Genomic mutational heterogeneity in cancer

Improved models and tools for driver gene detection

Martin Boström

Department of Medical Biochemistry and Cell Biology

Institute of Biomedicine

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2022

Cover illustration: Increased UV susceptibility due to ETS family transcription factor binding underlies promoter hotspot mutations in melanoma.

By Martin Boström. DNA pattern brush for Adobe Illustrator by James Hedberg used with permission.

Genomic mutational heterogeneity in cancer: Improved models and tools for driver gene detection

martin.bostrom@gu.se

ISBN 978-91-8009-578-5 (PRINT)
ISBN 978-91-8009-579-2 (PDF)

Printed in Borås, Sweden 2022
Printed by Stema Specialtryck AB

Till Antonia och Rufus

Genomic mutational heterogeneity in cancer

Improved models and tools for driver gene detection

Martin Boström

Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Cancer is a disease that is strongly related to evolution, as mutations that confer a benefit to individual cells face positive selection and eventually lead to tumorigenesis. As such, the search for genes that drive cancer development entails distinguishing positive selection from other sources of increased mutation rates, which requires detailed knowledge of how normal mutation rates vary across the genome. This thesis aims to improve that knowledge, as well as to provide novel methods of driver detection.

In cutaneous melanoma, there are mutational hotspots in promoters that coincide with the sequence motif "TTCCG". These hotspots could easily be misinterpreted as cancer drivers, but in the first paper of this thesis we show that they are in fact caused by increased UV damage susceptibility upon transcription factor binding, with some contribution from impaired DNA repair.

In the second paper, we study how the UV mutational signature varies between different genomic regions and show that the main difference is caused by the level of cytosine methylation, owing to its effect on UV damage formation. We also improve the traditional trinucleotide mutational signature by incorporating longer patterns, capturing the effect of TTCCG-related promoter mutations.

In the third paper, we demonstrate a novel method for driver detection that ignores recurrence signals, instead testing the likelihood of observing a particular combination of mutated tumours in a patient cohort. In addition to providing an orthogonal perspective on driver detection, this method is less sensitive to flaws in modelling some forms of mutational heterogeneity, such as the TTCCG hotspots.

In summary, this thesis improves our knowledge of mutational heterogeneity in cancer, in addition to describing a new driver detection test that is less sensitive to situations where that knowledge falls short. Both of these advances contribute to the search for genes that drive cancer development.

Keywords: Cancer, genomics, ultraviolet light, mutational heterogeneity

ISBN 978-91-8009-578-5 (PRINT)
ISBN 978-91-8009-579-2 (PDF)

SAMMANFATTNING PÅ SVENSKA

Cancer är en sjukdom med stark koppling till evolution. Mutationer som gynnar individuella celler utsätts för positiv selektion, och bidrar därmed till tumörbildning. Jakten efter gener som driver cancerutveckling innefattar därför att skilja på positiv selektion och andra orsaker till ökad mutationsfrekvens, vilket kräver detaljerad kännedom om hur mutationsfrekvensen normalt varierar i genomet. Den här avhandlingen ämnar öka den kunskapen, samt bidra med nya metoder för att hitta cancergener.

I malignt melanom i huden finns det positioner i genomet som har ovanligt hög mutationsfrekvens och som sammanfaller med sekvensmotivet ”TTCCG” i aktiva promotorer. Dessa positioner skulle kunna misstolkas som drivande för cancer, men i denna avhandlings första artikel visar vi att den underliggande orsaken är ökad känslighet för UV-relaterad skadebildning på DNA-molekylen vid bindning av transkriptionsfaktorer, med ett mindre bidrag från nedsatt DNA-reparation.

I den andra artikeln studerar vi hur mutationssignaturen från UV-ljus varierar mellan olika genomiska regioner. Vi visar att den största skillnaden är kopplad till metyleringsnivån av cytosin, på grund av effekten den har på UV-relaterad DNA-skadebildning. Vi förbättrar också den traditionella trinukleotidbaserade mutationssignaturen genom att inkorporera längre sekvenser, och får på så sätt med den mutationsökande effekten hos TTCCG-relaterade promotorregioner.

I den tredje artikeln demonstrerar vi en ny metod för detektion av cancerdrivande mutationer. Denna metod åsidosätter mutationsfrekvens, och utvärderar i stället sannolikheten att observera olika kombinationer av muterade tumörer. Metoden angriper detektion av cancerdrivande mutationer från en ny vinkel, och är dessutom mindre känslig för brister i underliggande mutationsmodeller, som annars leder till falska positiva resultat i regioner som de TTCCG-relaterade promotorerna.

Sammanfattningsvis bidrar den här avhandlingen till att öka kunskapen om mutationsheterogenitet i cancer, samt introducerar en ny metod för detektion av cancerdrivande mutationer. Dessa framsteg främjar jakten på gener som driver cancerutveckling.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals.

I. Elliott K*, Boström M*, Filges S, Lindberg M, Van den Eynden J, Ståhlberg A, Clausen A, Larsson E. **Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers**
PLOS Genetics 2018, 14(12)

* These authors contributed equally

II. Lindberg M, Boström M, Elliott K, Larsson E. **Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light**
PNAS 2019, 116 (41) 20411-20417

III. Boström M, Larsson E. **Mutation distribution skew in patient cohorts provides a novel signal for positive selection in cancer**
Manuscript

PAPERS NOT INCLUDED IN THIS THESIS

1. Kreisel K, Engqvist M, Kalm J, Thompson L, Boström M, Navarrete C, McDonald J, Larsson E, Woodgate R, Clausen A. **DNA polymerase η contributes to genome-wide lagging strand synthesis**
Nucleic Acids Res. 2019, 47(5):2425-2435

CONTENT

ABBREVIATIONS

6-4 PP	6-4 pyrimidine pyrimidone
A	Adenine
bp	Base pairs
C	Cytosine
COSMIC	Catalogue of Somatic Mutations in Cancer
CPD	Cyclobutane pyrimidine dimer
cSCC	Cutaneous squamous cell carcinoma
Cys	Cysteine
DNA	Deoxyribonucleic acid
dN/dS	Ratio between nonsynonymous and synonymous mutations
FDR	False discovery rate
G	Guanine
ICGC	International Cancer Genome Consortium
kb	Kilobases (Distance measurement in DNA – 1000 base pairs)
MMR	Mismatch repair
NER	Nucleotide excision repair
NMF	Nonnegative matrix factorisation
PCA	Principal component analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
RNA	Ribonucleic acid

SNP	Single-nucleotide polymorphism
SNV	Single-nucleotide variant
T	Thymine
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TFBS	Transcription factor binding site
TLS	Translesion synthesis
TSG	Tumour suppressor gene
TSS	Transcription start site
Tyr	Tyrosine
U	Uracil
UCEC	Uterine Corpus Endometrial Carcinoma
UV	Ultraviolet (light)
WGS	Whole genome sequencing
WXS	Whole exome sequencing
Y	Ambiguous base code for a pyrimidine, i.e., C or T

1 INTRODUCTION

For as long as there has been life, there has been evolution – the change across generations of heritable characteristics. Traits that are beneficial to the organism confer a selective advantage and are more likely to be passed on to the next generation. While evident in all life, we can observe this in real time most easily today in prokaryotes, where rapid generation cycles allow us to see the evolution of traits such as antibiotic resistance (1). In our early evolutionary history as single-cell organisms, positively selected traits would have included rate of proliferation, efficient metabolism, and countless others. As time progressed, multicellular life arose, and with it a stronger emphasis on the selective advantage of cooperation between cells, such as nutrient sharing and signalling. Today, humans and other highly complex organisms are composed of trillions of highly specialised cells (2), organised in different tissue types and organs. The selective pressure on the organism level entails different requirements for individual cells in the body than what single-cell organisms face.

And yet, the selective pressure that acts on single cells and started our journey toward becoming complex multicellular organisms never went away. While the cellular traits that are selected for in the human population emphasise cooperation and organisation, individual cells face very different selective pressures on the timescale of cellular generations. If a cell attains a trait that allows it to grow and divide faster than its neighbours, it may outcompete them and form a mass of descendant cells with uncontrolled growth (a tumour), if not hindered. Further selection of competitive advantages can result in the spread to other tissues through metastasis, eventually disrupting the functioning of the organism to the point of severe disease or death. This is cancer - a disease that is the product of positive selection in cell populations at the cost of the organism as a whole.

If the force behind cancer is the selective pressure that is simultaneously active in all of our trillions of cells, how can complex organisms exist in the first place? The answer is that we have evolved highly efficient defences against tumorigenesis. Cell proliferation is tightly regulated, and only allowed at the appropriate time. There are checkpoints in the cell cycle where cells with damaged DNA are forced to stop dividing, or even undergo apoptosis – controlled cell death (3). However, since cell proliferation is required for us to function, the cell cycle becomes a balancing act between allowing growth when needed and preventing tumorigenesis. Changes in cells that disrupt that balance in favour of tumorigenesis are what lead to cancer.

1.1 THE HALLMARKS OF CANCER

There are many changes in a cell that are positively selected for and can contribute to tumorigenesis, but typically most can be categorised according to what advantage they grant – or viewed another way, what anti-cancer defence they help the cell overcome. In 2000, Hanahan and Weinberg released a seminal review article describing six categories of biological capabilities (4), later updated to ten (5), that are common in cancer cells (Figure 1). These traits (marked in bold below) are typically attained through mutations, changes in the genetic code of the cell that can result in altered phenotypes.

In normal tissues, cell proliferation is controlled through modulation of growth-promoting signals. To attain the uncontrolled cell growth of cancer, **sustaining proliferative signalling** is critical. Bypassing the dependence on outer signals can be done in several ways. For instance, extra growth receptors can be expressed to achieve a stronger response to normal signalling (6), or the cell may cut out the middleman and produce its own growth signals (7). Alternatively, signalling can be sidestepped, at least in part, by constitutively activating the proteins downstream of the receptor (8). However, self-sufficiency in growth signals is in itself insufficient, as negative regulation of growth needs to be handled by **evading growth suppressors**, perhaps most commonly simply by inactivating them. Additionally, continuous growth requires nourishment, necessitating **sustained angiogenesis** (formation of new blood and lymphatic vessels) to provide oxygen and nutrients, as well as remove waste products (9).

When rampant cell growth or severe damage is detected in a cell, the most severe defensive response is triggering apoptosis, or controlled cell death. **Resisting cell death** is therefore another hallmark, achieved through various strategies, such as disrupting DNA damage sensors (10) or upregulating survival signals (11). Even with all these traits, unlimited proliferation is not possible without **enabling replicative immortality**. After a certain number of divisions, cells enter senescence, a state outside of the cell cycle with no further proliferation. Cells that manage to circumvent this eventually reach a state of crisis, due to the shortening of protective telomeres at the end of the chromosomes with each replication. This can be avoided by expressing telomerase, a DNA polymerase that lengthens telomeres, thereby staving off senescence and crisis-induced cell death (12).

A tumour that does not spread to other tissues is called benign. It is not until it has accomplished **tissue invasion and metastasis** that we call it cancerous. Eventually, most cancers metastasise, and the resulting invasion of other tissues is what leads to the vast majority of cancer deaths (4).

These six original hallmarks of cancer have since been joined by the **deregulation of cellular energetics**, which is an alteration of the metabolism of the cell to provide enough energy for sustained growth, and **avoidance of immune destruction**, by avoiding or hindering the response of the immune system to cancer cells. Finally, two hallmarks that facilitate attaining the others already listed are **genome instability**, where increased genomic alterations provide the potential for acquiring cancer-related traits, and **tumour-promoting inflammation**, where the immune system can paradoxically help tumorigenesis by supplying growth, survival, and angiogenic factors.

