The impact of germline variants on human somatic mutation processes

Mischan Vali Pour Jamnani

TESI DOCTORAL UPF / 2021

Director de la tesi

Dr. Ben Lehner
Genetic Systems Group
Systems Biology Department
Centre for Genomic Regulation (CRG)

DEPARTMENT OF EXPERIMENTAL AND HEALTH SCIENCES

Universitat
Pompeu Fabra
Barcelona

"What You Waiting For?" - Gwen Stefani

Acknowledgements

First and foremost, thanks a lot to my co-supervisors Ben and Fran for being the quickest people on earth when it comes to answering e-mails, reading drafts, and providing feedback. Thanks to Ben for giving me the opportunity to do this PhD, for always being available, and in particular, thank you for creating such a great lab environment to work in. Thanks to Fran for opening the world of mutations and crazy regressions to me, for always staying positive, and for being patient with me. Thanks to Solip for your always motivating words and your unique humor. Thanks to Donate for being a perceptive and helpful member in my thesis committee. And thanks to Natalie, my former PI, for once sparking my interest in the field of cancer genomics and pushing me to pursue a PhD in the field.

Further, I would like to thank all the people in the lab (and ex lab members), especially, the people in the lab who made my last year truly special. I feel extremely lucky to have met you and I am excited to welcome you in any place I will end up. If there is anything I've learned during the last years, it's that getting to know people like you can never be taken for granted. Words could not describe. Big thanks and hugs to Magda, Taylor, Taraneh, Toni, Maxi, Andre, Sarah, Thomas, Chenchun, Ignasi, Cristina, Albert, Aaron, Tiffany, Aina, Maria, Cici, and Júlia.

Last but not least, thanks to my family. Mum and Dad for showing me what you can do in life via education, hard work, and life-long learning. My brother(-in-law) Jan for always being supportive and my sister Mel, for being the person I could always rely on no matter what.

Abstract

Somatic mutations are an inevitable component of ageing and the most important cause of cancer. The rates and types of somatic mutation vary across individuals, but relatively few inherited influences on mutation processes are known. Here, we performed systematic studies investigating the influence of rare and common germline variants on somatic mutational processes. Firstly, we showed that independent component analysis and variational autoencoder neural networks can be utilized to extract biologically relevant mutational components from an input matrix covering different classes of mutations and genomic features from over 15,000 tumor genomes. Secondly, we identified via a gene-based rare damaging variant association study with diverse mutational processes, using human cancer genomes from over 11,000 individuals of European ancestry, that diverse genes associate with many different mutational processes. Further, we learned that a variance test can be utilized to compensate for inaccurate predictions of damaging variants by *in silico* predictors. Thirdly, in a genome-wide association study between common germline variants and different mutational processes, several hits at genome-wide significance were identified. Fourthly, significant heritable somatic mutational processes based on common variants were detected and heritability of the total mutation burden could be attributed to at least three different mutational processes. Overall, we suggest that mutational processes in our cells have an important heritable component, contributing to inter-individual differences in somatic mutation accumulation.

Resumen

Les mutacions somàtiques són un component inevitable de l'envelliment i la causa més important del càncer. Les taxes i els tipus de mutació somàtica varien entre els individus, però la base genètica d'aquesta variació encara no ha estat estudiada sistemàticament. Aquí, hem realitzat estudis sistemàtics per investigar la influència de variants genètiques rares i comunes de la línia germinal en els processos mutacionals somàtics. En primer lloc, hem demostrat que l'anàlisi de components independents i les xarxes neuronals d'autocodificadors variacionals es poden utilitzar per extreure components mutacionals biològicament rellevants d'una matriu d'entrada que cobreix diferents classes de mutacions i característiques genòmiques de més de 15,000 genomes tumorals. En segon lloc, mitjançant un estudi d'associació de variants perjudicials rares basat en gens amb diversos processos mutacionals, utilitzant genomes de càncer humà de més d'11,000 individus d'ascendència europea, hem identificat que diversos gens s'associen amb molts processos mutacionals diferents. A més, hem après que es pot utilitzar un test de variància per compensar prediccions inexactes de variants perjudicials mitjançant predictors *in silico*. En tercer lloc, en un estudi d'associació a tot el genoma entre variants comunes de línia germinal i diferents processos mutacionals, hem identificat diversos gens que arriben al llindar de significativitat a escala genòmica. En quart lloc, hem detectat processos mutacionals somàtics hereditaris significatius basats en variants comunes i hem atribuït l'heretabilitat en la càrrega total de mutacions a almenys tres processos mutacionals diferents. En general, suggerim que els processos mutacionals a les nostres cèl·lules tenen un component hereditari important, que contribueix a les diferències interindividuals en l'acumulació de mutacions somàtiques.

Contents

List of Figures

List of Tables

1 Introduction

Mutations in DNA cause disease, phenotypic variation and evolution. They are the result of alterations in the nucleotide sequence which remain unrepaired, predominantly due to replication errors and exogenous damaging agents. The advent of next-generation sequencing (NGS) data in the last 15 years has revolutionized the field of mutation research via the accumulation of variant data from 1,000[1], 10,000[2] and recently even over 100,000[3] human individuals.

In the field of cancer genomics the sources and distributions of somatic mutations have been extensively studied and catalogued[4,5,6]. Somatic mutations are DNA alterations which occur in cells of the body during cell divisions and most importantly are largely not passed on to the offspring. In large-scale sequencing efforts like the Cancer Genome Atlas Program (TCGA)[7], the Pan-Cancer Analysis of Whole Genomes (PCAWG)[8] and the Pan-cancer whole-genome analyses of metastatic solid tumours from the Hartwig Medical Foundation[9], tumor cells as well as patient-matched healthy cells were sequenced covering a variety of tissues and covering over 30 different tumor types. Analysis of the occurring somatic mutations in theses studies have helped to identify known and new genes promoting oncogenesis (driver genes)[10,11,12] and to catalogue the causal point mutations in these cancer driver genes (driver mutations)[4]. Studies mostly focussed on protein-coding genes and just recently also extended to the non-coding regions[13]. Driver mutations are outnumbered by mutations which do not lead to an apparent selective advantage of the cancer cell (passenger mutations) and it has been an ongoing challenge to distinguish driver mutations from passenger mutations. Even though passenger mutations do not result in the development of cancer, the number, type and distribution of these mutations have provided insights into the mechanisms generating mutations in the genome. As it will be described in more detail, single nucleotide substitutions and other mutation classes have been used to identify mutational patterns ('mutational signatures')[5], which have led to the identification of known, novel and unknown mutational processes[5,6,14]. While initially mutational signatures were mainly based on single nucleotide substitutions[5,6], by now mutational signatures have been generated based on insertions and deletions (indels)[14], double nucleotide substitutions[14], clustered mutations[15,16], copy number alterations[17] and recently even on structural variants[18], which are still more challenging to be precisely identified. In addition, it has been also described how different mutational processes can affect the distribution of mutations in a cancer genome with respect to different genomic properties such as chromatin marks[15,19,20], replication timing[21], chromatin accessibility[22], and the direction of replication[23] and transcription[24] to only name a few. Furthermore, somatic mutations have been used to analyse the effects of cancer therapies on the genome[25,26,27] and to identify drug-resistance mutations[28]. Additionally, mutational analysis has been used in the field of tumor evolution, which was significantly advanced by the emergence of NGS technologies. Even though this field will not be further covered in this thesis, fundamental insights into tumor evolution have been provided via the analysis of mutations in tumor genomes from patients at different time points and/or different regions of the same tumor. It was shown that cancer genomes are heterogenous even within an individual tumor and highly dynamic[29,30].

The accumulation of sequencing data has drastically increased our understanding of the mutational processes occurring in the tumor genome and also in healthy somatic cells. During the last years also the analysis of germline variants has been studied. In contrast to somatic mutations, germline variants occur in the germ line and thus, are passed on to the offspring. The number of germline variants in the genome is higher than the number of somatic mutations by several orders of magnitude. Studies have shown how germline variants can be highly informative for cancer risk prediction, prognosis and prediction of response to therapies[31]. Furthermore, a limited number of studies has shown that germline variants can affect mutational processes in the tumor[8,31,32]. This information will be crucial in the future to further improve cancer risk predictions of individuals and to better tailor a therapeutic strategy to the patient.

Hence, this PhD thesis was performed with the goal to get a better understanding of how germline variants affect somatic events in the tumor genome. The next sections of the introduction will to put this thesis into context. In the first section, the incidence of cancer, the hallmarks of cancer and the process of cancer evolution will be described (Section 1.1). Next, DNA repair pathways, DNA damaging agents and DNA damaging processes which play an important role in generating mutations the cancer genome will be described (Section 1.2). The different somatic mutational features which have been identified in the last years will be outlined (Section 1.3). Further, it will be explained to what extent the effects of germline variants on somatic events in the tumor genome have been studied, what was learned from these studies and which methods have been used (Section 1.4). Lastly, I will introduce the aim and objectives (Section 1.5) and the study design of the work presented in this thesis (Section 1.6).

1.1 Cancer in a Nutshell

1.1.1 Cancer Incidence

Cancer is still one of the most lethal diseases in the world. According to the International Agency for Research on Cancer (IARC), which is part of the World Health Organization (WHO), over 19 million new cases of cancer were reported in 2018 and almost 10 million people died from the consequences of the disease[33]. Most importantly, the number of cases are expected to rise to up to 30 million new cases per year and over 16 million deaths per year by 2040. The expectation of increasing number of cancer cases is connected to the growing lifespan[34].

1.1.2 The Hallmarks of Cancer

Cancer comprises a group of over 100 different diseases with different characteristics with respect to cell of origin, risk factors, incidence, mortality and therapeutic treatment to only name a few. Tremendous work has been performed since the early discoveries by David von Hansemann in 1890[35] and Theodor Boveri in 1914[36] that cancer cells exhibit chromosomal abnormalities. The common characteristics between all types of cancer have been summarised in six hallmarks of cancer in the landmark paper of Douglas Hanahan and Robert A. Weinberg[37] in 2000. These hallmarks were later updated in 2011 by two additional hallmarks and two enabling characteristics based on the advanced understanding of the diseases[38].

The six first hallmarks comprise resisting sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, in-

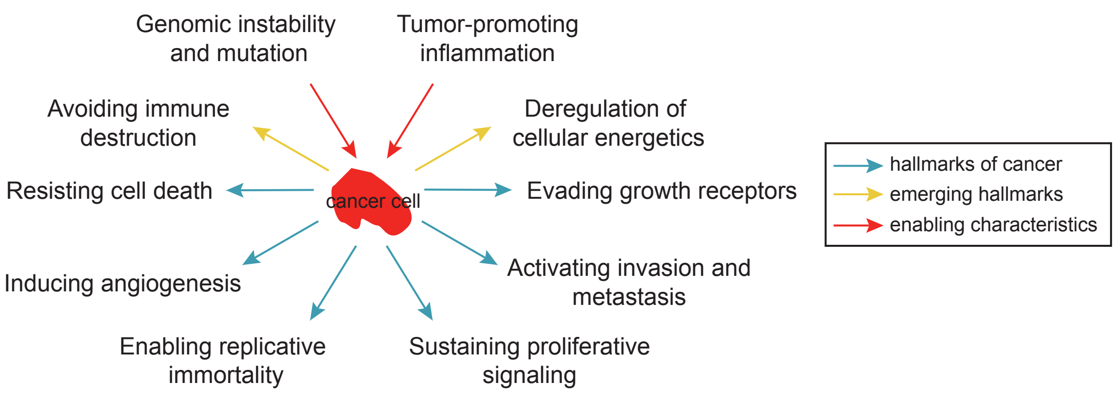

Genomic instability and mutation
Tumor-promoting inflammation
Avoiding immune destruction
Deregulation of cellular energetics
Resisting cell death
cancer cell
Evading growth receptors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality
Sustaining proliferative signaling
hallmarks of cancer
emerging hallmarks
enabling characteristics

Figure 1.1: The hallmarks of cancer. Figure inspired from ref[38].

ducing angiogenesis and activating invasion and metastasis[38].

- **Sustaining proliferative signaling**: An essential characteristic of cancer cells is their ability to grow infinitely which can be achieved via different mechanisms such as the production of growth factor ligands by the cancer cell itself or by increased levels of growth receptors on the cancer cell. Some cancer cells are also independent of growth factors through activating mutations in downstream components of the respective pathways[38].

- **Evading growth suppressors**: In the same way that cancer cell stimulate their growth by making use of growth stimulating pathways, they need to deactivate pathways which control and suppress cell proliferation. This is mostly achieved by inactivating mutations in tumor suppressor genes. Notable examples here are the retinoblastoma protein Rb and the tumor suppressor TP53[38].

- **Resisting cell death**: Cancer cells develop during their molecular evolution strategies to resist cell death (apoptosis). The most common mechanism is a loss-of-function of the tumor suppressor TP53. Other mechanisms involve the upregulation of antiapoptotic regulators, the increased expression of surviving signals or a decreased expression of proapoptotic regulators[38].

- **Enabling replicative immortality**: Normal cells go through a limited number of cell divisions and at last either enter the nonproliferative state of senescence or die via apoptosis. This limited number of cell divisions in normal cells has been attributed to the telomeres, which shorten after each cell division until chromosomes are not protected anymore, resulting in end-to-end fusions of chromosomes, and consequently leading to programmed cell death. Cancer cells found a way to circumvent this process. In over 90 % of the cases, cancer cells upregulate the expression of the enzyme telomerase, which adds telomeric tandem repeats to the chromosomal ends[38]. In rare cases, telomeric shortening is counteracted via a recombination-based mechanism[38].

- **Inducing angiogenesis**: To maintain their growth, tumor cells need to transport oxygen and nutrients into their cell and carry off carbon dioxide and other end products. Thus, cancer cells induce the growth and assembly of new blood vessels. This is performed for instance via the upregulation of angiogenesis inducers like vascular endothelial growth factor-A (VEGF-A), which can be promoted via oncogenic signalling[38].

- **Activating invasion and metastasis**: Another hallmark of many cancers, especially carcinomas, is the ability to change shape, invade other tissues and even spread to other places in the body. Studies so far have shown how gains and losses of cell-to-cell (e.g. E-cadherin) and cell-to-matrix proteins play a role here and overall the process of invasion and metastasis is believed to be a multistep process[38].

Overall, the six hallmarks of cancer describe how cancer cells manage to survive, proliferate and spread[37]. As discussed by Douglas Hanahan and Robert A. Weinberg, two enabling characteristics make the acquisition of these hallmarks possible in the first place: genomic instability and mutation, and tumor-promoting inflammation[38].

- **Genomic instability and mutation**: Many hallmarks of cancer are acquired via genomic instability and mutations in the cell and thus unexpectedly, many cancer types harbour defects in DNA repair and/or maintenance pathways. Loss-of-function in these pathways are essential for enabling a cell to acquire beneficial mutations, which can ultimately lead to tumor progression[38].

- **Tumor-promoting inflammation**: It has been reported how many tumors are infiltrated by immune cells and specific antibodies, which surprisingly promote tumorigenesis. Studies showed that inflammation can help the tumor cells to grow and to acquire different hallmarks by providing essential molecules such as proangiogenic factors and growth factors or by increasing mutation rates via the release of reactive oxygen species (ROS)[38].

In addition, two emerging hallmarks of cancer have been added to the initial six hallmarks: deregulating cellular energetics and avoiding immune destruction[38].

- **Deregulating cellular energetics**: Cancer cells have an altered energy metabolism to deal with the increased cell growth and increased number of cell divisions. The most frequently occurring metabolic switch in cancer cell is the 'Warburg-effect', which describes the observation that cancer cells use glycolysis for their main source of energy production instead of mitochondria. This state has been also termed as aerobic glycolysis. It is suspected that the intermediate molecules from glycolysis are used in cancer cells for biosynthesis of nucleotides, amino acids, proteins and organelles[38].

- **Avoiding immune destruction**: This hallmark has also been added due to the growing evidence that cancer cell growth is hold in check by the immune system. Cancer cells need to find ways to evade the surveillance of the immune system. Due to the complexity of the tumor microenviroment and the immune system, our understanding of the exact mechanisms is still rudimentary. Suggested mechanisms involve the recruitment of immunosuppressive cells by the cancer cells via promotion of inflammation and secretion of immunosuppressive factors such as the cytokine transforming growth factor beta (TGF-β)[38].

1.1.3 Cancer Evolution

Tumor progression is a multi-step progress in which, in very simple terms, cells with a favourable mutation outgrow the other cells. This process has been sometimes compared to Darwinian evolution[39]. In the end, cells acquiring a selective advantage via the mechanisms described in the 6-8 hallmarks of cancer will outgrow their neighbouring cells. The key point to remember here (especially for this thesis) is the importance of the generation of somatic mutations, which enable tumor transformation and progression in the first place, which is why it has been also described as one of the enabling characteristics of cancer[38]. Studies so far pointed out that depending on the cancer type, 2 to 10 mutations in cancer driver genes are enough to lead to tumorigenesis[10,40]. Even though mutations are the main focus of this thesis and most work so far, nongenetic factors can generate variation between cells leading to a selective advantage for specific cells as well[41].

1.2 Sources of DNA Mutations

Sources of DNA mutations can by endogenous (e.g. spontaneous deamination of cytosines, DNA replication errors) as well as exogenous (e.g. chemicals, radiation). While we will go through the different processes in more detail here separately, it should be noted that in the end the fixed mutations in the genome are a consequence of the interplay between DNA damage and DNA repair[42] (Figure 1.2).

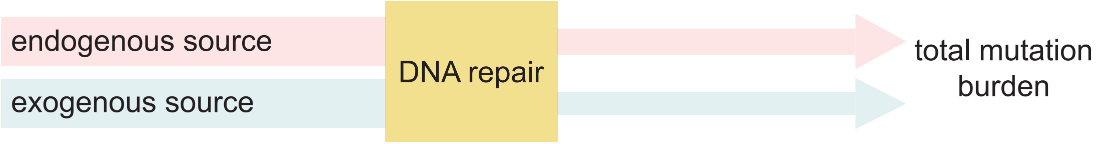


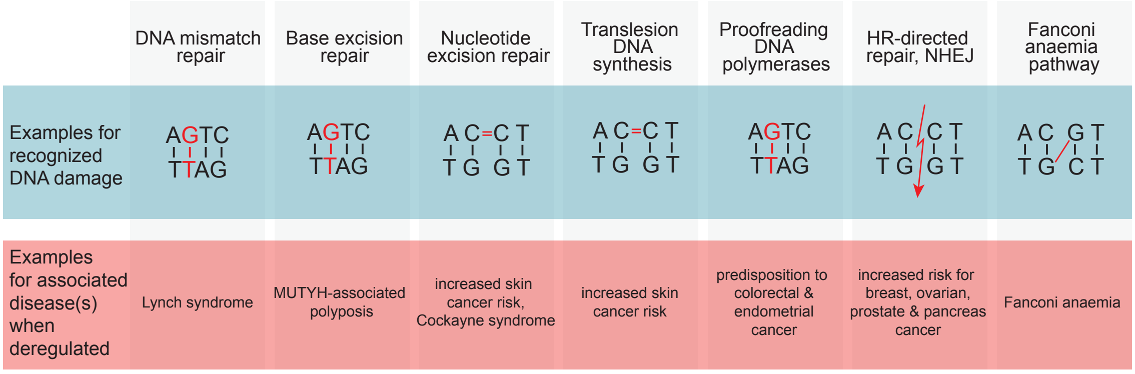
Figure 1.2: Illustration of the interplay between endogenous and exogeneous sources of DNA mutations and DNA repair. Normal cells as well as tumor cells are constantly affected by DNA damage. A reduced activity of DNA repair mechanisms or an excess of DNA damaging agents/processes will lead to an increase in the total mutation burden[132]. Figure inspired from ref[132].

1.2.1 DNA Repair Pathways

The different DNA repair pathways play a crucial role in identifying and repairing DNA damage and thus, preventing DNA mutations[43]. The different pathways have been extensively studied and each pathway has different characteristics concerning repair efficiencies, time of action and types of mutations which are recognized[44]. Deregulation of most of these pathways are not only often times connected to increased numbers of mutations in the tumor genome, but also frequently connected to specific diseases (DNA repair syndromes)[45,46] (Figure 1.3). Notably, DNA repair pathways are sometimes not only responsible for the repair of DNA mutations, but also a source of DNA mutations[47].

DNA mismatch repair

The DNA mismatch repair pathway (MMR) recognizes primarily base-base mismatches and indels. Its key function is the repair of mutations that are generated during replication. Thus, it prevents that replication-associated mutations get fixed in the genome. The pathway is highly conserved and was initially studied in detail in *Escherichia coli*. MMR is bidirectional (5’ to 3’ and 3’ to 5’),



	DNA mismatch repair	Base excision repair	Nucleotide excision repair	Translesion DNA synthesis	Proofreading DNA polymerases	HR-directed repair, NHEJ	Fanconi anaemia pathway
Examples for recognized DNA damage	AGTC TTAG	AGTC TTAG	A C=C T T G G T	A C=C T T G G T	AGTC TTAG	A C C T T G G T	A C G T T G C T
Examples for associated disease(s) when deregulated	Lynch syndrome	MUTYH-associated polyposis	increased skin cancer risk, Cockayne syndrome	increased skin cancer risk	predisposition to colorectal & endometrial cancer	increased risk for breast, ovarian, prostate & pancreas cancer	Fanconi anaemia

Figure 1.3: Illustration of DNA repair mechanisms, the DNA lesions they recognize, and associated dieases. While different DNA repair mechanisms recognize specific sets of DNA lesions, there is an overlap between DNA repair pathways. Deregulation of DNA repair pathway has been associated with different diseases. Showing here examples for each pathway.

strand-specific by recognizing the newly synthesized strand and has a broad substrate specificity[49].

Firstly, mismatches are recognized. It is still unclear how exactly MMR is able to discriminate the two strands from each other in eucaryotes. The most prominent explanation is the presence of a nick. Ultimately, repair is initiated either by MutSα or MutSβ. MutSα is a heterodimer consisting of MSH2 and MSH6 and MutSβ is a heterodimer consisting of MSH2 and MSH3. While MutSα preferentially binds base-base mismatches and indels with a length of 1 to 2 nucleotides[50,51], MutSβ has a preference for indels with size of 2 or more nucleotides[52,53]. Next, MutLα, which is a heterodimer comprised of MLH1 and PMS2, cuts the new DNA strand via its endonuclease activity and the mispaired bases and neighbouring bases are excised via EXO1. The gap is then re-filled via DNA polymerase ε and sealed via DNA Ligase 1. Furthermore, also other proteins have been reported to play a role in MMR such as RPA and PCNA[49]. RPA is a single-strand DNA binding protein and thought to protect ssDNA during MMR until re-synthesis. Before re-synthesis, RPA gets phosphorylated leading to a decreased affinity of RPA to DNA[55]. The DNA clamp PCNA has been reported to be required for the activations of the endonuclease activity of MutLα, which is crucial of MMR initiation[54].

Beside the essential function of MMR in the repair of replication associated mutations, MMR has also been reported to have non-canonical roles[56]. While in the replication-associated repair MMR can distinguish between the mother and daughter strand, this information is lost in the non-canonical MMR pathways. The non-canonical MMR pathway has been reported to play a role in somatic hypermutation at the immunoglobulin locus[57] and also in the repair of DNA lesions at A:T base pairs via the recruitment of the error-prone polymerase η (POLH)[58,15]. In addition, a recent study suggested the involvement of MMR in a replication independent, but error-free repair of deaminated 5-methyl cytosines[59]. It was proposed that the recruitment of the MMR machinery is dependent on histone mark H3K36me3, which has previously been reported to interact with the MutSα complex[20].

The importance of MMR becomes clear in the context of hereditary non-polyposis colorectal cancer (HNPCC), also termed as Lynch syndrome, which is caused by inherited rare damaging variants in MMR pathway components. Genetic analyses revealed that germline mutations in *MLH1*, *MSH2*, *MSH6* and *PMS2* cause the disease[60,61]. A characteristic of MMR deficiency is an increased number of mutations in genomic regions containing tandem repeats of DNA motifs (microsatellites), which has been termed as microsatellite instability (MSI)[62,63]. MSI has also been detected in sporadic cancers caused by somatically acquired mutations in these genes or by epigenetic silencing. The most frequent cause of sporadic colorectal cancers is the inactivation of *MLH1* by hypermethylation of its promoter[64]. Initially, MSI was detected via a PCR of specific microsatellite regions (Bethesda panel) or via the immunohistochemistry detection of the MMR proteins MSH2, MLH1, MSH6, and PMS2[65]. By now, different computational methods exist which are able to measure MSI such as MSISensor[66] and MANTIS[67]. A large-scale analysis of over 11,000 tumor samples covering 39 cancer types, revealed that 3.8 % of all cancers had a MSI phenotype and MSI phenotypes were found across 27 different cancer types. The highest fraction of MSI was reported in endometrial, colorectal and stomach adenocarcinomas[68]. Interestingly, even though MSI tumors have an increased number of somatic mutations in the tumor genome due to the decreased MMR activity, they have a better prognosis in comparison to microsatellite stable (MSS) tumors[69]. In addition, they respond better to immune checkpoint inhibitors due to the higher occurrence of neoantigens in these tumors[70].

Nucleotide excision repair

The nucleotide excision repair (NER) pathway is mostly responsible for the repair of bulky DNA lesions, which are caused by tobacco smoking, UV light and other carcinogens[71]. Inherited defects in NER components lead to the autosomal recessive disease xeroderma pigmentosum (XP). Patients display a high sensitivity to UV light and an almost 2000-fold increased risk for skin cancer due to the inability of the cells to repair UV-induced lesions[72]. Also other diseases such as Cockayne syndrome and trichothiodystrophy have been connected to damaging variants in NER genes. These diseases do not predispose to skin cancer but are connected to a broad number of diverse symptoms[71].

Two main modes of NER have been described so far: global-genome NER (GC-NER) and transcription-coupled NER (TC-NER). As described in the name, GC-NER occurs everywhere in the genome, while TC-NER is only active in the transcribed strand of expressed genes. The two processes are activated by different proteins but the excision is believed to be performed by the same core NER proteins. GC-NER is initiated by proteins which can sense DNA lesions via helix distortions (e.g. XPC-Rad23B) or via proteins which specifically recognize DNA damage such as damage specific DNA binding protein 2 (DDB2, also termed as XPE). TC-NER is started when the RNA polymerase stops transcription due to a lesion on its way. Other proteins such as CSA, CSB, and XAB2 are recruited then to initiate TC-NER. In the next steps, different proteins including transcription factor TFIIH bind at the site and excise a single-stranded DNA stretch (ssDNA) of around 30 nucleotides around the DNA lesion. The gap is then re-synthesized by a DNA polymerase (DNA polymerase α, ε or κ) and ligated by DNA Ligase 1 or 3[71].

Base excision repair

While NER is mostly responsible for bulky DNA lesions due to exogenous factors, base excision repair (BER) removes primarily small base lesions caused by endogenous processes. These include base lesions such as deaminated bases, alkylated bases or oxidized bases[73].

Initially, the damaged site is recognized by a DNA glycosylase, which subsequently cleaves the base. Two types of DNA glycosylases exist: monofunctional and bifunctional ones. Monofunctional DNA glycosylases leave an abasic site behind, which is then removed by an apurinic endonuclease (APE1). APE1 generates a 3'OH end and a 5' deoxyribosephosphate terminus, which is then excised by DNA polymerase β (Pol β). In the case of bifunctional DNA glycosylases, a single strand break (SSB) is generated via the associated lyase activity, which results either in a α,β-unsaturated aldehyde or a phosphate group in the 3' end. While the α,β-unsaturated aldehyde can be removed by the activity of APE1, the phosphate group is removed by polynucleotide kinase (PNKP). The missing base is inserted by Pol β and DNA ends are joined by DNA ligase 3, which is in a complex with the BER scaffold protein XRCC1. The repair of a single base via BER is also called short patch BER[73].

In long patch BER the gap is filled by DNA polymerase β, δ or ε. Even though it is still investigated under which conditions short patch BER is initiated and under which conditions long patch BER is initiated, it has been shown that long patch BER is initiated when the 5' sugar is modified since then it can't be processed by Pol β. In these cases, a DNA polymerase fills the one base gap and keeps on synthesizing. The nucleotides downstream of the abasic site are displaced and the created flap is cleaved via the endonuclease activity of FEN1. In this way a stretch of 2 to 13 nucleotides is removed from the damage site and re-filled again[73].

As DNA bases can be damaged via different mechanisms, different types of base modifications are recognized and repaired by different sets of specific DNA glycosylases as show in Table 1.1[73].

Table 1.1: Overview of DNA gycosylases and their substrates.

DNA gycosylase	Substrate/Target
SMUG1	uracil, 5-hydroxymethyluracil
UNG1	mitochondrial uracil
UNG2	uracil
MBD4	T:G mismatches
TDG	T:G mismatches, 5-carboxylcytosine
MPG	alkylated bases such as 3-methyladenine
OGG1	8-oxoguanine, FapyG
MUTYH	8-oxoguanine in mismatch with adenine
NTHL1	oxidized pyrimidines
NEIL1	oxidized pyrimidines, formamidopyrimidines and more
NEIL2	same as NEIL1, but preferentially in ssDNA
NEIL3	same as NEIL1

Inherited germline variants in different DNA glycosylases have been connected to different types of cancer and hereditary diseases[73]. For instance, germline variants in *MBD4* have been identified in stomach cancers in a Japanese cohort[74] and have been connected to an increased risk of lung and esophageal cancer in a Chinese cohort[75]. Individuals with inherited damaging germline variants in *MBD4* have been reported to have an increased number of C>T mutations at CpG contexts in several cancer types[8,76,77]. Furthermore, biallelic inactivation of the DNA glycosylase predisposes to MUTYH-associated polyposis (MAP). The disease is, in particular, connected with an increased risk of developing colorectal cancer[78]. In addition, biallelic germline mutations in *NTHL1* predispose to multiple cancer types as well such as adenomatous polyposis and colorectal cancer[79]. These examples of deficiencies in BER components causing carcinogenesis and/or increasing cancer risk are in line with conclusions coming from many mice studies, which showed that knock out of individual glycosylases can lead to an early onset of cancer[73].

Proofreading activities of the replicative DNA polymerases

Replicative DNA polymerase have the essential role of faithfully replicating the genetic material with a high accuracy. In mammalian cells, DNA polymerase δ (encoded by *POLD*) as well as DNA polymerase ε (encoded by *POLE*) have proofreading capabilities to increase fidelity. During DNA elongation, the incorporation of an incorrect base is recognized and subsequently, the incorrect base is excised via a 3' to 5' exonuclease activity[80]. Overall, it has been estimated that the polymerase selectivity leads to a error rate of 1 nucleotide in every 10^4 - 10^5 [81] nucleotides and that the proofreading activity increases the fidelity by 10^2 - 10^3 [82].

Thus, it is no surprise that inherited damaging mutations in the proofreading domains of *POLE* or *POLD1* predispose to certain kinds of cancer, such as to colorectal cancer[83]. Somatically acquired mutations in the exonuclease domain of *POLE* lead to a ultra-mutator MSS phenotype, which has been found in many sporadic colorectal as well as endometrial cancers[85,84]. Causative somatic mutations for a mutator phenotype have been also discovered in *POLD1*, but less frequently than in *POLE*. In contrast to *POLE*, somatic mutations in *POLD1* less frequently affect the exonuclease domain[84].

Translesion DNA synthesis

Translesion synthesis is a process which allows the cell to replicate DNA even in the presence of DNA lesions, which would otherwise block DNA elongation, and thus, cause replication stress, fork collapse and ultimately, apoptosis[86]. In the translesion synthesis process, the regular DNA polymerase during DNA elongation is switched out by a translesion polymerase when elongation is blocked. Translesion polymerases are capable of replicating over the base lesion due to a larger active site and are often specialized in inserting the correct base opposite specific base alterations. In the same time they are also more error-prone[87]. Still, the increased risk of additional mutations is less harmful for the cell than the consequences of fork collapse and DNA double-strand breaks. A common mechanism for polymerase switching is initiated via ubiquitination of PCNA, while also mechanisms independent of ubiquitinated PCNA have been reported[86].

As always the importance of these polymerases become more apparent in the context of diseases such as cancer. A notable example is DNA polymerase η, which is able to accurately replicate over UV-induced mutations[88], even though it is error-prone on undamaged DNA templates[87]. It has been shown that certain inherited variants in *POLH* (encoding DNA polymerase η) can predispose to skin cancer induced by UV light[86, 89].

Homologous recombination-directed repair

DNA double-strand breaks (DSBs) are caused by different factors such as ionizing radiation or chemotherapeutic drugs. They are predominantly repaired via two mechanisms[90]: homologous recombination (HR)- directed repair and non-homologous end joining. HR is a highly regulated process and is specifically active during the S and G_2 phases of the cell since the repair process is guided by the sister chromatid[90, 91].

Roughly, HR can be divided be into several steps. The first step involves the generation of 3'-single-stranded DNA performed by endonucleases from the MRN complex. Next, the ssDNA ends are bound by replication protein A (RPA) and then replaced by the recombinase RAD51, which is promoted by recombination proteins such as BRCA1 and BRCA2. Lastly, the ssDNA invades into the intact sist er chromatid (strand invasion), new DNA is synthesized by using the intact sister chromatid as template and consequently, the strand information is restored and DSB repair can be completed[92].

HR is important for DSB repair as well as chromosome segregation. Knockouts in HR genes like *BRCA1*, *BRCA2* or *RAD51* lead to embryonic lethality in mice. Heterozygous inactivations of these genes do not lead to embryonic lethality, but predispose to cancer[92]. In fact, *BRCA1* and *BRCA2* are frequently mutated in breast and ovarian cancer, *RAD54* and *CtIP* in lymphomas and *RECQL4* in different carcinomas to only name a few[90]. Furthermore, tumors with inactivations in HR genes like *BRCA1*, *BRCA2*, *RAD51C* and *PALB2* show distinct genomic alterations[93,94,95].

Interestingly, it has been shown that cells with inactivations in HR proteins such as BRCA1 and BRCA2 rely heavily on backup pathways. This observation has been utilized in cancer therapy by targeting HR deficient tumors with Poly(ADP-ribose) polymerase 1 (PARP1) inhibitors. One of the current hypothesis is that PARP1 inhibitors block repair of SSBs, causing the accumulation of DSBs via replication. This high accumulation of DSBs would be toxic for the cell and result in cell death. In this case, PARP inhibitors would be the first successful example of a synthetic lethal approach in cancer therapy[90].

Non-homologous end joining

Non-homologous end joining (NHEJ) is also a pathway responsible for the repair of DSBs. In contrast to homologous recombination directed repair, NHEJ occurs during the whole cell cycle. The pathway is also estimated to be faster than HR and importantly, more error-prone than HR-directed repair[96].

In NHEJ, the two ends of the DSB are held in proximity to each other and different proteins process the ends to make them compatible for subsequent DNA synthesis and ligation. End joining efficiency is increased when the two opposing strands have a stretch of 1-4 nucleotides of complementarity, also termed as microhomology[97]. End joining at microhomology regions can lead to the generation of several mutations at the joining site, which is also a feature often occurring in tumor genomes with mutations in components of the HR-directed repair pathway[14,95]. End joining can also occur when there is no microhomology existing between the two ends, even though it has been reported to be less preferred and less efficient[97]. Key components of the pathway include Ku70 and Ku86 for binding the ends, Artemis and DNA-PK$_{CS}$ for end-processing, and XRCC4 and DNA Ligase 4 for ligation[97].

Fanconi anaemia pathway

The Fanconi anaemia (FA) pathway is named after the disease Fanconi anaemia, which is caused by germline inactivations in any of the 19 core genes of this pathway. It is mostly responsible for the repair of interstrand crosslinks (ICLs), even though many of the components have also been reported to play roles in other pathways such as FANCD1 and FANCS in HR. ICLs can be caused by exogenous factors such as chemotherapeutic drugs or endogenous metabolites such as nitrous acid or oxygen radicals. Consequently, people with the disease

Fanconi anaemia develop predominantly cancer, especially leukemia, and are very sensitive to drugs which generate ICLs[98].

1.2.2 DNA Damaging Processes

DNA methylation and DNA replication induced mutagenesis

One of the most frequent mutations in the tumor genome[6] as well as in normal somatic cells[99] is the transition from C>T at CpG sites. In most cases, cytosines at CpG sites are methylated leading to a 5-methylcytosine (5mC). It has been shown *in vitro* that the mutation rate of C>T is increased at 5mCs in comparison to unmethylated cytosines[100]. While the deamination of 5mC leads to a thymine, the deamination of unmethylated cyotsine leads to a uracil (Figure 1.4). Uracil is a non naturally occurring base in DNA. Hence, uracils in the DNA due to spontaneously deaminated cytosines are efficiently recognized and repaired[101].

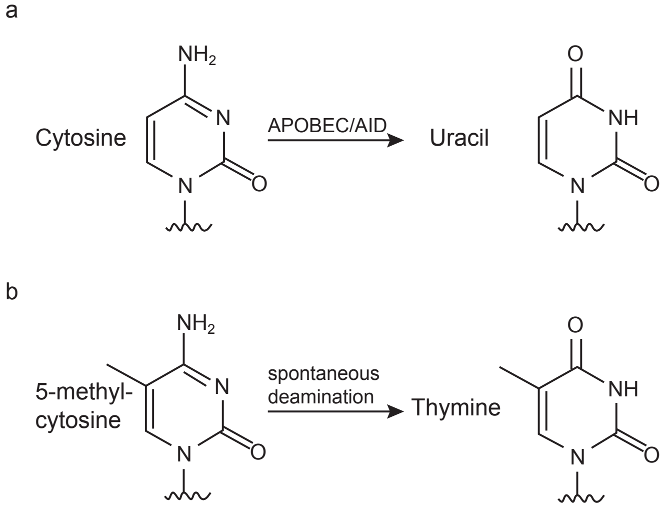


Figure 1.4: Deamination of cytosine and 5-methylcytosine. Cytosines in the genome can be deaminated via APOBEC/AID activity, while spontaneous deamination frequently occurs for 5-methylcytosines at CpG sites[103].

As described above, BER is the key player in repairing small base lesions. The DNA glycosylases UNG2, SMUG1, MBD4 and TDG have been reported to be involved in the repair of U:G mismatches[102]. The repair of the more commonly fixated mutation due to the deamination of 5mC, which results in a T:G mismatch, is repaired by the glycosylases TDG and MBD4[101]. In accordance (and as described before), tumors with germline inactivations in *MBD4* exhibit a drastic increase in C>T mutations at CpG sites, which further emphasizes the important role of BER in repairing these mutations[8,76,77]. In addition, a recent study also suggested the involvement of a non-canonical MMR pathway in the repair of T:G mismatches[59].

Since BER and the suggested non-canonical MMR pathway are replication independent, it would be expected that the distribution of C>T mutations in CpG contexts would not have a strong replication strand asymmetry. However, tumors with mutations in the exonuclease domain of *POLE* or tumors with MSI exhibit a high mutation rate of CpG>TpG with a strong replication strand asymmetry. Additionally, the CpG>TpG mutation rate has been reported to scale with the number of cell divisions which also indicates that replication plays a significant role in the generation of these mutations[103]. It has been hypot[8,]hesized that during DNA replication, adenines are misincorporated opposite of 5mCs by the replication machinery. The proofreading domain of DNA polymerase ε and MMR, which is a co-replicative process, seem to play a major role in repairing these mismatches in a replication dependent manner[104].

Thus, there seem to be several mechanisms involved in the repair of C>T mutations at CpG sites. Firstly, the spontaneous deamination of 5mCs are recognized and repaired by components of the BER pathway[101] as well as by a non-canonical MMR pathway independent of replication[59]. Secondly, errors during DNA replication at 5mC sites are repaired by the proofreading activity of DNA polymerase ε and by the canonical MMR pathway[103].

APOBEC and activation-induced deaminases AID

An important endogenous process contributing to mutagenesis are members of the apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) family as well as activation-induced deaminases (AID). APOBEC enzymes are part of the innate immunity by deaminating viral genomes and thus, protecting the cell from viral infections. AID plays an important role in the adaptive immunity by being crucial in the process of antibody diversification via somatic hypermutation and class switch recombination[103]. AID/APOBEC deaminases are active on ssDNA and create U:G mismatches via the deamination of cytosines. It has been reported that AID/APOBEC deaminases have a higher efficiency on unmethylated cytosines[106,103].

The role of AID/APOBEC deaminases has also been connected to tumorigenesis. In breast cancer genomes, an increased number of clustered C>T and C>G mutations at TpC sites were reported and these mutational patterns were suggested to be generated by the activity of AID/APOBEC deaminases[5]. These mutational patterns were also reported in many other cancer types[6, 15]. In accordance with the fact that these deaminases need ssDNA as a substrate, the mutational patterns connected with AID/APOBEC activity were found, in particular, in regions which are known to be single-stranded for some time such as regions near DSBs[107], the lagging strand during DNA replication[108], the non-transcribed strand during transcription[109] as well as in regions susceptible to certain secondary DNA structures[110]. Furthermore, it has been shown that carriers of germline copy number deletions of *APOBEC3B* have increased levels APOBEC directed mutagenesis[111].

Interestingly, it has been shown that APOBEC3 proteins play a role in creating locally hypermutated sites[5, 107]. These mutational showers, termed as kataegis (greek for thunderstorm)[111], consist of clustered mutations and were identified in many different cancer types[15, 8]. While the C>T mutations are likely replication-induced as described above, C>G mutations are suggested to be formed when the uracil is excised via the uracil-DNA glycosylase and a cytosine is inserted opposite the abasic site by the mutagenic translesion polymerase REV1. Deficiencies in these two genes (*UNG* and *REV1*) were reported in yeast experiments to lead to a significantly decreased number of kataegis loci while the total mutation load was increased[113]. Furthermore, recently a common mutational pattern has been described, which is generated by APOBEC3 proteins and directed via the activity of MMR. The mutational pattern consists of a diffuse set of clustered mutations, which is why it has been termed as omikli (greek for fog)[16].

1.2.3 DNA Damaging Agents

There are many agents causing mutations in our genomes and in this section I will focus on some frequent DNA damaging agents: ultraviolet radiation, tobacco smoking, reactive oxygen species, and chemotherapeutics (Figure 1.5). Mutational patterns caused by these factors have been identified in many large-scale mutational tumor analyses[6, 14, 27, 114, 115]. Still, it should be noted that many other DNA damaging agents exist. For instance, Kucab *et al.* performed in 2019 a comprehensive analysis of 79 different environmental drugs and showed that 41 out of 79 drugs generated specific substitution signatures in the genome[116].

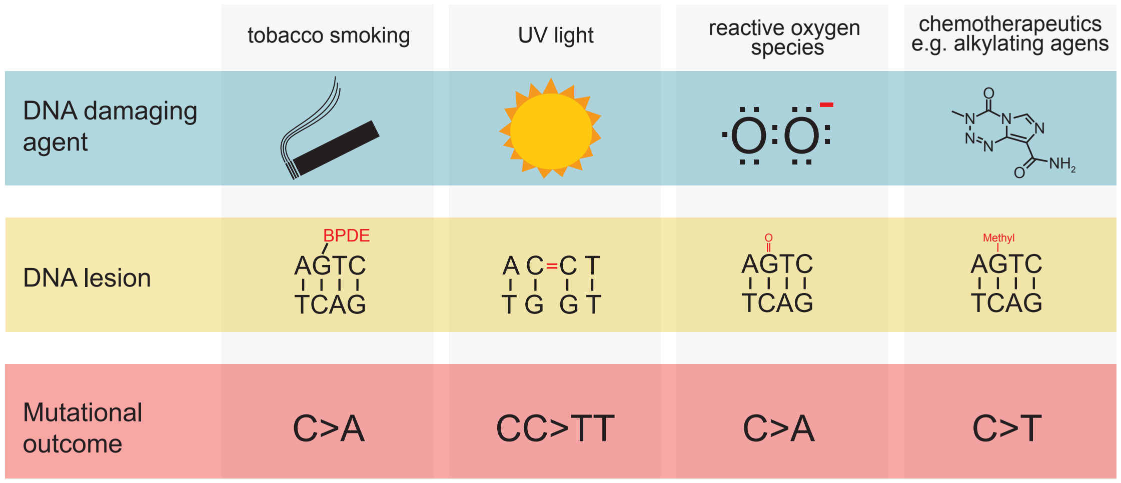


Figure 1.5: Overview of prominent DNA damaging agents. Showing different DNA damaging agents, their caused DNA lesion and a potential mutational outcome.

Ultraviolet radiation

UV-induced DNA damage leads primarily to the generation of cyclobutane pyrimidine dimers (CPD) and pyrimidine-pyrimidone (6–4) photoproducts (6-4PP). GC-NER as well as TC-NER are involved in the repair of UV-induced DNA lesions[103]. Notably, sun-exposed cancers exhibit a drastic increase in C>T and CC>TT mutations[117]. C>T mutations are enriched in a CCN and TCN context. Two factors are potentially responsible for this context dependency. First of all, the spontaneous deamination rate of cytosines and 5mCs are significantly increased at CPDs. It was shown that the deamination rate is context dependent with an increased deamination rate in a TCG context compared to a CCG context[118]. Secondly, repair of UV-induced mutations is performed by DNA polymerase η (POLH) via translesion synthesis. While POLH is able to synthesize efficiently over CPDs and thus, replicate UV-damaged DNA[88], it was reported that POLH preferentially creates C>T mutations in the TCG sequence context[119]. In fact, it was shown that mice with inactive POLH had a decreased number of C>T mutation in a TCG context, but overall an increased number of C>T mutations[119]. Thus, the context dependent formation of C>T mutations could be explained by the fact that the formation of CPDs is favoured at 5mCs and 5mCs quickly deaminate at CPDs (especially at TCGs[118]). Deaminated 5mCs result in a thymine and C>T mutations are caused by the "error-free" incorporation of an adenine opposite the thymine by POLH[119].

Tobacco smoking

Tobacco smoking causes bulky DNA adducts on guanines, which are reported to be generated by benzopyrene-7,8-diol-9,10-epoxide (BPDE). BPDE is the chemically active metabolite of benzo(a)pyrene, which can be commonly found in tobacco smoke[103]. BPDE binding on guanines is enhanced at methylated CpG sites and primarily causes C to A transversions[120]. Especially lung cancers often exhibit hotspot mutations in the tumor suppressor gene *TP53*, which have been reported to be induced via tobacco induced mutagenesis[121]. In addition, also other tobacco agents have been reported to enhance tumorigenesis such as 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) and N'-nitrosonornicotine (NNN)[122].

Reactive oxygen species

Reactive oxygen species (ROS) comprise highly reactive metabolites formed by oxygen (O_2) such as singlet oxygen ($^1O_2$) or hydrogen peroxide (H_2O_2). In healthy cells, ROS are in balance with antioxidants. Oxidative stress alters this balance leading to the generation of more ROS and/or less antioxidants. Oxidative stress is caused by exogenous factors such as UV light and viral infections, or by endogenous processes such as mitochondrial oxidative stress[123].

ROS have an important role in the context of cancer due to their high mutagenic potential and regulation of inflammation among others. The most common base alteration caused by ROS is 8-oxoguanine. Normally, 8-oxoguanine mispairs with an adenine, leading to a G>T substitution if unrepaired[124]. As described before (Table 1.1), the main players repairing this alteration are the DNA glycosylases OGG1 and MUTYH which are part of BER. Biallelic inactivation of *MUTYH* has been connected to the inherited disease MAP, which is characterized by an increased risk of developing colorectal cancer[78]. Furthermore, mutational analysis led the identification of specific mutational patterns occurring in tumor genomes with inactivations in *MUTYH*[125] and *OGG1*[126]. Interestingly, the mutational patterns generated by inactivations of these genes are defined by C>A mutations, but with enrichments in different trinucleotide contexts[125, 126].

Chemotherapeutics

Another important group of DNA damaging agents are chemotherapeutics, which play particularly an important role in the treatment of cancer. Lots of different types of chemotherapeutic agents exist (e.g. alkylating agents, antimetabolites and topoisomerase inhibitors) and they are commonly used as a non specific cancer treatment strategy leading to a broad range of side-effects. Furthermore, cancer resistance against many chemotherapeutic drugs is a frequent cause of treatment failure[127].

Many chemotherapeutic drugs induce DNA damage leading to programmed cell death such as alkylating agents. Alkylating agents such as temozolomide (TMZ) covalently bind to the DNA via their alkyl group resulting in methyl groups on guanines and adenines. In particular, methylation of guanine at O^6 can lead to the insertion of a thymine opposite the guanine, which can ultimately result in cell cycle arrest and apoptosis. TMZ has been widely used in the treatment of brain cancers such as glioblistoma multifome[128]. The main action of TMZ resistance has been connected to the O^6-methylguanine methyltransferase (MGMT), which can repair the TMZ induced base lesion O^6-methylguanine back to guanine. In accordance, a correlation between TMZ sensitivity and MGMT activity has been shown[129].

Another important group of chemotherapeutics are platinum-based drugs such as cisplatin and carboplatin. Platinum-based drugs induce different types of crosslinks in the DNA which ultimately lead to programmed cell death. They have been applied in cancer therapeutics since the 1970s for the treatment of diverse cancer types such a ovarian, lung, and bladder cancers[130].

Fluorouracil (5-FU) is another commonly used drug, especially in the treatment of colorectal as well as breast cancers. 5FU belongs to the group of antimetabolites. The exact mode of action is still unclear, but it is an analogue of uracil and it is believed to inhibit the thymidylate synthase (TS). Inhibition of TS leads to a decrease in deoxythymidine triphosphate (dTTP), resulting in a imbalance in the nucleotide pool. Lack of dTTP could impair DNA synthesis and consequently, lead to apoptosis[131].

Similarily like for the other DNA damaging agents, mutational patterns caused by chemotherapeutic drugs have been identified in tumor genomes[6,14,27,114,115]. In particular, comprehensive studies have described mutational patterns specific for different cancer therapeutics such as 5-FU, carboplatin, oxaliplatin, and even for specific drug combinations such as cisplatin+oxaliplatin[27,115].

1.3 Somatic Mutation Features in the Cancer Genome

The number, type and distribution of mutations in the genome are the consequence of the interplay(s) between different processes such as DNA repair and DNA damage[42]. Sometimes these patterns can be informative about the underlying mutational process(es)[132,144]. During the last years, the comprehensive mutational analysis of tumor genomes led to the discovery of many mutational patterns. The underlying cause(s) of some of these patterns have been identified, while the cause for many other mutational patterns remain to be elucidated[6,14,132,144]. In this section, the different identified patterns will be further described with respect to their characteristics, underlying cause, and occurrence in different cancer types. The focus will be set on the mutational features which were also used in this thesis.

1.3.1 Mutational Signatures

The concept of mutational signatures was initially introduced in 2012 using 21 whole-genome sequenced breast cancer samples[5] and further extended in 2013 using over 7,000 cancer samples covering 30 different cancer types[6]. Since then the methodology has been further improved, sample sizes have increased and a better understanding of the underlying mutational processes in the cell has been achieved[132].

Mutational signatures describe mutational patterns in the genome and it has been hypothesized that these patterns are the result of the mutational processes active in the cell. The initial mathematical framework was introduced by Alexandrov *et al.* in 2013[133]. The method involves the deconvolution of a mutational catalogue M into its signatures S and exposures E (Figure 1.6). The mutational catalogue M can be expressed as a matrix counting the different mutation types across all tumor genomes. While all kinds of mutation types can be utilized, there has been a focus on SNVs. The framework of Alexandrov *et al.*[133] involved 96 mutations types covering the 6 possible substitutions types (C>A, C>G, C>T, T>A, T>C, and T>G) and the sequence context using the proximal 5' and 3' base next to the mutated base (6*4*4 = 96 possibilites). The resulting signature matrix S contains the extracted mutational signatures and the contribution of each mutation type to each signature (e.g. Signature 1 in

Figure 1.7). The exposure matrix E involves the contribution of each signature to each sample. It considers how some mutational processes are more active in some tumors than in others (e.g. UV-directed damage). Ultimately, the total mutation burden is the sum of all operative mutational signatures weighted by the corresponding exposures ('activities'). In addition, it is expected that the mutational catalogue contains some noise in the form of sequencing errors and genuine signatures not captured by the NMF. Thus, the product of the two matrices approximates the total mutation burden in each tumor genome and is not exact (M $\approx$ E*S)[133].

$$
\begin{array}{c|ccccc}
 & \text{mutation type 1} & \text{mutation type 2} & \text{mutation type 3} & \text{mutation type 4} & \text{mutation type 5} \\
\hline
\text{tumor 1} & 23 & 4 & 20 & 2 & 1 \\
\text{tumor 2} & 22 & 4 & 0 & 1 & 2 \\
\text{tumor 3} & 0 & 33 & 2 & 1 & 3 \\
\text{tumor 4} & 2 & 1 & 1 & 3 & 3 \\
\text{tumor 5} & 4 & 0 & 0 & 0 & 2
\end{array}
\approx
\begin{array}{c|ccc}
 & \text{signature 1} & \text{signature 2} & \text{signature 3} \\
\hline
\text{tumor 1} & 85 & 9 & 4 \\
\text{tumor 2} & 0 & 42 & 13 \\
\text{tumor 3} & 0 & 0 & 118 \\
\text{tumor 4} & 3 & 12 & 4 \\
\text{tumor 5} & 0 & 10 & 0
\end{array}
*
\begin{array}{c|ccccc}
 & \text{mutation type 1} & \text{mutation type 2} & \text{mutation type 3} & \text{mutation type 4} & \text{mutation type 5} \\
\hline
\text{signature 1} & 0.2 & 0 & 0.2 & 0 & 0 \\
\text{signature 2} & 0.4 & 0 & 0 & 0.1 & 0.1 \\
\text{signature 3} & 0 & 0.3 & 0 & 0 & 0
\end{array}
$$

Mutational catalogue M   Exposures E   Signature Matrix S

Figure 1.6: Illustration of extraction of mutational signatures using NMF.. Mutational matrix M containing 5 tumors and measuring 5 different mutations types is deconvolved into an exposure matrix E and signature matrix S extracting 3 mutational signatures. Approach introduced in ref[5].

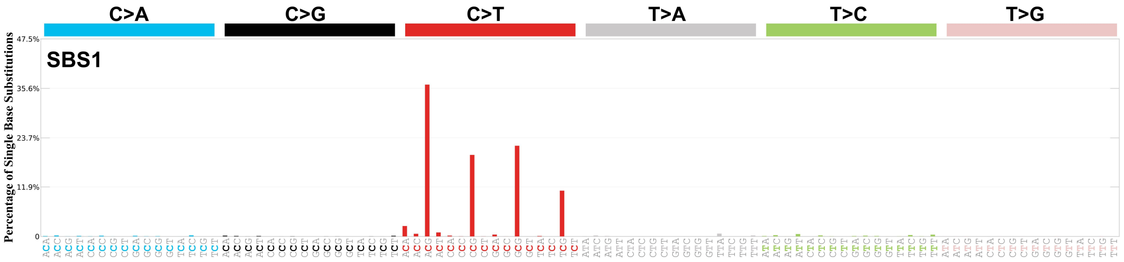


Figure 1.7: Mutational profile of COSMIC signature 1 (v3.2 - March 2021).. Signature profile showing percentage of contributions of each trinucleotide context (x-axis) to signature (y-axis). Signature 1 has been connected to spontaneous deamination of 5-methylcytosines. Figure downloaded from COSMIC[4].

Deconvolution is performed via non-negative matrix factorization (NMF). Techniques similar to NMF are princinpal component analysis (PCA) and independent component analysis (ICA). In contrast to PCA and ICA, NMF enforces all values to be non-negative and does not impose constraints such as orthogonality[134].

The application of NMF has been promoted by its success in other fields such as astronomy. Since the publication of the original mutational signature extraction framework by the Wellcome Sanger Institute[133], extensions such as bayesian NMF[135] as well as other frameworks, for instance based on independent probabilistic modelling[136] or denoising sparse autoencoder neural networks[137], have been developed.

An important number of the mutational signature extraction framework is the number of signatures, which needs to be defined *a priori*. Since the number of signatures is usually not known *a priori*, most extraction frameworks test several component extraction numbers and select the smallest number of signatures optimizing some metric such as silhouette index or Frobenius reconstruction error[133, 134]. Finding the optimal number of signatures is not a trivial task and also dependent on the selected statistical framework[134].

Nowadays, mutational signature analysis is a routine technique applied in the field of cancer genomics. Basically, two different approaches exist to assign mutational signatures to the samples present in ones dataset. In the first approach, signature extraction is performed from scratch ('de-novo extraction'). The identified signatures are then correlated with the existing set of signatures and assigned accordingly. Those are then fitted to the individual samples via 'signature fitting'. The second approach involves the usage of a set of already identified signatures (e.g. from Alexandrov *et al.*[14]) without performing the extraction and subsequent fitting to the mutational data in the dataset. As it will be explained in the next section, both approaches have some pitfalls which need to be taken into consideration[132].

Limitations and improvements

The study of mutational signatures in the last years shed light to the limitations of this approach. Major limitations, which have often been described are the following:

- **Power to extract signatures**: An important factor for the identification of mutational signatures is the sample size. For instance, while in the 2013 landmark paper[6] 21 SNV-based mutational signatures were extracted, in the 2020 landmark paper[14] 49 different SNV-based mutational signatures were extracted, which can be mainly attributed to the increased sample size ($\sim$9 times more whole-genome tumor samples). Another important factor to consider is the sequencing technology. Samples from whole-genome sequenced (WGS) tumor genomes usually have over thousands of SNVs and hundreds of indels. In comparison, the exome only makes up 1-3 % of the genome and thus, whole-exome sequenced (WES) samples have far less mutations. This makes it more difficult to reliably extract mutational signatures in WES data. For example, in the PCAWG study, only 17 SNV-based signatures were extracted using WES data, while 48 SNV-based signatures were extracted when using WGS data even though the sample size of the WGS data was approximately 4 times smaller[14].

- **Biases from the dataset**: Imbalanced datasets introduce potential biases by overrepresenting certain mutational patterns. This becomes increasingly important when tumor samples from different tissue types are aggregated. In this way, signatures only operative in a small fraction of the data can get lost. This problem was addressed in the work of Degasperi *et al.*, in which the authors suggest restricting the signature extraction to individual tissues[138]. Furthermore, cases in which samples from the same individual were taken at different time points or from different sites (e.g. biological replicates) should only be included once in the dataset for the de-novo extraction. Otherwise, this could also introduce a potential bias[132].

- **Flat signatures**: While some signatures have distinct peaks at specific mutation classes, other signatures are more flat and less distinctive (e.g. COSMIC signatures 3, 5, and 8)[139]. These signatures are more likely to be miss-assigned[132]. In these cases, it is necessary to further check other genomic features, which have been connected with the corresponding signature and check to what extent the signature assignment aligns with previous findings[139].

- **Localized processes**: Usually, the information about where in the genome the mutation happened gets lost. Hence, mutational processes, which affect a large fraction of the genome usually still get identified, while mutational processes only active in a specific region can get missed. It has been shown that highly localized processes such as the somatic hypermutation event, which occurs in B-cell receptors, can be identified by restricting signature extraction only to specific regions in the genome[139].

- **Signature bleeding**: Samples with an extremely high number of mutations (e.g. due to *POLE* exonuclease mutation), which are generated by a very mutagenic process, can sometimes *bleed* over into other signatures. For this reason, these hypermutated samples are often removed from the dataset or the mutation numbers are capped[139, 132].

- **Signature overfitting**: During signature fitting, signature overfitting can occur when the user uses all existing signatures. The existing strategies to extract mutational signatures are mathematical and do not consider biology[139]. Thus, it has been advised to only use those signatures for fitting, which have also been identified in the respective tissue[138]. Otherwise, a signatures which does not exist in the respective tissue might be assigned to a sample, leading to misinterpretations[132].

Signature extraction by Degasperi *et al.*[138]

Many of the major issues in mutational signature analysis have been addressed in the framework suggested by Degasperi *et al*[138]. De-novo mutational signature extraction was performed in each tissue/organ separately using over 3,000 WGS tumor samples, leading to the delineation of organ-specific signatures. These signatures were then fitted to the samples according to the matching tissue. In order to perform a global analysis and to compare the over 190 organ-specific signatures, all signatures were clustered together. In this way, a reference set of global signatures was generated (38 SNV+indel based signatures and 20 signatures based on structural variants), which were compared to the widely used COSMIC signatures[4]. Furthermore, signature fitting was performed via a bootstrap-based method, which produces a distribution of exposure assignments for each signature to each sample. The final exposures were calculated by taking the median values and importantly, exposures were set to 0 if they did not reach a specific statistical threshold. This approach was implemented to prevent miss-assignments. It was shown that the number of unassigned mutations rose when more signatures were used during signature fitting. This result implies that *a priori* knowledge of the contributing signatures in the corresponding tissue is crucial[138]. This framework of using organ-specific cancer signatures for signature fitting and converting them to a reference set for the global analysis, was also implemented in this thesis.

Mutational signatures based on different mutation classes

The whole concept of mutational signatures is based on the different types of mutations and the potential information they carry. Thus, the classes of mutations which are initially used can be crucial to extract biologically relevant signatures.

At the beginning, studies put the focus on SNVs and the 96 different mutational classes[6]. By now, mutational signatures have been also extracted based on indels[14], double nucleotide variants (DNVs)[14], clustered mutations[15, 16], copy number variants (CNVs)[17] and structural variants (SVs)[18]. In addition, mutational signatures based on SNVs extending the 96-channel trinucleotide context to a 1,536-channel pentanucleotide context have been generated as well[14]. Another interesting extension has been the consideration whether mutations occurred on the transcribed strand or on the untrabscribed strand[133]. On top of that, mutational signatures have been extracted combining the different mutational classes (SNVs[5, 6, 14], DNVs[14], indels[14], clustered mutations[15, 16], and SVs[18]) and also adding other genomic properties such as epigenetic marks and nucleosome states among others[115].

Single base substitution signatures

SNV-based mutational signatures are still the most widely extracted and used signatures. In 2013, 21 SNV-based signatures were reported[6], in 2020, 49 signatures were identified and in the latest release of COSMIC (v3.2, March 2021), 60 signatures were reported. During the time, some signatures got split into multiple ones and other signatures got further refined. For instance, the UV-induced signature SBS7, got split into SBS7a, SBS7b, SBS7c, and SBS7d in the latest release. Experimental efforts are onging to further validate these signatures[243, 126] and for many signatures it remains to be elucidated to which extent they cover real biology and not mathematical artefacts[132]. In particular, the causes for around 19 out of the 60 signatures are still unknown (based on COSMIC v3.2). An overview of the different signatures and their associated cause is shown in Table 1.2.

Interestingly, defective MMR has been associated with several signatures: SBS6, SBS15, SBS21, SBS26, and SBS44. In addition, a signature due a deficiency in MMR in combination with a defective DNA polymerase ε proofreading activity (SBS14) as well as a signature due a deficiency in MMR in combination with a defective DNA polymerase δ proofreading activity (SBS20) has been reported (COSMIC v3.2). The question remains whether these diverse signatures capture different types of MMR deficiencies, tissue specificities, interactions of MMR with other processes/drugs or whether they are just artefacts from the NMF framework[132]. In experimental work, knock-out cell lines covering different DNA repair genes (among others) were generated and mutational signatures were extracted

Table 1.2: Overview of SNV based signatures and their associated cause based on COSMIC v3.2. ROS: reactive oxygen species, AID: activation induced deaminase, TMZ: temozolomide.

Signature(s)	**Underlying Process/Cause**
SBS1	spontaneous deamination of 5-methylcytosine
SBS2 & SBS13	APOBEC activity
SBS3	defective HR
SBS4 & SBS92	tobacco smoking
SBS5	unknown, correlates with age
SBS6, SBS15, SBS21, SBS26 & SBS44	defective MMR
SBS7a, SBS7b, SBS7c & SBS7d	UV exposure
SBS9	DNA polymerase η
SBS10a & SBS10b	defective DNA polymerase ε proofreading
SBS10c & SBS10d	defective DNA polymerase δ proofreading
SBS11	TMZ treatment
SBS14	defective MMR + defective DNA polymerase ε proofreading
SBS18	ROS
SBS20	defective MMR + defective DNA polymerase δ proofreading
SBS22	aristolochic acid exposure
SBS24	aflatoxin exposure
SBS25	chemotherapy treatment
SBS29	tobacco chewing
SBS30	defective BER due to *NTHL1* mutations
SBS31 & SBS35	platinum chemotherapy treatment
SBS32	azathioprine exposure
SBS36	defective BER due to *MUTYH* mutations
SBS38	indirect UV exposure
SBS42	haloalkane exposure
SBS84	AID activity
SBS85	indirect effects of AID activity
SBS86	unknown chemotherapy treatment
SBS87	thiopurine chemotherapy treatment
SBS88	colibactin exposure
SBS90	duocarmycin exposure
rest	unknown

in absence of any exogenous process[126]. Interestingly, knowdowns of *MSH2*, *MSH6*, and *MLH1* generated similar mutational signatures, which correlated with SBS6, SBS20, and SBS44[126] (merged in Ref.Sig. MMR1 in Degasperi *et al.*[138]). In contrast, knockdown of *PMS2* led to a mutational signature correlating with SBS12 and SBS26[126] (merged in Ref.Sig. MMR2 in Degasperi *et al.*[138]). Thus, at least two different types of gene-specific MMR deficiencies exist, which lead to an increase in SNVs. In the future, experimental work combined with the analysis of cancer-derived signatures could support further illuminating the potentially different mechanisms.

Indel signatures

Indel signatures have been extracted based on 83 different indel types, considering the type of indel (insertion or deletion), length of indel, location of indel (homopolymers, repetitive tracks or microhomology sequence) and/or affected nucleotide (cytosine or adenine)[14,132]. In the current set of indel signatures based on COSMIC v3.2, 18 signatures have been discovered and 9 out of the 18 signatures have an unknown cause as shown in Table 1.3.

Table 1.3: Overview of indel based signatures and their associated cause based on COSMIC v3.2.

Signature(s)	Underlying Process/Cause
ID1	slippage of nascent strand during DNA replication
ID2	slippage of template strand during DNA replication
ID3	tobacco smoking
ID6	defective HR
ID7	defective MMR
ID8	DSB repair by NHEJ or *TOP2A* mutations
ID13	UV exposure
ID17	*TOP2A* mutations
ID18	colibactin exposure
rest	unknown

Double base substitution signatures

Signatures based on DNVs have been extracted using 78 different types of double base substitutions. DNVs only make up approximately 1 % of the SNVs, which is why it is more difficult to detect DNV based signatures reliably[14]. The current COSMIC v3.2 set reports 11 different DNV based signatures. An overview of the signatures and their connected mutational process is depicted in Table 1.4.

Table 1.4: Overview of DNVs based signatures and their associated cause based on COSMIC v3.2.

Signature(s)	Underlying Process/Cause
DBS1	UV exposure
DBS2	tobacco smoking or acetaldehyde exposure
DBS3	defective DNA polymerase ε proofreading
DBS5	platinum chemotherapy treatment
DBS7 & DBS10	defective MMR
rest	unknown

Copy number based signatures

Several cancer types such as ovarian, esopagheal, prostate, non-small-cell lung and triple-negative breast cancers are dominated by many copy number changes[140]. This has motivated the extraction of copy-number based signatures. So far most of these analyses restricted to specific cancer types[17,141]. In a notable study from Macintyre *et al.*[17], seven copy number signatures were extracted from over 100 WGS ovarian cancer samples. The different signatures could be associated to different mechanisms such as *CDK12* inactivation, or a deficiency in HR via mutations in *BRCA1* or *BRCA2*[17].

Structural variant based signatures

The extraction of signatures based on SVs has been hampered in the past by the difficulties in calling SVs accurately[142]. As part of the PCAWG study, SV based signatures were extracted and in total, 16 different signatures were identified. The PCAWG group used different classes of rearrangements such as tandem duplications and deletion events and further divided them by size, replication timing domain, and occurrence at fragile sites[18].

