Genetic screening and molecular characterisation of biomarkers in hepatocellular carcinoma

Inaugraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Viola Paradiso

Basel, 2021

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Luigi M. Terracciano

Dr. Salvatore Piscuoglio

Prof. Gerhard Christofori

PD. Dr. Marianna Kruithof-de Julio

Basel, 17.11.2020

Prof. Dr. Martin Spiess

Dekan

Was mich nicht umbringt,
macht mich stärker.
Friedrich Nietzsche

Abstract

Hepatocellular carcinoma (HCC) is the most common type of liver cancer that accounts for 4.7% of the total number of new cases of cancer worldwide every year. HCC is a highly heterogeneous and complex disease with an estimated 5-year survival rate of only 18%. A better understanding of the mechanisms involved in the development, progression and recurrence of this tumour could not only guide us in the improvement of preventive strategies but also in the expansion of alternative target therapies for HCC patients.

The aim of this thesis is to investigate new diagnostic and prognostic markers, both on genetic and molecular levels, in the context of HCC. The results section is divided in two, called **Chapter I** and **Chapter II**.

HCC presents a distinct mutational landscape and **Chapter I** describes how we developed a HCC-specific custom made sequencing panel, containing the genes most commonly affected by somatic mutations and copy number alterations (CNAs) in the disease. We created a panel that was tested in different kinds of patient biopsies: frozen tissues, formalin-fixed paraffin-embedded (FFPE) tissues and also liquid biopsies. Moreover, to have reliable and reproducible sequencing data, we created a solid and user friendly somatic variant calling pipeline specific for Ion Torrent sequencing data.

In **Chapter II**, we aimed to investigate the molecular mechanism of HMGA1 in HCC and to explore its molecular targets. HMGA1 is an architectural transcription factor that was found often overexpressed in HCCs. We explored its DNA-binding landscape and, after deregulating HMGA1 in a HCC *in vitro* environment, its expression signature both at the RNA and protein levels. With the analysis of the binding partners of HMGA1, we recognised the vast range of mechanisms of action of this complex protein. We identified several RNA regulators that bind HMGA1, including Alyref, which plays a role in the regulation of the transcription. Further work should aim to determine the non-canonical role of HMGA1 involved in the binding and the regulation not only at the DNA but also at the RNA level.

Both chapters describe the steps of this work on the identification and the functional understanding of HCC biomarkers. This may lead in the future to more individualised treatment

approaches, a need that in cancers with low survival rate such as HCC is not only highly desirable but is also a necessity.

List of abbreviations

AFB1: Aflatoxin 1
APC: Adenomatous polyposis coli
ARID1A: AT-rich interaction domain 1A
ARID1B: AT-rich interaction domain 1B
ARID2: AT-rich interaction domain 2
BAP1: BRCA1 associated protein 1
CDH26: Cadherin 26
CDKN2A: Cyclin dependent kinase inhibitor 2A
CDKN2B: Cyclin dependent kinase inhibitor 2B
cfDNA: Cell-free DNA
ChIP-seq: Chromatin immunoprecipitation sequencing
CLDN3: Claudin 3
CNA: Copy number alteration
ctDNA: Circulating tumour DNA
CTNNB1: Catenin beta 1
EGFR: Epidermal growth factor receptor
EMT: Epithelial to mesenchymal transition
EZH2: Enhancer of zeste 2
FFPE: Formalin-fixed paraffin-embedded
HBV: Hepatitis B virus
HCC: Hepatocellular carcinoma
HCV: Hepatitis C virus
HMGA1: High mobility group A 1
HNF1A: Hepatocyte nuclear factor 1
IRF2: Interferon regulatory factor 2
JAK: Janus kinase
KEAP1: Kelch like ECH associated protein 1
lncRNA: Long non-coding RNA
MLL3: Myeloid/lymphoid or mixed-lineage leukemia protein 3
MLL: Myeloid/lymphoid or mixed-lineage leukemia
mRNA: Messenger RNA
MS: Mass-spectrometry
NFE2L2: Nuclear factor erythroid 2 like 2

PBRM1: Polybromo 1

PIK3CA: Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha

siRNA: Small interfering RNA

SOX2: SRY-box transcription factor 2

STAT: Signal transducer and activator of transcription

TCGA: The Cancer Genome Atlas

TERT: Telomerase reverse transcriptase

WES: Whole exome sequencing

Table of contents

1- Introduction

I. The hallmarks of cancer

Nowadays cancer is still among the leading causes of death worldwide, with more than 9.5 million cancer-related deaths and ~18 million new cases worldwide in 2018 [1,2]. Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. Cancer has been demonstrated to be a multistep process [2] involving genetic and nongenetic alterations such as changes in the genotype (e.g. mutations in the DNA) or changes in the phenotype that do not involve DNA alteration, called epigenetic changes (e.g. DNA methylation, histone modifications, non-coding RNA mechanisms). These modifications affect how a gene is read by a cell and empower the gain of new capabilities.

When cells grow locally without invading adjacent tissues the tumour is classified as benign. Tumours that invade nearby tissues are called malignant; when a cell (or a group of cells) in the primary tumour gains the ability to extrude the initial tissue and disseminate into the body via the lymphatic system or through the bloodstream, the tumour is called invasive. Metastasis happens when the disseminated cells from the primary tumour seed and proliferate in a new distant site forming another tumour site [3]. Cancer cells are less specialised than normal cells, they can proliferate uncontrollably and avoid apoptosis. As tumours grow, the number of mutations will increase and the accumulation of mutations will confer survival advantages over time [4,5]. These advantages are biological capabilities (gain or loss of functions) that can describe the development of cancer and they can be combined into 10 groups, as Hanahan and Weinberg proposed [6], summarised shortly below. The hallmarks of cancer are, to date, the fundamentals for understanding the biology of cancer.

1. Sustaining proliferative signaling

One of the most critical abilities that a cell can gain is to sustain chronic proliferation by deregulating growth-promoting signals. Mutations in growth factors principally, or in any of the genes encoding for proteins involved in the subsequent intracellular signaling pathways that regulate progression, can influence not only the cell cycle and growth but also other cell-biological properties, such as cell survival and energy metabolism [7,8].

2. Evading growth suppressors

In addition to the sustaining proliferative signaling, cancer cells can also circumvent pathways that negatively regulate cell proliferation, with alterations in tumour suppressor genes. Tumour suppressor proteins function as gatekeepers of cell cycle progression, these proteins are involved

in the combination of inputs of stress and abnormality from extracellular and intracellular sources and regulate the entry of a cell through its growth and division cycle [9,10].

3. Enabling replicative immortality

This refers to the ability to avoid senescence and apoptosis. Normal cells have a limited number of cell division cycles assured by the telomeres, the tandem repeats at the ends of chromosomes. After reaching their maximum number of divisions, cells undergo senescence, a non-proliferative but viable state, and then enter into a crisis phase, which involves cell death/apoptosis [11,12]. Cancer cells develop an ability to maintain telomeric DNA at lengths sufficient to avoid triggering senescence or apoptosis. In the majority of the cases by upregulating expression of telomerase or, sometimes, using an alternative maintenance mechanism of recombination of telomeres [13].

4. Resisting cell death

This ability allows tumours to attenuate apoptosis. Apoptosis is triggered in cells in response to various physiologic stresses. Cancer cells can experience these brunts during the course of tumourigenesis or as a result of anticancer therapy [14]. The apoptosis machinery is composed by several regulators that can receive and process internal and external signals and can initiate a cascade of proteolysis involving effector components for the execution phase of apoptosis. When this fine counterbalance of pro- and antiapoptotic members is deregulated, cells may become resistant to inhibit apoptosis [14-16].

5. Avoiding immune destruction

Tumoural cells also gain the ability to avoid detection by the immune system. The immune surveillance is constantly monitoring cells in the body and is able to recognise and eliminate the majority of nascent cancer cells. The cells in a new tumour managed to avoid or limit the immunological detection and killing, thereby evading elimination [17,18].

6. Tumour promoting inflammation

Immune cells can not only recognise and eliminate cancer cells in the body, but they can also, paradoxically, enhance tumourigenesis and progression. Inflammation can contribute to the acquisition of multiple capabilities by supplying active molecules to the tumour microenvironment (growth, survival and angiogenic factors, inductive signals and extracellular matrix-modifying enzymes). Every tumour contains immune cells at different densities that can be used to promote aggressiveness and invasiveness [19,20].

7. Inducing angiogenesis

All tissues, including tumoural, need sustenance (nutrients and oxygen) and an evacuation system (for metabolic wastes and carbon dioxide). Big tumours acquire the ability to create new vasculature (angiogenesis) to address these needs by deregulation of signaling proteins that induce or inhibit angiogenic regulators. In this way, angiogenesis remains always activated and the tumour does not risk to undergo necrosis [21,22].

8. Deregulating cellular energetics

This indicates the ability of cancer cells to reprogram their glucose metabolism, and therefore their energy production, by limiting their energy metabolism mostly to glycolysis even in presence of oxygen, a state called “aerobic glycolysis”. The specific deregulation of oncogenes and tumour suppressor confers benefits for the reliance of glycolysis [23,24]. Today the rationale for this choice is still unclear, the most accredited theory states that the increase of glycolysis allows the deviation of glycolytic intermediates into various biosynthetic pathways; this supports the large-scale biosynthesis of the macromolecules programs that are required for active cell proliferation [25-27].

9. Activating invasion and metastasis

Tumour cells can gain the ability to develop alterations in shape and modification in attachment approaches to other cells and to the extracellular matrix. Deregulation or mutational inactivation and activation of key cell-to-cell and cell-to-extracellular matrix adhesion molecules is frequently observed in cancer and accelerates the capability for invasion and metastasis [28,29].

10. Genome instability and mutation

Last but not least, the ability of tumour cells to increase the rates of mutation is one of the hallmarks that confers major selective advantages on tumour clonal cells. The system in normal cells to detect and resolve defects in the DNA is extremely complicated but well-organised and interconnected. For this reason, the rate of spontaneous mutation is usually very low during a cell generation life. During tumourigenesis, cancer cells often increase the rates of mutation by disruption and break of one or more components in these genome maintenance systems (DNA damage, recognition and repair machinery factors). The accumulation of mutations sometimes force cells into senescence or apoptosis but can also trigger the acquisition of mutant genotypes that confer selective advantages and that can be maintained in the successive clonal expansions [30-33].

It is evident that the multistep process of human tumour pathogenesis needs the acquisition of new traits for normal cells to become tumourigenic and later malignant. The summary of these traits is well realised in the hallmarks of cancer proposed by Hanahan and Weinberg and an

illustrative summary is shown in Figure 1.1. The hallmarks contribute to provide a structure for the understanding of the complex biology of cancer.

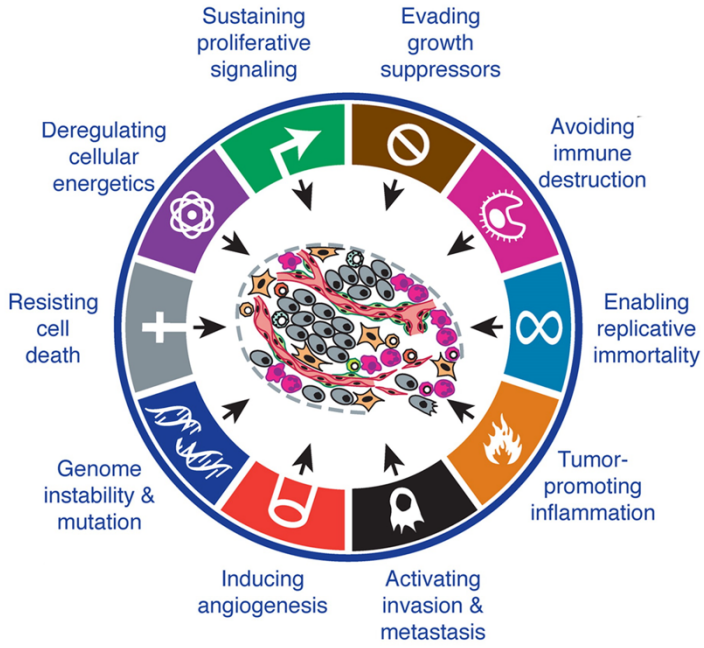


Figure 1.1: *Hallmarks of cancer. Taken and mod. from Hanahan and Weinberg 2011* [6]*. This illustration encompasses the ten hallmark capabilities of cancer proposed by Hanahan and Weinberg.*

II. Genetics of cancer

Considering the hallmarks of cancer, it is vital to understand the causes that lead to the uncontrolled growth of cells. One of the primary causes of cancer is genetic mutations [4,5]. The errors in the DNA of a cell can occur due to several factors (e.g. spontaneous errors during DNA replication process, heredity, radiation and chemicals) and can contribute to the dedifferentiation, the loss of the specificity of a cell variant.

Types of mutations

A mutation can be germline or somatic. Germline mutations are genetically inherited alterations that are present in the germ cells (sperm or eggs) and that are then contained in all cells of the individual [2,34]. Carrying germline mutations in cancer related genes increases the probability to develop cancer during a person's lifetime. On the other hand, somatic mutations are not inherited and they first appear in differentiated cells. Somatic mutations are therefore inherited by the

progeny of these particular cells during their divisions but they would not be inherited by the offspring of the person carrying the mutations.

Each person has a unique combination of genetic changes, but not all contribute to cancer progression. The accumulation of somatic genetic changes arising during a person's lifetime in a tissue can allow a tissue to acquire certain selective advantages compared to neighbouring cells, including the ability to drive tumourigenesis. The somatic alterations conferring a selective advantage on the cell, the so-called "driver" mutations, are kept during divisions, while the disadvantageous ones usually undergo negative selection [4,35-37]. It is hypothesised that the vast majority of mutated genes have no involvement in tumourigenesis; their mutations are passengers rather than drivers [38]. Thanks to the most recent next generation sequencing technologies, distinctive patterns of DNA mutations in different tumour types were revealed [39].

Genetic alterations can be classified into three main classes: (non-)synonymous mutations, structural variations and copy number variations (CNVs). The first category includes substitutions (e.g. silent, nonsense, missense, splice site) and insertions and deletions of one or more nucleotides (that result in frameshifts and in-frame mutations). The structural variation category consists of larger insertions, deletions, rearrangements or translocations of chromosomal regions. The CNVs are instead gene amplifications or losses, ending with a different number of copies of a locus. To identify the mutational pattern of a tumour we need to analyse the mutation types and the heterogeneity of mutation rate. One of the first and biggest consortium aimed at sequencing the exomes of thousands of tumours of more than thirty frequent cancer types is The Cancer Genome Atlas (TCGA) [40] and the Figure 1.2 represents an overview of their results, showing the number of mutations in coding regions divided per cancer type. Other pan-cancer projects followed and they led to some important findings in the discovery of driver mutations in driver genes in primary malignancies [41,42]. Recently, similar projects aim to reveal not only the landscape of driver alterations of advanced malignancies, but also to identify driver mutations in non-coding regions and regulatory sequences [41,43,44], as shown in the example in Figure 1.3. The importance of epigenetics as another main cause of tumourigenesis has been well acknowledged and a deeper elucidation of these epigenetic mechanisms might materially change our overall understanding of the means by which hallmark capabilities are acquired.

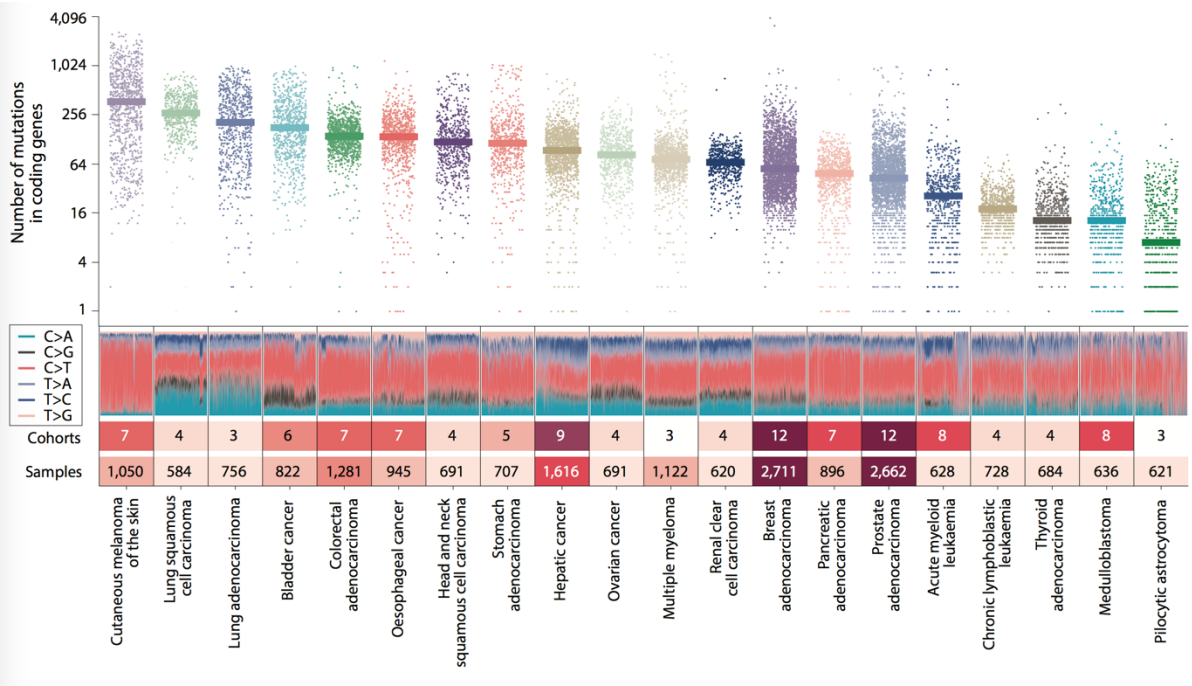


***Figure 1.2**: Number of mutations in coding genes per cancer type. Taken and mod. from Martínez-Jiménez et al. 2020 [45]. Mutation burden (top) and mutation type (bottom) of tumours from cancer types represented by at least two cohorts. Cohorts are coming from datasets of tumour mutations collected from the public domain and analysed with IntOGen pipeline. The number of cohorts and samples contributing to the distribution of each cancer type are shown below the plot. Adeno., adenocarcinoma; CLL, chronic lymphocytic leukaemia; Hartwig, Hartwig Medical Foundation; ICGC, International Cancer Genome Consortium; PCAWG, Pan-Cancer Analysis of Whole Genomes; St Jude, St Jude Children's Research Hospital; TARGET, Therapeutically Applicable Research to Generate Effective Treatments; TCGA, The Cancer Genome Atlas.*

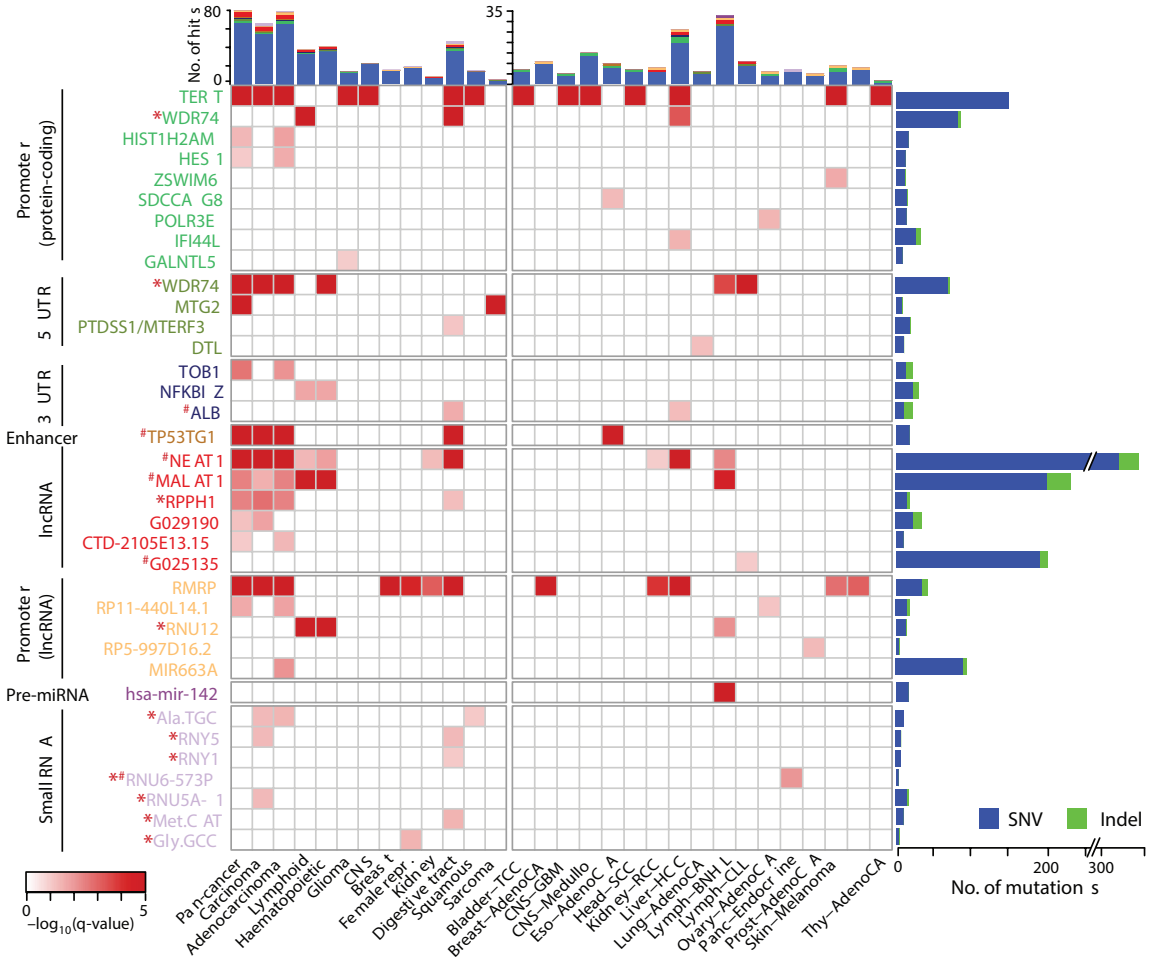


Figure 1.3: *Non-coding point mutations in Pan-Cancer Analysis of Whole Genomes (PCAWG). Taken and mod. from Rheinbay et al. 2020 [41]. Significant non-coding elements (Q < 0.1 of Brown's combined P values of up to 13 driver discovery methods) identified in cohorts with at least one hit. Colour represents significance levels. *Potential technical artefact; #targets affected by mutational processes. AdenoCA, adenocarcinoma; CNS, central nervous system; Eso, oesophageal; GBM, glioblastoma; HCC, hepatocellular carcinoma; Medullo, medulloblastoma; Panc, pancreatic; Prost, prostate; RCC, renal cell carcinoma; Repr., reproductive organs; SCC, squamous cell carcinoma; TCC, transitional cell carcinoma; Thy, thyroid. HIST1H2AM is also known as H2AC17; Ala.TGC as TRA-TGC3-1; Met.CAT as TRM-CAT1-1; and Gly.GCC as TRG-GCC2-3. PTDSS1/MTERF3 denotes that 5' UTR mutations in PTDSS1 also overlap the MTERF3 promoter.*

Cancer is a dynamic disease characterised by clonal evolution of cells, therefore usually it presents diverse characteristics inside the same site that are representative of its heterogeneity. One tumour can include a diverse collection of cells (subpopulations) harbouring distinct molecular signatures. This non-uniform distribution can be across and within disease sites (spatial heterogeneity) or can be due to temporal variations (temporal heterogeneity) [46]. Recognizing and comprehending the genetic, the epigenetic and the tumoural heterogeneity may provide important

insights not only into tumourigenesis but also into the drug resistance process that often results in differential levels of sensitivity to treatment for patients.

Oncogenes and tumour suppressors

Despite the genetic differences between cancer, two main classes of genes generally affected by somatic genetic alterations in cancers are recognised: proto-oncogenes and tumour suppressor genes. They typically encode for proteins involved in the major control of the pathways in the cells, and therefore play a key role in cancer development. A proto-oncogene generally encodes for a protein involved in the viability of the cell and enhances cell proliferation [47,48]. When mutated, damaged or amplified it becomes oncogenic and can enable tumour cells to circumvent the checks and balances that are in place during homeostasis to drive tumour growth. Tumour suppressor genes, on the other hand, encode for proteins implicated in the inhibition of uncontrolled cell proliferation and in the driving of cell death [49,50]. For this reason they are also often mutated in cancer; their proteins undergo loss of function and inactivation. Their inactivation might constitute driver events that are thought to occur in the earliest stages of carcinogenesis [49,50]. Once a gene is mutated, its product can be affected in different ways: the protein can not be functional anymore, the protein expression can be completely blocked, or the way of function of a protein can be modified. A mutation can also cause the activation of a gene that is not usually expressed in certain tissues or inactivate the expression of a gene important in normal conditions [2]. For example, *ras*, one of the most commonly mutated genes in all cancers, becomes oncogenic thanks to point mutations resulting in single amino acid substitutions at critical positions [51]. The first such mutation discovered was the substitution of valine for glycine at position 12. When *ras* shows this mutation, Ras protein is constitutively in the active GTP-bound conformation and drives unregulated cell proliferation [52]. Another example is about *APC* tumour suppressor gene. Hotspots mutations are mainly concentrated in the exon 15, often resulting in a truncated non-functional protein [53]. Moreover, it has been found that hypermethylation of the promoter region of this gene constitutes an alternative mechanism of gene inactivation [54]. DNA methylation is indeed one of the most commonly studied epigenetic mechanisms associated with the transcriptional silencing of tumour suppressor genes and the effectiveness of inactivation is the same with mutations [55].

To characterise and to understand the mutational and epigenetic changes in tumour samples is therefore not only changing research and clinical practice but has also driven the implementation of molecular testing into clinical practice.

III. Importance of next generation sequencing for the clinic

In the last two decades, sequencing technologies have allowed the discovery of distinctive patterns of mutations in different tumour types. Since then, many researchers have contributed to the exploration of the functions and mechanisms of the mutated gene products, their pathways and their implications in tumour processes. These discoveries led to the development of drugs that can target the mutated gene products (proteins or enzymes) and stop the informational cascade underneath [56,57]. Therefore, targeted drugs can turn off signals that make cancer cells grow, or can enhance internal tumoural signals leading to apoptosis. More deeply we understand about the mechanism resulting from these genetic changes, more precisely we might identify targets to develop strategic therapies [56,57]. Despite the enormous effort in this kind of research worldwide, so far only a few types of cancers are routinely treated using targeted drugs, and often they are in combination with other common therapies. Another reason might be the limited number of approved sequencing tests for diagnostic research [58,59].

The traditional genetic tests used in clinical practices have been replaced in the last decade by next generation sequencing. Not only this kind of technology is cost and time effective, but it allows the generation of bigger and still accurate and reliable information from the genome. A clinical next-generation sequencing test can target a panel of selected genes, the exome or, more rarely, the entire genome. Next-generation sequencing has revolutionised not only the oncology field, helping the identification of genetic variants in human cancers, but also the research for many other diseases, especially hereditary disorders in the pediatric area [60]. The main part of clinical tests are done on DNA samples from biopsy (for somatic mutations). The use of DNA from blood is instead mainly adopted in case of tumours of the haematopoietic and lymphoid malignancies and to reveal germline mutations.

Biopsy

Biopsies are samples of tissue taken from the body to get more information about possible anomalies. The information achievable from biopsies includes the presence, the cause or the extent of the disease. They are frequently used by pathologists to recognise if in the tissue there is a lesion, a mass or a tumour and they represent an invaluable source of biological material. There are two main approaches to store biopsies for extended duration by preserving the morphology and cellular details of the tissues: by formalin-fixation and paraffin-embedding (FFPE) and by snap freezing (Figure 1.4).

In the first case, the biopsy is fixed in formaldehyde, to preserve mostly the proteins and the structures of the tissue, important for immunohistochemistry. When a biopsy is rapidly frozen after

removal from the tissue, it is possible to preserve better all the components for molecular analysis [61]. Both methods show pros and cons and depending on the main use of the biopsies and the laboratory opportunities, one or both storage processes can be chosen. FFPE tissue samples can be stored at room temperatures and they do not need specialised equipment. This is a cost-effective storage method that makes it possible to keep a large collection of tissues accessible for a long time.

Even if the formalin and wax ensure that cell structures and proteins are well preserved, they are denatured and no longer biologically active. The fixation often leads to cross-linking, degradation and fragmentation of DNA and especially RNA molecules. These alterations inevitably affect the use of FFPE samples in molecular and genetic analysis and the results obtained are not comparable to the ones from frozen tissue samples [62].

Fresh tissue samples require a quicker but more expensive storage process. The need for specialised equipment, such as ultra-low temperature freezer, its preservation and maintenance with the problem of the rapid deterioration of the samples at room temperature, make this storage method not feasible for a big amount of samples for a long time. On the other hand, the proteins are still preserved in their native state and frozen tissue material is the gold standard especially for sequencing due to its superiority in preserving DNA and RNA [62,63].

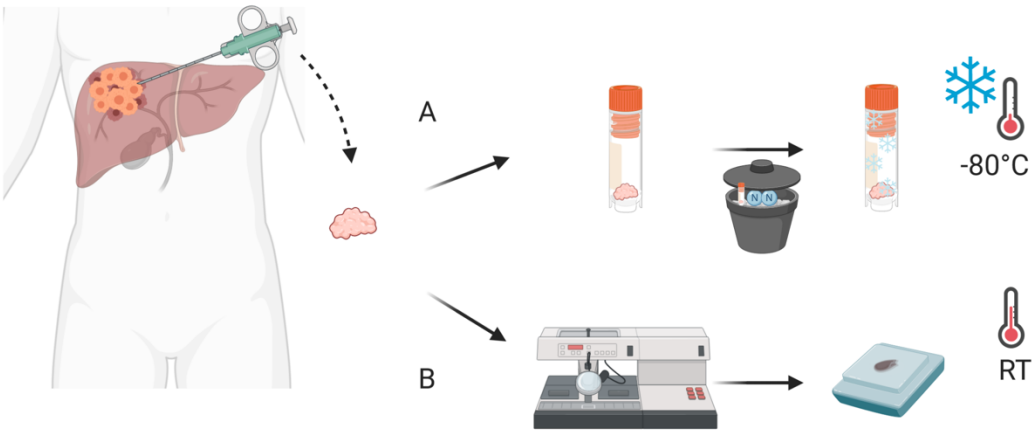


Figure 1.4: *Biopsy collection and storage.* ***A)*** *A small sample of tissue is taken from the tumour and is immediately snap frozen in liquid nitrogen and then stored at -80°C for long or short periods of time.* ***B)*** *A small sample of tissue is taken from the tumour and undergoes formalin-fixation and paraffin-embedding (FFPE) process. The FFPE tissue sample can be stored at room temperatures (RT).*

Liquid biopsy

The invasive nature of biopsy has encouraged investigations into the use of plasma liquid biopsy, a potentially minimally invasive alternative method that allows to explore plasma-derived cell-free DNA (cfDNA) for molecular profiling in several disease areas (Figure 1.5). When a healthy, inflamed or tumour cell undergoes apoptosis or necrosis, its content, including DNA, is released into the bloodstream. Circulating tumour DNA (ctDNA) is a small fraction of cfDNA found in the bloodstream and refers to DNA that comes from cancerous cells [64]. It is characterised by the presence of somatic variants representative of the tumoural genetic situation [65]. ctDNA levels change during disease progression and during chemotherapy, for this reason ctDNAs can be investigated not only as good candidate biomarkers for the screening of cancer patients but also for monitoring recurrence [66,67]. In recent years, the development of highly sensitive assays that can detect ctDNA from plasma contributed to make it an attractive investigative modality.

However, the applicability of liquid biopsy in the clinical routine, despite being a simple and non-invasive alternative for the patients to surgical biopsies, is still in an emerging state. Recently, the Food and Drug Administration approved some mutation tests for DNA from plasma of patients with cancer. For example, a gene mutation detection system working with both DNA from tissue samples or cfDNA from plasma, can help to personalise treatment of breast cancer patients with the identification of hotspots mutations in PIK3CA [68]. There is also a test for DNA from non-small cell lung cancer patients' plasma able to identify mutations in one of the most common mutated genes for this tumour, the epidermal growth factor receptor (EGFR), that can be responsible for resistance to therapy [69].

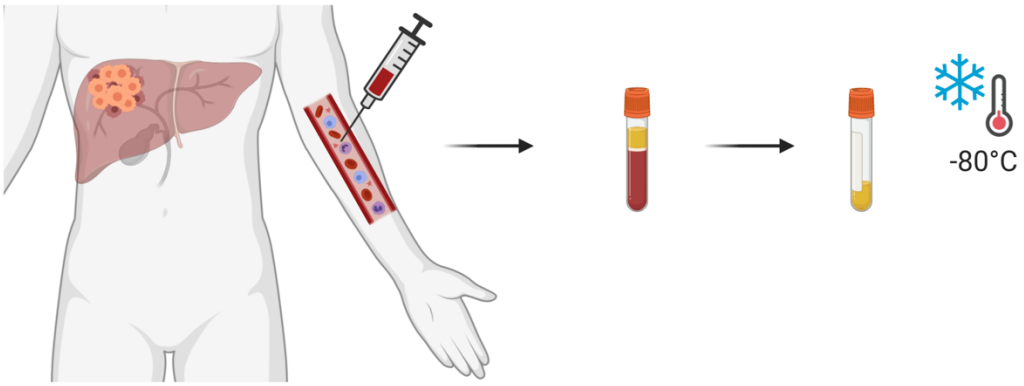


Figure 1.5: *Liquid biopsy collection and storage. Whole blood is drawn. The plasma is then separated from the red blood cells and other cellular components with centrifugation steps. Plasma can be stored at -80°C for long storage periods.*

In conclusion, nowadays it is possible to discover and analyse the complex mutational signature present in the tumoural DNA of a patient thanks to next generation sequencing techniques with minimal efforts and costs. These technologies have the great potential to uncover the clonal heterogeneity of tumours and to identify druggable targets, significant steps for targeted therapy advancement. Genetic panels developed by clinical molecular laboratories can assess multiple potential genetic causes of a tumour and meanwhile reduce the cost and time of diagnostic testing. In the clinic, up to the present time, approved and recognised sequencing panels that are tumour specific are being used; however, not all tumour patients have the possibility to be screened with a distinct approved genomic test. Among those patients, there are also liver cancer cases.

IV. Hepatocellular carcinoma

When a tumour cell starts to multiply and affect the liver in the first place, and it is not a metastasis developed from another part of the body, we talk about primary liver cancer. Primary liver cancers account for 4,7% of the total number of new cases of cancer worldwide every year, with >800000 cases in 2018. Liver cancer has its highest burden in Asian countries, where 72% of the new liver cancer cases worldwide per year are diagnosed [70,71]. Despite in Europe it constitutes only the 10%, the death rate is surmounting almost any other cancers, including breast, stomach and prostate cancers [1].

Hepatocellular carcinoma (HCC) is the most common type of liver cancer, it consists of approximately 85% of all primary liver cancers. The remaining 15% includes intrahepatic cholangiocarcinoma and extrahepatic bile-duct carcinoma [72,73]. HCC has been estimated to be the fourth most common cause of cancer-related death overall worldwide [74] Unfortunately, the large global disparity in the incidence and mortality from HCC is due to the existing differences in assessing the disease in early or late stages, healthcare resource availability and level of exposure to risk factors [74,75]. Less than half of the patients are eligible for curative treatments, that are represented by surgical resection or radiofrequency ablation of the tumour and, in the worst cases, liver transplantation. However these kinds of treatments are only possible if the patient is diagnosed at an early stage, otherwise receiving a palliative cure is the only opportunity. This is one of the reasons why the estimated 5-year survival rate for HCC is only 18% overall worldwide [76].

The majority of the HCC cases occur in patients with underlying liver diseases, mostly as a result of hepatitis B or C virus (HBV or HCV) infections, alcohol abuse or aflatoxin B1 (AFB1) exposure [77]. Even though HBV and HCV are the major causes of HCC, around 30% of patients do not show neither virus infection nor alcohol abuse, suggesting that other risk factors can have a big impact

on HCC development [78]. In the last few decades the importance of other etiologic factors such as metabolic syndromes, obesity and diabetes has been investigated and it has been shown a clear association with HCC [79].

Understanding the hidden process that goes between metabolic diseases, such as non-alcoholic fatty liver disease, autoimmune hepatitis, hereditary hemochromatosis and the development of liver tumour, but also the mechanism underlying cigarette smoking or alcohol abuse, can help us to increase the possibilities of cure for HCC patients.

Prevention

Nowadays the major part of prevention studies are focused on the main causes of HCC, the HBV and HCV infections.

The first line of prevention is a vaccination against HBV. This virus is considered to be the most critical environmental carcinogen to which humans are exposed. This is why the World Health Organization recommends the vaccination especially to the newborns and to high risk adult subjects in all countries [80]. Several studies have found evidence of efficacy of HBV vaccine and reduced incidence of HCC [81-83].

On the other hand, the infection with HCV is predominantly acquired in adulthood, mainly due to intravenous transfusion with contaminated products. Also in this case, antiviral treatments in patients with HCV-related cirrhosis result in lower risk of HCC development [84,85]. Interferon therapy is known to be used to reduce the risk of HCC in patients carrying HCV infection. However, it is not very efficient in patients with severe fibrosis or cirrhosis [86]. Alternatively, the development of direct-acting antiviral agents led to a high improvement both in the response rates and the tolerability of treatment of HCV infected patients [87]. They are protease or polymerase inhibitors that interfere with specific steps of the HCV replication process but the rapid evolution of the virus led to a wide variety of innate defence mechanisms, multidrug resistant-virus and in the end to the recurrence of HCC. For this reason, the debate about risk and benefit of the direct-acting antiviral agents is still challenging [88].

In first world countries, the only other evident line of prevention is a healthy lifestyle with low alcohol consumption. Patients having alcohol disorders can reduce the incidence of alcohol-associated cirrhosis by adopting abstinence behaviour.

In low-income countries, on the other hand, a great challenge would be to minimise both sources of contamination of *fungus Aspergillus*, the fungus that produced AFB1, and AFB1 contaminated food manufacturing. To change agricultural practices in regions of high dietary AFB1 intake, to improve storage methodologies and conditions and to screen food to search for contaminations would be long term and effective lines of prevention. However, due to some countries' resources

and/or education, those are difficult and ambitious solutions that cannot be easily accomplished and that certainly are important research areas.

Genomic landscape

As already mentioned, the treatment possibilities for HCC patients are not many and they often aim to extend lives more than cure the disease. This is because hepatocarcinogenesis is a complex cascade of multistep events that ends in a malignant transformation of a hepatocyte. When cirrhotic hyperplastic nodules are present in the liver, the regenerating hepatocytes can be subjected to genetic and morphological changes and form pre-malignant dysplastic lesions [89]. This kind of lesions alters the liver architecture and they are clearly recognisable because of their cytological characteristics, for example for their cellular changes in shape or for the presence of nuclear crowding [90]. Usually these lesions advance into HCC because of an accumulation of alterations in genes involved in one or more hallmarks of cancer pathways, creating unbalanced mechanisms in the cells. On average, HCCs harbour around 40 mutations in the exome, some are well known driver mutations and for others the role is uncertain [91,92]. Despite the characterisation of known modified oncogenes and suppressor genes, cell cycle regulators and immune response genes, not all molecular pathways that play a pivotal role in liver tumour development are fully identified. What we know about alterations in HCC was due to next generation sequencing techniques that were essential in the identification of key driver mutated genes. One of the main investigations up to the present is a massive sequencing study performed on more than three hundred HCC cases by exome sequencing performed by the TCGA network a couple of years ago [93], showing that up to 40% of patients present mutations in *TP53*, *PIK3CA*, and *CTNNB1* (ß-catenin). Other genes often mutated include *ARID1A, ARID1B, ARID2, BAP1, MLL, MLL3, PBRM1* (all involved in chromatin remodelling), *KEAP1, NFE2L2* (involved in the response to oxidative stress) and *AXIN1*, another component of the Wnt/β-catenin pathway (Figure 1.6) [93-95].

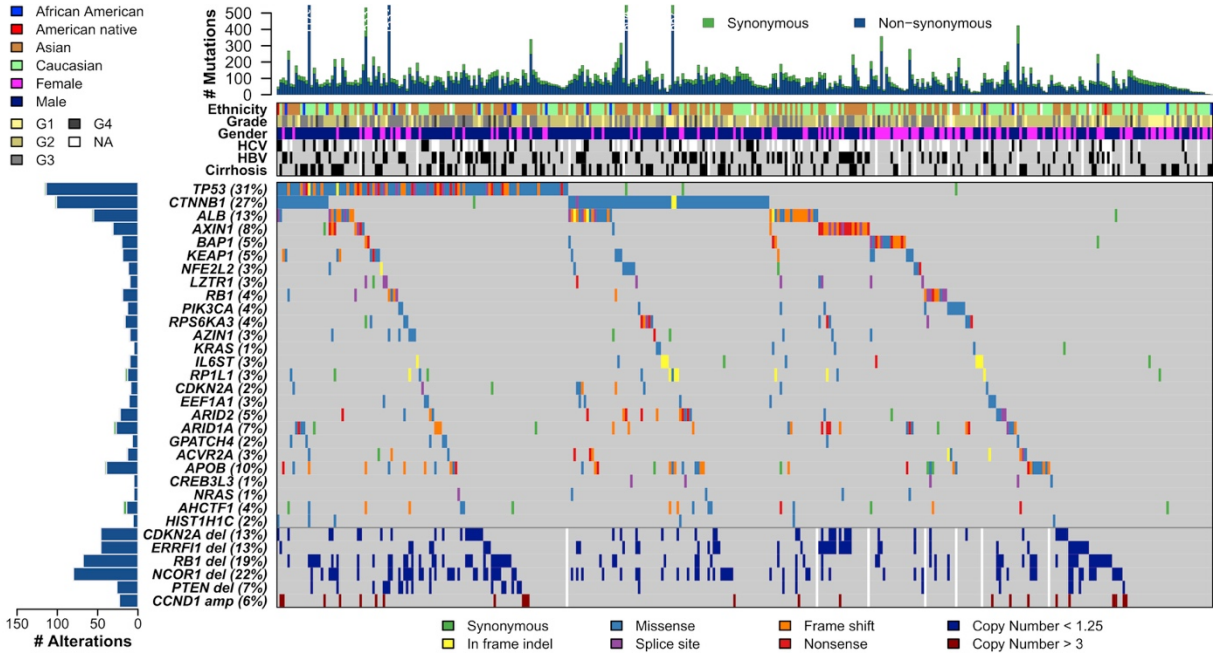


Figure 1.6: *The Genomic Landscape of Liver Hepatocellular Carcinoma. Taken from TCGA 2017* [93]*. The illustration summarises the mutational signatures in their HCC cohort (n=363). The top panel shows individual tumour mutation rates. The middle panel details ethnicity, tumour grade, age, gender, hepatitis C virus (HCV) and hepatitis B virus (HBV) infection status, and cirrhosis for each patient. Bottom panel shows genes with statistically significant levels of mutation (MutSig suite, FDR < 0.1). The bottom six rows display significant DNA copy number alterations in likely cancer driver genes. Mutation types are indicated in the legend at the bottom.*

Even if we were able to identify common alterations in cancer related genes in HCC patients, this type of tumour is characterised by high heterogeneity. For example, one of the causes is the presence or absence of hepatitis virus infections. It has been investigated that HBV-associated HCCs show mutations in the Wnt/β-catenin and JAK/STAT pathways in 65.2% and 45.5% of cases, respectively; while HCV-associated HCCs have 62.5% of mutations in *CTNNB1*. Alcohol-associated HCCs, instead, show mutations principally in chromatin remodelling genes [91,92]. One of the many interesting discoveries derived from the analysis of the mutational landscape is the high presence of mutations in the promoter of the telomerase reverse-transcriptase (*TERT*) gene, present in more than half of all HCC cases [96]. *TERT* gene codes for telomerase, one of the fundamental elements for aging in the cells, and therefore one of the molecular components involved in tumourigenesis. In the recent years mutations in *TERT* were found in several tumours [97,98] and they were also found to be associated with cirrhosis in humans [99]. These studies show that these mutations in the promoter could not only increase the promoter activity and therefore *TERT* transcription in general, but also create a potential binding site for other unusual transcription factors.

When talking about the HCC landscape, it is also important to mention the copy number alterations (CNAs). As for other, in HCC cases gains and losses are in specific chromosomal regions (for example, gain of chromosomes 1q, 5, 6p, 7, 8q, 17q, and 20, and loss of 1p, 4q, 6q, 8p, 13q, 16, 17p and 21 are the most frequent). Deletions in *CDKN2A-CDKN2B* were identified in 6.4% of cases, followed by deletions in *AXIN1* (3.2%) and *IRF2* (3.2%) [91,92].

Thanks to modern sequencing techniques, a significant number of HCCs has been analysed. The characterisation not only at genomic level but also at transcriptional level, opened the possibility to the identification of molecular subtypes for HCC. Several research were performed to address this question, but the complexity and the heterogeneity of this kind of tumour did not allow a recognised unique classification. There are classifications based on transcription data correlated with clinical and molecular features [100] and on differences in the rate of chromosomal instability [101]. The classification proposed by the most recent TCGA study, already mentioned previously [93], is based on genetic (CNAs) and transcription data. Each subgroup was characterized by clinical associations. Briefly, they divided HCC tumours in three subgroups: iClust 1 (main features: high vascular invasion and tumour grade, low rate of *CTNNB1*, *TERT* and *HNF1A* mutations); iClust2 (low vascular invasion and tumour grade, high rate of *CTNNB1*,*TERT* and *HNF1A* mutations); iClust3 (high rate not only of *CTNNB1*, *TERT* and *HNF1A* mutations, but also of *TP53* mutations and chromosome instability). Additionally to these components, they also added the correlation of epigenetic features, by analysis of DNA methylation profile, microRNA and protein expression, performed only in some of HCC cases of their cohort. To investigate the epigenetic of HCC is indeed the new frontier for the characterisation of HCC tumours.

In conclusion, the heterogeneity as well as the background aetiology might also be responsible for differential mutation rates of cancer drivers and associated pathways among different studies. The majority of genetic studies in HCC have been performed using comprehensive sequencing panels, or straight to whole exome sequencing [102-104]. So far, the majority of the genes often mutated only in HCC but rarely in other tumours (such as those important for hepatocyte differentiation and inflammatory response in liver) are not or partially targeted in commercially available panels.

V. HMGA1

HMGA protein family

The High Mobility Group A (HMGA) protein family is a group of small non-histone nuclear proteins known as ‘architectural transcriptional factors’. The *HMGA* gene family consists of the *HMGA1*

(human chr 6p21) and *HMGA2* (human chr 12q14) genes. The *HMGA1* gene contains 8 exons distributed over a region of about 10 kb while the *HMGA2* gene contains 5 exons distributed over a much larger genomic region of about 160 kb, because of its longer untranslated regions and introns [105].

HMGA1 and *HMGA2* together encode four proteins: HMGA1a, HMGA1b, HMGA1c and HMGA2 [106]. HMGA1a and HMGA1b are encoded by the *HMGA1* gene and are isoforms assembled through alternatively spliced mRNA that differ by 11 amino acid residues between the first and the second AT-hook domains (107 and 96 amino acids, respectively). HMGA1c (156 amino acid) is the rarest and most recently identified isoform [107]. It is also encoded by the *HMGA1* gene by alternative splicing using non-canonical splice donor and acceptor sites. This alternative splicing results in a frame shift such that HMGA1a and HMGA1c are identical in their first 65 amino acids but differ thereafter [107]. HMGA2 (109 amino acid) is encoded by the *HMGA2* gene and presents a structure very similar to HMGA1b, but contains a short peptide of 12 amino acid residues between the third AT-hook and the acidic C-terminal [108,109]. All the proteins in the HMGA family contain three basic "AT-hook" domains and an acidic C-terminal region that allow them to bind AT-rich DNA sequences in the minor groove of the double helix (Figure 1.7) [108,109].

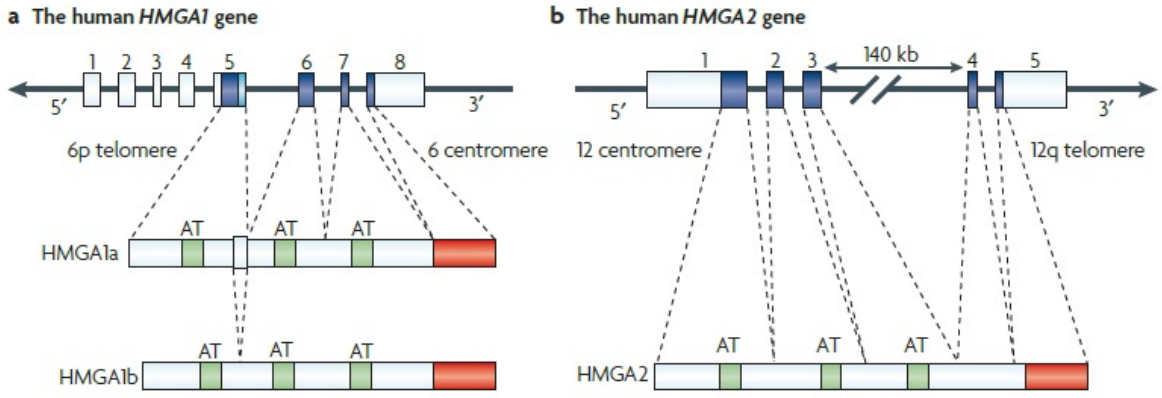


Figure 1.7: *Characteristics of the HMGA genes and proteins. Taken from Fusco 2007 [108].*

The expression of HMGA proteins is high during embryogenesis whereas it is undetectable or very low in differentiated adult tissues, except for a few specific tissues such as the testis and the thymus [105,110]. Nonetheless, especially HMGA1 was found to be again highly expressed in a broad range of malignancies [111-121]. For this reason, several studies have been tried to elucidate its role in cell transformation and its mechanism of action. To date, the discoveries reveal a complex situation, with numerous interaction systems for both DNA and transcription factors [122-127].

