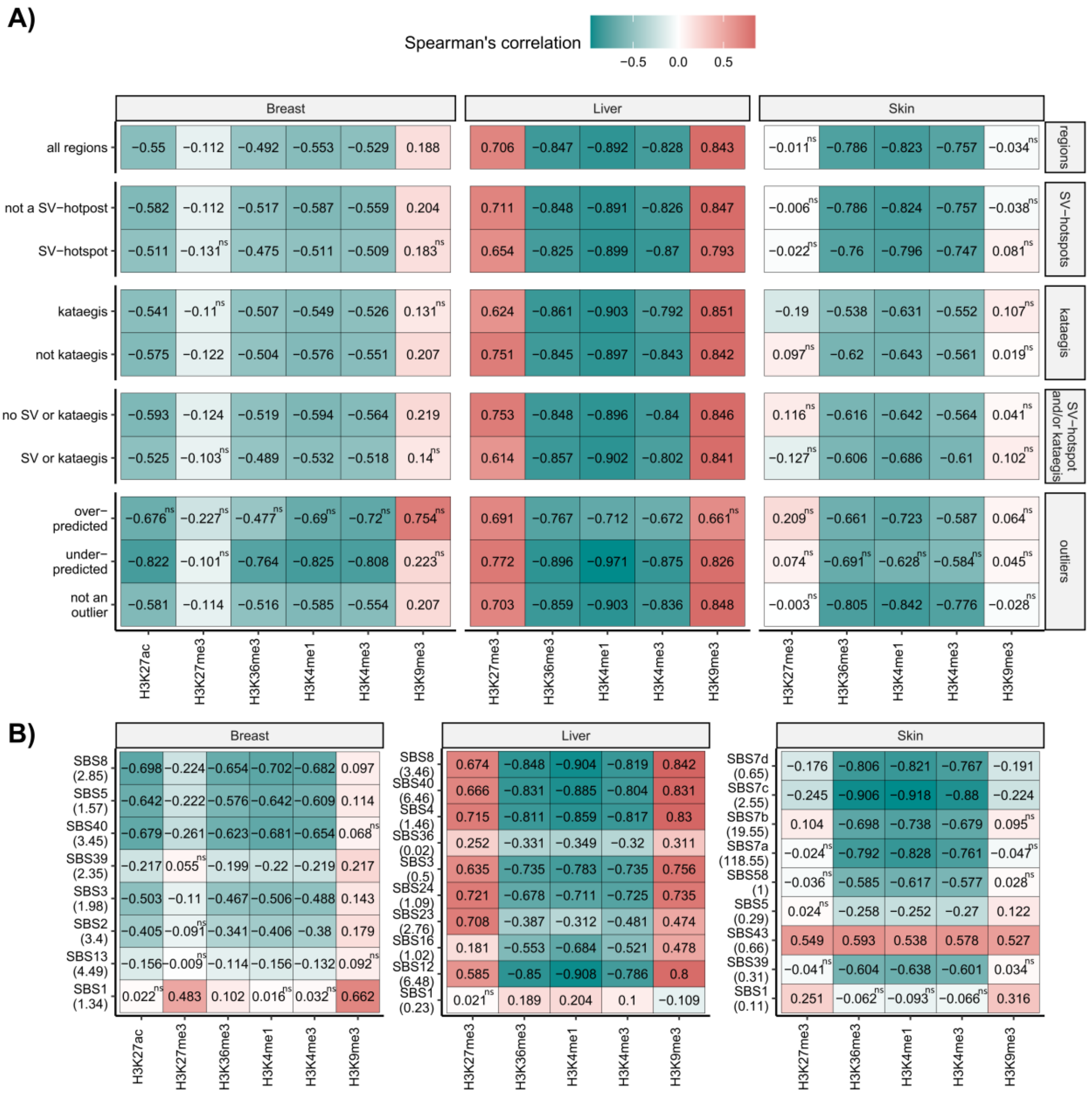


Figure 34. *A) Spearman's correlation of all mutations and histone modifications of cell-of-origin tissue on 1Mb genomic regions annotated based on different criteria (containing or not SV-hotspots or kataegis regions or both, based on outlier annotation) B) Spearman's of SBS mutational signatures and histone modifications of cell-of-origin tissue with number absolute counts divided by 10^6 of signatures in used 1Mb regions. Non-significant correlations, Benjamini-Hoechberg corrected p-value above 10^{-5} are labeled with ^{ns}.*

4.3.1.2. Improving the 1 Mb genomic region COO model

Since I showed that various genomic features and mutational signatures can influence error rates, I removed SV-hotspots and/or kataegis regions, as well as certain mutations originating from a specific mutational signature, to improve the prediction accuracy of the COO models. Only aggregated mutational profiles of breast cancer showed significant improvement in model accuracy compared to the original model with all regions (Wilcoxon test, p-value < 0.05) (Figure 35A). On the other hand, skin melanoma had the opposite effect where removal of especially kataegis affected regions significantly lowered the overall prediction rate with a drop from ~81% to ~64% (Wilcoxon test, p-value = 0.0003). In all models there has been a drop in individual patients COO predictions (Figure 35B) where the most significant one was for skin melanoma with removal of kataegis regions from to ~79% of correctly identified COOs ~63% (Chi-square test, p-value = 1.992*10^{-05}). Removal of the most associated SBS mutational signatures; SBS13 for breast, SBS36, SBS4, SBS8 and SBS40 for liver and SBS43 for skin cancer, only resulted in improved predictions for breast cancer. Although removal of both SV-hotspots and/or kataegis affected regions alongside removed SBS13 mutations resulted in highest breast aggregated COO accuracy of ~65%, the individual COO prediction rates were worse than the original COO model.

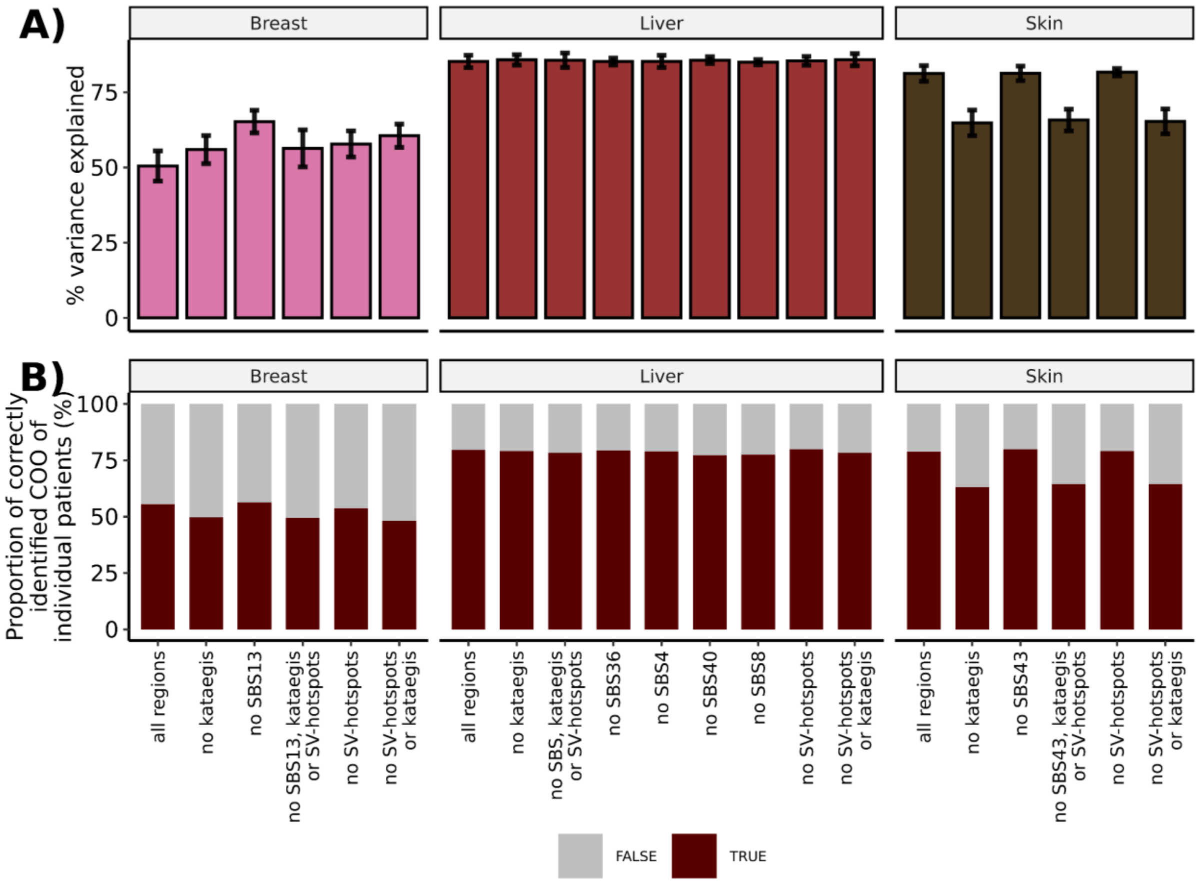


Figure 35. *A) Multiple linear regression models for the top prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WGS based on 1 Mb regions that were either removed due to being affected by SV-hotspots and/or kataegis or excluded certain mutations due to their SBS signature origin. The SBS mutations which were removed are SBS13 for breast, SBS36, SBS4, SBS8 and SBS40 for liver and SBS43 for skin cancer. The figure shows the overall explained variance, depicted as the mean with standard deviation, derived from a 10-fold cross-validation analysis. B) Proportion of correctly and incorrectly identified COO for individual patients based on results of corresponding COO models in A).*

Besides developing the COO random forest and extreme gradient boosting (xgbTree) using all 1Mb regions, I also developed a breast COO model with removed regions affected by SV-hotspots and/or kataegis with removed SBS13 mutations due to previously significantly better multiple-linear regression performance compared to other cancers. Newly obtained accuracies with more complex machine learning models did result in much higher explained variance compared to their respective multiple linear regression model (Figure 36A). The most significant drop in model accuracy was detected for breast cancer where both random forest and extreme gradient boosting COO models had lower variance on aggregated profiles and only ~20% and ~40% correctly identified COO for individual patients (Figure 36B). Moreover, the random forest model also resulted in lower number of correctly identified COO in liver and skin cancers compared to multiple linear regression and xgbTree models despite higher variance on aggregated profiles.

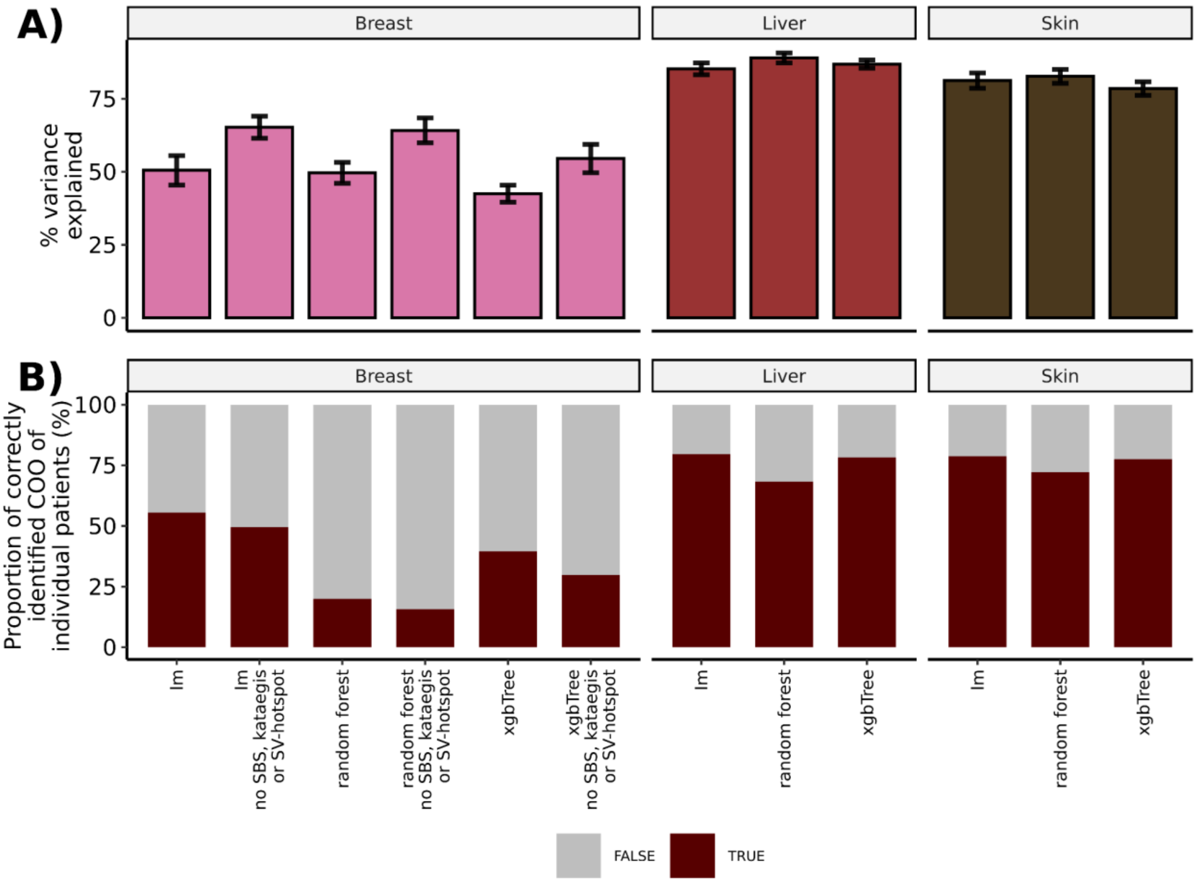


Figure 36. *A) Cell-of-origin models (multiple linear regression-lm, random forest, extreme gradient boosting-xgbTree) for the top prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WGS based on 1 Mb regions that were either removed due to being affected by SV-hotspots, kataegis and excluded certain mutations due to their SBS signature origin. The SBS mutations which were removed are SBS13 for breast due to only increased significant findings in the previous analysis. The figure shows the overall explained variance, depicted as the mean with standard deviation, derived from a 10-fold cross-validation analysis. B) Proportion of correctly and incorrectly identified COO for individual patients based on results of corresponding COO models in A).*

4.3.1.3. Individual patient 1 Mb genomic region COO model

The developed cell-of-origin (COO) models often assigned origins to patients that included brain-related tissues, immune cells, right atrium, muscle, and gastrointestinal tissues (muscle and mucosa), besides the expected correct origins for certain cancers (Figure 37). The models show a high degree of consistency in predicting the same COO tissues across all methods, particularly for liver and skin cancers. However, breast cancer patients experienced the highest rate of incorrect COO assignments. Many correctly identified breast cancer COO cases were misclassified as immune cells, gastrointestinal tissues, and other tissues in the random forest model. Additionally, a significant number of breast cancer patients were incorrectly predicted to have brain or right atrium as their COO across all model setups. In contrast, such observational shifts were less pronounced in other cancer types.

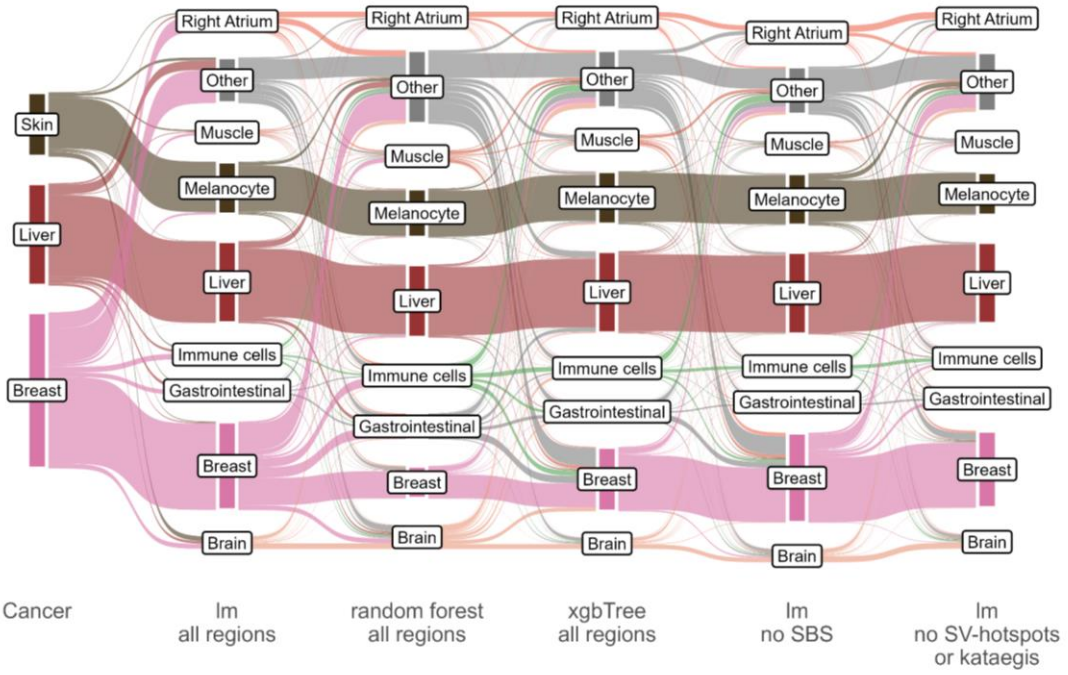


Figure 37. *Alluvial plot illustrating the cell-of-origin (COO) of individual patients across various COO model setups in both aggregated cancer data and individual patient mutational profiles. Assigned COO tissues that represent less than 4% in each cancer type are labeled as "Other". Single-base substitutions (SBS) which were removed are SBS13 for breast cancer, SBS4 for liver cancer and SBS43 for skin melanoma patients.*

When I separated the patients based on their correct cell-of-origin (COO) identification, I observed that in all cancer types, correctly identified COO patients exhibited significantly higher explained variance compared to incorrectly identified ones (Figure 38). Specifically, liver and skin melanoma cancers showed a significantly higher number of mutations in correctly identified patients, whereas this trend was not observed in breast cancer. Although the differences in the number of kataegis events were only marginally significant in liver cancer, incorrectly identified COO patients tended to have slightly more kataegis events. On the other hand, structural variation (SV) hotspots were slightly more prevalent in correctly identified COO patients for both breast and skin melanoma cancers. Conversely, in liver cancer, incorrectly identified COO patients had a higher occurrence of SV-hotspots. Despite these observations, the results for kataegis and SV-hotspots did not result in statistical significance according to the Wilcoxon test. Nonetheless, the trends suggest a potential link between the presence of kataegis and SV-hotspots with the accuracy of COO predictions.

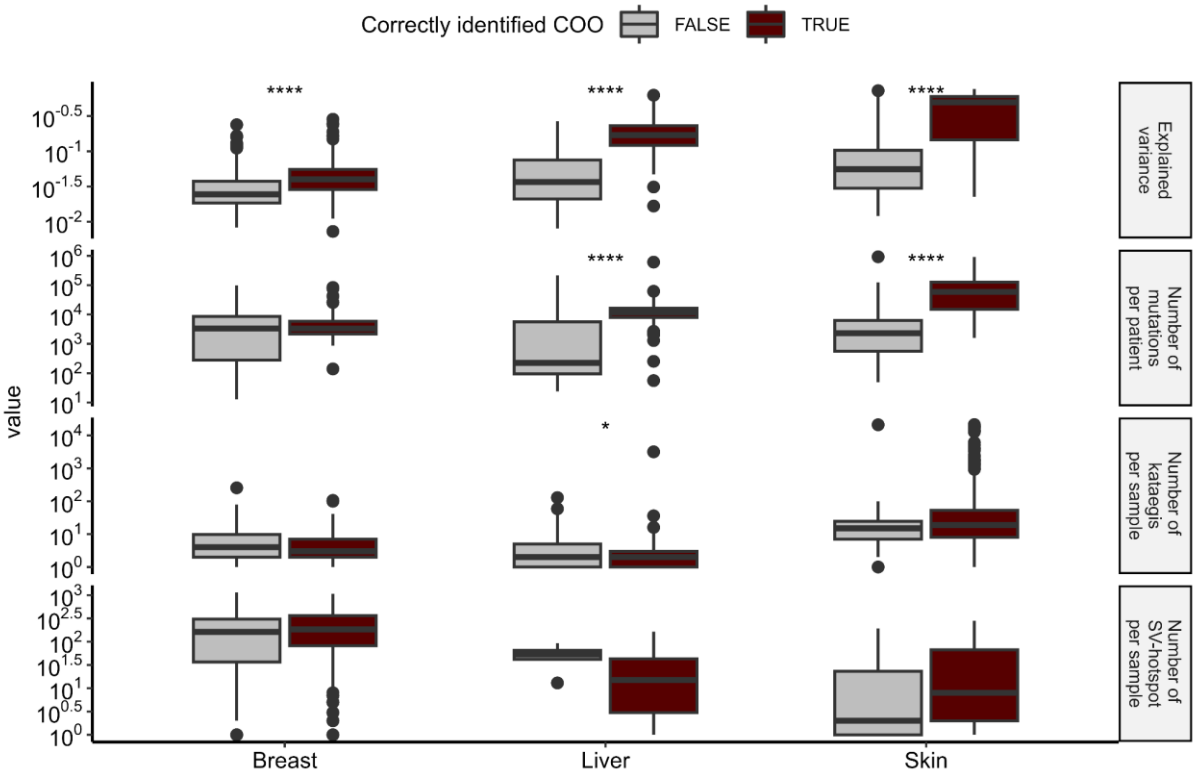


Figure 38. *The distribution of explained variance, the number of mutations, kataegis and SV-hotspots separated by correct or incorrect COO identification of individual patients. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

To further investigate the differences between correctly and incorrectly identified COO patients, I examined the proportion of SBS mutational signatures per patient that were found to be most prevalent in certain cancer type or showed strong correlations with more erroneous predictions in the 1 Mb genomic COO model (Figure 39). For breast cancer, incorrectly identified COO patients had a higher proportion of SBS13 signature (Wilcoxon test, p-value=0.0011). Although SBS2 and SBS3 did not show a statistically significant difference, they were also more prevalent in incorrectly identified patients. In contrast, correctly identified breast COO patients had more SBS40 and SBS1. For liver cancer, incorrectly identified COO patients had higher percentages of SBS16, SBS3, SBS8, SBS1, and SBS40 signatures. Conversely, correctly identified patients exhibited a higher proportion of the SBS4 signature. In skin melanoma, the most notable differences were the significantly higher proportions of the SBS7a signature in correctly identified COO patients (Wilcoxon test, p-value < $2.2*10^{-16}$) and lower proportions of SBS7b in incorrectly identified patients (Wilcoxon test, p-value < $2.2*10^{-16}$).

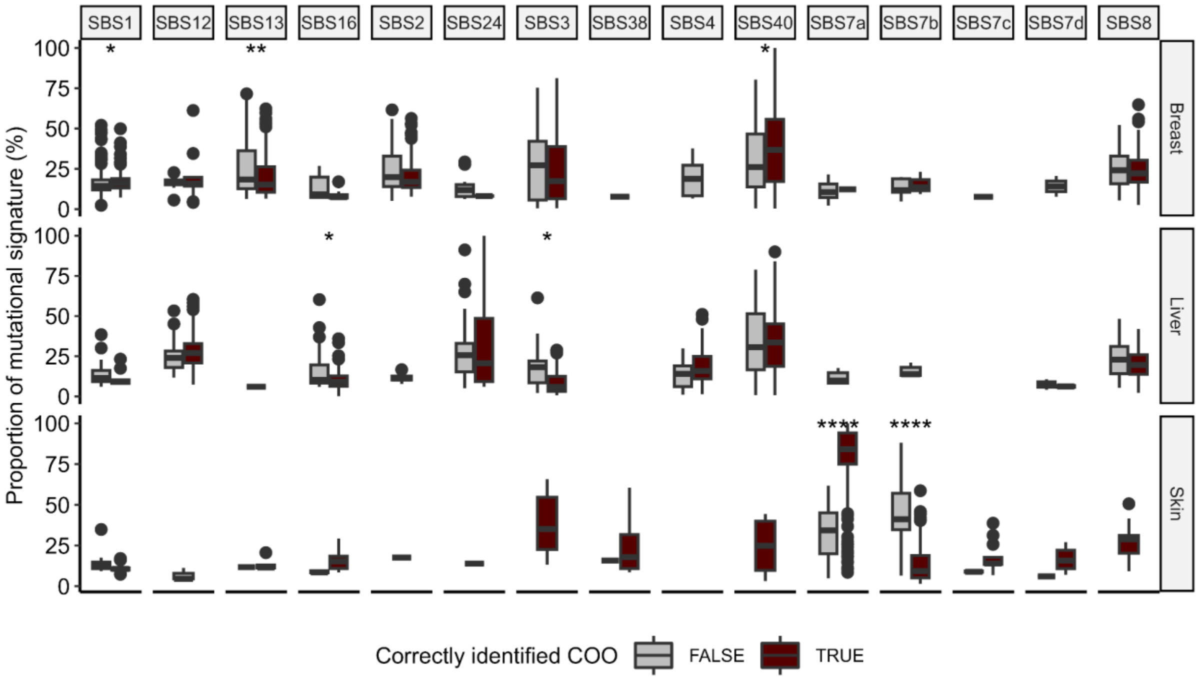


Figure 39. *Proportion of SBS mutational signatures per patient separated by correct or incorrect COO identification of individual patients. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.1.3.1 Breast cancer

I separated the breast cancer individual patients' COO results based on their histological type. As shown in Figure 40A, the breast cancer subtypes with the highest prediction accuracy did not necessarily have the highest number of mutations per patient. The calculated median of all percentages of correctly identified COO patients across all developed COO models was highest for the 12 duct micropapillary carcinoma patients, at approximately 75%, despite their mutational load being lower compared to other subtypes. Following this, the best-performing subtypes with a median above 60% included 5 metaplastic carcinoma patients with a higher mutational count, 4 carcinoma with apocrine differentiation patients, and 17 mucinous adenocarcinoma patients with a lower mutational count, as well as 91 patients with unknown histological subtypes who had a higher mutational count. Conversely, some breast cancer subtypes with lower median proportions of correct identification exhibited much larger variance in mutational count per patient, such as infiltrating duct carcinoma and lobular carcinoma, which generally had more patients. The worst-performing subtypes, each with only 1 or 2 patients, were neuroendocrine, medullary, pleomorphic, as well as duct and lobular carcinoma. Pleomorphic patients have the smallest mutational count.

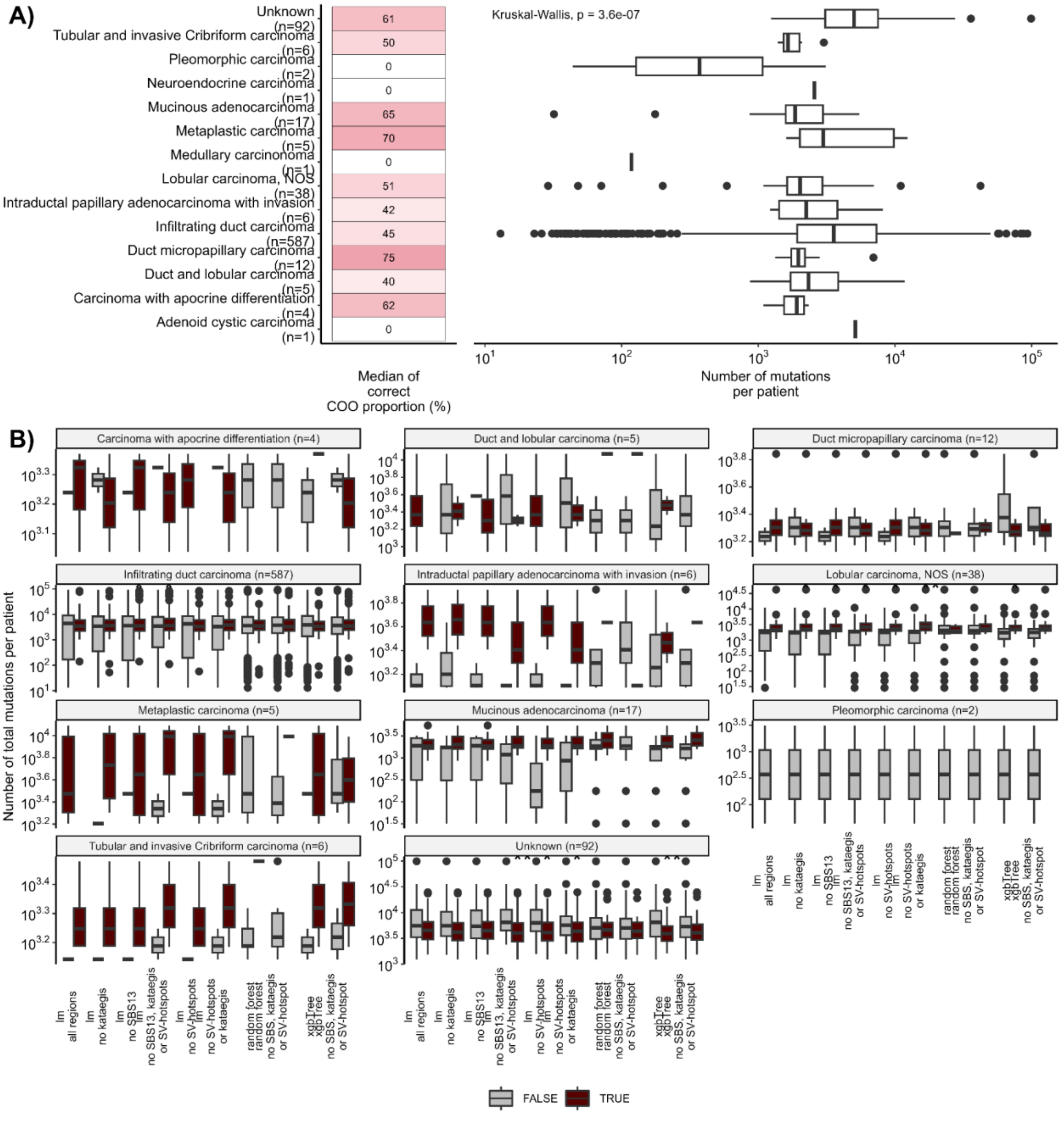


Figure 40. *A) Median of correctly identified cell-of-origin (COO) proportions across all developed COO model setups for breast cancer patients, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. B) Distribution of the number of mutations per patient separated by correctly identified COO for each histological subtype across various COO models. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

The number of mutations per patient was further analyzed by whether the COO was correctly identified across various COO models (Figure 40B). It was observed that breast cancer subtypes with a lower number of samples mostly suffered from incorrect COO assignments by random forest models like duct and lobular carcinoma. In all other models, correctly identified COO patients had significantly higher mutational counts compared to those incorrectly identified. However, this trend was not observed for infiltrating duct carcinoma and unknown breast cancer subtypes who also had the highest number of patients. Both pleomorphic carcinoma patients were not correctly identified by any COO model. Additionally, cancer types with lower patient counts suffered from a lack of statistical significance due to the smaller sample sizes.

When examining the percentage of explained variance, breast cancer patients had quite low variance, in majority cases below 10%, even for correctly identified COO patients (Supplementary Figure 9). In all breast cancer histological subtypes the correctly identified COO patients had higher variance explained, which was especially significantly in infiltrating duct carcinoma and unknown breast cancer types whose more higher percentage of explained variance compared in individual patients compared to other types.

Based on the provided PAM50 annotation for 86 breast patients from Kubler et al. (2019), almost all samples had incorrectly assigned COO shown regardless of the COO model setup shown as median of all percentages of correctly identified COO patients across all developed COO models (Figure 41A). Only a few samples, a maximum of 4, had correctly assigned COO in basal subtypes and one or 2 in luminal B type. There was a statistically significant difference in mutational count per patient between the PAM50 groups (Kruskal-Wallis, $p=5*10^{-4}$) with luminal A having the smallest mutational count and basal having the largest. In all PAM50 groups the most abundant histological type was infiltrating duct carcinoma.

Furthermore, out of 371 breast cancer patients with homologous recombination deficiency (HRD) annotated by either HRDetect or CHORD from Štancl et al. (2022), COOs of HRD samples were mostly incorrectly assigned (Figure 41B). The average correct COO prediction rate was 29% for HRD and 66% for non-HRD group. Also, HRD patients had significantly higher number of mutations compared to non-HRD group (Wilcoxon test, $p<2.2*10^{-16}$). Both groups had a majority of infiltrating duct carcinoma patients, but the non-

HRD group also included more cases of lobular carcinoma, mucinous adenocarcinoma, and duct micropapillary breast cancer.

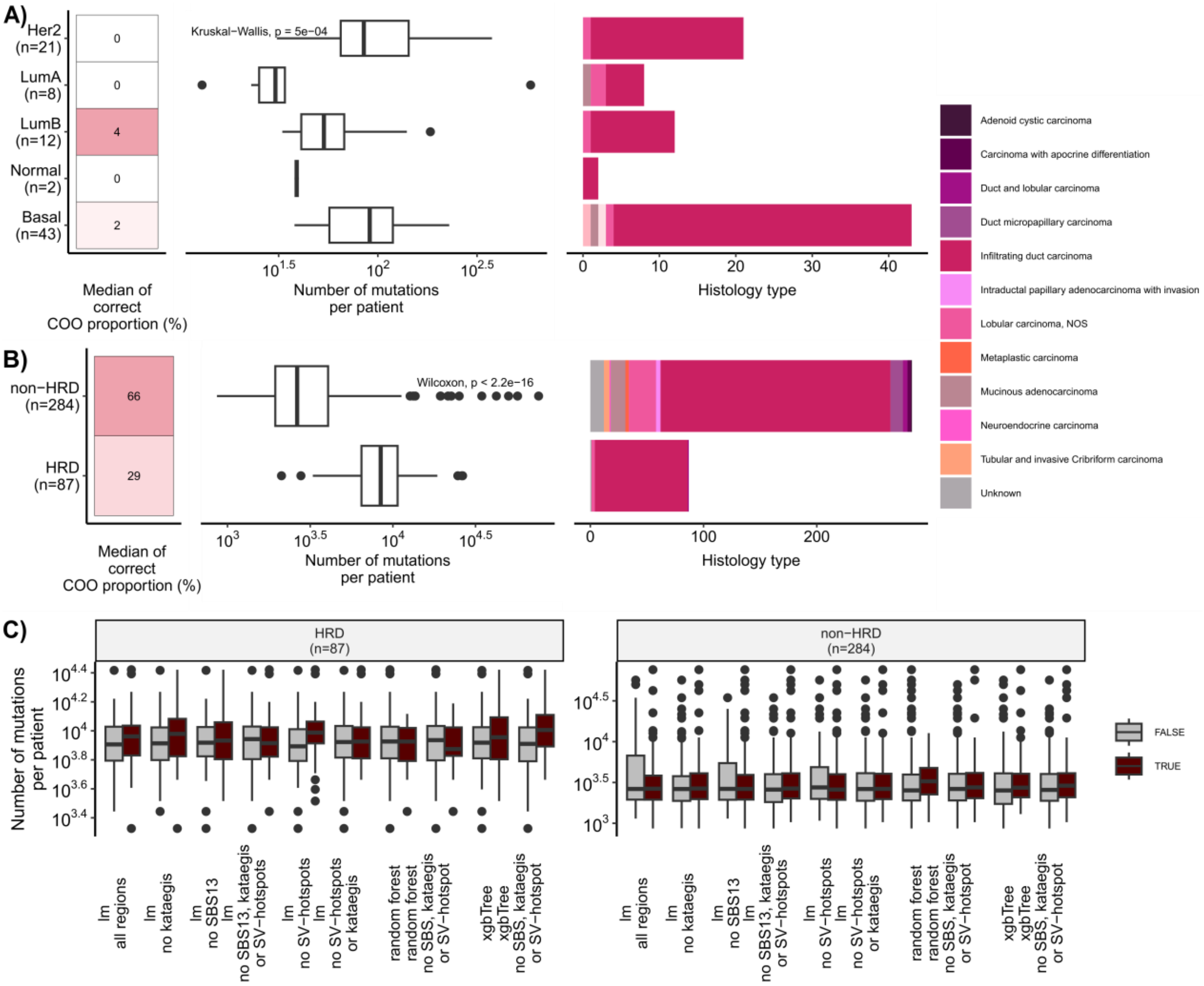


*Figure 41. A) Median of correctly identified cell-of-origin (COO) proportions across all developed COO model setups for breast cancer patients, separated by their PAM50 subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. B) Median of correctly identified cell-of-origin (COO) proportions across all developed COO model setups for breast cancer patients, separated by their HRD status. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

A more detailed view of the number of mutations per patient, separated by whether the COO was correctly identified across various COO models (Figure 41C), showed that correctly identified HRD patients had a significantly higher mutational count than non-HRD patients. This was particularly prominent for the COO model with removed SV-hotspots in both multiple linear regression and extreme gradient boosting models. On the other hand, non-HRD patients had more similar mutational count profiles regardless of correct COO identification. However, based on the explained variance, it was observed that correctly identified COO in both HRD and non-HRD patients had significantly higher variance explained than incorrectly identified ones.

4.3.1.3.2 Liver cancer

Separating liver cancer patients into histological types resulted in the majority of samples belonging to hepatocellular carcinoma (HCC), which displayed the highest and largest variance in the number of mutations per patient (Figure 42A). The highest median proportion of correctly identified cell-of-origin (COO) patients, above 80%, was detected for HCC, hepatocellular adenoma, and fibrolamellar hepatocellular carcinoma, each with only 5 samples.

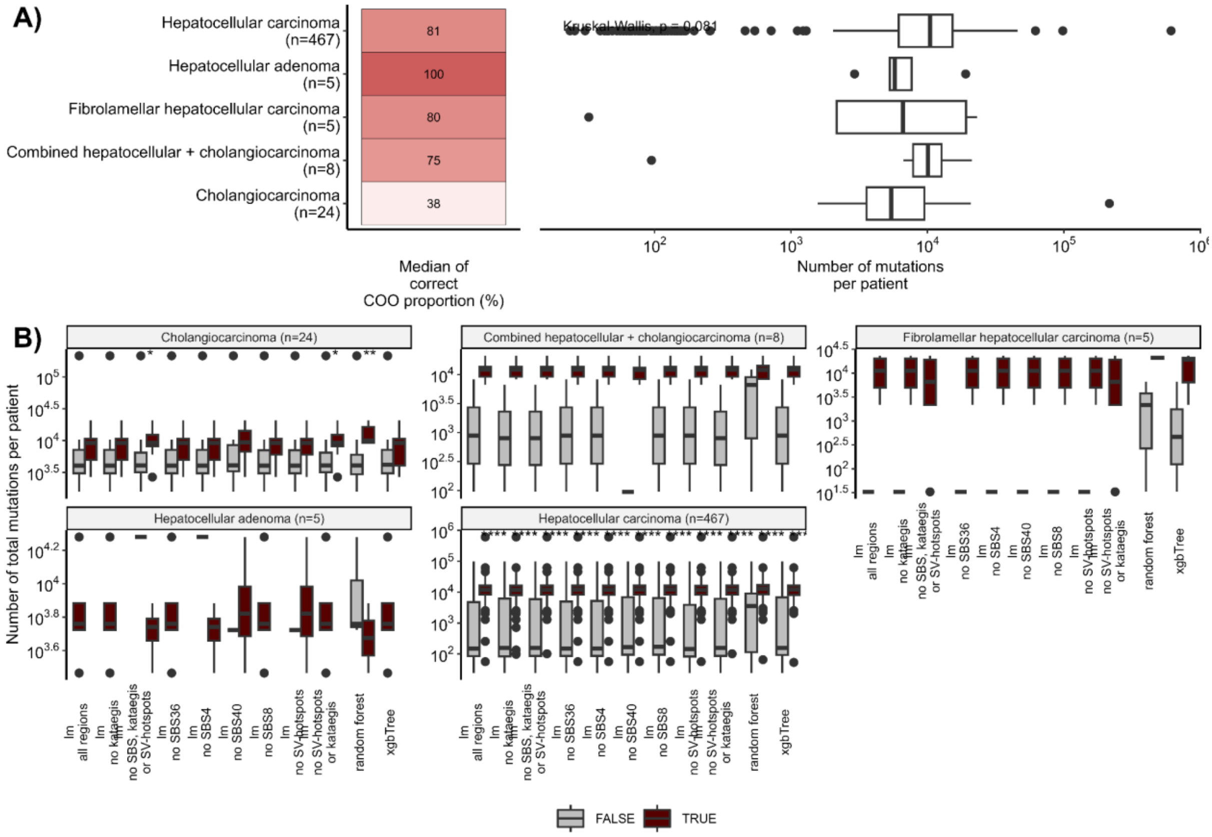


Figure 42. *A) Median of correctly identified cell-of-origin (COO) proportions across all developed COO model setups for liver cancer patients, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. B) Distribution of the number of mutations per patient separated by correctly identified COO for each histological subtype across various COO models. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

The worst performance in correct COO prediction for individual patients was observed in 8 combined hepatocellular and cholangiocarcinoma patients and 24 cholangiocarcinoma patients. Furthermore, cholangiocarcinoma had the smallest median number of mutations per patient. The number of mutations per patient was further analyzed by whether the COO was correctly identified across various COO models (Figure 42B). In all liver cancer subtypes and developed COO models, correctly identified patients had much higher mutational counts per patient compared to incorrectly labeled patients. However, in the developed random forest COO model for hepatocellular adenoma, more incorrectly annotated patients with higher mutational counts were detected compared to other COO model setups.

In all liver cancer histological subtypes the correctly identified COO patients had higher variance which was especially significantly in hepatocellular carcinoma and cholangiocarcinoma whose higher percentage of explained variance in individual patients compared to other histological types (Supplementary Figure 10).

4.3.1.3.3 Skin melanoma

The analysis of the COO predictions for skin melanoma patients, separated by their histological types, showed that the majority of histological types had quite high overall correct COO rate across all COO model setups (Figure 43A). The highest median proportions of correctly identified COO were observed in 2 lentigo malignant melanoma (100%), 10 unknown (95%), 79 nodular melanoma (93%), and 83 superficial spreading melanoma (92%) patients. These subtypes exhibited varying mutational loads, with some showing higher mutation counts than others. On the other hand, the lowest median proportions of correctly identified COO were for 5 mucosal lentiginous patients and 70 malignant melanoma, NOS. Those two types had lower mutational count per patients alongside 60 acral lentiginous melanoma patients with 63%. The Kruskal-Wallis test ($p < 2.2*10^{-16}$) confirms the statistical significance of mutational counts between skin melanoma histological types.

Moreover, I detected that correctly identified COO patients generally had a higher mutational load compared to incorrectly labeled patients across all types in all COO model setups (Figure 43B). Notably, subtypes such as malignant melanoma, NOS, and superficial spreading melanoma displayed significantly higher mutation counts in correctly identified COO patients. Conversely, certain subtypes like desmposlasmic and mucosal lentiginous melanoma had fewer samples, making the results less statistically robust. In the random forest COO model, incorrectly identified COO patients with mucosal lentiginous melanoma exhibited more mutations than those correctly identified.

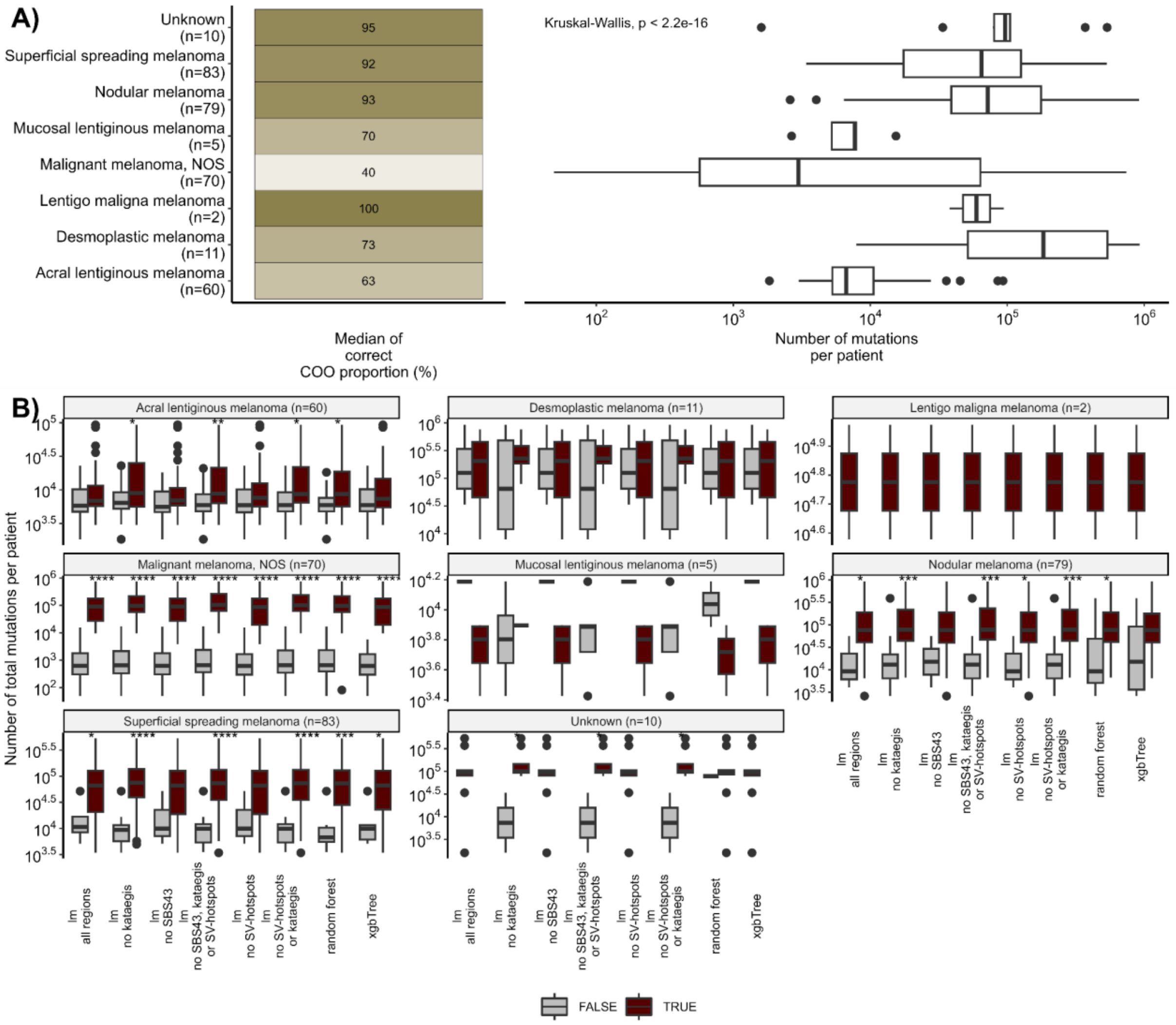


Figure 43. *A) Median of correctly identified cell-of-origin (COO) proportions across all developed COO model setups for skin melanoma cancer patients, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. B) Distribution of the number of mutations per patient separated by correctly identified COO for each histological subtype across various COO models. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Interestingly, the explained variance in correctly identified COO patients varied significantly across different melanoma subtypes and COO model setups (Supplementary Figure 11). Most melanoma types exhibited higher explained variance in correctly identified COO patients, indicating a strong relationship between mutational patterns and accurate COO prediction. However, desmoplastic melanoma showed an opposite trend across all developed COO model setups, with incorrectly identified COO patients having higher explained variance. Additionally, the removal of kataegis-affected regions from the skin melanoma COO model setups resulted in a significant reduction in explained variance compared to other setups.

4.3.2 Topologically-associated domains based cell-of-origin predictive model

To test the influence of 3D chromatin organization on the prediction accuracy of the models, I developed numerous regression models using various tissue-specific topologically associated domains from normal liver tissue, epidermal keratinocytes (NHEK), and human mammary epithelial cells (HMEC) called with various TAD callers for breast, liver, and skin cancers (Figure 44). The results of models trained on aggregated mutational profiles from each cancer type, regardless of the tissue-specific TADs on which the mutations were summarized, show that correctly identified COO was achieved in most WGS data (Figure 44A).

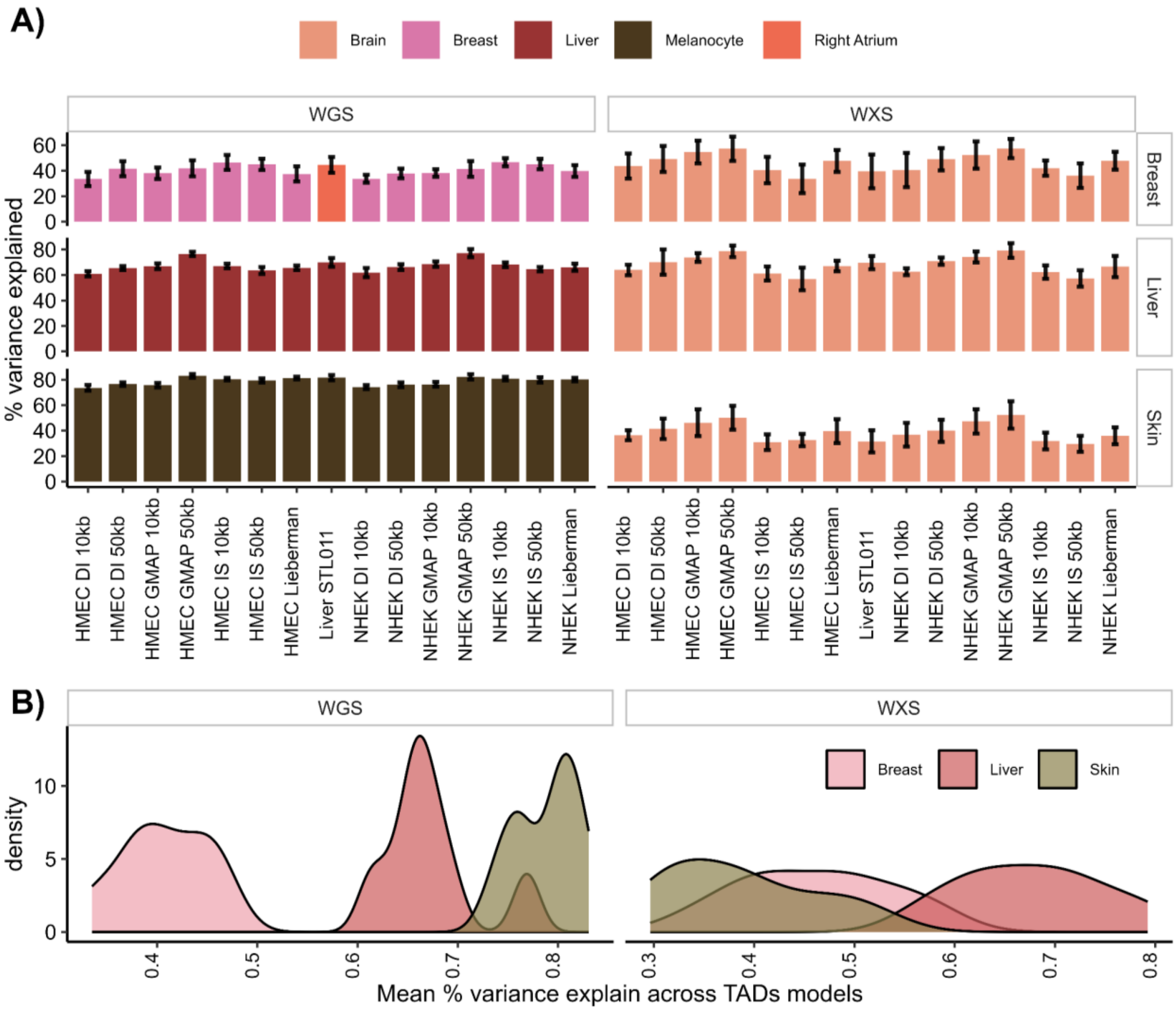


Figure 44. *A) Multiple linear regression models for the prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WGS from HMEC, NHEK cell-lines and liver tissue called using various programs from TADBK and 3D genome browser databases colored by their identified cell-of-origin. The models were trained on an extended set of 101 tissue sets but showing only the best one for each TADs. The overall explained variance is reported across the 10-fold cross-validation. B) Distribution of variance explained by each best topologically associated domains (TADs) model displayed as density plot*

For liver and skin melanoma aggregated mutational profiles, the TAD calling tool GMAP with a 50 kb resolution resulted in the highest percentage of explained variance, approximately 83% for melanoma and 77% for liver, irrespective of the cell-line used. In contrast, for breast cancer, the highest explained variance in the TAD-based COO models was 46%, achieved with the IS 10kb resolution, also independent of the cell-line. On the other hand, WXS data did not result in the correct COO prediction regardless of tissue specificity of used TADs, In WXS data, brain-related tissues were consistently the top predicted COO for all three cancers with generally lower variance explained across all models. Figure 44B contains the density plots of the mean percentage variance explained across TAD models for each cancer type. Out of the three cancers, skin cancer had the highest average accuracy across all TADs, with a mean of approximately 86%. Liver cancer followed with an average accuracy of around 67%, and breast cancer had the lowest average accuracy at around 40%.

First, I analyzed the number of mutations of aggregated cancer profiles normalized by TAD lengths for all used TADs (Figure 45A). The analysis revealed that breast, liver, and skin cancers exhibited similar mutational count patterns across different TAD calling methods. Notably, the mutational counts were somewhat higher for TADs identified using the IS 50kb method, while lower counts were observed for TADs called by the Lieberman method in both cell lines, HMEC and NHEK. Next, I examined the variation in TAD lengths across different TAD calling tools and cell lines or tissues (Figure 45B). The TAD length distributions varied significantly between different TAD calling tools, with less pronounced differences between cell lines. For instance, GMAP 50kb called TADs were substantially longer than IS 50kb TADs in both HMEC and NHEK cell lines, with median lengths around 10^6 bp for GMAP compared to 10^5 bp for IS. The DI method also exhibited longer TAD lengths at the 50kb resolution compared to the 10kb resolution, across all cell lines and tissues analyzed.

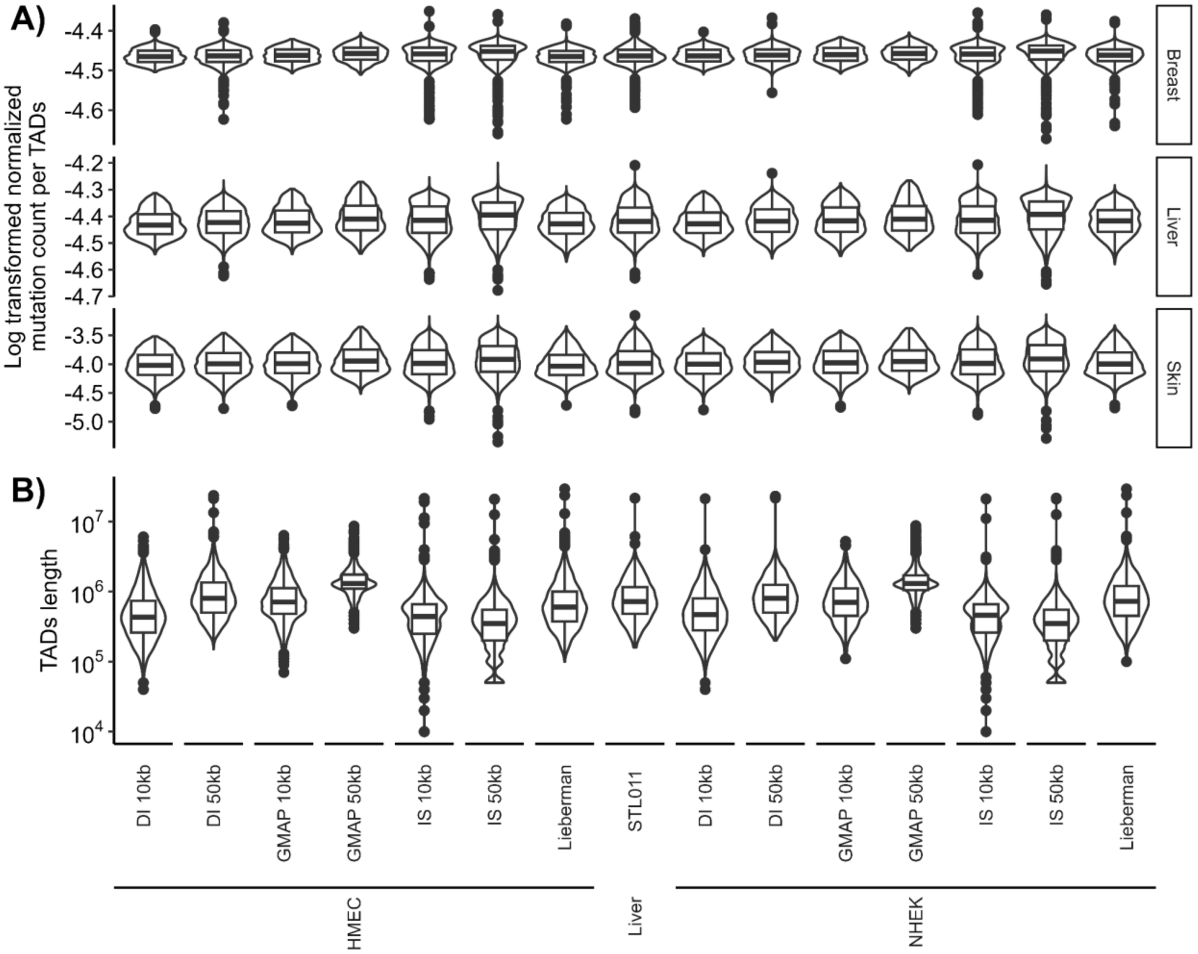


Figure 45. *A) Distribution of normalized mutation count per TAD length from HMEC, NHEK cell-lines and Liver tissue called using various programs from TADBK and 3D genome browser databases. B) Distribution of gene lengths used for development of COO models in breast cancer. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

Prediction of COO for individual patients resulted in varying percentages of correctly identified COO using different tissue specific TADs (Figure 46A). Breast cancer patients had the lowest percentage of correct COO predictions from around 10 to 20 %, while skin melanoma had the highest percentage from around 55 to 75%. On the contrary, liver patients had the most notable percentage fluctuations from around 20 to 65%. The fluctuations in COO prediction accuracy across all cancer types were more related to the tools used for calling the TADs rather than the tissue specificity of the TADs. This is evident from the results showing that both NHEK and HMEC TADs called with the DI tools had the lowest number of correctly identified COO patients, whereas the GMAP 50kb tool setting resulted in the highest accuracy for both liver and skin melanoma. Interestingly, this pattern was not observed for breast cancer patients. Correctly identified COO patients had a significantly higher mutational burden and explained variance by the best model compared to incorrectly classified patients across all tissue-specific TADs in all cancers (Figure 46BC). The most significant separation and difference between correctly and incorrectly identified COO patients in terms of these two features were detected in liver and skin melanoma using NHEK and HMEC TADs called with the GMAP 50kb tool.