TensorSignatures

While most signature detection tools extract signatures using each mutation class separately (e.g. SNV, DNV, CNV), recent advances have integrated different mutational classes to learn signatures across different classes of mutations[115,143]. In the case of TensorSignatures, signatures are learned utilizing SNVs, multi nucleotide variants (MNVs), indels, SVs, and several genomic features such as replication timing, direction of replication, and epigenetic marks[115]. Applying the TensorSignatures framework, 20 signatures were extracted from the PCAWG dataset ($\sim$2,700 whole genomes), out of which many signatures were replicated in an independent cohort[115]. In particular, it was shown how different mutational classes and genomic properties contributed to many mutational processes and thus, providing a more refined reflection of each mutational process[115] in comparison to the for example SNV-based signatures[6,14]. In addition, the framework has been implemented in the user-friendly and highly efficient TensorFlow backend, which will make it possible to run the tool in the future on even bigger WGS datasets and to further expand the list of genomic features[115].

1.3.2 Relative Mutation Rates across the Genome

Another important approach to extract mutational patterns are relative mutation rates, which can be measured from regional mutation densities[144]. It has been shown that due to diverse mutational processes, local mutation rates differ at multiple resolutions (from 1 bp to 10^6 bp) (Figure 1.8). Understanding relative mutation rates has helped to better understand the mechanisms creating mutations in the genome, especially, in sites increasing the risk of cancer such as oncogenes and tumor suppressor genes[144]. In this section, several features associated with variations in mutation rates will be described with a focus on the features which are also used in this thesis.

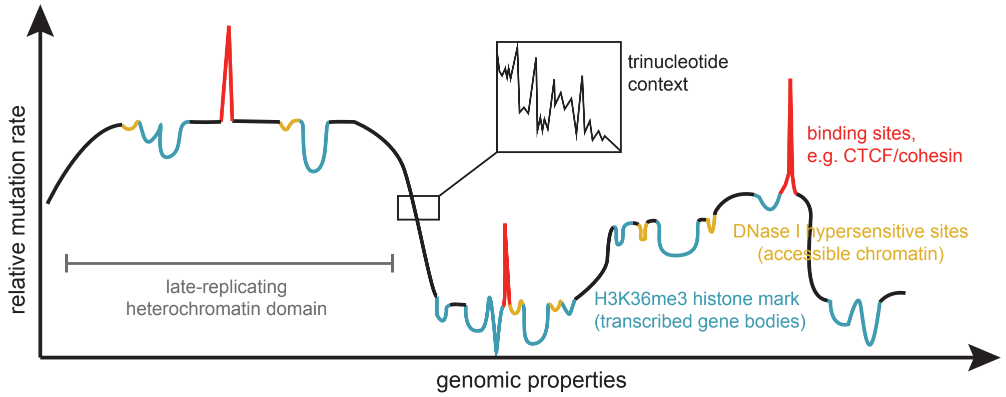

relative mutation rate
trinucleotide context
binding sites, e.g. CTCF/cohesin
DNase I hypersensitive sites (accessible chromatin)
H3K36me3 histone mark (transcribed gene bodies)
late-replicating heterochromatin domain
genomic properties

Figure 1.8: Relative mutation rates vary along the genome. Local mutation rates differ along the genome at different scales and have been associated with different genomic properties[144]. Figure adapted from ref[144].

Replication timing

Mutation rates vary at the scale of megabases, which has been shown to correlate robustly with replicating timing (RT) even after controlling for other confounders[15,21,145]. Mutation rates have been reported to be increased in late replicating regions and decreased in early replicating regions. The cause for this variability has been connected to MMR activity by showing that the variability gets lost in MMR deficient cells, leading to ’flat’ mutation distributions[21] (Figure

1.9). Since early replicating regions are associated with euchromatic regions and late replicating regions are associated with heterochromatic regions, chromatin accessibility has been also associated with this mutation rate variability[144]. However, just recently it has been shown that replication timing is responsible for the organization of the epigenome, and thus, further emphasizing the important role of replication timing[146].

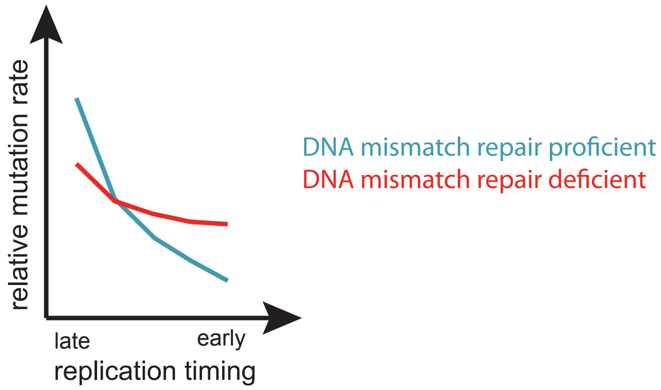


Figure 1.9: Relative mutation rates with respect to replication timing are affected by DNA mismatch repair. Deficiency of DNA mismatch repair leads to a 'flatter' genome, in particular, due to the decreased DNA repair in early replicating regions[21].

DNase Hypersensitivity I sites

DNase Hypersensitivity I sites (DHS), which mark accessible DNA regions, have also been connected with variations in mutation rates[22]. A reduced mutation rate was observed in DHS regions (intergenic, intronic, coding, and coding flanks) in comparison to flanking regions and the rest of the genome in skin, lung, colon, and bone marrow cancers[22]. The reduced mutation density in DHS regions also held true after accounting for replication timing, GC content, sequencing errors and other potential biases[22]. The activity of GG-NER was connected to the reduced mutation rates since melanoma with somatic mutations in different NER components showed increased mutations at DHS sites. Still, it should be noted that the study was performed with limited data (34 WGS samples) and it could not be excluded that also other mechanisms might play a role here[22] . In addition, a recent analysis, which utilized over 2,419 whole-genomes to measure mutation enrichments across different regulatory elements, reported an enrichment of mutations in open-chromatin regions in several cancer types such as breast, liver, and prostate cancers[147]. Thus, mutation rate variability at accessible DNA regions also seems to be tissue-specific[22, 147].

H3K36me3-marked regions

During gene transcription elongation, the methyltransferase SETD2 is recruited to the RNA polymerase II via its Ser2 phosphorylated C-terminal domain. Consequently, transcribed gene bodies often have H3K36me3 marks, which are enriched in the 3' ends[148]. Mutation rates have been reported to be decreased in H3K36me3-marked regions, which has been connected to MMR activity[20, 15]. Reportedly, histone mark H3K36me3 can recruit the mismatch recognition protein MutSα (heterodimer of MSH2 and MSH6) via binding of MSH6[20]. In this way, gene bodies are enriched during replication with components of the MMR pathway leading to an increased DNA repair efficiency in these regions[20, 15]. The causal role of H3K36me3 mark in increasing the DNA repair efficiency was further shown by controlling for other confounders such as replication timing and by showing that the effect was reduced in MMR deficient genomes[20].

Transcriptional strand asymmetry

Another frequent pattern found in cancer genomes is strand asymmetry with respect to the transcribed strand and the non-transcribed strand[24]. The strongest strand biases were reported in liver cancers, lung cancers and skins cancers. In skin cancers, UV-induced C>T mutations were decreased when the cytosine was on the transcribed strand, and in lung cancers tobacco smoking-induced C>A mutations were decreased when the guanine was on the transcribed strand. In both cases, strand asymmetry was attributed to the activity of TC-NER, which repairs lesions occurring on the transcribed strand. In liver cancers a strong strand asymmetry was reported with respect to T>C mutations. Interestingly, on top of having less T>C mutations when the adenine was on the transcribed strand, an increase in T>C mutations was reported with increasing gene expression levels. Normally, it would be expected that the number of mutations decrease with rising gene expression levels due to the activities of GG-NER and MMR, which are more efficient in high expressing regions. Thus, it has been suggested that the strong strand asymmetry for T>C mutations in liver cancer can be attributed to an expression-level dependent increase in DNA damage ('transcription-coupled damage') in combination with TC-NER, which might be independent of the DNA damage inducing process. The cause for the transcription-coupled damage is still unclear[24].

Replicative strand asymmetry

In a similar vein like in the case of transcriptional strand asymmetry, strand asymmetry has been also measured and observed with respect to the replication mode of the strand: leading and lagging[23,24]. While the leading strand is replicated continuously, the lagging strand is replicated discontinuously via Okazaki fragments. Mechanisms contributing to the replicative strand asymmetry are reportedly MMR, APOBEC, and the proofreading activites of DNA polymerase ε and DNA polymerase δ[24,108,149]. All cancer genomes with increased APOBEC activity such as breast, bladder and head and neck cancers were reported to have a strong replicative strand asymmetry with respect to C>T and C>G mutations when the cytosine was located on the lagging strand. The increase in APOBEC activity has been attributed to the increased exposure of the lagging strand to ssDNA[24,108], and additionally due to differential activity of MMR across the two strands[16]. Furthermore, strong strand asymmetry with respect to C>A mutations were observed in samples with mutations in the exonuclease domain of DNA polymerase ε when the cytosine was on the leading strand. This observation has been connected to the proofreading activity of DNA polymerase ε, which acts primarily on the leading strand[24,108]. The opposite effect was reported for samples with mutations in DNA polymerase δ, which acts primarily on the lagging strand[149]. MMR deficient cells showed similar replicative strand biases as samples which had mutations DNA polymerase δ and it has been suggested that MMR repairs mutations occurring on the lagging strand more efficiently than mismatches occurring on the leading strand[149].

Nucleosomes

An important feature of eucaryotic genomes are nucleosomes, which build the first layer of DNA packaging. In brief, DNA is wrapped around a histone octamer in 1.67 turns, forming one nucleosome. A nucleosome core is then connected via a stretch of DNA ('linker DNA'), to the next nucleosome. Somatic mutation rates have also been reported to correlate with nucleosome occupancy[150]. In a systematic analysis of over 3,400 tumors it was shown, that UV-induced mutations are enriched in nucleosomes in comparison to linkers, which has been attributed to the increased efficiency of GG-NER in the DNA linkers. In contrast, tobacco smoking-induced mutations are enriched in linkers, because binding of the chemically active agent BPDE is prevented at nucleosomes[150]. Furthermore, relative mutation rates also differ within nucleosomes with respect to the orientation of the DNA. DNA wrapped around nucleosomes can be divided into two structurally different regions: minor groove facing histones and minor groove facing out histones. For instance, UV-induced CPDs preferentially occur on the minor groove facing away from the histones. While these more exposed stretches also have a higher efficiency to be repaired via NER in comparison to the minor-in stretches, the effect of DNA damage outperforms the DNA repair activities and thus, resulting in increased mutations in minor-out stretches. Additionally, also other DNA damages such as ROS and tobacco-induced mutations leave patterns behind, which lead to differences in mutation rates within nucleosomes[150].

CTCF/cohesin-binding sites

A notable example for a mutational feature occurring at the sub-gene scale are CTCF/cohesin binding sites. These sites have been observed to be frequently mutated especially in a subset of colorectal cancers[151] and skin cancers[152]. Interestingly, these mutational hotspots were independent of MMR activity and decreased in samples with mutations in the exonuclease domain of *POLE*[151]. The exact mechanism responsible for the mutational peaks is still unknown and the types of mutations differ between different cancer types. In colorectal cancer, the peaks were enriched in SBS17 mutations. The underlying mechanism for SBS17 (specifically SB17a and SBS17b) is still unknown, but might be caused by oxidized guanines in the free nucleotide pool[153]. In skin cancer, the mutational pattern was enriched in UV-induced C>T mutations[144]. While not further covered in this thesis, it should be noted that also other transcription factor binding sites are frequently mutated in cancer genomes such as binding sites of the ETS family[144, 154].

X-chromosomal hypermutation

Another mysterious somatic mutational feature is X-chromosomal hypermutation, which was reported for the first time in 2013[155]. Initially, this feature was observed predominantly in diverse brain cancers and leukemia, and later also discovered in other adult cancer types[156]. The feature is characterized by up to 4 times higher mutation rates on the X chromosome in comparison to the autosomes, involving SNVs as well as indels. Interestingly, while mutation rates are increased on the X chromosomes, the mutational spectrum has been reported to be similar as on the autosomes. Furthermore, it was shown that the event takes place on the inactive X chromosome since the feature was mostly detected in female patients and in a male patient with two X chromosomes (Klinefelter syndrome). Consistently, X-chromosomal hypermutation was not observed in female patients lacking expression of *XIST*, which is the transcript responsible for the inactivation of the X chromosome. In addition, the feature was reported to occur early in tumor evolution and was not observed in nonmalignant genomes[155]. All in all, the characteristics of X-chromosomal hypermutation are known and the feature can be easily spotted in WGS data[155]. Still, the underlying mechanism remains to be elucidated.

1.3.3 Mutational Patterns in the Mitochondrial Genome

So far, all the mutational features described were based on patterns observed in the nuclear genome. It should not be forgotten that mitochondria carry genetic material. The mitochondrial genome is circular with a size of approximately 16.5 kb[157]. The number of copies of the mitochondrial genome per mitochondria fluctuate since mitochondria constantly divide and merge[158]. In tumor genomes, mitochondrial genomes have distinct mutational patterns, which have been connected to the unique genome replication occurring in mitochondria[157]. Most notably, specific signatures found in the nuclear genome, for instance due to UV or tobacco exposure have not been identified in the mitochondrial genomes. Thus, it is assumed that the mutational processes occurring in mitochondria are mostly different to those occurring in the nuclear genome[157].

In a comprehensive analysis of over 1,600 tumor genomes, at least one somatic mutation was detected in 50 % of the samples in the mitochondrial DNA (mtDNA). The number of mutations in the mtDNA varied across tumor types and was increased in particular in stomach, liver, prostate, and colocrectal cancers[157, 159]. The most frequent mutations in mtDNA are C>T and T>C mutations, which show a strong strand bias and can be detected across all cancer types. C>T mutations are enriched on the H-strand (H for heavy), and T>C mutations are enriched on the L-strand (L for light). The cause for these mutations have been connected to the replication process, which is mostly performed by the mtDNA poylmerase (POLG). It has been shown that the mtDNA polymerase preferentially causes C>T and A>G mutations on the H-strand, which would explain the observed mutational strand bias. Furthermore, the involvement of DNA repair processes has been suggested[157].

Another interesting feature of the mitochondrial genome is the copy number variation, which has been observed across cancer types[158]. In specific cancer types such as bladder, breast, and kidney cancer mtDNA is depleted in the tumor sample in comparison to the sample-matched healthy sample[158].

It could be expected that the decrease in the mtDNA would be connected to the preference of many cancer cells to use glycolysis as their primary source of energy (Warburg-effect). While a positive correlation between copy number of mtDNA and expression of mitochondrial respiratory genes was observed, it varied between cancer types. In addition, some samples with low mitochondrial copy number had high levels in the expression of metabolic mitochondrial genes. Thus, copy number alone can not be used as a predictor of mitochondrial activity and the consequences of this variation need to be further resolved[158].

1.3.4 Somatic Mutations in Normal Cells

Several somatic mutational features which have been identified in cancer genomes, have been also identified in normal (non-tumor) cells of the body[160,161,162]. Somatic mutation rates in normal human somatic cells have been estimated across different tissues and range between 2 to 10 mutations per cell division[163], which is higher than in germ cells[163,164]. The investigation of somatic mutations in normal cells has been mostly hindered by technical difficulties to expand many different types of healthy cells in culture and/or the high error-rate in single-cell sequencing[164].

Cancer signature analysis of somatic mutations identified in adult stem cells led to the identification of three signatures[160] which were previously identified in cancer genomes[6]: signature 5[6] (cause unknown, correlating with age[238]), signature 1[5] (spontaneous deamination of 5-methylcytosine, correlates with age[238]), and signature 18[6] (caused by damage by reactive oxygen species). Further, it was shown that healthy human cells with acquired somatic mutations in the exonuclease domain of *POLE* or *POLD1* have elevated somatic mutation rates[161] as it has been also reported in tumor genomes[14,85]. In addition, in a systematic analysis of somatic mutations (SNVs) from over 36 non-disease tissues using GTEx data several somatic features, which were reported previously in tumor genomes, were also detected in normal cells such as strand asymmetries with respect to DNA transcription (transcribed vs. non-transcribed strand) and associations with chromatin states (e.g. positive correlation of somatic mutation rate with heterochromatin mark H3K9me3)[162].

Thus, analyses so far suggest that many of the somatic mutational features, which were detected in tumor genomes, have likely been 'carried over' from mutational processes already occurring in healthy somatic cells such as the spontaneous deamination of 5-methylcytosines. In the future, more studies and data will help to elucidate the mutational processes occurring specifically in healthy cells for instance by also analyzing indels, copy number variants and structural variants on top of SNVs.

1.4 Inherited Variants Affecting Somatic Events in the Tumor

Somatic mutations play a crucial role in the emergence of cancer[39]. Another important role play inherited variants[31]. Initially, germline and somatic genomes were studied separately. Similarly as in the case of somatic mutations, germline variants have been associated with specific cancer subtypes and have been correlated with clinical outcomes[31]. Furthermore, germline variants in many different genes have bee reported to predispose to cancer[165]. On top of this, over 1,300 single-nucleotide polymorphisms (SNPs) have been identified via genome-wide association studies (GWAS) to increase cancer risk[166]. In an analysis of three large-scale cancer datasets, it has been shown that individuals with an increased number of rare damaging germline variants (RDGVs) in cancer hallmark genes have an earlier onset of cancer compared to individuals with low numbers of RDGVs. In line with this observation, the accumulation of somatic mutations plays a bigger role in patients, which have a low number of RDGVs, which is why these individuals tend to develop cancer at a higher age. The hypothesis is that specific germline variants increase the susceptibility to develop cancer[166].

During the last years, the analysis of germline variants has been integrated into the analysis of somatic mutations more frequently. It has been shown, that germline variants can affect which types of somatic events occur and get selected during tumorigenesis and affect somatic mutation rates[167,168,169]. While these interactions have been shown via large-scale systematic studies[167,168,169], the initial idea that germline variants could affect the occurrence and selection of somatic mutations in the tumor genome was initially observed and formulated by Alfred Knudson in 1971[170]. Knudson predicted based on a statistical analysis of retinoblastoma patients, that the inherited form of the disease would be caused by an inherited mutation (first hit) and subsequent second somatic mutation (second hit). Later it was confirmed and discovered that mutations need to occur on both alleles of the tumor suppressor gene *Rb* to lead to cancer onset[171]. Patients without an inherited pathogenic variant in the *Rb* gene rarely develop the disease since it would be statistically unlikely that both alleles get inactivated during the lifetime of an individual, which also explains the late onset of the nonhereditary form of retinoblastoma[170]. Knudson's two-hit hypothesis was further tested and validated once large numbers of tumor genomes were sequenced. In a systematic analysis, it was shown how this concept can be used to identify new cancer predisposition genes by investigating the combination of rare germline variants with somatic loss of heterozygosity (LOH)[167]. Further, it has been shown that tumor suppressor genes can also act as one-hit drivers and more importantly, that depending on the cancer type and genomic background a cancer gene can act either as a one-hit or two-hit driver[169].

Beyond Knudon's two-hit hypothesis and the interaction of germline variants with specific somatic mutations in either *cis* or *trans*[167,168,169], it has been also appreciated how germline variants in specific genes such as components of the DNA mismatch repair pathway can have an affect on global somatic mutations rates[31]. The investigation of how germline variants affect somatic mutations rates is also the main focus of this thesis and thus, the current understanding of this interplay will be further described in the following section.

1.4.1 Rare Variants affecting Somatic Events in the Tumor Genome

Most studies distinguish between rare and common variants. Disease alleles are often enriched in rare variants because they are under negative selection in human populations and exhibit large effects (Figure 1.10). Thus, the analysis of rare variants has been widely used to understand more about the causes of different diseases[172]. In the context of cancer and somatic mutations, rare germline variants in different genes have been reported to affect the occurrence of somatic mutations in the genome[31] (Table 1.5).

Damaging variants (including germline variants) in *BRCA1*, *BRCA2*, *PALB2*, and *RAD51C* have been reported in different studies to associate with several somatic mutation features[174]: signature 3 mutations[5], deletions of a size smaller than 50 bp at microhomology sites[5], duplications with a size of more than 1 kb[94,95], a specific copy number signature[17], and several structural rearrangement based

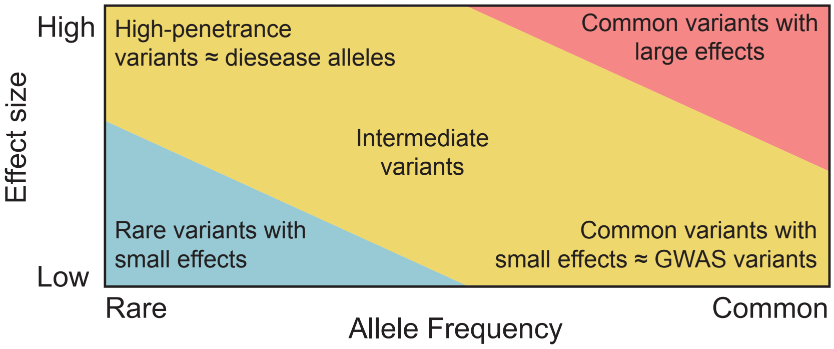

High
Low
Effect size
High-penetrance variants ≈ diesease alleles
Common variants with large effects
Intermediate variants
Rare variants with small effects
Common variants with small effects ≈ GWAS variants
Rare
Common
Allele Frequency

Figure 1.10: Illustration of genetic variant effects along allele frequencies. Rare variants (allele frequency < 0.5 %) are often enriched for disease causing alleles, while common variants (allele frequency > 1 %) which are typically identified in genome-wide association studies (GWAS), have often small effect sizes. Figure inspired from ref[173].

signatures[18] (Figure 1.11). These signatures are likely the consequence of the impairment of the cell to repair DSBs via homologous recombination, which is why the cell needs to use other more error-prone mechanisms such as microhomology-mediated end joining[175].

Table 1.5: Overview of genes in which damaging variants (including germline) have been associated with distinct somatic mutational features.

Gene	Associated Somatic Mutational Feature
BRCA1, *BRCA2*, *PALB2*, *RAD51C*	signature 3 mutations[5]
	deletions ≤50 bp at microhomology regions[5]
	duplications ≥ 1 kb[94,95]
	copy number signatures[17]
	structural rearrangement signatures[18]
MSH2, *MSH6*, *MLH1*, *PMS2*	SNV based signatures[6]
	indel based signatures[14]
	enrichment of SNVs in early replicating regions[21]
	enrichment of SNVs in regions with increased H3K36me3 marks[20]
	replicative strand asymmetry[149]
POLE	SNV based signatures SBS10a & SBS10b[14]
POLD1	SNV based signatures SBS10c & SBS10d[14]
MUTYH	SNV based signature SBS36[125]
NTHL1	SNV based signature SBS30[79]
MBD4	increase in C>T mutations at CpG sites[76,77,8]
TP53	massive chromosomal rearrangements ('chromothripsis')[178]

Similarly, pathogenic germline variants in components of the DNA mismatch repair pathway, which predispose to early-onset cancer of the colorectum and other organs (Lynch syndrome)[176], have been associated with several mutational signatures[31] as shown in Table 1.5. Pathogenic variants in genes such as *MSH2*, *MSH6*, *PMS2*, and *MLH1* have been associated with different kinds of small indels at microsatellite regions[14], different SNV-based mutational signatures[6],

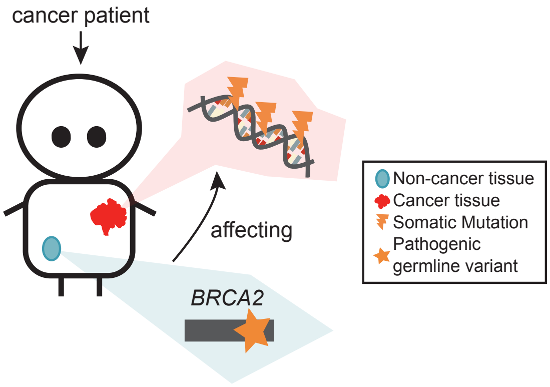


Figure 1.11: Inherited pathogenic variants can impact somatic events in the tumor. Inherited pathogenic variants in *BRCA2* have been reported among others to impact somatic mutational processes in the tumor[31].

an enrichment in early replicating regions[21], an enrichment in regions with high levels of H3K36me3 marks[20] and replicative strand asymmetry[149] (also see Sections 1.2.1 and 1.3.1).

Another interesting example is *TP53*. Inherited pathogenic variants in *TP53* cause Li-Fraumeni syndrome, which is a rare, autosomal dominant inherited disease resulting in a wide spectrum of different cancer types, which emerge before the age of 40[177]. Sequencing data from tumors from patients with inherited pathogenic variants in *TP53* uncovered how these germline variants can have disastrous effects on the genome. These tumors exhibit massive chromosomal rearrangements, termed as chromothripsis (*thripsis*: greek for shattered into pieces)[178]. Many of the listed associations influencing mutation processes include variants that cause familial cancer syndromes[165].

In addition, specific SNV based cancer signatures were extracted from patients with germline variants in the DNA glycosylases *NTLH1*[79] and *MUTYH*[125]. In the same way, cancer signatures were extracted from individuals with pathogenic mutations in either *POLE*[14] or *POLD1*[14].

So far, there have been only a limited number of studies, which explored the association of RDGVs with mutational patterns in a more systematic and unbiased approach. The biggest study to date was performed as part of the large-scale PCAWG study[8]. In the PCAWG study, over 1,600 whole-genome primary cancers of European ancestry were utilized for a rare variant association study. The association of RDGVs with different DNA rearrangement signatures, signature 1 (C>T mutations at CpG sites) and APOBEC signatures were tested and subsequently replicated in a validation dataset consisting of over 8,000 whole-exome primary cancers. RDGVs were defined as variants with a frequency of less than 0.5 %[8] (some studies use 0.1 %[167]) in the human population, which result in a protein-truncation via frameshift variants, nonsense variants, or splicing variants. For testing, a gene-based aggregation test was used by collapsing all RDGVs occurring the same gene (methodology will be described in more detail in the next section). RDGVs in *BRCA2* associated with a decreased number of C>T mutations at CpG sites and several copy number-based phenotypes[8]. RDGVs in the DNA glycosylase *MBD4* associated with an increased number of C>T mutations at CpG sites. This association was also found in several independent studies[76,77]. The *BRCA2* and *MBD4* associations with C>T mutations at CpG sites were both discovered at exome-wide significance in the PCAWG cohort and further replicated in a validation cohort (TCGA-WES)[8].

A similar approach was also applied at a smaller scale by a study which was focussed on breast cancer samples[32]. In the study, WES and WGS samples from around 1,000 individuals were utilized as the discovery cohort and a cohort consisting of around 170 individuals from Nigeria were used as validation. The association of RDGVs with 9 SNV-based mutational signatures was investigated using a predefined set of genes. While several genes were identified in the discovery cohort, none of the hits replicated at a p-value of 0.05 (without multiple testing correction) in the validation cohort. However, it should be noted that the study had a very limited sample size and that the initial discoveries would still suggest that there might be several pathogenic germline variants affecting mutational patterns, which might get replicated in higher powered studies[32].

All in all, currently only a handful of rare inherited pathogenic variants have been associated with distinct mutational patterns in the genome. A better understanding of this interaction will be crucial to better predict cancer risk, develop new therapeutic strategies and ultimately, tailor cancer treatment to individual patients. The major challenge to perform systematic rare variant association testing originates from the fact that they are rare. In the next section, I will further describe the different methodologies which can be utilized to perform a rare variant association study and how power can be increased even at lower sample sizes.

1.4.2 Methodologies for Gene-based Rare Variant Association Testing

It is challenging to perform association studies with individual variants if sample size or effect size is not big enough. Thus, it has been common practice to collapse rare variants based on different features such as domain, gene or connected pathway. In a classic burden test such as in the cohort allelic sum test (CAST)[179], rare variants within a predefined region are collapsed into a single binary value indicating whether a rare variant was found in the region or not. Subsequently, the value is regressed against the phenotype in order to test for association. This simple approach has been further extended in many different ways such as by counting the number of rare variants within a region, by collapsing the variants into several subgroups based on allele frequencies or by estimating a weighted sum based on specific assumptions[180]. A drawback of all these methods is that they assume that all variants within a region are causal and that all of them have the same effect size direction. In this way, classic burden tests are underpowered in cases in which these assumptions do not hold true. These problems have been partly addressed in so called adaptive burden tests which make less assumptions about the variants and incorporate the idea that variants within a region could have opposing effects (e.g. data adaptive sum method[181]). However, a major drawback of the adaptive burden tests is that they are computationally expensive since they make use of permutation tests[180].

A major advancement have been variance-component score tests, which evaluate the distribution of the test statistics. These tests are not burden tests anymore. First in class was the C-alpha test[182], which tests in case-control studies whether the allele frequencies have an altered variance in comparison to an expected variance. This idea was further generalised in the sequence kernel association test (SKAT)[183], which can also control for covariates and measure epistatic variant effects. SKAT assumes in the null hypothesis that the regression coefficient from each variant within a certain region follows a random distribution with the mean 0 and a variance of $weight_v * \tau$. Here $weight_v$ is the predefined weight for the respective variant and τ is the variance component. The null hypothesis tests whether $\tau = 0$. By default, the weight of each variant is set as follows: $\sqrt{w_v} = Beta(MAF_v; a_1, a_2)$. Here, *Beta* is the beta density distribution function, MAF_v the minor allele frequency (MAF) of the respective variant in the dataset and a_1 is set to 1 and a_2 is set to 25. Since it is not known *a priori*, which variants are causal, the weights are set in dependency on their allele frequencies. Thus, rare variants have a higher weight than more common variants. The default settings were chosen by the authors to give variants with a MAF between 1 % and 5 % a small, but nonzero weight. A good choice of the weights can further improve the power of the test[183]. Ultimately, it has been shown that SKAT outperforms burden tests when a region in the genome has many noncausal variants and/or variants with opposite estimate size directions. Vice versa, classic burden tests outperform SKAT in cases in which a genetic region has many causal variants with the same effect size direction[183].

Due to the different strengths of burden and nonburden tests, an approach combining the two tests has been suggested[184]. This unified approach has been executed in SKAT-O, which finds the optimal linear combination of both tests: burden test and SKAT. The test statistic is calculated as follows: $Q_\rho = \rho\ Q_B + (1 - \rho) Q_S$. Here, Q_B is the test statistic from the burden test, Q_S is the test statistic from SKAT, and ρ weights the two tests. If $\rho = 1$ the test statistic Q_ρ will follow the burden test, and if $\rho = 0$, the test statistic Q_ρ will follow SKAT. Since the optimal weight ρ is unknown, the default implementation of SKAT-O tests different settings of ρ and chooses the one which minimizes the p-value[184].

It should be noted, that no matter which test is used to find associations between rare variants with different phenotypes/traits, the initial variant set needs to be restricted when applying region-based tests. Even the nonburden tests require some information about how to weight the different rare variants. Thus, variant scoring algorithms which are able to predict pathogenicity[185,186] still have an important role in rare variant association studies in all cases in which rare variants cannot be studied independently and region-based aggregation tests need to be utilized. Most of these tools use various, diverse information to predict variant pathogenicity such as conservation data from multiple sequence alignments[187] or utilizing common variants from primate species[186]. In addition, ensemble and consensus methods have been applied such as CADD[185] and REVEL[188]. Recent approaches such as EVE (evolutionary model of variant effect)[189] or the autoregressive model by Shin *et al.*[190] make use of deep neural networks with generative properties (learn the distribution of the data). These tools help to restrict the initial set of variants and/or to assign a weight to each variant.

1.4.3 Common Variants affecting Somatic Events in the Tumor Genome

Rare variants are often studied with the goal to find disease alleles and to better understand the mechanism behind the disease. Based on evolutionary theory it is expected that highly penetrant disease alleles are rare[172]. Still, GWAS data suggests that while rare variants explain a fraction of the heritability of many diseases, they do not explain the majority[172]. This is also why the analysis of common variants can be equally important to better understand the genetic basis and mechanisms behind many diseases. In fact, via GWAS many thousands of common variants affecting disease risk have been identified and for complex diseases such as schizophrenia, the majority of the genetic variance is captured by common variants[172].

Similarly as for the analysis of rare variants affecting somatic mutations, the same research groups also explored the association of common variants with the same mutational patterns[8,32]. In the PCAWG study the association of common inherited variants (MAF $>$ 5 %) with signature 1 and APOBEC mutations was investigated. Several variants passed genome-wide significance and associated with APOBEC mutagenesis. These variants were also subsequently replicated in an independent cohort of Asian ancestry. The variant with the strongest signal (rs12628403) was reported to alter APOBEC-directed mutagenesis and to tag a germline *APOBEC3B* deletion[8], which was also identified in previous studies[111,112]. This deletion has also been reported to increase the risk of breast cancer[111]. In addition, a novel nearby QTL locus associating with APOBEC burden was identified as well[8]. In the smaller study conducted by Wang *et al.*[32] several common variants associating with APOBEC-directed mutagenesis were discovered as well. None of the top 30 SNPs replicated in the validation cohort and only 12 out of the 30 SNPs had a consistent direction in the effect size estimate. Here, it should be noted again that the limited sample size could be the reason for the negative replication of the hits. Notably, no association between common variants and other mutational signatures were identified in this study[32]. In addition to the PCAWG study[8] and the study by Wang *et al.*[32], the association of common variants with the total mutation burden was investigated by Sun *et al.* in 2021[191]. While no association reached genome-wide significance in the pan-cancer analysis, several hits were identified in specific cancer types such as breast and stomach cancer. The study design did not include a validation cohort, and thus, none of the hits were further followed up[191].

Our understanding of how common germline variants affect different mutational somatic processes is still limited. The studies conducted so far all selected a limited set of somatic mutational features and only associations between common variants and APOBEC-directed mutagenesis have been identified and replicated[111,112,8]. For the tumor-specific association between common variants and the total mutation burden, replication of the hits is still missing[191]. Furthermore, while many common cancer risk SNPs are known from case-control studies, the underlying mechanisms are largely unknown[176]. We hypothesized that some of them act via changes in mutation rates, which motivated us to test for associations with common variants.

1.4.4 Heritability of Somatic Mutational Processes

An important concept to better understand to which extent inherited variants affect a phenotype is heritability. It is used to get a better understanding of how much variance in a phenotype is attributed to the environment and how much to genetics in a particular population (at a particular time point). In the field, different definitions of heritability exist and in this study we will mostly focus on narrow-sense heritability. Narrow-sense heritability is defined as the total amount of variation in a phenotype, which can be explained via additive genetic effects. In contrast to broad-sense heritability, it does not consider genetic variance coming from epistatic effects and from dominance effects. It should also be noted here, that heritability estimates are not static. They are highly dependent on the population, the environment and they can change over time for the same population[192].

Initially, heritability was estimated by regressing the phenotypes of the parents against the one of the offspring and by looking at the difference between monozygotic and dizygotic twins. The era of genomics opened the avenue to estimate heritability from a large set of unrelated individuals. For instance, SNPs reaching significance in GWAS studies have been used to estimate to which extent they explain heritability of a trait. In most cases, SNPs which were identified from GWAS studies only explained a small amount of heritability. This could be due to the fact that the number of SNPs reaching significance in a GWAS study depends largely on the sample size and study design. Thus, many SNPs which potentially affect a trait might not be identified in a GWAS when they are too rare or the effect size is small. This is one of the major reasons why most current methods which estimate heritability from genomics data consider all (mostly common) SNPs[193]. A widely used method is GREML[194] which estimates heritability by using all SNPs and assuming an infinitesimal model. This model assumes that all SNPs contribute to heritability of the diseases via an extremely small contribution[195].

In the context of somatic mutations, a few studies have also estimated the heritability of several mutational patterns. As shown in Table 1.6, the heritability of different somatic mutational processes is still mostly unexplored.

Table 1.6: Estimated SNP-heritability of several somatic mutational features.

Somatic Feature	Heritability estimation via GREML	Study	Sample size
Total mutation burden	12.9 ± 4.7 %	Sun (2021)[191]	~7,000
APOBEC directed mutagenesis	43.2 ± 27.2 %	Wang (2019)[32]	~700
C>T mutations at CpG sites	40.0 ± 31.3 %	Wang (2019)[32]	~700
Deficiency in homologous recombination	15.5 ± 35.8 %	Wang (2019)[32]	~700

This can be largely attributed to the fact that the estimation of SNP heritability via methods like GREML requires a sample size of at least 1,000, since the error is approximately 318 divided by the sample size (error of ~32% at a sample size of 1,000). This becomes clear when looking at the estimated heritabilities from the study by Wang *et al.*, which were estimated based on ~700 individuals[32]. Interestingly, in a study which utilized pan-cancer data from over 7,000 individuals a heritability of around 13 % was estimated for the total mutation burden, which suggests that there is an important genetic component which contributes to the accumulation of somatic mutations in the genome[191].

All in all, there is a lack in understanding to which extent the different sources of mutations contain genetic components and, so far, it is not known which mutational mechanism(s) is (are) responsible for the heritability of the total mutation burden. This is also one of the questions with will be addressed in this thesis.

1.5 Aim and Objectives of this Study

Aim

While genetic screens in model organisms have shown that mutations in many different genes can influence mutational processes[196,197], our understanding of inherited variants influencing mutational processes in humans is still limited. Thus, the aim of this thesis was to gain an understanding of the extent to which inheritance plays a role in impacting mutational processes in somatic cells. For this purpose, association studies utilizing rare and common germline variants were performed and heritability of somatic mutational processes were estimated (Figure 1.12). These approaches have previously been shown to be applicable for discovering germline determinants of human somatic mutational processes[8,32,191].

Objectives

- Extract somatic mutational features/components, which recapitulate the different mutational processes occurring in humans cancers (feature engineering).
- Test for associations between rare damaging germline variants (RDGVs) and the compiled set of somatic mutational components in a gene-based testing approach.
- Perform a genome-wide association study (GWAS) testing for associations between common germline variants and somatic components.
- Assess the heritability of the different somatic mutational processes utilizing common variants.

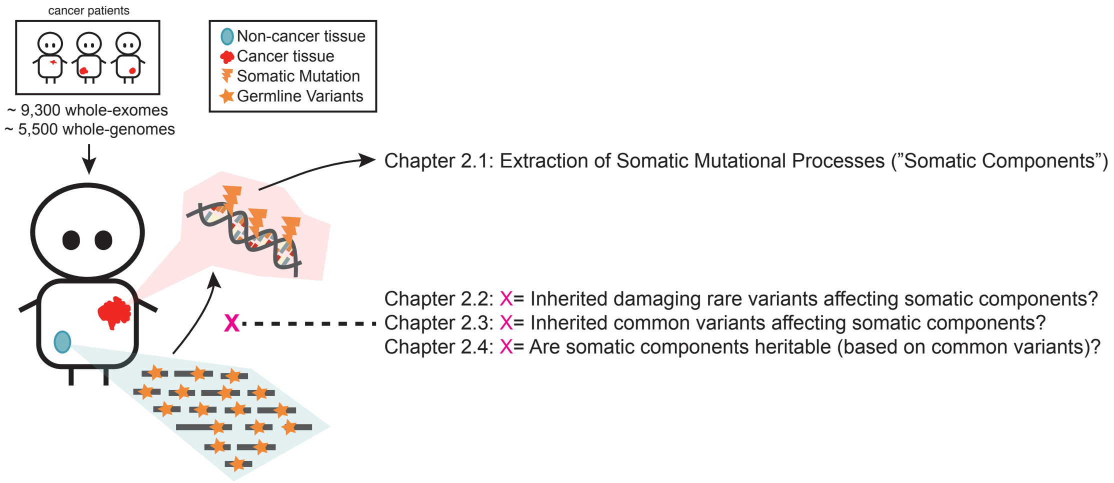

cancer patients
~ 9,300 whole-exomes
~ 5,500 whole-genomes
Non-cancer tissue
Cancer tissue
Somatic Mutation
Germline Variants
Chapter 2.1: Extraction of Somatic Mutational Processes ("Somatic Components")
X
Chapter 2.2: X= Inherited damaging rare variants affecting somatic components?
Chapter 2.3: X= Inherited common variants affecting somatic components?
Chapter 2.4: X= Are somatic components heritable (based on common variants)?

Figure 1.12: Illustration of the study design and structure of the results presented in the thesis.

1.6 Study Design

Genomic sequencing data from three large-scale projects were utilized: the Cancer Genome Atlas Program (TCGA)[7], the Pan-Cancer Analysis of Whole Genomes (PCAWG)[8], and the Hartwig Medical Foundation (Hartwig)[9]. Firstly, somatic components were extracted from around 15,000 cancer genomes. Next, associations between rare damaging germline variants (RDGVs) and somatic components were tested and associations between common germline variants and somatic components were tested (Figure 1.12). For both cases, associations were initially detected in the discovery cohort and hits reaching significance were re-tested in the validation cohort. TCGA whole-exome sequenced (WES) samples were used as the discovery cohort due to the bigger sample size and whole-genome sequenced (WGS) samples from PCAWG and Hartwig were aggregated and utilized as the validation cohort. Lastly, common variants were utilized to answer the question which mutational processes are heritable.

2 Results

2.1 Somatic Mutational Components

In this study the effects of inherited variants on different somatic mutational processes were studied. For this purpose, we generated a comprehensive list of somatic mutational features, that were partly based on the many mutational patterns, identified in cancer genomes[6, 14, 144]. Many of the generated features have been extracted from WGS data, but not from WES data[15, 17, 21, 138, 151] before. For this reason, we checked for all features whether we could detect the same effects in WES data as they were described before. Furthermore, since many of the generated features (partly) correlated with each other, different approaches were tested in order to embed all somatic mutational features in a lower dimensional space. In particular, principal component analysis (PCA), independent component analysis (ICA) and variational autoencoders (VAE) were tested. In this way, the goal was to (i) extract components reflecting the underlying causal mechanisms, (ii) increase interpretability, and (iii) improve power for subsequent association testing. Of note, parts of this work such as the extraction of somatic mutational components via ICA and VAE have been published on bioRxiv[198].

2.1.1 Features based on Single Nucleotide Variants

Different somatic mutational features were generated based on single nucleotide variants (SNVs). First of all, the total number of SNVs and the number of different single nucleotide substitutions were counted (7 types of substitutions: C>A, C>G, C>T at CpGs, C>T not at CpGs, T>C, T>A, and T>G). Further, the number of mutations in different trinucleotide contexts were counted by considering the 5’ and 3’ flanking nucleotide of the base substitution. This information was used to extract organ-specific mutational cancer signatures[138]. Next, mutation enrichments with respect to different genomic features were calculated.

Distribution of SNVs varied between individuals with respect to tumor tissue of origin and with respect to sequencing technology

As shown in Figure 2.1 the total number of SNVs and the total number of different types of SNVs differed between tissues and cancer cohorts. The median number of SNVs per sample was around 80 SNVs in TCGA, around 3,700 SNVs in PCAWG, and around 9,000 SNVs in Hartwig. The increased number of SNVs in Hartwig and PCAWG can be attributed to the differences between WGS and WES data. The differences in SNVs between the two WGS datasets PCAWG and Hartwig could be attributed to the different types of tumor samples. PCAWG mostly contains primary cancers[8], while the tumor samples in Hartwig are all metastatic[9]. In addition, many of the cancer types with high mutation load in PCAWG were not included in this study (e.g. colon, rectum, and lung) since they were also part of TCGA and we wanted to prevent an overlap between discovery and validation cohort for the subsequent association studies. Further, it can be seen how the median number of SNVs per individual was higher in tissues such as skin, colon, rectum, stomach, esophagus and lung, and lower in tissues such as thyroid and thymus in comparison to the other tissues. These observations matched results from previous studies[6,9].

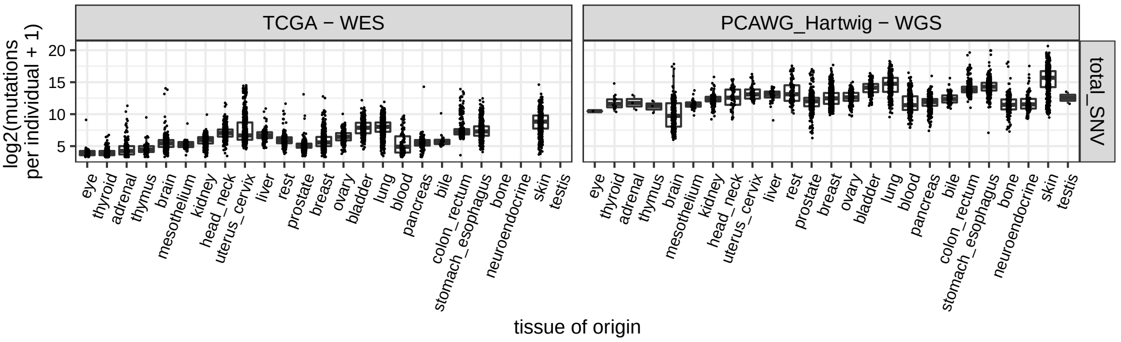


Figure 2.1: Distribution of SNVs varied across tissues and cancer cohorts in the nuclear genome.

The total number SNVs in the mitochondrial genome was calculated (Figure 2.2). In contrast to the distribution of SNVs across tissues in the nuclear genome, the number of SNVs in the mitochondrial genome was especially increased in cancers with the tissue of origin in colon, rectum, ovary, pancreas and blood. The average number of SNVs in the mitochondrial genome ($\sim$16,500 bp in length) was far lower than in the nuclear genome with less than 1 mutation per individual in Hartwig, around 1 mutation per individual in TCGA and around 2 mutations per individual in PCAWG. These differences in distributions of mutations across tissues were also in accordance with previous studies[157,159].

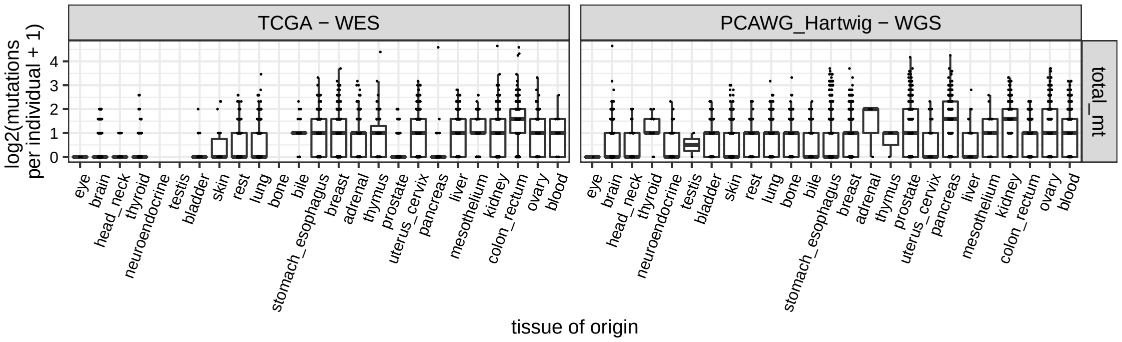


Figure 2.2: Distribution of SNVs varied across tissues in the mitochondrial genome.

Exposures of organ-specific signatures in whole exome sequenced samples recapitulated known aetiologies

For signature extraction, NMF-derived organ specific signatures were utilized[138] and signature exposures in each sample were estimated. In this way, exposures of signatures were only estimated for signatures which were also detected in the tissue of origin of the respective sample. Signature exposures were only assigned to samples when they reached a certain confidence threshold. Otherwise they were set to 0. Since this method was tested on WGS data[138], we investigated whether the exposure estimations for the TCGA -WES samples agreed with previous studies.

For TCGA -WES, the fraction of unassigned mutations per sample was increased with a median of 43 % in comparison to PCAWG_Hartwig -WGS, where the median fraction of unassigned mutations per sample was around 15 %. Unexpectedly, the fraction of unassigned mutations decreased with an increasing number of SNVs (Figure 2.3) with a Pearson correlation of R = -0.74 in TCGA -WES and a Pearson correlation of R = -0.46 in PCAWG_Hartwig -WGS.

Investigating the distribution of different signature mutations across tissues, known variations were captured as shown in Figure 2.4. Signature 1 (called Ref.Sig. 1 here) has been associated with the number of cell divisions[238] and was also found across all tissue of origins. This was also the case for Ref.Sig. 5. Other signatures were exclusively detected in a few tissues. For instance signature 7 mutations, which mark UV light exposure, were increased in skin cancers and sparsely detected in a few other tissues.

We further investigated whether signature exposures matched known causes.

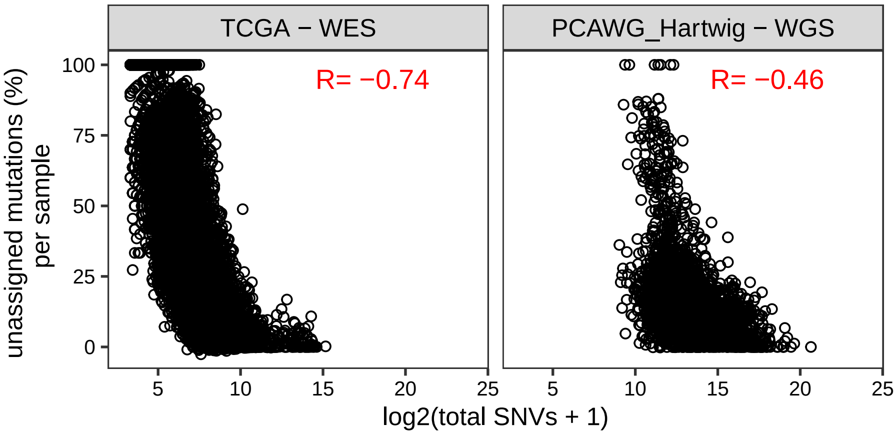


Figure 2.3: Fraction of unassigned mutations decreases with increasing number of total SNVs in a sample. Pearson correlation shown in red (top right).

First of all, we could show that samples which were assigned to have a high APOBEC activity (APOBEC high/not high assignment based on ratio r by counting specific SNVs as performed in ref[108]) also had an increased number of signature 2 and signature 13 mutations in the respective cancer types (Figure 2.5).

Similarly, samples with a somatic mutation in the exonuclease domain of *POLE* had an increased number of signature Ref.Sig. 10 mutations (Figure 2.6), which is the signature connected to POLE deficiency.

Also, samples which were reported to have microsatellite instability (MSI), which is caused by a deficiency in DNA mismatch repair, had a significantly increased number of either signature Ref.Sig. MMR1 or MMR2 mutations (Figure 2.7). Signature Ref.Sig. MMR1 mutations were increased in samples with MSI in endometrial, colon, rectal and stomach cancer, which are also the cancer types which have been previsously reported to have the highest fractions of MSI in the TCGA cohort[68]. Signature Ref.Sig MMR2 was not previously extracted from colon, rectal and stomach cancers, and was predominantly detected in liver, ovary, bone and soft tissue cancers[138]. While in liver cancers the number of Ref.Sig. MMR2 mutations were increased in MSI samples, this was not the case in ovary cancer.

Earlier studies have linked signature 11 to temozolomide (TMZ) treated cancers due to the similarities of the signature to the detected patterns after treating human cell lines with alkylating agents[6]. Further, this signature was associated

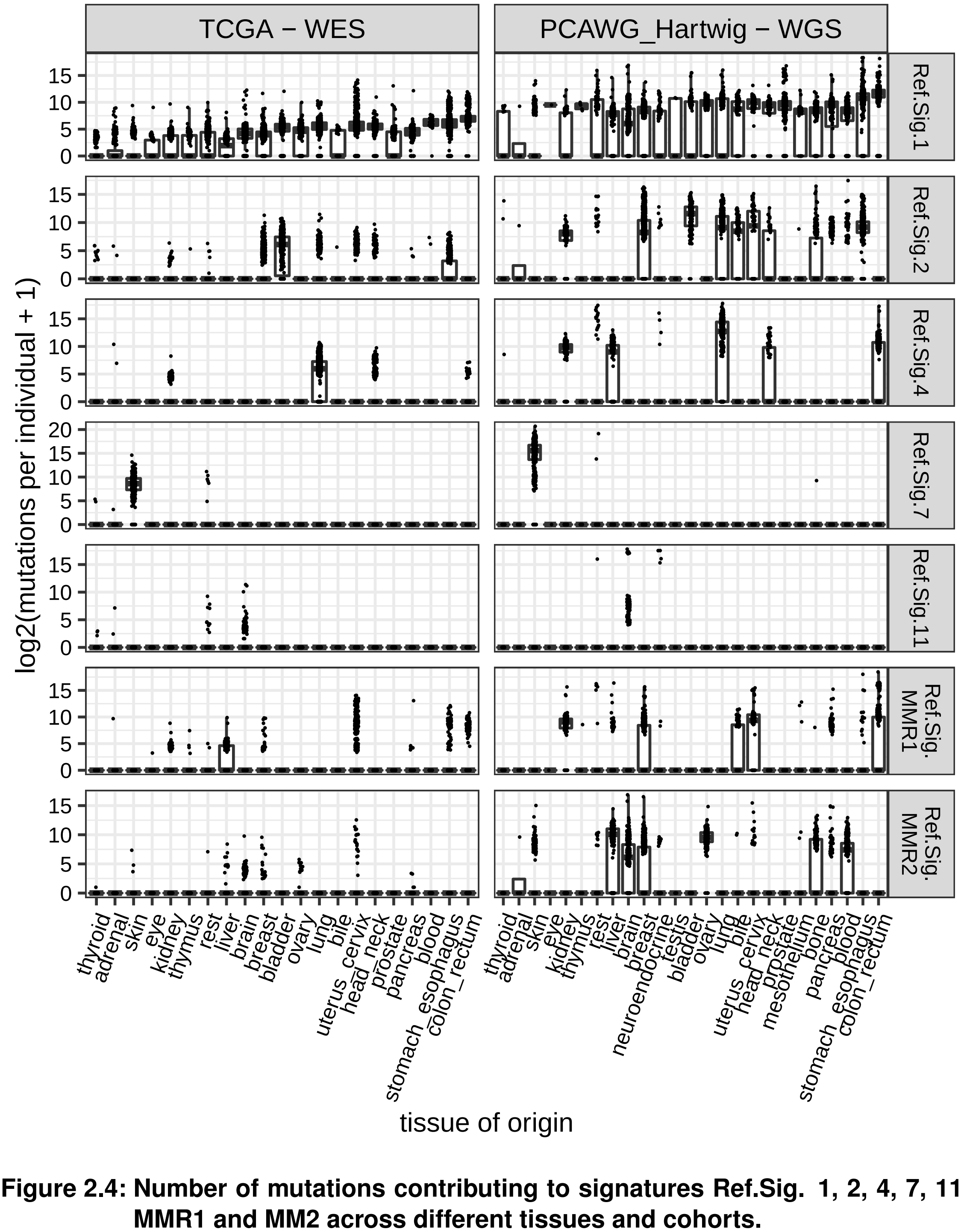


Figure 2.4: Number of mutations contributing to signatures Ref.Sig. 1, 2, 4, 7, 11, MMR1 and MM2 across different tissues and cohorts.

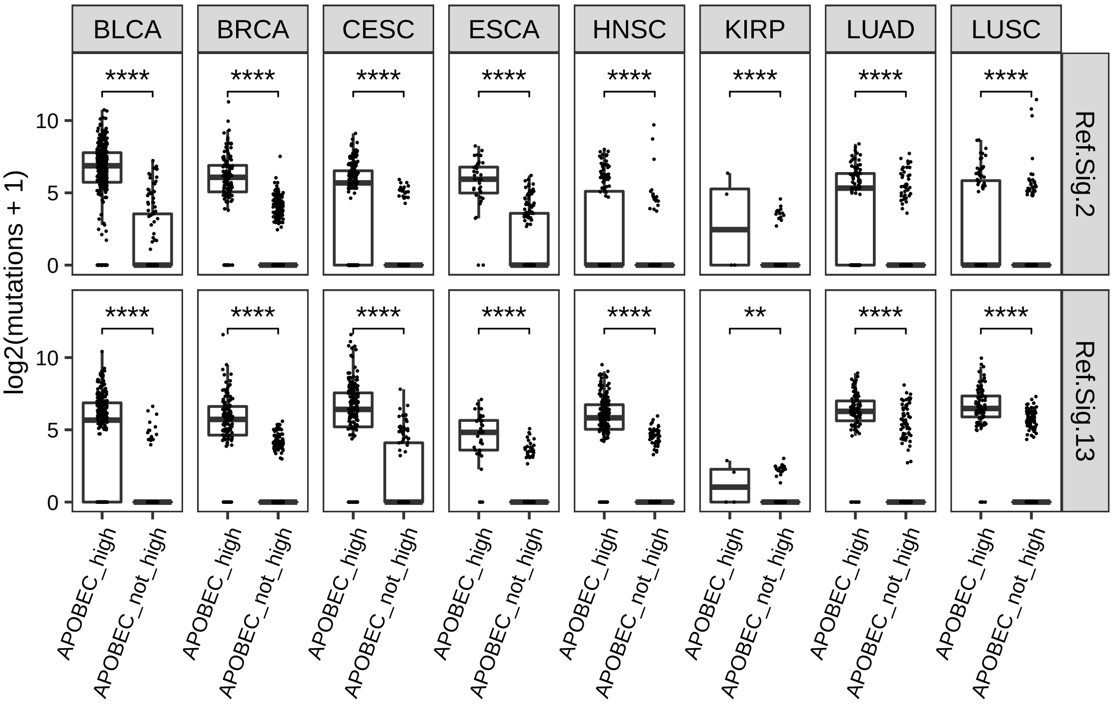


Figure 2.5: Signature Ref.Sig. 2 and 13 mutations were increased in samples with high APOBEC activity. APOBEC high/not high assignment based on ratio r by counting specific SNVs as performed in ref[108]. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

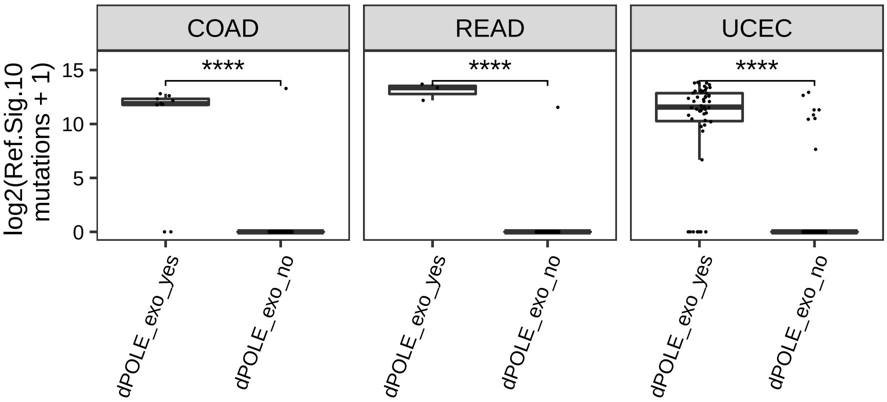


Figure 2.6: SNV-derived signature Ref.Sig. 10 mutations were increased in samples with somatic mutations in the exocnuclease domain of *POLE*. TCGA tumor abbreviations in Table 4.3. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

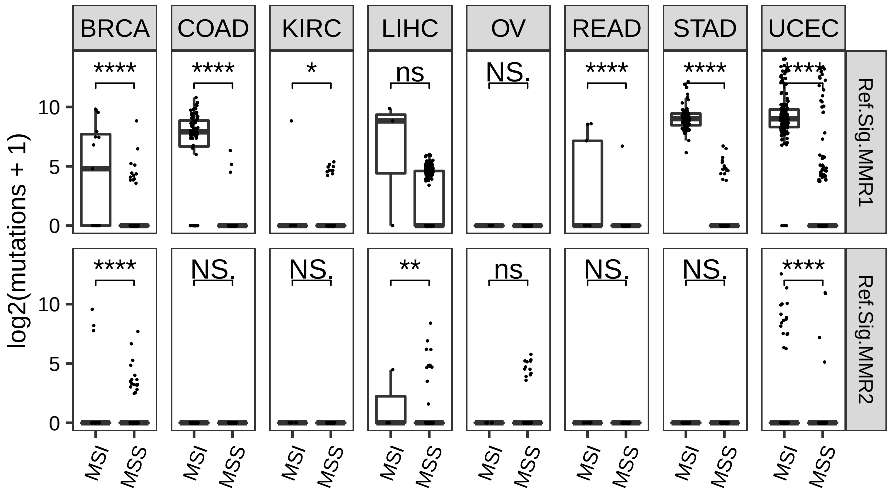


Figure 2.7: Enrichment of signature Ref.Sig. MMR1 and MMR2 mutations in microsatellite instable (MSI) samples across different cancer types. MSI was assigned via the MSI detection tool MANTIS[68]. TCGA tumor abbreviations in Table 4.3. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

with TMZ treated cancers in the presence of either MMR deficiency[199] or *MGMT* promoter methylation[27]. In addition, a highly similar signature was identified in cells treated with the DNA methylating agent 1,2-Dimethylhydrazine[116]. To further investigate this, we checked to which extent TMZ treated cancer had an increased number of signature Ref.Sig. 11 mutations in the presence of these factors. In neither TGCA-WES nor in Hartwig-WGS, TMZ treated cancer had a significantly increased number of signature 11 mutations in comparison to non-treated cancers. The number of mutations were also not increased in the presence of MMR deficiency (Figure 2.8) or *MGMT* promoter methylation (Figure 2.9).

All in all, the results showed that signature exposures, which were estimated based on the organ-derived signatures[138], recapitulated the known tissue distributions (e.g. UV-derived signatures) and known causes of some of the signatures (e.g. *POLE*, MSI, and APOBEC) also in the mutational data from WES samples. While we could not associate signature 11 exposures with TMZ treated cells and/or in combination with MMR and/or *MGMT* inactivations, previous studies also reported conflicting results[27, 116].

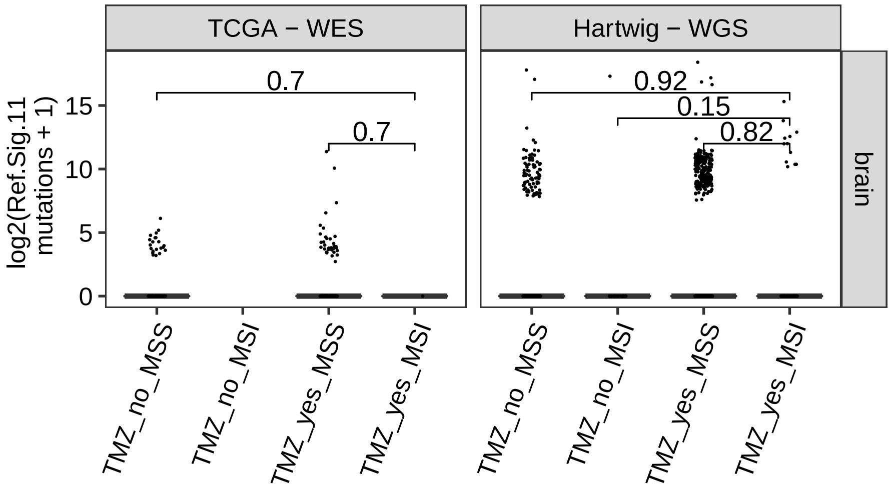


Figure 2.8: Signature Ref.Sig. 11 mutations were not increased in samples treated with temozolomide (TMZ) or dacarbazine irrespective of the occurrence of MSI. MSI assignment from MANTIS for TCGA[68] and from Hartwig flagship paper[9] for Hartwig. P-values from two-tailed Mann-Whitney U test.

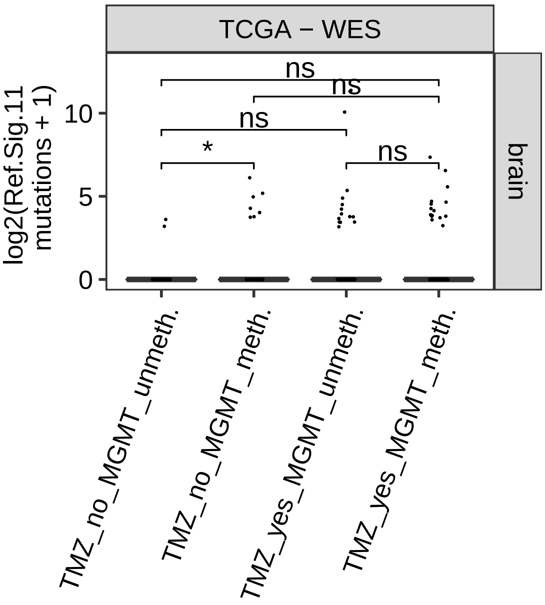


Figure 2.9: Signature Ref.Sig. 11 mutations were not increased in samples treated with TMZ irrespective of the somatic promoter methylation status of the O6-methylguanine-DNA methyltransferase (*MGMT*) gene. Promoter methylation status derived from ref[128]. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

Estimated transcriptional strand biases in whole exome sequencing data matched previous studies

We estimated transcriptional stand biases by dividing the number of mutations occurring on the untranscribed strand by the number of mutations occurring on the transcribed strand, while stratifying for the six different possible base substitutions. We investigated whether the estimations matched previous studies which were based on WGS data only[24]. Lung cancer types showed a strong transcriptional strand bias for C>A mutations, skin cancer types showed a strong bias for C>T mutations and liver cancer types showed a strong bias for T>C mutations (Figure 2.1). Thus, we were able to estimate this somatic feature from WES data.

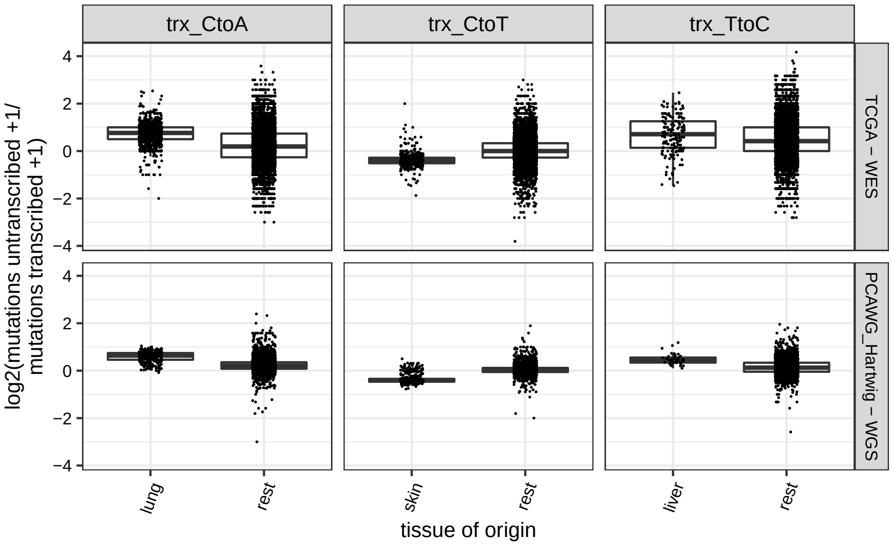


Figure 2.10: Strong transcriptional strand bias in lung, liver and skin cancers independent of the sequencing technology. trx: transcriptional strand bias.

Extraction of a somatic feature based on the reported replicative strand bias[149] associated with a deficiency in DNA mismatch repair

Similarly, we estimated the replicative strand bias by dividing the number of mutations occurring on the leading strand by the number of mutations occurring on the lagging strand. We focussed for this feature exclusively on T>C, T>G, G>A and C>A base substitutions, since those have been reported to have a strong replicative strand bias in connection with a deficiency MMR[149]. However, it should also be noted that APOBEC activity, a deficiency in DNA polymerase δ, and a deficiency in DNA polymerase ε can result in a replicative strand bias[108, 149]. As described in the methods, only a small fraction of the genome was covered by this feature (Table 4.1), in particular in WES data. Thus, we did not stratify for the different base substitutions, but combined them. The feature captured the replicative strand bias differences between MSI vs. MSS samples in several tissues (Figure 2.11). The differences were strongly visible in colon, rectal and endometrial cancers, which also matched the previous study[149]. Additionally, it can also be seen that this bias was also captured in other, previously not reported cancer types such as breast, prostate, kidney, stomach and esophagus cancers, even though with less significance, which can also be attributed to the lower occurrence of MSI in these tissues (e.g. kidney).