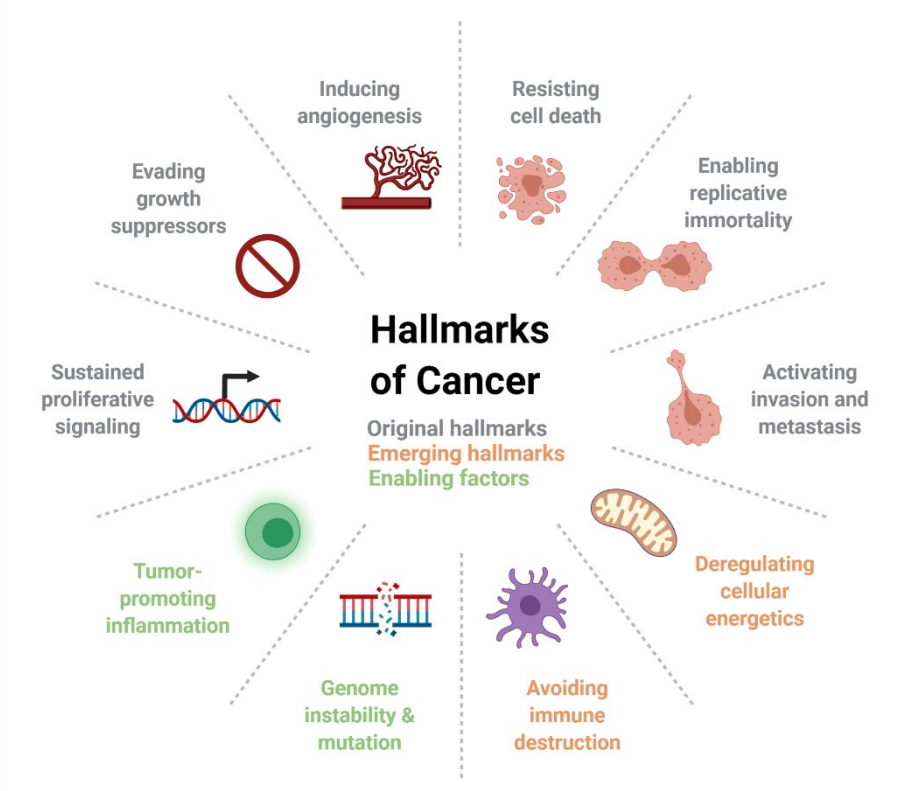

Figure 1. The Hallmarks of Cancer (5). Adapted from "Hallmarks of Cancer: Circle", by BioRender.com (2021). Retrieved from https://app.biorender.com/biorender-templates

1.2 MUTATIONS IN CANCER

The traits of cancer cells may be acquired in different ways. Some come from structural variations - genomic alterations involving segments of DNA larger than 1 kb (13). Examples include copy number variants, such as possessing multiple copies of DNA segments with oncogenic function, or deletions of those with protective roles (14). Viral integrations of DNA can also lead to cancer, with viruses implicated to some extent in 15-20 % of cancers, perhaps most notably exemplified by the human papilloma virus being detectable in nearly all cervical cancer (15).

In this thesis, the focus will be on single nucleotide variants (SNVs), where one nucleotide is exchanged for another. The effects of these mutations differ depending on the affected position in the genome. In protein-coding sequences, SNVs may alter protein function (Figure 2). As each subsequent triplet of nucleotides (or codon) in the protein-coding sequence of a gene encodes a certain amino acid, mutations that change the codon can alter what amino acid is incorporated during protein translation. SNVs that result in a different amino acid being encoded are called missense mutations, and they can dramatically alter the function of a protein if they occur in a critical position. SNVs may also introduce stop codons, which bring about the premature end of translation, often resulting in a non-functional protein. These are called nonsense mutations. Finally, an SNV can occur without a change in the encoded amino acid due to the degeneracy of the genetic code, resulting in a synonymous mutation.

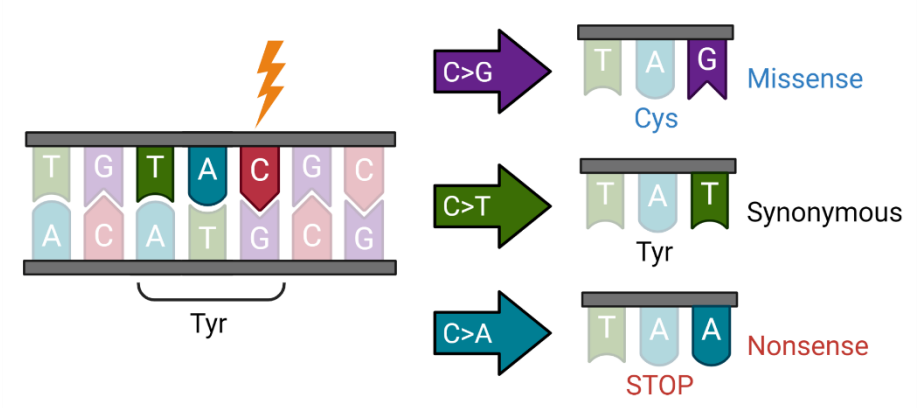

Figure 2. SNVs in coding sequences can have different effects on the encoded protein depending on the original and mutated codon. Created with BioRender.com.

Another class of mutation that commonly disrupts protein translation is the indel, where an insertion or a deletion of one or several nucleotides into a coding sequence can change which nucleotides belong to which codons in the

rest of the gene. These “frameshift” mutations normally result in a completely non-functional protein.

Cancer-related mutations can occur in either germline or somatic cells. Germline mutations occur in reproductive cells and can therefore be passed on to offspring, leading to hereditary cancer risk. Examples include inherited mutations in the *BRCA1* and *BRCA2* genes, which predispose to breast cancer (16). Essentially, a cancer-associated germline mutation can lower the bar for tumorigenesis by providing every cell in the body with an unfortunate head start. Mutations in somatic cells, by contrast, are not inherited. Through positive selection, cells with cancer-associated somatic mutations proliferate more than normal, resulting in a larger pool of cells where another beneficial somatic mutation can set off another wave of clonal expansion. In most cancers, tumour cells descend from a single cell with a cancer-associated somatic mutation (17).

1.2.1 ONCOGENES AND TUMOUR SUPPRESSORS

The kinds of somatic coding mutations that are selected for in tumour cells depend on the gene’s role in tumorigenesis. Genes that contribute to cellular growth are called oncogenes, and those that prevent it, thereby guarding against tumorigenesis, are called tumour suppressor genes (TSGs). Strictly speaking, an oncogene is the mutated form of a proto-oncogene, where the latter is the normal variant that merely has the potential to become oncogenic after attaining a tumorigenic mutation.

The type of mutation that is selected for in a TSG tends to be disruptive to protein function, as the outcome under positive selection is typically disabling or impairing the role of the protein. As such, nonsense and frameshift mutations are common, due to the fact that they often result in completely non-functional proteins. The most frequently mutated TSG is *TP53 (18)*, dubbed “the guardian of the genome”, which is mutated in more than 50 % of cancers (19). p53, the protein encoded by *TP53*, is heavily involved in arresting growth and inducing senescence or apoptosis in cells that exhibit DNA damage, shortened telomeres, or excessive activation of growth pathways, to name a few of its functions (20, 21).

Since oncogenes contribute to cancer growth, their mutations tend to activate them in some way. This can happen through mutations that reduce sensitivity to negative feedback, or that enhance the function or expression of the protein. Ras proteins are some of the most important oncogenes in cancer, with approximately 30 % of tumours containing some kind of Ras-activating mutation (22). Various forms of Ras are involved in signalling pathways to induce proliferation, such as the PI3K and Ras-Raf-MAPK pathways, where they contribute to the signalling cascade by activating downstream proteins (23). Several sites in Ras proteins are hotspots for mutations in different cancer

types, as they result in constitutive activation, causing constant activation of downstream proteins (8).

While the categorisation of genes as oncogenes or TSGs can be helpful, genes can have both roles at once, as exemplified by *TP53* (24). Some missense mutations in p53 not only hinder its tumour-suppressing capabilities, but also give it oncogenic function by allowing it to bind to and hinder its tumour-suppressing homologues p63 and p73, thereby promoting tumour invasion and metastasis (25).

1.2.2 NON-CODING MUTATIONS

Mutations that contribute to cancer are not limited to DNA sequences that encode proteins. In 2013, two seminal papers demonstrated that point mutations in the promoter region of the *TERT* gene formed new transcription factor binding sites (TFBSs), thereby increasing its expression (26, 27) (Figure 3). *TERT* encodes the reverse transcriptase subunit of telomerase, meaning its increased expression helps enable replicative immortality, as discussed in section 1.1. This discovery started a search for additional non-coding mutations, yet despite the promising start with the *TERT* mutations, few cancer-driving mutations have been found (28).

Since expression level changes of cancer genes are often seen in cancer, there are many elements in non-coding DNA that could plausibly host tumorigenic mutations, such as enhancers (29) and regulatory RNAs (30, 31). Some technical aspects could partially explain why so few non-coding cancer mutations in these elements have been found, such as the smaller amount of available whole-genome sequencing (WGS) data compared to whole-exome sequencing (WXS), and the reduced coverage often seen when sequencing regulatory regions (28, 32). On a biological level, the robustness of regulatory sequences to point mutations, unlike coding sequences, could provide another explanation (33).

While the relative lack of high-recurrence non-coding mutations is unlikely to change, the field is young compared to the analysis of coding sequences, and new discoveries are therefore still to be expected as new techniques and datasets become available (28).

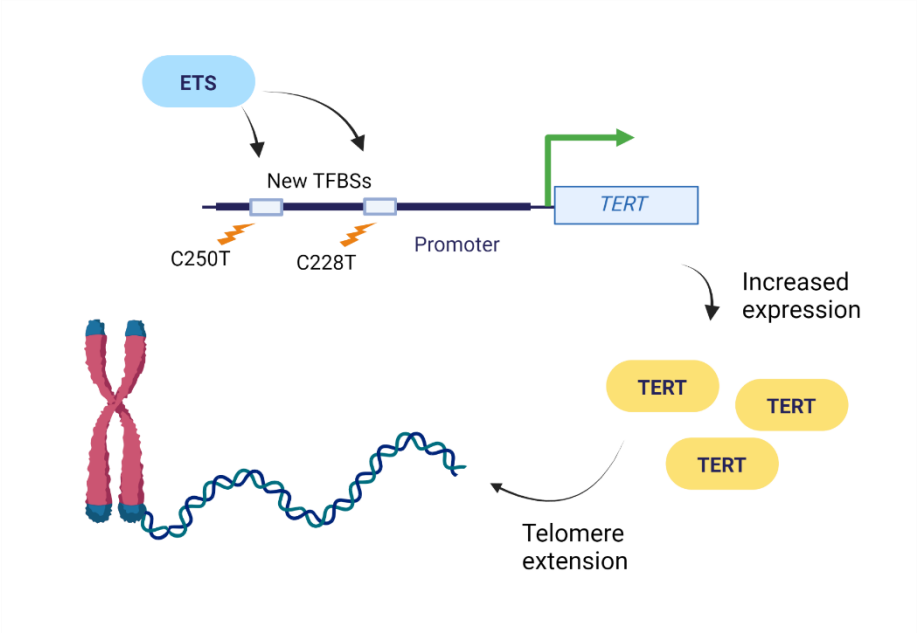

Figure 3. TERT promoter mutations create new binding sites for ETS transcription factors, increasing expression and thereby enabling telomere extension. Created with BioRender.com

1.3 WHAT CAUSES MUTATIONS?

The processes behind mutation formation differ greatly depending on the type of mutation. The causes of gene duplication, for instance, are not the same as those of point mutations, which are what we will focus on here. Mutations can be either induced, meaning they are caused by environmental factors, or spontaneous, i.e., resulting from normal or defective processes in the cell. Often, the initial mutagenic event is environmentally caused damage to the DNA, but with cellular processes leading to the actual mutation (34). Well-known examples of environmental mutagens include exposure to ultraviolet (UV) light and tobacco smoke (35). Major sources of mutations on the internal side are errors during DNA replication and repair. These can occur through simply incorporating the wrong nucleotide, as the error rate of fully functional DNA polymerases is not zero. Various factors can increase the error rate of DNA replication and repair, such as hereditary DNA repair defects and damage to the DNA due to external mutagens (34, 36). Other internal mutagenic factors include DNA damage from oxidative stress (37) and spontaneous nucleotide changes from chemical processes (38).