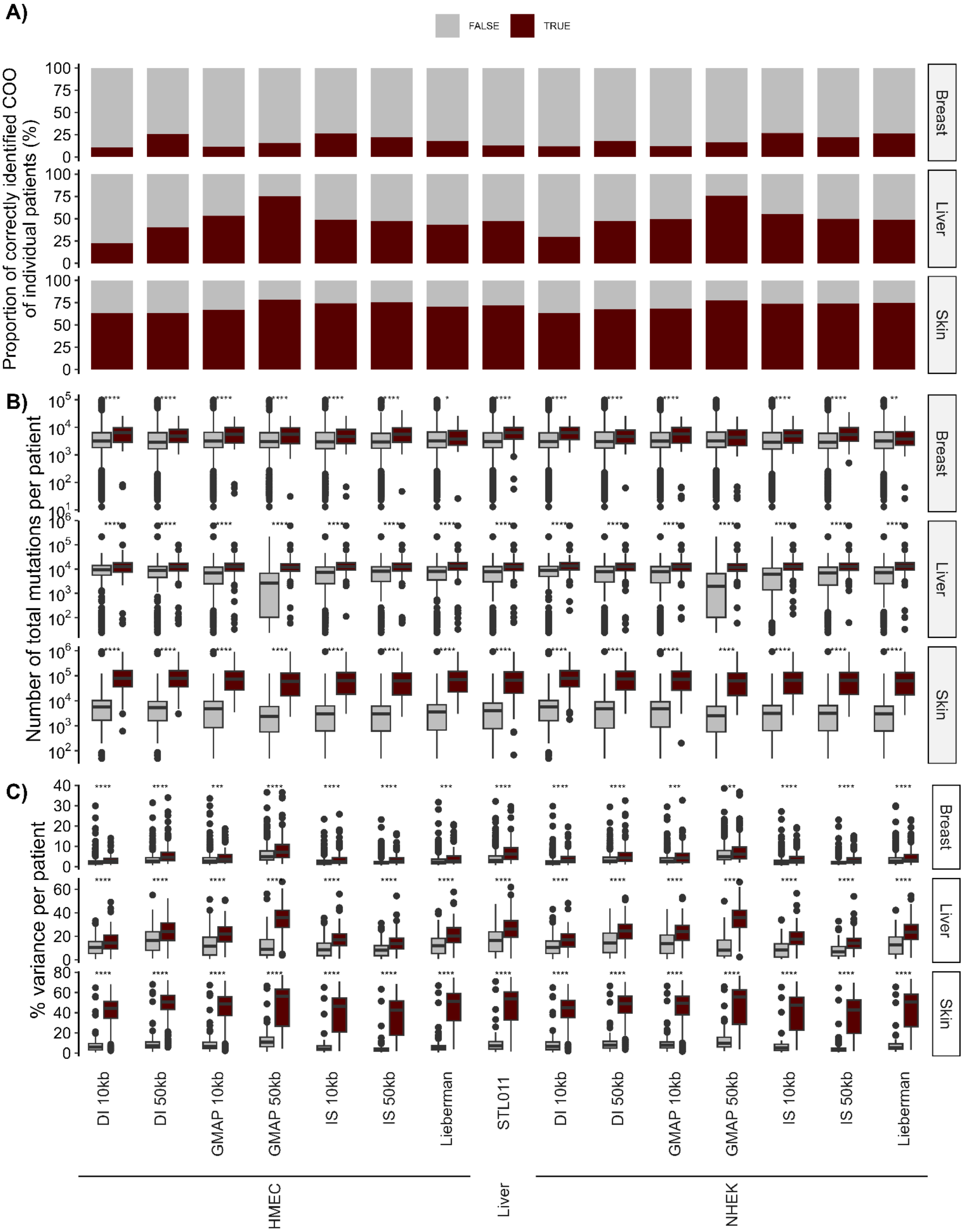
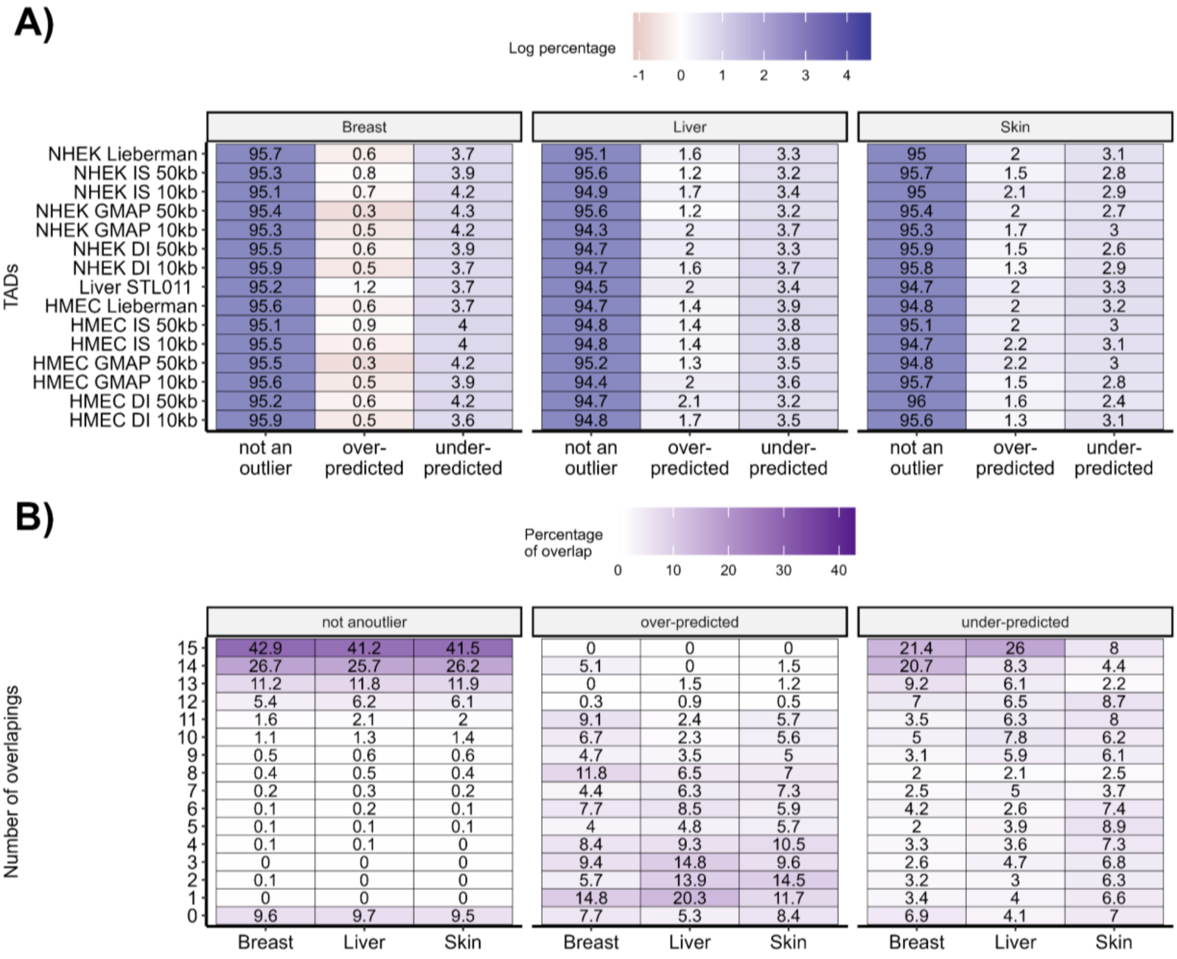


Figure 46. *A) Distribution of normalized mutation count per TAD length from HMEC, NHEK cell-lines and Liver tissue called using various programs from TADBK and 3D genome browser databases. B) gene lengths used for development of COO models in breast cancer. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Since tissue-specificity was not a crucial factor for correct prediction of COO, I took only one tissue-specific TAD corresponding to the cancer type to train the models with indel mutations. The COO model with aggregated indel counts per TADs did not result in correct COO for any of the analyzed cancer types when indels were used (Supplementary Figure 12). The top best models were either of adipose or immune origin with significantly lower explained variance compared to SBS COO models (Wilcoxon test, p-value = 0.001), with the lowest one being for skin melanoma of around 24%. Regarding the individual patients COO predictions, I detected that less than 10% of breast and less than 3% of liver and skin cancer patients had their COO correctly predicted by the model.

To determine if the annotated residual TADs significantly differed among cell lines and TAD calling tools within each cancer type, I calculated the total percentage of annotated TAD residuals and the number of overlapping annotated TAD residuals within TAD groups, defined by cell type/tissue and TAD calling tool (Figure 47).



TADs	Breast not an outlier	Breast over-predicted	Breast under-predicted	Liver not an outlier	Liver over-predicted	Liver under-predicted	Skin not an outlier	Skin over-predicted	Skin under-predicted
NHEK Lieberman	95.7	0.6	3.7	95.1	1.6	3.3	95	2	3.1
NHEK IS 50kb	95.3	0.8	3.9	95.6	1.2	3.2	95.7	1.5	2.8
NHEK IS 10kb	95.1	0.7	4.2	94.9	1.7	3.4	95	2.1	2.9
NHEK GMAP 50kb	95.4	0.3	4.3	95.6	1.2	3.2	95.4	2	2.7
NHEK GMAP 10kb	95.3	0.5	4.2	94.3	2	3.7	95.3	1.7	3
NHEK DI 50kb	95.5	0.6	3.9	94.7	2	3.3	95.9	1.5	2.6
NHEK DI 10kb	95.9	0.5	3.7	94.7	1.6	3.7	95.8	1.3	2.9
Liver STL011	95.2	1.2	3.7	94.5	2	3.4	94.7	2	3.3
HMEC Lieberman	95.6	0.6	3.7	94.7	1.4	3.9	94.8	2	3.2
HMEC IS 50kb	95.1	0.9	4	94.8	1.4	3.8	95.1	2	3
HMEC IS 10kb	95.5	0.6	4	94.8	1.4	3.8	94.7	2.2	3.1
HMEC GMAP 50kb	95.5	0.3	4.2	95.2	1.3	3.5	94.8	2.2	3
HMEC GMAP 10kb	95.6	0.5	3.9	94.4	2	3.6	95.7	1.5	2.8
HMEC DI 50kb	95.2	0.6	4.2	94.7	2.1	3.2	96	1.6	2.4
HMEC DI 10kb	95.9	0.5	3.6	94.8	1.7	3.5	95.6	1.3	3.1

Number of overlapings	not anoutlier Breast	not anoutlier Liver	not anoutlier Skin	over-predicted Breast	over-predicted Liver	over-predicted Skin	under-predicted Breast	under-predicted Liver	under-predicted Skin
15	42.9	41.2	41.5	0	0	0	21.4	26	8
14	26.7	25.7	26.2	5.1	0	1.5	20.7	8.3	4.4
13	11.2	11.8	11.9	0	1.5	1.2	9.2	6.1	2.2
12	5.4	6.2	6.1	0.3	0.9	0.5	7	6.5	8.7
11	1.6	2.1	2	9.1	2.4	5.7	3.5	6.3	8
10	1.1	1.3	1.4	6.7	2.3	5.6	5	7.8	6.2
9	0.5	0.6	0.6	4.7	3.5	5	3.1	5.9	6.1
8	0.4	0.5	0.4	11.8	6.5	7	2	2.1	2.5
7	0.2	0.3	0.2	4.4	6.3	7.3	2.5	5	3.7
6	0.1	0.2	0.1	7.7	8.5	5.9	4.2	2.6	7.4
5	0.1	0.1	0.1	4	4.8	5.7	2	3.9	8.9
4	0.1	0.1	0	8.4	9.3	10.5	3.3	3.6	7.3
3	0	0	0	9.4	14.8	9.6	2.6	4.7	6.8
2	0.1	0	0	5.7	13.9	14.5	3.2	3	6.3
1	0	0	0	14.8	20.3	11.7	3.4	4	6.6
0	9.6	9.7	9.5	7.7	5.3	8.4	6.9	4.1	7

Figure 47. *A) Percentage of annotated TAD from HMEC, NHEK cell-lines and Liver tissue called using various programs from TADBK and 3D genome browser databases based on their erroneous status; over-predicted are TADs with standard residuals lower than -2 and under-predicted TADs with standard residuals higher than 2. Tiles are colored by log transformed percentage.*

Based on the percentage of annotated residuals in each TAD calling tool and cell line, a consistent pattern of annotated residuals was observed across all three cancer types (Figure 47A). Out of all cancers, breast cancer had overall the lowest percentage of over-predicted TADs, ranging from 0.32% to 1.16%. In contrast, liver and skin melanoma exhibited slightly higher percentages of over-predicted residuals, from above 1% to around 2% in each TAD group. Conversely, breast cancer had higher percentages of under-predicted TADs across all TAD groups, with a maximum of around 4%, while liver and skin had around 2% to 3%.

To see if the annotated outliers were only specific for certain called TAD tissue and tool, I examined the overlaps between them. Based on the number of overlapping TAD residuals from 15 TAD groups for each individual annotated residual type and cancer, I detected varying patterns of overlap (Figure 47B). Under-predicted outliers showed a smaller percentage of overlap in all 15 groups. Breast cancer's under-predicted TADs mostly overlapped with 15 or 14 TAD groups, making up around 40% of all under-predicted TAD outliers across all TAD groups. Liver cancer had a slightly lower overlapping rate, with about 35% of under-predicted TADs overlapping with 15 or 14 TAD groups. Skin melanoma had the lowest overlap for under-predicted TADs among the analyzed cancers. Unlike non-outliers and under-predicted TAD outliers, over-predicted TAD outliers, which were generally less abundant, were mostly associated with only a few TAD groups or were unique to a single TAD group. This is evidenced by the very high percentage, around 14% and up to a maximum of 20%, for liver cancer.

4.3.2.1. Analysis of erroneous topologically-associated domains

Based on the COO model that demonstrated the highest percentage of explained variance across aggregated cancer profiles and achieved the highest accuracy in identifying COO of individual patients, I selected the following TADs for downstream analysis in each cancer type: HMEC IS 10kb for breast cancer, and NHEK GMAP 50kb for skin melanoma and liver cancer. Despite HMEC IS 10 kb having more called TADs than NHEK GMAP 50 kb setting, the same percentage of outliers (~4%) was detected in all cancer types. Breast cancer had a total of 220 outliers out of 4854 TADs; 28 over- and 192 under-predicted. Skin melanoma had 87 outliers out of 1876, comprising 22 over- and 60 under-predicted, while liver cancer had 82 out of 1876, consisting of 37 over- and 50 under-predicted (Figure 48A). Most enriched chromosomes with outliers where at least one outlier group was present in more than 5%, were breast chromosome 1, 8, 17 and 20; liver chromosome 5, 7, 8, and 15; and skin melanoma chromosome 1, 5, 7 and 9 (Figure 48B). Only on melanoma chromosomes 5 and 9 I found a more prominent enrichment of over-predicted outliers. In other chromosomes and cancers, under-predicted TADs outliers dominated. In agreement with the analysis conducted on 1 Mb regions, I found that only a small number of TADs were annotated as the same outlier type by models trained on different cancer types, suggesting a distinct trait of each outlier in its corresponding cancer type (Figure 48C). The highest overlap was found among non-outliers in all three cancer types.

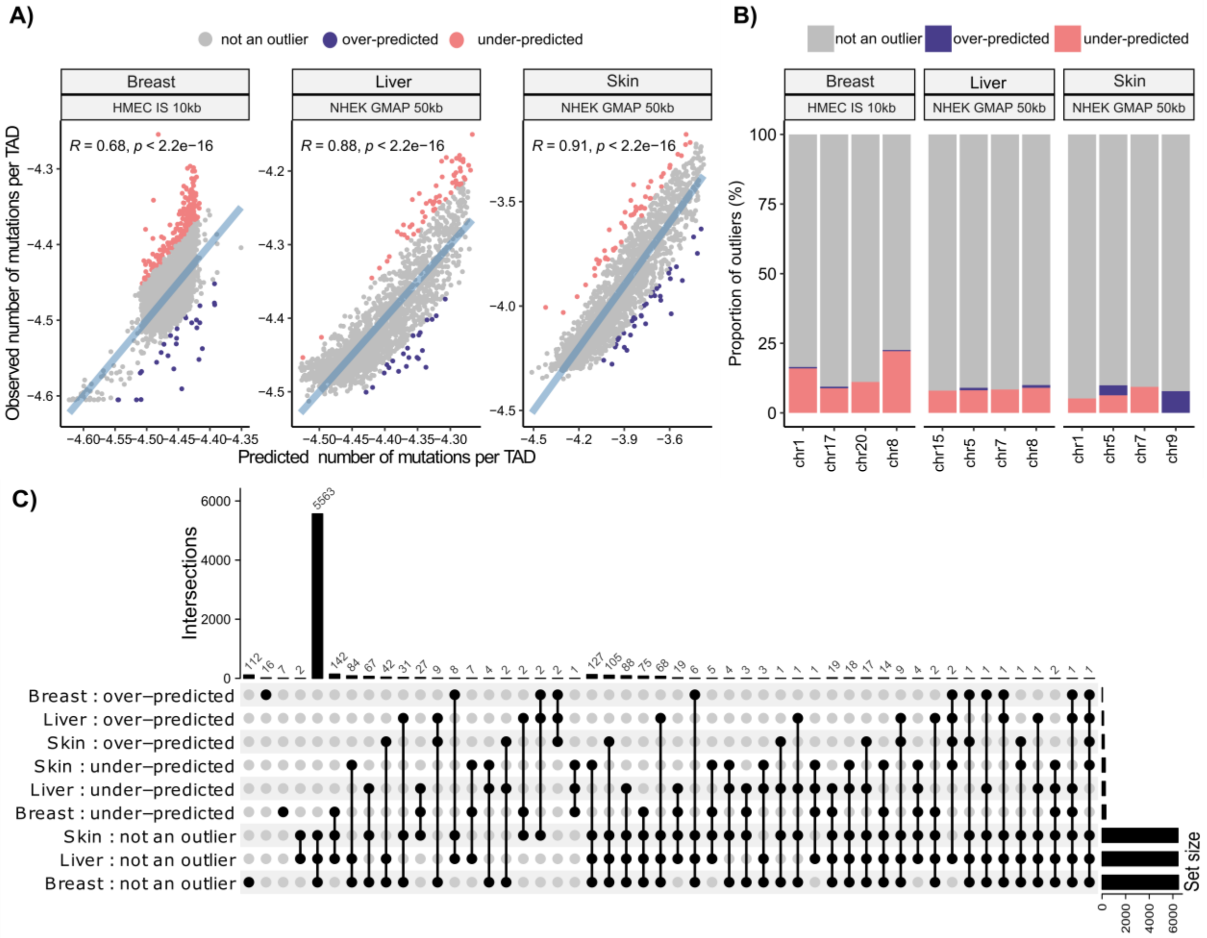


Figure 48. *A) Predicted vs observed mutations of aggregated tumor profiles of breast, liver and skin cancer from the best multiple linear regression models of WGS data in topologically associated domains (TADs). Colors denote different outlier classes. B) Proportion of annotated outliers based on their location in the genome (if more than 15% are outliers) in each cancer. C) UpsetR showing the overlaps of annotated TADs as over-, under-outliers or not an outlier between different cancer types.*

Afterwards, I analyzed the gene composition and regulatory features in those annotated outlier TADs. I found that CGC genes were more enriched in under-predicted outlier TADs in breast and skin melanoma cancers, whereas they were more prevalent in over-predicted regions of liver TADs (Figure 49A). Similar patterns were observed with TIME driver genes, with the exception that liver non-outlier TADs had a higher proportion of TIME genes compared to other outliers (Figure 49B). Only in breast cancer, the enrichment of TIME genes was significant, with a chi-square test p-value of $1.61*10^{-6}$. Tissue-specific super-enhancers showed higher enrichment in non-outliers TADs across all normal and cancerous tissues/cell lines in all three cancer types (Figure 49C). This was especially prominent in various liver SE tissues, where under-predicted TADs had the lowest proportion of SE-affected TADs. In skin melanoma, over-predicted TADs had the least proportion of SE-affected TADs across all cell lines.

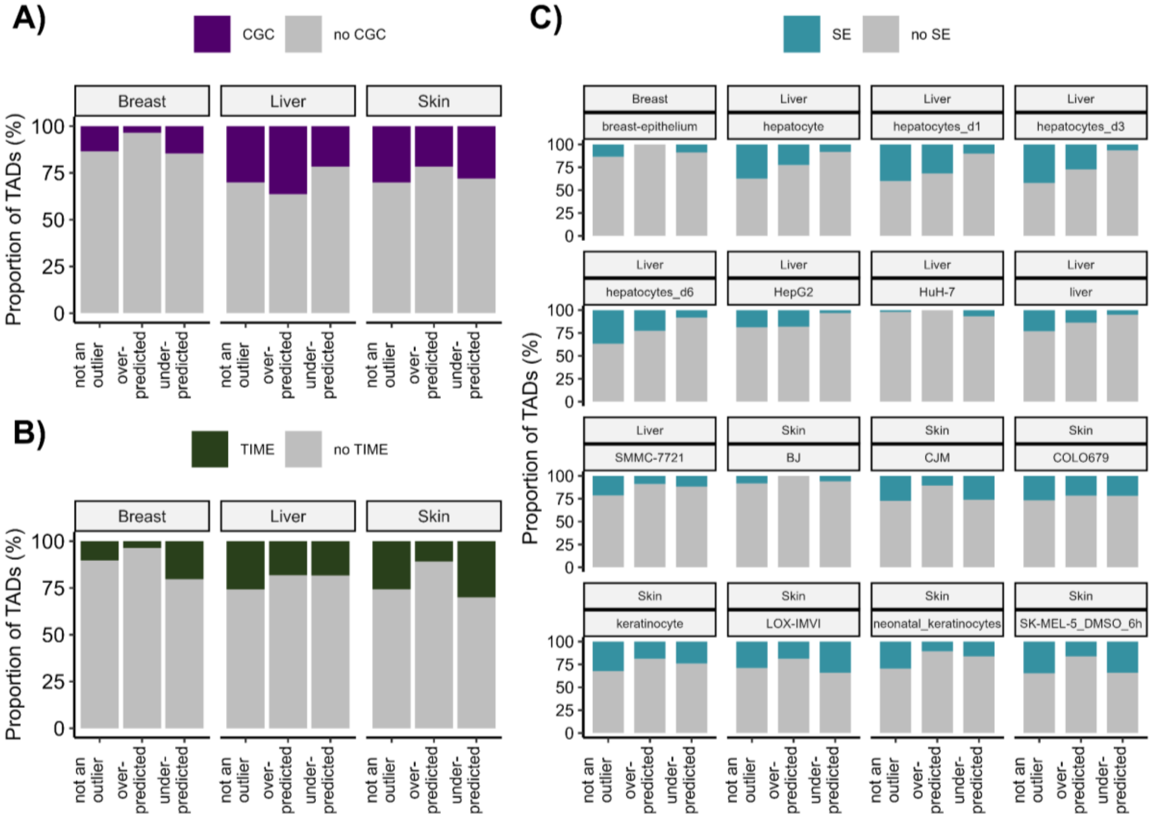


Figure 49. *Proportion of annotated topologically-associated domains (TADs) based on their erroneous status that are affected with A) Cancer Gene Consensus (CGC), B) tumor immune microenvironment (TIME) and C) tissue specific super-enhancers (SE) from both normal and tumorigenic cell-lines or tissues from the SEdb2.0 database*

Based on the over-presentation (ORA) analysis of GO terms, I found that the majority of terms were enriched in non-outlier TADs in all cancers (Figure 50A). In breast cancer, these pathways were associated with brain-related processes such as forebrain development, axonogenesis and actomyosin structure organization. In contrast, liver and skin melanoma cancers, on the other hand, were enriched in pathways related to post-transcriptional gene silencing, highlighting the role of RNA-mediated regulation. Additionally, under-predicted TADs in liver cancer showed enrichment for brain-related processes, as did over-predicted TADs in skin melanoma. Over-predicted TADs in the liver were also associated with the G protein-coupled purinergic nucleotide receptor signaling pathway.

The ORA of the Hallmark of Cancer pathways identified only one significant hallmark: IL6-JAK-STAT3 signaling, with two genes, *IL3RA* (ENSG00000185291) and *CSF2RA* (ENSG00000198223), found in over-predicted breast TADs (p-value = 0.014). Further ORA on databases of disease-gene associations (Jensen and DisGeNET) did not yield significant terms related to cancer development after adjusting for multiple hypothesis testing. However, examining the top 10 terms in the Jensen database for under-predicted TADs revealed associations with “Liver cancer” in liver TADs (27/597 genes, p-value = 0.21), “Skin cancer” in skin melanoma TADs (25/454 genes, p-value = 0.70), and “Breast cancer” in breast TADs (2/8 genes, p-value = 0.86). The two genes identified in under-predicted TADs for breast cancer were *MUC1* (ENSG00000185499) and *ERBB2* (ENSG00000141736).

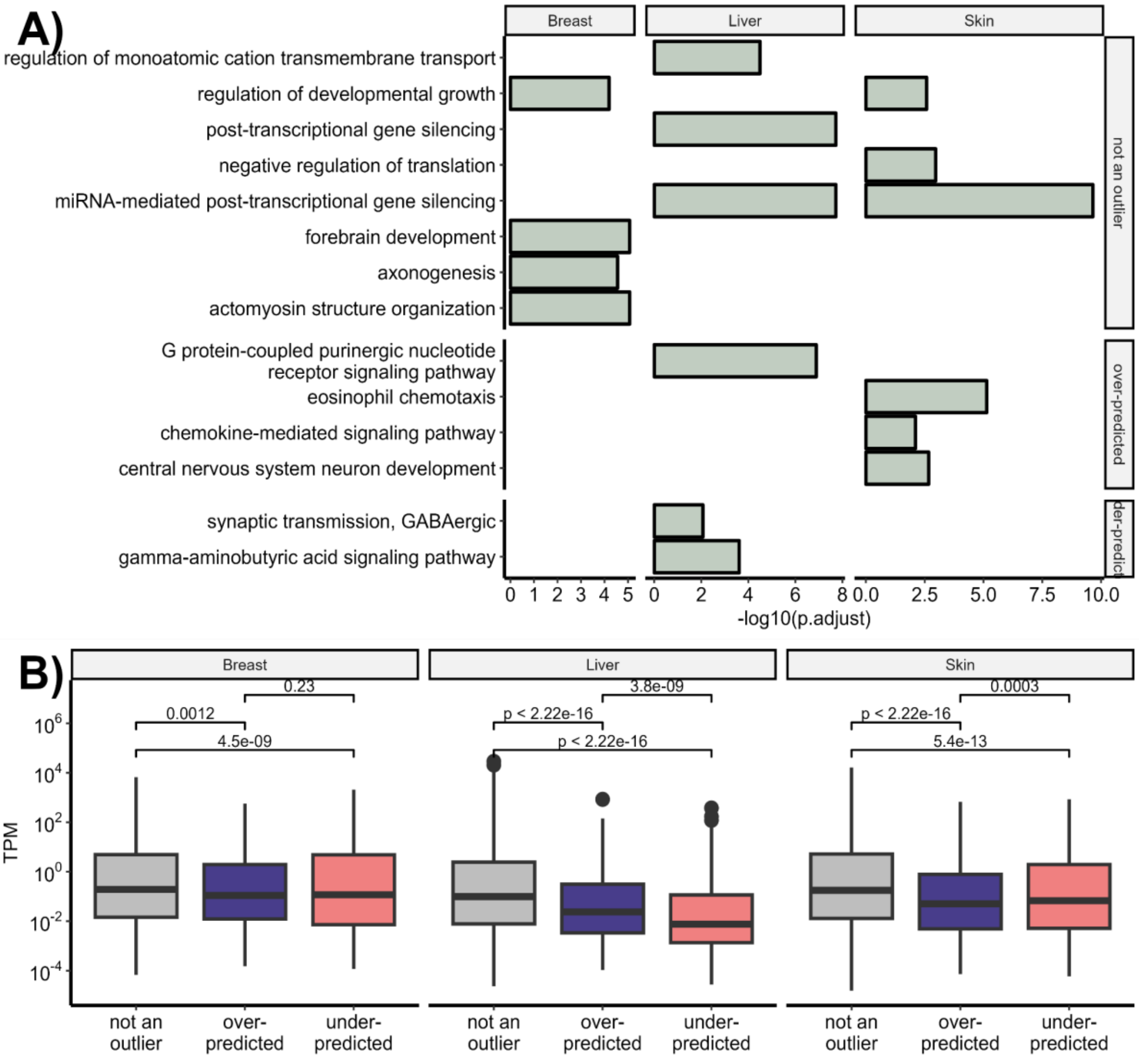


Figure 50. *A) Over-representation analysis of GO terms involved in biological processes. Top 3 most significant by filtering Benjamini-Hochberg p-value of 0.05 are shown for each category of annotated TADs based on their standard residual (erroneous prediction rate) B) TPM (transcripts per million) expression of genes in normal tissues and C) expression of tissue-specific genes from normal tissues of corresponding cancer type (breast, liver and skin cancer) from GTEX TPM 30 database separated by residual annotation. (Wilxcoxon test)*

When examining gene expression in normal tissues corresponding to the COO of each cancer type, outlier TADs exhibited overall lower gene expression compared to non-outliers (Figure 50B). In liver cancer, under-predicted TADs have significantly lower expression levels than over-predicted TADs ($p < 2.22*10^{-16}$) and non-outlier TADs ($p < 2.22*10^{-16}$). In contrast, for skin melanoma, over-predicted TADs have significantly lower gene expression levels compared to under-predicted TADs ($p = 0.0003$) and non-outlier TADs ($p < 2.22*10^{-16}$).

Furthermore, I classified the annotated TADs outliers by their chromatin states (Figure 51A). Chi-square test showed that there was a significant difference in chromatin states of annotated TADs outliers, p-value of $1.3*10^{-33}$, $1.69*10^{-26}$ and $7.77*10^{-7}$ for breast, liver and skin melanoma respectively. In breast and liver cancers, non-outlier TADs showed a higher proportion of repressed chromatin compared to over-predicted and under-predicted TADs. Also, in these cancers there was a higher proportion of over-predicted TADs with heterochromatin state. In skin melanoma, non-outlier TADs display a mix of low-active and repressed chromatin states similar to other two cancers, but under-predicted TADs are more characterized with a repressive state.

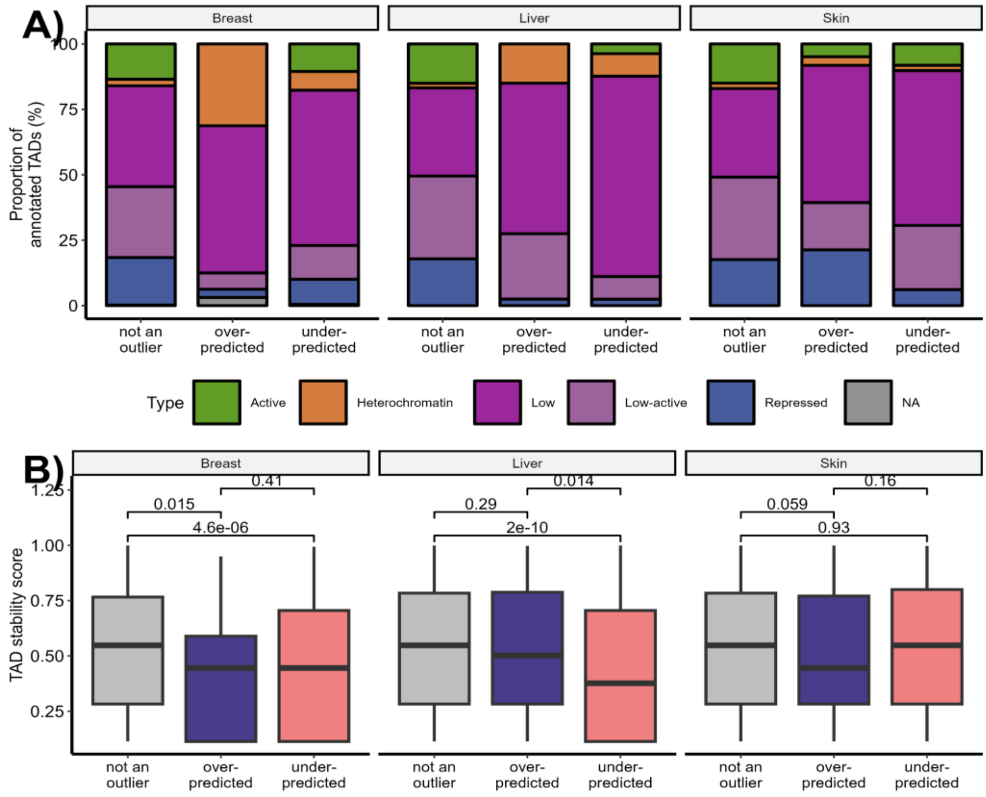


Figure 51. *A) TADs annotation by Akdemir et al. 2020. Based on active and inactive state (heterochromatin, low, low-active and repressed) B) TADs stability score of TADs boundaries distribution in annotated erroneous regions (Wilcoxon test).*

In breast cancer, the stability scores for over-predicted TADs were significantly lower compared to non-outlier TADs (p-value=0.015), and under-predicted TADs (p-value=4.6*10^{-6}) (Figure 51B). Liver cancer followed a similar trend, with significantly lower stability scores for under-predicted TADs than non-outliers (p-value=2*10^{-10}). Livers over-predicted TADs also showed reduced stability without statistical significance (p-value=0.29). Moreover, livers under-predicted TADs showed statistically significant lower stability than over-predicted ones (p-value=0.014). In skin melanoma, while the differences in stability scores among TAD categories were not statistically significant, there was a trend of lower stability in over-predicted compared to under-predicted (p-value=0.16) and non-outlier TADs (p-value=0.059).

Under-predicted TADs were significantly enriched with SV-hotspots than non-outliers and over-predicted TADs (Figure 52A). This enrichment was evident across all three cancer types, with p-values of 6.22*10^{-110} for breast cancer, 3.85*10^{-07} for liver cancer, and 0.0116 for skin melanoma, as determined by the Chi-squared test. Over-predicted TADs in breast cancer show a similar level of SV-hotspot enrichment to non-outliers, while they are less abundant in liver and skin melanoma. Breast *de novo* SV signature 4 hotspots were more abundant in both under- and over-predicted TAD outliers (Figure 52B). Liver *de novo* SV signature 3 hotspots were exclusively found in the liver, as this is the only SV-hotspot signature I detected for this cancer. For skin melanoma, under-predicted TADs commonly exhibit both *de novo* SV signature 3 and 4 hotspots, whereas over-predicted TADs are associated only with skin *de novo* SV signature 3.

Similar patterns were observed with kataegis regions, where under-predicted TADs in all cancer types demonstrate significant enrichment compared to other annotated TADs groups (Figure 52C). This enrichment is particularly pronounced in breast cancer (p-value 4.68*10^{-8}) and skin melanoma (p-value 3.42*10^{-14}), with liver cancer showing a p-value of 0.098 determined by Chi-squared test. Over-predicted outliers across all cancer types had a significantly lower percentage of them affected by kataegis. It appears that the severity of kataegis regions, determined by a higher number of mutations detected in kataegis regions, was positively correlated with the error rate of prediction on regions, absolute standard residual value (Figure 52D). This is apparently more profound in breast and skin melanoma cancers where the correlations are more positive in kategis defined regions compared to non-kataegis ones. The only stand out is liver cancer where the correlation was even lower in kataegis regions and not significant in both kataegis and non-kataegis regions unlike the other two cancer types.

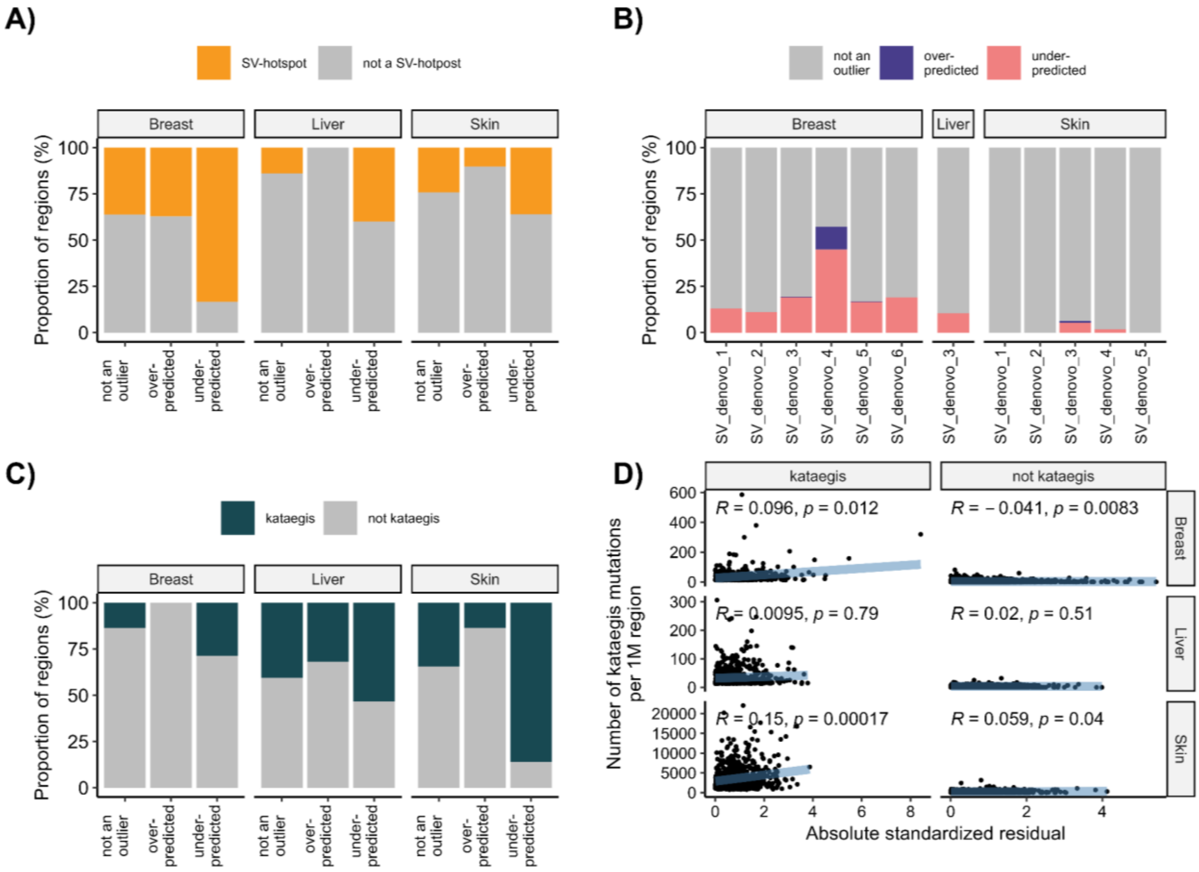


Figure 52. *A) Proportion of topologically associated domains (TADs) affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier B) Proportion of residual annotated TADs based on their SV-hotspot mutational signature C) Proportion of TADs affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier D) Spearman correlation between mutations affected by kataegis and erroneous prediction rate, as well as between non-kataegis mutations and erroneous prediction rate*

The results from Fisher's exact test or Pearson's Chi-square test indicate that there was no statistically significant difference in the presence of SV-hotspot and kataegis affected TADs across any defined outlier and non-outlier groups, with the exception of non-outliers in breast cancer, where the p-value was 1.74*10^{-17} (Figure 53).

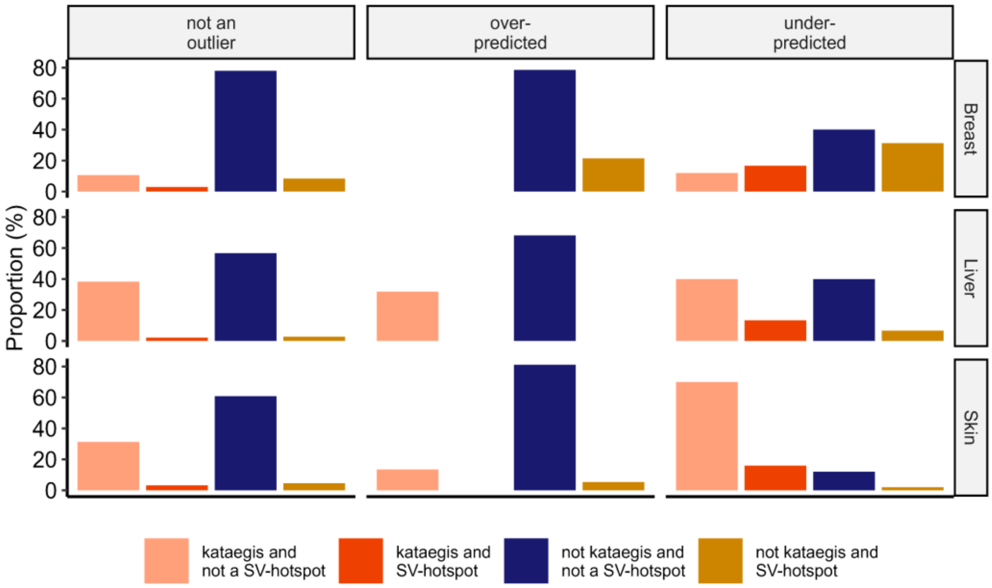


Figure 53. *Proportion of SV-hotspots and kataegis affected TADs of over-predicted, under-predicted and non-outlier regions across breast, liver and skin cancer. Each tested group is colored by Benjamini-Hoechberg corrected p-value of Fisher exact test or Pearson's Chi-square test results*

To conclude whether certain SBS mutational signatures contribute to higher prediction error rate in TADs COO model, I calculated Pearson's correlation of summarized SBS signatures in TADs and absolute standard residual (Table 10). After correcting for multiple hypotheses testing and filtering for significant correlations, I selected the top five most significant correlations and compared them to the correlations obtained using all mutations. In breast cancer, four out of the five shown SBS signatures had higher correlations than all mutations: SBS40, SBS8, SBS5, and SBS3. SBS40 and SBS8 were the most abundant, with proportions of 14.16% and 11.58%, respectively. For liver cancer, three out of the five SBS signatures had higher correlations than all mutations: SBS8, SBS40, and SBS12. All three SBS signatures were present in more than 10% of mutations, with SBS40 being the most frequent at 20.80%. In contrast, in skin melanoma, one of the most abundant signatures, SBS7a (81.06% of mutations), had a slightly higher positive correlation (~0.169) than all mutations (~0.162), while SBS7b had lower positive correlations with prediction error than all mutations.

Table 10. *Top 5 strongest Pearson's correlations of mutational signature present in more than 5% or all mutations and absolute standard residual per topologically-associated domain for breast and liver cancer alongside the correlation of all mutations with Benjamini-Hochberg corrected p-values lower than 0.05. Table also contains proportion of mutations belonging to each signature in aggregated tumor profiles*

Cancer type	SBS mutational signature	Pearson's correlation	p-value	Proportion of mutations (%)
Breast	All mutations	0.06163406	$5.43*10^{-05}$	100
	SBS40	0.09273914	$6.09*10^{-10}$	14.16
	SBS8	0.08685779	$7.88*10^{-09}$	11.58
	SBS5	0.07849807	$2.04*10^{-07}$	6.57
	SBS3	0.06441259	$2.45*10^{-05}$	8.11
	SBS1	0.0563627	$2.48*10^{-04}$	5.70
Liver	All mutations	0.23230902	$5.18*10^{-23}$	100
	SBS8	0.30046776	$2.00*10^{-38}$	10.99
	SBS40	0.25393662	$2.68*10^{-27}$	20.80
	SBS12	0.24209204	$7.44*10^{-25}$	20.49
	SBS92	0.22092379	$4.86*10^{-21}$	7.87
	SBS22	0.19700881	$7.14*10^{-17}$	5.44
Skin	All mutations	0.16185076	$1.14*10^{-11}$	100
	SBS7a	0.16944069	$1.01*10^{-12}$	81.06
	SBS7b	0.11643779	$1.63*10^{-06}$	13.67

Given that TADs with higher prediction error rates, as defined by standard residuals, were linked to SV-hotspots, kataegis regions, and specific mutational signatures, I calculated the correlation between mutations and histone modifications from normal tissues corresponding to the cell-of-origin for each cancer type. These calculations were grouped according to the influence of specific genomic features. Regardless of the specific groups of affected TADs, all mutations and closed chromatin modifications (H3K27me3 and H3K9me3) exhibited the strongest positive correlation for breast cancer (Figure 54A). On the other hand, COO open chromatin modifications had the strongest negative correlation in liver cancer regardless of the annotated TAD group. Skin melanoma showed the weakest correlations with closed chromatin histone modifications, with the highest positive correlations being 0.291 and 0.429 in under-predicted TADs. Over-predicted TADs generally had the worst correlation compared to other annotated outlier TAD groups across all cancer types. The most noticeable difference between TADs affected by SV-hotspots and/or kataegis was the increase of negative correlation with open chromatin marks compared to the regions not affected by those genomic features. This was particularly evident in breast cancer.

Analysis of the correlation of SBS mutational signature and normal epigenomes of cell-of-origin in TADs showed that the age-related SBS1 signature has a weaker correlation than expected in all cancers (Figure 54B). This was most apparent in liver and skin cancer where the SBS1 signature was present in smaller amounts compared to other analyzed signatures. In breast cancer, mutational signatures SBS8, SBS40, and SBS5 had the strongest correlations with both open and closed chromatin modifications. Although APOBEC-related mutational signatures SBS13 and SBS2 were quite abundant, they had weaker correlations with COO epigenome. In liver cancer, SBS24 and SBS23 signatures had the weakest negative correlations with open chromatin marks but very high positive correlations with H3K27me3. Interestingly, the abundant SBS12 signature in liver did not show a correlation with the same modification, H3K27me3. As for skin melanoma, SBS7b signature, which was less abundant than SBS7a, had stronger and statistically significant positive correlations with open chromatin.

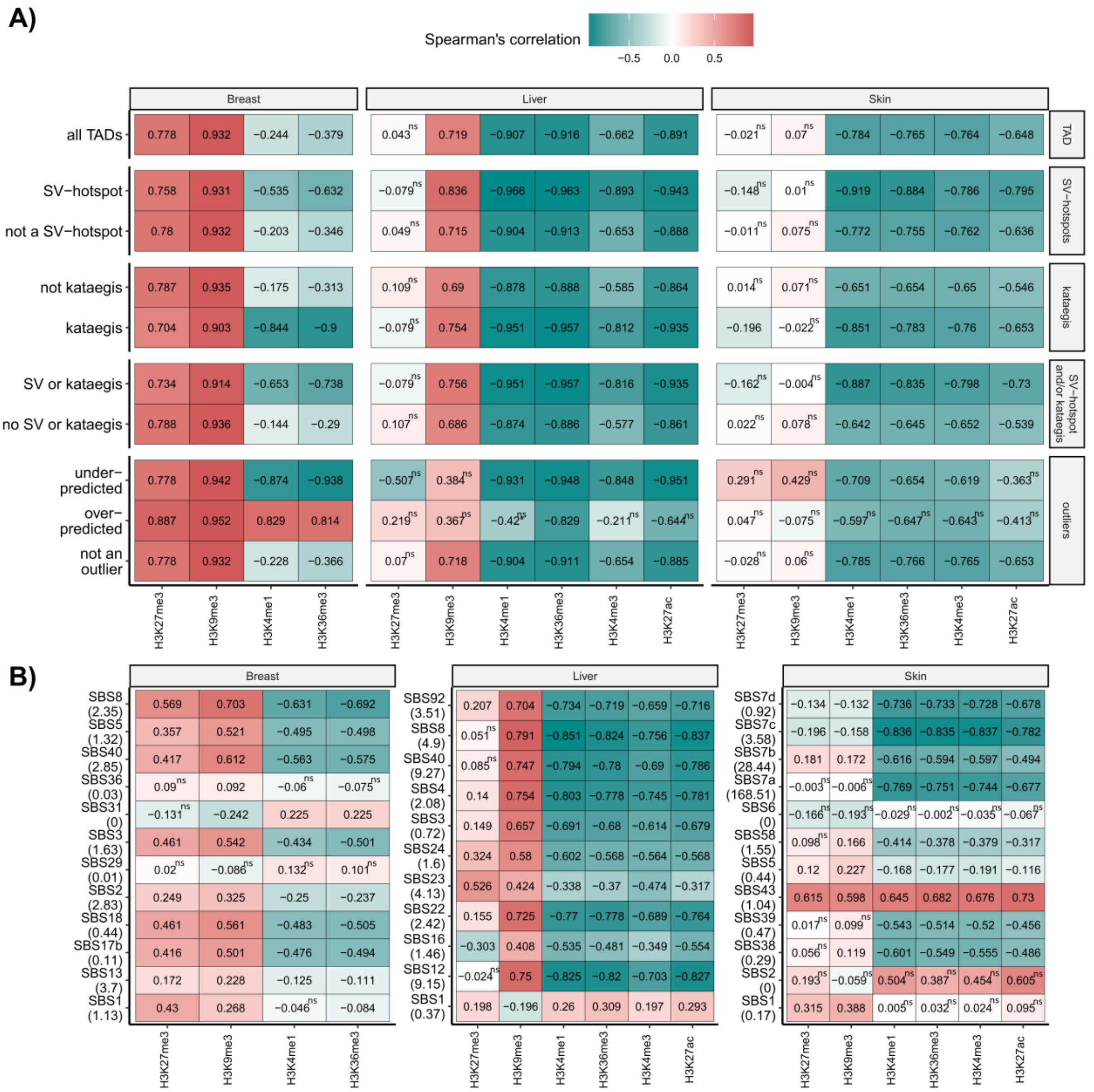


Figure 54. *A) Spearman's correlation of all mutations and histone modifications of cell-of-origin tissue on TADs annotated based on different criteria (containing or not SV-hotspots or kataegis regions or both, based on outlier annotation) B) Spearman's correlation of SBS mutational signatures and histone modifications of cell-of-origin tissue with number absolute counts divided by 10^6 written in brackets of signatures in used TADs. Non-significant correlations, Benjamini-Hoechberg corrected p-value above 10^{-5} are labeled with ^{ns}.*

4.3.2.2. Improving the TADs COO model

To enhance the accuracy of the TADs model, I excluded SV-hotspots, kataegis regions, and specific mutations associated with certain mutational signatures that were linked to higher error rates. In the aggregated breast cancer model, this adjustment led to an improvement in overall explained variance when kataegis- and SV-hotspot affected TADs were removed, with a Wilcoxon test p-value of 0.004 compared to the original model that included all TADs (Figure 55A). Removal of SBS mutational signatures did not results in better model accuracy. However, the proportion of correctly identified COO patients in breast cancer remained similar across all COO TADs model configurations (Figure 55B). For liver cancer, none of the TADs COO models setups resulted in significant improvement in either the explained variance of aggregated profiles or the accuracy of COO identification for individual patients. In contrast, the skin melanoma COO model experienced a notable decline in performance when kataegis-affected TADs were excluded from the models (Wilcoxon test, p-value=0.001).

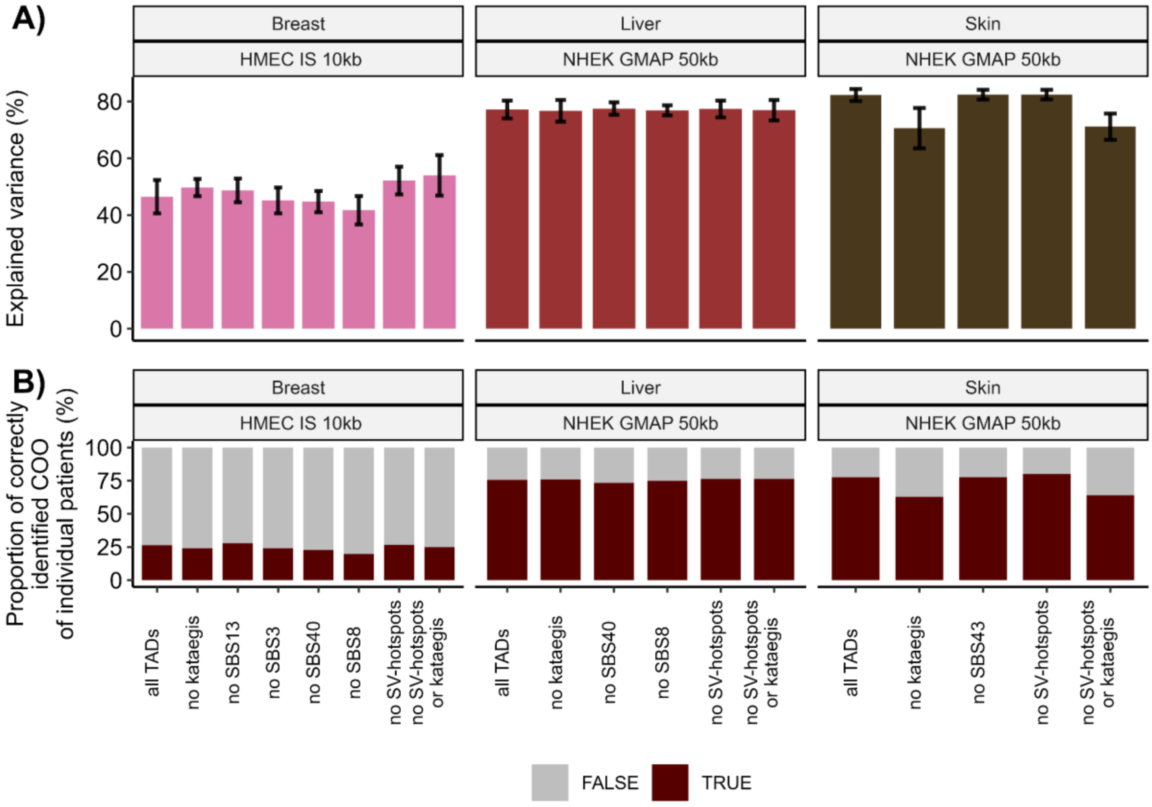


Figure 55. *A) Multiple linear regression models for the top prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WGS based on 1 Mb regions that were either removed due to being affected by SV-hotspots and/or kataegis or excluded certain mutations due to their SBS signature origin. The SBS mutations which were removed are SBS13, SBS3, SBS8 and SBS40 for breast, SBS40 and SBS8 for liver and SBS43 for skin cancer. The figure shows the overall explained variance, depicted as the mean with standard deviation, derived from a 10-fold cross-validation analysis. B) Proportion of correctly and incorrectly identified COO for individual patients based on results of corresponding COO models in A).*

Only the random forest modifications of the TADs COO model using all TADs and aggregated mutational cancer profiles resulted in a higher explained variance with correctly identified COOs (Figure 56A). Statistically significant improvements in correctly identified COOs on aggregated profiles were observed when comparing multiple linear regression to the random forest model for breast (p-value=0.001) and liver (p-value=3*10^{-07}) cancers. However, based on individual patient COO predictions, the random forest model performed similarly or even worse, as evidenced by the breast cancer results (Figure 56B). Unfortunately, the extreme gradient boost TADs COO model yielded lower explained variances and incorrectly predicted the cell-of-origin for breast cancer as keratinocyte. The poor performance of the current extreme gradient boost TADs COO model was also evident in its inability to run successfully for individual patients due to low variance in model splits.

