UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational approaches for utilizing mutational signatures for cancer treatment and cancer prevention

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioengineering

by

Erik N. Bergstrom

Committee in charge:

Professor Ludmil B. Alexandrov, Chair
Professor Vineet Bafna
Professor Lukas Chavez
Professor Olivier Harismendy
Professor Shankar Subramaniam

2022

The Dissertation of Erik N. Bergstrom is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To the amazing group of family, friends, and colleagues of whom

I have had the honor to interact and grow with.

EPIGRAPH

Now far ahead the Road has gone

And I must follow if I can,

Pursuing it with eager feet

Until it joins some larger way

Where many paths and errands meet.

And whither then? I cannot say

-J.R.R Tolkien*, The Lord of the Rings*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AA	Aristolochic acid
AID	Activation-induced deaminase
AUC	Area under the curve
AUROC	Area under the receiving operating curve
BER	Base excision repair pathway
BED	Browser Extensible Data
BFB	Breakage-fusion-bridge cycle
BIC	Bayesian information criteria
Bp	Base pair
CCF	Cancer cell fraction
CGC	Cancer Gene Census
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DBS	Double base substitution
DSB	Double-stranded break
ecDNA	Extrachromosomal circular DNA
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embeded
GDC	Genomic Data Commons
GDSC	Genomics of Drug Sensitivity in Cancer
H&E	Hematoxylin and eosin
HR	Homologous recombination or hazards ratio (with respect to survival analyses)

HRD	Homologous recombination deficiency
HRP	Homologous recombination proficiency
ID	Small insertion and deletion
IMD	Intermutational distance
Indel	Small insertion and deletion
Kb	Kilobase pair
LIRI	Liver cancer-hepatocellular carcinoma (virus associated)
LOH	Loss of heterozygosity
LST	Large-scale transitions
MAF	Mutation Annotation Format
Mb	Megabase pair
MBC	Metastatic breast cancer
MBS	Multi-base substitution
MEF	Mouse embryonic fibroblast
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MIL	Multiple-instance learning
MMEJ	Mircohomology-mediated end joining
MMR	Mismatch repair
MSI	Microsatellite instability
MSS	Microsatellite stable
NHEJ	Non-homologous end joining
PCA	Principal component analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes

ROC	Receiver operating characteristic curve
ROI	Region of interest
SBS	Single base substitution
ssDNA	Single-stranded DNA
SV	Structural variation
TAD	Topologically associating domain
TAI	Telomeric allelic imbalance
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
TC-NER	Transcription-couple nucleotide excision repair
TMB	Tumor mutational burden
VAF	Variant allele frequency
VCF	Variant Calling Format
VEP	ENSEMBLs Variant Effect Predictor
WSI	Whole-slide image

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents, Neil and Therese Bergstrom, for the unconditional support and love that they have provided throughout my entire life. No matter the scale or direction of my next goal, they have always been a constant force of unwavering encouragement. The successes of this dissertation would not have been possible without them. Further, I would like to thank Regan Briesacher for her enduring love and patience as I learned how to traverse the nuances and stresses of graduate school. She has always provided a listening ear and an uncompromised faith in my abilities and passions. I would also like to thank all other family and friends, past and present, who have been a part of this journey.

Importantly, I would like to sincerely thank my advisor and mentor, Ludmil Alexandrov, for his continuous source of guidance and motivation as I progressed through my work. It has been an honor to undergo my doctoral research under your support. I am proud to see how far the Alexandrov research group has evolved since our earliest days that sprouted from the development of a single script to plot SBS-96 spectra! These experiences under your mentorship paved the way for my growth into becoming an active participant in science and has presented me the opportunities for future successes of which I am most grateful.

I extend this gratitude to all former and current lab members who have endured my countless presentations, read the many iterations of manuscripts, and have attempted to make sense of my code. Further, I would like to thank all collaborators and coauthors who have made these contributions to our field what they are.

I would also like to thank all of my thesis committee members, Dr. Vineet Bafna, Dr. Lukas Chavez, Dr. Olivier Harismendy, and Dr. Shankar Subramaniam, for their insightful suggestions on the direction of my thesis.

Chapter 2, in full, is a reprint of the material as it appears in BMC Genomics 2019. Bergstrom, Erik N.; Huang, Mi Ni; Mahto, Uma; Barnes, Mark; Stratton, Michael R.; Rozen, Steven G.; Alexandrov, Ludmil B., Springer Nature, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in BMC Bioinformatics 2020. Bergstrom, Erik N.; Barnes, Mark; Martincorena, Inigo; Alexandrov, Ludmil B., Springer Nature, 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Bioinformatics 2022. Bergstrom, Erik N.; Kundu, Mousumy; Tbeileh, Noura; Alexandrov, Ludmil B., Oxford University Press, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in Nature 2022. Bergstrom, Erik N.; Luebeck, Jens; Petljak, Mia; Khandekar, Azhar; Barnes, Mark; Zhang, Tongwu; Steele, Christopher D.; Pillay, Nischalan; Landi, Maria Teresa; Bafna, Vineet; Mischel, Paul S.; Harris, Reuben S.; Alexandrov, Ludmil B., Springer Nature, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 6 contains unpublished material coauthored with Bergstrom, Erik N.; Diaz-Gay, Marcos; Abbasi, Ammal; Ladoire, Sylvain; and Alexandrov, Ludmil B. The dissertation author was the primary author of this chapter.

VITA

2017 Bachelor of Science in Bioengineering with a minor in Bioinformatics
University of California Santa Cruz

2022 Doctor of Philosophy in Bioengineering
University of California San Diego

PUBLICATIONS

Bergstrom EN, Kundu M, Tbeileh N, Alexandrov LB. Examining clustered somatic mutations with SigProfilerClusters, *Bioinformatics* (2022).

Bergstrom EN, Luebeck J, Petljak M, *et al*. Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* (2022).

Bergstrom EN, Barnes M, Martincorena I, & Alexandrov LB, Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator. *BMC Bioinformatics* 21, (2020).

Bergstrom EN, *et al*. SigProfilerMatrixGenerator: A tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* 20, (2019).

Petljak M, Dananberg A, Chu K, **Bergstrom EN**, *et al*. Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature* 607, 799–807 (2022).

Moody S, Senkin S, Islam SMA, Wang J, Nasrollahzadeh D, Penha PCC, Fitzgerald S, **Bergstrom EN**, *et al*. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. *Nature Genetics* 53, 1553–1563 (2021).

Makris EA, Sharma AK, **Bergstrom EN**, *et al*. Synchronous, yet genomically distinct, GIST offer new insights into precise targeting of tumor driver mutations. *JCO Precision Oncologoy* 5, 525-532 (2021).

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, **Bergstrom EN**, *et al*. The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101 (2020).

Consortium, ITPCAWG. Pan-cancer analysis of whole genomes. *Nature* 578, 82-93 (2020).

PREPRINTS

Bergstrom EN; Diaz-Gay M, Abbasi A, Ladoire S, and Alexandrov LB. Deep learning predicts response to platinum chemotherapy in breast and ovarian cancers. [preprint] (2022).

Otlu B, Díaz-Gay M, Vermes, I, **Bergstrom EN**, Barnes M, Alexandrov LB. Topography of mutational signatures in human cancer. bioRxiv, 2022. doi: https://doi.org/10.1101/2022.05.29.493921.

Li YR, Kandyba E, Halliwill K, Delrosario R, Tran Q, Bayani N, Wu D, Mirzoeva O, Reeves MM, Islam M, Riva L, **Bergstrom EN**, *et al*. The impact of carcinogens, obesity, and chronic inflammatory processes on mutational signatures and cancer risk in mouse tumour models. [preprint] (2022).

Mangiante L, Alcala N, Genova AD, Sexton-Oates A, Gonzalez-Perez A, Khandekar A, **Bergstrom EN**, *et al*. Disentangling heterogeneity of malignant pleural mesothelioma through deep integrative omics analyses. bioRxiv, 2021. doi: https://doi.org/10.1101/2021.09.27.461908.

Mcann JL, Cristini A, Law EK, Lee SY, Tellier M, Carpenter MA, Kim JJ, Jarvis MC, Beghè C, Salamango DJ, Brown MR, Murphy S, Temiz NA, **Bergstrom EN**, *et al.* R-loop homeostasis and cancer mutagenesis promoted by the DNA cytosine deaminase APOBEC3B (2021). doi: https://doi.org/10.1101/2021.08.30.458235.

Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, **Bergstrom EN**, *et al*. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. bioRxiv (2020). doi: https://doi.org/10.1101/2020.12.13.422570.

PATENTS

Alexandrov LB and Bergstrom EN (2021) Clustered Mutations for Treatment of Cancer. **U.S. Provisional Application Serial No. 63/289,601.**

Alexandrov LB and Bergstrom EN (2022) Artificial Intelligence Architecture for Predicting Cancer Biomarkers. **U.S Provisional Application Serial No. 63/269,033.**

ABSTRACT OF THE DISSERTATION

Computational approaches for utilizing mutational signatures for cancer treatment and cancer prevention

by

Erik N. Bergstrom

Doctor of Philosophy in Bioengineering

University of California San Diego, 2022

Professor Ludmil B. Alexandrov, Chair

The genome of a cancer cell is replete with somatic mutations imprinted by the activities of different endogenous and exogenous processes. Each mutational process exhibits a characteristic pattern of mutations, termed mutational signature. Prior work has shown that mutational signatures can be deciphered from a set of cancer genomes, thus, providing insight into the mutagenic processes that have been operative throughout the lineage of the cancer cell. Analysis of mutational signatures has had three major applications: (*i*) leveraging mutational signatures to identify environmental mutagens that cause cancer, thus, providing opportunities for developing cancer prevention strategies; (*ii*) using mutational signatures to better understand the biological mechanisms of DNA damage and repair processes; (*iii*) utilizing mutational signatures of failed DNA repair as biomarkers for targeted cancer treatment. However, the universal deployment of mutational signatures has been limited mainly by a reliance on whole-genome sequencing and downstream expert interpretation.

In this dissertation, we first develop three novel computational frameworks for exploring mutational signatures from large cohorts of cancer. We apply these approaches in a pan-cancer analysis to elucidate the mutational processes giving rise to clustered mutational events encompassing a plethora of operative endogenous and exogenous processes. Comprehensive characterization of these events reveals an enrichment within known driver genes. Importantly, clustered driver mutations are detectable from standard-of-care diagnostic tests and can serve as prognostic biomarkers for the overall survival of a cancer patient. Further, we introduce a novel form of oncogenesis, termed kyklonas, indicative of a repeated hypermutation of extrachromosomal circular DNA driven by the innate immune system.

Lastly, we propose an alternative sequencing-independent and cost-effective method for detecting mutational signatures by applying a deep learning approach to digital images of histopathological cancer slides. We demonstrate both the ability of this novel approach for detecting homologous recombination deficiency within breast and ovarian cancers as well as its clinical utility for predicting sensitivity to platinum treatment in individual cancer patients.

Chapter 1.

Introduction

1.1. The genomes of somatic and cancer cells

From the time of fertilization, each somatic cell in the human body will continuously acquire genetic mutations [1]. Most of these somatic mutations will have no effect on the relative fitness of a given cell but persist as "passengers" throughout the lineage of that cell [1-3]. A small proportion of these somatic mutations may confer a selective growth advantage for an individual cell leading to clonal expansions which could ultimately 'drive' the progression of a cancer [1]. Collectively, the accumulation of mutations across the genome of a somatic cell provides historical insight into the underlying mutational processes that have been active throughout the entire lineage dating back to the formation of the initial zygote [1]. In the case of cancer cells, subsets of these somatic mutations would have been generated prior to the occurrence of cancerous lesions while others could reflect mutator phenotypes triggered by neoplastic expansion [4, 5]. The relative mutation rates of most normal tissues are low; however, repeated exposures to potent mutagens such as ultraviolet radiation found in sunlight or benzo[a]pyrene found within tobacco smoke, can result in an elevated mutation rates, which in turn can lead to an increased risk for developing specific types of cancer indicative of individual lifestyle choices [6].

Upon cancer progression, the failure of key cellular processes, such as those involved in DNA repair, may result in a further elevation of mutation rates leading to an accelerated cancer evolution [7, 8]. Subsequent exposures, including ones due to anti-cancer therapies, such as platinum-based chemotherapy, may also sculpt the final landscape of genomic alterations [9, 10]. Thus, the final mutational landscape of a cancer genome reflects the additive effects of all mutational processes that have been operative during the lineage of the cancer cell [1].

1.2. Mutational patterns and mutational signatures

Each mutational process operative within a cell imprints a characteristic pattern of mutations, termed mutational signature, providing insight into the underlying etiology of a given cancer [4, 5]. Early studies focused on these patterns of mutations within the coding regions of *TP53*, the most mutated gene in human cancer [11], revealing distinct patterns of mutagenesis associated with exposure to ultraviolet radiation, smoking tobacco, ingestion of aflatoxin, and consumption of aristolochic acid, amongst others [12-15]. While these contributions were fundamental to our current understanding of mutational signatures, they had major limitations. For instance, these studies characterized the effects of specific mutagens but did not consider the activity of additional mutational processes that may be contributing to the final mutational pattern beyond the single process of interest [6]. Further, analyzing the mutations found within the coding regions of *TP53,* indicative of driver mutations, have likely been under strong selection throughout the evolution of the cancer and are ultimately superimposed on the associated mutational signature [1].

With the advent of next-generation sequencing technologies [16], the accumulation of large-scale cancer genomics datasets provided unprecedented opportunities for interrogating the mutational signatures across the spectrum of human cancers [4, 17-21]. Previous studies introduced a computational framework for deciphering the underlying mutational processes operative within large-scale genomic sequencing data by modelling the accumulation of somatic mutations in cancers as a blind-source separation problem using non-negative matrix factorization techniques [21]. This proposed solution was first applied to 21 whole-genome sequenced breast cancers [5] and subsequently to ~7,000 cancer genomes across 30 cancer types demonstrating the ability to extract the number of relevant processes operative within the cohort, while estimating the relative contributions of each mutational signature within each sample [4]. These studies revealed over 20 unique mutational signatures of single base substitutions [4, 5]. In a recent study, we performed mutational signature extraction across a combined ~4,600 whole-genome sequenced and ~19,000 whole-exome sequenced cancer genomes identifying 49 single base substitution, 11 doublet base substitution, and 17 small insertion and deletion signatures [18]. These findings implicated putative etiologies for a subset of these signatures including effects from both exogenous and endogenous processes in addition to processes unique to individual tissues. The details of each individual process is beyond the scope of this thesis; however, a detailed vignette for all single-base substitution, doublet-base substitution, small insertion and deletion, and copy number mutational signatures are found on the publicly available COSMIC website (https://cancer.sanger.ac.uk/signatures/).

1.2.1. DNA damage, repair, and replication

From a fundamental perspective, the mutational pattern observed in a cancer genome reflects the aggregated effects of the intrinsic infidelity of the DNA replication machinery and the combined result of DNA damage and repair [22]. While most damage inflicted on DNA by various mutagens is successfully repaired, a number of events are converted into permanent mutations that are found on the genome of a cell and the genomes of all offspring of that cell [1]. The local sequence context and topographical characteristics of each mutation are representative of the initial form of DNA damage and the subsequent process that acts upon it, which may include specific DNA replication and repair machinery [22-24]. Each of these mechanistic paths may result in various types of mutations. For instance, a damaged base pair may be depurinated leaving behind an abasic site [25]. While most of these sites will be successfully repaired through the base excision repair pathway (BER) [26, 27], a subset can result in the misincorporation of a nucleotide during replication, which commonly results in the incorporation of an adenine [28]. Alternatively, error-prone replicative machinery such as REV1 may be recruited, which incorporates cytosines opposite of abasic sites [23].

Depending on the initial form of damage, the resulting changes will leave behind a characteristic somatic mutation [22]. For example, if the initial abasic site was caused by the excision of a deaminated cytosine due to the activity of an APOBEC3 enzyme, the final mutational signature will be a combination of C:G>T:A transitions due to replication over an uncorrected abasic site (reflected by COSMIC signature SBS2) or C:G>G:C transversions due to the activity of REV1 (reflected by COSMIC signature SBS13) [23, 29]. Thus, extracting mutational signatures provides a lens to better understand the underlying processes that convert a specific type of DNA damage into the final mutation, which often includes the activity of multiple processes with alternative outcomes [22].

These final signatures are also sculpted by the proclivity of DNA repair processes to preferentially target specific strands of the DNA [23]. For instance, within coding regions of the genome, the transcription-coupled component of the nucleotide excision repair process (TC-NER) exclusively repairs damage that occurs on the transcribed strand [30]. This results in an increased mutational burden on the untranscribed strand, which is reflected in specific mutational signatures including exposure to carcinogens within tobacco smoke and exposure to ultraviolet radiation [4, 13, 18, 31]. A lack of transcriptional strand bias for these characteristic signatures may indicate a failure of transcription-coupled repair processes.

1.2.2. Environmental mutagens and somatic mutations

Mutational signatures are also driven by a plethora of external mutagenic factors spanning across physical, chemical, and biological components [22, 32]. Prior studies attributed mutational signatures to specific carcinogenic exposures within individual cancer types and further demonstrated actionable opportunities for cancer prevention [20, 33, 34]. For example, previous studies revealed that exposure to aristolochic acid (AA), a natural compound used in Chinese traditional medicine, results in a unique mutational signature (known as COSMIC SBS22) which is found in the genomes of upper urinary tract urothelial cell carcinomas [35-38], hepatocellular carcinomas [37], and renal cell carcinomas [39, 40]. Importantly, more than 50% of all liver cancers in China and Southeast Asia have been attributed to AA exposure, suggesting methods of prevention through regulation and increased screening of at-risk individuals [33, 41]. Similar methods for avoiding exposures to certain exogenous mutagenic carcinogens, such as exposure to ultraviolet light, tobacco smoke, and aflatoxin B1, have informed public health policies for limiting exposures that can ultimately lead to reducing cancer incidence rates [20, 42, 43].

While the mutational signatures of a subset of these exogenous exposures are well understood, there are many commonly observed signatures with unknown or poorly understood etiologies [18, 24, 33, 37, 43-49]. For example, while lacking an assigned etiology, COSMIC signatures SBS12 is ubiquitously found in hepatocellular carcinomas from Southern Asia but is generally absent in most cancers from North America and Western Europe [50]. This indicates an existence of previously unknown and geographically localized carcinogens that cause somatic mutations and contribute to cancer risk. Elucidating the processes underlying such mutational signatures can provide further opportunities for developing novel cancer prevention strategies [18, 51].

1.3. Mutational processes of clustered somatic mutations

Previous studies that characterized mutational signatures across human cancer considered individual somatic mutations as independent events contributing to the overall activity of a given mutational process [4, 5, 18, 19]. In practice, this assumption holds true for the majority of somatic single-base substitutions and small insertions and deletions, however, there are a subset of observed events that tend to cluster in non-random fashion [52, 53]. Clustered events, which encompass two or more simultaneously occurring somatic mutations, and their respective mutational patterns were first described in the form of doublet-base and multi-base substitutions, which are two or more adjacent base-pairs that get mutated at the same time [5, 18, 54-56]. These early studies described the mutagenic effects of DNA cross-linking caused by various exogenous factors including acetaldehyde and ultraviolet radiation, which lead to tandem base substitutions also known as doublet-base substitutions [14, 54-56]. Recent studies have illuminated additional forms of clustered mutagenesis and hypermutation, which will be summarized in the proceeding subsections [4, 5, 52, 57]. This clustering of mutations is often attributed to a combination of heterogeneous mutation rates across the genomic landscape, biophysical characteristics of exogenous carcinogens, dysregulation of endogenous processes, and the occurrence of larger events associated with genomic instability; amongst other [5, 14, 52, 57-65].

1.3.1. Variability of mutation rates across the genome

The distributions of mutations across the genome are sculpted by a range of different genomic features that act at varying resolutions [61]. At the megabase scale, each chromosome can be divided into partitioned domains with characteristic mutation rates [53, 66]. From a more refined mesoscale resolution, the occupancy of nucleosomes and the propensity of DNA to form secondary structures reveal variations in mutation frequencies often specific to individual mutational processes [67-72]. Lastly, zooming into the primary sequence of DNA reveals context motifs that, in most cases, are associated with different mutational signatures [4, 5, 18].

1.3.1.1. Chromosomal domains

At the macroscale, individual chromosomes are organized into megabase-sized chromatin domains [73]. Canonically referred to as *compartment A*, this domain is enriched for topologically associating domains (TADs) that are highly transcribed and comprised of gene rich euchromatin [73]. Further, the majority of these regions are replicated early during mitosis [74, 75]. In contrast, *compartment B* is full of TADs with gene-poor heterochromatin that are typically repressed from transcription and, in most cases, are replicated late [74, 75]. As a result of mismatch repair being coupled to DNA replication and preferentially active within early replicating, gene-rich regions, there is an enrichment of mutations found within *compartment B* [53, 57, 66]. However, inactivation of mismatch repair, such as in microsatellite unstable cancers, alleviates this discrepancy in mutation rate, indicating that the variability in mutation distribution is repair-dependent [76]. The increased availability of early replicating regions and the coupling of mismatch repair machinery to the replication fork is thought to account for the increased activity of repair within *compartment A* [76].

1.3.1.2. Mesoscale features and DNA secondary structures

Recent studies have characterized the effects on mutation rate at the mesoscale resolution [61, 67-72]. Typically, this scale encompasses shorter tracts of DNA including larger motifs and formation of secondary structures. For instance, the mutation rate within nucleosomes follows a periodicity of approximately 10 base pairs in length reflecting the alternating major and minor grooves facing toward and away from each histone [67, 68]. The relative mutation frequencies across these DNA tracts are also dependent on the underlying mutational process and are reflected by a decreased activity of nucleotide excision repair around DNA-bound proteins [68].

Additional variability of mutation rates at the mesoscale resolution is found around the formation of DNA secondary structures. For instance, regions that have a higher propensity to form non-canonical structures (non-B motifs), or those differing from the standard right-handed helical structure, have an increased rate of mutagenesis [24, 69, 77]. These regions are also enriched for recurrent hotspot mutations, which can bias downstream driver event detection algorithms [61]. Traditionally, recurrent hotspot events are implicated as being sites that have undergone positive selection in cancer development; however, the increased mutability of non-B motifs complicates the interpretability of recurrent events [69]. Specifically, palindromic sequences that have a higher propensity to form stems of hairpins and cruciforms result in an increased mutation rate mainly within the spacer sequences that form the loop structures [61]. These characteristic structures form single-stranded DNA (ssDNA) tracts that are easily damaged and accessible to enzymatic degradation resulting in the accumulation of a large number of passenger mutations that appear as hotspots across specific cancers [61].

Differences in mutation rate are also observed across complementary DNA strands associated with replication and transcription [70-72]. There is an enrichment of mutations found on the lagging strand during replication, which is driven by an increased availability of ssDNA at the replication fork [71, 72]. Damage caused to the ssDNA will be converted into permanent mutations in offspring cells given the lack of complementary strand information for correcting the DNA damage [71, 72]. As previously discussed, these strand asymmetries are lessened within early replicating regions due to the preferential activity of mismatch repair [57]. In comparison, mutation enrichments are also observed across the untranscribed strand within transcribed regions of the genome [4, 5, 18, 20, 23]. This asymmetry is attributed to the transcription-coupled nucleotide excision repair (TC-NER) complex that preferentially repairs actively transcribed genes containing bulky distortions of DNA [30]. Damage occurring on the untranscribed strand fails to halt the RNA polymerase likely escaping detection of the TC-NER complex and ultimately getting converted into a mutation after replication [30].

1.3.1.3. Primary DNA sequence motifs

The primary sequence of DNA also contributes to the variability of mutation frequencies across the genome. Specific mutational processes have a higher propensity of mutating DNA with a certain sequence context [4, 5, 19, 31]. For instance, aristolochic acid (AA) almost exclusively results in T:A>A:T transversions [33, 35]. This specificity of AA is further refined when considering the larger sequence context of the mutated base pair demonstrated by an enrichment of T:A>A:T transversions at CpTpG trinucleotides (mutated base underlined) [33, 35]. The characteristic probability of a mutational process affecting certain sequence motifs is the underpinnings of transforming mutational patterns into mutational signatures [21].

Additionally, the distribution of methylated cytosines across the genome can affect the relative mutation rate [78]. For instance, the deamination of 5-methylcytosine to thymine occurs due to spontaneous endogenous processes that results in an enrichment of C:G>T:A mutations at NpCpG sites [78]. In comparison, enzymatic deamination attributed to the APOBEC family of enzymes also results in an enrichment of C:G>T:A transitions but these are enriched at TpCpN trinucleotides and depleted at methylated CpG motifs [72].

1.3.2. Genome instability and large mutational events

On a more transient level, the mechanisms underlying larger chromosomal mutational events and other forms of genomic instability can affect the regional mutation rate leaving behind long tracts of clustered mutations [5, 52, 57-59, 65]. One of the most common forms of transient DNA damage are double-stranded breaks (DSBs). These types of lesions occur through both physiologic and pathologic mechanisms [79]. From a physiological perspective, DSBs occur as a consequence of somatic recombination and class switching during T and B cell maturation, which are repaired through the error-prone non-homologous end joining (NHEJ) pathway [79]. From a pathological perspective, the formation of DSBs can arise from both endogenous and exogenous factors including ionizing radiation, reactive oxygen species, replication across an unresolved nick in the DNA, enzymatic activity around fragile sites, and other forms of replicative and mechanical stress (reviewed here [79]). The resulting breaks from exposure to such processes are typically repaired through homologous recombination (HR) during the late S phase and G2 phase of the cell cycle [80]. During the remainder of the S phase, these DSBs are preferentially repaired through microhomology-mediated end joining (MMEJ) [80]. Lastly, NHEJ is used during both the early S phase and the G0/G1 phase [80].

To initiate the repair of a DSB through the HR pathway, a 5’ to 3’ resection of the DNA ends occurs leaving behind ssDNA [81, 82]. Any damage that occurs to these tracts of ssDNA have the potential to persist as mutations after DNA synthesis without a complementary strand to use as reference for correcting the damaged base pair [65]. As a result, long tracts of clustered mutations are commonly observed around the ends of DSBs [5, 65]. A similar 5’ to 3’ resection occurs at uncapped telomeres leaving behind highly mutable ssDNA [83]. Further, stalling or uncoupling of the DNA replication forks can leave behind temporary ssDNA which increases the rate of mutagenesis on the lagging strand during replication [70-72]. Alternative forms of genomic instability have also been shown to co-localize and to precede the formation of clustered mutations via the formation of ssDNA. For instance, chromothripsis, which involves large genomic rearrangements that are often caused during a single event and localized to specific regions of a chromosome, can co-localize with APOBEC-mediated hypermutation known as *kataegis* [84-86]. However, this co-localization is only partial since chromothriptic double-stranded breaks are preferentially reassembled through NHEJ repair, which does not perform 5’ to 3’ resection as heavily as HR repair [84].

1.3.3. Somatic mutations due to the AID/APOBEC deaminases

Another source of clustered mutagenesis arises from the activity of the APOBEC family of cytosine deaminases [52, 57, 58, 61, 87-89]. Typically, APOBEC deaminases are responsible for the anti-viral cellular response and for limiting the mobility of retrotransposon elements [90-96]. However, the APOBEC enzymes tend to also attack any available ssDNA, thus, making them a substantial contributor to the overall mutational burden of most human cancers [29, 57, 58, 88]. Specifically, the APOBEC3 enzymes give rise to at least two forms of clustered events requiring single-stranded DNA as a substrate including collections of long, processive mutations termed *kataegis* [5] as well as a more recently described collection of short, diffuse hypermutator events termed *omikli* [57]. These shorter *omikli* events are enriched in early replicating regions and are more prevalent in microsatellite stable (MSS) tumors indicating that mismatch repair (MMR) provides the opportunity for APOBEC3 mutagenesis by exposing short single-stranded DNA regions while processing mismatched bases during replication [57]. Further, the differential activity of MMR towards gene-rich regions results in an increased mutational burden of *omikli* mutations within common cancer driver genes [57].

Larger *kataegic* events are less prevalent than *omikli* events as they are likely dependent on longer tracks of single-stranded DNA, which are typically available during the repair of DSBs [58, 59, 65]. As such, there is an enrichment for *kataegic* events within 10 kilobases (kb) of detected breakpoints and have been observed in both recurrent and non-recurrent genomic regions across most cancers [60]. While a proportion of these events associate near structural breakpoints, there are a large portion of events that occur without any nearby breakpoint suggesting that there are additional sources of *kataegis* [60].

Another member of the same family of cytosine deaminases is the activation-induced deaminase (AID), which has been shown to give rise to somatic hypermutation within the immunoglobulin loci of developing B cells [97]. This reflects a highly controlled mechanism responsible for diversifying antibody affinities and isotopes to combat foreign organisms [97]. These tracts of AID-induced somatic hypermutation have also been observed in B-cell lymphomas outside of the targeted immunoglobulin loci resulting in larger *kataegic*-like events [98]. Direct replication over AID-induced lesions results in an enrichment of C:G>T:A or C:G>G:C clustered mutations at WpRpCpY contexts, where *W* reflects an adenine or thymine base, *R* reflects any purine base, and *Y* reflects any pyrimidine base [98]. This canonical AID-induced signature demonstrates the off-target effects which preferentially targets transcriptional start sites of highly transcribed genes and is on-going throughout a tumor's evolution reflecting characteristics of somatic hypermutation phenotypes [98]. Alternatively, AID-induced lesions can be processed by the mismatch repair pathway that recruits the error-prone DNA polymerase η resulting in non-canonical AID mutations, which tend to occur earlier in tumor evolution [98].

1.3.4. Exogenous exposures

Additionally, exposures to various exogenous mutation-causing carcinogens including ultraviolet radiation, benzo[a]pyrene, and platinum chemotherapeutics have been shown to cause clustered somatic mutations including previously reported doublet-base and multi-base substitutions in addition to short diffuse hypermutated events similar to APOBEC-driven *omikli* events [18, 52, 57]. While the mutational burden of these processes is high in certain cancers such as doublet-base substitutions from ultraviolet radiation exposure in skin cancers [52, 57], these events appear to have a lower burden within oncogenic-associated events compared with AID/APOBEC-induced lesions [52, 57]. However, exposure to these exogenous factors is thought to promote the recruitment of error-prone repair directed towards active genes, thus indirectly contributing to oncogenesis [52].

1.4. Clinical applications of mutational signatures for cancer treatment

Traditionally, the concept of precision oncology evolved around identifying individual mutations and other forms of genetic alterations at specific genomic coordinates that indicate failures or alterations of the function(s) of a specific protein and/or a particular pathway [99]. While these approaches show benefit for identifying individuals who would benefit from a specific therapy, they have several limitations. Typically, the phenotype of a tumor is reflective of a concert of genetic modifications that influence the functionality of a given pathway [100, 101]. These alterations can manifest as various forms of mutations as well as through epigenetic silencing or from interactions with peripheral pathways. When a deficiency arises in a DNA repair pathway, in most cases, a characteristic mutational signature develops based upon the inability of the cell to correct specific forms of DNA damage [22]. The detection of such signatures within a patient's tumor can be leveraged as a biomarker for targeting these deficient repair processes independent of the underlying cause, ultimately informing and improving subsequent personalized treatments [22, 102-107].

1.4.1. Somatic mutations due to mismatch repair deficiency

Mismatch repair corrects mis-incorporated nucleotides that occur during replication, recombination, and after some forms of DNA damage [108, 109]. As discussed in the prior sections, this pathway is preferentially active within early replicating and gene-rich regions [76, 110]. Tumors that are deficient in MMR are associated with the activity of multiple mutational signatures resulting in a strong inflammatory response due to a high burden of non-synonymous mutations known as microsatellite instability (MSI) [111]. These deficiencies can arise from inherited predispositions such as having germline variants within the MSH2 or MLH1 mismatch repair genes [112]. Alternatively, mismatch repair genes can be inactivated through sporadic somatic mutations or through methylation of regulatory regions [112]. Individuals with cancer exhibiting MSI mutational signatures benefit from treatment with immunotherapy, regardless of the underlying cause of MMR deficiency [106, 107]. Thus, detectable MMR-associated signatures present an actionable clinical biomarker for these types of treatments.

1.4.2. Somatic mutations due to deficiency of DNA polymerase epsilon

Analogous to MSI tumors, defects within DNA polymerases, such as mutations in the proofreading exonuclease domain of polymerase epsilon (POLE), can result in a hypermutator phenotype associated with mutational signatures unique from other canonical signatures [113, 114]. Specifically, defects in the exonuclease domain of POLE results in a hypermutation of C:G>A:T transversions almost exclusively at TpCpT contexts and C:G>T:A transitions at TpCpG contexts [4, 18]. In most cases, tumors harboring these defects also benefit from immunotherapy presenting an additional biomarker based on mutational signatures [115].

1.4.3. Homologous recombination repair deficiency

As previously discussed, HR is responsible for maintaining the integrity of DNA after double strand breaks [82]. Specifically, HR relies on homologous regions with identical nucleotide sequences for accurate repair of double strand breaks [82]. Upon the loss of function in this repair pathway, the cell recruits other more error-prone mechanisms including NHEJ and MMEJ to repair DSBs, which leave behind characteristic patterns of somatic mutations [116]. These patterns manifest in the forms of single base substitutions, small insertions and deletions, structural variation, and copy number mutational signatures [4, 17, 18, 24].

The detection of deficiencies within the HR pathway have traditionally been focused on the functional status of key genes including BRCA1 and BRCA2, amongst several others [47, 103, 117]. Individuals with germline defects in these genes harbor a predisposition for developing ovarian and breast cancers [118, 119]. From a clinical perspective, individuals who develop HR deficiency (HRD) benefit from treatment with compounds that increase the demand of DSB repair [120]. These compounds include PARP inhibitors or platinum chemotherapies, which lead to replication fork collapse or a direct increase in DNA damage leading to DSBs, respectively [121]. The accumulation of DSBs without the ability to repair these lesions results in selective cell death [122, 123]. The use of mutational signatures has shown promise as a more universal biomarker to detect HRD as the collection of associated signatures reflect the combined effects from all types of failures within the HR repair pathway [24, 47, 103]. These may include germline or somatic mutations to key genes, epigenetic silencing, or the alteration of genes of unknown relevance.

1.5. Contributions

Historically, analysis of mutational signatures has had three major applications: (*i*) leveraging mutational signatures to identify environmental mutagens that cause cancer, thus, providing opportunities for developing cancer prevention strategies; (*ii*) using mutational signatures to better understand the biological mechanisms of DNA damage and repair processes; (*iii*) utilizing mutational signatures of failed DNA repair as biomarkers for targeted cancer treatment. However, the universal deployment of mutational signatures within the clinical setting or for large-scale epidemiological studies has been limited by the reliance on whole-genome or whole-exome sequencing followed by extensive expert-performed bioinformatics analysis. In this dissertation, I introduce computational approaches that provide efficient, accurate, and cost-effective frameworks for detecting mutational signatures from cancer patients. Specifically in *Chapter 2*, I describe a standardized framework for classifying and exploring mutational patterns across small mutational events providing an essential preliminary step for deciphering the underlying mutational signatures. In *Chapter 3*, I outline a method for efficiently simulating synthetic and realistic background models of cancer genomes that can be used for downstream hypothesis testing. In *Chapter 4*, I introduce a method that seamlessly integrates the functionality of the tools from the previous two chapters for identifying and characterizing all classes of clustered single base substitutions and clustered small insertions and deletions. In *Chapter 5*, I demonstrate the applicability of these approaches across more than 2,500 whole-genome sequenced cancers spanning 30 different cancer types by providing a comprehensive map of clustered mutational signatures in human cancer. Additionally, I describe novel biomarkers of clustered mutations that are validated across ~10,000 whole-exome sequenced cancers and ~60,000 targeted-sequenced cancers providing an immediate translation into existing standard-of-care clinical platforms. Lastly, I introduce a novel form of oncogenesis characterized by the repeated attack of extrachromosomal circular DNA (ecDNA) via APOBEC3 deamination. In *Chapter 6*, I describe an alternative sequencing-independent method for detecting mutational signatures using routinely sampled histopathological slides from tumor samples, which circumvents the financial bottleneck of sequencing within the clinic. I demonstrate both the ability of this novel method for detecting homologous recombination deficiency within breast and ovarian cancers as well as its clinical utility for predicting sensitivity to platinum treatment in individual cancer patients. In *Chapter 7*, I summarize the main findings of this dissertation, while outlining the direct implications and current limitations of the work. Lastly, I discuss future work necessary to expand on the clinical implications underlying clustered mutagenesis and the applicability of deploying artificial intelligence-based methods within the clinic to detect mutational signatures.

Chapter 2.

SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events

Abstract

Background: Cancer genomes are peppered with somatic mutations imprinted by different mutational processes. The mutational pattern of a cancer genome can be used to identify and understand the etiology of the underlying mutational processes. A plethora of prior research has focused on examining mutational signatures and mutational patterns from single base substitutions and their immediate sequencing context. We recently demonstrated that further classification of small mutational events (including substitutions, insertions, deletions, and doublet substitutions) can be used to provide a deeper understanding of the mutational processes that have molded a cancer genome. However, there has been no standard tool that allows fast, accurate, and comprehensive classification for all types of small mutational events.

Results: Here, we present SigProfilerMatrixGenerator, a computational tool designed for optimized exploration and visualization of mutational patterns for all types of small mutational events. SigProfilerMatrixGenerator is written in Python with an R wrapper package provided for users that prefer working in an R environment. SigProfilerMatrixGenerator produces fourteen distinct matrices by considering transcriptional strand bias of individual events and by incorporating distinct classifications for single base substitutions, doublet base substitutions, and small insertions and deletions. While the tool provides a comprehensive classification of mutations, SigProfilerMatrixGenerator is also faster and more memory efficient than existing tools that generate only a single matrix.

Conclusions: SigProfilerMatrixGenerator provides a standardized method for classifying small mutational events that is both efficient and scalable to large datasets. In addition to extending the classification of single base substitutions, the tool is the first to provide support for classifying doublet base substitutions and small insertions and deletions.

SigProfilerMatrixGenerator is freely available at https://github.com/AlexandrovLab/SigProfilerMatrixGenerator with an extensive documentation at https://osf.io/s93d5/wiki/home/.

2.1. Background

Analysis of somatic mutational patterns is a powerful tool for understanding the etiology of human cancers [124]. The examination of mutational patterns can trace its origin to seminal studies that evaluated the patterns of mutations imprinted in the coding regions of *TP53* [125], the most commonly mutated gene in human cancer [11]. These early reports were able to identify characteristic patterns of single point substitutions imprinted due to smoking tobacco cigarettes, exposure to ultraviolet light, consumption of aflatoxin, intake of products containing aristolochic acid, amongst others [12-15]. The advent of massively parallel sequencing technologies [16] allowed cheap and efficient evaluation of the somatic mutations in a cancer genome. This provided an unprecedented opportunity to examine somatic mutational patterns by sequencing multiple cancer-associated genes, by sequencing all coding regions of the human genome (i.e., usually referred to as whole-exome sequencing), or even by interrogating the complete sequence of a cancer genome (i.e., an approach known as whole-genome sequencing).

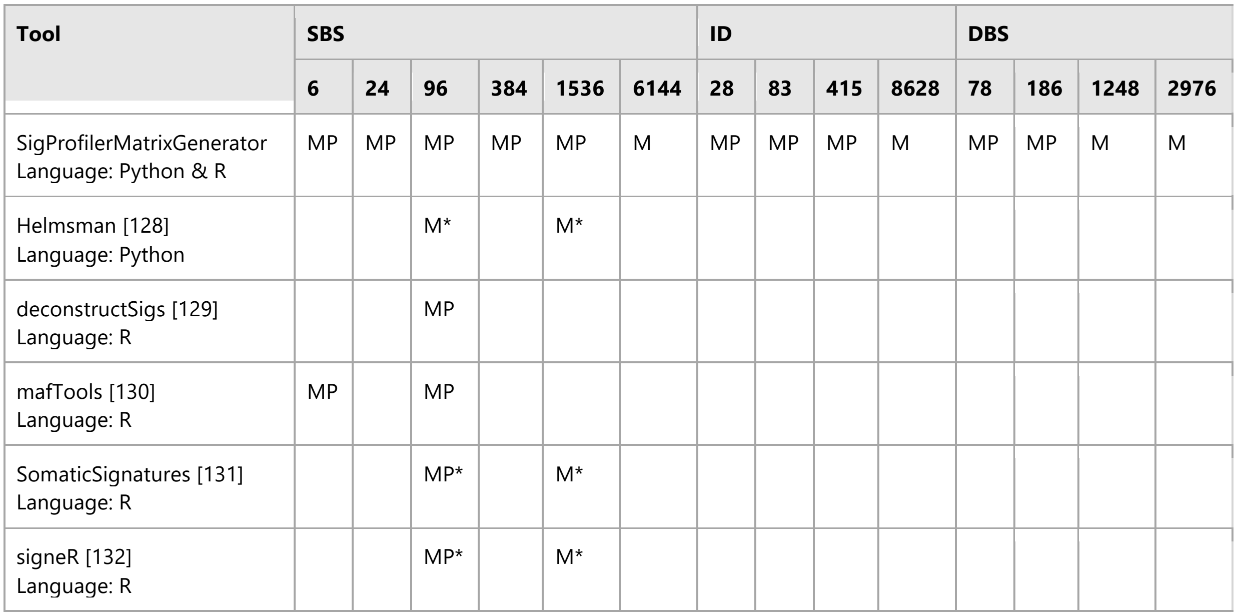
Examinations of mutational patterns from whole-genome and whole-exome sequenced cancers confirmed prior results derived from evaluating the mutations in the coding regions of *TP53* [2]. For example, the cancer genome of a lung cancer patient with a long history of tobacco smoking was peppered with somatic mutations exhibiting predominately cytosine to adenine single base substitutions [31]; the same mutational pattern was previously reported by examining mutations in *TP53* in lung cancers of tobacco smokers [13, 126]. In addition to confirming prior observations, whole-exome and whole-genome sequencing data provided a unique opportunity for identifying all of the mutational processes that have been active in the lineage of a cancer cell [19]. By utilizing mathematical modelling and computational analysis, we previously created the concept of mutational signatures and provided tools for deciphering mutational signatures from massively parallel sequencing data [21]. It should be noted that a mutational signature is mathematically and conceptually distinct from a mutational pattern of a cancer genome. While a mutational pattern of a cancer genome can be directly observed from sequencing data, a mutational signature is, in most cases, not directly observable. Rather, a mutational signature corresponds to a mathematical abstraction (i.e., a probability mass function) derived through a series of numerical approximations. From a biological perspective, a mutational signature describes a characteristic set of mutation types reflecting the activity of endogenous and/or exogenous mutational processes [19]. By examining the directly observed mutational patterns of thousands of cancer genomes, we were able to identify 49 single point substitution, 11 doublet base substitution, and 17 small insertion and deletion signatures [18] in human cancer and to propose a putative etiology for a number of these signatures.

Since we presented the very first bioinformatics framework for deciphering mutational signatures in cancer genomes [4, 21], a number of computational tools have been developed for the analysis of mutational signatures (recently reviewed in [127]). All of these tools perform a matrix factorization or leverage an approach mathematically equivalent to a matrix factorization. As such, each of these tools directly or indirectly requires generating a correct initial input matrix for subsequent analysis of mutational signatures. In principle, creating an input matrix can be examined as a transformation of the mutational catalogues of a set of cancer genomes to a matrix where each sample has a fixed number of mutation classes (also, known as mutation channels).

The majority of existing tools have focused on analyzing data using 96 mutation classes corresponding to a single base substitution and the 5′ and 3′ bases immediately adjacent to the mutated substitution. While this simple classification has proven powerful, additional classifications are required to yield greater understanding of the operative mutational processes in a set of cancer genomes [19].

Here, we present SigProfilerMatrixGenerator, a computational package that allows efficient exploration and visualization of mutational patterns. SigProfilerMatrixGenerator is written in Python with an R wrapper package provided for users that prefer working in an R environment. The tool can read somatic mutational data in most commonly used data formats such as Variant Calling Format (VCF) and Mutation Annotation Format (MAF) and it provides support for analyzing all types of small mutational events: single bases substitutions, doublet base substitutions, and small insertions and deletions. SigProfilerMatrixGenerator generates fourteen distinct matrices including ones with extended sequencing context and transcriptional strand bias, while providing publication ready visualization for the majority of these matrices. Further, the tool is the first to provide standard support for the classification of small insertions and deletions as well as the classification of doublet base substitutions that were recently used to derive the next generation of mutational signatures [18]. While SigProfilerMatrixGenerator provides much more functionality (Table 2.1), in almost all cases, it is more computationally efficient than existing approaches. Lastly, SigProfilerMatrixGenerator comes with extensive Wiki-page documentation and can be easily integrated with existing packages for analysis of mutational signatures.

Table 2.1: Matrix generation and visualization functionality of six commonly used tools. M corresponds to providing functionality to only generate a mutational matrix; MP corresponds to providing functionality to both generate and plot a mutational matrix. * indicates that a tool can perform only one of the actions in a single run; for example, Helmsman can either generate a 96 or a 1536 mutational matrix but not both in a single run.



Tool	**SBS**						**ID**				**DBS**			
	6	**24**	**96**	**384**	**1536**	**6144**	**28**	**83**	**415**	**8628**	**78**	**186**	**1248**	**2976**
SigProfilerMatrixGenerator Language: Python & R	MP	MP	MP	MP	MP	M	MP	MP	MP	M	MP	MP	M	M
Helmsman [128] Language: Python			M*		M*									
deconstructSigs [129] Language: R			MP											
mafTools [130] Language: R	MP		MP											
SomaticSignatures [131] Language: R			MP*		M*									
signeR [132] Language: R			MP*		M*									

2.1.1. Implementation

2.1.1.1. Classification of Single Base substitutions (SBSs)

A single base substitution (SBS) is a mutation in which a single DNA base-pair is substituted with another single DNA base-pair. An example of an SBS is a C:G base-pair mutating to an A:T base-pair; this is usually denoted as a C:G > A:T. The most basic classification catalogues SBSs into six distinct categories, including: C:G > A:T, C:G > G:C, C:G > T:A, T:A > A:T, T:A > C:G, and T:A > G:C. In practice, this notation has proven to be bulky and, in most cases, SBSs are referred to by either the purine or the pyrimidine base of the Watson-Crick base-pair. Thus, one can denote a C:G > A:T substitution as either a C > A mutation using the pyrimidine base or as a G > T mutation using the purine base. While all three notations are equivalent, prior research on mutational signatures [4, 5, 21] has made the pyrimidine base of the Watson-Crick base-pair a community standard. As such, the most commonly used SBS-6 classification of single base substitutions can be written as: C > A, C > G, C > T, T > A, T > C, and T > G. The classification SBS-6 should not be confused with signature SBS6, a mutational signature attributed to microsatellite instability [4].