Mechanisms of action

The most studied mechanism of action of HMGA proteins is represented by the interaction and the binding with the DNA and the following recruitment of transcription factors. HMGA proteins bind the DNA, at that point they bind transcription factors, both with a direct and indirect binding, to form macromolecular complexes and together they can promote or repress the expression of target genes [128,129]. Another way HMGA proteins use to modify gene transcription is by binding with a transcription factor, therefore they are able to modify its conformation and improve its DNA binding affinity [108,130]. Finally, HMGA proteins can alter the chromatin structure [131]. For example, it has been shown that HMGA1 competes with histone H1, is able to displace it, to open the minor groove and to facilitate the recruitment of transcription factors [132]. For these reasons, whilst HMGA1 does not have transcriptional activity *per se*, its overexpression could trigger the deregulation of oncogene and tumour suppressor gene expression, leading to transformation and cancer progression [108,115]. It is indeed acknowledged that HMGA1 proteins participates in a myriad of cellular processes implicated by all hallmarks of cancer [133], including cell cycle regulation and chromosomal changes, DNA replication and repair, apoptosis, but also mitochondrial function and retroviral integration [134-140]. Figure 1.8 simplifies the complexity of HMGA1 networks in cancer development and tumour progression and its direct and indirect targets.

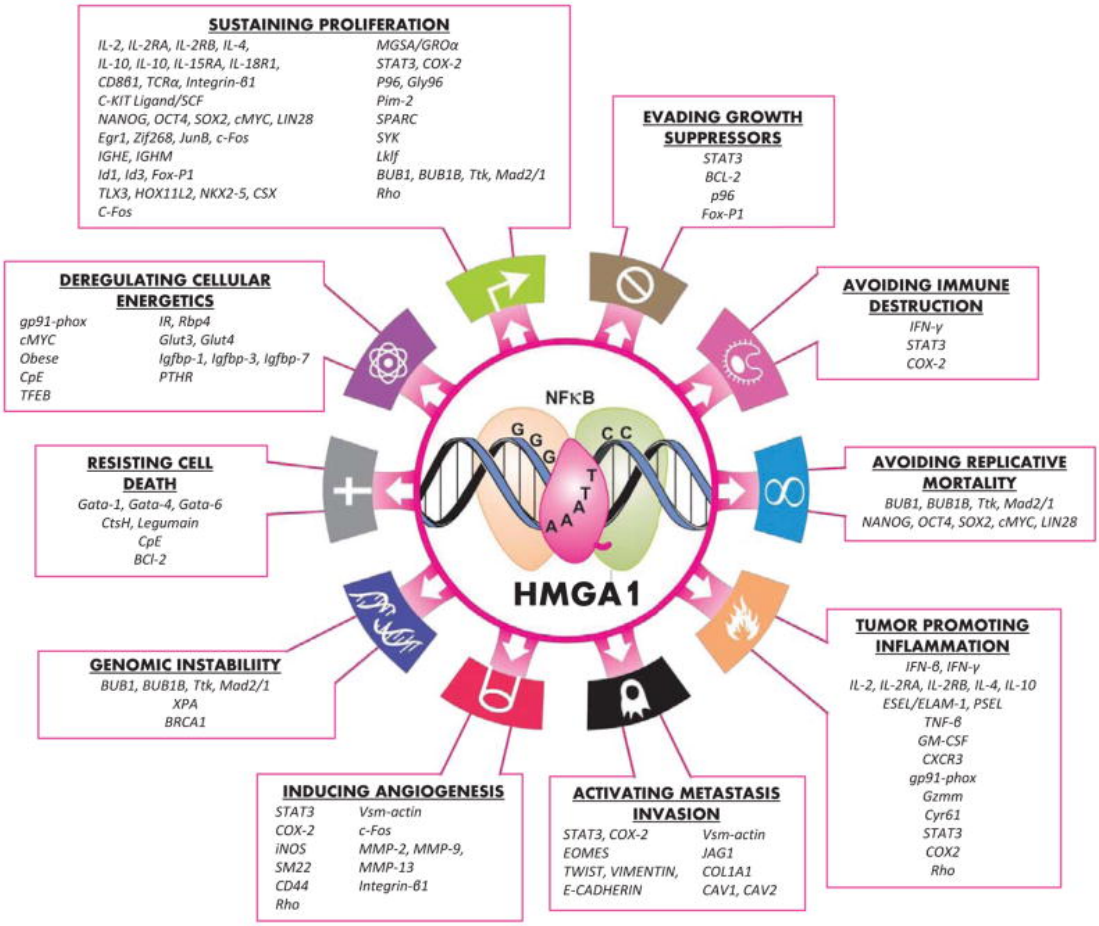


Figure 1.8: *HMGA1 networks involve all hallmarks of cancer. Taken from Sumter et al. 2016* [133].

HMGA1 and its role in carcinogenesis

HMGA1 overexpression was first associated with the neoplastic phenotype in rat thyroid transformed cells [141] and it has been described since in many human carcinomas, including those of the colon, breast, pancreas, ovary, lung, oesophagus, amongst others [111-120]. Importantly, the expression level of HMGA1 has been found to correlate with the aggressiveness of colorectal carcinomas [110,112]. HMGA1 overexpression is associated with invasion-positive and advanced staged colorectal carcinomas and with the presence of distant metastasis. To further support the role of HMGA1 in tumour progression, its expression levels have been found to be associated with histologic grades of breast and ovarian carcinomas, where HMGA1 expression increases progressively from no expression in normal breast tissue, to moderate expression in hyperplastic lesions to strong overexpression in ductal carcinomas [114], and from a weak expression in ovarian carcinomas with low invasive potential to high expression in invasive carcinomas [115].

Regarding its molecular targets, it was recently demonstrated that HMGA1 binds to the AT-rich promoter of osteopontin, a protein involved in the acquisition of fully transformed features in

human carcinomas [142]. Not only is HMGA1 able to bind the osteopontin promoter, it is also able to compete and interfere with other regulators of transcription present in the same sites of this promoter [142]. Conversely, HMGA1 binding to the promoter regions of *TP53* results in negative transcriptional regulation [143]. Furthermore, several studies have demonstrated that HMGA1 directly activates specific gene-subsets involved in tumour growth, migration, invasion, resistance to drug-induced cell death as well as to epithelial-mesenchymal transition in cancer cells [142,144,145]. Another study describes how HMGA1 promotes an undifferentiated pluripotent stem-like cell state throughout the induction of several genes including *SOX2*, *LIN28* and *cMYC* [143,146]. In human embryonic stem cells, HMGA1 binds to the promoters of these genes, thus suggesting that it can directly regulate their expression [147]. In fact, direct evidence of the role of HMGA1 in carcinogenesis, tumour progression and induction of stem-like properties has been provided in several experimental animal models [148-151]. These data demonstrate the multi-faceted function of HMGA1 and the intricate ways it regulates other cellular processes implicated in carcinogenesis. Despite the molecular characterisation of HMGA1 function carried out to date, a systematic analysis of the genes directly and/or indirectly regulated by HMGA1 have not been performed.

HMGA1 and HCC

The locus where *HMGA1* is located is gained in around 40% of HCCs [91]. An early study suggested that HMGA1 is expressed in 30% of primary HCC on the mRNA level and 13% on the protein level [152]. It should be noted that in this study, HMGA1 expression was only assessed in treated HCCs and thus HMGA1 levels may have been altered as a result of treatment. Moreover, earlier iterations of HMGA1 antibodies were fraught with issues of specificity. More recently, the expression of HMGA1 in HCC cases was investigated by my team in our laboratory [121]. HMGA1 expression was evaluated in two independent cohorts of 59 and 192 HCC cases through gene expression microarray and immunohistochemistry. We demonstrated that HMGA1 levels increase through progression stages from normal liver to HCC, both at mRNA and protein levels. Furthermore, we showed that more than 50% of HCCs are HMGA1-positive and this high expression is associated with poor prognosis. Finally, functional examinations supported the involvement of HMGA1 in cell growth and migration in liver cancer cells [121].

All these findings demonstrated not only an overexpression of HMGA1 in the HCC context, but also evidence that HMGA1 confers a neoplastic advantage to liver cancer cell lines. However, given its multifaceted functions, HMGA1 cannot currently be exploited as a therapeutic target and further characterisation in liver biology will provide novel insights into its mechanisms of action in driving disease progression.

2- Rationale and Aims of the Thesis

The main objective of my project was to investigate new diagnostic and prognostic markers, both on a genetic and molecular level, in the context of hepatocellular carcinoma (HCC). A better understanding of these mechanisms may guide alternative target therapies for HCC patients.

In **Chapter I**, we aimed to develop an HCC-specific custom made sequencing panel using the Ion Torrent platform, containing the genes and loci most commonly affected by somatic mutations and copy number alterations (CNAs) in HCC. We wanted to create a panel that generates reliable data using all kinds of patient biopsies, from frozen to formalin-fixed paraffin-embedded (FFPE) tissues to liquid biopsies. Moreover, we wanted to make the analysis user friendly but reliable and reproducible and, with the work of the bioinformaticians in our laboratory, we created a somatic variant calling pipeline specific for Ion Torrent sequencing data.

In **Chapter II**, we focused our work on the investigation of the oncogenic role of HMGA1 in HCC. This protein is often overexpressed in many types of cancers including HCC, in which we demonstrated that HMGA1 levels increase through progression stages from normal liver to HCC [121]. To better understand the significance of HMGA1 overexpression, we aimed to molecularly characterise HMGA1 and to explore its molecular targets in the HCC *in vitro* environment.

Both chapters underline the importance of the discovery and the functional understanding of tumour markers and their molecular mechanisms in HCC. The findings may lead to more individualised treatment approaches for HCC patients.

3- Results

3.1- Chapter I

Design and validation of a custom made sequencing panel for the screening of HCC somatic mutations

Hepatocellular carcinoma (HCC), as the other type of cancer, presents a distinct mutational landscape. There are numerous genes commonly mutated in HCC but not frequently in other tumours that are currently not targeted, or are only partially targeted, in commercial sequencing panels. Our objective was to construct a high-throughput and cost effective sequencing panel specifically to screen for the most common somatic alterations in HCC. We wanted to develop a sequencing-panel applicable for frozen tissues but also with low input material such as formalin-fixation and paraffin-embedded (FFPE). Moreover, we wanted to be valuable also if used with plasma derived cell-free DNA (cfDNA). Last but not least, we developed a high sensitive and specific somatic variant calling pipeline to use to analyse this kind of sequencing data.

This chapter contains my work on the design, the validation and the feasibility of this HCC specific sequencing panel and its use to identify alterations in HCC patients. It is divided in three parts, resulting in three manuscripts reported after the summary of the comprehensive work.

The first one, “Diagnostic Targeted Sequencing Panel for Hepatocellular Carcinoma Genomic Screening”, aimed to design an amplicon-based sequencing panel for Ion Torrent technology, the most available and economical sequencing methods in diagnostic laboratories. The design of the panel was performed using genomic regions frequently altered in HCC according to publicly available data, as explained in details in the manuscript. The panel testing was achieved using a cohort of fresh frozen and FFPE biopsies of HCC tissue samples and the results were compared to the results obtained by whole exome sequencing (WES) performed on the same samples. All but one mutation identified from WES were detected by using our custom HCC panel. Additional mutations within the coding regions were identified thanks to the higher depth of the sequencing obtained with the panel compared to WES. Moreover, several mutations detected with the HCC panel were within the promoter and long non-coding RNA (lncRNA) regions, so not possible to be found by WES and not currently targeted by commercial panels. We demonstrated that our custom panel is high-throughput and cost effective and allows the screening for somatic alterations specific for HCC samples even with the use of low-input DNA. Furthermore, we demonstrated that, using this kind of samples, it is also possible to detect copy number variations in genes commonly gained or lost in HCC.

Because of the invasive nature of tissue biopsies, plasma-derived cfDNA is becoming a new potential alternative to tissue biopsies for the screening of mutations for detection and surveillance of the tumour. In the second part of this chapter, “Genetic profiling using plasma-derived cell-free DNA in therapy-naïve HCC patients: a pilot study”, we explored whether somatic mutations in HCC driver genes could be detected with high confidence using our custom amplicon-based sequencing panel in the cfDNA of HCC patients who have not undergone systemic therapy. We

used blood samples synchronously collected with a core needle tumour biopsy from a prospective cohort study and we determined if the range of mutations in the cfDNA is representative of the tumour biopsy. The potential of liquid biopsy-based biomarker identification led us to a publication of another study aimed to evaluate the feasibility of cfDNA extraction and somatic mutation assessment in 30-year-old sera that had been collected from patients with breast cancer (see Annex). Our conclusions support the robustness of current next generation sequencing to accurately sequence cfDNA to detect cancer-specific mutations in these old samples, despite the long cryopreservation and repeated changes of storage location. These findings encourage the use of long-term storage of biological samples in longitudinal studies prior to analysis, with the possibility to assess the prognostic role of pathogenic mutations in cfDNA present at diagnosis by comparing overall and relapse-free survival between patients with or without specific mutations.

The third part, “PipeIT: A Singularity Container for Molecular Diagnostic Somatic Variant Calling on the Ion Torrent Next-Generation Sequencing Platform”, is our solution to the extensive manual review of the results required for a diagnostic laboratory to analyse somatic mutations data obtained by Ion Torrent sequencing platforms. Moreover, the lack of optimised analysis workflows for custom targeted sequencing panels usually lead to poor reproducibility and portability. Thanks to our bioinformaticians, we developed PipeIT, a stand-alone singularity container of a somatic mutation calling and filtering pipeline for matched tumour-normal Ion Torrent sequencing data, able to generate data with high positive predictive value and high sensitivity. This pipeline ensures the reproducibility of data and reduces the need for manual curation of the results.

Diagnostic Targeted Sequencing Panel for Hepatocellular Carcinoma Genomic Screening

Viola Paradiso*, Andrea Garofoli*, Nadia Tosti, Manuela Lanzafame, Valeria Perrina, Luca Quagliata, Matthias S. Matter, Stefan Wieland,y Markus H. Heim,yz Salvatore Piscuoglio, Charlotte K.Y. Ng, and Luigi M. Terracciano

2018 JMD CME Program

The Journal of Molecular Diagnostics, Vol. 20, No. 6, November 2018

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

Diagnostic Targeted Sequencing Panel for Hepatocellular Carcinoma Genomic Screening

Viola Paradiso,* Andrea Garofoli,* Nadia Tosti,* Manuela Lanzafame,* Valeria Perrina,* Luca Quagliata,* Matthias S. Matter,* Stefan Wieland,† Markus H. Heim,†‡ Salvatore Piscuoglio,* Charlotte K.Y. Ng,*† and Luigi M. Terracciano*

From the Institute of Pathology, University Hospital Basel, Basel; the Department of Biomedicine,† University of Basel, Basel; and the Department of Gastroenterology and Hepatology,‡ University Hospital Basel, Basel, Switzerland*

CME Accreditation Statement: This activity ("JMD 2018 CME Program in Molecular Diagnostics") has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education (ACCME) through the joint providership of the American Society for Clinical Pathology (ASCP) and the American Society for Investigative Pathology (ASIP). ASCP is accredited by the ACCME to provide continuing medical education for physicians.

The ASCP designates this journal-based CME activity ("JMD 2018 CME Program in Molecular Diagnostics") for a maximum of 18.0 AMA PRA Category 1 Credit(s)™. Physicians should claim only credit commensurate with the extent of their participation in the activity.

CME Disclosures: The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Accepted for publication July 2, 2018.

Address correspondence to Luigi M. Terracciano, M.D., Institute of Pathology, University Hospital Basel, Schoenbeinstrasse 40, 4031 Basel, Switzerland. E-mail: luigi.terracciano@usb.ch.

Commercially available targeted panels miss genomic regions frequently altered in hepatocellular carcinoma (HCC). We sought to design and benchmark a sequencing assay for genomic screening of HCC. We designed an AmpliSeq custom panel targeting all exons of 33 protein-coding and two long non-coding RNA genes frequently mutated in HCC, *TERT* promoter, and nine genes with frequent copy number alterations. By using this panel, the profiling of DNA from fresh-frozen ($n = 10$, 1495×) and/or formalin-fixed, paraffin-embedded (FFPE) tumors with low-input DNA ($n = 36$, 530×) from 39 HCCs identified at least one somatic mutation in 90% of the cases. Median of 2.5 (range, 0 to 74) and 3 (range, 0 to 76) mutations were identified in fresh-frozen and FFPE tumors, respectively. Benchmarked against the mutations identified from Illumina whole-exome sequencing (WES) of the corresponding fresh-frozen tumors (105×), 98% (61 of 62) and 100% (104 of 104) of the mutations from WES were detected in the 10 fresh-frozen tumors and the 36 FFPE tumors, respectively, using the HCC panel. In addition, 18 and 70 somatic mutations in coding and noncoding genes, respectively, not found by WES were identified by using our HCC panel. Copy number alterations between WES and our HCC panel showed an overall concordance of 86%. In conclusion, we established a cost-effective assay for the detection of genomic alterations in HCC. *(J Mol Diagn 2018, 20: 836—848; https://doi.org/10.1016/j.jmoldx.2018.07.003)*

Sequencing technologies have allowed the discovery of genetic alterations essential in the diagnosis and treatment of human cancer or approval of new targeted therapies.[1] In addition, the presence of subclonal mutations has direct implications in the development of drug resistance.[2,3] In the era of precision medicine, the development of rapid, accurate, high-throughput, and cost-effective genomic assays to accommodate the increasingly genotype-based therapeutic approaches is required.[4,5] Currently, the costs of whole-genome and whole-exome sequencing (WES) are still prohibitive in the clinical setting, especially for small institutions. Furthermore, although DNA from fresh-frozen

Supported in part by the Swiss Cancer League (Oncosuisse) grants KLS-3639-02-2015 (L.M.T.) and KFS-3995-08-2016 (S.P.), Krebsliga beider Basel project KLbB-4183-03-2017 (C.K.Y.N.), Swiss National Science Foundation Ambizione grant PZ00P3_168165 (S.P.), the Swiss Centre for Applied Human Toxicology (SCAHT; V.Pa.), and the European Research Council ERC Synergy grant 609883 (C.K.Y.N. and M.H.H.).

V.Pa. and A.G. contributed equally to this work.

C.K.Y.N. and L.M.T. contributed equally to this work as senior authors.

Disclosures: None declared.

Funding bodies had no role in the design of the study, collection, analysis, and interpretation of the data or the writing of the manuscript.

https://doi.org/10.1016/j.jmoldx.2018.07.003

tissue is ideal for genomic screening, it is not part of routine diagnostic practice at most hospitals and institutions. Instead, DNA from formalin-fixed paraffin-embedded (FFPE) material is frequently the only option. Moreover, DNA from small tumors, after reserving materials for histopathologic analyses, may be extremely limited. For research institutes, being able to exploit and revisit archival materials associated with long-term follow-up but whose DNA may potentially be degraded is also highly desirable. Given these limitations, PCR-based sequencing panels may be more broadly applicable than capture-based solutions.

Existing commercial sequencing panels, such as the amplicon-based Ion Torrent Oncomine Comprehensive Assay version 3 (Thermo Fisher Scientific, Waltham, MA) and the capture-based Foundation Medicine FoundationOne assay, are broadly applicable to common cancer types. Compared with other common cancer types, however, hepatocellular carcinoma (HCC) has a distinct mutational profile. Although HCC driver genes *TP53* and *CTNNB1* are also frequently mutated in cancers such as those of the lungs, the breasts, and colon,[6] genes such as *APOB*, *ALB*, *HNF1A*, and *HNF4A* are significantly mutated only in HCC.[7–17] The distinct mutational landscape of HCC is likely a result of the unique biology of hepatocyte differentiation and liver functions. Of note, the frequently altered *APOB*, *ALB*, and *HNF4A* are not targeted by most commercial assays. In the noncoding regions, recent commercially available panels include *TERT* promoter mutation hotspot (c.-124C>T). However, long noncoding RNA (lncRNA) genes frequently mutated in HCC, such as *MALAT1* and *NEAT1*,[16] have yet to be included in commercial panels or in exome capture panels. Recent whole-genome studies have also uncovered mutation clusters in promoter regions of genes such as *MED16*, *WDR74*, and *TFPI2*[16,18] that are not covered in commercial panels.

In this study, we designed a high-throughput and cost-effective amplicon-based sequencing panel specifically to screen for somatic mutations and copy number alterations (CNAs) in HCC. Our panel includes genes and regions frequently altered in HCC, including those not currently covered by commercial panels. We tested the sequencing panel by using fresh-frozen and FFPE materials with low-input DNA to evaluate the feasibility of this panel in routine diagnostics.

Materials and Methods

Targeted Panel Design and Generation

A custom targeted sequencing panel that focused on the most frequently altered genes in HCC[7–18] was designed by using Ion Ampliseq Designer (Thermo Fisher Scientific). The panel (hereafter the HCC panel) covers all exons of 33 protein-coding genes; recurrently mutated lncRNA genes *MALAT1* and *NEAT1*; and the recurrently mutated promoter regions of *TERT*, *WDR74*, *MED16*, and *TFPI2* (Figure 1A and Supplemental Table S1).[7–18] Nine genes frequently altered by CNAs and mutation hotspots in seven cancer genes are also covered (Figure 1A and Supplemental Table S1).[7–18] The HCC panel was designed by using the FFPE option for smaller amplicon size. The nine genes for CNA profiling were designed to be covered by at least 10 non-overlapping amplicons evenly distributed across the length of the genes. The designed panel was further inspected by

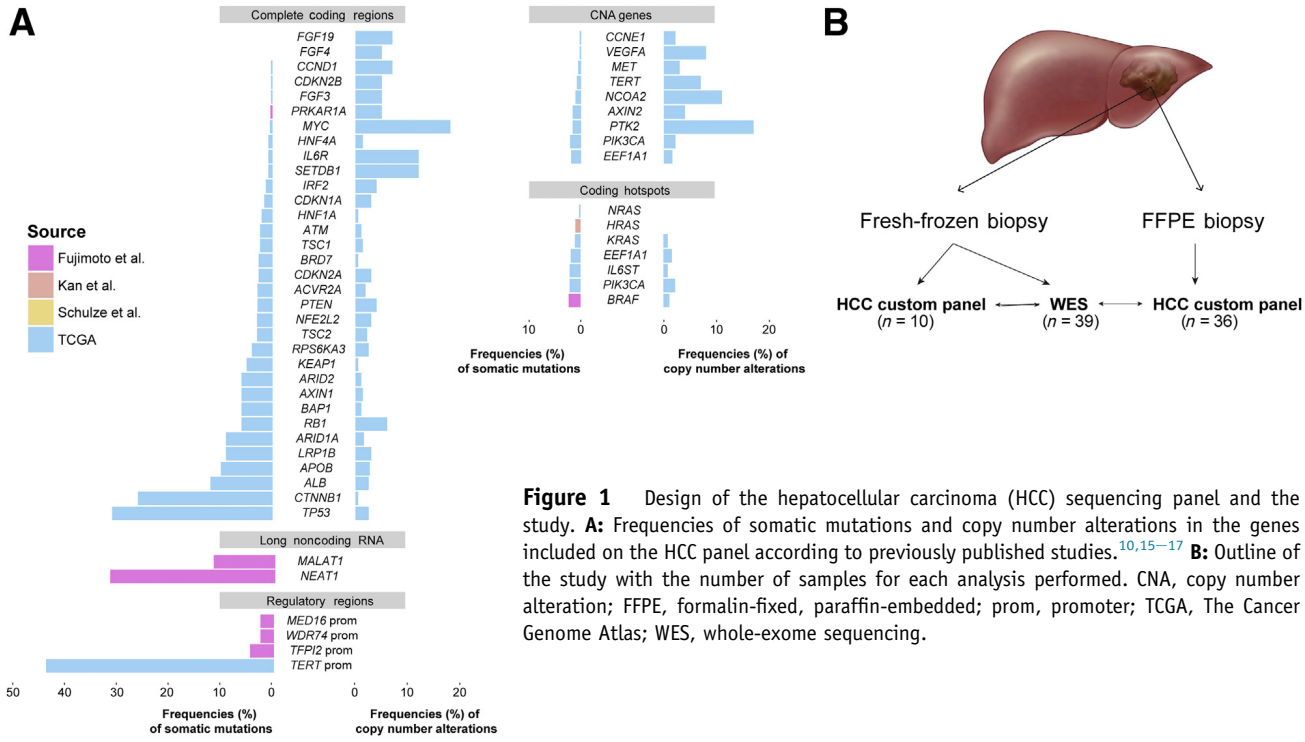


Figure 1 Design of the hepatocellular carcinoma (HCC) sequencing panel and the study. **A:** Frequencies of somatic mutations and copy number alterations in the genes included on the HCC panel according to previously published studies.[10,15–17] **B:** Outline of the study with the number of samples for each analysis performed. CNA, copy number alteration; FFPE, formalin-fixed, paraffin-embedded; prom, promoter; TCGA, The Cancer Genome Atlas; WES, whole-exome sequencing.

the white glove service (Thermo Fisher Scientific) for primer specificity in a multiplex PCR reaction. The HCC panel consists of 2120 amplicons split into two primer pools and covers genomic regions of approximately 203 kb.

Tissue Samples

Human tissues were obtained from patients undergoing diagnostic liver biopsy at the University Hospital Basel, Basel, Switzerland. Written informed consent was obtained from all included patients. Ultrasound-guided needle biopsies were obtained from tumor lesion(s) and adjacent nontumoral liver tissue (Figure 1B). The study was approved by the ethics committee of the northwestern part of Switzerland (protocol EKNZ 2014-099). For all patients except cases 2, 6, 7, and 9, a single tumor biopsy was included (Supplemental Table S2). For cases 6 and 7, two tumor biopsies were included, and for cases 2 and 9, three tumor biopsies were included. A portion of each biopsy was FFPE for clinical purposes, and the remaining portion of each biopsy was snap-frozen and stored at −80°C for research purposes. For this study, 45 fresh-frozen tumor biopsies and 39 fresh-frozen nontumor biopsies from 39 patients were included. FFPE tissue samples that remained after diagnostic routine (36 tumor biopsies and 31 nontumor biopsies from 36 patients) were included. Pathologic assessment of tumor content was performed by two expert hepatopathologists (M.S.M. and L.M.T.) with the use of diagnostic hematoxylin and eosin slides.

DNA Extraction

DNA from fresh-frozen biopsies was extracted by using the ZR-Duet DNA/RNA MiniPrep Plus kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. Before extraction, tissue samples were crushed in liquid nitrogen to facilitate lysis. For DNA extraction from FFPE samples, one 5-μm−thick slide was cut directly in the tube, and DNA was extracted with the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions as previously described.[19,20] DNA was quantified by using the Qubit Fluorometer (Thermo Fisher Scientific).

Library Preparation and Deep Sequencing Using the HCC Panel

Library preparation for the HCC panel was performed by using the Ion AmpliSeq library kit version 2.0 (Thermo Fisher Scientific) according to the manufacturer's guidelines. For cases 2, 6, 7, and 9, DNA extracted from multiple fresh-frozen tumor biopsies was pooled equimolar before library preparation (Supplemental Table S2). In total, 20 fresh-frozen samples (10 tumor samples and 10 nontumoral counterparts) and 67 FFPE samples (36 tumor biopsies and 31 nontumoral counterparts) were sequenced by using the HCC panel.

The HCC panel consists of two pools of amplification primers. Ten nanograms of DNA per sample was used for library preparation for each pool. Amplification was performed according to the manufacturer's guidelines. The amplicons from the two pools were combined and treated to digest the primers and to phosphorylate the amplicons. The amplicons were then ligated to Ion Adapters (Thermo Fisher Scientific) by using DNA ligase. Finally, cleaning and purification of the generated libraries were performed with Agencourt AMPure XP (Beckman Coulter, Brea, CA) according to the manufacturer's guidelines. Quantification and quality control were performed with the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific). Samples were diluted to reach the concentration of 40 pmol and then were pooled for sequencing. Twenty-five μL of the pooled libraries was loaded on Ion 530 Chip (Thermo Fisher Scientific) and processed in Ion Chef Instrument (Thermo Fisher Scientific). Sequencing was performed on Ion S5 XL system (Thermo Fisher Scientific).

Sequence Data Analysis for the HCC Panel

Sequence reads were aligned to the human reference genome hg19 by using TMAP within the Torrent Suite Software version 5.4 (Thermo Fisher Scientific; *https://github.com/iontorrent/TS*) for the Ion S5XL system. Coverage analysis was performed by using Picard's CollectTargetedPcrMetrics tool version 2.4.1 (*http://broadinstitute.github.io/picard*) (Supplemental Table S3). Uniformity of sequencing was defined as the proportion of target bases covered at >20% of mean amplicon coverage for a given sample. Comparison of the coverage for the two primer pools was performed by using paired Wilcoxon test.

Somatic mutations were identified with Torrent Variant Caller version 5.0.3 (Thermo Fisher Scientific; *https://github.com/iontorrent/TS*). For fresh-frozen samples, the corresponding fresh-frozen nontumoral samples were used as the germline control. For FFPE samples, FFPE nontumoral samples were used as the matched germline sample when available. When FFPE nontumoral samples were not available, the corresponding fresh-frozen nontumoral samples were used as germline control. Mutations at hotspot residues were whitelisted.[21,22] Mutations supported by <8 reads, and/or those covered by <10 reads in the tumor or <10 reads in the matched nontumoral counterpart were filtered out. Only those for which the tumor variant allele fraction (VAF) was >10 times that of the matched nontumoral VAF were retained to ensure the somatic nature of the variants. Because of the repetitive nature and the high GC content of the *TERT* promoter region, *TERT* mutation hotspots (chr5:1295228 and chr5:1295250) were additionally screened. *TERT* promoter mutations were considered present if supported by at least five reads or VAF of at least 5%. All mutations were manually inspected by using the Integrative Genomics Viewer version 2.3.69 (*https://software.broadinstitute.org/software/igv*).[23]

CNAs were defined as follows. For each sample, end-to-end sequence reads were extracted separately for the two

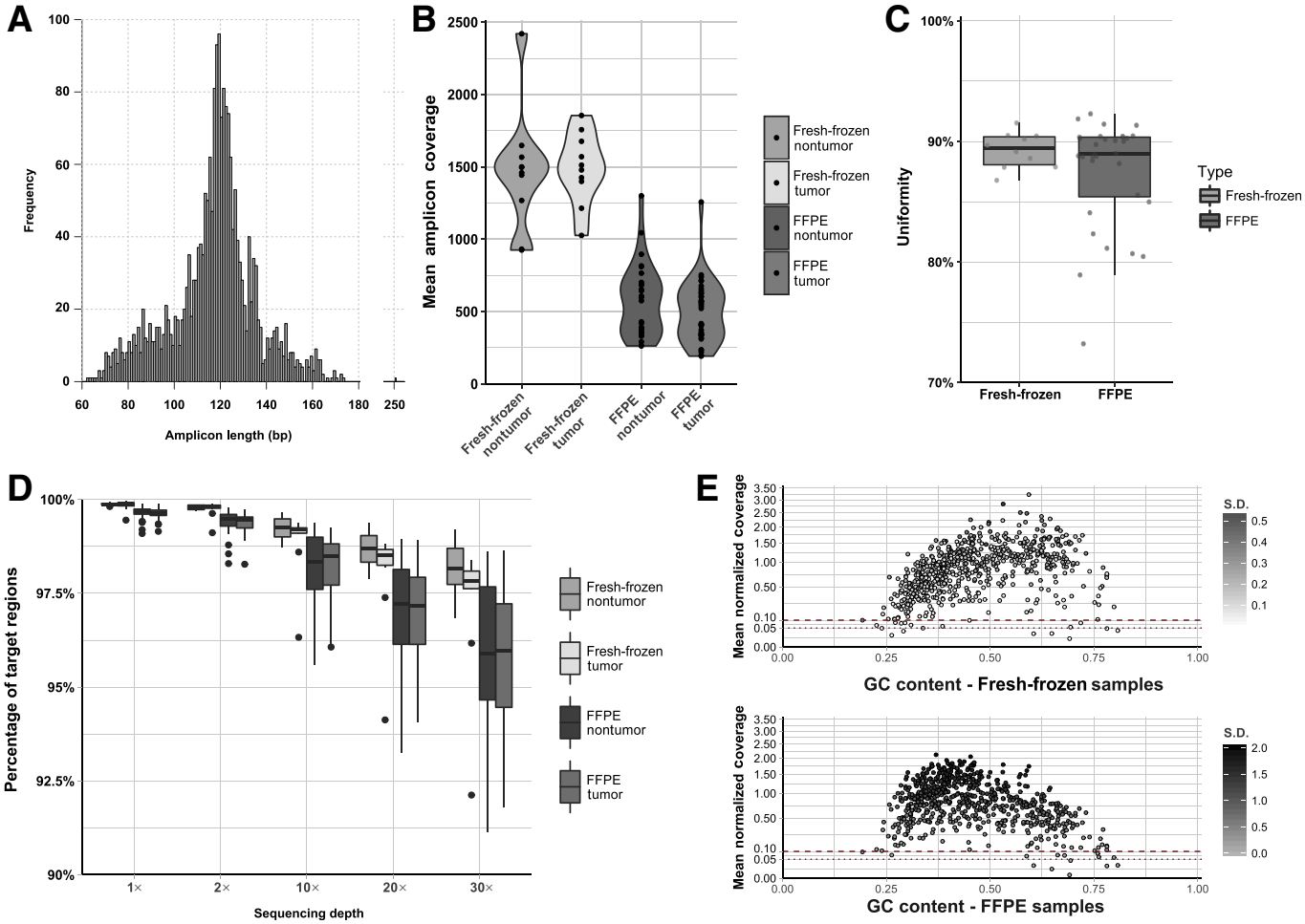


Figure 2 Coverage analyses and statistics of the hepatocellular carcinoma (HCC) panel. **A:** Distribution of the amplicon sizes on the HCC panel. **B:** Violin plots of the mean amplicon coverage across fresh-frozen nontumor; fresh-frozen tumor; formalin-fixed, paraffin-embedded (FFPE) nontumor; and FFPE tumor samples. **C:** Coverage uniformity, defined as the percentage of target bases covered at >20% of the mean coverage, in fresh-frozen and FFPE nontumor samples. **D:** Percentages of target regions covered at various depths (1×, 2×, 10×, 20×, and 30×) across fresh-frozen nontumor, fresh-frozen tumor, FFPE nontumor, and FFPE tumor samples. **E:** Scatter plot of GC content and mean normalized coverage for all amplicons in fresh-frozen and FFPE samples. Color of the dots indicates the SD of mean normalized coverage within each group. **Dashed red lines** indicate the mean normalized coverage at 0.1 and 0.05.

amplicon pools. A copy number reference for each pool was generated by using all nontumoral samples to estimate overall read depth, $\log_2$ ratio, and variability by using the reference function from CNVkit version 0.9.0 (*https://github.com/etal/cnvkit*).[24] Amplicons with <100 read depth, absolute $\log_2$ ratio >1.5, or spread >1 were removed from copy number analysis. Protein-coding genes for which the complete coding region was included in the panel or for which amplicons were specifically designed for copy number analysis were included. Samples with excessive residual copy number $\log_2$ ratio (segment interquartile range >0.8) were excluded, as previously described.[25]

For each tumor/nontumor pair, $\log_2$ ratio was computed for each amplicon, separately for the two amplicon pools by using Varscan2 version 2.4.3 (*https://github.com/dkoboldt/varscan*).[26] Log_2 ratios for the two pools were separately centered then merged for segmentation by using circular binary segmentation.[27] CNAs were determined, adopting a previously described approach.[20] In brief, SD of the $\log_2$ ratios of the 40% of the central positions ordered by their $\log_2$ ratios was computed. Copy number gains and amplifications/high gains were defined as +2 SDs and +6 SDs, respectively. Copy number losses and deep deletions were defined as −2.5 SDs and −7 SDs, respectively. All gene amplifications and deep deletions were visually inspected by using $\log_2$ ratio plots.

To evaluate the impact of tumor purity on CNA analysis, an *in silico* simulation was performed on 12 cases (six frozen and six FFPE, selected on the basis of the presence of gene amplification/high gain or deep deletion), by replacing tumor reads with reads sampled from the normal samples to simulate tumor content 5%, 10%, 20% up to the actual tumor content for the samples. CNA analysis was performed as described above.

WES

WES was performed for DNA extracted from the 45 tumor biopsies and 39 nontumoral counterparts from the 39

patients (Supplemental Table S2). Whole-exome capture was performed by using the SureSelectXT Clinical Research Exome (Agilent, Santa Clara, CA) platform according to the manufacturer's guidelines. Sequencing (2 × 101 bp) was performed at the Genomics Facility of ETH Zurich Department of Biosystems Science and Engineering (Basel, Switzerland) by using Illumina HiSeq 2500 (Illumina, San Diego, CA) according to the manufacturer's guidelines. Sequence reads were aligned to the reference human genome GRCh37 by using Burrows-Wheeler Aligner-MEM version 0.7.12 (*http://bio-bwa.sourceforge.net*).[28] Local realignment, duplicate removal, and base quality adjustment were performed by using the Genome Analysis Toolkit version 3.6 (*https://software.broadinstitute.org/gatk*)[29] and Picard version 2.4.1 (*http://broadinstitute.github.io/picard*).

For WES samples, sequence reads overlapping with the target regions of the HCC panel were extracted for further comparative analyses. Sequencing statistics were evaluated for the overlap of the target regions of the WES and the HCC panel. For cases 2, 6, 7, and 9, for which DNA from multiple fresh-frozen tumor biopsies was pooled before sequencing by using the HCC panel, WES reads from the multiple biopsies were merged to facilitate downstream comparisons. For all four cases, the number of reads obtained from WES of individual biopsies was comparable (Supplemental Table S3).

Somatic single nucleotide variants and small insertions and deletions (indels) were detected by using MuTect version 1.1.4 (*https://software.broadinstitute.org/cancer/cga/mutect*)[30] and Strelka version 1.0.15 (*https://github.com/Illumina/strelka*),[31] respectively. Single nucleotide variants and small indels outside of the target regions, those with VAF of <1%, and/or those supported by <3 reads were filtered out. Only variants for which the tumor VAF was >5 times that of the matched nontumoral VAF were retained. Further, variants identified in at least two of a panel of 123 nontumoral liver tissue samples, using the artifact detection mode of MuTect2 implemented in Genome Analysis Toolkit version 3.6 were excluded,[29] where the panel of 123 nontumoral liver tissue samples included the 39 nontumoral samples in the present study and were captured and sequenced with the same protocols. All indels were manually inspected by using the Integrative Genomics Viewer.[23] Copy number analysis was performed with FACETS version 0.5.13 (*https://github.com/mskcc/facets*),[32] and genes targeted by amplifications or deep deletions were defined by using the same thresholds as above.

Pairwise Comparisons between Mutations Identified by WES, Fresh-Frozen and FFPE Tissues

Pairwise comparisons of the somatic mutations identified by WES and by the HCC panel were performed, according to the originating biopsies (Supplemental Table S2). Discordant variants were reevaluated and interrogated for their presence by supplying Torrent Variant Caller version 5.0.3 with their positions as the hotspot list (for Ion Torrent sequencing) or by Genome Analysis Toolkit version 3.6 Unified Genotyper by using the GENOTYPE_GIVEN_ALLELES mode (for WES).

Sanger Sequencing

To validate the discordant variants, Sanger sequencing was performed on both DNA from the fresh-frozen and the corresponding FFPE tumor biopsies. PCR amplification of 5 ng of genomic DNA was performed with the AmpliTaq 360 Master Mix Kit (Thermo Fisher Scientific) on a Veriti Thermal Cycler (Thermo Fisher Scientific) as previously described (Supplemental Table S4).[20] PCR fragments were purified with ExoSAP-IT (Thermo Fisher Scientific). Sequencing reactions were performed on a 3500 Series Genetic Analyzer instrument by using the ABI BigDye Terminator chemistry version 3.1 (Thermo Fisher Scientific) according to the manufacturer's instructions. All analyses were performed in duplicate. Sequences of the forward and reverse strands were analyzed with MacVector software version 15.1.3 (MacVector, Inc., Apex, NC).[20]

Analysis of TCGA Data

To determine the frequencies of high-level copy number gains/focal amplifications and deep deletions/focal homozygous deletions in HCC, the GISTIC 2.0 copy number calls for The Cancer Genome Atlas (TCGA) HCC cohort from the cBioPortal were obtained.[33] High-level gains and deep deletions were defined as those with GISTIC copy number state 2 and −2, respectively. Focal amplifications and focal homozygous deletions were defined as high-level gains and deep deletions that affected <25% of a given chromosome arm. For the 37 genes included in the copy number analysis, the frequencies of high-level gains/deep deletions and of focal amplifications/focal homozygous deletions were computed.

Statistical Analysis

Correlation analyses were performed with Pearson's r and r^2. Statistical analyses were performed in R version 3.4.2 (The R Foundation, Vienna, Austria).

Results

HCC-Specific Custom Targeted Sequencing Panel Design and Quality Assessment

An HCC sequencing panel was designed to specifically target genes and genomic regions frequently altered in HCC[7–18] (Figure 1A and Supplemental Table S1). The HCC panel consisted of complete coding regions of 33 genes involved in several pathways implicated in HCC pathogenesis, including the WNT pathway (*CTNNB1*, *AXIN1*), chromatin remodeling (*ARID1A*, *ARID2*, and *BAP1*), cell cycle regulation (*CDKN1A*, *CDKN2A*, *CDKN2B*, *CCND1*, *RPS6KA3*, *RB1*, and *TP53*),

inflammatory response (*IL6R*, *IL6ST*), and hepatocyte differentiation (*ALB*, *APOB*, *HNF1A*, and *HNF4A*). In addition, the HCC panel also targeted recurrently mutated lncRNA genes *MALAT1* and *NEAT1* and recurrently mutated promoter regions of *TERT*, *WDR74*, *MED16*, and *TFPI2*. Genes frequently altered by CNAs (eg, *CCNE1*, *VEGFA*, *TERT*) and mutation hotspots in *BRAF*, *EEF1A1*, *HRAS*, *IL6ST*, *KRAS*, *NRAS*, and *PIK3CA* were also targeted. To enable the efficient profiling of DNA samples derived from potentially degraded FFPE materials, the panel was designed by using the FFPE option for smaller amplicon size, with a mean amplicon size of 118 bp (range, 63 to 252 bp) (Figure 2A). The HCC panel was tested on the DNA extracted from 20 fresh-frozen samples (10 from tumor biopsies and 10 from nontumoral counterparts) and 67 FFPE samples (36 from tumor biopsies and 31 from nontumoral counterparts) obtained from 39 patients (Figure 1B and Supplemental Table S2).

A coverage analysis of the HCC panel was performed with the 10 fresh-frozen and 31 FFPE nontumoral DNA samples. In the fresh-frozen and FFPE nontumoral DNA samples, a mean coverage of 1478× (range, 925× to 2420×) and 580× (range, 263× to 1300×), respectively, were achieved (Figure 2B and Supplemental Table S3). No difference was found between the depth of coverage of the two pools of amplicons ($P = 0.9879$, paired Wilcoxon test) (Supplemental Figure S1A). At least 96.8% and 91.1% of the amplicons were covered at >30× and at least 98.7% and 95.6% of the amplicons were covered at >10× in the fresh-frozen and FFPE nontumor samples, respectively (Figure 2C and Supplemental Figure S1B). Median uniformity (defined as the proportion of target bases covered at >20% of the mean amplicon coverage of a given sample) was 89.9% (range, 86.8% to 91.5%) in the fresh-frozen samples and 89.0% (range, 73.3% to 92.3%) in the FFPE samples (Figure 2D). As expected, depth of sequencing of the amplicons was associated with GC content, with reduced depth at extreme GC content (Figure 2E).

HCC Panel Captures Somatic Mutations Concordant with WES and Identifies Additional Mutations

Next, the somatic mutations identified in the 10 fresh-frozen tumor/nontumoral pairs sequenced with the HCC panel were evaluated. A median sequencing depth of 1495× (range, 1026× to 1855×) in the tumor samples was achieved (Figure 2B and Supplemental Table S3). A median of 2.5 somatic mutations (range, 0 to 74 somatic mutations) were identified, including a median of 2 mutations (range, 0 to 52 mutation) in protein-coding genes (Figure 3A and Supplemental Table S4). No somatic mutations were identified for 2 of 10 cases (cases 3 and 12), although both cases had ≥50% tumor cell content (Supplemental Table S2). One case (case 9) exhibited a hypermutator phenotype with 74 somatic mutations identified.

To evaluate the somatic mutations defined with the HCC panel, the somatic mutations derived from WES, generated on the orthogonal Illumina technology, of the same DNA aliquots from the fresh-frozen tumors and matched nontumor samples were used as a benchmark (Figure 1B). By considering only the coding regions covered by the HCC panel, the median depths of WES were 114× (range, 92× to 345×) and 51× (range, 45× to 84×) in the fresh-frozen tumors and matched nontumor samples, respectively (Supplemental Table S3). WES analysis confirmed that no mutations were present within the targeted protein-coding regions in cases 3 and 12 and that case 9 was hypermutated (Figure 3B). Of the 62 mutations in the coding region identified from WES analysis, 61 (98%) were also called by the HCC panel analysis (Figure 3B). One *NRAS* Q61K hotspot mutation (case 6) was missed by using the HCC panel analysis. Manual review of this position revealed that the mutation had VAF of 2.5% by WES and 2.0% by the HCC panel (Supplemental Figure S2 and Supplemental Table S4). Note, however, that 2% is close to the detection limit of the current sequencing technologies.

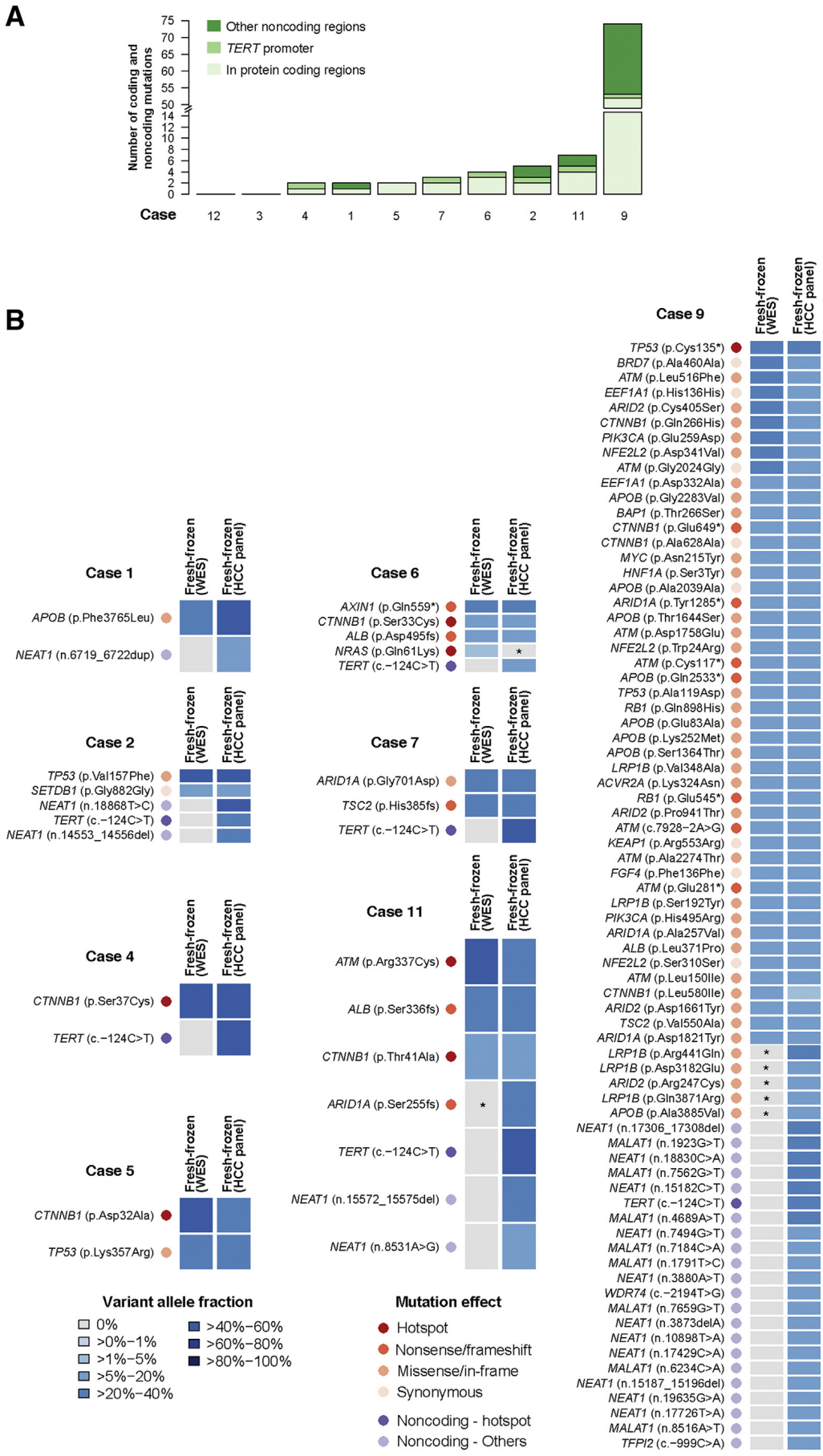
Compared with the WES analysis, the HCC panel analysis revealed an additional six mutations in the coding regions, including five in case 9 and one in case 11 (Figure 3B). Manual review of the WES data showed that all six mutations were in fact supported by at least one read in WES, but those positions were covered at reduced depth, with 4 of 6 covered by ≤40 reads (including three in *LRP1B*) and 5 of 6 ≤80 reads (Supplemental Figure S2C and Supplemental Table S4). This suggested that the increased sensitivity in the HCC panel analysis was likely due to the increased depth achieved.