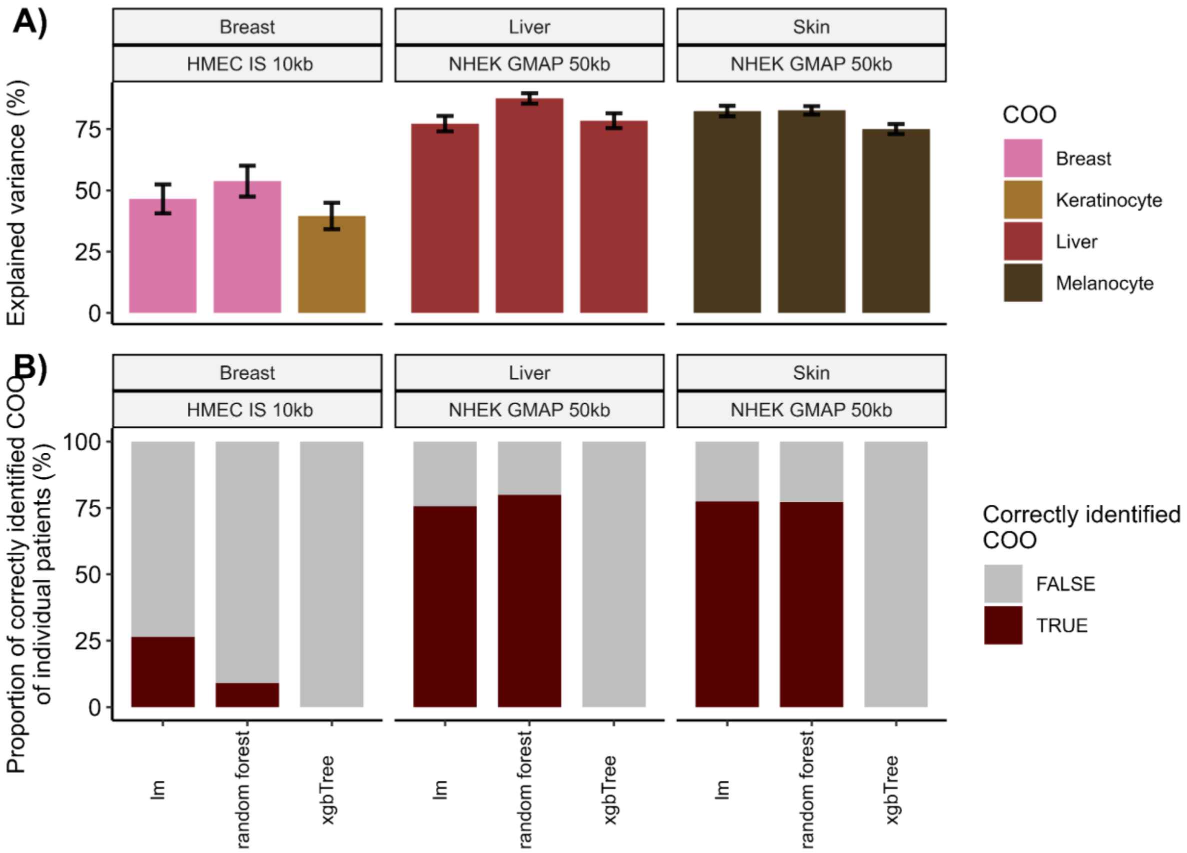


Figure 56. *A) COO models (multiple linear regression, random forest and extreme gradient boosting) for the top prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WGS based on TADs. The figure shows the overall explained variance, depicted as the mean with standard deviation, derived from a 10-fold cross-validation analysis. B) Proportion of correctly and incorrectly identified COO for individual patients based on results of corresponding COO models in A).*

4.3.2.3. Individual patient TADs COO model

Incorrectly assigned COOs for individual patients included brain-related tissues, right atrium, keratinocytes, gastrointestinal tissues, esophagus and immune cells (Figure 57). The models show a high degree of consistency in predicting the same COO tissues across all methods, particularly for liver and skin cancers.

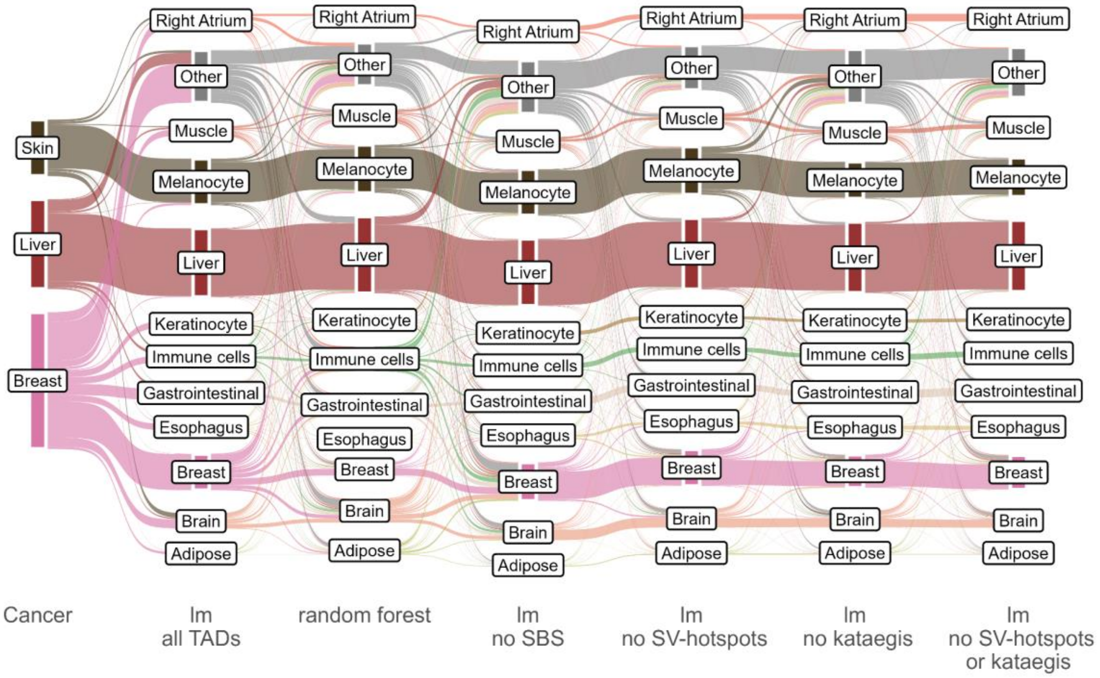


Figure 57. *Alluvial plot illustrating the predicted cell-of-origin (COO) of individual patients across various COO model setups in both aggregated cancer data and individual patient mutational profiles. Assigned COO tissues that represent less than 4% in each cancer type are labeled as "Other". Single-base substitutions (SBS) which were removed are SBS13 for breast cancer, SBS40 for liver cancer and SBS43 for skin melanoma patients.*

The majority of correctly identified COO patients across all cancer types exhibited significantly higher explained variance and a greater number of mutations per patient in TADs (Figure 58). However, only the correctly identified COO patients with skin melanoma demonstrated a higher number of kataegis in TADs per patient. In contrast, breast cancer patients with incorrectly identified COOs showed a higher number of both kataegis and SV-hotspots in TADs. For liver cancer, while the number of kataegis was higher in incorrectly identified COOs, the number of SV-hotspots did not differ significantly between the two groups of patients.

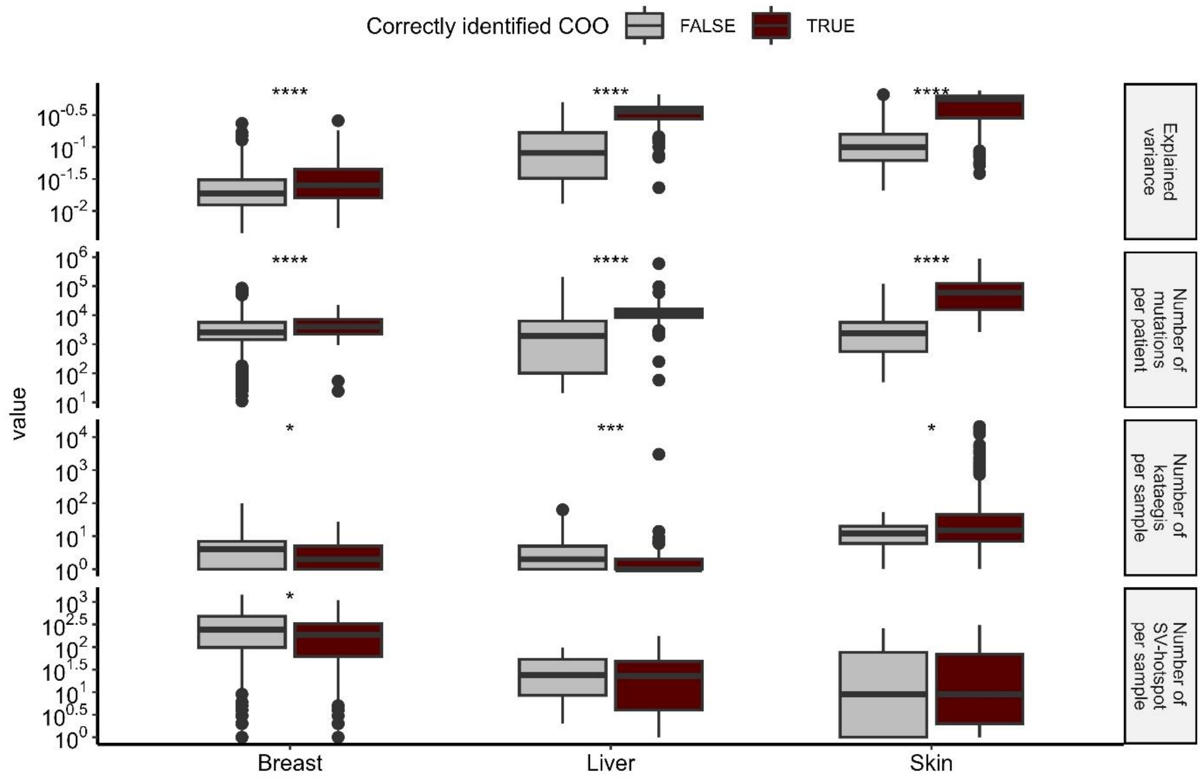


Figure 58. *The distribution of explained variance, the number of mutations, kataegis and SV-hotspots separated by correct or incorrect COO identification of individual patients with TADs. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Regarding the differences in the proportion of SBS mutational signatures in TADs between correctly and incorrectly identified COO patients using the TADs COO model, I found that SBS1 tended to be higher in incorrectly identified COO patients, with a significant difference detected in liver cancer (Figure 59). In breast cancer, SBS13 and SBS3 were significantly enriched in incorrectly identified COO patients, while SBS3 was enriched in correctly identified COO patients. Although not statistically significant, SBS2 and SBS40 tended to be higher in incorrectly identified COO breast cancer patients. In liver cancer, most signatures tended to have higher abundance in incorrectly identified patients, except for SBS12, which was significantly higher in correctly identified COO patients. In skin melanoma, the most noticeable differences were the statistically significant differences in SBS7a and SBS7b, which showed opposite enrichments between the two groups of patients. SBS7a was more abundant in correctly identified COO patients, while SBS7b was more abundant in incorrectly identified ones, in agreement with 1 Mb models. Other more abundant signatures in skin melanoma, such as SBS3, SBS38, SBS40, and SBS8, showed higher proportions in correctly identified COO patients.

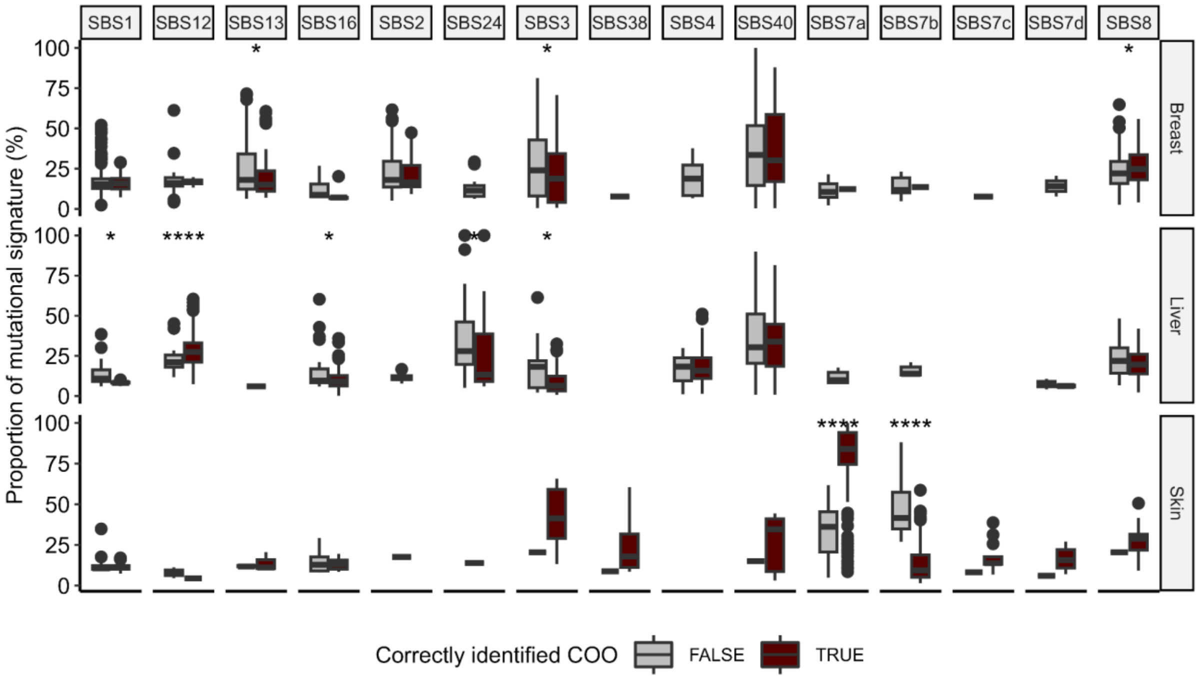


*__Figure 59.__ Proportion of SBS mutational signatures per patient separated by correct or incorrect COO identification of individual patients with TADs. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.2.3.1 Breast cancer

The accuracy of correctly identifying the COO using TADs COO models resulted in a low proportion of correctly identified COO in all histological subtypes of breast cancer (Figure 60). Adenoid cystic carcinoma, with only one patient, achieved the highest accuracy, with 100% of patients correctly identified but with only 1 sample. In contrast, several subtypes, including tubular and invasive cribriform carcinoma, pleomorphic carcinoma, neuroendocrine carcinoma, and medullary carcinoma, had no patients with correctly identified COO. Among the more common subtypes, mucinous adenocarcinoma, infiltrating duct carcinoma, lobular carcinoma (NOS), and duct micropapillary carcinoma showed varying accuracy rates of 35%, 27%, 26%, and 33%, respectively. For correctly identified COO patients, the number of mutations per patient and the explained variance were higher across several subtypes. Specifically, in infiltrating duct carcinoma, the difference in mutational count and explained variance between correctly and incorrectly identified patients was statistically significant.

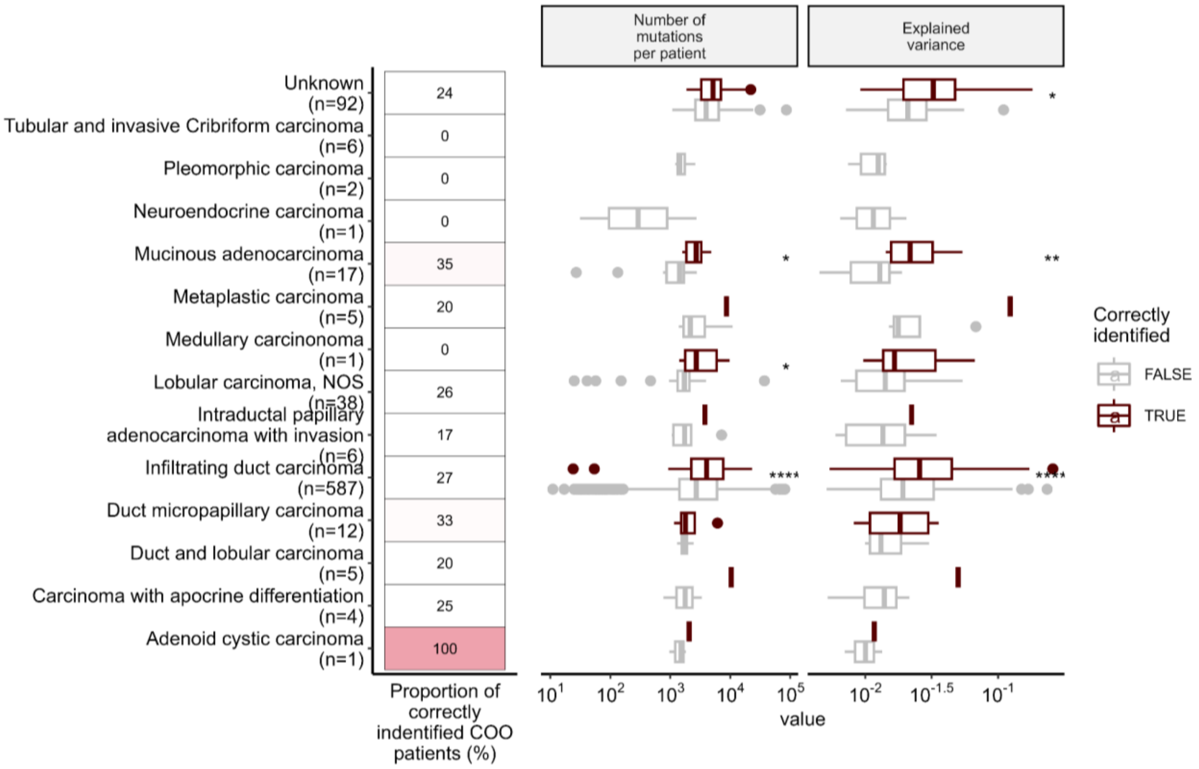


Figure 60. *A) Proportion of correctly identified cell-of-origin (COO) proportions using the topologically associated domains (TADs) separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Despite a significant difference in the number of mutations per patient among subtypes (Kruskal-Wallis test, p = 0.00084), only one patient from the Her2 subtype had its COO correctly identified using TADs COO models (Figure 61A). The majority of PAM50 subtypes were classified as infiltrating duct carcinoma. Based on their HRD status, HRD patients showed a higher proportion of correctly identified COO patients (41%) compared to non-HRD patients (30%) (Figure 61B). Additionally, the number of mutations per patient was significantly higher in HRD patients, who were predominantly of the infiltrating duct carcinoma subtype.

Regardless of the HRD status, correctly identified COO patients exhibited a higher total number of mutations per patient in the used TADs compared to incorrectly identified ones (Figure 61C). HRD breast cancer patients had overall higher values of both mutation count and explained variance in the COO model compared to non-HRD patients.

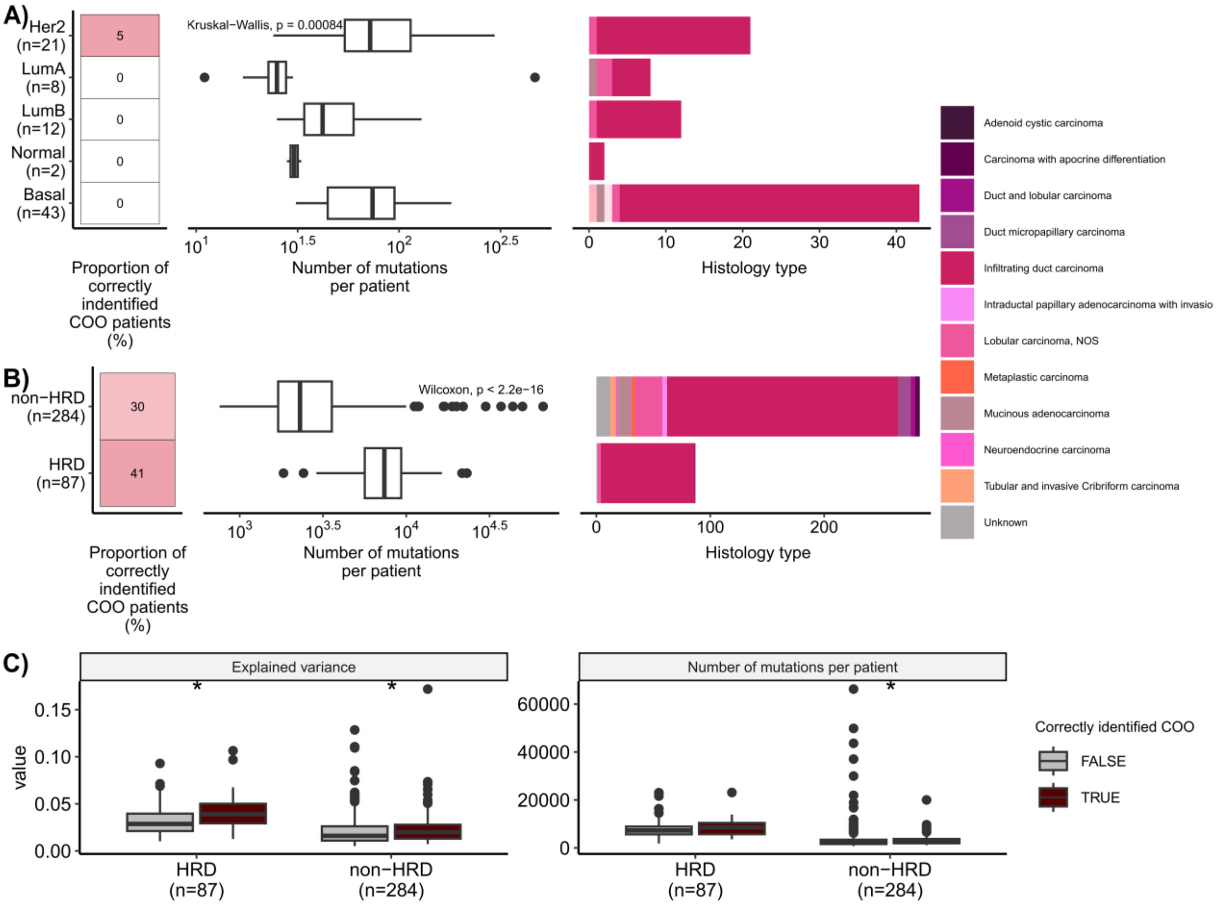


Figure 61. *A) Proportion of correctly identified cell-of-origin (COO) proportions using TADs, separated by their PAM50 subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. B) Proportion of correctly identified cell-of-origin (COO) proportions using the top 40% mutated protein-coding genes, separated by their PAM50 subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.2.3.2 Liver cancer

A high proportion of individual patients had its COO correctly identified with TADs COO model in the majority of liver cancer subtypes (Figure 62). In hepatocellular carcinoma, which is the most common subtype with 467 patients, 78% of patients were correctly identified. Hepatocellular adenoma and fibrolamellar hepatocellular carcinoma, each with 5 patients, had the highest accuracy, with 100% and 80% of patients correctly identified, respectively. Combined hepatocellular and cholangiocarcinoma subtypes, with 8 patients, showed a 75% accuracy, while cholangiocarcinoma, with 24 patients, had the lowest accuracy of 25% in COO identification. Correctly identified COO patients consistently had a significantly higher number of mutations per patient across almost all subtypes, with the most pronounced difference observed in hepatocellular carcinoma (p < 0.0001). Additionally, correctly identified patients also had a higher explained variance in the models compared to incorrectly identified COO patients in all subtypes.

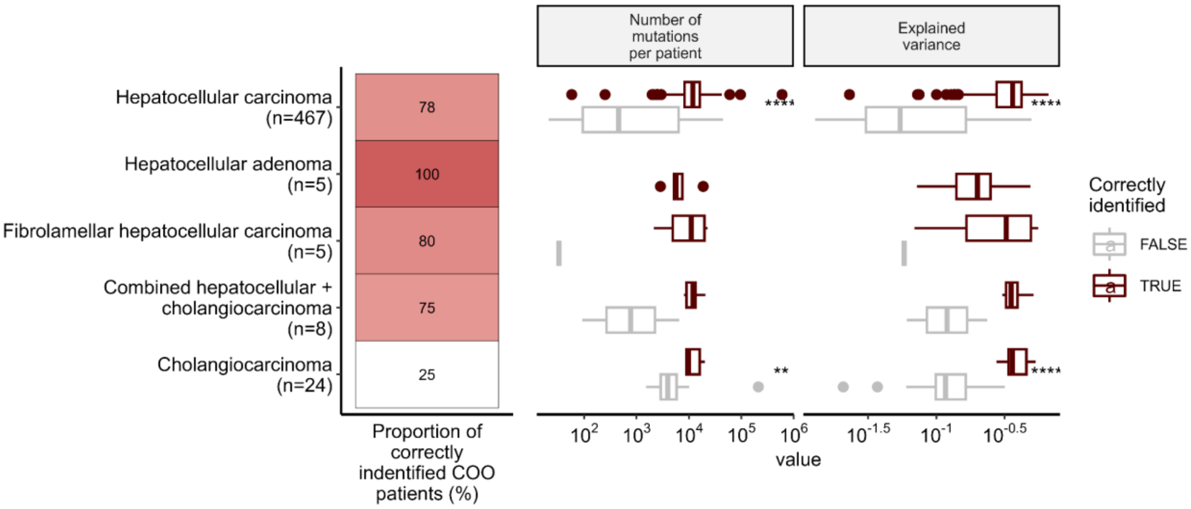


Figure 62. *A) Proportion of correctly identified cell-of-origin (COO) proportions using the top 40% mutated protein-coding genes, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.2.3.3 Skin cancer

The analysis of the histological subtypes of skin melanoma showed varying proportions of correctly identified COO patients across the different subtypes (Figure 63). The mucosal lentiginous melanoma and malignant melanoma (NOS) subtypes had the lowest proportion of correctly identified COO patients, at 39% and 31%, respectively. In contrast, the lentigo malignant melanoma with two and unknown melanoma subtypes with 10 patients achieved the highest accuracy, with 100% and 90% of patients correctly identified, respectively. Superficial spreading melanoma and nodular melanoma also displayed high accuracy, with 96% and 96% of patients correctly identified. Patients with correctly identified COO consistently exhibited a higher number of mutations per patient and higher explained variance compared to incorrectly identified COO patients across all subtypes. This observed difference was most notable and statistically significant in malignant melanoma (NOS) patients.

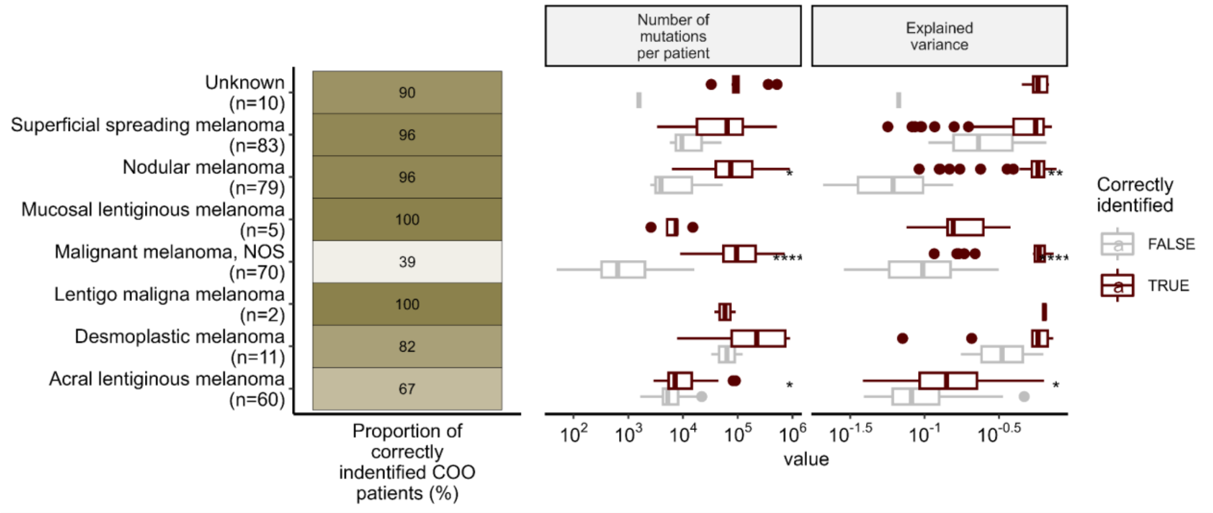


Figure 63. *A) Proportion of correctly identified cell-of-origin (COO) proportions using the top 20% mutated protein-coding genes, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.3 Gene-based cell-of-origin predictive models

In the development of the gene-based COO model I made numerous preliminary multiple linear models on aggregated gene mutational profiles of breast, liver and skin cancer in order to select the best model setting based on various COO model setups.

4.3.3.1. Lack of intron enrichment affects gene-based COO models

In previous analyses, I detected a significant difference between WGS and WXS mutational profiles, as well as a difference in performance of COO models based on 1 Mb genomic regions and TADs. These differences were especially prominent in TCGA cohorts. For this reason, I first calculated the proportion of mutations in introns and exons to see if the performance differences were caused by varying enrichment between the datasets. As it is clearly shown on Figure 64, most of the WGS datasets were more enriched in introns, except the TCGA cohorts (BRCA-US, LIHC-US, SKCM-US), while WXS datasets were more enriched in exons, with the exception of the BRCA-UK cohort.

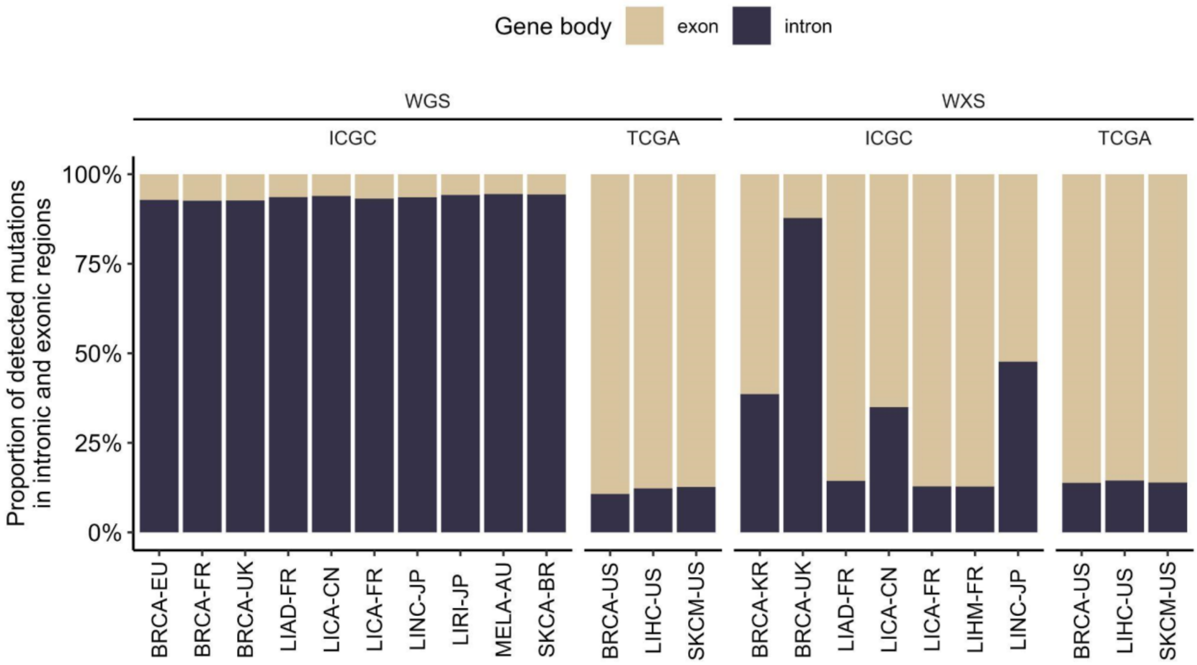


Figure 64. *Proportion of detected mutations in intronic and exonic regions in all genes (human genome version hg19) from WGS and WXS data of breast, liver and skin cancer individual cohorts. (Chi-square test, p-value < $2.2*10^{-16}$ for both WGS and WXS datasets)*

As there was a significant difference in intron enrichment between sequencing technologies and TCGA cohorts, I evaluated the performance of the top COO model of various settings where I used different gene sets and different normalizations (summed exonic mutations normalized by gene length, *Gene exon* model; summed exonic and intronic mutations normalized by gene length, *Gene exon+intron* model; and summed exonic mutations normalized by coding-sequence length, *CDS* model). For the WXS data, COO models did not result in correct cell-of-origin prediction in any of the settings for breast and liver cancers (Figure 65), for which the majority of models predicted brain-related tissues as the COO. Only the COO of skin melanoma was correctly predicted in all selected gene groups with the CDS model setting, with the exception of TIME genes. The highest explained variance (35%) was obtained with the use of protein coding genes.

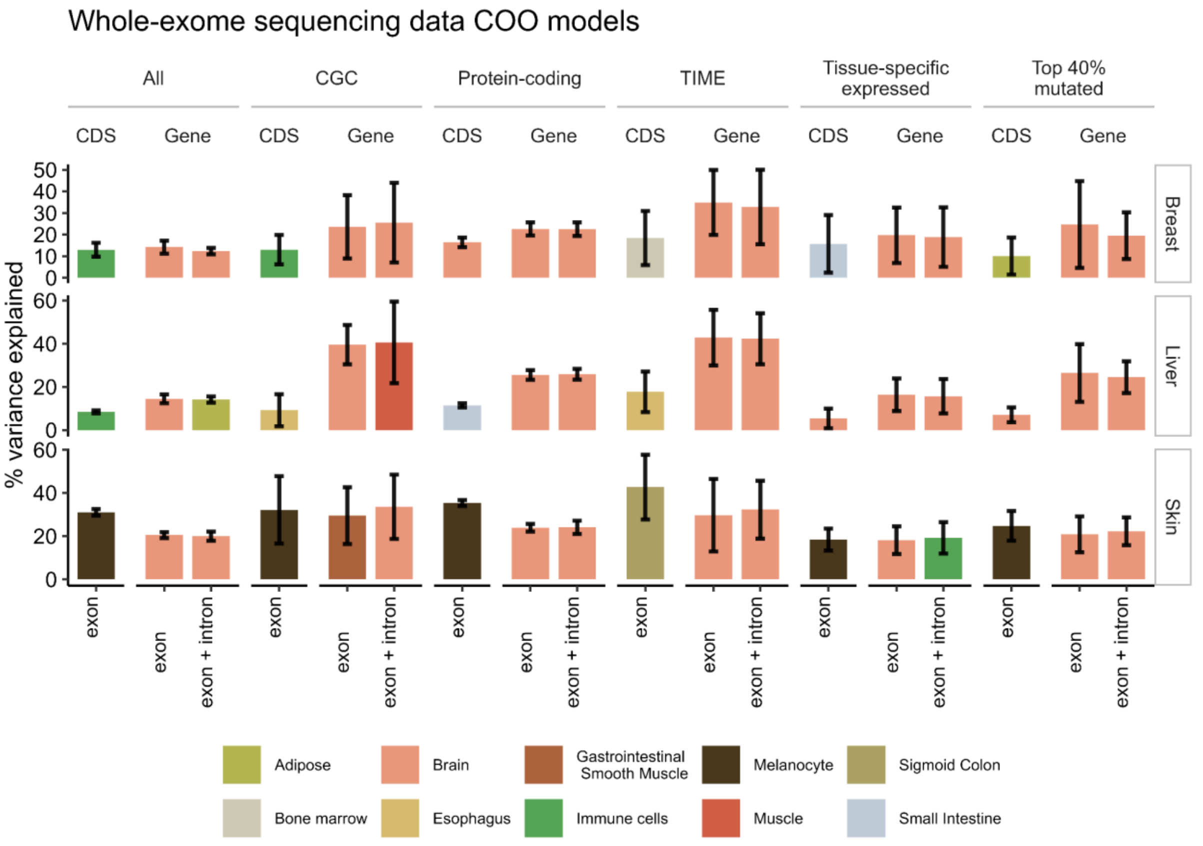


Figure 65. *The top performing multiple linear regression models for the prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WXS in various gene settings. Coding-sequence (CDS) settings represent the count of mutations per exon normalized length by total CDS length. Gene setting implies summing mutations per exon or exon plus intron normalized by total gene length. The overall explained variance is reported across the 10-fold cross-validation.*

Given that the WXS-based COO model failed to accurately predict the COO in two cancer types, I conducted an additional assessment of the significance of intronic mutations in the COO models. I ran WGS-based *Gene exon*, *Gene exon+intron* and *CDS* models and observed a substantial decline in the accuracy of COO predictions and a significant reduction in the explained variance when intronic mutations were not present, which was most pronounced in the liver and skin cancer COO models (Figure 66). Among all cancer types, skin melanoma exhibited the highest overall percentage of explained variance with correct COO identification across all gene-based COO model settings, except *Gene exon* models. In liver and breast cancer, the correct COO was not identified using either all genes or tissue-specific genes, with the best models suggesting immune cells or adipose tissue as the likely COO. All cancer types showed the highest variance explained and correct COO identification when using the top 40% of all mutated genes specific to each cancer type. Consequently, the following sub-chapters focus on the characterization of the top N% mutated genes from WGS data and a detailed analysis of the WGS-based COO models for each cancer type.

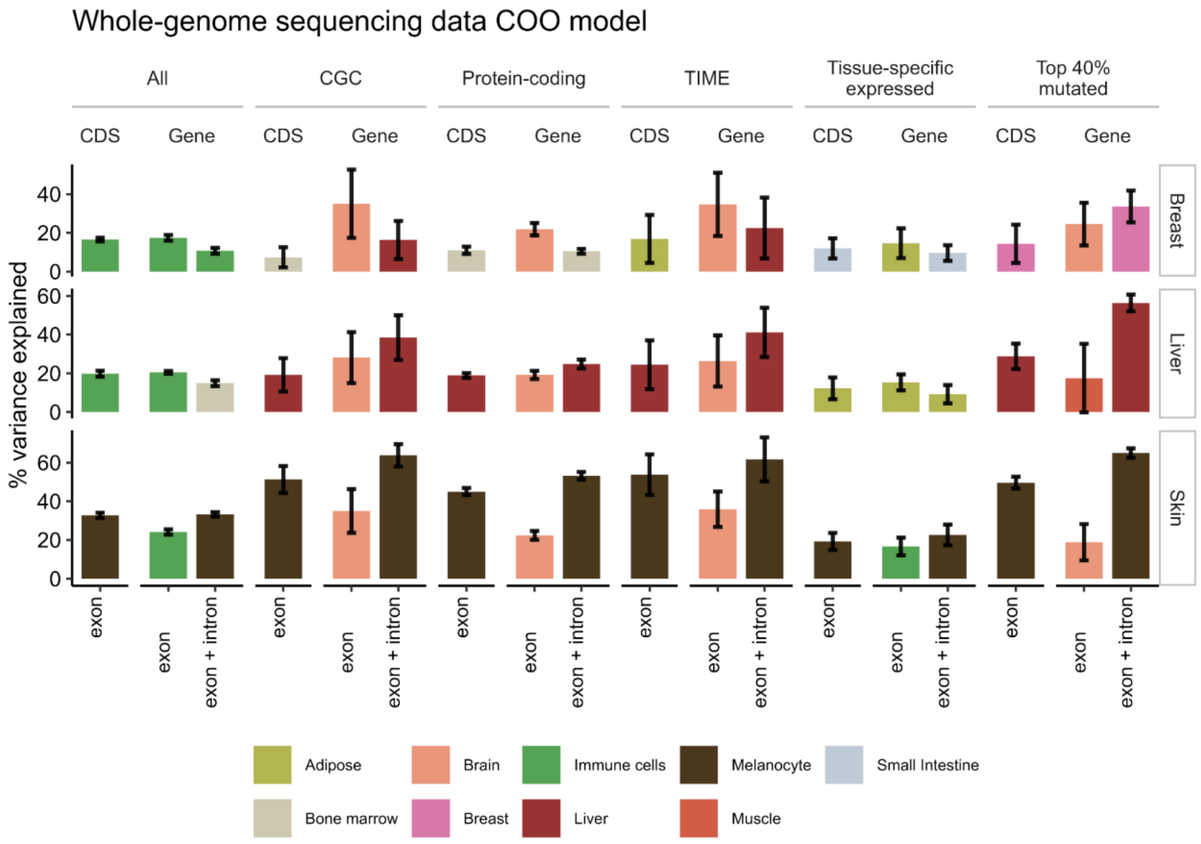


Figure 66. *The top performing multiple linear regression models for the prediction of mutation density of aggregated tumor profiles of breast, liver and skin cancer WGS in various gene settings. Coding-sequence (CDS) settings represent the count of mutations per exon normalized length by total CDS length. Gene setting implies summing mutations per exon or exon plus intron normalized by total gene length. The overall explained variance is reported across the 10-fold cross-validation.*

Preliminary downstream analysis of disease pathways indicated that outlier genes with higher than expected prediction errors (measured as standardized residuals) were implicated in pathways associated with male reproduction, such as spermatogenic failure and oligospermia, particularly in skin melanoma (ORA, $p < 0.05$). Given the unexpected nature of these findings, I investigated potential biases in standardized residual values with respect to chromosomal location. It was found that genes on the Y chromosome exhibited the highest absolute standardized residual values, followed by those on the X chromosome (Figure 67). This bias was notably pronounced in liver cancer, where genes on both sex chromosomes had higher error rates compared to the median value across all genes (Wilcoxon test, p-value < 0.05). Consequently, genes located on sex chromosomes were excluded from subsequent analyses to mitigate this bias.

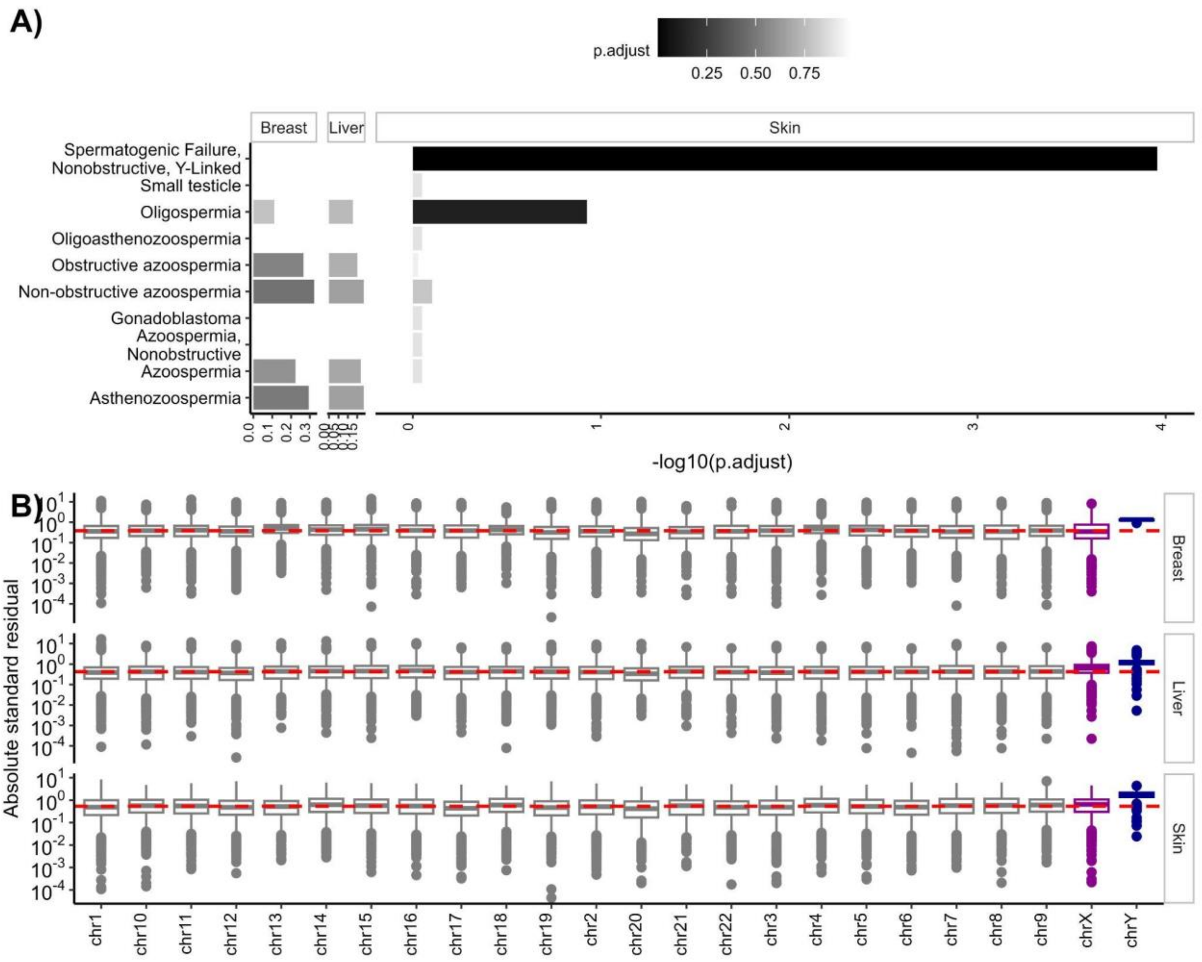


Figure 67. *A) Disease over-representation analysis (ORA) from DisGeNET of all WGS data outliers genes, absolute residual value greater than 2, determined in corresponding correct COO tissue. B) Log transformed absolute standard residual values per all genes. Autosomal chromosomes are colored grey while X and Y chromosomes are colored darkmagenta and darkblue. Dashed red line represents the median value of absolute standard residual values across all genes in each cancer type for the correct COO model. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

4.3.3.2 Gene-based COO predictive model on WGS data

Firstly, I characterized the top N% most frequently mutated genes in breast, liver, and skin cancers, excluding genes from the sex chromosomes. Subsequently, I re-ran the gene-based COO models with these excluded genes to maintain consistency across all cancer types. The analysis and interpretation were focused on gene-based COO model results obtained from mutations detected by whole-genome sequencing data, due to the significantly superior performance observed across all three cancer types, particularly with the use of the top mutated genes.

4.3.3.2.1. Frequently mutated genes in different cancers are involved in brain-specific processes

Frequently mutated genes in cancers are often reported based on their known implications in cancer or by focusing solely on driver mutations. However, since background mutations also accumulate in cancers, in addition to driver mutations, I examined the most mutated genes, considering all mutations regardless of their location in introns or exons. Analysis of all genes and protein-coding genes showed that, compared to breast and skin cancer, liver cancer had twice as many genes among 5% of most frequently mutated genes (Figure 68A). As the percentage of top mutated genes increases, the number of top mutated genes in skin cancer surpasses those in breast and liver cancers. The majority of the genes were shared among cancers, with 468 overlapping genes in all cancer types for top 40% mutated genes (Figure 68B). Almost all of the top 40% mutated genes in breast and liver cancer overlap with the top all 40% mutated genes identified in skin cancer.

The over-representation analysis (ORA) using Gene Ontology (GO) terms for the top 40% mutated genes within each cancer type separately shows that most pathways are shared among cancers (Figure 68C). These pathways primarily involve brain-related processes such as the regulation of neuron projection development and synaptic transmission. Additionally, skin melanoma is enriched in the process of axonogenesis, while breast and liver cancers include locomotory behavior as enriched terms.

Further validation of these findings was performed by re-running the ORA on the 468 overlapping genes across all cancer types. The analysis confirmed that these genes are predominantly involved in brain-specific processes, including synaptic transmission, protein localization to synapse, and the regulation of neuron projection development (Figure 68D). Moreover, certain processes were uniquely enriched in most frequently mutated genes in breast cancer, such as the development of primary female sexual characteristics.

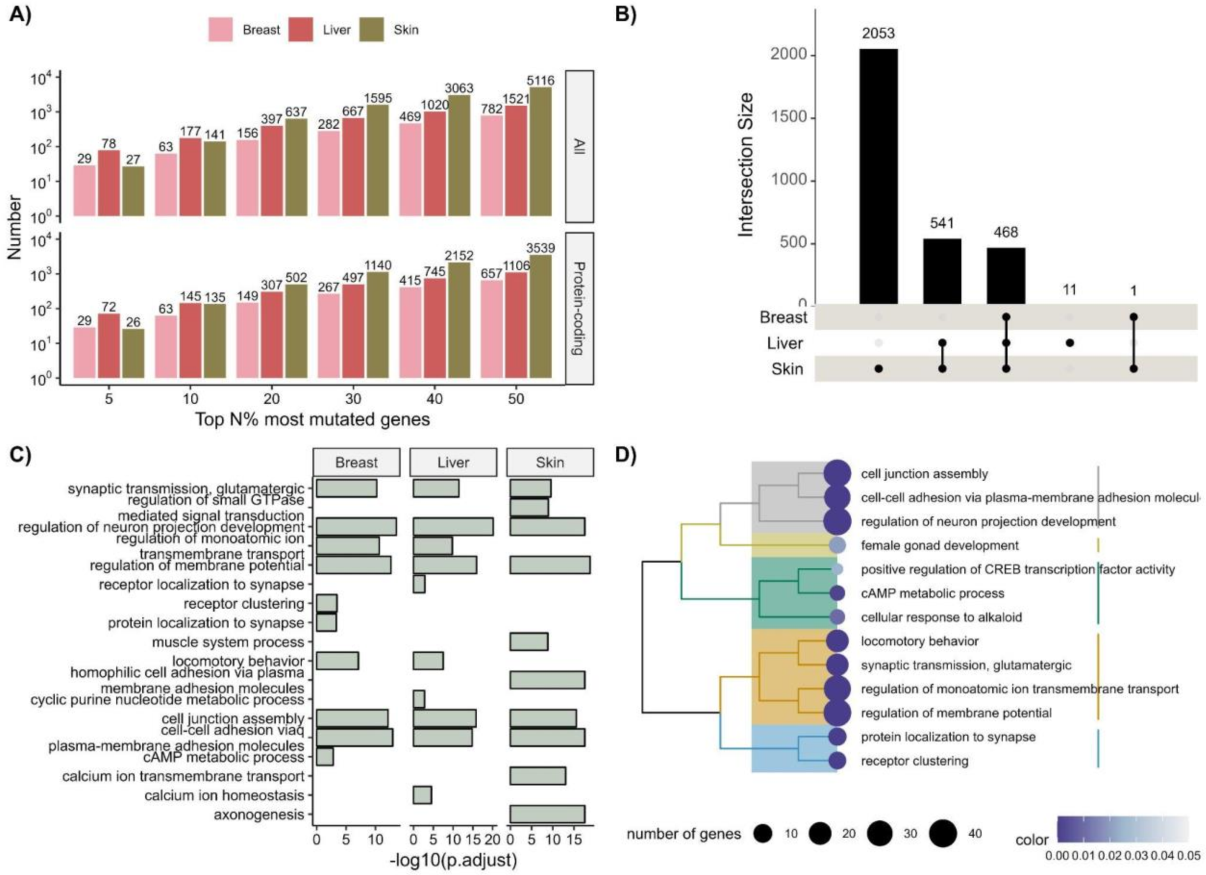


Figure 68. *A) Number of top N% frequently mutated genes (all and protein-coding) in breast, liver and skin cancer. The x-axis represents the top N% of mutated genes, where N ranges from 5% to 50%. The y-axis shows the corresponding number of genes within each percentage category. B) Upset plot showing the overlap of top 40% mutated genes across different cancer types. C) Gene ontology (GO) over-representation analysis showing the top 10 enriched pathways in each cancer type of top 40% mutated genes. D) Hierarchical clustering of enriched biological processes associated with 468 overlapping 40% all mutated genes found in breast, liver and skin melanoma cancers. Each row represents a biological process from GO terms, and processes are color-coded according to their statistical significance (p.adjust values), where darker blue shades represent more significant enrichments. The size of the circles corresponds to the number of genes associated with each process.*

Additionally, I examined the expression levels of the 468 overlapping genes in normal tissues using GTEx TPM data across 30 tissues. I found that these genes were more highly expressed in brain, nerve, and testis tissues (Figure 69). In contrast, their expression was significantly lower in normal breast ($p=5.710*^{-14}$), liver ($p=1.7*10^{-58}$), and skin ($p=1.6*10^{-20}$) tissues compared to the highest expression levels in brain tissue, as determined by the Wilcoxon test.

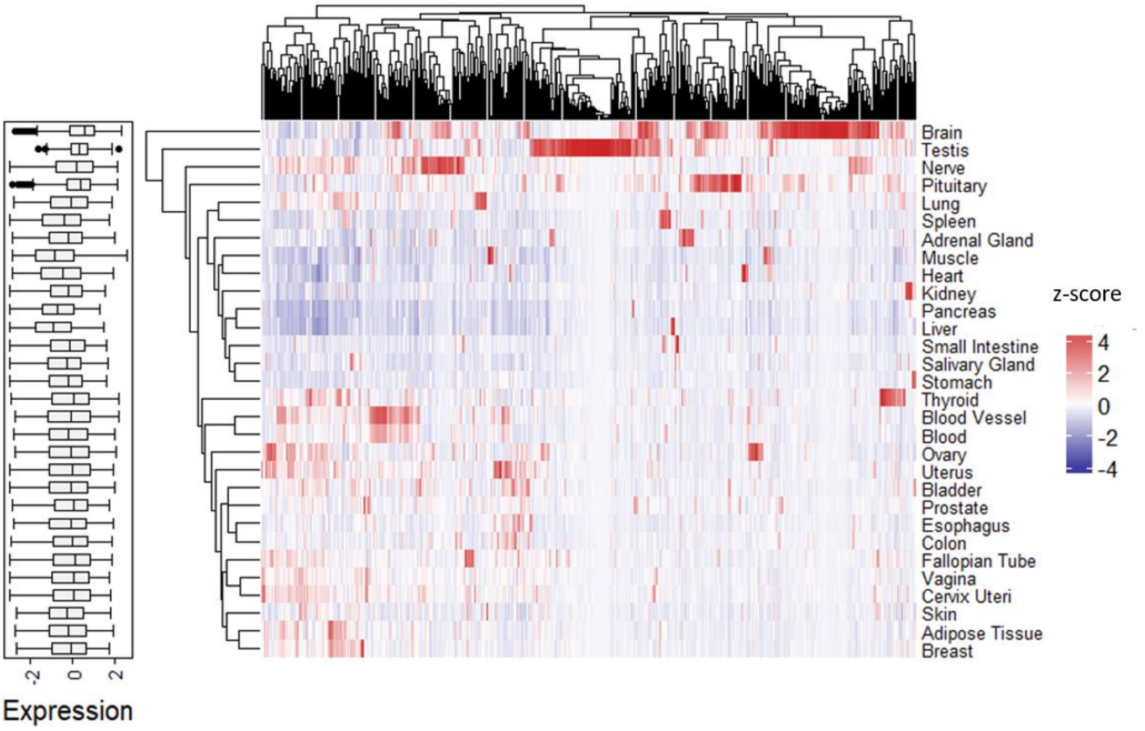


Figure 69. *Unsupervised clustering (average) using Euclidean distance of GTEX TPM normalized 30 tissues of 468 overlapping 40% all mutated genes found in breast, liver and skin carcinoma tissues. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR of normalized expression TPM values for each normal tissue.*

4.3.3.2.2 Breast cancer COO gene-based predictive model

To investigate how genes can be used to predict the cell-of-origin (COO) based on the breast cancer gene mutational profiles and the epigenomes of normal tissues, I created various gene subgroups: all genes, protein-coding genes, tissue-specific genes, driver genes, genes grouped according to HMEC IS 10kb topologically-associated domains (TAD-grouped genes), and the top N% mutated genes. The multiple linear regression analysis using these defined gene groups demonstrated that the correct COO of aggregated breast cancer mutational profiles could only be reliably identified using either TAD-grouped or top N% mutated genes (Figure 70). Additionally, these same groups exhibited the highest explained variance. For instance, TAD-grouped protein-coding genes had an explained variance of approximately 28%, while the top 10% mutated protein-coding gene group showed an explained variance of around 54%. Other defined gene subgroups had about 10% explained variance when the COO was incorrectly identified, except for driver genes, which had an incorrectly identified COO with more than 20% explained variance. Moreover, the second-best COO model, which was not the correct COO in any model setup, performed approximately 4% worse on average than the top COO model, with larger discrepancies observed in the top N% of mutated genes. Gene ontology and the top 5% mutated genes failed in COO modeling due to the lower number of genes and overall mutations in them, so they were not included in downstream analysis.

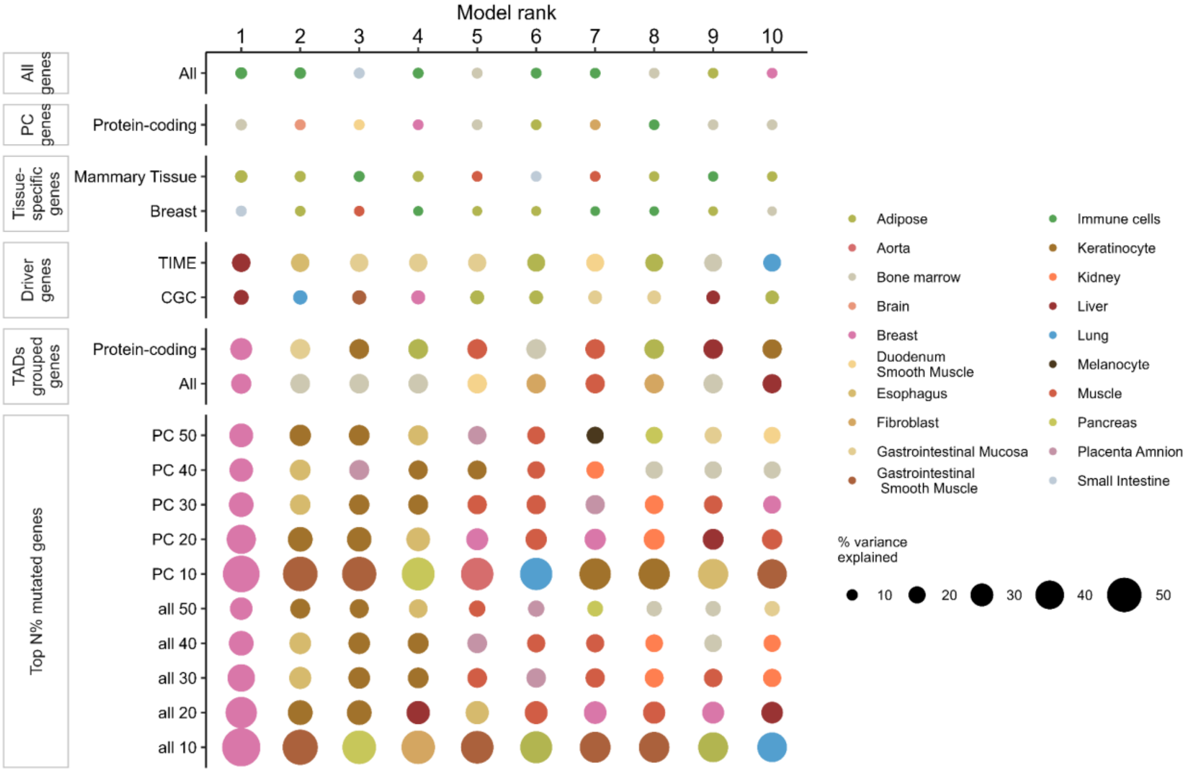


__Figure 70.__ Multiple linear regression models for the prediction of mutation density of aggregated tumor profiles of breast WGS were trained on an extended set of 101 tissue sets but showing only the top 10 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation.