The simplicity of the SBS-6 classification allows capturing the predominant mutational patterns when only a few somatic mutations are available. As such, this classification was commonly used in analyzing mutational patterns derived from sequencing TP53 [13, 126]. The SBS-6 classification can be further expanded by taking into account the base-pairs immediately adjacent 5′ and 3′ to the somatic mutation. A commonly used classification for analysis of mutational signatures is SBS-96, where each of the classes in SBS-6 is further elaborated using one base adjacent at the 5′ of the mutation and one base adjacent at the 3′ of the mutation. Thus, for a C > A mutation, there are sixteen possible trinucleotide (4 types of 5′ base ∗ 4 types of 3′ base): $A\underline{C}A > A\underline{A}A$, $A\underline{C}C > A\underline{A}C$, $A\underline{C}G > A\underline{A}G$, $A\underline{C}T{>}A\underline{A}T$, $C\underline{C}A > C\underline{A}A$, $C\underline{C}C > C\underline{A}C$, $C\underline{C}G > C\underline{A}G$, $C\underline{C}T > C\underline{A}T$, $G\underline{C}A > G\underline{A}A$, $G\underline{C}C > G\underline{A}C$, $G\underline{C}G > G\underline{A}G$, $G\underline{C}T > G\underline{A}T$, $T\underline{C}A > T\underline{A}A$, $T\underline{C}C > T\underline{A}C$, $T\underline{C}G > T\underline{A}G$, and $T\underline{C}T > T\underline{A}T$ (mutated based is underlined). Each of the six single base substitutions in SBS-6 has sixteen possible trinucleotides resulting in a classification with 96 possible channels (Figure 2.1a). In this notation, the mutated base is underlined and the pyrimidine base of the Watson-Crick base-pair is used to refer to each SBS. Please note that using the purine base of the Watson-Crick base-pair for classifying mutation types will require taking the reverse complement sequence of each of the classes of SBS-96. For example, $A\underline{C}G{:}T\underline{G}C > A\underline{A}G{:}T\underline{T}C$ can be written as $A\underline{C}G > A\underline{A}G$ using the pyrimidine base and as CGT > CTT using the purine base (i.e., the reverse complement sequence of the pyrimidine classification). Similarly, an $A\underline{G}C{:}T\underline{C}G > A\underline{A}C{:}T\underline{T}G$ mutation can be written as $A\underline{G}C > A\underline{A}C$

using the purine base and GCT > GTT using the pyrimidine base (i.e., the reverse complement sequence of the purine classification). In principle, somatic mutations are generally reported based on the reference strand of the human genome thus requiring converting to either the purine or the pyrimidine base of the Watson-Crick base-pair. Prior work on mutational signatures [4, 5, 21] has established the pyrimidine base as a standard for analysis of somatic mutational patterns.

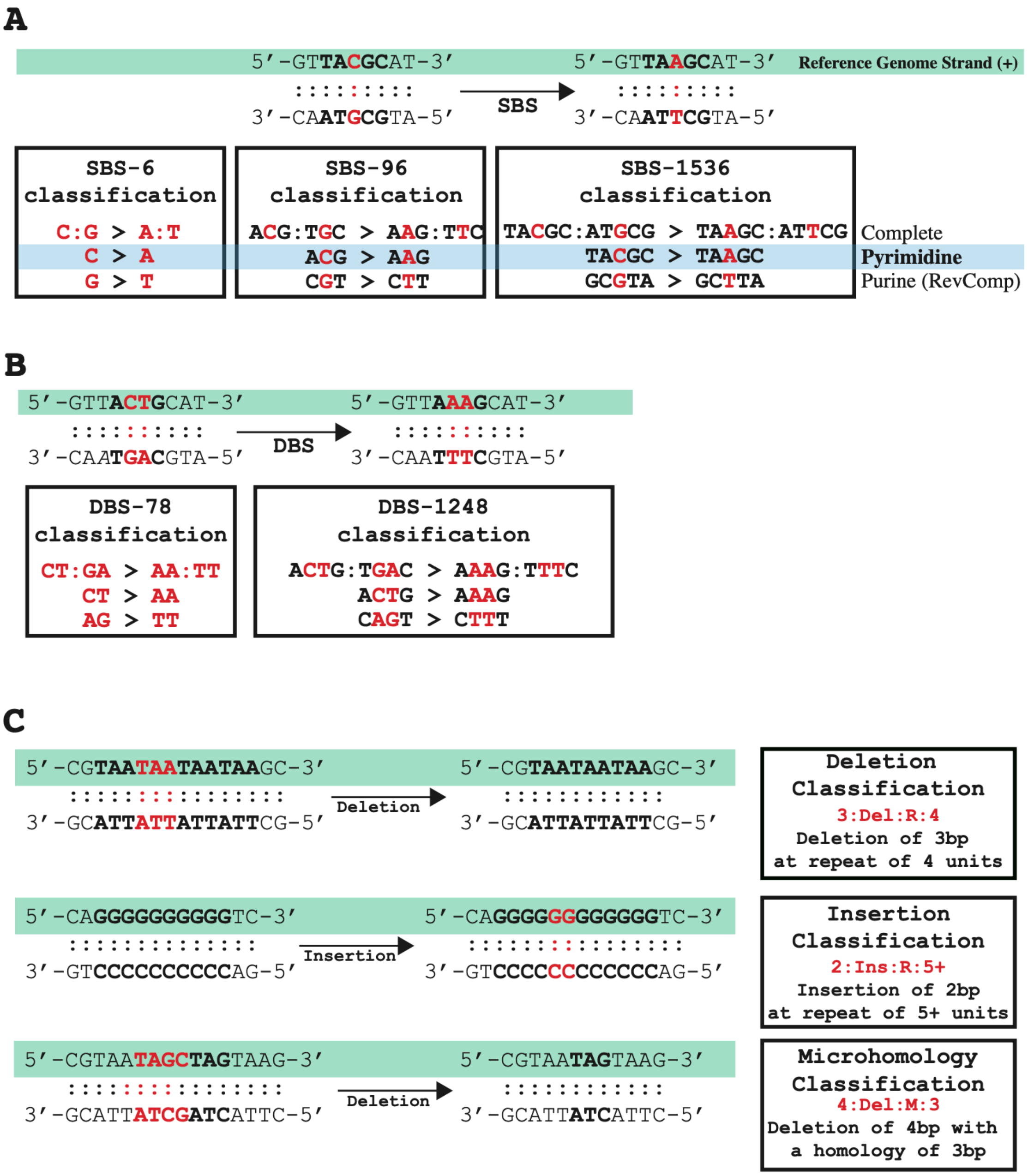


Figure 2.1: Classifications of single base substitutions, doublet base substitutions, and indels. a Classification of single base substitutions (SBSs). The complete classification of an SBS includes both bases in the Watson-Crick base-pairing. To simplify this notation, one can use either the purine or the pyrimidine base. SigProfilerMatrixGenerator uses as a standard the pyrimidine classification. b Classification of doublet base substitutions (DBSs). The complete classification of a DBS includes bases on both strands. To simplify this notation, in most cases, SigProfilerMatrixGenerator uses the maximum number of pyrimidines. c Classification of small insertions and deletions. The complete classification includes the length of the indel and the number of repeated units surrounding the sequence. For deletions at microhomologies, the length of the homology, rather than the number of repeat units surrounding the indel, is used in the classification.

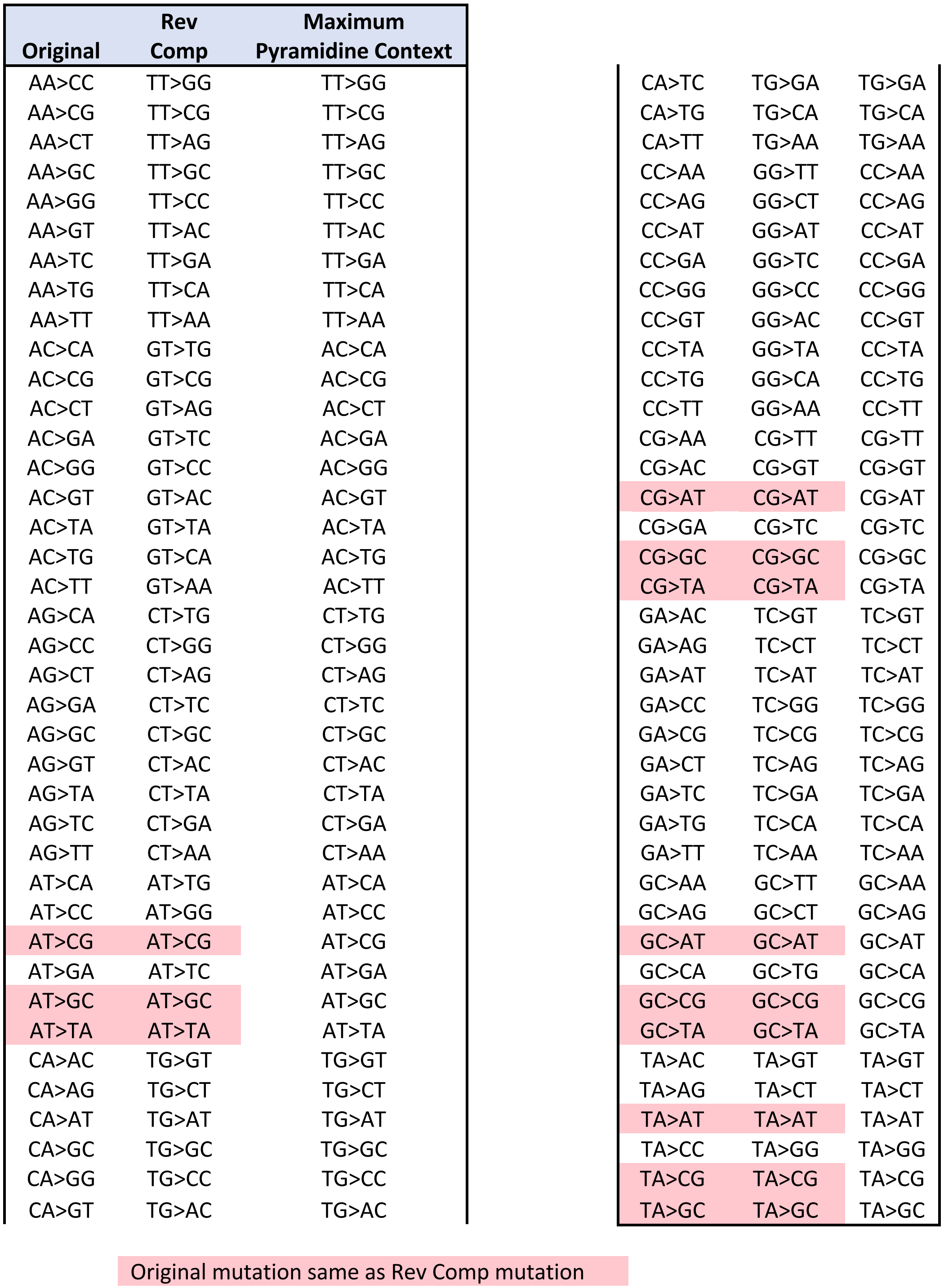
The SBS-96 has proven particularly useful for analysis of data from both whole-exome and whole-genome sequencing data [5]. This classification is both simple enough to allow visual inspection of mutational patterns and yet sufficiently complicated for separating different sources of the same type of an SBS. For example, mutational signatures analysis has identified at least 15 distinct patterns of C > T mutations each of which has been associated with different mutational processes (e.g., exposure to ultraviolet light [34], activity of the APOBEC family of deaminases [58], failure of base excision repair [133], etc.). SBS-96 can be further elaborated by including additional sequencing context. Simply by including additional 5′ and 3′ adjacent context, one can increase the resolution. For example, considering two bases 5′ and two bases 3′ of a mutation results in 256 possible classes for each SBS (16 types of two 5′ bases ∗ 16 types of two 3′ bases). Each of the six single base substitutions in SBS-6 has 256 possible pentanucleotides resulting in a classification with 1536 possible channels. Since we first introduced SBS-1536 [21], this classification has found limited use in analysis of mutational patterns. The increased number of mutational channels requires a large number of somatic mutations, which can be generally found only in whole-genome sequenced cancer exhibiting a high mutational burden (usually > 2 mutations per megabase). Nevertheless, SBS-1536 has been used to further elaborate the mutational patterns exhibited by several mutagenic processes, for example, the aberrant activity of DNA polymerase epsilon [18] or the ectopic action of the APOBEC family of cytidine deaminases [18, 21].

SigProfilerMatrixGenerator provides matrix generation support for SBS-6, SBS-96, and SBS-1536 using the commonly accepted pyrimidine base of the Watson-Crick base-pair. Further, the tool allows interrogation of transcriptional strand bias for each of these classifications and provides a harmonized visualization for all three matrices.

2.1.1.2. Classification of Doublet Base substitutions (DBSs)

A doublet base substitution (DBS) is a somatic mutation in which a set of two adjacent DNA base-pairs is simultaneously substituted with another set of two adjacent DNA base-pairs. An example of a DBS is a set of CT:GA base-pairs mutating to a set of AA:TT base-pairs, which is usually denoted as CT:GA > AA:TT (Figure 2.1b). It should be noted that a CT:GA > AA:TT mutation can be equivalently written as either a CT > AA mutation or an AG > TT mutation (note that AG > TT is the reverse complement of CT > AA). Similar to the SBSs, the complete notation for DBS has proven bulky. As such, we have previously defined a canonical set of DBSs and used this set to interrogate both mutational patterns and mutational signatures [14]. In this canonical set, DBSs are referred to using the maximum number of pyrimidine nucleotides of the Watson-Crick base-pairs; for example, an AA:TT > GT:CA mutation is usually denoted as TT > AC as this notation contains three pyrimidine nucleotides rather than the alternative AA>GT notation, which contains only a single pyrimidine nucleotide. There are several DBSs with the equivalent number of pyrimidine nucleotide in each context (e.g., AA:TT > CC:GG), in such cases, one of these notations was selected. Further, it should be noted, that some DBSs are palindromic. For example, an AT:TA > CG:GC can be written only as AT>CG since the reverse complement of 5′-AT-3′ > 5′-CG-3′ is again 5′-AT-3′ > 5′-CG-3′. Overall, the basic classification catalogues DBSs into 78 distinct categories denoted as the DBS-78 matrix (Table 2.2).

Table 2.2: DBS classification. Double Base Substitutions are classified into 78 mutational channels. The complete list of possible DBS is bulky, therefore, previous studies use the maximum pyrimidine context to collapse the number of possible mutation types. Reproduced with permission from Alexandrov *et al.*, doi:10.1101/322859.



Original	Rev Comp	Maximum Pyramidine Context
AA>CC	TT>GG	TT>GG
AA>CG	TT>CG	TT>CG
AA>CT	TT>AG	TT>AG
AA>GC	TT>GC	TT>GC
AA>GG	TT>CC	TT>CC
AA>GT	TT>AC	TT>AC
AA>TC	TT>GA	TT>GA
AA>TG	TT>CA	TT>CA
AA>TT	TT>AA	TT>AA
AC>CA	GT>TG	AC>CA
AC>CG	GT>CG	AC>CG
AC>CT	GT>AG	AC>CT
AC>GA	GT>TC	AC>GA
AC>GG	GT>CC	AC>GG
AC>GT	GT>AC	AC>GT
AC>TA	GT>TA	AC>TA
AC>TG	GT>CA	AC>TG
AC>TT	GT>AA	AC>TT
AG>CA	CT>TG	CT>TG
AG>CC	CT>GG	CT>GG
AG>CT	CT>AG	CT>AG
AG>GA	CT>TC	CT>TC
AG>GC	CT>GC	CT>GC
AG>GT	CT>AC	CT>AC
AG>TA	CT>TA	CT>TA
AG>TC	CT>GA	CT>GA
AG>TT	CT>AA	CT>AA
AT>CA	AT>TG	AT>CA
AT>CC	AT>GG	AT>CC
AT>CG	AT>CG	AT>CG
AT>GA	AT>TC	AT>GA
AT>GC	AT>GC	AT>GC
AT>TA	AT>TA	AT>TA
CA>AC	TG>GT	TG>GT
CA>AG	TG>CT	TG>CT
CA>AT	TG>AT	TG>AT
CA>GC	TG>GC	TG>GC
CA>GG	TG>CC	TG>CC
CA>GT	TG>AC	TG>AC
CA>TC	TG>GA	TG>GA
CA>TG	TG>CA	TG>CA
CA>TT	TG>AA	TG>AA
CC>AA	GG>TT	CC>AA
CC>AG	GG>CT	CC>AG
CC>AT	GG>AT	CC>AT
CC>GA	GG>TC	CC>GA
CC>GG	GG>CC	CC>GG
CC>GT	GG>AC	CC>GT
CC>TA	GG>TA	CC>TA
CC>TG	GG>CA	CC>TG
CC>TT	GG>AA	CC>TT
CG>AA	CG>TT	CG>TT
CG>AC	CG>GT	CG>GT
CG>AT	CG>AT	CG>AT
CG>GA	CG>TC	CG>TC
CG>GC	CG>GC	CG>GC
CG>TA	CG>TA	CG>TA
GA>AC	TC>GT	TC>GT
GA>AG	TC>CT	TC>CT
GA>AT	TC>AT	TC>AT
GA>CC	TC>GG	TC>GG
GA>CG	TC>CG	TC>CG
GA>CT	TC>AG	TC>AG
GA>TC	TC>GA	TC>GA
GA>TG	TC>CA	TC>CA
GA>TT	TC>AA	TC>AA
GC>AA	GC>TT	GC>AA
GC>AG	GC>CT	GC>AG
GC>AT	GC>AT	GC>AT
GC>CA	GC>TG	GC>CA
GC>CG	GC>CG	GC>CG
GC>TA	GC>TA	GC>TA
TA>AC	TA>GT	TA>GT
TA>AG	TA>CT	TA>CT
TA>AT	TA>AT	TA>AT
TA>CC	TA>GG	TA>GG
TA>CG	TA>CG	TA>CG
TA>GC	TA>GC	TA>GC

🟥 Original mutation same as Rev Comp mutation

While the prevalence of DBSs in a cancer genome is relatively low, on average a hundred times less than SBSs [18], we have previously demonstrated that a doublet base substitution is not two single base substitutions occurring simply by chance next to one another [18]. While such events are possible, across most human cancers, they will account for less than 0.1% of all observed DBSs [18]. Further, certain mutational processes have been shown to specifically generate high levels of DBSs. A flagship example is the exposure to ultraviolet light, which causes large numbers of CC > TT mutations in cancers of the skin [14]. Other notable examples are DBSs accumulating due to defects in DNA mismatch repair [18], exposure to platinum chemotherapeutics [134], tobacco smoking [20], and many others [18].

Similar to the classification of SBSs, we can expand the characterization of DBS mutations by considering the 5′ and 3′ adjacent contexts. By taking one base on the 5′ end and one base on the 3′ end of the dinucleotide mutation, we establish the DBS-1248 context. For example, a CC > TT mutation has 16 possible tetranucleotides: ACCA>ATTA, ACCC>ATTC, ACCG>ATTG, ACCT>ATTT, CCCA>CTTA, CCCC>CTTC, CCCG>CTTG, CCCT>CTTT, GCCA>GTTA, GCCC>GTTC, GCCG>GTTG, GCCT>GTTT, TCCA>TTTA, TCCC>TTTC, TCCG>TTTG, and TCCT>TTTT (mutated bases are underlined). With seventy-eight possible DBS mutations having sixteen possible tetranucleotides each, this context expansion results in 1248 possible channels denoted as the DBS-1248 context. While this classification is provided as part of SigProfilerMatrixGenerator, it has yet to be thoroughly leveraged for analysis of mutational patterns. Further, it should be noted that for most samples, the low numbers of DBSs in a single sample will make the DBS-1248 classification impractical. Nevertheless, we expect that this classification will be useful for examining hypermutated and ultra-hypermutated human cancers.

SigProfilerMatrixGenerator generates matrices for DBS-78 and DBS-1248 by predominately using the maximum pyrimidine context of the Watson-Crick base-pairs. The matrix generator also supports the incorporation of transcriptional strand bias with an integrated display of the DBS-78 mutational patterns.

2.1.1.3. Classification of small insertions and deletions (IDs)

A somatic insertion is an event that has incorporated an additional set of base-pairs that lengthens a chromosome at a given location. In contrast, a somatic deletion is an event that has removed a set of existing base-pairs from a given location of a chromosome. Collectively, when these insertions and deletions are short (usually < 100 base-pairs), they are commonly referred as small insertions and deletions (often abbreviated as indels). In some cases, indels can be complicated events in which the observed result is both a set of deleted base-pairs and a set of inserted base-pairs. For example, 5′-ATCCG-3′ mutating to 5′-ATAAAG-3′ is a deletion of CC:GG and an insertion of AAA:TTT. Such events are usually annotated as complex indels.

Indel classification is not a straightforward task and it cannot be performed analogously to SBS or DBS classifications, where the immediate sequencing context flanking each mutation was utilized to subclassify these mutational events. For example, determining the flanking sequences for deleting (or inserting) a cytosine from the sequence 5′-ATCCCCCCG-3′ is not possible as one cannot unambiguously identify which cytosine has been deleted. We recently developed a novel way to classify indels and used this classification to perform the first pan-cancer analysis of indel mutational signatures (Table 2.3) [18]. More specifically, indels (IDs) were classified as single base-pair events or longer events. A single base-pair event can be further subclassified as either a C:G or a T:A indel; usually abbreviated based on the pyrimidine base as a C or a T indel. The longer indels can also be subclassified based on their lengths: 2 bp, 3 bp, 4 bp, and 5 + bp. For example, if the sequence ACA is deleted from 5′-ATTACA[GGCGC-3′ we denote this as a deletion with length 3. Similarly, if a genomic region mutates from 5′-ATTACAGGCGC-3′ to 5′-ATTACA**CCTG**GGCGC-3′, this will be denoted as an insertion with length 4 (Figure 2.1c).

Table 2.3: ID classification. Small insertions and deletions are classified into 83 mutational channels. This classification considers the size of the ID and the repeat size surrounding the event. Events that are 1bp in length are classified by their pyrimidine base (C or T) and the number of repeated bases surrounding the event. Indels longer than 1bp are classified by the length of the event and the number of surrounding repeated units. Reproduced with permission from Alexandrov et al., doi:10.1101/322859.

Mutation class number	1bp deletions			
	Type	Base Pair	Repeat Size	Example
1	Del	T:A	1	ACCCC|T|CGCGGC (delete 1 T from a stretch of 1 Ts)
2	Del	C:G	1	ACCAA|C|TGCGGC
3	Del	T:A	2	ACCCC|T|**T**GCGGC (delete 1 T from a stretch of 2 Ts)
4	Del	C:G	2	ACCAA|C|**C**GCGGC
5	Del	T:A	3	ACCCC|T|**TT**GCGGC
6	Del	C:G	3	ACCAA|C|**CC**GCGGC
7	Del	T:A	4	ACCCC|T|**TTT**GCGGC
8	Del	C:G	4	ACCAA|C|**CCC**GCGGC
9	Del	T:A	5	ACCCC|T|**TTTT**GCGGC
10	Del	C:G	5	ACCAA|C|**CCCC**GCGGC
11	Del	T:A	6+	ACCCC|T|**TTTTT**GCGGC
12	Del	C:G	6+	ACCAA|C|**CCCCC**GCGGC

	1bp insertion			
	Type	Base Pair	Repeat Size	Example
13	Ins	T:A	0	ACCCC|T|CGCGGC (insert 1 T with no neighboring Ts)
14	Ins	C:G	0	ACCAA|C|TGCGGC
15	Ins	T:A	1	ACCCC|T|**T**GCGGC (insert 1 T with 1 neighboring T)
16	Ins	C:G	1	ACCAA|C|**C**GCGGC
17	Ins	T:A	2	ACCCC|T|**TT**GCGGC (insert 1 T with 2 neighboring Ts)
18	Ins	C:G	2	ACCAA|C|**CC**GCGGC
19	Ins	T:A	3	ACCCC|T|**TTT**GCGGC
20	Ins	C:G	3	ACCAA|C|**CCC**GCGGC
21	Ins	T:A	4	ACCCC|T|**TTTT**GCGGC
22	Ins	C:G	4	ACCAA|C|**CCCC**GCGGC
23	Ins	T:A	5+	ACCCC|T|**TTTTT**GCGGC
24	Ins	C:G	5+	ACCAA|C|**CCCCC**GCGGC

	>=2bp deletions			
	Type	Deletion size	Repeat Size	Example
25	Del	2bp	1	ACCAA|TC|AAGCGGC (delete a single 2-bp sequence with no microhomology)
26	Del	2bp	2	ACCCC|TC|**TC**GCGGC (delete a single 2-bp sequence from repeat of 2 2-bp units)
27	Del	2bp	3	ACCCC|TC|**TCTC**GCGGC
28	Del	2bp	4	ACCCC|TC|**TCTCTC**GCGGC
29	Del	2bp	5	ACCCC|TC|**TCTCTCTC**GCGGC
30	Del	2bp	6+	ACCCC|TC|**TCTCTCTCTC**GCGGC

Table 2.3: ID Classification (Continued). Small insertions and deletions are classified into 83 mutational channels. This classification considers the size of the ID and the repeat size surrounding the event. Events that are 1bp in length are classified by their pyrimidine base (C or T) and the number of repeated bases surrounding the event. Indels longer than 1bp are classified by the length of the event and the number of surrounding repeated units. Reproduced with permission from Alexandrov et al., doi:10.1101/322859.

31	Del	3bp	1	ACCAAA|TTC|AAAGCGGC
32	Del	3bp	2	ACCAAA|TTC|**TTC**AAAGCGGC
33	Del	3bp	3	ACCAAA|TTC|**TTCTTC**AAAGCGGC
34	Del	3bp	4	ACCAAA|TTC|**TTCTTCTTC**AAAGCGGC
35	Del	3bp	5	ACCAAA|TTC|**TTCTTCTTCTTC**AAAGCGGC
36	Del	3bp	6+	ACCAAA|TTC|**TTCTTCTTCTTCTTC**AAAGCGGC
37	Del	4bp	1	ACCAAAA|TCTC|AAAAGCGGC
38	Del	4bp	2	ACCAAAA|TCTC|**TCTC**AAAAGCGGC
39	Del	4bp	3	ACCAAAA|TCTC|**TCTCTCTC**AAAAGCGGC
40	Del	4bp	4	ACCAAAA|TCTC|**TCTCTCTCTCTC**AAAAGCGGC
41	Del	4bp	5	ACCAAAA|TCTC|**TCTCTCTCTCTCTCTC**AAAAGCGGC
42	Del	4bp	6+	ACCAAAA|TCTC|**TCTCTCTCTCTCTCTCTCTC**AAAAGCGGC
43	Del	5+bp	1	ACCAAAAA|TCATC|AAAAAGCGGC
44	Del	5+bp	2	ACCAAAAA|TCATC|**TCATC**AAAAAGCGGC
45	Del	5+bp	3	ACCAAAAA|TCATC|**TCATTCATC**AAAAAGCGGC
46	Del	5+bp	4	ACCAAAAA|TCATC|**TCATTCATTCATC**AAAAAGCGGC
47	Del	5+bp	5	ACCAAAAA|TCATC|**TCATTCATTCATTCATC**AAAAAGCGGC
48	Del	5+bp	6+	ACCAAAAA|TCATC|**TCATTCATTCATTCATTCATC**AAAAAGCGGC

	>=2bp insertions			
	Type	**Deletion size**	**Repeat Size**	**Example**
49	Ins	2bp	0	ACCAA|TC|AAGCGGC
50	Ins	2bp	1	ACCCC|TC|**TC**GCGGC
51	Ins	2bp	2	ACCCC|TC|**TCTC**GCGGC
52	Ins	2bp	3	ACCCC|TC|**TCTCTC**GCGGC
53	Ins	2bp	4	ACCCC|TC|**TCTCTCTC**GCGGC
54	Ins	2bp	5+	ACCCC|TC|**TCTCTCTCTC**GCGGC
55	Ins	3bp	0	ACCAAA|TTC|AAAGCGGC
56	Ins	3bp	1	ACCAAA|TTC|**TTC**AAAGCGGC
57	Ins	3bp	2	ACCAAA|TTC|**TTCTTC**AAAGCGGC
58	Ins	3bp	3	ACCAAA|TTC|**TTCTTCTTC**AAAGCGGC
59	Ins	3bp	4	ACCAAA|TTC|**TTCTTCTTCTTC**AAAGCGGC
60	Ins	3bp	5+	ACCAAA|TTC|**TTCTTCTTCTTCTTC**AAAGCGGC
61	Ins	4bp	0	ACCAAAA|TCTC|AAAAGCGGC
62	Ins	4bp	1	ACCAAAA|TCTC|**TCTC**AAAAGCGGC
63	Ins	4bp	2	ACCAAAA|TCTC|**TCTCTCTC**AAAAGCGGC
64	Ins	4bp	3	ACCAAAA|TCTC|**TCTCTCTCTCTC**AAAAGCGGC
65	Ins	4bp	4	ACCAAAA|TCTC|**TCTCTCTCTCTCTCTC**AAAAGCGGC
66	Ins	4bp	5+	ACCAAAA|TCTC|**TCTCTCTCTCTCTCTCTCTC**AAAAGCGGC
67	Ins	5+bp	0	ACCAAAAA|TCATC|AAAAAGCGGC
68	Ins	5+bp	1	ACCAAAAA|TCATC|**TCATC**AAAAAGCGGC
69	Ins	5+bp	2	ACCAAAAA|TCATC|**TCATTCATC**AAAAAGCGGC
70	Ins	5+bp	3	ACCAAAAA|TCATC|**TCATTCATTCATC**AAAAAGCGGC

Table 2.3: ID Classification (Continued). Small insertions and deletions are classified into 83 mutational channels. This classification considers the size of the ID and the repeat size surrounding the event. Events that are 1bp in length are classified by their pyrimidine base (C or T) and the number of repeated bases surrounding the event. Indels longer than 1bp are classified by the length of the event and the number of surrounding repeated units. Reproduced with permission from Alexandrov et al., doi:10.1101/322859.

71	Ins	5+bp	4	ACCAAAAA|TCATC|**TCATTCATTCATTCATC**AAAAAGCGGC
72	Ins	5+bp	5+	ACCAAAAA|TCATC|**TCATTCATTCATTCATTCATC**AAAAAGCGGC

	>=2bp deletions at micro-homologies			
	Type	**Deletion size**	**Homology Size**	**Example**
73	Del	2bp	1bp	ACCAA|TC|**T**AGCGGC or ACAA**C**|TC|AAGCGGC
74	Del	3bp	1bp	ACCCA|TTC|**T**AGCGGC or ACCC**C**|TTC|AAGCGGC
75	Del	3bp	2bp	ACCCA|TTC|**TT**AGCGGC or ACCC**TC**|TTC|AAGCGGC
76	Del	4bp	1bp	ACCCA|TATC|**T**TAGCGGC or ACCCA**C**|TATC|AAGCGGC
77	Del	4bp	2bp	ACCCA|TATC|**TA**AGCGGC or ACCCG**TC**|TATC|AAGCGGC
78	Del	4bp	3bp	ACCCA|TATC|**TAT**AGCGGC or ACCC**ATC**|TATC|AAGCGGC
79	Del	5+bp	1bp	ACCCA|TAGTC|**T**TAGCGGC or ACCCA**C**|TAGTC|AAGCGGC
80	Del	5+bp	2bp	ACCCA|TAGTC|**TA**AGCGGC or ACCCC**TC**|TAGTC|AAGCGGC
81	Del	5+bp	3bp	ACCCA|TAGTC|**TAG**AGCGGC or ACCC**GTC**|TAGTC|AAGCGGC
82	Del	5+bp	4bp	ACCCA|TAGTC|**TAGT**AGCGGC or ACCC**AGTC**|TAGTC|AAGCGGC
83	Del	5+bp	5+bp	ACCCA|TAGCCTC|**TAGCCT**AGCGGC or ACCC**AGCCTC**|TAGCCTC|AAGCGGC

Indels were further subclassified into ones at repetitive regions and ones with microhomologies (i.e., partial overlap of an indel). Note that microhomologies are not defined for indels with lengths of 1 bp as partial overlaps are not possible. For indels with lengths of 1 bp, the subclassification relied on repetitive regions that are stretches of the same base-pair referred to as homopolymers. The repeat sizes of insertions were subclassified based on their sizes of 0 bp, 1 bp, 2 bp, 3 bp, 4 bp, 5 + bp; while the repeat sizes of deletions were subclassified as 1 bp, 2 bp, 3 bp, 4 bp, 5 bp, 6 + bp (note that one cannot have a deletion with a repeat size of 0 bp). For example, if the sequence ACA is deleted from 5′-ATTACA[GGCGC-3′, this will be denotated as a deletion with length 3 at a repeat unit of 2 since there are two adjacent copies of ACAACA and only one of these copies has been deleted. Similarly, if a genomic region mutates from 5′-ATTACAGGCGC-3′ to 5′-ATTACA**CCTG**GGCGC-3′, this will be denoted as an insertion with length 4 at a repeat unit of 0 since the adjacent sequences are not repeated.

In addition to classifying indels as ones occurring at repetitive regions, a classification was performed to identify the long indels with microhomologies (i.e., partially overlapping sequences). Since almost no insertions with microhomologies were identified across more than 20,000 human cancers [18], this classification was limited to long deletions at microhomologies. Microhomologies were classified based on the length of the short identical sequence of bases adjacent to the variation. For example, if TAGTC is deleted from the sequence 5′-ACCCA TAGTAGCGGC-3′, this will be classified as a deletion of length five occurring at a microhomology site of length four because of the identical sequence TAGT located at the 3′ end of the deletion. Similarly, if TAGTC is deleted from the sequence 5′- ACCCAGTC AAGCGGC-3′, this will also be classified as a deletion of length five occurring at a microhomology site of length four because of the identical sequence AGTC located at the 5′ end of the deletion. The classification does not distinguish (i.e., subclassify) between 3′ and 5′ microhomologies since these tend to be dependent on the mutation calling algorithms. For example, 5′-ACCCA TAGTAGCGGC-3′ is the same event as 5′-ACCCATAG CGGC-3′ since in both cases a 5 bp sequence is deleted from a reference sequence 5′-ACCCATAGTCTAGTAGCGGC-3’and the result is 5′-ACCCATAGCGGC-3′. While somatic mutation callers may report different indels, our classification will annotate these indels as exactly the same mutational event.

The classification of small insertions and deletions was developed to reflect previously observed indel mutational processes. More specifically, the large numbers of small insertions and deletions at repetitive regions were observed in micro-satellite unstable tumors [112] as well as the large numbers of deletions were observed in tumors with deficient DNA double-strand break repair by homologous recombination [24]. Our classification was previously used to identify 17 indel signatures across the spectrum of human cancers [18]. SigProfilerMatrixGenerator allows generation of multiple mutational matrices of indels including ID-28 and ID-83. Importantly, the tool also generates an ID-8628 matrix that extends the ID-83 classification by providing complete information about the indel sequence for indels at repetitive regions with lengths of less than 6 bp. While SigProfilerMatrixGenerator provides this extensive indel classification, ID-8628 has yet to be thoroughly utilized for analysis of indel mutational patterns. Further, it should be noted that for most samples, the low number of indels in a single sample will make the ID-8628 classification impractical. Nevertheless, we expect that this classification will be useful for examining cancers with large numbers of indels and especially ones with deficient DNA repair. The matrix generator also supports the incorporation of transcriptional strand bias for ID-83 and the generation of plots for most of the indel matrices.

2.1.2. Incorporation of Transcription Strand bias (TSB)

The mutational classifications described above provide a detailed characterization of mutational patterns of single base substitutions, doublet base substitutions, and small insertions and deletions. Nevertheless, these classifications can be further elaborated by incorporating additional features. Strand bias is one commonly used feature that we and others have incorporated in prior analyses [4, 5, 18, 21]. While one cannot distinguish the strand of a mutation, one expects that mutations from the same type will be equally distributed across the two DNA strands. For example, given a mutational process that causes purely C:G > T:A mutations and a long repetitive sequence 5′-CGCGCGCGCGCGCGCGCCG-3′ on the reference genome, one would expect to see an equal number of C > T and G > A mutations. However, in many cases an asymmetric number of mutations are observed due to either one of the strands being preferentially repaired or one of the strands having a higher propensity for being damaged. Common examples of strand bias are transcription strand bias in which transcription-couple nucleotide excision repair (TC-NER) fixes DNA damage on one strand as part of the transcriptional process [135] and replicational strand bias in which the DNA replication process may result in preferential mutagenesis of one of the strands [70]. Strand bias can be measured by orienting mutations based on the reference strand. In the above-mentioned example, observing exclusively C > A mutations (and no G > A mutations) in the reference genome sequence 5′-CGCGCGCGCGCGCGCGCCG-3′ may mean that: (*i*) the guanine on the reference strand is protected; (*ii*) the cytosine on the reference strand is preferentially damaged; (*iii*) the guanine on the non-reference strand is preferentially damaged; (*iv*) the cytosine on the non-reference strand is protected; or (*v*) a combination of the previous four examples. In principle, a strand bias reveals additional strand-specific molecular mechanisms related to DNA damage, repair, and mutagenesis.

SigProfilerMatrixGenerator provides a standard support for examining transcriptional strand bias for single base substitutions, doublet base substitutions, and small indels. The tool evaluates whether a mutation occurs on the transcribed or the non-transcribed strand of well-annotated protein coding genes of a reference genome. Mutations found in the transcribed regions of the genome are further subclassified as: (*i*) transcribed, (*ii*) un-transcribed, (*iii*) bi-directional, or (*iv*) unknown. In all cases, mutations are oriented based on the reference strand and their pyrimidine context.

To sub-classify mutations based on their transcriptional strand bias, we consider the pyrimidine orientation with respect to the locations of well-annotated protein coding genes on a genome. For instance, when the coding strand (i.e., the strand containing the coding sequence of a gene; also known as the un-transcribed strand) matches the reference strand, a T:A > A:T will be reported as an untranscribed T > A (abbreviated as U:T > A; Figure 2.2). In this case, the template strand (i.e., the strand NOT containing the coding sequence of a gene; also known as the transcribed strand) will be complementary to the reference strand and a G:C > C:G mutation will be reported as a transcribed C > G (abbreviated as T:C > G; Figure 2.2). In rare cases, both strands of a genomic region code for a gene. Such mutations are annotated as bidirectional based on their pyrimidine context. For example, both a T:A > C:G and a A:T > G:C mutations in regions of bidirectional transcription will both be annotated as a bidirectional T > C (abbreviated as B:T > C). The outlined notations are applicable when describing mutations that are located within the transcribed regions of the genome. When a mutation is located outside of these regions, it will be classified as non-transcribed. For example, both a C:G > T:A and a G:C > A:T mutations in non-transcribed regions will be annotated as a non-transcribed C > T (abbreviated as N:C > T).

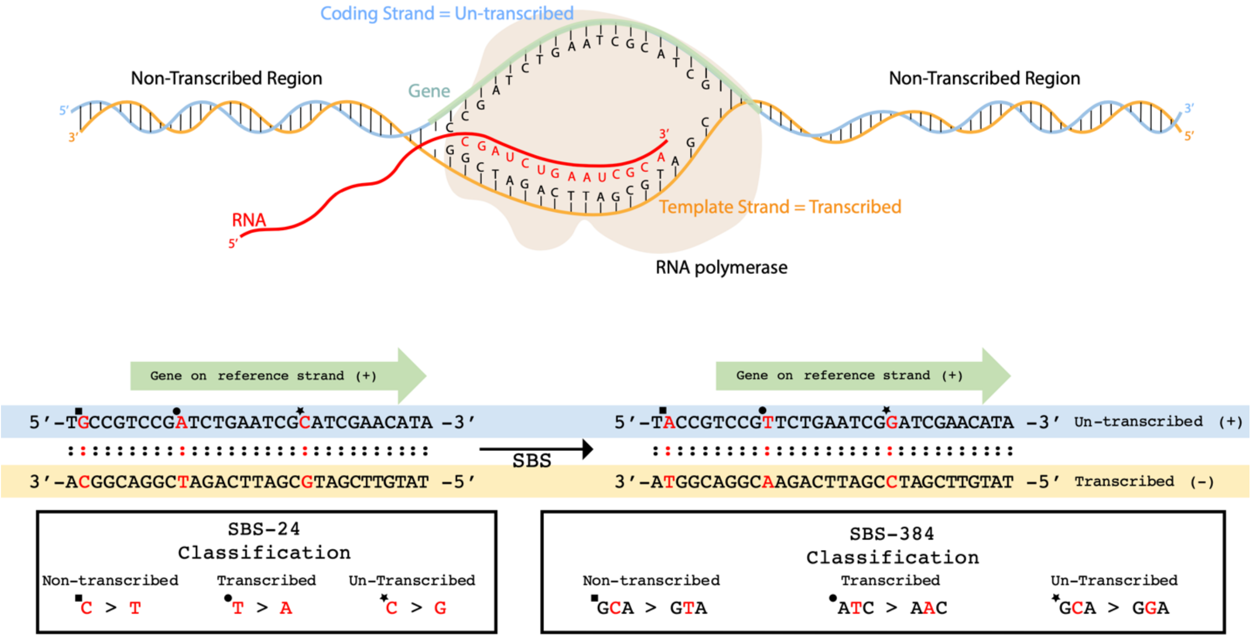


Figure 2.2: Classifications of transcriptional strand bias. a RNA polymerase uses the template strand to transcribe DNA into RNA. The strand upon which the gene is located is referred to as the coding strand. All regions outside of the footprint of a gene are referred to as non-transcribed regions. b Single point substitutions are oriented based on their pyrimidine base and the strand of the reference genome. When a gene is found on the reference strand an A:T > T:A substitution in the footprint of the gene is classified as transcribed T > A (example indicated by circle) while a C:G > G:C substitution in the footprint of the gene is classified as un-transcribed C > G (example indicated by star). Mutations outside of the footprints of genes are classified as non-transcribed (example indicated by square). Classification of single base substitutions is shown both in regard to SBS-24 and SBS-384.

When considering doublet base substitutions or small indels in transcribed regions, for certain mutational events, it is not possible to unambiguously orient these mutations. More specifically, mutations containing both pyrimidine and purine bases cannot be unequivocally attributed to a strand. For example, a TA > AT doublet substitution or a 5′-CATG-3′ deletion cannot be oriented based on the pyrimidine context as both strands contain purine and pyrimidine bases. In contrast, a GG > TT doublet substitution or a 5′-CTTCC-3′ deletion can be oriented as one of the strands is a pure stretch of pyrimidines. Somatic mutations with ambiguous strand orientation have been classified in a separate unknown category (e.g., a TA > AT doublet substitution in a transcribed region is abbreviated as Q:TA > AT). In contrast, the classification of somatic indels and DBSs with clear strand orientation has been conducted in a manner similar to the one outlined for single base substitutions.

2.1.3. Generation of mutational matrices and additional features

Prior to performing analyses, the tool requires installing a reference genome. By default, the tool supports five reference genomes and allows manually installing any additional reference genome. Installing a reference genome removes the dependency for connecting to an external database, allows for quick and simultaneous queries to retrieve information for sequence context and transcriptional strand bias, and increases the overall performance of the tool.

After successful installation, SigProfilerMatrixGenerator can be applied to a set of files containing somatic mutations from different samples. The tool supports multiple commonly used input formats and, by default, transforms the mutational catalogues of these samples to the above-described mutational matrices and outputs them as text files in a pre-specified output folder.

In addition to generating and plotting matrices from mutational catalogues, SigProfilerMatrixGenerator allows examining patterns of somatic mutations only in selected regions of the genome. The tool can be used to generate mutational matrices separately for: each individual chromosome, for the exome part of the genome, and for custom regions of the genome specified by a BED file. SigProfilerMatrixGenerator can also perform statistical analysis for significance of transcriptional strand bias for each of the examined samples with the appropriate corrections for multiple hypothesis testing using the false discovery rate (FDR) method. Overall, the tool supports the examination of significantly more mutational matrices than prior tools (Table 2.1) while still exhibiting a better performance (Figure 2.3).

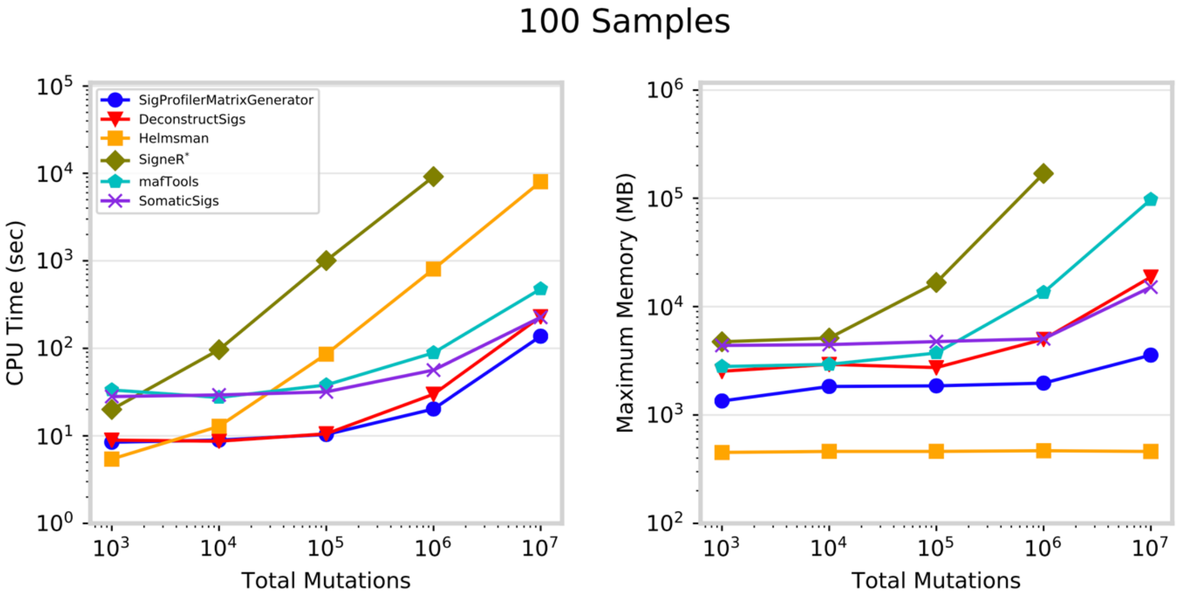


Figure 2.3: Performance for matrix generation across six commonly used tools (mutations). Each tool was evaluated separately using 100 VCF files, each corresponding to an individual cancer genome, containing total somatic mutations between 1000 and 10 million. a CPU runtime recorded in seconds (log-scale) and b maximum memory usage in megabytes (log-scale). *SigneR was unable to generate a matrix for 10^7 mutations as it exceeded the available memory of 192 gigabytes. Performance metrics exclude visualization.

2.1.4. Computational optimization

In addition to its extensive functionality (Table 2.1), the performance of SigProfilerMatrixGenerator has been optimized for analysis of large mutational datasets. More specifically, as part of the installation process, each chromosome of a given reference genome is pre-processed in a binary format to decrease subsequent query times. This pre-processing reduces a genomic base-pair to a single byte with binary flags that allow immediately identifying the reference base, its immediate sequence context, and its transcriptional strand bias. A single binary file is saved for each reference chromosome on the hard-drive; note that these binary files have similar sizes to ones of FASTA files containing the letter sequences of chromosomes.

When SigProfilerMatrixGenerator is applied to a set of input files, the tool first reformats all input files into a single file per chromosome sorted by the chromosomal positions, e.g., for a human reference genome a total of 25 files are generated: 22 files are generated for the autosomes, two files for the sex chromosomes, and one file for the genome of the mitochondria. Then, the tool processes the input data one chromosome at a time. For example, for a human reference genome, it first loads the reference binary file for chromosome one (~ 250 megabytes) and all mutations located on chromosome one across all samples are assigned to their appropriate bins in the most extensive classification (e.g., SBS-6144 for single base substitutions). Note that the binary pre-processing of the reference chromosomes makes this a linear operation with identifying the appropriate category for each mutation being a simple binary check against a binary array. After processing all mutations for a particular chromosome, the tool unloads the chromosomal data from memory and proceeds to the next chromosome. When all chromosomes have been processed, the most extensive classification is saved and iteratively collapsed to all other classifications of interests. For example, for single base substitutions, the SBS-6144 is first saved on the hard-drive and then collapsed to SBS-1536 and SBS-384. Then, SBS-1536 and SBS384 are saved on the hard-drive and collapsed, respectively, to SBS-96 and SBS-24. Similarly, SBS-96 and SBS-24 are saved on the hard-drive with SBS-24 being also collapsed to SBS-6, which is also recorded on the hard-drive. Overall, the computational improvements in SigProfilerMatrixGenerator rely on binary pre-processing of reference genomes, iterative analysis of individual chromosomes, and iterative collapsing of output matrices. These computational improvements have allowed computationally outperforming five other commonly used tools.