1.3.1 MUTATIONS IN CUTANEOUS MELANOMA

As an example of how several processes can be involved in mutagenesis, we will examine how UV-induced DNA damage causes mutations. UV-induced damage primarily causes C>T transitions and is the main source of mutations in cutaneous melanoma, a cancer type that is of particular interest in this thesis. Normally, pyrimidines base-pair with purines (C with G and T with A) on the opposite strand. When UV light hits the DNA molecule, the absorbed energy can lead to the formation of bonds between adjacent pyrimidines (C or T) on the same strand (Figure 4). The most notable photoproducts formed this way are cyclobutane pyrimidine dimers (CPDs) and 6-4 pyrimidine pyrimidones (6-4 PPs), with CPDs being the most numerous. These bulky lesions can be repaired by nucleotide excision repair (NER), where the damaged DNA is removed, and the resulting gap is filled using the complement strand as a template. However, if replication of the affected DNA is attempted before repair is finished (or if NER is defective, as in those afflicted with xeroderma pigmentosum (39)), the replication fork would stall, potentially leading to fatal double-strand breaks. To avoid this, the cell can attempt to continue replication past the DNA damage using special DNA polymerases that perform translesion synthesis (TLS). It is in this step that most UV-induced mutations arise, through one of two different models (40) (Figure 5).

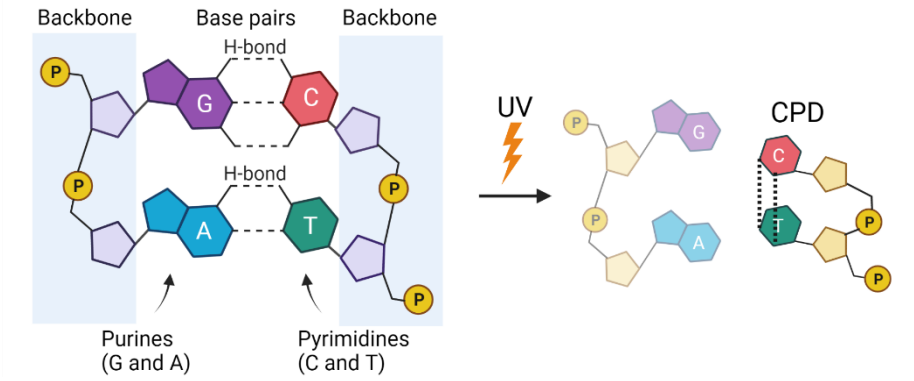

Figure 4. Regular base-paired bases, with CPD formation following UV exposure. Created with BioRender.com

In the first model, an error-prone TLS polymerase inserts an adenine opposite the lesion, resulting in C>T mutations. Adenine appears to be the most frequently inserted base by the error-prone polymerase, and it is also the base that is the easiest to extend the DNA strand from when opposite a lesion, for some polymerases. The second model relies on the fact that cytosine can spontaneously deaminate into uracil, a process that is much more prone to occurring in cytosines that are part of CPDs than when regularly base-paired. Uracil is the RNA equivalent of thymine and base-pairs with adenine rather than guanine, just like thymine. During TLS, polymerase η correctly pairs an adenine with the uracil, resulting in a C>T mutation after replication. Through the same process, CC>TT mutations can occur when a CPD is formed between two cytosines that both deaminate (40).

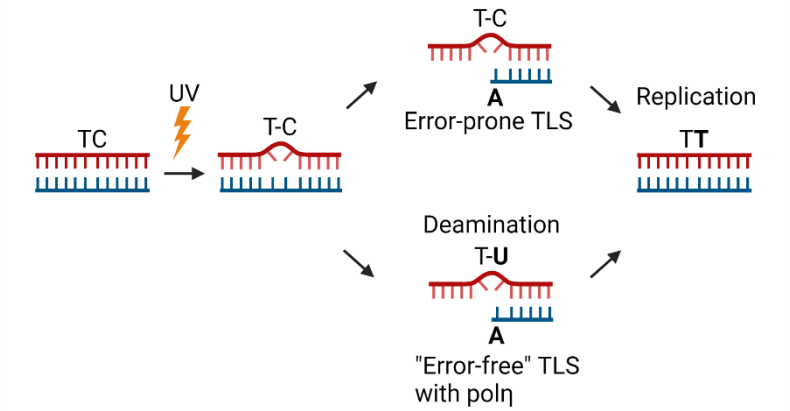

Figure 5. UV-induced CPD formation can cause C>T mutations through error-prone TLS, or through spontaneous cytosine deamination followed by error-free TLS with polη. Created with BioRender.com

Mutagenesis is further affected by the presence or absence of a G following a dipyrimidine with a C in the second position (i.e., YCG, where Y indicates a pyrimidine). CpG sites are highly methylated in the genome, and methylated cytosines form CPDs more readily than non-methylated cytosines under UVB exposure (41, 42), the wavelength of light responsible for most of the mutations caused by sunlight (40). This is an example of how neighbouring bases can affect the mutagenicity of a genomic site.

1.3.2 MUTATIONAL SIGNATURES

The highly specific mutation types generated by UV light, along with the effect of neighbouring bases, forms a kind of fingerprint or signature by which the origin of the mutations can be discerned. While the UV signature tends to be quite dominant in cutaneous melanoma, other cancer types are more accurately characterised as mosaics of different mutational processes, presenting the problem of how to separate those processes in a cohort of tumours in which they may be active to different degrees. In 2012, Nik-Zainal et al. published a landmark paper proposing a method for extracting signatures using nonnegative matrix factorisation (NMF) and applied it to a cohort of 21 breast cancer genomes (43). In this method, mutations are classified according to their substitution type (C>A, C>G, C>T, T>A, T>C, or T>G) and the immediate neighbouring bases, for a total of 96 combinations. Through NMF, different trinucleotide signatures could be separated in the cohort, and the degree to which the causative mutational processes were active in the different tumours could be discerned.

Following up on this discovery, another paper was published applying the method to more genomes in different cancer types, establishing a catalogue of signatures that are present in tumours (44). In this catalogue, the dominant signature in cutaneous melanoma matched previously known facts about UV mutagenesis well, being almost exclusively C>T mutations at dipyrimidines (Figure 6). The preference for YCG trinucleotides is also captured by the signature, albeit slightly obscured due to the fact that the traditional representation of the signatures is not normalised by the frequency with which the trinucleotides occur in the genome. The CG dinucleotide is quite rare in most of the human genome, with the notable exception of promoter regions, owing to methylation and subsequent spontaneous deamination causing C>T mutations, resulting in CpG depletion over evolutionary time (45).

While the link between signature 7 and UV radiation is quite clear, many signatures have unknown aetiology, as signature extraction does not inherently provide any information about the underlying process. Other signatures with known aetiology include signature 4, which is related to tobacco smoking (Figure 6). Tobacco-induced mutations occur through bulky adducts on the DNA molecule that, similar to UV-induced mutations, need to be repaired with NER or bypassed with TLS, but other differences cause the signatures to look

completely different, with most of the signature 4 mutations being C>A instead of C>T (46).

Since the original publication of the compendium of signatures, new signatures have been added, and some signatures have been split into multiple components. The UV-caused signature 7 belongs to the latter category, having been split into four signatures (7a-d) in the latest release of COSMIC. There is some uncertainty about the processes behind each of these signatures.

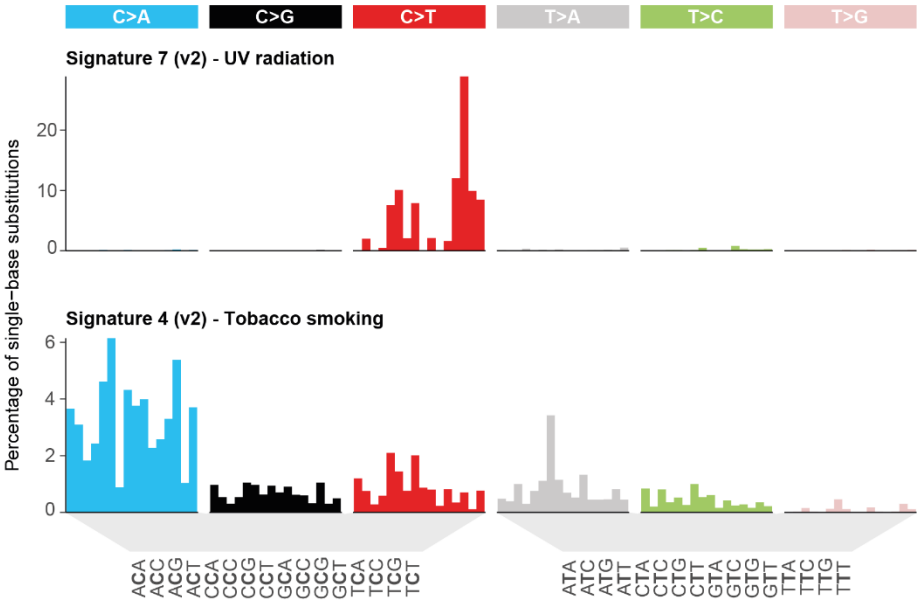

Figure 6. Signatures associated with UV radiation and tobacco smoking for comparison. Each bar shows the percentage of single-base substitutions belonging to the given substitution type and trinucleotide context for the mutational process in the human genome. The signature data was downloaded from cancer.sanger.ac.uk/signatures (version 2).

1.4 FINDING DRIVERS IN GENOMIC DATA

Discovering new cancer genes through bioinformatical methods involves sifting through somatic mutation data from tumour-normal pairs looking for mutations that are under positive selection - commonly referred to as "drivers". Most tumours contain a handful of known driver mutations, but these are vastly outnumbered by "passenger" mutations that do not contribute to cancer (47, 48). Separating the drivers from the passengers is a big area of research in cancer genomics, with the goal of gaining a better understanding of cancer biology and ultimately finding new drug targets.

The most basic approach in the search for cancer genes is to look at the recurrence of mutations in tumours. If a mutation is beneficial to a cancer cell, it stands to reason that it would be encountered more frequently in patients as a result of positive selection. Some driver mutations in the very strongest cancer genes are so recurrent that this approach is feasible without further complications. As previously mentioned, *TP53* is mutated in more than 50 % of tumours, providing strong support for *TP53* mutations as drivers. For most drivers, however, detailed mutational models are necessary to determine whether a mutation is under positive selection or not. This is not only so that weakly recurrent driver mutations may be found, but also to avoid false positives due to normal mutation rates that are underestimated and misinterpreted as a sign of positive selection (Figure 7).

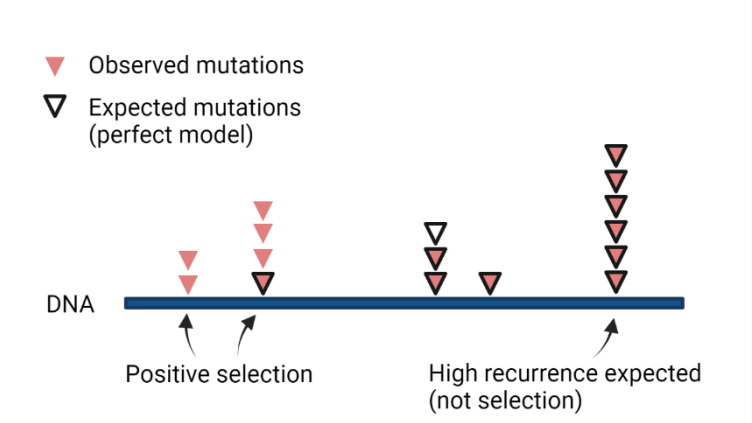

Figure 7. Accurate mutational models are required to tell whether a highly recurrent mutation is caused by normal mutational processes, or whether it is under positive selection. Created with BioRender.com.

A decade ago, most tools aimed at finding cancer genes assumed a flat background mutation rate based on the average mutation rate in the cancer type, with only some adjustments for the relative frequencies of different mutation types (49). In a paper describing the new driver detection tool MutSigCV, Lawrence et al. showed that the lack of sophistication in the mutational models in use at the time was becoming a problem, with more and more false positives among the putative cancer genes reported by tools as cohort sizes grew (49). Many olfactory receptor genes, a group not particularly plausible as cancer genes to begin with, were reported as significantly mutated in cancer types where they were not even expressed. The problem with the models was attributed to genomic mutational heterogeneity, or how mutational processes induce mutations at varying rates across the genome. In MutSigCV, gene expression levels and replication timing were included in the mutational model to account for the variance in the background mutation rate, with excellent results. In the example of olfactory receptors, low expression levels and late replication timing explained why their mutation rates were higher than previously expected. Since the release of MutSigCV, many more different sources of mutation rate variation have been characterised and included in the mutational models of driver detection tools.