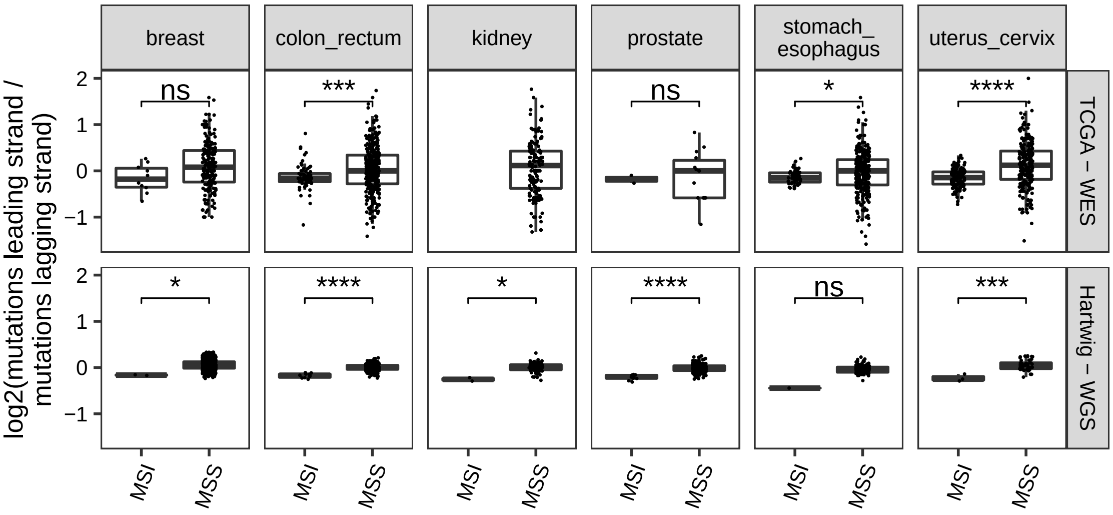


Figure 2.11: Replicative strand bias differed between MSI and MSS samples across tissues. Number of T>C, T>G, G>A and C>A mutations on the leading strand were divided by the number of those mutations occurring on the lagging strand. MSI assignment from MANTIS for TCGA[68] and from Hartwig flagship paper[9] for Hartwig. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

Mutation enrichment calculations with regards to replication timing, histone mark H3K36me3, DNase I hypersensitive sites, and RNA-seq expression

The distribution of mutations along the genome can change for instance due to a deficiency in DNA mismatch repair[21,144]. For this reason, we generated somatic features estimating the mutation enrichment with regards to different genomic properties. A feature calculating the enrichment of mutations in early replicating regions in comparison to late replicating regions, a feature measuring the enrichment of mutations in regions with a high amount of histone mark H3K36me3 in comparison to regions with no amount of this histone mark, a feature measuring the enrichment of mutations in regions with a high amount of DNase I hypersensitive site in comparison to regions with no DNase I sites and a feature corresponding to the enrichment of mutations in regions which are highly expressed in comparison to regions which are not expressed, was generated. We calculated these features via negative binomial regression and also controlled for the different base substitution types (Section 4.1.13).

To check whether the estimation of these features worked, we looked again at the differences between MSI and MSS samples as shown in Figure 2.12. It can be seen how there were significant differences between MSI and MSS samples for replication timing coefficients and histone mark H3K36me3 coefficients, especially in colon, rectal and endometrial cancers. This matched our expectations and previous studies[15,21]. Significant differences between MSI and MSS samples for the DNase and expression feature were not expected, but were detected in a few cancer types (e.g. colon and rectal), even though effect sizes were lower in comparison.

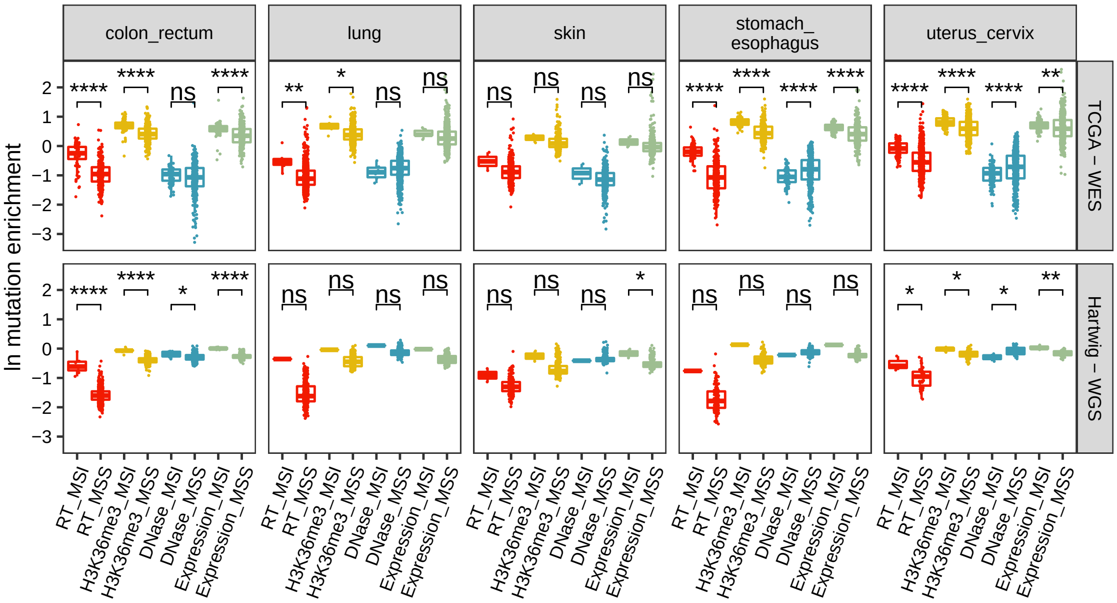


Figure 2.12: Mutation enrichment estimations with regards to replication timing, histone mark H3K36me3, DNase I hypersensitive sites and RNA-seq expression across selected tissues. Samples were further split into MSI vs. MSS samples. MSI assignment from MANTIS for TCGA[68] and from Hartwig flagship paper[9] for Hartwig. RT: replication timing. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test. DNase: DNase I hypersensitive sites, Expression: RNA-seq expression, H3K36me3: histone mark H3K36me3, RT: replication timing.

To further confirm that the coefficients from the regression correctly measured mutation enrichments and that they were not influenced by the total number of mutations in a sample, we also investigated the differences between MSI and MSS samples in the presence of mutations in the exonuclease domain of *POLE*, which lead to a hypermutator phenotype[85] (Figure 2.13). Except for one case in endometrial cancer for the DNase feature, there were no significant differences in the coefficients due to the hypermutator phenotype caused by mutations in *POLE*.

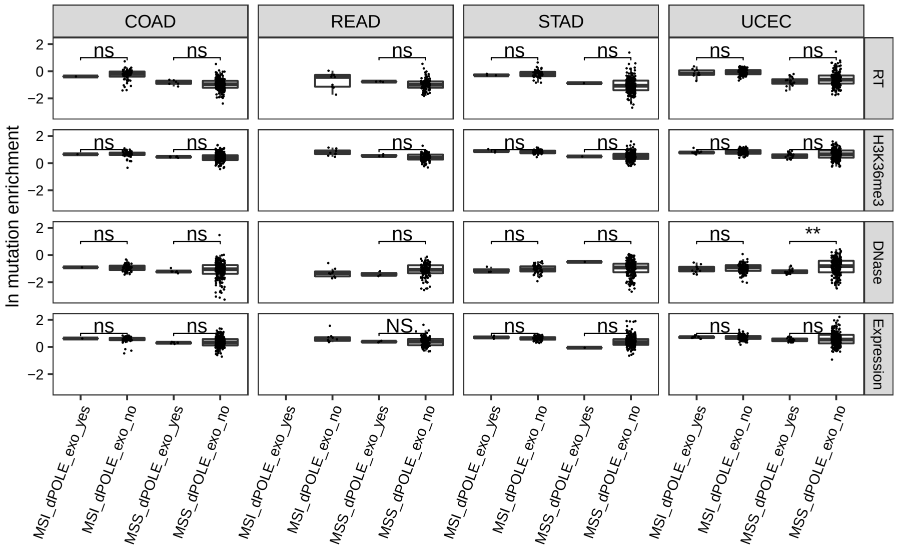


Figure 2.13: Mutation enrichment differences between MSI and MSS sampes were not affected by the presence of a hypermutator phenotype. Samples were grouped into MSI and MSS samples[68] and further stratified by the presence of a somatic mutation in the exonuclease domain of *POLE*. TCGA data only shown here. TCGA tumor abbreviations in Table 4.3. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05.$, **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

CTCF somatic feature only showed moderate differences between MSI and MSS samples

CTCF/cohesin binding sites have mutation peaks in many cancer types[151,152] and most importantly are largely unaffected by a deficiency in DNA mismatch repair[151]. Inspired by this observation we generated a somatic feature dividing the number of mutations occurring in CTCF/cohesin binding sites by the number of mutations occurring in the flanking sites. It was anticipated that in samples with a deficiency in DNA mismatch repair the number of mutations in the flanking regions would be increased while the number of mutations in the binding site would be unaffected. In TCGA, in stomach and esophagus cancers it could be seen how the ratio was decreased in MSI samples in comparison to MSS samples (Figure 2.14). In many other cancer types, the trend was visible but not strong enough (e.g. prostate cancer in Hartwig -WGS) or lacking in sample size (e.g. kidney in TCGA -WES) to be significant. In some other cancer types the effect was not detected, such as in breast cancer and uterus and cervix tissue derived cancers in Hartwig -WGS.

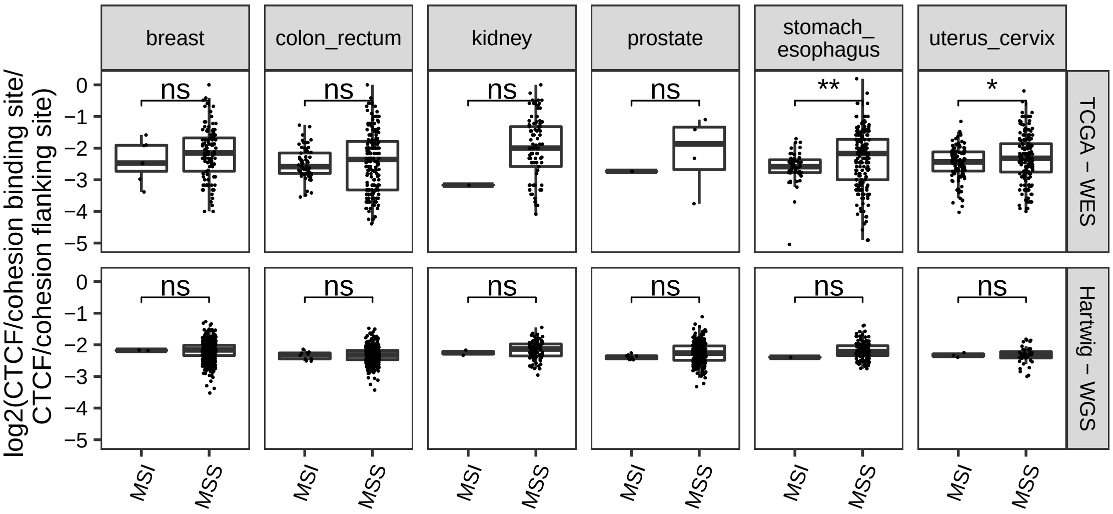


Figure 2.14: CTCF somatic feature only showed moderate differences between MSI vs. MSS samples. Samples were split into MSI vs. MSS samples. MSI was assigned via the MSI detection tool MANTIS for the TCGA cohort. MSI assignments for Hartwig were extracted from the flagship paper[9]. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

X-chromosomal hypermutation predominantly in brain and blood cancers in females

We also generated a feature to measure X-chromosomal hypermutation, which has been predominantly identified in brain cancers, blood derived cancers[155] and head and neck cancers[156] in females. The hypermutation phenotype on the X chromosome has been largely attributed to an increasing number of mutations on the inactive X chromosome, explaining why it has been only identified in females[155]. To measure this feature, we divided the number of mutations occurring on the X chromosome by the average number of mutations occurring on the autosomes. As shown in Figure 2.15, the mutational load was mostly increased in the respective cancer types in females in comparison to males even after correcting for the missing inactive X chromosome in males by multiplying the number of mutations times 2. Still, in pediatric brain cancers in the PCAWG cohort we could not recapitulate the feature.

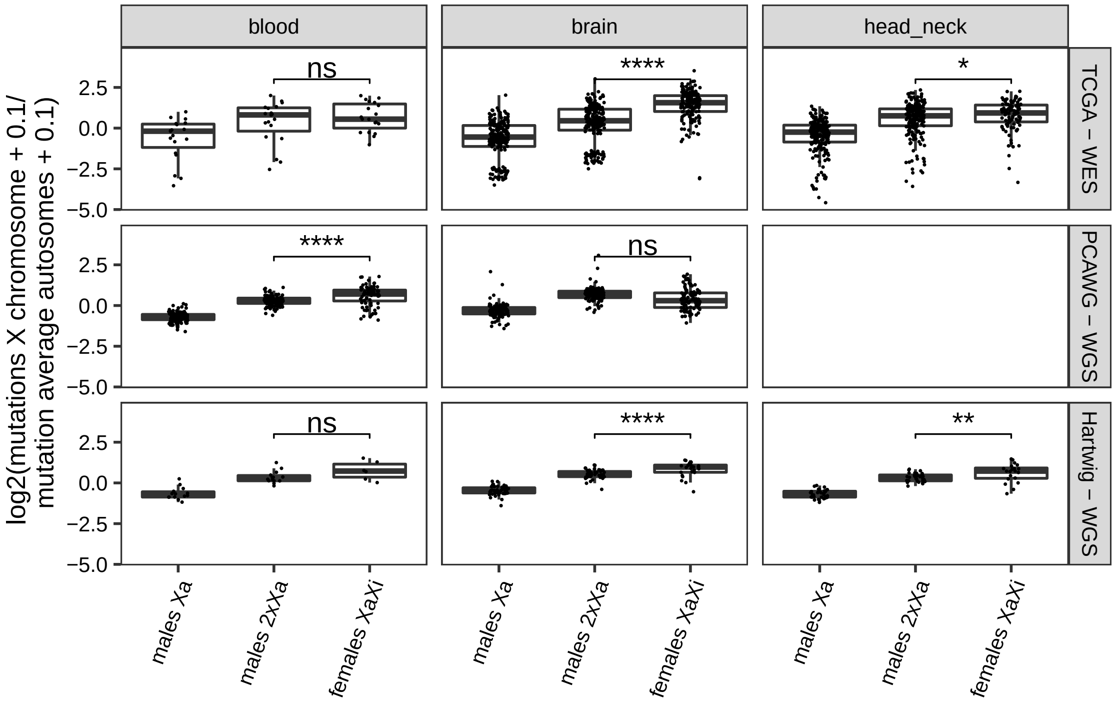


Figure 2.15: Mutation load on the X chromosome predominantly increased in brain cancers, blood cancers, and head and neck cancers in females compared to males. Feature was calculated in males, who only have one active X chromosome (Xa) and females, who have an inactive and an active X chromosome (XaXi). To control for the lack of a second X chromosome in males, we also calculated the feature after multiplying the number of mutations occurring on the X chromosome in males times 2 (2xXa). Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

2.1.2 Features based on Double Nucleotide Variants

Double base substitution signatures (DBS) were extracted as described in Section 4.1.14 and de-novo signatures were fitted to the established COSMIC signatures[4], which were then used to estimate exposures in each sample. Four DNV-derived signatures, which were found in both cohorts TCGA-WES and PCAWG_Hartwig-WGS were kept for further analysis. DBS1 has been connected to UV-directed mutagenesis and the highest exposures were also detected in our analysis in skin cancer (Figure 2.16). DBS2 has been attributed to tobacco smoking and also here it can be seen that the highest exposures were measured in lung cancers. DBS4 has been suggested to be generated by an endogenous process occurring in normal human cells, but has not been validated[14]. As it can be seen in Figure 2.16 this signature was not significantly enriched in a specific tissue type, but rather detected across all tissues. DBS9 was in particular enriched in adrenal and kidney cancers in PCAWG_Hartwig -WGS, but was also observed in many other tissues. At the moment there is no existing aetiology for this signature. In general, it can be seen how especially in TCGA -WES, most samples did not contain any DBS signature mutation, which was expected since DNVs are much rarer in comparison to SNVs[14].

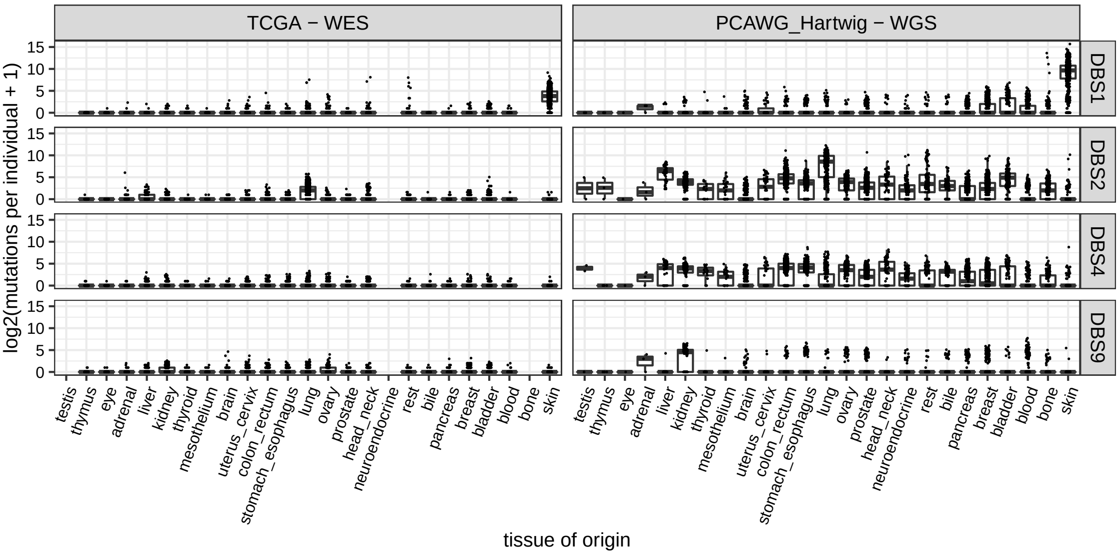


Figure 2.16: Distribution of double substitution signatures across tissues: DBS1, DBS2, DBS4 and DBS9.

2.1.3 Features based on Insertions and Deletions

Similarly as for the DNVs, we also generated indel and deletion based signatures. Indel signature ID2, ID3, ID4 and ID8 were extracted in both cohorts. ID2 has been connected to a deficiency in DNA mismatch repair, ID3 has been connected to tobacco smoking, the aetiology for ID4 is unknown, and ID8 has been primarily connected to repair of DNA double strand breaks by non-homologous end joining[14]. The distribution of mutations can be seen in Figure 2.17.

Furthermore, based on a previous study, which reported informative features to predict homolgous recombination-directed repair[95], we generated a feature counting the number of deletions bigger than or equal to 10 bp, a feature counting

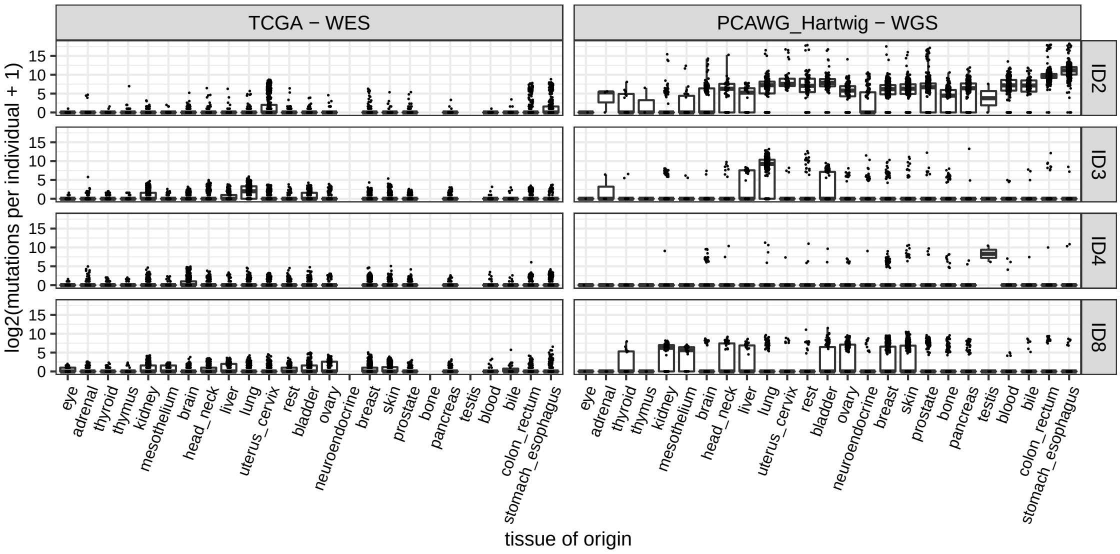


Figure 2.17: Distribution of insertion and deletion signatures across tissues: ID2, ID3, ID4 and ID8.

the number of deletions in microhomology-flanked regions with a size of 1 bp and a feature counting the number deletions in microhomology flanked regions with a size bigger than 1 bp.

Next, more features potentially covering a deficiency in DNA mismatch repair were generated. The number of indels were counted in total as well as of different sizes (1 bp or 2 to 5 bp) stratified by different genomic regions (microsatellite regions and non-microsatellite regions). In addition, we also counted the number of indels with a size of 6 to 10 bp. To check whether these features could be useful to detect a deficiency in MMR, the differences in MSI vs. MSS samples across tissues were checked (Figure 2.18). It can be seen that even counting the total number of indels showed stronger differences in MSI vs. MSS samples in comparison to the same analysis performed with SNV-derived somatic features (Figures 2.7, 2.11 and 2.14). Furthermore, signature ID2 and the feature counting the number of indels of 1 bp in size at microsatellite regions, showed consistently the strongest effect across tissues. In some tissues such as in prostate cancer in TCGA -WES or in brain cancer in Hartwig -WGS, the effect was less strong when looking at the number of indels of a size of 2 to 5 bp in non microsatellite regions in comparison to indels in microsatellite regions of the same size. This was also expected since the MSI assignment was based on a MSI detection tool, which counted mutations in microsatellite regions[68].

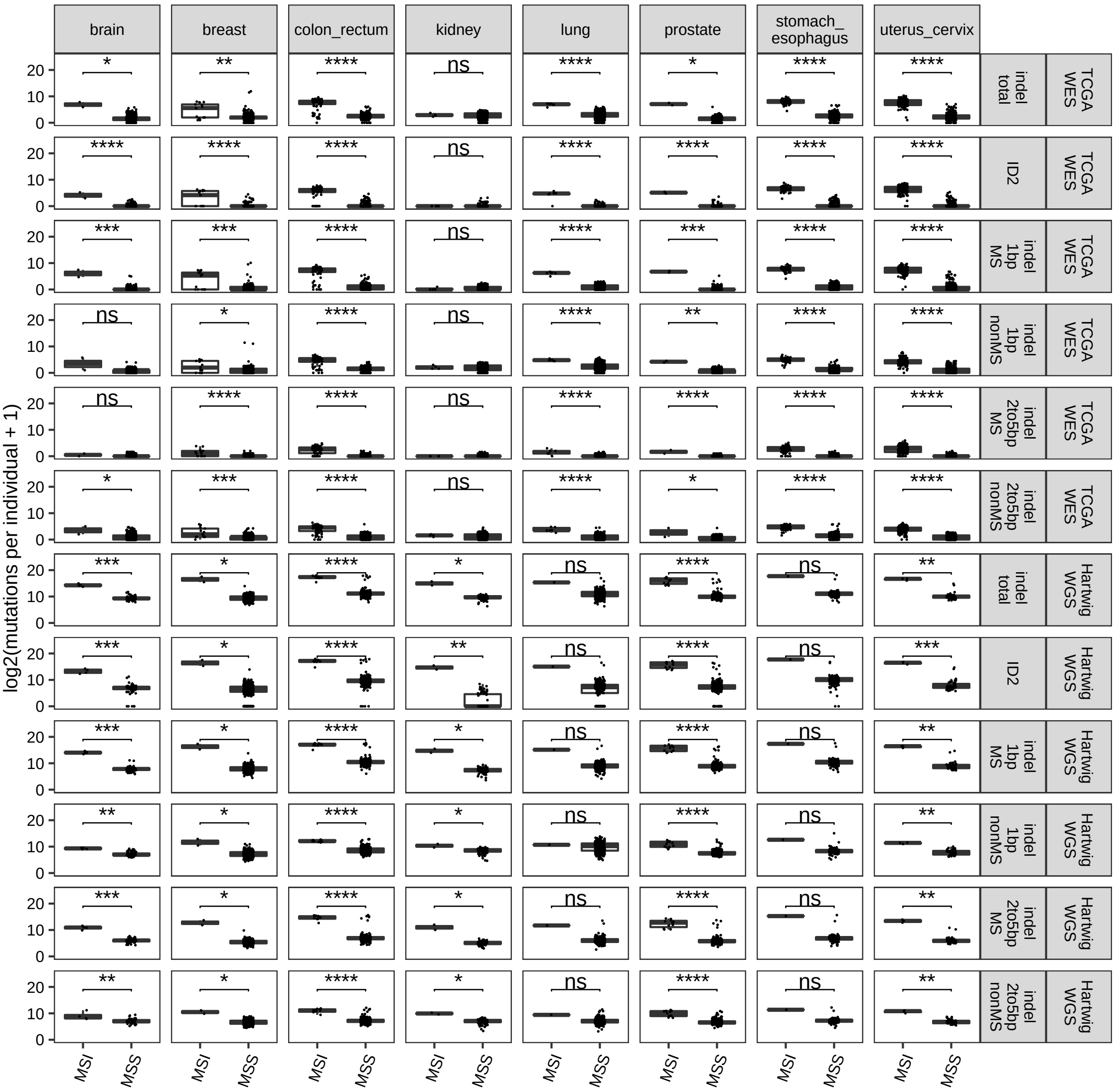


Figure 2.18: A deficiency in DNA mismatch repair can be measured by counting small indels in repetitive elements. Samples were split into MSI vs. MSS samples. MSI assignment from MANTIS for TCGA[68] and from Hartwig flagship paper[9] for Hartwig. Total number of indels, number of signature ID2 mutations, number of indels of different sizes (1 bp or 2 to 5 bp) stratified by genomic regions (microsatellite (MS) regions and non microsatellite regions (nonMS)) are shown. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

2.1.4 Features based on Copy Number Variants

Features based on copy number alterations were split into amplification and deletion events stratified by different sizes as it has been suggested in other studies[18, 94]. Amplification events were split into four groups (1 to 10 kb, 10 to 100 kb, 100 kb to 1 mb and > 1 mb) and deletion events were split into three groups (1 to 10 kb, 10 to 100 kb and > 100 kb). In accordance with previous studies[18], the highest amount of amplifications were detected in esophagus and ovary cancers 2.19.

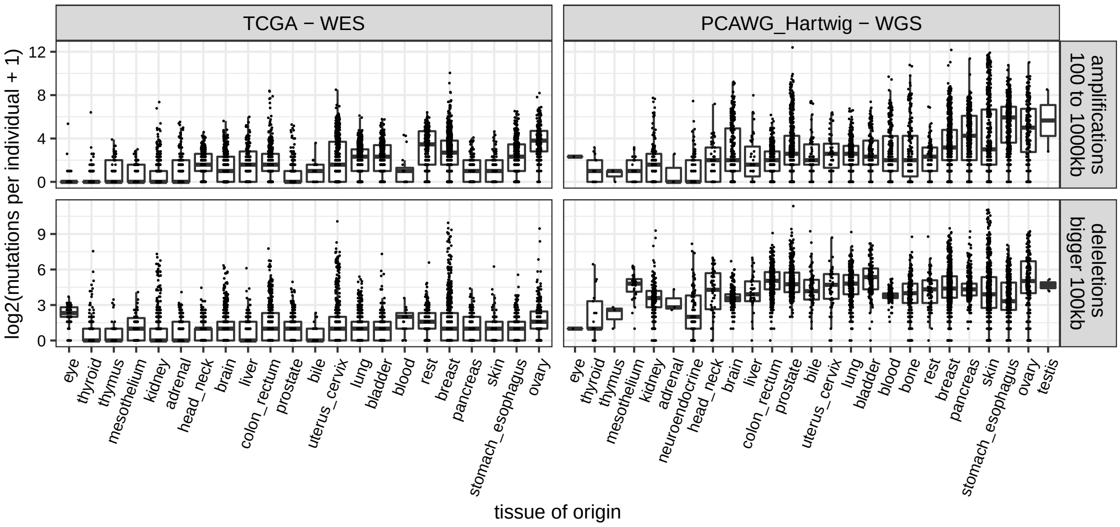


Figure 2.19: Distribution of amplification and deletion events across tissues. Showing the number of amplifications with a size of 100 kb to 1 mb and the number deletions with a size > 100 kb per individual.

2.1.5 Overview of all Somatic Features

All in all, somatic features were extracted from different mutation classes (SNVs, DNVs, indels, and CNVs) utilizing different cancer cohorts (TCGA -WES cohort in Figure 2.20 and PCAWG_Hartwig -WGS cohort in Figure 2.21).

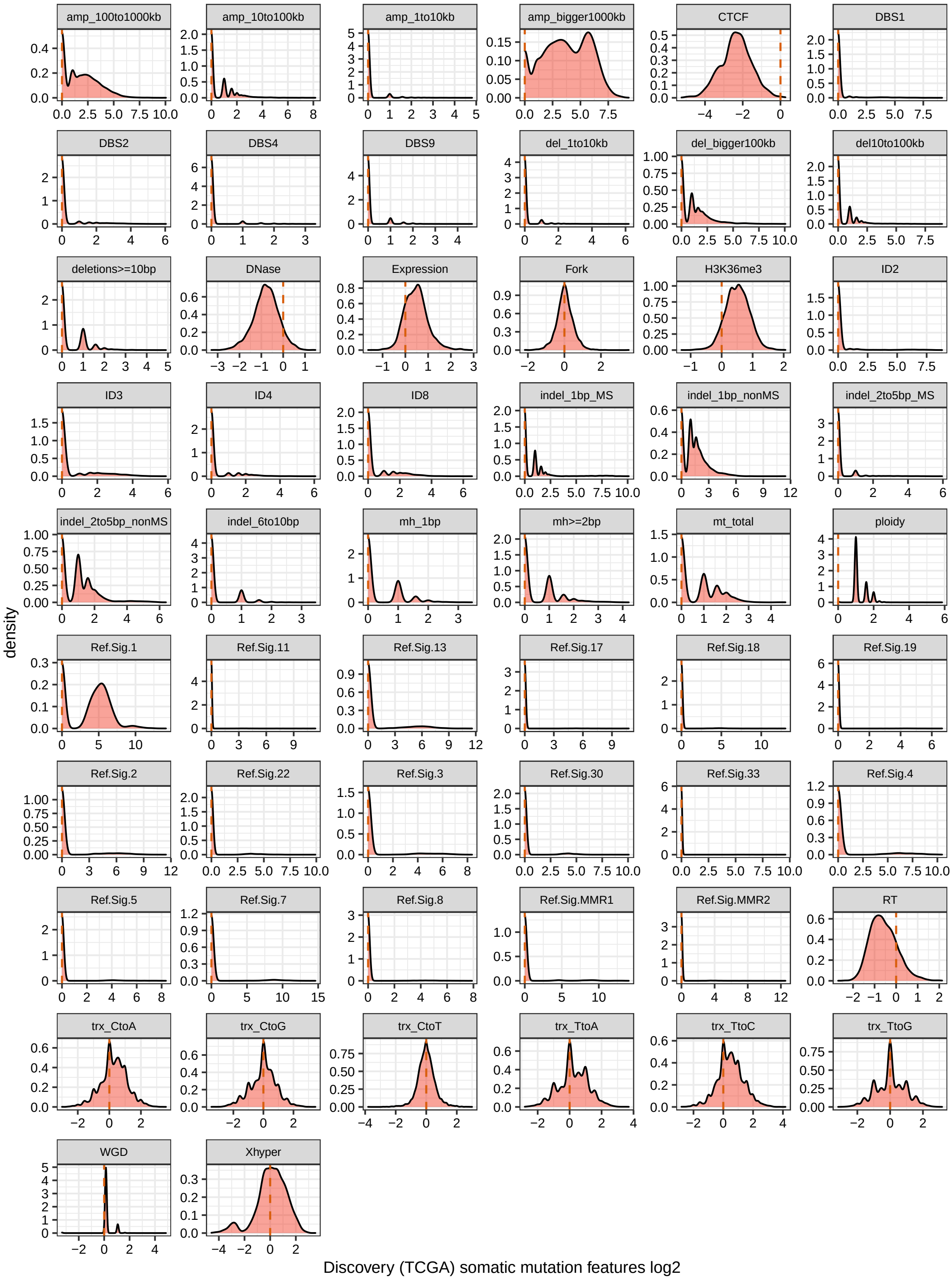


Figure 2.20: Distribution of 56 somatic features in TCGA-WES. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation.

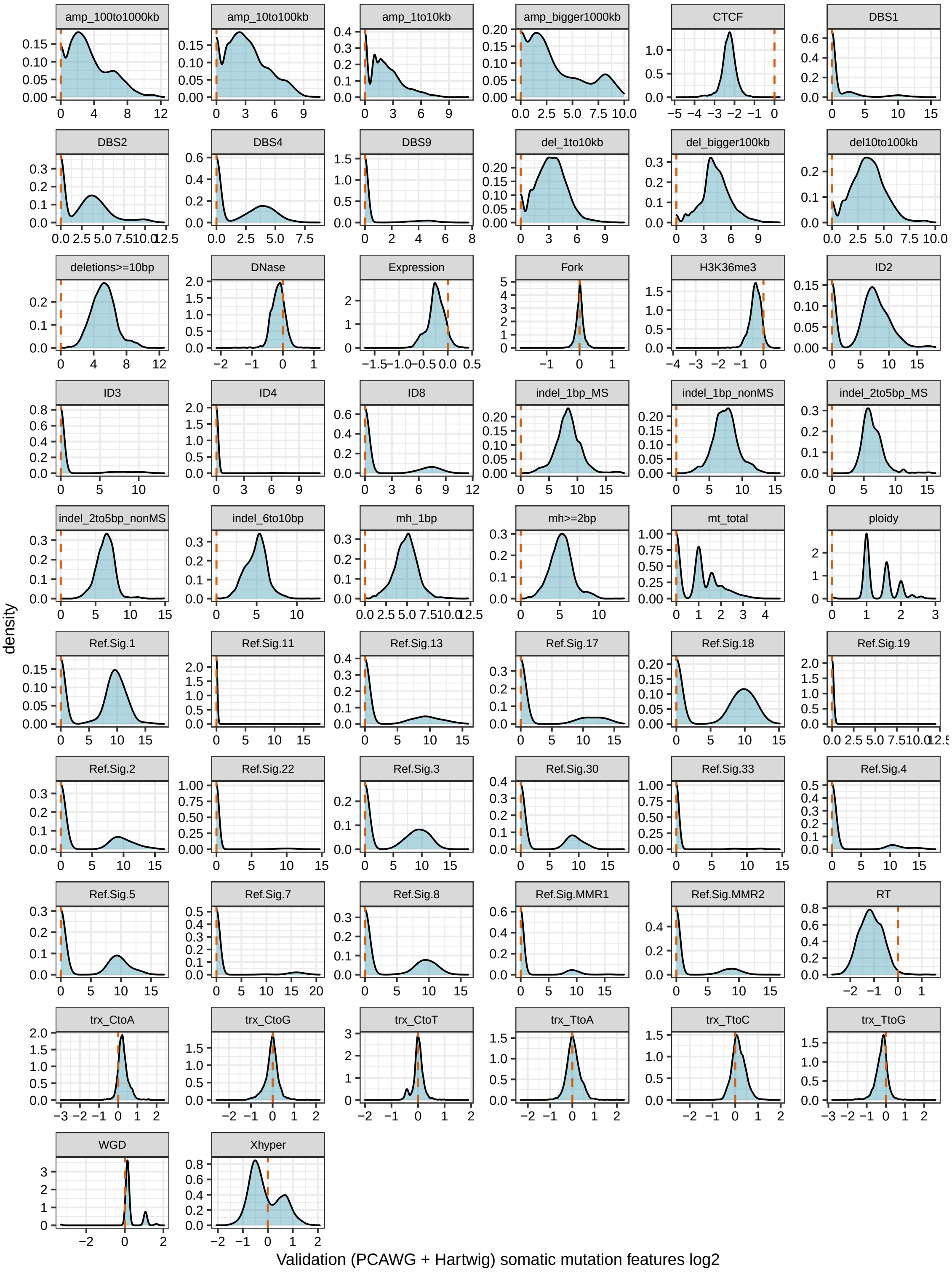


Figure 2.21: Distribution of 56 somatic features in PCAWG_Hartwig-WGS. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation.

2.1.6 Comparison between Whole-Exome and Whole-Genome extracted Somatic Features

As described earlier, several somatic features such as the organ-specific SNV-derived signatures[138], transcriptional strand bias[24], relative mutation rates with respect to replication timing, expression and histone mark H3K36me3[15, 21, 144] and mutation peaks at CTCF/cohesin binding sites[151] were only extracted from WGS data so far. We showed that most features recapitulated known effects also in WES data by checking reported effects (APOBEC signatures in Figure 2.5, signature Ref.Sig. 10 in Figure 2.6, transcriptional strand bias in Figure 2.10, relative mutations rates in Figure 2.12, and CTCF/cohesin feature in Figure 2.14). Only the effect of TMZ treatment on Ref.Sig.11 was not replicated in WES data (Figure 2.8 and 2.9), but was also not replicated in WGS data (Figure 2.8). Thus, in this case the missing effect could not be explained by the sequencing technology.

Still, it should be noted that WES data usually covers around $\sim 2\,\%$ of the genome. Thus, we wondered how the somatic features change when extracting them from WES data compared to WGS data. For this purpose, we extracted the same somatic features from WGS samples from TCGA, which were part of PCAWG (Section 4.1.3) and compared them to the somatic features which were extracted from TCGA -WES from the same individuals. In total, somatic features were estimated for tumor genomes from around 530 individuals for which somatic calls were called via WES and WGS.

As shown in Figure 2.22, Pearson correlation was higher than 0.5 for 31 out of 65 somatic features and higher than 0.8 for 11 out of 65 somatic features. There was a significant correlation ($p < 0.05$ after correcting for multiple testing via Bonferroni) between WGS and WES for 59 out of 65 somatic features. No significant correlations were observed for the following six somatic features: del_1to10kb, del_bigger100kb, amp_1to10kb, DBS4, replicative strand bias (Fork), and CTCF/cohesin binding sites (CTCF). Especially, the low correlation between WES and WGS for the somatic features capturing the replicative strand bias (Fork) and CTCF/cohesin binding sites (CTCF) was not surprising, since these features could only be measured at few loci (Table 4.1).

Taken together, while the correlation between WES and WGS extracted components was low ($R < 0.5$) for the majority of the features, most features still had significant correlations between each other indicating that the general trends were still captured in WES data extracted components. This is supported by the fact that global genomic events such as APOBEC activity or a deficiency in DNA mismatch repair (leading to MSI) were captured in WES-extracted somatic features.

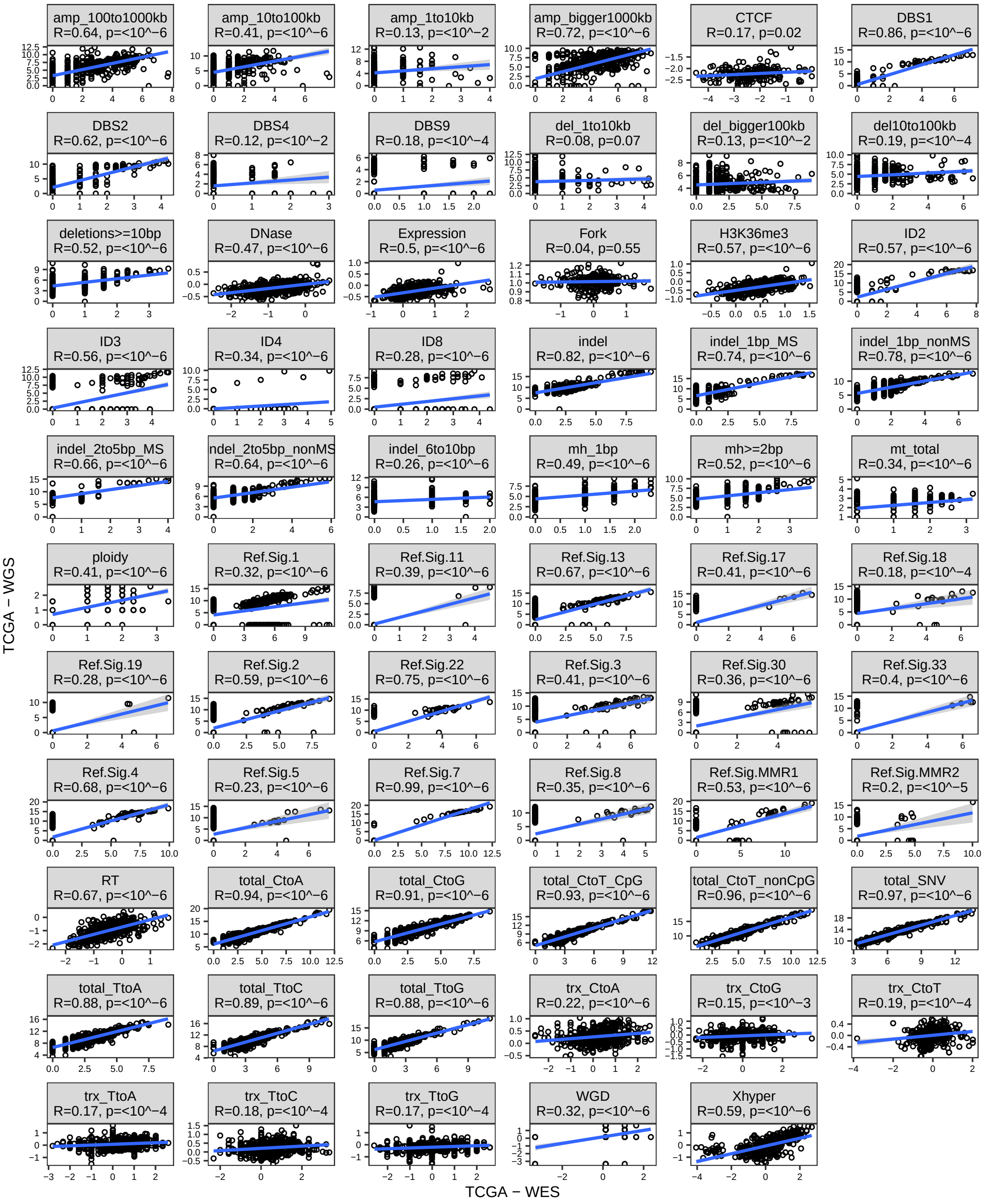


Figure 2.22: Comparison between WES and WGS extracted somatic features in TCGA. Somatic features were extracted from around 530 individuals from TCGA for which WES as well as WGS data was available. Somatic feature measured in TCGA-WES on the x-axis and in TCGA-WGS on the y-axis. One plot for each somatic features with Pearson correlation and associated p-value. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation.

2.1.7 Application of Dimensionality Reduction Techniques

After generating the different somatic features, we tested different high-dimensionality reduction techniques to embed the somatic features into a lower dimensional space. We aimed to remove the redundancy in the extracted somatic features, extract components which better reflect the underlying causal mechanisms, and to improve the statistical power to detect genetic associations by reducing the multiple testing burden. For this purpose, we utilized 56 somatic features as inputs (Sections 4.1.18-4.1.20) as shown in Table 2.1. We tested three different techniques: principal component analysis (Section 4.1.18), independent component analysis (Section 4.1.19) and a variational autoencoder neural network (Section 4.1.20).

Table 2.1: Input somatic features for component extraction. 56 different somatic mutation features were estimated in each cancer genome, covering different types of mutations.

Mutation Class	Features	#	Feature Name
SNV	Organ-specific mutation signatures based on trinucleotide context	17	Ref.Sig.X
	Transcriptive strand asymmetry	6	Transcriptional strand bias XtoX
	Replicative strand asymmetry	6	Replicative strand bias
	Relative mutation rates with respect to different genomic regions	5	Replication timing, H3K36me3, DNase, CTCF, Expression
	X-chromosomal hypermutation	1	Chromosome X hypermutation
	Mutation count on mitochondrial genome	1	Mitochondrial genome
DNV	Mutational Signatures	4	DBS1, DBS2, DBS4, DBS9
Indels	Mutational Signatures	4	ID2, ID3, ID4, ID8
	Deletions ≧ 10 bp	1	Deletions ≧ 10 bp
	Deletions of different lengths at microhomology flanking sites	2	Mh_1bp, Mh≧2bp
	Indels in/outside microsatellite regions of different lengths	4	Indels_X_MS, Indels_X_nonMS
CNV	Amplifications of different lengths	4	Amp_X
	Deletions of different lengths	4	Del_X
	Ploidy/Whole genome duplications	2	Ploidy, WGS

2.1.8 Extraction of Principal Components

The first 17 components from the PCA explained more than 1.78 % of the variance (expectation 1/56 of the variance explained; Figure 2.23). Further, we looked into the first principal components and into the features which had the highest contribution and correlation to a respective component (Figure 2.24). Several indel and deletion features contributed highly to principal component 1 such as indels at microhomology flanking regions and indels of different sizes. While the short indels are characteristic for a deficiency in DNA mismatch repair[6], deletions at microhomology flanking regions are characteristic for a deficiency in homologous recombination directed repair[5]. Thus, this principal component did not capture a specific mechanism but rather a high mutation load in general. For the other principal components we observed that two different mechanisms (or more) were often captured in one principal component with opposite signs. For instance principal component 2 anticorrelated with a deficiency in DNA mismatch repair (ID2, small indels in microsatellite regions, Ref.Sig. MMR1) and correlated with amplification events. Principal component 7 correlated with APOBEC activity (Ref.Sig. 2 and 13) and anticorrelated with a mix of different signatures (Ref.Sig. 8, 19 and 3). Based on the current understanding of cancer signatures, the principal components grouped different mechanisms into one principal component and did not separate them into individual components.

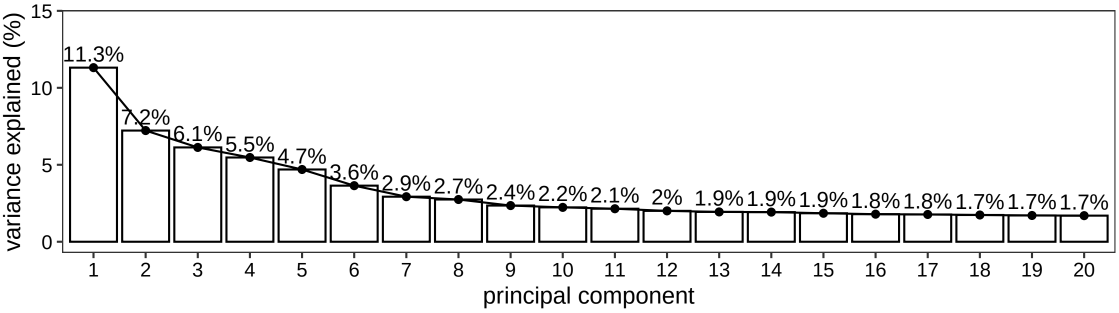


Figure 2.23: Dimensionality reduction via principal component analysis. Showing the variance explained in % for the first 20 principal components.

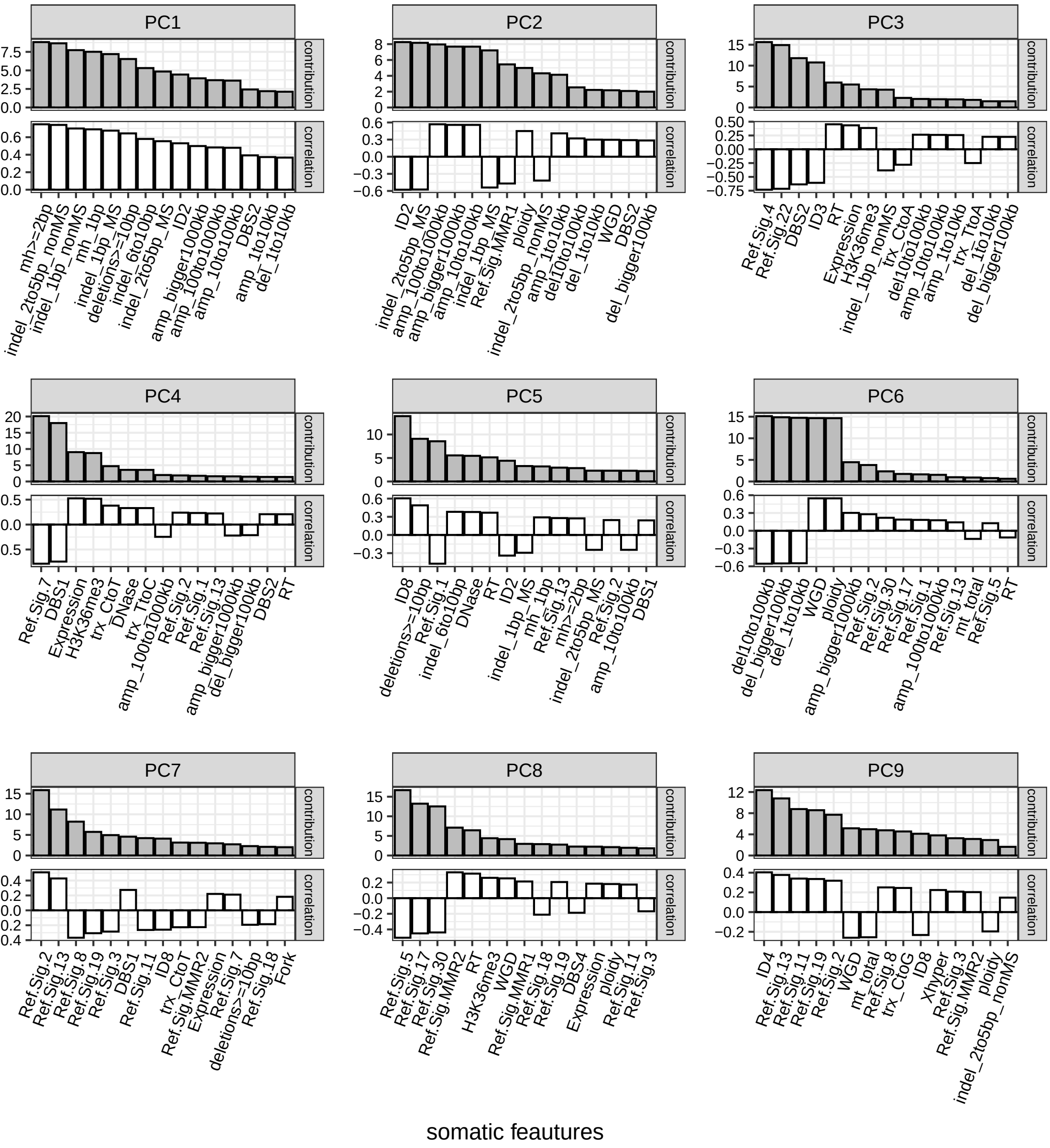


Figure 2.24: Overview of strongest contributing features to the first nine principal components. Showing the Pearson correlation (white bars) and contribution (in %) (grey bars) of the 15 strongest somatic features to the respective principal components. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation.

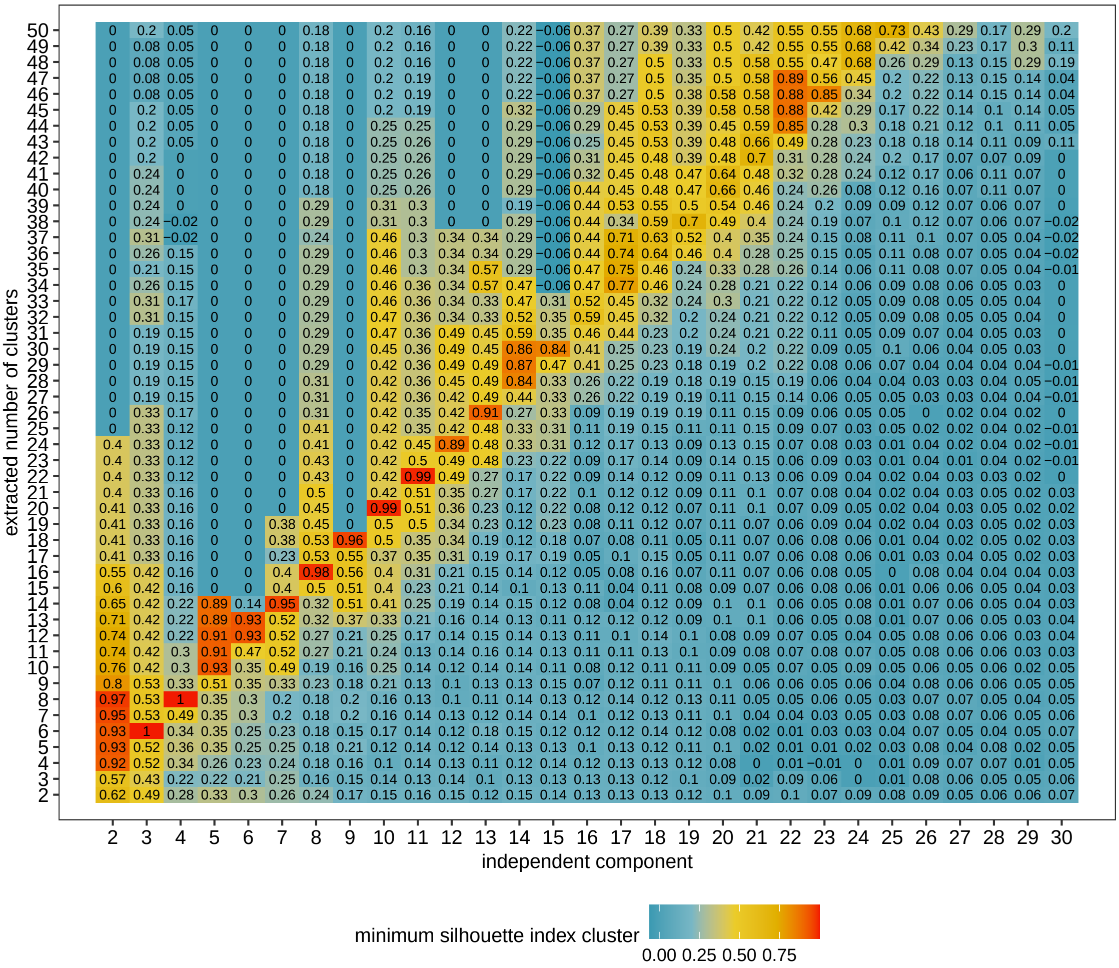
2.1.9 Extraction of Independent Components

Next, since PCA did not lead to the goal of having individual sources of mutations grouped into separate components, we tested ICA. For the ICA it is required to set the number of components to be extracted *a priori*. Since the optimal number was unknown, we ran the ICA at varying numbers of components (2 to 30). Each component extraction was run 200 times at different random seeds and k-medoid clustering was subsequently used to extract clusters (Section 4.1.19). The optimal number of components was extracted by using the silhouette index for guidance. The optimal number of clusters for a component extraction was always times two the number of components (Figure 2.25). This was expected since the sign of a component can randomly change with each extraction. In this way each component was often present with opposite signs within the 200 extractions for a set number of components.

Looking more closely at the average, lowest and second lowest silhouette index of the clusters (Figure 2.26), when using the number of clusters equal to the number of components times two, a steep drop in the silhouette index was observed from 15 to 16 components. Thus, for the downstream analysis steps 15 independent components were used. Those were identified using k-medoid clustering with k = 30. As shown in Figure 2.27, two clusters (recapitulating one component with opposite signs) always correlated with each other with a Pearson coefficient of almost -1. One cluster of each pair was retained for further analysis.

To obtain a better overview of the different components, the contributions and correlations of the input somatic features with the respective components were visualized (Figure 2.28). Except for independent component 2, 7 and 15, all components had at least one somatic feature which correlated with a Pearson coefficient of more than 0.65 with one of the components. Some components such as component 10 and 13 had a strong contribution from one somatic feature (e.g. DBS9 for IC10 and ID4 for IC13) and other components such as component 3 and 12 had a strong contribution from several features.

In contrast to the components extracted from the principal component analysis, many components could be linked to a single underlying cause of mutagenesis. For instance independent component 3 could be linked to a deficiency in DNA mismatch repair due to the high contributions from small indels in microsatellite regions, DNA mismatch repair linked indel signature ID2 and DNA mismatch



	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
50	0	0.2	0.05	0	0	0	0.18	0	0.2	0.16	0	0	0.22	−0.06	0.37	0.27	0.39	0.33	0.5	0.42	0.55	0.55	0.68	0.73	0.43	0.29	0.17	0.29	0.2
49	0	0.08	0.05	0	0	0	0.18	0	0.2	0.16	0	0	0.22	−0.06	0.37	0.27	0.39	0.33	0.5	0.42	0.55	0.55	0.68	0.42	0.34	0.23	0.17	0.3	0.11
48	0	0.08	0.05	0	0	0	0.18	0	0.2	0.16	0	0	0.22	−0.06	0.37	0.27	0.5	0.33	0.5	0.58	0.55	0.47	0.68	0.26	0.29	0.13	0.15	0.29	0.19
47	0	0.08	0.05	0	0	0	0.18	0	0.2	0.19	0	0	0.22	−0.06	0.37	0.27	0.5	0.35	0.5	0.58	0.89	0.56	0.45	0.2	0.22	0.13	0.15	0.14	0.04
46	0	0.08	0.05	0	0	0	0.18	0	0.2	0.19	0	0	0.22	−0.06	0.37	0.27	0.5	0.38	0.58	0.58	0.88	0.85	0.34	0.2	0.22	0.14	0.15	0.14	0.04
45	0	0.2	0.05	0	0	0	0.18	0	0.2	0.19	0	0	0.32	−0.06	0.29	0.45	0.53	0.39	0.58	0.58	0.88	0.42	0.29	0.17	0.22	0.14	0.1	0.14	0.05
44	0	0.2	0.05	0	0	0	0.18	0	0.25	0.25	0	0	0.29	−0.06	0.29	0.45	0.53	0.39	0.45	0.59	0.85	0.28	0.3	0.18	0.21	0.12	0.1	0.11	0.05
43	0	0.2	0.05	0	0	0	0.18	0	0.25	0.26	0	0	0.29	−0.06	0.25	0.45	0.53	0.39	0.48	0.66	0.49	0.28	0.23	0.18	0.18	0.14	0.11	0.09	0.11
42	0	0.2	0	0	0	0	0.18	0	0.25	0.26	0	0	0.29	−0.06	0.31	0.45	0.48	0.39	0.48	0.7	0.31	0.28	0.24	0.2	0.17	0.07	0.07	0.09	0
41	0	0.24	0	0	0	0	0.18	0	0.25	0.26	0	0	0.29	−0.06	0.32	0.45	0.48	0.47	0.64	0.48	0.32	0.28	0.24	0.12	0.17	0.06	0.11	0.07	0
40	0	0.24	0	0	0	0	0.18	0	0.25	0.26	0	0	0.29	−0.06	0.44	0.45	0.48	0.47	0.66	0.46	0.24	0.26	0.08	0.12	0.16	0.07	0.11	0.07	0
39	0	0.24	0	0	0	0	0.29	0	0.31	0.3	0	0	0.19	−0.06	0.44	0.53	0.55	0.5	0.54	0.46	0.24	0.2	0.09	0.09	0.12	0.07	0.06	0.07	0
38	0	0.24	−0.02	0	0	0	0.29	0	0.31	0.3	0	0	0.29	−0.06	0.44	0.34	0.59	0.7	0.49	0.4	0.24	0.19	0.07	0.1	0.12	0.07	0.06	0.07	−0.02
37	0	0.31	−0.02	0	0	0	0.24	0	0.46	0.3	0.34	0.34	0.29	−0.06	0.44	0.71	0.63	0.52	0.4	0.35	0.24	0.15	0.08	0.11	0.1	0.07	0.05	0.04	−0.02
36	0	0.26	0.15	0	0	0	0.29	0	0.46	0.3	0.34	0.34	0.29	−0.06	0.44	0.74	0.64	0.46	0.4	0.28	0.25	0.15	0.05	0.11	0.08	0.07	0.05	0.04	−0.02
35	0	0.21	0.15	0	0	0	0.29	0	0.46	0.3	0.34	0.57	0.29	−0.06	0.47	0.75	0.46	0.24	0.33	0.28	0.26	0.14	0.06	0.11	0.08	0.07	0.05	0.04	−0.01
34	0	0.26	0.15	0	0	0	0.29	0	0.46	0.36	0.34	0.57	0.47	−0.06	0.47	0.77	0.46	0.24	0.28	0.21	0.22	0.14	0.06	0.09	0.08	0.06	0.05	0.03	0
33	0	0.31	0.17	0	0	0	0.29	0	0.46	0.36	0.34	0.33	0.47	0.31	0.52	0.45	0.32	0.24	0.3	0.21	0.22	0.12	0.05	0.09	0.08	0.05	0.05	0.04	0
32	0	0.31	0.15	0	0	0	0.29	0	0.47	0.36	0.34	0.33	0.52	0.35	0.59	0.45	0.32	0.2	0.24	0.21	0.22	0.12	0.05	0.09	0.08	0.05	0.05	0.04	0
31	0	0.19	0.15	0	0	0	0.29	0	0.47	0.36	0.49	0.45	0.59	0.35	0.46	0.44	0.23	0.2	0.24	0.21	0.22	0.11	0.05	0.09	0.07	0.04	0.05	0.03	0
30	0	0.19	0.15	0	0	0	0.29	0	0.45	0.36	0.49	0.45	0.86	0.84	0.41	0.25	0.23	0.19	0.24	0.2	0.22	0.09	0.05	0.1	0.06	0.04	0.05	0.03	0
29	0	0.19	0.15	0	0	0	0.29	0	0.42	0.36	0.49	0.49	0.87	0.47	0.41	0.25	0.23	0.18	0.19	0.2	0.22	0.08	0.06	0.07	0.04	0.04	0.04	0.04	−0.01
28	0	0.19	0.15	0	0	0	0.31	0	0.42	0.36	0.45	0.49	0.84	0.33	0.26	0.22	0.19	0.18	0.19	0.15	0.19	0.06	0.04	0.04	0.03	0.03	0.04	0.05	−0.01
27	0	0.19	0.15	0	0	0	0.31	0	0.42	0.36	0.42	0.49	0.44	0.33	0.26	0.22	0.19	0.19	0.11	0.15	0.14	0.06	0.05	0.05	0.03	0.03	0.04	0.04	−0.01
26	0	0.33	0.17	0	0	0	0.31	0	0.42	0.35	0.42	0.91	0.27	0.33	0.09	0.19	0.19	0.19	0.11	0.15	0.09	0.06	0.05	0.05	0	0.02	0.04	0.02	0
25	0	0.33	0.12	0	0	0	0.41	0	0.42	0.35	0.42	0.48	0.33	0.31	0.11	0.19	0.15	0.11	0.11	0.15	0.09	0.07	0.03	0.05	0.02	0.02	0.04	0.02	−0.01
24	0.4	0.33	0.12	0	0	0	0.41	0	0.42	0.45	0.89	0.48	0.33	0.31	0.12	0.17	0.13	0.09	0.13	0.15	0.07	0.08	0.03	0.01	0.04	0.02	0.04	0.02	−0.01
23	0.4	0.33	0.12	0	0	0	0.43	0	0.42	0.5	0.49	0.48	0.23	0.22	0.09	0.17	0.14	0.09	0.14	0.15	0.06	0.09	0.03	0.01	0.04	0.01	0.04	0.02	−0.01
22	0.4	0.33	0.12	0	0	0	0.43	0	0.42	0.99	0.49	0.27	0.17	0.22	0.09	0.14	0.12	0.09	0.11	0.13	0.06	0.09	0.04	0.02	0.04	0.03	0.03	0.02	0
21	0.4	0.33	0.16	0	0	0	0.5	0	0.42	0.51	0.35	0.27	0.17	0.22	0.1	0.12	0.12	0.09	0.11	0.1	0.07	0.08	0.04	0.02	0.04	0.03	0.05	0.02	0.03
20	0.41	0.33	0.16	0	0	0	0.45	0	0.99	0.51	0.36	0.23	0.12	0.22	0.08	0.12	0.12	0.07	0.11	0.1	0.07	0.09	0.05	0.02	0.04	0.03	0.05	0.02	0.02
19	0.41	0.33	0.16	0	0	0.38	0.45	0	0.5	0.5	0.34	0.23	0.12	0.23	0.08	0.11	0.12	0.07	0.11	0.07	0.06	0.09	0.04	0.02	0.04	0.03	0.05	0.02	0.03
18	0.41	0.33	0.16	0	0	0.38	0.53	0.96	0.5	0.35	0.34	0.19	0.12	0.18	0.07	0.08	0.11	0.05	0.11	0.07	0.06	0.08	0.06	0.01	0.04	0.02	0.05	0.02	0.03
17	0.41	0.33	0.16	0	0	0.23	0.53	0.55	0.37	0.35	0.31	0.19	0.17	0.19	0.05	0.1	0.15	0.05	0.11	0.07	0.06	0.08	0.06	0.01	0.03	0.04	0.05	0.02	0.03
16	0.55	0.42	0.16	0	0	0.4	0.98	0.56	0.4	0.31	0.21	0.15	0.14	0.12	0.05	0.08	0.16	0.07	0.11	0.07	0.06	0.08	0.05	0	0.08	0.04	0.04	0.04	0.03
15	0.6	0.42	0.16	0	0	0.4	0.5	0.51	0.4	0.23	0.21	0.14	0.1	0.13	0.11	0.04	0.11	0.08	0.09	0.07	0.06	0.08	0.06	0.01	0.06	0.06	0.05	0.04	0.03
14	0.65	0.42	0.22	0.89	0.14	0.95	0.32	0.51	0.41	0.25	0.19	0.14	0.15	0.12	0.08	0.04	0.12	0.09	0.1	0.1	0.06	0.05	0.08	0.01	0.07	0.06	0.05	0.04	0.03
13	0.71	0.42	0.22	0.89	0.93	0.52	0.32	0.37	0.33	0.21	0.16	0.14	0.13	0.11	0.12	0.12	0.12	0.09	0.1	0.1	0.06	0.05	0.08	0.01	0.07	0.06	0.05	0.03	0.04
12	0.74	0.42	0.22	0.91	0.93	0.52	0.27	0.21	0.25	0.17	0.14	0.15	0.14	0.13	0.11	0.1	0.14	0.1	0.08	0.09	0.07	0.05	0.04	0.05	0.08	0.06	0.06	0.03	0.04
11	0.74	0.42	0.3	0.91	0.47	0.52	0.27	0.21	0.24	0.13	0.14	0.16	0.14	0.13	0.11	0.11	0.13	0.1	0.09	0.08	0.07	0.08	0.07	0.05	0.08	0.06	0.06	0.03	0.03
10	0.76	0.42	0.3	0.93	0.35	0.49	0.19	0.16	0.25	0.14	0.12	0.14	0.14	0.11	0.08	0.12	0.11	0.11	0.09	0.05	0.07	0.05	0.09	0.05	0.06	0.05	0.06	0.02	0.05
9	0.8	0.53	0.33	0.51	0.35	0.33	0.23	0.18	0.21	0.13	0.1	0.13	0.13	0.15	0.07	0.12	0.11	0.11	0.1	0.06	0.06	0.05	0.06	0.04	0.08	0.06	0.06	0.05	0.05
8	0.97	0.53	1	0.35	0.3	0.2	0.18	0.2	0.16	0.13	0.1	0.11	0.14	0.13	0.12	0.14	0.12	0.13	0.11	0.05	0.05	0.06	0.05	0.03	0.07	0.07	0.05	0.04	0.05
7	0.95	0.53	0.49	0.35	0.3	0.2	0.18	0.2	0.16	0.14	0.13	0.12	0.14	0.14	0.1	0.12	0.13	0.11	0.1	0.04	0.04	0.03	0.05	0.03	0.08	0.07	0.06	0.05	0.06
6	0.93	1	0.34	0.35	0.25	0.23	0.18	0.15	0.17	0.14	0.12	0.18	0.15	0.12	0.12	0.12	0.14	0.12	0.08	0.02	0.01	0.03	0.03	0.04	0.07	0.05	0.04	0.05	0.07
5	0.93	0.52	0.36	0.35	0.25	0.25	0.18	0.21	0.12	0.14	0.12	0.14	0.13	0.13	0.1	0.13	0.12	0.11	0.1	0.02	0.01	0.01	0.02	0.03	0.08	0.04	0.08	0.02	0.05
4	0.92	0.52	0.34	0.26	0.23	0.24	0.18	0.16	0.1	0.14	0.13	0.11	0.12	0.14	0.12	0.13	0.13	0.12	0.08	0	0.01	−0.01	0	0.01	0.09	0.07	0.07	0.01	0.05
3	0.57	0.43	0.22	0.22	0.21	0.25	0.16	0.15	0.14	0.13	0.14	0.1	0.13	0.13	0.13	0.13	0.12	0.1	0.09	0.02	0.09	0.06	0	0.01	0.08	0.06	0.05	0.05	0.06
2	0.62	0.49	0.28	0.33	0.3	0.26	0.24	0.17	0.15	0.16	0.15	0.12	0.15	0.14	0.13	0.13	0.13	0.12	0.1	0.09	0.1	0.07	0.09	0.08	0.09	0.05	0.06	0.06	0.07

Figure 2.25: Finding the optimal number of independent components. Independent component extraction was run with increasing number of independent components (x-axis). Each extraction was performed 200 times and then further clustered using k-medoid clustering with increasing number of clusters (y-axis). Color code as well as number of each tile shows the minimum silhouette index of a cluster.

