Computational approaches to improve precision oncology

Inaugraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Andrea Garofoli

2021

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Luigi M. Terracciano

Prof. Michael N. Hall

Prof. Julia E. Vogt

Basel, 15.12.2020

Prof. Dr. Martin Spiess

Dekan

Without deviation from the norm,

progress is not possible.

Frank Zappa

Abstract

The word “cancer” identifies a collection of remarkably diverse range of diseases whose common trait is the presence of accelerated and unregulated cell proliferation that escalates into the development of so-called “tumoral tissue”. Molecular profiling of cancers has uncovered the presence of a vast diversity between cancers, laying the foundations for the use of case-by-case defined clinical decisions. The philosophy of precision oncology is based on the idea that patient care must take into account their molecular characteristics, to define the best therapy possible. The rise of big data and the computational approaches able to dissect it has enabled the profiling of an extraordinary number of diseases, whose characterization can be the stepping stone of precision oncology itself.

The aim of this project is the development of computational methodologies to support modern precision oncology and to help expand its modern implementations. Results are divided in two sections called **Chapter I** and **Chapter II**.

In the **Chapter I** we present PipeIT, a somatic variant caller we have developed to help researchers and clinicians to detect potential driver mutations in patients. PipeIT has been specifically designed to process data obtained from Ion Torrent, a sequencing platform frequently used in diagnostic settings but, compared to the other sequencing platforms, with few analysis tools. The novelty brought by PipeIT is its Singularity container nature, which ensures reproducibility of its analyses and enhances its ease of use. Two different PipeIT versions were developed. PipeIT was designed to perform variant calling on tumor-germline matched data. PipeIT2 was later developed to enable variant calling analysis of tumor only data, to broaden its use in the typical clinical setting. PipeIT2 takes advantage of publicly accessible databases and on panels of unmatched normals to account for the absence of a matched germline control. Both PipeIT pipelines were able to detect important driver genomic variants, proving to be a powerful tool for modern precision oncology.

In **Chapter II** we investigated the role of gene expression data as an alternative to DNA biomarkers to detect the presence of oncogenic molecular processes in cancer patients. Based on the assumption that the activation of oncogenic pathways caused by driver mutations can produce a specific transcriptional profile, we designed a machine learning classifier able to extract said profile from patients with driver hotspot mutations and infer its presence in patients who do not have the same hotspot mutations. The classifier was first

tested on one of the most frequently mutated oncogenes, *PIK3CA*, using publicly accessible TCGA pan-cancer data. The classifier was able to detect the presence of *PIK3CA* hotspot driver mutations on a testing data obtaining a ROC score of 0.87. The approach was further tested on 15 different oncogenes, demonstrating good results for the more commonly mutated oncogenes and underperforming for more rarely mutated ones. Finally, the *PIK3CA* model was used on an external set of TCGA samples to determine whether the classifier was also able to infer the presence of additional *PIK3CA* oncogenic mutations. This project highlighted the importance of novel AI based approaches on cancer data and the potential applications of transcriptomic data as biomarker to further improve precision oncology.

List of abbreviations

AUC ROC: Area Under the Curve Receiver Operating Characteristic

BAM: Binary Alignment Map

BRCA: Breast Invasive Carcinoma

BED: Browser Extensible Data

DNA: Deoxyribonucleic Acid

FDA: Food and Drug Administration

GATK: Genome Analysis Toolkit

HCC: Hepatocellular Carcinoma

ICGC: International Cancer Genome Consortium

NGS: Next Generation Sequencing

PCAWG: Pan-Cancer Analysis of Whole Genomes

PoN: Panel of Normal

RNA: Ribonucleic Acid

SMOTE: Synthetic Minority Over-sampling Technique

TCGA: The Cancer Genome Atlas

TVC: Torrent Variant Caller

VAF: Variant Allele Frequency

VCF: Variant Call Format

WES: Whole Exome Sequencing

WGS: Whole Genome Sequencing

List of genes

AKT1: AKT Serine/Threonine Kinase 1

BRAF: B-Raf Proto-Oncogene, Serine/Threonine Kinase

CTNNB1: Catenin Beta 1

EGFR: Epidermal Growth Factor Receptor

ERBB2: Erb-B2 Receptor Tyrosine Kinase 2

FGFR3: Fibroblast Growth Factor Receptor 3

GNA11: G Protein Subunit Alpha 11

GNAQ: G Protein Subunit Alpha Q

HRAS: HRas Proto-Oncogene, GTPase

IDH1: Isocitrate Dehydrogenase 1

KRAS: KRAS Proto-Oncogene, GTPase

MAP2K1: Mitogen-Activated Protein Kinase Kinase 1

NFE2L2: Nuclear Factor, Erythroid 2 Like 2

NRAS: NRAS Proto-Oncogene, GTPase

PIK3CA: Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha

SF3B1: Splicing Factor 3b Subunit 1

TERT: Telomerase Reverse Transcriptase

TP53 : Tumor Protein P53

Table of contents

Abstract I

List of abbreviations III

List of genes IV

Table of contents V

1- Introduction 1

I. Precision Medicine 2

The Role of Computational Science in Medicine 3

II. Cancer and Precision Oncology 5

DNA Alterations in Precision Oncology 5

The Role of Next Generation Sequencing on Precision Oncology 6

DNA Mutations 6

DNA Sequencing 9

Liquid Biopsies 11

Non-Coding variants 11

Molecular Tumor Boards and Basket Trials 12

III. The Landscape of Biomarkers Outside DNA sequencing 13

Gene Expression as Clinical Biomarker 14

IV. Bioinformatics in Precision Oncology 16

Variant Callers 17

Machine Learning Applications in Precision Oncology 18

2- Rationale and Aims of the Thesis 21

3- Results 22

3.1- Chapter I 23

Development of somatic variant calling pipelines for the detection of oncogenic mutations and to drive precision medicine 23

PipeIT: A Singularity Container for Molecular Diagnostic Somatic Variant Calling on the Ion Torrent Next-Generation Sequencing Platform 24

PipeIT2: Singularity Container for Tumor-Only Molecular Diagnostic Somatic Variant Calling on Ion Torrent NGS Platform 36

ABSTRACT 37

INTRODUCTION 38

MATERIALS AND METHODS 39

The PipeIT2 Tumor-Only Workflow 39

Generation of the Panel of Normals (PoN) VCF file 42

Building the PipeIT2 Singularity Container Image 42

Evaluation of the PipeIT2 tumor-only workflow 42

Visualization of BAM files 43

RESULTS 43
Running the PipeIT2 tumor-only workflow 43
Evaluation of the PipeIT2 tumor-only workflow 44
Validation of false positive mutations 47
DISCUSSION 48
ACKNOWLEDGMENTS 50
AUTHOR CONTRIBUTION 50
FIGURES AND FIGURE LEGENDS 52
SUPPLEMENTARY FIGURE AND LEGENDS 56

3.2- Chapter II 59

A machine learning approach to extract oncogenic transcriptional profiles and to expand precision oncology 59
ABSTRACT 60
INTRODUCTION 61
MATERIALS AND METHODS 64
Downloading of the TCGA data 64
Development of a logistic regression classifier 64
Defining the gene interactors 66
Correcting for unbalanced classes 66
Evaluation of classifier performance 67
Prediction on the Discovery Dataset 67
RESULTS 67
Development of a logistic regression classifier 67
Training the PIK3CA classifier 68
Using SMOTE to overcome class imbalance 71
The benefits of pan-cancer data over cancer specific data 72
Evaluating the PIK3CA classifier 73
Development and evaluation of classifiers for 15 additional oncogenes 74
Inference of pathway activation status 77
DISCUSSION 78
FIGURES 80
TABLES 85

4- Discussions and Outlook 95

Bibliography 100

Annex 109

Acknowledgments 123

1- Introduction

I. Precision Medicine

The unprecedented influx of data obtained in the first two decades of the new millennium is one of the most critical factors that is revolutionizing modern medicine.[1] The information provided by high-throughput and high resolution omic technologies such as DNA and RNA sequencing, proteomics, microarrays and epigenomics has aided the discovery of novel molecular mechanisms.

While diversity such as lifestyle, environment factors and family health disorders history, has always been acknowledged in everyday clinical care to identify differences between patients, the new molecular data unveiled a whole new level of heterogeneity within diseases. This discovery demonstrates how a single disease can potentially be, on the molecular level, a collection of multiple diseases with a diverse range of molecular aberrations that converge on the phenotypic level. This highlights the reasons behind the different responses observed from standardized therapies in patients suffering, apparently, from the same disease.

Precision Medicine moves the paradigm from this kind of standardized “one size fits all” treatments to “case by case” therapeutic scenarios (**Figure 1**).[2,3] The rationale behind it is that by gathering, understanding and profiling the information collected and processed from large cohorts of patients, it is possible to personalize, or “tailor”, aspects of treatments to better fit each individual’s needs and and to improve outcome. This paved the way to the development of new therapies, able to target the specific aberrations observed in patients.[1,4,5]

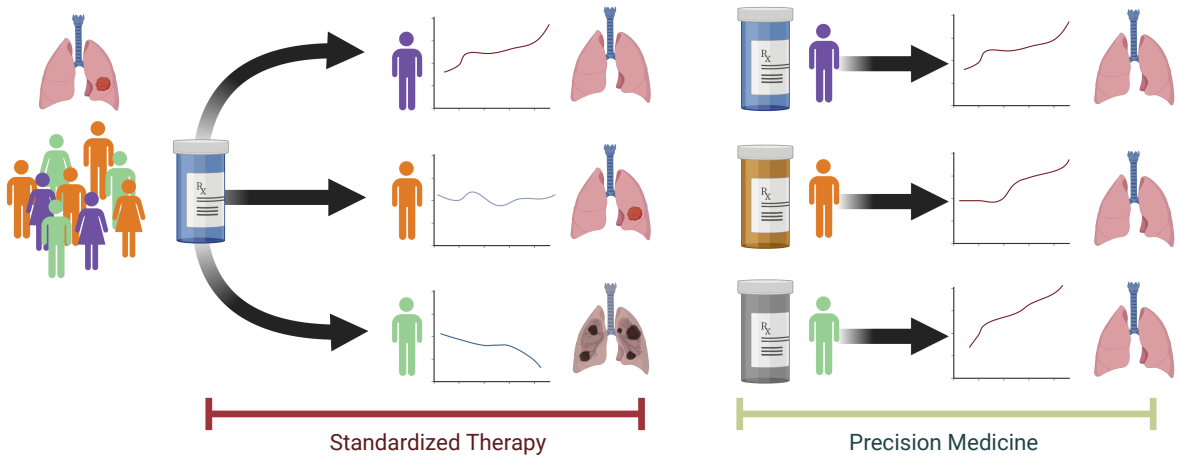


***Figure 1**. Visual description of the Precision Medicine philosophy. A disease can be the result of a convergent phenotype caused by heterogeneous molecular mechanisms. This explains why standard treatments can lead to different outcomes. The aim of Precision Medicine is to tailor clinical decisions for individual patients.*

Many initiatives have been launched all around the world to promote Precision Medicine. In 2013 the European Union funded the Personalized Medicine (PerMed) project, which laid the foundations for the following International Consortium of Personalized Medicine (ICPerMed), involving over 30 different members between countries and private organizations, even outside the European continent, bounded by the shared goal: to coordinate and fund research that will eventually have a real impact on the clinical level.[6] In 2015, the USA National Institute of Health (NIH) promoted the All-of-US project, whose aim is to collect and study health data from a wide range of patients.[7] In Asia, China funded the China Precision Medicine Initiative in 2017 with an investment of US$9.2 billions.[8] In the same year, 5 Switzerland university hospitals and other public research institutions joined forces and created the Swiss Personalized Health Network (SPHN), under the leadership of the Swiss Academy of Medical Sciences and in collaboration with the Swiss Institute of Bioinformatics.[9,10] These are only a few examples of the massive projects launched worldwide, but they highlight the high hopes and anticipation that advances in precision medicine will have wide-ranging impact on clinical cancer care worldwide.

The Role of Computational Science in Medicine

The advent of the high-throughput big data made it clear how the idea of a manual curation of the novel information was inconceivable. The support of computational technologies became indispensable to provide a solid infrastructure to store, manage, distribute and, ultimately, analyze the huge influx of data produced. This is particularly true in a precision medicine context where a real time and easy way to obtain information can be critical for clinicians to strategize the decision making.[11]

Bioinformatics stems from the utility informatics science brought and merges it with statistics, mathematics and, of course, medical and biological knowledge.[12,13] The goal behind this rise was to find a new way (the *in silico* one) to investigate the molecular mechanisms compared to the more classic experimental (*in vitro* or *in vivo*) and clinical studies.[14,15] It was easy to see how Bioinformatics was able to open the door to unprecedented scenarios. However, this field also faces a number of challenges. The first, and probably most obvious, lies in its multidisciplinary nature; the contribution Bioinformatics brings to each project or study

relies on a good comprehension of all the different fields it is based on. The others cover more technical aspects. It is mandatory for investigations based on computational analyses to ensure data traceability, sharing and reproducibility. The lack of said aspects can significantly undermine the credibility of the study.

Translational bioinformatics research led to the development of new approaches able to leverage on the influx of novel omic data to dissect disease molecular networks, identify new treatment biomarkers and estimate changes between healthy and unhealthy cells to find correlations between convergent phenotypes, pathway aberrations and their activating key events.[14,16]

II. Cancer and Precision Oncology

Oncology is at the forefront of Precision Medicine.[17,18] The reasons are easy to understand. The word “cancer” identifies a collection of extremely heterogeneous diseases whose common theme is the presence of abnormally proliferative and long-living cells. They can result in a mass of tumoral tissue able to take over the original tissue, in case of solid cancers, or in the loss of the normal tissue functions, for example when the disease involves the bone marrow, blood or lymph nodes and results in specific cancer types named lymphoma and leukemia. The incredible survival capacities of cancer cells can also allow them to escape the original tissue and, thanks to the bloodstream or the lymph system, travel through to different areas of the body, resulting in the development of metastasis.

What are the causes behind the rise of such abnormal cells? Cancer is often defined as “a disease of the genome”. DNA alterations in strategic genes can disrupt cellular molecular pathways, altering the way the cells behave, grow and proliferate.[19,20] Said alterations can belong to the ‘germline’ and the ‘somatic’ groups. The former mutations are inherited from parents’ sperm and egg cells, which means they are present in every cell in the body. Only a small fraction of cancer types has been definitely associated with germline variants. The “triple negative” breast cancer subtype, for example, where frequencies of *BRCA1* and *BRCA2* (BReast CAncer gene 1 and 2) germline mutations are higher compared to other breast cancer subtypes. The latter class accounts for the vast majority of the known oncogenic alterations (causing 80% - 95% of the cancers, worldwide) and are accumulated in the DNA over the lifetime of the patients.[21,22] There are a number of different factors behind the development of said mutations. Environmental factors in individual lifestyle have a huge impact. Exposure to tobacco, alcohol, radiations and other cancerous substances can induce genetic alterations in specific tissues and, possibly, affect the genes involved in oncogenesis. Somatic mutations can also be the result of processes unrelated to lifestyles, like base mismatches during cell replications.[23]

DNA Alterations in Precision Oncology

The impact that DNA alterations have on oncogenesis highlights how Precision Oncology is based on the assumption that the response to individual treatments is mostly derived by the genetic profile of the patients. Aberrations such as single nucleotide variants, insertions and deletions, copy number variations and gene fusions can act as biomarkers and their identification can tip the scale in favor of treatments specifically designed to counter well defined molecular processes.[17] Ultimately, patients with the same genetic biomarkers, the

same cancer type and a similar background (such as lifestyle and family diseases history) should have the same response to therapies.

The Role of Next Generation Sequencing on Precision Oncology

Precision Oncology makes intensive use of DNA sequencing techniques to investigate the genome. Nonetheless, there are a number of other sequencing options available for both the research and the clinical communities.[24] Messenger RNA (mRNA) sequencing is needed to measure changes in the transcriptome, Chromatin Immunoprecipitation (ChIP) sequencing is able to give insights on DNA-protein interactions, and methylome sequencing to profile the presence of methylation, just to cite some of the most prominent ones. [24–26]

In the first two decades of the 21st century the Next Generation Sequencing (NGS) technologies underwent a considerable evolution. The increasing accuracy, higher resolution, decreasing costs and the faster outputs allowed sequencing to become one of the, if not the, main protagonist of routine oncology diagnostics practice. [27,28]

DNA Mutations

Genes whose aberrations are connected to cancer growth are either classified as “oncogenes” or “tumor suppressor”.[29] Mutations in the former class can lead to an oncogenic protein whose novel gain-of-function can activate pathways that culminate in the disease phenotype. Genes in the latter class are translated into proteins in charge of “housekeeping” tasks such as DNA repairing, apoptosis promotion or cell division control. Oncogenic mutations can lead to the loss of these functions and promote the unregulated growth quintessential for tumoral cells. [29,30] A few exceptions, can both act as tumor suppressors and oncogenes, *TP53* being the most famous example.[31] *TP53* encodes the p53, a transcription factor able to regulate the cell cycle and to activate tumor suppressor processes such as DNA repair and apoptosis (**Figure 2**). Unlike what can be expected from classic tumor suppressor genes, gain-of-function oncogenic mutations in the *TP53* gene (such as R282W) have been observed in cancer patients, providing to this gene a dual nature.

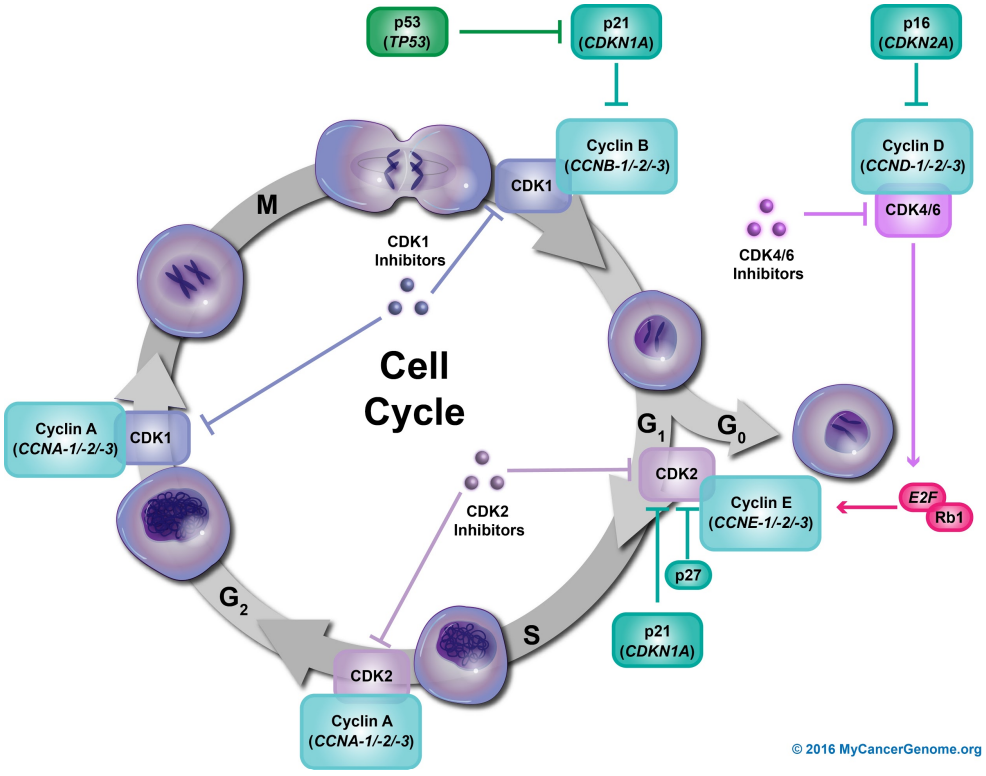


***Figure 2**. Summary of the cell life cycle process. The p53 tumor suppressor protein plays a role in both G1/S (gap phase 1 and DNA synthesis phase) and G2/M (gap phase 2 and mitosis phase) checkpoints. Said checkpoints are triggered by DNA damage and are able to promote the programmed cell death pathway. Image obtained from MyCancerGenome.org.*

Obviously, not all the mutations found in these genes are directly tied to carcinogenesis. Cancer cells contain a mixture of mutations which lead to oncogenic events and mutations with no real impact on cancer growth. They are called “driver” and “passenger” mutations, respectively.[19] In order to assign a mutation to one of these groups it is important to first interpret their impact on oncogenic pathways, a significantly more challenging task than simply detecting them.

Mutations are likely to have different repercussions depending on the tissue they are generated in, which means that the driver nature of a mutation is often tied to specific cancer types. Mutations in the *PIK3CA* (Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha) gene, for example, are most frequently found in patients with breast, thyroid, endometrial and head and neck carcinomas. To a lesser, but still significant extent, *PIK3CA* is also frequently mutated in colon and lung carcinomas.[32,33] The gene encodes for the catalytic domain (p110α) of the phosphatidylinositol-3-kinase (PI3K), a receptor tyrosine kinase involved in processes such as cell growth, survivability and proliferation. The PI3K

catalytic domain is able to convert its phosphatidylinositol-4,5-bisphosphate substrate (PIP2) into a phosphatidylinositol-3,4,5-bisphosphate (PIP3) substrate and, ultimately, trigger several signaling cascades (**Figure 3**). The activation of the AKT/mTOR pathway is one of the most well known consequences. When the PI3K catalytic domain is mutated, the AKT/mTOR pathway is exacerbated and leads to an extraordinary cell survivability.[34] Nonetheless, it has been observed that not all the genetic mutations in *PIK3CA* promote tumorigenesis.[35,36]

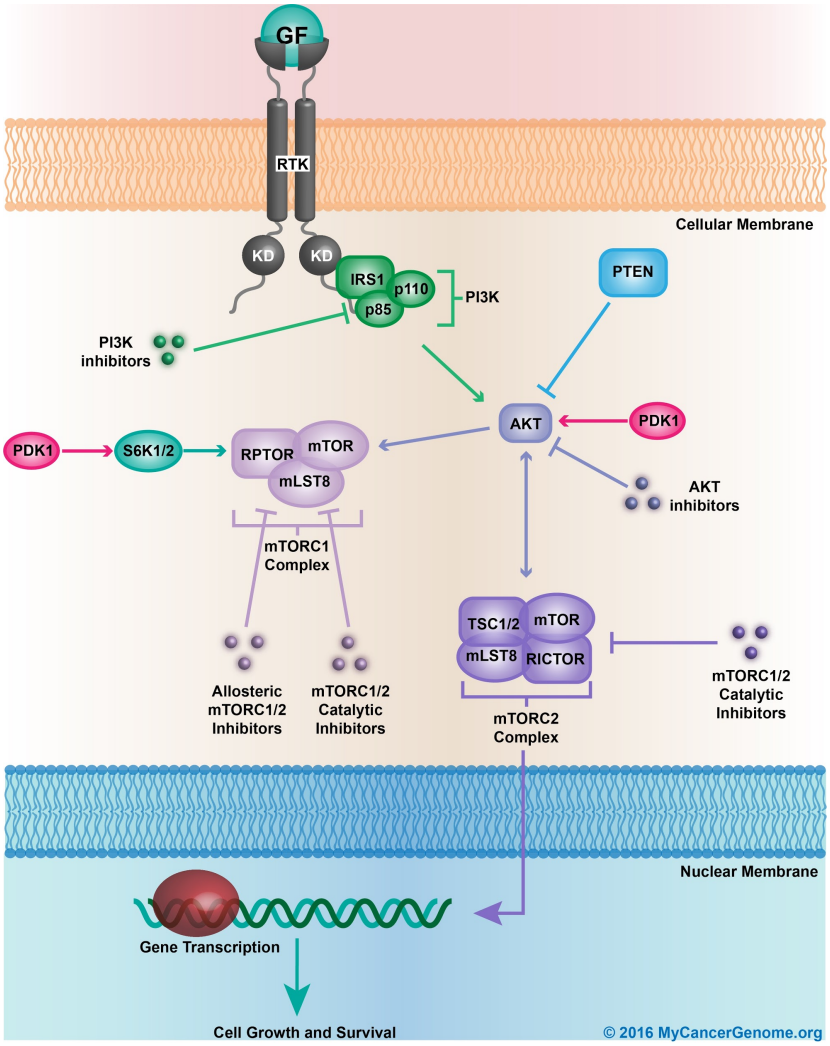


***Figure 3**. Schematic summary of the PI3K/AKT/mTOR pathway. Tyrosine kinase receptor (RTK) is activated by the binding of growth factors and promotes the enzymatic activity of the PI3K p110α subunit allowing the transphosphorylation from phosphatidylinositol-4,5-bisphosphate (PIP2) to* phosphatidylinositol-3,4,5-bisphosphate (*PIP3). PIP3, in turn, activates AKT1 and the downstream mTORC1 and mTORC2 (mammalian target of rapamycin complex 1 and 2) complexes, promoting cell growth and survival. Image obtained from MyCancerGenome.org.*

DNA Sequencing

Depending on the regions of interest, there are several options for DNA sequencing. Whole Genome Sequencing (WGS) or Whole Exome Sequencing (WES) analyses can be performed. The strength of WGS lies in the ability to investigate non-coding portions of DNA like introns, whose alterations can potentially impact the RNA transcription and produce aberrant proteins, or regulatory regions with potential alterations in the transcriptional levels.[37] WGS analyses enabled the researchers to perform Genome Wide Association Studies (GWAS), whose aim is to better investigate the role of genomic elements across complete genomes of several individuals and try to detect potential biomarkers associated with specific types of diseases.[38,39] While WGS offers the possibility of retrieval of every last genetic anomaly from the samples, normally unobtainable by other sequencing approaches, this option also comes with considerable drawbacks. First, WGS is still significantly more expensive and slower to obtain. Second, the enormous amount of bases studied comes at the expense of the coverage, relatively lower at individual loci, making it overall less accurate. Last, WGS sequencing can be much harder to review and study, even with computational approaches, due to the fact that the non-coding regions of the human genome are much less well characterised. In conclusion, while it offers a rich data source for specific, in depth studies, WGS is far from being optimal in a routine clinical setting. WES tries to overcome these limits by focusing on the protein-coding regions, which account for 2% of the whole human genome. Exomes have been intensively studied in the past years making it easier to interpret the role of variants found in the coding regions, compared to the ones in the remaining 98% of the genome. Moreover, the lower costs and easier to interpret results made WES more appropriate for clinical applications.[40]

Sequencing panels have been developed to further aid the diagnostic and research laboratories. The idea behind said panels is to focus on a small subset of genes and other genetic regions with established oncogenic aberrations in order to further decrease the sequencing costs, increase the feasibility of targeted analyses and maximize the output's sequencing depth and coverage compared to analyses based on both WGS and WES.[41] Many sequencing panels were clinically approved and made commercially available. Some of them, such as the Ion Torrent Oncomine Comprehensive Assay version 3 (Thermo Fisher Scientific, Waltham, MA)[42] and the capture-based Foundation Medicine FoundationOne assay, are commonly used to identify the genetic cause behind the disease and include a list of regions whose mutations are known to have oncogenic effects on a broad range of cancer types. While most of the genetic aberrations behave differently in diverse cancers, as previously explained, mutations in genes such as *TP53*, *CTNNB1* and *TERT* promoter are

well known examples of driver events shared across many different cancer types.[43] Other panels, the Illumina breast cancer specific AmpliSeq panel or the Ion Torrent lung and colon AmpliSeq panel, for instance, are cancer type specific and include a list of biomarker mutations particularly important in the selected diseases.