2.2. Results

The performance of SigProfilerMatrixGenerator was benchmarked amongst five commonly used packages: deconstructSigs [129], mafTools [130], SomaticSignatures [131], signeR [132], and Helmsman [128]. While some of these packages can perform various additional tasks (e.g., extraction/decomposition of mutational signatures), the benchmarking considered only the generation of mutational matrices. The performance was evaluated by measuring the CPU time and maximum memory necessary to generate mutational matrices based on randomly generated VCF files for 100 samples (one file per sample) with different total numbers of somatic mutations: 10^3, 10^4, 10^5, 10^6, and 10^7. To maintain consistency, each test was independently performed on a dedicated computational node with an Intel® Xeon® Gold 6132 Processor (19.25 M Cache, 2.60 GHz) and 192GB of shared DDR4–2666 RAM. In all cases, the tools generated identical SBS-96 matrices.

In addition to generating an SBS-96 matrix, SigProfilerMatrixGenerator also generates another twelve matrices including ones for indels and doublet base substitutions (Table 2.1). In contrast, all other tools can only generate a single mutational matrix exclusively for single base substitutions (Table 2.1). While offering additional functionality, SigProfilerMatrixGenerator exhibits an optimal performance and, in almost all cases, outperforms other existing tools (Figure 2.3). For example, for more than one million mutations, the tool is between 1.5 and 2 times faster compared to the next fastest tool, deconstructSigs. With the exception of Helmsman, SigProfilerMatrixGenerator requires less memory than any of the other tools making it scalable to large numbers of somatic mutations (Figure 2.3). Helmsman's low memory footprint comes at a price of a significantly slower performance for larger datasets (Figure 2.3).

Lastly, we evaluated whether the exhibited performance is independent of the number of samples by comparing the tools using a total of 100,000 somatic mutations distributed across: 10, 100, and 1000 samples (Figure 2.4). SigProfilerMatrixGenerator, deconstructSigs, Helmsman, and mafTools demonstrated an independence of sample number with respect to both CPU runtime and maximum memory usage. The memory usage of SomaticSigs is independent of sample count, however, the runtime increases linearly with the number of samples. The runtime of SigneR is somewhat independent of sample count, however, the memory increases linearly with the number of samples.

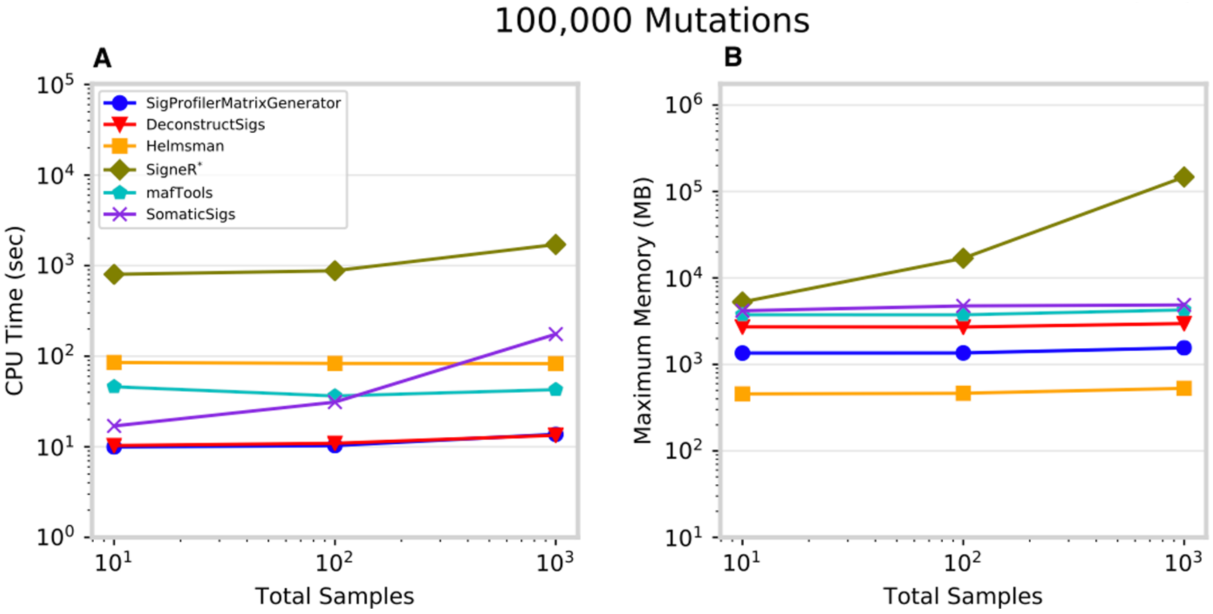


Figure 2.4: Performance for matrix generation across six commonly used tools (samples). Each tool was evaluated separately using 10, 100, and 1,000 VCF files, each corresponding to an individual cancer genome, containing a total of 100,000 somatic mutations A) CPU runtime recorded in seconds (log-scale) and B) maximum memory usage in megabytes (log-scale). Performance metrics exclude visualization.

2.3. Discussion

SigProfilerMatrixGenerator transforms a set of mutational catalogues from cancer genomes into fourteen mutational matrices by utilizing computationally and memory efficient algorithms. Indeed, in almost all cases, the tool is able to outperform other tools that generate only a single mutational matrix. SigProfilerMatrixGenerator also provides an extensive plotting functionality that seamlessly integrates with matrix generation to visualize the majority of output in a single analysis (Figure 2.5). In contrast, most other tools have plotting capabilities solely for displaying an SBS-96 matrix (Table 2.1). Currently, SigProfilerMatrixGenerator supports only classifications of small mutational events (i.e., single base substitutions, doublet base substitutions, and small insertions and deletions) as we have previously demonstrated that these classifications generalize across all types of human cancer [18]. While classifications for large mutational events (e.g., copy-number changes and structural rearrangements) have been explored by us and others [24, 136, 137] such classifications have been restricted to individual cancer types and it is unclear whether they will generalize in a pan-tissue setting.

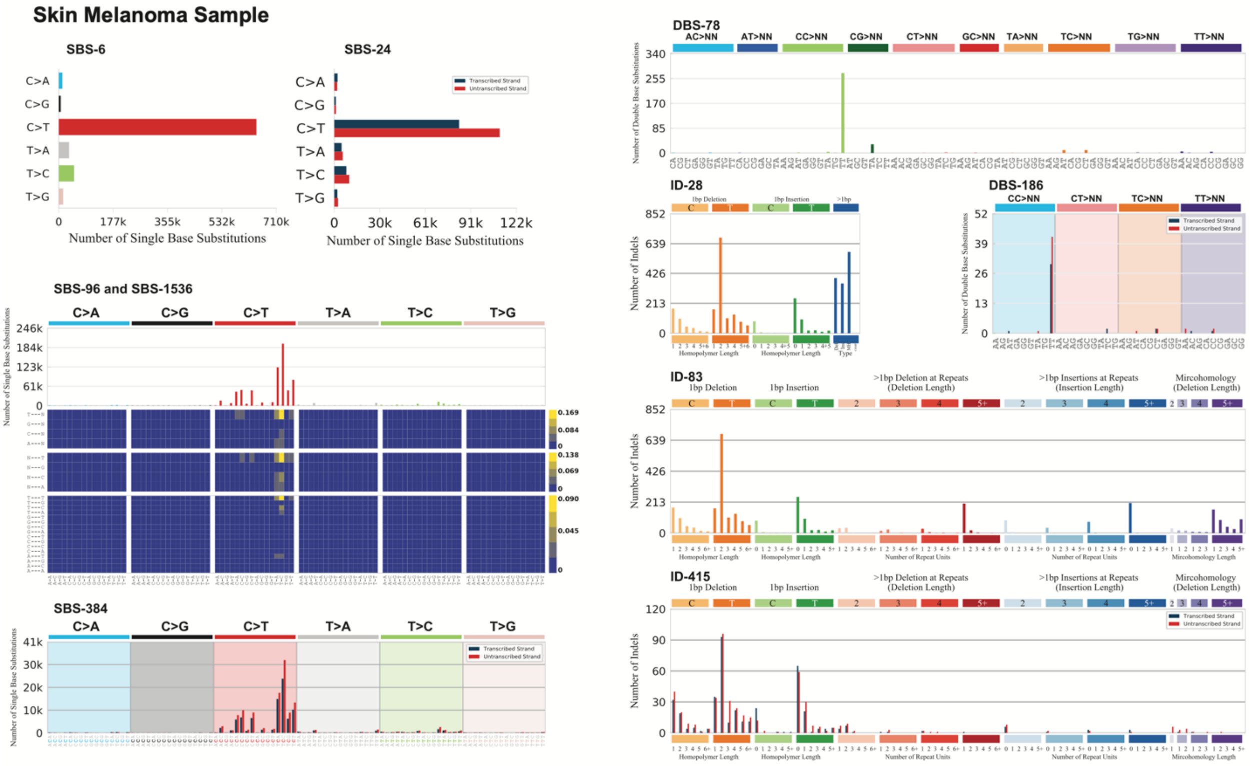


Figure 2.5: Portrait of a cancer sample. SigProfilerMatrixGenerator provides a seamless integration to visualize the majority of generated matrices. One such functionality allows the user to display all mutational plots for a sample in a single portrait. The portrait includes displaying of each of the following classifications: SBS-6, SBS-24, SBS-96, SBS-384, SBS-1536, DBS-78, DBS-186, ID-28, ID-83, and ID-415. Each of the displayed plots can also be generated in a separate file. Detailed documentation explaining each of the plots can be found at: https://osf.io/2aj6t/wiki/home/.

Importantly, SigProfilerMatrixGenerator is not a tool for analysis of mutational signatures. Rather, SigProfilerMatrixGenerator allows exploration and visualization of mutational patterns as well as generation of mutational matrices that subsequently can be subjected to mutational signatures analysis. While many previously developed tools provide support for examining the SBS-96 classification of single base substitutions, SigProfilerMatrixGenerator is the first tool to provide extended classification of single base substitutions as well as the first tool to provide support for classifying doublet base substitutions and small insertions and deletions.

2.4. Conclusions

A breadth of computational tools was developed and applied to explore mutational patterns and mutational signatures based on the SBS-96 classification of somatic single base substitutions. While the SBS-96 has yielded significant biological insights, we recently demonstrated that further classifications of single base substitutions, doublet base substitutions, and indels provide the means to better elucidate and understand the mutational processes operative in human cancer. SigProfilerMatrixGenerator is the first tool to provide an extensive classification and comprehensive visualization for all types of small mutational events in human cancer. The tool is computationally optimized to scale to large datasets and will serve as foundation to future analysis of both mutational patterns and mutational signatures. SigProfilerMatrixGenerator is freely available at https://github.com/AlexandrovLab/SigProfilerMatrixGenerator with an extensive documentation at https://osf.io/s93d5/wiki/home/.

2.5. Availability and requirements

Project name: SigProfilerMatrixGenerator.

Project home page: https://github.com/AlexandrovLab/SigProfilerMatrixGenerator

Operating system(s): Unix, Linux, and Windows.

Programming language: Python 3; R wrapper.

Other requirements: None.

License: BSD 2-Clause “Simplified” License.

Any restrictions to use by non-academics: None.

2.6. Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

2.7. Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in BMC Genomics 2019. Bergstrom, Erik N.; Huang, Mi Ni; Mahto, Uma; Barnes, Mark; Stratton, Michael R.; Rozen, Steven G.; Alexandrov, Ludmil B., Springer Nature, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 3.

Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator

Abstract

Background: Performing a statistical test requires a null hypothesis. In cancer genomics, a key challenge is the fast generation of accurate somatic mutational landscapes that can be used as a realistic null hypothesis for making biological discoveries.

Results: Here we present SigProfilerSimulator, a powerful tool that is capable of simulating the mutational landscapes of thousands of cancer genomes at different resolutions within seconds. Applying SigProfilerSimulator to 2144 whole-genome sequenced cancers reveals: (*i*) that most doublet base substitutions are not due to two adjacent single base substitutions but likely occur as single genomic events; (*ii*) that an extended sequencing context of ± 2 bp is required to more completely capture the patterns of substitution mutational signatures in human cancer; (*iii*) information on false-positive discovery rate of commonly used bioinformatics tools for detecting driver genes.

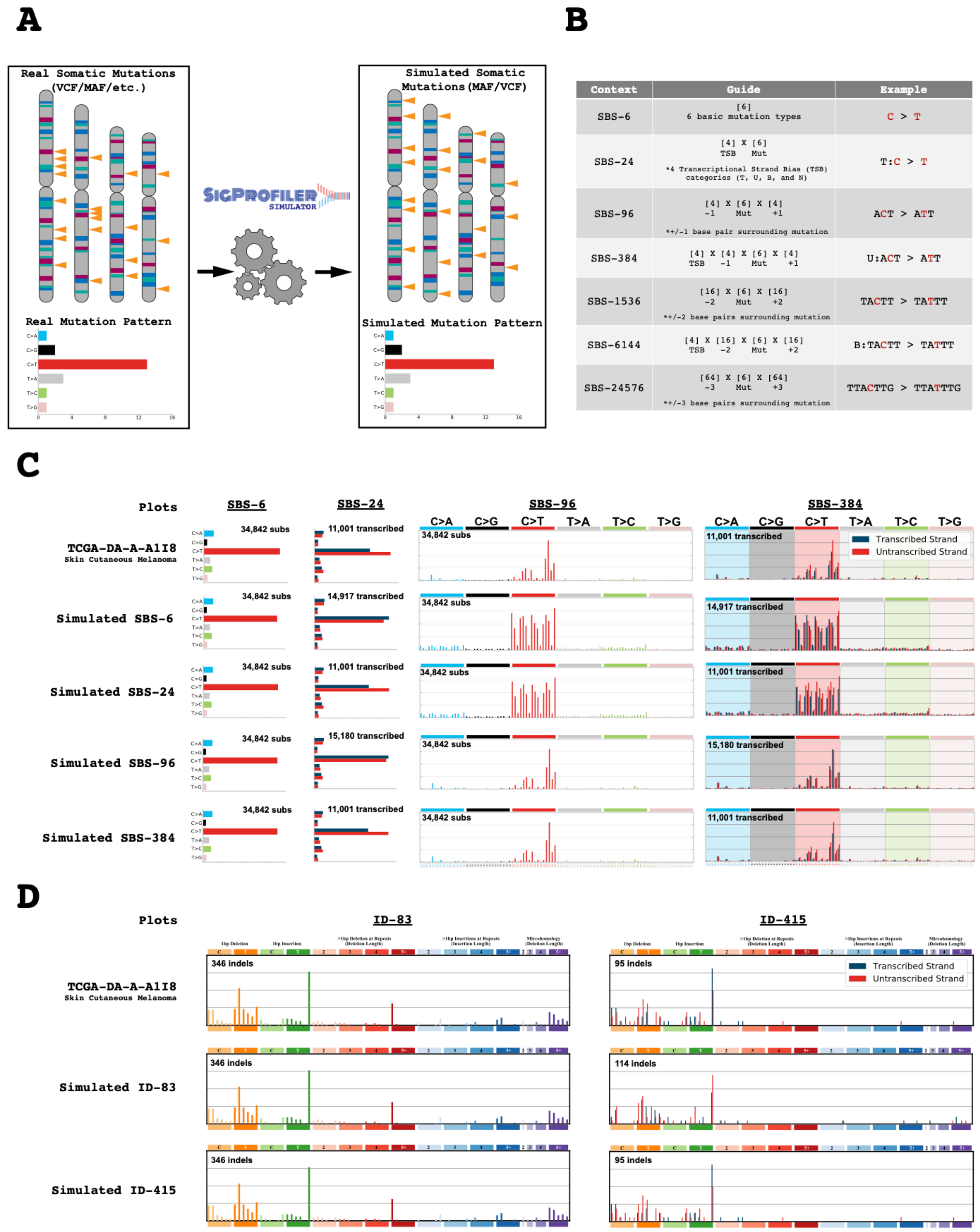
Conclusions: SigProfilerSimulator's breadth of features allows one to construct a tailored null hypothesis and use it for evaluating the accuracy of other bioinformatics tools or for downstream statistical analysis for biological discoveries. SigProfilerSimulator is freely available at https://github.com/AlexandrovLab/SigProfilerSimulator with an extensive documentation at https://osf.io/usxjz/wiki/home/.

3.1. Background

Performing a statistical evaluation to determine whether an observation is seen by chance necessitates the construction of a null hypothesis corresponding with the expected default position. An observation is generally considered statistically significant if it reflects an unlikely outcome of the null hypothesis. In most practical applications, observations seen in less than 5% of outcomes from a null distribution are considered statistically significant.

Large-scale computational analyses of cancer genomes use background mutational models to evaluate driver mutations [53, 62, 138-141], mutational signatures [4], and topographical accumulation of somatic mutations [23]. In almost all cases, a null hypothesis model of the background mutation rate is implicitly incorporated into a bioinformatics tool [21, 53, 142] and used to report statistically significant results. Here we present SigProfilerSimulator, a computationally efficient bioinformatics tool for generating sample specific mutational landscapes that match the mutational signatures operative in each sample (Figure 3.1a). SigProfilerSimulator provides a framework for generating a background mutational model for downstream statistical analyses and hypothesis testing. The tool supports generation of simulated single base substitutions (SBSs), small insertions and deletions (IDs), and doublet base substitutions (DBSs) while maintaining their patterns at different resolutions. SigProfilerSimulator is available as both a Python and an R package, provides support for commonly used data formats, and is extensively documented. To demonstrate the wide applicability of SigProfilerSimulator, we illustrate its basic functionality using a single cancer genome and then apply the tool to 2144 whole-genome sequenced cancers and to 1,024 whole-exome sequenced breast cancers to address three different questions in cancer genomics.

Figure 3.1: High-level overview illustrating the functionality of SigProfilerSimulator. **a** Schematic depiction of SigProfilerSimulator’s general functionality. The tool transforms the real somatic mutational catalog of a cancer genome into a simulated mutational catalog, while maintaining the mutational burden and the mutational pattern at a preselected resolution. **b** Summary of SBS classifications with corresponding examples at each resolution. The guide summarizes how the number of mutational channels is derived for each classification. **c** Comparing the simulated catalogues of a single cancer genome at different resolutions. Adding additional sequence context to a simulation creates a more specific and complex mutational model (e.g., SBS-96 provides greater resolution than SBS-6). Similarly, one can preserve the number of mutations in both genic and intergenic regions as well as the transcriptional strand bias by simulating with either SBS-24, SBS-384, or SBS-6144 classifications. Simulating a more complex classification of the data results in matching catalogs for all collapsed versions of the higher matrix (i.e., simulating SBS-384 ensures that the SBS-6, SBS-24, and SBS-96 simulated catalogs match the original data). **d** Comparing the simulated catalogues of a single cancer genome at different resolutions for small insertions and deletions. One can preserve the number of small insertions and deletions in genic and intergenic regions as well as the transcriptional strand bias by simulating the ID-415 classification.



Context	Guide	Example
SBS-6	[6] 6 basic mutation types	C > T
SBS-24	[4] X [6] TSB Mut *4 Transcriptional Strand Bias (TSB) categories (T, U, B, and N)	T:C > T
SBS-96	[4] X [6] X [4] -1 Mut +1 *+/-1 base pair surrounding mutation	ACT > ATT
SBS-384	[4] X [4] X [6] X [4] TSB -1 Mut +1	U:ACT > ATT
SBS-1536	[16] X [6] X [16] -2 Mut +2 *+/-2 base pairs surrounding mutation	TACTT > TATTT
SBS-6144	[4] X [16] X [6] X [16] TSB -2 Mut +2	B:TACTT > TATTT
SBS-24576	[64] X [6] X [64] -3 Mut +3 *+/-3 base pairs surrounding mutation	TTACTTG > TTATTTG

3.2. Implementation

The mutational pattern of a cancer genome can be described using distinct classification schemes reflecting the activity of mutational processes at different resolutions [143]. For example, single base substitutions can be described using only the mutated base-pair (6 possible mutational channels; known as SBS-6 classification), or the mutated base-pair with ± 1 bp context (SBS-96), or the mutated base-pair with ± 2 bp context (SBS-1536), or the mutated base-pair with ± 3 bp context (SBS-24576), etc. (Figure 3.1b) [18]. Each of these classifications can be subsequently elaborated by considering additional features [18, 143]. For example, SBS-24 extends the SBS-6 classification by including four subtypes for the six possible single base substitutions: substitutions are first split into ones in non-transcribed/intergenic regions and ones in genic regions; substitutions in genetic regions are further subclassified as ones occurring on the transcribed strand, untranscribed strand, or in regions of bi-directional transcription [143]. Similarly, the SBS-384 and SBS-6144 classifications extend SBS-96 and SBS-1536, respectively, by subclassifying each mutational channel into four: non-transcribed, transcribed, untranscribed, and bi-directional [143]. Note that, conventionally, these classifications have been displayed using the mutations only on the transcribed and untranscribed strands (e.g., 192 channel depiction for SBS-384) [4, 19, 21] since, historically, mutational patterns have been predominately investigated in whole-exome sequenced samples that provide little information about mutations outside of transcribed protein coding regions.

By preserving the pattern of mutations at a preselected resolution, SigProfilerSimulator converts a set of real somatic mutations from a cancer genome into another set of randomly generated somatic mutations (Figure 3.1a). Maintaining the mutational pattern provides an assurance that the same mutational processes are observed in both the real and the simulated cancer genome. By default, the tool projects these mutations as statistically independent events onto each chromosome by proportionately assigning mutations based on the observed rate of each mutational channel across the complete length of a preselected reference genome. The number of mutations is proportionally assigned to each chromosome based on the number of mutational channels (e.g., 96 channels reflecting trinucleotides) found on that chromosome. The tool also provides a variety of custom options for simulating mutations, including: (*i*) gender of the sample allowing appropriate incorporation of sex chromosomes; (*ii*) transcriptional strand bias allowing accurate distribution of mutations to account for the activity of transcription-coupled nucleotide excision repair; (*iii*) considering mutations as dependent sequential events where each mutation updates the observed rate of a mutational channel for a preselected reference genome, e.g., a C > T mutation at ACT trinucleotide will reduce the number of ACT trinucleotides in the reference genome by one and increase the number of ATT trinucleotides in the reference genome by one, thus, each mutation will modify the overall observed rate of a mutational channel in the genome and affect subsequent mutations; (*iv*) preserving mutational burden and mutational patterns for each chromosome instead of the complete genome, thus, the number and type of mutations assigned to each chromosome match exactly the ones observed in the original sample (Figure 3.2); (*v*) exome simulations that generate mutations only in the protein coding regions of the genome; (*vi*) adding Poisson noise to the number of mutations in each mutational channel of the original data; (*vii*) allowing the use of a probability mask that can decrease or increase the opportunity for mutations in certain parts of the genome (Figure 3.3); and several other options.

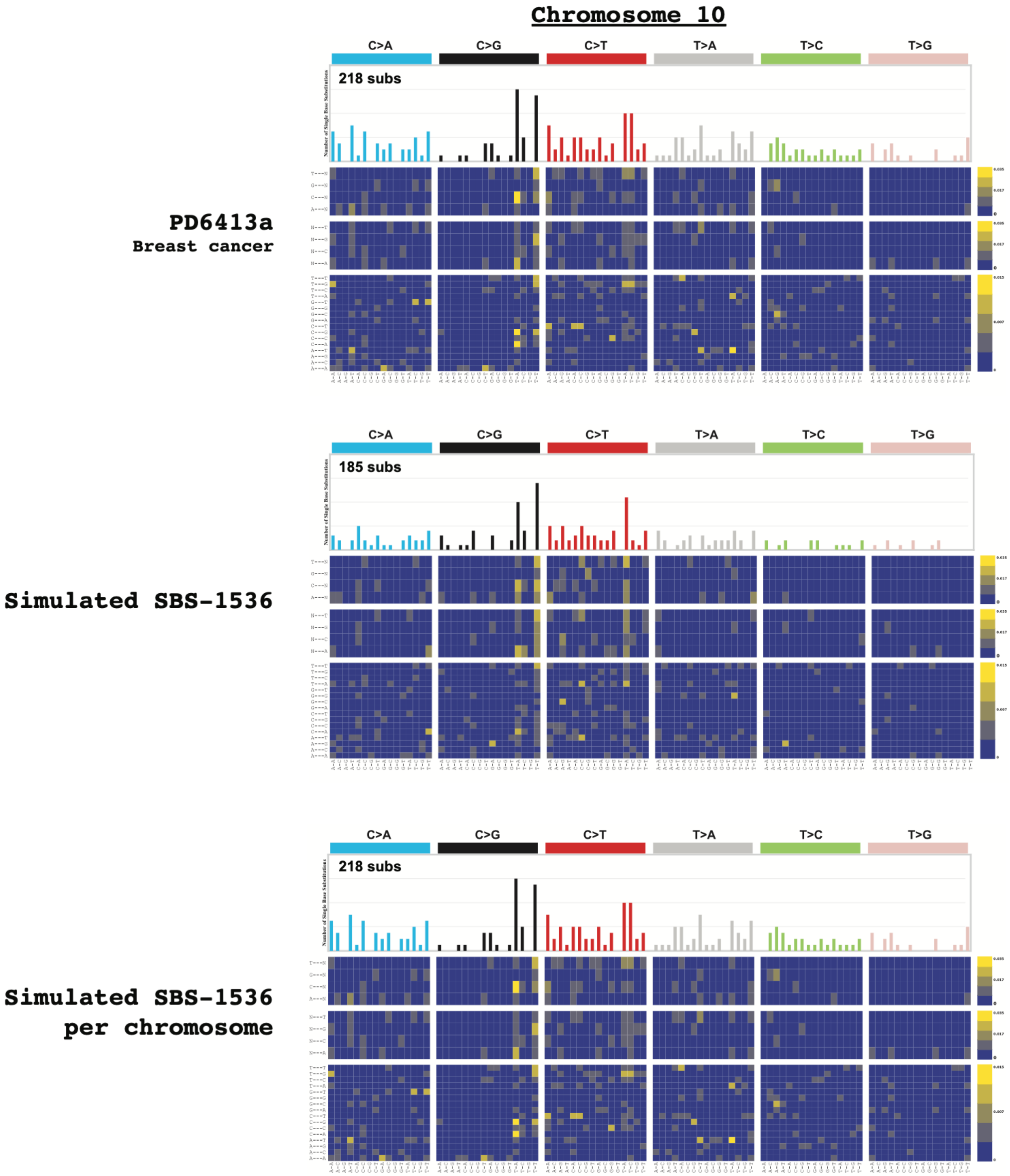


Figure 3.2: Example of an additional resolution for simulating mutational patterns supported by SigProfilerSimulator. The example illustrates the resulting patterns when maintaining the mutational burden on each chromosome and when only relying on proportionate allocation based upon the nucleotide context distribution of the reference genome. Comparison is provided for a single breast cancer sample simulated at an SBS-1536 resolution.

Additionally, one can simulate the germline variants in a (matched-)normal sample(s), which can be used for subsequent comparisons against tumor samples. With this collection of features, one can easily tailor an appropriate background mutational model for testing different biological hypotheses or for evaluating existing bioinformatics tools. Importantly, SigProfilerSimulator is computationally efficient. For example, the tool can simulate ~ 37 million somatic mutations found in the 2144 whole-genome sequenced cancers generated by Pan-cancer Analysis of Whole Genomes (PCAWG) initiative [60] within 90 s.

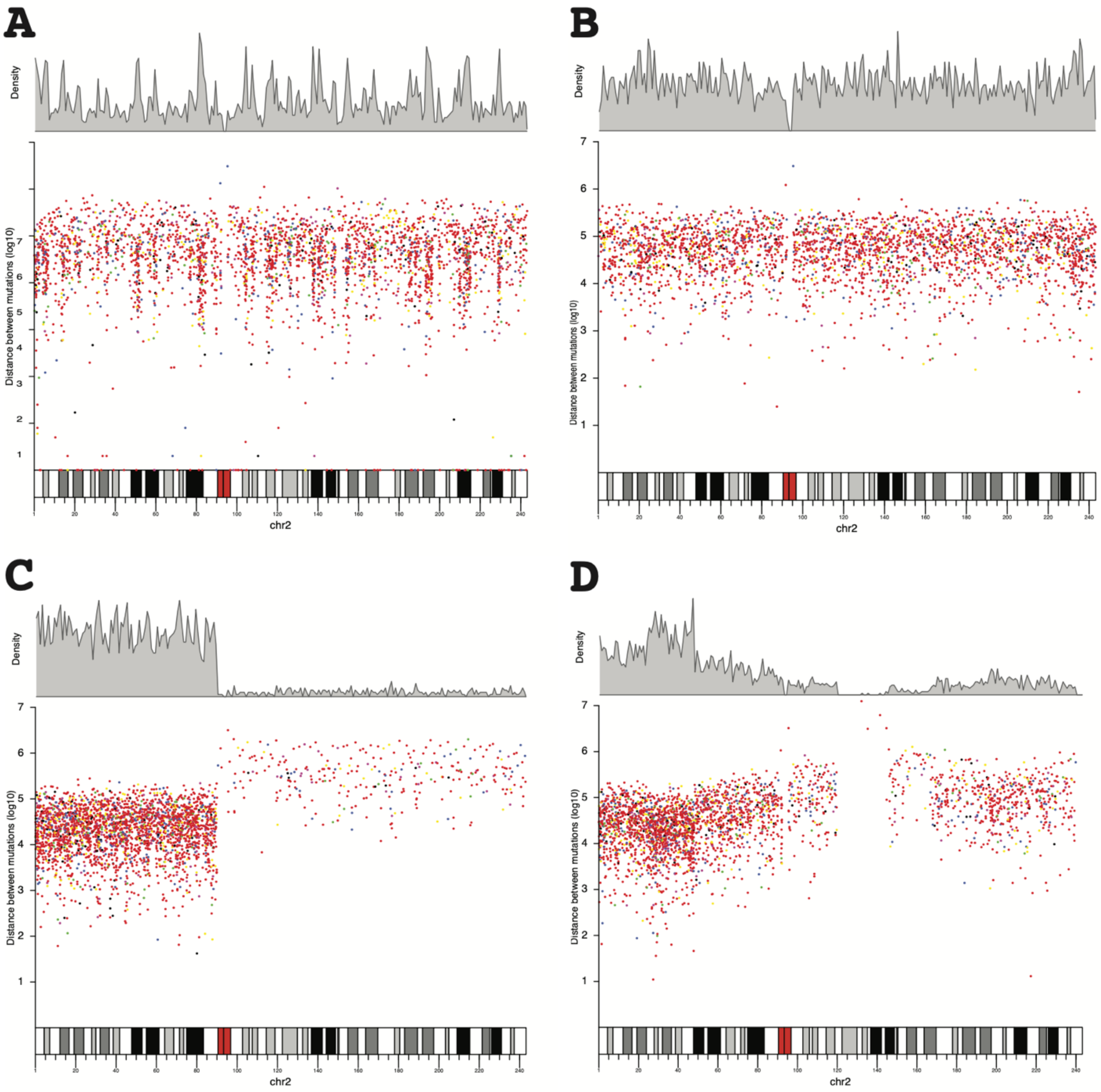


Figure 3.3: Simulating cancer genomics data using a probability mask. An example rainfall plot visualization when simulating a single TCGA melanoma sample, TCGA-DA-A-A1I8, with and without a probability mask on chromosome 2. A) Distribution of single base substitutions across chromosome 2 as found in the original sample. B) Distribution of single base substitutions across chromosome 2 when simulating the sample with default parameters. C) Distribution of single base substitutions across chromosome 2 when simulating the sample using a probability mask with 90% probability for mutations on the p arm and a 10% on the q arm. D) Distribution of single base substitutions across chromosome 2 when simulating the sample with a probability mask that varies in weights across the chromosome. All rainfall plots generated using karyoploteR [144]. Y-axes reflect log-scaled distances between adjacent mutations. X-axes reflect positions on chromosome 2 in TCGA-DA-A-A1I8. Each dot reflects a single base substitution colored using the default coloring scheme of karyoploteR.

3.3. Results

To illustrate several of SigProfilerSimulator's features, we provide a detailed visualization for a single TCGA melanoma sample: TCGA-DA-A-A1I8. Simulating TCGA-DA-A-A1I8 using the SBS-6 classification maintains the original sample's pattern for the six possible types of single base mutations, however, it also results in completely different patterns for classifications at higher resolutions (Figure 3.1c). Simulating an extended sequence context (SBS-96; trinucleotides) results in a perfect match with the original landscape when including ± 1 adjacent bases; however, it does not reflect the transcriptional strand bias observed in the sample (Figure 3.1c). As such, one can further elaborate these simulations by incorporating transcriptional strand bias (Figure 3.1c), by considering ± 2 adjacent bases (Figure 3.2), or by preserving the mutational burden and mutational patterns on each chromosome (Figure 3.2). Similarly, simulations can be performed for the different classification types for small insertions and deletions (ID-83 and ID-415; Figure 3.1d). Each of these simulations can be subsequently used to test different hypotheses. To demonstrate this capability, we applied SigProfilerSimulator to three questions in cancer genomics.

First, we used simulations to evaluate whether doublet base substitutions (e.g., CC:GG > TT:AA mutations) are two subsequent single base substitutions occurring simply by chance in adjacent genomic positions. We constructed a null hypothesis by applying the tool to the 2144 PCAWG cancer genomes. Simulations were performed considering SBSs as both statistically independent events (non-updating—simulating with replacement; each mutation has no effect on the observed rate of mutational channels) and dependent events (updating—simulating without replacement; each mutation updates the observed rate of mutational channels). Each sample was simulated 1000 times providing a distribution of doublet base substitutions. After simulating the SBS-96 context for each PCAWG sample, we examined the number of single base substitutions occurring next to one another on the genome simply by chance. For example, in the sample SP99325 (LIRI), we observed on average approximately 23 pairs of adjacent SBSs when considering mutations as statistically independent events and 14 pairs of adjacent SBSs when considering mutations as dependent events (Figure 3.4a). In contrast, the actual sample contains 303 doublet base substitutions indicating a 22-fold and a 13-fold enrichment compared to the null hypothesis, respectively.

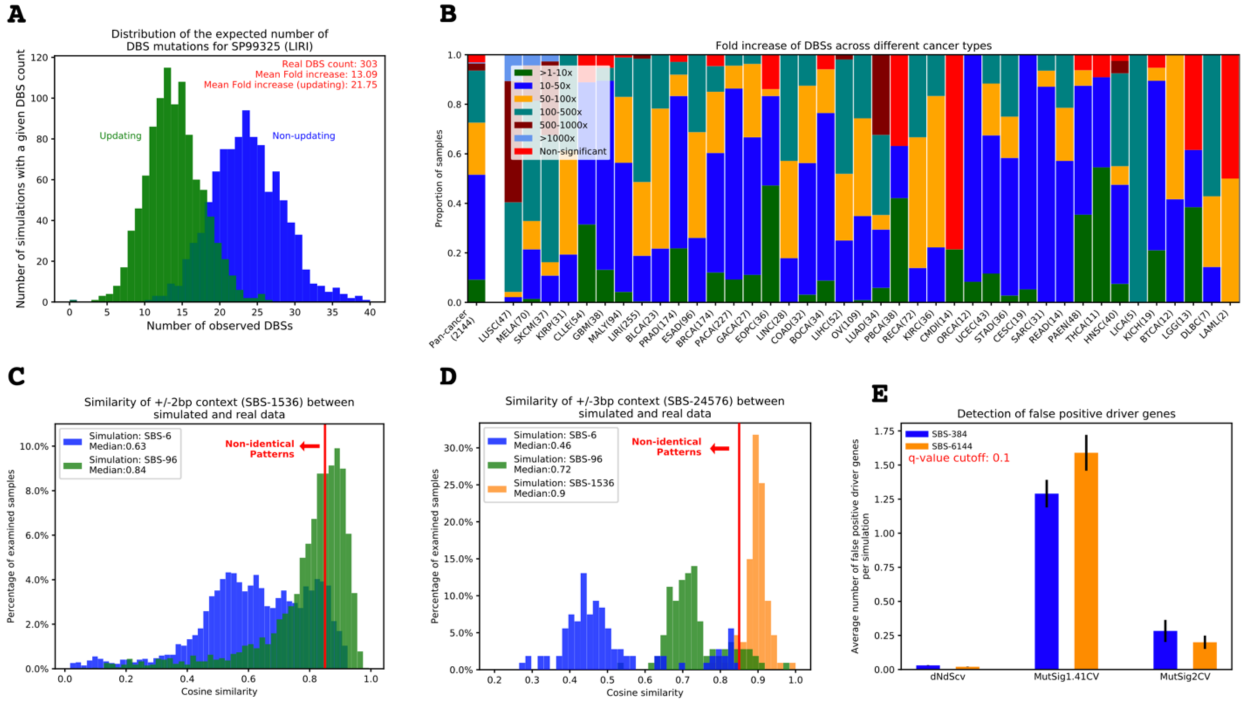


Figure 3.4: Applying SigProfilerSimulator to three distinct cancer genomics problems. **a** Distribution of the expected number of doublet base substitutions (DBSs) due to the adjacent single base substitutions (SBSs) observed by chance for the PCAWG sample SP99325. The distributions represent the results from 1000 simulations of the mutational pattern of SP99325 treating mutations as statistically independent events (blue) and 1000 simulations of the mutational pattern of SP99325 treating mutations as dependent events. **b** The fold increase of DBSs observed in the original PCAWG samples and the average number of DBSs observed in our simulations. The mutational pattern of each sample was generated 1000 times considering somatic mutations as statistically independent events. **c** Comparing the similarities of mutational patterns at ± 2 bp context (SBS-1536) between real and simulated PCAWG samples. Simulations were performed at SBS-6 and SBS-96 resolutions. **d** Comparing the similarities of mutational patterns at ± 3 bp context (SBS-24576) between real and simulated PCAWG samples. Simulations were performed at SBS-6, SBS-96, and SBS-1536 resolutions. **e** Evaluating the false-positive rates of MutSigCV1.41, MutSigCV2, and dNdScv driver detection tools using SigProfilerSimulator. All TCGA breast cancer WES samples were simulated 100 times and examined for driver mutations using both MutSigCV and dNdScv. The average number of significant driver genes are plotted using a recommended q-value cutoff of 0.10.

The results indicate that it is highly unlikely that the majority of observed doublet base substitutions in SP99325 are the result of two adjacent SBS events. Applying the same approach to all PCAWG samples reveals between 10- and 1000-fold increase of the real number of DBSs compared to simulated data (Figure 3.4b, Figure 3.5). These results confirm the belief that the vast majority of doublet base substitutions in human cancer are not due to adjacent single base substitutions. Rather, doublet base substitutions are likely due either to single genomic events or to higher mutagenic propensities of certain regions of the human genome. Indeed, we recently derived mutational signatures of doublet base substitutions across the PCAWG dataset [18]. Nevertheless, it is important to remember that, especially for hyper-mutated samples, some of the observed DBSs may be due to having two single base substitutions occurring by chance in adjacent positions (Figure 3.4a).

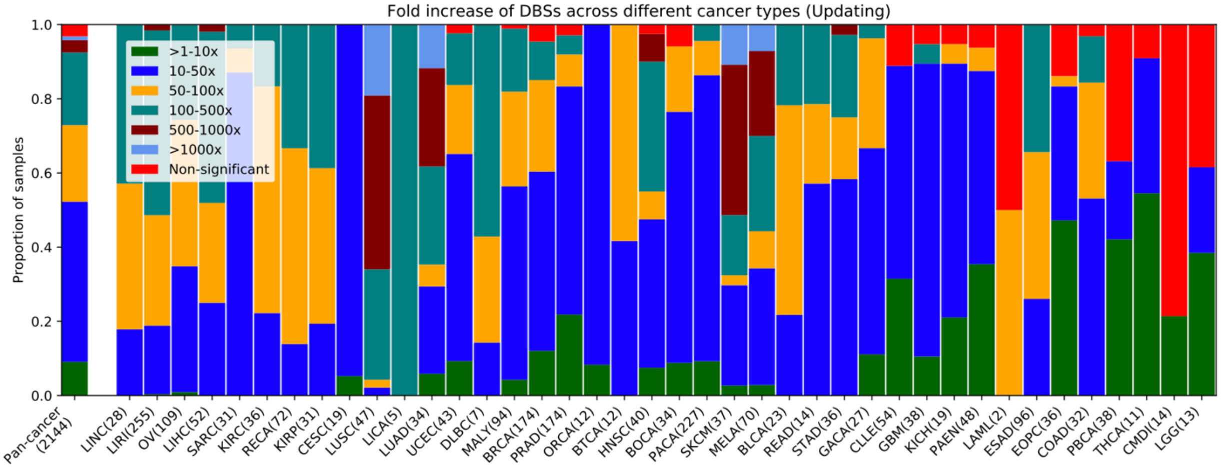


Figure 3.5: Evaluating the expected rates of DBSs for mutations simulated as dependent events. The fold increase of DBSs observed in the original PCAWG samples and the average number of DBSs observed in our simulations. The mutational pattern of each sample was generated 1000 times considering somatic mutations as dependent events.

Second, we evaluated whether incorporating additional sequence context 5′ and 3′ of single base substitutions increases the specificity of the mutational patterns observed in cancer genomes [18]. Here, we considered two mutational patterns to be the same if their cosine similarity is more than 0.85 (Figure 3.6; Methods). Specifically, we simulated the PCAWG dataset at different resolutions (viz., SBS-6, SBS-96, and SBS-1536; Figure 3.1b) and compared them to the patterns of mutations observed in the real samples (Figure 3.4c,d). Comparing the ± 2 bp context of data simulated using SBS-6 to the ± 2 bp context of the real data demonstrated that for almost all samples the SBS-6 simulations do not capture the ± 2 bp context as 91% of samples exhibited a cosine similarity below 0.85. Similarly, only half of the samples simulated using SBS-96 (i.e., ± 1 bp) had consistent ± 2 bp context when compared to the real data (44% below 0.85; Figure 3.4c). This demonstrates that the mutational patterns of the examined cancer genomes exhibit additional specificity for ± 2 bp adjacent to single base substitutions; note that ± 2 bp contains within itself the ± 1 bp classification. In contrast, comparing the ± 3 bp context of data simulated using SBS-1536 demonstrated that the ± − 2 bp context captures the patterns observed at ± 3 bp for almost all samples (only 6.5% of samples below 0.85; Figure 3.4d). Overall, these results suggest that the SBS-1536 classification is necessary to capture additional information for a set of signatures beyond SBS-6 and SBS-96. Moreover, extending this classification to ± 3 bp (SBS-24576) is likely not necessary as the SBS-1536 classification already captures the patterns of ± 3 bp for majority of the examined cancer samples.

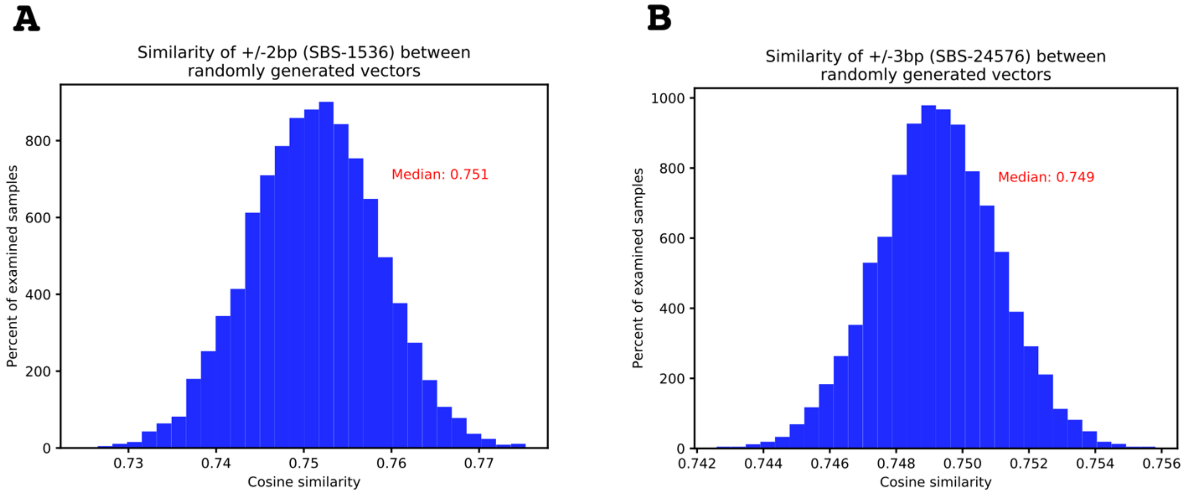


Figure 3.6: Evaluating the average similarity of random nonnegative vectors. A) Comparing the cosine similarities amongst 10,000 randomly generated nonnegative vectors, where each vector has 1536 mutational channels. B) Comparing the cosine similarities amongst 10,000 randomly generated nonnegative vectors, where each vector has 24,576 mutational channels.

Third, we evaluated the false-positive rates of tools commonly used for discovery of cancer driver genes. More specifically, we simulated the somatic mutations observed in the 1024 whole-exome sequenced breast cancers reported in the TCGA MC3 release [145]. The simulations were repeated 100 times and each of these 100 repetitions was analyzed for driver genes using MutSigCV1.41 and MutSigCV2 [53] as well as dNdScv [142]. In principle, since SigProfilerSimulator randomly shuffles somatic mutations, one would not expect to find any genes under selection. However, each of the tools found significantly mutated genes within the simulations using the recommended cutoff threshold of q-value < 0.10 (Figure 3.4e). On average MutSig1.41CV found between 1.3 and 1.6 false-positive driver genes per simulation when examining data generated using the SBS-384 and SBS-6144 mutational classifications, respectively. In contrast, MutSig2CV found between 0.3 and 0.2 false-positive driver genes per simulation using SBS-384 and SBS-6144, respectively. Lastly, dNdScv found between 0.03 and 0.02 false-positive driver genes per simulation using SBS-384 and SBS-6144, respectively. Note that by chance, when using a q-value cutoff of 0.1, one would expect to observe less than 0.1 false-positive driver genes per simulation. Lowering the threshold for statistical significance to 0.01 eliminates all false-positive results from dNdScv and MutSig2CV but not for MutSig1.41CV.

3.4. Conclusions

Increasingly, there is a need to develop reliable background models of cancer mutational landscapes to allow downstream statistical analysis for biological discoveries. Currently, to the best of our knowledge, there is no tool that allows explicitly simulating accurate background mutational landscapes. This report presents SigProfilerSimulator, a method that allows fast generation of mutational landscapes at different resolutions. As demonstrated by our analyses, SigProfilerSimulator can be used to evaluate the accuracy of other bioinformatics tools or it can be leveraged for making novel discoveries. SigProfilerSimulator's breadth of features allows one to construct a tailored null hypothesis of mutational landscapes and to identify significance levels of the subsequent results. Overall, SigProfilerSimulator will be a useful tool for any researcher that performs statistical analysis based on mutational data derived from the sequencing of cancer or normal somatic tissues.

3.5. Methods

Tool implementation: SigProfilerSimulator is developed as a computationally efficient Python package and it is available for installation through PyPI. Further, an R-wrapper is available through GitHub. The tool leverages a PCG random number generator that provides a simple, fast, and space-efficient algorithm for generating random numbers with high statistical quality [146]. The tool uses a Monte Carlo approach for randomly generating somatic mutations while considering the observed frequency of a preselected reference genome. More specifically, SigProfilerSimulator randomly shuffles mutations by using the precomputed observed rates of mutational channels in a reference genome. The tool works in unison with SigProfilerMatrixGenerator [143] to first classify a catalog of somatic mutations prior to simulating it. The final mutational catalog is outputted into commonly used mutation data formats including mutation annotation format (MAF) files and variant annotation format (VCF) files. SigProfilerSimulator is freely available and has been extensively documented.