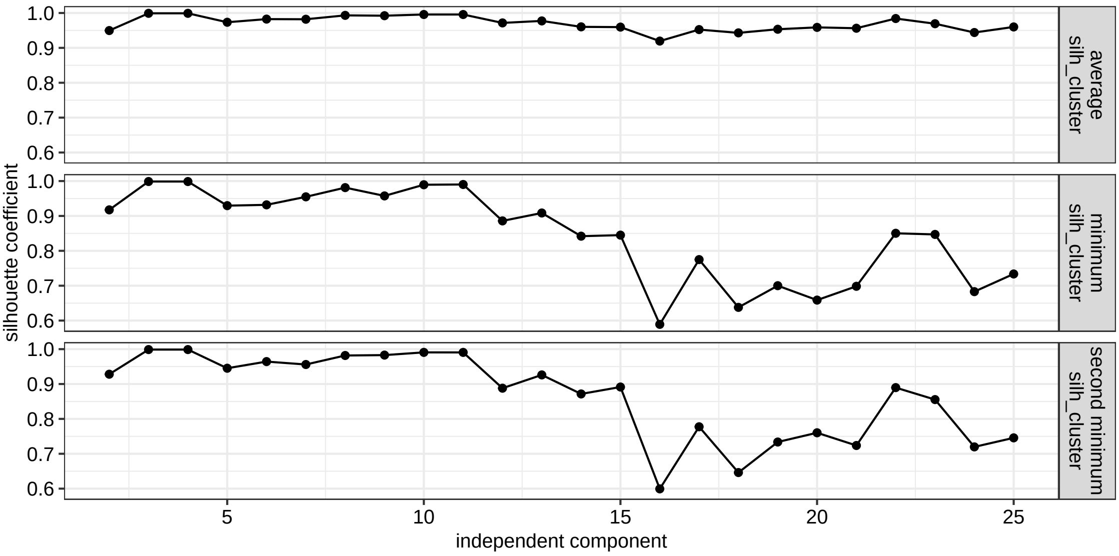


Figure 2.26: Selection of 15 independent components for further analysis. Showing the average, minimum and second minimum silhouette index of the clusters when extracting 2 times more clusters for a set number of components.

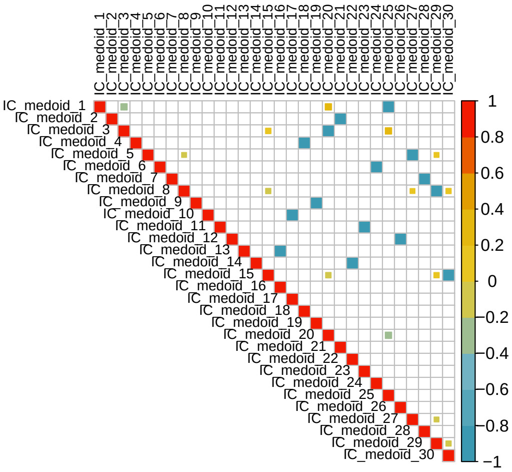


Figure 2.27: Pearson correlations between all 30 independent components which were extracted using 15 components and k-medoid clustering with k = 30. Each component occurred twice with opposite signs.

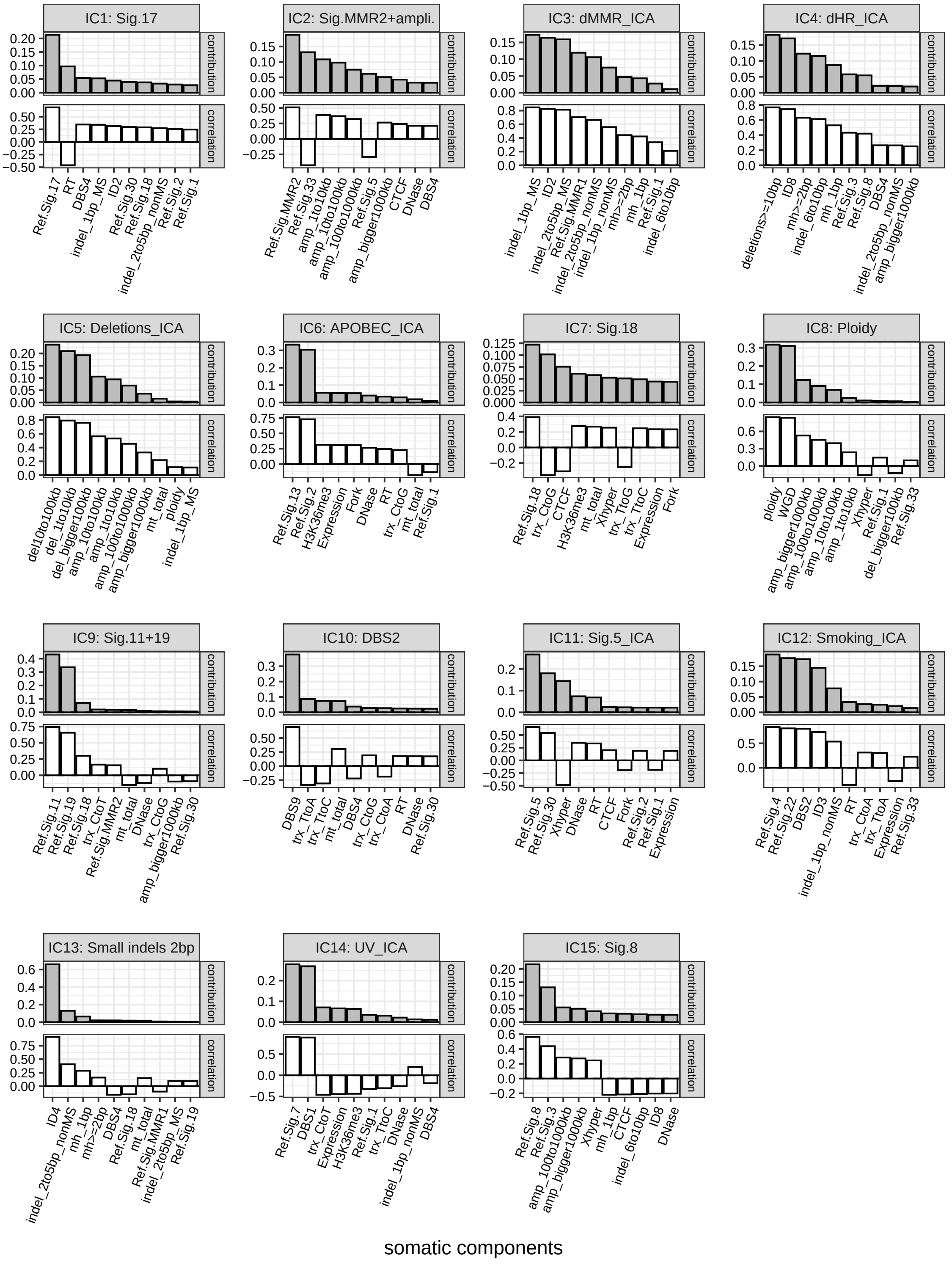


Figure 2.28: Overview of strongest contributing features to the independent components. Showing the Pearson correlation (white bars) and contribution (fraction of 1) (grey bars) of the 10 strongest somatic features to the respective components. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation. Components were renamed based on strongest correlating somatic features.

repair linked SNV signature Ref.Sig. MMR1. This was also shown by checkking the scores for independent component 3 in MSI vs. MSS samples across different tissues. MSI samples had significantly increased estimates of this component in comparison to MSS samples across several tissues (Figure 2.29).

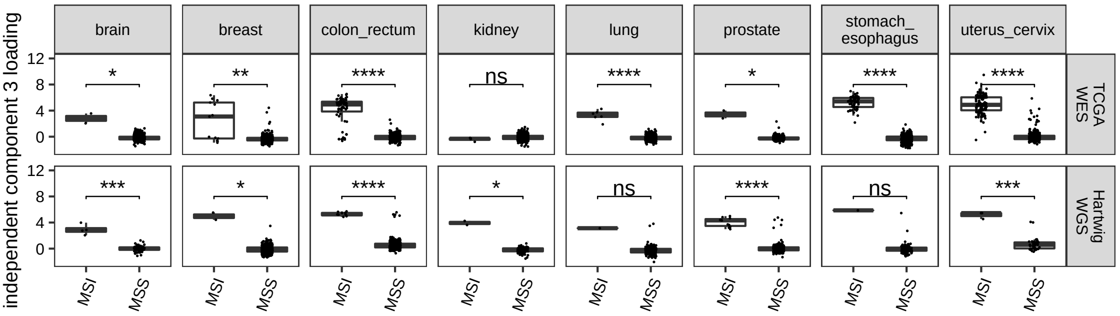


Figure 2.29: Independent component 3 scores were increased in samples with a deficiency DNA mismatch repair. Samples were split into MSI vs. MSS samples across different tissues. MSI was assigned via the MSI detection tool MANTIS for the TCGA cohort. MSI assignments for Hartwig were extracted from the flagship paper[9]. Significance given by two-tailed Mann-Whitney U test: ns: $p > 0.05$, *: $p \leqq 0.05$., **: $p \leqq 0.01$, ***: $p \leqq 0.001$, ****: $p \leqq 0.0001$ and NS: not enough observations to test.

Further, based on the previously reported signatures for different sources of mutations[6,14], component 3 could be linked to a deficiency in homologous recombination-directed repair, independent component 6 to APOBEC activity, independent component 12 to smoking and independent component 14 to UV exposure. Notably, somatic features based on different mutation classes were often grouped together. For instance component 12 which was linked to smoking, grouped a SNV based signature, a DNV based signature and a indel based signature. Also features measuring strand biases and regional enrichments grouped together with other cancer signatures, such as the mutation enrichment with regards to replication timing grouped together with SNV based signature 17 in indepencent component 1. Copy number based features in particular all grouped together in components 5 and 8, but also grouped together with other mutation classes in components 2 and 15.

Next we also investigated to which extent some components were enriched in specific tissues. It was tested within each cohort for each cancer type whether its independent component scores were significantly different from the rest via a unpaired Welch t-test (Section 4.1.21). Effect sizes were estimated by calculating

Cohen's *d* (difference in means divided by the standard deviation) and then grouped by tissues and ordered by the mean value within a tissue (Figure 2.30). As expected, component 14 (UV exposure) was strongly enriched in skin tissue derived cancers, component 12 (smoking) was strongly enriched in lung tissue derived tumors and component 3 scores (deficiency in DNA mismatch repair) were increased in colon and rectal tissue derived cancers. Further, component 9 showed a clear enrichment in brain tissue derived cancers. We were not able to link this component to a mutational cause. Some components such as components 7, 8, 13, and 15 did not show a clear enrichment in a specific tissue.

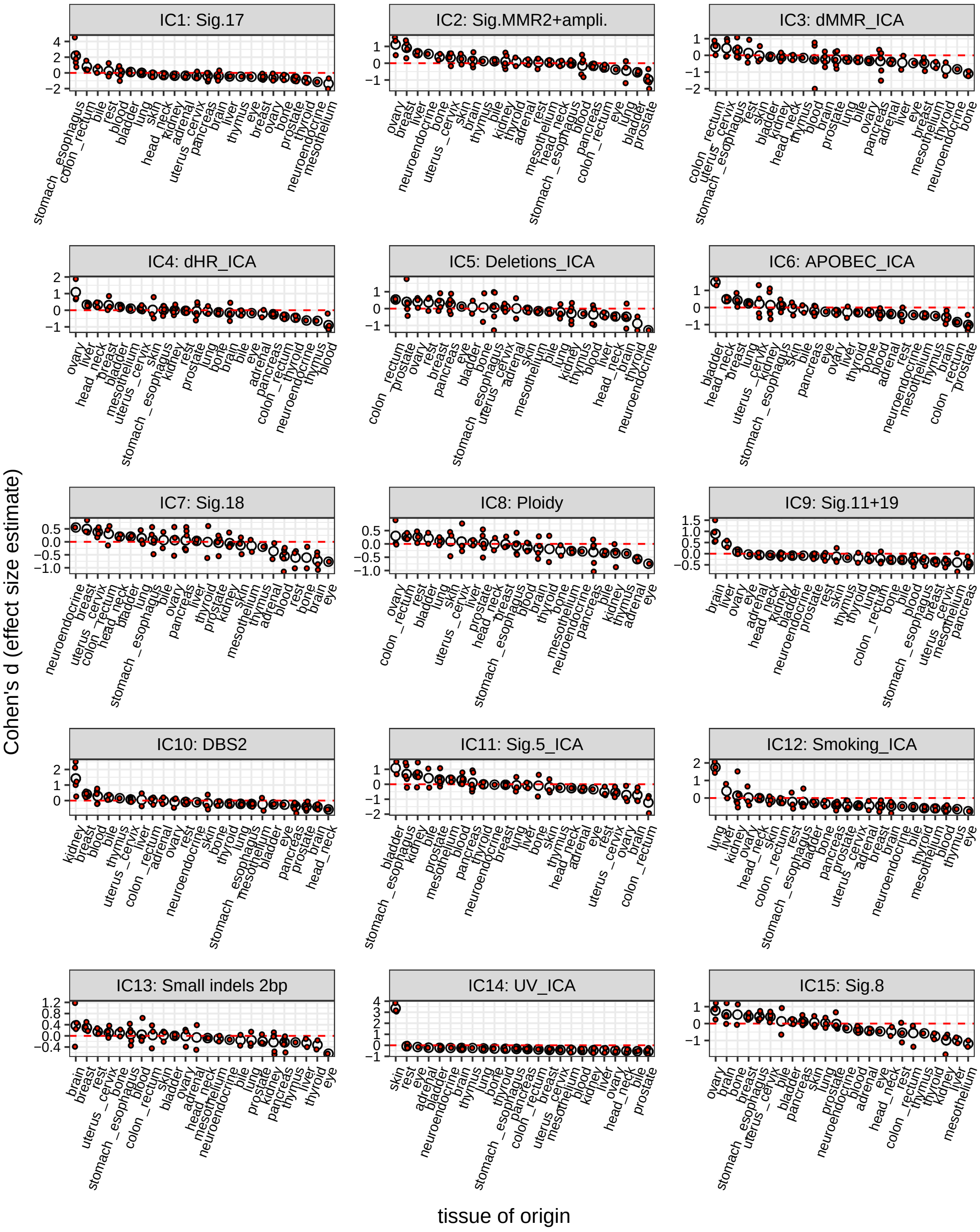


Figure 2.30: Several independent component scores were enriched in specific tissue of origins. Cohen's *d* (effect size estimate) was calculated for each cancer type, grouped by tissue of origin and, then the average value was estimated. Average effect size estimates were ordered by decreasing value for each independent component. Components were renamed based on strongest correlating somatic features.

2.1.10 Extracting Components using Variational Autoencoder Neural Network

With the ICA we were able to extract some components which covered known underlying causal mechanisms of mutagenesis. Since during the last years VAE neural networks have shown promising results in capturing biologically relevant representations[267], we also tested VAEs as an alternative way of deconvolving the input features into somatic mutational components. Our approach of component extraction was inspired by the work of Way *et al.*[267], who used a VAE to compress gene expression data. Firstly, a parameter sweep was performed to find the optimal settings for the hyperparameters (Section 4.1.20). The data was split into 90 % training data and 10 % validation data and stratified for gender and cancer type. Two different measurements were evaluated to access performance: the correlation of the reconstructed input with the initial input and the average correlation of the VAE-derived components, which had the highest correlation with the four ICs which covered known mutational sources (UV$_{ICA}$, smoking$_{ICA}$, dMMR$_{ICA}$ and dHR$_{ICA}$). The correlation with the input increased with increasing number of epochs, with increasing number of components and was the highest at a batch size of 50 samples (Figure 2.31). At a high number of epochs, the tested learning rate, number of extra hidden layers and the factor kappa only had negligible effects on the correlation with the input. The correlation was always above 0.95. In contrast, the average correlation with the four ICs quickly reached saturation after 75 epochs and increased with higher batch sizes. Further, the correlation was always higher when only using one hidden layer compared to using two hidden layers. Here, the factor kappa again only had small effects and correlations were higher on average at a learning rate of 0.0005.

Based on these results, we ran the VAE with increasing number of components with the following settings: learning rate = 0.0005, batch size = 200, kappa = 0.5, and number of epochs = 200. Since we wanted to check whether the extra layer led to new, potentially interesting components, we ran the VAE using 1 and 2 hidden layer(s). For each component extracting we re-ran the VAE five times with different random initiations. We evaluated the results again by looking at the average correlation of the VAE-derived components, which had the highest correlation with the four mentioned ICs since we already knew that these ICs covered known mutational mechanisms and biological representation was more important for our study than minimizing the reconstruction loss. As shown in Figure 2.32, the average correlation increased quickly with increasing number

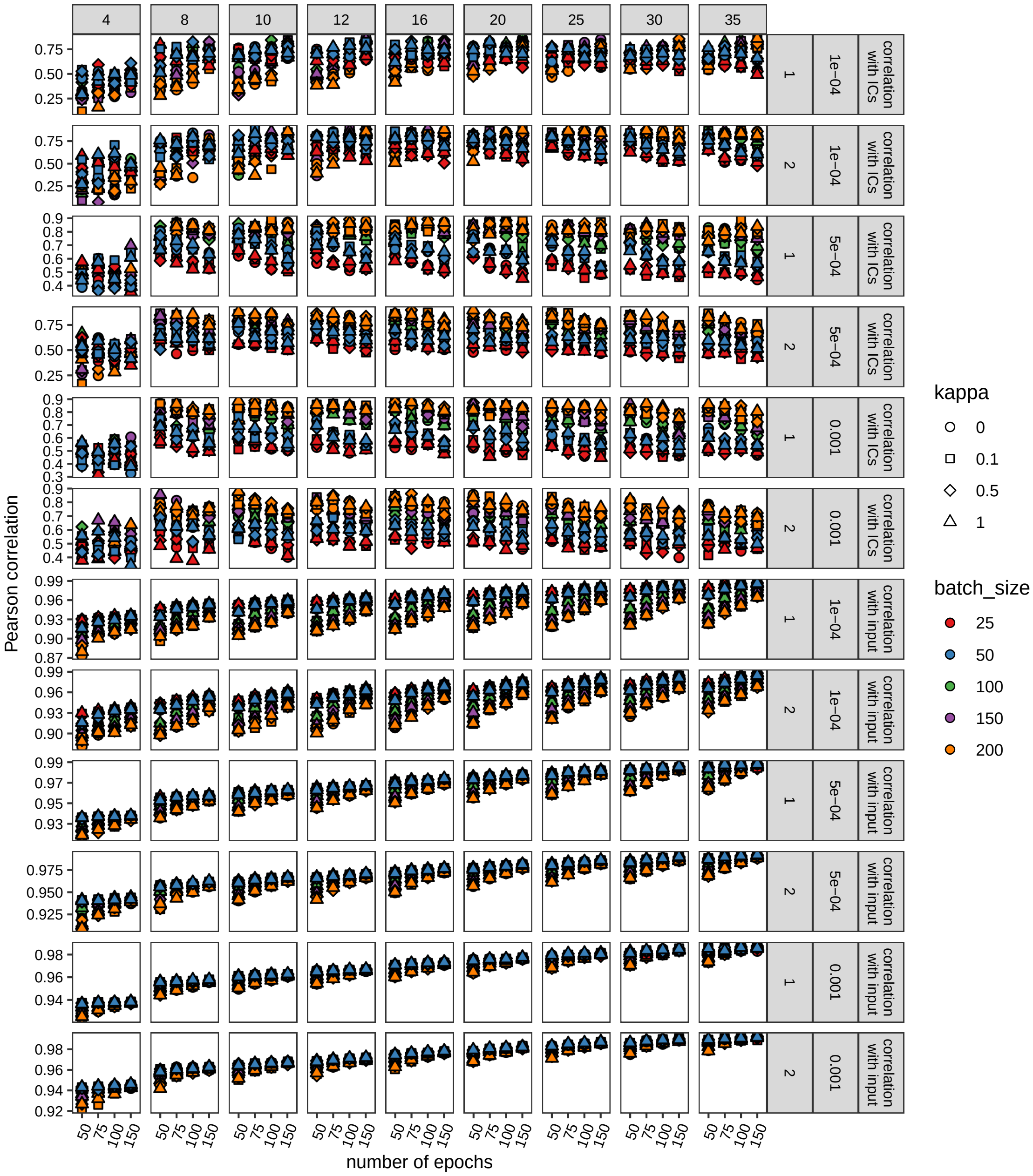


Figure 2.31: Finding the optimal hyperparameters for the variational autoencoder. Increasing number of epochs (x-axis), different kappa factors (point shape), increasing batch sizes (point color), three different learning rates (0.001, 0.0005, and 0.0001), different number of hidden layers between latent space and input/output (either 1 or 2) and different number of components (4, 8, 10, 12, 16, 20, 25, 30, and 35) were tested. Evalutation by measuring the average Pearson correlation with four ICs (UV, smoking, dMMR, and dHR ICs) and by calculating the Pearson of the reconstructed input with the initial input (y-axis).

of components, reached saturation around 14 components and then started to decrease again. The inital increase in the correlation was expected since it was unlikely to capture the four ICs with a latent space of less than 4. We observed an optimum at 14 components for both scenarios (1 or 2 hidden layers between latent space and input/output).

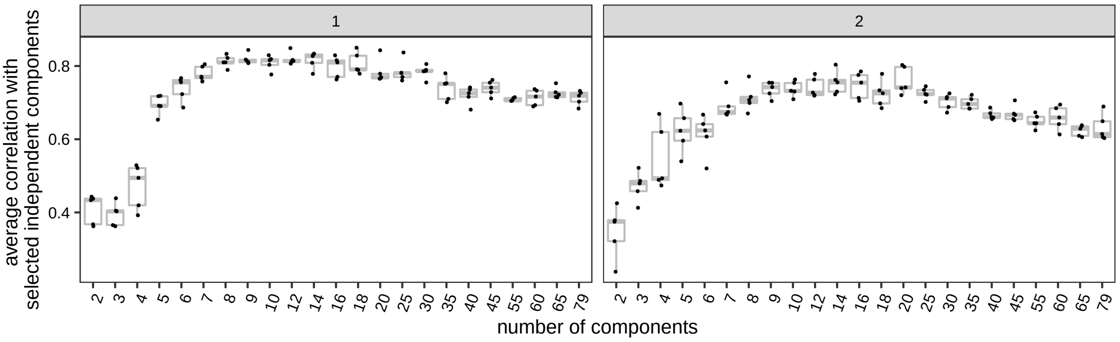


Figure 2.32: Correlation with biologically relevant components increased with increasing number of component extractions and quickly reached saturation. Component extractions were run 5 times for each set component number with different random initiations. Number of components (x-axis) are shown against the average Pearson correlation with four biologically relevant IC components (UV, smoking, dMMR, dHR). Facet for either using 1 hidden layer or 2 hidden layers between latent space and input/output.

Next, it was checked whether the components derived from the VAE using either 1 or 2 hidden layers between latent space and input/output differed. When looking at the Pearson correlation between then different components it can be seen that many components were extracted in both cases and showed strong correlations between each other (Figure 2.33) such as VAE_depth1_1 with VAE_depth2_7, VAE_depth1_2 with VAE_depth2_2 and VAE_depth1_6 with VAE_depth2_14. 7 ouf 14 components had a Pearson correlation higher than 0.8, 9 out of 14 had a Pearson correlation higher than 0.7, and 14 out of 14 components has a Pearson correlation higher than 0.5 with at least one component from the VAE component extraction with a different number of hidden layers. Thus, no component was uniquely found with only one component extraction. For this reason, we further focussed on the component extraction using only one hidden layer since a simpler architecture was favored. From here one, all VAE-derived components were extracted using 1 hidden layer between latent space and input/ouput.

Having a closer look at the VAE-derived components (Figure 2.34), we ob-

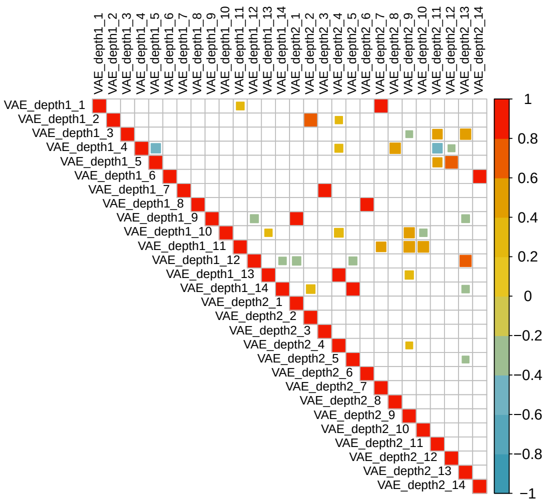


Figure 2.33: Number of hidden layers barely made a difference on the extracted components in the latent space of the variational autoencoder. Pearson correlation between the 14 extracted components with the variational autoencoder using either 1 hidden layer (depth = 1) or 2 hidden layers (depth = 2) between latent space and input/output.

served as it was also optimized for, a component capturing tobacco smoking induced mutagenesis (VAE_9), a components capturing UV-induced mutagenesis (VAE_5), components capturing a deficiency in DNA mismatch repair (VAE_10 and VAE_13), and components capuring a deficiency in homologous recombination-directed repair (VAE_6 and VAE_8). Further, VAE_4 correlated strongly with cancer signatures Ref.Sig. 1 and 5 and anti-correlated with the somatic feature for X-chromosomal hypermutation. VAE_7 showed a strong correlation with the somatic feature measuring the number of mutations in the mitochondrial genome and VAE_12 anticorrelated with several amplification events, small indels in non microsatellite regions and some other types of indels.

Next, similarly as in the ICA, we also investigated tissues enrichments of the individual components (Figure 2.35). Expectably, the VAE-derived component capturing UV-induced mutagenesis (VAE_5) was enriched in skin cancers and and the VAE-derived component capturing tobacco smoking-induced mutagenesis (VAE_9) was enriched in lung cancers. Component VAE_10, which covered a form of dMMR was enriched in kidney cancers, while VAE_13, which also covered a form of dMMR, was enriched in stomach, esophageal, colon, and rectal cancers.

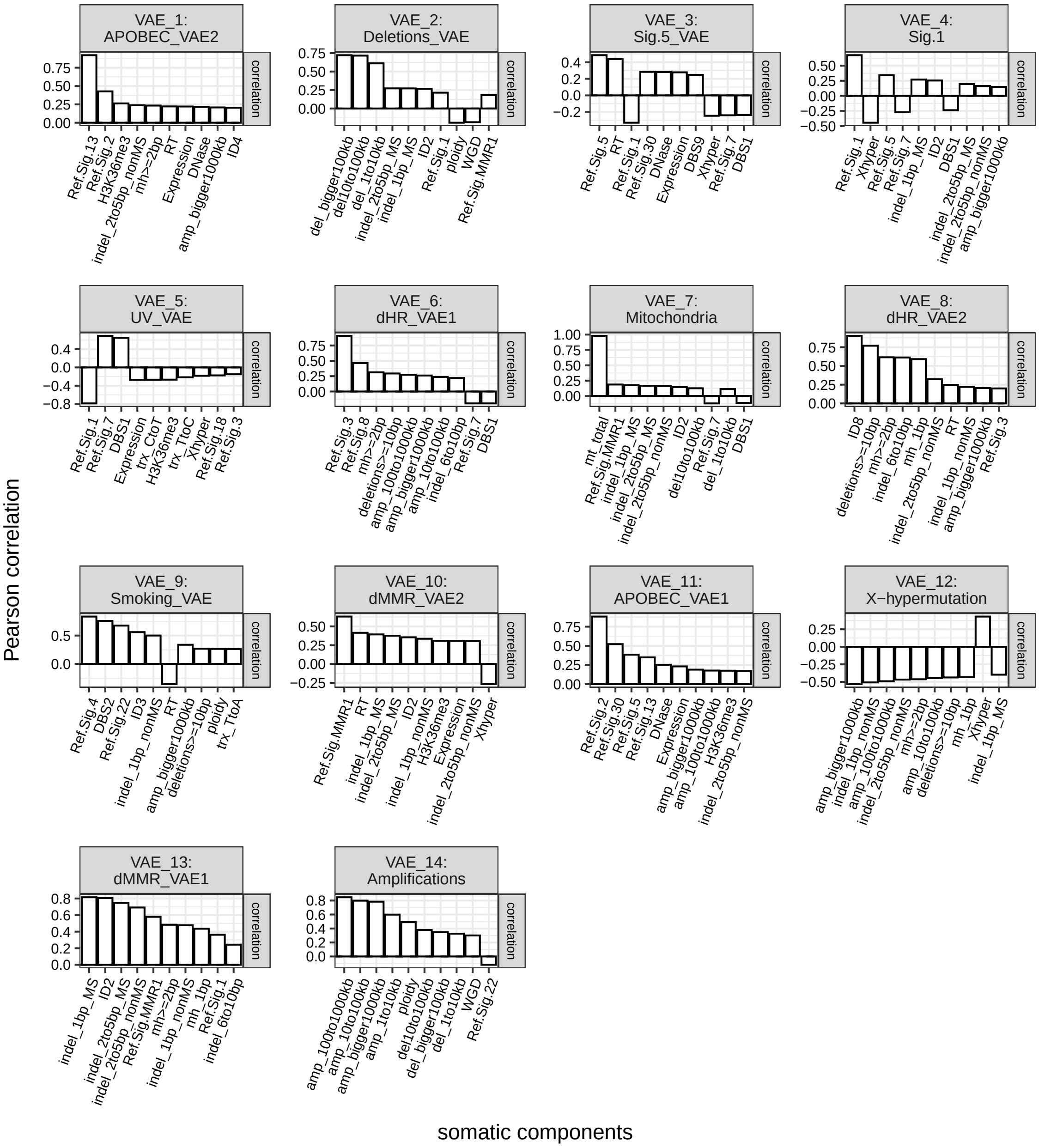


Figure 2.34: Overview of strongest contributing features to the variational autoencoder derived components. Showing the Pearson correlation (white bars) of the 10 strongest somatic features to the respective components, which were extracted via 1 hidden layer between latent space and input/output. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation. Components were renamed based on strongest correlating somatic features.

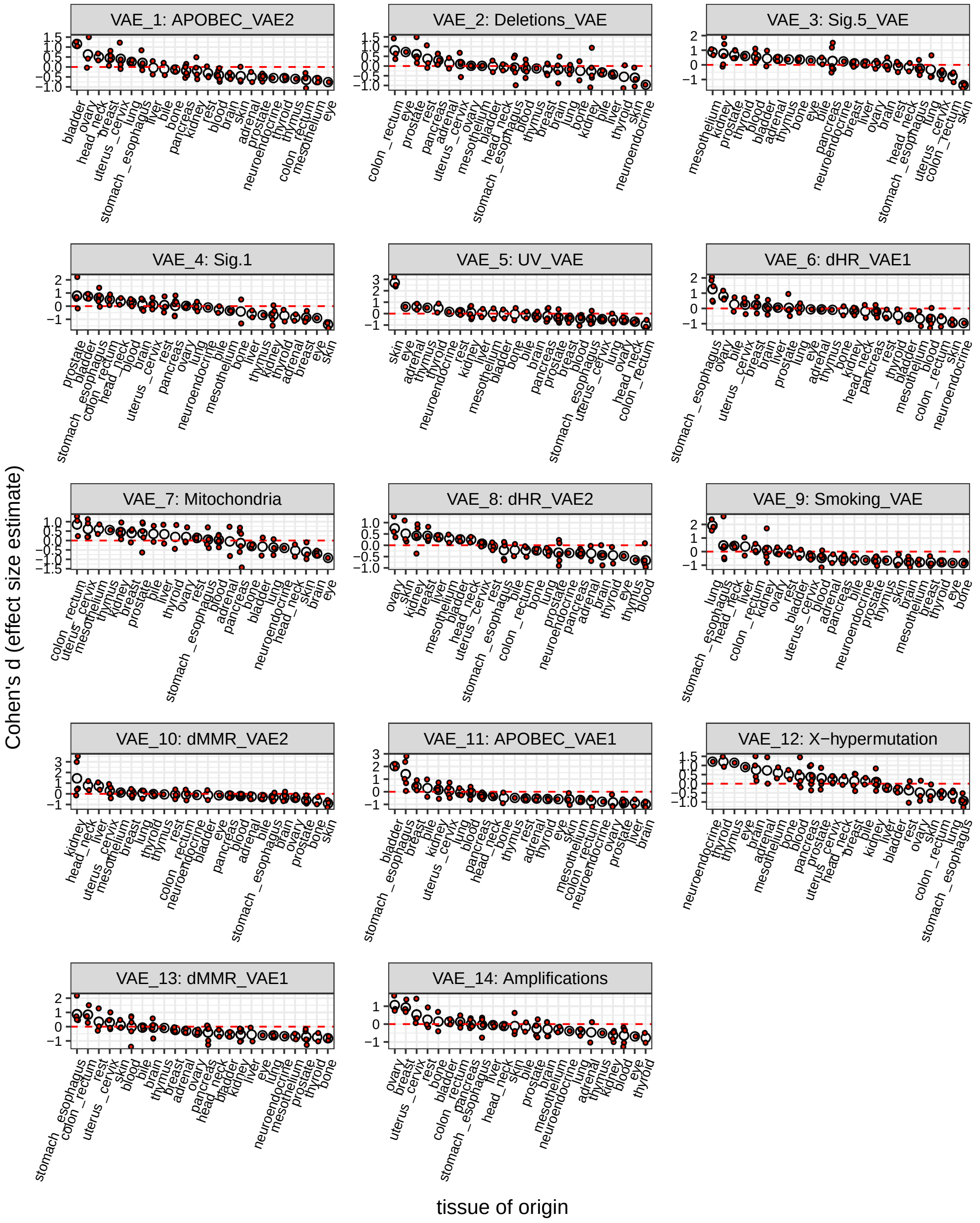


Figure 2.35: Several VAE-derived component scores were enriched in specific tissue of origins. Cohen's *d* (effect size estimate) was calculated for each cancer type, grouped by tissue of origin and, then the average value was estimated. Average effect size estimates were ordered by decreasing value for each VAE-derived component. Components were renamed based on strongest correlating somatic features.

2.1.11 Final Set of Somatic Components for Association Testing

We further investigated which approach (ICA or VAE) was better suitable for association testing. When looking at the Pearson correlations between the 14 VAE-derived components and 15 ICA-derived components (Figure 2.36), we observed strong correlations between several components such as VAE_13 with IC3 (dMMR), VAE_8 with IC3 (dHR), or VAE_9 with IC12 (smoking). Interestingly, the components VAE_4 , VAE_7 and VAE_12 extracted from the VAE did not have a strong correlation (R = < 0.5 and > -0.5) with any of the components extracted from the ICA. Thus, some components were unique to the respective technique.

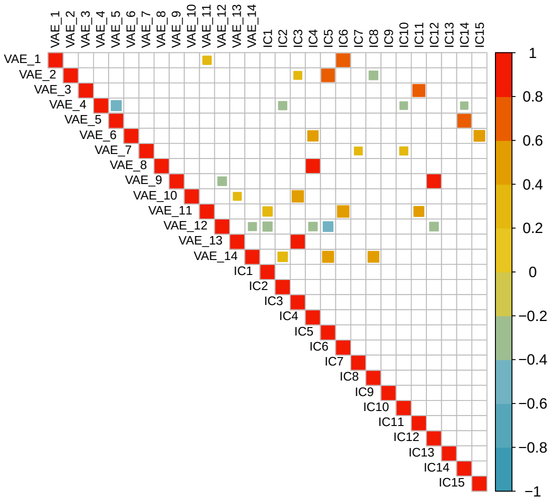


Figure 2.36: Some VAE-derived components were not captured in the independent component analysis. Showing the Pearson correlation between the VAE-derived components using 1 hidden layer (depth = 1) and the ICs.

Since it was not clear which set of components would be more appropriate for associationg testing, both were kept (Figure 2.37). Thus, 14 VAE-derived components and 15 ICA-derived components were retained for association testing. The distribution of component scores across the discovery and validation cohort are shown in Figures 2.38 and 2.39.

To get a better overview of all somatic compenents, which were extracted in this study from over 15,000 cancer genomes, the components were clustered together and renamed based on the underlying mutational source of the feature(s)

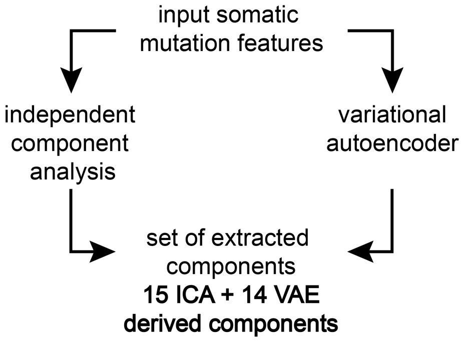

input somatic mutation features
independent component analysis
variational autoencoder
set of extracted components
15 ICA + 14 VAE derived components

Figure 2.37: Approach to extract final set of somatic components. Final set of somatic components was extracted by applying two methods to the input matrix (samples as rows and somatic input features as columns): independent component analysis (ICA) and a variational autoencoder (VAE). 15 ICA-derived and 14 VAE-derived components were extracted.

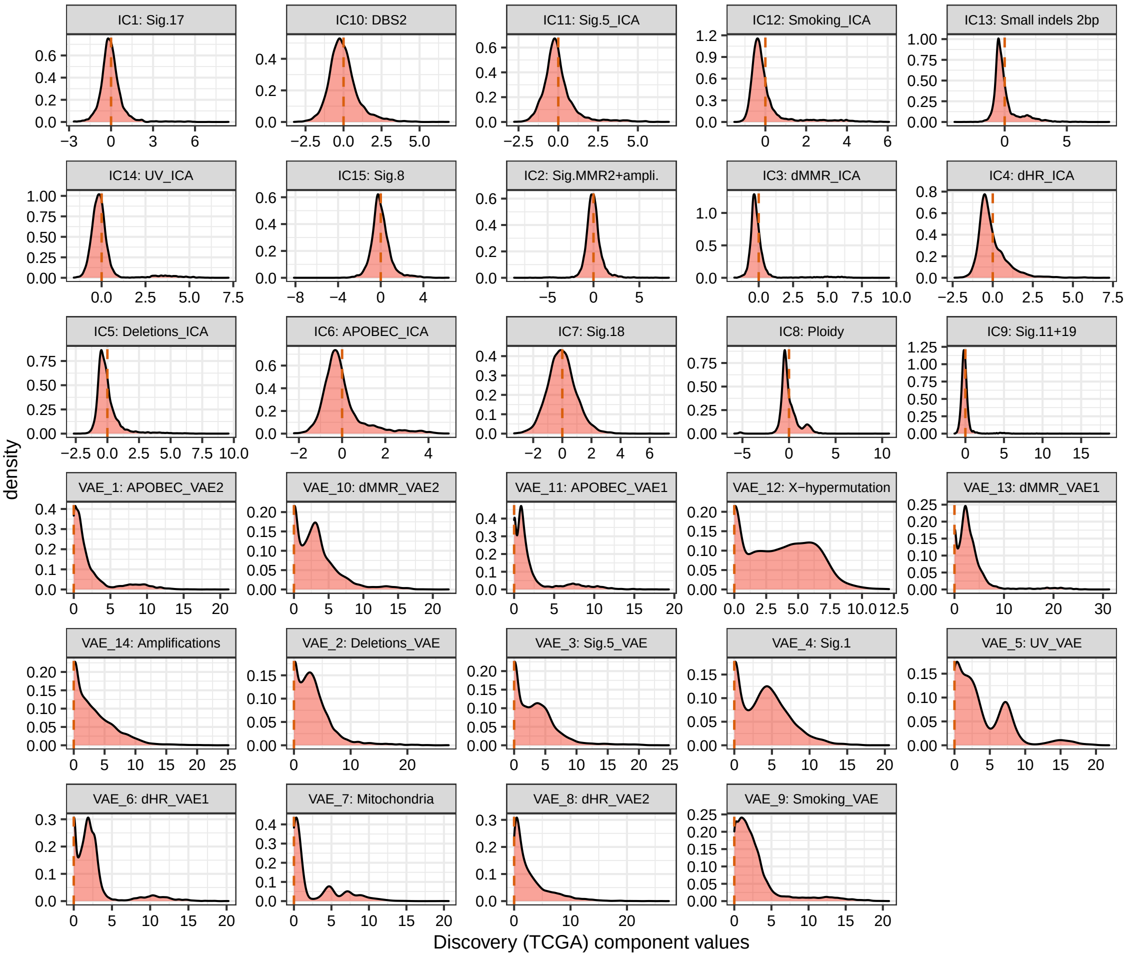


Figure 2.38: Distribution of all 29 somatic components in TCGA-WES.

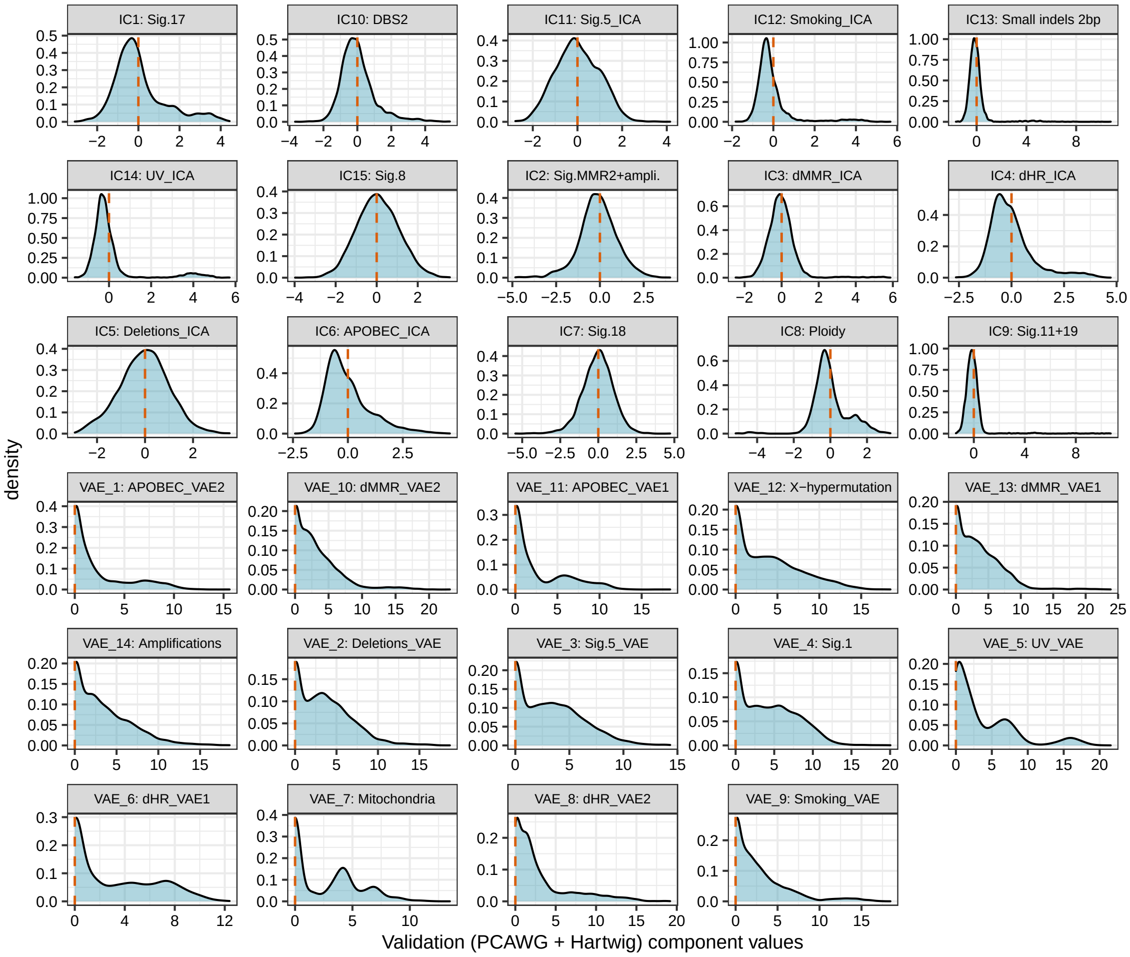


Figure 2.39: Distribution of all 29 somatic components in PCAWG_Hartwig -WGS.

which had the strongest correlation with a somatic component (Figure 2.40). Thirteen of the 29 components captured known mutagenic mechanisms, including UV radiation exposure (UV$_{ICA}$ and UV$_{VAE}$, including CC>TT substitutions), tobacco smoking (Smoking$_{ICA}$ and Smoking$_{VAE}$), deficiencies in MMR (dMMR$_{ICA}$, dMMR$_{VAE1}$, and dMMR$_{VAE2}$), deficiency in the repair of DSBs via homologous recombination (dHR$_{ICA}$, dHR$_{VAE1}$, and dHR$_{VAE2}$), and APOBEC-directed mutagenesis (APOBEC$_{ICA}$, APOBEC$_{VAE1}$, and APOBEC$_{VAE2}$). Many of the components combined different classes of mutational features. For instance, dMMR$_{VAE2}$, has a high correlation with the SNV signature Ref.Sig. MMR, several types of short indels at microsatellite loci and the relative mutation rate with respect to replication timing. The remaining 16 components did not have a known mechanistic cause but could be further described via the features with which they are strongly correlated. For instance, we extracted components covering X-chromosomal hypermutation (X-hypermutation), a component covering mitochondrial SNVs (Mitochondria), and two components related to SNV-signature 5 mutations (Ref.Sig. 5$_{ICA}$ and Ref.Sig.5$_{VAE}$).

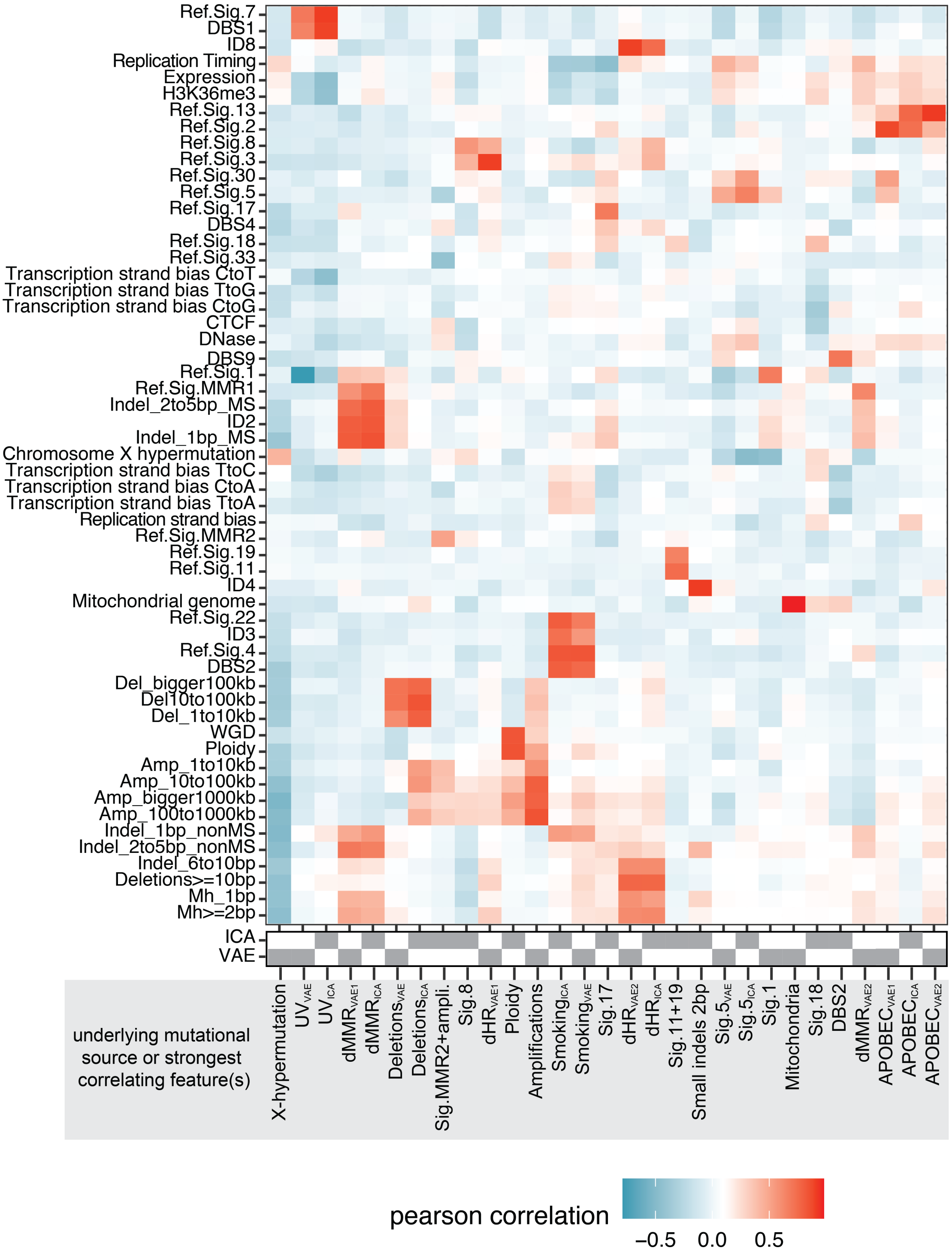


Figure 2.40: Overview of somatic mutation components extracted from 15,000 human tumors. Overview of extracted somatic components (x-axis) and their Pearson correlation (color code) with the input somatic features (y-axis). Grey strip at the bottom displays whether the component was extracted via ICA or VAE. Components were named based on the underlying mutational process or strongest correlating input feature(s).

2.2 Rare Damaging Germline Variant Association Testing

After extracting the different somatic mutational components as described above, we investigated to which extent inherited variants impact the distribution of these components across individuals. For this purpose, we started with rare damaging germline variants (RDGVs) and performed RDGV association testing (Figure 2.41). Of note, all the results from this section have been published on bioRxiv[198].

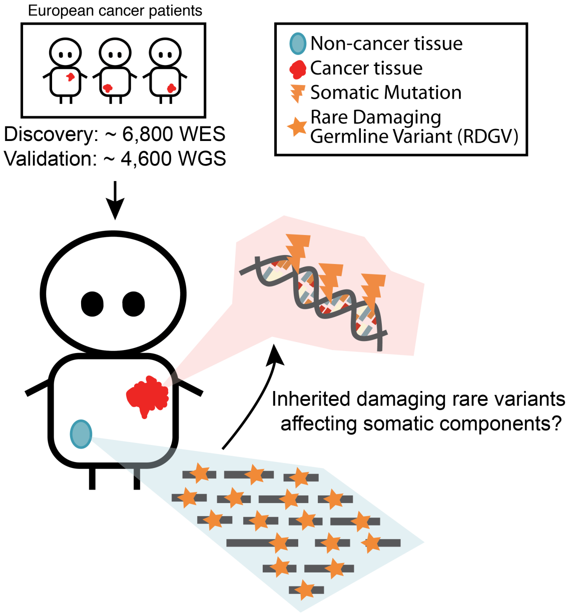


Figure 2.41: Illustration of rare damaging variant association testing. Associations were identified in the discovery cohort (TCGA-WES) and replicated in the validation cohort (PCAWG + Hartwig-WGS).

First of all, sample level quality control was performed and to control for population structure individuals of European ancestry were extracted since this was the biggest group in our cohort (Section 2.2.1). RDGV association testing was performed in the discovery cohort TCGA-WES and identified hits were re-tested in the fully independent validation cohort PCAWG_Hartwig-WGS. Since RDGV could not be tested individually due to the sample size, we performed gene-level based rare variant association testing (Section 2.2.2).

2.2.1 Sample Level Quality Control and Extraction of Individuals of European Ancestry

Sample level filtering was performed to remove potential biases before performing associating testing. To minimize population-specific effects, individuals of European ancestry were extracted by utilizing the common variants (MAF >5 %) in the datasets and a PCA was performed to use the first PCs to control for population substructure by using them as covariates. In line with the study design of having a discovery and validation cohort, these steps were performed for the discovery cohort TCGA -WES and the validation cohort PCAWG_Hartwig -WGS separately.

First of all, individuals with an outlying total number of variants (Figures 2.42a and 2.43a and b) or outlying heterozygosity rate (Figures 2.42b and 2.43c and d) were discarded. Next, duplicated or highly related samples were removed by checking the proportion of IBD between all pairwise combinations of samples. As it can be seen in Figures 2.42c or 2.42e, most samples were unrelated. In TCGA -WES 542 samples and in PCAWG_Hartwig -WGS 479 samples with an IBD > 0.185 (expectation for individuals between third- and second-degree relatives) were discarded.

Next, a PCA was performed on the common variants (MAF >5 %) to check for major confounding and to extract individuals of European ancestry. The first 6 PCs did not show any confounding due to the tissue of origin, sequencing center, whole genome amplification status, sequencing system, gender, or age at diagnosis (Figure 2.44). The same analysis was performed in PCAWG_Hartwig - WGS as well and no confounding was detected when checking for cancer type, center/study where the germline variants were called, gender, or age of diagnosis (Figure 2.45).

As expected, the first two PCs in particular (as well as other PCs) recapitulated the population structure (Figures 2.46a and 2.47a). To extract individuals of European ancestry, the first 10 PCs were used for clustering and different population ancestries were mostly captured in individual clusters as shown in Figures 2.47b and 2.47c for TCGA -WES and in Figures 2.47b and 2.47c for PCAWG_Hartwig -WGS. Subsequently, individuals in the respective clusters, which were enriched for individuals with reported European ancestry, were kept. A PCA was performed on the remaining samples and the estimated PCs were

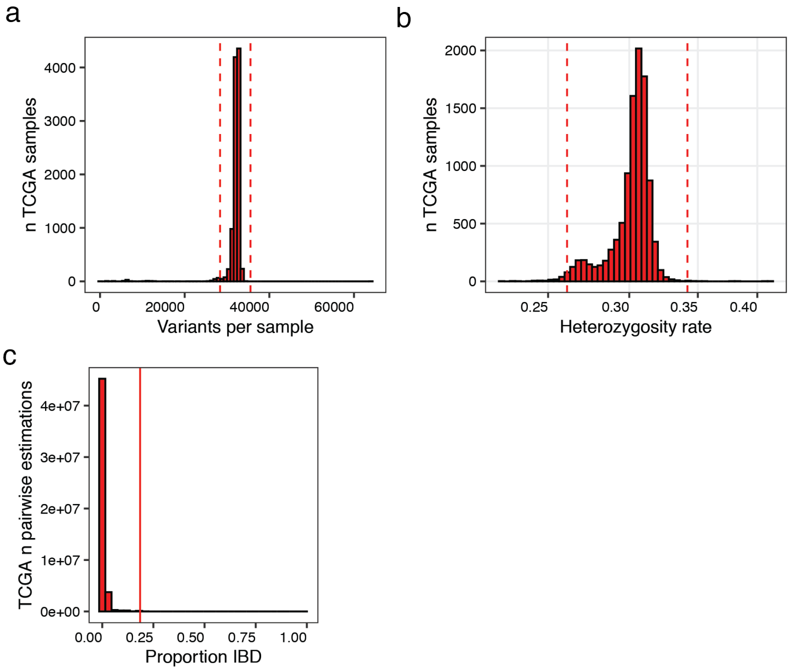


Figure 2.42: Identification of individuals with outlying total number of variants, outlying heterozygosity rate or high relatedness in TCGA -WES. (a) Distribution of total number of variants across samples. Red dashed lines at 1.5 standard deviations away from the mean. (b) Distribution of heterozygosity rate across samples. Red dashed lines at 3 standard deviations away from the mean. (c) Proportion of identity-by-descent (IBD) across all sample pairs. Solid red line at 0.185.

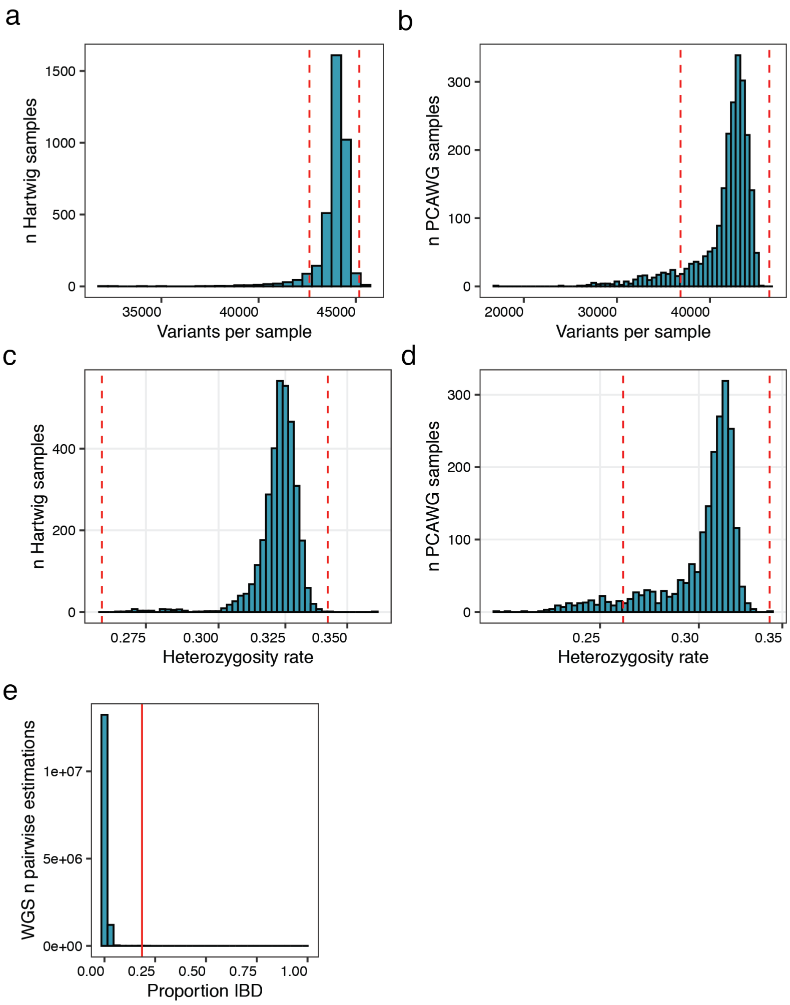


Figure 2.43: Identification of individuals with outlying total number of variants, outlying heterozygosity rate or high relatedness in PCAWG_Hartwig-WGS. Distribution of total number of variants across samples in (a) Hartwig and (b) PCAWG. Red dashed lines at 1.5 standard deviations away from the mean. Distribution of heterozygosity rate across samples in (c) Hartwig and (d) PCAWG. Red dashed lines at 3 standard deviations away from the mean. (e) Proportion of identity-by-descent (IBD) across all sample pairs (PCAWG and Hartwig merged). Solid red line at 0.185.

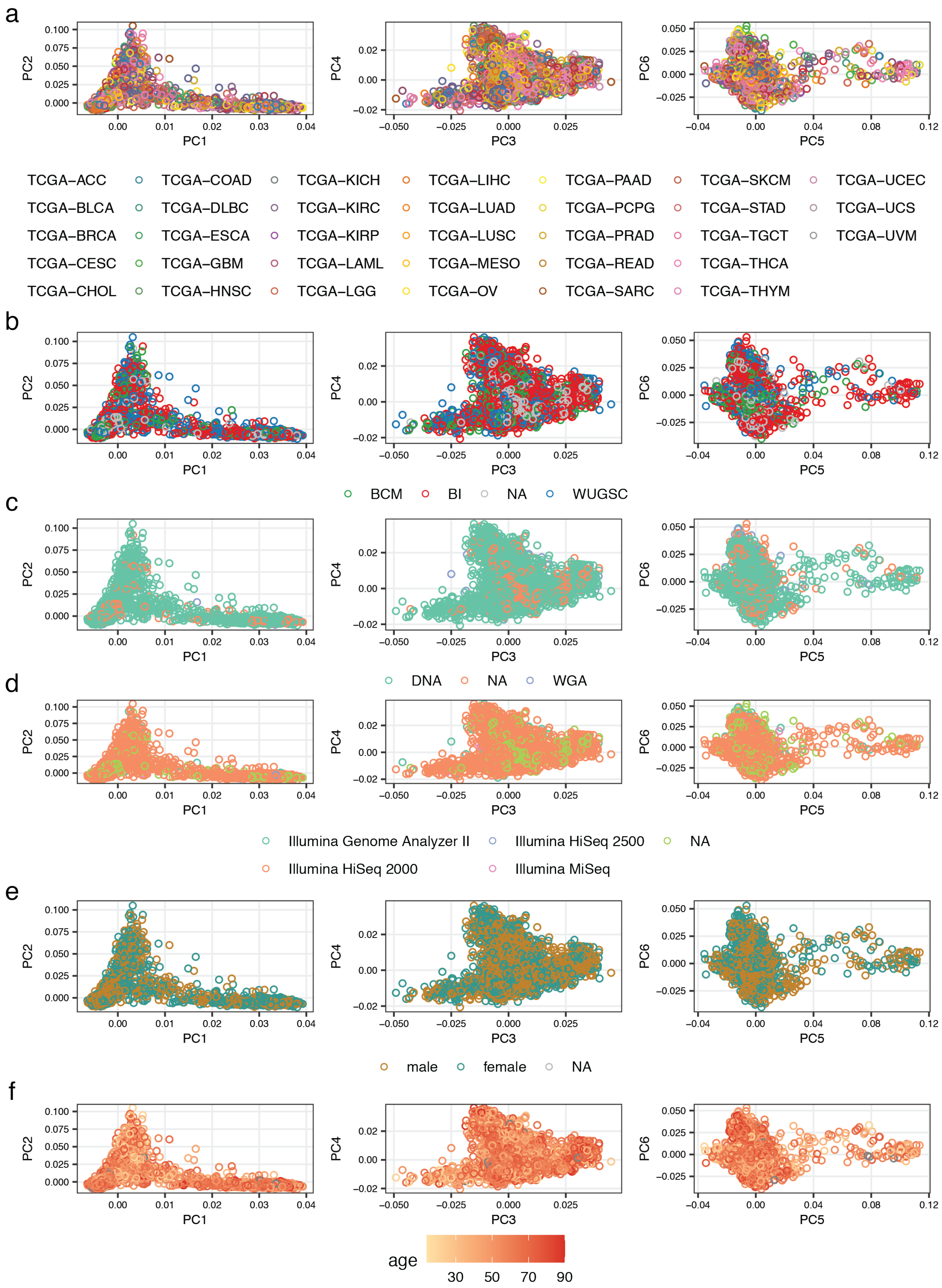


Figure 2.44: Principal component analysis on common germline variants in TCGA-WES. Principal components 1 to 6 color coded by (a) TCGA project id, (b) sequencing center, (c) whole genome amplification (WGA) status prior to sequencing, (d) sequencing system, (e) gender, and (f) age of diagnosis.

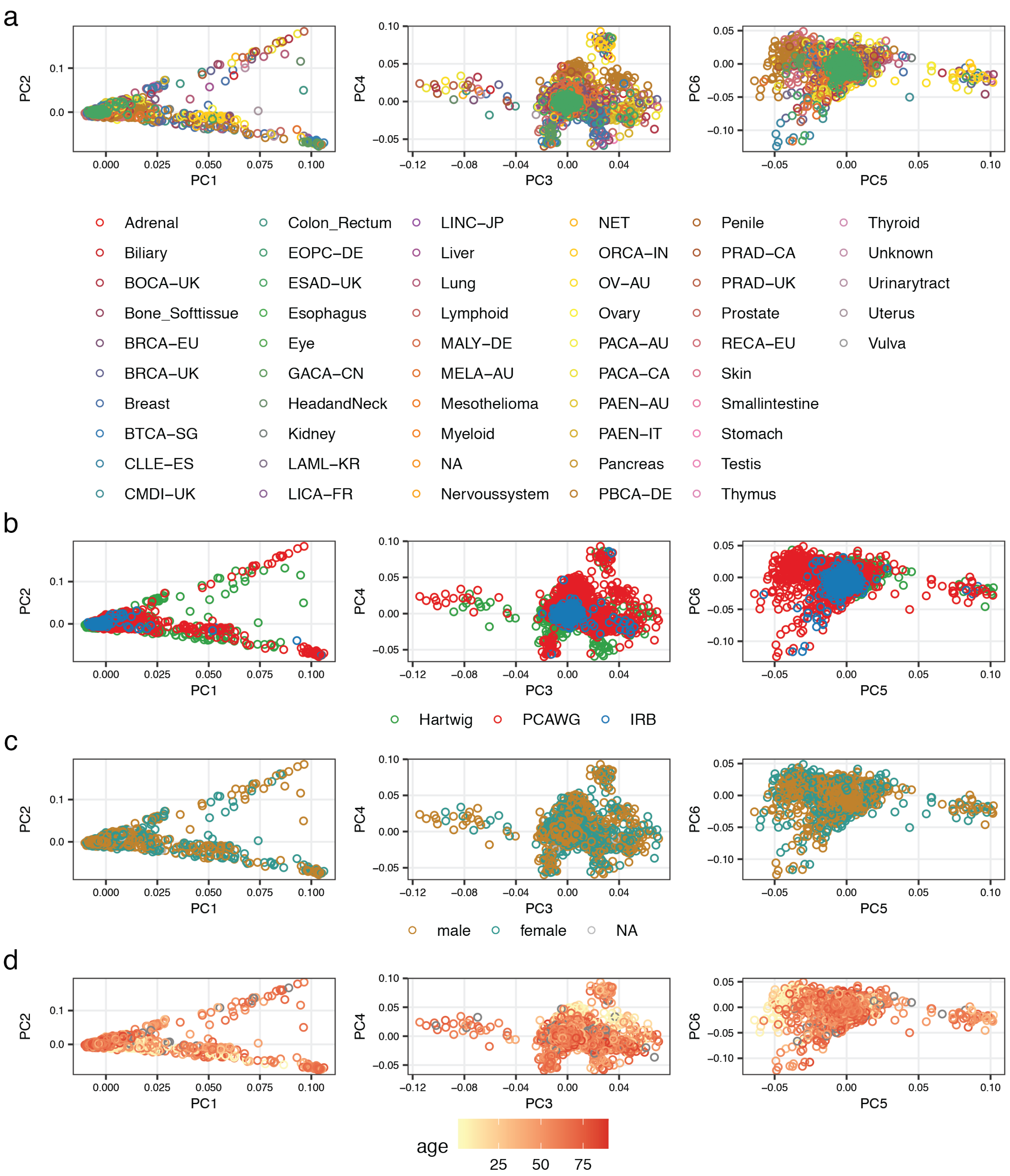


Figure 2.45: Principal component analysis on common germline variants in PCAWG_Hartwig -WGS. Principal components 1 to 6 color coded by (a) PCAWG project id or Hartwig tissue of origin, (b) center/study where germline variants were called, (c) gender, (d) age of diagnosis.

used in association testing as covariates to control for population substructure within the individuals of European ancestry.