Panels are further divided into DNA targeted ones, whose role is to investigate the presence of mutations, copy number variations and indels, and RNA targeted ones, mainly used to investigate the presence of gene fusions. In 2018 we published our targeted sequencing panel for hepatocellular carcinoma (HCC). The panel includes many important biomarkers for HCC: all the exons of 33 genes, such as the HCC specific *APOB*, *ALB*, *HNF1A* and *HNF4A*, 2 long non-coding RNA genes, *TERT* promoter and 9 further genes for the detection of copy number alterations (**Figure 4**).[44]

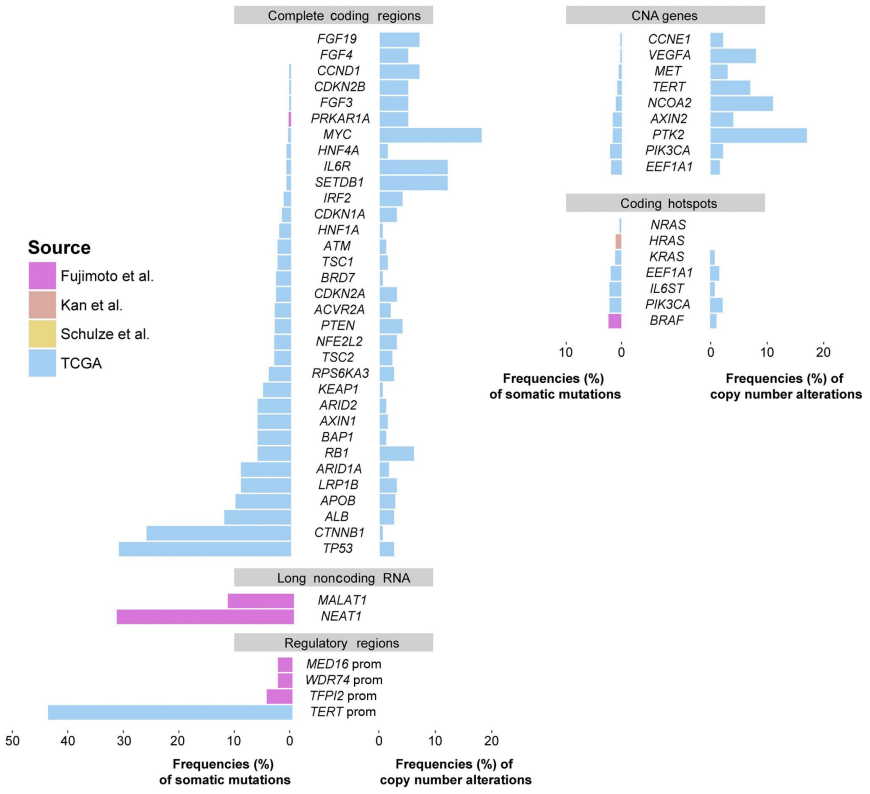


***Figure 4**. Design of our hepatocellular carcinoma (HCC) sequencing panel. Regions included in the panel are based on the frequencies of somatic mutations and copy number alterations observed in The Cancer Genome Atlas (TCGA) database or described in previously published studies. Image retrieved and adapted from "Diagnostic Targeted Sequencing Panel for Hepatocellular Carcinoma Genomic Screening".[44]*

Liquid Biopsies

In recent years researchers tried to infer whether liquid biopsies, such as blood, could offer an alternative, less invasive source of genetic material for sequencing analysis. The identification of cell-free circulating DNA (cfDNA) derived from necrotic and apoptotic cancer cells in blood comes with pros and cons.[45] Despite providing a practical solution for the identification in the clinical setting for prognostic mutations or other genomic aberrations, the analysis of circulating genetic material is not necessarily straightforward, given that even driver alterations may only be present in small fractions of the cfDNA and are therefore inherently difficult to detect. Technological advancements in molecular barcoding and error correction have started making substantial improvements in our ability to accurate profiling of cfDNA.[46]

Non-Coding variants

The importance of hotspot mutations in the aforementioned *TERT* promoter highlights the potential oncogenic driver role of non-coding variants. *TERT* (gene symbol for the gene Telomerase Reverse Transcriptase) plays an important role in telomere maintenance; its aberrant production caused by a mutated promoter severely hampers telomere attrition and ultimately leads to replicative immortality (i.e. unlimited potential for cellular proliferation).[47] *TERT* promoter is only one of the many examples of driver events in non-coding regions found over the past few years. Other important oncogenic aberrations were found in long non-coding and microRNA molecules, promoters, enhancers and other regulatory elements, proving the limits of WES or exome specific targeted assays in cancer.

On the 5th of February 2020, TCGA and the International Cancer Genome Consortium (ICGC) released the results of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project with a collection of 21 articles where they tried to determine the role of said non-exonic regions by studying a total of over 2600 cancer samples across 38 different cancer types.[43] Some of the investigated topics were the impact of aberrations such as chromothripsis and kataegis, respectively the abnormal rearrangement of big sections of a chromosome and the presence of hypermutations in small genomic regions. 25% of the tumors in the PCAWG cohort (n=2583) had at least one putative non-coding driver mutation so non-coding driver mutations were found to be rarer than their coding counterparts, with the exception of *TERT* being found mutated in 9% of the samples included in the whole cohort, but still critical in the profiling of the driver mutation landscape.

Molecular Tumor Boards and Basket Trials

Molecular Tumor Boards (MTB) and Basket Trials have been established to help the decision making and to define the treatment strategy that best suits the patient.[48] The classic workflow starts with the retrieval of biopsies from tumoral tissues and the sequencing of genetic material. Multiple biopsies can be obtained from different stages of the disease to better evaluate the progression. Sequencing data is then used to detect anomalies able to explain the reasons behind the disease. Patient data is then interpreted to discover pharmacologically actionable features and the therapy that better matches the profile obtained (**Figure 5**).[49,50] Computational approaches play an important role in all said phases, by being able to improve the resolution and the accuracy of the data retrieved from the patient and compelling methodologies to analyse it.

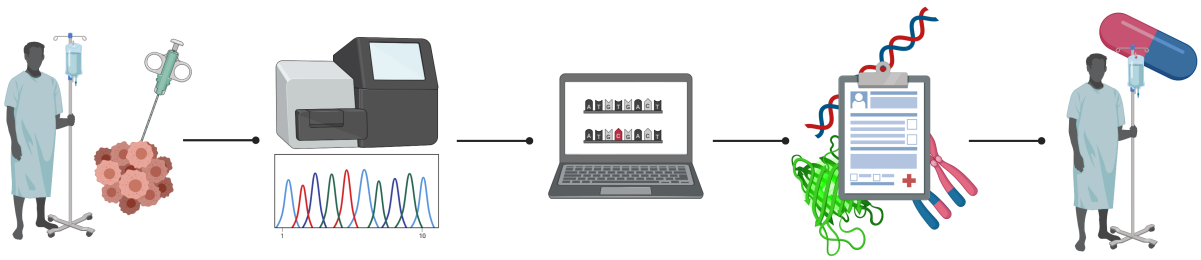


***Figure 5**. Schematic summary of the standard MTB routine. Biopsies are first retrieved from cancer patients. Genetic material is then extracted, sequenced and computationally analyzed. Results from said analyses are used to define a clinical profile which, in turn, is used by the tumor board to find and define the therapy that will likely lead to good response.*

III. The Landscape of Biomarkers Outside DNA sequencing

DNA alterations have a predominant importance in cancer.[19] However, the interplay with other factors must also be taken into account to properly profile a patient, especially in absence of easy to interpret genetic anomalies. One of the most interesting results shown in the aforementioned PCAWG study was that the average number of driver mutations identified across all the samples and both the coding and non-coding regions was between 4 and 5, while in around 5% of the 2583 samples (so approximately 130 samples) no known driver aberration was identified, suggesting that there are still several, although relatively rarer, driver mutations are yet to be identified (**Figure 6**).[43] The PCAWG results revealed that there is still a significant fraction of cancer patients without any known genetic therapeutic target. This is why the attention of clinicians and researchers cannot be exclusively focused on DNA.

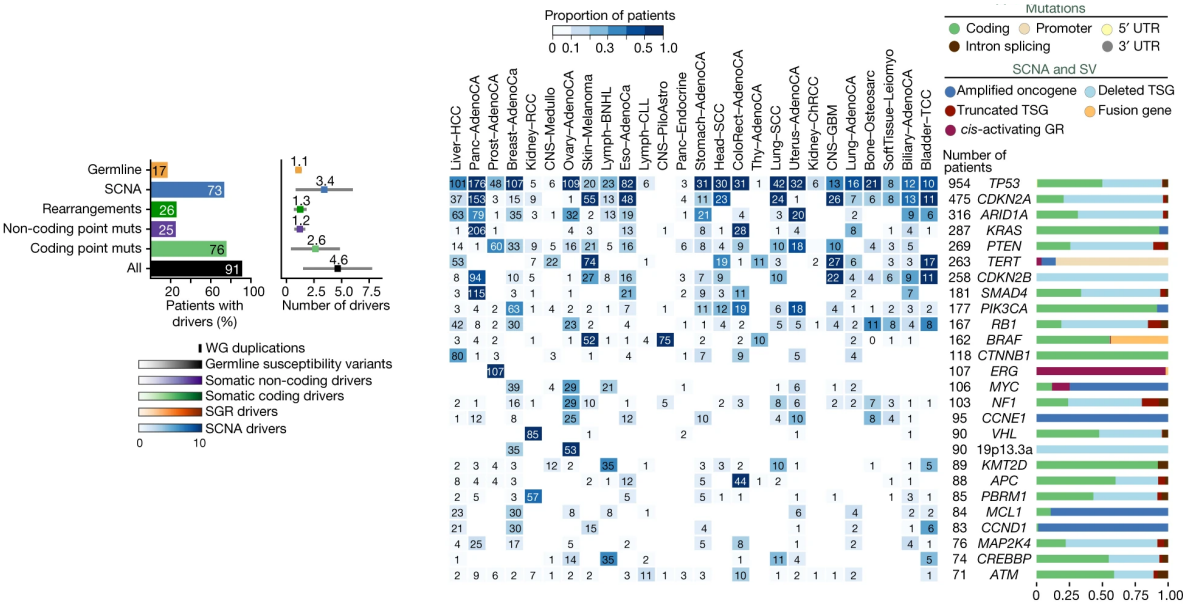


***Figure 6**. Left: quantification of different driver events observed in the PCAWG cohort (n = 2,583). The horizontal bar plot shows the percentages, the dot plot shows the average number of each type of driver events across the whole cohort. Right: Heatmap with the frequencies of the most common driver alterations observed in the PCAWG cohort across different cancer types. Proportions of each event type are also reported as a barplot. Figure obtained and adapted from "Pan-cancer analysis of whole genomes".*

Similar to DNA sequencing, RNA sequencing has a critical role in oncology diagnostics. Some of the most critical information provided by RNA sequencing in the diagnostic setting involve genomic rearrangements. It is known that chromosomal rearrangements can produce

chimeric gene fusions. Just like mutations, fusions can be driver events that vary largely between cancer types and can either promote oncogenic gain-of-functions or inactivate tumor suppressors. Their presence has initially been acknowledged in hematological cancers and soft tissue sarcomas, a rare and heterogeneous subgroup of cancers that originate from mesenchymal cells. From a therapeutic point of view, their detection in solid tumors is usually not clinically relevant. Lung and prostate cancers are among the most eminent exceptions. In lung cancer, specifically in the non-small Cell subtype, the clinical role of the fusion between the Echinoderm Microtubule-associated Protein-like 4 (*EML4*) and the Anaplastic Lymphoma Kinase (*ALK*) genes is able to provide to clinicians a clear understanding of the oncogenic process in patients. In the prostate cancer the fusion between the Transmembrane protease, serine 2 (*TMPRSS2*) and ETS-related gene (*ERG*) genes, found in 40%-78% of the prostate cancers studied to date, making it one, if not the, most important biomarkers in this cancer type.[51–53]

Gene Expression as Clinical Biomarker

RNA sequencing is starting to become embedded in precision oncology thanks to the ability to retrieve transcriptional rearrangements with potential oncogenic effects. Transcriptomic data can be further used to perform gene expression profiling in the inspected tissue. By observing the uneven coverage values, it is possible to deduct the transcriptional levels across all the observed genes, prior proper normalization based on different factors such as gene lengths.

The general idea is that by comparing the different levels of expressions between healthy tissues and tumors, it is possible to obtain a better insight on the oncogenic molecular processes.[54] A well-known application of gene expression profiling can be seen in breast cancer subtype classification, where gene expression values are usually collected from gene expression microarrays or Reverse transcription polymerase chain reaction. Breast cancer has 4 molecular subtypes: luminal A, luminal B, HER2, and triple negative.[55] The expression levels of the estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2 (HER2), and proliferation genes are major protagonists for this classification and, by crossing results with additional analyses such as immunohistochemistry, can help clinicians in their decisions. For example, a breast cancer showing an amplification of the *HER2* gene can be treated with HER2 inhibitors.[56]

Gene expression data has also been used in the past years to evaluate cell populations in tumor microenvironments. In particular, it is possible to infer the presence of tumor infiltrating

lymphocytes, which play a role in cancer growth and in determining response to therapy. The concept behind this idea is the following: the presence of tumor infiltrating lymphocytes has an expected impact on gene expression alterations; by profiling said impact to the gene expression values observed in a tumoral tissue, it is possible to infer proportions of cells in the tumor microenvironment. The computational approaches able to perform such analysis are based on deconvolutional algorithms (in this case, algorithms able to assume tumor infiltrating lymphocytes proportions by correlating their potential role in the gene expression values).[57–59]

In 2019 Rodon et al. published the results obtained from the WINTHER clinical trial.[60] In this trial 107 cancer patients were enrolled. 69 were treated accordingly to the presence of actionable DNA driver alterations (DNA arm), the remaining 38 had not such genomic biomarkers and were selected for RNA driven therapies (RNA arm). Gene expression data was obtained from the patients in the RNA arm, processed by comparing gene expression differences between tumor and healthy biopsies, and computationally analysed to identify potential drugs able to provide a good response. For example, one patient with a refractory gastrointestinal neuroendocrine tumor showed a clear overexpression of *AKT2* and *AKT3* genes, whose roles in the oncogenic phosphatidylinositol 3 kinase - AKT - mechanistic target of rapamycin (PI3K/AKT/MTOR) cell signaling pathway are well known. The patient was treated with an mTOR inhibitor and showed a good response. Median overall survival of patients treated using tailored therapies based on either a DNA actionable alteration or on a computationally designed RNA profile was significant better than overall survival of patients treated with non-precision medicine based therapies like adjuvant chemotherapy (25.8 months for the former, versus 4.5 months for the latter), suggesting that DNA and RNA sequencing data can both be used by clinicians to profile a patient's disease and better drive treatments.

IV. Bioinformatics in Precision Oncology

The development of several open-access data sharing platforms able to provide an user friendly way to investigate cancer patients cohorts is critical in both the clinical and the research landscapes. cBio Cancer Genomics Portal[61] and XenaBrowser[62], for example, offer a way to visualize non-synonymous mutations, DNA copy-number data, gene expression data, protein-level and phosphoprotein level data, DNA methylation data and de-identified clinical cancer data, while also providing some user friendly features to perform quick pre-defined analyses.

Large-scale sequencing projects and Their Role in Precision Oncology

Both cBio Cancer Genomics Portal and XenaBrowser rely on data provided by a number of public datasets. The Cancer Genome Atlas (TCGA) project, launched in 2005 under the supervision of the NIH, and the International Cancer Genome Consortium (ICGC) project, launched in 2008, are the two most important cancer specific and publicly accessible resources to date. The aim of these projects is to provide large sources of multi omics data obtained from different cancer type cohorts.

They feature a collection of over 20,000 samples, including primary tumors, metastatic and germline samples, spanning over different cancer types (33 and 20, respectively for TCGA and ICGC). Offered information covers several data categories, including the most important omic sources. To cite a few of them, in TCGA we can retrieve clinical, DNA alterations, imaging and protein expression data.

The utility brought by these resources can enable crowd-sourced analyses, making it possible for researchers with no access to novel patients data to validate their theories and for clinicians to perform quick analyses with the significant statistical power provided by the large number of patients (**Figure 7**). These are just some easy examples to prove how public available datasets can affect diagnostic, research, disease gene discovery and therapies development and, ultimately, have a real impact on the improvement of precision oncology. It is important to point out that said resources can be truly crowd-sourced once individual researchers and research groups actively submit the information gathered on their samples. The continuous influx of data can increase the depth and statistical power of these resources.

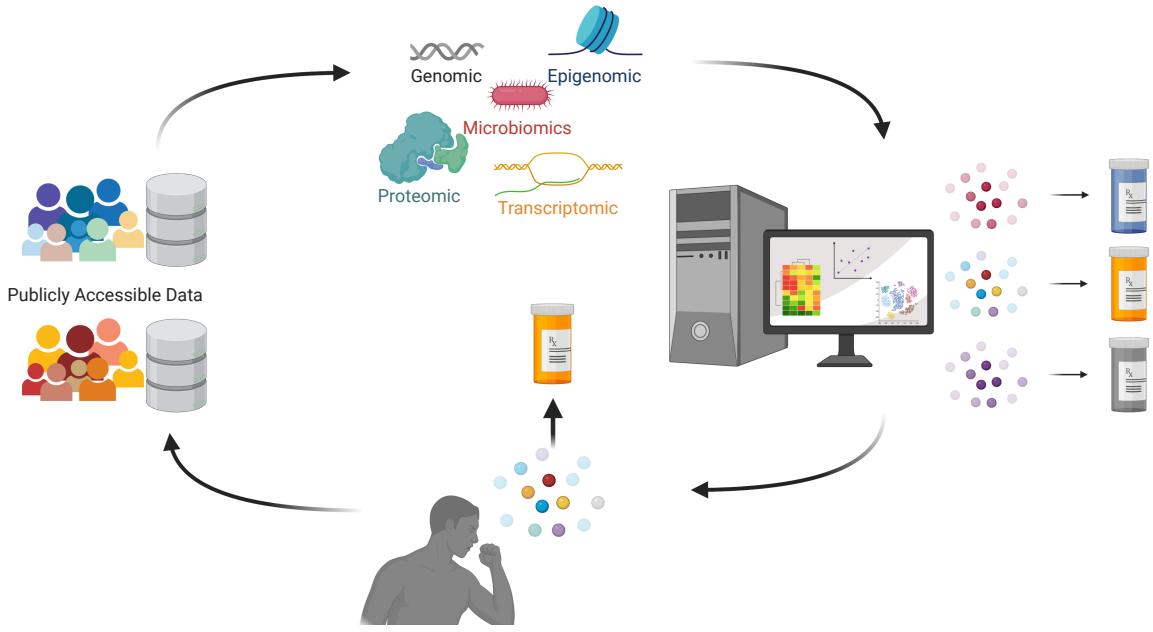


***Figure 7**. Example of a crowd-sourced clinical analysis. Omics data is collected from publicly accessible databases and used to identify biomarkers able to predict therapies response. Patients profiles are then produced by clinicians; if the same biomarkers are found it is possible to define the best way of action. Ideally, the data obtained from the patient can be de-identified and added to the public datasets, to provide further statistical power for future analyses.*

Variant Callers

Bioinformatics has a key role in the identification of genomic variants from sequencing data. The number of bases covered from sequencing experiments, even for relatively small targeted panel based sequencing, is simply too large to expect a manual review, making computational methods the only realistic way to identify mutations. Said methods need not only to be able to process such a large amount of data, but they also need to be accurate at the base resolution, to identify single base mutations. Moreover, they also need to be reproducible and able to filter sequencing noise and artefacts, artificially induced mutations produced either during the preparation of the samples or during the sequencing.[63]

The sequencing of a healthy tissue to act as a germline control for the data obtained from the tumoral tissue can be a meaningful addition for the analyses, able to remove germline mutations, when needed, and to remove potential sequencing artifacts.[37,64] Differentiating somatic and germline mutations can be a challenging task due to the fact that the latter class of variants vastly outnumber the former, making it easy to mistake a somatic mutation as a germline variant. Strelka[65,66] and Mutect[67] are two of the most frequently used somatic

variant callers today, both of which rely on matched tumor-normal data. Moreover, their performance on modern NGS sequencing data is well acknowledged, but their use on Ion Torrent data, one of the most frequently used platforms in the clinical setting, is not as reliable.[68]

While the importance of the sequencing of a germline control is widely recognised, it is not always performed in the routine diagnostic to reduce costs and have faster results. Moreover, it is not unusual for researchers to analyse old samples, usually kept either as fresh frozen or stored in formalin-fixed paraffin-embedded (FFPE) blocks, making it impossible to retrieve a proper control healthy tissue processed in the same conditions as the tumor tissue. A combination of bioinformatics and big data can provide enough statistical power to offer a solution for this issue, while not as optimal as a proper control. Mutations frequently found in population databases are likely not to be driver events, but a further manual review of these mutations is needed in order not to improperly remove hotspot driver mutations.

Machine Learning Applications in Precision Oncology

Artificial intelligence approaches such as Machine Learning and Deep Learning can further revolutionize the study of omics data. Machine learning is able to take advantage of advanced mathematical functions to build prediction models based on the interplay of the complex associations identified computationally within the input data. In a precision medicine context this could mean, for example, expanding patient care by combining all the data recovered from patients to better profile their disease and help the strategization of the therapies. The novelty brought by machine learning based studies, over the classic statistics based ones, is the ability to detect new, unexpected connections between every bit of digitized healthcare information retrieved from huge cohorts of patients profiles at once, something cannot be easily obtained from the classic univariate statistical studies.

There are many different types of machine learning models. One of the classic ways to classify said models is using the ‘unsupervised’ and ‘supervised’ classes.[69] The former class is based on the assumption that the model is able to automatically identify meaningful inferences from data that lacks a prior human-made classification, to perform an unrestrained prediction.[70] The aim of this kind of models is to gain unprecedented and novel insights from the data. However, these unsupervised approaches provide results that can be hard to interpret. This lack of understanding, a natural consequence of the ‘black box’ nature of this kind of methodologies, does not always translate well in a patient care setting.[69] The latter class includes all the approaches able to build a prediction model using previously obtained

information as a training dataset. In other words, the supervised model is able to learn how to perform the predictions from humans and, ideally, outperform them.[71,72] For example it is theoretically possible to predict an optimal therapy response by observing the responses to therapies previously decided by clinicians and processing the omics profiles of the involved patients (**Figure 8**). It is also possible to let a classifier define the stage, type and subtype from histopathology slides, and provide faster and more accurate results.[72,73] The general workflow of supervised methods starts from the selection of a training dataset, that has to be processed in order to make the information as polished as possible, by selecting the best possible subset of features and samples to be used in the next phases. The trained model will be tested and validated in order to obtain the most accurate results. Finally, the model is ready to be used on new data.

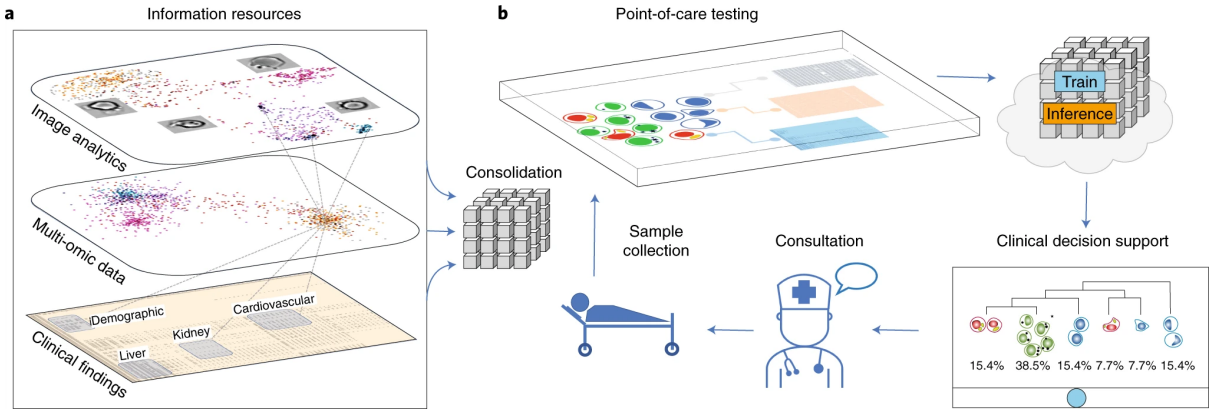


***Figure 8**. Diverse and plentiful information can be efficiently integrated by machine learning to assist clinical decision making. a) Incorporation of integrated health data (histopathology images, omic datasets, and clinical tests, for example) can be used as inputs for network analysis or machine learning algorithms to identify relevant connections (grey lines) between the diverse information and help diagnosis. b) Pre-trained machine learning models allow feature extraction, data integration and inference to assist the clinical decisions. Data obtained from the patient will be integrated in the previously obtained one, to increase the information and, potentially, detect new connections. Image obtained from "Leveraging machine vision in cell-based diagnostics to do more with less".*

Machine learning based approaches are also having a huge impact on many aspects of drug development. This starts from the very first step, the discovery of the molecule. Here these computational approaches can define novel potential molecules, and help the pharmacokinetics analyses, where models can predict aspects such as absorption,

distribution, metabolism, excretion and toxicology. This not only can help the drug development process as a whole, but it also makes it faster.[69]

Despite the significant improvement artificial intelligence based approaches are bringing to precision medicine, they also come with significant shortcomings. First, they heavily rely on a large amount of data for both the training and evaluation of the models. A shortage of information might lead to ‘overfitted’ models, unable to properly process data from new patients. Second, a model will never be able to be 100% accurate in real world scenarios.[73] This means that even approaches able to perform excellently on population-scale can potentially fail on a single patient, making the supervision of properly trained experts still mandatory in the proper healthcare process. However, implementation of artificial intelligence is only in its early days and the continuous improvements, both on the methodologies and on the computational power at their disposal, are likely to redefine medicine itself in the upcoming years.

2- Rationale and Aims of the Thesis

The main objective of my project was to develop computational approaches to help expand precision oncology.

Chapter I

Modern precision oncology relies heavily on the detection of genomic variants. The Ion Torrent platform is frequently used in the diagnostic setting due to its low costs, fast execution and modest requirement in terms of genetic material, but lacks optimized analysis workflows for custom targeted sequencing panels. In the first section of this chapter we describe PipeIT, the tumor-germline matched pipeline we have developed to offer to Ion Torrent users a reliable variant caller that only needs minimal manual curation. The tool is based on the Singularity container technology, that ensures easy to perform and reproducible results. In the second part we describe PipeIT2, the tumor-only variant caller pipeline as an extension to PipeIT. Matched germline sequencing data is not always available to clinicians and researchers, PipeIT2 was developed to eliminate the need of this germline control and still provide trustworthy results. In both sections we discuss in depth about both the pipelines and their validation on different cohorts of cancer sequencing data.

Chapter II

We investigated the potential role of gene expression data to detect the presence of oncogenic pathways activation in patients. In order to do so, we developed and tested a machine learning classifier, based on logistic regression, to extract transcriptomic profiles associated with hotspot driver mutations in oncogenes using TCGA publicly available data for the training and testing. The first tests were done using the *PIK3CA* oncogene to determine the prediction performances of the approach. Next, the same methodology was tested on 15 additional oncogenes to investigate the results obtained from a diverse landscape of mutation frequencies and roles in different cancer types. Finally, the model trained on *PIK3CA* data was used to infer the same transcriptomic profiles of driver hotspot mutations in an external set of samples, to infer the presence of the same oncogenic pathways activation and to determine whether gene expression data can actually be used as an alternative to genomic data to determine the presence of distinct oncogenic molecular processes in patients.