Additional to the mutations in the protein-coding regions, the HCC panel also targeted the lncRNA genes *MALAT1* and *NEAT1* and the promoter regions of *TERT*, *WDR74*, *MED16*, and *TFPI2* (Figure 1A). Within these noncoding regions, an additional 32 mutations were identified across the 10 cases, representing a 48% gain of information compared with sequencing the protein-coding genes alone (Figure 3B). *TERT* promoter mutations were found in 60% (6 of 10) of cases and 16 somatic mutations in the lncRNA gene *NEAT1* were identified in 40% (4 of 10) of cases (Figure 3B and Supplemental Table S4).

Taken together, for the protein-coding genes frequently mutated in HCC, the HCC panel analysis produced highly reliable results compared with WES. Given the increased sequencing depth achieved by using the HCC panel, somatic mutations that were missed by WES were identified. Of importance, the HCC panel analysis enabled us to identify somatic mutations in promoter regions and frequently mutated lncRNA genes.

HCC Panel Analysis Identifies Somatic Mutations in FFPE Diagnostic Biopsies with Low-Input DNA

Nucleic acids from diagnostic specimens are frequently derived from small FFPE samples. Therefore, it would be important to determine whether the HCC panel could also be used for somatic mutational screening on low-input DNA


A
Number of coding and noncoding mutations
Other noncoding regions
TERT promoter
In protein coding regions
Case
12
3
4
1
5
7
6
2
11
9
B
Case 1
APOB (p.Phe3765Leu)
NEAT1 (n.6719_6722dup)
Case 2
TP53 (p.Val157Phe)
SETDB1 (p.Gly882Gly)
NEAT1 (n.18868T>C)
TERT (c.−124C>T)
NEAT1 (n.14553_14556del)
Case 4
CTNNB1 (p.Ser37Cys)
TERT (c.−124C>T)
Case 5
CTNNB1 (p.Asp32Ala)
TP53 (p.Lys357Arg)
Case 6
AXIN1 (p.Gln559*)
CTNNB1 (p.Ser33Cys)
ALB (p.Asp495fs)
NRAS (p.Gln61Lys)
TERT (c.−124C>T)
Case 7
ARID1A (p.Gly701Asp)
TSC2 (p.His385fs)
TERT (c.−124C>T)
Case 11
ATM (p.Arg337Cys)
ALB (p.Ser336fs)
CTNNB1 (p.Thr41Ala)
ARID1A (p.Ser255fs)
TERT (c.−124C>T)
NEAT1 (n.15572_15575del)
NEAT1 (n.8531A>G)
Case 9
TP53 (p.Cys135*)
BRD7 (p.Ala460Ala)
ATM (p.Leu516Phe)
EEF1A1 (p.His136His)
ARID2 (p.Cys405Ser)
CTNNB1 (p.Gln266His)
PIK3CA (p.Glu259Asp)
NFE2L2 (p.Asp341Val)
ATM (p.Gly2024Gly)
EEF1A1 (p.Asp332Ala)
APOB (p.Gly2283Val)
BAP1 (p.Thr266Ser)
CTNNB1 (p.Glu649*)
CTNNB1 (p.Ala628Ala)
MYC (p.Asn215Tyr)
HNF1A (p.Ser3Tyr)
APOB (p.Ala2039Ala)
ARID1A (p.Tyr1285*)
APOB (p.Thr1644Ser)
ATM (p.Asp1758Glu)
NFE2L2 (p.Trp24Arg)
ATM (p.Cys117*)
APOB (p.Gln2533*)
TP53 (p.Ala119Asp)
RB1 (p.Gln898His)
APOB (p.Glu83Ala)
APOB (p.Lys252Met)
APOB (p.Ser1364Thr)
LRP1B (p.Val348Ala)
ACVR2A (p.Lys324Asn)
RB1 (p.Glu545*)
ARID2 (p.Pro941Thr)
ATM (c.7928−2A>G)
KEAP1 (p.Arg553Arg)
ATM (p.Ala2274Thr)
FGF4 (p.Phe136Phe)
ATM (p.Glu281*)
LRP1B (p.Ser192Tyr)
PIK3CA (p.His495Arg)
ARID1A (p.Ala257Val)
ALB (p.Leu371Pro)
NFE2L2 (p.Ser310Ser)
ATM (p.Leu150Ile)
CTNNB1 (p.Leu580Ile)
ARID2 (p.Asp1661Tyr)
TSC2 (p.Val550Ala)
ARID1A (p.Asp1821Tyr)
LRP1B (p.Arg441Gln)
LRP1B (p.Asp3182Glu)
ARID2 (p.Arg247Cys)
LRP1B (p.Gln3871Arg)
APOB (p.Ala3885Val)
NEAT1 (n.17306_17308del)
MALAT1 (n.1923G>T)
NEAT1 (n.18830C>A)
MALAT1 (n.7562G>T)
NEAT1 (n.15182C>T)
TERT (c.−124C>T)
MALAT1 (n.4689A>T)
NEAT1 (n.7494G>T)
MALAT1 (n.7184C>A)
MALAT1 (n.1791T>C)
NEAT1 (n.3880A>T)
WDR74 (c.−2194T>G)
MALAT1 (n.7659G>T)
NEAT1 (n.3873delA)
NEAT1 (n.10898T>A)
NEAT1 (n.17429C>A)
MALAT1 (n.6234C>A)
NEAT1 (n.15187_15196del)
NEAT1 (n.19635G>A)
NEAT1 (n.17726T>A)
MALAT1 (n.8516A>T)
TFPI2 (c.−999C>A)
Fresh-frozen (WES)
Fresh-frozen (HCC panel)
Variant allele fraction
0%
>0%−1%
>1%−5%
>5%−20%
>20%−40%
>40%−60%
>60%−80%
>80%−100%
Mutation effect
Hotspot
Nonsense/frameshift
Missense/in-frame
Synonymous
Noncoding - hotspot
Noncoding - Others

(20 ng) extracted from FFPE samples. The DNA extracted from 36 diagnostic FFPE tumor biopsies was subjected to HCC panel sequencing to a median depth of 530× (range, 192× to 1257×) (Figures 1A and 2, B and C, and Supplemental Table S3). The median tumor content for these 36 cases was 90% (range, 5% to 100%) (Supplemental Table S2), thus representative of the distribution of tumor content in diagnostic samples in clinical practice. A median of three mutations (range, 0 to 76 mutations) per sample, including a median of two mutations (range, 0 to 53 mutations) in the coding regions was identified (Figure 4, Supplemental Figure S3, and Supplemental Table S4). No somatic mutations were identified for 8% (3 of 36) of cases (cases 7, 12, and 37), indicating that at least one somatic mutation could be detected in 92% of HCC diagnostic samples. Of note, although somatic mutations in the one biopsy with 5% tumor content could not be detected, somatic alterations in samples with 30% to 40% tumor content were detected.

The mutations identified in protein-coding genes from these 36 FFPE diagnostic biopsies were compared with those identified by WES of the DNA from the corresponding fresh-frozen biopsies. All 104 mutations identified from WES analysis were also called based on the HCC panel analysis (Figure 4 and Supplemental Figure S3), with 21 of 36 cases (58%) harboring *CTNNB1* mutations, a higher proportion than the TCGA and other HCC cohorts that was likely due to the higher percentage of alcohol-associated HCC (Supplemental Tables S1 and S2).[15] In addition, analysis of the HCC panel identified 18 mutations in the coding regions that were not found in the WES analysis in 11 cases. Of these 18 mutations, 13 were evident in WES but were not identified as mutations in the WES analysis, predominantly because of low sequencing depth (Supplemental Figures S2D and S3). The remaining five mutations were verified to be present in the corresponding FFPE samples but absent in the fresh-frozen samples by Sanger sequencing (Supplemental Figure S4 and Supplemental Table S4), indicating that they were genuine discordances between the fresh-frozen and FFPE DNA and not false positive calls from the HCC panel assay. Of note, two of five mutations validated to be absent from the fresh-frozen DNA affected mutation hotspots in *CTNNB1* (D32N and S45A) (Figure 4 and Supplemental Figure S4). The increased number of detected mutations by the HCC panel analysis was likely due to a combination of intratumor heterogeneity and the higher sequencing depth achieved.

Considering the 36 FFPE diagnostic biopsies, the HCC panel identified 70 somatic mutations in lncRNA genes and promoter regions, including 22 *TERT* promoter mutations (Figure 4 and Supplemental Table S4). Somatic mutations in lncRNA genes and promoter regions accounted for 37% of the total number of somatic mutations identified in the FFPE samples.

Compared with the high correlation of VAF between the sequencing platforms used in the fresh-frozen samples ($r = 0.89$, $r^2 = 0.79$, Pearson correlation), the correlation between WES from fresh-frozen samples and HCC panel by using FFPE samples was more modest ($r = 0.67$, $r^2 = 0.45$, Pearson correlation) (Supplemental Figure S2, A and B). Mutations with large deviations in VAFs between the sequencing platforms used in the fresh-frozen samples tended to be covered at reduced depths on either platform (Supplemental Figure S2C). Similar observations could be made between VAFs of exome (fresh-frozen) and HCC panel (FFPE) (Supplemental Figure S2D). The deviations in the latter may be more noticeable by the overall lower depth achieved in the FFPE samples than in the HCC panel sequencing of the fresh-frozen samples. Intratumor heterogeneity between the fresh-frozen and FFPE aliquots likely contributed to the reduced correlation.

Taken together these results suggested that the HCC panel analysis has high specificity and sensitivity in somatic mutation detection. Furthermore, somatic mutations in promoter regions (*TERT* promoter) and lncRNA genes (*MALAT1* and *NEAT1*) highly mutated in HCC could also be detected.

Copy Number Analysis of the HCC Panel Reveals High Concordance with WES

To determine whether the HCC panel could also be used to detect CNAs, 42 genes whose coding regions were entirely covered or were tiled across the lengths of the genes for CNA detection were evaluated (Figure 1A and Supplemental Table S1). Using the 41 nontumoral samples, the variability of the depth of coverage in the amplicons targeting the 42 genes was assessed (*Materials and Methods*). After removing amplicons with low depth of coverage or high variability, 1483 amplicons were used for CNA profiling. To assess the ability to detect per-gene CNA, each nontumoral sample was further paired with two other randomly selected, sex-matched nontumoral samples. The copy number $\log_2$ ratio of five genes, namely *LRP1B*, *ALB*, *BRD7*, *ACVR2A*, and *IRF2*, was variable (SD $>$ 0.3); therefore, these genes were excluded from further CNA analyses. Thirty-seven genes were included in the CNA analysis.

Figure 3 Comparison of somatic mutations defined by whole-exome sequencing (WES) and hepatocellular carcinoma (HCC) panel in fresh-frozen tissues. **A:** Number of coding and noncoding mutations per case identified in 10 fresh-frozen biopsies by using the HCC panel. **B:** Comparison of somatic coding and noncoding mutations found by WES and the HCC panel in the fresh-frozen samples. Heatmaps indicate the variant allele fractions of the somatic mutations (blue, see color key) or their absence (gray) in the eight cases in which at least one somatic mutation was identified. Mutation types are indicated as colored dots according to the color key. Mutations that were not called by mutation caller but were supported by at least one sequencing read are indicated by **asterisks**.

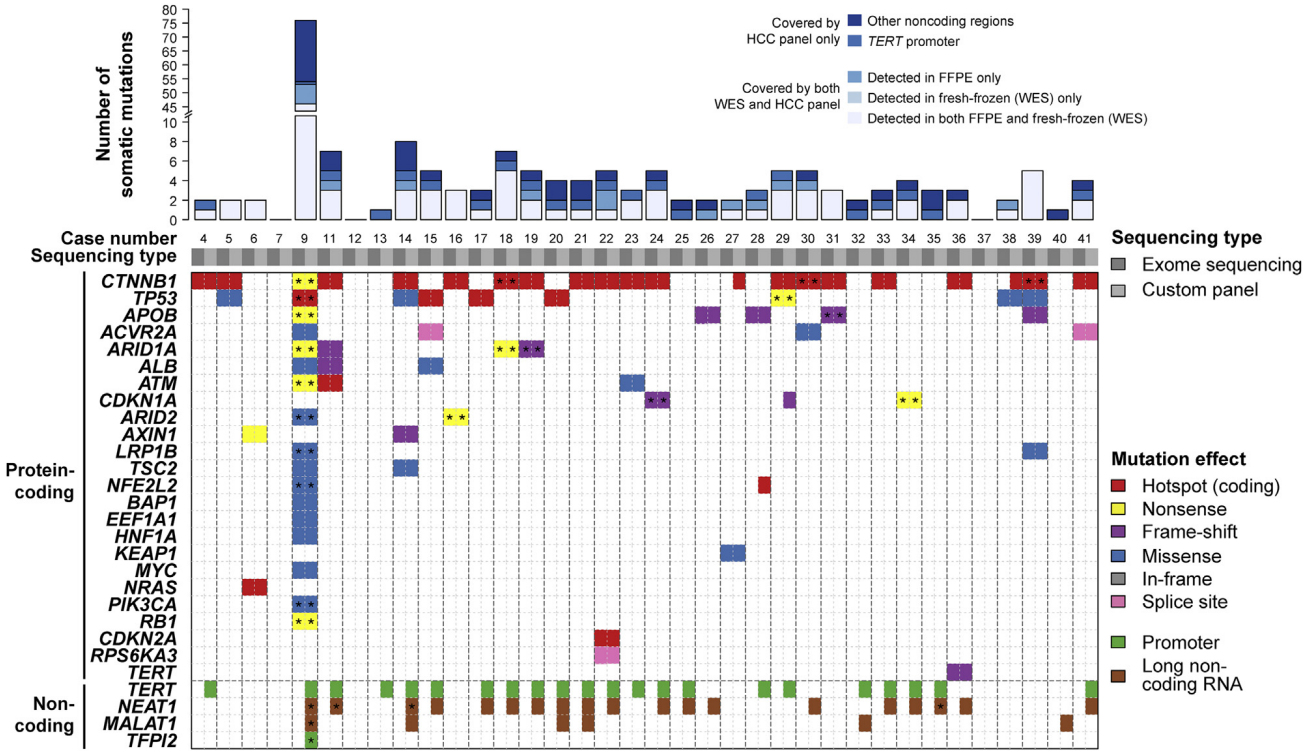


Figure 4 Comparison of somatic mutations defined by whole-exome sequencing (WES) and hepatocellular carcinoma (HCC) panel in formalin-fixed paraffin-embedded (FFPE) tissue. Barplot illustrates the number of somatic coding and noncoding mutations found in 36 FFPE tumor biopsies by using the HCC panel. In the main panel, each row represents a gene on the HCC panel and each column represents a sample. The mutations identified by WES in the fresh-frozen biopsies and those defined by sequencing the corresponding FFPE samples by using the HCC panel are placed next to each other. Mutation types are color coded according to the color key. The presence of multiple mutations in the same gene is illustrated by **asterisks**. Noncoding regions below the **dashed line** were not covered by WES.

The copy number profiles of matched fresh-frozen tumor/nontumor pairs and those derived from WES were compared. Of the 10 fresh-frozen pairs sequenced by using the HCC panel, one was excluded for excessive residual copy number $\log_2$ ratio (segment interquartile range, >0.8).[25] For the nine evaluable samples, a correlation of $r = 0.80$ ($r^2 = 0.64$) was found between the copy number $\log_2$ ratio of the two platforms (Figure 5A). When the copy number profiles of the 34 evaluable FFPE tumors were compared with the matched profiles from WES, a correlation of $r = 0.73$ ($r^2 = 0.54$) was observed between the copy number $\log_2$ ratios (Figure 5A). Overall, 86% of the evaluable genes had concordant copy number states (Figure 5B).

It has previously been reported that tumor purity had an impact on the ability to make CNA calls.[25,34] The impact of tumor purity on CNA analysis was therefore evaluated by using an *in silico* simulation on 12 cases (six fresh-frozen and six FFPE, selected on the basis of the presence of gene amplification/high gain or deep deletion), by replacing tumor reads with reads sampled from the normal samples to simulate tumor content 5%, 10%, 20% up to the actual tumor content for the samples. It was observed that amplifications/high gains were readily detected at 5% tumor content in many cases and at 20% in all cases (Supplemental Figure S5). In this cohort, deep deletions could not be detected at tumor content <40%.

Taken together, these results demonstrated that, despite profiling only a small number of genes, the HCC panel was able to detect CNAs in genes frequently gained or lost in HCC in both fresh-frozen and FFPE tumor samples with low-input DNA.

Discussion

HCC has a distinct mutational landscape compared with the major tumor entities. Numerous genes have been found to be mutated frequently in HCC but rarely in other tumors, such as those important for hepatocyte differentiation (*ALB*, *APOB*, *HNF1A*, *HNF4A*) and inflammatory response (*IL6R*, *IL6ST*). Given the relative rarity of HCC, these genes are currently not targeted or are only partially targeted in commercial panels [eg, Oncomine Comprehensive Panel version 3 (Thermo Fisher Scientific)] and in panels used by sequencing services [eg, FoundationOne assay (Foundation Medicine, Cambridge, MA)] (Supplemental Table S1). Thus, the currently available commercial assays for genomic profiling have suboptimal utility for HCC, and a targeted sequencing panel specifically designed for HCC is warranted.

In this study, we designed a custom Ion Torrent AmpliSeq sequencing panel, targeting all exons of 33 protein-coding genes, two lncRNA genes, promoter regions of four

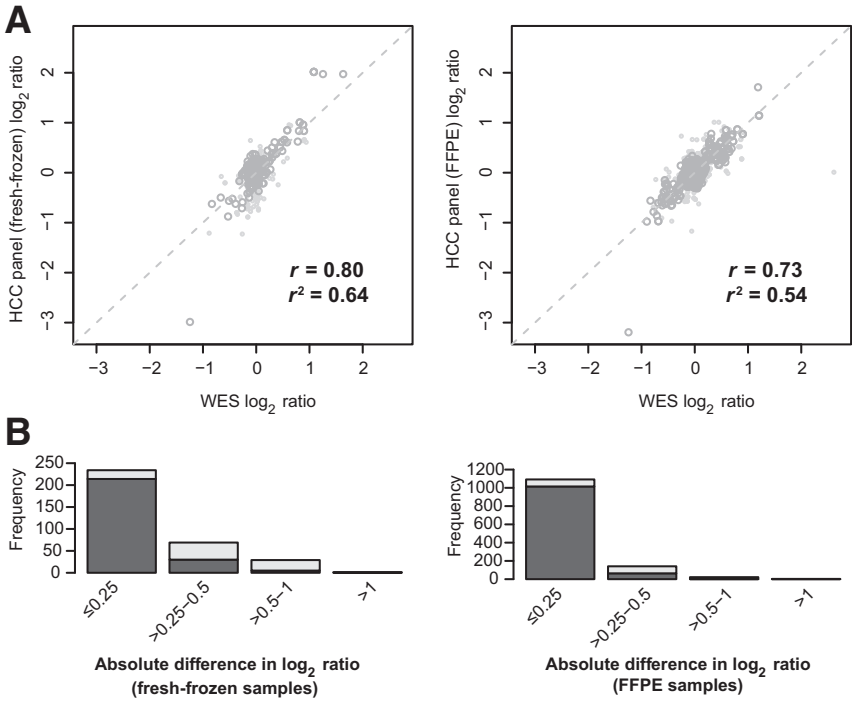


Figure 5 Copy number profiling by using the hepatocellular carcinoma (HCC) panel. **A:** Scatter plots illustrate the copy number $\log_2$ ratio of whole-exome sequencing (WES) and HCC panel sequencing of the fresh-frozen and the formalin-fixed, paraffin-embedded (FFPE) tumor samples. **B:** Barplots illustrate the number of genes with concordant (dark gray) or discordant (light gray) copy number states, binned by the absolute difference in copy number $\log_2$ ratio between WES and HCC panel sequencing of the fresh-frozen and FFPE samples.

genes previously found to be recurrently mutated in HCC, nine genes frequently affected by CNAs, and mutation hotspots in seven cancer genes.[7–17] Of importance, a number of the genes targeted by using the HCC panel are not currently on these two commercial panels. Of the 39 cases profiled with the HCC panel (including both fresh-frozen and FFPE samples), at least one somatic mutation was detected in 90% (35 of 39) of the cases. Of the mutations in coding genes found using this panel, 22% (42 of 189) would have been missed by both Oncomine Comprehensive Panel version 3 and the FoundationOne assay. In addition, recent whole-genome studies of HCC have revealed frequent mutations in lncRNA genes *NEAT1* and *MALAT1*, both of which are not currently targeted by commercial panels. In fact, it was found that approximately one-third of the mutations on the HCC panel were within the promoter and lncRNA regions.

Mutation screening and copy number profiling results from the HCC panel were benchmarked against those obtained from WES by the orthogonal Illumina sequencing technology. All but one mutation identified from WES were detected by using the HCC panel. An additional 10% to 15% of mutations within the coding regions were identified. Most of these additional mutations were in fact supported by few reads by WES; thus, the increased sensitivity was likely a direct result of the increased sequencing depth of both the tumor and the matched normal samples achieved. Crucially, however, evidence of intratumor genetic heterogeneity between adjacent fresh-frozen and FFPE biopsies, including two *CTNNB1* mutations, was found, suggesting that in these cases the *CTNNB1* mutations were not trunk mutations.

Although CNA detection using capture-based methods has been successful for targeted sequencing panel of several hundred genes,[35] CNA detection using amplicon-based targeted sequencing has proven more difficult. A recent study investigated the use of an amplicon-based sequencing strategy that targeted all exons of 113 genes related to DNA repair.[25] The researchers demonstrated that, with an appropriate analysis strategy and quality control, amplicon-based sequencing strategy is feasible and cost-effective for CNA profiling in FFPE samples.[25] In the present study, the strategy of computing and centering the $\log_2$ ratios for the primer two pools separately, before merging and segmentation proved to be an effective strategy in resolving issues associated with variable amplification efficiencies, with 86% of the genes showing concordant copy number states. Considering few studies have investigated the use of small targeted sequencing panel for CNA profiling, further benchmarking studies comparing analysis strategies and including larger sample size will likely improve the accuracies.

In the clinical setting, the quality, type, and amount of input materials for genomic profiling are crucial considerations, particularly in light of the smaller tumors being detected in screening programs. Here, we demonstrated that the HCC panel could be used for genomic screening with high sensitivity and specificity with low-input DNA (20 ng) derived from FFPE samples without compromising the results. Although based on an analysis of the TCGA HCC cases, 92% and 85% of the cases would have exhibited at least one nonsynonymous mutation by using the FoundationOne and the Oncomine assays, respectively, the HCC panel holds the advantage of much lower input requirement

than that required for commercial panels (eg, >40-μm tissue samples for the FoundationOne assay) and for capture-based targeted sequencing strategies.[35] In addition, somatic genetic alterations (somatic mutations and amplifications) could be detected from tumor samples with as low as 30% tumor content. Considering that mutations in the one sample with 5% tumor content could not be detected, 30% may be the lower limit of successful genomic profiling. Although lower limits (approximately 20%) have also been reported,[36] samples were not available to verify this. The samples included in this study are *de facto* samples obtained from routine diagnostic practice, and it was demonstrated that the low-input DNA requirement facilitates genomic profiling from small biopsies.

Driver genetic alterations have not yet become a tangible tool in clinical decision making for the treatment of HCC; thus, the immediate clinical application of our panel may be limited. However, recent studies have described the association of *TERT* promoter and *CTNNB1* exon 3 mutations with increased risk of malignant transformation of hepatocellular adenomas,[37,38] more frequent *HNF1A* and *IL6ST* mutations in hepatocellular adenomas than HCCs,[37] as well as *TP53* mutation as a poor prognostic indicator in HCC.[39–41] These associations suggest a potential utility of genomic profiling in prognostication for hepatocellular adenomas and HCCs, in tissues or even in cell-free DNA.[41,42] In terms of potential targetable alterations, three somatic mutations identified in our cohort of HCC are molecular targets in other cancer types according to OncoKB.[43] These include *ATM* loss of function mutation using olaparib in prostate cancer (level 4; biological evidence), *NRAS* hotspot mutation with binimetinib or in combination with ribociclib in melanoma (level 3; clinical evidence), and *TSC2* mutation with everolimus in central nervous system cancer (level 2; standard of care).[43] Application of our panel in clinical decision may become feasible in the future.

This study has several limitations. First, the targeted nature of the HCC panel means that copy number profiling is not genome-wide and is restricted to the genes included on the panel. Clinically, focal amplifications, compared with gains of chromosome arm, are more likely to be true driver genetic event and may be considered drug targets. The targeted nature of the HCC panel makes it difficult to distinguish the two scenarios. However, a re-analysis of the TCGA data suggests that high-level gains of chr11q13.3 (encompassing *CCND1*, *FGF19*, *FGF3*, *FGF4*) are almost always focal amplifications (>93%), whereas 50% to 70% of high-level gains of *TERT* and *VEGFA* are focal amplifications (Supplemental Table S5). By contrast, high-level gains of chr1q (*SETDB1* and *IL6R*) and chr8q (*NCOA2*, *MYC*, and *PTK2*) are frequently nonfocal (<10%), consistent with the frequent high-level gain of entire arms of chr1q and chr8q.[17] For deletions, most deep deletions are focal deletions, including all deletions (100%) in *ARID2*, *AXIN1*, *CDKN2A/B*, *PTEN*, and *TSC1/2*. These results suggest that CNAs affecting some of the most promising drug targets on the HCC panel are frequently true focal CNAs. Second, given that a median of two to three mutations per tumor were identified, tumor mutational burden, a putative biomarker for response to immune therapy, may not be accurately defined.[44] Third, the HCC panel does not include unique molecular identifiers, which would be useful to assess library complexity, particularly for samples with low-input DNA. We envisage that the addition of unique molecular identifiers would be particularly beneficial for the study of cell-free DNA from HCC patients.[41,42] Fourth, we designed the panel specific for HCC. Recent studies have revealed that mixed HCC/cholangiocarcinoma and cholangiocarcinoma have recurrent mutations in genes such as *IDH1/2*,[45] whereas *FRK* mutations decrease in frequency from hepatocellular adenoma to HCC.[37] These genes are not covered by the HCC panel. However, as an amplicon-based sequencing panel, adding amplicons to include genes that may assist in the differential diagnosis of HCC is straightforward.

Conclusion

This study demonstrated that the HCC panel is a cost-effective strategy for mutation screening and copy number profiling for routine diagnostic HCC samples with low-input DNA.

Acknowledgments

S.P., C.K.Y.N., and L.M.T. conceived and supervised the study; L.Q., M.S.M., S.P., C.K.Y.N., and L.M.T. performed literature search and designed the sequencing panel; S.W. and M.H.H. provided the samples and the whole-exome sequencing data; V.Pa., N.T., M.L., V.Pe., and S.P. performed DNA extraction and sequencing and prepared the library; A.G. and C.K.Y.N. developed the bioinformatics pipeline for mutation calling; V.Pa., A.G., S.P., C.K.Y.N., and L.M.T. analyzed the results and wrote the manuscript.

Supplemental Data

Supplemental material for this article can be found at *https://doi.org/10.1016/j.jmoldx.2018.07.003*.

References

1. Chin L, Andersen JN, Futreal PA: Cancer genomics: from discovery science to personalized medicine. Nat Med 2011, 17:297–303
2. Mok TS, Wu YL, Ahn MJ, Garassino MC, Kim HR, Ramalingam SS, Shepherd FA, He Y, Akamatsu H, Theelen WS, Lee CK, Sebastian M, Templeton A, Mann H, Marotti M, Ghiorghiu S, Papadimitrakopoulou VA; AURA3 Investigators: Osimertinib or platinum-pemetrexed in EGFR T790M-positive lung cancer. N Engl J Med 2017, 376:629–640
3. Toy W, Weir H, Razavi P, Lawson M, Goeppert AU, Mazzola AM, Smith A, Wilson J, Morrow C, Wong WL, De Stanchina E,

Carlson KE, Martin TS, Uddin S, Li Z, Fanning S, Katzenellenbogen JA, Greene G, Baselga J, Chandarlapaty S: Activating ESR1 mutations differentially affect the efficacy of ER antagonists. Cancer Discov 2017, 7:277–287

4. Kris MG, Johnson BE, Berry LD, Kwiatkowski DJ, Iafrate AJ, Wistuba II, Varella-Garcia M, Franklin WA, Aronson SL, Su PF, Shyr Y, Camidge DR, Sequist LV, Glisson BS, Khuri FR, Garon EB, Pao W, Rudin C, Schiller J, Haura EB, Socinski M, Shirai K, Chen H, Giaccone G, Ladanyi M, Kugler K, Minna JD, Bunn PA: Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA 2014, 311:1998–2006
5. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN, Brannon AR, O'Reilly C, Sadowska J, Casanova J, Yannes A, Hechtman JF, Yao J, Song W, Ross DS, Oultache A, Dogan S, Borsu L, Hameed M, Nafa K, Arcila ME, Ladanyi M, Berger MF: Memorial Sloan Kettering-Integrated Mutation Profiling Of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagn 2015, 17:251–264
6. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L: Mutational landscape and significance across 12 major cancer types. Nature 2013, 502:333–339
7. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, Clement B, Balabaud C, Chevet E, Laurent A, Couchy G, Letouze E, Calvo F, Zucman-Rossi J: Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet 2012, 44:694–698
8. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, et al: Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet 2012, 44:760–764
9. Cleary SP, Jeck WR, Zhao X, Chen K, Selitsky SR, Savich GL, Tan TX, Wu MC, Getz G, Lawrence MS, Parker JS, Li J, Powers S, Kim H, Fischer S, Guindi M, Ghanekar A, Chiang DY: Identification of driver genes in hepatocellular carcinoma by exome sequencing. Hepatology 2013, 58:1693–1702
10. Kan Z, Zheng H, Liu X, Li S, Barber TD, Gong Z, et al: Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res 2013, 23:1422–1433
11. Ahn SM, Jang SJ, Shim JH, Kim D, Hong SM, Sung CO, Baek D, Haq F, Ansari AA, Lee SY, Chun SM, Choi S, Choi HJ, Kim J, Kim S, Hwang S, Lee YJ, Lee JE, Jung WR, Jang HY, Yang E, Sung WK, Lee NP, Mao M, Lee C, Zucman-Rossi J, Yu E, Lee HC, Kong G: Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. Hepatology 2014, 60:1972–1982
12. Jhunjhunwala S, Jiang Z, Stawiski EW, Gnad F, Liu J, Mayba O, Du P, Diao J, Johnson S, Wong KF, Gao Z, Li Y, Wu TD, Kapadia SB, Modrusan Z, French DM, Luk JM, Seshagiri S, Zhang Z: Diverse modes of genomic alteration in hepatocellular carcinoma. Genome Biol 2014, 15:436
13. Totoki Y, Tatsuno K, Covington KR, Ueda H, Creighton CJ, Kato M, et al: Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat Genet 2014, 46:1267–1273
14. Shiraishi Y, Fujimoto A, Furuta M, Tanaka H, Chiba K, Boroevich KA, Abe T, Kawakami Y, Ueno M, Gotoh K, Ariizumi S, Shibuya T, Nakano K, Sasaki A, Maejima K, Kitada R, Hayami S, Shigekawa Y, Marubashi S, Yamada T, Kubo M, Ishikawa O, Aikata H, Arihiro K, Ohdan H, Yamamoto M, Yamaue H, Chayama K, Tsunoda T, Miyano S, Nakagawa H: Integrated analysis of whole genome and transcriptome sequencing reveals diverse transcriptomic aberrations driven by somatic genomic changes in liver cancers. PLoS One 2014, 9:e114263
15. Schulze K, Imbeaud S, Letouze E, Alexandrov LB, Calderaro J, Rebouissou S, Couchy G, Meiller C, Shinde J, Soysouvanh F, Calatayud AL, Pinyol R, Pelletier L, Balabaud C, Laurent A, Blanc JF, Mazzaferro V, Calvo F, Villanueva A, Nault JC, Bioulac-Sage P, Stratton MR, Llovet JM, Zucman-Rossi J: Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet 2015, 47:505–511
16. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, et al: Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet 2016, 48:500–509
17. Cancer Genome Atlas Research Network: Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 2017, 169:1327–1341.e23
18. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W: Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet 2014, 46:1160–1165
19. Ng CK, Piscuoglio S, Geyer FC, Burke KA, Pareja F, Eberle C, Lim R, Natrajan R, Riaz N, Mariani O, Norton L, Vincent-Salomon A, Wen YH, Weigelt B, Reis-Filho JS: The landscape of somatic genetic alterations in metaplastic breast carcinomas. Clin Cancer Res 2017, 23:3859–3870
20. Piscuoglio S, Ng CK, Murray MP, Guerini-Rocco E, Martelotto LG, Geyer FC, Bidard FC, Berman S, Fusco N, Sakr RA, Eberle CA, De Mattos-Arruda L, Macedo GS, Akram M, Baslan T, Hicks JB, King TA, Brogi E, Norton L, Weigelt B, Hudis CA, Reis-Filho JS: The genomic landscape of male breast cancers. Clin Cancer Res 2016, 22:4045–4056
21. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, Gao J, Socci ND, Solit DB, Olshen AB, Schultz N, Taylor BS: Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol 2016, 34:155–163
22. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, Zhang H, Solit DB, Taylor BS, Schultz N, Sander C: 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med 2017, 9:4
23. Thorvaldsdottir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013, 14:178–192
24. Talevich E, Shain AH, Botton T, Bastian BC: CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol 2016, 12:e1004873
25. Seed G, Yuan W, Mateo J, Carreira S, Bertan C, Lambros M, Boysen G, Ferraldeschi R, Miranda S, Figueiredo I, Riisnaes R, Crespo M, Rodrigues DN, Talevich E, Robinson DR, Kunju LP, Wu YM, Lonigro R, Sandhu S, Chinnayan A, de Bono JS: Gene copy number estimation from targeted next-generation sequencing of prostate cancer biopsies: analytic validation and clinical qualification. Clin Cancer Res 2017, 23:6070–6077
26. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012, 22:568–576
27. Olshen AB, Venkatraman ES, Lucito R, Wigler M: Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004, 5:557–572
28. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25:1754–1760
29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297–1303
30. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013, 31:213–219

31. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK: Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 2012, 28: 1811−1817
32. Shen R, Seshan VE: FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res 2016, 44:e131
33. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N: Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013, 6:pl1
34. Grasso C, Butler T, Rhodes K, Quist M, Neff TL, Moore S, Tomlins SA, Reinig E, Beadling C, Andersen M, Corless CL: Assessing copy number alterations in targeted, amplicon-based next-generation sequencing data. J Mol Diagn 2015, 17:53−63
35. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 2017, 23:703−713
36. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al: Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nat Biotechnol 2013, 31:1023−1031
37. Pilati C, Letouze E, Nault JC, Imbeaud S, Boulai A, Calderaro J, Poussin K, Franconi A, Couchy G, Morcrette G, Mallet M, Taouji S, Balabaud C, Terris B, Canal F, Paradis V, Scoazec JY, de Muret A, Guettier C, Bioulac-Sage P, Chevet E, Calvo F, Zucman-Rossi J: Genomic profiling of hepatocellular adenomas reveals recurrent FRK-activating mutations and the mechanisms of malignant transformation. Cancer Cell 2014, 25:428−441
38. Nault JC, Couchy G, Balabaud C, Morcrette G, Caruso S, Blanc JF, Bacq Y, Calderaro J, Paradis V, Ramos J, Scoazec JY, Gnemmi V, Sturm N, Guettier C, Fabre M, Savier E, Chiche L, Labrune P, Selves J, Wendum D, Pilati C, Laurent A, De Muret A, Le Bail B, Rebouissou S, Imbeaud S; GENTHEP Investigators, Bioulac-Sage P, Letouze E, Zucman-Rossi J: Molecular classification of hepatocellular adenoma associates with risk factors, bleeding, and malignant transformation. Gastroenterology 2017, 152:880−894.e6
39. Goossens N, Sun X, Hoshida Y: Molecular classification of hepatocellular carcinoma: potential therapeutic implications. Hepat Oncol 2015, 2:371−379
40. Desert R, Rohart F, Canal F, Sicard M, Desille M, Renaud S, Turlin B, Bellaud P, Perret C, Clement B, Le Cao KA, Musso O: Human hepatocellular carcinomas with a periportal phenotype have the lowest potential for early recurrence after curative resection. Hepatology 2017, 66:1502−1518
41. Kancherla V, Abdullazade S, Matter MS, Lanzafame M, Quagliata L, Roma G, Hoshida Y, Terracciano LM, Ng CKY, Piscuoglio S: Genomic analysis revealed new oncogenic signatures in TP53-mutant hepatocellular carcinoma. Front Genet 2018, 9:2
42. Ng CKY, Di Costanzo GG, Tosti N, Paradiso V, Coto-Llerena M, Roscigno G, Perrina V, Quintavalle C, Boldanova T, Wieland S, Marino-Marsilia G, Lanzafame M, Quagliata L, Condorelli G, Matter MS, Tortora R, Heim MH, Terracciano LM, Piscuoglio S: Genetic profiling using plasma-derived cell-free DNA in therapy-naive hepatocellular carcinoma patients: a pilot study. Ann Oncol 2018, 29:1286−1291
43. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al: OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017, 2017
44. Goodman AM, Kato S, Bazhenova L, Patel SP, Frampton GM, Miller V, Stephens PJ, Daniels GA, Kurzrock R: Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. Mol Cancer Ther 2017, 16:2598−2608
45. Farshidfar F, Zheng S, Gingras MC, Newton Y, Shih J, Robertson AG, et al: Integrative genomic analysis of cholangiocarcinoma identifies distinct IDH-mutant molecular profiles. Cell Rep 2017, 19:2878−2880

Genetic profiling using plasma-derived cell-free DNA in therapy-naïve hepatocellular carcinoma patients: a pilot study

Charlotte K.Y. Ng, Giovan G. Di Costanzo, Nadia Tosti, **Viola Paradiso**, Mairene Coto-Llerena, Giuseppina Roscigno, Valeria Perrina, Cristina Quintavalle, Tujana Boldanova, Stefan Wieland, Giuseppina Marino-Marsilia, Manuela Lanzafame, Luca Quagliata, Gerolama Condorelli, Matthias S. Matter, Raffaella Tortora, Marcus H. Heim, Luigi M. Terracciano, and Salvatore Piscuoglio

ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

Annals of Oncology 0: 1–6, 2018
doi:10.1093/annonc/mdy083
Published online 2 March 2018

ORIGINAL ARTICLE

Genetic profiling using plasma-derived cell-free DNA in therapy-naïve hepatocellular carcinoma patients: a pilot study

C. K. Y. Ng[1,2*], G. G. Di Costanzo[3], N. Tosti[1], V. Paradiso[1], M. Coto-Llerena[2], G. Roscigno[4], V. Perrina[1], C. Quintavalle[1], T. Boldanova[2,5], S. Wieland[2], G. Marino-Marsilia[6], M. Lanzafame[1], L. Quagliata[1], G. Condorelli[4], M. S. Matter[1], R. Tortora[3], M. H. Heim[2,5], L. M. Terracciano[1] & S. Piscuoglio[1*]

[1]Institute of Pathology, University Hospital Basel, Basel; [2]Hepatology Laboratory, Department of Biomedicine, University of Basel, Basel, Switzerland; [3]Department of Transplantation – Liver Unit, Cardarelli Hospital, Naples; [4]Department of Molecular Medicine and Medical Biotechnology, "Federico II" University of Naples, Naples, Italy; [5]Division of Gastroenterology and Hepatology, University Hospital Basel, Basel, Switzerland; [6]Pathology Unit, Cardarelli Hospital, Naples, Italy

Correspondence to: Dr Charlotte K. Y. Ng, Institute of Pathology, University Hospital Basel, Schoenbeinstrasse 40, 4031 Basel, Switzerland. Tel: +41-613286874; Fax: +41-612653194; E-mail: kiuyancharlotte.ng@usb.ch

Dr Salvatore Piscuoglio, Institute of Pathology, University Hospital Basel, Schoenbeinstrasse 40, 4031 Basel, Switzerland. Tel: +41-613286874; Fax: +41-612653194; E-mail: salvatore.piscuoglio@usb.ch

Background: Hepatocellular carcinomas (HCCs) are not routinely biopsied, resulting in a lack of tumor materials for molecular profiling. Here we sought to determine whether plasma-derived cell-free DNA (cfDNA) captures the genetic alterations of HCC in patients who have not undergone systemic therapy.

Patients and methods: Frozen biopsies from the primary tumor and plasma were synchronously collected from 30 prospectively recruited, systemic treatment-naïve HCC patients. Deep sequencing of the DNA from the biopsies, plasma-derived cfDNA and matched germline was carried out using a panel targeting 46 coding and non-coding genes frequently altered in HCCs.

Results: In 26/30 patients, at least one somatic mutation was detected in biopsy and/or cfDNA. Somatic mutations in HCC-associated genes were present in the cfDNA of 63% (19/30) of the patients and could be detected 'de novo' without prior knowledge of the mutations present in the biopsy in 27% (8/30) of the patients. Mutational load and the variant allele fraction of the mutations detected in the cfDNA positively correlated with tumor size and Edmondson grade. Crucially, among the seven patients in whom the largest tumor was ≥5 cm or was associated with metastasis, at least one mutation was detected 'de novo' in the cfDNA of 86% (6/7) of the cases. In these patients, cfDNA and tumor DNA captured 87% (80/92) and 95% (87/92) of the mutations, suggesting that cfDNA and tumor DNA captured similar proportions of somatic mutations.

Conclusion: In patients with high disease burden, the use of cfDNA for genetic profiling when biopsy is unavailable may be feasible. Our results support further investigations into the clinical utility of cfDNA in a larger cohort of patients.

Key words: hepatocellular carcinoma, cell-free DNA, circulating tumor DNA, somatic mutations, liquid biopsy, mutation screening

Introduction

The invasive nature of biopsy has prompted investigations into the use of plasma-derived cell-free DNA (cfDNA) as a potential minimally invasive surrogate for molecular profiling in several cancer types [1–4]. In contrast to most solid tumor types, hepatocellular carcinoma (HCC) diagnosis is frequently on the basis of radiology alone and in the absence of tumor biopsy. Therefore, nucleic acids for genetic profiling of HCC are typically obtained from tumor resection, a procedure that is only carried out in patients with limited, early-stage disease. In unresectable HCC patients, should the need for molecular profiling arises, the tumor materials would have to be collected in a non-routine invasive

procedure. The lack of routinely collected tumor materials is a hurdle for wider adoption of tumor profiling.

Studies have found that cfDNA concentration in serum or plasma of HCC patients is 3–4 times higher than in patients with chronic hepatitis and is up to 20 times higher than in healthy individuals [5–7]. Moreover, cfDNA concentration was found to be associated with tumor size, portal vein invasion and may be prognostic [5–8]. Molecular studies of circulating tumor DNA (ctDNA) in HCCs have investigated the size profiles of ctDNA [9], or were mutational studies of few cases, of resected materials, carried out at very low depth or investigated few mutation hotspots [10–14]. The use of resected materials, however, restricts molecular analyses to patients with early-stage, resectable disease. Given the correlation of cfDNA concentration and tumor size, one may speculate that patients with later stage disease would have higher mutational burden in cfDNA, as has been shown in other cancer types [4, 15].

Restricting molecular studies of ctDNA to mutation hotspots risks missing a substantial number of mutations, as most somatic mutations in HCC, even those in HCC-associated driver genes, do not fall into mutation hotspots [16–21]. Besides *TP53* (p53), *CTNNB1* (β-catenin) and *TERT* promoter, a wide range of HCC-associated driver genes and recurrently mutated promoter regions have been discovered, including those involved in chromatin remodeling (e.g. *ARID1A*, *ARID1B*, *ARID2*, *BAP1*), Wnt/β-catenin pathway (e.g. *AXIN1, FGF19*), and response to oxidative stress (e.g. *KEAP1*, *NFE2L2*) [16–21]. Additionally, long non-coding RNA genes (lncRNA, e.g. *NEAT1*, *MALAT1*) and promoter regions of *WDR74, TFPI2 and MED16* are also recurrently mutated [18, 19, 22].

In this exploratory study, we sought to determine whether somatic mutations in HCC driver genes can be detected with high confidence using next-generation sequencing in the plasma-derived cfDNA of HCC patients who have not undergone systemic therapy, and if the repertoire of mutations in the cfDNA is representative of the synchronously collected tumor biopsy. To address these questions, we prospectively recruited 30 HCC patients from whom we synchronously collected diagnostic core needle tumor biopsy and whole blood (supplementary Table S1, available at *Annals of Oncology* online) and carried out deep sequencing targeting HCC driver genes and mutation hotspots (supplementary Table S2, available at *Annals of Oncology* online).

Patients and methods

Patients

Thirty patients diagnosed with HCC at the University Hospital Basel, Basel, Switzerland or at Ospedale Cardarelli, Naples, Italy, were prospectively recruited for this study after written informed consent (supplementary Table S1, available at *Annals of Oncology* online). Patients who had previous systemic therapy for HCC were excluded. One patient was treated with radio-frequency thermal ablation 21 months before sample collection. From each patient undergoing diagnostic liver biopsy, two ultrasound-guided core needle biopsies of the primary tumor and whole blood were collected at diagnosis at the same time. Of the two primary tumor biopsies, one was processed and embedded in paraffin for clinical purposes and the other one was snap-frozen and stored at −80°C for research purposes. Ten millilitres of whole blood was collected in a 10 ml Cell-Free DNA Blood Collection Tube (BCT, Streck) and processed immediately (supplementary methods, available at *Annals of Oncology* online). Plasma was stored at −80°C until cfDNA extraction.

Tumor size, tumor location, macrovascular invasion, multifocality, and extrahepatic spread of each patient were assessed radiologically. Clinical staging of the patients was determined according to the Barcelona Clinic Liver Cancer (BCLC) staging system [23]. Sex of the patients, serum alpha-fetoprotein (AFP) levels, primary risk factors (hepatitis B/C virus infection, alcoholic liver disease, non-alcoholic fatty liver disease) were retrieved from clinical files. Histologic grading was carried out according to the 4-point scale Edmondson and Steiner system [24] (supplementary methods, available at *Annals of Oncology* online). Approval for the use of these samples has been granted by the ethics committee (Protocol Number EKNZ 2014-099).

Targeted sequencing and analysis

Tumor and germline DNA was extracted from fresh frozen biopsies and peripheral blood leukocytes ('buffy coat'). Circulating cfDNA was extracted from 3 to 6 ml of plasma (supplementary methods, available at *Annals of Oncology* online). DNA samples from the tumors, plasma-derived cfDNA and germline DNA were subjected to targeted sequencing using an Ampliseq panel targeting all exons of 33 liver cancer-associated protein-coding genes, all exons of the recurrently mutated lncRNA genes *MALAT1* and *NEAT1*, recurrently mutated promoter region of *TERT, WDR74, TFPI2* and *MED16*, as well as hotspots mutations in an additional seven cancer genes (supplementary Table S2, available at *Annals of Oncology* online). Sequencing was carried out on an Ion 530 chip using the Ion S5 XL system (Thermo Fisher Scientific, supplementary methods and Table S3, available at *Annals of Oncology* online). Sequencing data have been deposited in the Sequence Read Archive under the accession SRP115181.

Sequence reads were aligned to the human reference genome hg19 using TMAP. Somatic mutations were defined using Torrent Variant Caller (TVC) v5.0.3. We filtered out mutations supported by $\leq$8 reads, and/or those covered by $<$10 reads in the tumor/cfDNA or $<$10 reads in the matched germline. We only retained mutations for which the tumor variant allele fraction (VAF) was at least 10 times that of the matched normal VAF to ensure we kept only the somatic variants (supplementary methods, available at *Annals of Oncology* online). Due to the repetitive nature and the high GC content of the *TERT* promoter region, *TERT* mutation hotspots (chr5: 1295228 and chr5: 1295250) were additionally screened, and were considered present if supported by at least 5 reads or VAF of at least 5%. Mutations identified using the above steps are referred to as those found by 'de novo' methods.

To account for somatic mutations that may be present at low VAF in either the tumor biopsy or the matched cfDNA samples but not both, all somatic mutations identified using the 'de novo' methods in one of the two samples were interrogated for their presence in the matched sample by supplying TVC with their positions as the 'hotspot list'. Mutations supported by at least 2 reads were considered to be present. Mutations identified using the above steps are referred to as those found by 'interrogation'. Clinical actionability was assessed using OncoKB [25].

Statistical analysis

All statistical analyses were carried out in R v3.3.1. Correlations between the number of mutations, cfDNA concentrations and continuous/ordinal clinical variables (supplementary methods, available at *Annals of Oncology* online) were assessed using the Spearman's ρ. Comparisons of continuous/ordinal clinical variables between patients with and without somatic mutations in the cfDNA were carried out using Mann–Whitney *U* tests. Comparisons of categorical clinical variables and between patients with and without somatic mutations in the cfDNA were carried out using Fisher's exact tests. All statistical tests were two-tailed and $P < 0.05$ was considered statistically significant; 95% confidence intervals (CIs)

were estimated by leaving out 20% of the data points, computed over 100 runs.

Results

Of the 30 patients prospectively recruited into this study, 33% (10/30) had BCLC stages B/C/D disease (Table 1; supplementary Table S1, available at *Annals of Oncology* online). Multifocal and metastatic diseases were seen in 11 and 1 patients, respectively. Median diameter of the largest tumor was 34 mm (range 13–220 mm). At least one primary risk factor was identified for all patients (except HPU025 for whom the information is unavailable). Cirrhosis was seen in 87% (26/30) of the cases.