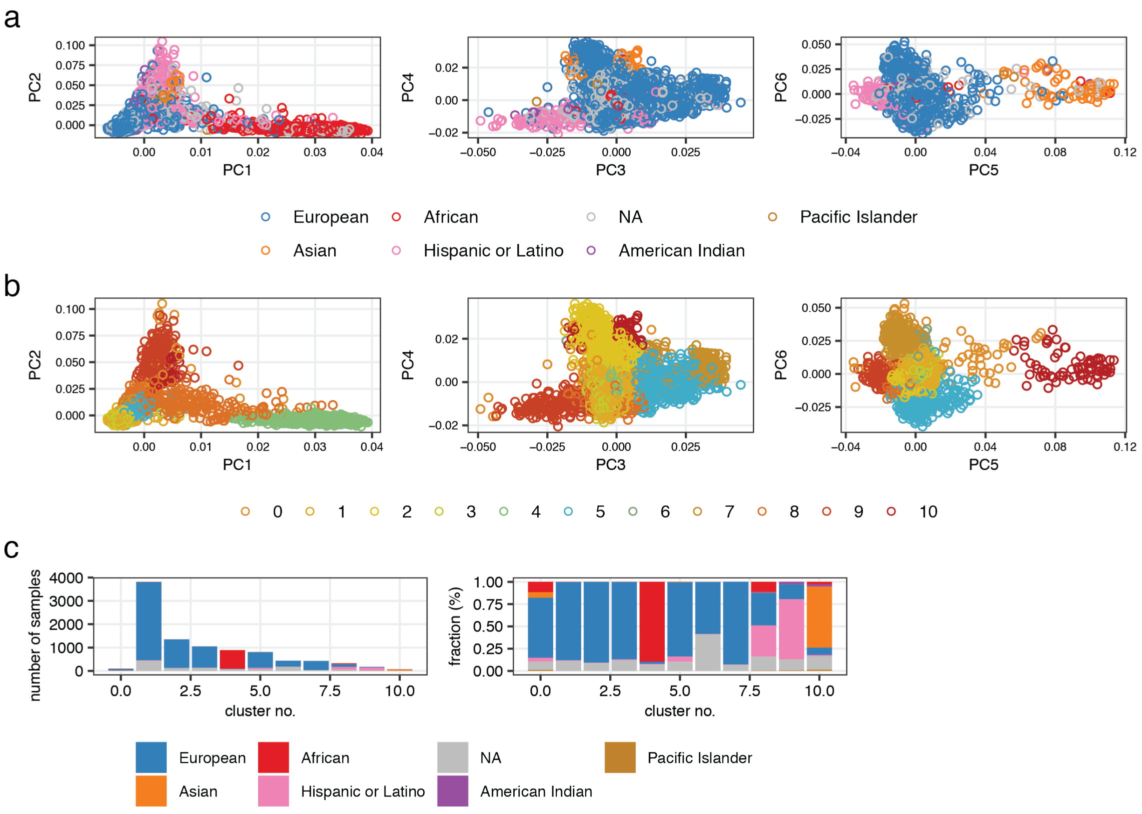


Figure 2.46: Extraction of European individuals in TCGA-WES. Principal components 1 to 6 color coded by (a) reported ethnicities and (b) clustering results using the first 10 principal components. (c) Overview of clustering results. Samples which could not be assigned to a cluster (cluster no. 0) were excluded.

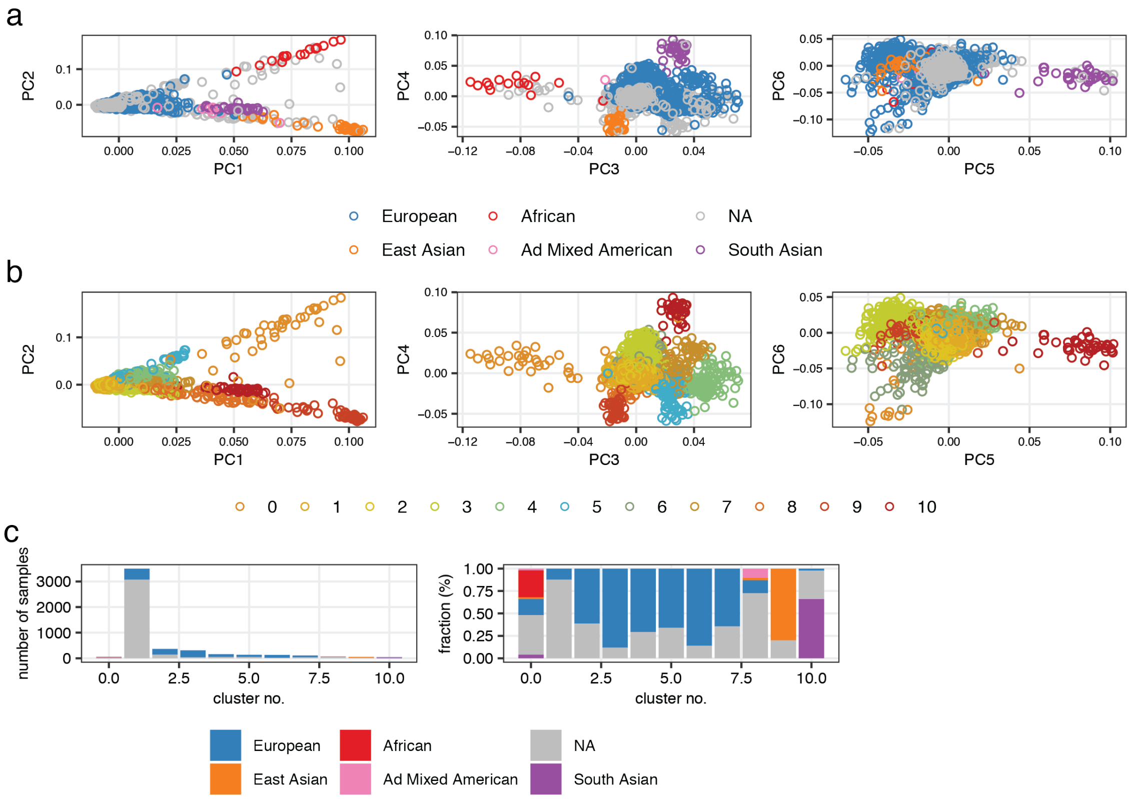


Figure 2.47: Extraction of European individuals in PCAWG_Hartwig -WGS. Principal components 1 to 6 color coded by (a) reported ethnicities and (b) clustering results using the first 10 principal components. (c) Overview of clustering results. Samples which could not be assigned to a cluster (cluster no. 0) were excluded.

2.2.2 Gene-Based Rare Variant Association Testing

Rare variant association with a combined burden and variance test

To identify genes with rare germline variants that impact somatic mutational processes, we defined five different sets of RDGVs using varying approaches and stringency criteria for identifying causal variants, and tested three models of inheritance by also considering RDGVs in combination with somatic loss-of-heterozygosity (LOH)[167]. In total, 15 different models were tested (Figure 2.48 top). To increase statistical power, we restricted testing to a set of 892 genes including known cancer predisposition genes, DNA repair genes and chromatin modifiers. The combined test SKAT-O[184], which unifies burden testing and the SKAT variance test[182,183], was utilized for testing (Figure 2.48 bottom). In brief, the test statistic in SKAT-O is the weighted sum of the test statistic from a burden test and a SKAT test. While in burden testing the variants are aggregated first and then jointly regressed against a phenotype, in SKAT the individual variants in a gene are regressed against the phenotype, and then the distribution of the individual variant score statistics is tested. Importantly, the burden test is more powerful when all RDGVs in a gene are causal, while SKAT is more powerful when some RDGVs are not causal or when RDGVs are causal but with effects in opposite directions[184]. In SKAT-O the parameter ρ controls the contribution of the two tests and corresponds to the smallest reported p-value[184], indicating whether the burden or the variance test was used to identify the association.

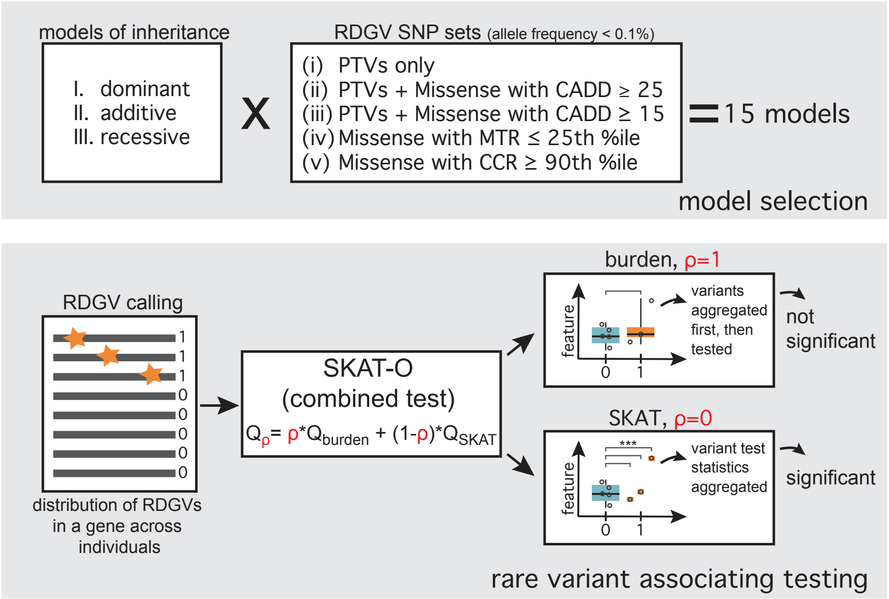

models of inheritance
I. dominant
II. additive
III. recessive
X
RDGV SNP sets (allele frequency < 0.1%)
(i) PTVs only
(ii) PTVs + Missense with CADD ≥ 25
(iii) PTVs + Missense with CADD ≥ 15
(iv) Missense with MTR ≤ 25th %ile
(v) Missense with CCR ≥ 90th %ile
= 15 models
model selection
RDGV calling
1
1
1
0
0
0
0
0
distribution of RDGVs in a gene across individuals
SKAT-O
(combined test)
Qρ= ρ*Qburden + (1-ρ)*QSKAT
burden, ρ=1
feature
0
1
variants aggregated first, then tested
not significant
SKAT, ρ=0
***
feature
0
1
variant test statistics aggregated
significant
rare variant associating testing

Figure 2.48: Rare damaging germline variant association testing study design. Associations were tested via 15 models in total, by utilizing 3 models of inheritance and 5 (differently prioritized) SNP sets of rare variants (all with population allele frequency $< 0.1\,\%$) (top). The combined test SKAT-O was applied, which calculates a weighted sum between a burden test statistic and the SKAT variance test statistic. When $\rho = 1$, the test reduces to a burden test, and when $\rho = 0$, the test reduces to the variance (SKAT) test. SKAT is more powerful when a fraction of the variants in the SNP set are non-causal, while the burden test has higher power when all variants are causal[184].

42 genes robustly associated with somatic mutation phenotypes

Testing was performed in the discovery cohort (TCGA) across 6,799 individuals of European ancestry and 12 different cancer types as well as in a pan-cancer analysis ('pancan') for all 15 models. Genes were only tested via the dominant or additive model when at least 2 individuals carried a RDGV in that gene. For the recessive model, genes were only tested when the gene was biallelically affected in at least 2 samples either by a biallelic RDGV or via a RDGV + LOH (see Methods). In total 594,462 tests were conducted. The tests showed little evidence of inflation when considering models in which at least 100 genes were tested. Overall there was slight deflation (median: 0.78; 1st quartile: 0.55; 3rd quartile: 0.97; max: 2.27) (Figure 2.49), suggesting conservatively biased test results. Inflated tests were discarded (cut-off at lambda $\geqq$ 1.5; 19 out of 1,909 discarded). We further estimated false discovery rates (FDRs) by randomization. The link between somatic components and individuals was broken by randomly shuffling the somatic component estimates of the individuals within each cancer type. Empirical FDRs were estimated by comparing the observed p-value distribution against the random one (Section 4.4.9 and Figure 2.50). As an additional negative control, we considered a random set of genes, comparing the number of replicated hits at a certain empirical FDR with the random gene set to the number with our candidate gene list (Figure 2.50). It should be noted that this yields a conservative upper limit since the random gene lists may also include genes which affect somatic mutation processes.

In total, we identified 6,488 associations (out of 591,302 tests) in the discovery phase at an empirical (randomization-based) FDR of 1 % (Figure 2.51a-d). Out of the 6,488 hits, 3,807 had a sufficient number of RDGVs in the matching cancer type (see Section 4.4.6) to allow re-testing in an independent validation cohort (merged PCAWG and Hartwig) in the matching cancer type, consisting of 4,683 patients of European ancestry. 207 associations replicated in the validation cohort at an empirical FDR of 1 %, covering 42 individual genes, 15 mutational components, 46 unique gene-cancer type pairs, and 65 unique gene-cancer type-component combinations (Figure 2.52). We also checked the number of replicated associations at a more permissive FDR of 2 %. At an FDR of 2 %, 12,480 hits were detected in the discovery cohort, 7,290 hits were re-tested in the validation cohort, out of which 356 associations were replicated covering 86 individual genes, 24 mutational components, 105 unique gene-cancer type pairs, and 140 unique gene-cancer type-component combinations (Figure 2.53). No-

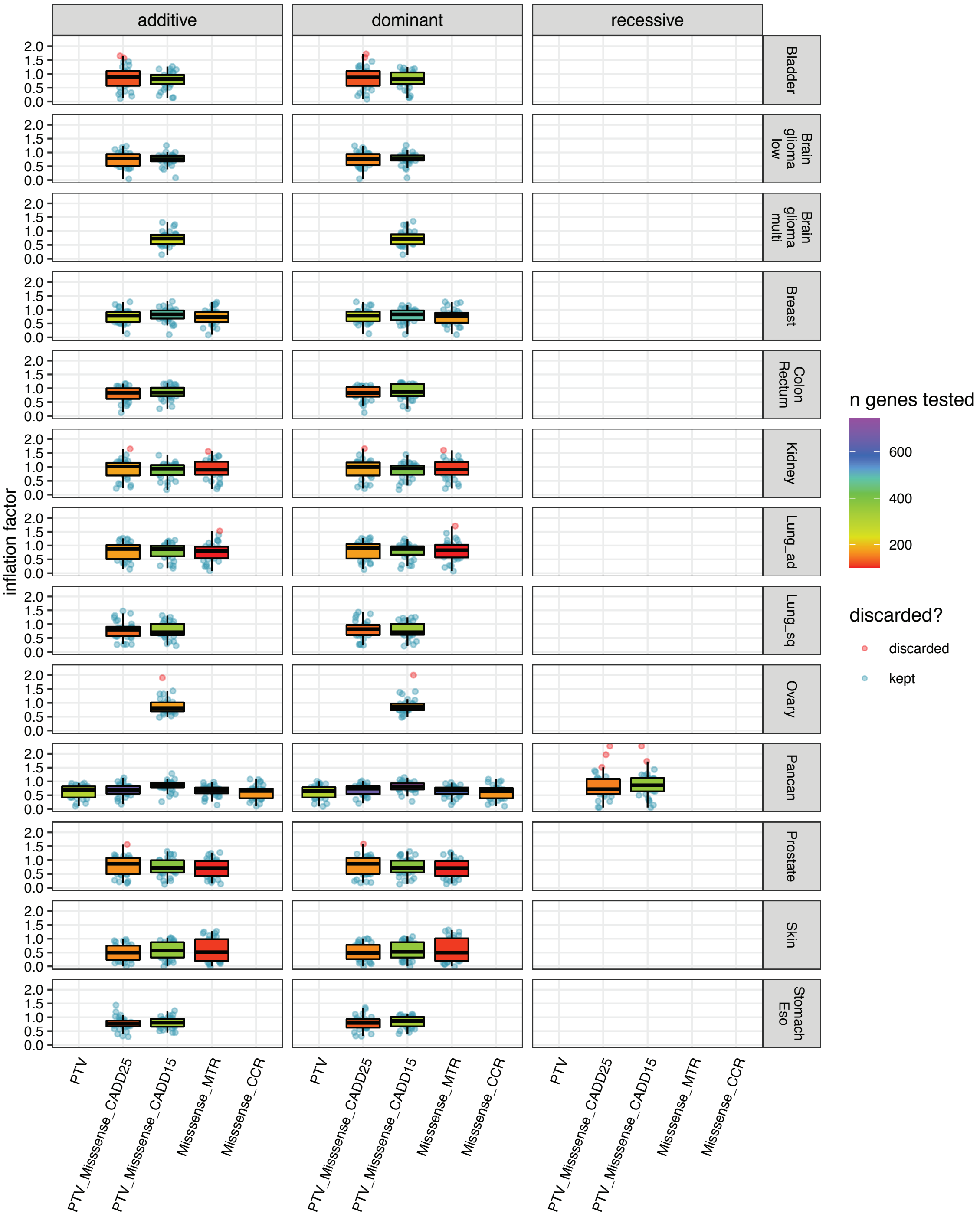


Figure 2.49: Inflation analysis. Overview of inflation factors (y-axis) across RDGV sets (x-axis), across cancer types (rows), and across models of inheritance (columns). Color code for box plots illustrates the number of tested genes for the respective scenario. Inflation factors were only calculated when at least 100 genes were tested, and inflation factors ≧ 1.5 were discarded (red point). Each data point represents the calculated inflation factor for one somatic component.

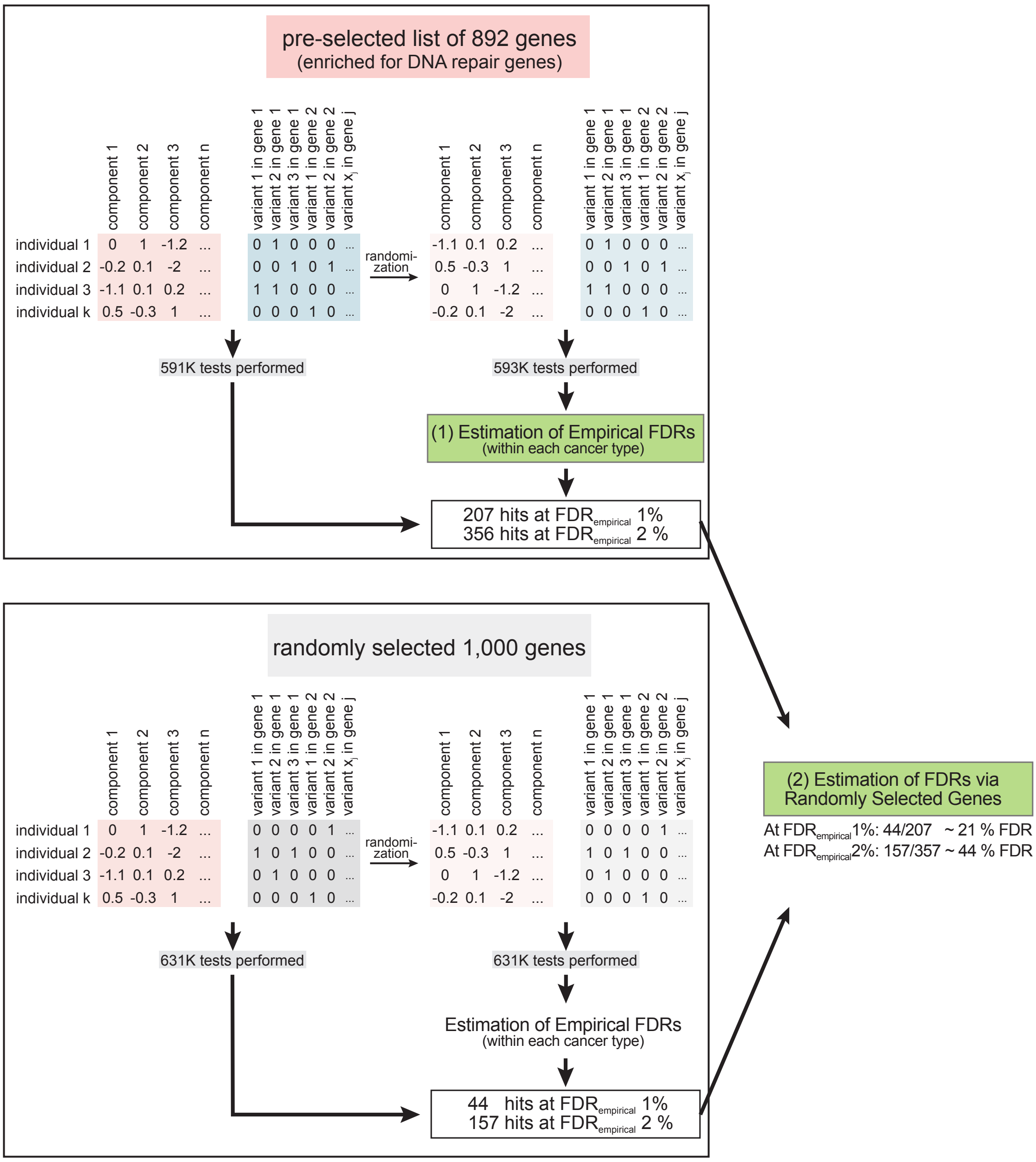

pre-selected list of 892 genes
(enriched for DNA repair genes)
component 1
component 2
component 3
component n
variant 1 in gene 1
variant 2 in gene 1
variant 3 in gene 1
variant 1 in gene 2
variant 2 in gene 2
variant x_j in gene j
individual 1
individual 2
individual 3
individual k
randomi-zation
591K tests performed
593K tests performed
(1) Estimation of Empirical FDRs
(within each cancer type)
207 hits at FDR_empirical 1%
356 hits at FDR_empirical 2 %
randomly selected 1,000 genes
631K tests performed
631K tests performed
Estimation of Empirical FDRs
(within each cancer type)
44 hits at FDR_empirical 1%
157 hits at FDR_empirical 2 %
(2) Estimation of FDRs via Randomly Selected Genes
At FDR_empirical 1%: 44/207 ~ 21 % FDR
At FDR_empirical 2%: 157/357 ~ 44 % FDR

Figure 2.50: Estimation of false discovery rates. Schematic illustration of the approach. Firstly, testing was performed using the pre-selected 874 genes. Then, randomization was performed shuffling the rows within cancer types, effectively breaking down the link between individuals and somatic components. Testing was performed with the randomized somatic component matrix as well and empirical FDRs were calculated based on the randomization for each cancer type (top half of plot). The same approach was repeated with a random set of 1,000 genes after excluding the pre-selected gene list and any gene interacting with a gene from the pre-selected gene list (bottom half of plot). The number of genes replicating via the randomly selected list of genes at a specific FDR was divided by the number of genes replicating with the pre-selected list to get a conservative FDR estimate.

tably, 7 genes associated across more than one cancer type, of which 3 (*BRCA1*, *EP300*, *MTOR*) associated with the same somatic mutational component across two different cancer types (Figure 2.54).

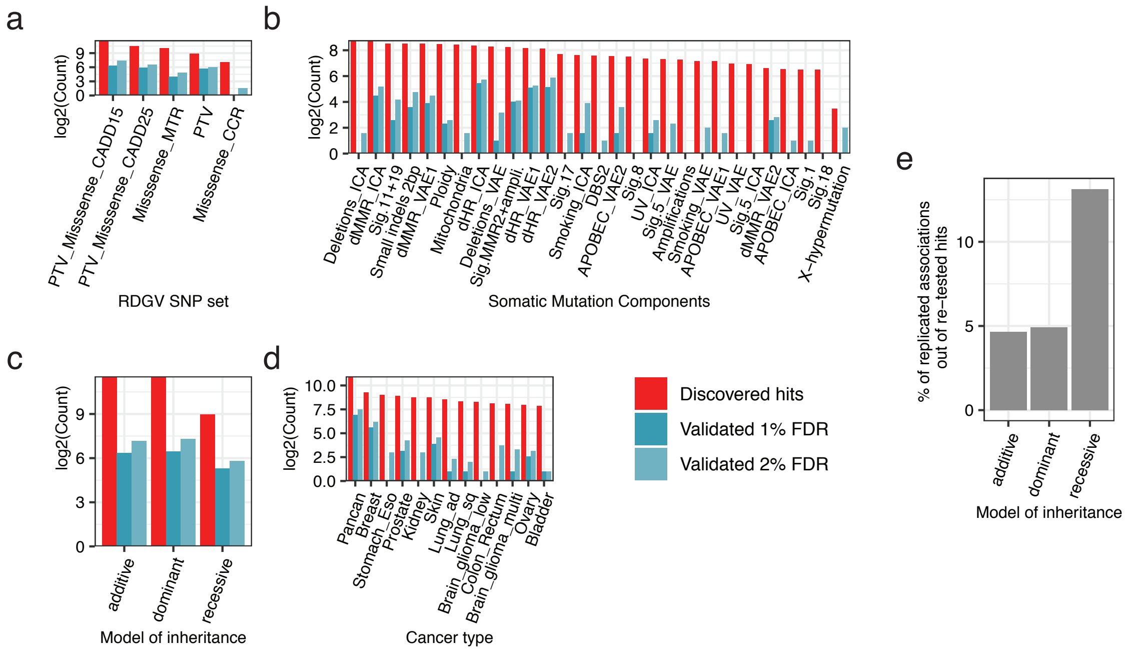


Figure 2.51: Overview of number of discovered and validated hits. (a) Number of discovered hits, number of replicated hits at a FDR of 1 % and number replicated hits at a FDR of 2 % across RDGV sets, (b) somatic components, (c) models of inheritance, and (d) cancer types. Log2 counts shown on the y-axis for panels a-d. (e) Amount of replicated hits out the re-tested discovered hits at a FDR of 1 % across different models of inheritance.

At an FDR of 1 %, most of the replicated hits were identified in the pan-cancer analysis (57 %), followed by breast cancer (24 %), skin cancer (7 %), and prostate cancer (4 %) (Figure 2.55a), reflecting differential sample sizes between cancer types (Figure 2.56). Furthermore, approximately half of the components (15 out of 29) were associated with at least one gene-cancer type pair (Figure 2.55b). Many replicated hits were associated with features related with dHR (dHR$_{ICA}$: 21 %, dHR$_{VAE1}$: 17 %; dHRVAE2: 16 %), followed by dMMR (dMMR$_{ICA}$: 11 %; dMMR$_{VAE1}$: 7 %), consistent with well-established roles of HR and MMR failures in accelerating mutation rates in tumors[31]. Notably, 25 genes were only identified via an ICA derived component, while 8 genes were only identified via a VAE derived component (Figure 2.55c), suggesting a complementary role of the two approaches to summarize mutation processes.

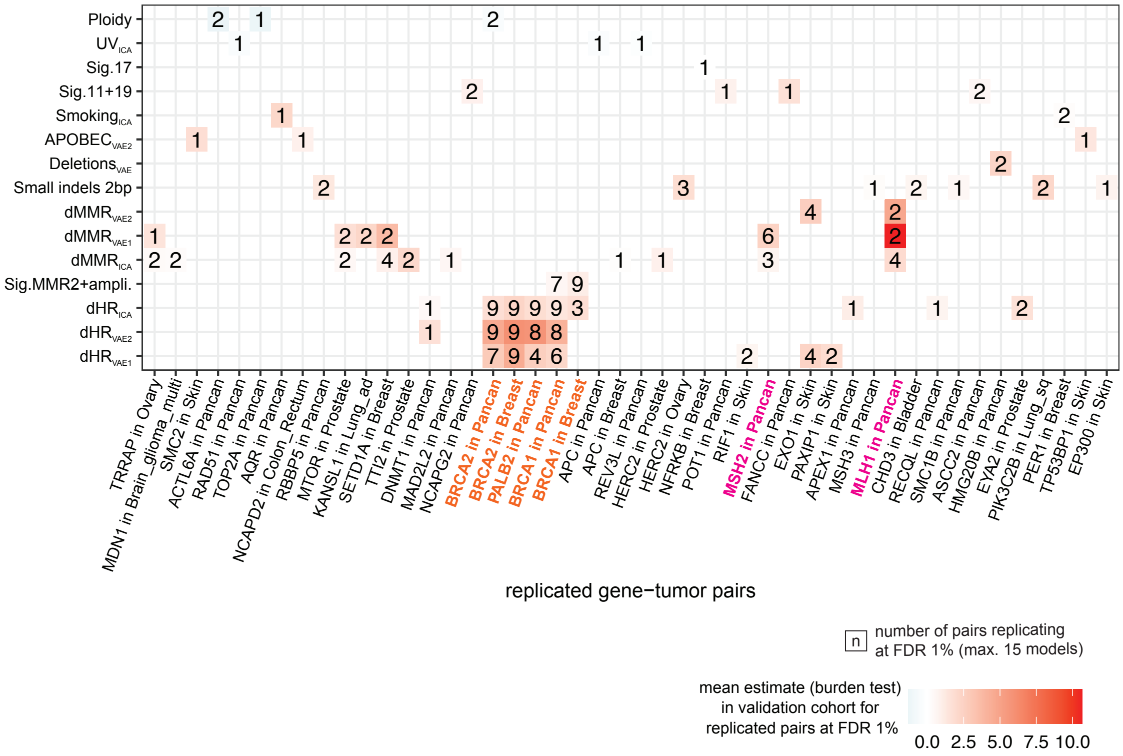


Figure 2.52: Overview of replicated hits at a FDR of 1 %. Showing gene-cancer type pairs (x-axis), the corresponding somatic mutational component (y-axis), and the number of times they replicated at a FDR of 1 % (maximum of 15 models for each gene-cancer type-somatic component tuple). Color represents the mean estimate of the regression coefficient from burden test for all replicated hits at a FDR of 1 %, for the respective gene-cancer type-somatic component combination. Previously associated dHR genes in orange and dMMR genes in pink. Genes on the x-axis were ordered based on hierarchical clustering results using DepMap[201] CRISPR-derived genetic fitness (Chronos[202]) scores.

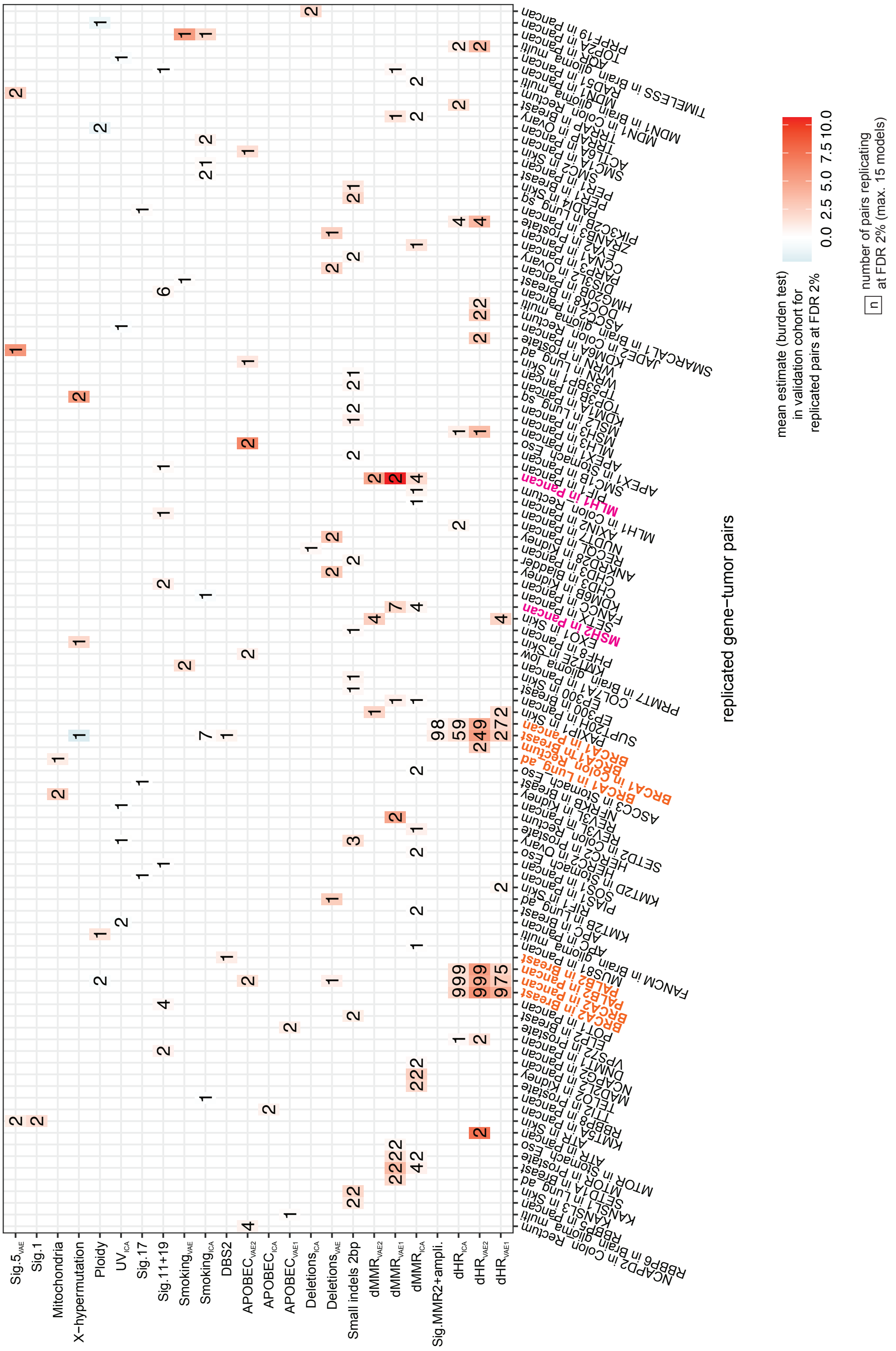


Figure 2.53: Overview of replicated hits at a FDR of 2 %. Showing gene-cancer type pairs (x-axis), the corresponding somatic component (y-axis), and the number of times they replicated at a FDR of 2 % (maximum of 15 models for each gene-cancer type-somatic component tuple). Color code represents the mean estimate of the regression coefficient from burden test for all replicated hits at a FDR of 2 % for the respective gene-cancer type-somatic component tuple. Previously associated dHR genes in orange and dMMR genes in pink. Genes on the x-axis were ordered based on hierarchical clustering results using DepMap[201] CRISPR-derived genetic fitness (Chronos[202]) scores.

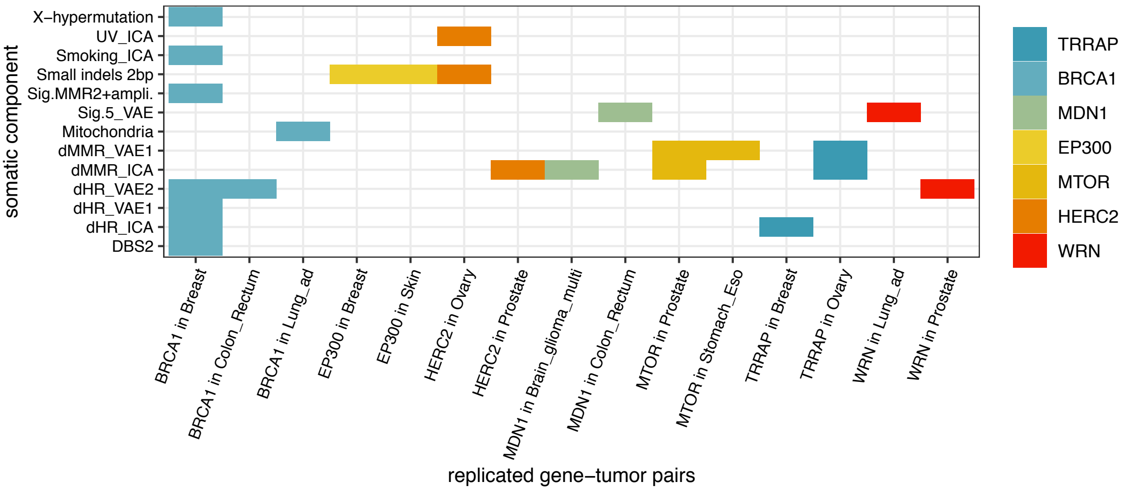


Figure 2.54: Seven genes associated with a somatic component in $\geqq$ 1 cancer type. Showing gene-cancer type pairs (x-axis) and the corresponding somatic component they associated with at a FDR of 2 % (y-axis). Color code for each gene. Results from pan-can analysis excluded.

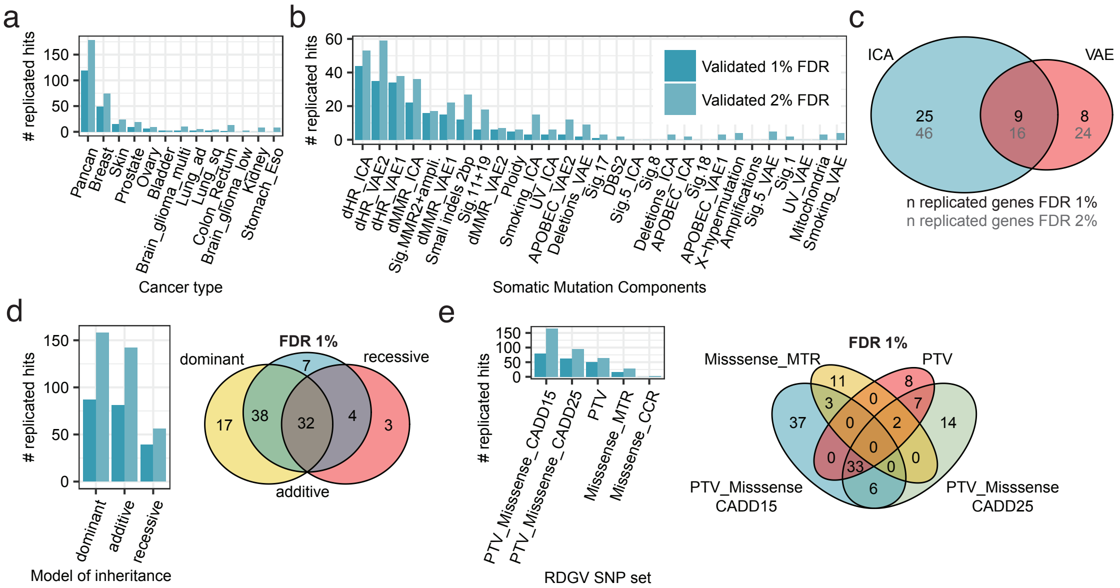


Figure 2.55: Discovery and validation of rare damaging germline variants (RDGVs) associating with somatic components via a gene-based combined burden and variance test. (a) Number of replicated hits at a FDR of 1 % and 2 % across cancer types and (b), across somatic mutational components. (c) Overlap of number of genes replicating at a FDR of % and 2 % via the two different dimensionality reduction methods. (d) Number of replicated hits at a FDR of 1 % and 2 % across models of inheritance (left) and overlap of replicated hits between models at a FDR of 1 % (right). (e) Number of replicated hits at a FDR of 1 % and 2 % across RDGVs sets (left) and overlap of replicated hits between RDGV sets at a FDR of 1 % (right).

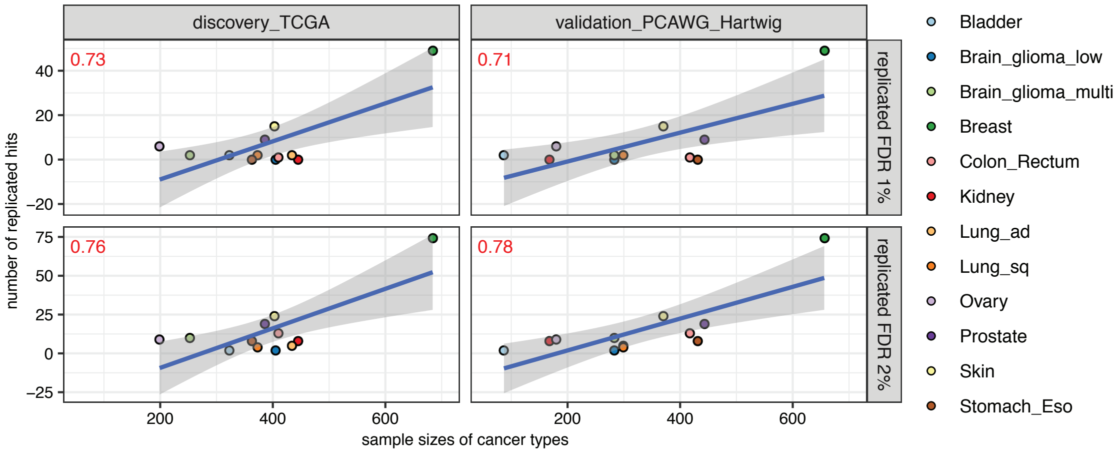


Figure 2.56: Number of replicated hits increased with sample size. Number of replicated hits (y-axis) versus sample sizes of the corresponding cancer types in which they replicated (x-axis). Columns represent the two cohorts, and rows the applied FDR. Color code for the different cancer types. Pearson correlation shown on the top left corner in red and linear regression fitted through each plot (blue line). Shaded band illustrating 95 % confidence interval. Pan-can analysis was excluded.

We further tested three different models of inheritance for association testing (Figure 2.48 top). While we only considered RDGVs for the dominant model, we also tested to which extent an inactivation of a gene on both alleles would affect a somatic mutational component via the additive and recessive model (Section 4.4.6). Many of the replicated associations were identified via the dominant (42 %) and (39 %) additive model (Figure 2.55d), suggesting that heterozygous variants can alter mutation rates in humans, as it was suggested in a yeast screen[196]. The comparatively lower number of replicated hits of the recessive model can be largely attributed to the fact that RDGV combined with somatic LOH events are considerably less frequent and thus associations could not be tested for many genes (only 4 % of the 591,302 tests performed in the discovery phase came from the recessive model). Considering the proportion of replicated hits to the number of re-tested hits, the validation rate was ~2.5 times higher via the recessive model (Figure 2.51e), which was expected since many DNA repair genes are believed to be haplosufficient[200].

We further considered the number of replicated associations using different approaches and stringency thresholds for declaring a variant to be pathogenic. The highest number of hits replicated using the more permissive thresholds, using protein-truncating variants (PTVs) + missense variants at a CADD[185] score ≧ 15 (79/207, 38 %), followed by PTVs + missense variants at a CADD score ≧

25 (62/207, 30 %) and PTVs only (50/207, 24 %) (Fig. 2g). This suggests that some missense variants that were assigned a lower pathogenicity score - likely due to difficulties in assessing variant pathogenicity *in silico*[203] - can nonetheless bear on somatic mutation phenotypes. We further tested by only considering RDGVs in conserved gene segments (via "constrained coding regions"[272] and "missense tolerance ratio"[271] methods), however this yielded few replicated hits (Figure 2.51e). It should be noted, however, that some hits were only identified when using the PTV-only set and were not recovered in more permissive RDGV sets, suggesting that very few missense RDGVs in those genes are causal.

In summary, with regards to the model of inheritance, RDGV set, and component (mutational process) extraction method, there was no single best model and most models added unique associations to the results.

More permissive thresholds for variant pathogenicity increase the utility of a variance test over a burden test

The SKAT-O test we employed combines burden testing and a variance test component (SKAT)[184]. The contribution of each test to the total test statistic (Figure 2.48 bottom) is controlled via the parameter ρ. When $\rho = 0$ the test becomes a variance test (SKAT) and when $\rho = 1$ the test becomes a burden test. Examining the SKAT-O parameter ρ for the 207 validated hits, in both the discovery and the validation cohort, revealed that most hits replicated via the variance test ($\rho < 0.5$ in 393/414 tests) (Figure 2.57a). The variance test is the more powerful test of the two when many variants in the tested set are not causal[184]. We hypothesized that a common reason why allegedly pathogenic RDGVs would not be causal is because of inaccurate prediction of damaging variants by *in silico* predictors[204]. If so, at the more stringent settings more hits would replicate via the burden test (which has higher power when many variants in the set are causal), while at the less stringent settings more hits would replicate via the variance test (which is robust to inclusion of non-causal variants). Indeed, several hits replicated via the burden test when using the most stringent RDGV set (PTVs only; Figure 2.57b), including *MLH1*, *BRCA1*, and *BRCA2*. For the more permissive RDGV sets, the number of hits replicating via the burden test decreased and all of the replicated hits had a ρ lower than 0.25 (meaning, they used nearly exclusively the variance component) for the RDGV set including missense variants at CADD $\geqq$ 15. The positive control genes

BRCA1 and *BRCA2* still replicated in the PTV+missense CADD $\geqq$ 15 RDGV set, but with a ρ of 0 (variance test exclusively used), suggesting that this variant set included many non-causal variants.

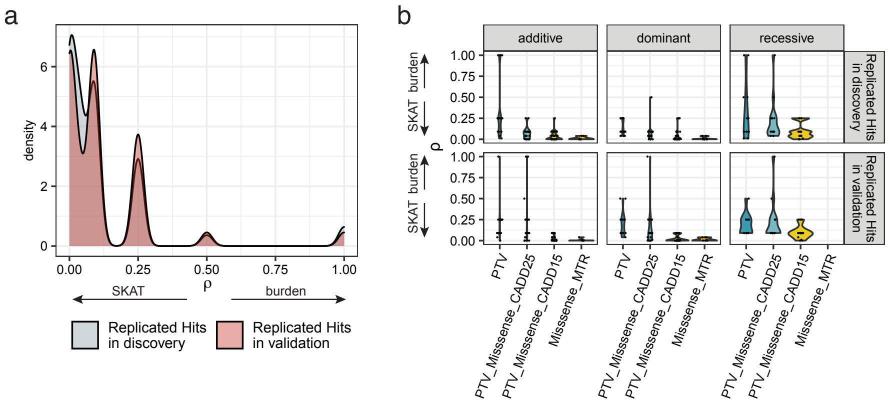


Figure 2.57: Hits mostly replicating via variance (SKAT) test. (a) Distribution of ρ values from SKAT-O test (x-axis) for the 207 hits, which replicated at a FDR of 1 %, in the discovery (grey) and validation cohort (red). (b) Distribution of ρ values from SKAT-O test (y-axis) for the 207 hits, which replicated at a FDR of 1 %, in the discovery (top row) and validation cohort (bottom row), across models of inheritance (columns) and RDGV sets (x-axis).

In summary, many hits were recovered even with more permissive RDGV sets by utilizing the combined testing approach of the SKAT-O method, suggesting the variance (SKAT) component can partially compensate for the inaccuracy of the *in silico* predictors. Most of the replicated hits would not have been identified by use of classical burden testing in a data set of this size.

Novel genes associating with defects in homologous recombination

Within the set of 207 replicated associations at an FDR of 1 %, 117 (57 %) involved associations of *BRCA1*, *BRCA2*, and *PALB2* with various mutational components associated with dHR (Figure 2.52), consistent with the known role of these genes in the error-free repair of DSBs. All three genes associated with features of defective HR, such as deletions at microhomology-flanked sites (dHR$_{ICA}$ and dHR$_{VAE2}$) and SNV signature 3 mutations (dHR$_{VAE1}$). In addition, *BRCA1*, but not *BRCA2*, associated with component Sig.MMR2+ampli., reflecting an increased number of amplification events. This is in accordance with a recent report, in which BRCA1-type dHR vs. BRCA2-type dHR were differentiated via the presence of duplication events[95].

We also detected additional genes associating with these dHR mutational components. In skin cancer, *PAXIP1*, *EXO1*, and *RIF1* associated with dHR$_{VAE1}$, the component correlating with SNV signature 3 mutations. In support of this, *PAXIP1* and *RIF1* have been implicated in the repair of DNA DSBs[205,206,207] and interact with each other[208]. Thus, these associations suggest that individuals carrying damaging variants in either gene have an increase in signature 3 mutations, potentially reflecting a downstream effect of disrupted DSB repair. Additionally, *EXO1* knockout in a cell line model[126] was reported to result in a mutational signature correlating with signatures 3 (Pearson R = 0.71) and 5 (R = 0.71), supporting our association observed in tumors.

Furthermore, we identified pan-cancer replicated associations of *APEX1*, *RECQL*, and *DNMT1* with dHR$_{ICA}$ (with *DNMT1* additionally associating with dHR$_{VAE2}$). These associations with a microhomology deletion mutation phenotype are diagnostic of an increased activity of the microhomology-mediated end joining (MMEJ), a highly error-prone DSB repair pathway, suggesting that variants in these genes may disrupt normal functioning of the less error-prone HR and/or NHEJ pathways.

Five additional genes (*ATR*, *JADE2*, *SMARCAL1*, *TIMELESS*, and *WRN*) were identified at a more permissive threshold, associating with at least one dHR-related component (dHR$_{ICA}$ and/or dHR$_{VAE2}$). Notably, ATR and WRN physically interact with BRCA1 (Figure 2.58) and play known roles in repair of DSBs[209,210,211], which would support these associations. In particular, pathogenic recessive variants in *WRN* cause Werner syndrome[212] and it has been suggested that the WRN helicase is crucial for the repair of MMR-induced DSBs (MMR-failure can induce DSBs at AT repeats)[213,214]. Additionally, SMARCAL1 and TIMELESS directly interact with ATR (Figure 2.58).

Our analyses therefore replicate well-known associations between rare inherited variants in HR genes and somatic mutational components, as well as identifying new associations with additional genes.

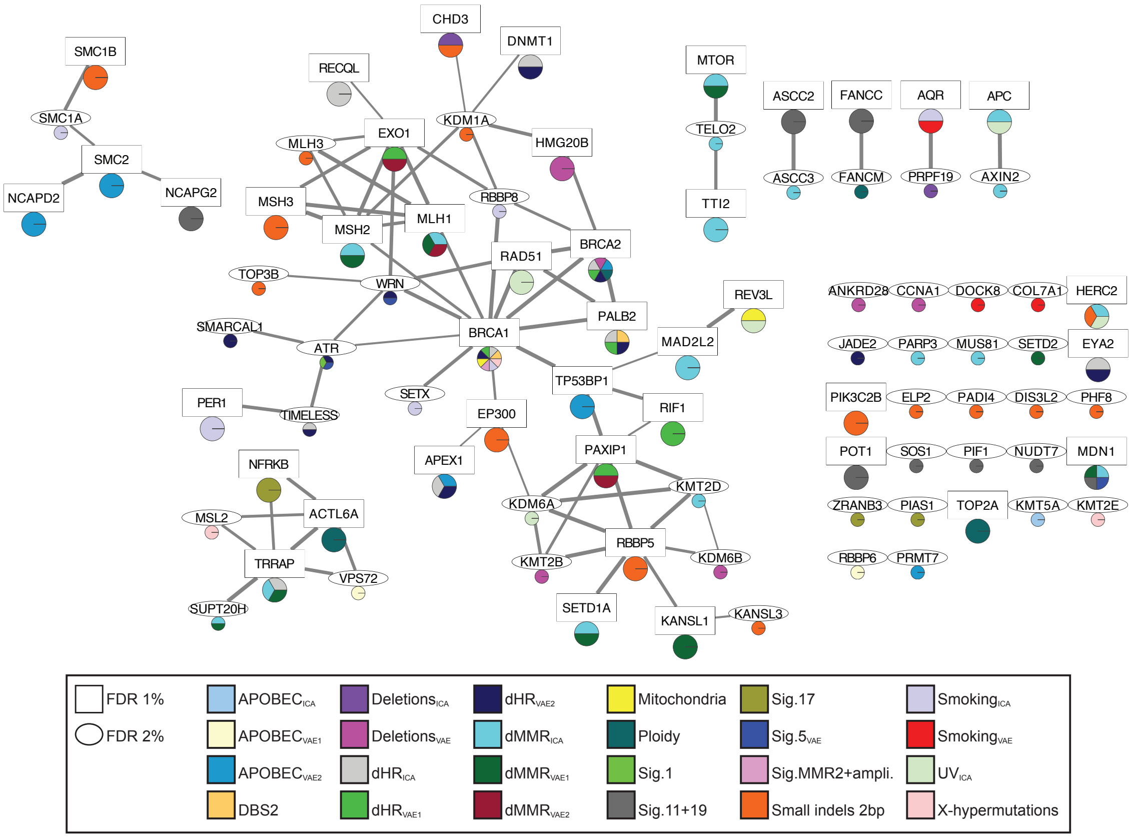


Figure 2.58: Overview of replicated hits in STRING network. Visualisation of physical interactions between proteins for genes replicating at a FDR of 1 % (square) and genes replicating at a FDR of 2 % (ellipse). Color code in pie chart shows the somatic components the corresponding gene was associated with (bottom panel). Line width corresponds to combined (experimental, database, and text mining) STRING physical interaction score.

MTOR and interacting protein variants associate with mismatch repair phenotypes

In the context of Lynch syndrome, germline variants in *MLH1*, *MSH2*, *MSH6* and *PMS2*[215] affect somatic mutation patterns via impairment of the DNA mismatch repair pathway, observed as microsatellite instability (MSI, indels at simple DNA repeats)[62,63]. MSI was also later associated with mutational signatures derived from SNVs[6], as well as with a 'redistribution' of mutations across replication timing domains[21]. In accordance with this, we detected associations of RDGVs in *MLH1* and *MSH2* with multiple dMMR-related components i.e. those having a high contribution of small indels at microsatellite regions (dMMR$_{ICA}$ and dMMR$_{VAE1}$), and with SNV-derived signature MMR1 mutations and replication timing (dMMR$_{VAE2}$; for *MLH1*; Figure 2.52).

Beyond the known Lynch Syndrome genes, we also discovered associations between variation in *EXO1*, which has an established role in MMR[216] and increases the frequency of 1 bp indels when inactivated in cultured cells[126], and dMMR$_{VAE1}$ and dMMR$_{VAE2}$. However, *EXO1* also associated with dHR-related components, suggesting a more pleiotropic role for the encoded exonuclease in shaping somatic mutational processes in human tumors. Consistent with the association with dHR components, it was reported in yeast as well as human cell lines that EXO1 processes DSB ends[217] and is required for the repair of DSBs via HR[218].

Multiple other genes were associated with dMMR-directed phenotypes (all associated with dMMR$_{ICA}$ and dMMR$_{VAE1}$), including the chromatin modifying enzyme genes *TRAAP* in ovarian and *SETD1A* in breast, and the major growth signalling gene *MTOR* in prostate cancer (and in stomach+esophagus cancer with dMMR$_{VAE1}$ only at a FDR of 2 %). Additionally, *TTI2* in prostate, *APC* in breast, *MAD2L2* in pan-cancer, *HERC2* in prostate, and *MDN1* in brain cancer associated with mutation component dMMR$_{ICA}$. There is additional evidence supporting these associations for some of these genes from prior studies. *MTOR* was identified as one of four genes that regulate MSH2 protein stability[278]. Thus, a possible mechanism explaining the identified association of *MTOR* with dMMR-linked components could be a decreased stability of MSH2 leading to dMMR and consequently, an increased number of indels. A similar mechanism could be speculated for TTI2, which binds MTOR via the TTT complex (TELO2-TTI1-TTI2) and is important for mTOR maturation[219]. This hypothesis is further supported by *TELO2* associating with the same component (dMMR$_{ICA}$) in kidney cancer at a more permissive FDR of 2 % (Figure 2.53). Furthermore, *SETD2* associated in colorectal cancer with dMMR$_{VAE1}$ at a FDR of 2 %. It has been shown in previous studies[20], including in cancer genomes[15], that the encoded methyltransferase SETD2 regulates MMR activity by recruiting the MSH2-MSH6 complex to H3K36me3 marked regions.

Taken together, we recovered known associations of MMR genes with somatic mutational patterns and identified additional genes where germline variants are associated with MMR phenotypes, suggesting that a broad network of genes cooperates to maintain MMR efficiency in human cells.

MSH3 and additional genes associate with a distinct dMMR phenotype

Interestingly, we identified associations between RDGVs in several genes and a somatic mutational component (Small indels 2 bp) that reflects indels of a size of 2 bp and longer, which is in contrast to the predominantly 1 bp long indels caused by dMMR. Furthermore, this component does not have any contribution from SNV features, indicating that it is specifically capturing indels (Figure 2.40 and 2.28). Among others, the MMR gene *MSH3* associated with this component in the pan-cancer analysis. In contrast to the DNA mismatch repair genes *PMS2*, *MLH1*, *MSH2*, and *MSH6*, germline variants in *MSH3* have not been identified in patients with Lynch syndrome, even though they were reported to increase cancer risk[176]. The MSH2-MSH3 complex has a role in repairing insertion/deletion loops rather than for base-base mismatches[52, 220, 221]. This is in contrast to the MSH2-MSH6 complex, which repairs base-base mismatches and indels shorter than 2 nucleotides[50, 51]. These prior mechanistic studies support our association and suggest that loss of MSH3 in cancer cells results in an increased rate of accumulation of indels of 2 bp and longer. Other genes associating with this component were *CHD3* in bladder cancer, *HERC2* in ovary cancer, *PIK3C2B* in lung squamous cell cancer, *EP300* in skin cancer (and breast cancer at a FDR of 2 %), *RBBP5* in pan-can, and *SMC1B* in pan-can. Additionally, *MLH3* associated with the same component at an FDR of 2 %. The *MLH3* protein is a paralog of *MLH1* that interacts with other MMR proteins (Figure 2.58) and was previously associated with microsatellite instability[222].

Overall, we detected associations between germline variants in *MSH3* and several other genes and somatic indels of at least 2 bp, suggesting a causal role for *MSH3* variants in a specific subtype of MMR failure which does not markedly increase SNV rates.

Genes associating with a somatic feature enriched in brain and liver cancer

Beyond the dHR and dMMR-related components, the component associated with the largest number of genes was component Sig.11+19, which is enriched for SNV signatures Ref.Sig. 11 and 19[138] (Figure 2.55b). This component is enriched in brain and liver cancers (Figure 2.30). Signature 11 has been reported to be enriched in brain cancers, associated with temozolomide treatment[6], and is similar to the signature which results from the treatment with the DNA methylating agent 1,2-Dimethylhydrazine[223]. The cause of Ref.Sig. signature 19 is unknown and it has been mostly identified in brain, liver and blood cancers[138]. At a FDR of 1 %, the genes *ASCC2*, *FANCC*, *NCAPG2* and *POT1* associated with this component in the pan-cancer analysis, as do *NUDT7*, *PIF1*, and *SOS1* at a more permissive 2 % FDR. *POT1* and *PIF1* interact with each other[280] (Figure 2.62) and both have functions in telomere maintenance[225,226], but we did not detect any correlation between this component and reported telomere features[227] (Figure 2.59).

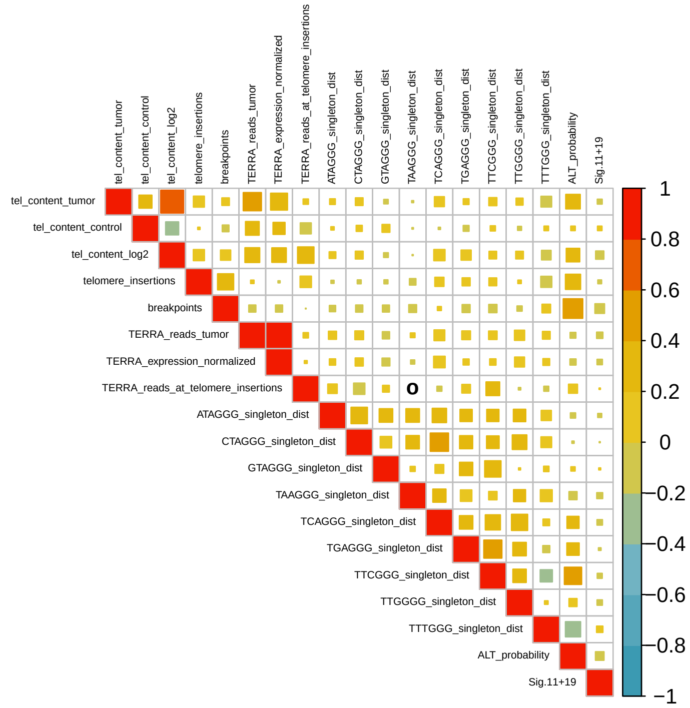


Figure 2.59: Pearson correlation between component Sig.11+19 (IC9) and telomere features. Correlations estimated based on 1,254 samples from PCAWG. Telomere features were downloaded from ref[227].

Variants in APEX1 associate with increased level of APOBEC-directed mutagenesis

We discovered and replicated associations between *APEX1* and three different somatic components. *APEX1* encodes for a purinic/apyrimidinic (AP) endonuclease that cleaves at abasic sites, which can be formed spontaneously or during base excision repair pathway by a DNA glycosylase[73]. At a FDR of 1 %, *APEX1* associated with dHR$_{ICA}$ in pan-can (Figure 2.52), and at a FDR of 2 % it associated with dHR$_{VAE2}$ in pan-can and with APOBEC$_{VAE2}$ in stomach/esophagus cancer (Figure 2.53). The somatic components dHR$_{ICA}$ and dHR$_{VAE2}$ are enriched for deletions at microhomology-flanked regions. Prior studies showed that the encoded protein APE1 protein plays a role in the repair of DSBs and that depletion of APE1 leads to an decrease of HR-directed repair[228], suggesting a higher reliance on alternative pathways.

The APOBEC$_{VAE2}$ component is enriched for SNV signature 13 (C>G) mutations[6]. These can be formed when the APOBEC-induced uracil is excised via the uracil-DNA glycosylase UNG and a cytosine is inserted opposite the abasic site by the mutagenic translesion polymerase REV1[113]. Conceivably, a mechanism underlying the higher burden of C>G mutations in tumors of individuals with inherited damaging variants in *APEX1* could be due to a decreased activity leading to a slower repair of the abasic site and consequently, a preference for lesion bypass via the error-prone REV1.

Network analysis reinforces the role of rare germline variants in somatic mutation processes

The previously known dHR genes encode proteins that physically interact as part of the same protein complexes[229]. Similarly, the products of the known dMMR genes also physically interact[229]. We used protein-protein interactions curated in the STRING[279] database to test whether the genes identified as having rare germline variants associating with somatic mutational phenotypes also encode physically interacting proteins. Such 'guilt by association' network analysis has been used to support associations between somatic mutations and cancer[231,232] and between common variants and disease phenotypes[233] but has not yet been widely adopted for the analysis of rare variants.

We first considered genes associated with somatic mutation phenotypes at a FDR of 1 %. These genes are strongly enriched for encoding proteins with physical interactions (Figure 2.60a; median difference in interactions between observed value and randomization = 17 and P = 0.002 by randomisation, controlling for interaction node degree). This also held true after removing genes with previously reported associations between RDGVs and somatic mutational processes (Figure 2.60c; median difference = 7 and P = 0.032 by randomisation). Secondly, we considered the 44 genes with moderate statistical support of association with somatic mutation phenotypes (those replicating at a FDR of 2 %). 21 of the encoded proteins interact with at least one of the proteins encoded by the more stringent FDR 1 % genes. This is again higher than expected by chance (Figure 2.60b; median difference = 6 and P = 0.021 by randomisation), further prioritising these 21 genes for additional study. This also held true after removing previously known genes (Figure 2.60d; median difference = 5 and P = 0.033 by randomisation). Similar results were seen using the HumanNet gene network[280] that incorporates many data sources to predict functionally-related genes (Figures 2.61 and 2.62).

Thus, genes with replicated associations with somatic mutation phenotypes preferentially encode proteins that physically interact in cellular networks with genes replicating at a more permissive FDR also often connected to the same sub-networks, illustrating the potential for network-based analyses to provide supporting evidence in rare variant association studies.

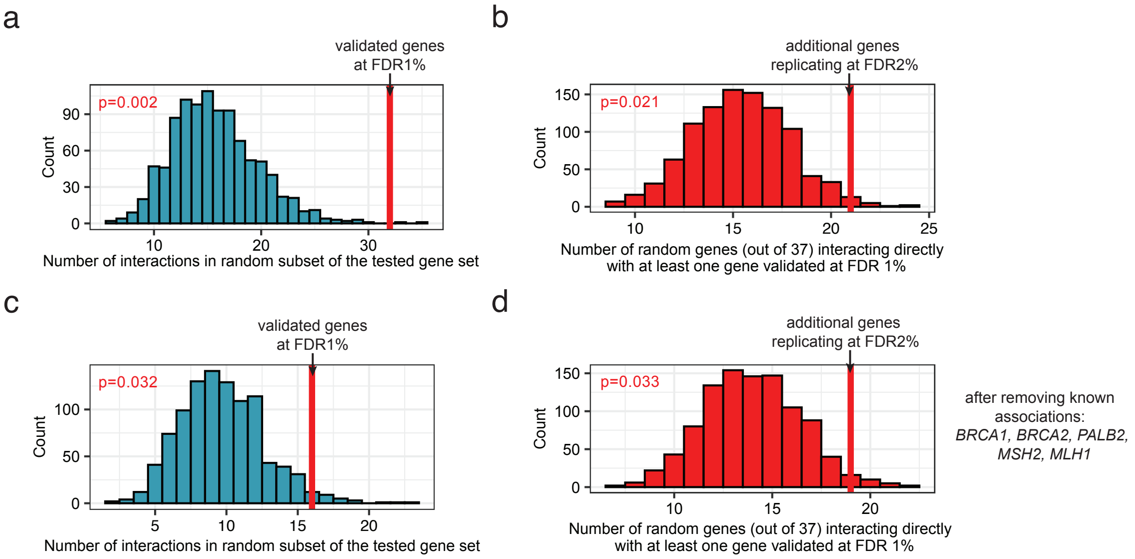


Figure 2.60: Network analysis supports the role of rare germline variation in somatic mutational processes. All panels in this figure were generated using physical interactions from the STRING database having a combined score $\geqq 80\,\%$. (a) Number of physical interactions in a random subset of the tested gene set (controlled for interaction node degree) (x-axis). Red line shows the number of interactions within genes which replicated at a FDR of 1 %. (b) Number of randomly selected genes from the tested gene set interacting with at least one gene, which replicated at a FDR of 1 % (x-axis), (controlled for interaction node degree). Red line shows the number of genes, out of the ones which additionally replicated at a FDR of 2 %, interacting with at least one gene replicating at a FDR of 1 %. (c) Same as in panel a, after excluding known genes from the analysis (*BRCA1*, *BRCA2*, *PALB2*, *MSH2*, and *MLH1*). (d) Same as in b after excluding known genes from the analysis (*BRCA1*, *BRCA2*, *PALB2*, *MSH2*, and *MLH1*).