3- Results

3.1- Chapter I

Development of somatic variant calling pipelines for the detection of oncogenic mutations and to drive precision medicine

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

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

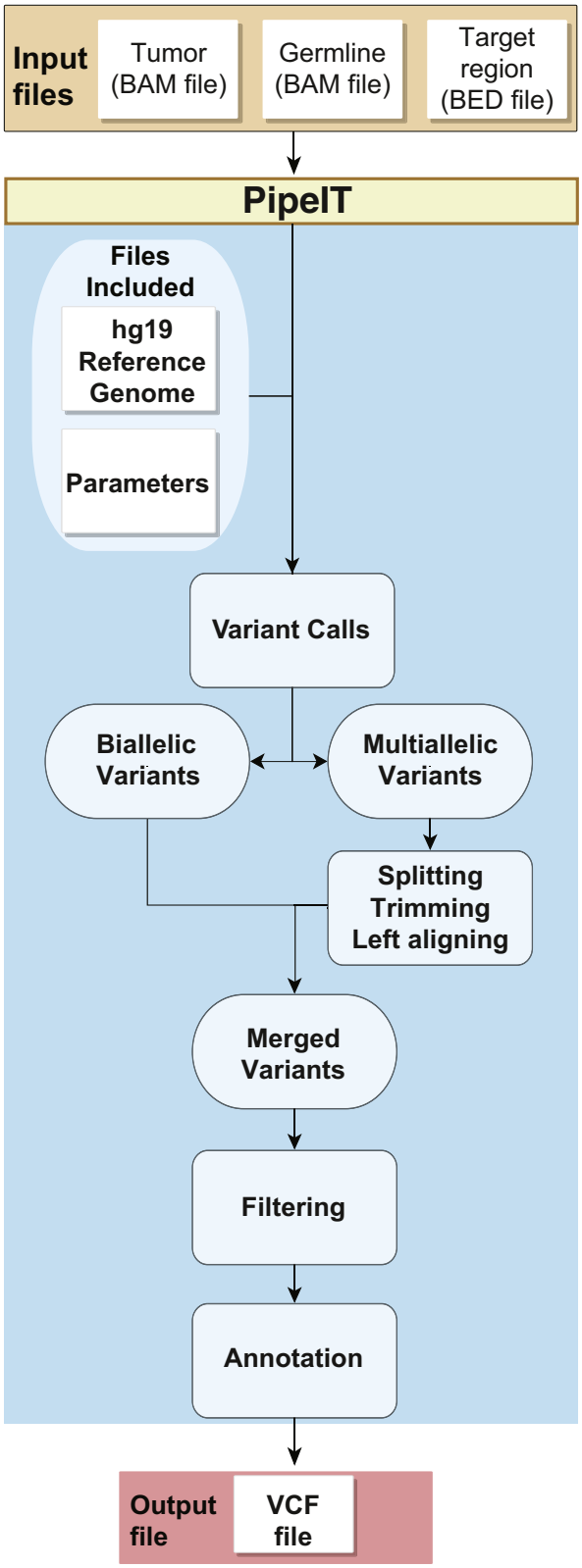


Figure 1 Overview of the PipeIT container. Flowchart showing the execution of PipeIT where the users need to provide only three files [Binary Alignment Map (BAM) files for tumor and normal samples and the target Browser Extensible Data (BED) file]. Variant calling is then performed using the Torrent 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

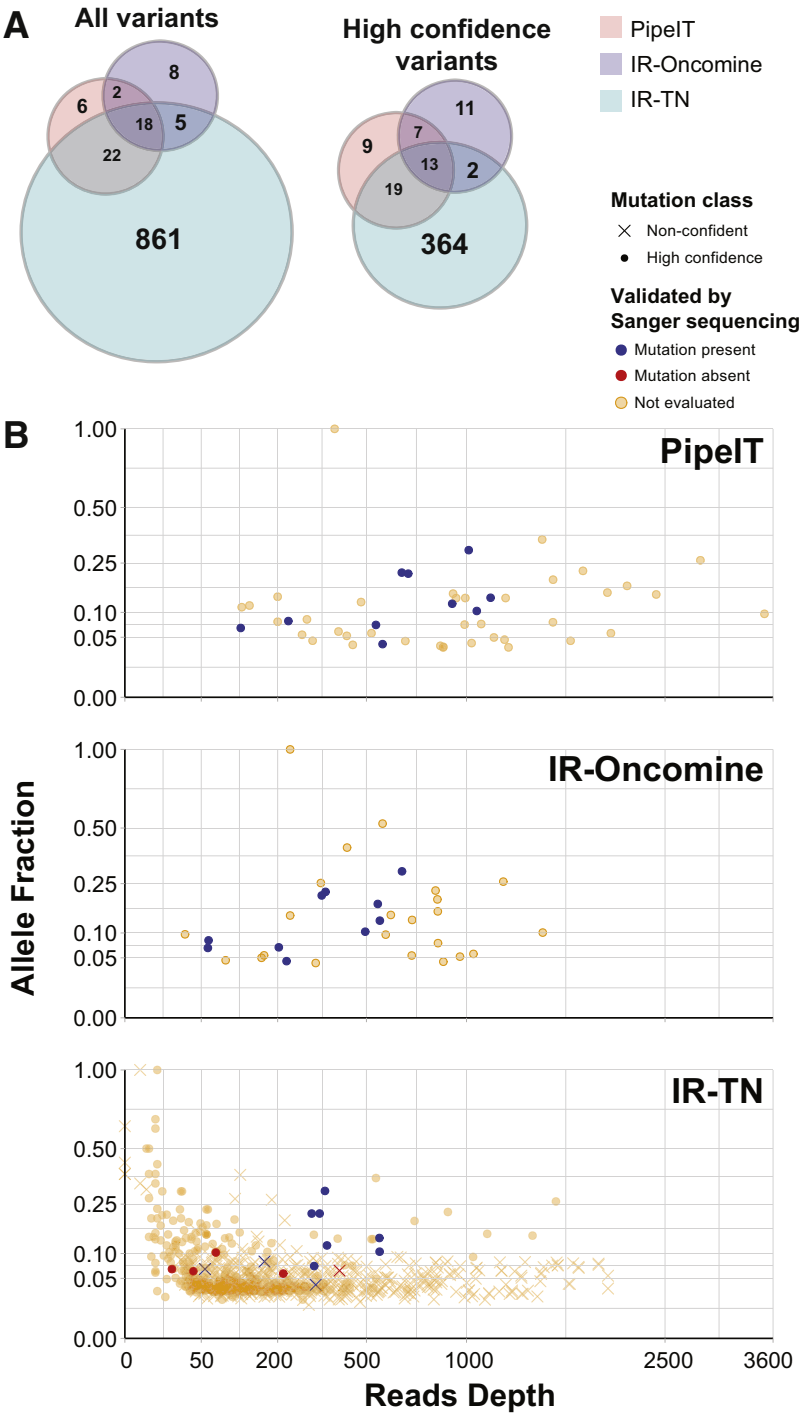


Figure 2 Comparison of mutation calls from PipeIT, Ion Reporter (IR) Oncomine Comprehensive Panel workflow (IR-Oncomine), and IR in a tumor-normal DNA analysis workflow (IR-TN) in 15 colon adenomas sequenced using the commercial Oncomine Comprehensive Panel v3. **A:** Venn diagrams showing the overlap 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.

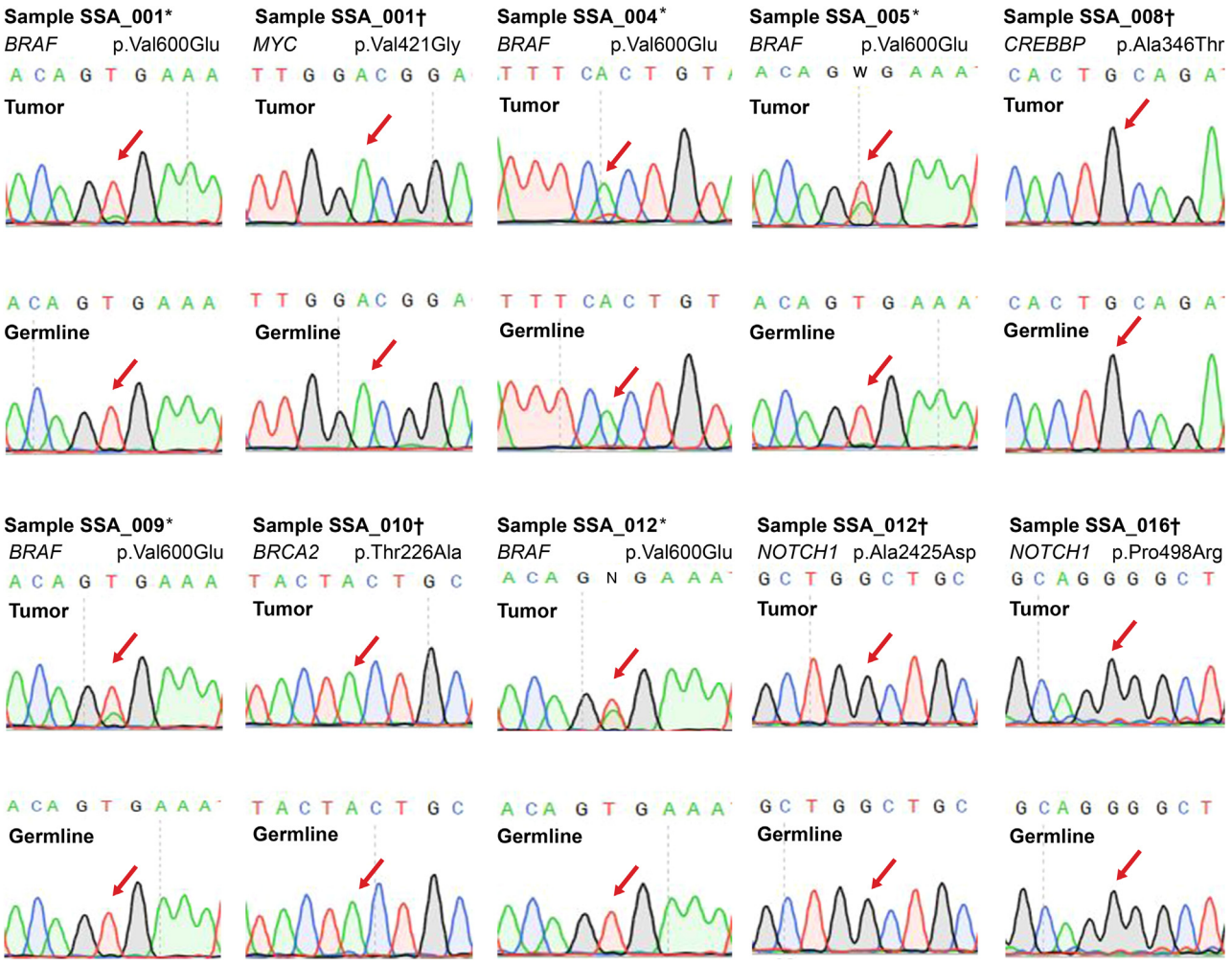


Figure 3 Validation of selected somatic mutations using Sanger sequencing. Sanger sequencing chromatograms of selected mutations being validated. **Red arrows** indicate the specific nucleotide investigated for the presence or absence of a specific somatic mutation. **Asterisks** indicate *BRAF* V600E 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

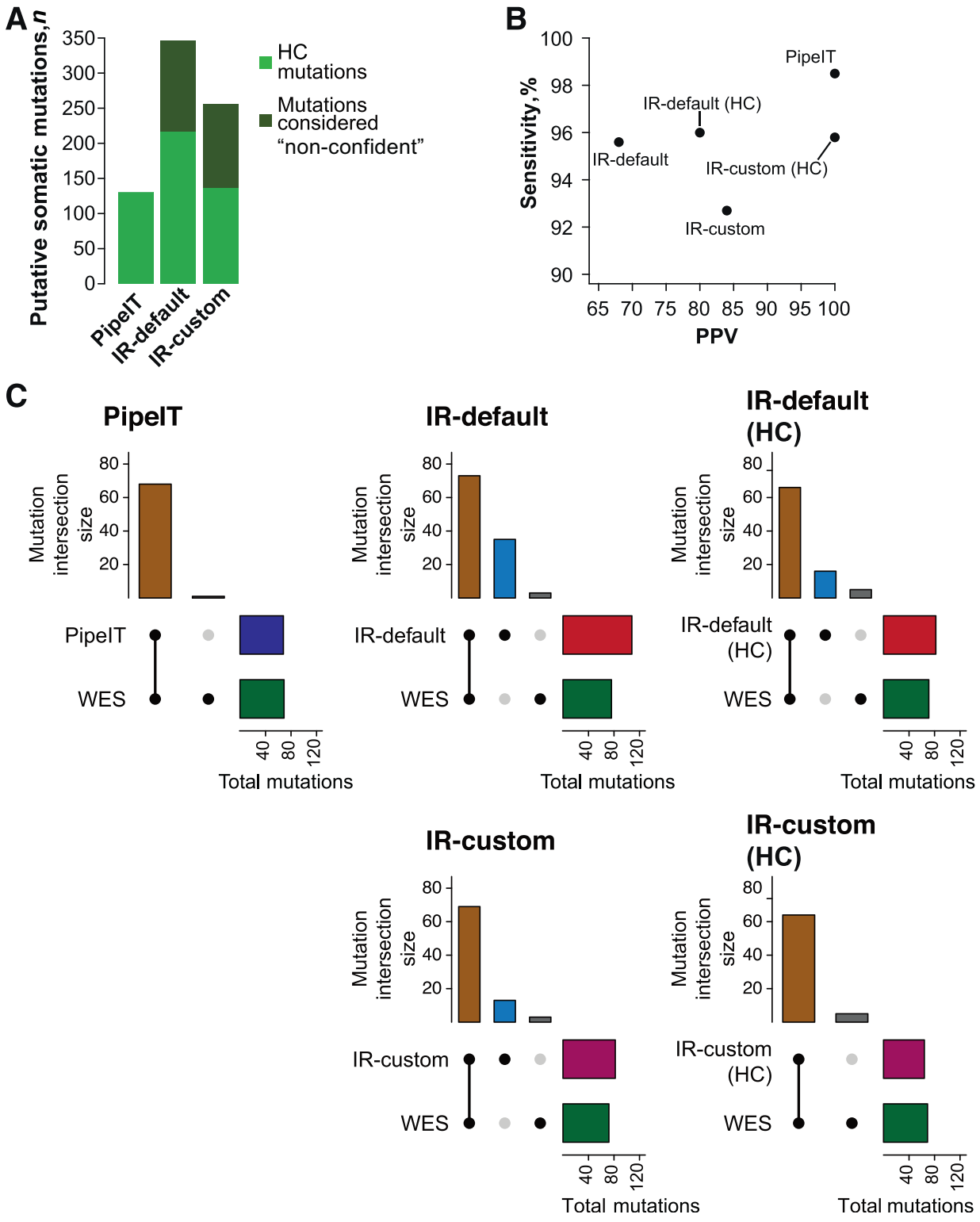


Figure 4 Comparison of mutation calls from PipeIT, IR analysis using the default parameters (IR-default), and IR analysis using the custom set of variant calling parameters used in PipeIT (IR-custom) in 10 hepatocellular carcinomas sequenced using a custom AmpliSeq panel. **A:** Bar plot shows the number 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

PipeIT2: Singularity Container for Tumor-Only Molecular Diagnostic Somatic Variant Calling on Ion Torrent NGS Platform

Andrea Garofoli*, Guglielmo Roma[$], Luigi M. Terracciano*, Gunnar Raetsch[%], Mark A. Rubin^, Salvatore Piscuoglio*,[#] and Charlotte K. Y. Ng^

*Institute of Pathology, University Hospital Basel, Basel Switzerland;

[$]Department of Biology, University of Naples Federico II, Naples, Italy;

[%]Department of Computer Science, ETH Zurich

^Department for BioMedical Research, University of Bern, Bern, Switzerland

[#]Visceral surgery research laboratory, Clarunis, Department of Biomedicine, University of Basel, Basel, Switzerland.

Running title: Ion Torrent somatic variants pipeline

Disclosures: None declared

Correspondence: Dr. Charlotte K. Y. Ng. Department for BioMedical Research, University of Bern, Murtenstrasse 40, Bern, 3008, Switzerland. Tel: +41 31 632 8779; E-mail: charlotte.ng@dbmr.unibe.ch

ABSTRACT

Modern precision oncology relies on the identification of somatic mutations in cancer patients. While the sequencing of the tumoral tissue is frequently part of routine clinical care, the additional sequencing of healthy tissue from the same patient, needed to properly exclude germline mutations, is rarely performed. We previously published PipeIT, a somatic variant caller enclosed in a Singularity container, specific for Ion Torrent sequencing data. PipeIT combines an user friendly execution, easily reproducible analysis and sensible mutation identification ability, but relies on the presence of matched germline sequencing data. To expand the original PipeIT, we have developed and here we present PipeIT2, a variant caller pipeline able to identify somatic mutations in the absence of germline control data, taking advantage of publicly accessible population databases and panels of unmatched germline samples. We show that PipeIT2 detects driver mutations in oncogenes and filters out most of the germline mutations and sequencing artefacts. Similar to its predecessor, PipeIT2 ensures reproducibility and can be rapidly executed with a single command.

Keywords: Ion Torrent, somatic mutations, next-generation sequencing, singularity, tumor-only variant calling.

INTRODUCTION

Detection of genomic alterations is becoming a critical component in the standard-of-care in modern oncology.[74,75] Typically, the detection of genomic alterations is performed using targeted sequencing panels to profile previously described cancer and actionable gene regions. The Ion Torrent sequencing platform is frequently used for targeted sequencing in the diagnostic setting due to its relatively low costs, ability to profile limited genetic material and rapid turnaround.[76]

While Ion Torrent library preparation and sequencing are relatively straightforward, the methods for sequencing data analysis are not very well-developed. Due to the differences between Ion Torrent and other sequencing platforms, most of the variant calling tools previously tested, validated and extensively used by the community are not suited for Ion Torrent data.[77] However, Thermo fisher provided its own analysis platform, Ion Reporter. We previously benchmarked the variant calling analysis of Ion Reporter, using both standard parameters provided by the machine and a set of custom, previously tested parameters. In both cases, Ion Reporter was indeed able to detect real mutations (validated by analysing WES analysis and Sanger sequencing on two different matched tumor-germline cohorts), but it also showed the presence of several false positives, notably when the analysis was performed using the standard, non optimized parameters provided by the machine.[78] This not only highlighted the need of intensive manual review, but also the potential lack of standardized and reproducible results, critical in clinical care.

We recently published PipeIT, a pipeline to detect somatic variants in matched tumor-germline samples from Ion Torrent sequencing data,[78] providing a reliable and automated workflow to perform variant calling analysis. To ensure reproducibility and ease of deployment, PipeIT was built as a Singularity container[79] image file that can be easily executed with a single command, without the need of additional software other than the Singularity platform. The

main drawback of PipeIT is the need for germline matched control data. When the goal is to identify somatic mutations, the sequencing of normal controls can be critical in order to remove germline mutations.[74,80,81] In routine clinical care, however, the sequencing of tumor-only tissue is often preferred, for time, costs and sample availability reasons. Moreover, researchers might want to analyse old, archived samples, for which matched germline controls may not be available. These scenarios significantly narrow down the contexts where PipeIT can be used and, ultimately, prevent the software from fully achieving its original aim.

Here we present PipeIT2, an extension of PipeIT to enable variant calling analyses on tumor samples without matched germline controls with a single command. PipeIT2 identifies and filters likely germline mutations by leveraging their allele frequencies in population databases and by detecting their presence in unmatched Panel of Normal (PoN) samples. We demonstrate that PipeIT2 was able to detect clinically relevant somatic mutations, while correctly identifying and removing most of the germline genomic alterations.

MATERIALS AND METHODS

The PipeIT2 Tumor-Only Workflow

PipeIT2 requires the following input files: a Binary Alignment Map (BAM)[82] file for the tumor sample, a Browser Extensible Data (BED)[83] file defining the target sequenced regions, Annovar[84] annotation files comprising of data collected in populations databases, and a Variant Call Format (VCF)[85] file with the description of the mutations found in the samples included in a PoN (**Figure 1**). The BAM format is a binary encoded version of alignment data obtained. The Ion Torrent Server automatically performs sequence alignment using the Torrent Mapping Alignment Program (TMAP) aligner against the hg19 genome and provides a BAM file as the final output. The BED format describes the regions covered by the targeted

sequencing panel. The Annovar annotation files are textual tables with the list of variants detected from the sequencing performed on large cohorts of individuals and stored in public accessible databases. Finally, the VCF file is a textual description of variants called. The PipeIT2 tumor-only default workflow comprises the following steps: 1) variant calling; 2) post-processing variants; 3) variant annotation; 4) annotation-based variant filtering; and 5) optional PoN-based variant filtering (**Figure 1**).

The variant calling step (step 1) is performed using the Torrent Variant Caller (TVC, v5.0.3, Thermo Fisher Scientific) using the same low stringency parameters used in the tumor-germline workflow (i.e. the original PipeIT)[78], adapted from the standard parameters used in Ion Reporter analyses. Specifically, using a quality threshold of 6.5, a variant score equal or higher than 10, a minimum coverage of 8 reads for somatic nucleotide variants (SNVs) and 15 reads for small insertion/deletions (indels), a variant count of at least 4 and, finally, a variant allele frequency (VAF) of 5% for long assembled indels. The parameters for TVC are stored within the Singularity container in a JSON file. Users can also submit different TVC parameters by providing a JSON file.

The post-processing step (step 2) is performed to facilitate downstream filtering and annotation. It comprises the same operations described in PipeIT.[78] First, the 'norm' function in BCFtools is used to split multiallelic variants into monoallelic variants. The variants are then left-aligned using the Genome Analysis Toolkit (GATK) 'LeftAlignAndTrimVariants' tool.[82,86]

The annotation step (step 3) is performed by Annovar and by GATK.[84,86] Annovar is used to annotate the allele frequencies of the variants in the population databases. PipeIT2 (version 1.2.15) uses Annovar databases files from the 1000 Genomes Project[87], the Exome Aggregation Consortium[88], the NHLBI Exome Sequencing Project[89] and the Genome Aggregation Database[90]. The GATK 'VariantAnnotator' tool is then used to add information

regarding the presence and the length of homopolymer regions. These annotations are added to the INFO field of the post-processed (step 2) VCF files.

The annotation-based variant filtering step (step 4) is implemented using the GATK 'VariantFiltration' tool to flag the variants based on read counts and on the annotations from step 3. By default, PipeIT2 removes variants that do not meet all of the following requirements: a minimum depth of 20 total reads (corresponding to the INFO field FDP), a minimum 6 reads supporting the variant (FAO) and a VAF (FAO/FDP) of at least 0.1. Moreover, the variant allele must be observed in at least 3 forward (FSAF) and 3 reverse reads (FSAR), with a strand bias (FSAF/FSAR) smaller than 0.2, in either direction. Based on the information annotated in step 3, a variant is removed whenever it is observed with a frequency equal to or higher than 0.5% in any of the four population-level databases (see step 3). A variant with a VAF between 0.4 and 0.6, or greater than 0.9 is also removed if it is also found at any allele frequency in any of the four population-level datasets. By default, PipeIT2 also removes variants in homopolymer regions of length > 4. Finally, synonymous and non coding mutations are removed from the results, unless stated otherwise by the user. Hotspot variants are whitelisted and kept even when they do not meet the above filtering criteria, due to their likely role in cancer development.[91,92]

The final step is the optional PoN-based variant filtering (step 5). As an additional way to remove likely false positive variants, which may include germline variants not removed in step 4 and systematic sequencing artefacts, PipeIT2 can use a user-submitted VCF file obtained from a panel of (unmatched) germline samples (see below)Variants identified in the PoN VCF are filtered to obtain the final list of variants.

Generation of the Panel of Normals (PoN) VCF file

The PoN VCF file is produced by steps 1 and 2 of PipeIT2 as described above from the BAM files of a panel of (un-matched) gemline samples. The VCF files are then merged using GATK 'CombineVariants' function using the UNIQUIFY option and retaining the mutations found in at least 2 of the input germline samples.

Building the PipeIT2 Singularity Container Image

The original PipeIT Singularity container has been updated to include the PipeIT2 tumor-only workflow described above. The file is a read-only squashfs file system Singularity image built on a CentOS7 Docker image as a base, as previously described.[78] PipeIT2 provides the entry points to perform both the matched tumor-germline and the new tumor-only workflow.

Evaluation of the PipeIT2 tumor-only workflow

Sequencing data for 15 formalin-fixed paraffin-embedded colon adenocarcinomas[93] and for 10 frozen samples of hepatocellular carcinoma (HCC)[94] were retrieved from our previous publication.[78] To evaluate the performance of PipeIT2 and the contribution of the PoN-based variant filtering step (step 5 above) 4 different analyses have been performed on HCC and colon adenocarcinoma samples using either PoN VCF files obtained from an increasing number of randomly chosen unmatched germline samples (2, 4 and 8) or without submitting any PoN file.

Using the list of mutations from the PipeIT publication[78] as the benchmark, the mutations detected in PipeIT2 were evaluated as: true positive mutations (mutations called by both workflows), false positives (mutations called by the tumor-only workflow, but not by the tumor-germline workflow), and false negatives (mutations detected by the tumor-germline workflow, but not by tumor-only workflow). The similarities observed between the number of mutations

called by the different analyses were quantified using the Jaccard Index, defined as the size of the intersection divided by the size of the union.

Visualization of BAM files

Integrative Genomics Viewer (IGV) [95] was used to visualize the BAM files and search for the presence of false positive mutations across the original tumor-germline matched pairs and the unmatched germline samples used to build the PoN files used to run the analyses.

SOFTWARE AVAILABILITY: PipeIT2 is freely available at http://oncogenomicslab.org/software-downloads/

RESULTS

Running the PipeIT2 tumor-only workflow

To provide an effective somatic variant calling analysis on tumor data originated from Ion Torrent platform in the absence of a matched germline, we updated the original PipeIT functionality to allow the users to choose between the classic tumor-germline (PipeIT) and the new tumor-only (PipeIT2) analyses. Similar to the PipeIT, the new PipeIT2 Singularity image provides most of the data needed to perform the whole analysis, with the only exceptions of the public datasets data and the PoN VCF file. The PipeIT2 tumor-only workflow can be executed in a single command as follows:

```
singularity run PipeIT.img -t path/to/tumor.bam -e path/to/region.bed -c path/to/annovar/humandb/folder (-d path/to/PoN/file.vcf)
```

The new tumor-only PipeIT2 workflow takes advantage of two resources in order to remove likely germline and artefactual variants: population sequencing data and an unmatched PoN. PipeIT2 makes use of four publicly accessible population-level variant databases, namely, the

1000 Genomes Project[87], the Exome Aggregation Consortium[88], the NHLBI Exome Sequencing Project[89] and the Genome Aggregation Database[90]. PipeIT2 can be used to retrieve the population data with the following commands:

```
singularity exec PipeIT.img annotate_variation.pl -downdb -webfrom annovar -buildver hg19 esp6500siv2_all humandb/
singularity exec PipeIT.img annotate_variation.pl -downdb -webfrom annovar -buildver hg19 1000g2015aug humandb/
singularity exec PipeIT.img annotate_variation.pl -downdb -webfrom annovar -buildver hg19 exac03 humandb/
singularity exec PipeIT.img annotate_variation.pl -downdb -webfrom annovar -buildver hg19 gnomad_genome humandb/
```

The PoN VCF is generated from the BAM files sequenced on the same platform used for the tumor samples, the Ion S5XL system, and can be generated using the --pon-build parameter followed by a text file with the paths for each of the germline BAM files. The whole command shall be:

```
singularity run PipeIT.img -t path/to/tumor.bam -e path/to/region.bed -c path/to/annovar/humandb/folder (--pon-build path/to/list.txt)
```

Evaluation of the PipeIT2 tumor-only workflow

We used the PipeIT2 tumor-only workflow, with standard parameters, to analyze the 15 colon adenocarcinomas and 10 HCCs previously published.[78] The 15 FFPE colon adenocarcinomas, together with their matched germline counterparts, were sequenced using the Oncomine Comprehensive Panel v3, while the 10 fresh frozen HCCs and their germline counterparts were sequenced using a previously published custom HCC targeted sequencing panel.[94] We compared the results from the PipeIT2 tumor-only workflow against the set of non-synonymous somatic mutations previously defined by the matched tumor-germline workflow of PipeIT.[78] In addition, we wanted to investigate whether the use of a larger PoN

VCF file could outperform the use of a PoN VCF file obtained from a smaller pool of germline samples or not using a PoN at all. For each of the 25 samples 3 PoN VCF were generated from, respectively, 2, 4 and 8 randomly chosen unmatched germline samples from other samples of the same cancer cohort (i.e. excluding the matched germline). We analyzed each of these 25 samples without a PoN and with each of the 3 PoN VCFs. PipeIT2 performance was evaluated in terms of true positive, false negative, and false positive mutations.

Across the analyses performed on the colon adenocarcinoma samples using PoN generated from 0, 2, 4 and 8 unmatched germline samples (they will be called, respectively, PoN0, PoN2, PoN4, and PoN8), the number of true positives detected was almost identical (**Figure 2 A**). The first three analyses each identified the same 24 true positive mutations, while the PoN8 analysis led to the discovery of 23 true positives (Jaccard index (JI) = 1 between PoN0 and PoN2, PoN2 and PoN4, JI = 0.958 between PoN4 and PoN8). The one missing mutation in the PoN8 analysis, a missense variant in the *BRAF* oncogene, was removed because it was also present in at least two of the randomly chosen germline samples used to generate the PoN.

Next, the amount of detected false positives was investigated (**Figure 2 A**). Both PoN0 and PoN2 led to a total of 20 false positive mutations across the 15 samples, the PoN4 led to 18 mutations and, finally, PoN8 led to 12 mutations (JI = 1 between the PoN0 and the PoN2, JI = 0.904 between the PoN2 and PoN4, JI = 0.631 between the PoN4 and PoN8, JI = 0.571 between the PoN0 and PoN8). It should be noted that one mutation called in the SSA005 sample, a mutation in the *NOTCH2 [96]*, was also called by the PipeIT tumor-normal workflow but was manually reviewed due to the low allele fraction. Moreover, the same mutation, manually reviewed and removed due to its low AF as explained in the previous analysis,[78] was called by PipeIT the sample SSA014 but correctly filtered by PipeIT2 (**Supplementary Figure S1**).