Python code: https://github.com/AlexandrovLab/SigProfilerSimulator

R wrapper: https://github.com/AlexandrovLab/SigProfilerSimulatorR

Documentation: https://osf.io/usxjz/wiki/home/

3.6. Computational benchmarking

The computational efficiency of SigProfilerSimulator was benchmarked by simulating the freely available PCAWG dataset, consisting of 2,144 samples with 36,876,213 single base substitutions, for a single iteration using the default parameters. Simulating the complete dataset took approximately 90 s. Simulations were performed on a dedicated computational node with a dual Intel® Xeon® Gold 6132 Processors (19.25 M Cache, 2.60 GHz) and 192 GB of shared DDR4-2666 RAM.

3.6.1. Analysis of doublet base substitutions

We simulated the PCAWG dataset using the SBS-96 classification. Each simulation was performed 1,000 times considering mutations as both statistically independent events (non-updating; each mutation has no effect on the observed rate of mutational channels) and dependent events (updating; each mutation updates the observed rate of mutational channels). To calculate the number of DBS mutations occurring by chance in each sample, we generated the mutational catalogs for DBS-78 using SigProfilerMatrixGenerator [143]. The resulting counts for DBSs were used to plot the distributions of the expected number of DBSs due to two adjacent SBSs happening purely by chance. The fold change was calculated by dividing the mean DBS count observed across the simulations by the total number of DBSs found in the original sample. Derivation of q-values was performed by applying the Benjamini and Hochberg false discovery rate correction to p-values calculated using z-tests based on the DBS distributions found in the simulations and the numbers of DBSs observed in the real data.

3.6.2. Sequence context analysis for mutational signatures

The PCAWG dataset was simulated using the SBS-6, SBS-96, and SBS-1536 classifications while ensuring the respective mutational patterns and mutational burdens on each chromosome match the ones observed in the real data. SigProfilerMatrixGenerator was used to derive the mutational vectors for each sample including vectors incorporating three bases 5′ and three bases 3′ of each mutation, resulting in a classification with 24,576 mutational channels. To avoid comparisons of sparse binary vectors, only samples that had at least 2 mutations per mutational channel were included in the comparative analyses. The simulated and real mutational patterns of a cancer genome were considered the same if their cosine similarity was at least 0.85. Note that the average cosine similarity between two random nonnegative vectors is 0.75 (Figure 3.6). The chance of two nonnegative vectors with 1,536 mutational channels or 24,576 mutational channels to have a similarity of 0.85 simply by chance is less than 10–6 (Figure 3.6).

3.6.3. Benchmarking false-positive driver genes detected by MutSigCV and dNdScv

All whole-exome sequenced breast cancer samples part of the TCGA MC3 release were simulated using SBS-384 and SBS-6144 contexts while maintaining the mutational burden on each chromosome. As recommended [142], 23 samples with more than 500 exonic mutations were excluded from the analysis. Each simulation was repeated 100 times with different random seeds. The variant annotation predictor [147] was used to annotate simulated mutations with the appropriate gene name for compatibility with MutSigCV1.41 and MutSigCV2 [53]. We ran MutSigCV1.41 and MutSigCV2 using the recommended default parameters in conjunction with the genome reference sequence for hg19, mutation dictionary file, exome coverage file, and gene covariates file as found at https://software.broadinstitute.org/cancer/cga/mutsig_run. We ran dNdScv [142] using the default library parameters and filtered out the significant genes using the recommended q-value cutoff of less than 0.10. All rainfall plots were generated using karyoploteR [144].

3.7. Availability and requirements

Project name: SigProfilerSimulator

Project home page: https://github.com/AlexandrovLab/SigProfilerSimulator

Operating system(s): Unix, Linux, and Windows

Programming language: Python 3; R wrapper

Other requirements: None

License: BSD 2-Clause "Simplified" License

Any restrictions to use by non-academics: None

3.8. Availability of data and materials

No novel data were generated as part of this study. All source code is freely available and can be downloaded from the links below. Python code: https://github.com/AlexandrovLab/SigProfilerSimulator. R wrapper:

https://github.com/AlexandrovLab/SigProfilerSimulatorR. Documentation: https://osf.io/usxjz/wiki/home/

3.9. Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in BMC Bioinformatics 2020. Bergstrom, Erik N.; Barnes, Mark; Martincorena, Inigo; Alexandrov, Ludmil B., Springer Nature, 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 4.

Examining clustered somatic mutations with SigProfilerClusters

Abstract

Motivation: Clustered mutations are found in the human germline as well as in the genomes of cancer and normal somatic cells. Clustered events can be imprinted by a multitude of mutational processes, and they have been implicated in both cancer evolution and development disorders. Existing tools for identifying clustered mutations have been optimized for a particular subtype of clustered event and, in most cases, relied on a predefined intermutational distance (IMD) cutoff combined with a piecewise linear regression analysis.

Results: Here, we present SigProfilerClusters, an automated tool for detecting all types of clustered mutations by calculating a sample-dependent IMD threshold using a simulated background model that takes into account extended sequence context, transcriptional strand asymmetries and regional mutation densities. SigProfilerClusters disentangles all types of clustered events from non-clustered mutations and annotates each clustered event into an established subclass, including the widely used classes of doublet-base substitutions, multi-base substitutions, omikli and kataegis. SigProfilerClusters outputs non-clustered mutations and clustered events using standard data formats as well as provides multiple visualizations for exploring the distributions and patterns of clustered mutations across the genome.

Availability and implementation: SigProfilerClusters is supported across most operating systems and made freely available at https://github.com/AlexandrovLab/SigProfilerClusters with an extensive documentation located at https://osf.io/qpmzw/wiki/home/.

4.1. Introduction

Mutations are found on the genomes of all cells in the human body [1, 148]. Most single-base substitutions and small insertions and deletions (indels) accumulate independently across the genome, but a subset of the mutations cluster in a non-random manner [52, 53]. Previous studies have revealed that clustered mutations are imprinted by a plethora of endogenous and exogenous mutational processes [5, 14, 18, 24, 52, 54, 56-61, 65, 89, 149-151]. Some clustered mutations have been implicated in cancer evolution [52, 57, 58, 60, 150, 152], while de novo clustered mutations have been identified in the human germline and shown to contribute to developmental disorders [153, 154]. In recent years, sets of simultaneously occurring clustered substitutions have been further subclassified into independent events [57, 152], including (*i*) doublet-base substitutions (DBSs); (*ii*) multi-base substitutions (MBSs); (*iii*) diffuse hypermutation termed omikli; (*iv*) longer strand-coordinated events termed kataegis and (*v*) recurrent hypermutation of extra-chromosomal DNA (ecDNA) termed kyklonas.

Traditional methods separate clustered mutations based on a predefined inter-mutational distance (IMD) threshold typically between 1 and 2 kilobases [4, 18, 24, 58, 59, 85, 155]. Many of these approaches utilize a piecewise linear regression to segment each chromosome, which, in most cases, is optimized for calling larger strand-coordinated kataegic events (Figure 4.1) [4, 156, 157]. Most existing methods have also ignored confounding effects attributed to localized differences in mutation rates, copy number alterations or the mutational burden across each chromosome within a given sample leading to an accumulation of false-positive clustered events (Figure 4.1). Further, the majority of existing tools focus on detecting only a specific class of clustered events including doublet-base substitutions and multi-nucleotide variants [54, 150, 151], kataegis [58, 155, 156] or APOBEC3-associated events [5, 59] while ignoring the larger landscape of clustered mutations. For example, a recent study [57] developed an algorithm focused on the detection of APOBEC3-associated omikli and kataegis events in cancer genomes by incorporating simulations of somatic mutations and estimates of cancer cell fractions.

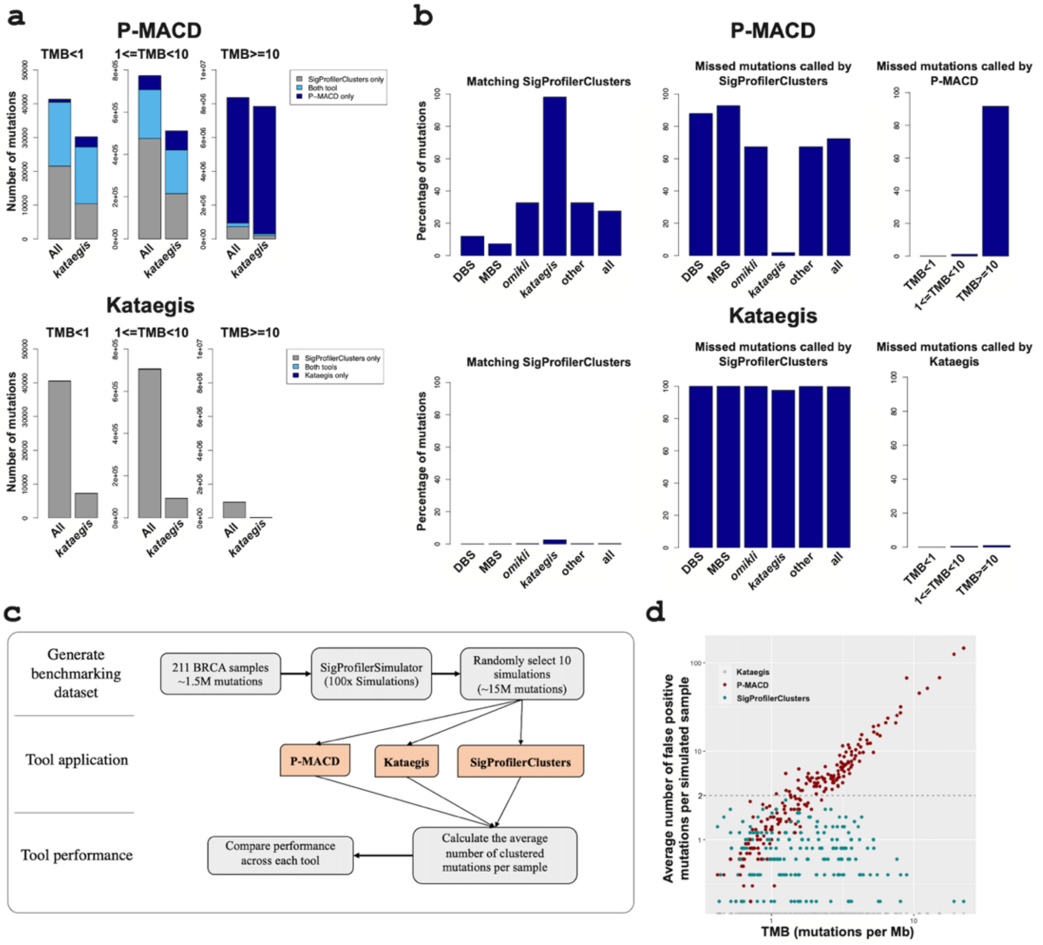


Figure 4.1: Benchmarking of existing tools for detecting clustered mutations. a) Assessing the degree of overlap between two tools that detect clustered mutations and SigProfilerCusters. P-MACD locates regions of clustered mutations by implementing a negative binomial distribution to model the probability of observing a given cluster within a given window of the genome. When calculating p-values, this model assumes that all mutations occur randomly across the genome limiting clustered events to those that occur at most 10 kilobases (kb) from one another. Kataegis implement a piece-wise linear model with set thresholds requiring all clustered events to be composed of 6 or more mutations with an average inter-mutational distance of 1 kb between adjacent mutations. All tools were applied to 2,703 whole-genome sequenced samples from PCAWG and were split into low (TMB<1), intermediate (1<=TMB<10), and high (TMB>=10) TMB. b) The percentage of mutations overlapping each subclass of clustered events called by SigProfilerClusters (left), the percentage of mutations that were missed by each of the two tools that were called by SigProfilerClusters (middle), and the percentage of total mutations called by each tool that were missed by SigProfilerClusters separated by low, intermediate, and high TMB (right). c) Schematic workflow to assess the false positive rate of each tool using simulations of 211 breast cancer genomes from PCAWG. d) The average number of false positive mutations per simulated sample detected across the simulated breast cancer dataset from PCAWG. Kataegis did not detect any kataegic clustered events within the simulated data using the tools' default parameters. Similarly, SigProfilerClusters did not detect any kataegic clustered events within these simulated data.

Separation and classification of clustered events are required to fully elucidate the mutational processes operating in cancer and normal somatic cells [52, 152]. Here, we present SigProfilerClusters, a tool to comprehensively characterize and subclassify clustered mutations from the complete catalog of mutations within the genome of a single sample (Figure 4.2a). SigProfilerClusters classifies all types of clustered mutations, including (*i*) doublet-base substitutions; (*ii*) multi-base substitutions; (*iii*) omikli; (*iv*) kataegis and (*v*) clustered small insertions and deletions (indels). The tool calculates a sample-dependent IMD threshold that considers regional differences in mutation rates, variant allele fractions and cancer cell fractions of adjacent mutations to reduce the false positive rate and provides visualizations for downstream analyses (Figure 4.2b and c; Figure 4.1). Further, SigProfilerClusters integrates within the larger suite of SigProfiler tools [143, 158, 159] to facilitate downstream mutational signature analysis of both non-clustered and clustered single-base substitutions and indels, thus, allowing the accurate detection of mutational processes giving rise to even low levels of clustered events (Figure 4.2d) [143, 158, 159].

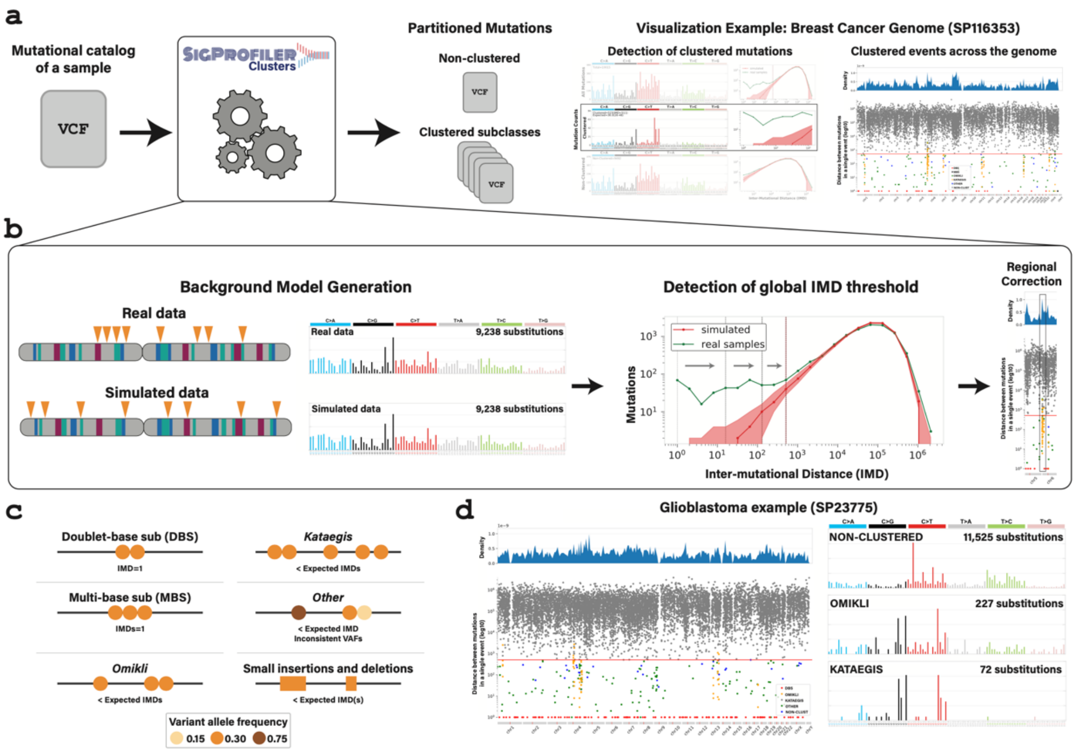


Figure 4.2: Detection and characterization of clustered mutations with SigProfilerClusters. (a) An example workflow used to detect clustered mutations in a single cancer genome. As an input, SigProfilerClusters accepts common formats for mutations, such as ones in the variant calling format (VCF), and the tool separates all clustered mutations from the complete mutational catalog of the provided sample. Final partitions of mutations in the sample are outputted as VCF files and visualized using the mutational spectra of all mutations, only clustered mutations and only non-clustered mutations along with a rainfall plot commonly used to show the distribution of inter-mutational distances across a cancer genome [4, 5, 152]. (b) Schematic demonstrating the process of calculating a sample-dependent IMD threshold to separate clustered from non-clustered mutations across each genome. A binary search algorithm is used to efficiently detect the optimal global IMD threshold for each sample. Detection of the global IMD threshold is illustrated using gray arrows. Regional corrections are performed to identify local IMD thresholds based on variance of mutation rates across the genome. (c) Every clustered mutation is classified into a single subcategory of clustered event. (d) Rainfall plot illustrating the distribution of IMDs across a single glioblastoma sample (left). The mutational spectra for omikli and kataegic events reveal a different mutational pattern compared to the pattern of all non-clustered somatic mutations (right).

4.2. Material and methods

SigProfilerClusters derives an IMD cutoff that is unlikely to occur purely by chance given the observed mutational burden and the mutational patterns within the genome of a given sample. To calculate the genome-dependent IMD, the tool leverages SigProfilerSimulator [158] to generate background models by randomizing the distribution of mutations across the genome. By default, the genome of each sample is simulated 100 times in order to derive 95% confidence intervals for the expected genomic mutational landscape, with every simulation maintaining the penta-nucleotide sequence context for each substitution, the ratio of all mutations in genic and inter-genic regions, the transcriptional strand asymmetries of all mutations in genic regions and the mutational burden on each chromosome [143, 158]. Importantly, this randomization procedure is highly customizable [158] and can be altered based on the needs of a given study design, thus, allowing the incorporation of other factors that affect the accumulation of mutations such as nucleosome occupancy, presence of histone modifications and many others. A binary search algorithm is implemented to efficiently derive the global IMD threshold for each genome. The final global IMD threshold is selected by ensuring that 90% of mutations below the chosen cutoff are unlikely to appear by chance given the simulated distribution of mutations (q-value < 0.01; Figure 4.1) with a maximum global IMD cutoff of 10 kilobases. The algorithm also considers regional heterogeneities of mutation rates, generally associated with replication timing [75] or differential gene expression [53, 61, 160-162], by correcting for variance in clonality as well as variance in both mutation-dense and mutation-poor regions using a sliding genomic window (default size of 1 megabase). Specifically, an additional regional IMD cutoff is corrected within each genomic window based on the fold difference between the number of real and the number of simulated mutations, while maintaining the original criteria of <10% of mutations below the IMD cutoff appearing by chance (q-value < 0.01). Lastly, when data are available, SigProfilerClusters ensures that adjacent mutations are in the same cells by introducing a maximum difference in variant allele frequencies (VAF) or cancer cell fraction (CCF), which incorporates copy number changes, below a certain threshold (default cutoff value of 0.10 and 0.25; respectively).

After identifying the set of clustered mutations, SigProfilerClusters subclassifies each clustered substitution into a single category of previously established clustered events [57, 152]. Briefly, all clustered substitutions with consistent VAFs or consistent CCFs are classified into one of four categories. Two mutations with an IMD of 1 are classified as doublet-base substitutions, while clusters of three or more adjacent mutations each with an IMD of 1 are classified as multi-base substitutions. Clusters of two or three mutations with IMDs less than the sample-dependent cutoff and with at least a single IMD greater than 1 are classified as omikli [152], while clusters of four or more mutations with IMDs less than the sample-dependent cutoff and with at least a single IMD greater than 1 are classified as kataegis [152]. All remaining clustered mutations with inconsistent VAFs or CCFs are classified as other. Clustered indels are not subclassified into different categories due to a lack of previously defined subtypes.

4.3. Usage

SigProfilerClusters is freely available as a Python package, distributed under the permissive BSD-2 clause license and can be used on most operating systems including Windows, MacOS and Linux-based machines. The tool is compatible with large-scale deployments on high-performance computing clusters as well as on cloud infrastructures such as Amazon Web Services. Input data can be provided in the form of common mutation formats including the Variant Call Format (VCF), the Mutation Annotation Format or in the form of a simple text file. The output of SigProfilerClusters results in the partitioning of all mutations into a clustered or non-clustered directory. All clustered mutations are then classified into distinct subcategories of events and provided individually in VCF files for downstream visualization and analyses. The output for each subclass of the clustered event can be directly utilized by additional SigProfiler tools including SigProfilerExtractor for mutational signature analysis [159] and SigProfilerPlotting for examining patterns of somatic mutations [143]. The results for each sample are also summarized using two individual visualizations that include: (*i*) a rainfall plot depicting the minimum global IMD between all adjacent mutations, where each individual set of adjacent mutations is colored based on its clustered classification; and (*ii*) a multi-panel figure that displays the mutational patterns across all mutations, clustered mutations and non-clustered mutations, separately along with the distribution of IMDs across the real and simulated data for each sample (Figure 4.2a).

4.4. Conclusion

Elucidating the compendium of clustered somatic mutations in the genome of a sample allows further understanding of the mutational process that give rises to these events and can provide novel insights into disease etiology [52, 57, 152]. Previous studies have traditionally interrogated the complete mutational catalogs of cancer genomes, which can lead to the inability to detect processes active at low levels or those which have been transiently activated. Our prior analysis of clustered mutations [152] has revealed an enrichment of clustered mutations within known cancer driver events, hypermutation of extra-chromosomal DNA fueling the evolution of cancers, and ultimately, resulting in a differential patient outcome. Here, we provide SigProfilerClusters, an automated and freely available Python-based tool that comprehensively identifies and classifies clustered mutations enabling users to interrogate the mutational processes giving rise to such events.

4.5. Data availability

No data were generated for this publication.

4.6. Acknowledgements

Chapter 4, in full, is a reprint of the material as it appears in Bioinformatics 2022. Bergstrom, Erik N.; Kundu, Mousumy; Tbeileh, Noura; Alexandrov, Ludmil B., Oxford University Press, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 5.

Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA

Abstract

Clustered somatic mutations are common in cancer genomes and previous analyses reveal several types of clustered single-base substitutions, which include doublet- and multi-base substitutions [5, 18, 54-56], diffuse hypermutation termed omikli [57], and longer strand-coordinated events termed kataegis [5, 58, 59, 149]. Here we provide a comprehensive characterization of clustered substitutions and clustered small insertions and deletions (indels) across 2,583 whole-genome-sequenced cancers from 30 types of cancer [60]. Clustered mutations were highly enriched in driver genes and associated with differential gene expression and changes in overall survival. Several distinct mutational processes gave rise to clustered indels, including signatures that were enriched in tobacco smokers and homologous-recombination-deficient cancers. Doublet-base substitutions were caused by at least 12 mutational processes, whereas most multi-base substitutions were generated by either tobacco smoking or exposure to ultraviolet light. Omikli events, which have previously been attributed to APOBEC3 activity [57], accounted for a large proportion of clustered substitutions; however, only 16.2% of omikli matched APOBEC3 patterns. Kataegis was generated by multiple mutational processes, and 76.1% of all kataegic events exhibited mutational patterns that are associated with the activation-induced deaminase (AID) and APOBEC3 family of deaminases. Co-occurrence of APOBEC3 kataegis and extrachromosomal DNA (ecDNA), termed kyklonas (Greek for cyclone), was found in 31% of samples with ecDNA. Multiple distinct kyklonic events were observed on most mutated ecDNA. ecDNA containing known cancer genes exhibited both positive selection and kyklonic hypermutation. Our results reveal the diversity of clustered mutational processes in human cancer and the role of APOBEC3 in recurrently mutating and fuelling the evolution of ecDNA.

5.1. Main

Cancer genomes contain somatic mutations that are imprinted by different mutational processes [4, 18]. Most single-base substitutions and small indels are independently scattered across the genome; however, a subset of substitutions and indels tend to cluster [52, 53]. This clustering has been attributed to a combination of heterogeneous mutation rates across the genome, biophysical characteristics of exogenous carcinogens, dysregulation of endogenous processes and larger mutational events associated with genome instability—amongst others [5, 14, 52, 54, 57-62, 64, 65, 142]. Previous analyses of clustered mutations have focused on single-base substitutions and revealed several classes of clustered events, including doublet- and multi-base substitutions [5, 18, 54-56] (DBSs and MBSs, respectively), diffuse hypermutation (omikli) [57] and longer events (kataegis) [5, 58, 59, 149]. Most kataegic events were found to be strand-coordinated, defined as sharing the same strand and reference allele [4, 5]. Previous studies have also revealed nine clustered signatures [52] and clustered driver substitutions due to APOBEC3-associated mutagenesis [57] or carcinogenic-triggered *POLH* mutagenesis [52].

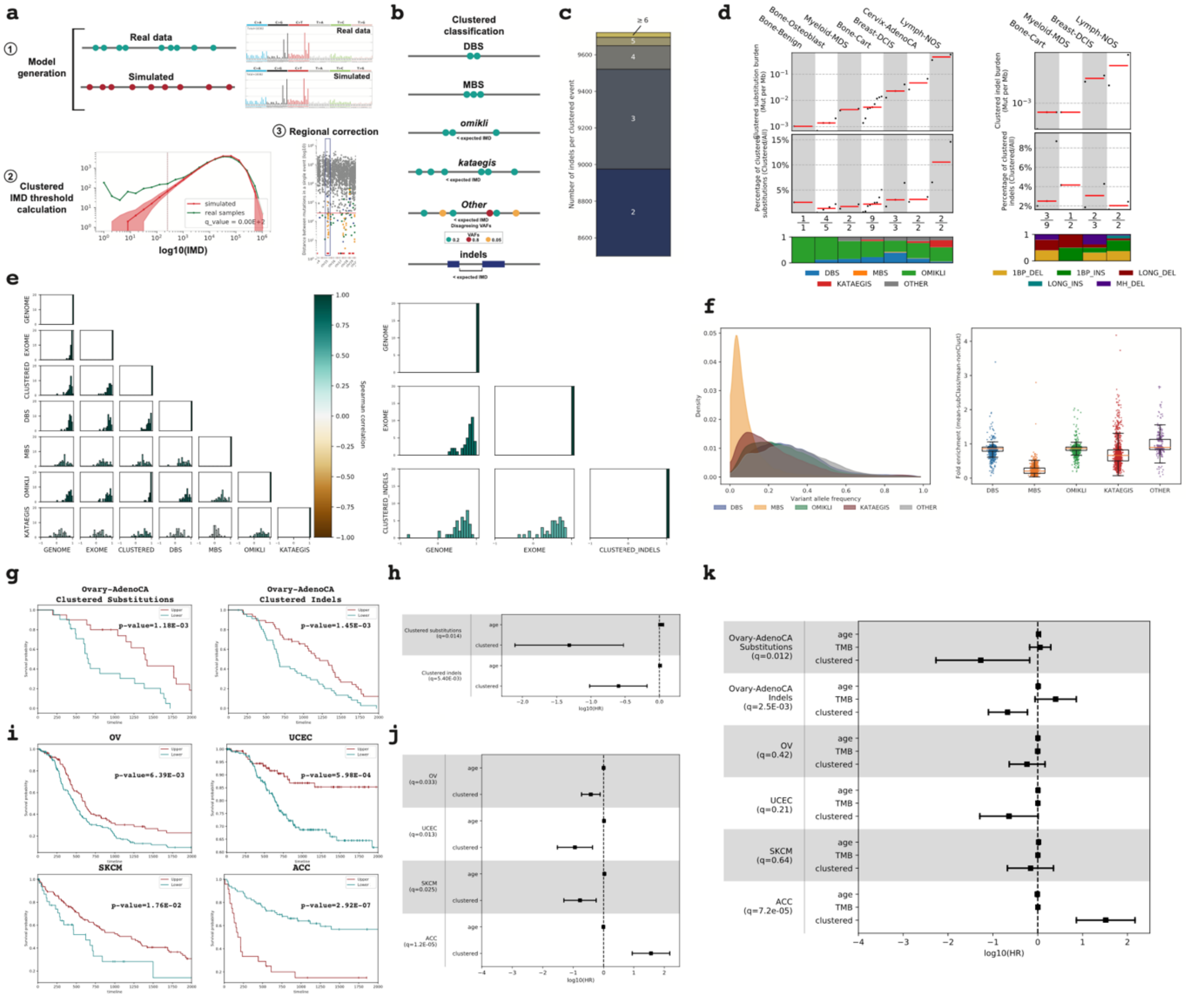
DBSs have been extensively examined, revealing multiple endogenous and exogenous processes that can cause these events, including failure of DNA repair pathways and exposure to environmental mutagens [4, 5, 18]. By contrast, MBSs have not been comprehensively investigated, presumably owing to their small numbers in cancer genomes. Moreover, only a handful of processes have been associated with omikli and kataegic events, with most processes attributed to the AID and APOBEC3 family of deaminases [5, 52, 57-59, 61, 87-89, 98]. Specifically, the APOBEC3 enzymes, which are typically responsible for antiviral responses [90-96], give rise to omikli and kataegis by requiring single-stranded DNA as a substrate [57, 58, 85, 88]. Omikli were found to be enriched in early replicating regions and more prevalent in microsatellite stable tumours, indicating that mismatch repair has a role in exposing short single-stranded DNA regions [57]. The differential activity of mismatch repair towards gene-rich regions results in increased omikli events within cancer genes [57]. Kataegis is less prevalent than omikli as it is likely to depend on longer tracks of single-stranded DNA [58, 59, 65]. Such tracks are typically available during the repair of double-strand breaks and most kataegis has been observed within 10 kb of detected breakpoints [60].

Amplification of known cancer genes is known to drive tumorigenesis in many types of cancer [163]. Studies have shown high copy-number states of circular ecDNAs, which often contain known cancer genes and are found in most cancers [163-166]. The circular nature of ecDNAs and their rapid replication mimic double-stranded DNA viral pathogens, which indicates that they could be substrates for APOBEC3 mutagenesis; this may contribute to the evolution of tumours that contain ecDNA through accelerated diversification of extrachromosomal oncoproteins.

5.1.1. The landscape of clustered mutations

To identify clustered mutations, a sample-dependent intra-mutational distance (IMD) cut-off was derived in which mutations below the cut-off were unlikely to occur by chance (q-value < 0.01). A statistical approach using the IMD cut-off, variant allele frequencies (VAFs) and corrections for local sequence context was applied to each specimen (Methods, Figure 5.1a). Clustered mutations with consistent VAFs were subclassified into four categories (Figure 5.1b). DBSs and MBSs were characterized as two adjacent mutations (DBSs) and as three or more adjacent mutations (MBSs) (IMD = 1). Multiple substitutions each with IMD > 1 bp and below the sample-dependent cut-off were characterized as either omikli (two to three substitutions) or kataegis (four or more substitutions) (Figure 5.2). Clustered substitutions with inconsistent VAFs were classified as 'other'. Although clustered indels were not subclassified into different categories, most events resembled diffuse hypermutation, with 92.3% of events having only two indels (Figure 5.1c).

Figure 5.1: Identification and clinical associations of clustered events. **a**. Schematic depiction for separating clustered mutations for a sample. **b.** Subclassification of clustered substitutions and indels. Expected IMD derived using steps 2 and 3 (**a**). **c.** Distribution of indels present in a single clustered event. **d.** Distribution of clustered substitutions (left) and indels (right) across cancers with less than 10 samples subclassified into different categories. **e.** Correlations between TMB of each sample, the TMB within the exome, or the TMB for each class of clustered substitutions (left) and indels (right). **f.** Distribution of VAFs for all clustered substitution classes (left; DBS: 1,215 samples; MBS: 851; omikli:1,466; kataegis: 1,108; other: 335) with the average fold enrichment compared against non-clustered mutations (right). For each boxplot, the middle line reflects the median, the lower and upper bounds correspond to the first and third quartiles, and the lower and upper whiskers extend from the box by 1.5x the inter-quartile range (IQR). **g.** Kaplan–Meier curves between samples with high (top 80th percentile) and low (bottom 20th percentile) clustered substitution (left) or indel (right) burdens in PCAWG ovarian cancer. **h.** Cox regressions performed for PCAWG cancer types while correcting for age (n = 20 upper and n = 21 lower clustered substitutions; n = 49 upper and n = 49 lower clustered indels). **i.** Kaplan–Meier survival curves for TCGA cancer types with a differential patient outcome associated with the detection of any clustered mutations. **j. k.** Cox regressions performed for TCGA samples while correcting for age (**j**) and total mutational burden (**k**) (OV: n = 111 upper, n = 159 lower clustered substitutions; UCEC: n = 322 upper, n = 64 lower; ACC: n = 24 upper, n = 67 lower). PCAWG ovarian cancers were included in **k**. Centre of measure for each Cox regression reflects the log10(Hazards ratios) with the 95% confidence intervals in **h**–**k**).


a
① Model generation
Real data
Simulated
② Clustered IMD threshold calculation
simulated
real samples
q_value = 0.00E+2
log10(IMD)
③ Regional correction
b
Clustered classification
DBS
MBS
omikli
< expected IMD
kataegis
< expected IMD
Other
< expected IMD
Disagreeing VAFs
VAFs
0.2
0.8
0.05
indels
< expected IMD
c
≥ 6
5
4
3
2
Number of indels per clustered event
d
Bone-Benign
Bone-Osteoblast
Myeloid-MDS
Bone-Cart
Breast-DCIS
Cervix-AdenoCA
Lymph-NOS
Clustered substitution burden (Mut per Mb)
Percentage of clustered substitutions (Clustered/All)
DBS
MBS
OMIKLI
KATAEGIS
OTHER
Bone-Cart
Myeloid-MDS
Breast-DCIS
Lymph-NOS
Clustered indel burden (Mut per Mb)
Percentage of clustered indels (Clustered/All)
1BP_DEL
1BP_INS
LONG_DEL
LONG_INS
MH_DEL
e
GENOME
EXOME
CLUSTERED
DBS
MBS
OMIKLI
KATAEGIS
Spearman correlation
GENOME
EXOME
CLUSTERED_INDELS
f
Density
Variant allele frequency
DBS
MBS
OMIKLI
KATAEGIS
OTHER
Fold enrichment (mean-subClass/mean-nonClust)
DBS
MBS
OMIKLI
KATAEGIS
OTHER
g
Ovary-AdenoCA Clustered Substitutions
p-value=1.18E-03
Ovary-AdenoCA Clustered Indels
p-value=1.45E-03
Upper
Lower
h
Clustered substitutions (q=0.014)
Clustered indels (q=5.40E-03)
age
clustered
log10(HR)
i
OV
p-value=6.39E-03
UCEC
p-value=5.98E-04
SKCM
p-value=1.76E-02
ACC
p-value=2.92E-07
Upper
Lower
Survival probability
timeline
j
OV (q=0.033)
UCEC (q=0.013)
SKCM (q=0.025)
ACC (q=1.2E-05)
age
clustered
log10(HR)
k
Ovary-AdenoCA Substitutions (q=0.012)
Ovary-AdenoCA Indels (q=2.5E-03)
OV (q=0.42)
UCEC (q=0.21)
SKCM (q=0.64)
ACC (q=7.2e-05)
age
TMB
clustered
log10(HR)

Examining 2,583 whole-genome-sequenced cancers from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project revealed a total of 1,686,013 clustered single-base substitutions and 21,368 clustered indels (Figure 5.3, Figure 5.1d). DBSs, MBSs, omikli and kataegis comprised 45.7%, 0.7%, 37.2% and 7.0% of clustered substitutions across all samples, respectively, and their distributions varied greatly within and across cancer types. For example, melanoma had the highest clustered substitution burden, with ultraviolet light associated doublets (CC>TT) accounting for 74.2% of clustered mutations; however, these contributed only 5.3% of all substitutions in melanoma (Figure 5.3a). By contrast, 11.5% of all substitutions in bone leiomyosarcomas were clustered, and omikli and kataegis constituted 43.8% and 46.7% of these mutations, respectively (Figure 5.3a). Clustered indels exhibited similarly diverse patterns within and across cancer types (Figure 5.3b). For example, the highest mutational burden of clustered indels was observed in lung and ovarian cancers. Clustered indels in lung cancer accounted for only 2.6% of all indels and were characterized by 1-bp deletions. By contrast, clustered long indels at microhomologies were commonly found in ovarian and breast cancers and contributed more than 10% of all indels in a subset of samples (Figure 5.3b). Correlations between the total number of mutations and the number of clustered mutations were observed for DBSs and omikli but not for MBSs, kataegis or indels (Figure 5.1e). In most cancers, DBSs and omikli had VAFs consistent with those of non-clustered mutations, whereas MBSs and kataegis tended to have lower VAFs (Figure 5.1f). Kataegic events contained 4 to 44 mutations and 81% of events were strand-coordinated, indicative of damage or enzymatic changes on a single DNA strand.

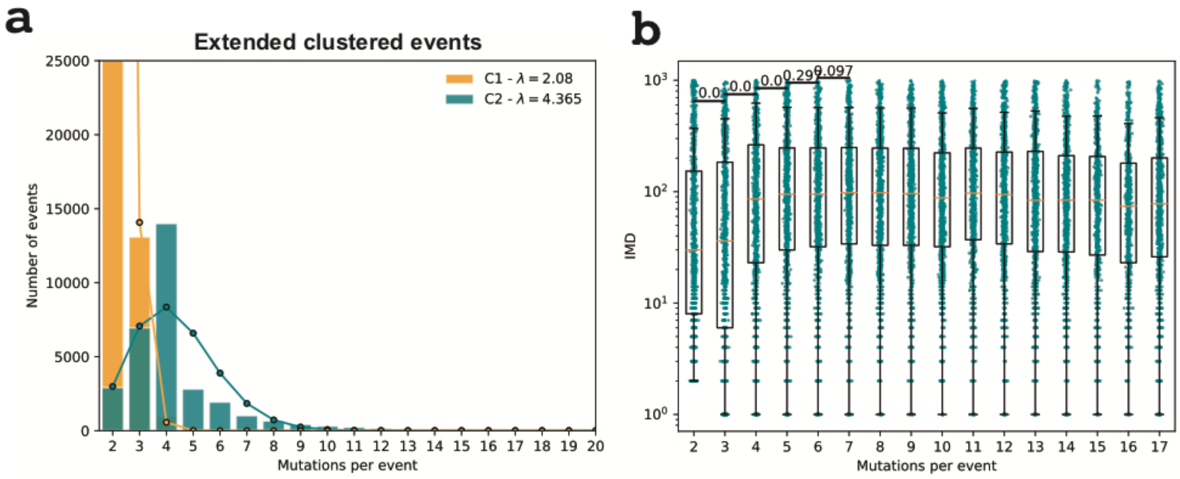


Figure 5.2: Determining the number of mutations differentiating between omikli and kataegis. **a)** Modeling the number of mutations per event using a mixture of two Poisson distributions. The first component, representative of omikli, has an average IMD of 2.1, while the second component, representative of kataegis, has an average IMD of 4.4. The estimated contribution of mutations of each component are depicted as bars for each corresponding event size **b)** The distribution of IMDs per event across different sized events (n=199,912 events with 2 mutations; n=35,576 events with 3 mutations; n=15,320 events with 4 mutations; n=9,613 events with 5 mutations). The chosen cutoff between omikli and kataegis was four mutations. For each boxplot, the middle line reflects the median, the lower and upper bounds of the box correspond to the first and third quartiles, and the lower and upper whiskers extend from the box by 1.5x the inter-quartile range (IQR).

The overall survival was compared between patients with cancers containing high and low numbers of clustered mutations within whole-genome-sequenced PCAWG and whole-exome sequenced The Cancer Genome Atlas (TCGA) cancer types [167]. Better overall survival was observed only in whole-genome-sequenced ovarian cancers that contained high-levels of clustered substitutions or clustered indels (q-values < 0.05) (Figure 5.1g,h). Conversely, whole-exome-sequenced adrenocortical carcinomas containing clustered substitutions were associated with a worse overall survival (q-value = 7.2 × 10−5) (Figure 5.1i-k).

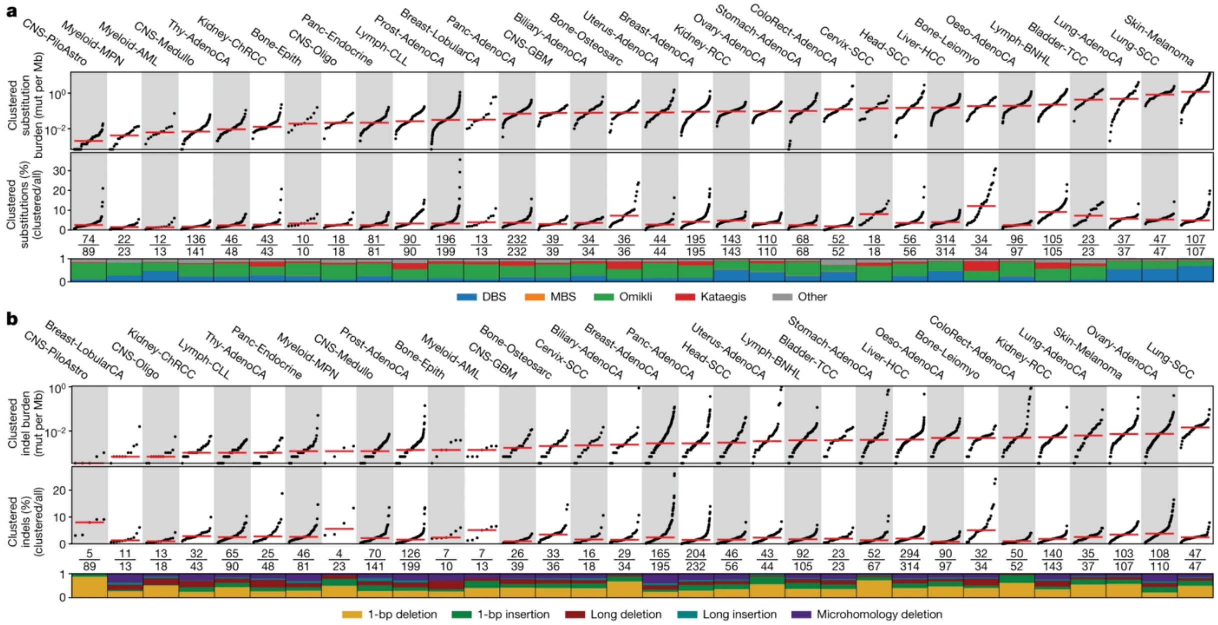


Figure 5.3: The landscape of clustered mutations across human cancer. Pan-cancer distribution of clustered substitutions subclassified into DBSs, MBSs, omikli, kataegis and other clustered mutations (**a**). Top, each black dot represents a single cancer genome. Red bars reflect the median clustered TMB (mutations (mut) per Mb) for cancer types. Middle, the clustered TMB normalized to the genome-wide TMB reflecting the contribution of clustered mutations to the overall TMB of a given sample. Red bars reflect the median contribution for cancer types. Bottom, the proportion of each subclass of clustered events for a given cancer type with the total number of samples having at least a single clustered event over the total number of samples within a given cancer cohort. **b.** Pan-cancer distribution of clustered small indels. The top and middle panels have the same information as **a**. Bottom, the proportion of each cluster type of indel for a given cancer type with the total number of samples having at least a single clustered indel over the total number of samples within a given cancer cohort. All 2,583 whole-genome-sequenced samples from PCAWG are included in the analysis; however, cancers with fewer than 10 samples were removed from the main figure and included in Figure 5.1d. For definitions of abbreviations for cancer types used in the figures, see 'Cancer-type abbreviations' in Methods.

5.1.1.1. Determining the number of clustered mutations in omikli and kataegic events

To determine the cutoff of the number of mutations in an omikli versus a kataegic event, we modelled the distribution of clustered event sizes (excluding DBSs, MBSs, and other clustered events with disagreeable variant allele frequencies) using a mixture of two Poisson distributions (Figure 5.2a). The modelling also excluded clustered mutations from skin melanomas, that contribute a disproportionate number of DBS events, and clustered mutations from lymphomas, that contribute a large proportion of canonical and non-canonical AID kataegis. The first component, corresponding with omikli events (gold), had an average of 2.1 mutations per event, while the second component, corresponding to larger kataegic events (teal), had an average of 4.4 mutations per events. Using the posterior probabilities of each distribution, we calculated the likelihood of a given clustered event belonging to a specific component. Events comprised of four or more mutations were attributed to the kataegic component with >95% probability. Further, we assessed the IMD distributions of different sized events revealing approximately a 2-fold increase in average IMD between events possessing 3 and 4 mutations supporting the activity of two separate mutational processes (Figure 5.2b).

5.1.2. Signatures of clustered mutations

Mutational signature analysis was performed for each category of clustered events, which enabled the identification of 12 DBS, 5 MBS, 17 omikli, 9 kataegic and 6 clustered indel signatures (Figure 5.4). Although DBS signatures have previously been described [18], previous analysis combined DBSs and MBSs into a single class [18]. Separating these events into individual classes showed that a multitude of processes can give rise to DBSs, whereas most MBSs are attributable to signatures associated with tobacco smoking (SBS4) or ultraviolet light (SBS7). Additional DBS and MBS signatures were found within a small subset of cancer types (Figure 5.5).

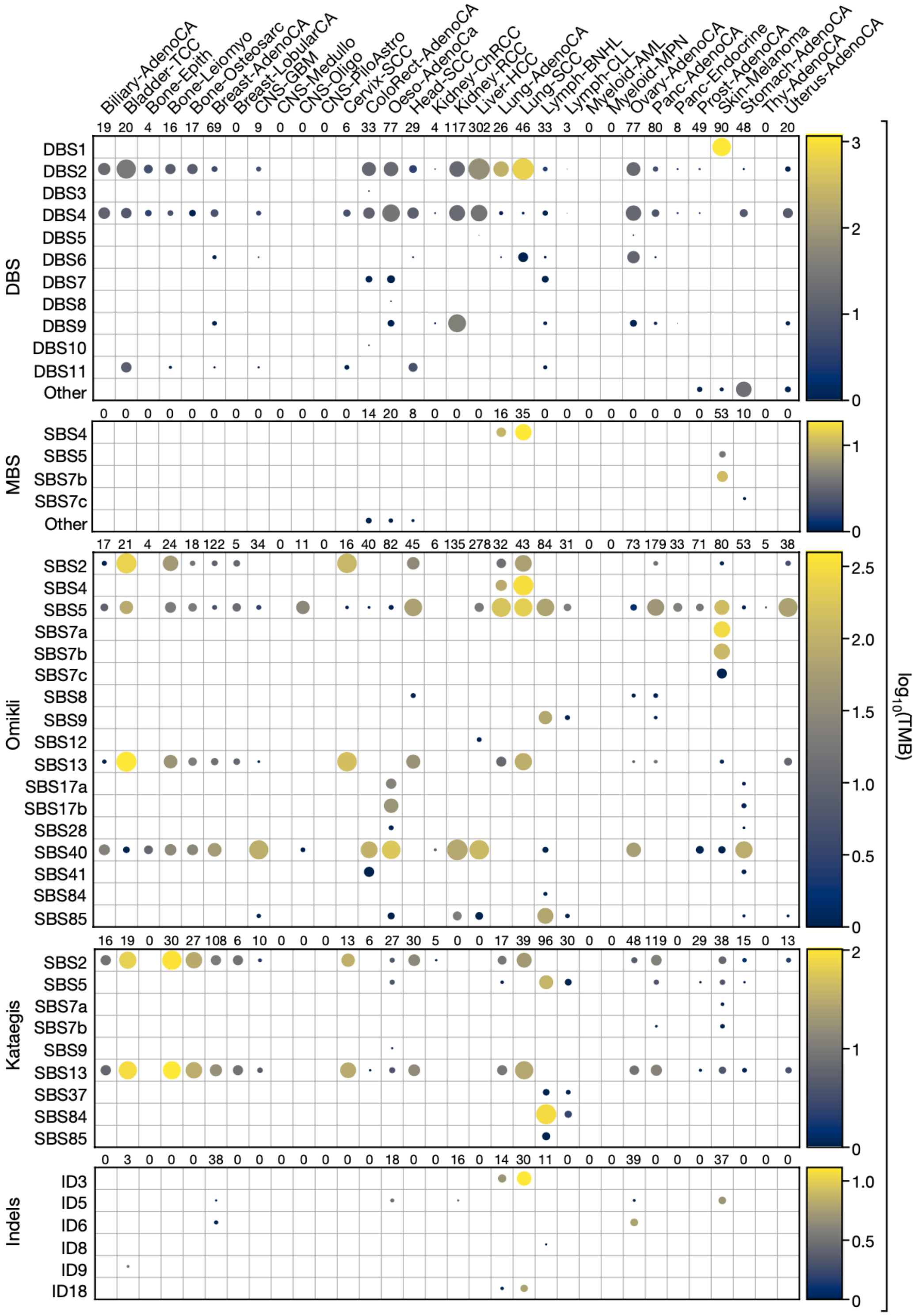


Figure 5.4: Mutational processes that underlie clustered events. Each circle represents the activity of a signature for a given cancer type. The radius of the circle determines the proportion of samples with greater than a given number of mutations specific to each subclass; the colour reflects the median number of mutations per cancer type. A minimum of two samples are required per cancer type for visualization (Methods).