From each patient undergoing diagnostic liver biopsy, a core needle biopsy and whole blood were collected at the same time for targeted sequencing. A median of 94.6 ng (range 19.8–1710 ng) of plasma cfDNA was obtained from 10 ml of whole blood per patient (supplementary Table S1, available at *Annals of Oncology* online). We carried out deep sequencing of the HCC biopsies, cfDNA and matched germline using an in-house custom-made panel targeting 46 coding and lncRNA genes frequently altered in HCCs (median 1339× in biopsies and plasma, range 703–9385×, supplementary Tables S2 and S3, available at *Annals of Oncology* online). To mimic the potential use of plasma-derived cfDNA in the absence of available resected tumor material or a core needle biopsy in a clinical setting, we defined the somatic mutations for each HCC and cfDNA samples independently without prior knowledge of the repertoire of mutations present in the biopsy/cfDNA counterpart following a stringent set of analysis criteria (or 'de novo'). Additionally, to account for mutations that may be present at frequencies below the detection limit of the 'de novo' approach, we further examined the sequencing data of the biopsies for all mutations detected in the cfDNA (or 'by interrogation'), and *vice versa*. In 26/30 patients, at least one somatic mutation was detected in the biopsy and/or cfDNA (Figure 1 and supplementary Table S4, available at *Annals of Oncology* online).

Table 1. Clinicopathologic parameters of 30 therapy-naïve HCC included in this study

Age ($N = 30$)	Median years	72 (49–86)
Gender (N=30)	Female	10
	Male	20
BCLC classification (N=30)	A	20
	B	8
	C	1
	D	1
Associated with cirrhosis (N=30)	Yes	26
	No	4
Edmondson grade (N=28)	2	17
	3	7
	4	4
Largest tumor diameter (mm) (N=30)	Median (mm)	34 (13–220)
Macrovascular invasion (N=29)	Absent	28
	Present	1
Presence of metastasis (N=30)	Absent	29
	Present	1
Multifocal (N=30)	Absent	19
	Present	11
AFP (ng/ml) (N=29)	Median (ng/ml)	9 (1.6–7852)
Macrovascular invasion (N=29)	Absent	28
	Present	1
HBV (N=29)	Absent	27
	Present	2
HCV (N=29)	Absent	12
	Present	17
ALD (N=29)	Absent	19
	Present	10
NAFLD (N=29)	Absent	26
	Present	3
Prior treatment (N=30)	No	29
	Yes	1 (RFTA)

AFP, alpha-fetoprotein; ALD, alcoholic liver disease; BCLC, Barcelona Clinic Liver Cancer; HBV, hepatitis B virus; HCV, hepatitis C virus; NAFLD, non-alcoholic fatty liver disease; RFTA, radio-frequency thermal ablation.

Using the 'de novo' approach, we detected at least one somatic mutation in the cfDNA of 27% (8/30) of the patients (median 3, blue/gold bars, Figure 1). Considering the 7 non-hypermutator cases with at least one detectable mutation in the cfDNA, 81% (17/21) of the mutations detected in the cfDNA were also independently detected in the biopsy counterparts. In the hypermutator case (HPU207), 97% (64/66) of the mutations detected in the cfDNA were also independently detected in the biopsy counterpart (blue bars, Figure 1). All six apparently cfDNA-specific mutations were found to be present at low frequencies in their biopsy counterparts by interrogation (gold bars, Figure 1), suggesting that, in accordance with a recent study [13], cfDNA may be useful in overcoming intra-tumor genetic heterogeneity within the biopsies in therapy-naïve HCC patients. On the other hand, of all mutations detected in the non-hypermutator and the hypermutator cases, 78% (78/100) and 7% (5/71), respectively, were detected only in the HCC biopsies using the 'de novo' approach (dark/light red bars, Figure 1). However, 31% (24/78) and 100% (5/5) of these mutations could in fact be detected in the cfDNA by interrogation (dark red bars, Figure 1). Taken together, these results demonstrate that at least one somatic mutation can be detected in the cfDNA without prior knowledge of the repertoire of mutations in the HCC biopsies in 27% (8/30) of HCC patients and that at least one mutation was present, including those identified by interrogation, in 63% (19/30) of the cases.

Comparing the clinicopathologic parameters, we found that the 8 cases for whom at least one somatic mutation was detected in the cfDNA using the 'de novo' approach were associated with larger tumors (diameter of the largest tumor) and increasing Edmondson grade ($P = 0.012$ and $P = 0.010$, Mann–Whitney U tests, Figure 1; supplementary Table S5, available at *Annals of Oncology* online). Across all patients, the number of mutations detected 'de novo' in the cfDNA was positively correlated with the diameter of the largest tumor and Edmondson grade ($r = 0.482$, $P = 0.007$ and $r = 0.470$, $P = 0.012$, respectively, Spearman's ρ). The diameter of the largest tumor and Edmondson grade were also correlated with the maximum variant allele fractions of the mutations detected in the cfDNA ($r = 0.496$, $P = 0.005$ and $r = 0.502$, $P = 0.007$, respectively,

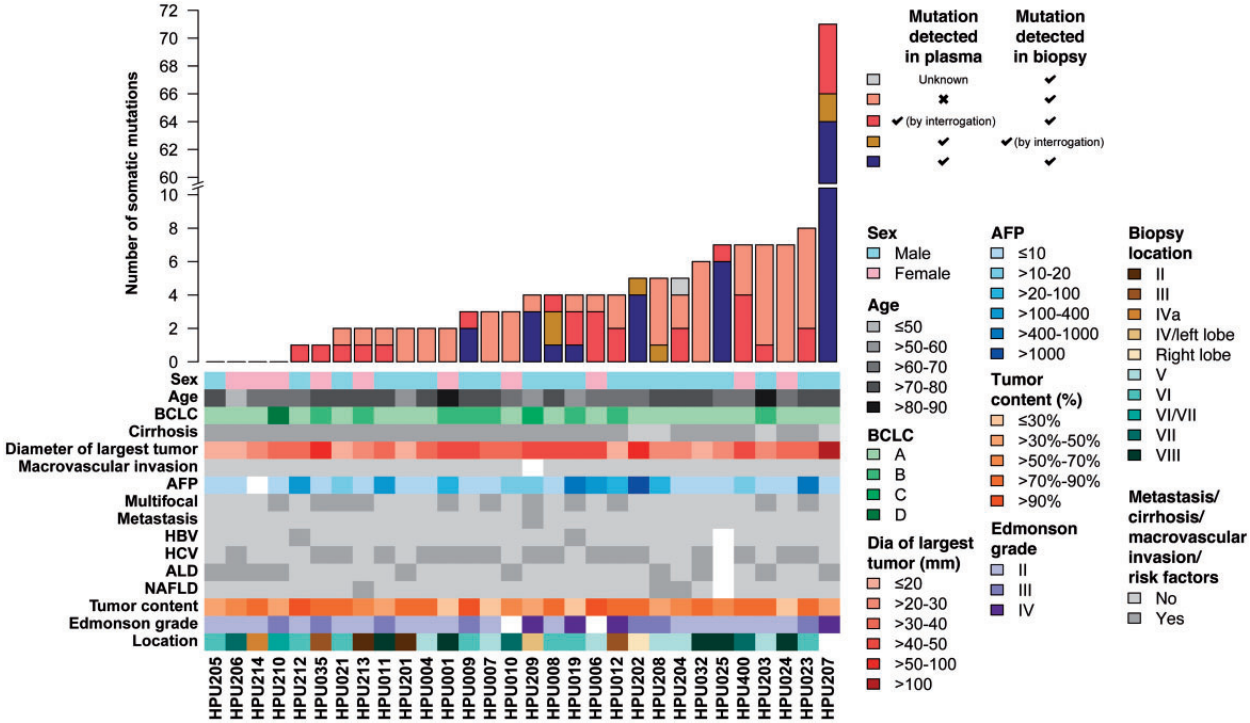


Figure 1. Number of somatic mutations detected in plasma-derived cell-free DNA and clinicopathologic information of the 30 patients with therapy-naïve hepatocellular carcinoma. The number of somatic mutations were categorized based on whether they were detected 'de novo' or 'by interrogation' (i.e. without or with prior knowledge of the repertoire of mutations in the biopsy/cfDNA counterpart, respectively, see color key). Clinicopathologic information is color-coded. White indicates unavailable information.

Spearman's ρ) and cfDNA concentration ($r = 0.889$, $P < 0.001$ and $r = 0.439$, $P = 0.020$, respectively, Spearman's ρ; supplementary Table S5, available at *Annals of Oncology* online). Additionally, at least one mutation was detected 'de novo' in the cfDNA in 40% (8/20) of male patients compared with 0% (0/10) of female patients, and in 75% (3/4) of HCCs not associated with cirrhosis compared with 19% (5/26) of HCCs with cirrhosis ($P = 0.029$ and $P = 0.048$, respectively, Fisher's exact tests; supplementary Table S5, available at *Annals of Oncology* online).

Among the seven cases in whom the largest tumor was ≥5 cm or was associated with metastasis, at least one mutation was detected 'de novo' in the cfDNA of 86% (6/7) of the cases, with a median of 75% (range 0%–100%) of the mutations detected in the cfDNA (Figure 2). Importantly, 87% (80/92, 95% CI 84% to 91%) of the mutations were detected 'de novo', and all but two remaining mutations could be detected by interrogation in the cfDNA counterparts. Conversely, 95% (87/92, 95% CI 93% to 97%) of the mutations were detected 'de novo' in the tumor biopsies, suggesting that mutation profiling of cfDNA in these patients captured similar proportion of mutations as tumor profiling would. By contrast, only 9% (7/78, 95% CI 6% to 11%) of the mutations were detected in the cfDNA of the remaining 23 patients with small (largest tumor ≤5 cm), non-metastatic HCC. These results suggest that in most HCC patients with high tumor burden, somatic mutations can be detected in the cfDNA with high confidence and that the repertoire of somatic mutations detected in cfDNA is representative of that in the primary HCC biopsy.

Discussion

HCC differs from most other tumor types in that biopsies are rarely carried out as they are usually not required for diagnosis. Thus, in patients not eligible for tumor resection (i.e. patients with large or metastatic disease and/or with poor performance status), tumor materials are usually unavailable for molecular profiling. Here we describe a prospective study to investigate the utility of cfDNA collected at the time of biopsy for molecular profiling in HCC patients. Targeting the most significantly mutated genes and regions in HCCs, we found that, even without the prior knowledge of the somatic mutations in the HCCs, high-depth sequencing analysis of plasma-derived cfDNA revealed that at least one somatic mutation in HCC driver genes can be detected in 27% (8/30) of therapy-naïve HCC patients. In an additional 11 cases, cfDNA captured mutations present below 'de novo' detection limit in the biopsies, demonstrating that somatic mutations were present in the cfDNA of 63% (19/30) of HCC patients at diagnosis. Importantly, among the patients with high disease burden (large tumor or metastasis) and most likely to be ineligible for resection, cfDNA profiling captured nearly as many mutations as primary tumor biopsy profiling alone. Of note, a *TSC2* frameshift mutation detected in the cfDNA and the primary tumor of the metastatic patient HPU209 is targetable by everolimus in cancers of the central nervous system as standard of care (supplementary Table S4, available at *Annals of Oncology* online). Taken together, our results demonstrate that the repertoire of mutations in HCC-associated genes identified in the cfDNA is representative of that in the biopsy.

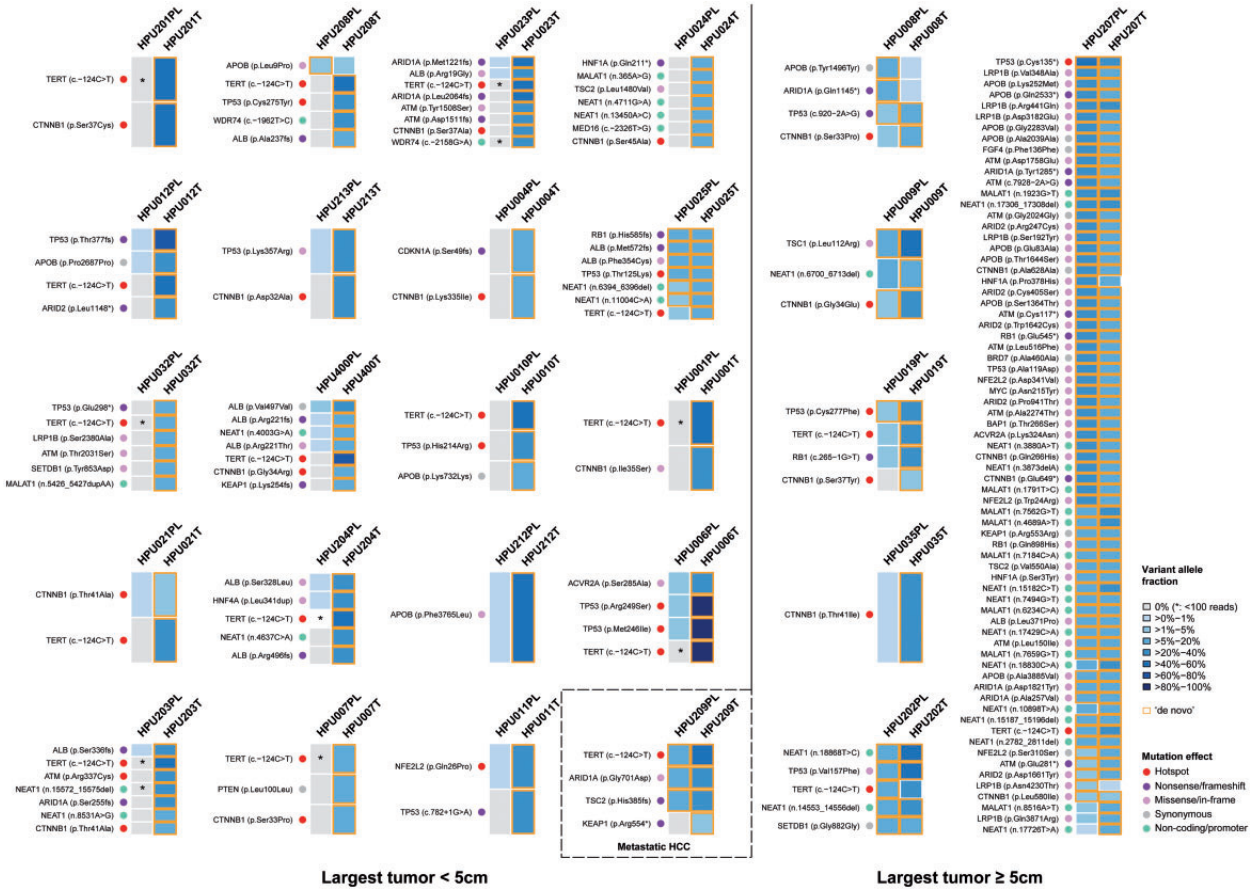


Figure 2. Somatic mutations found in cfDNA and in their primary tumor biopsies. Heatmaps indicate the variant allele fractions of the somatic mutations (blue, see color key) or their absence (grey) in the 26 pairs of tumor biopsy and cfDNA for which at least one somatic mutation was identified. Mutation types are indicated as colored dots. Orange boxes denote the mutations detected using the 'de novo' approach (i.e. without prior knowledge of the mutations in the biopsy/cfDNA counterpart). Mutations not detected by the 'de novo' approach but were covered by <100 reads are indicated by an asterisk. Cases are grouped according to the diameter of the largest tumor. T, tumor; PL, plasma.

Many HCC patients present with multifocal or metastatic disease and variable levels of heterogeneity with branched and parallel evolutionary patterns have been detected in HCC patients [13, 14]. Here we found a number of mutations that were detected with high confidence in the cfDNA but could only be detected by interrogation in the biopsy counterparts, reinforcing the notion that genetic analysis of a single diagnostic biopsy of the primary tumor may not be representative of the disease. Studies into the use of cfDNA as a minimally invasive surrogate for molecular profiling in HCC patients are therefore of particular clinical relevance.

Our study was limited in cohort size but as a proof of principle study and interpreted in the context of other tumor types [1–3], we found strong evidence that somatic mutations can be reliably detected in patients with high disease burden. As a prospective study, we have not assessed the prognostic significance of our findings. Furthermore, our filtering steps for the 'de novo' approach were deliberately stringent to closely recapitulate a potential clinical scenario. It is plausible that the limited sensitivity in detecting mutations 'de novo' in patients with low tumor burden is related to stringent filters. In fact, the number of mutations detected by interrogation suggests that advanced sequencing technologies incorporating molecular barcoding or alternative high-fidelity sequencing techniques will likely increase detection sensitivity in the clinical setting. Despite these limitations, the observed correlation of detectable somatic mutations and disease burden has important implications in the implementation of precision medicine [3]. Our results point toward the use of cfDNA for genetic profiling in HCC patients ineligible for resection and provide an argument for not subjecting patients with high disease burden to otherwise diagnostically unnecessary invasive procedure. Our results support further investigations into the clinical utility of cfDNA in a larger cohort of patients.

Funding

This work was supported by the Krebsliga beider Basel (KLbB-4183-03-2017 to CKYN). Additional financial support was provided by the Swiss Cancer League (Oncosuisse) (KLS-3639-02-2015 to LMT and KFS-3995-08-2016 to SP); the Swiss National Science Foundation (Ambizione PZ00P3_168165 to SP); the European Research Council (ERC Synergy Grant 609883 to MHH).

Disclosure

The authors have declared no conflicts of interest.

References

1. Dawson SJ, Tsui DW, Murtaza M et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med 2013; 368(13): 1199–1209.
2. Siravegna G, Mussolin B, Buscarino M et al. Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. Nat Med 2015; 21(7): 795–801.
3. Bidard FC, Weigelt B, Reis-Filho JS. Going with the flow: from circulating tumor cells to DNA. Sci Transl Med 2013; 5(207): 207ps14.
4. Bettegowda C, Sausen M, Leary RJ et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014; 6: 224ra224.
5. Huang Z, Hua D, Hu Y et al. Quantitation of plasma circulating DNA using quantitative PCR for the detection of hepatocellular carcinoma. Pathol Oncol Res 2012; 18(2): 271–276.
6. Tokuhisa Y, Iizuka N, Sakaida I et al. Circulating cell-free DNA as a predictive marker for distant metastasis of hepatitis C virus-related hepatocellular carcinoma. Br J Cancer 2007; 97(10): 1399–1403.
7. Yang YJ, Chen H, Huang P et al. Quantification of plasma *hTERT* DNA in hepatocellular carcinoma patients by quantitative fluorescent polymerase chain reaction. CIM 2011; 34(4): 238.
8. Iizuka N, Sakaida I, Moribe T et al. Elevated levels of circulating cell-free DNA in the blood of patients with hepatitis C virus-associated hepatocellular carcinoma. Anticancer Res 2006; 26: 4713–4719.
9. Jiang P, Chan CW, Chan KC et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci USA 2015; 112(11): E1317–E1325.
10. Chan KC, Jiang P, Zheng YW et al. Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. Clin Chem 2013; 59(1): 211–224.
11. Kirk GD, Lesi OA, Mendy M et al. 249(ser) *TP53* mutation in plasma DNA, hepatitis B viral infection, and risk of hepatocellular carcinoma. Oncogene 2005; 24(38): 5858–5867.
12. Liao W, Yang H, Xu H et al. Noninvasive detection of tumor-associated mutations from circulating cell-free DNA in hepatocellular carcinoma patients by targeted deep sequencing. Oncotarget 2016; 7(26): 40481–40490.
13. Huang A, Zhao X, Yang XR et al. Circumventing intratumoral heterogeneity to identify potential therapeutic targets in hepatocellular carcinoma. J Hepatol 2017; 67(2): 293–301.
14. Huang A, Zhang X, Zhou SL et al. Detecting circulating tumor DNA in hepatocellular carcinoma patients using droplet digital PCR is feasible and reflects intratumoral heterogeneity. J Cancer 2016; 7(13): 1907–1914.
15. Diaz LA Jr, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. J Clin Oncol 2014; 32(6): 579–586.
16. Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 2017; 169: 1327–1341 e1323.
17. Cleary SP, Jeck WR, Zhao X et al. Identification of driver genes in hepatocellular carcinoma by exome sequencing. Hepatology 2013; 58(5): 1693–1702.
18. Fujimoto A, Furuta M, Totoki Y et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet 2016; 48(5): 500–509.
19. Fujimoto A, Totoki Y, Abe T et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet 2012; 44(7): 760–764.
20. Guichard C, Amaddeo G, Imbeaud S et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet 2012; 44(6): 694–698.
21. Kan Z, Zheng H, Liu X et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. Genome Res 2013; 23(9): 1422–1433.
22. Weinhold N, Jacobsen A, Schultz N et al. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet 2014; 46(11): 1160–1165.
23. Llovet JM, Bru C, Bruix J. Prognosis of hepatocellular carcinoma: the BCLC staging classification. Semin Liver Dis 1999; 19(03): 329–338.
24. Edmondson HA, Steiner PE. Primary carcinoma of the liver: a study of 100 cases among 48, 900 necropsies. Cancer 1954; 7(3): 462–503.
25. Chakravarty D, Gao J, Phillips SM et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017; 1–16.

PipeIT: A Singularity Container for Molecular Diagnostic Somatic Variant Calling on the Ion Torrent Next-Generation Sequencing Platform

Andrea Garofoli*, **Viola Paradiso***, Hesam Montazeri, Philip M. Jermann, Guglielmo Roma, Luigi Tornillo, Luigi M. Terracciano, Salvatore Piscuoglio, and Charlotte K.Y. Ng

2019 JMD CME Program

The Journal of Molecular Diagnostics, Vol. 21, No. 5, September 2019

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

PipeIT

Check for updates

A Singularity Container for Molecular Diagnostic Somatic Variant Calling on the Ion Torrent Next-Generation Sequencing Platform

Andrea Garofoli,* Viola Paradiso,* Hesam Montazeri,*† Philip M. Jermann,* Guglielmo Roma,‡ Luigi Tornillo,*§ Luigi M. Terracciano,* Salvatore Piscuoglio,*¶ and Charlotte K.Y. Ng*‖

From the Institute of Pathology, University Hospital Basel, Basel, Switzerland; the Department of Bioinformatics,† Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran; the Department of Biology,‡ University of Naples Federico II, Naples, Italy; the GILAB AG,§ Allschwil, Switzerland; the Visceral Surgery Research Laboratory,¶ Clarunis, Department of Biomedicine, University of Basel, Basel, Switzerland; and the Department for Biomedical Research,‖ University of Bern, Bern, Switzerland*

CME Accreditation Statement: This activity ("JMD 2019 CME Program in Molecular Diagnostics") has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education (ACCME) through the joint providership of the American Society for Clinical Pathology (ASCP) and the American Society for Investigative Pathology (ASIP). ASCP is accredited by the ACCME to provide continuing medical education for physicians.

The ASCP designates this journal-based CME activity ("JMD 2019 CME Program in Molecular Diagnostics") for a maximum of 18.0 AMA PRA Category 1 Credit(s)™. Physicians should claim only credit commensurate with the extent of their participation in the activity.

CME Disclosures: The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Accepted for publication May 16, 2019.

Address correspondence to Charlotte K.Y. Ng, Ph.D., Department for Biomedical Research, University of Bern, Murtenstrasse 40, Bern 3008, Switzerland.
E-mail: charlotte.ng@dbmr.unibe.ch.

The accurate identification of somatic mutations has become a pivotal component of tumor profiling and precision medicine. In molecular diagnostics laboratories, somatic mutation analyses on the Ion Torrent sequencing platform are typically performed on the Ion Reporter platform, which requires extensive manual review of the results and lacks optimized analysis workflows for custom targeted sequencing panels. Alternative solutions that involve custom bioinformatics pipelines involve the sequential execution of software tools with numerous parameters, leading to poor reproducibility and portability. We describe PipeIT, a stand-alone Singularity container of a somatic mutation calling and filtering pipeline for matched tumor-normal Ion Torrent sequencing data. PipeIT is able to identify pathogenic variants in *BRAF*, *KRAS*, *PIK3CA*, *CTNNB1*, *TP53*, and other cancer genes that the clinical-grade Oncomine workflow identified. In addition, PipeIT analysis of tumor-normal paired data generated on a custom targeted sequencing panel achieved 100% positive predictive value and 99% sensitivity compared with the 68% to 80% positive predictive value and 92% to 96% sensitivity using the default tumor-normal paired Ion Reporter workflow, substantially reducing the need for manual curation of the results. PipeIT can be rapidly deployed to and ensures reproducible results in any laboratory and can be executed with a single command with minimal input files from the users. *(J Mol Diagn 2019, 21: 884—894; https://doi.org/10.1016/j.jmoldx.2019.05.001)*

The significant breakthrough in next-generation sequencing (NGS) of the last decade has provided an unprecedented opportunity to investigate human genetic variation and its role in health and disease. Spearheading these international, large-scale efforts are The Cancer Genome Atlas and the

Supported by Krebsliga beider Basel grant KLbB-4183-03-2017 (C.K.Y.N.), Swiss Cancer League grants KLS-3639-02-2015 (L.M.T.) and KFS-3995-08-2016 (S.P.), Swiss National Science Foundation grant PZ00P3_168165 (S.P.), and the Swiss Centre for Applied Human Toxicology (V.P.).

A.G. and V.P. contributed equally to this work.

Disclosures: None declared.

https://doi.org/10.1016/j.jmoldx.2019.05.001

International Cancer Genome Consortium. The efforts by these two consortia have led to a comprehensive molecular portrait of human cancers and their molecular pathogenesis.[1,2] Among the major findings is the unbiased discovery of genes mutated at rates significantly higher than the expected background level,[3] forming a significant group of the so-called driver genes. The discovery of these driver genes has provided the essential background knowledge for the design of cost-effective genomic assays that form the critical foundations of cancer diagnostics, therapeutics, clinical trial design, and selection of rational combination therapies. The accurate identification of somatic mutations has become a pivotal component of tumor profiling and precision medicine.

For tumor profiling in the research setting, the Illumina sequencing technology is by far the most commonly used. As a result, most of the research on error modeling, error correction, and the accurate calling of somatic mutations has been performed on the Illumina platform. There is a general consensus on the best practices for Illumina sequencing data analysis. In the diagnostic setting, however, the Ion Torrent technology is often used because of its relatively low costs, its fast turnaround time, and the availability of sequencing panels that require little DNA or RNA input. Ion Torrent sequencers are most frequently used for surveying cancer mutation hotspots and/or a limited number of cancer genes in molecular diagnostics laboratories. However, there is a lack of consensus on how to perform somatic mutation analysis for Ion Torrent data.[4,5]

A typical approach to perform somatic mutation calling on the Ion Torrent platform is through the proprietary browser-based Ion Reporter (IR) interface. The underlying variant calling engine of the IR is the Torrent Variant Caller (TVC), which generally achieves better specificity than tools not designed to consider the Ion Torrent–specific flow space.[4] However, the IR has several notable shortcomings. First, a recent comparison of variant calling methods reported that although the IR was the preferred solution, it suffered from an approximately 50% false-positive (FP) rate.[5] The high FP rate mandates lengthy and careful expert manual review of the results, thus introducing human-induced variability. Second, given the diversity in the landscape of somatic alterations among tumor types,[6] molecular diagnostics laboratories and researchers are increasingly creating customized targeted sequencing panels to address specific questions or tasks. However, IR analysis support for assays (ie, targeted sequencing panels and associated analysis procedures) other than the commercially released Ion Torrent assays is limited.

The importance of properly developed and maintained NGS bioinformatics pipelines in patient care cannot be understated.[7] NGS analysis pipelines typically involve the consecutive execution of tools.[8] Ensuring reproducible analyses and validating analysis pipelines would require the execution of multiple tools while locking down software versions and configurations.[7] In addition, many software tools have complex prerequisites (eg, the stand-alone version of TVC), adding time for software installation, maintenance, and testing to ensure compatibility. To ensure reproducibility and to ease software deployment, container technologies are being adopted by the bioinformatics community as prebuilt packages in which the necessary software is already installed, tested, and ready to be executed. In the context of NGS analysis pipelines in the diagnostic setting, container technology facilitates pipeline validation when transferred from one laboratory to another because a containerized pipeline gives the same results regardless of the hardware configurations and operating systems. Docker, firstly released in 2013, is the gold standard of container technologies, and today one can find Docker containers for many commonly used bioinformatics tools. For instance, the Genome Analysis Toolkit (GATK; Broad Institute, Cambridge, MA),[9] one of the most well-maintained NGS analysis packages, has been releasing Docker images since 2016. However, Docker images usually require root privileges to be executed, making them impractical for regular users in shared high-performance computing clusters. To overcome this limitation, Singularity[10] was created as an alternative for distributed environments.

We recently reported on a diagnostic targeted sequencing assay designed for hepatocellular carcinoma (HCC) with results benchmarked against whole-exome sequencing (WES) on an orthogonal sequencing platform.[11] We present the analysis pipeline as PipeIT, a Singularity container image that can be rapidly deployed and executed from end-to-end using a single command, from aligned Binary Alignment Map (BAM) files automatically generated by the Torrent Server to the final list of somatic mutations with high sensitivity and specificity.

Materials and Methods

Tissue Samples, Library Preparation, and Sequencing

Fifteen formalin-fixed, paraffin-embedded (FFPE) colon adenomas were obtained from the archive at the Institute of Pathology, University Hospital Basel, Basel, Switzerland. The adenoma tissue and matched germline control were microdissected separately from the same slide, and DNA was extracted as previously described.[12] DNA was quantified using the Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA). Approval for the use of these samples has been granted from the local ethics committee. Library preparation for the colon adenomas and their matched germline controls was performed using the Ion Torrent DNA Oncomine Comprehensive Panel v3M (Thermo Fisher Scientific) as previously described.[11,13] Quantification was performed using the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific), and sequencing was performed on an Ion S5XL system (Thermo Fisher Scientific).

Sequencing data for 10 frozen samples of HCC with matched germline sequenced using a custom AmpliSeq

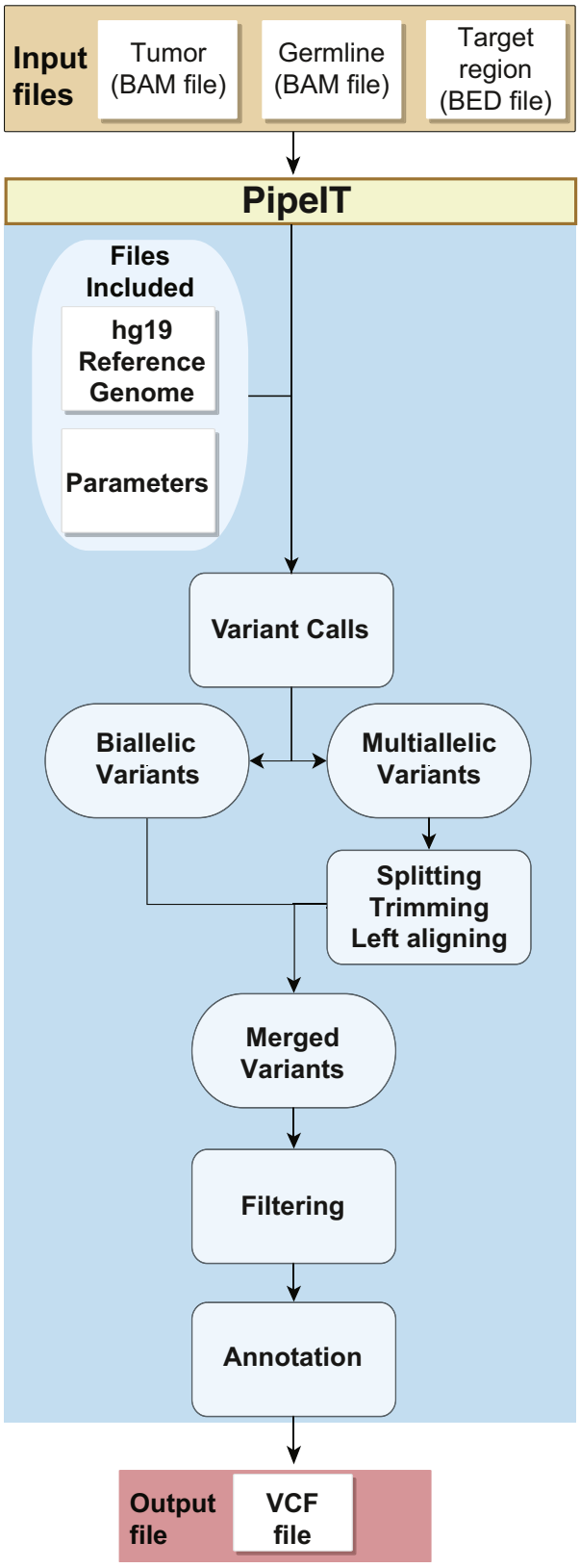


Figure 1 Overview of the PipeIT container. Flowchart showing the execution of PipeIT where the users need to provide only three files [Binary Alignment Map (BAM) files for tumor and normal samples and the target Browser Extensible Data (BED) file]. Variant calling is then performed using the Torrent Variant Caller with the packaged parameters file. The filtered and annotated mutations are then returned as output Variant Call Format (VCF) files.

targeted sequencing panel (Thermo Fisher Scientific) designed to focus on the most frequently altered genes in HCC were obtained from our previously published study.[11] The custom HCC panel includes 33 complete coding genes, two long noncoding RNA genes, four gene promoter regions, and mutation hotspots in seven genes, covering genomic regions of approximately 203 kb.[11] Sequencing was performed on an Ion S5XL system (Thermo Fisher Scientific). These samples had previously been subjected to WES using the SureSelectXT Clinical Research Exome (Agilent, Santa Clara, CA) platform and sequenced on an Illumina HiSeq2500 (Illumina, San Diego, CA).[11]

The PipeIT Workflow

As mandatory input files, BAM[14] files for the tumor and the matched germline samples, and a Browser Extensible Data (BED)[15] file specifying the target regions are required (Figure 1). The BAM files consist of sequencing reads aligned to the reference genome using the TMAP aligner and are generated as part of the standard automated data processing on the Torrent Server as sequencing data are generated. The BED file specifies the design of the targeted sequencing panel and comes with every panel design. A second BED file of the unmerged detailed version of the design BED file may be provided. If this file is not provided, PipeIT will create it automatically. The PipeIT workflow comprises the following steps: i) variant calling, ii) post-processing variants, iii) variant filtering, and iv) variant annotation (Figure 1).

The variant calling step (step 1) is performed using TVC version 5.0.3 (Thermo Fisher Scientific) as the variant calling engine, using a set of parameters modified from the set of default somatic, low-stringency parameters for AmpliSeq panels sequenced on the Personal Genome Machine (Thermo Fisher Scientific). Some of the most important modifications include a quality threshold of 6.5, a minimum variant score of 10, a minimum coverage of 8 and 15 for somatic nucleotide variants (SNVs) and small insertion/deletions (indels), respectively, and a minimum variant count of 4 and variant allele frequency (VAF) of 5% for long assembled indels. The modifications were made on the basis of the values recommended in IR. As with the original set of parameters for somatic analysis for AmpliSeq panels, SNVs, indels, and multinucleotide variants are reported, whereas complex variants are not reported. These parameters were used in a benchmarking study[11] and in another study in which variant detection was performed in cell-free DNA in patients with HCC.[13] The JSON file containing the benchmarked parameters is packaged within the container, but PipeIT also allows user-specified TVC parameters provided as a JSON file.

The postprocessing step (step 2) is performed to facilitate downstream filtering and annotation. This step is only required for multiallelic variants and consists of two parts. First, multiallelic variants are split into monoallelic variants

Table 1 Software Installed within the PipeIT Container, Including the Main Tools Used by the PipeIT Pipeline and the Dependencies Needed by the Main Tools

Main software	TVC version 5.0.3 (Thermo Fisher Scientific, Waltham, MA) BAMtools version 2.4.0 (*https://github.com/pezmaster31/bamtools*) SAMtools version 1.3.1 (*http://www.htslib.org*)[14] BCFtools version 1.5 (*http://www.htslib.org*) IGVtools version 2.3.60 (*https://software.broadinstitute.org/software/igv/igvtools*)[20] VCFtools version 0.1.14 (*https://vcftools.github.io*)[21] GATK version 3.6 (*https://software.broadinstitute.org/gatk*) SnpEff and SnpSift version 4.1l (*http://snpeff.sourceforge.net*)[19] HTSlib version 1.3.1 (*http://www.htslib.org*)[14]
Additional dependencies	armadillo, atlas, autoconf, automake, blas, boost, bzip2, cmake, epel, gcc, gcc-c++, git, igvtools, java, kernel-debug, lapack, libbz2, libopenblas, make, ncurses, openblas, unzip, wget, xz, zlib

using the BCFtools *norm* function. Second, each monoallelic variant is then left-aligned using the GATK *LeftAlignAndTrimVariants* tool. Multiallelic variants are therefore treated as individual monoallelic variants for downstream analysis and filtering. These postprocessing steps are particularly important for indels because TVC frequently reports several indels at a given locus, including ones that are not actually detected, within homopolymer or repeated regions.[16]

The variant filtering step (step 3) is implemented using *VariantFiltration* in GATK. Variants outside the target regions are removed. Hotspot variants[17,18] are then whitelisted. Variants covered by fewer than the specified number of reads (default to 10) in either the tumor or the matched normal sample or supported by fewer than the specified number of reads (default to 8) are removed. Furthermore, variants not likely to be somatic based on the ratio of VAF between tumor and normal (default to minimum 10:1) are also removed. PipeIT also allows user-specified values for the above filters. Given the clinical significance of many hotspot mutations, hotspot mutations were whitelisted even if they did not pass all read count and/or VAF filters. Reviewing the whitelisted hotspot variants that did not pass the above read count and/or VAF filters is recommended. Finally, variants passing the filters are annotated using the *ann* command of SnpEff[19] (step 4) using the canonical transcripts (defined as the longest protein coding transcript) from the genome version GRCh37.75. The final output is a Variant Call Format (VCF) file, with gene, transcript, and amino acid annotations in the *INFO* field.

Building the PipeIT Singularity Container Image

The PipeIT somatic variant detection workflow described above was implemented in a Singularity container[10] in the form of a compressed, read-only squashfs file system. Using a CentOS7 Docker image as a base, the software and tools required to execute the PipeIT workflow, including the standalone version of TVC and its dependencies for variant calling (step 1), BAMtools, SAMools, BCFtools, IGVtools, GATK, Tabix, and SnpSift for VCF file manipulation (steps 2 and 3), and SnpEff for variant annotation (step 4) (Figure 1 and Table 1) were installed and configured. The installation process defines the environment variables to ensure that the tools can be executed seamlessly. The JSON parameters file for TVC and the human hg19 reference genome compatible with the version used by the Torrent Server for alignment were added to the PipeIT container. Finally, a script that executes all the steps in the workflow described above was included to streamline the entire workflow into a single command.

Sequencing Data Analysis by the IR

Sequence reads were aligned to the human reference genome hg19 using TMAP within the Torrent Suite Software version 5.4 (Thermo Fisher Scientific) for the Ion S5XL system. Aligned BAM files were uploaded to the IR version 5.6 (Thermo Fisher Scientific) for analysis. For the analysis of the Oncomine Comprehensive Panel, the analysis was performed using the recommended workflow for the Oncomine Comprehensive Panel v3 (*Oncomine Comprehensive v3 - w3.1.1 - DNA - Single Sample* workflow, hereafter *IR-Oncomine*). Specifically, this tumor-only DNA analysis workflow uses the Oncomine Comprehensive DNA v3 Regions v1.0 file and the Oncomine Comprehensive DNA v3 Hotspots v1.0 file as target and hotspot regions, respectively, and hg19 as the reference genome. Default variant calling parameters, all annotation sets, no report template, and the Oncomine Variants v5.6 filter chain were used. The analysis was also performed using IR in a tumor-normal DNA analysis workflow (*IR-TN*), using the Oncomine Comprehensive DNA v3 Regions v1.0 file and the Oncomine Comprehensive DNA v3 Hotspots v1.1 file as target and hotspot regions, respectively, and hg19 as the reference genome. Furthermore, the default variant calling parameters, all annotation sets, the Default Variants View v5.6 filter chain, and no report template were used.

For the analysis of the HCC-targeted sequencing panel, two IR tumor-normal DNA analysis workflows were generated using the design BED file as the target regions and hg19 as the reference genome. In the first workflow (*IR-default*), the default variant calling parameters were used. In the second workflow (*IR-custom*), the set of custom parameters included in PipeIT (see above) was used and a

BED file that covered the mutation hotspots[17,18] within the target regions was included as the hotspot regions. For both workflows, all annotation sets, no report template, and the Default Variants View v5.6 filter chain were used.

The analyses for the 15 colon adenoma–normal pairs and the 10 HCC tumor-normal pairs were set up manually and sequentially. The filtered results of each analysis were downloaded as TSV files. For clarity, mutations not marked as *Non-confident* by the IR in tumor-normal DNA workflows (ie, IR-TN for the Comprehensive Cancer Panel and IR-default and IR-custom for the HCC targeted sequencing panel) were considered high confidence (HC) in this study.

Sanger Sequencing

To validate selected discordant variants among the mutation calling pipelines, Sanger sequencing was performed. Primer

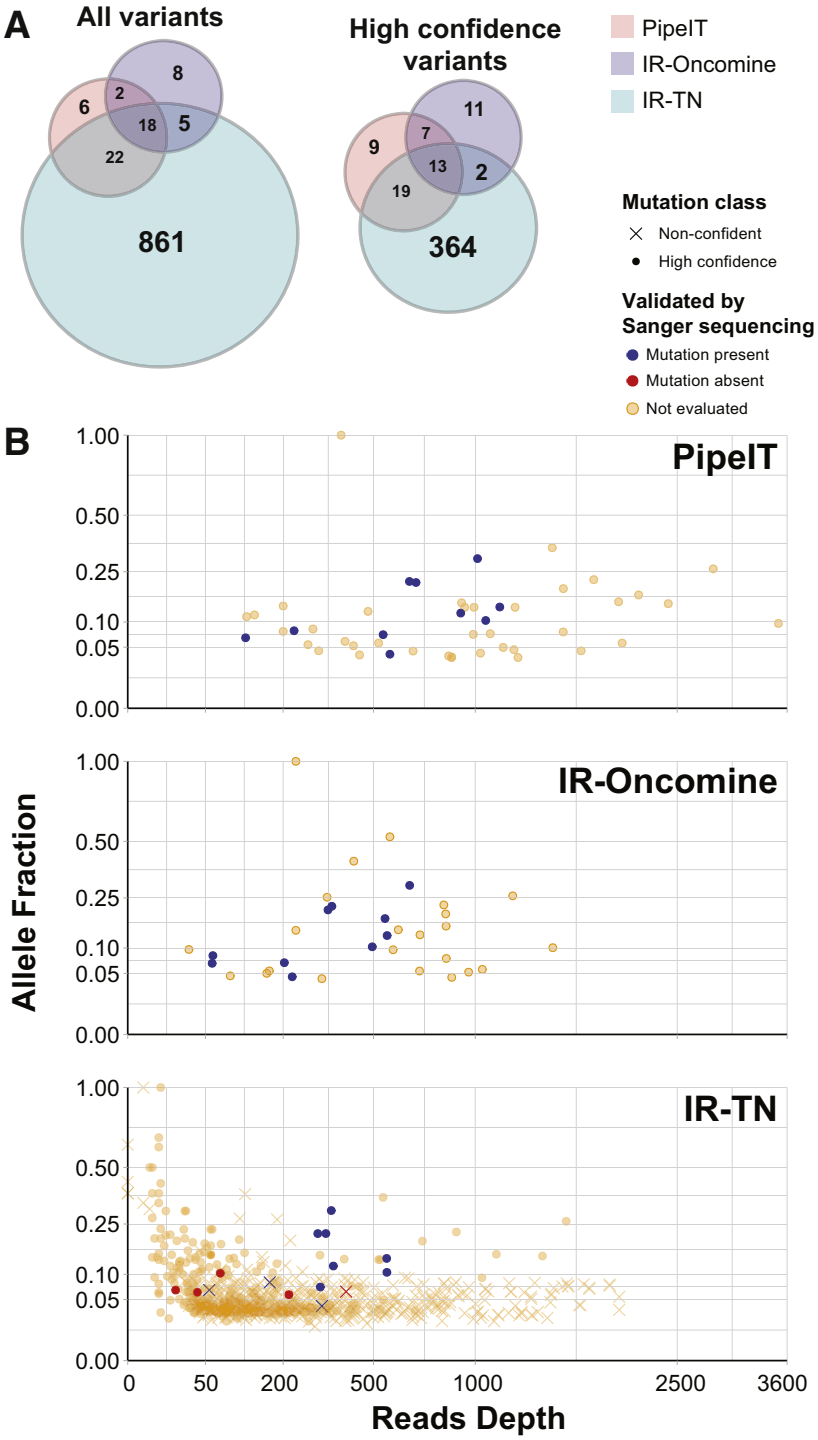


Figure 2 Comparison of mutation calls from PipeIT, Ion Reporter (IR) Oncomine Comprehensive Panel workflow (IR-Oncomine), and IR in a tumor-normal DNA analysis workflow (IR-TN) in 15 colon adenomas sequenced using the commercial Oncomine Comprehensive Panel v3. **A:** Venn diagrams showing the overlap of the mutation calls among PipeIT, IR-Oncomine, and IR-TN (**left panel**) and among PipeIT, IR-Oncomine, and IR-TN (high confidence; **right panel**). **B:** Scatterplots illustrating the variant allele fractions against read depth of the putative mutations identified by the three workflows. Mutations that were confirmed to be present by Sanger sequencing are colored in purple, and mutations that could not be confirmed by Sanger sequencing are colored in red. Mutations marked as non-confident by IR-TN analysis are indicated with crosses. Mutations in IR (both IR-Oncomine and IR-TN) appeared to have lower overall depth than PipeIT because the downsampling of the reads during variant calling.

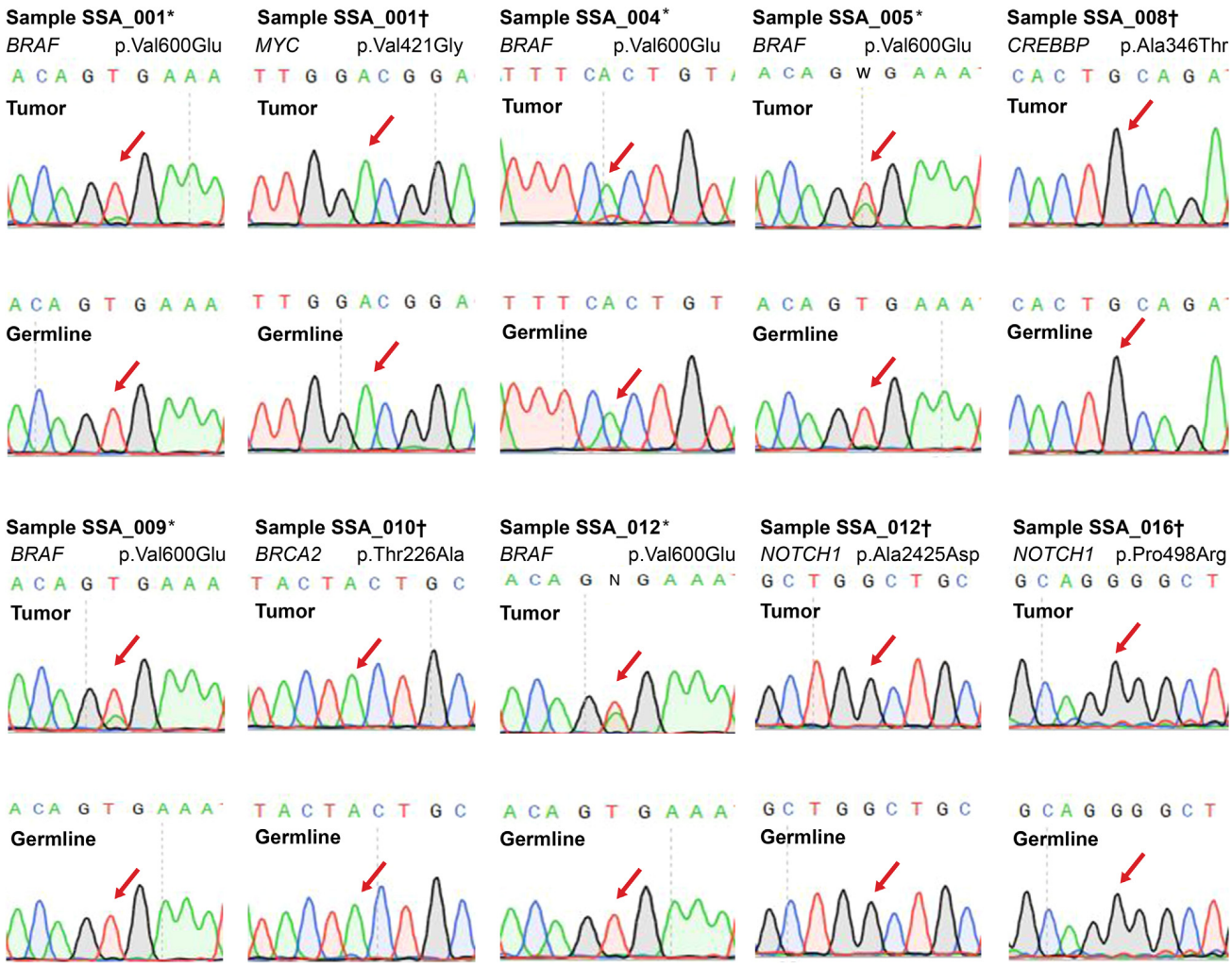


Figure 3 Validation of selected somatic mutations using Sanger sequencing. Sanger sequencing chromatograms of selected mutations being validated. **Red arrows** indicate the specific nucleotide investigated for the presence or absence of a specific somatic mutation. **Asterisks** indicate *BRAF* V600E mutations identified by PipeIT, Ion Reporter (IR) Oncomine Comprehensive Panel workflow (IR-Oncomine), and IR in a tumor-normal DNA analysis workflow (IR-TN). **Daggers** indicate putative mutations identified by IR-TN.

sets were designed as previously described[12] and reported in Table 2. PCR amplification of 5 ng of genomic DNA was performed with the AmpliTaq 360 Master Mix Kit (Thermo Fisher Scientific) on a Veriti Thermal Cycler (Thermo Fisher Scientific) as previously described.[12] PCR fragments were purified with ExoSAP-IT (Thermo Fisher Scientific). Sequencing reactions were performed on a 3500 Series Genetic Analyzer instrument by using the ABI BigDye Terminator chemistry version 3.1 (Thermo Fisher Scientific) according to the manufacturer's instructions. All analyses were performed in duplicate. Sequences of the forward and reverse strands were analyzed with SnapGene Viewer software version 4.0.2 (GSL Biotech LLC, Chicago, IL).