Lastly, we investigated the amount of somatic mutations that were not called by PipeIT2 (i.e. false negative, **Figure 2 B**) and the reasons that led the pipeline to their removal as likely somatic variants (**Figure 2 C**). The amount of false negative mutations was coherent with our results on true positives: the numbers of false negatives was the same across the 4 different analyses (JI = 1 between PoN0, PoN2 and PoN4, JI = 0.928 between PoN4 and PoN8), with the aforementioned *BRAF* mutation being the only exception. The total numbers of false negative mutations across the colon adenocarcinoma samples were 13 for the PoN0, PoN2 and PoN4 analyses. 14 false negative mutations were found in the PoN8 analysis. Looking at the results obtained from the PoN8 analysis, the principal reason for their removal was their low VAF, leading to the removal of 9 mutations. The 5 remaining mutations were filtered out because of the PoN filter step (1 mutation, *BRAF*), the low number of reads (1 mutation) or a combination of low allele fraction with either the insufficient number of reads (2 mutations) or the presence of the mutation in the population datasets (1 mutation).

We then investigated the HCC samples. It is worth to mention that one of the HCC samples (HPU207) was previously identified as hypermutated compared to the average case sequenced with the help of a sequencing panel [94], offering us the opportunity to benchmark a more challenging sample. The size of the PoN had no effect on true positive mutations detection (JI = 1 for all the pairings). 47 mutations were detected across the 10 HCCs, 34 of them belonging to the HPU207 sample (**Figure 2 D**). When a PoN VCF was provided, its size did not alter the amount of false positive mutations (JI = 1 for all the pairings, excluding PoN0) (**Figure 2 D**). On the other hand, the lack of a PoN had a significant impact on the number of false positives detected. 9 of the 10 HCC samples (HPU202 being the only exception) had between 1 and 3 additional germline mutations wrongly detected as somatic, for a total of 23 (JI = 0.238 between PoN0 and any of the other analyses) (**Supplementary Figure S2**). The absence and the size and of the PoN file did not impact the detection of false negative mutations (JI = 1 for all the pairings) (**Figure 2 E**). 12 mutations (9 from the HPU207 sample) previously identified by the PipeIT matched tumor-normal workflow were missed by PipeIT2.

Finally, we once more discerned the reasons behind the filtering of the false negative mutations, showing again the importance of allele fraction, which caused the removal of all mutations. All 12 mutations missed from the results returned by PipeIT2 were detected with an allele fraction lower than the 0.1 threshold value (**Figure 2 F**).

The size of the PoN files had almost no impact on the amount of true positive and false negative mutations in both the cohorts. False positive mutations observed, however, were diminishing significantly as the size of the PoN files was increasing, with a difference of 8 false positive mutations between the PoN0 and PoN8 analyses in the colon adenocarcinoma cohort and 18 between the PoN0 and PoN8 analyses in the HCC cohort.

Validation of false positive mutations

We finally investigated whether the false positive mutations detected by PipeIT2 could be seen in the original germline matched control BAM file (**Figure 3**). We used IGV [95] to manually detect the presence of said false positive mutations in said matched tumor and germline BAM files and in the randomly chosen unmatched germline samples used to build the PoN VCF files used to perform the analysis. All the false positive mutations proved to be genuine germline variants. **Figure 3** shows the rationale of this validation on 4 different, randomly chosen mutations detected by PipeIT2: a mutation found in the *ARID1A* gene from the HPU201 sample, a mutation in *MYC* from HPU202, a mutation in *RICTOR* from SSA004, and a mutation in *ATRX* from SSA011. All the aforementioned mutations were observed in both the tumor and germline matched BAM files, but were not present in any of the samples randomly detected to create the PoN files.

DISCUSSION

Precision oncology care is increasingly reliant on the identification of somatic DNA alterations in cancer patients. DNA sequencing of tumor tissues with targeted genomic assays represent, up to date, the best means to retrieve this information.[97,98] Furthermore, the additional sequencing of a healthy tissue sample from the same cancer patient is the most definitive way to determine which of the genetic alterations found in the tumor tissue are likely to somatic.[81]

Ion Torrent is one of the most popular sequencing platforms in the routine diagnostic setting due to its low costs and low sample input requirements, but it lacks a streamlined data analysis. In 2019 we developed and published PipeIT, a somatic variant caller specific for Ion Torrent sequencing data enclosed in a Singularity image file.[78] The strength of PipeIT lies in its ease of deployment and use, reproducible results, and demonstrated accuracy. On the other hand, the need for tumor-germline matched sequencing data limits the use of PipeIT in the clinical setting where germline samples are frequently not sequenced. The main reasons for the lack of sequencing data of a matched normal sample time and costs, as well as the availability of such samples. In order to address this shortcoming, we updated the original PipeIT to include an additional tumor-only pipeline as PipeIT2.

To overcome the challenges associated with the lack of a matched gemline control, PipeIT2 leverages three filtering steps. The first filter relies on more stringent filtering threshold values, compared to the ones used in PipeIT, including a threshold VAF of 10%, compared to the previous 5%. The second makes use of data obtained from the 1000 Genomes Project[87], the Exome Aggregation Consortium[88], the NHLBI Exome Sequencing Project[89] and the Genome Aggregation Database[90]. PipeIT2 removes from the final output the mutations detected in at least 0.5% (or any other user-submitted percentage) of the samples in any of these databases. The last filter needs a list of user-submitted mutations obtained from unmatched normal samples, in the shape of a Panel of Normals VCF file. This second step is

not mandatory, to make it possible for users to use PipeIT2 even in the absence of a proper PoN VCF.

To evaluate the performances of PipeIT2 performances, all the mutations identified by PipeIT2 from 15 colon adenocarcinoma and 10 hepatocellular carcinoma samples were compared to the ones identified, and previously validated, by running the original PipeIT on the same samples.[78] Using different panels of 8 randomly picked unmatched normals for each sample, a total of 70 coding mutations, some of whom important clinical biomarkers, were properly detected across the two cohorts. Nevertheless, 26 mutations were mistakenly removed from PipeIT2 output. The primary culprit of the removals (80.7% of the observed cases) was exclusively the low allele fraction detected in these mutations (which comprises a threshold allele fraction value of 10%, the standard value provided by PipeIT2). Moreover, 17 germline mutations were identified and incorrectly labeled as somatic across the 25 samples.

While both the analyses performed on the colon adenocarcinoma and the HCC showed that, with only 1 *BRAF* mutation wrongly removed in a PoN8 across 25 samples, the size of the PoN files had little to no impact on the true positive and false negative detection capacities of PipeIT2. However, we observed different results with false positives across the two cohorts regarding the detection of false positive mutations. In the colon adenocarcinoma cohort the size of the PoN had a significant impact in the removal of germline mutations wrongly classified as somatic and, therefore, included in the final results.

We also tried to determine whether the size and the presence of a PoN could significantly alter the output. We repeated the analyses for all the samples using either no PoN VCF or a PoN with 2, 4 or 8 randomly chosen unmatched germline samples (called, respectively, PoN0, PoN2, PoN4 and PoN8). Regarding the mutations detected by both the original PipeIT and new PipeIT2 workflows, the only difference observed across the analyses based on the different sized PoN files was a single *BRAF* mutation in the whole colon adenocarcinoma

cohort erroneously removed only in the PoN8, proving that it is possible, despite being rare, for PoNs to cause the removal of real somatic mutations. While the size of the PoN file had little to no impact on the number of false negative somatic mutations detected by the original PipeIT workflow, the amount of false positive mutations misclassified as somatic changed significantly. False positive mutations steadily decreased as the size of the submitted PoN file was increasing in the colon adenocarcinoma cohort. In the HCC cohort PoN2, PoN4 and PoN8 analyses coherently removed the same mutations regardless of the amount of germline samples in the PoN. However, in the PoN0 analysis the amount of false positives increased dramatically. (**Supplementary Figure S1 and Supplementary Figure S2**).

By providing a variant calling analysis able to detect somatic mutations in tumor samples lacking a proper matched germline control, PipeIT2 offers an important improvement over the original PipeIT. Thanks to filters based on population allele frequencies and on lists of variants found in panels of unmatched germline samples, PipeIT2 was able to detect and properly classify most of the somatic mutations previously identified in the tumor-germline matched analysis, including several important clinical biomarkers. In conclusion, PipeIT2 offers a powerful, user friendly and easily reproducible tool specific for Ion Torrent targeted sequencing analyses.

ACKNOWLEDGMENTS

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

AUTHOR CONTRIBUTION

S.P. and C.K.Y.N. conceived and supervised the study. A.G. and C.K.Y.N. developed the methodology. G.R., L.M.T., G.R and M.A.R. provided critical review of the results. A.G., S.P.

and C.K.Y.N. interpreted the results and wrote the manuscript. All authors agreed to the final version of the manuscript.

FIGURES AND FIGURE LEGENDS

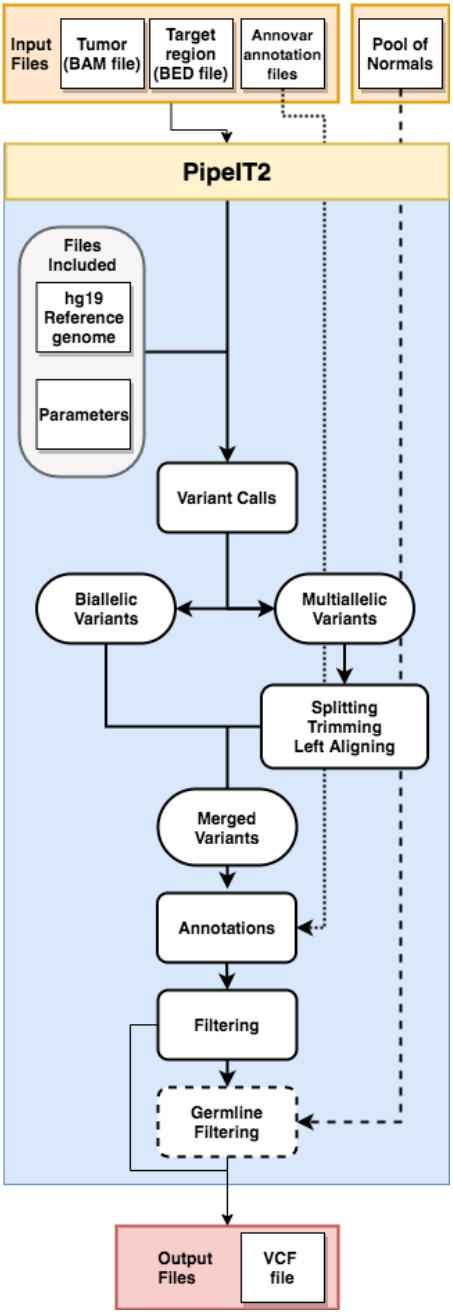


Figure 1. Overview of the PipeIT2 workflow. Flowchart showing the different steps of the PipeIT2 workflow. The user needs to provide a BAM file for the tumor sample, the BED file for the target regions and the Annovar datasets for the 1000 Genomes Project, the Exome

Aggregation Consortium, the NHLBI Exome Sequencing Project and the Genome Aggregation Database. Variant calling is then performed using the Torrent Variant Caller with the packaged parameters file. Mutations are filtered using said Annovar datasets and, when provided, a Panel of Normals. Output is returned as a VCF file.

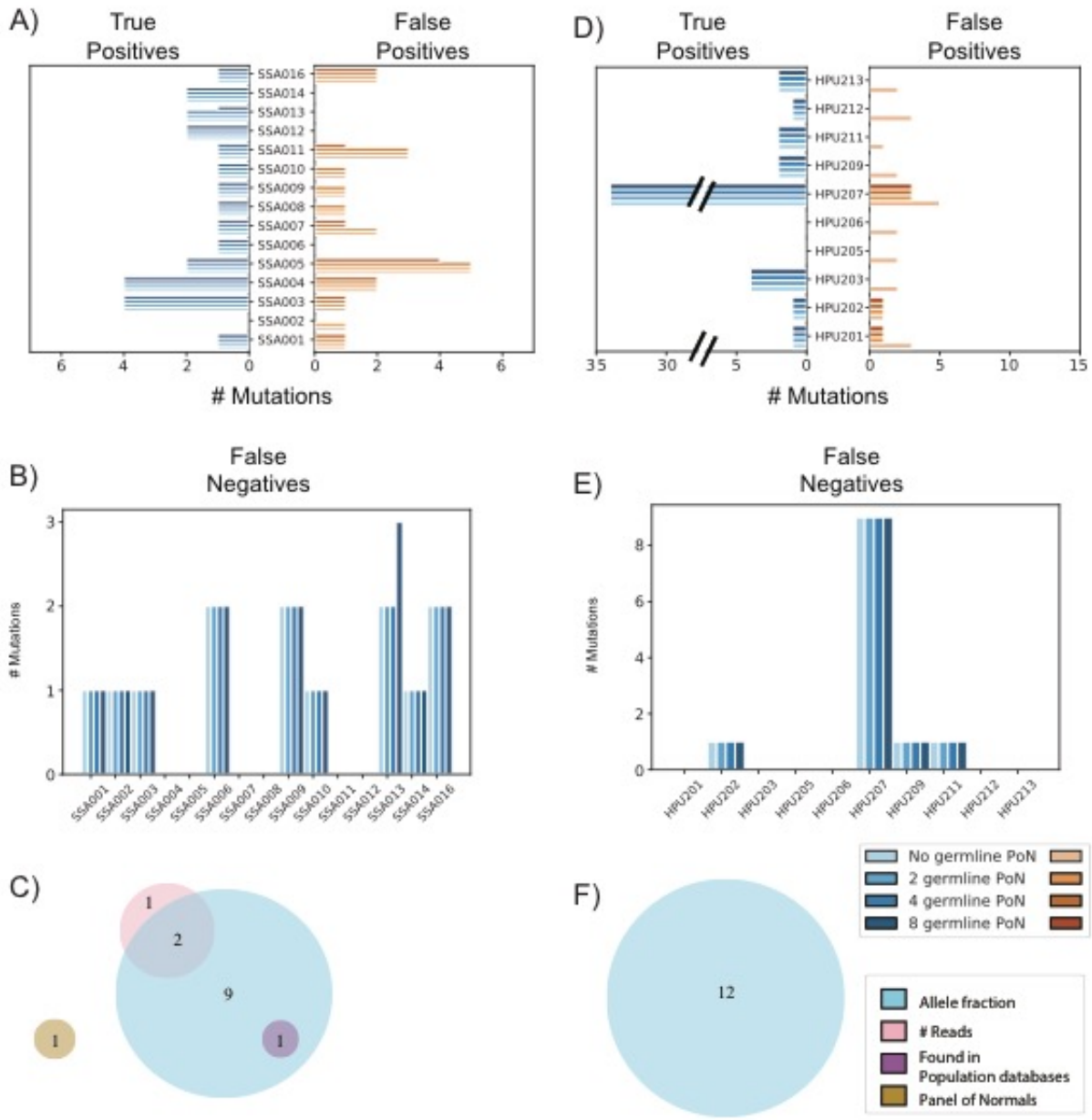


Figure 2. Mutations comparisons. A) and D) Comparison between True Positive and False Positive mutations identified by PipeIT2 using different sized PoN files on the HCC and the colon adenocarcinoma cohorts. PipeIT results were used as the golden standard. B) and E) Number of False Negative mutations identified by PipeIT2 on the HCC and colon adenocarcinoma cohorts using different sized PoN files. C) and F) Number of Somatic mutations removed for each filter type.

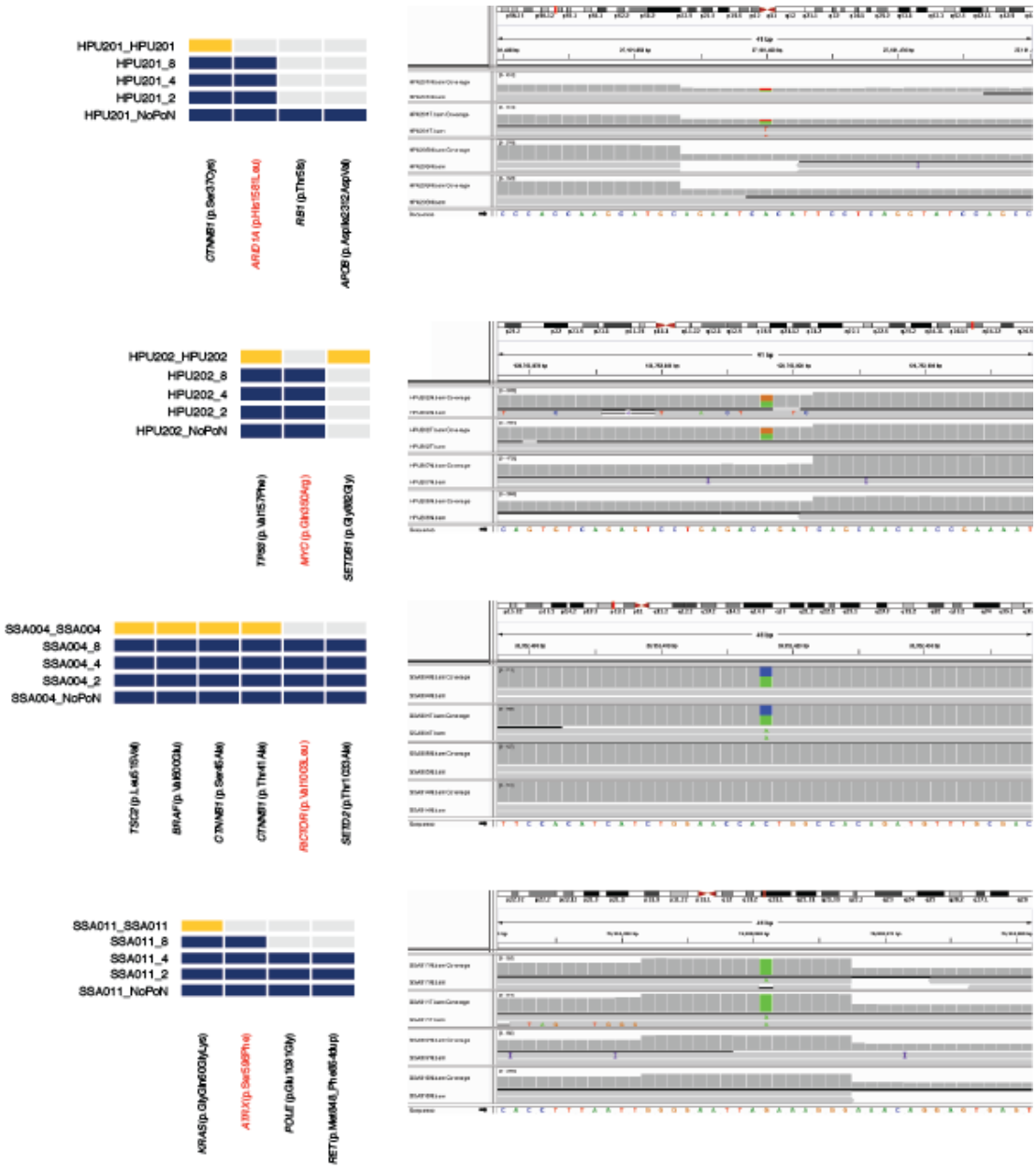
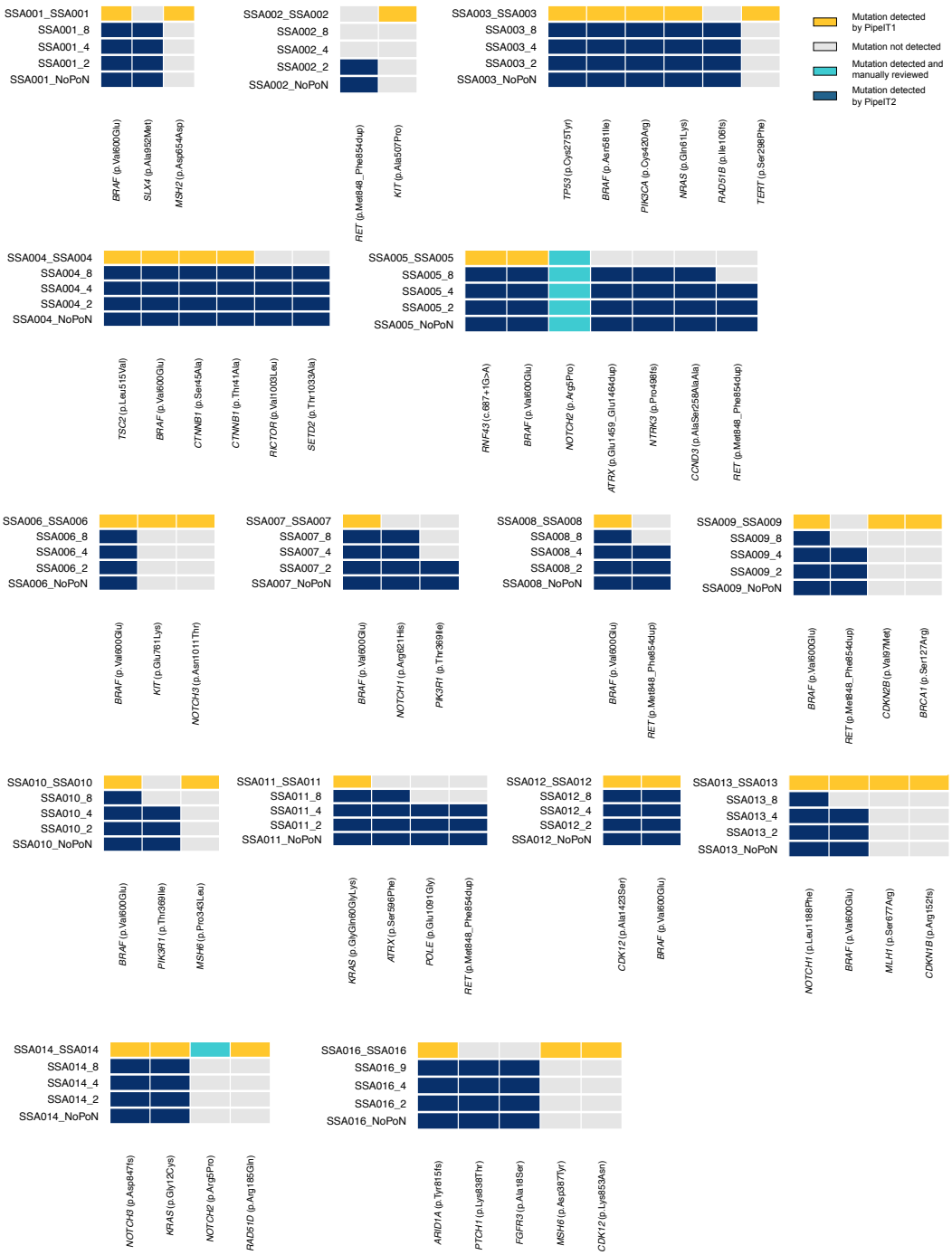
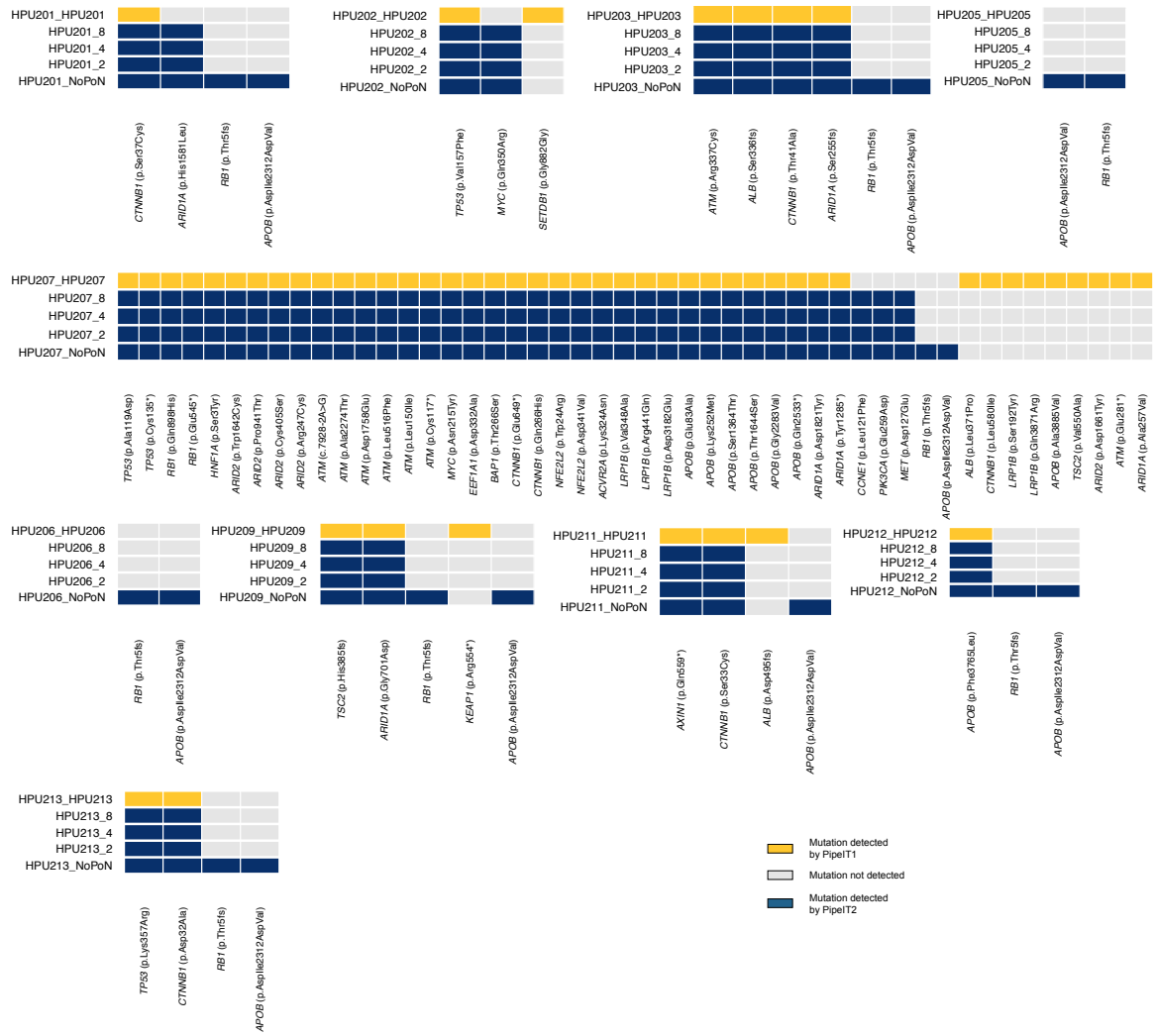


Figure 3. Germline mutations. False positive mutations were investigated by visualizing the BAM files of the original tumor-germline matched samples and unmatched germline samples used to build the PoN used to analyze the tumor samples (SSA004, SSA011, HPU201, and HPU202) using the Integrative Genomic Viewer. The heatmaps on the left show the presence of mutations identified by PipeIT2 (blue cells), using different sized PoN files, and PipeIT (yellow cells). The names of the investigated false positive mutations are highlighted in red. The corresponding IGV screens are shown on the right. Colored bars show the presence of mutated alleles, grey bars show the lack of genomic alterations.

SUPPLEMENTARY FIGURE AND LEGENDS



Supplementary Figure S1. Mutations in the colon adenocarcinoma cohort. Mutations identified by PipeIT2 (dark blue cells), using different sized PoN files, and PipeIT (yellow cells) across the colon adenocarcinoma samples. Cyan cells identify mutations manually reviewed and removed.



Supplementary Figure S2. Mutations in the HCC cohort. Mutations identified by PipeIT2 (dark blue cells), using different sized PoN files, and PipeIT (yellow cells) across the different HCC samples.