Prediction on individual breast cancer patients across all COO model setups using various gene subsets resulted in the majority of patients having incorrectly identified COO. Only in the top N% mutated gene subsets, both for all genes and protein-coding genes, were the models able to correctly identify the COO for up to 10% of patients (Figure 71A), with the model trained on top 40% of all mutated genes achieving highest accuracy, and models trained on all genes, all TADs-grouped genes and protein-coding TADs-grouped genes achieving lowest accuracy. Across all models, the cell-of-origin assigned to individual breast cancer patients was identified as either immune cells or brain-related tissues (Figure 71B). Immune cells were predominantly assigned in models that included all genes. In contrast, when the analysis was restricted to protein-coding genes, there was a higher misclassification rate involving brain-related tissues. Additionally, a significant proportion of incorrectly identified COO were assigned as originating from the thymus.

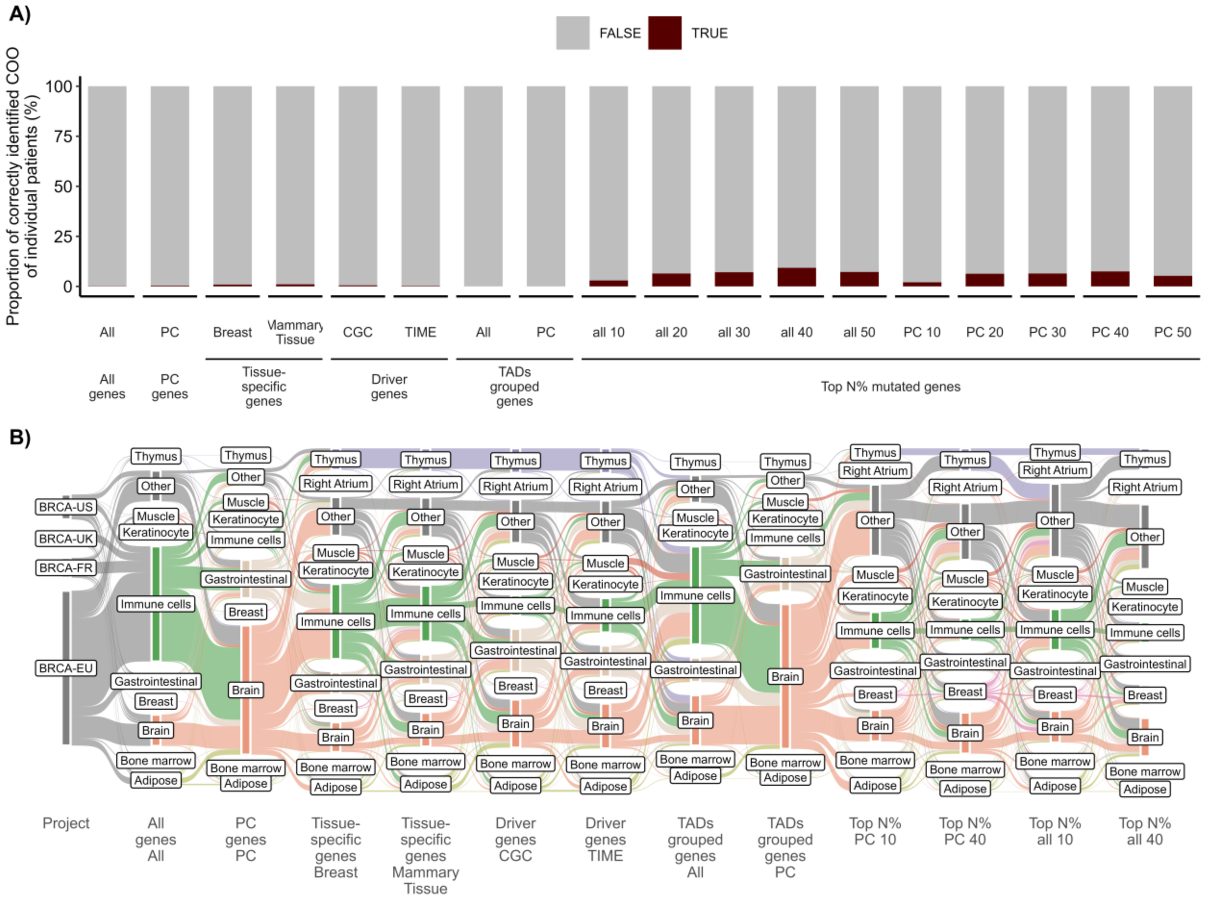


Figure 71. *A) The proportion of individual samples in which the prediction on an individual level matches the correct cell of origin of the highest explaining models in different modeling setups using various subsets of genes for breast cancer. B) Alluvial plot illustrating the cell-of-origin (COO) of individual patients across various COO model setups based on different gene subsets*

To investigate the significant differences observed between developed breast cancer cell-of-origin (COO) models based on different gene subsets, I first examined the gene length and the number of mutations within these subsets. I found that the mutation counts, normalized by gene length, exhibit considerable variability across different gene subgroups (Figure 72A). The top N% mutated genes showed the highest normalized mutational count, followed by all genes, protein-coding genes, and TIME genes. The Cancer Gene Consensus set had one of the lowest normalized mutational counts. When I examined gene length distributions, I found that they varied widely across different groups (Figure 72B). For instance, the all genes and all TADs grouped genes subsets displayed a wide range of gene lengths, from very short genes (around 10^2 bp) to very long genes (up to 10^7 bp). In contrast, the top N% mutated gene subsets exhibited a narrower range of longer gene lengths, with an increasing trend towards longer lengths in the most frequently mutated genes. Specifically, the top 10% and top 20% mutated gene subsets had the longest gene lengths on average.

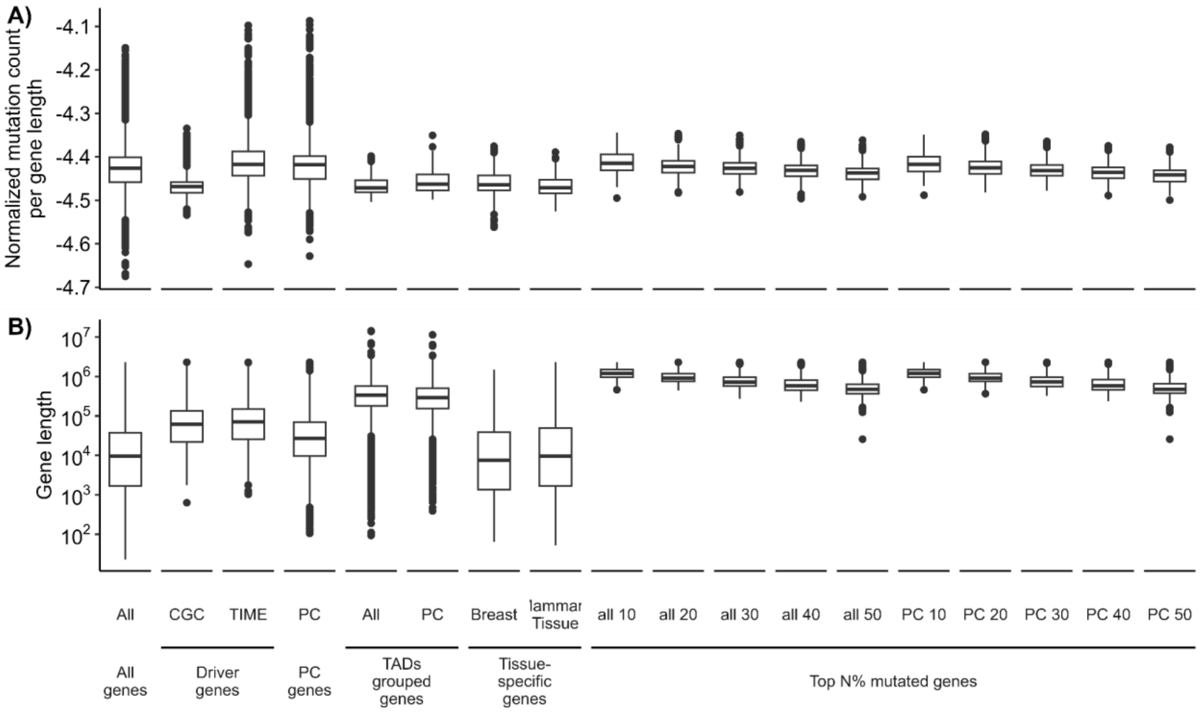


Figure 72. *A) Distribution of normalized mutation count per gene length of various groups of genes and B) gene lengths used for development of COO models in breast cancer. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

Moreover, based on the Spearman's correlation with the best assigned COO model epigenome, which indicates the potential cell-of-origin of breast cancer, I found that both the TADs grouped gene subsets and the top N% mutated genes exhibited the strongest positive correlation with repressive histone marks H3K9me3 and H3K27me3(Figure 73). While TADs grouped genes had higher positive correlation, top N% mutated had higher negative correlation with repressive chromatin marks. Specifically, the top 20% mutated protein-coding genes showed a strong positive correlation with H3K9me3 (0.732) and H3K27me3 (0.651) and strong negative correlations with H3K4me1 (-0.547) and H3K36me3 (-0.654). Similar trends, but weaker, were observed for other top N% mutated gene subsets, such as the top 40%, top 30%, and top 20% mutated genes, indicating a robust relationship between these repressive marks and the mutational profiles of breast cancer genes. Driver genes also exhibited higher correlations with the epigenome for their assigned COO, with correlations above 0.600 for H3K27me and above 0.490 for H3K9me3. However, unlike the top N% mutated gene subsets, driver genes with incorrectly identified COO demonstrated lower negative associations with open histone marks, H3K4me1 and H3K36me3. Also, all genes and tissue-specific gene subsets exhibited the weakest correlations, and their associations with repressive chromatin marks were in different directions compared to the other gene subgroups.

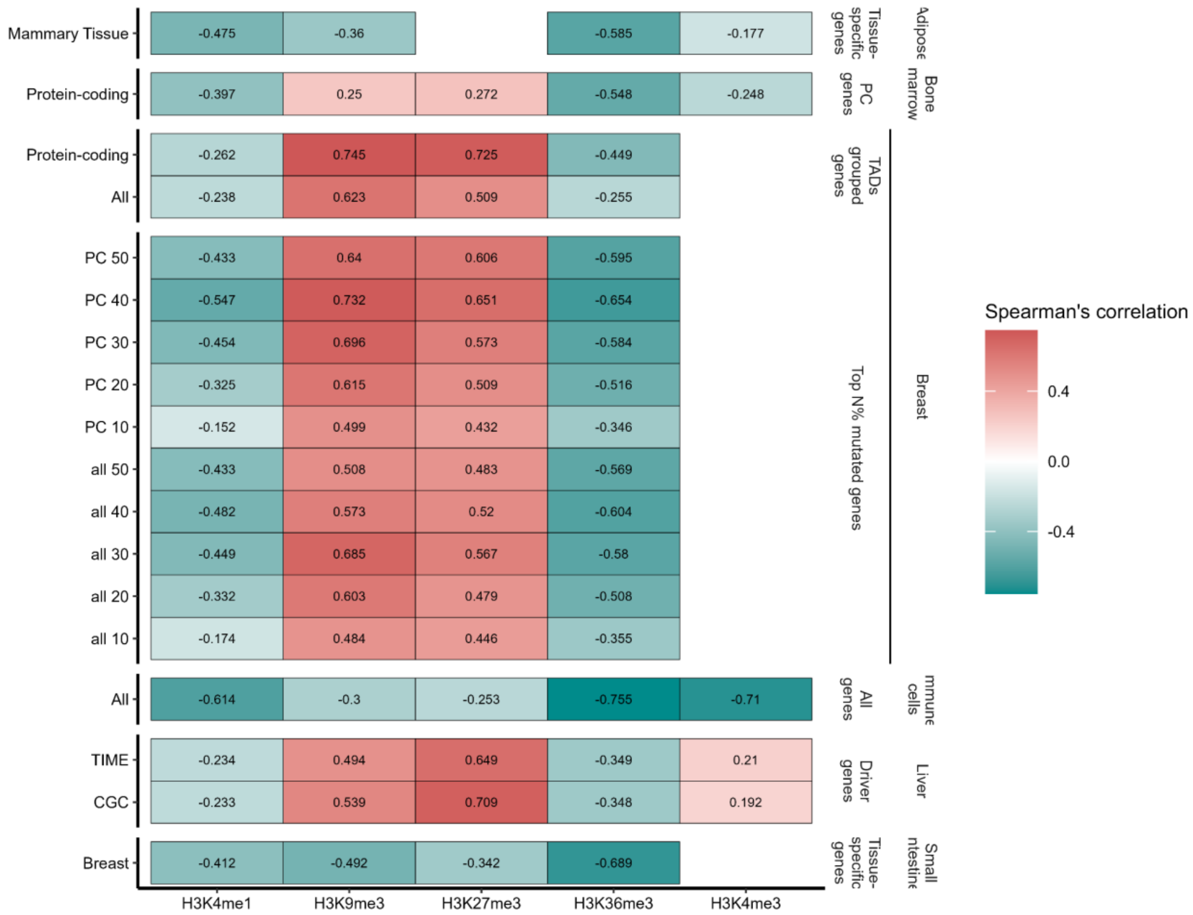


Figure 73. *Spearman's correlation of all mutations and histone modifications of top identified cell-of-origin tissue in various gene subgroups for breast cancer. White blank field represent not available histone marks in the dataset for specific epigenome of normal tissue*

There was no statistically significant difference between the SBS mutational profiles of various defined gene subgroups in breast cancer (Figure 74A). All defined gene subgroups were enriched with SBS12, SBS2, SBS40, and SBS8, which together accounted for more than 60% of the total breast cancer gene mutational landscape in each subgroup. The top N% mutated genes were more affected by kataegis and SV-hotspots than other groups, but none were associated with breast-epithelium super-enhancers (SE) (Figure 74B). The driver gene groups, CGC and TIME, had over 25% of genes with SV-hotspots or kataegis. Unlike the top N% mutated genes, they had the highest proportion of breast-epithelium SE.

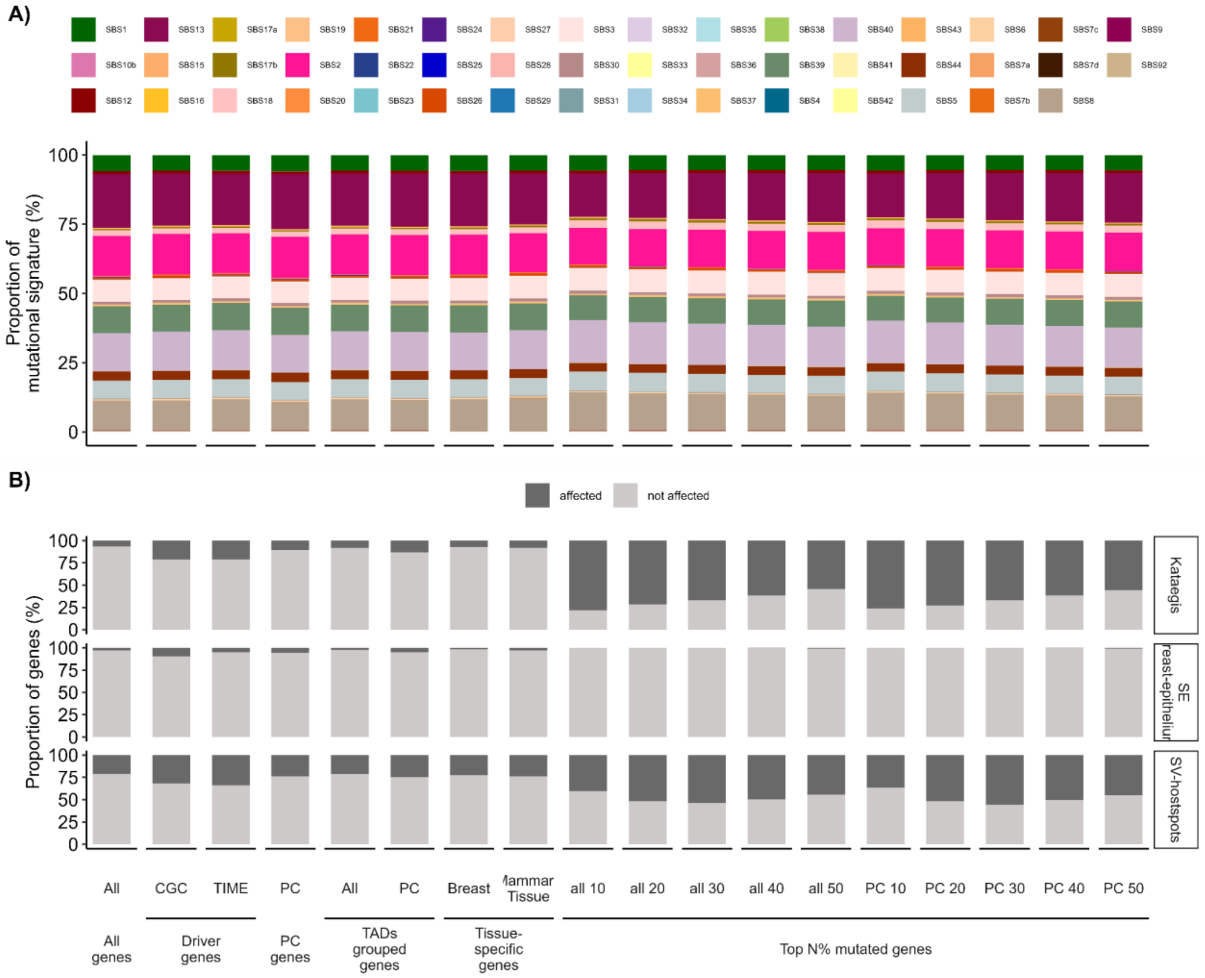


Figure 74. *A) Proportion of mutational signature in genes separated by various gene subgroups of breast cancer B) Proportion of affected genes by either kataegis, SV-hotspot or are proximal to tissue specific super-enhancers (SE) from breast-epithelium tissue*

4.3.3.2.2.1 Top 40% mutated breast genes as best gene model

I selected the top 40% most mutated genes, which produced the best results in the multiple linear regression COO gene-based model, and then ran random forest and extreme gradient boosting models to determine if the results would significantly improve. Unfortunately, the extreme gradient boosting model failed to run successfully and did not produce any COO predictions due to low mutation count and not enough variation in the splits of the advanced models.

In contrast, compared to linear regression, the random forest COO model applied to aggregated breast cancer profiles resulted in a higher explained variance (~30%) for both all genes and protein-coding genes (Figure 75A). Additionally, other breast tissues ranked second in the model trained on top 40% protein-coding genes.

However, predictions on individual patients performed worse than the multiple linear regression model (Figure 75B), with 72 patients correctly identified by the multiple linear regression model and only 55 patients correctly identified by the random forest model.

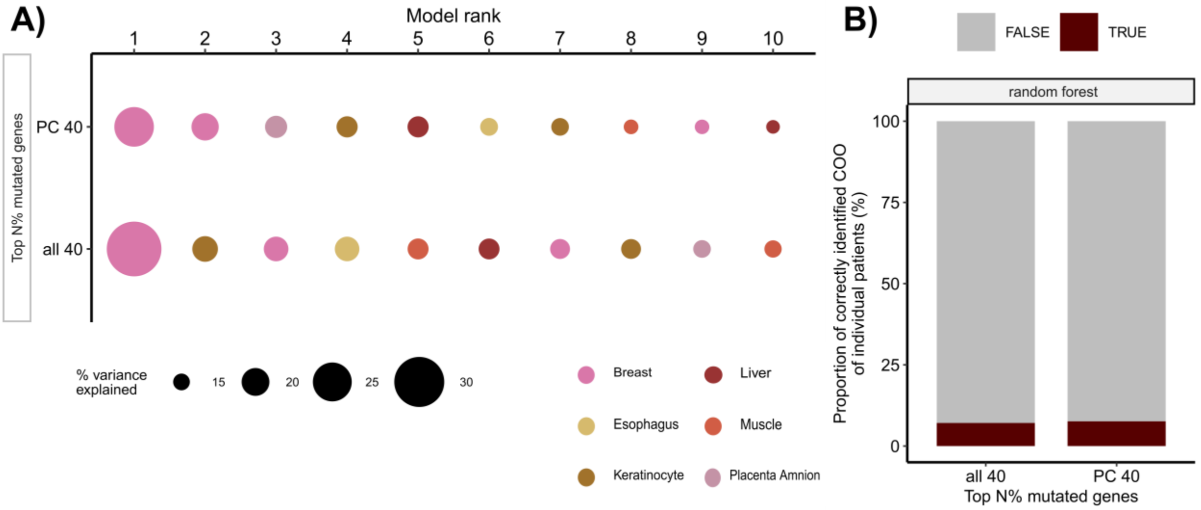


Figure 75. *A) Random forest models for the prediction of mutation density of aggregated cancer profiles in top 40% mutated genes of breast WGS were trained on an extended set of 101 tissue sets but showing only the top 10 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation. B) Proportion of correctly and incorrectly identified COO of individual patients using random forest model with top 40% mutated genes*

I decided to analyze in greater detail the top 40% protein-coding genes to identify potential new driver genes for breast cancer. Overall, I detected 20 under-predicted genes, 2 over-predicted genes, and 393 non-outlier genes (Figure76A). Under-predicted genes had higher proportions of SV-hotspots and kataegis compared to non-outliers and over-predicted genes (Figure 76BC).

Over-representation analysis (ORA) on databases of disease-gene associations using under-predicted genes did not yield significant terms related to breast cancer development after adjusting for multiple hypothesis testing. However, when searching for breast cancer-related diseases in the results, I found some terms that were present but did not reach statistical significance: DisGeNET term “C0678222 Breast Carcinoma” with 6 genes (p-value = 0.95): *CTNNA2*, *EIF3E*, *TRPS1*, *ANKRD30A*, *ZFPM2*, and *TSHZ2*.

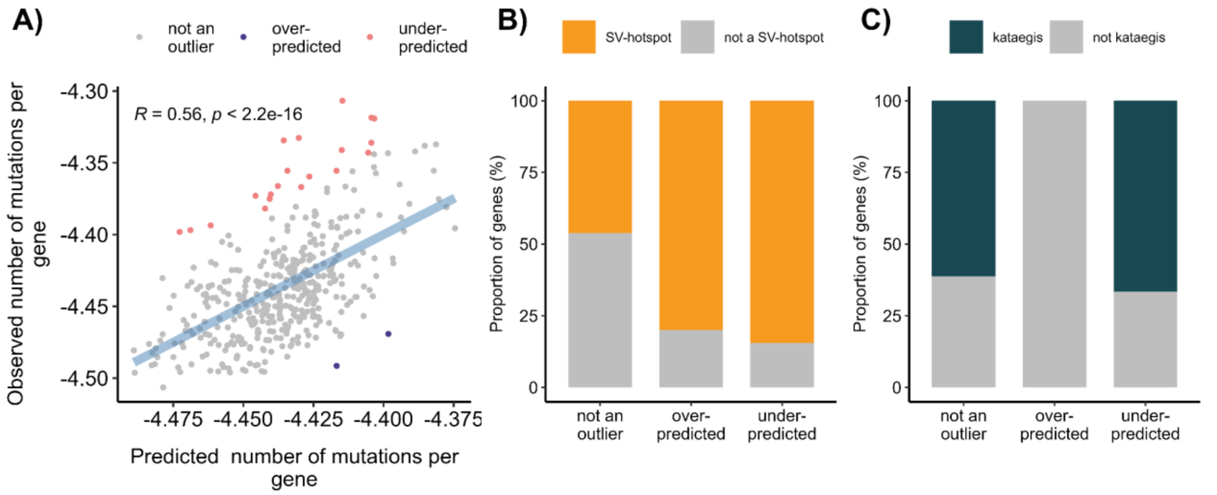


Figure 76. *A) Spearman correlation of observed vs predicted normalized number of mutations per gene in top 40% mutated protein-coding genes in breast cancer. B) Proportion of regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier C) Proportion of regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier*

Afterwards, I examined TAD regions in which these genes occurred. The two over-predicted genes had larger TADs stability score, which reflects the conservation of TAD boundaries across many cell types, than non-outlier and under-predicted genes, although this difference was not statistically significant (Figure 77A). I found a significantly different landscape of different outlier genes based on their location in certain TADs domain (Chi-square test, p-value=0.00238) (Figure 77B). Active genes were only found in non-outlier breast genes, while outliers were more enriched in more closed, repressive and heterochromatin TADs.

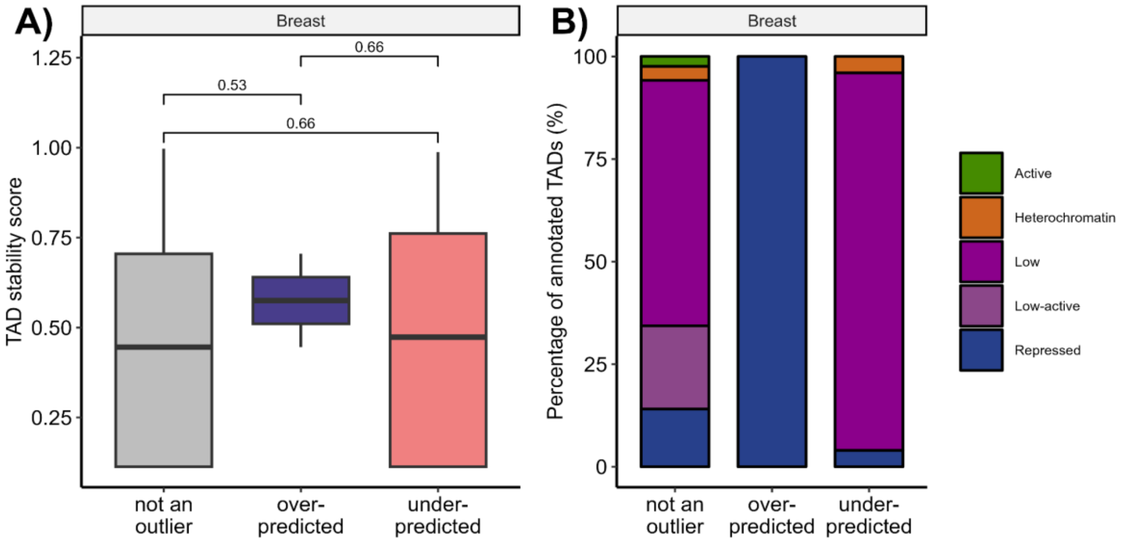


Figure 77. *A) TADs stability score of TADs boundries distribution in annotated erroneous regions B) TADs annotation by Akdemir et al. 2020. Based on active and inactive state (heterochromatin, low, low-active and repressed)*

4.3.3.2.2.2 Patient characteristics based on the best gene COO model

I analyzed the prediction results on individual patients from top 40% protein-coding gene-based COO models developed using multiple linear regression. Although the number of mutations per patient was significantly higher in correctly identified COO patients using the top 40% protein-coding genes, the explained variance was significantly lower for these correctly identified COO patients (Figure 78A). Despite the lack of statistical significance, there was a trend for SV-hotspots to be more prevalent in correctly identified COO patients, while kataegis occurrences were quite similar between correctly and incorrectly identified COO patients.

When examining the proportion of SBS mutational signatures per patient in the top 40% protein-coding genes, which were among the most abundant and had noticeable effects in other developed COO models, I found that incorrectly identified COO patients had significantly higher proportions of SBS1, SBS3, SBS8 and SBS40 (Figure 78B). Additionally, SBS13 tended to be higher and SBS2 lower in incorrectly identified COO patients.

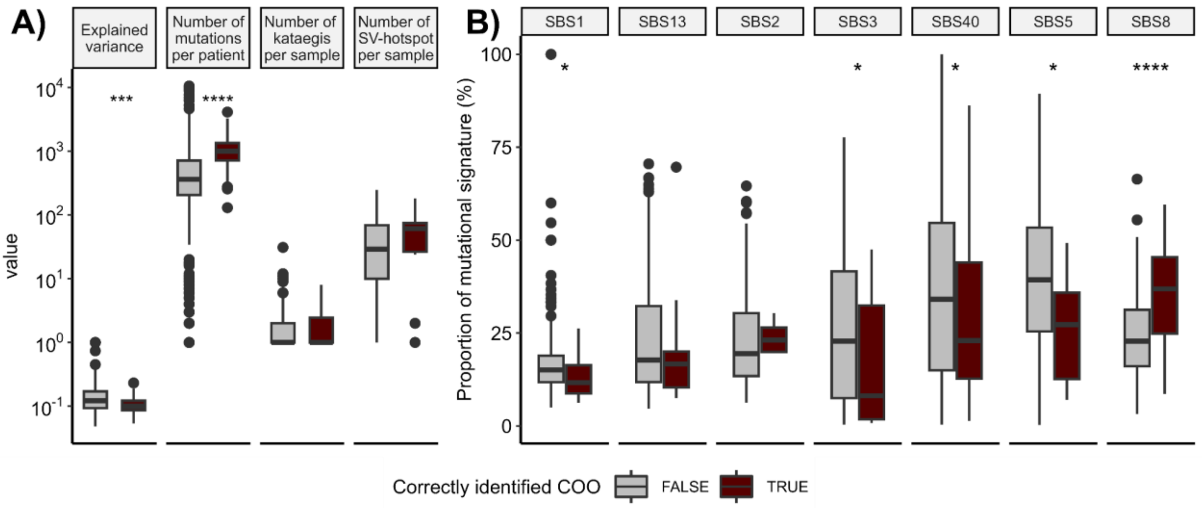


Figure 78. *A) The distribution of explained variance and the number of mutations separated by correct or incorrect COO identification of individual breast cancer patients. B) Proportion of SBS mutational signatures per breast cancer patient separated by correct or incorrect COO identification of individual patients. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Furthermore, I separated the breast cancer individual patients' COO results based on their histological type (Figure 79). Notably, metaplastic carcinoma had the highest proportion of correctly identified COO patients at 40%, followed by mucinous adenocarcinoma and infiltrating duct carcinoma at 6% and 8%, respectively. Patients with unknown cancer types also had around 8% correctly identified COO. Other subtypes, such as tubular and invasive cribriform carcinoma, pleomorphic carcinoma, and neuroendocrine carcinoma, showed no correctly identified COO patients.

Patients with correctly identified COO had a higher number of mutations and lower explained variance, particularly in infiltrating duct carcinoma, where the difference was statistically significant. Only the unknown subtypes showed a trend of higher explained variance alongside a higher mutational burden, but this was not statistically significant.

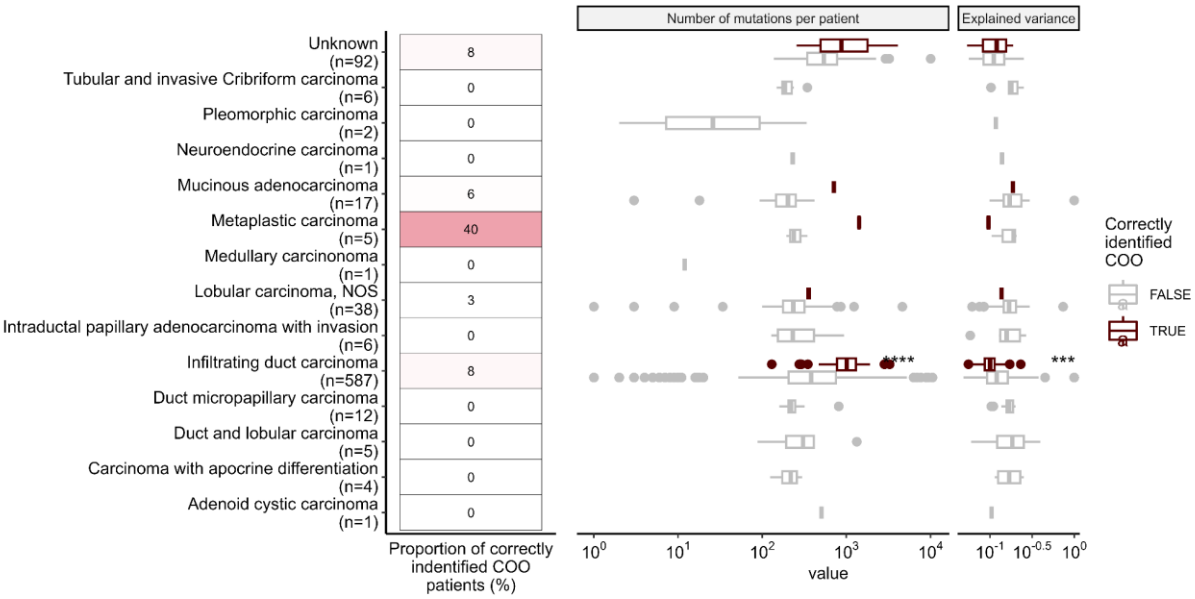


Figure 79. *A) Proportion of correctly identified cell-of-origin (COO) proportions using the top 40% mutated protein-coding genes, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Figure 80 contains the relationship between mutation counts, histological types, and homologous recombination deficiency (HRD) status in predicting the cell-of-origin (COO) for breast cancer patients. Despite a significant difference in the number of mutations per patient among subtypes (Kruskal-Wallis test, p = 0.023), none of the subtypes, including Her2, LumA, LumB, Normal, and Basal, showed any correctly identified COO patients (Figure 80A). Most of the PAM50 subtypes were infiltrating duct carcinoma.

Based on their HRD status, HRD patients exhibited a higher proportion of correctly identified COO patients (14%) compared to non-HRD patients (4%) (Figure 80B). Additionally, the number of mutations per patient was significantly higher in HRD patients, who were mostly of the infiltrating duct carcinoma subtype. A more detailed examination of the number of mutations per patient and explained variance using the top 40% protein-coding models showed that HRD patients with correctly identified COO had a significantly higher mutation burden but lower explained variance compared to those incorrectly identified (Figure 80C). Similarly, non-HRD patients with correctly identified COO also showed a significantly higher mutation burden and slightly higher explained variance, although the latter was not statistically significant.

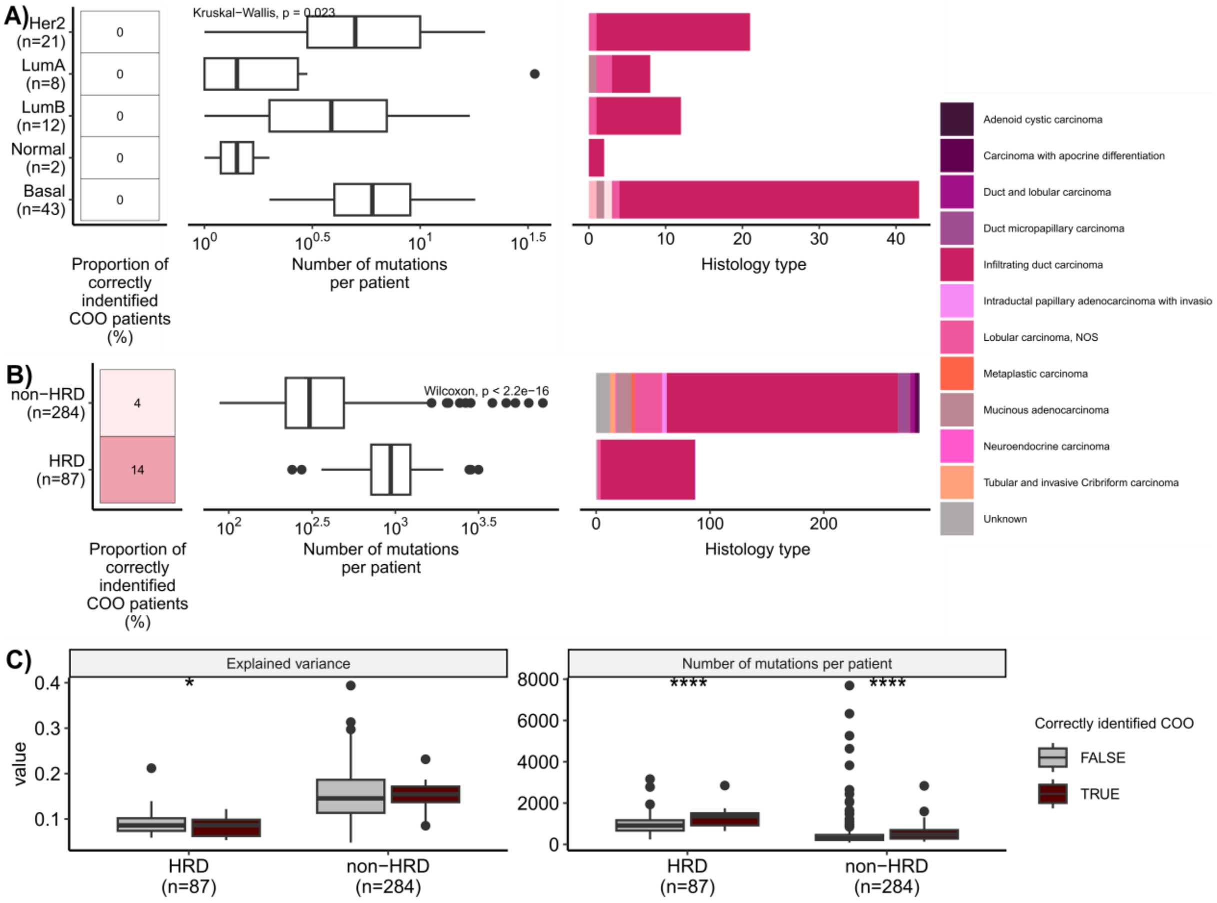


***Figure 80.** A) Proportion of correctly identified cell-of-origin (COO) proportions using the top 40% mutated protein-coding genes, separated by their PAM50 subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. B) Proportion of correctly identified cell-of-origin (COO) proportions using the top 40% mutated protein-coding genes, separated by their PAM50 subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.3.2.3 Liver cancer COO gene-based predictive model

Gene-based multiple linear regression COO models using various gene subgroups on aggregated liver cancer mutational profiles were able to identify the correct COO as liver tissue in almost all defined gene subgroups, except all and liver tissue-specific genes (Figure 81). The highest variance explained was achieved with top 40% protein-coding genes (~56%). The driver gene subsets, including the CGC and TIME genes, exhibit higher variances compared to other subsets, with values reaching up to 30% and 40%, respectively. The liver-specific gene subset showed a smaller explained variance of around 10% for the best model, incorrectly predicting adipose tissue as the COO. In general, the top incorrectly identified COO for aggregated profiles in each gene subgroup had overall lower explained variance.

Moreover, the next best COO model which was not liver COO performed approximately 7% worse on average than the top COO model, with larger discrepancies observed in the top N% of mutated genes. Gene ontology and the top 5% mutated genes failed in COO modeling due to the lower number of genes and overall mutations in them, so they were not included in downstream analysis.

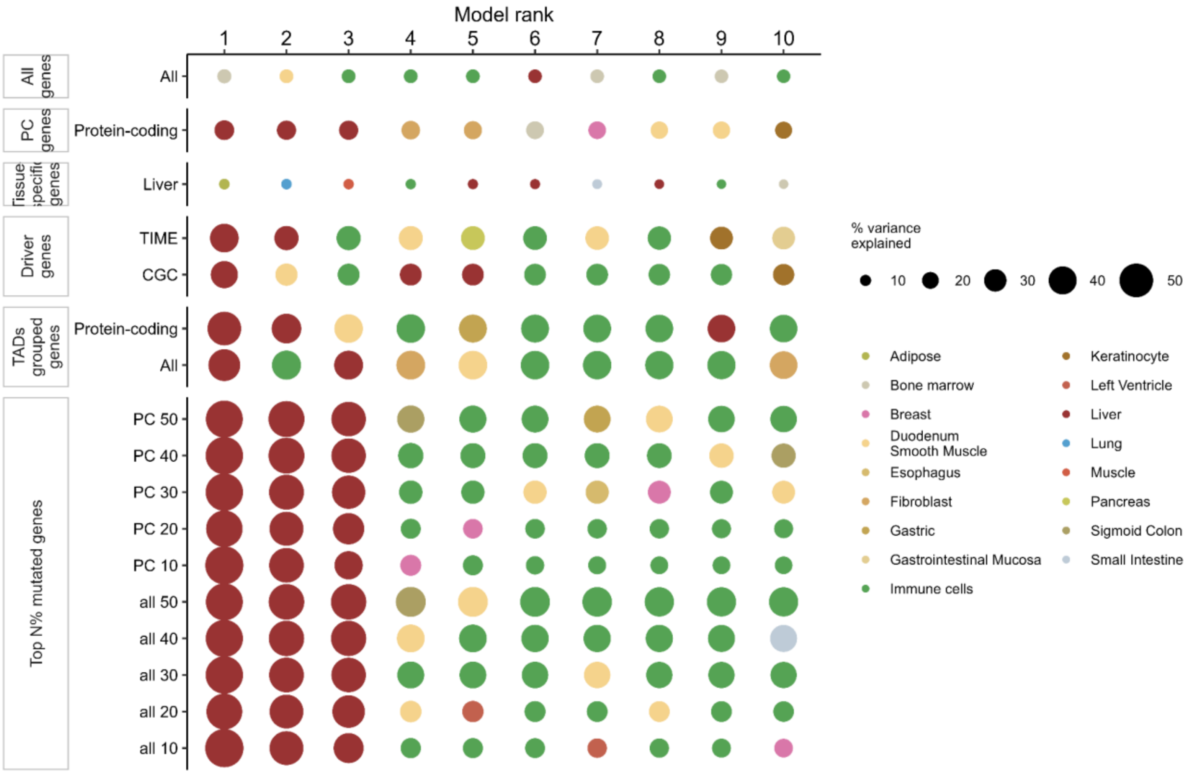


Figure 81. *Multiple linear regression models for the prediction of mutation density of aggregated tumor profiles of liver WGS were trained on an extended set of 101 tissue sets but showing only the top 10 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation.*

Prediction on individual liver cancer patients across all COO model setups using various gene subsets resulted in the majority of patients having incorrectly identified COO. The overall trend revealed that the top N% mutated genes subsets (particularly the top 10%, 20%, 30%, 40%, and 50%) exhibit a higher proportion of correctly identified COO patients compared to other gene subsets (Figure 82A). Specifically, the highest proportion of correctly identified COO liver cancer patients was detected for the top 40% all and protein-coding genes, around ~70%. The gene subgroups with the lowest number of correctly identified patients included the comprehensive gene sets (all genes and protein-coding genes) as well as the TADs-grouped gene sets (both all genes and protein-coding genes within TADs).

Most of the misclassified COO on individual patients was labeled as either brain-related tissues or immune cells (Figure 82B). Immune cells were predominantly assigned in models that included all and liver tissue-specific genes. Driver genes lead to increase of gastrointestinal tissue as primary identified COO for individual patients. Only when I used the top N% mutated gene subsets for COO model, did those misclassifications reduce. Depending on the top N% mutated gene subset, I still detected different proportions of incorrectly identified COO patients as brain, immune and even thymus as their COO.

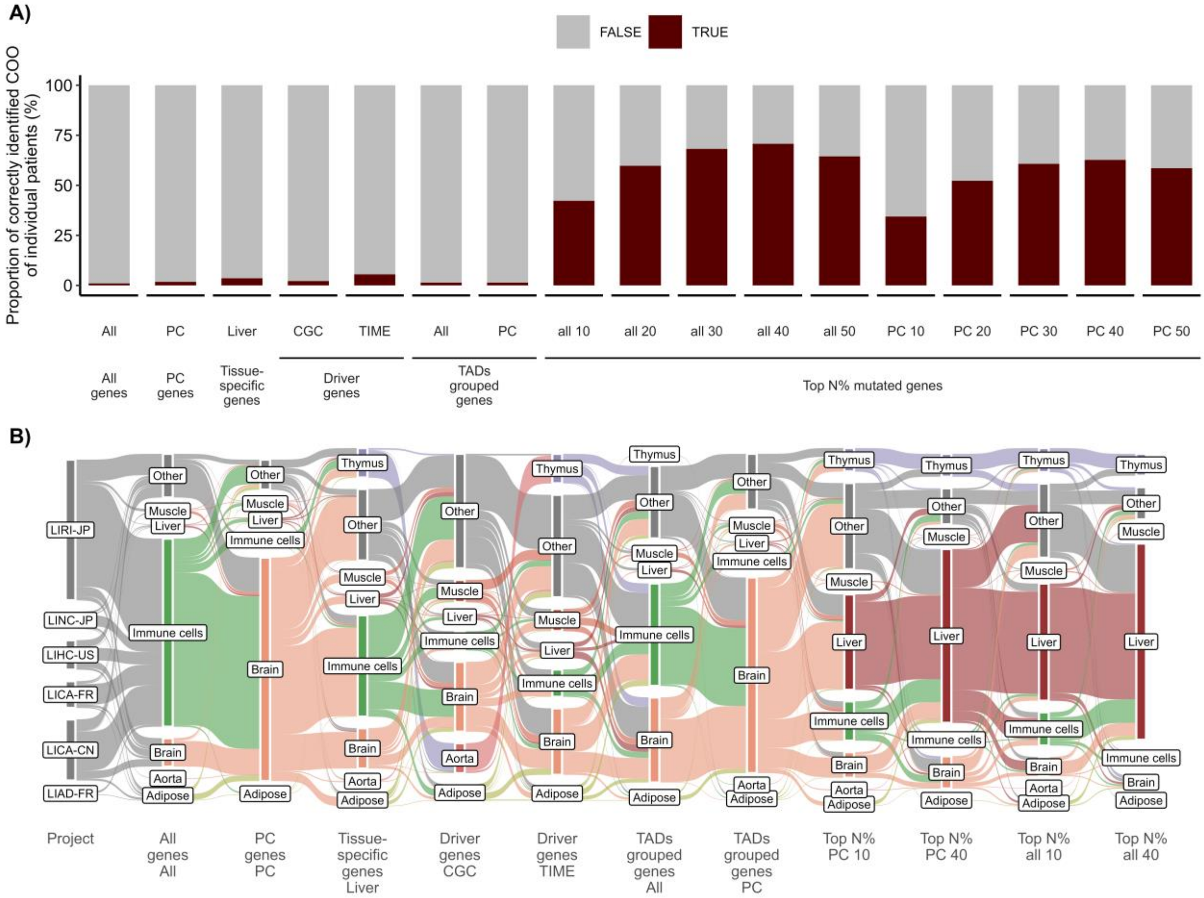


Figure 82. *A) The proportion of individual samples in which the prediction on an individual level matches the correct cell of origin of the highest explaining models in different modeling setups using various subsets of genes for liver cancer. B) Alluvial plot illustrating the cell-of-origin (COO) of individual patients across various COO model setups based on different gene subsets*

The top N% mutated liver cancer genes exhibited the highest number of mutations per gene when normalized by gene length (Figure 83A). In contrast, protein-coding genes, CGC driver genes, and liver tissue-specific genes had the lowest mutational counts. When examining gene lengths, the top N% mutated liver cancer genes again stood out with the longest gene lengths compared to other groups (Figure 83B). TADs-grouped genes followed, with longer total lengths but a higher number of detected outliers. Driver genes, including those from CGC and TIME, also had higher gene lengths. Conversely, liver tissue-specific genes and the set of all genes had the shortest gene lengths among all the subgroups.

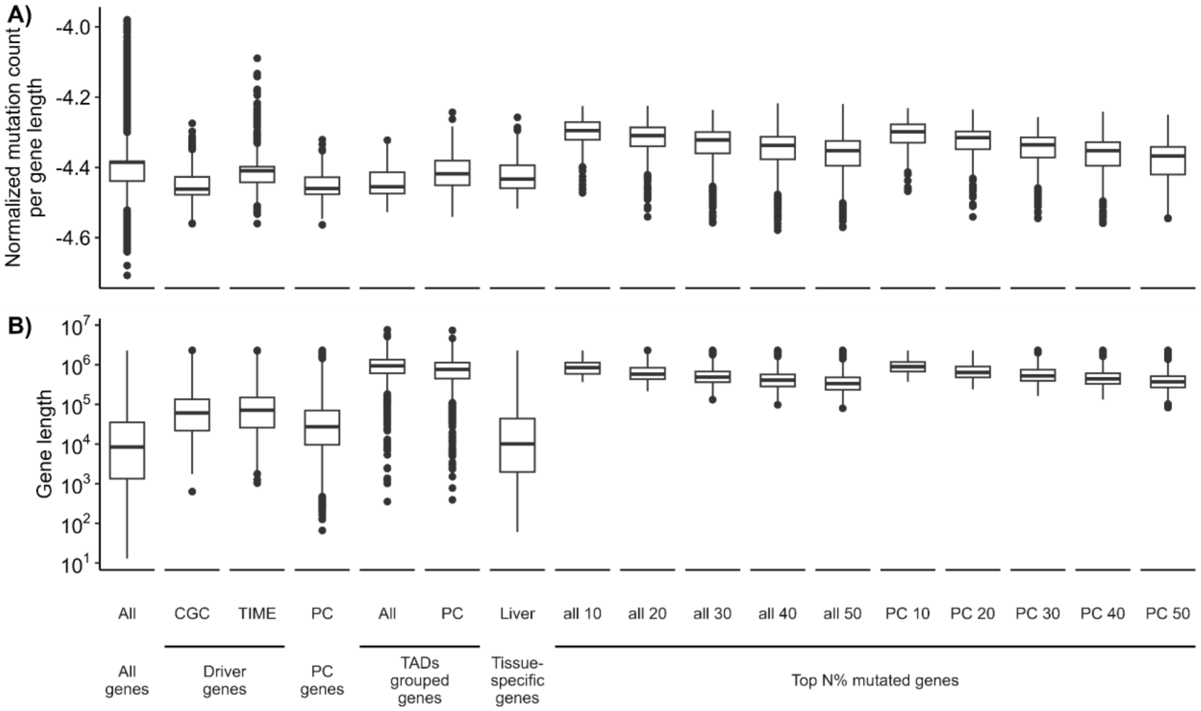


Figure 83. *A) Distribution of normalized mutation count per gene length of various groups of genes and B) gene lengths used for development of COO models in liver cancer. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

Furthermore, I observed that the correlation of aggregated liver cancer mutations with epigenomes for incorrectly identified COO, specifically as adipose for liver tissue-specific genes and bone marrow for all genes, exhibited much weaker positive correlations with closed chromatin modifications H3K9me3 and H3K27me3 (Figure 84). The strongest positive correlations with these closed chromatin modifications were found in driver genes and TADs-grouped protein-coding genes, with correlations exceeding ~0.7. Conversely, the strongest negative correlations with open chromatin modifications from liver tissues were detected for the top N% mutated genes, particularly the top 30% protein-coding genes, with correlations nearing -0.7 for H3K4me1 and H3K36me3. Histone modification H3K4me3 showed the weakest negative correlation across all gene subgroups, except in the group of all genes.

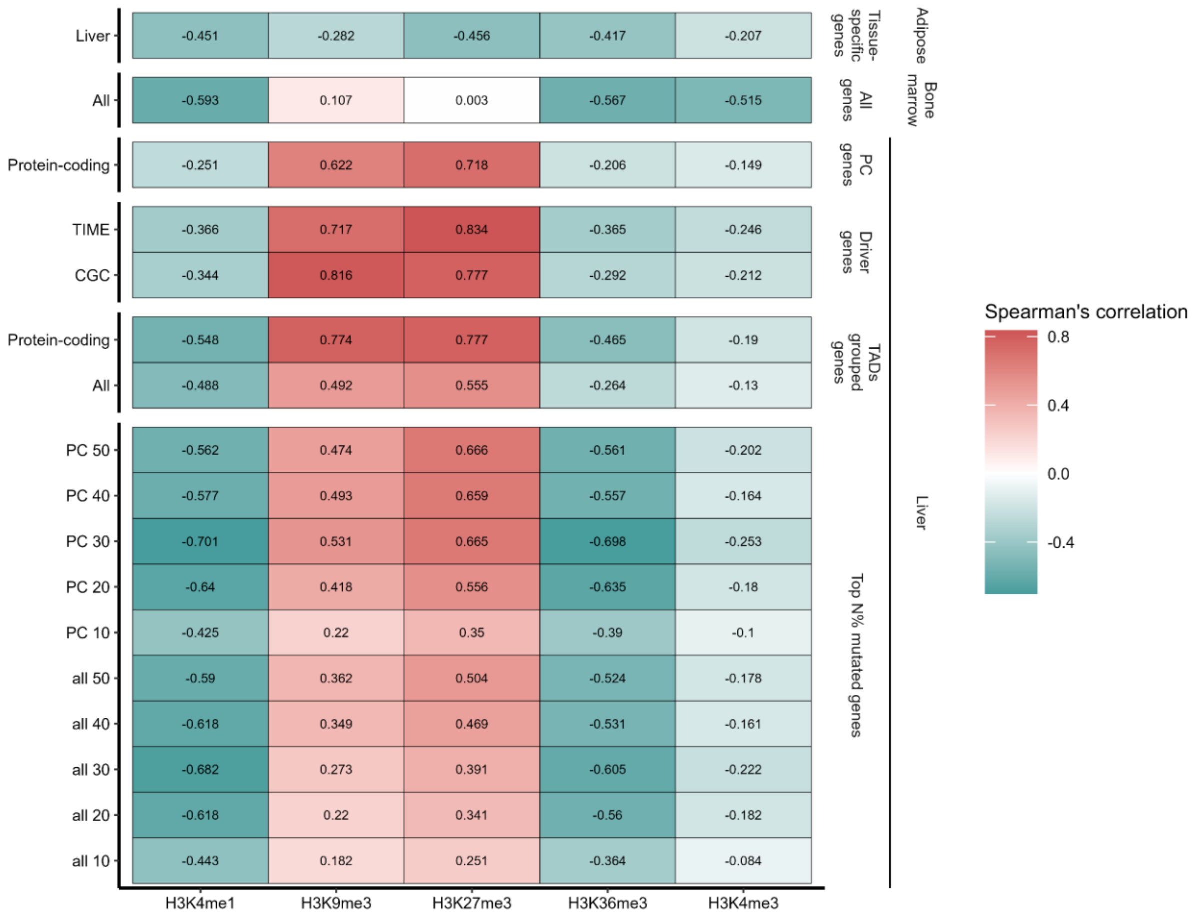


Figure 84. *Spearman's correlation of all mutations and histone modifications of top identified cell-of-origin tissue in various gene subgroups. White blank field represent not available histone marks in the dataset for specific epigenome of normal tissue*

The gene mutational landscape of liver cancer was relatively consistent across all gene subgroups shown in Figure 85A. The predominant mutational signatures in all groups were SBS40, SBS12, SBS23, SBS8, and SBS16. However, there was a notable difference in the genes affected by kataegis and SV-hotspots (Figure 85B). Kataegis were particularly abundant in the top N% mutated genes, affecting more than 50% of these genes, compared to other gene subgroups. Additionally, around 20% of the driver genes, including CGC and TIME, were affected by kataegis. Although SV-hotspots also impacted more top N% mutated genes than other subgroups, they affected only about 15% of these genes compared to kataegis. Super-enhancers, regardless of the cell line, affected only a small proportion of CGC, TIME, and protein-coding genes and did not affect the top N% mutated genes at all.

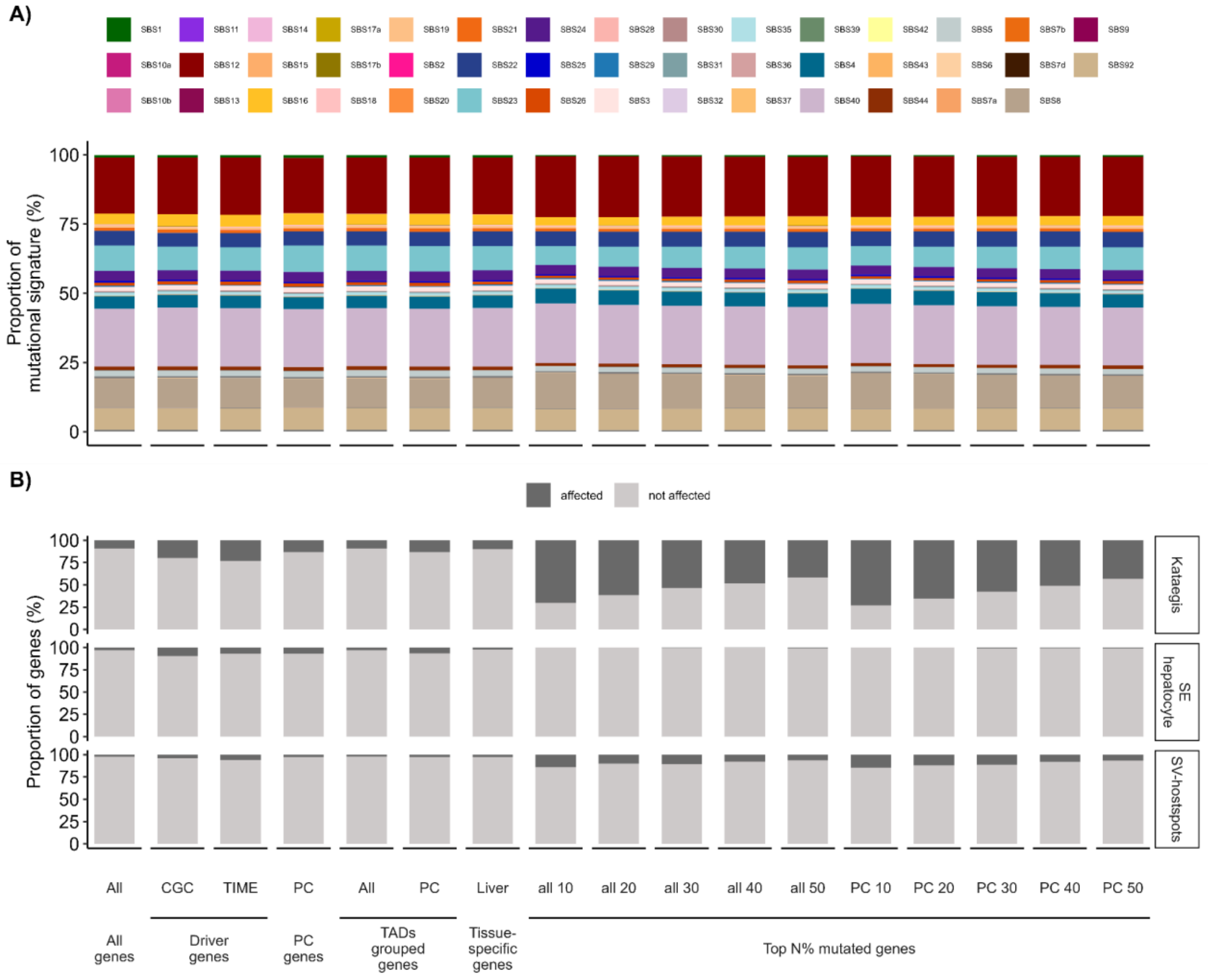


Figure 85. *A) Proportion of mutational signature in genes separated by various gene subgroups of liver cancer B) Proportion of affected genes by either kataegis, SV-hotspot or are proximal to tissue specific super-enhancers (SE) from hepatocytes*

4.3.3.2.3.1 Top 40% mutated liver genes as best gene model

For downstream analysis, I selected the top 40% most mutated liver genes due to their high accuracy in predicting the correct COO in both aggregated mutational profiles and individual patients. I re-ran the COO models using random forest and extreme gradient boosting, instead of the initial multiple linear regression. Only the random forest model produced COO predictions for both aggregated and individual patients, while the extreme gradient boosting model failed to run successfully with the current parameters.

