Computational exploration of cancer genomes

Joakim Karlsson

UNIVERSITY OF GOTHENBURG

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

Gothenburg 2019

Computational exploration of cancer genomes

joakim.karlsson@gu.se

ISBN 978-91-7833-290-8 (PRINT)
ISBN 978-91-7833-291-5 (PDF)
http://hdl.handle.net/2077/58233

Printed in Gothenburg, Sweden 2019
Printed by BrandFactory

Dedicated to Evert Jonsson

"To secure ourselves against defeat lies in our own hands, but the opportunity of defeating the enemy is provided by the enemy himself."

- Sun Tzu

Abstract

Cancer evolves due to changes in DNA that give a cell an advantage at the expense of the remaining organism. These alterations range from individual base substitutions to broad losses or duplications of chromosomal material. This thesis explores how DNA and RNA sequencing can guide discovery of altered genes responsible for cancer development, profile the immune landscapes of tumors and support the diagnosis of difficult cases.

In the first of three studies, we examined DNA and RNA from the tumors of a patient with metastatic cancer but an uncertain diagnosis. We discovered that these tumors harbored a mutational signature associated with ultraviolet radiation. This restricted the possible sites of origin to those that can be exposed to sunlight. To confirm this, gene expression estimates were then compared to a large database of multiple cancer types. This gave a perfect match to cutaneous melanoma, thus enabling a certain diagnosis.

The second study established a method for searching candidate cancer genes that are altered by genomic copy number changes. The method integrates estimates of copy number changes with gene expression to prioritize genes concurrently and consistently altered with respect to both, putting greater emphasis on copy number changes comprising smaller chromosomal regions, which tend to exclude unselected genes from consideration. This system was able to retrieve known cancer genes as top candidates in several cancer types. In addition, this method also implemented a way to examine regions of DNA where genes are currently not known to exist.

In the final study, we molecularly profiled metastatic uveal melanoma (UM), a rare but difficult to treat eye cancer. We reintroduced a functional version of the tumor suppressor *BAP1* into one deficient tumor, resulting in a global transcriptional shift towards a less metastatic subtype. We also found one tumor harboring a specific mutational signature that has not previously been observed in UM, and which might suggest a new risk factor. Next, we narrowed down a set of candidate genes potentially influencing tumor behavior via broad copy number changes, which could possibly be drug targets. Finally, we transcriptomically profiled tumor-infiltrating T-cells and found these to be in exhausted states, possibly explaining the failures of immunotherapy in UM. Despite this, they were in several cases capable of tumor recognition.

In conclusion, this thesis explores molecular data of cancers from a number of different angles. The results should have relevance for diagnostic principles and may suggest candidate genes for future functional studies.

Keywords: Cancer genomics, transcriptomics, driver genes, copy number changes, immunogenomics, cancer of unknown primary, uveal melanoma.

Sammanfattning på svenska

Cancer utvecklas på grund av förändringar i DNA som ger en cell en fördel på bekostnad av den övriga organismen. Dessa förändringar sträcker sig från substitutioner av enskilda nukleotider till större förluster eller dupliceringar av kromosomalt material. Denna avhandling undersöker hur DNA- och RNA-sekvensering kan styra upptäckten av förändrade gener som ansvarar för cancerutveckling, profilera tumörers interaktioner med immunsystemet och informera diagnos i svåra fall.

I den första av tre studier undersökte vi DNA och RNA från tumörer hos en patient med metastatisk cancer men en osäker diagnos. Vi fann att dessa tumörer hade ett specifikt mönster av mutationer som associerats med ultraviolett strålning. Detta begränsade de möjliga ställen i kroppen i vilka den primära tumören kunde ha uppstått, till de som kan utsättas för solljus. För att bekräfta detta jämfördes genuttrycksnivåer sedan med en stor databas över flera cancertyper. Detta gav en perfekt matchning mot hudmelanom, vilket möjliggjorde en slutgiltig diagnos.

Den andra studien etablerade en metod för att söka efter möjliga cancergener som påverkas av förändringar i antalet underliggande DNA-kopior. Metoden integrerar estimerade kopietalsförändringar med genuttryck för att prioritera gener som samtidigt och konsekvent förändras med avseende på båda. Den lägger samtidigt större vikt vid kopietalsförändringar som innefattar mindre kromosomala regioner, eftersom dessa tenderar att utesluta gener som inte selekteras av tumörer. Detta system kunde åternominera kända gener som toppkandidater i flera cancertyper. Utöver detta implementerade metoden också ett sätt att undersöka regioner av DNA där gener för närvarande inte är kända att existera.

I den slutliga studien genomförde vi en molekylär profilering av metastatiskt uvealt melanom (UM), en sällsynt men svårbehandlad ögoncancer. Vi återinförde en funktionell version av tumörsuppressorn *BAP1* i en tumör där denna gen inaktiverats av en mutation, vilket resulterade i ett globalt skifte i genuttryck i riktning mot en subtyp av denna cancer som har mindre metastatisk kapacitet. Vi fann också en tumör som hade en specifik mutationssignatur som inte tidigare har observerats i denna cancer, och som skulle kunna indikera en ny riskfaktor. Därefter kartlade vi gener som potentiellt skulle kunna påverka tumörers beteende via förändringar i antalet kopior av kromosomarmar, vilka eventuellt skulle kunna vara mål för riktade läkemedel. Slutligen profilerade vi genuttryck i individuella tumörinfiltrerande T-celler och fann dessa vara i utmattade tillstånd, en möjlig

förklaring till varför immunterapi inte fungerar i denna cancerform. Trots detta vara T-cellerna i flera fall kapabla att känna igen tumörer.

Sammanfattningsvis innefattar denna avhandling en undersökning av molekylär data från tumörer utifrån ett antal olika vinklar. Resultaten bör ha relevans för diagnostiska principer och kan möjligtvis föreslå gener av intresse för framtida funktionella studier.

List of papers

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

I Mutational signature and transcriptomic classification analyses as the decisive diagnostic tools for a cancer of unknown primary

*Olofsson Bagge R, *Demir A, *Karlsson J, Alaei-Mahabadi B, Einarsdottir BO, Jespersen H, Lindberg MF, Muth A, Nilsson LM, Persson M, Svensson JB, Söderberg EMV, de Krijger RR. Nilsson O, Larsson E, Stenman B and Nilsson JA. *JCO Precision Oncology*. 2018.

II FocalScan: Scanning for altered genes in cancer based on coordinated DNA and RNA change

Karlsson J and Larsson E. *Nucleic Acids Research*. 2016; 44 (19): e150.

III Molecular profiling of driver events and infiltrating T-cells in metastatic uveal melanoma

Karlsson J, Nilsson LM, Forsberg EMV, Mitra S, Alsén S, Stierner U, All-Eriksson C, Green L, Einarsdottir B.O, Jespersen H, Belgrano V, Nilsson Wassen O, Ny L, Lindnér P, Larsson E, Olofsson Bagge R and Nilsson J.A. *Manuscript*

Papers not included in this thesis

I A patient-derived xenograft pre-clinical trial reveals treatment responses and a resistance mechanism to karonudib in metastatic melanoma

Einarsdottir BO, *Karlsson J, *Söderberg EMV, Lindberg MF, Funck-Brentano E, Jespersen H, Brynjolfsson SF, Olofsson Bagge R, Carstam L, Scobie M, Koolmeister T, Wallner O, Stierner U, Warpman Berglund U, Ny L, Nilsson LM, Larsson E, Helleday T and Nilsson JA. *Cell Death & Disease*. 2018; 9: 810.

II H19 induces abdominal aortic aneurysm development and progression

Li DY, Busch A, Jin H, Chernogubova E, Pelisek J, Karlsson J, Sennblad B, Liu S, Lao S, Hofmann P, Bäcklund A, Eken SM, Roy J, Eriksson P, Dacken B, Ramanujam D, Dueck A, Engelhardt S, Boon RA, Eckstein HH, Spin JM, Tsao PS, and Maegdefessel L. *Circulation*. 2018; 138 (15): 1551–1568.

III Transcriptomic characterization of the human cell cycle in individual unsynchronized cells

Karlsson J, Kroneis T, Jonasson E, *Larsson E and *Ståhlberg A. *Journal of Molecular Biology*. 2017; 429 (24): 3909-3924.

IV Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers

Alaei-Mahabadi B, Bhadury J, Karlsson J, Nilsson JA, and Larsson E. *Proceedings of the National Academy of Sciences of the United States of America*. 2016; 113 (48): 13768-13773.

Content

Abbreviations

A	Adenine
APC	Antigen-presenting cell
C	Cytosine
CAR-T	Chimeric antigen receptor T-cells
CM	Cutaneous melanoma
CNV	Copy number variant
CUP	Cancer of unknown primary
DC	Dendritic cell
DNA	Deoxyribonucleic acid
DNA-seq	DNA-sequencing
FFPE	Formalin-fixed paraffin embedded
G	Guanine
HLA	Human leukocyte antigen
Indel	Insertion or deletion
LOH	Loss of heterozygosity
MHC	Major histocompatibility complex
mRNA	Messenger RNA
NK	Natural killer (cell)
PDX	Patient-derived xenograft
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
SNV	Single nucleotide variant
T	Thymine
TCGA	The Cancer Genome Atlas
TCR	T-cell receptor
U	Uracil
UM	Uveal melanoma
UV	Ultraviolet (radiation)
WES	Whole exome sequencing
WGS	Whole genome sequencing

TCGA

ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Diffuse large B-cell lymphoma
GBM	Glioblastoma multiforme
KICH	Chromophobe renal cell carcinoma
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LGG	Brain lower-grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumor
THCA	Thyroid carcinoma
THYM	Thymoma
UCS	Uterine carcinosarcoma
UCEC	Uterine corpus endometrial carcinoma
UVM	Uveal melanoma

1 Introduction

Cancer arises due to the corruption of normal cells by changes in the cellular DNA. These changes can occur due to inherited DNA maintenance deficiencies, age-associated accumulation of mutations or exposure to exogenous mutagens, such as certain chemicals or radiation. Cancer cells grow in a disordered fashion to form masses called tumors, due to the inactivation of control mechanisms that normally prevent this. The host immune system has the capacity to recognize most rogue cells and eliminate them. However, cancer cells eventually tend to develop evasion mechanisms and become resistant. As the cells keep dividing, they gradually acquire additional genomic changes that may provide them with the ability to migrate and settle in new locations of the body: they metastasize.

Metastatic disease is very difficult to treat. On one hand, cancers that spread may hide in obscure places. On the other, they often evolve independent drug and immune resistance mechanisms. For these reasons, treatments may eliminate a fraction of the tumors, while unresponsive clones often remain in the body. In some cases, cancers may metastasize before the original tumor is discovered, leaving few effective treatment options. In a fraction of these instances, the original tumor and the type of cancer the patient is affected by cannot be determined.

Cancers arising from different tissues and in different patients are unique diseases. The underlying causes, mutations and cellular behaviors vary, and drugs found to be appropriate for one cancer may have no use on another. The success of cancer treatment is also dependent on the status of the patient's immune system. Eventually, mechanisms that are in place to prevent harmful long-term immune reactions activate, and consequently reduce the ability to destroy the cancer cells. Therefore, treatment needs to be tailored to the unique conditions of each patient for an optimal response.

The goal of cancer genomics is to profile the genetic changes that occur in tumors, to discover the mutations that cause the cells to behave erratically, determine the underlying risk factors and to search for tumor vulnerabilities that may provide an opportunity to eliminate them.

1.1 The human genome

The human genome is composed of approximately 3 billion pairs of nucleobases, which constitute the double stranded nucleic acids known as DNA (deoxyribonucleic acid). The nucleobases are each members of the four-letter alphabet adenine (A), cytosine (C), guanine (G) and thymine (T) (**Fig. 1a**)[1]. A and G derive from a class of compounds termed purines, while C and T correspond to pyrimidines. DNA is double-stranded and composed of pairs between A, C, G, T in one strand and T, G, C, A in the other complementary strand, respectively (**Fig. 1b-d**)[1]. Human DNA is partitioned into separate large units termed chromosomes, of which there are 23 pairs, in addition to a separate sequence of mitochondrial DNA.

The genome encodes instructions for making the proteins that govern the biochemical reactions, and thereby the appearance and behavior, (phenotype) of the cell, in units called genes[1]. These genes are transcribed by enzymes called RNA polymerases that create ribonucleic acid (RNA) sequences, also called transcripts, reflecting their original sequences. RNA is a similar molecule to DNA, but tends to be single-stranded and also substitutes the base T for uracil (U) (**Fig. 1d**)[1]. The set of RNAs produced from the genome is termed the transcriptome. A large fraction of these transcripts are then translated to protein sequences (**Fig. 1d**) by molecular complexes called ribosomes. These are termed messenger RNAs (mRNAs). Those that are not are called non-coding RNAs. Translation uses triplets of nucleotides, which can theoretically be decoded in three partially overlapping reading frames, to determine which amino acids are incorporated into proteins. A given gene can encode instructions for multiple different variants of a protein (isoforms), which is possible due to a mechanism called alternative splicing. This process functions by joining the different coding regions, exons, of a gene in different ways. The consequence is the exclusion or inclusion of different sets of exons, leading to the production of different mRNA sequences[1].

The genome is replicated and partitioned into two daughter cells at each cell division. Although, a certain number of small errors always occurs, leading to mutations accumulating during aging. At each cell division, the ends of the chromosomes, telomeres, also progressively shorten, since they are difficult to fully replicate each time[2]. When the telomeres are short enough, the cells enter a state called senescence, and cease to divide. This is beneficial, since cell division and improper repair mechanisms acting on chromosomes with compromised telomeres can lead to genomic instability, which could eventually could cause cancer[2]. Cancerous cells, however, tend to avoid senescence by expressing proteins that regenerate telomere sequences, which are otherwise only expressed in specific cell types, such as stem cells[2].

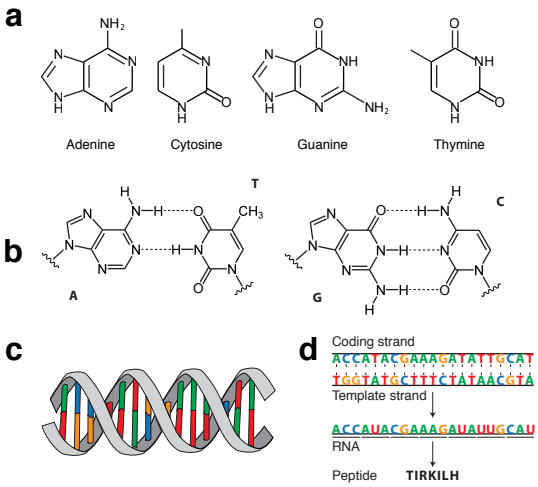

Figure 1. The elements in the basic structure of DNA. **a)** The chemical structure of the four constituent nucleobases. **b)** The two standard types of base pairs that form between these bases. **c)** The double-helical structure of DNA. No shown here: nucleobases gain sugar and phosphate groups to become nucleotides before joining in DNA chains. **d)** DNA is transcribed to RNA and subsequently translated to protein (peptides: protein fragments).