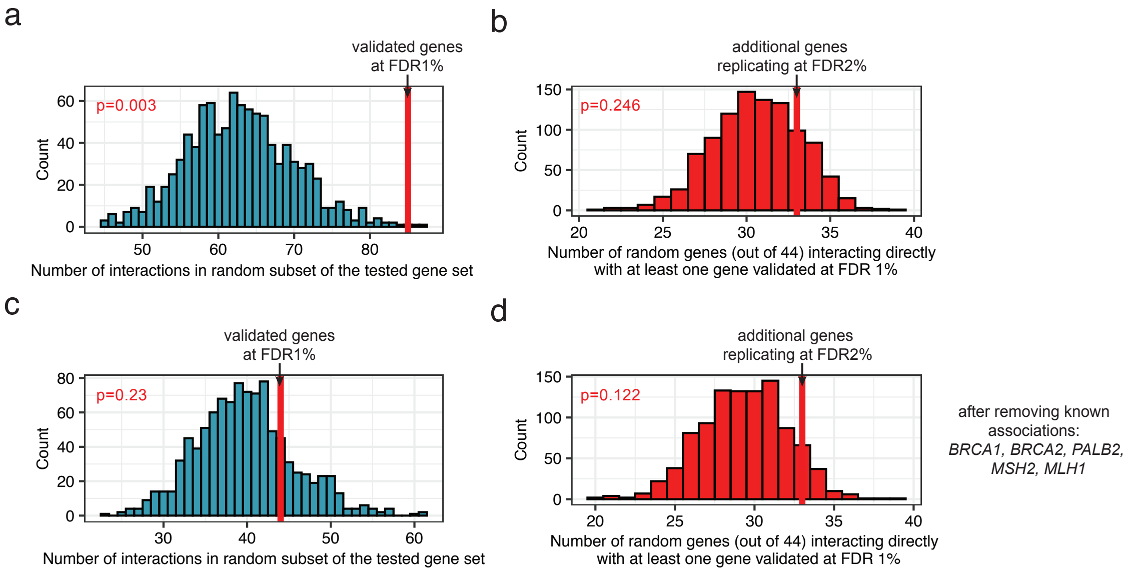


Figure 2.61: Network analysis (HumanNet) supports the role of rare germline variation in somatic mutational processes. All panels in this figure were generated using the functional gene network from HumanNet. (a) Number of physical interactions in a random subset of the tested gene set (controlled for interaction node degree) (x-axis). Red line shows the number of interactions within genes which replicated at a FDR of 1 %. (b) Number of randomly selected genes from the tested gene set interacting with at least one gene, which replicated at a FDR of 1 % (x-axis), (controlled for interaction node degree). Red line shows the number of genes, out of the ones which additionally replicated at a FDR of 2 %, interacting with at least one gene replicating at a FDR of 1 %. (c) Same as in panel a, after excluding known genes from the analysis (*BRCA1*, *BRCA2*, *PALB2*, *MSH2*, and *MLH1*). (d) Same as in b after excluding known genes from the analysis (*BRCA1*, *BRCA2*, *PALB2*, *MSH2*, and *MLH1*).

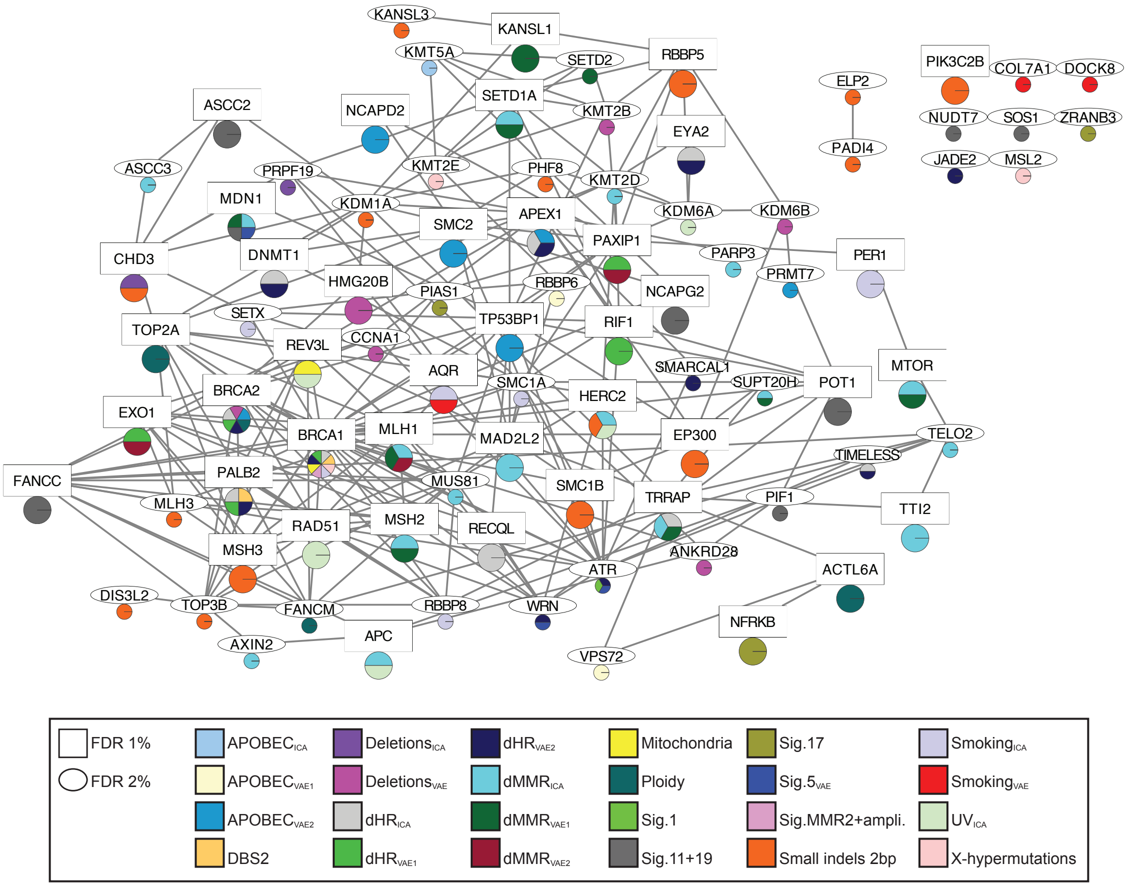


Figure 2.62: Overview of replicated hits in HumanNet network. Visualisation of physical interactions between proteins for genes replicating at a FDR of 1 % (square) and genes replicating at a FDR of 2 % (ellipse). Color code in pie chart shows the somatic components the corresponding gene was associated with (bottom panel). Line width corresponding to interaction score.

Prevalence of damaging germline variants in genes associated with somatic mutational phenotypes

To better estimate the contribution of RDGVs to differences in somatic mutational processes, we counted how many individuals in our dataset had certain RDGVs and compared this to randomly selected protein-coding genes while controlling for covered gene length (Figure 2.63). Considering known mutator genes, 44 individuals (0.6 %) had PTVs in Lynch syndrome dMMR genes (*MSH2*, *MLH1*, *MSH6*, *PMS2*), and 100 (1.5 %) had PTVs in in dHR genes (*BRCA1*, *BRCA2*, *PALB2*, *RAD51C*) in the discovery cohort (TCGA). Considering only the newly associated genes, 107 individuals (1.6 %) had a PTV in genes that replicated at a FDR of 1 %, and 166 (2.4 %) in genes which replicated at a FDR of 2 %. A similarly high prevalence of damaging variants in newly-discovered genes, relative to known mutator genes, was seen in prioritized missense variants, via the CADD score at stringent (≧25) and permissive thresholds (≧15; Figure 2.63). Additionally, when comparing this with prevalence of deleterious variants in control sets of length-matched genes, there is an excess of damaging missense variants in the known dHR and dMMR genes as well as in the newly-discovered genes at 1 % and 2 % FDR thresholds (Figure 2.63).

Taken together, these results suggest that the novel candidate mutator genes are affected by deleterious variants in a higher fraction of the population of cancer patients than the known human germline dMMR and dHR genes.

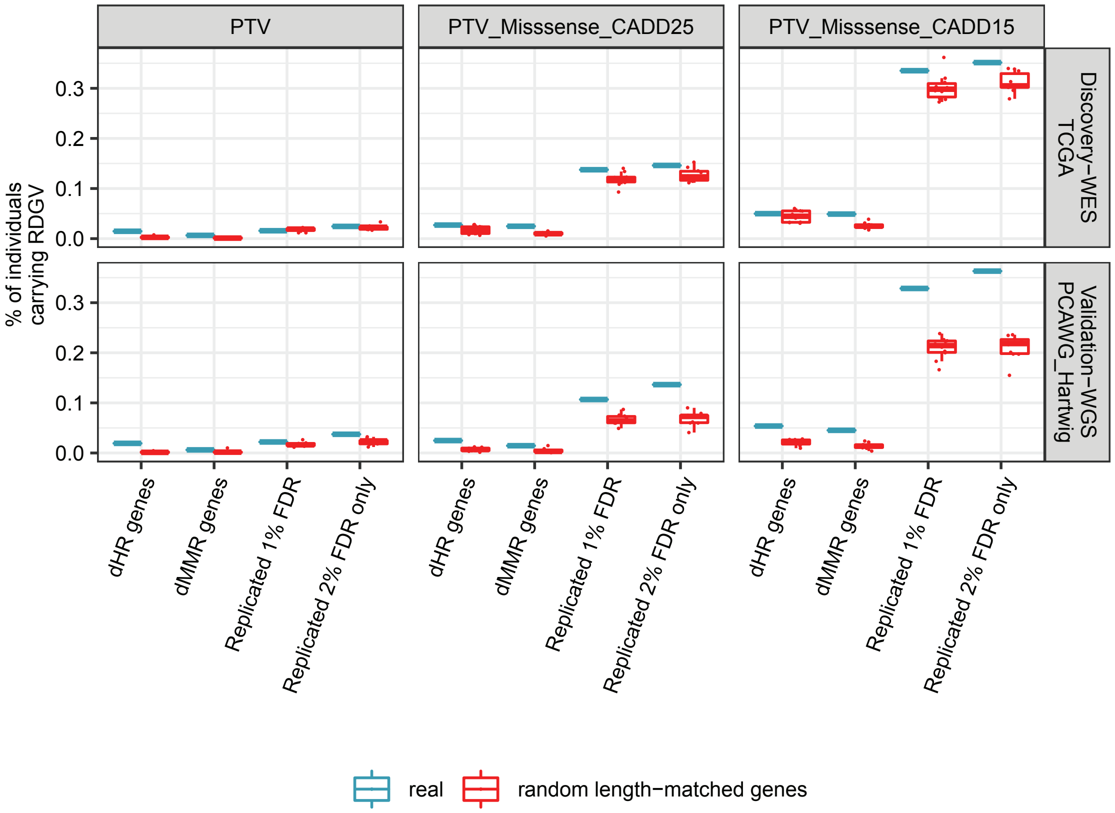


Figure 2.63: Frequency of RDGVs across cohorts. Showing the frequencies of RDGVs within the individuals (y-axis) of the discovery cohort (TCGA-WES) and validation cohort (PCAWG+Hartwig-WGS) (rows) across different RDGV sets (columns) for different gene sets (x-axis). Known dHR gene set includes *BRCA1*, *BRCA2*, *PALB2*, and *RAD51C*, known dMMR gene set includes *MSH2*, *MSH6*, *MLH1*, and *PMS2*, the replicated 1 % FDR set includes all genes replicating at a FDR of 1 % after excluding known dMMR and dHR genes, and the replicated 2 % FDR only set includes all additional genes which replicated at a FDR of 2 %. Color code for the real gene sets (blue) and length-matched, randomly selected protein-coding gene sets (red). Random selection for length-matched protein-coding genes was performed 10 times, and distribution shown in boxplot.

2.3 Common Variant Association Testing

Next, we investigated to which extent common variants affect the different extracted somatic mutational components (Figure 2.64). In contrast to rare variants, common variants can be tested independently. Similarly as for the rare variant association testing we performed quality control steps and extracted individuals of European ancestry. In addition, missing SNPs were imputed (Section 4.7.1) In the discovery cohort TCGA common variants were extracted from SNP array data, while in the validation cohort PCAWG_Hartwig common variants were extracted from WGS data. After checking data quality we performed genome-wide common variant association testing.

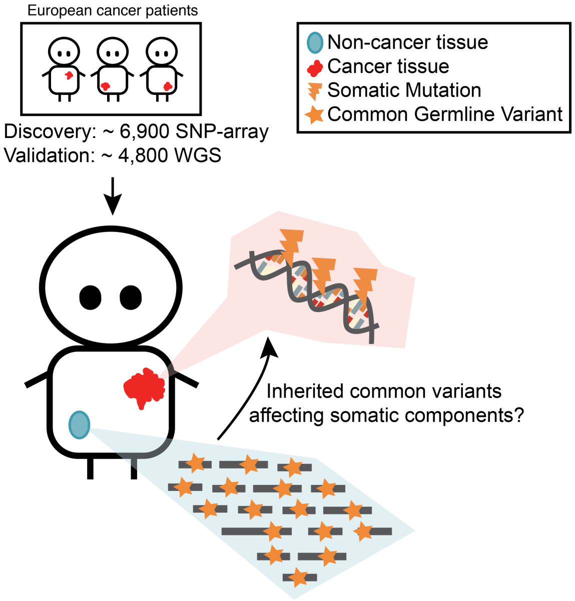


Figure 2.64: Illustration of common variant association testing. Associations were identified in the discovery cohort (TCGA -SNP array) and replicated in the validation cohort (PCAWG + Hartwig -WGS).

2.3.1 Quality Control

As described in the methods (Section 4.7), 7,886 individuals of European ancestry and 484,843 SNP with an allele frequency of at least 1 % were extracted from the SNP array data in the discovery cohort TCGA. To investigate for confounding, we performed a PCA and checked the first 10 PCs. As shown in Figure 2.65, we did not detect any confounding with respect to the TCGA project id (Figure 2.65a), gender (Figure 2.65b), or age or onset of disease (Figure 2.65d). In the validation cohort PCAWG_Hartwig-WGS, 4,831 individuals of European ancestry and 481,488 SNPs (same SNPs as in TCGA) were extracted from the WGS data with an allele frequency at least 1 %. Similarly, as in TCGA, we did not detect any confounding in the PCA with the respect to the project id (Figure 2.66a), gender (Figure 2.66c), or age at diagnosis (Figure 2.66e). For the validation cohort germline variants were called in 3 different centers: PCAWG, Hartwig, and at the IRB (Section 4.7). Here, we could also not detect any major batch effects (Figure 2.66b).

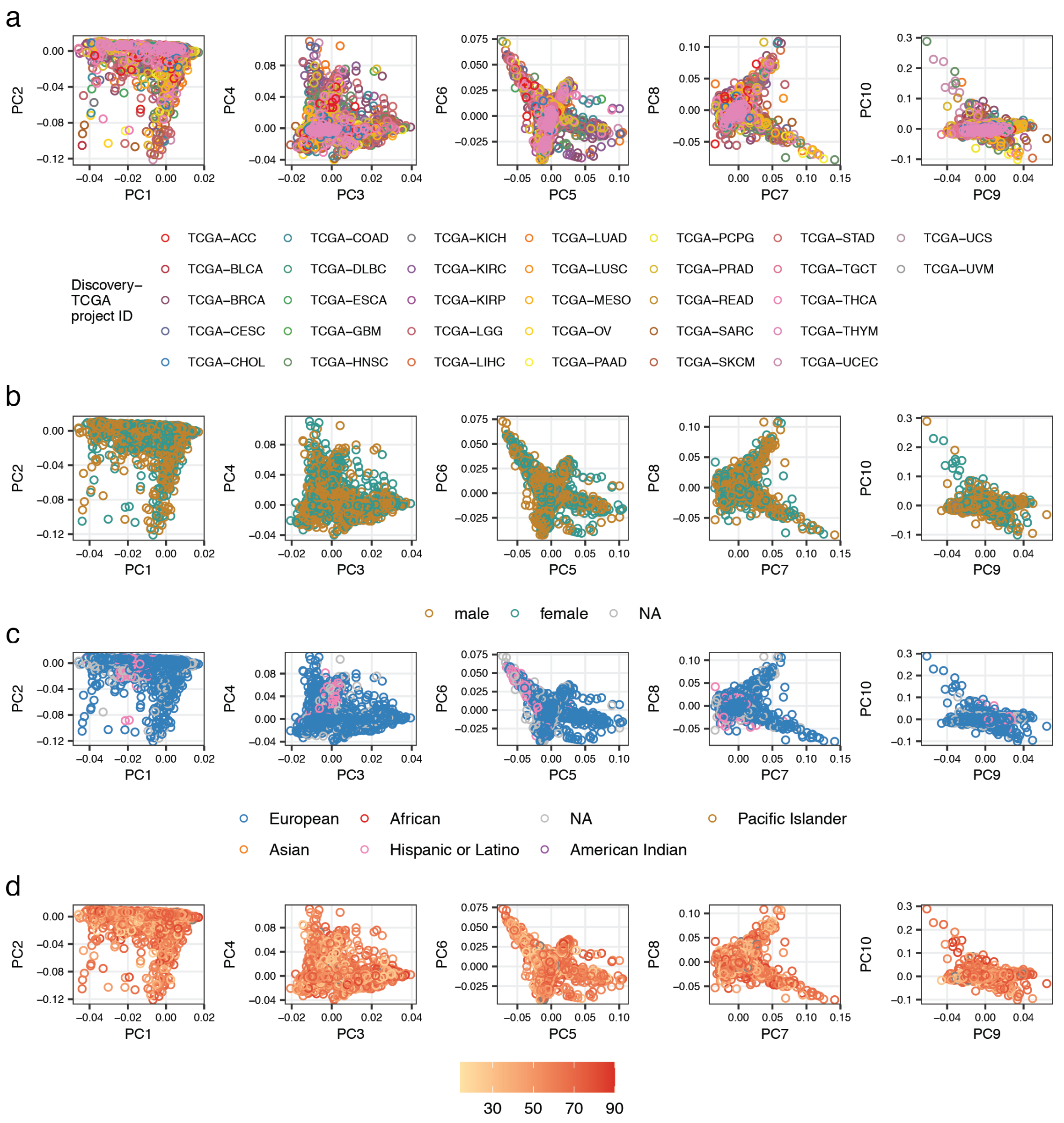


Figure 2.65: Principal component analysis on common germline variants in TCGA SNP array data. Principal components 1 to 10 color coded by (a) TCGA project id, (b) gender, (c) ethnic, and (d) age at diagnosis.

Next, we checked for batch effects between the discovery and validation cohort. Common variants were extracted from SNP array data in the discovery cohort and from WGS data in the validation cohort. In order to have a reference for comparison, we performed a PCA merging the data from the discovery cohort, validation cohort and from 1000 genomes (1KG). SNPs in 1KG were called from WGS data. As shown, in Figure 2.67, there was no major batch effect directly visible, meaning that the genotyped SNPs from the two different technologies did not separate (Figure 2.67a). The first components mostly captured differences with respect to ethnicity in the European population such as people with Finish ancestry vs. the rest (Figure 2.67c). Still, we observed in several PCs a wider distribution of PC values in the SNP array data compared to the WGS data (Figure 2.67c). This became in particular visible when looking at PCs 3, 5, and 7 (Figure 2.68).

Taken together, even though we did not detect major batch effects between the two data types (SNP array data and WGS), there was a higher variance visible in the SNP array data compared to the WGS data. This increased variance was most likely not connected to ethnicity, but to the genotyping technology, since the distributions of the PCs from 1KG and the validation cohort were much more similar, than the distributions from the discovery cohort to those two.

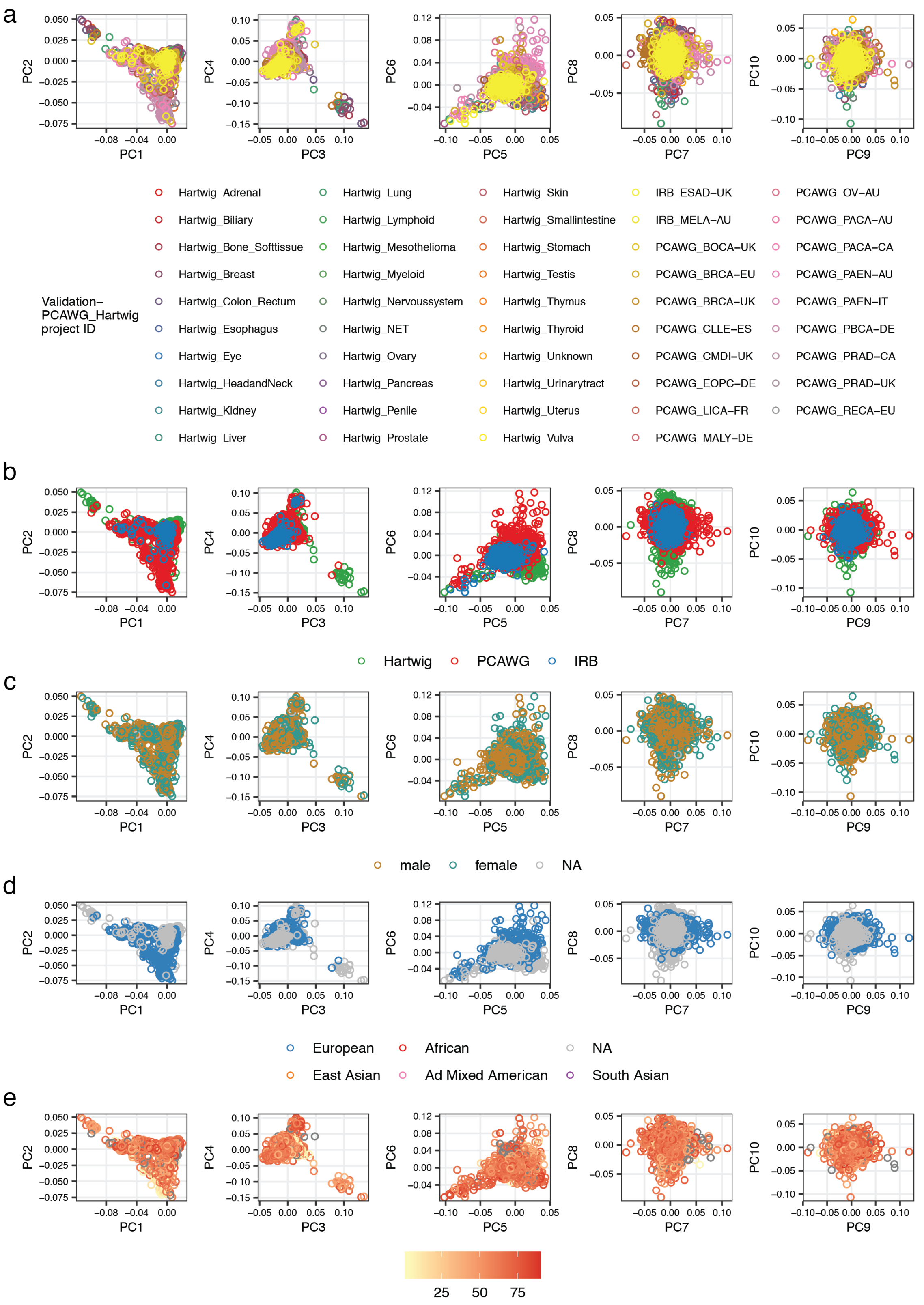


Figure 2.66: Principal component analysis on common germline variants in PCAWG-Hartwig -WGS data. Principal components 1 to 10 color coded by (a) cancer project id, (b) center/project where common variants were called, (c) gender, (d) ethnic, and (e) age at diagnosis.

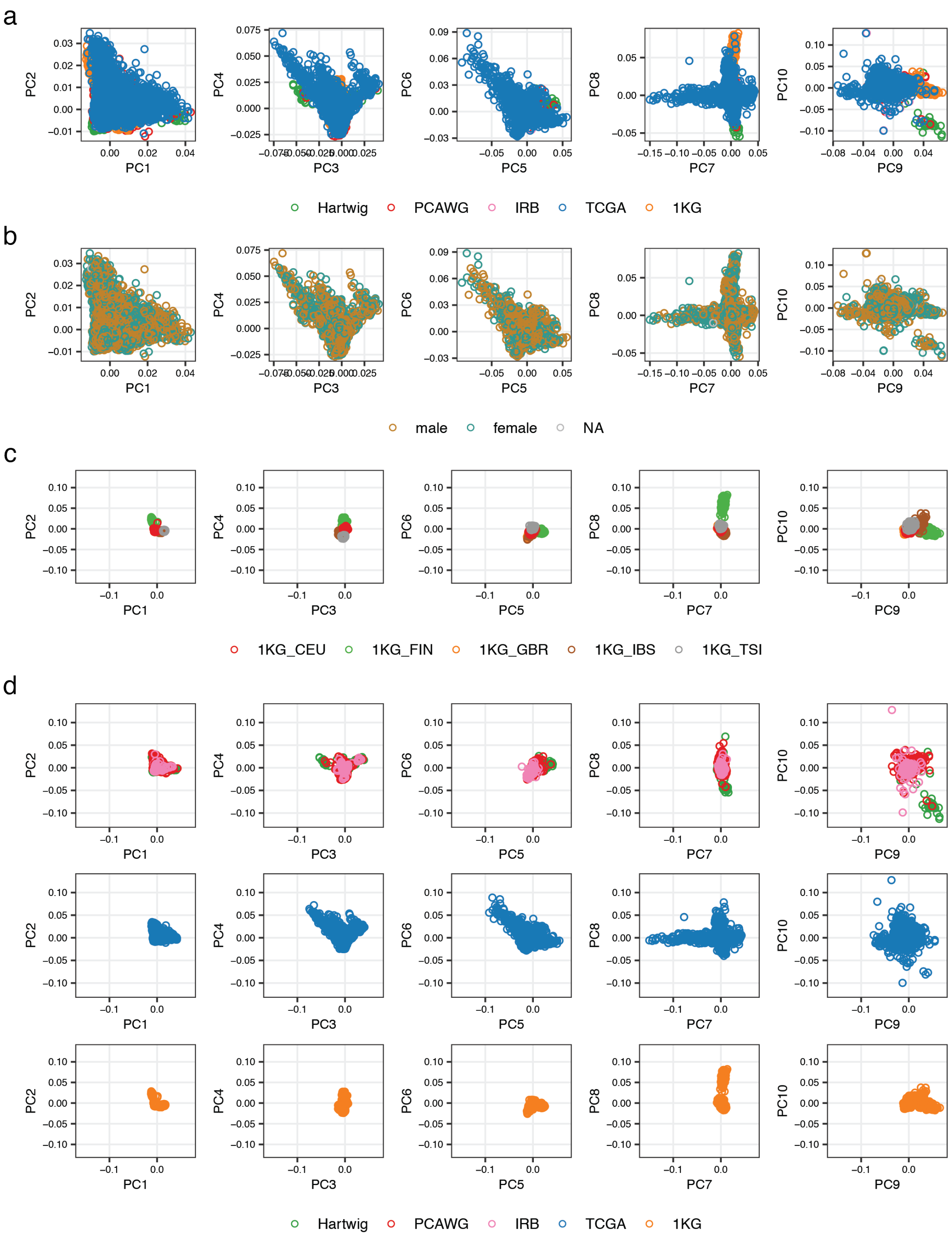


Figure 2.67: Principal component analysis on common germline variants on TCGA-WES SNP array data, PCAWG-Hartwig-WGS data and 1000 genomes (1KG)-WGS data together. Principal components 1 to 10 color coded by (a) center/project where common variants were called, (b) gender, (c) reported ethnicity in 1KG (only showing 1KG), and (d) different data sources separately (top: PCAWG_Hartwig, middle: TCGA, bottom: 1KG).

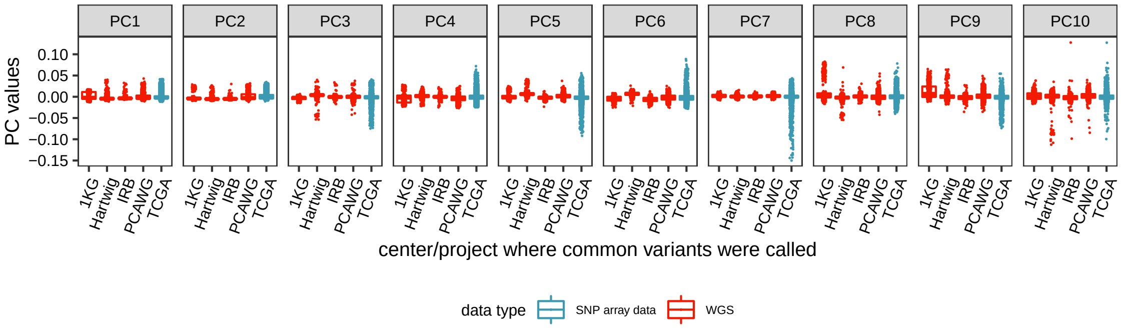


Figure 2.68: Wider distributions of principal components from samples which were genotyped by SNP arrays in comparison to WGS data. Showing distribution of PC values (y-axis) based on center/cohort where common germline variants were called (x-axis) for the first 10 PCs. Colored by data type which was used for genotyping.

2.3.2 Genome-Wide Association Study (GWAS)

For the GWAS, we tested for associations between around 484k SNPs and all 29 somatic mutational components as it was performed before for the rare variant association testing. Testing was only performed for cancer types in which the sample size was above 200 in the discovery and in the validation. In total, testing was performed in 8 different cancer types (brain glioma, breast, colon + rectum, lung adenocarcinoma, lung squamous carcinoma, prostate, skin, stomach + esophagus) and in pan-can. Inflation factors λ were all distributed between 0.97 and 1.02 with the median at 1.0, indicating that the tests were well calibrated (Figure 2.69).

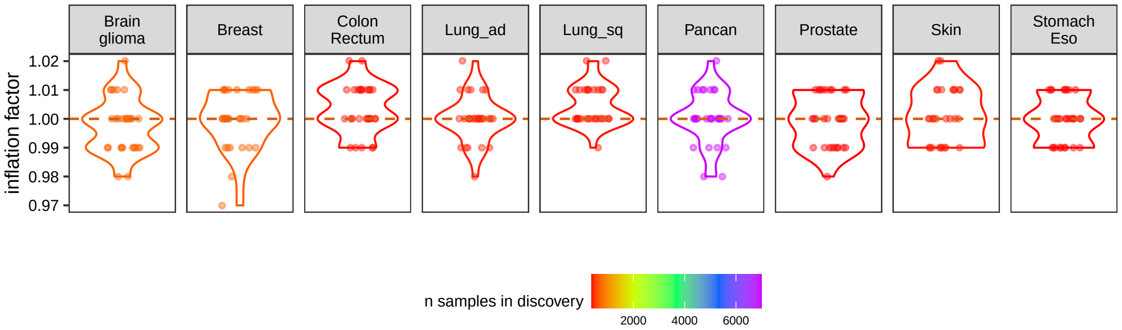


Figure 2.69: GWAS were well-calibrated. Showing inflation factors (y-axis) across all cancer types and pan-can. Each point represents one estimated inflation factor for one component. 29 components were tested in each cancer type or in pan-can. Color code for the sample size in the discovery cohort TCGA - WES.

In total, 10 SNPs reached genome-wide significance ($p < 5*10^{-8}$; p-value after performing Bonferroni correction on all possible common independent human SNPs[234]) covering 5 out of 9 cancer types and 9 out of 29 somatic mutational features. All 10 hits were re-tested in the validation cohort in the corresponding cancer type. One SNP (rs17177814) had an allele freuency of less than 1 % in the validation cohort. None of the 10 re-tested hits replicated in the validation cohort at a p-value of 0.05 (corrected by multiple testing via Bonferroni). Still, one hit reached at least a p-value < 0.05 (SNP id: rs635332). This SNP is located in the intron of the gene *TSPAN9* and associated in colorectal cancer with an increase of a somatic component (Sig.MMR2+ampli.), which has large contributions from the SNV-derived cancer signature Ref.Sig. MMR2[138] and amplification events. This gene encodes a cell surface protein which has cell developmental functions and has no clear link to DNA repair pathways. In addition, only 5 out of 10 of the re-tested hits had the same effect size direction as in the discovery cohort. We also evaluated whether any of the 10 hits and associated SNPs (by clumping, Section 4.7.3) have been previously reported in the GWAS catalog from EMBL-EBI (October 2020) as a cancer risk SNP, which was not the case.

Table 2.2: Overview of common SNPs reaching genome-wide significance in association study and asociations in validation cohort

								Discovery			Validation		
Chr	Location	SNP	Ref	Alt	Gene	Components	Cancer type	P-value	Effect	Alt freq.	P-value	Effect	Alt freq.
18	56,739,983	rs9960014	C	A	*OACYLP*	UV_ICA	Colon_Rectum	4.04e-08	-0.42	0.38	0.71	-0.03	0.35
12	3,386,273	rs635332	G	A	*TSPAN9*	Sig.MMR2+ampli.	Colon_Rectum	1.04e-08	0.47	0.31	0.04	0.16	0.30
9	133,279,776	rs4740385	G	A	-	Deletions_ICA	Colon_Rectum	4.33e-08	0.53	0.15	0.80	0.03	0.13
3	110,962,617	rs2958070	A	C	*PVRL3*	dHR_VAE1	Colon_Rectum	1.51e-08	0.40	0.55	0.86	-0.01	0.52
3	195,929,252	rs1105977	G	T	*ZDHHC19*	Amplifications	Lung_sq	1.75e-08	-0.43	0.27	0.96	0.004	0.29
4	114,619,923	rs17593296	A	C	*CAMK2D*	DBS2	Pan-can	4.45e-08	-0.14	0.14	0.24	0.04	0.12
2	173,446,266	rs1869922	A	G	*PDK1*	Sig.11+19	Pan-can	1.39e-08	0.14	0.12	0.67	0.01	0.14
21	20,290,110	rs17177814	G	T	-	Sig.1	Skin	5.87e-09	1.48	0.01	0.38	0.29	0.009
7	77,501,962	rs2471585	T	C	*PHTF2*	Sig.1	Skin	6.11e-09	0.50	0.14	0.83	-0.02	0.14
13	100,114,342	rs17656454	A	C	-	APOBEC_VAE1	Stomach_Eso	5.87e-09	0.68	0.10	0.59	-0.06	0.10

To further investigate potential reasons why all of the hits did not replicate in the validation cohort, we performed a power analysis (Section 4.7.4). In GWAS, the statistical power depends on different factors: sample size, effect size of the causal genetic variant, allele frequency, and amount of linkage disequilibrium (LD) between the genotyped SNP and causal SNP[247]. As shown in Figure 2.70, the statistical power to detect any association with an effect size of around 0.1 would be at the sample size of the discovery cohort at an allele frequency of 50 % at ∼70 % and at an allele frequency of 25 % at ∼40 %. For the individual cancer types the statistical power would be even lower. Lower powered studies do not only increase the chance of false-negatives, but also the probability of finding false-positives[235]. Thus, this study was underpowered to detect potential small true effects. In particular, when comparing with effect sizes of cancer risks SNPs, one would expect that most SNPs would have low effect sizes[246].

All in all, we performed GWAS to find associations between common SNPs and different types of somatic mutational patterns. Several hits reached genome-wide significance in the discovery cohort, but none of them replicated in the validation cohort after correcting for multiple testing. Furthermore, there was no overlap between the hits and known cancer risk SNPs. As shown in the power analysis, one of the reasons could be due to low statistical power[235]. Other potential reasons for the lack in replication could originate from batch effects due to the two different genotyping technologies which were utilized or differences in the estimation of the somatic features (WES in discovery, WGS in validation).

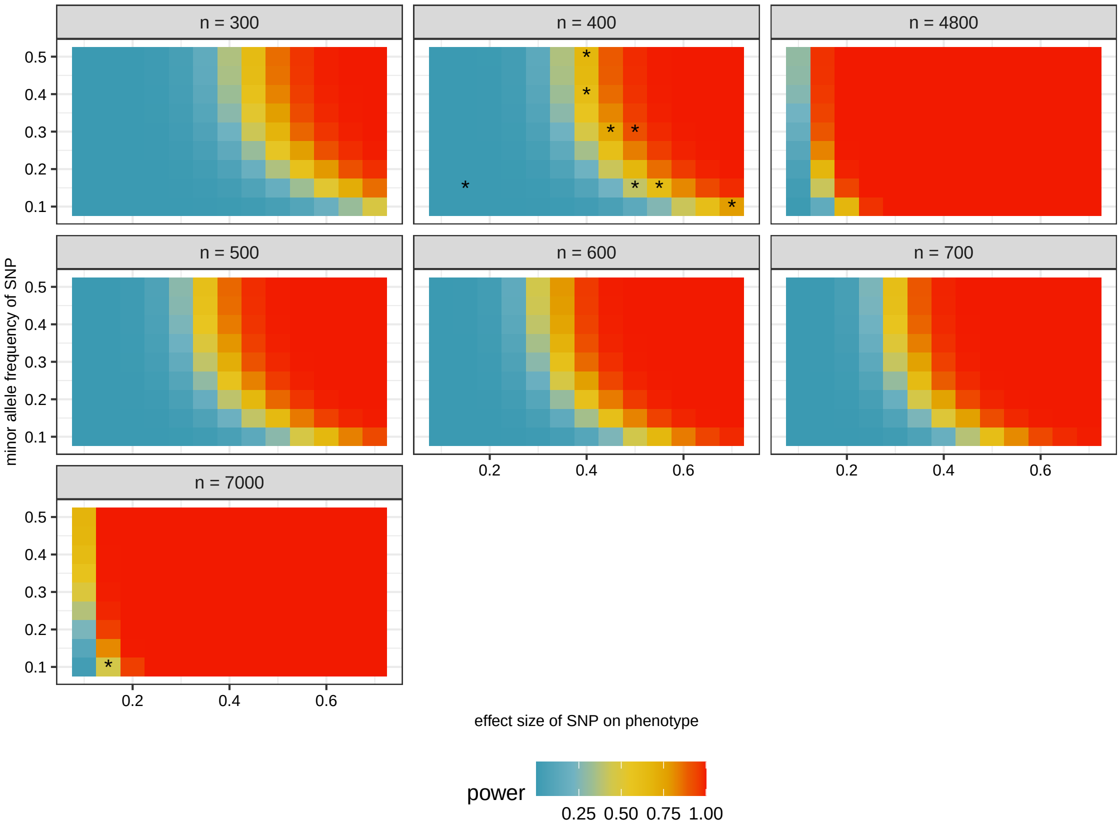


Figure 2.70: GWAS power analysis. Theoretical statistical power for genome-wide significance was estimated (color code) at varying allele frequencies (y-axis) and effect sizes (x-axis) for different sample sizes (subplots). Nine hits, which were identified in the discovery cohort (Table 2.2) are marked on the plots based on with which effect size, allele frequency and in which cancer type they were identified. One hit (rs17177814) not marked, but expected power at 100 %.

2.4 SNP-Heritability of Somatic Mutational Processes

We further asked the question to which extent the different somatic features and components are heritable (Figure 2.71). For this purpose, with estimated SNP heritabilities utilizing common SNPs with two commonly used methods GREML[194] and LDSC[284]. Both methods make use of all common SNPs with the assumption of an infinitesimal model[236]. The infinitesimal model assumes that a trait is influenced by an infinitely high number of genes and that each gene makes a small contribution to the variation of a trait.

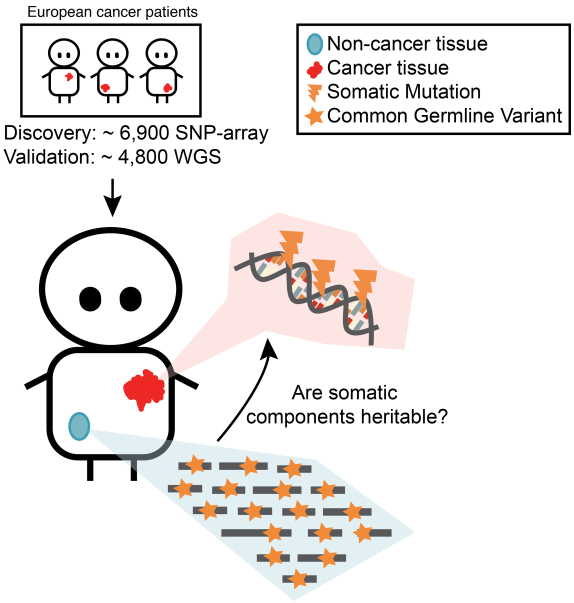


Figure 2.71: Illustration of estimation of SNP-heritabiites. SNP-heritabilities were estimated in the TCGA cohort (SNP-array data) and in PCAWG + Hartwig (WGS data).

2.4.1 Signature Ref.Sig. MMR1 Mutations and the Total Number of C>T Mutations have a Heritable Component

First of all, SNP heritabilities of the input somatic features were estimated with both methods across the two cohorts. We only calculated SNP heritabilities in pan-can, while controlling for the individual cancer types. Since errors in the estimation of the SNP-heritability by GREML are around 318 divided by the sample size[191] (~32 % at a sample size of 1,000), sample sizes were too low to estimate heritabilites for the specific cancer types (n < 1,000; Table 4.7). Still, errors in the estimation of SNP heritabilities were high due to the sample sizes (around 6,900 in TCGA and 4,800 in PCAWG_Hartwig). For GREML, the median error of SNP heritability was at 4.6 % (1st quantile: 4.6 %, 3rd quantile: 4.8 %) in TCGA-WES and at 6.6 % (1st quantile: 6.5 %, 3rd quantile: 6.6 %) in PCAWG_Hartwig-WGS. For LDSC, the median error of SNP heritability was at 7.5 % (1st quantile: 7.1 %, 3rd quantile: 8.0 %) in TCGA-WES and at 10.6 % (1st quantile: 10.3 %, 3rd quantile: 10.9 %) in PCAWG_Hartwig-WGS.

There were differences in the estimated heritabilities between the two cohorts and methods (Figure 2.74). When comparing the estimated heritabilities between LDSC and GREML within the same cohort, Pearson correlation was at 0.37 for TCGA-WES and at 0.32 for PCAWG_Hartwig-WGS (Figure 2.72). The Pearson correlation for the estimated heritabilities between the two cohorts was at 0.09 for GREML and 0.17 for LDSC (Figure 2.73). For instance, in TCGA-WES amplification events of a size of between 10 and 100 kb had the highest heritability via GREML with around 13 %, while in PCAWG_Hartwig-WGS the same feature had a heritability of around 3 %. In PCAWG_Hartwig-WGS, very high heritabilities were estimated for SNV-derived signatures 11 (around 35 %) and 19 (around 28 %) via GREML. The same features had low heritabilities in TCGA-WES with 5.3 % and 0.3 %, respectively. There could be several reasons why SNP heritabilities varied so much between the cohorts and methods such as sample size, data type to extract common variants (SNP array vs WGS), data source to estimate somatic features (WES in TCGA and WGS in PCAWG_Hartwig), or differences in the number and distribution of cancer types in each cohort. The potential reasons will be covered in more detail in the discussion (Section 3.4).

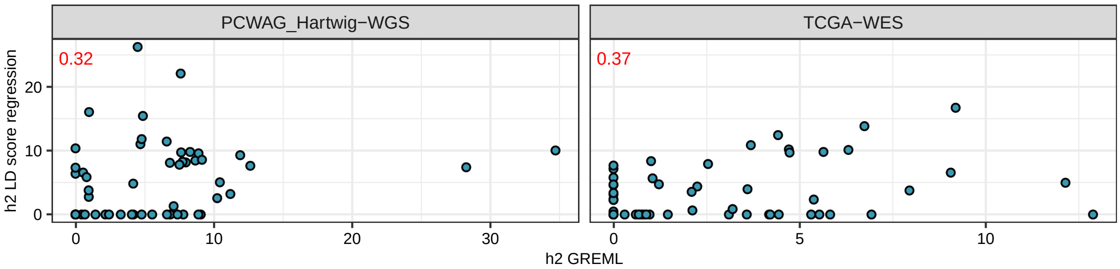


Figure 2.72: Correlations of heritability estimates between methods. Pearson correlation (red) shown for heritability estimates between the methods for both cohorts separately.

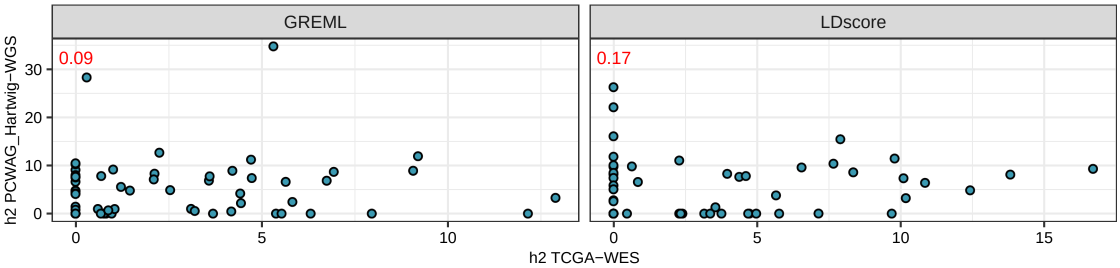


Figure 2.73: Correlations of heritability estimates between cohorts. Pearson correlation (red) shown for heritability estimates between the cohorts for both methods separately.

Next, we focussed on the features, that had increased heritabilities in both cohorts. We combined p-values from each method between the two cohorts via Fisher's method for the 65 somatic mutation features and then adjusted for multiple testing. For GREML, 6 out of 65 features were significant at a FDR of 20 % and 2 features were significant at a FDR of 5 % (Figure 2.75). While four of them (Ref.Sig 11, Ref.Sig 19, amp_10to100kb, trx_TtoA) were only significant due to the very high heritabilities in one of the two cohorts, the somatic features covering the total number of C>T mutations and SNV-derived cancer signature Ref.Sig. MMR1 had increased heritabilities in both cohorts with at least 8 % in GREML. Signature Reg.Sig. MMR1 has been connected to defective mismatch repair, and in particular, to the signature generated by knockout of *MSH6*[138]. Hereditary colorectal cancer (Lynch syndrome) is caused by a deficiency in MMR and so far, several common variants have been reported to increase the risk of this disease[176]. Thus, the increased heritability of the somatic cancer signature Ref.Sig. MMR1 could be explained via common SNPs which affect DNA mismatch repair. The increased heritability of C>T mutations could be attributed to many different sources of mutagenesis such as smoking, reactive oxygen species, dMMR, dHR, and many more[14]. It is unclear which mutational processes cause this increased heritability. The total mutation burden (total_SNV) was consistently increased in both cohorts as well with a SNP heritability of around 5.6 % in TCGA-WES and of around 6.6 % in PCAWG_Hartwig-WGS for GREML, which was lower than the previously estimated heritability of around 12.9 %[191].

For LDSC, none of the features were significant after adjusting for multiple testing, which can be attributed to the lower power of this method compared to GREML. Still, it could be seen that the somatic feature for C>T mutations, similarly as in GREML, had an increased heritability in both cohorts (16.7 % in TCGA and 9.3 % PCAWG_Hartwig) as well as signature MMR1 (6.6 % in TCGA

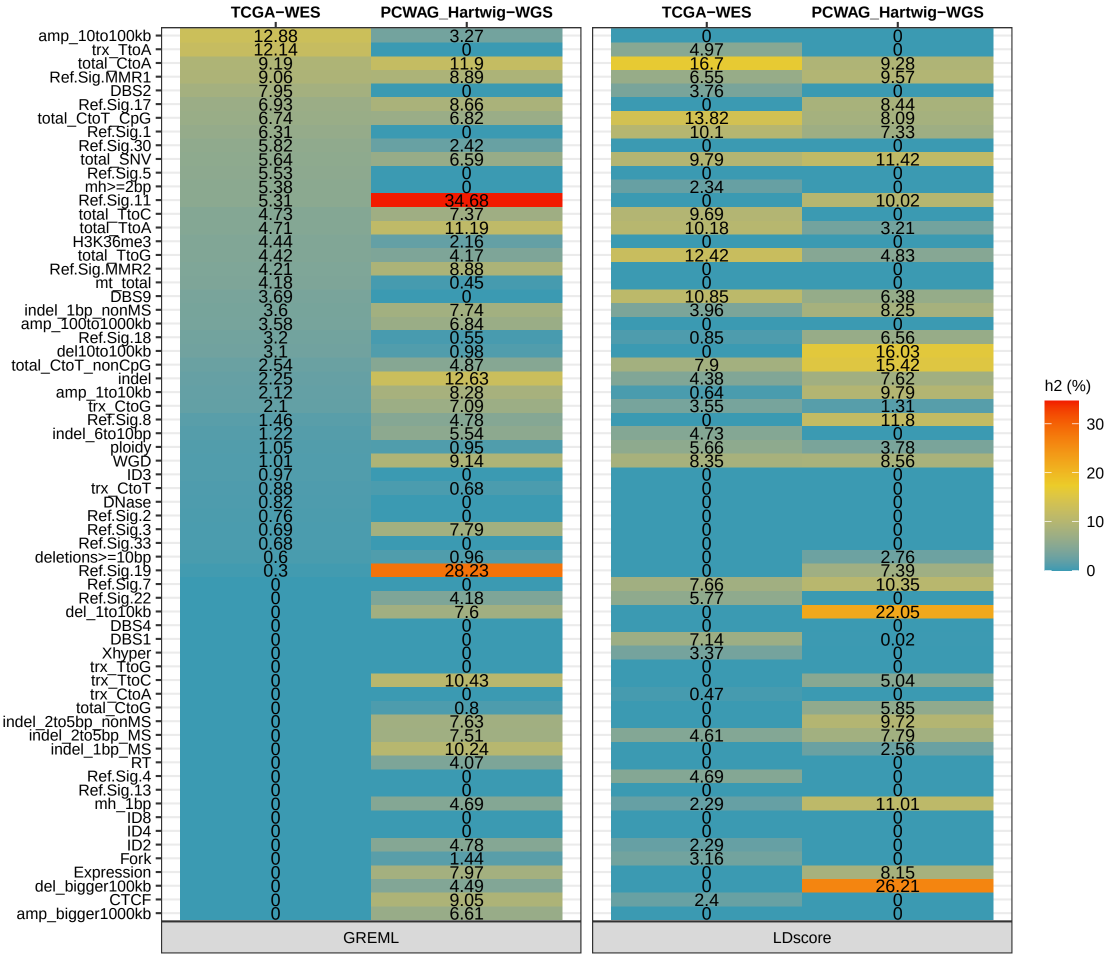


Figure 2.74: Overview of estimated SNP heritabilities for somatic input features. SNP heritabilities were calculated via GREML and LDSC for the two cohorts TCGA-WES and PCAWG_Hartwig-WGS (x-axis) for 65 somatic features (y-axis). Color code shows estimated SNP-heritability h^2 in %. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation.

and 9.6 % PCAWG_Hartwig). In addition, heritability for the total mutation burden was increased in both cohorts (9.8 % in TCGA and 11.4 % PCAWG_Hartwig), which was closer to the results from a previous study[191].

Taken together, we detected in particular two somatic features with increased heritability estimates in both cohorts: signature Ref.Sig. MMR1 and the total number of C>T mutations. The estimated heritability for the total mutation burden was consistently increased across cohorts and methods, but lower in comparison to what was reported before[191].

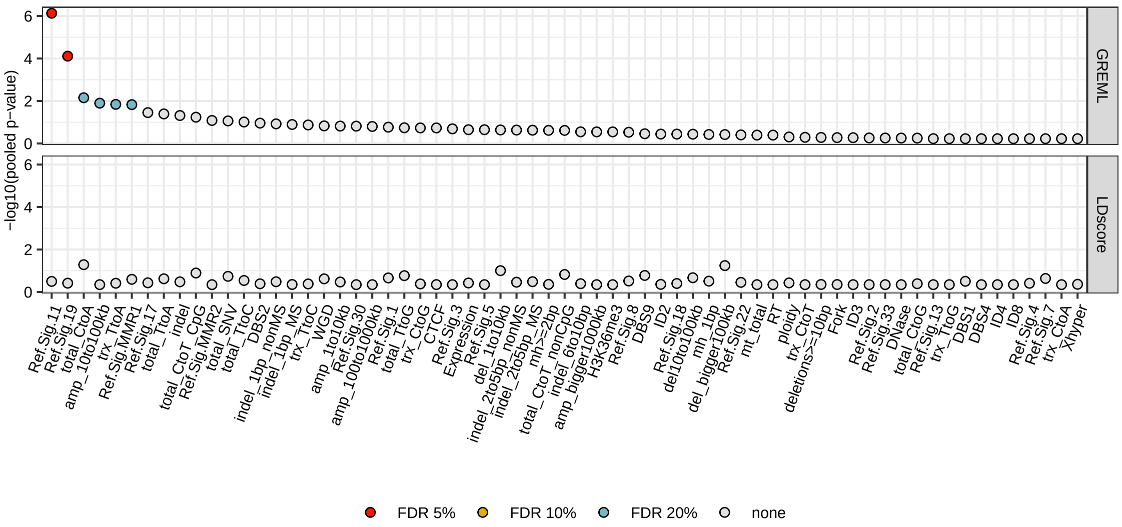


Figure 2.75: Pooled p-values for SNP-heritabilities of somatic input features. P-values from estimated heritabilities across cohorts were combined via Fisher's method across all features (x-axis) for both methods GREML and LDSC. Transformed p-value shown on the y-axis and color code multiple testing correction via Benjamini-Hochberg. Fork: replicative strand bias, RT: replication timing, trx: transcription strand bias, Xhyper: Chromosome X hypermutation.

2.4.2 Heritability Estimates for Somatic Components Varied across Cohorts and Methods

We also investigated the heritability estimates for our extracted somatic components. The estimated errors were in the same range as for the somatic features. There was no correlation between the estimated heritabilities between the two cohorts: -0.018 for GREML and -0.017 for LDSC. Correlations were even lower than for the raw somatic features and heritabilities varied between the different cohorts and methods (Figure 2.76).

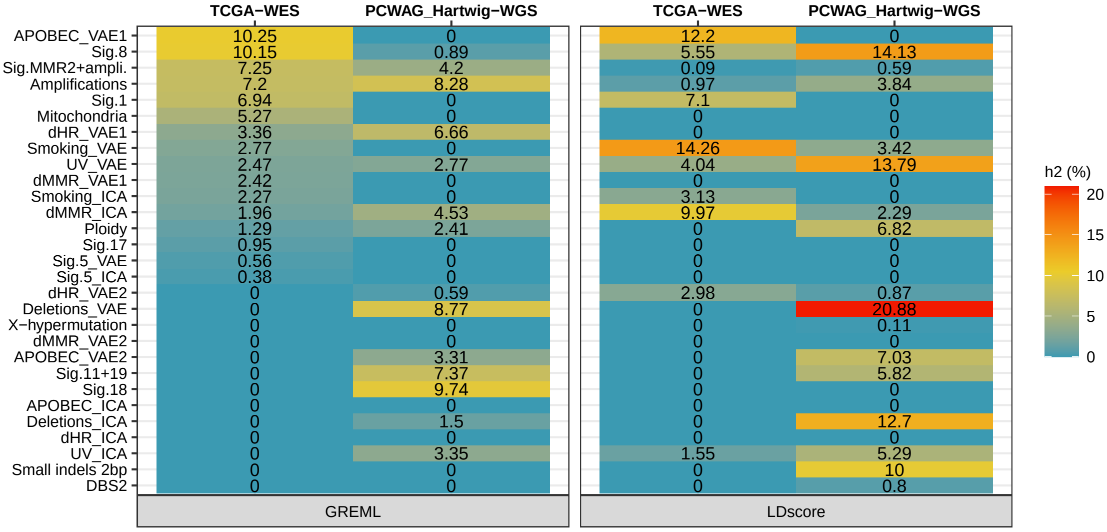


Figure 2.76: Overview of estimated SNP heritabilities for somatic components. SNP heritabilities were calculated via GREML and LDSC for the two cohorts TCGA-WES and PCAWG_Hartwig-WGS (x-axis) for the 29 extracted somatic components (y-axis). Color code shows estimated SNP-heritability h^2 in %.

We checked by combining p-values from the heritability estimates from the two cohorts via Fisher's method whether any components had consistently increased heritabilities. For both methods, none of the components were significant at a FDR lower or equal to 20 % (Figure 2.77).

In conclusion, we did not extract somatic components which consistently had an increased heritability across cohorts and/or methods. This was in contrast to the raw somatic features, where we found several features with significantly increased heritability estimates. Still, independent of the somatic feature or somatic component which was utilized, most heritability estimates for somatic mutational processes based on common SNPs were below 15 % (Figures 2.74 and 2.76).

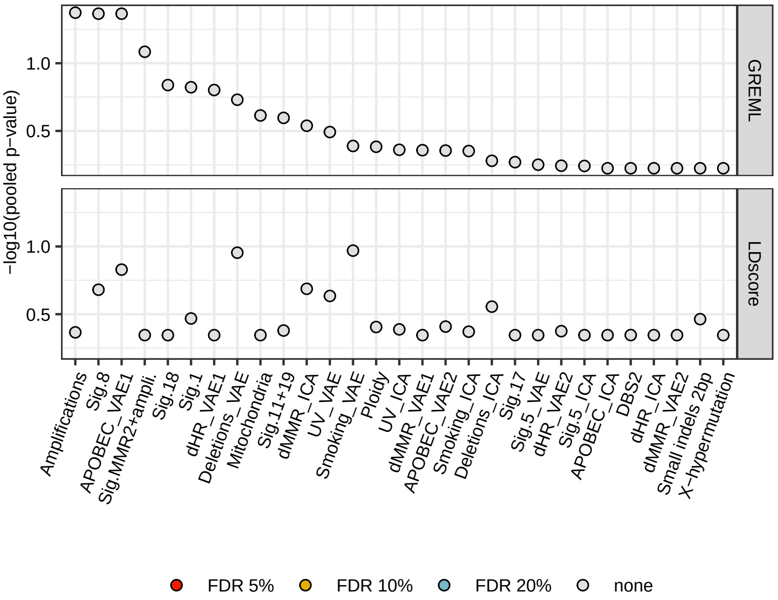


Figure 2.77: Pooled p-values for SNP-heritabilities of somatic components. P-values from estimated heritabilities across cohorts were combined via Fisher's method across all features (x-axis) for both methods GREML and LDSC. Transformed p-value shown on the y-axis and color code multiple testing correction via Benjamini-Hochberg.

2.4.3 Contribution of Heritability to Total Mutation Burden originating from different Cancer Types

To further investigate which mutational processes contribute to the heritability of the total SNV-based mutation burden, we removed individual cancer types from the analysis and then calculated the difference in heritability in comparison to the total SNV heritability before dropping out the respective cancer type. To check whether the difference was significant, we randomly removed the same number of individuals from the analysis 1,000 times to get a null distribution.

For several cancer types, the heritability estimate changed significantly when it was dropped (Figure 2.78). For instance, the heritability of the total mutation burden decreased when removing biliary cancer, ER-positive_HER2-negative breast cancer, or stomach cancer. The heritability of the total mutation burden increased when removing ER-negative_HER2-positive breast cancer, cancer types of the nervous system, or head and neck cancer. A decrease in the heritability after removing a cancer type would indicate that the corresponding cancer type contributes significantly to the total mutation heritability. This information could be utilized to pin down which mutational processes are responsible for the increased heritability and/or exclude several mutational processes. For instance, if mutations due to UV-light exposure would have had a high contribution to the heritability of the total mutation burden, one would have expected that dropping skin cancers of the analysis would lead to a decreased heritability estimate.

As shown in Figures 2.79 and 2.80, we pooled p-values either from an upper- or lower-tailed test for the corresponding cancer types within the same tissue or origin to examine whether a specific tissue contributes significantly to the total mutation heritability. None of the tissues were significant after controlling for multiple testing.

Taken together, while we observed that several cancer types contributed significantly to the heritability of the total mutation burden, we could not pin down which mutational processes are responsible and we also did not find specific tissues contributing to the total heritability. For some cancer types such as skin or lung cancer, the majority of samples can be connected to a specific mutational process (such as smoking or UV light), while other cancer types comprise a much more diverse set of underlying mutational processes[237]. This makes it more challenging to pin down the underlying mechanisms when solely dropping samples based on cancer types.

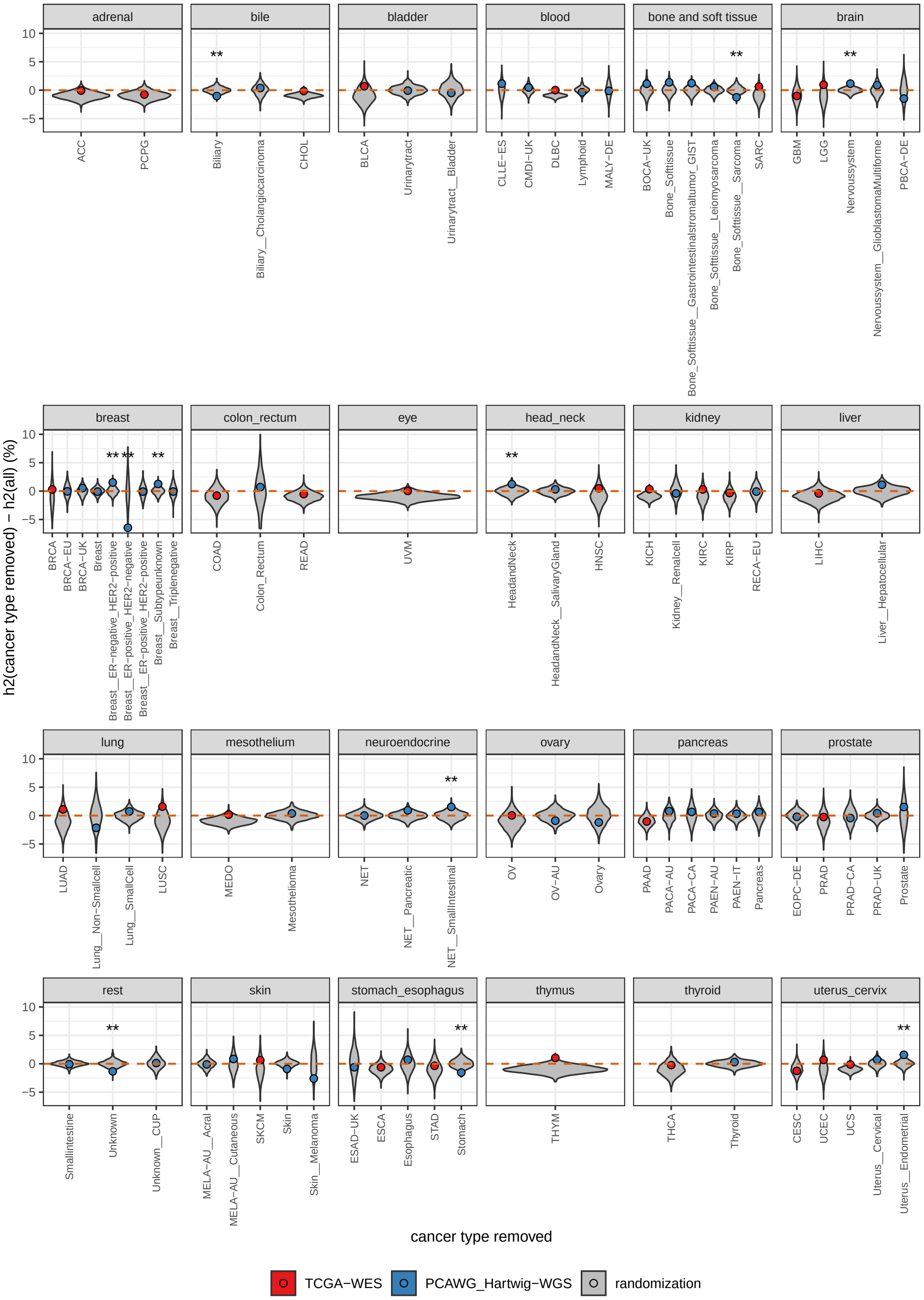


Figure 2.78: Heritability for total mutation burden when removing individual cancer types. Heritabilities were estimated via GREML. Y-axis showing difference in heritability for total mutation burden after removing an individual cancer type. Color code for the two cohorts (red, blue) and randomization (grey; 1,000 data points). Faceting for tissue of origin. Cancer types annotated with ** when p-value < 0.05 either for upper- or lower-tailed test.

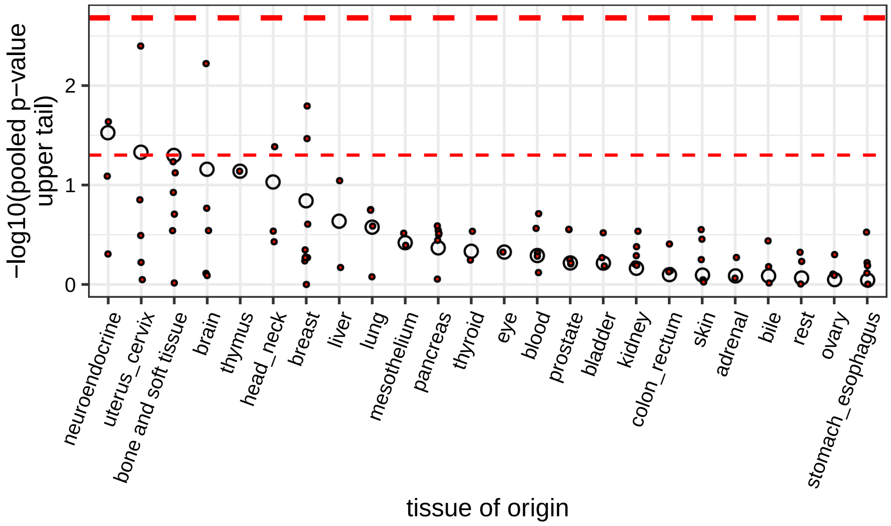


Figure 2.79: Pooled p-values for change in heritability of total mutation burden estimates across tissues - upper tailed test. P-values (upper tailed test) from individual cancer types (small points) for influence on total mutation burden were pooled via Fisher's method (y-axis) across tissues (x-axis). Thin dashed line at p-value of 0.05 and thick dashed line at 0.05/24 (number of tissues).

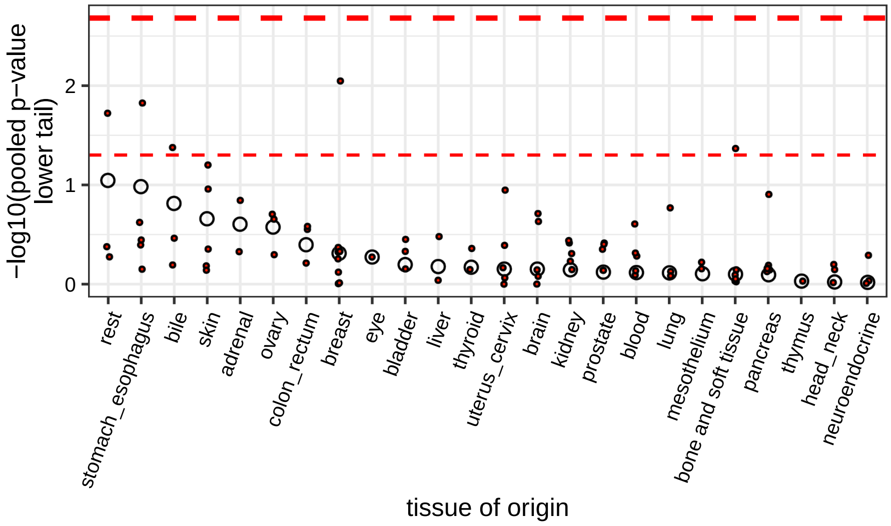


Figure 2.80: Pooled p-values for change in heritability of total mutation burden estimates across tissues - lower tailed test. P-values (lower tailed test) from individual cancer types (small points) for influence on total mutation burden were pooled via Fisher's method (y-axis) across tissues (x-axis). Thin dashed line at p-value of 0.05 and thick dashed line at 0.05/24 (number of tissues).