Evaluation of PipeIT and IR Results against WES

The somatic mutations identified in the HCC samples by PipeIT, IR-default, and IR-custom were compared with those identified by WES.[11] To account for the possibility that variants identified by PipeIT, IR-default, and IR-custom but not WES might have been detected but were not called in the WES analysis, discordant variants were reevaluated and interrogated for their presence in the WES data using the GATK version 3.6 UnifiedGenotyper by using the GENOTYPE_GIVEN_ALLELES mode. Mutations concordant with WES were considered true-positive (TP) results, mutations not found by WES were considered FP results, and mutations called in the WES analysis but not by PipeIT, IR-default, or IR-custom were considered false-negative (FN) results. Evaluation of performance of PipeIT and IR was then performed by computing positive predictive value (PPV, also known as precision), defined as TP/(TP + FP), and sensitivity, defined as TP/(TP + FN).

Software Availability

The PipeIT pipeline is freely available from Oncogenomics Laboratory (Basel, Switzerland; *http://oncogenomicslab.org/software-downloads*, last accessed December 12, 2018).

Table 2 Primer Sets Used to Perform Sanger Sequencing Validation of Selected Mutations

Gene	Mutation	Forward	Reverse
BRAF	p.Val600Glu	5′-AGCCTCAATTCTTACCATCCACA-3′	5′-ACTGTTTTCCTTTACTTACTACACCT-3′
RNF43	p.Thr20fs	5′-GGTCCATTTTCAAGGGGATCAC-3′	5′-ATGGTTGAAGTGCATTGCTG-3′
MYC	p.Val421Gly	5′-GTGACCAGATCCCGGAGTTG-3′	5′-CGCACAAGAGTTCCGTAGCT-3′
CREBBP	p.Ala346Thr	5′-GCTTGCTCTCGTCTCTGACA-3′	5′-CTTGGAACTCTGAGAGGTTAAAGT-3′
BRCA2	p.Thr226Ala	5′-TGCATTCTAGTGATAATATACAATACACA-3′	5′-TGTAAGATAAATAATTTAACAAGGCATTCC-3′
NOTCH1	p.Ala2425Asp	5′-GCTCTCCTGGGGCAGAATAG-3′	5′-CAGCAAACATCCAGCAGCAG-3′
NOTCH1	p.Pro498Arg	5′-GCCAGGGTGCAGACGACC-3′	5′-CCCTCACTGTTGCCCCAC-3′

Results

To streamline the somatic mutation analysis for matched tumor-germline DNA sequencing data generated on the Ion Torrent platform, PipeIT was built, implementing the workflow previously used in our diagnostic HCC assay (Figure 1).[11] This workflow has been benchmarked in samples sequenced from approximately 200× to approximately 1600× depth in both fresh-frozen and FFPE samples against results from WES on an orthogonal sequencing platform and were shown to be highly concordant.[11]

PipeIT was built as a Singularity container image that can be executed in a single command, eliminating the need for the individual execution of variant calling, postprocessing steps, filtering, and annotation (Figure 1). Importantly, as a container image, PipeIT is easily portable to any laboratory and always produces the same results. To execute the complete somatic mutation calling workflow of PipeIT, the single command *singularity run PipeIT.img -t path/to/tumor.bam -n path/to/normal.bam -e path/to/region.bed* is needed. Additional optional parameters, such as TVC parameters and thresholds for variant filtering, may also be specified, allowing individual laboratories to customize their own analyses. Since Singularity is high-performance computing compatible, PipeIT can be used to execute many analyses in parallel without cumbersome and labor-intensive analysis setup.

PipeIT was tested on 15 colon adenomas and 10 HCCs. The 15 colon adenomas consisted of adenoma-normal pairs of FFPE colon adenoma sequenced using the Oncomine Comprehensive Panel v3, covering approximately 349 kb to a median depth of 569× (range, 301× to 834×), whereas the 10 HCCs consisted of the previously published 10 tumor-normal pairs of fresh-frozen HCCs sequenced using a custom HCC targeted sequencing panel covering approximately 203 kb to a median depth of 1495× (range, 1026× to 1855×).[11] On a machine with Intel Xeon 2.6 Hz processor with four threads and 32 GB of memory, the PipeIT analysis took a mean of approximately 15 minutes for each colon adenoma-normal pair and a mean of approximately 45 minutes for each HCC tumor-normal pair.

PipeIT Identifies Pathogenic Somatic Mutations on the Commercial Oncomine Comprehensive Panel

To compare PipeIT to the IR, the routinely used interface for mutation calling in clinical diagnostic laboratories, PipeIT was first evaluated on the 15 colon adenomas sequenced on the commercially available Oncomine Comprehensive Panel v3. PipeIT was first compared to the IR out-of-the-box tumor-only workflow optimized for the Oncomine Comprehensive Panel v3 for diagnostic use (*IR-Oncomine*). PipeIT and IR-Oncomine identified 48 and 33 mutations, respectively (Figure 2A), with a median of 3 (range, 1 to 6 somatic mutations) and 2 (range, 0 to 5 somatic mutations) per sample, respectively. Both PipeIT and IR-Oncomine identified bona fide pathogenic variants, including *BRAF* V600E mutation in 10 cases, all of which were confirmed by Sanger sequencing (Figure 3, Table 2, and Supplemental Figure S1A). Furthermore, PipeIT identified the additional pathogenic variants that IR-Oncomine found, including *NRAS* Q61K, *KRAS* G12C, and Q61K; *PIK3CA* C420R, *CTNNB1* T41A, and S45A; and *TP53* C275Y, *ARID1A* Y815fs, and *CDKN1B* R152fs mutations (Supplemental Table S1). Of note, two of the *BRAF* V600E mutations were flagged for review by PipeIT because of the presence of small number of variant reads (ie, the presence of some adenoma cells) in the matched germline samples (Table 2 and Supplemental Figure S1A).

On the other hand, 13 variants were found only by IR-Oncomine but not PipeIT (Figure 2A). On inspection of the variants and the sequence reads, two were found to be germline heterozygous variants (*BRCA2* K3326* and *MET* R988C, both with minor allele frequency >0.1% in the general population),[22] nine were present at low VAF in matched tumor and normal samples, and one was in a poorly aligned region, highlighting the advantage of performing matched tumor-normal analysis as opposed to tumor-only analysis in removing germline variants and systematic artifacts. Lastly, an *RNF43* large frameshift deletion found only by IR-Oncomine but not PipeIT was shown to be FP by Sanger sequencing (Table 2 and Supplemental Figure S1B). However, the tumor-only IR-Oncomine workflow only conservatively reports the subset of mutations cataloged in its internal database as likely somatic, therefore likely

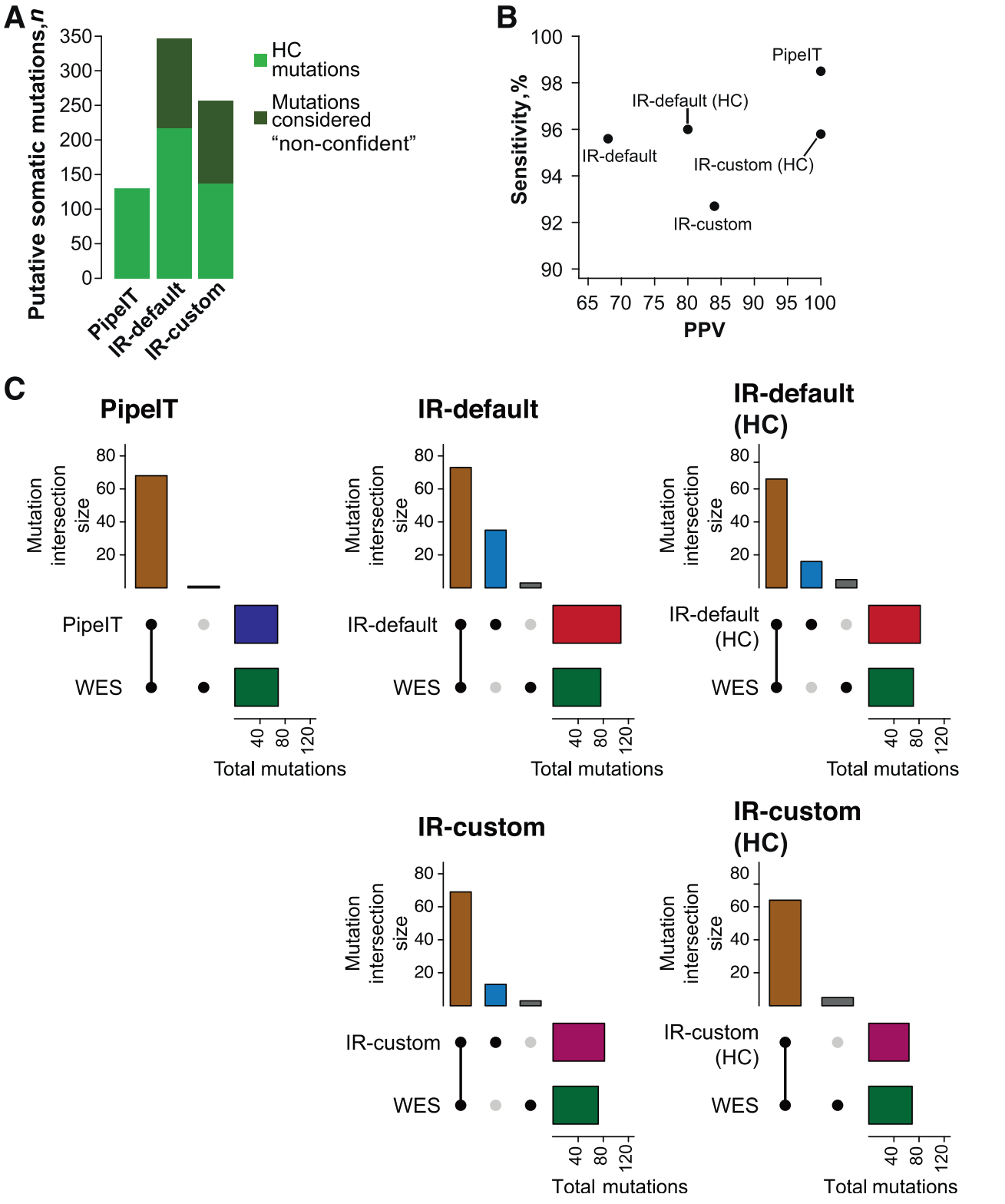


Figure 4 Comparison of mutation calls from PipeIT, IR analysis using the default parameters (IR-default), and IR analysis using the custom set of variant calling parameters used in PipeIT (IR-custom) in 10 hepatocellular carcinomas sequenced using a custom AmpliSeq panel. **A:** Bar plot shows the number of putative somatic mutations (including all protein-coding and noncoding mutations). **B:** The positive predictive values and the sensitivity are plotted for each analysis pipelines. **C:** UpSet[21] plots show the number of protein-coding and splice site mutations identified by each of PipeIT, IR-default, or IR-custom compared with whole-exome sequencing (WES). Vertical bars represent, from the leftmost to the rightmost, the numbers of mutations at the intersection, the ones called by PipeIT, IR-default or IR-custom only, and the ones called in the WES only. Horizontal bars represent the total number of mutations called by PipeIT, IR-default, or IR-custom or in WES. For IR-default and IR-custom, two plots were made, one with all the variants called and one with the subset of high-confidence (HC) mutations.

omitting genuine but rare somatic variants, in particular those in tumor suppressor genes. For instance, an *RNF43* splice site mutation at 35% VAF was reported by PipeIT but not IR-Oncomine.

Given that the tumor-only IR-Oncomine identified a number of germline variants and false variants that could have been removed using a tumor-normal approach, PipeIT was further evaluated against an IR matched tumor-normal (*IR-TN*) workflow. The IR-TN workflow identified 906 (of which 398 were HC) mutations, with a median of 52 (range, 36 to 114; or median, 18; range, 6 to 65 for HC mutations) mutations per sample (Figure 2A and Supplemental Table S1). IR-TN identified 861 putative variants (or 364 counting only HC variants) that were not detected by PipeIT or IR-Oncomine (Figure 2A). Five of these IR-TN–specific mutations with VAF >5% were randomly selected for validation by Sanger sequencing, including four mutations that were considered to be HC. All five mutations in *MYC*, *CREBBP*, *BRCA2*, and *NOTCH1* were absent by Sanger sequencing (Figure 3 and Table 2), indicating that these were not variants that were missed by PipeIT. Compared with PipeIT and IR-Oncomine, IR-TN identified many more mutations with low VAF and/or low depth (Figure 2B). Many of the IR-TN variants were flagged as non-confident, primarily because they were detected at low VAF in both the tumor and the corresponding normal samples. Among the 508 non-confident variants called by IR-TN were three *BRAF* V600E mutations, highlighting the need for careful manual curation of the non-confident IR-TN results. Taken together, these results indicate that PipeIT was able to identify pathogenic variants that were detected using the IR-Oncomine as would have been done in the diagnostic setting.

PipeIT Accurately Identifies Somatic Mutations on a Custom AmpliSeq Panel

For custom sequencing panels, IR does not provide optimized analysis workflows. For the 10 HCCs sequenced on a custom AmpliSeq panel, PipeIT was first evaluated against an IR tumor-normal analysis workflow using default parameters and default variants filter chain (*IR-default*) (*Materials and Methods*). In addition, a second workflow that used the custom variant calling parameter set used in PipeIT and hotspot regions[17,18] curated from the literature (*IR-custom*) was generated to mimic the setup of PipeIT. Across the 10 HCCs, PipeIT, IR-default, and IR-custom identified 139, 346 (of which 217 were HC), and 256 (of which 137 were HC) somatic mutations, respectively, with 134 (128 counting only HC mutations from IR-default and IR-custom) identified by all three analyses (Figure 4A,[23] Supplemental Figure S2, and Supplemental Table S2). PipeIT, IR-default, and IR-custom identified a median of 2.5 (range, 0 to 112), 25 (range, 16 to 137; or median, 11.5; range, 6 to 117 for HC mutations), and 13.5 (range, 7 to 130; or median, 3; range, 0 to 109 for HC mutations) somatic mutations, respectively, per sample. As previously reported,[11] one of the cases displayed a hypermutator phenotype with >50% of all mutations coming from this single case (HPU207T). IR (IR-default and IR-custom) did not appear to recognize the noncoding gene *NEAT1* (Supplemental Table S2).

The exonic mutations (in protein coding genes, including splice site mutations) calls obtained from PipeIT, IR-default, and IR-custom were compared with those obtained from WES on the Illumina platform.[11] All 68 exonic mutations identified by PipeIT were confirmed to be present and somatic by WES, giving a PPV of 100% (Figure 4, B and C), including two with <5% VAF and 14 with <10% VAF. Compared with PipeIT, IR-default identified more putative exonic mutations ($n = 108$, of which 82 were HC) but with a far inferior PPV (68%, or 80% counting only HC variants). On the other hand, IR-custom identified 82 (of which 64 were HC) but had a PPV more similar to PipeIT (84%, or 100% counting only HC variants). Compared with the variability in PPV among the various workflows, all workflows achieved >92% sensitivity, with 99% sensitivity for PipeIT outperforming all other workflows (93% to 96%) (Figure 4B).

Taken together, benchmarked against the mutations identified from WES and compared with the IR, PipeIT identified more known mutations while maintaining excellent PPV, including mutations at low VAF. Of note, customizing the variant calling parameters alone in the IR (as in IR-custom) raised the PPV substantially compared with the default tumor-normal DNA analysis parameters in IR-default.

Discussion

Modern clinical molecular diagnostics are becoming increasingly reliant on the identification of somatic genetic alterations using NGS. Owing to its relative low costs and fast turnaround, the Ion Torrent platform is one of the main sequencing platforms used in the clinical setting. Although the workflow for sample and library preparation, as well as for sequencing, is well standardized and streamlined, data analysis remains cumbersome, and it is difficult to obtain consistent and reliable results. A properly developed analysis pipeline is critical to ensuring adequate patient care.[7] The most common approach to analyzing Ion Torrent sequencing data is to use Thermo Fisher Scientific's proprietary IR software interface. The IR is highly customizable but also suffers from a number of drawbacks. Although optimized tumor-only analysis parameters for clinical-grade Oncomine panels are available out of the box, the default solutions for custom panels suffer from a high FP rate, requiring tuning of variant calling parameters, defining optimized filters, and/or extensive manual post-IR filtering.

To overcome these limits, PipeIT, a Singularity container for diagnostic somatic variant calling on the Ion Torrent platform, applicable for both Oncomine and custom targeted

sequencing panels, was developed. The pipeline was designed to account for the requirements of somatic variant calling analysis in a diagnostic setting. First, sensitivity and PPV are both important. In particular, a high PPV would reduce the workload in curating the results and increase reproducibility by minimizing variability associated with manual review of the results. Second, the ability of high-throughput analysis of many tumor-normal sample pairs by executing a single shell command, either on a desktop computer or in parallel in a high-performance computing environment, is desirable. Third, although the PipeIT workflow was designed to be run from start to finish, it is also possible to execute individual components (Table 1) instead of the complete PipeIT workflow. Fourth, reproducibility and portability are enabled by the use of the Singularity container technology, thus removing the hassle of complex software setup and ensuring that results are reproducible in any hardware and operating system configuration. This in turn facilitates the pipeline validation process that is necessary when pipelines are deployed in a new laboratory.[7] The read-only nature of a Singularity container also prevents unintentional alterations of the software setup by the users.

The performance of PipeIT was demonstrated using two data sets. In the first set of 15 FFPE colon adenomas sequenced using the Oncomine Comprehensive Panel, PipeIT was able to identify the bona fide pathogenic mutations identified by IR-Oncomine, the optimized IR workflow for diagnostic use. PipeIT did not identify a number of germline variants called by IR-Oncomine and the many IR-TN−specific variants enriched for low VAF and/or low depth, many of which are likely to have been fixation artifacts or are otherwise FP results, as shown by the enrichment of C>T mutations at the low VAF range (Supplemental Figure S3).[24] In 10 fresh-frozen HCCs sequenced using a custom AmpliSeq panel, PipeIT has excellent PPV compared with IR solutions. Benchmarked against the mutations identified by WES of the HCCs, PipeIT identified the most known mutations while maintaining excellent PPV compared with both IR-default and IR-custom, including variants at low VAF. Interestingly, although IR-default (HC) suffered from poor PPV of <80%, IR-custom (HC) had 100% PPV, and its performance was comparable to PipeIT. This observation underlines the necessity of molecular diagnostics laboratories to customize their own analysis parameters and filters; PipeIT provides a tested and easy-to-implement solution.

In conclusion, PipeIT offers a fully automated, self-contained pipeline for somatic variant calling for Ion Torrent sequencing, with minimal input requirements. The excellent PPV of PipeIT significantly reduces the need for extensive expert manual review. PipeIT is a useful addition to molecular diagnostics laboratories, especially for custom targeted sequencing panels, as well as for researchers seeking a workflow to analyze somatic mutations from Ion Torrent data.

Acknowledgments

Development of PipeIT was performed at the sciCORE scientific computing center at the University of Basel, Basel, Switzerland.

S.P. and C.K.Y.N. conceived and supervised the study; A.G., H.M., and C.K.Y.N. developed the methods; V.P. performed the sequencing experiments; P.M.J. provided expertise in the IR data analysis; G.R. critically reviewed the results; L.T. and L.M.T. provided the sequencing data and critically discussed the results; A.G., V.P., S.P., and C.K.Y.N. interpreted the results and wrote the manuscript; all authors agreed to the final version of the manuscript.

Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.jmoldx.2019.05.001*.

References

1. Joyner MJ, Paneth N: Seven questions for personalized medicine. JAMA 2015, 314:999−1000
2. Lander ES: Initial impact of the sequencing of the human genome. Nature 2011, 470:187−197
3. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G: Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014, 505:495−501
4. Shin S, Lee H, Son H, Paik S, Kim S: AIRVF: a filtering toolbox for precise variant calling in Ion Torrent sequencing. Bioinformatics 2018, 34:1232−1234
5. Deshpande A, Lang W, McDowell T, Sivakumar S, Zhang J, Wang J, San Lucas FA, Fowler J, Kadara H, Scheet P: Strategies for identification of somatic variants using the Ion Torrent deep targeted sequencing platform. BMC Bioinformatics 2018, 19:5
6. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, MC3 Working Group, et al: Comprehensive characterization of cancer driver genes and mutations. Cell 2018, 173:371−385
7. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB: Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn 2018, 20:4−27
8. Singer J, Ruscheweyh H-J, Hofmann AL, Thurnherr T, Singer F, Toussaint NC, Ng CKY, Piscuoglio S, Beisel C, Christofori G, Dummer R, Hall MN, Krek W, Levesque MP, Manz MG, Moch H, Papassotiropoulos A, Stekhoven DJ, Wild P, Wüst T, Rinn B, Beerenwinkel N: NGS-pipe: a flexible, easily extendable and highly configurable framework for NGS analysis. Bioinformatics 2018, 34:107−108
9. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010, 20:1297−1303
10. Kurtzer GM, Sochat V, Bauer MW: Singularity: scientific containers for mobility of compute. PLoS One 2017, 12:e0177459
11. Paradiso V, Garofoli A, Tosti N, Lanzafame M, Perrina V, Quagliata L: Diagnostic targeted sequencing panel for hepatocellular carcinoma genomic screening. J Mol Diagn 2018, 20:836−848

12. Piscuoglio S, Ng CKY, Murray MP, Guerini-Rocco E, Martelotto LG, Geyer FC, Bidard F-C, Berman S, Fusco N, Sakr RA, Eberle CA, De Mattos-Arruda L, Macedo GS, Akram M, Baslan T, Hicks JB, King TA, Brogi E, Norton L, Weigelt B, Hudis CA, Reis-Filho JS: The Genomic Landscape of Male Breast Cancers. Clin Cancer Res 2016, 22:4045−4056
13. Ng CKY, Di Costanzo GG, Tosti N, Paradiso V, Coto-Llerena M, Roscigno G, Perrina V, Quintavalle C, Boldanova T, Wieland S, Marino-Marsilia G, Lanzafame M, Quagliata L, Condorelli G, Matter MS, Tortora R, Heim MH, Terracciano LM, Piscuoglio S: Genetic profiling using plasma-derived cell-free DNA in therapy-naïve hepatocellular carcinoma patients: a pilot study. Ann Oncol 2018, 29: 1286−1291
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25:2078−2079
15. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010, 26:841−842
16. Tan A, Abecasis GR, Kang HM: Unified representation of genetic variants. Bioinformatics 2015, 31:2202−2204
17. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, Gao J, Socci ND, Solit DB, Olshen AB, Schultz N, Taylor BS: Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol 2016, 34:155−163
18. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, Zhang H, Solit DB, Taylor BS, Schultz N, Sander C: 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med 2017, 9:4
19. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012, 6:80−92
20. Thorvaldsdottir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013, 14:178−192
21. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group: The variant call format and VCFtools. Bioinformatics 2011, 27:2156−2158
22. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al: Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016, 536:285−291
23. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H: UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph 2014, 20:1983−1992
24. Do H, Dobrovic A: Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase. Oncotarget 2012, 3:546−558

3.2- Chapter II

Identification of HMGA1 molecular targets in hepatocellular carcinomas

HMGA1 is a protein known as an architectural transcription factor because it can bind to and modify the DNA. Even if it does not have transcriptional activity *per se*, it can assemble functional transcription units and by that can affect transcription [108,109]. HMGA1 is known to be involved in several cellular processes implicated in cancer development and tumour progression. It has also been found to be highly expressed in a broad range of tumours whilst it is not or lowly expressed in normal tissue. Even if the mechanisms of action are quite well studied for some of its targets, a broad characterisation of its molecular interactors and its specific role with them is not yet discovered, especially in the context of HCC. After our previous study showing an overexpression of HMGA1 in more than 50% of HCC patients in both transcription and protein levels [121], we decided to focus our attention on the role of HMGA1 in HCC and to elucidate its molecular targets.

This chapter contains my experimental work performed on HMGA1 on *in vitro* HCC models to explore more in depth the general role of this protein and, above all, to highlight its molecular targets of potential clinical utility.

Our first aim was to investigate the binding landscape of HMGA1 at the DNA level by performing chromatin immunoprecipitation sequencing (ChIP-seq) on endogenous levels of HMGA1 in HCC cells. The second aim was to define a gene and protein expression signature of HMGA1 deregulation. We performed RNA-sequencing on HCC cell lines after overexpression and silencing of *HMGA1* and mass spectrometry after silencing of *HMGA1*. The third aim was to identify HMGA1 molecular partners by mass spectrometry after immunoprecipitation.

Materials and Methods

Cell lines

All HCC-derived cell lines were maintained in a 5% CO2-humidified atmosphere at 37°C and cultured in DMEM supplemented with 10% FBS, 1% Pen/Strep (Bio-Concept) and 1% MEM-NEAA (MEM non-essential amino acids, Thermo Fisher Scientific). All cell lines were confirmed negative for mycoplasma infection using the PCR-based Universal Mycoplasma Detection kit (American Type Culture Collection) according to standard protocol.

Plasmids and transfection

The vector for the overexpression of HMGA1 was designed and ordered on GenScript Biotech (Figure 3.2.1). As empty control vector was used pCMV-mir EGFP. $2x10^5$ cells per 6-well were plated 24h before transfection. The expression vectors were transfected using the jetPRIME transfection reagent (Polyplus) following the manufacturer's instructions. Briefly, 2µg of plasmid were added in 200µl of jetPRIME buffer. Successively, 4µl of jetPRIME reagent were added. The transfection mix was then properly mixed and was incubated at room temperature for 10 minutes. The transfection mix was added dropwise onto the cells and distributed evenly. To avoid cytotoxicity, the transfection medium was replaced after 4 hours by complete medium. The expression of the plasmids was evaluated by western blot and qRT-PCR analysis 48h after transfection.

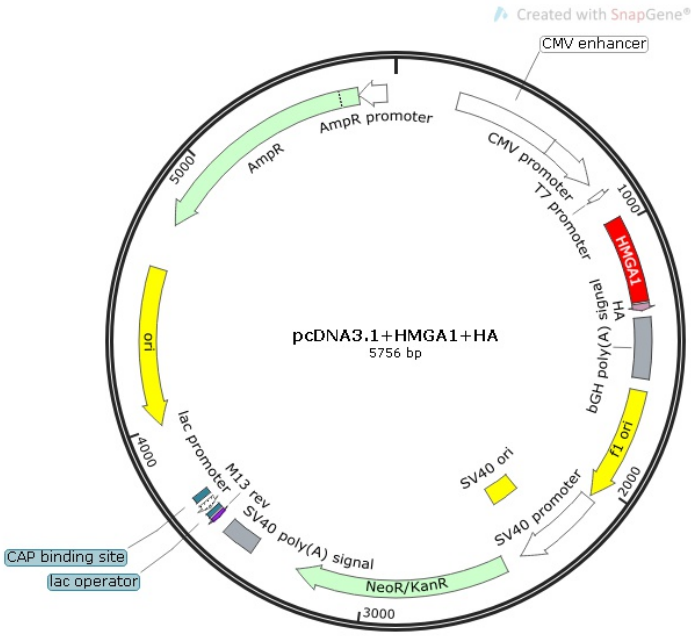


Figure 3.2.1: *Construct map of HMGA1 cloning vector by GenScript Biotech.*

siRNA and transfection

The silencing of HMGA1 was performed using ON-TARGET plus siRNA transfection. ON-TARGET plus SMARTpool siRNAs against human HMGA1 and ON-TARGET plus SMARTpool non-targeting control and DharmaFECT transfection reagent were all purchased from GE Dharmacon.

Catalog ID for ON-TARGET plus SMARTpool siRNAs against human HMGA1: L-004597-00-0005

Catalog ID for ON-TARGET plus SMARTpool non-targeting control: D-001810-10-05

Transfection was performed according to the manufacturer's protocol. Briefly, HCC cancer cells were seeded at approximately 60% confluence. Because residual serum affects the knockdown efficiency of ON-TARGET plus siRNAs, growth medium was removed as much as possible and replaced by serum-free medium (Opti-MEM). siRNAs were added to a final concentration of 25nM. Cells were incubated at 37°C in 5% CO2 for 48 hours or for 48 and 72 hours (for western blotting after subcellular fractionation). To avoid cytotoxicity, the transfection medium was replaced with a complete medium after 24 hours.

RNA extraction and qRT-PCR

RNA extraction was performed using the TRIzol Reagent (ThermoFisher) followed by an additional DNase treatment using DNA-free DNA Removal Kit (ThermoFisher) according to the manufacturer's specifications. RNA concentrations were quantified using Qubit Fluorometer (Life Technologies).

1µg of RNA was retro-transcribed using the SuperScript VILO cDNA synthesis kit (Thermo Fisher Scientific). Quantitative real-time PCR for the expression levels of *HMGA1* was performed with Sybr Green method. When GAPDH was used as a housekeeping gene, the fold changes in gene expression were calculated using the standard ΔΔCt method [153]. To quantitate HMGA1 transcript levels, dilutions of the HMGA1 vector were used as standard curves (dilutions ranged from 2.5 to 1 million copies of plasmid).

HMGA1 primers:

Forward: CAACTCCAGGAAGGAAAC

Reverse: AGGACTCCTGCGAGATGC

GAPDH primers:

Forward: AGGTGAAGGTCGGAGTCAACG

Reverse: TGGAAGATGGTGATGGGATTT

Protein extraction and Western Blot

Proteins were extracted using Co-IP buffer (100mmol/L NaCl, 50mmol/L Tris pH 7.5, 1mmol/L EDTA, 0.1% Triton X-100) supplemented with 1x protease inhibitors (cOmplete Mini, EDTA-free Protease Inhibitor Cocktail, Roche) and 1x phosphatase inhibitors (PhosSTOP, Merck) at 4°C for 30 minutes. Cell lysates were quantified by the Bradford protein assay (Bio-Rad) using Bovine Serum Albumin Acetylated (Promega) as standard in seven diluted concentrations. Same concentrations of the cell lysates were then treated with 1x reducing agent (NuPAGE Sample Reducing Agent, Invitrogen), 1x loading buffer (NuPAGE LDS Sample Buffer, Invitrogen), boiled at 70°C for 10 minutes and loaded into neutral pH, pre-cast, discontinuous SDS-PAGE mini-gel system (NuPAGE 4-12% Bis-Tris Protein Gels, Thermo Fisher). The proteins were then transferred to nitrocellulose membranes using the Trans-Blot Turbo Transfer System (Bio-Rad). The membranes were blocked for 1 hour with 3% Sure Block (Lubio science) and then probed with primary antibodies overnight at 4°C. Next day, the membranes were incubated for 1 hour at room temperature with fluorescent secondary goat anti-mouse (IRDye 680) or anti-rabbit (IRDye 800) antibodies (both from LI-COR Biosciences). Blots were scanned using the Odyssey Infrared Imaging System (LICOR Biosciences) and band intensity was quantified using ImageJ software. The ratio of proteins of interest/loading control in treated samples were normalised to their counterparts in control cells.

Antibodies

Antibody	**Company**	**Catalog number**	**Application**	**Dilution / Concentration**
HMGA1	abcam	ab4078	ChIP-seq	5 µg
HMGA1	Cell Signaling	D6A4 – 7777	WB, IHC, IP	1:1000, 1:500, 4 µg
Alyref	abcam	ab202894	WB, IP	1:2000, 4 µg

Rabbit IgG	Cell Signaling	2729	IP	4 µg
β-actin	Sigma Aldrich	A5441	WB	1:5000
MEK1/2	Cell Signaling	D145 – 8727	WB	1:1000
AIF	Cell Signaling	D39D2 – 5318	WB	1:1000
Vimentin	Cell Signaling	D21H3 – 5741	WB	1:1000
Histone H3	Cell Signaling	D1H2 – 4499	WB	1:1000

Chromatin Immunoprecipitation (ChIP) – sequencing

In collaboration with Song Shuang and Prof. Patrick Matthias from the Friedrich Miescher Institute for Biomedical Research, and Dr. Fengyuan Tang from the Department of Biomedicine in Basel, we established a protocol to perform the ChIP-seq for HMGA1. In brief, more than 50 million cells per sample were crosslinked and subjected to sonication to obtain DNA fragments of approximately 300 bp. Chromatin was immunoprecipitated using a new ChIP-grade anti-HMGA1 antibody (Abcam). Library construction from immunoprecipitated DNA and the input (50ng each) was performed using NEBNext Ultra II DNA Library preparation kit (New England BioLabs) according to the standard protocol. Massively parallel sequencing was performed on an Illumina NextSeq 550 in our institute according to manufacturer's instructions. We obtained ~30 million reads per sample. The study was performed on 2 biological replicates for each cell line.

Analysis of ChIP-seq

ChIP-seq analysis was performed by the bioinformaticians in our group, in particular Dr. Gallon and Dr. De Filippo. ChIP-seq reads were aligned to the human reference genome hg19 using the BWA aligner. Peaks were called using MACS2 [154] and subsequently combined across replicates, into consensus peak lists, using MSPC [155]. A second analysis for identification of enriched regions was done by counting reads overlapping regions using countOverlaps from the GenomicRanges package in R. Pathway analysis was carried out on consensus peak lists for each cell line whereby peaks were annotated using the ChIPseeker package, before KEGG pathway analysis was performed using the clusterProfiler package in R.

RNA sequencing

Extracted RNA from cells was subjected to quality control before library preparation. The RNA integrity number was > 9 for all the samples. 100ng RNA were used for each sample for the construction of libraries, according to the poly-A tail selection TruSeq Stranded mRNA protocol (Illumina). The libraries were prepared and sequenced on an Illumina NovaSeq 6000 by CeGaT GmbH. We obtained > 30 million reads for each sample. The study was performed on 2 biological replicates for each cell line.

Analysis of RNA-seq

RNA-seq analysis was performed by Dr. Gallon. Sequencing reads generated were aligned to hg19 and assigned to genes using STAR 2.5 [156]. Analysis of differential gene expression between HMGA1 dysregulated and control samples was performed using DESeq2 [157] and Gene Set Enrichment Analysis was then carried out using the moderated T statistics from this analysis to detect differential expression of gene sets defined in the MsigDB database, using the 'fgsea' package in R [158].

Mass-spectrometry (MS)

Extracted proteins from whole cells ($2.5x10^8$ per sample) were lysed in 50µL of lysis buffer (1% Sodium deoxycholate (SDC), 10mM TCEP, 100mM Tris pH=8.5) using 10 cycles of sonication (30 seconds on, 30 seconds off per cycle) on a Bioruptor (Dianode). Following sonication, proteins in the cell lysates were reduced by TCEP at 95°C for 10 minutes. Proteins were then alkylated using 15mM chloroacetamide at 37°C for 30 minutes and further digested using sequencing-grade modified trypsin (1/50, w/w, trypsin/protein; Promega) at 37°C overnight. After digestion, the samples were acidified with TFA to a final concentration of 1%. Peptide Desalting was performed using iST cartridges (PreOmics) following the manufacturer's instructions. Peptides were dried under vacuum. Sample aliquots comprising 12.5µg of peptides were resuspended in 10µl labelling buffer (2M urea, 0.2 M HEPES pH 8.3) by sonication and labelled with isobaric tandem mass tags (TMTpro 16-plex, Thermo Fisher Scientific). For that, 2.5µL of each TMT reagent were added to the individual peptide samples followed by a 1 hour incubation at 25°C shaking at 500 rpm. To control for ratio distortion during quantification, a peptide calibration mixture consisting of six digested standard proteins mixed in different amounts was added to each sample before TMT labelling (for details see [159]). To quench the labelling reaction, 0.75µL aqueous 1.5M hydroxylamine solution was added and samples were incubated for 5 minutes at 25°C shaking at 500 rpm followed by pooling of all samples. The pH of the sample pool was

increased to 11.9 by addition of 1M phosphate buffer (pH 12) and incubated for 20 minutes at 25°C shaking at 500 rpm to remove TMT labels linked to peptide hydroxyl groups. Subsequently, the reaction was stopped by addition of 2M hydrochloric acid until a pH < 2 was reached. Peptide samples were further acidified using 5% TFA and desalted using BioPureSPN Macro spin columns (The Nest Group) according to the manufacturer's instructions and dried under vacuum.

TMT-labelled peptides were fractionated by high-pH reversed phase separation using a Xbridge Peptide BEH C18 column (3.5µm, 130Å, 1mm x 150mm, Waters) on an Agilent 1260 Infinity HPLC system. Peptides were loaded on column in buffer A (20mM ammonium formate in water, pH 10) and eluted using a two-step linear gradient from 2% to 10% in 5 minutes and then to 50% buffer B (20mM ammonium formate in 90% acetonitrile, pH 10) over 55 minutes at a flow rate of 42 µl/min. Elution of peptides was monitored with a UV detector (215nm, 254nm) and a total of 36 fractions were collected, pooled into 12 fractions using a post-concatenation strategy as previously described (Ahrné et al. 2016) and dried under vacuum. Dried peptides were resuspended in 0.1% aqueous formic acid and subjected to LC–MS/MS analysis by the team of Proteomics Core Facility, Biozentrum, University of Basel, using a Q Exactive HF Mass Spectrometer fitted with an EASY-nLC 1000 (both Thermo Fisher Scientific) and a custom-made column heater set to 60°C.

Analysis of Mass-spectrometry (MS)

The acquired MS raw-files were analysed by the team of Proteomics Core Facility, Biozentrum, University of Basel, using the SpectroMine software (Biognosis AG). Spectra were searched against a human database consisting of 20767 protein sequences (downloaded from Uniprot on 20190307) including commonly observed contaminants. Standard Pulsar search settings for TMT 16 pro ("TMTpro_Quantification") were used and resulting identifications and corresponding quantitative values were exported on the PSM level using the "Export Report" function. Acquired reporter ion intensities in the experiments were employed for automated quantification and statistical analysis using an in-house developed SafeQuant R script (v2.3 [159]). This analysis included adjustment of reporter ion intensities, global data normalization by equalizing the total reporter ion intensity across all channels, summation of reporter ion intensities per protein and channel, calculation of protein abundance ratios and testing for differential abundance using empirical Bayes moderated t-statistics. Finally, the calculated p-values were corrected for multiple testing using the Benjamini−Hochberg method.

All LC-MS analysis runs were acquired from three independent biological samples. To meet additional assumptions (normality and homoscedasticity) underlying the use of linear regression

models and Student t-Test MS-intensity signals were transformed from the linear to the log-scale. Linear regression was performed using the ordinary least square (OLS) method as implemented in the base package of R v.3.1.2 (http://www.R-project.org/). The Proteomics Core Facility performed the analysis assuming a within-group MS-signal Coefficient of Variation of 10% in a sample size of three biological replicates. They applied a two-sample, two-sided Student t test that gives adequate power (80%) to detect protein abundance fold changes higher than 1.65, per statistical test. The statistical package used to assess protein abundance changes, SafeQuant, employs a moderated t-Test, which has been shown to provide higher power than the Student t-test.

Immunoprecipitation (IP)

Cells from 10cm plates (one for regular IP, four for IP-MS) were lysed with Co-IP buffer (100mmol/L NaCl, 50mmol/L Tris pH 7.5, 1mmol/L EDTA, 0.1% Triton X-100) supplemented with 1x protease inhibitors (cOmplete Mini, EDTA-free Protease Inhibitor Cocktail, Roche) and 1x phosphatase inhibitors (PhosSTOP, Merck) at 4°C for 30 minutes. The lysates were first cleared by spinning at 16,000g at 4°C for 30 minutes to remove cell debris, pre-cleared using Dynabeads Protein G (Thermo Fisher) for 1 hour and then the protein amount was split and used for immunoprecipitation with 4µg of either anti-HMGA1 antibody or Rabbit-IgG antibody (for the control). The binding of the targets to the antibody occurred at 4°C overnight. Ips were performed at 4°C for 1hour and 30 minutes adding Dynabeads Protein G (Thermo Fisher) and then washed five times in lysis buffer. The proteins were eluted with SDS sample buffer (NuPAGE LDS Sample Buffer, Invitrogen) and heated at 70°C for 10 minutes for regular IP samples and eluted with glycine 0.2M ph 2.7 for IP-MS samples.

Mass spectrometry after Immunoprecipitation (IP-MS)

Glycine eluted samples after immunoprecipitation were subjected to on-bead digestion (adapted from [160]) by trypsin (5µg/ml, Promega) in 1.6M urea / 0.1M ammonium bicarbonate buffer at 27°C for 30 minutes. Supernatant eluates containing active trypsin were further incubated with 10mM TCEP and 15mM Chloroacetamide at 37°C overnight in order to achieve complete digestion and carbamidomethylation of cysteines. The tryptic digest was acidified (pH < 3) using TFA and desalted using C18 reversed phase spin columns (Harvard Apparatus) according to the protocol of the manufacturer. Dried peptides were dissolved in 0.1% aqueous formic acid solution at a concentration of 0.2mg/ml prior to injection into the mass spectrometer. Aliquots of 0.4µg of total peptides (in 0.1% aqueous formic acid) were subjected to LC-MS analysis again by the team of Proteomics Core Facility, Biozentrum, University of Basel, using a dual pressure LTQ-Orbitrap

Elite mass spectrometer connected to an electrospray ion source (both Thermo Fisher Scientific) and a custom-made column heater set to 60°C.

Analysis of IP-MS

The acquired MS raw-files were imported into the Progenesis QI software (v2.0, Nonlinear Dynamics Limited), which was used to extract peptide precursor ion intensities across all samples applying the default parameters, and analysed by the team of Proteomics Core Facility, Biozentrum, University of Basel. The generated mgf files were searched using MASCOT against a decoy database containing normal and reverse sequences of the concatenated Homo sapiens (UniProt, May 2018) proteome including commonly observed contaminants (in total 41534 sequences) generated using the SequenceReverser tool from the MaxQuant software (Version 1.0.13.13). The following search criteria were used: full tryptic specificity was required (cleavage after lysine or arginine residues, unless followed by proline); 3 missed cleavages were allowed; carbamidomethylation I was set as fixed modification; oxidation (M) and protein N-terminal acetylation were applied as variable modifications; mass tolerance of 10ppm (precursor) and 0.6Da (fragments) was set. The database search results were filtered using the ion score to set the false discovery rate (FDR) to 1% on the peptide and protein level, respectively, based on the number of reverse protein sequence hits in the datasets. Quantitative analysis results from label-free quantification were normalised and statically analysed using the SafeQuant R package v.2.3.4 (https://github.com/eahrne/SafeQuant/; [159]) to obtain relative protein abundances. This analysis included summation of peak areas per protein and LC MS/MS run followed by calculation of protein abundance ratios. Only isoform specific peptide ion signals were considered for quantification. The summarised protein expression values were used for statistical testing of differentially abundant proteins between conditions. Empirical Bayes moderated T-tests were applied, as implemented in the R/Bioconductor limma package (http://bioconductor.org/packages/release/bioc/html/limma.html). The resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg method.

g:Profiler

Pathway enrichment analysis was performed on g:Profiler using the KEGG and the Gene Ontology (GO) datasets. Enriched pathways were used for the screening of the molecular partners of HMGA1.

Immunohistochemistry

For immunocytochemistry, cultured cells were washed with PBS and fixed in 4% paraformaldehyde overnight at 4°C. After fixation and further washing, cells were encapsulated in a drop of Richard-Allan Scientific HistoGel (Thermo Fisher), histoprocessed and embedded. The FFPE block was then cut and dried on glass slides. Antigen retrieval and immunohistochemical staining were performed on the glass slides on an automated Benchmark Leica Bond III immunohistochemistry staining system (Leica Biosystems) using a BOND polymer refine detection kit (Leica Biosystems). Heat-induced epitope retrieval (HIER) was performed on all slides in Bond Epitope Retrieval Solution 1, pH 6 (Leica Biosystems) for 20 minutes. As a primary antibody, anti-HMGA1 (Cell Signaling) was applied and incubated for 20 minutes at room temperature at 1:500 dilution. Images were acquired using an Olympus BX46 microscope.

Subcellular fractionation

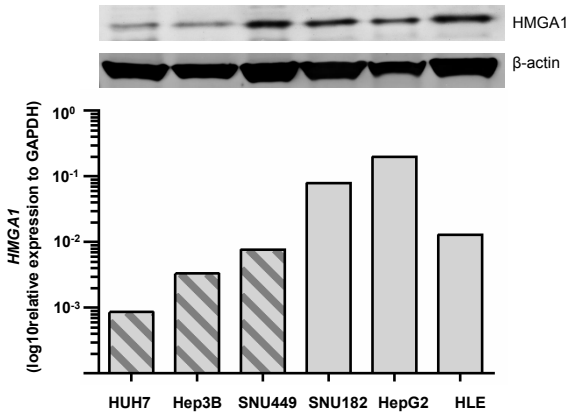
The isolation of different cellular compartments was performed using $2.5x10^6$ cells per sample following the manufacturer's instructions of Cell Fractionation Kit (Cell Signaling). All steps were performed at 4°C. 30µl of all fractions were then treated with 1x loading buffer (NuPAGE LDS Sample Buffer, Invitrogen), boiled at 70°C for 10 minutes and loaded into a SDS-PAGE mini-gel system (NuPAGE 4-12% Bis-Tris Protein Gels, Thermo Fisher) to proceed with the blotting.

Results

I. HMGA1 genome-wide DNA-binding landscape in HCC

As previously stated, the exact role of HMGA1 is only partially known. One of its main roles is acting as an architectural transcription factor, however, its binding sites have not yet been systematically investigated. For this reason our first aim was to explore HMGA1 binding landscape at the DNA level in an *in vitro* model of HCC. To achieve this goal, we performed chromatin immunoprecipitation sequencing (ChIP-seq). Between a panel of six HCC cell lines available in the laboratory we selected three (Hep3B, HUH7, SNU449) with different characteristics (HBV positive and negative, different tumour grades, different patient's age and different HMGA1 endogenous levels) to represent the disease (Figure 3.2.2). The ChIP-seq experiments were run in duplicate for each cell line.

A



B

	HUH7	Hep3B	SNU449
HBV infection	no	yes	yes
Tumour grade	II	II	III/IV
Age group	adult	child	adult

Figure 3.2.2: *HMGA1 screening and characteristic of the chosen HCC cell lines.* ***A)*** *Six HCC cell lines are shown to represent the endogenous amount of HMGA1 at the protein (upper part) and mRNA levels (lower part). In oblique lines pattern the cell lines used for the ChIP-seq analysis (HUH7, Hep3B, SNU449).* ***B)*** *A table showing the characteristics (HBV infection, tumour grade and age group) of the three HCC cell lines (HUH7, Hep3B, SNU449) chosen for the ChIP-seq analysis.*

HMGA1 is a DNA-binding protein with low complexity motif (AT-rich regions); it might bind to any open region in the chromatin containing an enrichment of AT bases. Our results indeed confirmed the binding of HMGA1 protein on distinctive genomic regions in dependency of a specific DNA-binding pattern. ChIP-seq analysis revealed on average high occupancy of HMGA1 protein toward the entire AT-rich DNA (average of 60% binding in AT regions, Table 3.2.1), in accordance with a recent study (D. F. Colombo et al. 2017). Broad regions with high enrichment of HMGA1 binding sites had high AT content and they were consistent between replicates. This consistency provided direct support for the specificity of HMGA1-binding for AT-regions independently of the function of DNA domains. In addition to this, we found similar percentages of AT content through HMGA1 binding sites between cell lines. The difference of total AT content between cell lines, despite not being significant, suggested a correlation with the amount of endogenous HMGA1.

Sample	**Mean AT content (%)**	**SD of AT content**
HUH7_1	73.60	0.0932
HUH7_2	69.27	0.1179
Hep3B_1	66.28	0.0987
Hep3B_2	66.26	0.0901
SNU449_1	60.62	0.1181
SNU449_2	63.37	0.1313

Table 3.2.1 *AT content mean and standard deviation (SD) of binding sites from input normalised data from all samples and their replicates obtained by ChIP-seq analysis.*

With the exploration of the DNA domains of the HMGA1 bindings, we found that the majority of the peaks were in introns and distal intergenic regions and fewer than 20% of the peaks were in promoter regions, while 10% were found in the first intron, only 0.6% were found in exons (Figure 3.2.3A). Furthermore, 14% of the peaks were located in the 6Kb spanning promoters (Figure 3.2.3B). Notable also between the cell lines were differences in the distribution of HMGA1 binding sites. Hep3B showed the greatest proportion of peaks binding close to transcription start sites

(TSS'), with 15% of HMGA1 falling within 6 Kb of promoters, while only 12.6% of HMGA1 peaks were within this distance of promoters in the SNU449 line. This is also demonstrated by a greater percentage of Hep3B HMGA1 peaks as being nearer TSS' than in either SNU449 or HUH7 cells. This finding was identified again in the level of HMGA1 ChIP-seq coverage around TSS' in Hep3B compared to SNU449, with Hep3B showing greater normalised coverage around TSS' than both SNU449 and HUH7 (Figure 3.2.3C). It is interesting to note that there appears to be a substantial drop in HMGA1 coverage around TSS' in both Hep3B and HUH7, while this was less evident in SNU449. Taken together the mapping results of the ChIP sequencing suggests a non-conventional role of this protein as transcription factors.