Besides genes, the human genome also contains a large number of regulatory elements. These include the core promoter regions adjacent to the transcription start sites of genes, where transcription factors and polymerase bind to initiate transcription[3]. In addition, regions known as enhancers exist, which also bind transcription factors, and can increase the transcription frequency of nearby genes[3]. Silencers, on the other hand, can instead recruit repressive factors[3]. The regions of the genome influenced by such elements are restricted by the presence of insulator sequences, which act as boundaries[3]. Various types of so called epigenetic DNA modifications, for instance methylation and chromatin conformation changes, also influence the transcription of genes. These modifications can be either activating or repressive[4]. Combined, these elements and alterations provide a wide variety of possibilities to regulate expression, making it possible for cells from different tissue types to display widely different phenotypes, despite sharing the same underlying DNA.

The basic sequence of the human genome was determined by the Human Genome Project and published in a first preliminary version in 2001[5]. This reference sequence continues to be improved as technological enhancements are made that allow for more precise reconstruction. However, each individual has a unique genome, containing a large number of inherited

changes, some of which are particular to certain populations[6]. The sequencing of many additional genomes in recent years, such as by the 1000 Genomes Project, have made it possible to construct databases of normal human DNA variants[7]. Understanding the structure of the normal human genome is of fundamental importance for identifying the genomic changes that give rise to cancer cells.

1.2 The cancer genome

Damage to DNA during cell division or due to external factors can cause any of the bases in the normal genome to be substituted for another one, called single nucleotide variants (SNVs, **Fig. 2a**), or additional bases to be inserted or deleted (indels, **Fig. 2b**). The latter can cause a shift in the natural reading frame of a gene, leading to irrelevant amino acids being incorporated into the translated protein. Some mutations can occur in germ cells and be inherited. Those that occur elsewhere in the body are termed somatic mutations. Many base substitutions do not alter protein sequences, since multiple codons exist that correspond to the same amino acids, i.e., they are degenerate. These mutations are called synonymous, whereas those that influence protein composition are termed non-synonymous.

New variants of the proteins encoded by mutated genes may have different properties. In some cases, they could provoke the cell to constantly signal for cell division. Altered genes that operate in this manner are called oncogenes, whereas their normal counterparts are called proto-oncogenes. In other cases, proteins that act as natural breaks on replication or contribute to other mechanisms preventing cancer development, such as through DNA repair, may lose their function[8]. These are termed tumor suppressors.

A cell can also suffer a larger error during cell division, which in turn, may cause wide sections of chromosomes to be lost, or gained in multiple copies. Such DNA copy number changes can be broad, sometimes affecting entire chromosome arms, or focal, limited to a relatively small number of genes (**Fig. 2c-d**). The result can be overexpression or loss of expression of the affected genes[8]. In the latter case, such events may also be coupled with mutations, which together inactivate a tumor suppressor and thereby triggers unrestrained growth, termed loss of heterozygosity (LOH). Genomic instability can also lead to the creation of fusion genes, as a consequence of genomic rearrangements (**Fig. 2e**). This can result in a new abnormal protein that changes cell behavior, and which can thereby act as an oncogene. However, such events can also lead to tumor suppressor dysfunction[9]. In addition, tumor suppressors may also be inactivated, or oncogenes overexpressed, as a result of viral infections. These viruses may also carry their own oncogenes into the cell[10].

Identifying oncogenes and tumor suppressors is of great importance, since the proteins they encode or the biochemical signaling pathways they affect can potentially be targeted with drugs. Targeting unique mechanistic dependencies in cancer cells can help to eliminate tumors while sparing normal cells. Identification of such genes largely demands that the genomes of many tumors are compared side by side to find consistent alterations in a genomic

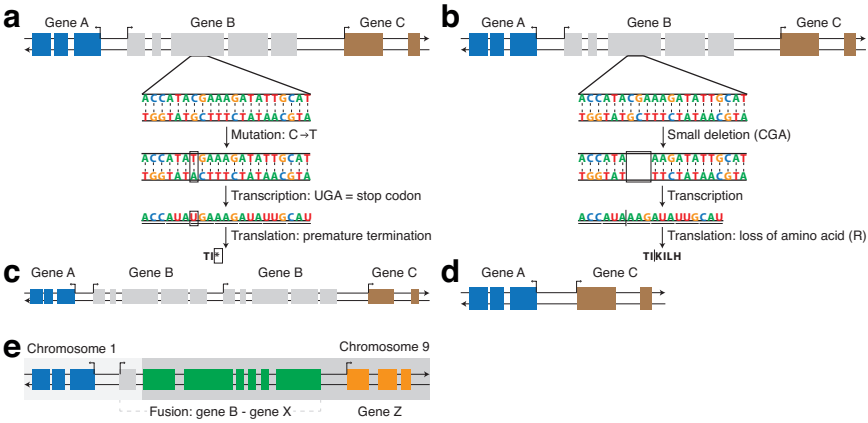

Figure 2. Alterations in the cancer genome that affect the coding sequences of genes. Transcription start sites indicated by arrows. **a)** Substitution of a single nucleotide (C to T) in a hypothetical gene "B" leading to a stop codon, which truncates the translated peptide. **b)** A small deletion in gene B leading to the loss a codon corresponding to the amino acid arginine (R), and therefore loss of this amino acid from the resulting peptide. **c)** Duplication of the entire gene B, which could influence its overall expression level. **d)** Deletion of the entire gene B. **e)** Translocation of a fragment from one chromosome to another, in this case resulting in a fusion between gene B and gene X from the other chromosome.

region. Overrepresented genomic events can indicate that these give the tumor a selective advantage, which in turn could suggest that the tumor might depend on the affected genes[11]. For instance, in cutaneous melanoma it has been found that the gene *BRAF* is mutated at the exact same nucleotide in over half of all patients, giving rise to a new hyperactive version of the protein it produces, which can promote cell division[12]. This has led to the development of a compound that inhibits the activity of the mutated protein, which has been successful in prolonging the survival of patients[13].

Cellular development is also influenced by the epigenetic state of DNA. This refers to non-nucleotide modifications of the genome, such as changes in its three-dimensional organization or the attachment of certain molecules to DNA, which can determine what regions that are open for transcription or alter the activity of regulatory elements, such as enhancers. Common epigenetic modifications are those that attach or remove methyl, acetyl or phosphoryl groups from structural units called histones. Besides regulating transcription, these changes can also have an impact on DNA repair. Several known oncogenes and tumor suppressors are involved in such modifications[14]. Since cellular specialization in the formation of different organs largely depends on changes in epigenetic state, cancers that arise from different organs and tissues also tend to possess different epigenetic landscapes.

In addition, differences can often also be found among subtypes of the same cancer[15,16]. Some tumors contain relatively few, if any, recurrent mutations, but instead seem dominated by epigenetic alterations. This is the case for a number of pediatric cancers[14]. Epigenetic state transitions are more common among cells in very young children, due to the active development of organs, which could potentially explain the overrepresentation of seemingly epigenetically driven cancers in this group.

An important American initiative, The Cancer Genome Atlas (TCGA), has performed sequencing of over 10000 genomes from 33 cancer types, and profiled their gene expression patterns and epigenetic states. This research has been made available to researchers worldwide and has tremendously progressed our understanding for the genomic characteristics of cancer cells. It has also made it possible to identify several new causal genes, for which drugs either have been developed or might be in the future. In addition, this data has led to more accurate ways of classifying tumors, on a genetic basis, which can improve prognostics and clinical decision-making[15,17].

1.3 Determining genome and transcriptome composition

The composition of DNA and RNA can be determined in a high-throughput manner with sequencing technologies. In a common approach, DNA is first shattered into short fragments, adapter and index sequences added to first facilitate amplification of the fragments, then immobilization within the instrument and later identification of the sample that the sequences originated from, in cases where multiple samples are analyzed simultaneously. Sequences complementary to the immobilized and amplified single-stranded fragments are then synthesized one nucleotide at a time. The nucleotides are tagged with chemical groups that lead to the generation of light signals as they are added, which allows the instrument to register their identity. Some methods use other approaches, such as the emission of hydrogen ions, in order to register added nucleotides. This allows reconstruction of the original sequences. RNA is analyzed similarly, but first needs to be reverse transcribed to complementary DNA[18]. The results are millions of short "reads". Alternatively, each fragment can also be sequenced from both ends, yielding pairs of reads.

These then need to be mapped to a reference sequence of the genome of the studied organism, using computational methods. Mismatches to the reference may indicate somatic or inherited DNA variants. The paired-end approach can be beneficial for reconstructing splice variants of RNA, determination of structural rearrangements in DNA, expressed fusion genes or viruses that have integrated themselves into the genome.

Different strategies can be applied for genome sequencing. The two methods employed in this work are whole exome sequencing (WES) and whole genome sequencing (WGS). WES is targeted towards the coding regions of genes, the exons. WGS targets the whole human genome, including the intronic regions of genes, which do not carry over to the protein sequence, as well as "intergenic" regions, which do not contain known genes. While WGS is theoretically preferable, using this technique is not always economically feasible. Still, the cost of sequencing a whole genome has decreased substantially from over $0.5-1 billion spent on the first human genome to less than $2000 presently[6]. For clinical purposes, it is also possible to perform targeted sequencing of specific genes that are both known to be altered in cancer and therapeutically actionable.

For transcriptome analysis, the corresponding method is termed RNA-seq and only covers actively transcribed exons. Different strategies can be used to prioritize different subsets of RNA molecules. For instance, one may deplete

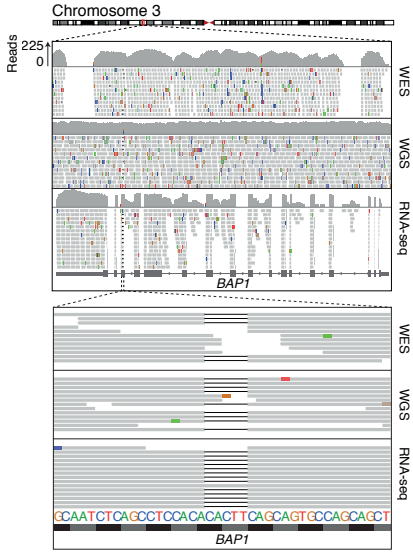

Figure 3. Reads (gray) aligned to the human genome, from exome (WES), whole genome (WGS) and RNA sequencing (RNA-seq), respectively, at the position of the tumor suppressor *BAP1*. The lower panel shows a zoomed-in view of a position with a small deletion in this gene, which leads to a frameshift upon translation. The consequence is very likely to be a non-functional protein.

ribosomal RNAs before sequencing, since these tend to be very abundant and rarely of interest. Another option is to only sequence transcripts with a poly-A tail, which will capture the majority of protein-coding RNAs, but miss many non-coding ones[19]. An earlier method for gene expression analysis was the microarray, which gives quantitative output, but not the sequences of transcripts. A visual comparison of WES, WGS and RNA-seq reads aligned to the human genome is shown in **Fig. 3**, highlighting a frameshift deletion in the tumor suppressor *BAP1*.

1.4 Genetic alterations that drive cancer development

The genes or genomic alterations that are responsible for the creation of tumors are collectively referred to as drivers[11]. These may allow the cancer cell to replicate beyond normal cellular limits, metastasize to new locations in the body, promote the growth of new blood vessels supplying a tumor with nutrients, reprogram its metabolism, or make it unresponsive to attacks by the immune system or drugs used for cancer treatment. In addition, the altered activity of such genes may compromise the genetic stability of the cell, facilitating new mutations and genomic rearrangements that can alter cellular behavior and allow it to adapt to new circumstances[20]. The characteristic biological rewiring that occurs in cancer cells have been divided into what is known as the hallmarks of cancer, summarized in **Table 1**[20]. To gain these hallmarks, only a few important genes may need to be altered[11,21]. The majority of mutations that occur in cancer are inconsequential, and referred to as passengers. Driver genes can, in addition to the hallmarks in which they participate, also be subdivided based on the types of genomic changes by which they are activated or inactivated. The following sections describe this in more detail, focusing on the most common classes of driver alterations, as well as approaches that may be used to identify genes of interest.

1.4.1 Base substitutions and indels

Oncogenes and tumor suppressors altered by base substitutions or indels are referred to as mutational drivers. In the case of oncogenes, these events commonly occur at specific "hotspots" in the gene, which may, for instance, disrupt a protein domain critical for interaction with a negative regulator, make them independent of otherwise necessary ligands for activation or prevent them from binding inhibitory drugs[8,21]. They may also give the gene entirely new functions[21]. Tumor suppressors, on the other hand, tend to be affected by a wider range of mutations that either cause dysfunction of a specific domain, a large truncation of the protein or indel-induced shifts in reading frame[22]. However, biologically relevant inactivation typically requires that both alleles are mutated, whereas one allelic event can be sufficient to activate an oncogene[11,23]. The exceptions are haploinsufficient tumor suppressors, where only one allele needs to be altered in order for the cancer to gain an advantage[24].

Mutations can also alter the splicing of transcripts, leading to retention of intronic sequences, which can prohibit complete translation. Such mutations may also cause exons to be either added or spliced out of the resulting transcript, which can alter protein function[25-27]. In addition, a

Table 1. Characteristic traits of cancer cells that have become known as the hallmarks of cancer, as a result of an influential review article of the same name[20].

Hallmark	Description
Induction of proliferative signaling	Growth factor availability limits cell division. Cancer cells may reduce their dependence on these, produce their own, provoke other cells to produce them or inactivate negative feedback mechanisms restraining their use.
Inactivation of anti-proliferative mechanisms	Checkpoint mechanisms prevent inappropriate replication if the right conditions are not met, for instance, if the genome is damaged. Cancer cells can deactivate such mechanisms.
Replicative immortality	Cells are limited in the number of replications they can undergo, closely related to chromosomal telomere length. Upregulation of enzymes regenerating telomeric sequences can counteract this.
Resistance to cell death	Mechanisms are in place to trigger cell death upon various cellular crisis states, which can become deactivated in cancer cells.
Genomic instability	Cancer development and evolution relies on the accumulation of mutations and other genomic alterations, caused by failures in genome maintenance and mutagenic exposures.
Angiogenesis	Tumors can produce factors that induce the growth of new blood vessels, which provide nutrients and oxygen needed for cellular metabolism and replication. It additionally allows them to get rid of metabolic waste products.
Metabolic rewiring	Cancer cells adapt their metabolism to constantly supply energy required for replication, commonly by high rates of glycolysis (even under aerobic conditions).
Inflammation	Inflammation-associated factors can promote growth, survival, angiogenesis and an epithelial to mesenchymal transition (associated with metastasis).
Immune evasion	The immune system normally recognizes and kills transformed cells. Expression of T-cell inhibitory molecules or defective antigen presentation can interfere with this[28].
Invasion and metastasis	Loss of intercellular adherence and specific changes in phenotype enables migration and adaptation to new environments, which makes it possible for the tumors to spread.

mutation may a create a new splice site within an exon, leading to a truncated version being included. Splicing defects can also be a mechanism that activates oncogenes, for instance by the skipping of exons that encode protein domains enforcing ligand-dependence[27]. An interesting case illustrating the potential consequences of alternative splicing is the oncogene *BRAF*^{V600E}, where the preferential expression of a particular transcript variant has been found associated with drug resistance[29].