In cancer genomes, omikli were previously attributed to APOBEC3 mutagenesis [57] with some indirect evidence from experimental models [88, 168, 169]. Our analysis of sequencing data [170] from the clonally expanded breast cancer cell line BT-474 with active APOBEC3 mutagenesis experimentally confirmed the existence of APOBEC3-associated omikli events (cosine similarity: 0.99) (Figure 5.6a). Only 16.2% of omikli events across the 2,583 cancer genomes matched the APOBEC3 mutational pattern, suggesting that a variety of other processes can give rise to diffuse clustered hypermutation. Notably, our analysis revealed omikli due to tobacco smoking (SBS4), clock-like mutational processes (SBS5), ultraviolet light (SBS7), both direct and indirect mutations from AID (SBS9 and SBS85), and multiple mutational signatures with unknown aetiology in different cancer types (SBS8, SBS12, SBS17a/b, SBS28, SBS40 and SBS41) (Figure 5.4). Cell lines previously exposed to benzo[a]pyrene [171] and ultraviolet light [172] confirmed the generation of omikli events as a result of these two environmental exposures (cosine similarities: 0.86 and 0.84, respectively) (Figure 5.6a).

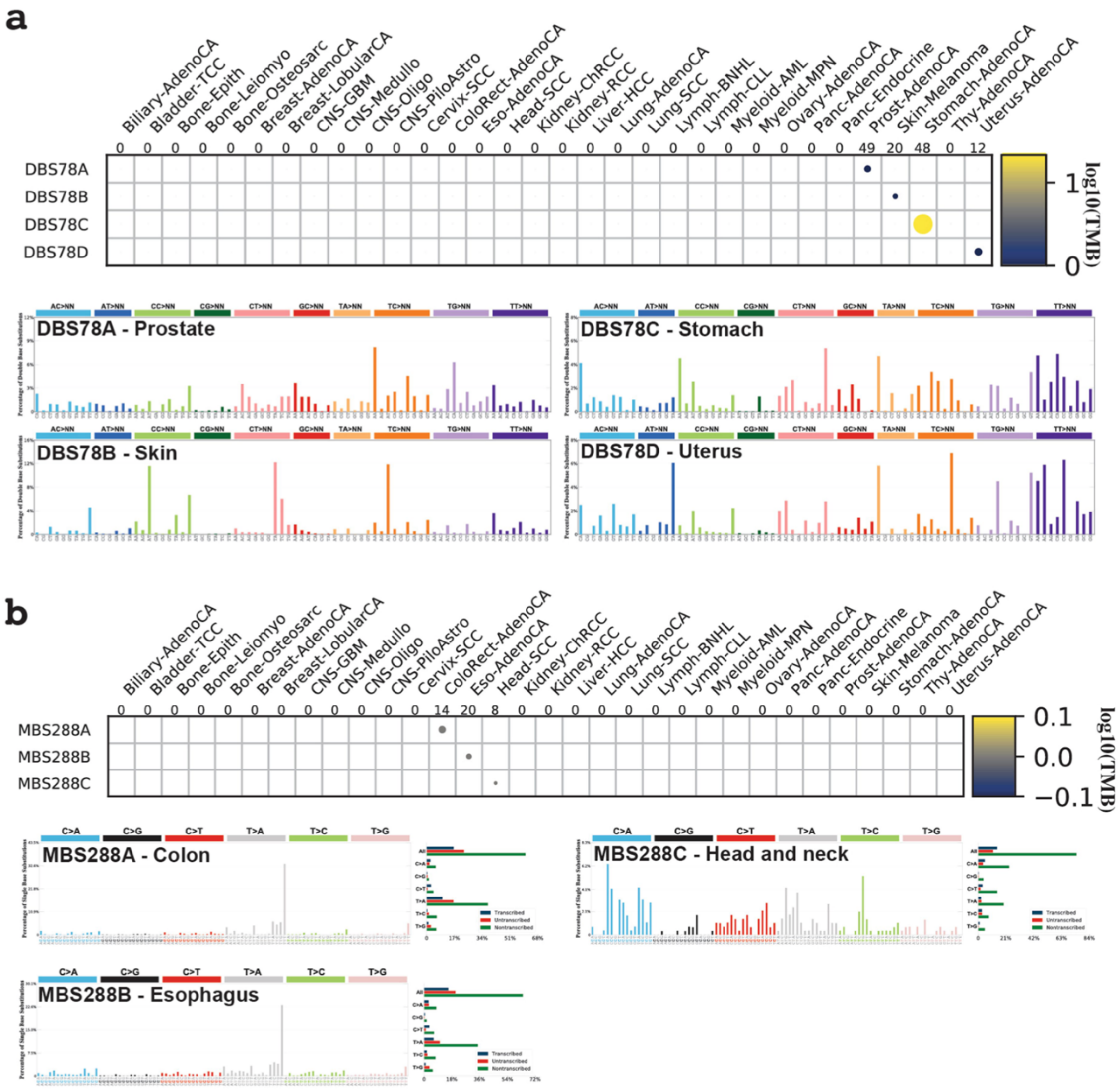


Figure 5.5: De novo signatures of DBS and MBS signatures. **a.** The activity of DBS de novo signatures (top) and the corresponding signatures extracted from prostate, skin, stomach, and uterine cancers that could not be accurately reconstructed using known COSMIC mutational signatures (bottom; Methods). **b.** The activity of MBS de novo signatures (top) and the corresponding signatures extracted from colon, oesophagus, and head and neck cancers that could not be accurately reconstructed using known COSMIC mutational signatures (bottom; Methods).

Of the nine kataegic signatures, four have been reported previously, including two associated with APOBEC3 deaminases (SBS2 and SBS13) and two associated with canonical or non-canonical AID activities (SBS84 and SBS85) (Figure 5.4). SBS5 (clock-like mutagenesis) accounted for 15.0% of kataegis, with most events occurring in the vicinity of AID kataegis within B cell lymphomas. The remaining four kataegic signatures accounted for only 8.9% of kataegic mutations and included SBS7a/b (ultraviolet light), SBS9 (indirect mutations from AID) and SBS37 (unknown aetiology). Most kataegic signatures were strand-coordinated (Figure 5.6b). Some samples exhibited consistent whereas others exhibited distinct signatures of clustered and non-clustered mutagenesis (Figure 5.7). For example, in SP56533 (lung squamous cell carcinoma), most non-clustered and omikli substitutions were caused by tobacco signature SBS4, whereas kataegic events were generated by the APOBEC3 signatures (Figure 5.7a). By contrast, the pattern of non-clustered substitutions in SP24815 (glioblastoma) was due to clock-like signatures SBS1 and SBS5, whereas omikli and kataegic events were mostly attributable to APOBEC3 (Figure 5.7a).

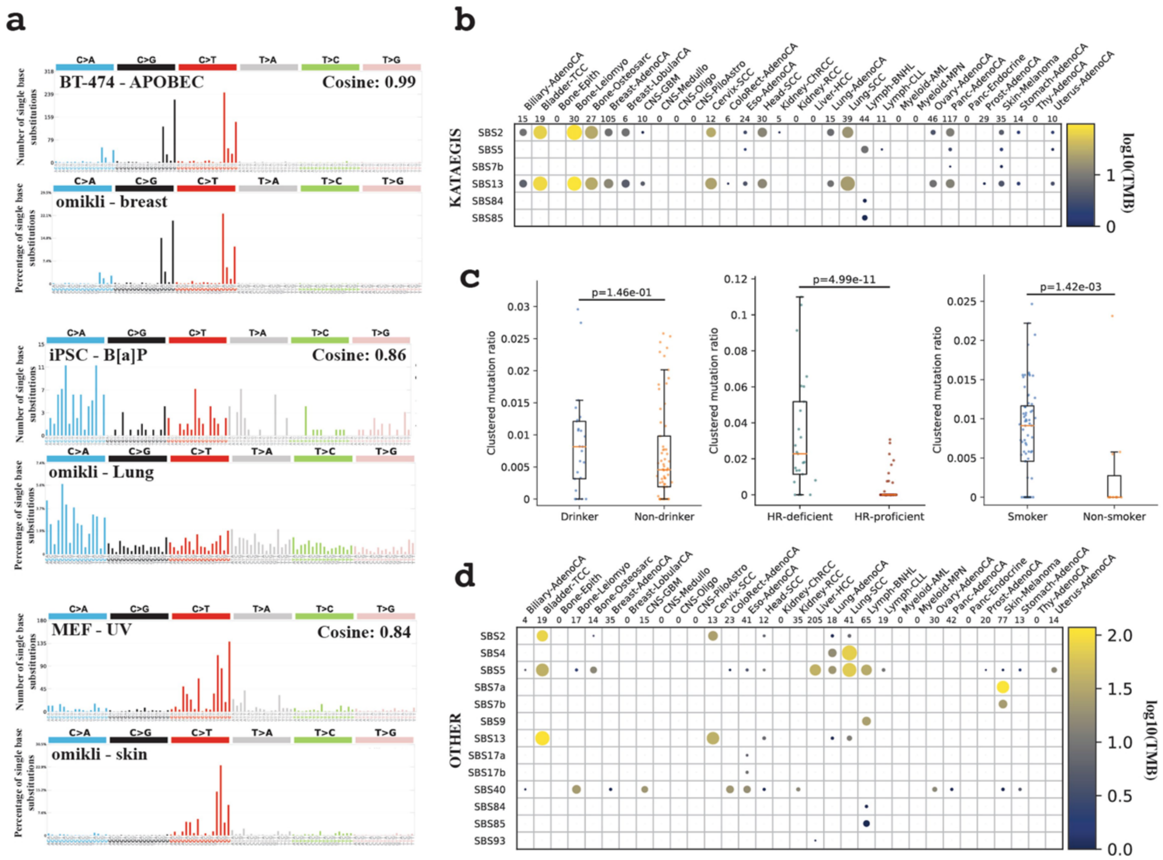


Figure 5.6: Experimental validation and epidemiological associations of clustered mutational processes. **a.** Experimental validation of three omikli processes. Specifically, APOBEC3-associated omikli were validated using a clonally expanded BT-474 breast cancer cell line (top), omikli events resulting from exposure to benzo[a]pyrene were validated using iPS cells (middle), and omikli events resulting from exposure to ultraviolet light were validated using iPS cells (bottom). **b.** Mutational processes of strand-coordinated kataegic events. **c.** Epidemiological associations comparing the ratio of clustered TMB to the total TMB for a given sample between: drinkers (n = 25) and non-drinkers (n = 61); smokers (n = 68) and non-smokers (n = 11); homologous-recombination deficient (HR-deficient; n = 25) and homologous-recombination proficient samples (HR-proficient; n = 64). For each boxplot, the middle line reflects the median, the lower and upper bounds of the box correspond to the first and third quartiles, and the lower and upper whiskers extend from the box by 1.5x the inter-quartile range (IQR). P-values were calculated using a two-tailed Mann–Whitney U-test. **d.** Mutational processes of clustered events with inconsistent VAFs classified as other clustered substitutions. A minimum of two samples are required per cancer type for visualization (Methods).

The remaining 'other' clustered substitutions exhibited inconsistent VAFs that probably represent mutations at highly mutable genomic regions or the effects of co-occurring large mutational events such as copy number alterations (Figure 5.6d).

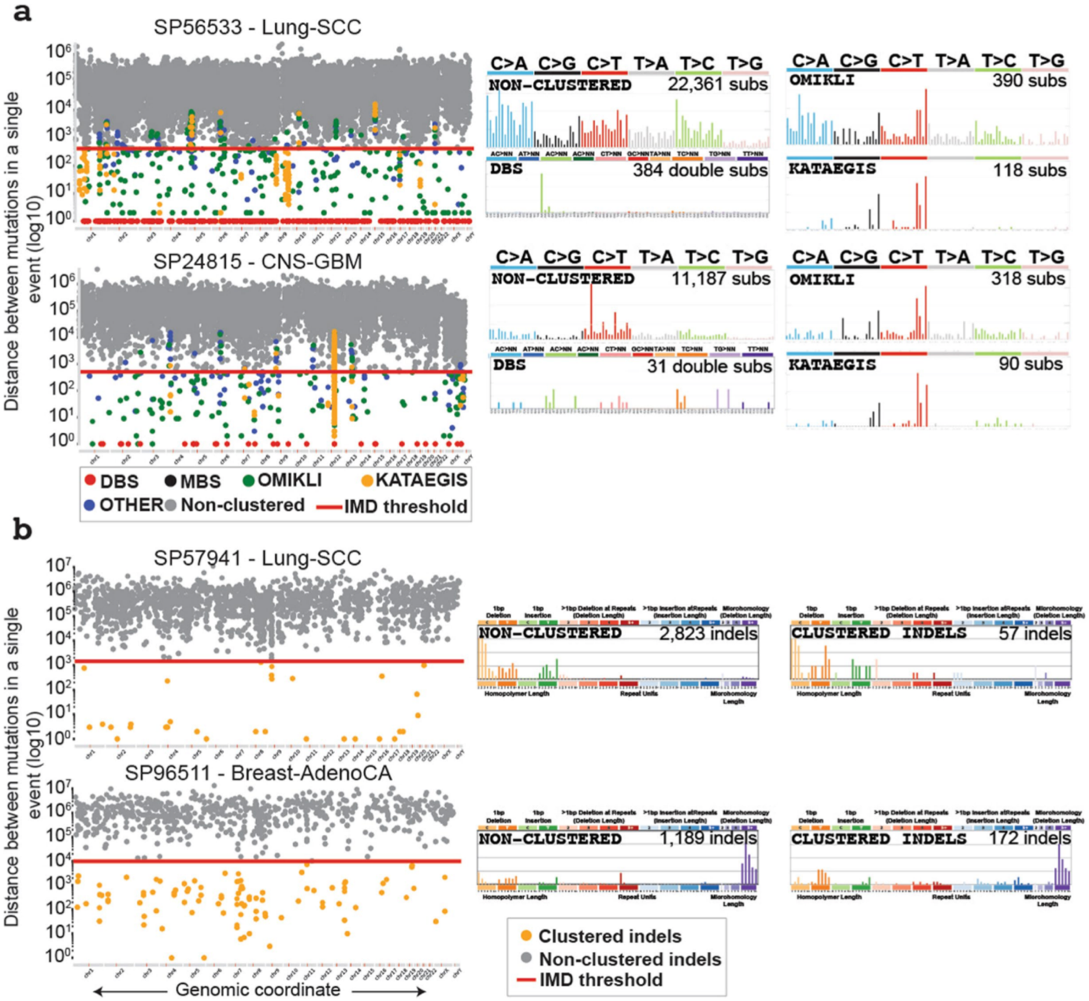


Figure 5.7: Examples of clustered mutational signatures. **a.** Two samples depicting the intra-mutational distance (IMD) distributions of substitutions across genomic coordinates, where each dot represents the minimum distance to adjacent mutations for a selected mutation coloured based on the corresponding subclassification of event (rainfall plot; left). The red lines depict the sample-dependent IMD threshold for each sample. Specific clustered mutations may be above this threshold based on corrections for regional mutation density. The mutational spectra for the different catalogues of clustered and non-clustered substitutions for each sample (right; MBS are not shown). **b.** Two samples illustrating the IMD distributions of indels across the given genomes, with the IMD indel thresholds shown in red (left). The non-clustered and clustered indel catalogues for each sample (right).

Different cancers showed distinct tendencies of clustered indel mutagenesis (Figure 5.4). For instance, clustered indels attributed to ID3 (tobacco smoking; characterized by 1-bp deletions) were found predominately in lung cancers and were significantly increased in smokers compared to non-smokers (P = 0.0014) (Figure 5.6c, Figure 5.7b). Clustered indels due to signatures ID6 and ID8—both attributed to homologous recombination deficiency and characterized by long indels at microhomologies—were found in breast and ovarian cancers and were highly increased in cancers with known deficiencies in homologous recombination genes (P = 4.9 × 10−11) (Figure 5.6c, Figure 5.7b).

5.1.3. Panorama of clustered driver mutations

The PCAWG project elucidated a constellation of mutations that putatively drive cancer development [60]. Our current analysis reveals significant enrichments of clustered substitutions and clustered indels amongst these driver mutations. Specifically, whereas only 3.7% of all substitutions and 0.9% of all indels are clustered events, they contribute 8.4% and 6.9% of substitution and indel drivers, respectively (q-values < 1 × 10−5; Fisher's exact tests) (Figure 5.8a,b). Omikli accounted for 50.5% of all clustered substitution drivers, whereas DBSs, kataegis and other clustered events each contributed between 14% and 18% (Figure 5.8c). Clustered driver substitutions varied greatly between genes and across different cancers (Figure 5.8c, Figure 5.9a) with a 2.4-fold enrichment of clustered events within oncogenes compared to tumour suppressors (P = 5.79 × 10−3) (Figure 5.9b,c). In some cancer genes, only a small percentage of driver events are due to clustered substitutions; examples include TP53 (4.5% clustered driver substitutions), KRAS (3.7%) and PIK3CA (2.2%). In other genes, most detected substitution drivers were clustered events; examples include: BTG1 (73.1%), SGK1 (66.6%), EBF1 (60.0%) and NOTCH2 (38.5%). Notably, the contribution from each class of clustered events varied across driver substitutions in different genes (Figure 5.8c). For instance, ultraviolet-light-associated DBSs comprised 93% of clustered BRAF driver events, omikli contributed 63% of clustered BTG1 driver events and kataegis accounted for 100% of clustered NOTCH2 driver substitutions (Figure 5.8c). Similar behaviour was observed for clustered indel drivers, with 48.7% being single-base pair indels (Figure 5.8d). In some cancer genes, clustered indel drivers were rare (for example, 2.4% of indel drivers in TP53 were clustered), whereas in others they were common (for example, 76.6% in ALB) (Figure 5.8d). Clustered driver substitutions were enriched in stop-lost mutations (q-value = 1.9 × 10−2) and depleted in stop-gained mutations (q-value = 3.3 × 10−3) when compared to non-clustered drivers (Figure 5.8e). Furthermore, driver genes that contained clustered events were often differentially expressed compared to those containing non-clustered events (Figure 5.9d). For instance, clustered events within CTNNB1 and BTG1 associated with an increased expression compared to both non-clustered and wild-type expression levels for each gene (q-values < 0.05). Opposite effects were observed in STAT6 and RFTN1 (q-values < 0.05). Collectively, these driver events were induced by the activity of multiple mutational processes including exposure to ultraviolet light, tobacco smoke, platinum chemotherapy and AID and APOBEC3 activity; amongst others (Figure 5.9e).

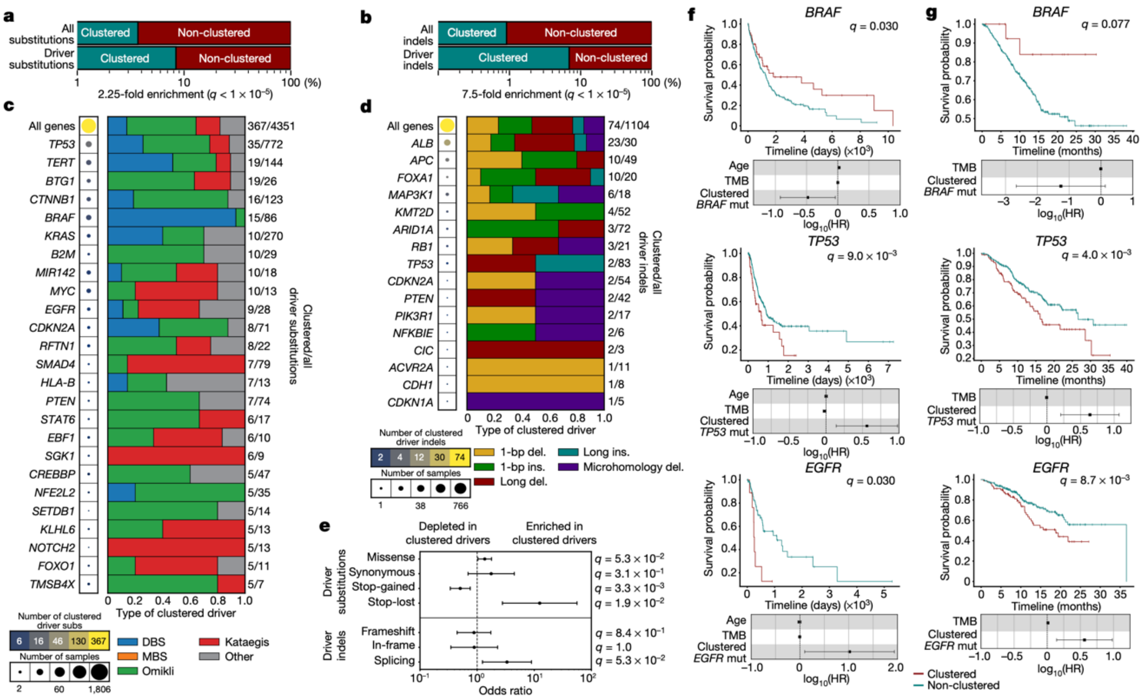


Figure 5.8: Panorama of clustered driver mutations in human cancer. **a, b.** Percentage of clustered mutations (top) compared to the percentage of clustered driver events (bottom) for substitutions (**a**) and indels (**b**). **c.** The frequency of clustered driver events across known cancer genes. The radius of the circle is proportional to the number of samples with a clustered driver mutation within a gene; the colour reflects the clustered mutational burden. All clustered driver events are classified into one of the five clustered classes, with the number of clustered driver substitutions and the total number of driver substitutions shown on the right. **d.** Clustered indel drivers are shown in a similar manner to **c, e.** The odds ratio of clustered substitutions (top) and indels (bottom) resulting in deleterious (n = 192 clustered substitutions; n = 54 clustered indels) or synonymous changes (n = 5 clustered substitutions; n = 5 clustered indels) within a given driver gene compared to non-clustered driver mutations (n = 771 deleterious and n = 237 synonymous substitutions; n = 111 deleterious and n = 50 synonymous indels). All events were overlapped with the PCAWG consensus list of driver events and were annotated using the ENSEMBL Variant Effect Predictor (VEP). The odds ratios are shown with their 95% confidence intervals. **f.** Kaplan–Meier survival curves comparing the outcome of samples with clustered versus non-clustered mutations in BRAF (top), TP53 (middle) and EGFR (bottom) across TCGA cohorts. Only cohorts with more than five samples containing a clustered mutation within the given gene were included. **g.** Kaplan–Meier survival curves comparing the outcome of samples with clustered versus non-clustered mutations in the same genes across the MSK-IMPACT cohort. The log10-transformed hazards ratios (log10(HR)) are shown with their 95% confidence intervals in **f, g.** Cox regressions were corrected for age (TCGA only), mutational burden and cancer type (Methods). Q values in **a, b, e.** were calculated using a two-tailed Fisher’s exact test and corrected for multiple hypothesis testing.

The clinical utility of detecting clustered events in driver genes was evaluated by comparing the survival amongst individuals with clustered mutations versus individuals with non-clustered mutations within each driver gene across all whole-exome-sequenced samples in TCGA. For each of these comparisons, we performed Cox regressions considering the effects from age and tumour mutational burden (TMB) while correcting for cancer type and multiple hypothesis testing. These results were validated in targeted panel sequencing data from the Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) cohort [173, 174]. These analyses revealed a significant difference in survival between individuals with clustered and individuals with non-clustered mutations detected in TP53, EGFR and BRAF. Specifically, individuals with clustered events within BRAF had a better overall survival compared to individuals with non-clustered events (q-values < 0.05) (Figure 5.8f,g). Conversely, in both TCGA and MSK-IMPACT, individuals with clustered mutations in TP53 or EGFR exhibited a significantly worse outcome compared to individuals with non-clustered mutations in each of these genes (q-values < 0.05) (Figure 5.8f, g).

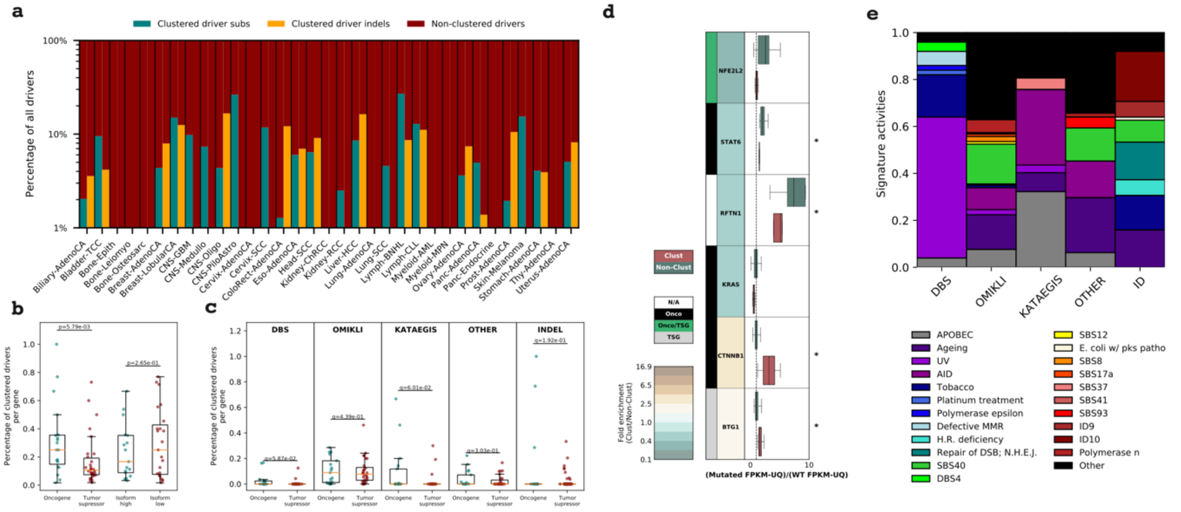


Figure 5.9: Mutational processes of clustered driver events. **a.** The percentage of clustered driver substitutions and indels within each cancer type. All samples 2,583 whole-genome sequenced samples from PCAWG with a detected driver event are included; however, cancer types with fewer than 10 samples are not presented. **b.** The proportion of clustered driver mutations per cancer gene compared between oncogenes (n = 19 genes) versus tumour suppressor genes (n = 30 genes) and genes with high numbers of isoforms (n = 17) versus genes with low numbers of isoforms (n = 23; upper and lower quartiles of isoforms across all cancer drivers). **c.** The proportion of clustered driver mutations for a given subclass per cancer gene compared between oncogenes (n = 17 genes with clustered substitutions and n = 13 with for clustered indels) versus tumour suppressor genes (n = 28 genes with clustered substitutions and n = 70 genes with clustered indels). **d.** The relative expression of driver genes containing clustered (copper) versus non-clustered events (green). All expression values were normalized using FPKM normalization and upper quartile normalization obtained from the official PCAWG release and were subsequently normalized using the average expression of the wild-type gene. A value of 1 (dashed lined) reflects no difference in expression compared to the wild-type gene. **e.** The proportional activity of mutational signatures contributing to clustered driver events within each subclass. MBSs did not contribute to any reported driver events. For analyses in **b–d**, p-values were generated using a two-tailed Mann–Whitney U-test (*P < 0.05; p = 0.03 for STAT6; p = 0.04 for CTNNB1; p = 0.02 for BTG1). For each boxplot, the middle line reflects the median, the lower and upper bounds of the box correspond to the first and third quartiles, and the lower and upper whiskers extend from the box by 1.5x the inter-quartile range (IQR).

5.1.4. Kataegic events and focal amplifications

In each sample, kataegic mutations were separated into distinct events on the basis of consistent VAFs across adjacent mutations and IMD distances greater than the sample-dependent IMD threshold (Methods). Our analysis revealed that 36.2% of all kataegic events occurred within 10 kb of a structural breakpoint but not on detected focal amplifications (Figure 5.10a). In addition, 21.8% of all kataegic events occurred either on a detected focal amplification or within 10 kb of a focal amplification's structural breakpoints: 9.6% on circular ecDNA, 6.3% on linear rearrangements, 3.3% within heavily rearranged events and 2.6% associated with breakage–fusion–bridge cycles (BFBs) (Figure 5.10a). Finally, 42.0% of kataegic events were neither within 10 kb of a structural breakpoint nor on a detected focal amplification. Modelling the distribution of the distances between kataegic events and the nearest structural variations revealed a multi-modal distribution with three components (Figure 5.10b): kataegis within 10 kb, around 1 Mb, or more than 1.5 Mb of a detected breakpoint. Of note, ecDNA-associated kataegis—termed kyklonas (Greek for cyclone)—had an average distance from the nearest breakpoint of around 750 kb, with only 0.35% of kyklonic events occurring both on ecDNA and within 10 kb of a breakpoint (Figure 5.10b). These results indicate that kyklonic events are not likely to have occurred because of structural rearrangements during the formation of ecDNA. In most cancer types, DBSs, MBSs, omikli and other cluster events were not found in the vicinity of structural variations (Figure 5.11a,b).

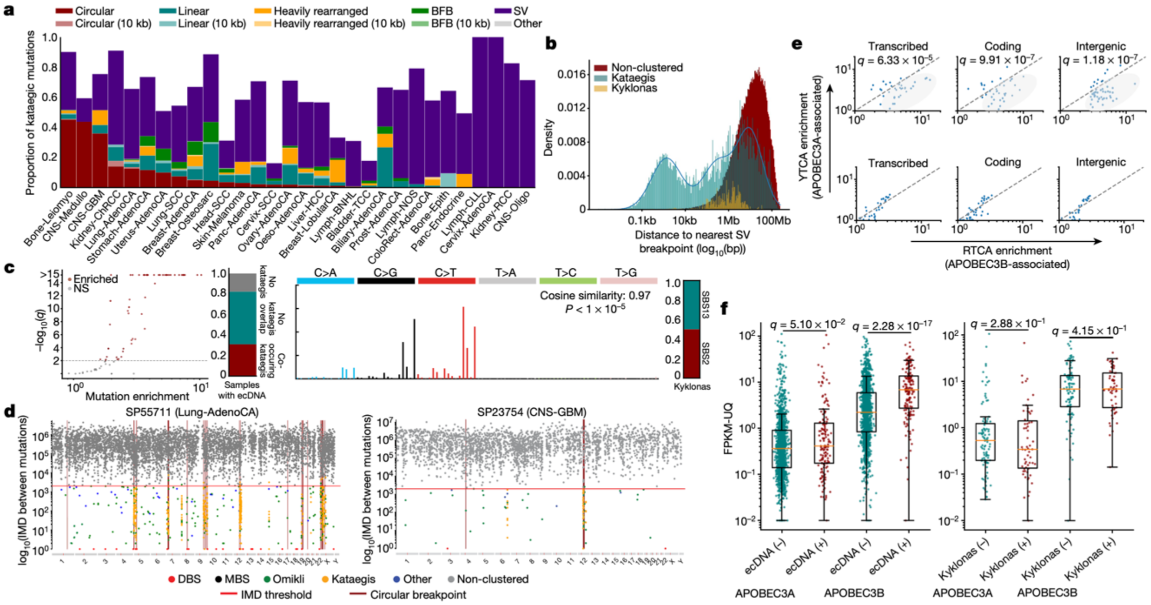


Figure 5.10: Kataegic events co-locate with most forms of structural variation. **a.** Proportion of all kataegic events per cancer type overlapping different amplifications or structural variations. **b.** Distance to the nearest breakpoint for all kataegic mutations (teal), kyklonas (gold) and non-clustered mutations (red). Kataegic distances were modelled as a Gaussian mixture with three components (blue line). **c.** Left, volcano plot depicting samples that are statistically enriched for kyklonas (red; q-values from a false discovery rate (FDR)-corrected z-test; not significant (NS)). Middle left, proportion of samples with ecDNA co-occurring with kataegis. Middle right, mutational spectrum of all kyklonas. Right, proportion of kyklonic events attributed to SBS2 and SBS13. Cosine similarity was calculated between the kyklonic and the reconstructed spectra composed using SBS2 and SBS13 (P value from a Z-score test). **d.** Rainfall plots illustrating the IMD distribution for a given sample with the genomic locations of ecDNA breakpoints (maroon). **e.** Top, YTCA versus RTCA enrichments per sample with kyklonas, in which YTCA or RTCA enrichment is suggestive of higher APOBEC3A or APOBEC3B activity, respectively. Genic mutations were divided into transcribed (template strand) and coding mutations. The RTCA/YTCA fold enrichments were compared to those of non-clustered mutations (bottom). **f.** Relative expression of APOBEC3A and APOBEC3B in samples containing ecDNA (n = 157) compared to samples without ecDNA (n = 1,364) (left), and in samples with ecDNA that have kyklonas (n = 59) compared to samples without kyklonas (n = 98) (right). Expression values were normalized using fragments per kilobase of exon per million mapped fragment (FPKM) and upper quartile (UQ) normalization obtained from the PCAWG release. Q values in **e, f.** were calculated using a two-tailed Mann–Whitney U-test and FDR corrected using the Benjamini–Hochberg procedure. For box plots, the middle line reflects the median, the lower and upper bounds of the box correspond to the first and third quartiles, and the lower and upper whiskers extend from the box by 1.5× the interquartile range.

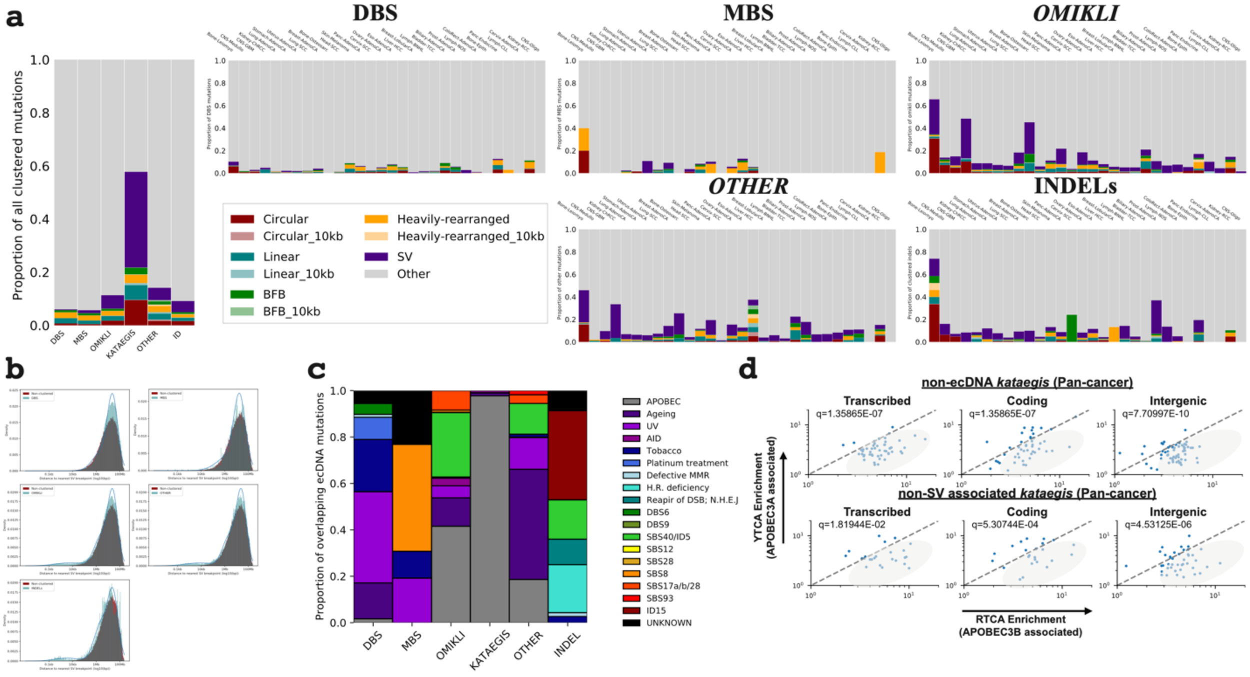


Figure 5.11: Clustered events and structural variations. **a.** The proportion of all clustered events co-locating with structural variations across all cancer types (left) and across each cancer type (right). **b.** The distance to the nearest structural variation for each class of clustered mutations (teal), and non-clustered mutations (red). The distribution for each class of clustered events were modelled using a Gaussian mixture (blue line). DBSs and MBSs were modelled using a single distribution, whereas omikli, other, and indels were modelled using two components reflecting the minimal distribution of overlap with structural variations. **c.** The mutational signatures active in ecDNA clustered events. **d.** YTCA versus RTCA enrichments per sample within non-ecDNA kataegis (top) and non-SV associated kataegis (bottom), where YTCA and RTCA enrichment is suggestive of APOBEC3A or APOBEC3B activity, respectively. Genic mutations were divided into transcribed (template strand) and coding mutations. The RTCA/YTCA fold enrichments were compared to the fold enrichments of non-clustered mutations (p-values calculated using two-tailed Mann–Whitney U-tests and corrected for multiple hypothesis testing using the Benjamini–Hochberg FDR procedure).

5.1.5. Recurrent kyklonic mutagenesis of ecDNA

Although only 9.6% of kataegic events occur within ecDNA regions, more than 30% of ecDNAs had one or more associated kyklonic events (Figure 5.10c). The mutations within these ecDNA regions were dominated by the APOBEC3 patterns, which are characterized by strand-coordinated C>G and C>T mutations in the TpCpW context and attributed to signatures SBS2 and SBS13 (P <1 × 10−5) (Figure 5.10c, d, Figure 5.11c). These APOBEC3-associated events contributed 97.8% of all kyklonic events, whereas the remaining mutations were attributed to clock-like signature SBS5 (1.2%) and other signatures (1.0%) (Figure 5.11c). Furthermore, kyklonic events exhibited an enrichment of C>T and C>G mutations at APOBEC3B-preferred RTCA compared to APOBEC3A-preferred YTCA contexts (underlining reflects the mutated nucleotide) [59], indicating that APOBEC3B is likely to have an important role in the mutagenesis of circular DNA bodies (Figure 5.10e). Similar levels of enrichment for RTCA contexts were also observed in both non-ecDNA kataegis and non-structural variant (SV)-associated kataegis, suggesting that APOBEC3B generally gives rise to many of the strand-coordinated kataegic events (Figure 5.11d). An increase in the expression of APOBEC3B—but not APOBEC3A—was observed in cancers with ecDNA compared to samples without ecDNA (3.1-fold; q-value < 1 × 10−5) (Figure 5.10f). Within cancers containing ecDNA, no differences were observed in the expression of APOBEC3A or APOBEC3B between samples with and without kyklonic events (Figure 5.10f).

More recurrent APOBEC3 kataegis was observed across circular ecDNA regions compared to other forms of structural variation (Figure 5.12a). An average of 2.5 kyklonic events were observed within ecDNA regions (range: 0–64 kyklonic events; 0–505 mutations). Recurrent kyklonas was widespread across cancer types (Figure 5.13a,b). For example, glioblastomas and sarcomas exhibited an average of 5 and 86 kyklonic mutations, respectively. The average VAF of kyklonas was significantly lower than both non-ecDNA associated kataegis and all other clustered events (q-values < 1 × 10−5 Figure 5.12b). Notably, a subset of kyklonas exhibited VAFs above 0.80, which is likely to reflect early mutagenesis of genomic regions that have subsequently amplified as ecDNA. Moreover, kyklonic events with high VAFs occurred more commonly on ecDNA that contained known cancer genes, suggesting a mechanism of positive selection (Figure 5.12b). Approximately 7.2% of kyklonas occurred early in the evolution of a given ecDNA population within a tumour (VAF > 0.80), whereas the majority of kyklonic events (around 82.5%; VAF < 0.5) have probably occurred after clonal amplification by recurrent APOBEC3 mutagenesis.

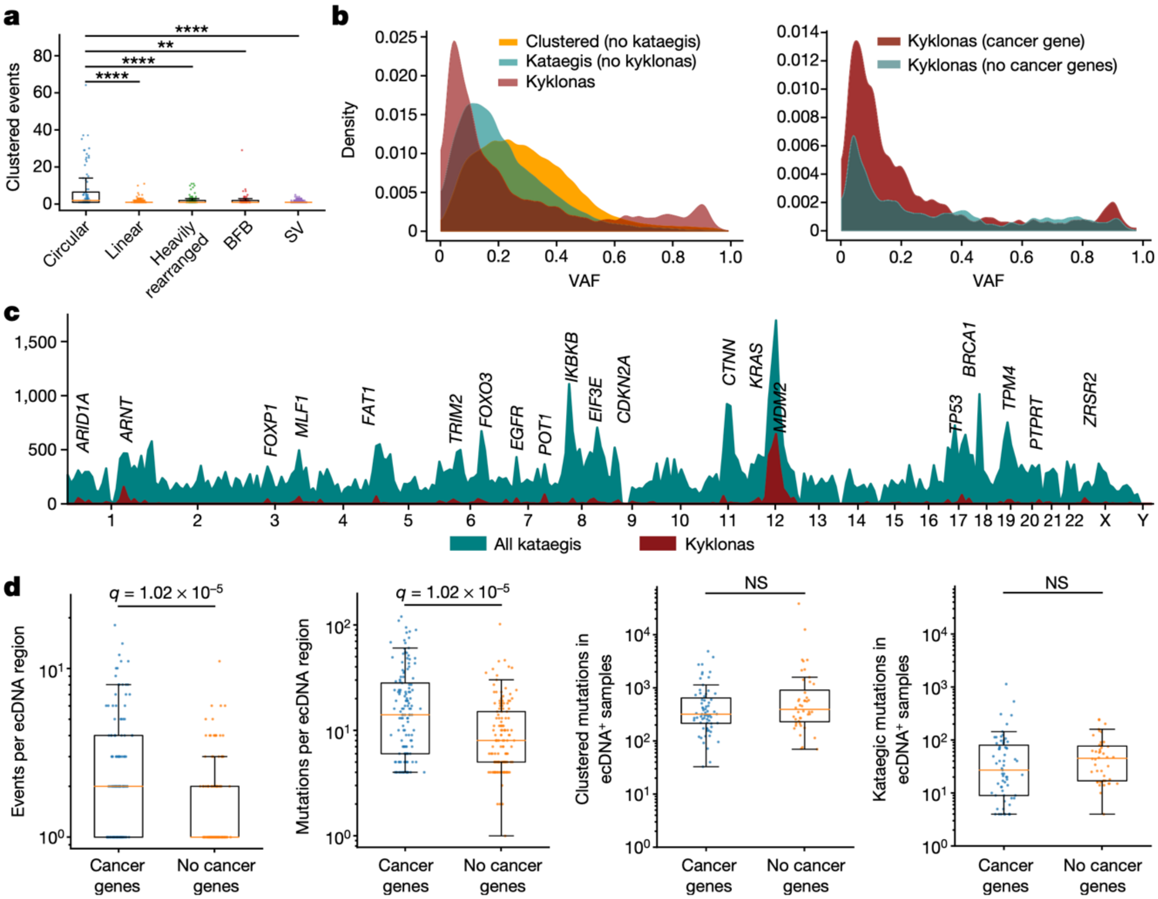


Figure 5.12: Recurrent APOBEC3 hypermutation of ecDNA. **a.** Number of clustered events overlapping a single amplicon or SV event; each dot represents an amplicon or SV (n = 84 circular; n = 275 linear; n = 111 heavily rearranged; n = 62 BFB; and n = 11,139 SV). A 10-kb window was used to determine the co-occurrence of kataegis with SV breakpoints (**q < 0.01, ****q < 0.0001). **b.** Left, normalized distributions of the VAFs for all clustered mutations excluding kataegis (orange), all non-ecDNA kataegis (teal), and kyklonas (red). Right, normalized VAF distributions for kyklonic ecDNA containing cancer genes and for kyklonic ecDNA without cancer genes. **c.** Frequency of recurrence for all kataegis (teal) and kyklonas (red) using a sliding genomic window of 10 Mb. **d.** Number of kyklonic events and kyklonic mutations per ecDNA region containing cancer genes (n = 137) or without cancer genes (n = 134; left and right, respectively). **e.** Total number of clustered and kataegic mutations found in samples with ecDNAs containing cancer genes (n = 67 samples) compared to samples with ecDNAs without cancer genes (n = 44; left and right, respectively). Q values in **a, d, e.** were calculated using a two-tailed Mann–Whitney U-test and FDR-corrected using the Benjamini–Hochberg procedure. Box plot parameters as in Figure 5.10.

Recurrent kyklonic events were increased within or near known cancer-associated genes including TP53, CDK4 and MDM2, amongst others (Figure 5.12c). These recurrent kyklonas were observed across many cancers including glioblastomas, sarcomas, head and neck carcinomas and lung adenocarcinomas (Figure 5.13c,d). For example, in a sarcoma sample (SP121828), 10 distinct kyklonic events overlapped a single ecDNA region with recurrent APOBEC3 activity in proximity to MDM2, resulting in a missense L230F mutation (Figure 5.13c). The same ecDNA region contained additional kyklonic events occurring within intergenic regions that have distinguishable VAF distributions, implicating recurrent mutagenesis (Figure 5.13c). Similarly, two distinct kyklonic events occurred on an ecDNA containing EGFR, resulting in a missense mutation D191N within a head and neck cancer (Figure 5.13d). Of note, ecDNA regions with known cancer-associated genes had significantly higher numbers of kyklonic events and mutational burdens of kyklonas compared to ecDNA regions without any known cancer-associated genes (q-values < 1 × 10−5) (Figure 5.12d). Furthermore, we observed a higher co-occurrence of kyklonas with known cancer-associated genes, which were mutated 2.5 times more than ecDNA without cancer-associated genes (P = 1.2 × 10−5; Fisher's exact test). Overall, 41% of kyklonic events were found within the footprints of known cancer driver genes (P < 1 × 10−5). These enrichments cannot be accounted for either by an increase in the overall mutations or by an increase in the overall clustered mutations in these samples (Figure 5.12e). To understand the functional effect of kyklonas, we annotated the predicted consequence of each mutation. In total, 2,247 kyklonic mutations overlapped putative cancer-associated genes, of which 4.3% occur within coding regions (Figure 5.13e). Specifically, 63 resulted in missense mutations, 29 resulted in synonymous mutations, 4 introduced premature stop codons and 1 removed a stop codon. These downstream consequences of APOBEC3 mutagenesis suggest a contribution to the oncogenic evolution of specific ecDNA populations.