Using aggregated mutational profiles of the top 40% most mutated genes, the random forest COO model achieved a high explained variance of approximately 60% in identifying the correct COO for liver cancer (Figure 86A). However, predictions on individual patients yielded lower numbers of patients with correctly identified COO (360 compared to 332 for multiple linear regression and random forest, respectively; Figure 86B).

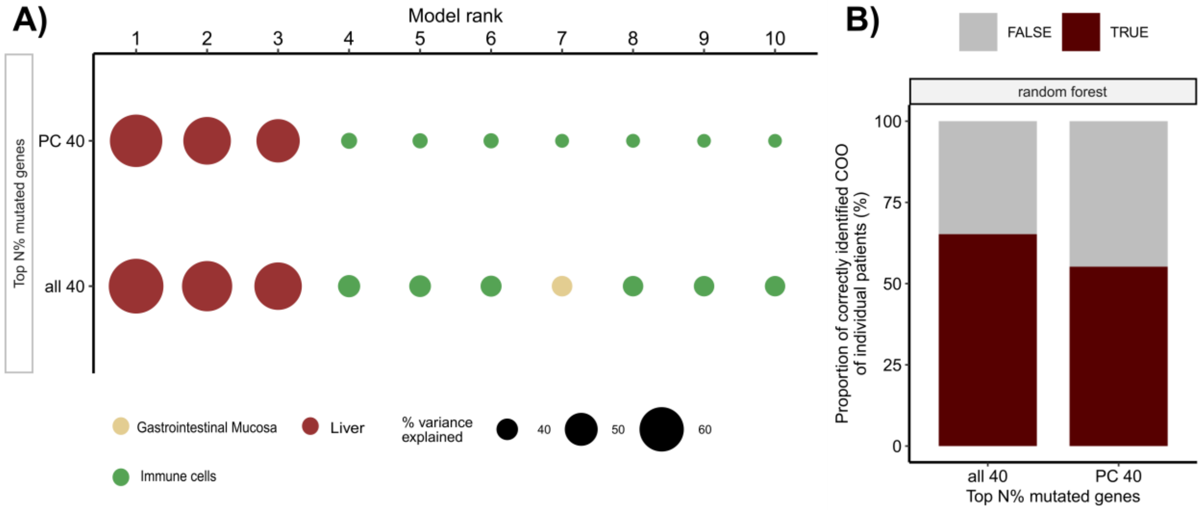


Figure 86. *A) Random forest models for the prediction of mutation density of aggregated cancer profiles in top 40% mutated genes of liver WGS were trained on an extended set of 101 tissue sets but showing only the top 10 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation. B) Proportion of correctly and incorrectly identified COO of individual patients using random forest model with top 40% mutated genes*

Moreover, I analyzed the top 40% protein-coding genes to characterize their association with liver cancer. Out of 745 genes, 6 were identified as over-predicted and 24 as under-predicted outliers based on their standardized residuals (Figure 87A). Only the under-predicted genes were affected by SV-hotspots and kataegis, whereas the over-predicted genes were not affected at all (Figure 87BC). However, although under-predicted genes were more frequently associated with SV-hotspots compared to non-outliers, this was not the case for kataegis enrichment.

Over-representation analysis (ORA) on databases of disease-gene associations with under-predicted genes did not yield significant terms related to liver cancer development after adjusting for multiple hypothesis testing. However, when searching for liver cancer-related diseases in the results, I found some terms that were present but did not reach statistical significance: DisGeNET term “C2239176 Liver carcinoma” with 9 genes (p-value = 0.69): *EPHA3*, *PREX2*, *SEMA3A*, *FNDC3B*, *NCOA2*, *CACNA2D1*, *MMP16*, *ZFPM2* and *BASP1*.

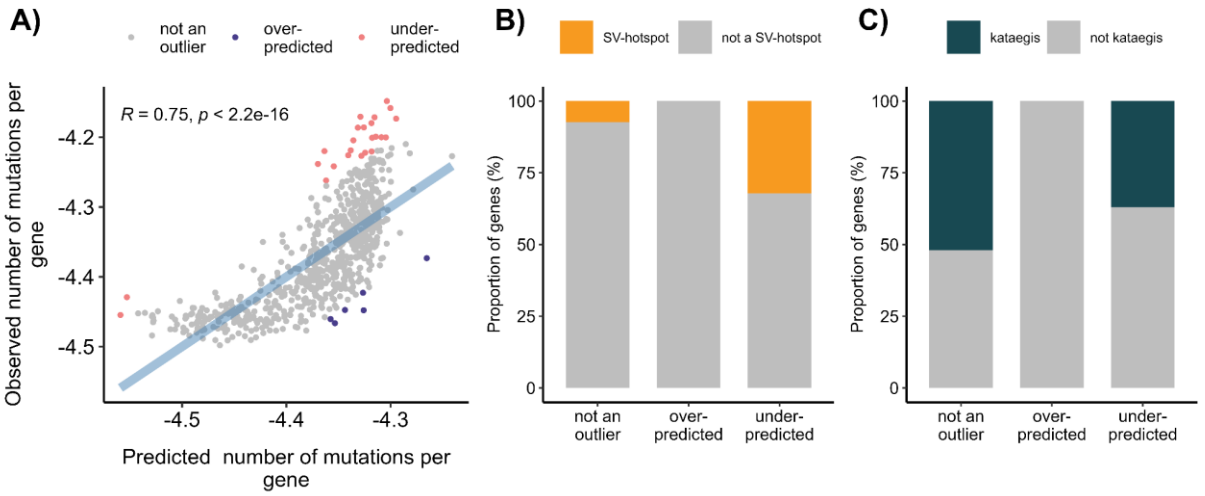


Figure 87. A) Spearman correlation of observed vs predicted normalized number of mutations per gene in top 40% mutated protein-coding genes in liver cancer B) Proportion of regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier C) Proportion of regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier

Afterwards, I examined the TADs regions where the outliers occurred. TADs with under-predicted genes had significantly lower stability scores than those with over-predicted genes, indicating that these TADs are less conserved across cell types (Figure 88A). Active chromatin states were only found in TADs with non-outlier liver genes, while outliers were more enriched in more closed, repressive and heterochromatin TADs (Figure 88B).

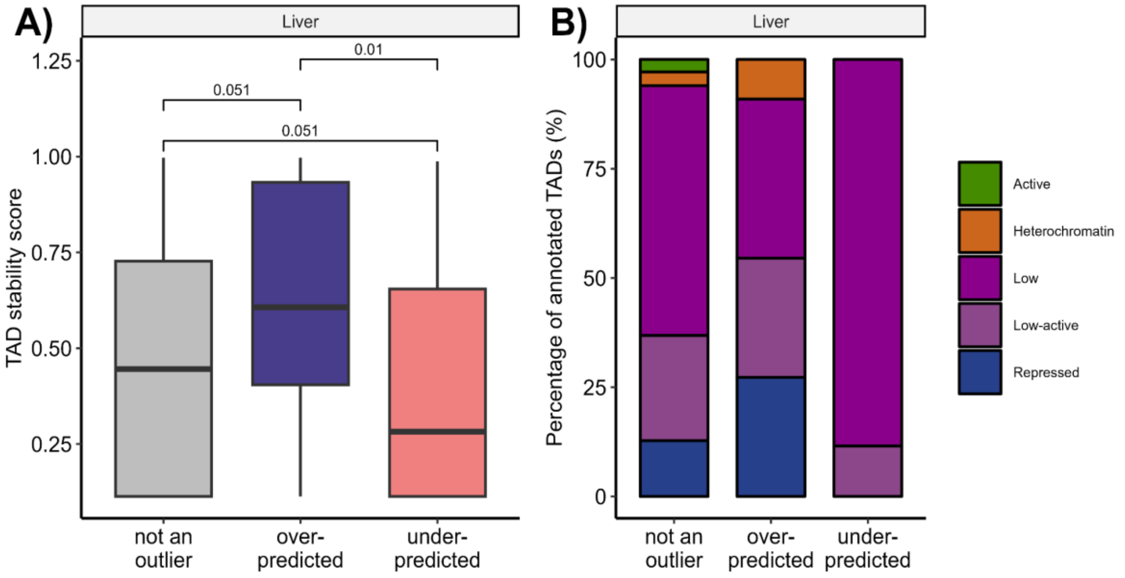


Figure 88. A) TADs stability score of TADs boundaries distribution in annotated erroneous regions B) TADs annotation by Akdemir et al. 2020. Based on active and inactive state (heterochromatin, low, low-active and repressed)

4.3.3.2.3.2 Patient characteristics based on the best gene COO model

When separating patients based on their correct or incorrect COO status as determined by the multiple linear regression COO model using the top 40% genes, I observed a significantly higher explained variance and number of mutations in correctly identified COO patients (Figure 89A). The number of kataegis per patient in this 40% gene set was similar for both correctly and incorrectly identified COO patients. Although statistical significance was not detected for SV-hotspots, correctly identified COO patients tended to have a higher number of SV-hotspots than incorrectly identified ones.

Examining the proportions of SBS signatures per patient in the top 40% protein-coding genes between the two groups of COO liver cancer patients, I found that incorrectly identified patients had a significantly higher proportion of SBS16 signature (Figure 89B). On the other hand, correctly identified COO patients had a significantly higher proportion of signature SBS12. All other signatures were quite similar between the two groups.

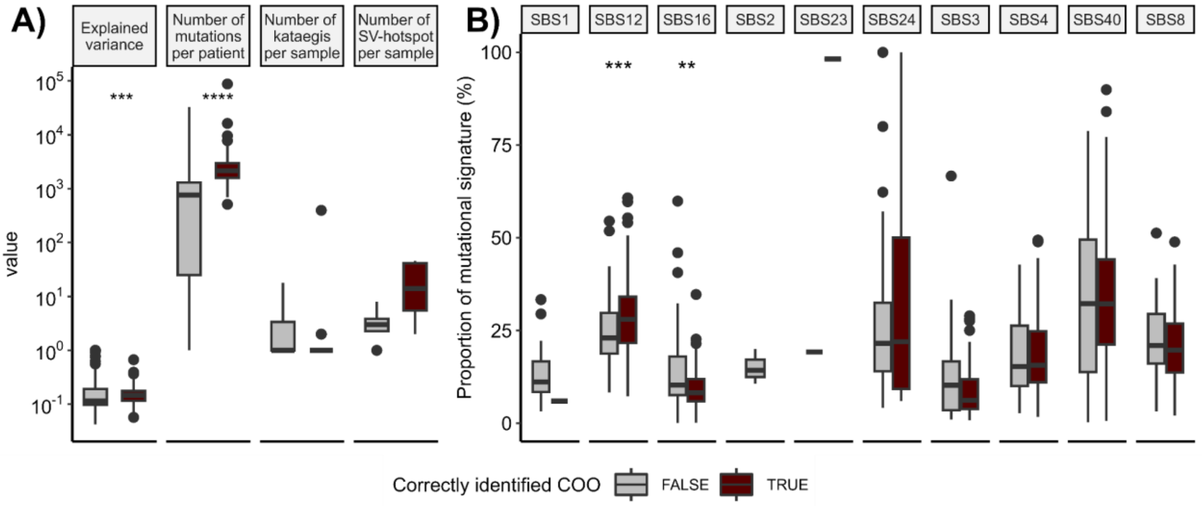


Figure 89. *A) The distribution of explained variance and the number of mutations separated by correct or incorrect COO identification of individual liver cancer patients. B) Proportion of SBS mutational signatures per liver cancer patient separated by correct or incorrect COO identification of individual patients. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

When patients were separated based on their histological types, I observed varying percentages of correctly identified COO across different types (Figure 90). Hepatocellular carcinoma, the most common subtype with 467 patients, showed a 65% accuracy in COO identification. In contrast, hepatocellular adenoma and fibrolamellar hepatocellular carcinoma, each with 5 patients, had lower proportions of 20% and 40%, respectively. Combined hepatocellular and cholangiocarcinoma, with 8 patients, has a 50% accuracy, while cholangiocarcinoma, with 24 patients, shows a 29% accuracy in COO identification. Only for hepatocellular carcinoma, correctly identified COO patients had significantly higher mutational count per patient and a higher explained variance compared to incorrectly identified patients. This trend was consistent across all subtypes, with correctly identified COO patients generally exhibiting higher mutation counts and explained variance, although the differences are more pronounced in some subtypes than others.

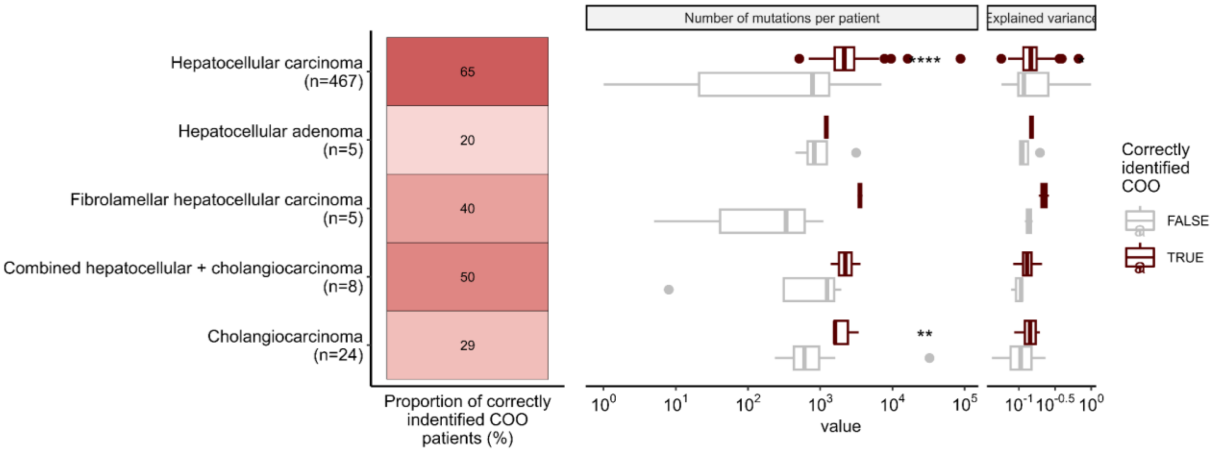


Figure 90. *A) Proportion of correctly identified cell-of-origin (COO) proportions using the top 40% mutated protein-coding genes, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.3.2.4 Skin melanoma COO gene-based predictive model

The gene-based multiple linear regression COO models, developed using various gene subgroups using aggregated skin melanoma mutational profiles, successfully identified the correct COO as melanocyte across nearly all defined gene subgroups as the top model with the highest explained variance (Figure 91). The highest variances were obtained using top N% mutated skin melanoma genes and driver genes. On the other hand, skin tissue-specific genes had the lowest variance explained out of all developed modes. Out of all defined skin tissue-specific genes, skin not exposed genes had the highest variance explained.

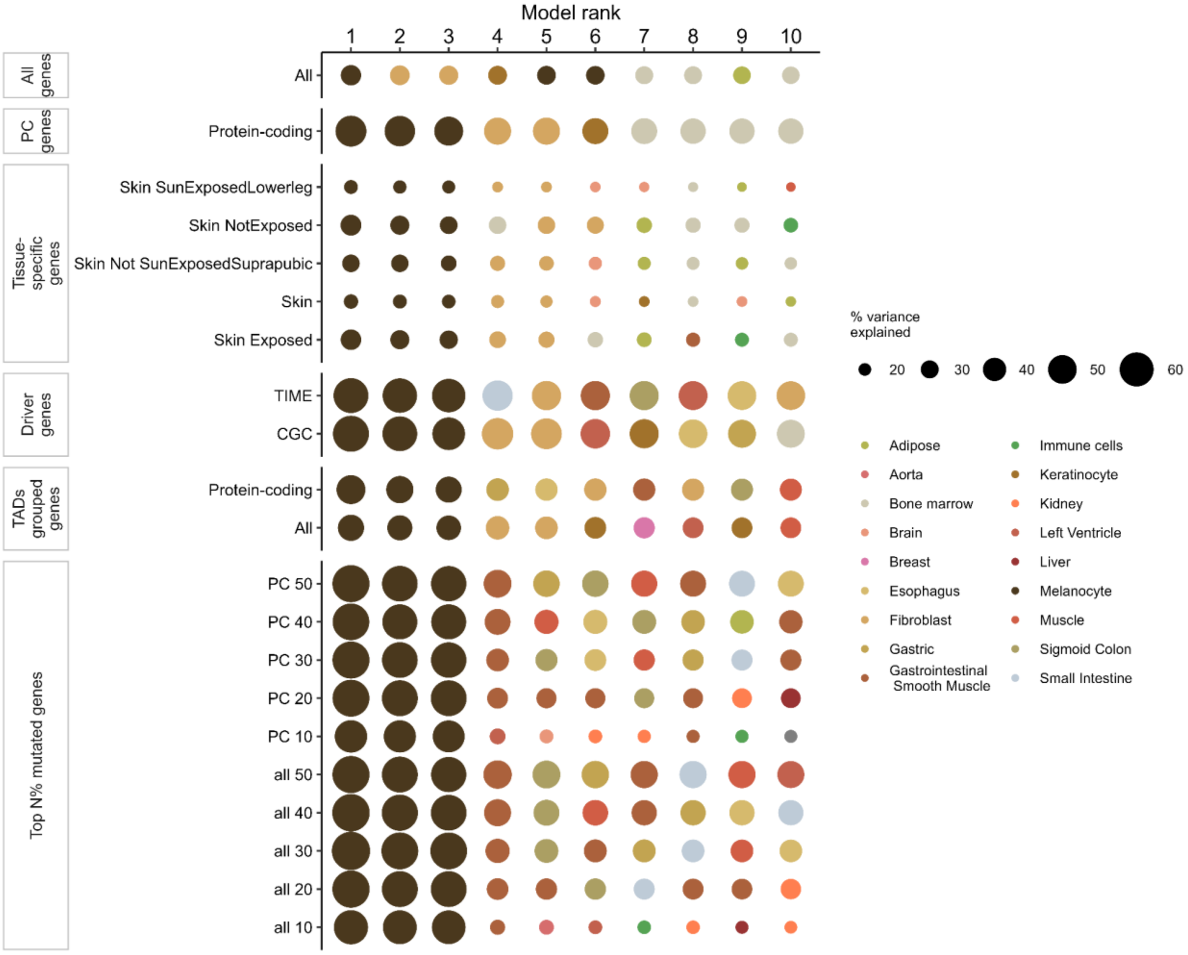


Figure 91. *Multiple linear regression models for the prediction of mutation density of aggregated tumor profiles of breast WGS were trained on an extended set of 101 tissue sets but showing only the top 10 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation.*

Prediction on individual skin melanoma patients across the majority of gene-based COO model setups using various gene subsets resulted in the majority of patients having incorrectly identified COO (Figure 92A). The only exception were the top N% mutated genes where we had more than 50% of patients correctly identified, the highest being in top 20% mutated genes. Besides the most mutated gene groups, TADs-grouped genes and both driver genes groups had close to ~20% patients with correct COO.

The majority of incorrectly identified COO patients had their COO assigned to brain-related tissues or immune cells (Figure 92B). Immune cells were predominantly assigned in models that included all genes and various skin tissue-specific gene groups. Driver genes lead to increase of gastrointestinal tissue as primary identified COO for skin melanoma patients. Only when I used the top N% mutated gene subsets for COO model, did those misclassifications reduce. Depending on the top N% mutated gene subset, I still detected different proportions of incorrectly identified COO patients mostly annotated as certain brain tissue.

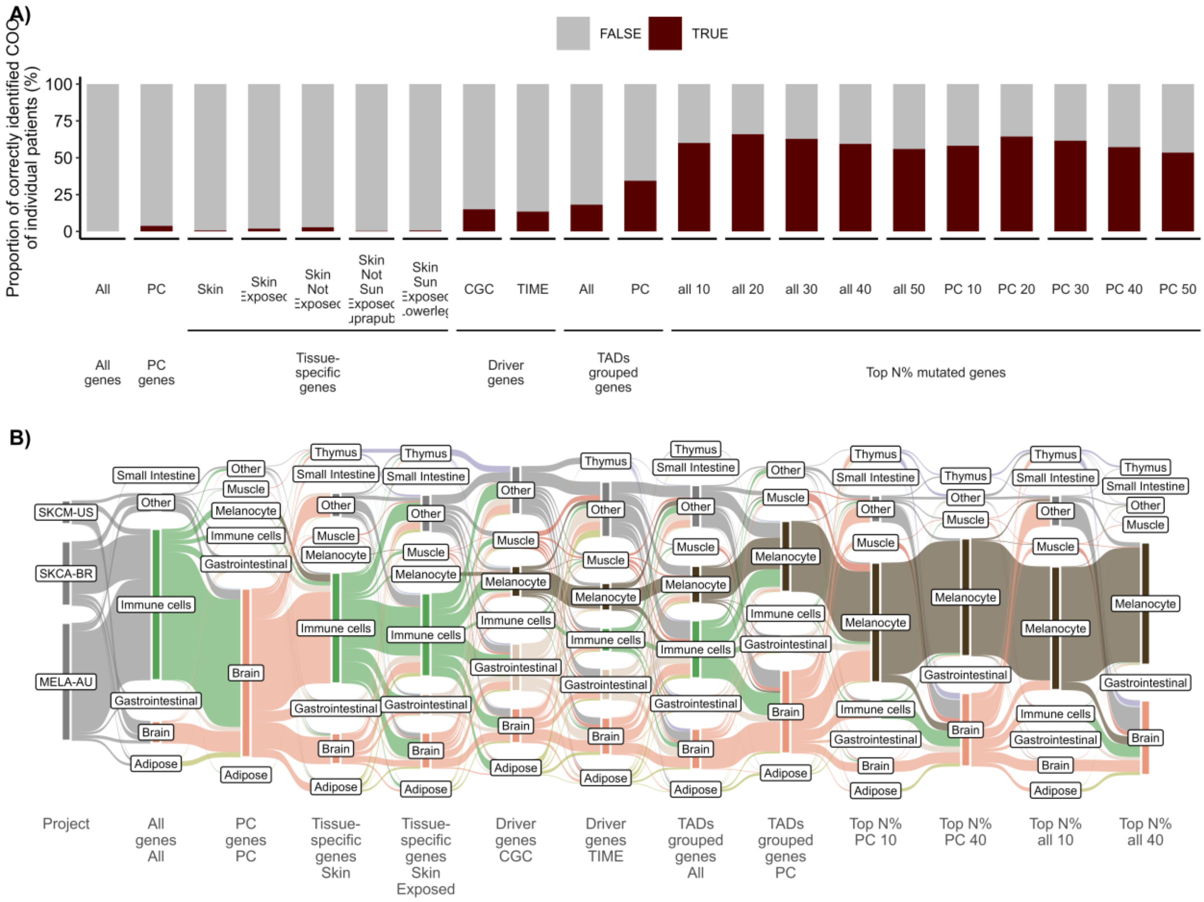


Figure 92. *A) The proportion of individual samples in which the prediction on an individual level matches the correct cell of origin of the highest explaining models in different modeling setups using various subsets of genes for skin melanoma. B) Alluvial plot illustrating the cell-of-origin (COO) of individual patients across various COO model setups based on different gene subsets*

Furthermore, I analyzed the normalized mutational count per patient in each specific gene group and found that the top N% mutated genes had the highest mutational counts and the longest genes (Figure 93AB). The skin-specific groups, both sun-exposed and non-exposed, had the lowest mutational counts per patient and among the shortest gene lengths. Specifically, the skin lower leg and suprapubic tissue groups had even shorter gene lengths but higher mutational counts compared to other genes. Interestingly, the TADs-grouped genes, despite their longer lengths, did not result in higher mutational counts than the other groups.

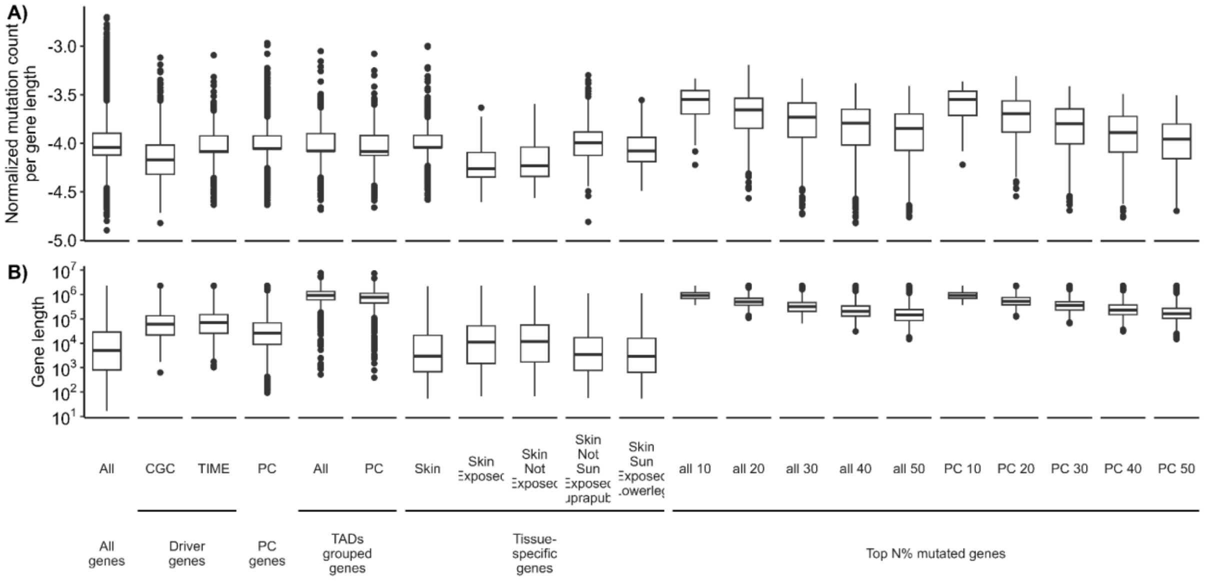


Figure 93. *A) Distribution of normalized mutation count per gene length of various groups of genes and B) gene lengths used for development of COO models in liver cancer. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*

Based on the correlation analysis of correctly identified COO tissues, the top N% mutated genes exhibit the strongest negative correlations with open chromatin marks (H3K4me1, H3K36me3, H3K4me3) and positive correlations with closed chromatin marks (H3K9me3, H3K27me3) (Figure 94). As the percentage of top mutated genes decreases (from top 50% to top 10%), the negative correlations with H3K4me1 and H3K36me3 become more pronounced. Conversely, the positive correlations with H3K9me3 and H3K27me3 vary, increasing or decreasing depending on the gene group.

In general, the positive correlations with open chromatin modifications were much weaker in the top N% mutated genes and TADs-grouped genes compared to driver genes or even tissue-specific genes. Skin tissue-specific gene groups exhibited higher and stronger positive correlations with closed chromatin modifications, surpassing those of other non-tissue-specific gene groups. However, these skin-specific groups also demonstrated the weakest negative correlations with open chromatin marks. This suggests a unique chromatin landscape in skin tissue-specific genes, characterized by stronger associations with repressive chromatin states and weaker associations with active chromatin states.

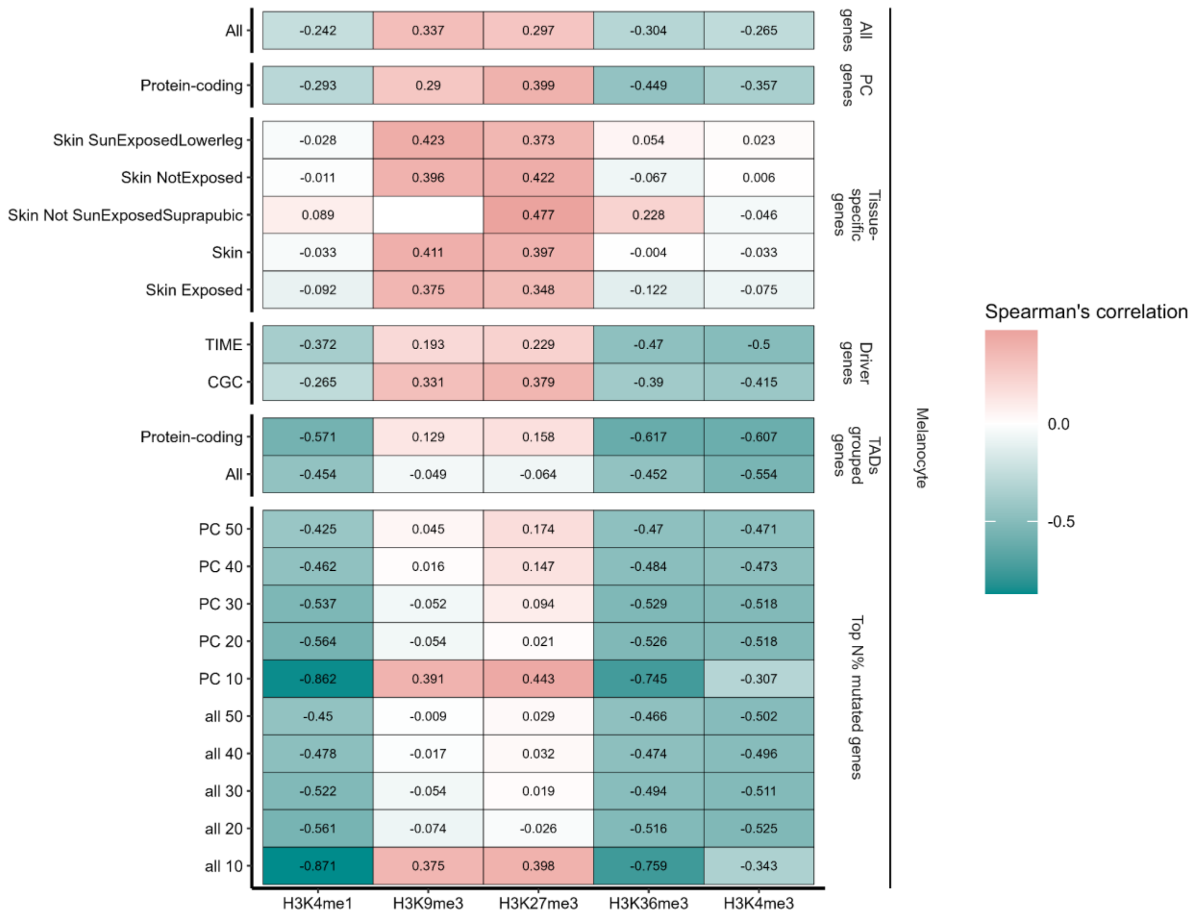


Figure 94. *Spearman's correlation of all mutations and histone modifications of top identified cell-of-origin tissue in various gene subgroups. White blank field represent not available histone marks in the dataset for specific epigenome of normal tissue*

The gene mutational signature landscape of skin melanoma was relatively consistent across all gene subgroups shown in Figure 95A. SBS7a and SBS7b overwhelmingly dominated the gene mutational landscape in all groups. All gene groups were significantly affected by kataegis, with nearly 100% of the top N% mutated genes having detected kataegis regions (Figure 95B). Similarly, SV-hotspots were more prevalent in the top N% mutated genes, though their maximum proportion was around 30%. Super-enhancers, regardless of the cell line, influenced only a small proportion of CGC, TIME, protein-coding, and TAD-grouped PC genes, and had little to no effect on the top N% mutated genes.

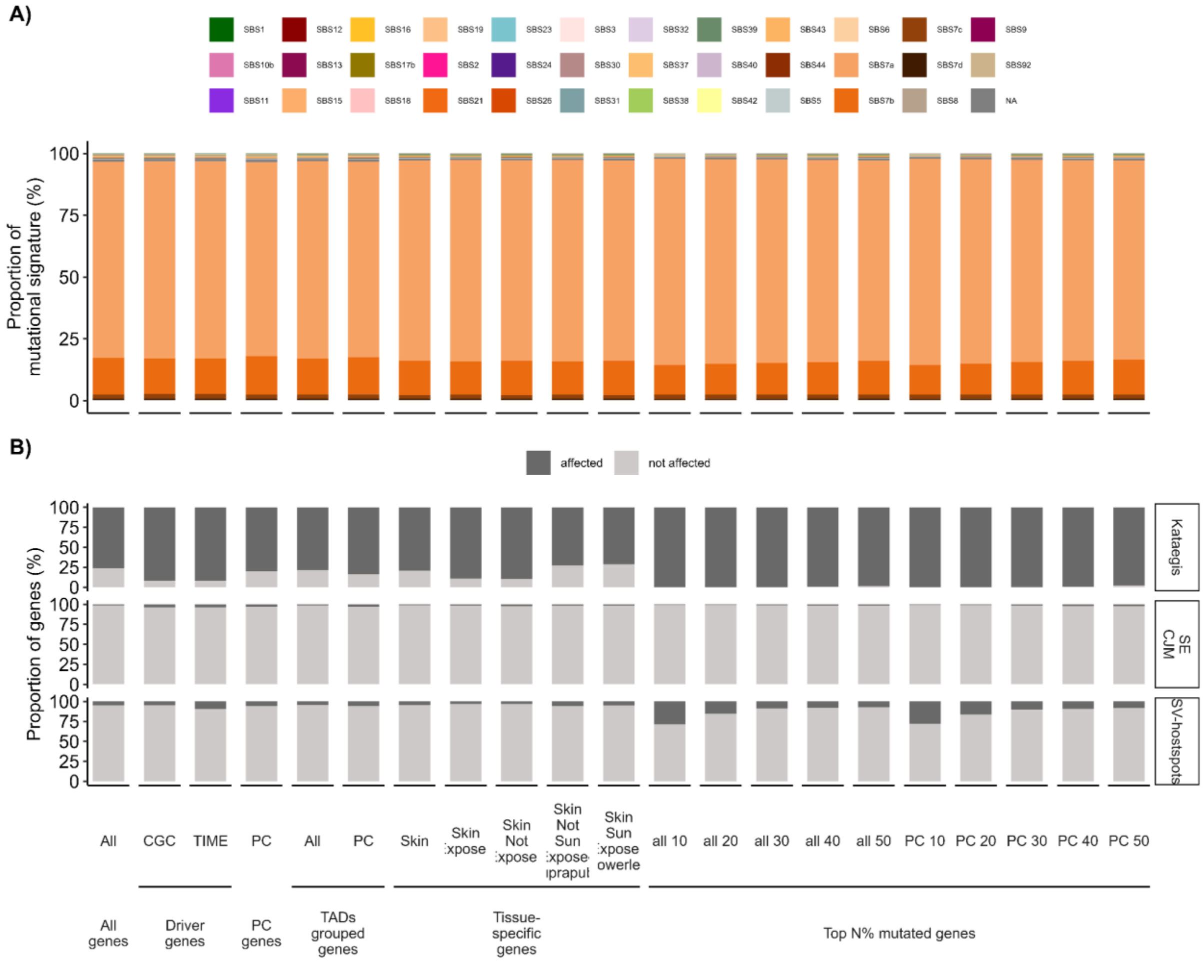


Figure 95. *A) Proportion of mutational signature in genes separated by various gene subgroups of skin melanoma cancer B) Proportion of affected genes by either kataegis, SV-hotspot or are proximal to tissue specific super-enhancers (SE) from breast-epithelium tissue*

4.3.3.2.4.1 Top 20% mutated skin melanoma genes as best gene model

Since the highest accuracy of the gene-based model was achieved using the top 20% mutated genes, I selected these genes for downstream analysis. More complex machine learning models, such as random forest and extreme gradient boosting, were applied using the top 20% mutated genes (Figure 96). However, these models did not result in better predictions for individual skin melanoma patients. The extreme gradient boosting model even failed to run completely. On the other hand, the random forest model showed similarly high explained variance on aggregated mutational profiles, but a lower number of correctly identified COOs for individual patients (198 for random forest, compared to 211 for multiple linear regression).

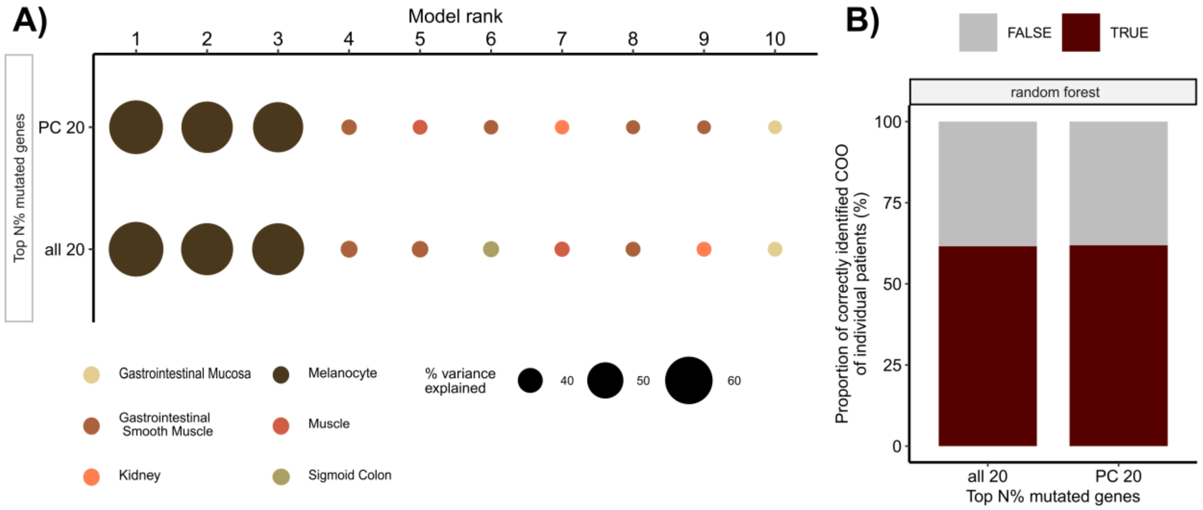


Figure 96. *A) Random forest models for the prediction of mutation density of aggregated tumor profiles in top 20% mutated genes of skin melanoma WGS were trained on an extended set of 101 tissue sets but showing only the top 10 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation. B) Proportion of correctly and incorrectly identified COO of individual patients using random forest model with top 20% mutated genes*

I focused my analysis on top 20% protein-coding genes to get a better understanding why these genes have a positive contribution to gene-based COO models in skin melanoma. Out of 502 genes, 5 were identified as over-predicted and 14 as under-predicted outliers based on their standardized residuals (Figure 97A). SV-hotspots were only found in some non-outlier genes (Figure 97B), while kataegis affected all genes regardless of their outlier status (Figure 97C).

Over-representation analysis (ORA) on databases of disease-gene associations with under-predicted genes did not yield significant terms related to liver cancer development after adjusting for multiple hypothesis testing. However, when searching for skin and melanoma cancer-related diseases in the results, I found some terms that were present but did not reach statistical significance: “C0025202 melanoma” with 5 genes (p-value=0.38); *HDAC9*, *GHR*, *CACNA1A*, *PDE1C* and *PRKCB*.

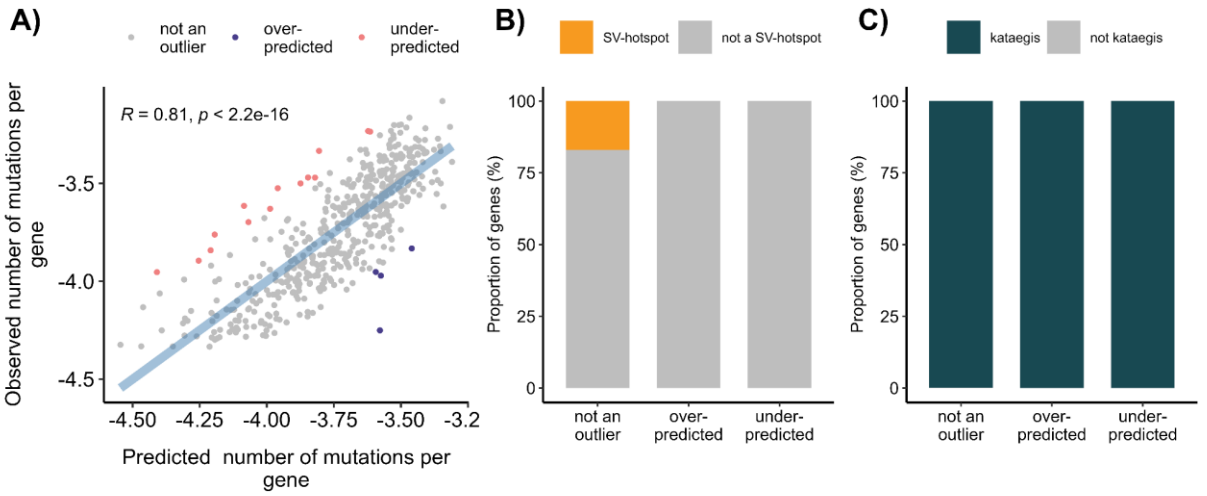


Figure 97. *A) Spearman correlation of observed vs predicted normalized number of mutations per gene in top 20% mutated protein-coding genes in skin melanoma B) Proportion of regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier C) Proportion of regions affected by SV-hotspot or not based on their annotation as over-, under-outliers or not an outlier*

Moreover, I examined in which TADs regions the outliers occurred. Over-predicted gene outliers were mostly found in TADs with lower TAD stability score despite the lacking statistical significance (Figure 98A). All annotated outliers had quite similar profiles chromatin states of TADs (Chi-square, p-value = 0.65) (Figure 98B).

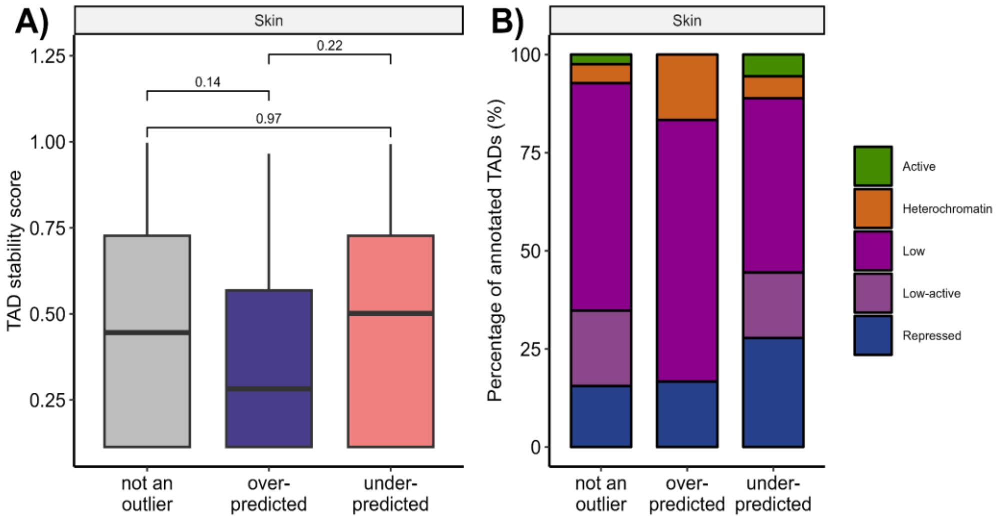


Figure 98. *A) TADs stability score of TADs boundaries distribution in annotated erroneous regions B) TADs annotation by Akdemir et al. 2020. Based on active and inactive state (heterochromatin, low, low-active and repressed) Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

4.3.3.2.4.2 Patient characteristics based on the best gene COO model

Correctly identified COO skin melanoma patients had a significantly higher mutational count in the top 40% mutated protein-coding genes used in the COO model, as well as a significantly higher explained variance compared to incorrectly identified patients (Figure 99A). While the number of kataegis regions was significantly enriched in correctly identified COO patients, there was no significant difference in the number of SV-hotspots per patient in the selected genes. Examining the gene mutational signature landscape, I observed significantly higher proportions of SBS7a, SBS7c, and SBS38 in correctly identified COO patients (Figure 99B). Conversely, SBS7b was significantly more enriched in incorrectly identified COO patients. Age-related signature SBS1 was distributed similarly between both groups of skin melanoma patients.

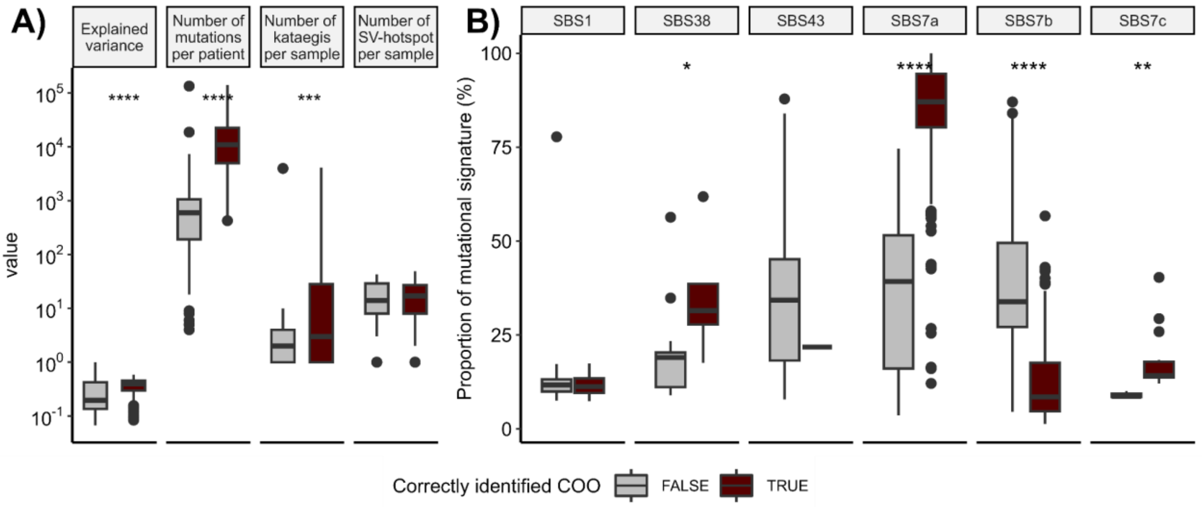


*Figure 99. A) The distribution of explained variance and the number of mutations separated by correct or incorrect COO identification of individual skin melanoma cancer patient. B) Proportion of SBS mutational signatures per skin melanoma cancer patient separated by correct or incorrect COO identification of individual patients. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Among all histological types of skin melanoma, mucosal lentiginous melanoma and malignant melanoma (NOS) had the lowest proportion of correctly identified COO patients, with 40% and 31%, respectively (Figure 100). In contrast, lentigo malignant melanoma and the unknown melanoma subtypes achieved the highest accuracy, with 100% and 90% of patients correctly identified, respectively. Superficial spreading melanoma and nodular melanoma also displayed high accuracy, with 84% and 91% of patients correctly identified. Correctly identified COO patients consistently exhibited a higher number of mutations per patient and a higher explained variance compared to incorrectly identified COO patients across almost all subtypes.

For acral lentiginous melanoma and malignant melanoma (NOS), despite the significantly higher mutational counts in correctly identified COO patients, there was no significant difference in explained variance between the correctly and incorrectly identified groups. Conversely, correctly identified COO patients in the mucosal lentiginous melanoma subtype exhibited both lower mutational counts and explained variance compared to their incorrectly identified counterparts.

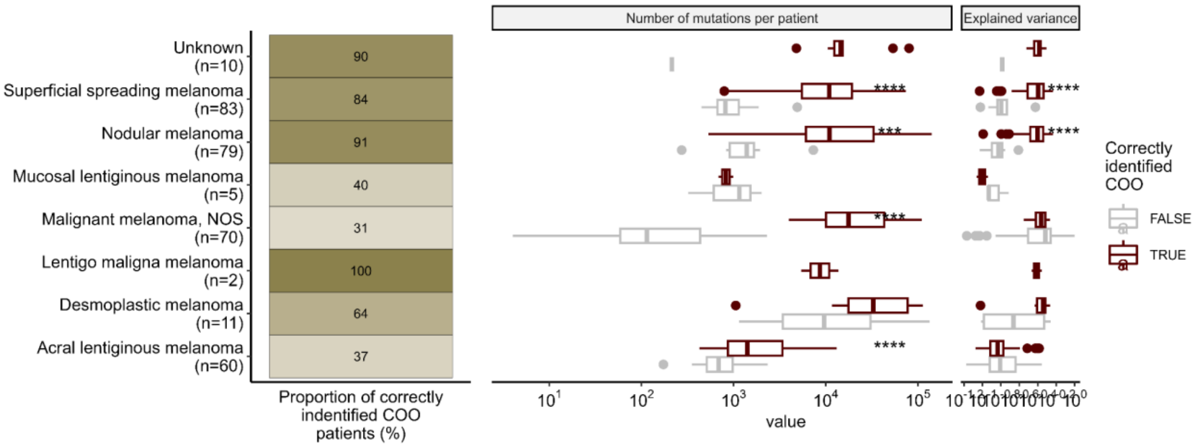


Figure 100. A) Proportion of correctly identified cell-of-origin (COO) proportions using the top 20% mutated protein-coding genes, separated by their histological subtype. The distribution of the number of mutations per patient is visualized as boxplots for each histological subtype. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001

4.4 RNA-seq single-nucleotide variants for prediction of the cell-of-origin

To evaluate whether mutations called from RNA-seq data can be integrated into the previously developed gene-based COO models using the gene mutational landscapes and epigenomic features of genes, I analyzed RNA-seq from breast, liver, and skin melanoma. This chapter provides a summary of the mutations called from RNA-seq data and presents gene-based COO models utilizing the same gene subsets as those in section 4.3.3 Gene-based cell-of-origin predictive models.

4.4.1 Mutational landscape obtained by RNA-seq data

After applying multiple filtering steps to RNA-seq called single-nucleotide variants, the median and standard deviation of RNA-seq detected mutations were 309 ± 616 for breast cancer, 390 ± 135 for liver cancer, and 31,691 ± 9,771 for skin melanoma. The number of mutations detected from RNA-seq data was consistently higher across all cancer types compared to WXS, and lower than WGS (Figure 101A). In breast and liver cancers, where normal matching tissue was available and multiple variant-calling tools were utilized, the mutation profiles from RNA-seq data were more closely aligned with those obtained from WXS than with those from WGS. For skin melanoma, the distribution of mutations identified by RNA-seq, which were called exclusively by Mutect2, was similar to that of WGS data (Wilcoxon test, p-value = 0.53).

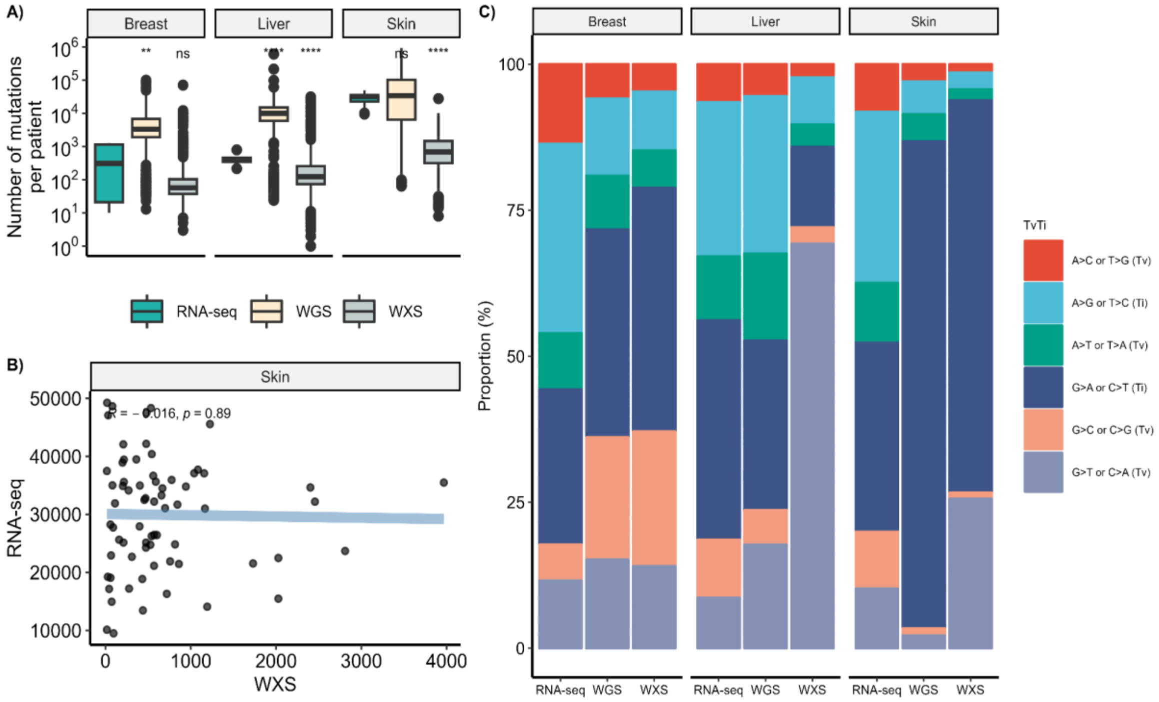


Figure 101. *A) Distribution of single-nucleotide variant (SNV) per patient across breast, liver and skin melanoma obtained by RNA-seq, whole-genome (WGS) or whole-exome sequencing (WXS). B) Pearson correlation analysis of the number of mutations identified via RNA-seq and WXS in the same 69 skin melanoma patients from the SKCM-US cohort. C) Proportion of various types of transversions (Tv) and transitions (Ti) in aggregated mutational profiles for RNA-seq, WGS, and WXS across breast, liver, and skin melanoma cancers. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test to the reference group of RNA-seq, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*

Moreover, the correlation between RNA-seq and WXS mutation counts with same skin melanoma patients from the SKCM-US cohort was found to be insignificant and slightly negative (R = -0.016, p-value = 0.89) (Figure 101B). The aggregated mutational landscape, defined by transversions (Tv) and transitions (Ti), revealed that RNA-seq mutations were predominantly enriched with A>G or T>C and G>A or C>T transitions (Figure 101C). Furthermore, the RNA-seq aggregated mutational profiles from all three cancer types were more similar to each other (Chi-square, p-value=$1.21*10^{-53}$) than to other sequencing technologies within their respective cancer types, where the p-value was below $2.2*10^{-308}$. In breast cancer, the mutational profiles of WGS and WXS were more similar to each other than those of RNA-seq. In liver cancer, the RNA-seq profile was much more similar to WGS, while WXS showed a distinctly unique pattern enriched with G>T transversions. As for skin melanoma, the WGS and WXS show much more abundant enrichment with G>A or C>T transitions than RNA-seq mutations.

4.4.2 Gene mutational landscape obtained by RNA-seq data

I analyzed the number and distribution of various gene types affected by mutations detected through RNA-seq compared to WGS and WXS in the following gene categories: immunoglobulin genes, non-coding genes, pseudogenes, mitochondrial genes, protein-coding genes, and ribosomal RNA (rRNA) genes (Figure 102A). Across breast, liver, and skin melanoma, the majority of mutations identified by RNA-seq, WGS, and WXS were located within protein-coding genes, which accounted for approximately 40% to 80% of the mutations. The most notable differences were a higher proportion of non-coding genes and pseudogenes in WGS than RNA-seq or WXS.

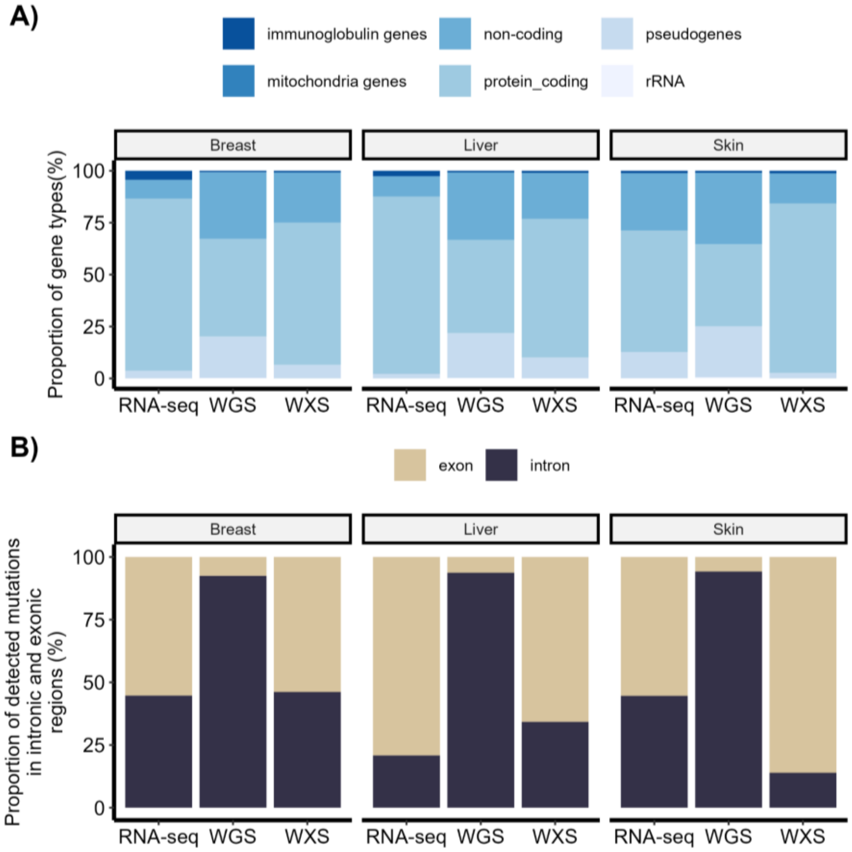


Figure 102. *A) Proportion of various gene types affected by SNVs identified through RNA-seq, WGS and WXS across breast, liver and skin melanoma cancers. The gene types include protein-coding genes, non-coding genes, pseudogenes, ribosomal RNA (rRNA) genes, immunoglobulin genes, and mitochondrial genes. B) Proportion of detected mutations found in exonic and intronic regions for each sequencing method (RNA-seq, WGS, WXS) across the same cancer types.*

Additionally, RNA-seq detected a significantly higher proportion of mutations in immunoglobulin genes, particularly in breast and liver cancers, where approximately 4% and 2% of the mutations, respectively, were found in these genes. In contrast, skin melanoma had around 1% of its mutations in immunoglobulin genes. Subsequently, I examined the enrichment of SNVs in exonic and intronic regions as detected by RNA-seq in comparison to WGS and WXS (Figure 102B). Across all three cancer types, RNA-seq and WXS exhibited a similar enrichment of SNVs in exonic regions, with over 55% of the mutations being localized in exons. Difference in enrichment was not detected for breast cancer RNA-seq and WXS mutations (Chi-square test p-value = 0.11), while other comparisons of RNA-seq and WXS in other cancers showed significant differences (p-values <= 0.05). In contrast, WGS detected a higher proportion of SNVs in intronic regions across all cancer types.