Mutations can be detected in a tumor by mismatches that occur as sequencing reads are aligned to a reference genome. However, artifactual base changes can occur in reads due to issues with sequencing or alignment[30]. Accurate alignment is also challenging in certain sequence contexts of the human genome. For these reasons, methods need to be used that can statistically model which estimated mutant allele frequencies are relevant for considering a variant a true positive in different contexts[30].

In addition, a large number of discovered mismatches to the reference will be normal population variants that are not specific to the cancer. Therefore, it is beneficial to also sequence normal cell material from a patient, which makes it possible to filter out healthy genome variation and determine which mutations are somatic. Such control samples can also aid in detecting artifactual variant calls. Databases of common population variants are also available[7,31], which can allow further discrimination, or partially substitute when normal material is not available.

A challenge here is also the fact that tumors tend to contain varying amounts of normal cells from adjacent tissue, as well as immune cells. This can cause some mutations to occur in a smaller fraction of reads than expected for a pure sample, making them difficult to detect with accuracy. Furthermore, tumors often contain various subpopulations, known as sub-clonal cells, which possess different sets of mutations, contributing to the issue. Mutant allele frequencies are also influenced by DNA copy number changes. For this reason, methods developed for genotype calling of healthy individuals are generally not sufficient. Fortunately, however, a number of specialized tools have been developed that are able to statistically assess the influence of these confounding factors[32-34].

Next is the issue of discriminating driver mutations from the numerous passenger events that occur in cancer genomes. For this purpose, one may statistically test whether or not a given genomic element is mutated more frequently than can be expected by chance, which could indicate that the event is under positive selection. However, a factor to be accounted for is that different genomic contexts have different probabilities to undergo mutation. For instance, long genes and late replicating genomic regions will tend to accumulate a larger total number of mutations than others[35].

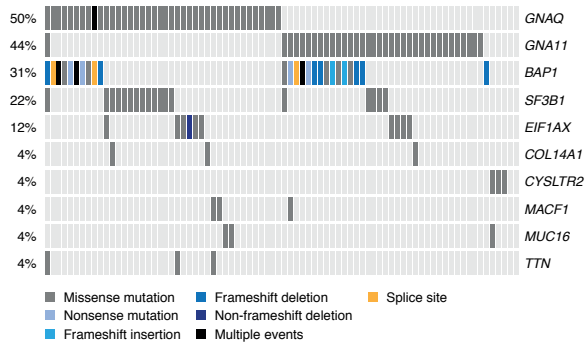

Figure 4. Recurrency and mutual exclusivity patterns of non-synonymous mutations in primary uveal melanoma (data from TCGA). The ten most frequently mutated genes in the cohort of 80 patients (columns) are shown. *GNAQ*, *GNA11*, *BAP1*, *SF3B1*, *EIF1AX* and *CYSLTR2* are considered mutated more frequently than can be expected by chance, taking into account covariates such as gene length. The mutual exclusivity patterns displayed by some of these also indicate that they may have effects that can substitute for each other in driving development of this cancer.

Therefore, indications of unexpectedly high recurrence may also be found if the model of what to expect is not accurate. Some methods used for nominating events under potential selection therefore also take into account a number of covariates that influence mutation rates[35,36]. However, these methods tend to require large sample sizes to be successful, which are not always available.

Significantly recurrent mutations in a given cancer type frequently display patterns of mutual exclusivity (**Fig. 4**). This can indicate that only one of them in a given tumor is sufficient to provide the cancer with the necessary advantage. Often such patterns have been found for genes that participate in the same biological pathways. For this reason, mutual exclusivity can provide additional evidence that a mutation or gene may be a cancer driver, in addition to potentially suggesting new mechanisms the gene could be involved in[37].

1.4.2 Copy number changes

Besides mutations, a frequent means of altering gene activity is via their duplication or loss in DNA. Detection of such copy number changes can be done with either WGS or WES. Another common approach uses single nucleotide polymorphism (SNP) microarrays. In the former, differences in genomic read coverage compared to normal control samples are used to discover changes[38]. In the latter, fragmented and fluorescently labeled DNA binds to oligonucleotide probes of common heterozygous genetic variants along the genome that are immobilized on a microarray. Signal intensities associated with binding to probes corresponding to different alleles can then

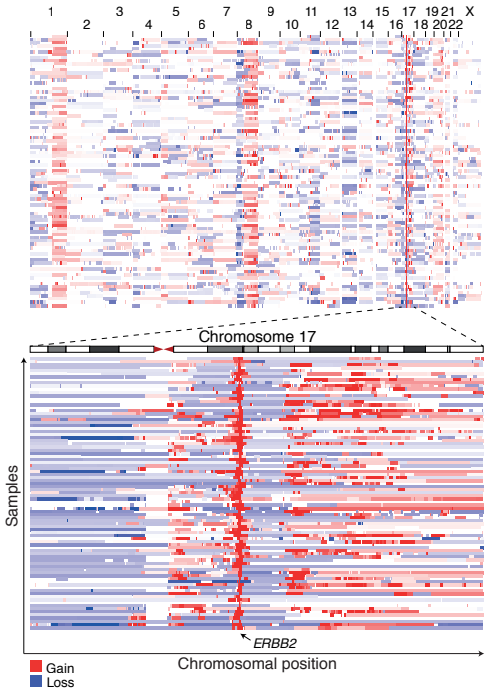

Figure 5. Copy number changes in breast cancer genomes sequenced by TCGA. Red indicates genomic regions that have been gained in multiple copies, with color intensity proportional to the number of additional sequences, whereas blue indicates regions where either one or two alleles have been lost. Each row represents one tumor and patient. The large number of aberrations indicates high levels of genomic instability. Some events occur at nearly the same position in a large number of cases. As an example (lower panel), recurrent focal copy number gains occur on chromosome 17 at the position of the oncogene *ERBB2*.

be used to assess the presence of broad shifts in estimated allele frequencies[39,40]. With both technologies, ratios relative to the expected normal scenarios are then segmented using a specialized algorithm in order to determine continuous copy number-altered regions.

These events can cause overexpression of oncogenes, underexpression, LOH or complete deletion of tumor suppressors. As with recurrent base substitutions, recurrency can also be indicative of genes under selection within regions of somatic copy number variants (CNVs)[41]. However, CNVs tend encompass a large number of genes, making it difficult to discriminate drivers from passengers. To better pinpoint those genes, one may take advantage of the fact that some tumors of a given cancer types may possess more size-limited (focal) events, which exclude most of the irrelevant candidates. In large cohorts, one may spot patterns that tend to narrow in on

a very specific region, containing only a few genes. Such is the case for *ERBB2* in breast cancer (**Fig. 5**), which has since been identified as an oncogene activated by the resulting overexpression in about 20-30% of tumors[42]. This, in turn, has led to the development of a successful drug that inhibits the encoded protein[42].

However, not all recurrent copy number changes are indicative of positive selection. Tumors with high levels of genomic instability may display such patterns by chance. To avoid false positives, as with mutations, one may therefore statistically assess whether the changes in a given region are more frequent than random expectation. One commonly used algorithm for this purpose first calculates an empirical model of somatic CNV background rates, to which different genomic regions are then compared to derive a significance estimate. This is followed by a second algorithm that delineates specific regions of peak recurrence[43].

Although this provides a statistical basis for determining copy number aberrations under selection and can lead to the nomination of convincing targets in some cases, many peak regions still encompass a large number of genes. Therefore, approaches that integrate additional data for each gene may be required to narrow the search space further. Commonly, one tests for correlations between changes in gene expression and copy number, since this is the most likely way that a copy number aberration will mediate a selective advantage.

A remaining difficulty is defining potential candidates in the cancer types that almost exclusively display chromosome arm-wide copy number changes, such as uveal melanoma[16]. In these cases, no informative regions of minimal overlap exist. There is also the possibility that these events do not target single genes, but rather groups that participate in specific pathways[44]. Yet another hypothesis is that they have no target at all, but rather confer an advantage via aneuploidy alone, although some literature suggests that the latter might have a negative impact on fitness[44-47].

1.4.3 Gene fusions

A third major class of genetic drivers arises from the merging of material of different genes. These so-called fusion genes can encode chimeric proteins that either have entirely new functions or the ability to be expressed at higher levels. They may also maintain constitutive activation due to acquired or lost protein domains that naturally either promote or restrain their actions[48,49]. Fusions are caused by chromosomal rearrangements, which may occur due to adverse events during DNA repair or cell division[50]. Such events may also cause the translocation of genes to regions that are more or less permissive to

transcription, as determined by nearby regulatory elements[48]. Less well studied is the phenomenon of transcription-induced gene fusions, where mistakes in transcription or splicing merges material from two, most often nearby, transcripts[48]. These do not appear particularly recurrent or cancer-specific, however, making it unlikely that they have any major role[48]. Many fusion genes encode kinases, which may be possible to target with existing inhibitors, making their identification a priority[49].

Recurrent fusions can be found in several cancer types. For instance, *BCR* and *ABL1* are often joined in myeloid leukemia, whereas fusions between *FGFR3* and *TACC3* are prevalent in glioblastoma multiforme, as well as other cancers, while *TMPRSS2* is fused to *ERG* in up to 38% percent of prostate adenocarcinomas sequenced by TCGA[9].

Fusions can be detected by analysis on either WGS or RNA-seq data, although using both approaches can improve specificity[9,48]. In these cases, it is beneficial to use paired-end sequencing, since the different ends of a given fragment may be found to map to exons of two different genes. As always, it is also useful to sequence a non-cancer sample from the same patient, in order to determine whether the discovered fusions are somatic. Another way these methods complement each other is that RNA-seq allows for assessing effects on expression of each partner gene, whereas WGS can reveal structural rearrangements that do not manifest as gene fusions.

1.4.4 Cancer viruses

Latent cancer genes can also be activated or deactivated by viral infections. Currently, seven, or potentially eight, viruses are known to be involved in human cancers[10,51]. To corrupt the cell into an efficient factory of viral particles, and in some cases integrate into its DNA, they must overcome the host's defense mechanisms. Some of these mechanisms are also important in preventing tumor formation. For instance, most of these cancer viruses express oncogenes that inhibit cell cycle and DNA damage checkpoints controlled by the tumor suppressors *RB1* and *P53*, the latter being the most commonly mutated gene in human cancer and a regulator of several pathways relevant to cancer hallmarks[10,52]. Accumulation of genomic damage as a result of dysfunctional checkpoint mechanisms can further enable mutation-induced activation of other cancer-associated genes.

There are also other means by which viruses can induce mutations, for instance via provoking aberrant activity of activation-induced cytidine deaminase (AID), a cellular protein with the explicit function of mutating DNA, which is normally used for diversifying the sequences of B-cell receptors. Inappropriate AID activity can be seen in several cancers, and is

associated with specific patterns of C > G and C > T mutations across the genome[53,54]. Some viral proteins can also promote migration and invasion[10], possibly enabling further distribution of viral particles. Tumor development can also be enabled by virus-associated chronic inflammation[10].

In the case of those that are able to integrate into the host genome, the viral sequences may contain enhancers or promoters that lead to the overexpression of oncogenes close to the integration site, achieve the same end goal by disrupting local regulatory elements, or transform proto-oncogenes via fusions with viral sequences[51,55-58]. This is known as insertional mutagenesis. Similarly, regulatory or sequence alterations in tumor suppressors may lead to their inactivation. Furthermore, some viruses may integrate cellular proto-oncogenes into their own genetic material and transform them to actively oncogenic variants[55,58].

The presence of viral sequences in cancer cells can be detected by examining reads that do not fully align to the human genome, which can then be searched for better matches among genetic material from viruses[56,57,59]. Reference independent (*de novo*) assembly of non-human reads can also first be done to find contiguous sequences, which may be more easily mapped foreign genomes[60]. Virus detection would also be a situation were paired-end reads are more valuable, since they would better allow the detection of fused human-viral sequences.

Finally, while computational methods can identify genes, mutations, chromosomal rearrangements and viruses that are likely to be positively selected by tumors, functional experiments are always required to confirm their hypothesized roles in cancer development and progression.

1.5 Causes of mutations

Damage to cellular DNA can occur as a result of external genotoxic exposures, or due to internal biological processes. Among the external factors are certain types of radiation, chemicals and viral infections. Among internal causes are inherited defects in DNA replication or repair, reactive oxygen species associated with metabolic stress or inflammation, as well as natural errors that occur during DNA replication in every cell division[61]. Most mutations are of little consequence, but accumulate with age, increasing the risk of cell eventual cell transformation. A number of these factors can be influenced by personal decisions regarding, mainly, sun exposure, smoking, alcohol consumption, physical activity and diet[61].

The implication of risk exposures has traditionally relied on epidemiological studies and findings that specific occupations or habits have been overrepresented among victims of certain types of cancer, which, for instance, led to the discovery of compounds in cigarette smoke as potent carcinogens[58]. However, the recent contributions from large-scale sequencing projects have enabled new ways to study the potential causes of cellular mutations.

Besides the possibility of statistically assessing relations between cancer incidence and individual heritable variants across the human genome, it has also been found that different carcinogens and biological processes are associated with non-random occurrences of mutations along the genome. This has made it possible to define unique genomic mutational signatures that recur across many tumors. Combined with information about known exposures and genomic profiling of inherited variants, these patterns can be highly suggestive about potential underlying causes[53].

1.5.1 Mutational processes and their signatures

Passenger mutations may not contribute directly to cancer development, but their genomic distributions can serve as fingerprints for the processes that have been operative[53]. For instance, C > T and CC > TT mutations have long been known to be common in cutaneous melanoma and other skin cancers[53,62]. These are caused by UV-induced dimer formation between adjacent pyrimidines (C or T), and subsequent failure to repair these lesions before DNA replication[63]. The most common variants are cyclobutane pyrimidine dimers (CPD) and 6–4 pyrimidine–pyrimidone (6-4PP)[63]. Since their repair partly depends on a mechanism coupled to transcription, a bias can be seen for mutations on the untranscribed strand[53,64]. Individuals with inherited

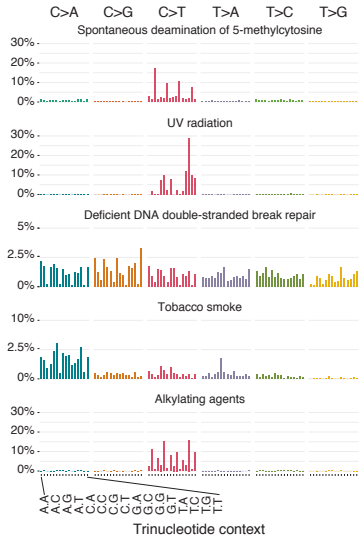

Figure 6. Five mutational signatures which have been found associated with specific mutational processes and/or carcinogens. The y-axis indicates the two bases adjacent to the one that is substituted, which together are referred to as a trinucleotide. Different trinucleotide contexts have different probabilities to undergo mutations, depending on the process generating the substitutions. This can make some signatures useful for discovering new carcinogens or as a biomarker for what may have caused the cancer, and in some cases also where in the body the tumor may have originated from. Data for this figure was obtained from *https://cancer.sanger.ac.uk/cosmic/signatures* (accessed 04-12-2018).

defects in nucleotide excision repair also have a predisposition for developing various skin cancers[63].