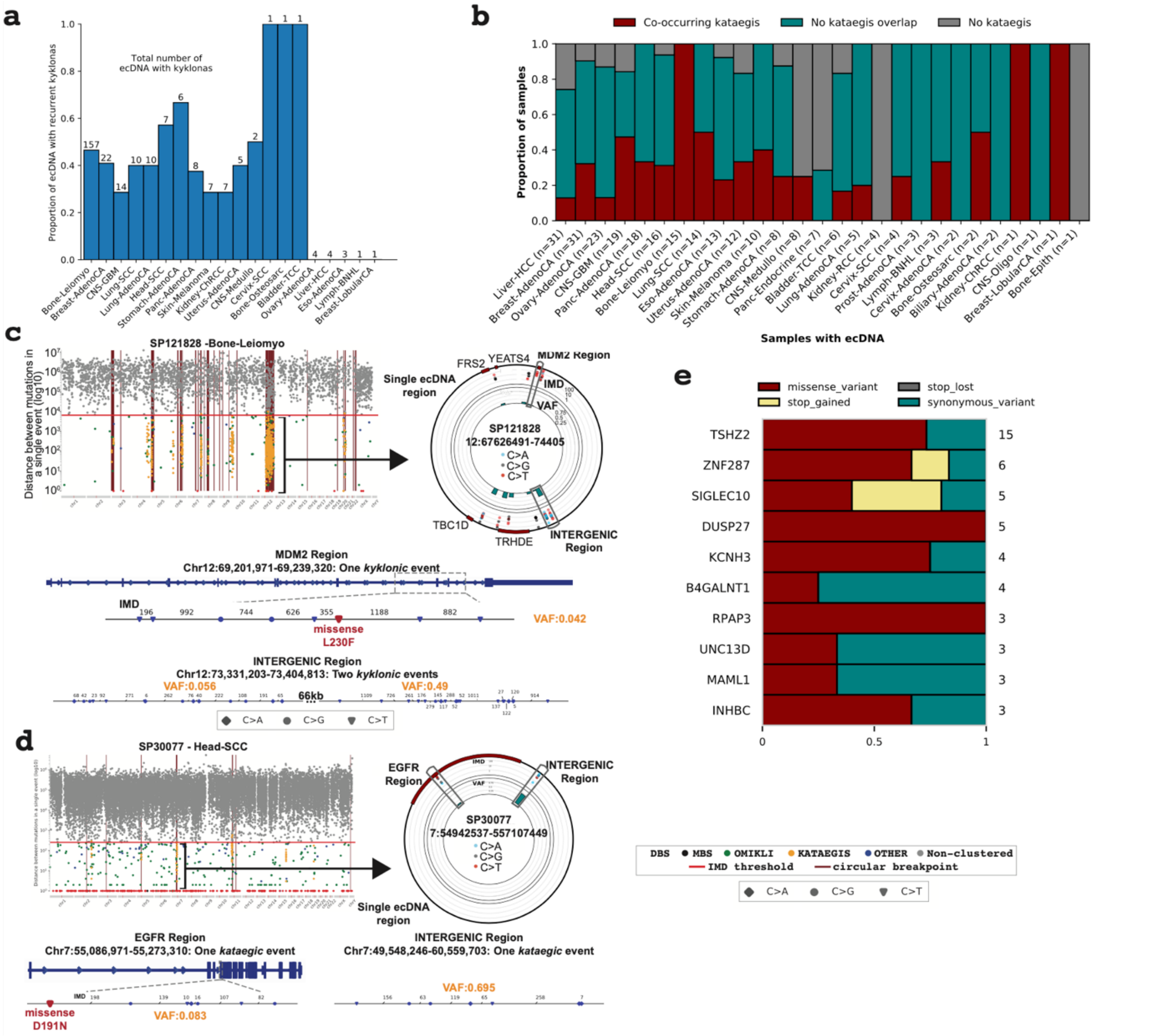


Figure 5.13: Recurrent mutagenesis and functional effects of kyklonas. **a.** The total number of recurrently mutated ecDNA displayed as a proportion of the total number of ecDNA with kyklonas for a given cancer type. The total number of ecDNA with kyklonas are displayed above each bar plot for each cancer type. All ecDNA with recurrent hypermutation were considered enriched for kyklonic events after correcting for multiple hypothesis testing (Z-score test; q-values < 0.05). **b.** Proportion of samples containing ecDNA divided exclusively into those with co-occurring kataegis, no kataegis overlap, and no detected kataegis across the entire genome. The number of samples included in each cancer type are listed. For certain cancer types, as few as a single sample may represent the entire proportional breakdown (for example, Bone-Osteosarc or Bone-Epith). **c.** A single sarcoma genome and **d,** a single head squamous cell carcinoma genome depicting the overlap of kataegis with ecDNA regions displayed as a rainfall (top left) with a single zoomed in ecDNA represented using a circos plot (top right). Bottom: Two regions of the ecDNA with overlapping kyklonic events. VAFs are shown per event (orange). **e.** Kyklonic substitutions resulting in recurrent coding mutations within known cancer genes.

5.1.6. Validation of kyklonic events in ecDNA

Kyklonic events were further investigated across 3 additional independent cohorts, including 61 sarcomas [136], 280 lung cancers [175] and 186 oesophageal squamous cell carcinomas [176]. Comparable rates of clustered mutagenesis were found for both substitutions and indels to the rates reported in PCAWG, with a 2.4- and 5.0-fold enrichment of clustered substitutions and indels within driver events, respectively (Figure 5.14a). Across the three cohorts, 31% of samples with ecDNA exhibited kyklonas within the sarcomas, 14% within the oesophageal cancers and 28% within the lung cancers, supporting the rates observed in PCAWG (Figure 5.10c, Figure 5.13b, Figure 5.14c). Similar to the rate observed in PCAWG (36.2%), approximately 30.1% of all kataegis occurred within 10 kb of the nearest breakpoint in the validation cohort (Figure 5.15a). In addition, only 0.34% of kyklonic events in the validation dataset occurred closer to SVs than expected by chance, which closely resembles the observations in the PCAWG data (0.35%) (Figure 5.15b). Kyklonic mutations were predominantly attributed to APOBEC3 signatures SBS2 and SBS13 (P < 1 × 10−5) (Figure 5.14b, Methods) with an enrichment of mutations at the RTCA context supporting the role of APOBEC3B (Figure 5.14d). A widespread recurrence of kyklonic events was observed across the sarcomas, oesophageal and lung cancers, with 45%, 28% and 46% of samples with ecDNA containing multiple, distinct kyklonic events (Figure 5.14e). An example from each cohort was selected to illustrate multiple kyklonic events occurring within single ecDNAs, validating the recurrent APOBEC3 hypermutation of ecDNA (Figure 5.16).

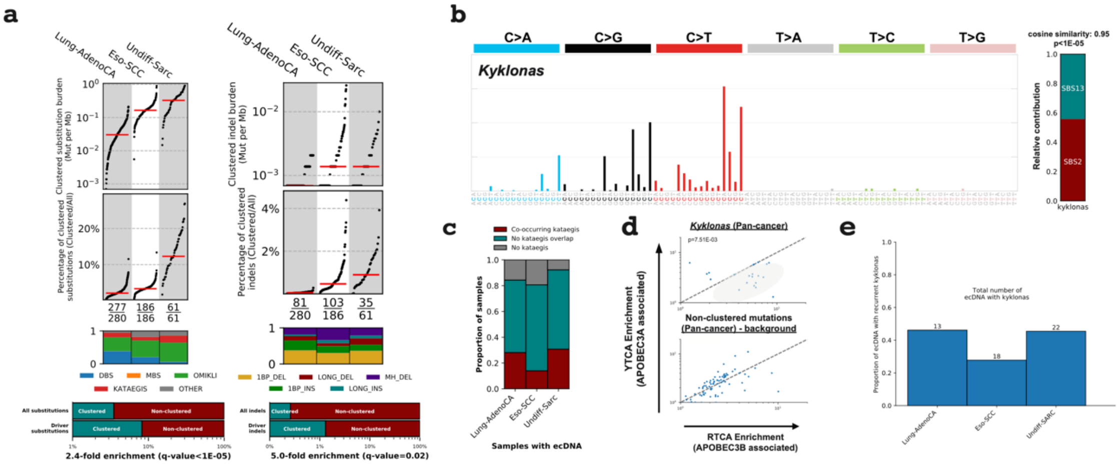


Figure 5.14: Validation of APOBEC3 hypermutation of ecDNA in three independent cohorts. **a.** Distribution of clustered substitutions (left) and clustered indels (right) across three validation cohorts. Clustered substitutions were subclassified into DBSs, MBSs, *omikli*, kataegis, and other clustered mutations. Top: Each black dot represents a single cancer genome. Red bars reflect the median clustered TMB and the percentage of clustered mutations contributing to the overall TMB of a given sample for each cancer type. Middle: The proportion of each subclass of clustered events for a given cancer type with the total number of samples having at least a single clustered event over the total number of samples within a given cancer cohort. Bottom: Percentage of clustered mutations compared to the percentage of clustered driver events for substitutions (left) and indels (right). P-values were calculated using a Fisher's exact test and corrected for multiple hypothesis testing using Benjamini–Hochberg FDR procedure. **b.** Left: The mutational spectrum of all kyklonas across the validation cohorts. Right: The proportion of kyklonic events attributed to SBS2 and SBS13 (p-value determined using a Z-score test; Methods). **c.** The proportion of samples with ecDNA that co-occur with kataegis, do not co-occur with kataegis, or do not have any detected kataegic activity across each cohort. **d.** YTCA versus RTCA enrichments per sample with kyklonas, where YTCA and RTCA enrichment is suggestive of higher APOBEC3A or APOBEC3B activity, respectively. The RTCA/YTCA fold enrichments were compared to the fold enrichments of non-clustered mutations (p-values calculated using a two-tailed Mann–Whitney U-test). **e.** The proportion of ecDNA with kyklonas that contain multiple kyklonic events. The total number of ecDNA with kyklonas are displayed above each bar plot for each cancer type.

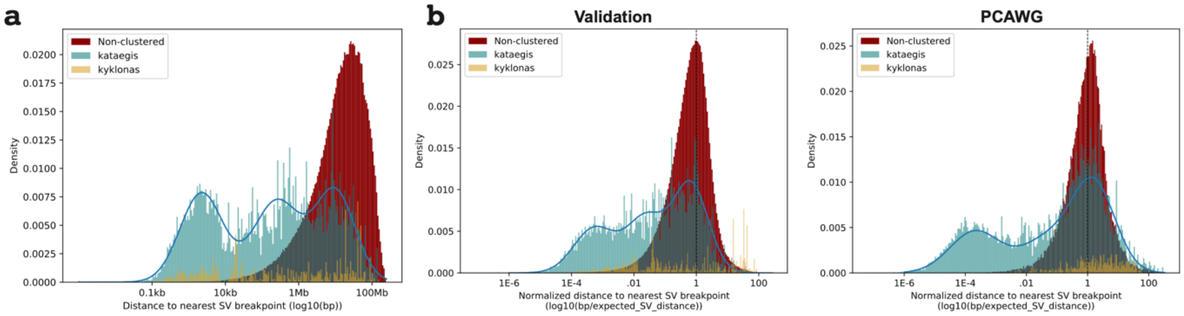


Figure 5.15: Kyklonas occur distally from structural breakpoints across three independent cohorts. **a.** The distance to the nearest breakpoint for all kataegic mutations (teal), kyklonas (gold), and non-clustered mutations (red) across the three validation cohorts. **b.** Distances to the nearest SV breakpoints were normalized by calculating the expected distance a mutation would fall from a breakpoint given the number of breakpoints detected per chromosome and the overall length of the chromosome across the validation cohorts (left) and PCAWG (right). A value of 1 (dashed line) reflects a distance that one would expect based on the random placement of a mutation across the chromosome, whereas a value less than 1 reflects a mutation occurring closer than what is expected by random chance. The distributions of kataegic mutations were modelled using Gaussian mixture models (blue lines) with an automatic selection criterion for the number of components using the minimum Bayesian information criteria (BIC).

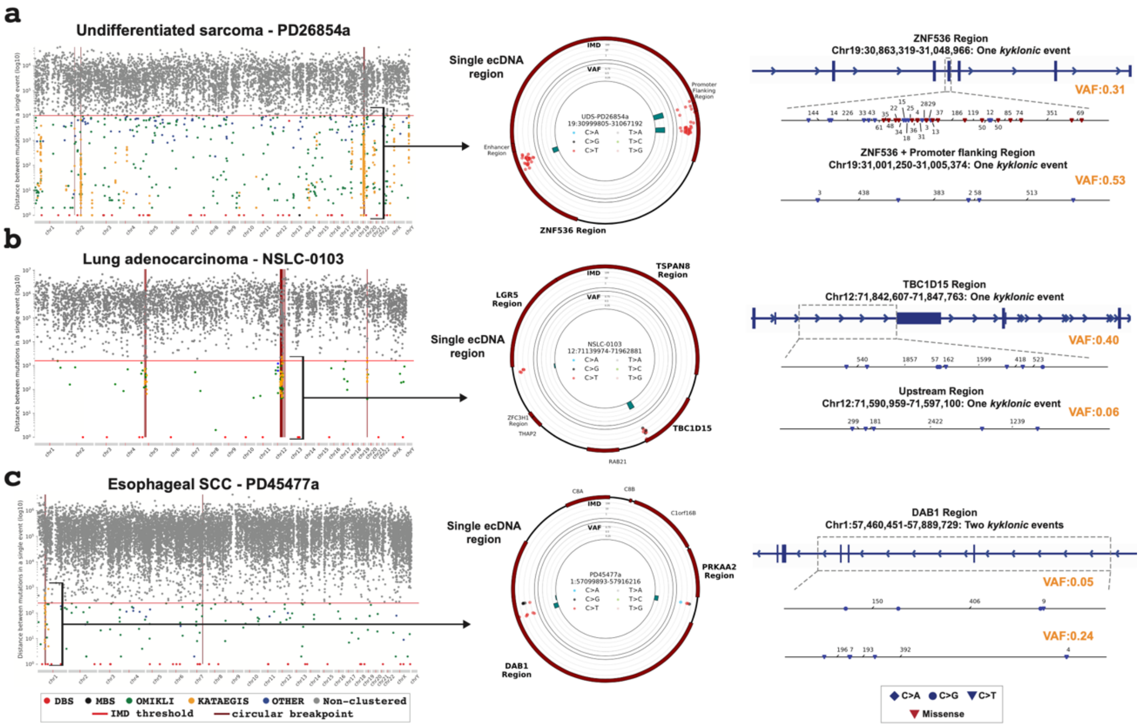


Figure 5.16: Examples of kyklonas in three independent cohorts. **a.** A single undifferentiated sarcoma genome depicting the overlap of kataegis with ecDNA regions displayed as a rainfall (left) with a single zoomed in ecDNA represented using a circos plot (middle). The outer track of the circos plot represents the reference genome of the ecDNA with proximal known cancer driver genes. The middle track reflects a circular rainfall plot where each dot represents the IMD around a single mutation coloured based on the substitution change. The innermost track shows the average VAF for each kyklonic event. Right: Two smaller regions of the selected ecDNA including a single kyklonic event within ZNF536 region resulting in a plethora of missense and stop-gained mutations, and a single kyklonic event within a promoter flanking with the average VAFs per event (orange). **b.** A single lung adenocarcinoma genome depicting the overlap of kataegis with ecDNA regions (left) with a single zoomed in ecDNA containing TBC1D15 and two distinct kyklonic events represented using a circos plot (middle). Right: Two kyklonic events overlapping an upstream region and TBC1D15. **c.** A single oesophageal squamous cell carcinoma genome depicting the overlap of kataegis with ecDNA regions (left) with a single zoomed in ecDNA containing PRKAA2 and DAB1 and three distinct kyklonic events (middle). Right: Two kyklonic events overlapping DAB1.

5.2. Discussion

Clustered mutagenesis in cancer can occur through different mutational processes, with AID and APOBEC3 deaminases having the most prominent role. In addition to enzymatic deamination, other endogenous and exogenous sources imprint many of the observed clustered indels and substitutions. A multitude of mutational processes can give rise to omikli events, including tobacco carcinogens and exposure to ultraviolet light. Clustered substitutions and indels were highly enriched in driver events and associated with differential gene expression, implicating them in cancer development and cancer evolution. Some clustered mutational signatures are associated with known cancer risk factors or the activity or failure of DNA repair processes. Notably, clustered mutations in TP53, EGFR and BRAF associated with changes in overall survival and can be detected in most types of sequencing data, including clinically actionable targeted panels such as MSK-IMPACT.

A large proportion of kataegic events occur within 10 kb of detected SV breakpoints with a mutational pattern, suggesting the activity of APOBEC3. Multiple distinct kataegic events, independent of detected breakpoints, were observed on circular ecDNA; such events—termed kyklonas—suggest recurrent APOBEC3 mutagenesis. The circular topology of ecDNAs [177] and their rapid replication patterns are reminiscent of the structure and behaviour of the circular genomes of several double-stranded-DNA based, pathogens including herpesviruses, papillomaviruses and polyomaviruses [163-166]. Previous pan-virome studies have shown that these double-stranded DNA viral genomes often manifest mutations from APOBEC3 enzymes [178-180]. As such, recurrent APOBEC3 mutagenesis on ecDNA is likely to be representative of an antiviral response in which the ecDNA viral-like structure is treated as an infectious agent and attacked by APOBEC3 enzymes. ecDNAs contain a plethora of cancer-associated genes and are responsible for many gene amplification events that can accelerate tumour evolution. Repeated mutagenic attacks of these ecDNAs reveal functional effects within known oncogenes and implicate additional modes of oncogenesis that may ultimately contribute to subclonal tumour evolution, subsequent evasion of therapy and clinical outcome. Further investigations with large-scale clinically annotated whole-genome-sequenced cancers are required to fully understand the clinical implications of clustered mutations and kyklonas.

5.3. Methods

5.3.1. Data sources

Somatic variant calls of single-base substitutions, small indels and structural variations were downloaded for the 2,583 white-listed whole-genome-sequenced samples from PCAWG along with the corresponding list of consensus driver events [60]. Epidemiological and clinical features for all available samples were downloaded from the official PCAWG release (https://dcc.icgc.org/releases/PCAWG). The collection of whole-exome-sequenced samples from TCGA along with all available clinical features were downloaded from the Genomic Data Commons (GDC; https://gdc.cancer.gov/). The MSK-IMPACT Clinical Sequencing Cohort [174] composed of 10,000 clinical cases was downloaded from cBioPortal (https://www.cbioportal.org/study/summary?id=msk_impact_2017). The subclassification of focal amplifications comprised circular ecDNA, linear amplifications, BFBs and heavily rearranged events, and their corresponding genomic locations were obtained for a subset of samples (n = 1,291) as reported [166].

Experimental models used to validate clustered events were derived from previous studies using primary Hupki mouse embryonic fibroblasts (MEFs) exposed to ultraviolet light [172], human induced pluripotent stem cells (iPS cells) exposed to benzo[a]pyrene [171], and a clonally expanded BT-474 human breast cancer cell line with episodically active APOBEC3 [170].

Independent cohorts used to validate kyklonic events were collected from multiple sources. The 61 undifferentiated sarcomas [136] and 187 high-confidence oesophageal squamous cell carcinomas [176] were downloaded from the European Genome-phenome Archive (EGAD00001004162 and EGAD00001006868, respectively). The 280 lung adenocarcinomas [175] were downloaded from dbGaP under the accession number (phs001697.v1.p1). Clustered mutations in validation samples were analysed using the same approach as the one used in the original cohort.

5.3.2. Detection of clustered events

SigProfilerSimulator (v.1.0.2) was used to derive an IMD cut-off [158] that is unlikely to occur by chance based on the TMB and the mutational patterns for a given sample. Specifically, each tumour sample was simulated while maintaining the sample's mutational burden on each chromosome, the ±2 bp sequence context for each mutation and the transcriptional strand bias ratios across all mutations. All mutations in each sample were simulated 100 times and the IMD cut-off was calculated such that 90% of the mutations below this cut-off could not appear by chance ($q\text{-value} < 0.01$). For example, in a sample with an IMD threshold of 500bp, one may observe 1,000 mutations within this threshold with no more than 100 mutations expected based on the simulated data ($q\text{-value} < 0.01$). P values were calculated using z-tests by comparing the number of real mutations and the distribution of simulated mutations that occur below the same IMD threshold. A maximum cut-off of 10 kb was used for all IMD thresholds. By generating a background distribution that reflects the random distribution of events used to reduce the false positive rate, this model also considers regional heterogeneities of mutation rates, partially attributed to replication timing and expression, and variances in clonality by correcting for mutation-rich regions and mutation-poor regions within 1-Mb windows. The 1-Mb window size has been used and established as an appropriate scale when considering the variability in mutation rates associated with chromatin structure, replication timing and genome architecture [61, 160, 162]. The 1-Mb window ensures that subsequent mutations are likely to have occurred as single events using a maximum cut-off of 0.10 for differences in the VAFs. The regional IMD cut-off was determined using a sliding window approach that calculated the fold enrichment between the real and simulated mutation densities within 1-Mb windows across the genome. The IMD cut-offs were further increased, for regions that had higher than ninefold enrichments of clustered mutations and where more than 90% of the clustered mutations were found within the original data, to capture additional clustered events while maintaining the original criteria (less than 10% of the mutations below this cut-off appear by chance; q-value < 0.01). Last, as VAF of mutations may confound the definition of clustered events in ecDNA, we calculated the distribution of inter-event distances within recurrently mutated ecDNA while disregarding the VAF of individual mutations. This resulted in the exact same separation of kataegic events using only the inter-event distances as a criterion for the grouping of mutations into a single event.

Subsequently, all clustered mutations with consistent VAFs were classified into one of four categories (Extended Data Fig. 1a). Two adjacent mutations with an IMD of 1 were classified as DBSs. Three or more adjacent mutations each with an IMD of 1 were classified as MBSs. Two or three mutations with IMDs less than the sample-dependent threshold and with at least a single IMD greater than 1 were classified as omikli. Four or more mutations with IMDs less than the sample-dependent threshold and with at least a single IMD greater than 1 were classified as kataegis. A cut-off of four mutations for kataegis was chosen by fitting a Poisson mixture model to the number of mutations involved in a single event across all extended clustered events excluding DBSs and MBSs (Figure 5.2). This model comprised two distributions with C1 = 2.08 and C2 = 4.37 representing omikli and kataegis, respectively. A cut-off of four mutations was used for kataegis on the basis of a contribution of greater than 95% from the kataegis-associated distribution with events of four or more mutations. Note that there is certain ambiguity for events with two or three mutations. Although the majority of these events are omikli, some of these events are likely to be short kataegic events (Figure 5.2). All remaining clustered mutations with inconsistent VAFs were classified as other. Clustered indels were not classified into different classes. We also performed additional quality-checks to ensure that the majority of clustered indels were mapped to high confidence regions of the genome (Figure 5.17). Specifically, all clustered indels were aligned against a consensus list of blacklisted genomic regions developed by ENCODE [181] revealing that only 0.5% of all clustered indels overlapped regions with low mappability scores.

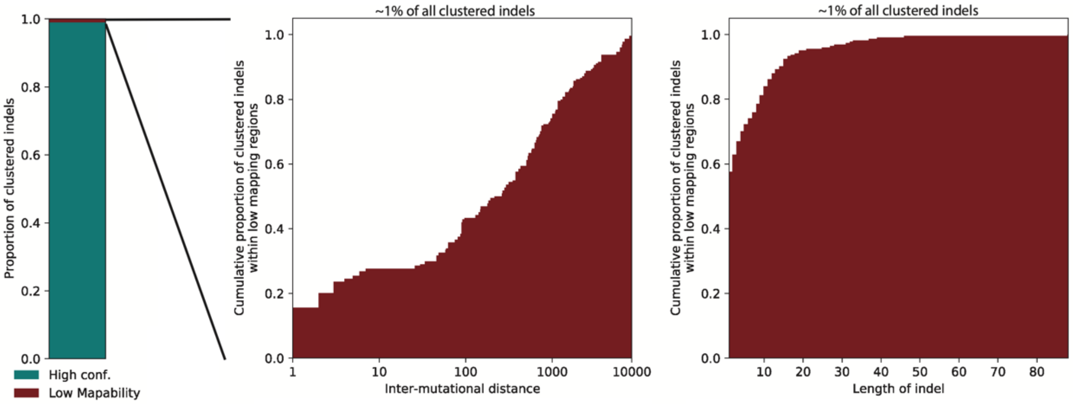


Figure 5.17: The distribution of low confidence clustered indels. The number of clustered indels falling within regions of the genome with low mapping scores consists of approximately 1% of all clustered indels. Within these 1% of mutations with low mapping scores, only 30% of events have an inter-mutational distance less than 10 (0.3% of all clustered indels), while indels of lbp falling within low mapping regions comprise only 0.5% of all clustered indels.

5.3.3. Clustered mutational signatures analysis

The clustered mutational catalogues of the examined samples were summarized in SBS288 and ID83 matrices using SigProfilerMatrixGenerator [143] (v.1.2.0) for each tissue type and each category of clustered events. For example, six matrices were constructed for clustered mutations found in Breast-AdenoCA: one matrix for DBSs, one matrix for MBSs, one matrix for omikli, one matrix for kataegis, one matrix for other clustered substitutions and one matrix for clustered indels. The SBS288 classification considers the 5′ and 3′ bases immediately flanking each single-base substitution (referred to using the pyrimidine base in the Watson–Crick base pair) resulting in 96 individual mutation channels. In addition, this classification considers the strand orientation for mutations that occur within genic regions resulting in three possible categories: (1) transcribed; pyrimidine base occurs on the template strand; (2) untranscribed; pyrimidine base occurs on the coding strand; or (3) non-transcribed; pyrimidine base occurs in an intergenic region. Mutations in genic regions that are bidirectionally transcribed were evenly split amongst the coding and template strand channels. Combined, this results in a classification consisting of 288 mutation channels, which were used as input for de novo signature extraction of clustered substitutions. The ID83 mutational classification has previously been described [143].

Mutational signatures were extracted from the generated matrices using SigProfilerExtractor (v.1.1.0), a Python-based tool that uses non-negative matrix factorization to decipher both the number of operative processes within a given cohort and the relative activities of each process within each sample [159]. The algorithm was initialized using random initialization and by applying multiplicative updates using the Kullback–Leibler divergence with 500 replicates. Each de novo extracted mutational signature was subsequently decomposed into the COSMIC (v.3) set of signatures (https://cancer.sanger.ac.uk/signatures/) requiring a minimum cosine similarity of 0.80 for all reconstructed signatures. All de novo extractions and subsequent decomposition were visually inspected and, as previously done1, manual corrections were performed for 2.2% of extractions (4 out of 180 extractions) in which the total number of operative signatures was adjusted ±1. Consistent with prior visualizations [60], we have included all cancer types within the PCAWG cohort, which may comprise as few as one sample for certain cancer types. Similarly, consistent with prior visualizations1, decomposed signature activity plots required that each cancer type have more than 2 samples and used mutation thresholds for each clustered category; 25 mutations per sample were required for DBSs, omikli events and other clustered mutations; 15 mutations per sample were required for MBSs and kataegic events; and 10 mutations were required per sample for clustered indels.

5.3.4. Experimental validation

A subset of clustered mutational signatures was validated using previously sequenced in vitro cell line models. As done for PCAWG samples, we generated a background model using SigProfilerSimulator [158] to calculate the clustered IMD cut-off for each sample and partitioned each substitution into the appropriate category of clustered events. Mutational spectra were generated for each subclass within each sample using SigProfilerMatrixGenerator [143] and were compared against the de novo signatures extracted from human cancer. The cosine similarity between the in vitro mutational spectra and de novo observed clustered signatures was calculated to assess the degree of similarity. The average cosine similarity between two random non-negative vectors is 0.75, and the cosine similarities above 0.81 reflect P values below 0.01 (ref. [158]).

5.3.5. Associations with cancer risk factors

Homologous recombination (HR) deficiency was defined for breast cancers using the status of BRCA1, BRCA2, RAD51C and PALB2 [47]. Samples with a germline, somatic or epigenetic alteration in one of these genes were considered HR-deficient, whereas samples without any known alterations in these genes were considered HR-proficient. The number of clustered indels was compared between HR-deficient and HR-proficient samples. The smoking status of lung cancers was determined using the clinical annotation from TCGA (https://portal.gdc.cancer.gov/repository). The number of clustered indels associated with tobacco smoking (ID6) was compared between samples annotated as lifelong non-smokers and samples annotated as current and reformed smokers. The status of alcohol consumption was determined using the annotations from the official PCAWG release (https://dcc.icgc.org/releases/PCAWG). The total number of clustered indels was compared in samples annotated with no alcohol consumption and those annotated as daily and weekly drinkers.

5.3.6. Expression of driver genes

All RNA-seq expression data were downloaded as a part of the official PCAWG release (https://dcc.icgc.org/releases/PCAWG). The relative expression data found within this release were normalized using FPKM normalization and upper quartile normalization. The relative expression of a gene was compared between those containing clustered or non-clustered events. Each distribution was then normalized to the average expression of the wild-type gene. Only genes with at least 10 total events (that is, clustered and non-clustered mutations) including at least 5 clustered events were considered for examination.

5.3.7. SVs and clustered events

The distance to the nearest structural variation breakpoint was calculated for each mutation in each subclass using the minimum distance to the nearest adjacent upstream or downstream breakpoint. Each distribution was modelled using a Gaussian mixture with an automatic selection criterion for the number of components ranging between one and five components using the minimum Bayesian information criteria (BIC) across all iterations. Modelling of kataegic events resulted in an optimal fit of three components, which was used to separate kataegic substitutions into SV-associated and non-SV associated mutations. DBSs and MBSs were both modelled using a single Gaussian distribution relating to non-SV associated mutations, whereas omikli and other clustered mutations were modelled using a mixture of two components, probably reflecting leakage of smaller kataegic events contributing to a weak SV-associated distribution. To account for the frequency of breakpoints across each sample, we normalized the minimum distance of each mutation to the nearest SV by calculating the expected distance between a mutation and SV for each sample using the total number of breakpoints and the overall length of a given chromosome (Extended Data Fig. 9a, b). After normalizing the kataegic events, we observed an optimal solution of two components with one SV-associated distribution (on average each mutation occurs within one-thousandth of the expected distance to nearest structural variation) and one non-SV associated distribution (on average occurring within the expected distance to the nearest structural variation). The normalized kyklonic events are consistent with the non-SV associated distribution reflecting kataegic events that occur on ecDNA typically of lengths 1–10 Mb (ref. [165]).

5.3.8. APOBEC3A and APOBEC3B enrichment analysis

The enrichment score of RTCA and YTCA penta-nucleotides quantifies the frequency for which each TpCpA>TpKpA mutation occurs at either an RTCA or a YTCA context. To account for motif availability, this score is calculated using the ±20 bp sequence context around each mutation and normalized by the number of cytosine bases and C>N mutations within the set of 41-mers surrounding each mutation of interest [59].

5.3.9. APOBEC3 gene expression and kyklonas

All RNA-seq expression data were downloaded as a part of the official PCAWG release (https://dcc.icgc.org/releases/PCAWG). The relative expression data found within this release were normalized using FPKM normalization and upper quartile normalization. The APOBEC3A/B normalized expression was compared between samples containing ecDNA versus samples with no detected ecDNA and between samples with kyklonas and without kyklonas. All P values were generated using a Mann–Whitney U-test and were corrected for multiple hypothesis testing using the Benjamini–Hochberg FDR procedure.

5.3.10. Circular ecDNA and kataegis

The collection of ecDNA ranges was intersected with the catalogue of clustered mutations, which was used to determine the overlapped mutational burden for each subclass of clustered event and the mutational spectra of overlapping kataegic events. Enrichments of events were calculated using statistical background models generated using SigProfilerSimulator [158] that shuffled the dominant mutation in each clustered event across the genome (that is, the most frequent mutation type in a single event). The decomposed kyklonic mutational spectra were generated using the decomposition module within SigProfilerExtractor [159]. Only mutational signatures that increased the overall cosine similarity by at least 0.01 were used. In both the original and validation cohorts, SBS2 and SBS13 were sufficient to explain the kyklonic mutational spectra with no other known mutational signature increasing the cosine similarity by more than 0.01. Comparisons between ecDNA with and without cancer genes were performed using the set of cancer genes from the Cancer Gene Census (CGC) [182]. All statistical comparisons and P values were calculated using a two-tailed Mann–Whitney U-test unless otherwise specified. For each set of tests, P values were corrected for multiple hypothesis testing using the Benjamini–Hochberg FDR procedure. The predicted effect of each overlapping variant was determined using ENSEMBL's Variant Effect Predictor tool by reporting only the most severe consequence [147].

5.3.11. Overall survival and clustered mutations

All survival analyses, including the generation of Kaplan–Meier curves, Cox regressions and log-rank tests, were performed using the Lifelines Python package (v.0.24.4). Across the 30 distinct whole-genome-sequenced cancer types included in the PCAWG study, only 6 cancer types contained enough samples to examine the associations between survival and overall number of clustered mutations. The sufficient sample size criteria required more than 50 samples with survival end-points with at least 30 of the samples with an observed clustered event. Each cancer type was analysed separately by comparing the survival of samples with a high clustered mutational burden (top 80th percentile across a given cancer type) to the survival of samples with a low clustered mutational burden (bottom 20th percentile across a given cancer type).

Analysis of whole-exome-sequenced samples from TCGA was altered to reflect the limited resolution for identifying clustered mutations within the exome. Specifically, SigProfilerSimulator (v.1.0.2) [158] was used to derive an IMD cut-off for each sample based on the TMB within the exome and the mutational patterns for a given sample. Mutations were randomly shuffled while maintaining the mutational burden within the exome of each chromosome, the ±2 bp sequence context for each mutation and the transcriptional strand bias ratios across all mutations. Each sample was simulated 100 times and an IMD cut-off was calculated using the same methods as outlined for the detection of clustered events within PCAWG. Owing to the limited number of detected events, 22 cancer types had sufficient data to perform survival analysis. Each cancer type was analysed separately by comparing samples with at least a single clustered event to samples with no detected clustered events within the exome.

For both PCAWG and TCGA analyses, survival distributions within a given cancer type were compared using a log-rank test. Cox regressions were performed to determine hazards ratios and to correct for age and total mutational burden. All P values were also corrected for multiple hypothesis testing using the Benjamini–Hochberg FDR procedure.

To investigate differential survival associated with the detection of clustered events within cancer driver genes, Kaplan–Meier survival curves were compared between individuals with clustered versus non-clustered mutations within a given cancer driver gene. The distributions were compared using a log-rank test. Cox regressions were performed to determine the hazards ratios and to correct for age, total mutational burden and cancer type across TCGA. Cox regressions performed for the MSK-IMPACT cohort were corrected for total mutational burden and cancer type. No corrections were performed for age as these metadata were not available for the MSK-IMPACT cohort. All P values were also corrected for multiple hypothesis testing using the Benjamini–Hochberg FDR procedure.

5.3.12. Validation of kyklonas in three cohorts

All three validation cohorts were analysed analogous to the PCAWG cohorts. Specifically, clustered mutations were classified by calculating a sample-dependent IMD threshold for clustered versus non-clustered mutations using a background model generated by SigProfilerSimulator [158]. All clustered mutations were subclassified into DBS, MBS, omikli, kataegis or other mutations. AmpliconArchitect (v.1.2) was used to determine regions of focal amplifications [183], which were used for subsequent validation of kyklonic events by overlapping kataegic events with all detected focal amplifications. The decomposed kyklonic mutational spectra were generated using the decomposition module within SigProfilerExtractor [159]. Only mutational signatures that increased the overall cosine similarity by at least 0.01 were used. In both the original and validation cohorts, SBS2 and SBS13 were sufficient to explain the kyklonic mutational spectra with no other known mutational signature increasing the cosine similarity by more than 0.01.

5.3.13. Cancer-type abbreviations

Biliary-AdenoCA, biliary adenocarcinoma; Bladder-TCC, bladder transitional cell carcinoma; Bone-Epith, bone epithelioid; Bone-Leiomyo, bone leiomyosarcoma; Bone-Osteosarc, bone osteosarcoma; Breast-AdenoCA, breast adenocarcinoma; Breast-LobularCA, breast lobular carcinoma; CNS-GBM, glioblastoma (central nervous system); CNS-Medullo, medulloblastoma (central nervous system); CNS-Oligo, oligodendroglioma (central nervous system); CNS-PiloAstro, pilocytic astrocytoma (central nervous system); Cervix-AdenoCA, cervix adenocarcinoma; Cervix-SCC, cervix squamous cell carcinoma; ColoRect-AdenoCA, colorectal adenocarcinoma; Head-SCC, head and neck squamous cell carcinoma; Kidney-ChRCC, chromophobe renal cell carcinoma; Kidney-RCC, renal cell carcinoma; Liver-HCC, hepatocellular carcinoma; Lung-AdenoCA, lung adenocarcinoma; Lung-SCC, lung squamous cell carcinoma; Lymph-BNHL, B-cell non-Hodgkin lymphoma; Lymph-CLL, chronic lymphocytic leukaemia; Lymph-NOS, metastatic lymphoma; Myeloid-AML, acute myeloid leukaemia; Myeloid-MPN, myeloproliferative neoplasm; Oeso-AdenoCA, oesophageal adenocarcinoma; Ovary-AdenoCA, ovary adenocarcinoma; Panc-AdenoCA, pancreatic adenocarcinoma; Panc-Endocrine, pancreatic neuroendocrine carcinoma; Prost-AdenoCA, prostate adenocarcinoma; Skin-Melanoma, malignant melanoma; Stomach-AdenoCA, stomach adenocarcinoma; Thy-AdenoCA, thyroid adenocarcinoma; Uterus-AdenoCA, uterine adenocarcinoma.

5.4. Data availability

No data were generated specifically for this study. All data were and can be downloaded from the appropriate links, repositories and references. Specifically, for the discovery cohort, all data and metadata were obtained from the official PCAWG release (https://dcc.icgc.org/releases/PCAWG). All data and metadata for TCGA samples were obtained from the GDC (https://gdc.cancer.gov/). Genomics data for clonally expanded cell lines were downloaded from the European Genome-phenome Archive (EGAD00001004201, EGAD00001004203 and EGAD00001004583). For the three validation cohorts, datasets were downloaded as submitted by the original publications and genomics data were downloaded from their respective repositories: EGAD00001004162 for 61 undifferentiated sarcomas [136] (European Genome-phenome Archive); EGAD00001006868 for 187 high-confidence oesophageal squamous cell carcinomas [176] (European Genome-phenome Archive); and phs001697.v1.p1 for 280 lung adenocarcinomas [175] (dbGaP). Somatic mutations and metadata for the MSK-IMPACT Clinical Sequencing Cohort composed of 10,000 clinical cases [173] were downloaded from cBioPortal (https://www.cbioportal.org/study/summary?id=msk_impact_2017).

5.5. Code availability

The SigProfiler compendium of tools are developed as Python packages and are freely available for installation through PyPI or directly through GitHub (https://github.com/AlexandrovLab/). For all tools, each package is fully functional, free and open sourced distributed under the permissive 2-Clause BSD License and is accompanied by extensive documentation: (1) SigProfilerMatrixGenerator [143] (v.1.2.0; https://github.com/AlexandrovLab/SigProfilerMatrixGenerator); (2) SigProfilerSimulator [158] (v.1.0.2; https://github.com/AlexandrovLab/SigProfilerSimulator); and (3) SigProfilerExtractor [159] (v.1.1.0; https://github.com/AlexandrovLab/SigProfilerExtractor). Each SigProfiler tool also has an R wrapper available for installation through the GitHub repositories. AmpliconArchitect [166] (v.1.2) is also freely available and can downloaded from https://github.com/virajbdeshpande/AmpliconArchitect. The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v.3.0, which allows for reuse and distribution.

5.6. Acknowledgements

Chapter 5, in full, is a reprint of the material as it appears in Nature 2022. Bergstrom, Erik N.; Luebeck, Jens; Petljak, Mia; Khandekar, Azhar; Barnes, Mark; Zhang, Tongwu; Steele, Christopher D.; Pillay, Nischalan; Landi, Maria Teresa; Bafna, Vineet; Mischel, Paul S.; Harris, Reuben S.; Alexandrov, Ludmil B., Springer Nature, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 6.

Deep learning predicts response to platinum chemotherapy in breast and ovarian cancers

6.1. Main

The advent of next-generation sequencing technologies resulted in an explosion of cancer genomics data, which provided unprecedented opportunities to interrogate the underlying molecular landscape of cancer [5, 24, 31, 60, 140, 167, 175, 176, 184-191]. The expansive exploration of these large consortiums such as the International Cancer Genome Consortium (ICGC) [60] and The Cancer Genome Atlas (TCGA) [167] led to key discoveries that provided insight into preventative strategies and potential therapeutic interventions when approaching cancer diagnosis and treatment. One such example includes the discovery of hereditary predispositions to breast and ovarian cancer attributed to germline mutations in the family of BRCA genes [118, 119, 192-194], which are involved in the repair of double-stranded DNA breaks through homologous recombination repair (HR) [195]. Complete inactivation of these genes results in deficiency of HR repair (HRD) ultimately leading to increased genomic instability [196]. Further investigations using large-scale cohorts of cancer genomes revealed unique mutational signatures of HRD which can be found in both germline and sporadic breast cancers in addition to other cancer types including ovarian, pancreatic, and prostate cancer, amongst others [4, 5, 24, 47, 102, 103]. From a clinical perspective, patients harboring these genomic scars of HRD are sensitive to certain therapeutics such as platinum chemotherapies and PARP inhibitors which increase the demand on double-stranded DNA break repair leading to selective tumor cell death [120, 122, 123].

As a result of these seminal studies, a plethora of methods have been developed to detect the “BRCAness” HRD phenotype with the goal of treating a wider range of individuals beyond only those harboring germline predispositions [102, 103, 197-199]. Several of these approaches show potential for direct translation into the clinic; however, they rely heavily on whole-genome or whole-exome sequencing for accurate predictions, which remains largely unavailable in most clinical settings across the world [200, 201]. To circumvent these limitations, additional methods are being developed to predict HRD using targeted sequencing panels but are still reliant on genomic sequencing and are rudimentary in deployment to the clinic [199]. Further, this dependence on sequencing is thought to be exacerbating known racial disparities in health care [202, 203].

In fact, recent reports reveal that the incidence of breast cancer is similar between White and Black populations in the United States; however, Black women are more likely to die of breast cancer than any other ethnic group [204, 205]. This racial disparity is believed to be linked to several risk factors including genetic and molecular differences in the cancer, average age of screening and diagnosis, socioeconomic status, and differential accessibility of adequate health care associated with geographical and financial barriers [206, 207]. From a research perspective, there has been systemic omission of under-served populations during the accumulation of large-scale genomics and medical data [202, 208, 209], which has compounded the racial biases of downstream algorithm development [210, 211]. Modern consortiums are seeking to alleviate some of these disparities; however, these studies are centered around sequencing and molecular diagnostics, which remain as a bottleneck both geographically and financially for many populations [212-216].

While the availability of sequencing-based diagnostics and personalized treatment regimens are limited in accessibility, tissue biopsies are routinely sampled for cancer diagnostics. In combination with recent advances in computer vision, artificial intelligence-based deep learning models allow for both prognostic and diagnostic predictions using only histopathological tissue slides [217]. Here we demonstrate the ability to detect HRD directly from digitalized hematoxylin and eosin (H&E) tissue slides with direct implications of addressing these socioeconomic disparities in the diagnoses of breast and ovarian cancers.

To train a downstream model capable of predicting HR status directly from digitalized tissues slides, we implemented a weakly supervised convolutional neural network architecture based upon the fundamental assumptions of multiple-instance learning (MIL; Figure 6.1) [218]. Specifically, all partitioned regions of a slide, or tiles, within a whole-slide image (WSI) are assigned a weak label based upon the slide-level classification for each sample. It is assumed that all tiles within a negatively labeled slide are HR-proficient (HRP), whereas at least a single tile must exhibit an HRD phenotype within a positively labeled slide. These assumptions allow the model to be trained using only a single classification label for an entire image or patient without the need for detailed manual annotations from a pathologist, which currently do not exist for characterizing HRD. HRD scores were calculated using the combined aggregated score of the telomeric allelic imbalance score (TAI), loss of heterozygosity score (LOH), and large-scale transitions score (LST) for each patient (Methods) [219]. Traditionally, an HRD score greater than 42 was used to determine eligibility for treatment with PARP in triple negative breast cancers to ensure a sensitivity of greater than 95% [219]. For our ground truth, we incorporate soft labelling during training to prevent the model from becoming overconfident with a single image by centering the HRD score cutoff at the median score across all breast cancer samples (HRD=30). All samples with an HRD score above 50 are considered deficient, while all samples with an HRD score below 20 are considered proficient. The remaining samples with an HRD score between 20 and 50 are modelled based upon a probability distribution centered at 30 where there is an equal probability of being deficient or proficient (Methods).

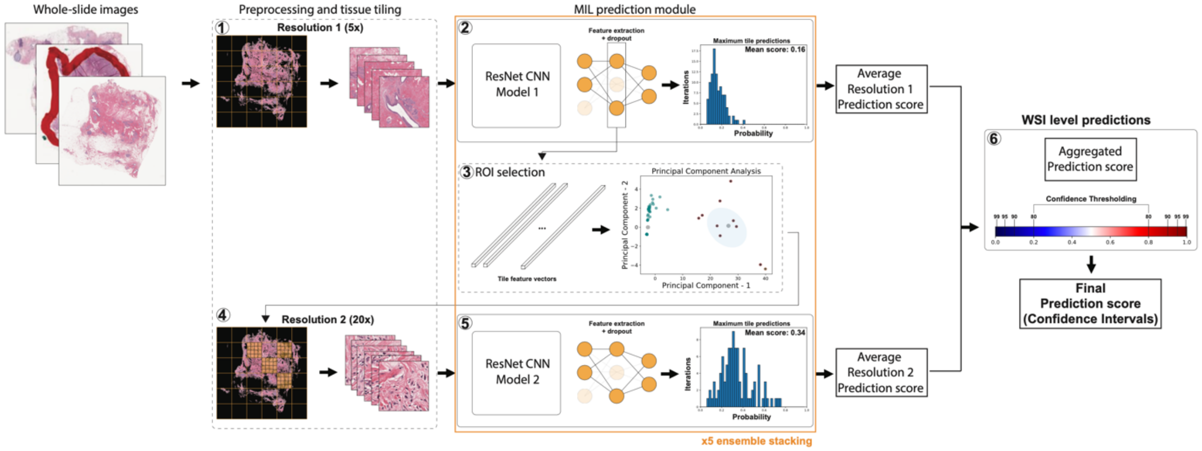


Figure 6.1: Multi-resolution convolutional neural network architecture to detect homologous recombination deficiency from histopathological tissue slides. For each whole-slide image (WSI), a single prediction score is estimated based on the detection of HRD. Specifically, each WSI undergoes preprocessing and quality control (1). This module consists of tissue segmentation, filtering for non-focused tissue, and final tiling of regions that contain tissue at a set resolution (i.e. 5x magnification). All tiles for a single image are processed through the first multiple instance learning (MIL) ResNet18 convolutional neural network (2). This architecture uses the average of the top 25 predicted tile scores as the WSI predicted score. Dropout is incorporated into the fully connected layers in the feature extraction module to reduce overfitting during training. The same dropout technique is also incorporated during inference to simulate Monte Carlo dropout used to calculate confidence intervals in the final WSI prediction. The tile feature vectors from the penultimate layer of the feature extraction are used to automatically select regions of interest (ROI) from the original WSI for additional assessment (3). The feature vectors are reduced in dimensions using principal component analysis and a custom k-means clustering module determines the optimal number of clusters per sample. The selected tiles are then resampled at a higher magnification (i.e. 20x; 4). These sets of tiles are used to train a second MIL-ResNet18 model (5) using an identical architecture to the one used previously in (2). The average predictions across both models are aggregated for a single WSI (6). The resulting distribution of scores are used to calculate confidence intervals and establish a threshold of confidence for a final prediction.

Our proposed model is composed of a multi-resolution decision, which performs an initial prediction on a low magnification (i.e., 5x magnification) of localized regions of interest (ROI) that are automatically selected and performs a secondary prediction on an enhanced magnification within the selected ROIs (i.e., 20x magnification; Figure 6.1). The final model encompasses an ensemble of five identical architectures, which each produce multi-resolution prediction scores. The average of these scores is used to make a final prediction for each tissue slide. Due to the computational cost associated with processing an entire WSI, each slide is first segmented into smaller tiles at a given resolution. For the first stage of the model, each slide is tiled at a 5x magnification with 256x256 pixels per tile with approximately 2µm of tissue per pixel. Blurred tiles and those with less than 80% of pixels containing tissue were removed from the analysis (Figure 6.1; Methods). For both stages of the model, ResNet18 convolutional neural networks were trained to extract features from the collection of tiles composing a single WSI. The resulting encoded features from the penultimate fully connected layer were used to automatically select ROIs at the 5x resolution. Specifically, Principal Component Analysis (PCA) was used to project the encoded features into a latent space encompassing the greatest variance. K-means clustering was then used to group each tile representation. The cluster containing the tile with the maximum prediction probability was selected along with all tiles in the same cluster having a Silhouette coefficient greater than the 95% quantile of all Silhouette scores across the WSI (Methods). The final ROIs were tiled at 20x magnification (0.5µm per pixel) and used to train and test the second model. The top 25 tiles were averaged to calculate a final prediction score at a given resolution during an inference pass of a WSI.