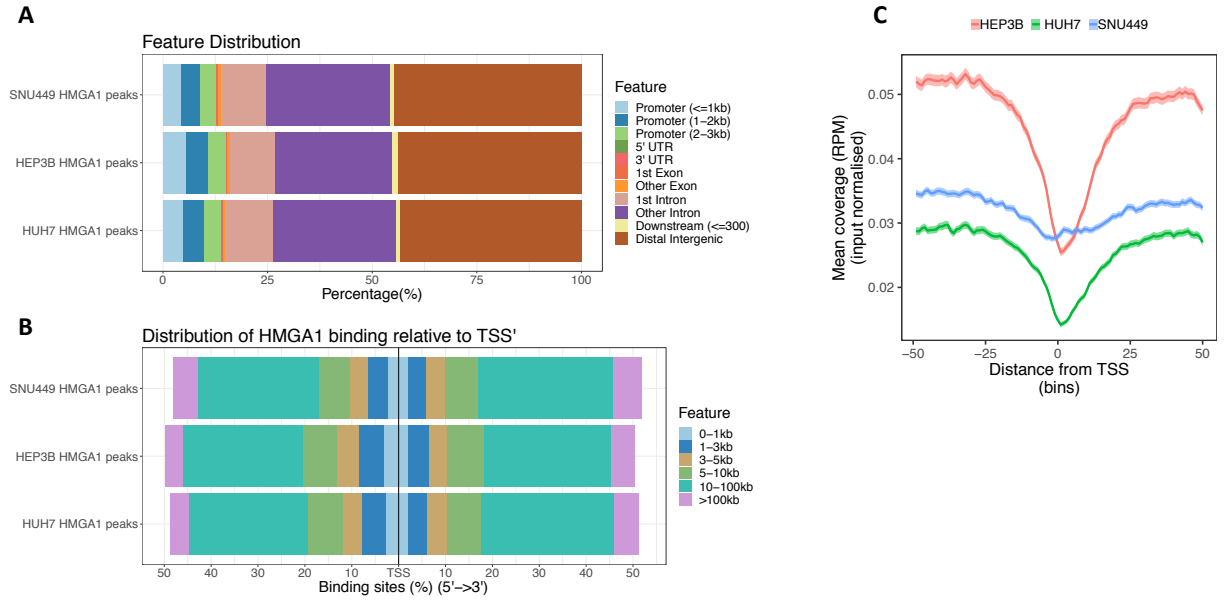


Figure 3.2.3: *Characterisation of HMGA1 DNA binding.* ***A)*** *Proportionate distribution of MACS2 called HMGA1 peaks, in each of the cell lines analysed, according to the nearest genomic feature.* ***B)*** *Distribution of HMGA1 peaks relative to transcription start sites (TSS'), defined by the UCSC database.* ***C)*** *HMGA1 ChIP-seq coverage, normalised against input, in 3 Kb region around TSS', for each cell line analysed.*

To gain insight into the signalling pathways that may be regulated by HMGA1 at the transcriptional level, we performed a KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathway analysis using the genes nearest to the called peaks in each cell line (Figure 3.2.4). Although Hep3B did not show enrichment for any pathways in its peak distribution, in the two HCC cell lines HUH7 and SNU449 there was an enrichment of binding sites for HMGA1 in proximity to non-canonical genes involved in the epithelial to mesenchymal transition (EMT) such as *CLDN3*, *CDH26*, *EZH2*. For example, the "Axon guidance" pathway includes Rho GTPases and cyclines (such as *Cycline A2*) that are involved in the microtubule and cytoskeletal organization [162-164]. Furthermore, both cell lines showed enrichment in the "Rap1 signalling pathway", implicated in cell-adhesion in response to various membrane receptors and in migration and invasion [165,166]. More specifically,

SNU449 also showed an enrichment of peaks closed to genes connected to adherens junctions while in HUH7 there was significant enrichment for genes related to gap junctions.

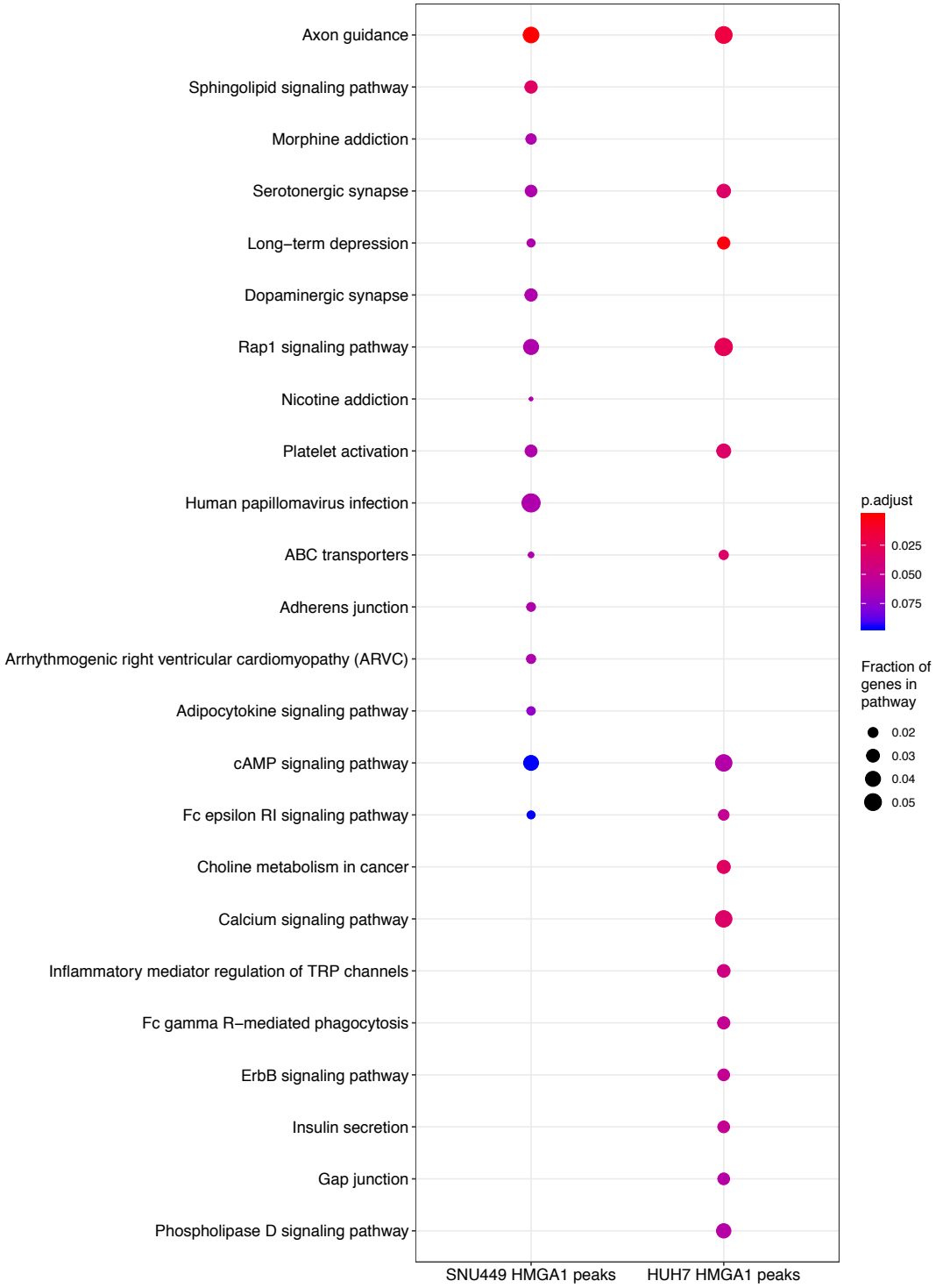


Figure 3.2.4: *KEGG pathway enrichment analysis on normalised data peaks obtained by ChIP-seq in each of the cell lines analysed. Hep3B did not show any enriched pathway. Points are coloured according to -log10 (Benjamini Hochberg adjusted* P *value).*

II. HMGA1 expression signature in HCC

To identify the “direct” (i.e. directly bound to HMGA1) and “indirect” (i.e. not bound to HMGA1) targets of HMGA1, we performed RNA-seq on HCC cell lines after dysregulation of HMGA1. We screened the above mentioned panel of liver cancer cell lines and we selected two with the lowest endogenous level of HMGA1, HUH7 and Hep3B, for protein overexpression, and three with high endogenous levels, SNU182, HLE and SNU449, for silencing (Figure 3.2.5A). RNA-seq was performed on the chosen 5 HCC cell lines, the dysregulation of HMGA1 in all samples was confirmed using both western blot analysis and quantitative real-time PCR (qRT-PCR) (Figure 3.2.5B-F). The RNA-seq was executed in collaboration with CeGaT facility.

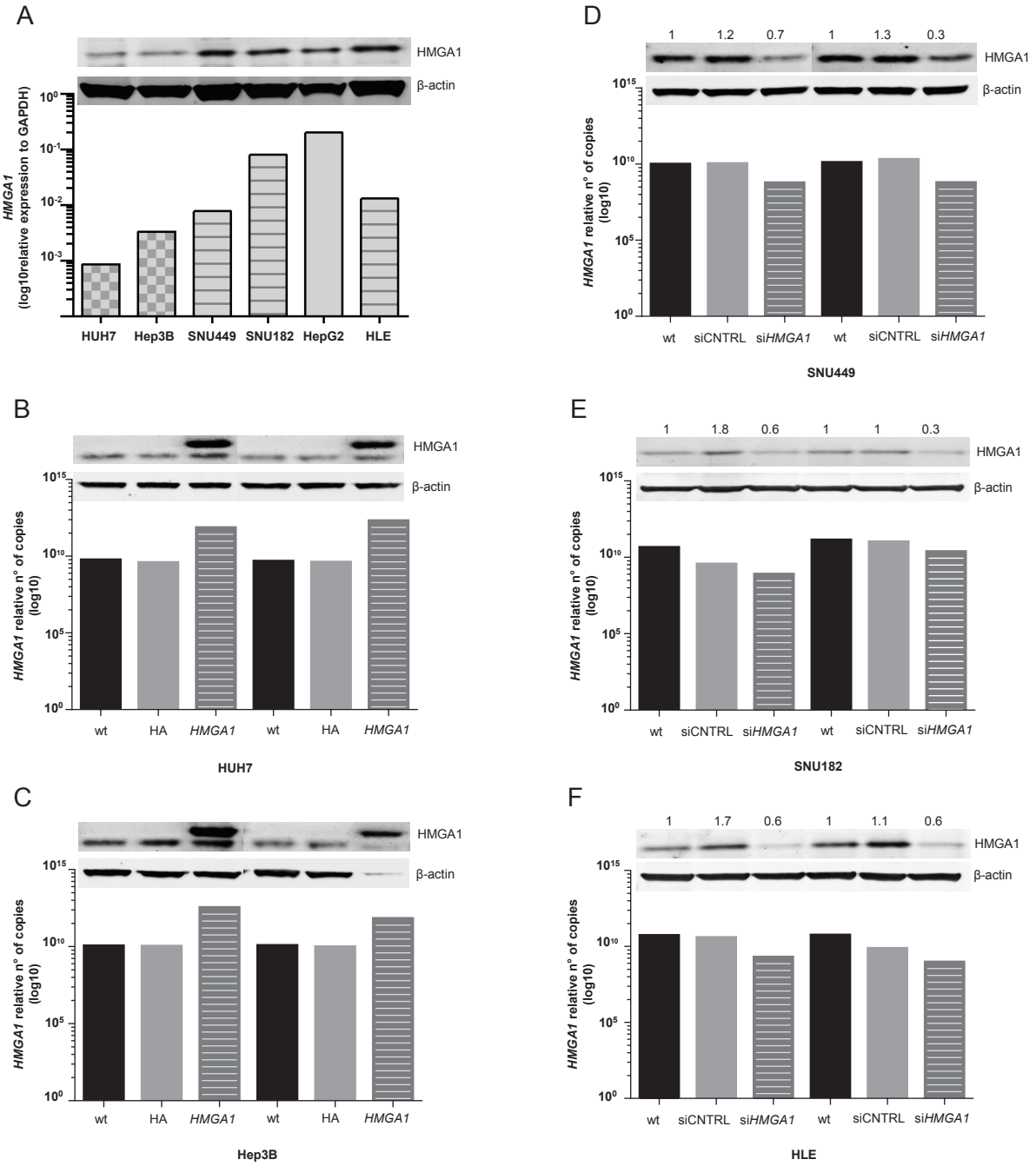


Figure 3.2.5: *HCC cell lines screening and expression level of HMGA1 after overexpression and silencing.* ***A)*** *The same data as in Figure 3.2.2A. Six HCC cell lines are shown to represent the endogenous amount of HMGA1 at the protein (upper part) and mRNA levels (lower part). In checkered pattern the cell lines used for HMGA1 overexpression (HUH7, Hep3B), in striped pattern the ones for silencing (SNU449, SNU182, HLE)* ***B)*** *Expression of HMGA1 in HUH7 wild type (wt), in HUH7 after transfection with the plasmid control (HA) and with the HMGA1-overexpressing plasmid (HMGA1) at the protein and mRNA levels.* ***C)*** *Expression of HMGA1 in Hep3B wild type (wt), in Hep3B after transfection with the plasmid control (HA) and with the HMGA1-overexpressing plasmid (HMGA1) at the protein and mRNA levels.* ***D)*** *Expression of HMGA1 in SNU449 wild type (wt) and SNU449 after transfection with the control pool of siRNA (siCNTRL) and the HMGA1 pool of siRNA (siHMGA1) at protein and mRNA level.* ***E)*** *Expression of HMGA1 in SNU182*

wild type (wt) and SNU182 after transfection with the control pool of siRNA (siCNTRL) and the HMGA1 pool of siRNA (siHMGA1) at protein and mRNA level. ***F)*** *Expression of HMGA1 in HLE wild type (wt) and HLE after transfection with the control pool of siRNA (siCNTRL) and the HMGA1 pool of siRNA (siHMGA1) at the protein and mRNA levels. Each experiment was repeated twice and both replicates are shown for each cell line.*

Unexpectedly, there was not a substantial effect on gene expression level between the HMGA1 dysregulated vs control samples, as shown in the volcano plots (Figure 3.2.6). It is although noticeable that *HMGA1* is the most significant upregulated/downregulated gene in each cell line compared to control cells, with a log2 fold change in expression of -4.41, -4.33 and -3.37 when silenced in SNU449, SNU182 and HLE respectively, and 5.00 and 6.26 when overexpressed in Hep3B and HUH7 respectively.

To have a broader overview on the effect of HMGA1 dysregulation, we also examined the changes in response to HMGA1 alteration at the protein level. We performed mass spectrometry analysis on the three transiently silenced HCC cell lines. Consistent with the results obtained by RNA-seq, no substantial effect was observed at the protein level between the control and the silenced conditions (Figure 3.2.7). HMGA1 was significantly downregulated in the three cell lines; especially SNU449 and HLE showed substantial differences compared to the endogenous levels. In the SNU182 the low adjusted p value (Q value) showed that there was a consistent decrease in HMGA1 expression, but the magnitude of this change was just less than the stringent threshold of a log2FC ≥ 1 which was applied (-0.9). However, none of these results were able to unveil the role of HMGA1 in the context of HCC, so we decided to focus our attention on the molecular partners of our protein.

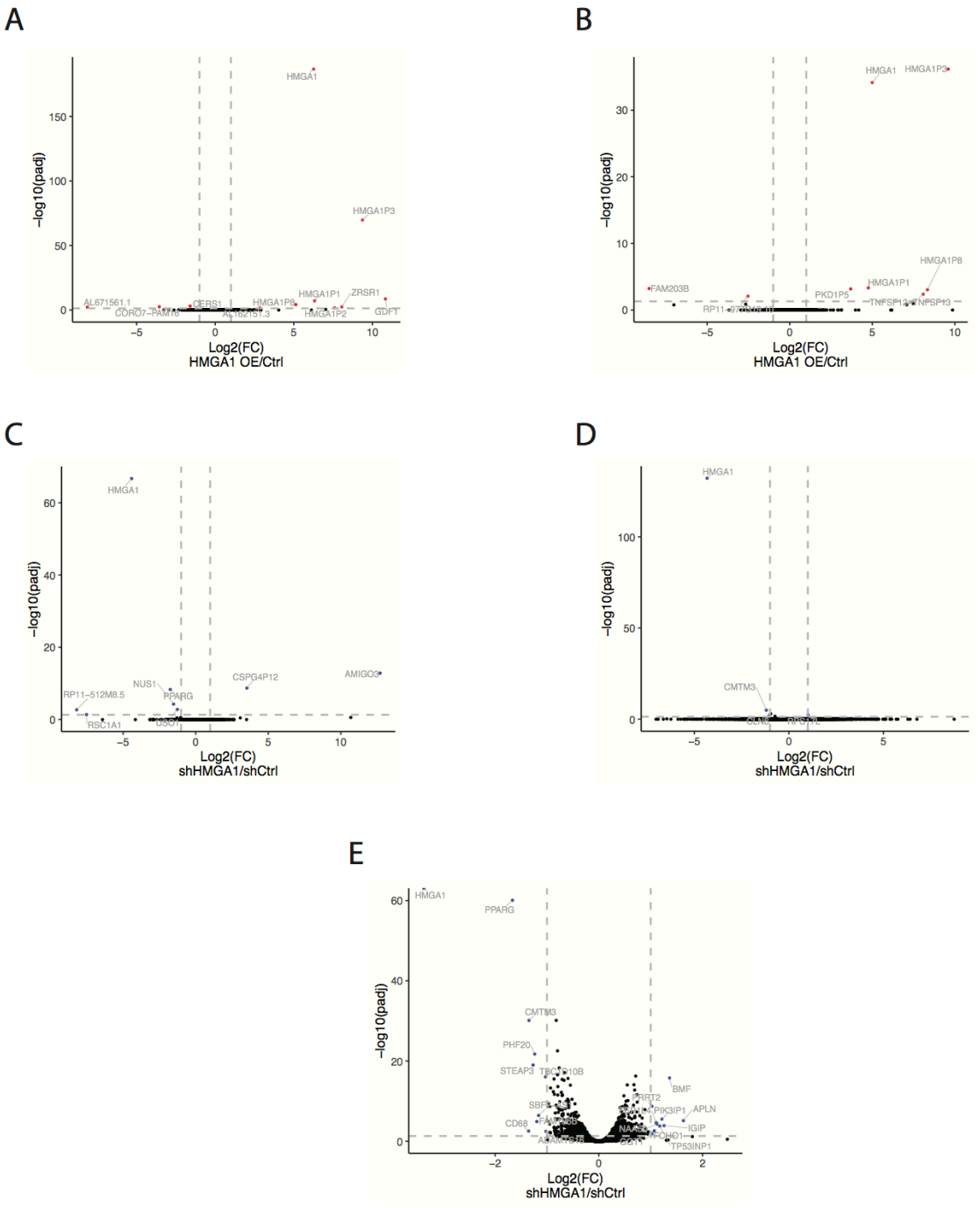


Figure 3.2.6: *Differential gene expression on overexpression and silencing of HMGA1. Volcano plots showing -log10 (FDR, false discovery rate) against log2FC (fold change) in gene expression for all genes analysed. Genes showing significantly different expression (log2FC > ± 1 and FDR < 0.05) are labelled in red in the analysis in the cell lines designed for the overexpression (**A)** HUH7, **B)** Hep3B), in blue for the silencing (**C)** SNU449, **D)** SNU182, **E)** HLE).* HMGA1 *is the most significant dysregulated in all the five cell lines.*

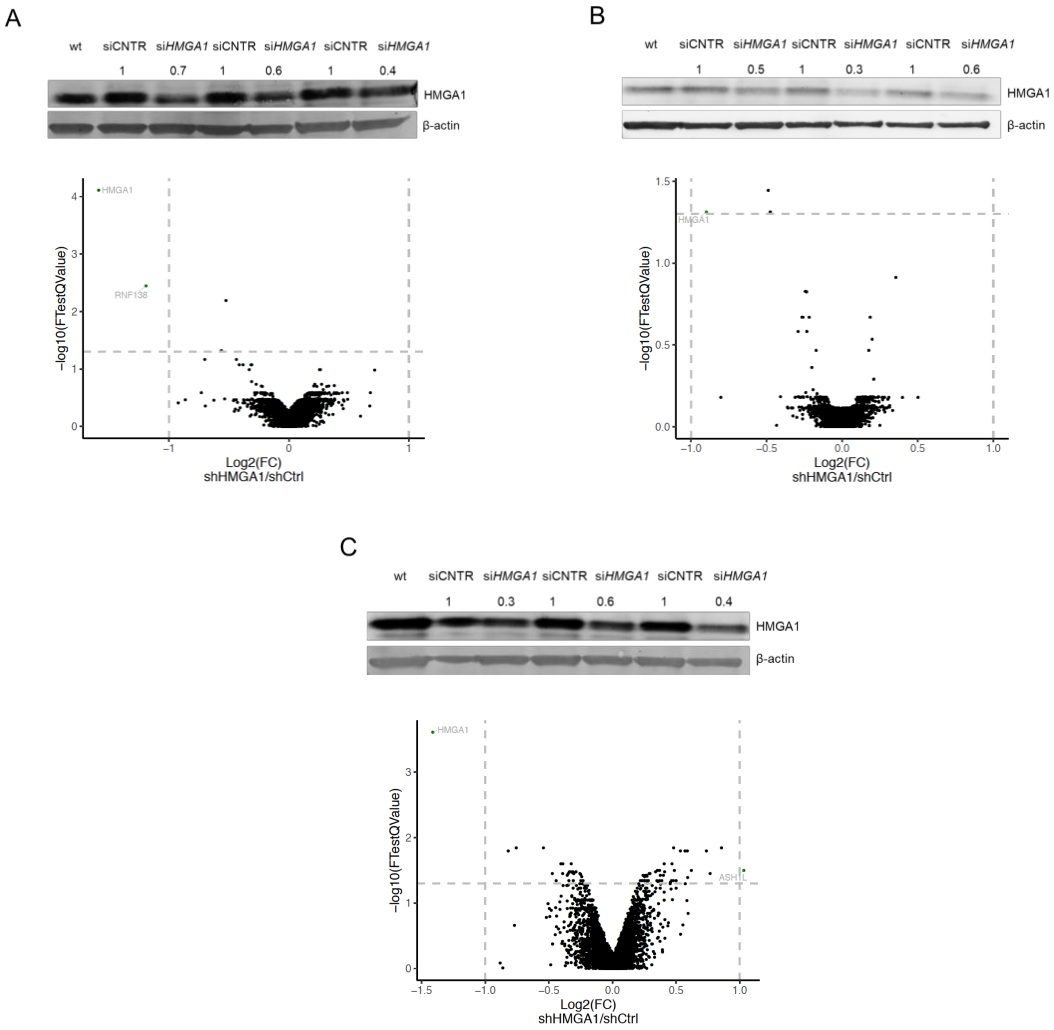


Figure 3.2.7: *Response of cell lines (SNU449, SNU182, HLE) on protein level after HMGA1 silencing, compared to control.* ***A)*** *Expression of HMGA1 in SNU449 wild type (wt) and SNU449 after transfection with the control pool of siRNA (siCNTRL) and the HMGA1 pool of siRNA (siHMGA1) at protein level (upper part). Volcano plot of Benjamini-Hochberg adjusted* P *value (Q-value, -log10scale) vs fold-change (log2scale) generated for HMGA1 silenced cells vs control. Each dot represents a protein element. Dashed lines represent the thresholds used and only significant elements are labelled and coloured (lower part).* ***B)*** *Expression of HMGA1 in SNU182 wild type (wt) and SNU182 after transfection with the control pool of siRNA (siCNTRL) and the HMGA1 pool of siRNA (siHMGA1) at protein level (upper part). Volcano plot of Benjamini-Hochberg adjusted* P *value (Q-value, -log10scale) vs fold-change (log2scale) generated for HMGA1 silenced cells vs control. Each dot represents a protein element. Dashed lines represent the thresholds used and only significant elements are labelled and coloured (lower part).* ***C)*** *Expression of HMGA1 in HLE wild type (wt) and HLE after transfection with the control pool of siRNA (siCNTRL) and the HMGA1 pool of siRNA (siHMGA1) at protein level (upper part). Volcano plot of Benjamini-Hochberg adjusted* P *value (Q-value, -log10scale) vs fold-change (log2scale) generated for HMGA1 silenced cells vs control. Each dot represents a protein element. Dashed lines represent the thresholds used and only significant elements are labelled and coloured (lower part).*

III. Identification of molecular partners of HMGA1

For an accurate molecular characterisation of biological systems, it is critical to decipher the dynamics of protein interactions in complex networks. To elucidate the role of HMGA1, we used HMGA1 immunoprecipitation followed by mass-spectrometry (IP-MS) to reveal its interacting partners. The immunoprecipitation was performed on three HCC cell lines with high levels of endogenous HMGA1 (SNU449, SNU182, HLE) and these samples underwent MS analysis (Figure 3.2.8A). We performed three independent replicates and an equal number of matched IgG control samples were prepared. The IP-MS method identified 95 proteins in the SNU449, 189 in the SNU182 and 166 in the HLE cell line with positive enrichment at q-value of 0.01. HMGA1 was one of the two most significantly enriched proteins in all three cell lines identified with 12 unique peptides. The portion of proteotypic peptides associated with each protein that are observed by MS analysis are indicators of the accuracy of the prediction. HMGA1 protein, with its ~100 aa, was detected with high accuracy without distinction between its isoforms. We found the known co-regulators (NPM1 [167] and several members of the histone H1 family [168,169]) and several new binding factors with significant enrichment in the HMGA1 IP samples compared to the controls (Figure 3.2.8B). We identified 71 proteins in common between the three cell lines (Table 3.2.2). Notably 41% (n= 29) were ribosomal proteins, 11% (n=8) were histone proteins and the remaining ones were mostly nuclear proteins involved in different networks, such as RNA binding or maturation of ribosomes. The KEGG pathway analysis and the Gene Ontology enrichment analysis performed on the common set of HMGA1 binding proteins reveal a putative function of HMGA1 in protein translation, showing also the highest enrichment for RNA and rRNA binding in addition to chromatin and DNA binding, as already well known (Figure 3.2.9).

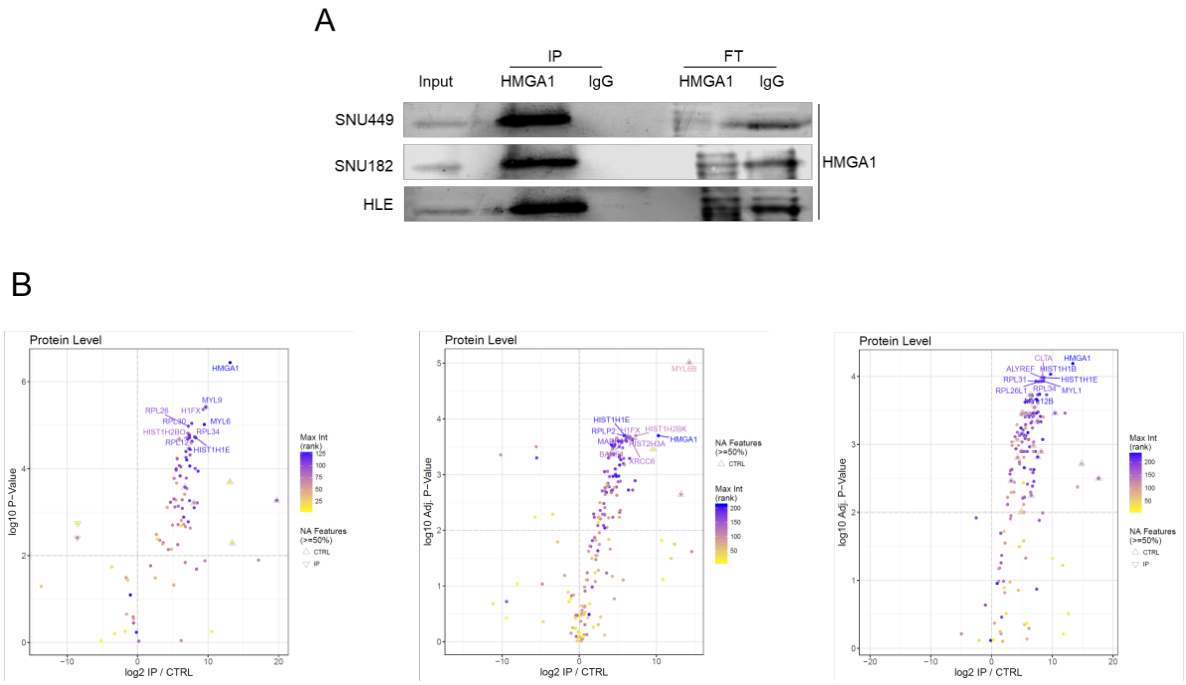
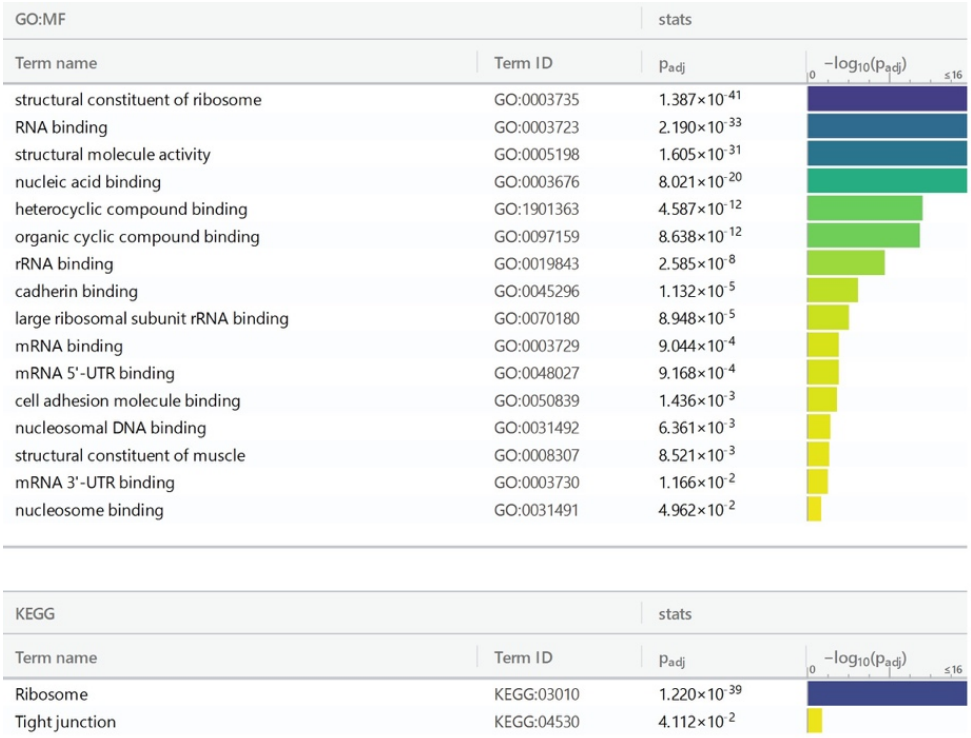


Figure 3.2.8: *Identification of interaction partners of HMGA1 by IP-MS.* ***A)*** *Immunoprecipitation of endogenous HMGA1 in the 3 cell lines (SNU449, SNU182, HLE). Immunoprecipitation with HMGA1 antibody or IgG was performed. IgG pull-down was used for IP control. Inputs and Flow-Through (FT) were investigated in parallel. The inputs, IP and FT fractions were immunoblotted with HMGA1 antibody.* ***B)*** *Volcano plot of Benajmini-Hochberg adjusted* P *value (Q-value, log10scale) vs fold-change (log2scale) generated for IP with HMGA1 antibody vs control (IP with IgG) in SNU449, SNU182 and HLE, respectively, showing the quantitative results of the IP-MS. HMGA1 and nine of its most significantly enriched interactors are labelled. The colour gradient (Max Int: maximum intensity) refers to an intensity ranking of the displayed proteins, from the blue ones with highest and the yellow ones with lowest intensity in the experiment. More than 50% of the quantified features imputed, not based on a real measurement due to missing values (NA Features), are represented with an up-pointing triangle for the control condition and a down-pointing triangle for the immunoprecipitation.*

Ribosomal proteins	Histone proteins	Myosin proteins	Remaining
RPL12	H1F0	MYH9	ACTB
RPL13	H1FX	MYL12B	ALYREF
RPL14	H2AFZ	MYL6	BANF1
RPL21	HIST1H1B	MYL9	C11orf98
RPL22	HIST1H1C		CAVIN1
RPL23	HIST1H1E		CCL17
RPL23A	HIST1H2BC		CHTOP
RPL24	HIST1H4A		FLG2
RPL26			GRN
RPL27			IGHG1
RPL28			MRPL11
RPL30			HSPA9
RPL31			HNRNPA0
RPL32			HNRNPA1
RPL34			HNRNPC
RPL35			HRNR
RPL35A			NCL
RPL36			NPM1
RPL37A			PLEC
RPL7A			PRSS1
RPL8			RBMX
RPLP2			SERBP1
RPS11			SRP14
RPS14			SSBP1
RPS17			STAU1
RPS19			TMOD3
RPS23			TPM1
RPS25			TPM3
RPS7			YBX3
			YWHAZ

Table 3.2.2: *Interaction partners of HMGA1 by IP-MS in common between the three cell lines (SNU449, SNU182, HLE).*



GO:MF		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
structural constituent of ribosome	GO:0003735	1.387×10^{-41}	
RNA binding	GO:0003723	2.190×10^{-33}	
structural molecule activity	GO:0005198	1.605×10^{-31}	
nucleic acid binding	GO:0003676	8.021×10^{-20}	
heterocyclic compound binding	GO:1901363	4.587×10^{-12}	
organic cyclic compound binding	GO:0097159	8.638×10^{-12}	
rRNA binding	GO:0019843	2.585×10^{-8}	
cadherin binding	GO:0045296	1.132×10^{-5}	
large ribosomal subunit rRNA binding	GO:0070180	8.948×10^{-5}	
mRNA binding	GO:0003729	9.044×10^{-4}	
mRNA 5'-UTR binding	GO:0048027	9.168×10^{-4}	
cell adhesion molecule binding	GO:0050839	1.436×10^{-3}	
nucleosomal DNA binding	GO:0031492	6.361×10^{-3}	
structural constituent of muscle	GO:0008307	8.521×10^{-3}	
mRNA 3'-UTR binding	GO:0003730	1.166×10^{-2}	
nucleosome binding	GO:0031491	4.962×10^{-2}	

KEGG		stats	
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16
Ribosome	KEGG:03010	1.220×10^{-39}	
Tight junction	KEGG:04530	4.112×10^{-2}	

Figure 3.2.9: *Identification of relations between the common set of HMGA1 binding proteins and their activity. The molecular function by Gene Ontology (GO:MF; upper panel) shows the activities at molecular level performed by the binding partners of HMGA1 in common between the 3 HCC cell lines. KEGG pathway enrichment analysis (bottom panel) shows the interaction networks between the common binding partners of HMGA1. Image obtained with g:Profiler.*

IV. HMGA1 and the translational regulation

We decided then to focus our attention on these intriguing results about HMGA1 interaction proteins involved in translational processes. We noticed that there were only 9 proteins in common between the three cell lines if isolating the most significantly enriched 50 proteins in the IP-MS analysis. These common results are histone and ribosomal proteins, with the exception of Alyref. Alyref is a protein involved in the control of the transcription and in mRNA transport and stabilization [170-173]. It has been shown that Alyref binds RNAs because it is able to recognise their 5-methylcytosine (m5C) modifications [174]. This m5C modification, as the other post-transcriptional RNA modifications discovered, can affect mRNA metabolism [175-178]. m5C was first identified in stable and highly abundant tRNAs and rRNAs, but in the last decades many other sites on coding and non-coding RNAs have been discovered [179-181]. m5C promotes mRNA export thanks to the regulatory proteins NSUN2 (“writer”, methyltransferase) and Alyref (“reader”, RNA binding and chaperone protein) [174]. The idea that HMGA1 could be involved in the translation and mRNA binding is entirely new, but it was supported by the finding as common interaction partners in the three cell lines of other translational regulators: CHTOP - a paralog of Alyref part of the TREX

complex [174,182], YBX3 - a RNA-binding protein that regulates transport and abundance of mRNAs [183], and HNRNPA0 - another RNA-binding protein involved in mRNA metabolism and transport [184,185]. Furthermore, both YBX and hnRNP families consist of “readers” of RNA modifications; YBX1 is another m5C reader [186], while HNRNPG interacts with m6A modified RNAs [187]. After verification by IP of the direct binding between HMGA1 and Alyref (Figure 3.2.10A), we considered the possibility that HMGA1 could be involved in the coordination of the processing of mRNAs into mature proteins as a consequence of export regulation when binding Alyref. This could also imply that HMGA1 might be partially cytoplasmic. Therefore, we first investigated HMGA1 expression in the nucleus and in the cytoplasm by immunoblotting of subcellular fractions and immunohistochemistry (IHC) (Figure 3.2.10B, C). In contrast to previous reports, we found evidence of the presence of HMGA1 in the cytoplasm. Additionally, HMGA1 cytoplasmic expression differences between cell lines found by immunoblotting matched with the levels shown by IHC.

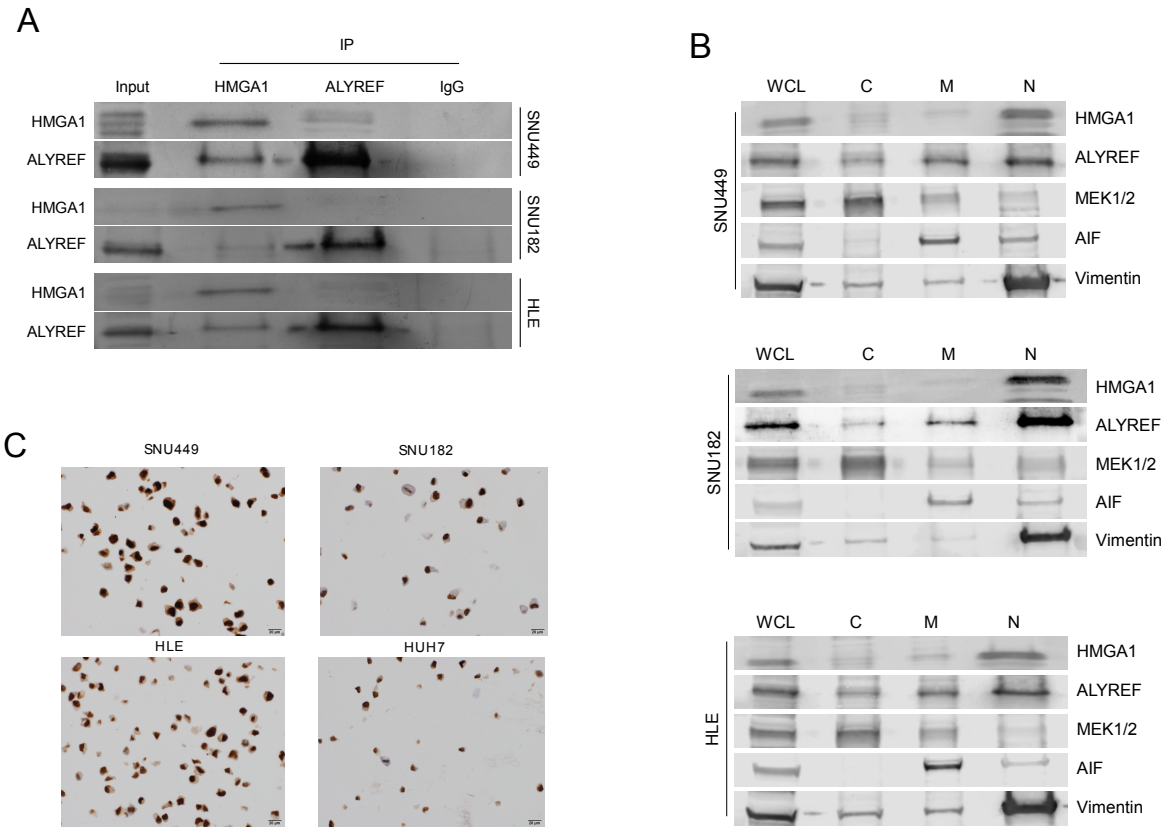


Figure 3.2.10: *Cytoplasmic presence of HMGA1* ***A)*** *Immunoprecipitation (IP) of endogenous HMGA1 and Alyref in the 3 HCC cell lines (SNU449, SNU182, HLE) with HMGA1 or Alyref pull-down. IgG pull-down was used for control IP. The IP fractions and inputs were immunoblotted with HMGA1 and Alyref antibodies.* ***B)*** *Western blot analysis of cell fractions in the 3 cell lines showing cytoplasmic (C), membrane/organellular (M) and nuclear/cytoskeletal (N) localization. Whole cell lysates (WCS) were used to represent total protein abundance. In addition to HMGA1 and Alyref, Mek1/2, AIF and Vimentin antibodies were used as markers for cytoplasm, mitochondria and cytoskeleton, respectively.* ***C)*** *Immunohistochemistry of HMGA1 on 4 HCC cell lines (SNU449, SNU182, HLE and HUH7, as control). Nuclear staining is high in all cell lines but HUH7, cytoplasmic presence of HMGA1 is noticeable in the SNU449, and to a smaller extent in the other cell lines.*

V. Alyref and HMGA1

At this point we were evaluating the possibility to examine a new putative role of HMGA1 that, while binding Alyref, might regulate comprehensively the activation of oncogenes in HCC cell lines involved in a mechanism of m5C-mediated export (Figure 3.2.11A). A second hypothesis considers the idea that only the binding with some specific common targets could be better stabilised when HMGA1 binds Alyref; HMGA1 is therefore involved in a target specific regulation (Figure 3.2.11B). These hypotheses open a new panel of questions to be answered. First of all, it is important to investigate if the Alyref shuttling is regulated by HMGA1. Furthermore, it would

be interesting to explore if the binding of Alyref on m5C is HMGA1-dependent and thus if HMGA1 is involved in m5C mRNA export regulation. We already started investigating our hypothesis checking the role of HMGA1 in the regulation of the nuclear-cytoplasmic shuttling of Alyref. We compared the expressions of both proteins at the nuclear and cytoplasmic levels in endogenous conditions and after HMGA1 silencing in each of the three HCC cell lines used for the previous analysis (SNU449, SNU182, HLE) (Figure 3.2.12). Despite the substantial efficiency of HMGA1 downregulation, we did not find considerable differences in Alyref cellular distribution when HMGA1 was silenced for 48 and 72h. This led us to be more inclined in our second hypothesis: HMGA1 could stabilise only a few mRNA targets of Alyref involved in the tumourigenesis. To verify this hypothesis, omics techniques are needed. The next steps foresee a RNA bisulfite sequencing analysis (RNA-BisSeq), to analyse Alyref mRNA targets showing a significant decrease in m5C methylation following HMGA1 silencing, and/or RNA immunoprecipitation sequencing (RIP-seq) with Alyref pull down in normal vs HMGA1 silenced conditions.

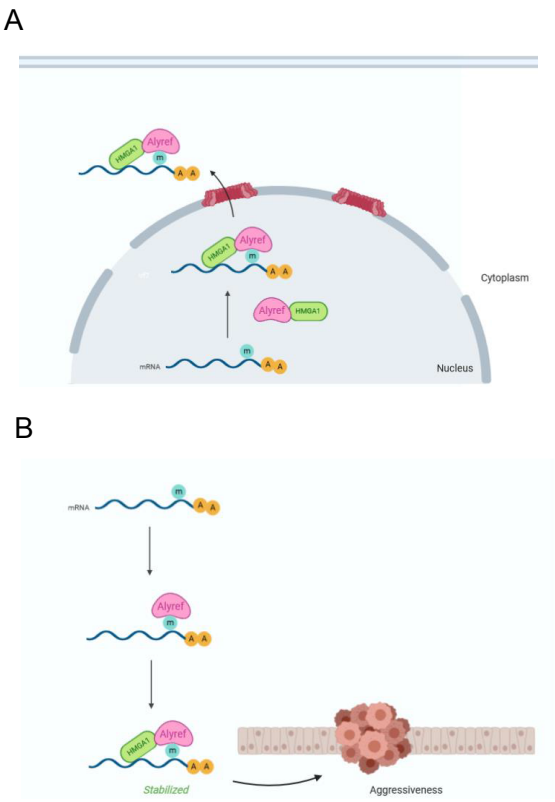


Figure 3.2.11: *HMGA1 and Alyref interactions hypothesis* ***A)*** *Broad regulation of HMGA1 on Alyref targets. HMGA1 regulates nuclear-cytoplasmic shuttling and m5C mediated RNA-binding ability of Alyref. This promotes the translation of oncogenes and their consecutive activation.* ***B)*** *Alyref targets' specific regulation of HMGA1. HMGA1 stabilises specific m5C-mRNAs targets of Alyref to promote their translation and activation, as well as aggressiveness in case of targeted oncogenes.*

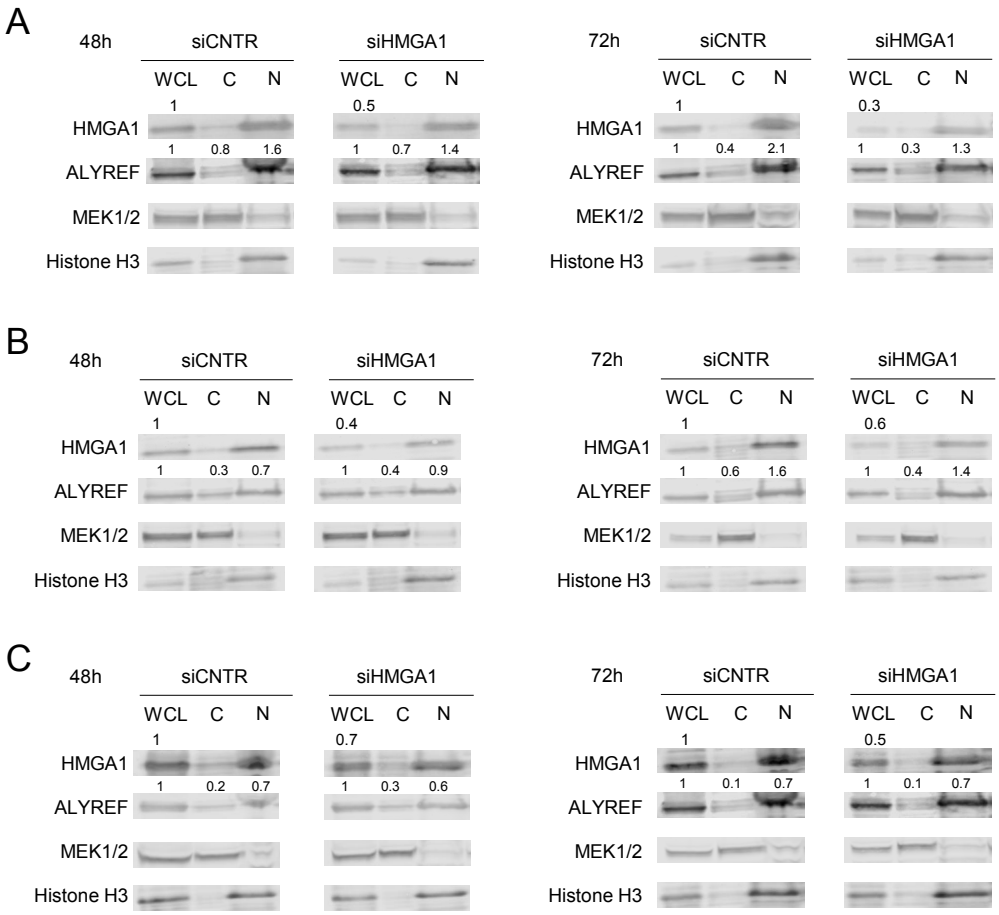


Figure 3.2.12: *HMGA1 and Alyref localization in HCC cells. Western blot analysis of cell fractions in the 3 cell lines (SNU449* ***A)****, SNU182* ***B)****, HLE* ***C)****) showing cytoplasmic (C) and nuclear/cytoskeletal (N) localization after transfection with the control pool of siRNA (siCNTR) and the HMGA1 pool of siRNA (siHMGA1). Whole cell lysates (WCS) were used to represent total protein abundance. In addition to HMGA1 and Alyref, Mek1/2 and Histone H3 antibodies were used as markers for cytoplasm and nucleus, respectively. The analysis was performed at 48h and 72h after transfection with siRNAs.*

Discussion

Although most patients with HCC present high expression levels of HMGA1 that correlate with advanced disease, tumour progression and invasion, the mechanistic explanations of its role in tumourigenesis are poorly understood. In this study, we examined the binding profile of this architectural transcription factor along the genome and its expression signature in both RNA and protein levels in HCC *in vitro* models using several omics techniques (ChIP-seq, RNA-seq, MS). The AT-dependence in the binding of HMGA1 on the DNA revealed a high abundance of detected peaks all over the genome, not only close to transcription start sites, and explains the overall lack of focal binding. However, the KEGG pathway analysis on the genes nearest to HMGA1 peaks revealed an overall enrichment of binding sites for HMGA1 in proximity to unconventional genes involved in the EMT. The role of HMGA1 in this process has already been established in the literature *in vitro* and *in vivo* in other types of cancer [188,189].

We also evaluated the expression signature of deregulated HMGA1 in HCC cell lines at the transcription and translation levels. We found few significant changes in the altered conditions vs the control. One of the explanations could be due to the limitations of the model used. With transient transfections we can identify molecular changes happening in a short range of time. In case of complex mechanisms to study, as the examination of dysregulated HMGA1 expression signatures, a stable transfection might have been a more appropriate method to capture the overall changes. At the beginning of this project we indeed planned to carry our study using HMGA1 stable clones. The overexpressing cell lines, chosen after four weeks of genetic selection with Geneticin antibiotic, presented a higher level of HMGA1 but only for a few passages. A stable overexpression of HMGA1 in HCC cell lines induced stress responses and the cells were dying not long after. We tried to optimise the generation of stable cell lines with changes of the concentration levels of the antibiotic, medium conditions and transfection methods. Nonetheless, we always observed the same ending point. We concluded that overexpression of HMGA1 for a long period is not feasible for the survival of these HCC cell lines. We also tried to generate stable clones with knock-down of HMGA1 using a well-known short hairpin RNA construct, already present in the literature for HMGA1 silencing in glioblastoma and colon cancer cell lines [143,190,191]. Despite several attempts, HMGA1 was not silenced successfully and we hypothesised that this kind of construct used for RNA interference is not functional in HCC cell lines. For this reason we proceeded with a transient transfection of HMGA1 in all previous experiments, well aware of the limitations of the method.