However, there are also other processes that can cause C > T mutations. For instance, methylated cytosines (5-methylcytosine) may undergo deamination spontaneously, transforming them to thymine, which may then fail to be corrected by repair mechanisms[65]. In addition, it is likely that multiple mutational processes have operated on the genetic material of a given tumor. In order to separate the contributions of such processes, one may take advantage of the fact that the local sequence context can often influence the probability of substitutions[66]. A landmark study examined the mutation frequencies of each among the 96 possible triplets of adjacent nucleotides that contain the altered base as its central component in over 7000 cancers of diverse types[53]. By a clustering technique known as non-negative matrix factorization (NMF), they were able to distinguish 21 independent components of the overall mutation spectra of the tumors, and their associations with different cancer types.

One of these matched well with prior knowledge of UV-induced mutations, whereas a separate pattern emerged that was compatible with spontaneous deamination of methylated cytosines, the latter of which was also found correlated with the age of diagnosis[53]. Moreover, the analysis also revealed signatures that preferentially occurred in individuals with the habit of tobacco smoking, those exposed to the alkylating anti-cancer drug temozolomide, carriers of mutations in the DNA repair-associated genes *BRCA1* or *BRCA2*, aberrant activity of the cytidine deaminase AID and similar proteins, as well as other patterns that could be tied to specific exposures[53]. Some of these are shown in **Fig. 6**.

Later studies have since extended this work by defining novel signatures or associations with potential risk factors[15,67]. Important to keep in mind, however, is that an association is not a proof of cause. Although the mechanisms that give rise to some of these signatures are well established, further experiments are needed to delineate the exact processes behind the majority. In addition, while the NMF method may be successful with cohorts the size of TCGA, smaller scale studies may not enable the full separation of patterns that are simultaneously present in tumors[53].

Nonetheless, screening of tumors for already known signatures can still be done with as little as one sample, and given the occasionally strong associations with particular exposures or cancer types, these patterns may be repurposed as biomarkers. To do so, one method is to search for the non-negative linear combination of known signatures that best explains the overall mutation spectrum in a tumor[68].

1.6 Transcriptomics in cancer research

The gene expression patterns in cancer cells are shaped by a multitude of factors, including the cell origin, copy number alterations and the activities of oncogenes and tumor suppressors[15]. Different cells may share the same genome, but epigenetic diversification during development, such as methylation or other histone modifications, can lead to large transcriptomic differences. Besides protein coding genes, the genome also encodes at least as many actively transcribed non-coding RNAs (although some studies argue that sizeable fractions of them may in fact produce small peptides[69,70]), which are also often expressed in a tissue-dependent manner[71]. As the convergence of all these factors, the transcriptome can be argued to yield a snapshot of the cell's phenotype at a given time point. RNA-seq can therefore be used to answer a wide range of biological questions.

A frequent use case is the comparison of differences in gene expression between groups of samples, commonly cell lines, subjected to varying experimental treatments. This may be done to determine mechanisms involved in response to drugs, genetic perturbations or for discovering biomarkers. RNA-based biomarkers may be measured in patient material and used for prognosis and determining the most suitable course of action. As an example, the eye cancer uveal melanoma can be divided into two broad subtypes, which are associated with greatly different likelihoods of future metastasis. For some, removing the primary tumor is often enough, whereas for the rest, metastasis is almost a given. A group of genes have been found that are consistently expressed at different levels between these subtypes, and which have therefore been used to develop a classification algorithm, based on a so-called support vector machine, which can distinguish between them. This approach has been found highly predictive of patient outcome[72,73]. As a result, measuring the expression of these genes can identify individuals that will need further surveillance.

Another common use case is the identification of new tumor subtypes. Traditional approaches have relied on histological appearance, that is, the microscopical structure of sections taken from tumors, and their associated expression of established protein biomarkers. In recent years, transcriptome-based approaches on large cohorts have led to refinements of previous classifications for many tumor types, as well as the discovery of associations between these and specific genomic alterations[15,16,74,75]. For the purpose of subtype determination using gene expression data, unsupervised clustering on sample correlations is the most common approach[76]. The term unsupervised implies that no prior knowledge about tumor classifications is used to inform the analysis, as opposed to supervised clustering, which is used to classify

samples based on knowledge of true class memberships in a reference dataset. In the latter case, one may for instance desire establishing the diagnosis of a difficult case by comparing with gene expression data from diverse cancer types.

Clustering can also be used in the inverse fashion, on genes with respect to samples. This can determine subsets of co-expressed genes, which have often been found to participate in similar biological pathways[77], enabling the discovery of new gene functions. Moreover, it is also possible to combine gene expression measurements with other data types in integrative clustering approaches, to find subgroups that share similarities also with respect to copy number changes and DNA methylation profiles, for instance[76].

In addition, RNA-seq may also be used for the determination of which mutations are expressed in a given tumor. Related this is the prediction of potential antigens presented specifically by tumors, which could be targets for immunotherapy approaches[78]. In bulk tumor material, it is also possible to use gene expression data to dissect interactions between tumors and immune cells, since a given tumor is a complex aggregate of multiple cell types, including immune cells. These express very specific sets of genes, which can be used to identify their presence and cellular states. This can be accomplished with computational approaches that are termed deconvolution, which use reference expression profiles associated with specific cell types to estimate their proportions in the sequenced material[79,80]. For instance, the sequencing-based estimation tumor immune infiltrate composition has been found predictive of survival in cutaneous melanoma[81,82].

The complexity of tumor material can also be studied with methods that profile expression in single cells. Recent protocols enable the sequencing of several thousand cells simultaneously, which can be highly informative about intra-tumor heterogeneity, the evolution of sub-clones and interactions with the microenvironment[83].

1.7 The immune system in cancer

The body has a natural defense against cancer development via the immune system, which can monitor cells that acquire extensive mutational burdens and grow too fast. The primary effectors are cytotoxic T-cells, which can recognize antigens presented on the surface of tumor cells via their T-cell receptors (TCR) and induce apoptosis, a cell death mechanism, in those that do not resemble normal cells[84]. T-cells that recognize antigens derived from normally expressed peptides tend to be depleted by the immune system in order to limit autoimmunity, whereas those that recognize foreign material are expanded into clones that target an infected or cancerous cell.

It is likely that this has protected us from a number of naturally occurring pre-cancerous cells that have arisen during our lifetimes. However, cancer cells can eventually become unresponsive to attacks from the immune system. Alternatively, the immune cells may cease to function properly, thereby allowing the transformed cells to spread into systemic disease[84,85]. Restoring the responsiveness of T-cells to tumors is the goal of immunotherapy, which shows promise as a potentially more effective way to treat cancer than traditional approaches[78]. However, tumor-immune interactions are complex, and critical to fully understand in order to improve response rates. As a result, immunogenomics has developed as a novel discipline, which utilizes sequencing data as a basis to dissect these relations[86].

1.7.1 Tumor antigens and T-cell recognition

Tumor cells can betray themselves to the immune system due to fragments of mutated proteins being presented on their surfaces via the major histocompatibility complex (MHC), which is sensed by receptors on T-cells. Antigens derived from intracellular proteins, including any viral products that may be expressed, are presented on class I MHC molecules, whereas those of cell-external origin, such as those derived from bacteria, are preferentially presented on MHC class II[87]. Thus, class I and its associated antigens tend to be the most relevant in a cancer context. Although, it is also possible for MHC II to present endogenous peptides under some circumstances, such as via endocytosis of membrane components or autophagy of internal proteins[88]. MHC I activates CD8+ T-cells (those expressing the protein CD8, a TCR co-receptor), which are the main subset responsible for anti-tumor activity, whereas MHC II activates CD4+ T-cells. However, the latter can also exert cytotoxic activity against tumor cells to some extent. MHC II is, however, not expressed as widely by tumor cells as MHC I, but rather tends to be utilized by antigen-presenting immune cells to indirectly prime responses[88]. For these

reasons, most studies in cancer so far have focused on antigens presented on MHC I.

The main MHC I genes are *HLA-A*, *HLA-B* and *HLA-C*. Their sequence composition is highly variable between individuals, which consequently affects the range of antigens they can present[87]. The high variability is likely due to an evolutionary need to constantly adapt to new pathogens. Similarly, TCR sequences are also highly variable, although to an even greater extent. T-cell receptors are rearranged somatically at the sequence level, which give rise to a broad repertoire of T-cells capable of recognizing a wide range of potential antigens[89]. To limit harmful autoimmune responses, T-cells are therefore subjected to a process that causes undesirable self-recognizing cells to undergo apoptosis.

The process by which TCR sequences are rearranged is termed V(D)J recombination, deriving from the names of the composite fragments, a variable (J), diversity (D) and joining (J) region[90]. These fragments are combined together with a constant sequence component to form the two chains that are the basis for TCRs, α and β, with only the β chain including the D fragment, although a minority of T-cells instead present TCRs that utilize invariant γ and δ chains. The region at the junction of V (D) and J is termed complementarity-determining region 3 (CDR3) and is highly variable in composition, partly due to the combinatorial joining of the constituent fragments, but also a result of the possibility to add or delete nucleotides at their joining ends[91]. This region is the most critical for antigen recognition.

It is currently possible to determine the sequence composition of TCRs at a single-cell level, as a result of recent technological advancements. Since TCR composition essentially tags unique T-cell clones, this makes it possible to study clonal T-cell dynamics in tumor immune infiltrates. Paired with transcriptome sequencing, this further enables determination of the respective activation states of each clone, aiding in the discovery of subsets that have become activated as a result of tumor antigen-derived stimulation[92,93].

Antigens derived from mutated proteins, which are exclusively presented by the tumors, can offer a way to specifically target the latter, while sparing normal cells[78]. Such "neoantigens" can be detected by computational methods from sequencing data (**Fig. 7**). The steps for detection involve establishing the somatic mutations in the tumor, determining the genotypes of the HLA genes in an individual, assessing which mutations are expressed and predicting the binding affinity of the mutated protein fragments to each HLA complex[86]. The peptides recognized by MHC class I are usually around 9 amino acids long and their composition determines the likelihood of binding[94]. Different methods exist for predicting binding affinity. A commonly used algorithm utilizes an artificial neural network model trained

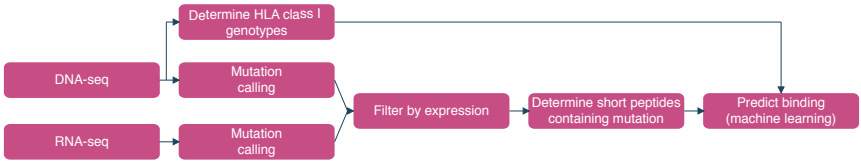

Figure 7. Workflow for computationally predicting neoantigens that may be presented by tumors. After determining expressed peptides containing a given mutation, these are screened for estimated binding affinity to HLA class I molecules, determined from genome or RNA sequencing of the patient. A current state-of-the-art method to predict binding uses an artificial neural network model for this purpose. Biological validation experiments are required to confirm antigen presentation.

on features of peptides known to be presented by specific HLA molecules[94]. However, current methods still tend to suffer from high rates of false positives[95]. Besides inaccuracy in binding predictions, this is partly also due to other complex factors influencing presentation, including whether or not the cellular protein degradation machinery generates the predicted peptides, and whether the cell can successfully transport them to MHC.

In addition to neoantigens, tumors can also be targeted with some specificity by the immune system if they express genes that are not normally expressed by most cell types, such a cancer germline antigens (also known as cancer testis antigens), which are normally restricted to germline cells and trophoblasts[93]. Frequently, it is also found that tumors contain infiltrating T-cells that recognize other highly expressed lineage-specific markers, such those involved in melanogenesis in melanoma[96]. However, therapies activating a response towards such antigens may induce harmful autoimmune responses, as they are also expressed in a fraction of normal cells[96].

1.7.2 Immune evasion

Malignant cells that form tumors and spread are able to do so because they have developed ways to avoid an effective immune response. This can occur via a number of means (**Table 2**). The immune system has a number of checkpoints in place to prevent persistent inflammation and autoimmune problems. One type of checkpoint consists of certain of receptors that can be expressed on the surface of T-cells, which occurs after prolonged antigen stimulation[97]. When bound to their respective ligands, these T-cells refrain from killing their targets. Checkpoint ligands can be expressed by other immune cells in the tumor microenvironment, but in some cases also by the cancer cells themselves. As this begins to occur, T-cells start to enter a state known as exhaustion, limiting their capacity.

In addition, suppression of immunity also occurs as a result of a the activity of a specific class of T-cells known as regulatory T-cells, which are

often found highly represented near the tumors[98]. Other suppressive cell types can also contribute to a failed anti-tumor response, such as M2 macrophages and myeloid-derived suppressor cells. Transition of macrophages towards the suppressive M2 state can additionally be promoted by lactate production from cancer cells. Lactate is a byproduct of glycolysis, which most cancer cells are highly reliant on[98].

Furthermore, cancer cells can downregulate or mutate genes used for antigen presentation (*HLA-A*, *HLA-B* and *HLA-C*), genes involved in the transport of peptides to MHC (*TAP1*), or β_2 microglobulin (*B2M*), which is a critical component of functional MHC I[85]. As a result, T-cells are no longer able to identify the cancer cells as unusual. Normally, NK cells can recognize loss of HLA expression and target the tumor cells[85]. However, tumors can also express molecules inhibiting NK cell activity, including both checkpoint ligands such as PD-L1 and soluble ligands that bind to the receptors used by NK cells to identify cells that are missing antigen presentation, essentially diverting their attention[99].

Moreover, tumors can also promote the exclusion of immune cells from their local environment. This can occur through the expression of certain chemokines, molecules influencing cell migration, which act to keep subsets of cytotoxic immune cells from entering[98,100]. The same chemokines, some of

Table 2. Common immune evasion mechanisms. Intrinsic refers to mechanisms mediated by the cancer cells themselves, whereas extrinsic refers to factors of the tumor microenvironment. Most of these aspects are possible to study using either RNA or DNA sequencing data.

Intrinsic	Extrinsic
• Expression of immune checkpoint ligands (for instance PD-L1).	• T-cell exhaustion, anergy or senescence.
• Resistance to apoptosis, which prevents T-cells from inducing cancer cell death.	• Regulatory T-cell presence, which suppresses the activity of cytotoxic T-cells.
• Defective antigen presentation, which can occur due to downregulation or mutation of critical genes, for instance *B2M*, *TAP1* and *HLA-A*, *-B* and *-C*.	• Expression of checkpoint ligands by other immune cell types in the tumor microenvironment, including myeloid-derived suppressor cells, tumor-associated macrophages and dendritic cells.
• Cancer cell expression of chemokines contributing to exclusion of cytotoxic T-cells from the tumor microenvironment.	• Expression of chemokines contributing to exclusion of cytotoxic T-cells by other cells in the tumor microenvironment.
• Production of lactate, which can induce transition of tumor-associated macrophages towards the suppressive M2 phenotype.	• Physical exclusion of immune cells from the tumor environment via, for instance, a thick extracellular matrix created by cancer-associated fibroblasts.

which can also be expressed by other cells in the microenvironment, may, however, still allow for the entrance of immune suppressive cell types, including regulatory T-cells. Physical exclusion may also occur due to development of an impenetrable extracellular matrix, which cancer-associated fibroblasts can contribute to[98].