During training, random dropout of nodes within the fully connected layers of the ResNet architecture was incorporated to prevent overfitting of the training dataset. This same dropout technique was applied during inference of each WSI, known as Monte Carlo dropout, to provide an estimation of the model uncertainty by performing multiple inference passes of a single WSI [220]. The resulting distribution of predictions were averaged to calculate a final score encompassing any epistemic uncertainty and were used to calculate confidence thresholds for a given sample (Figure 6.1; Methods).

Specifically, we trained a multi-resolution model to detect HRD using the collection of flash frozen tissue slides from the TCGA breast cancer cohort using 85% of the samples for training and the remaining 15% of samples to internally validate the final models (n=1,055 samples; Figure 6.2a). Prior to training, the number of HRD and HRP samples in each cancer subtype were balanced to prevent the models from learning features specific to individual subtype histology rather than those directly associated with HRD. During training, a validation set composed of 15% of the samples was used to select the final models and to adjust the prediction threshold. The final models were then tested on the held-out internal test set to assess the overall performance resulting in an area under the receiving operating curve (AUROC or AUC) of 0.81 ([0.77-0.85] 95% CI; Figure 6.3a).

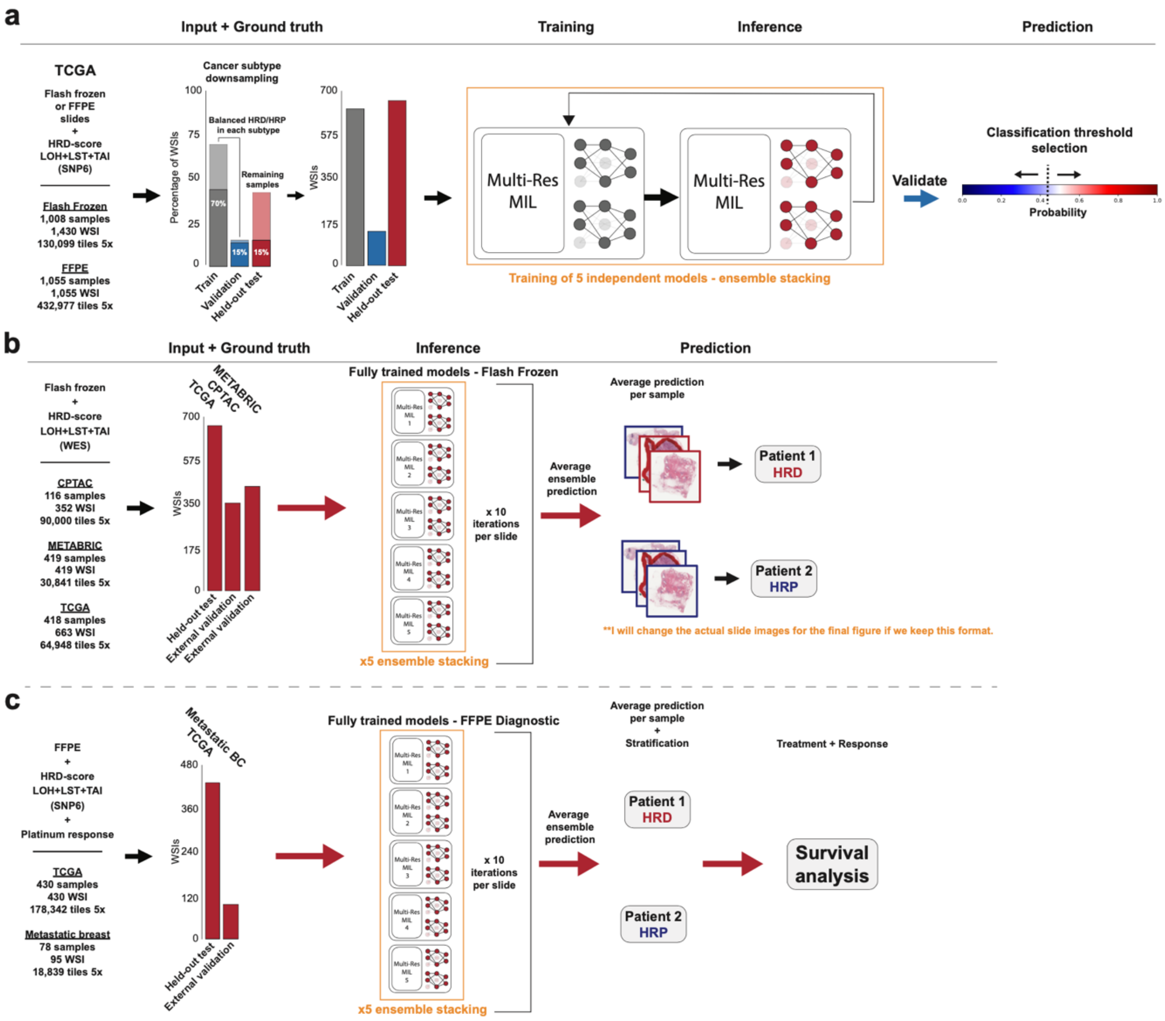


Figure 6.2: Training and testing workflow. **a.** The collection of breast cancer tissue slides comprised of either flash frozen or FFPE-derived images from TCGA were used to train the multi-resolution model. HRD scores are calculated from SNP6 sequencing data and used as the ground truth. The number of HRD/HRP samples within each breast cancer subtype were balanced within the training and validation set. The randomly under-sampled images were added to the held-out test set. **b.** Testing of the trained model was performed using the held-out TCGA samples along with two external validation cohorts. **c**. The final model was used to predict metastatic breast cancer responders (HRD) from non-responders (HRP) after treatment with platinum chemotherapy. The final prediction probabilities were used to stratify HRD from HRP patients and were used for downstream survival analysis.

To assess the generalizability of the model, we performed an external validation using the collection of breast cancer slides from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [221] and the Molecular Taxonomy of Breast Cancer International Consortium

(METABRIC; Figure 6.2b) [222] resulting in an AUC of 0.76 ([0.71-0.82] 95% CI; Figure 6.3a). Using the aggregated prediction score from the Monte Carlo dropout, we can remove samples that the model was less confident in. For instance, after removing all predictions that fell within 5% of the classification cutoff retained 72% of samples in TCGA resulting in an AUC of 0.83 ([0.78-0.88] 95% CI; Figure 6.3b). Applying this thresholding to CPTAC retained 59% of samples with an AUC of 0.83 ([0.76-0.89] 95% CI). This threshold can be made more stringent to provide confident predictions for a subset of patients. Specifically, removing predictions that fell within 10% of the prediction cutoff retained 42% of samples in TCGA with an AUC of 0.86 ([0.80-0.92] 95% CI) and 36% of samples in CPTAC with an AUC of 0.86 ([0.77-0.93] 95% CI; Figure 6.3c). Further, HRD is enriched in luminal B, basal-like, and Her2 enriched breast cancers; however, our model distinguished HR deficiency and proficiency across all subtypes (Figure 6.3d).

While flash frozen tissue slides are commonly used for downstream molecular analyses, formalin-fixed paraffin-embedded (FFPE) tissue slides are the standard for clinical diagnostics. Therefore, we trained an independent model to classify HRD directly from FFPE slides from the TCGA breast cancer cohort following the same procedure as previously described for training on flash frozen tissue images (Figure 6.2a&c; Methods). The final base model resulted in an AUC of 0.81 ([0.77-0.85] 95% CI; Figure 6.3e). Removing samples with a prediction that fell within 5% of the classification threshold retained 72% of samples with an AUC of 0.83 ([0.77-0.85] 95% CI; Figure 6.3c). These results indicate that the fixation procedure has minimal effect on the performance of predicting HRD status directly from breast cancer tissue slides. Further, the FFPE model was capable of distinguishing metastatic breast cancer (MBC) samples that had a complete response to platinum chemotherapy from those having only a partial or no response to treatment with an AUC of 0.76 ([0.57-0.93] 95% CI; Figure 6.3e) [223]. Separating the MBC samples treated with platinum based upon the model prediction probabilities reveal a difference in survival between the HRD and HRP-predicted samples with a median survival of 23.4 months for HRD patients and 9.4 months for HRP patients (p-value=0.023, Log-rank test; Figure 6.3e). The model's predictive value was consistent after correcting for the subtype and age of each cancer with a hazard ratio of 0.51 ([0.27-0.94] 95% CI, q-value=0.030; Cox Proportional Hazards regression).

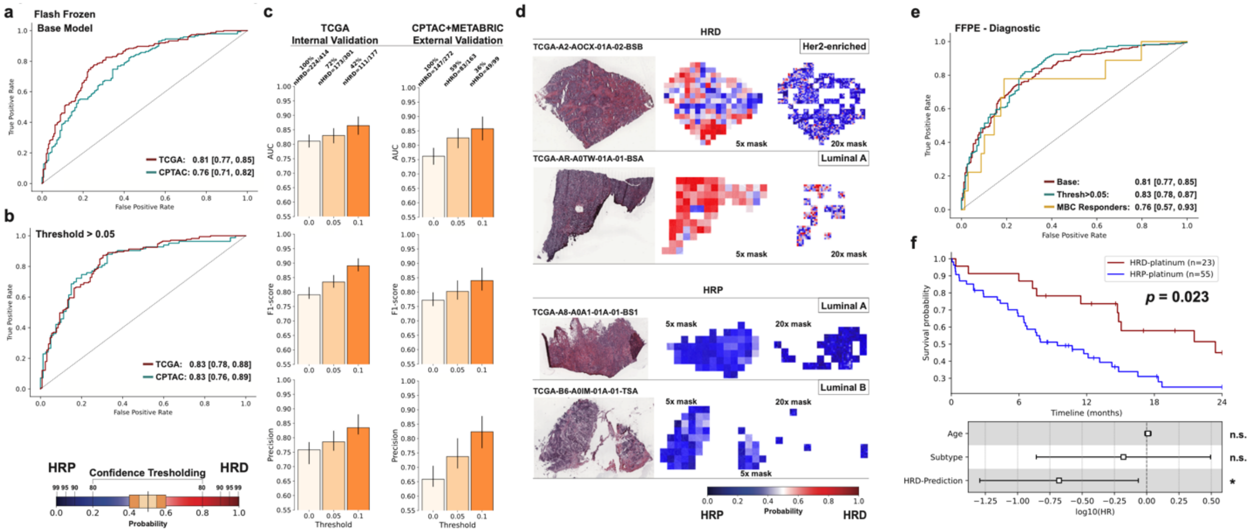


Figure 6.3: Performance of detecting HRD from breast cancer slides. **a**) The receiver operating characteristic curve (ROC) for classifying HRD in the TCGA test set, the CPTAC cohort, and the METABRIC cohort using the base model. **b**) ROC after removing the raw prediction probabilities that were within 5% of the classification threshold. **c**) The area under the ROC (AUC), F1-scores, and precision across the TCGA test set and CPTAC and METABRIC cohorts using varying confidence thresholds. Standard error bars are shown along with the number of HRD samples included after each filtering threshold (nHRD). **d**) Representative TCGA tissue slides are shown for both HRD and HRP samples across multiple breast cancer subtypes along with the resulting predictions for each segmented tile at the 5x and 20x stages of the model. **e**) ROC for the FFPE diagnostic base model, after removing samples within 5% of the threshold, and for classifying metastatic breast cancer responders (MBC). Bootstrapped 95% confidence intervals are provided for all ROC curves. **f**) Kaplan-Meier survival curves for MBC patients treated with platinum chemotherapy separated by the HRD model predictions (top). Cox regression showing the log10-transformed hazards ratios are shown with their 95% confidence intervals (bottom).

To assess the generalizability of the proposed method in application to other cancers, we performed transfer learning on the TCGA ovarian cancer cohort (n=589 samples). Previous studies show that patients with ovarian cancers have the highest median HRD score and those with a score greater than 63 have a better overall prognosis [224]. Further, individuals with ovarian cancer have traditionally received platinum chemotherapies as the first-line standard of care making this cohort ideal to test whether HRD predictions from tissue slides may have a direct clinical benefit. Specifically, we trained an independent model to predict HRD status (HRD score >63) from flash frozen slides using the TCGA ovarian cohort. Due to a smaller cohort size, the ovarian model was initiated using the pretrained weights and biases generated from the breast cancer model with the convolutional weights and biases frozen during training (Figure 6.4a). The final model was applied to a held-out test set of TCGA ovarian cancers to assess the ability of the model in separating individuals who benefit from treatment with platinum chemotherapy (Figure 6.4a). Of the 117 patients in the test set, 74 had a record of receiving platinum chemotherapy. Separating these individuals by their multi-resolution HRD prediction resulted in a differential median survival between HRD and HRP predicted patients (Figure 6.4b). Specifically, patients predicted to be HRD had a median survival of 4.6 years, while those predicted to be HRP had a median survival of 3.2 years (q-value=0.034) with a hazard ratio of 0.49 after correcting for the stage of the cancer, age, and the HRD score ([.25-0.96] 95% CI; Cox Proportional Hazards ratio; Figure 6.4b).

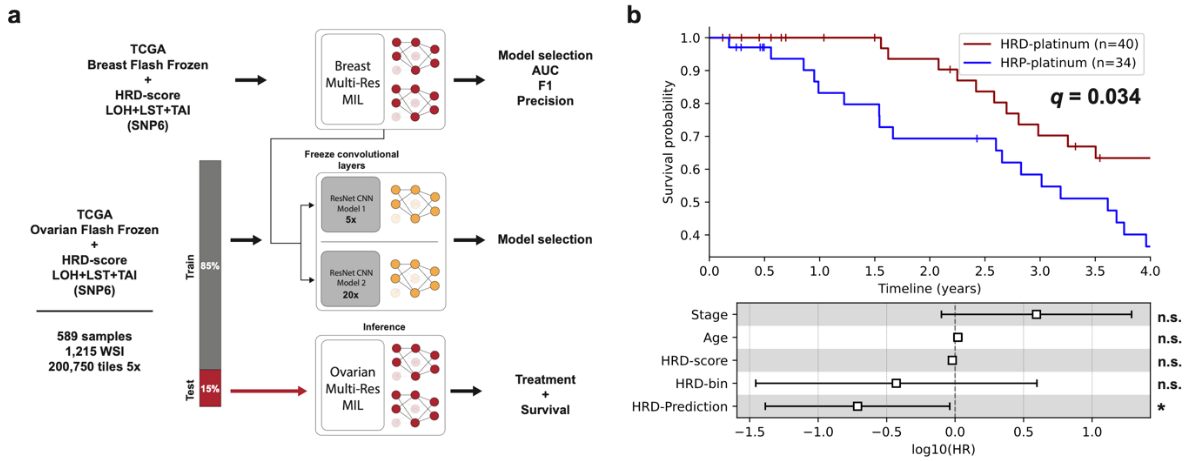


Figure 6.4: Transfer learning in ovarian cancer predicts response to treatment. **a**) Schematic demonstrating the transfer learning method to train an ovarian HRD model using a pretrained breast HRD model. The pretrained breast models are used to initiate the weights and biases of all parameters in the ovarian model. **b**) Kaplan-meier survival curves comparing the outcome of patients treated with platinum chemotherapy split by the HRD model prediction. Q-value is corrected for after considering stage, age, the HRD score (HRD-score), and the binary HRD classification score >63 (HRD-bin; *q<0.01). The log10-transformed hazards ratios (log10(HR)) are shown with their 95% confidence intervals.

The development of these HRD prediction models on both breast and ovarian cancers demonstrates the practicality of employing artificial intelligence-based guidance into clinical diagnostics. The models are generalizable across different cancers, subtypes, and tissue fixation procedures using routinely sampled tissue blocks and can predict overall patient outcome after treatment with chemotherapeutics. Further, each model is adjustable to ensure a specific level of confidence in the final patient prediction, ultimately circumventing the reliance on additional sequencing information when diagnosing HRD in ~75% of patients. Removing the sequencing bottleneck traditionally used for calculating an HRD score allows for this method to be more readily deployable into the clinic and provides greater accessibility to the standard of care for a larger proportion of the population across a diverse collection of socioeconomic groups.

6.2. Methods

6.2.1. Data sources

The collection of flash frozen and FFPE slides from TCGA along with all clinical features were downloaded from the Genomic Data Commons (GDC; https://gdc.cancer.gov/). The collection of flash frozen slides from CPTAC were downloaded from The Cancer Imaging Archive (TCIA) [225], and the genomics data was downloaded from the GDC. The collection of images from METABRIC and the associated SNP6 data were downloaded from EGA (EGAD00010000270 and EGAD00010000266). The predicted cancer subtype for a subset of the TCGA breast cancer cohort were obtained from the previous study using the 50-gene PAM50 model [226]. HRD scores for the TCGA breast and ovarian cancers were obtained from the previous study [102]. The slides and clinical data for the metastatic breast cancers were obtained from SRA repository under BioProject accession number PRJNA793752 [223].

6.2.2. Data preprocessing

Each WSI was segmented into 256x256 tiles at 5x and 20x magnifications containing 2µm per pixel and 0.5µm per pixel, respectively. Blurry tiles and those with less than 80% of pixels representing tissue were removed from all training and testing cohorts. To filter blurry tiles, a Laplacian filter was applied to each tile using a 3x3 kernel, and all tiles with a variance less than 0.02 were removed from the remaining analysis. All green, red, and blue pen marks and other annotation artifacts were removed by thresholding on the RGB color channels within each pixel.

6.2.3. Calculating HRD scores

HRD scores were calculated as previously reported [102] using scarHRD. Specifically, the HRD score is the summation of the loss of heterozygosity (LOH) [227], the telomeric allelic imbalance (TAI) [228], and the large-scale state transitions (LST) [229] scores calculated using copy number calls derived from SNP6 arrays using ASCAT. The HRD scores for the CPTAC breast cancer samples were calculated based on copy number calls derived from whole-exome sequencing using Sequenza [230], which has been shown to result in analogous distributions of HRD scores. HRD scores above 50 were considered HR-deficient and scores below 20 were considered proficient in the breast cancer cohorts. All intermediate scores were modelled as a probability of being deficient or proficient with an equal probability of both conditions at an HRD score of 30. Within the TCGA ovarian cohort, HRD scores about 72 were considered deficient and scores below 52 were considered proficient with the intermediate probabilities centered at 63.

6.2.4. Model training and testing

Prior to training, the number of HRD and HRP samples were balanced in all breast cancer subtypes using the PAM50 model classifications to normalize for specific subtypes enriched and depleted of HRD. All samples without annotated PAM50 subtype labels were considered as missing and were also balanced for the number of HRD and HRP cases. Soft labelling was incorporated to prevent overfitting during training and to account for ambiguity in the ability of the HRD score to classify true HRD samples. The entirety of training and testing was performed using the machine learning Python framework Pytorch (v.1.5.0). For both resolution models, the Adam optimizer was used for training with a learning rate of 1E-03, a weight decay of 1E-04, and minibatches consisting of 64 tiles. Each model was initiated using the ResNet18 architecture that was pretrained on the ImageNet (http://www.image-net.org/) database and was trained for 200 epochs. All convolutional weights were frozen during training. Early stoppage was incorporated to prevent overfitting.

After training the 5x resolution models, a final inference pass is performed on all slides. All features from a single WSI were selected from the penultimate layer of the feature extractor and projected into a lower dimensional latent space using PCA. K-means clustering was used to automatically select ROIs for retiling at 20x magnification. The number of clusters was determined by selecting the solution with the maximum silhouette coefficient. The cluster containing the tile with the highest prediction probability was used to select the ROIs. All tiles belonging to this cluster, and which had a silhouette score greater than the 95% quantile of all silhouette scores for the given WSI were chosen as the final ROIs. Each ROI was then tiled into 256x256 pixel sub-tiles at 20x magnification. This results in 16 tiles at 20x magnification for each ROI at a 5x magnification. To perform an inference pass of the model, a single WSI image is processed across 10 iterations with a random dropout probability of 0.2 for all nodes within the fully-connected layers.

The weights collected from the final models trained to detect HRD from flash frozen breast slides were used to initiate the model weights for the ovarian model known as transfer learning. The held-out internal validation set was used to perform survival analysis based upon prior treatment with platinum chemotherapy. There were not enough FFPE slides for the ovarian cohort for training and testing (results not shown).

6.2.5. Survival analysis

Survival analysis was performed using the Lifelines Python package (v.0.24.4.). For both the metastatic breast cancer (MBC) and the TCGA ovarian cohorts, samples were partitioned based upon the prediction probabilities from each respective model using the HRD threshold cutoff that gave the highest F1-score in the internal validation sets. Only samples that were treated with platinum chemotherapy were considered in the survival comparisons. Survival curves were compared using a log-rank test. Hazards ratios were calculated from Cox regressions after correcting for age and subtype within the MBC cohort and stage, age, and HRD score within the TCGA ovarian cohort. Median survival was calculated as the time at which the chance of surviving beyond that point is 50%.

6.2.6. Statistics

All performance metrics including AUCs for each ROC curve, F1-scores, and precision metrics were calculated using the scikit-learn Python package (v.0.22.1). Confidence intervals were calculated using non-parametric resampling. Standard error bars were calculated using the Numpy Python package (v.1.18.1).

6.3. Acknowledgements

Data used in this publication were generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC). Chapter 6 contains unpublished material coauthored with Bergstrom, Erik N.; Diaz-Gay, Marcos; Abbasi, Ammal; Ladoire, Sylvain; and Alexandrov, Ludmil B. The dissertation author was the primary author of this chapter.

Chapter 7.

Conclusion and future work

In this dissertation I introduced three novel computational approaches to explore and decipher the mutational processes underlying the development of human cancer. These approaches seamlessly integrate providing a streamlined toolbox for reproducibility and large-scale deployment within epidemiological studies and within the clinic. Through the application of this standardized framework, I provide a comprehensive characterization of clustered somatic mutations across human cancer revealing a plethora of mutational processes giving rise to such events. Several of these processes result in clustered mutations that are enriched within known driver genes providing prognostic biomarkers for patient survival in certain cancer types. The extensive characterization of these processes in coordination with the localization of different focal amplifications, revealed a novel form of oncogenesis reflecting the repeated hypermutation of ecDNA by APOBEC3 deaminases.

While the analysis of mutational signatures provides insight into the underlying etiology of a given cancer, the universal application within the clinic is currently limited due to financial bottlenecks and due to the need for downstream expert analysis. To address these current limitations, this dissertation also proposed an alternative approach for detecting mutational signatures using artificial intelligence-based models applied to histopathological cancer slides that are routinely sampled for cancer samples within the clinic. The deep learning architecture was applied to breast and ovarian cancers demonstrating its ability to predict homologous recombination repair deficiency and ultimately differential patient response to platinum chemotherapy.

7.1. Implications

The findings of this thesis provide insights into fundamental cancer biology and biomarker discovery while illuminating broader implications to motivate future studies. First, the comprehensive mapping of clustered mutagenesis across human cancer revealed a cancer-dependent enrichment of clustered events implicating novel components of oncogenesis and tumor evolution. These findings demonstrate similarities between viral and ecDNA-driven cancers suggesting a direct role of the innate immune response in accelerating the evolution of some tumors.

Second, exploration of clustered biomarkers and alternative diagnostic platforms implicates that mutational signature detection can be directly incorporated into existing clinical infrastructures. Specifically, the analysis of clustered mutations within whole-genome sequenced tumors revealed novel biomarkers attributed to an enrichment of clustered mutational events within known driver genes. Extensive validation of several of these biomarkers across both whole-exome and targeted sequencing panels demonstrates the applicability of using these prognostic biomarkers within a clinical setting. Specifically, the targeted sequencing cohorts were retrospectively collected and generated using cancer gene panels that are still used in contemporary clinical settings for precision medicine. Thus, the detection of clustered biomarkers can be easily incorporated into existing clinical pipelines. Additionally, the ability to detect mutational signatures directly from histopathological slides requires only the digitalization of a given tissue block using either an internal or tertiary scanning service. The algorithm itself is light-weight and can be deployed with minimal computational resources.

Lastly, the adoption of these proposed artificial intelligence-based diagnostic platforms ideally removes the reliance on traditional sequencing technologies, which currently hinders broad accessibility to the standard of care and compounds disparities observed across healthcare and, especially, across cancer diagnostics. The elaboration and careful incorporation of these tools can introduce state-of-the-art diagnostics to communities that have been historically underserved due to both geographical and/or financial barriers.

7.2. Limitations

While this thesis provides a standardized framework for the efficient and cost-effective analysis of mutational signatures, there are limitations to the current approaches and their subsequent implications. Mainly, clustered mutations are ubiquitous across human cancer; however, they occur at low numbers in each sample. This restricts the ability to decipher the mutational signatures of clustered events to data derived from whole-genome sequenced cancers, which are not as plentiful as whole-exome or targeted sequencing cohorts. Until whole-genome sequencing becomes more readily accessible, the current benefits of analyzing clustered mutations within the clinic will be limited to detecting single events within known driver genes as prognostic or predictive biomarkers.

Further, the ability to expand on the oncogenic implications of hypermutated ecDNA is largely restricted by the dearth of detailed clinical annotations paired with whole-genome sequenced cancers. The pilot studied described in this thesis was unable to elaborate on the effects of observing *kyklonic* events on survival and patient outcome due to a lack of clinical endpoints for most cancer samples. These analyses were also limited by the current algorithmic approaches for detecting ecDNA, which also require whole-genome sequencing data. Significant efforts are ongoing to improve the stability of locating *bona fide* ecDNA regions, which are often missed due to the lack of concordance in copy number callers and ambiguity in the reconstructed substructure graphs associated with short-read sequencing. Ideally, long-read sequencing technologies should be used to refine the resolution at which focal amplifications are located and resolved into their complete structures. This would enhance the accuracy of detecting ecDNA and ultimately provide a higher resolution of the mutational landscape for downstream signature analysis.

Lastly, the development of alternative approaches for detecting mutational signatures presents a double-edged sword in terms of translational research. For example, deep learning technologies show promise in circumventing financially burdensome protocols and clinical testing; however, they are typically trained on retrospective datasets that lack patient diversity necessary for universal deployment. This issue stems from a systemic omission of underserved populations within previous large-scale sequencing efforts. Without careful considerations of the training framework and data inclusion criteria, technological advances can further exacerbate existing disparities in health care. Recent efforts from multiple consortia are now actively working to alleviate these disparities within medical records and public datasets; however, this will take time. For these reasons, it is essential to acknowledge the limitations and breadth of scope of existing algorithms.

7.3. Future work

Initial efforts outlined in this thesis revealed translational opportunities of leveraging the detection of mutational signatures and clustered events to further our understanding of cancer biology and to allow better clinical management of cancer patients. The proposed approaches were largely limited by the accessibility of large-scale datasets with extensive clinical annotations. Current efforts are being made to generate and collect additional better annotated cohorts that will allow further interrogation of specific cancer types enriched for clustered events and for detecting additional mutational signatures from digitalized tissue slides from cancer tissues.

7.3.1. Clinical integration of clustered mutations

The ability to detect clustered mutations within known cancer genes sequenced through standard-of-care targeted panels allows for the immediate translation of any clustered mutations that predict response to drug treatment or that provide prognostic information about a patient's overall survival. While this study utilized a large collection of samples that were sequenced with targeted cancer gene panels, the number of publicly available genomes are continuing to drastically increase. Since the publication of the initial study, the sequencing data for the cancers of over ~100,000 individual patients have become available including detailed survival annotations across most human cancers. The expansion of these cohorts will provide an opportunity to explore all forms of detectable clustered events with downstream prognostic value across the most common panels utilized in clinical settings.

Further, there is now an opportunity to investigate the diagnostic potential of clustered mutations in coordination with specific treatments. As an exploratory study, the Genomics of Drug Sensitivity in Cancer (GDSC) dataset can be utilized to determine whether key clustered events associate with better overall survival probability after treatment across different types of compounds. The GDSC contains detailed screening of 1,000 human cancer cell lines with over 600 different compounds that target 24 different molecular pathways.

7.3.2. Investigating the clinical implications of *kyklonas*

This thesis introduced a novel type of clustered hypermutation that occurs on ecDNA, coined *kyklonas*. While these events appear to be the direct attack of enzymes involved in the innate immune response driving an accelerated diversification of the ecDNA mutational landscape, our understanding of this process remains limited. With current efforts to generate and collect additional whole-genome sequenced cancers including cohorts of pediatric cancers that are enriched for ecDNA, there will be opportunities to further expand our understanding of the mechanistic underpinnings of *kyklonic* events. This will include investigating the evolutionary trajectory of individual ecDNA populations using the occurrence of distinct *kyklonic* events, the effect of recurrently mutated ecDNA on overall patient survival, and the activity of APOBEC3 deaminases in response to treatment and subsequent therapeutic resistance in the presence of ecDNA.

7.3.3. Detecting multiple mutational signatures from digital tissue slides

The demonstration of detecting homologous recombination deficiency within breast and ovarian cancer using the proposed deep learning architecture revealed the ability to predict individual patient sensitivity to treatment. While breast and ovarian cancers have traditionally been associated with homologous recombination deficiency, pan-cancer studies have revealed a prevalence of mutated HRD genes across many different cancers, including pancreatic and prostate cancers, amongst others. The *BRCAness* phenotype found within additional cancers suggests that additional individuals may also benefit from the predictive capabilities of deep learning histology approaches; however, the publicly available datasets for these other cancer types are extremely limited in size. Additional efforts must be made in obtaining the relevant digitalized slides necessary for training and validating additional models for extended applications.

Further, while this study provided an initial proof-of-concept in detecting a clinically relevant mutational signature with actionable treatment options, there are additional diagnostic signatures that are of interest to detect. These include signatures associated with *1)* microsatellite instability, which convey a sensitivity to immune checkpoint inhibitors; *2)* signatures associated with mutations in the proofreading exonuclease domain of *POLE*, which are also sensitive to immune checkpoint inhibitors; and *3)* signatures associated with the activity of APOBEC3 deaminases, which typically occur later in the progression of a cancer and might induce resistance to certain therapeutics such as treatment with tamoxifen.

This list is not comprehensive; however, it provides a logical progression of expanding on our ability to detect mutational signatures using alternative technologies. For instance, MSI and *POLE* deficient tumors are common across colon cancer, which have an ample collection of samples available for training and validation. This is also true for APOBEC3 activity that commonly occurs among tissues with growing public repositories including kidney, breast, and lung, amongst others. With the largest bottleneck simply being the digitalization of tissue slides, the potential for expanding on our repertoire of detectable signatures will continue to expand.

REFERENCES

1. Stratton MR, Campbell PJ, Futreal PA: **The cancer genome**. *Nature* 2009, **458**(7239):719-724.

2. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR: **Patterns of somatic mutation in human cancer genomes**. *Nature* 2007, **446**(7132):153-158.

3. Rubin AF, Green P: **Mutation patterns in cancer genomes**. *Proc Natl Acad Sci U S A* 2009, **106**(51):21766-21770.

4. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Australian Pancreatic Cancer Genome I, Consortium IBC, Consortium IM-S, PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR: **Signatures of mutational processes in human cancer**. *Nature* 2013, **500**(7463):415-421.

5. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR, Breast Cancer Working Group of the International Cancer Genome C: **Mutational processes molding the genomes of 21 breast cancers**. *Cell* 2012, **149**(5):979-993.

6. Olivier M, Hussain SP, Caron de Fromentel C, Hainaut P, Harris CC: **TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer**. *IARC Sci Publ* 2004(157):247-270.

7. Loeb LA, Bielas JH, Beckman RA: **Cancers exhibit a mutator phenotype: clinical implications**. *Cancer Res* 2008, **68**(10):3551-3557; discussion 3557.

8. Lengauer C, Kinzler KW, Vogelstein B: **Genetic instabilities in human cancers**. *Nature* 1998, **396**(6712):643-649.

9. Cahill DP, Levine KK, Betensky RA, Codd PJ, Romany CA, Reavie LB, Batchelor TT, Futreal PA, Stratton MR, Curry WT, Iafrate AJ, Louis DN: **Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment**. *Clin Cancer Res* 2007, **13**(7):2038-2045.

10. Hunter C, Smith R, Cahill DP, Stephens P, Stevens C, Teague J, Greenman C, Edkins S, Bignell G, Davies H, O'Meara S, Parker A, Avis T, Barthorpe S, Brackenbury L, Buck G, Butler A, Clements J, Cole J, Dicks E, Forbes S, Gorton M, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Kosmidou V, Laman R, Lugg R, Menzies A, Perry J, Petty R, Raine K, Richardson D, Shepherd R, Small A, Solomon H, Tofts C, Varian J, West S, Widaa S, Yates A, Easton DF, Riggins G, Roy JE, Levine KK, Mueller W, Batchelor TT, Louis DN, Stratton MR, Futreal PA, Wooster R: **A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy**. *Cancer Res* 2006, **66**(8):3987-3991.

11. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L: **Mutational landscape and significance across 12 major cancer types**. *Nature* 2013, **502**(7471):333-339.

12. Nedelko T, Arlt VM, Phillips DH, Hollstein M: **TP53 mutation signature supports involvement of aristolochic acid in the aetiology of endemic nephropathy-associated tumours**. *Int J Cancer* 2009, **124**(4):987-990.

13. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P: **Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers**. *Oncogene* 2002, **21**(48):7435-7451.

14. Pfeifer GP, You YH, Besaratinia A: **Mutations induced by ultraviolet light**. *Mutat Res* 2005, **571**(1-2):19-31.

15. Mace K, Aguilar F, Wang JS, Vautravers P, Gomez-Lechon M, Gonzalez FJ, Groopman J, Harris CC, Pfeifer AM: **Aflatoxin B1-induced DNA adduct formation and p53 mutations in CYP450-expressing human liver cell lines**. *Carcinogenesis* 1997, **18**(7):1291-1297.

16. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.

17. Steele CD, Abbasi A, Islam SMA, Khandekar A, Haase K, Hames S, Tarabichi M, Lesluyes T, Flanagan AM, Mertens F, Van Loo P, Alexandrov LB, Pillay N: **Signatures of copy number alterations in human cancer**. *bioRxiv* 2021:2021.2004.2030.441940.

18. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Group PMSW, Getz G, Rozen SG, Stratton MR, Consortium P: **The repertoire of mutational signatures in human cancer**. *Nature* 2020, **578**(7793):94-101.

19. Alexandrov LB, Stratton MR: **Mutational signatures: the patterns of somatic mutations hidden in cancer genomes**. *Curr Opin Genet Dev* 2014, **24**:52-60.

20. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, Campbell PJ, Vineis P, Phillips DH, Stratton MR: **Mutational signatures associated with tobacco smoking in human cancer**. *Science* 2016, **354**(6312):618-622.

21. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR: **Deciphering signatures of mutational processes operative in human cancer**. *Cell Rep* 2013, **3**(1):246-259.

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

23. Morganella S, Alexandrov LB, Glodzik D, Zou X, Davies H, Staaf J, Sieuwerts AM, Brinkman AB, Martin S, Ramakrishna M, Butler A, Kim HY, Borg A, Sotiriou C, Futreal PA, Campbell PJ, Span PN, Van Laere S, Lakhani SR, Eyfjord JE, Thompson AM, Stunnenberg HG, van de Vijver MJ, Martens JW, Borresen-Dale AL, Richardson AL, Kong G, Thomas G, Sale J, Rada C, Stratton MR, Birney E, Nik-Zainal S: **The topography of mutational processes in breast cancer genomes**. *Nat Commun* 2016, **7**:11383.

24. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Van Loo P, Ju YS, Smid M, Brinkman AB, Morganella S, Aure MR, Lingjaerde OC, Langerod A, Ringner M, Ahn SM, Boyault S, Brock JE, Broeks A, Butler A, Desmedt C, Dirix L, Dronov S, Fatima A, Foekens JA, Gerstung M, Hooijer GKJ, Jang SJ, Jones DR, Kim HY, King TA, Krishnamurthy S, Lee HJ, Lee JY, Li Y, McLaren S, Menzies A, Mustonen V, O'Meara S, Pauporte I, Pivot X, Purdie CA, Raine K, Ramakrishnan K, Rodriguez-Gonzalez FG, Romieu G, Sieuwerts AM, Simpson PT, Shepherd R, Stebbings L, Stefansson OA, Teague J, Tommasi S, Treilleux I, Van den Eynden GG, Vermeulen P, Vincent-Salomon A, Yates L, Caldas C, Van't Veer L, Tutt A, Knappskog S, Tan BKT, Jonkers J, Borg A, Ueno NT, Sotiriou C, Viari A, Futreal PA, Campbell PJ, Span PN, Van Laere S, Lakhani SR, Eyfjord JE, Thompson AM, Birney E, Stunnenberg HG, van de Vijver MJ, Martens JWM, Borresen-Dale AL, Richardson AL, Kong G, Thomas G, Stratton MR: **Author Correction: Landscape of somatic mutations in 560 breast cancer whole-genome sequences**. *Nature* 2019, **566**(7742):E1.

25. Krokan HE, Bjoras M: **Base excision repair**. *Cold Spring Harb Perspect Biol* 2013, **5**(4):a012583.

26. Robertson AB, Klungland A, Rognes T, Leiros I: **DNA repair in mammalian cells: Base excision repair: the long and short of it**. *Cell Mol Life Sci* 2009, **66**(6):981-993.

27. Wilson DM, 3rd, Bohr VA: **The mechanics of base excision repair, and its relationship to aging and disease**. *DNA Repair (Amst)* 2007, **6**(4):544-559.

28. Strauss BS: **The "A" rule revisited: polymerases as determinants of mutational specificity**. *DNA Repair (Amst)* 2002, **1**(2):125-135.

29. Petljak M, Dananberg A, Chu K, Bergstrom EN, Striepen J, von Morgen P, Chen Y, Shah H, Sale JE, Alexandrov LB, Stratton MR, Maciejowski J: **Mechanisms of APOBEC3 mutagenesis in human cancer cells**. *Nature* 2022, **607**(7920):799-807.

30. Hanawalt PC, Spivak G: **Transcription-coupled DNA repair: two decades of progress and surprises**. *Nat Rev Mol Cell Biol* 2008, **9**(12):958-970.

31. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordonez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, Birney E, Rhodes MD, McKernan KJ, Stratton

MR, Futreal PA, Campbell PJ: **A small-cell lung cancer genome with complex signatures of tobacco exposure**. *Nature* 2010, **463**(7278):184-190.

32. Poon SL, McPherson JR, Tan P, Teh BT, Rozen SG: **Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention**. *Genome Med* 2014, **6**(3):24.

33. Poon SL, Huang MN, Choo Y, McPherson JR, Yu W, Heng HL, Gan A, Myint SS, Siew EY, Ler LD, Ng LG, Weng WH, Chuang CK, Yuen JS, Pang ST, Tan P, Teh BT, Rozen SG: **Mutation signatures implicate aristolochic acid in bladder cancer development**. *Genome Med* 2015, **7**(1):38.

34. Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, Weninger A, Hollstein M, Stratton MR, Phillips DH: **The genome as a record of environmental exposure**. *Mutagenesis* 2015, **30**(6):763-770.

35. Hoang ML, Chen CH, Sidorenko VS, He J, Dickman KG, Yun BH, Moriya M, Niknafs N, Douville C, Karchin R, Turesky RJ, Pu YS, Vogelstein B, Papadopoulos N, Grollman AP, Kinzler KW, Rosenquist TA: **Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing**. *Sci Transl Med* 2013, **5**(197):197ra102.

36. Hollstein M, Moriya M, Grollman AP, Olivier M: **Analysis of TP53 mutation spectra reveals the fingerprint of the potent environmental carcinogen, aristolochic acid**. *Mutat Res* 2013, **753**(1):41-49.

37. Poon SL, Pang ST, McPherson JR, Yu W, Huang KK, Guan P, Weng WH, Siew EY, Liu Y, Heng HL, Chong SC, Gan A, Tay ST, Lim WK, Cutcutache I, Huang D, Ler LD, Nairismagi ML, Lee MH, Chang YH, Yu KJ, Chan-On W, Li BK, Yuan YF, Qian CN, Ng KF, Wu CF, Hsu CL, Bunte RM, Stratton MR, Futreal PA, Sung WK, Chuang CK, Ong CK, Rozen SG, Tan P, Teh BT: **Genome-wide mutational signatures of aristolochic acid and its application as a screening tool**. *Sci Transl Med* 2013, **5**(197):197ra101.

38. Chen CH, Dickman KG, Moriya M, Zavadil J, Sidorenko VS, Edwards KL, Gnatenko DV, Wu L, Turesky RJ, Wu XR, Pu YS, Grollman AP: **Aristolochic acid-associated urothelial cancer in Taiwan**. *Proc Natl Acad Sci U S A* 2012, **109**(21):8241-8246.

39. Scelo G, Riazalhosseini Y, Greger L, Letourneau L, Gonzalez-Porta M, Wozniak MB, Bourgey M, Harnden P, Egevad L, Jackson SM, Karimzadeh M, Arseneault M, Lepage P, How-Kit A, Daunay A, Renault V, Blanche H, Tubacher E, Sehmoun J, Viksna J, Celms E, Opmanis M, Zarins A, Vasudev NS, Seywright M, Abedi-Ardekani B, Carreira

C, Selby PJ, Cartledge JJ, Byrnes G, Zavadil J, Su J, Holcatova I, Brisuda A, Zaridze D, Moukeria A, Foretova L, Navratilova M, Mates D, Jinga V, Artemov A, Nedoluzhko A, Mazur A, Rastorguev S, Boulygina E, Heath S, Gut M, Bihoreau MT, Lechner D, Foglio M, Gut IG, Skryabin K, Prokhortchouk E, Cambon-Thomsen A, Rung J, Bourque G, Brennan P, Tost J, Banks RE, Brazma A, Lathrop GM: **Variation in genomic landscape of clear cell renal cell carcinoma across Europe**. *Nat Commun* 2014, **5**:5135.

40. Jelakovic B, Castells X, Tomic K, Ardin M, Karanovic S, Zavadil J: **Renal cell carcinomas of chronic kidney disease patients harbor the mutational signature of carcinogenic aristolochic acid**. *Int J Cancer* 2015, **136**(12):2967-2972.

41. Ng AWT, Poon SL, Huang MN, Lim JQ, Boot A, Yu W, Suzuki Y, Thangaraju S, Ng CCY, Tan P, Pang ST, Huang HY, Yu MC, Lee PH, Hsieh SY, Chang AY, Teh BT, Rozen SG: **Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia**. *Sci Transl Med* 2017, **9**(412).

42. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, Wedge DC, Fullam A, Alexandrov LB, Tubio JM, Stebbings L, Menzies A, Widaa S, Stratton MR, Jones PH, Campbell PJ: **Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin**. *Science* 2015, **348**(6237):880-886.

43. Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, Couchy G, Meiller C, Shinde J, Soysouvanh F, Calatayud AL, Pinyol R, Pelletier L, Balabaud C, Laurent A, Blanc JF, Mazzaferro V, Calvo F, Villanueva A, Nault JC, Bioulac-Sage P, Stratton MR, Llovet JM, Zucman-Rossi J: **Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets**. *Nat Genet* 2015, **47**(5):505-511.

44. Petljak M, Alexandrov LB: **Understanding mutagenesis through delineation of mutational signatures in human cancer**. *Carcinogenesis* 2016, **37**(6):531-540.

45. Kim J, Mouw KW, Polak P, Braunstein LZ, Kamburov A, Kwiatkowski DJ, Rosenberg JE, Van Allen EM, D'Andrea A, Getz G: **Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors**. *Nat Genet* 2016, **48**(6):600-606.

46. Merlevede J, Droin N, Qin T, Meldi K, Yoshida K, Morabito M, Chautard E, Auboeuf D, Fenaux P, Braun T, Itzykson R, de Botton S, Quesnel B, Commes T, Jourdan E, Vainchenker W, Bernard O, Pata-Merci N, Solier S, Gayevskiy V, Dinger ME, Cowley MJ, Selimoglu-Buet D, Meyer V, Artiguenave F, Deleuze JF, Preudhomme C, Stratton

MR, Alexandrov LB, Padron E, Ogawa S, Koscielny S, Figueroa M, Solary E: **Mutation allele burden remains unchanged in chronic myelomonocytic leukaemia responding to hypomethylating agents**. *Nat Commun* 2016, **7**:10767.

47. Polak P, Kim J, Braunstein LZ, Karlic R, Haradhavala NJ, Tiao G, Rosebrock D, Livitz D, Kubler K, Mouw KW, Kamburov A, Maruvka YE, Leshchiner I, Lander ES, Golub TR, Zick A, Orthwein A, Lawrence MS, Batra RN, Caldas C, Haber DA, Laird PW, Shen H, Ellisen LW, D'Andrea AD, Chanock SJ, Foulkes WD, Getz G: **A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer**. *Nat Genet* 2017, **49**(10):1476-1486.

48. Mimaki S, Totsuka Y, Suzuki Y, Nakai C, Goto M, Kojima M, Arakawa H, Takemura S, Tanaka S, Marubashi S, Kinoshita M, Matsuda T, Shibata T, Nakagama H, Ochiai A, Kubo S, Nakamori S, Esumi H, Tsuchihara K: **Hypermutation and unique mutational signatures of occupational cholangiocarcinoma in printing workers exposed to haloalkanes**. *Carcinogenesis* 2016, **37**(8):817-826.

49. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, Patch AM, Kakavand H, Alexandrov LB, Burke H, Jakrot V, Kazakoff S, Holmes O, Leonard C, Sabarinathan R, Mularoni L, Wood S, Xu Q, Waddell N, Tembe V, Pupo GM, De Paoli-Iseppi R, Vilain RE, Shang P, Lau LMS, Dagg RA, Schramm SJ, Pritchard A, Dutton-Regester K, Newell F, Fitzgerald A, Shang CA, Grimmond SM, Pickett HA, Yang JY, Stretch JR, Behren A, Kefford RF, Hersey P, Long GV, Cebon J, Shackleton M, Spillane AJ, Saw RPM, Lopez-Bigas N, Pearson JV, Thompson JF, Scolyer RA, Mann GJ: **Whole-genome landscapes of major melanoma subtypes**. *Nature* 2017, **545**(7653):175-180.

50. Letouze E, Shinde J, Renault V, Couchy G, Blanc JF, Tubacher E, Bayard Q, Bacq D, Meyer V, Semhoun J, Bioulac-Sage P, Prevot S, Azoulay D, Paradis V, Imbeaud S, Deleuze JF, Zucman-Rossi J: **Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis**. *Nat Commun* 2017, **8**(1):1315.

51. Van Hoeck A, Tjoonk NH, van Boxtel R, Cuppen E: **Portrait of a cancer: mutational signature analyses for cancer diagnostics**. *BMC Cancer* 2019, **19**(1):457.

52. Supek F, Lehner B: **Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes**. *Cell* 2017, **170**(3):534-547 e523.

53. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y,

Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortes ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G: **Mutational heterogeneity in cancer and the search for new cancer-associated genes**. *Nature* 2013, **499**(7457):214-218.

54. Matsuda T, Kawanishi M, Yagi T, Matsui S, Takebe H: **Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases**. *Nucleic Acids Res* 1998, **26**(7):1769-1774.

55. de Gruijl FR, van Kranen HJ, Mullenders LH: **UV-induced DNA damage, repair, mutations and oncogenic pathways in skin cancer**. *J Photochem Photobiol B* 2001, **63**(1-3):19-27.