1.4.1 GENOMIC MUTATIONAL HETEROGENEITY

The background mutation rate is affected by phenomena that are active at different scales, from the megabase level all the way down to single nucleotides (50, 51). We have already explored how different mutational processes leave characteristic imprints on the genome through their trinucleotide signatures. Even without extracting signatures from a tumour cohort, a simple trinucleotide-based mutational model goes a long way toward modelling the mutation rate of the active mutational processes on the single-nucleotide scale. What it misses are effects related to longer sequences, and variations in different genomic regions, which we will cover here.

CHROMATIN

To fit the DNA molecule inside a cell, compact packing without entanglement is required. To accomplish this, DNA is wound twice around octamers of histone proteins. Each such DNA-histone complex is called a nucleosome (technically a nucleosome core), and they are separated from each other by stretches of unwound DNA. This first loose level of packing is called euchromatin. The nucleosomes also allow for tighter packing of DNA, known as heterochromatin. Which regions of DNA are in different forms of chromatin can change depending on when access is required, for instance during replication and for gene expression.

The chromatin structure of DNA has an effect on mutational heterogeneity on both small and large scales, and it interacts in different ways with various

mutational processes. On the megabase scale, heterochromatin accumulates more mutations than more loosely packed DNA (52), largely due to restricted accessibility for DNA repair mechanisms (53) (Figure 8). The same concept holds true on the scale of individual nucleosomes, where the DNA linking DNA-histone complexes together is more easily accessible to repair than the DNA wrapped around the histones (54). On the other hand, some mutational processes are also hampered by tightly packed DNA. Spontaneous deamination of methylated cytosines is reduced in nucleosomes, leading to fewer C>T mutations caused by this process (55). While the overall mutation rate is still generally increased in tightly packed DNA, it is important to remember that the mutational processes active in our genomes are affected in different ways by features such as chromatin structure.

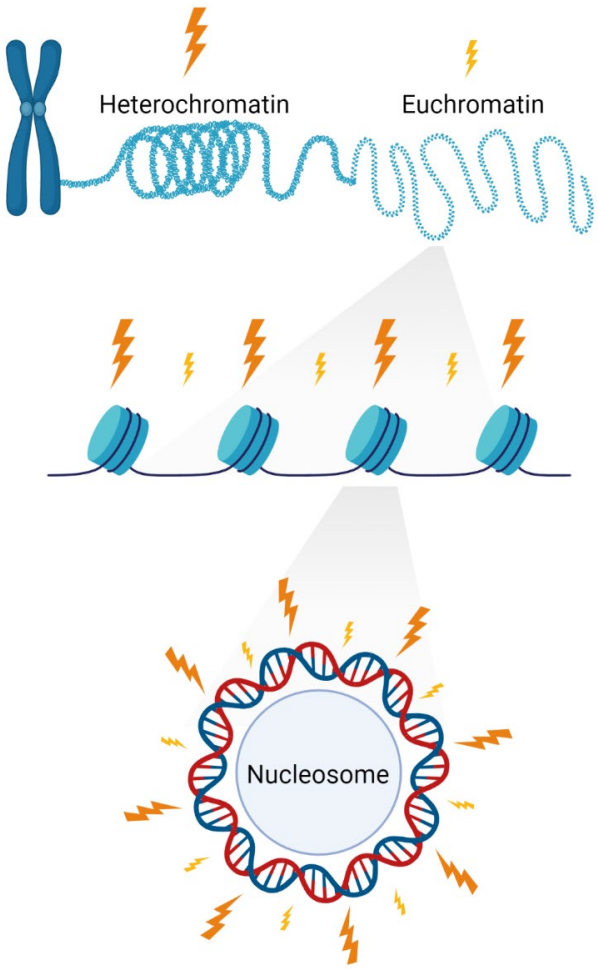

Figure 8. The effect of chromatin structures on the background mutation rate at different scales. Created with BioRender.com

Even within the DNA wound around the histone proteins in a nucleosome, there is mutational heterogeneity. This is related to whether the major or the minor groove of the DNA molecule is positioned outward, resulting in a periodicity in mutation rate due to the way the DNA is wound around the histones. UV-induced CPD formation, which is otherwise unaffected by chromatin state, has been shown to correlate with this periodicity in a way that cannot be explained by sequence context, and DNA repair accessibility is also affected (56).

EXPRESSION AND REPLICATION TIMING

The level of gene expression is anticorrelated with mutation rate. This is strongly related to the aforementioned chromatin effects, as high expression requires loose packing and vice versa. There are also expression-related repair mechanisms at play, resulting in lower mutation rates in exons (57), and particularly on the transcribed strand (58, 59). The strand bias is not only caused by differential repair, as damage formation has been observed to increase on the coding strand (59, 60).

Replication timing is similarly related to chromatin state, with early-replicating regions receiving fewer mutations than late-replicating regions. This has been attributed to differential mismatch repair (MMR), where early-replicating euchromatin regions are repaired more effectively (61). Another possible contributing factor is the depletion of the nucleotide pool towards the end of replication, which could result in increased mutation rates as vulnerable single-strand DNA is exposed for longer as the replication fork slows or stalls (62). Replication-related strand bias has also been reported in mutational signatures both related and unrelated to DNA repair (60, 63, 64).

TRANSCRIPTION FACTOR BINDING SITES

Several studies have shown that mutation rates are increased at transcription factor binding sites (TFBSs) in some cancer types, most notably in melanoma (65-67). The main cause for this appears to be bound TFs blocking access for NER. This is backed up by the observations that the effect is observed in active, but not inactive, TFBSs, indicating that TF binding is required, and that xeroderma pigmentosum patients, who have deficient NER, do not show the same pattern. Furthermore, NER maps show reduced repair at bound TFBSs following UV damage formation (65, 66). The fact that the phenomenon is most notable in melanoma is consistent with the extensive use of NER to repair UV-induced damage. Similarly, the bulky adducts caused by tobacco smoking, also repaired through NER, explain why the mutation rate at TFBSs is also increased in lung cancer (66).

Differential DNA damage formation has also been reported at occupied TFBSs (56, 68, 69). Unlike NER, where TF binding mostly blocks access, the effect

of a bound TF appears to increase or decrease damage formation in a manner that is dependent on both the mutagen and the TF. For instance, UV-induced CPD formation at active TFBSs is reduced in some groups of TFs and increased in others.

TTCCG-RELATED PROMOTER MUTATIONS IN MELANOMA

In cutaneous melanoma, there are a number of promoters with highly recurrent mutations at specific sites. Mutations in the *TERT* promoter are confirmed drivers as discussed in section 1.2.2, but the other recurrent promoter mutations do not appear to be under positive selection. Previous explanations for this phenomenon tend to focus on differential DNA repair due to TF binding (65), as described above.

In 2017, Fredriksson et al. published a paper highlighting the fact that very nearly all of the recurrent promoter mutations in melanoma, except the *TERT* mutations, occurred in or immediately upstream of the sequence TTCCG (70) (Table 1), a motif matching the consensus binding sequence of the ETS family of transcription factors (71). Interestingly, the TTCCG-related mutations occurred only in UV-related cancers and sun-exposed skin, suggesting UV mutagenesis rather than positive selection as the cause. In further support of this explanation, the number of hotspot mutations in each tumour correlated with that tumour's mutation burden. The mutations could even be induced in cells following UV exposure.

In light of the fact that both differential repair and damage formation have been observed after transcription factor binding (56, 65-69), either or both could be the cause of the recurrent mutations. Fredriksson et al. noted that UV-exposed tumours lacking global NER were still mutated in the hotspots, albeit to a lesser degree, suggesting that differential repair does not provide the full explanation. Increased UV damage formation seemed likely to contribute as well, which will be discussed at length in paper **I**.

Table 1. Sequence context of melanoma promoter mutations recurrent in at least 5 tumours. Mutated base in bold, with TTCCG sequence highlighted. Adapted from Fredriksson et al. (70)

Recurrence	Gene	Sequence context
11	*RPL13A*	TCCGGACATT**C**TTCCGGTTGG
10	***TERT***	CCCGACCCCT**C**CCGGGTCCCC
7	*C16orf59*	AGCCACGCCC**C**TTCCGGGAGG
7	***TERT***	GCCCAGCCCC**C**TCCGGGCCCT
5	*ASXL2*	CGCCCCCGCC**C**TTCCGGTCTC
5	*PDCD11*	CAAATCCCGC**C**CTTCCGATTC
5	*FTH1*	GAGCCCGCTC**C**TTCCGGTGGG
5	*FTH1*	CGAGCCCGCT**C**CTTCCGGTGG
5	*FUBP3*	CCGGCTTTCC**C**TTCCGCCGGA
5	*ALYREF*	CGCGTGAGGC**C**TTCCGGTGCC
5	*RNF185*	AAATTAACCT**C**TTCCGGTTGG
5	*MRPS31*	CCCGCCCTCT**C**TTCCGCTTCC
5	*DPH3*	AGGACTAGCC**C**TTCCGGCGCA
5	*RPL18A*	GAGGGCGGGT**C**TTCCGGTAGT
5	*C16orf59*	GAGCCACGCC**C**CTTCCGGGAG
5	*DERL1*	CGAAACTTCC**C**CTTCCGGCGA
5	***TERT***	CTCCCGGGTC**C**CCGGCCCAGC

1.4.2 DRIVER DETECTION METHODS

Including genomic mutational heterogeneity in a mutational model allows for more accurate prediction of mutation rates, which is a requirement for cancer driver detection, as we have discussed previously. The kinds of heterogeneity that are incorporated varies between methods, and depends in part on the underlying approach to detecting positive selection. The approaches in use today rely on a few different concepts, which we will summarise here (Figure 9).

EXCESS MUTATIONS

The most straightforward way of searching for driver mutations is by testing whether the number of mutations in a region exceeds expectations. A major difference between methods is how the background mutation rate is modelled (49, 72, 73). As recurrence-based approaches are sensitive to model flaws on both local and larger scales, incorrect mutation rate expectations could result in both false positives and false negatives.

dN/dS RATIO

Both oncogenes and TSGs tend to have a higher proportion of nonsynonymous mutations than non-cancer genes. As discussed in section 1.2.1, oncogenes are often activated by missense mutations, while TSGs are made defunct through nonsense mutations. By analysing the ratio between nonsynonymous and synonymous mutations (dN/dS) in a gene, positively selected genes may be found (48). This approach assumes that most synonymous mutations are passengers. While synonymous mutations have been shown to contribute to cancer (74, 75), the dN/dS ratio still manages to detect positive selection of cancer genes (48). dN/dS methods require mutational models that accurately handle mutational probabilities within a gene, but the effects of megabase-scale genomic mutational heterogeneity are reduced due to the fact that the count of synonymous mutations provides a local estimate of the mutation rate (76). Nevertheless, dNdScv, the most prominent dN/dS-based method, still takes large-scale genomic mutational heterogeneity into account to improve sensitivity (48).

POSITIONAL CLUSTERING

Positively selected mutations often occur in functionally relevant positions in a protein. For instance, a mutation in an active site is more likely to alter the function of a protein than one in another region, resulting in clustering of positively selected mutations (77). This can be exploited for driver detection, by searching for positional clustering of mutations in genes (78-83). The kinds of clustering detected by these methods can be divided into three groups (84). Linear clustering (78) is simply the distance in the primary structure of a protein, i.e., how many amino acid residues separate mutations. In domain clustering, specific regions such as SH2 and kinase domains are analysed for enrichment of mutations (79, 81). Finally, 3D clustering takes the tertiary structure of a protein into account, and tests for clustering of mutations in 3D, even if the mutations are far apart in the primary structure (80, 82).

FUNCTIONAL IMPACT

Functional impact-based methods work by evaluating the effect that a mutation has and relies on the assumption that positively selected mutations tend to skew more towards disruptive effects than passengers do (85-87). The exact impact of a mutation is generally not known, but it can be estimated through various methods (88-90). A simple example in the case of a nonsynonymous mutation in a coding region would be comparing how similar the substituted amino acid is in terms of polarity to the previous one, where a big change would constitute a larger functional impact.

OncodriveFML is an example of a notable functional impact method that works on both coding and non-coding sequences, as long as a suitable functional impact score is provided (91). In this method, the observed average

functional impact of mutations in a region is compared to the expected value, provided by mutation simulations. By simulating the same number of mutations as observed in the studied region and cohort of tumours, the test eliminates recurrence signals and provides an orthogonal approach to driver detection. While still vulnerable to flaws in the mutational model pertaining to individual positions in the region of interest, this approach elegantly sidesteps the issue of genomic mutational heterogeneity on larger scales.