1.7.3 Immunotherapy

Knowledge of the mechanisms whereby T-cells eventually fail to eradicate tumors has led to the development of therapies aimed at reinvigorating their activity. Among the most successful of current immunotherapies are those that target immune checkpoint receptors or ligands, most commonly PD-1, PD-L1 or CTLA-4, through the use of inhibitory antibodies[78]. A number of cancer types can display great responses to these, most prominently cutaneous melanoma, but far from every patient receives any benefit. On the pessimistic side of the spectrum are cancers where these treatments almost completely lack effect, for instance uveal melanoma[101]. An increased understanding of the operative immune evasion mechanisms that determine outcome will be required to design more effective options for these patients.

Another immunotherapy approach utilizes chimeric antigen receptor T-cells (CAR-T), where T-cells are modified to express TCRs that recognize tumor specific antigens[78]. With this option, it is of importance to ensure that other cells in the body are not targeted, which might be the case if the antigens are not exclusive to the tumors. It also remains a risk that the TCRs used may cross-react with unknown antigens presented by normal cells[78]. Potentially, neoantigens could be useful as targets with CAR-T, although most studies have focused on other ones derived from genes expressed with specificity in tumors[102]. A third class of immunotherapies are vaccine-based. With these, antigenic peptides themselves are used to prime the immune system, which can be accomplished by a variety of means. For instance, dendritic cells (DCs), which present antigens to T-cells, can be loaded with such peptides and transferred to patients to prime subsequent responses. It is also possible to deliver tumor-specific peptides themselves into the patient, aimed at achieving a similar indirect effect. Responses with these therapies have not been as promising as for checkpoint inhibition, however. Although, it is possible that combinations of them may improve outcomes[103].

Factors known to associate with the success of immunotherapies include, among others, total mutational load, indirectly implying neoantigen load; the proportion of CD8+ T-cells infiltrating the tumor, the expression of immune checkpoint genes, presence of genomic alterations affecting antigen presentation components, as well as expression signatures associated with

immune cell exhaustion or exclusion[28,81]. All of these factors are possible to study to some extent with genome and transcriptome sequencing[93].

2 Aims

This thesis aims to investigate the range of answers that can be derived from the genomes and transcriptional material of tumor cells, to questions concerning cancer development. The included papers focus on a range of topics in cancer genomics, from tumor classification to profiling of driver events and immune landscapes. Material from large public resources of tumor sequencing data is utilized, as well as patient material from Sahlgrenska University Hospital. The focus of each paper is as follows:

I Determination of the origin of a metastatic cancer of unknown primary based on DNA and RNA sequencing.

II Development of a method to prioritize genes of interest in focal copy number changes based on integration with gene expression data.

III In-depth molecular characterization of uveal melanoma metastases, concerning driver mutations, mutational spectrum, genes of interest in recurrent broad copy number changes and the phenotypes of tumor-infiltrating T-cells.

3 Results and discussion

3.1 Paper I

Cancer of unknown primary (CUP) is the diagnosis arrived at when the patient presents with metastatic disease but it is not possible to determine the primary cancer type or its site of origin. Knowledge of this is important for deciding an appropriate and effective treatment. As a result, these patients are most commonly treated with general cytotoxic chemotherapies, which tend to be unsuccessful and can have a wide range of undesirable side effects. Naturally, mortality within this group of patients is very high[104]. However, the tumors may have specific driver mutations for which targeted drugs could be available. This alone would justify some level of genomic profiling. But even so, these drugs may not work equally well in all cancer types driven by the same mutation[105], and some drugs have only been approved for use in specific cancers. Therefore, tumor classification is also a priority.

In this study, we encountered a patient that had been misdiagnosed three times previously. The tumor was originally thought to be a lung cancer with an oncogenic fusion involving *ALK*, but later re-evaluation discovered this to be a false positive finding. Due to prominent neuroendocrine characteristics, the diagnosis was changed to a neuroendocrine lung cancer. A third re-evaluation confirmed the neuroendocrine-like phenotype, but favored the diagnosis of paraganglioma, albeit an unusual case with a $BRAF^{V600E}$ mutation. This mutation is very common in melanoma[12]. However, all decisive melanoma markers examined by immunohistochemistry on both a brain and subcutaneous metastasis were negative. Fortunately, this mutation is possible to treat with a BRAF inhibitor, which was also done here with temporary success, although resistance eventually developed. At this point in time, whole genome and transcriptome sequencing was performed on both of the metastases.

3.1.1 Establishing the origin of a metastasis with sequencing data

Genomic profiling uncovered two potential resistance mutations in *MAP2K1* and a homozygous focal deletion of the tumor suppressor *CDKN2A*. Transcriptome analysis confirmed high expression of neuroendocrine and neural crest markers. No known oncogenic fusions were discovered, of relevance since some paragangliomas have been found positive for fusions involving *MAML3*[106].

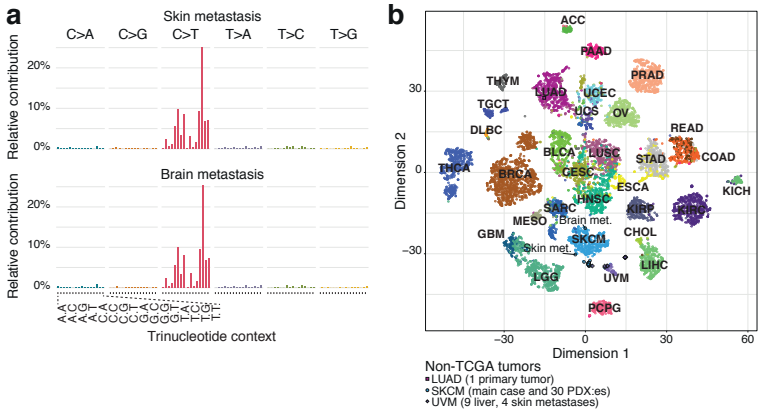

Figure 8. a) Mutational spectrum, as measued by overall trinucleotide substitution frequencies in a subcutaneous and brain metastasis of the patient (compare with **Fig. 6**). **b)** Gene expression data from both metastases were compared with 9583 tumors from 32 cancer types available from TCGA. t-SNE analyses grouped both samples with cutaneous melanomas. The same cancer type was also predicted using an independent 6-nearest nighbor classification based on Spearman correlations.

Following this, an analysis of the mutational spectrum of the tumor, considering synonymous and non-synonymous exonic variants, revealed a prominent mutational signature associated with UV-induced damage (**Fig. 8a**), questioning the previous diagnosis. The contribution of this signature to the overall mutational load was estimated to 80%. The pattern was identical in both the skin and brain metastasis from the patient. This precluded a primary origin in the lung region. A transcriptomic comparison to gene expression profiles of over 9500 TCGA tumors from 32 different cancer types showed strong matches to cutaneous melanoma (CM, **Fig 8b**). As a result we could firmly establish the diagnosis for this patient.

The strong expression of neural markers was likely related to the fact that melanocytes derive from neural crest or, alternatively, Schwann cell precursors[107]. This, and the lack of expression for decisive melanocyte markers may be explained by either a dedifferentiation towards an earlier developmental state of the cell or transformation of a precursor present in the skin.

3.1.2 The utility of DNA and RNA-seq for cancers of unknown primary

These findings demonstrate that DNA and RNA-seq analysis can be highly useful in the study of cancers that are difficult to categorize with traditional methods. A novel aspect here is that the overall mutational spectrum in the tumor can reveal important clues, since a signature of carcinogens only expected to be encountered in very specific parts of the body can exclude a range of alternatives. In fact, we found that the UV signature could even discriminate between TCGA head and neck squamous cell carcinomas that arise on the lip from those that arise elsewhere in the oral region. Furthermore, analysis of the TCGA lung squamous cell carcinoma cohort also revealed three samples with prominent UV signatures, arguing that they were actually metastases from elsewhere. Although not all known mutational signatures are likely to be equally informative, the analysis is relatively straightforward if sequencing data is already available. The potential discovery of therapeutically actionable driver mutations, copy number changes and fusion genes is already a strong argument for performing genome and transcriptome sequencing. The latter can also be used for mutation discovery and, when paired with DNA-seq, further allow the exclusion of false positive variants. Standardized analysis pipelines are possible to implement and can potentially yield important information in a short time period.

Furthermore, recent global sequencing projects, i.e. TCGA, have made available massive datasets for numerous cancer types, which are possible to use as a reference in transcriptomic comparisons. Such analyses can evidently uncover details missed by traditional immunohistochemical staining approaches. We evaluated this method of classification using leave-one-out cross-validation on each tumor in TCGA and found accuracy to be 95%, with 84% accuracy on an independent dataset.

The latter was mainly composed of uveal melanoma (UM) metastases and CM patient-derived xenograft (PDX) models, which may be argued to be somewhat limited. The lower performance on this dataset, however, highlighted some important flaws with the approach that one should be mindful of. Namely, the purities of metastases are likely to influence the results, since the presence of hepatocytes in a fraction of UM liver metastases led to misclassification of some as liver hepatocellular carcinoma. In addition, the TCGA cross-validation also showed that tumors deriving from similar cell types had higher misclassification rates. Yet another issue to be aware of is that the TCGA data is not fully comprehensive, in that it lacks both some common cancer types, such as cutaneous squamous and basal cell carcinoma, as well as a number of rare cancers. The latter may be alleviated with time as future

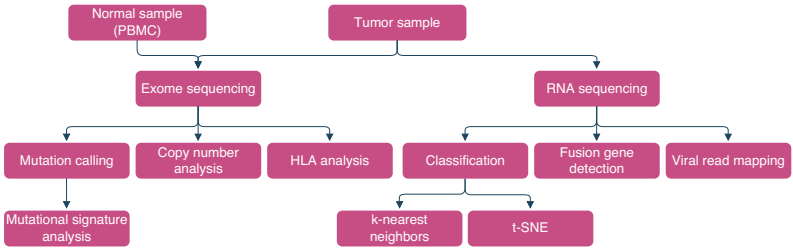

Figure 9. Illustration of information that can be extracted from RNA and exome sequencing, which may aid in tumor classification and treatment selection. PBMC: peripheral blood mononuclear cells, commonly used as a control sample of normal cells.

sequencing studies contribute more data. A workaround for purity issues could be to develop an integrative classification approach that also considers mutations and copy number profiles, since these also possess some amount of distinctive qualities.

For the application of this framework on CUP, additional factors need to be taken into account. Material from such tumors presently tends to be stored under formalin-fixed paraffin-embedded (FFPE) conditions, which is associated with artifactual base substitutions, as well as RNA degradation[108]. Epigenetic assays have been described that may work better for such samples[109], although the two methods remain to be compared. Epigenetic profiling is also likely to be cheaper. Ideally, however, DNA and RNA-seq can provide additional information that could be highly useful, such as the detection of targetable fusion oncogenes, mutations, presence of viruses and immunogenomic profiling (**Fig. 9**). This could potentially justify the additional costs associated with sequencing. Naturally, one could also argue for improved storage conditions of CUP samples going forward.

In all, this study highlighted the fact that much more can likely be done for the group of patients whose tumors are currently classified as indeterminable and, in most cases, largely untreatable. We can most likely assign this group to more effective options by comprehensively profiling their disease molecularly, however one may decide to do this in the most cost-effective way.

3.2 Paper II

Genomic instability can cause the accidental duplication or loss of genetic material. Such copy number changes can activate oncogenes or inactivate tumor suppressors. If these events provide a selective advantage, those cells will expand in numbers relative to those that lack them. It is commonly found that tumors from multiple patients have somatic copy number changes affecting the same genomic regions. This implies that those changes may have contributed to the formation and survival of the cancer. Studying these recurrent events can give clues about potential driver genes, which may suggest new therapeutic targets[11].

However, the affected regions in a given tumor often span many genes, making it challenging to identify the ones that are most relevant. Therefore, one typically searches for regions of minimal overlap between recurrent changes, which indicate genes that are altered in the majority of tumors. However, random inconsequential events may also overlap such a locus, shifting the minimally overlapping region away from the actual driver gene, and additionally, such regions may still contain multiple candidates[43].

3.2.1 Recurrent focal copy number and gene expression changes

The primary means by which copy number changes can drive cancer is via changes in transcriptional output. However, not all genes that experience copy number gains or losses are markedly affected in terms of expression. Therefore, to further limit the search space for genes of interest, one may integrate the information obtained from DNA with that of RNA-seq. The natural approach for this would be to assess correlations between copy number and expression for each gene. This has a number of weaknesses, however. On one hand, the effects may not always conform to the types of trends assessed by correlation methods. On the other hand, larger simultaneous changes in copy number and expression, or smaller amplitude gains with large effects, would be expected to provide stronger evidence than others, but these instances will not be rewarded to any additional extent in the calculation of a correlation coefficient. Furthermore, these approaches do not take into account the fact that more size-limited alterations are more informative about the relevance of each gene in the overall region that is altered across tumors.

Here, we developed a screening tool that instead uses an approach similar to covariance, where stronger dosage effects will have a greater relative contribution to the final score of a gene. As a basis for this calculation, changes

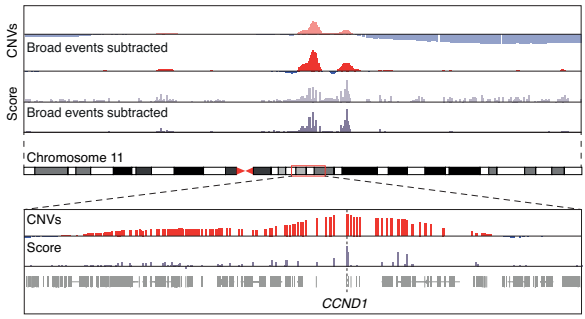

Figure 10. Illustration of scores calculated by the developed method (FocalScan) in regions of recurrent copy number change, based on breast cancer data from TCGA. Upper panel: summed $\log_2$ ratios of copy number changes relative to a diploid chromosome, contrasted with the corresponding profile for copy number changes where broad events have been subtracted. The respective scores calculated for each gene by integrating these changes with alterations in gene expression are shown below. Lower panel: zoomed view of the region surrounding the gene *CCND1*, which was the most recurrent amplification event on this chromosome.

in gene expression or copy number relative to the median values in samples diploid at a given genomic position is used. In addition, the scoring system prioritizes focal events by subtracting broad changes from the copy number amplitudes in each sample prior the calculation. The hypothesis is that the best-ranked genes in each region are more likely to be driver genes than lesser-ranked ones. The method can be used either on gene-level or on continuous intervals across the genome (**Fig. 10**), facilitating analysis also of unknown transcripts. To automatically find regions across the genome with clustered high scores and select the strongest candidates, a peak detection algorithm is then used.