Given the limited number of patients in the liver and breast RNA-seq cohorts, I focused on identifying the top 5 frequently mutated protein-coding genes that were mutated in more than 50% of the samples (Table 11). None of these genes were found in the Cancer Gene Consensus or TIME gene lists. I found that only the *UTRN* gene, which was frequently mutated in breast cancer as detected by RNA-seq, had also been previously identified as highly mutated in WGS breast cancer datasets. Additionally, only one liver gene, *CYP2E1*, was found to be influenced by super enhancers in hepatocyte cell lines.

Table 11. *Most frequently mutated protein-coding genes detected by RNA-seq in 13 liver and 5 breast cancer patients.*

Cancer	Gene name	Ensembl gene	Percentage of patients with mutated gene	WGS most frequently mutated group	Super-enhancer affected tissue
Breast	*CAPZB*	ENSG00000077549	0.6	x	x
	SDC3	ENSG00000162512	0.6	x	x
	RNF115	ENSG00000121848	0.6	x	x
	DPYSL3	ENSG00000113657	0.6	x	x
	UTRN	ENSG00000152818	0.6	40%	x
	PGM5	ENSG00000154330	0.6	x	x
Liver	*TMBIM6*	ENSG00000139644	0.6923076923	x	x
	ALB	ENSG00000163631	0.6153846154	x	x
	SPRN	ENSG00000203772	0.5384615385	x	x
	CYP2E1	ENSG00000130649	0.5384615385	x	hepatocytes

On the other hand, in skin melanoma I detected a substantial number of top N% mutated genes which I overlapped with the ones identified with WGS. First of all, the number of mutated genes increases substantially with the percentage of the top mutated genes considered (Figure 103A). For instance, when considering the top 5% of mutated genes, 575 genes are identified. This number increases incrementally, with 9779 genes being identified at the 50% threshold. Secondly, when examining the overlap of top 10% mutated genes from both RNA-seq and WGS, I noticed that only 4 genes were overlapping, while the majority of the rest were unique to RNA-seq (Figure 103B). When I performed ORA of GO, I found significantly enriched terms related to in utero embryonic development, regulation of embryonic development, and response to oxygen levels (Figure 103C). These processes indicate that the top mutated genes are involved in critical developmental and regulatory functions. Additional processes such as cellular response to peptide hormone stimulus, lipid import into cells, and RNA localization suggest diverse functional roles for these genes.

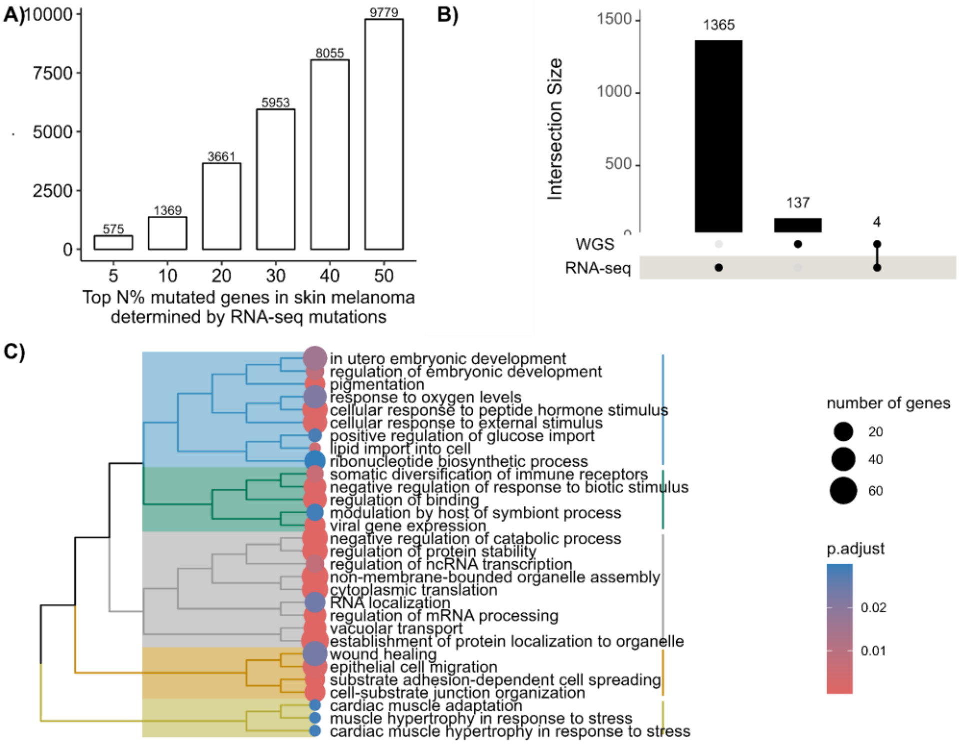


Figure 103. *Analysis of the top N% mutated genes in skin melanoma, determined by RNA-seq mutations. A) Bar plot depicting the number of top N% mutated genes in skin melanoma identified through RNA-seq data. The x-axis represents the top N% of mutated genes, where N ranges from 5% to 50%. The y-axis shows the corresponding number of genes within each percentage category. B) UpsetR plot showing the overlap sizes of top 10% mutated genes between RNA-seq and whole-genome sequencing (WGS). C) Hierarchical clustering of enriched biological processes associated with the top N% mutated genes identified through RNA-seq in skin melanoma. Each row represents a biological process from Gene Ontology (GO) terms, and processes are color-coded according to their statistical significance (p.adjust values), where darker blue shades represent more significant enrichments. The size of the circles corresponds to the number of genes associated with each process.*

4.4.3 Gene-based COO predictive models using RNA-seq SNVs

Despite encountering a substantial number of false positive mutations in the RNA-seq datasets, particularly within the skin melanoma cohort, I proceeded to develop the cell-of-origin models using the same methodologies applied to WGS and WXS gene-based COO models. Figure 104 shows the performance of these models across various gene groups, including all genes, protein-coding genes, driver genes, and the top 40% and 50% most frequently mutated protein-coding genes, based on aggregated mutational profiles from RNA-seq data.

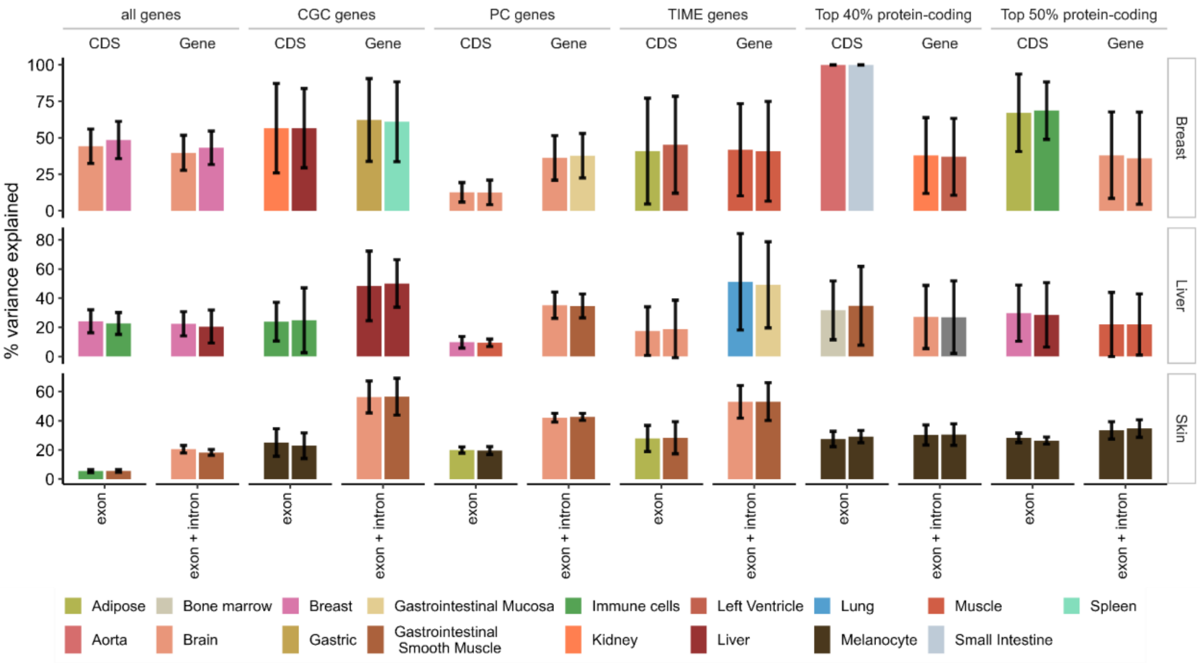


Figure 104. *Multiple linear regression models for the prediction of mutation density of aggregated cancer profiles of breast, liver and skin cancer RNA-seq called single-nucleotide variants (SNVs) were trained on an extended set of 101 tissue sets but showing only the top 2 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation. Coding-sequence (CDS) settings represent the count of mutations per exon normalized length by total CDS length. Gene setting implies summing mutations per exon or exon plus intron normalized by total gene length. The overall explained variance is reported across the 10-fold cross-validation.*

In breast cancer, the models using all genes were able to correctly identify the COO, irrespective of whether genes were normalized by coding sequence (CDS) length or overall gene length. The explained variance was approximately 43% for gene-length normalization and 48% for CDS normalization. However, when compared to the next best tissue, which was not the correct COO for breast cancer, there was no significant difference in the model's explanatory power (Wilcoxon test, p-value = 0.94). Other gene groups, including driver genes and the top percentage of mutated protein-coding genes, did not perform well in predicting the COO for breast cancer.

In the case of liver cancer, the most accurate and correct COO identification was achieved using the CGC genes normalized by overall gene length, with a high explained variance of approximately 50%. Additionally, the second best-performing tissue in the model was also another type of liver tissue, indicating a strong tissue-specific signal for liver cancer in the CGC gene group.

Contrasting with breast and liver cancers, the skin melanoma models exhibited a different pattern. A significant number of gene groups, including the top 40% and 50% most frequently mutated protein-coding genes, were able to identify the correct COO. Despite this, the explained variance for these models was generally below 30%, indicating that while these models can identify the correct COO, the level of confidence or strength of the prediction is relatively low.

Furthermore, when I used tissue-specific genes I was able to predict the correct COO using models based on aggregated mutational profiles in breast and skin melanoma cancers (Figure 105). In breast cancer, models developed using breast-specific genes correctly identified the tissue of origin with high explained variance, approximately 64% when normalized by gene length and 57% when normalized by CDS length. However, the models based on breast mammary tissue-specific genes failed to correctly predict the COO. In liver cancer, models using liver-specific genes did not successfully predict the correct COO. For skin melanoma, the use of skin-specific genes resulted in accurate COO prediction with models normalized by CDS length across all groups of skin tissues. However, gene-based normalization for skin-specific genes failed to predict the correct COO. The explained variance for skin cancer models was consistently below 20% across all setups. Furthermore, there were no significant differences between the explained variance of the best COO model and the next non-COO model result across all tissue-specific gene groups for each cancer type, as determined by the Wilcoxon test (p-value = 0.912).

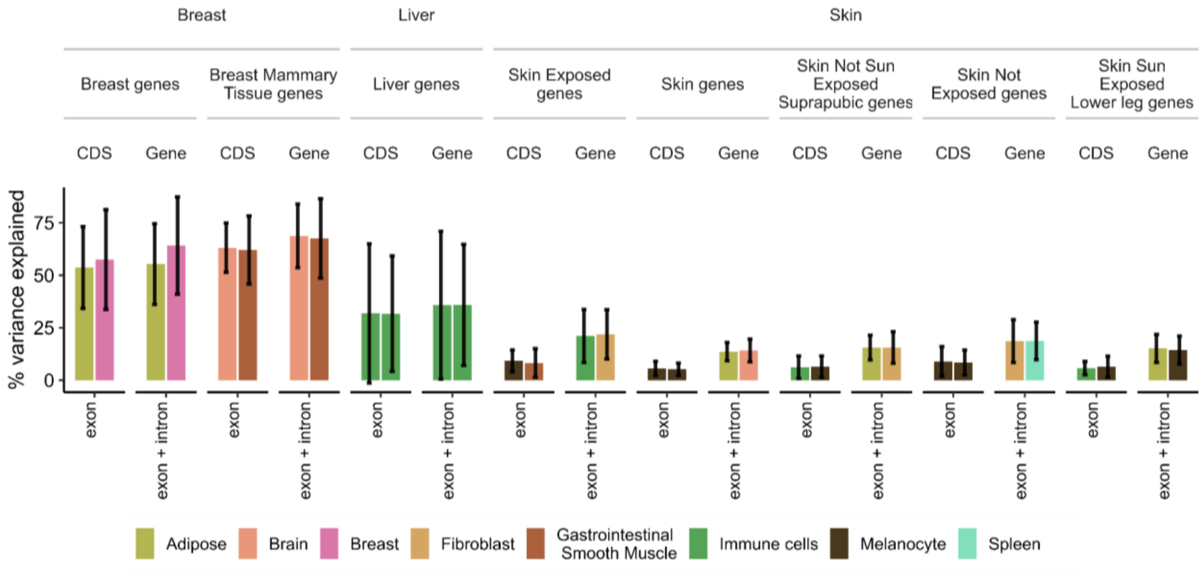


Figure 105. *Multiple linear regression models for the prediction of mutation density of aggregated cancer profiles of breast, liver and skin cancer RNA-seq called single-nucleotide variants (SNVs) were trained on an extended set of 101 tissue sets but showing only the top 2 in tissue-specific gene groups. The overall explained variance is reported across the 10-fold cross-validation. Coding-sequence (CDS) settings represent the count of mutations per exon normalized length by total CDS length. Gene setting implies summing mutations per exon or exon plus intron normalized by total gene length. The overall explained variance is reported across the 10-fold cross-validation.*

Unfortunately, the COO prediction rates on individual patients were extremely low, where for breast and liver mostly one or two samples had the correct identified COO using all or tissue-specific genes, while skin melanoma RNA-seq had more patients but their overall proportion in the analysed cohort was extremely small (Figure 106).

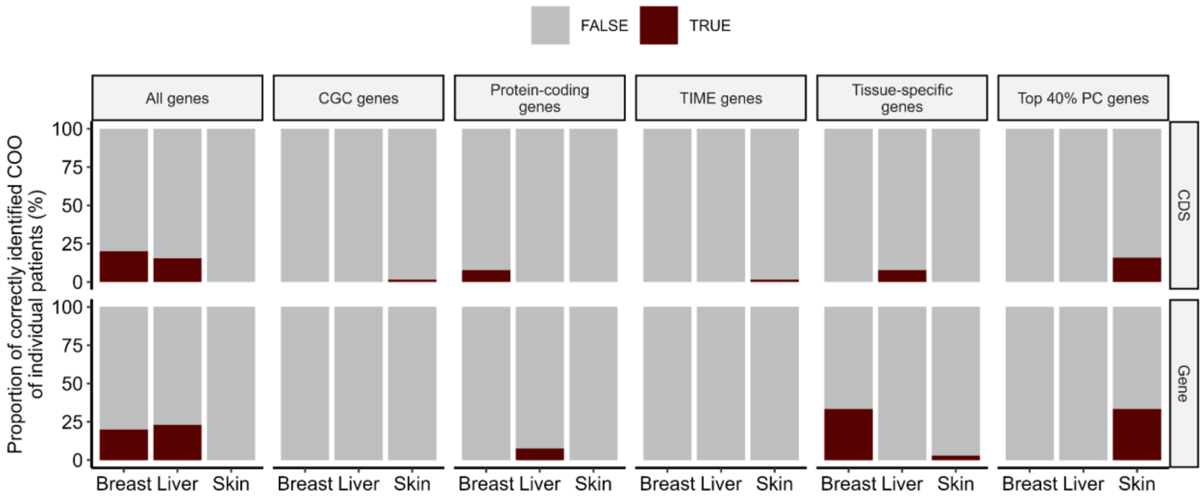


Figure 106. *Proportion of correctly and incorrectly identified COO of individual patients using multiple linear regression model with top 40% mutated genes of breast, liver and skin cancer RNA-seq called single-nucleotide variants (SNVs). Coding-sequence (CDS) settings represent the count of mutations per exon normalized length by total CDS length.*

5 Discussion

The ongoing endeavor to outsmart cancer by ensuring early detection and effective treatment remains a dynamic and relentless challenge for scientists. To get a better understanding of carcinogenesis mechanisms, international consortia like the International Cancer Genome Consortium (ICGC), The Cancer Genome Atlas (TCGA), and the Pan-Cancer Analysis of Whole Genomes (PCAWG) have collected extensive genomic data from diverse cancer patients using next-generation sequencing (NGS). Recent initiatives, such as ICGC-ARGO, are expanding this work by including over 10,000 cancer cases with detailed clinical information, aiming to discover novel biomarkers and therapeutic targets. Identifying the cell of origin (COO) remains a significant challenge, as it influences treatment strategies and prognosis. Consequently, large datasets from these consortia are being used to develop machine learning tools to predict COO, though their performance varies across cancer types due to differing mutational landscapes (Kübler et al., 2019; Liu et al., 2020; Nguyen et al., 2022; Polak et al., 2015). However, these predictive tools vary in their performance across different cancer types, a discrepancy that may be attributed to the distinct mutational landscapes present in various cancers.

So, I began my research by analyzing the genomic features of breast, liver and skin melanoma cancers from these cohorts prior to developing and improving the COO models based on the same principle as Polak et al. (2015).

5.1 Mutational landscape in breast, liver and skin melanoma cancers

Out of all the analysed cancers, skin melanoma had the highest mutational burden of single-nucleotides (SNVs) and indels in general. Skin melanoma is widely considered to be one of the most mutated cancer, especially compared to breast and liver cancers, primarily due to its significant exposure to ultraviolet (UV) radiation. UV radiation from sunlight induces direct DNA damage, leading to the formation of complex lesions cyclobutane pyrimidine dimers and 6-4 photoproducts, which leave highly abundant C>T transitions associated with skin cancers (Alexandrov et al., 2020; Brash et al., 1991). Specifically, these C>T transitions often occur at the TCN and CCN sequence contexts, generating distinct mutational signatures such as SBS7a and SBS7b (Alexandrov et al., 2013, 2020). In addition to SBS7a and SBS7b, the detected less abundant UV-related mutational signatures, SBS7c and SBS7d, arise from rarer types of DNA damage. SBS7c is characterized by thymine to adenine (T>A) transversions at NTT trinucleotide contexts, resulting from the misincorporation of adenine opposite to thymine residues in cyclobutane pyrimidine dimers. SBS7d, on the other hand, involves thymine to cytosine (T>C) transitions at NTT contexts, likely due to the erroneous insertion of guanine or cytosine during DNA repair processes (Alexandrov et al., 2020; Brash, 2015; Tate et al., 2019). Furthermore, due to the highly mutated landscape caused by extensive UV radiation exposure, skin melanomas are significantly enriched with kataegis regions. This abundance of mutations at smaller scales, as defined by kataegis, contributes to a higher detected number of these events. In a comprehensive pan-cancer analysis of 38 cancer types, kataegis events were detected in 60.5% of the analyzed cancers, with notably high frequencies in lung squamous cell carcinoma, bladder cancer, acral melanoma, and sarcomas (Aaltonen et al., 2020). In skin melanoma, kataegis regions are predominantly enriched with SBS7a and SBS7b signatures, which are the primary UV-induced mutational signatures that are driving tumorigenesis in this cancer type. UV-driven mutations include not only single nucleotide variants but also insertions and deletions (indels) that accumulate as a result of DNA damage induced by UV radiation. Both cyclobutane pyrimidine dimers and 6-4 photoproducts can lead to double-strand breaks, prompting extensive DNA repair processes that contribute to the accumulation of deletions. One such repair mechanism is non-homologous end joining (NHEJ), an error-prone process that often results in the accumulation of deletions within the genome. This is represented by the highly abundant indel signature in skin melanoma ID8 which is characterized by deletions of ≥5 base pairs. This signature is associated with the repair of DNA double-strand breaks by NHEJ and is correlated with age at diagnosis, reflecting the cumulative effects of UV exposure over time (Alexandrov et al., 2013; Tate et al., 2019). Additionally, another highly prevalent UV-associated indel signature in skin melanoma, ID13, has been identified in cancers from sun-exposed skin areas, further emphasizing the significant impact of UV radiation on the genomic landscape of melanoma (Tate et al., 2019).

In contrast to skin melanoma, where the mutational landscape is heavily influenced by exogenous factors such as UV radiation, breast cancer is primarily driven by endogenous factors. The most prominent of these is DNA damage caused by the activity of the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like (APOBEC) enzymes. This enzyme family contributes significantly to the mutational profile of breast cancer, as evidenced by the presence of mutational signatures SBS13 and SBS2 (Petljak & Alexandrov, 2016). APOBEC-associated mutations are not limited to only breast cancer; they are a common mutational process observed in approximately 75% of cancer types and more than 50% of all cancers analyzed (Alexandrov et al., 2020). In my analysis, I confirmed that the majority of kataegis foci in breast cancer are indeed generated by APOBEC activity. These regions of hypermutation are characterized by clusters of cytosine to thymine (C>T) transitions and cytosine to guanine (C>G) transversions, reflecting the mutagenic action of APOBEC enzymes (Alexandrov et al., 2020; Petljak & Alexandrov, 2016; Taylor et al., 2013). Additionally, another highly abundant single base substitution (SBS) signature found in both breast and liver cancers, and frequently associated with kataegis regions, is SBS40. The etiology of SBS40 remains unknown, although there is evidence suggesting a correlation between the frequency of these mutations and the age of patients in certain tumor types (Tate et al., 2019). SBS40 is known as one of the flat mutational signatures because it lacks a strong context-specific pattern, meaning that the mutations it comprises are relatively evenly distributed across different trinucleotide contexts, unlike APOBEC or UV-induced mutational signatures. This signature is observed across multiple cancer types and bears a resemblance to SBS5, further complicating the identification of its etiology (Alexandrov et al., 2020). Moreover, flat signatures like SBS5 and SBS40 present significant challenges for accurate extraction and quantification using various mutation-calling software, due to their lack of distinct pattern (Wu et al., 2022). Therefore, any results regarding the impact of SBS40 on the accuracy of cell-of-origin models should be interpreted with caution, acknowledging the inherent difficulties in detecting and analyzing such flat signatures. Similar observation also applies to the flat SBS39 signature with high abundance in kataegis regions and unknown aetiology. Other highly abundant signatures in breast cancer are also flat signature SBS3, associated with homologous recombination deficiency in cancers, as well as SBS8 with unknown aetiology.

Unlike skin melanoma and breast cancer, liver cancer exhibits a more complex and varied mutational landscape. While the highly abundant endogenous SBS40 signature is prevalent in liver cancer, other significant contributions to the mutational profile include SBS12, SBS16, SBS23, SBS24, SBS8, and SBS93. The etiologies of several of these signatures; SBS40, SBS8, SBS23, SBS16, and SBS93, remain largely unknown, raising questions about their origins. It is unclear whether these signatures are indeed intrinsic to liver cancer or if they represent artifacts introduced by the mutational signature calling tools used, potentially contaminated by other flat signatures such as SBS5, which is known for its broad presence across various cancers. In contrast, SBS12 and SBS24 are considered to be more liver-specific signatures. SBS12, despite its unknown etiology, is observed at relatively low frequencies in liver cancer patients and may be indicative of liver-specific mutational processes (Tate et al., 2019). SBS24, on the other hand, is directly associated with exposure to aflatoxin, a potent hepatocarcinogen, which was also observed in experimental system exposed to it (Alexandrov et al., 2020; Tate et al., 2019).

Breast and liver cancers exhibit a more similar indel mutational landscape compared to skin melanoma, which is predominantly influenced by UV radiation. In both breast and liver cancers, substantial amounts of indel signatures ID1, ID2, and ID8 were detected. ID1 and ID2 are common across a wide range of cancer types and are also found in normal cells, suggesting they represent background mutational processes that accumulate over time often called clock-like signatures (Tate et al., 2019). ID8, on the other hand, is associated with the error-prone non-homologous end joining (NHEJ) repair pathway, which is often employed as an alternative repair mechanism in cancers with deficient homologous recombination repair, such as those with *BRCA* mutations (Davies et al., 2017; Tate et al., 2019). Liver cancer had the highest proportion of ID5, another clock-like signature, and ID3, which is strongly associated with exposure to tobacco smoking (Tate et al., 2019).

These significant findings were primarily detected using whole-genome sequencing data, which offers a superior ability to detect a broader spectrum of mutations, providing a more comprehensive view on the mutagenic processes across all tumors. WGS captures mutations throughout the entire genome, whereas whole-exome sequencing focuses only on the coding regions of genes. As a result, the number of mutations detected in WXS datasets is significantly lower than those detected by WGS, which is expected due to the more limited genomic coverage of WXS. Interestingly, WGS has also proven to be more effective than WXS in detecting variants even within the exome (Belkadi et al., 2015). his may explain why the mutational landscapes analyzed from WGS and WXS of the same cancer type and even the same patients can show significant differences.

While the observed mutational landscapes were most consistent for single nucleotide variants (SNVs) and single base substitution (SBS) mutational signatures in liver and skin cancer, liver cancer exhibited an abnormally high number of SBS29, a signature associated with tobacco chewing (Tate et al., 2019). In the WXS data for skin melanoma, there was a higher abundance of SBS38, a possible UV-associated signature characterized by C>A transversions in the CCA context (Tate et al., 2019). The detection of SBS38 in coding regions may suggest a previously unrecognized mechanism of UV-induced DNA damage that specifically affects genes, potentially expanding our understanding of how UV radiation contributes to mutagenesis in skin cells. Since insertions and deletions are relatively rare occurrences in genes, the distinct differences in the indel mutational landscapes observed between WGS and WXS datasets may be attributed to the unreliable extraction of indel (ID) signatures from WXS due to the smaller number of detected indels

The *de novo* extraction of structural variant signatures using the Palimpsest tool led to the identification of up to seven novel SV signatures across the analyzed cancers. These novel signatures exhibited low cosine similarity to the annotated SV signatures in the COSMIC database, which is likely due to differences in the methodologies used for extraction and the distinct classification systems for structural variants employed by each tool. Palimpsest uses a more detailed 38-class classification system for SVs, compared to the 32-class system used by COSMIC (Shinde et al., 2018; Tate et al., 2019). Despite these differences, two of the de novo SV signatures identified in skin cancer, namely skin SV3 and skin SV2, showed the highest cosine similarity to COSMIC's SV4 and SV6, respectively. The COSMIC SV4 signature is primarily composed of clustered (complex) translocations, whereas SV6 includes a diverse array of very large complex rearrangements, such as deletions, tandem duplications, and inversions (Tate et al., 2019). In mucosal melanoma, the presence of SV4 and SV6, along with a high number of kataegis regions, effectively separated the samples into two distinct groups (Newell et al., 2019). Breast cancer SV identification using the Palimpsest tool showed the poorest performance. The highest similarity was between breast *de novo* SV1 and COSMIC SV1 signature which is characterized by long tandem duplications and common in breast, ovarian, and uterine cancers (Tate et al., 2019). However, the well-known COSMIC SV3 signature, which is associated with homologous recombination deficiency and is prevalent in breast cancer (Nik-Zainal et al., 2016), was not reliably detected in the *de novo* SV signatures identified in breast cancer. The closest *de novo* extracted signature to COMISC SV3 was breast SV2, although the cosine similarity was not the highest for breast *de novo* SV2. This discrepancy is likely due to the different classification systems used by the tools, which may heavily affect the detection and characterization of such signatures. In liver cancer, the de novo liver SV5 signature showed the closest resemblance to COSMIC's SV1, while liver SV6 was most similar to COSMIC's SV3 with the highest overall cosine similarities.

Most studies on structural variants have primarily focused on their impact on the expression of nearby genes, which can be mediated through mechanisms such as enhancer hijacking or the disruption of topologically associated domains (Aganezov et al., 2020; Zhang et al., 2018, 2021). They are not confined to hotspot analyses, as they aim to explore a broader range of mechanisms that contribute to genomic alterations and their functional consequences. Consequently, research specifically targeting SV hotspots in cancer is relatively limited. However, there is one study that focused on the analysis of tandem-duplication COSMIC SV1 and SV3 rearrangement, or as they refer to them as RS1 and RS3, hotspots in 560 breast cancers (Glodzik et al., 2017). They hypothesized that these hotspots represent foci that are more susceptible to double-strand damage leading to NHEJ repairs generating large tandem duplications. While I found 227 breast de novo SV1 (most similar to COSMIC SV1) and 537 SV2 (most similar to COSMIC SV3) hotspots, they detected only 4 SV3 hotspots and 33 SV1 hotspots. The differences may be again due to different classification systems as well as different algorithms for calling peaks. More detailed assessment and optimization of the used SV-HotSpot algorithm (Eteleeb et al., 2020), as well as the implementation of others, is required to obtain more consensus SV hotspot profiles that reliably reflect the processes leading up to cancer.

For the accurate analysis of various genomic features across different cancer types, and even among patients with the same cancer, it is crucial to consider not only the next-generation sequencing technology used for the data but also the specific computational tools employed. To achieve a more robust annotation of SVs, I recommend using multiple tools in addition to Palimpsest to identify novel SV signatures. It is also beneficial to incorporate refitting methods with existing COSMIC signatures to ensure consistency with the cataloged data in the COSMIC database. This approach would facilitate a more comprehensive and reliable representation of the SV landscape, enhancing the validity and accuracy of the analysis. In assessing SBS mutational signature calling, I evaluated multiple tools, all of which demonstrated extremely low reconstruction error, particularly for WGS data. However, there were still a lot of signatures with unknown aetiology and some which could represent artifacts. To further refine SBS calling, it is recommended to focus on tissue-specific signatures rather than using the broad set of all COSMIC signatures. Tissue-specific signatures have been thoroughly analyzed in specific cancers and provide a better biological interpretation of the processes driving tumorigenesis (Koh et al., 2020). For instance, using tissue-specific signatures can help identify and interpret the mutational processes that are particularly relevant to the cancer type under study, thus providing more targeted insights into its genomic landscape. However, it is important to acknowledge that the use of a tissue-specific set of signatures may hinder the detection of rare mutational processes that are present only in a small number of cancer samples and are not included in the existing set. As a result, these rare processes may go undetected, highlighting a significant trade-off in the analysis. Researchers must carefully consider these trade-offs in the context of their specific research questions and the broader implications for cancer genomics studies.

Therefore, to identify even rarer mutational signatures that could significantly impact cell-of-origin models, I decided to use a broad spectrum of COSMIC signatures. This comprehensive approach helps to ensure that less common signatures, which may play critical roles in tumorigenesis, are not overlooked. However, before analyzing how certain genomic characteristics influence these models, it was necessary to adjust the SBS calling to obtain profiles for 1 Mb genomic regions, topologically associated domains, and specific genes. Most mutational calling methods have a notable limitation: they can estimate the absolute and relative abundance of each mutational signature across patients but do not provide information on the specific locations of these mutations within the patient's genome. Only certain tools, like the Palimpsest tool used in this study, can assign mutational signatures to individual mutations (Shinde et al., 2018). However, this feature demonstrated the lowest performance among all the standard tools used. For other tools, I had to split the mutational context matrix by patients for each genomic feature or by genomic features (such as 1 Mb regions, TADs, and genes) and evaluated the mutational calling by calculating the reconstruction error. This approach significantly reduced the overall accuracy of the tools in reconstructing the original mutational contexts, as dividing the data by these features decreased the number of mutations within each context, thereby significantly reducing the power to accurately evaluate mutational signatures. Using a tissue-specific set of mutational signatures might improve the calling of signatures per feature, an approach that warrants further exploration. Additionally, reassessing the obtained results with newer tools such as SigProfilerAssignment (Díaz-Gay et al., 2023), which has shown high accuracy in assigning mutational signatures to individual mutations, could be beneficial. However, the performance of this tool has not been compared directly with the Palimpsest settings. Therefore, re-running and comparing the reconstruction error and mutational profiles using SigProfilerAssignment against those obtained with Palimpsest may result in more precise insights into the genomic landscape of cancer and its implications for cell-of-origin models.

5.2 Cell-of-origin models using whole-genome and whole-exome sequencing

The cell-of-origin approach developed by Paz et al. (2015) utilizing WGS leverages the biologically meaningful correlation between the tumor's mutational landscape and the normal epigenome to accurately predict the COO of each patient. However, when I applied their models to WXS data, the COO prediction was unsuccessful for both aggregated data and individual patient cancer profiles across the analyzed cancers. This discrepancy can be attributed to the lower number of detectable mutations in WXS data, as it does not provide the complete genomic coverage like WGS (Guan et al., 2012). As a result, the WXS-based COO models, developed using 1 Mb genomic regions and topologically associated domains, were inadequate for accurate COO prediction.

To my surprise, even the WXS gene-based COO model failed to predict the COO successfully. I suspect this failure is due to the quality of the WXS data itself. Notably, the COO models for breast and liver cancers were unable to correctly identify the COO using either gene or CDS profiles. However, when simulated gene and CDS profiles from the WGS were used, the predictions were accurate. Only the WXS COO models for skin melanoma, using CDS profiles, managed to predict the COO correctly. This success might be attributed to the higher mutational burden in skin melanoma compared to the other two cancer types, resulting in increased statistical power for the model. Furthermore, I found that COO models utilizing both intronic and exonic mutations in genes resulted in a better COO performance. Introns make up a substantial portion of the human genome, leading to a higher absolute number of mutations in these regions compared to the smaller exonic regions (Lander et al., 2001). Furthermore, mutation rates are often higher in intronic regions than in coding regions, likely because intronic regions can accumulate mutations without necessarily causing deleterious effects on protein function, allowing them to serve as better markers for COO identification (Hodgkinson & Eyre-Walker, 2011). These findings highlight the impact of data quality and quantity on the performance of COO models. Therefore, when designing WXS experiments to identify the cell-of-origin using this particular COO model, it is crucial to be aware of the platform's limitations, particularly in terms of exome capture (Clark et al., 2011), and modify the research accordingly. Additionally, these limitations may also contribute to the significantly lower COO prediction accuracy observed in WGS data of TCGA cohorts across various settings, potentially resulting in inadequate coverage of extensive genomic regions.

However, other COO models, that rely solely on the variants inside genes and across multiple different cancer patients, have successfully predicted the COO across a wide range of samples and various cancer types using more or less the same TCGA/ICGC sample (Chakraborty et al., 2021; Dietlein & Eschner, 2014). Although most of these models' performance is much lower than models developed by WGS, some can achieve model accuracies up more than 90% by applying deep neural network machine learning (Sun et al., 2019). Other ways in which these WXS COO models improve their prediction rates is by using additional features such as copy-number variation information per patient or by resorting to the integration of multi-omics data to achieve (Cai et al., 2022; Marquard et al., 2015). Many of these developed models rely on utilizing advanced machine learning approaches to select the used gene subset with questionable interpretability for each cancer type or rely on the predefined set known to be involved in carcinogenesis (He et al., 2020; Marquard et al., 2015). This discrepancy in WXS data usage suggests that the principle underlying my COO model is particularly sensitive to the highlighted concerns regarding exome capture limitations and data comprehensiveness. As a result, the primary conclusions of the COO model, which rely on a detailed analysis of the mutational landscape and its correlation with normal epigenomes, were based on WGS data.

In addition to the quality of mutational data and the selection of next-generation sequencing techniques, I found that the choice of tissue-specific topologically associated domains did not significantly impact the accuracy of cell-of-origin predictions. This finding aligns with the fact that TADs are largely conserved across different cell types (Boltsis et al., 2021; McArthur & Capra, 2021), which reduces their variability as a factor in improving COO model performance. I found that longer topologically associated domains containing more mutations improved the accuracy of COO models. However, it is the boundaries of TADs that play a crucial role in regulating the expression of nearby genes. These boundaries are known to contribute significantly to complex-trait heritability, particularly for traits related to immunologic, hematologic, and metabolic functions (Boltsis et al., 2021). Given the regulatory importance of TAD boundaries, it would be valuable to expand and improve COO models by incorporating these boundaries to better capture the structural and functional genome organization.

5.2.1 Genomic feature and their contribution to COO models

Developed models COO models using all single-nucleotide variant and epigenome of normal tissues across 1 Mb genomic regions and different topologically-associated domains confirmed the observations detected in previous researches using similar settings (Kübler et al., 2019; Polak et al., 2015).

The use of small insertions and deletions instead of SNVs in the developed COO models did not result in successful identification of the COO, with the exception of the 1 Mb genomic region in skin melanoma, which achieved an accuracy of approximately 30%. This was significantly lower compared to the 80% accuracy obtained with SNVs, and it allowed for the correct identification of only a few patients. This finding aligns with our previous work (Bakšić, 2022), with the notable difference that the random forest algorithm used in this study was unable to predict the correct COO for any melanoma cohorts. Indels are inherently less frequent in the genome compared to SNVs, primarily due to their distinct mutational mechanisms and the selective pressures they are subjected to. Indels, especially those in coding sequences, can have a significant impact on gene function by disrupting reading frames and altering protein structures. Consequently, they are more likely to be subject to purifying selection to maintain essential protein functions (Gagliano et al., 2019) making them unsuitable to use in this developed COO model.

Other published cell-of-origin models seldom, if ever, incorporate indels as significant predictors for determining the COO. Nguyen et al. (2022) found that the mutational load of indels did not rank among the top 15 most significant features for predicting the COO across various cancer types. However, they identified an exception in pilocytic astrocytoma (CNS-PiloAstro), where the mutational load of indels was the second most critical feature, just after regional mutational density (RMD). This suggests that while indels generally have limited predictive power in COO models for most cancers, such as liver, skin melanoma, and breast cancer examined in this research, they may play a more critical role in certain cancer types. Further research is needed to explore whether the indel count in other cancer types, such as pilocytic astrocytoma and other brain cancers, could serve as a valuable resource to enhance the accuracy of COO model predictions based on correlations with normal epigenomes.

5.2.1.1 Mutational signatures

Other important genomic features that have been assessed in predicting the COO are single base substitution (SBS) mutational signatures. Since the developed COO models heavily depend on the correlation between mutations and the normal epigenomes within defined genomic features, it is crucial to establish the relationship between SBS mutational signatures and the epigenome, as well as their impact on prediction accuracy in certain used 1 Mb regions or TADs.

Interestingly, across all cancer types, the clock-like SBS1 signature showed the poorest correlation with the epigenomes of normal cells-of-origin for the corresponding cancer types. The correlations were notably better within TADs, which represent biologically meaningful 3D structural features critical for gene regulation and the maintenance of genomic stability (Long et al., 2022). In contrast, 1 Mb genomic regions are more arbitrary and do not provide the same level of functional insight. Although SBS1 was not the signature with the strongest correlation to erroneous regions, it was documented that using SBS1 in conjunction with chromatin accessibility regions, defined by ATAC-seq, resulted in overall lower accuracy for COO models (Ocsenas & Reimand, 2022). Additionally, they found that the SBS1 signature in liver cancers showed a much stronger association with liver cancer chromatin accessibility profiles compared to normal tissues. In both liver and skin melanoma, COO predictions for patients within 1 Mb regions and TADs had a slightly higher proportion of SBS1, but its overall abundance was negligible compared to other signatures.

Although I did not test whether removing SBS1 mutations would significantly change the accuracy of COO predictions, I focused on another age-related signature with a higher correlation with erroneous predictions and higher abundance: the SBS40 signature. The SBS40 signature has been associated with transcriptionally active domains (Akdemir et al., 2020). Despite Ocsenas and Reimand (2022) showing that the inclusion of SBS40 in prediction models resulted in higher accuracies, my analysis found that removing this "erroneous" SBS40 signature from breast and liver TADs did not significantly impact COO predictions in either aggregated or individual profiles. Nevertheless, I identified that correctly identified COO patients did have a higher proportion of SBS40 compared to incorrectly identified ones, supporting their findings.

The most striking "erroneous" mutational signature identified in the 1 Mb genomic region model was the APOBEC-generated SBS13 signature in breast cancer. This signature exhibited the strongest correlation with erroneous 1 Mb regions and one of the weakest correlations with the normal cell-of-origin epigenome in both 1 Mb regions and TADs. The same APOBEC signature was associated with lower COO prediction accuracy in the study by Despite Ocsenas and Reimand (2022), but it was found to be among the top 15 most important features by Nguyen et al. (2022). Unfortunately, removing the APOBEC SBS13 signature significantly improved the COO model only in aggregated breast cancer profiles; while the predictions for individual patients remained largely unchanged. In breast cancer patients whose mutational profiles are predominantly characterized by APOBEC-associated mutations, the COO prediction tends to be less accurate. While these patients may still achieve a correct COO prediction in some instances, the overall impact of APOBEC mutations is insufficient to significantly enhance prediction accuracy across all individual cases. This limitation is particularly evident in the context of the COO model's objective, which is to provide reliable predictions for each patient. The same can be said for HRD signature SBS3, another less accurate signature identified by Ocsenas and Reimand (2022). Moreover, unlike the APOBEC signature, which is found in early-replicating euchromatic regions, I observed that the SBS8 signature, which occurs in late-replicating heterochromatic domains, is more enriched in TADs and even within the top 40% most mutated breast cancer patients with correctly identified COO. This suggests that mutations arising from uncorrected late-replication errors during cancer progression may be more relevant for COO model prediction.

The liver-specific SBS12 signature has been identified as a crucial feature for COO prediction in multiple studies (Nguyen et al., 2022; Ocsenas & Reimand, 2022). My analysis confirmed its significance, as I detected a higher abundance of SBS12 in patients who were correctly identified as having liver cancer. Conversely, the SBS16 signature, another liver-associated mutational signature, was more frequently observed in patients whose COO was incorrectly predicted. This suggests that while SBS12 is a reliable indicator of liver cancer origin, SBS16 may be less specific or indicative of different mutational processes that complicate accurate COO determination.

In skin melanoma, the worst correlations with the melanocyte epigenome were observed for SBS43 in both 1 Mb regions and TADs. Currently, SBS43 is considered a potential sequencing artifact (Tate et al., 2019), and its removal did not significantly improve the COO model. Interestingly, I found that correctly identified COO patients were significantly enriched with the SBS7a signature, while incorrectly identified patients had a higher proportion of the SBS7b signature in both 1 Mb regions and TADs. Both signatures are UV-induced and result from photoproducts, making it challenging to determine precisely what differentiates these groups and affects the COO models without further experimental validation and mechanistic insight.

In conclusion, the interpretation and application of SBS mutational signatures in COO models should be approached with caution, as their effects can vary significantly depending on the specific context and settings of the COO model.

5.2.1.2 Under-predicted outliers affected by hotspots

Catastrophic genomic events, such as kataegis and structural variant (SV) hotspots, have been frequently associated with regions of high mutational instability. These events are often observed in regions that are under-predicted outliers in both 1 Mb genomic regions and TADs across various cancer types. This pattern is particularly evident in breast cancer, which exhibits the highest number of SV hotspots among the cancers analyzed. Specific analysis of individual SV signature hotspots is challenging due to previously highlighted limitations in SV analysis. Notably, the COSMIC HRD-associated signature SV3, which closely resembles the *de novo breast* SV1, was predominantly enriched with non-outlier regions in breast cancer models of 1 Mb regions and TADs. In contrast, skin *de novo* SV3 (COSMIC SV4) within TADs had higher proportion of under-predicted TADs, while skin SV2 (COSMIC SV6) was significantly enriched with over-predicted 1 Mb outlier regions. These findings underscore the influence of cancer-specific SV hotspots on the erroneous prediction landscape, highlighting the need for more detailed interpretations of SV signature impacts in COO models.

Kataegis regions present essentially hotspots of hypermutation that can distort the mutational landscape, making it difficult for models to accurately predict the true mutational burden. Therefore, COO models might predict a lower number of mutations than what is actually present, which can affect the precision of cancer origin predictions. Kataegis events are frequently linked with SV breaks, as demonstrated in multiple studies (Aaltonen et al., 2020; Nik-Zainal et al., 2012; Roberts et al., 2012). In my analysis, the presence of kataegis was particularly notable in under-predicted breast cancer 1 Mb genomic regions, where 42% of these regions also contained SV hotspots. This overlap suggests that areas with both kataegis and SVs are more error-prone, leading to inaccuracies in mutational burden estimates when using conventional COO models. Such regions, by concentrating a high number of mutations in a small genomic space, complicate the assessment of the overall mutational load.

I detected, especially in breast cancer but also others, that TADs affected by kataegis and/or SV hotspots exhibit a stronger association of single nucleotide variants (SNVs) with regions marked by open chromatin and specific histone modifications. This observation can be attributed to APOBEC-associated mutations (SBS13 and SBS2), which constitute the majority of kataegis events and are typically concentrated in these accessible regions of the genome. APOBEC enzymes are known to induce mutations preferentially in areas of open chromatin, which are more transcriptionally active and have less compacted DNA, leading to higher mutation rates in these regions and TADs (Akdemir et al., 2020; Kazanov et al., 2015). The absence of a strong correlation when examining solely SBS signatures may be due to the predominance of flat signatures such as SBS5, SBS40, and SBS8. These signatures are less specific and can lead to misannotation when using tools like Palimpsest or applying other stringent signature analysis methods in the future. More complex mutational landscapes often require a more detailed approach to signature analysis to accurately reflect the underlying biological processes.

Interestingly, removing regions affected by kataegis and/or SV-hotspots improved the accuracy of the COO model for aggregated profiles but did not have the same effect for individual patients in breast cancer. The removal of kataegis regions in skin melanoma, for instance, reduced the COO model's accuracy, which is understandable given that melanomas are hypermutated tumors that are likely to exhibit more kataegis regions due to their high mutational burden. On the other hand, in breast cancer, kataegis and SVs represent complex regions where multiple repair mechanisms and proteins may be involved, indicating the need for further investigation to understand the impact on COO predictions fully.

In Cancer of Unknown Primary Location Resolver (CUPLR) model (Nguyen et al., 2022) addition of structural variant information significantly improved the COO prediction for cancer that are lacking informative features including central nervous system pilocytic astrocytoma, lung non-small cell carcinoma, and prostate cancer. This shows that adding SV-related information could increase in certain tumors which needs further research in these developed COO model.

Moreover, super-enhancers (SEs) were typically either sparse or completely absent in over-predicted regions and TADs. This lack of super-enhancers in these regions correlates with lower expression levels of tissue-specific genes compared to under-predicted and non-outlier regions, which generally contained a higher number of SEs. The study by Yang et al. (2023) highlights that chromatin marks at or near tissue-specific enhancer regions undergo significant changes during the progression to cancer. This is particularly evident in cases of Barret’s metaplasia and esophageal adenocarcinoma genomes, where the loss of tissue-specific enhancers and dynamic changes in somatic mutation patterns in these regions lead to a reduced correlation with the original COO chromatin marks. However, in breast, liver, and skin melanoma, the non-outlier regions or TADs with the best prediction accuracy either had the highest abundance or showed similar proportions to the more erroneous under-predicted outliers. This suggests that in these cancers, tissue-specific enhancers may not play as significant a role as they tissue-specific ones do in precancerous stated for Barrett’s metaplasia and esophageal adenocarcinoma. The same study by Yang et al. (2023) demonstrated that the implementation of more advanced machine learning algorithms, such as extreme gradient boosting, significantly enhanced the COO prediction accuracy for colorectal and esophageal cancers. However, the prediction accuracy for liver cancer remained largely unchanged. In the present study, I observed similar findings: using an extreme gradient boosting approach did not significantly increase the COO prediction accuracy. Especially when using topologically associated domains to predict individual patients or in gene-based COO models, the algorithm failed to execute successfully. This was mainly due to the limited variance in data splits, which hindered the model's ability to capture distinct patterns necessary for accurate predictions. Further optimization of random forest and extreme gradient boosting parameters is required to improve the prediction accuracy across different genomic features used in COO models.

To further elaborate, under-predicted 1 Mb regions and topologically associated domains demonstrated a slightly higher enrichment of known driver genes, as identified by the Cancer Gene Census (CGC) or Tumor Immune Microenvironment (TIME), across most cancer types and their respective COO model setups. This suggests that these regions might contain critical genomic elements associated with cancer development and progression. Example of detected important cancer drive in under-predicted breast TAD is *ERBB2*, known as Her2, which is amplified and overexpressed in more than 15% of invasive breast cancers (Ng et al., 2015).

Under-predicted features could potentially harbor newly identified cancer genes that have not been extensively studied in these specific cancer types which was also done by other studies (Ocsenas & Reimand, 2022). This highlights the potential for these regions to contain novel oncogenes or tumor suppressors that could expand our understanding of cancer genomics. For this reason, investigating gene-based COO models in these under-predicted regions could provide valuable insights into the identification of new driver genes and enhance our current knowledge of cancer biology.

5.2.2 Gene-based COO models

The selected gene groups based on specific gene characteristics were unable to accurately identify the COO of most individual patients, even though they showed decent COO prediction based on aggregated mutational profiles. The prediction using all of the gene sets was significantly worse than the COO models developed using 1 Mb regions or TADs for each analyzed cancer type. The majority of wrongly identified COO patients were mistakenly categorized as having immune cells or brain-related tissues as their COO. This misassignment might be due to contamination of tumor microenvironment with immune or even recently discovered nerve cells (Jeong et al., 2018) whose mutational profiles of all used genes can maybe outpower the cancers. However, this is a speculative theory that requires further analysis and validation.

Most of the analysis of frequently mutated genes in cancer cohorts rely on driver, non-synonymous mutations, that in some way lead to carcinogenesis or increase cancer fitness. However, background mutations as we already saw with mutational signatures can also tell us a lot about the consequences and processes leading up to cancer. For this reason, for the selection of top mutated genes, I did not filter mutations based on their impact and found them to be brain-related genes with little to no expression in normal tissues of corresponding cancer. These genes were significantly longer than other groups of selected genes, which allows for the accumulation of a greater number of mutations simply due to their increased length. To account for this potential bias, the COO models were normalized by gene length. This normalization ensures that the gene length does not influence the biologically meaningful interpretation of the results, allowing for a more accurate assessment of the relationship between gene mutations and cancer development. I hypothesize that these genes are in lowly expressed regions of the genome where DNA repair is not operating leading to the accumulation of mutations based on the original findings between mutational rate and chromatin (Polak et al., 2015). The low expression of these genes in normal tissues was also confirmed by the absence of tissue-specific enhancers and a high prevalence of top N% mutated genes affected by SV-hotspots and/or kataegis across various cancers. These genes were frequently found within closed chromatin regions, as annotated by TADs, which are less accessible and transcriptionally inactive.

Incorrectly identified COO patients using the best top N% mutated gene subset COO models, showed similar enrichments with SBS mutational signatures in those subset of genes as did they in 1 Mb region and TADs model. In skin melanoma, SBS7a and SBS7c were characteristic for correctly identified one, while SBS7b for incorrectly. In breast cancer the most significant difference was higher enrichment of SBS8 and even SBS2 in correct COO patients, while other signatures especially age relate: SBS1, SBS40 and SBS5 were more abundant in incorrect ones. For liver cancer, the liver-specific mutational signature SBS12 was enriched in correctly identified patients. On the other hand, signatures like SBS16 and SBS3 were more common among incorrectly predicted cases.

Using logistic regression models to analyze mutational signatures in genes has proven to be an effective method for determining cancer’s origin, achieving area under the curve

(AUC) values ranging from 0.76 to 0.93 (Wang et al., 2022). They also found that combining somatic mutation data with cancer-type-specific mutational patterns derived from circulating free DNA (cfDNA) resulted in a high prediction accuracy of 90% for breast and prostate cancers This indicates high accuracy in identifying the COO across various cancers using only an abundance of mutational signatures as predictors. For that reason, mutational signatures still represent a worthy challenge to investigate even further and implement in the COO model based on any genomic feature used in this research.

I hypothesized that maybe early arising signatures as aging ones such as SBS1 and SBS40/5 poorly reflect the cell-of-origin. Signatures generated by endogenous or exogenous sources are those driving and allowing for correct COO identifications. For instance, although the SBS8 proposed aetiology is largely unknown, it is suspected to be involved with NER or HR repair processes (Tate et al., 2019). Various processes, including DNA repair mechanism-related damages characterized by SBS8, accumulate in non-active, brain-related genes that are not crucial for the function of the analyzed tumors. These genes generally follow the normal epigenome until significant driver mutations and genes trigger tumor development. Therefore, it would be valuable to investigate the origins of driver mutations in these genes to determine the initial events that trigger tumorigenesis. For example, the E542K mutation in the PIK3CA gene has been associated with the APOBEC SBS2 mutational signature (Temko et al., 2018), which is indicative of the role of the APOBEC family of cytidine deaminases in generating this mutation. Similarly, in melanomas, the KIT K642E mutation is linked with the clock-like SBS5 signature (Temko et al., 2018), suggesting that this mutation accumulates over time as a result of normal cellular processes rather than exogenous mutagens.

In addition to using only somatic mutations or mutational signatures, incorporating gene expression data into the models has significantly improved their accuracy, achieving up to 97% accuracy in some cases (Abraham et al., 2021; He, Dai, et al., 2020; He, Lang, et al., 2020). This suggests that COO models, which are based on the correlation between mutations and epigenome modifications within genes, can be substantially enhanced by including RNA-seq expression data from normal tissues. Furthermore, including promoter regions in these models can also improve COO prediction accuracy. Promoter regions are key regulatory elements, and certain histone modifications, such as H3K27ac and H3K4me3, are particularly specific to these regions, marking active transcription sites and enhancing the expression of associated genes (Herrera-Uribe et al., 2020; Roadmap Epigenomics Consortium et al., 2015).

5.2.3 Patients characteristic affected by the models

Among the various histological subtypes of liver cancer, cholangiocarcinoma patients often had the cell-of-origin incorrectly assigned to liver tissue in predictive models. This discrepancy arises because the true COO for cholangiocarcinoma is the cholangiocyte, which is an epithelial cell lining the bile ducts (Goral, 2017). These cells are distinct from hepatocytes, which are typically the primary cell type involved in other forms of liver cancer, such as hepatocellular carcinoma (HCC). During liver regeneration, pluripotent progenitor cells can differentiate into either hepatocytes or cholangiocytes, highlighting the complex cellular dynamics within the liver (Duncan et al., 2009). A COO model on 12 extrahepatic biliary tract cholangiocarcinoma samples (BTCAs) showed stomach tissues as COO for BTCAs, in contrast to hepatocytic predicted COO for hepatocellular carcinomas (HCCs) (Ha et al., 2020). Also, for the mixed hepatocellular carcinoma/intrahepatic cholangiocarcinoma subtype, the cell of origin was predominantly hepatocytic, despite the presence of mixed histological features. This suggests that, while these cancers share some characteristics with cholangiocarcinoma, their primary cellular origin aligns more closely with hepatocytes. Hepatocellular adenoma, a rare benign neoplasm of the liver, predominantly occurs in young women who have a history of oral contraceptive use (Wang & Zhang, 2022). This neoplasm was accurately identified as hepatic in origin using COO models that employed 1 Mb genomic regions and TADs. This successful identification underscores the capability of COO models to detect early, benign stages of liver neoplasms before they potentially progress to more malignant forms.

In skin melanoma, the two histological subtypes with the poorest COO model predictions were diagnosed with malignant melanoma, not otherwise specified (NOS), and acral lentiginous melanoma (ALM), which typically occurs on the hands and feet. ALM is a rare subtype of melanoma that predominantly affects individuals of African American, Hispanic, and Asian descent, and is associated with a worse prognosis compared to non-acral melanomas (Bradford et al., 2009). The lower prediction accuracy for these melanoma types can largely be attributed to their lower mutational account compared to other types.

As for breast carcinoma, a higher mutational burden did not necessarily correlate with the accurate prediction of the cell-of-origin, particularly in cases of invasive ductal carcinoma (IDC) and in patients with an unknown histological subtype. In contrast, patients with invasive lobular carcinoma (ILC) who were correctly identified by the COO model generally exhibited a higher mutational load. IDC and ILC are the two most prevalent types of breast cancer, each characterized by distinct molecular and histological features, suggesting that their unique mutational landscapes may influence COO predictions differently. ILC is well-known for the loss E-cadherin loss, as well as showing *CDH1* and *PTEN* loss, *AKT* activation, and mutations in *TBX3* and *FOXA1* (Ciriello et al., 2015). ILC has a more favorable prognosis due to its hormone receptor positivity, lower histological grade, HER2 negativity, and better response to endocrine therapy than IDC (Barroso-Sousa & Metzger-Filho, 2016; Filho et al., 2015). However, ILC patients with high-risk indicators, such as those who are hormone receptor-negative and lymph node-positive, have shown worse overall survival (OS) compared to IDC patients in subgroup analyses, highlighting the complexity of prognostic factors within these subtypes (C. Yang et al., 2020).