2.4.4 Several Mutational Processes contribute to the Heritability of the Total Mutation Burden

Since we could not pin down the underlying mutational processes contributing to the heritability of the total mutation by dropping individual cancer types, we applied a different approach. This time, we took the whole data set and removed individual mutation types in their trinucleotide context from an individual cancer type from the whole data set. For instance, we removed all C>T mutations in a CCC context from lung cancer samples, estimated the heritability, and calculated the difference to the total heritability when counting all SNVs. We performed this analysis for all 96 possible SNV-trinucleotide combinations across all cancer types (Figure 2.81).

To find specific mutational processes contributing to the total SNV heritability, we performed a PCA on the dataset after capping an outlier h^2 value (Section 4.8.3). The input matrix consisted of 96 rows (for each SNV-trinucleotide pair) and 104 columns (different cancer types). The first PCs explained a significant amount of the variance in the data (expectation at $\approx 1\,\%$) and, in particular, the first four PCs were able to explain around $40\,\%$ of the variance (Figure 2.82).

We investigated the individual cancer types contributing to the first components (Figure 2.83), the SNV-trinucleotide contexts having the highest contributions to the PCs (Figure 2.84), and the correlations between the SNV-trinucleotide contributions to the individual PCs and the established COSMIC signatures[4] (Figure 2.85). In this way, we aimed to examine to which extent the different components captured specific mutational processes, aiming to identify specific mutational processes contributing to the heritability of the total mutation burden.

PC1 had high negative correlations with C>G and C>T mutations in TCA and TCT contexts, which are typical for APOBEC-directed mutagenesis signatures (Signature 2 and 13[6]). Accordingly, this component also had high negative correlations with cancer types that are enriched for APOBEC-directed mutagenesis (e.g. Hartwig-Biliary and Hartwig-Breast_ER-positive_HER2-positive). In addition, PC2 had a high correlation with APOBEC activity connected COSMIC signature 2.

PC2 was enriched for a diverse set of cancer types such as Hartwig-Pancreas and Hartwig-Breast_Subtypeunknown and had high positive correlations with

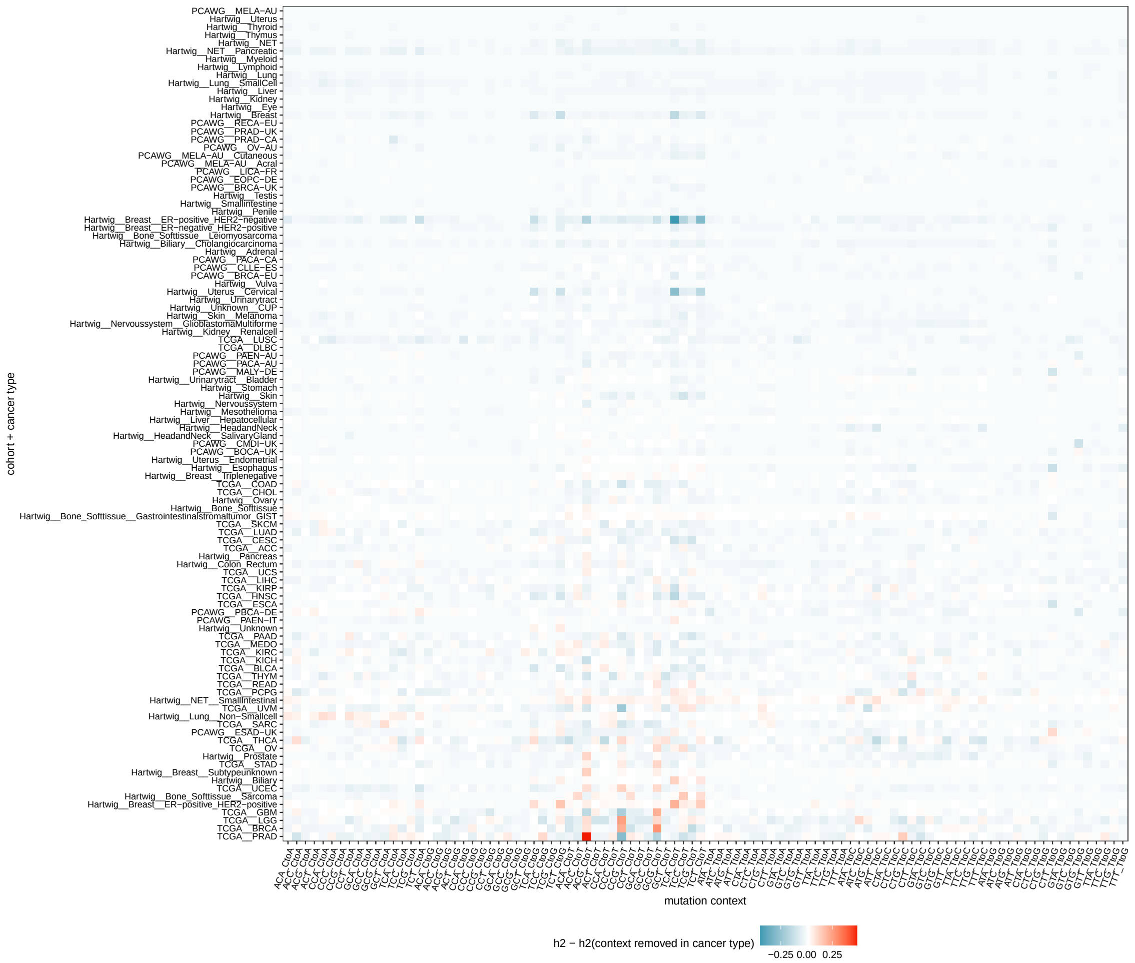


Figure 2.81: Change in total mutation burden heritabiity when dropping specific trinucleotide contexts from cancer types. Showing heritability difference between total mutation burden heritability and calculated heritability when dropping a specific SNV-trinucleotide context from a cancer type via GREML (color code). Different cancer types on y-axis (104 in total) and dropped SNV in trinucleotide context on x-axis (96 in total).

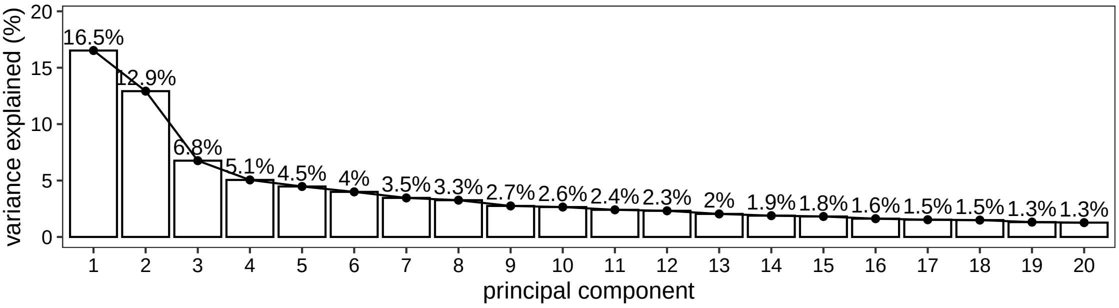


Figure 2.82: Principal component analysis on change in total mutation burden heriability estimates after dropping SNVs within trinucleotide contexts across cancer types. Showing the variance explained in % for the first 20 principal components.

C>T mutations at ACG, CCG, GCG, and TCG contexts. Expectably, clock-like COSMIC signature 1, which has been reported to correlate with age and the number of cells divisions in some tissues[238], had the highest correlation with this component.

PC3 had the highest contributions from Hartwig-Lung_Non-Smallcell and Hartwig-Liver_Hepatocellular. This component correlated with several COSMIC signatures such as signature 92 (tobacco smoking), signature 16 (unknown), signature 5 (unknown, clock-like), and signature 19 (unknown). It also had increased contributions from other lung and liver cancer. In total lung cancers contributed to this component by 17 % and liver cancers by 12 %. This would speculatively connect this PC to an environmental and/or metabolic process.

PC4 could be connected to COSMIC signature 17b. The underlying mutational process causing this signature has not been identified yet, but it has been speculated that this signature might be caused by increased levels of 8-oxo-dGTP via bile acids and/or gastric acids, which can lead to A>C transversions[239,240]. In accordance, PC4 had high contributions from signature 17 enriched cancer types such as Hartwig-Esophagus and TCGA-ESCA and in total esophagus cancers contributions to this component were around 20 %.

The remaining PCs could not be clearly connected to specific mutational processes and/or COSMIC signatures.

Taken together, we identified several mutational processes contributing to the heritability of the total mutation burden via dropping all possible SNV-trinucleotide contexts from each cancer type and estimating the change in the heritability. Via a PCA, we identified several components which could be connected to different mutational processes. We identified a component capturing APOBEC-directed mutagenesis (PC1), a component capturing cancer signature 1 (PC2), a component being enriched in lung and liver cancer types (PC3) and a component capturing signature 17b (PC4). These results would suggest that several mutational processes contribute to the heritability of the total mutation burden. In the future, higher sample sizes and analyses of individual cancer types will help to further obtain a more fine-grained view of which variants contribute to the heritability of which mutational somatic process.

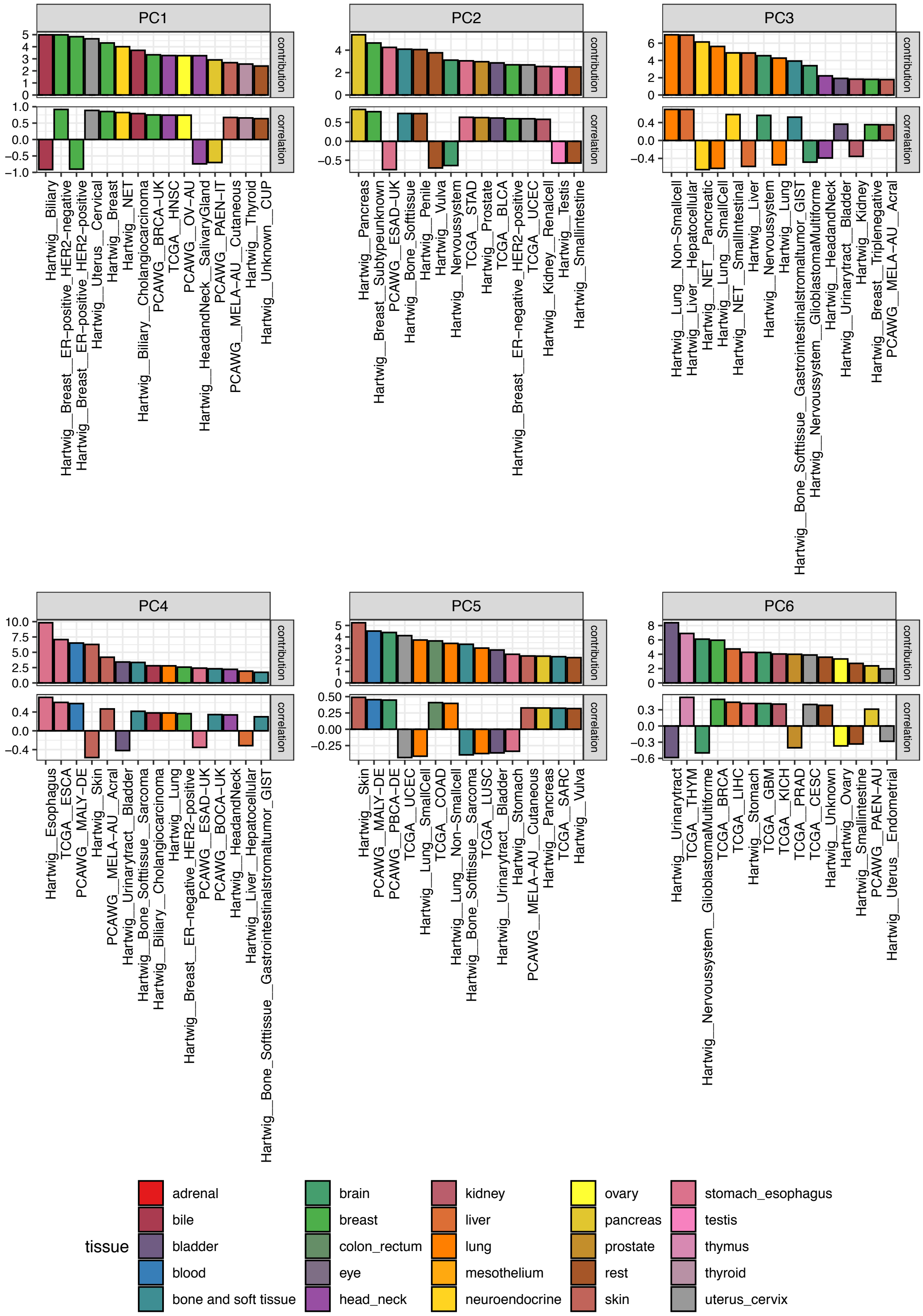


Figure 2.83: Overview of first 10 PCs and the corresponding 15 cancer types with the strongest contribution to the respective component. .Showing the contribution and correlation (y-axis) of each cancer type with the respective component (y-axis). Color code for tissue of origin.

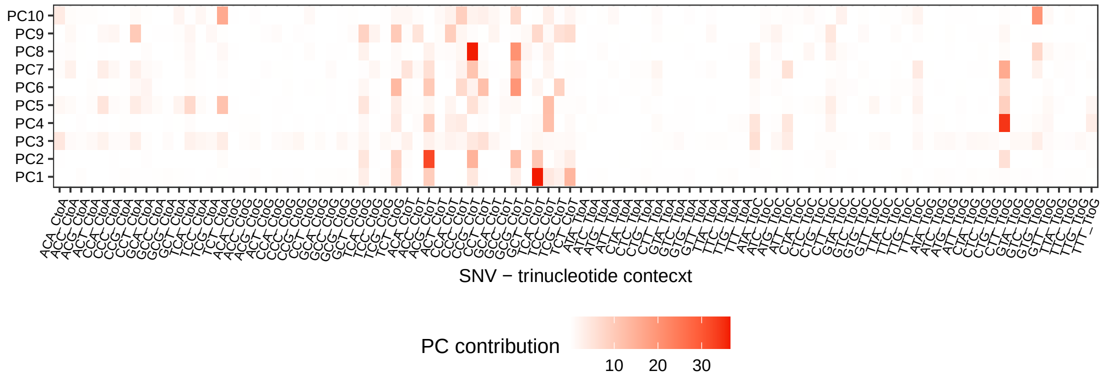


Figure 2.84: Overview of SNV-trinucleotide contexts and correlations with the first 10 PCs. Showing SNV-trinucleotide context on the x-axis, PC on y-axis and PC value encoded in color (0 equals white).

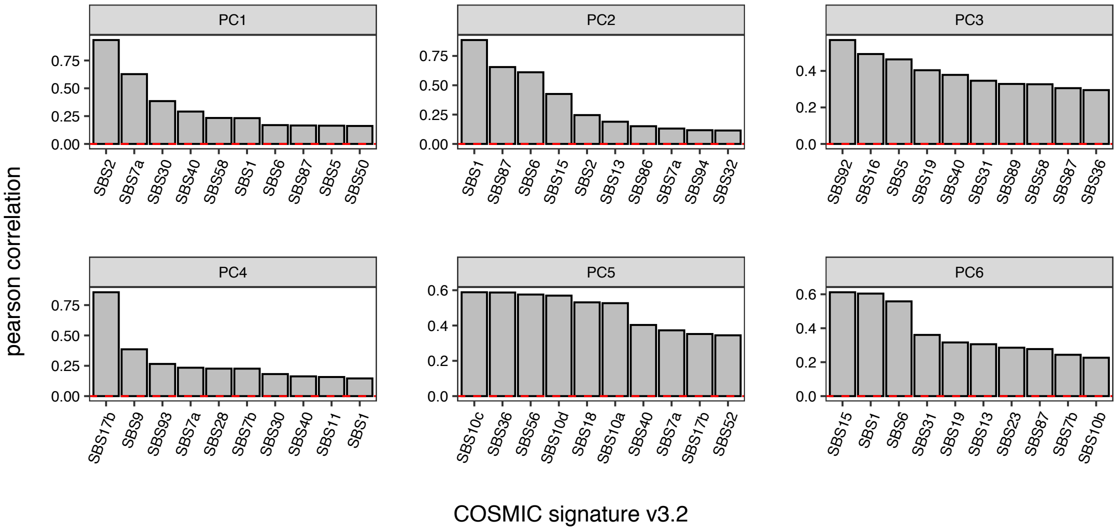


Figure 2.85: Pearson correlation between the 96 SNV-trinucleotide contexts contributing to the PCs and COSMIC signatures. Showing in each facet the 10 COSMIC signatures v3.2 (x-axis) with highest Pearson correlation (y-axis) to the respective component.

3 Discussion

In this study, the goal was to gain understanding of the role inheritance plays in shaping somatic mutational processes. Studying this question was encouraged by genetic screens from model organisms (yeast and bacteria) which showed that many different genes impact mutational processes[196,197]. Further, association studies utilizing rare and common germline variants to investigate this question showed how this approach would be applicable to find novel germline determinants of somatic mutational processes[8,32]. To comprehensively investigate this question, we extracted a set of mutational features covering a broad spectrum of somatic mutational processes from ∼15,000 tumor genomes. To increase interpretability, remove redundancy and increase statistical power, we applied different dimensionality reduction techniques to extract informative somatic mutational components (Section 2.1). These components were then used to answer the question whether inherited variants have an affect on a specific somatic mutational process. We performed for this purpose a gene-based rare variant association study (MAF < 0.1 %) (Section 2.2) and a genome-wide association study utilizing common variants (MAF > 1 %) (Section 2.3). Further, we examined to which extent somatic mutational processes contain a heritable component (Section 2.4). The main outcomes from this analysis will be further discussed in the next sections.

3.1 Extraction of Somatic Mutational Components capturing Somatic Mutational Processes

Extraction of informative somatic mutational features from whole-exome and whole-genome sequencing data

To extract somatic components covering different somatic mutational processes, somatic features based on previous reports were extracted based on SNVs[5,6,14,132], DNVs[14], indels[14], and copy number variants[17,18]. Further we also generated features incorporating relative mutation rates[144] with respect to different genomic properties such as replication timing[21], chromatin mark H3K36me3[15,20], the direction of DNA replication (leading vs lagging strand)[23,24], the direction of transcription (transcribed vs. untranscribed strand)[24], chromatin accessibility via DNase I hypersensitive sites[22], CTCF/cohesin binding sites[151,152], and the inactive X-chromosome[155]. Moreover, mutations occurring on the mitochondrial genome were counted[157,159] (Table 2.1).

Since many of these features were not extracted from WES data before, we tested whether the described effects in WGS data could also be seen in WES data. We showed that except for Ref.Sig. 11, the reported effects in WGS data were also replicated in WES data (Figures 2.5, 2.6, 2.10, 2.12, and 2.14). The missing effect between TMZ treatment and Ref.Sig. 11 (Figure 2.8) would most likely not be due to the sequencing technology since it was also not detected in WGS data (Figure 2.9). Associations in combination with *MGMT* promoter methylation or MMR inactivations, which were previously reported[27,199], were not replicated as well. Potential reasons could be missing patient information about TMZ treatment as it was suggested in ref[27], difficulties in assigning this signature to samples as it was reported in the same study or too small sample sizes (e.g. samples with *MGMT* promotor methylation + TMZ treated).

Furthermore, we showed that while the correlation was significant when comparing features extracted from the same patients from WES and WGS data (59 out of 65), the correlation coefficient was only higher than 0.8 for 11 out of 65 somatic features. In particular, somatic features which could only be measured from a limited number of loci such as CTCF/cohesin-binding site mutation peaks and replicative strand asymmetry had correlation coefficients smaller than 0.2 (Figure 2.22). Still, even here significant mutational differences between MSI and MSS samples were detected in several tissues (Figures 2.11 and 2.14).

Thus, while extracting somatic features from WES data will lead to less accurate estimations of features in comparison to extracting them from WGS data, global mutational processes such as APOBEC activity or a deficiency in DNA mismatch repair can still be detected. These results suggest that WES data is a viable source to investigate many different kinds of somatic mutational patterns, especially since openly available WGS data is still a limiting factor, and for association studies high sample sizes are required.

Extraction of somatic mutational components via ICA and VAE

Dimensionality reduction techniques were applied to deconvolve the 56 input somatic mutation features (Table 2.1) into 'somatic components'. The aim was to find components that are biologically relevant by capturing the underlying mutational process, reducing the dimensionality to increase statistical power, and increase interpretability. Basically, there were two factors to consider here: the algorithm and the number of components to extract. Selecting the 'best' algorithm and the optimal number of components is not trivial since most methods extract components by optimizing a function (e.g. neg-entropy in ICA), which does not necessarily result in biologically relevant representations.

In a study by Way *et al.*[267], gene expression data was compressed using five different compression methods including ICA and VAE and it was comprehensively tested with which algorithm and which number of component extractions the most number of biologically relevant components were extracted[267]. It was shown that there was no single best algorithm[267] and that even the same algorithm could capture different biological relevant components when using different numbers of components[267]. While these results would imply that using many different algorithms and including more components would increase the chances of capturing many different (aiming at all) mutational processes, this would also decrease the statistical power by increasing the burden of multiple testing correction, which would have been unfeasible in this study due to the main limitation coming from the sample size.

In consideration of the previous study[267], we started by testing three different methods (PCA, ICA, and VAE) and selected for the ICA the number of components optimizing the silhouette index (Figure 2.26) and for the VAE the number of components which optimized the correlation with four biologically relevant independent components (Figure 2.32). The ability of the components to capture mutational processes was assessed before performing the germline association studies by checking the somatic input features which had the highest correlations with the respective components (Figures 2.24, 2.28, and 2.34) and by investigating tissue enrichments (Figures 2.30 and 2.35) of the components, and comparing it to our current understanding of which features contribute to a specific mutational process[6, 14, 95, 21] (Section 1.3). Based on this examination, principal components were not further tested since we observed that often different mutational processes were grouped into one component with opposite signs (Figure 2.24). With the ICA and VAE-derived components 13 out of 29 components (5 for the ICA and 8 for the VAE) covered known, individual biologically relevant mutational processes such as UV-light exposure and tobacco smoking (Figure 2.40).

It can be also noted that we used existing signature detection tools[14,138] to extract mutational signatures for each mutation class separately and then utilized them as input next to other generated somatic features to extract components via ICA and VAE. This approach offers flexibility since any somatic feature could be included in the input matrix. In the past, mutational signatures have been extracted based on each mutational class separately (SNVs[5], indels[14], DNVs[14], clustered mutations[15,16], CNVs[17] and structural variants[18]) via NMF[133], updated versions of NMF[135,248] or via other approaches such as a probabilistic modelling approach[136], denoising sparse autoencoder neural networks[137], topic modelling[143], or mixture modelling[249]. More recent approaches have integrated different mutational classes to jointly learn mutational signatures and to characterize the different mutational processes more precisely[115,143]. These newer approaches could also be used in the future as an alternative approach to extract somatic mutational components, which combine different types of mutations and mutations occurring in different genomic regions together.

All in all, we showed here how ICA and VAE neural networks can be applied to extract biologically relevant mutational processes from high-dimensional data covering different mutational classes and including somatic features based on different genomic properties.

3.2 Rare Damaging Germline Variants Associating with Somatic Mutational Processes

We used the extracted somatic mutational components to test whether rare damaging germline variants impact any mutational process. We showed via a gene-based rare damaging germline variant association study that rare inherited variants in diverse genes associate with different mutational processes. Our approach incorporated a variance-based test via SKAT-O[184], two different dimensionality reduction algorithms to extract somatic mutation patterns, the usage of different *in silico* variant prioritization tools[185,272,271] and the use of different models of inheritance for association testing. This experimental design allowed us to identify multiple new replicating associations between genes and somatic mutation phenotypes.

Most of the associations we identified were replicated only via the variance-based test SKAT[183], which suggests that variants predicted to be damaging still contain many non-causal variants. More accurate variant effect prediction tools should further increase the power of these kinds of analyses[204,189,241]. We also found that using the two techniques to derive informative somatic mutation components (ICA and VAE) identified more replicated associations than using either approach alone. This is consistent with findings in other fields, where different algorithms have also been found to capture complementary information, for example in gene expression analysis[267] and genetic variant analysis from genomic data[242].

We identified novel genes associating with dHR-related repair (e.g. *RIF1*, *PAXIP1*, *WRN*, *EXO1*, and *ATR*) and with components connected to dMMR (e.g. *MTOR*, *TTI2*, *SETD2*, *EXO1*, *MSH3*, and *MLH3*). Several novel associations are supported by strong evidence from prior studies such as *EXO1* with dHR[217,218] and dMMR[243,216,126], *SETD2* with dMMR[20,15] and MSH3 with a different form of dMMR[126,52,220]. On top of the associations with dHR- and dMMR-related components, we also identified an association of *APEX1* with APOBEC-directed mutagenesis (as well as dHR), and additionally several genes associating with a component enriched in brain and liver cancers with an unknown underlying mechanism. 'Guilt by association' network analysis has not yet been widely adopted in rare variant association studies but we found that it was useful for both connecting high stringency replicating genes to each other and for connecting lower confidence hits to the high confidence genes. These interactions are useful for prioritising the newly associated genes and provide specific hypotheses connecting to known germline mutator genes.

Interestingly, the genetic associations distinguish between two different dMMR mutational phenotypes. Firstly, the common dMMR signature, enriched for 1 bp indels and the SNV-signature MMR1; these associations involved e.g. the Lynch syndrome genes *MSH2* and *MLH1*, and some additional genes e.g. *MTOR*, and *SETD2*. Secondly, a distinct set of associations involved a mutational component enriched for 2 bp and longer indels, but did not encompass a notable increase in SNVs, e.g. involving the core MMR gene *MSH3*, and additionally *MLH3*, *EP300*, and *PIK3C2B*.

Our findings support observations from genetic screens in model organisms suggesting that mutational processes can be affected by variation in diverse genes[196,197]. In particular, in a previous study heterozygous mutations in diverse genes in yeast were reported to cause genetic instability[196]. Out of the 50 identified human homologs in that study[196], RDGVs in 3 genes (*MSH2*, *MDN1*, *MTOR*) also associated in our study with a somatic component at a FDR of 1 % and 6 genes (*MSH2*, *MUS81*, *MDN1*, *MTOR*, *NUDT7*, *RBBP6*) at a FDR of 2 %. Further, in a different screen performed in bacteria, 284 human homologs were reported to cause DNA damage when upregulated[197]. Out of the 284 genes[197], 42 of them were also tested in our study, and RDGVs in 7 genes (*MSH2*, *MSH3*, *REV3L*, *DNMT1*, *RECQL*, *SMC2*, *TOP2A*) also associated in our study with a somatic component at a FDR of 1 % and 12 genes (*ASCC3*, *FANCM*, *MSH2*, *MSH3*, *REV3L*, *WRN*, *DNMT1*, *RECQL*, *SMC1A*, *SMC2*, *TOP2A*, *TOP3B*) at a FDR of 2 %.

This study has some limitations resulting from technical factors. The design is likely to result in a conservative bias in the number of replicated hits, because the discovery and validation cohorts were based on different sequencing technologies (WES versus WGS, respectively). WES data yields more noisy somatic mutation features, as it covers ∼2 % of the genome and some features (e.g. replicative strand asymmetry, mutations at CTCF/cohesin binding sites) are measurable at few loci (Table 4.1) and so enrichments are difficult to estimate due to low mutation counts. Moreover the power to call germline variants at certain loci may be different for WGS and WES data. The TCGA WES data also has batch effects originating from the different sequencing centers and sequencing technologies[244,245]. To offset this risk, we only extracted germline variants from regions with enough coverage in each of three sequencing centers as previously shown[167]. This limited the number of RDGVs, and thus potentially also the number of discoveries.

In order to increase the sample size and thus power, we combined the cancer cohorts that contained both primary[7,8] and metastatic[9] cancers, as well as treatment-naive and pretreated. Similarly, in the pan-cancer analyses, we aggregated data from all cancer types, with the result that the distribution of cancer types between the discovery and validation cohort was somewhat different. It is possible that some hits did not replicate due to these differences in cancer type composition.

Our initial set of somatic mutational features was largely motivated by recent reports[5,6,14,20,21,22,23,24,94,95,132,144,149,151,155,157]. Consideration of additional, complementary features could identify additional associations in future studies. Lastly, our analysis was performed on samples with European ancestry since this was the most numerous group and including sequencing data from more diverse populations is also likely to identify additional associations.

In the future, larger sample sizes with WGS data and better variant pathogenicity prediction tools will enable higher-powered association studies. Recent advances in the field of variant pathogenicity prediction include EVE (evolutionary model of variant effect), which is based on a deep generative model (VAE) and only needs the multiple sequence alignment of the protein of interest as input[189]. Another new tool is based on an autoregressive generative model, which does not need any alignment as input and is even able to predict indels[190]. These advances will further help elucidating the potentially very numerous set of genes which determine human somatic mutation rates. The identification of additional genes altering human mutation processes may have important implications for understanding, preventing and treating cancer and other somatic mutation-associated disorders.

3.3 Common Germline Variants Associating with Somatic Mutational Processes

After showing how rare damaging germline variants in many different genes can have an impact on somatic mutational processes, the effects of common germline variants on the same somatic components were investigated. Previously, a common deletion polymorphism in the coding region of *APOBEC3B*, altering APOBEC-signature mutagenesis, was discovered and replicated in several studies[111,112]. In addition, in the PCAWG study, another nearby but independent SNP was identified which affected APOBEC mutagenesis[8]. In another pan-cancer study, which did not include a validation cohort in the study design, no genome-wide significant association between common variants and total mutation burden were identified, but several hits in specific cancer types such as breast and stomach cancer were reported[191]. Furthermore, in a smaller study, which focussed on breast cancer, several associations between common variants and APOBEC-directed mutagenesis were identified. While none of the hits replicated in the validation cohort, several hits had matching effect size directions in the validation cohort[32].

Thus, to date, the association between common variants and somatic mutational processes has only been performed with a few somatic processes. Here, we identified 10 associations at genome-wide significance in the discovery cohort covering 4 different cancer types (including pan-cancer) and 9 different somatic mutational features (Table 2.2). None of the hits replicated in the validation cohort after correcting for multiple testing. One hit had a p-value < 0.05 and 5 out of 10 hits shared the same effect size direction as in the discovery (Table 2.2). In addition, none of the ten hits were previously reported as cancer risk SNPs in the GWAS catalog from EMBL-EBI.

Consequently, the impact of common germline variants on somatic mutation processes remains inconclusive. There could be several reasons, apart from the technical limitations which were already described for the rare variant association study (e.g. composition of cohorts, inaccuracies in estimating somatic features from WES; Section 3.2). Common germline variants were identified via different technologies. In the discovery cohort, common germline variants were called from SNP-array data, while in the validation cohort WGS data was utilized. A PCA performed on both cohorts together including 1000 genomes data showed that while all cohorts overlapped with each other in the first PCs, the PC scores from samples in which germline variants were called via SNP-array data were much wider compared to the ones which were extracted from WGS data (Figure 2.68). Further, GWAS hits usually have small effect sizes[173] and this also holds true for GWAS hits which were reported to alter cancer risk[246]. Small effect sizes require high sample sizes to be detected and our power analysis showed that even in pan-cancer this study was underpowered to detect small effect SNPs. Thus, potential small true effects could not have been detected in this study. Fourthly, genetic architectures vary between cancer types[246], and thus, one could expect that cancer-specific associations would not be detected in the pan-cancer analysis due to the lacking effects in the other cancer types. Further, they would also not be identified in the cancer-specific GWAS due to the low sample sizes of the individual cancer types (Table 4.7).

All in all, several associations between common germline variants and different mutational processes were identified, but none of them replicated in an independent cohort. The most likely explanations would be low sample size (leading to an underpowered study) and differences in sequencing technology (SNP-arrays vs. WGS) and component extractions (WES vs. WGS). In the future, larger sample sizes will be required to perform association studies like these in pan-cancer and across all cancer types to further investigate to which extent inherited common germline variants affect somatic mutational processes.

3.4 Heritability of Somatic Mutational Processes

While we did not find many common variants affecting mutational processes, we further asked to which extent the different mutational processes would be heritable. If the different mutational processes would be polygenic, as it was suggested for the total mutation burden[191], it would be reasonable to expect that many SNP contribute to the overall heritability by a tiny fraction, which is why they would not be detected in the genome-wide association study (since very small effect size), but in the SNP-heritability estimate. If a somatic feature or somatic mutational component had an increased heritability it would suggest a genetic cause or genetic contribution on the corresponding feature/component.

To date, only the heritability of the total mutation burden has been estimated in a pan-cancer analysis and has been calculated to be around 13 % (Table 1.6). The heritability of APOBEC-directed mutagenesis, C>T mutations at CpG sites and dHR has been only estimated in a breast cancer cohort with high errors (> 20 %) due to the low sample size. Thus, the heritability of the different mutational processes has not been investigated on a considerable sample size before.

Here, we calculated SNP-heritabilities by two different approaches. Firstly, heritabilities of the 65 somatic mutational features and of the 29 extracted somatic components were calculated across the two cohorts (TCGA and PCAWG_Hartwig) and compared. There were high differences in the heritability estimates between the two cohorts (R <0.2). Still, two mutational somatic features had SNP-heritabilities above 8 % in both cohorts: signature Ref.Sig. MMR1 and the total number of C>A mutations (Figure 2.74). In a previous meta-analysis of all reported SNPs contributing to the risk of colorectal cancer, several SNPs within DNA mismatch repair genes such as *MLH1* and *MSH3* were identified[176]. Thus, speculatively common variants in DNA repair genes resulting in inefficiencies of the DNA mismatch repair machinery could explain the increased heritability of somatic signature Ref.Sig. MMR1 mutations. Heritability of the total number C>A mutations could be connected to many different mutational processes. One of them could be tobacco smoking since it dominantly results in C>A mutations[6] and also in a previous study a genetic correlation between smoking initiation and the total mutation burden was identified[191]. Another potential mechanism which predominantly creates C>A mutations in the genome would be oxidative damage[124] (signatures 18[6,126] and 36[14,125]). The total mutation burden itself also had increased heritabilities in both cohorts but to a lesser extent than it was reported before with approximately 6 %. Further, APOBEC signatures, which were previously predicted to have a heritability of $>20\,\%$ in a breast cancer cohort (sample size ≈ 700)[32], had a heritability of 0 % in our pan-cancer study for both APOBEC signatures (Ref.Sig.2 & Ref.Sig.13) via both tools.

Secondly, individual mutation types in their trinucleotide context (e.g. C>T in CCC context) were dropped from the total mutation burden from a specific cancer type and the difference in the SNP-heritability was calculated. With the second approach, the idea was to find the mutational processes contributing to the heritability of the total mutation burden. A PCA was performed on the dataset and it was investigated which cancer types contributed the strongest (Figure 2.83) and which signatures had the highest correlations (Figure 2.85) with the first PCs. At least three different mutational processes were identified, which contributed to the heritability of the total mutation burden: APOBEC mutagenesis, cancer signature 1 (correlating with age and number of cell divisions across many tissues[238]), and signature 17b (most likely due to oxidative damage in the nucleotide pool[239,240]). Another PC captured a mutational process with the strongest contributions coming from lung and liver cancers, which could speculatively point to an environmental factor or metabolic process. The contexts with the strongest contributions to this PC were T>C in ATA (6.1 %), C>T in CCT (5.8 %), and C>A in CCT (4.5 %), resulting in a positive correlation with COSMIC signature 92 (tobacco smoking) and signature 16 (unknown, enriched in liver cancer among others).

The main limitation of this analysis was the sample size. Heritability can't be robustly measured at sample sizes below 1,000 since on average the errors of the SNP-heritability estimates via GREML scale by 318/n[191] (n equals the sample size), which has not made it possible to measure heritability in the respective cancer types separately. This would be an important analysis in the future since different mutational processes drive different cancer types[237] and the genetic architecture has also been reported to be different across cancer types[246]. Thus, heritability estimates from the pan-cancer analysis can be regarded as averages and it would be highly likely that specific processes have higher or lower heritabilities in individual cancer (sub)types. Moreover, even in the pan-cancer analysis the heritability estimates had high errors. The high errors in the estimation could be one reason why the heritability estimates between the two cohorts had low correlations. Another highly likely reason could be the low correlation between WES- and WGS-extracted somatic features (Figure 2.22). Other limitations are similar to the ones discussed before (Sections 3.2 and 3.3), such as differences originating from different sequencing technologies (SNP-arrays vs. WGS).

All in all, we showed that despite limitations originating mostly from the low sample sizes, cancer signature Ref.Sig. MMR1, the total number of C>A mutations and the total mutation burden have a heritable component. Furthermore, heritability of the total mutation burden could be attributed to at least three different mutational processes (APOBEC, signature 1, and signature 17b). In the future, higher sample sizes will make it possible to accurately estimate the heritability of different somatic mutational processes across cancer types. This will be an important step to better understand the factor inheritance plays next to environmental factors in gaining somatic mutations, which could possibly lead to tumorigenesis or other somatic mutation-associated disorders. In addition, this approach shows how estimating heritability of somatic mutational features or of traits in general can be utilized to infer a genetic cause/contribution.

4 Methods

4.1 Extraction of Somatic Mutational Features

4.1.1 Data Sources in the Discovery Cohort

For the somatic features which were based on single nucleotide variants (SNVs), double nucleotide variants (DNVs), and insertions or deletions (indels), the somatic calls from the MC3 Project[250] were used. For the somatic features based on copy number alterations (CNAs), TCGA exome data was downloaded from the GDC Data Portal[252] and processed as described in ref[253]. Copy numbers were identified with the tool FACETS[251]. The tool used as input data the BAM file of the tumor sample, the BAM file of the sample-matched normal sample, and a vcf file of common human SNPs. Furthermore, 93 individuals, which were reported to be positive for human papillomaviruses in head and neck cancer samples[254], were excluded from the analysis. In total, this yielded somatic calls from 10,033 individuals.

4.1.2 Data Sources in the Validation Cohort

Mutation calls for PCAWG[8] were obtained from the ICGC data portal. Somatic mutation calls and copy number calls were obtained from the DKFZ/EMBL variant call pipeline. All samples were downloaded except for ESAD-UK, MELA-AU and all project id's ending with '-US' in order to prevent an overlap with the discovery cohort. In total, samples from 1,662 donors were downloaded. In short, single nucleotide variants were called via samtools[255] and bcftools 0.1.19[256], and indels were called via Platypus 0.7.4[257]. Copy number alterations were estimated with ACEseq v1.0.189[258] (Supplementary information in PCAWG flagship paper[8]). Data access to the estimated somatic nucleotide variants and copy number variants from Hartwig were acquired as well (`https://www.hartwigmedicalfoundation.nl/en/`), making up 3,613 samples in total. In Hartwig nucleotide variants were called with Strelka[259] 1.0.14 and copy number alteration with the Purple tool[9]. BAM files for the melanoma dataset MELA-

AU (dataset ID: EGAD00001003388; 183 individuals) and the esophagus dataset ESAD-UK (dataset ID: EGAD00001003580; 303 individuals) were downloaded from the European Genome-Phenome Archive (EGA). Somatic mutations were called via Strelka[260] 2.9.10 and copy number alterations were extracted as described above with the tool FACETS[251].

4.1.3 Data Sources to check Overlap between WES and WGS Features

To check how somatic features changed when extracting them from WES data compared to WGS data, we also downloaded WGS-based somatic calls from TCGA, which were generated within the PCAWG project[8]. Somatic calls were obtained from the ICGC portal in the same way as it was performed for the other PCAWG samples. It should be noted that these samples were not used for any other analysis in order to keep the discovery and validation cohort independent of each other.

4.1.4 Further Processing of Somatic Calls

For all datasets, regions which are known to be difficult to be aligned were excluded as well as regions which have been blacklisted by the UCSC Genome Browser[261]. As described previously[21,15] blacklisted regions by Duke and DAC were removed and the CRG75 alignability track was applied to only keep regions where 75-mers in the genome can be uniquely aligned in the human reference genome hg19.

4.1.5 Single Nucleotide Variants - Total Mutation Counts

Based on the number of SNVs in the nuclear genome, 8 different somatic mutational somatic features were estimated: the total number of SNVs, the number of C>A substitutions, the number of C>G substitutions, the number C>T substitutions in regions where the 3' flanking site was not a G (non CpGs), the number of C>T substitutions in regions where the 3' flanking site was a G (CpGs), the number of T>A substitutions, the number of T>C substitutions and the number of T>G substitutions. The number of C>T substitutions was divided into two groups (at CpG sites vs. non-CpGs sites) due to the effect of CpG sites on mutation rates (due to DNA methylation)[103]. A pseudocount of 1 was added to each somatic mutational feature and all features were log transformed to the base 2.

4.1.6 Single Nucleotide Variants in Mitochondrial DNA - Total Mutation Counts

As other studies have pointed, WES data can be used to extract mutations occurring in the mitochondrial DNA, due to the large amount of off-target reads[157, 158]. The coverage file of each sample was used to estimate to which extent the mitochondrial genome in each sample was sequenced. Only samples in which at least 50 % of the mitochondrial genome were covered by at least 4 reads were kept for further analysis. Furthermore, following a previous study[157], only variants were kept which had an allele frequency of at least 3 % in order to remove potential false-positive calls. For the cancer cohorts Hartwig, ESAD-UK and MELA-AU, which were all based on WGS data, somatic variants in the mtDNA with a frequency of less than 3 % were filtered out as well. After filtering, the total number of SNVs in the mtDNA in each sample was calculated. For PCAWG, mutation calls on the mitochondrial genome were downloaded from the respective study (`https://ibl.mdanderson.org/tcma/mutation.html`)[159]. At last, a pseudocount of 1 was added to each individual and the feature was log transformed to the base 2.

4.1.7 Single Nucleotide Variants - NMF-derived Organ-specific Signatures

First of all, the python tool SigProfilerMatrixGenerator[262] was used to generate for each dataset a matrix counting all mutations in the 96 possible trinucleotide contexts by considering the adjacent 5' and 3' base of the somatic variant (16 trinucleotides for each single nucleotide variant). Next, exposures of the organ-specific signatures, which were derived in the work of Degasperi *et al.*[138] were assigned to each sample as described in ref[138]. Organ-specific signature exposures were estimated by selecting for each sample the respective organ-specific signature set based on the tissue it was derived from. In cases in which no organ-specific signature set was existing due to its low sample size (e.g. mesothelioma, thymoma, penile, and vulva), the reference mutational signature set was used. In short, this aims to only fit signatures to a sample which were also identified in the according tissue. The tool uses a bootstrap-based method to only assign signatures to a sample when they reach a specific threshold ($p < 0.05$), otherwise they are set to 0. The goal of this approach is to decrease the probability of overfitting and miss-assignment of signatures[138]. In the discovery cohort the median fraction of unassigned mutations was 47 % and in the validation cohort 15 %, which is likely due to the low number of somatic mutations in the discovery cohort. To have a common set of signatures, all signature exposures were then converted to the reference signature set via the conversion matrix provided in ref[138]. For further analysis we only kept 17 signatures, which had in the discovery and in the validation cohort an activity of > 5 % in at least one matching cancer type or in the pan-cancer analysis 'pancan': Ref.Sig.1, Ref.Sig.2, Ref.Sig.3, Ref.Sig.4, Ref.Sig.5, Ref.Sig.7, Ref.Sig.8, Ref.Sig.11, Ref.Sig.13, Ref.Sig.17, Ref.Sig.18, Ref.Sig.19, Ref.Sig.22, Ref.Sig.30, Ref.Sig.33, Ref.Sig.MMR1 and Ref.Sig.MMR2. A pseudocount of 1 was added and each estimated signature count was log transformed to the base 2.

4.1.8 Single Nucleotide Variants - Transcriptive Strand Bias

To estimate this strand bias[24], the number of mutations occurring on the untranscribed strand and on the transcribed strand were calculated. This was performed by the python tool SigProfilerMatrixGenerator[262]. Based on the six possible base substitutions, six different somatic features were generated (C>A, C>T, C>G, T>A, T>C, T>G). For each one, the number of base substitutions occurring on the untranscribed strand were divided by the number of mutations occurring on the transcribed strand. A pseudocount of 1 was added to the numerator and denominator before division and the resulting quotient was log transformed to the base 2.

4.1.9 Single Nucleotide Variants - Replicative Strand Bias

To estimate this strand bias[23], replication timing data from lymphoblastoid cell lines was downloaded (`http://mccarrolllab.org/resources/`)[263]. The fork polarity, which is a derivative of the replication timing estimate, was estimated as described by Seplyarskiy *et al.*[108]. In brief, the slope/derivative at each coordinate of the replication timing landscape was calculated by considering the region approximately ± 5 kb of the coordinate. The fork polarity value reflects whether the reference strand is more likely to be replicated as the leading strand (fork polarity > 0) or as the lagging strand (fork polarity < 0). Next, the genome was divided into equal sized bins of the length of 10 kb and the average fork polarity in each bin was calculated. Further, the whole genome was split into 10 equal sized bins. To calculate the replicative strand bias, we only considered the two lowest bins (reference strand more frequently replicated as the lagging strand) and the two highest bins (reference strand more frequently replicated as the leading strand). From the perspective of the reference strand, we divided the total number of T>C, T>G, G>A, and C>A mutations occurring on the leading strand by the total number of T>C, T>G, G>A, and C>A mutations occurring on the lagging strand. This would mean for instance that a A>G mutation occurring on the leading strand was counted as a mutation occurring on the lagging strand (since T>C on the other strand). We focused on these four mutation types since replicative strand biases have been previously reported for these in connection with a deficiency in DNA mismatch repair[149]. This feature was only calculated in samples, in which at least 20 of the 4 single substitutions types were counted within the covered region. The estimated values were log transformed to the base 2.

4.1.10 Single Nucleotide Variants - X-Chromosomal Hypermutation

For generating a somatic mutational feature for X-Chromosomal hypermutation[155], first of all the total number of single nucleotide variants per MB on each chromosome was counted. Next, the number of mutations per MB occurring on the X chromosome was divided by the average number of mutations per MB occurring on the autosomes. A pseudocount of 0.1 was added to the numerator and denominator before division and the resulting quotient was log transformed to the base 2.

4.1.11 Single Nucleotide Variants - CTCF/Cohesin Binding Sites

CTCF/cohesin binding sites are often mutated in cancer[151,152]. To capture this somatic mutational feature, we counted the number of single nucleotide variants occurring in CTCF/cohesion binding site and divided them by the number of mutations occurring in the flanking site (± 500 bp) of the binding site. CTCF/cohesin binding sites were obtained from Roadmap and averaged over 8 cell types[264]. Genomic regions, that were bound by CTCF in at least one cell type and by cohesin in at least two cell types were set as CTCF/cohesin binding sites. All sites ± 500 bp of the sites that were bound by CTCF in at least one cell type were set as the flanking site. Length of covered genomic regions can be found in Table 4.1. This somatic feature was only estimated in samples which had at least 10 SNVs counted in total within the CTCF/cohesin binding and/or flanking site. At last, we were able to calculate the CTCF somatic feature for 38 % of the samples in the discovery cohort and 98 % of the samples in the validation cohort. The ratio was log transformed to the base 2.

4.1.12 Extraction of Genomic Region Densities of Expression, Histone Mark H3K36me3, Replication Timing and DNase I Hypersensitive Sites

Features measuring mutation rate variation with regards to expression, histone mark H3K36me3, replication timing, and DNase I hypersensitive sites were calculated using negative binomial regression in order to reduce the correlation of these features with each other and to control for mutation substitution types. For this purpose, regional data from a previously published study[15] was used. In brief, levels of histone mark H3K36me3 (averaged over 8 cell types) and DNase I hypersensitive sites were downloaded from Roadmap Epigenomics[264]. Genomic regions with no signal for the corresponding feature were set as 'bin 0' and the remaining genomic regions were split into 5 equal-sized bins with increasing signal. In this way, genomic regions with the highest amount of histone mark H3K36me3 were put into bin 5, regions with the lowest amount into bin 1 and regions with no signal into bin 0. Replication timing information was derived from the ENCODE project using the average over 8 cell lines. Genomic regions were split into 6 equal sized bins, where bin 1 corresponded to the latest replicating region and bin 6 to the earliest replicating region. Expression levels were based on RNA-seq data, which was obtained from Roadmap[264] and averaged over 8 cell types as well. Bin 0 represented regions with no expression (RPKM = 0) and the remaining 5 bins were split equally by increasing expression levels. All these genomic masks were further processed by applying the CRG75 alignability track. For the whole- exome sequencing data specifically, the masks were intersected with the coverage mask from the MC3 project, since the somatic WES mutation calls were derived from there. Furthermore, the 4 masks (expression, histone mark H3K36me3, replication timing, and DNase I hypersensitive sites) were intersected with each other for the subsequent regression. Several bins extracted from the whole exome mask covered only a small region in the genome (< 5 MB), which was expected since the exomic regions in the genome are known to be enriched for early replicating regions and histone mark H3K36me3. Since we observed that the regression often failed when bin sizes were too small, some bins were merged: replicating timing bins 1 and 2, histone mark H3K36me3 bins 1 and 2, expression bins 0 and 1, and DNase I hypersensitive site bins 1 and 2. This step was not performed for the whole-genome masks since the covered regions for each bin were big enough. Length of covered genomic regions can be found in Table 4.1.

4.1.13 Single Nucleotide Variants - Mutation Enrichment Calculations with regards to Expression, Histone Mark H3K36me3, Replication Timing and DNase I Hypersensitive Sites

The individual features corresponding to the enrichment of mutations in a particular genomic region were calculated via negative binomial regression using the function *glm.nb* from the R package MASS (version 7.3.53) in R 3.5.0. The regression was performed for the different features in each tumor sample as follows:

(i) mutation count $\sim$ replication timing + mutation type + offset

(ii) mutation count $\sim$ replication timing + DNase + mutation type + offset

(iii) mutation count $\sim$ replication timing + expression + mutation type + offset

(iv) mutation count $\sim$ replication timing + H3K36me3 + mutation type + offset

In the discovery cohort (WES only) the mutation type variable had 7 possible encodings (C>A, C>T at CpG sites, C>T at non-CpG sites, C>G, T>A, T>C and T>G), and in the validation cohort (WGS only) the mutation type variable encompassed all 96 possible substitutions within the trinucleotide context (e.g. C>A mutation within ACA context). The offset represents the nucleotide-at-risk and is the natural log of the number of nucleotides covering the respective region. As described previously[15], the coefficients obtained from the regression for the different genomic regions represent the log enrichment of mutations in each bin in comparison to the reference bin. For replication timing, the latest replicating bin was set as the reference, for expression the lowest expressing bin was set as the reference and for histone mark H3K36me3 and DNase I hypersensitive sites the bins with no signal were set as the reference. This would mean that for instance the coefficient obtained from regression (iv) for bin 5 from the histone mark H3K36me3 variable describes the log enrichment of mutations in regions with a high signal of this histone mark in comparison to regions with no histone mark signal, while controlling for replication timing and the mutational context. In this way we aimed to control for the correlation of expression levels, histone mark H3K36me3 and DNase I hypersensitive sites with replication timing and the mutational context. Especially, for WES data this approach was limited by the reduced covered genomic region and the decreased number of mutations in comparison to whole-genome sequencing data. The regression was only performed in samples, that had at least 30 SNVs counted. The coefficient obtained in regression (i) for the earliest replicating bin was extracted for the replication timing (RT) feature, the coefficient obtained in regression (ii) for the bin with the highest amount of signal in DNase I hypersensitive sites was extracted for the DNase I hypersensitive site (DNase) feature, the coefficient obtained in regression (iii) for the bin with the highest expressing regions was extracted for the expression (Expression) feature, and the coefficient obtained in regression (iv) for the bin with highest amount of signal in histone mark H3K36me3 was extracted for the H3K36me3 (H3K36me3) feature. High errors in the regression coefficients (standard error > 100) indicated that the regression failed to converge for the corresponding coefficient and thus, were removed. In the discovery cohort, 7,650 RT coefficients, 7,684 H3K36me3 coefficients, 7,471 DNase coefficients and, 7,664 Expression coefficients were extracted in total. In the validation cohort, 5,759 RT coefficients, 5,749 H3K36me3 coefficients, 5,752 DNase coefficients and, 5,759 Expression coefficients were extracted in total.

4.1.14 Double Nucleotide Variants - NMF-derived Signatures and Fitting

Double nucleotide variants were extracted with the python tool SigProfilerMatrixGenerator[262]. The tool counted the occurrence of 78 double nucleotide variants (AC, AT, CC, CG, CT, GC, TA, TC, TG, or TT to NN). The matrix was used as an input to extract Double Base Substitution (DBS) signatures using the python tool SigProfilerExtractor[14]. In brief, the tool uses non-negative matrix factorization (NMF) to extract mutation signatures. Since the exact number of mutation signatures is not known, the tool extracted 1 to 25 signatures. For each signature extraction 100 iterations were performed adding poisson noise to the samples during each iteration. For the discovery cohort the optimal solution were 3 signatures and for the validation cohort 11. Next, the tool fitted the established DBS signatures from COSMIC[4] v3.2 to the extracted de-novo signatures. Then, signature exposures were estimated by fitting the extracted COMISC signatures to each sample. In the discovery cohort the COSMIC[4] DBS signatures DBS1, DBS2, DBS4, DBS9 and DBS10 were extracted and in the validation cohort the DBS signatures DBS1, DBS2, DBS4, DBS5, DBS6, DBS7 and DBS9 were extracted. The 4 DBS signatures which were found in both cohorts were kept for association testing: DBS1, DBS2, DBS4 and DBS9. Next, a pseudocount of 1 was added to each estimated signature exposure and each estimated exposure was log transformed to the base 2.

4.1.15 Insertions and Deletions - Total Mutation Counts

Different insertion and deletion somatic mutational features were generated. First of all, the total number of indels occurring in each sample was counted. Next, the number of indels in microsatellite (MS) regions was counted due to its frequent occurrence in sample with deficient mismatch repair[6,265]. For this purpose, the number of indels with a length of 1 bp and the number of indels with a length of 2 to 5 bp were counted within and outside MS regions. MS locations were identified via the tandem repeat search tool Phobos (`https://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm`). Next, the total number of indels with a length of 6 to 10 bp was counted. Due to the low number of indels of this length, especially in whole-exome sequencing data, this feature was not further split into MS vs non-MS regions. Furthermore, since deletions have often been reported to be predictive of deficiency in homologous recombination[95], different deletion features were created. The total number of deletions with a length of bigger than or equal to 10 bp was created. Also, the number of deletions at flanking microhomology sites of either 1 bp or more than 1 bp was counted by using the output matrix from the python tool SigProfilerMatrixGenerator[262]. A pseudocount of 1 was added to each feature and each feature was log transformed to the base 2.

4.1.16 Insertions and Deletions - NMF-derived Signatures and Fitting

Small insertion and deletion (ID) signatures were extracted in the same way as described for the DBS signatures. For the discovery cohort the optimal solution were 4 signatures and for the validation cohort 10. The COSMIC[4] ID signatures were fit to the de-novo signatures and in the discovery cohort COSMIC[4] ID signatures ID2, ID3, ID4, ID7, ID8 and ID15 were extracted and in the validation cohort ID signatures ID1, ID2, ID3, ID4, ID5, ID6, ID8, ID9, ID10, ID12, ID13 and ID14 were extracted. The 4 ID signatures which were found in both cohorts were kept for further association testing: ID2, ID3, ID4 and ID8. Next, a pseudocount of 1 was added to each estimated signature exposure and each estimated exposure was log transformed to the base 2.

4.1.17 Copy Number Variants - Total Mutation Counts, Ploidy and Whole Genome Duplications

Copy number based features were generated by splitting amplification and deletion events by different sizes. The number of amplifications with a size of 1 to 10 kb, 10 to 100 kb, 100 to 1000 kb, and bigger than 1000 kb were counted. Similarly, the number of deletions with a size of 1 to 10 kb, 10 to 100 kb, and bigger than 100 kb were counted. Next, a feature was generated based on the estimated ploidy of the tumor sample from the corresponding copy number detection tool. The number of whole genome duplication (WGD) events were calculated by dividing the ploidy by 2 via integer division. A pseudocount of 1 was added to the amplification and deletion based features, a pseudocount of 0.1 was added to the WGD feature and no pseudocount was added to the ploidy feature since ploidy can never be 0. At last, each feature was log transformed to the base 2.

4.1.18 Principal Component Analysis

For the PCA all somatic features described above were used except for the following 9 somatic features: total number of SNVs, total number of indels and total number of the 7 different single mutation substitutions types. These were excluded since they were already represented by the different NMF- derived signatures. Further, all samples were removed in which 20 % of the features were not estimated due to low mutation counts. Thus, 9,235/9,425 samples were left in the discovery cohort and 5,597/5,613 samples were left in the validation cohort. Next, missing values were replaced by the median value of the respective columns and each feature was centered and standardized to a mean of 0 and standard deviation of 1. This step was performed for the somatic features, which were extracted from three different cohorts (TCGA, Hartwig, PCAWG), separately in order to control for potential biases. Then, the three matrices were merged and used as input for the PCA (samples as rows, features as columns). PCs were extracted using the funcion *PCA* in the R package FactoMineR (version 2.4). Correlations were estimated by calculating the pearson correlation of each input somatic feature with each estimated score of each PC. Contributions were calculated by squaring the estimated loading matrix and dividing the squared loading by the sum of the loadings for the respective PC. Thus, the sum of the contributions (56 somatic input features for each PC) for each PC equals 1 (100 %).

4.1.19 Independent Component Analysis

The ICA was run on the same 56 somatic feature using the input matrix as described above. Similarly, as for the NMF, the number of ICs needs to be set before running the ICA. The methodology to extract the optimal number of components was adapted from the methodology applied previously[15] to extract the optimal number of NMF derived components. For the extraction of ICs the R package fastICA (version 1.2.1) in R 3.5.0 was used. The ICA was run by varying the number of extracted components from 2 to 30. For each component extraction the ICA was run 200 times and the seed for the random number generator was changed before every iteration. In each iteration the ICA decomposes the input matrix into a loadings matrix (corresponding to the components and their attributed weight from each somatic feature) and a scoring matrix (also called source matrix; samples projected to new component axes). After 200 iterations, the 200 loadings matrices were combined and clustered using k-medoids clustering with varying k from 2 to 50. Clustering was performed with the function *pam* from the R package cluster (version 2.0.6). For each clustering the average of the mean silhouette indexes of each cluster were saved as well as the lowest and second lowest mean silhouette index of a cluster extraction. Later, extracted summary silhouette indexes for different extracted IC numbers were plotted against the different number of extracted clusters (Figure 2.25). The optimal number of components was decided on visually (Figure 2.26). For a given extracted number of ICs, the optimal number of clusters was always times 2 since during each iteration signs flipped randomly and thus, each component always had a ‘mirrored’ counterpart with opposite signs (Figure 2.27). In the end always one component of the mirrored pair was kept. For the ICA with 56 somatic features as input, 15 individual ICs (using 30 clusters) were extracted. Correlations and contributions of each input somatic feature were estimated as described above (Figure 2.28).

4.1.20 Extraction of Components via a Variational Autoencoder

The architecture of the VAE was adapted from studies from Way *et al.*[266,267] (`https://github.com/greenelab/tybalt/blob/master/tybalt_vae.ipynb`), where they applied a VAE to compress gene expression data to extract biologically relevant representations. The script was modified for our purposes. In short, it is a simple ladder-VAE architecture[268] consisting of one encoding and one decoding layer to generate a generalizable representation of the input and to use this representation to reconstruct the input. Batch normalization was performed in the encoding layer before applying the activation function *ReLu*. In the encoding layer the VAE learned a distribution of means and standard deviations to generate the latent space. This latent representation was then decoded in the decoding layer by applying the *tanh* function as the final activation function. Weights were initialized via the Glorot uniform initializer[269]. We also tested adding an additional layer between the input and the encoding layer and between the latent space and the decoding layer. The extra layer always had 2 times more dimensions than the latent space and involved a batch normalization step before applying the *ReLu* activation function. The reconstruction loss was the sum of the mean squared error and the KL-divergence loss. To encourage learning, the ladder-VAE makes use of a so called *warm* start, meaning that it starts training without the KL divergence loss and linearly increases the contribution of the KL divergence loss after each cycle via the parameter *beta* (mean squared error+*beta**KL divergence loss). The linear increase of the contribution of the KL divergence loss was controlled via the parameter *kappa*.

In contrast to a previous VAE architecture[266,267], we applied the *tanh* function in the final decoding layer and used the mean squared error as part of the reconstruction loss since our input was not binary. To reconstruct the input via the *tanh* function, all the somatic features were transformed to a range of -1 to 1 prior to running the VAE. The data was split into 90 % training data and 10 % validation data and stratified by gender and cancer type. Performance was evaluated by checking the mean correlation of the reconstructed validation set with the validation input set and by calculating the correlation with selected ICs, which were shown to represent biologically relevant components. For this purpose, we calculated the maximum correlation of the components from the latent space of the VAE to the ICs 3 (dMMR_ICA), 4 (dHR_ICA), 12 (Smoking_ICA) and 14 (UV_ICA) and then calculated the average. To find the optimal hyperparameters we performed a grid search testing over 4,300 hyperparameter combinations (Figure 2.31). After finding the optimal hyperparameters, the VAE was run for different latent space dimensionalities 5 times with different random initializations (Figure 2.32). In the end, the results from using a latent space with 14 dimensions was extracted for further downstream analysis using the architecture with no extra layer between input and encoder and with no extra layer between decoder and output (Figures 2.33 and 2.34).

The VAE was run in a singularity container. A docker file was generated based on the docker image *tensorflow/tensorflow:1.15.5-gpu-py3-jupyter* and the python modules *scipy*, *scikit-learn*, and *seaborn* were added. The resulting docker image was then uploaded into Docker Hub and run in a singularity container. Python version 3.6.9, keras version 2.2.4 and tensorflow version 1.15.5 were used in this environment.

4.1.21 Estimation of Tissue Enrichments of Components

Tissue enrichments of individual components (Figures 2.30 and 2.35) were calculated as follows. For each component it was tested whether the component scores from one cancer type were significantly different to the scores of the remaining cancer types via a two-sided Welch's t-test. In addition, Cohen's d statistic was calculated between the two groups. This test was performed for each cancer type and separately for the two cohorts (TCGA and PCAWG + Hartwig). Cancer types were then grouped into their corresponding tissue of origin and the average Cohen's d statistic was calculated.

4.2 Identification of Rare Damaging Germline Variants

4.2.1 Extraction of Rare Germline Variants in the Discovery Cohort

TCGA bam files were downloaded as described here[253]. Strelka[260] 2.9.7 was run on TCGA WES normal and tumor samples to extract germline variants. Germline variants called in the tumor samples (will be a mix of germline and somatic mutations) were used later in a downstream step to only keep germline variants which were identified in the normal and tumor tissue. In this way, we aimed to remove potential false-positive germline calls in the normal sample and to remove variants which were selected out in the tumor and thus, irrelevant for our association analysis. Germline variants which were called in the normal sample with the filter PASS were kept as well as variants which were called with the filter LowGQX but had a GQX of at least 10. Variants which were found inside gnomAD[3] with the filter PASS and had a GQX of at least 10 were kept as well as variants which were not found inside gnomAD[3], but had a GQX of at least 20. Next, variants were annotated via ANNOVAR[270] (version 2019-10-24), CADD[185] v.1.6 scoring was added, and only exonic, and splicing variants were kept. Furthermore, only variants which had allele frequency of less than 0.1 % in gnomAD[3] (overall and in each subpopulation) were kept as well as variants which were not found inside gnomAD[3]. Variants with a frequency equal to or higher than 1 % within the cohort were removed. Additionally, rare germline variants were only kept when they were also found in the matching tumor sample.

4.2.2 Generation of a Coverage File for TCGA

We used the same methodology as described in previous work[167] to only extract genomic regions with sufficient coverage to be sure that regions in which no damaging germline variant was called was not due to lacking coverage. In brief, within each sequencing center (BI, WU, and BCM) 100 coverage files were randomly selected. Genomic regions which were covered by at least 8 reads in 90 % of the samples within each sequencing center were kept. Next, the coverage masks of the 3 sequencing centers were intersected, making up in total a genomic mask of 60 MB in length. Only genomic regions within these sites were kept for further analysis.

4.2.3 Extraction of Germline Variants in the Validation Cohort

Germline variants from PCAWG, Hartwig, ESAD-UK and MELA-AU were all processed in the same way if not indicated otherwise. Each cohort was processed at the beginning separately due to the different formats. The files were combined in the end. While germline calls from PCAWG and Hartwig were obtained as described above, germline variants in ESAD-UK and MELA-AU were called via Strelka [260] 2.9.10 (same approach as in TCGA), and derived from the same datasets from which the somatic calls were obtained as well. Thus, for ESAD-UK and MELA-AU the same approach as for TCGA was applied. For PCAWG and Hartwig, germline calls with the filter PASS by the respective germline detection tool were kept. Next, variants which were found inside gnomAD[3] and had the filter PASS were kept as well as variants which were not found inside gnomAD (rare singletons). Variants were annotated via ANNOVAR[270] (2019-10- 24). All variants which were found inside gnomAD[3] were required to have an allele frequency of less than 0.1 % (overall and in each subpopulation). Exonic and splicing variants were extracted. Furthermore, variants outside the CRG75 alignability mask were filtered out and variants with a frequency equal to or higher than 1 % within each cohort were discarded as well. The rare germline calls from the different cohorts were combined. Further, in all cases in which germline calls were also available for the matching tumor sample, variants were filtered out if they were not found in the matching tumor sample. Germline calls for matching tumor samples were available for whole PCAWG, $\sim$80 % of Hartwig, and not available for ESAD-UK and MELA-AU.

4.2.4 Definition of Rare Damaging Germline Variants

In this study 5 definitions of Rare Damaging Germline Variants (RDGVs) were applied in addition to requiring an allele frequency of <0.1 % (described above):

(i) RDGV = protein truncating variants (PTVs)

(ii) RDGV = PTVs + Missense variants with a CADD[185] $\geqq 25$

(iii) RDGV = PTVs + Missense variants with a CADD[185] $\geqq 15$

(iv) RDGV = Missense variants with a 'missense tolerance ratio'[271] $\leqq$ 25th percentile

(v) RDGV = Missense variants with a 'constrained coding region'[272] value $\geqq$ 90th percentile

For case (i) only PTVs were considered. PTVs comprised in this study frameshift deletions, frameshift insertions, stoploss variants, stopgain variants, startloss variants and splicing variants. Splicing variants comprise the canonical splice variants annotated by ANNOVAR[270] (version 2019-10-24) and variants with a predicted donor loss or acceptor loss higher than 0.8 by SpliceAI[273]. Pre-computed SpliceAI score files were downloaded from Illumina Basespace and annotations were added to each variant (hg38 for the discovery cohort and hg19 for the validation cohort). For cases (ii) and (iii) potentially damaging missense SNVs were added on top of the PTVs. Deleteriousness was assigned via the phred-scaled CADD[185] scores. For case (iv) we only considered missense SNVs with a missense tolerance ratio (MTR)[271] lower or equal to the 25th percentile and for case (v) we only considered missense SNVs with a constrained coding region (CCR)[272] value equal or bigger than the 90th percentile. On top of these variant filtering steps, two additional filtering steps were applied to all five RDGV sets in order to discard potential false-positive RDGVs: the proportion expressed across transcripts (PEXT) metric[274] and the terminal truncating exon rule[3].

4.2.5 Filtering out Non-Expressed Variants via the PEXT Metric

The PEXT score was introduced in one of the gnomAD articles and in brief, estimates to which extent a variant is expressed in a tissue based on isoform transcription levels from RNA-seq data. PEXT scores were estimated using over 11,000 tissue samples from GTEx[274]. Thus, PEXT scores were downloaded and added to the variant annotations. Since hg38 was used for the germline calls in the discovery cohort, PEXT annotations were first converted from hg19 to hg38 via the liftover tool from UCSC[261] (version 021620). This step was not necessary for the validation cohort. Variants were only kept when they had a PEXT value higher than 0.1 in the matching GTEx tissue. Matching a cancer type with the most appropriate GTEx tissue was mostly guided by a previous study[275]. For cases in which no matching GTEx tissue was available for a cancer type, the mean PEXT value was used. This filter was applied to all variants not affecting splicing since many splicing variants are close to exon borders and thus, don't have a PEXT score.