3.2- Chapter II

A machine learning approach to extract oncogenic transcriptional profiles and to expand precision oncology

ABSTRACT

The accurate identification of somatic mutations has become a pivotal component of tumor profiling and precision medicine. Yet, there is still a fraction of cancer patients without any known genomic biomarker, demonstrating the need for alternative biomarkers to help clinical decision making. We reasoned that driver mutations activating specific downstream signaling pathways are manifested as transcriptional signatures that can be leveraged to predict the potential pathogenicity and actionability of rare mutations. Therefore, we developed logistic regression classifiers to learn the transcriptomic profiles associated with hotspot driver mutations in 16 oncogenes using data obtained from The Cancer Genome Atlas (TCGA) and used the classifiers to infer pathway activation status in cancers without such hotspot driver mutations. In particular, our approach incorporated the Synthetic Minority Over-sampling Technique (SMOTE) to overcome the imbalance of the input classes, as a result of the general rarity of samples with hotspot driver mutations. Our approach was first tested on the *PIK3CA* oncogene and its E542, E545 and H1047 driver hotspot mutations leading to a mean area under receiver operator curve (ROC) score of 0.87 on a validation dataset. The same approach was then further applied to an additional 15 oncogenes, demonstrating a correlation between the sensitivity of the models and the fraction of samples with hotspot mutations in the training dataset. Finally, using the model on samples with *PIK3CA* non-hotspot mutations and *PIK3CA* known interactors, leading to the identification of the transcriptomic profile associated to the *PIK3CA* hotspot driver mutations in samples with other known oncogenic mutations. Results obtained show that transcriptomic data can be used to infer the presence of oncogenic transcriptional profiles in patients and can potentially be leveraged to help expand precision medicine.

INTRODUCTION

Modern precision medicine is based on the concept that clinical decisions must take into account the variability between individuals in order to optimize treatment.[2,99] Due to the heterogeneity between cancers, even within a given cancer type, precision medicine has the potential to dramatically improve oncology care. Cancer is known to be a disease of the genome[100], where intrinsically (e.g. the aging process) or extrinsically (e.g. environmental carcinogens) induced genomic alterations are known to play an important role in tumorigenesis and cancer progression.[101] Therefore the identification of anomalies in the genome has become one of the pillars of precision oncology.

Modern Next Generation Sequencing (NGS) technologies have enabled large-scale sequencing projects such as those by The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC), that have characterized tens of thousands of cancer genomes to the base-pair resolution.[102,103] The resulting efforts have uncovered the extensive inter-tumor genetic heterogeneity and have helped to identify hundreds of thousands of novel somatic mutations and other genetic alterations. However, detection of the genetic alterations is only the first step; the interpretation of their pathogenicity is not always obvious. Nowadays it is understood that even established oncogenes can contain both mutations that are likely to play a role in carcinogenesis (also known as ‘driver’ mutations) and mutations that are likely to be bystanders (‘passenger’ mutations).[74,104] Whilst the phenotypic manifestations of the major highly recurrent (or ‘hotspot’) somatic mutations in oncogenes and their effects on the major signaling pathways are reasonably well characterized, many of the non-hotspot mutations, even in cancer genes, are of unknown clinical and biological significance. Nonsense or frameshift mutations can generally be expected to significantly alter or abrogate protein function, but it is not always clear if missense mutations outside of hotspot residues in oncogenes are activating and/or can be targeted with the same agents used in tumors with well characterized, constitutively activating

mutations.[105] Understanding the role of these mutations across the different cancer types is one of the main focuses of modern cancer research, as it is crucial to understand these differences to effectively target such alterations.

From the clinical perspective, the development of small molecule inhibitors has enabled the targeting of specific genetic alterations or signaling pathways (i.e. 'actionability') and forms the basis for the so-called genomics-guided oncology care, in which genomic characterization of the tumors is used to guide treatment decisions. Actionable genomic alterations are used to predict effectiveness to the molecule inhibitors, whose development goes through several preclinical and clinical phases and are potentially approved by agencies such as the US Food and Drug Administration (FDA) or the European Medicines Agency (EMA) for specific indications (i.e. for specific cancer types).[106] The use of these drugs in this manner is generally referred to as "on-label" to demonstrate this recognition of effectiveness on specific cancer types and in the presence of specific biomarkers. Conversely, the use of these drugs outside of the specific indications is generally defined as "off-label" use. "Off-label" therapies are not automatically a sign of ineffectiveness or poor clinical care, nonetheless.[107,108] There is evidence of effectiveness of drugs administered to patients with a different cancer type than the original intended indications, but potential new uses have not been thoroughly investigated. For example, there are proofs of good responses from patients with thyroid cancer to dabrafenib (BRAF inhibitor) and trametinib (MEK inhibitor) drugs even before their official approval from the FDA, occurred in 2018 (clinical trial ID: NCT02034110).[109–111]

Many cancers do not harbor any actionable genomic biomarker, limiting the benefits of precision oncology.[103] While the systematic functional characterization of every possible mutation is costly and labor-intensive, other omic data sources are being investigated as alternative biomarkers. Transciptomics is one of them. The use of gene expression data is based on the hypothesis that driver mutations activating a specific downstream signaling pathway are manifested as transcriptional signatures that can be leveraged to infer pathway

activation and hence actionability.[112,113] From a precision oncology perspective, this hints at the possibility that response may be observed in patients whose disease does not harbor the specific genomic biomarkers. WINTHER, a clinical trial completed in February 2019, tested this hypothesis on real cancer patients.[60] In WINTHER, the patients were split in 2 arms: arm A (DNA-guided), the subset of patients with known actionable driver mutations, identified through NGS, and arm B (RNA-guided), the subset of patients without recognized actionable genomic biomarkers, for whom microarray and gene expression values were obtained. While arm A patients were treated according to FDA-approved targeted therapies, a computational methodology was used to perform therapy response prediction for arm B. This method was based on the profiling and scoring of the differential expression between tumoral and normal biopsies of a literature-derived list of genes. Arm B patients were then treated with drugs based on these profiles and scores. The rate of stable disease ≥6 months and partial or complete response observed across the 107 patients enrolled in the study was 26.2% (arm A: 23.2%; arm B: 31.6% ($P = 0.37$)). The comparable response of arm A and B patients demonstrated that transcriptomic profiles can be used in adjunction to classic genomic profiles to tailor precision oncology treatments, particularly for patients without known actionable genetic alterations.[60]

To expand the pool of actionable patients in precision oncology, we inferred the presence of molecular oncogenic processes in patients from their transcriptional data. To do so, we developed a machine learning approach, based on an elastic net regularized logistic regression classifier, able to detect gene expression profiles associated with the activation of specific oncogenic pathways by integrating data from the TCGA dataset. The approach was first tested on *PIK3CA*, then expanded to 15 other oncogenes to evaluate whether the classifier was able to retrieve the transcriptomic profiles associated with the hotspot cancer

driver mutations in these genes. Finally, the classifier was used on an external subset of cancer patients, devoid of hotspot mutations, to infer the presence of the same oncogenic pathways activation in patients lacking genomic biomarkers.

MATERIALS AND METHODS

Downloading of the TCGA data

Open access molecular data were obtained from the TCGA database.[102] (**Figure 1 A**) Gene expression and mutation data were collected from Genome Data Commons (GDC) https://gdc.cancer.gov/about-data/publications/pancanatlas.[114] Cancer type and copy number alteration data were collected from the UCSC Xena https://xenabrowser.net/datapages/?cohort=TCGA%20Pan-Cancer%20(PANCAN).[62] The gene expression data comprise 8184 samples across 33 cancer types and gene expression values (MapSplice + RSEM, then normalised by setting the upper-quartile to 1,000) from 20,502 genes. The mutation data comprise 3.6 million mutations across 10,295 samples. The copy number alteration data comprise copy number values (estimated using the GISTIC2 threshold method[115]) for 24,776 genes across 10845 samples. Information regarding the cancer and the tissue types was available for 12804 samples. The 8,184 samples with the complete set of mutation, gene expression, copy number alteration and cancer type information were used for further analysis.

Development of a logistic regression classifier

The aim of the classifier is to identify the downstream effects of driver mutations in a given gene on the transcriptomic level that may be associated with the specific oncogenic pathway activation. The classifier is based on a logistic regression model, with an elastic net regularization.[116] To develop the classifier based on the presence of genetic alterations (i.e.

non-synonymous mutations and copy number alterations amplifications or deletions) in the gene and its neighboring genes, the TCGA samples were assigned to one of four mutually exclusive classes, in the following order (**Figure 1 A**): 1) 'interactor-mutant' samples: samples with any genetic alterations in any of the interactors of the selected oncogene (see section 'defining the gene interactors' below); 2) 'driver hotspot' samples: samples with selected, known hotspot mutations in the selected oncogene; 3) 'wild-type' samples: samples without non-synonymous mutations in the selected oncogene; 4) 'non-hotspot' samples: samples with non hotspot mutations in the selected oncogene.

The gene expression data of the samples in the 'driver hotspot' and the 'wild-type' classes were used as input data for the training of the model. These samples were first merged into a single gene expression matrix then randomly split in a 2:1 proportion in a stratified manner as the training and the testing subsets, respectively. In order to selective informative features and to avoid a potentially overfitted model, the most differentially expressed genes, based on the p-value obtained from t-tests performed on the gene expression data between the 'driver hotspot' and the 'wild-type' samples,[117] were selected for model training. (**Figure 1 A**) The expression data for the selected genes were z-score transformed prior to model training. Cancer type information for each sample was then added to the gene expression matrix as additional features (as a boolean matrix where each cancer type is a feature). This matrix was then used as input data for the training of the machine learning model, based on a logistic regression model with an elastic net regularization.[116] (**Figure 1 B**) The Logistic Regression model is trained to identify the transcriptomic profile that differentiates the 'driver hotspot' and the 'wild-type' classes by assigning a weight score that minimizes the loss function:

$$\min_{w} \frac{1}{2n_{\text{samples}}} ||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

The *alpha* (α) and l_1 *ratio* (ρ) values are hyperparameters that can be adjusted to optimise the accuracy of the model. In order to do so, a 5 fold cross validation was performed on the training data. Specifically, the training data was randomly split into 5 evenly split subsets of samples.

A model was then trained using only 4 of these subsets (i.e. 80% of the whole training dataset) and validated on the remaining subset to evaluate the performance of the classifier defined as the Area Under the Receiver Operating Characteristic Curve (ROC AUC) score. This step was iterated 5 times, so that every subset was used as the validation data exactly once. Moreover, the tuning of the hyperparameters was iterated using the Pipeline Sklearn function with different combinations of *alpha* and *l1 ratio* values. The combination of hyperparameters that led to the best 5-fold cross validation scores was then used to train the final classifier.

The machine learning classifier was implemented using Python3 and the Pandas[118], SciPy[119] and scikit-learn (Sklearn)[120] Python libraries.

Defining the gene interactors

For each index gene, the lists of interactors were retrieved from the StringDB website. The 10 interactors with the higher confidence score (derived by benchmarking the performance of the predictions against a common reference set of trusted, true associations) with experimental evidence retrieved were selected to define the potentially confounding genes.[121]

Correcting for unbalanced classes

Ideally, a multi-class classifier wants evenly distributed input data. Uneven proportions can potentially result in a model that completely ignores the less represented class or an overfitted model, depending on the actual proportions. In the context of the TCGA data, for the vast majority of genes, the ‘wild-type’ class would significantly outnumber the ‘driver hotspot’ class, leading to an unbalanced training dataset. To address this potential issue, we implemented the Synthetic Minority Over-sampling Technique (SMOTE)[122] before model training. In brief, SMOTE addresses the unbalanced classes issue as follows: a random sample in an underrepresented class is chosen and its five nearest neighbours within the same class, (i.e. the five samples with the most similar gene expression profiles), are picked. Artificial samples

are then introduced by averaging the gene expression values of the index sample and each of its five neighbours, resulting in five new artificial samples. This process is iterated until an even ratio between all the classes is reached. (**Figure 1 C**)

Evaluation of classifier performance

Classifier performance was evaluated using the following statistical measurements both on training and testing datasets: sensitivity (the ability to identify the samples with the mutations of interest), specificity (the ability to identify the samples without the mutations of interest), accuracy (the ability to correctly classify the samples based on their status of the mutations of interest), and the ROC AUC score (the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance). (**Figure 1 D**)

Prediction on the Discovery Dataset

A combination of the samples from the 'non-hotspot' and the 'interactor-mutant' classes were used as discovery datasets for the model trained with the pan-cancer and SMOTE adjusted data to predict the presence of the transcriptomic profile associated to the selected 'driver hotspot' mutations. (**Figure 1 E**)

RESULTS

Development of a logistic regression classifier

We developed an *in silico* approach using a logistic regression classifier with an elastic net regularization to predict the presence of specific oncogenic pathway activation by extracting transcriptomic profiles from patients with driver hotspot mutations and detect the same profiles in patients devoid of the same genomic alterations. This approach was based on the assumption that the activation of oncogenic pathways is reflected on a transcriptomic level. Therefore, the identification of a specific gene expression profile can be leveraged to drive

treatments in patients without clear genomic biomarkers and, ultimately, expand modern precision oncology.

A summary of the workflow is the following: for a given gene, we trained a logistic regression model using samples with known activating mutations in the gene and wild-type samples (i.e. without any genetic alteration in the gene). To avoid potential confounding effects from the inclusion of samples with genetic alterations in the interactors of our gene of interest, we excluded the samples with either non-synonymous mutations or copy number variations in any of the interactor genes. To select informative features and to avoid overfitting the model, the 1000 most differentially expressed genes between the two classes have been retained through a t-test correlation.[123,124] Finally, we use the trained model to predict pathway activation in all remaining samples that were not in the training data. (**Figure 1**) The workflow will be explained in greater detail with the illustrative example of *PIK3CA*.

Training the *PIK3CA* classifier

We first tested our approach by training and testing a classifier for *PIK3CA*. *PIK3CA* is the gene that encodes the catalytic subunit (p110α) of the Phosphatidylinositol-3-kinase (PI3K), a receptor tyrosine kinase. The stimulation of p110α causes a signaling cascade which starts with the conversion of the lipid substrate PIP2 (phosphatidylinositol-4,5-bisphosphate) into PIP3 (phosphatidylinositol-3,4,5-bisphosphate). One of the pathways activated by this process is the PI3K/mTOR/AKT pathway, which is involved in cell growth, cell cycle, survival, proliferation, and motility.[34] The helical domain of *PIK3CA*, translated from exon 9, harbors two hotspot sites, E542K and E545K (**Figure 2 A**). Another well known hotspot site is the H1047 (**Figure 2 A**), found within exon 20, in the kinase domain. This mutation often causes the substitution of the Histidine with an Arginine or, less frequently but still more commonly than most of the other *PIK3CA* mutations, in a Leucine. *PIK3CA* is one of the most commonly mutated genes in a broad range of cancer types, in particular in endometrial carcinoma

(UCEC, mutated in 51.1% of the samples in the TCGA cohort), uterine carcinosarcoma (UCS, 35.1%), breast invasive carcinoma (BRCA, 34.4%), colon adenocarcinoma (COAD, 31.0%), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, 29.4%) and others. (**Figure 2 B**) Other than being one of the cancer types with the highest frequencies of *PIK3CA*-mutated samples, BRCA is the only disease with FDA-approved drugs in the OncoKB database.[125–127]

We used the TCGA cohort to train a classifier to detect the activation of the oncogenic pathway caused by the *PIK3CA* hotspot mutations E542K, E545K and H1047R/L (**Table 1**). The cohort consisted of 9234 samples with complete gene expression, mutation, copy number alteration and cancer type data from 33 cancer types. Here we used the samples with any of the *PIK3CA* driver hotspot mutations (i.e. 'driver hotspot') as the positive set and the *PIK3CA*-wild-type samples ('wild-type') as the negative set. To minimize the presence of mutations that may result in gene expression alterations potentially similar to the ones resulting from *PIK3CA* driver mutations, which can potentially confound the training of the machine learning model, we excluded samples with any genetic alteration (i.e. non-synonymous mutations, copy number amplifications or deletions) in the interactors of *PIK3CA* ('interactor-mutant'). Here we considered the interactors as the top 10 protein-coding genes with experimentally proven interactions with *PIK3CA* according to the String database: *PIK3R1*, *PIK3R3*, *HRAS*, *NRAS*, *MRAS*, *KRAS*, *PIK3R2*, *ERAS*, *RRAS2* and *RRAS* (**Table 2**). After the training of the classifier model, we used the model to infer pathway activation status for the samples in the 'neighbor-mutant' class, as well as the samples with 'non-hotspot' *PIK3CA* mutations (variants of unknown significance). Across the TCGA cohort, the distribution of the samples across the 4 classes is the following: 426 samples (5.2%) had 'driver hotspot' mutations, 5343 samples (65.3%) were 'wild-type' for *PIK3CA*, 290 samples (3.5%) had 'non-hotspot' *PIK3CA* mutations and 2125 samples (26%) were 'interactor-mutant'. (**Figure 2 C**)

We investigated the distributions of the three hotspots across the 33 different cancer types defined in the TCGA database. The 'driver hotspot' class accounted for at least 10% of the samples in 3 cancer types: uterine carcinosarcoma (UCS, ~12%), breast invasive carcinoma (BRCA, ~21%), and cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC, ~15%). In particular, the E545K and E542K mutations are more frequent in the CESC samples (respectively, ~9% and ~5.2%), while the H1047R mutation is found more frequently in BRCA samples (~11%). The H1047L mutation is rarer compared to the other hotspot mutations and has been found more frequently in cholangiocarcinoma (CHOL) samples (~2.8%) compared to any other cancer types. (**Figure 2 D**) Of note, the distribution of *PIK3CA* hotspot mutations is not the same across the cancer types. In particular, we note that *PIK3CA* mutations among UCEC samples are more diverse with less frequent E542K, E545K and H1047R/L mutations (5.9%) compared to BRCA (~21%).

Samples in the 'driver hotspot' and 'wild-type' classes were randomly divided in the 'training' and the 'testing' subsets with a 2:1 ratio. The training data were used to train a logistic regression classifier, regularised with an elastic net normalization, to extract the transcriptional profile observed in cancers with at least one of the E545K, E542K, and H1047R/L hotspot mutations and to produce gene-specific weight scores used to predict oncogenic pathway activation in cancers from their transcriptomic profiles.

Two comparative analyses were performed to assert 1) whether the significantly different proportions of samples in the 'driver hotspot' and in the 'wild-type' classes would hinder the performances of the classifier and, if that is the case, how to address this issue and 2) whether a model trained on a pan-cancer dataset could outperform a model built on BRCA data alone.

Using SMOTE to overcome class imbalance

In an ideal scenario, a machine learning classifier would need input training elements equally distributed among the various classes. This was not the case for the *PIK3CA* training data. Here the two classes used as input for the training of the classifier were the samples that carried the known activating mutations in *PIK3CA* (i.e. 'driver-hotspot') and the samples that did not have genetic alterations at all in *PIK3CA* (i.e. 'wild-type'). Given the overall rarity of samples with oncogenic mutations, the number of 'wild-type' samples vastly outnumbered the number of 'driver-hotspot' samples. This could be a problem in the training of the classifier since unbalanced proportions can potentially result in a model that completely ignores the underrepresented class or an overfitted model. A classic approach used to address this problem is the resampling of the input dataset. This means either the downsampling of the overrepresented classes, by removing samples until the sizes of the classes are equal, or the oversampling of the underrepresented classes, by cloning the samples, for example. Both these approaches come with noticeable shortcomings. The downsampling approach removes information, which might impact the performance of the model, while oversampling by cloning can lead to overfitting, due to the redundant information provided. Here we tested a more sophisticated oversampling approach, the Synthetic Minority Over-sampling Technique (SMOTE),[122] to overcome the problems associated in unbalanced classes. This is a brief summary of the SMOTE approach: a sample in the underrepresented class is randomly chosen and then its five nearest neighbours (i.e. the five samples with the most similar gene expression values in the same class) are identified. Five artificial samples are then added to the training cohort by averaging the gene expression values of the index sample and each of its five neighbours. This process is iterated until equal distribution between the classes is reached.

We compared the performance of the *PIK3CA* classifier with and without the SMOTE adjustment to determine whether the use of SMOTE would affect the performance. For each of the two approaches, we first performed 5-fold cross validation, followed by a further

validation on the testing subset. Performance was quantified as sensitivity, specificity, accuracy and ROC scores. Due to the random nature of some of the steps (for example, the splitting of the training and the testing subsets), both analyses were iterated 100 times and the performance measures averaged.

When we evaluated the classifier trained on the unbalanced training data, we observed a mean sensitivity of 0.94 (range: 0.83 - 0.99), a mean specificity of 0.97 (0.93 - 0.98) and a mean ROC score of 0.99 (0.97 - 0.99), from the training subset, and a mean sensitivity of 0.52 (0.38 - 0.6), a mean specificity of 0.9 (0.87 - 0.94) and a mean ROC score of 0.81 (0.77 - 0.83), from the testing subset. Conversely, when the SMOTE adjustment was performed, we observed a mean sensitivity of 0.9 (range: 0.88 - 0.91), a specificity of 0.88 (0.87 - 0.9) and a ROC score of 0.96 (0.95 - 0.96), from the training subset, and a mean sensitivity of 0.88 (range: 0.84 - 0.91), a specificity of 0.7 (0.65 - 0.75) and a mean ROC score of 0.87 (0.86 - 0.87), from the testing subset. The substantial difference observed in the training and testing sets of scores obtained from the use of unbalanced training data hints at an overfitted model, whereas the sets of scores obtained from the model trained on the SMOTE-adjusted data were much closer. These results demonstrate the benefits from the use of SMOTE.

The benefits of pan-cancer data over cancer specific data

The second question we tried to answer was whether the classifier would work better with a pan-cancer or with a cancer type-specific training dataset. For *PIK3CA*, the E542K, E545K and H1047L/R driver mutations are frequently found in some cancer types (e.g. breast cancer) but not others. In particular, the hotspot mutations were not found in any sample in 13 cancer types. This raised the question whether the inclusion of data from diseases in which none of the samples had any of the *PIK3CA* hotspot mutations could interfere with the training of the classifier. Based on this, we tested and compared the performance of the *PIK3CA* models trained using either the whole TCGA pan-cancer dataset or only the subset of BRCA samples.

Given that the pre-processing of the training data with SMOTE improved the performance, here we also used SMOTE to correct class imbalance in both analyses.

The model built on BRCA specific training data led to a mean sensitivity of 0.83 (range: 0.56 - 1), a specificity of 0.89 (0.7 - 1) and a ROC score of 0.91 (0.77 - 1), from the training subset, and a mean sensitivity of 0.67 (range: 0.35 - 0.77), a specificity of 0.59 (0.54 - 0.79) and a mean ROC score of 0.68 (0.63 - 0.7), from the testing subset. When the pan-cancer data was used, we observed a mean sensitivity of 0.9 (range: 0.88 - 0.91), a specificity of 0.88 (0.87 - 0.9) and a ROC score of 0.96 (0.95 - 0.96), from the training subset, and a mean sensitivity of 0.88 (range: 0.84 - 0.91), a specificity of 0.7 (0.65 - 0.75) and a mean ROC score of 0.87 (0.86 - 0.87), from the testing subset, demonstrating that the larger pan-cancer cohort was a better training dataset than the BRCA-specific training dataset. The reduction in performance caused by the smaller number of samples used to train the BRCA-specific model (218 compared to the 426 used in the pan-cancer model) is comparable to the reduction observed in the classifier trained on unbalanced dataset (**Figure 2 E**) (**Table 5**) .

Evaluating the *PIK3CA* classifier

Lastly, we wanted to investigate more in deep samples classified by the model built to recognize *PIK3CA* driver hotspot mutations using the pan-cancer, SMOTE adjusted dataset. We first looked at the distributions of ‘driver hotspot’ and ‘wild-type’ inferred classifications observed using the whole samples assigned to these 2 classes. Out of 426 ‘driver hotspot’ samples, a mean of 374 (87.7%, range: 357 - 388) were properly classified, while a mean of 52 (12.2%, range: 38 - 69) were misclassified as ‘wild-type’. In the ‘wild-type’ class, a mean of 1603 (30%, range: 1336 - 1764) out of 5343 samples were misclassified as ‘driver hotspot’, but a mean of 3740 (70%, range: 3579 - 4007) samples were properly classified. Due to the role of E542K, E545K and H1047R/L mutations in the activation of the PI3K/mTOR/AKT pathway, we retrieved the list of mutations in the genes involved in this pathway from the

misclassified ‘wild type’ samples. Said mutations can potentially affect the pathway in a similar manner as the *PIK3CA* hotspot mutations and result in a comparable gene expression profile: *AKT1*, *AKT2*, *AKT3*, *CRKL*, *IRS2*, *MTOR*, *PIK3CG*, *PIK3R1*, *PIK3R2*, *PTEN*, *RICTOR*, *RNF43*, *RPTOR*, *TSC1* and *TSC2*.[128] These genes have been found mutated in a mean of 16.5% of the misclassified ‘wild type’ samples, *PTEN* being the most frequently mutated (*AKT1* (1%), *AKT2* (0.3%), *AKT3* (0.5%), *CRKL* (0.2%), *IRS2* (0.3%), *MTOR* (2.1%), *PIK3CG* (1.8%), *PTEN* (4.5%), *RICTOR* (1.1%), *RNF43* (0.9%), *RPTOR* (1%), *TSC1* (1.2%) and *TSC2* (1.4%)). *PIK3R1* and *PIK3R2* were previously selected as known *PIK3CA* iteractors and samples with alterations in these two genes were previously removed from this class of samples.

Development and evaluation of classifiers for 15 additional oncogenes

Next we proceeded to test the approach on 15 additional oncogenes to determine whether the same approach could be used to train classifiers to distinguish samples with and without known oncogenic mutations based on gene expression. In this analysis, we included 15 oncogenes obtained from the Oncokb database[129] and comprises the following genes: *SF3B1*, *BRAF*, *CTNNB1*, *EGFR*, *ERBB2*, *FGFR3*, *GNA11*, *GNAQ*, *HRAS*, *IDH1*, *KRAS*, *NFE2L2*, *NRAS*, *AKT1*, *MAP2K1*. We chose these genes for different reasons. First, they all are acknowledged oncogenes, so they suit the scope of this study. Next, they offer a broad range of variety in terms of mutational frequencies and cancer types in which they are known to have a role in tumorigenesis. For each of them, we trained each classifier using the 'driver hotspot' and 'wild-type' samples across all cancer types (**Table 1**), excluding the samples with alterations in the interactors (**Table 2**), pre-processed the data with SMOTE, performed feature selection, trained the classifiers and evaluated the classifiers in the same way as was performed for *PIK3CA*.

First we compared the distribution of the samples across the 4 different classes (i.e. ‘driver hotspot’, ‘wild-type’, ‘interactor-mutant’ and ‘non-hotspot’) (**Figure 3 A**). The general trend is roughly the same across all the oncogenes; the wild-type samples represent the largest

fraction, followed by the 'interactor-mutant' ones, while the samples with either hotspot or non-hotspot mutations account for significantly smaller proportions in all 15 oncogenes. Moreover, in some of them (*MAP2K1*, *HRAS* and *ERBB2*, for example) the samples with driver hotspot mutations were so scarce that, even with the use of SMOTE, we did not expect to achieve good performance or even being able to reach the minimum amount of data needed for the training of the model. In addition to this, we have decided to investigate, for each oncogene, the distribution of the samples with the selected driver hotspot mutations across all the cancer types defined in TCGA (**Figure 3 B**). This highlighted whether said genes were predominant in specific cancer type (for example *GNA11* and *GNAQ* found almost exclusively in uveal melanoma (UVM) samples) or were frequently found across different diseases (like *KRAS*, found in at least 10% of the samples in pancreatic adenocarcinoma (PAAD), colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD) and rectum adenocarcinoma (READ)).

Analogous to the comparison of using the PanCancer vs BRCA training sets for *PIK3CA*, we tested whether the models would perform better with the whole PanCancer dataset, or with more targeted datasets consisting of specific cancer types (**Table 3**). For most of the oncogenes, multiple cancer types, chosen based on both the literature and the observed mutation frequencies, have been selected. The models were then tested and compared over 20 iterations.