56. Brash DE: **UV signature mutations**. *Photochem Photobiol* 2015, **91**(1):15-26.

57. Mas-Ponte D, Supek F: **DNA mismatch repair promotes APOBEC3-mediated diffuse hypermutation in human cancers**. *Nat Genet* 2020, **52**(9):958-968.

58. Taylor BJ, Nik-Zainal S, Wu YL, Stebbings LA, Raine K, Campbell PJ, Rada C, Stratton MR, Neuberger MS: **DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis**. *Elife* 2013, **2**:e00534.

59. Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, Kim J, Kwiatkowski DJ, Fargo DC, Mieczkowski PA, Getz G, Gordenin DA: **An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers**. *Nat Genet* 2015, **47**(9):1067-1072.

60. Consortium ITP-CAoWG: **Pan-cancer analysis of whole genomes**. *Nature* 2020, **578**(7793):82-93.

61. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, Lawrence MS: **Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features**. *Science* 2019, **364**(6447).

62. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF: **Statistical analysis of pathogenicity of somatic mutations in cancer**. *Genetics* 2006, **173**(4):2187-2198.

63. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ: **Universal Patterns of Selection in Cancer and Somatic Tissues**. *Cell* 2018, **173**(7):1823.

64. Hainaut P, Pfeifer GP: **Patterns of p53 G-->T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke**. *Carcinogenesis* 2001, **22**(3):367-374.

65. Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak LJ, Kryukov GV, Malc E, Mieczkowski PA, Resnick MA, Gordenin DA: **Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions**. *Mol Cell* 2012, **46**(4):424-435.

66. Hodgkinson A, Eyre-Walker A: **Variation in the mutation rate across mammalian genomes**. *Nat Rev Genet* 2011, **12**(11):756-766.

67. Pich O, Muinos F, Sabarinathan R, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N: **Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes**. *Cell* 2018, **175**(4):1074-1087 e1018.

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

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

70. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ, Kamburov A, Hanawalt PC, Wheeler DA, Koren A, Lawrence MS, Getz G: **Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair**. *Cell* 2016, **164**(3):538-549.

71. Hoopes JI, Cortez LM, Mertz TM, Malc EP, Mieczkowski PA, Roberts SA: **APOBEC3A and APOBEC3B Preferentially Deaminate the Lagging Strand Template during DNA Replication**. *Cell Rep* 2016, **14**(6):1273-1282.

72. Seplyarskiy VB, Soldatov RA, Popadin KY, Antonarakis SE, Bazykin GA, Nikolaev SI: **APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication**. *Genome Res* 2016, **26**(2):174-182.

73. Dixon JR, Gorkin DU, Ren B: **Chromatin Domains: The Unit of Chromosome Organization**. *Mol Cell* 2016, **62**(5):668-680.

74. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome**. *Science* 2009, **326**(5950):289-293.

75. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR: **Human mutation rate associated with DNA replication timing**. *Nat Genet* 2009, **41**(4):393-395.

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

77. Zou X, Morganella S, Glodzik D, Davies H, Li Y, Stratton MR, Nik-Zainal S: **Short inverted repeats contribute to localized mutability in human somatic cells**. *Nucleic Acids Res* 2017, **45**(19):11213-11221.

78. Pfeifer GP: **Mutagenesis at methylated CpG sequences**. *Curr Top Microbiol Immunol* 2006, **301**:259-281.

79. Lieber MR: **The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway**. *Annu Rev Biochem* 2010, **79**:181-211.

80. Errol C. Friedberg GCW, Wolfram Siede, Richard D. Wood, Roger A. Schultz, Tome Ellenberger: **DNA repair and mutagenesis**. In: *DNA Repair and Mutagenesis.* 2006: 1-1.

81. Mimitou EP, Symington LS: **DNA end resection--unraveling the tail**. *DNA Repair (Amst)* 2011, **10**(3):344-348.

82. San Filippo J, Sung P, Klein H: **Mechanism of eukaryotic homologous recombination**. *Annu Rev Biochem* 2008, **77**:229-257.

83. Burrell RA, McGranahan N, Bartek J, Swanton C: **The causes and consequences of genetic heterogeneity in cancer evolution**. *Nature* 2013, **501**(7467):338-345.

84. Cortes-Ciriano I, Lee JJ, Xi R, Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang CZ, Pellman DS, Group PSVW, Park PJ, Consortium P: **Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing**. *Nat Genet* 2020, **52**(3):331-341.

85. Maciejowski J, Chatzipli A, Dananberg A, Chu K, Toufektchan E, Klimczak LJ, Gordenin DA, Campbell PJ, de Lange T: **APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis**. *Nat Genet* 2020, **52**(9):884-890.

86. Sakofsky CJ, Roberts SA, Malc E, Mieczkowski PA, Resnick MA, Gordenin DA, Malkova A: **Break-induced replication is a source of mutation clusters underlying kataegis**. *Cell Rep* 2014, **7**(5):1640-1648.

87. Burns MB, Temiz NA, Harris RS: **Evidence for APOBEC3B mutagenesis in multiple human cancers**. *Nat Genet* 2013, **45**(9):977-983.

88. Petljak M, Chu K, Dananberg A, Bergstrom EN, Morgen Pv, Alexandrov LB, Stratton MR, Maciejowski J: **The APOBEC3A deaminase drives episodic mutagenesis in cancer cells**. *bioRxiv* 2021:2021.2002.2014.431145.

89. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, Harris S, Shah RR, Resnick MA, Getz G, Gordenin DA: **An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers**. *Nat Genet* 2013, **45**(9):970-976.

90. Bogerd HP, Wiegand HL, Doehle BP, Lueders KK, Cullen BR: **APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells**. *Nucleic Acids Res* 2006, **34**(1):89-95.

91. Malim MH: **Natural resistance to HIV infection: The Vif-APOBEC interaction**. *C R Biol* 2006, **329**(11):871-875.

92. Malim MH, Bieniasz PD: **HIV Restriction Factors and Mechanisms of Evasion**. *Cold Spring Harb Perspect Med* 2012, **2**(5):a006940.

93. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, Neuberger MS, Malim MH: **DNA deamination mediates innate immunity to retroviral infection**. *Cell* 2003, **113**(6):803-809.

94. Venkatesan S, Rosenthal R, Kanu N, McGranahan N, Bartek J, Quezada SA, Hare J, Harris RS, Swanton C: **Perspective: APOBEC mutagenesis in drug resistance and immune escape in HIV and cancer evolution**. *Ann Oncol* 2018, **29**(3):563-572.

95. Harris RS, Dudley JP: **APOBECs and virus restriction**. *Virology* 2015, **479-480**:131-145.

96. Chen H, Lilley CE, Yu Q, Lee DV, Chou J, Narvaiza I, Landau NR, Weitzman MD: **APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons**. *Curr Biol* 2006, **16**(5):480-485.

97. Maul RW, Gearhart PJ: **AID and somatic hypermutation**. *Adv Immunol* 2010, **105**:159-191.

98. Kasar S, Kim J, Improgo R, Tiao G, Polak P, Haradhvala N, Lawrence MS, Kiezun A, Fernandes SM, Bahl S, Sougnez C, Gabriel S, Lander ES, Kim HT, Getz G, Brown JR: **Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution**. *Nat Commun* 2015, **6**:8866.

99. Garraway LA: **Genomics-driven oncology: framework for an emerging paradigm**. *J Clin Oncol* 2013, **31**(15):1806-1814.

100. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.

101. Hanahan D: **Hallmarks of Cancer: New Dimensions**. *Cancer Discov* 2022, **12**(1):31-46.

102. Marquard AM, Eklund AC, Joshi T, Krzystanek M, Favero F, Wang ZC, Richardson AL, Silver DP, Szallasi Z, Birkbak NJ: **Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs**. *Biomark Res* 2015, **3**:9.

103. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, Ramakrishna M, Martin S, Boyault S, Sieuwerts AM, Simpson PT, King TA, Raine K, Eyfjord JE, Kong G, Borg

A, Birney E, Stunnenberg HG, van de Vijver MJ, Borresen-Dale AL, Martens JW, Span PN, Lakhani SR, Vincent-Salomon A, Sotiriou C, Tutt A, Thompson AM, Van Laere S, Richardson AL, Viari A, Campbell PJ, Stratton MR, Nik-Zainal S: **HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures**. *Nat Med* 2017, **23**(4):517-525.

104. Sztupinszki Z, Diossy M, Krzystanek M, Borcsok J, Pomerantz MM, Tisza V, Spisak S, Rusz O, Csabai I, Freedman ML, Szallasi Z: **Detection of Molecular Signatures of Homologous Recombination Deficiency in Prostate Cancer with or without BRCA1/2 Mutations**. *Clin Cancer Res* 2020, **26**(11):2673-2680.

105. Borcsok J, Diossy M, Sztupinszki Z, Prosz A, Tisza V, Spisak S, Rusz O, Stormoen DR, Pappot H, Csabai I, Brunak S, Mouw KW, Szallasi Z: **Detection of Molecular Signatures of Homologous Recombination Deficiency in Bladder Cancer**. *Clin Cancer Res* 2021, **27**(13):3734-3743.

106. Lee V, Murphy A, Le DT, Diaz LA, Jr.: **Mismatch Repair Deficiency and Response to Immune Checkpoint Blockade**. *Oncologist* 2016, **21**(10):1200-1211.

107. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, Skora AD, Luber BS, Azad NS, Laheru D, Biedrzycki B, Donehower RC, Zaheer A, Fisher GA, Crocenzi TS, Lee JJ, Duffy SM, Goldberg RM, de la Chapelle A, Koshiji M, Bhaijee F, Huebner T, Hruban RH, Wood LD, Cuka N, Pardoll DM, Papadopoulos N, Kinzler KW, Zhou S, Cornish TC, Taube JM, Anders RA, Eshleman JR, Vogelstein B, Diaz LA, Jr.: **PD-1 Blockade in Tumors with Mismatch-Repair Deficiency**. *N Engl J Med* 2015, **372**(26):2509-2520.

108. Jiricny J: **The multifaceted mismatch-repair system**. *Nat Rev Mol Cell Biol* 2006, **7**(5):335-346.

109. Pena-Diaz J, Jiricny J: **Mammalian mismatch repair: error-free or error-prone?** *Trends Biochem Sci* 2012, **37**(5):206-214.

110. Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, Collisson EA, Vemula SS, Naujokas A, Chiotti KE, Cheng JB, Fassihi H, Blumberg AJ, Bailey CV, Fudem GM, Mihm FG, Cunningham BB, Neuhaus IM, Liao W, Oh DH, Cleaver JE, LeBoit PE, Costello JF, Lehmann AR, Gray JW, Spellman PT, Arron ST, Huh N, Purdom E, Cho RJ: **Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes**. *Cell Rep* 2014, **9**(4):1228-1234.

111. Bouffet E, Larouche V, Campbell BB, Merico D, de Borja R, Aronson M, Durno C, Krueger J, Cabric V, Ramaswamy V, Zhukova N, Mason G, Farah R, Afzal S, Yalon M, Rechavi G, Magimairajan V, Walsh MF, Constantini S, Dvir R, Elhasid R, Reddy A, Osborn M, Sullivan M, Hansford J, Dodgshun A, Klauber-Demore N, Peterson L, Patel S, Lindhorst S, Atkinson J, Cohen Z, Laframboise R, Dirks P, Taylor M, Malkin D, Albrecht S, Dudley RW, Jabado N, Hawkins CE, Shlien A, Tabori U: **Immune Checkpoint Inhibition for Hypermutant Glioblastoma Multiforme Resulting From Germline Biallelic Mismatch Repair Deficiency**. *J Clin Oncol* 2016, **34**(19):2206-2211.

112. Boland CR, Goel A: **Microsatellite instability in colorectal cancer**. *Gastroenterology* 2010, **138**(6):2073-2087 e2073.

113. Cancer Genome Atlas N: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.

114. Cancer Genome Atlas Research N, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, Shen H, Robertson AG, Pashtan I, Shen R, Benz CC, Yau C, Laird PW, Ding L, Zhang W, Mills GB, Kucherlapati R, Mardis ER, Levine DA: **Integrated genomic characterization of endometrial carcinoma**. *Nature* 2013, **497**(7447):67-73.

115. Johanns TM, Miller CA, Dorward IG, Tsien C, Chang E, Perry A, Uppaluri R, Ferguson C, Schmidt RE, Dahiya S, Ansstas G, Mardis ER, Dunn GP: **Immunogenomics of Hypermutated Glioblastoma: A Patient with Germline POLE Deficiency Treated with Checkpoint Blockade Immunotherapy**. *Cancer Discov* 2016, **6**(11):1230-1236.

116. Ceccaldi R, Rondinelli B, D'Andrea AD: **Repair Pathway Choices and Consequences at the Double-Strand Break**. *Trends Cell Biol* 2016, **26**(1):52-64.

117. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, Liu Y, Akbani R, Feng B, Donehower LA, Miller C, Shen Y, Karimi M, Chen H, Kim P, Jia P, Shinbrot E, Zhang S, Liu J, Hu H, Bailey MH, Yau C, Wolf D, Zhao Z, Weinstein JN, Li L, Ding L, Mills GB, Laird PW, Wheeler DA, Shmulevich I, Cancer Genome Atlas Research N, Monnat RJ, Jr., Xiao Y, Wang C: **Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas**. *Cell Rep* 2018, **23**(1):239-254 e236.

118. Couch FJ, Nathanson KL, Offit K: **Two decades after BRCA: setting paradigms in personalized cancer care and prevention**. *Science* 2014, **343**(6178):1466-1470.

119. King MC: **"The race" to clone BRCA1**. *Science* 2014, **343**(6178):1462-1465.

120. Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C, Martin NM, Jackson SP, Smith GC, Ashworth A: **Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy**. *Nature* 2005, **434**(7035):917-921.

121. Prakash R, Zhang Y, Feng W, Jasin M: **Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins**. *Cold Spring Harb Perspect Biol* 2015, **7**(4):a016600.

122. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, Kyle S, Meuth M, Curtin NJ, Helleday T: **Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase**. *Nature* 2005, **434**(7035):913-917.

123. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, Mortimer P, Swaisland H, Lau A, O'Connor MJ, Ashworth A, Carmichael J, Kaye SB, Schellens JH, de Bono JS: **Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers**. *N Engl J Med* 2009, **361**(2):123-134.

124. Greenblatt MS, Bennett WP, Hollstein M, Harris CC: **Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis**. *Cancer Res* 1994, **54**(18):4855-4878.

125. Hainaut P, Hollstein M: **p53 and human cancer: the first ten thousand mutations**. *Adv Cancer Res* 2000, **77**:81-137.

126. Le Calvez F, Mukeria A, Hunt JD, Kelm O, Hung RJ, Taniere P, Brennan P, Boffetta P, Zaridze DG, Hainaut P: **TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: distinct patterns in never, former, and current smokers**. *Cancer Res* 2005, **65**(12):5076-5083.

127. Omichessan H, Severi G, Perduca V: **Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance**. *bioRxiv* 2018.

128. Carlson J, Li JZ, Zollner S: **Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets**. *BMC Genomics* 2018, **19**(1):845.

129. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C: **DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution**. *Genome Biol* 2016, **17**:31.

130. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP: **Maftools: efficient and comprehensive analysis of somatic variants in cancer**. *Genome Res* 2018, **28**(11):1747-1756.

131. Gehring JS, Fischer B, Lawrence M, Huber W: **SomaticSignatures: inferring mutational signatures from single-nucleotide variants**. *Bioinformatics* 2015, **31**(22):3673-3675.

132. Rosales RA, Drummond RD, Valieris R, Dias-Neto E, da Silva IT: **signeR: an empirical Bayesian approach to mutational signature discovery**. *Bioinformatics* 2017, **33**(1):8-16.

133. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, de Ligt J, Behjati S, Grolleman JE, van Wezel T, Nik-Zainal S, Kuiper RP, Cuppen E, Clevers H: **Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer**. *Science* 2017, **358**(6360):234-238.

134. Boot A, Huang MN, Ng AWT, Ho SC, Lim JQ, Kawakami Y, Chayama K, Teh BT, Nakagawa H, Rozen SG: **In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors**. *Genome Res* 2018, **28**(5):654-665.

135. van Zeeland AA, Vreeswijk MP, de Gruijl FR, van Kranen HJ, Vrieling H, Mullenders LF: **Transcription-coupled repair: impact on UV-induced mutagenesis in cultured rodent cells and mouse skin tumors**. *Mutat Res* 2005, **577**(1-2):170-178.

136. Steele CD, Tarabichi M, Oukrif D, Webster AP, Ye H, Fittall M, Lombard P, Martincorena I, Tarpey PS, Collord G, Haase K, Strauss SJ, Berisha F, Vaikkinen H, Dhami P, Jansen M, Behjati S, Amary MF, Tirabosco R, Feber A, Campbell PJ, Alexandrov LB, Van Loo P, Flanagan AM, Pillay N: **Undifferentiated Sarcomas Develop through Distinct Evolutionary Pathways**. *Cancer Cell* 2019, **35**(3):441-456 e448.

137. Macintyre G, Goranova TE, De Silva D, Ennis D, Piskorz AM, Eldridge M, Sie D, Lewsley LA, Hanif A, Wilson C, Dowson S, Glasspool RM, Lockley M, Brockbank E, Montes A, Walther A, Sundar S, Edmondson R, Hall GD, Clamp A, Gourley C, Hall M, Fotopoulou C, Gabra H, Paul J, Supernat A, Millan D, Hoyle A, Bryson G, Nourse C, Mincarelli L, Sanchez LN, Ylstra B, Jimenez-Linan M, Moore L, Hofmann O, Markowetz F, McNeish IA, Brenton JD: **Copy number signatures and mutational processes in ovarian carcinoma**. *Nat Genet* 2018, **50**(9):1262-1270.

138. Youn A, Simon R: **Identifying cancer driver genes in tumor genome sequencing studies**. *Bioinformatics* 2011, **27**(2):175-181.

139. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: **The consensus coding sequences of human breast and colorectal cancers**. *Science* 2006, **314**(5797):268-274.

140. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba, II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK: **Somatic mutations affect key pathways in lung adenocarcinoma**. *Nature* 2008, **455**(7216):1069-1075.

141. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VE, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M, Seshagiri S: **Diverse somatic mutation patterns and pathway alterations in human cancers**. *Nature* 2010, **466**(7308):869-873.

142. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ: **Universal Patterns of Selection in Cancer and Somatic Tissues**. *Cell* 2017, **171**(5):1029-1041 e1021.

143. Bergstrom EN, Huang MN, Mahto U, Barnes M, Stratton MR, Rozen SG, Alexandrov LB: **SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events**. *BMC Genomics* 2019, **20**(1):685.

144. Gel B, Serra E: **karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data**. *Bioinformatics* 2017, **33**(19):3088-3090.

145. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, Sofia HJ, Hutter C, Getz G, Wheeler D, Ding L, Grp MW, Network CGAR: **Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines**. *Cell Syst* 2018, **6**(3):271-+.

146. Lemire D: **Fast Random Integer Generation in an Interval**. *ACM Trans Model Comput Simul* 2019, **29**(1):1-12.

147. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F: **The Ensembl Variant Effect Predictor**. *Genome Biol* 2016, **17**(1):122.

148. Martincorena I, Campbell PJ: **Somatic mutation in cancer and normal cells**. *Science* 2015, **349**(6255):1483-1489.

149. Boichard A, Tsigelny IF, Kurzrock R: **High expression of PD-1 ligands is associated with kataegis mutational signature and APOBEC3 alterations**. *Oncoimmunology* 2017, **6**(3):e1284719.

150. Chen JM, Ferec C, Cooper DN: **Patterns and mutational signatures of tandem base substitutions causing human inherited disease**. *Hum Mutat* 2013, **34**(8):1119-1130.

151. Wang Q, Pierce-Hoffman E, Cummings BB, Alfoldi J, Francioli LC, Gauthier LD, Hill AJ, O'Donnell-Luria AH, Genome Aggregation Database Production T, Genome Aggregation Database C, Karczewski KJ, MacArthur DG: **Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes**. *Nat Commun* 2020, **11**(1):2539.

152. Bergstrom EN, Luebeck J, Petljak M, Khandekar A, Barnes M, Zhang T, Steele CD, Pillay N, Landi MT, Bafna V, Mischel PS, Harris RS, Alexandrov LB: **Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA**. *Nature* 2022.

153. Kaplanis J, Akawi N, Gallone G, McRae JF, Prigmore E, Wright CF, Fitzpatrick DR, Firth HV, Barrett JC, Hurles ME, Deciphering Developmental Disorders s: **Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations**. *Genome Res* 2019, **29**(7):1047-1056.

154. Veltman JA, Brunner HG: **De novo mutations in human genetic disease**. *Nat Rev Genet* 2012, **13**(8):565-575.

155. D'Antonio M, Tamayo P, Mesirov JP, Frazer KA: **Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels**. *Cell Rep* 2016, **16**(3):672-683.

156. Lin X, Hua Y, Gu S, Lv L, Li X, Chen P, Dai P, Hu Y, Liu A, Li J: **kataegis: an R package for identification and visualization of the genomic localized hypermutation regions using high-throughput sequencing**. *BMC Genomics* 2021, **22**(1):440.

157. Yin X, Bi R, Ma P, Zhang S, Zhang Y, Sun Y, Zhang Y, Jing Y, Yu M, Wang W, Tan L, Di W, Zhuang G, Cai MC: **Multiregion whole-genome sequencing depicts intratumour heterogeneity and punctuated evolution in ovarian clear cell carcinoma**. *J Med Genet* 2020, **57**(9):605-609.

158. Bergstrom EN, Barnes M, Martincorena I, Alexandrov LB: **Generating realistic null hypothesis of cancer mutational landscapes using SigProfilerSimulator**. *BMC Bioinformatics* 2020, **21**(1):438.

159. Islam SMA, Wu Y, Díaz-Gay M, Bergstrom EN, He Y, Barnes M, Vella M, Wang J, Teague JW, Clapham P, Moody S, Senkin S, Li YR, Riva L, Zhang T, Gruber AJ, Vangara R, Steele CD, Otlu B, Khandekar A, Abbasi A, Humphreys L, Syulyukina N, Brady SW, Alexandrov BS, Pillay N, Zhang J, Adams DJ, Marticorena I, Wedge DC, Landi MT, Brennan P, Stratton MR, Rozen SG, Alexandrov LB: **Uncovering novel mutational signatures by <em>de novo</em> extraction with SigProfilerExtractor**. *bioRxiv* 2020:2020.2012.2013.422570.

160. Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, Lawrence MS, Getz G: **Passenger Hotspot Mutations in Cancer**. *Cancer Cell* 2019, **36**(3):288-301 e214.

161. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR, Ye K, Alipaz J, Bauer MJ, Beare D, Butler A, Carter RJ, Chen L, Cox AJ, Edkins S, Kokko-Gonzales PI, Gormley NA, Grocock RJ, Haudenschild CD, Hims MM, James T, Jia M, Kingsbury Z, Leroy C, Marshall J, Menzies A, Mudie LJ, Ning Z, Royce T, Schulz-Trieglaff OB, Spiridou A, Stebbings LA, Szajkowski L, Teague J, Williamson D, Chin L, Ross MT, Campbell PJ, Bentley DR, Futreal PA, Stratton MR: **A comprehensive catalogue of somatic mutations from a human cancer genome**. *Nature* 2010, **463**(7278):191-196.

162. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence M, Reynolds A, Rynes E, Vlahovicek K, Stamatoyannopoulos JA, Sunyaev SR: **Cell-of-origin chromatin**

organization shapes the mutational landscape of cancer. *Nature* 2015, **518**(7539):360-364.

163. Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, Li B, Arden K, Ren B, Nathanson DA, Kornblum HI, Taylor MD, Kaushal S, Cavenee WK, Wechsler-Reya R, Furnari FB, Vandenberg SR, Rao PN, Wahl GM, Bafna V, Mischel PS: **Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity**. *Nature* 2017, **543**(7643):122-125.

164. Koche RP, Rodriguez-Fos E, Helmsauer K, Burkert M, MacArthur IC, Maag J, Chamorro R, Munoz-Perez N, Puiggros M, Dorado Garcia H, Bei Y, Roefzaad C, Bardinet V, Szymansky A, Winkler A, Thole T, Timme N, Kasack K, Fuchs S, Klironomos F, Thiessen N, Blanc E, Schmelz K, Kunkele A, Hundsdorfer P, Rosswog C, Theissen J, Beule D, Deubzer H, Sauer S, Toedling J, Fischer M, Hertwig F, Schwarz RF, Eggert A, Torrents D, Schulte JH, Henssen AG: **Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma**. *Nat Genet* 2020, **52**(1):29-34.

165. Verhaak RGW, Bafna V, Mischel PS: **Extrachromosomal oncogene amplification in tumour pathogenesis and evolution**. *Nat Rev Cancer* 2019, **19**(5):283-288.

166. Kim H, Nguyen NP, Turner K, Wu S, Gujar AD, Luebeck J, Liu J, Deshpande V, Rajkumar U, Namburi S, Amin SB, Yi E, Menghi F, Schulte JH, Henssen AG, Chang HY, Beck CR, Mischel PS, Bafna V, Verhaak RGW: **Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers**. *Nat Genet* 2020, **52**(9):891-897.

167. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* 2013, **45**(10):1113-1120.

168. Green AM, Landry S, Budagyan K, Avgousti DC, Shalhout S, Bhagwat AS, Weitzman MD: **APOBEC3A damages the cellular genome during DNA replication**. *Cell Cycle* 2016, **15**(7):998-1008.

169. Stenglein MD, Burns MB, Li M, Lengyel J, Harris RS: **APOBEC3 proteins mediate the clearance of foreign DNA from human cells**. *Nat Struct Mol Biol* 2010, **17**(2):222-229.

170. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, Dawson KJ, Ju YS, Iorio F, Tubio JMC, Koh CC, Georgakopoulos-Soares I, Rodriguez-Martin B, Otlu B, O'Meara S, Butler AP, Menzies A, Bhosle SG, Raine K, Jones DR, Teague JW, Beal K, Latimer C, O'Neill L, Zamora J, Anderson E, Patel N, Maddison M, Ng BL,

Graham J, Garnett MJ, McDermott U, Nik-Zainal S, Campbell PJ, Stratton MR: **Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis**. *Cell* 2019, **176**(6):1282-1294 e1220.

171. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP, Arlt VM, Phillips DH, Nik-Zainal S: **A Compendium of Mutational Signatures of Environmental Agents**. *Cell* 2019, **177**(4):821-836 e816.

172. Liu Z, Hergenhahn M, Schmeiser HH, Wogan GN, Hong A, Hollstein M: **Human tumor p53 mutations are selected for in mouse embryonic fibroblasts harboring a humanized p53 gene**. *Proc Natl Acad Sci U S A* 2004, **101**(9):2963-2968.

173. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Brannon AR, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman JF, Yao J, Song W, Ross DS, Oultache A, Dogan S, Borsu L, Hameed M, Nafa K, Arcila ME, Ladanyi M, Berger MF: **Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology**. *J Mol Diagn* 2015, **17**(3):251-264.

174. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, Srinivasan P, Gao J, Chakravarty D, Devlin SM, Hellmann MD, Barron DA, Schram AM, Hameed M, Dogan S, Ross DS, Hechtman JF, DeLair DF, Yao J, Mandelker DL, Cheng DT, Chandramohan R, Mohanty AS, Ptashkin RN, Jayakumaran G, Prasad M, Syed MH, Rema AB, Liu ZY, Nafa K, Borsu L, Sadowska J, Casanova J, Bacares R, Kiecka IJ, Razumova A, Son JB, Stewart L, Baldi T, Mullaney KA, Al-Ahmadie H, Vakiani E, Abeshouse AA, Penson AV, Jonsson P, Camacho N, Chang MT, Won HH, Gross BE, Kundra R, Heins ZJ, Chen HW, Phillips S, Zhang H, Wang J, Ochoa A, Wills J, Eubank M, Thomas SB, Gardos SM, Reales DN, Galle J, Durany R, Cambria R, Abida W, Cercek A, Feldman DR, Gounder MM, Hakimi AA, Harding JJ, Iyer G, Janjigian YY, Jordan EJ, Kelly CM, Lowery MA, Morris LGT, Omuro AM, Raj N, Razavi P, Shoushtari AN, Shukla N, Soumerai TE, Varghese AM, Yaeger R, Coleman J, Bochner B, Riely GJ, Saltz LB, Scher HI, Sabbatini PJ, Robson ME, Klimstra DS, Taylor BS, Baselga J, Schultz N, Hyman DM, Arcila ME, Solit DB, Ladanyi M, Berger MF: **Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients**. *Nat Med* 2017, **23**(6):703-713.

175. Zhang T, Joubert P, Ansari-Pour N, Zhao W, Hoang PH, Lokanga R, Moye AL, Rosenbaum J, Gonzalez-Perez A, Martinez-Jimenez F, Castro A, Muscarella LA, Hofman P, Consonni D, Pesatori AC, Kebede M, Li M, Gould Rothberg BE, Peneva I, Schabath MB, Poeta ML, Costantini M, Hirsch D, Heselmeyer-Haddad K, Hutchinson A, Olanich M, Lawrence SM, Lenz P, Duggan M, Bhawsar PMS, Sang J, Kim J, Mendoza L, Saini N, Klimczak LJ, Islam SMA, Otlu B, Khandekar A, Cole N, Stewart DR, Choi J,

Brown KM, Caporaso NE, Wilson SH, Pommier Y, Lan Q, Rothman N, Almeida JS, Carter H, Ried T, Kim CF, Lopez-Bigas N, Garcia-Closas M, Shi J, Bosse Y, Zhu B, Gordenin DA, Alexandrov LB, Chanock SJ, Wedge DC, Landi MT: **Genomic and evolutionary classification of lung cancer in never smokers**. *Nat Genet* 2021, **53**(9):1348-1359.

176. Moody S, Senkin S, Islam SMA, Wang J, Nasrollahzadeh D, Penha RCC, Fitzgerald S, Bergstrom EN, Atkins J, He Y, Khandekar A, Smith-Byrne K, Carreira C, Gaborieau V, Latimer C, Thomas E, Abnizova I, Bucciarelli PE, Jones D, Teague JW, Abedi-Ardekani B, Serra S, Scoazec J-Y, Saffar H, Azmoudeh-Ardelan F, Sotoudeh M, Nikmanesh A, Eden M, Richman P, Campos LS, Fitzgerald RC, Ribeiro LF, Dzamalala C, Mmbaga BT, Shibata T, Menya D, Goldstein AM, Hu N, Malekzadeh R, Fazel A, McCormack V, McKay J, Perdomo S, Scelo G, Chanudet E, Humphreys L, Alexandrov LB, Brennan P, Stratton MR: **Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence**. *medRxiv* 2021:2021.2004.2029.21255920.

177. Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J, Luebeck J, Rajkumar U, Diao Y, Li B, Zhang W, Jameson N, Corces MR, Granja JM, Chen X, Coruh C, Abnousi A, Houston J, Ye Z, Hu R, Yu M, Kim H, Law JA, Verhaak RGW, Hu M, Furnari FB, Chang HY, Ren B, Bafna V, Mischel PS: **Circular ecDNA promotes accessible chromatin and high oncogene expression**. *Nature* 2019, **575**(7784):699-703.

178. Cheng AZ, Yockteng-Melgar J, Jarvis MC, Malik-Soni N, Borozan I, Carpenter MA, McCann JL, Ebrahimi D, Shaban NM, Marcon E, Greenblatt J, Brown WL, Frappier L, Harris RS: **Epstein-Barr virus BORF2 inhibits cellular APOBEC3B to preserve viral genome integrity**. *Nat Microbiol* 2019, **4**(1):78-88.

179. Poulain F, Lejeune N, Willemart K, Gillet NA: **Footprint of the host restriction factors APOBEC3 on the genome of human viruses**. *PLoS Pathog* 2020, **16**(8):e1008718.

180. Zhu B, Xiao Y, Yeager M, Clifford G, Wentzensen N, Cullen M, Boland JF, Bass S, Steinberg MK, Raine-Bennett T, Lee D, Burk RD, Pinheiro M, Song L, Dean M, Nelson CW, Burdett L, Yu K, Roberson D, Lorey T, Franceschi S, Castle PE, Walker J, Zuna R, Schiffman M, Mirabello L: **Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance**. *Nat Commun* 2020, **11**(1):886.

181. Amemiya HM, Kundaje A, Boyle AP: **The ENCODE Blacklist: Identification of Problematic Regions of the Genome**. *Sci Rep* 2019, **9**(1):9354.

182. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA: **The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers**. *Nat Rev Cancer* 2018, **18**(11):696-705.

183. Deshpande V, Luebeck J, Nguyen ND, Bakhtiari M, Turner KM, Schwab R, Carter H, Mischel PS, Bafna V: **Exploring the landscape of focal amplifications in cancer using AmpliconArchitect**. *Nat Commun* 2019, **10**(1):392.

184. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, Garraway LA: **The genomic complexity of primary human prostate cancer**. *Nature* 2011, **470**(7333):214-220.

185. Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K, Koboldt DC, Fulton RS, Delehaunty KD, McGrath SD, Fulton LA, Locke DP, Magrini VJ, Abbott RM, Vickery TL, Reed JS, Robinson JS, Wylie T, Smith SM, Carmichael L, Eldred JM, Harris CC, Walker J, Peck JB, Du F, Dukes AF, Sanderson GE, Brummett AM, Clark E, McMichael JF, Meyer RJ, Schindler JK, Pohl CS, Wallis JW, Shi X, Lin L, Schmidt H, Tang Y, Haipek C, Wiechert ME, Ivy JV, Kalicki J, Elliott G, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson MA, Baty J, Heath S, Shannon WD, Nagarajan R, Link DC, Walter MJ, Graubert TA, DiPersio JF, Wilson RK, Ley TJ: **Recurring mutations found by sequencing an acute myeloid leukemia genome**. *N Engl J Med* 2009, **361**(11):1058-1066.

186. Tao Y, Ruan J, Yeh SH, Lu X, Wang Y, Zhai W, Cai J, Ling S, Gong Q, Chong Z, Qu Z, Li Q, Liu J, Yang J, Zheng C, Zeng C, Wang HY, Zhang J, Wang SH, Hao L, Dong L, Li W, Sun M, Zou W, Yu C, Li C, Liu G, Jiang L, Xu J, Huang H, Li C, Mi S, Zhang B, Chen B, Zhao W, Hu S, Zhuang SM, Shen Y, Shi S, Brown C, White KP, Chen DS, Chen PJ, Wu CI: **Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data**. *Proc Natl Acad Sci U S A* 2011, **108**(29):12042-12047.

187. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S: **Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution**. *Nature* 2009, **461**(7265):809-813.

188. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, Ha C, Johnson S, Kennemer MI, Mohan S, Nazarenko I, Watanabe C, Sparks AB, Shames DS, Gentleman R, de Sauvage FJ, Stern H, Pandita A, Ballinger DG, Drmanac R, Modrusan Z, Seshagiri S, Zhang Z: **The mutation spectrum revealed by paired genome sequences from a lung cancer patient**. *Nature* 2010, **465**(7297):473-477.

189. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, Onofrio R, Pugh TJ, Rajkumar SV, Ramos AH, Siegel DS, Sivachenko A, Stewart AK, Trudel S, Vij R, Voet D, Winckler W, Zimmerman T, Carpten J, Trent J, Hahn WC, Garraway LA, Meyerson M, Lander ES, Getz G, Golub TR: **Initial genome sequencing and analysis of multiple myeloma**. *Nature* 2011, **471**(7339):467-472.

190. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, McLellan MD, Fulton RS, Fulton LL, Abbott RM, Hoog J, Dooling DJ, Koboldt DC, Schmidt H, Kalicki J, Zhang Q, Chen L, Lin L, Wendl MC, McMichael JF, Magrini VJ, Cook L, McGrath SD, Vickery TL, Appelbaum E, Deschryver K, Davies S, Guintoli T, Lin L, Crowder R, Tao Y, Snider JE, Smith SM, Dukes AF, Sanderson GE, Pohl CS, Delehaunty KD, Fronick CC, Pape KA, Reed JS, Robinson JS, Hodges JS, Schierding W, Dees ND, Shen D, Locke DP, Wiechert ME, Eldred JM, Peck JB, Oberkfell BJ, Lolofie JT, Du F, Hawkins AE, O'Laughlin MD, Bernard KE, Cunningham M, Elliott G, Mason MD, Thompson DM, Jr., Ivanovich JL, Goodfellow PJ, Perou CM, Weinstock GM, Aft R, Watson M, Ley TJ, Wilson RK, Mardis ER: **Genome remodelling in a basal-like breast cancer metastasis and xenograft**. *Nature* 2010, **464**(7291):999-1005.

191. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome**. *Nature* 2008, **456**(7218):66-72.

192. John EM, Miron A, Gong G, Phipps AI, Felberg A, Li FP, West DW, Whittemore AS: **Prevalence of pathogenic BRCA1 mutation carriers in 5 US racial/ethnic groups**. *JAMA* 2007, **298**(24):2869-2876.

193. Malone KE, Daling JR, Doody DR, Hsu L, Bernstein L, Coates RJ, Marchbanks PA, Simon MS, McDonald JA, Norman SA, Strom BL, Burkman RT, Ursin G, Deapen D, Weiss LK, Folger S, Madeoy JJ, Friedrichsen DM, Suter NM, Humphrey MC, Spirtas R,

Ostrander EA: **Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 years**. *Cancer Res* 2006, **66**(16):8297-8308.

194. **Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group**. *Br J Cancer* 2000, **83**(10):1301-1308.

195. Lord CJ, Ashworth A: **The DNA damage response and cancer therapy**. *Nature* 2012, **481**(7381):287-294.

196. Venkitaraman AR: **Cancer suppression by the chromosome custodians, BRCA1 and BRCA2**. *Science* 2014, **343**(6178):1470-1475.

197. Eeckhoutte A, Houy A, Manie E, Reverdy M, Bieche I, Marangoni E, Goundiam O, Vincent-Salomon A, Stoppa-Lyonnet D, Bidard FC, Stern MH, Popova T: **ShallowHRD: detection of homologous recombination deficiency from shallow whole genome sequencing**. *Bioinformatics* 2020, **36**(12):3888-3889.

198. Sztupinszki Z, Diossy M, Krzystanek M, Reiniger L, Csabai I, Favero F, Birkbak NJ, Eklund AC, Syed A, Szallasi Z: **Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer**. *NPJ Breast Cancer* 2018, **4**:16.

199. Gulhan DC, Lee JJ, Melloni GEM, Cortes-Ciriano I, Park PJ: **Detecting the mutational signature of homologous recombination deficiency in clinical samples**. *Nat Genet* 2019, **51**(5):912-919.

200. Weymann D, Laskin J, Roscoe R, Schrader KA, Chia S, Yip S, Cheung WY, Gelmon KA, Karsan A, Renouf DJ, Marra M, Regier DA: **The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers**. *Mol Genet Genomic Med* 2017, **5**(3):251-260.

201. Schwarze K, Buchanan J, Fermont JM, Dreau H, Tilley MW, Taylor JM, Antoniou P, Knight SJL, Camps C, Pentony MM, Kvikstad EM, Harris S, Popitsch N, Pagnamenta AT, Schuh A, Taylor JC, Wordsworth S: **The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom**. *Genet Med* 2020, **22**(1):85-94.

202. Popejoy AB, Fullerton SM: **Genomics is failing on diversity**. *Nature* 2016, **538**(7624):161-164.

203. Spratt DE, Chan T, Waldron L, Speers C, Feng FY, Ogunwobi OO, Osborne JR: **Racial/Ethnic Disparities in Genomic Sequencing**. *JAMA Oncol* 2016, **2**(8):1070-1074.

204. Copeland G, Lake A, Firth R: **Cancer in North America: 2005-2009. Volume One: Combined Cancer Incidence for the United States, Canada and North America. Springfield, IL: North American Association of Central Cancer Registries**. *Inc June* 2012.

205. Howlader N, Noone A-M, Krapcho M, Garshell J, Neyman N, Altekruse S, Kosary C, Yu M, Ruhl J, Tatalovich Z: **SEER cancer statistics review, 1975–2010**. *National Cancer Institute* 2014.

206. Yedjou CG, Sims JN, Miele L, Noubissi F, Lowe L, Fonseca DD, Alo RA, Payton M, Tchounwou PB: **Health and Racial Disparity in Breast Cancer**. *Adv Exp Med Biol* 2019, **1152**:31-49.

207. Hirschman J, Whitman S, Ansell D: **The black:white disparity in breast cancer mortality: the example of Chicago**. *Cancer Causes Control* 2007, **18**(3):323-333.

208. Oh SS, Galanter J, Thakur N, Pino-Yanes M, Barcelo NE, White MJ, de Bruin DM, Greenblatt RM, Bibbins-Domingo K, Wu AH, Borrell LN, Gunter C, Powe NR, Burchard EG: **Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled**. *PLoS Med* 2015, **12**(12):e1001918.

209. Haas JS, Hill DA, Wellman RD, Hubbard RA, Lee CI, Wernli KJ, Stout NK, Tosteson AN, Henderson LM, Alford-Teaster JA, Onega TL: **Disparities in the use of screening magnetic resonance imaging of the breast in community practice by race, ethnicity, and socioeconomic status**. *Cancer* 2016, **122**(4):611-617.

210. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G: **Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data**. *JAMA Internal Medicine* 2018, **178**(11):1544-1547.

211. Chen IY, Joshi S, Ghassemi M: **Treating health disparities with artificial intelligence**. *Nature Medicine* 2020, **26**(1):16-17.

212. Davis M, Martini R, Newman L, Elemento O, White J, Verma A, Datta I, Adrianto I, Chen Y, Gardner K, Kim HG, Colomb WD, Eltoum IE, Frost AR, Grizzle WE, Sboner A, Manne U, Yates C: **Identification of Distinct Heterogenic Subtypes and Molecular Signatures Associated with African Ancestry in Triple Negative Breast Cancer Using Quantified Genetic Ancestry Models in Admixed Race Populations**. *Cancers (Basel)* 2020, **12**(5).

213. Pitt JJ, Riester M, Zheng Y, Yoshimatsu TF, Sanni A, Oluwasola O, Veloso A, Labrot E, Wang S, Odetunde A, Ademola A, Okedere B, Mahan S, Leary R, Macomber M, Ajani M, Johnson RS, Fitzgerald D, Grundstad AJ, Tuteja JH, Khramtsova G, Zhang J, Sveen E, Hwang B, Clayton W, Nkwodimmah C, Famooto B, Obasi E, Aderoju V, Oludara M, Omodele F, Akinyele O, Adeoye A, Ogundiran T, Babalola C, MacIsaac K, Popoola A, Morrissey MP, Chen LS, Wang J, Olopade CO, Falusi AG, Winckler W, Haase K, Van Loo P, Obafunwa J, Papoutsakis D, Ojengbede O, Weber B, Ibrahim N, White KP, Huo D, Olopade OI, Barretina J: **Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features**. *Nat Commun* 2018, **9**(1):4181.

214. Qian J, Nie W, Lu J, Zhang L, Zhang Y, Zhang B, Wang S, Hu M, Xu J, Lou Y, Dong Y, Niu Y, Yan B, Zhong R, Zhang W, Chu T, Zhong H, Han B: **Racial differences in characteristics and prognoses between Asian and white patients with nonsmall cell lung cancer receiving atezolizumab: An ancillary analysis of the POPLAR and OAK studies**. *Int J Cancer* 2020, **146**(11):3124-3133.

215. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, Levin AM, Eng C, Yazdanbakhsh M, Wilson JG, Marrugo J, Lange LA, Williams LK, Watson H, Ware LB, Olopade CO, Olopade O, Oliveira RR, Ober C, Nicolae DL, Meyers DA, Mayorga A, Knight-Madden J, Hartert T, Hansel NN, Foreman MG, Ford JG, Faruque MU, Dunston GM, Caraballo L, Burchard EG, Bleecker ER, Araujo MI, Herrera-Paz EF, Campbell M, Foster C, Taub MA, Beaty TH, Ruczinski I, Mathias RA, Barnes KC, Salzberg SL: **Assembly of a pan-genome from deep sequencing of 910 humans of African descent**. *Nat Genet* 2019, **51**(1):30-35.

216. Abul-Husn NS, Soper ER, Braganza GT, Rodriguez JE, Zeid N, Cullina S, Bobo D, Moscati A, Merkelson A, Loos RJF, Cho JH, Belbin GM, Suckiel SA, Kenny EE: **Implementing genomic screening in diverse populations**. *Genome Med* 2021, **13**(1):17.

217. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN: **Deep learning in cancer pathology: a new generation of clinical biomarkers**. *Br J Cancer* 2021, **124**(4):686-696.

218. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ: **Clinical-grade computational pathology using weakly supervised deep learning on whole slide images**. *Nat Med* 2019, **25**(8):1301-1309.

219. Telli ML, Timms KM, Reid J, Hennessy B, Mills GB, Jensen KC, Szallasi Z, Barry WT, Winer EP, Tung NM, Isakoff SJ, Ryan PD, Greene-Colozzi A, Gutin A, Sangale Z, Iliev D, Neff C, Abkevich V, Jones JT, Lanchbury JS, Hartman AR, Garber JE, Ford JM, Silver DP, Richardson AL: **Homologous Recombination Deficiency (HRD) Score Predicts Response to Platinum-Containing Neoadjuvant Chemotherapy in Patients with Triple-Negative Breast Cancer**. *Clin Cancer Res* 2016, **22**(15):3764-3773.

220. Yarin Gal ZG: **Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning**. *International Conference on Machine Learning* 2016, **48**.

221. (CPTAC) NCICPTAC: **The Clinical Proteomic Tumor Analysis Consortium Breast Invasive Carcinoma Collection (CPTAC-BRCA) (Version 1) [Data set].** *The Cancer Imaging Archive* 2020.

222. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Graf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Group M, Langerod A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Borresen-Dale AL, Brenton JD, Tavare S, Caldas C, Aparicio S: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups**. *Nature* 2012, **486**(7403):346-352.

223. Galland L, Ballot E, Mananet H, Boidot R, Lecuelle J, Albuisson J, Arnould L, Desmoulins I, Mayeur D, Kaderbhai C, Ilie S, Hennequin A, Bergeron A, Derangere V, Ghiringhelli F, Truntzer C, Ladoire S: **Efficacy of platinum-based chemotherapy in metastatic breast cancer and HRD biomarkers: utility of exome sequencing**. *NPJ Breast Cancer* 2022, **8**(1):28.

224. Takaya H, Nakai H, Takamatsu S, Mandai M, Matsumura N: **Homologous recombination deficiency status-based classification of high-grade serous ovarian carcinoma**. *Sci Rep* 2020, **10**(1):2757.

225. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F: **The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository**. *J Digit Imaging* 2013, **26**(6):1045-1057.

226. Cancer Genome Atlas N: **Comprehensive molecular portraits of human breast tumours**. *Nature* 2012, **490**(7418):61-70.

227. Abkevich V, Timms KM, Hennessy BT, Potter J, Carey MS, Meyer LA, Smith-McCune K, Broaddus R, Lu KH, Chen J, Tran TV, Williams D, Iliev D, Jammulapati S, FitzGerald LM, Krivak T, DeLoia JA, Gutin A, Mills GB, Lanchbury JS: **Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer**. *Br J Cancer* 2012, **107**(10):1776-1782.

228. Birkbak NJ, Wang ZC, Kim JY, Eklund AC, Li Q, Tian R, Bowman-Colin C, Li Y, Greene-Colozzi A, Iglehart JD, Tung N, Ryan PD, Garber JE, Silver DP, Szallasi Z, Richardson AL: **Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents**. *Cancer Discov* 2012, **2**(4):366-375.

229. Popova T, Manie E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, Delattre O, Sigal-Zafrani B, Bollet M, Longy M, Houdayer C, Sastre-Garau X, Vincent-Salomon A, Stoppa-Lyonnet D, Stern MH: **Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation**. *Cancer Res* 2012, **72**(21):5454-5462.

230. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, Szallasi Z, Eklund AC: **Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data**. *Ann Oncol* 2015, **26**(1):64-70.