COMBINING METHODS

The different approaches outlined above may be combined to attack the problem of driver detection from multiple angles at once. As an example, Dietlein et al. recently published a paper describing how MutPanning combines recurrence and functional impact by testing the likelihood of observing a given number of mutations in a region, and the likelihood that those mutations occur in their respective trinucleotide contexts (92). The trinucleotide context portion is an indirect proxy for functional impact, as functionally important positions are not likely to have trinucleotide contexts that match the most mutated contexts in the signatures of the mutational processes active in the tumour cohort.

Frameworks have also been established where multiple tools are run on the same dataset (84, 93). By combining orthogonal approaches to driver detection, high-confidence compendiums of cancer genes may be generated.

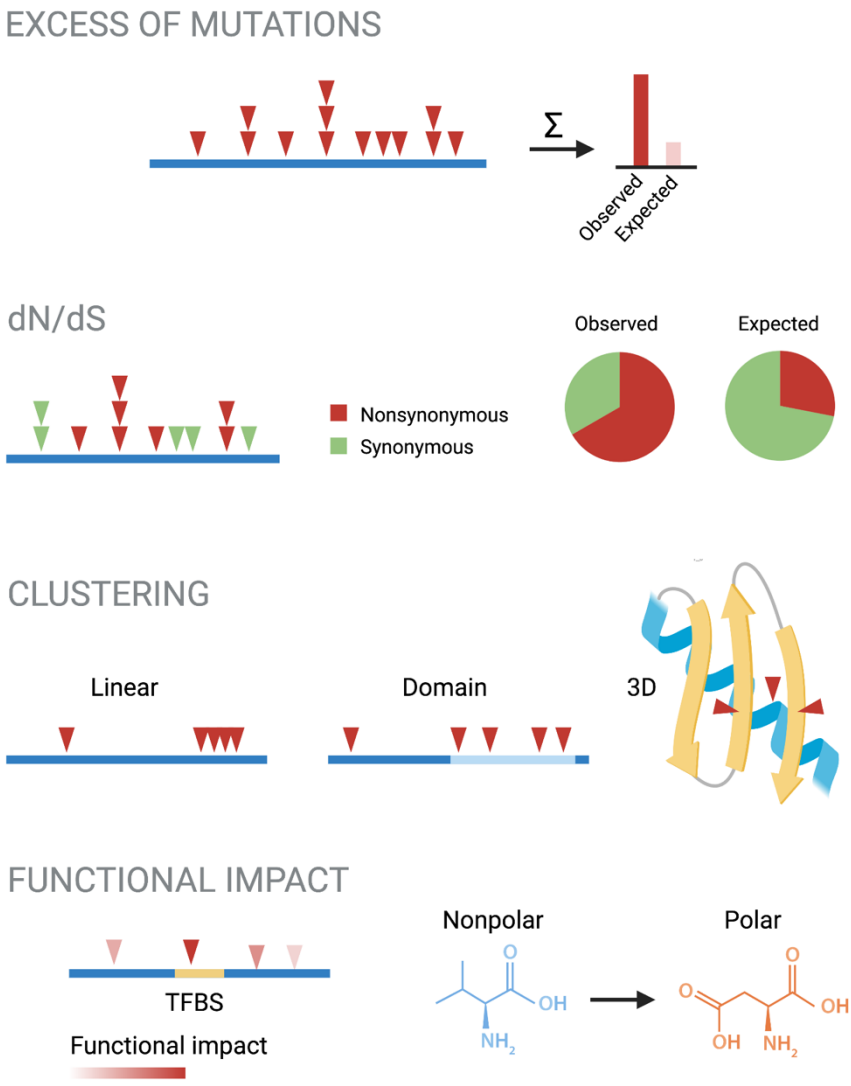

Figure 9. Some different approaches in cancer driver detection. Created with BioRender.com.

1.4.3 DNA SEQUENCING TO MAP MUTATIONS

All driver detection methods discussed here rely on the availability of somatic mutation data from tumours. Such data is generated by using high-throughput technologies to sequence tumour DNA and aligning the resulting short reads to the human genome. In order to filter out germline variants, non-tumour DNA is also sequenced, and variants that are present in the tumour but not in the normal sample are presented as the results of this somatic mutation calling. The process is complicated by the fact that tumour DNA may be of different purity, meaning the tumour sample will have some amount of non-tumour DNA as well, which has an effect on the expected number of sequencing reads that will contain a variant base. Another problem is the varying mappability of the genome, where repetitive regions are more difficult to detect mutations in. The end result of these complications is that somatic mutation calling is imperfect, with both missed somatic mutations and misclassified germline mutations being issues that have to be taken into consideration in downstream analyses. The latter can result in false positives in driver detection, and for that reason variants commonly occurring in the human population (SNPs, short for single-nucleotide polymorphisms) are often filtered out prior to driver analysis.

The area of the genome covered when sequencing depends on the technique used. Most of the publicly available mutation data to date comes from WXS, where only the exome is sequenced. This is done by isolating exonic DNA, for instance through array-based capture, where single-stranded DNA from the target regions is used to bind exonic DNA before sequencing. As the coding part constitutes only 1 % of the full genome, this process is cheaper than sequencing the whole genome, and allows for deeper sequencing. As much of the search for cancer drivers is focused on protein-coding genes, WXS mutation calls are of great use. The availability of WXS somatic mutations is also excellent, in no small part due to The Cancer Genome Atlas (TCGA). This project has characterised a massive amount of genomic data, starting with glioblastoma in 2008 (94) and now covering 33 cancer types.

WGS data presents an obvious advantage in that it allows for the analysis of non-coding regions as well, but the larger number of mutations covered can also help with modelling the background mutation rate. This is important for driver detection, as we have discussed at length. In particular, tumour-specific mutational models are much more feasible with the greater coverage, and therefore higher number of mutations, of WGS data. High-quality WGS somatic mutation calls have been generated by the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium (47), using WGS data from both TCGA and the International Cancer Genome Consortium (ICGC) (95). As the cost of sequencing goes down, WGS is becoming more and more common compared to WXS.

1.4.4 DAMAGE AND REPAIR MAPS

When attempting to characterise genomic mutational heterogeneity, somatic mutation data only provides part of the picture. As discussed in section 1.4 and 1.4.1, mutations in tumours are the end result of DNA damage, repair, and selection. Detangling the relative roles of each of these processes is aided by mapping of not only mutations, but of damage formation and repair. There are many different methods available to this end.

DAMAGE MAPS

Of the many different damage mapping techniques available, we will briefly touch on Excision-seq, Damage-seq/HS-Damage-seq, and CPD-seq. In Excision-seq, damaged bases are excised using a base excision repair enzyme (96). Given a large enough amount of damage, this will split the DNA molecule into small double-stranded fragments that can be sequenced and mapped. The damaged sites will then correspond to the base just before the mapped read. If the amount of DNA damage is insufficient for this approach, an alternative version exists, where excision repair is instead used to destroy damaged DNA, leading to only undamaged DNA remaining to be mapped to the genome.

In Damage-seq (97), and the improved version called HS-Damage-seq (High Sensitivity) (69), the fact that bulky adducts block most DNA polymerases is utilised. In HS-Damage-seq, fragments containing bulky lesions are partially copied, with the copies ending at the damaged sites. Copies of non-damaged fragments are discarded, allowing for amplification of relevant fragments only. These are then mapped to the genome, with the damaged site corresponding to the position right before the read.

The most important damage mapping technique for this thesis is CPD-seq (56), used to map the positions of CPDs. In this method, the ends of DNA fragments are blocked with specific primers, followed by cleavage of the CPD site with T4 endonuclease V and subsequent end repair with another endonuclease. This leaves an unblocked end to which another primer can be ligated, and because of the initial blocking step, all such unblocked ends will be at CPD positions. Through amplification, sequencing, and mapping, the positions of CPDs can then be determined. As in the other techniques, the position just before the mapped read corresponds to the damaged site.

REPAIR MAPS

NER can be mapped using excision repair sequencing (XR-seq) (98, 99). In this technique, excised DNA fragments containing bulky lesions are captured using immunoprecipitation, typically with antibodies that only bind to one type of damage, such as CPDs. Adapters are added to the fragments, after which the damage is either reversed or bypassed with TLS. For instance, CPDs can be separated into regular pyrimidines using photolyase. After amplification of the

resulting undamaged DNA, the fragments can be sequenced and mapped to the genome in order to find where NER is active and to what extent. This technique has been used extensively to study differential repair of UV damage (65-67), as discussed in section 1.4.1. While highly useful, XR-seq does not show the exact position of the damage being repaired.

An indirect way of mapping repair is to make damage maps at different time points. By calculating the difference in remaining damage between two time points, the level of repair in that time can be inferred (68). One of the main benefits of this approach is that it captures damage as well as repair. If only repair mapping is performed, it is not possible to tell whether high repair activity is caused by large amounts of DNA damage or because of high accessibility for DNA repair, for instance. Another benefit is that nucleotide-resolution repair maps are possible provided that the chosen damage mapping technique has that property.

2 AIM

The aim of this thesis is to investigate genomic mutational heterogeneity, particularly in cutaneous melanoma, with an aim to improving the mutational models required for driver detection in cancer. Commonly used trinucleotide models fail to capture effects related to longer sequence contexts, and do not account for mutational signature variability in different genomic regions – oversights that can result in both false positives and false negatives in the search for cancer genes. In three papers, we address some of these problems by:

I. Investigating the cause of recurrent promoter mutations in UV-induced melanoma.

II. Studying the sequence properties and epigenetic factors that contribute to variable levels of CPD formation, and its effect on the mutational signature of UV light.

III. Developing a recurrence-independent method demonstrating a novel concept for driver detection, which is less sensitive to some forms of mutational heterogeneity.

3 RESULTS AND DISCUSSION

3.1 INCREASED UV DAMAGE FORMATION UNDERLIES PROMOTER HOTSPOT MUTATIONS IN MELANOMA (PAPER **I**)

As discussed in section 1.4.1, the majority of all recurrent mutations in cutaneous melanoma occur near TFBSs, specifically in or immediately upstream of the sequence context TTCCG, a sequence matching the binding motif of the ETS family of TFs. The high level of recurrence could easily be misinterpreted as positive selection, when in fact there are elements of genomic mutational heterogeneity at play. Several studies point to differential DNA repair due to TF binding being the cause of these mutations (65, 66), while Fredriksson et al. suggested that the main contributor might be increased UV damage formation (70), which has previously been shown to be modulated by protein binding (56, 68, 69, 100, 101). Following up on this, we set out to investigate the contributions of UV damage formation (specifically CPDs) and DNA repair in promoter mutation hotspots in melanoma.

We started by characterising the hotspots in terms of mutations. For this, we used a cohort of 221 melanoma whole genomes, combining data from TCGA and ICGC (102, 103). With the exception of some notable driver genes, most highly recurrent mutations occurred near TFBSs, and the majority of those within 10 bp of the TTCCG context. The most recurrently mutated site was in the *RPL13A* promoter just upstream of a TTCCG motif, with mutations in 58 tumours, outnumbering even the cancer-driving *TERT* promoter mutations (26, 27). If the *TERT* mutations are excluded, the fraction of melanoma promoter mutations that are TTCCG-related increases with the recurrence of the mutations, illustrating the dominance of this phenomenon in promoters. Somewhat similarly, the number of mutated hotspots in each tumour was strongly correlated with their mutation burden, replicating the results of Fredriksson et al (70). This indicates that the hotspots are passenger mutations resulting from a mutational process, as opposed to driver mutations, since the selection of a mutation occurs irrespective of the overall burden in a tumour. We will revisit this concept in paper **III**, where it is applied to driver detection.

3.1.1 UV DAMAGE FORMATION

In order to study the relationship between UV damage formation and the hotspot mutations, we generated a map of CPD damage across the genome, adapting the CPD-seq technique previously used by Mao et al. in yeast (56) to work with Illumina sequencing. This method yields the position of one CPD per read pair, requiring extensive sequencing to attain high coverage. By

comparison, sequencing where each base in the read is informative (for instance for mutation calling) covers two orders of magnitude more positions per read pair.