4.2.6 Exclusion of Terminal Truncating Exon Variants (with exceptions)

Terminal truncating variants might not have a deleterious loss-of-function effect since they can escape non-sense mediated decay ref and still be functional. For these reasons, they have been also removed in the loss-of-function transcript effect estimator (LOFTEE) of gnomAD[3]. Hence, variants occurring in the terminal exon were removed. This filter was not applied in cases in which the variant was predicted to have a deleterious effect by CADD $\geqq$ 15 or in cases in which the variant was predicted to have a splicing effect. In this way, we aimed to reduce the risk of loosing potentially harmful variants, which as described in the gnomAD flagship paper[3], can be the case when the C-terminal domain of a protein exerts a crucial function. To identify variants occurring in the last exon, gene coordinates were downloaded from UCSC[261] using the NCBI RefSeq track[276]. Exon coordinates of the last exon of the longest transcript were kept. These coordinates were then intersected with the variant coordinates in order to detect variants occurring in terminal exons.

4.3 Detecting and Assigning putative Loss of Heterozygosity (LOH)

4.3.1 Detecting and Assigning putative LOH in the Discovery Cohort TCGA

To detect LOH, we considered the copy number calls from FACETS[251]. FACETS calls were available for 9,814 samples. We extracted all 'LOH' and 'DUP-LOH' calls and assigned them to genes by intersecting the extracted coordinates with gene coordinates from NCBI Refseq[276] hg38. We assigned LOH to a gene in samples in which LOH was called via FACETS + the variant allele frequency of the RDGV was not higher in the normal sample than in the tumor sample and the variant allele frequency of the RDGV was not higher than 0.8 in the tumor and sample-matched normal sample. In this way, we aimed to only consider LOH events, when the putative RDGV of interest got enriched in the tumor via LOH since this was the tested hypothesis for the recessive and additive model. For 441 samples for which we did not have any FACETS calls, we assigned LOH to a gene in a sample when the difference in the variant allele frequency of the putative RDGV between tumor and normal sample was higher than 0.25 and when the variant allele frequency of the putative RDGV was higher than 0.8 in the tumor and sample-matched normal sample.

4.3.2 Detecting and Assigning putative LOH in the Validation Cohort

For PCAWG (excluding ESAD-UK and MELA-AU), CNV calls from ACEseq[258] v1.0.189 were further processed. All passed calls with the assignments 'LOH', 'LOHgain' or 'cnLOH' were extracted and genes were assigned to the LOH events as before (using NCBI Refseq[276] hg37). We excluded LOH calls when the corresponding RDGV in the respective gene had a lower allele frequency in the tumor than in the sample-matched normal sample and the allele frequency was not higher than 0.8 in both tissues.

For ESAD-UK and MELA-AU, CNV calls were available via FACETS[251] and LOH was called as described for TCGA. In contrast to the steps performed for TCGA, germline calls from the tumor tissue were not available for ESAD-UK and MELA-AU. Thus, LOH calls were not further filtered.

For Hartwig, CNV calls were provided via the tool Purple[9]. LOH was assigned to locations in which the minor allele ploidy was lower than 0.4. LOH calls were excluded in cases in which the allele frequency of the RDGV was lower in the tumor than in the sample-matched normal tissue and the allele frequency of the RDGV was not higher than 0.8 in the normal and tumor tissue. This was only applicable to the samples in which germline calls from the tumor genome were available (678 samples with germline calls from tumor genomes not available).

4.4 Gene-Based Rare Variant Association Testing

4.4.1 Extraction of Common Germline Variants and Sample-level Quality Control

Common variants were extracted from the normal samples to apply some sample-level quality control as well as to prepare the data to perform a PCA for extracting population ancestry. The following steps were performed for the discovery cohort (TCGA) and the validation cohort (PCAWG and Hartwig) separately. Germline variants which were called with the filter PASS were kept. Also, in accordance with the extraction of rare germline variants, variants with the filter LowGQX but a $\text{GQX} \geqq 10$ were kept in the respective cohorts (TCGA, ESAD-UK and MELA-AU). Common variants were extracted by only keeping variants which were identified inside gnomAD[3] with the filter PASS and with an allele frequency $> 5\,\%$ within the overall population. In TCGA all variants within the generated genomic mask were retained and in the other cohorts all variants within the CRG75 alignability mask were retained. Loci, in which more than 2 alleles existed were removed. The total number of common variants inside each sample was calculated and within each cohort (TCGA, Hartwig, PCAWG) samples in which the total number of variants was 1.5 standard deviations away from the mean were discarded (214 samples in TCGA, 212 samples in Hartwig, 204 samples in PCAWG) (Figures 2.42a and 2.43a). Next, common variants for each cohort were uploaded into PLINK v1.90b6.1 and further processed there. Missing genotypes were set as homozygous for the reference allele. Only variants with a $\text{MAF} > 5\,\%$ were retained and samples with a heterozygosity rate ± 3 standard deviations away from the mean were removed (127 samples in TCGA, 54 samples in Hartwig, 39 samples in PCAWG) (Figures 2.42b and 2.43b). For the following steps, variants on the sex chromosomes, on the mitochondrial chromosome and within regions with high amount of linkage disequilibrium (LD) (https://github.com/meyer-lab-cshl/plinkQC/tree/master/inst/extdata) were removed. Also, variants extensively deviating from the Hardy-Weinberg-equilibrium with $p < 10^{-6}$ were excluded.

4.4.2 Identification of Duplicated or Related Individuals

The dataset was pruned on the discovery cohort (TCGA) and on the merged validation cohort (PCAWG and Hartwig) separately, applying a window size of 50 bp, a step size of 5 and a r^2 threshold of 0.2. The identity-by-state (IBS) matrix was calculated for all pairs of individuals within each cohort. Within all pairs of individuals with identity-by-descent (IBD) > 0.185 (0.185 would be the expected value for individuals between third- and second-degree relatives) one individual was removed (542 samples in TCGA, and 479 samples in PCAWG and Hartwig) (Figures 2.42c and 2.43c).

4.4.3 Extraction of European Individuals

To extract individuals of European ancestry the pruned dataset was used and a principal component analysis (PCA) was performed. The PCA was run on the discovery cohort and on the merged validation cohort (Figures 2.44 and 2.45). The first ten principal components were used for clustering using the R package tclust (version 1.4.2), which trimmed 1 % of the outlying samples as described previously[167]. Individuals were grouped into $k = 10$ clusters and European groups were selected based on the reported TCGA/PCAWG annotations. In total 7,864 individuals were retained in the discovery cohort and 4,691 individuals were retained in the validation cohort. The PCA was repeated on the pruned dataset for the individuals of European ancestry in the respective cohorts to extract the PCs, which were used as covariates in the association testing (Figures 2.46 and 2.47).

4.4.4 Gene-Based Rare Variant Burden Testing

As described above 29 somatic mutational components were extracted from the discovery and validation cohort from the tumor genomes. Rare damaging germline variants (RDGVs) were extracted from the sample-matched normal samples. Gene-based rare variant burden testing was only performed on samples which survived the quality control filters (as described above). We limited the analysis to individuals with European ancestry due to the bigger sample size. In addition, only samples in which at least 10 SNVs were counted were kept. In total 6,799 samples were left in the discovery cohort for testing and 4,683 samples were left in the validation cohort for testing.

4.4.5 Gene Set

For testing, RGDVs occurring in 892 different genes were extracted. The gene set covered DNA damage response genes[277], known cancer predisposition genes[165], genes involved in chromatin organization (https://pathcards.genecards.org), genes involved in DNA double strand repair (https://pathcards.genecards.org), genes which were reported to regulate MSH2 stability[278], and human homologs of genes, in which heterozygous mutations cause genetic instability in *Saccharomyces cerevisiae*[196]. Effectively, out of the 892 individual genes 746 genes were tested in the most permissive RDGV set (set iii) in pan-can. The remaining genes were not tested in the discovery cohort since not enough RDGVs were identified in these genes to test them.

4.4.6 Association Testing via SKAT-O

Association testing was performed in each cancer type separately and with all cancer types together (pan-can). The effect of a gene on a somatic component was only tested when a RDGV in that gene was identified in at least two individuals. Testing was performed across 12 cancer types as shown in Table 4.5. Accordingly, depending on the cancer type different numbers of genes were tested in total.

Association testing was conducted via the unified testing approach of SKAT-O[184]. In short, SKAT-O combines the tests SKAT and burden via a weighted mean:

$$Q_\rho = \rho\, Q_B + (1\text{-}\rho) Q_S.$$

Here, Q_ρ is the final statistic from the weighted mean of the burden statistic Q_B and SKAT statistic Q_S. The parameter ρ influences how strongly each test is weighted. SKAT-O testing was performed via the R package SKAT[184]. For testing, the covariates were firstly regressed against the somatic components with the function *SKAT_Null_Model*. When applicable, age of diagnosis, sex, ancestry (first 6 PCs) and cancer type were used as covariates. Categorical variables were encoded as dummy variables with the R package dummy. Missing age information was imputed by taking the median value in the respective cohort. After initializing the null model, SKAT-O was run by using the function *SKAT* and setting the method to *SKATO*. The function ran SKAT-O with 10 different values of ρ (from 0 to 1) and reported the ρ value which led to the lowest p-value.

Three models of inheritance were tested in total and individual variants were encoded as follows:

(i) **Dominant**: no RDGV = 0; RDGV = 1

(ii) **Additive**: no RDGV = 0; RDGV = 1; RDGV + somatic LOH or biallelic RDGV = 2

(iii) **Recessive**: no RDGV = 0; RDGV + somatic LOH or biallelic RDGV = 1; RDGV without somatic LOH = excluded sample

Taken together, 3 models of inheritance were tested with 5 different RDGV sets, making up in total 15 models to test across 12 different cancer types and pan-can. In total, 15*12*29 = 5,655 model scenarios could have been tested at most. Ultimately, 4,693/5,655 out of scenarios were tested in the discovery phase, since in particular the recessive model could not be tested in many cancer types since not enough RDGV + somatic LOH events in the selected genes were detected.

4.4.7 Estimation of Effect Sizes via Burden Testing

Since no effect sizes were reported in SKAT-O, we also performed gene-based burden testing (aggregating variants occurring in the same gene) applying the same models as above. Association testing was performed via linear regression with the *lm* function of the R base package stats in R 3.5.0 as follows:

(i) Somatic Component $\sim$ Gene + Covariates

The somatic components were coded as quantitative variables as described above. The gene variable was encoded as a binary categorical variable depending on the model of inheritance (additive, recessive, dominant). When applicable, we controlled for age of diagnosis, sex, cancer type and ancestry (first 6 PCs) as covariates. In total, burden testing was performed for each scenario which was also tested via SKAT-O.

4.4.8 Quantile-Quantile Plots for Quality Control

To check for potential biases in testing, we plotted quantile-quantile plots (QQ-plots) for each somatic component tested for each scenerio (model of inheritance, RDGV set) in the respective cancer type and calculated the corresponding inflation factor λ. For the QQ-plots, the expected p-value was calculated by ranking all tested genes and dividing the rank of a gene by the total number of genes tested. The idea behind the QQ-plots was that most genes were expected to not have an effect on a somatic feature and thus, most p-values would be distributed randomly and fall on a linear line when ordered. The inflation factor λ was calculated to check for inflation, which would be indicated by $\lambda > 1$. The inflation factor λ was estimated by dividing the median of the chi-squared test statistic of the p-values (in R: *qchisq(1-p_values,1)*) by the expected median of the chi-squared distribution, which would be a chi-squared distribution with one degree of freedom (in R: *qchisq(0.5,1)*). QQ-plots with no inflation would have an inflation factor of $\lambda \approx 1$ and deflated QQ-plots would have an inflation factor of $\lambda < 1$. Ultimately, we excluded model scenarios in which at least 100 genes were tested and the inflation factor was $\geqq 1.5$ (19 ot ouf 1,909).

4.4.9 Estimation of False Discovery Rates

We calculated false discovery rates (FDRs) via two approaches: empirical FDR and via a random set of genes. To estimate the empirical FDR, the somatic component matrix (somatic components as columns and sample IDs as rows) were randomly shuffled within each cancer type. Importantly, the link between individuals and somatic components was broken, but correlation structure between features was conserved. Then, with the randomized somatic component matrix testing was performed in the same way as it was performed before. We calculated empirical FDR thresholds for each cancer type (or pan-can) separately. For instance, the p-value at which 1 % of the associations from the randomized run would have been called as a hit (false discovery) corresponds to a FDR of 1 %.

For our second approach, we repeated the whole analysis using 1,000 random genes. We generated a list of genes, which were not in our pre-selected gene list of 892 genes and in which RDGVs according to RDGV set (iii) were identified in at least 2 samples. In addition, we discarded all genes which were reported to have a physical interaction with any gene from our pre-selected gene list according to the reported physical interactions from STRING v11.5[279] with a combined score of at least 50 %. Out of 11,408 remaining genes, 1,000 genes were randomly selected and used for testing. Next, we performed the same steps as it was performed for the pre-selected list of genes, including the calculation of empirical FDRs via randomization and the exclusion of model scenarios with high inflation factors (31 out of 1,885). Based on the conservative hypothesis that there would be no real associations from the random list of genes, we calculated FDRs at different empirical FDR thresholds by dividing the number of hits, which were detected via the random list of genes by the number of genes detected at the same empirical FDR with our pre-selected list of genes. For instance, at an empirical FDR of 1 % we identified 44 hits with our random list of genes and 207 hits with out pre-selected list of genes. Thus, we estimated a FDR of $44/207 \approx 21\,\%$ at our empirical FDR of 1 %. This is a very conservative estimate of the FDR since the random gene lists may also include genes which affect somatic mutation processes.

4.4.10 Identification of Associations in the Discovery Cohort and Re-Testing in the Validation Cohort

Hits were identified in the discovery cohort when they were significant either at a FDR of 1 % or 2 % based on the estimation of the empirical FDR. These were then re-tested in the matching cancer type based on the tissue of origin (see Table 4.3). In total, for 12 individual cancer types a matching cancer type based on the tissue of origin was available in the validation cohort with a sample size of at least 50 samples: bladder cancer, brain glioma multiforme, low-grade glioma, breast cancer colorectal cancer, kidney cancer, lung adenocarcinoma, lung squamous carcinoma, ovary cancer, prostate cancer, skin cancer, stomach and esophagous cancer. Hits which were identified with all cancer types together (pan-can) were re-tested in the validation cohort in the same way. We called a hit as replicated when it reached the empirical FDR of either 1 % or 2 % and had the same estimate effect direction as in the discovery cohort. Effect size directions were extracted from the performed burden tests.

4.5 Network anaysis

For the network analysis, we downloaded protein network data from STRING v11.5[279] involving only physical links, and from HumanNet v3[280] the functional gene network (HumanNet-FN). From STRING we only kept interactions which had a combined confidence score (based on experimental, database, and text mining) of at least 80 %. The following steps were performed for each protein network separately.

Firstly, we extracted all interactions which involved interactions between genes from our pre-selected gene list of 892 genes. We calculated the total number of interactions our replicated genes had at an empirical FDR of 1 % with each other. It was tested via randomization whether this number was higher than one would expect at random. For this purpose, we selected randomly the same number of genes and calculated the total number of interactions these genes had with each other. We controlled for the total number of interactions each gene had, since some genes (e.g. *BRCA1*) have in general a lot of physical interactions, which would confound our results. To control for this, we counted the total number of interactions our replicated genes had, split them into 10 equal sized bins, assigned all our pre-selected genes a bin, and then selected randomly the same number of genes from each bin. Randomization was performed 1,000 times.

Next, we counted how many genes, which only replicated at an FDR of 2 %, had at least one interaction with a gene which replicated at an FDR of 1 %. Here, we applied the same approach. We counted the total number of interactions each gene, which only replicated at an FDR of 2 %, had in total and split the number of interactions into 10 equal sized bins. Each gene from our list of genes was assigned a bin and then we randomly selected 1,000 times the same number of genes from each bin and performed the same calculation.

4.6 Calculation of Frequency of RDGVs in Length Matched Randomly Selected Genes

To calculate the number of RDGVs occuring in a control set of genes, we matched each replicated gene randomly with a gene covering the same length. For this pupose, we intersected the TCGA coverage file (see Methods 4.2.2) with the reported exonic coordinates provided by NCBI RefSeq track[276] hg38. We only considered protein-coding genes. The covered length of each gene was calculated in kilobases and each replicated gene was randomly matched 10 times with a gene, which covered the same length in our data. Subsequently, RDVGs based on different sets (see Methods 4.2.4) were counted in the replicated gene sets as well as in the length matched control genes. For the validation cohort PCAWG_Hartwig-WGS, the same approach was applied. Here, the coordinates from the CRG75 alignability track were intersected with the exonic coordinates provided by NCBI RefSeq track[276] hg19 to determine the length of the coding region for a gene.

4.7 Common Germline Variant Association Testing

4.7.1 Discovery cohort - TCGA SNP Array Data

TCGA genotype data from Affymetrix 6.0 SNP arrays were downloaded from the Genome Data Commons (GDC) legacy archive. Genotypes from the GDC were called via the tool Birdseed and files from blood derived normal tissue or normal solid tissue were further processed (TCGA sample codes 10-14). Genotypes called with an error rate > 10 % were set as missing and unplaced SNPs were removed. Files were uploaded into PLINK v1.9 via a custom R script and all subsequent steps were handled in PLINK v1.9 starting with 905,420 SNPs and 10,128 samples.

Quality Control Steps on the SNP Array Data

Further processing of the data was performed following the recommendations of previous studies[281]. Samples with a SNP missing rate > 5% were removed and samples with an altered heterozygosity rate +/- 3 standard deviations from the mean were removed. SNPs with a missing rate > 5% across the samples were discarded. SNPs with a MAF < 1% and a significant deviation from the Hardy-Weinberg equilibrium P < 10^{-6} were filtered out. The SNP dataset was reduced via LD pruning to calculate the homozygosity rate and in order to calculate the identity-by-state (IBS) matrix. The data was pruned applying a window size of 50 bp, a step size of 5 bp and r^2 threshold of $r^2 > 0.2$. Individuals with discordant sex information with respect to the TCGA annotation were removed based on the homozygosity rate estimation. The IBS matrix was used to remove one individual from each pair removed when the calculated relatedness was > 0.185 (between second- and third-degree relatives). Next, SNPs on the sex chromosomes or mitochondrial chromosome were discarded from the not-pruned SNP set. SNPs in high LD regions ref were filtered out as well as SNPs from regions which are difficult to be aligned according to the CRG75 alignability track. In the end, 9,806 individuals and 567,769 SNPs were retained for the next steps.

Extraction of European Individuals

To extract individuals of European ancestry, the SNP set was pruned again as described above and a PCA was performed. The first ten principal components were used for clustering using the R package tclust (version 1.4.2), which trimmed 1 % of the outlying samples as described previously[167]. Individuals were grouped into k = 10 clusters and European groups were selected based on the reported TCGA annotation. In total 1,920 individuals were discarded.

Concordance Check with 1000 Genomes and GnomAD

1,000 genomes was used as the reference panel for imputation. Thus, it was checked whether SNP annotations from the array matched the reference. A/T and C/G SNPs, which were not concordant with 1000 genomes, were removed. Strand orientation of SNPs were flipped if necessary in order to match the strand orientation of the reference panel. Furthermore, a small number of SNPs that were triallelic in 1000 genomes (n = 70) were removed. Minor allele frequencies of all SNPs were calculated and compared against allele frequencies of the same SNPs in the non-Finnish European population in gnomAD v2 ref. SNPs with an allele frequencies difference bigger than 5 % in comparison to whole-genome gnomAD were removed (n = 12,332 SNPs). Next, SNPs with a MAF < 1 % were removed. In total, 7,886 individuals of European ancestry and 488,613 SNPs were left.

Imputation

Missing SNPs were imputed via Beagle[282] phase 3 version 5a (hg19), using the corresponding genetic maps provided by Beagle. Imputation was performed by Tiffany Delhomme from the Genome Data Science group at IRB. 503 samples from individuals of European ancestry from 1,000 genomes were used as the reference panel.

Post-Imputation Filtering and Extraction of Principal Components

Imputed SNPs with a dosage squared $r^2 < 0.3$ were removed from further analysis. LD pruning was performed as described above and a PCA was performed to extract the first 6 principal components, which were used in later steps as covariates.

4.7.2 Validation Cohort - Common Variant Processing

Extraction of Common Variants

Similar steps were performed as described above for the discovery cohort. In contrast to the discovery cohort, WGS data was used for genotyping since data from SNP arrays was not available. Variants called with the filter PASS were set according to their determined genotype, variants which were not called as PASS were set as missing and variants which were not in the sample file but in the dataset were set as homozygous using the reference allele. Only biallelic SNPs were kept. Since the validation is constrained by the SNPs which were tested in the discovery cohort, only SNPs which were extracted from the TCGA SNP array data, were extracted (n = 488,613). Next, within each cohort, samples with a missing SNP rate of > 5 % were removed and samples with an altered heterozygosity rate +/- 3 standard deviations from the mean were removed as well. All samples were loaded into one PLINK file, making up in total 5,569 samples and 488,592 SNPs. SNPs with an missing rate > 50 % and SNPs deviating from the Hardy-Weinberg equilibrium with $P < 10^{-6}$ were excluded from further analysis. Furthermore, the next steps were performed as described before in the discovery cohort. Highly related individuals were excluded, individuals of European ancestry were retained using the available PCAWG annotation[8] and SNPs were flipped if necessary in order to match 1000 genomes. SNPs with an altered MAF with respect to gnomAD were removed and the remaining 4,831 samples and 475,044 SNPs were used as input for imputation. Imputation was performed as described above and after imputation, SNPs with a dosage squared $r^2 > 0.3$ were kept. A PCA was performed on the imputed, LD pruned set of SNPs and the first 6 principal components were extracted, which were used during association testing as confounders.

4.7.3 Genome-Wide Association Analysis

First of all, all samples were retained for which somatic components were estimated. Similarly as for the rare variant association study, TCGA SNP array data was treated as the discovery cohort and PCAWG and Hartwig WGS data were treated as the validation cohort. The following steps were conducted for the two cohorts separately. Each somatic mutation phenotype was regressed against the confounders (age at onset of disease, sex, ancestry (first 6 PCs), and cancer type if applicable) and residuals were transformed via the rank based inverse normal transformation (RINT). Thus, we performed association testing via a indirect RINT-based approach, which has been reported to be robust when phenotypes are not normally distributed[283].

Association testing was performed via PLINK v2.00a3LM via the additive model. Only variants having an allele frequency of at least 1 % in the respective cancer type were tested. For the association tests the RINT transformed phenotypes were used and it was controlled for sex, cancer type, age at onset of disease, and ancestry (first 6 PCs). Association test were only performed in cancer types which had a sample size of at least 200 in the discovery cohort and in the validation cohort. These were the following: breast cancer (TCGA ID: BRCA), brain gliomas (TCGA ID: GBM + LGG), colorectal cancer (TCGA ID: COAD + READ), lung adenocarcinomas (TCGA ID: LUAD), lung squamous carcinomas (TCGA ID: LUSC), prostate cancer (TCGA ID: PRAD), skin cancer (TCGA ID: SKCM), stomach and esophagus cancer (TCGA ID: STAD + ESCA), and pan-can (TCGA ID: all combined). Matching cancer types in the validation cohort are shown in Table 4.5. In total, samples from 6,993 individuals and 484,781 SNPs were left for testing in pan-can in the discovery cohort. In the validation cohort, samples from 4,827 individuals and 481,475 SNPs were left.

Association testing was performed initially in the discovery cohort across all cancer types and somatic mutational components. SNP clumping was performed after association testing to report uncorrelated hits. SNPs within a radius of 500 kb, a p-value not bigger than $1*10^{-5}$ and a r^2 of at least 0.5 were clumped together by taking the hit with the lowest p-value as the central variant. Hits reaching genome-wide significance ($p < 5*10^{-8}$) we re-tested in the validation cohort.

4.7.4 GWAS Power Analysis

GWAS power analysis was performed with an openly available R function from GitHub (`https://github.com/kaustubhad/gwas-power`), which is based on power calculations formulae presented by Visscher *et al.*[247]. The theoretical power was estimated to identify a hit at genome-wide significance ($p < 5*10^{-8}$) at varying effect sizes (0.1-0.7), varying minor allele frequencies (0.1-0.5), and varying sample sizes (300, 400, 500, 600, 700, 4800, and 7000. Tested effect sizes were selected based on the effect sizes of the individual cancer types and pan-can (Table 4.7).

4.8 SNP Heritability Estimation

ISNP heritability was estimated via two methods: single-component GREML[194] and LD score regression[284]. GREML is higher powered than LD score regression and requires a sample size of at least 1,000 since the error at sample size of 1,000 is approximately 32 %, at a sample size of 5,000 samples around 6 % and at sample size of 7,000 samples around 5 % (318 divided by the sample size[191]. Thus, SNP heritabilities were primarily only calculated for pan-can and not for the individual cancer types since sample sizes were too low.

4.8.1 Single Component GREML

SNP heritabilities via single-component GREML were calculated via the published software package in GCTA[194]. In short, based on the extracted SNPs a genetic relationship matrix was calculated, which was then used for the main GREML analysis. RINT transformed somatic components were used as the phenotype inputs and sex, age at diagnosis, ancestry (first 6 PCs), and cancer type were used as covariates if applicable.

4.8.2 LD Score Regression

LD score regression (LDSC)[284] was used as the second method for calculating SNP heritabilities. LDSC based heritabilites were calculated by utilizing the published tool `https://github.com/bulik/ldsc`. The tool used as input the GWAS summary statistics from the conducted GWAS to regress the GWAS statistics against the LD scores. Pre-calculated LD scores based on individuals of European ancestry from 1000 genomes were utilized.

4.8.3 Estimation of SNP-heritabilities changes

First of all, individual cancer types from the respective cohorts (TCGA and PCAWG+Hartwig) were dropped, heritability of the total mutation burden was estimated via GREML, and the difference to the heritability of the total mutation burden without dropping the cancer types was calculated. A null distribution was obtained via randomization. For each tested cancer type, the same number of samples was randomly selected in the dataset and dropped and the heritability was estimated in the same way as described above. Randomization was performed for each cancer type 1,000 times. A p-value was estimated via a lower- and upper-tailed test, by calculating the probability to obtain the same difference in heritability or a more extreme value. P-values from cancer types derived from the same tissues were combined via Fisher's method.

Secondly, as shown in Section 2.4.4, SNV-based counts for each mutation type in its trinucleotide context (96 possibilities, e.g. C>A in CCC context) were dropped from each cancer type and the resulting somatic feature was used to estimate the heritability via GREML. Again, the difference in heritability was calculated by subtracting the estimated heritability, from the heritability which was measured when using the total mutation burden with all SNVs. In total, there were 96 different contexts and 104 cancer types, leading to 96*104 = 9,984 heritability estimations. A PCA was run on the matrix (Figure 2.81), after capping a clear outlier value (C>T in ACG context in TCGA_PRAD estimated difference of 49 %) at the second-highest value, since 99 % of the estimated differences in heritability were $\leqq$ 6 %.

4.9 Supplementary Tables

4.9.1 Genomic Regions

Table 4.1: Covered Genomic Regions with WES and WGS Masks. Estimated lengths of different genomic regions in megabases after applying CRG75 alignability mask on WES and WGS data respectively.

Region	Bin	Size in WES in Megbases	Size in WGS in Megbases
RT	1of6 (late)	1.62	345
	2of6	4.03	382
	3of6	7.94	381
	4of6	12.0	379
	5of6	16.7	373
	6of6 (early)	29.2	356
H3K36me3	0of5 (no marks)	71.0	3,705
	1of5	6.63	143
	2of5	8.54	143
	3of5	11.3	144
	4of5	16.1	147
	5of5 (high density of marks)	29.4	152
Expression	0of5 (no expression)	7.54	2,050
	1of5	14.9	374
	2of5	17.0	425
	3of5	24.2	523
	4of5	34.9	531
	5of5 (high expression)	44.4	531
DNase I	0of5 (no marks)	95.7	3,758
	1of5	6.89	137
	2of5	7.68	136
	3of5	8.85	135
	4of5	9.95	134
	5of5 (high density of marks)	9.95	134
CTCF/cohesin	flanking site ±500 bp	7.04	83.2
	binding site	1.56	17.38
Fork polarity	1of10 (lagging strand)	7.71	199
	2of10	7.65	196
	9of10	6.75	195
	10of10 (leading strand)	5.54	200

4.9.2 Somatic Components

Table 4.2: Somatic Component Names.

Component	Name
IC1	Sig.17
IC2	Sig.MMR2+ampli.
IC3	dMMR$_{ICA}$
IC4	dHR$_{ICA}$
IC5	Deletions$_{ICA}$
IC6	APOBEC$_{ICA}$
IC7	Sig.18
IC8	Ploidy
IC9	Sig.11+19
IC10	DBS2
IC11	Sig.5$_{ICA}$
IC12	Smoking$_{ICA}$
IC13	Small indels 2bp
IC14	UV$_{ICA}$
IC15	Sig.8
VAE_1	APOBEC$_{VAE2}$
VAE_2	Deletions$_{VAE}$
VAE_3	Sig.5$_{VAE}$
VAE_4	Sig.1
VAE_5	UV$_{VAE}$
VAE_6	dHR$_{VAE1}$
VAE_7	Mitochondria
VAE_8	dHR$_{VAE2}$
VAE_9	Smoking$_{VAE}$
VAE_10	dMMR$_{VAE2}$
VAE_11	APOBEC$_{VAE1}$
VAE_12	X-hypermutation
VAE_13	dMMR$_{VAE1}$
VAE_14	Amplifications

4.9.3 Cancer Types

Table 4.3: TCGA Study Abbreviations. `https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations`. List of studies, which were ultimately used in association testing (after filtering steps).

Study Abbreviation	Study Name
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma

Table 4.4: PCAWG Study Abbreviations. `https://dcc.icgc.org/projects/details`. List of studies, which were ultimately used in association testing (after filtering steps).

Study Abbreviation	Study Name
BOCA-UK	Bone Cancer - UK
BRCA-EU	Breast ER+ and HER2- Cancer - EU/UK
BRCA-UK	Breast Triple Negative/Lobular Cancer - UK
BTCA-SG	Biliary Tract Cancer - SG
CLLE-ES	Chronic Lymphocytic Leukemia - ES
CMDI-UK	Chronic Myeloid Disorders - UK
EOPC-DE	Early Onset Prostate Cancer - DE
ESAD-UK	Esophageal Adenocarcinoma - UK
LICA-FR	Liver Cancer - FR
MALY-DE	Malignant Lymphoma - DE
MELA-AU	Skin Cancer - AU
OV-AU	Ovarian Cancer - AU
PACA-AU	Pancreatic Cancer - AU
PACA-CA	Pancreatic Cancer - CA
PAEN-AU	Pancreatic Cancer Endocrine neoplasms - AU
PAEN-IT	Pancreatic Endocrine Neoplasms - IT
PBCA-DE	Pediatric Brain Cancer - DE
PRAD-CA	Prostate Adenocarcinoma - CA
PRAD-UK	Prostate Adenocarcinoma - UK
RECA-EU	Renal Cell Cancer - EU/FR

Table 4.5: Cancer Type Names. Cancer type names used in this study and respective cancer types from TCGA, PCAWG, and Hartwig which were assigned to it. EAC: oesophageal adenocarcinoma.

Cancer Type Name	Discovery		Validation
	TCGA Cancer Type	PCAWG Project ID(s)	Hartwig Cancer Type(s)
Bladder	BLCA	-	Urinarytract
Brain_glioma_low	LGG	PBCA-DE	Nervoussystem_Gliomas or _NA
Brain_glioma_multi	GBM	PBCA-DE	Nervoussystem_Gliomas or _NA
Breast	BRCA	BRCA-EU, BRCA-UK	Breast
Colon_Rectum	COAD, READ	-	Colon_Rectum
Kidney	KIRC, KIRP	RECA-EU	Kidney
Lung_ad	LUAD	-	Lung
Lung_sq	LUSC	-	Lung
Ovary	OV	OV-AU	Ovary
Prostate	PRAD	PRAD-CA, PRAD-UK	Prostate
Skin	SKCM	MELA-AU_Cutaneous	Skin_Melanoma or _NA
Stomach_Eso	STAD, ESCA (EAC only[285])	GACA-CN, ESAD-UK	Stomach, Esophagus

Table 4.6: Overview of sample sizes in rare variant association study. Corresponding cancer types for the cancer type names can be found in Table 4.5.

Cancer Type Name	Discovery cohort sample size	Validation cohort sample size
Bladder	323	87
Brain_glioma_low	405	283
Brain_glioma_multi	253	283
Breast	684	656
Colon_Rectum	410	417
Kidney	445	168
Lung_ad	434	299
Lung_sq	373	299
Ovary	199	180
Prostate	386	443
Skin	403	370
Stomach_Eso	363	431
Pan-can	6,799	4,683

Table 4.7: Overview of sample sizes in common variant association study. Corresponding cancer types for the cancer type names can be found in Table 4.5.c

Cancer Type Name	Discovery cohort sample size	Validation cohort sample size
Brain_glioma	714	249
Breast	688	680
Colon_Rectum	395	435
Lung_ad	427	312
Lung_sq	403	312
Prostate	378	448
Skin	414	423
Stomach_Eso	381	453
Pan-can	6,993	4,827

5 Conclusions

- While somatic mutational features can be more accurately measured via whole-genome sequencing data, whole-exome sequencing data is also a viable source to extract somatic mutational patterns.
- ICA and VAE neural networks can be utilized to extract biologically relevant mutational components from an input matrix containing mutation counts and ratios based on different mutation classes and genomic properties.
- Rare damaging inherited variants in diverse genes associated with many different mutational processes.
- Novel genes associated in the rare germline variant association study with dHR-related repair (e.g. *RIF1*, *PAXIP1*, *WRN*, *EXO1*, and *ATR*), with dMMR (e.g. *MSH3*, *MTOR*, *TTI2*, *SETD2*, *EXO1*, and *MLH3*), and with APOBEC-directed mutagenesis (e.g. *APEX1*) among others.
- Network analysis in rare variant association testing is a useful approach for prioritising newly associated genes and generating hypotheses.
- In SNP-set level tests, the variance-based test SKAT can be utilized to compensate for inaccurate predictions of damaging variants by *in silico* predictors.
- Several hits were identified in the common germline variant association studies, but none of them replicated, potentially due to low power, batch effects from genotyping (SNP-arrays vs. WGS), or differences in measuring somatic components (WES vs. WGS).
- Sample sizes for common germline variant association studies were too low to identify potential small true effect SNPs.
- Heritability estimates showed that at least somatic mutations due to DNA mismatch repair deficiencies, the total burden of C>A mutations, and the total mutation burden have a heritable component originating from common germline variants, but mostly $\leqq 10\%$.

- Heritability of the total mutation burden can be attributed to at least three different mutational processes: APOBEC activity, signature 1 ($\sim$ number of cell divisions), and signature 17b ($\sim$ oxidative damage in the nucleotide pool).

- In the future, larger sample sizes and more whole-genome sequencing data will make it possible to study the effects of germline variants on somatic mutational processes at an even higher resolution across individual cancer (sub)types and populations.

6 References

1. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

2. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).

3. Karczewski, K.J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434-443 (2020).

4. Tate, J.G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).

5. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-93 (2012).

6. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).

7. Bailey, M.H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034-1035 (2018).

8. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).

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

10. Martincorena, I. *et al*. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e21 (2017).

11. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat Genet* **49**, 1785-1788 (2017).

12. Dietlein, F. *et al*. Identification of cancer driver genes based on nucleotide context. *Nat Genet* **52**, 208-218 (2020).

13. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102-111 (2020).

14. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).

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

16. Mas-Ponte, D. & Supek, F. DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers. *Nat Genet* **52**, 958-968 (2020).

17. Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet* **50**, 1262-1270 (2018).

18. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121 (2020).

19. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504-7 (2012).

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

21. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81-4 (2015).

22. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat Biotechnol* **32**, 71-5 (2014).

23. Shinbrot, E. *et al.* Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res* **24**, 1740-50 (2014).

24. Haradhvala, N.J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-49 (2016).

25. Cancer Genome Atlas Research, N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).

26. Hunter, C. *et al*. A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res* **66**, 3987-91 (2006).

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

28. Thomas, R.K. *et al.* Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat Med* **12**, 852-5 (2006).

29. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892 (2012).

30. Black, J.R.M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat Rev Cancer* **21**, 379-392 (2021).

31. Ramroop, J.R., Gerber, M.M. & Toland, A.E. Germline Variants Impact Somatic Events during Tumorigenesis. *Trends Genet* **35**, 515-526 (2019).

32. Wang, S. *et al*. Germline variants and somatic mutation signatures of breast cancer across populations of African and European ancestry in the US and Nigeria. *Int J Cancer* **145**, 3321-3333 (2019).

33. International Agency for Research on Cancer (IARC). All cancers fact sheet. (2020). `https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf`.

34. National Cancer Institute (NIH). Cancer Statistics. (2020). `https://www.cancer.gov/about-cancer/understanding/statistics`.

35. Hansemann, D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Archiv f. pathol. Anat.* **119**, 299–326 (1890).

36. Boveri, T. Zur Frage der Entstehung Maligner Tumoren. 1-64 (Gustav Fischer, 1914).

37. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

38. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-74 (2011).

39. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* 458, 719-24 (2009).

40. Tomasetti, C., Marchionni, L., Nowak, M.A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A* **112**, 118-23 (2015).

41. Brock, A., Chang, H. & Huang, S. Non-genetic heterogeneity–a mutation-independent driving force for the somatic evolution of tumours. *Nat Rev Genet* **10**, 336-42 (2009).

42. Volkova, N.V. *et al*. Mutational signatures are jointly shaped by DNA damage and repair. *Nat Commun* **11**, 2169 (2020).

43. Kiwerska, K. & Szyfter, K. DNA repair in cancer initiation, progression, and therapy-a double-edged sword. *J Appl Genet* **60**, 329-334 (2019).

44. Chatterjee, N. & Walker, G.C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen* **58**, 235-263 (2017).

45. Knoch, J., Kamenisch, Y., Kubisch, C. & Berneburg, M. Rare hereditary diseases with defects in DNA-repair. *Eur J Dermatol* **22**, 443-55 (2012).

46. Sharma, R., Lewis, S. & Wlodarski, M.W. DNA Repair Syndromes and Cancer: Insights Into Genetics and Phenotype Patterns. *Front Pediatr* **8**, 570084 (2020).

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

48. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**, 484-92 (2012).

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

50. Drummond, J.T., Li, G.M., Longley, M.J. & Modrich, P. Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. *Science* **268**, 1909-12 (1995).

51. Palombo, F. *et al*. GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. *Science* **268**, 1912-4 (1995).

52. Habraken, Y., Sung, P., Prakash, L. & Prakash, S. Binding of insertion/deletion DNA mismatches by the heterodimer of yeast mismatch repair proteins MSH2 and MSH3. *Curr Biol* **6**, 1185-7 (1996).

53. Palombo, F. *et al.* hMutSβ, a heterodimer of hMSH2 and hMSH3, binds to insertion/deletion loops in DNA. *Curr Biol* **6**, 1181-4 (1996).

54. Kadyrov, F.A., Dzantiev, L., Constantin, N. & Modrich, P. Endonucleolytic function of MutLα in human mismatch repair. *Cell* **126**, 297-308 (2006).

55. Guo, S. *et al.* Regulation of replication protein A functions in DNA mismatch repair by phosphorylation. *J Biol Chem* **281**, 21607-21616 (2006).

56. Crouse, G.F. Non-canonical actions of mismatch repair. *DNA Repair (Amst)* **38**, 102-109 (2016).

57. Cascalho, M., Wong, J., Steinberg, C. & Wabl, M. Mismatch repair co-opted by hypermutation. *Science* **279**, 1207-10 (1998).

58. Pena-Diaz, J. *et al.* Noncanonical mismatch repair as a source of genomic instability in human cells. *Mol Cell* **47**, 669-80 (2012).

59. Fang, H., Zhu, X., Oh, J., Barbour, J. & Wong, J. Deficiency in DNA mismatch repair of methylation damage is a major mutational process in cancer. *bioRxiv* (2020).

60. Truninger, K. *et al.* Immunohistochemical analysis reveals high frequency of PMS2 defects in colorectal cancer. *Gastroenterology* **128**, 1160-71 (2005).

61. Liccardo, R., De Rosa, M., Izzo, P. & Duraturo, F. Novel Implications in Molecular Diagnosis of Lynch Syndrome. *Gastroenterol Res Pract* **2017**, 2595098 (2017).

62. Thibodeau, S.N., Bren, G. & Schaid, D. Microsatellite instability in cancer of the proximal colon. *Science* **260**, 816-9 (1993).

63. Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363**, 558-61 (1993).

64. Esteller, M., Levine, R., Baylin, S.B., Ellenson, L.H. & Herman, J.G. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. *Oncogene* **17**, 2413-7 (1998).

65. Boland, C.R. *et al.* A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* **58**, 5248-57 (1998).

66. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015-6 (2014).

67. Kautto, E.A. *et al.* Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* **8**, 7452-7463 (2017).

68. Bonneville, R. *et al.* Landscape of Microsatellite Instability Across 39 Cancer Types. *JCO Precis Oncol* **2017** (2017).

69. Buckowitz, A. *et al.* Microsatellite instability in colorectal cancer is associated with local lymphocyte infiltration and low frequency of distant metastases. *Br J Cancer* **92**, 1746-53 (2005).

70. Le, D.T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**, 2509-20 (2015).

71. Scharer, O.D. Nucleotide excision repair in eukaryotes. *Cold Spring Harb Perspect Biol* **5**, a012609 (2013).

72. DiGiovanna, J.J. & Kraemer, K.H. Shining a light on xeroderma pigmentosum. *J Invest Dermatol* **132**, 785-96 (2012).

73. Wallace, S.S., Murphy, D.L. & Sweasy, J.B. Base excision repair and cancer. *Cancer Lett* **327**, 73-89 (2012).

74. Yamada, T. *et al.* Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability. *Cancer Lett* **181**, 115-20 (2002).

75. Hao, B. *et al.* Identification of genetic variants in base excision repair pathway and their associations with risk of esophageal squamous cell carcinoma. *Cancer Res* **6**4, 4378-84 (2004).

76. Rodrigues, M. *et al.* Outlier response to anti-PD1 in uveal melanoma reveals germline MBD4 mutations in hypermutated tumors. *Nat Commun* **9**, 1866 (2018).

77. Sanders, M.A. et al. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. *Blood* **132**, 1526-1534 (2018).

78. Cheadle, J.P. & Sampson, J.R. Exposing the MYtH about base excision repair and human inherited disease. *Hum Mol Genet* **12 Spec No 2**, R159-65 (2003).

79. Grolleman, J.E. *et al*. Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell* **35**, 256-266 e5 (2019).

80. Bebenek, A. & Ziuzia-Graczyk, I. Fidelity of DNA replication-a matter of proofreading. *Curr Genet* **64**, 985-996 (2018).

81. Echols, H. & Goodman, M.F. Fidelity mechanisms in DNA replication. *Annu Rev Biochem* **60**, 477-511 (1991).

82. Kunkel, T.A. Evolving views of DNA replication (in)fidelity. *Cold Spring Harb Symp Quant Biol* **74**, 91-101 (2009).

83. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* **45**, 136-44 (2013).

84. Barbari, S.R. & Shcherbakova, P.V. Replicative DNA polymerase defects in human cancers: Consequences, mechanisms, and implications for therapy. *DNA Repair (Amst)* **56**, 16-25 (2017).

85. Imboden, S. *et al.* Phenotype of POLE-mutated endometrial cancer. *PLoS One* **14**, e0214318 (2019).

86. Zafar, M.K. & Eoff, R.L. Translesion DNA Synthesis in Cancer: Molecular Mechanisms and Therapeutic Opportunities. *Chem Res Toxicol* **30**, 1942-1955 (2017).

87. Goodman, M.F. & Woodgate, R. Translesion DNA polymerases. *Cold Spring Harb Perspect Biol* **5**, a010363 (2013).

88. Johnson, R.E., Washington, M.T., Prakash, S. & Prakash, L. Fidelity of human DNA polymerase eta. *J Biol Chem* **275**, 7447-50 (2000).

89. Masutani, C. *et al*. The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase eta. *Nature* **399**, 700-4 (1999).

90. Helleday, T. Homologous recombination in cancer development, treatment and development of drug resistance. *Carcinogenesis* **31**, 955-60 (2010).

91. Lieber, M.R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem* **79**, 181-211 (2010).

92. Krajewska, M., Fehrmann, R.S., de Vries, E.G. & van Vugt, M.A. Regulators of homologous recombination repair as novel targets for cancer treatment. *Front Genet* **6**, 96 (2015).

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

94. Menghi, F. *et al*. The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell* **34**, 197-210 e5 (2018).

95. Nguyen, L., J, W.M.M., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nat Commun* **11**, 5584 (2020).

96. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. Comparison of non-homologous end joining and homologous recombination in human cells. *DNA Repair (Amst)* **7**, 1765-71 (2008).

97. Lieber, M.R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* **4**, 712-20 (2003).

98. Ceccaldi, R., Sarangi, P. & D'Andrea, A.D. The Fanconi anaemia pathway: new players and new functions. *Nat Rev Mol Cell Biol* **17**, 337-49 (2016).

99. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264 (2016).

100. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**, 3405-10 (1974).

101. Bellacosa, A. & Drohat, A.C. Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites. *DNA Repair (Amst)* **32**, 33-42 (2015).

102. Krokan, H.E. *et al*. Error-free versus mutagenic processing of genomic uracil–relevance to cancer. *DNA Repair (Amst)* **19**, 38-47 (2014).

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

104. Tomkova, M., McClellan, M., Kriaucionis, S. & Schuster-Bockler, B. DNA Replication and associated repair pathways are involved in the mutagenesis of methylated cytosine. *DNA Repair (Amst)* **62**, 1-7 (2018).

105. Moris, A., Murray, S. & Cardinaud, S. AID and APOBECs span the gap between innate and adaptive immunity. *Front Microbiol* **5**, 534 (2014).

106. Nabel, C.S. *et al.* AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. *Nat Chem Biol* **8**, 751-8 (2012).

107. Roberts, S.A. *et al*. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* **46**, 424-35 (2012).

108. Seplyarskiy, V.B. *et al.* APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res* **26**, 174-82 (2016).

109. Nordentoft, I. *et al*. Mutational context and diverse clonal development in early and late bladder cancer. *Cell Rep* **7**, 1649-1663 (2014).

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

111. Nik-Zainal, S. *et al*. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genet* **46**, 487-91 (2014).

112. Middlebrooks, C.D. *et al*. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat Genet* **48**, 1330-1338 (2016).

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

114. Pleasance, E. *et al*. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nature Cancer 1*, 452–468 (2020).

115. Vohringer, H., Hoeck, A.V., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat Commun* **12**, 3628 (2021).

116. Kucab, J.E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836 e16 (2019).

117. Brash, D.E. UV signature mutations. *Photochem Photobiol* **91**, 15-26 (2015).

118. Cannistraro, V.J. & Taylor, J.S. Acceleration of 5-methylcytosine deamination in cyclobutane dimers by G and its implications for UV-induced C-to-T mutation hotspots. *J Mol Biol 392*, 1145-57 (2009).

119. Ikehata, H., Chang, Y., Yokoi, M., Yamamoto, M. & Hanaoka, F. Remarkable induction of UV-signature mutations at the 3'-cytosine of dipyrimidine sites except at 5'-TCG-3' in the UVB-exposed skin epidermis of xeroderma pigmentosum variant model mice. *DNA Repair (Amst)* **22**, 112-22 (2014).

120. Denissenko, M.F., Chen, J.X., Tang, M.S. & Pfeifer, G.P. Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proc Natl Acad Sci U S A* **94**, 3893-8 (1997).

121. Hussain, S.P. *et al.* Mutability of p53 hotspot codons to benzo(a)pyrene diol epoxide (BPDE) and the frequency of p53 mutations in nontumorous human lung. *Cancer Res* **61**, 6350-5 (2001).

122. Xue, J., Yang, S. & Seng, S. Mechanisms of Cancer Induction by Tobacco-Specific NNK and NNN. *Cancers (Basel)* **6**, 1138-56 (2014).

123. Waris, G. & Ahsan, H. Reactive oxygen species: role in the development of cancer and various chronic conditions. *J Carcinog* **5**, 14 (2006).

124. Higinbotham, K.G. *et al*. GGT to GTT transversions in codon 12 of the K-ras oncogene in rat renal sarcomas induced with nickel subsulfide or nickel subsulfide/iron are consistent with oxidative damage to DNA. *Cancer Res* **52**, 4747-51 (1992).

125. Pilati, C. *et al*. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol* **242**, 10-15 (2017).

126. Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nat Cancer* **2**, 643-657 (2021).

127. Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. & Baradaran, B. The Different Mechanisms of Cancer Drug Resistance: A Brief Review. *Adv Pharm Bull* **7**, 339-348 (2017).

128. Brennan, C.W. *et al*. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462-77 (2013).

129. Lee, S.Y. Temozolomide resistance in glioblastoma multiforme. *Genes Dis* **3**, 198-210 (2016).

130. Dasari, S. & Tchounwou, P.B. Cisplatin in cancer therapy: molecular mechanisms of action. *Eur J Pharmacol* **740**, 364-78 (2014).

131. Longley, D.B., Harkin, D.P. & Johnston, P.G. 5-fluorouracil: mechanisms of action and clinical strategies. *Nat Rev Cancer* **3**, 330-8 (2003).

132. Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer* (2021).

133. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-59 (2013).

134. Baez-Ortega, A. & Gori, K. Computational approaches for discovery of mutational signatures in cancer. *Brief Bioinform* **20**, 77-88 (2019).

135. Kasar, S. *et al*. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat Commun* **6**, 8866 (2015).

136. Fischer, A., Illingworth, C.J., Campbell, P.J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol* **14**, R39 (2013).

137. Pei, G., Hu, R., Dai, Y., Zhao, Z. & Jia, P. Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network. *Oncogene* **39**, 5031-5041 (2020).

138. Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat Cancer* **1**, 249-263 (2020).

139. Maura, F. *et al.* A practical guide for mutational signature analysis in hematological malignancies. *Nat Commun* **10**, 2969 (2019).

140. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-33 (2013).

141. Wang, S. *et al.* Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. *PLoS Genet* **17**, e1009557 (2021).

142. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome Biol* **20**, 246 (2019).

143. Funnell, T. *et al.* Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput Biol* **15**, e1006799 (2019).

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

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

146. Klein, K.N. *et al.* Replication timing maintains the global epigenetic state in human cells. *Science* **372**, 371-378 (2021).

147. Lee, C.A., Abd-Rabbo, D. & Reimand, J. Functional and genetic determinants of mutation rate variability in regulatory elements of cancer genomes. *Genome Biol* **22**, 133 (2021).

148. Sun, Z. *et al.* H3K36me3, message from chromatin to DNA damage repair. *Cell Biosci* **10**, 9 (2020).

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

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

151. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-21 (2015).

152. Poulos, R.C. *et al.* Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. *Cell Rep* **17**, 2865-2872 (2016).

153. Inoue, M. *et al.* Induction of chromosomal gene mutations in *Escherichia coli* by direct incorporation of oxidatively damaged nucleotides. New evaluation method for mutagenesis by damaged DNA precursors in vivo. *J Biol Chem* **273**, 11069-74 (1998).

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

155. Jager, N. *et al.* Hypermutation of the inactive X chromosome is a frequent event in cancer. *Cell* **155**, 567-81 (2013).

156. Akdemir, K.C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat Genet* **52**, 1178-1188 (2020).

157. Ju, Y.S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife* **3** (2014).

158. Reznik, E. *et al.* Mitochondrial DNA copy number variation across human cancers. *Elife* **5** (2016).

159. Yuan, Y. *et al.* Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat Genet* **52**, 342-352 (2020).

160. Blokzijl, F. *et al*. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264 (2016).

161. Robinson, P.S. *et al*. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat Genet* **53**, 1434-1442 (2021).

162. Garcia-Nieto, P.E., Morrison, A.J. & Fraser, H.B. The somatic mutation landscape of the human body. *Genome Biol* **20**, 298 (2019).

163. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).

164. Martincorena, I. & Campbell, P.J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483-9 (2015).

165. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302-8 (2014).

166. Qing, T. *et al*. Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. *Nat Commun* **11**, 2438 (2020).

167. Park, S., Supek, F. & Lehner, B. Systematic discovery of germline cancer predisposition genes through the identification of somatic second hits. *Nat Commun* **9**, 2601 (2018).

168. Carter, H. *et al*. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov* **7**, 410-423 (2017).

169. Park, S., Supek, F. & Lehner, B. Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers. *Nat Commun* **12**, 7051 (2021).

170. Knudson, A.G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-3 (1971).

171. Friend, S.H. et al. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**, 643-6 (1986).

172. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-45 (2012).

173. Tam, V. *et al*. Benefits and limitations of genome-wide association studies. *Nat Rev Genet* **20**, 467-484 (2019).

174. Lord, C.J. & Ashworth, A. BRCAness revisited. *Nat Rev Cancer* **16**, 110-20 (2016).

175. Sfeir, A. & Symington, L.S. Microhomology-Mediated End Joining: A Backup Survival Mechanism or Dedicated Pathway? *Trends Biochem Sci* **40**, 701-714 (2015).

176. Ma, X., Zhang, B. & Zheng, W. Genetic variants associated with colorectal cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Gut* **63**, 326-36 (2014).

177. Malkin, D. Li-fraumeni syndrome. *Genes Cancer 2*, 475-84 (2011).

178. Rausch, T. *et al.* Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell* **148**, 59-71 (2012).

179. Morgenthaler, S. & Thilly, W.G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* **615**, 28-56 (2007).

180. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23 (2014).

181. Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* **70**, 42-54 (2010).

182. Neale, B.M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet* **7**, e1001322 (2011).

183. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).

184. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).

185. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med* **13**, 31 (2021).

186. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**, 1161-1170 (2018).

187. Ng, P.C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-4 (2003).

188. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016).

189. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91-95 (2021).

190. Shin, J.E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nat Commun* **12**, 2403 (2021).

191. Sun, X. *et al.* Tumor Mutational Burden Is Polygenic and Genetically Associated with Complex Traits and Diseases. *Cancer Res* **81**, 1230-1239 (2021).

192. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet* **9**, 255-66 (2008).

193. Zhu, H. & Zhou, X. Statistical methods for SNP heritability estimation and partition: A review. *Comput Struct Biotechnol J* **18**, 1557-1568 (2020).

194. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).

195. Young, A.I. Solving the missing heritability problem. *PLoS Genet* **15**, e1008222 (2019).

196. Coelho, M.C., Pinto, R.M. & Murray, A.W. Heterozygous mutations cause genetic instability in a yeast model of cancer evolution. *Nature* **566**, 275-278 (2019).

197. Xia, J. *et al.* Bacteria-to-Human Protein Networks Reveal Origins of Endogenous DNA Damage. *Cell* **176**, 127-143 e24 (2019).

198. Vali Pour, M., Lehner, B. & Supek, F. The impact of rare germline variants on human somatic mutation processes. *bioRxiv* (2021).

199. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nat Genet* **48**, 768-76 (2016).

200. Jonsson, P. *et al.* Tumour lineage shapes BRCA-mediated phenotypes. *Nature* **571**, 576-579 (2019).

201. Dempster J. *et al.* Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv* (2019).

202. Dempster J. *et al.* Chronos: a CRISPR cell population dynamics model. *bioRxiv* (2021).

203. Starita, L.M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet* **101**, 315-325 (2017).

204. Livesey, B.J. & Marsh, J.A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol Syst Biol* **16**, e9380 (2020).

205. Wang, X., Takenaka, K. & Takeda, S. PTIP promotes DNA double-strand break repair through homologous recombination. *Genes Cells* **15**, 243-54 (2010).

206. Wu, J., Prindle, M.J., Dressler, G.R. & Yu, X. PTIP regulates 53BP1 and SMC1 at the DNA damage sites. *J Biol Chem* **284**, 18078-84 (2009).

207. Fontana, G.A. *et al.* Rif1 S-acylation mediates DNA double-strand break repair at the inner nuclear membrane. *Nat Commun* **10**, 2535 (2019).

208. Callen, E. *et al.* 53BP1 mediates productive and mutagenic DNA repair through distinct phosphoprotein interactions. *Cell* **153**, 1266-80 (2013).

209. Marechal, A. & Zou, L. DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb Perspect Biol* **5**(2013).

210. Sakamoto, S. *et al.* Werner helicase relocates into nuclear foci in response to DNA damaging agents and co-localizes with RPA and Rad51. *Genes Cells* **6**, 421-30 (2001).

211. Chen, L. *et al.* WRN, the protein deficient in Werner syndrome, plays a critical structural role in optimizing DNA repair. *Aging Cell* **2**, 191-9 (2003).

212. Oshima, J., Sidorova, J.M. & Monnat, R.J., Jr. Werner syndrome: Clinical features, pathogenesis and potential therapeutic interventions. *Ageing Res Rev* **33**, 105-114 (2017).

213. Zecevic, A. *et al.* WRN helicase promotes repair of DNA double-strand breaks caused by aberrant mismatch repair of chromium-DNA adducts. *Cell Cycle* **8**, 2769-78 (2009).

214. Chan, E.M. *et al.* WRN helicase is a synthetic lethal target in microsatellite unstable cancers. *Nature* **568**, 551-556 (2019).

215. Sehgal, R. *et al*. Lynch syndrome: an updated review. *Genes (Basel)* **5**, 497-507 (2014).

216. Goellner, E.M., Putnam, C.D. & Kolodner, R.D. Exonuclease 1-dependent and independent mismatch repair. *DNA Repair (Amst)* **32**, 24-32 (2015).

217. Tsubouchi, H. & Ogawa, H. Exo1 roles for repair of DNA double-strand breaks and meiotic crossing over in Saccharomyces cerevisiae. *Mol Biol Cell* **11**, 2221-33 (2000).

218. Bolderson, E. *et al*. Phosphorylation of Exo1 modulates homologous recombination repair of DNA double-strand breaks. *Nucleic Acids Res* **38**, 1821-31 (2010).

219. Takai, H., Xie, Y., de Lange, T. & Pavletich, N.P. Tel2 structure and function in the Hsp90-dependent maturation of mTOR and ATR complexes. *Genes Dev* **24**, 2019-30 (2010).

220. Sanders M. *et al.* Life without mismatch repair. *bioRxiv* (2021).

221. Edelmann, W. *et al*. The DNA mismatch repair genes Msh3 and Msh6 cooperate in intestinal tumor suppression.*Cancer Res* **60**, 803-7 (2000).

222. Lipkin, S.M. *et al*. MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability. *Nat Genet* **24**, 27-35 (2000).

223. Kucab, J.E. *et al*. A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836 e16 (2019).

224. Hwang, S. *et al*. HumanNet v2: human gene networks for disease research. *Nucleic Acids Res* **47**, D573-D580 (2019).

225. Wu, Y., Poulos, R.C. & Reddel, R.R. Role of POT1 in Human Cancer. *Cancers (Basel)* **12** (2020).

226. Stinus, S., Paeschke, K. & Chang, M. Telomerase regulation by the Pif1 helicase: a length-dependent effect? *Curr Genet* **64**, 509-513 (2018).

227. Sieverling, L. *et al.* Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat Commun* **11**, 733 (2020).

228. Strobel, T.*et al*. Ape1 guides DNA repair pathway choice that is associated with drug tolerance in glioblastoma. *Sci Rep* **7**, 9674 (2017).

229. Krajewska, M., Fehrmann, R.S., de Vries, E.G. & van Vugt, M.A. Regulators of homologous recombination repair as novel targets for cancer treatment. *Front Genet* **6**, 96 (2015).

230. Szklarczyk, D. *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).

231. Leiserson, M.D. *et al*. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106-14 (2015).

232. Cho, A. *et al*. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol* **17**, 129 (2016).

233. Escala-Garcia, M. *et al*. A network analysis to identify mediators of germline-driven differences in breast cancer prognosis. *Nat Commun* **11**, 312 (2020).

234. Xu, C. *et al*. Estimating genome-wide significance for whole-genome sequencing studies. *Genet Epidemiol* **38**, 281-90 (2014).

235. Button, K.S. *et al*. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* **14**, 365-76 (2013).

236. Evans, L.M. *et al*. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat Genet* **50**, 737-745 (2018).

237. Cannataro V., Mandell J. & J., T. Attribution of Cancer Origins to Endogenous, Exogenous, and Actionable Mutational Processes. *bioRxiv* (2021).

238. Alexandrov, L.B. *et al*. Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-7 (2015).

239. Dvorak, K. *et al.* Bile acids in combination with low pH induce oxidative stress and oxidative DNA damage: relevance to the pathogenesis of Barrett's oesophagus. *Gut* **56**, 763-71 (2007).

240. Inoue, M. *et al.* Induction of chromosomal gene mutations in Escherichia coli by direct incorporation of oxidatively damaged nucleotides. New evaluation method for mutagenesis by damaged DNA precursors in vivo. *J Biol Chem* **273**, 11069-74 (1998).

241. Wu, Y., Li, R., Sun, S., Weile, J. & Roth, F.P. Improved pathogenicity prediction for rare human missense variants. *Am J Hum Genet* **108**, 1891-1906 (2021).

242. Zarate, S. *et al*. Parliament2: Accurate structural variant calling at scale. *Gigascience* **9** (2020).

243. Zou, X. *et al*. Validating the concept of mutational signatures with isogenic cell models. *Nat Commun* **9**, 1744 (2018).

244. Buckley, A.R. *et al*. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC Genomics* **18**, 458 (2017).

245. Rasnic, R., Brandes, N., Zuk, O. & Linial, M. Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants. *BMC Cancer* **19**, 783 (2019).

246. Zhang, Y.D. *et al*. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat Commun* **11**, 3353 (2020).

247. Visscher, P.M. *et al*. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).

248. Rosales, R.A., Drummond, R.D., Valieris, R., Dias-Neto, E. & da Silva, I.T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8-16 (2017).

249. Sason, I., Chen, Y., Leiserson, M.D.M. & Sharan, R. A mixture model for signature discovery from sparse mutation data. *Genome Med* **13**, 173 (2021).

250. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281 e7 (2018).

251. Shen, R. & Seshan, V.E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* **44**, e131 (2016).

252. Grossman, R.L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**, 1109-12 (2016).

253. Levatic, J., Salvadores, M., Fuster-Tormo, F. & Supek, F. Mutational signatures are markers of drug sensitivity of cancer cells. *bioRxiv* (2021).

254. Eckhardt, M. *et al.* Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network. *Cancer Discov* **8**, 1474-1489 (2018).

255. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).

256. Danecek, P. & McCarthy, S.A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037-2039 (2017).

257. Rimmer, A. *et al*. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**, 912-918 (2014).

258. Kleinheinz, K. *et al.* ACEseq – allele specific copy number estimation from whole genome sequencing. *bioRxiv* (2017).

259. Saunders, C.T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28, 1811-7 (2012).

260. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594 (2018).

261. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51-4 (2003).

262. Bergstrom, E.N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).

263. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* **91**, 1033-40 (2012).

264. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-8 (2010).

265. Bhattacharyya, N.P., Skandalis, A., Ganesh, A., Groden, J. & Meuth, M. Mutator phenotypes in human colorectal carcinoma cell lines. *Proc Natl Acad Sci U S A* **91**, 6319-23 (1994).

266. Way, G.P. & Greene, C.S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac Symp Biocomput* **23**, 80-91 (2018).

267. Way, G.P., Zietz, M., Rubinetti, V., Himmelstein, D.S. & Greene, C.S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biol* **21**, 109 (2020).

268. Sønderby C., Raiko T., Maaløe L., Sønderby S. & O., W. Ladder Variational Autoencoders. *arXiv* (2016).

269. Glorot X. & Y., B. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of Machine Learning Research* (2010).

270. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

271. Traynelis, J. *et al.* Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res* **27**, 1715-1729 (2017).

272. Havrilla, J.M., Pedersen, B.S., Layer, R.M. & Quinlan, A.R. A map of constrained coding regions in the human genome. *Nat Genet* **51**, 88-95 (2019).

273. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e24 (2019).

274. Cummings, B.B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452-458 (2020).

275. Zeng, W.Z.D., Glicksberg, B.S., Li, Y. & Chen, B. Selecting precise reference normal tissue samples for cancer research using a deep learning approach. *BMC Med Genomics* **12**, 21 (2019).

276. O'Leary, N.A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-45 (2016).

277. Pearl, L.H., Schierz, A.C., Ward, S.E., Al-Lazikani, B. & Pearl, F.M. Therapeutic opportunities within the DNA damage response. *Nat Rev Cancer* **15**, 166-80 (2015).

278. Diouf, B. *et al.* Somatic deletions of genes regulating MSH2 protein stability cause DNA mismatch repair deficiency and drug resistance in human leukemia cells. *Nat Med* **17**, 1298-303 (2011).

279. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).

280. Hwang, S. *et al.* HumanNet v2: human gene networks for disease research. *Nucleic Acids Res 47*, D573-D580 (2019).

281. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-73 (2010).

282. Browning, B.L., Zhou, Y. & Browning, S.R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**, 338-348 (2018).

283. McCaw, Z.R., Lane, J.M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**, 1262-1272 (2020).

284. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).

285. Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169-175 (2017).

7 Acronyms

1KG 1000 genomes

5-FU fluorouracil

5mC 5-methylcytosine

6-4PP pyrimidine-pyrimidone (6–4) photoproduct

AID activation-induced deaminase

APE1 apurinic endonuclease 1

APOBEC apolipoprotein B mRNA editing enzyme catalytic polypeptide-like

BER base excision repair

BPDE benzopyrene-78-diol-910-epoxide

CCR constrained coding region

CNV copy number variant

CPD cyclobutane pyrimidine dimer

dHR deficient homologous recombination

DHS DNase Hypersensitivity I sites

dMMR deficient DNA mismatch repair

DNV double nucleotide variant

DSB double-strand break

dTTP deoxythymidine triphosphate

FA Fanconi anaemia

FDR False Discovery Rate

GC-NER global-genome nucleotide excision repair

GWAS genome-wide association study

HNPCC hereditary non-polyposis colorectal cancer

HR homologous recombination

IBD identity-by-descent

IC independent component

ICA independent component analysis

ICL interstrand crosslink

KL Kullback-Leibler

LD linkage disequilibrium

LDSC LD score regression

LOH loss of heterozygosity

MAF minor allele frequency

MAP MUTYH-associated polyposis

MGMT O^6-methylguanine methyltransferase

MMR DNA mismatch repair

MNV multi nucleotide variant

MSI microsatellite instability

MSS microsatellite stable

mtDNA mitochondrial DNA

MTR missense tolerance ratio

NER nucleotide excision repair

NGS next-generation sequencing

NHEJ non-homologous end joining

NMF non-negative matrix factorization

NNK 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone

NNN N'-nitrosonornicotine

PARP1 Poly(ADP-ribose) polymerase 1

PC principal component

PCA principal component analysis

PNKP polynucleotide kinase

Pol β DNA polymerase β

POLH DNA polymerase η

PTV protein-truncating variant

RDGV rare damaging germline variant

RINT rank based inverse normal transformation

ROS reactive oxygen species

RT replication timing

SKAT sequence kernel association test

SNP single-nucleotide polymorphism

SSB single strand break

ssDNA single-stranded DNA

SV structural variant

TC-NER transcription-coupled nucleotide excision repair

TGF-β transforming growth factor beta

TMZ temozolomide

TS thymidylate synthase

VAE variational autoencoder

WES whole-exome sequencing

WGS whole-genome sequencing

XP xeroderma pigmentosum