For 12 out of 16 oncogenes, we observed better ROC score for the models trained on the PanCancer data (**Table 4** and **Table 5**). The only exceptions were *SF3B1* (pan-cancer mean ROC= 0.97 (range: 0.96 - 0.97), cancer specific mean ROC= 0.98 (range: 0.98 - 0.98)), *IDH1* (pan-cancer mean ROC= 0.96 (range: 0.96 - 0.97), cancer specific mean ROC= 0.99 (range: 0.98 - 0.99) and *NFE2L2* (pan-cancer mean ROC= 0.95 (range: 0.94 - 0.95), cancer specific mean ROC= 0.95 (range: 0.95 - 0.95)), for which the cancer types specific models slightly outperformed compared to the pan-cancer models (**Table 5**). Nevertheless, ROC scores were

not always good metrics for the 'driver hotspot' detection performances of the models. As expected, this classification approach did not perform well for some of the oncogenes with very infrequent occurrences of driver hotspot mutations (**Figure 3 C**). For example, *MAP2K1* did not reach the minimal threshold of hotspot mutated samples (n=12) making it impossible to train a model. Moreover, ROC scores of models for rarely mutated oncogenes such as *ERBB2* (mean ROC = 0.71 (range: 0.69 - 0.73)) and *HRAS* (mean ROC = 0.57 (range: 0.5 - 0.63)) were below 80%, suggesting that either even SMOTE could not adequately account for the class imbalance or that these genes did not have strong transcriptional signatures. We then asked whether the fraction of the 'driver hotspot' class might account for the difference in the sensitivity of the models. The comparison of the different accuracy scores highlighted how while some accuracies scores were coherent for some oncogenes (i.e. *BRAF* with a mean ROC= 0.97, mean sensitivity= 0.9 and mean specificity= 0.95) some of the models with high ROC scores were biased toward the 'wild-type' class (i.e. *SF3B1* with a mean ROC= 0.96, mean sensitivity= 0.08 and mean specificity= 0.99) (**Figure 3 C**) (**Table 5**). We observed a significant correlation between the fraction of the 'driver hotspot' class and the sensitivity scores for the respective models (R= 0.71, p.value= 0.0021, Spearman correlation test) (**Figure 3 D**). *PIK3CA*, *KRAS*, *BRAF*, *IDH1*, and *NRAS* were the only oncogenes where the percentage of driver hotspot samples was higher than 1% and were the ones that achieved significantly higher sensitivity scores (*PIK3CA*= 0.88 (range: 0.84 - 0.91), *KRAS*= 0.83 (0.68 - 0.92), *BRAF*= 0.9 (0.88 - 0.9), *IDH1*= 0.92 (0.9 - 0.92), *NRAS*= 0.87 (0.8 - 0.91)) compared to the remaining oncogenes (mean= 0,31 (0 - 0.86), p=0.001, Mann-Whitney rank test). Interestingly, 3 of the oncogenes with fewer than 1% 'driver hotspot' samples, *EGFR*, *FGFR3*, and *GNA11*, reached sensitivity scores above the 70% (mean sensitivity for *EGFR*= 0.73 (0.58 - 0.86), mean sensitivity for *FGFR3*= 0.72 (0.7 - 0.73) and mean sensitivity for *GNA11*= 0.72 (0.6 - 0.85)), which may suggest that the activation of the respective oncogenic pathways leads to relatively easy-to-recognize transcriptional signatures, despite the small positive training dataset.

Inference of pathway activation status

Lastly, we have applied the same model to infer the PI3K pathway activation status on the samples from the two classes of samples not used for model training: the samples with 'non-hotspot' *PIK3CA* mutations and the 'interactor-mutant' samples. The aim of this step is to determine if the transcriptional profile identified by the model can be detected and, therefore, suggest pathway activation in these samples. Of the 2125 'interactor-mutant' samples, a mean of 284 (13.4%, range: 276 - 295) samples were classified as having the 'driver hotspot' transcriptional profile, while a mean of 1841 (86.6%, range: 1830 - 1849) were classified as having the 'wild-type' profile. Among the 290 'non-hotspot' samples, we observed a mean of 186 samples (64.1%, range: 179 - 188) classified as 'driver hotspot' and 104 (35.9%, range: 102 - 111) classified as 'wild type'. Next, we extracted all the *PIK3CA* mutations found in these two classes and annotated their oncogenic status based on the OncoKB database (**Figure 4**).[129] Of the 131 mutations in samples classified as 'driver hotspot', 106 (80.9%) were either 'Oncogenic' or 'Likely oncogenic' according to OncoKB, compared to the 16 'Oncogenic' or 'Likely oncogenic' mutations seen in samples classified as 'wild-type'. 25 (19.1%) mutations in samples classified as 'driver' hotspot' were either unannotated or were annotated as 'Inconclusive' according to OncoKB, compared to 7 (28%) mutations in samples classified as 'wild-type'.

Moreover, across the samples included in the 'interactor mutant' class, 31 *PIK3CA* hotspot mutations (6 E542K, 10 E545K, 13 H1047R and 2 H1047L) were found in patients classified as 'driver hotspot' by the model, while 6 mutations (1 E545K and 5 H1047R) were found in patients classified as 'wild-type'. These results can potentially hint that the transcriptomic profile associated with the *PIK3CA* non hotspot oncogenic mutations was recognized by the classifier.

DISCUSSION

It has been demonstrated that gene expression data can be used in adjunction to the more classic genomic sequencing to drive precision oncology therapies.[60] What is still needed is a technique able to methodologically leverage transcriptomic data to identify oncogenic molecular processes in patients. We developed a logistic regression machine learning approach able to detect and extract a transcriptomic signature from samples with cancer driver and driver hotspot mutations, to infer the pathway activation status in patients without the hotspot mutations. This is important as many cancers do not harbor known cancer drivers.[103]

We initially tested the approach on the *PIK3CA* oncogene. As a proof of concept, we tried to train a classifier able to detect the oncogenic pathway activation caused by the most common *PIK3CA* hotspot driver mutations (E545K, E542K, and H1047R/L). After implementing a computational algorithm (SMOTE) to address the class imbalance of the training data and demonstrating that the use of a pan-cancer dataset leads to a better classifier than one based on a breast cancer specific dataset, the logistic regression model was able to predict the presence or the lack of these hotspot mutations with sensitivity and specificity scores of, respectively, 88% and 70% and a ROC AUC score of 87%. Performing comparative analyses showed how leaving class imbalance not addressed and using a cancer type-specific subset as the training data (as opposed to using pan-cancer training data) effectively lead to the removal of information which may ultimately have reduced model performance.

The approach was applied to train additional classifiers for 15 other oncogenes. Including *PIK3CA*, 12 of these classifiers were able to achieve good prediction scores (ROC score higher than 85%), while 4 were not performing as well (ROC score lower than 85%). Our main hypothesis, supported by the different frequencies of actionable mutations in the TCGA cohort, was that some oncogenes did not reach the necessary numbers of actionable samples needed

to properly train the models. A correlation between these numbers and the sensitivity scores obtained from the respective classifiers showed a strong correlation (R= 0.71), suggesting that the better performing models were associated with having a higher percentage of samples with actionable mutations. This conclusion is coherent with the assumption that the power of machine learning lies in its ability to learn from large amounts of data. Underperforming models trained on the rarer oncogenes, compared to the other selected ones, can be further tested once more once additional data from cancer patients is obtained to improve their classification accuracies.

Finally, we used the classifier trained with *PIK3CA* hotspot mutations to infer the presence of the same transcriptomic profile and the relative oncogenic pathway activation in a subset of TCGA samples with either non-hotspot mutations or with alterations in genes that are known *PIK3CA* interactors. This analysis was performed to achieve two aims. First, it was done to test whether the classifier was able or not to identify established non-hotspot oncogenic mutations. Second, it was used to infer the oncogenic and actionable nature of mutations with unknown role in cancer development. The model was able to infer the presence of the same transcriptional profile associated to the *PIK3CA* hotspot driver mutations in several samples with previously studied *PIK3CA* mutations, annotated as oncogenic or likely oncogenic in the OncoKB database (106 out of 122).[129] From a precision oncology perspective, cancer drugs used to treat patients with *PIK3CA* oncogenic mutations, such as the combination of Alpelisib and Fulvestrant [125], rely on the presence of a broad range of oncogenic mutations rather than a few selection like what is seen, for example, in BRAF inhibitors, specifically used on samples with a V600E mutations in the *BRAF* gene.[130] This hints at the fact that many oncogenic *PIK3CA* mutations lead to the same downstream oncogenic effect. The ability of the model to infer the same transcriptomic profile is coherent to this assumption, together with the assumption that gene expression profiling can be used as an alternative to genetic data to drive precision oncology[60], points to a possibility to use this approach to potentially reveal hidden responders in the cancer patients. A potential follow-up study can focus on a similar

methodology to train models able to detect transcriptomic profiles associated with clinically meaningful genetic biomarkers to further investigate and validate this theory.

In conclusion, we developed a computational approach based on a logistic regression model able to predict oncogenic pathways activation using TCGA gene expression data. In the absence of clinically approved genomic biomarkers, our classifier can be used to find potentially novel responders to cancer therapies and expand modern precision oncology. One of the limits of our study is the dependence on a large training dataset, able to provide enough information to train even models specific to less frequently mutated oncogenes. As the number of cancer patient data increases, the detection of the approach will increase as well, expanding the range of oncogenes that could benefit from this methodology.

FIGURES

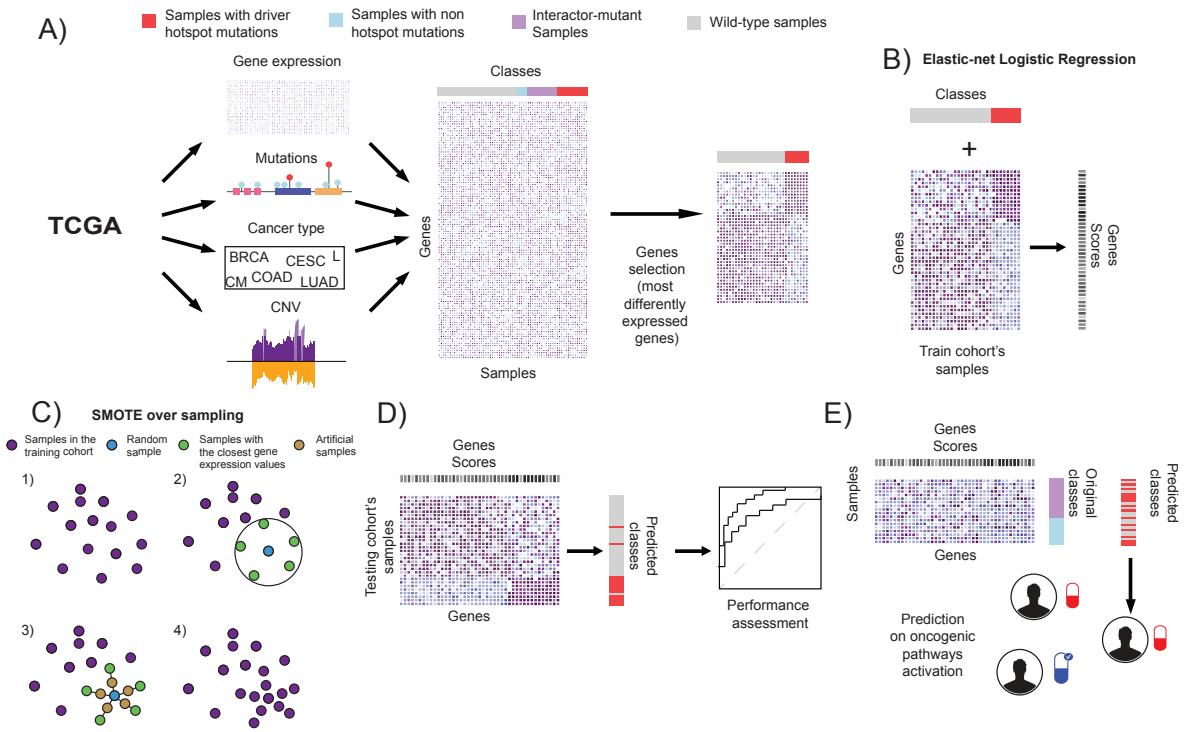


Figure 1. Development of a logistic regression classifier to infer pathway activation from oncogenic mutations. Flowchart showing the different steps of the workflow. A) Data

was retrieved from the publicly accessible TCGA database. Information obtained includes gene expression, mutation, copy number alteration and cancer type data. Samples were classified based on this information: samples with 'interactor-mutant' alterations; samples with 'driver hotspot' mutations in the selected oncogene; the 'non-hotspot' samples with non-hotspot mutations in the selected oncogene; samples 'wild-type' for the selected oncogene. The 1000 most differentially expressed genes between the 'driver hotspot' and the 'wild-type' classes were identified and used to generate a data matrix used for the training and the testing of the classifier. B) The classifier was trained using logistic regression, with an elastic net regularization, by assigning a specific weight score to each gene. C) Brief description of the SMOTE algorithm: 1) samples in the 'driver hotspot' class were collected, 2) a sample was randomly chosen and the 5 neighbours samples in the same class with the closest gene expression values were selected, 3) artificial samples were generated by averaging the gene expression values of the initial sample and each of its neighbours, 4) the artificial samples were included in the 'driver hotspot' class. D) The accuracy of the trained classifier was evaluated. E) The classifier was used on the 'non-hotspot' class to infer the presence of the oncogenic pathway activation.

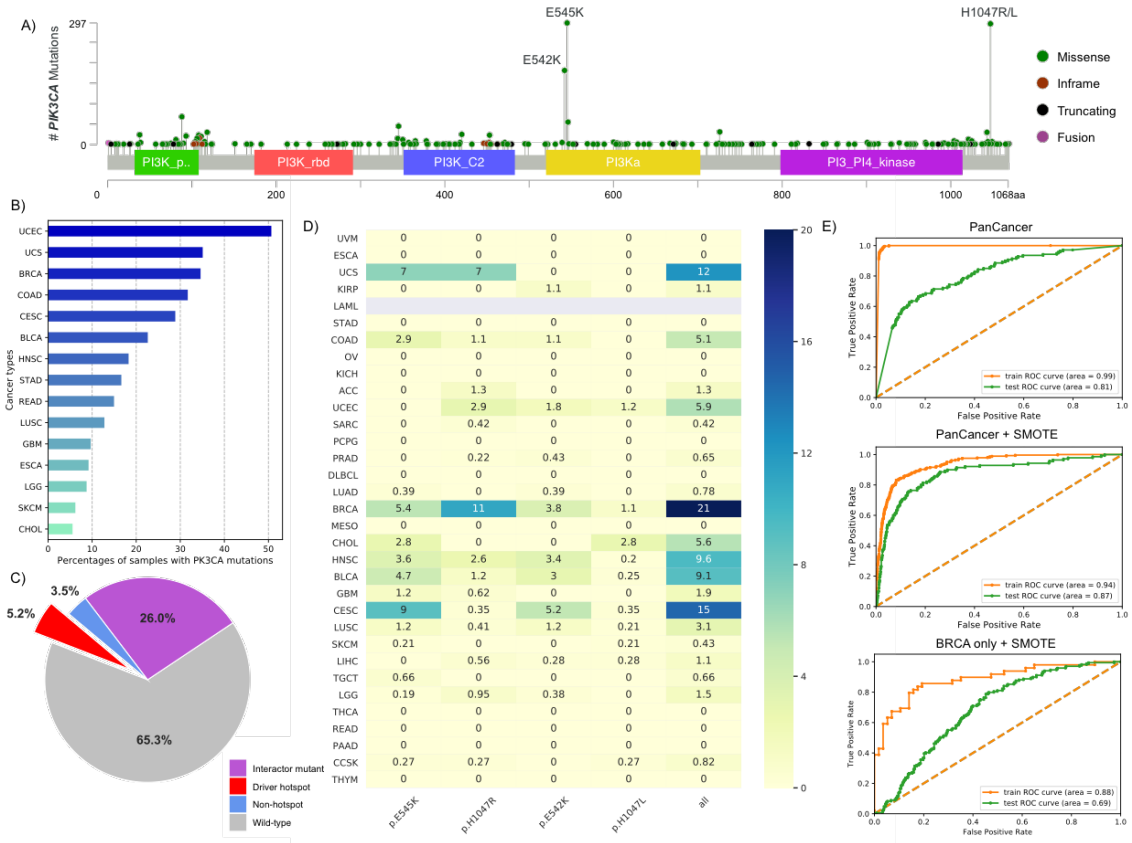


Figure 2. Training of the *PIK3CA* classifier. A) Distribution of *PIK3CA* mutations along the *PIK3CA* gene across the pan-cancer TCGA dataset. Each dot identifies a specific mutation. The height corresponds to the relative observed frequency, the color corresponds to the mutation type. B) Percentages of *PIK3CA*-mutated samples across the TCGA dataset. Only the 15 most frequently mutated cancer types are shown. C)Percentages of samples included in each of the 4 classes defined in the approach. D) Percentages of samples with *PIK3CA* 'driver hotspot' mutations across the 33 cancer types included in TCGA. The first 4 columns show the percentages for each individual mutation, the last column show the percentage for all of them. Data for Acute Myeloid Leukemia (LAML) was removed due to the lack of gene expression data for samples included in this cancer type. E) Receiver Operating Characteristic (ROC) curves obtained from 3 differentially trained *PIK3CA* classifiers. Topmost ROC was obtained from a model trained on pan-cancer data, without SMOTE samples adjustment. Middle ROC was obtained from a model trained with pan-cancer data, adjusted with SMOTE. Bottommost ROC was obtained from a model trained with Breast invasive carcinoma (BRCA) data, adjusted with SMOTE.

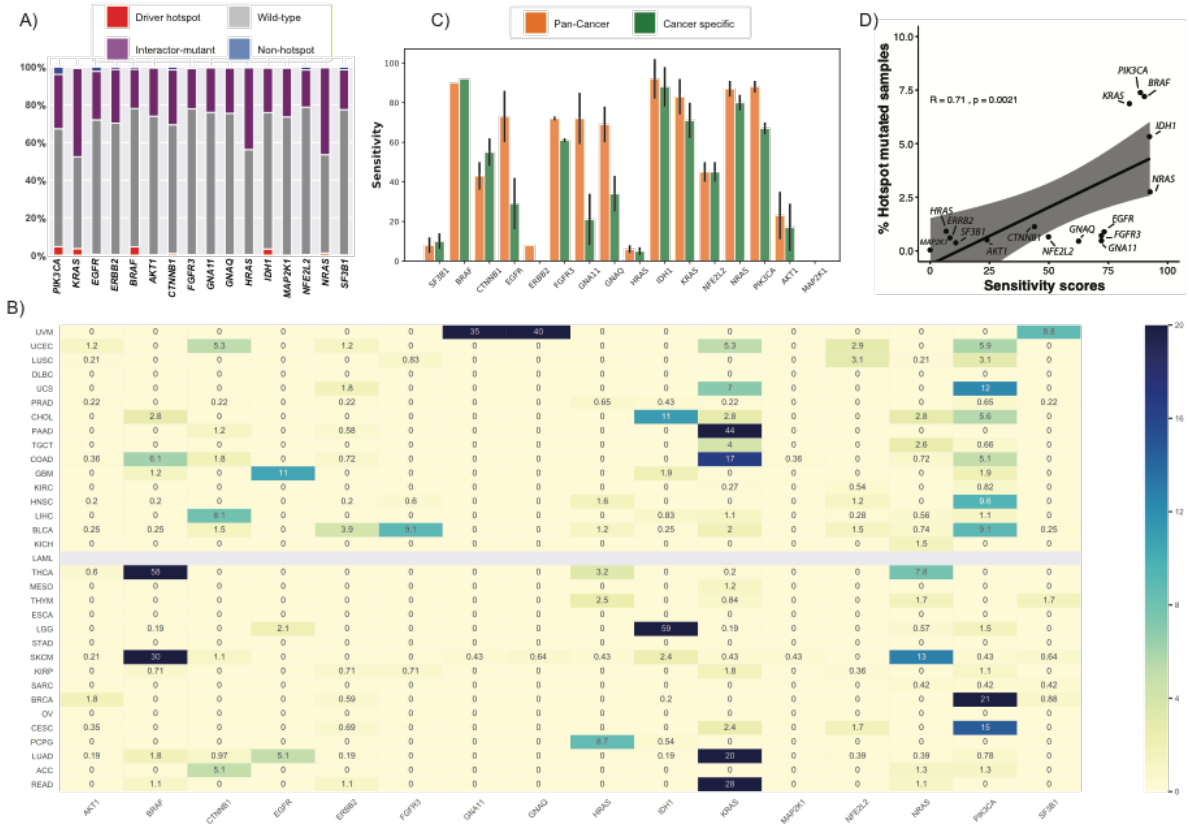


Figure 3. Classifiers trained on 16 different oncogenes. 16 classifiers were trained to identify the transcriptional signatures for 16 oncogenes: *PIK3CA*, *SF3B1*, *BRAF*, *CTNNB1*, *EGFR*, *ERBB2*, *FGFR3*, *GNA11*, *GNAQ*, *HRAS*, *IDH1*, *KRAS*, *NFE2L2*, *NRAS*, *AKT1*, *MAP2K1*. A) Distributions of 4 classes of samples: samples with 'driver hotspot' mutated samples, 'wild-type' samples for the selected oncogene, samples with 'interactor-mutant' alterations and samples with 'non-hotspot' mutations in the selected oncogene. B) Percentages of 'driver hotspot' mutated samples for each oncogene across the 33 TCGA cancer types. C) ROC scores obtained from each classifier using either pan-cancer data or cancer specific data. D) Spearman correlation between sensitivity scores obtained from each classifier (using pan-cancer training data) and the respective percentage of 'driver hotspot' mutated samples used for the training of the models.

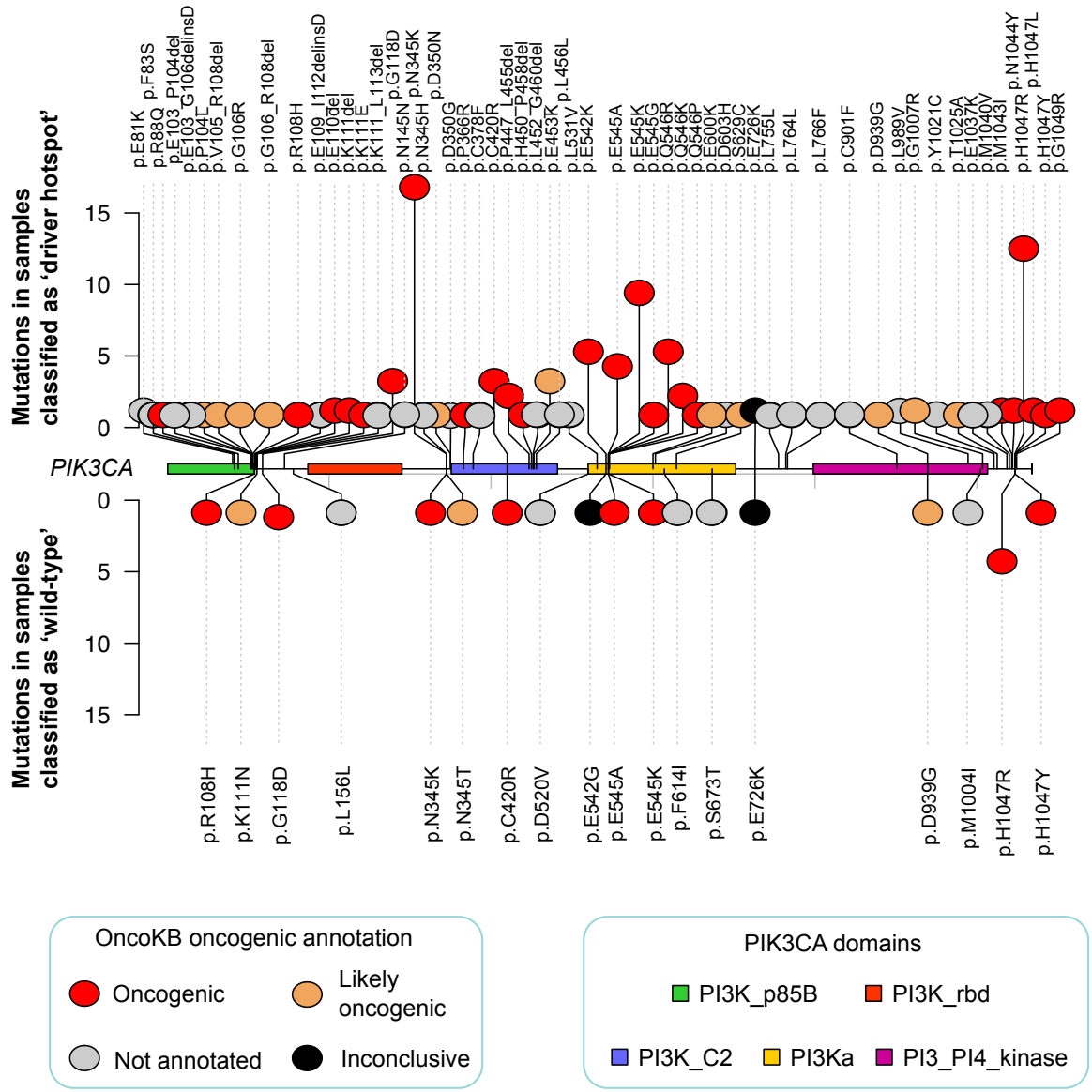


Figure 4. Oncogenicity prediction on *PIK3CA* mutations. *PIK3CA* mutations retrieved from 'non-hotspot' and 'interactor-mutant' samples classified by the pan-cancer, SMOTE adjusted *PIK3CA* model. Top mutations were observed in patients classified as 'hotspot mutated', bottom mutations were observed in samples classified as 'wild-type'. Mutations were colored based on the oncogenicity effects annotated on the OncoKB database.[129]

TABLES

Oncogene	Hotspot mutations
PIK3CA	p.E545K, p.H1047R, p.E542K, p.H1047L
SF3B1	p.K700E, p.R625H, p.R625C
BRAF	p.V600E, p.V600M
CTNNB1	p.S37F, p.T41A, p.S45P, p.S37C, p.S33C, p.S33F, p.S45F, p.G34R, p.T41I
EGFR	p.E746_A750del, p.A289V, p.L858R, p.G598V
ERBB2	p.S310F, p.V842I, p.R678Q, p.L755S, p.V777L
FGFR3	p.S249C, p.Y375C
GNA11	p.Q209L
GNAQ	p.Q209P, p.Q209L
HRAS	p.Q61R, p.G13V, p.Q61K, p.G13R
IDH1	p.R132H, p.R132C, p.R132G
KRAS	p.G12D,p.G12V,p.G12C,p.G13D,p.G12A,p.G12R,p.G12S
NFE2L2	p.R34G, p.D29H, p.E79Q, p.E82D
NRAS	p.Q61R, p.Q61K, p.Q61L
AKT1	p.E17K
MAP2K1	p.P124S

Table 1. Driver hotspots across 16 oncogenes. List of the selected hotspot mutations used to classified samples into the ‘driver hotspot’ class for each of the 16 selected oncogenes.

Oncogenes	Interactors
PIK3CA	*PIK3R1, PIK3R3, HRAS, NRAS, MRAS, KRAS, PIK3R2, ERAS, RRAS2, RRAS*
SF3B1	*SF3B5, PHF5A, SF3B3, SF3B2, SF3B4, SF3A2, CDC5L, SNRPA1, DDX42*
BRAF	*MAP2K1, MAP2K2, RPS6KB2, PRKACA, MAP3K1, PRKCA, PRKCE, RAP1A, HRAS*
CTNNB1	*CDH1, CTNNA1, FYN, FER, CTNNBIP1, APC, BCL9, BTRC, LEF1, RAPGEF2*
EGFR	*CBL, EGF, CBLB, SHC1, GRB2, PIK3R1, ERBB3, TGFA, EREG, ERRFI1*
ERBB2	*ERRFI1, GABPB1, GRB2, SHC1, PTPN11, CBL, PIK3R1, PIK3R2, PTK2, SH3BGRL*
FGFR3	*FGF8, FGF2, FGF9, FGF1, PLCG2, PLCG1, FGF16, FGF20, FGFR2, C6orf47*
GNA11	*GNB1, GNB2, GNB4, GNG2, RHOA, RHOC, RHOB, GNB3, RGS14, RGS3*
GNAQ	*GNB1, GNB2, GNB4, RHOA, GNG2, GNB3, RHOC, RHOB, RGS13, RGS4*
HRAS	*RAF1, PIK3CG, SOS1, MLLT4, NRAS, KRAS, RASA1, PIK3CD, PIK3CA, RASSF5*
IDH1	*IDH2, IDH3B, IDH3G, FAM49B, CEP57L1, IDH3A, MDH2, DAZAP1, PHB2, LTA4H*
KRAS	*ARAF, BRAF, RAF1, HRAS, NRAS, PIK3CG, PIK3CA, SOS1, EGFR, ERBB2*
NFE2L2	*NFE2, NFE2L3, CREB3, MAFG, MAFF, ENC1, KEAP1, MAFK, MAF, ATF4*
NRAS	*HRAS, KRAS, PIK3CG, PIK3CA, RAF1, ARAF, PIK3CD, BRAF, MLLT4, PIK3CB*
AKT1	*PDPK1, GSK3B, PPP2CA, VHL, EGLN1, RPS6KB1, MTOR, PDK1, AKT2, MAP3K5*
MAP2K1	*BRAF, KSR2, MAPK3, MAPK1, KSR1, MAP2K2, ARAF, RAF1, RNASE9, MAPK14*

Table 2. Interactors of the 16 oncogenes. List of the top 10 interactors retrieved from StringDB for each of the 16 selected oncogenes.