To generate the map, we irradiated A375 melanoma cells with UV light, immediately followed by CPD-seq to ensure that there would be no time for DNA repair before sequencing, thus isolating the damage formation phenomenon. The UV exposure was performed on both cellular and naked DNA, as the UV damage formation in the former is affected by protein binding, for which naked DNA acted as a control. To this we added a sample with no UV exposure as a control for the method itself. We mapped 200 million CPDs to the genome in the cellular DNA sample, creating the most extensive map of CPD damage to date.

As even this large dataset did not provide quantitative CPD data for individual regions, all of the TTCCG-related hotspots mutated in at least 5 tumours were aggregated. By centring a window on the TTCCG motif for each hotspot, we could compare mutations and CPD formation in the hotspots (Figure 10). The expected mutational peaks were accompanied by similarly strong CPD peaks, but only in cellular DNA. Neither the UV-exposed naked DNA nor the non-exposed control sample showed any increase in CPD formation at the hotspot sites. The fact that UV damage formation was increased only in cellular DNA indicates that the binding of TFs to the TFBSs increases the sensitivity to UV light at these sites, ultimately resulting in C>T mutations. The dominant role of this phenomenon in explaining the promoter hotspots is further cemented by the strong correlation between the level of recurrence of mutations with the amount of CPD formation. It is further supported by another paper by Mao et al. that came out the same year (104).

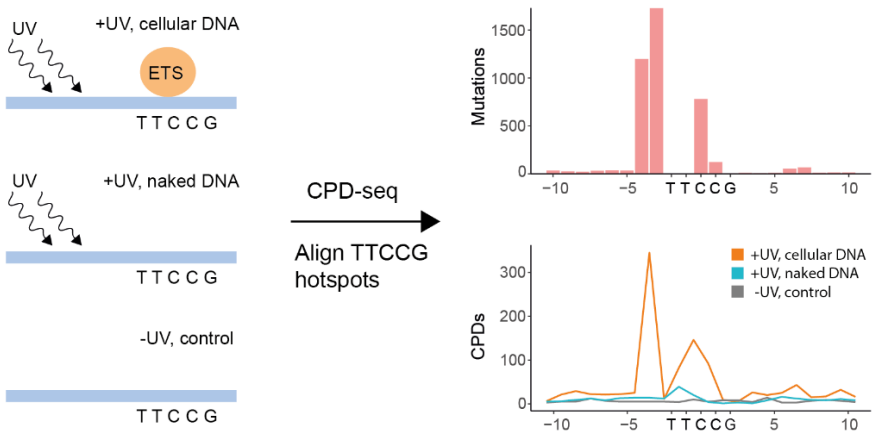

Figure 10. Comparison of mutations in 221 melanoma genomes with CPD formation following UV exposure with and without bound proteins, as well as a control sample without UV exposure, in aggregated hotspot positions.

3.1.2 THE ROLE OF REPAIR

To directly test the role of DNA repair versus damage formation, we UV-irradiated both A375 cells with functioning NER and fibroblasts with homozygous mutations in several NER-related proteins and looked for resulting mutations in the *RPL13A* hotspot. The UV dose had to be limited to only 20 J/m^2 (as compared to 1000 J/m^2 for the CPD map), as the repair-deficient cells were sensitive to UV light, but even at this low exposure mutations at the hotspot site stood out after sequencing with the ultrasensitive SiMSen-Seq technique (105). Notably, mutations at the hotspot site were observed in all of the repair-deficient cell lines, arguing against differential DNA repair as the main explanation of the hotspots.

As restricted access for NER to occupied TFBSs has been demonstrated to cause increased mutation rates (65-67), we decided to assess whether CPD formation also plays a part in this phenomenon. Therefore, we compared mutation rates around TFBSs in our melanoma cohort with CPD formation. While the same increased mutation rate observed by others was apparent, no great changes from expectations were evident in CPD formation. Instead, the TFBS-centred mutation peak matched well with impaired NER, mapped with XR-seq. Additionally, we observed a complete lack of increased mutation rate around TFBSs in cutaneous squamous cell carcinomas (cSCCs) deficient in global NER, indicating that NER is indeed responsible for these mutations. Interestingly, when we filtered the TFBSs to remove TTCCG-related promoters, there was a marked decrease in mutations in melanoma, but not in the NER-deficient cSCCs, suggesting that impaired NER contributes to mutations in TTCCG-related promoters as well.

In conclusion, while impaired NER does indeed appear to be the source of most mutations around TFBSs, including in TTCCG-related promoters, the extraordinary recurrence of mutations at specific positions in and near the TTCCG motif are caused by increased UV damage susceptibility upon TF binding. This phenomenon illustrates the need for mutational models that accommodate longer sequence patterns than the normally used trinucleotides, as well as how seemingly promising putative drivers may just be passengers, misidentified because of an insufficient understanding of genomic mutational heterogeneity.

3.2 VARIATION OF THE UV MUTATIONAL SIGNATURE IN THE GENOME (PAPER **II**)

Trinucleotide signatures are widely used in modelling mutation rates for driver detection, but as demonstrated in paper **I**, they fail to capture effects related to extended sequences. The TTCCG motif that is central to the UV-induced melanoma hotspots is problematic not only because it is longer than trinucleotide models can accommodate, but also because it only affects mutation rates in active promoters. In this paper, we set out to study how the UV trinucleotide signature varies in different regions of the genome, as well as to propose an extended trinucleotide-based mutational model that incorporates longer sequences.

3.2.1 TRINUCLEOTIDE SIGNATURE VARIATION

The mutational profile of cutaneous melanomas is dominated by the UV-associated signature 7, making it ideal for investigating variation across the genome, as no deconvolution of signatures is required. For this reason, we selected 130 melanoma whole genomes with a high fraction of UV-related mutations (>80 % C>T or CC>TT in dipyrimidines contexts, and at least 10,000 mutations) from the cohort used in paper **I**. Unsurprisingly, the mutational profile of this dataset closely resembled the UV signature. To study signature variation across the genome, we utilised a ChromHMM model based on RoadMap epigenomic data (106, 107) dividing the genome into 15 chromatin states, such as TSSs, transcribed regions, and heterochromatin. While the mutational signature in these regions was mostly highly similar to the UV signature, principal component analysis (PCA) separated regions related to active transcription start sites (TSS) from the others. These regions had lower cosine similarity to the UV signature, with the biggest difference being a strongly decreased number of mutations in the TCG trinucleotide context. While TCG is an uncommon trinucleotide in the genome due to methylation-mediated C>T mutations, as discussed in section 1.3.2, it has the highest weight in the UV signature. Previous observations confirm a lower number of TCG mutations in promoters which varied with methylation level, leading to a lower mutation rate in these regions (66). To confirm the connection between lower methylation levels and reduced TCG mutation rates, we used bisulfite sequencing data (107) to calculate the average methylation level of the chromatin states (low in TSS-associated regions and high in the rest of the genome), as well as in annotated promoters. In both cases, there was a clear correlation between the level of methylation and the amount of TCG mutations, explaining why active TSSs have fewer TCG mutations than the rest of the genome, and the deviation from the UV signature.

One plausible explanation for the relationship between TCG mutations and methylation levels is the documented propensity for CPD formation at methylated cytosines upon exposure to UVB light (41, 42). To investigate this possible cause, we generated a new CPD map using UVB wavelengths, as the CPD dataset in paper **I** used UVC light. The UVC CPD dataset was also included for comparison. In order to compare the CPD data to the mutational signature, we defined a similar trinucleotide signature for CPDs, including the CPD dinucleotide and one additional base on the 3' side. This allowed us to study methylated cytosines in CPDs, by including both the CPD and overlapping CpG's (YCG). Repeating the methylation level analysis performed for mutational signatures, we observed that CPD formation at YCG trinucleotides was similarly correlated with high methylation, though only for UVB and not UVC light (Figure 11a). In further support of methylation-related CPD formation as the cause of the variation in the mutational signature, a comparison of promoter and non-promoter regions showed significantly lower YCG CPD formation in promoters, again only for UVB light (Figure 11b). While CCG mutations did not have decreased weight in the mutational signature in promoters, this could simply be because of low frequency and the use of relative signature weights.

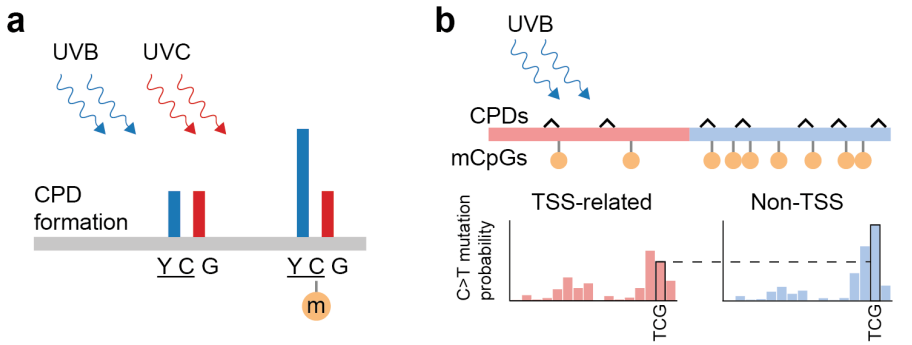

Figure 11. The effect of methylation levels on the UV signature. (a) Methylated cytosines are more prone to CPD formation than unmethylated cytosines when exposed to UVB, but not UVC light. (b) The low level of methylation at CpG's in TSS-related regions compared to the rest of the genome results in fewer CPDs forming in YCG contexts. This causes a significant decrease in the weight of C>T mutations in the TCG context in the UV mutational signature.

3.2.2 EXTENDED UV SIGNATURE

In an effort to model the effect of sequence contexts longer than trinucleotides on mutation rates in different regions, we extended the traditional trinucleotide signature model to include the presence or absence of selected pentamers within 10 bp on either side. For each chromatin state, pentamers frequently co-occurring with C>T mutations were selected, and the top-ranking pentamers from each region were included in a regression model together with the regular

trinucleotides. Using this model, we found both stimulating and attenuating effects from proximal pentamers. Most relevant to this thesis, the TTCCG motif was found to increase the probability of mutations, but only in TSS-related regions. When applying this model to the TTCCG-related hotspots in paper **I**, the expected mutation rates were considerably higher than predicted by traditional trinucleotide signatures, thereby better modelling the mutational heterogeneity observed in these regions.

3.3 RECURRENCE-INDEPENDENT DRIVER DETECTION (PAPER **III**)

In paper **I**, we argued that the correlation between the number of TTCCG hotspot mutations in each individual tumour and that tumour's mutation burden indicated that the hotspots were passenger mutations. Essentially, if the probability of observing a mutation in a given region is directly proportional to the exposure of the mutational process that would cause it, then there is no evidence of positive selection. This is because positive selection only comes into play after a mutation has been formed, taking no heed of the process that brought it about. As such, a mutation that is beneficial to a tumour should generally be under the same degree of positive selection regardless of the overall mutation burden of the tumour, thereby disrupting the correlation between mutation burden and the occurrence of the mutation in tumours.

In this paper, we set out to use this concept in the opposite way of how it was used in paper **I**. Instead of confirming passenger status upon observing the correlation, we developed a test that identifies cancer drivers by searching for disrupted correlation. In a cohort of patients, this would manifest as a higher-than-expected portion of the mutations in a region occurring in low-burden tumours. There are several potential benefits to a driver test based on this principle. First of all, it is an orthogonal approach to driver detection compared to currently used methods, and as such could help identify cancer genes that are missed by other approaches. Secondly, an approach that bypasses recurrence effects like this could prove to be less sensitive to the false positives that plague driver tests due to inaccuracies in the mutational model, such as the TTCCG hotspots.

3.3.1 IMPLEMENTATION OF THE METHOD

Our method uses a simple trinucleotide-based mutational model that only takes SNVs into account, where the probability of observing a given mutation depends on the frequency of that mutation type in the genome. The model is entirely patient-specific when using WGS data, but for WXS data it is calculated based on the whole cohort and then scaled to each tumour using their mutation burden, due to the lower coverage of this data type. Using this model, the probability of observing one or more mutations in a region, typically a gene, can be calculated for each tumour. We then simulate thousands of cohorts with the same number of mutated tumours as observed in the region of interest in the real cohort, with the probability of each tumour being mutated determined by our model. By comparing the likelihood of the combination of mutated tumours in the real cohort to the simulated ones, we can determine whether it is likely to be the result of positive selection (Figure 12).