Kubler et al. (2019) showed that both ductal and lobular carcinomas have the same mature cell-of-origin despite their molecular and histological differences, suggesting that factors beyond the primary cell of origin influence their development and progression. Their study also found that homologous recombination deficiency (HRD), characterized by inactivation of *BRCA1*, *BRCA2*, or *RAD51C*, did not significantly influence the COO predictions when assessed using aggregated mutational profiles. My approach leveraged HRD information from both HRD classifiers, CHORD and HRDetect, to increase the number of HRD patients to assess the influence on COO model prediction (Štancl et al., 2022). Interestingly, HRD patients were less correctly identified than non-HRD patients using the 1 Mb region model. However, models utilizing TADs and the top 40% of most mutated genes resulted in better COO prediction accuracy for HRD patients. TADs could provide a more functionally relevant and structurally intact framework that captures the spatial and regulatory interactions influencing the genomic instability seen in HRD. While top 40% mutated genes may unravel new key driver genes for HRD. Unfortunately, none of the PAM50 groups resulted in correctly identifying the cell-of-origin for nearly any patients, regardless of the genomic feature used. For these numerous reasons, further detailed research is required to explore these findings and refine the predictive models for various breast cancer types.

5.3 RNA-seq called mutations in gene-based cell-of-origin model

RNA-seq called mutations resulted in significant high rate of false positive mutations, especially when using only tumor mode calling with Mutect2 of skin-melanoma. While paired both tumor and normal tissues alongside multiple variants calling algorithms resulted in a much more similar mutational count number to WXS than WGS technologies, aligned with good practices in cancer genomics (Goode et al., 2013; Koboldt, 2020; Van der Auwera et al., 2013). There was still a significant bias in RNA-seq called mutations which is apparent from RNA-seq profiles of transversions and transitions from different cancers were more similar to one another than to their corresponding cancers profiles detected by other NGS technologies. This was also apparent by high abundance of called A-to-G transitions. These transitions are caused by adenosine deaminases acting on RNA (ADARs) proteins which modify genetically encoded A to inosine (I) in double-stranded RNA (dsRNA) substrates (Bass, 2002; Savva et al., 2012). That results in increased number of A-to-G transitions in mRNA-sequencing data when aligned to the reference genome, which is quite abundant in human genes (Bazak et al., 2014). Although I did remove all the RNA-editing sites annotated in DARNED and RADAR databases, apparently majority of possibly new ones remained in analyzed dataset. Further improvement of calling the RNA-seq variants can be done by creating RNA-based panel of normal using for instance multiple normal tissues found in the GTEx database to filter out various RNA artifacts (Long et al., 2022). They also included in their pipeline filtered out of mutations found in immunoglobulin, pseudogenes and non-coding RNA alignments, as well as removing intronic mutations to improve the overall quality of the RNA-seq called mutations. I detected those enrichment od immunoglobulin genes and intronic mutations in my data, although the RNA-seq intronic mutation was very similar in breast cancers RNA-seq and WXS mutations probably due to the designed library used in WXS to capture both gene body regions. As I wanted to compare the best predictive COO models based on different gene groups, I did not do the additional filtering of intronic mutations.

Most frequently mutated genes also significantly differed from RNA-seq mutated ones to ones detected by WGS. Although the top ones, mutated in over 50% of sample in breast and liver cancer were not found in Cancer Gene Census, some of them were known to be involved in carcinogenesis. For breast cancer RNA-seq, I detected a *UTRN* gene, which encodes utrophin, a dystrophin-related protein, to also be presented in top 40% mutated genes by WGS. *UTRN* is a tumor suppressor gene found to be mutated in multiple cancers and to contain somatic truncating mutations in primary breast cancer (Li et al., 2007). In liver cancer, *CYP2E1* gene under regulation of hepatocytes super-enhancer is known to also act as a tumor suppressor by regulating Wnt/Dvl2/β-catenin (Zhu et al., 2022). Both of these detected tumor suppressors are downregulated in their corresponding detected cancers. As for skin melanoma RNA-seq mutations, majority of top mutated genes showed very little overlap to the WGS identified, with also varying enriched Gene Ontology terms for which these genes are involved.

Out of all used gene groups for developing the COO models, the tissue-specific ones managed to capture the highest model accuracy in breast cancer, while melanoma’s best model was with top N% mutated genes regardless of the CDS or gene length normalization. However, since the second non-COO model was very similar to the best identified one, this shows insufficient power and reliability in COO predictions which can be further tested using multiple runs with different seed settings. Individual predictions failed to identify the correct COO in most of the patients implying that further improvements and reevaluation of the whole gene-base RNA-seq variant mutational COO models need to do in the future.

6 Conclusion

In this thesis I developed multiple cell-of-origin prediction (COO) models based on mutational landscape and epigenomic profiles in 1 Mb genomic regions, topologically-associated domain and genes for breast, liver and skin melanoma cancers and found following:

1. High-quality mutational data, particularly data that includes intronic mutations, enhances the prediction accuracy of COO models, regardless of the genomic feature used
2. Insertions and deletions (indels) alone do not provide sufficient power for reliable COO prediction across different cancer types
3. The prediction of COO for any cancer type is not significantly influenced by tissue-specific TADs, as TADs are conserved across different cell types and do not necessarily reflect tissue-specific genomic features
4. Longer genomic features tend to contain more mutations, which contributes to a more accurate COO prediction
5. APOBEC-generated mutations, which are primarily associated with kataegis regions, reduce the accuracy of COO models based on arbitrarily selected 1 Mb genomic regions in breast cancers and are enriched in genomic features of breast patients with incorrectly identified COO
6. The removal of kataegis-affected regions significantly decreases COO prediction accuracy in hypermutated cancers like skin melanoma
7. Structural variant (SV) hotspots within TADs reduce the COO model's prediction accuracy for aggregated mutational profiles in breast cancer
8. Under-predicted genomic features are enriched with cancer driver genes, SV-hotspots and kataegis events while regions where COO models perform well tend to have a higher presence of super-enhancers
9. The most frequently mutated genes, involved in brain-related processes and typically found in more closed chromatin regions, offer the highest COO prediction accuracy in gene-based COO models
10. Different histopathological subtypes of cancer exhibit varying levels of accuracy in COO model predictions
11. Advanced machine learning methods, such as random forest and extreme gradient boosting, generally yield similar COO prediction accuracy. However, in cases with

a low number of mutations and insufficient variation, these models struggle to predict COO due to their complex data splitting mechanisms

12. Mutations identified through RNA-Seq data often result in a high number of false positives, though they can still predict COO for cancer on aggregated mutational profiles

To sum up, this thesis shows that the accuracy of predicting the cell-of-origin is greatly improved by high-quality mutational data, especially when considering intronic mutations and longer genomic features. Nevertheless, challenges persist in certain contexts, such as the impact of kataegis regions and structural variant hotspots, emphasizing the need for careful selection of genomic features and advanced machine learning approaches tailored to the complexities of cancer genomes.

7 References

Aaltonen, L. A., Abascal, F., Abeshouse, A., Aburatani, H., Adams, D. J., Agrawal, N., Ahn, K. S., Ahn, S.-M., Aikata, H., Akbani, R., Akdemir, K. C., Al-Ahmadie, H., Al-Sedairy, S. T., Al-Shahrour, F., Alawi, M., Albert, M., Aldape, K., Alexandrov, L. B., Ally, A., … von Mering, C. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Abraham, J., Heimberger, A. B., Marshall, J., Heath, E., Drabick, J., Helmstetter, A., Xiu, J., Magee, D., Stafford, P., Nabhan, C., Antani, S., Johnston, C., Oberley, M., Korn, W. M., & Spetzler, D. (2021). Machine learning analysis using 77,044 genomic and transcriptomic profiles to accurately predict tumor type. *Translational Oncology*, *14*(3). https://doi.org/10.1016/j.tranon.2021.101016

Aganezov, S., Goodwin, S., Sherman, R. M., Sedlazeck, F. J., Arun, G., Bhatia, S., Lee, I., Kirsche, M., Wappel, R., Kramer, M., Kostroff, K., Spector, D. L., Timp, W., Richard McCombie, W., & Schatz, M. C. (2020). Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Research*, *30*(9), 1258–1273. https://doi.org/10.1101/GR.260497.119

Ahmad, M., Weiswald, L. B., Poulain, L., Denoyelle, C., & Meryet-Figuiere, M. (2023). Involvement of lncRNAs in cancer cells migration, invasion and metastasis: cytoskeleton and ECM crosstalk. In *Journal of Experimental and Clinical Cancer Research* (Vol. 42, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13046-023-02741-x

Akdemir, K. C., Le, V. T., Chandran, S., Li, Y., Verhaak, R. G., Beroukhim, R., Campbell, P. J., Chin, L., Dixon, J. R., Futreal, P. A., Akdemir, K. C., Alvarez, E. G., Baez-Ortega, A., Boutros, P. C., Bowtell, D. D. L., Brors, B., Burns, K. H., Campbell, P. J., Chan, K., … von Mering, C. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nature Genetics*, *52*(3), 294–305. https://doi.org/10.1038/s41588-019-0564-y

Akdemir, K. C., Le, V. T., Kim, J. M., Killcoyne, S., King, D. A., Lin, Y. P., Tian, Y., Inoue, A., Amin, S. B., Robinson, F. S., Nimmakayalu, M., Herrera, R. E., Lynn, E. J., Chan, K., Seth, S., Klimczak, L. J., Gerstung, M., Gordenin, D. A., O'Brien, J., … Andrew Futreal, P. (2020). Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nature Genetics*, *52*(11), 1178–1188. https://doi.org/10.1038/s41588-020-0708-0

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Boutros, P., … Yu, W. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J.

A., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alhamlan, F. S., Bakheet, D. M., Bohol, M. F., Alsanea, M. S., Alahideb, B. M., Alhadeq, F. M., Alsuwairi, F. A., Al-Abdulkareem, M. A., Asiri, M. S., Almaghrabi, R. S., Altamimi, S. A., Mutabagani, M. S., Althawadi, S. I., & Al-Qahtani, A. A. (2023). SARS-CoV-2 spike gene Sanger sequencing methodology to identify variants of concern. *BioTechniques*, *74*(2), 69–75. https://doi.org/10.2144/btn-2021-0114

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. In *Genome Biology* (Vol. 21, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-020-1935-5

Antoniou, A. C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkäs, K., Roberts, J., Lee, A., Subramanian, D., De Leeneer, K., Fostira, F., Tomiak, E., Neuhausen, S. L., Teo, Z. L., Khan, S., Aittomäki, K., Moilanen, J. S., Turnbull, C., Seal, S., Mannermaa, A., … Tischkowitz, M. (2014). Breast-Cancer Risk in Families with Mutations in PALB2 . *New England Journal of Medicine*, *371*(6), 497–506. https://doi.org/10.1056/nejmoa1400382

Aran, D., Camarda, R., Odegaard, J., Paik, H., Oskotsky, B., Krings, G., Goga, A., Sirota, M., & Butte, A. J. (2017). Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-01027-z

Aylon, Y., & Oren, M. (2011). New plays in the p53 theater. In *Current Opinion in Genetics and Development* (Vol. 21, Issue 1, pp. 86–92). https://doi.org/10.1016/j.gde.2010.10.002

Baker, M. (2012). Structural variation: The genome's hidden architecture. *Nature Methods*, *9*(2), 133–137. https://doi.org/10.1038/nmeth.1858

Bakšić, I. (2022). *Određivanje ishodišne stanice tumora na temelju raspodjele različitih tipova mutacija* [Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet]. https://urn.nsk.hr/urn:nbn:hr:217:265845

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., Kostareva, A. A., Glotov, O. S., & Predeus, A. V. (2020). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-59026-y

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, *41*(D1). https://doi.org/10.1093/nar/gks1193

Barroso-Sousa, R., & Metzger-Filho, O. (2016). Differences between invasive lobular and invasive ductal carcinoma of the breast: Results and therapeutic implications. In *Therapeutic Advances in Medical Oncology* (Vol. 8, Issue 4, pp. 261–266). SAGE Publications Inc. https://doi.org/10.1177/1758834016644156

Barutcu, A. R., Lajoie, B. R., McCord, R. P., Tye, C. E., Hong, D., Messier, T. L., Browne, G., van Wijnen, A. J., Lian, J. B., Stein, J. L., Dekker, J., Imbalzano, A. N., & Stein, G. S.

(2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0768-0

Bass, B. L. (2002). RNA Editing by Adenosine Deaminases That Act on RNA. *Annual Review of Biochemistry*, *71*, 817–846. https://doi.org/10.1146/annurev.biochem.71.110601.135501

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F. J., Rechavi, G., Li, J. B., Eisenberg, E., & Levanon, E. Y. (2014). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Research*, *24*(3), 365–376. https://doi.org/10.1101/gr.164749.113

Beauchamp, K., Moran, B., O'Brien, T., Brennan, D., Crown, J., Sheahan, K., & Cotter, M. B. (2023). Carcinoma of unknown primary (CUP): an update for histopathologists. In *Cancer and Metastasis Reviews* (Vol. 42, Issue 4, pp. 1189–1200). Springer. https://doi.org/10.1007/s10555-023-10101-6

Beddoe, A. M., Nair, N., & Dottino, P. (2016). Challenges to Cancer Program Development in Low- and Middle-Income Countries. In *Annals of Global Health* (Vol. 82, Issue 4, pp. 614–620). Elsevier USA. https://doi.org/10.1016/j.aogh.2016.08.001

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J. L., & Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(17), 5473–5478. https://doi.org/10.1073/pnas.1418631112

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. *BioRxiv*. https://doi.org/10.1101/861054

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature, 321*.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boltsis, I., Grosveld, F., Giraud, G., & Kolovos, P. (2021). Chromatin Conformation in Development and Disease. In *Frontiers in Cell and Developmental Biology* (Vol. 9). Frontiers Media S.A. https://doi.org/10.3389/fcell.2021.723859

Borkiewicz, L. (2021). Histone 3 lysine 27 trimethylation signature in breast cancer. In *International Journal of Molecular Sciences* (Vol. 22, Issue 23). MDPI. https://doi.org/10.3390/ijms222312853

Bouffet, E., Larouche, V., Campbell, B. B., Merico, D., De Borja, R., Aronson, M., Durn, C., Krueger, J., Cabric, V., Ramaswamy, V., Zhukova, N., Mason, G., Farah, R., Afzal, S., Yalon, M., Rechavi, G., Magimairajan, V., Walsh, M. F., Constantini, S., … Tabori, U. (2016). Immune checkpoint inhibition for hypermutant glioblastoma multiforme resulting from germline biallelic mismatch repair deficiency. *Journal of Clinical Oncology*, *34*(19), 2206–2211. https://doi.org/10.1200/JCO.2016.66.6552

Bradford, P. T., Goldstein, A. M., McMaster, M. L., & Tucker, M. A. (2009). Acral lentiginous melanoma: Incidence and survival patterns in the United States, 1986-2005. *Archives of Dermatology*, *145*(4), 427–434. https://doi.org/10.1001/archdermatol.2008.609

Brash, D. E. (2015). UV signature mutations. In *Photochemistry and Photobiology* (Vol. 91, Issue 1, pp. 15–26). Blackwell Publishing Inc. https://doi.org/10.1111/php.12377

Brash, D. E., Rudolph, J. A., Simon, J. A., Lin, A., Mckenna, G. J., Badent, H. P., Halperin, A. J., & Pontin$, J. (1991). A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma (UV light/tumor suppressor genes). *Proceedings of the National Academy of Sciences*, *88*(22), 10124–10128.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*. https://doi.org/10.3322/caac.21834

Breasted, J. H. (1930). *The Edwin Smith Surgical Papyrus*. University of Chicago Press.

Bressac, B., Puisieux, A., Kew, M., Volkmann, M., Krebsforschungszentrum, D., Bozcall, S., Bella Mura, J., de la Monte, S., Carlson, R., Blum, H., Wands, J., Takahashi, H., von Weizsacker, F., Galun, E., General, M., Kar, S., Carr, B. I., Schroder, C. H., Erken, E., … Tang, Z. (1991). p53 mutation in hepatocellular carcinoma after aflatoxin exposure. *The Lancet*, *338*(8779), 1356–1359.

Brown, J. S., Amend, S. R., Austin, R. H., Gatenby, R. A., Hammarlund, E. U., & Pienta, K. J. (2023). Updating the Definition of Cancer. *Molecular Cancer Research*, *21*(11), 1142–1147. https://doi.org/10.1158/1541-7786.MCR-23-0411

Cai, L., Yuan, W., Zhang, Z., He, L., & Chou, K. C. (2016). In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Scientific Reports*, *6*. https://doi.org/10.1038/srep36540

Cai, Z., Poulos, R. C., Liu, J., & Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *IScience*, *25*(2). https://doi.org/10.1016/j.isci

Chakraborty, S., Begg, C. B., & Shen, R. (2021). Using the "Hidden" genome to improve classification of cancer types. *Biometrics*, *77*(4), 1445–1455. https://doi.org/10.1111/biom.13367

Chen, Z., Yuan, Y., Chen, X., Chen, J., Lin, S., Li, X., & Du, H. (2020). Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-60559-5

Chilamakuri, C. S. R., Lorenz, S., Madoui, M. A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., & Meza-Zepeda, L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics*, *15*(1). https://doi.org/10.1186/1471-2164-15-449

Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumno, P., Nayir, A., Bakkaloglu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Fahri, A., Mane, S., & Lifron, R. P. (2009). *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing* (Vol. 106, Issue 45).

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., & Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219. https://doi.org/10.1038/nbt.2514

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., Bowlby, R., Shen, H., Hayat, S., Fieldhouse, R., Lester, S. C., Tse, G. M. K., Factor, R. E., Collins, L. C., Allison, K. H., … Perou, C. M. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*, *163*(2), 506–519. https://doi.org/10.1016/j.cell.2015.09.033

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., Butte, A. J., & Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, *29*(10), 908–916. https://doi.org/10.1038/nbt.1975

Costa, P. M. da S., Sales, S. L. A., Pinheiro, D. P., Pontes, L. Q., Maranhão, S. S., Pessoa, C. do Ó., Furtado, G. P., & Furtado, C. L. M. (2023). Epigenetic reprogramming in cancer: From diagnosis to treatment. In *Frontiers in Cell and Developmental Biology* (Vol. 11). Frontiers Media S.A. https://doi.org/10.3389/fcell.2023.1116805

Cremer, T., & Cremer, M. (2010). Chromosome territories. In *Cold Spring Harbor perspectives in biology* (Vol. 2, Issue 3). https://doi.org/10.1101/cshperspect.a003889

da Fonseca, L. G., Reig, M., & Bruix, J. (2020). Tyrosine Kinase Inhibitors and Hepatocellular Carcinoma. In *Clinics in Liver Disease* (Vol. 24, Issue 4, pp. 719–737). W.B. Saunders. https://doi.org/10.1016/j.cld.2020.07.012

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., … Futreal, & P. A. (2002). *Mutations of the BRAF gene in human cancer*. www.nature.com/nature

Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., Sieuwerts, A. M., Simpson, P. T., King, T. A., Raine, K., Eyfjord, J. E., Kong, G., Borg, Å., Birney, E., Stunnenberg, H. G., Van De Vijver, M. J., … Nik-Zainal, S. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, *23*(4), 517–525. https://doi.org/10.1038/nm.4292

Depledge, D. P., Srinivas, K. P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D. G., Mohr, I., & Wilson, A. C. (2019). Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-08734-9

Díaz-Gay, M., Vangara, R., Barnes, M., Wang, X., Islam, S. M. A., Vermes, I., Duke, S., Narasimman, N. B., Yang, T., Jiang, Z., Moody, S., Senkin, S., Brennan, P., Stratton, M. R., & Alexandrov, L. B. (2023). Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. *Bioinformatics*, *39*(12). https://doi.org/10.1093/bioinformatics/btad756

Dietlein, F., & Eschner, W. (2014). Inferring primary tumor sites from mutation spectra: A meta-analysis of histology-specific aberrations in cancer-derived cell lines. *Human Molecular Genetics*, *23*(6), 1527–1537. https://doi.org/10.1093/hmg/ddt539

Divate, M., Tyagi, A., Richard, D. J., Prasad, P. A., Gowda, H., & Nagaraj, S. H. (2022). Deep Learning-Based Pan-Cancer Classification Model Reveals Tissue-of-Origin Specific Gene Expression Signatures. *Cancers, 14*(5). https://doi.org/10.3390/cancers14051185

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://code.google.com

Dulbecco, R. (1986). A turning point in cancer research: sequencing the human genome. *Science, 231*(4742), 1055–1056. https://doi.org/10.1126/science.3945817

Duncan, A. W., Dorrell, C., & Grompe, M. (2009). Stem Cells and Liver Regeneration. In *Gastroenterology* (Vol. 137, Issue 2, pp. 466–481). W.B. Saunders. https://doi.org/10.1053/j.gastro.2009.05.044

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*(16), 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Korlach, J., & Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science, 323*(5910), 130–133. https://doi.org/10.1126/science.1166256

Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics*, *14*(5), 305–314. https://doi.org/10.1093/bfgp/elv014

Eteleeb, A. M., Quigley, D. A., Zhao, S. G., Pham, D., Yang, R., Dehm, S. M., Luo, J., Feng, F. Y., Dang, H. X., & Maher, C. A. (2020). SV-HotSpot: detection and visualization of hotspots targeted by structural variants associated with gene expression. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-71168-7

Fagerberg, L., Hallstrom, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjostedt, E., Lundberg, E., Szigyarto, C. A. K., Skogs, M., Ottosson Takanen, J., Berling, H., Tegel, H., Mulder, J., … Uhlen, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular and Cellular Proteomics*, *13*(2), 397–406. https://doi.org/10.1074/mcp.M113.035600

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. In *Nature Reviews Genetics* (Vol. 7, Issue 2, pp. 85–97). https://doi.org/10.1038/nrg1767

Filho, O. M., Giobbie-Hurder, A., Mallon, E., Gusterson, B., Viale, G., Winer, E. P., Thürlimann, B., Gelber, R. D., Colleoni, M., Ejlertsen, B., Debled, M., Price, K. N., Regan, M. M., Coates, A. S., & Goldhirsch, A. (2015). Relative effectiveness of letrozole compared with tamoxifen for patients with lobular carcinoma in the BIG 1-98 Trial.

Journal of Clinical Oncology, *33*(25), 2772–2778. https://doi.org/10.1200/JCO.2015.60.8133

Flaherty, K. T., Infante, J. R., Daud, A., Gonzalez, R., Kefford, R. F., Sosman, J., Hamid, O., Schuchter, L., Cebon, J., Ibrahim, N., Kudchadkar, R., Burris, H. A., Falchook, G., Algazi, A., Lewis, K., Long, G. V., Puzanov, I., Lebowitz, P., Singh, A., … Weber, J. (2012). Combined BRAF and MEK Inhibition in Melanoma with BRAF V600 Mutations. *New England Journal of Medicine*, *367*(18), 1694–1703. https://doi.org/10.1056/nejmoa1210093

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kä hä ri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., … Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Research*, *39*(SUPPL. 1). https://doi.org/10.1093/nar/gkq1064

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., & Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, *7*(6), 461–465. https://doi.org/10.1038/nmeth.1459

Fraga, M. F., Ballestar, E., Villar-Garea, A., Boix-Chornet, M., Espada, J., Schotta, G., Bonaldi, T., Haydon, C., Ropero, S., Petrie, K., Iyer, N. G., Pérez-Rosado, A., Calvo, E., Lopez, J. A., Cano, A., Calasanz, M. J., Colomer, D., Piris, M. Á., Ahn, N., … Esteller, M. (2005). Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genetics*, *37*(4), 391–400. https://doi.org/10.1038/ng1531

Friedenreich, C. M., Ryder-Burbidge, C., & McNeil, J. (2021). Physical activity, obesity and sedentary behavior in cancer etiology: epidemiologic evidence and biologic mechanisms. In *Molecular Oncology* (Vol. 15, Issue 3, pp. 790–800). John Wiley and Sons Ltd. https://doi.org/10.1002/1878-0261.12772

Gabrielaite, M., Torp, M. H., Rasmussen, M. S., Andreu-Sánchez, S., Vieira, F. G., Pedersen, C. B., Kinalis, S., Madsen, M. B., Kodama, M., Demircan, G. S., Simonyan, A., Yde, C. W., Olsen, L. R., Marvig, R. L., Østrup, O., Rossing, M., Nielsen, F. C., Winther, O., & Bagger, F. O. (2021). A comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers*, *13*(24). https://doi.org/10.3390/cancers13246283

Gagliano, S. A., Sengupta, S., Sidore, C., Maschio, A., Cucca, F., Schlessinger, D., & Abecasis, G. R. (2019). Relative impact of indels versus SNPs on complex disease. *Genetic Epidemiology*, *43*(1), 112–117. https://doi.org/10.1002/gepi.22175

Glodzik, D., Morganella, S., Davies, H., Simpson, P. T., Li, Y., Zou, X., Diez-Perez, J., Staaf, J., Alexandrov, L. B., Smid, M., Brinkman, A. B., Rye, I. H., Russnes, H., Raine, K., Purdie, C. A., Lakhani, S. R., Thompson, A. M., Birney, E., Stunnenberg, H. G., … Nik-Zainal, S. (2017). A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nature Genetics*, *49*(3), 341–348. https://doi.org/10.1038/ng.3771

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., & Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, *14 Suppl 3*. https://doi.org/10.1186/1471-2164-14-s3-s7

Goode, D. L., Hunter, S. M., Doyle, M. A., Ma, T., Rowley, S. M., Choong, D., Ryland, G. L., & Campbell, I. G. (2013). A simple consensus approach improves somatic mutation prediction accuracy. *Genome Medicine*, *5*(9). https://doi.org/10.1186/gm494

Goral, V. (2017). Cholangiocarcinoma: New insights. In *Asian Pacific Journal of Cancer Prevention* (Vol. 18, Issue 6, pp. 1469–1473). Asian Pacific Organization for Cancer Prevention. https://doi.org/10.22034/APJCP.2017.18.6.1469

Greco, A. F., & Hainsworth, J. D. (1992). Tumors of Unknown Origin. *CA: A Cancer Journal for Clinicians*, *42*(96). https://doi.org/10.3322/canjclin.42.2.96

Guan, Y. F., Li, G. R., Wang, R. J., Yi, Y. T., Yang, L., Jiang, D., Zhang, X. P., & Peng, Y. (2012). Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. In *Chinese Journal of Cancer* (Vol. 31, Issue 10, pp. 463–470). https://doi.org/10.5732/cjc.012.10216

Ha, K., Fujita, M., Karlić, R., Yang, S., Xue, R., Zhang, C., Bai, F., Zhang, N., Hoshida, Y., Polak, P., Nakagawa, H., Kim, H. G., & Lee, H. (2020). Somatic mutation landscape reveals differential variability of cell-of-origin for primary liver cancer. *Heliyon*, *6*(2). https://doi.org/10.1016/j.heliyon.2020.e03350

Hainaut, P., & Pfeifer, G. (2016). *Somatic TP53 Mutations in the Era of Genome Sequencing* (G. Lozanzo & A. Levine, Eds.; 1st edn). Cold Spring Harbor Laboratory Press: Cold Spring Harbor.

Hajdu, S. I. (2011). A note from history: Landmarks in history of cancer, part 1. In *Cancer* (Vol. 117, Issue 5, pp. 1097–1102). https://doi.org/10.1002/cncr.25553

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. In *Cancer Discovery* (Vol. 12, Issue 1, pp. 31–46). American Association for Cancer Research Inc. https://doi.org/10.1158/2159-8290.CD-21-1059

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre. In *Cell* (Vol. 100).

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. In *Cell* (Vol. 144, Issue 5, pp. 646–674). https://doi.org/10.1016/j.cell.2011.02.013

He, B., Dai, C., Lang, J., Bing, P., Tian, G., Wang, B., & Yang, J. (2020). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1866*(11). https://doi.org/10.1016/j.bbadis.2020.165916

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., Gao, W., Bing, P., Tian, G., & Yang, J. (2020). TOOme: A Novel Computational Framework to Infer Cancer Tissue-of-Origin by Integrating Both Gene Mutation and Expression. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.00394

Herrera-Uribe, J., Liu, H., Byrne, K. A., Bond, Z. F., Loving, C. L., & Tuggle, C. K. (2020). Changes in H3K27ac at Gene Regulatory Regions in Porcine Alveolar Macrophages Following LPS or PolyIC Exposure. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00817

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, Charles. H., Goldmann, J., Lajoie, B. R., Fan Peng, Z., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J., & Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, *351*(6280), 1454–1458. https://doi.org/10.1126/science.aad2257

Ho, S. S., Urban, A. E., & Mills, R. E. (2020). Structural variation in the sequencing era. In *Nature Reviews Genetics* (Vol. 21, Issue 3, pp. 171–189). Nature Research. https://doi.org/10.1038/s41576-019-0180-9

Hodgkinson, A., & Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. In *Nature Reviews Genetics* (Vol. 12, Issue 11, pp. 756–766). https://doi.org/10.1038/nrg3098

Hsu, F., Kent, J. W., Clawson, H., Kuhn, R. M., Diekhans, M., & Haussler, D. (2006). The UCSC known genes. *Bioinformatics*, *22*(9), 1036–1046. https://doi.org/10.1093/bioinformatics/btl048

Huang, X., Wojtowicz, D., & Przytycka, T. M. (2018). Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics*, *34*(2), 330–337. https://doi.org/10.1093/bioinformatics/btx604

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusuda, J., Lane, D. P., … Wainwright, B. J. (2010). International network of cancer genome projects. In *Nature* (Vol. 464, Issue 7291, pp. 993–998). https://doi.org/10.1038/nature08987

Illumina.Inc. (2010). *Illumina Sequencing Technology*. https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Inoue, K., & Fry, E. A. (2017). Haploinsufficient tumor suppressor genes. *Advances in Experimental Medicine and Biology*, *118*, 83–122.

International Agency for Research on Cancer. (2024, March 13). *GLOBOCAN 2020*. Https://Gco.Iarc.Fr/Today/Home. https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer/incidence#heading-One

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology, 17*(1). https://doi.org/10.1186/s13059-016-1103-0

Jeong, S., Zheng, B., Wang, H., Xia, Q., & Chen, L. (2018). Nervous system and primary liver cancer. In *Biochimica et Biophysica Acta - Reviews on Cancer* (Vol. 1869, Issue 2, pp. 286–292). Elsevier B.V. https://doi.org/10.1016/j.bbcan.2018.04.002

Jia, Y., Chng, W. J., & Zhou, J. (2019). Super-enhancers: Critical roles and therapeutic targets in hematologic malignancies. In *Journal of Hematology and Oncology* (Vol. 12, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13045-019-0757-y

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D. M., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., & Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, *502*(7471), 333–339. https://doi.org/10.1038/nature12634

Karlić, R., Chung, H. R., Lasserre, J., Vlahoviček, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(7), 2926–2931. https://doi.org/10.1073/pnas.0909344107

Kastan, M. B., Canman, C. E., & Leonard, C. J. (1995). P53, cell cycle control and apoptosis: Implications for cancer. In *Cancer and Metastasis Reviews* (Vol. 14).

Kazanov, M. D., Roberts, S. A., Polak, P., Stamatoyannopoulos, J., Klimczak, L. J., Gordenin, D. A., & Sunyaev, S. R. (2015). APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions. *Cell Reports*, *13*(6), 1103–1109. https://doi.org/10.1016/j.celrep.2015.09.077

Kchouk, M., Gibrat, J. F., & Elloumi, M. (2017). Generations of Sequencing Technologies: From First to Next Generation. *Biology and Medicine*, *09*(03). https://doi.org/10.4172/0974-8369.1000395

Kim, J., Mouw, K. W., Polak, P., Braunstein, L. Z., Kamburov, A., Tiao, G., Kwiatkowski, D. J., Rosenberg, J. E., Van Allen, E. M., D'Andrea, A. D., & Getz, G. (2016). Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nature Genetics*, *48*(6), 600–606. https://doi.org/10.1038/ng.3557

Kim, S., Scheffler, K., Halpern, A. L., Bekritsky, M. A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., & Saunders, C. T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, *15*(8), 591–594. https://doi.org/10.1038/s41592-018-0051-x

Kim, T., & Croce, C. M. (2023). MicroRNA: trends in clinical trials of cancer diagnosis and therapy strategies. In *Experimental and Molecular Medicine* (Vol. 55, Issue 7, pp. 1314–1321). Springer Nature. https://doi.org/10.1038/s12276-023-01050-9

Kiran, A., & Baranov, P. V. (2010). DARNED: A DAtabase of RNa editing in humans. *Bioinformatics*, *26*(14), 1772–1776. https://doi.org/10.1093/bioinformatics/btq285

Knudson, A. G., & Knudson, A. G. (1996). Hereditary cancer: two hits revisited. In *J Cancer Res Clin Oncol* (Vol. 122). Springer-Verlag.

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. In *Genome Medicine* (Vol. 12, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-020-00791-w

Koh, G., Degasperi, A., Zou, X., Momen, S., & Nik-Zainal, S. (2021). Mutational signatures: emerging concepts, caveats and clinical applications. In *Nature Reviews Cancer* (Vol. 21, Issue 10, pp. 619–637). Nature Research. https://doi.org/10.1038/s41568-021-00377-7

Koh, G., Zou, X., & Nik-Zainal, S. (2020). Mutational signatures: Experimental design and analytical framework. In *Genome Biology* (Vol. 21, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-020-1951-5

Kondo, Y., Shen, L., Suzuki, S., Kurokawa, T., Masuko, K., Tanaka, Y., Kato, H., Mizuno, Y., Yokoe, M., Sugauchi, F., Hirashima, N., Orito, E., Osada, H., Ueda, R., Guo, Y., Chen, X., Issa, J. P. J., & Sekido, Y. (2007). Alterations of DNA methylation and histone modifications contribute to gene silencing in hepatocellular carcinomas. *Hepatology Research*, *37*(11), 974–983. https://doi.org/10.1111/j.1872-034X.2007.00141.x

Kouzarides, T. (2007). Chromatin Modifications and Their Function. In *Cell* (Vol. 128, Issue 4, pp. 693–705). https://doi.org/10.1016/j.cell.2007.02.005

Kryuchkova-Mostacci, N., & Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, *18*(2), 205–214. https://doi.org/10.1093/bib/bbw008

Kübler, K., Karlić, R., Haradhvala, N. J., Ha, K., Kim, J., Kuzman, M., Jiao, W., Gakkhar, S., Mouw, K. W., Braunstein, L. Z., Elemento, O., Biankin, A. V, Rooman, I., Miller, M., Karthaus, W. R., Nogiec, C. D., Juvenson, E., Curry, E., Mino-Kenudson, M., … Getz, G. (2019). Tumor mutational landscape is a record of the pre-malignant state. *BioRxiv*. https://doi.org/10.1101/517565

Kuhn, M. (2008). *Journal of Statistical Software Building Predictive Models in R Using the caret Package*. http://www.jstatsoft.org/

Land, H., Parada, L. F., & Weinberg, R. A. (1983). Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. *Nature*, *304*(5927), 596–602. https://doi.org/https://doi.org/10.1038/304596a0

Lander, S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … Yeh, R.-F. (2001). Initial sequencing and analysis of the human genome. *NATURE*, *409*. www.nature.com

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1113

Lane, D. P. (1992). p53, guardian of the genome. *Nature*, *358*(6381), 15–16. https://doi.org/10.1038/358015a0

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Law, E. K., Sieuwerts, A. M., Lapara, K., Leonard, B., Starrett, G. J., Molan, A. M., Temiz, N. A., Vogel, R. I., Meijer-Van Gelder, M. E., Sweep, F. C. G. J., Span, P. N., Foekens, J. A., Martens, J. W. M., Yee, D., & Harris, R. S. (2016). The DNA cytosine deaminase

APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer. *Science Advances*, *2*(10). https://www.science.org

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218. https://doi.org/10.1038/nature12213

Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B. R., Aulakh, L. K., Lu, S., Kemberling, H., Wilt, C., Luber, B. S., Wong, F., Azad, N. S., Rucki, A. A., Laheru, D., Donehower, R., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., … Diaz, L. A. (2017). Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*, *357*, 409–413. https://www.science.org

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., Biedrzycki, B., Donehower, R. C., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Duffy, S. M., Goldberg, R. M., de la Chapelle, A., … Diaz, L. A. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, *372*(26), 2509–2520. https://doi.org/10.1056/nejmoa1500596

Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J. F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., Bioulac-Sage, P., Prévôt, S., Azoulay, D., Paradis, V., Imbeaud, S., Deleuze, J. F., & Zucman-Rossi, J. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-01358-x

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Y., Huang, J., Zhao, Y. L., He, J., Wang, W., Davies, K. E., Nosé, V., & Xiao, S. (2007). UTRN on chromosome 6q24 is mutated in multiple tumors. *Oncogene*, *26*(42), 6220–6228. https://doi.org/10.1038/sj.onc.1210432

Li, Y., & Luo, Y. (2020). Performance-weighted-voting model: an ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quantitative Biology*, *8*(4), 347–358. https://doi.org/10.1007/s40484-020-0226-1

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., & Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322. https://doi.org/10.1038/nature08514

Liu, T., Porter, J., Zhao, C., Zhu, H., Wang, N., Sun, Z., Mo, Y. Y., & Wang, Z. (2019). TADKB: Family classification and a knowledge base of topologically associating domains. *BMC Genomics*, *20*(1). https://doi.org/10.1186/s12864-019-5551-2

Liu, X., Li, L., Peng, L., Wang, B., Lang, J., Lu, Q., Zhang, X., Sun, Y., Tian, G., Zhang, H., & Zhou, L. (2020). Predicting Cancer Tissue-of-Origin by a Machine Learning Method Using DNA Somatic Mutation Data. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00674

Long, H. S., Greenaway, S., Powell, G., Mallon, A. M., Lindgren, C. M., & Simon, M. M. (2022). Making sense of the linear genome, gene function and TADs. *Epigenetics and Chromatin*, *15*(1). https://doi.org/10.1186/s13072-022-00436-9

Long, Q., Yuan, Y., & Li, M. (2022). RNA-SSNV: A Reliable Somatic Single Nucleotide Variant Identification Framework for Bulk RNA-Seq Data. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.865313

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. In *Nature Genetics* (Vol. 45, Issue 6, pp. 580–585). https://doi.org/10.1038/ng.2653

Lord, C. J., & Ashworth, A. (2016). BRCAness revisited. *Nature Reviews Cancer*, *16*(2), 110–120. https://doi.org/10.1038/nrc.2015.21

Lu, Y., Chan, Y. T., Tan, H. Y., Li, S., Wang, N., & Feng, Y. (2020). Epigenetic regulation in human cancer: The potential role of epi-drug in cancer therapy. In *Molecular Cancer* (Vol. 19, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12943-020-01197-3

Luger, K., Mä Der, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A ˚ resolution. *Nature, 389.*

Lüleci, H. B., & Yılmaz, A. (2022). Robust and rigorous identification of tissue-specific genes by statistically extending tau score. *BioData Mining*, *15*(1). https://doi.org/10.1186/s13040-022-00315-9

Mandelker, D., & Ceyhan-Birsoy, O. (2020). Evolving Significance of Tumor-Normal Sequencing in Cancer Care. In *Trends in Cancer* (Vol. 6, Issue 1, pp. 31–39). Cell Press. https://doi.org/10.1016/j.trecan.2019.11.006

Manders, F., Brandsma, A. M., de Kanter, J., Verheul, M., Oka, R., van Roosmalen, M. J., van der Roest, B., van Hoeck, A., Cuppen, E., & van Boxtel, R. (2022). MutationalPatterns: the one stop shop for the analysis of mutational processes. *BMC Genomics*, *23*(1). https://doi.org/10.1186/s12864-022-08357-3

Mantovani, F., Collavin, L., & Del Sal, G. (2019). Mutant p53 as a guardian of the cancer cell. In *Cell Death and Differentiation* (Vol. 26, Issue 2, pp. 199–212). Nature Publishing Group. https://doi.org/10.1038/s41418-018-0246-9

Manzano, J. L., Layos, L., Bugés, C., De los Llanos Gil, M., Vila, L., Martínez-Balibrea, E., & Martínez-Cardús, A. (2016). Resistant mechanisms to BRAF inhibitors in melanoma. *Annals of Translational Medicine*, *4*(12), 1–9. https://doi.org/10.21037/atm.2016.06.07

Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., Ferté, C., Jamal-Hanjani, M., Wilson, G. A., Shafi, S., Swanton, C., André, F., Szallasi, Z., & Eklund, A. C. (2015). TumorTracer: A method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Medical Genomics*, *8*(1). https://doi.org/10.1186/s12920-015-0130-0

Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., Royo, R., Ziccheddu, B., Puente, X. S., Avet-Loiseau, H., Cambell, P. J., Nik-Zainal, S., Campo, E., Munshi, N., & Bolli, N. (2019). A practical guide for mutational signature analysis in hematological malignancies. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-11037-8

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. In *Biochemistry* (Vol. 74, Issue 2).

Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

McArthur, E., & Capra, J. A. (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *American Journal of Human Genetics*, *108*(2), 269–283. https://doi.org/10.1016/j.ajhg.2021.01.001

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meienberg, J., Zerjavic, K., Keller, I., Okoniewski, M., Patrignani, A., Ludin, K., Xu, Z., Steinmann, B., Carrel, T., Röthlisberger, B., Schlapbach, R., Bruggmann, R., & Matyas, G. (2015). New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Research*, *43*(11). https://doi.org/10.1093/nar/gkv216

Mercier-Darty, M., Boutolleau, D., Rodriguez, C., & Burrel, S. (2019). Added value of ultra-deep sequencing (UDS) approach for detection of genotypic antiviral resistance of herpes simplex virus (HSV). *Antiviral Research*, *168*, 128–133. https://doi.org/10.1016/j.antiviral.2019.05.017

Meyerson, W., Meyerson, W., Leisman, J., Navarro, F. C. P., Navarro, F. C. P., & Gerstein, M. (2020). Origins and characterization of variants shared between databases of somatic and germline human mutations. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-3508-8

Mirmohammadsadegh, A., Marini, A., Nambiar, S., Hassan, M., Tannapfel, A., Ruzicka, T., & Hengge, U. R. (2006). Epigenetic silencing of the PTEN gene in melanoma. *Cancer Research*, *66*(13), 6546–6552. https://doi.org/10.1158/0008-5472.CAN-06-0384

Misetic, H., Keddar, M. R., Jeannon, J. P., & Ciccarelli, F. D. (2023). Mechanistic insights into the interactions between cancer drivers and the tumour immune microenvironment. *Genome Medicine*, *15*(1). https://doi.org/10.1186/s13073-023-01197-0

Mittal, P., & Roberts, C. W. M. (2020). The SWI/SNF complex in cancer — biology, biomarkers and therapy. In *Nature Reviews Clinical Oncology* (Vol. 17, Issue 7, pp. 435–448). Nature Research. https://doi.org/10.1038/s41571-020-0357-3

National Cancer Institute. (2024). *NCI Dictionary of Cancer Terms* . https://www.cancer.gov/publications/dictionaries/cancer-terms

National Human Genome Research Institute. (2024, March 14). *The Human Genome Project (HGP)*. https://www.genome.gov/human-genome-project

Nepali, K., & Liou, J. P. (2021). Recent developments in epigenetic cancer therapeutics: clinical advancement and emerging trends. In *Journal of Biomedical Science* (Vol. 28, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12929-021-00721-x

Nesta, A. V., Tafur, D., & Beck, C. R. (2021). Hotspots of Human Mutation. In *Trends in Genetics* (Vol. 37, Issue 8, pp. 717–729). Elsevier Ltd. https://doi.org/10.1016/j.tig.2020.10.003

Newell, F., Kong, Y., Wilmott, J. S., Johansson, P. A., Ferguson, P. M., Cui, C., Li, Z., Kazakoff, S. H., Burke, H., Dodds, T. J., Patch, A. M., Nones, K., Tembe, V., Shang, P., van der Weyden, L., Wong, K., Holmes, O., Lo, S., Leonard, C., … Scolyer, R. A. (2019). Whole-genome landscape of mucosal melanoma reveals diverse drivers and therapeutic targets. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-11107-x

Ng, C. K. Y., Martelotto, L. G., Gauthier, A., Wen, H. C., Piscuoglio, S., Lim, R. S., Cowell, C. F., Wilkerson, P. M., Wai, P., Rodrigues, D. N., Arnould, L., Geyer, F. C., Bromberg, S. E., Lacroix-Triki, M., Penault-Llorca, F., Giard, S., Sastre-Garau, X., Natrajan, R., Norton, L., … Reis-Filho, J. S. (2015). Intra-tumor genetic heterogeneity and alternative driver genetic alterations in breast cancers with heterogeneous HER2 gene amplification. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0657-6

Nguyen, L., Van Hoeck, A., & Cuppen, E. (2022). Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-31666-w

Nicolussi, A., Belardinilli, F., Mahdavian, Y., Colicchia, V., D'Inzeo, S., Petroni, M., Zani, M., Ferraro, S., Valentini, V., Ottini, L., Giannini, G., Capalbo, C., & Coppa, A. (2019). Next-generation sequencing of BRCA1 and BRCA2 genes for rapid detection of germline mutations in hereditary breast/ovarian cancer. *PeerJ*, *7*. https://doi.org/10.7717/peerj.6661

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., … Stratton, M. R. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, *149*(5), 979–993. https://doi.org/10.1016/j.cell.2012.04.024

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjærde, O. C., Langerød, A., Ringnér, M., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54. https://doi.org/10.1038/nature17676

Nik-Zainal, S., Kucab, J. E., Morganella, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., Weninger, A., Hollstein, M., Stratton, M. R., & Phillips, D. H. (2015). The genome as a record of environmental exposure. *Mutagenesis*, *30*(6), 763–770. https://doi.org/10.1093/mutage/gev073

NobelPrize.org. (2024). *The Nobel Prize in Chemistry 1980*. Nobel Prize Outreach AB. https://www.nobelprize.org/prizes/chemistry/1980/summary/

Nourbakhsh, M., Degn, K., Saksager, A. B., Tiberti, M., & Papaleo, E. (2024). Prediction of cancer driver genes and mutations: the potential of integrative computational frameworks. In *Briefings in Bioinformatics* (Vol. 25, Issue 2). Oxford University Press. https://doi.org/10.1093/bib/bbad519

Ocsenas, O., & Reimand, J. (2022). Chromatin accessibility of primary human cancers ties regional mutational processes and signatures with tissues of origin. *PLoS Computational Biology*, *18*(8). https://doi.org/10.1371/journal.pcbi.1010393

Oh, E., & Lee, H. (2023). Transcriptomic data in tumor-adjacent normal tissues harbor prognostic information on multiple cancer types. *Cancer Medicine*, *12*(10), 11960–11970. https://doi.org/10.1002/cam4.5864

Pandey, P., Arora, S., & Rosen, G. L. (2022). MetaMutationalSigs: comparison of mutational signature refitting results made easy. *Bioinformatics*, *38*(8), 2344–2347. https://doi.org/10.1093/bioinformatics/btac091

Patel, A. (2020). Benign vs Malignant Tumors. In *JAMA Oncology* (Vol. 6, Issue 9, p. 1488). American Medical Association. https://doi.org/10.1001/jamaoncol.2020.2592

Pavlidis, N., Briasoulis, E., Hainsworth, J., & Greco, F. A. (2003). Diagnostic and therapeutic management of cancer of an unknown primary. *European Journal of Cancer*, *39*(14), 1990–2005. https://doi.org/10.1016/S0959-8049(03)00547-1

Pavlidis, N., & Fizazi, K. (2009). Carcinoma of unknown primary (CUP). In *Critical Reviews in Oncology/Hematology* (Vol. 69, Issue 3, pp. 271–278). https://doi.org/10.1016/j.critrevonc.2008.09.005

Petljak, M., & Alexandrov, L. B. (2016). Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*, *37*(6), 531–540. https://doi.org/10.1093/carcin/bgw055

Petljak, M., Dananberg, A., Chu, K., Bergstrom, E. N., Striepen, J., von Morgen, P., Chen, Y., Shah, H., Sale, J. E., Alexandrov, L. B., Stratton, M. R., & Maciejowski, J. (2022). Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature*, *607*(7920), 799–807. https://doi.org/10.1038/s41586-022-04972-y

Pfeifer, G. P., Denissenko, M. F., Olivier, M., Tretyakova, N., Hecht, S. S., & Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, *21*, 7435–7451. https://doi.org/10.1038/sj.onc

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-29325-6

Piñero, J., Bravo, Á., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., & Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, *45*(D1), D833–D839. https://doi.org/10.1093/nar/gkw943

Pisani, P. (2011). The cancer burden and cancer control in developing countries. *Environmental Health: A Global Access Science Source*, *10*(SUPPL. 1). https://doi.org/10.1186/1476-069X-10-S1-S2

Polak, P., Karlic, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M. S., Reynolds, A., Rynes, E., Vlahovicek, K., Stamatoyannopoulos, J. A., & Sunyaev, S. R. (2015). Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, *518*(7539), 360–364. https://doi.org/10.1038/nature14221

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K. W., Kamburov, A., Maruvka, Y. E., Leshchiner, I., Lander, E. S., Golub, T. R., Zick, A., Orthwein, A., Lawrence, M. S., Batra, R. N., … Getz, G. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*, *49*(10), 1476–1486. https://doi.org/10.1038/ng.3934

Pon, J. R., & Marra, M. A. (2015). Driver and passenger mutations in cancer. *Annual Review of Pathology: Mechanisms of Disease*, *10*, 25–50. https://doi.org/10.1146/annurev-pathol-012414-040312

Pott, S., & Lieb, J. D. (2015). What are super-enhancers? In *Nature Genetics* (Vol. 47, Issue 1, pp. 8–12). Nature Publishing Group. https://doi.org/10.1038/ng.3167

Pruitt, K. D., Tatusova, T., Klimke, W., & Maglott, D. R. (2009). NCBI reference sequences: Current status, policy and new initiatives. *Nucleic Acids Research*, *37*(SUPPL. 1). https://doi.org/10.1093/nar/gkn721

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biology and Medicine*, *16*(1), 4–10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055

Rajderkar, S., Barozzi, I., Zhu, Y., Hu, R., Zhang, Y., Li, B., Alcaina Caro, A., Fukuda-Yuzawa, Y., Kelman, G., Akeza, A., Blow, M. J., Pham, Q., Harrington, A. N., Godoy, J., Meky, E. M., von Maydell, K., Hunter, R. D., Akiyama, J. A., Novak, C. S., … Pennacchio, L. A. (2023). Topologically associating domain boundaries are required for normal genome function. *Communications Biology*, *6*(1). https://doi.org/10.1038/s42003-023-04819-w

Ramaswami, G., & Li, J. B. (2014). RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt996

Reddy, B. Y., Miller, D. M., & Tsao, H. (2017). Somatic driver mutations in melanoma. In *Cancer* (Vol. 123, pp. 2104–2117). John Wiley and Sons Inc. https://doi.org/10.1002/cncr.30593

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317–329. https://doi.org/10.1038/nature14248

Roberts, S. A., Sterling, J., Thompson, C., Harris, S., Mav, D., Shah, R., Klimczak, L. J., Kryukov, G. V., Malc, E., Mieczkowski, P. A., Resnick, M. A., & Gordenin, D. A. (2012). Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Molecular Cell, 46*(4), 424–435. https://doi.org/10.1016/j.molcel.2012.03.030

Ruscito, I., Dimitrova, D., Vasconcelos, I., Gellhaus, K., Schwachula, T., Bellati, F., Zeillinger, R., Benedetti-Panici, P., Vergote, I., Mahner, S., Cacsire-Tong, D., Concin, N., Darb-Esfahani, S., Lambrechts, S., Sehouli, J., Olek, S., & Braicu, E. I. (2014). BRCA1 gene promoter methylation status in high-grade serous ovarian cancer patients - A study of the tumour Bank ovarian cancer (TOC) and ovarian cancer diagnosis consortium (OVCAD). *European Journal of Cancer*, *50*(12), 2090–2098. https://doi.org/10.1016/j.ejca.2014.05.001

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). *DNA sequencing with chain-terminating inhibitors* (Vol. 74, Issue 12).

Sasidharan Nair, V., El Salhat, H., Taha, R. Z., John, A., Ali, B. R., & Elkord, E. (2018). DNA methylation and repressive H3K9 and H3K27 trimethylation in the promoter regions of PD-1, CTLA-4, TIM-3, LAG-3, TIGIT, and PD-L1 genes in human primary breast cancer. *Clinical Epigenetics, 10*(1). https://doi.org/10.1186/s13148-018-0512-1

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics, 28*(14), 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Savva, Y. A., Rieder, L. E., & Reenan, R. A. (2012). The ADAR protein family. In *Genome biology* (Vol. 13, Issue 12, p. 252). https://doi.org/10.1186/gb-2012-13-12-252

Schoenfelder, S., & Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. In *Nature Reviews Genetics* (Vol. 20, Issue 8, pp. 437–455). Nature Research. https://doi.org/10.1038/s41576-019-0128-0

Schofield, J. B., Wells, T., Trust, N., Oien, K., & Wilkins, B. (2018). *Standards and datasets for reporting cancers Dataset for histopathological reporting of cancer of unknown primary (CUP) and malignancy of unknown primary origin (MUO) Unique document number G167 Document name Dataset for histopathological reporting of cancer of unknown primary (CUP) and malignancy of unknown origin (MUO)*.

Seligson, D. B., Horvath, S., Shi, T., Yu, H., Tze, S., Grunstein, M., & Kurdistani, S. K. (2005). Global histone modification patterns predict risk of prostate cancer recurrence. *Nature, 435*(7046), 1262–1266. https://doi.org/10.1038/nature03672

Sharma, S., Kelly, T. K., & Jones, P. A. (2009). Epigenetics in cancer. In *Carcinogenesis* (Vol. 31, Issue 1, pp. 27–36). https://doi.org/10.1093/carcin/bgp220

Shaw-Smith, C., Redon, R., Rickman, L., Rio, M., Willatt, L., Fiegler, H., Firth, H., Sanlaville, D., Winter, R., Colleaux, L., Bobrow, M., & Carter, N. P. (2004). Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *Journal of Medical Genetics*, *41*(4), 241–248. https://doi.org/10.1136/jmg.2003.017731

Shinde, J., Bayard, Q., Imbeaud, S., Hirsch, T. Z., Liu, F., Renault, V., Zucman-Rossi, J., & Letouzé, E. (2018). Palimpsest: An R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*, *34*(19), 3380–3381. https://doi.org/10.1093/bioinformatics/bty388

Sieuwerts, A. M., Willis, S., Burns, M. B., Look, M. P., Gelder, M. E. M. Van, Schlicker, A., Heideman, M. R., Jacobs, H., Wessels, L., Leyland-Jones, B., Gray, K. P., Foekens, J. A., Harris, R. S., & Martens, J. W. M. (2014). Elevated APOBEC3B Correlates with Poor Outcomes for Estrogen-Receptor-Positive Breast Cancers. *Hormones and Cancer*, *5*(6), 405–413. https://doi.org/10.1007/s12672-014-0196-8

Singh, R. R. (2022). Target Enrichment Approaches for Next-Generation Sequencing Applications in Oncology. In *Diagnostics* (Vol. 12, Issue 7). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/diagnostics12071539

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H., & Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*, 674–679.