Oncogenes	Cancer types
PIK3CA	BRCA
SF3B1	UVM, THYM, BRCA
BRAF	SKCM, THCA
CTNNB1	ACC, UCEC, LIHC
EGFR	LUAD, GBM
ERBB2	BLCA
FGFR3	BLCA
GNA11	UVM, SKCM
GNAQ	UVM, SKCM
HRAS	PCPG, THCA, THYM
IDH1	LGG, CHOL
KRAS	COAD, LUAD, PAAD, READ
NFE2L2	UCEC, HNSC, BLCA, CESC, LUSC
NRAS	SKCM, THCA
AKT1	UCEC, BRCA, THCA, COAD, CESC
MAP2K1	COAD, SKCM

Table 3. Cancer types for the cancer type-specific models for the 16 oncogenes. List of TCGA cancer types selected as training datasets for the cancer type-specific models. The cancer type for which the driver hotspot mutations in each of the 16 selected oncogenes (Table 1) were the most frequently observed were selected. Multiple cancer types were selected if a single type did not provide enough samples to train the model. Breast invasive carcinoma (BRCA), Uveal Melanoma (UVM), Thymoma (THYM), Skin Cutaneous Melanoma (SKCM), Thyroid carcinoma (THCA), Adrenocortical carcinoma (ACC), Uterine Corpus Endometrial

Carcinoma (UCEC), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Glioblastoma multiforme (GBM), Bladder Urothelial Carcinoma (BLCA), Pheochromocytoma and Paraganglioma (PCPG), Brain Lower Grade Glioma (LGG), Cholangiocarcinoma (CHOL), Colon adenocarcinoma (COAD), Pancreatic adenocarcinoma (PAAD), Rectum adenocarcinoma (READ), Head and Neck squamous cell carcinoma (HNSC), Lung squamous cell carcinoma (LUSC) and Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC).

		Sensitivity	Specificity	Accuracy	ROC score
PIK3CA	*pan-cancer*	0.9 (0.88 - 0.91)	0.88 (0.87 - 0.9)	0.88 (0.87 - 0.9)	0.96 (0.95 - 0.96)
	cancer-specific	0.83 (0.56 - 1)	0.89 (0.7 - 1)	0.86 (0.71 - 1)	0.91 (0.77 - 1)
SF3B1	*pan-cancer*	1 (1 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	0.99 (0.99 - 1)
	cancer-specific	1 (1 - 1)	0.98 (0.96 - 1)	0.98 (0.96 - 1)	0.99 (0.98 - 1)
BRAF	*pan-cancer*	0.96 (0.96 - 0.96)	0.98 (0.98 - 0.98)	0.98 (0.98 - 0.98)	0.99 (0.99 - 0.99)
	cancer-specific	0.99 (0.99 - 1)	1 (1 - 1)	0.99 (0.99 - 1)	1 (1 - 1)
CTNNB1	*pan-cancer*	1 (1 - 1)	0.99 (0.98 - 0.99)	0.99 (0.98 - 0.99)	0.99 (0.99 - 0.99)
	cancer-specific	0.99 (0.9 - 1)	0.98 (0.89 - 1)	0.98 (0.89 - 1)	0.99 (0.99 - 1)
EGFR	*pan-cancer*	1 (1 - 1)	0.99 (0.98 - 0.99)	0.99 (0.98 - 0.99)	0.99 (0.99 - 0.99)
	cancer-specific	0.98 (0.85 - 1)	1 (1 - 1)	0.99 (0.97 - 1)	1 (1 - 1)
ERBB2	*pan-cancer*	1 (1 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	0.99 (0.99 - 1)
	cancer-specific	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
FGFR3	*pan-cancer*	1 (1 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	0.99 (0.99 - 1)
	cancer-specific	1 (1 - 1)	0.99 (0.99 - 1)	0.99 (0.99 - 1)	1 (1 - 1)
GNA11	*pan-cancer*	1 (1 - 1)	0.99 (0.98 - 0.99)	0.99 (0.98 - 0.99)	0.99 (0.99 - 0.99)
	cancer-	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)

	specific				
GNAQ	*pan-cancer*	1 (1 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)
	cancer-specific	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
HRAS	*pan-cancer*	0.94 (0.94 - 0.94)	0.98 (0.98 - 0.99)	0.98 (0.98 - 0.98)	0.99 (0.99 - 0.99)
	cancer-specific	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
IDH1	*pan-cancer*	0.96 (0.95 - 0.97)	0.98 (0.98 - 0.99)	0.98 (0.98 - 0.98)	0.99 (0.99 - 0.99)
	cancer-specific	0.93 (0.91 - 0.94)	0.99 (0.98 - 1)	0.95 (0.94 - 0.96)	0.98 (0.97 - 0.99)
KRAS	*pan-cancer*	0.94 (0.92 - 0.97)	0.96 (0.95 - 0.97)	0.96 (0.95 - 0.97)	0.98 (0.98 - 0.98)
	cancer-specific	0.99 (0.97 - 1)	1 (1 - 1)	0.99 (0.98 - 1)	1 (1 - 1)
NFE2L2	*pan-cancer*	1 (1 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	1 (1 - 1)
	cancer-specific	1 (1 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	1 (1 - 1)
NRAS	*pan-cancer*	0.97 (0.97 - 0.97)	0.98 (0.98 - 0.98)	0.98 (0.98 - 0.98)	0.99 (0.99 - 0.99)
	cancer-specific	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)	1 (1 - 1)
AKT1	*pan-cancer*	1 (1 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)
	cancer-specific	1 (1 - 1)	0.99 (0.99 - 1)	0.99 (0.99 - 1)	1 (1 - 1)
MAP2K1	*pan-cancer*	x	x	x	x
	cancer-specific	x	x	x	x

Table 4. Evaluation of classifier performance on the training subset. List of performance metrics (sensitivity, specificity, accuracy and ROC score) on the training subset for the classifiers trained on either the whole pan-cancer cohort or the cancer-specific cohorts.

		Sensitivity	Specificity	Accuracy	ROC score
PIK3CA	*pan-cancer*	0.88 (0.84 - 0.91)	0.7 (0.67 - 0.75)	0.72 (0.69 - 0.75)	0.87 (0.86 - 0.87)
	cancer-specific	0.67 (0.35 - 0.77)	0.59 (0.54 - 0.79)	0.61 (0.59 - 0.67)	0.68 (0.63 - 0.7)
SF3B1	*pan-cancer*	0.08 (0 - 0.12)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	0.97 (0.96 - 0.97)
	cancer-specific	0.1 (0 - 0.2)	0.99 (0.99 - 0.99)	0.98 (0.98 - 0.98)	0.98 (0.98 - 0.98)
BRAF	*pan-cancer*	0.9 (0.88 - 0.9)	0.95 (0.94 - 0.96)	0.95 (0.94 - 0.95)	0.97 (0.97 - 0.97)
	cancer-specific	0.92 (0.91 - 0.93)	0.7 (0.66 - 0.75)	0.81 (0.79 - 0.83)	0.92 (0.91 - 0.92)
CTNNB1	*pan-cancer*	0.43 (0.38 - 0.5)	0.98 (0.97 - 0.98)	0.97 (0.97 - 0.98)	0.92 (0.89 - 0.94)
	cancer-specific	0.55 (0.38 - 0.66)	0.91 (0.85 - 0.95)	0.87 (0.8 - 0.9)	0.87 (0.79 - 0.92)
EGFR	*pan-cancer*	0.73 (0.58 - 0.86)	0.96 (0.96 - 0.98)	0.96 (0.95 - 0.97)	0.96 (0.96 - 0.96)
	cancer-specific	0.29 (0.19 - 0.47)	0.9 (0.75 - 0.95)	0.84 (0.72 - 0.88)	0.68 (0.62 - 0.71)
ERBB2	*pan-cancer*	0.08 (0.08 - 0.08)	0.99 (0.99 - 0.99)	0.98 (0.98 - 0.99)	0.71 (0.69 - 0.73)
	cancer-specific	0 (0 - 0)	0.96 (0.95 - 0.97)	0.91 (0.9 - 0.91)	0.41 (0.36 - 0.46)
FGFR3	*pan-cancer*	0.72 (0.7 - 0.73)	0.99 (0.99 - 0.99)	0.99 (0.98 - 0.99)	0.97 (0.96 - 0.98)
	cancer-specific	0.61 (0.43 - 0.73)	0.93 (0.9 - 0.94)	0.87 (0.83 - 0.9)	0.87 (0.84 - 0.89)
GNA11	*pan-cancer*	0.72 (0.6 - 0.85)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)

	cancer-specific	0.21 (0.2 - 0.25)	0.99 (0.99 - 0.99)	0.94 (0.94 - 0.94)	0.7 (0.68 - 0.71)
GNAQ	*pan-cancer*	0.69 (0.65 - 0.78)	0.99 (0.99 - 1)	0.99 (0.99 - 0.99)	0.99 (0.99 - 0.99)
	cancer-specific	0.34 (0.3 - 0.39)	0.98 (0.98 - 0.98)	0.93 (0.93 - 0.94)	0.86 (0.85 - 0.86)
HRAS	*pan-cancer*	0.06 (0.05 - 0.08)	0.99 (0.99 - 0.99)	0.98 (0.98 - 0.98)	0.57 (0.5 - 0.63)
	cancer-specific	0.05 (0.05 - 0.05)	0.99 (0.99 - 1)	0.95 (0.95 - 0.95)	0.54 (0.49 - 0.57)
IDH1	*pan-cancer*	0.92 (0.9 - 0.92)	0.99 (0.99 - 0.99)	0.98 (0.98 - 0.98)	0.96 (0.96 - 0.97)
	cancer-specific	0.88 (0.86 - 0.92)	0.97 (0.95 - 0.97)	0.9 (0.88 - 0.93)	0.99 (0.98 - 0.99)
KRAS	*pan-cancer*	0.83 (0.68 - 0.92)	0.9 (0.86 - 0.94)	0.89 (0.86 - 0.92)	0.94 (0.93 - 0.94)
	cancer-specific	0.71 (0.68 - 0.75)	0.59 (0.55 - 0.62)	0.64 (0.61 - 0.65)	0.73 (0.73 - 0.74)
NFE2L2	*pan-cancer*	0.45 (0.41 - 0.5)	0.97 (0.97 - 0.98)	0.96 (0.96 - 0.96)	0.95 (0.94 - 0.95)
	cancer-specific	0.45 (0.41 - 0.5)	0.97 (0.97 - 0.98)	0.96 (0.96 - 0.96)	0.95 (0.95 - 0.95)
NRAS	*pan-cancer*	0.87 (0.8 - 0.91)	0.93 (0.92 - 0.94)	0.93 (0.92 - 0.93)	0.96 (0.96 - 0.96)
	cancer-specific	0.8 (0.74 - 0.87)	0.56 (0.51 - 0.62)	0.63 (0.61 - 0.67)	0.75 (0.71 - 0.78)
AKT1	*pan-cancer*	0.23 (0.15 - 0.35)	0.96 (0.94 - 0.98)	0.96 (0.94 - 0.97)	0.8 (0.78 - 0.81)
	cancer-specific	0.17 (0.17 - 0.17)	0.95 (0.94 - 0.96)	0.95 (0.93 - 0.95)	0.66 (0.63 - 0.68)
MAP2K1	*pan-cancer*	x	x	x	x
	cancer-specific	x	x	x	x

Table 5. Evaluation of classifier performance on the testing subset. List of performance metrics (sensitivity, specificity, accuracy and ROC score) on the testing subset for the classifiers trained on either the whole pan-cancer cohort or the cancer-specific cohorts.

4- Discussions and Outlook

In the recent past we have witnessed the advent of precision medicine, which has made it possible to move from the “standardized treatment”-based clinical care to one that is more customized to the needs of each patient. One of the main protagonists that made it possible to have this paradigm shift is the rise of “Big Data” into the healthcare setting.[131] Information brought by big data was not only exceptional in size and in variety, but it was also novel in terms of new kinds of omic information we were able to generate and process to an unprecedented level of accuracy and detail. As the amount and the complexity of data increases, the need for computational approaches able to manage and investigate this data increases. This has laid the foundations for Bioinformatics to become more and more essential both in the research scene, promoting a complementary framework to the *in vitro* and *in vivo* experiments, and in the clinical setting, where computational approaches can offer clinicians new ways to use patient data to help strategize treatments decisions. Translational bioinformatics research has enabled the study of these novel omic data to characterize molecular networks tied to diseases and to identify new treatment biomarkers.

The precision medicine philosophy easily translates to oncology. Cancer is an extremely heterogeneous spectrum of different diseases for which the common trait is the uncontrolled division of cells, resulting in the formation of tumors able to invade the original tissues.[19] Not only does cancer offer a rich landscape of diverse molecular processes whose deregulation leads to the same convergent disease phenotype, making the personalization of treatments essential, but it also presents a straightforward kind of clinically meaningful biomarkers: genomic mutations. In fact, precision oncology is, currently, driven by the use of treatments against specific mutations, called ‘driver’ mutations, able to promote carcinogenesis.

The constant evolution of DNA and RNA sequencing technologies has enabled the detection of said driver mutations in patients and the profiling of their diseases, making sequencing one of the pillars of precision oncology.[27] While whole genome sequencing and whole exome sequencing have proved to be powerful tools in the research setting, targeted sequencing panels are being adopted as a cheaper and faster alternative in clinical care, able to focus on regions with potential biomarkers tied to specific diseases. Bioinformatics has proven to be indispensable in the identification of genomic variants from sequencing data, which are simply too large for a manual approach to the identification of mutations. The Ion Torrent sequencing platform is commonly used in the routine diagnostic setting due to its relatively low costs and fast execution.[76] On the other side, it lacks streamlined data analysis tools. In chapter I of the Results we discussed how we addressed this problem with the development of PipeIT, a somatic variant caller pipeline, enclosed in a Singularity image, specifically designed to analyze Ion Torrent data. The strength of PipeIT is not limited to the accuracy of its somatic

mutation detection but also on its ease of use and its ability to ensure perfect reproducibility of the results, both of which are the direct consequences of the containerized nature of Singularity. PipeIT was published in 2019 in The Journal of Molecular Diagnostics. The main drawback of PipeIT is the need of the sequencing of a germline sample matched to the tumor counterpart. It is a drawback because, despite the important role of the germline sample to definitely identify somatic variants, control healthy samples are not always sequenced, limiting the usage of PipeIT to relatively rare occasions in the clinical setting. We addressed this limitation with the development of PipeIT2. This new pipeline was specifically designed to work on tumor-only sequencing data, thanks to a combination of more stringent filters, annotations on population data and panel of unmatched normals. Population data and panels of normals can compensate for the lack of a matched normal control by providing statistical inferences of the non-somatic nature of mutations, based on the assumption that mutations found in significant proportions in large non-disease-specific datasets are likely to be germline mutations. Both PipeIT and PipeIT2 pipelines were tested and validated on two different cohorts of cancer samples, showing the ability to properly detect important biomarkers. For example, they were able to detect *BRAF* Val600Glu mutations in colon adenoma samples. This mutation is a well-known cancer driver genomic alteration, frequently found in a diverse range of cancer types such as melanomas and thyroid, lung, and colon carcinomas (present in ~10% metastatic colorectal cancer in The Cancer Genome Atlas dataset) and is associated with poor survival. The *BRAF* Val600Glu gain-of-function mutation alters serine/threonine kinase domain and ultimately deregulates the mitogen-activated protein kinase (MAPK) pathway, promoting cell proliferation and inhibiting apoptosis.[132] Vemurafenib, dabrafenib, and encorafenib are just a few of the commercially available BRAF inhibitors specifically used to target this mutation.[133–135] Therefore, by detecting this mutation, PipeIT would have been able to drive clinical care for these patients. This example highlights the potential role of PipeIT and variant detection in precision oncology.

While the detection of genetic alterations is extremely important in oncology, there are other aspects of the disease that can be profiled to better understand the underlying molecular processes. The development of new evidence-based treatment predictors enabled by omic data is critical for precision medicine to evolve.[136] One of the main reasons is that there are still a substantial proportion of patients for whom genetic predictive biomarkers cannot be identified, suggesting it is likely that there are still genomic driver mutations yet to be discovered.[43] Gene expression data have been studied in the recent years as an additional profilable information that can either integrate the classic DNA sequencing profiling or offer an alternative to it, when driver mutations are not detected. The WINTHER clinical trial demonstrated that patients who were treated based on therapies designed on transcriptomic

data had good responses, comparable to patients who were treated according to more well established therapies based on DNA biomarkers.[60] This finding opens the door to a potential clinical revolution. In chapter II we discussed how we investigated this new landscape with the use of machine learning, which is showing promise in a diverse range of applications from image analysis in radiology and pathology to prognosis in oncology. Machine learning is one of the main branches of artificial intelligence. Its role is to dissect big and complex data that are impossible to be directly processed by humans to find underlying connections and patterns. It is easy to see how machine learning is perfectly suited to dissect the large variety of interconnected big data in healthcare to search for patterns.[73]

In Chapter II of the Results, we investigated whether it was possible to detect the presence of oncogenic molecular processes directly from gene expression data. To do so, we developed a machine learning classifier to extract the transcriptomic profiles associated with specific hotspot driver mutations and to infer the presence of the same oncogenic pathway activation in patients from transcriptomic data. *PIK3CA* hotspot mutations were initially used as a first testing oncogene to define the approach and to validate the performance of the machine learning model. Information retrieved from the open access TCGA database was used as the training and testing datasets for this project, highlighting the impact the large-scale sequencing projects have in the modern research setting. With a ROC score of 87% and a sensitivity of 88%, the model proved to be able to infer the presence of *PIK3CA* driver hotspot mutations in patients. Furthermore, by comparing the different tests performed on *PIK3CA* we observed the importance of computational methods to address class imbalance, thanks to the SMOTE algorithm, and how the training of the model using a more comprehensive pan-cancer input data led to better performance than model training on a more restrictive, cancer-specific data. Next, we applied the same methodology on 15 additional oncogenes to determine the performance of the models trained for genes that cover a diverse range of mutational frequencies and roles in carcinogenesis. Some models (in particular the ones for *BRAF*, *IDH1*, *KRAS* and *NRAS*) proved to be as good as or even better than the one trained for *PIK3CA*. On the other hand, the approach failed to achieve good performance on other less frequently mutated genes (for example, *MAP2K1*, *ERBB2* and *HRAS*), indicating the correlation of performance and the fraction of mutated samples. This finding is coherent to what is known to be one of the limits of machine learning methodologies, the need of a large training dataset. Finally, we tested the model built on *PIK3CA* on an external pool of TCGA samples to assess whether the model was also able to infer the oncogenicity of non-hotspot *PIK3CA* mutations. 80.9% of the known oncogenic and likely oncogenic, but non-hotspot, mutations were found in patients in which the model inferred the presence of the same pathway activation observed in the driver hotspot mutated samples. This could mean that these patients, along with the

samples with unannotated *PIK3CA* mutations also classified by the model in the same way, could benefit from the same therapies adopted for patients whose disease harbor the driver hotspot mutations. However, the apparently false classifications also demonstrate that although machine learning approaches are powerful tools, they are not 100% accurate. In a follow up study we could focus on drugs with specific genetic biomarkers. By studying the predictions made by these new classifiers, it may be possible to effectively expand modern precision medicine.

The studies performed in this thesis showed how computational methodologies able to process data retrieved from patients are extremely important. These methodologies are not only able to provide answers to modern questions but can, for example thanks to artificial intelligence based approaches, generate brand new hypotheses and revolutionize healthcare in the near future.

Bibliography

1. Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine. 2016. pp. 1216–1219. doi:10.1056/nejmp1606181

2. Jameson JL, Larry Jameson J, Longo DL. Precision Medicine—Personalized, Problematic, and Promising. Obstetrical & Gynecological Survey. 2015. pp. 612–614. doi:10.1097/01.ogx.0000472121.21647.38

3. Mirnezami R, Nicholson J, Darzi A. Preparing for Precision Medicine. New England Journal of Medicine. 2012. pp. 489–491. doi:10.1056/nejmp1114866

4. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Medical Genomics. 2015. doi:10.1186/s12920-015-0108-y

5. Cirillo D, Valencia A. Big data analytics for personalized medicine. Current Opinion in Biotechnology. 2019. pp. 161–167. doi:10.1016/j.copbio.2019.03.004

6. Nimmesgern E, Benediktsson I, Norstedt I. Personalized Medicine in Europe. Clinical and Translational Science. 2017. pp. 61–63. doi:10.1111/cts.12446

7. Investigators TA of URP, The All of Us Research Program Investigators. The “All of Us” Research Program. New England Journal of Medicine. 2019. pp. 668–676. doi:10.1056/nejmsr1809937

8. Zhang XD. Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships. Journal of Pharmacogenomics & Pharmacoproteomics. 2015. doi:10.4172/2153-0645.1000e144

9. Swiss Personalized Health Network (SPHN): Die nationale Initiative im Überblick. Schweizerische Ärztezeitung. 2017. pp. 595–596. doi:10.4414/saez.2017.05640

10. Website. [cited 14 Oct 2020]. doi:10.3929/ethz-b-000274911 approach to precision medicine

11. The SIB Swiss Institute of Bioinformatics’ resources: focus on curated databases. Nucleic Acids Research. 2016. pp. D27–D37. doi:10.1093/nar/gkv1310

12. Benton D. Bioinformatics — principles and potential of a new multidisciplinary tool. Trends in Biotechnology. 1996. pp. 261–272. doi:10.1016/0167-7799(96)10037-8

13. Ranganathan S. Bioinformatics Education—Perspectives and Challenges. PLoS Computational Biology. 2005. p. e52. doi:10.1371/journal.pcbi.0010052

14. Lill MA. In silico drug discovery and design. In Silico Drug Discovery and Design. 2013. pp. 2–5. doi:10.4155/ebo.13.272

15. Rostami-Hodjegan A, Tucker G. “In silico” simulations to assess the “in vivo” consequences of “in vitro” metabolic drug–drug interactions. Drug Discovery Today:

Technologies. 2004. pp. 441–448. doi:10.1016/j.ddtec.2004.10.002

16. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics. 2011. pp. 2323–2323. doi:10.1093/bioinformatics/btr408

17. Garraway LA, Verweij J, Ballman KV. Precision Oncology: An Overview. Journal of Clinical Oncology. 2013. pp. 1803–1805. doi:10.1200/jco.2013.49.4799

18. Mendelsohn J. Personalizing Oncology: Perspectives and Prospects. Journal of Clinical Oncology. 2013. pp. 1904–1911. doi:10.1200/jco.2012.45.3605

19. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013. pp. 1546–1558. doi:10.1126/science.1235122

20. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013. pp. 333–339. doi:10.1038/nature12634

21. Berchuck A, Heron KA, Carney ME, Lancaster JM, Fraser EG, Vinson VL, et al. Frequency of germline and somatic BRCA1 mutations in ovarian cancer. Clin Cancer Res. 1998;4: 2433–2437.

22. Qing T, Mohsen H, Marczyk M, Ye Y, O'Meara T, Zhao H, et al. Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. Nat Commun. 2020;11: 2438.

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

24. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366: 883–892.

25. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes Dev. 2010;24: 2343–2364.

26. Bachman KE, Park BH, Rhee I, Rajagopalan H, Herman JG, Baylin SB, et al. Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. Cancer Cell. 2003;3: 89–95.

27. Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet. 2016;24: 1515.

28. Good BM, Ainscough BJ, McMichael JF, Su AI, Griffith OL. Organizing knowledge to enable personalization of medicine in cancer. Genome Biol. 2014;15: 438.

29. Levine AJ, Puzio-Kuter AM. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. Science. 2010;330: 1340–1344.

30. Croce CM. Oncogenes and Cancer. New England Journal of Medicine. 2008. pp. 502–511. doi:10.1056/nejmra072367

31. Strano S, Dell'Orso S, Di Agostino S, Fontemaggi G, Sacchi A, Blandino G. Mutant p53: an oncogenic transcription factor. Oncogene. 2007. pp. 2212–2219. doi:10.1038/sj.onc.1210296

32. Huang C-H, Mandelker D, Gabelli SB, Amzel LM. Insights into the oncogenic effects of PIK3CA mutations from the structure of p110alpha/p85alpha. Cell Cycle. 2008;7: 1151–1156.

33. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;174: 1034–1035.

34. Bader AG, Kang S, Zhao L, Vogt PK. Oncogenic PI3K deregulates transcription and translation. Nat Rev Cancer. 2005;5: 921–929.

35. Ikenoue T, Kanai F, Hikiba Y, Obata T, Tanaka Y, Imamura J, et al. Functional analysis of PIK3CA gene mutations in human colorectal cancer. Cancer Res. 2005;65: 4562–4567.

36. Barbareschi M, Buttitta F, Felicioni L, Cotrupi S, Barassi F, Del Grammastro M, et al. Different prognostic roles of mutations in the helical and kinase domains of the PIK3CA gene in breast carcinomas. Clin Cancer Res. 2007;13: 6064–6069.

37. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6: 10001.

38. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9: 356–369.

39. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics. 2005. pp. 95–108. doi:10.1038/nrg1521

40. Beltran H, Eng K, Mosquera JM, Sigaras A, Romanel A, Rennert H, et al. Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. JAMA Oncol. 2015;1: 466–474.

41. Hamblin A, Wordsworth S, Fermont JM, Page S, Kaur K, Camps C, et al. Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: Retrospective validation and prospective audit in the UK National Health Service. PLoS Med. 2017;14: e1002230.

42. Hovelson DH, McDaniel AS, Cani AK, Johnson B, Rhodes K, Williams PD, et al. Development and Validation of a Scalable Next-Generation Sequencing System for Assessing Relevant Somatic Variants in Solid Tumors. Neoplasia. 2015. pp. 385–399. doi:10.1016/j.neo.2015.03.004

43. Consortium TIP-CA of WG, The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020. pp. 82–93. doi:10.1038/s41586-020-1969-6

44. Paradiso V, Garofoli A, Tosti N, Lanzafame M, Perrina V, Quagliata L, et al. Diagnostic Targeted Sequencing Panel for Hepatocellular Carcinoma Genomic Screening. J Mol Diagn. 2018;20: 836–848.

45. 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: 347–376.

46. Lanman RB, Mortimer SA, Zill OA, Sebisanovic D, Lopez R, Blau S, et al. Analytical and

Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA. PLoS One. 2015;10: e0140712.

47. Vinagre J, Almeida A, Pópulo H, Batista R, Lyra J, Pinto V, et al. Frequency of TERT promoter mutations in human cancers. Nature Communications. 2013. doi:10.1038/ncomms3185

48. Meric-Bernstam F, Brusco L, Shaw K, Horombe C, Kopetz S, Davies MA, et al. Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. J Clin Oncol. 2015;33: 2753–2762.

49. Singer J, Irmisch A, Ruscheweyh H-J, Singer F, Toussaint NC, Levesque MP, et al. Bioinformatics for precision oncology. Briefings in Bioinformatics. 2019. pp. 778–788. doi:10.1093/bib/bbx143

50. Le Tourneau C, Kamal M, Tsimberidou A-M, Bedard P, Pierron G, Callens C, et al. Treatment Algorithms Based on Tumor Molecular Profiling: The Essence of Precision Medicine Trials. J Natl Cancer Inst. 2016;108. doi:10.1093/jnci/djv362

51. Parker BC, Zhang W. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. Chinese Journal of Cancer. 2013. pp. 594–603. doi:10.5732/cjc.013.10178

52. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. 2018;23: 227–238.e3.

53. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM. Recurrent gene fusions in prostate cancer. Nature Reviews Cancer. 2008. pp. 497–511. doi:10.1038/nrc2402

54. Murohashi M, Hinohara K, Kuroda M, Isagawa T, Tsuji S, Kobayashi S, et al. Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells. Br J Cancer. 2010;102: 206–212.