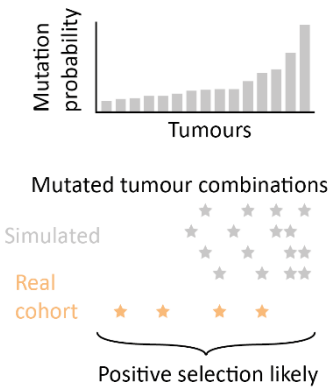

Figure 12. Positive selection test concept. The probability of a region (e.g., a gene) being mutated is calculated for each tumour in a cohort. This is used to determine the likelihood of different combinations of mutated tumours. The combination in the real cohort is compared to simulated cohorts with the same number of mutated tumours to test for positive selection.

3.3.2 DETECTING DRIVERS IN MELANOMA

To evaluate the method, we used somatic SNVs from 466 melanoma whole-exome tumour genomes from TCGA. We tested all genes with mutations in at least 3 tumours, resulting in 6 significant genes at a false discovery rate (FDR) of 5 %. At the top of this list were *BRAF* and *NRAS*, both well-known drivers in melanoma, and mutated in a large portion of the cohort. The other significant genes were *PTEN*, *GNAQ*, *MAP2K1*, and *KIT*, all known cancer genes, with perhaps the most interesting result being *GNAQ*, with only 7 mutated tumours. This gene is a cancer driver in uveal melanoma and has been detected in melanoma subtypes that are not UV-related (102, 108, 109). Upon closer inspection, we found that the low-burden tumours with mutations in *GNAQ* had a much lower percentage of UV-typical mutations (C>T in dipyrimidines contexts) than most of the tumours. The same was true for *KIT* and the not quite significant *SF3B1*, both of which have also been reported in non-UV-related melanomas. This result suggests that our method might be particularly suitable for finding cancer drivers in genes that are atypical of the cancer type being analysed.

To test the method's sensitivity to flaws in the mutational model, we next evaluated mutations in promoter regions in the cohort of 221 melanoma whole genomes used in paper **I**. In methods that do not include the TTCCG hotspot effect in their models, for instance through an extended trinucleotide signature as discussed in paper **II**, TTCCG-related sites would likely fill the list of significant results with false positives, interspersed with real cancer drivers such as *TERT* promoter mutations. However, since TTCCG hotspot mutations

correlate with UV-induced mutation burden, our method did not attribute positive selection to the TTCCG-related promoters despite not modelling the TTCCG phenomenon, instead only identifying the *TERT* promoter mutations as drivers. To illustrate the difference, we compared the *TERT* promoter with the most recurrently mutated TTCCG-related promoter, that of *RPL13A*, and noted that *TERT* promoter mutations skewed toward low-burden tumours in a way that *RPL13A* promoter mutations did not (Figure 13). As a representative for recurrence-based methods, ActiveDriverWGS (73) found both the *TERT* promoter and the TTCCG hotspots to be significant, with the *RPL13A* promoter edging out *TERT* as the most significant. This demonstrates how our method's independence of recurrence reduces the impact of model flaws, avoiding problems with some types of genomic mutational heterogeneity. In theory, this should be applicable to any model flaws, provided that they affect the tumours in a cohort equally.

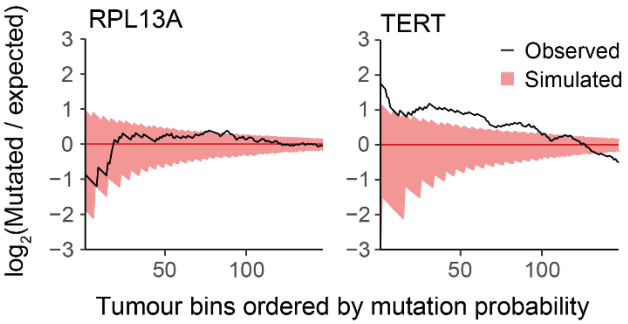

Figure 13. Log_2 ratio of the number of observed vs expected mutated tumours in 75 tumour bins in a sliding window, for the RPL13A and TERT promoters in melanoma. The coloured area indicates the least extreme 90 % of the simulated cohorts. The skew of mutations toward low-burden tumours in the TERT promoter is evidence of positive selection and is absent in the RPL13A promoter.

3.3.3 DETECTING CANCER DRIVERS IN DIFFERENT CANCER TYPES

After our initial tests in melanoma, we evaluated the method's performance in additional cancer types using whole-exome tumour genomes from TCGA. Out of all cancer types with enough mutations for inclusion (≥ 1000 genes with at least 3 mutated tumours), endometrioid uterine corpus carcinoma (UCEC) performed best, with some 50 significant genes, most of which are canonical cancer genes (110) or identified in other cancer gene catalogues (48, 49, 92, 93). This was attributed to the high mutational burden of UCEC, a trait shared by most cancer types that our method performed well with. Of all the genes identified in this test, several were promising putative driver genes not previously touted as such, including *BMF* (a pro-apoptotic protein) and *DOK1* (a negative regulator of the insulin signalling pathway) in UCEC, as well as *DCLK1* (a marker for tuft cells in intestines) in stomach cancer.

In conclusion, our method introduces a novel concept for driver detection that is orthogonal to currently used approaches, allowing it to detect some drivers that are missed by other tools. The method's uncoupling from recurrence signals shields it from false positives caused by some forms of poorly modelled genomic mutational heterogeneity, such as the TTCCG hotpots. As no model is ever perfect, this is a valuable property that should make our method a useful complement to other tools.

4 CONCLUSIONS AND FUTURE PERSPECTIVES

Searching for cancer drivers in genomic data generally entails finding regions that are more mutated than expected. The strongest drivers are easily discernible without advanced mutational models, but detailed modelling of genomic mutational heterogeneity is necessary to find more subtle drivers, as well as to avoid false positives. Mutational heterogeneity exists on all scales, from megabase-level effects all the way down to single nucleotides, and it varies between different mutational processes. In this thesis, we have endeavoured to characterise heterogeneity in cutaneous melanoma related to the variable formation rate of UV-induced DNA damage, specifically TTCCG-related mutation hotspots in promoters and the variation of the UV mutational signature due to methylation levels. While these advances will undoubtedly be useful, they cover only a fraction of the heterogeneity that mutational models have to contend with. In an attempt to circumvent the problem of flawed models, we developed a driver detection test that is not affected by recurrence signals, instead only testing whether the combination of mutated tumours in a cohort is likely while ignoring the number of mutations. This approach proved resilient to poorly modelled mutational heterogeneity in the form of the TTCCG promoter hotspots and should be similarly useful for other heterogeneity that affects the entire cohort equally.

While the new driver detection method showed great promise, there are several avenues for improvement that are worth exploring. Firstly, the mutational model was quite simplistic, and could be improved. The test is robust concerning model flaws that affect all tumours in a cohort equally, but it is sensitive to differences between tumours. Better modelling of the mutation rates in individual tumours could potentially be achieved with tumour-specific expression scaling, for instance, and would be expected to improve driver detection results. Secondly, many tools use multiple driver detection concepts synergistically, and we believe our method could fit well into such approaches. For instance, studying how the dN/dS ratio or the functional impact of mutations varies across a cohort of tumours could yield more information than just treating the cohort as one data point.

In conclusion, handling genomic mutational heterogeneity is of great importance for the discovery of cancer drivers. As the amount of available genomic data increases, ever better models as well as improved driver detection methods will be required to find drivers without being inundated with false positives. In this regard, we believe we have contributed by adding some small pieces to the puzzle.

ACKNOWLEDGEMENTS

Först och främst vill jag förstås tacka dig, **Erik**, för att jag har fått vara en del av labbet de senaste fem åren. När jag först ansökte om att doktorera hos dig tittade jag på bilder på gruppen på labbhemsidan, och noterade hur avslappnad och trevlig stämningen verkade vara. Det intrycket visade sig vara helt rätt, och jag tror inte det är någon slump att just din grupp blev så. Jag har lärt mig mycket om bioinformatik, skrivande, presentationer, figurdesign och allt möjligt av dig, men det allra bästa har varit din förmåga att entusiasmera. Det har alltid räckt att ta ett möte med dig när ett projekt känns tungrott för att det ska kännas kul igen, och det har varit ovärderligt. Tack! Tack också till min bihandledare **Anders**, inte minst för samarbetet på polη-artikeln.

Secondly, a big thank you to all members of the Larsson Lab, past and present. **Kerryn**, you are the glue that holds the lab together, both socially and scientifically. You're both a good friend and something of a bonus supervisor. Thank you for all the good times, and for always being willing to give me input on anything, be it figures, scientific writing (including this thesis!), or dog names. **Markus**, du har stenkoll på vilket håll alla dörrar öppnas åt, och blandar bara ibland ihop M&M. Att ha fått dela hela doktorandresan med en vän som dig är jag mycket glad för. Den av oss som får nytt jobb först får påtala behovet av en extra bioinformatiker så att vi kan fortsätta på nästa arbetsplats med. **Ari**, you have helped me all throughout my PhD, from the ISP on my very first day all the way to the tangled mess of forms and deadlines for the dissertation. This would have been a lot more stressful without you, and I'm very grateful for your help and support. **Isabella**, **Vinod**, and **Tom**, on the rare occasions when I actually show up at the lab in person these days, it's always great to see you. There will be more of that now that the thesis writing is done, I think, and I'm looking forward to it! **Arman**, **Alireza**, and **Emma**, I'm very happy that you were part of the lab, and I hope to see you again soon. **Babak**, **Swaraj**, **Susanne**, **Joakim**, and **Jimmy**, you were there when my PhD journey started, and it wouldn't have been the same without you. Finally, **Katrin**, you are an honorary Larsson Lab member at this point, and a very welcome addition to all pre-pandemic lunches along with the rest of the Clausen lab. I could go on much more about everyone mentioned here, but in the interest of keeping this somewhat brief: thank you all for making the lab such a welcoming and fun place to work. Better lab mates I could not ask for.

Anna och **Erik (**Kristiansson), tack för exjobbet som lät mig prova på bioinformatik. Det passade mig så mycket bättre än allt annat på utbildningen och ni var fantastiska handledare. Det är tack vare er som jag kunde fortsätta med en doktorandtjänst!

Mamma och **Pappa**, tack för allt stöd under doktorandperioden och de 27 åren innan det. Ni har alltid haft tid att prata när jag har behövt det, oavsett om det gäller vilken kaffekokare jag ska köpa eller de lite större livsbesluten. Tack också till resten av familjen (**Jonas**, **Karin**, **Anne**, **Åke**, **August**, **Astrid** och **Mormor**) för att ni finns och är den bästa familjen man kan önska sig.

Louise och **Anders**, tack för att ni har tagit hand om Rufus ännu mer än vanligt under avhandlingsskrivandet. Det har underlättat enormt! **Rufus**, tack för att du är en fin hund. Sitt.

Sist, men inte minst, **Antonia**. Tack för ditt tålamod de senaste månaderna när jag har jobbat kvällar och helger med avhandling och artiklar. Tack för all uppmuntran när det har varit svårt, inte bara inför disputationen utan under hela doktorandperioden. Att få komma hem till dig varje dag har gjort de senaste fem åren så mycket bättre.