3.2.2 Peak detection

Detecting the type of genomic peaks resulting from these scores is a challenge, since the difference between a peak and a sub-peak is not easy to define in a way that consistently works across the genome and across different cancer types. It is also dependent on the desired level of sensitivity and specificity, since one can theoretically call each individual score a peak, and miss no genes, but be unable to interpret the results. Conversely, one may also call an entire chromosome a peak and retrieve only the highest ranked gene within it as a candidate. Most evaluated algorithms for peak detection, mainly developed for signal processing, were also found to rely on assumptions that all peaks conform to certain statistical distributions, which is not the case here, since genomic elements are unevenly distributed and genes often have uncorrelated

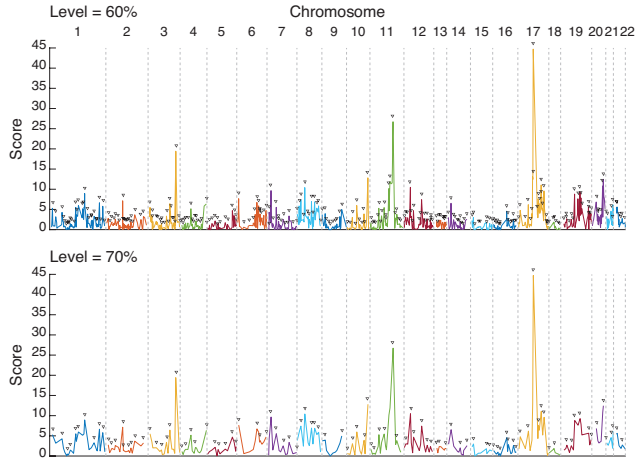

Figure 11. Peak detection performed on scores derived from TCGA breast cancers. Detection can be performed on multiple levels of granularity (specified as a percentage), two of which are shown here. A higher value gives fewer peaks, but may miss genes of interest that form peaks in close regions, whereas a lower value will nominate more candidates by dissecting larger peaks into sub-peaks, but this may also yield noisier results.

levels of transcription. These methods also required a large number of user-defined parameters. Settings that worked well on one cancer cohort often failed on another.

The proposed solution to this problem works by first considering each gene that has a higher score than its two neighbors as a peak. Then, a higher level of peaks is defined as those peaks from the previous level that are greater than their two neighbors. This process repeats until only one gene on the entire chromosome is considered a peak. Depending on the desired degree of specificity, one can then select an intermediate level of choice to obtain a reasonable number of peaks across the genome (**Fig. 11**). The peaks formed this way are independent of the distances between genes and between sub-peaks, and the method does not assume any specific statistical distribution. The only parameter needed is the desired choice of specificity, which can be specified as a percentage relative to the total number of levels generated. The latter also made it possible to run the algorithm with comparable performance on different cohorts using the same setting.

The combined use of a scoring system that weighs recurrence, focality and coordinated changes in copy number and gene expression together with this peak detection approach led the method to outperform other similar methods designed for ranking potential driver candidates, as defined according

to enrichment of known cancer driver genes within the ranked list of putative targets generated.

3.2.3 Recurrently altered unknown transcripts

Recurrent focal copy number changes can sometimes be found in regions where no known genes exist, which are also relevant to consider. This was not an option with other current approaches. Therefore, we added the ability to calculate scores for continuous small sections of the genome, thus making the method optionally independent of a reference genome annotation.

Fig. 12 shows one such unknown transcript nominated by the method as recurrently focally amplified and upregulated in TCGA breast cancers. This occurred in an intronic region of *PVT1*, but did not correspond to any alternative exon of this gene. *PVT1* is adjacent to the oncogene *MYC*, but the nominated transcript received a higher score than either. Later studies have shown that the *PVT1* locus contains a cluster of enhancers, some of which can control *MYC* expression[110]. One such enhancer region, as determined with a CRISPR interference screen, corresponded precisely to the location of this peak[110]. The transcript produced here is likely an enhancer RNA (eRNA), which is a class of non-coding transcripts produced at enhancers that have been found correlated with their activity[111]. Recurrent focal amplifications of enhancers that lead to *MYC* overexpression have recently been found in lung adenocarcinoma, endometrial carcinoma, T-cell acute lymphoblastic leukemia and neuroblastoma[112-114]. This illustrates that the method is capable of nominating candidates for further experimental studies also in regions without known genes.

A weakness of the approach is that these ranked lists are only indicative of the relative strengths of association of each genomic element with factors that relate to selection, but are not direct statistical tests for selection as such. Thus, the output is not a statement that a given element is under selection, but merely an aid in the prioritization of genes within regions of recurrent copy number changes. There does not currently seem to exist any integrative method that statistically assesses positive selection. Most likely, this is due to the difficulty in defining a null model for what constitutes RNA levels that are not under selection. To achieve a more statistically based assessment of possible selection using the above described method, one may focus the analysis on regions where copy number aberrations have been determined to occur more frequently than expected by chance, for which there are existing methods available to assess[43].

In addition, the method benefits greatly from a large number of samples being used, preferably in the hundreds, which is available for TCGA

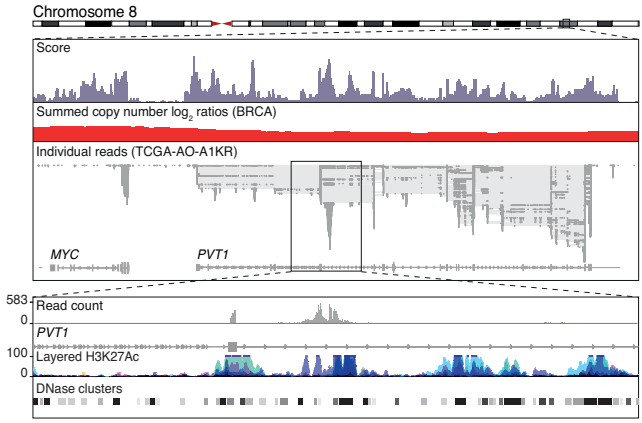

Figure 12. Non gene-centric scores in a focally amplified region, followed by peak detection, nominates an unknown transcript intronic to *PVT1*. The transcript coincides with histone 3 lysine 27 acetylation and DNase clusters indicative of an active enhancer.

data, but typically not with in-house generated datasets. From a usability standpoint, the method could also be improved by outputting a merged list of candidates across different peak calling levels to enable a comparison of all top candidates that may be nominated with different settings in a given region, since some true driver genes present in what could be considered sub-peaks may otherwise be missed. A further limitation is the concentration on focal events. Several cancer types display frequent, if not exclusively, broad chromosome arm-level changes. Currently, few, if any, studies have been made that properly investigate candidates within these, simply due to the vast number of genes they contain and the absence of any minimally overlapping region to guide the analysis. We attempt to address this issue to some extent in Paper III.

3.3 Paper III

Uveal melanoma (UM) is a rare eye cancer affecting the choroid or ciliary body, or less commonly the iris, which together constitute the uvea. However, it is also the most common form of eye cancer[115]. The biology of UM is distinct from cutaneous melanoma (CM), with mostly unknown underlying causes, as well as driven by a different set of oncogenes and tumor suppressors[12,16,115]. Approximately 50% of patients develop metastases, mainly to the liver, which is closely tied to the presence of inactivating genomic aberrations in the tumor suppressor *BAP1*[115]. Treatment of primary UM most commonly involves enucleation of the affected eye or destruction of the tumor with radiation[116]. In the metastatic state, however, no effective treatments exist. While immunotherapy has been highly successful in CM, little to no responses are seen in UM. Current targeted drugs and general cytotoxic therapies also lack efficacy[115,116].

To gain additional insights into the genomic characteristics of metastatic UM, we have examined the whole genomes and transcriptomes of metastases from 24 patients. We additionally profiled tumor-infiltrating T-cells at the single-cell level from eight patients with respect to their transcriptomes and T-cell receptors (TCRs), to better understand their phenotypes and how they may relate to the failure of immunotherapy.

3.3.1 *BAP1* loss is frequent and drives transcriptomic changes towards a metastatic phenotype

We discovered mutations in the established driver genes *GNAQ* and *GNA11* to be present in approximately half of the samples each, in addition to less frequent mutations of *SF3B1*, *CYSLTR2* and *PLCB4*. Furthermore, we found inactivating mutations in *BAP1* in 23/24 samples (96%), consistent with the well-known association of *BAP1* loss-of-function and metastasis. In the vast majority of cases, this was also coupled with loss of one copy of chromosome 3 (monosomy 3), where *BAP1* is located, thus contributing to gene dysfunction via LOH. These events are known to frequently co-occur in poor-prognosis tumors[116].

Some of the mutations in this gene also affected splice sites. The pairs of nucleotides adjacent to each exon are considered the ones most critical to splicing[117]. Two tumors had base substitutions affecting these regions, and RNA-seq from the same tumors displayed elevated intron retention around the affected exons. The result of this can be transcripts that are unable to become successfully translated into functional proteins[26]. In another sample, an intronic mutation was found directly adjacent to a part of an intron that

was retained, but distant from any splice site. The sequence motif at this position also suggested the creation of a new intronic splice site. Common variant filtering practices frequently disregard intronic non-splice site mutations, but this finding would serve as an example for further considering them. This may be especially important if a single gene inactivation is predictive of future metastasis, as is the case in UM.

BAP1 has also been discovered as a tumor suppressor in a number of other cancer types, including renal cell carcinoma and mesothelioma[118]. Several mechanisms for its anti-tumor activity have been proposed, including, but not limited to, participation in the DNA repair machinery, cell cycle and metabolic regulation, apoptosis, ferroptosis (a cell death program induced by oxidative stress) and regulation of differentiation[119-127]. Given the variety of mechanisms proposed in different cancer types, it is possible that the role of *BAP1* is context specific. This is likely, since *BAP1* is a histone deubiquitinase, which enables potential epigenetic regulation of a large number of genes[128]. The set of genes susceptible to regulation is likely to vary between cell types due to differences in underlying epigenetics states. Furthermore, it has been found that small interfering RNA (siRNA)-mediated knock-down of *BAP1* in UM cell lines leads to downregulation of melanocyte lineage markers, suggesting dedifferentiation, which could be counteracted by inhibiting the histone deacetylase *HDAC1*, which also implies an epigenetic mechanism[122].

To gain further insights, we reintroduced a functional copy of the *BAP1* gene, using a viral vector, into a tumor-derived cell line possessing a dysfunctional variant due to a homozygous frameshift deletion. We then compared gene expression in this tumor to cells transfected with an empty vector control. A large number of genes were significantly differentially expressed. However, we also noted that 7/12 genes included in an expression-based clinical prognostic test that accurately discriminates between the two major UM subtypes were among these[72,73]. These two subtypes are commonly referred to as class I and II, respectively. These genes were all altered in the same directions prescribed by the test as indicating a good-prognosis subtype upon *BAP1* reintroduction. We examined this further by testing for the enrichment of two larger sets of genes previously identified as differentially expressed between these subtypes[129]. This revealed a strong enrichment for class I genes among those upregulated upon reintroduction, and equally strong enrichment for class II genes among those downregulated. This suggests a broad transcriptomic shift towards a better-prognosis phenotype, or conversely that loss drives cells towards the metastatic phenotype. Furthermore, it also implies that the factors that determine the two major UM subtypes are regulated by *BAP1*. The broad response, comprising a varied set

of cellular pathways, would also be compatible with epigenetically mediated regulation.

3.3.2 Evidence of UV-induced damage in an iris melanoma

The reasons UM develops are still unknown, but a number of risk factors have been identified. These include light eye color, Caucasian ethnicity, geographical latitude and welding[130]. While sun exposure may be easily suspected based on some of these, and due to its role in CM, it has never been conclusively shown that this is a true risk factor. Some of the known risk factors could also have an origin in a genetic predisposition common to Caucasians. Adding to this is the results of several UM genomic studies that have assessed the mutational spectrum of the tumors, and found no indication of the well-known UV signature (**Fig. 6**)[131,132]. This suggests that UV-induced damage is unlikely to play a major role in the development of UM.

We investigated the contribution of established mutational signatures to each tumor in this cohort, and included four simultaneously sequenced CMs for reference. When clustering all tumors on the estimated relative contributions of each signature, we found that UMs possessed distinct profiles compared to CM. The main contributing signatures in UM were associated with aging and certain DNA repair deficiencies, but one was also found that is not currently assigned any known underlying mutational process. CMs, on the other hand, were mainly characterized by contributions from the UV-associated signature, as expected. There was, however, one exceptional UM that clustered distinctly together with the CMs and which displayed a prominent contribution from the UV signature, which has not before been observed for UM tumors. By additional analyses, including a transcriptomic classification similarly to Paper I, we were able to exclude the possibility of a sample mix-up or misdiagnosis. The likely explanation for this case is that the tumor was an exceedingly rare specimen that developed in the iris, which is estimated to only occur in about 5% of cases[130]. Arguably, the iris is more exposed to UV light than the choroid or ciliary body of the uvea. This provides DNA-level evidence for the involvement of UV damage, but suggests that it may only be a potential risk factor in iris melanomas.

3.3.3 Copy number and gene expression changes associated with metastasis

UM predominantly harbors arm-level copy number changes, with no oncogenes or tumor suppressors described as targeted by recurrent focal changes in recent large-scale sequencing studies, besides *BAP1*[16,116,132]. However, several of these broad changes are highly recurrent and some, such

as gain of 8q and loss of 6q, besides loss of 3, are also significantly associated with metastasis[116,132-135]. Consistent with this, we also found that gains of 8q and losses of 6q were significantly overrepresented in the metastasis cohort compared to TCGA tumors, which are all primary. In addition, we also found 17p loss and 5p gain overrepresented. Notably, we also discovered focal deletions affecting the tumor suppressor *CDKN2A* in two tumors, which have not been described in other recent UM genomics studies[16,131,132,136]. Since most studies have focused on primary tumors, the occurrences of this event here could suggest that they are late events in the evolution of the tumors. One may speculate that loss of *CDKN2A* could potentially influence metastatic progression.

The presence of clinically prognostic arm-level copy number changes motivated us to search for genes consistently altered in expression within these, and prioritize among them on basis of protein-interaction data to find genes that may exert broader effects on cellular phenotypes when perturbed. The latter would be expected for any gene putatively conferring a selective advantage to the tumor cell. It should be noted here, however, that copy number changes might also inactivate tumor suppressors through LOH combined with a mutation, or potentially result in fusion genes at their breakpoints. However, no significantly recurrent mutations or fusions have been observed for genes in the regions subjected to these aberrations, leaving expression changes as the most probable mechanism for mediating an advantage.

Based on these assumptions, we screened for genes that had consistent associations between expression and copy number in both our own cohort and TCGA UMs. We then ranked them based on their numbers of interacting partners in a curated protein-protein interaction database[137], and binary membership in pathways previously identified as perturbed overall in TCGA UMs[16,138]. The former database would constitute pre-existing evidence suggesting a potential for impact. Use of the latter method relies on the *a posteriori* argument that pathways objectively altered by genomic events in UM contain these genes, and therefore members of these pathways would be more likely to influence mechanisms of relevance than unrelated genes, even though the latter may still have large theoretical interactomes. Lastly, we also considered univariate expression associations with survival in the TCGA cohort, but only used this to further rank within the list already established from the previous analyses. The motivation for not considering survival data as stronger evidence was that it is easily confounded by numerous genomic and clinical parameters, which are difficult to adequately adjust for, hence also the choice of univariate survival statistics.