Staaf, J., Glodzik, D., Bosch, A., Vallon-Christersson, J., Reuterswärd, C., Häkkinen, J., Degasperi, A., Amarante, T. D., Saal, L. H., Hegardt, C., Stobart, H., Ehinger, A., Larsson, C., Rydén, L., Loman, N., Malmberg, M., Kvist, A., Ehrencrona, H., Davies, H. R., … Nik-Zainal, S. (2019). Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nature Medicine*, *25*(10), 1526–1533. https://doi.org/10.1038/s41591-019-0582-4

Štancl, P., Hamel, N., Sigel, K. M., Foulkes, W. D., Karlić, R., & Polak, P. (2022). The Great Majority of Homologous Recombination Repair-Deficient Tumors Are Accounted for by Established Causes. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.852159

Štancl, P., & Karlić, R. (2023). Machine learning for pan-cancer classification based on RNA sequencing data. In *Frontiers in Molecular Biosciences* (Vol. 10). Frontiers Media SA. https://doi.org/10.3389/fmolb.2023.1285795

Stephenson, W., Razaghi, R., Busan, S., Weeks, K. M., Timp, W., & Smibert, P. (2022). Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. *Cell Genomics*, *2*(2). https://doi.org/10.1016/j.xgen.2022.100097

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. In *Nature* (Vol. 458, Issue 7239, pp. 719–724). https://doi.org/10.1038/nature07943

Sun, Y., Zhu, S., Ma, K., Liu, W., Yue, Y., Hu, G., Lu, H., & Chen, W. (2019). Identification of 12 cancer types through genome deep learning. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-53989-3

Swanton, C., McGranahan, N., Starrett, G. J., & Harris, R. S. (2015). APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. In *Cancer discovery* (Vol. 5, Issue 7, pp. 704–712). https://doi.org/10.1158/2159-8290.CD-15-0344

Taberlay, P. C., Achinger-Kawecka, J., Lun, A. T. L., Buske, F. A., Sabir, K., Gould, C. M., Zotenko, E., Bert, S. A., Giles, K. A., Bauer, D. C., Smyth, G. K., Stirzaker, C., O'Donoghue, S. I., & Clark, S. J. (2016). Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Research*, *26*(6), 719–731. https://doi.org/10.1101/gr.201517.115

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

Taylor, B. J. M., Nik-Zainal, S., Wu, Y. L., Stebbings, L. A., Raine, K., Campbell, P. J., Rada, C., Stratton, M. R., & Neuberger, M. S. (2013). DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *ELife*, *2013*(2). https://doi.org/10.7554/eLife.00534

Temko, D., Tomlinson, I. P. M., Severini, S., Schuster-Böckler, B., & Graham, T. A. (2018). The effects of mutational processes and selection on driver mutations across cancer types. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-04208-6

Theisen, A. (2008). Microarray-based Comparative Genomic Hybridization (aCGH). *Nature Education*, *2008*(1), 45.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *SUPL.43*. https://doi.org/10.1002/0471250953.bi1110s43

Van Hoeck, A., Tjoonk, N. H., Van Boxtel, R., & Cuppen, E. (2019). Portrait of a cancer: Mutational signature analyses for cancer diagnostics. In *BMC Cancer* (Vol. 19, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12885-019-5677-2

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*. http://science.sciencemag.org/

Vestergaard, L. K., Oliveira, D. N. P., Høgdall, C. K., & Høgdall, E. V. (2021). Next generation sequencing technology in the clinic and its challenges. In *Cancers* (Vol. 13, Issue 8). MDPI AG. https://doi.org/10.3390/cancers13081751

Visvader, J. E. (2011). Cells of origin in cancer. In *Nature* (Vol. 469, Issue 7330, pp. 314–322). https://doi.org/10.1038/nature09781

Wang, D., Wang, J., Ding, N., Li, Y., Yang, Y., Fang, X., & Zhao, H. (2016). MAGE-A1 promotes melanoma proliferation and migration through C-JUN activation. *Biochemical and Biophysical Research Communications*, *473*(4), 959–965. https://doi.org/10.1016/j.bbrc.2016.03.161

Wang, X., & Zhang, X. (2022). Hepatocellular adenoma: Where are we now? In *World Journal of Gastroenterology* (Vol. 28, Issue 14, pp. 1384–1393). Baishideng Publishing Group Inc. https://doi.org/10.3748/wjg.v28.i14.1384

Wang, Y., Song, C., Zhao, J., Zhang, Y., Zhao, X., Feng, C., Zhang, G., Zhu, J., Wang, F., Qian, F., Zhou, L., Zhang, J., Bai, X., Ai, B., Liu, X., Wang, Q., & Li, C. (2023). SEdb 2.0: a comprehensive super-enhancer database of human and mouse. *Nucleic Acids Research*, *51*(1), D280–D290. https://doi.org/10.1093/nar/gkac968

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M. N. K., Li, Y., Hu, M., Hardison, R., Wang, T., & Yue, F. (2018). The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1519-9

Wang, Z., Zhang, T., Wu, W., Wu, L., Li, J., Huang, B., Liang, Y., Li, Y., Li, P., Li, K., Wang, W., Guo, R., & Wang, Q. (2022). Detection and Localization of Solid Tumors Utilizing the Cancer-Type-Specific Mutational Signatures. *Frontiers in Bioengineering and Biotechnology*, *10*. https://doi.org/10.3389/fbioe.2022.883791

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., & Watson, M. (2015). Exome sequencing: Current and future perspectives. *G3: Genes, Genomes, Genetics*, *5*(8), 1543–1550. https://doi.org/10.1534/g3.115.018564

Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C., & Chinnaiyan, A. M. (2014). RNA-Seq Accurately Identifies Cancer Biomarker Signatures to Distinguish Tissue of Origin. *Neoplasia (United States)*, *16*(11), 918–927. https://doi.org/10.1016/j.neo.2014.09.007

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., … Kling, T. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Wong, J. K. L., Aichmüller, C., Schulze, M., Hlevnjak, M., Elgaafary, S., Lichter, P., & Zapatka, M. (2022). Association of mutation signature effectuating processes with mutation hotspots in driver genes and non-coding regions. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-021-27792-6

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Wu, Y., Chua, E. H. Z., Ng, A. W. T., Boot, A., & Rozen, S. G. (2022). Accuracy of mutational signature software on correlated signatures. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-021-04207-6

Xiao, T., & Zhou, W. (2020). The third generation sequencing: The advanced approach to genetic diseases. In *Translational Pediatrics* (Vol. 9, Issue 2, pp. 163–173). AME Publishing Company. https://doi.org/10.21037/TP.2020.03.06

Yang, C., Lei, C., Zhang, Y., Zhang, J., Ji, F., Pan, W., Zhang, L., Gao, H., Yang, M., Li, J., & Wang, K. (2020). Comparison of Overall Survival Between Invasive Lobular Breast Carcinoma and Invasive Ductal Breast Carcinoma: A Propensity Score Matching Study Based on SEER Database. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.590643

Yang, S., Ha, K., Song, W., Fujita, M., Kübler, K., Polak, P., Hiyama, E., Nakagawa, H., Kim, H. G., & Lee, H. (2023a). COOBoostR: An Extreme Gradient Boosting-Based Tool for Robust Tissue or Cell-of-Origin Prediction of Tumors. *Life*, *13*(1). https://doi.org/10.3390/life13010071

Yang, S., Ha, K., Song, W., Fujita, M., Kübler, K., Polak, P., Hiyama, E., Nakagawa, H., Kim, H. G., & Lee, H. (2023b). COOBoostR: An Extreme Gradient Boosting-Based Tool for Robust Tissue or Cell-of-Origin Prediction of Tumors. *Life*, *13*(1). https://doi.org/10.3390/life13010071

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang, F.-L., & Li, D.-Q. (2022). Targeting Chromatin-Remodeling Factors in Cancer Cells: Promising Molecules in Cancer Therapy. In *International Journal of Molecular Sciences* (Vol. 23, Issue 21). MDPI. https://doi.org/10.3390/ijms232112815

Zhang, H., Weng, X., Ye, J., He, L., Zhou, D., & Liu, Y. (2015). Promoter hypermethylation of TERT is associated with hepatocellular carcinoma in the Han Chinese population. *Clinics and Research in Hepatology and Gastroenterology*, *39*(5), 600–609. https://doi.org/10.1016/j.clinre.2015.01.002

Zhang, Y., Chen, F., & Creighton, C. J. (2021). SVExpress: identifying gene features altered recurrently in expression with nearby structural variant breakpoints. *BMC Bioinformatics*, *22*(1). https://doi.org/10.1186/s12859-021-04072-0

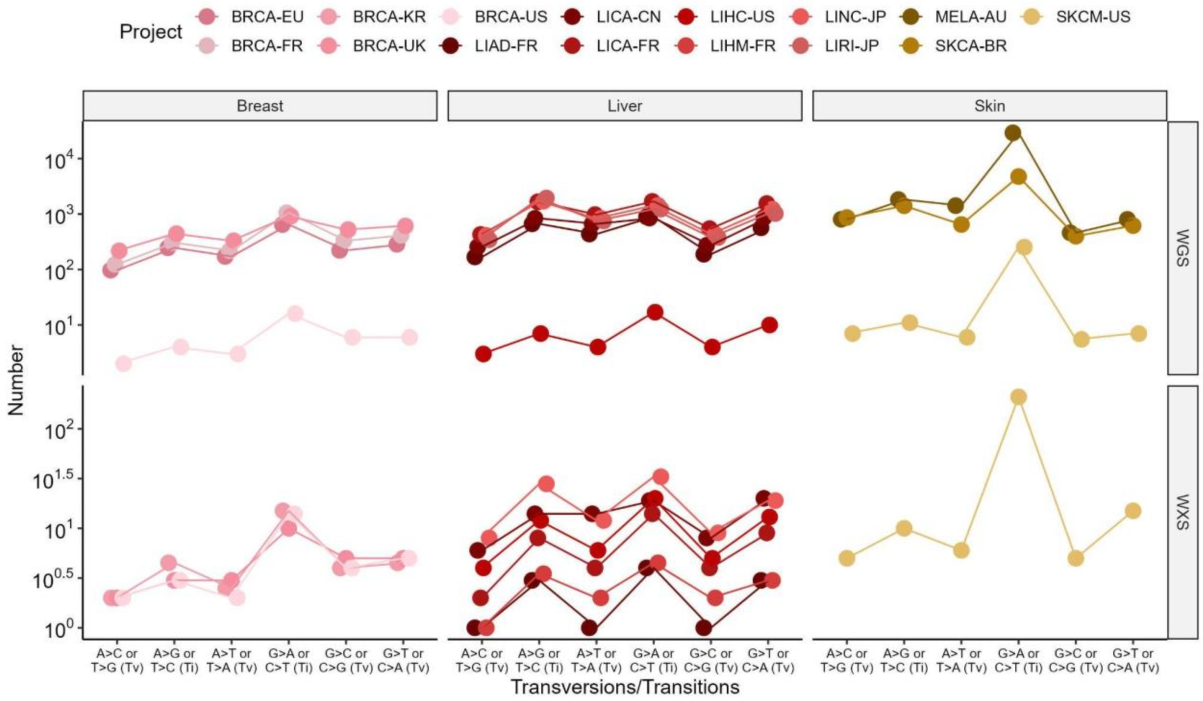
Zhang, Y., Yang, L., Kucherlapati, M., Chen, F., Hadjipanayis, A., Pantazi, A., Bristow, C. A., Lee, E. A., Mahadeshwar, H. S., Tang, J., Zhang, J., Seth, S., Lee, S., Ren, X., Song, X., Sun, H., Seidman, J., Luquette, L. J., Xi, R., … Creighton, C. J. (2018). A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Reports*, *24*(2), 515–527. https://doi.org/10.1016/j.celrep.2018.06.025

Zhao, Y., Pan, Z., Namburi, S., Pattison, A., Posner, A., Balachander, S., Paisie, C. A., Reddi, H. V., Rueter, J., Gill, A. J., Fox, S., Raghav, K. P. S., Flynn, W. F., Tothill, R. W., Li, S., Karuturi, R. K. M., & George, J. (2020). CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine*, *61*. https://doi.org/10.1016/j.ebiom.2020.103030

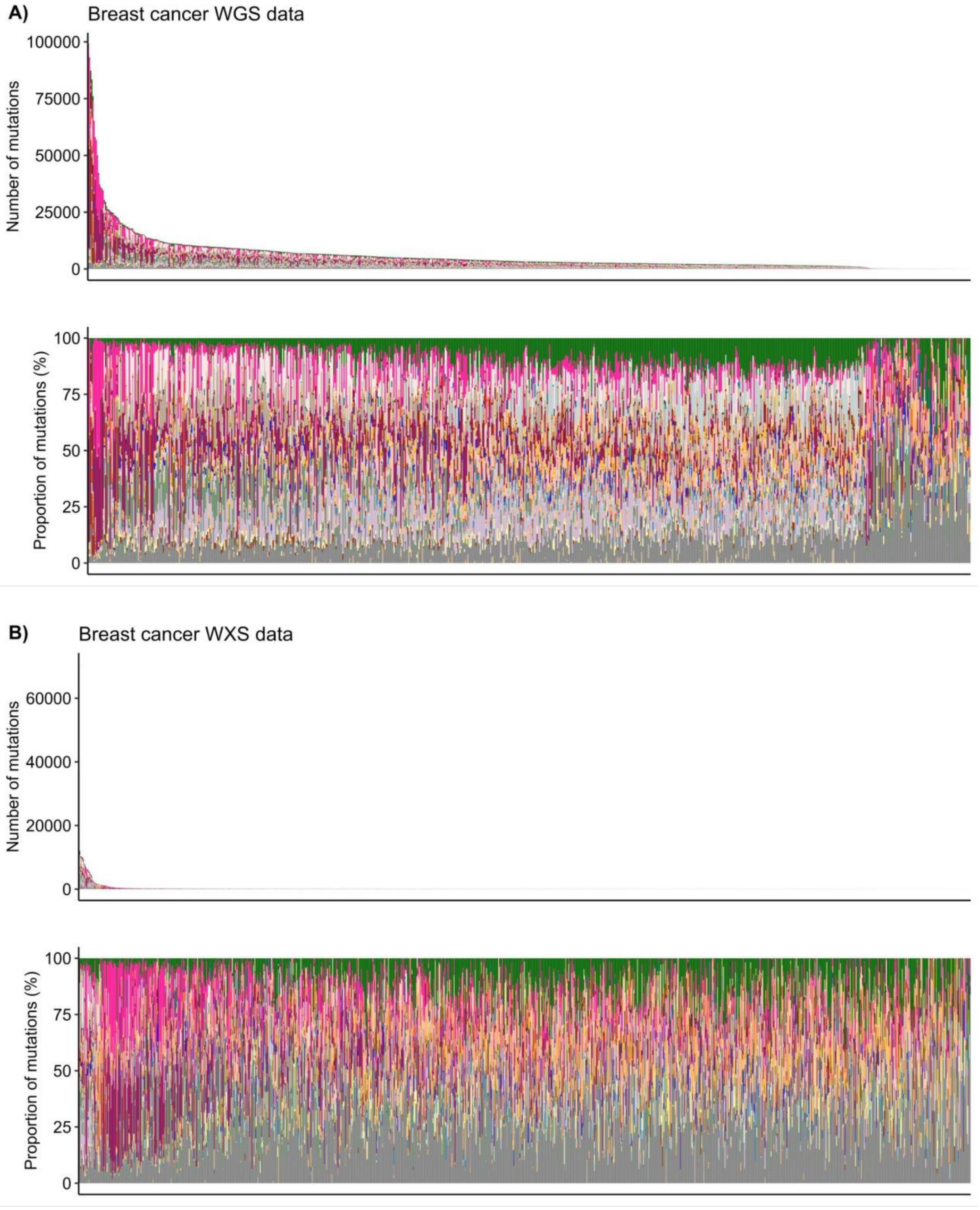
Zheng, C., & Xu, R. (2020). Predicting cancer origins with a DNA methylation-based deep neural network model. *PLoS ONE, 15*(5). https://doi.org/10.1371/journal.pone.0226461

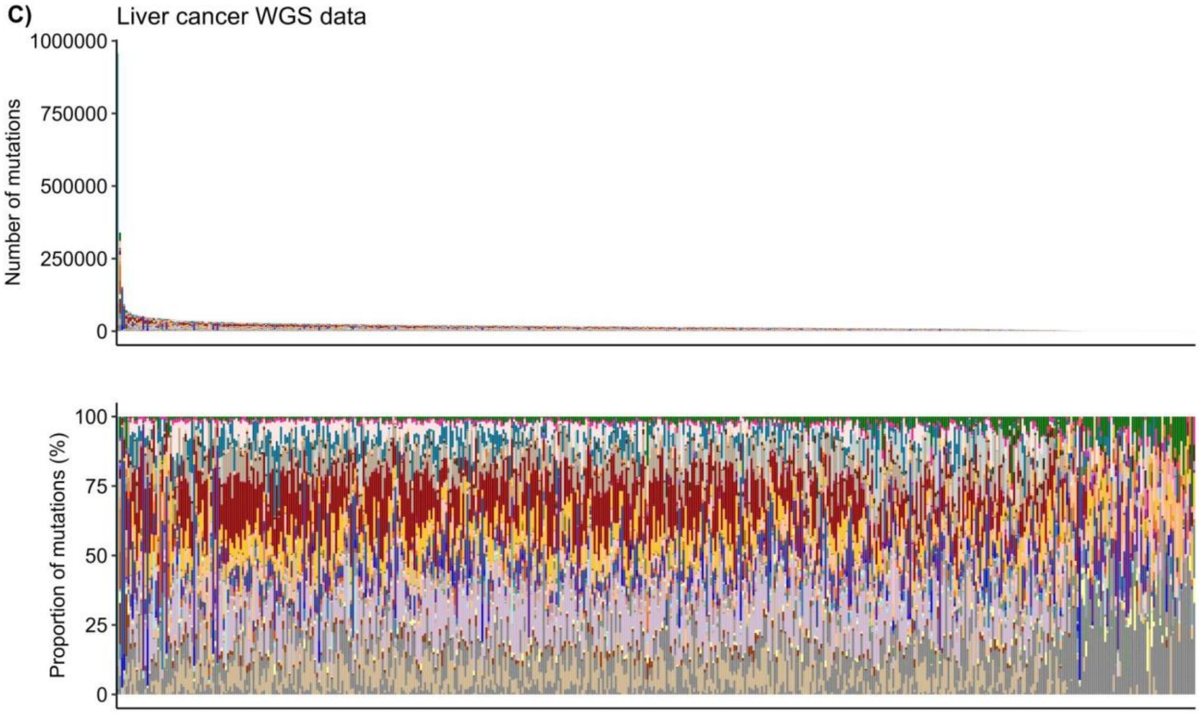
Zhu, L., Yang, X., Feng, J., Mao, J., Zhang, Q., He, M., Mi, Y., Mei, Y., Jin, G., & Zhang, H. (2022). CYP2E1 plays a suppressive role in hepatocellular carcinoma by regulating Wnt/Dvl2/β-catenin signaling. *Journal of Translational Medicine*, *20*(1). https://doi.org/10.1186/s12967-022-03396-6

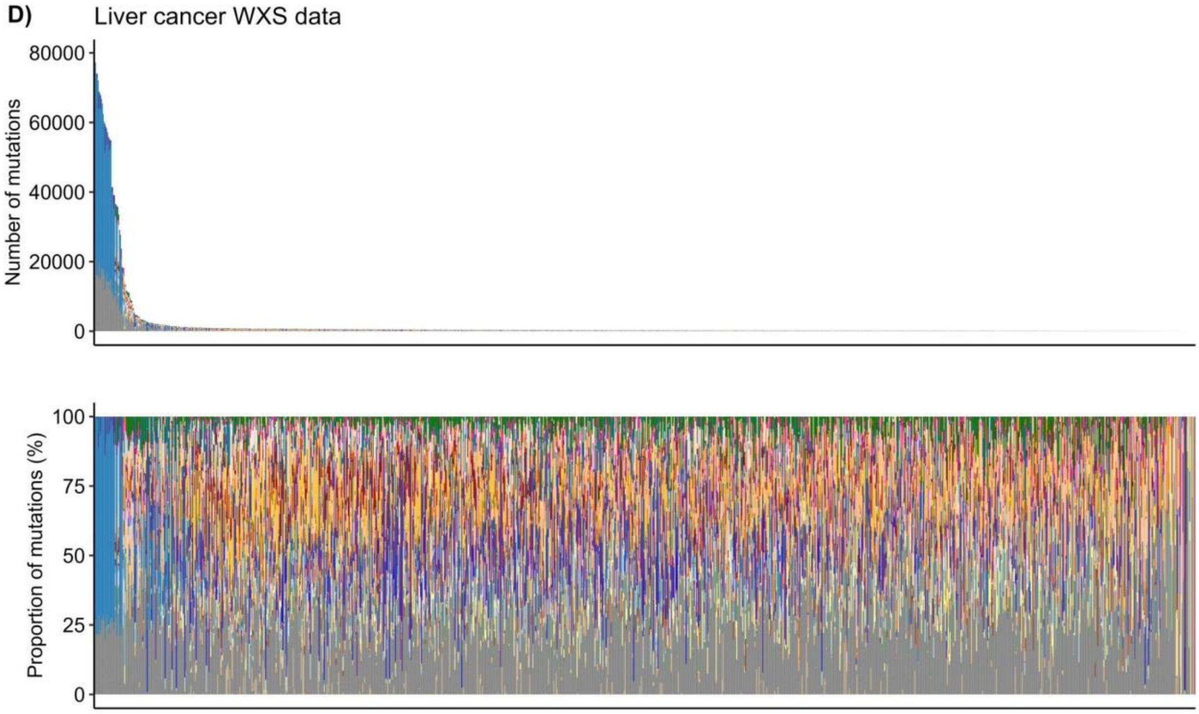
8 Supplementary

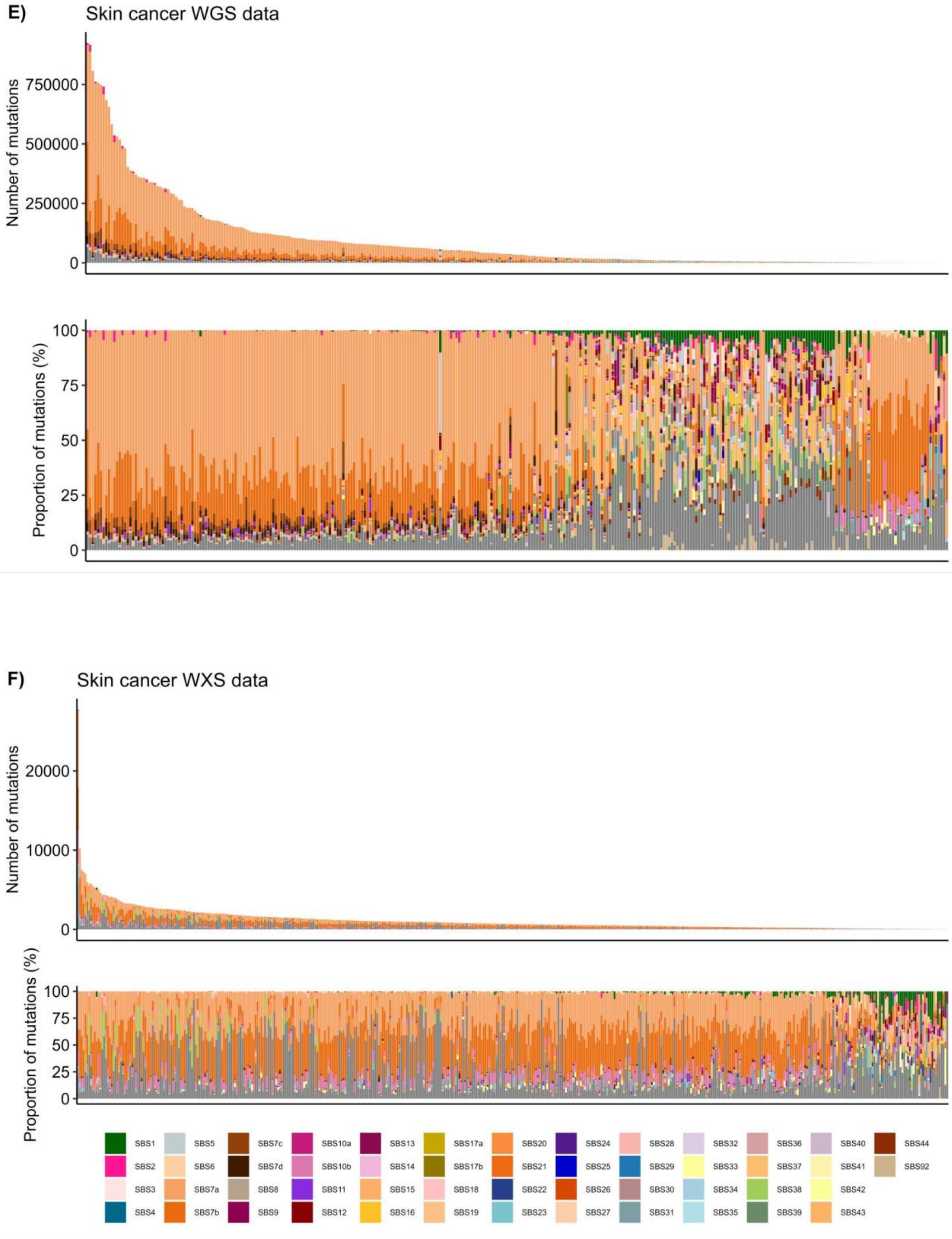


Supplementary figure 1. *Median transversions and transitions across all breast, liver and skin melanoma cancer type cohorts*

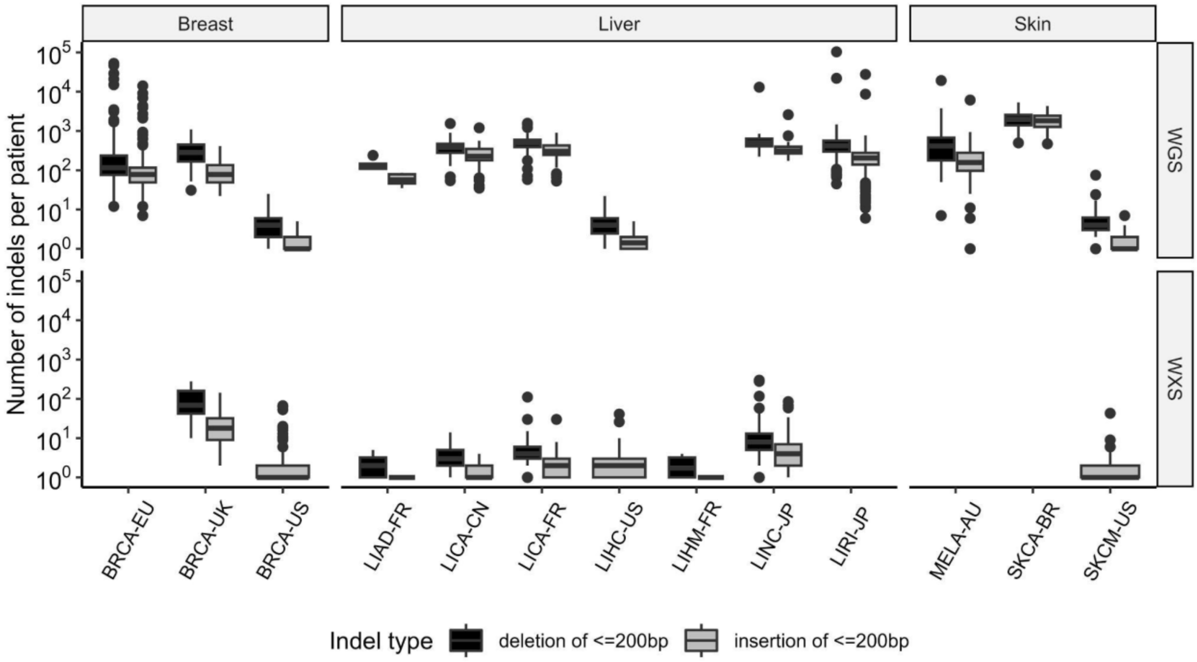

A)
Breast cancer WGS data
Number of mutations
100000
75000
50000
25000
0
Proportion of mutations (%)
100
75
50
25
0
B)
Breast cancer WXS data
Number of mutations
60000
40000
20000
0
Proportion of mutations (%)
100
75
50
25
0


C)
Liver cancer WGS data
Number of mutations
1000000
750000
500000
250000
0
Proportion of mutations (%)
100
75
50
25
0

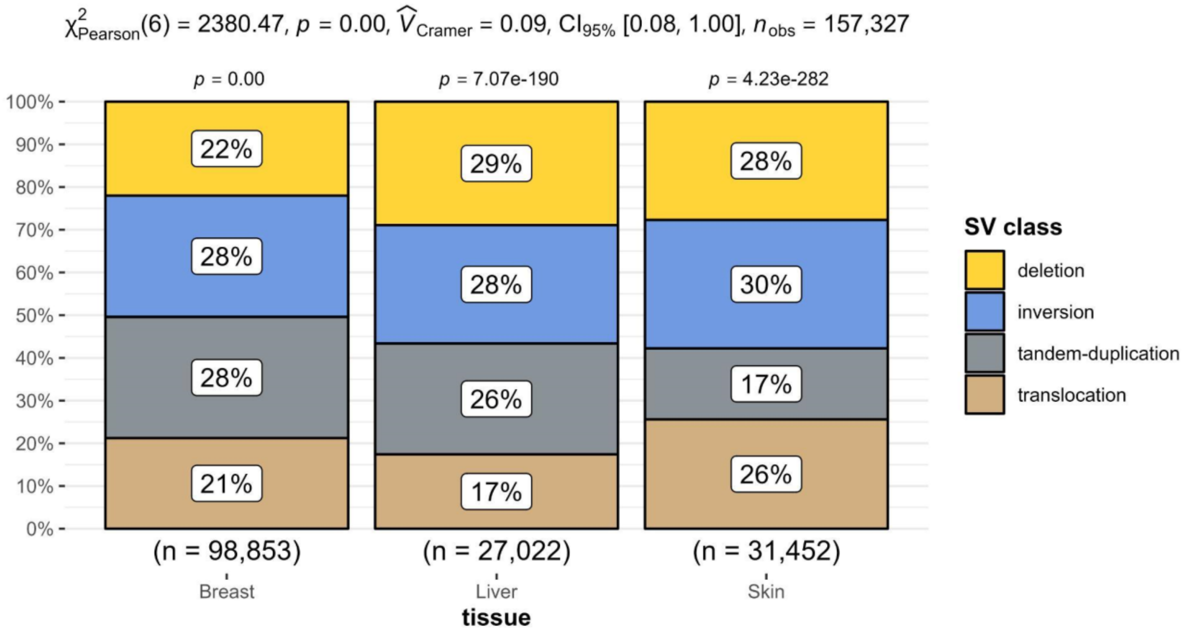

D)
Liver cancer WXS data
Number of mutations
80000
60000
40000
20000
0
Proportion of mutations (%)
100
75
50
25
0



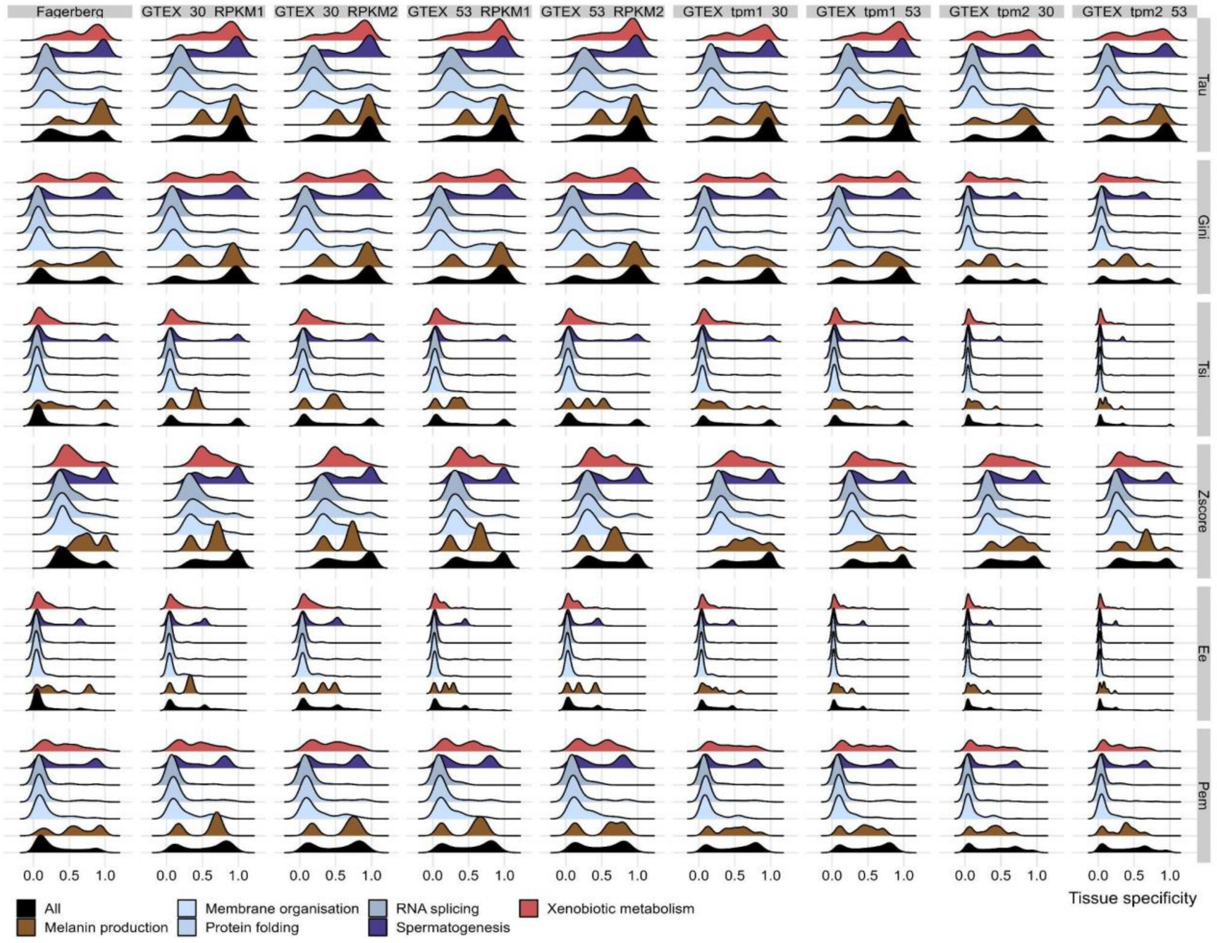
Supplementary figure 2. *A-F) Summary of somatic point mutations from different cancer types (breast, liver and skin cancer) from WGS or WXS data. The top panel contains the total number of point mutations in the samples. The bottom panel shows the somatic point-mutation signature compositions of the triplet mutational spectra. Samples are ordered by the total amount of mutations.*



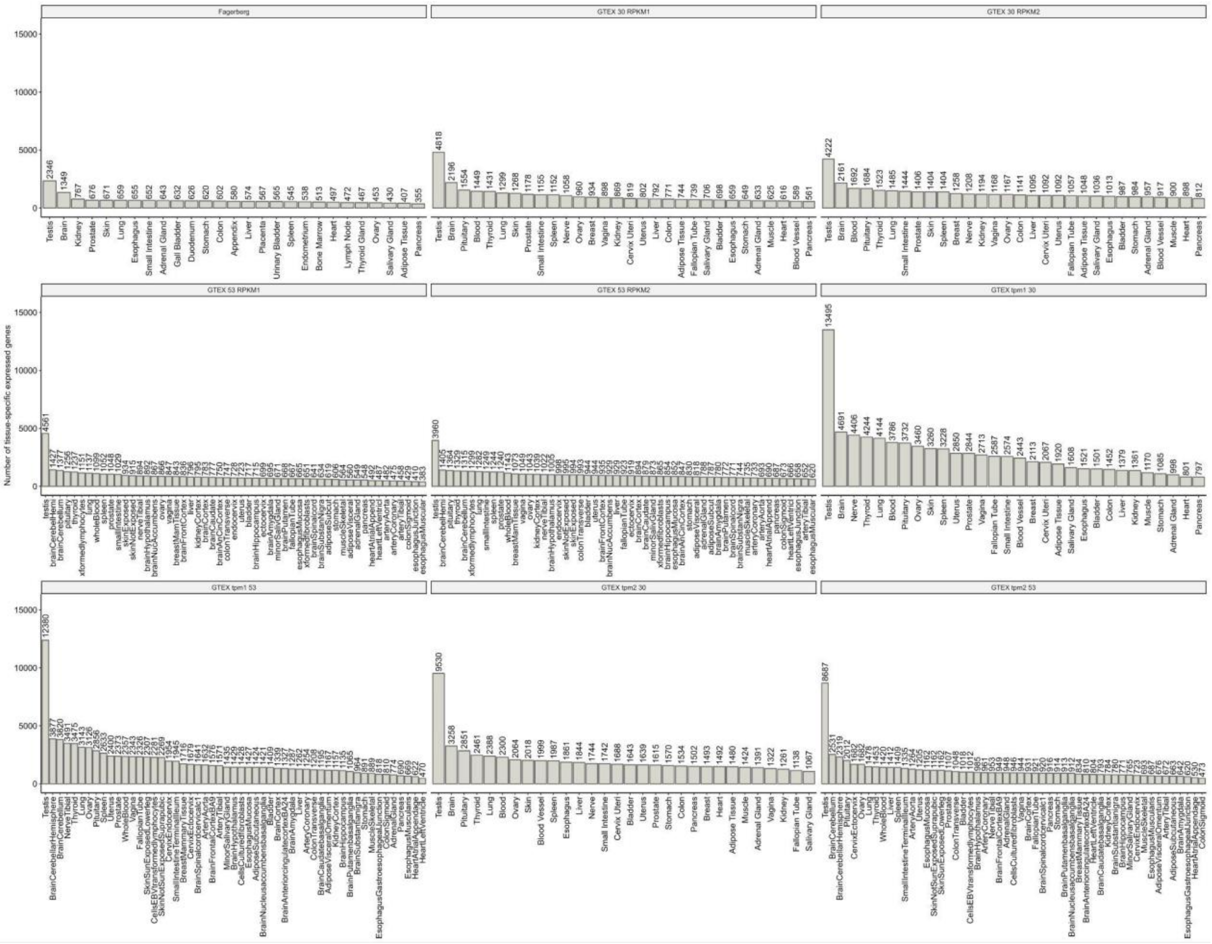
Supplementary figure 3. *Indel type per cohort across different cancer types and sequencing technologies. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR.*



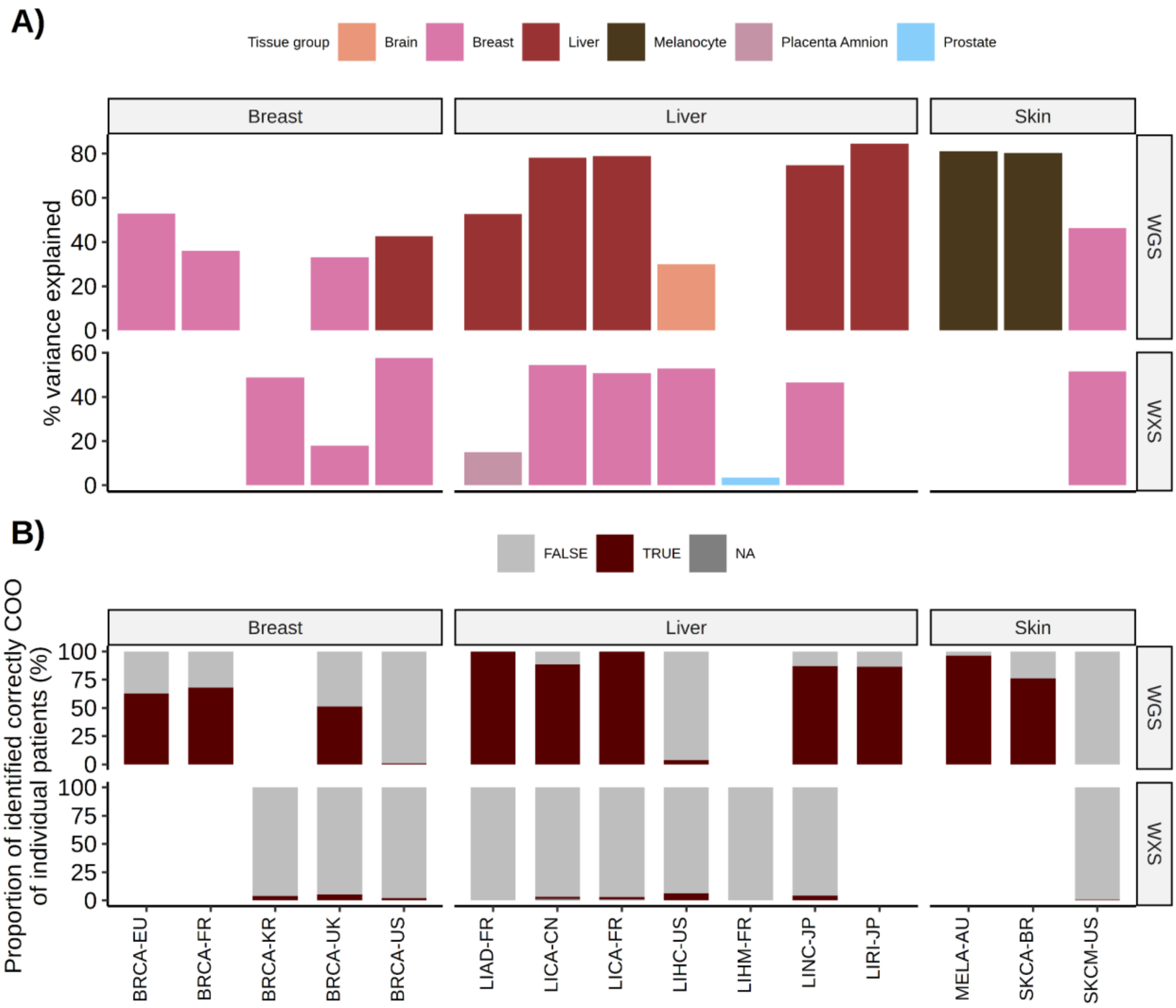
Supplementary figure 4. *Proportion of SV classes per breast, liver and skin cancer (Chi-square test, p-value = 0)*



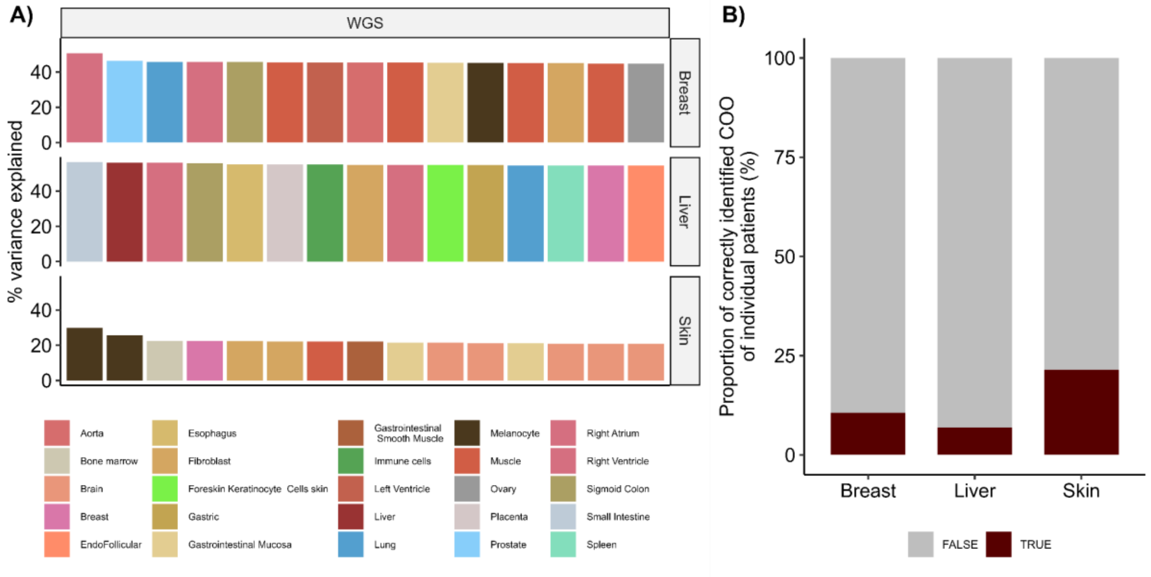
Supplementary figure 5. *Multiple tissue specific metric across all analyzed datasets separated by on tissue specific and broad Gene Ontology (GO) terms*



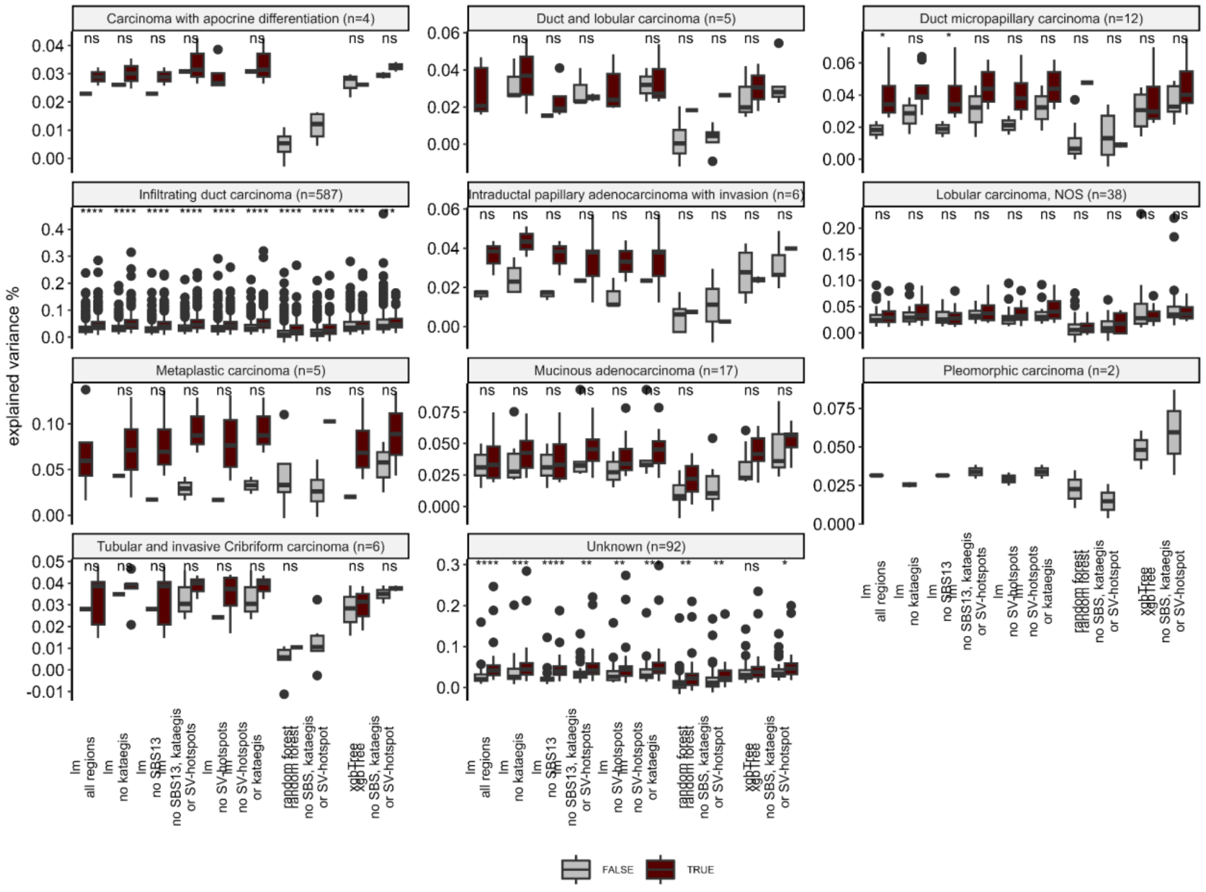
Supplementary figure 6. *Number of tissue-specific genes for each RNA-seq dataset of normal tissue based on the extended Tau index.*



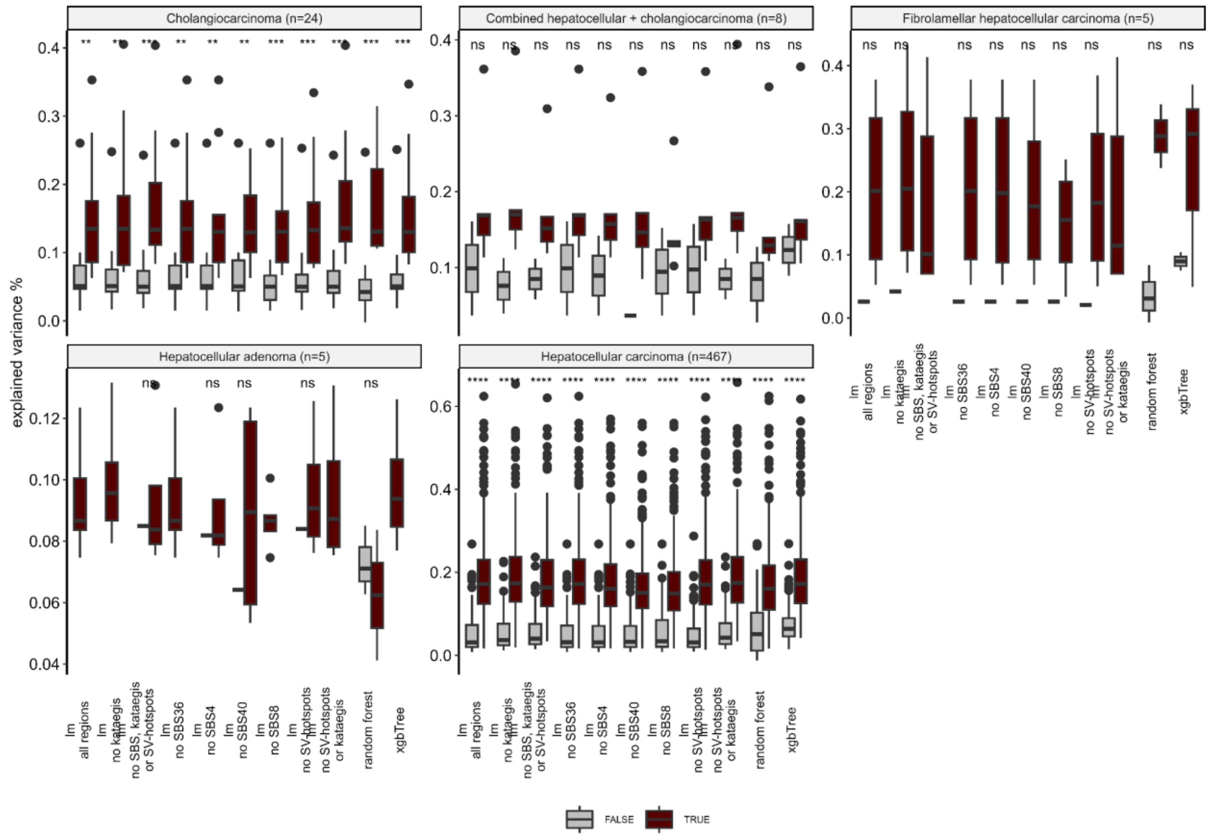
Supplementary figure 7. *A) Multiple linear regression models for the prediction of mutation density of aggregated tumor profiles in 1 Mb genomic regions of breast, liver and skin cancer WGS and WXS were trained on an extended set of 101 tissue sets but showing only the top one. The overall explained variance is reported across the 10-fold cross-validation. B) Proportion of correctly and incorrectly identified cell-of-origin (COO) of individual patients colored by their belonging cohort.*



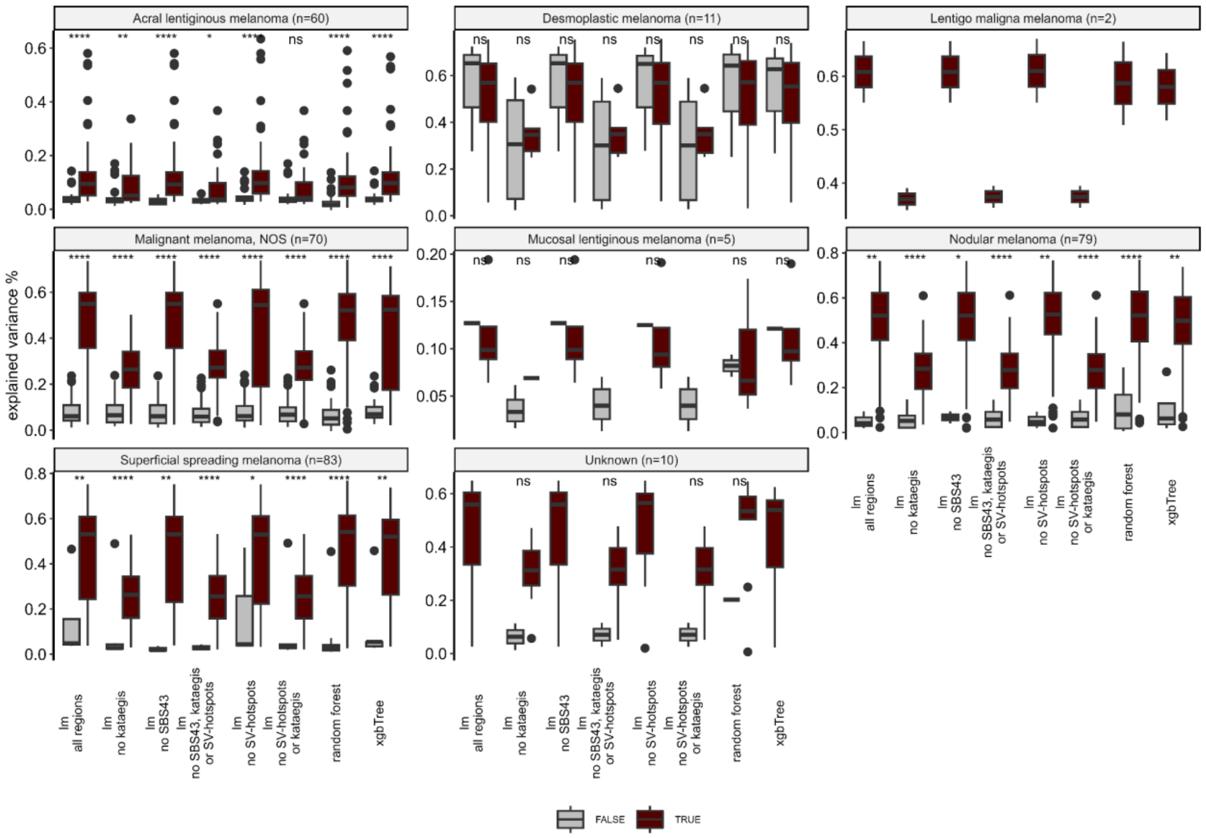
Supplementary figure 8. *Multiple linear regression models for the prediction of mutation density of indel aggregated tumor profiles in 1MB genomic regions of breast, liver and skin cancer WGS and WXS were trained on an extended set of 101 tissue sets but showing only the top 15 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation.*



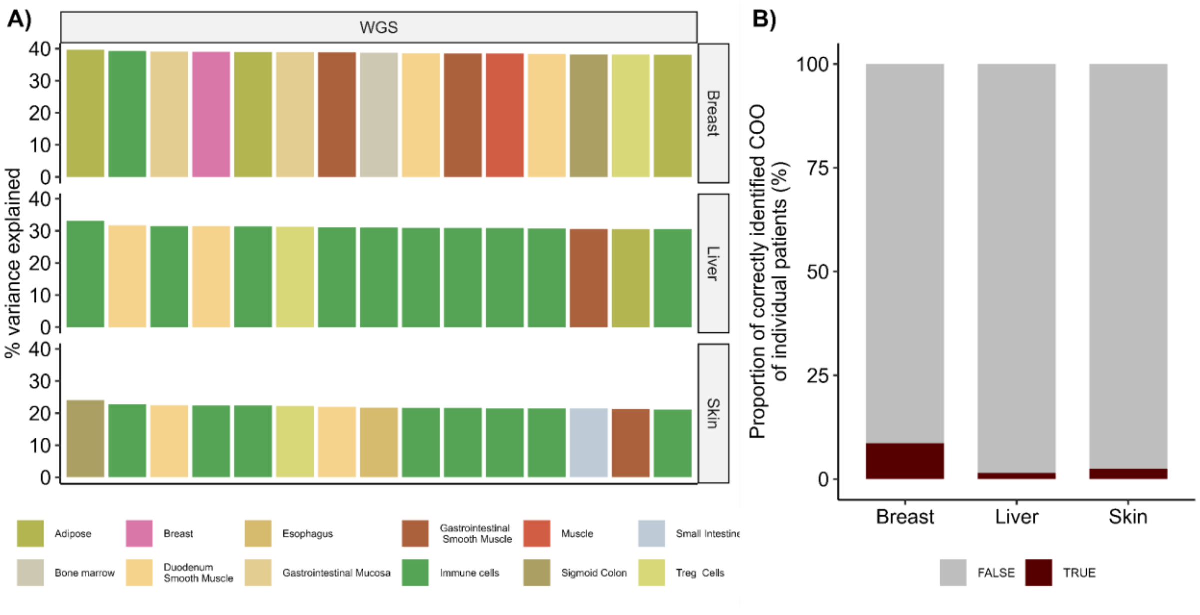
Supplementary figure 9. *Explained variance for correctly and incorrectly identified COO of individual patients with* 1 MB genomic regions COO *multiple models for different breast cancer histological types. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*



Supplementary figure 10. *Explained variance for correctly and incorrectly identified COO of individual patients with* 1 MB genomic regions COO *multiple models for different liver cancer histological types. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two*-sided Wilcoxon test, *ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*



Supplementary figure 11. *Explained variance for correctly and incorrectly identified COO of individual patients with* 1 MB genomic regions COO *multiple models for different skin melanoma histological types. Box plots show the median value, interquartile range as a box, and the whiskers extend to IQR±1.5*IQR. Two-sided Wilcoxon test, ns: p > 0.05 *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001*



Supplementary figure 12. *Multiple linear regression models for the prediction of mutation density of indel aggregated tumor profiles in TADs of breast, liver and skin cancer WGS and WXS were trained on an extended set of 101 tissue sets but showing only the top 15 in each defined subgroup of genes. The overall explained variance is reported across the 10-fold cross-validation.*

9 Curriculum vitae

Paula Štancl was born on June 22, 1996 in Zagreb, where she finished elementary school and high school. In 2015, she enrolled in the undergraduate study of molecular biology at the Faculty of Science (PMF) in Zagreb. She won a Rector's award in 2018. for her research under title "Računalna analiza sljedova ogulinske špiljske spužvice (*Eunapius subterraneus* Sket & Velikonja, 1984) prikupljenih tehnologijom sekvenciranja nanoporama". In 2020 she graduated from the graduate study of molecular biology at the PMF in Zagreb.

Since October 2020, she has been employed as an assistant in the Bioinformatics group at the Department of Molecular Biology, PMF, University of Zagreb, where in the same year she enrolled in a doctoral course in Biology. She completed her doctoral dissertation as a collaborator on the HRZZ project PREDI-COO (*A statistical modeling approach to predict the cell-of-origin and investigate mechanisms of cancer development*) (Project number: IP-2019-04-9308) under the mentorship of professor Ph.D. Rosa Karlić. Moreover, she worked on a project *Understanding the History of Mutations and Cancer Cells* financed by The Chan Zuckerberg Initiative. She has published 5 scientific papers and has taken part in numerous international meetings, presenting posters or giving short oral presentations as the first author. She received the best poster award at the conference *HDIR-6: Targeting Cancer* in 2022 and the best abstract award at the 5th Belgrade Bioinformatics conference in 2024.

In addition, she participates in lectures and practical teaching as part of the course Algorithms and programming, Machine learning and statistics, Computational biology, Translational genomics, Experimental design of high-flow experiments and Laboratory professional practices. She has also co-mentored three master's theses.