REFERENCES

1. Baym M, Lieberman TD, Kelsic ED, Chait R, Gross R, Yelin I, et al. Spatiotemporal microbial evolution on antibiotic landscapes. Science. 2016;353(6304):1147-51.
2. Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, et al. An estimation of the number of cells in the human body. Annals of Human Biology. 2013;40(6):463-71.
3. Matthews HK, Bertoli C, De Bruin RAM. Cell cycle control in cancer. Nature Reviews Molecular Cell Biology. 2021.
4. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.
5. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011;144(5):646-74.
6. Witsch E, Sela M, Yarden Y. Roles for Growth Factors in Cancer Progression. Physiology. 2010;25(2):85-101.
7. Sporn MB, Todaro GJ. Autocrine secretion and malignant transformation of cells. N Engl J Med. 1980;303(15):878-80.
8. Fernandez-Medarde A, Santos E. Ras in Cancer and Developmental Diseases. Genes & Cancer. 2011;2(3):344-58.
9. Nishida N, Yano H, Nishida T, Kamura T, Kojiro M. Angiogenesis in cancer. Vascular Health and Risk Management. 2006;2(3):213-9.
10. Harris CC. p53 tumor suppressor gene: from the basic research laboratory to the clinic--an abridged historical perspective. Carcinogenesis. 1996;17(6):1187-98.
11. Downward J. Mechanisms and consequences of activation of protein kinase B/Akt. Current Opinion in Cell Biology. 1998;10(2):262-7.
12. Okamoto K, Seimiya H. Revisiting Telomere Shortening in Cancer. Cells. 2019;8(2):107.
13. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics. 2006;7(2):85-97.
14. Shlien A, Malkin D. Copy number variations and cancer. Genome Medicine. 2009;1(6):62.
15. Chan CK, Aimagambetova G, Ukybassova T, Kongrtay K, Azizan A. Human Papillomavirus Infection and Cervical Cancer: Epidemiology, Screening, and Vaccination—Review of Current Perspectives. Journal of Oncology. 2019;2019:1-11.
16. Yang X, Lippman ME. BRCA1 and BRCA2 in breast cancer. Breast Cancer Res Treat. 1999;54(1):1-10.
17. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8.

18. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb Perspect Biol. 2010;2(1):a001008.
19. Hamzehloie T, Mojarrad M, Hasanzadeh Nazarabadi M, Shekouhi S. The role of tumor protein 53 mutations in common human cancers and targeting the murine double minute 2-p53 interaction for cancer therapy. Iran J Med Sci. 2012;37(1):3-8.
20. Aubrey BJ, Strasser A, Kelly GL. Tumor-Suppressor Functions of the TP53 Pathway. Cold Spring Harbor Perspectives in Medicine. 2016;6(5):a026062.
21. Itahana K, Dimri G, Campisi J. Regulation of cellular senescence by p53. Eur J Biochem. 2001;268(10):2784-91.
22. Murugan AK, Grieco M, Tsuchida N. RAS mutations in human cancers: Roles in precision medicine. Seminars in Cancer Biology. 2019;59:23-35.
23. Malumbres M, Barbacid M. RAS oncogenes: the first 30 years. Nature Reviews Cancer. 2003;3(6):459-65.
24. Soussi T, Wiman KG. TP53: an oncogene in disguise. Cell Death & Differentiation. 2015;22(8):1239-49.
25. Pitolli C, Wang Y, Mancini M, Shi Y, Melino G, Amelio I. Do Mutations Turn p53 into an Oncogene? Int J Mol Sci. 2019;20(24):6241.
26. Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, et al. TERT Promoter Mutations in Familial and Sporadic Melanoma. Science. 2013;339(6122):959-61.
27. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly recurrent TERT promoter mutations in human melanoma. Science (New York, NY). 2013;339(6122):957-9.
28. Elliott K, Larsson E. Non-coding driver mutations in human cancer. Nature Reviews Cancer. 2021;21(8):500-9.
29. Corona RI, Seo J-H, Lin X, Hazelett DJ, Reddy J, Fonseca MAS, et al. Non-coding somatic mutations converge on the PAX8 pathway in ovarian cancer. Nature communications. 2020;11(1):2020-.
30. Shuai S, Suzuki H, Diaz-Navarro A, Nadeu F, Kumar SA, Gutierrez-Fernandez A, et al. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature. 2019;574(7780):712-6.
31. Suzuki H, Kumar SA, Shuai S, Diaz-Navarro A, Gutierrez-Fernandez A, De Antonellis P, et al. Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma. Nature. 2019;574(7780):707-11.
32. Wang W, Wei Z, Lam TW, Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci Rep. 2011;1:55.
33. Payne JL, Wagner A. Mechanisms of mutational robustness in transcriptional regulation. Front Genet. 2015;6:322.

34. Liu B, Xue Q, Tang Y, Cao J, Guengerich FP, Zhang H. Mechanisms of mutagenesis: DNA replication in the presence of DNA damage. Mutation Research/Reviews in Mutation Research. 2016;768:53-67.
35. Wu S, Zhu W, Thompson P, Hannun YA. Evaluating intrinsic and non-intrinsic cancer risk factors. Nat Commun. 2018;9(1):3490.
36. Sharma R, Lewis S, Wlodarski MW. DNA Repair Syndromes and Cancer: Insights Into Genetics and Phenotype Patterns. Frontiers in Pediatrics. 2020;8(683).
37. Ames BN, Shigenaga MK, Hagen TM. Oxidants, antioxidants, and the degenerative diseases of aging. Proc Natl Acad Sci U S A. 1993;90(17):7915-22.
38. Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980;287(5782):560-1.
39. Abeti R, Zeitlberger A, Peelo C, Fassihi H, Sarkany RPE, Lehmann AR, et al. Xeroderma pigmentosum: overview of pharmacology and novel therapeutic strategies for neurological symptoms. Br J Pharmacol. 2019;176(22):4293-301.
40. Ikehata H, Ono T. The Mechanisms of UV Mutagenesis. Journal of Radiation Research. 2011;52(2):115-25.
41. Tommasi S, Denissenko MF, Pfeifer GP. Sunlight induces pyrimidine dimers preferentially at 5-methylcytosine bases. Cancer research. 1997;57(21):4727-30.
42. Drouin R, Therrien JP. UVB-induced cyclobutane pyrimidine dimer frequency correlates with skin cancer mutational hotspots in p53. Photochem Photobiol. 1997;66(5):719-26.
43. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979-93.
44. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SaJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415-21.
45. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321(6067):209-13.
46. 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):7435-51.
47. ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82-93.
48. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171(5):1029-41.e21.
49. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8.

50. Poulos RC, Wong JWH. Finding cancer driver mutations in the era of big data research. Biophys Rev. 2019;11(1):21-9.
51. Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome. Cell. 2019;177(1):101-14.
52. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012;488(7412):504-7.
53. Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, et al. Transcription Restores DNA Repair to Heterochromatin, Determining Regional Mutation Rates in Cancer Genomes. Cell Reports. 2014;9(4):1228-34.
54. Yazdi PG, Pedersen BA, Taylor JF, Khattab OS, Chen Y-H, Chen Y, et al. Increasing Nucleosome Occupancy Is Correlated with an Increasing Mutation Rate so Long as DNA Repair Machinery Is Intact. PLOS ONE. 2015;10(8):e0136574.
55. Chen X, Chen Z, Chen H, Su Z, Yang J, Lin F, et al. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science. 2012;335(6073):1235-8.
56. Mao P, Smerdon MJ, Roberts SA, Wyrick JJ. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(32):9057-62.
57. Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. Reduced mutation rate in exons due to differential mismatch repair. Nature Genetics. 2017;49:1684.
58. Vrieling H, Venema J, van Rooyen ML, van Hoffen A, Menichini P, Zdzienicka MZ, et al. Strand specificity for UV-induced DNA repair and mutations in the Chinese hamster HPRT gene. Nucleic Acids Res. 1991;19(9):2411-5.
59. Mugal CF, von Grünberg H-H, Peifer M. Transcription-Induced Mutational Strand Bias and Its Effect on Substitution Rates in Human Genes. Molecular Biology and Evolution. 2008;26(1):131-42.
60. Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell. 2016;164(3):538-49.
61. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature. 2015;521(7550):81-4.
62. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. Human mutation rate associated with DNA replication timing. Nature Genetics. 2009;41(4):393-5.
63. Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biology. 2018;19(1):129.

64. Kreisel K, Engqvist MKM, Kalm J, Thompson LJ, Boström M, Navarrete C, et al. DNA polymerase η contributes to genome-wide lagging strand synthesis. Nucleic Acids Research. 2018;47(5):2425-35.
65. 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):264-7.
66. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JWH. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature. 2016;532(7598):259-63.
67. Poulos RC, Thoms JAI, Guan YF, Unnikrishnan A, Pimanda JE, Wong JWH. Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. Cell Reports. 2016;17(11):2865-72.
68. Frigola J, Sabarinathan R, Gonzalez-Perez A, Lopez-Bigas N. Variable interplay of UV-induced DNA damage and repair at transcription factor binding sites. Nucleic Acids Research. 2020;49(2):891-901.
69. Hu J, Adebali O, Adar S, Sancar A. Dynamic maps of UV damage formation and repair for the human genome. Proceedings of the National Academy of Sciences. 2017:201706522-.
70. Fredriksson NJ, Elliott K, Filges S, Van den Eynden J, Ståhlberg A, Larsson E. Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. PLoS Genetics. 2017;13(5).
71. Wei G-H, Badis G, Berger MF, Kivioja T, Palin K, Enge M, et al. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. The EMBO Journal. 2010;29(13):2147-60.
72. Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res. 2015;43(17):8123-34.
73. Zhu H, Uusküla-Reimand L, Isaev K, Wadi L, Alizada A, Shuai S, et al. Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks. Molecular Cell. 2020;77(6):1307-21.e10.
74. Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, et al. A pan-cancer analysis of synonymous mutations. Nature Communications. 2019;10(1).
75. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell. 2014;156(6):1324-35.
76. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348(6237):880-6.
77. Wagner A. Rapid detection of positive selection in genes and genomes through variation clusters. Genetics. 2007;176(4):2451-63.

78. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics (Oxford, England). 2013;29(18):2238-44.
79. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013;9:637.
80. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci U S A. 2015;112(40):E5486-95.
81. Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. Bioinformatics (Oxford, England). 2014;30(21):3109-14.
82. Tokheim C, Bhattacharya R, Niknafs N, Gygax DM, Kim R, Ryan M, et al. Exome-Scale Discovery of Hotspot Mutation Regions in Human Cancer Using 3D Protein Structure. Cancer research. 2016;76(13):3719-31.
83. Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet. 2016;48(8):827-37.
84. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nature Reviews Cancer. 2020;20(10):555-72.
85. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer research. 2009;69(16):6660-7.
86. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39(17):e118.
87. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012;40(21):e169.
88. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009;4(7):1073-81.
89. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812-4.
90. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
91. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome biology. 2016;17(1):128-.

92. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, et al. Identification of cancer driver genes based on nucleotide context. Nature Genetics. 2020.
93. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173(2):371-85.e18.
94. The Cancer Genome Atlas Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8.
95. The International Cancer Genome Consortium. International network of cancer genome projects. Nature. 2010;464:993.
96. Bryan DS, Ransom M, Adane B, York K, Hesselberth JR. High resolution mapping of modified DNA nucleobases using excision repair enzymes. Genome Res. 2014;24(9):1534-42.
97. Hu J, Lieb JD, Sancar A, Adar S. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2016;113(41):11507-12.
98. Hu J, Adar S, Selby CP, Lieb JD, Sancar A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes & Development. 2015;29(9):948-60.
99. Hu J, Li W, Adebali O, Yang Y, Oztas O, Selby CP, et al. Genome-wide mapping of nucleotide excision repair with XR-seq. Nat Protoc. 2019;14(1):248-82.
100. Pfeifer GP, Drouin R, Riggs AD, Holmquist GP. Binding of transcription factors creates hot spots for UV photoproducts in vivo. Molecular and cellular biology. 1992;12(4):1798-804.
101. Tornaletti S, Pfeifer GP. UV light as a footprinting agent: modulation of UV-induced DNA damage by transcription factors bound at the promoters of three human genes. Journal of molecular biology. 1995;249(4):714-28.
102. Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. Nature. 2017;545:175.
103. The Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. Cell. 2015;161(7):1681-96.
104. Mao P, Brown AJ, Esaki S, Lockwood S, Poon GMK, Smerdon MJ, et al. ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. Nature Communications. 2018;9:2626.
105. Ståhlberg A, Krzyzanowski PM, Jackson JB, Egyud M, Stein L, Godfrey TE. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. Nucleic Acids Research. 2016;44(11):e105-e.

106. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215-6.
107. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
108. Kim C-Y, Kim DW, Kim K, Curry J, Torres-Cabala C, Patel S. GNAQmutation in a patient with metastatic mucosal melanoma. BMC Cancer. 2014;14(1):516.
109. Livingstone E, Zaremba A, Horn S, Ugurel S, Casalini B, Schlaak M, et al. GNAQ and GNA 11 mutant nonuveal melanoma: a subtype distinct from both cutaneous and uveal melanoma. British Journal of Dermatology. 2020;183(5):928-39.
110. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research. 2019;47(D1):D941-D7.