Consequently, we also cannot exclude the possibility that the few molecular changes detected by RNA-seq and MS between different conditions could not be a reflection of the reality. It might be

possible that other molecular changes could be detected past the investigated time points. In our experiments we proceeded with the analysis of the expression signature after 48 hours of dysregulated HMGA1. This was the first of the investigated time points where we obtained ~50% either increase or decrease in the amount of HMGA1 compared to its endogenous levels.

Our last examination was on the molecular partners binding HMGA1. The relevance of our study is underlined by the discovery of Alyref and other direct targets of HMGA1 involved in the regulation of the translation and the RNA binding. Alyref is a m5C reader that promotes mRNA export [174] and allows the translation of its targets. This could lead to a new putative role of HMGA1 in the post-transcriptional RNA modification mechanisms that affect mRNA metabolism. Furthermore, most of the studies on HMGA1 have focused on its nuclear accumulation and therefore its role in the nucleus. Our results, instead, showed how HMGA1 is also present in the cytoplasm of HCC cell lines. The implications of the presence and the role of HMGA1 in other compartments of the cells, aside from the nucleus, have been investigated in recent years. HMGA1 has been found to be extracellular in breast tumour invasive cells [192]. The extracellular fraction of HMGA1 has been found to be secreted through non-canonical secretion and mediates migration and invasion in the extracellular space with the activation of pERK signaling pathway. The novelty of this kind of research makes us realise how much is still undiscovered about the HMGA1 mechanistic process in tumourigenesis.
Our future goals would be to identify HMGA1-Alyref targets and to examine the molecular mechanism used by this complex to mediate mRNA export in the cytoplasmic space to stabilise oncogenic targets that will alter cell pathways promoting growth and invasiveness of the tumour.

4- Discussions and Outlook

The death rate of hepatocellular carcinoma (HCC) is rising faster both in European and American countries in the last decades and is already a burden for Asian countries [79]. Surgery, radiofrequency ablation and radiation therapy are the most common strategies used in case of a diagnosis of an early stage HCC. Today, targeted therapies for HCC are used when cancer is in a late stage, and they are mostly adopted to improve patients' lives. The goal of new treatments is to focus on slowing the growth and relieving symptoms to improve quality of life. Few targeted therapies that demonstrated a survival benefit are currently in use in the clinic, including regorafenib, cabozantinib, and ramucirumab alongside the most investigated protein kinase inhibitor, the drug sorafenib, but the results remain modest [76]. One of the reasons is that only parts of the molecular mechanisms responsible for the HCC tumourigenesis are well known. Despite some of these processes share genetic and molecular features with other types of tumours, there are some characteristics that are specific for this organ, as the transformation of hepatocytes. The screening for genes with specific HCC alterations can increase the survival rate in HCC patients; not only because of tumour surveillance but also because it can facilitate the discovery of new molecular biomarkers. The exploration and characterisation of putative biomarkers are the major answers for our urgent need to improve our current treatment possibilities for HCC patients.

I. Clinical screening of mutations in HCC: Considerations

As already discussed, at the moment the best way to increase the survival rate in patients with HCC is the surveillance. However, the effectiveness of HCC surveillance in clinical practice is limited also by poor HCC specificity of existing commercial sequencing panels. We designed an amplicon-based sequencing panel specifically to screen for somatic mutations and copy number alterations in HCC. The most significant variants in HCC, confirmed through our own study [193], included *TERT* promoter mutation, *TP53*, *CTNNB1*, *ALB*, *AXIN1*, *RB1* and chromatin remodelling genes (*ARID1A, ARID2, BAP1*). All these variants are all included in our panel, as well as those not currently covered by commercially available ones. We tested the sequencing panel by using biopsies (fresh-frozen, formalin-fixed paraffin-embedded (FFPE) materials) and liquid biopsy (plasma-derived cell-free DNA (cfDNA)) to evaluate the feasibility of the panel in routine diagnostics. We also designed a somatic mutation calling pipeline, PipeIT, that is practical and simple to use and ensures reproducible results in any laboratory.

Among the limitations of the study already discussed in the manuscripts, we need to consider the importance of methylation changes on the side of genetic alterations. Epigenetic altered pathways may be a consequence of aging process, persistent viral infection and chronic inflammation. They

are characterised by three main mechanisms: DNA hypermethylation leading to gene inactivation, DNA hypomethylation causing genomic instability, histone modifications affecting chromatin conformation [194]. DNA methylation occurs in different stages of liver disease (non-cirrhosis, cirrhosis and HCC) and both DNA hypomethylation and CpG hypermethylation are the dominant event during HCC development and progression [194,195]. Therefore, epigenetic changes may serve as indicators or biomarkers for screening of patients with an increased risk for HCC and they are not possible to be detected with a genetic sequencing panel.

However, we were able to identify somatic mutations from plasma-derived cfDNA, even without prior knowledge of the alteration pattern in the HCCs. This is a great outcome considering the potential of liquid biopsy. Biopsies in HCCs are rarely carried out. Thus, in patients not eligible for tumour resection, meaning patients with high disease burden, tumour materials are usually unavailable for molecular profiling. The possibility to capture nearly as many mutations as in primary tumour biopsy using only cfDNA profiling may be highly beneficial in the choice of therapeutic strategies for HCC patients.

The advantage of the adoption in diagnostic routine analysis of an accessible, cost-effective, reliable sequencing panel to use with all kinds of patients' biopsies is the possibility to accelerate the identification and validation of biomarkers to create a more personalised and optimised therapy based on putative oncogenic drivers.

II. HMGA1 study: limitations and outlooks

The increasing interest in identifying new putative biomarkers made relevant the exploration of the role of HMGA1 in HCC. HMGA1 is a protein not only highly involved in the chromatin network, but also connected with a huge number of other macromolecular complexes with diverse roles. These complexes enhance the expression of genes that are involved in several biological processes, from embryogenesis to virus integration to neoplastic transformation. HMGA1 protein is expressed in embryonic cells and in several tumour tissues but is absent in normal adult tissues [110,196]. Among the tumours with the majority of cases displaying a high expression of HMGA1 there is also HCC [91,121,152].The oncogenic mechanism of HMGA1 has been demonstrated to be mostly due to its ability to modulate chromatin structure and to bind to different transcription factors, facilitating the expression of genes involved in tumour progression and metastasis. However, the multi-facety of this protein has not yet been comprehensively revealed.

We investigated the potential oncogenic role of HMGA1 in HCC and we identified its molecular targets using *in vitro* models. Our study has several limitations. The use of cell lines may fail to recapitulate key features of HCC, taking into account also interactions between cells, three-dimensional tumour architecture and cellular heterogeneity. The recent development of organoid technology might overcome these limitations. Organoids are superior in maintaining cancer tissue architecture and they allow differentiation of tissue stem cells into functional organ-like structures [197]. Tumour organoids from biopsy of HCC patients have been established [198] and the creation of an organoids biobank containing patients with high and low levels of HMGA1 in our laboratory is one of our outlook. This model might better represent the patients' features and it might be of great interest to analyse the expression signatures and differences, for example possible divergent epithelial to mesenchymal transition (EMT) and stemness characteristics.

In my study all the observations were made using transient transfection approaches. This may be as well a limitation. In contrast to other types of cancer cell lines, HCC cells are less prone to grow and survive with dysregulated levels of HMGA1. Especially for the overexpression, we noticed that they do not survive after 3-4 passages. HMGA1 overexpression might result in a saturation of the system that does not allow us to observe drastic differences in the functionality of this protein. Also the silencing could be sensitive to several constraints and parameters. Changes in the availability of HMGA1 only for a certain amount of time could alter the potency of its target, it might be that specific conditions of abundance must be met for changes to affect silenced HMGA1-mediated targets. For this reason the tumourigenic role of HMGA1 in the hepatocellular cells might not be accentuated with the temporary differential expression of the gene. Also in this case the establishment of organoids biobank and the characterisation of HMGA1 in more patients' representative models might help the exploration of the function of the protein in HCC.

As already mentioned before, it has been shown that HCCs harbour a multitude of epigenetic aberrations, in conjunction with the genetic ones, involved in the process of liver carcinogenesis [199,200]. Thus, to analyse the epigenomics of hepatocellular cells might also help to have a deeper understanding of the genomic targets of HMGA1. We investigated the genome-wide DNA binding profile of HMGA1 and we demonstrated the protein preference for AT-rich regions. However, the enrichment in a broad range of higher AT content binding sites, and not specifically in regulatory regions, did not help to highlight the targets of HMGA1. To acquire information about heterochromatic regions compatibility and chromatin modifications across the genome when HMGA1 is deregulated, we already started to assess its genome-wide chromatin accessibility. An initial test on two HCC cell lines with silenced HMGA1 vs control by ATAC-seq (Assay for transposase-accessible chromatin with high-throughput sequencing) is ongoing. We are excited

to explore the results and, with a settled preparation protocol, we will proceed with more specimens.

Finally, our results highlight the potential relevance of a non-canonical role of HMGA1 binding mRNAs. The direct binding with Alyref and other proteins involved in translational regulation shows a new aspect of HMGA1. The proteins complex can lead to the stabilization and therefore the regulation of specific m5C-mRNAs targets of Alyref to promote their translation and activation, ending in promotion of specific tumour features. The concept and the discovery of atypical roles of HMGA1 is not completely new. Aside from the study already mentioned showing an extracellular role of secreted HMGA1 in triple negative breast cancer cells promoting tumour progression [192], another recent research described the binding of HMGA1 with RNAs in breast cancer cells [201]. This research group showed that HMGA1 is able to increase mRNA level of estrogen receptor alpha (ERα) because of its ability to induce exon-skipping. Furthermore, they confirm the participation of HMGA1 in alternative splicing because they demonstrated that the HMGA1 RNA decoy inhibits ERα46 expression and increases cell viability, sensitises the cells to tamoxifen and induces tumour formation *in vivo* [202]. The same group had recently shown the aberrant exon-skipping caused by HMGA1-RNA-protein complex on target sites adjacent to authentic 5' splice sites. This causes the overexpression of genes involved in neuronal cells degeneration of patients with sporadic Alzheimer's disease [203]. New experiments adequate to discover the atypical role of HMGA1-Alyref complex in stabilization of mRNA targets, like RNA immunoprecipitation sequencing (RIP-seq), could add another piece in the knowledge of the role of HMGA1 in tumourigenesis.

III. Conclusions

HCC is an aggressive and deadly type of cancer, with an incidence rate increasing every year also in western countries [79]. A better understanding of the mechanisms underlying the development and progression of the tumour could be a great starting point in the development of preventive strategies and specific targeted therapies. The main research objective of my project was to investigate biomarkers in HCC.

In **Chapter I**, this aim was followed based on the genetics of the disease. The possibility to screen patients' samples for genetic alterations can give us an overview of the genes dysregulated and help us to focus our attention on the appropriate and advantageous targets. We developed a HCC specific sequencing panel, with the most common somatic mutations and copy number alterations (CNAs) evaluated in HCC. It was tested for different kinds of patients' biopsies, frozen tissues,

FFPEs and also liquid biopsies. We created a somatic variant calling pipeline specific for these sequencing data as well, to have a reliable and reproducible analysis in each laboratory.

We also explored the role of one of the biomarkers of prognosis and survival of liver carcinogenesis. In **Chapter II**, we focused on the molecular characterisation of HMGA1 and the identification of its targets. We deregulated the protein in a HCC *in vitro* environment and we evaluated the DNA-binding landscape and its expression signature at RNA and protein level. We also identified the binding partners of HMGA1 and we recognised several RNA regulators, including Alyref. This discovery opens new research possibilities to seek a non-canonical role of HMGA1 binding mRNAs and involved in translational regulation.

The steps collected with my study are small but they underline the necessity to keep researching on markers in HCC and to define their clinical significance. Furthermore, it is important to create a method to evaluate the genetic landscape of a patient's tumour that may be feasible for diagnostic routine practice. This may lead in the nearest future to a deeper knowledge of the tumourigenicity of HCC and to more individualised treatment approaches.

Bibliography

1. IARC data, International Agency for Research on Cancer. http://gco.iarc.fr/today/home.
2. Weinberg & A., R. *The Biology of Cancer: Second International Student Edition*. (W.W. Norton & Company, 2013).
3. Gupta, G. P. & Massagué, J. Cancer metastasis: building a framework. *Cell* **127**, 679–695 (2006).
4. Stratton, M. R., Campbell, P. J. & Andrew Futreal, P. The cancer genome. *Nature* vol. 458 719–724 (2009).
5. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331**, 1553–1558 (2011).
6. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* vol. 144 646–674 (2011).
7. Cheng, N., Chytil, A., Shyr, Y., Joly, A. & Moses, H. L. Transforming growth factor-beta signaling-deficient fibroblasts enhance hepatocyte growth factor signaling in mammary carcinoma cells to promote scattering and invasion. *Mol. Cancer Res.* **6**, 1521–1533 (2008).
8. Bhowmick, N. A., Neilson, E. G. & Moses, H. L. Stromal fibroblasts in cancer initiation and progression. *Nature* **432**, 332–337 (2004).
9. Burkhart, D. L. & Sage, J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nature Reviews Cancer* vol. 8 671–682 (2008).
10. Deshpande, A., Sicinski, P. & Hinds, P. W. Cyclins and cdks in development and cancer: a perspective. *Oncogene* vol. 24 2909–2915 (2005).
11. Blasco, M. A. Telomeres and human disease: ageing, cancer and beyond. *Nat. Rev. Genet.* **6**, 611–622 (2005).
12. Shay, J. W. & Wright, W. E. Hayflick, his limit, and cellular ageing. *Nat. Rev. Mol. Cell Biol.* **1**, 72–76 (2000).
13. Artandi, S. E. & DePinho, R. A. Telomeres and telomerase in cancer. *Carcinogenesis* **31**, 9–18 (2010).

14. Adams, J. M. & Cory, S. The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene* vol. 26 1324–1337 (2007).
15. Junttila, M. R. & Evan, G. I. p53 — a Jack of all trades but master of none. *Nature Reviews Cancer* vol. 9 821–829 (2009).
16. Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432**, 307–315 (2004).
17. Teng, M. W. L., Swann, J. B., Koebel, C. M., Schreiber, R. D. & Smyth, M. J. Immune-mediated dormancy: an equilibrium with cancer. *J. Leukoc. Biol.* **84**, 988–993 (2008).
18. Kim, R., Emi, M. & Tanabe, K. Cancer immunoediting from immune surveillance to immune escape. *Immunology* **121**, 1–14 (2007).
19. Greten, F. R. & Grivennikov, S. I. Inflammation and Cancer: Triggers, Mechanisms, and Consequences. *Immunity* **51**, 27–41 (2019).
20. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, inflammation, and cancer. *Cell* **140**, 883–899 (2010).
21. Baeriswyl, V. & Christofori, G. The angiogenic switch in carcinogenesis. *Seminars in Cancer Biology* vol. 19 329–337 (2009).
22. Raica, M., Cimpean, A. M. & Ribatti, D. Angiogenesis in pre-malignant conditions. *European Journal of Cancer* vol. 45 1924–1934 (2009).
23. Jones, R. G. & Thompson, C. B. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes Dev.* **23**, 537–548 (2009).
24. DeBerardinis, R. J., Lum, J. J., Hatzivassiliou, G. & Thompson, C. B. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.* **7**, 11–20 (2008).
25. Kennedy, K. M. & Dewhirst, M. W. Tumor metabolism of lactate: the influence and therapeutic potential for MCT and CD147 regulation. *Future Oncology* vol. 6 127–148 (2010).
26. Feron, O. Pyruvate into lactate and back: From the Warburg effect to symbiotic energy fuel exchange in cancer cells. *Radiotherapy and Oncology* vol. 92 329–333 (2009).
27. Semenza, G. L. Tumor metabolism: cancer cells give and take lactate. *The Journal of*

clinical investigation vol. 118 3835–3837 (2008).

28. Cavallaro, U. & Christofori, G. Cell adhesion and signalling by cadherins and Ig-CAMs in cancer. *Nature Reviews Cancer* vol. 4 118–132 (2004).
29. Talmadge, J. E. & Fidler, I. J. AACR Centennial Series: The Biology of Cancer Metastasis: Historical Perspective. *Cancer Research* vol. 70 5649–5669 (2010).
30. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability — an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology* vol. 11 220–228 (2010).
31. Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol. Cell* **40**, 179–204 (2010).
32. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* vol. 461 1071–1078 (2009).
33. Friedberg, E. C. *et al.* DNA repair: From molecular mechanism to human disease. *DNA Repair* vol. 5 986–996 (2006).
34. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
35. Sirica, A. E. Tumor Progression and the Clonal Evolution of Neoplasia. *The Pathobiology of Neoplasia* 217–229 (1989) doi:10.1007/978-1-4684-5523-6_11.
36. Alberts, B. *et al.* Molecular Biology of the Cell. (2007) doi:10.1201/9780203833445.
37. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **173**, 1823 (2018).
38. Rubin, A. F. & Green, P. Comment on 'The consensus coding sequences of human breast and colorectal cancers'. *Science* vol. 317 1500 (2007).
39. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
40. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
41. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes.

Nature **578**, 102–111 (2020).

42. International Cancer Genome Consortium *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
43. Priestley, P. *et al.* Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
44. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).
45. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
46. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology* vol. 15 81–94 (2018).
47. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
48. Beroukhim, R., Meyerson, M., Garraway, L. & Prensner, J. Abstract 5759: The landscape of copy-number changes across multiple human cancer types. *Cellular and Molecular Biology* (2010) doi:10.1158/1538-7445.am10-5759.
49. Chen, Y., Widschwendter, M. & Teschendorff, A. E. Systems-epigenomics inference of transcription factor activity implicates aryl-hydrocarbon-receptor inactivation as a key event in lung cancer development. *Genome Biol.* **18**, 236 (2017).
50. Zheng, S. C., Widschwendter, M. & Teschendorff, A. E. Epigenetic drift, epigenetic clocks and cancer risk. *Epigenomics* **8**, 705–719 (2016).
51. Prior, I. A., Lewis, P. D. & Mattos, C. A comprehensive survey of Ras mutations in cancer. *Cancer Res.* **72**, 2457–2467 (2012).
52. Muñoz-Maldonado, C., Zimmer, Y. & Medová, M. A Comparative Analysis of Individual RAS Mutations in Cancer Biology. *Front. Oncol.* **9**, 1088 (2019).
53. Gayther, S. A. *et al.* Regionally clustered APC mutations are associated with a severe phenotype and occur at a high frequency in new mutation cases of adenomatous polyposis coli. *Human Molecular Genetics* vol. 3 53–56 (1994).

54. Esteller, M. *et al.* Analysis of adenomatous polyposis coli promoter hypermethylation in human cancer. *Cancer Res.* **60**, 4366–4371 (2000).
55. Garinis, G. A., Patrinos, G. P., Spanakis, N. E. & Menounos, P. G. DNA hypermethylation: when tumour suppressor genes go silent. *Hum. Genet.* **111**, 115–127 (2002).
56. Huang, M., Shen, A., Ding, J. & Geng, M. Molecularly targeted cancer therapy: some lessons from the past decade. *Trends Pharmacol. Sci.* **35**, 41–50 (2014).
57. DiMasi, J. A. & Grabowski, H. G. Economics of New Oncology Drug Development. *Journal of Clinical Oncology* vol. 25 209–216 (2007).
58. Bai, X. *et al.* CMTTdb: the cancer molecular targeted therapy database. *Ann Transl Med* **7**, 667 (2019).
59. Li, Y. H. *et al.* Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* **46**, D1121–D1127 (2018).
60. Meng, L. *et al.* Use of Exome Sequencing for Infants in Intensive Care Units: Ascertainment of Severe Single-Gene Disorders and Effect on Medical Management. *JAMA Pediatr.* **171**, e173438 (2017).
61. Gaffney, E. F., Riegman, P. H., Grizzle, W. E. & Watson, P. H. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. *Biotechnic & Histochemistry* vol. 93 373–386 (2018).
62. Gao, X. H. *et al.* Comparison of Fresh Frozen Tissue With Formalin-Fixed Paraffin-Embedded Tissue for Mutation Analysis Using a Multi-Gene Panel in Patients With Colorectal Cancer. *Front. Oncol.* **10**, 310 (2020).
63. Wimmer, I. *et al.* Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fixed paraffin-embedded tissue samples. *Sci. Rep.* **8**, 6351 (2018).
64. Siravegna, G., Marsoni, S., Siena, S. & Bardelli, A. Integrating liquid biopsies into the management of cancer. *Nat. Rev. Clin. Oncol.* **14**, 531–548 (2017).
65. Crowley, E., Di Nicolantonio, F., Loupakis, F. & Bardelli, A. Liquid biopsy: monitoring

cancer-genetics in the blood. *Nature Reviews Clinical Oncology* vol. 10 472–484 (2013).

66. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
67. Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci. Transl. Med.* **7**, 302ra133 (2015).
68. Center for Devices & Radiological Health. The therascreen PIK3CA RGQ PCR Kit - P190001 and P190004. https://www.fda.gov/medical-devices/recently-approved-devices/therascreen-pik3ca-rgq-pcr-kit-p190001-and-p190004 (2019).
69. Malapelle, U. *et al.* Profile of the Roche cobas® EGFR mutation test v2 for non-small cell lung cancer. *Expert Review of Molecular Diagnostics* vol. 17 209–215 (2017).
70. Meijer, G. A. GLOBOCAN 1: Cancer Incidence and Mortality Worldwide. : By J Ferlay, D M Parkin, P Pisani. ($90.00.) International Agency for Research on Cancer, 1998. *Journal of Clinical Pathology* vol. 53 164–a (2000).
71. Bosch, F. X. & Ribes, J. The epidemiology of primary liver cancer: global epidemiology. *Viruses and Liver Cancer* 1–16 (2002).
72. Kumar, M., Zhao, X. & Wang, X. W. Molecular carcinogenesis of hepatocellular carcinoma and intrahepatic cholangiocarcinoma: one step closer to personalized medicine? *Cell Biosci.* **1**, 5 (2011).
73. El–Serag, H. B. & Lenhard Rudolph, K. Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. *Gastroenterology* vol. 132 2557–2576 (2007).
74. Global Burden of Disease Cancer Collaboration *et al.* Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol* **3**, 524–548 (2017).
75. Yang, J. D. *et al.* A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 589–604 (2019).
76. Yarchoan, M. *et al.* Recent Developments and Therapeutic Strategies against Hepatocellular Carcinoma. *Cancer Res.* **79**, 4326–4330 (2019).

77. Llovet, J. M., Burroughs, A. & Bruix, J. Hepatocellular carcinoma. *Lancet* **362**, 1907–1917 (2003).

78. Davila, J. A., Morgan, R. O., Shaib, Y., McGlynn, K. A. & El–Serag, H. B. Hepatitis C infection and the increasing incidence of hepatocellular carcinoma: A population-based study. *Gastroenterology* vol. 127 1372–1380 (2004).

79. El-Serag, H. B. & Kanwal, F. Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go? *Hepatology* **60**, 1767–1775 (2014).

80. World Health Organization. Hepatitis B vaccines: WHO position paper, July 2017 - Recommendations. *Vaccine* **37**, 223–225 (2019).

81. Colombo, M. & Lleo, A. The impact of antiviral therapy on hepatocellular carcinoma epidemiology. *Hepat Oncol* **5**, HEP03 (2018).

82. Chiang, C.-J. *et al.* Significant reduction in end-stage liver diseases burden through the national viral hepatitis therapy program in Taiwan. *Hepatology* **61**, 1154–1162 (2015).

83. Kao, J.-H. Hepatitis B vaccination and prevention of hepatocellular carcinoma. *Best Pract. Res. Clin. Gastroenterol.* **29**, 907–917 (2015).

84. Bruno, S. *et al.* Sustained virological response to interferon-alpha is associated with improved outcome in HCV-related cirrhosis: a retrospective study. *Hepatology* **45**, 579–587 (2007).

85. Veldt, B. J. *et al.* Sustained virologic response and clinical outcomes in patients with chronic hepatitis C and advanced fibrosis. *Ann. Intern. Med.* **147**, 677–684 (2007).

86. Nagata, H. *et al.* Effect of interferon-based and -free therapy on early occurrence and recurrence of hepatocellular carcinoma in chronic hepatitis C. *J. Hepatol.* **67**, 933–939 (2017).

87. Dore, G. J. & Feld, J. J. Hepatitis C virus therapeutic development: in pursuit of 'perfectovir'. *Clin. Infect. Dis.* **60**, 1829–1836 (2015).

88. Feld, J. J. & Foster, G. R. Second generation direct-acting antivirals - Do we expect major improvements? *J. Hepatol.* **65**, S130–S142 (2016).

89. Llovet, J. M. *et al.* Hepatocellular carcinoma. *Nat Rev Dis Primers* **2**, 16018 (2016).

90. Llovet, J. M., Villanueva, A., Lachenmayer, A. & Finn, R. S. Advances in targeted therapies for hepatocellular carcinoma in the genomic era. *Nat. Rev. Clin. Oncol.* **12**, 436 (2015).
91. Guichard, C. *et al.* Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet.* **44**, 694–698 (2012).
92. Cleary, S. P. *et al.* Identification of driver genes in hepatocellular carcinoma by exome sequencing. *Hepatology* **58**, 1693–1702 (2013).
93. Cancer Genome Atlas Research Network. Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327–1341.e23 (2017).
94. Villanueva, A., Newell, P., Chiang, D. Y., Friedman, S. L. & Llovet, J. M. Genomics and signaling pathways in hepatocellular carcinoma. *Semin. Liver Dis.* **27**, 55–76 (2007).
95. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764 (2012).
96. Nault, J. C. *et al.* High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. *Nat. Commun.* **4**, 2218 (2013).
97. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
98. Killela, P. J. *et al.* TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proceedings of the National Academy of Sciences* vol. 110 6021–6026 (2013).
99. Hartmann, D. *et al.* Telomerase gene mutations are associated with cirrhosis formation. *Hepatology* **53**, 1608–1617 (2011).
100. Hoshida, Y. *et al.* Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.* **69**, 7385–7392 (2009).
101. Boyault, S. *et al.* Transcriptome classification of HCC is related to gene alterations and to

new therapeutic targets. *Hepatology* **45**, 42–52 (2007).

102. Schulze, K. *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).

103. Ahn, S.-M. *et al.* Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. *Hepatology* **60**, 1972–1982 (2014).

104. Totoki, Y. *et al.* High-resolution characterization of a hepatocellular carcinoma genome. *Nature Genetics* vol. 43 464–469 (2011).

105. Friedmann, M., Holth, L. T., Zoghbi, H. Y. & Reeves, R. Organization, inducible-expression and chromosome localization of the human HMG-I(Y) nonhistone protein gene. *Nucleic Acids Res.* **21**, 4259–4267 (1993).

106. Cleynen, I. & Van de Ven, W. J. M. The HMGA proteins: a myriad of functions (Review). *Int. J. Oncol.* **32**, 289–305 (2008).

107. Peluso, S. & Chiappetta, G. High-Mobility Group A (HMGA) Proteins and Breast Cancer. *Breast Care* vol. 5 81–85 (2010).

108. Fusco, A. & Fedele, M. Roles of HMGA proteins in cancer. *Nature Reviews Cancer* vol. 7 899–910 (2007).

109. Reeves, R. High mobility group (HMG) proteins: Modulators of chromatin structure and DNA repair in mammalian cells. *DNA Repair* vol. 36 122–136 (2015).

110. Chiappetta, G. *et al.* High level expression of the HMGI (Y) gene during embryonic development. *Oncogene* **13**, 2439–2446 (1996).

111. Abe, N. *et al.* Diagnostic Significance of High Mobility Group I(Y) Protein Expression in Intraductal Papillary Mucinous Tumors of the Pancreas. *Pancreas* vol. 25 198–204 (2002).

112. Chiappetta, G. *et al.* High mobility group HMGI(Y) protein expression in human colorectal hyperplastic and neoplastic diseases. *International Journal of Cancer* vol. 91 147–151 (2001).

113. Piscuoglio, S. *et al.* HMGA1 and HMGA2 protein expression correlates with advanced tumour grade and lymph node metastasis in pancreatic adenocarcinoma. *Histopathology* **60**, 397–404 (2012).

114. Chiappetta, G. *et al.* HMGA1 protein overexpression in human breast carcinomas: correlation with ErbB2 expression. *Clin. Cancer Res.* **10**, 7637–7644 (2004).

115. Hristov, A. C. *et al.* HMGA1 correlates with advanced tumor grade and decreased survival in pancreatic ductal adenocarcinoma. *Mod. Pathol.* **23**, 98–104 (2010).

116. Masciullo, V. *et al.* HMGA1 protein over-expression is a frequent feature of epithelial ovarian carcinomas. *Carcinogenesis* **24**, 1191–1198 (2003).

117. Zhang, Z., Wang, Q., Chen, F. & Liu, J. Elevated expression of HMGA1 correlates with the malignant status and prognosis of non-small cell lung cancer. *Tumour Biol.* **36**, 1213–1219 (2015).

118. Franco, R. *et al.* Detection of high-mobility group proteins A1 and A2 represents a valid diagnostic marker in post-pubertal testicular germ cell tumours. *J. Pathol.* **214**, 58–64 (2008).

119. Sepe, R. *et al.* HMGA1 overexpression is associated with a particular subset of human breast carcinomas. *J. Clin. Pathol.* **69**, 117–121 (2016).

120. Chen, X. Expression of the High Mobility Group Proteins HMGI(Y) Correlates with Malignant Progression in Barrett's Metaplasia. *Cancer Epidemiology Biomarkers & Prevention* vol. 13 30–33 (2004).

121. Andreozzi, M. *et al.* HMGA1 Expression in Human Hepatocellular Carcinoma Correlates with Poor Prognosis and Promotes Tumor Growth and Migration in in vitro Models. *Neoplasia* **18**, 724–731 (2016).

122. Reeves, R. & Beckerbauer, L. HMGI/Y proteins: flexible regulators of transcription and chromatin structure. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* vol. 1519 13–29 (2001).

123. Battista, S. *et al.* Loss of Hmga1 gene function affects embryonic stem cell lymphohematopoietic differentiation. *The FASEB Journal* vol. 17 1–27 (2003).

124. Zanin, R. *et al.* HMGA1 promotes breast cancer angiogenesis supporting the stability, nuclear localization and transcriptional activity of FOXM1. *Journal of Experimental & Clinical Cancer Research* vol. 38 (2019).

125. Cheng, Y., Cheng, T., Zhao, Y. & Qu, Y. HMGA1 exacerbates tumor progression by activating miR-222 through PI3K/Akt/MMP-9 signaling pathway in uveal melanoma. *Cellular Signalling* vol. 63 109386 (2019).

126. Xian, L. *et al.* HMGA1 amplifies Wnt signalling and expands the intestinal stem cell compartment and Paneth cell niche. *Nature Communications* vol. 8 (2017).

127. Penzo *et al.* HMGA1 Modulates Gene Transcription Sustaining a Tumor Signalling Pathway Acting on the Epigenetic Status of Triple-Negative Breast Cancer Cells. *Cancers* vol. 11 1105 (2019).

128. Thanos, D. & Maniatis, T. The High Mobility Group protein HMG I(Y) is required for NF-κB-dependent virus induction of the human IFN-β gene. *Cell* vol. 71 777–789 (1992).

129. Thanos, D., Du, W. & Maniatis, T. The high mobility group protein HMG I(Y) is an essential structural component of a virus-inducible enhancer complex. *Cold Spring Harb. Symp. Quant. Biol.* **58**, 73–81 (1993).

130. Chin, M. T. *et al.* Enhancement of Serum-response Factor-dependent Transcription and DNA Binding by the Architectural Transcription Factor HMG-I(Y). *Journal of Biological Chemistry* vol. 273 9755–9760 (1998).

131. Galande, S. Chromatin (dis)organization and cancer: BUR-binding proteins as biomarkers for cancer. *Curr. Cancer Drug Targets* **2**, 157–190 (2002).

132. Zhao, K., Käs, E., Gonzalez, E. & Laemmli, U. K. SAR-dependent mobilization of histone H1 by HMG-I/Y in vitro: HMG-I/Y is enriched in H1-depleted chromatin. *The EMBO Journal* vol. 12 3237–3247 (1993).

133. Sumter, T. F. *et al.* The High Mobility Group A1 (HMGA1) Transcriptome in Cancer and Development. *Curr. Mol. Med.* **16**, 353–393 (2016).

134. Bianchi, M. E. & Beltrame, M. Upwardly mobile proteins. Workshop: the role of HMG proteins in chromatin structure, gene expression and neoplasia. *EMBO Rep.* **1**, 109–114 (2000).

135. Postnikov, Y. V. & Bustin, M. Reconstitution of high mobility group 14/17 proteins into nucleosomes and chromatin. *Methods Enzymol.* **304**, 133–155 (1999).

136. Reeves, R. Structure and Function of the HMGI(Y) Family of Architectural Transcription Factors. *Environmental Health Perspectives* vol. 108 803 (2000).

137. Reeves, R. Molecular biology of HMGA proteins: hubs of nuclear function. *Gene* **277**, 63–81 (2001).

138. Thomas, J. O. HMG1 and 2: architectural DNA-binding proteins. *Biochemical Society Transactions* vol. 29 395–401 (2001).

139. Farnet, C. M. & Bushman, F. D. HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. *Cell* **88**, 483–492 (1997).

140. Mao, L. *et al.* HMGA1 levels influence mitochondrial function and mitochondrial DNA repair efficiency. *Mol. Cell. Biol.* **29**, 5426–5440 (2009).

141. Giancotti, V. *et al.* Changes in nuclear proteins on transformation of rat epithelial thyroid cells by a murine sarcoma retrovirus. *Cancer Res.* **45**, 6051–6057 (1985).

142. Sepe, R. *et al.* CBX7 and HMGA1b proteins act in opposite way on the regulation of the SPP1 gene expression. *Oncotarget* **6**, 2680–2692 (2015).

143. Puca, F. *et al.* HMGA1 silencing restores normal stem cell characteristics in colon cancer stem cells by increasing p53 levels. *Oncotarget* **5**, 3234–3245 (2014).

144. Hillion, J. *et al.* Upregulation of MMP-2 by HMGA1 promotes transformation in undifferentiated, large-cell lung cancer. *Mol. Cancer Res.* **7**, 1803–1812 (2009).

145. Takaha, N., Resar, L. M. S., Vindivich, D. & Coffey, D. S. High mobility group protein HMGI(Y) enhances tumor cell growth, invasion, and matrix metalloproteinase-2 expression in prostate cancer cells. *Prostate* **60**, 160–167 (2004).

146. Shah, S. N. *et al.* HMGA1 reprograms somatic cells into pluripotent stem cells by inducing stem cell transcriptional networks. *PLoS One* **7**, e48533 (2012).

147. Tesfaye, A. *et al.* The high-mobility group A1 gene up-regulates cyclooxygenase 2 expression in uterine tumorigenesis. *Cancer Res.* **67**, 3998–4004 (2007).

148. Baldassarre, G. *et al.* Onset of natural killer cell lymphomas in transgenic mice carrying a truncated HMGI-C gene by the chronic stimulation of the IL-2 and IL-15 pathway. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 7970–7975 (2001).

149. Xu, Y. *et al.* The HMG-I oncogene causes highly penetrant, aggressive lymphoid malignancy in transgenic mice and is overexpressed in human leukemia. *Cancer Res.* **64**, 3371–3375 (2004).

150. Fedele, M. *et al.* Transgenic mice overexpressing the wild-type form of the HMGA1 gene develop mixed growth hormone/prolactin cell pituitary adenomas and natural killer cell lymphomas. *Oncogene* **24**, 3427–3435 (2005).

151. Belton, A. *et al.* HMGA1 induces intestinal polyposis in transgenic mice and drives tumor progression and stem cell properties in colon cancer cells. *PLoS One* **7**, e30034 (2012).

152. Chang, Z.-G. *et al.* Determination of high mobility group A1 (HMGA1) expression in hepatocellular carcinoma: a potential prognostic marker. *Dig. Dis. Sci.* **50**, 1764–1770 (2005).

153. Livak, K. J. & Schmittgen, T. D. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* vol. 25 402–408 (2001).

154. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* vol. 9 R137 (2008).

155. Jalili, V., Matteucci, M., Masseroli, M. & Morelli, M. J. Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics* vol. 34 2338–2338 (2018).

156. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* vol. 29 15–21 (2013).

157. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* vol. 15 (2014).

158. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. *BioRxiv* doi:10.1101/060012.

159. Ahrné, E. *et al.* Evaluation and Improvement of Quantification Accuracy in Isobaric Mass Tag-Based Protein Quantification Experiments. *J. Proteome Res.* **15**, 2537–2547 (2016).

160. Hubner, N. C. *et al.* Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *J. Cell Biol.* **189**, 739–754 (2010).

161. Colombo, D. F., Burger, L., Baubec, T. & Schübeler, D. Binding of high mobility group A

proteins to the mammalian genome occurs as a function of AT-content. *PLOS Genetics* vol. 13 e1007102 (2017).

162. Bendris, N., Arsic, N., Lemmers, B. & Blanchard, J. M. Cyclin A2, Rho GTPases and EMT. *Small GTPases* **3**, 225–228 (2012).

163. Parri, M. & Chiarugi, P. Rac and Rho GTPases in cancer cell motility control. *Cell Communication and Signaling* vol. 8 (2010).

164. Ungefroren, H., Witte, D. & Lehnert, H. The role of small GTPases of the Rho/Rac family in TGF-β-induced EMT and cell motility in cancer. *Developmental Dynamics* vol. 247 451–461 (2018).

165. Zhang, Y.-L., Wang, R.-C., Cheng, K., Ring, B. Z. & Su, L. Roles of Rap1 signaling in tumor cell migration and invasion. *Cancer Biol Med* **14**, 90–99 (2017).

166. Ma, X.-L. *et al.* CD73 promotes hepatocellular carcinoma progression and metastasis via activating PI3K/AKT signaling by inducing Rap1-mediated membrane localization of P110β and predicts poor prognosis. *J. Hematol. Oncol.* **12**, 37 (2019).

167. Arnoldo, L. *et al.* A novel mechanism of post-translational modulation of HMGA functions by the histone chaperone nucleophosmin. *Sci. Rep.* **5**, 8552 (2015).

168. Catez, F. *et al.* Network of dynamic interactions between histone H1 and high-mobility-group proteins in chromatin. *Mol. Cell. Biol.* **24**, 4321–4328 (2004).

169. Senigagliesi, B. *et al.* The High Mobility Group A1 (HMGA1) Chromatin Architectural Factor Modulates Nuclear Stiffness in Breast Cancer Cells. *Int. J. Mol. Sci.* **20**, (2019).

170. Hung, M.-L., Hautbergue, G. M., Snijders, A. P. L., Dickman, M. J. & Wilson, S. A. Arginine methylation of REF/ALY promotes efficient handover of mRNA to TAP/NXF1. *Nucleic Acids Res.* **38**, 3351–3361 (2010).

171. Hautbergue, G. M., Hung, M.-L., Golovanov, A. P., Lian, L.-Y. & Wilson, S. A. Mutually exclusive interactions drive handover of mRNA from export adaptors to TAP. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5154–5159 (2008).

172. Stubbs, S. H. & Conrad, N. K. Depletion of REF/Aly alters gene expression and reduces RNA polymerase II occupancy. *Nucleic Acids Res.* **43**, 504–519 (2015).

173. Stubbs, S. H., Hunter, O. V., Hoover, A. & Conrad, N. K. Viral factors reveal a role for REF/Aly in nuclear RNA stability. *Mol. Cell. Biol.* **32**, 1260–1270 (2012).

174. Yang, X. *et al.* 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an mC reader. *Cell Res.* **27**, 606–625 (2017).

175. Zhao, X. *et al.* FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res.* **24**, 1403–1419 (2014).

176. Meyer, K. D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635–1646 (2012).

177. Meyer, K. D. *et al.* 5′ UTR m6A Promotes Cap-Independent Translation. *Cell* vol. 163 999–1010 (2015).

178. Yang, Y. *et al.* Dynamic m 6 A modification and its emerging regulatory role in mRNA splicing. *Science Bulletin* vol. 60 21–32 (2015).

179. Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Research* vol. 37 e12–e12 (2008).

180. Squires, J. E. *et al.* Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* **40**, 5023–5033 (2012).

181. Khoddami, V. & Cairns, B. R. Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol.* **31**, 458–464 (2013).

182. Chang, C.-T. *et al.* Chtop is a component of the dynamic TREX mRNA export complex. *EMBO J.* **32**, 473–486 (2013).

183. Cooke, A. *et al.* The RNA-Binding Protein YBX3 Controls Amino Acid Levels by Regulating SLC mRNA Abundance. *Cell Rep.* **27**, 3097–3106.e5 (2019).

184. Zhang, J. *et al.* hnRNPs and ELAVL1 cooperate with uORFs to inhibit protein translation. *Nucleic Acids Res.* **45**, 2849–2864 (2017).

185. Rousseau, S. *et al.* Inhibition of SAPK2a/p38 prevents hnRNP A0 phosphorylation by MAPKAP-K2 and its interaction with cytokine mRNAs. *EMBO J.* **21**, 6505–6514 (2002).

186. Chen, X. *et al.* 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. *Nat. Cell Biol.* **21**, 978–990 (2019).

187. Zhou, K. I. *et al.* Regulation of Co-transcriptional Pre-mRNA Splicing by mA through the Low-Complexity Protein hnRNPG. *Mol. Cell* **76**, 70–81.e9 (2019).

188. Pegoraro, S. *et al.* HMGA1 promotes metastatic processes in basal-like breast cancer regulating EMT and stemness. *Oncotarget* **4**, 1293–1308 (2013).

189. Zhong, J. *et al.* TGF-β1 induces HMGA1 expression: The role of HMGA1 in thyroid cancer proliferation and invasion. *Int. J. Oncol.* **50**, 1567–1578 (2017).

190. Puca, F. *et al.* HMGA1 negatively regulates NUMB expression at transcriptional and post transcriptional level in glioblastoma stem cells. *Cell Cycle* **18**, 1446–1457 (2019).

191. Colamaio, M. *et al.* HMGA1 silencing reduces stemness and temozolomide resistance in glioblastoma stem cells. *Expert Opin. Ther. Targets* **20**, 1169–1179 (2016).

192. Méndez, O. *et al.* Extracellular HMGA1 Promotes Tumor Invasion and Metastasis in Triple-Negative Breast Cancer. *Clin. Cancer Res.* **24**, 6367–6382 (2018).

193. Paradiso, V. *et al.* Diagnostic Targeted Sequencing Panel for Hepatocellular Carcinoma Genomic Screening. *J. Mol. Diagn.* **20**, 836–848 (2018).

194. Su, P.-F. *et al.* Differential DNA methylation associated with hepatitis B virus infection in hepatocellular carcinoma. *Int. J. Cancer* **121**, 1257–1264 (2007).

195. Tischoff, I. & Tannapfe, A. DNA methylation in hepatocellular carcinoma. *World J. Gastroenterol.* **14**, 1741–1748 (2008).

196. Rogalla, P. *et al.* HMGI-C expression patterns in human tissues. Implications for the genesis of frequent mesenchymal tumors. *Am. J. Pathol.* **149**, 775–779 (1996).

197. Clevers, H. Modeling Development and Disease with Organoids. *Cell* **165**, 1586–1597 (2016).

198. Nuciforo, S. *et al.* Organoid Models of Human Liver Cancers Derived from Tumor Needle Biopsies. *Cell Rep.* **24**, 1363–1376 (2018).

199. Liu, W.-R., Shi, Y.-H., Peng, Y.-F. & Fan, J. Epigenetics of hepatocellular carcinoma: a new horizon. *Chin. Med. J.* **125**, 2349–2360 (2012).

200. Toh, T. B., Lim, J. J. & Chow, E. K.-H. Epigenetics of hepatocellular carcinoma. *Clin. Transl. Med.* **8**, 13 (2019).

201. Ohe, K. *et al.* HMGA1a Induces Alternative Splicing of the Estrogen Receptor-αlpha Gene by Trapping U1 snRNP to an Upstream Pseudo-5′ Splice Site. *Frontiers in Molecular Biosciences* vol. 5 (2018).
202. Ohe, K. *et al.* HMGA1a induces alternative splicing of estrogen receptor alpha in MCF-7 human breast cancer cells. *J. Steroid Biochem. Mol. Biol.* **182**, 21–26 (2018).
203. Ohe, K. & Mayeda, A. HMGA1a trapping of U1 snRNP at an authentic 5’ splice site induces aberrant exon skipping in sporadic Alzheimer's disease. *Mol. Cell. Biol.* **30**, 2220–2228 (2010).

Annex

Original Study

Identification of Somatic Mutations in Thirty-year-old Serum Cell-free DNA From Patients With Breast Cancer: A Feasibility Study

Mathilde Ritter,[1,2] Viola Paradiso,[3] Patrik Widmer,[3] Andrea Garofoli,[3,4] Luca Quagliata,[3] Serenella Eppenberger-Castori,[3] Savas D. Soysal,[4] Simone Muenst,[3] Charlotte K.Y. Ng,[5] Salvatore Piscuoglio,[3,4] Walter Weber,[6] Walter P. Weber[1,2]

Abstract

The aim of this study was to assess the feasibility of cell-free DNA extraction and circulating tumor DNA sequencing in 30-year-old serum samples of patients with breast cancer. Cell-free DNA extraction was successful in 52 of 52 patients, and 24 cancer-specific mutations were found in 22 of 25 samples undergoing sequencing. This study shows that next-generation sequencing technology is sufficiently robust and specific to analyze 30-year-old serum.

Introduction: The aim of this study was to assess the feasibility of cell-free DNA (cfDNA) extraction and circulating tumor DNA sequencing in 30-year-old serum samples. **Materials and Methods:** We evaluated serum samples from 52 patients with breast cancer, which were collected between 1983 and 1991, with correlating clinicopathologic data. cfDNA was extracted by using the QIAamp Circulating Nucleic Acid Extraction Kit (Qiagen). Of these 52 cfDNA samples, 10 were randomly selected and sequenced with the Oncomine Breast cfDNA Assay (A31183). In a second step, high-depth targeted sequencing of 15 additional cfDNA samples was performed using a custom Ampliseq Ion Torrent panel targeting breast cancer-related genes. **Results:** cfDNA extraction was successful in 52 (100%) of 52 patients with a total concentration of 0.2 to 54 ng/uL. A total of 24 cancer-specific mutations were found in 22 (88%) of the 25 samples undergoing sequencing. Of the 52 patients, 32 (62%) had died from breast cancer after a median follow-up of 7.9 years (interquartile range, 3.7-15.5 years). **Conclusion:** The present study shows that current next generation sequencing technology is sufficiently robust and specific to analyze 30-year-old serum. Therefore, longitudinal studies can be designed with storage of serum samples over many years, thereby obviating the need for timely and continuous cfDNA extraction and sequencing. The samples can be pooled and processed at once with the most modern technology available at the end of the study, when accumulation of events allows correlation of clinical outcomes with adequate power.

Clinical Breast Cancer, Vol. ■, No. ■, ■-■

Keywords: Breast cancer, Circulating tumor DNA, Serum, Somatic mutations

M.R. and V.P. contributed equally to this work as first authors. W.P.W. and W.W. contributed equally to this work as last authors.

[1]Breast Center, University Hospital of Basel and University of Basel, Basel, Switzerland
[2]Department of Breast Surgery, University Hospital of Basel, Basel, Switzerland
[3]Institute of Medical Genetics and Pathology, University Hospital Basel, Basel, Switzerland
[4]Visceral Surgery Research Laboratory, Department of Biomedicine, University of Basel, Basel, Switzerland
[5]Department for BioMedical Research, University of Bern, Bern, Switzerland
[6]Clinical Cancer Etiology Unit, Basel, Switzerland

Submitted: Jan 25, 2020; Revised: Apr 4, 2020; Accepted: Apr 7, 2020

Addresses for correspondence: Mathilde Ritter, MD, Division of Breast Surgery, University Hospital Basel, Spitalstrasse 21, CH-4031, Basel, Switzerland; or Walter P. Weber, MD, Chief, Division of Breast Surgery, Breast Surgeon SSO, University Hospital Basel, Spitalstrasse 21, CH-4031, Basel, Switzerland
E-mail contact: mathilde.ritter@usb.ch; walter.weber@usb.ch

1526-8209/
http://dx.doi.org/10.1016/j.clbc.2020.04.005

Introduction

A breakthrough in next-generation sequencing (NGS) in the past decade provided an unprecedented opportunity to investigate genetic variations in humans and their roles in health and disease. In particular, large-scale efforts such as The Cancer Genome Atlas and the International Cancer Genome Consortium have provided a comprehensive molecular portrait of human cancers.[1,2] The discovery of the so-called 'driver genes' has provided the basis for the development of the concept of precision medicine, where the identification of targetable alterations guides the therapeutic approach in treating patients with cancer. Nowadays, the decreasing costs of massively parallel sequencing have resulted in increased adoption of genomic profiling as part of the standard diagnostic procedures in most tumor types.[3,4]

In patients with cancer, nucleic acids obtained from tumor biopsies and resections remain the main source for molecular profiling. However, these procedures are invasive, costly, time-consuming, and have only limited potential to be repeated in longitudinal studies.[5] Their relevance is further limited by the prevalence of intra-tumor genetic heterogeneity as shown in multiple sequencing studies over the past decade.[5,6] Thus, a single biopsy of the primary tumor is not likely to be genetically representative of the whole tumor. To overcome these challenges, circulating cell-free DNA (cfDNA) has been proposed as an alternative because it can be collected less invasively compared with conventional biopsies.[7,8] Circulating cfDNA is a type of cell-free nucleic acid that derives from apoptotic and necrotic cells or is released from living eukaryotic cells.[9] The detection of DNA in the blood originating from tumors in patients with cancer has been described decades ago.[10-12] The fraction of cfDNA derived from tumor is termed circulating tumor DNA (ctDNA).[7,8] ctDNA can be considered a new source for the detection and surveillance of major cancers because it is more likely to be present in patients with cancer.[7,8]

The potential of using cfDNA as an indicator of disease burden with prognostic implication and clinical applicability during follow-up and monitoring in both the curative and palliative setting has been investigated in numerous studies.[13,14] Cancer-specific mutations, copy number alterations, and genomic rearrangements assessed in ctDNA demonstrated potential prognostic and predictive significance.[15-20] To further evaluate prognostic and predictive biomarkers in cfDNA and assess its value as a disease monitoring tool, longitudinal studies with long follow-up are necessary. The utility of the technology depends on its capability to assess sequential samples that have been collected and stored over a long period of time. This study aims to assess the feasibility of cfDNA extraction and somatic mutation assessment in 30-year-old serum that has been collected from patients with breast cancer between 1983 and 1991.