By thus ranking genes based on these evidence classes, a number of candidates emerged, which by definition were consistently associated with the underlying copy number events and possessed a potential for a wider phenotypic impact. Examining the top candidates, we found genes in regions of loss to be enriched for apoptosis pathways, whereas genes in gained regions were enriched for hemostasis, integrin signaling and a few other pathways that have some relation to *GNAQ*/*GNA11* activity. Furthermore, we noted a significant overrepresentation for genes upregulated in class II tumors[129] in regions of gain and class II downregulated genes in regions of loss, indirectly validating the associations in a third independent dataset. Altogether, this suggests that the recurrent broad copy number changes in UM are unlikely to be functionally inert, and argues against the notion that these events may only influence tumor phenotypes via general side effects of aneuploidy.

Among the first-ranked genes for each region were some of particular interest. For instance, *CASP9*, which initiates the apoptotic cascade[139], experienced recurrent loss and downregulation on chromosome 1p. The highly prevalent gain of 8q was associated with persistent upregulation of *PTK2* (focal adhesion kinase), which suppresses an apoptotic program initiated as cells lose contact with their surroundings (anoikis), thus enabling anchorage independent growth, a hallmark of metastasis[140,141]. Consequently, *PTK2* overexpression has been associated with metastasis and worse survival in a number of cancer types[140]. Such a role would also be consistent with our observation that each tumor in the metastasis cohort had 8q gains, and with previous studies establishing their association with UM metastasis. Furthermore, *CDH1* (E-cadherin) underwent loss and downregulation on 16q; a gene that is important in contact inhibition, a mechanism that suppresses proliferation and migration when the cell is in contact with other cells[142]. Naturally, loss of *CDH1* has also been associated with metastasis in a several cancers[142]. Given that metastatic UM currently has no effective targeted treatment, further functional studies of the above genes may be of interest[143,144].

3.3.4 Phenotypes of tumor-infiltrating T-cells

The high failure rate for immunotherapy in UM motivated us to profile the phenotypes of tumor-infiltrating T-cells, the primary immune cells that mediate anti-tumor activity. To do so, we performed paired transcriptome and TCR sequencing of close to 30000 individual T-cells infiltrating the tumors of eight patients. By transcriptomically classifying immune cell subpopulations using reference gene sets derived from pure cell types, we discovered heterogeneous immune cell communities. One sample in particular

hardly contained any infiltrating cytotoxic CD8+ T-cells, but was rather dominated by CD4+ cells, some of which were regulatory T-cells that have a suppressive role. This tumor also had the lowest expression of the MHC class I genes, offering a potential explanation for the low level of infiltrating CD8+ cells, since the former are responsible for attracting the latter to their targets.

A theme that recurred across all samples was widespread expression of inhibitory receptors on T-cells. Notably, expression of the molecules LAG3, TIM-3 and TIGIT was more prominent than PD-1 or CTLA-4, against which current checkpoint blockade immunotherapies are directed. This additionally indicated that the T-cells were in an exhausted state. In support of this, we also found lower levels of cell cycle gene signature expression in some of the samples with the highest fractions of T-cells expressing these receptors.

However, the presence of certain exhaustion markers is also indicative of past antigen recognition and tumor-reactivity. Adding to this were greater levels of clonal expansion within tumors with more exhausted communities, as inferred by studying subsets of cells with identical TCRs. While we did not discover any solid evidence suggesting the involvement of neoantigens in such tumor recognition, we were able to determine binding to common melanoma antigens by T-cells from four tumors, using functional assays.

Thus, we could conclude that UM metastases do harbor tumor-recognizing infiltrating T-cells, but that they tend to have highly exhausted phenotypes. It is also possible that PD-1 and CTLA-4-directed immunotherapies fail due to expression of the additional checkpoint receptors LAG3, TIM-3 and TIGIT.

In summary, this study has offered insights into the consequences of *BAP1* loss, highlighted a potential new risk factor for UM development, nominated genes of interest in recurrent broad copy number aberrations and provided a characterization of the landscape of infiltrating T-cells in metastatic tumors. This knowledge may open up new opportunities for research on potential therapeutic targets and alternative strategies for immunotherapy in UM.

4 Conclusions and future perspectives

Ultimately, cancer is an evolutionary issue. Cancer arises because a few cells in our body develop a selective advantage due to genomic changes. These then carry out their own egotistical lives without care for whether or not the remaining cells in the organism agree with it. They spread and colonize anywhere they can, eventually causing the dysfunction of critical organs. Surgically removing the tumors they form eventually becomes fruitless, since they may be deeply infiltrating vital sites that cannot be compromised by such procedures, or remain hidden in obscure places. Cancer is also the consequence of defects in our own biology, which lead to the progressive accumulation of mutations and other genomic aberrations as cells divide during aging. Such accumulation can also be potently accelerated by various external carcinogens, some of which we have a more or less inherited tendency to become addicted of exposing ourselves to.

As human evolution only cares about whether or not we successfully reproduce and raise our offspring before we die, there is no advantage to be gained from keeping us alive after this mission is accomplished. Therefore, we have not developed stronger defense mechanisms against this class of disease, which tends to present itself at advanced ages. If we wish to extend our healthy lifespans, defeating cancer is critical. To do so, extensive knowledge about human genomic evolutionary mechanisms is required. Massive advances have been made as a result of the development of high-throughput sequencing technologies. These, in turn, generate enormous amounts of data in laboratories across the globe. Fully interpreting this data requires, above all else, efficient interaction between biologists, clinicians, and those of computational predisposition.

This thesis has focused on bioinformatical explorations, the intersect of biology and information science. The works included demonstrate how DNA and RNA-sequencing can be used to track the origin of tumors, nominate candidate genes that may influence cancer phenotypes, discover processes that generate cellular mutations and dissect the immune environments of tumors. A desirable end goal in line with the present work would be the standardized implementation of a number of these approaches to comprehensively genomically profile the tumors of each patient, in order to discover any opportunities presented by our enemy. This should ideally be followed by the rational customization of treatments for each unique condition and an instant win accompanied by a suitable victory theme.

Acknowledgements

The following individuals deserve acknowledgements for important contributions to projects involving the author of this thesis under the terms of the present employment, listed here in the glorious font of Helvetica. Names are ordered according to position of formal authority over the one who wrote this document, authorship and the initial involvement date of the previously referenced thesis-maker in each project. While this list is of limited scope, credit is also well deserved by remaining collaborators, who will no doubt experience a certain level of upset upon discovering their absence herein, should they ever decide to peruse the pages bound within this tome. Finally, apologies are reserved for all those unfortunate beings accidentally excluded from or misrepresented in this list due to memory issues or logical errors emerging from an overheating central processing unit.

Supervisor:	Erik Larsson Lekholm
Co-supervisor:	Anders Ståhlberg
Collaborators (main project leaders):	Per Lindahl
	Jonas Nilsson
	Roger Olofsson-Bagge
	Göran Landberg
Collaborators (main contributions, beyond those of the former category):	Thomas Kroneis
	Emma Jonasson
	Berglind Einarsdottir
	Elin Forsberg
	Akif Demir
	Nadia Gul
	Lisa Nilsson
	Suman Mitra
	Samuel Alsén
	Paul Fitzpatrick
	Lars Ny
	Larissa Rizzo
	Valerio Belgrano
	Tobias Hofving
	Mohammad Arabpour
Recommended me for this position:	Bengt Sennblad

This work would not have been possible without the contributions of those directly or indirectly affected by this disease, or the resources shared by The Cancer Genome Atlas. Neither would it have been possible without the contributions of the Institute of Biomedicine, Department of Medical Biochemistry and Cell Biology at the University of Gothenburg, or my family.

References

1 Lodish, H. F. Molecular cell biology, 6 Ed. (2008).

2 O'Sullivan, R. J. & Karlseder, J. Telomeres: protecting chromosomes against genome instability. *Nat Rev Mol Cell Biol* **11**, 171-181, doi:10.1038/nrm2848 (2010).

3 Spitz, F. & Furlong, E. E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-626, doi:10.1038/nrg3207 (2012).

4 Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **17**, 487-500, doi:10.1038/nrg.2016.59 (2016).

5 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

6 Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-throughput sequencing technologies. *Mol Cell* **58**, 586-597, doi:10.1016/j.molcel.2015.05.004 (2015).

7 The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

8 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

9 Gao, Q. *et al.* Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep* **23**, 227-238 e223, doi:10.1016/j.celrep.2018.03.050 (2018).

10 Krump, N. A. & You, J. Molecular mechanisms of viral oncogenesis in humans. *Nat Rev Microbiol* **16**, 684-698, doi:10.1038/s41579-018-0064-6 (2018).

11 Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331**, 1553-1558, doi:10.1126/science.1204040 (2011).

12 The Cancer Genome Atlas Research Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).

13 Morris, V. & Kopetz, S. BRAF inhibitors in clinical oncology. *F1000Prime Rep* **5**, 11, doi:10.12703/P5-11 (2013).

14 Feinberg, A. P., Koldobskiy, M. A. & Gondor, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet* **17**, 284-299, doi:10.1038/nrg.2016.13 (2016).

15 Hoadley, K. A. *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291-304 e296, doi:10.1016/j.cell.2018.03.022 (2018).

16 Robertson, A. G. *et al.* Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer Cell* **32**, 204-220 e215, doi:10.1016/j.ccell.2017.07.003 (2017).

17 Liu, J. *et al.* An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400-416 e411, doi:10.1016/j.cell.2018.02.052 (2018).

18 Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1-8, doi:10.1016/j.ygeno.2015.11.003 (2016).

19 Hrdlickova, R., Toloue, M. & Tian, B. RNA-seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA* **8**, doi:10.1002/wrna.1364 (2017).

20 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

21 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).
22 Field, M. G. *et al.* Punctuated evolution of canonical genomic aberrations in uveal melanoma. *Nat Commun* **9**, 116, doi:10.1038/s41467-017-02428-w (2018).
23 Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-823, doi:10.1073/pnas.68.4.820 (1971).
24 Sherr, C. J. Principles of tumor suppression. *Cell* **116**, 235-246, doi:10.1016/S0092-8674(03)01075-4 (2004).
25 Jayasinghe, R. G. *et al.* Systematic analysis of splice-site-creating mutations in cancer. *Cell Rep* **23**, 270-281 e273, doi:10.1016/j.celrep.2018.03.052 (2018).
26 Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* **47**, 1242-1248, doi:10.1038/ng.3414 (2015).
27 Zhang, J. & Manley, J. L. Misregulation of pre-mRNA alternative splicing in cancer. *Cancer Discov* **3**, 1228-1237, doi:10.1158/2159-8290.CD-13-0253 (2013).
28 Cogdill, A. P., Andrews, M. C. & Wargo, J. A. Hallmarks of response to immune checkpoint blockade. *Br J Cancer* **117**, 1-7, doi:10.1038/bjc.2017.136 (2017).
29 Poulikakos, P. I. *et al.* RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature* **480**, 387-390, doi:10.1038/nature10662 (2011).
30 Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443-451, doi:10.1038/nrg2986 (2011).
31 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311, doi:10.1093/nar/29.1.308 (2001).
32 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).
33 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).
34 Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811-1817, doi:10.1093/bioinformatics/bts271 (2012).
35 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).
36 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).
37 Leiserson, M. D., Wu, H. T., Vandin, F. & Raphael, B. J. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol* **16**, 160, doi:10.1186/s13059-015-0700-7 (2015).
38 Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* **26**, 64-70, doi:10.1093/annonc/mdu479 (2015).
39 LaFramboise, T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* **37**, 4181-4193, doi:10.1093/nar/gkp552 (2009).
40 Lin, C. F., Naj, A. C. & Wang, L. S. Analyzing copy number variation using SNP array data: protocols for calling CNV and association tests. *Curr Protoc Hum Genet* **79**, Unit 1 27, doi:10.1002/0471142905.hg0127s79 (2013).

41 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).
42 Mitri, Z., Constantine, T. & O'Regan, R. The HER2 receptor in breast cancer: pathophysiology, clinical use, and new advances in therapy. *Chemother Res Pract* **2012**, 743193, doi:10.1155/2012/743193 (2012).
43 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).
44 Graham, N. A. *et al.* Recurrent patterns of DNA copy number alterations in tumors reflect metabolic selection pressures. *Mol Syst Biol* **13**, 914, doi:10.15252/msb.20167159 (2017).
45 Sansregret, L. & Swanton, C. The role of aneuploidy in cancer evolution. *Cold Spring Harb Perspect Med* **7**, 1-18, doi:10.1101/cshperspect.a028373 (2017).
46 Seton-Rogers, S. Tumorigenesis: Fitness penalties of aneuploidy. *Nat Rev Cancer* **17**, 142-143, doi:10.1038/nrc.2017.9 (2017).
47 Sheltzer, J. M. *et al.* Single-chromosome gains commonly function as tumor suppressors. *Cancer Cell* **31**, 240-255, doi:10.1016/j.ccell.2016.12.004 (2017).
48 Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene fusions in cancer. *Nat Rev Cancer* **15**, 371-381, doi:10.1038/nrc3947 (2015).
49 Stransky, N., Cerami, E., Schalm, S., Kim, J. L. & Lengauer, C. The landscape of kinase fusions in cancer. *Nat Commun* **5**, 4846, doi:10.1038/ncomms5846 (2014).
50 Mani, R. S. & Chinnaiyan, A. M. Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. *Nat Rev Genet* **11**, 819-829, doi:10.1038/nrg2883 (2010).
51 Nault, J. C. *et al.* Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat Genet* **47**, 1187-1193, doi:10.1038/ng.3389 (2015).
52 Kastenhuber, E. R. & Lowe, S. W. Putting p53 in context. *Cell* **170**, 1062-1078, doi:10.1016/j.cell.2017.08.028 (2017).
53 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).
54 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993, doi:10.1016/j.cell.2012.04.024 (2012).
55 Dudley, J. P., Mertz, J. A., Rajan, L., Lozano, M. & Broussard, D. R. What retroviruses teach us about the involvement of c-Myc in leukemias and lymphomas. *Leukemia* **16**, 1086-1098, doi:10.1038/sj.leu.2402451 (2002).
56 Khoury, J. D. *et al.* Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. *J Virol* **87**, 8916-8926, doi:10.1128/JVI.00340-13 (2013).
57 Tang, K. W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat Commun* **4**, 2513, doi:10.1038/ncomms3513 (2013).
58 Weinberg, R. The biology of cancer, 6 Ed. (2013).
59 Chen, Y. *et al.* VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* **29**, 266-267, doi:10.1093/bioinformatics/bts665 (2013).
60 Tang, K. W. & Larsson, E. Tumour virology in the era of high-throughput genomics. *Philos Trans R Soc Lond B Biol Sci* **372**, 1-7, doi:10.1098/rstb.2016.0265 (2017).