55. Carey LA, Dees EC, Sawyer L, Gatti L, Moore DT, Collichio F, et al. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. Clin Cancer Res. 2007;13: 2329–2334.

56. Bose R, Kavuri SM, Searleman AC, Shen W, Shen D, Koboldt DC, et al. Activating HER2 mutations in HER2 gene amplification negative breast cancer. Cancer Discov. 2013;3: 224–237.

57. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res. 2017;77: e108–e110.

58. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res. 2020;48: W509–W514.

59. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12: 453–457.

60. Rodon J, Soria J-C, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med. 2019;25: 751–758.

61. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.

Cancer Discov. 2012;2: 401–404.

62. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38: 675–678.

63. Peng Q, Vijaya Satya R, Lewis M, Randad P, Wang Y. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC Genomics. 2015;16: 589.

64. Sun JX, He Y, Sanford E, Montesion M, Frampton GM, Vignot S, et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput Biol. 2018;14: e1005965.

65. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Keira Cheetham R. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics. 2012. pp. 1811–1817. doi:10.1093/bioinformatics/bts271

66. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15: 591–594.

67. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31: 213–219.

68. Wang Q, Kotoula V, Hsu P-C, Papadopoulou K, Ho JWK, Fountzilas G, et al. Comparison of somatic variant detection algorithms using Ion Torrent targeted deep sequencing data. BMC Med Genomics. 2019;12: 181.

69. Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S. Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine. OMICS. 2018;22: 630–636.

70. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18: 435–441.

71. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318: 2199–2210.

72. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med. 2018;24: 1559–1567.

73. Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nature Materials. 2019. pp. 410–414. doi:10.1038/s41563-019-0345-0

74. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339: 1546–1558.

75. Garraway LA, Verweij J, Ballman KV. Precision oncology: an overview. J Clin Oncol. 2013;31: 1803–1805.

76. Singh RR, Patel KP, Routbort MJ, Reddy NG, Barkoh BA, Handal B, et al. Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. J Mol Diagn. 2013;15: 607–622.

77. Deshpande A, Lang W, McDowell T, Sivakumar S, Zhang J, Wang J, et al. Strategies for identification of somatic variants using the Ion Torrent deep targeted sequencing platform. BMC Bioinformatics. 2018;19: 5.

78. Garofoli A, Paradiso V, Montazeri H, Jermann PM, Roma G, Tornillo L, 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: 884–894.

79. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12: e0177459.

80. Oh S, Geistlinger L, Ramos M, Morgan M, Waldron L, Riester M. Reliable analysis of clinical tumor-only whole exome sequencing data. doi:10.1101/552711

81. Schrader KA, Cheng DT, Joseph V, Prasad M, Walsh M, Zehir A, et al. Germline Variants in Targeted Tumor Sequencing Using Matched Normal DNA. JAMA Oncology. 2016. p. 104. doi:10.1001/jamaoncol.2015.5208

82. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

83. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

84. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38: e164.

85. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158.

86. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

87. Consortium T 1000 GP, The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015. pp. 68–74. doi:10.1038/nature15393

88. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res. 2017;45: D840–D845.

89. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493: 216–220.

90. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581: 434–443.

91. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34: 155–163.

92. Gao J, Chang MT, Johnsen HC, Gao SP, Sylvester BE, Sumer SO, et al. 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. 2017;9: 4.

93. Piscuoglio S, Ng CKY, Murray MP, Guerini-Rocco E, Martelotto LG, Geyer FC, et al. The Genomic Landscape of Male Breast Cancers. Clin Cancer Res. 2016;22: 4045–4056.

94. 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.

95. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14: 178–192.

96. Sikandar SS, Pate KT, Anderson S, Dizon D, Edwards RA, Waterman ML, et al. NOTCH signaling is required for formation and self-renewal of tumor-initiating cells and for repression of secretory cell differentiation in colon cancer. Cancer Res. 2010;70: 1469–1478.

97. Kruglyak KM, Lin E, Ong FS. Next-generation sequencing in precision oncology: challenges and opportunities. Expert Review of Molecular Diagnostics. 2014. pp. 635–637. doi:10.1586/14737159.2014.916213

98. Kadri S, Long BC, Mujacic I, Zhen CJ, Wurst MN, Sharma S, et al. Clinical Validation of a Next-Generation Sequencing Genomic Oncology Panel via Cross-Platform Benchmarking against Established Amplicon Sequencing Assays. J Mol Diagn. 2017;19: 43–56.

99. Ashley EA. Towards precision medicine. Nature Reviews Genetics. 2016. pp. 507–522. doi:10.1038/nrg.2016.86

100. Macconaill LE, Garraway LA. Clinical implications of the cancer genome. J Clin Oncol. 2010;28: 5219–5228.

101. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500: 415–421.

102. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Współczesna Onkologia. 2015. pp. 68–77. doi:10.5114/wo.2014.47136

103. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578: 82–93.

104. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458: 719–724.

105. Tsongalis GJ, Peterson JD, de Abreu FB, Tunkey CD, Gallagher TL, Strausbaugh LD, et al. Routine use of the Ion Torrent AmpliSeq™ Cancer Hotspot Panel for identification of clinically actionable somatic mutations. Clin Chem Lab Med. 2014;52: 707–714.

106. Frueh FW, Amur S, Mummaneni P, Epstein RS, Aubert RE, DeLuca TM, et al. Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. Pharmacotherapy. 2008;28: 992–998.

107. Stafford RS. Regulating Off-Label Drug Use — Rethinking the Role of the FDA. New

England Journal of Medicine. 2008. pp. 1427–1429. doi:10.1056/nejmp0802107

108. Levêque D. Off-label use of anticancer drugs. The Lancet Oncology. 2008. pp. 1102–1107. doi:10.1016/s1470-2045(08)70280-8

109. Subbiah V, Kreitman RJ, Wainberg ZA, Cho JY, Schellens JHM, Soria JC, et al. Dabrafenib and Trametinib Treatment in Patients With Locally Advanced or Metastatic BRAF V600–Mutant Anaplastic Thyroid Cancer. Journal of Clinical Oncology. 2018. pp. 7–13. doi:10.1200/jco.2017.73.6785

110. White PS, Pudusseri A, Lee SL, Eton O. Intermittent Dosing of Dabrafenib and Trametinib in Metastatic BRAFV600E Mutated Papillary Thyroid Cancer: Two Case Reports. Thyroid. 2017. pp. 1201–1205. doi:10.1089/thy.2017.0106

111. Odogwu L, Mathieu L, Blumenthal G, Larkins E, Goldberg KB, Griffin N, et al. FDA Approval Summary: Dabrafenib and Trametinib for the Treatment of Metastatic Non-Small Cell Lung Cancers Harboring BRAF V600E Mutations. The Oncologist. 2018. pp. 740–745. doi:10.1634/theoncologist.2017-0642

112. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Rep. 2018;23: 172–180.e3.

113. Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet. 2018;19: 93–109.

114. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016;375: 1109–1112.

115. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12: R41.

116. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005. pp. 301–320. doi:10.1111/j.1467-9868.2005.00503.x

117. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018. pp. 70–79. doi:10.1016/j.neucom.2017.11.077

118. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. SciPy; 2010. pp. 56–61.

119. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272.

120. Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd; 2013.

121. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47: D607–D613.

122. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002. pp. 321–

357. doi:10.1613/jair.953

123. Loi S, Haibe-Kains B, Majjaj S, Lallemand F, Durbecq V, Larsimont D, et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. Proc Natl Acad Sci U S A. 2010;107: 10208–10213.

124. Dietterich T. Overfitting and undercomputing in machine learning. ACM Computing Surveys. 1995. pp. 326–327. doi:10.1145/212094.212114

125. Juric D, Janku F, Rodón J, Burris HA, Mayer IA, Schuler M, et al. Alpelisib Plus Fulvestrant in PIK3CA-Altered and PIK3CA-Wild-Type Estrogen Receptor-Positive Advanced Breast Cancer: A Phase 1b Clinical Trial. JAMA Oncol. 2019;5: e184475.

126. André F, Ciruelos E, Rubovszky G, Campone M, Loibl S, Rugo HS, et al. Alpelisib for -Mutated, Hormone Receptor-Positive Advanced Breast Cancer. N Engl J Med. 2019;380: 1929–1940.

127. Bosch A, Li Z, Bergamaschi A, Ellis H, Toska E, Prat A, et al. PI3K inhibition results in enhanced estrogen receptor function and dependence in hormone receptor-positive breast cancer. Sci Transl Med. 2015;7: 283ra51.

128. Riquelme I, Tapia O, Espinoza JA, Leal P, Buchegger K, Sandoval A, et al. The Gene Expression Status of the PI3K/AKT/mTOR Pathway in Gastric Cancer Tissues and Cell Lines. Pathol Oncol Res. 2016;22: 797–805.

129. 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. doi:10.1200/PO.17.00011

130. Planchard D, Smit EF, Groen HJM, Mazieres J, Besse B, Helland Å, et al. Dabrafenib plus trametinib in patients with previously untreated BRAF-mutant metastatic non-small-cell lung cancer: an open-label, phase 2 trial. Lancet Oncol. 2017;18: 1307–1316.

131. Khoury MJ, Ioannidis JPA. Medicine. Big data meets public health. Science. 2014;346: 1054–1055.

132. Samowitz WS, Sweeney C, Herrick J, Albertsen H, Levin TR, Murtaugh MA, et al. Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. Cancer Res. 2005;65: 6063–6069.

133. Long GV, Flaherty KT, Stroyakovskiy D, Gogas H, Levchenko E, de Braud F, et al. Dabrafenib plus trametinib versus dabrafenib monotherapy in patients with metastatic BRAF V600E/K-mutant melanoma: long-term survival and safety analysis of a phase 3 study. Ann Oncol. 2017;28: 1631–1639.

134. Eberlein TJ. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. Yearbook of Surgery. 2012. pp. 353–356. doi:10.1016/j.ysur.2011.09.017

135. Encorafenib, Binimetinib, and Cetuximab in BRAF V600E–Mutated Colorectal Cancer. New England Journal of Medicine. 2020. pp. 876–878. doi:10.1056/nejmc1915676

136. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, et al. Criteria for the use of omics-based predictors in clinical trials: explanation and elaboration. BMC Med. 2013;11: 220.

Annex

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 nonoverlapping amplicons evenly distributed across the length of the genes. The designed panel was further inspected by

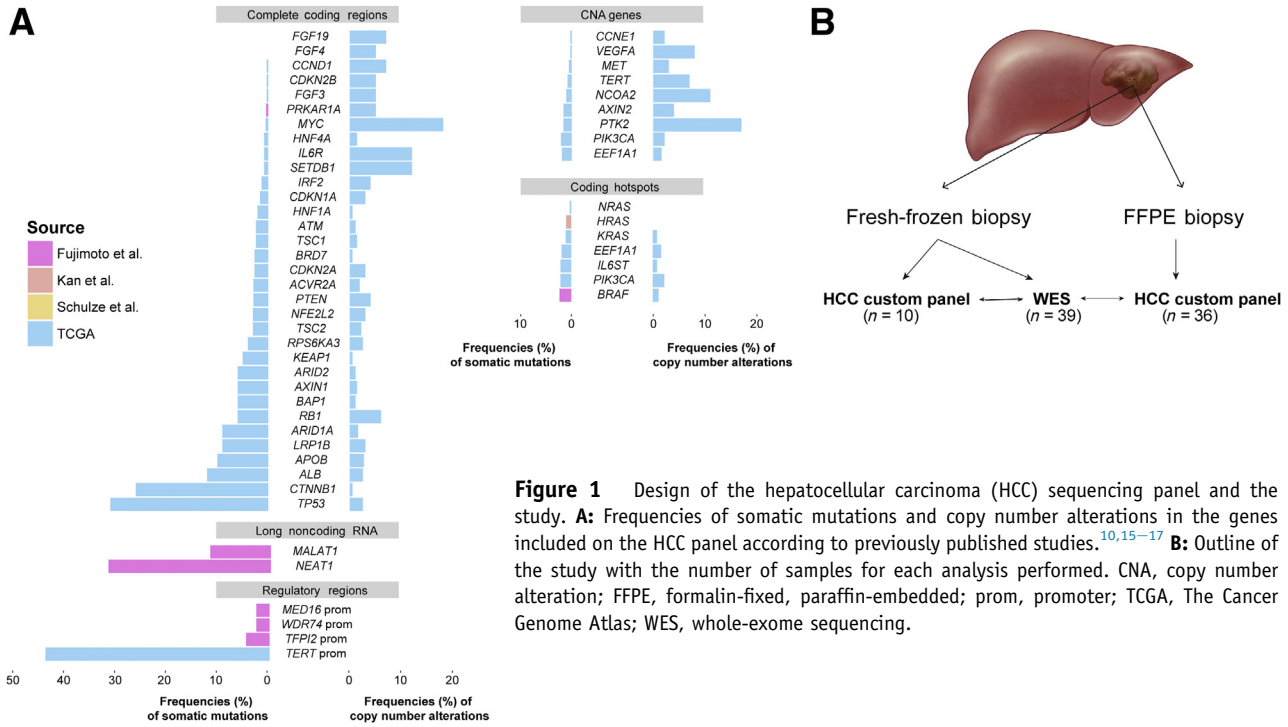


Figure 1 Design of the hepatocellular carcinoma (HCC) sequencing panel and the study. **A:** Frequencies of somatic mutations and copy number alterations in the genes included on the HCC panel according to previously published studies.[10,15–17] **B:** Outline of the study with the number of samples 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

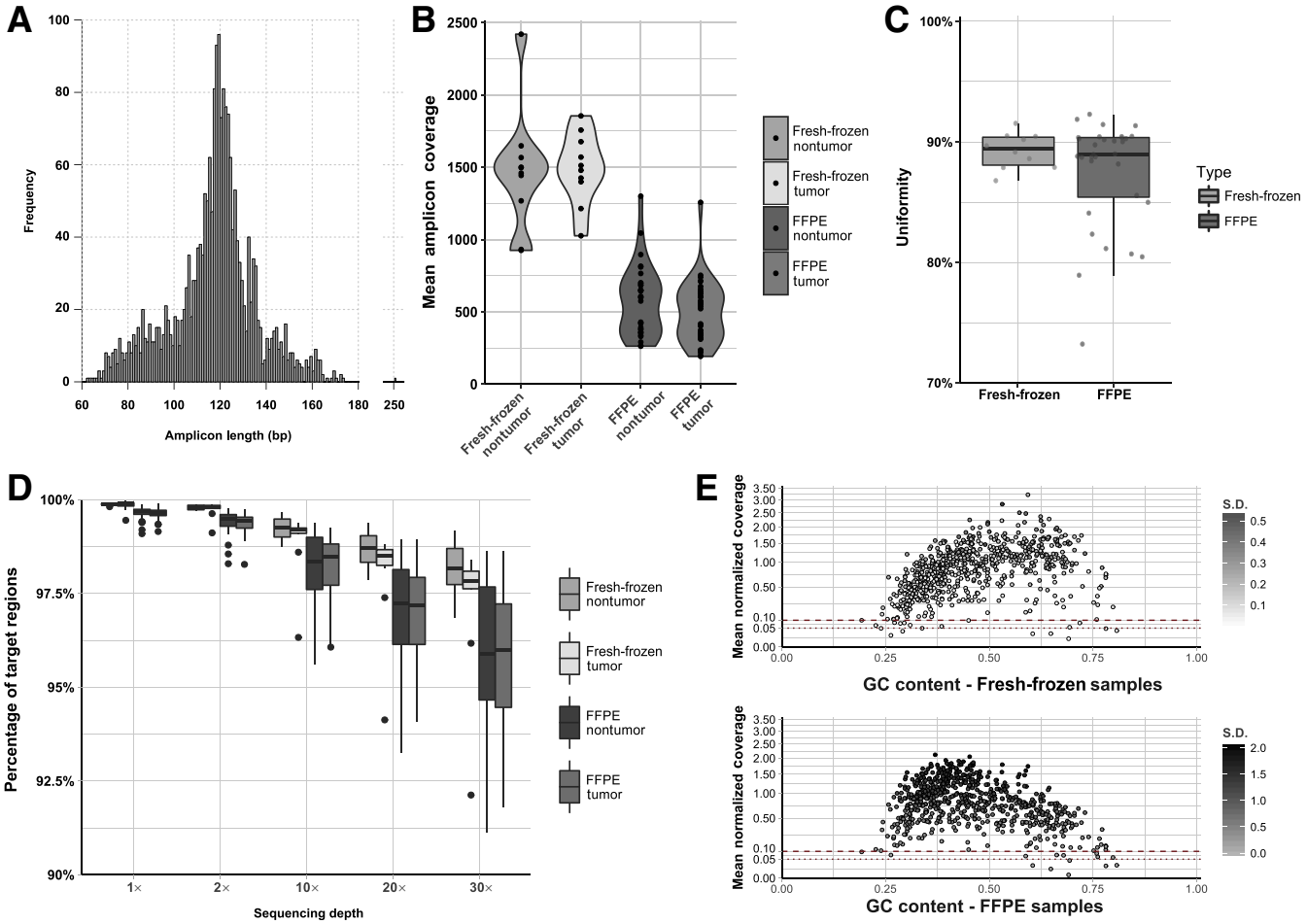


Figure 2 Coverage analyses and statistics of the hepatocellular carcinoma (HCC) panel. **A:** Distribution of the amplicon sizes on the HCC panel. **B:** Violin plots of the mean amplicon coverage across fresh-frozen nontumor; fresh-frozen tumor; formalin-fixed, paraffin-embedded (FFPE) nontumor; and 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.

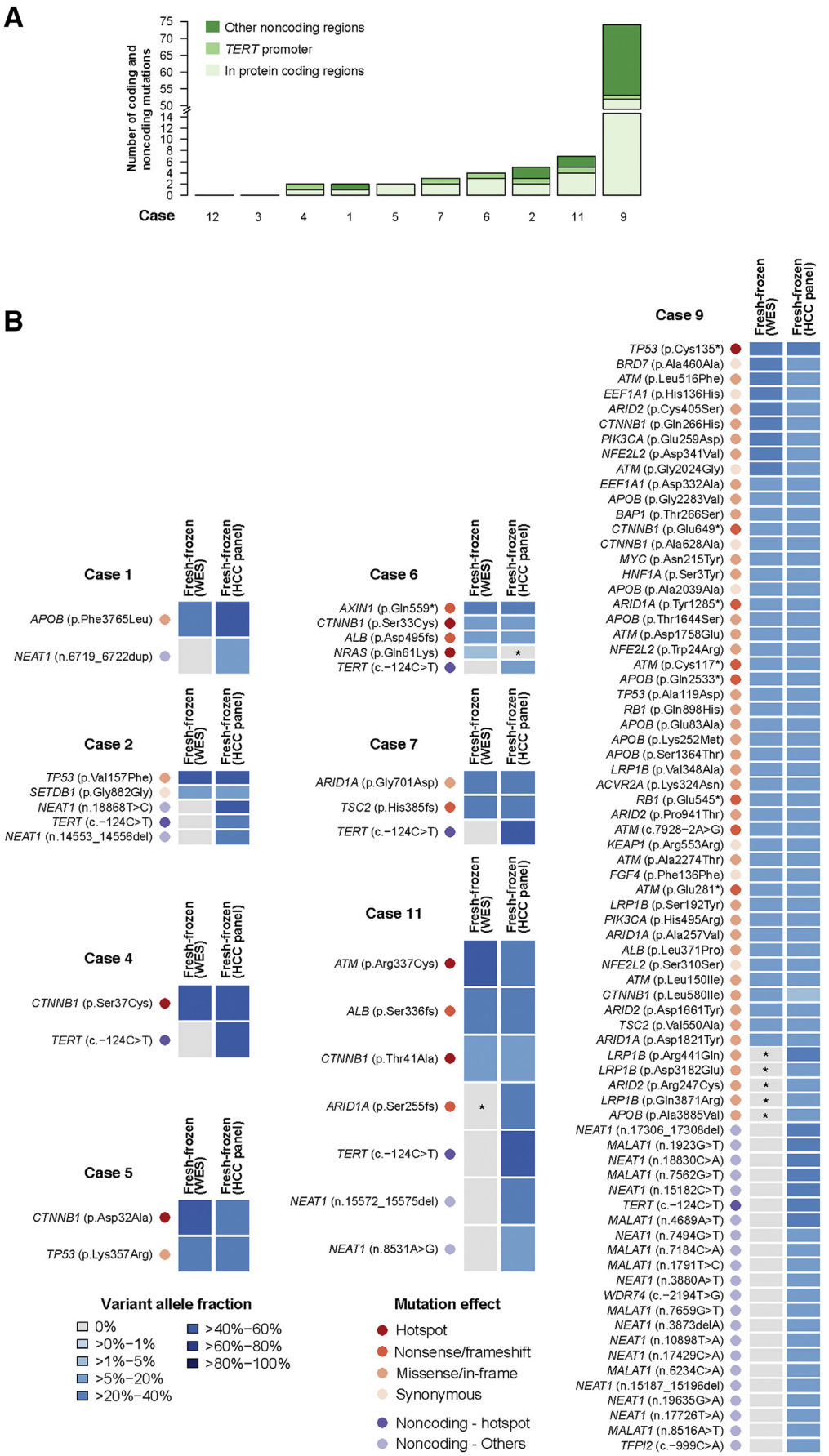
Compared with the WES analysis, the HCC panel analysis revealed an additional six mutations in the coding regions, including five in case 9 and one in case 11 (Figure 3B). Manual review of the WES data showed that all six mutations were in fact supported by at least one read in WES, but those positions 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
Other noncoding regions
TERT promoter
In protein coding regions
Number of coding and noncoding mutations
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**.

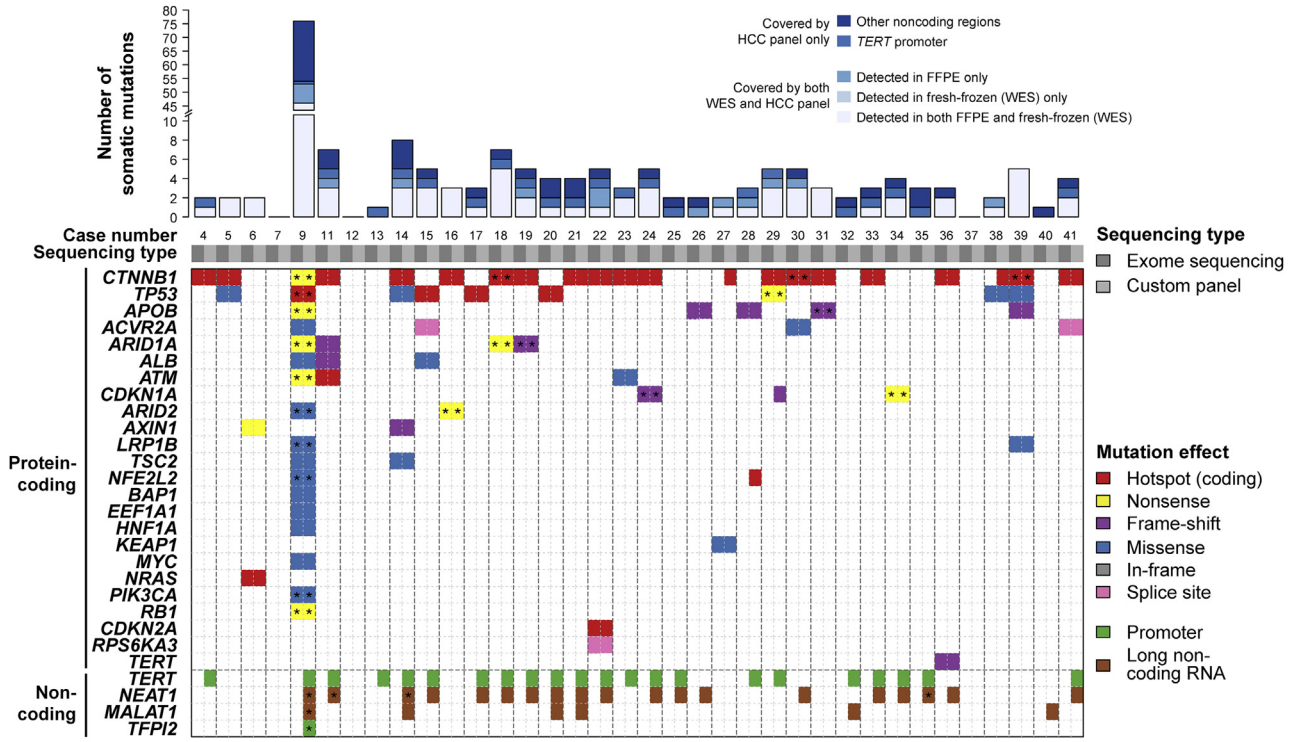


Figure 4 Comparison of somatic mutations defined by whole-exome sequencing (WES) and hepatocellular carcinoma (HCC) panel in formalin-fixed paraffin-embedded (FFPE) tissue. Barplot illustrates the number of somatic coding and noncoding mutations found in 36 FFPE tumor biopsies by using the HCC panel. 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

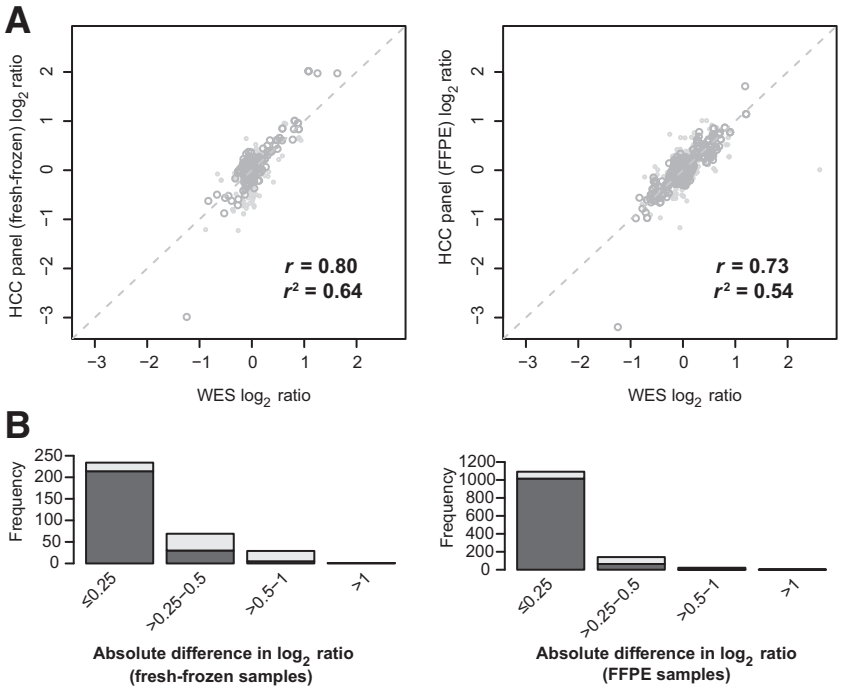


Figure 5 Copy number profiling by using the hepatocellular carcinoma (HCC) panel. **A:** Scatter plots illustrate the copy number $\log_2$ ratio of whole-exome sequencing (WES) and HCC panel sequencing of the fresh-frozen and the formalin-fixed, paraffin-embedded (FFPE) tumor samples. **B:** Barplots 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

Acknowledgments

There are many people who made it possible for me to complete this work.

I would like to thank all the members of the group, the present ones and the past ones, for being fellow adventurers in this new chapter of my life, in particular the other PhD students with whom I have shared the good and the bad of our trip toward the doctorate degree. Thanks to Dr. Piscuoglio for welcoming me in his group and for helping me navigate my way through my new life in the institute of pathology. A special thanks goes to Dr. Ng, for relentlessly teaching me so much about bioinformatics and, above all, for being someone who has always genuinely strived to be the best guidance she could be.

I would also like to thank Prof. Luigi M. Terracciano, Prof. Michael N. Hall and Prof. Julia E. Vogt for accepting to be part of my PhD committee, judge my work and for giving me new insights and different points of view during the past meetings.

Next, I want to thank the new friends I made here in Basel, who enriched my life and helped me survive during the last ~4 years, the ones I left in Terni, who make me feel like I never left whenever I go back home, and the ones all over the world, who were here even when they were not here.

Last, I thank my parents, brothers and my family, who love me like no one else and which I love like no one else, despite not saying it often enough. If it wasn't for you all, I would have not even begun this new chapter.