Materials and Methods

Patients and Serum

For this study, we had access to serum samples from 753 patients with cancer, which were collected between 1983 and 1991 in an oncologic private practice in Basel, Switzerland. Of 753 patients, 152 were females with breast cancer. The patients were referred to the medical oncologist either after surgery of the primary tumor or after the diagnosis of local/regional recurrence and/or distant metastases. After obtaining informed consent, 10 mL of native venous blood were collected in a 10-mL BD Vacutainer blood collection tube and centrifuged in a Hettich centrifuge at 5000 rpm for 10 minutes. The serum samples were immediately frozen and stored at −70°C to −80°C in 3 Nunc Cryogenic tubes per patient (Gibko AG) at a private office during the first 9 years; thereafter, the samples have been transferred to the Institute of Immunobiology in Freiburg, Germany, by using transportable refrigerating boxes to avoid thawing. In 1999, the samples were relocated to the Laboratory for Medical Genetics of the University of Basel, Switzerland, and stored until processing and analysis. Clinicopathologic variables regarding patient demographics, primary tumor, treatment, recurrence, and survival were retrieved from clinical files. Approval for the use of these samples and correlating data has been granted by the responsible ethics committee (approval number: eknz-2018-00252).

cfDNA Extraction

Circulating DNA was extracted from 2 to 4 mL of isolated serum from 52 randomly selected patients with breast cancer with the QIAamp Circulating Nucleic Acid Kit (Qiagen) as previously described.[21] DNA was quantified using the Qubit Fluorometer (Invitrogen) and analyzed using the 2200 TapeStation system (Agilent Technologies) with the High Sensitivity DNA Analysis Kit.

Targeted Sequencing and Library Preparation

Sequencing was performed using 2 different amplicon-based targeted sequencing panels. The first 10 randomly selected samples were sequenced with the Oncomine Breast cfDNA Assay (A31183, Thermo Fisher Scientific). This panel covers 152 hotspot mutations in 10 genes (*AKT1*, *EGFR*, *ERBB2*, *ERBB3*, *ESR1*, *FBXW7*, *KRAS*, *PIK3CA*, *SF3B1*, and *TP53*) across 26 amplicons. This integrates the TagSeq technology (molecular barcode) and allows detection of rare variants present at 0.1% allelic frequency. Library preparation, molecular barcoding, and sequencing were performed according to the instructions and guidelines provided by Thermo Fisher, using 5 ng of DNA as input. Briefly, the library preparation protocol was based on a 2-step cycle multiplex touchdown polymerase chain reaction (PCR) with a temperature ranging from 64 °C to 58 °C, which allowed to amplify target regions and to introduce unique molecular identifiers. The obtained tagged amplicons of around 100 to 140 bp length were then cleaned up using Agencourt AMPure XP (Beckman Coulter), then eluted in 24 µl low TE buffer. A second round of PCR (18 cycles) was performed in a total volume of 50 µl to amplify the purified amplicons and to introduce Ion Torrent Tag-Sequencing adapters containing sample-specific barcodes. The resulting library of target DNA fragments was purified by performing a 2-step cleanup using Agencourt AMPure XP (Beckman Coulter). The purified libraries were then diluted 1:1000 and quantified by qPCR using the Ion Universal Quantitation Kit (Thermo Fisher Scientific). The quantified stock libraries were then diluted to 100 pM for downstream template preparation. Subsequentially, sequencing runs were planned on the Torrent Suite Software v5.2, and libraries were pooled and loaded on an Ion 540 chip using the Ion Chef Instrument (Thermo Fisher Scientific). The loaded chip was then sequenced using 500 flows on an S5 system (Thermo Fisher Scientific).

Table 1 Clinicopathologic Parameters of 52 Patients With Breast Cancer

Survival and Recurrences	n (%) or Median (IQR)
No. patients with clinical data and serum	52
Median age at first diagnosis, y	49.5 (45.5-60.5)
Median follow-up time from diagnosis to death or last follow-up, y	7.9 (3.7-15.5)
Median time from initial diagnosis to date of sample collection, y	1.6 (1.0-4.8)
No. breast cancer-specific deaths	32 (62)
No. deaths unrelated to breast cancer	4 (7)
Cause of death unknown	2 (3)
Median overall survival from diagnosis to death, y	6.8 (3.2-13.9)
Median disease-free survival diagnosis to local/regional or distant recurrence, second breast cancer or death, y	2.7 (1.5-6.4)
Treatment at first diagnosis	
Neoadjuvant treatment	2 (3)
Surgery	50 (96)
Adjuvant radiation	18 (35)
Adjuvant tamoxifen	18 (35)
Adjuvant CMF (±v ±p) or LMF (±vp) chemotherapies	15 (28)
Other adjuvant systemic therapies	3 (5)
Clinicopathological parameters at first diagnosis	
Female	52 (100)
Laterality	
Left	25 (48)
Right	24 (47)
Bilateral	3 (5)
Grade[a]	
1	1 (2)
2	14 (27)
3	37 (71)
Hormone receptor status[a]	
Positive	42 (82)
Negative	10 (18)
T stage	
1	14 (27)
2	28 (54)
3	5 (10)
4	4 (7)
X	1 (2)
N stage	
0	16 (31)
1	26 (50)
2	6 (11)
3	1 (2)
X	3 (6)
M stage	
0	50 (97)
1	2 (3)

Abbreviations: C = cyclophosphamide; F = fluorouracil; IQR = interquartile range; L = chlorambucil (leukeran); M = methotrexate; P = prednisone; V = vincristine.
[a]Had not been assessed routinely at the time.

In a second round, 15 randomly selected samples were sequenced with a custom targeted sequencing panel focusing on the most frequently altered genes in breast cancer previously described.[22] Library preparation for the breast panel was performed using the Ion AmpliSeq library kit 2.0 (Thermo Fisher Scientific) according to the manufacturer's guidelines. The panel consists of 2 pools of amplification primers. Ten ng of DNA per sample were used for library preparation for each pool. Amplification was performed according to the manufacturer's guidelines. The amplicons from the 2 pools were combined and treated to digest the primers and to phosphorylate the amplicons. The amplicons were then ligated to Ion Xpress Barcode Adapters (Thermo Fisher Scientific) using DNA ligase. Finally, cleaning and purification of the generated libraries were performed with Agencourt AMPure XP (Beckman Coulter) according to the manufacturer's guidelines. Quantification and quality control were performed with Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific). Samples were diluted to reach the concentration of 40 pmol and then were pooled for sequencing. Twenty-five µl of the pooled libraries were loaded on Ion 540 Chip (Thermo Fisher Scientific) and processed in Ion Chef Instrument (Thermo Fisher Scientific). Sequencing was performed on Ion S5XL system (Thermo Fisher Scientific).[23]

Somatic Variants Identification

Raw data were processed automatically on the Torrent Server and aligned to the reference hg19 genome. The analysis pipeline included signal processing, base calling, quality score assignment, adapter trimming, PCR duplicate removal, and control of mapping quality. All samples passed the quality check and met the requirements of a minimum molecular average depth. The first round of samples sequencing data (n = 10) was uploaded in BAM format to the Ion Reporter Analysis Server for variant calling and annotation. Variant calling was performed on Ion Reporter (IR) Analysis Software v5.2 using the Oncomine TagSeq Breast Liquid Biopsy w2.0 workflow. Coverage metrics for each amplicon were obtained by running the Coverage Analysis Plugin software v5.2.1 (Thermo Fisher Scientific). Identified variants were only considered if the variant had a molecular coverage of at least 3, indicating that the variant was detected in 3 independent template molecules. Finally, all candidate mutations were manually reviewed using the Integrative Genomics Viewer37.

Mutation Identification in 30-year-old Serum

In the second round of sequenced samples (n = 15), variant calling was performed with TVC version 5.0.3 (Torrent Variant Caller, Thermo Fisher Scientific) using low-stringency parameters previously described.[24,25] Briefly, mutations detected by TVC were subsequently filtered by the following steps. First, all the multiallelic variants have been split and left aligned. Moreover, the presence and the relative length of homopolymer sequences were annotated to take into account the presence of possible wrongly aligned sequencing reads and, therefore, false-positive variants. Second, because the 15 samples had no matched germline samples, all the variants have been annotated using 3 databases: the 1000 Genomes Project, the Exome Aggregation Consortium, and the NHLBI GO Exome Sequencing Project.[26,27] All the mutations identified by TVC that were also present within the databases in significative frequencies (> 5%) have been flagged as probable germline mutations. Furthermore, a pool of 16 germline samples collected from an independent cohort was used to provide an additional list of likely germline mutations that, together with the ones previously flagged, have been filtered out from the final output list. To avoid the removal of clinically relevant information, mutations found in known cancer driver hotspots have been whitelisted and kept even when they met the criteria for the aforementioned filtering steps.

Results

cfDNA Extraction From 30-year-old Serum of Sufficient Quality for Sequencing Analysis

We randomly selected 52 of the 152 patients with breast cancer (clinicopathologic characteristics of patients are shown in Table 1) to perform cfDNA extraction (Table 2). cfDNA extraction was successful in all patients, and cfDNA levels were determined for each sample with a fluorometric quantitation system. We obtained a range of concentrations from 0.2 to 54 ng/ul (Table 2). To assess the serum-derived cfDNA integrity and quality, we performed a capillary electrophoretic separation using the TapeStation system (Agilent Technology). Electropherograms were generated for each sample and the fragment size of the cfDNA measured between 2 markers against fluorescence intensity. The mean of cfDNA fragment size distribution ranged from 106 to 216 bp (average, 136 bp), with no significant differences in cfDNA fragment size between all samples (Figure 1). Even though the serum samples showed contamination with high molecular weight genomic DNA in comparison to samples extracted from plasma (gDNA) (Figure 1), it was observed that the amount of cfDNA was more than the gDNA (Figure 1).

Taken together, cfDNA was successfully extracted from all the samples with sufficient quality for further sequencing analysis, suggesting the feasibility of the use of long-storage serum for molecular analysis.

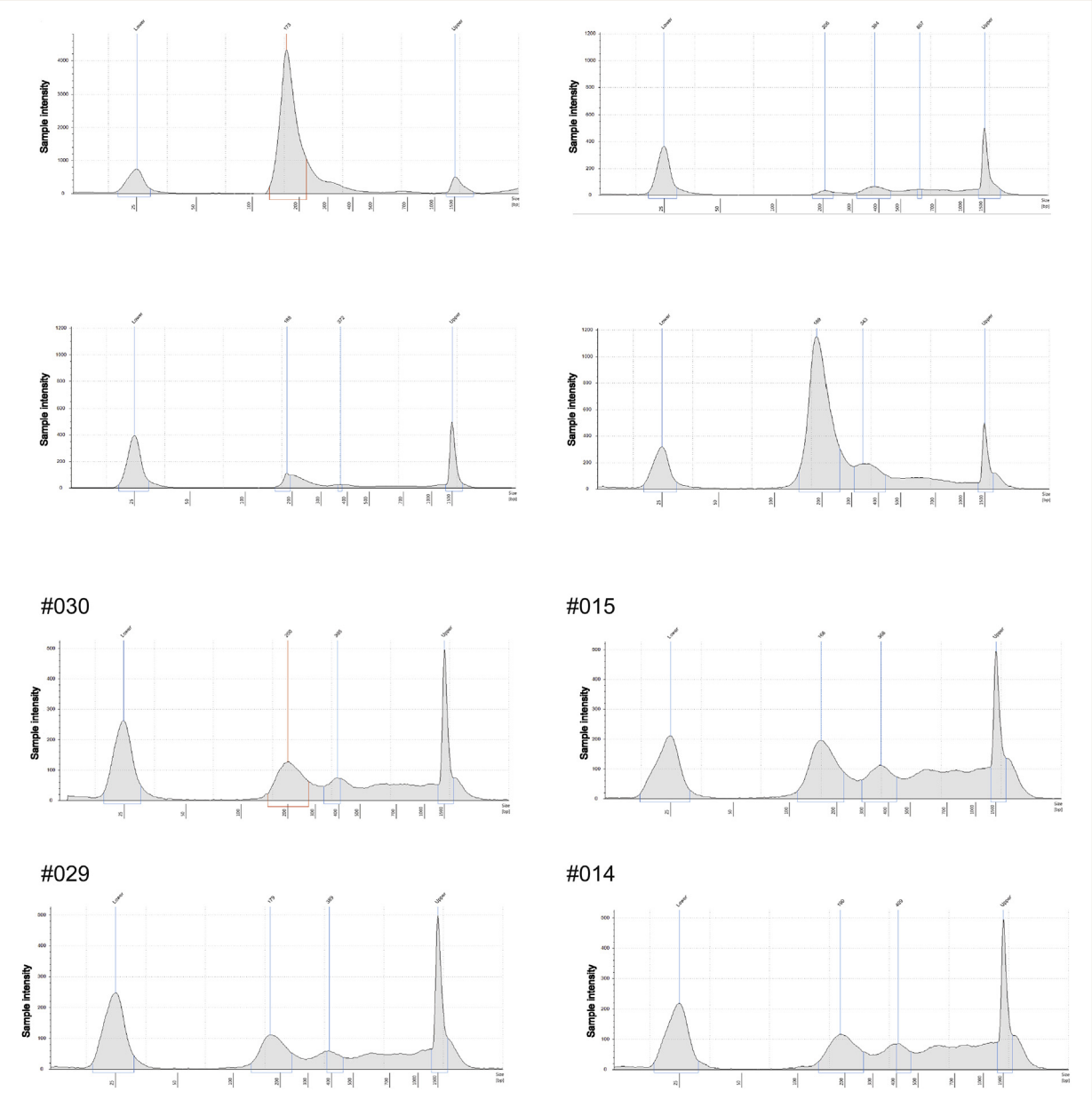
Targeted Sequencing Showed Breast Cancer-specific Somatic Mutations in cfDNA

From the 52 extracted cfDNA samples, we randomly selected 25 for subsequent mutation investigation. Sequencing was performed using 2 different targeted panels. Ten samples were sequenced using the Oncomine Breast cfDNA Assay, which covers the most common hotspots in 10 highly mutated genes in breast cancer.

Table 2 Circulating Free DNA Extraction Data

Sample Name	Serum, mL	Concentration, ng/uL
#001	3	1.5
#002	4	2.5
#003	3.2	9.6
#004	3	0.6
#005	3.5	8.8
#006	3.5	10.3
#007	2.5	12.1
#008	2	1.4
#009	3	3.7
#010	3	3.4
#011	3.5	10
#012	4	10.5
#013	3.2	6.7
#014	2.5	6.5
#015	2.5	5.4
#016	3	3.7
#017	3	1.2
#018	3	0.25
#019	3.5	3.4
#020	3	3.4
#021	3.2	7.5
#022	2	2.5
#023	4	1.1
#024	3	2.5
#025	4	28
#026	3	3.9
#027	3.5	4.9
#028	4	8.1
#029	3	9.1
#030	3	2.9
#031	3	2.3
#032	3.5	1.7
#033	4	2.8
#034	3	1.4
#035	3	5.3
#036	3	2.3
#037	3	0.8
#038	2.5	7.9
#039	3	1.8
#040	3	4.2
#041	4	19.3
#042	3.5	4.2
#043	2.5	9.2
#044	2.5	17.8
#045	2.5	2.4
#046	3.5	2.7
#047	2.5	1.6
#048	4	54
#049	2.5	6.4
#050	3	1.9
#051	3	15.3
#052	2.5	0.7

Figure 1 Circulating Free DNA Analysis Using the TapeStation System. Representative Electropherograms of Total Extracted Circulating Free DNA From 4 Patients Selected for Sequencing (#030, #015, #029, #014)



Owing to the molecular barcoding, this panel allows for the identification of mutations at a very low allelic frequency. We obtained a mean sequencing depth of 55,612X (ranging from 5227 to 108,393) and identified somatic mutations in 8 of the tested samples encompassing *KRAS, TP53* (Table 3A and Figure 2). The other 15 samples were instead sequenced with a custom targeted sequencing panel that covers all exons of 27 protein-coding genes as well as mutation hotspots in 3 cancer genes and the recurrently mutated lncRNA genes *MALAT1* and *NEAT1*[28] (see Supplemental Table 1 in the online version). In this second round of sequencing, we obtained a mean sequencing depth of 2891X (ranging from 903 to 18,210), and we identified somatic mutations in 12 of the tested samples (Table 3B and Figure 2). Mutations were detected in some of the most commonly mutated genes in breast cancer, as *TP53*(p.Arg248Trp) and *PTEN*(p.Phe278Leu).

Taken together, a total of 24 cancer-specific mutations in 10 of the most commonly mutated breast cancer genes were found in 22 (88%) of 25 randomly selected 30-year-old serum-derived cfDNA from patients with breast cancer (Figure 2), suggesting the feasibility of using very old serum samples for mutational profiles.

Discussion

The present cohort of patients with breast cancer with complete long-term follow-up and available blood samples taken 30 years ago

Table 3 Depth of Sequencing

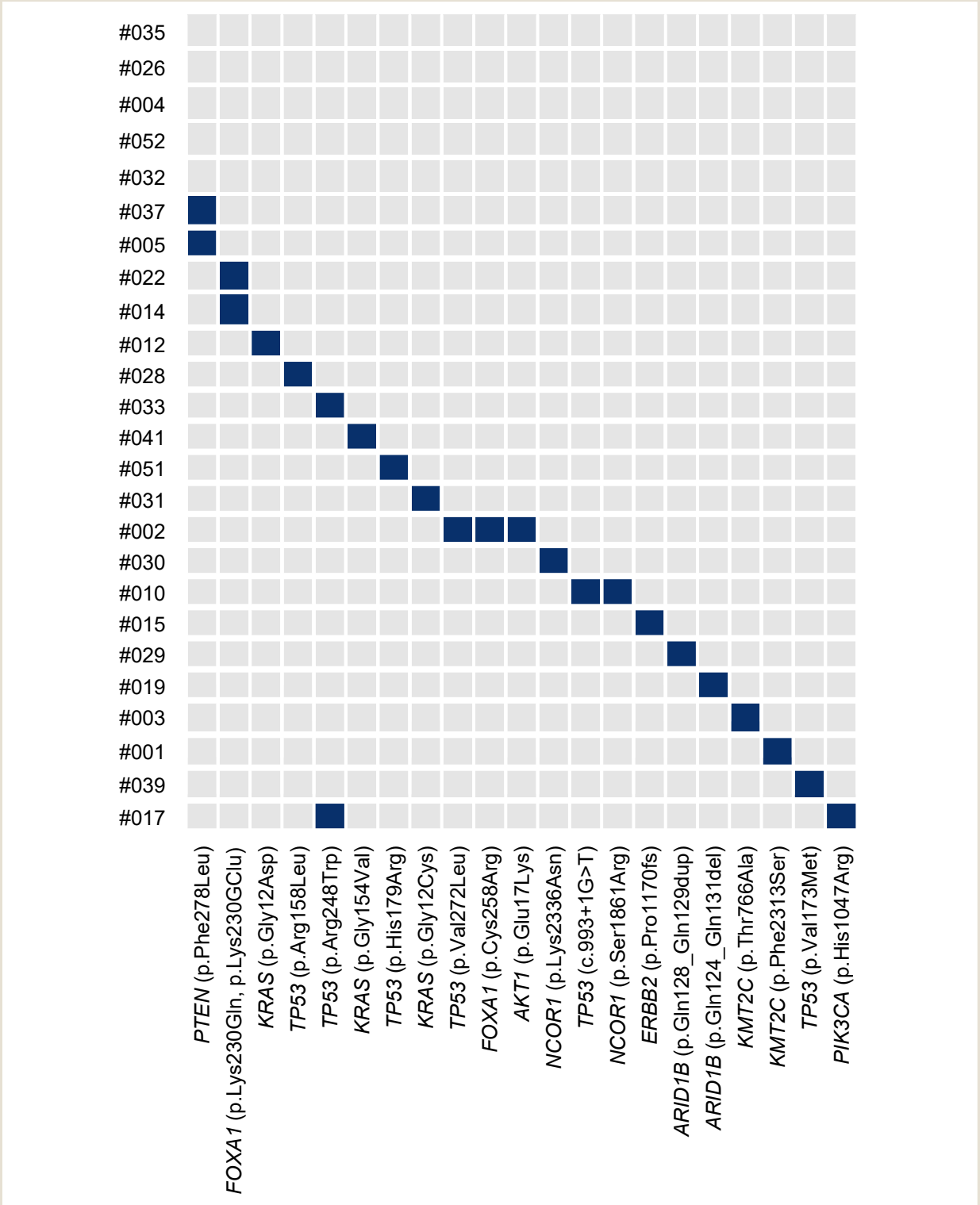
A, Sequencing with Oncomine™ Breast Circulating Free DNA Assay

Sample Name	Mapped Reads	On Target	Mean Coverage	KRAS	TP53	PIK3CA
#012	822,290	95.84	32,400	p.G12D (0.1%)		
#017	1,513,454	96.98	60,944		p.R248W (0.1%)	p.H1047R (2.84%)
#026	1,430,359	95.98	56,798			
#028	17,27,543	97.76	73,289		p.R158L (1.58%)	
#031	1,719,712	96.01	69,907	p.G12C (0.1%)		
#033	978,141	96.31	39,142		p.R248W (0.1%)	
#035	1,084,084	95.86	43,583			
#039	135,999	97.44	5227		p.V173M (0.46%)	
#041	1,626,548	96.08	66,433	p.G154V (0.06%)		
#051	2,656,056	96.21	108,393		p.H179R (0.14%)	

B, Sequencing With Custom Ampliseq Ion Torrent Panel

Sample	Mapped Reads	On Target	Mean Coverage	PTEN	TP53	AKT1	FOXA1	ARID1B	NCOR1	ERBB2	KMT2C
#001	1,695,052	91.83	904								p.Phe2313Ser (0.51)
#002	33,435,937	92.57	18,210		p.Val272Leu (0.08)	p.Glu17Lys (0.13)	p.Cys258Arg (0.4)				
#003	3,398,689	93.08	1879								p.Thr766Ala (0.51)
#004	3,357,088	92.24	1729								
#005	3,924,417	92.69	2151	p.Phe278Leu (0.25)							
#010	3,138,700	93.12	1589		c.993+1G>T (0.56)				p.Ser1861Arg (0.16)		
#014	2,284,741	93.62	1265				p.Lys230Gln (0.16)				
#015	6,629,734	94.07	3529							p.Pro1170fs (0.1)	
#019	4,542,595	93.35	2516					p.Gln124_Gln131del (0.07)			
#022	2,169,116	93.45	1160				p.Lys230Gln (0.13)				
#029	3,605,091	93.79	1975					p.Gln128_Gln129dup (0.23)			
#030	3,090,414	94.6	1606						p.Lys2336Asn (0.48)		
#032	3,743,126	91.21	2031								
#037	2,094,065	91.87	1093	p.Phe278Leu (0.27)							
#052	3,098,154	93.8	1689								

Figure 2 Heatmap of the Comparison of all the Somatic Mutations Found in 25 Randomly Selected Samples Across the Whole Cohort. Rows Indicate the Specific Samples, Columns Indicate the Specific Mutations, Annotated With the Gene Name and the Specific Aminoacidic Change. The Heatmap Shows the Presence (Blue Cells) or the Lack (Grey Cells) of Mutations



represents an exceptional opportunity to study the potential of liquid biopsy-based biomarker identification. The patients were treated in curative or palliative intent for stage I to IV breast cancer. Treatment heterogeneity was limited at that time because only tamoxifen and CMF or LMF (cyclophosphamide- or chlorambucil-methotrexate-fluorouracil) chemotherapy were used in most patients who received systemic treatment. The majority of the patient population recurred at some point, which was the reason why 35 (62%) of the 52 patients died from breast cancer in this cohort. The collection of serum started in 1983, whereas the dates of first diagnosis and treatment go back to 1967. cfDNA was obtained in sufficient quantity and quality for sequencing in all 52 patients. The sample size was too small to make any firm conclusions on differences between patients with short and long-term survival. However, the present results suggest that this modern technology can be used to accurately extract and sequence ctDNA to detect cancer-specific mutations in these old samples, despite the long cryopreservation and repeated changes of storage location.

These findings support the use of long-term storage of biological samples in longitudinal studies prior to analysis, which, in turn, will increase feasibility by making the study protocols less depending on consecutive and timely processing at the centralized high-depth targeted sequencing unit. The principle of long-term storage may facilitate the performance of large international studies that assess the prognostic role of cancer-associated pathogenic mutations in serum cfDNA present at diagnosis by comparing overall and relapse-free survival between patients with or without specific mutations. Therefore, another potential value of using samples of patients diagnosed a long time ago is to increase the number of events (eg, relapse, deaths) and increase the statistical power for survival analyses (study of the prognostic value of the identified mutations in ctDNA). The predictive power of response-associated mutations can then be assessed based on in silico mutation effect predictors and curated databases of cancer- and response-associated variants.[29-35] One would hypothesize that patients with detectable mutations in the cfDNA would have higher tumor burden and/or tumor cells with a higher tendency to shed into the bloodstream and, therefore, poorer outcome than patients without detectable mutations in the cfDNA. Candidate somatic mutations can be further evaluated in vitro and in vivo by using xenograft models. For instance, patient-derived breast tumor cells can be engineered to express the same mutation found to be associated with resistance and test their sensitivity to the same targeted therapy in xenograft models compared with control cancer cells (ie, wild-type in the corresponding allele).

The blood samples analyzed here were taken at a time when physicians could not anticipate NGS approaches. Nevertheless, substantial efforts were made to collect the samples under the assumption that someday technology would have advanced to the point where relevant research could be performed with a few mL of serum and matched clinical data. Storing blood samples over the entire duration of longitudinal studies allows newly developed technology to analyze cfDNA more thoroughly and homogenously. Hence, the most modern state-of-the-art technology for nucleic acids extraction, sequencing, data analysis, and new targeted panels that may only become available at the end of the study can be applied to all serial blood samples, which increases data quality and comparability. Innovative studies can be designed to track the evolution of disease-associated mutations in the serum cfDNA. This would allow to evaluate if variations in the tumor allele fractions of the mutations mirror the genetic heterogeneity in the tumors and to determine if disease progression is associated with the emergence of additional somatic mutations. This, in turn, may help to assess whether mutational evolution reflects radiologically determined disease burden, recurrence, or metastasis.

This study has several limitations. First, some of the mutations may have been germline variants, especially those at high allelic frequencies. We cannot exclude this possibility owing to the lack of germline controls or clonal hematopoiesis. However, the primary aim of the study was to determine whether it was possible to identify mutations in 30-year-old serum, and the exclusion of germline variants can be achieved by using germline control. Second, the custom panel we used in this study was not optimized for mutation detection in cfDNA, because some of the amplicons are bigger than the average size of cfDNA fragments. We may thus have missed some mutations, and the use of a panel with smaller amplicon size will likely increase the number of mutations that can be detected. Third, another important limitation is the small sample size that precluded any analyses on associations between mutations and clinical endpoints. For example, it would be interesting to see if patients with detectable mutations in the cfDNA have a higher tumor stage and therefore poorer outcome than patients without detectable mutations. This has been shown in patients with late-stage gastric cancer, where patients with detectable mutations had a 5.6% 5-year overall survival rate compared with 31.5% in patients without detectable mutations.[36] TP53 is one of the most frequently mutated genes in breast cancer, and, being a tumor suppressor and usually associated with the loss of the wild-type allele, TP53 mutations are likely to be more readily detectable in cfDNA than activating oncogenic mutations. In fact, 7 of 24 detected mutations in this series were TP53 mutations. However, it would be very challenging to adjust for the selection bias in this series of high-risk patients referred to the medical oncologist for systemic treatment even if a higher sample size could have been achieved. Nevertheless, evaluating associations between mutations and clinical endpoints is an area of high potential relevance, particularly when DNA from matched archival tissue of primary tumors or distant metastases are available. As outlook for future projects, we plan to assess the prognostic role of cancer-associated mutations in the serum cfDNA at diagnosis with the extensive follow-up information and clinicopathologic parameters available for our unique cohort of patients.

In conclusion, the present study shows that current NGS technology is sufficiently robust and specific to analyze 30-year-old serum. Based on this finding, longitudinal studies can be designed to be more feasible and flexible by storing biological samples over a long period of time. This allows for uniform sequencing with the most modern technology and adequate statistical power by cumulating oncologic events. Our study supports the value of liquid biopsies in assessing the dynamic changes of genetic heterogeneity over time and in the validation of new cfDNA biomarkers for breast cancer.

Clinical Practice Points

- The potential of using cfDNA as an indicator of disease burden with prognostic implication and clinical applicability during follow-up and monitoring in both the curative and palliative

setting has been investigated in numerous studies. Cancer-specific mutations, copy number alterations, and genomic rearrangements assessed in ctDNA demonstrated potential prognostic and predictive significance.

- To further evaluate prognostic and predictive biomarkers in cfDNA and assess its value as a disease monitoring tool, longitudinal studies with long follow-up are necessary. The utility of the technology depends on its capability to assess sequential samples that have been collected and stored over a long period of time.
- The aim of this study was to assess the feasibility of cfDNA extraction and somatic mutation assessment in 30-year-old serum. We evaluated samples from 52 patients with breast cancer, which were collected between 1983 and 1991. cfDNA extraction was successful in 52 of 52 patients, and 24 cancer-specific mutations were found in 22 of 25 samples undergoing sequencing.
- Our results suggest that current NGS technology is sufficiently robust and specific to analyze 30-year-old serum. Based on this finding, longitudinal studies can be designed to be more feasible and flexible by storing biological samples over a long period of time. This allows for uniform sequencing with the most modern technology and adequate statistical power by cumulating oncologic events. Our study supports the value of liquid biopsies in assessing the dynamic changes of genetic heterogeneity over time and in the validation of new cfDNA biomarkers for breast cancer.

Acknowledgments

The authors thank Michelle Attenhofer and Stefan Herms for their assistance with sample and data storage over the long period of time. The authors thank Philip Martin Jermann for critical review of the manuscript. The authors also thank Basel Cancer League for paying the first freezer in 1983.

Partially, this work was funded by the "OPO-Stiftung" and the "Frieder Locher-Hofmann-Stiftung." This funding source had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Disclosure

The authors have stated that they have no conflicts of interest.

Supplemental Data

Supplemental table accompanying this article can be found in the online version at https://doi.org/10.1016/j.clbc.2020.04.005.

References

1. Joyner MJ, Paneth N. Seven questions for personalized medicine. *JAMA* 2015; 314:999-1000.
2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; 490:61-70.
3. Cheng DT, Mitchell TN, Zehir A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J Mol Diagn* 2015; 17:251-64.
4. Chen K, Meric-Bernstam F, Zhao H, et al. Clinical actionability enhanced through deep targeted sequencing of solid tumors. *Clin Chem* 2015; 61:544-53.
5. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012; 366: 883-92.
6. Yates LR, Gerstung M, Knappskog S, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* 2015; 21:751-9.
7. Wan JC, Massie C, Garcia-Corbacho J, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* 2017; 17: 223-38.
8. Bidard FC, Weigelt B, Reis-Filho JS. Going with the flow: from circulating tumor cells to DNA. *Sci Transl Med* 2013; 5:207ps14.
9. Jung K, Fleischhacker M, Rabien A. Cell-free DNA in the blood as a solid tumor biomarker–a critical appraisal of the literature. *Clin Chim Acta* 2010; 411:1611-24.
10. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res* 1977; 37:646-50.
11. Stroun M, Anker P, Lyautey J, Lederrey C, Maurice PA. Isolation and characterization of DNA from the plasma of cancer patients. *Eur J Cancer Clin Oncol* 1987; 23:707-12.
12. Thierry AR, El Messaoudi S, Gahan PB, Anker P, Stroun M. Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev* 2016; 35: 347-76.
13. Schwarzenbach H, Pantel K. Circulating DNA as biomarker in breast cancer. *Breast Cancer Res* 2015; 17:136.
14. Umetani N, Giuliano AE, Hiramatsu SH, et al. Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J Clin Oncol* 2006; 24: 4270-6.
15. Dawson SJ, Rosenfeld N, Caldas C. Circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 2013; 369:93-4.
16. Dawson SJ, Tsui DW, Murtaza M, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* 2013; 368:1199-209.
17. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat Rev Cancer* 2011; 11:426-37.
18. Garcia-Murillas I, Schiavon G, Weigelt B, et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci Transl Med* 2015; 7: 302ra133.
19. De Mattos-Arruda L, Weigelt B, Cortes J, et al. Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor DNA: a proof-of-principle. *Ann Oncol* 2014; 25:1729-35.
20. Schiavon G, Hrebien S, Garcia-Murillas I, et al. Analysis of ESR1 mutation in circulating tumor DNA demonstrates evolution during therapy for metastatic breast cancer. *Sci Transl Med* 2015; 7:313ra182.
21. De Mattos-Arruda L, Mayor R, Ng CKY, et al. Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nat Commun* 2015; 6:8839.
22. Luen S, Virassamy B, Savas P, Salgado R, Loi S. The genomic landscape of breast cancer and its interaction with host immunity. *Breast* 2016; 29:241-50.
23. Paradiso V, Garofoli A, Tosti N, et al. Diagnostic targeted sequencing panel for hepatocellular carcinoma genomic screening. *J Mol Diagn* 2018; 20: 836-48.
24. Garofoli A, Paradiso V, Montazeri H, et al. PipeIT: a singularity container for molecular diagnostic somatic variant calling on the Ion Torrent next-generation sequencing platform. *J Mol Diagn* 2019; 21:884-94.
25. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature* 2015; 526:68-74.
26. Karczewski KJ, Weisburd B, Thomas B, et al. The ExAC browser: displaying reference data information from over 60,000 exomes. *Nucleic Acids Res* 2017; 45: D840-5.
27. National Heart, Lung, and Blood Institute. NHLBI Grand Opportunity Exome Sequencing Project (ESP), Available at: https://esp.gs.washington.edu/drupal/. accessed January 8, 2020.
28. Soysal SD, Ng CKY, Costa L, et al. Genetic alterations in benign breast biopsies of subsequent breast cancer patients. *Front Med (Lausanne)* 2019; 6:166.
29. Carter H, Chen S, Isik L, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* 2009; 69:6660-7.
30. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013; 34:57-65.
31. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010; 7:575-6.
32. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* 2012; 7:e46688.
33. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* 2017; 2017.
34. Wagner AH, Coffman AC, Ainscough BJ, et al. DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res* 2016; 44:D1036-44.
35. Griffith M, Griffith OL, Coffman AC, et al. DGIdb: mining the druggable genome. *Nat Methods* 2013; 10:1209-10.
36. Fang WL, Lan YT, Huang KH, et al. Clinical significance of circulating plasma DNA in gastric cancer. *Int J Cancer* 2016; 138:2974-83.

Supplemental Data

Supplemental Table 1 Gene Coverage of the Custom Targeted Sequencing Panel

Gene Name	Chromosome	Start	End	Cytoband	Remarks
ARID1A	1	27022524	27108595	p36.11	Complete coding region
NRAS	1	115247090	115259515	p13.2	Hotspot residues 12,13 and 61 only
SETD2	3	47057919	47205457	p21.31	Complete coding region
PIK3CA	3	178865902	178957881	q26.32	Complete coding region
FBXW7	4	153242410	153457253	q31.3	Complete coding region
MAP3K1	5	56111401	56191979	q11.2	Complete coding region
PIK3R1	5	67511548	67597649	q13.1	Complete coding region
ARID1B	6	157099063	157531913	q25.3	Complete coding region
EGFR	7	55086714	55324313	p11.2	Complete coding region
KMT2C	7	151832010	152133090	q36.1	Complete coding region
PTPRD	9	8314246	10612723	p23	Complete coding region
GAT A3	10	8095567	8117161	p14	Complete coding region
PTEN	10	89622870	89731687	q23.31	Complete coding region
HRAS	11	532242	537287	p15.5	Hotspot residues 12, 13, and 61 only
NEAT1	11	65190245	65213011	q13.1	Complete coding region
MALAT1	11	65265233	65273940	q13.1	Complete coding region
ATM	11	108093211	108239829	q22.3	Complete coding region
KRAS	12	25357723	25403870	p12.1	Hotspot residues 12, 13, and 61 only
ERBB3	12	56473641	56497289	q13.2	Complete coding region
TBX3	12	115108059	115121969	q24.21	Complete coding region
RBI	13	48877887	49056122	q14.2	Complete coding region
F0XA1	14	38059189	38069245	q21.1	Complete coding region
AKT1	14	105235686	105262088	q32.33	Complete coding region
CBFB	16	67063019	67134961	q22.1	Complete coding region
CTCF	16	67596310	67673086	q22.1	Complete coding region
CDH1	16	68771128	68869451	q22.1	Complete coding region
TP53	17	7565097	7590856	p13.1	Complete coding region
MAP2K4	17	11924141	12047147	P12	Complete coding region
NCOR1	17	15932471	16121499	p11.2	Complete coding region
NF1	17	29421945	29709134	q11.2	Complete coding region
ERBB2	17	37844167	37886679	q12	Complete coding region
RUNX1	21	36160098	37376965	q22.12	Complete coding region

Acknowledgments

This work would have not been possible without the contribution of many people, for whom I am deeply grateful.

I would like to thank my supervisors Prof. Terracciano and Dr. Piscuoglio for their guidance for the past four years and for the opportunity to work on some exciting projects; Dr. Ng for her time and patience helping me through bioinformatic and stylistic problems and to challenge my critical thinking; and all the past and present members of Dr. Piscuoglio's group for introducing me to the not so distant worlds of biology and politics.

I would also like to thank Prof. Christofori, PD. Dr. Kruithof-de Julio and Prof. Odermatt for accepting to judge my work and to be part of my PhD committee.

Not less important, I would like to thank *so many* other people - friends - not only in Basel but all over the world without whom I could have not done my job and overcome all the adversities. You are the best, I am blessed to have you in my life.

However, I have to thank especially all my friends from the pathology institute. From the first one I made coming here in Basel alone, to the last one, arrived not so long ago; from the people who left already to the people who will stay after me. To thank you one by one I would need a longer space than the entire thesis, because to describe your qualities, values and support in a proper way could take me ages and you know I can ramble sometimes! But I hope you know already how significant and special you are for me, with all your amazing differences between each other, so please be always aware of your value and never allow anyone to mine you, personally as professionally. Working with you was an honour and drinking and laughing with you was it even more. Also, discussing with you was an essential part of my growth and without you I would be never here, where I am right now. So thank you very much, I deeply love you!

And of course I cannot exclude from my group of friends my best friend and partner in crime in the last 2 years. You are what I could have never imagined a person to be, you're much more than just supportive, kind, curious and smart. You expanded my world, my perspectives and my expectations. You are the best robot I ever met and I am extremely grateful to have you in my life.

Last but not least, I would like to thank my parents and my family for being supportive and patient. Feeling your love despite the distance allowed me to never be alone and I am so lucky to have such special people forever linked to my life. Thank you for having always believed in me, I dedicate this thesis to you, and especially to you, brother.

Curriculum vitae

VIOLA PARADISO

PERSONAL DETAILS

Address:
Mülhauserstrasse 112
4056 Basel

Tel: +41766039270
+393339879298

Email:
violaparadiso89@gmail.com

COMPETENCIES

- Experience in Translational Oncology Research, Execution and Operations
- Stakeholder management
- Project management
- Documentation and Reporting Skills
- Proficient in MS Office

LANGUAGE SKILLS

- Italian: native
- English: fluent (spoken and written)
- French: basic (spoken and written)
- German: basic (spoken and written)

WORKING EXPERIENCE

Dec 2016 – present: Doctoral Researcher
Laboratory of Prof Luigi Maria Terracciano and Dr Salvatore Piscuoglio, Institute of Pathology and Medical Genetics, University Hospital of Basel **(Switzerland)**

- Oncology research focused on the identification of new molecular targets and their molecular validation
- NGS expertise with both Illumina and Ion Torrent
- Training and supervision of master students and research courses

Oct 2015 – Nov 2016: Research assistant
Institut für Humangenetik, Laboratory of Gynecology, Universitätsklinikum Halle (Saale) **(Germany)**

- Genomic characterization in oncologic patients with gene expression arrays and NGS
- Maintenance of patients data for longitudinal studies

Feb 2015 - Sept 2015: Internship
Laboratory of Molecular Genetics, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) "Giovanni Paolo II" Bari **(Italy)**

- Genomic characterization of breast cancer patients with NGS

Jan 2014 - Jul 2014: Internship
Laboratory of Homologous Recombination and Cancer, Institut Curie - Research Centre, Orsay **(France)**

- Molecular characterization of unknown genetic alterations in oncogenic genes with assay development and investigation of mechanisms

Sep 2011 - Dec 2011: Internship
Laboratory of Molecular Genetics, IRCCS "Giovanni Paolo II" Bari **(Italy)**

- Genomic characterization of breast cancer patients with NGS

EDUCATION

December 2016 – *December 2020*: PhD candidate
Institute of Pathology and Medical Genetics, **University Hospital Basel (Switzerland)**.
Medical – Biological research Thesis: "*Genetic screening and molecular characterisation of biomarkers in hepatocellular carcinoma*".
Scientific supervisor: Prof. Luigi M. Terracciano and Dr. Salvatore Piscuoglio.

Sept 2012 - Oct 2014: Master's Double Degree in Functional Genomics and Genetics
(Final grade: 110/110). **University of Trieste (Italy)** and **University of Paris Diderot (France)**.
Molecular Genetic and Oncology Thesis: "*Functional characterization of BRCA2 variants identified in families at high risk of breast cancer*".
Scientific supervisor: Prof. Alberto Manfioletti and Dr. Aura Carreira.
Laboratory of Homologous Recombination and Cancer, Institut Curie - Research Centre, Orsay (France)

Sept 2008 - Dec 2011: Bachelor Degree in Biological Science – Molecular biology curriculum
(Final grade: 110/110) **University of Ferrara (Italy)**.
Medical Genetics Thesis: "*Genetic characterization of hereditary breast cancer susceptibility genes BRCA1 and BRCA2*".
Scientific supervisor: Prof. Chiara Scapoli and Dr. Stefania Tommasi.
Laboratory of Molecular Genetics, IRCCS "Giovanni Paolo II" Bari (Italy).

OTHER COURSES & QUALIFICATIONS

- 2020 – 4 days: Essentials in Drug Development & Clinical Trials Course, University of Basel
- 2020 – 2 days: Good Clinical Practice for Investigators and Study Teams Course, Swiss Tropical and Public Health Institute
- 2019 – 1 year: Antelope program-Novartis for career path for female doctoral students that includes mentorship and workshops, University of Basel and Novartis
- 2019 – 2 days: Practical Personalized Medicine Course, Universitätsspital Basel
- 2018 - 1 day: Research integrity, University of Basel
- 2018 - 1 day: Inferring gene regulatory networks from high-throughput data with ISMARA, Swiss institute of bioinformatics
- 2018 – 3 day: International PhD course "Frontiers in Metastasis Biology", Department of Biomedicine Basel
- 2018 – 1 day: Project Management for Researchers Course, University of Basel
- 2018 – 2 days: Learning How to Lead and to Build a Successful Work Environment, University of Basel
- 2017 – 1 semester: Molecular medicine II, University of Basel
- 2017 – 1 semester: Translational cancer research, University of Basel
- 2017 – 1 semester: Translational control and post-translational protein modification, University of Basel
- 2017 – 1 week: Bioinformatics: Computer Methods in Molecular and Systems Biology, ICGB Trieste

CONFERENCES & AWARDS

- June 2019: Personalized oncology event; Basel, Switzerland
- April 2018: AACR Annual Meeting 2018, Chicago, USA
- January 2018: 17th Hepatobiliary and Gastrointestinal research retreat; Les Diablerets, Switzerland
- November 2017: 83rd Annual congress of the Swiss society of pathology; Thun, Switzerland
- Sep 2016: EORTC PathoBiology Group travel award for the best abstract for oral presentation in the group of the young researchers, in Rotterdam
- Dec 2015: Degree award in memory of Dr. Federica Ziller for the best thesis on cancer field in the current academic year, in Trieste

PUBLICATIONS

- Prautsch K, Schmidt A, **Paradiso V**, Schaefer D, Guzman R, Kalbermatten D, Madduri S. "Modulation of human adipose stem cells' neurotrophic potency using a variety of growth factors for neural tissue engineering applications: axonal growth, transcriptional and phosphoproteomic analyses". *Cells. August 2020*
- Rodriguez A, Gallon J, Akhoundova D, Maletti S, Ferguson A, Cyrta J, Amstutz U, Garofoli A, **Paradiso V**, Tomlins SA, Hewer E, Genitsch V, Fleischmann A, Rushing EJ, Grobholz R, Fischer I, Jochum W, Cathomas G, Bubendorf L, Moch H, Ng CKY, Gillessen Sommer S, Piscuoglio S and Rubin MA. "The Genomic Landscape of Prostate Cancer Brain Metastases". *Under review to Nat Genet. May 2020*
- Koessler T, **Paradiso V**, Piscuoglio S, Nienhold R, Ho L, Christinat Y, Terracciano LM, Cathomas G, Wicki A, McKee T, and Nouspikel T. "Reliability of liquid biopsy analysis: an inter-laboratory comparison of circulating tumor DNA extraction and sequencing with different platforms". *Lab Investigation. July 2020*
- Ritter M*, **Paradiso V***, Widmer P, Garofoli A, Quagliata L, Eppenberger-Castori S, Soysal S, Muenst S, Ng CKY, Salvatore Piscuoglio S, Weber W and Weber WP. "Identification of somatic mutations in thirty-year-old serum cell-free DNA from patients with breast cancer: a feasibility study". *Clin Breast Cancer. 2020 January*

- Montazeri H, Coto-Llerena M, Bianco G, Zangeneh E, Taha-Mehlitz S, **Paradiso V**, Srivatsa S, de Weck A, Roma G, Lanzafame M, Bolli M, Beerenwinkel N, von Flüe M, Terracciano LM, Piscuoglio S, Ng CKY. "Systematic Identification of Novel Cancer Genes through Analysis of Deep shRNA Perturbation Screens". *bioRxiv. 2019 December*
- Soysal SD, Ng CKY, Costa L, Weber WP, **Paradiso V**, Piscuoglio S, Muenst S. "Genetic Alterations in Benign Breast Biopsies of Subsequent Breast Cancer Patients". *Front Med (Lausanne). 2019 July*
- Garofoli A*, **Paradiso V***, Montazeri H, Jermann PM, Roma G, Tornillo L, Terracciano LM, Piscuoglio S, Ng CKY. "PipeIT: A Singularity Container for Molecular Diagnostic Somatic Variant Calling on the Ion Torrent Next-Generation Sequencing Platform". *J Mol Diagn. 2019 September*
- Lackner C, Quagliata L, Cross W, Ribi S, Heinimann K, **Paradiso V**, Quintavalle C, Kovacova M, Baumhoer D, Piscuoglio S, Terracciano LM, Kovac M. "Convergent Evolution of Copy Number Alterations in Multi-Centric Hepatocellular Carcinoma". *Sci Rep. 2019 March*
- Montagna G, Ng CKY, Vlajnic T, **Paradiso V**, Dellas S, Reina H, Kind A, Weber WP, Piscuoglio S, Kurzeder C. "Fibroepithelial Breast Lesion: When Sequencing Can Help to Make a Clinical Decision. A Case Report". *Clin Breast Cancer. 2019 February*
- **Paradiso V***, Garofoli A*, Tosti N, Lanzafame M, Perrina V, Quagliata L, Matter MS, Wieland S, Heim MH, Piscuoglio S, Ng CKY, Terracciano LM. "Diagnostic Targeted Sequencing Panel for Hepatocellular Carcinoma Genomic Screening". *J Mol Diagn. 2018 November*
- Ng CKY, Di Costanzo GG, Tosti N, **Paradiso V**, Coto-Llerena M, Roscigno G, Perrina V, Quintavalle C, Boldanova T, Wieland S, Marino-Marsilia G, Lanzafame M, Quagliata L, Condorelli G, Matter MS, Tortora R, Heim MH, Terracciano LM, Piscuoglio S "Genetic profiling using plasma-derived cell-free DNA in therapy-naïve hepatocellular carcinoma patients: a pilot study". *Ann Oncol. 2018 March*
- Alisch F, Weichert A, Kalache K, **Paradiso V**, Longardt AC, Dame C, Hoffmann K, Horn D. "A familial case of Gordon syndrome is associated with a PIEZO2 mutation." *Am J Med Genet A. 2016 October*
- De Summa S, Pinto R, Sambiasi D, Petriella D, **Paradiso V**, Paradiso A, Tommasi S. "BRCAness: a deeper insight into basal-like breast tumors". *Ann Oncol. 2013 November*

REFERENCES

Luigi M. Terracciano, PD Dr
Head of the Molecular Pathology Division
luigi.terracciano@usb.ch

Salvatore Piscuoglio, Dr
Head of Research & Principal Research Investigator
Institute of Pathology and Medical Genetics
University Hospital Basel
salvatore.piscuoglio@usb.ch

Martina Vetter, Dr
Laboratory of Gynaecology
Universitätsklinikum
Halle (Saale) (Germany)
martina.vetter@medizin.uni-halle.de