61 Wu, S., Zhu, W., Thompson, P. & Hannun, Y. A. Evaluating intrinsic and non-intrinsic cancer risk factors. *Nat Commun* **9**, 3490, doi:10.1038/s41467-018-05467-z (2018).
62 Brash, D. E. *et al.* A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Genetics* **88**, 10124-10128 (1991).
63 Pfeifer, G. P., You, Y. H. & Besaratinia, A. Mutations induced by ultraviolet light. *Mutat Res* **571**, 19-31, doi:10.1016/j.mrfmmm.2004.06.057 (2005).
64 Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat Rev Mol Cell Biol* **15**, 465-481, doi:10.1038/nrm3822 (2014).
65 Lewis, C. A., Jr., Crayle, J., Zhou, S., Swanstrom, R. & Wolfenden, R. Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proc Natl Acad Sci U S A* **113**, 8194-8199, doi:10.1073/pnas.1607580113 (2016).
66 Ellegren, H., Smith, N. G. & Webster, M. T. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* **13**, 562-568, doi:10.1016/j.gde.2003.10.008 (2003).
67 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).
68 Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**, 33, doi:10.1186/s13073-018-0539-0 (2018).
69 Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**, 1365-1379, doi:10.1016/j.celrep.2014.07.045 (2014).
70 Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4**, e08890, doi:10.7554/eLife.08890 (2015).
71 Cieslik, M. & Chinnaiyan, A. M. Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet* **19**, 93-109, doi:10.1038/nrg.2017.96 (2018).
72 Onken, M. D. *et al.* Collaborative Ocular Oncology Group report number 1: prospective validation of a multi-gene prognostic assay in uveal melanoma. *Ophthalmology* **119**, 1596-1603, doi:10.1016/j.ophtha.2012.02.017 (2012).
73 Worley, L. A. *et al.* Transcriptomic versus chromosomal prognostic markers and clinical outcome in uveal melanoma. *Clin Cancer Res* **13**, 1466-1471, doi:10.1158/1078-0432.CCR-06-2401 (2007).
74 The Cancer Genome Atlas Research Network *et al.* Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N Engl J Med* **372**, 2481-2498, doi:10.1056/NEJMoa1402121 (2015).
75 The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).
76 Kristensen, V. N. *et al.* Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* **14**, 299-313, doi:10.1038/nrc3721 (2014).
77 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).
78 Wang, R. F. & Wang, H. Y. Immune targets and neoantigens for cancer immunotherapy and precision medicine. *Cell Res* **27**, 11-37, doi:10.1038/cr.2016.155 (2017).

79 Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18**, 220, doi:10.1186/s13059-017-1349-1 (2017).
80 Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457, doi:10.1038/nmeth.3337 (2015).
81 Auslander, N. *et al.* Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat Med* **24**, 1545-1549, doi:10.1038/s41591-018-0157-9 (2018).
82 Jiang, P. *et al.* Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* **24**, 1550-1558, doi:10.1038/s41591-018-0136-1 (2018).
83 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049, doi:10.1038/ncomms14049 (2017).
84 Schietinger, A. & Greenberg, P. D. Tolerance and exhaustion: defining mechanisms of T cell dysfunction. *Trends Immunol* **35**, 51-60, doi:10.1016/j.it.2013.10.001 (2014).
85 Hicklin, D. J., Marincola, F. M. & Ferrone, S. HLA class I antigen downregulation in human cancers: T-cell immunotherapy revives an old story. *Mol Med Today* **5**, 178-186, doi:10.1016/S1357-4310(99)01451-3 (1999).
86 Liu, X. S. & Mardis, E. R. Applications of immunogenomics to cancer. *Cell* **168**, 600-612, doi:10.1016/j.cell.2017.01.014 (2017).
87 Neefjes, J. & Ovaa, H. A peptide's perspective on antigen presentation to the immune system. *Nat Chem Biol* **9**, 769-775, doi:10.1038/nchembio.1391 (2013).
88 Thibodeau, J., Bourgeois-Daigneault, M. C. & Lapointe, R. Targeting the MHC class II antigen presentation pathway in cancer immunotherapy. *Oncoimmunology* **1**, 908-916, doi:10.4161/onci.21205 (2012).
89 La Gruta, N. L., Gras, S., Daley, S. R., Thomas, P. G. & Rossjohn, J. Understanding the drivers of MHC restriction of T cell receptors. *Nat Rev Immunol* **18**, 467-478, doi:10.1038/s41577-018-0007-5 (2018).
90 Alcover, A., Alarcon, B. & Di Bartolo, V. Cell biology of T cell receptor expression and regulation. *Annu Rev Immunol* **36**, 103-125, doi:10.1146/annurev-immunol-042617-053429 (2018).
91 Cabaniols, J. P., Fazilleau, N., Casrouge, A., Kourilsky, P. & Kanellopoulos, J. M. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* **194**, 1385-1390, doi:10.1084/jem.194.9.1385 (2001).
92 Azizi, E. *et al.* Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293-1308 e1236, doi:10.1016/j.cell.2018.05.060 (2018).
93 Hackl, H., Charoentong, P., Finotello, F. & Trajanoski, Z. Computational genomics tools for dissecting tumour-immune cell interactions. *Nat Rev Genet* **17**, 441-458, doi:10.1038/nrg.2016.67 (2016).
94 Jurtz, V. *et al.* NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* **199**, 3360-3368, doi:10.4049/jimmunol.1700893 (2017).
95 Vitiello, A. & Zanetti, M. Neoantigen prediction and the need for validation. *Nat Biotechnol* **35**, 815-817, doi:10.1038/nbt.3932 (2017).
96 Pitcovski, J., Shahar, E., Aizenshtein, E. & Gorodetsky, R. Melanoma antigens and related immunological markers. *Crit Rev Oncol Hematol* **115**, 36-49, doi:10.1016/j.critrevonc.2017.05.001 (2017).

97 Wherry, E. J. & Kurachi, M. Molecular and cellular insights into T cell exhaustion. *Nat Rev Immunol* **15**, 486-499, doi:10.1038/nri3862 (2015).
98 Joyce, J. A. & Fearon, D. T. T cell exclusion, immune privilege, and the tumor microenvironment. *Science* **348**, 74-80, doi:10.1126/science.aaa6204 (2015).
99 Chiossone, L., Dumas, P. Y., Vienne, M. & Vivier, E. Natural killer cells and other innate lymphoid cells in cancer. *Nat Rev Immunol* **18**, 671-688, doi:10.1038/s41577-018-0061-z (2018).
100 Vicari, A. P. & Caux, C. Chemokines in cancer. *Cytokine Growth Factor Rev* **13**, 143-154, doi:10.1016/S1359-6101(01)00033-8 (2002).
101 Yang, J., Manson, D. K., Marr, B. P. & Carvajal, R. D. Treatment of uveal melanoma: where are we now? *Ther Adv Med Oncol* **10**, 1758834018757175, doi:10.1177/1758834018757175 (2018).
102 Gill, S., Maus, M. V. & Porter, D. L. Chimeric antigen receptor T cell therapy: 25 years in the making. *Blood Rev* **30**, 157-167, doi:10.1016/j.blre.2015.10.003 (2016).
103 Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nat Rev Clin Oncol* **15**, 353-365, doi:10.1038/s41571-018-0002-6 (2018).
104 Golfinopoulos, V. *et al.* Comparative survival with diverse chemotherapy regimens for cancer of unknown primary site: multiple-treatments meta-analysis. *Cancer Treat Rev* **35**, 570-573, doi:10.1016/j.ctrv.2009.05.005 (2009).
105 Kopetz, S. *et al.* Phase II pilot study of vemurafenib in patients with metastatic BRAF-mutated colorectal cancer. *J Clin Oncol* **33**, 4032-4038, doi:10.1200/JCO.2015.63.2497 (2015).
106 Fishbein, L. *et al.* Comprehensive molecular characterization of pheochromocytoma and paraganglioma. *Cancer Cell*, 181-193, doi:10.1016/j.ccell.2017.01.001 (2017).
107 Adameyko, I. *et al.* Schwann cell precursors from nerve innervation are a cellular origin of melanocytes in skin. *Cell* **139**, 366-379, doi:10.1016/j.cell.2009.07.049 (2009).
108 Greytak, S. R., Engel, K. B., Bass, B. P. & Moore, H. M. Accuracy of molecular data generated with FFPE biospecimens: lessons from the literature. *Cancer Res* **75**, 1541-1547, doi:10.1158/0008-5472.CAN-14-2378 (2015).
109 Moran, S. *et al.* Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* **17**, 1386-1395, doi:10.1016/S1470-2045(16)30297-2 (2016).
110 Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773, doi:10.1126/science.aag2445 (2016).
111 Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annu Rev Genet* **46**, 1-19, doi:10.1146/annurev-genet-110711-155459 (2012).
112 Herranz, D. *et al.* A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat Med* **20**, 1130-1137, doi:10.1038/nm.3665 (2014).
113 Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48**, 176-182, doi:10.1038/ng.3470 (2016).
114 Zimmerman, M. W. *et al.* MYC drives a subset of high-risk pediatric neuroblastomas and is activated through mechanisms including enhancer hijacking and focal enhancer amplification. *Cancer Discov* **8**, 320-335, doi:10.1158/2159-8290.CD-17-0993 (2018).

115 Carvajal, R. D. *et al.* Metastatic disease from uveal melanoma: treatment options and future prospects. *Br J Ophthalmol* **101**, 38-44, doi:10.1136/bjophthalmol-2016-309034 (2017).
116 Amirouchene-Angelozzi, N. *et al.* Upcoming translational challenges for uveal melanoma. *Br J Cancer* **113**, 1249-1253, doi:10.1038/bjc.2015.269 (2015).
117 Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* **28**, 4364-4375, doi:10.1093/nar/28.21.4364 (2000).
118 Masoomian, B., Shields, J. A. & Shields, C. L. Overview of BAP1 cancer predisposition syndrome and the relationship to uveal melanoma. *J Curr Ophthalmol* **30**, 102-109, doi:10.1016/j.joco.2018.02.005 (2018).
119 Bononi, A. *et al.* BAP1 regulates IP3R3-mediated Ca(2+) flux to mitochondria suppressing cell transformation. *Nature* **546**, 549-553, doi:10.1038/nature22798 (2017).
120 Eletr, Z. M. & Wilkinson, K. D. An emerging model for BAP1's role in regulating cell cycle progression. *Cell Biochem Biophys* **60**, 3-11, doi:10.1007/s12013-011-9184-6 (2011).
121 Harbour, J. W. *et al.* Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science* **330**, 1410-1413, doi:10.1126/science.1194472 (2010).
122 Matatall, K. A. *et al.* BAP1 deficiency causes loss of melanocytic cell identity in uveal melanoma. *BMC Cancer* **13**, 371, doi:10.1186/1471-2407-13-371 (2013).
123 Nishikawa, H. *et al.* BRCA1-associated protein 1 interferes with BRCA1/BARD1 RING heterodimer activity. *Cancer Res* **69**, 111-119, doi:10.1158/0008-5472.CAN-08-3355 (2009).
124 Ruan, H. B. *et al.* O-GlcNAc transferase/host cell factor C1 complex regulates gluconeogenesis by modulating PGC-1alpha stability. *Cell Metab* **16**, 226-237, doi:10.1016/j.cmet.2012.07.006 (2012).
125 Ventii, K. H. *et al.* BRCA1-associated protein-1 is a tumor suppressor that requires deubiquitinating activity and nuclear localization. *Cancer Res* **68**, 6953-6962, doi:10.1158/0008-5472.CAN-08-0365 (2008).
126 Yu, H. *et al.* The ubiquitin carboxyl hydrolase BAP1 forms a ternary complex with YY1 and HCF-1 and is a critical regulator of gene expression. *Mol Cell Biol* **30**, 5071-5085, doi:10.1128/MCB.00396-10 (2010).
127 Zhang, Y. *et al.* BAP1 links metabolic regulation of ferroptosis to tumour suppression. *Nat Cell Biol* **20**, 1181-1192, doi:10.1038/s41556-018-0178-0 (2018).
128 Scheuermann, J. C. *et al.* Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. *Nature* **465**, 243-247, doi:10.1038/nature08966 (2010).
129 Onken, M. D. *et al.* Functional gene expression analysis uncovers phenotypic switch in aggressive uveal melanomas. *Cancer Res* **66**, 4602-4609, doi:10.1158/0008-5472.CAN-05-4196 (2006).
130 Krantz, B. A., Dave, N., Komatsubara, K. M., Marr, B. P. & Carvajal, R. D. Uveal melanoma: epidemiology, etiology, and treatment of primary disease. *Clinical Ophthalmology* **Volume 11**, 279-289, doi:10.2147/opth.S89591 (2017).
131 Johansson, P. *et al.* Deep sequencing of uveal melanoma identifies a recurrent mutation in PLCB4. *Oncotarget* **7**, 4624-4631, doi:10.18632/oncotarget.6614 (2016).

132 Royer-Bertrand, B. *et al.* Comprehensive genetic landscape of uveal melanoma by whole-genome sequencing. *Am J Hum Genet* **99**, 1190-1198, doi:10.1016/j.ajhg.2016.09.008 (2016).
133 Damato, B., Dopierala, J. A. & Coupland, S. E. Genotypic profiling of 452 choroidal melanomas with multiplex ligation-dependent probe amplification. *Clin Cancer Res* **16**, 6083-6092, doi:10.1158/1078-0432.CCR-10-2076 (2010).
134 Prescher, G. *et al.* Prognostic implications of monosomy 3 in uveal melanoma. *Lancet* **347**, 1222-1225, doi:10.1016/S0140-6736(96)90736-9 (1996).
135 Tschentscher, F. *et al.* Tumor classification based on gene expression profiling shows that uveal melanomas with and without monosomy 3 represent two distinct entities. *Cancer Res* **63**, 2578-2584 (2003).
136 Moore, A. R. *et al.* Recurrent activating mutations of G-protein-coupled receptor CYSLTR2 in uveal melanoma. *Nat Genet* **48**, 675-680, doi:10.1038/ng.3549 (2016).
137 Keshava Prasad, T. S. *et al.* Human Protein Reference Database—2009 update. *Nucleic Acids Res* **37**, D767-772, doi:10.1093/nar/gkn892 (2009).
138 Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-245, doi:10.1093/bioinformatics/btq182 (2010).
139 Li, P. *et al.* Cytochrome c and dATP-dependent formation of Apaf-1/caspase-9 complex initiates an apoptotic protease cascade. *Cell* **91**, 479-489 (1997).
140 Mehlen, P. & Puisieux, A. Metastasis: a question of life or death. *Nat Rev Cancer* **6**, 449-458, doi:10.1038/nrc1886 (2006).
141 Owens, L. V. *et al.* Overexpression of the focal adhesion kinase (p125FAK) in invasive human tumors. *Cancer Res* **55**, 2752-2755 (1995).
142 Mendonsa, A. M., Na, T. Y. & Gumbiner, B. M. E-cadherin in contact inhibition and cancer. *Oncogene* **37**, 4769-4780, doi:10.1038/s41388-018-0304-2 (2018).
143 Golubovskaya, V. M. Targeting FAK in human cancer: from finding to first clinical trials. *Front Biosci (Landmark Ed)* **19**, 687-706, doi:10.2741/4236 (2014).
144 Jiang, H. *et al.* Targeting focal adhesion kinase renders pancreatic cancers responsive to checkpoint immunotherapy. *Nat Med* **22**, 851-860, doi:10.1038/nm.4123 (2016).